**Данные исследования**

Для проведения данного исследования, а именно построение модели прогнозирования интернет-продаж в РФ, были отобраны различные показатели, оказывающие возможное влияние на динамику рынка интернет-торговли, за период с 2016 по 2021 год по субъектам РФ (было отобран 81 субъект РФ) из различных источников (преимущественно ФСС (Росстат). В общей сложности по каждому показателю было проведено 486 наблюдений (за 6 лет по 81 субъекту). Все переменные были разделены на пять групп – зависимая переменная (y), социально-демографические факторы (соц x), экономические факторы (эк x), технологические факторы (тех x), логистические факторы (лог x). Список всех показателей представлен в таблице А.1.

В качестве зависимой переменной предлагается рассмотреть показатель продаж через Интернет в абсолютном выражении.

На основании этого можно предложить следующую гипотезу исследования: наибольшее влияние на развитие интернет-торговли оказало распространение безналичного расчета с помощью карт, эпидемия коронавируса, рост ВВП. Также можно предположить, что рынок интернет-торговли будет иметь и дальше положительную динамику. Наибольшую точность должны показать анасамблевая модель.

Таким образом, для исследования был отобран 20 показателей по различным группам факторов за 2016-2021 гг. по 81 субъекту РФ.

**Анализ описательной статистики**

На первом этапе исследования необходимо провести предварительный анализ данных. В таблице В.3 представлены некоторые основные результаты описательной статистики для всех 81 субъектов РФ за 2016-2021 гг. по выделенным показателям за исключением covid, так как это фиктивная переменная.

Отобразим данные для всех 81 субъектов РФ за 2016-2021 гг. по зависимой переменной esales по разделению фиктивной переменной covid на рисунке 1.

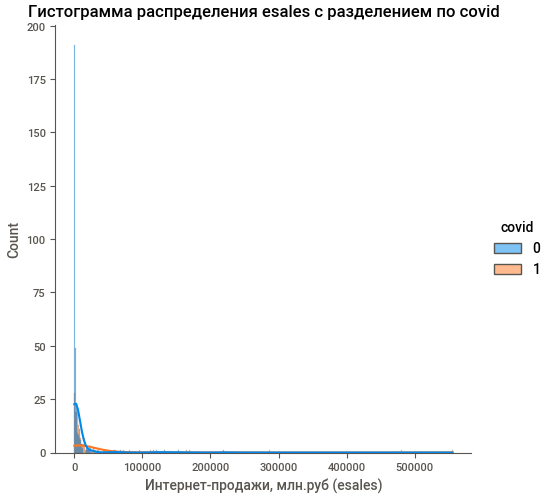


Рисунок 1 – Гистограмма распределения esales с разделением по covid

На графика распределение можно наблюдать большое количество выбросов и неструктурированность данных. Большинство значение находится около 0 и только несколько значений имеют большие продажи.

Таким образом, для дальнейшей работы необходима обработка данных. Используем нормализацию для удобства дальнейшего анализа. На рисунке 2 представим график распределения по нормализированным данным.

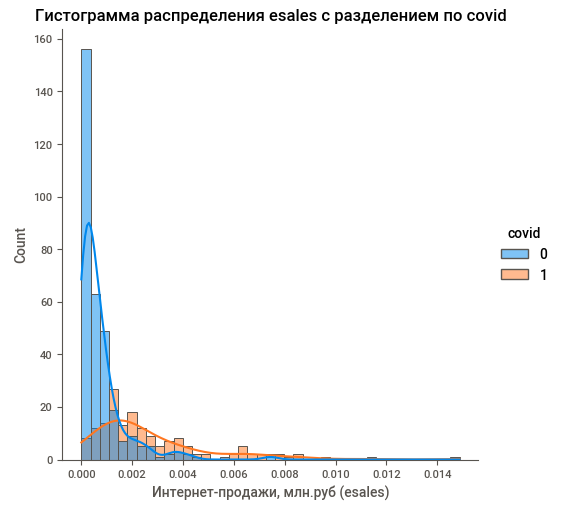


Рисунок 2 – Гистограмма распределения esales с разделением по covid по нормализированным данным

После нормализации данных видно, что разброс значений уменьшился и график принял адекватный вид для анализа. Можно наблюдать, что среднее и максимальное значения в период коронавируса больше, чем в период до него. Также стандартное отклонение выше во время эпидемии.

Построим графики ящиков с усами по рассматриваемой переменной, где отобразим основные статистические значения распределения данных.

На рисунке 3 отобразим график по переменной esales по нормализированным данным.

