МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИМЕНИ М. В. ЛОМОНОСОВА

Факультет «Вычислительной Математики и Кибернетики»

Направление подготовки «Математической Статистики»

01.03.02 (бакалавр)

**Борщ Максим Валерьевич**

студент учебной группы МС-316

Отчёт по домашним заданиям за первую половину семестра.

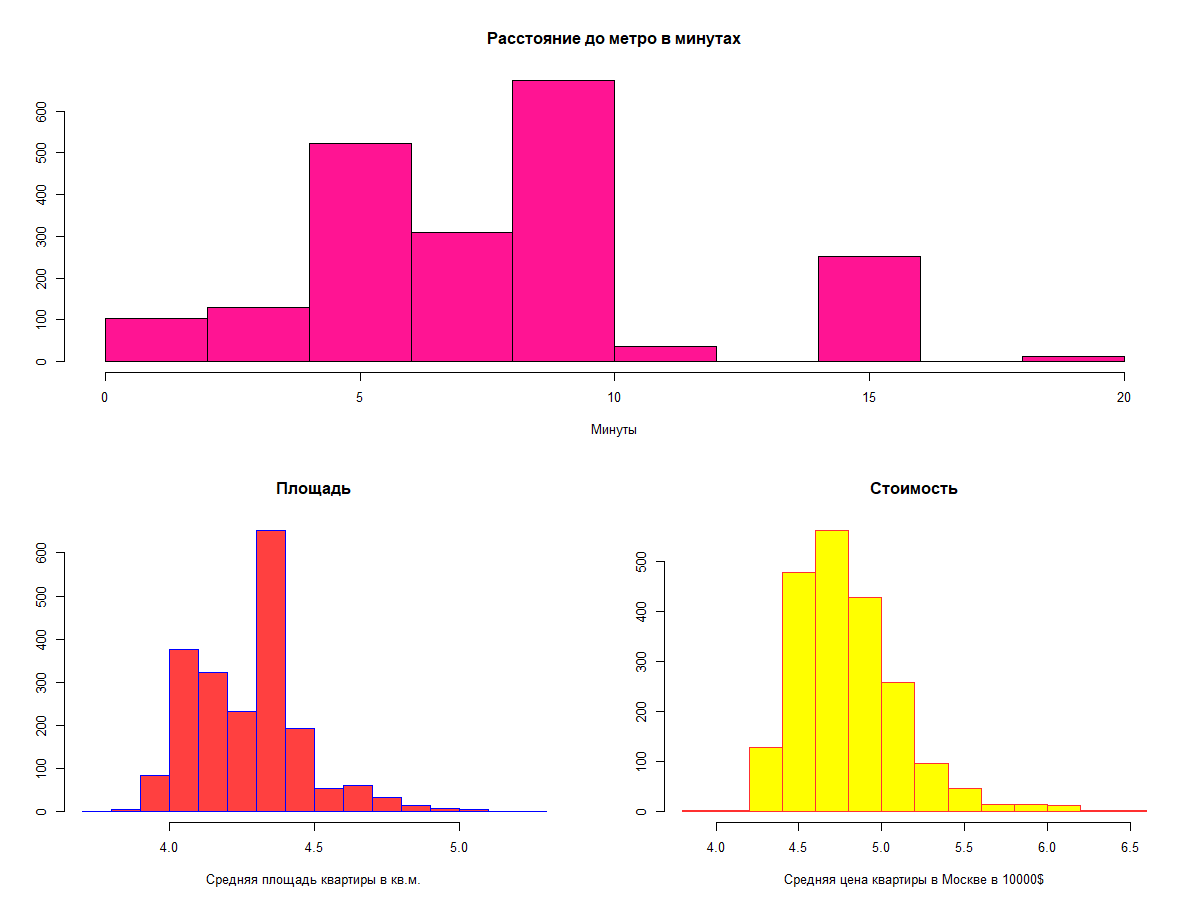
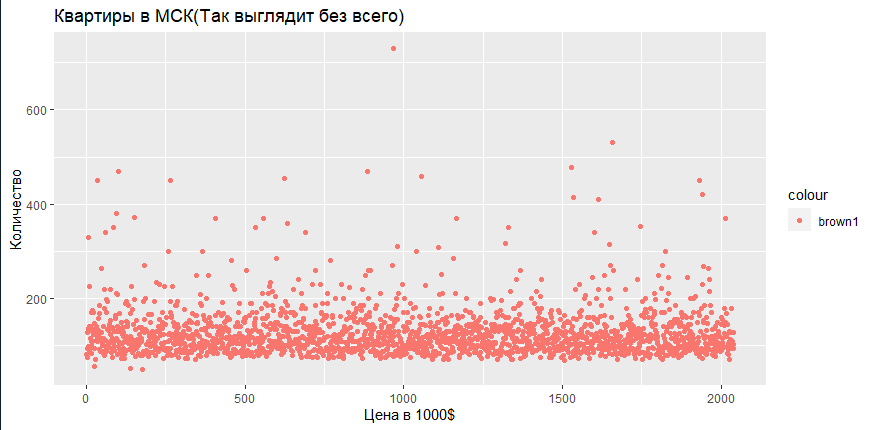
Проверил: кандидат физико-математических наук Горшенин Андрей Константинович

