Национальный исследовательский университет
«Высшая школа экономики»
Факультет городского и регионального развития
Бакалаврская программа "Городское планирование"

Отчет о самостоятельной работе
по дисциплине «Анализ количественных данных в социальных науках»
по теме “Анализ нагрузки на сеть московского метрополитена”

Группа БГП203
Комаров Олег Александрович
Ледовских Валерия Александровна
Репина Елизавета Андреевна

Москва 2022

Оглавление

1. Общая постановка задачи

Наше исследование направлено на изучение количественных метрик, таких как количество входов и выходов на всех станциях московского метрополитена за I и II кварталы 2022 года. Этот анализ позволит сделать выводы о нагрузке московского метрополитена и предложить альтернативные варианты его развития или же подкрепить гипотезы и результаты существующим проектом развития метро до 2030 года.

Анализ проводится по следующей схеме:

1. На естественном языке формулируется ряд гипотез об указанной взаимосвязи. В состав гипотез входят как простые гипотезы (о направлении связи), так и сложные гипотезы, учитывающие нелинейный характер связи.
2. Происходит сбор данных о входах и выходах пассажиров на каждой станции метро. Выбираются качественные и количественные переменные, определяется их зависимость и независимость.
3. Проводится предварительный анализ данных путем построения гистограмм, диаграмм рассеивания и простой графической визуализации.
4. Строится, т.е. специфицируется и оценивается, базовая модель — модель множественной линейной регрессии целевой переменной на объясняющие. Анализируются ее свойства. При необходимости корректируется состав объясняющих переменных.
5. Выполняется пошаговая корректировка спецификации базовой модели для учета всех возможных комбинаций сформулированных сложных гипотез. Модифицированные модели оцениваются и выполняется проверка соответствующих гипотез.
6. На основании стандартных критериев анализируется качество всех построенных моделей, включая базовую, и выбирается наилучшая. Происходит окончательное подтверждение или опровержение гипотез.

1.1. Описание прикладной области данных

Исследование имеет прикладную направленность. В ходе него будет изучена нагрузка на сеть московского метрополитена. Мы ставим перед собой цель изучить пассажиропоток за I и II кварталы 2022 года и гипотетическую неравномерность его распределения между разными частями столичного метро для составления плана дальнейшего развития инфраструктуры с учетом текущих особенностей сети и сравнить получившиеся выводы с актуальным планом развития московского метрополитена. Теоретически, построенный в ходе работы план может служить альтернативным планом развития метро при несовпадении с актуальным, или основой для дополнительных мер по повышению эффективности сети в обратном случае.

Все данные собраны из открытых источников:

- Интернет ресурс Реклама в метро Москвы / rus-metro.ru [Электронный

ресурс]. Режим доступа: https://www.rus-metro.ru/russia/moscow/statisticheskie-dannye.htm#

- Интернет ресурс Мособлреклама / mosoblreclama.ru [Электронный ресурс]. Режим доступа: https://www.mosoblreclama.ru/auxpage_passazhiropotok_metro
- Интернет ресурс Портал открытых данных / data.mos.ru [Электронный ресурс]. Режим доступа: https://data.mos.ru/opendata/6274

Переменная	**Тип переменной**	**Зависимость**	**Целевая/объясняющая**
Наименование ветки метрополитена	Качественная	Независимая	Объясняющая
Вход пассажиров	Количественная	Зависимая	Целевая
Выход пассажиров	Количественная	Зависимая	Целевая
Численность населения района	Количественная	Независимая	Объясняющая
Округ	Качественная	Независимая	Объясняющая
Численность населения округа	Количественная	Независимая	Объясняющая
Количество пересадок в ТПУ	Количественная	Независимая	Объясняющая
Протяженность ветки	Количественная	Независимая	Объясняющая
Нахождение в ЦАО	Качественная категориальная	Независимая	Объясняющая

Таблица 1. Описание переменных

1.2. Основные гипотезы, которые планируется проверить в рамках исследования

Для исследования были составлены 4 гипотезы, 2 из которых простые и 2 - сложные. Гипотезы затрагивают все объясняющие переменные и устанавливают связи между ними.

Простая гипотеза: Пассажиропоток на станции зависит от численности

населения округа, в котором находиться станция, от расположения в пределах или за границами ЦАО, от количества пересадок на другие ветки.

Простая гипотеза: Нагрузка на ветку метро в целом растет с ростом ее протяженности.

Сложная гипотеза: Количество пересадок больше влияет на входы в пределах ЦАО.

