**Национальный исследовательский университет**

**«Высшая школа экономики»**

**Московский институт электроники и математики им. А.Н. Тихонова, Москва**

Направление 09.04.01. Информатика и вычислительная техника Бакалаврская программа "Прикладная математика"

Отчет о самостоятельной работе по

дисциплине «Методы анализа стохастических взаимосвязей»

# **Бригада № 12:**

# Варфоломеева Анастасия Андреевна, 3 курс, БПМ213

# Мирный Игорь Дмитриевич, 3 курс, БПМ213

Москва 2024

# Оглавление

1. [Общая постановка задачи 3](#_30j0zll)
   1. [Описание прикладной области и данных 4](#_3znysh7)
   2. [Основные гипотезы, которые планируется проверить в рамках исследования](#_tyjcwt)

[5](#_tyjcwt)

1. [Предварительный анализ собранных данных 5](#_1t3h5sf)
   1. [Анализ особенностей данных: потенциальные ошибки и пропущенные](#_2s8eyo1) [значения, группы и выбросы 5](#_2s8eyo1)
      1. [Анализ количественных переменных 5](#_3rdcrjn)

[2.2.2. Анализ качественных переменных 7](#_lnxbz9)

* 1. [Анализ статистической связи. 8](#_1ksv4uv)
     1. [Графический анализ пары «целевая переменная – качественная](#_2jxsxqh) [объясняющая переменная» 8](#_2jxsxqh)
     2. [Графический анализ пары «числовая зависимая переменная – числовая](#_3j2qqm3) [независимая переменная» 9](#_3j2qqm3)
     3. [Анализ статистической взаимосвязи между независимыми](#_4i7ojhp) [переменными 10](#_4i7ojhp)
     4. [Предварительная проверка гипотез 10](#_1ci93xb)

1. [Проверка гипотез с помощью моделирования 11](#_2bn6wsx)
   1. [Построение базовой модели. 11](#_3as4poj)
   2. [Проверка гипотез с помощью моделирования 11](#_49x2ik5)
   3. [Оптимизация итоговой модели, сравнение качества моделей. 11](#_147n2zr)
   4. [Проверка прогностических способностей модели 12](#_23ckvvd)

[4. Заключение 12](#_32hioqz)

# Общая постановка задачи

## Описание прикладной области и данных

Данные о "прокате велосипедов", предоставляют полный набор информации, относящейся к краткосрочной аренде велосипедов компании BoomBikes из США за 2018 и 2019 годы

Данные были взяты с[Kaggle](https://www.kaggle.com/datasets/yasserh/bike-sharing-dataset), выборка состоит из 730 строк, где каждая строка соответствует каждому дню.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Характеристика объекта/явления | Название  переменной | Шкала измерения (одна из четырех) | Роль:  целевая/объясняющая |
| 1 | Время года (1 – зима,  2 – весна,  3 – лето,  4 – осень) | season | номинальная | объясняющая |
| 2 | Праздничный день (да/нет) | holiday | номинальная | объясняющая |
| 3 | Рабочий день (да/нет) | working\_day | номинальная | объясняющая |
| 4 | Погода (1 – солнечная,  2 – пасмурная,  3 – дождливая) | weathersit | порядковая | объясняющая |
| 5 | Средняя суточная температура, °C | temp | относительная | объясняющая |
| 6 | Относительная влажность воздуха | hum | относительная | объясняющая |
| 7 | Скорость ветра, км/ч | windspeed | относительная | объясняющая |
| 8 | Количество зарегистрированных поездок | cnt | относительная | целевая |

Таблица 1. Описание факторов, учтенных в анализе.

## Основные гипотезы, которые планируется проверить в рамках исследования

Простая гипотеза:

При увеличении температуры увеличивается количество поездок. Это связано с тем, что для велосипедиста улучшается условия поездки, соответственно его желание совершить поездку увеличится.

Сложные гипотеза:

В пасмурную погоду с увеличением скорости ветра количество поездок уменьшается сильнее, чем в ясную погоду. Это связано с тем, что при высокой скорости ветра в пасмурную погода вероятно дальнейшее ее ухудшение, что может насторожить потенциального клиента.

