

Saint Petersburg State University
Graduate School of Management

Master of Business Analytics and Big Data

**BIG DATA ANALYTICS IN TRAVEL INDUSTRY: CASE OF AFFILIATE
MARKETING CHANNEL AT AVIASALES COMPANY**

Consulting project for Aviasales

Master's Thesis by the 2nd year students
Concentration — BM.5783.2019
Master in Business Analytics and Big Data

Markov Danil

Timerbaev Dmitry

Research Advisor:
Zhukova Sofia
Professor, Information Technologies in Management Department

Saint Petersburg

2021

ЗАЯВЛЕНИЕ О САМОСТОЯТЕЛЬНОМ ХАРАКТЕРЕ ВЫПОЛНЕНИЯ ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Мы, Марков Данил, студент 2 курса магистратуры ВМ.5783.2019 «Бизнес-аналитика и большие данные (Master in Business Analytics and Big Data – MiBA)», и Тимербаев Дмитрий, студент 2 курса магистратуры ВМ.5783.2019 «Бизнес-аналитика и большие данные (Master in Business Analytics and Big Data – MiBA)», подтверждаем, что в нашей магистерской диссертации на тему «Аналитика больших данных в индустрии путешествий: кейс аффилиатного маркетингового канала компании Aviasales», представленной в службу обеспечения программ магистратуры для последующей передачи в государственную аттестационную комиссию для публичной защиты, не содержится элементов плагиата.

Все прямые заимствования из печатных и электронных источников, а также из защищенных ранее курсовых и выпускных квалификационных работ, кандидатских и докторских диссертаций имеют соответствующие ссылки.

Нам известно содержание п. 9.7.1 Правил обучения по основным образовательным программам высшего и среднего профессионального образования в СПбГУ о том, что «ВКР выполняется индивидуально каждым студентом под руководством назначенного ему научного руководителя», и п. 51 Устава федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Санкт-Петербургский государственный университет» о том, что «студент подлежит отчислению из Санкт-Петербургского университета за представление курсовой или выпускной квалификационной работы, выполненной другим лицом (лицами)».

 01.06.2021

Марков Данил

 01.06.2021

Тимербаев Дмитрий

STATEMENT ABOUT THE INDEPENDENT CHARACTER OF THE MASTER THESIS

We, Markov Danil, 2nd year master student of the BM.5783.2019 master program “Master in Business Analytics and Big Data - MiBA”, and Timerbaev Dmitry, 2nd year master student of the BM.5783.2019 master program “Business Analytics and Big Data (Master in Business Analytics and Big Data - MiBA) ”, state that our master thesis on the topic “Big Data Analytics in Travel Industry: case of affiliate marketing channel at Aviasales company” which is presented to the Master Office to be submitted to the Official Defense Committee for the public defense, does not contain any elements of plagiarism.

All direct borrowings from printed and electronic sources, as well as from master theses, PhD and doctorate theses which were defended earlier, have appropriate references.

We are aware that according to paragraph 9.7.1. of Guidelines for instruction in major curriculum programs of higher and secondary professional education at St. Petersburg University «A master thesis must be completed by each of the degree candidates individually under the supervision of his or her advisor», and according to paragraph 51 of Charter of the Federal State Institution of Higher Professional Education Saint-Petersburg State University «a student can be expelled from St. Petersburg University for submitting of the course or graduation qualification work developed by other person (persons)».

 01.06.2021

Марков Данил

 01.06.2021

Тимербаев Дмитрий

АННОТАЦИЯ

Авторы	Марков Данил Тимербаев Дмитрий
Название магистерской диссертации	Аналитика больших данных в индустрии путешествий: кейс аффилиатного маркетингового канала компании Aviasales
Факультет	Высшая школа менеджмента
Направление подготовки	Бизнес-аналитика и большие данные
Год	2021
Научный руководитель	Жукова София
Описание цели, задач и основных результатов	Цель настоящего исследования заключалась в создании визуализации состояния глобального рынка аффилиатного (партнерского) маркетинга на рынке путешествий в 2020-2021 годах с целью оценки положения компании Aviasales относительно конкурентов. С использованием инструментов аналитики больших данных и машинного обучения в Python, была проведена обработка, моделирование и интеграция в единый массив данных, предоставленных компанией Aviasales. Финальный массив данных был использован для создания аналитического инструмента (дэшборда) в Power BI, с целью оценить характеристики и конкурентную среду мирового аффилиатного маркетинга в индустрии путешествий. Этот аналитический инструмент позволил создать управленческие рекомендации для компании Aviasales касательно развития их аффилиатной программы.
Ключевые слова	аффилиатный маркетинг, партнерский маркетинг, индустрия путешествий, аналитика больших данных

ABSTRACT

Master Student Names	Markov Danil Timerbaev Dmitry
Master Thesis Title	Big Data Analytics in Travel Industry: case of affiliate marketing channel at Aviasales company
Title Faculty	Graduate School of Management
Main field of study	Business Analytics and Big Data
Year	2021
Academic Advisor's Name	Zhukova Sofia
Description of the goals, tasks and main results	The goal of this research was to visualize the state of the global travel affiliate market 2020-2021 in order to assess the competitive position of Aviasales relative to its competitors. With the use of Python's big data analytics and machine learning tools, the data provided by Aviasales company was processed, modelled and integrated into a single dataset. Final dataset was used to create an analytic tool (dashboard) in Power BI for assessing the global affiliate marketing characteristics and competitive landscape. This analytic tool allowed for making managerial recommendations for the Aviasales company regarding the development of their affiliate program.
Keywords	affiliate marketing, travel industry, big data analytics

TABLE OF CONTENTS

Introduction	8
Chapter 1. Business Understanding	12
1.1. Description of affiliate marketing channel	12
1.2. Company and industry information	13
1.3. Research problems and goal	15
1.3.1. Gap analysis	15
1.3.2. Research goal & tasks	15
1.4. Assessment of situation	16
1.4.1. Research assumptions & limitations	16
1.4.2. Overview of available IT resources	17
1.4.3. Data problems	18
1.5. Project specifications	19
1.5.1. Data mining outputs expectations	19
1.5.2. Research objectives	19
1.6. Project plan	23
1.6.1. Project timeline	23
1.6.2. Project team	23
1.6.3. Research framework	24
1.6.4. Methodological framework	24
Chapter 2. Data Mining & BI Tool Development Process	25
2.1. Data description and exploration	25
2.1.1. General overview	25
2.1.2. Data specification	26
2.1.3. Primary data description	27
2.1.4. Supplementary data description	30
2.1.5. Data problems (detailed description)	32
2.2. Data construction of the direct advertisers' datasets	34
2.3. Model building for travel advertisers' classification	35
2.3.1. Selecting modeling technique	36
2.3.2. Model design	39
2.3.3. Model building	41
2.3.4. Model assessment	42
2.4. Data construction for network advertisers' datasets	43
2.4.1. Domain search by deep link analysis	43
2.4.2. Domain search by advertiser ID (key)	44
2.4.3. Travel class retrieval from affiliate network links	45
2.5. Data construction of the travel verticals classification	46
2.6. Data integration of the final dataset	58
	6

2.6.1. Standardization of datasets	59
2.6.2. Join of datasets	60
2.7. BI dashboard preparation	61
2.7.1. Defining affiliate market characteristics	61
2.7.2. BI tool requirements	62
2.7.3. BI tool framework selection	63
2.7.4. Dashboard design	65
2.7.5. Dashboard content	68
2.8. Conclusions on data mining & BI tool development process	77
Chapter 3. Evaluation and Deployment	77
3.1. Validation of results	78
3.2. Affiliate market analysis	80
3.3. Competitive analysis	86
3.3.1. Competitive quadrants	86
3.3.2. Aviasales competition	89
3.3.3. Aviasales affiliate ecosystem	93
3.4. Business recommendations	97
Conclusion	98
References	100
Appendix	103

Introduction

More and more travelers are searching for airplane tickets, hotels or car rentals using the Internet creating a huge demand for the services and products offered by online travel companies. Oxford Economics estimated the global tourism industry to grow 3.9% annually over the next decade, while a study conducted by Deloitte pointed out that digital technologies would open new opportunities for emerging travel companies and redefine customers' experience. (Travelpayouts Blog, 2017)

Competition among online travel companies is getting tougher as more players are entering the market offering their customers better prices and wider selection of available products/services. Tough competition requires travel tech companies to look for various ways of marketing their services and generating additional conversion. After accounting for the COVID-19 pandemic impact on the travel industry, the competition can be expected to increase even more.

One of the channels that travel brands actively employ for promotion is affiliate marketing. Affiliate marketing is a type of online marketing whereby a firm (an advertiser or a merchant) makes an agreement with another firm/individual (a publisher or an affiliate) to feature a link from its websites on affiliates sites.

The goals of affiliate marketing are to promote the product and generate sales through additional distribution channels, as well as to increase web traffic to the advertiser's website in exchange for commission. Commission payable to affiliates is determined by the individual advertiser rewards model, and is usually based on a certain percentage of sale generated by the affiliate. (Dwivedi, et al., 2017) Affiliate programs are employed by a vast majority of travel tech companies, allowing them to generate additional traffic and revenue at comparatively low cost in a highly competitive and price-sensitive environment of the travel industry. That's why travel companies should thoroughly analyze the potential gains of launching their affiliate programs and understand the current competitive landscape within affiliate marketing channel.

Despite the fact that affiliate marketing existed for some time, there are very few academic papers that were written in the last years regarding analysis and research of the affiliate marketing channel. The first academic literature review was conducted only in 2017 and most of the available studies relied on limited data. (Dwivedi, et al., 2017) Literature review conducted by Dwivedi, et al. in 2017 showed that most of the sources used by researchers were either purely conceptual or included limited case studies, interviews and secondary data. Few academic studies and lack of quality data clearly shows that the features of an affiliate marketing channel in the travel industry are poorly researched and understood.

This research provides an expanded overview on the state of an affiliate marketing channel in the travel industry compared to other studies. It relies on the big data analysis of major characteristics of the affiliate marketing channel in the travel industry and accounts for the effect of the pandemic.

The main research goal of this study was to visualize the state of the global travel affiliate market 2020-2021 in order to assess the position of Aviasales relative to its competitors using data collected by Aviasales team.

Because the data was unstructured, disintegrated and lacked certain descriptive parameters, in order to achieve the research goal, it was decided to initiate the project to tackle the following tasks:

1. Combine all datasets into single data file for further analysis by employing data cleaning, data construction and data integration procedures (data mining stage)
2. Choose characteristics to analyze and appropriate visualizations; prepare a BI dashboard based on combined data file and with the chosen visualizations (visualizations development stage)
3. Analyze broad affiliate marketing conditions and Aviasales competitive position to derive managerial implications (data analytics stage)

In pursuing those tasks, various tools of data mining and data analytics were employed, including web cloud storage (Amazon Web Services), JupyterHub environment with Python libraries (NLTK, Scikit-learn, PySpark) and business intelligence software (Microsoft Power BI).

This research has theoretical as well as practical value. In terms of theoretical value, it is the first quantitative assessment of an affiliate marketing channel in the travel industry in academia. As for the practical value, this research formed the basis for the analysis of the competitive environment in travel affiliate marketing that can be used by the management of the Aviasales company for better decision-making and understanding of the channel.

Advertisers and affiliates in affiliate marketing were the objects in this study, while general processes, parameters and characteristics of the advertisers and affiliates were the subject.

Overall approach for structuring work related to data mining was defined by the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. In addition, several natural language processing and machine learning methods were used for modeling purposes, including text tokenization and lemmatization, term-frequency inverse-document-frequency and k-means clustering algorithms.

This research relied on primary data on affiliate links and advertisers provided by Aviasales company. Theoretical background regarding affiliate marketing, data mining, natural language processing and machine learning methods were supplied by the academic papers, articles and studies. Frameworks and methodology for analyzing an affiliate marketing channel were developed in collaboration with Aviasales industry experts. Finally, this research relied on technical documentation for Python programming language NLP, ML and Big Data libraries, such as NLTK, Scikit-learn and PySpark.

Structurally, the research is divided into three chapters. The first chapter provides an overview of the business-related background. It defines research the goal, tasks, objectives, limitations and resources for all stages of the project - data mining, visualization development and data analytics. The second chapter provides a detailed overview of the data mining and visualization development work completed. The last chapter is focused on utilizing the created BI dashboard to describe global travel affiliate market characteristics and develop managerial recommendations for Aviasales based on the competition analysis.

Data mining work for the research is organized according to the CRISP-DM methodology principles outlined in the work of Chapman, et. al in 2000 (see Table 1 below).

Table 1. Thesis structure within CRISP-DM methodology

CRISP-DM Phase	CRISP-DM tasks with related thesis chapter
Business Understanding	Determine Business Objectives 1.3. Research problems and goal Assess Situation 1.4. Assessment of situation Determine Data Mining Goals 1.5. Project specifications Produce Project Plan 1.6. Project plan
Data Understanding	Describe & Explore Data Verify Data Quality 2.1. Data description and exploration
Data Preparation	Select & Clean Data 2.3. Model building for travel advertisers' classification Construct, Integrate & Format Data 2.2. Data construction of the direct advertisers' datasets 2.4. Data construction for network advertisers' datasets 2.5. Data construction of the travel verticals classification 2.6. Data integration of the final dataset
Modeling	Select Modeling Technique & Generate Test Design Build & Assess Model 2.3. Model building for travel advertisers' classification

Evaluation	Evaluate Results, Review Process 2.8. Conclusions on data mining & BI tool development process 3.1. Validation of results
Deployment	Plan Deployment & Produce Final Report 2.7. BI dashboard preparation 3.2. Affiliate market analysis 3.3. Competitive analysis 3.4. Business recommendations

Source: [authors research]

The project was collectively performed by the authors of the paper - Dmitry Timerbaev and Danil Markov. Each task was addressed in a collaborative way, meaning that there was no strict division of project work by the authors. Table 2 below represents the main types of work and the share estimate of workload invested in them by the authors.

Table 2. Overview of the project workload completed by each author

Type of work	Student	Approx. workload
Data description and exploration	Dmitry Timerbaev	50%
	Danil Markov	50%
Datasets construction	Dmitry Timerbaev	50%
	Danil Markov	50%
ML model building	Dmitry Timerbaev	50%
	Danil Markov	50%
Methodological frameworks development	Dmitry Timerbaev	60%
	Danil Markov	40%
Dataset integration	Dmitry Timerbaev	40%
	Danil Markov	60%
BI dashboard preparation	Dmitry Timerbaev	40%
	Danil Markov	60%
Validation of results, evaluation & recommendations	Dmitry Timerbaev	50%
	Danil Markov	50%

Source: [authors research]

Chapter 1. Business Understanding

1.1. Description of affiliate marketing channel

Affiliate marketing can be defined as a type of online marketing, where one company (known as an advertiser) signs an agreement with another company (known as an affiliate) to include a link from its websites on the affiliate's website in exchange for commission if the customer follows the link and makes the purchase. The goals of affiliate marketing are to promote and sell products or services through the additional channel, increase web traffic to advertisers, and generate additional conversion. (Dwivedi, et.al., 2017). Information technology made affiliate marketing a relatively inexpensive way for the online firms to earn additional profit and increase awareness of their brand. Because of the commission based financial model, affiliate marketing is considered an exceptionally cost-efficient channel for the advertisers (Jurisova, 2013).

According to Hee and Patrick, 2019, the major factors that influence advertisers to use an affiliate marketing channel are relative advantages of affiliate marketing, observability and compatibility. Affiliate marketing offers relative advantage over other online marketing tools – in terms of cost-efficiency, large scope and low complexity. Observability is important as results of using affiliate marketing are clearly observable and thus more travel advertisers decide to implement it. Finally, compatibility refers to the fact that affiliate marketing fits well within the overall marketing strategy of most advertisers, and does not require significant investment in infrastructure and development (Patrick, Hee, 2019).

The key players within the affiliate marketing channel are advertisers, affiliates and affiliate networks. Advertisers are companies (or brands) that offer a product or a service and conduct their business through the internet. Affiliates are the marketing partners of the advertisers. Any website can be an affiliate (blog, review website, informational website, coupon website, etc.). Affiliate networks are firms that operate affiliate programs on behalf of their clients (advertisers). Affiliate networks give advertisers access to their base of affiliates and allow affiliates to join various brands' affiliate programs, acting as intermediaries within this marketing channel (Acceleration Partners, 2017). Certain advertisers may choose to establish the affiliate program on their own platform and attract affiliates directly (those advertisers are called direct advertisers), while others may choose to use the services of affiliate networks (thus they are called network advertisers). The interaction between key players in the affiliate marketing channel is summarized in the Figure 1 below.

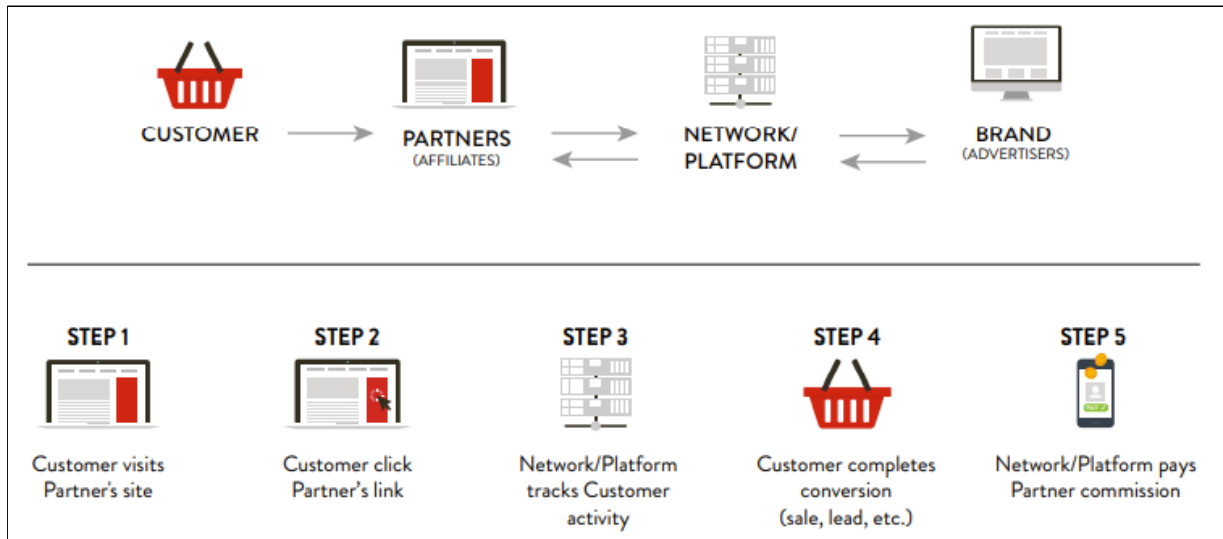


Figure 1. Interactions within affiliate marketing channel

Source: [Acceleration Partners, 2017]

The literature review conducted by Dwivedi, et.al. in 2017 concluded that the amount of research in the area of affiliate marketing is limited. Most of the current research consists of case studies, conceptual studies and studies based on secondary data sources. Few studies involved primary data sources (interviews and questionnaires), and there was no quantitative analysis based on the actual web data (Dwivedi, et.al., 2017). No research on affiliate marketing in the travel industry has ever been conducted before.

1.2. Company and industry information

Aviasales is a Russian online travel metasearch engine founded in 2007 that provides airline ticket offerings aggregation, allowing its customers to search for the best ticket deals across multiple airlines. Aviasales is a major player in Russia and CIS countries and has an audience of 15 million users per month as of 2020. Aviasales has been developing its affiliate program since 2011 through Travelpayouts affiliate network. Since 2011, Aviasales has paid over 1.6 billion rubles of commissions to its affiliates (Aviasales, n.d.).

On a global scale, travel affiliate marketing is represented by a large number of well-known travel brands, such as Booking, Airbnb, Expedia and others. Those brands operate in various travel sub-industries, also known as travel verticals. Vertical is a term used to describe a specific industry that focuses on a particular market niche (Indeed, 2021). Affiliate marketing is featured in many global travel verticals, such as hotel booking, airline tickets, car rentals, and cruises.

At the moment, Aviasales is trying to understand its place in the global travel affiliate market – how it performs against other travel brands and what niches are there within the affiliate channel. This information should allow Aviasales management team to understand the effectiveness of the current affiliate program and if there are any available options for improvement. In pursuing this goal, Aviasales team has gathered large amounts of unique data on global travel affiliate links and advertisers from the worldwide web. However, this data is unstructured and deficient – it cannot be used in its current state to perform any type of analysis or drawing any managerial implications.

Brief overview of the datasets gathered by Aviasales is presented on Table 3 below.

Table 3. Aviasales dataset collections overview

Dataset collection	ID	Description	Number of files
Data on all travel industry direct advertisers' redirect web links	1	Contains parameters of all travel industry direct advertisers' redirect web links (such as ahrefs rank, traffic, number of backlinks, etc)	123 CSV files
Data on all network advertisers' affiliate links placed through world's largest affiliate networks	2	Contains parameters of all network advertisers' affiliate web links (such as ahrefs rank, traffic, number of backlinks, etc.)	28 CSV files
Data on the text content of homepages of all network advertisers placed through world's largest affiliate networks	3	Contains parsed HTML data of the text presented on all network advertisers' homepages	809 CSV files
Data on network advertisers product categories	4	Contains data provided by the affiliate networks (Travelpayouts, Admitad and others) on the product categories to which each participating advertiser belongs (such as travel, clothing, consumer goods, etc.)	4 CSV files
Advertiser IDs (keys) data for all network advertisers	5	Contains data on all active advertiser IDs or keys (item within the URL used by the affiliate network to identify specific participating advertiser) for a number of affiliate networks (CityAds, TradeTracker, Awin, CJ, Shareasales)	7 CSV files

Source: [authors research]

1.3. Research problems and goal

1.3.1. Gap analysis

After the analysis of research aspects of affiliate marketing channel, certain research gaps were identified:

- Few academic studies conducted in the field of affiliate marketing
- No quantitative parameters to analyze the market are available
- Current academic understanding of an affiliate marketing channel is based on limited case studies and surveys (Dwivedi, et. al., 2017)

In addition to research gaps, certain problems with Aviasales data were identified. The following problems are compromising Aviasales's management ability to assess company's position within global affiliate marketing landscape using the collected datasets:

- Data gathered by Aviasales is unstructured and divided into multiple datasets - it should be integrated into a single data model.
- Data lacks certain descriptive parameters that can be used to analyze affiliate marketing - such as advertisers travel vertical type and whether advertisers belong to travel industry or not
- Data is raw and unorganized - there is a need to define characteristics that describe affiliate market and BI tool for visualizing the data.

1.3.2. Research goal & tasks

To expand the academic knowledge base of affiliate marketing channel in travel industry and solve the business problem of Aviasales, this research will be focusing on addressing the following research goal:

- **Research goal** – visualize the state of the global travel affiliate market 2020-2021 in order to assess the competitive position of Aviasales relative to its competitors using datasets collected by Aviasales team

To achieve the stated research goal, several tasks have to be fulfilled during the research process:

- Research task 1. Create the unified and structured data model
- Research task 2. Visualize the affiliate market with the use of BI tool
- Research task 3. Use created BI tool (dashboard) to analyze the global affiliate market and the competitive position of Aviasales

In order to achieve the stated research goal and complete research tasks, it was decided to initiate a project that includes both data mining and BI dashboard development tasks (Research tasks 1,2) with the results to be used to fulfill Research task 3. The process of data mining will be guided by the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. CRISP-DM is an industry-wide standard for conducting data mining projects, thus the following research will be structured according to the CRISP-DM principles (Chapman, et. al, 2000).

1.4. Assessment of situation

Before developing a list of objectives that have to be completed to address the research tasks, it is necessary to outline general research assumptions, methodological limitations, IT resources and data problems.

1.4.1. Research assumptions & limitations

The following is the list of assumptions that were accepted at the start of the project:

- Samples presented in the Aviasales dataset collections are unbiased, relatively full and relatively representative of the global travel affiliate market, meaning that average parameters and conceptual conclusions drawn from the sample should be statistically significant for the population as well.
- Effect of the COVID-19 pandemic is already within the data as the samples were collected during Fall 2020. It is also assumed that the COVID-19 pandemic affected all companies and affiliates within the travel industry to the roughly same degree.
- An affiliate marketing channel in the travel industry is relatively stable in the middle-run and is expected to change proportionally to the change in overall travel industry conditions. In addition, although the numbers of active affiliates and links may change day-to-day, on average, broad characteristics and trends of the affiliate marketing channel in the travel industry are expected to be relatively static and change over time proportionally to the change in overall travel industry conditions.

Limitations, risk level associated with each of them to the research integrity and options to mitigate those risks were also identified (see Table 4 below).

Table 4. Initial research limitations overview

Limitation	Risk level	Risk mitigation factors and options
Preliminary assessment of datasets showed that the majority of links, affiliates and advertisers belong to the regions of North America, Europe, Middle East, Latin America, India and Oceania. Data lacks significant representation of the East Asian and African markets.	Medium	The limitation is acknowledged in one of the assumptions. For the research purposes, it is assumed that data is only relatively full. This limitation is not violating the assumption as the vast majority of travel brands and regional markets with significant impact to the global travel affiliate channel are still covered by the available data. Overall conclusions drawn from the sample are not expected to change drastically if East Asian and African markets were to be taken into account
Certain links, affiliates or brands may not be active by the time this research is completed, thus reducing the correctness and relevance of the information presented.	Medium	The limitation is acknowledged in one of the assumptions. For the research purposes and simplicity, it is assumed that the travel affiliate channel is relatively stable and all meaningful changes occur within the range of the overall travel industry fluctuations. Dynamic nature of the world wide web indexing makes it impossible to account for such changes within the scope of this research. However, given the large number of links, affiliates and brands involved in the analysis, possible changes as a proportion of total composition should be negligible. Impact on the results and overall conclusions drawn from the analysis should thus be within the expected confidence interval.
Data may contain certain erroneous or unrepresentative items due to data collection, modeling, processing or integration deficiencies and constraints	Low	To mitigate this problem, the aggregated results of the data analysis should be constantly checked for problems. Consultations with Aviasales specialists for additional viability checks can also be conducted. Overall, given the amount and complexity of the sample, isolated data deficiencies are not expected to affect the general research conclusions in any significant manner.

Source: [authors research]

1.4.2. Overview of available IT resources

Regarding the resources for data mining and analytics, there were several employed for this research (see Table 5 below).

Table 5. IT resources used for the research

Resource	Description	Application
Amazon Web Services Cloud storage	On-demand cloud computing and data storage platform	AWS cloud storage was used to store the Aviasales datasets, as well as temporary data files generated during data mining process
Jupyter Hub with Spark environment	Web-based interactive development environment (IDE) for Jupyter notebooks, code, and data	Jupyter Hub notebooks were used as a computational environment for machine learning and data mining tasks in Python programming language and its libraries, including, but not limited to NLTK, Scipy, Numpy, Scikit-learn, Matplotlib, Pandas, PySpark
Microsoft Power BI	Business intelligence and interactive visualization software	Microsoft Power BI was used as a platform for the implementation of final BI dashboard
Microsoft Excel & Google Sheets	Spreadsheet programs	Spreadsheet programs were used for miscellaneous data work
Atlassian Confluence	Web-based collaboration software	Confluence was used to coordinate and organize project workflow

Source: [authors research]

1.4.3. Data problems

Finally, after identifying research assumptions, limitations and tools to be used, the assessment of related dataset collections' problems was conducted. Table 6 below summarizes the problems with the dataset collections which ultimately translated into the definition of research objectives and the required outputs (for detailed description of data problems see 2.1.5. "Data problems (detailed description)").

Table 6. Summary of the problems within Aviasales dataset collections

Dataset collection №	Problems
1	<ul style="list-style-type: none"> • Not all direct links within the datasets are affiliate links • Markers for affiliate links are not identified
2	<ul style="list-style-type: none"> • Not all affiliate network links belong to the travel industry brands • Advertiser domain is not clearly identifiable
1,2	<ul style="list-style-type: none"> • Datasets lack information on advertisers' travel vertical
3	<ul style="list-style-type: none"> • Texts require preprocessing • Sample imbalance exists, as there are not many texts of the travel

	industry brands within the dataset <ul style="list-style-type: none"> • Dataset contains text in multiple languages
4	<ul style="list-style-type: none"> • Not all relevant travel brands are categorized
5	<ul style="list-style-type: none"> • Not all advertiser IDs are within the available datasets

Source: [authors research]

1.5. Project specifications

1.5.1. Data mining outputs expectations

After all assumptions, limitations, resources and data problems were assessed, the list of data outputs needed as a result of the data mining (Research task 1) was prepared in relation to dataset collections or previously created outputs involved (see Table 7 below).

Table 7. Overview of the outputs required for completing Research task 1

Data mining output needed	Output dataset №	Dataset collection № and/or Output dataset № to be used
Dataset that includes all affiliate web links and their parameters for direct advertisers	1	Dataset collection № 1
Dataset that includes a list of all travel-related network advertisers domains	2	Dataset collections № 3,4
Dataset that includes all affiliate web links and their parameters for network advertisers	3	Dataset collections № 2,5; Output dataset № 2
Dataset that includes classification of all travel brands found in output № 1 and 2 by travel vertical	4	Output datasets № 1,2
Unified dataset with cleaned and standardized data that includes outputs № 1, 3, 4	5	Output datasets № 1,3,4

Source: [authors research]

1.5.2. Research objectives

Finally, after listing all required outputs at the data mining stage (Research task 1), the list of all objectives for the whole project was prepared. Table 8 below contains a numbered list of objectives, the output that is expected to be generated as a result of the objective completion, the brief description of the objective, its constraints from technical perspective and the success criteria.