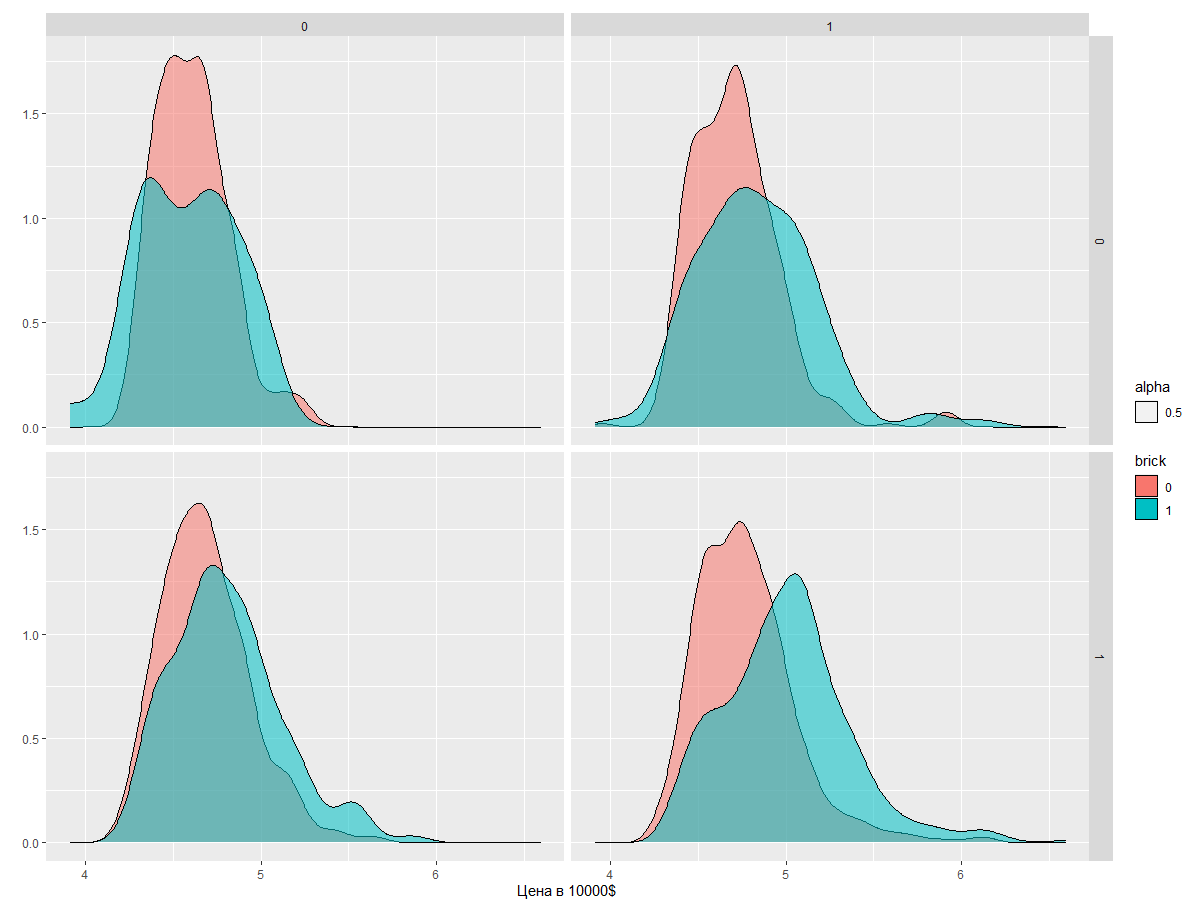
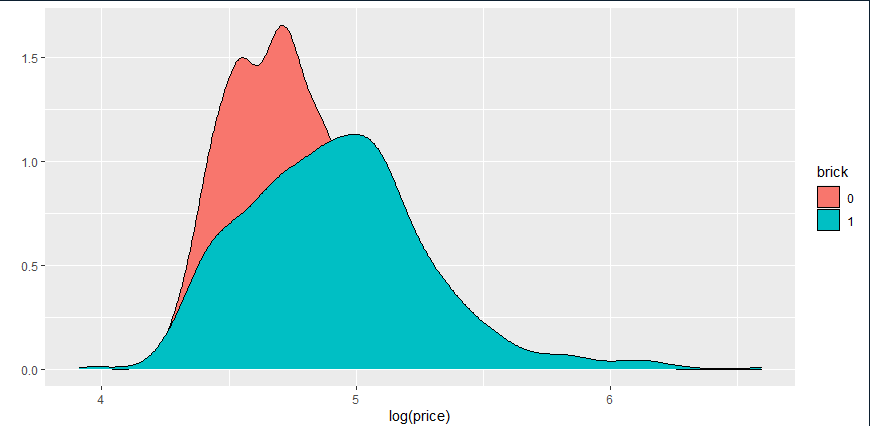
Москва, 2020

Отчёт по домашним заданиям за первую половину семестра.

В ходе выполнения домашних работ использовал базу данных по стоимости квартир в Москве за 2019 год.

