Министерство науки и высшего образования РФ  
ФГАОУ ВО

Национальный исследовательский технологический университет «МИСИС»

Институт Информационных технологий и компьютерных наук (ИТКН)

Автоматизация систем управления (АСУ)

**Курсовая работа**

по дисциплине «Прикладной статистический анализ»

на тему «Разработка модели прогнозирования прогноза пассажиропотока в транспорте в Москве»

Выполнил:

студент группы: \_\_\_\_\_БИВТ-21-2\_\_\_\_\_\_

ФИО: \_\_\_\_\_\_\_\_\_\_\_\_Винокуров Н.Р\_\_\_\_\_\_\_\_

Подпись: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Проверил:

ФИО: \_\_\_\_\_\_\_\_\_\_Маркарян Е.В.\_\_\_\_\_\_\_

Оценка: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата: «\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20\_\_\_ г.

Подпись: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Москва, 2023

**Оглавление**

[Введение 2](#_Toc154053442)

[Анализ характеристик объекта исследования 3](#_Toc154053443)

[1. Описание объекта исследования 3](#_Toc154053444)

[2. Анализ объекта исследования с помощью статистических показателей и выявление связей 4](#_Toc154053445)

[3. Постановка задачи моделирования. 6](#_Toc154053446)

[Моделирование статистических зависимостей 6](#_Toc154053447)

[4. Проверка гипотезы о нормальном распределении 6](#_Toc154053448)

[5. Корреляционный анализ 7](#_Toc154053449)

[6. Построение регрессионной модели 8](#_Toc154053450)

[Исследование модели 9](#_Toc154053451)

[7. Анализ полученных данных модели 9](#_Toc154053452)

[8. Исследование мультиколлинеарности факторов. 10](#_Toc154053453)

[Программная реализация 10](#_Toc154053454)

[9. Описание выбора и описание программного обеспечения. 10](#_Toc154053455)

[Выводы 11](#_Toc154053456)

[Источники 12](#_Toc154053457)

# **Введение**

Современные города сталкиваются с вызовами, связанными с постоянным ростом населения и увеличением транспортной мобильности. В контексте города Москвы, являющегося одним из крупнейших мегаполисов мира, эффективное управление пассажиропотоком в общественном транспорте становится стратегическим приоритетом для обеспечения комфортной и эффективной передвижения граждан.

Объем пассажиропото ка не только отражает динамику городской активности, но и является ключевым показателем, влияющим на планирование и развитие транспортной инфраструктуры. В этом контексте, разработка надежных моделей прогнозирования пассажиропотока становится неотъемлемой частью стратегии управления транспортной системой.

Целью данной курсовой работы является разработка модели прогнозирования пассажиропотока в транспорте Москвы. Эта задача включает в себя анализ существующих данных о пассажиропотоке, разработку математических моделей прогнозирования и оценку их эффективности. Полученные результаты позволят не только предсказывать изменения в пассажиропотоке, но и оптимизировать управление общественным транспортом, повышая качество услуг и удовлетворение потребностей городского населения.

Исследование в данной области не только актуально с практической точки зрения для городского планирования, но и вносит вклад в развитие теории прогнозирования пассажиропотока в контексте мегаполисов. Кроме того, результаты могут быть использованы для создания более устойчивых и эффективных городских транспортных систем, что является важным аспектом развития современных городов.

# **Анализ характеристик объекта исследования**

1. **Описание объекта исследования**

За объект исследования было решено взять данные из метрополитена города Москвы, которые выглядят следующим образом.

