

St. Petersburg State University  
Graduate School of Management

Master in Business Analytics and Big Data

Analysis of the Relationship between the Sales of Product Categories  
and their Distribution: case of Procter & Gamble

Master's Thesis by the 2<sup>nd</sup> year students

Daria S. Dobrego

 Daria Dobrego

---

Zamira Magomedova

 Magomedova Zamira

---

Research advisor:

**Elvira V. Strakhovich,**

Associate Professor of the Department of  
Informational Technologies in Management

---


St. Petersburg, 2021


## ЗАЯВЛЕНИЕ О САМОСТОЯТЕЛЬНОМ ХАРАКТЕРЕ ВЫПОЛНЕНИЯ ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Мы, Добрего Дарья Сергеевна и Магомедова Замира, студенты второго курса магистратуры направления «Менеджмент», заявляем, что в нашей магистерской диссертации на тему «Анализ взаимосвязи между продажами категорий товаров и их дистрибьюции», 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 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)».

  
\_\_\_\_\_ Добрего Дарья (Добрего Д.С.)

  
\_\_\_\_\_ Магомедова Замира (Магомедова З.)

## STATEMENT ABOUT THE INDEPENDENT CHARACTER OF THE MASTER THESIS

We, Daria Dobrego and Zamira Magomedova, second year master students, program «Management», state that our master thesis on the topic ‘Analysis of the Relationship between the Sales of Product Categories and their Distribution’, 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)».

  
\_\_\_\_\_ Daria Dobrego (D. Dobrego)

  
\_\_\_\_\_ Magomedova Zamira (Z. Magomedova)

## АННОТАЦИЯ

Автор	Доброго Дарья Сергеевна, Магомедова Замира
Название ВКР	Анализ взаимосвязи между продажами категорий товаров и их дистрибьюции
Образовательная программа	Менеджмент
Направление подготовки	Бизнес-аналитика и большие данные
Год	2021
Научный руководитель	Страхович Эльвира Витаутасовна
Описание цели, задач и основных результатов	<p>Цель проекта – анализ взаимоотношений уровня продаж и дистрибьюции и выявление факторов, влияющих на эти отношения.</p> <p>Задачами проекта являются:</p> <ol style="list-style-type: none"><li>1. Сделать обзор литературы на тему взаимоотношений между продажами и дистрибьюцией;</li><li>2. Определить модели и методы машинного обучения, применяющиеся в ритейл аналитике;</li><li>3. Проанализировать данные на наличие линейной связи между уровнем продаж и дистрибьюцией;</li><li>4. Определить, что влияет на характер отношений между продажами и дистрибьюцией;</li><li>5. Определить наиболее оптимальный метод для анализа связи между уровнем продаж и дистрибьюцией;</li></ol> <p>В результате работы были выявлены факторы, влияющие на характер отношений между продажами и дистрибьюцией, были предложены методы для анализа отношений в будущем, и предложены рекомендации на основе результатов.</p>
Ключевые слова	Дистрибьюция, продажи, анализ взаимоотношений, линейная регрессия, Random Forest, трансформации

## ABSTRACT

Master Students' Name	Dobrego Daria, Magomedova Zamira
Master Thesis Title	Analysis of the Relationship between the Sales of Product Categories and their Distribution
Educational Program	Management
Main field of study	Master in Business Analytics and Big Data
Year	2021
Academic Advisor's Name	Elvira V. Strakhovich
Description of the goal, tasks and main results	<p>The goal of the project is to analyze the relationship between sales and distribution and determine which factors affect this relationship.</p> <p>Objectives of the project are:</p> <ol style="list-style-type: none"> <li>1. Conduct literature analysis on the topic of relationship between sales and distribution;</li> <li>2. Determine what machine learning methods and models are used in retail analytics;</li> <li>3. Analyze the data to determine whether there is a linear relationship between sales and distribution;</li> <li>4. Determine factors affecting the nature of the relationship between sales and distribution;</li> <li>5. Determine optimal methodology for conducting analysis of the relationship between sales and distribution.</li> </ol> <p>As a result, we have identified the factors that affect the relationship between sales and distribution, proposed methodology to use for this analysis in the future, and made recommendations based on our findings.</p>
Keywords	Distribution, sales value, product categories, regression, relationship analysis, Linear Regression, Random Forest, transformations

<b>INTRODUCTION</b>	<b>6</b>
<b>1. LITERATURE REVIEW</b>	<b>9</b>
1. CONCEPT OF DISTRIBUTION	9
2. CATEGORY MANAGEMENT	13
3. INTERDEPENDENCE OF SALES AND DISTRIBUTION	15
4. MACHINE LEARNING IN RETAIL	16
5. SUMMARY OF CHAPTER 1	27
<b>2. EMPIRICAL STUDY</b>	<b>28</b>
1. PROJECT DESCRIPTION	28
2. ANALYSIS	30
3. SUMMARY OF CHAPTER 2	62
<b>CONCLUSION</b>	<b>64</b>
<b>REFERENCES</b>	<b>67</b>
<b>APPENDICES</b>	<b>70</b>

## INTRODUCTION

Determining the effects and their sizes of various marketing variables on sales has been among the most important activities in academic marketing research. There is a lot of evidence of the effects that advertising and pricing have, but the same cannot be said for the distribution.<sup>1</sup> Hanssens et. al. say that “distribution is one of the most potent marketing contributors to sales and market share” and add that “its elasticity can be substantially greater than one”<sup>2</sup>, but empirical studies on distribution are few and far between.

The distribution of products is considered one of the key drivers of the sales growth. Several studies have shown a convex relationship between distribution and sales in cross-sectional data.<sup>3</sup> Generally, when considering the interdependent relationship of sales and distribution, it seems obvious and logical that there is one. The more of the product available for purchase, the more it is purchased, and the more it is purchased and in demand, the broader the distribution, but the theoretical and empirical evidence of such a relationship is limited.

The distribution landscape is changing very fast, and several factors are impacting its evolution. There are the more obvious one, like e-commerce and new generations that are making up the majority of the economically active population. There are also the black swans that come out of nowhere and drastically change every industry, bringing about the new normal. That is why companies need to be able to analyze different elements of their marketing mix, and see how each one affects their sales volume, so that they can adjust it to bring maximum possible profit.

Analysis of the relevant literature on the subject of distribution helped us identify a major research gap. Mainly, it is the type of the relationship that distribution and sales have. Logically, wider distribution causes higher sales volume, but there are other extraneous factors that must be accounted for, such as prices, promotions, or even product characteristics, so everything is not as easy as it seems at first. Also, categories of the products play a role, as the literature indicates that the actual strength of the relationship between sales and distribution is moderated by the product type and category, as well as the growth stage of the category.<sup>4</sup>

In this research, the main focus will be on the relationship that distribution and sales of category products have. There is some research that shows the quantitative effect that distribution

---

<sup>1</sup> Bucklin, Randolph & Siddarth, S. & Silva-Risso, Jorge. (2008). Distribution Intensity and New Car Choice. *Journal of Marketing Research - J MARKET RES-CHICAGO*. 45. 473-486. 10.1509/jmkr.45.4.473.

<sup>2</sup> Hanssens, Dominique M., Leonard J. Parsons, and Randall L. Schultz (2001), Market Response Models: Econometric and Time Series Analysis, 2<sup>nd</sup> edition, Kluwer Academic Publishers, p. 347

<sup>3</sup> Wilbur, K. C. and Farris, P. W. (2014). Distribution and market share. *Journal of Retailing*, 90(2):154{167.

<sup>4</sup> PERFORMANCE IMPACT OF DISTRIBUTION EXPANSION: A REVIEW AND RESEARCH AGENDA (Hibbard, et al)

has on sales in general, but only few that focus on specific product categories, and that is why our focus will be on one specific category.

The goal of our project is to analyze the type of relationship that sales and distribution have, whether it is linear or non-linear, so that the company can better analyze the effects that changing distribution can have on sales. To reach this goal, we have formulated questions that we will need to answer, which are the following:

- Is there an interdependent relationship between sales and distribution?
- What type of relationship is there between sales and distribution, linear or non-linear?
- What differentiates the products with linear relationship between sales and distribution from those that demonstrate non-linear relationship?
- What is the most suitable model to analyze this relationship?

First of all, we start with theoretical part. In the first chapter, we do a literature review on distribution, category management, and the relationship between sales and distribution. We then discuss machine learning in retail, its tasks, issues, and limitations. Later, we discuss the cases of machine learning application in business, focusing on machine learning in retail, specifically its usage in Magnit, Perekrestok, Pyaterochka, and M-Video, and the results that using machine learning has yielded those companies. Afterwards, we move on to machine learning methods, and go deeper on the methods that we have chosen for our project.

In the second chapter, we present the empirical part of our project. We start with describing the company for whom this analysis is conducted, Procter & Gamble. Then describe the business task, and set more detailed goals and objectives based on this business tasks. We also create a step-by-step analysis with brief description of each step. We start our analysis with exploratory data analysis, where we familiarize ourselves with data, clean it of outliers and drop missing data, do some basic analyses, such as correlation analysis. We also calculate several new variables that will aid us in our analysis further. Then we look at some products separately to understand whether there is a linear relationship between sales and distribution, with the help of scatter plots. Afterwards, we describe our challenges, which are computational power of our machines, data, especially lack of data on other elements of marketing mix, and different time periods in which products are sold, as some of them have been selling for 5 years, and they have more data, while some of them have only been on the market for three months, for example. Then, we preprocess data, and move on to applying different machine learning models, namely, linear regression and random forest. We also apply different types of transformations in linear regression to see which ones would yield highest results. After this, we analyze the results of both models, and try to formulate what differentiates products with linear relationship between sales and distribution from those without.

As an expected result, we intend to find out when products demonstrate linear relationship between sales and distribution, and when they do not do that, and what product characteristics differentiate those two groups from one another.

There were two people working on this project, and responsibilities were divided in this way:

- Daria Dobrego was responsible for literature analysis of category management, exploratory data analysis, baseline linear regression and analysis of the results, and building random forest and analyzing and interpreting results of this model;
- Zamira Magomedova was responsible for literature analysis of distribution and machine learning in retail, exploratory data analysis, linear regression with transformations and analysis and interpretation of the results, and writing the text of the thesis.



## **1. LITERATURE REVIEW**

### **1. Concept of distribution**

Distribution of goods is an important activity in the marketing strategy of a company, as it moves goods and services from the manufacturer to the ultimate consumer. Some authors put distribution in the 'place' element of the marketing mix (e.g., Wearne and Morrison, 1998), while some authors consider it the same as the 'place' element (e.g., Doyle and Stern, 2006). Distribution helps bring the goods and services at the time and place where they are needed to who needs them. Distribution is a set of activities undertaken by a company so that the products are made available and accessible to the target consumer. Distribution encompasses a system of all activities that are related to the transfer of economic goods between manufacturers and consumers.<sup>5</sup> Distribution is fundamental to a company, as choosing the right distribution strategy plays significant role in company's sales, because a product is of no use if it is not sold to the right consumer who will appreciate its value. And the best way to reach this right consumer is by choosing a right distribution strategy and channel.

Distribution of a company consists of two major elements: distribution channels and physical distribution. We will discuss those components in following sections.

#### **1. Physical distribution**

Physical distribution includes all the functions of movement and handling of goods, particularly transportation services (trucking, freight rail, air freight, marine shipping, and pipelines), transshipment and warehousing services (e.g. consignment, storage, inventory management), trade, wholesale and, in principle, retail.<sup>6</sup> Physical distribution is concerned with physical handling of goods and providing customer service in terms of delivery. Its aim is to deliver right products to the right place in the right condition.

The elements of a physical distribution channel are: (1) customer service, (2) order processing, (3) inventory control, (4) warehousing, (5) transportation mode, (6) materials handling.<sup>7</sup>

When talking about customer service, physical distribution affects (1) the length of the processing and delivery of the order, (2) the percent of 'out-of-stock' orders, (3) the condition with

---

<sup>5</sup> Segetlija, Zdenko et al. "Importance of Distribution Channels - Marketing Channels - for National Economy." (2011).

<sup>6</sup> Kim R. Fowler, Chapter 16 - Logistics, Distribution, and Support, Editor(s): Kim R. Fowler, Craig L. Silver, Developing and Managing Embedded Systems and Products, Newnes, 2015, Pages 649-672, <https://doi.org/10.1016/B978-0-12-405879-8.00016-7>.

<sup>7</sup> <https://www.yourarticlelibrary.com/marketing/distribution-channels/top-6-elements-of-physical-distribution-channels-with-diagram/48313>

which the ordered goods are delivered to the customer, (4) how willing is the manufacturer to replace the defected goods.<sup>8</sup>

## **2. Distribution channels**

Distribution channels, which are sometimes called marketing channels, are sets of interdependent organizations involved in the process of making a product or service available for use or consumption.<sup>9</sup> Increasing the number of ways a consumer can find goods can increase the sales, but it can also make the distribution management too complex.

Distribution channels can be divided into two major categories: direct and indirect channels. Direct distribution channel, which Kotler calls 'zero-level' channel, means that there are no middlemen, and all marketing functions are realized by the manufacturer themselves, and they pass their goods and services directly to the consumer. As for the indirect distribution channel, it comprises at least one level, aptly called 'one-level' channel, between a manufacturer and an ultimate consumer, meaning that the goods and services go through someone else, for example a wholesaler or a dealer, before they reach the target consumer. This distinction is a bit different when it comes to Procter & Gamble, for whom direct distribution means selling to stores such as Magnit, or Perekrestok, and indirect distribution means first selling to a distributor, who then sells products on their own terms.

The choice of a channel depends on the company's marketing strategy and how it is positioning itself. For example, a manufacturer of luxury goods would be better suited to choose an exclusive channel of distribution, while an FMCG manufacturer would gain more from an intensive distribution channel. Also, the chosen distribution model must add value to the consumer. Some products require a salesperson, some require a trial, and others require no hassle at all, so it all must be accounted for when choosing a distribution model.

## **3. Measuring distribution**

Distribution can be measured in terms of numeric and weighted distribution. Numeric, or physical, distribution simply shows the percentage of the stores that sell the item in a selected location. Weighted distribution shows the percentage of the stores that sell the product weighted by the sales of the product category in the store.

There is also an All-Commodity-Value or -Volume, or % ACV, which shows store's total sales of all products relative to the sales of all relevant retailers in a given territory. It helps a

---

<sup>8</sup> T. N. Duening, R. D. Hisrich, M. A. Lechter, Chapter 11 - Going to Market and the Marketing Plan, Editor(s): Thomas N. Duening, Robert D. Hisrich, Michael A. Lechter, Technology Entrepreneurship, Academic Press, 2010, Pages 351-386, ISBN 9780123745026, <https://doi.org/10.1016/B978-0-12-374502-6.00011-0>.

<sup>9</sup> Kotler P. & Armstrong G. (2006). Principles of marketing, (11th Ed.) Upper Saddle River: New Jersey: Prentice-Hall.

company decide which stores to prioritize, as it shows how well is the store at getting the products off their shelves compared to all the stores in a selected area.

Another measure is a Product class value, PCV, which shows how well a specific product category is sold in a store compared to all the other stores in a given territory. It shows where customers go to buy a specific category, and can help a company choose a good location to sell specific category products.

Using this measure helps companies see where to direct its distribution efforts. For example, when the numeric distribution is high, but weighted distribution is low, the company can understand that while its products are sold in a lot of stores, they are not being sold in the right stores that matter most to the business. And vice versa, when weighted distribution is high, but the numeric is low, it shows that while the company's products are not present in a lot of stores, they are present in the ones that have a high impact on its business and sales.

#### **4. Distribution in FMCG companies**

Distribution is vital to the success of the FMCG companies, and delivery of the goods in the right amount, the right place and the right time, and in excellent condition is an integral part of the sales process. Distribution in FMCG is affected by following features of the FMCG industry: short product life cycle, which creates a high turnover of goods, and spontaneous purchases made unthinkingly by consumers. That is why it is highly important for FMCG companies to find the most efficient strategy of distribution.

FMCG companies have three main objectives when it comes to distribution<sup>10</sup>:

- availability of the brand: a consumer cannot buy the product if it is not on the shelf, hence the products of the right brands must be present at the right place in the right time, and made available to the right consumer;
- quality of the product: companies must ensure that the product presented to the consumer is of the highest available quality in terms of freshness, quality, and packaging;
- effectiveness: making the product available to the right consumer in the most effective and efficient way possible plays a major role in the quality of distribution.

FMCG companies operate in quite a volatile industry, so every business there strives to retain their customers by investing in innovative and better techniques to provide a flawless experience to the end consumer<sup>11</sup>. Technology and digitization are playing a major role in how FMCG companies are innovating their operations. Their distribution operations are moving online,

---

<sup>10</sup> Olariu, Ioana. (2009). FMCG companies specific distribution channels. Studies and Scientific Researches - Economic Edition. 10.29358/scoco.v0i14.48.

<sup>11</sup> <https://www.fieldassist.in/blog/fmcg-distribution-network/>

which helps companies improve productivity, ensure uninterrupted delivery and respond faster to market and consumer demand. The continuous rise of e-commerce is also challenging the companies to come up with solutions that will respond to the current trend of going digital.

## **5. Future of distribution**

There is a number of new technologies that either already have or will affect the distribution models. With the development of Internet, new models of distribution have emerged. For example, e-commerce is widely used as an emerging distribution channel. E-commerce has revolutionized and reshaped business relationships and has caused dramatic shifts in channel power as information and communication imbalances disappear.<sup>12</sup>

Another example is the use of RFID tags, Radio frequency identification, which help keep track of products, when they are being moved between different company locations (a wide range of tags is available, with different areas of coverage and temperature tolerance<sup>13</sup>), or shipped to distributor or end-consumer. RFID tags can help synchronize supply chain, as they allow to track the product throughout the distribution system. A company can collect RFID data from many sources and know the exact location of products and their number, so that they can adjust their production. That same technology can help make warehousing automated, as there would be no need for inventory because all of the data would be available in real-time.

Yet another example of how technology is changing distribution, is the usage of drones and driverless trucks to deliver orders. These technologies are still being tested and are not widely used, but they have the potential to change the landscape of distribution even more.

And of course, the pandemic is teaching a few lessons to the distribution. For instance, it has shown that omni-channel distribution strategy is highly important to boost sales, and research has shown that distributors with strong e-commerce platforms or a digital presence have fared better in the crisis.<sup>14</sup>

Overall, companies that embrace digitization and use transaction data across products, brands, and customer segments, are significantly more resilient than others. Companies can achieve higher value if they align their offering with customer needs and willingness to pay, and address their pain points in time.<sup>15</sup>

---

<sup>12</sup> Muhamad, Jantan & Ndubisi, Nelson & Yean, Ong. (2003). Viability of e-commerce as an alternative distribution channel. *Logistics Information Management*. 16. 427-439.

<sup>13</sup> <https://www.corerfid.com/rfid-applications/rfid-in-distribution/>

<sup>14</sup> <https://www.mckinsey.com/industries/advanced-electronics/our-insights/covid-19-crisis-how-distributors-can-emerge-stronger-than-before>

<sup>15</sup> Ibid.

## 2. Category management

### 1. The concept of category management

When analyzing possible options of distribution of various products in stores, first thing that comes to mind is defining how different SKUs (stock keeping units) are distributed within the store. The way the products are placed in a store always has a specific logic behind it, thus, there has arisen an autonomous discipline called category management. The main goal of category management is to make a move from product-centered approach to customer-centric one. The strategy of product placement should always consider customer demand and split numerous SKUs into subgroups or classes, which are called categories. These categories can be very diverse in their size, from small to large ones.

In such approach each category is treated separately, almost as a business unit, and managers' aim is to maximize customer demand within each of them.

ECR (Efficient Consumer Response – an organization in Europe) has provided the most accepted definition of category management—a retailer/supplier process of managing categories as strategic business units, producing enhanced business results by focusing on delivering consumer value<sup>16</sup>.

Category management includes developing the best possible customer proposition, with competitive pricing, attractive promotions, relevant assortments, and appealing visual merchandising, planogramming.

However, it is not always as easy as it sounds, - some categories might have hundreds of SKUs, and it would be tough to define what product distribution, placement or price optimization, will deliver maximum demand and the best possible profit.

Category management began as an approach to restructure the purchasing organizations of retailers<sup>17 18</sup>. At some time retailers are in need of taking care of hundreds of categories, and retail managers cannot give equal attention and resources to all of them. Thus, the fundamental idea of category management is that manufacturers and retailers collaborate to improve each category to meet customer needs.

There are two key accents in the category management definition: product/service category as seen by customers is a central point of business activities, and it should be a joint process of retailers and suppliers. This approach to category management is seen as marketing category

---

<sup>16</sup> ECR Europe (2014) Glossary. Retrieved 25 Nov 2014, from <http://www.ecr-europe.org/toolbox/glossary>

<sup>17</sup> ACNielsen, Karolefski J, Heller A (2006) Consumer-centric category management: how to increase profits by managing categories based on consumer needs. John Wiley & Sons, Inc, Hoboken

<sup>18</sup> D. Dujak, Z. Segetlija, J. Mesarić, Efficient Demand Management in Retailing Through Category Management

management, and sometimes it is also called micro-marketing, shelf management, space management, schematic development or fact-based selling.<sup>19</sup>

## **2. Category management in retail**

There are two ways in the supply chain that can be outlined and treated as special activities of category management in retail:

to customers/consumers — retail management fuels demand for the category by increasing customer traffic in the store (the number of customers who visit the store), and/or increasing the probability of purchasing in the category for the customers who are already in the store

to suppliers/manufacturers — the management strives to improve the supply and logistics by providing information on demand and other information required to ensure optimization of the chain and requesting additional work and information about the category from supplier.

## **3. Category management in regard to distribution**

It is hard to name any other industry where the amounts of data daily aggregated is as large as in retail, especially FMCG. Unfortunately, this also means that there should be a very specific logic behind variables that are taken into account when doing analysis. Doubtlessly, category here plays a major role.