Table 8. Overview of the project objectives, constraints and success criteria

Research Task 1 - Create the unified and structured data model (Data Mining Stage)				
Objective №	Objective output(s)	Objective description	Technical constraints	Success criteria
1	Output dataset № 1	Retrieve affiliate link markers from the datasets with direct advertisers' redirect links (dataset collection № 1). Use these markers to distinguish affiliate links from non-affiliate redirect links. Prepare output dataset № 1 based on retrieved affiliate markers.	None	Affiliate link markers for all direct advertisers identified and tested for correctness with the help of Aviasales specialists. Output dataset № 1 created.
2	Output dataset № 2	Create an unsupervised text classifier to sort out network advertisers related to the travel industry from those that are unrelated. (using dataset collection № 3) For the advertisers that cannot be included into the classifier due to language related constraints (see Table 17 for details) – make best effort to classify them based on the product category data provided by the affiliate networks (dataset collection № 4). Prepare output dataset № 2 based on resulting classification.	Time and computing power constraints	All network advertisers must be classified as related/not related to travel industry. Accuracy is assessed by randomly selecting a sample of 100 advertiser texts within a predicted grouping 3 times and manually checking if predicted class is correct. More than 95% of advertisers picked for each randomized check must be classified correctly for the classifier prediction to be valid. Output dataset № 2 created.

3	Output dataset № 3	Perform preprocessing of affiliate network links (dataset collection № 2) to retrieve advertisers' domain names using either deep link analysis or advertiser ID data (dataset collection № 5). Merge output dataset № 2 with the affiliate links datasets for network advertisers (dataset collection № 2). Prepare output dataset № 3 as a result.	Time and computing power constraints	All datasets are preprocessed and merged correctly without any erroneous output. Output dataset № 3 created.
4	Output dataset № 4	Create a methodology for classifying advertisers by travel vertical, as there is no industry accepted methodology for vertical classification. Classify travel advertisers – both direct and network from output datasets № 1 and 2. Prepare output dataset № 4 as a result.	None	Created methodology is approved by Aviasales specialists. All travel brands are assigned to travel vertical correctly according to developed methodology. Output dataset № 4 created.
5	Output dataset № 5	Merge output datasets № 1, 3, 4. Standardize and clean the data. Prepare output dataset № 5 as a result.	Time and computing power constraints	All datasets are merged correctly. All necessary data cleaning and standardization completed. Output dataset № 5 created.
Research Task 2 - Visualize the affiliate market with the use of BI tool (Visualization Development Stage)				
6	List of questions, indicators and competitive metrics that can be used for market and competitive landscape analysis	Collaborate with Aviasales industry experts to select characteristics that can be used to describe global affiliate market	None	All characteristics are approved by Aviasales industry experts

7	List of visualizations to be used in BI dashboard. List of technical and functional requirements for the BI tool and dashboard.	Choose the appropriate visualizations to represent defined characteristics from objective № 6. Define technical and functional requirements of the BI tool and dashboard.	None	All visualizations are appropriate and approved by Aviasales industry experts. Technical and functional requirements for BI tool and dashboard are clearly defined by Aviasales representatives.
8	Description of analytical framework(s) to be used for analysis	Prepare analytical framework(s) for company's competitive segmentation	None	Developed framework(s) are approved by Aviasales industry experts
9	Developed dashboard	Choose appropriate BI tool and develop dashboard with visualizations and frameworks defined in objectives № 7,8	None	BI tool and final dashboard satisfy company's requirements
Research Task 3 - Use created BI dashboard to analyze the global affiliate market and the competitive position of Aviasales (Data Analytics Stage)				
10	Written analysis answering questions and defining indicators outlined in output of objective № 6	Analyze characteristics of global affiliate market in travel defined in objective № 6 using the BI dashboard	None	All questions and indicators outlined in output of objective № 6 are covered
11	Written analysis defining Aviasales competitive position using frameworks from output of objective № 8, as well as the managerial recommendations	Use the BI dashboard to define competitive position of Aviasales in the market using the frameworks developed in objective № 8. Draw managerial recommendations from the conducted analysis.	None	All developed frameworks are used in the analysis
12	Result approval by Aviasales representatives	Present the results (outputs of objectives № 9, 10, 11) to Aviasales and project sponsors	None	Presented results are approved by Aviasales representatives

Source: [authors research]

1.6. Project plan

1.6.1. Project timeline

Our project plan for completing research objectives can be summarized by the Gantt chart below (Figure 2). Blue lines represent data mining objectives (Research task 1), orange lines represent visualization development objectives (Research task 2), and green lines represent data analytics objectives (Research task 3). Numbers of thesis chapters where objectives were covered are also included. The whole project was planned to be completed in 30 weeks.

Objective №	Chapter №	Objectives	Weeks																																															
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30																		
	2.	Research task 1: Create the unified and structured data model	█																																															
1	2.2.	Filter affiliate links from non-affiliate in direct advertisers datasets	█	█	█																																													
2	2.3.	Clusterize travel advertisers from non-travel in network advertisers datasets		█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█																	
3	2.4.	Retrieve advertisers' domain names in network advertisers datasets																																																
4	2.5.	Prepare methodology for travel vertical classification and classify all advertisers																																																
5	2.6.	Integrate all the datasets in clean and standardized format																																																
	2.7.	Research task 2: Visualize the affiliate market																																																
6	2.7.1.	Select the characteristics for the analysis of affiliate market																																																
7	2.7.2.	Choose the appropriate visualization types for characteristics visualization																																																
8	2.7.5.; 3.3.1.	Prepare the analytical framework for company's competitive segmentation																																																
9	2.7.5.	Develop the BI dashboard for the market visualization																																																
	3.	Research task 3: Apply the BI tool to analyze the global affiliate market and the competitive position of Aviasales																																																
10	3.2.	Analyze the characteristics of global affiliate market in travel																																																
11	3.3.	Define the competitive position of Aviasales on the market																																																
12	-	Present the results to Aviasales team																																																

Figure 2. Data mining project schedule (Gantt chart)

Source: [authors research]

1.6.2. Project team

Table 9 represents the main participants in the research project, information about their roles and contact details.

Table 9. Overview of the project team

Name	Role	Email
Dmitry Timerbaev	Researcher	dmitry.xnt@gmail.com
Danil Markov	Researcher	danilmrkov@gmail.com
Sofia Zhukova	Academic supervisor	sofia.v.zhukova@gsom.spbu.ru
Vasiliy Garshin	IT supervisor	vgarshin@gsom.spbu.ru
Sergei Pitinov	Aviasales representative Business Development & Partner Relations at Travelepayouts	sergey.pitinov@aviasales.ru

Elizaveta Rudykh	Sponsor Head of Marketing at Travelpayouts	elizaveta.rudykh@aviasales.ru
Tatiana Buyanova	Sponsor Head of Business Development & Partner Relations at Travelpayouts	tatiana.buyanova@travelpayouts.com

Source: [authors research]

1.6.3. Research framework

In relation to CRISP-DM methodology our data mining objectives can be divided into four major groups, with each objective covered in its respective sub-chapter (see Table 10 below).

Table 10. Data mining project objectives within CRISP-DM methodology

Objective(s) №	CRISP-DM Phase	Chapter
-	Data understanding (description and exploration)	2.1. Data description and exploration
1	Data construction	2.2. Data construction of the direct advertisers' datasets
4		2.4. Data construction for network advertisers' datasets
3		2.5. Data construction of the travel verticals classification
2	Model building	2.3. Model building for travel advertisers' classification
5	Data integration	2.6. Data integration of the final dataset
6-12	Preparation of final report (BI dashboard in this case)	2.7. BI dashboard preparation

Source: [authors research]

1.6.4. Methodological framework

The development of methodology for travel verticals classification (objective № 4) involved interviews with the Aviasales industry experts, as well as the market analysis of the companies operating within the travel industry - their products, business models and distinctive features. In addition, travel vertical classification models from other sources were evaluated and utilized during methodology development (Revfine, Jai, K., n.d.).

As for the development of the BI dashboard (objectives № 6, 7, 9), the general methodological approach included the following steps (see details in 2.7. BI dashboard preparation):

- *Performance requirements* - after the analysis of processing capacity of data and the technical requirements of available tools on the market, the format of BI dashboard was chosen. In addition, this format was coordinated during the interview with Aviasales team to verify the compatibility and convenience of use in the company.
- *Company functionality expectations* - based on the developed business, functional and non-functional requirements, the desired content of the dashboard was developed. The list of functionalities included all the desired components to be displayed on the slide: charts, statistics, slicers and supporting dashboard elements.
- *Template standardization* - before the creation of the dashboard content, the unified dashboard template was developed. The dashboard template served as a basis for individual pages, which included the unified design, navigation and slicer panels, page titles and other template components.

Finally, the development of the frameworks for the analysis of travel brands competitive position (objective № 8) was driven by empirical evaluation of the affiliate links profile and consultations with the Aviasales industry experts. Companies' affiliate program size and quality of affiliates were evaluated against averaged metrics for the whole travel affiliate market. Relative performance of the travel brands against the market in those two dimensions defined the framework for measuring the brands overall competitive positioning and laid the ground for managerial implications.

The following chapter provides detailed description and exploration of data used during the project, as well as the in-depth description of each objective completed.

Chapter 2. Data Mining & BI Tool Development Process

2.1. Data description and exploration

2.1.1. General overview

The general scope of the project included the analysis of the major datasets, provided by the Aviasales company (see Table 3). As discussed in the previous chapter, the data contains information on affiliate marketing links in the travel industry, its parameters and the

supplementary specification files for dataset refinement and modelling. Overall, the whole set of files can be divided into two functional groups:

1. *Primary data* - the group includes dataset collections № 1 and 2, containing all the affiliate link base with link parameters for direct and network advertisers. The primary data constitute the basis of analysis in the project and has a key importance for fulfilling research objectives.
2. *Supplementary data* - the group includes dataset collections № 3, 4 and 5, complementing primary data in classification and filtration. As stated in the previous chapter (see 1.4.3. “Data problems”) and will be put into details further (see 2.1.5. “Data problems (detailed description)”), the primary data requires certain preprocessing and filtering that can be achieved with the help of the supplementary data. The supplementary data serves for primary data improvement and may increase the accuracy of data modeling and construction.

Based on the allocated functional groups, the data varies in its nature, source and characteristics, which will be discussed in the following paragraphs.

2.1.2. Data specification

All the listed dataset collections were collected and provided by the Aviasales specialists. The following table represents the description of the data and its collection methods, based on the distinguished functional groups (see Table 11 below).

Table 11. Data specification

Primary data	Files	151 CSV files (each CSV file contains data on one individual advertiser domain): <ul style="list-style-type: none"> ● 123 CSV files of direct advertisers’ affiliate links (dataset collection № 1) ● 28 CSV files of network advertisers’ affiliate links (dataset collection № 2)
	Size	117 582 616 links, 55 parameters: <ul style="list-style-type: none"> ● 90 139 295 links of network advertisers ● 27 443 321 links of direct advertisers
	Collection method	Aviasales acquired the primary data from the Ahrefs Pte. Ltd., a multinational website explorer company, which was responsible for the entire retrieval of link base from the Internet. Ahrefs’ collection algorithm includes the application of its own developed web crawler (or bot) that visits millions of websites globally to retrieve information and store it in Ahrefs records. The bot iteratively parses web link characteristics, including domain rating, backlinks quality and other features (see 2.1.3. “Primary data description” for details). According to Ahrefs, their crawler collects the data from the

		entire world wide web, constantly updating their world’s largest third party index of backlinks. (Ahrefs by the numbers, n.d.) The collected Ahrefs data was transferred to Aviasales company, which performed further validity and quality checks, as well as additional processing of the received data.
Supplementary data	Files	815 CSV, 5 XLSX: <ul style="list-style-type: none"> • 809 CSV of dataset collection № 3 • 4 CSV files of dataset collection № 4 • 7 CSV files of dataset collection № 5
	Collection method	<ul style="list-style-type: none"> • <i>Dataset collection 3</i> - using a website parser, the html data of network advertisers’ homepage texts was collected (a total of 218,061 network advertisers) The list of selected advertisers included the unique advertiser domains from affiliate network links (dataset collection № 2). The process of dataset collection was held by Aviasales company. • <i>Dataset collection 4</i> - using a website parser, data on advertisers’ product categories from affiliate networks Travepayouts, Skimlinks, Admitad and Awin was collected. The covered brands were limited to the four mentioned networks. The process of dataset collection was held by Aviasales company. • <i>Dataset collection 5</i> - using a website parser, advertiser IDs for network advertiser links were collected by joining the redirected advertiser domain with unique advertiser IDs (keys) retrieved from the affiliate links. Each network advertiser is assigned a unique advertiser ID (key). The process of dataset collection was held by Aviasales company.

Source: [authors research]

2.1.3. Primary data description

The primary data contains information about the affiliate links that were received by the Ahrefs web crawler through the search of the global Internet. The dataset includes both direct and network affiliate advertisers. Total number of links is 117 582 616 (network: 90 139 295 links, direct: 27 443 321 links). Each link is characterized by 55 parameters (Ahrefs API, n.d.), the description of which is presented in the following table (see Table 12 below).

Table 12. Affiliate link parameters

Parameter	Description and data type
url_from	the URL where the referral link was found and from which the information was collected (string)
refdomain	the main domain from which the information was collected (string)
ahrefs_rank	domain rating according to the Ahrefs methodology, which takes into account the size and quality of the backlinks (integer)

domain_rating	an indicator of the quality of the domain's backlinks relative to other domains in the Ahrefs database (integer)
ahrefs_top	domain rating relative to all sites in the Ahrefs database (integer)
ip_from	domain IP address (string)
links_internal	number of links to web pages within the domain (integer)
links_external	number of links to external web pages (integer)
page_size	size web page in bytes (integer)
encoding	the encoding type of the content of the web page (string)
title	the title of the web page (string)
language	the language of the web page (string)
url_to	referral link to the advertiser's website (string)
first_seen	the date when the page first appeared in the network register (string/datetime)
last_visited	the date of the last visit to the page by the search bot (string/datetime)
prev_visited	date of the penultimate page visit by the search bot (string/datetime)
deleted_at	marker whether the page was deleted (boolean)
original	marker original web page (boolean)
content	token content of the web pages (boolean)
refdomains	the number of referral domains (integer)
linked_root_domains	the number of root domains (integer)
positions	the position of a web page in search results (integer)
traffic	traffic for website (float)
redirect_code	code move the web page (integer)
alt	attribute alternative text for images (string)
anchor	the anchor link from search results (string)
del_reason	code the reason for the removal of the web page (string)
lost_redirect_reason	code of the reason for moving the web page (string)
lost_redirect_source	source code move the web page (string)
lost_redirect_new_target	code new goals after you move the web page (string)
drop_reason	reason code resolve web pages from search results (string)

in_raw	the indicator of net links (boolean)
in_rendered	indicator links to the previous render (boolean)
text_pre	raw text indicator (string)
text_post	processed text indicator (string)
redirect_chain	link redirection indicator (string)
http_code	http code indicator (integer)
url_from_first_seen	the date when the page first appeared in the network register (string/datetime)
first_origin	the source of the original data page (string)
last_origin	the source of the last data page (string)
total_backlinks	aggregated number of backlinks of the domain (integer)
powered_by	pointer of the server where the website is hosted (string)
all	contains any HTML elements (boolean)
text	text elements (boolean)
image	contains images (boolean)
nofollow	contains the token rel="nofollow" (boolean)
ugc	contains the token of user generated content (boolean)
sponsored	contains sponsor token (boolean)
redirect	contains the token redirect (boolean)
dofollow	contains the token rel="follow" (boolean)
canonical	contains a canonical element in the HTML (boolean)
gov	contains generic top level domain for the government of the United States (boolean)
edu	contains a common top level domain for educational institutions (boolean)
rss	contains an enriched summary of the site (boolean)
alternate	contains alternative domains (boolean)

Source: [authors research]

The data on affiliate network links (dataset collection № 2) included 16 unique affiliate networks (see Table 13 below).

Table 13. Affiliate networks in dataset collection № 2

Affiliate network domain	Available subordinate domains
admitad.com	admitad.com
affilired.com	affilired.com
avantlink.com	avantlink.com
awin1.com	awin1.com
cityads.com	cityadspix.com, nfemo.com, pwieu.com, hskwq.com
cj.com	tkqlhce.com, jdoqocy.com, anrdoezrs.net, dpbolvw.net, qzyfj.com
impact.com	2lka.net, 7eer.net, pxf.io, sjv.io, evyy.net, ojrj.net
linkshare.com	click.linksynergy.com
maxbounty.com	maxbounty.com
partnerize.com	prf.hn
shareasale.com	shareasale.com
skimlinks.com	go.skimresources.com
tc.tradetracker.net	tc.tradetracker.net
tradedoubler.com	tradedoubler.com
travelpayouts.com	travelpayouts.com
viglink.com	viglink.com

Source: [authors research]

2.1.4. Supplementary data description

Further exploration provides the description of dataset collections № 3-5, which include data of the text content of homepages, data on network advertisers product categories and data on advertiser IDs for network advertisers.

Dataset collection № 3 contains the texts of the home webpages of each network advertiser collected by the parser. The parser selected the advertiser domains, required page languages, collected texts in HTML and retrieved other related parameters. The dataset includes texts in 47 languages. The webpage texts dataset consists of 218 161 texts, collected from the unique network advertiser domains. The purpose of this dataset was to develop an unsupervised classifier to categorize travel websites in dataset collection № 2 (see 2.1.5 “Data problems (detailed description)”). The dataset includes 6 parameters, presented in the Table 14 below.

Table 14. Dataset collection № 3 parameters

Parameter	Description and data type
url	network advertiser domain (string)
flag	technical marker - whether the data is uploaded by url or not (boolean)
code	technical marker - HTML code (integer)
file	link to the original file where the HTML code of the network advertiser's home page was uploaded (string)
text	text of the network advertiser's home page (string)
language	language code (string)

Source: [authors research]

Dataset collection № 4 contains information on product categories of advertiser brands, included in affiliate networks travelpayouts.com, admitad.com, awin1.com and skimlinks.com. The list included more than 335 categories for multiple languages. Each brand could have been assigned to more than 1 product category (e.g. “Lighting, Furniture, Lingerie Sleepwear, Soft Furnishings, Men's Shoes”). The categories are developed and assigned by each affiliate network to brands by their own methodologies. The total categorical list consisted of 32 569 unique brand domains. Along with the previous dataset (dataset collection № 3), the purpose of this dataset was to additionally categorize travel websites in dataset collection № 2 (see 2.1.5 “Data problems (detailed description)”).

Table 15 represents the general parameters included in dataset collection № 4:

Table 15. Dataset collection № 4 parameters

Parameter	Description and data type
id	advertiser ID number (integer)
domain	domain of advertiser’s website (string)
country	country name of a domain (string)
categories	list of categories included in advertiser domain (string)

Source: [authors research]

Dataset collection № 5 included the data on advertiser IDs, which are uniquely set by the affiliate networks for each advertiser. The purpose of this dataset was to retrieve advertiser domains from the certain network advertisers’ links, Those links do not contain identifying information on a particular advertiser domain, but include only its advertiser ID. For instance, the URL from tradetracker.net affiliate network

“<https://tc.tradetracker.net/?c=14252&m=582122&a=358080&r=&u=>” contains unique advertiser ID **c=14252**, which stands for “*qatarairways.com*” in Tradetracker advertiser ID base. The dataset collection covered all affiliate networks, such as awin1.com, cityads.com, cj.com, shareasale.com, tradetracker.com, etc. The dataset included the following general parameters (see Table 16 below).

Table 16. Dataset collection № 5 parameters

Parameter	Description and data type
adv_id	unique advertiser ID of an advertiser in the affiliate network (integer)
advertiser	domain of advertiser’s website (string)

Source: [authors research]

2.1.5. Data problems (detailed description)

As discussed in the first chapter, the received dataset collections contained individual challenges and issues, which impeded the straightway data analysis. For this purpose, the data mining objectives and their outputs were set (see 1.5.1. “Data mining outputs expectations”). Namely, the encountered problems in data can be divided by dataset collections, presented by the following table (see Table 17 below).

Table 17. Detailed description of the problems within Aviasales dataset collections

Dataset collection	Problem	Description	Solution
Dataset col. № 1 <i>Data on all travel industry direct advertisers’ redirect web links</i>	Not affiliate links	The data may contain ordinary links that do not belong to any affiliate program. Such link URLs do not contain unique affiliate markers	Filter all direct advertisers’ links by using unique affiliate markers
	Unknown affiliate markers	Unique affiliate markers for direct links were not provided - need to find affiliate markers for each advertiser	Separate all direct link parameters for each travel advertiser, identify the parameter responsible for affiliate program (affiliate marker)
	Travel vertical is not known	Travel vertical that describes the type of market niche that brand specializes in is not available within the data	Prepare methodology for travel vertical classification and manually assign travel vertical to each advertiser
Dataset col. № 2 <i>Data on all network advertisers’ affiliate links</i>	Not travel links	The data may contain affiliate network links not belonging to travel industry	Classify travel brands: 1. Joining with available categorization of brands (Dataset col. № 4) 2. Building a category

<i>placed through world largest affiliate networks</i>			classifier using texts of website homepages (Dataset col. № 3)
	Unknown advertiser domains	Affiliate network links lead to the website of an affiliate network (indirect link) and do not visibly contain the domain of an advertiser	Allocate advertiser domain: 1. Joining with available affiliate advertiser IDs for networks (Dataset col. № 5) 2. Finding advertiser domains in a deep link (e.g. advertiser domain marked in blue: https://clk.tradedoubler.com/click?p=224460&a=2474116&g=0&url=https://www.skyscanner.fi)
	Travel vertical is not known	Travel vertical that describes the type of market niche that brand specializes in is not available within the data	Prepare methodology for travel vertical classification and manually assign travel vertical to each advertiser
Dataset col. № 3 <i>Data on the text content of homepages of all network advertisers placed through world largest affiliate networks</i>	Sample imbalance	Dataset contains low share of travel websites, so NLP classification model poorly trains on highly imbalanced samples - classification of travel class is inefficient	Solve sample imbalance problem by filtering dataset texts by popular travel keywords
	Multi-language texts	Dataset contains 47 languages, which require individual language models to be built	Narrow language model training to 2 most popular languages in dataset collection: English and Russian (account for 80% of all advertisers within the dataset collection)
Dataset col. № 4 <i>Data on network advertisers product categories</i>	Not all brands are present	The dataset categories do not cover all the listed advertiser brands in dataset collection № 2	Match only the available brands by categories, classify the rest by the built NLP model
Dataset col. № 5 <i>Advertiser IDs (keys) data for all network advertisers</i>	Not all advertiser IDs may be present	The dataset advertiser IDs may not cover all the listed affiliate networks in dataset collection № 2	Match only the available affiliate networks by affiliate IDs. Employ deep link analysis to recover remaining advertiser domains

Source: [authors research]

2.2. Data construction of the direct advertisers' datasets

The first step in the data mining process involved retrieving affiliate link markers from the datasets with direct advertisers' redirect links and using them to separate affiliate links from non-affiliate redirect links (see objective № 1 in Table 8).

To achieve the objective, the following algorithm was developed (using the example of the advertiser "rentalcars.com"):

- 1) CSV file "rentalcars.com.csv" from dataset collection № 1 was loaded with Python's Pandas library in JupyterHub IDE
- 2) Special function was applied on column 'url_to'. This function decomposed all URL links in the column and returned the total count of all URL markers found (for function code see Figure 1 in Appendix)
- 3) Each found URL marker was tested if (1) it is evident that the marker identifies affiliate relation (see Table 18 below for example)

Table 18. Marker identification explanation 1

Marker	URL	Explanation
affiliateCode	https://www.rentalcars.com/CityLandingPage.do?countryCode=my&place=Meridi&affiliateCode=myvasocom950&preflang=en	The name of the marker hints that it is related to advertiser's affiliate program
preflang	https://www.rentalcars.com/CityLandingPage.do?countryCode=us&place=Berthold&preflang=es	This marker is not related to the affiliate program. It is the preferred language marker followed by the language code.

Source: [authors research]

and/or (2) it is inside a URL that redirects to the purchase/booking window and contains an affiliate ID that can be used by the advertiser to identify a specific affiliate (see Table 19 below for example)

Table 19. Marker identification explanation 2

URL	Explanation
https://www.rentalcars.com/CityLandingPage.do?countryCode=my&place=Meridi&affiliateCode=myvasocom950&preflang=en	The URL redirects to the car rental booking window. It also contains the affiliate ID after the marker ("myvasocom950" - an ID for "my10airport-hotels.com" affiliate websites)

http://www.rentalcars.com/Home.do?affiliateCode=distancecities&prelang=EN&adcamp=result&adplat=distancecities	The URL redirects to the car rental booking window. It also contains the affiliate ID after the marker ("distancecities" - an ID for "distance-cities.com" affiliate websites)
---	--

Source: [authors research]

- 4) The marker that satisfied the criterias from the previous step ("affiliateCode" in case of "rentalcars.com") was checked by the Aviasales specialist. After specialist's approval, the marker was considered to be a true affiliate marker of the "rentalcars.com".

Other 123 CSV files (i.e. advertiser domains) were processed using the same algorithm outlined above. Finally, after all affiliate markers for direct advertisers were found, each CSV file was filtered by the respective affiliate marker in the 'url_to' column. The resulting 123 filtered samples were merged into the final dataset (output dataset № 1) that contained all affiliate links and their parameters for each direct advertiser. 11.2 million links (around 41% of total direct advertisers' redirect links) were identified as affiliate links.

2.3. Model building for travel advertisers' classification

Next step in the data mining process required building a text classification model based on parsed texts of the network advertisers' homepages (see objective № 2 in Table 8). This task was required in order to distinguish advertisers related to the travel industry in the dataset collection № 2 from those that are not related (see Table 17 for details). The idea is that homepages of the travel brands contain specific words (such as "travel", "hotels", "air tickets", etc.) that can be used to form a group of similar homepages that are strictly related to travel services. The general process involves three major steps to be taken:

- **Text feature extraction.** Transform text data into appropriate form that can be fed into the machine learning model
- **Modeling.** Design a model with appropriate algorithm, where it is possible to load transformed data and get the groupings of similar homepages
- **Testing.** Observe resulting groupings and choose the one that includes travel-related advertisers

According to Provost, et.al, the general strategy for transforming text into appropriate data format requires tokenization of all words, followed by applying vectorization technique. In the context of this research, each advertiser's homepage represents a document – one piece

of text consisting of words (tokens). The whole collection of documents is referred to as corpus, while every document in corpus is treated as a collection of individual words. Preprocessing stage requires that each collection of words be cleaned of punctuation, stop words and lemmatized. After preprocessing, an appropriate vectorization algorithm is applied to extract the features from the text (Provost, 2013, pp. 250-252).

In regards to the algorithms that can detect similarity within the text, Provost et.al. suggest using various unsupervised clustering algorithms, such as hierarchical clustering or k-means clustering, that are employed to detect certain natural (but not directly observed) groupings within the data. The idea is that those algorithms are trying to divide similar data points according to the defined distance function. Number of clusters is either defined by the researcher (in case of k-means clustering) and clusters are iteratively formed until all data points are assigned to each cluster and cluster centroids stop shifting; or as in case with hierarchical clustering, defined by the clustering algorithm itself (Provost, 2013, pp. 141-169).

2.3.1. Selecting modeling technique

In selecting optimal modeling techniques for objective № 2, several peculiarities have to be considered:

- **Technical constraints.** Time and computational resources available for the whole project were limited. Due to the large amount of data that had to be processed, resource-demanding modeling techniques could not be accepted.
- **Sample imbalance.** Preliminary analysis of the dataset collection showed that there exists an imbalance between travel advertisers and non-travel advertisers. Therefore, optimal modeling technique had to be flexible and efficient enough to allow quick implementation of additional steps to mitigate sample imbalance.
- **Multi-language dataset.** Dataset collection contained documents in 47 different languages. Because 80% of the dataset collection was covered by English and Russian languages – it was decided to prepare only two models for those two languages, with other languages to be classified using supplementary data from dataset collection № 4. Therefore, an optimal modeling technique had to be flexible and efficient enough to quickly implement two language models with consideration of technical constraints.

After extensive analysis of several vectorization and clustering techniques, it was decided to choose TF-IDF (for vectorization) and K-Means (for clustering) algorithms.