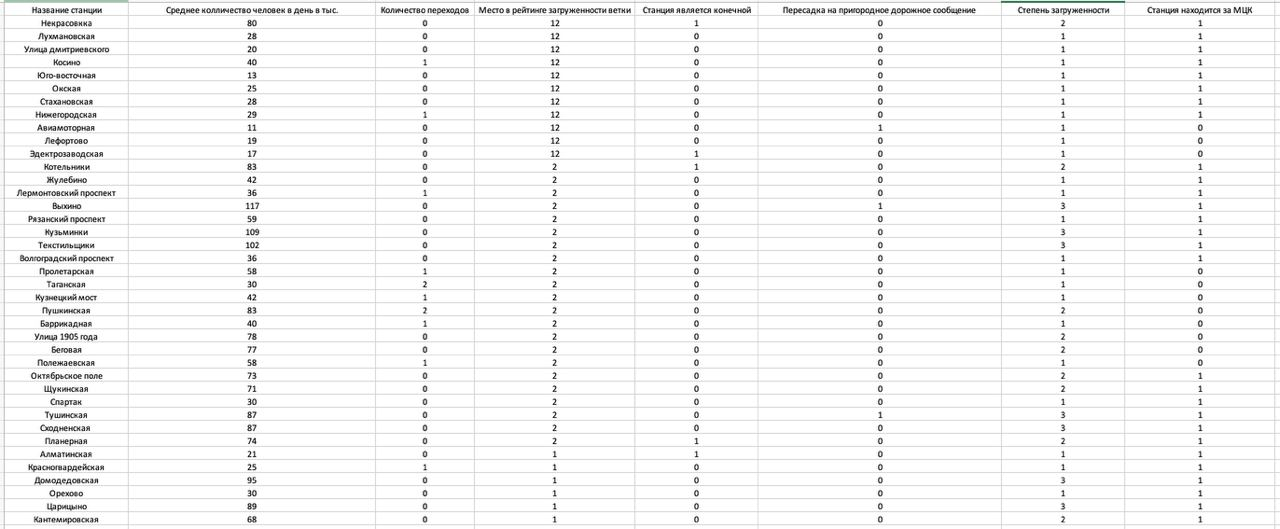


Рисунок 1. Вырезка из датасета объекта исследования

Объект исследования представляет собой набор данных, описывающих различные станции в системе общественного транспорта или железнодорожной сети. Каждая строка данных представляет отдельную станцию, а столбцы представляют различные характеристики и параметры, связанные с этими станциями. Вот краткое описание каждого столбца:

* Название станции: Содержит названия станций общественного транспорта или железнодорожных станций, которые могут включать в себя информацию о местоположении или других характеристиках.
* Среднее количество человек в день в тыс.: Показывает среднюю посещаемость станции в тысячах человек в день, что может быть важным параметром для оценки загруженности станции.
* Количество переходов: Отражает количество переходов (подземных или наземных) на данной станции. Это может влиять на общую доступность и удобство использования станции.
* Место в рейтинге загруженности ветки: Указывает на позицию данной станции в рейтинге по загруженности на соответствующей транспортной ветке. Это может быть важным для определения важности станции в сети общественного транспорта.
* Станция является конечной: Булева переменная, которая указывает, является ли данная станция конечной точкой (конечной станцией) для определенной линии или ветки общественного транспорта.
* Пересадка на пригородное дорожное сообщение: Индикатор того, предоставляет ли данная станция возможность пересадки на пригородные поезда или другие виды пригородного транспорта.
* Степень загруженности: характеризует уровень загруженности станции, на основе на основе критерия (ЕСЛИ(xi>85; 3; ЕСЛИ(xi>60; 2; 1))).
* Станция находится за МЦК: говорит о том находится ли станция внутри Московского Центрального Кольца.

Эти параметры могут быть использованы для анализа и оптимизации работы системы общественного транспорта, оценки эффективности станций, а также для принятия решений в плане развития и улучшения общественного транспорт.

1. **Анализ объекта исследования с помощью статистических показателей и выявление связей**

Ознакомимся с рейтингом загруженности линий метрополитена, на который отсылается наш датасет. Из статистических показателей видно, что загруженность станции может коррелировать с тем, на какой линии метро находится станция.