Для начала продемонстрирую несколько графиков, чтобы можно было составить небольшое представление о данных:





n – номер квартиры по порядку

price – цена квартиры в $10000

totsp – общая площадь квартиры, кв.м.

livesp жилая площадь квартиры, кв.м.

kitsp – площадь кухни, кв.м.

dist – расстояние от центра в км.

metrdist – расстояние до метро в минутах

walk – 1 – пешком от метро, 0 – на транспорте

brick 1 – кирпичный, монолит ж/б, 0 – другой

floor 1 – этаж кроме первого и последнего, 0 – иначе.

code – число от 1 до 8, при помощи которого мы группируем наблюдения по

подвыборкам:

1. Наблюдения сгруппированы на севере, вокруг Калужско-Рижской линии

метрополитена

2. Север, вокруг Серпуховско-Тимирязевской линии метрополитена

3. Северо-запад, вокруг Замоскворецкой линии метрополитена

4. Северо-запад, вокруг Таганско-Краснопресненской линии метрополитена

5. Юго-восток, вокруг Люблинской линии метрополитена

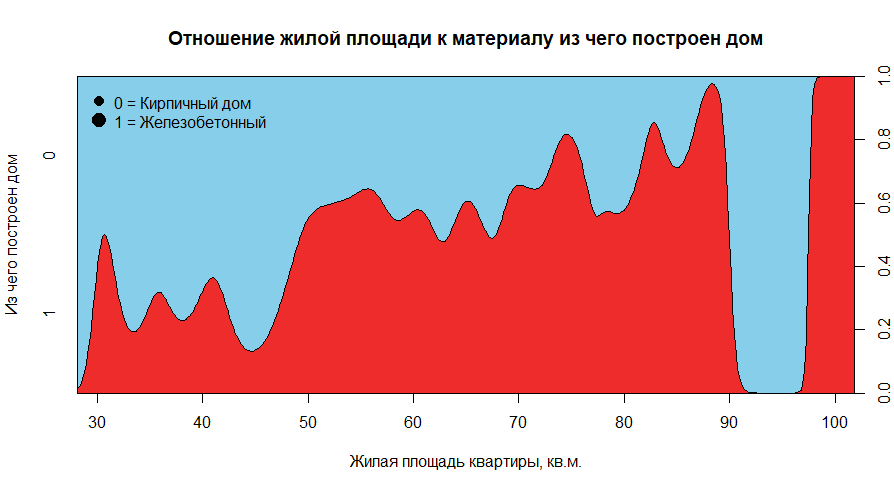
6. Юго-восток, вокруг Таганско-Краснопресненской линии метрополитена

7. Восток, вокруг Калиниской линии метрополитена

8. Восток, вокруг Арбатско-Покровской линии метрополитена

metrdist – дистанция дома ближайшей станции метро в минутах пешком

Всего 2040 квартир

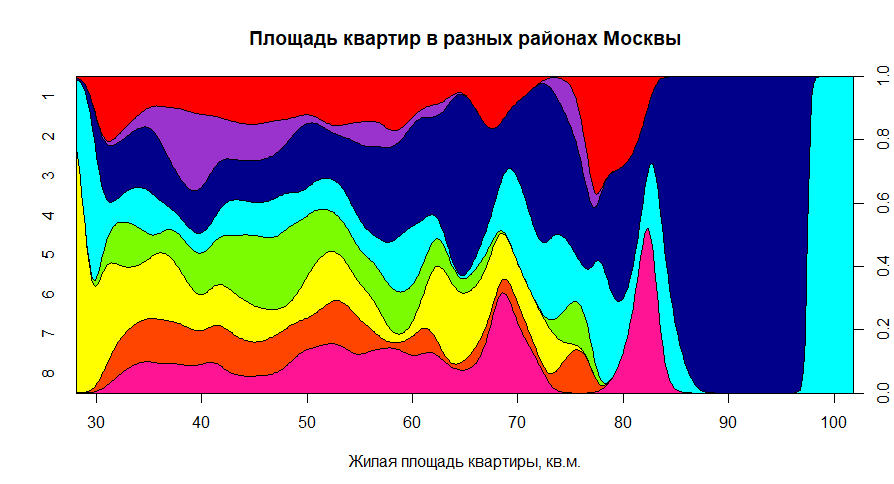


Из графика выше можно сделать следующий вывод:

- Нету квартир с жилой площадью от 91 до 96 кв м построенных из железобетона

- Большую часть квартир из среднего значения от 47 до 90 это квартиры сделанные из железобетона

- На рынке если есть квартиры, построенные из другого материала, то их значение мало по сравнению с кирпичными и железобетонными квартирами



