Санкт-Петербургский государственный университет
Кафедра математической теории игр и статистических решений

Ковалев Денис Александрович

Выпускная квалификационная работа бакалавра

Анализ факторов, влияющих на уровень преступности в Санкт-Петербурге

Направление 020302

Фундаментальная информатика и

информационные технологии

Научный руководитель,
кандидат физ.-мат. наук,
доцент
Панкратова Я. Б.

Санкт-Петербург

2021

Содержание

Введение

Полное и точное понимание состояния преступности, эффективности мер общественного контроля, деятельности органов уголовного правосудия, а также разработка государственной политики по борьбе с преступностью возможны только при наиболее полном статистическом анализе списка зарегистрированных преступлений. Руководители правоохранительных органов традиционно оценивают эффективность работы правоохранительных органов по формальным показателям, основным из которых остается раскрытие преступлений.

Основная функция статистики в сфере правосудия - это первичный учет преступлений и фиксация заявлений граждан для организации последующей работы правоохранительных органов. Статистические исследования преступности имеют особое значение для криминологии, уголовного права, уголовного процесса, криминалистики, исправительно-трудового и административного права. Ни одна из юридических наук не может обходиться, не говоря уже о развитии и формировании, без использования данных и статистики.

В криминологии статистические исследования позволяют выявлять основные показатели состояния преступности, ее причины и условия, способствующие совершению преступлений. Статистическое исследование включает наблюдение, сводку, группировку и анализ. Полученный в результате разработки статистический материал часто требует визуального изображения. В настоящее время в России в основные задачи правовой статистики входит не только информирование об истинном положении дел в сфере борьбы с преступностью в стране, но и выявление взаимосвязи преступности и правонарушений с политическим, социально-

экономическим и моральным положением в стране.

Интерпретация полученных результатов позволяет обосновывать принятие управленческих решений на всех уровнях власти, поэтому задача выявления и моделирования основных тенденций в изменении обобщающих показателей преступности не теряет своей актуальности.

Постановка задачи

В данной работе основное внимание уделено исследованию основных факторов, которые влияют на уровень преступности в Санкт-Петербурге. Произведен анализ взаимосвязи количества преступлений в г. Санкт-Петербург от социальных и экономических показателей. Для этого рассмотрена статистика и динамика преступлений в России в 2010 - 2018 годах и проведен корреляционно-регрессионный анализ факторов.

В работе также проводится анализ временного ряда и моделирование численности зарегистрированных преступлений в Санкт-Петербурге в период 2010 - 2021 гг. После этого, на основе данных, полученных с сайта прокуратуры г. Санкт-Петербург, строится аналитическая панель с интерактивными графиками. Для выполнения поставленных целей необходимо решить ряд задач:

- Построить модель линейной регрессии
- Дать интерпретацию построенной модели и её коэффициентам, оценить ее качество
- Построить модель SARIMA для моделирования временного ряда
- Спрогнозировать количество преступлений на несколько шагов вперед
- Подготовить данные для аналитической панели
- Связать множество интерактивных компонентов в единую наглядную систему

Обзор литературы

Для написания данной работы была использована научная литература, а также публикации из научных изданий и интернет-источники.

Основные положения о факторах, влияющих на преступность и причинах совершения преступлений были взяты из книги «Причины преступности», автора Антонян Ю.М. [3] Дополнительным источником являлась статья «Зависимость уровня преступности от экономических и социальных факторов в регионах РФ» автора Барышниковой А.В. [6]

Основными источниками для изучения регрессионного анализа были книги «Introductory Econometrics for Finance» автора Chris B. [1] и курс лекций по машинному обучению автора Воронцова К. В. [5]. Для изучения множественной регрессии (многофакторного анализа) был использован интернет-ресурс Studme («Многофакторный регрессионный анализ») [11].

Для изучения существующих вариантов построения регрессионных моделей для анализа уровня преступности были изучены статьи «Статистическое изучение уровня преступности в Российской Федерации» автора Красиковой Е.М. [7] и «Исследование основных факторов, влияющих на уровень преступности в России» автора Гусевой М.В. [8]

Основным источником для изучения анализа временных рядов и моделей для их прогнозирования была книга «Introduction to Time Series Analysis and Forecasting» авторов Montgomery, D.C., Jennings C.L., Kulahci M. [2]

Для изучения существующих работ по анализу временных рядов на тему преступности были изучены статьи «Анализ и прогнозирование динамики зарегистрированных преступлений в России на основе времен-

ного ряда 1991–2019 гг.» авторов Шумилина О.В. и Мячина Н.В. [9] и «Методика моделирования и прогнозирования преступности в Российской Федерации» авторов Богдановой М.В., Паршинцевой Л.С. и Квачко В.Ю. [10]

Глава 1. Основные понятия и определения

1.1. Определения

Определение 1.1. Регрессионная модель

$$y = f(x, b) + \varepsilon,\ E(\varepsilon) = 0,$$

где b – параметры модели, ε – случайная ошибка модели; называется *линейной регрессией*, если функция регрессии $f(x, b)$ имеет вид

$$f(x, b) = b_0 + b_1 x_1 + b_2 x_2 + .. + b_k x_k,$$

где где b_j – параметры (коэффициенты) регрессии, x_j – регрессоры (факторы модели), k – количество факторов модели

Коэффициенты линейной регрессии показывают скорость изменения зависимой переменной по данному фактору, при фиксированных остальных факторах. Таким образом, линейная регрессия является методом восстановления зависимости объясняемой переменной y от другой или нескольких других переменных-регрессоров x с линейной функцией зависимости.

Определение 1.2. Временной ряд – последовательность значений показателя или признака, упорядоченная в хронологическом порядке, т.е. в порядке возрастания временного параметра. Отдельные наблюдения временного ряда называются *уровнями* этого ряда.

Каждый временной ряд содержит два элемента: значения времени и соответствующие им значения уровней ряда. В отличие от пространственных данных, уровни временного ряда, как правило, не являются статистически независимыми и одинаково распределенными.

Определение 1.3 Стационарный временной ряд – временной ряд, эле-

менты которого являются случайными величинами с постоянным математическим ожиданием и постоянной дисперсией.

Определение 1.4 ARIMA – (англ. autoregressive integrated moving average) — модель и методология анализа временных рядов. Модель $ARIMA(p, d, q)$ для нестационарного временного ряда X_t имеет вид:

$$\Delta^d X_t = c + \sum_{i=1}^{p} a_i \Delta^d X_{t-i} + \sum_{j=1}^{q} b_j \varepsilon_{t-j} + \varepsilon_t,$$

где ε_t - стационарный временной ряд; c, a_i, b_j - параметры модели; Δ^d - оператор разности временного ряда порядка d

Определение 1.5 Дашборд – это современный формат сбора и визуального представления массивов данных. Это аналитическая панель с понятным интерфейсом для интерактивного взаимодействия с огромным количеством постоянно изменяющихся показателей.

Глава 2. Построение регрессионной модели, анализ показателей

2.1. Модель линейной регрессии

В этой главе проведем корреляционно-регрессионный анализ факторов, влияющих на уровень преступности в Санкт-Петербурге за 2010–2018 гг. Для этого построим модель линейной регрессии, то есть модель зависимости целевой переменной от одной или нескольких других независимых переменных (факторов) с линейной функцией зависимости, используя средства языка программирования Python.