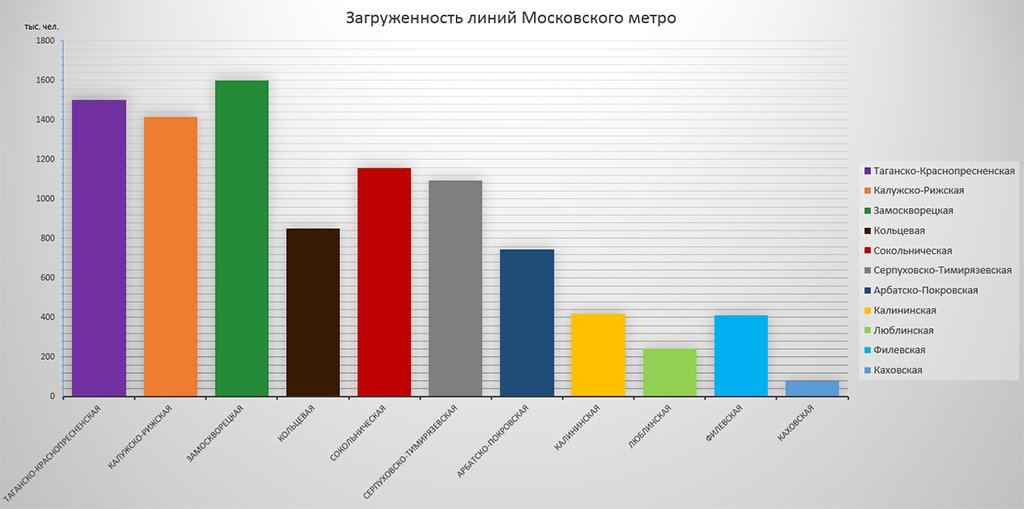


Рисунок 2 Рейтинг загруженности линий Московского метро

Ознакомимся с местонахождением станций и количеством человек в день на каждой, из наглядного анализа можно понять, что возможна корреляция загруженности и нахождения станции вне МЦК, количества пересадок.