1. Наблюдения сгруппированы на севере, вокруг Калужско-Рижской линии

метрополитена

2. Север, вокруг Серпуховско-Тимирязевской линии метрополитена

3. Северо-запад, вокруг Замоскворецкой линии метрополитена

4. Северо-запад, вокруг Таганско-Краснопресненской линии метрополитена

5. Юго-восток, вокруг Люблинской линии метрополитена

6. Юго-восток, вокруг Таганско-Краснопресненской линии метрополитена

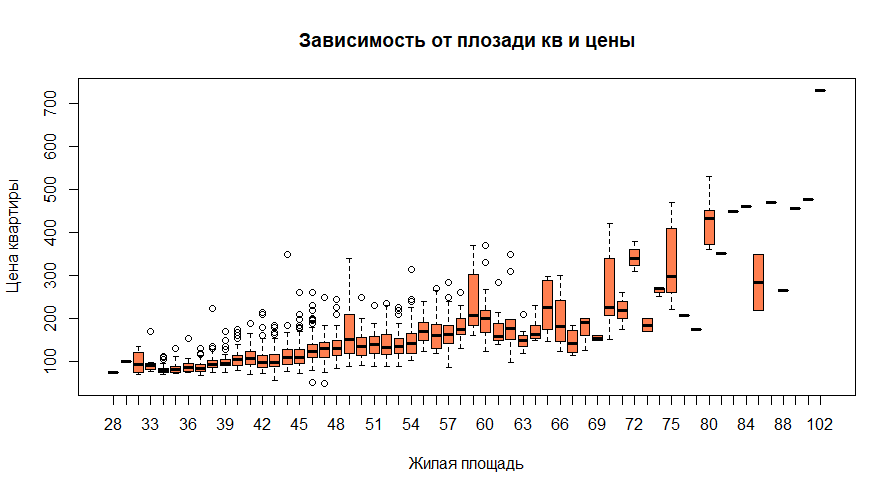
7. Восток, вокруг Калиниской линии метрополитена

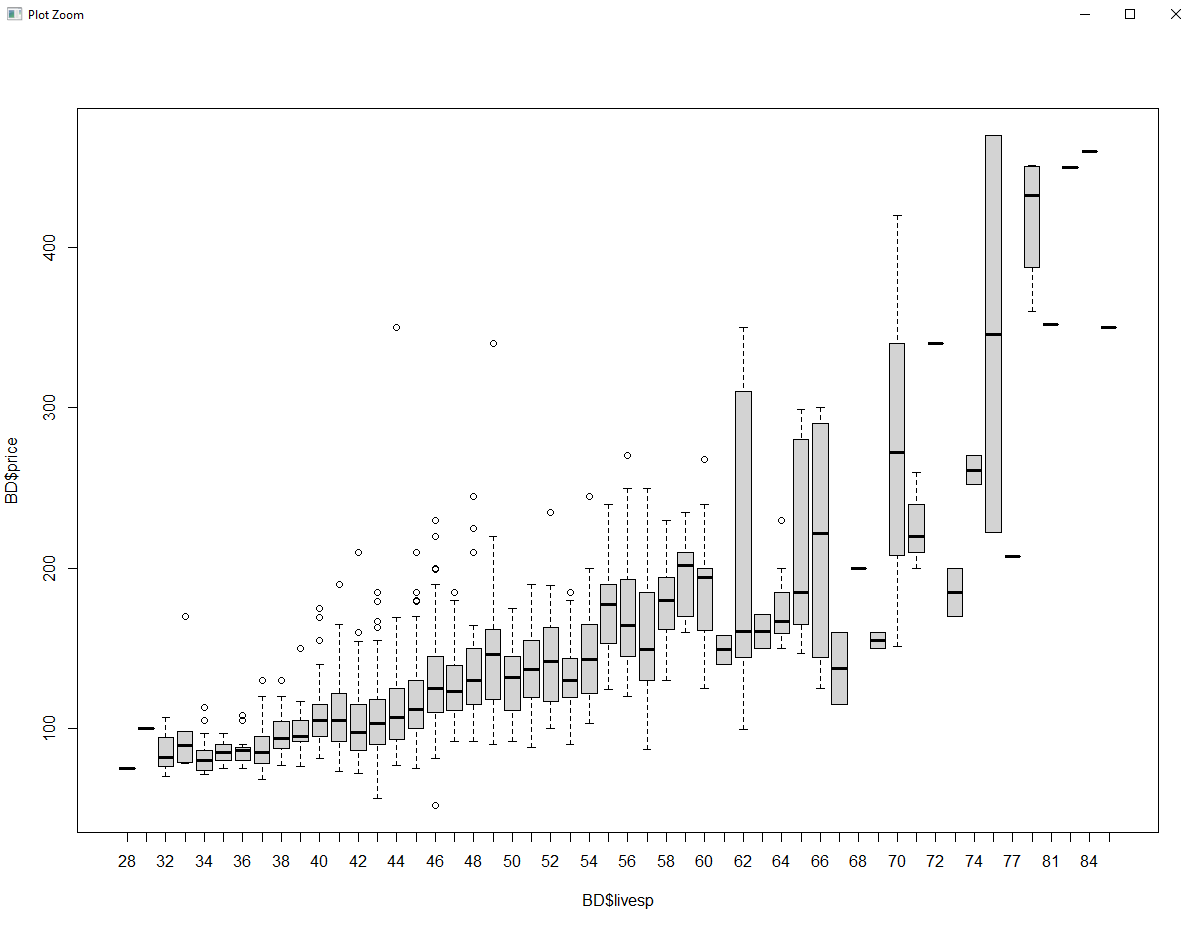
8. Восток, вокруг Арбатско-Покровской линии метрополитена

Отсюда можно сделать следующие выводы, что в 8 районе нет квартир от 90 + кв м

Весь рынок квартир от 90 до 96 км м занимают квартиры в 3 районе

Квартиры больше 96 кв м продаются только в 4 районе



Вообще выбросы можно показать с помощью boxplot

Черная черточка — это медиана (среднее значение)