2. Предварительный анализ данных

2.1 Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выбросы

2.1.1 Анализ количественных переменных

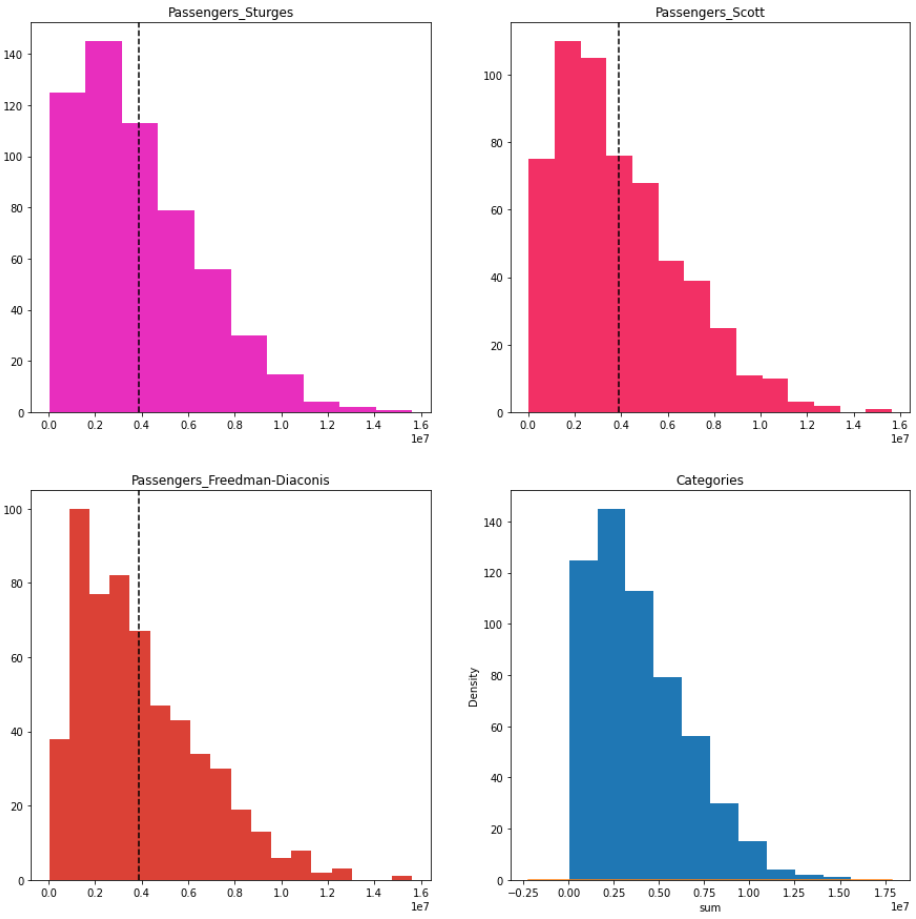


Рисунок 1. Частота станций с разным диапазоном пассажиропотока

На гистограмме показана частота станций с определенным диапазоном пассажиропотока. Было использовано 4 вида гистограмм, построенных по разным моделям, все они отражают приблизительно одинаковую картину. Здесь видно одномодальное распределение в трех случаях и в гистограмме Фридмана Диакониса есть незначительная бимодальность. Везде наблюдается сдвиг вправо в область больших значений. Присутствует выброс по модели Скотта и

Фридмана Диакониса. Видно, что распределение не приближено к нормальному и есть перевес в первой четверти, что означает, что большинство станций имеют не самую большую нагрузку, а станций, где пассажиропоток в 4-5 раз больше - гораздо меньше.

Для того чтобы понять, какие станции являются самыми нагруженными и где они расположены, необходимо проанализировать население округов и количество станций в каждом округе.