2.2. Исследуемые факторы

Поведение преступника представляет собой сложную и динамичную систему различных проявлений человеческой активности и пассивности, которая определяется объективными и субъективными факторами до, во время и после совершения преступления. Эти факторы отражаются в объективной реальности, в образе жизни и в сознании человека в виде материальных и идеальных следов.

Преступление – это социально опасный и аморальный акт, который может быть основан на различных экономических, политических, социальных и культурных факторах. В результате достаточно сложно сказать, какой фактор больше всего влияет на уровень преступности.

Выявим влияние на изменение общего числа преступлений, совершенных в Санкт-Петербурге за 2010–2018 гг. следующих факторов:

- Доход населения

- Количество беспризорных
- Миграционный прирост
- Безработица
- Алкоголизация населения
- Наркотизация населения

Введем переменные:

1. $MONEY$ – Соотношение среднедушевых денежных доходов населения с величиной прожиточного минимума (процент)
2. $HOMELESS$ – Количество выявленных беспризорных и безнадзорных несовершеннолетних (человек)
3. $MIGRATION$ – Коэффициент миграционного прироста (на 10 тыс. человек)
4. $UNEMPL$ – Численность безработных в возрасте 15–72 лет в Санкт-Петербурге (тыс. человек)
5. $BEER\_SOLD$ – Продажа пива и пивных напитков населению в Российской Федерации в натуральном выражении (млн дкл)
6. $VODKA\_SOLD$ – Продажа водки и ликероводочных изделий населению в Российской Федерации в натуральном выражении (млн дкл)
7. $INCOME$ – Динамика среднедушевых доходов населения по Российской Федерации (рублей в месяц)

8. $POVERTY$ – Доля населения с денежными доходами ниже величины прожиточного минимума, установленной в г. Санкт-Петербург (значение показателя за год)

9. $ALCO\_MED$ – Численность больных алкоголизмом и алкогольными психозами, состоящих на учете в лечебно-профилактических организациях (тыс. человек)

10. $NARCO\_MED$ – Численность больных наркоманией, состоящих на учете в лечебно-профилактических организациях (тыс. человек)

Зависимую переменную назовем $CRIME\_TOTAL$. Она обозначает общее количество преступлений, совершенных в Санкт-Петербурге.

2.3. Источники данных

Статистическую информацию получим из открытых государственных источников.

Данные для параметров $MONEY$, $HOMELESS$, $MIGRATION$, $POVERTY$ возьмем с сайта федеральной службы государственной статистики Росстат. [13] Международная экспертиза признала статистические данные Федеральной службы государственной статистики надежными.

Данные для параметров $UNEMPL$, $INCOME$, $ALCO\_MED$, $NARCO\_MED$, $BEER\_SOLD$, $VODKA\_SOLD$, $CRIME\_TOTAL$ возьмем с сайта ЕМИСС (единой межведомственной информационно-статистической системы). [12] ЕМИСС содержит официальную статистическую информацию, формируемую субъектами официального статистического учета в рамках Федерального плана статистических работ.

2.4. Устранение мультиколлинеарности

В регрессионном анализе мультиколлинеарность означает наличие линейной зависимости между объясняющими переменными регрессионной модели. Другими словами, это корреляция независимых переменных, которая затрудняет оценку, анализ и интерпретацию общего результата.

Из-за мультиколлинеарности факторов математическая модель регрессии будет содержать избыточные переменные, в результате:

- осложняется интерпретация параметров множественной регрессии как величин действия факторов, и параметры регрессии теряют смысл
- оценки параметров становятся ненадежны из-за больших стандартных ошибок, которые меняются с изменением объема наблюдений, что делает модель регрессии непригодной для прогнозирования

Для обнаружения мультиколлинеарности факторов проанализируем непосредственно корреляционную матрицу факторов. Наличие больших по модулю (выше 0.7-0.8) значений коэффициентов парной корреляции свидетельствует о возможных проблемах с качеством получаемых оценок.

Построим корреляционную матрицу (Рисунок 1).

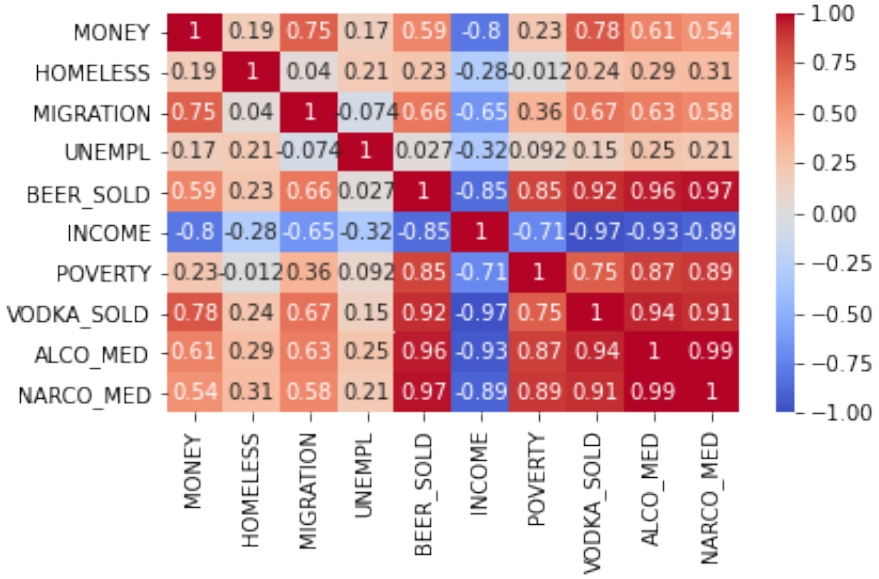

Рис. 1: Корреляционная матрица всех изначальных переменных

Видно, что переменные $BEER\_SOLD, VODKA\_SOLD, INCOME, POVERTY, ALCO\_MED, NARCO\_MED$ сильно коррелируют друг с другом (Рисунок 2).

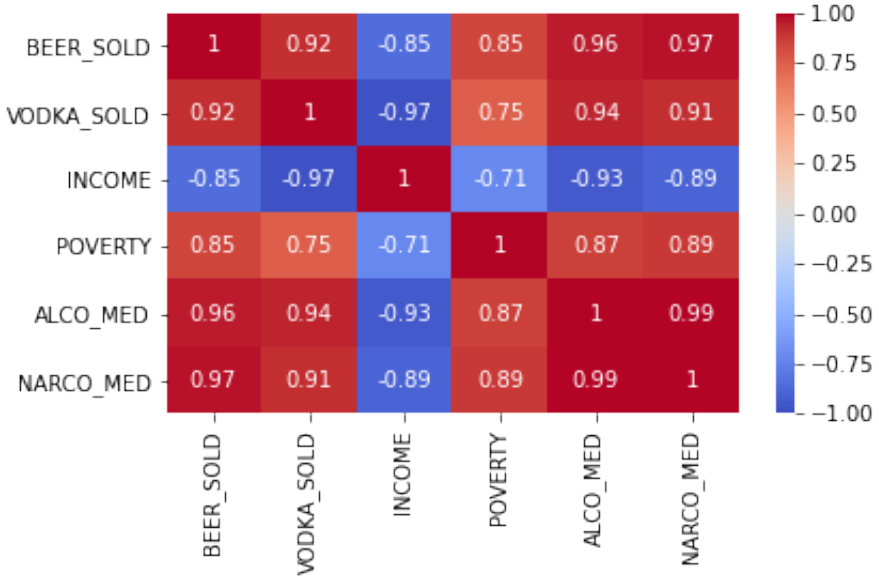

Рис. 2: Сильно коррелированные признаки

Включим в модель фактор $INCOME$, а остальные коррелирующие

с ним факторы удалим (Рисунок 3).