то, что в прямоугольнике вверх и в низ серым цветом это 25% данных

Пунктир это 75% моих данных

Все остальное это выбросы, которые обозначаться кружками (аномальные данные от которых нужно избавляться)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

База данных по ценам на квартиры в Москве оказалась удобной для применения критериев хи-квадрат и точного теста Фишера. В нём проверялись гипотезы о влиянии на цену квартиры таких вещей, как:

- Общая жилая площадь

- Дистанция до станции метро

- Площадь кухни

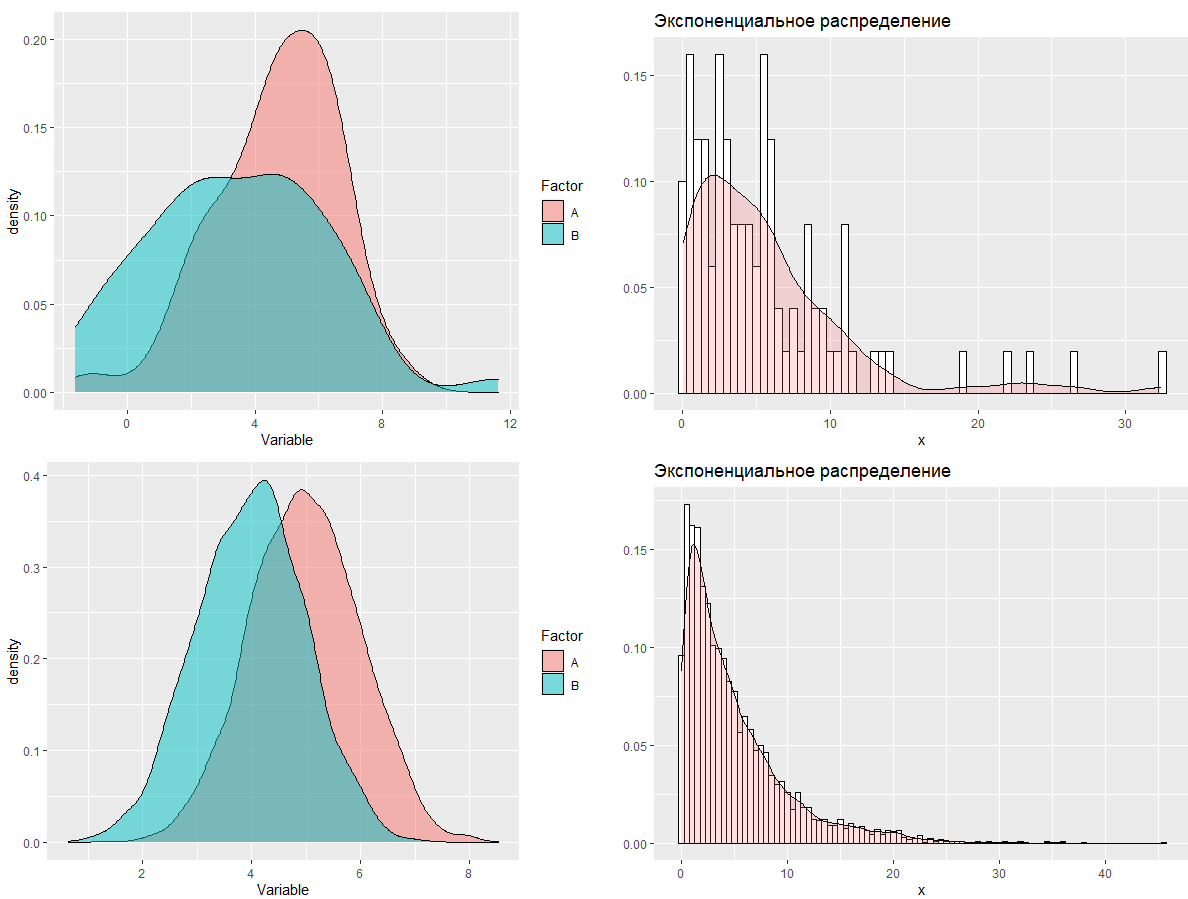
- Из чего построен дом.

Оказалось, что на ближайшую станцию метро люди практически не смотрят, влияние на то, из чего построен дом на цену тоже не подтверждается (pvalue ≈ 0.5), а зависимость от площади кухни уже подвергается к сомнению: pvalue = 0.05989. Практически точно такие же результаты показал точный тест Фишера.

Один раз применялись синтетические данные — в домашнем задании, в котором нужно было использовать тесты на нормальность.

Их итоги были вполне предсказуемы: нормальное распределение нормально, а Вейбулловское - не очень нормально. Тест на нормальность проходит и смесь нормальных распределений, если их параметры довольно близки.