In category management, it's customers who decide which SKU will be included in a certain category by their purchase behavior. Retail corporations usually put effort to observe such behavior and make relevant conclusions and implement them then in the category structure. All products that satisfy the consumer's need in more or less equal matter or those that meet their secondary need linked to the primary one (e.g. coffee, tea, cookies, sugar) constitute a specific category.<sup>20</sup>

Category management is straightly referring to customer experience but, unfortunately, most retailers lack customer insights to make proper relevant decisions.

First off, retail analysts should always keep track of the data that they collect and use wide array of data inputs to drive unique insights. It is important to cover all categories in-depth, for example, have records by brand, pack size, sub segment, store cluster, and so on. The final analysis should perform a plan for each category that defines the steps, required investments, and expected financial or operational benefits.

Time-series variables are also needed to include a better understanding of past product performance and consumer needs than currently exist for most retailers.

---

<sup>19</sup> Ursin C (2004) Facing facts: what category management—or fact-based selling—can do for you. In: Beverage Dynamics. Bev-AI Communications, Inc. Retrieved 10 Apr 2007, from <http://www.beveragenet.net/bd/2004/0406/0406cm.asp>

<sup>20</sup> Oliver Wyman, Making category management work, 2012

Secondly, managers should take a closer look at loyalty, social media or other sources of data to better understand how their core customers shop their stores now and identify growth points in decreasing segments and figure out where new customers can be acquired by satisfying unmet needs. For example, a retailer could define that a main client segment makes purchases primarily among four categories. Using category management and client-centric approach, the retailer is capable of identifying categories in which the needs of that customer were not unmet and fixing the gap. Retailers such as Target, Kroger and Walmart have seen impressive results from revamping category management, including a 2% to 4% increase in sales, a 2% to 3% increase in margin and a 10% to 15% increase in inventory productivity.<sup>21</sup>

Finally, according to the data analyses category managers make decisions on the share and power of each individual SKU and therefore propose strategies to replace, maintain, reduce or increase SKU in a category or the whole category's distribution in the store.

Overall, category as a factor can play critical role in store's sales. As an example, some products can be popular at the specific district and can drive no demand in the other. Collaboration between retailers and suppliers is dedicated to continuous work to ensure timely supply of retail stores of those products that are needed in the quickest and cheapest way.

### **3. Interdependence of sales and distribution**

According to some studies, distribution and price are two of the marketing mix elements that have the greatest impact on brand sales.<sup>22</sup> Other studies have shown that the link between retail distribution and sales is statistically important, and it's been found that the total (short- and long-term) elasticity of sales with respect to distribution is much higher than that of advertising (0.74 to 0.13).<sup>23</sup> Several studies have shown a convex relation between sales and retail distribution.<sup>24</sup> However, despite the obvious importance of distribution for sales, there is little evidence of the significance and magnitude of the effect that the former has on the latter. And it is obvious that the effect runs both ways: the product is bought because it is available for purchase, and it is available for purchase because the seller expects it to sell well. There, of course, might be other factors affecting this relationship, such as discounts or promotions, but we will not be focusing on those factors.

---

<sup>21</sup> Kurt Salmon, Seven Facets of Modern Category Management

<sup>22</sup> Marketing mix and brand sales in global markets: Examining the contingent role of country-market characteristics (S. Cem BAHADIR, et al)

<sup>23</sup> Ataman, M. B., Van Heerde, H. J., and Mela, C. F. (2010). The long-term effect of marketing strategy on brand sales. *Journal of Marketing Research*, 47(5):866-882.

<sup>24</sup> Wilbur and Farris (2014)

Distribution alone doesn't affect sales, as brand and product itself play a major role. Making more of a product available won't automatically sell more, as the quality and reputation of the brand are still very important to the consumer.

Still, Reibstein and Farris (1995) find a positive relationship between distribution intensity and market share in consumer-packaged goods, and Bucklin et al. (2008) expand on this study and find a positive relationship between distribution intensity and market share in consumer durables, such as cars. Pancras et al. (2012) show the positive effect of an increasing distribution intensity (in terms of numeric distribution) on overall sales.

Friberg and Sanctuary (2017) have found that sales are convex in turnover weighted retail distribution: the wider the distribution, the greater the percentage change in sales volume from further expansion. They have also found that as product's distribution widens, it competes with successively fewer products, and this product takes a larger share of a smaller pie.

Intuitively, we can say that distribution is one of the most potent marketing contributors to sales and market share<sup>25</sup>, but the empirical evidence of such an assumption is very scant, as most studies focus on other elements of the marketing mix and their effect on sales and market share.

We will be focusing on the relationship between distribution, mainly the weighted distribution, and category products' sales, as the literature indicates that the actual strength of the relationship is moderated by the product type and category, as well as the growth stage of the category.<sup>26</sup>

#### **4. Machine learning in retail**

Retail industry is the big data industry by definition. According to estimates, Walmart alone collects around 2.5 petabytes of data about transactions, behavior of customers, locations, etc. every hour.<sup>27</sup> Data analysis offers insights that can help increase customer engagement, retainment, and loyalty, but with the amount of data generated today, it can be hard to accurately analyze large data without sufficient means. That is where machine learning comes in.

In Russia, companies such as Wildberries and Perekrestok use machine learning algorithms for recommendations, and for customer analysis, such as frequency and value of purchases, lifestyle, favourite product categories, and so on.<sup>28</sup> Companies all over the world integrate machine learning into their operations to increase efficiency, increase customer loyalty, analyze customers, and generate new insights and ideas based on them.

---

<sup>25</sup> Hanssens, Dominique M., Leonard J. Parsons, and Randall L. Schultz (2001), Market Response Models: Econometric and Time Series Analysis, 2<sup>nd</sup> edition, Kluwer Academic Publishers.

<sup>26</sup> PERFORMANCE IMPACT OF DISTRIBUTION EXPANSION: A REVIEW AND RESEARCH AGENDA (Hibbard, et al)

<sup>27</sup> McAfee, Andrew, Erik Brynjolfsson, Thomas H. Davenport, D.J. Patil and Dominic Barton (2012), "Big Data: The Management Revolution," *Harvard Business Review*, 90 (10), 61–7.

<sup>28</sup> <https://mcs.mail.ru/blog/kak-machine-learning-povyshaet-prodazhi>



Machine learning is an application of artificial intelligence which focuses on learning from data and improving its accuracy without being explicitly programmed to do so.<sup>29</sup> In machine learning, algorithms are trained to find patterns and features in large amounts of data so as to make decisions and predictions based on new data.<sup>30</sup> The better the algorithm, the more accurate will be the decisions and predictions, as it processes more and more data.

Typically, a machine learning application consists of four basic steps:

1. selecting and preparing a training data set: data should be preprocessed before being used in an application;
2. choosing an algorithm: algorithm choice depends on whether or not data is labeled, and with labeled data regression and decision trees are used, while with unlabeled data – clustering algorithms and neural networks;
3. training the algorithm on training data set and evaluating it: it's an iterative process where different hyperparameters (a parameter whose value is used to control the learning process) are adjusted to increase accuracy of the model;
4. using the model: finally, the final model can be used on new data, and improved if necessary.<sup>31</sup>

In this chapter, we are going to discuss what the applications of machine learning algorithms, and how it is applied in retail industry specifically. We are also going to talk about different machine learning methods, and which methods we have chosen for this project.

## **1. Main machine learning tasks**

All machine learning methods can be categorized into four primary categories:

- Supervised learning: machine learning algorithm is trained on a labeled dataset; it requires less data than other categories, and makes training easier because the results can be compared to actual data;
- Unsupervised learning: machine learning algorithm is trained on a huge unlabeled dataset, and the algorithm has to extract meaningful connections and features itself; it's more suitable for identifying features and trends that human eye might miss;
- Semi-supervised learning: a small amount of data is labeled in this case, and machine learning algorithm finds features from these data, and then classifies remaining data accordingly; it is the most common type of method used;

---

<sup>29</sup> <https://www.ibm.com/cloud/learn/machine-learning>

<sup>30</sup> Ibid.

<sup>31</sup> Ibid.

- Reinforcement learning: a behavioral machine learning method where algorithm is learning as it goes, by trial and error; a sequence of successful outcomes will be reinforced to develop best predictions or recommendations.

Essential machine learning techniques are regression, classification, clustering, decision trees, and neural networks. Regression, decision trees, and classification belong to supervised learning category, while clustering and neural networks belong to unsupervised learning category.

Regression is used to predict a specific numerical value, or explain the effect that an independent variable has on a dependent variable. Classification algorithms are used to predict a class value. Clustering algorithms are used to group data points according to similar characteristic. Decision trees can be used both for regression and classification, and they split data according to some features and conditions. They can be used to predict whether the buyer will become a repeat customer, and they can also be used to forecast the value of purchases. Finally, neural networks mimic the structure of the brain, where artificial neurons connect with other neurons, and create a complex structure. Neural networks are also known as a deep learning algorithm.

Even though machine learning is applied in different spheres, from healthcare, where it's used for diagnostics, to banking, where it can be used for fraud detection or making decisions on loan risks, all machine learning tasks belong to one of the following categories:

- Classification task: diagnostics in healthcare, computer vision, repeat customer identification, etc.;
- Regression task: sales forecast, apartment price, shares price, weather, quality assurance, etc.;
- Clustering: customer segmentation, deciding whether a cosmic body is a star or a plane, etc.;
- Dimensionality reduction: compressing the data while preserving the structure and statistical properties as much as possible;
- Anomaly detection (identifying outliers in datasets): a bit different from classification, because an anomaly is not much present in datasets, so training a machine learning model to detect anomalies using a training dataset is very hard, and an example of this would be fraud detection.

## **2. Machine learning issues and limitations**

Machine learning is rapidly gaining in popularity, and it is applied in our day-to-day lives, even if we don't notice it. But it is not perfect, and we have to consider its issues and limitations when applying machine learning algorithms to solve our tasks.

First of all, machine learning algorithms require a massive amount of data to train and learn, in order for the prediction to be as accurate as possible. The data must be of good quality, and sometimes data preprocessing takes more time than model training itself.

Then, there is the fact that machine learning algorithms require a lot of time and resources to learn and develop, and it might take much more computational power than one could have expected.

There is also a problem of overfitting and underfitting. The former happens when the model overlearns, and its predictions are perfect for the training data, while of lower quality for the test, and especially new, data. Underfitting happens when the model is not complex enough, so it cannot capture relations between dataset's features and dependent variables. There is a fine balance between building a model that is complex enough to make accurate predictions, but not too complex so that it doesn't become too attuned to the training data.

Another limitation of machine learning is the approximation of results. No matter how good a model is, it doesn't produce a 100% accurate result, as there is always room for errors and biases.

Finally, it can be hard to interpret models, what they actually do and learn on the inside. It can be a problem when one is trying to explain how machine learning model achieved its tasks, but as long as results are accurate enough, not many people really consider it to be a big problem.

Beside technical issues and limitations, there are more abstract questions too, like the question of ethics, or responsibility and accountability of models' developers and users. But those are questions that are not actually relevant in this project, as we are not working with sensitive data, so we will not be discussing them here.

### **3. Business cases of ML/Regression implication**

In this section we will discuss how machine learning is being applied in business, focusing on specific companies and scenarios where they use machine learning algorithms. We will be talking about applications of machine learning in retail.

#### **1. Sales and demand forecasting**

Machine learning algorithm for sales and demand forecasting finds and analyzes patterns and relationships between different products, analyzes historical data on sales, analyzes external factors such as competitors and the market situation, and it all results in a sales and demand forecast.

The model can analyze how discount for one product would affect the sales of other products of similar characteristics, or how different weather conditions would analyze the sales of specific product categories, for example, hot weather's effect on the sales of lemonade. It allows retailers to optimize their stocks and prepare for the upswing in demand, or vice versa.

An example of this application of machine learning would be X5 Retail Group, who implemented machine learning for demand forecasting in Perekrestok. The model analyzes around 200 factors affecting demand, such as price elasticity, advertising, sales of other products, competitor's activity, etc. The model analyzes receipts to identify whether the product is available on shelves, and to have a full history of sales, which increases the accuracy of the model. The model updates the forecast every single day, which increases the availability of the product in demand on shelves, and allows for flexibility of the whole supply chain.<sup>32</sup>

Another case of machine learning for demand forecasting would be Magnit. In Magnit, they apply neural networks to analyze demand and optimize the proposition of the products. This resulted in the 5% increase of the accuracy of the forecast, which in turn is expected to increase the revenue by 4 bln rubles a year by decreasing the product deficit by 2%.<sup>33</sup>

## **2. Marketing and advertising optimization**

Machine learning algorithms allow to optimize advertising spending and increase revenue from marketing activities by stopping advertising activities that are not efficient, and focusing on those that bring the most value. For example, a machine learning model can be used for finding the best fit of place-product-promotion, to increase the efficiency of such an activity.

In Pyaterochka, they have used a reinforcement learning model for customized promotions and discounts. The model analyzed previous promotions and offered the most relevant one for customers based on this historical data. All customers, who were all members of a loyalty program, were divided into 27 groups based on their purchase history. The model analyzed how each group reacted to a promotion, how it affected the sales volume, and whether this promotion was lucrative at all. As a result, the project team identified the most profitable promotion recipes, and the most efficient communications for Pyaterochka.<sup>34</sup>

Machine learning algorithm is much better at segmenting customers than simple marketing analysis, as it allows for deeper analysis and pattern identification. Machine learning allows to group customers into segments by identifying hidden, not obvious patterns, for example, young fathers who buy diapers with beers. This segmentation can also be applied to products themselves, identifying products that are usually bought together.

M.Video uses machine learning to segment customers based on their values, like ambitious people who only want the best, or family-oriented people, and promote different products based on this segmentation.<sup>35</sup>

## **3. Recommender systems**

---

<sup>32</sup> [https://www.cnews.ru/news/line/2020-06-08\\_v\\_seti\\_perekrestok\\_vnedrena](https://www.cnews.ru/news/line/2020-06-08_v_seti_perekrestok_vnedrena)

<sup>33</sup> <https://mcs.mail.ru/blog/kak-machine-learning-povyshaet-prodazhi>

<sup>34</sup> <https://www.retail.ru/news/x5-primenit-mashinnoe-obuchenie-v-marketinge/>

<sup>35</sup> <https://www.retail.ru/articles/kak-m-video-ispolzuet-dannye-pokupatelya-dlya-promo-meropriyatiy/>

Recommender systems are a class of machine learning algorithms that offer relevant suggestions to users. Recommender systems can be used anywhere: Netflix uses them to offer what to watch based on previous history and the history of users with similar tastes; Amazon uses them to offer products that were bought by users with similar characteristics; YouTube and Vk use them to suggest videos based on user's tastes and tastes of their friends and others with similar tastes. Recommender systems are also used to personalize offerings to customers, especially via email marketing.

Perekrestok uses recommender systems to group products together and propose them to customers based on data about their behavior and relevancy of the products at that very moment. Perekrestok tried to divide their products into as specific categories as possible, for example, lactose free milk was also divided by country of origin, so that the offer could be as customized as possible. As a result, the user chooses a product, and then sees other similar products, which were offered not only based on user behavior, but also product attributes. The model then can evaluate the choice of the user, and then adjust its recommendation based on results of evaluation.<sup>36</sup>

Another popular example of recommender system application is Amazon, who among the first started saving customer history and using it to make customized offerings. Today, Amazon has integrated recommendations into nearly every step of the purchasing process, thus increasing its sales. Even more, 35% of what people buy on Amazon comes from recommendation algorithms implemented in their processes.<sup>37</sup> They offer personal recommendations, based on what user has previously bought, recommendations of products usually bought together, and products that a user has recently viewed.

#### **4. Literature review on methods**

In this section, we will discuss machine learning methods that are used in retail analytics. Considering that our objective is to analyze the effect that one variable (distribution) has on another (sales), it is more sensible to focus on methods that help do that. The methods that are used to analyze relationships between variables are all a part of category of regression methods, so here we will talk about regression methods in machine learning, and discuss their basics, advantages and disadvantages.

Wang et al<sup>38</sup>. compared four machine learning algorithms used for forecast: ordinary least squares, support vector machine, regression trees, and bagged trees. OLS is a linear, parametric model used for testing relationships between different variables, while the remaining three models

---

<sup>36</sup> <https://vc.ru/trade/143587-keysy-personalizacii-kak-internet-magazin-perekrestok-ispolzuet-tovarnye-rekomendacii-dlya-rosta-onlayn-prodazh>

<sup>37</sup> <https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers>

<sup>38</sup> Xin (Shane) Wang, Jun Hyun (Joseph) Ryoo, Neil Bendle, Praveen K. Kopalle, The role of machine learning analytics and metrics in retailing research, *Journal of Retailing*, 2020

are non-linear and non-parametric models. They found that OLS performed the worst out of four, with R-squared of 0.76, and SVM, regression trees, and bagged trees had R-squared of 0.9, 0.91, and 0.94, respectively.

### 1. Ordinary least squares

Ordinary least squares is a method in linear regression, used to estimate parameters in the model, which aims at minimizing the sum of squared errors between observed data and predicted data. To use the OLS method, the following formula is applied:

$$m = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$
$$b = \bar{y} - m * \bar{x},$$

where m is the slope, b is the line intercept, x and y are independent and dependent variables respectively, and  $\bar{x}$  and  $\bar{y}$  are the average of independent and dependent variables respectively.

OLS is a relatively easy method to estimate a relationship between variables, which doesn't require too much computational power. Also, according to Gauss-Markov theorem, OLS can provide best linear unbiased estimators. Some disadvantages include sensitivity to outliers, sensitivity to too many variables, and a number of conditions to satisfy, like multicollinearity and normal distribution.

### 2. Support vector machine

Support vector machine is a popular machine learning algorithm used for classification and regression. Support vector regression uses the same principle as SVM, as it will find an appropriate line or a hyperplane to fit the data.

It differs from OLS in that it aims at minimizing the coefficients, not the squared errors. In SVR, the error term,  $\epsilon$  (epsilon), can be tuned to increase the accuracy of the model. Training the SVR model means solving following equation:

$$\text{minimize } \frac{1}{2} \|w\|^2,$$

with constraints  $|y_i - w_i x_i| \leq \epsilon,$

where w is the coefficient,  $y_i$  and  $x_i$  are dependent and independent variable respectively, and  $\epsilon$  is the error term. Another parameter that can be tuned is the tolerance of values that fall outside the acceptable error threshold.

SVR is a powerful algorithm, whose advantage is that it gives flexibility to define how much error is acceptable in a model, and also define the tolerance of the values that are beyond the error margin.

### **3. Regression trees**

Regression tree is a type of a decision tree, used when a decision has a continuous target variable, e.g. price. Basic regression tree divides a data set into smaller subgroups, and then fits a simple constant for each observation in the subgroup. Regression trees consist of roots, which are a starting point of any tree, leaves, and nodes. It is easier to interpret the results of decision trees, compared to OLS, but the accuracy is lost. That is why a common method for decision trees would be bagging (bootstrap aggregating) trees, or random forests, to increase accuracy of the model.

The decision tree model begins with the entire dataset, and searches distinct value of each input variable to find the predictor and further split data into two regions, so that their overall sum of squares is minimized. The splitting is then repeated on each region, and so on, until some criterion for stopping is reached.

Decision trees might become too complex, which might lead to great accuracy, but also leads to overfitting the data, and performing poorly on new data. The advantages of regression trees include interpretability, understanding of which variables are important for the prediction, and fast processes. But overall, single decision trees do not provide high accuracy, so it's better to use them in ensembles.

### **4. Bagged trees**

Bootstrap aggregation, or bagging, is a procedure for reducing the variance of a statistical learning method. The algorithm constructs  $N$  regression trees using  $N$  bootstrapped training sets, and average the result of all trees. The trees go as deep as they can go, no pruning is done, and that is why each tree has high variance, but averaging the results allows to reduce variance, and simultaneously have low bias.

Bagging results in higher accuracy of a model, as it trains a lot of trees, but the major drawback of bagging is that the result is difficult to interpret. It means that it is no longer clear which variables are the most important for the model.

So, while bagging helps improve accuracy of the model, it does so at the expense of interpretability.

### **5. Chosen methods for the project**

The goal of this work is mainly to analyze the relationship between product distribution and sales, and the best method for this kind of task would be regression analysis, as it allows to measure the impact that different independent factors, such as temperature or competitors' promotions, might have on a dependent variable, such as sales. Regression analysis is a go-to method in companies that want to explain a phenomenon (e.g. sales drop last month), forecast future (e.g. what will sales be like in the next three months), or choose an action plan (e.g. which

promotion is supposed to bring more profits).<sup>39</sup> Based on this, and the literature review, we have chosen linear regression and regression trees random forest for our project. In this section, we will look deeper into both methods.

### 1. Linear regression

Linear regression attempts to model the relationship between two variables by way of fitting a linear equation to the data. One variable is considered a dependent variable, while the other is an explanatory variable. Before fitting a linear model to observed data, it is important to first determine whether there is any relationship between dependent and independent variables. Tools that can be used for this step include scatterplots and correlation analysis, as they both help determine the strength of a relationship between the variables of interest.

A linear regression line takes the form of a following equation:

$$Y = a + b * X,$$

where Y is the dependent variable, X is the explanatory variable, a is the intercept (when x is equal to 0), and b is the slope of the regression line.

To employ linear regression, several assumptions must be satisfied:

- linear relationship: there exists a linear relationship between dependent variable, y, and independent variable, x;
- independence: the residuals are independent;
- homoscedasticity: the residuals have constant variance at every level of x;
- normality: the residuals of the model are normally distributed.

If one or more of these assumptions is not met, then the results of the linear regression model will not be wholly reliable.

There are several techniques for estimating unknown parameters in linear regression model:

- least squares estimations techniques, which include ordinary, generalized, and weighted least squares;
- maximum likelihood estimations;
- Bayesian linear regression, etc.