TF-IDF (term frequency – inverse term frequency) algorithm was chosen for several reasons:

- 1) TF-IDF vectorizer is intuitive, simple to compute and frequently used in a business environment. It captures only the importance of specific words to the document and corpus, relying on word's frequency. It does not capture semantics or word positioning, but it's not required as part of the objective, since only lexical characteristics of documents are used for classification (whether homepage contains travel-related words or not) (Provost 2013, pp. 252-256)
- 2) Traditional sparse vectorizers (such as TF-IDF) usually outperform neural word and character embedding models by an average margin of 3-5%, especially in general classification tasks (see Figure 3 below). Moreover, sparse vectorizers require much less time and computational resources to process the data. (Arora, et. al., 2019)

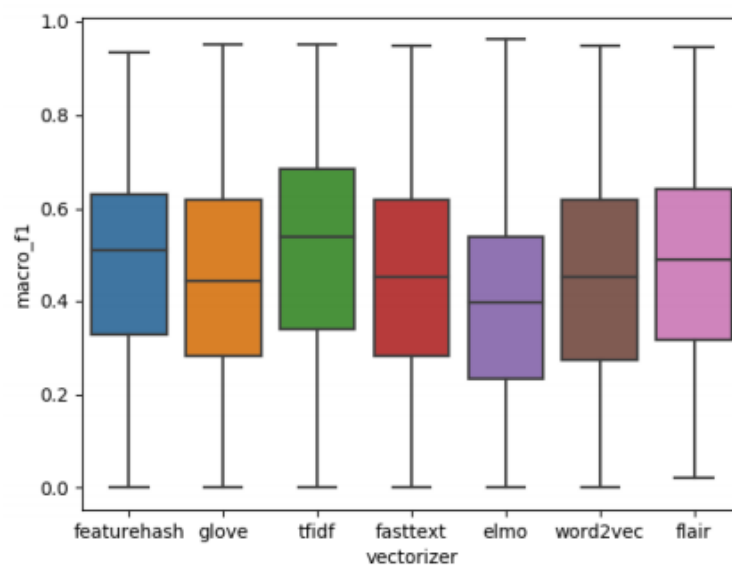


Figure 3. Vectorizers F1 score comparison across testing datasets

Source: [Arora, et.al., 2019]

Similarly, the K-Means algorithm was chosen for the following reasons:

- 1) K-Means is simple to implement, generalizes to the clusters of any shapes and sizes, takes less time to converge compared to other algorithms and is not computationally demanding. (Google Developers, 2021)
- 2) K-Means scales well to large datasets, and on average performs better with large datasets compared to other algorithms (Abbas, 2008)
- 3) K-Means performs better on average compared to DBSCAN and OPTICS algorithms. It performs worse than model-based clustering algorithms (such as EM), but is more computationally effective. Most of the dimensionality drawbacks of K-Means

algorithm can be mitigated through the use of PCA dimensionality reduction (Ding, Xiaofeng., 2004)

- 4) Despite the fact that K-Means requires manual setting of clusters, being a fast algorithm allows testing of multiple k values and parameter tuning, thus allowing to choose optimal composition depending on the preferred outcome. It also tends to give good results in tasks of general data segmentation (Amancio, et. al. 2019). The accuracy of K-Means on average is approximately similar compared to more sophisticated computationally-intensive algorithms, especially on datasets that contain large number of features – just like in the case of the dataset collection № 2 (see Figure 4 below)

#	Algorithm	DB10C200F		
		ARI_{def} (%)	ARI_{best_p} (%)	ARI_{best_r} (%)
1	subspace	33.1	79.0	-
3	EM	33.6	77.5	100.0
2	spectral	76.7	83.0	100.0
4	clara	56.1	69.5	93.1
5	hcmode	34.1	95.1	100.0
6	k-means	60.9	90.7	99.5
7	hierarchical	0.04	92.1	100.0
8	optics	54.9	81.4	83.1
9	dbscan	69.0	87.6	88.5

Figure 4. Clustering algorithms performance on 200 features artificial datasets

* ARI_{def} represents the average accuracy obtained when considering the default parameters of the algorithms. ARI_{best_p} represents the average of the best accuracies obtained when varying a single parameter. ARI_{best_r} represents the average of the best accuracies obtained when parameters are randomly selected.

Source: [Amancio, et.al., 2019]

The selected techniques were implemented in Jupyter Hub using Python’s machine learning and natural language processing libraries. Below is the summary of used libraries and modules (Table 20).

Table 20. Description of data modeling tasks and instruments

Preprocessing	
Task	Library and module used
Stop words removal	nltk.corpus.stopwords (NLTK.org, 2019)

Lemmatization (WordNet, English)	nlTK.stem.WordNetLemmatizer (NLTK.org, n.d.)
Lemmatization (MorphAnalyzer, Russian)	pymorphy2.MorphAnalyzer (Pymorphy2, n.d.)
Modeling	
Task	Library and module used
TF-IDF vectorization	sklearn.feature_extraction.text.TfidfVectorizer (Scikit-learn.org. Feature extraction, n.d.)
K-Means clustering	sklearn.cluster.KMeans (Scikit-learn.org. Clustering, n.d.)
Principal component analysis	sklearn.decomposition.PCA(Scikit-learn.org. Principal component analysis (PCA), n.d.)

Source: [authors research]

2.3.2. Model design

After finalizing the techniques and Python modules to be used for modeling, the general design of the model was prepared (see Figure 5 below).

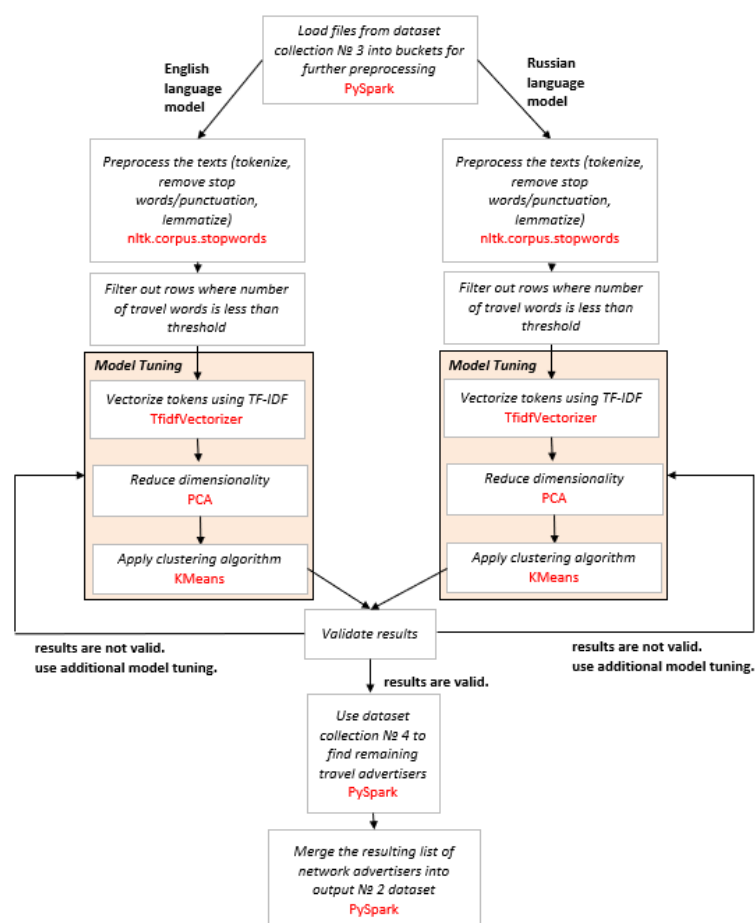


Figure 5. General model design (red color denotes the tool used)

Source: [authors research]

Overall, two language models have to be created - separate one for English and Russian languages. The appropriate data is loaded from dataset collection № 3 into PySpark buckets for necessary preprocessing (tokenization, stop words removal, punctuation removal, lemmatization). After preprocessing, the rows that do not contain specific words related to the travel industry are removed from the model to mitigate the effect of sample imbalance. Lists of travel-related keywords were approved by the Aviasales specialists. The threshold or cut-off values (CV) were chosen to be CV=3 in order to account for possible random occurrences of words, but at the same time to be strict enough to allow for maximum sample imbalance correction. If the number of travel words was less than the threshold, the row was dropped (see Table 21 below).

Table 21. Overview of travel-related keywords

Language	List of travel keywords	Cut-off value (CV)
Russian	'билет', 'отель', 'гостиница', 'авиабилет', 'аренда', 'прокат', 'автобус', 'поезд', 'самолет', 'самолёт', 'путевка', 'путёвка', 'travel', 'trip', 'путешествие', 'путешественник', 'бронирование', 'автобусный', 'расписание', 'забронировать', 'железнодорожный', 'турист', 'туроператор', 'туристический', 'туризм', 'виза', 'страна', 'аэропорт', 'курорт', 'шенген'	3
English	'ticket', 'hotel', 'rent', 'rental', 'bus', 'train', 'plane', 'airplane', 'voucher', 'travel', 'trip', 'traveler', 'booking', 'schedule', 'tourist', 'tour', 'tourism', 'visa', 'country', 'airport', 'resort', 'shengen', 'vacation', 'flight'	3

Source: [authors research]

Next, filtered tokens were to be vectorized, their dimensionality reduced using PCA (if needed) and then the clustering algorithm applied. Resulting clusters are validated through random sampling of 100 texts from each cluster and manual check if those texts indeed belong to travel advertisers. Random sampling is done three times for each cluster. For the results to be validated, certain criteria has to be satisfied:

- 1) Clusters that are assumed to contain travel advertisers on average have to contain no less than 95% of texts belonging to travel industry as a result of 3 random sampling

checks (the remaining non-travel advertisers were to be dropped from the dataset during completion of objective № 4)

- 2) Clusters that are assumed not to contain travel advertisers on average have to contain no more than 5% of texts belonging to the travel industry as a result of 3 random sampling checks (those 5% were considered noise and were expected to be picked later using supplemental information from dataset collection № 4).

If the results are considered not valid, the model is returned for further tuning. After getting valid results, the remaining travel advertisers within dataset collection № 3 (non-English and non-Russian) and within non-travel clusters are classified using the data on affiliate advertisers product categories from dataset collection № 4. The resulting lists of network advertisers in travel are aggregated in the output dataset № 2.

2.3.3. Model building

Both language models were initiated, preprocessed and filtered using travel keywords (see Appendix Figures 7-10 for code examples). During the modeling and tuning phase, multiple setups were tested, but the final model was built with the following parameters (see Table 22 below).

Table 22. Overview of language model parameters

Algorithm	English language model parameters	Russian language model parameters
TF-IDF	max_df=0.8 min_df=0.01 use_idf=True ngram_range(1,2)	max_df=0.8 min_df=0.01 use_idf=True
PCA	Applied at 1000 components, as those explained 60% cumulative variance. The acceptable variance explained in factor analysis for a construct to be valid is 60% (Hair, et. al., 2014, p.112), thus it is expected that reduction to 1000 components should be appropriate for the clusters to retain enough information (see Appendix Figure 11)	Not applied. PCA was considered redundant, due to already small corpus size as compared to the English language model.
KMeans	Default parameters, tested with k = [2,3,4,5,6,7,8,9,10]. In the final model k = 8 (gives	Default parameters, tested with k = [2,3,4,5] In the final model k = 2 (gives

	the lowest ratio of average % of travel advertisers in non-travel clusters / average % travel advertisers in travel clusters) (see Table 23 below)	the lowest ratio of average % of travel advertisers in non-travel clusters / average % travel advertisers in travel clusters) (see Table 23 below)
--	---	---

Source: [authors research]

Table 23. Overview of cluster size used and resulting criteria ratios

	<i>k=2</i>	<i>k=3</i>	<i>k=4</i>	<i>k=5</i>	<i>k=6</i>	<i>k=7</i>	<i>k=8</i>	<i>k=9</i>	<i>k=10</i>
EN model	0.259 (22%/85%)	0.118 (11%/89%)	0.078 (7%/94%)	0.096 (9%/97%)	0.100 (10%/96%)	0.068 (7%/97%)	0.025 (2%/97%)	0.046 (4%/95%)	0.047 (4%/89%)
RU model	0.021 (2%/95%)	0.042 (4%/94%)	0.022 (2%/92%)	0.088 (11%/91%)	-	-	-	-	-

Source: [authors research]

2.3.4. Model assessment

The resulting language models were accepted as valid. All the criterias were met (see Table 24 below).

Table 24. Final clusters obtained (green highlighting denotes clusters containing travel advertisers)

<i>English language model cluster and % of travel related advertisers within</i>	<i>Russian language model cluster and % of travel related advertisers within</i>
<ul style="list-style-type: none"> ● Cluster 1 - 98% ● Cluster 2 - 98% ● Cluster 3 - 95% ● Cluster 4 - 4% ● Cluster 5 - 4% ● Cluster 6 - 3% ● Cluster 7 - 1% ● Cluster 8 - 0% 	<ul style="list-style-type: none"> ● Cluster 1 - 95% ● Cluster 2 - 2%

Source: [authors research]

Additionally, other travel advertisers, not covered by the language models or those that were incorrectly placed into non-travel clusters, were filtered out from dataset collection № 3 using supplemental product categories data from dataset collection № 4. Finally, 3254 network advertisers were classified as related to the travel industry - output dataset № 2 with those advertisers was successfully created with the advertiser domains placed in column ‘advertiser’. Results presented in the output dataset № 2 were approved by the Aviasales specialists.

2.4. Data construction for network advertisers' datasets

As was previously stated in the Data problems description (see 2.1. “Data description and exploration”), the data on network advertisers’ links differs from direct advertisers’ links in a way that referring domain is set to affiliate network website (instead of advertiser website), and the domain of the network advertiser cannot be identified.

The two applied methods to retrieve advertiser domains from affiliate network links include the deep link analysis and the advertiser domain search by the advertiser ID (keys). Each method was individually chosen for each affiliate network, depending on the general form of a link that affiliate networks use in their business. The following Table 25 represents the affiliate networks and methods applied to each of them to retrieve advertiser domains.

Table 25. Advertiser domain retrieval method by affiliate networks

Method	Affiliate networks
Domain search by deep link analysis	<ol style="list-style-type: none"> 1. avantlink.com 2. cityads.com 3. linkshare.com 4. tradedoubler.com 5. partnerize.com 6. skimlinks.com 7. viglink.com 8. impact.com 9. affilired.com.csv
Domain search by advertiser ID (key)	<ol style="list-style-type: none"> 1. awin1.com 2. shareasale.com 3. tradedoubler.com 4. admitad.com.csv 5. cj.com 6. tc.tradetracker.net 7. travelpayouts.com

Source: [authors research]

These two methods will be further discussed in detail.

2.4.1. Domain search by deep link analysis

Deep link analysis represents one of the ways to retrieve the advertiser domain from the affiliate link. Deep link can be defined as a web link that contains additional internal links inside (Bray, 2002). In the example provided below (Figure 6), the link “<http://go.skimresources.com>” cannot be classified as a deep link, since it refers to the homepage and does not contain any additional internal links inside of it. The second link presented is the deep link that contains the internal link to the domain

“<http://www.booking.com/>” - this is the advertiser’s domain the user will be redirected, once clicked on this referral link.



Figure 6. Deep link in affiliate network links

Source: [authors research]

For the extraction of an advertiser domains from affiliate network deep links, the following procedures were taken:

1. The list of affiliate networks with advertiser domains in deep links was selected, analyzed and approved by Aviasales
2. Based on the selection, CSV files of affiliate networks from dataset collection № 2 were loaded with Python’s PySpark library in JupyterHub IDE
3. Special function was applied on column ‘**url_to**’. This function searched the position of the advertiser domain argument within deep links, retrieved advertiser domain (using Python regex library), unquoted and extracted the advertiser domain in a standardized form (using Python urllib.parse and tldextract libraries). For function code see Figure 2 in Appendix.
4. Each extracted advertiser domain was assigned to a new column ‘**advertiser**’

2.4.2. Domain search by advertiser ID (key)

The second case of advertiser domain retrieval includes the search of advertiser IDs in network affiliate links. Similar to the previous case, an advertiser ID can provide information on a specific advertiser domain, though encrypted in a special key, which affiliate networks individually set for each advertiser domain. In the example provided below (Figure 7), the advertiser ID “*a=253678*” stands for an advertiser “*hotelspecials.se*” in the affiliate network “*tc.tradetracker.net*”. However, not all the affiliate networks use encryption for advertiser IDs - as in the previous case, sometimes an advertiser domain is explicitly set within a deep link.

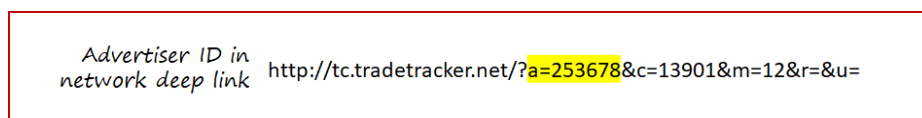


Figure 7. Advertiser ID in affiliate network links

Source: [authors research]

Thus, each network advertiser can be identified using affiliate links, if the advertiser ID (keys) are available. With the help of the dataset collection № 5 (see 2.1.4. “Supplementary data description” for details) the following procedures took place to retrieve the advertiser domains:

1. Affiliate networks with advertiser IDs in affiliate links were selected, based on the available advertiser IDs in the dataset collection № 5
2. Based on the selection, CSV files of network advertisers links from dataset collection № 2 and advertiser IDs from dataset collection № 5 were loaded with Python’s PySpark library in JupyterHub IDE
3. For dataset collection № 2 files, a special function was applied on column ‘**url_to**’. This function searched the position of an advertiser ID argument in affiliate link and retrieved this ID (using Python regex library). For function code see Figure 3 in Appendix.
4. Each extracted advertiser ID was assigned to a new column ‘**adv_id**’
5. The selected links from the dataset collection № 2, with extracted advertiser IDs, were joined with dataset collection № 5 by the column “**adv_id**” to get the advertiser domain for each ID (using Python PySpark.join function). For function code see Figure 4 in Appendix.
6. Each retrieved advertiser domain was assigned to the column “**advertiser**”

In the end, application of both methods allowed to create a preprocessed dataset of affiliate network links with known advertiser domain for each link. In this format, the dataset can be further filtered by the advertisers that belong to the travel industry.

2.4.3. Travel class retrieval from affiliate network links

The complete list of classified travel brands presented in the output dataset № 2 (see 2.3. “Model building for travel advertisers’ classification” for details) and previously prepared dataset with affiliate network links, made it possible to solve the problem of having non-travel network links in our dataset.

The process of travel links retrieval took the following procedures:

1. CSV files of all network advertisers links from dataset collection № 2 and a list of travel advertisers from output dataset № 2 were loaded with Python’s PySpark library in JupyterHub IDE

2. Two datasets - affiliate network links and travel advertiser domains, are joined by the ‘**advertiser**’ column (using Python PySpark.join function). For function code see Figure 5 in Appendix.
3. Links that were not joined are dropped from the final dataset

As a result of the described dataset preprocessing and travel domains retrieval procedures, the output dataset № 3 was created (see Table 8 for details). The output dataset № 3 contains network affiliate links and their parameters of the travel-related advertisers (brands). In further stages of the project, this output will also be merged with the direct affiliate links (output dataset № 1) and travel verticals classification (output dataset № 4) to create a unified dataset of affiliate links in travel.

2.5. Data construction of the travel verticals classification

Next step involved the development of a methodology for classification of all advertisers in the affiliate channel (both direct and network) by travel vertical (see objective № 4 in Table 8). As was previously mentioned, travel vertical represents a specific niche, sub-industry or concentration within travel services, such as flights, hotels, tours and activities, etc. Preliminary analysis showed that there is no industry-accepted standard of vertical classification. Given the lack of standard methodology for classifying travel brands by specific vertical, it was decided to prepare an original methodological document in collaboration with the industry experts from Aviasales company, and with support of the outside market research sources (Revfine, Jai, K., n.d.).

Methodology, that was created as a result, is based on two-tier division of travel verticals. Tier 1 travel vertical represents highest order vertical concentration that splits into specific niches on a Tier 2 level (see Table 26 below). The methodology also includes two types of travel vertical attributes - Primary and Secondary attributes. Primary attribute represents the main business concentration for the travel advertiser, while the Secondary attributes are assigned if the advertiser offers additional travel services that are not part of its core business activities (for example - a hotel offering a transfer to the airport). Additionally, the specific travel vertical - "Aggregators", was introduced for travel advertisers that do not provide goods/services on their behalf (and also do not provide direct support to providers) and act as an intermediary in aggregating the offers of other travel brands.

All travel advertisers were assigned to a specific Primary and Secondary (if applicable) attributes of Tier 2 verticals based upon satisfying specific criterias (see Table 27 below). There were two general rules when applying vertical criterias:

- 1) Travel advertiser was assigned a certain Primary attribute if more than 50% of criterias for the attribute were satisfied, except for advertisers that satisfy Primary attribute criteria 1 of the "Aggregator" Tier 2 vertical. Those advertisers are classified as "Aggregator" in the Primary attribute by default
- 2) "Aggregator" vertical cannot be assigned as a Secondary attribute for advertisers that are not classified as an "Aggregator" in the Primary attribute

Table 26. Summary of Tier 1 and Tier 2 travel verticals

Tier 1 Vertical	Description	Tier 2 Verticals
Transportation	all verticals that involve transportation services (air, ground, water, rail)	<ul style="list-style-type: none"> ● Flights ● Buses ● Car rentals ● Trains ● Transfers ● Water Transport
Accommodation	all verticals that involve lodging or traveling space and related services	<ul style="list-style-type: none"> ● Hotels ● Vacation rentals ● Hostels ● Camping
Leisure Services	all verticals related to various leisure activities	<ul style="list-style-type: none"> ● Cruises ● Package tours ● Tours and activities ● Outdoors ● Food & Beverage ● Shopping
Auxiliary Services	all vertical related to providing supporting services to travelers and tourists	<ul style="list-style-type: none"> ● Aggregator ● Insurance ● Financial services ● Parking ● Information ● Legal services ● Loyalty programs ● Other

Source: [authors research]

Table 27. Description of criterias for assigning vertical to travel brands

Transportation Verticals			
Tier 2 Vertical	Description	Criteria for placing in Primary attribute	Criteria for placing a Secondary attribute mark
Flights	Air transportation - regular flights and private air transport. Companies offer regular flight tickets or provide booking for private jet / helicopter flight.	<ol style="list-style-type: none"> 1) customer is able to buy a ticket for a regular flight or book a private flight on-site 2) the company is directly involved in the business of air transportation 3) the company focuses primarily on transport services (air transportation from point A to point B) 	customer is able to buy a ticket for a regular flight or book a private flight without redirecting to other organizations' websites (for aggregators - on-site offers aggregation functionality is a minimum requirement)
Buses	Primarily long-haul, regular, ground transportation using bus-type vehicles. Companies offer booking of bus tickets.	<ol style="list-style-type: none"> 1) customer is able to buy a bus ticket on-site 2) the company is directly involved in the business of bus transportation 3) the company focuses primarily on long-haul transport services (bus service from point A to point B, primarily between settlements or certain sites) 	customer is able to buy a long-haul bus ticket without redirecting to other organizations' websites (for aggregators - on-site offers aggregation functionality is a minimum requirement)
Car Rentals	Automobile rentals, including camper trailers and car-sharing. Companies offer booking of vehicle rentals from their own fleet.	<ol style="list-style-type: none"> 1) customer is able to book an automobile/camper trailer rental on-site 2) the company owns vehicle fleet or has explicit control over it 3) the company focuses on renting automobiles and camper trailers from its vehicle fleet 	customer is able to book a car/camper trailer rental without redirecting to other organizations' websites (for aggregators - on-site offers aggregation functionality is a minimum requirement)
Trains	Regular railway transportation (excluding trams). Companies offer booking of train tickets.	<ol style="list-style-type: none"> 1) customer is able to buy a train ticket on-site 2) the company is directly involved in the business of train transportation 3) the company focuses primarily on transport services (train service from point A to point B) 	customer is able to buy a rail transport ticket (except for trams) without redirecting to other other organizations' websites (for aggregators - on-site offers aggregation)

			functionality is a minimum requirement)
Transfers	Specialized non-regular short-distance ground transportation services - transfers from/to airport, accomodation, tourist attraction, etc. Also includes taxi and chauffeur services. Companies offer booking and organization of transfer services.	<p>1) customer is able to book a short-distance ground transportation to the specified destination point - to the airport/accommodation/railway station, etc, or to book a custom transfer to the desired destination (for example - group transportation to the tourist attractions, taxi or chauffeur services) on-site</p> <p>2) the company is directly involved in the short-distance transfer business</p> <p>3) vehicle fleet may be represented by any kind of passenger ground vehicles - cars, minivans, buses, etc.</p> <p>4) the company focuses primarily on ground transport services (train service from point A to point B)</p>	customer is able to book a transfer to/from the specified destination point - to/from the airport/accommodation/railway station/attraction, etc, or to book a custom transfer to the desired destination (for example - group transportation to the tourist attractions, taxi or chauffeur services) without redirecting to other organizations' websites (for aggregators - on-site offers aggregation functionality is a minimum requirement)
Water Transport	Water transportation - ferries and short-distance passenger ships/boats. Companies offer ferry/short-distance passenger ship/boat tickets.	<p>1) customer is able to buy a ferry and/or passenger ship/boats ticket on-site</p> <p>2) the company is directly involved in the business of water transportation</p> <p>3) the company focuses primarily on transport services (ferry/boat service from point A to point B)</p>	customer is able to buy a water transport ticket (ferries and passenger ships/boats) without redirecting to other organizations' websites (for aggregators - on-site offers aggregation functionality is a minimum requirement)
Accommodation Verticals			
Tier 2 Vertical	Description	Criteria for placing in Primary attribute	Criteria for placing a Secondary attribute mark

<p>Hotels</p>	<p>Hotels, motels, inns, vacation resorts, etc. Companies offer booking of hotel-type rooms at their premises, including standard rooms, deluxe, joint, suites, etc. Companies offer accommodation booking at the living spaces under their explicit control.</p>	<p>1) customer is able to book a hotel-type room on-site 2) the company has explicit control over the facilities in which the clients stay (ownership of real estate, or the right of disposal, which includes control by the company over real estate in matters of cleaning, renovation, additional services, design, rules of residence, advertising, etc. without having ownership rights) 3) the company focuses primarily on the sale of its hotel-type rooms. Company actively promotes its hotel-type room stock.</p>	<p>customer is able to book hotel-type accommodation without redirecting to other organization's website (for aggregators - on-site offers aggregation functionality is a minimum requirement)</p>
<p>Vacation rentals</p>	<p>Lodging rentals that do not fit into common hotel-type room specifications. Those include -private home rooms, apartments, residential hotels, houses, villas, cottages, bungalows, lodges, cabins, bed & breakfast, etc. Companies offer accommodation booking at the living spaces under their explicit control.</p>	<p>1) customer is able to book a lodging rental that do not fit into common hotel-type room specifications on-site 2) the company has explicit control over the facilities in which the clients stay (ownership of real estate, or the right of disposal, which includes control by the company over real estate in matters of cleaning, renovation, additional services, design, rules of residence, advertising, etc. without having ownership rights) 3) the company focuses primarily on the sale of its lodging rentals that do not fit into common hotel-type room specifications. Actively promotes lodging options that cannot be classified as common hotel-type rooms.</p>	<p>customer is able to book accommodation that do not fit into common hotel-type room specifications without redirecting to other organizations' websites (for aggregators - on-site offers aggregation functionality is a minimum requirement)</p>

Hostels	Lodging rentals in hostels and similar types of shared accommodation. Companies offer accommodation booking at the living spaces under their explicit control.	<p>1) customer is able to book a hostel or similar type of shared accommodation on-site</p> <p>2) the company has explicit control over the facilities in which the clients stay (ownership of real estate, or the right of disposal, which includes control by the company over real estate in matters of cleaning, renovation, additional services, design, rules of residence, advertising, etc. without having ownership rights)</p> <p>3) the company focuses primarily on renting out its hostel spaces or spaces in similar types of shared accommodation. Company characterizes its offerings as hostel or hostel-type shared accommodation</p>	customer is able to book hostel accommodation or other similar type of shared accommodation without redirecting to other organizations' websites (for aggregators - on-site offers aggregation functionality is a minimum requirement)
Camping	Campsite rentals and other types of accommodation at the camping grounds. Companies offer campsite booking as well as other related services (tents and other camping equipment rentals)	<p>1) customer is able to book a campsite rental or similar type of accommodation at the camping grounds on-site</p> <p>2) the company has explicit control over the camping grounds/facilities where clients stay (ownership of territory, or the right of disposal, which includes control by the company over the camping grounds in matters of cleaning, renovation, additional services, design, rules of conduct, advertising, etc. without having ownership rights)</p> <p>3) the company focuses primarily on renting out campsites. Company characterizes its offerings as camping, campsites rentals, camping rental services.</p>	customer is able to book a campsite or other similar type of accommodation at the camping grounds without redirecting to other organizations' websites (for aggregators - on-site offers aggregation functionality is a minimum requirement)
Leisure Services Verticals			
Tier 2 Vertical	Description	Criteria for placing in Primary attribute	Criteria for placing a Secondary attribute mark

Cruises	Sea and river travels on cruise ships that include on board accommodation, dining and entertainment. Companies offer cruise trip tickets.	<p>1) customer is able to buy a cruise trip ticket on-site</p> <p>2) the company is directly involved in the business of cruise travel (must be either an owner or manager of cruise ships fleet)</p> <p>3) the company focuses on full cruise service including on-board accommodation, dining and entertainment where applicable</p>	customer is able to buy a cruise trip ticket without redirecting to other organizations' websites (for aggregators - on-site offers aggregation functionality is a minimum requirement)
Package tours	Vacation packages from online travel agencies (OTAs) that include a full travel bundle - any combination of transportation option, accommodation, extra services. Companies offer packages with predefined destination, dates and conditions.	<p>1) customer is able to buy a vacation package on-site (any combination of transportation option, accommodation, extra services)</p> <p>2) the company is directly responsible for organization of packaged services - all transportation options, accommodation and extra services must be arranged by the company with the actual providers on behalf of the client</p> <p>3) the company do not additionally aggregate any other offers with redirects to other organizations' websites</p> <p>4) the company is focused on packaged travel options as its main product</p>	availability to purchase vacation packages that may include several combinations of transportation option + accommodation + extra services from a list of services provided by the company; also includes functionality for booking/purchasing those services through company's website without redirecting to other organizations' websites (for aggregators - on-site offers aggregation functionality is a minimum requirement)

<p>Tours and Activities</p>	<p>Various tours and activities for travelers. The sector includes the following groups of activities:</p> <ul style="list-style-type: none"> - Excursions & Sightseeing (excursions/sightseeing tours to the historical, natural and cultural landmarks, either on foot or using vehicles. Includes tours to exotic places or tours organized in uncommon settings) - Wellness & Relax (spa, massage, baths, saunas, beaches, various wellness and relaxation activities, etc.) - Leisure Rentals (rental of bicycles, motorcycles, ATVs, yachts, boats, jet skis, snowmobiles, etc., including tours using that equipment) - Active Recreation (outdoor sports, hiking, surfing, rock climbing, diving, fishing, hunting, etc.) - Events (sporting events, concerts, exhibitions, fairs, parades, as well as scientific and business conferences, weddings, banquets, etc.) - Cultural and Sports Facilities (theaters, museums, libraries, galleries, tennis courts, gyms, pools, etc.) - Entertainment (nightclubs, amusement parks, casinos, shooting ranges, etc.) <p>Companies offer organization, booking, or distribution of tickets / passes for these activities.</p>	<p>1) website allows customer to book/buy a ticket/pass for offered tours and activities on-site or at least contains information on how those services can be purchased from the provider</p> <p>2) the company is a direct provider of offered tours and activities, or has a direct affiliation with the tour or activity provider (that may include direct support in organization, promotion, distribution of tickets/passes, marketing and logistics)</p> <p>3) the company is focused on providing tours and activities from the presented list of activity groups</p>	<p>company explicitly offers activities from the presented groups of activities; also includes functionality for booking/purchasing those activities through company's website without redirecting to other organizations' websites:</p> <ul style="list-style-type: none"> - Excursions & Sightseeing - Wellness & Relax - Leisure Rentals - Active Recreation - Events - Cultural & Sports Facilities - Entertainment <p>(for aggregators - on-site offers aggregation functionality is a minimum requirement)</p>
------------------------------------	---	---	---