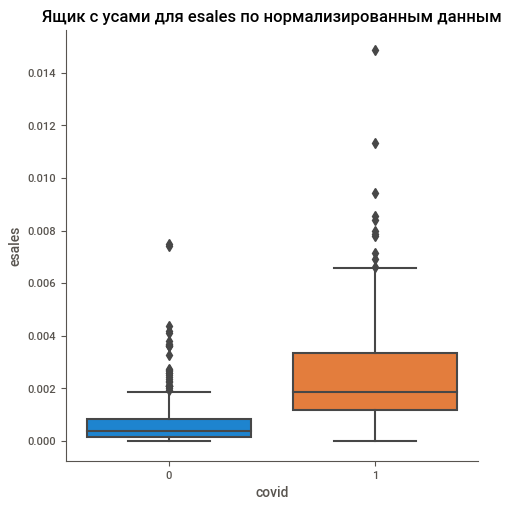


Рисунок 3 – Графики «ящик с усами» esales с разделением по covid по нормализированным данным

На графике видно, что имеется большое количество выбросов, которые во многом превышают средние значения по группам. Также можно наблюдать, что медианное значение в период коронавируса выше, чем до него. Помимо этого, разброс значений (межквартальный размах) в период коронавируса гораздо больше, чем до вспышки. Максимальное значение также выше в случае коронавируса. Можно сделать предположение, что вспышка коронавируса все же повлияла на увеличение доли продаж в Интернете, так как значения по выборке гораздо выше, чем в выборке без эпидемии.

Рассмотрим также описательную статистику для выделенной обучающей выборки, включающей значения с 2016 по 2020 гг., и для тестовой выборке с данными за 2021 года в динамике. Данные представлены в таблицах В.4 и В.5. Также отобразим динамику распределения на рисунке 4.

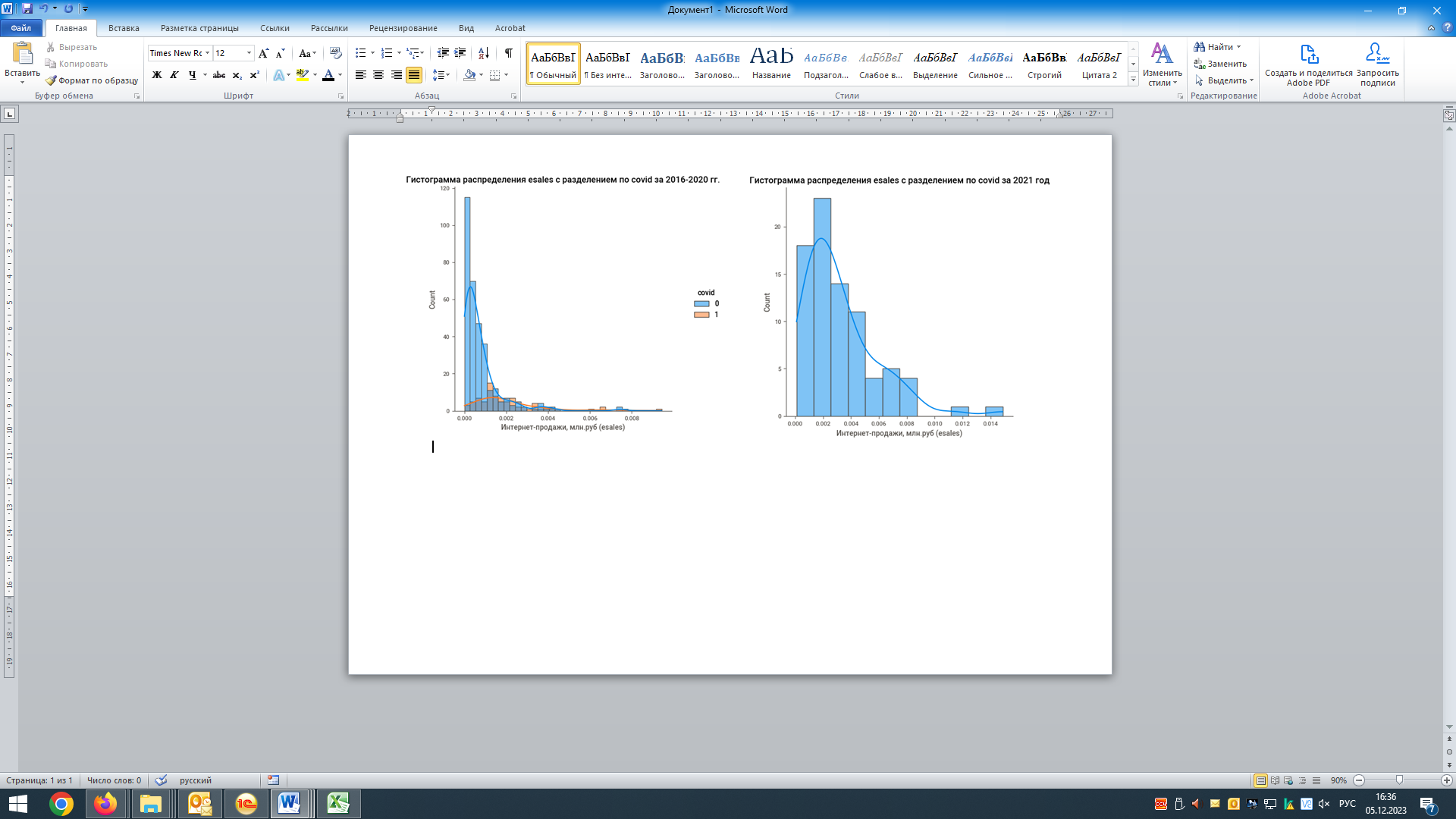


Рисунок 4 - Распределения esales по нормализированным данным по обучающей (2016-2020 гг.) и тестовой выборках (2021 г.)