Least squares is a method for approximately determining the unknown parameters in a linear regression model. It does so while minimizing the errors, which are the difference between predicted and actual data points. Figure 1 shows the formula for ordinary least squares, where the squared value of errors is minimized.

---

<sup>39</sup> <https://hbr.org/2015/11/a-refresher-on-regression-analysis>



$$\hat{\alpha} = \min_{\alpha} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 = \min_{\alpha} \sum_{i=1}^n \varepsilon_i^2$$

$$\hat{\beta} = \min_{\beta} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 = \min_{\beta} \sum_{i=1}^n \varepsilon_i^2$$

**Fig. 1** Ordinary least squares method

Source: [Towards Data Science, 2019]

OLS is the most popular optimization method in linear regression, since the outputs of the regression are unbiased estimators of the real values of alpha and beta. Furthermore, according to the Gauss-Markov theorem, the OLS estimators are the Best Linear Unbiased Estimators of the real values of alpha and beta, if the assumptions of the linear regression model are met.

We will also be applying transformations to variables in linear regression. Transformations in linear regressions are used to:

- linearize regression model;
- stabilize variance, by reducing heterogeneity of variance;
- normalize variables.

Our main purpose for using non-linear transformations is to linearize the relationship between sales and distribution.

There are several ways to carry out transformation: by transforming only the dependent variable, only independent variable, or by transforming both dependent and independent variables. The main drawback of applying transformation is that the results of the model have to be interpreted with this transformation in mind, as there have to be made more calculations to understand how exactly changes in the independent variable affect the dependent variable.

There are several types of transformations, and the most popular ones are log transformation, square root transformation, and reciprocal transformation, all known as non-linear transformations.

For log transformation, we take a logarithm of a dependent or independent variable, and when interpreting the coefficients, we have to calculate the exponential of the coefficient, subtract one from it, and multiply by 100 if only the dependent variable was transformed, divide the coefficient by 100 in case where only predictor variable was transformed, or simply interpret the

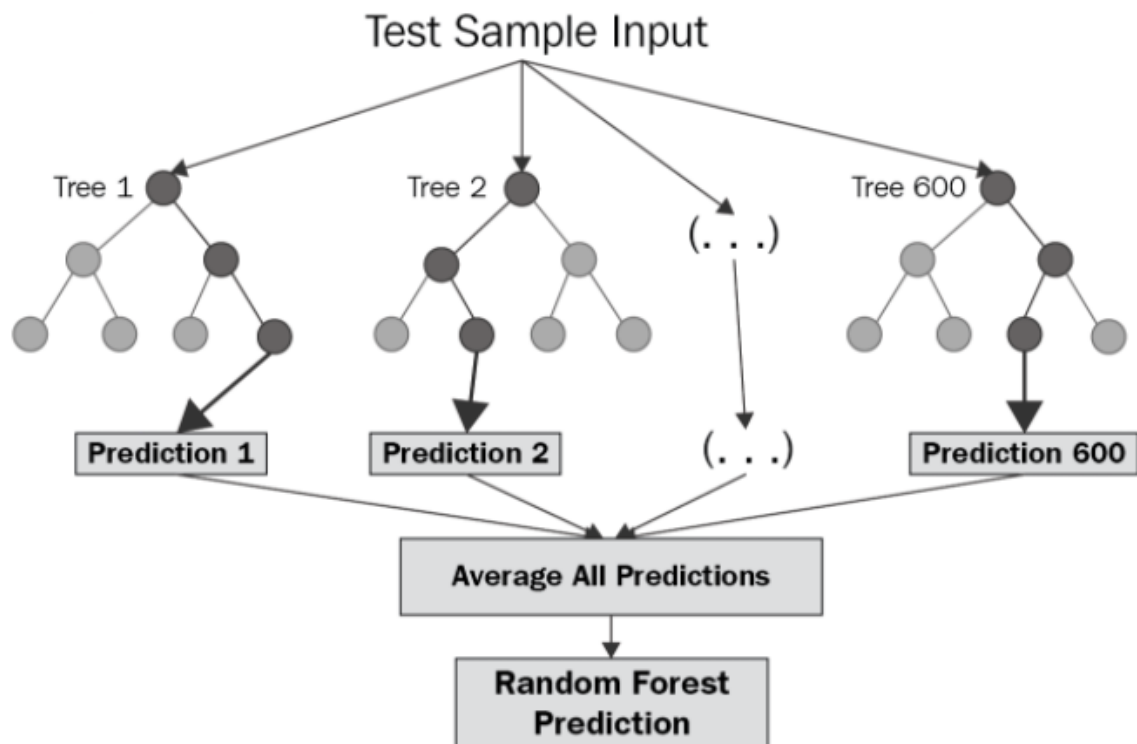
coefficient as the percent increase in the dependent variable if both independent and dependent variables were transformed.

For square root transformation, we take square roots of dependent or independent variables, or both. For reciprocal transformation, we will take  $(1/\text{dependent variable})$ ,  $(1/\text{independent variable})$ , or both.

Linear regression model helps us solve two issues. On one hand, it helps us identify products which demonstrate linear relationship between sales and distribution. On the other hand, it gives us estimates of how, exactly, distribution affects sales, in numerical terms, as in, how sales value changes when distribution is increased or decreased by 1%.

## 2. Random forest

Random forest regression is an extension of regression tree model. Random forest is a tree-based algorithm that uses qualities and features of multiple regression trees for making decisions. It solves the problem of overfitting, which is very common when using single regression trees. What random forest does is it merges the output of multiple regression trees to generate the final result. It is also applicable in case of non-linear data, which is the reason we have chosen this method. Figure 2 shows the structure of a random forest.



**Fig. 2** Structure of a random forest

Source: [Level Up Coding, 2020]

The trees in a random forest are all trained in parallel to each other and independently, on different sub-samples.

Random forests have its advantages and disadvantages. As we have stated earlier, random forest provides more accurate results, and escapes overfitting, a problem from which single decision trees usually suffer. As for its disadvantages, random forest requires a lot of computational power, and they can be too slow for real-time predictions. Also, random forests can be harder to interpret and explain than other models, but this challenge can be overcome.

### **Summary of Chapter 1**

We have analyzed academic literature on three topics: distribution, category management, the relationship between distribution and sales volume, and machine learning. We discussed the concept of distribution, two elements of distribution (physical distribution and the distribution channels), common ways to measure distribution, specifics of distribution in FMCG companies and the future of distribution. As for the category management, we discussed its concept, category management in retail and the category management data usage in retail analytics.

As for the relationship between sales and distribution, we have found that there is not a lot of empirical studies confirming the widespread belief that distribution does indeed drive sales, so we are aiming to aid in filling this gap with our thesis, with focus on products' category.

Finally, we presented machine learning, discussed the main tasks that machine learning is aiming to solve, talked about some common issues and limitations of applying machine learning. We then discussed the cases of machine learning applications in retail, with examples both from Russian companies, and some foreign companies. We also conducted a short literature review on common algorithms and models of machine learning, focused specifically on models that are applicable in retail, and our cases, and discussed limitations of those models and algorithms.

Based on this analysis, we have set following research questions:

- Is there an interdependent relationship between sales and distribution?
- What is the model of relationship between product value share and its weighted distribution?
- Does this model of relationship stay the same for all FMCG products, or does it change based on factors such as product category, or brand?

And the aim of this study is, therefore, to find out whether there is indeed a statistically significant interdependence relationship between category products sales and their distribution.

## **2. EMPIRICAL STUDY**

In this chapter, we will talk about the company for whom this consulting project is done, describe the business task of this project, describe available data, and clarify goal and objectives that we aim to achieve with this work. Then, we will move on to analysis, starting with exploratory data analysis, data preprocessing, and then the models and algorithms that we have chosen for this project, and the process of application of those models. Afterwards, we will discuss our findings, and offer suggestions based on obtained results.

### **1. Project description**

Before we move on to the analysis, it is important to talk about the project: who it is done for, what is the business task of this project, what are the goals and objectives. There is both a technical need, and a business need, and the analysis conducted in this paper aims to satisfy both. We will start by discussing the company, and the industry where it operates, then we will talk about business task, goal and objectives of the project, and further, we will describe data that we have, and finally, we will move on to the analysis.

#### **1. Procter & Gamble**

This project is done in cooperation with Procter & Gamble, an American multinational company that operates in packaged consumer goods industry, and is one of the leaders in said industry. The company not only produces and realizes consumer packaged goods, but a lot of its activities are concerned with retail analytics, ranging from analysis of marketing activities to supply chain analytics, all with one aim: improving the customer experience. In the age of Big Data, there are a lot of opportunities for companies such as Procter and Gamble to analyze their clients and gain insights, and Procter & Gamble does exactly that, with the help of cutting-edge technologies and data science methods, staying proactive and innovative in its solutions.

Procter & Gamble are leaders in the FMCG industry all over the world. The company managed to achieve this status not just with high product quality and innovations, but also with an excellent marketing mix, especially the place component of said mix. Procter & Gamble products are available almost all over the world, and they are insanely easy to purchase: it is quite hard to find a supermarket that doesn't sell products under one of Procter & Gamble's brands.

Procter & Gamble is highly effective in its use of its wide distribution system all over the world, a system which includes manufacturing firms, distributors, and retailers.<sup>40</sup> That is why the products of Procter & Gamble are distributed widely and intensively, and they are sold in all kinds of stores, ranging from discount stores and mom-and-pops to hypermarkets. The nature of the products also helps, seeing as they are fast moving consumer goods, which are replaced frequently.

---

<sup>40</sup> Iacobucci, D 2013, MM4, Mason, Ohio: South-Western: Andover: Cengage Learning [distribution], [2013], p. 127

Of course, Procter & Gamble's success cannot be attributed only to its intensive distribution network, because the company has specialized in consumer experience since the start<sup>41</sup>, and is always striving to attract more clients with promotions and campaigns, and retain loyalty. Due to the combination of all those factors, and Procter & Gamble has succeeded in becoming the world-renowned company that it is today.

## **2. Business task description**

As was stated in the previous section, Procter & Gamble has a very effective distribution system. It is also widely accepted that extensive and intensive distribution drive up the sales, but there is a very scant amount of research to clarify exactly how distribution and sales are connected. It seems only logical that the wider the distribution – the higher the sales. But there is a consensus in the industry that distribution is one of the main drivers of the sales, and what little research is available, it proves this consensus (refer to section 3 of the 1st chapter).

For a company that directs a lot of resources to maintaining a wide and extensive distribution network, it is highly important to know how exactly that wide network drives sales, and whether it drives them at all.

FMCG is not really famous for consumer loyalty to specific brands, so if one brand of some product in the store is not available, consumer will simply buy another brand, not leave the store in search of this specific brand. There are a lot of options from which to choose for a consumer, and that is why, at least in FMCG, distribution drives sales, because if the product is not available, it is simply not bought, but its substitute is. This is true for both same categories of the products, and for the inner characteristics of the products, such as size of the package, or the taste, if applicable.

So, if the product is unique in its category, of course the relationship between distribution and sales would be linear: simply selling the product in the stores would lead to increase in its sales. But as we are dealing with consumer packaged goods, there are a lot of choices of products within the same category, and just as much within the same product forms. This substitutability is what increases the complexity of the problem. It is assumed that some consumer packaged goods do not have a linear relationship between sales and distribution. But if not linear, then what kind of relationship do they have? And what are the characteristics that determine which products do or do not have this linearity in the sales-distribution relationship? Or maybe this relationship depends on whether or not the product is new in its category?

To conclude, what this project aims to do is find out the nature of the relationship between sales and distribution. We will apply linear regression model to determine the products whose sales

---

<sup>41</sup><https://cloud.google.com/blog/products/data-analytics/how-procter-gamble-improves-consumer-experiences-with-data>

and distribution demonstrate linear relationship, determine which factors distinguish those products from those that do not demonstrate linearity, and then we will try to determine what kind of a relationship the latter have between their sales and distribution.

### **3. Goal and objectives**

Based on the needs of the company, the goal of this project is to analyze the relationship between sales and distribution of a specific category, and if possible, extrapolate this relationship on other categories, as well. With this goal in mind, the following objectives for the project were set:

1. To preprocess the data and use machine learning algorithms to analyze the relationship between sales and distribution of a specific category, and to provide quantitative effect of distribution on sales;
2. To analyze what differentiates the products with linear relationship between sales and distribution from those that do not demonstrate such a relationship: it might be a new product (whose distribution was zero at the beginning), or it might be some characteristics of the product, such as its segment, or its size;
3. To analyze the products that exhibit non-linear relationship between sales and distribution with machine learning algorithms designed for non-linear relationships, and to provide quantitative effects that distribution has on sales, if possible.

To achieve this goal, the main instrument that will be used is Jupyter Notebook application, an interactive shell where different programming languages, such as Python, or R, can be applied.

## **2. Analysis**

### **1. Analysis plan**

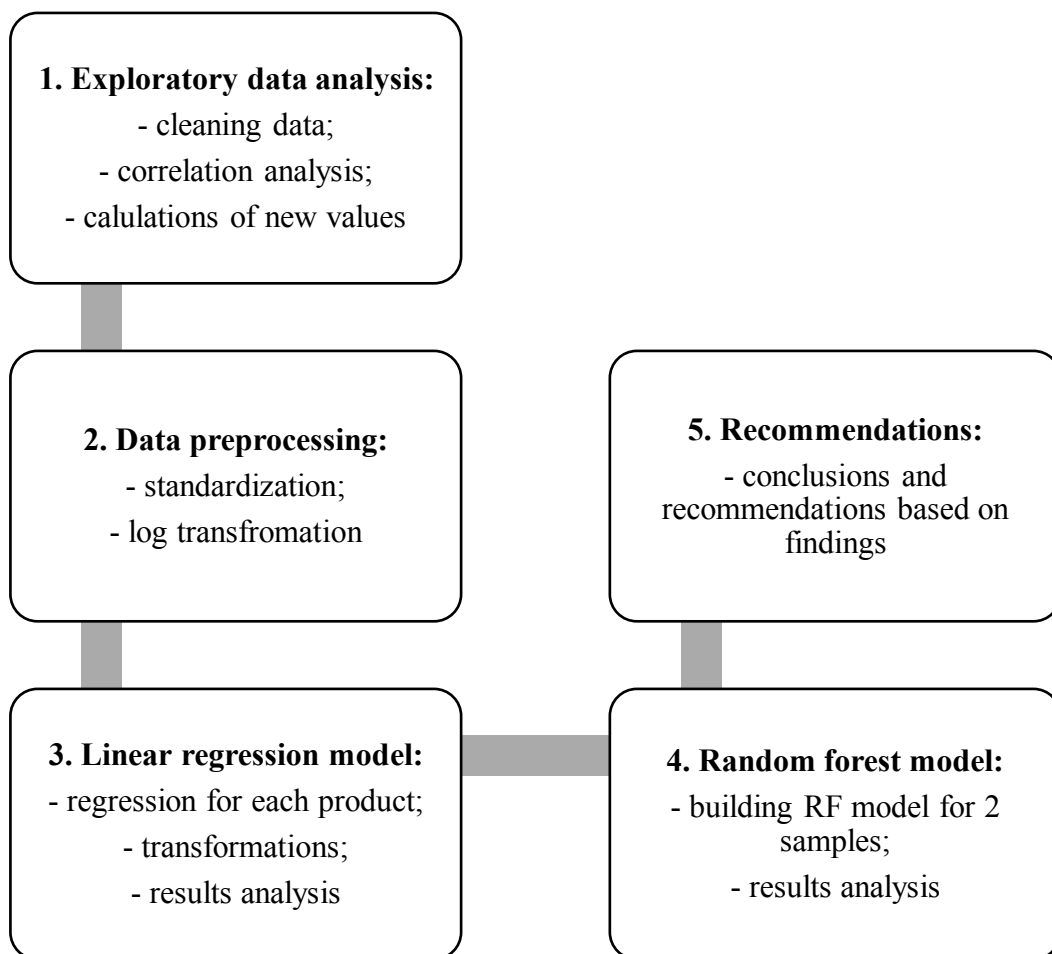
Based on the goal and objectives that were identified in previous sections, as well as based on the relevant literature, the following plan for the implementation of the project has been developed. The step-wise project plan is presented on Figure 3.

Step 1 will consist of exploratory data analysis, where we familiarize ourselves with data, clean it, identify challenges that may arise when working with the data, and identify some main trends that are present in the dataset.

Step 2 will consist of preprocessing data as necessary for the implementation of chosen machine learning models and algorithms. Preprocessing might include feature transformation techniques, such as standardization or normalization, and log transformation. Data transformation is essential when dealing with data, as it can help when working with skewed data, or data in different formats, which can lead to skewed results.

Step 3 will consist of building a linear regression model for each item separately, to identify which products demonstrate linear relationship between their sales and distribution, and those that do not. Then, we will analyze what differentiates former from the latter, to identify some common characteristics that can help the company in the future to identify whether a product will demonstrate linearity between its sales and distribution. We will also apply different types of transformations to see which one gives best results.

Step 4 will consist of analysis of those product that do not demonstrate linear relationship between sales and distribution. Mainly, we will try to analyze what form a relationship between sales and distribution of such products takes. For this stage, we will use random forest regression trees, as single decision trees are known to be weak, but random forest amplifies single tree's accuracy.



**Fig. 3** Step-wise analysis plan

Final fifth step consists of making recommendations based on our analysis and findings.

## 2. Exploratory data analysis

There are four major measures that make a start off for our analysis, and these are Trade Panel data for Weighted Distribution, Numerical Distribution, Value Sales (measured in local currency), and Volume Sales (measured in statistical unit of measure, which helps to compare volume sales between forms and categories proportional to product pack size). Geographically, these measures represent one country, and distribution includes three channels. As for the time period, it spans approximately five years, and the data are available for each month, from December, 2015, to August, 2020.

There are also around 4000 items (keys for matching with data from above), and each of these items has following characteristics: category of the product, company that produces it, brand under which it is sold, form of a product, two segments, which stand for some specific features of the product, and the size of the product.

In Table 1 and Table 2 presented are the samples of the data. Information about specific products, namely their characteristics, is presented in Table 1, while trade panel data is presented in Table 2.

**Table 1** Master data

Item	Category	Company	Brand	Form	Segment 1	Segment 2	Size
Item_1	Category_1	Company_1	Brand_1	Form_1	Segment_1	Segment_1	Size_23 9
Item_2	Category_1	Company_1	Brand_1	Form_1	Segment_2	Segment_2	Size_22 4
Item_3	Category_1	Company_2	Brand_2	Form_2	Segment_3	Segment_3	Size_37
Item_4	Category_1	Company_2	Brand_3	Form_3	Segment_4	Segment_4	Size_16
Item_5	Category_1	Company_2	Brand_3	Form_1	Segment_2	Segment_5	Size_20 5

Source: [Procter & Gamble, 2020]



**Table 2** Trade panel data

<b>Item</b>	<b>Area</b>	<b>Month (TP)</b>	<b>TP Weighted Distribution</b>	<b>TP Numerical Distribution</b>	<b>TP Value Sales (MLC)</b>	<b>TP Volume Sales (MSU)</b>
ITEM_1	Area_1	2015M12	11.27	0.77	602.72	0.65
ITEM_2	Total	2015M12	0	0.01	1.86	0
ITEM_3	Area_1	2015M12	14.15	1.71	1041.76	1.06
ITEM_4	Area_1	2015M12	0.6	1.73	44.59	0.01
ITEM_5	Area_2	2015M12	0.01	0.05	4.64	0

Source: [Procter & Gamble, 2020]

First, we will start with cleaning the data. There are several conditions which must be accounted for when cleaning these data:

- when the item's price is not zero, we assume that at least 1 item has been sold - so that the real SKU is tracked in the database. More formally this would mean: if TP Value Sales  $> 0$ , then its TP Volume Sales should be  $> 0$  as well;
- if any item has been sold, its price (value) has to be higher than zero as we do not give any items for free. if TP Volume Sales  $> 0$ , then its TP Value Sales should be  $> 0$  as well;
- TP Weighted Distribution and TP Numerical Distribution are stated as percent measures, so these cannot exceed 100.

We clean data that doesn't correspond to the rules above. We started with more than 140,000 data points, and after we are done cleaning out the data that does not satisfy the conditions, we are left with around 96,000 data points.

There are many missing values in the dataset: nearly 15,000 data points. We have 2 options what to do with this:

- clean all missing values, or
- change missing values to the minimum nearest value.

We have decided to clear all missing data so as not to skew the analysis results with potentially biased or wrong data.

Before building any models, we have to analyze whether there is any correlation between distribution, both weighted and numerical, and sales volume and sales value, and whether it is significant. So, the hypothesis is that there is no correlation between distribution (both weighted and numerical), and sales (both value and volume). The alternative hypothesis is that there is correlation between all those variables. To check these hypotheses, we will use Pearson's

correlation coefficient, calculated in Jupyter Notebook using Python programming language. The results are presented in Table 3. To determine whether we fail to reject the null hypothesis or not, we will be setting significance level at 0.05.