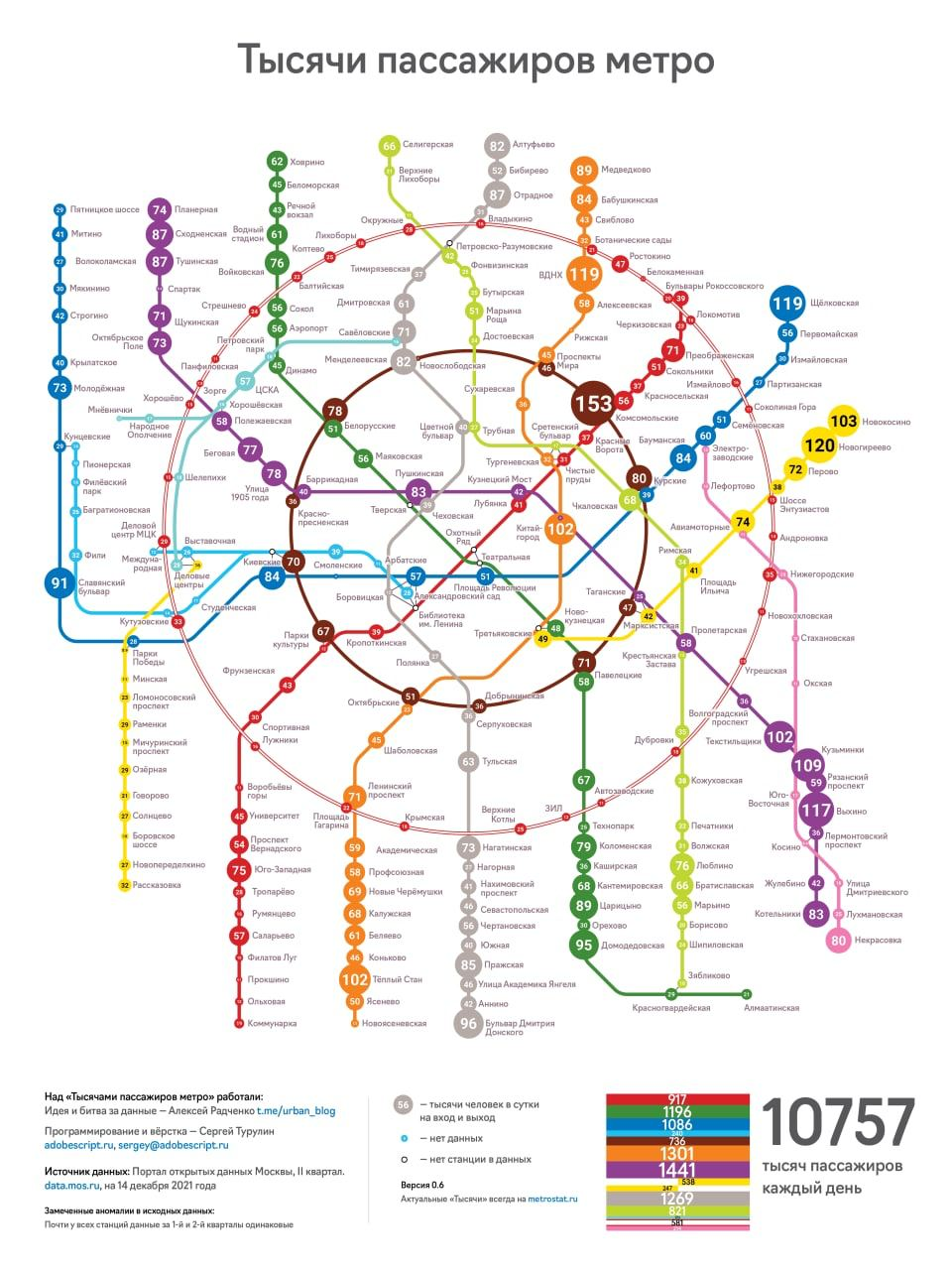


Рисунок 3 Количество человек на каждой из станций московского метро.

После зрительного анализа перейдем к постановке задачи и выберем критерии, по которым будем обучать модель

1. **Постановка задачи моделирования.**

Внимательно проанализировав данные, было выбрано решение обучить модель предсказывать загруженность станций по трехбалльной системе, таким образом мы сможем добиться более точных результатов и более эффективно предсказывать загруженность станций метро. Трехбалльная система построена следующим образом:

* (3 балла) количество человек в день на станции превышает 85 тыс.
* (2 балла) количество человек в день на станции превышает 60 тыс.
* (1 балл) количество человек в день не превышает 60 тыс.

Для написания модели будем использовать библиотеку sklearn на языке Python,

модель будем считать успешной если вероятность предсказание будет больше 65%

# **Моделирование статистических зависимостей**

1. **Проверка гипотезы о нормальном распределении**

Построили гистограмму для столбца «Среднее количество человек в день в тыс.»