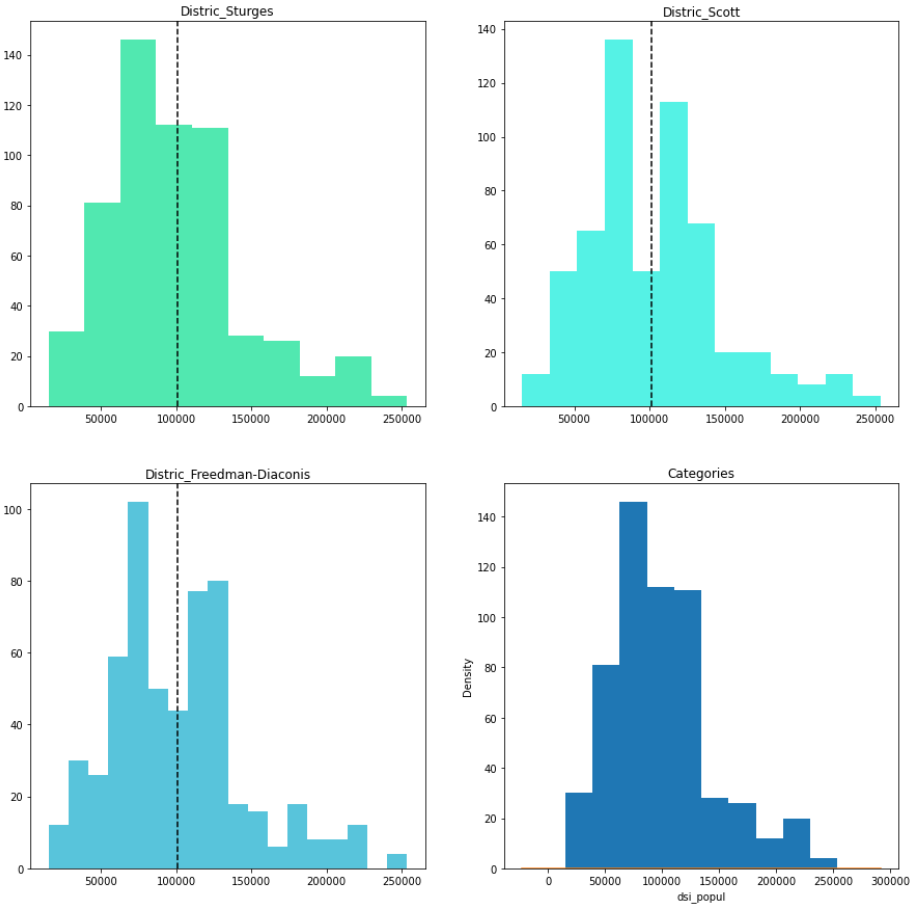


Рисунок 2. Распределение населения районов

На гистограмме показано распределение населения районов. Для каждой станции метро был прописан район, в котором она находится и его население, поэтому частота здесь связана с количеством упоминаний каждого района за весь период наблюдений. В модели Скотта и Фридмана Диакониса видна бимодальность гистограммы, по Стерджесу полимодальности нет. Также всем гистограммам свойственна асимметрия вправо. Вид их далек от нормального распределения, есть выбросы. Все это может говорить о неравномерном распределении населения по районам Москвы или заметной дифференциации между ними в обеспеченности станциями метро. Исходя из базовых знаний предметной области, мы склоняемся ко второму объяснению.

В дальнейшем анализ названий районов использоваться не будет, а рассмотрение нагрузки будет исключительно через количественный показатель населения районов (из-за слишком высокого уровня уникальных качественных переменных, создаваемых из названий в модели, для которых ей не хватает объема данных), а также на уровне округов и их населении.

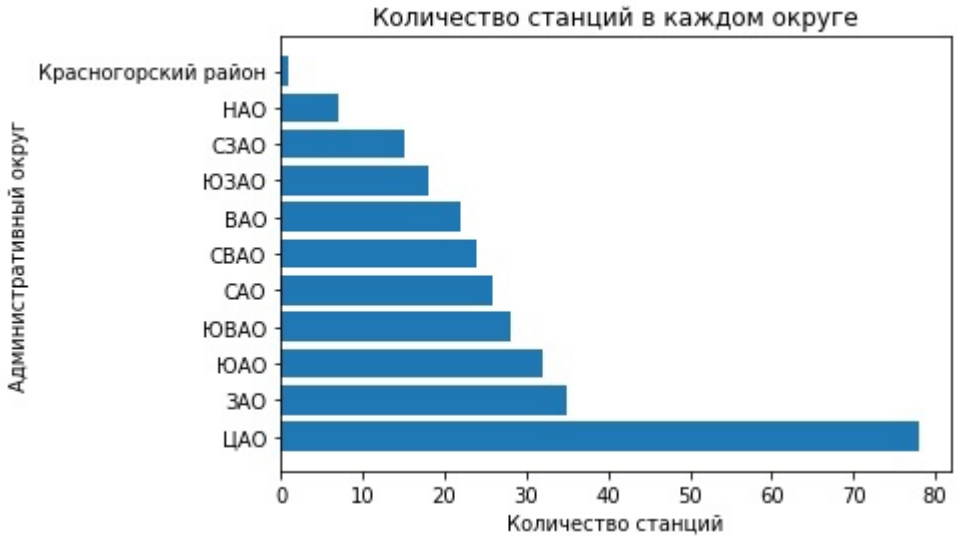


Рисунок 3. Количество станций в округе

На диаграмме показано население округов, в которых есть станции метро. Видно, что задействованы не все округа Москвы, отсутствует Троицкий административный округ, однако есть район из Московской области. Наибольшее население в целом сосредоточено в южной и восточной частях Москвы, в силу этого далее будет рассмотрено количество станций в каждом из округов, а также косвенная нагрузка на каждый из округов относительно их населения.

