Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»

Факультет компьютерных систем и сетей

Кафедра ПОИТ

ОТЧЁТ
по дисциплине
«Модели и методы обработки больших объёмов данных»

Тема проекта
«Анализ данных об аренде квартир»

Выполнила:
студент гр. 156301
Карнаух Д. М.

Проверил:
доц. каф. информатики
Стержанов М. В.

МИНСК 2021

ОГЛАВЛЕНИЕ

1. ПОСТАНОВКА ЗАДАЧИ

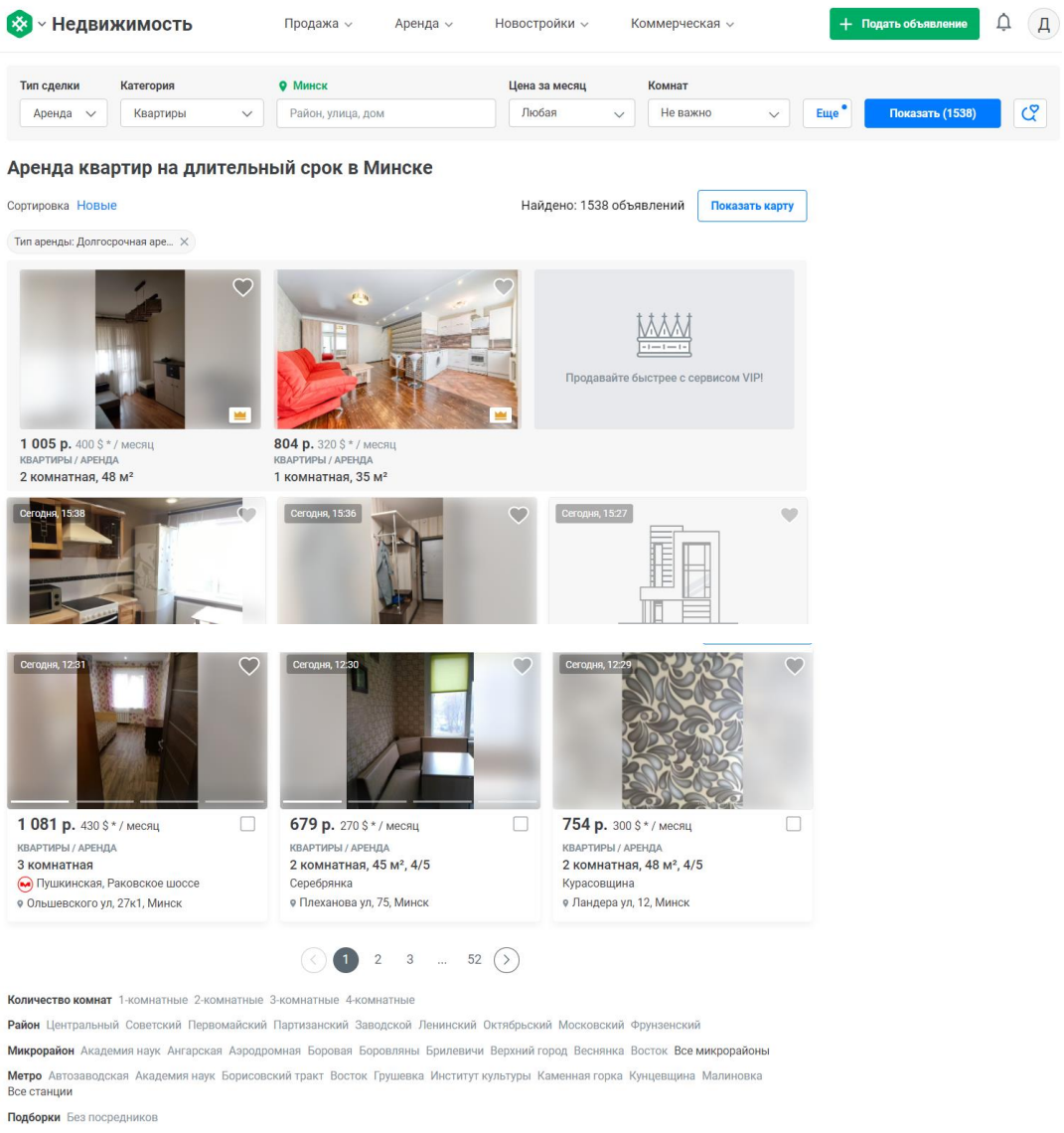
Тема проекта «Анализ данных об аренде квартир». Для анализа данных используются данные с сайта kufar.by. Было решено использовать данные о квартирах только в Минске, так как на цену аренды влияет город, поэтому для упрощения задачи анализа был взят только один большой город.

Анализ данных о квартирах представляет собой анализ о их обустройстве, расположении, ремонте, площади и их влияние на цену.

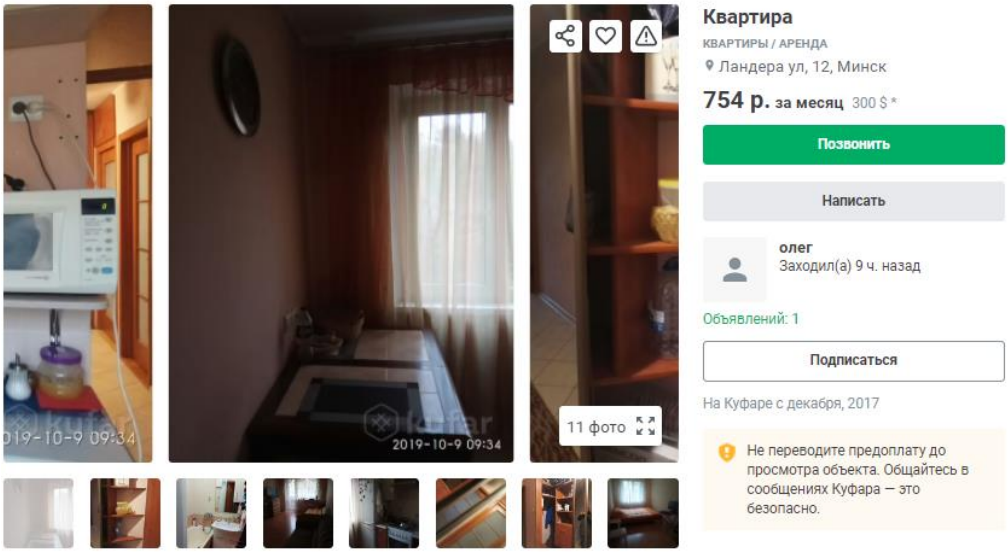
Также кроме анализа будет предсказана цена на основе собранных данных, используя простейшие модели машинного обучения, и будут сделаны выводы о ценности характеристик для модели.

2. ПОЛУЧЕНИЕ ДАННЫХ

Сайт Куфар с выставленными параметрами по поиску квартир выглядит следующим образом:



Каждый элемент представляет собой фотографию с кратким описанием квартиры и ссылкой на полное описание квартиры. Страницы указаны снизу, поэтому переход по ним осуществляется через стрелочку. Страница с описанием квартиры выглядит так:



Квартира

КВАРТИРЫ / АРЕНДА

Ландера ул, 12, Минск

754 р. за месяц 300 $ *

Позвонить

Написать

олег
Заходил(а) 9 ч. назад

Объявлений: 1

Подписаться

На Куфаре с декабря, 2017

Не переводите предоплату до просмотра объекта. Общайтесь в сообщениях Куфара — это безопасно.

Вчера, 12:29 Просмотр панорамы

Характеристики

Город / Район	Октябрьский
Микрорайон	Курасовщина
Количество комнат	2
Тип аренды	Долгосрочная
Количество спальных мест	3
Общая площадь	48 м²
Жилая площадь	32 м²
Площадь кухни	7 м²
Санузел	Раздельный
Ремонт	Косметический
Обустройство быта	Телефон, Wi-Fi, Телевизор
В ванной	Стиральная машина
На кухне	Газовая плита, Посуда/столовые приборы, СВЧ-печь, Холодильник
Кому сдается	Парам, С детьми
Предоплата	2 месяца
Этаж	4
Этажность дома	5
Материал стен	Панельный
Состояние	Вторичное жилье

Описание

Сдам в аренду на длительный срок.