Со временем прослеживается рост среднего значения распределения, а также повышение максимального значения. Так за 2016-2020 гг. среднее значение по распределению показателя esales составляло 8 166,65 млн. руб, а максимум – 481 413,77 млн. руб, а в 2021 году среднее значение составило 24 759,04 млн. руб., а максимум – 555 569,3 млн. руб. Вместе с этим возросло и стандартное отклонение с 32 404,81 млн. руб. до 71 541,7 млн. руб.

Рассмотрим также форму распределения данных на рисунке 5.

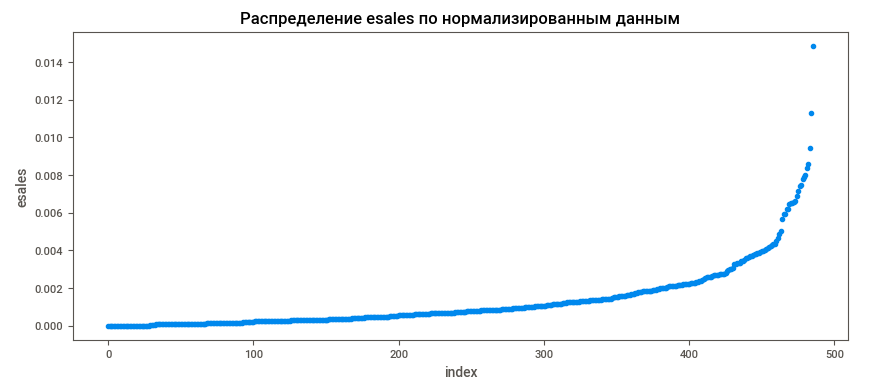


Рисунок 5 – Распределения esales по нормализированным данным

Четко можно наблюдать, что данные имеют экспоненциальное распределение, которое точно описывает изменение значений во времени. В нашем случае по годам.

Далее рассмотрим распределение по признакам.

На рисунке 6 представлено количество уникальных значений в каждом признаке.

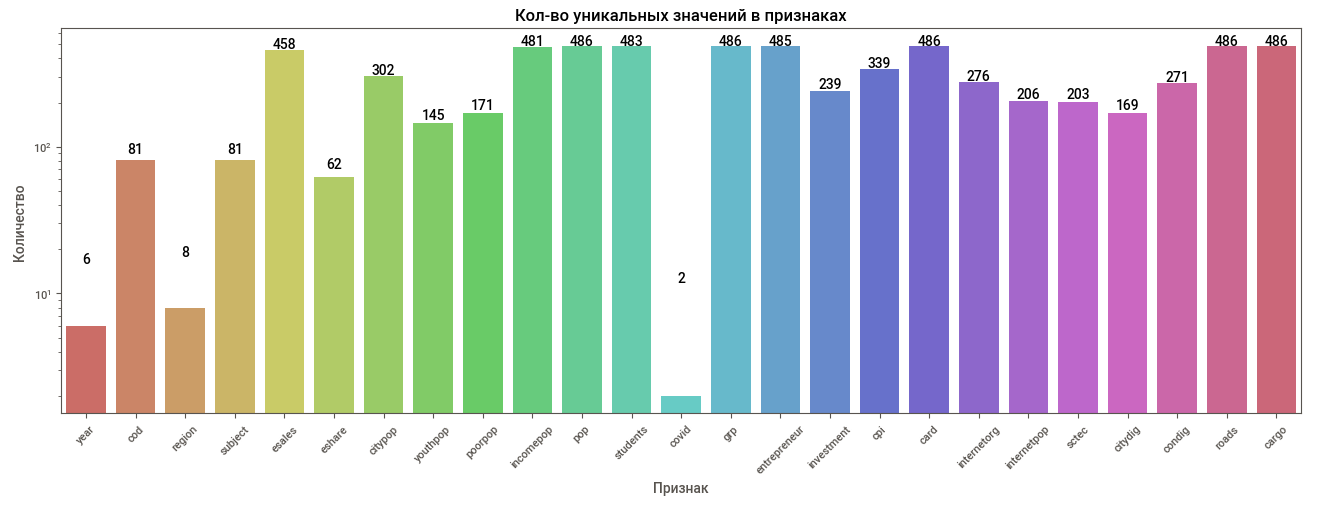


Рисунок 6 – Количество уникальных значений по каждому признаку датасета

Преимущественная часть переменных имеет большое количество значений, а где-то оно совпадает с количеством данных субъектов по 6 годам.