2.1.2 Анализ качественных переменных

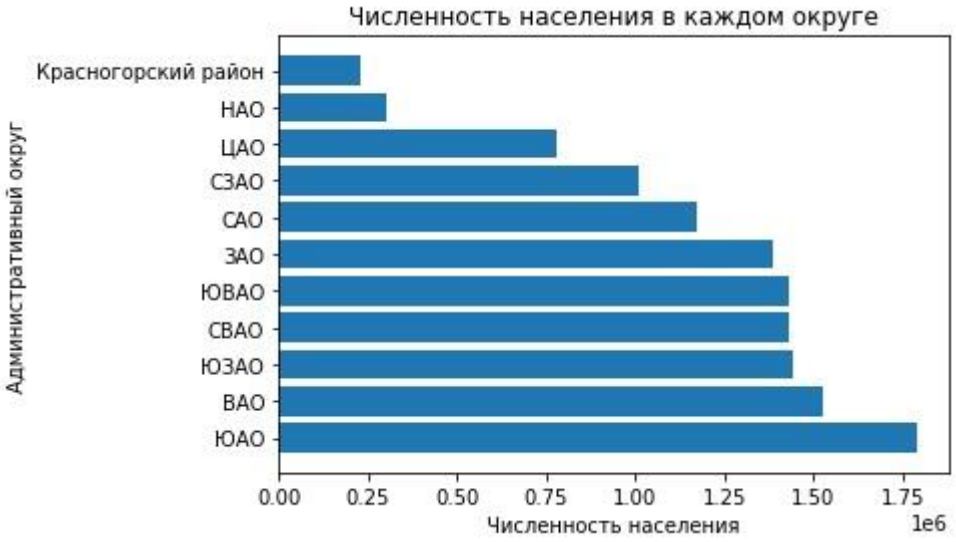


Рисунок 4. Численность населения округов

В этом пункте показан анализ качественных переменных. Мы составили диаграмму количества станций в каждом округе. Можно заметить, что в центральном административном округе сосредоточено больше всего станций и

как будет видно далее, пассажиропоток на станциях центрального округа гораздо выше, чем в других округах. Более того, ранее наблюдение о значительно высоком населении в южной части города связано теперь с тем, что в южных округах действительно суммарно станций больше, чем в других (за исключением ЦАО).

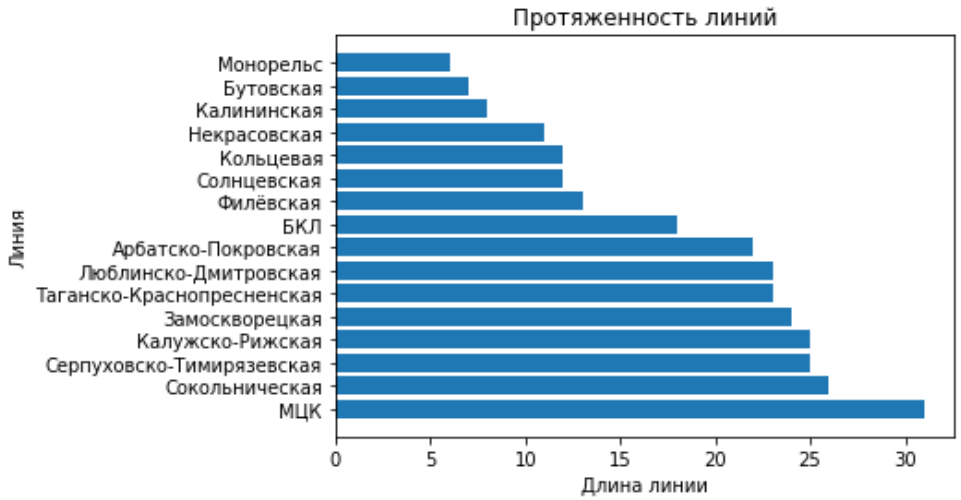


Рисунок 5. Протяженность линий метрополитена

Одна из гипотез связана с тем, что нагрузка на ветку метро в целом растет с ростом ее протяженности, чтобы проверить эту гипотезу мы создали соответствующую переменную и составили диаграмму протяженности каждой ветки. В предварительной проверке гипотез по графическим материалам, она опровергается. Заметим, что наиболее протяженными ветками являются Московское центральной кольцо и Сокольническая линия, которая уходит в Новую Москву. Монорельс выполняет экскурсионную функцию, имеет самую низкую протяженность и низкую нагрузку, Бутовская линия является продолжением Серпуховско-Тимирязевской линии и уходит за МКАД, интегрируясь в транспортную сеть отдаленные районы ЮАО. Калининская линия выполняет соединяющую функцию, она объединяет жителей ВАО с центром, обеспечивая быстрый маршрут и пересадки на все ближайшие соседние ветки.

2.2 Анализ статистической связи

2.2.1 Графический анализ пары “целевая переменная - качественная переменная”