Так же мы сгенерировали плотности с разным количеством наблюдений



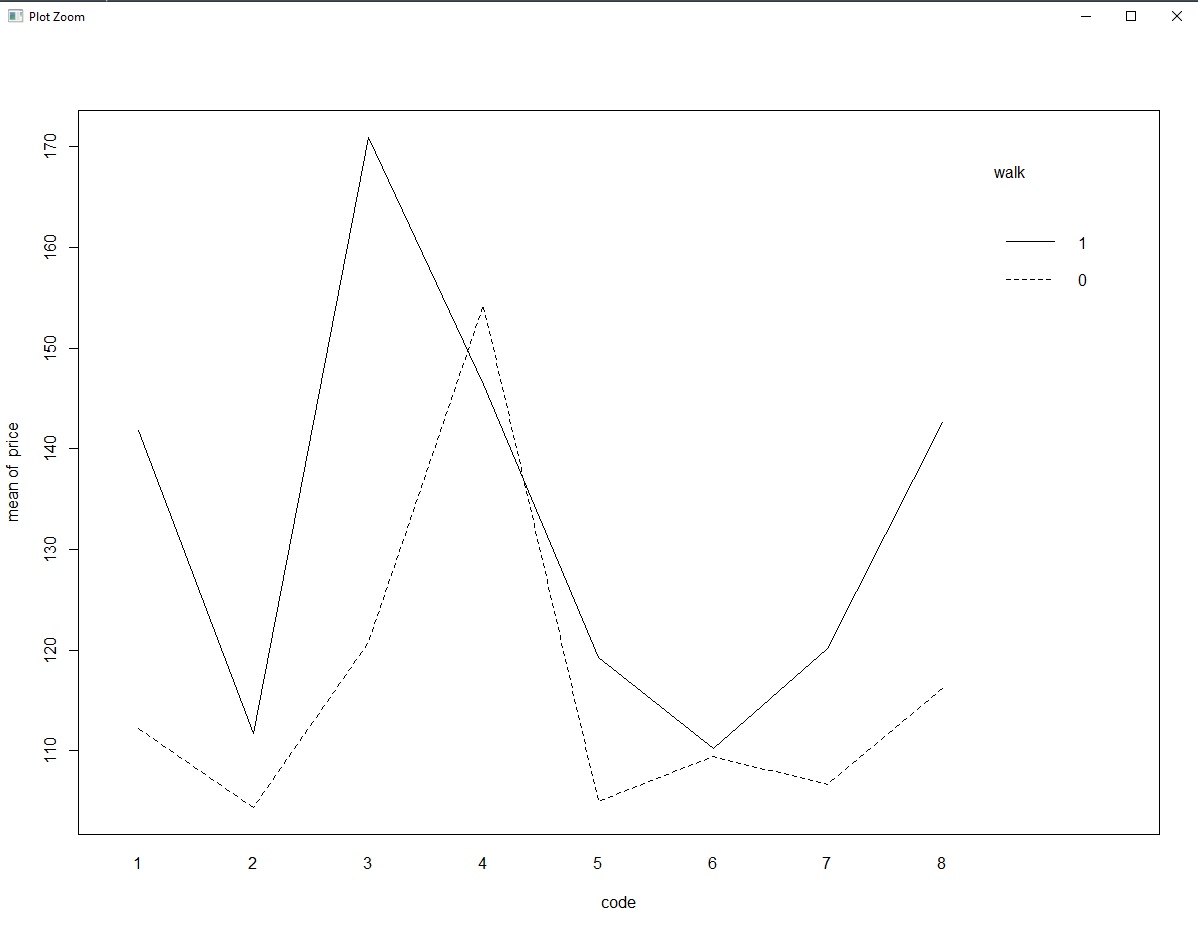
Из всего выше показанного можно слетать следующий вывод:

Что при увеличении количества наблюдений, любое распределение стримится к своему нормальному виду (по которому можно сказать, что это за распределение)

При чем при малом количестве наблюдений мы может догадываться что это за распределение (к примеру, гипергеометрическое распределение, нормальное и биномиальное распределение, при малом количестве данных они похожи между собой) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дисперсионный анализ – это форма линейной регрессии, поэтому в идеале су-  
ществует линейная связь между независимыми переменными и зависимой пере-  
менной. Одним из источников нелинейной связи является взаимодействие меж-  
ду двумя независимыми переменными: когда одна переменная меняет значение,  
другая меняет свою связь с зависимой переменной. Проверка взаимодействия  
между независимыми переменными является базовой диагностикой.

При отображении зависимости переменной code от переменной price мы ищем прямые, параллельные линии, которые указывают на линейную связь.  
Тем не менее, используя функцию interaction.plot, мы получаем диаграмму

Глядя на нее, видно, что что-то не так:

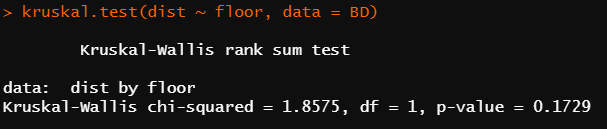
Каждая линяя на диаграмме отображает зависимость переменной price от пере-  
менной code. Разница между линиями заключается в том, что каждая линия пред-  
назначена для разных значений переменной walk.

Очевидно, что различие в значении walk «исказило» линии, внося нелинейность в

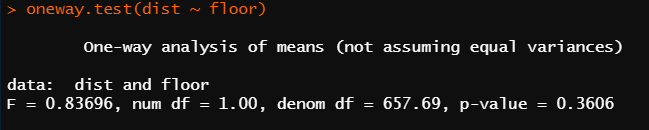
связь между code и price.