Собираемая информация:

- Цена в $;
- Количество фотографий;
- Характеристики;
- Описание.

Дополнительно в датасет добавляется ссылка на описание, чтобы после нового сбора информации не добавлять один объект дважды. И дата сбора, чтобы в перспективе можно было отслеживать изменение цен. Дата подачи

объявления не собирается, так как на куфаре часто обновляют объявления, чтобы быть выше в сортировке по новизне объявлений.

Алгоритм сбора информации:

1. Доступ к сайту через URL https://re.kufar.by/l/minsk/snyat/kvartiru-dolgosrochno?cur=USD. Проверка полученного HTML и перевод в формат парсинга библиотекой BeautifulSoup.
2. Получение количества страниц через класс kf-eaFC-63a11, проверка результата.
3. Цикл по страницам: получение URL следующей страницы через класс kf-eaFC-63a11 kf-easz-33922.
4. Поиск элементов - классов со ссылками - объявления на странице, проверка на наличие результатов поиска. Проверка, не встречалась ли эта ссылка раньше в парсинге.
5. Цикл по элементам.
6. Переход по ссылке в элементе на страницу объявления, проверка результата HTML и перевод в формат парсинга библиотекой BeautifulSoup.
7. Формирование словаря характеристика : значение с проверками результата:
    - В графе характеристик класс kf-RrYL-321f6, в котором текст класса kf-RrYf-35f08 – характеристика, а текст kf-RrYE-854c8 / kf-RrYZ-38305 – значение.
    - Класс kf-HHCF-ae1ac содержит описание квартиры.
    - Цена в долларах в тексте класса kf-RtPp-f9db0.
    - Количество фото содержится в классе kf-PJXK-96099, при их отсутствии проверяется наличие класса kf-PJXk-8e17, отвечающего за отображение отсутствия фотографии. Если фото отсутствуют, то вносится значение «0 фото».
    - Дополнительное: URL : URL страницы.
    - Дополнительное: Date : даса сбора информации.
8. Переход к следующему элементу.
9. Переход к следующей странице.
10. Вывод количества элементов и времени, потраченном на сбор информации (в среднем 35 мин, по 1,4с на каждый элемент), или сообщения об ошибке.

Следующим этапом массив из словарей преобразуется в датасет DataFrame библиотеки pandas. Проверяется наличие файла с предыдущей собранной информацией. Если он есть, то считывается и дополняется новой информацией по новым URL и перезаписывается. Если файла нет, то просто выполняется запись сырых данных.

3. ОЧИСТКА ДАТАСЕТА

Датасет содержит 1498 элементов с 39 характеристиками. Некоторые характеристики содержат пропущенные значения. Их процент представлен в таблице:

	Feature	% nans	Dtype
38	До ближайшей станции на транспорте	99.9	object
37	Пешком до ближайшей станции	99.9	object
36	Рассрочка от продавца	98.8	object
31	Возможна рассрочка	96.7	object
35	Возможен обмен	95.1	object
34	Без мебели	93.3	object
33	Студия	90.2	object
23	В новостройке	88.0	object
32	Есть проходная комната	87.1	object
15	Кому сдается	69.0	object
29	Номер и дата договора	68.4	object
22	Обустройство дома	61.1	object
6	Количество спальных мест	55.7	object
3	Метро	55.7	object
21	Окна выходят	54.7	object
16	Предоплата	44.7	object
30	Микрорайон	42.1	object
20	Год постройки	34.8	object
13	В ванной	33.5	object
9	Площадь кухни	32.8	object
11	Ремонт	31.2	object
19	Материал стен	27.5	object
14	На кухне	26.0	object
12	Обустройство быта	24.6	object
8	Жилая площадь	23.1	object
7	Общая площадь	15.1	object
28	Балкон	12.8	object
18	Этажность дома	12.7	object
17	Этаж	8.7	object
10	Санузел	4.6	object
2	Город / Район	0.1	object
4	Количество комнат	0.0	object
5	Тип аренды	0.0	object
1	Date	0.0	object
27	Количество фото	0.0	object
26	Цена $	0.0	object
25	Описание	0.0	object
24	Состояние	0.0	object
0	URL	0.0	object

Некоторые автоматически собранные характеристики не несут информации, например, количество спальных мест или проходная комната. Поэтому будут выбраны только важные описания. Также будет опущено описание с сайта от арендодателя из-за наличия практически полного описания в классах, которые собираются автоматически. Описание собиралось для возможного применения в будущем для более точного предсказания цены.

Вся информация представляется в строковом виде, поэтому необходимо извлечь числа из строк в следующих характеристиках: "Цена $", "Общая площадь", "Количество комнат", "Количество фото" (применяется приведение к 21 при бОльшем количестве), "Этажность дома", "Этаж", "Год

постройки". Удаляются элементы с ценой 5000$+, так как это выброс в данных.

Далее приводятся к общему виду следующие характеристики:

- «Агенство» – новый признак, основанный на «Номер и дата договора». Если договор есть, то 1, иначе 0.
- "В новостройке" – бинарный признак, считается за 1, если указано, за 0, если не указано. Также если возраст дома, расчитанный из года простройки, менее 10, то считается новостройкой.
- "Без мебели" – бинарный признак. Если указано, что без мебели («Да»), то 1, иначе, если не указано или написано «нет», будет 0.
- "Состояние" – бинарный признак, если указано новое, то 1, иначе 0.
- "Ремонт" – категориальный признак. «Строительная отделка» заменена на «без отделки».
- Количество "Студия" очень мало, поэтому признак удаляется, а "количество комнат" ставится в 0.
- "Санузел" – категориальный признак, замена отсутствующего значении на «раздельный», так как подразумевается, что при совместном будет указано.
- "Балкон" – количественный признак. Описание по словам переведены в цифры от 0 до 2, пропущенные значения считаются за отсутствие.
- "Предоплата" – количественный признак, обозначающий сколько месяцев она составляет. Залог считается за 1 месяц.

Оставляются после преобразования эти признаки: «Город / Район», «Метро», «Количество комнат», «Общая площадь», «Санузел», «Ремонт», «Обустройство быта», «В ванной », «На кухне», «Предоплата», «Этаж», «Этажность дома», «Обустройство до ма», «В новостройке», «Состояние», «Цена $», «Количество фото», «Балкон», «Микрорайон», «Без мебели», «Агенство».

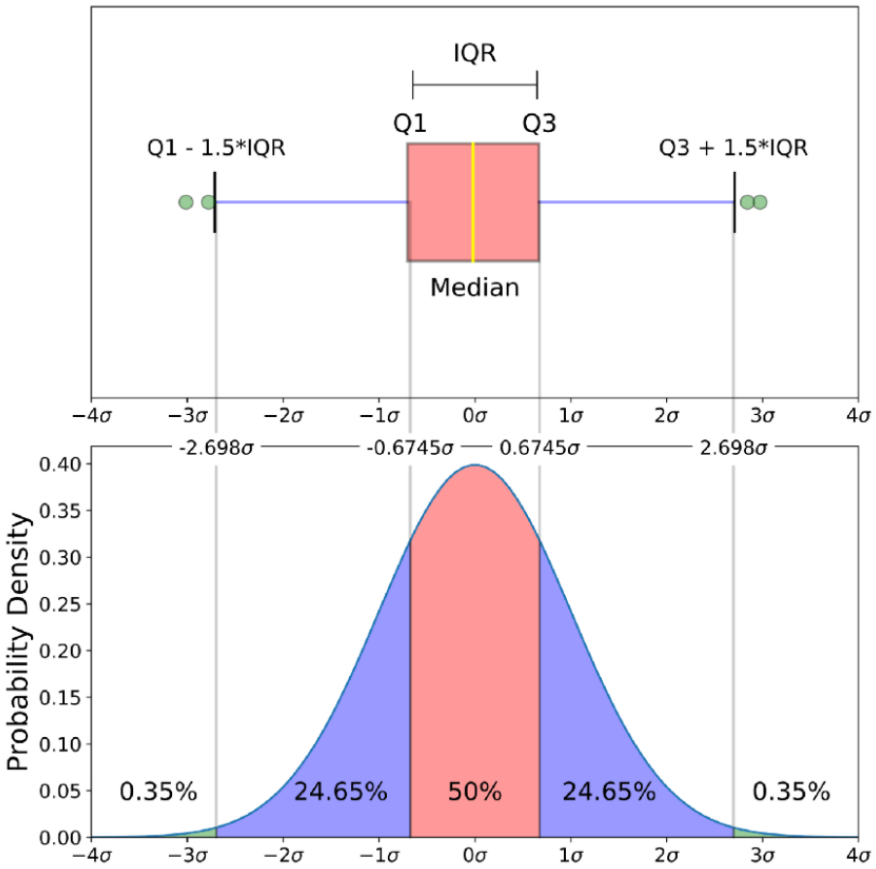
Также есть признаки, состоящие из списка предметов быта. Из них оставлены только те, что подчёркнуты, так как представляют сколько-нибудь ценную информацию о типе квартиры, в отличие от остальных:

- "Обустройство быта" - 'Утюг', 'Мебель', 'Телевизор', 'Wi-Fi', 'Телефон', 'Кондиционер'.
- "На кухне" - 'Посуда/столовые приборы', 'Посудомоечная машина', 'Холодильник' 'СВЧ-печь', 'Газовая плита', 'Электрическая плита', 'Кофеварка'.
- "В ванной" - 'Стиральная машина', 'Душевая кабина', 'Комплект полотенец', 'Джакузи', 'Фен'.
- "Обустройство дома" - 'Лифт', 'Мусоропровод', 'Огороженная территория', 'Домофон', 'Видеонаблюдение', 'Стояночное место', 'Парковка крытая', 'Пандус', 'Подвал'.

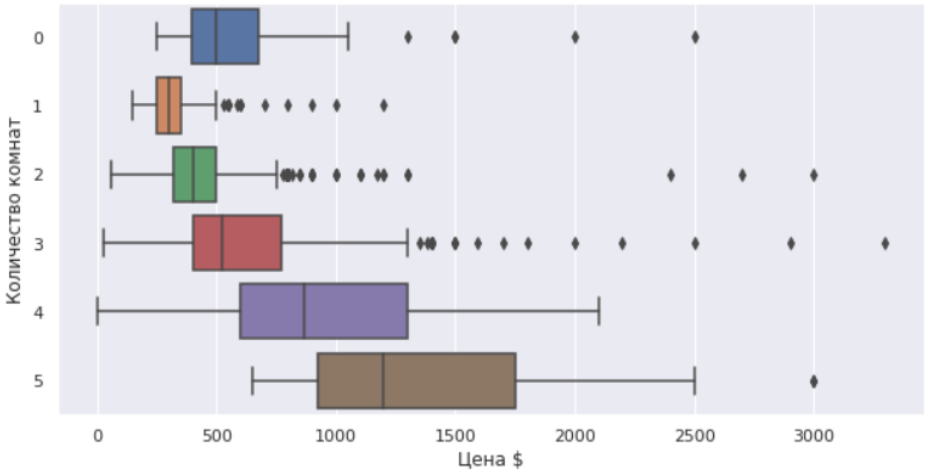
4. АНАЛИЗ ДАННЫХ

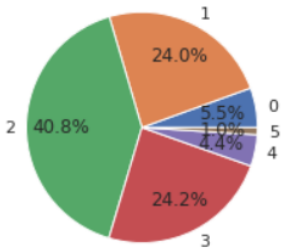
Первым делом необходимо посмотреть на распределение цены. Для корректного отображения графика будут представлены только квартиры до 1600$. Только 23 квартиры стоимостью от 1600$ до 5000$, что составляет только 1,5%. Как видно, основной диапазон от 200$ до 700$, далее идёт только элитное жильё.

Необходимо сказать про тип графика бокс-плот. Наиболее понятно он изображён на рисунке.

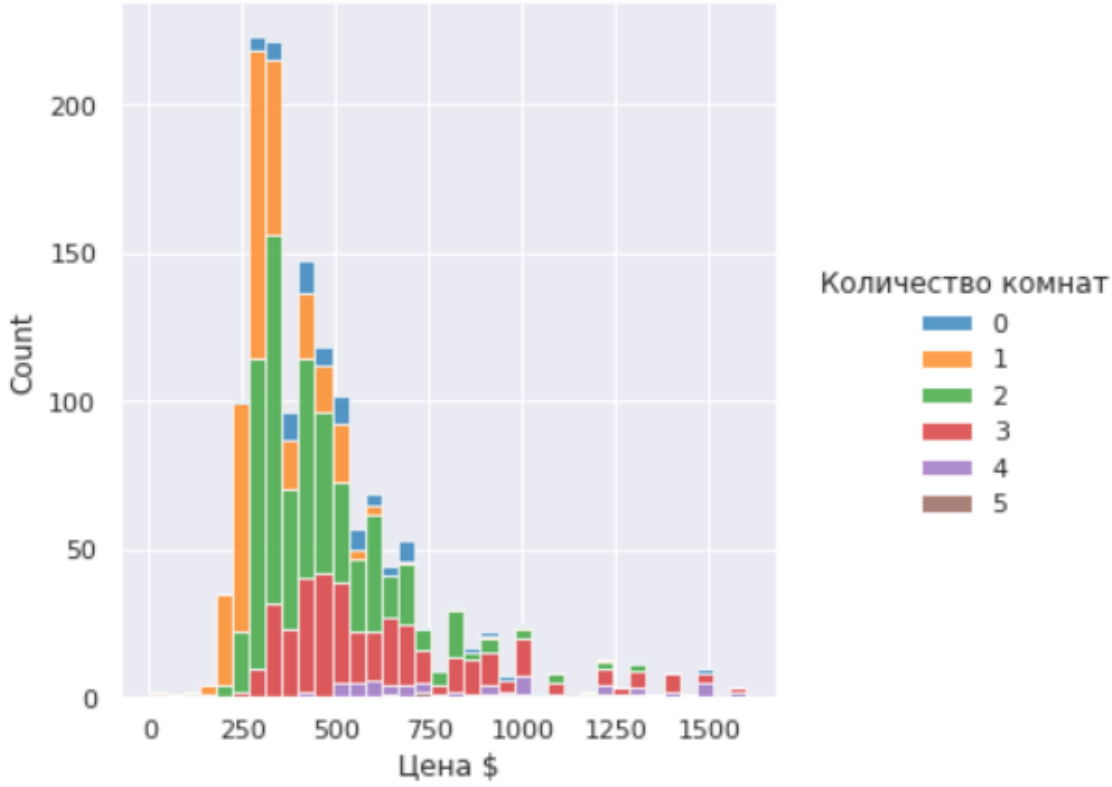


Далее влияние количества комнат на цену, где 0 – студия. Основные квартиры 1-3 комнаты. Примечательно, что стоимость студии больше, чем полноценной одной квартиры. И соответственно, цена растёт по мере увеличения количества комнат.