<p>Outdoor s</p>	<p>Ski resorts, golf resorts/clubs and natural reserves/parks. Companies offer packaged services that may include entrance, various outdoor activities, accommodation and dining at the resort, club, or natural reserve/park premises.</p>	<p>1) only ski resorts, golf resorts/clubs, natural reserves and parks included 2) customer is able to purchase a full package (that includes accommodation, dining and additional services) and/or book certain services separately (given that those services are provided at the resort/natural reserve location) on-site 3) the company has explicit control over the resort's/natural reserve's territory/facilities (ownership of property, or the right of disposal, which includes control by the company over the territory/facilities of the resort/natural reserve in terms of operations, marketing, maintenance and other matters relevant to proper functioning of the location, regardless of ownership status) 4) the company is positioning itself as ski resort, golf resort/club or natural reserve/park (clear focus on main activities is required - for example, for ski resort that might be skiing/snowboarding, for golf resort/club - playing golf, for natural reserve/park - hiking, outdoor activities, etc.)</p>	<p>availability of mountain skiing/golf facilities as well as the access to natural reserves and parks; also includes functionality for booking/purchasing services provided by those locations through company's website without redirecting to other organizations' websites (for aggregators - on-site offers aggregation functionality is a minimum requirement)</p>
<p>Food & Beverag e</p>	<p>Restaurants, bars, cafes, dining/banquet halls, etc. Companies offering food/drinks preparation and catering services.</p>	<p>1) website provides table/catering booking information at minimum (for example - phone number or email) 2) the company is directly involved in the business of food/drinks preparation or catering 3) the company focuses on food/drinks preparation or catering</p>	<p>availability of restaurants, bars, cafes, dining/banquet halls, etc. in a list of services provided by the company; also includes functionality for booking/purchasing those services through company's website without redirecting to other organizations' websites (for aggregators - on-site offers aggregation functionality is a minimum requirement)</p>

Shopping	Stores (including ecommerce), duty free shops, souvenir shops, shopping malls, etc catering to travelers/tourists. Companies offer consumer goods for sale.	1) the company is directly involved in the business of selling goods (either as manufacturer or retailer) 2) the company focuses on selling consumer goods	availability of stores (including ecommerce), duty free shops, souvenir shops, shopping malls, etc. in a list of services provided by the company; also includes functionality for booking/purchasing those services (including goods from the listed shopping facilities) through company's website without redirecting to other organizations' websites (for aggregators - on-site offers aggregation functionality is a minimum requirement)
-----------------	---	---	---

Auxiliary Services Verticals

Tier 2 Vertical	Description	Criteria for placing in Primary attribute	Criteria for placing a Secondary attribute mark
Aggregator	Companies aggregating offers from other organizations. Aggregators act as an intermediary for customers searching for specific travel services and goods. Can be represented by multi-vertical online travel agencies or by the metasearch engines (“aggregators of aggregators”)	1) website provides functionality for search and aggregation of goods/services offered by third party providers and/or functionality to purchase/book those goods/services on-site 2) website content has explicit commercial intent - any non-commercial informational materials do not represent the core feature of the website 3) the company does not produce, or provide offered goods/services directly and does not provide direct support to goods/services providers in the matters of their operations, marketing or logistics or own any tangibles related to the offered goods/services	only if Aggregator vertical is already set as a Primary Sector

Insurance	Insurance companies offer various types of insurance for travelers/tourists, such as medical insurance and auto insurance. Companies provide options to purchase or apply for insurance on their website	1) the company is directly involved in the business of travel insurance and allows on-site purchase or application filling 2) the company focuses on selling travel insurance	customer is able to apply/purchase any type of travel insurance without redirecting to other organization's website (for aggregators - on-site offers aggregation functionality is a minimum requirement)
Financial services	Banking and financial services for tourists/travelers. Companies provide offers for bank cards, payment systems and currency exchange.	1) the company is directly involved in the business of travel-related financial services 2) the company focuses on providing financial services to travelers/tourists	customer is able to apply for bank card, sign up for payments system or order foreign currency without redirecting to other organization's website (for aggregators - on-site offers aggregation functionality is a minimum requirement)
Parking	Parking spaces and lots. Operating companies offer booking options for parking spaces	1) the company is directly involved in the business of providing parking spaces (possesses the ownership or disposal rights over parking facilities) 2) the company focuses on providing parking spaces or/and other related services	customer is able to book a parking space without redirecting to other organization's website (for aggregators - on-site offers aggregation functionality is a minimum requirement)

Information	<p>Companies offering information services for tourists/travelers:</p> <ul style="list-style-type: none"> - Online publishers and information platforms (news and entertainment websites dedicated to tourism and travel, also travel guides and social networks for tourists/travelers) - Transportation facilities websites (information websites of the airports, railway stations and metro systems) - Tourism information portals of cities and regions 	<p>1) the company is not involved directly in offering goods/services displayed on its website or explicit commercial aggregation of those offers, except for informational support only</p> <p>2) significant proportion of website's content is of informational or entertainment nature (articles, reviews, guides, users' comments, forums are all indicators of informational type of content)</p> <p>3) the company is focused on either distribution of travel-related information (including information on travel-related goods/services), or on operating an information-sharing platform for its' clients and partners covering travel-related themes</p>	<p>significant presence of information materials for tourists/travelers that are not directly related to company's primary business activities</p>
Legal Services	<p>Companies offering various legal services for travelers/tourists. Those services include legal consulting abroad, visa application assistance, legal help in lawsuits against airline companies and travel agencies.</p>	<p>1) the company is directly involved in the business of travel-related legal services</p> <p>2) the company focuses on providing legal services to travelers/tourists</p>	<p>customer is able to purchase travel legal services without redirecting to other organization's website (for aggregators - on-site offers aggregation functionality is a minimum requirement)</p>
Loyalty programs	<p>Companies that operate various rewards and loyalty programs for travel brands. Companies provide options to sign up for the rewards/loyalty programs.</p>	<p>1) the company is directly involved in the business of operating loyalty/rewards programs for related travel brands</p> <p>2) the company focuses on operating loyalty/rewards programs for travel brands</p>	<p>customer is able to apply for loyalty/rewards card (or sign up for any similar types of bonus programs) without redirecting to other organization's website (for aggregators - on-site offers aggregation functionality is a minimum requirement)</p>

Other	Companies offering other auxiliary services for tourists/travelers, such as airport lounges booking, delivery and storage of luggage, SIM-cards, mobile apps for trip planning and toll road payment.	1) company's activities are not covered by any other vertical criteria 2) company's activities are related to travel industry 3) company is a direct provider (or a direct affiliate of the provider) of offered goods/services and is focused on offering those goods/services	customer is able to purchase/book auxiliary travel services that do not belong to any other vertical without redirecting to other organization's website (for aggregators - on-site offers aggregation functionality is a minimum requirement)
--------------	---	---	--

Source: [authors research]

For proper classification, all 3377 advertisers found in output datasets № 1 and 2 were analyzed through multiple dimensions:

- Additional check on whether advertiser is indeed related to the travel industry
- Analysis of content of the advertiser’s main website (what kind of goods/services are available for purchase/booking on the website; any explicit mentioning of offering certain goods/services, etc)
- Identification of whether advertiser’s offerings are provided by the advertiser directly, provided by other company but with direct support from the advertiser or aggregated from third-party providers
- Analysis of additional information provided about the advertiser and its’ business model on the advertiser’s website and in the third-party sources

As a result of the analysis and application of criterias, specified within the developed methodology, 298 advertisers were concluded to be unrelated to the travel industry. Other 3079 advertisers were confirmed to be travel advertisers and were assigned Primary and Secondary attributes of Tier 2 travel verticals. The results were approved by the Aviasales industry specialists and the output dataset № 4 was successfully created.

2.6. Data integration of the final dataset

The final integration of data included the collection of output datasets № 1 (direct affiliate travel links), № 3 (network affiliate travel links) and № 4 (travel vertical classification of brands) into a single data model. The process of final dataset integration also included the standardization of the received datasets and the joining of datasets.

2.6.1. Standardization of datasets

The main purpose of dataset standardization is to reach the required dimensionality of different sets of data and effectively join them in one entity. In this case, the vertical join of datasets required choosing similar parameters (columns) in joining parts. The choice of parameters was mainly based on their functionality and integrity - whether a parameter would serve any use for further analysis, as well as the completeness of such parameter (insignificant share of missing values for a parameter).

Table 28 below represents the main standardization procedures taken during the final dataset integration.

Table 28. Datasets standardization procedures

Procedure	Description
Drop operational columns	It was necessary to drop any intermediary columns that were created during the preprocessing stage. For example, the number of datasets from network affiliate advertisers were set as “ adv_id ”, which was required to find the advertiser domain by advertiser ID. Once the domains are found and assigned to the “ advertiser ” column, the parameter “ adv_id ” has no more use and is dropped.
Drop columns with many missing values	The number of parameters included observations with a considerable amount of missing values, which invalidates any use of such parameters. Such parameters were identified and dropped from the final dataset.
Drop technical web parameters	The list of affiliate link parameters (see Table 12) included 55 parameters, where a considerable part described the technical parameters of a web page, such as “ http_code ”, “ encoding ”, “ page_size ” and many others. Such parameters were not considered as valuable for further analysis, therefore they were dropped from the final dataset.
Add “ brand ” column	The majority of travel links were highly fragmented to regional top-level domains. For instance, the brand Airbnb was presented in 40 different regional domains, which made links “airbnb.com”, “airbnb.co”, “airbnb.es” and many others seem as different players in the travel industry. By joint decision of Aviasales and the project team, it was decided to drop the regional domains and look at the brands as a whole (e.g. “Airbnb”, “Skyscanner”, etc.). For this purpose, the “ brand ” column was added to output datasets № 1, 3 and 4 by dropping the top level domains (“.com”, “.ru”, etc.) in “ advertiser ” column using Python built-in functions (for function code see Figure 6 in Appendix).
Add “ source ” column	In order to track the source of affiliate links, it was decided to add a column “ source ”, which assigned the name of an affiliate network that the link belongs to or “ <i>direct advertiser</i> ” if the affiliate link is direct. Python built-in functions were used.

Source: [authors research]

The standardized parameters of output datasets № 1 and 3 included the following columns: (1) **url_from**, (2) **refdomain**, (3) **ahrefs_rank**, (4) **domain_rating**, (5) **ahrefs_top**,

(6) **ip_from**, (7) **links_internal**, (8) **links_external**, (9) **language**, (10) **url_to**, (11) **last_visited**, (12) **refdomains**, (13) **linked_root_domains**, (14) **traffic**, (15) **total_backlinks**, (16) **advertiser**, (17) **brand**, (18) **source**.

The standardized parameters of output dataset № 4 (travel vertical classification of brands) included the following columns: (1) **advertiser**, (2) **brand**, (3) **primary_attribute**, (4-28) *list of 24 parameters by names of secondary attributes* (for details see Table 29).

2.6.2. Join of datasets

After the standardization procedures, the output datasets № 1, 3 and 4 were joined to form a unified final dataset. The outputs joining procedure took two steps:

1. *Vertical join* - standardized output dataset № 1 (direct affiliate travel links) and output dataset № 3 (network affiliate travel links) were vertically concatenated using the PySpark build-in function.
2. *Horizontal join* - The concatenated dataset of output dataset № 1 and 3 were horizontally joined with output dataset № 4 by “**brand**” column using PySpark build-in function.

As a result of datasets standardization and joining procedure the output dataset № 5 was created - a unified dataset with cleaned and standardized data, which will be used further in BI dashboard development.

Table 29. Final dataset parameters

Parameter	Description and data type
url_from	the URL where the referral link was found and from which the information was collected (string)
refdomain	the main domain from which the information was collected (string)
ahrefs_rank	domain rating according to the Ahrefs methodology, which takes into account the size and quality of the backlinks (integer)
domain_rating	an indicator of the quality of the domain's backlinks relative to other domains in the Ahrefs database (integer)
ahrefs_top	domain rating relative to all sites in the Ahrefs database (integer)
ip_from	domain IP address (string)
links_internal	number of links to web pages within the domain (integer)
links_external	number of links to external web pages (integer)
language	the language of the web page (string)

url_to	referral link to the advertiser's website (string)
last_visited	the date of the last visit to the page by the search bot (string/datetime)
refdomains	the number of referral domains (integer)
linked_root_domains	the number of root domains (integer)
traffic	traffic for website (float)
total_backlinks	aggregated number of backlinks of the domain (integer)
advertiser (<i>new</i>)	The domain name of an advertiser (string)
brand (<i>new</i>)	The brand name of an advertiser (string)
source (<i>new</i>)	The name of the affiliate link source (string)
primary_attribute (<i>new</i>)	The name of the primary tier 2 travel vertical to which the travel brand belongs (string)
<i>list of 24 parameters by names of secondary attributes (new)</i>	1 - if travel brand has the secondary attribute (boolean)

Source: [authors research]

2.7. BI dashboard preparation

The second stage in this project is related to the construction of BI dashboard, which will provide visualization and insights on the competitive landscape, features and structure of the affiliate channel in the global travel industry.

2.7.1. Defining affiliate market characteristics

In order to construct the BI dashboard, most relevant characteristics that describe the current state of the global affiliate marketing channel were collected from Aviasales industry experts. In particular, Aviasales experts pointed to the following questions of interest and related indicators that could be derived from the data:

1. Which travel brands and travel verticals are actively using an affiliate marketing channel?
 - a. Top travel brands by total unique affiliate links / total affiliate partners
 - b. Top travel verticals by travel brands / affiliates
2. What languages prevail within the structure of the affiliate marketing channel in the travel industry?
 - a. Top languages by total unique affiliate links
3. How many travel affiliate programs do affiliates participate in?

- a. Minimum, median, average, maximum number of travel partners of affiliates links
 - b. Share of affiliates by total advertiser partnerships
4. What combinations of affiliate programs are used by affiliates, if any?
- a. Top travel vertical program combinations used by affiliates

In addition, Aviasales industry experts stated that competition within affiliate marketing channel is defined primarily by the affiliate program size (in terms of number of unique affiliates and links) and the domain rating of the affiliates. Brands with larger affiliate program size and high quality affiliates (measured by the domain rating) are considered the market leaders. Those metrics laid the ground for development of analytical frameworks - affiliate map, competitive quadrants, portfolio development ratio and affiliate ecosystem that are discussed in detail in part 3.3. “Competitive analysis”

2.7.2. BI tool requirements

In order to visualize the state of the global affiliate market and it’s competitive landscape, a number of visualizations were considered for implementation, such as single and clustered bar charts, pie/donut charts, tables, funnel charts, flow diagrams, scatter/bubble charts, network diagrams. All the visualizations were chosen according to the optimal chart selection scheme developed by Dr. Abela A. in 2006 (see Figure 8 below).

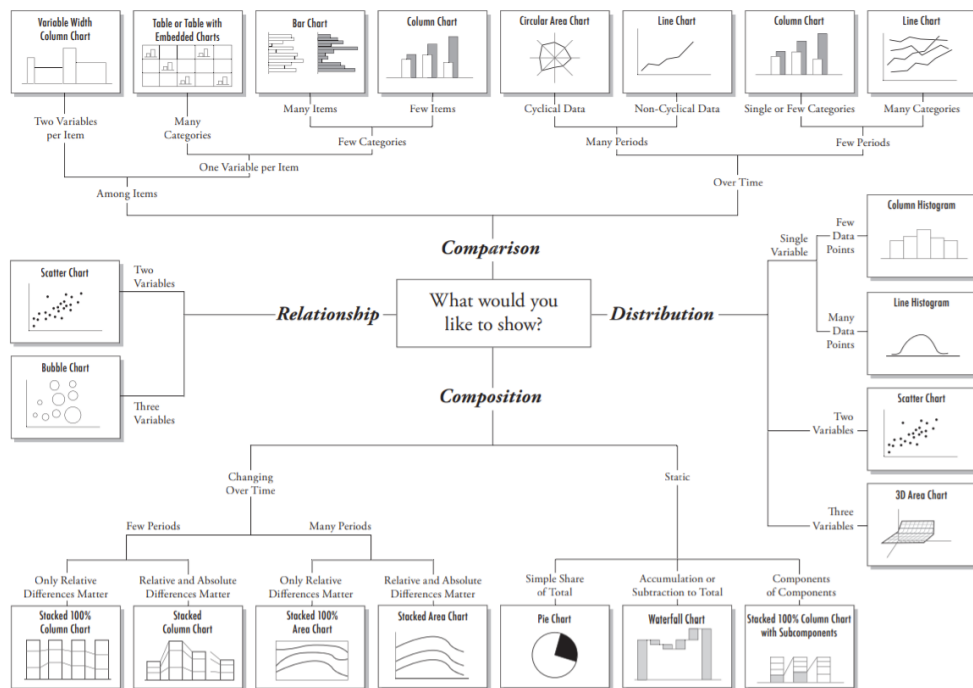


Figure 8. Chart Suggestions - A Thought-Starter

Source: [Abela, A., 2006]

Finally, the technical requirements regarding the BI tool and dashboard were defined in coordination with Aviasales specialists. Aviasales requested the following functional features of the future dashboard:

- User can launch the instrument and get ready-to-use statistics and visualizations on travel affiliate market characteristics
- User can navigate through the pages, which will contain the statistics with different areas of focus of travel affiliate market
- User can filter the content of the dashboard by applying the statistic slicers and filters by travel brands
- User can proceed to the detalization page for any individual travel brand, affiliate partner or travel vertical and get detailed information on it
- User can easily export the statistics content of the dashboard in CSV format

Functional and non-functional requirements for the BI tool were also identified (see Table 30 below).

Table 30. Functional and non-functional requirements of the BI tool

Functional requirements	Non-functional requirements
<ul style="list-style-type: none"> ● Data is presented in charts and tables ● Data can be cross-filtered by different parameters ● Each travel brand, affiliate and travel vertical is provided with detalization page ● Statistics data can be exported as CSV file 	<ul style="list-style-type: none"> ● Instrument can run on average local PC (MS Windows OS or MacOS) ● Interaction with visualization and slicers is time-efficient (no more than 3 seconds for loading) ● Instrument has a form of single file or a web version

Source: [authors research]

2.7.3. BI tool framework selection

Based on the collected functional and technical requirements, the BI dashboard must have the following characteristics:

1. *Convenience of use* - easy learning, use, error prevention and recovery, interaction efficiency, and accessibility features
2. *High performance* - good efficiency for working with large amounts of data, average-user system requirements
3. *Security* - physical security, data & software protection, low possibility of data loss or damage that may result from the use of the product

The process of BI platform selection starts from the analysis of the current market of business intelligence tools and its major players. As of February 2021, Gartner conducted

research indicating the main factors in BI tools industry, which are presented in a detailed matrix with narratives on the Figure 9 below (Gartner, 2021).

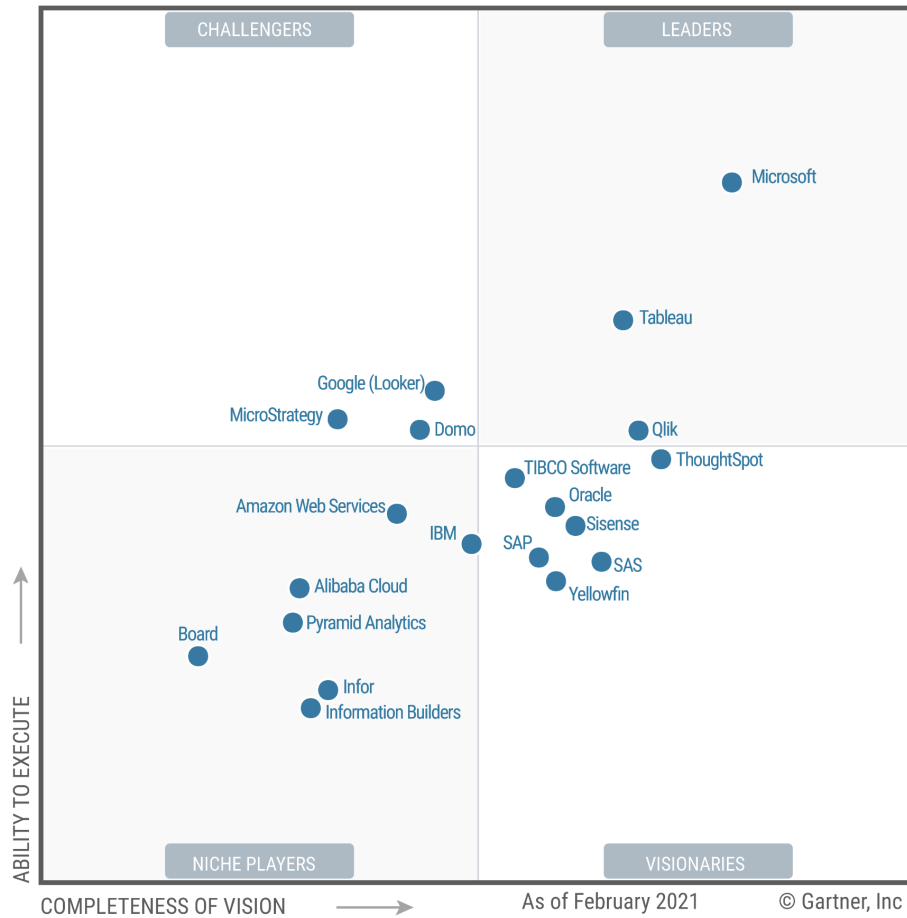







Figure 9. Gartner Magic Quadrant for Analytics and BI Platforms 2021

Source: [Gartner, 2021]

According to the research, the global market of analytical BI platforms is led by Microsoft Power BI, Tableau and Qlik Sense. The list of runner ups includes ThoughtSpot, Looker (Google) and other tools. In order to narrow down the search, the following table was prepared (see Table 31 below). The table provides a detailed comparison of the listed leading BI tools by their major functionalities (see functionalities description in Appendix Table 1)

Table 31. BI tools comparison

					
	Power BI	Tableau	Qlik Sense	ThoughtSpot	Looker
Full-featured free version	Yes	Separate tool	Separate tool	No	No
R/Python supported	Yes	Yes	Yes	R only	Yes
Dynamic cross-filtering	Yes	Yes	Yes	No	No
AI-enabled analytics	Yes	Yes	Yes	Yes	No
Search analytics with NLP	Yes	Yes	Yes	Yes	No
Data prep tools	Yes	Separate tool	Separate tool	Yes	No
Data modelling tools	Yes	Separate tool	Yes	Yes	Yes
Database independent	Yes	Yes	Yes	Yes	No
Built-in row level security	Yes	Yes	Yes	No	No
Mixed model types	Yes	No	No	Yes	No
Third-party data model access	Yes	No	No	No	No
Commenting & Collaboration	Yes	Yes	Yes	Yes	No
Embedded analytics	Yes	Yes	Yes	Yes	Yes
Open-source custom visualizations	Yes	No	Yes	No	Yes
Native mobile app	Yes	Yes	Yes	Yes	No

Source: [Petrossian, 2020]

Comparison results demonstrated the functional advantage of Microsoft Power BI over its major competitors. As the majority of the selected features are integral for successful visualization and effective reporting, Microsoft Power BI tool was chosen as the major BI platform to use.

Based on the BI platform choice, the following sections will discuss the dashboard design and construction in Power BI tool.

2.7.4. Dashboard design

The first step in analytical dashboard construction was the design of dashboard template, which includes the general structure of a page, the main components and applied functionalities. Figures 10 and 11 below present the general view of the dashboard template, which was designed and implemented in the instrument.

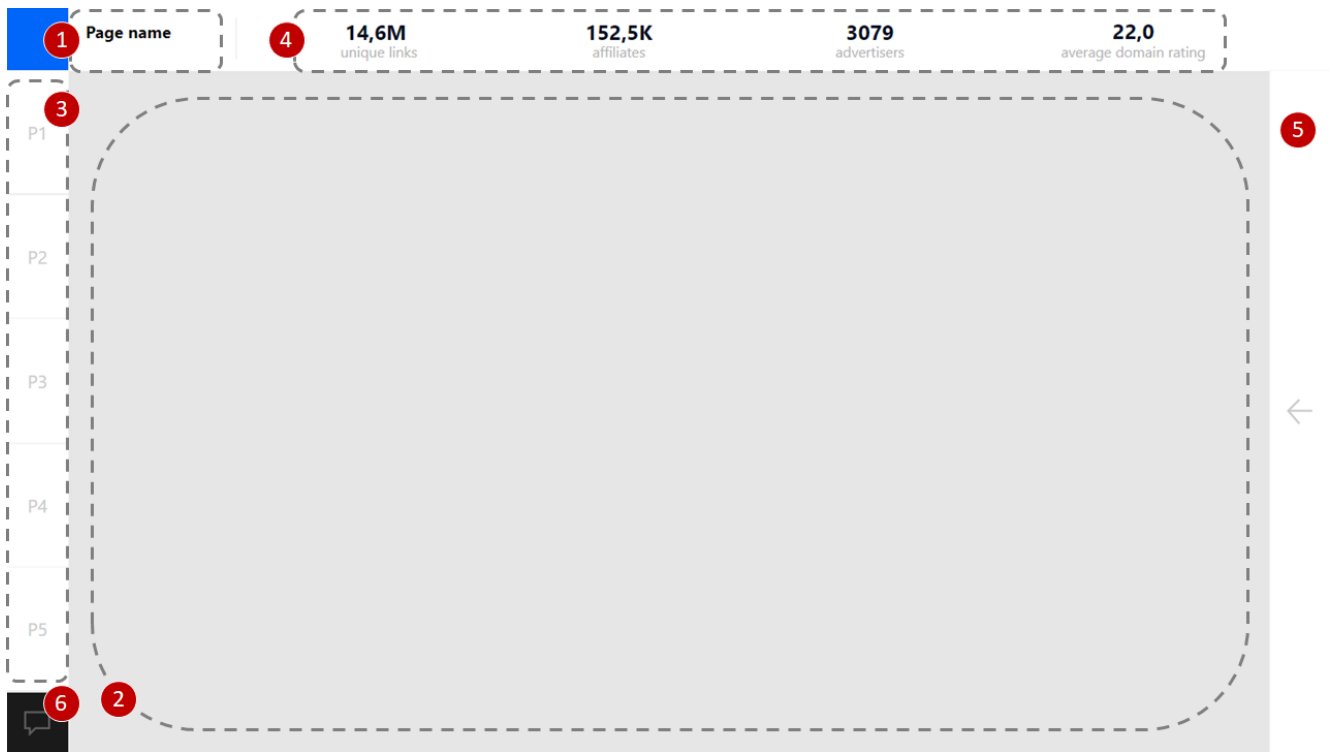


Figure 10. Dashboard design

Source: [authors research]

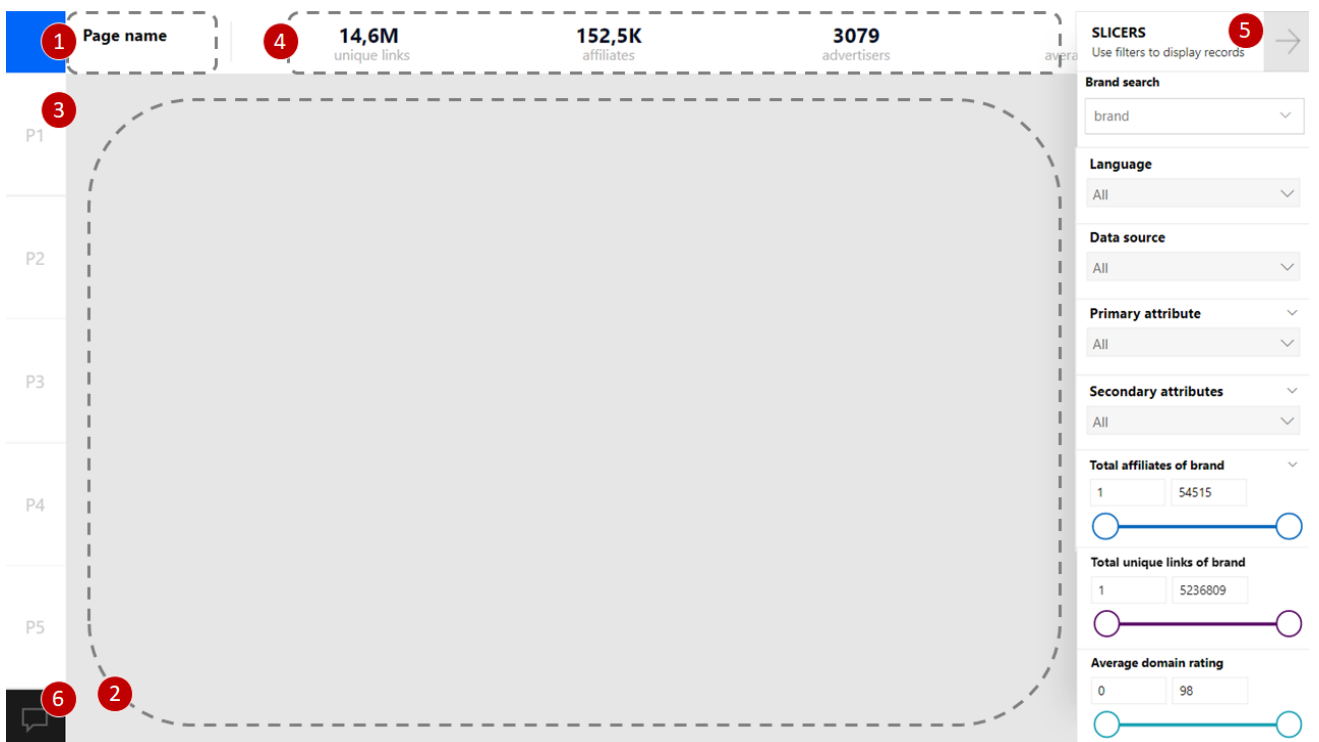


Figure 11. Dashboard design - slicers on

Source: [authors research]

Overall, the structure of the dashboard page includes 6 main components (see the numbered components on Figures 10-11):

- 1) The name of the object of the analysis presented on the page
- 2) The main content of the page, including charts, statistics and supporting information
- 3) The navigation panel to navigate through dashboard pages
- 4) The interactive descriptive statistics - shows statistics on the whole dataset or on the selected items (if any filters applied)
- 5) The interactive slicer panel to filter the main content by the preferred characteristics. The panel can be opened and closed using special buttons.
- 6) Search analytics using built-in NLP algorithm (e.g. the search “*show me top advertiser by the number of refdomains*” gives “booking.com” with 53 214 affiliates)

The main methods applied during the dashboard construction:

- *Dynamic cross-filtering* - interaction with charts by clicking on categories and brands will filter other charts and statistics presented on this page
- *Bookmarks* - interaction with buttons with bookmarks will change the visibility of content on the page (e.g. opening/closing slicer panel)
- *Tooltips* - hovering the cursor over brands and categories will pop up a small window with supporting statistics
- *Drill through* - brands and categories are provided with detail pages using drill through option on the charts
- *Measures* - the calculation of supporting variables based on the initial dataset

Based on the presented dashboard template, functionalities and methods, the following section provides the description of the main dashboard content and the questions of interest covered.