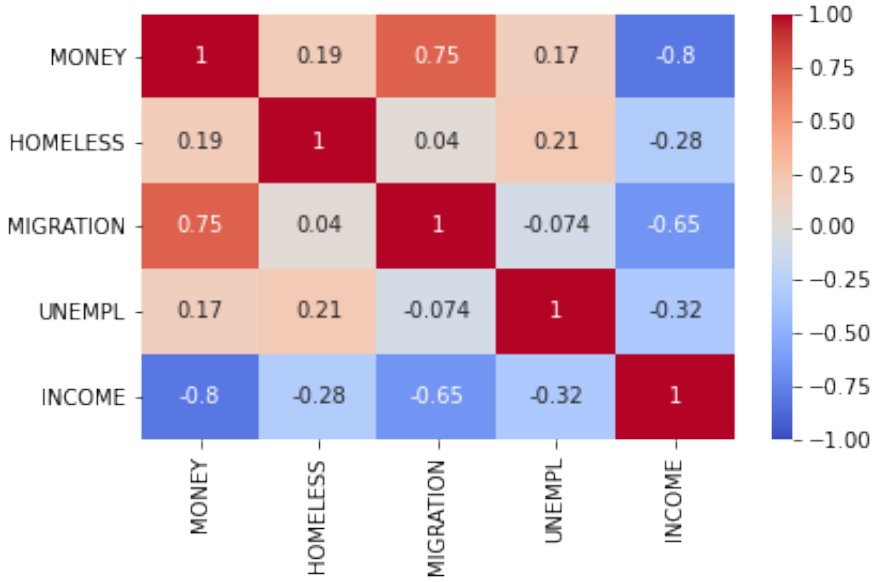

Рис. 3: Корреляционная матрица после удаления 5 признаков

Параметр $MONEY$ сильно коррелирует с параметрами $MIGRATION$ и $INCOME$. Исключим фактор $MONEY$ из модели (Рисунок 4):

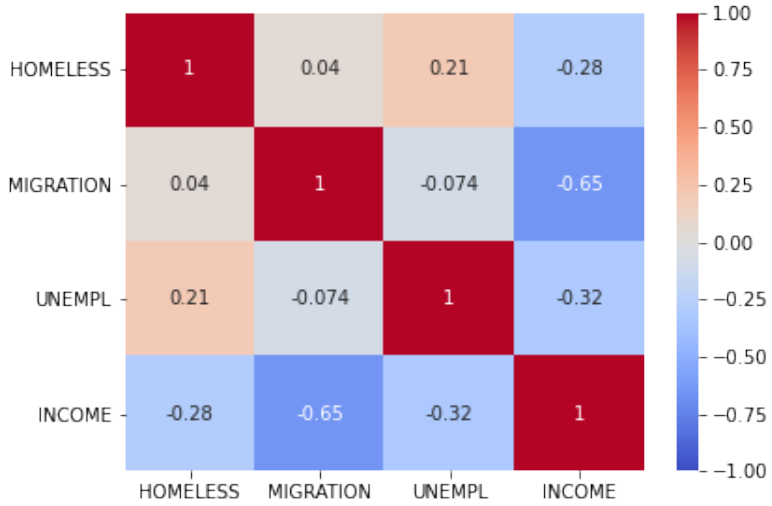

Рис. 4: Итоговая корреляционная матрица

Таким образом, удалось избавиться от мультиколлинеарности факторов.

2.5. Построение модели регрессии с помощью Python

Для наиболее точной оценки влияния факторов, включенных в модель, проведем регрессионный анализ с помощью языка программирования Python и модуля statsmodels. Результаты анализа показаны на Рисунке 5.

OLS Regression Results

Dep. Variable:	CRIME_TOTAL	R-squared:	0.963
Model:	OLS	Adj. R-squared:	0.925
Method:	Least Squares	F-statistic:	25.71
Date:	Thu, 27 May 2021	Prob (F-statistic):	0.00410
Time:	16:28:59	Log-Likelihood:	-72.232
No. Observations:	9	AIC:	154.5
Df Residuals:	4	BIC:	155.5
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.408e+04	6400.153	10.012	0.001	4.63e+04	8.18e+04
INCOME	-0.6512	0.124	-5.255	0.006	-0.995	-0.307
UNEMPL	97.5118	36.066	2.704	0.054	-2.623	197.647
HOMELESS	2.4355	1.111	2.191	0.094	-0.650	5.521
MIGRATION	-7.9822	11.603	-0.688	0.529	-40.198	24.234

Omnibus:	1.346	Durbin-Watson:	1.045
Prob(Omnibus):	0.510	Jarque-Bera (JB):	0.049
Skew:	-0.155	Prob(JB):	0.976
Kurtosis:	3.188	Cond. No.	4.73e+05

Рис. 5: Результаты регрессии по 4 переменным

P-значение – вероятность, позволяющая определить значимость коэффициента регрессии. В случаях, когда P-значение больше 0.05, коэффициент может считаться нулевым, что означает, что соответствующая независимая переменная статистически не влияет на зависимую переменную.

P-значения для переменной $MIGRATION$ значительно превышают 0.05, следовательно, удалим её и проведем регрессионный анализ еще раз. Результаты анализа показаны на Рисунке 6.

OLS Regression Results

Dep. Variable:	CRIME_TOTAL	R-squared:	0.958
Model:	OLS	Adj. R-squared:	0.933
Method:	Least Squares	F-statistic:	38.14
Date:	Thu, 27 May 2021	Prob (F-statistic):	0.000720
Time:	16:33:32	Log-Likelihood:	-72.735
No. Observations:	9	AIC:	153.5
Df Residuals:	5	BIC:	154.3
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.078e+04	4012.954	15.146	0.000	5.05e+04	7.11e+04
INCOME	-0.5899	0.081	-7.242	0.001	-0.799	-0.381
UNEMPL	106.7633	31.653	3.373	0.020	25.397	188.130
HOMELESS	2.5567	1.038	2.463	0.057	-0.112	5.225

Omnibus:	0.476	Durbin-Watson:	1.228
Prob(Omnibus):	0.788	Jarque-Bera (JB):	0.229
Skew:	0.322	Prob(JB):	0.892
Kurtosis:	2.557	Cond. No.	3.14e+05

Рис. 6: Результаты регрессии по переменным INCOME, UNEMPL и HOMELESS

Теперь исключим из модели параметр *HOMELESS* из-за высокого P-значения и проведем регрессию по двум переменным. Результаты анализа представлены на Рисунке 7.