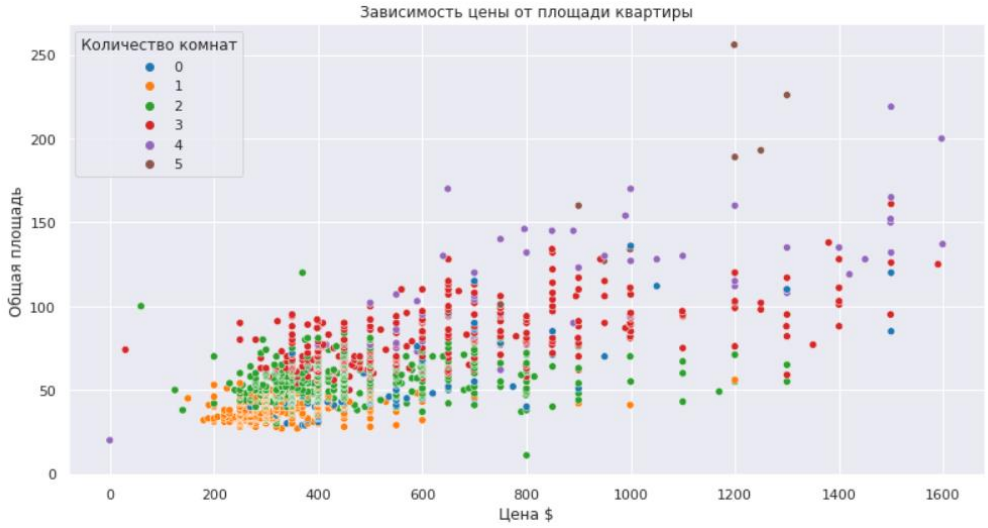


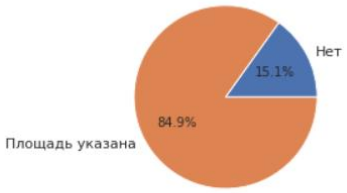


Также можно посмотреть на распределение количества комнат в каждом ценовом сегменте. Как видно с этого графика, резко уменьшается количество однокомнатных квартир с ростом цены, а студии расположены практически равномерно. Это показывает, что важно не только количество комнат, но и ремонт и оснащение квартиры.

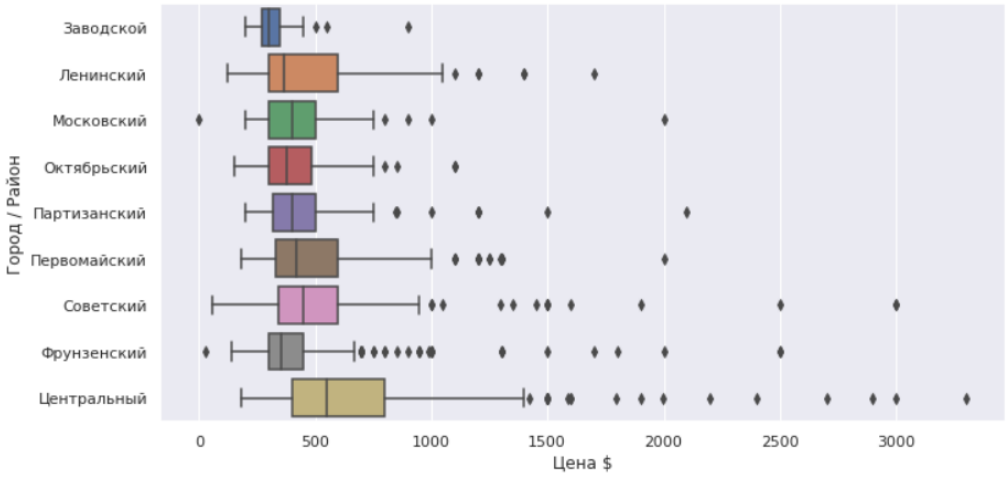


Кроме количества комнат можно посмотреть на площадь квартиры. Она не указана в 15% объявлений. По цвету точек на графике площади видно, как сильно разбросана площадь квартиры при одном количестве комнат. Также можно заметить сильную корреляцию площади с ценой и уже на основании этого графика быть уверенным в полезности этого признака.

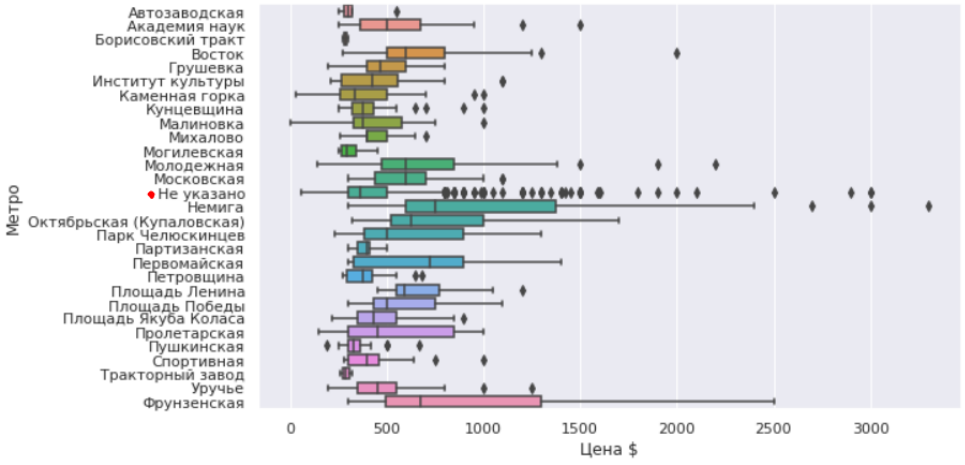


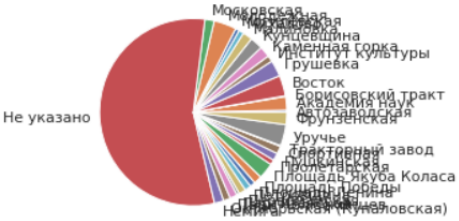


По распределению районов видно, что они не особо влияют на цену, однако наибольшая цена в Центральном, наименьшая в Заводском.

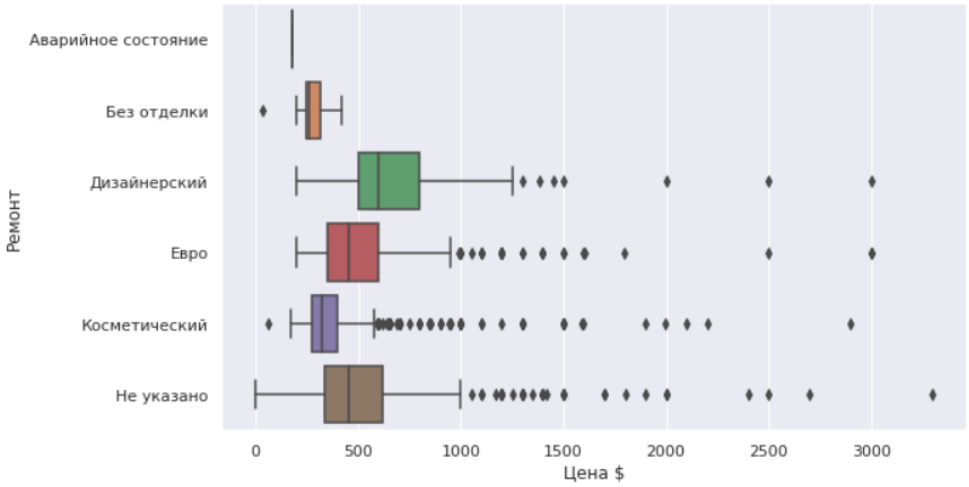


Немного более половины объявлений без указания метро, поэтому этот признак плохо иллюстрирует зависимость цены от станции, потому что слишком мало элементов для каждой станции.