Рассмотрим также распределение целевого признака esales по регионам и субъектам РФ.

На рисунке 7 представлено распределение esales по регионам в среднем за 6 лет с 2016 года по 2021 год.

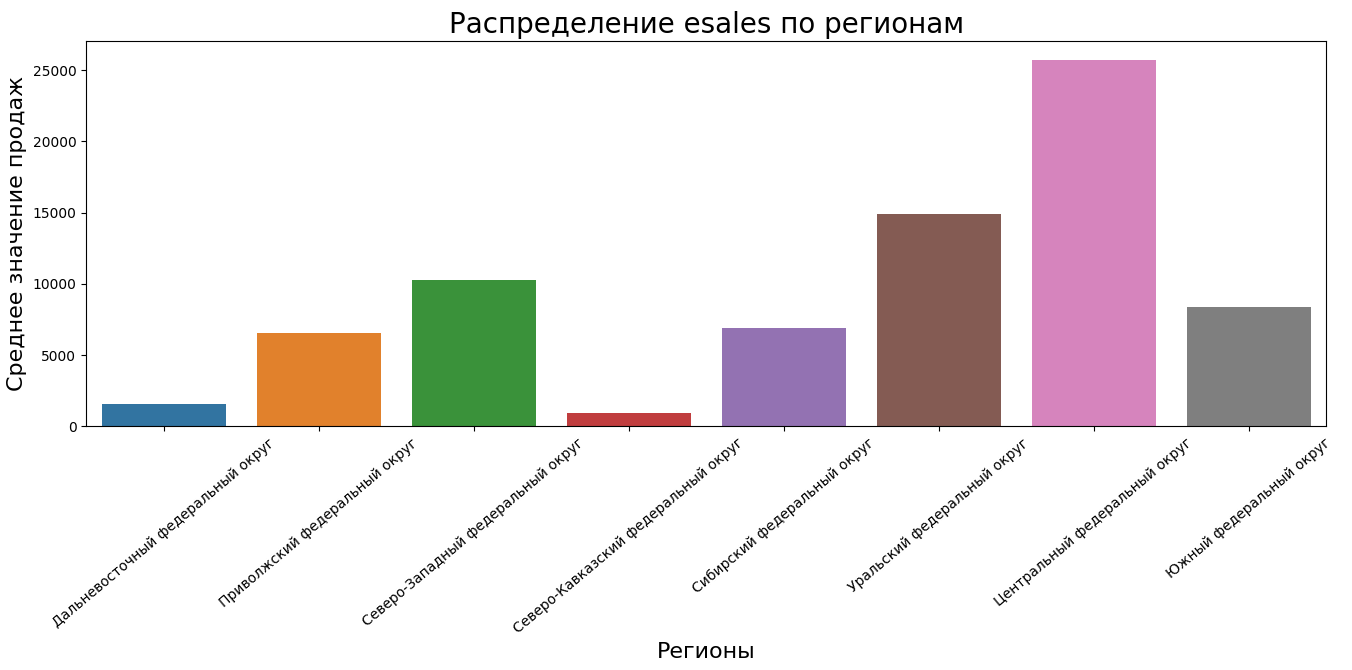


Рисунок 7 – Распределение esales по регионам в среднем за 6 лет с 2016 года по 2021 год

Как и следовало ожидать, наибольшее значение имеет Центральный федеральный округ, так как он включает в себя г. Москва и Московскую область – столицу России, где сосредоточен бюджет страны и где самые высокие зарплаты в стране и самая развитая инфраструктура.

На рисунке 8 представлено распределение населения по регионам.



Рисунок 8 – распределение населения по регионам в среднем за 6 лет с 2016 года по 2021 год

Наибольшее количество населения сосредоточено в Уральском федеральном округе.

Построим также графики «ящик с усами» для анализа распределения по регионам по нормированным значениям esales (рисунок 9) и значениям pop (рисунок 10).