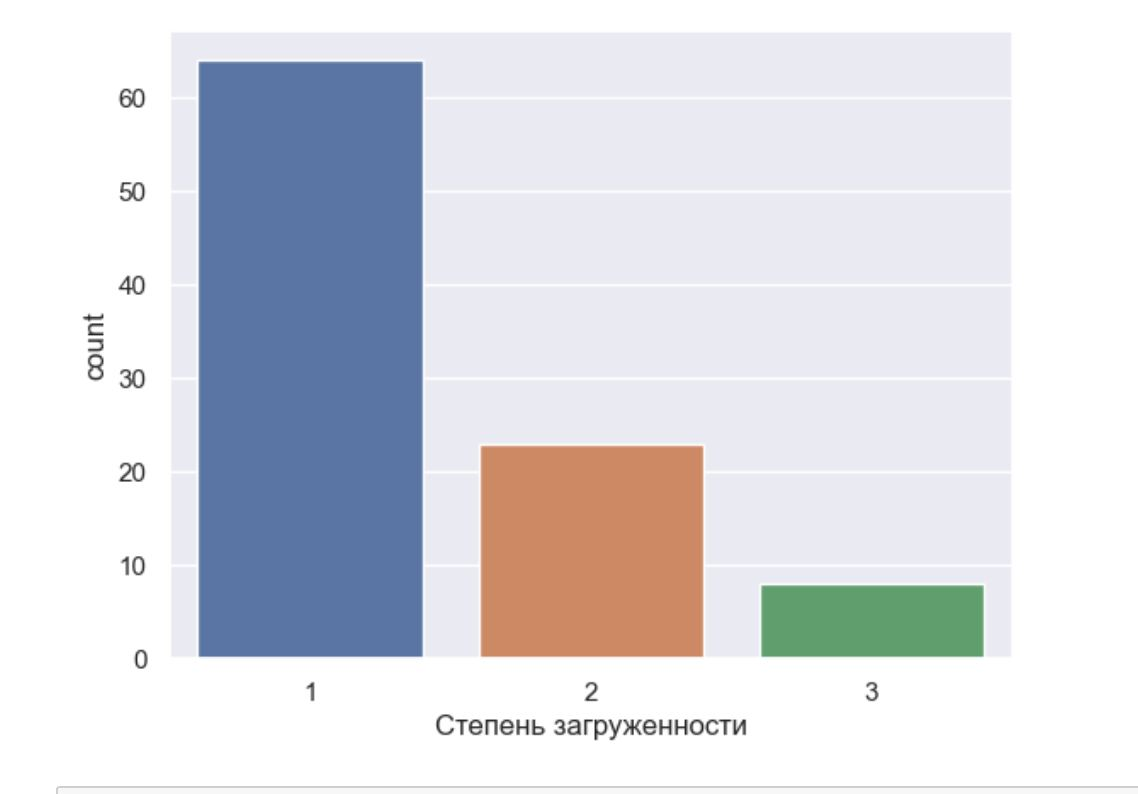
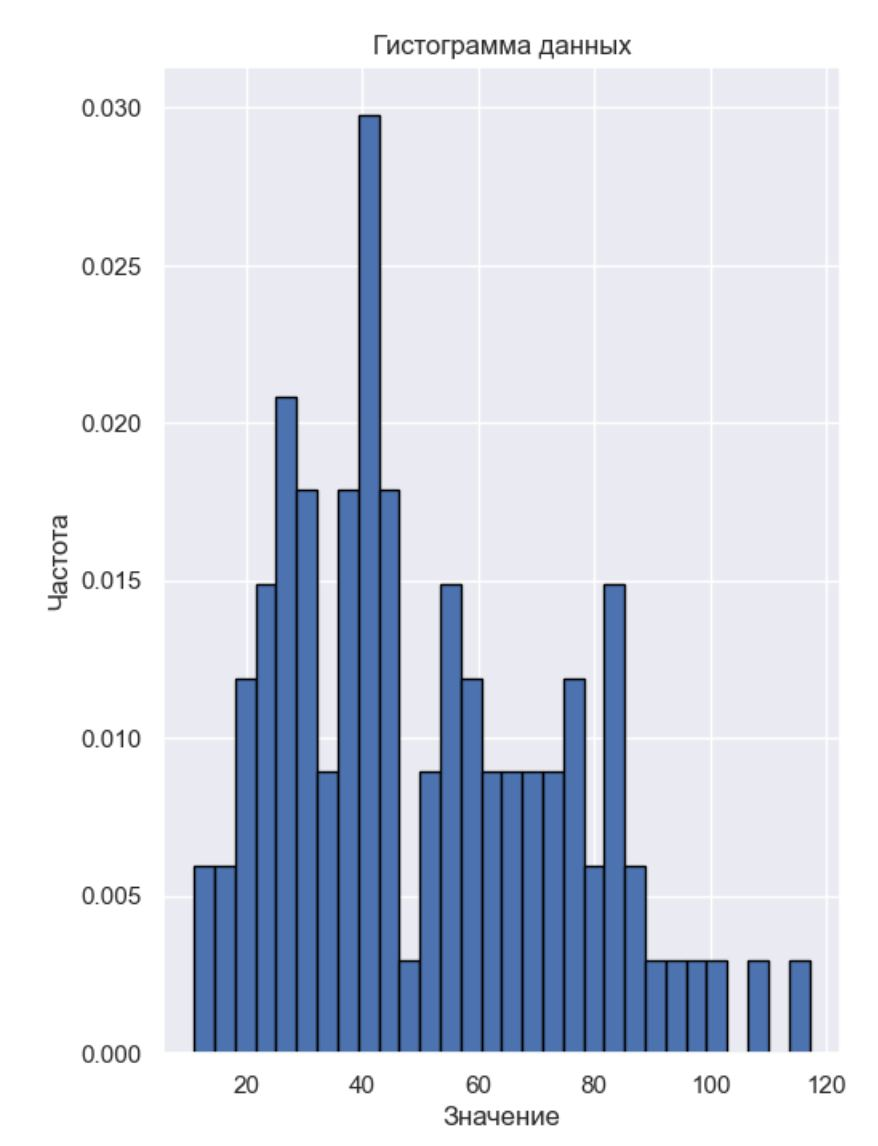


Рисунок 4 Распределение величин столбца «Среднее количество человек в день в тыс.» и «Степень загруженности»

Гистограммы имеют несимметричную форму, отвергаем гипотезу о нормальном распределении

1. **Корреляционный анализ**

Построим таблицу корреляции всех наших данных для установления зависимостей между значениями наших данных в датасете. Поскольку нас интересует столбец «Среднее количество человек в день в тыс.», будем смотреть на корреляцию остальных столбцов к этому.