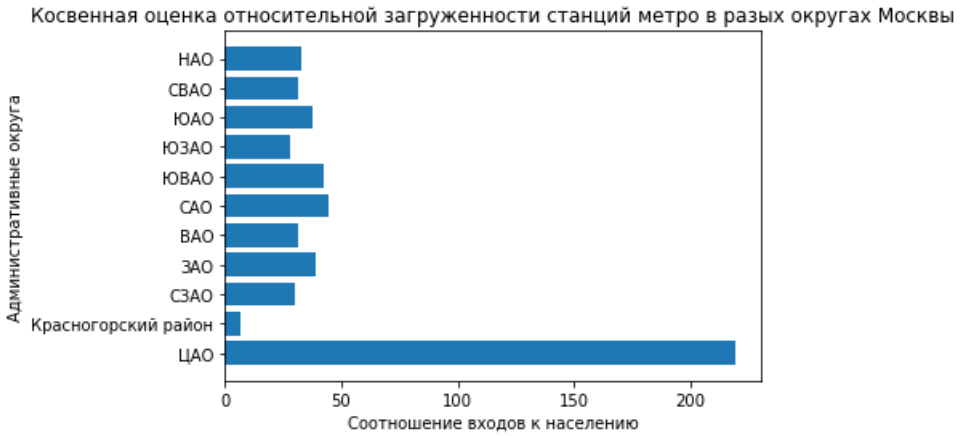
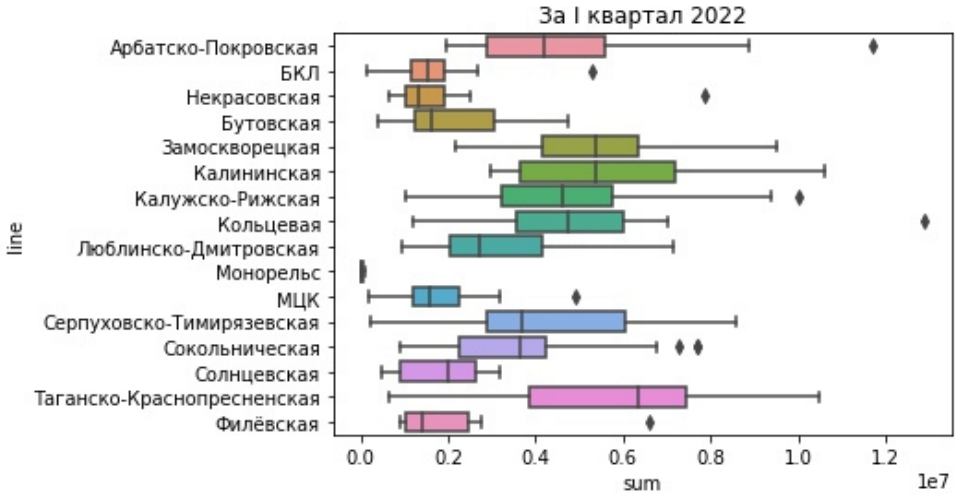


Рисунок 6. Относительная загруженность станций метро по округам

Из диаграммы видно, что самая большая нагрузка относительно населения округа в ЦАО. Это объясняется наибольшим количеством станций и не самым большим населением. Станции в Красногорском районе не принимают на себя большую нагрузку. Остальные округи имеют равномерную и почти одинаковую нагрузку относительно своего населения. Эта диаграмма стала причиной формулирования сложной гипотезы о том, что наличие пересадок и расположение в пределах ЦАО сильнее влияет на рост пассажиропотока, чем



население округа.

Рисунок 7. Загруженность линий за I квартал 2022

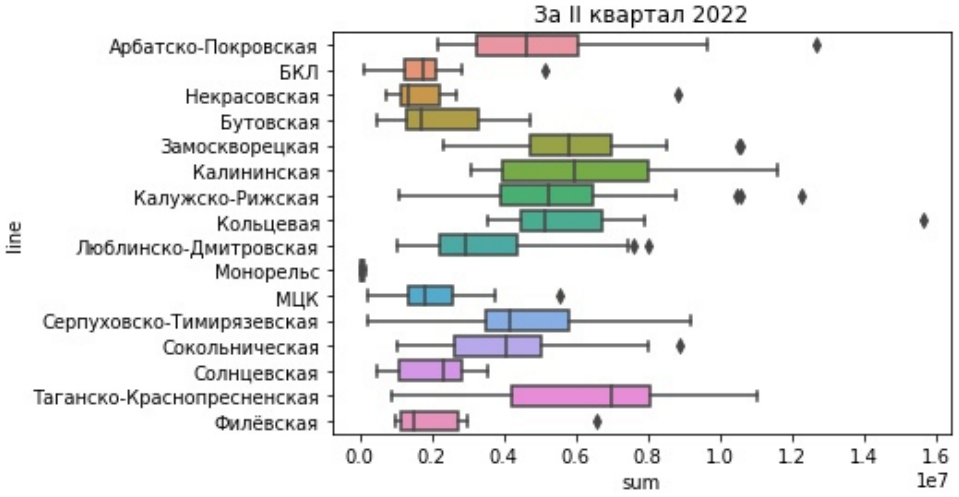


Рисунок 8. Загруженность линий за II квартал 2022

Самая загруженная ветка - Таганско-Краснопресненская. Далее - Калининская и Замоскворецкая. Наибольший выброс - кольцевая линия.

Монорельс, Некрасовская, Большая кольцевая линия - наименее загруженные (БКЛ находится в процессе строительства). Некрасовская линия является ответвлением Таганско-Краснопресненской линии (как Бутовская и Серпуховско-Тимирязевская), поэтому нагрузка с Некрасовки перетекает на Таганско-Краснопресненскую линию в станциях Косино - Лермонтовский проспект.