Это сигнализирует о возможном взаимодействии, которое мы должны прове-  
рить. В случае с этими данными просто так получается, что да, взаимодействие  
есть, и нет, оно не является статистически значимым. Мораль ясна: визуальная  
проверка полезна, но не надежна.

Обычный дисперсионный анализ предполагает, что ваши данные имеют нор-  
мальное распределение. Он может допускать некоторое отклонение от нормаль-  
ности, но крайние отклонения приведут к бессмысленным p-значениям.  
Критерий Краскела–Уоллиса представляет собой непараметрическую версию  
ANOVA. Это означает, что он не предполагает нормальности. Тем не менее он  
предполагает распределения одинаковой формы. Критерий Краскела–Уоллиса  
следует использовать, если распределение ваших данных ненормально или просто  
неизвестно.  
Нулевая гипотеза состоит в том, что все группы имеют одинаковую медиану.  
Когда мы отвергаем нулевую гипотезу (когда p < 0,05), это не значит, что все груп-  
пы разные, а предполагает, что две или более групп различны.

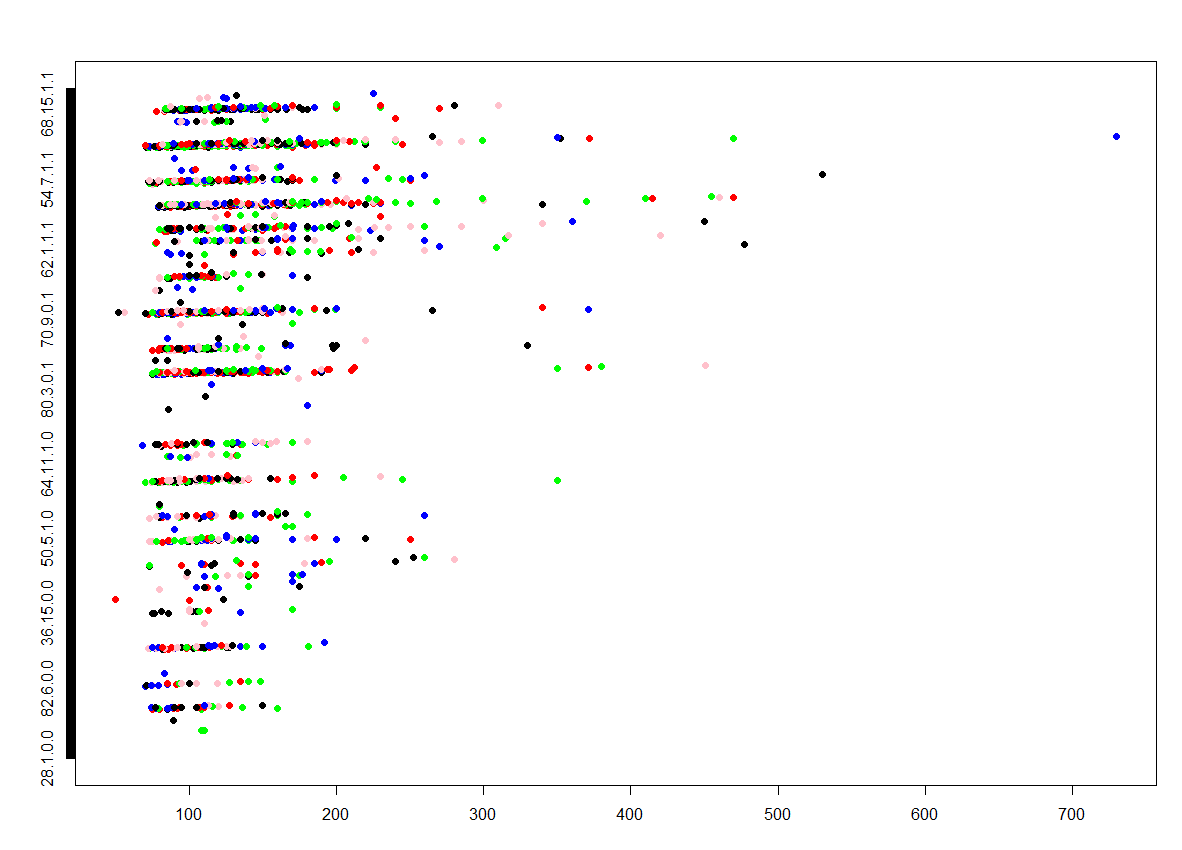


Однофакторный

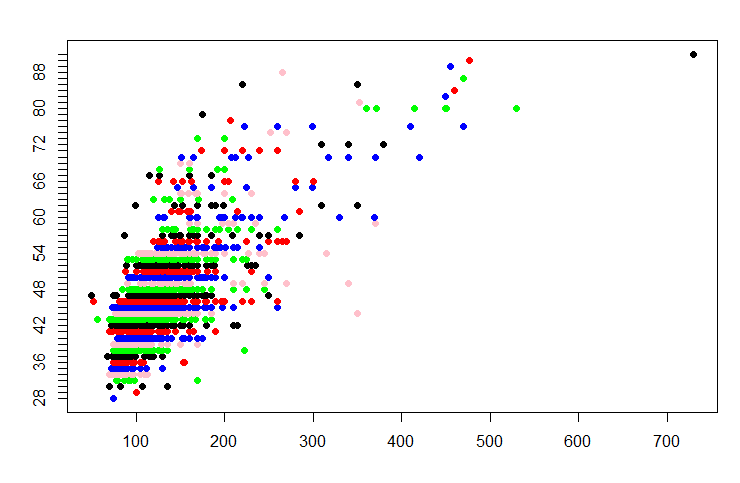


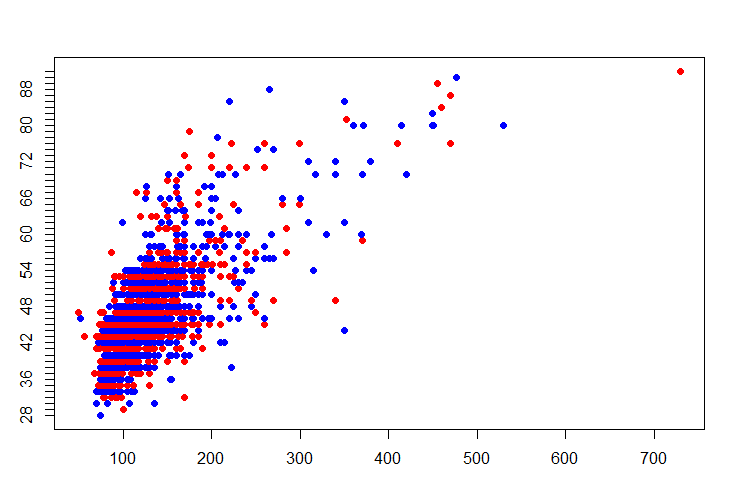
Но к моим данных можно и в лоб применить дисперсионный анализ

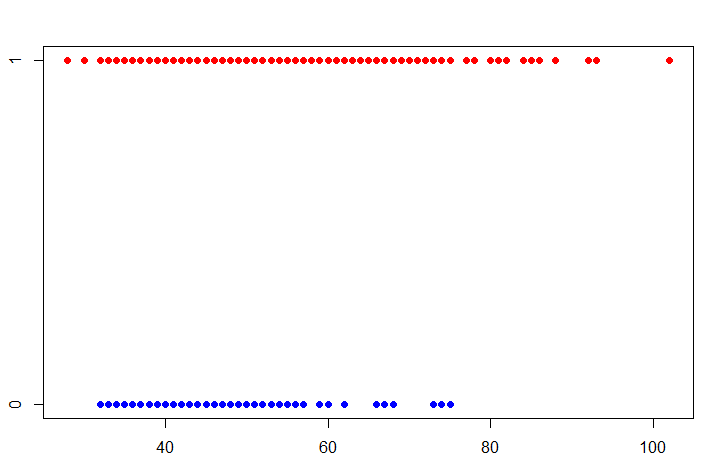
Но ничего хорошего из этого не выйдет

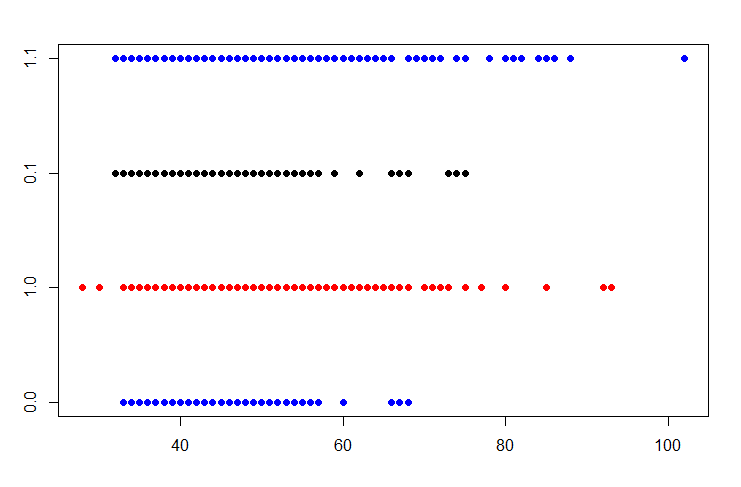