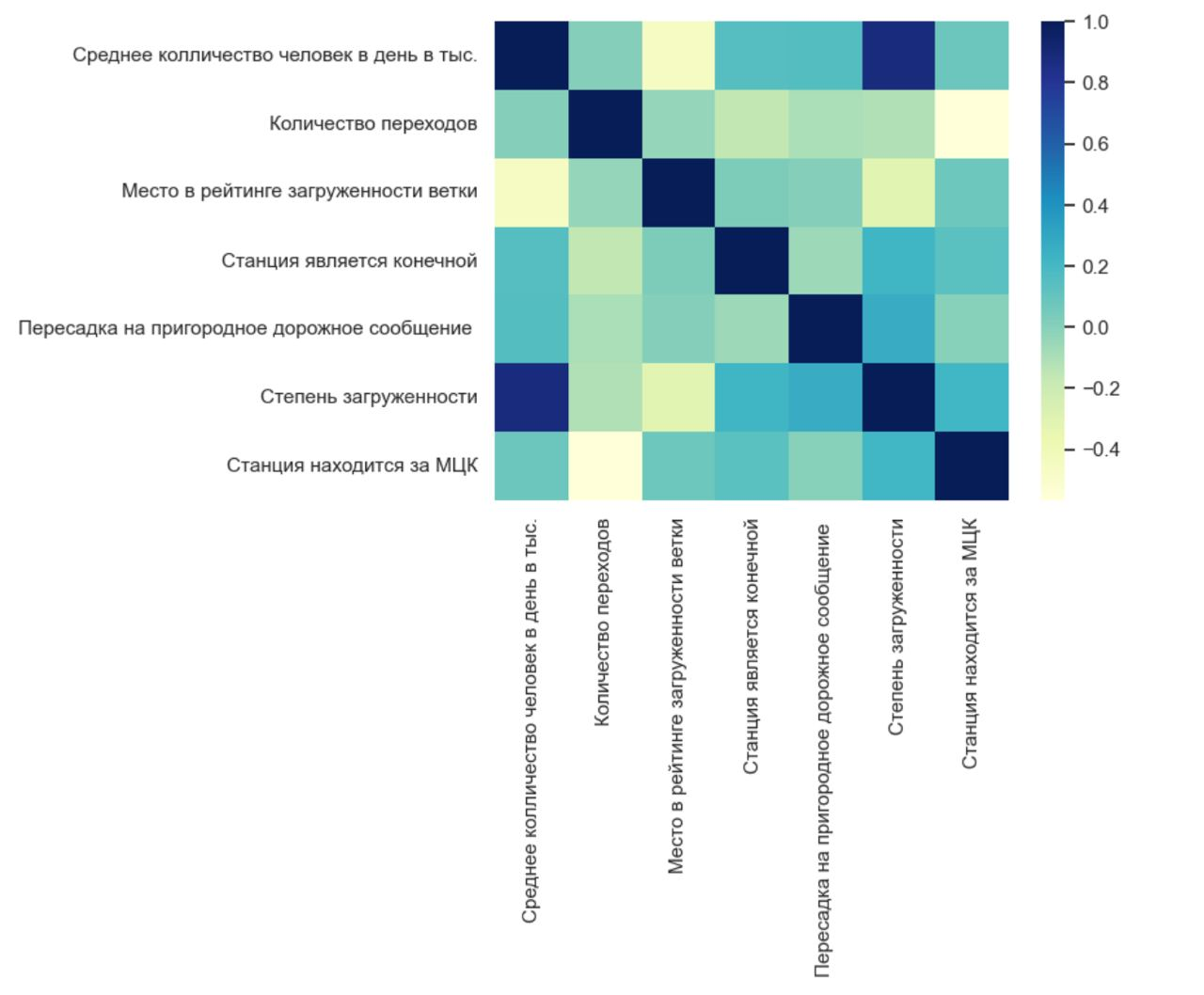


Рисунок 5 Таблица корреляции

По таблице видно, что степень загруженности станции сильно коррелирует со средним количеством человек в день на станции поскольку эти являются зависимыми. Наравне коррелируют столбцы «Станция является конечной» и «Пересадка на пригородное дорожное сообщение», отрицательную корреляцию можно увидеть со столбцом «Место загруженности ветки», поскольку чем больше человек на линии метро, тем выше она в рейтинге, тем ниже цифра места.