OLS Regression Results

Dep. Variable:	CRIME_TOTAL	R-squared:	0.907
Model:	OLS	Adj. R-squared:	0.876
Method:	Least Squares	F-statistic:	29.38
Date:	Thu, 27 May 2021	Prob (F-statistic):	0.000796
Time:	16:41:07	Log-Likelihood:	-76.311
No. Observations:	9	AIC:	158.6
Df Residuals:	6	BIC:	159.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.722e+04	4133.989	16.260	0.000	5.71e+04	7.73e+04
INCOME	-0.6365	0.108	-5.915	0.001	-0.900	-0.373
UNEMPL	117.1472	42.605	2.750	0.033	12.897	221.398

Omnibus:	1.773	Durbin-Watson:	1.876
Prob(Omnibus):	0.412	Jarque-Bera (JB):	1.030
Skew:	0.771	Prob(JB):	0.597
Kurtosis:	2.394	Cond. No.	2.37e+05

Рис. 7: Результаты регрессии по переменным INCOME и UNEMPL

2.6. Интерпретация результатов

По результатам регрессионного анализа получено следующее уравнение:

$$CRIME\_TOTAL = 67220 - 0.6365 * INCOME + 117.1472 * UNEMPL$$

Данное уравнение показывает, что в среднем с увеличением среднедушевых доходов населения на 1 рубль в месяц будет наблюдаться снижение количества преступлений на 0.6365, а с повышением численности безработных в Санкт-Петербурге на 1000 человек будет наблюдаться повышение числа преступлений на 117.1472.

Проверка адекватности модели осуществляется с помощью расчета F-критерия Фишера. В данном случае $p(F) < 0.05$, и уравнение статистически значимо.

P-значения для всех параметров модели меньше 0.05, следовательно, все параметры являются статистически значимыми.

Глава 3. Построение модели SARIMA, анализ временного ряда

3.1. SARIMA для прогнозирования временных рядов

Сезонное авторегрессионное интегрированное скользящее среднее, SARIMA или Seasonal ARIMA, является расширением ARIMA, которое явно поддерживает одномерные данные временных рядов с сезонным компонентом.

Эта модель добавляет к ARIMA три новых гиперпараметра для указания авторегрессии (AR), разности (I) и скользящего среднего (MA) для сезонной составляющей ряда, а также дополнительный параметр для периода сезонности. Таким образом, сезонная модель ARIMA формируется путем включения дополнительных сезонных компонентов в ARIMA. Сезонная часть модели состоит из компонентов, которые очень похожи на несезонные компоненты модели, но включают обратные сдвиги сезонного периода.

Данные о динамике преступлений для анализа возьмем с сайта ЕМИСС. [12] Данные представляют собой временной ряд общего количества совершенных преступлений в Санкт-Петербурге с января 2010 г. по апрель 2021 г. с периодичностью по месяцам.

3.2. Стационарность

Проведем анализ временного ряда. График временного ряда выглядит следующим образом (Рисунок 8):

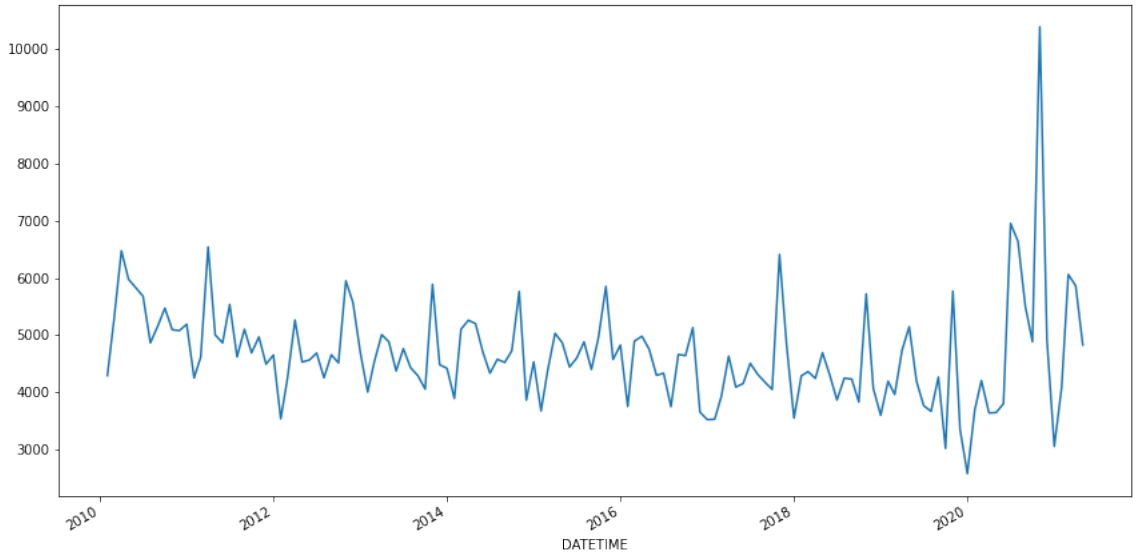

Рис. 8: График временного ряда

Перед тем, как перейти к моделированию, стоит сказать о таком важном свойстве временного ряда, как стационарность. Под стационарностью понимают свойство процесса не менять своих статистических характеристик с течением времени, а именно постоянство матожидания, постоянство дисперсии и независимость ковариационной функции от времени. По стационарному ряду проще строить прогноз, так как предполагается, что его будущие статистические характеристики не будут отличаться от наблюдаемых текущих.

3.3. Устранение нестационарности

Основная особенность нестационарного временного ряда заключается в том, что он может обладать трендом и содержать циклическую составляющую. Декомпозиция временных рядов – это процесс разделения данных временных рядов на их основные компоненты. Эти компоненты включают потенциальный тренд (общий рост или падение среднего), сезон-

ность (повторяющийся цикл) и оставшийся случайный остаток. Прежде чем применять модель, необходимо идентифицировать и отделить тенденции и сезонность от данных временного ряда. В библиотеке Python *statsmodels* есть метод декомпозиции временных рядов, который называется *seasonal_decompose*(). Применив его ко временному ряду количества преступлений, можно явно увидеть тренд и сезонную составляющую с периодом в 12 месяцев (Рисунок 9).