При низкой температуре воздуха количество поездок в рабочий день будет больше, чем в выходной день с такой же температурой воздуха. Мы связываем эту гипотезу с тем, что в рабочий день велосипед рассматривают как вид транспорта, а в выходные – как развлечение. В связи с этим клиент может совершить поездку в менее приятных для себя условиях.

Здесь необходимо сформулировать три гипотезы о статистической взаимосвязи целевой переменной и объясняющими. Гипотезы могут быть простыми и сложными. Простые гипотезы формулируются как предположения о корреляционной связи (направлении влияния) – «с ростом независимой переменной зависимая переменная растет/уменьшается». Сложная гипотеза содержит предположение о зависимости корреляционной связи от значений некоторых переменных в число которых может входить и рассматриваемая. В частности, это может быть гипотеза об изменении направления корреляционной связи или об изменении ее силы. Например — «до определенного возраста доход возрастает, а после него не меняется или даже снижается». Это гипотеза о существовании «пика карьеры». Или, «с возрастом скорость роста заработной платы для мужчин не равна скорости роста зарплаты для женщин». Это гипотеза о гендерном неравенстве в карьерном росте. В число гипотез должно входить не более одной простой гипотезы.

# Предварительный анализ собранных данных

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

* + 1. Анализ количественных переменных

Здесь необходимо построить и проанализировать гистограммы для всех количественных (интервальных и относительных) переменных в анализе. Необходимо охарактеризовать вид распределения по отношению к нормальному распределению — асимметрию, эксцесс, полимодальность. Для этого следует привести график гистограммы совместно с графиком плотности нормального распределения, а также таблицу основных статистик.

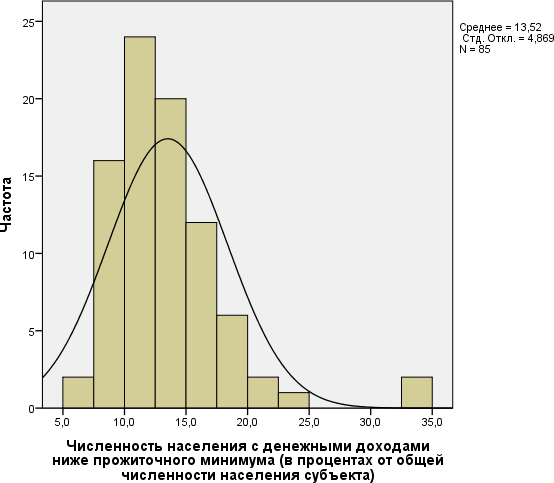


Рисунок 1. Численность населения ….

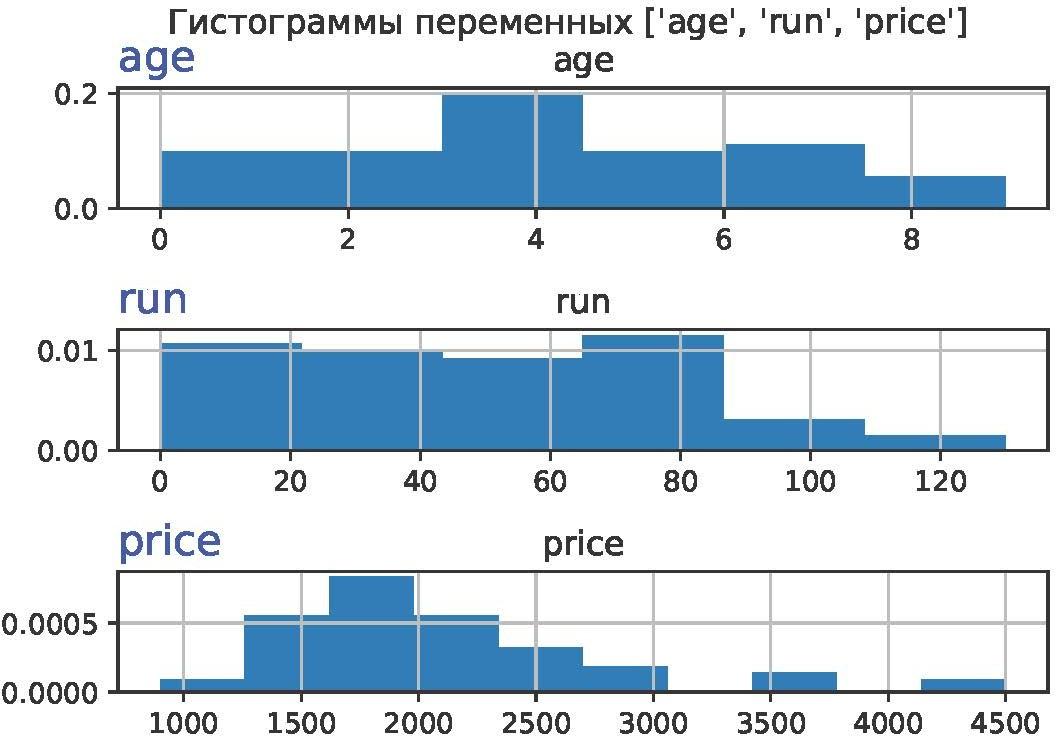
|  |  |
| --- | --- |
| Статистика | Значение |
| Среднее |  |
| Медиана |  |
| Стандартное отклонение |  |
| Межквартильный размах |  |
| Верхняя квартиль |  |
| Нижняя квартиль |  |
| Коэффициент асимметрии |  |
| Коэффициент эксцесса |  |
| Минимальное значение |  |
| Максимальное значение |  |
| Количество наблюдений |  |
| Количество пропущенных значений |  |