1. **Построение регрессионной модели**

Для постановки регрессионной модели используем библиотеку sklearn для языка Python. Прокомментировали каждую часть кода. Первым шагом разделим данные на те, которые будем прогнозировать в переменную Y, и на те, которые помогут модели в предсказании столбца Y. Удалим столбцы «Название станции», поскольку этот столбец является бесполезным для нашей математической модели так как он никаким образом не может коррелировать со степенью загруженности, «Степень загруженности» поскольку этот столбец мы переместили в Y, «Среднее количество человек в день в тыс.» поскольку это зависимые данные со степенью загруженности.

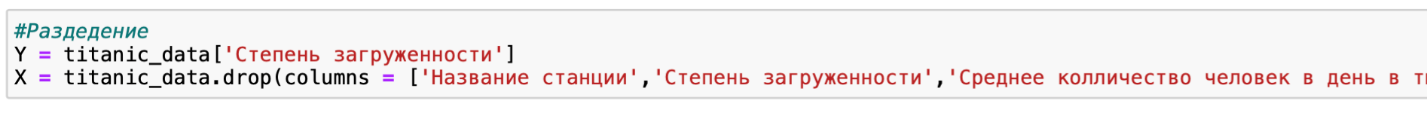
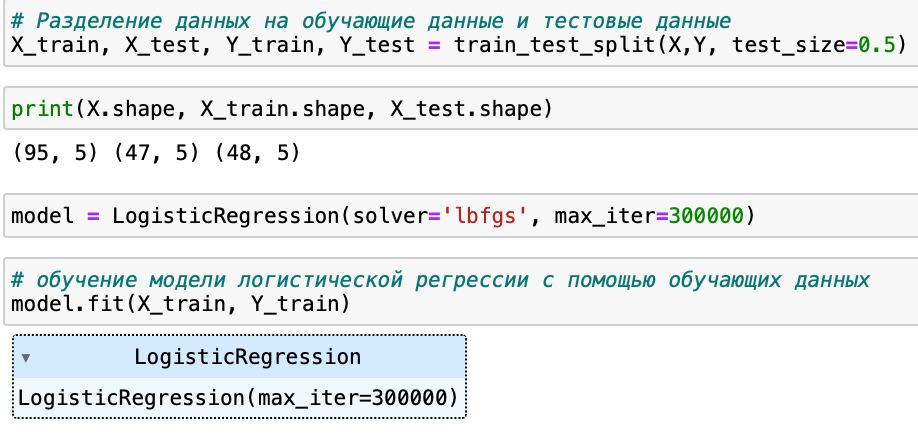


Рисунок 6 Разделение данных, удаление ненужных столбцов

Разделили данные на тестовые и обучающие пополам, это означает что мы даем нашей модели в обучение 50% нашего датасета, а на остальных 50% мы будем проводить тест ответов нашей модели.

Рисунок 7 Построение регрессионной модели

# **Исследование модели**

1. **Анализ полученных данных модели**

После прогонки нашей модели порядка десяти раз, сделали вывод, что средняя точность предсказывания – 66.6%, следовательно модель считается успешной и цель работы была выполнена.



Рисунок 8 Анализ полученных данных модели

1. **Исследование мультиколлинеарности факторов.**

Мультиколлинеарность — это явление, при котором два или более фактора в модели машинного обучения сильно коррелированы между собой, что может вызывать проблемы при интерпретации результатов и оценке влияния каждого фактора на зависимую переменную.

В нашем случае, зависимые переменные были «Степень загруженности» и «Среднее количество человек в день в тыс.», мы избавились от мультиколлинеарности на стадии подготовки модели, удалив столбец ««Среднее количество человек в день в тыс.».

# **Программная реализация**

1. **Описание выбора и описание программного обеспечения.**

В процессе разработки модели прогнозирования пассажиропотока в городском общественном транспорте в Москве было принято решение использовать язык программирования Python в связке с несколькими ключевыми библиотеками для обработки данных, визуализации и построения модели.