2.7.5. Dashboard content

Exact contents of the dashboard are presented within Table 32 below.

Table 32. Dashboard content description

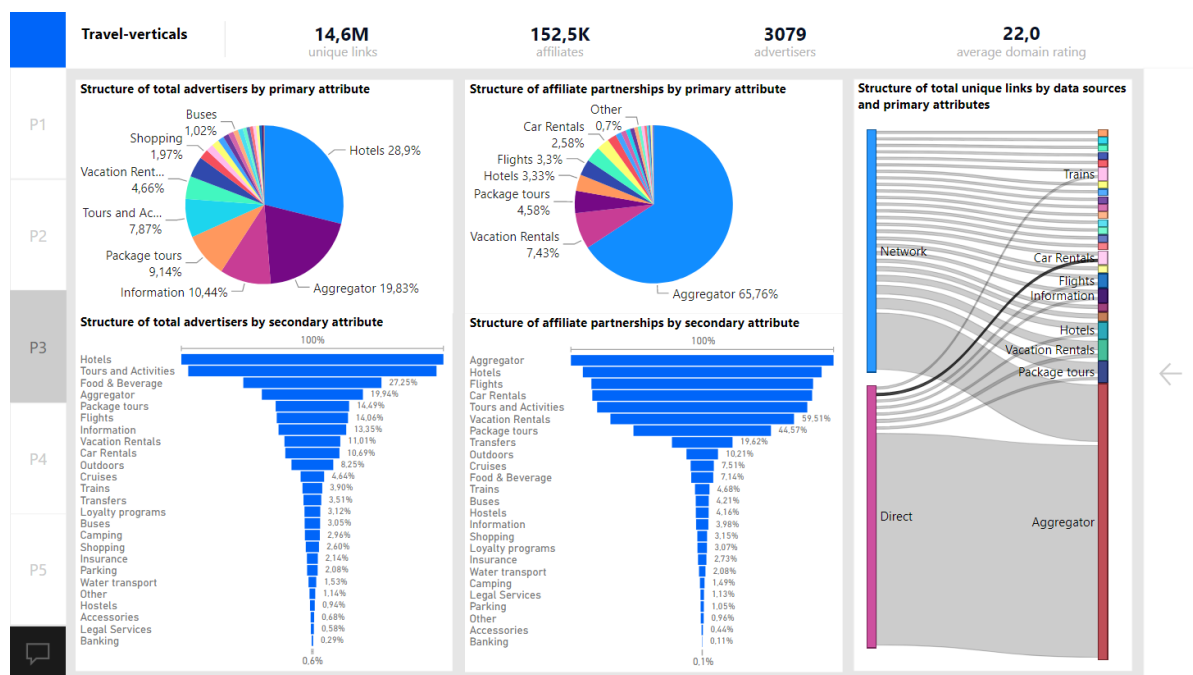
<i>Page 1 - Advertisers</i>																																																																							
Advertisers	14,6M unique links																																																																						
152,5K	3079 advertisers																																																																						
22,0	average domain rating																																																																						
P1	Top advertisers by total unique links 																																																																						
P2	<table border="1"> <thead> <tr> <th>Advertiser</th> <th>Unique links</th> <th>Affiliates</th> <th>Average domain rating</th> <th>Unique links/affiliates</th> </tr> </thead> <tbody> <tr><td>booking</td><td>5 236,8K</td><td>54,5K</td><td>26,58</td><td>96,06</td></tr> <tr><td>agoda</td><td>2 891,4K</td><td>19,8K</td><td>16,70</td><td>145,78</td></tr> <tr><td>hotelscombined</td><td>982,6K</td><td>12,4K</td><td>13,91</td><td>79,13</td></tr> <tr><td>rentalcars</td><td>793,2K</td><td>13,8K</td><td>26,46</td><td>57,52</td></tr> <tr><td>aviasales</td><td>679,0K</td><td>11,9K</td><td>9,26</td><td>57,20</td></tr> <tr><td>getyourguide</td><td>353,3K</td><td>7,8K</td><td>34,77</td><td>45,32</td></tr> <tr><td>skyscanner</td><td>340,2K</td><td>6,6K</td><td>8,56</td><td>52,57</td></tr> <tr><td>kiwi</td><td>307,0K</td><td>1,6K</td><td>6,29</td><td>194,18</td></tr> <tr><td>airbnb</td><td>281,3K</td><td>12,1K</td><td>57,23</td><td>23,21</td></tr> <tr><td>hotels</td><td>242,4K</td><td>4,8K</td><td>32,74</td><td>50,36</td></tr> <tr><td>tui</td><td>215,4K</td><td>2,8K</td><td>14,92</td><td>76,69</td></tr> <tr><td>viator</td><td>156,2K</td><td>1,5K</td><td>11,39</td><td>106,52</td></tr> <tr> <td>Total / Average</td> <td>14 554,5K</td> <td>152,5K</td> <td>21,97</td> <td>95,46</td> </tr> </tbody> </table>	Advertiser	Unique links	Affiliates	Average domain rating	Unique links/affiliates	booking	5 236,8K	54,5K	26,58	96,06	agoda	2 891,4K	19,8K	16,70	145,78	hotelscombined	982,6K	12,4K	13,91	79,13	rentalcars	793,2K	13,8K	26,46	57,52	aviasales	679,0K	11,9K	9,26	57,20	getyourguide	353,3K	7,8K	34,77	45,32	skyscanner	340,2K	6,6K	8,56	52,57	kiwi	307,0K	1,6K	6,29	194,18	airbnb	281,3K	12,1K	57,23	23,21	hotels	242,4K	4,8K	32,74	50,36	tui	215,4K	2,8K	14,92	76,69	viator	156,2K	1,5K	11,39	106,52	Total / Average	14 554,5K	152,5K	21,97	95,46
Advertiser	Unique links	Affiliates	Average domain rating	Unique links/affiliates																																																																			
booking	5 236,8K	54,5K	26,58	96,06																																																																			
agoda	2 891,4K	19,8K	16,70	145,78																																																																			
hotelscombined	982,6K	12,4K	13,91	79,13																																																																			
rentalcars	793,2K	13,8K	26,46	57,52																																																																			
aviasales	679,0K	11,9K	9,26	57,20																																																																			
getyourguide	353,3K	7,8K	34,77	45,32																																																																			
skyscanner	340,2K	6,6K	8,56	52,57																																																																			
kiwi	307,0K	1,6K	6,29	194,18																																																																			
airbnb	281,3K	12,1K	57,23	23,21																																																																			
hotels	242,4K	4,8K	32,74	50,36																																																																			
tui	215,4K	2,8K	14,92	76,69																																																																			
viator	156,2K	1,5K	11,39	106,52																																																																			
Total / Average	14 554,5K	152,5K	21,97	95,46																																																																			
P3	Top advertisers by total affiliates 																																																																						
P4	Language structure 																																																																						
P5	Data source structure 																																																																						
Description	<p>The page is devoted to the affiliate marketing activity by advertiser brands:</p> <ul style="list-style-type: none"> • Top left bar chart describes the top travel brands by the number of unique affiliate links • Bottom left bar chart describes the top travel brands by the number of unique affiliate partners • Top right table provides more detailed statistics on the number of unique affiliate links, unique affiliate partners, average domain rating of a partner and the average number of unique links per partner • Bottom left donut chart presents the structure of languages of the total unique affiliate links • Bottom right donut chart presents the structure of data sources (direct and network links) of the total unique affiliate links 																																																																						
Questions of interest covered	Questions 1 & 2																																																																						
<i>Page 2 - Affiliates</i>																																																																							

	<p>mode, average numbers of travel brands</p> <ul style="list-style-type: none"> ● Top center line & clustered column chart represents the share of total number of affiliates by quantity range of travel brand partners (in blue bar), the share of total number of unique links by the quantity range of travel brand partners (in purple bar) and the affiliates average domain rating by the quantity range of travel brand partners (in pink line) ● Bottom left donut chart represents the structure of unique affiliate links by travel vertical (primary attribute) ● Bottom right donut chart represents the structure of unique affiliate links by related travel brands <p><i>View 2:</i></p> <ul style="list-style-type: none"> ● Cloud of words represents the top travel brands by the number of unique affiliate partners
--	---

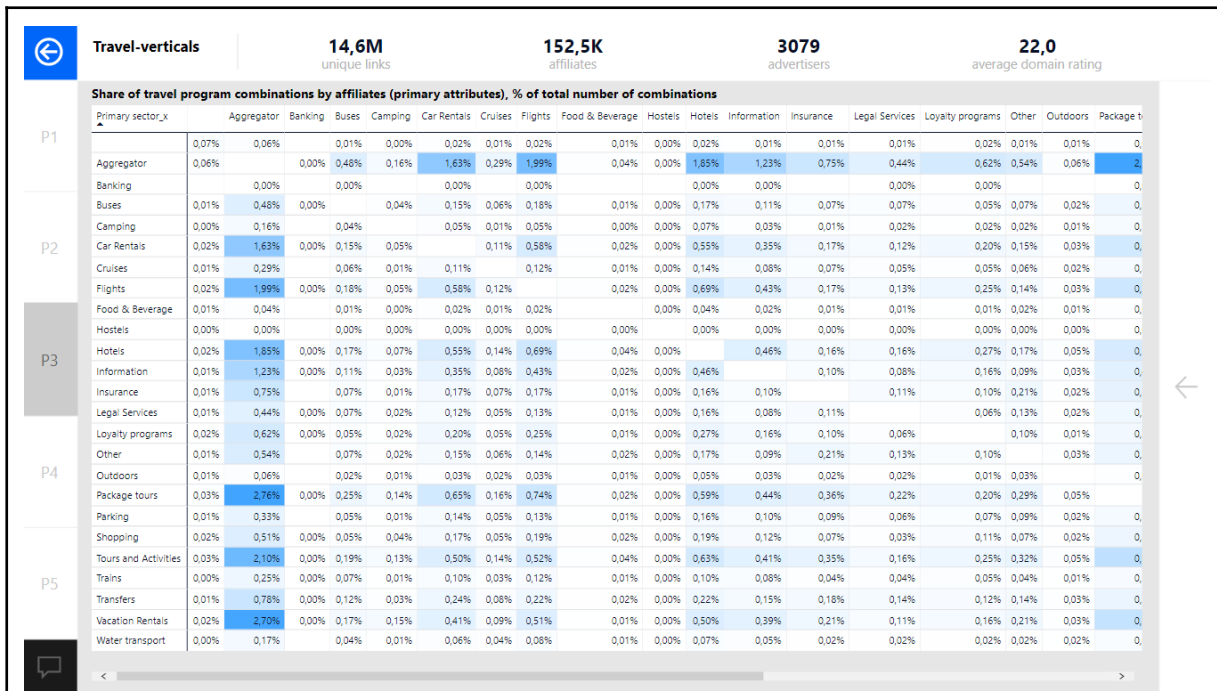
<p>Questions of interest covered</p>	<p>Questions 3 & 4</p>
---	----------------------------

Page 3 - Travel-verticals

View 1:



View 2:



Description

The page is devoted to the affiliate marketing activity by travel-verticals:
View 1:

- Top left pie chart represents the structure of travel-verticals (primary attribute) by the number of unique travel brands
- Bottom left funnel chart represents the averaged composition of travel brands by travel-vertical secondary attributes
- Top right pie chart represents the structure of travel-verticals (primary attribute) by the number of unique affiliates
- Bottom right funnel chart represents the averaged composition of affiliate partnerships with advertisers by travel-vertical secondary attributes
- Right flow diagram represents the distribution of unique affiliate links by their sources (direct/network) and travel-verticals (primary attribute)

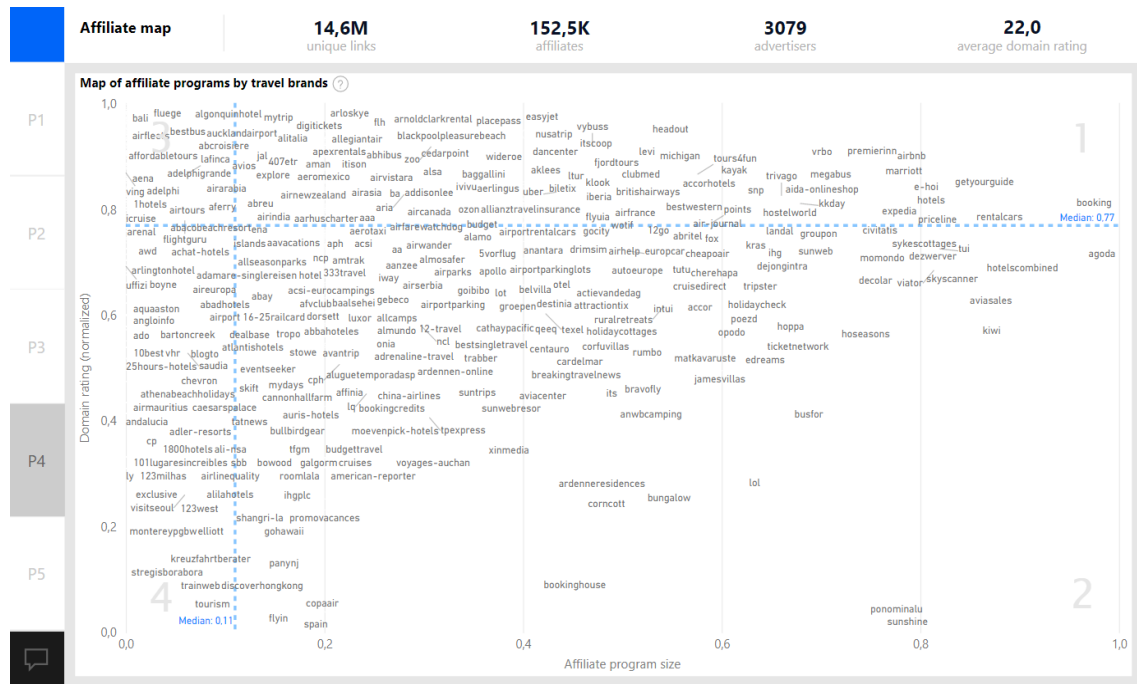
View 2:

- The matrix shows the top affiliate program combinations that affiliates use when partnering with travel brands
- The intersection of rows and columns shows the share (the popularity) of such combination from the total number of travel-vertical combinations in the samples

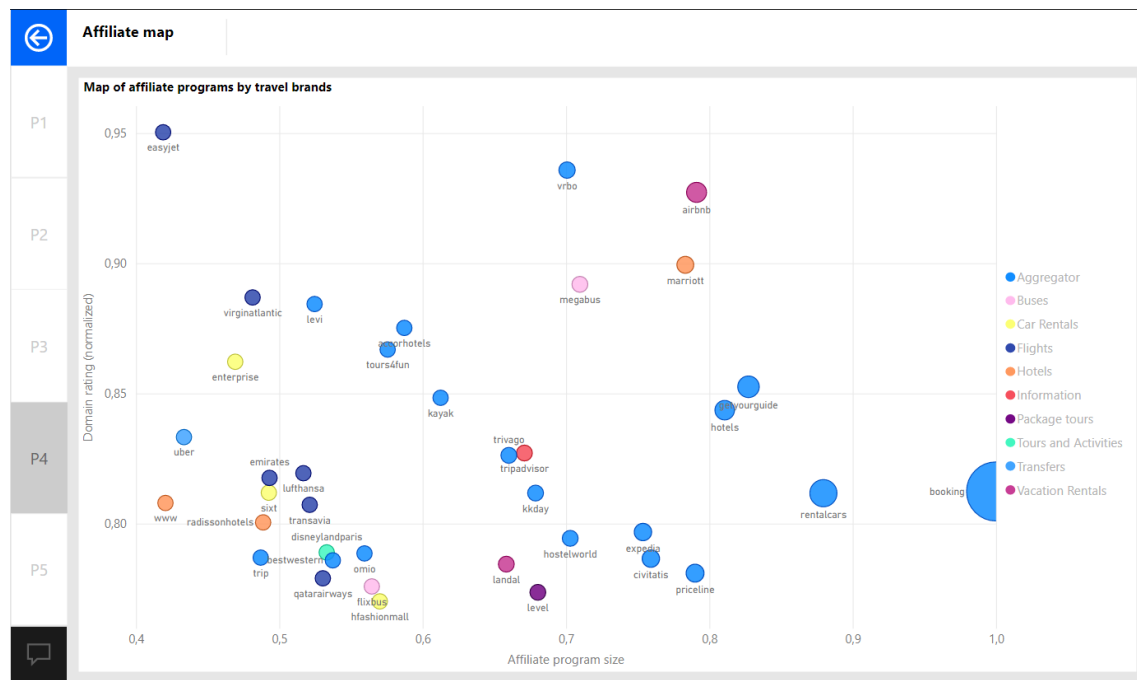
Questions of interest covered

Question 1, 4

View 1:



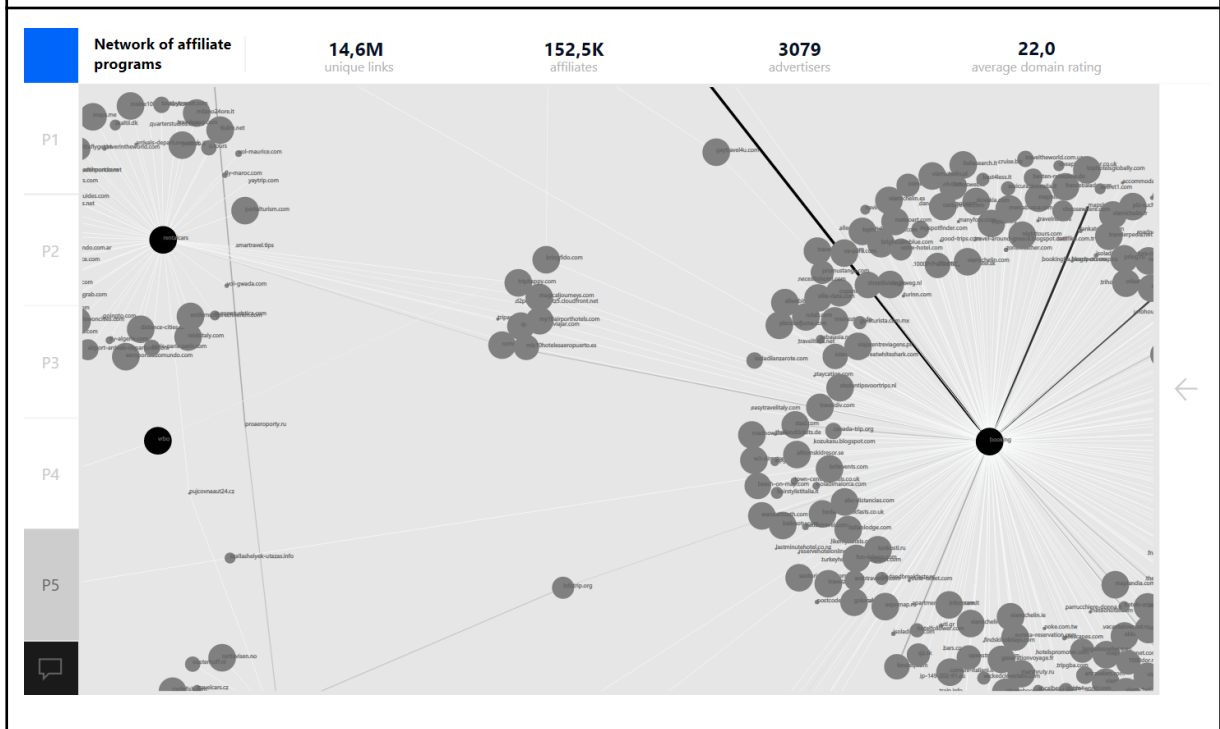
View 2:



	<p>advertiser</p> <p><i>View 3:</i></p> <ul style="list-style-type: none"> • The map represents the portfolio development ratio, which is the ratio of affiliate program size to median domain rating. $\text{Portfolio development ratio} = \text{NORM}(\text{Domain rating}) / \text{Affiliate program size}$ • The closer the ratio is to 1, the more balanced the program is: <ul style="list-style-type: none"> ○ Ratios with up to 25% variance (from 0.75 to 1.25) are well balanced ○ Ratios with 25-50% variance (from 0.5 to 0.75 or from 1.25 to 1.5) are moderately imbalanced ○ Ratios with >50% variance (<0.5 or >1.5) are strongly imbalanced • Advertisers in the lower part (ratio <1) have larger affiliate programs with lower affiliate ratings, whereas advertisers in the upper part (ratio >1) have smaller affiliate programs with higher median affiliate ratings
--	--

<p>Questions of interest covered</p>	<p>Question 1</p>
---	-------------------

Page 5 - Network of affiliate programs (affiliate ecosystem)



Description	<p>The page is devoted to the analysis of affiliate programs connections between the top travel brands (for details see 3.3.3. “Aviasales affiliate ecosystem”).</p> <ul style="list-style-type: none"> ● Black bubbles represent the advertisers. The size of a bubble indicates the total number of links that belong to the advertiser ● Grey bubbles represent the affiliates. The size of a bubble indicates the domain rating of this affiliate. ● Lines represent the connections between the advertisers and the affiliates. The darker the line is - the more links the advertiser has with this affiliate.
Questions of interest covered	Question 1

Drill through pages (Detailed overview)

View 1:

Advertiser detailization
aviasales

679,0K
unique links

11,9K
affiliates

9,3
average domain rating

Primary & Secondary attributes of advertiser

P1 | **Aggregator**

 | Aggregator

P2 | Flights

 | Hotels

P3

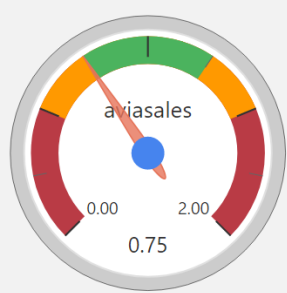
P4

P5

Top affiliates of advertiser

Affiliate	Unique links	Average DR	Partners
proaeroporty.ru	204517	0.00	2
pavliko.ru	58860	4.00	1
aviability.com	48958	19.00	1
aviastation.ru	44550	17.00	1
garlandus.ru	39000	0.00	1
mapsroad.ru	38652	4.00	3
goodwidgets.ru	37860	0.00	1
airetic.ru	16165	0.00	3
vandrouki.ru	15276	36.00	39
100500miles.ru	12642	25.00	21
travel-g.ru	8738	0.00	20
d3z.ru	8710	0.00	1
imigo.ru	8253	10.00	8
pirates.travel	7276	33.00	19
blograte.ru	6763	38.00	1
avia3.ru	6584	16.00	23
subscribe.ru	6342	78.00	19
lowcost.pro	4888	0.00	2
t.me	4464	94.00	72
triptodream.ru	4354	11.00	24
bookingroom.ru	3840	0.00	1
trip4you.ru	3566	15.00	19
sites.google.com	3434	93.00	128
Total / Average	678984	9,26	

Portfolio development ratio of advertiser



Aviasales

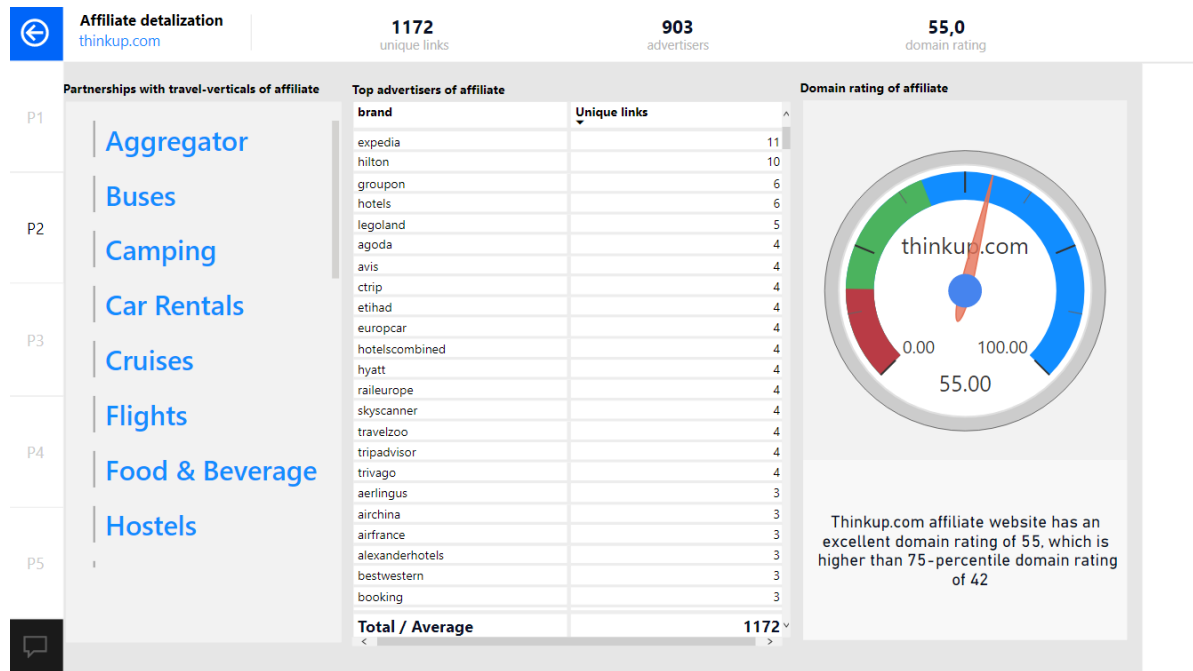
0.00 2.00

0.75

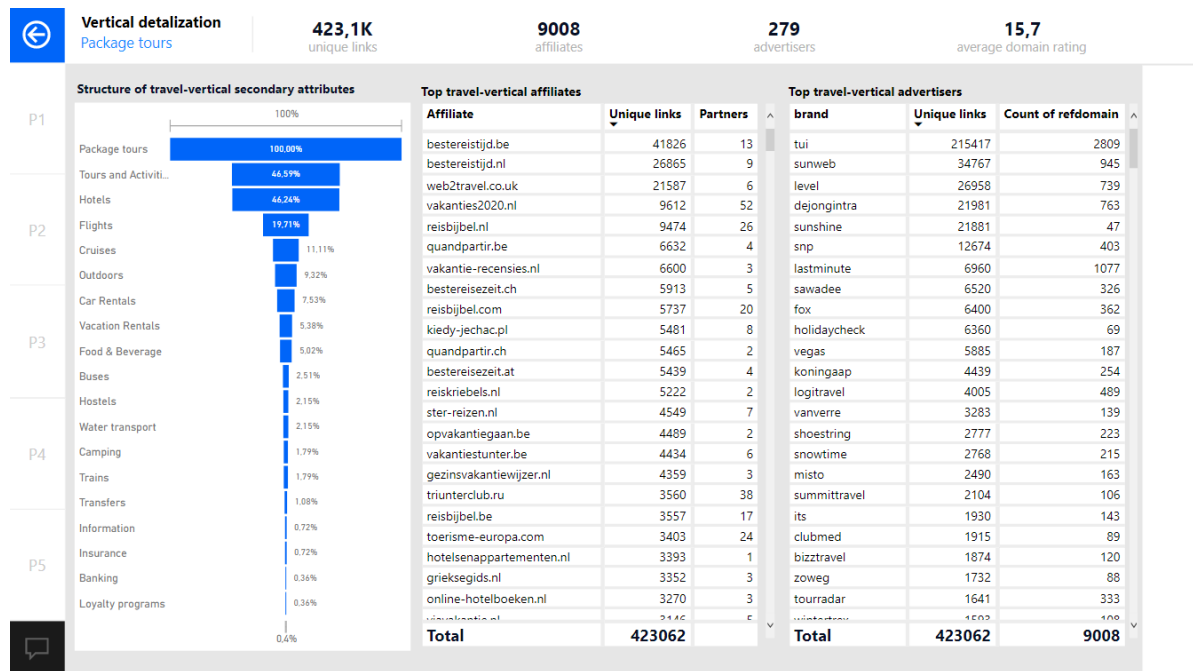
Aviasales company has a well balanced portfolio of affiliate partners.

At average affiliate domain rating of 9.26 the portfolio size of Aviasales constitute 11871 affiliates.

View 2:



View 3:



Description

The drill through (detailed overview) pages are devoted to the detailed analysis of the chosen travel company, affiliate partner or travel vertical according to the main affiliate characteristics.