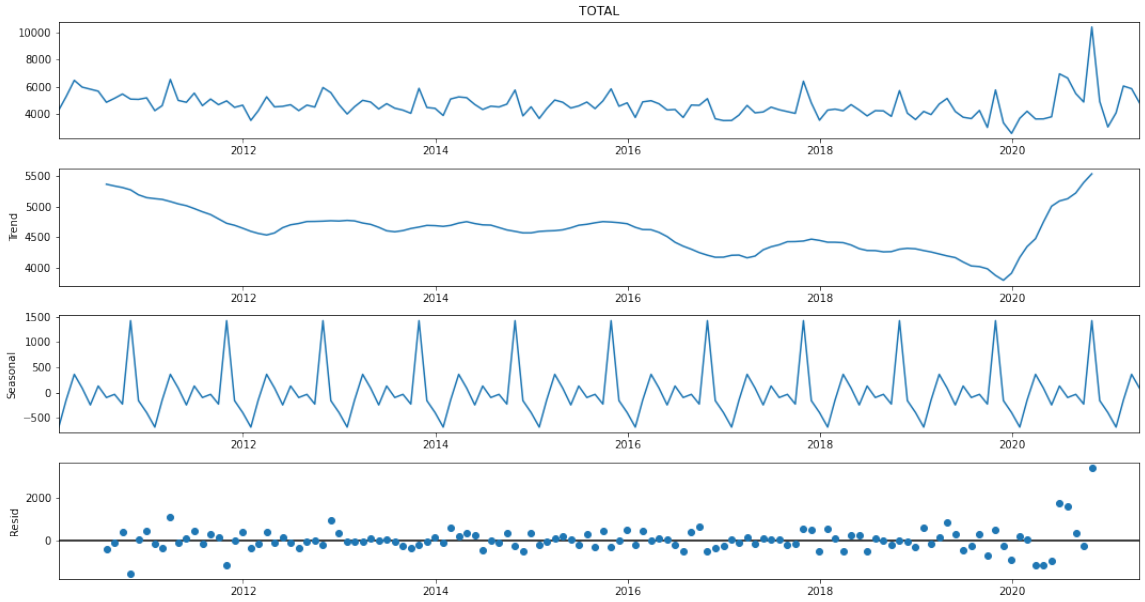

Рис. 9: Декомпозиция временного ряда

Перед тем, как построить ARIMA модель для количества преступлений, приведем ряд к стационарному виду. Для проверки ряда на стационарность используем *тест Дики-Фулера* – методику, используемую в прикладной статистике и эконометрике в целях анализа временных рядов.

Критерий Дики-Фуллера для исходного ряда: $p = 0.166649$. Как и следовало ожидать, исходный ряд стационарным не является, критерий Дики-Фуллера не отверг нулевую гипотезу о наличии единичного корня.

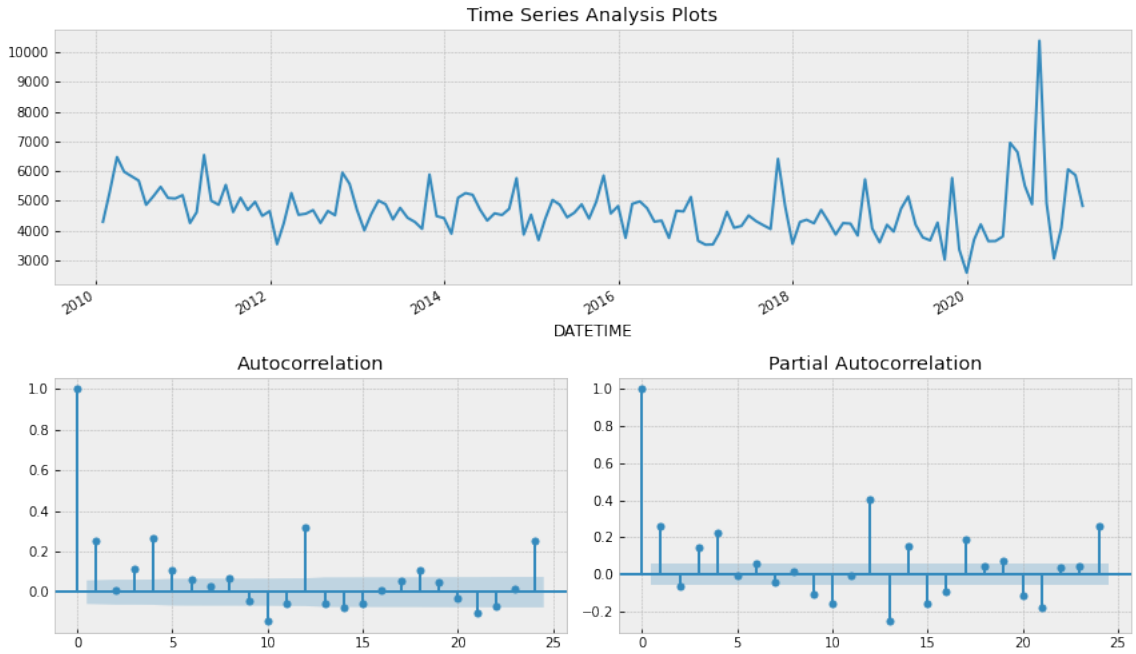

Рис. 10: ACF и PACF исходного ряда

По автокорреляционной функции можно судить о сезонности порядка 12 месяцев в данном ряде. Возьмём у ряда сезонные разности с лагом 12:

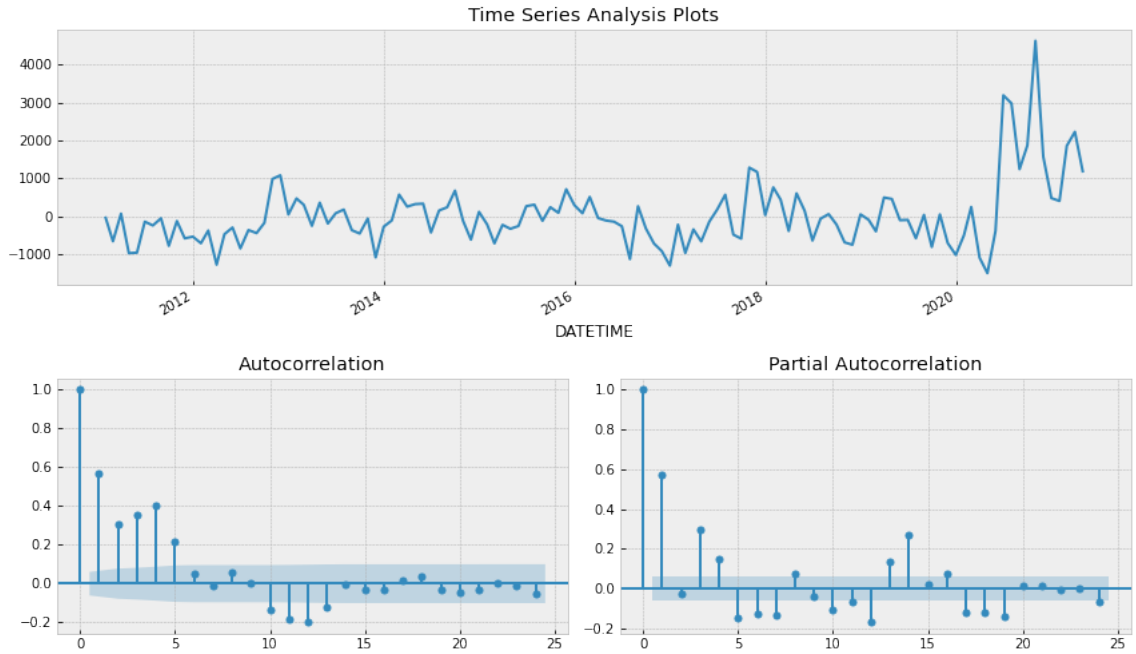

Рис. 11: ACF и PACF ряда с устраненной сезонностью

Критерий Дики-Фуллера продифференцированного ряда: $p = 0.505319$. Критерий Дики-Фуллера по-прежнему отвергает нулевую гипотезу о нестационарности, и автокорреляционная функция имеет большое количество значимых лагов. Стоит взять еще первые разности, чтобы привести ряд к стационарному виду.