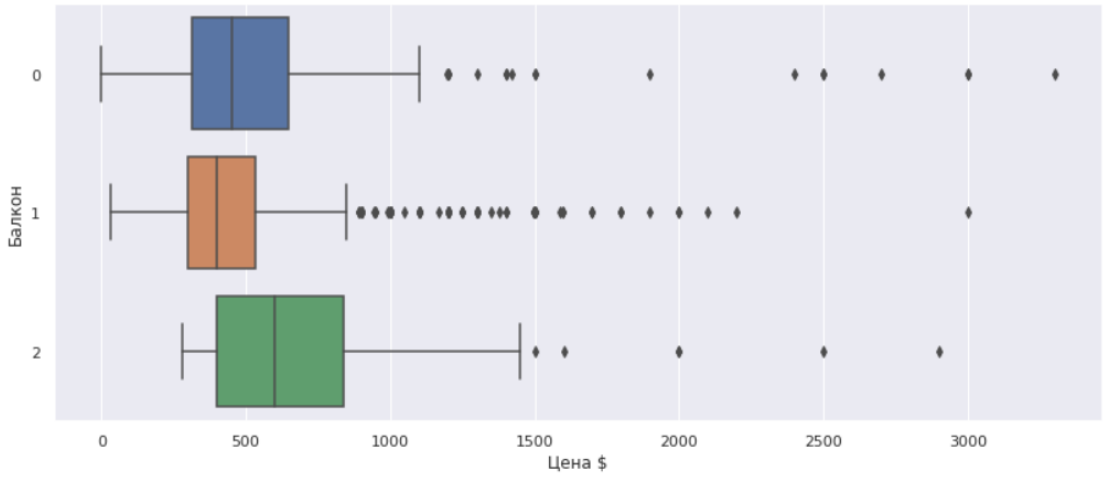


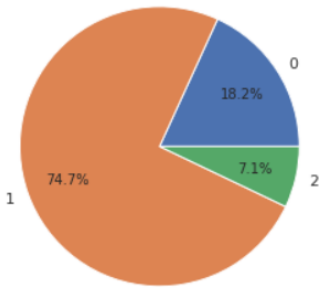


Следующий признак – ремонт. Только один элемент в аварийном состоянии. 31% не указали ремонт, поэтому можно предположить, что он «обычный», т.е. косметический. Ожидаемо, что без отделки квартиры дешевле, а с дизайнерским ремонтом дороже, хотя выбросы не зависят от ремонта.

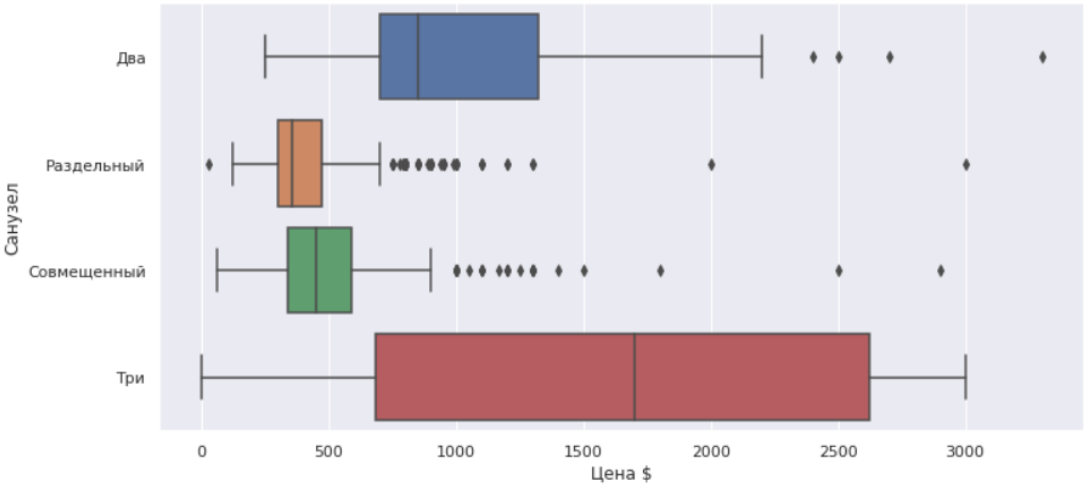


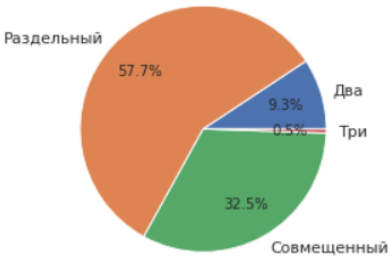
Ещё одной характеристикой квартиры служит балкон. В 75% квартир он есть, в 18% нет, но его наличие не сильно влияет на цену, кроме как их два. Но в этом случае обратная зависимость: влияют не балконы, а размер квартиры, который подразумевает несколько балконов.



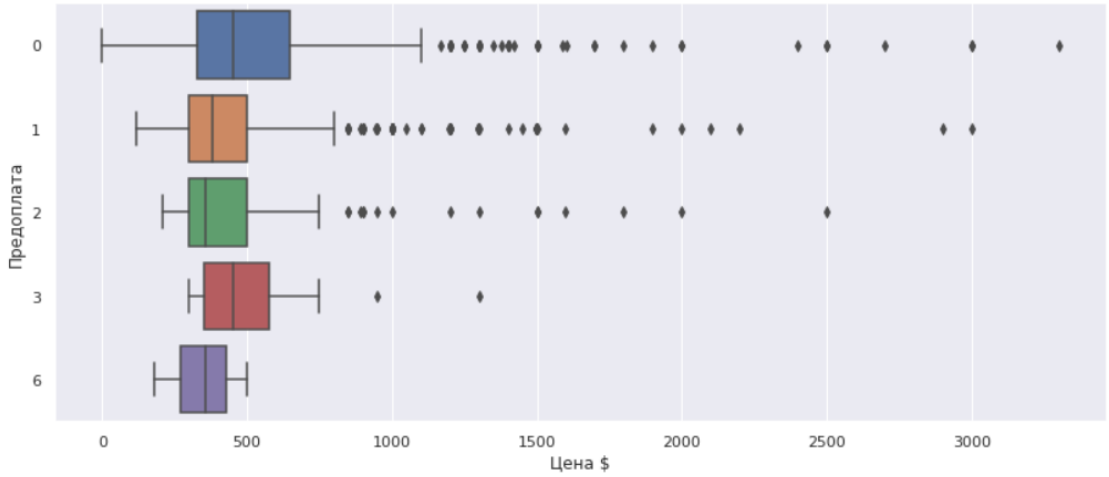


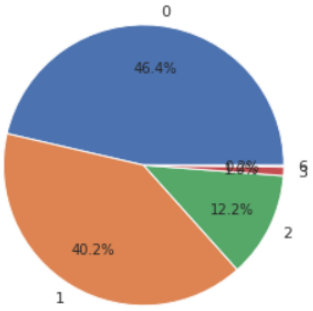
Интересно посмотреть, как наличие совмещённого санузла влияет на стоимость. Он есть преимущественно в однокомнатных квартирах, поэтому удивительно, почему для них цена в среднем выше, чем для раздельных. Хотя процент совмещённого санузла довольно высок – 32,5%, можно предположить, что в аренду часто сдаются старые дома-хрущёвки, а цена в них выше от ремонта.



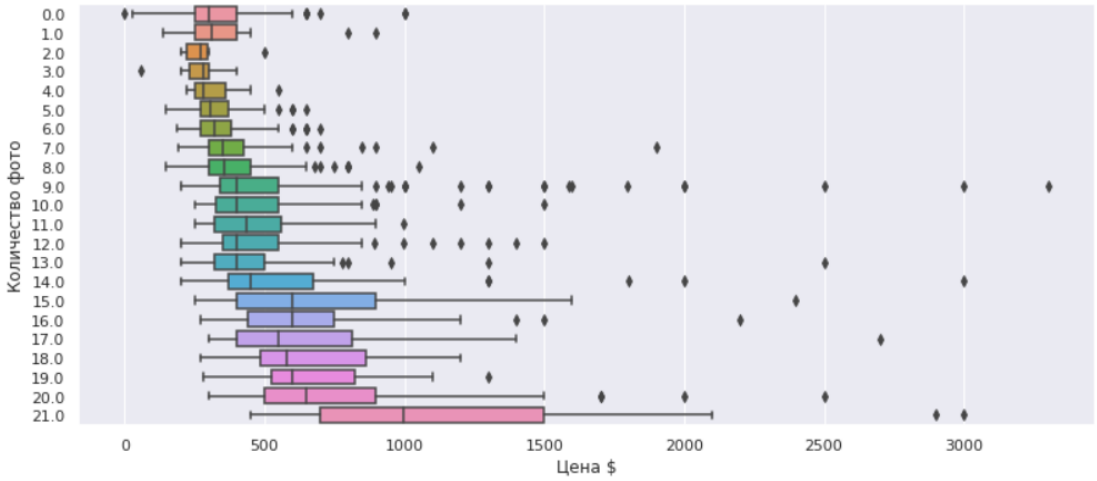


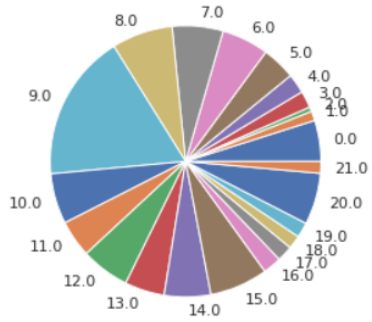
Предоплата – самый ненадёжный признак, так как мы предположили о размере при пропущенном значении (отображается как 0).