2.2.2 Графический анализ пары «числовая зависимая переменная – числовая независимая переменная»

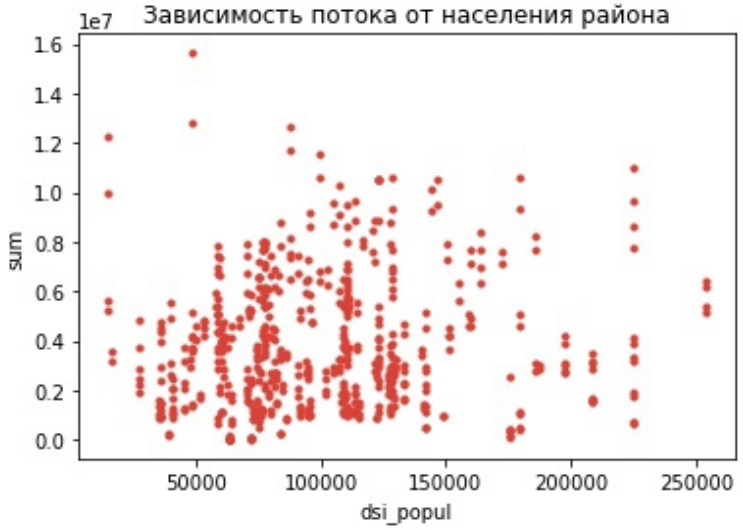


Рисунок 9. Зависимость пассажиропотока от населения района

По диаграмме рассеивания видно, что население района слабо влияет на увеличение пассажиропотока. Безусловно видны выбросы, в которых

густонаселенные районы обладают малым пассажиропотоком, но также есть и обратное наблюдение о высоком пассажиропотоке при малом населении.

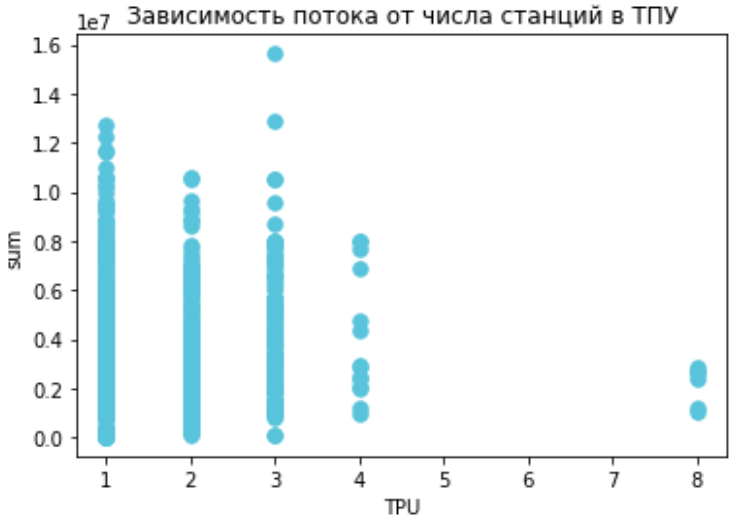


Рисунок 10. Зависимость пассажиропотока от числа пересадочных станций

Число станций в ТПУ - это количество пересадочных станций, в том числе кросс-платформенные и наземные пересадки. Можно заметить, что фактор увеличения станций пересадки не влияет на увеличение пассажиропотока, а скорее имеет обратную зависимость. Есть выброс в 3 пересадках, где достигается максимальное значение пассажиропотока. Также видно, что распределение загруженности для станций с 1 пересадкой более равномерное, чем для остальных.

2.2.3 Предварительная проверка гипотез

Гипотеза о прямой зависимости протяженности линии и ее нагруженности требует проверки моделированием, так как графический материал может дать неточный результат. Наиболее протяженные линии (МЦК, Сокольническая) далеко не самые нагруженные. Калининская линия, которая является одной из самых коротких, наоборот принимает на себя поток в разы больше, чем Сокольническая линия или МЦК.

Гипотеза про наличие пересадок и населения округа опровергается предварительным анализом графических материалов и далее будет проверяться моделированием.

3. Проверка гипотез с помощью моделирования

3.1. Построение базовой модели

Изначальный вид модели включал в себя целевую переменную в виде входов и все взятые нами показатели в виде названий станций, веток метро, дифференциации по времени на разные кварталы 2022 года, названия и населения районов и округов, где находиться станция, количество пересадок в транспортно-пересадочном узле (т.е. включая саму станцию) и протяженность ветки (в количестве станций).

Однако после формирования фиктивных переменных из всех качественных и первого прогона модели выяснилось, что такое количество переменных слишком избыточно и столь сложная модель не может дать адекватных результатов на основе базы данных в ≈600 измерений.

Поэтому модель была доработана. Из нее были убраны наименование станций, переменная, содержащая названия районов, в которых находиться станция метро, была также исключена переменная, содержащая квартал, в котором был совершен замер. Однако это исключение дало совершенно неадекватный вид модели, в котором абсолютно все объясняющие переменные оказались незначимыми.

На следующей итерации доработки модели методом подбора была исключена еще одна переменная, являвшаяся незначимой в изначальной модели - население округов, в котором находиться станция метро. Данное изменение дало наиболее устроившую нас модель из всех опробованных, она и была использована для анализа.

3.2. Проверка гипотез с помощью моделирования

Большинство наших гипотетических предположений было отброшено на этапах проверки корреляции между количественными переменными и проверке качества самой базовой модели. Так по модели Пирсона значимы оказалась корреляция между входами и населением района, в которой расположена станция, а также длина линии, к которой принадлежит станция. По Спирмену сюда же добавилось нахождение в пределах ЦАО, которая в первой модели находилось на грани попадания в 5% диапазон.