Таблица 2. Статистические свойства количественных факторов.

Необходимо дать интерпретацию статистических свойств количественных переменных в контексте предметной области. Например, (см. Рисунок 1), на основании гистограммы и числовых характеристик распределения можно сделать вывод о наличии небольшого количества субъектов федерации с очень большой долей бедного населения.

Также, для целевой переменной следует проанализировать наличие выбросов на основании правила «трех-сигм». Следует отметить в базе все выбросы и на основании сравнения соответствующих значений объясняющих переменных с их средними/медианными значениями объяснить, почему эти наблюдения могут интерпретироваться как выбросы.

Желательно формировать комбинированные визуализации — на одном листе приводить графики для целевой переменной и нескольких объясняющих переменных для того, чтобы была возможность сравнить их статистические свойства.



2.2.2. Анализ качественных переменных.

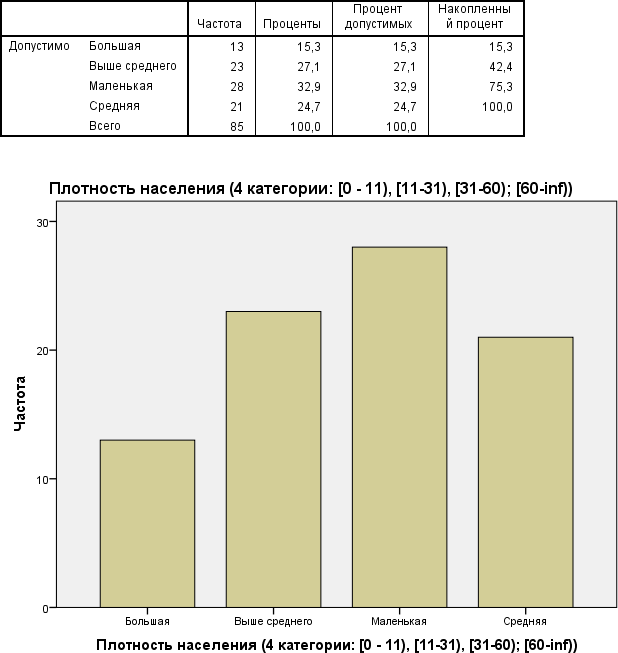
Здесь следует привести столбчатые диаграммы, которые отражают количество измерений с разными уровнями для данной переменной. Приветствуются комбинированные визуализации.

Рисунок 2. Плотность населения.