View 1:

- The advertiser's drill through shows the detailed information on the advertiser's travel vertical (primary + secondary attributes), the top list of partnered affiliates and the general performance of the advertiser affiliate program

	<ul style="list-style-type: none"> • The performance of an advertiser is based on the Portfolio development ratio (see Page 4 - View 3 in table for detail) <p><i>View 2:</i></p> <ul style="list-style-type: none"> • The affiliates drill through shows the detailed information on the travel verticals, with which the affiliate works (only primary attribute), the top travel brands that works with the affiliate, and the performance of the affiliate based on the affiliate's domain rating • The analyzed domain rating is compared against the 1-st, 2-nd and 3-rd quartiles of median domain rating of the sample <p><i>View 3:</i></p> <ul style="list-style-type: none"> • The travel verticals drill through shows the information about the chosen primary attribute: its average composition of secondary attributes, the top affiliates that work with this travel vertical, and top travel brands belonging to this travel vertical.
Questions of interest covered	Question 1

Source: [authors research]

2.8. Conclusions on data mining & BI tool development process

This chapter discussed the main objectives completed in data mining and visualization development stages of the project. Major activities in this chapter included:

- 1) exploration of the main characteristics, features and problems of the initial datasets
- 2) discussion of the ways how the datasets of both direct and network affiliate links were preprocessed and integrated together
- 3) discussion of affiliate market characteristics to be analyzed; frameworks that were developed; BI software and dashboard requirements
- 4) development of Power BI dashboard, which became the basis for further data analysis and recommendation proposal.

The following Chapter 3 will cover the implementation of the prepared dashboard for assessing the state of global affiliate marketing in travel and providing Aviasales company with business recommendations regarding their affiliate marketing program.

Chapter 3. Evaluation and Deployment

The last chapter will focus on the use of a prepared dashboard for completing objectives № 10-12 (see Table 8). Based on the analysis of main characteristics of an affiliate marketing channel in travel and the analysis of competition, which Aviasales company

experiences in this channel, the corresponding recommendations and managerial implications will be derived.

As a main methodology for the analysis the EIC framework principles were used (see Figure 12 below). First, the affiliate marketing will be considered on the economy level, which includes the description of general characteristics of this channel and the main players on the global scale; on this level the stated questions of interest and metrics (see 2.7.1. “Defining affiliate market characteristics”) will be discussed and provided with descriptive statistics. Second, the market will be considered on the industrial level, which narrows down the scope to the particular travel verticals, in which Aviasales participates - on this level the main competitors of Aviasales will be analyzed and compared, as well as the possible connections of their affiliate programs. At third level, the analysis will be narrowed to the company level, which includes the descriptive statistics and interpretation of Aviasales affiliate marketing activity, as well as the strong and weak points of their program.

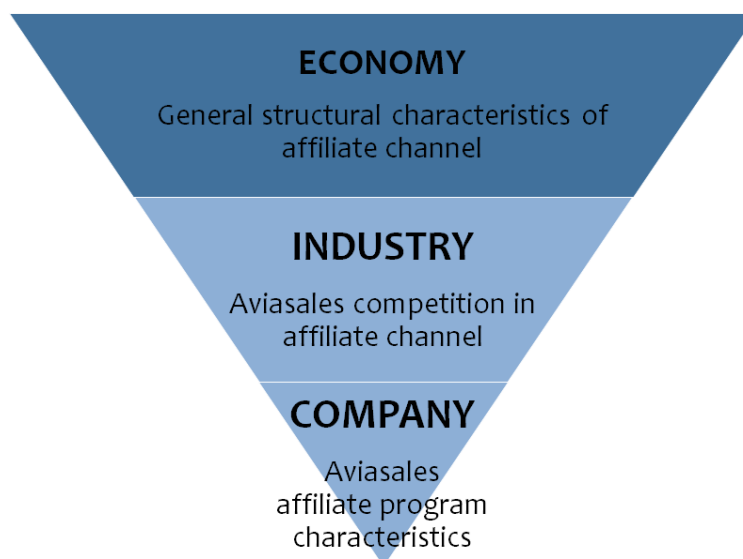


Figure 12. EIC framework
Source: [authors research]

Before moving to the analysis of the global affiliate marketing activity, the following section will describe the preliminary validation of the results received, the potential areas of concern and their validation procedures.

3.1. Validation of results

Final results of the data mining and visualization development stages were validated with the use of additional data sources and the expert evaluation from the Aviasales specialists. Based on the list of research limitations and assumptions outlined in section 1.4.1.

“Research assumptions & limitations”, the following summary of conducted validity checks was prepared (see Table 33 below).

Table 33. Summary of results validation

Area of concern	Validation procedure and outcome
Overall representability of the global affiliate marketing channel within the data	Representability was measured using the recent data (as of March 1, 2021) provided by Similarweb, on the global website traffic to world’s largest websites in the travel category. Analysis of the data showed that brands presented in the final dashboard cover around 60% of the total worldwide traffic to travel websites. Moreover, if adjusted for the travel brands that do not have affiliate programs, the brands presented in the dashboard account for more than 80% of the global travel traffic, with the largest brands presented. (Similarweb, 2021) Therefore, it can be concluded that the results of the analysis are valid and representative of the global travel industry.
Fullness and correctness of the total affiliate links data collected	Fullness and correctness of the total links data was confirmed by the experts from Aviasales and Ahrefs. Ahrefs representatives confirmed that their website crawler constantly updates their database with the data collected from the entire world wide web, and only a limited amount of links could have been missed (due to link having extremely low domain rating, being not indexed before, or having crawling restrictions) (Ahrefs, How often is Ahrefs links database updated?, n.d.). Aviasales representatives confirmed that the Ahrefs data mostly matches their internal data regarding Aviasales links parameters. Therefore, as was stated in the assumptions and limitations section of the research, possible number of unrepresentative/missing links (due to links becoming inactive or not being indexed by the Ahrefs crawler) is not expected to be significant enough to influence the overall conclusions and managerial implications drawn from the final dashboard.
Validity of the summary information, visuals and interpretations presented in the final dashboard	Summary information and dashboard’s visuals validity were approved by the Aviasales industry experts. Interpretations, assumptions and frameworks applied throughout the analysis were communicated to the Aviasales team, and their approval received. Thus, the information and visuals presented within the dashboard can be assumed to be correct, precise and appropriate for business decision-making.
COVID-19 impact and duration of results relevance	As was previously mentioned, the impact of COVID-19 pandemic is already built-in within the data, as it was collected in Fall 2020, with affiliate marketing players already adapted to the new market conditions. In the long-run, the industry is expected to stabilize and return to average growth of around 10% annually, as was observed before the pandemic (Mhojhos Research, 2020). Regarding the duration of data relevance, the general conditions within an affiliate marketing channel are expected to react proportionally to brands’ web traffic dynamics - as affiliate marketing is in essence a traffic-generating tool. In order to make projections on the web traffic dynamics, the traffic data for top-10 travel

	<p>brands in the affiliate channel was taken from Similarweb. Monthly traffic changes (both desktop and mobile versions) were evaluated for the period from November 2020 to March 2021. (Similarweb, 2021) Correlation matrix between monthly traffic for all brands shows relatively strong correlation of 66.5% (see Figure 12 in the Appendix section). Therefore, it can be assumed that all travel brands traffic dynamics must have similar statistical parameters, with those top-10 brands having the most impact over the global affiliate channel. This assumption provided the basis for conducting a Monte Carlo simulation to test the possible time period, within which the data will be relevant enough to support the managerial decision making. The cumulative weighted traffic change to top-10 brands was iterated randomly 10000 times using normal distribution, and tested against the critical threshold of 2 standard deviations from the historical monthly mean change for all top-10 brands. It was assumed that cumulative change in traffic for all top-10 brands in excess of 2 standard deviations will be indicative of structural changes within an affiliate marketing channel and will require updating the data within the dashboard. The results of the simulation indicate that the current data can be assumed to be relevant within a range of 5-16 months, with mean duration of 16.5 months and mode of 8 months (see Figure 13 in the Appendix section). Results and methodology were approved by the Aviasales industry experts. Thus, the information presented within the dashboard can be assumed to be representative of pandemic impact and rigid enough to stay relevant within reasonable time, allowing for conducting additional analysis and management decisions.</p>
--	---

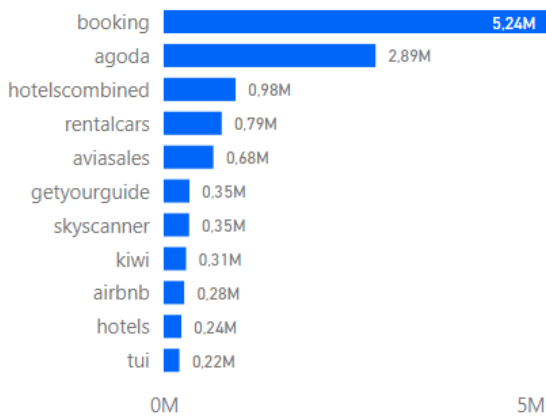
Source: [authors research]

Overall, all of the major areas of concern within this research were addressed and checked for validity. Final results are accepted for answering the questions and addressing the business objectives of Aviasales.

3.2. Affiliate market analysis

As a result of the analysis, a total of 3079 travel brands were identified as actively using the affiliate marketing channel. The top travel brands by the number of unique affiliate links and the number of affiliates are presented in the Figure 13 below.

Top advertisers by total unique links



Top advertisers by total affiliates

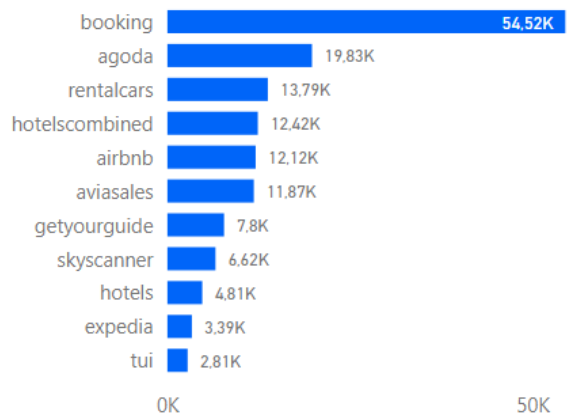


Figure 13. Top advertisers by total unique links and affiliates

Source: [authors research]

The top 11 travel brands from the above Figure cover 84% of all unique affiliate links and 61% of total affiliates.

In regards to travel verticals, affiliate marketing is mostly used by hotels (28.9%), travel aggregators (19.8%) and travel information websites (10.4%) (see Figure 14 below).

STRUCTURE OF TOTAL ADVERTISERS BY PRIMARY ATTRIBUTE

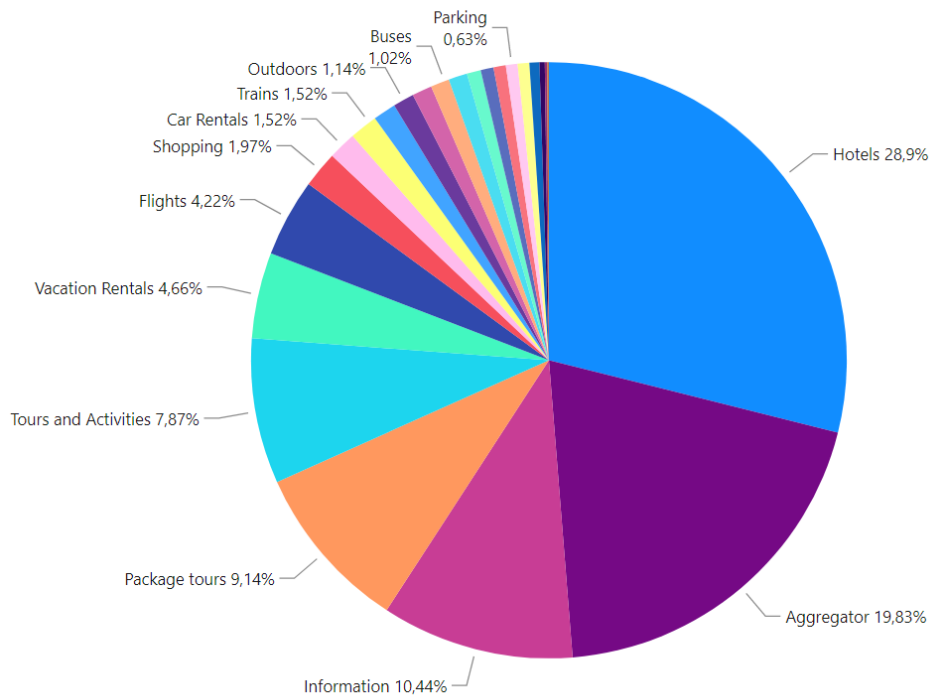


Figure 14. Structure of travel advertisers by vertical (primary attribute)

Source: [authors research]

In addition to those primary services, most of the travel brands within affiliate marketing channel also provide hotel bookings (51.7%), tours & activities (49%) and food & beverages (27.3%) (see Figure 15 below).

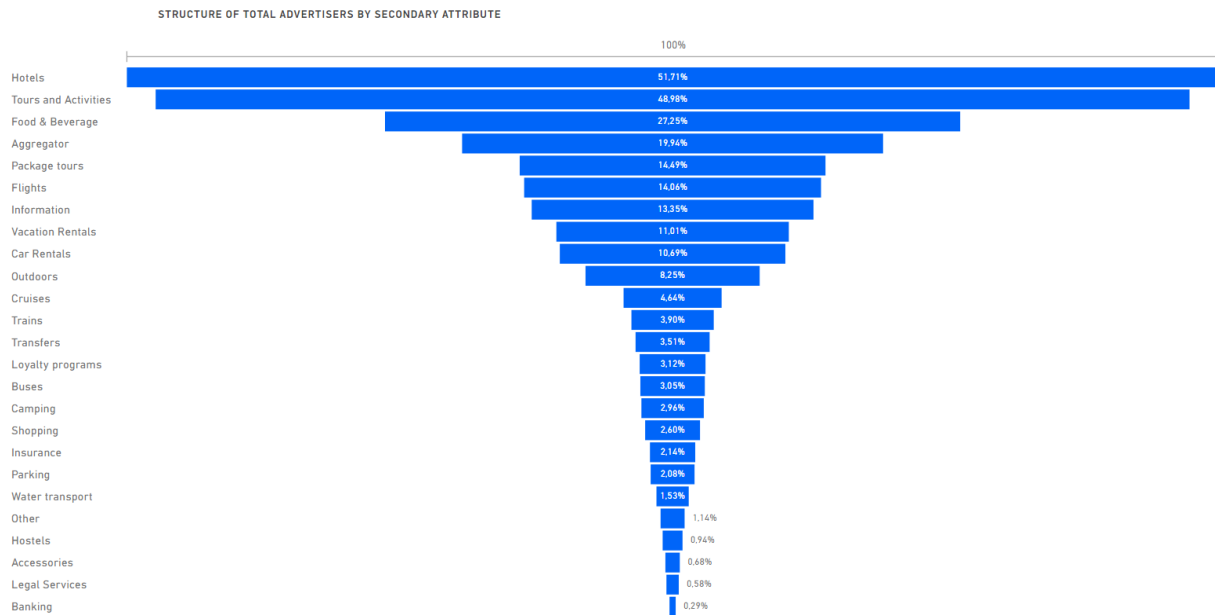


Figure 15. Structure of travel advertisers by vertical (secondary attribute)

Source: [authors research]

As for the affiliates, the vast majority of them partner with travel aggregators (65.8%), followed by vacation rentals (7.4%) and package tours (4.6%) (see Figure 16 below).

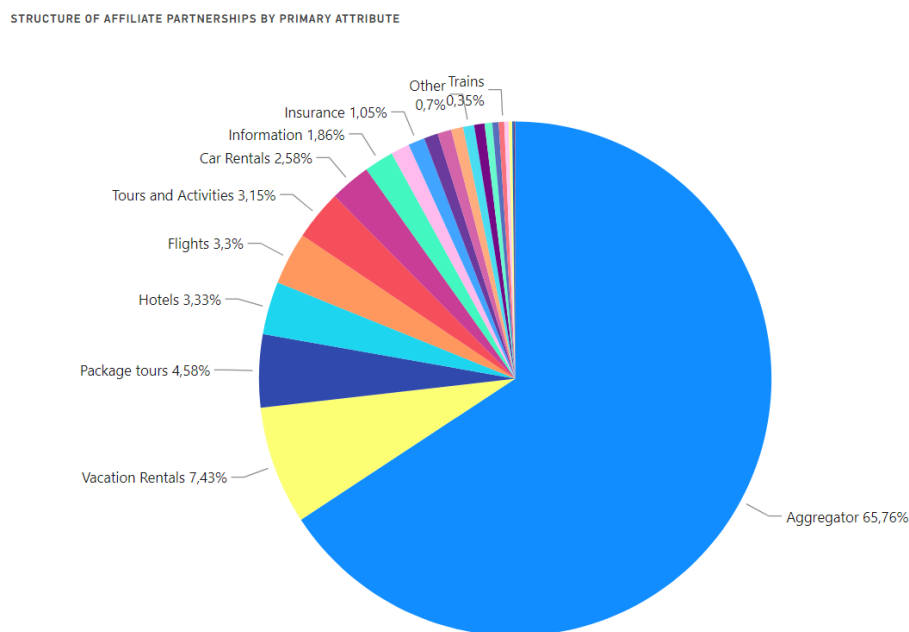


Figure 16. Structure of affiliate partnerships by vertical (primary attribute)

Source: [authors research]

The most popular secondary travel services that dominate affiliate partnerships include hotel bookings (77.4%), flight bookings (71.9%) and car rentals (71.3%) (see Figure 17 below).

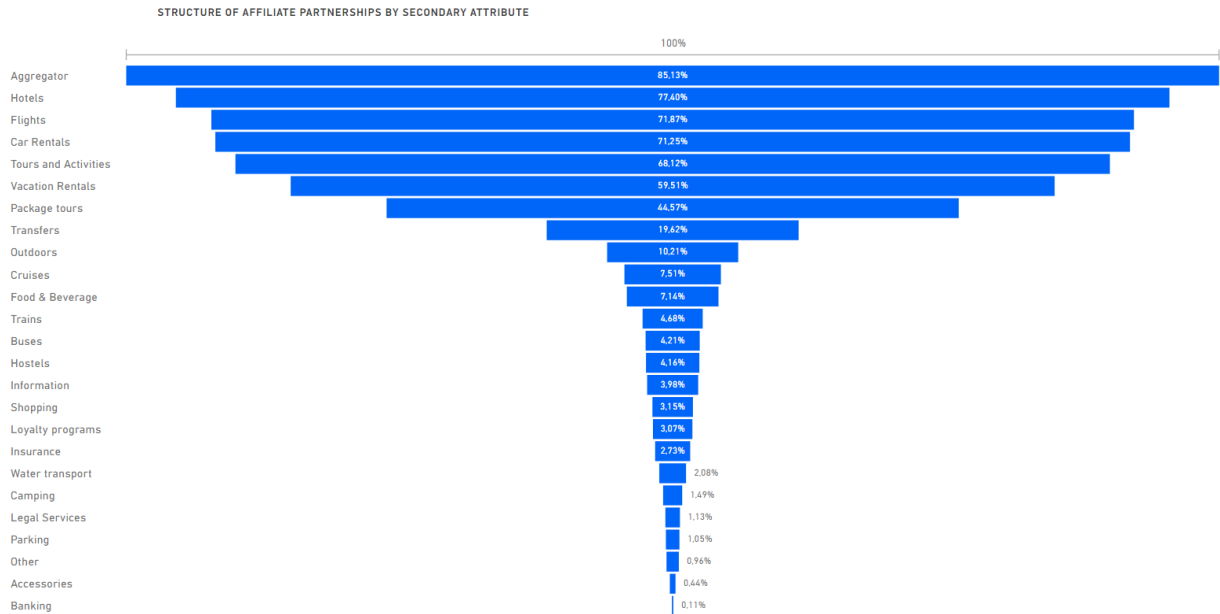


Figure 17. Structure of affiliate partnerships by vertical (secondary attribute)

Source: [authors research]

Finally, the majority of the unique affiliate links belong to the travel aggregators (90.5%), followed by the package tours (2.9%), and vacation rentals (2.5%) (see Figure 18 below).

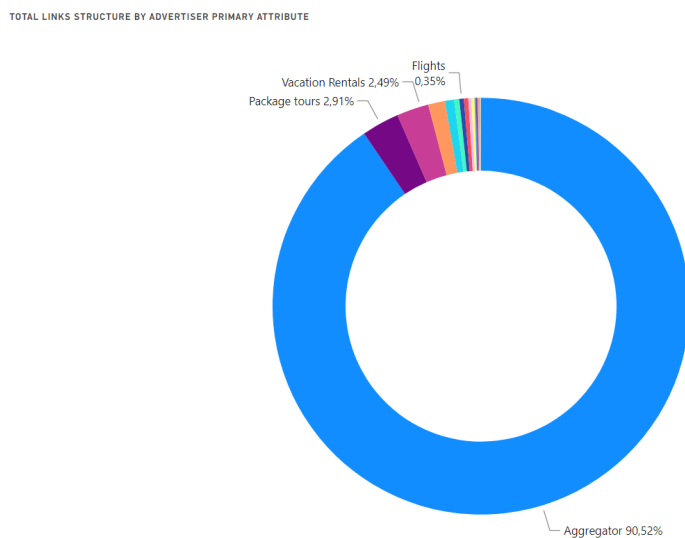


Figure 18. Structure of unique affiliate links by vertical (primary attribute)

Source: [authors research]

The analysis of data revealed that most of the affiliate links are featured at the English-speaking websites (47.5%) followed by the Russian (13.3%), German (6.2%), Dutch (4.3%) and Spanish (4.2%) websites (see Figure 19 below).

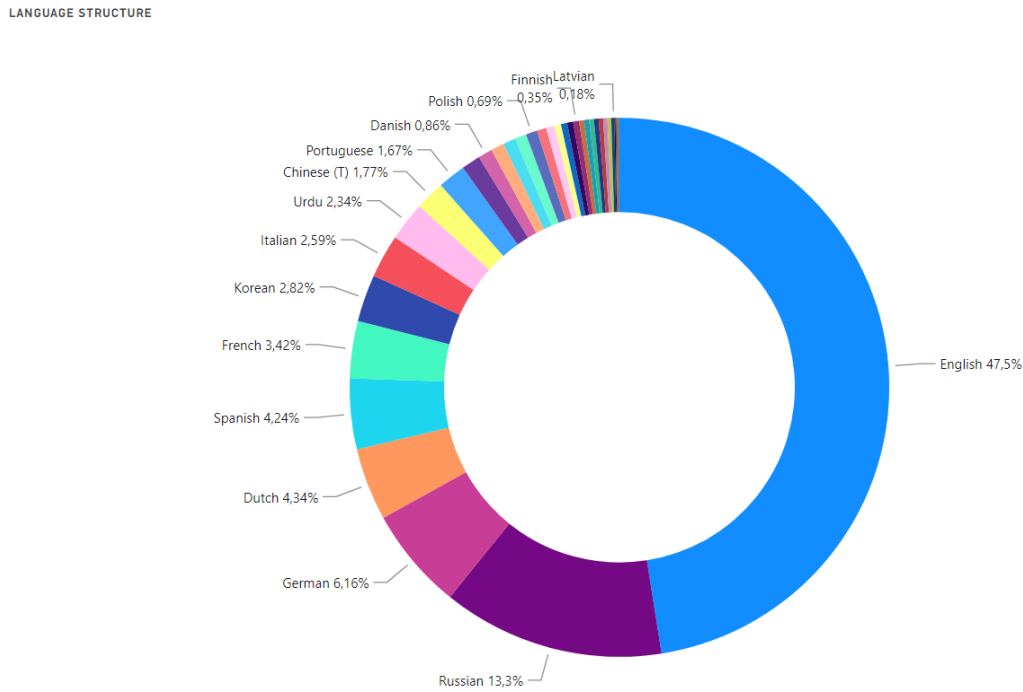


Figure 19. Structure of unique affiliate links by language

Source: [authors research]

Overall, it can be concluded that travel affiliate marketing is a popular extra income source among European, English-speaking (UK, US, Canada, Australia) and Spanish-speaking (Spain & Latin America) affiliate websites.

Furthermore, the number of affiliate programs in which affiliates participate is ranging from 1 to 903, with the average number of partnerships being 1.93.

The majority of affiliate partnerships feature only one advertiser partner (75.1% of all affiliates have only 1 travel brand in partnership), which covers 42.6% of all affiliate links. In addition, 98.3% of all affiliates have partnerships with 1 to 10 travel brands that cover 90.2% of total unique affiliate links. Finally, the analysis determined the positive correlation between the number of travel advertisers in partnership with the affiliate and the average domain rating of the affiliate websites - the more partners the affiliate has, the higher is the rating of this affiliate (see Figure 20 below).

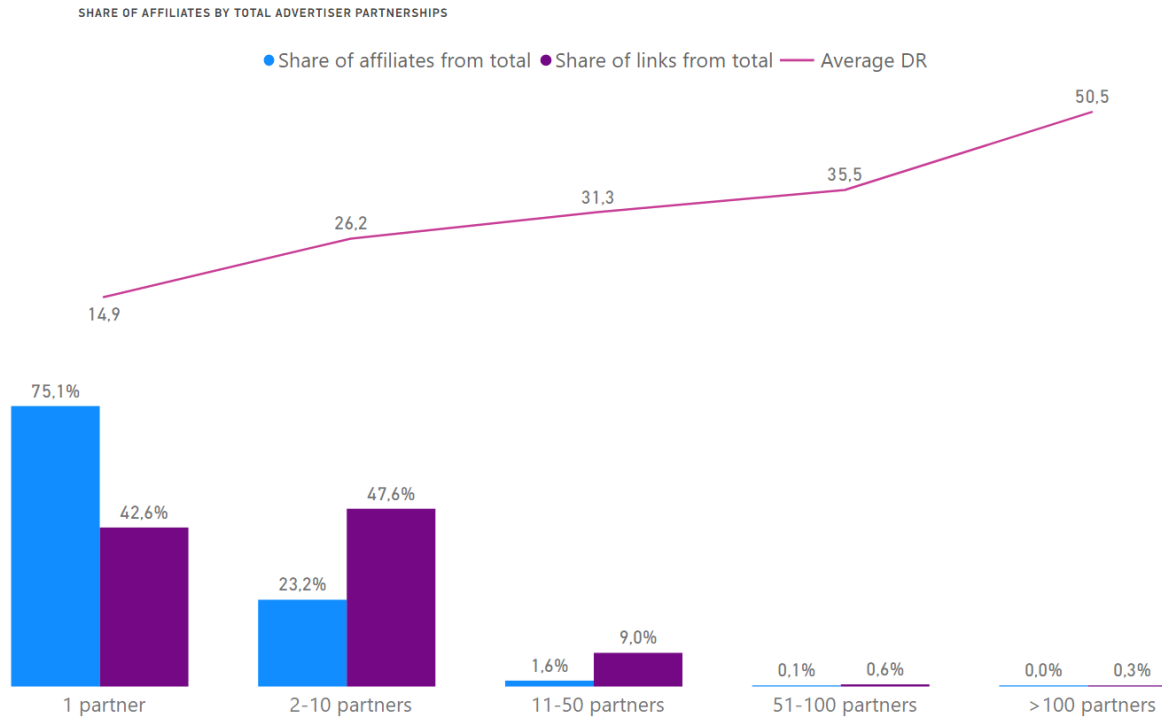


Figure 20. Share of affiliates by total advertiser partnerships

Source: [authors research]

To sum up, it can be concluded that the majority of affiliate programs in travel are small and feature from 1 to 10 unique affiliates. In addition, those brands that manage larger affiliate programs have, on average, a better quality profile of affiliates (measured by their domain rating).

Finally, it was determined that there are a total of 552 various combinations of travel verticals existing in affiliate partnerships. The most popular one is a combination of programs from travel aggregators and online travel agencies that offer packaged products (3.12% of all total existing combinations). Travel aggregators programs are also frequently combined with direct hotel bookings (3.11%), direct flight ticket bookings (3.11% of all total existing combinations) and direct tours & activities bookings (2.83% of all total existing combinations). Among non-aggregator programs, the most frequently used combinations include direct hotel & flight bookings (1.04% of all total existing combinations), travel packages & flight bookings (0.9% of all total existing combinations) and hotels & travel packages bookings (0.86% of all total existing combinations) (see Figure 21 below).

Vertical combination	% of all vertical combinations
Aggregator-Package tours	3.12%
Aggregator-Hotels	3.11%
Aggregator-Flights	2.83%
Aggregator-Tours and Activities	1.71%
Aggregator-Information	1.38%
Aggregator-Car Rentals	1.34%
Flights-Hotels	1.04%
Aggregator-Vacation Rentals	1.01%
Flights-Package tours	0.90%
Package tours-Hotels	0.86%
Package tours-Tours and Activities	0.78%
Aggregator-Buses	0.74%
Hotels-Tours and Activities	0.68%
Car Rentals-Hotels	0.62%
Hotels-Information	0.61%

Figure 21. Most popular combinations of travel verticals used in affiliate marketing channel

Source: [authors research]

3.3. Competitive analysis

3.3.1. Competitive quadrants

The affiliate map developed for the dashboard became one of the foundations in assessment of competitive positions of travel affiliate programs. As discussed in the 2.7.5. “Dashboard content”, the travel brands are presented on the map with the size of their affiliate program on X axis and the median domain rating of their affiliates on Y axis. Resulting visualization provides an overview of the competitive landscape in the affiliate marketing channel, where each brand can be set against the others to compare its relative position within the market.

In order to separate the brands according to their competitive profile, a special framework was developed - it was called “competitive quadrants”. The concept of the framework involves dividing the total affiliate market into the separate sub sectors, known as quadrants, which represent the groups of market participants with comparable profiles. The median values of axes are taken as the thresholds for market separation, which form the four major quadrants (see Figure 22 below).