Оценка P-уровня модели также свидетельствует о высокой значимости протяженности линии. Что подтверждает нашу вторую гипотезу.

В то же самое время этих действий достаточно, чтобы опровергнуть первую гипотезу, так как часть показателей из нее оказываются сами по себе незначимыми.

Для третьей гипотезы мы прописали отдельную модель, содержащую новую переменную, которая отражает связь и влияние количества пересадок на целевую переменную в зависимости от нахождения станции в пределах ЦАО или за его границами. Итогами теста которой стало подтверждение третьей гипотезы. Притом данная переменная даже несколько улучшила качество исходной модели.

3.3. Оптимизация итоговой модели, сравнение качества моделей.

После изначального подбора базовой модели дальнейших изменений внесено не было, так как последующее исключение переменных прироста качества модели не давало, изменяя модифицированный R^2 в пределах одной тысячной и практически не меняя критерий Акаике. Соответственно, мы просто теряли объясняющую переменную без улучшения модели.

3.4. Проверка прогностических способностей модели

На финальном этапе работы с построенной моделью мы проверили ее прогностические способности, опробовав на тестовом множестве, сформированном из исходных данных на начальном этапе проекта. Для чего был построен график прогноза поведения замеров из тестового множества и графики границ доверительного интервала этого прогноза. Кроме того, мы построили график, отображающий поведение непосредственно замеров из тестового множества.

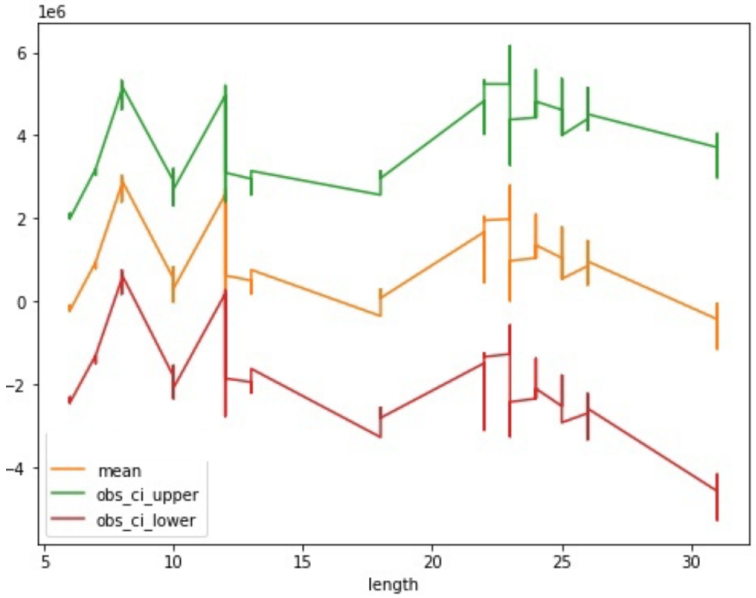


Рисунок 11. График границ доверительного интервала

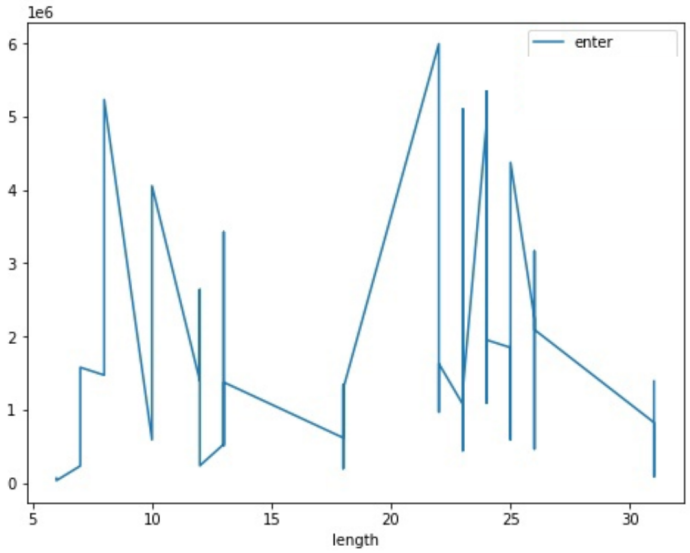


Рисунок 12. График прогноза поведения замеров из тестового множества

Как видно, прогностические способности модели находятся на удовлетворительном уровне. Большинство реальных замеров попадает в доверительный интервал модели, хотя в районе показателя длины в 22 станции наблюдается явный выход реального показания за пределы доверительного интервала.

3.5. Диагностика регрессионной модели

Таким образом, из моделирования следует, что на пассажиропоток на станциях московского метрополитена оказывает влияние протяженность линии метро, к которой относится станция, население района, в котором расположена станция, и, предположительно, на станциях, расположенных в пределах центрального административного округа пассажиропоток, также выше, чем на станциях с аналогичными исходными, но в других округах. Но, как минимум, с уверенностью можно заявить, что на пассажиропоток влияет количество пересадок со станции, если та построена в пределах ЦАО.