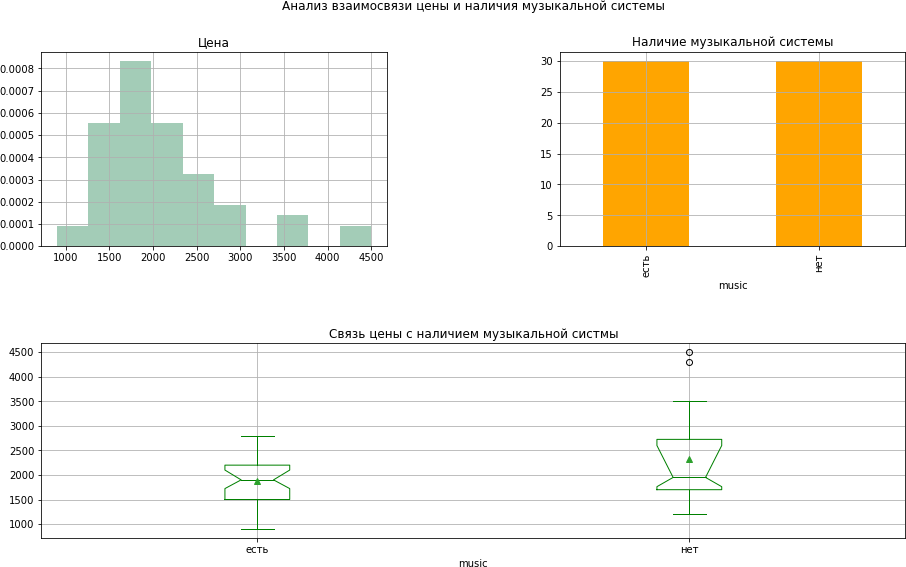
Необходимо проанализировать степень представленности всех уровней и при необходимости (наличии уровней с долей менее 5%) произвести укрупнение уровней. Результат привести на новых диаграммах. Принцип укрупнения пояснить.

* 1. Анализ статистической связи.
     1. Графический анализ пары «целевая переменная – качественная объясняющая переменная».

Здесь для каждой пары {количественная зависимая переменная – качественная независимая переменная} необходимо построить категорированную диаграмму Бокса- Уискера (Box-Whisker).

Рисунок 3. Зависимость заболеваемости от уровня преступности.

На основании анализа диаграммы следует охарактеризовать связь среднего значения и разброса количественной зависимой переменной с уровнями качественной независимой переменной. Интерпретацию дать в контексте предметной области.

Рекомендуется использовать комбинированные визуализации

Для формальной проверки гипотезы о наличии статистической связи следует выполнить непараметрический дисперсионный анализ (критерий Крускала-Уоллиса).

* + 1. Графический анализ пары «числовая зависимая переменная – числовая независимая переменная».

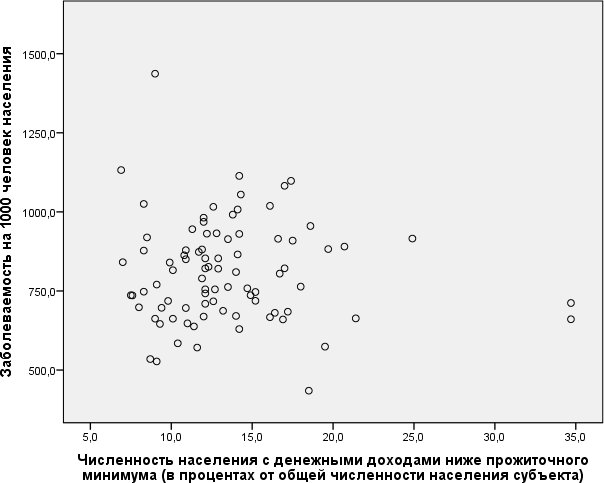
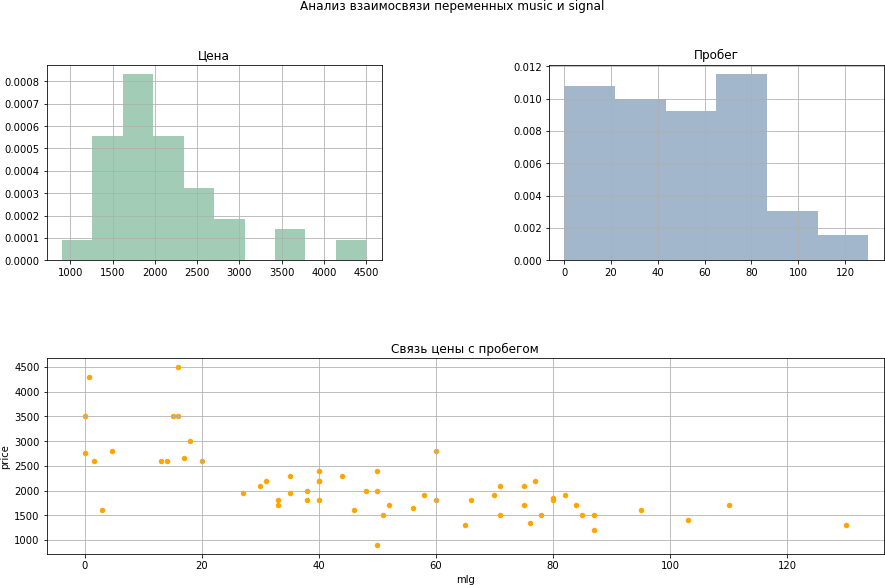
Здесь для каждой пары {количественная зависимая переменная – количественная независимая переменная} необходимо построить диаграммы рассеивания (Scatter plot).