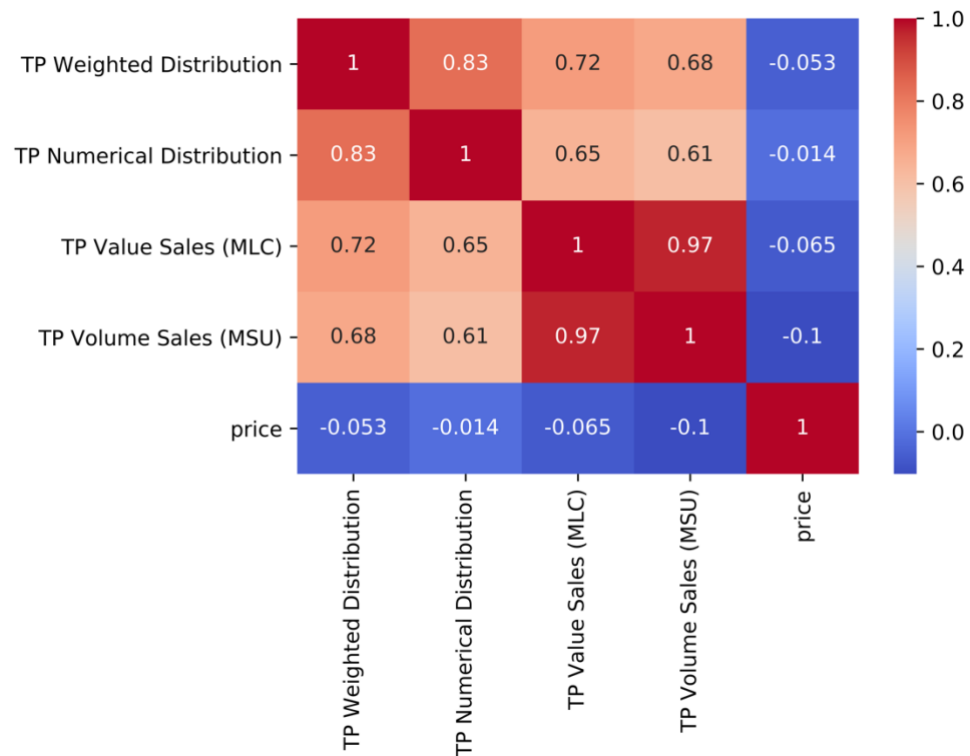
**Table 3** Correlation analysis results

	Weighted distribution	Numerical distribution
Value sales	0.725, p-value = 0.0	0.661, p-value = 0.0
Volume sales	0.69, p-value = 0.0	0.619, p-value = 0.0

As it is stated in Table 3, for each pair of variables on which the correlation analysis was conducted, Pearson’s p-value is 0.0, which is lower than significance level 0.05, so we reject the null hypothesis, which states that there is no correlation between those variables, in favor of the alternative hypothesis, which states that there is a correlation.

The data shows reasonably high positive correlation with the highest one being between weighted distribution and sales value. Basic distribution theory states that the weighted distribution relates to the product being presented in the right store for the right consumer, so if the brand is presented carefully, then its sales value (and volume too) increases. We make the following conclusion based on a theory: if an item's weighted distribution is higher than its numerical distribution, then the item is well distributed as the item is located in the stores that attract relevant people that would probably buy the item. If the opposite inequality happens, then the product is wrongly distributed and needs reconsideration.

Afterwards, we calculate the price of each product, by dividing value of sales by volume of sales. Figure 4 shows correlation analysis which includes correlations between price and all the other variables, and also it includes correlation between numerical and weighted distributions, as well as sales volume and value sales.



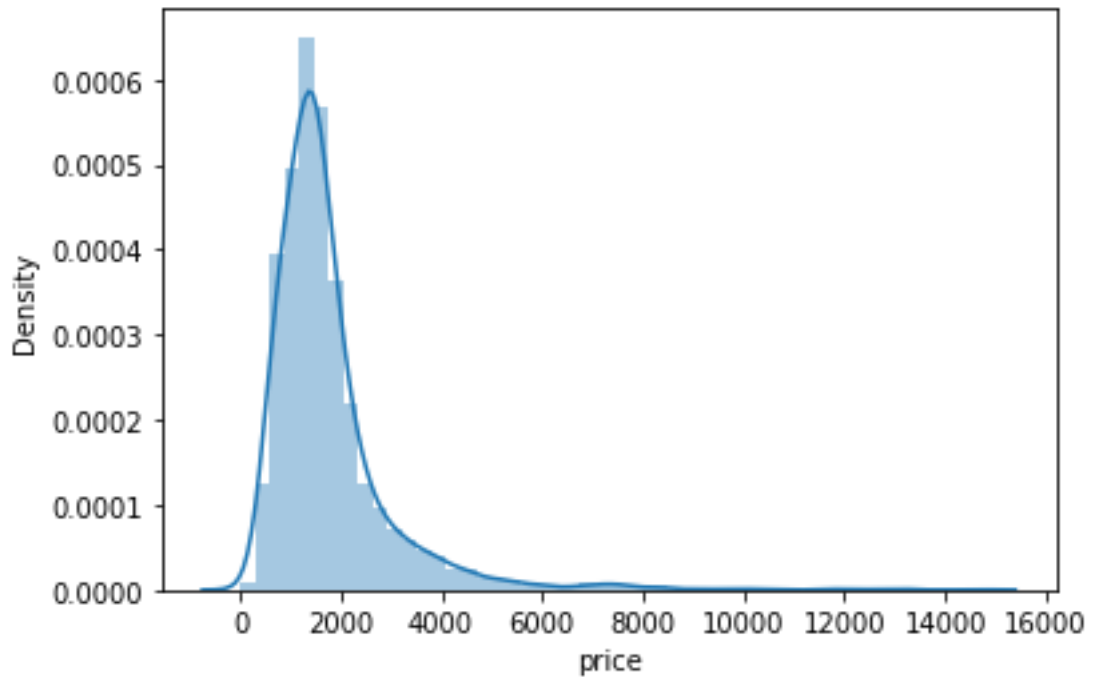
**Fig. 4** Correlation matrix

The mean price in the clean dataset is 1771.623, and the 75<sup>th</sup> percentile is equal to 2074.000. However, the highest price equals to 16,514, which might be an outlier, considering the industry and the data available to us.

Then, based on calculated prices, we divide products into four basic segments, which were decided roughly by the percentiles. So, the first segment is all the products with price equal to, or lower than 1000. Second segment includes all the products that cost between 1000 and 1500, and the third segment consists of all the products whose price ranges from 1500 to 2000. All the products whose price is higher than 2000, belong to the fourth segment. The distribution of products among segments is roughly equal, with 714 items belonging to the first, 988 items – to the second, 787 items – to the third, and 833 items belonging to the fourth segments.

Afterwards, we grouped prices by each item’s mean price, so that we have one price, weighted distribution, numerical distribution, value sales and volume sales values for each unique item, and to make further calculations easier.

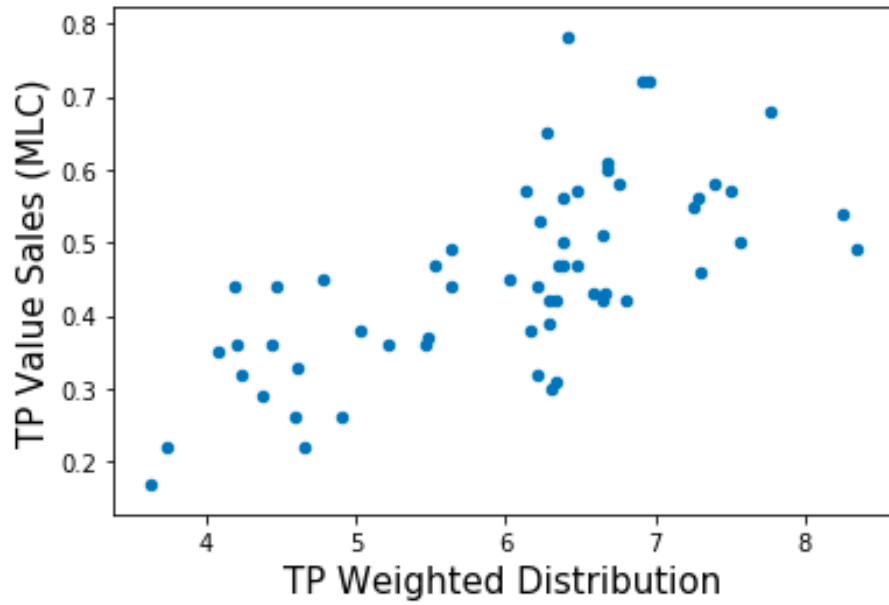
On Figure 5, we can see that the data points mainly lie in the middle-price segment - it is concentrated within 1000-2000 price range.



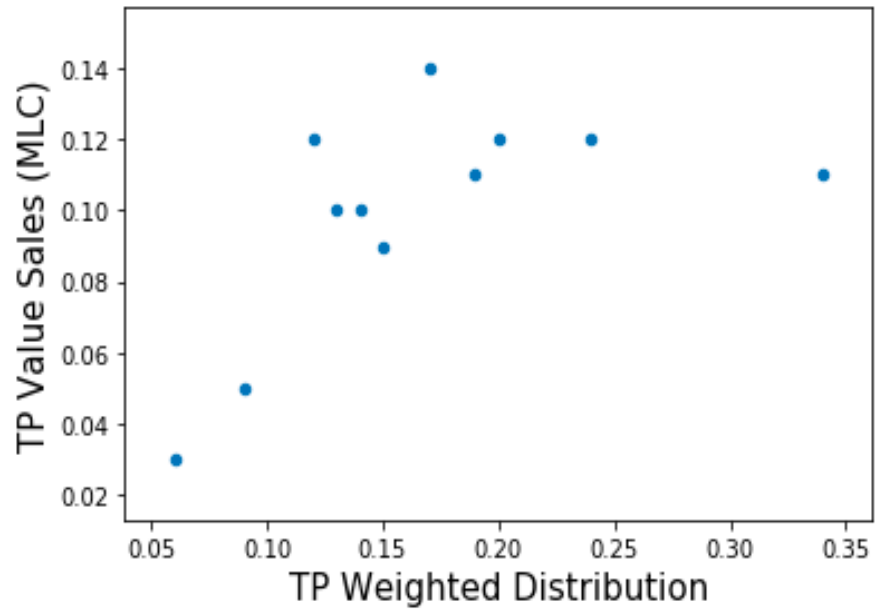
**Fig. 5** Distribution of products by price

Then, we will look at the relationship between sales and distribution of different items, chosen randomly, by creating scatter plots, and deciding whether linear relationship is present in any of them. Figure 6 shows such a relationship between sales and distribution of Item 1, and we can see that there is some linearity present, although it is not a perfect straight line. Figure 7 shows the relationship between distribution and sales of Item 100, and there is not linearity in sight, moreover, the relationship looks like a logarithmic function. Figure 8 shows the relationship between those two variables of an Item 3333, and the relationship there also has some linearity to it. Finally, Figure 9 shows the relationship between sales and distribution of Item 568, and there seems to be no discernible relationship at all.

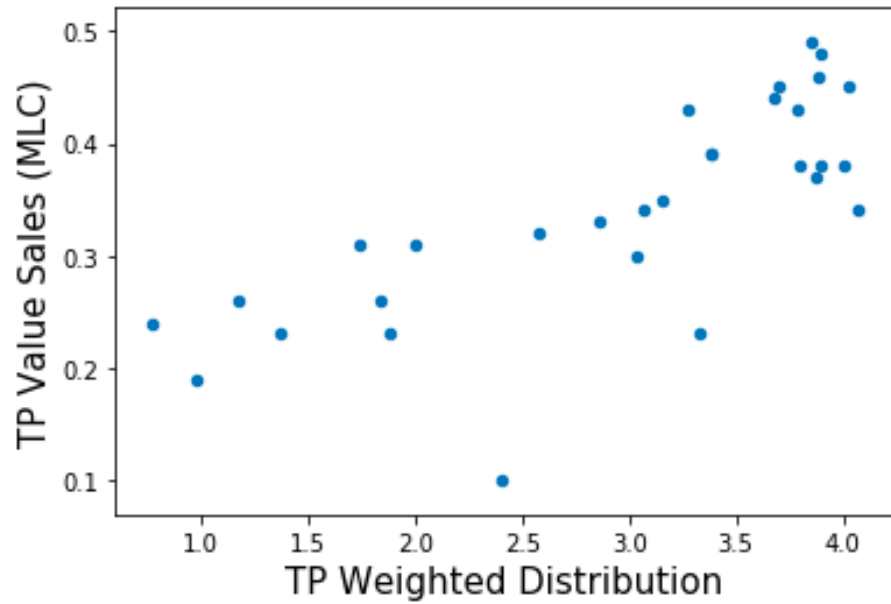
This short analysis has shown us that while some products might demonstrate a linear relationship between their value sales and distribution, other products tend to demonstrate non-linear relationship, while others yet – no relationship at all.



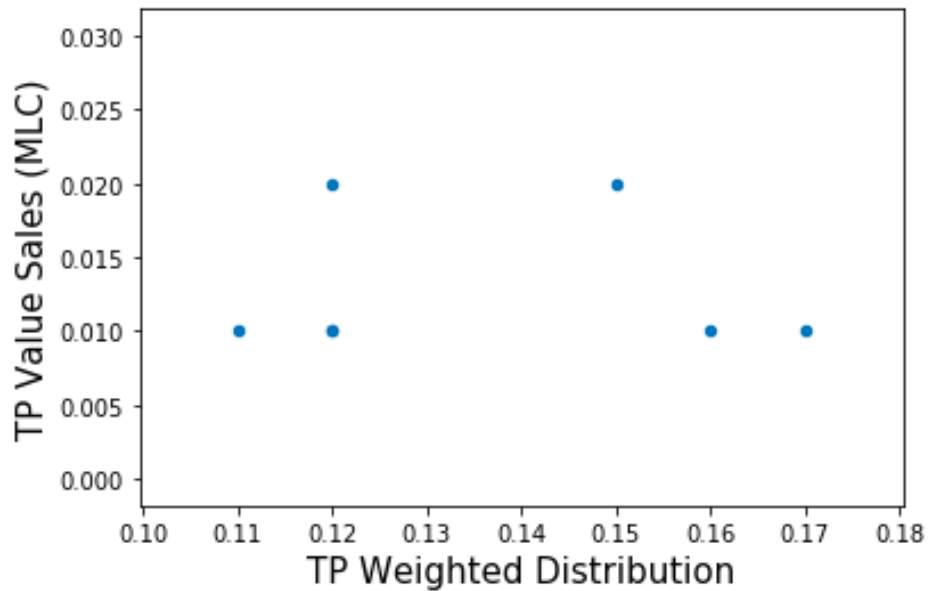
**Fig. 6** Relationship between Weighted distribution and Value Sales of Item 1



**Fig. 7** Relationship between Weighted distribution and Value Sales of Item 100



**Fig. 8** Relationship between Weighted distribution and Value Sales of Item 3333

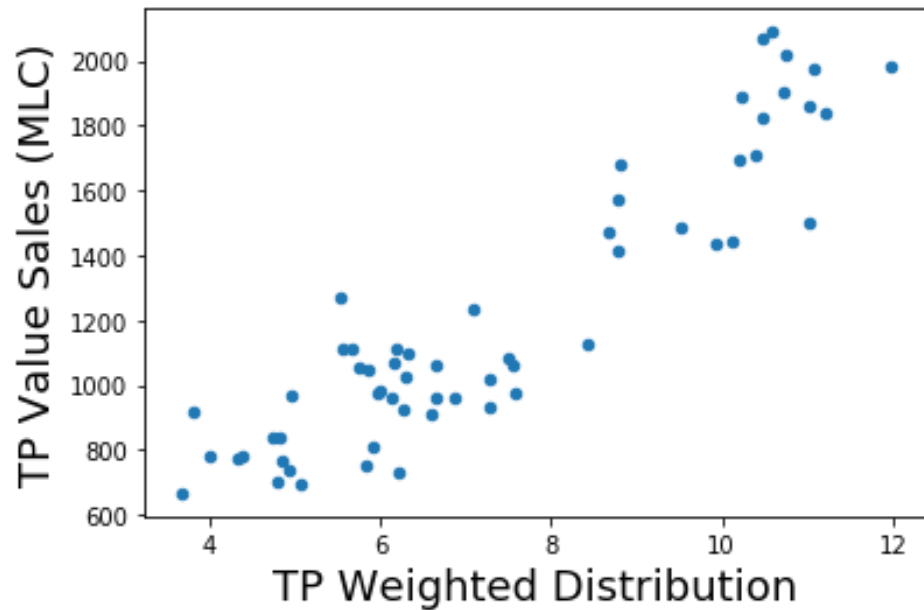


**Fig. 9** Relationship between Weighted distribution and Value Sales of Item 568

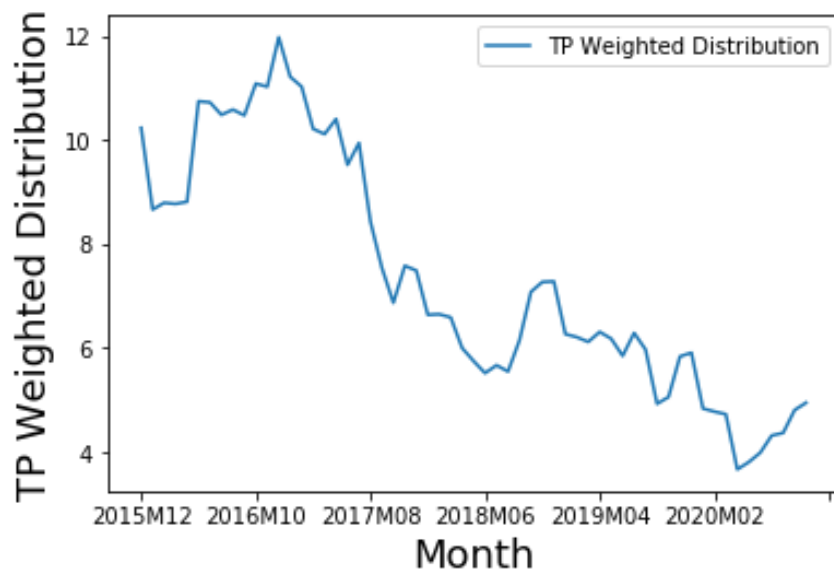
We will further look at different scatter plots to see what differentiates products with possibly linear relationship between sales and distribution from those that demonstrate non-linear distribution between those two variables. We will be selecting products randomly, and then looking at them deeper.

Figure 10 shows relationship between sales and distribution of Item 3. We can see that there is a linear trend there, as higher weighted distribution equals to higher value sales. When looking at its weighted distribution over time, we can also see that it has a clear downward trend, with some spikes here and there. Item 3 is also one of the products that has been sold for 59 months, throughout the whole time period available in our data. But Item 1 has also been sold for this same

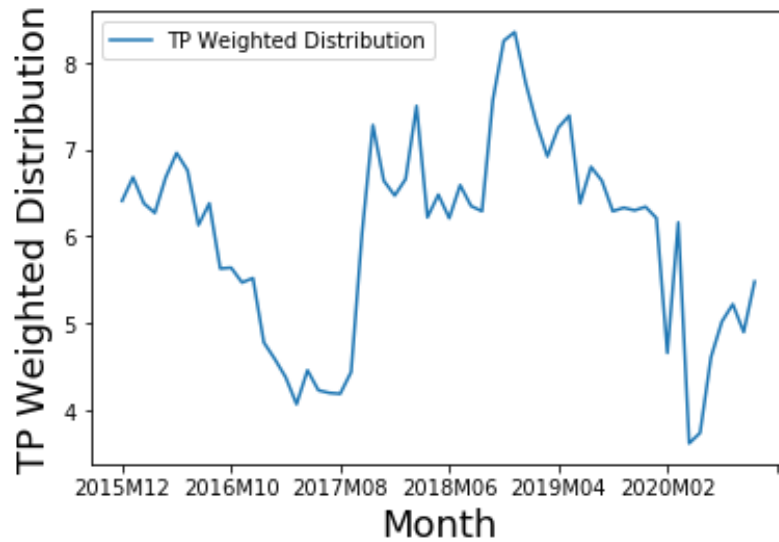
time period, and it does not portray such clear linear trend. The difference between those two items is that weighted distribution of Item 1 has not been stable, or followed any trend line over this time period, as can be seen on Figure 12.



**Fig. 10** Relationship between Weighted distribution and Value Sales of Item 3

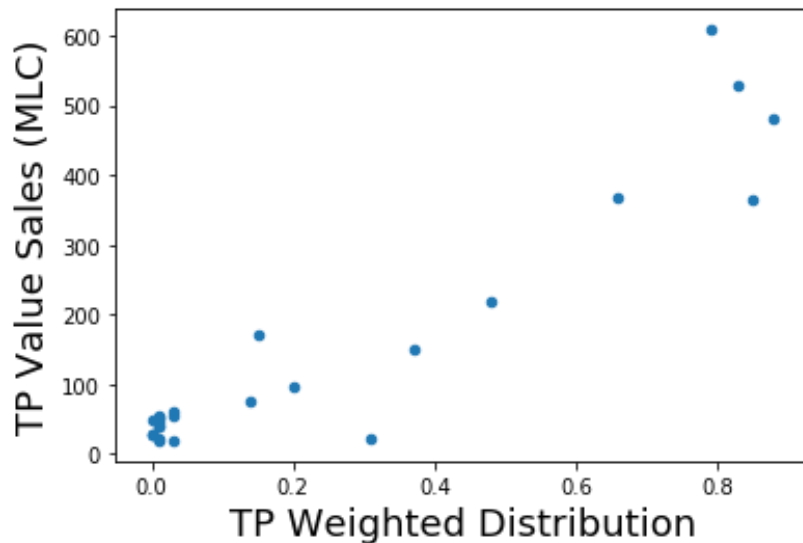


**Fig. 11** Weighted distribution of Item 3



**Fig. 12** *Weighted distribution of Item 1*

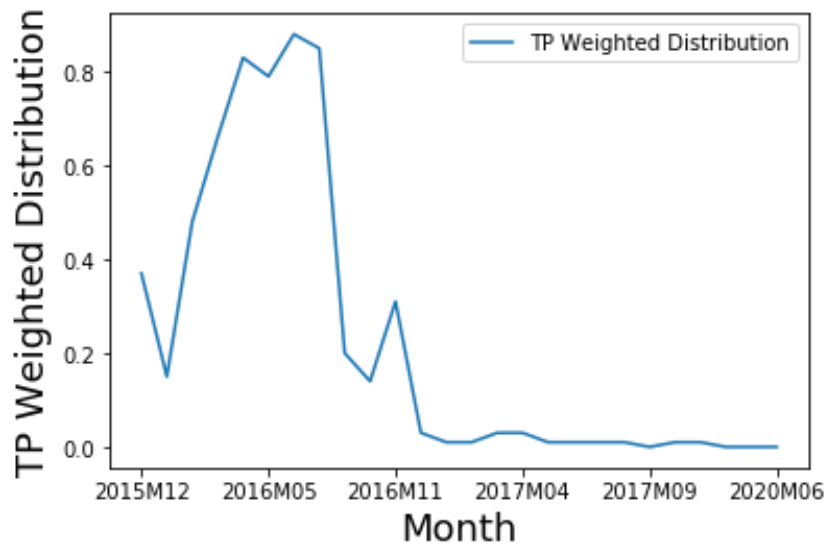
Another example of a product with a linear trend between sales and distribution is Item 443, shown on the Figure 13. Figure 14 shows that its weighted distribution hasn't been very stable over the years, with some steep rises and falls, but it has been sold for 26 months, more than 2 years.