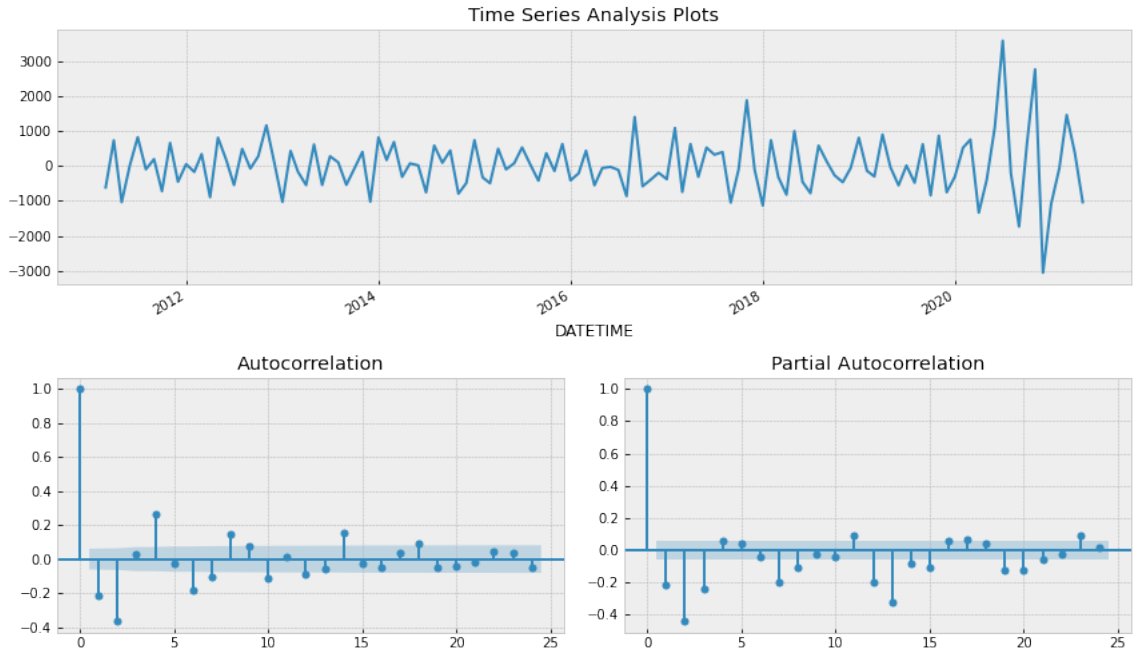

Рис. 12: ACF и PACF ряда первых разностей

Критерий Дики-Фуллера после взятия первых разностей: $p = 0.000001$. После взятия сезонных разностей с лагом 12 и дополнительно разностей порядка 1, получили стационарный ряд.

3.4. Поиск параметров по сетке

Один из подходов к настройке модели, использующей быстрое и параллельное современное оборудование, заключается в поиске в сетке набора конфигураций гиперпараметров, чтобы определить, какая конфигурация работает лучше всего. Часто этот процесс может выявить неинтуи-

тивные конфигурации моделей, которые приводят к более низкой ошибке прогноза, чем конфигурации, указанные в результате тщательного анализа.

Модель ARIMA имеет три параметра тренда, которые требуют настройки, в частности:

- p: Порядок авторегрессии тренда
- d: Порядок изменения тренда
- q: Тренд скользящей средней

Также есть четыре сезонных параметра, не являющиеся частью ARIMA, но входящие в модель SARIMA, которые должны быть настроены, а именно:

- P: Сезонный порядок авторегрессии
- D: Порядок сезонных разниц
- Q: Сезонный порядок скользящих средних
- m: Количество временных шагов за один сезонный период

Вместе обозначение для модели SARIMA задается как:

$$SARIMA(p, d, q)(P, D, Q, m)$$

Если в результате предварительного анализа о задаче известно достаточное количество информации, чтобы указать один или несколько из этих параметров, то имеет смысл указать их сразу. В ходе приведения ряда к стационарному ряд был продифференцирован один раз с порядком 12 и один раз с порядком 1, поэтому установим параметры $d = 1$, $D = 1$, $m = 12$

Обычно конфигурации предполагают, что каждый из компонентов AR, MA и I для тренда и сезонности имеет низкий порядок, например лежит в [0, 2]. Можно расширить эти диапазоны, если сделать предположение, что порядок может быть выше.

Для поиска лучшей модели зададим сетку гиперпараметров следующего вида: $p \in [0; 3]$, $d = 1$, $q \in [0; 3]$, $P \in [0; 3]$, $D = 1$, $Q \in [0; 3]$, $m = 12$,

Будем искать лучшую модель по параметру AIC (информационный критерий Акаике). После перебора всех параметров в сетке обнаружена лучшая конфигурация для модели с наименьшим AIC: $p = 2$, $d = 1$, $q = 2$, $P = 0$, $D = 1$, $Q = 1$, $m = 12$, которая соответсвует модели $SARIMA(2, 1, 2)(0, 1, 1, 12)$

Результаты модели показаны на рисунке 13:

SARIMAX Results

Dep. Variable:	TOTAL	No. Observations:	136
Model:	SARIMAX(2, 1, 2)x(0, 1, [1], 12)	Log Likelihood	-974.409
Date:	Sat, 29 May 2021	AIC	1960.819
Time:	15:03:13	BIC	1977.692
Sample:	0	HQIC	1967.673
	- 136		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2612	0.121	2.154	0.031	0.023	0.499
ar.L2	-0.7800	0.083	-9.414	0.000	-0.942	-0.618
ma.L1	-0.6629	0.141	-4.690	0.000	-0.940	-0.386
ma.L2	0.6961	0.122	5.713	0.000	0.457	0.935
ma.S.L12	-0.3065	0.154	-1.990	0.047	-0.608	-0.005
sigma2	4.31e+05	3.03e+04	14.229	0.000	3.72e+05	4.9e+05

Ljung-Box (L1) (Q):	0.22	Jarque-Bera (JB):	185.38
Prob(Q):	0.64	Prob(JB):	0.00
Heteroskedasticity (H):	2.56	Skew:	0.83
Prob(H) (two-sided):	0.00	Kurtosis:	8.78

Рис. 13: Результаты модели

P-значения для всех параметров модели меньше 0.05, следовательно, все параметры являются статистически значимыми. Модель соотносится с реальными данными следующим образом (Рисунок 14):

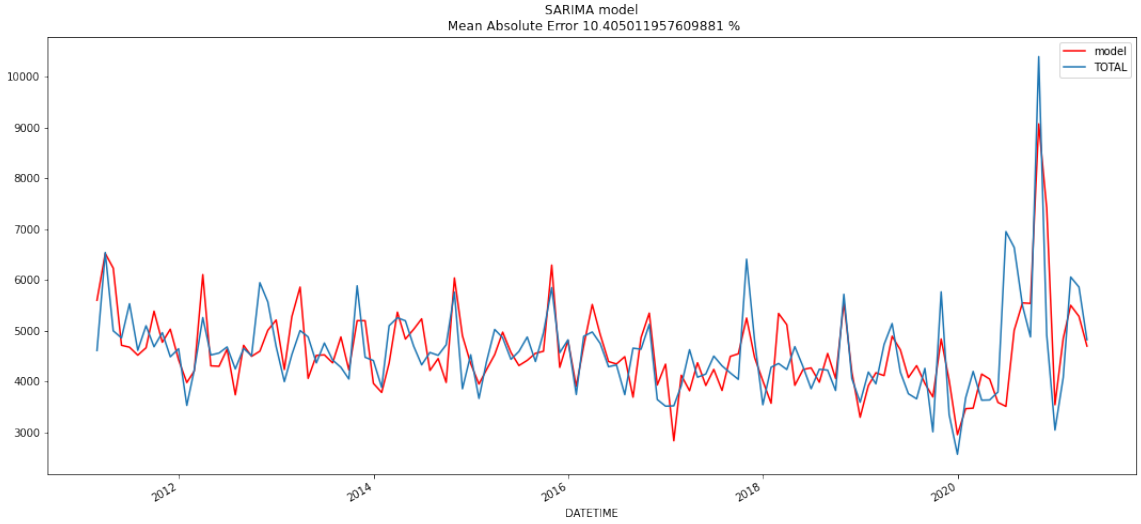

Рис. 14: Результаты модели

Тест Дики-Фулера на остатках модели показал, что остатки являются стационарными и соответсвуют «белому шуму», следовательно, модель получила всю полезную информацию из данных.