Кроме того, при рассмотрении P-уровня модели, построенной для работы, выясняется, что на уровень входов на станции также влияют отдельные показатели из качественных переменных, конкретные округа и ветки метро сами по себе обладают большим пассажиропотоком. Так, например, высокая значимость оказалась у СВАО и ЮВАО, а также у Таганско-Краснопресненской и Арбатско-Покровской веток метро. Все это может означать пространственную дифференциацию в городе, которая дополнительно к более общим показателям, рассмотренным в модели, влияет на потоки людей в метро.

4. Заключение

Новая схема развития метрополитена до 2030 года имеет подтверждение наших расчетов и мы нашли 3 ключевых изменения.

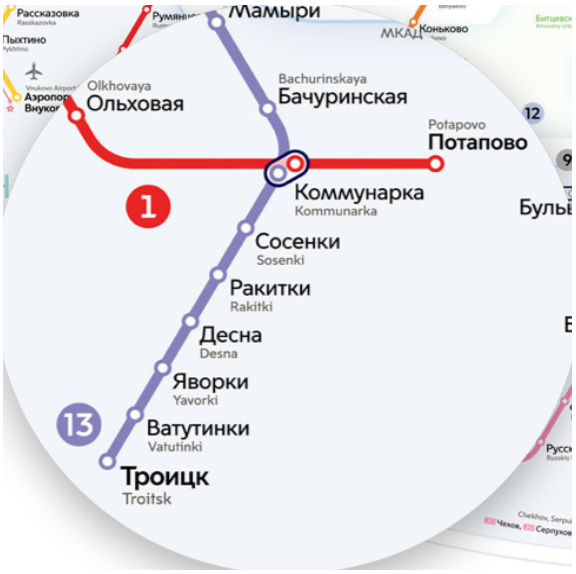


Рисунок 13. Продление 13 ветки метро

Продление 13 ветки до Троицка и включение Новой Москвы в систему метро. Это должно снизить нагрузку на станцию Коммунарка. Далее эта ветка будет проходить до станции Некрасовка.

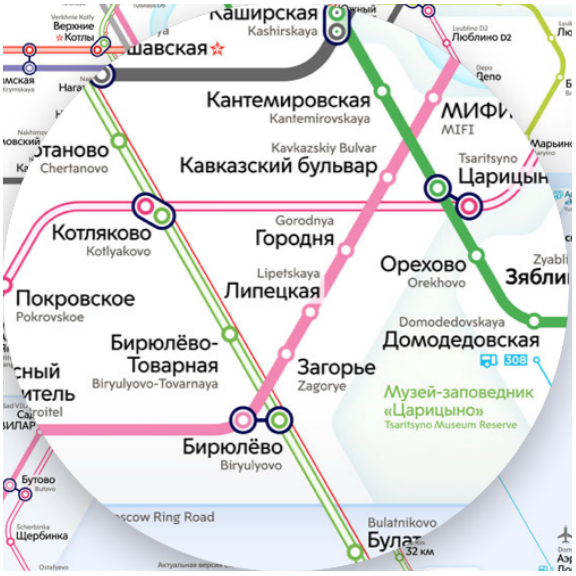


Рисунок 14. Строительство новых станций

Обеспечение отдаленных районов ЮАО станциями метро (в Бирюлево нет ни одной станции на сегодняшний день). Эта ветка будет уходить в Западное Подмосковье в район Рублево-Архангельское.

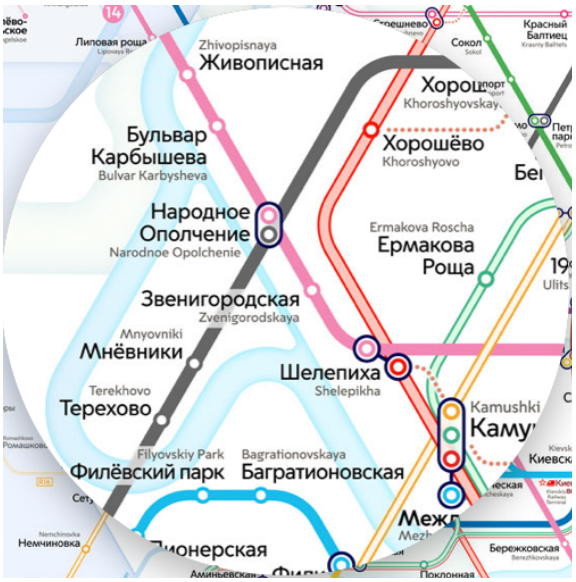


Рисунок 15. Развитие 14 ветки

Строительство станций в СЗАО и ЗАО, создание ТПУ Камушки (14 ветка из Раздолье в Рублево через Бирюлево). Станции “Бульвар Карбышева” и “Живописная” обеспечат транспортной доступностью районы на границе ЗАО и СЗАО, которые сейчас исключены из сети метрополитена из-за водных объектов.

Из 3 наших гипотез подтвердились 2, и одна подтверждения не получила. Наличие пересадок и численность населения округа не влияют на загруженность метрополитена, в то время как нахождение в ЦАО, протяженность ветки и население района дислокации становятся значимыми факторами для увеличения нагрузки на станцию.