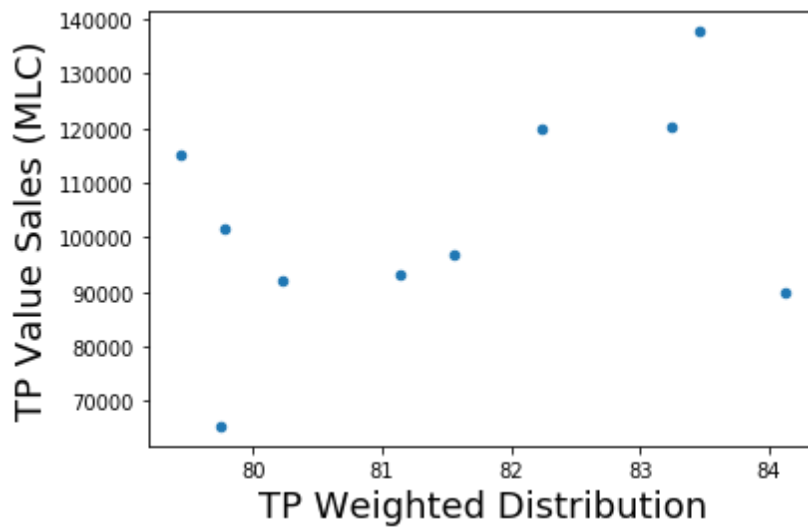
**Fig. 13** *Relationship between Weighted distribution and Value Sales of Item 443*

Another example of a product that does not demonstrate linear relationship between sales and distribution is Item 1899, as shown on Figure 15. Its weighted distribution has also not been following any obvious trend, with sharp increases and decreases, and it has also been on market only for 10 months.

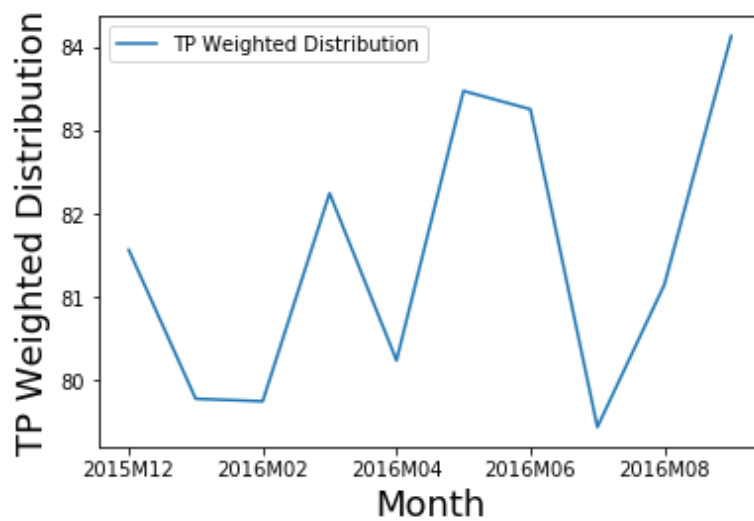




**Fig. 14** Weighted distribution of Item 443

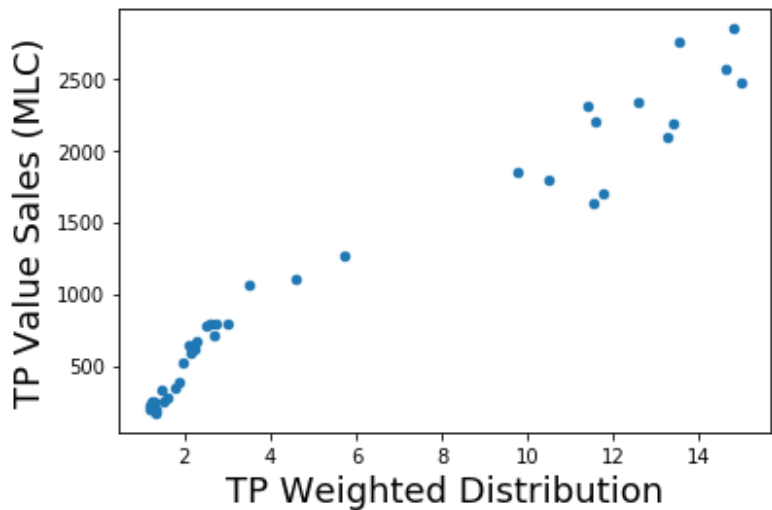


**Fig. 15** Relationship between Weighted distribution and Value Sales of Item 1899

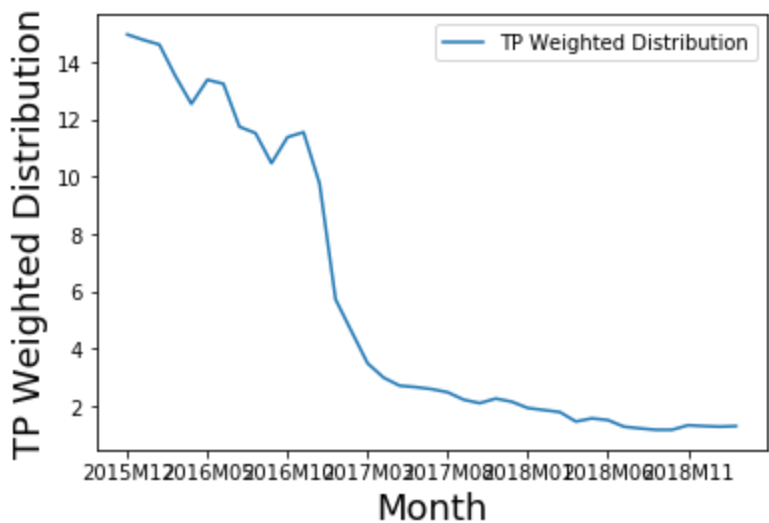


**Fig. 16** Weighted distribution of Item 1899

Looking at some more products, Figure 17 shows relationship between sales and distribution of Item 987, and there also seems to be a linear trend there. Figure 18 shows this product's weighted distribution, which also seems to be following a downward trend, with a steep decline over couple of months. This item has been on market for 39 months.

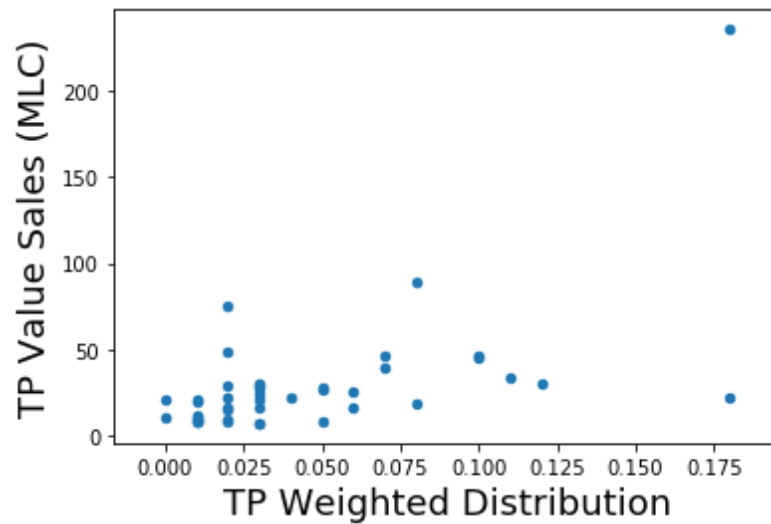


**Fig. 17** Relationship between Weighted distribution and Value Sales of Item 987

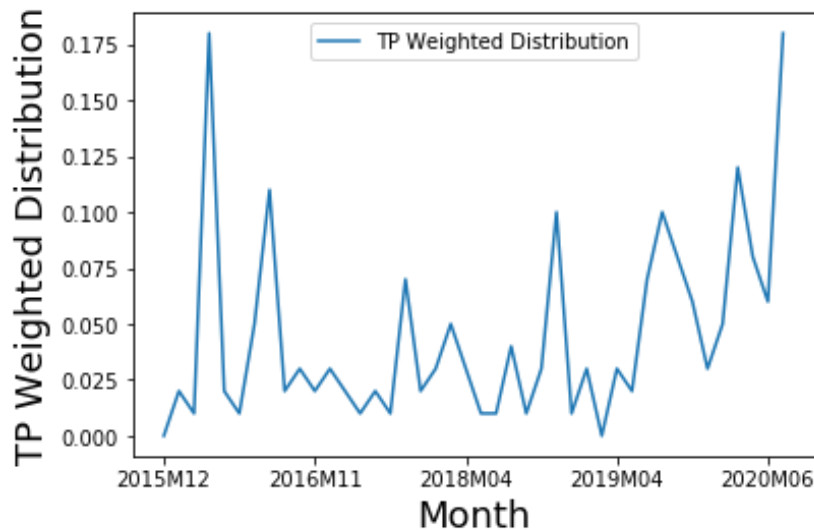


**Fig. 18** Weighted distribution of Item 987

Relationship between sales and distribution of Item 10, which is shown on Figure 19, does not follow an obvious linear trend. Figure 20 shows this product's weighted distribution over available time period, which is equal to 42 months, with some months missing in-between. Its weighted distribution has been quite unstable over this period, with constant rises and declines.



**Fig. 19** Relationship between Weighted distribution and Value Sales of Item 10



**Fig. 20** Weighted Distribution of Item 10

We can assume from this short analysis that products that are likely to demonstrate linear relationship between sales and distribution, are the products that have been on market for around one year at least, so that they are well-established, and products whose weighted distribution has not been all over the place, but pretty consistent. When those two assumptions are met simultaneously, as we have seen that some well-established products do not have linear relationship between sales and distribution, the product is more likely to be demonstrating linearity in the relationship between sales and distribution. A possible explanation for this could be the fact that if a product is sold for an extended period of time, and is distributed pretty consistently over this period of time, then consumers have time to get used to this product and start buying it consistently too. Products with random spikes in their distribution are more likely to create a sense of scarcity, or drive consumers to forget about the product and simply find its substitute.

### **3. Challenges**

The main challenge with our dataset is that some products have a lot more data about them, simply because they have been selling for a much longer time period. On the other hand, it gives us ample opportunity to analyze whether the novelty of the product somehow affects the relationship that its sales and distribution demonstrate. From the technical point of view, it is a bit of a problem as it might skew the analysis, but from the business perspective, it is no problem at all, seeing as how it might give us insights that were unaccounted for.

There is also the problem of missing data, and data that was incorrectly transferred to the dataset. Some Items have no information on their value sales and volume sales, while still having information about their distribution. As for the second case, for some products data on the value sales is available, but the volume sales come in the form of zero, and it can lead to infinities in the table, so we have decided to drop data points like this. For the former case, we have decided to drop those data points altogether. We have been informed that the second case is not due to some specific way of calculations that company employs (although it is known that Procter & Gamble does employ a specific way of calculating volume sales by standardizing products across the categories, to be able to compare apples to oranges, so to speak), so deletion of such data points does not result in loss of some important information.

Another challenge that might arise, should we decide to run models that are more computationally expensive than simple regressions, or even trees, is the computational abilities of our machines. Fortunately, there are ways to deal with this challenge, from connecting to stronger sources, such as Google Colaboratory, or reducing data, which comes with the risk of reducing important data, which might lead to skewed results.

Finally, one of the biggest challenges is interpreting the results. As we have stated earlier, distribution alone does not drive the sales, although it is assumed that distribution is the main driver. Other marketing mix elements, such as product, price, and promotion, also influence the sales volume, because they are also almost always at play when it comes to sales. And since we have no data on other marketing mix elements, our results will be limited by this. In addition, we will be using random forest for products that do not demonstrate linear relationship between variable of interest, a black box model that is known for being hard to interpret. But all of those challenges have been discussed with Procter & Gamble, and this research is done with full acknowledgment of such limitations.

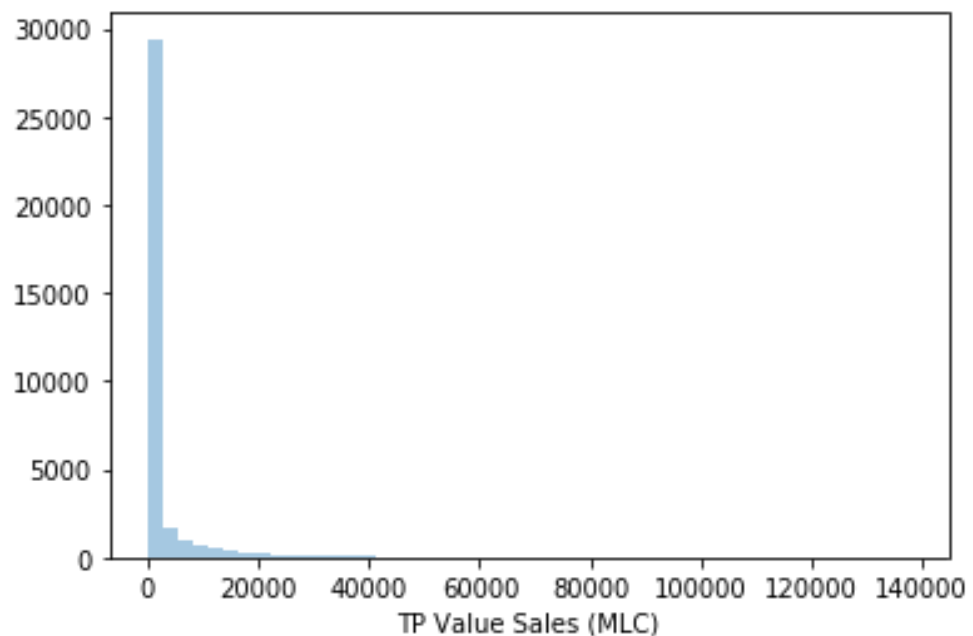
### **4. Linear regression: preliminary results and challenges**

As was stated in our step-wise project plan, we first need to transform our data to have more accurate results. Our dependent value is value sales, and our independent value is weighted distribution. First, we will look at the distribution of the value sales variable, which is presented

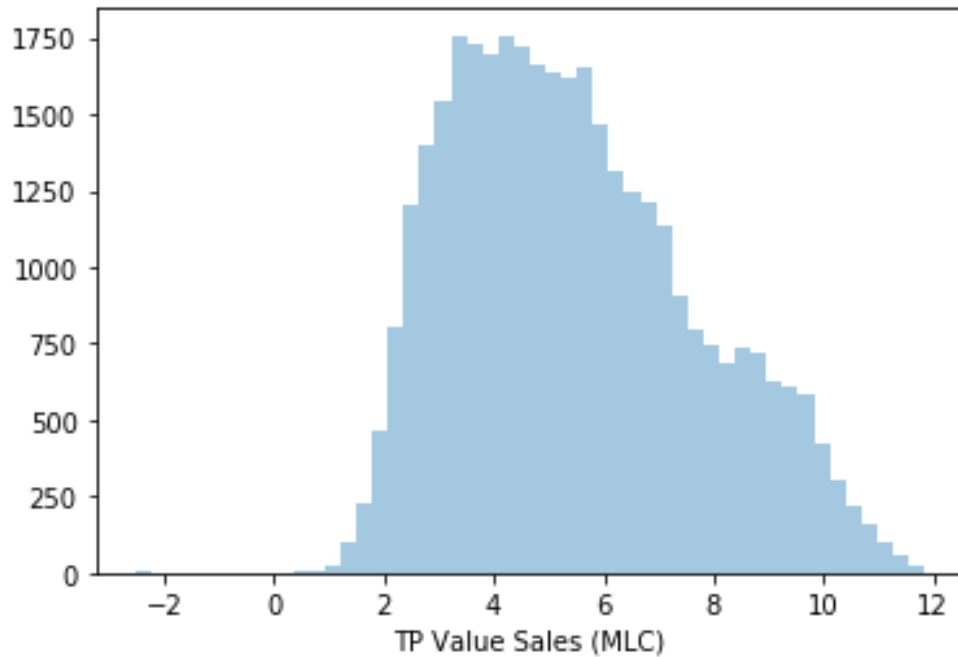
on Figure 21. We can see that the distribution of the variable is not normal, and that might hinder our analysis, so we will apply log transformation on that variable to increase the accuracy of our model. Figure 22 shows distribution of value sales variable after we have applied log transformation to this variable, and we can see that the distribution is closer to normal now. In general, log transformation helps the analysis in several ways:

- makes the data more normal, more symmetric;
- helps meet the assumption of normality;
- helps meet the assumption of constant variance when it comes to linear modeling;
- helps make a non-linear relationship more linear.

All of those ways are helpful in our case. We will have to transform the results a bit when we will be interpreting them, but it is not a big issue.



**Fig. 21** Distribution of Value sales before log transformation



**Fig. 22** Distribution of Value sales after log transformation

We start our analysis with linear regression, as it is one of the best methods available to determine how one variable affects the other, and also to determine whether there is a linear relationship between dependent and independent variables.

We have run linear regression model on all four thousand products, and we will be using adjusted R-squared to determine whether or not model the model is accurate, and whether or not there is linear relationship between sales and distribution of each product. We have decided that products, whose R-squared is 75%, or higher, are those that demonstrate the linear relationship between their sales and distribution. An acceptable value of R-squared depends wholly on the research, and we have decided on that number because we realize that distribution is usually not the only element of marketing mix in play, but if it explains more than 75% of the variance of the dependent variable, than it is reasonable to assume that it is a main driver of the sales of such products.

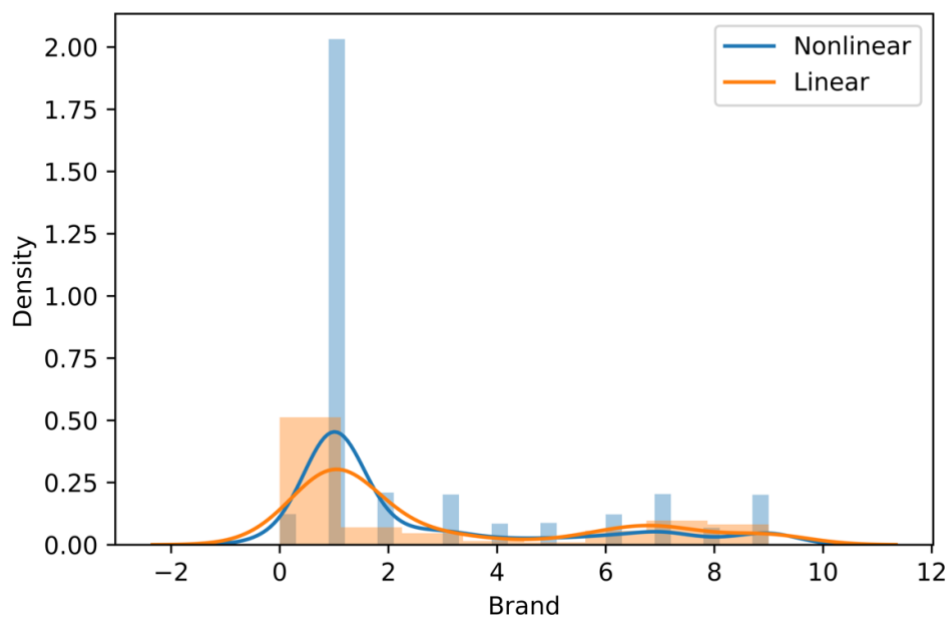
An important part of the research is to check whether the Gauss-Markov theorem assumptions. As for the first assumption, we have already assessed the relationship between dependent variable, which is sales value, and independent variable, which is weighted distribution, in our case. There are some products that seem to demonstrate linear relationship between variables of interest, and there are also products that do not do that. Seeing as we are using linear regression model to determine which products demonstrate linear relationship between sales and distribution in the first place, it seems redundant to be assessing this assumption in this case.

As for other assumptions, we ran a Jarque-Bera residual normality test which performed well on the majority of the data points. Also, we ran a Breusch-Pagan heteroskedasticity test and

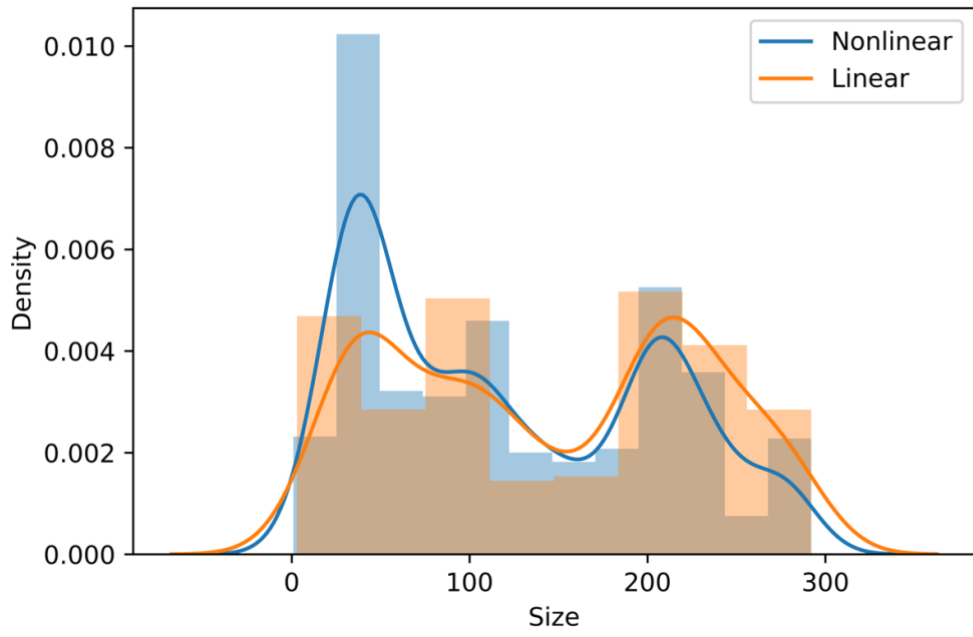
over 80% of the data demonstrated  $p\text{-value} > 0.05$ , confirming the null about the homoscedasticity. Our tests also showed that the expectation of residuals is indeed 0. Autocorrelation test revealed that the residuals mostly do not have one – the metric results lie within the norm.