3.5. Прогнозирование

Для прогнозирования используем библиотеку $forecast$. Сделаем прогноз на 1 шаг вперед (Рисунок 15).

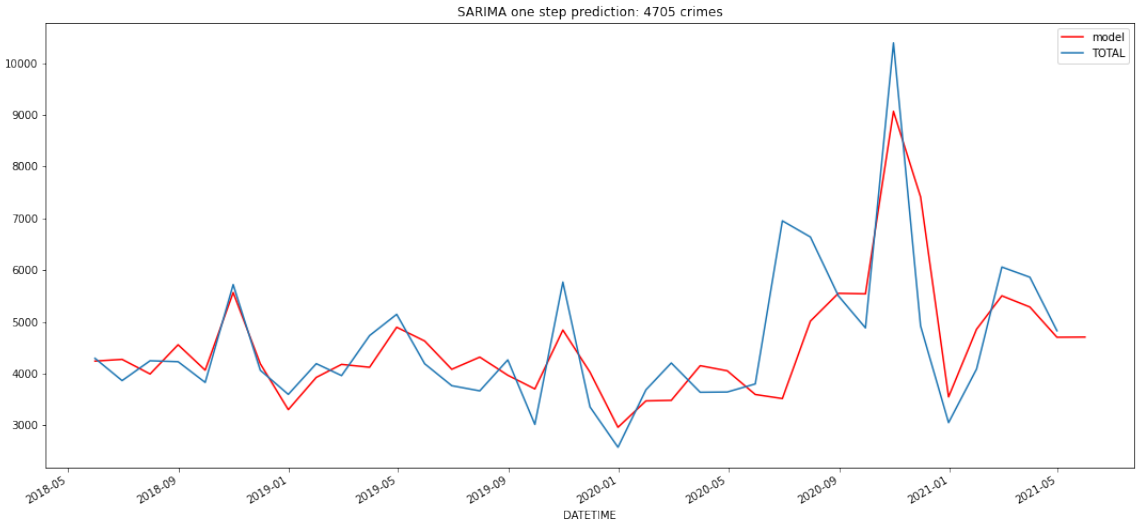

Рис. 15: Прогноз на 1 месяц вперед

Прогноз модели: в мае 2021 года в Санкт-Петербурге общее число совершенных преступлений будет равняться 4705.

Библиотека *forecast* позволяет делать и многошаговые прогнозы. Сделаем прогноз на 3 месяца вперед (Рисунок 16):

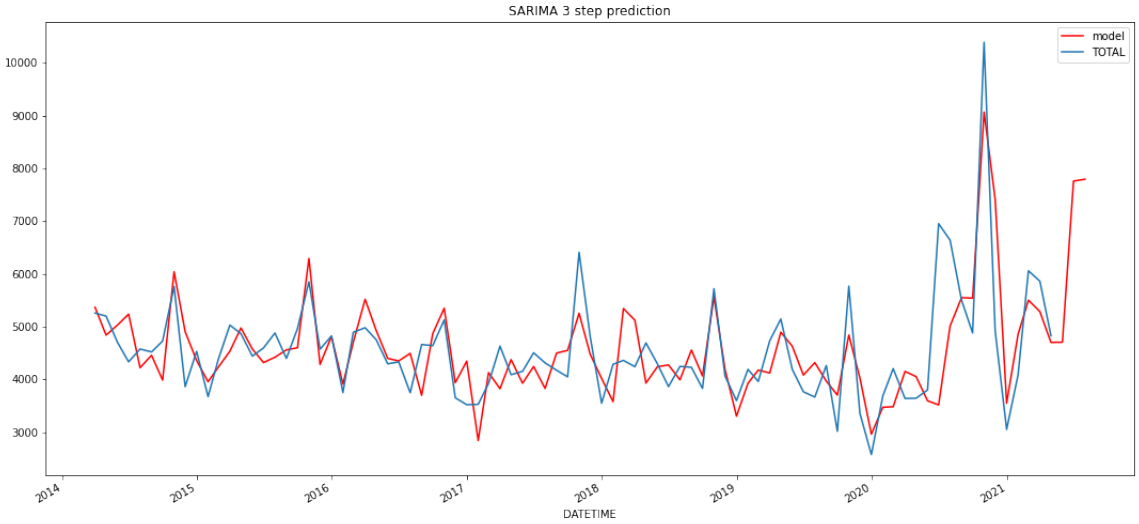

Рис. 16: Прогноз на 3 месяца

Глава 4. Построение аналитической панели

4.1. Основные библиотеки

С помощью библиотек языка python был построен дашборд с интерактивной и наглядной инфографикой. Источником сведений о состоянии преступности является официальный сайт прокуратуры Санкт-Петербурга. [13] Ежемесячно прокуратура Санкт-Петербурга публикует актуальные данные о состоянии преступности за прошедший месяц, включая данные о различных типах совершенных преступлений, их тяжести, а также о социально-криминологической характеристике преступности.

С помощью библиотеки *openpyxl* написан скрипт, объединяющий документы excel с данными о преступности в единый датасет. Затем с помощью библиотек *pandas* и *plotly* эти данные были использованы для построения дашборда с показателями преступности в г. Санкт-Петербург.

Библиотека *plotly* в Python делает интерактивные графики онлайн и позволяет сохранять их в автономном режиме, если это необходимо. У *plotly* есть несколько функций, которые делают её лучше, чем другие графические библиотеки:

- По умолчанию она интерактивна
- Диаграммы не сохраняются как изображения, а сериализуются как JSON, что делает их открытыми для чтения с помощью R, MATLAB, Julia
- Компонентами легко манипулировать и встраивать в Интернет

4.2. Компоненты дашборда

Библиотека *plotly* содержит различные компоненты для интерактивного взаимодействия с инфографикой. В построенном дашборде есть компоненты выбора года, месяца и вида преступлений (Рисунок 17).