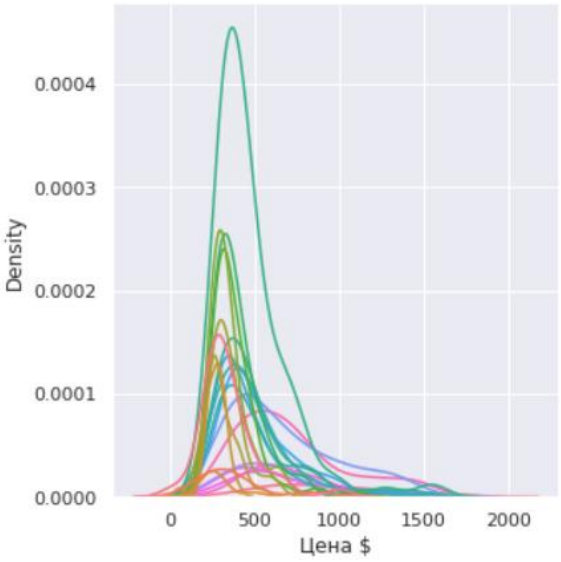


Также интересно посмотреть на зависимость цены от количества фото. Очень чётко видна зависимость цены от количества фото. Это обуславливается необходимостью хорошо показать квартиру, которую хотят сдать с большей стоимостью.

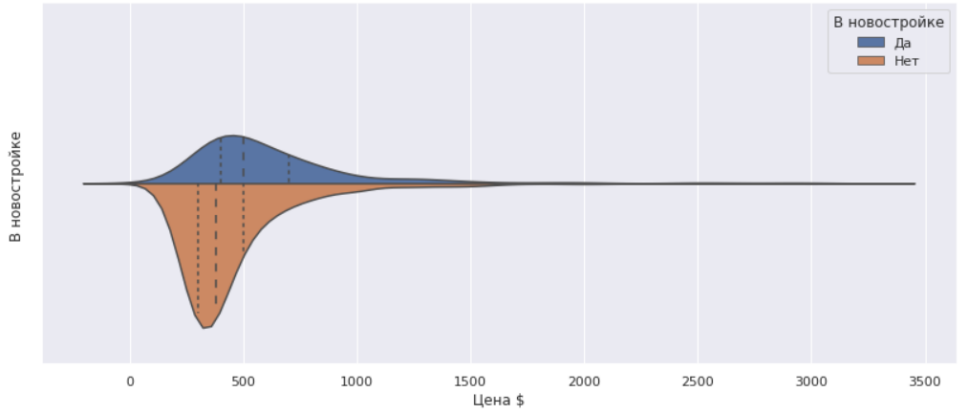


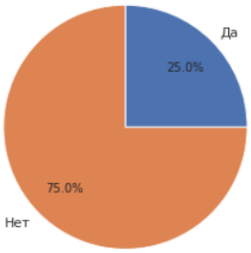


На этом графике видно, как распределено количество фото. Они все стремятся к цифре 250$. И только розовые линии (соответствуют большему количеству фото) отклоняются к большей цене.

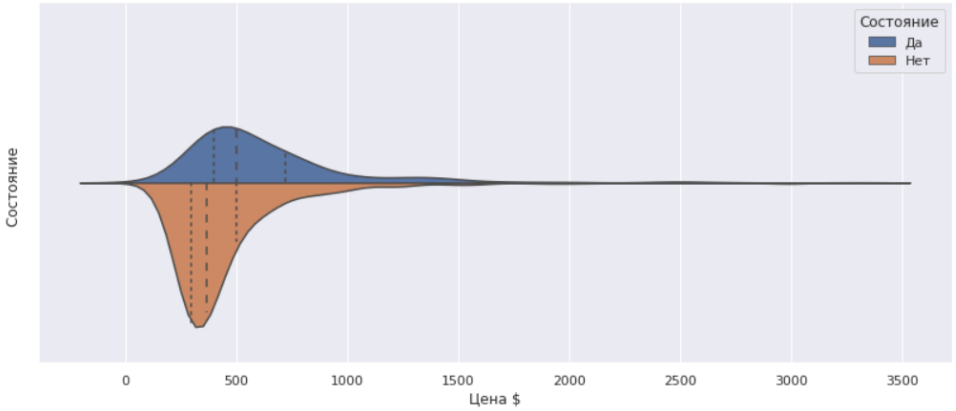


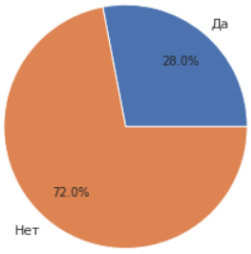
Далее следующий тип графиков. На данном изображено, как зависит цена от новостройки. Только 25% домов считаются новостройками, но цены в низ не выше, чем в старых домах.



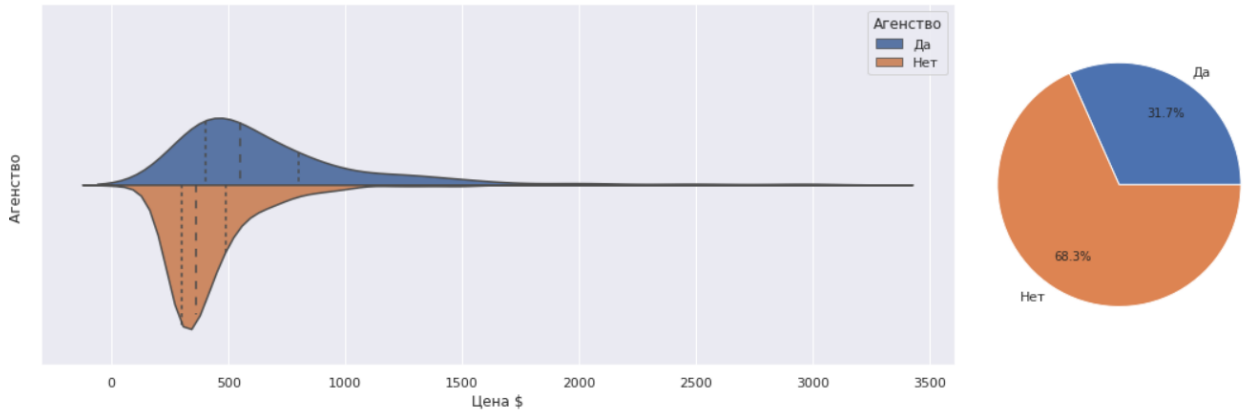


Новая квартира, видимо, с хорошим ремонтом, стоит в среднем не намного дороже обычной вторичной.