After we have run a linear regression model on our data, and then split data by the value of R-squared, there were 516 products that demonstrated linear relationship between their sales and distribution, and 1852 products that did not demonstrate linearity in the relationship between variables of interest.

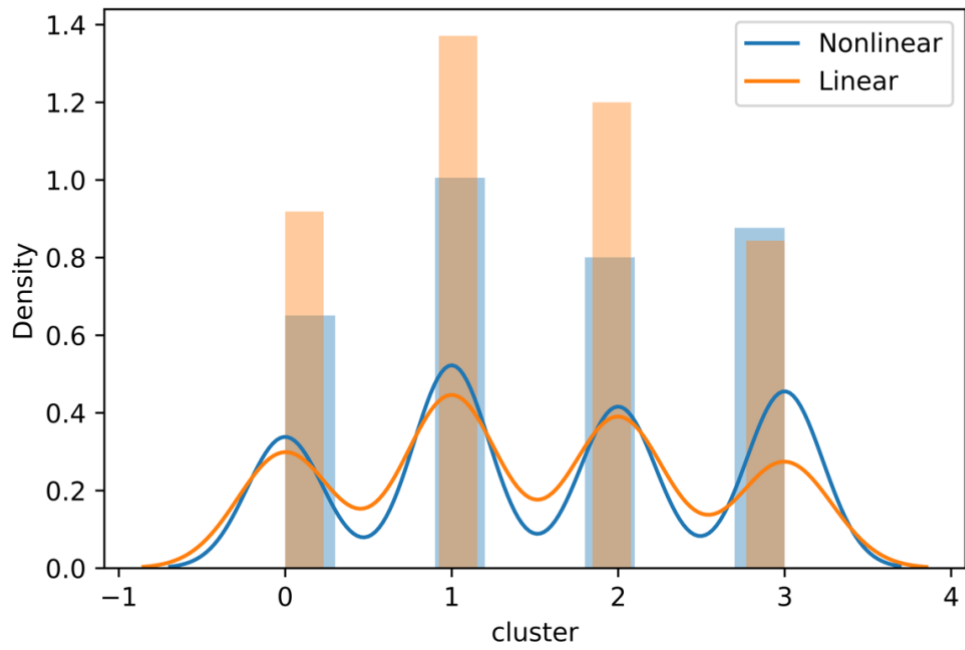
The next step that needs to be taken is the analysis of what differentiates products that demonstrate linear relationship between their sales and distribution from those that do not do that. We will start with comparison of categorical variables, which are: company, brand, form, segment 1 and segment 2, and size. To do that, we will use density charts to see whether any of the options in each categorical variable is more present than others in either sample of products with or without apparent linear relationship between sales and distribution.



**Fig. 23** Distribution of brands across samples

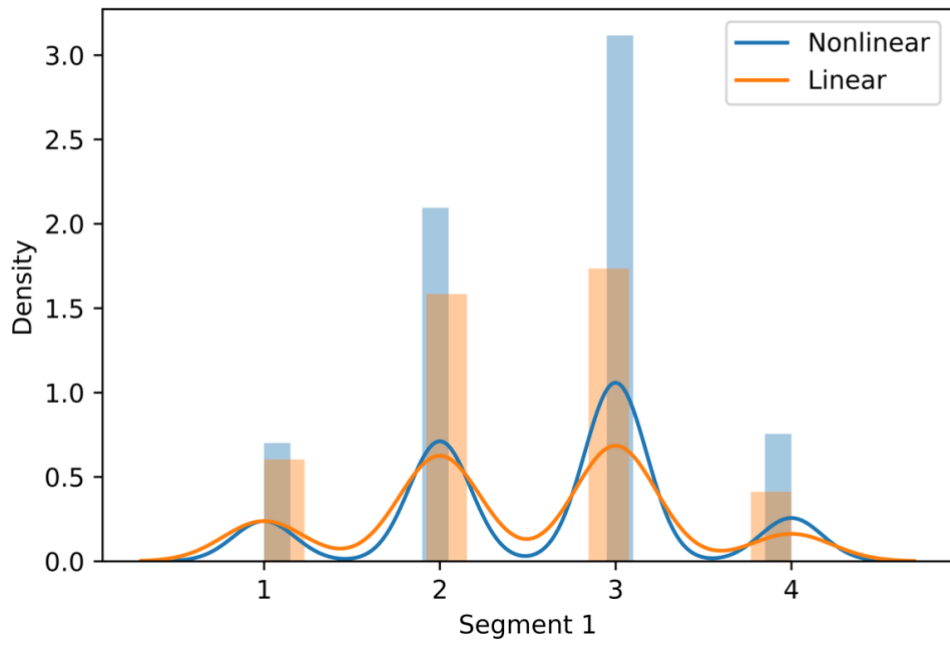


**Fig. 24** Distribution of sizes across samples

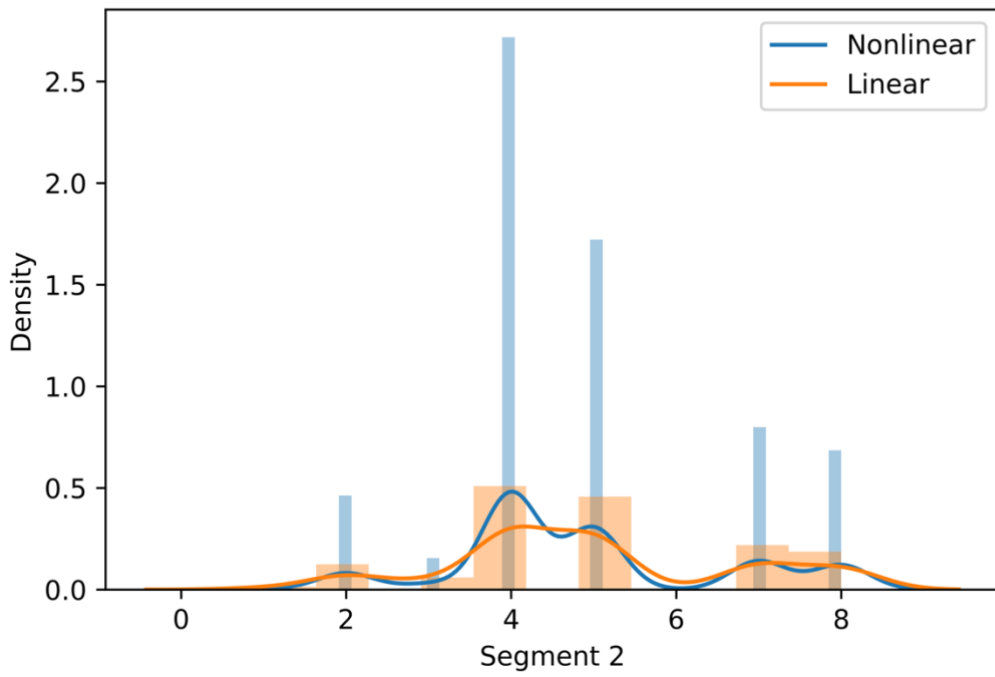


**Fig. 25** Distribution of clusters across samples

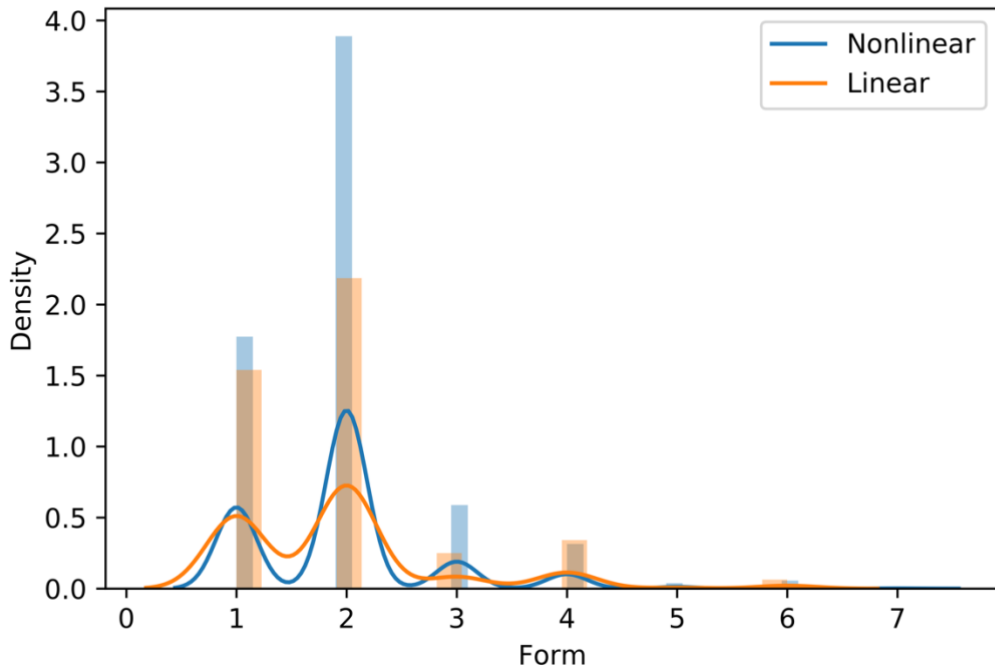




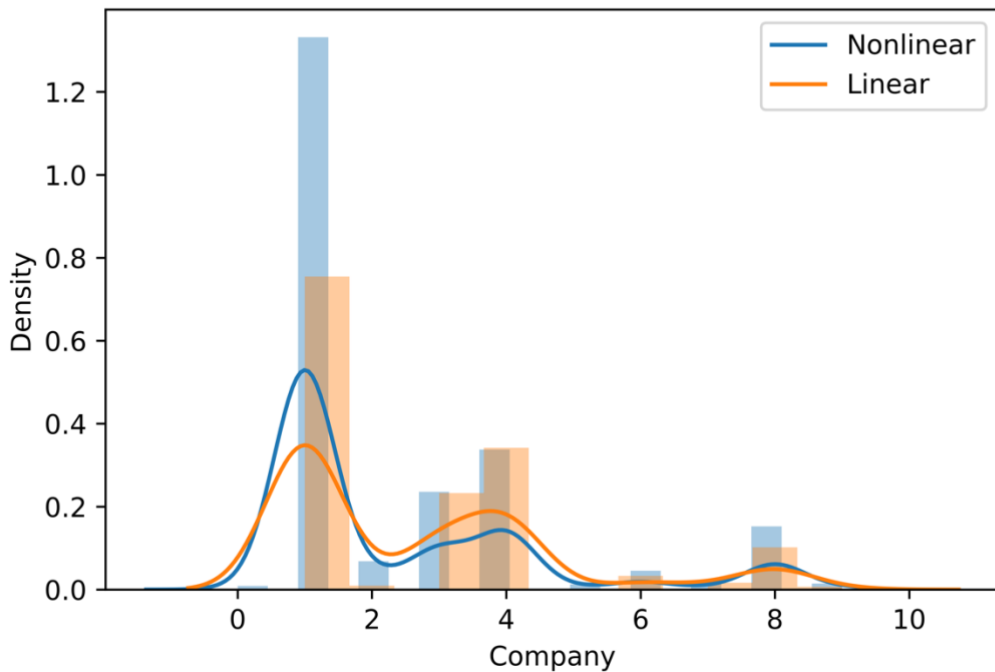
**Fig. 26** Distribution of segment 1 across samples



**Fig. 27** Distribution of segment 2 across samples



**Fig. 28** Distribution of forms across samples



**Fig. 29** Distribution of companies across samples

Figures 23-29 show how different categorical variables are distributed across samples, with color blue representing products with non-linear relationship between sales and distribution, and color orange – products with linear relationship between sales and distribution. Figure 23 shows

distribution of brands, and while the distribution seems similar for both samples, there is a definite spike of products of Brand\_1 in the non-linear relationship sample. Figure 24 shows the distribution of sizes, and there is an obvious difference there, as we can see from the chart that there are much more products with sizes 40-50 in the non-linear sample. Figure 25 shows distribution of clusters, which is equal in both samples. Figure 26 and Figure 27 show distributions of segments, and there isn't an obvious difference there, although there is a little spike of Segment\_3 in non-linear relationship sample in the former chart. Figure 28 shows the distribution of forms, and again, there seems to be a spike of Form\_2 in non-linear relationship sample. Finally, Figure 29 shows the distribution of companies across the samples, and there also seems to be obvious difference between the two samples.

From these charts, we can assume that the categories that influence whether the product shows linear relationship between sales and distribution are size and brand, and a smaller effect can be attributed to the form of the product. As all the data has been anonymized, we cannot say exactly what those sizes and brands are, but the client company has that information and will be able to apply results when necessary.

Now, we will look at the coefficients of Weighted Distribution of products that demonstrate linear relationship between their sales and distribution. The values of coefficient range from (-112.27) to 137.05, with the median being 2.37, and the mean - 7.6. The negative values of the coefficient belong to ten items, but almost all of those items have a high adjusted R-squared, which can be seen in Table 4. We can see that all of those items are products of one company and brand, and the majority of them is in the third cluster.

**Table 4** Items with negative distribution coefficients

Item	R2	Comp.	Brand	Form	Size	Price	Cluster	Constant	Distribution
ITEM_1101	0.835051	1	1	2	56	3406.4	3	4.381651	-10.524258
ITEM_168	0.999843	1	1	3	205	13126.3	3	7.592536	-5.353944
ITEM_389	0.995958	1	1	1	239	1810.0	2	3.708787	-50.157878
ITEM_2173	0.896504	1	1	2	41	611.0	0	2.022196	-16.145884
ITEM_1589	0.752339	1	1	2	146	1346.3	1	4.809245	-88.069844
ITEM_1317	0.840864	1	1	1	245	2882.5	3	6.368686	-32.674472
ITEM_704	0.992551	1	1	2	68	2697.13	3	5.744718	-52.669019
ITEM_3486	0.968055	1	1	1	205	530.4	0	5.419120	-19.871418
ITEM_609	0.965457	1	1	1	239	1334.0	1	4.121126	-112.266461

Item	R2	Comp.	Brand	Form	Size	Price	Cluster	Constant	Distribution
ITEM_1458	0.989731	1	1	1	224	529.75	0	2.015260	-8.769199

The highest coefficient is 137.056, of Item\_3283, with an adjusted R-squared equal to 0.99.

This product is also of the same company and brand, like products from above, but it could be simply due to the fact that there are much more instances of that brand than any others. To understand true estimates, we have to recalculate coefficients, as the regressions were built using log transformed values of the dependent variables. We have to calculate the exponential of the coefficient, subtract one from it, and multiply by 100, to see the percentage change of sales value if the distribution is increased by 1 unit (in our case – percent). There are items for which the change of the sales value is not so dramatic, for example, for Item\_2340, the sales value would increase by 12% if the weighted distribution increased by 1%. In general, almost all coefficients of distribution are statistically significant (p-value < 0.05), with only 60 items' coefficients statistically insignificant.

### 5. Linear regression transformations: results and challenges

We are going to apply different transformations techniques to our linear regression model, to see if we can improve the results. We are going to use logarithmic transformation, reciprocal transformation, and square root transformation, as those are the most popular and commonly used transformations.

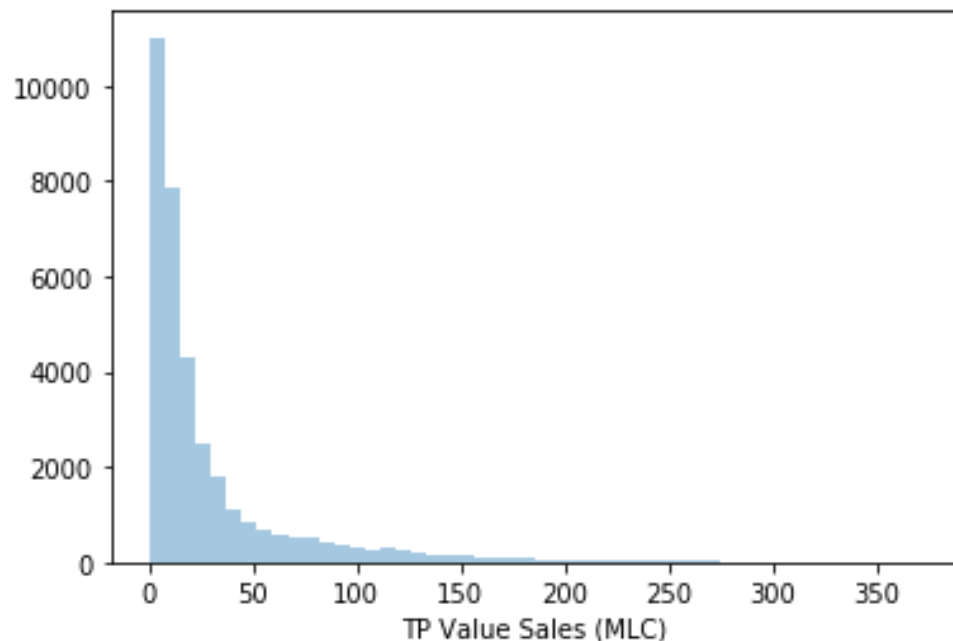
We have applied 3 different transformations in three different ways: only on the dependent variable (sales value), only on the independent variable (weighted distribution), and on both dependent and independent variables. We started out with 3319 items, but during the regression analysis around 400 of them were dropped, because their R-squared values were equal to NaNs, which could possibly mean that there is not enough data to make predictions for those items (their period must be too short).

**Table 5** Results of applying different transformations

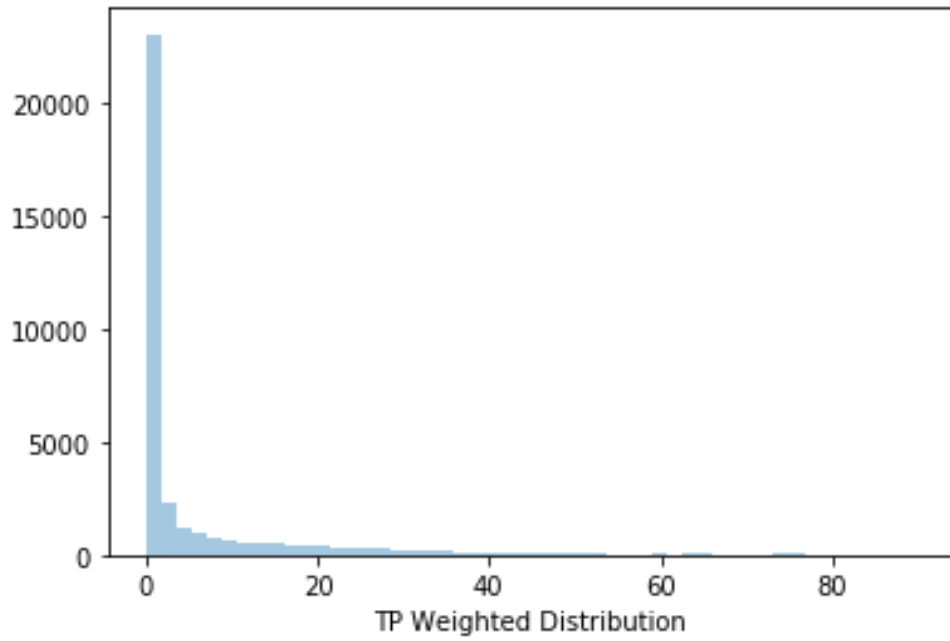
Transformation	Linear count	Non-linear count	Sum	%
Logarithmic Both	574	1751	2325	24.688172
Logarithmic X	383	1796	2179	17.5768701
Logarithmic Y	516	1876	2392	21.5719064
No transformation	499	1695	2194	22.7438469
Reciprocal Both	287	1766	2053	13.9795421
Reciprocal X	491	1649	2140	22.9439252

Transformation	Linear count	Non-linear count	Sum	%
Reciprocal Y	292	1824	2116	13.7996219
Square root Both	575	1763	2338	32.3781009
Square root X	463	1735	2198	21.0646042
Square root Y	586	1769	2355	24.8832272

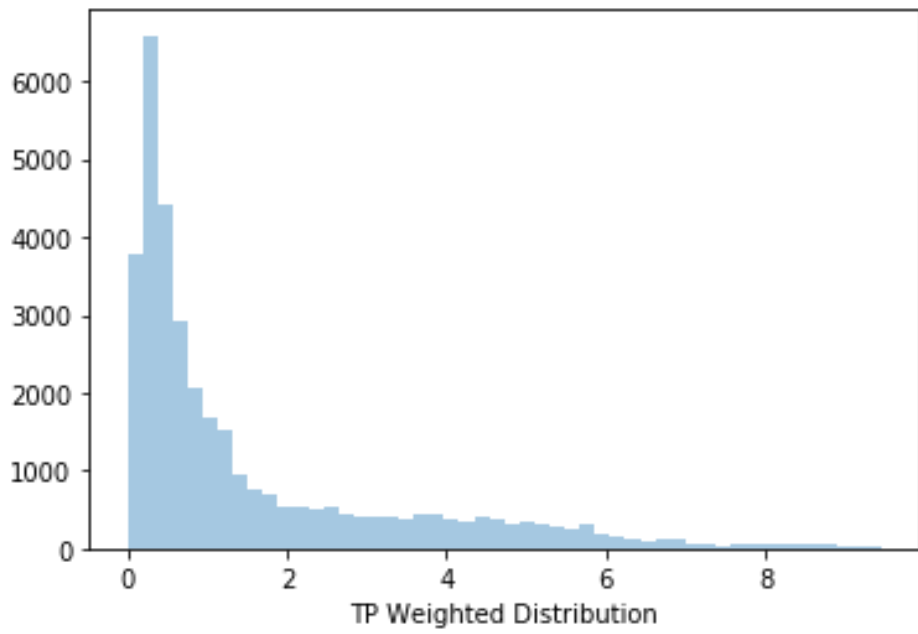
The best results, which for we defined as the highest number of products that demonstrate linear relationship between sales and distribution, based on the condition of R-squared being higher or equal to 0.75, were given by applying log transformation on both dependent and independent variables, using square root transformation on either both variables, or only the dependent variable. The result of all the transformations are present in Table 5. The worst results were given by applying reciprocal transformation on either dependent variable, or both variables, and the logarithmic transformation of independent variable. Using no transformation at all also gives medium results, as does reciprocal transformation of independent variable, as they fall right in the middle segment of all the results that were acquired by applying different types of transformation. Figures 21-22, and Figures 30-33 show the distribution of sales and the distribution of weighted distribution without any transformations, and with transformations that gave the highest results.