Рисунок 4. Зависимость заболеваемости от …

На основании визуального анализа диаграммы следует сделать предположение о наличии и характере статистической взаимосвязи. Интерпретацию результатов дать в контексте предметной области.

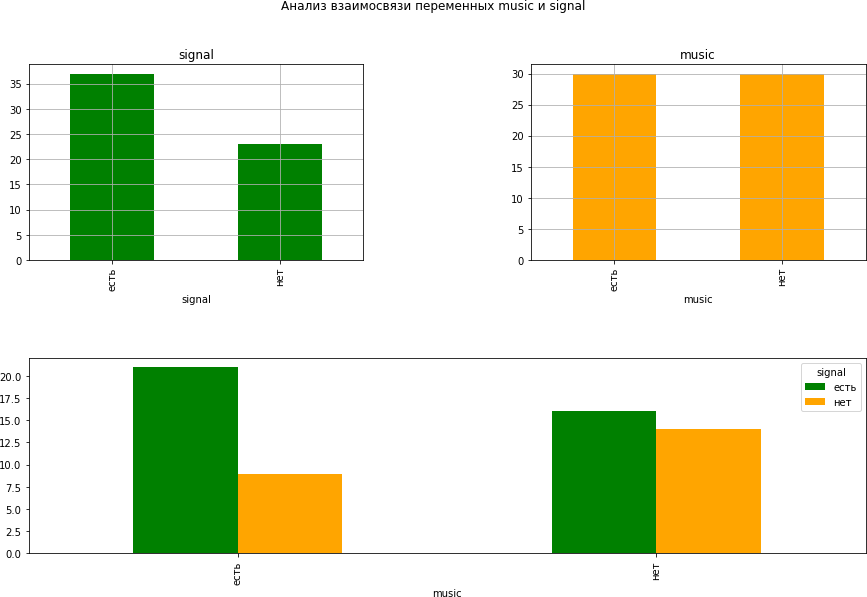
Рекомендуется использовать комбинированные визуализации вида



Для формальной проверки гипотезы о наличии связи следует подсчитать коэффициенты корреляции Пирсона и Спирмена, а также тау Кендала и привести результаты проверки их значимости.

* + 1. Анализ статистической взаимосвязи между независимыми переменными.

Следует проанализировать силу связи между независимыми переменными, используя инструменты пп. 3.2.1 и 3.2.2. Для анализа силы связи между качественными переменными следует использовать анализ таблиц: необходимо привести таблицу кросс- табуляции, значения статистики хи-квадрат и V-Крамера.

Рекомендуется использовать комбинированные визуализации

* + 1. Предварительная проверка гипотез

Здесь необходимо рассказать о результатах проверки гипотез из п.1.3 на основании предварительного анализа данных.

# Проверка гипотез с помощью моделирования

Данный раздел предполагает проверку прогностических способностей построенной модели. В связи с этим исходную выборку следует случайным образом разделить на обучающую и тестовую в пропорции 80:20. На обучающей выборке будет осуществляться построение моделей, тестовая выборка будет использоваться для проверки прогностических способностей.