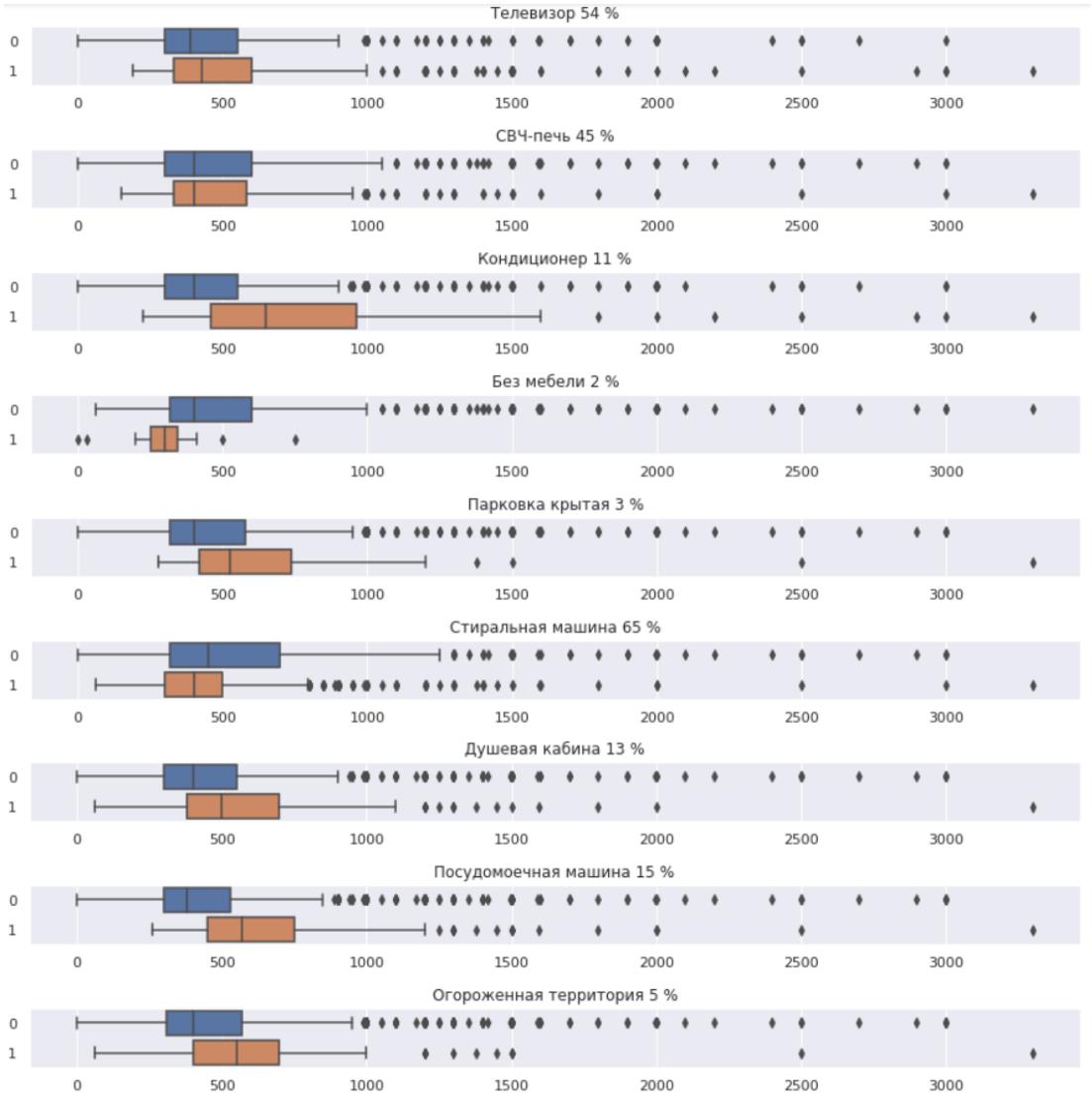




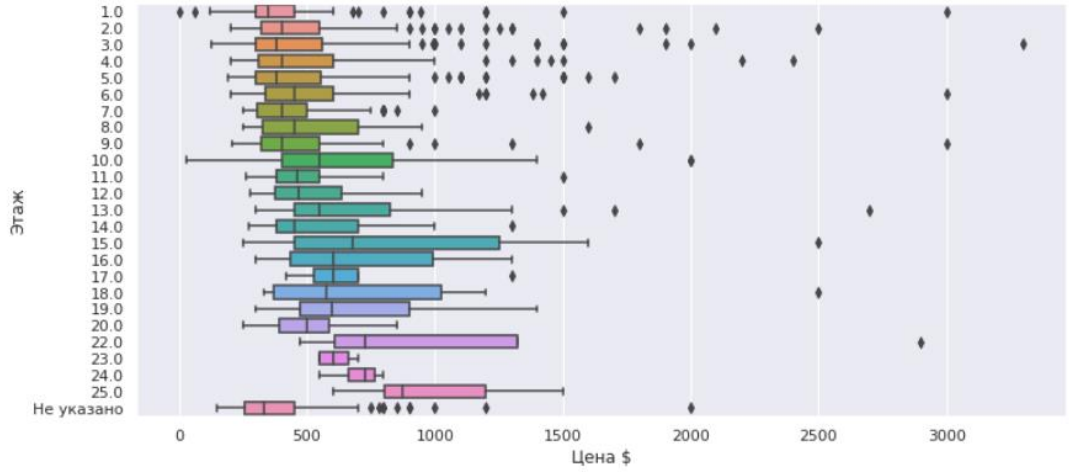
Тот же график распределения и для агентства. Треть рынка занимают агенты, а не владельцы квартир. Стоит учитывать, что многие не указывают напрямую, что они агентства.

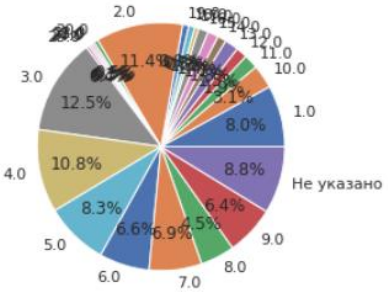


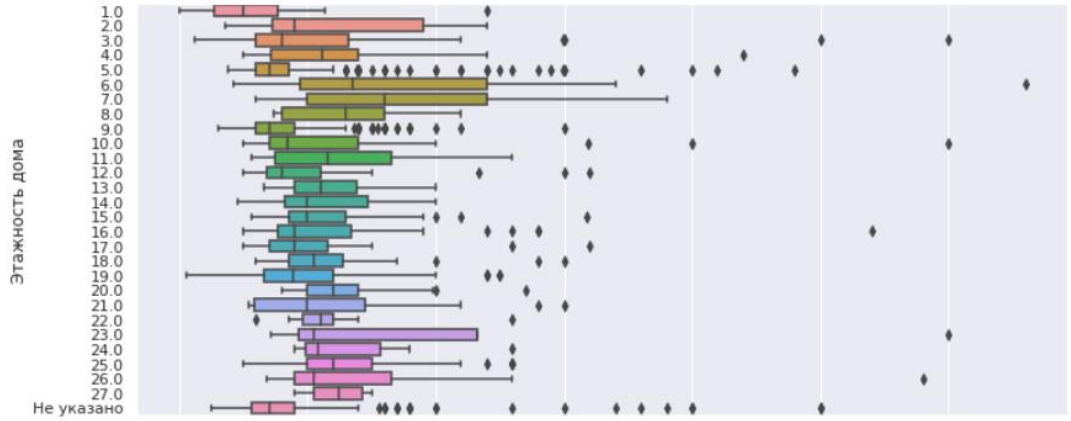
Последнее идёт оснащение квартиры. На графиках указано, в каком проценте объявлений указано о наличии этот оснащения.

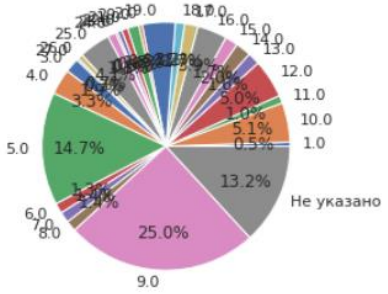


Что касается этажа и этажности дома, то не прослеживается логики зависимости цены. Хотя, чем выше этаж, тем в среднем больше цена.