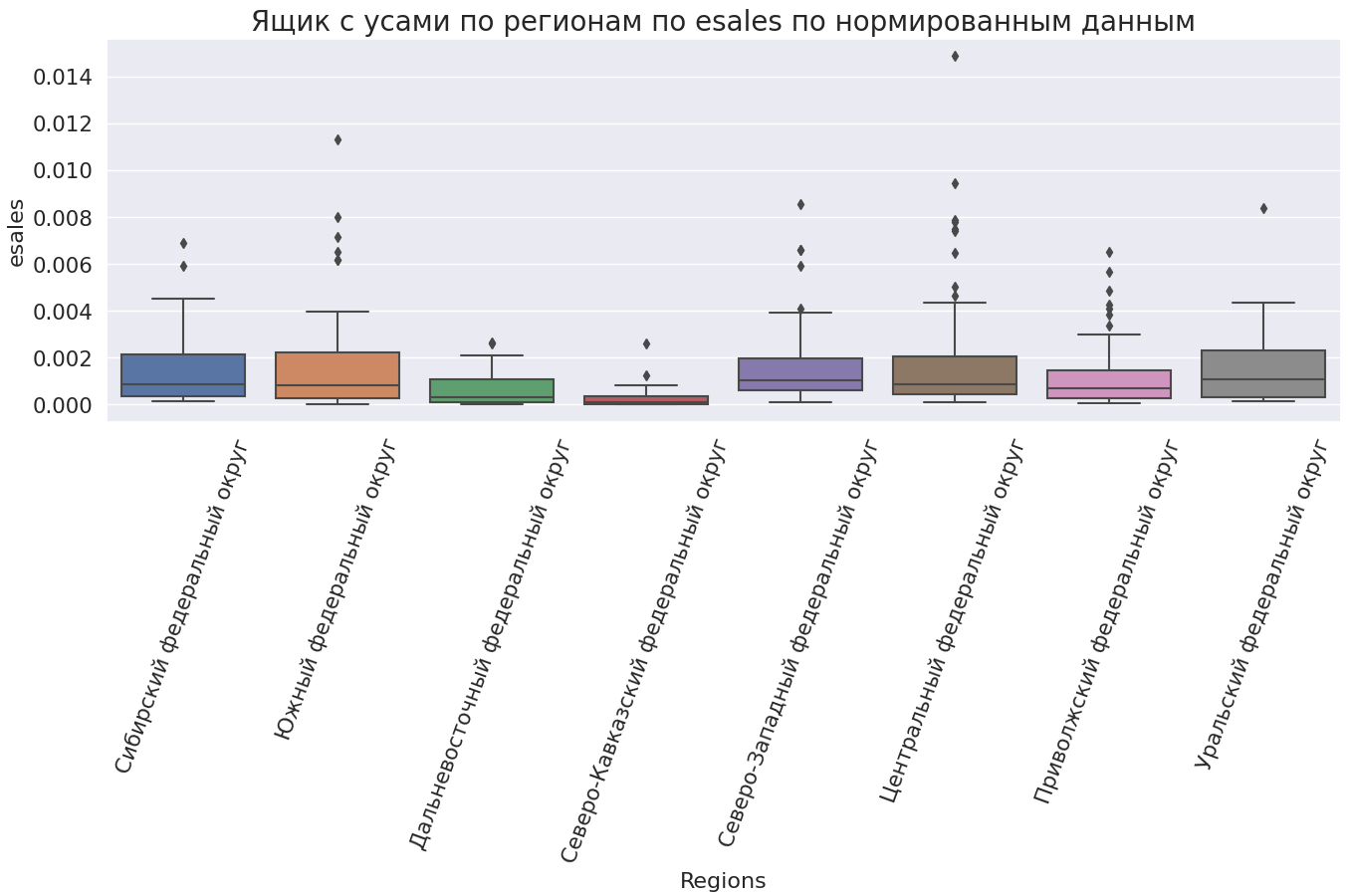


Рисунок 9- График «Ящик с усами» по регионам по esales   
по нормированным данным

Как упоминалось ранее, наибольшее значение выбросов имеет Центральный федеральный округ, а также регионы, в которых находятся города федерального значения (Санкт-Петербург, Севастополь), а также города-мегаполисы. Наименьшее значение по интернет-продажам наблюдается в Северо-Кавказском округе.