Да в моей БД есть выбросы, и они влияют на картину в целом, если их убрать это уже будет совершенно другая БД и уже тогда будет вопрос про достоверность полученных прогнозов и закономерностей.





И тяжело как то судить что происходит по этим графикам 



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Общий вывод:

Рынок недвижимости в Москве и Подмосковье в последнее время стабилизировался. Падение цен на квартиры, наблюдавшееся с началом экономического кризиса, стали сокращаться и к осени стали минимальны. Вероятней всего, в ближайшие месяцы, произойдет окончательная стабилизация рынка и произойдет постепенное восстановление спроса на недвижимость. Если улучшится экономическая ситуация в стране и начнется постепенный выход из кризиса, то и цена на недвижимость, через некоторое время, будет возвращаться к прежним позициям. Но нельзя и полностью исключать другое развитие событий: вследствие новой волны экономического кризиса стоимость квадратного метра может продолжить свое падение. В любом случае, как продажа квартир, так и их покупка – останутся неизменным процессом, требующим определенных навыков и знаний. Квартиры в Москве могут служить не только для комфортного проживания, но и как эффективное средство капиталовложения. Купить квартиру в Москве, на первый взгляд, очень просто, но когда углубляешься в этот процесс, становится понятно, что все просто только на первый взгляд.