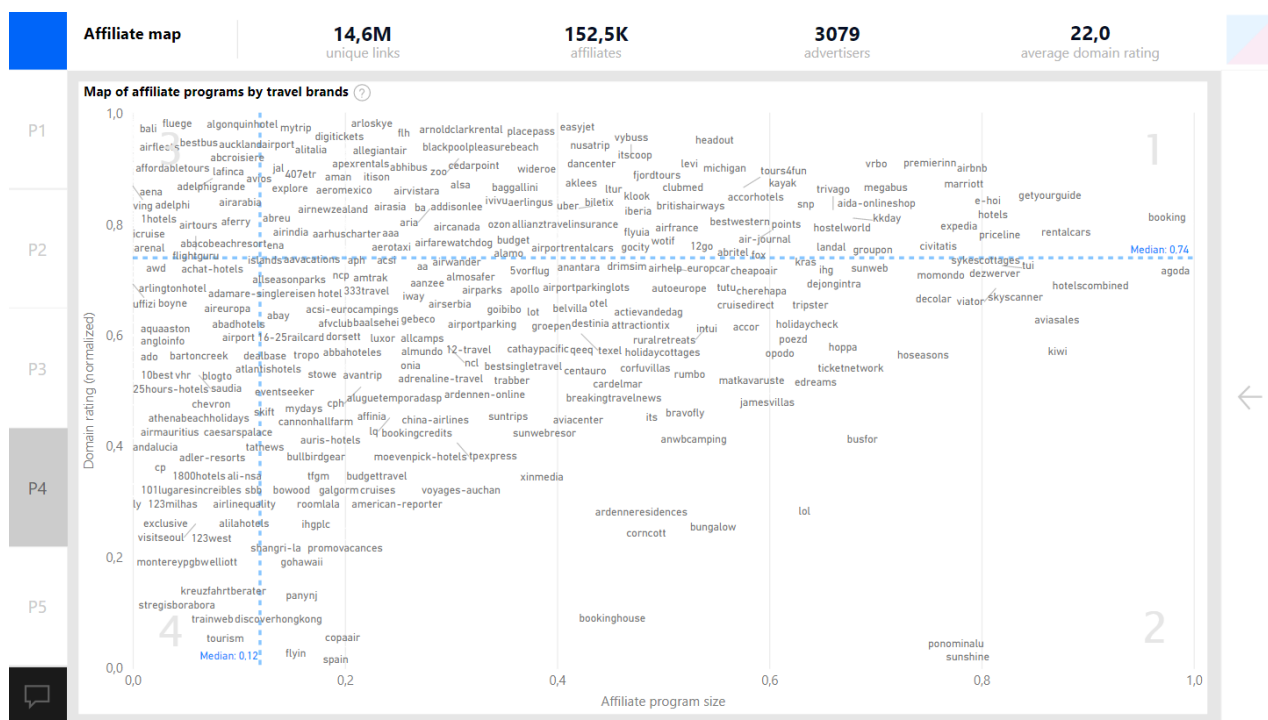


Figure 22. Competitive quadrants on affiliate map

Source: [authors research]

The following table describes the characteristics of the four competitive quadrants with regards to the total affiliate marketing channel (see Table 34 below).

Table 34. Competitive quadrants characteristics

Competitive quadrant	Details	Description
Quadrant №1 Leaders	Quadrant definition	The quadrant includes leading companies in the market, with big portfolios of affiliates and high median affiliate ratings. The quadrant is mostly occupied by the global well-known travel brands.
	Descriptive statistics	<ul style="list-style-type: none"> ➤ 11.1 mln unique links ➤ 118.4K affiliates ➤ 859 advertisers ➤ 25.4 median affiliate domain rating
	Featuring participants	Booking, Airbnb, Expedia, Agoda, etc.
Quadrant № 2 Runners-up	Definition	The quadrant includes market runners-up companies, featured by big portfolios of affiliates, but with lower median affiliate ratings. The quadrant features some well-known global travel brands, as well as the brands with more regional presence.
	Descriptive statistics	<ul style="list-style-type: none"> ➤ 3.5 mln unique links ➤ 56.1K affiliates

		<ul style="list-style-type: none"> ➤ 686 advertisers ➤ 11.0 median affiliate domain rating
	Featuring participants	Aviasales, Tui, Skyscanner, Kiwi, etc.
Quadrant №3 Niche Players	Definition	The quadrant includes small companies, featured by small portfolios of affiliates, but with high median affiliate ratings. The quadrant does not feature any globally popular travel brands, but rather small brands with local operations.
	Descriptive statistics	<ul style="list-style-type: none"> ➤ 1629 unique links ➤ 409 affiliates ➤ 694 advertisers ➤ 37.2 median affiliate domain rating
	Featuring participants	Ocean-Florida, Islandbuses, Dollarflightclub, etc.
Quadrant №4 Underdogs	Definition	The quadrant includes market outsiders, characterized by small portfolios of affiliates and low median affiliate ratings. The quadrant does not feature any globally popular travel brands, but rather small brands with local operations.
	Descriptive statistics	<ul style="list-style-type: none"> ➤ 1703 unique links ➤ 422 affiliates ➤ 840 advertisers ➤ 3.8 median affiliate domain rating
	Featuring participants	Travelmarket, K-west, Flyswoop, Adamare, etc.

Source: [authors research]

As a supplement to the competitive quadrants framework, the companies on the affiliate map are analyzed according to their *portfolio development ratio* (see details in 2.7.5. “Dashboard content”), which indicates how balanced the affiliate program is according to its size and the quality (median domain rating of affiliates). As mentioned previously:

- The companies that are close to the ratio line (ratio ~1) have balanced programs
- The companies that are significantly above the ratio line (ratio >1) have unbalanced programs with smaller affiliate portfolios, but higher ratings of affiliates
- The companies that are significantly below the ratio line (ratio <1) have unbalanced programs with bigger affiliate portfolios, but lower ratings of affiliates

As a part of the developed methodology, the analysis of the best practices of leading travel brands in the sample showed that the balance between the quantity and the quality of affiliates may dictate the effectiveness of the affiliate programs. According to the analysis, neither of the top leading travel companies in the affiliate channel have any significant

imbalance in their programs (see Figure 23 below). The imbalance may cause the inefficiency of the program and therefore lower potential profitability.

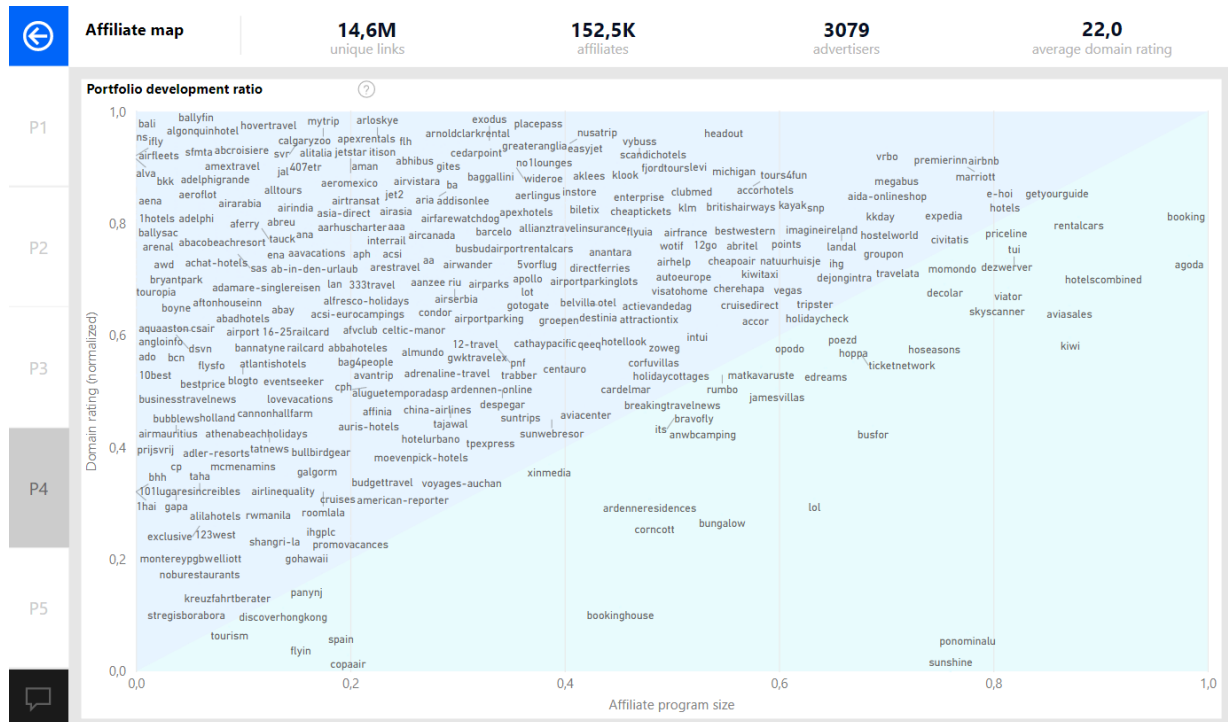


Figure 23. Affiliate map - portfolio development ratio

Source: [authors research]

3.3.2. Aviasales competition

During the development of travel vertical methodology (see 2.5. “Data construction of the travel verticals classification”) and the following classification of advertisers, each travel brand was assigned to Primary and Secondary attributes, which can define the scope and direction of their operations within the market. Based on the attributes we can differentiate companies by their core business, which will define the borders of direct competition for these companies.

For Aviasales, the competition will be considered among the travel aggregators (Primary attribute) that sell flight tickets (Secondary attribute). As a remark, since Aviasales still has the Secondary attribute of “Hotels” category, it will not be accounted for in the competitive analysis, as this category generates the minority share of sales and is not reflected in the company’s major positioning. In addition, the analysis of competition will be considered both in the contexts of Russian and global markets.

In accordance with the selected features for the competitive analysis, the general landscape of the affiliate map takes the following form (see Figure 24 below).

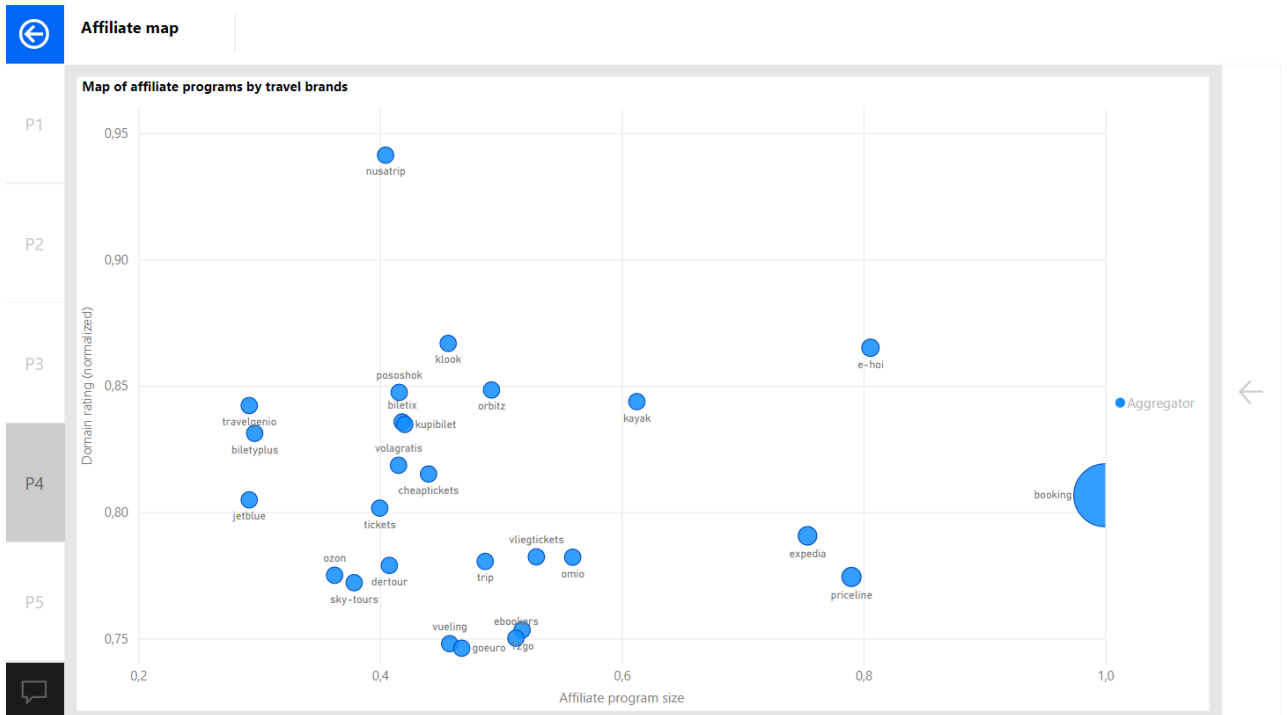


Figure 25. Competitive quadrant №1 - Global sector of flight aggregators

Source: [authors research]

As for the second quadrant, the list of the companies included skyscanner.ru, kiwi.com, tutu.ru, momondo.ru, onetwotrip.com and sletat.ru (see Figure 26 below).

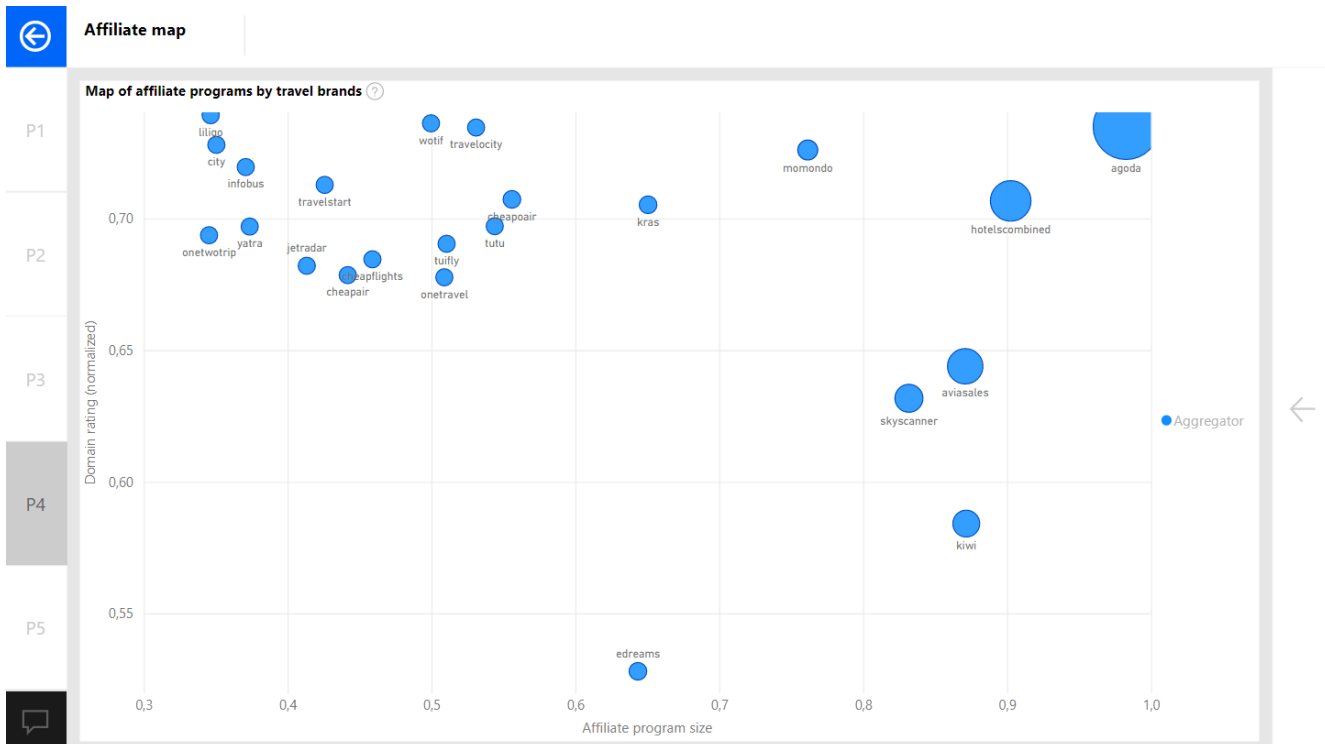


Figure 26. Competitive quadrant №2 - Global sector of flight aggregators

Source: [authors research]

Affiliate program profiles of identified Aviasales competitors were analyzed through several dimensions: number of unique links, number of unique affiliates, average domain rating of affiliates and portfolio development ratio. Table 35 below presents the comparison of Aviasales with the selected main competitors.

Table 35. Characteristics of Aviasales competitors

Advertiser	Unique links		Affiliates		Average domain rating		Portfolio development ratio	
	Global	Russia	Global	Russia	Global	Russia	Global	Russia
aviasales	678 984	568 741	11 871	9 818	9.26	9.67	0.75	0.77
skyscanner	348 249	1 228	6 624	176	8.56	54.68	0.77	2.03
kiwi	307 005	243 935	1 581	119	6.29	0.12	0.68	0.00
momondo	75 084	1 696	1 001	64	15.77	20.90	0.96	1.49
omio	7 662	2 204	923	125	22.68	20.93	1.41	1.50
tutu	5 170	3 635	529	382	13.07	10.74	1.30	1.29
biletix	617	479	85	59	32.08	32.90	2.01	2.07
kupibilet	598	469	74	55	31.88	32.67	2.00	2.06
tickets	562	146	113	50	25.72	23.26	2.02	2.57
ozon	515	351	258	169	21.65	21.79	2.16	2.28
onetwotrip	418	168	246	87	12.78	9.23	2.03	2.17
sletat	26	20	11	9	7.38	6.65	3.00	3.17

Source: [authors research]

As a result of the analysis of competitors' characteristics, the following major conclusions were derived:

- Aviasales has a considerable competitive advantage in terms of total number of unique links and the number of affiliates both in Russia and globally
- The portfolio development ratio indicates that Aviasales program is well balanced, and is close to its direct global competitors - Skyscanner and Momondo, with only slight imbalance towards quantitative factor

- The median affiliate domain rating of Aviasales is one of the lowest among the competitors
- The majority of competitors tend to have portfolio development ratios higher than 1, meaning that on average they are positioned better in their affiliate domain ratings than in the program size.

As a result, the top-5 competitors of Aviasales were identified: Skyscanner, Kiwi, Momondo, Omio and Tutu. They may constitute visible competition for the Aviasales within the affiliate marketing channel today or in future. In further analysis, the listed companies will be directly compared to Aviasales using the affiliate ecosystem framework.

3.3.3. Aviasales affiliate ecosystem

Potential connections of affiliate programs owned by different companies were analyzed. The analysis takes the form of a network, where each node represents a separate company or affiliate, all of which are connected with the lines. The black bubbles represent separate travel brands. The grey bubbles represent separate affiliates, the size of which indicate its domain rating - the smaller the node, the lower the rating. The connections between the affiliates and advertisers are set by lines, where the darker lines signify stronger connections (many links between the advertiser and affiliate).

The exhibit below presents the example of Aviasales affiliate program, which is visualized as an isolated ecosystem (see Figure 27 below).

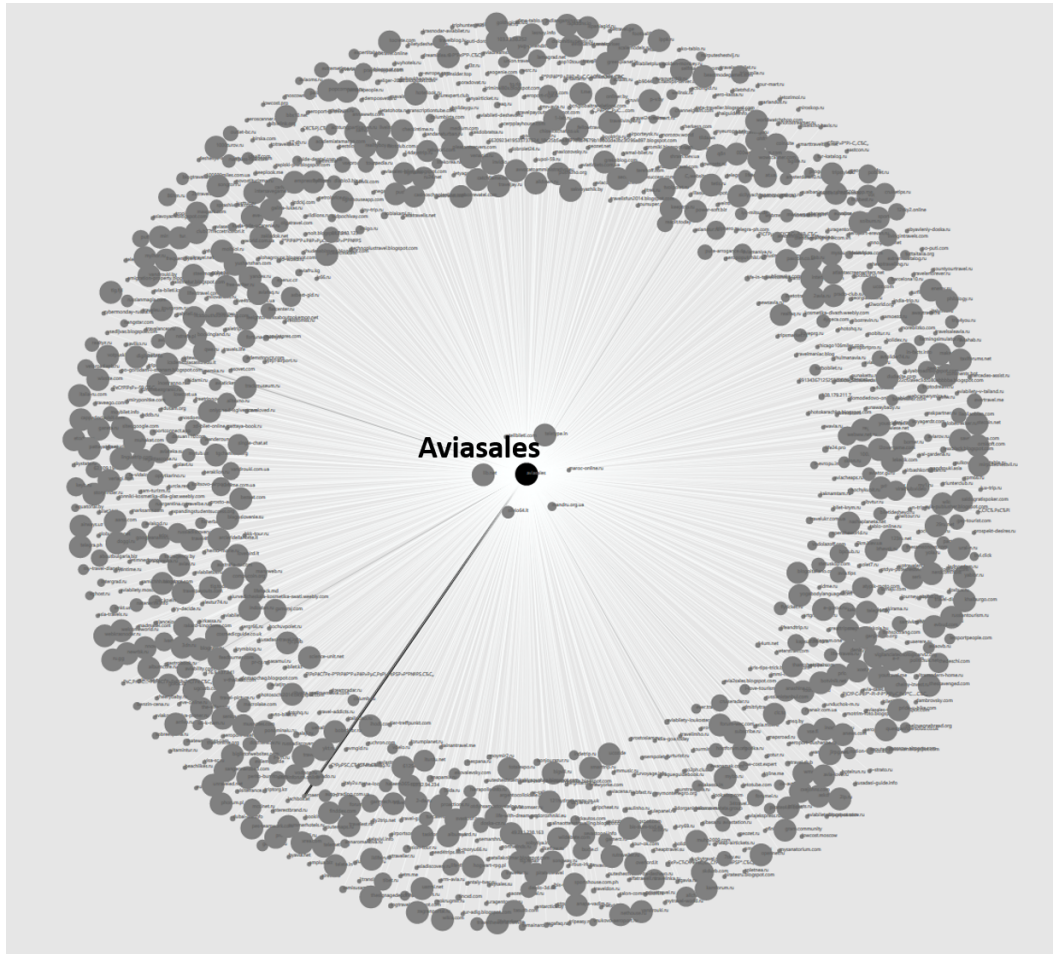


Figure 27. Aviasales affiliate program ecosystem

Source: [authors research]

As seen from the chart, circles representing affiliates are varying in size, which indicates the domain rating heterogeneity of Aviasales affiliate portfolio. In addition, a distinctly strong connection of Aviasales with a certain affiliate is visible - website “proaerportunity.ru” with domain rating 0, on which Aviasales has published over 204k unique affiliate links in total. The rest of the link connections range from low to moderate. Addressing the problem of the high number of affiliates with low ratings, it is possible to analyze the examples of affiliate ecosystems of the leading global travel brands, which tend to have balanced portfolios (see example in Appendix Figure 14).

Based on the analysis of the affiliate ecosystems of Aviasales’ competitors, the certain interconnections between the affiliate programs can be identified. As presented on Figure 28 below, the ecosystem of Aviasales has distinct groups of affiliates that are shared with its direct competitors on the Russian market.

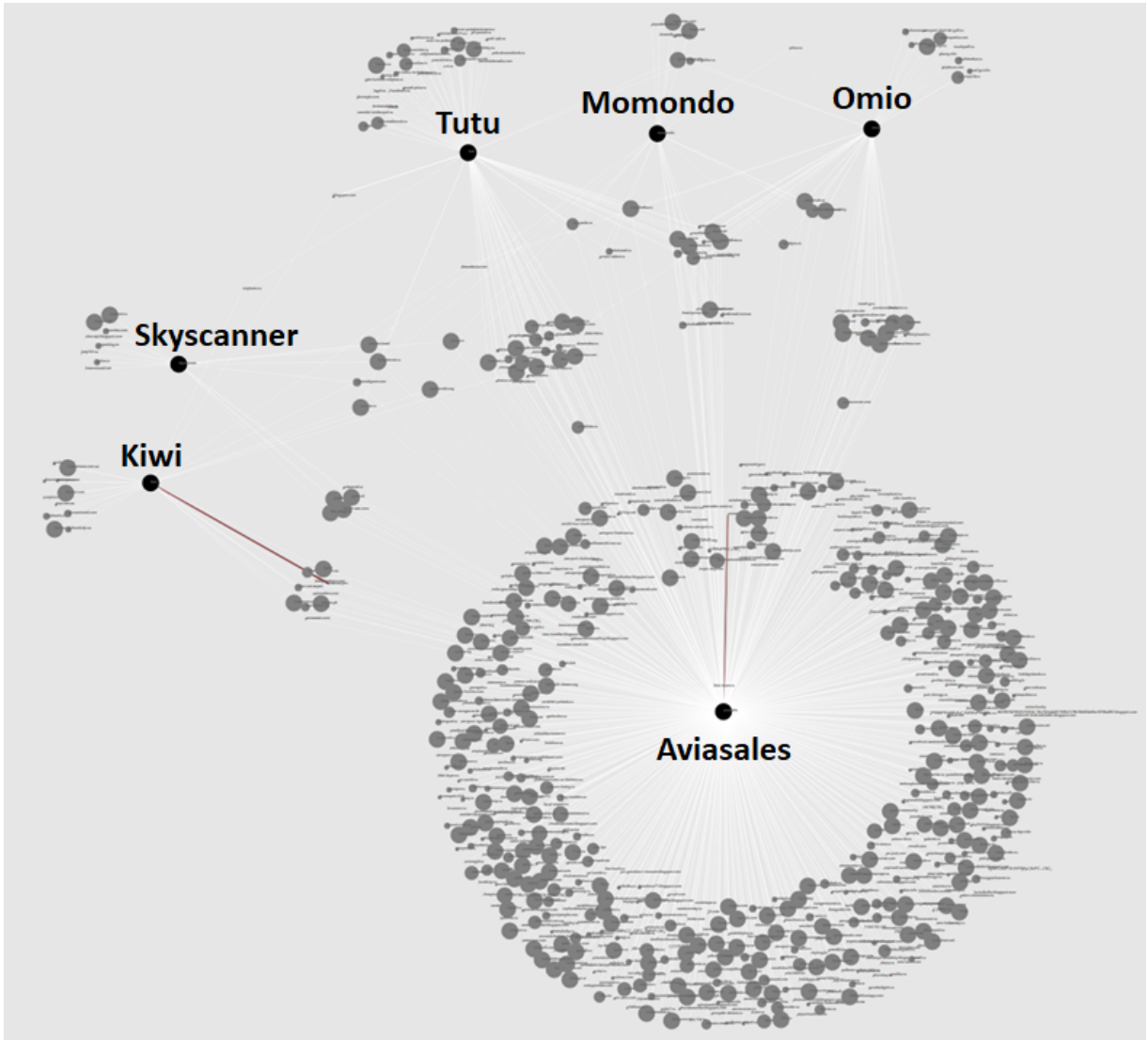


Figure 28. Aviasales affiliate competition ecosystem - Russian market

Source: [authors research]

Similarly, the ecosystem of Aviasales affiliate portfolio is analyzed against its competitors on the global scale, which is presented on Figure 29 below.

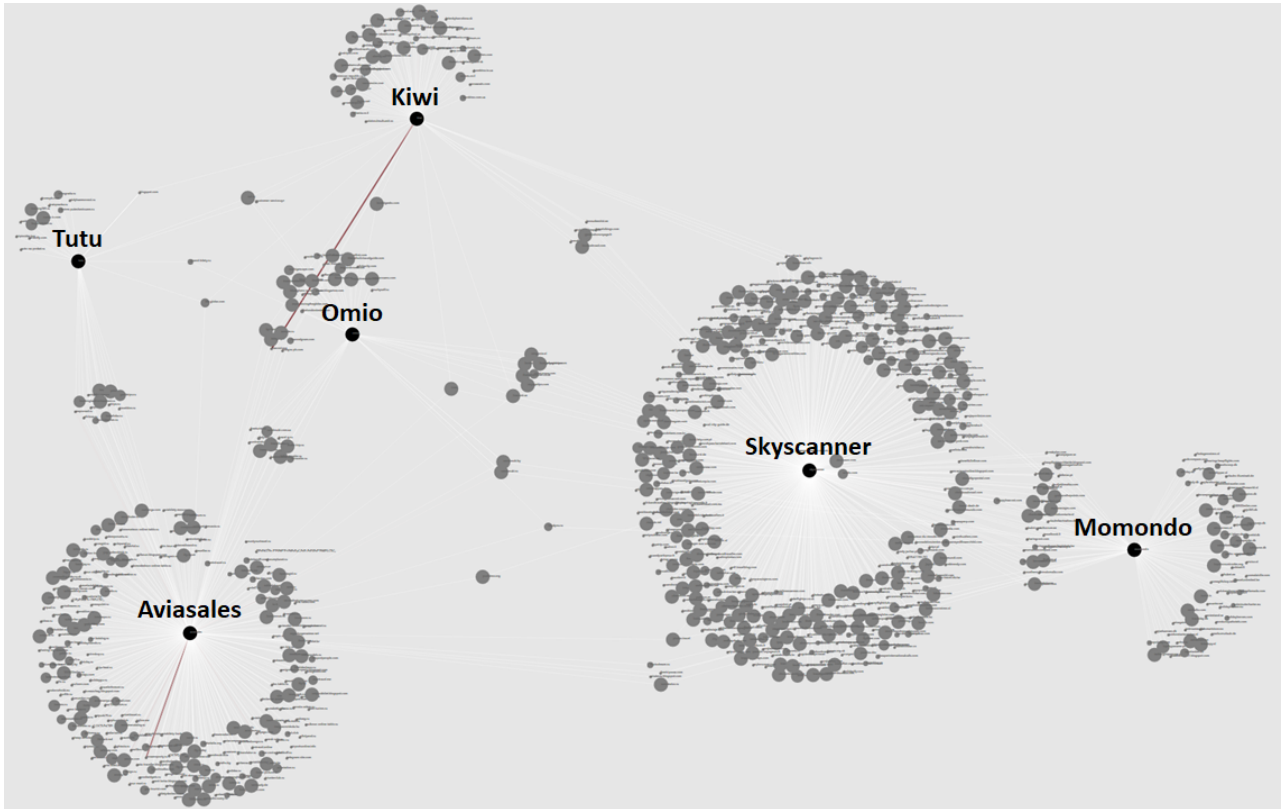


Figure 29. Aviasales affiliate competition ecosystem - Global market

Source: [authors research]

Based on the identified groups of affiliates, which are shared with Aviasales competitors, it is possible to analyze the characteristics of such groups, which are presented in the following table (see Table 36 below).