* 1. Построение базовой модели.

Базовая модель служит для анализа изменения качества моделирования при учете сформулированных гипотез. В качестве базовой модели следует использовать модель линейной регрессии целевой переменной на все объясняющие. Для базовой модели следует проверить значимость всех объясняющих переменных, а также уровень мультиколлинеарности (показатель VIF) и наличие гетероскедастичности (критерий Уайта). Исходная базовая модель и результаты ее анализа включается в отчет.

Далее необходимо оптимизировать структуру модели для повышения ее качества и возможного снижения уровня мультиколлинеарности. Для этого следует пошагово удалять незначимые переменные, переоценивая модель после каждого удаления. Необходимо также пошагово удалять переменные, которые демонстрируют высокую взаимосвязь с другими переменными (VIF > 3). Оценку мультиколлинеарности и гетероскедастичности следует выполнять на каждом шаге оптимизации. В отчете следует привести один промежуточный и итоговый вариант, который не содержит незначимых объясняющих переменных и имеет удовлетворительный уровень мультиколлинеарности. Следует привести оценку мультиколлинеарности вошедших в модель переменных и оценку наличия гетероскедастичности. Необходимо также привести оценку качества полученной модели (критерий Akaike, R-sq и adjusted R-sq).

В ходе оптимизации следует оставить в модели объясняющие переменные, которые необходимы для проверки гипотез даже, если они незначимы или имеют высокое значение показателя VIF. Это следует отметить в отчете.

* 1. Проверка гипотез с помощью моделирования

Для проверки выдвинутых в п. 1.2. сложных гипотез выполняется модификация оптимизированной базовой модели поэтапно для каждого сочетания сформулированных гипотез. Сначала модифицируют базовую модель для каждой сложной гипотезы отдельно, далее для всевозможных пар и т.д. Для простых гипотез модификация не требуется.

Модифицированные модели оцениваются и выполняется проверка как сложных, так и простых гипотез. Методология проверки каждой гипотезы должна быть описана в отчете в виде ограничений на коэффициенты и пары статистических гипотез. Результаты использования каждой модифицированной модели включаются в отчет.

Модель, которая учитывает все сформулированные гипотезы объявляется итоговой.

* 1. Оптимизация итоговой модели, сравнение качества моделей.

Итоговая модель подвергается оптимизации за счет пошагового удаления незначимых переменных. На каждом шаге модель переоценивается. Для финального варианта оценивается качество модели с использованием критерия Akaike и adjusted R-sq. Оптимизированная итоговая модель и результаты ее анализа включаются в отчет.

По результатам работы формируется таблица с перечнем моделей включенных в отчет и оценками их качества — значениями критерия Akaike, R-sq и adjusted R-sq

|  |  |  |  |
| --- | --- | --- | --- |
| Номер\Критерий | R^2 | Adj R^2 | Akaike |
|  |  |  |  |

Таблица 3. Сравнение качества построенных моделей. Таблицу следует прокомментировать.

* 1. Проверка прогностических способностей модели

Проверка прогностических способностей осуществляется для всех включенных в отчет моделей. Необходимо подсчитать значения прогнозов для элементов тестовой выборки и построить для них центральные доверительные интервалы на основе нормального распределения для доверительной вероятности 95%. Для результатов следует рассчитать среднеквадратическую погрешность прогнозирования и максимальную абсолютную погрешность прогнозирования, а также эмпирическую оценку доверительной вероятности. Результаты следует представить в виде таблицы

|  |  |  |  |
| --- | --- | --- | --- |
| Номер\Критерий | Среднеквадратическая  погрешность | Абсолютная  погрешность | Доверительная  вероятность |
|  |  |  |  |

Таблица 4. Сравнение прогностических способностей моделей

Таблицу следует прокомментировать, в частности, оценку доверительной вероятности. Результаты, представленные в таблице, следует сопоставить с оценками качества данных моделей.

4. Заключение

В данном разделе следует перечислить результаты проверки сформулированных гипотез в различных сочетаниях и проверки прогностических способностей моделей.