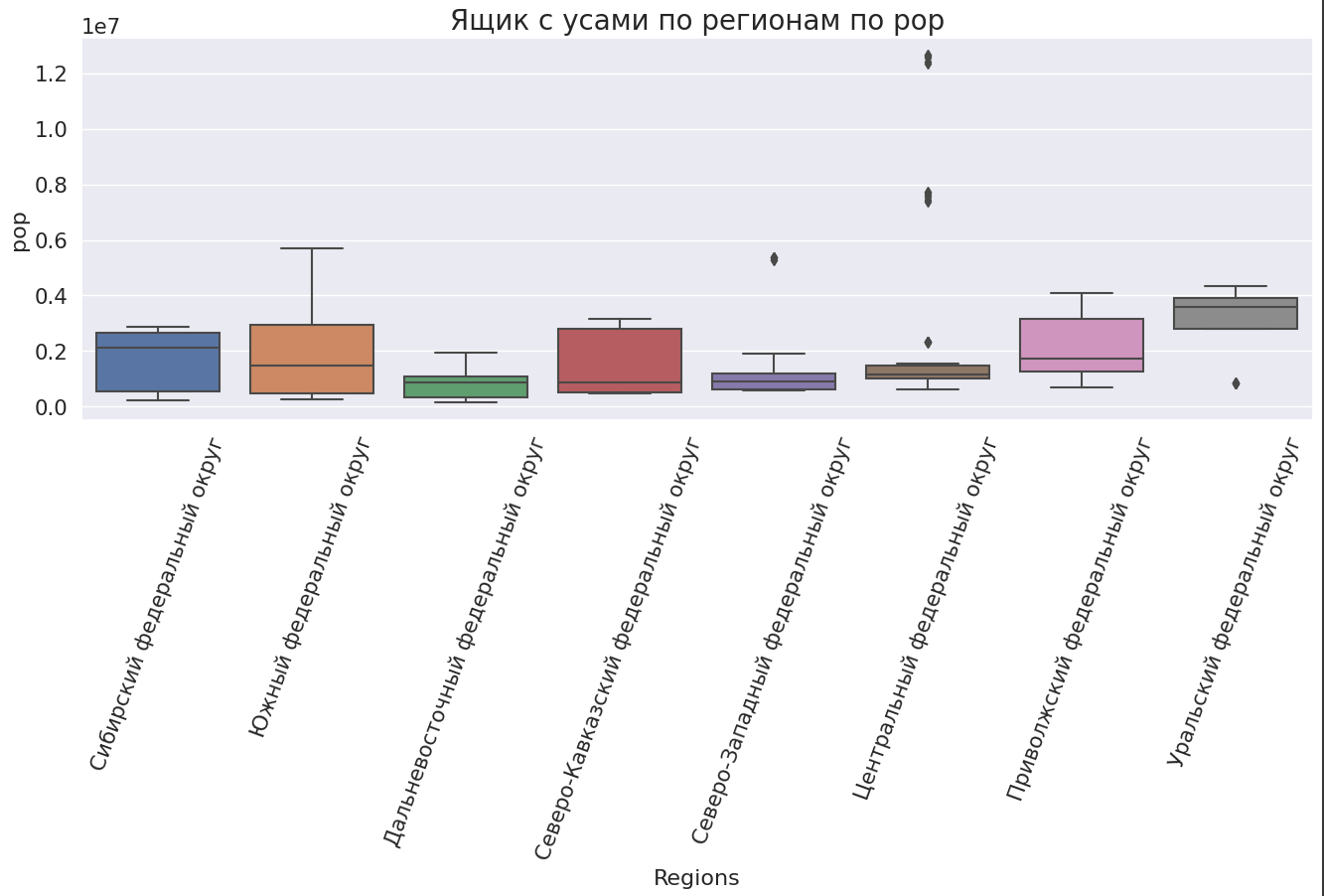


Рисунок 10 – График «Ящик с усами» по регионам по pop

по нормированным данным

По распределению населения также наблюдается наибольшее значение выбросов по Центральному региону. Самое высокое среднее значение по населению наблюдается в Уральском федеральном округе. Самый большой межквартальный размах наблюдается в Южном федеральном округе (скорее всего из-за неравномерного распределения между субъектами – основная концентрация в Краснодарском округе).

Рассмотрим также распределения отдельно по субъектам РФ.

На рисунке 11 представлен топ-5 субъектов РФ по среднему значению esales за 2016-2021 гг.

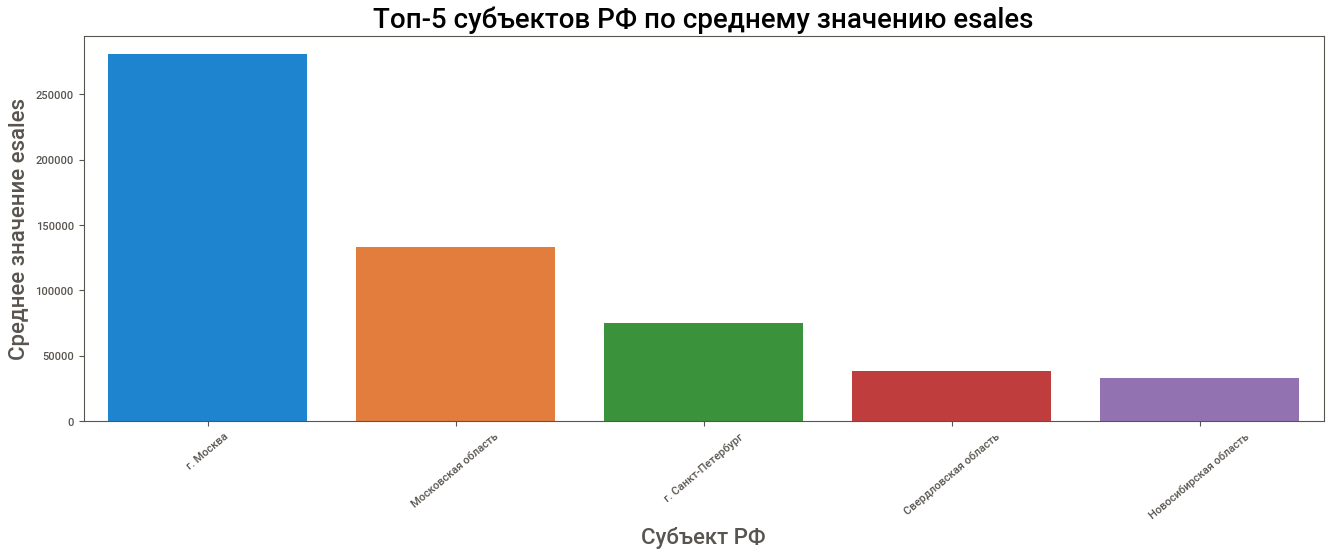


Рисунок 11 – Топ-5 субъектов РФ по среднему значению esales   
за 2016-2021 гг.