Использованные библиотеки:

* NumPy: NumPy предоставляет мощные функции для работы с многомерными массивами и математическими функциями. Был использован для эффективной работы с числовыми данными.
* Pandas: Pandas предоставляет удобные структуры данных и функции для анализа и манипуляции табличных данных. С его помощью была осуществлена предобработка и анализ данных.
* Matplotlib и Seaborn: Эти библиотеки были использованы для визуализации данных. Matplotlib предоставляет базовые графические возможности, а Seaborn — дополнительные стили и улучшенные функции визуализации.
* Scikit-learn (sklearn): Scikit-learn предоставляет широкий набор инструментов для машинного обучения, включая модели регрессии, классификации, кластеризации и инструменты для оценки моделей. Был использован для построения и оценки модели прогнозирования.

Выбор языка программирования:

Python был выбран в виду своей популярности, обширного сообщества разработчиков и богатого экосистемы библиотек для анализа данных и машинного обучения. Python также предоставляет удобство и читаемость кода, что является важным фактором в разработке и исследовательском анализе данных.

Обоснование выбора:

Эффективность и гибкость: Python обеспечивает высокую эффективность в разработке, а его гибкость позволяет легко интегрировать различные библиотеки и инструменты.

* Машинное обучение в одном месте: Библиотеки NumPy, Pandas и Scikit-learn предоставляют все необходимые инструменты для работы с данными и построения моделей машинного обучения.
* Визуализация данных: Matplotlib и Seaborn предоставляют мощные средства визуализации данных, что важно для понимания распределения и характеристик данных.

Этот стек программного обеспечения был выбран с целью обеспечения эффективной разработки, четкой визуализации данных и построения точных моделей прогнозирования пассажиропотока в городском общественном транспорте Москвы.

# **Выводы**

В ходе исследования прогнозирования пассажиропотока в городском метрополитене Москвы был проведен комплексный анализ, включающий сбор и предобработку данных, построение модели прогнозирования, а также исследование мультиколлинеарности факторов. Все эти этапы позволили получить более глубокое понимание динамики пассажиропотока и факторов, влияющих на него.

* **Сбор и предобработка данных:**

Эффективное сбор данных о пассажиропотоке и внешних факторах, таких как погодные условия и события, позволило создать базу данных, достаточно репрезентативную для анализа.

* **Построение модели прогнозирования:**

Выбранная модель, опирающаяся на методы машинного обучения, дала возможность эффективного прогнозирования пассажиропотока с учетом различных факторов в городском общественном транспорте.

* **Анализ мультиколлинеарности:**

Исследование мультиколлинеарности выявило наличие сильной корреляции между некоторыми факторами, что могло повлиять на стабильность и интерпретируемость модели.

* **Оптимизация модели:**

Путем удаления избыточных факторов и использования методов регуляризации была достигнута оптимизация модели, сохраняя при этом ее способность точно предсказывать пассажиропоток.

* **Выводы и перспективы:**

Работа предоставляет фундаментальные выводы по динамике пассажиропотока и факторам, влияющим на него. Дальнейшие исследования и улучшения модели могут включать в себя учет дополнительных факторов, более сложные методы обработки данных и периодическое обновление модели с учетом изменяющихся условий.

В итоге, данная работа не только предоставляет практическую модель для прогнозирования пассажиропотока в городском общественном транспорте Москвы на примере метрополитена, но и раскрывает важность анализа мультиколлинеарности в процессе построения таких моделей, призывая к постоянному совершенствованию и адаптации к изменяющимся условиям.

# **Источники**

* <https://www.mosoblreclama.ru/auxpage_passazhiropotok_metro>
* <https://moskvichmag.ru/gorod/novaya-karta-tysyachi-passazhirov-metro-pokazyvaet-zagruzhennost-pochti-kazhdoj-stantsii/>
* <https://www.rus-metro.ru/russia/moscow/statisticheskie-dannye.htm>
* https://tproger.ru/articles/gajd-po-obrabotke-dannyh-s-pomoshhyu-pandas-chast-pervaya
* https://datastart.ru/blog/read/chto-takoe-scikit-learn-gayd-po-populyarnoy-biblioteke-python-dlya-nachinayuschih