Table 36. Aviasales shared affiliate groups

Affiliate group	Total affiliates shared with competitor		% of competitor's portfolio		Average domain rating of the group	
	Global	Russia	Global	Russia	Global	Russia
Aviasales - Tutu	249	198	47.07%	51.83%	18.35	19.15
Aviasales - Kiwi	133	69	8.41%	57.98%	26.35	23.65
Aviasales - Omio	119	90	12.89%	72.00%	19.41	19.44
Aviasales - Skyscanner	114	57	1.72%	32.39%	39.97	32.02
Aviasales - Momondo	77	53	7.69%	82.81%	22.87	16.17

Source: [authors research]

The analysis of the shared affiliate groups has led to the following conclusions:

- The competitors of Aviasales tend to have a considerable share of affiliates that already exist in Aviasales portfolio - from 32.39% to 72% of competitor's portfolio on the Russian market and from 1.72% to 47.07% on the global scale.
- The average domain rating of the shared affiliates is relatively high (from 16.17 to 39.97 of average domain rating), while Aviasales' total average indicator barely reaches 9.3.
- Considering the context of analyzed competitors, on the Russian market Aviasales has 92.86% of exclusive affiliates (the only partner - Aviasales), the average rating of which constitute 5.95; on the global market the share of exclusive affiliates increases to 95.79%, as well as their average domain rating of 7.09.

Currently, the analyzed shared affiliate groups do not represent a threat to Aviasales, taking into account that they constitute the minor share of the Aviasales portfolio. However, this may create a potential threat in future, as the listed competitors already do select affiliates with higher domain ratings to build stronger ecosystems (though not so many affiliates in total yet), which may affect the overall performance of Aviasales affiliate program in future.

3.4. Business recommendations

The analysis of Aviasales position within the affiliate channel through EIC framework can be summarized through the following points:

- Domain rating of the Aviasales affiliate program is lower than the median market benchmark. At the same time, Aviasales has one of the largest affiliate programs in industry in terms of both the number of unique affiliates and the number of unique links per affiliate. Due to this issue, Aviasales is currently placed into the Runners-up quadrant with travel brands that have acquired a sufficient network of affiliates, but didn't develop enough quality.
- Aviasales has a well-balanced affiliate portfolio compared to its rivals in Russian and global markets. However, the domain rating of its affiliates is below the market median (Portfolio development index < 1).
- Aviasales shares its highest quality affiliates with the competitors, which may adversely affect the profitability of the affiliate program in the future, if the competitors (such as Skyscanner, Tutu or Kiwi) decide to be more aggressive in development of their own affiliate portfolios.

In order to address those issues, and improve profitability and efficiency of Aviasales affiliate program, the following set of recommendations was prepared (see Table 37 below).

Table 37. Strategic recommendations summary

Strategic recommendation	Description
Implement additional mechanism for affiliates screening	Additional screening procedures should be focused on keeping affiliates with the most relevant types of content (related to travel) and high domain rating (affiliate websites must have good backlink profile, traffic and indexing positions according to Ahrefs methodology). Those affiliates have the highest potential of providing Aviasales with additional customer traffic and conversion. In the future, Aviasales must focus on attracting about 1800 new unique affiliates with average domain rating of 14.5 in order to move into the first (Leaders) quadrant, while keeping its affiliate portfolio well-balanced.
Establish quality KPIs for affiliates and relate it to the commissions they receive	Quality KPIs and metrics for the existing affiliates may be established. Those KPIs can include: average 6-month domain rating, average 6-month organic traffic/unique visitors, average 6-month content updates, etc. Affiliates performing well on those KPIs will receive higher commissions from every sale provided to Aviasales. This initiative can encourage existing affiliates to improve the quality of their websites that will ultimately translate into overall improvement of the Aviasales affiliate portfolio.
Improve features of existing program to attract higher ranking affiliates and create more isolated affiliate ecosystem	Improving features of the existing affiliate program (such as adding better quality widgets, commissions structure and product selection) should provide incentives for high ranking affiliates to build closer partnerships with Aviasales. This will reduce Aviasales affiliate ecosystem connections with the competitors, making partnerships more efficient in bringing traffic and conversion to Aviasales offerings.

Source: [authors research]

Overall, the above recommendations can help Aviasales build more competitive affiliate program and gain leadership positions in travel affiliate marketing both in Russia/CIS and the world.

Conclusion

This research was devoted to visualization of an affiliate marketing channel in the travel industry, and providing managerial recommendations to Aviasales on development of its own affiliate program. Achievement of this goal required three tasks to be completed -

creating single data model for the Aviasales dataset collections; choosing characteristics of affiliate market in travel and visualizing them in a form of BI dashboard; and conducting data analytics work to assess the state of global affiliate marketing and drawing managerial recommendations for Aviasales regarding their own affiliate program (see Table 8 for details).

The dataset collections provided by Aviasales were unstructured and contained multiple problems (see Table 17 for details). In order to structure the data for analysis and manage the data problems, the special project was initiated, designed with the use of CRISP-DM methodology, and completed using Python capabilities. As part of the data mining stage, direct advertisers were identified, and network travel advertisers were retrieved using natural language processing and machine learning techniques. Additionally, methodology for travel vertical classification was established, and all travel advertisers were assigned a Primary or Secondary attribute according to their main business concentration. Finally, all data mining outputs were merged into a unified dataset in order to create an BI dashboard. Dashboard was designed and built in Power BI according to specific requirements. Visualizations and frameworks presented in the dashboard were developed in collaboration with Aviasales team.

Completed BI dashboard provided insights into the structure of affiliate marketing in the travel industry that were not previously studied in other academic works. The research identified the largest travel affiliate programs; the most popular languages, verticals and vertical combinations within the affiliate marketing channel. Moreover, three measures for assessing competitive positioning of affiliate programs were established - the affiliate map/competitive quadrants framework, the portfolio development ratio and affiliate program ecosystem.

Those measures were used to assess the strengths and weaknesses of the Aviasales affiliate program. It was concluded that Aviasales has a well-balanced program with one of the world's largest affiliate portfolios in terms of size. However, Aviasales affiliate program was not qualified for the leading quadrant due to low domain rating of its affiliates. The business recommendations for the Aviasales management team focused on improving affiliates domain rating while preserving the balance of the overall portfolio. Those recommendations included implementing additional mechanisms for the affiliates screening; establishing quality KPIs for the affiliates that are tied to commissions; and improving features of the program to attract high ranking affiliates and isolating Aviasales affiliate ecosystem from competitors.

Overall, the research goal was achieved - all tasks and objectives were completed successfully. Aviasales team approved the final result (see Figure 15 in Appendix).

References

1. Abbas, O.A. (2008). Comparisons Between Data Clustering Algorithms. *International Arab Journal of Information Technology*, 5(3), 320-325.
https://www.researchgate.net/publication/220413756_Comparisons_Between_Data_Clustering_Algorithms
2. Abela, A. (2006). *Charts*. The Extreme Presentation Method.
<https://extremepresentation.com/design/7-charts/>
3. Acceleration Partners. (2017). *The Ultimate Guide to The Affiliate Marketing Model*.
https://www.accelerationpartners.com/wp-content/uploads/2017/05/AP_Ebook_UltimateAffiliateGuide_Final.pdf
4. Ahrefs. (n.d.). *Ahrefs API*. Retrieved April 20, 2021,
<https://ahrefs.com/api/documentation>
5. Ahrefs. (n.d.). *Ahrefs by the numbers*. Retrieved April 20, 2021, from
<https://ahrefs.com/big-data>
6. Ahrefs. (n.d.). *How often is Ahrefs links database updated?* Retrieved April 20, 2021, from <https://help.ahrefs.com/en/articles/78052-how-often-is-ahrefs-links-database-updated>
7. Ahrefs. (n.d.). *What is Domain Rating (DR)?*. Retrieved April 20, 2021,
<https://help.ahrefs.com/en/articles/1409408-what-is-domain-rating-dr>
8. Amancio, D.R., Casanova, D., Comi C.H., et. al. (2019). Clustering algorithms: A comparative approach. *PLoS ONE 14(1): e0210236*.
<https://doi.org/10.1371/journal.pone.0210236>
9. Arora, K., Chakraborty, R., & Elhence, A. (2019). Sparse Victory – A Large Scale Systematic Comparison of count-based and prediction-based vectorizers for text classification. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 188-197.
<https://www.aclweb.org/anthology/R19-1022.pdf>
10. Aviasales. (n.d.). *Affiliate program*. Retrieved April 20, 2021, from
<https://www.aviasales.ru/affiliateprogram>
11. Bray, T. (2002, September 2). *"Deep Linking" in the World Wide Web*. W3.org.
<http://www.textuality.com/tag/DeepLinking.html>
12. Chapman P., Clinton J., Kerber R., et. al. (2000). *CRISP-DM 1.0. Step-by-step data mining guide*. SPSS.
<https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf>
13. Ding, C., Xiaofeng, H. (2004). K-means Clustering via Principal Component Analysis. *Proceedings of the twenty-first international conference on Machine learning*.
<https://doi.org/10.1145/1015330.1015408>

14. Dwivedi, Y., Rana, N. & Alryalat, M. (2017). Affiliate marketing: An overview and analysis of emerging literature. *The Marketing Review*, 17(1), 33-50, doi: 10.1362/146934717X14909733966092
15. Gartner. (February, 2021). *2021 Gartner Magic Quadrant for Analytics and Business Intelligence Platforms*.
<https://info.microsoft.com/ww-Landing-2021-Gartner-MQ-for-Analytics-and-Business-Intelligence-Power-BI.html>
16. Google Developers. (2021, January 13). *K-Means Advantages and Disadvantages*.
<https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages>
17. Hair, J., Black, W., Babin, B., et. al. (2014). *Multivariate data analysis*. Pearson.
18. Indeed. (2021, April 2). *What is a Business Vertical? Definition, Benefits and Examples*.
<https://www.indeed.com/career-advice/career-development/business-vertical>
19. Jai, K. (n.d.). *Tour and Travel - A way out of Corona*. Craft Driven Market Research. Retrieved April 20, 2021, from <https://www.revfine.com/travel-industry/>
20. Jurisova, V. (2013). Affiliate marketing in the context of online marketing. *Review of Applied Socio-Economic Research*, 5(1), 106-110, Retrieved from:
<https://EconPapers.repec.org/RePEc:rse:wpaper:v:5:y:2013:i:1:p:106-111>
21. Mhojhos Research. (2020, May 8). *Affiliate Marketing. Market Size and Total Addressable market*.
<https://mhojhosresearch.com/2020/05/08/affiliate-marketing-is-it-profitable/>
22. NLTK.org. (2019, September 4). *Accessing Text Corpora and Lexical Resources*.
<https://www.nltk.org/book/ch02.html>
23. NLTK.org. (n.d.). *Nltk.stem.package*. Retrieved April 20, 2021, from <https://www.nltk.org/api/nltk.stem.html>
24. Patrick, Z., & Hee, O. C. (2019). Factors Influencing the Intention to Use Affiliate Marketing: A Conceptual Analysis. *International Journal of Academic Research in Business and Social Sciences*, 9(2), 701– 710, doi: 10.6007/IJARBSS/v9-i2/5608
25. Petrossian, G. (2020, April 13). *Business Intelligence Tools: A Pros and Cons Comparison Chart*. CSGPro.com.
<https://www.csgpro.com/blog/business-intelligence-tools-comparison-chart/>
26. Provost, F., & Fawcett, T. (2013). *Data science for business*. O'Reilly
27. Pymorphy2 (n.d.). *Morphological analyzer pymorphy2*. Retrieved April 20, 2021, from <https://pymorphy2.readthedocs.io/en/stable/>
28. Revfine. (n.d.). *What is the Travel Industry?* Retrieved April 20, 2021, from <https://www.revfine.com/travel-industry/>
29. Scikit-learn.org. (n.d.). *Clustering*. Retrieved April 20, 2021, from <https://scikit-learn.org/stable/modules/clustering.html>

30. Scikit-learn.org. (n.d.). *Feature extraction*. Retrieved April 20, 2021, from https://scikit-learn.org/stable/modules/feature_extraction.html
31. Scikit-learn.org. (n.d.). *Principal component analysis (PCA)*. Retrieved April 20, 2021, from <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>
32. Similarweb. (2021, March 1). *Top sites ranking for Travel And Tourism in the world*. <https://www.similarweb.com/top-websites/category/travel-and-tourism/>
33. Travelpayouts Blog. (2017, June 9). *Travel Trends of Today: The Current State of the Travel Affiliate Market*. <https://blog.travelpayouts.com/en/travel-affiliate-market-trends/>

Appendix

```
import re

name = 'rentalcars.com'

get_object_response = s3.get_object(Bucket=AVSLS_BUCKET, Key='backlinks/rentalcars.com.csv')

links = pd.read_csv(get_object_response['Body'], sep=',')

f = []

def find_arguments(row):
    m = re.search("\\/(?:[^\\/]+)\\/?$", row['url_to'])
    if m is not None:
        m = m.group()
        if '?' in m:
            m = re.search("\?(.*?)", m).group(1)
            split = m.split('&')
            for i in split:
                if '=' in i:
                    arg = re.search('(.*?)=', i)
                    f.append(arg.group(1))

links.apply(lambda row: find_arguments(row), axis=1)

args = pd.DataFrame(columns=['arg'], data=f)
m = args['arg'].value_counts()
m = pd.DataFrame(columns=['arg'], data = m)
m.to_csv('arg_' + name)
```

Figure 1. Direct advertisers argument finding function

Source: [authors research]

```
import re
import urllib.parse
import tldextract

def find_url(row):
    for url in row:
        if url is not None and 'destination' in url:
            m = re.search('destination(?:%3A)(.*)', url)
            if m:
                quoted_url = m.group(2)
                unquoted_url = urllib.parse.unquote(quoted_url)
                extracted_domain = tldextract.extract(unquoted_url)
                domain = "{}.{}".format(extracted_domain.domain, extracted_domain.suffix)

                return domain
            else:
                return None
    else:
        return None
```

Figure 2. Advertiser domain extraction from network deep link

** Arguments of "re.search" method vary based on individual affiliate networks*

Source: [authors research]

```

import re

def find_mid(row):
    for url in row:
        if url is not None:
            url = url.lower()
            m = re.search('((\?|&)m=|merchantid=)(\d{1,10})', url)
            if m is not None:
                return m.group(3)
            else:
                return None
        else:
            return None

```

Figure 3. Advertiser domain extraction from advertiser ID - searching advertiser ID

Source: [authors research]

```

df = awin_keys.join(dataset, awin_keys.mid == dataset.adv_id, how='right')

```

Figure 4. Advertiser domain extraction from advertiser ID - joining dataset with ID keys

Source: [authors research]

```

joined = df.join(keys, df.advertiser == keys.advertiser, how='inner')

```

Figure 5. Travel class retrieval from affiliate network links -
joining affiliate network dataset with classified travel domains

Source: [authors research]

```

def find_brand(row):
    for domen in row:
        if domen is not None:
            brand = domen.split('.')[0]
            return brand

```

Figure 6. Brand name retrieval from advertiser domain

Source: [authors research]


```

# конвертируем данные в лист
data = df_ru['text'].values.tolist()

# очищаем от лишних символов
data = [re.sub('\s+', ' ', sent) for sent in data]
data = [re.sub("\'", "", sent) for sent in data]

# препроцессинг
morph = pymorphy2.MorphAnalyzer() # анализатор морфологии
stemmer = SnowballStemmer("russian") # русский стеммер
russian_stopwords = stopwords.words("russian") # русские стопворды
russian_stopwords.extend(['руб', 'год', 'продукт', 'магазин', 'аксессуар', 'товар', 'купить', 'оплата', 'набор', 'акция', 'кредит', 'карта', 'программа'])

# очищаем от пунктуации
def punct(sentences):
    for sentence in sentences:
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True)) # deacc=True убирает пунктуацию

# лемматизация
def lemmatization(data):
    lemmatized = []
    for i in data:
        function_words = {'INTJ', 'PRCL', 'CONJ', 'PREP'}
        i = list(map(lambda word: morph.parse(word)[0], i))
        result = []
        for word in i:
            if word.tag.POS not in function_words:
                result.append(word.normal_form)
        lemmatized.append(result)
    return lemmatized

# убираем стоп-ворды
def stop_words_removal(data):
    data = [[word for word in simple_preprocess(str(doc)) if word not in russian_stopwords] for doc in data]
    return data

```

Figure 7. Words preprocessing for Russian language model

Source: [authors research]

```

# функция для подсчета слов связанных с тревелом
def travel_words(text):
    travel_words = ['билет', 'отель', 'гостиница', 'авиабилет', 'аренда', 'прокат', 'автобус', 'поезд', 'самолет', 'самолёт', 'путевка', 'путёвка', 'travel', 'trip', 'путешествие', 'путешественник', 'бронирование', 'автобусный', 'расписание', 'забронировать', 'железнодорожный', 'турист', 'туроператор', 'туристический', 'туризм', 'виза', 'страна', 'аэропорт', 'курорт', 'шенген']
    num_of_words = []
    for i in text:
        count = 0
        for t in i:
            if t not in travel_words:
                continue
            else:
                count = count + 1
        num_of_words.append(count)
    return num_of_words

```

Figure 8. Russian travel keywords counter function

Source: [authors research]

```

def preprocessing(df):
    lemmatizer = WordNetLemmatizer() # лемматизация
    english_stopwords = stopwords.words("english") # английские стопворды
    tokens = []

    # лемматизируем, токенизируем, обрабатываем стоп ворды
    for i in df:

        lemmatized_words = list(map(lambda word: lemmatizer.lemmatize(word.lower()), i.split()))

        clean_tokens = [word for word in lemmatized_words if not word in english_stopwords]
        clean_tokens = [''.join(c for c in s if c not in string.punctuation) for s in clean_tokens]
        clean_tokens = [x for x in clean_tokens if x != '']

        tokens.append(clean_tokens)

    data_final = []
    for i in tokens:
        i = " ".join(i)
        data_final.append(i)

    return pd.DataFrame(data_final)

```

Figure 9. Words preprocessing for English language model

Source: [authors research]

```

# функция для подсчета слов связанных с тревелом
def travel_words(text):
    travel_words = ['ticket', 'hotel', 'rent', 'rental', 'bus', 'train', 'plane', 'airplane', 'voucher', 'travel', 'trip', 'traveler', 'booking', 'schedule', 'tourist', 'tour', 'tourism', 'visa', 'country', 'airp
    num_of_words = []
    for i in text:
        count = 0
        for t in i.split():
            if t not in travel_words:
                continue
            else:
                count = count + 1
        num_of_words.append(count)
    return num_of_words

```

Figure 10. English travel keywords counter function

Source: [authors research]

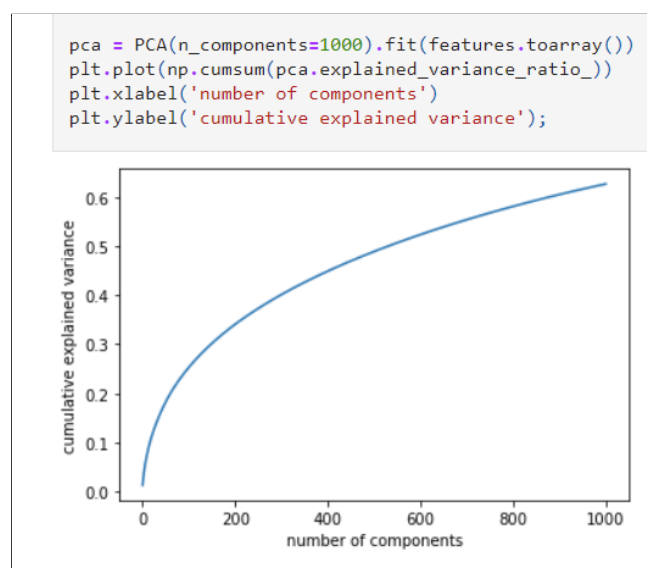


Figure 11. Principal components and cumulative variance explained

Source: [authors research]

Table 1. BI platform functionalities

Functionality	Description
Full-featured free version	Full-functional BI tool for free (not trial)
R/Python supported	Integration with programming languages R/Python
Dynamic cross-filtering	Dynamic use of slicers and filters on the report level
AI-enabled analytics	AI-supported visualizations and data insights
Search analytics with NLP	Ability to ask questions in natural language to receive analytics
Data prep tools	Ability to connect to data sources, transform and clean the data
Data modelling tools	Ability to introduce relationships between the data tables
Database independent	Does not require a database to work
Built-in row level security	Ability to create security roles: different job functions see only related to them data in report
Mixed model types	Ability to work with mixed data connections as livestream direct connection or data import
Third-party data model access	The data model can be extended (just as a data warehouse) using third-party tools or languages using
Commenting & Collaboration	Taking notes in BI reports and on-fly collaboration with other users
Embedded analytics	Ability to embed the analytical report as a white-label to your own portal: website, product, service offerings, etc.
Open-source visualizations	Ability to download new visualizations from open source libraries or own custom visualizations
Native mobile app	Ability to download a branded mobile application on iOS or Android and receive the reports

Source: [authors research]

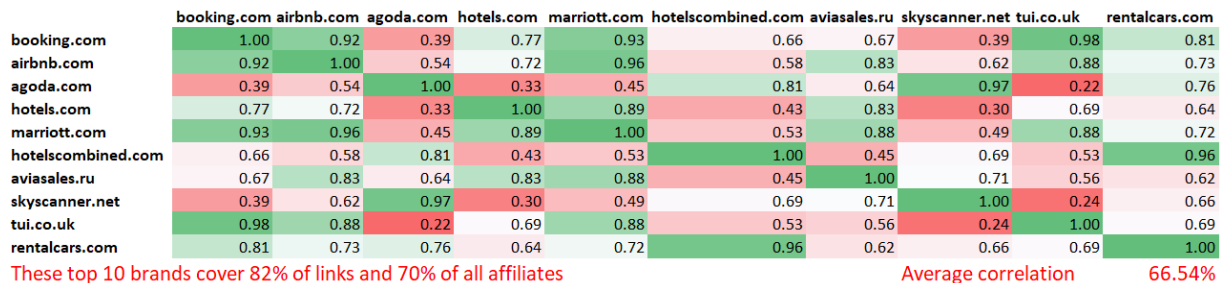


Figure 12. Web traffic change correlation matrix for top-10 travel brands

Source: [authors research]

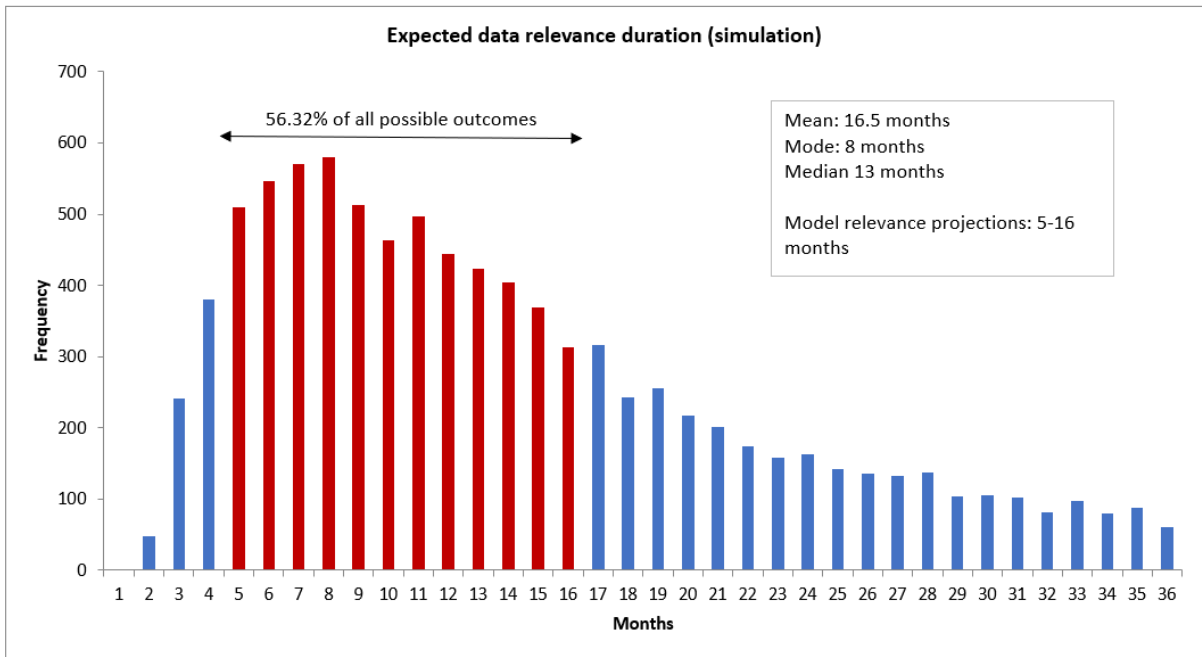


Figure 13. Monte Carlo simulation results

Source: [authors research]

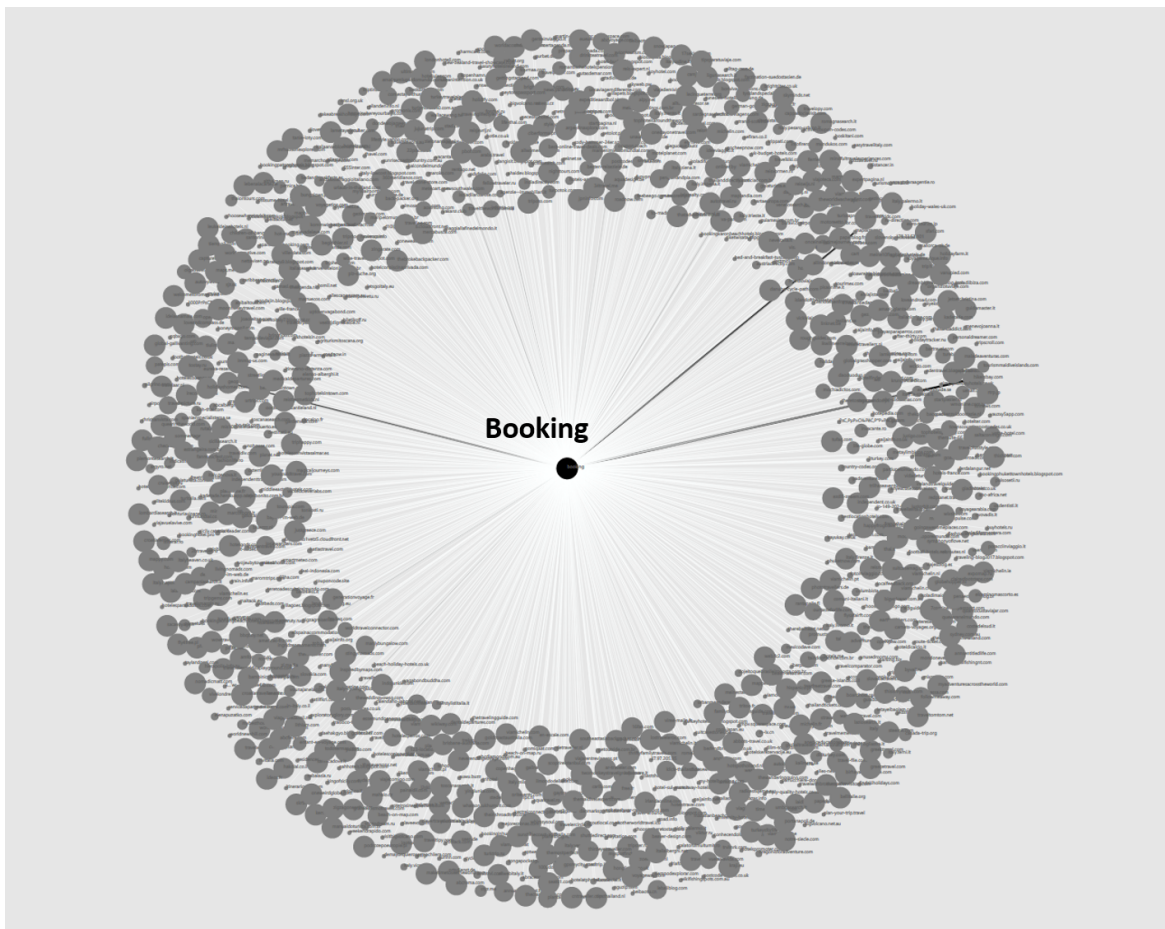


Figure 14. Booking.com affiliate ecosystem

Source: [authors research]

РЕКОМЕНДАТЕЛЬНОЕ ПИСЬМО

Марков Данил Владимирович и Тимербаев Дмитрий Евгеньевич, студенты 2го курса программы магистратуры ВШМ СПбГУ «Бизнес Аналитика и Большие Данные», выполняли проект (исследование) по анализу больших данных для ООО «Авиасейлс Медиа» в рамках дипломной работы.

В рамках проекта студенты показали выдающиеся способности работы с большими данными, применяя разнообразные библиотеки и инструменты аналитики. Это позволило им успешно решить несомненно сложные задачи проекта. Результаты их работы будут использованы для составления большого исследования, публикации о котором появятся во всех основных англоязычных изданиях индустрии путешествий.

Помимо этого, разработанный Данилом и Дмитрием инструмент (дэшборд), является неисчерпаемым источником данных по структуре партнерского маркетинга в сфере путешествий. Отдельно бы хотелось отметить качество визуализации и удобство работы с дэшбордом. Он позволяет выявить стратегические точки роста собственной сети партнерского маркетинга Aviasales и обосновать принятие решение, опираясь на эти точные данные.

Студенты отлично зарекомендовали себя в качестве аналитиков больших данных. Все задачи были выполнены на уровне выше наших ожиданий, сроки всегда соблюдались.

24 мая 2021 г.



ООО «Авиасейлс», ОГРН 1187847259192
197136, г. Санкт-Петербург, ул. Большая Морская, д. 30 литера А, пом. Ч.П. 3-Н, 194, 196-217, 219, этаж 5

Figure 15. Aviasales project completion recommendation letter

Source: [authors research]