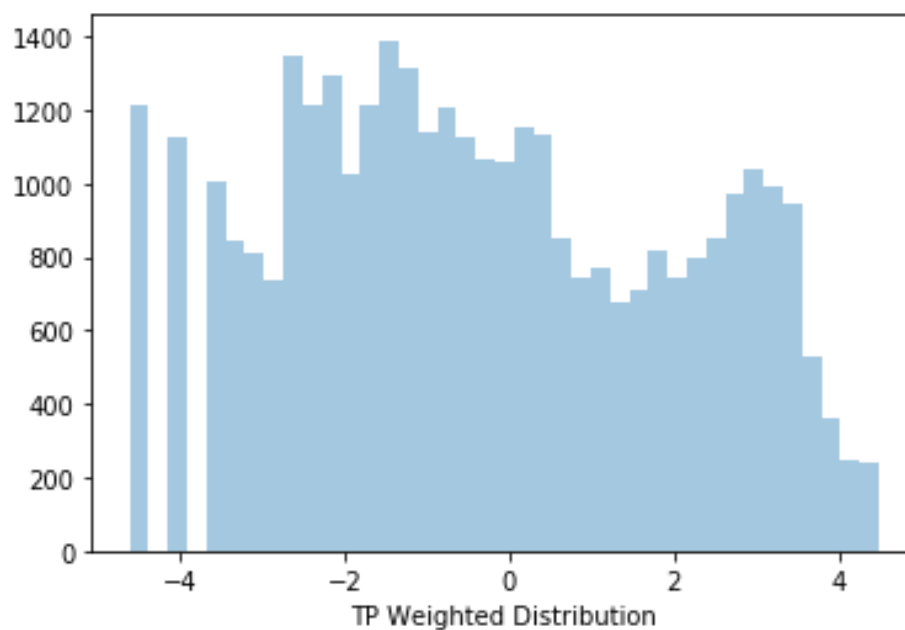
**Fig. 30** *Distribution of Value Sales with Square root transformation*



**Fig. 31** *Distribution of Weighted Distribution without transformations*



**Fig. 32** *Distribution of Weighted Distribution with Square root transformation*



**Fig. 33** *Distribution of Weighted Distribution with Log transformation*

Figure 33, for example, shows distribution of the independent variable after applying log transformation to it, and we can see that it doesn't really follow the normal distribution. Original variable is showing right skewed distribution, on Figure 31, and while square root transformed variable also shows right skewed distribution, on Figure 32, it is better than the former. Also, considering the industry, it is normal that the majority of the products are not extensively distributed, seeing as there are a lot of alternatives on the market, and each one is fighting for its place on the shelf.

It can be seen from Table 5 that applying square root transformation of both dependent and independent variables gives highest results, with square root transformation of only the independent variable in the second place. The problem with square root transformations of only one variable is that it does not aid in interpretation. In the case of only applying square root transformation on dependent variable, to interpret the effect of independent variable on the target, we would have to square both sides of the equation, leading to obfuscation of the interpretation of the coefficients in the root scale<sup>42</sup>. Furthermore, the square root transformation changes the functional relationship between predictors and outcome and the distribution of the errors.<sup>43</sup> So, while square root transformation is great at reducing right-skewedness, and linearizing the relationship between dependent and independent variables, it results in virtually uninterpretable regression coefficients in the square root scale.<sup>44</sup> But square root transformation can be useful for

<sup>42</sup> Pek, Jolynn & Wong, Augustine & Wong, Octavia. (2017). *Data Transformations for Inference with Linear Regression: Clarifications and Recommendations*. Practical Assessment. 22(9)

<sup>43</sup> Ibid.

<sup>44</sup> Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel hierarchical models*. New York, NY: Cambridge University Press.

prediction of the results, while log-transformation is good for interpretation of the coefficients.<sup>45</sup> That is why, we will be discussing the results of applying log and square root transformations to both variables, considering that former also gives one of the highest results out of all transformations.

We can further look at the results of the regression where both target and predictor variables were log transformed, to see how changes in weighted distribution affect the changes in sales value. Considering that both variables here were log transformed, we do not need to carry out any additional calculations to find value of coefficients of linear regression, we can simply look at those values. Figure 34 shows a sample of items that demonstrate linear relationship between sales and distribution. The mean value of the coefficient of distribution is 1.2477, and 75% of items have coefficients that are lower than 1.41. There are 69 items with coefficients of distribution higher than 2, and the highest coefficient is equal to 13.574445, belonging to Item\_3585, which was sold only for 3 months in 2016. Columns ‘Constant T’ and ‘Distribution T’ represent t-statistics of each variable in linear regression. As for p-values, they are described in Table 6. 55 items’ distribution coefficient’s p-value is higher than 0.05, meaning that when it comes to the distribution’s effect on sales of remaining items, it is statistically significant.

As for squared root transformation, the sample of items with linear relationship between sales and distribution is shown in Figure 35, and p-values of coefficients are described in Table 7. 51 items’ distribution’s coefficient is not statistically significant (p-value > 0.05), and 356 items’ constant is not statistically significant (p-value is > 0.05).

	Item	R2	Category	Company	Brand	Form	Segment 1	Segment 2	Size	Constant	Distribution	Constant T	Distribution T
0	ITEM_3	0.806058	CATEGORY_1	COMPANY_2	BRAND_2	FORM_2	SEGMENT_3	SEGMENT_3	SIZE_37	5.220290	0.938494	43.899696	15.558245
1	ITEM_14	0.798112	CATEGORY_1	COMPANY_1	BRAND_1	FORM_2	SEGMENT_3	SEGMENT_3	SIZE_32	5.804018	1.354986	38.821955	13.953831
2	ITEM_37	0.782947	CATEGORY_1	COMPANY_1	BRAND_1	FORM_3	SEGMENT_4	SEGMENT_8	SIZE_12	5.726233	0.941421	89.535062	11.750427
3	ITEM_58	0.955516	CATEGORY_1	COMPANY_4	BRAND_9	FORM_2	SEGMENT_3	SEGMENT_3	SIZE_37	6.163850	0.783399	105.922341	35.310492
4	ITEM_64	0.969783	CATEGORY_1	COMPANY_1	BRAND_1	FORM_3	SEGMENT_4	SEGMENT_8	SIZE_182	3.420988	1.269983	59.022599	34.936420
5	ITEM_69	0.893451	CATEGORY_1	COMPANY_2	BRAND_4	FORM_2	SEGMENT_3	SEGMENT_7	SIZE_37	6.137662	1.828581	41.665598	14.221386
6	ITEM_77	0.970602	CATEGORY_1	COMPANY_3	BRAND_11	FORM_4	SEGMENT_4	SEGMENT_3	SIZE_270	5.996412	0.974714	63.257089	33.023121
7	ITEM_101	0.936889	CATEGORY_1	COMPANY_1	BRAND_1	FORM_1	SEGMENT_2	SEGMENT_8	SIZE_190	5.306613	0.866237	48.757520	15.444202
8	ITEM_104	0.755105	CATEGORY_1	COMPANY_2	BRAND_13	FORM_2	SEGMENT_3	SEGMENT_7	SIZE_37	7.341184	1.674601	13.844563	6.873923
9	ITEM_105	0.786914	CATEGORY_1	COMPANY_3	BRAND_11	FORM_4	SEGMENT_4	SEGMENT_3	SIZE_273	3.558835	1.783226	7.959228	14.669381
10	ITEM_112	0.782997	CATEGORY_1	COMPANY_3	BRAND_11	FORM_4	SEGMENT_4	SEGMENT_3	SIZE_273	4.194665	1.709083	15.300376	14.500928
11	ITEM_122	0.752478	CATEGORY_1	COMPANY_1	BRAND_1	FORM_2	SEGMENT_3	SEGMENT_2	SIZE_87	5.724436	1.068918	18.110284	9.114884
12	ITEM_127	0.819666	CATEGORY_1	COMPANY_1	BRAND_1	FORM_2	SEGMENT_3	SEGMENT_2	SIZE_87	6.025870	1.257054	16.751368	8.586278
13	ITEM_126	0.967475	CATEGORY_1	COMPANY_1	BRAND_1	FORM_1	SEGMENT_2	SEGMENT_8	SIZE_205	4.705633	0.766815	116.646270	29.889372
14	ITEM_131	0.817575	CATEGORY_1	COMPANY_1	BRAND_1	FORM_2	SEGMENT_3	SEGMENT_2	SIZE_87	5.169737	0.868991	16.523752	7.092166

**Fig. 34** Sample of items that show linear relationship between sales and distribution with log transformation

<sup>45</sup> <https://vulstats.ucsd.edu/pdf/Gelman.ch-04.regression-transformations.pdf>



**Table 6** Description of p-values

	Distribution	Constant
count	574.000000	574.000000
mean	0.012179	0.015298
std	0.032944	0.071666
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000182	0.000000
75%	0.006428	0.000395
max	0.224421	0.804996

	Item	R2	Constant	Distribution	Constant T	Distribution T	Constant P	Distribution P
0	ITEM_2509	0.807699	3.120987	8.171676	2.622389	4.690506	0.058652	0.009374
1	ITEM_2888	0.857351	-16.864094	28.287974	-1.665863	5.572342	0.171070	0.005082
2	ITEM_2274	0.805695	-77.852418	84.694878	-3.527512	6.190235	0.007760	0.000262
3	ITEM_1272	0.878706	-14.188993	44.600945	-2.648802	6.668315	0.045492	0.001145
4	ITEM_3950	0.897937	-2.479770	7.555241	-3.127407	10.323500	0.009622	0.000001
5	ITEM_822	0.996181	2.152032	3.325525	34.870239	22.861397	0.018252	0.027829
6	ITEM_3869	0.867296	-2.057962	26.044549	-1.473993	5.209815	0.236920	0.013747
7	ITEM_3874	0.762876	-2.013542	13.924166	-1.345025	7.018400	0.200002	0.000006
8	ITEM_2904	0.881037	-8.931914	17.862638	-1.049169	6.166825	0.353307	0.003510
9	ITEM_2555	0.914656	49.487533	13.904124	5.590516	7.388246	0.005023	0.001789
10	ITEM_1437	0.983881	0.815496	11.882020	1.358388	13.568799	0.307271	0.005388
11	ITEM_3994	0.939645	-5.524435	10.458789	-4.972536	11.204829	0.001615	0.000010
12	ITEM_2637	0.989770	0.112441	14.632018	0.495687	17.066169	0.669226	0.003416
13	ITEM_2913	0.756495	4.619903	12.797635	0.580450	4.066139	0.592741	0.015271
14	ITEM_1269	0.774633	-7.237533	34.724925	-1.777454	5.947435	0.109214	0.000216

**Fig. 35** Sample of items that show linear relationship between sales and distribution with square root transformation**Table 7** Description of p-values

	Distribution	Constant
count	575.000000	575.000000
mean	0.014906	0.239524
std	0.036712	0.284061
min	0.000000	0.000000
25%	0.000004	0.013192

	Distribution	Constant
50%	0.000677	0.111982
75%	0.008851	0.386976
max	0.221312	0.994394

Log transformation and square root transformation have given us the best results, as former is a transformation that aims at linearizing the relationship between variables, while latter explicitly helps with right-skewed data. Although, square root transformation leads to uninterpretabiltiy of the results, so if it is important to interpret the coefficients, it is suggested to use log transformation to both dependent and independent variable, as it is the best way to linearize the relationship between sales and distribution, although the majority of items will still have non-linear relationship between sales and distribution; and applying square root transformation is recommended when interpretation of the results is not important, but prediction of the target variable is needed.

## 6. Random Forest: results and challenges

With the next iteration we decided to upgrade the current linear model to the more complex one – random forest. Our approach was to test two strategies: 1) build a model using only two variables – Weighted Distribution and Sales Value; 2) build a model using other categorical variables as well (such as price, volume share, size, etc.) – many different combinations are possible in this case.

After cleaning all the data that doesn't correspond with Gauss-Markov assumptions, we are left with 1852 nonlinear data points and 515 ones that could bear linear relationship. This data is split into 2 datasets. One of the strategies often used in regression tasks is feature scaling. It is explained as a step of data preprocessing which is applied to independent variables or features of data. It basically helps to normalize the data within a particular range.

Table 8 shows the results of a model built on products showing linear relationship between sales and distribution. The data in that model has been scaled beforehand. We can see that R-squared of this model is 0.81, and it is a relatively high value.

**Table 8** Model 1. Two variables. Linear Dataset. Scaler

Mean Absolute Error:	0.7198701966086399
Mean Squared Error:	0.9207590187736855
Root Mean Squared Error:	0.9595618889752163
R Squared Score is:	0.8125084300527634

Table 9 shows the same model as in Table 6, the only difference is that no scaler was applied there. This model actually performed a little bit better, with root mean squared error lower, and R-squared score higher than that of a previous model.

**Table 9** Model 1. Two variables. Linear Dataset. No scaler

Mean Absolute Error:	0.6987572469001453
Mean Squared Error:	0.8368199216368706
Root Mean Squared Error:	0.9147786189220158
R Squared Score is:	0.8283558506997916

In general, the intermediate result doesn't show much difference between the scaled model and the original one. Both models show relatively high R-squared and a low RMSE – which is an indication of a good random forest model. Going further, we calculate metrics for nonlinear dataset as well.

Table 10 shows the random forest model build on data that showed no linear relationship between sales and distribution, with scaling applied to data points. Again, this model performed a little bit better than two previous ones, but it is not a dramatic difference.

**Table 10** Model 1. Two variables. Nonlinear Dataset. Scaler

Mean Absolute Error:	0.6962017803087331
Mean Squared Error:	0.7806550671756768
Root Mean Squared Error:	0.8835468675603331
R Squared Score is:	0.8452660802626787

Table 11 shows the same model as in Table 10, only no scaler was applied to data points here. This model performed as well as the previous model.

**Table 11** Model 1. Two variables. Nonlinear Dataset. No scaler

Mean Absolute Error:	0.7036849052090645
Mean Squared Error:	0.7887535361049883
Root Mean Squared Error:	0.8881179742044344
R Squared Score is:	0.8376094561683674

Feature Scaling still doesn't show any significant difference in results. However, the R-squared in random forest model is much higher than the R-squared for this data point in the linear regression model (all of them were split based on lower than 0.75 R2 score).

For testing the second model we include all variables from the dataset except those that have obvious and proven multicollinearity, and variables which are non-numerical. Thus, the final set of variables excludes TP Value Sales (MLC), TP Volume Sales (MSU), TP Numerical Distribution, Area, and Month (TP).

Table 12 shows the results of the model with all the variables, and scaler applied to the data. It obviously outperformed the previous models, with RMSE almost twice as low as that of previous models, and R-squared score equal to 0.95.

**Table 12** Model 2. Multiple variables. Linear Dataset. Scaler

Mean Absolute Error:	0.3372747906783991
Mean Squared Error:	0.23976408681151315
Root Mean Squared Error:	0.4896571114683347
R Squared Score is:	0.950206254213176

Table 13 shows variables and their importance in the random forest model. Weighted distribution remains the most important feature in this set. However, it is noticeable that the next important features tend to be price and product size, which corresponds to our findings from linear regression analysis, which showed that size was the categorical variable that differed the most across two different data samples. The two variables price\_x and price\_y are a little bit different, in that price\_x shows the price of a product in each month, while price\_y shows the average price of a product throughout the time period that is in our dataset.

**Table 13** Variables and their importance

Variable	Importance
TP Weighted Distribution	0.866051
price_y	0.030375
Size	0.021060
price_x	0.020342
Segment 2	0.009243
Company	0.009159
Form	0.007851
Segment 1	0.006905
Brand	0.005066
cluster	0.001502
Category	0.000000

Table 14 shows the results of the same model, only this time no scaler was applied. The model in this case performed a little bit worse than the previous model, but again, the difference

here is not so dramatic so as to make some conclusions as to importance of a scaler in this particular case.

**Table 14** Model 2. Multiple variables. Linear Dataset. No scaler

Mean Absolute Error:	0.3366202012989508
Mean Squared Error:	0.24576011396073014
Root Mean Squared Error:	0.49574198325412194
R Squared Score is:	0.948355342156652

Table 15 shows the results of a random forest model which was built on non-linear dataset, with the application of scaler. This model performs almost as well as the model in Table 11.

**Table 15** Model 2. Multiple variables. Non-linear Dataset. Scaler

Mean Absolute Error:	0.3525944387059927
Mean Squared Error:	0.24477666218291372
Root Mean Squared Error:	0.49474909012843443
R Squared Score is:	0.9498643787182609

Table 16 shows the results of model which was built on non-linear dataset that was not scaled. This model performs almost the same as the model in Table 12 does, and again, the application of scaler has not affected the results of the model dramatically.

**Table 16** Model 2. Multiple variables. Nonlinear Dataset. No scaler

Mean Absolute Error:	0.33988097508276455
Mean Squared Error:	0.22912280124080722
Root Mean Squared Error:	0.4786677357424534
R Squared Score is:	0.9532476544254329

Table 17 shows the importance of variables that were included in the random forest model. Weighted distribution remains the most important variable, with its importance equal to 0.856, and after it the most important are the price and the size of the product, and this also corresponds with our results from the linear regression analysis, which showed the importance of the size.

**Table 17** Variables and their importance

Variable	Importance
TP Weighted Distribution	0.856223
price_y	0.044068
price_x	0.023856
Size	0.023764

<b>Variable</b>	<b>Importance</b>
Company	0.007138
Segment 1	0.007075
Segment 2	0.007014
Form	0.004304
Brand	0.004274
cluster	0.001141
Category	0.000000

In conclusion, random forest model shows higher R-squared in both linear and nonlinear datasets together with demonstrating very low RMSE.

## **Summary of Chapter 2**

In chapter 2, we described the project and its business value, conducted exploratory data analysis, preprocessed data, and implemented linear regression model and a random forest model.

Exploratory data analysis has shown us that there is high statistically significant correlation between distribution, both weighted and numerical, and sales. We have also calculated prices of products, to later include them in the analysis. We have also assumed that among other possible factors, what influences the type of relationship that product's sales value and distribution demonstrate, is whether its weighted distribution is consistent over the years, with minimum steep increases and falls, and whether it has been on a market for an extended period of time, so as to be a well-established product.

We then applied log transformation to the dependent variable, sales value, because the distribution of this variable was not normal, so we had to transform it for better analysis.

Linear regression model has helped us identify products that demonstrate linear relationship between sales and distribution, based on the adjusted R-squared value (we have settled on adjusted R-squared higher than 0.75 to identify those products). We then analyzed each categorical variable to identify what differentiates products with linear relationship between sales and distribution from those without one. Our comparison has shown that size is the variable that differs the most between two samples. It could be explained by the fact that big packages usually cost proportionally less than small ones do, as in one unit of the product in a large package costs less than one unit of the same product in a small package. Large sized products usually take longer to be finished, so they are bought less than small sized products of the same category. Higher

distribution makes sure that more products are bought, but only up to a certain point, where people already have enough of that product, because they bought in a large size, so there is a longer period of time between purchases of this product, and this explains the non-linear relationship between sales and distribution, and how size of the product affects this relationship. Unfortunately, we cannot say if the size in our case is a large or a small one, so we can only assume.

We also applied different kind of transformations in different ways to see which one would give us the highest number of products with linear relationship between sales and distribution, and our analysis has shown us that applying log transformation to both dependent and independent variables yields the best optimal results for when interpretation of coefficients is important, while square root transformation is useful in case of prediction of the target variable.

Afterwards, we analyzed our dataset using random forest model. We have divided the dataset into two different samples, based on the linear regression analysis. Random forest regression models are used for non-linear datasets, and considering that results of both samples were quite high in terms of accuracy (R-squared), it can mean that the relationship between sales and distribution is better described as non-linear, even though the linear regression analysis has shown us that some products do demonstrate linear relationship between sales and distribution.

Our research shows that even through on some of the data simple linear regression model performs well, more complex random forest model outperforms at both this data and data that had no result in linear models. Also, when using linear regression, it might be more suitable to apply log transformation to both target and predictor variables, as this transformation linearizes the relationship between variables.

We have also used random forest regression model to include more categorical variables in the analysis, and we have identified that weighted distribution makes up for around 85% of sales value, with price and the size of the product making up for 4% and 2% respectively. Our major limitation is that we have no data on other elements of marketing mix, except for price, which was calculated bases on available data, so we cannot analyze the effect that all those elements have on each other, and how they affect sales together.

## CONCLUSION

We have completed the research for Procter and Gamble, and here we will talk about our main findings, managerial implications of those findings, limitations of our research and further research.