Именно значения имеют выбросы по всем данным, так как они являются городами мегаполисами.

Похожая тенденция наблюдается по показателю pop (рисунок 12).

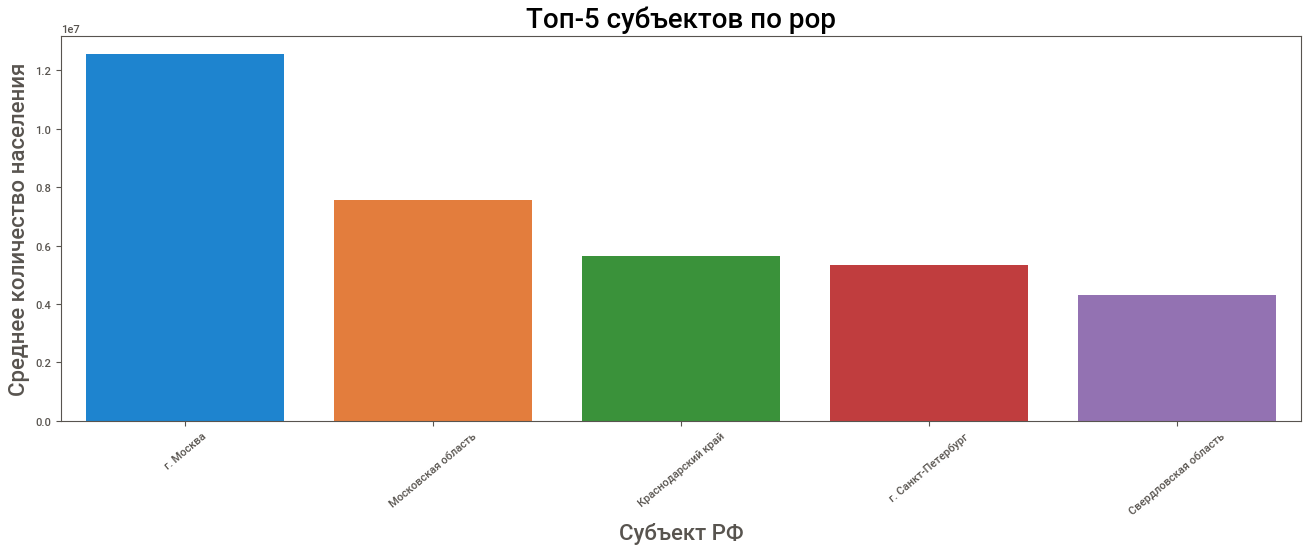


Рисунок 12 – Топ-5 субъектов РФ по среднему значению pop за 2016-2021 гг.

Анализ по остальным показателям представлен в приложении (рисунки В.2-В.20).

Постоим также корреляционную матрицу между отобранными признаками. Рассмотрим две различных матрицы: по коэффициенту корреляции Спирмена и Пирсона. Коэффициент Пирсона больше подходит для переменных, принадлежащих к интервальной шкале. Это можно увидеть по полученным результатам, так как по коэффициенту Пирсона найдено больше связей между переменными. На рисунке 13 отображены значения коэффициентов выше 0.5.

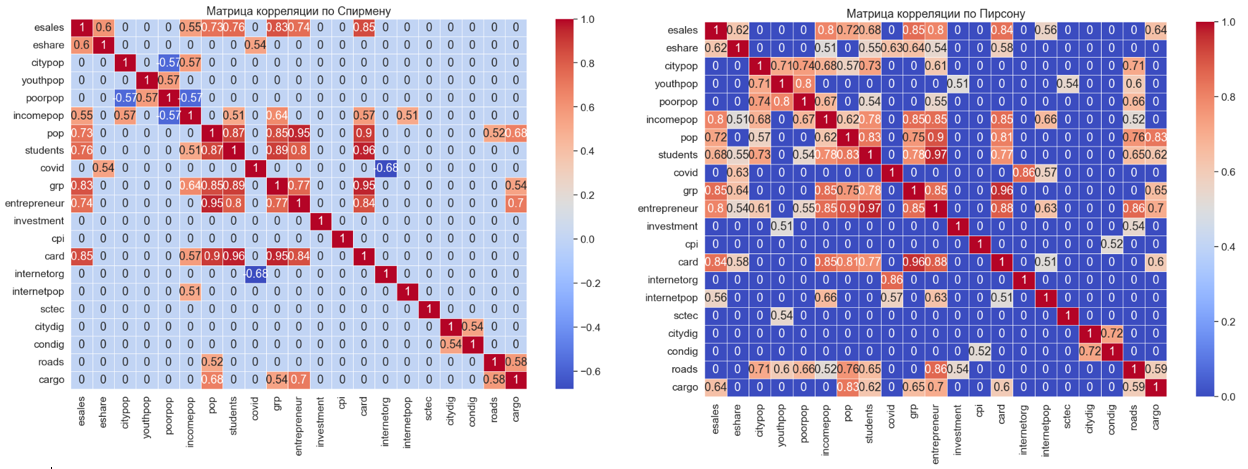


Рисунок 13 – Корреляционные матрицы по коэффициентам   
Спирмена и Пирсона

Сильную корреляцию можно наблюдать по следующим переменным:

- между pop и entrepreneur;

- между pop и card;

- между students и card;

- между students и entrepreneur;

- между grp и card.

Таким образом, eshare необходимо убрать из данных, так как целевой показатель esales напрямую рассчитывается из данной переменной. Также исключим показатель pop, так как он сильно коррелирует с переменными, связанными с численностью населения.

Рассмотрим также общий график интернет-продаж по РФ на рисунке.

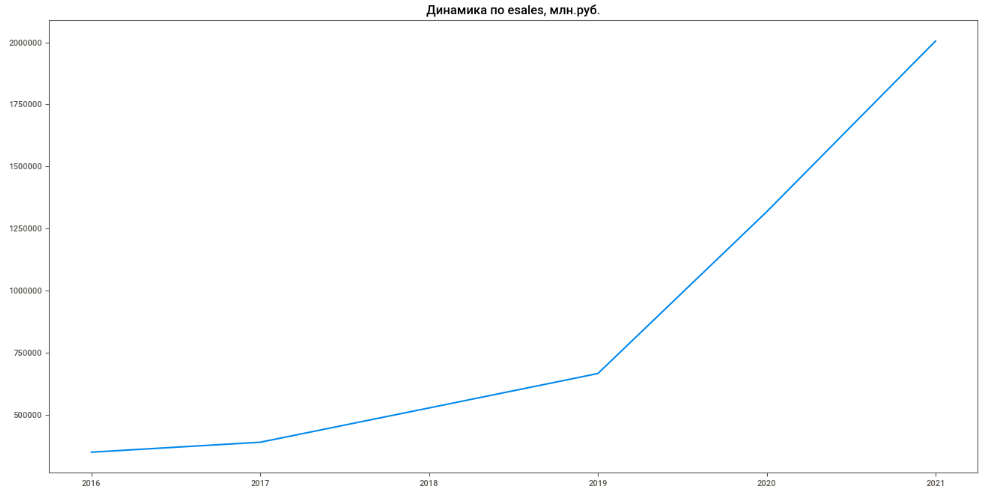


Рисунок 14 – Динамика интернет-продаж за 2016-2021 гг. по РФ

На графике наблюдается резкий прирост интернет-продаж в 2020 году. Это можно связать с распространением эпидемии коронавируса. То есть данный рынок и так имел некоторую тенденцию роста, но описанная ситуация позволила увеличить темпы. В 2021 году данный тренд также продолжает наблюдаться.

Так как данные представляют конкретное значение за один год без разбиения по месяцам, то невозможно рассмотреть какую-либо сезонность значений. Поэтому по проведенным тестам Дики-Фуллера (ADF = 10.3133, p-value = 1) и KPSS-тесту (KPSS = 0.3765, p-value = 0.0873) ряд является стационарным и значения по интернет-продажам не зависят от времени.

Но при этом, если построить график по автокорреляционной функции (рисунок 15), мы можем заметить, что в данных присутствует автокорреляция, которая указывает на присутствие четкого тренда.

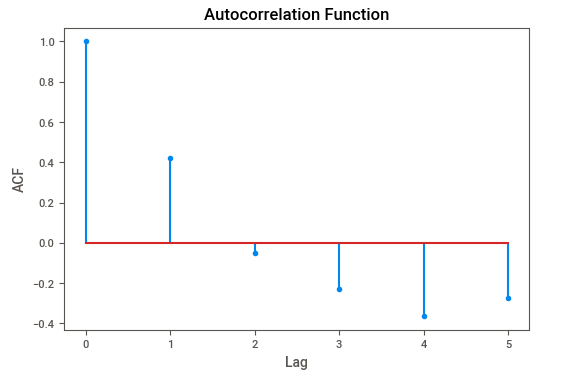


Рисунок 15 – График автокорреляционной функции по показателю esales

Таким образом, рассмотрев результаты анализа описательной статистики, можно сделать выводы о том, что отобранные данные являются несбалансированными, так как имеют большое количество выбросов, разные единицы измерения по признакам, мультиколлиниарность и автокорреляцию. Именно поэтому для работы был выбран алгоритм градиентного бусинга, который не чувствителен ко всем перечисленным выше проблемам и способен при этом показывать наилучшие результаты.