**Отчёт по итоговому домашнему заданию  
по предмету “**[**Прикладные методы математической статистики**](https://lms.hse.ru/student.php?lessons_ID=143541)**”**

Выполнен студентами группы БПИ204:  
Пеганов Никита  
Куляхтина Алёна  
Завалинский Даниил

**Описание данных и особенности работы с ними**

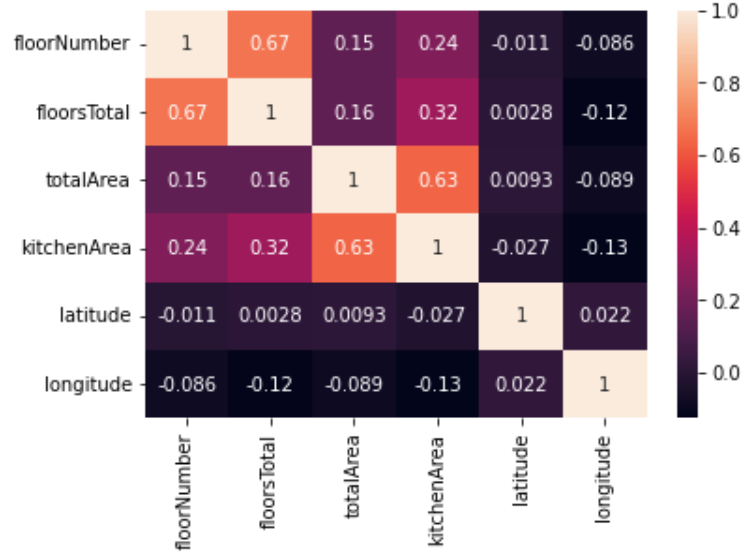
Выбранные нами данные взяты из статьи <https://medium.com/@max.bobkov/machine-learning-moscow-flats-appraising-25a1e9f171db>. По словам автора, они получены в результате парсинга сайта [www.cian.ru](http://www.cian.ru), специализирующегося на продаже квартир. Также данные очищены «от дублей, объявлений с одинаковым текстом в описании, от строк с нулевой площадью кухни, без материала стен и этажей в доме».

Для оценки полученных результатов данные были разделены на две выборки: обучающую и тестовую в пропорции 70 к 30. Обработка данных происходила в среде jupyter notebook на языке python. Код решения можно найти в файле processing.ipynb.

**Описание переменных, используемых в анализе**

Рассмотрим переменные, использованные в датасете:  
**wallsMaterial** — **материал стен  
floorNumber** — **на каком этаже расположена квартира**  
**floorsTotal** — общее число этажей в доме  
**totalArea** — общая площадь квартиры  
**kitchenArea** — площадь кухни  
**latitude** — широта расположения дома на карте  
**longitude** — долгота расположения дома на карте  
**price** — цена квартиры

Проверим объясняющие переменные на мультиколлинеарность. Для этого составим матрицу корреляции числовых переменных:



Матрица корреляции показывает, что значительных попарных корреляций между переменными в данных нет. Они не превосходят 0,7. Это означает, что возможных проблем с качеством получаемых оценок не предвидится.

**Описание полученных регрессионных моделей**

С помощью модуля анализа данных Excel посчитаны коэффициенты линейной регрессии. Ее вид:

Где - число, соответствующее виду материала, – номер этажа квартиры, – общее число этажей в квартире, – площадь квартиры, – площадь кухни, – широта расположения квартиры, – долгота расположения квартиры.

При изменении номера этажа на единицу цена меняется на в ту же сторону, следовательно, чем выше этаж, тем квартира дороже.

При изменении количества этажей на единицу цена изменяется на в обратную изменению кол-ва этажей сторону, квартиры в малоэтажных домах дороже.

При изменении площади квартиры на единицу цена меняется на в ту же сторону.

При изменении площади кухни на 1 квадратный метр цена меняется на в обратную сторону, квадратный метр площади кухни дешевле, чем квадратный метр жилой площади.

При изменении широты координат квартиры на карте на 0,001 цена изменится на в ту же сторону, недвижимость на севере Москвы дороже.

При изменении долготы координат квартиры на карте на 0,001 цена изменится на в ту же сторону, недвижимость на востоке Москвы дороже

Затем в программе Excel взяты натуральные логарифмы от значений всех объясняющих переменных, а также от цены квартир. На них также построена линейная регрессия с помощью модуля анализа данных. Полученная логарифмическая регрессия имеет вид:

|  |
| --- |
| 0,999145 |
| 1,00025 |
| 0,999877 |
| 1,011146 |
| 1,002266 |
| 1,521203 |
| 1,024284 |

При увеличении номера этажа на 1% цена уменьшается на 0,025%. При увеличении количества этажей на 1% цена увеличивается на 0,012258%. При увеличении площади квартиры на 1% цена уменьшается на 1,1146%. При увеличении площади кухни на 1% цена увеличивается на 0,2266%. При увеличении широты квартиры на 1% цена увеличивается на 52,12%. При увеличении долготы на 1% цена уменьшается на 2,4283%.

**Описание проведенных тестов - Тест Чоу**

Для проведения теста Чоу выборка была разделена на две подвыборки – дома, в которых не более 12ти этажей, и те, в которых 13 и больше – подвыборки A и B соответственно.

n = 63931 – общее число наблюдений в полной выборке

nA = 29550 – число наблюдений в выборке А

nB = 34381 – число наблюдений в выборке B

k = 7 – число регрессоров без константы модели

SL = 5% - уровень значимости

1. Оцениваем регрессию для полной выборки и находим сумму квадратов остатков

RSSall = 25 206 816 903 486 800 000

1. Оцениваем регрессии по подвыборкам A и B и находим сумму квадратов остатков для каждой

RSSA = 17 432 539 921 279 500 000

RSSB = 5 507 061 025 055 580 000

1. Рассчитываем значение статистики:
2. Критическая точка:
3. Получим, что Тнабл ∈ [0; Ткр].

**Вывод: гипотеза не отвергается, и мы оцениваем единую модель по всей выборке.**