### **Main findings**

The main findings of our research are the following:

- We have analyzed the existing literature, and realized that there is a research gap in regards to the relationship between sales and distribution. It is assumed that there is indeed a significant relationship, but existing research doesn't often focus on distribution's effects on sales, and what research does focus on this relationship, there is not a lot of it;
- Our literature analysis has shown us that the most suitable option to analyze relationships between variables is a regression model, which shows numeric effects of independent variables on target variables, and its derivatives, like random forest;
- Analysis of the scatter plots of a number of items demonstrates linear trends between sales and distribution, but majorly only for items that have been on the market for more than 12 months, and items whose weighted distribution hasn't gone through sharp increases and decreases in this period, so we believe that items have to be well established in the market to develop a linear relationship between sales and distribution;
- 25% of items in our data show linear relationship between sales and distribution, and while some of the do not follow the assumption made previously, a number of them do, and vice versa, some items that do follow that assumption, did not have an R-squared score  $> 0.75$  (which was arbitrary chosen by us to define linearity);
- Size of the product is the categorical variable that differs the most between products with linear and products with non-linear relationship between sales and distribution, with brand and form following;
- Applying log transformations to both dependent and independent variables gives the best results and interpretability, compared to other transformations (meaning that the highest number of items show a relatively high R-squared and therefore a linear relationship), while square root transformation gives higher results at the expense of interpretability, but it can be used when prediction is of higher importance;



- The mean value of the coefficient of weighted distribution in linear relationship is 1.25, meaning that for those items who demonstrate linear relationship between sales and distribution, the sales value grows by 1.25% in average when distribution increases by 1%;
- Applying random forest regression gives high results (R-squared  $>0.8$  for models with only distribution as independent variable, and R-squared  $>0.9$  for models with all categorical variables also included, and a relatively low RMSE). There is no big difference in results between applying standardization and not applying it;
- RF gives 85% of importance to weighted distribution, and the second in place is the price of the product. The main challenge with these results is that it will be difficult to extrapolate them on new data, as the RF model can overfit on training data, but trimming the trees so they don't go too deep can help with this challenge.

### **Managerial implications of our findings**

As for the managerial implications of our results and findings, we can divide them into more technical recommendations, and business-oriented recommendations.

First of all, we will discuss technical recommendations, specifically the models that we have used, and what the results of their application mean for the company. We have found that applying log transformation to both dependent and independent variables in linear regression yields best results for analysis of the relationship between sales and distribution, so if Procter and Gamble decide that they need the highest possible results, they should apply log transformations. High results were also given by applying square root transformation, and while this transformation is virtually uninterpretable, it can be used when all that is needed is the prediction of sales value by weighted distribution. Procter and Gamble can use one of those transformation depending on the goal of the analysis: log transformation for measuring the effect of weighted distribution on the sales, and square root transformation – for predicting the sales value based on weighted distribution. Random forest regression model also gives high results in the analysis of the relationship between sales and distribution, but it does so at the expense of extrapolation, as it can be difficult to use the trained model on absolutely new data. So, using a more complex model can give the company a better understanding of the effect of distribution on sales, and with some adjustments, Random Forest regression can be used on absolutely new data.

Now, we will discuss the more business-oriented recommendations. As we have found, the size is the product characteristics that can possibly affect the form of the relationship between sales and distribution, so it has to be accounted for when either predicting the sales value, or measuring the effect of distribution on sales. Also, we have found that weighted distribution highly affects

the sales value, so it must be included when forecasting sales of different products. It all leads to increased efficiency of planning sales and distribution, as we the company now knows the mean effect of distribution growth on sales. Finally, we find that items that have had the chance to become established on the market, as in their weighted distribution has been relatively stable and they have been on the market for at least a year, are much more likely to have a linear relationship between sales and distribution. So, if the company aims to have this linear relationship between sales and distribution, first it should make sure that the products have a chance to become well known at the market, and be distributed in a stable manner, before there is possibility to develop a linear relationship between those variables, depending on product characteristics.

To sum up, from the technical point of view, we suggest applying two types of transformations, depending on the goals of the analysis. From the business point of view, we suggest including some characteristics of the product, such as size, when forecasting sales based on distribution, but first letting products become established on the market before doing any analysis or forecast at all. Analysts from Procter and Gamble now will be able to spend less time on adjusting models when performing tasks on analyzing distribution, which in turn will let them focus on other tasks, thus reducing costs.

### **Limitations and further research**

Our most major limitation is that we do not have any data on promotion, and our results are mostly relevant when promotion is not considered in the sales. Also, we have only worked with one product category, and we cannot say how our finding will be relevant for other product categories. We also cannot give concrete suggestions based on, for example, the fact that the size differs the most between two subsets of data, as we do not know what size it is exactly, but the company does have this information, so it is only a limitation on our side.

For further research, we would suggest looking at the interplay of all the elements of marketing mix, to analyze how they interact with each other, what effect they have on each other particularly, and on the sales value in general. Different product categories should also be included in the further research, to define for what categories our findings are relevant, and what categories are different.

To conclude, we have answered all the questions that were posed before us for this research. We have stayed in touch with the representative of Procter & Gamble throughout this whole process, and received and applied suggestions to our work, so we can say that they were a major part of our research. We have also presented our results to them, and they accepted those results and remarked that they will be using the results in their future work.

## REFERENCES

1. Abdelnour, A., Devignes, J-C., Randery, T., Rogers, J. (2020). COVID-19 crisis: How distributors can emerge stronger than before. Retrieved from: <https://www.mckinsey.com/industries/advanced-electronics/our-insights/covid-19-crisis-how-distributors-can-emerge-stronger-than-before>
2. ACNielsen, Karolefski, J., Heller, A. (2006) Consumer-centric category management: how to increase profits by managing categories based on consumer needs. John Wiley & Sons, Inc, Hoboken
3. Agarwal, R. (2016). Top 6 Elements of Physical Distribution Channels. Retrieved from: <https://www.yourarticlelibrary.com/marketing/distribution-channels/top-6-elements-of-physical-distribution-channels-with-diagram/48313>
4. Ataman, M. B., Van Heerde, H. J., Mela, C. F. (2010). The long-term effect of marketing strategy on brand sales. *Journal of Marketing Research*, 47(5). 866-882.
5. Bahadir, S. C., Bharadwaj, S. (2015). Marketing mix and brand sales in global markets: Examining the contingent role of country-market characteristics. *Journal of International Business Studies*, 46(5)
6. Bucklin, R. E., Siddarth, S., Silva-Risso, J. M. (2008). Distribution intensity and new car choice. *Journal of Marketing Research*, 45(4). 473-486.
7. CNews. (2020) Machine Learning-based demand forecast system is implemented in Perekrestok. Retrieved from: [https://www.cnews.ru/news/line/2020-06-08\\_v\\_seti\\_perekrestok\\_vnedrena](https://www.cnews.ru/news/line/2020-06-08_v_seti_perekrestok_vnedrena)
8. Core RFID. RFID Systems in Distribution | Managing Shipments. Retrieved from: <https://www.corerfid.com/rfid-applications/rfid-in-distribution/>
9. Duening, T. N., Hisrich, R. D., Lechter, M. A. (2010). Going to Market and the Marketing Plan. In: *Technology Entrepreneurship*. Academic Press.
10. Dujak, D., Segetlija, Z., Mesarić, J. (2016). Efficient Demand Management in Retailing Through Category Management. In: *Efficiency in Sustainable Supply Chain*. Springer
11. Dynamic Yield Russia. (2020). Customization cases: how Perekrestok online shop uses product recommendations to increase online sales. Retrieved from: <https://vc.ru/trade/143587-keysy-personalizacii-kak-internet-magazin-perekrestok-ispolzuet-tovarnye-rekomendacii-dlya-rosta-onlayn-prodazh>
12. ECR Europe (2014) Glossary. Retrieved from: <http://www.ecr-europe.org/toolbox/glossary>
13. Fowler, K. R. (2015) Logistics, Distribution, and Support. In: *Developing and Managing Embedded Systems and Products*. Newnes

14. Friberg, R., Sanctuary, M. (2017). The Effect of Retail Distribution on Sales of Alcoholic Beverages. *Marketing Science*, 36(4).
15. Gallo, A. (2015). A Refresher on Regression Analysis. Retrieved from: <https://hbr.org/2015/11/a-refresher-on-regression-analysis>
16. Gelman, A., Hill, J. (2007). *Data analysis using regression and multilevel hierarchical models*. New York, NY: Cambridge University Press.
17. Google Cloud. (2020). How Procter & Gamble uses Google Cloud to improve consumer experience. Retrieved from: <https://cloud.google.com/blog/products/data-analytics/how-procter-gamble-improves-consumer-experiences-with-data>
18. Hanssens, D. M., Parsons, L. J., Schultz, R.L. (2001). *Market Response Models: Econometric and Time Series Analysis*, 2nd edition. Kluwer Academic Publishers.
19. Hibbard, J., Sadeh, F. (2017) Performance Impact of Distribution Expansion: A Review and Research Agenda. In: *Handbook of research on distribution channels*. Edward Elgar Publishing
20. Iacobucci, D. (2013). *MM4*. Cengage Learning.
21. IBM Cloud Education. (2020). Machine Learning. Retrieved from: <https://www.ibm.com/cloud/learn/machine-learning>
22. Jantan, M., Nelson, N., Ong, Y. (2003). Viability of e-commerce as an alternative distribution channel. *Logistics Information Management*, 16, 427-439.
23. Kotler P., Armstrong G. (2006). *Principles of marketing*, (11th Ed.). Upper Saddle River: New Jersey: Prentice-Hall.
24. Kushnir, E. (2019). How machine learning increases sales. Retrieved from: <https://mcs.mail.ru/blog/kak-machine-learning-povyshaet-prodazhi>
25. MacKenzie, I., Meyer, C., Noble, S. (2013). How retailers can keep up with consumers. Retrieved from: <https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers>
26. Manoj, K. (2020). How FMCG Brands are Deploying Technology to Streamline their Distribution? Retrieved from: <https://www.fieldassist.in/blog/fmcg-distribution-network/>
27. McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D.J., Barton, D. (2012). *Big Data: The Management Revolution*. Retrieved from: <https://hbr.org/2012/10/big-data-the-management-revolution>
28. Olariu, I. (2009). *FMCG companies specific distribution channels*. Studies and Scientific Researches - Economic Edition.
29. Pek, J., Wong, A., Wong, O. (2017). Data Transformations for Inference with Linear Regression: Clarifications and Recommendations. *Practical Assessment*, 22(9)

30. Reibstein, D. J., Farris, P. W. (1995). Market Share and Distribution: A Generalization, a Speculation, and Some Implications. *Marketing Science*, 14(3). 190-202
31. Retail.ru. (2017). How M-Video uses consumer data for promotions. Retrieved from: <https://www.retail.ru/articles/kak-m-video-ispolzuet-dannye-pokupatelya-dlya-promo-meropriyatiy/>
32. Retail.ru. (2018). X5 will apply machine learning in marketing. Retrieved from: <https://www.retail.ru/news/x5-primenit-mashinnoe-obuchenie-v-marketinge/>
33. Salmon, K. Seven Facets of Modern Category Management
34. Segetlija, Z., Mesarić, J., Dujak, D. (2011). Importance of Distribution Channels - Marketing Channels - for National Economy
35. Ursin, C. (2004). Facing facts: what category management—or fact-based selling—can do for you. In: *Beverage Dynamics*. Bev-AI Communications, Inc.
36. Wang, X., Ryoo, J.H., Bendle, N., Kopalle, P.K. (2020). The role of machine learning analytics and metrics in retailing research. *Journal of Retailing*, 96(4)
37. Wilbur, K. C. and Farris, P. W. (2014). Distribution and market share. *Journal of Retailing*, 90(2), 154-167.
38. Wyman, O. (2012). Making category management work. Retrieved from: [https://www.oliverwyman.com/content/dam/oliver-wyman/global/en/2014/jul/2012\\_OW\\_Category%20Management\\_ENG.pdf](https://www.oliverwyman.com/content/dam/oliver-wyman/global/en/2014/jul/2012_OW_Category%20Management_ENG.pdf)

## APPENDICES

```
In [2]: data = pd.read_csv('TP_Data_to_share.csv')
```

```
In [3]: master = pd.read_excel('Master_data_to_share.xlsx')
```

### *Appendix 1: Data Preprocessing*

```
In [17]: data = data.dropna()
```

```
In [18]: weighted = data['TP Weighted Distribution']
numerical = data['TP Numerical Distribution']
value_sales = data['TP Value Sales (MLC)']
volume_sales = data['TP Volume Sales (MSU)']
```

```
In [19]: coefPears, pvPears=stats.pearsonr(weighted, value_sales) #correlation test
print("Pearson's correlation = ", round(coefPears,3))
print("Pearson's p-value = ", round(pvPears,3))

Pearson's correlation = 0.725
Pearson's p-value = 0.0
```

```
In [20]: coefPears, pvPears=stats.pearsonr(numerical, value_sales) #correlation test
print("Pearson's correlation = ", round(coefPears,3))
print("Pearson's p-value = ", round(pvPears,3))

Pearson's correlation = 0.661
Pearson's p-value = 0.0
```

```
In [21]: coefPears, pvPears=stats.pearsonr(volume_sales, weighted) #correlation test
print("Pearson's correlation = ", round(coefPears,3))
print("Pearson's p-value = ", round(pvPears,3))

Pearson's correlation = 0.69
Pearson's p-value = 0.0
```

```
In [22]: coefPears, pvPears=stats.pearsonr(volume_sales, numerical) #correlation test
print("Pearson's correlation = ", round(coefPears,3))
print("Pearson's p-value = ", round(pvPears,3))

Pearson's correlation = 0.619
Pearson's p-value = 0.0
```

### *Appendix 2: Preliminary correlation tests*

```
In [7]: data['price'] = data['TP Value Sales (MLC)'] / data['TP Volume Sales (MSU)']
```

```
In [9]: #data['price'].isnull()
data['price'] = np.nan_to_num(data['price'])
```

```
In [10]: data['price'].isnull().sum()
```

```
Out[10]: 0
```

### *Appendix 3: Introducing "price" as a new variable*

```
In [37]: data = data.drop(data[(data['TP Value Sales (MLC)'] > 0) & (data['TP Volume Sales (MSU)'] == 0)].index)
```

### *Appendix 4: Running logical assumptions*

```
In [42]: prices = data.groupby('Item').mean()
```

### *Appendix 5: Adding mean price to each dataset item*

```
In [296]: import statsmodels.api as sm
from statsmodels.compat import lzip
import statsmodels.stats.api as sms
import math as m
from sklearn.preprocessing import StandardScaler
Heteroskedasticity = []
NormalityRes = []
MeanRes = []
Autocorr = []

r2 = {}
params = {}
tvalues = {}
pvalues = {}
s = 0

for i in items:
    item = combined_data[combined_data["Item"]==i]
    X = item['TP Weighted Distribution']
    y = np.log(item['TP Value Sales (MLC)'])
    X = sm.add_constant(X)
    model = sm.OLS(y, X)
    model = model.fit()
    r2_i = model.rsquared_adj
    params_i = model.params
    tvalues_i = model.tvalues
    pvalues_i = model.pvalues
    residmean = model.resid.mean()

    #Breusch Pagan
    heterosk = sms.het_breuschpagan(model.resid, model.model.exog)

    #Jarque-Bera
    residd = sms.jarque_bera(model.resid)
```

```
#perform Durbin-Watson test
autocorr = sms.durbin_watson(model.resid)
#print(autocorr)

if not m.isnan(r2_i):
    s += 1
    if heterosk[1] > 0.05 and residd[1] > 0.05 and autocorr > 0.5:
        MeanRes.append(round(residmean, 2))
        Heteroskedasticity.append(heterosk[1])
        NormalityRes.append(residd[1])
        Autocorr.append(autocorr)
        n = {i: r2_i}
        r2.setdefault(i, r2_i)
        params.setdefault(i, params_i)
        tvalues.setdefault(i, tvalues_i)
        pvalues.setdefault(i, pvalues_i)
```

### Appendix 6: Linear regression model with Gauss-Markov assumptions check

```
In [58]: pvalues=pd.DataFrame(list(pvalues.items()),columns = ['Item','P-values'])
tvalues = pd.DataFrame(list(tvalues.items()),columns = ['Item','T-values'])
```

```
In [63]: linear = r_sq[r_sq['R2'] >= 0.75]
nonlinear = r_sq[r_sq['R2'] < 0.75]
```

```
In [60]: r_sq = pd.DataFrame(list(r2.items()),columns = ['Item','R2'])
r_sq['R2'].isna().sum()
```

Out[60]: 445

```
In [65]: linear = linear.merge(master, how='left', on='Item')
nonlinear = nonlinear.merge(master, how='left', on='Item')

nonlinear.replace([np.inf, -np.inf], np.nan, inplace=True)
nonlinear=nonlinear.dropna()
```

```
In [67]: linear = linear.merge(paramets, how='left', on='Item')
linear = linear.merge(tvalues, how='left', on='Item')
linear = linear.merge(pvalues, how='left', on='Item')
nonlinear=nonlinear.merge(paramets, how='left', on='Item')
```

## Appendix 7: Splitting the dataset into two – linear and nonlinear

```
In [72]: linear
```

```
Out[72]:
```

	Item	R2	Category	Company	Brand	Form	Segment 1	Segment 2	Size	price	cluster	Parameters	T-values	P-values
0	ITEM_3574	0.916737	1	1	1	2	3	4	41	679.433333	0	const 1.541402 TP Weigh...	const 5.217382 TP Weight...	const 0.120557 TP Weight...
1	ITEM_2509	0.840776	1	4	9	2	3	4	76	1683.554422	2	const 3.511139 TP Weight...	const 22.538586 TP Weight...	const 0.000023 TP Weight...
2	ITEM_2888	0.847767	1	1	5	2	3	4	51	1733.365623	2	const 5.895571 TP Weight...	const 21.770609 TP Weight...	const 0.000026 TP Weight...
3	ITEM_2274	0.774417	1	8	0	2	3	5	149	944.731891	0	const 5.758003 TP Weight...	const 13.828169 TP Weight...	const 7.227690e-07 TP We...

## Appendix 8: Linear dataset snapshot

```
In [56]: r2 = dict((k, v) for k, v in r2.items() if v >=-1 and v<1)
print(len(r2))
...
```

```
In [70]: linear_relation = dict((k, v) for k, v in r2.items() if v >= 0.75)
print(len(linear_relation))
...
```

```
In [71]: nonlinear_relation = dict((k, v) for k, v in r2.items() if v < 0.75)
print(len(nonlinear_relation))
...
```

```
In [73]: linear_rel = pd.DataFrame(linear_relation.items(), columns=['Item', 'R2'])
nonlinear_rel = pd.DataFrame(nonlinear_relation.items(), columns=['Item', 'R2'])
linear_rel[:10]
```

```
In [74]: combined_linear = pd.merge(linear_rel, master, how='left', left_on=['Item'], right_on=['Item'], validate = 'm:1')
combined_nonlinear = pd.merge(nonlinear_rel, master, how='left', left_on=['Item'], right_on=['Item'], validate = 'm:1')
```

```
In [75]: combined_nonlinear
```

```
Out[75]:
```

	Item	R2	Category	Company	Brand	Form	Segment 1	Segment 2	Size	price	cluster
0	ITEM_192	0.424976	1	1	1	1	2	5	205	585.429457	0
1	ITEM_3933	0.712155	1	4	2	2	2	2	66	1557.933162	2
2	ITEM_2309	0.253249	1	8	3	2	3	5	32	2156.144739	3

## Appendix 9: Preparing datasets for Random Forest model

```
In [76]: from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.metrics import r2_score
```

```
In [78]: linear_rf = pd.merge(combined_linear, combined_data, how='left', left_on=['Item'], right_on=['Item'],
                             validate = 'm:m')
nonlinear_rf = pd.merge(combined_nonlinear, combined_data, how='left', left_on=['Item'], right_on=['Item'],
                          validate = 'm:m')
```

```
In [65]: X = nonlinear_rf.iloc[:,1:18].drop(['TP Value Sales (MLC)', 'TP Volume Sales (MSU)', 'TP Numerical Distribution',
                                           'Area', 'Month (TP)'], axis=1).values
y = np.log(nonlinear_rf.iloc[:,15].values)
```

## Appendix 10: Choosing variables for the Random Forest model



```
In [288]: X = X.reshape(-1, 1)
X
```

```
Out[288]: array([[ 0.05],
 [ 0.05],
 [ 0.07],
 ...,
 [41.3 ],
 [41.2 ],
 [40.29]])
```

```
In [67]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```
In [68]: # Feature Scaling
from sklearn.preprocessing import StandardScaler

sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

```
In [69]: from sklearn.ensemble import RandomForestRegressor

regressor = RandomForestRegressor(n_estimators=2000, random_state=50)
regressor.fit(X_train, y_train)
y_pred = regressor.predict(X_test)
```

```
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
print('R Squared Score is:', r2_score(y_test, y_pred))
```

```
Mean Absolute Error: 0.3491165190070628
Mean Squared Error: 0.2308443970344131
Root Mean Squared Error: 0.4804626905748386
R Squared Score is: 0.9519888337565839
```

```
In [72]: pd.DataFrame({'Variable':linear_rf.iloc[:,1:18].drop(['TP Value Sales (MLC)', 'TP Volume Sales (MSU)',
 'TP Numerical Distribution', 'Area', 'Month (TP)'], axis=1).columns,
 'Importance':regressor.feature_importances_}).sort_values('Importance', ascending=False)
```

Out[72]:

	Variable	Importance
10	TP Weighted Distribution	0.856377
11	price_y	0.044033
8	price_x	0.024591
7	Size	0.023698
0	R2	0.021176
5	Segment 1	0.007701
2	Company	0.006958
6	Segment 2	0.006441
3	Brand	0.004362
4	Form	0.003538
9	cluster	0.001127
1	Category	0.000000

*Appendix 11: Example of Random Forest result for nonlinear dataset*