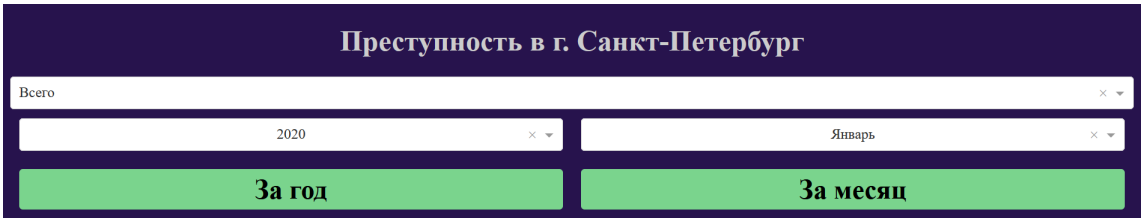

Рис. 17: Компонент дашборда. Выбор года, месяца и вида преступлений

Далее находятся компоненты, содержащие основные числовые показатели преступности (количество зарегистрированных и раскрытых преступлений за отчётный период, рост или падение показателя относительно предыдущего месяца) и социально-криминологическую характеристику преступности (количество преступлений, совершенных несовершеннолетними, совершенных лицами, уже совершавшими преступления, совершенных группой, в состоянии алкогольного или наркотического опьянения, процент от общего числа раскрытых преступлений).

Рис. 18: Компонент дашборда. Количественные показатели

Далее расположены различные графики. Присутствует график количества выбранного типа преступлений по месяцам за 1 год (Рисунок 19).

Рис. 19: Динамика за 2020 год

Следующие графики - круговые диаграммы, демонстрирующие процентное соотношение разных типов совершенных преступлений: кража, грабёж, убийство, и т.д. (Рисунок 20).

Рис. 20: Компонент дашборда. Соотношение преступлений по типам

Наконец, присутствуют графики, демонстрирующие следующие особенности совершенных преступлений: количество тяжких и особо тяжких преступлений, преступлений, совершенных в особо крупном размере,

преступлений экономического характера, преступлений, совершенных в общественных местах или на улицах города. Визуально показано соотношение количества зарегистрированных и раскрытых преступлений таких видов к общему числу зарегистрированных преступлений (Рисунок 21).

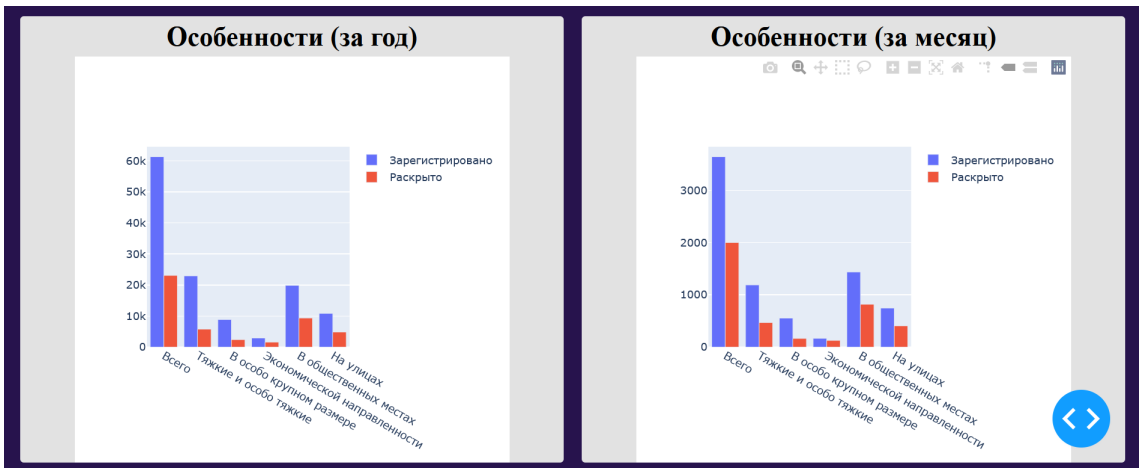

Рис. 21: Компонент дашборда. Соотношение преступлений по особенностям

4.3. Некоторые выводы

Визуализиция статистических данных позволяет быстро увидеть, что:

1. Значительная часть преступлений (около 45 процентов) совершается лицами, ранее уже их совершавшими.

2. С июня 2020 года резко возросло количество случаев мошенничества более чем в 3 раза.

3. Значительная часть преступлений (около 94 процентов) приходится на 3 следующих вида: кража, мошенничество, незаконный оборот наркотиков. Примерно половина из всех зарегистрированных преступлений это кражи.

4. Тяжких и особо тяжких преступлений совершается меньше, чем преступлений небольшой и средней тяжести.

5. Примерно каждое третье преступление совершается в общественных местах.

Заключение

В данной работе были выполнены все поставленные задачи. В главе 1 были изучены все необходимые понятия с математической стороны.

В главе 2 была построена модель линейной регрессии, и выявлены основные факторы, влияющие на уровень преступности в Санкт-Петербурге.

В главе 3 была найдена оптимальная конфигурация модели SARIMA для моделирования временного ряда и произведен прогноз количества преступлений на несколько месяцев вперед.

В главе 4 были описаны компоненты созданной аналитической панели.

В целом, в работе показано как можно использовать язык Python для анализа данных.

Список литературы

1. Brooks, C., 2008. Introductory Econometrics for Finance. Cambridge University Press.

2. Montgomery, D.C., C.L. Jennings and M. Kulahci, 2008. Introduction to Time Series Analysis and Forecasting. John Wiley Sons. Inc.

3. Антонян, Ю.М., 2006. Причины преступности. Москва: ИД "Камерон".

4. Буре В. М., Парилина Е. М., Седаков А. А. Методы прикладной статистики в R и Excel. 3 изд. Лань, 2018. 152 с..

5. Машинное обучение (курс лекций) // MachineLearning http://www.machinelearning.ru/wiki

6. Барышникова, А.В., 2016. Зависимость уровня преступности от экономических и социальных факторов в регионах РФ, ФГБОУ ВПО "Пермский национальный исследовательский технологический политехнический университет".

7. Красикова, Е. М. Статистическое изучение уровня преступности в Российской Федерации / Е. М. Красикова. — Текст : непосредственный // Молодой ученый. — 2017. — № 16 (150). — С. 269-272. — URL: https://moluch.ru/archive/150/42387/

8. Гусева, М.В., 2015. Исследование основных факторов, влияющих на уровень преступности в России, ФГАОУ ВО "Уральский федеральный университет имени первого Президента России Б.Н.Ельцина".

9. Шумилин, О.В. and Н.В. Мячин, 2020. Анализ и прогнозирование динамики зарегистрированных преступлений в России на основе временного ряда 1991–2019 гг.. Вестник Уральского юридического института МВД России, 4. Date Views 01.05.2021 cyberleninka.ru/article/n/analiz-i-prognozirovanie-dinamiki-zaregistrirovannyh-prestupleniy-v-rossii-na-osnove-vremennogo-ryada-1991-2019-gg.

10. Богданова, М.В., Л.С. Паршинцева and В.Ю. Квачко, 2019. Методика моделирования и прогнозирования преступности в Российской Федерации. Правовая информатика, 4. Date Views 02.05.2021 cyberleninka.ru/article/n/metodika-modelirovaniya-i-prognozirovaniya-prestupnosti-v-rossiyskoy-federatsii.

11. Многофакторный регрессионный анализ // Studme https://studme.org/

12. Единая межведомственная информационно-статистическая система (ЕМИСС) https://www.fedstat.ru/

13. Федеральная служба государственной статистики (Росстат) https://rosstat.gov.ru/

14. Прокуратура Санкт-Петербурга https://epp.genproc.gov.ru/web/proc_78