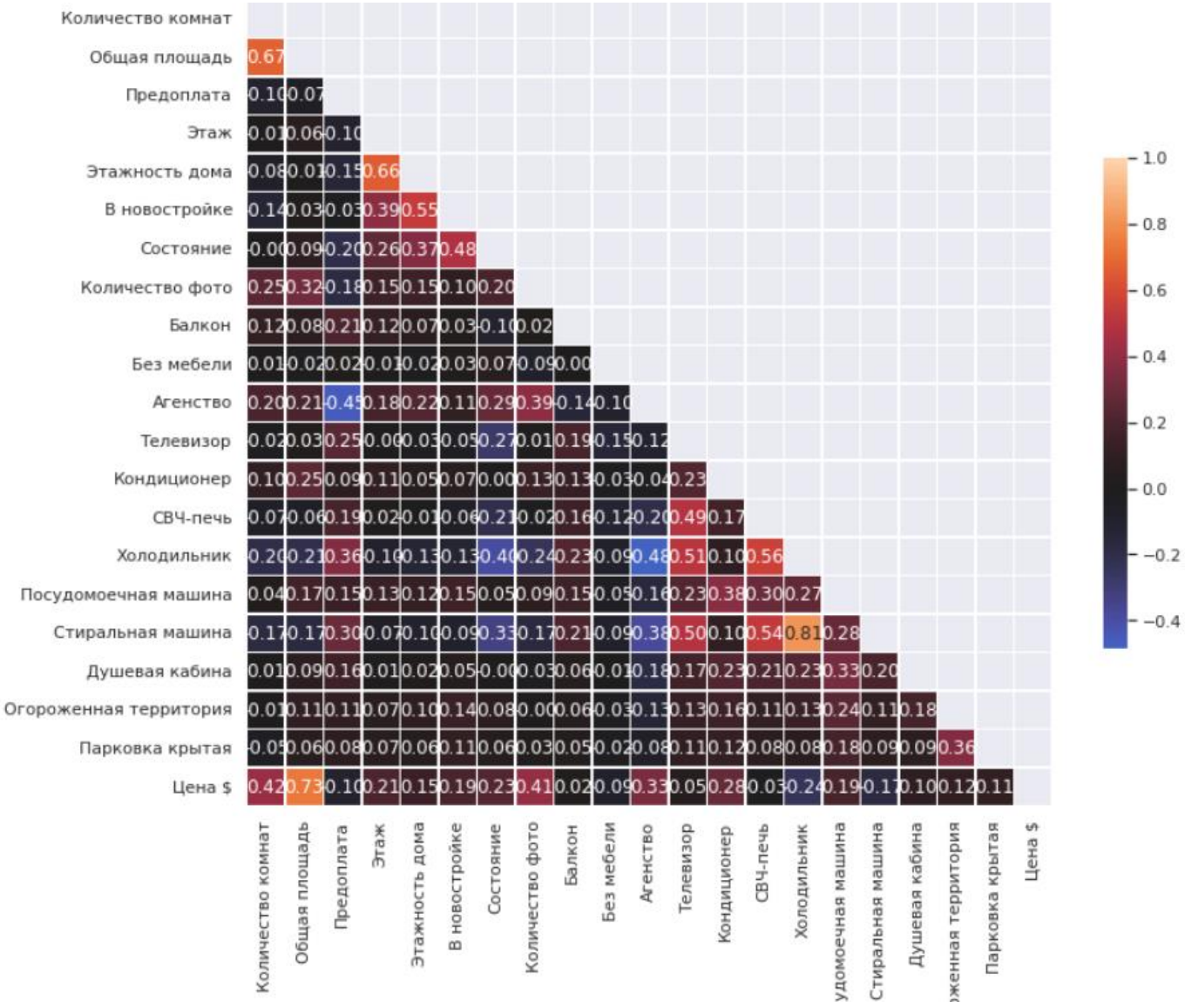




5. ПРЕДСКАЗАНИЕ ЦЕНЫ

Так как некоторые признаки до сих по содержали пропущенные значения, эти значения необходимо восстановить. Это признаки площади, этажа и этажности, которые заполняются средними по их классу значениями (для площади) и средними по этажам и этажности.

Чтобы лучше понять данные, была построена матрица корреляции. Как мы и предполагали, основную роль в формировании цены играет площадь квартиры, которая к тому же хорошо коррелирует с количеством комнат. Также корреляция есть с количеством фото. Лучше понять влияние признаков поможет их анализ после построения моделей.



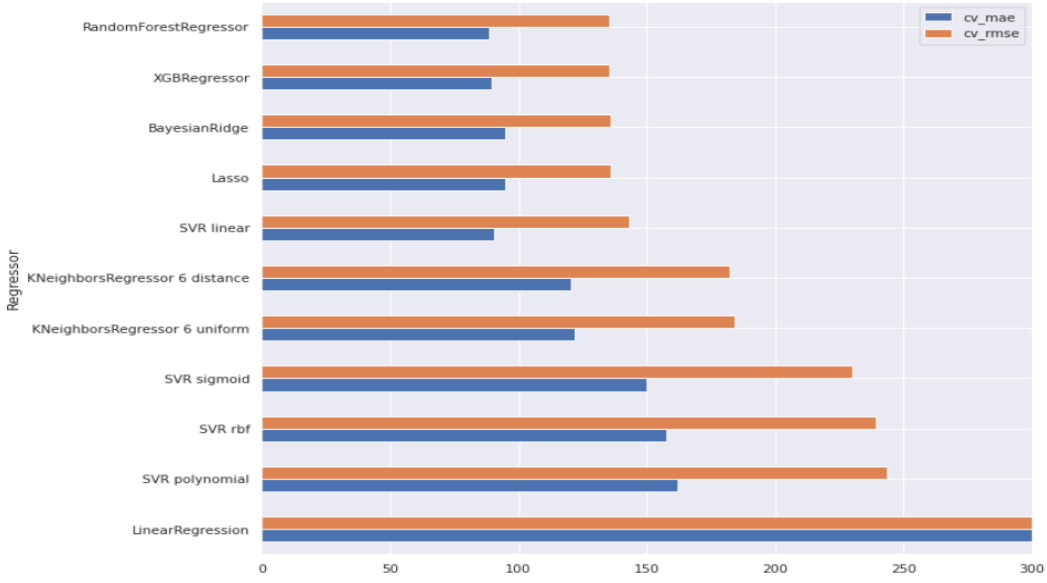
Так как предсказание – регрессия, метрики оценки классические: средняя абсолютная ошибка, средняя квадратичная ошибка и средняя среднеквадратичная ошибка.

С помощью StandardScaler данные масштабируются.

Датасет был разбит на две выборки: 70% и 30% - обучающая и тестовая соответственно. Также для сравнения и точного результата использовалась кросс-валидация с 5 разбиениями (для теста получается 20%).

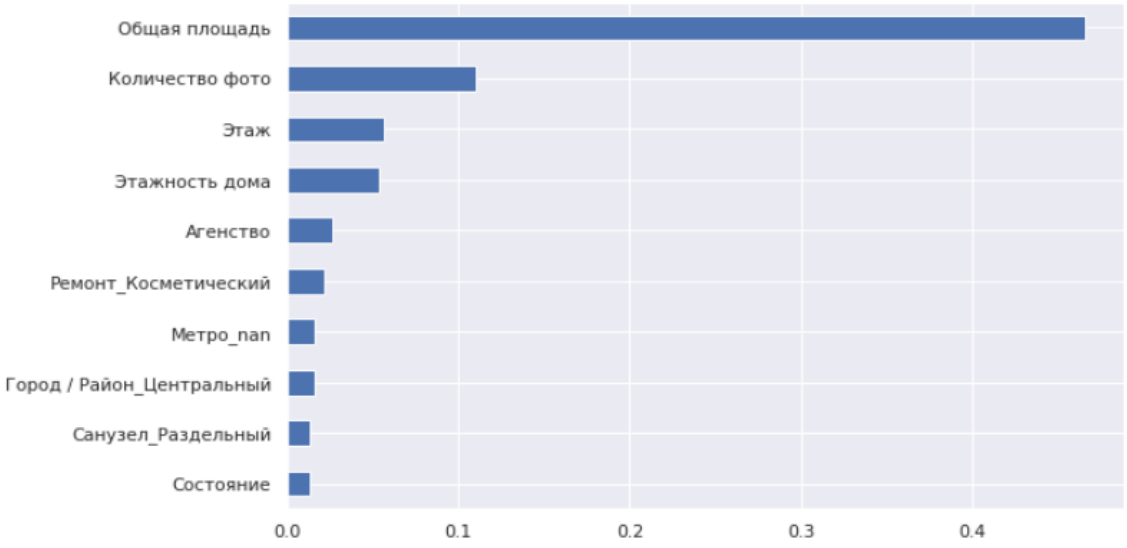
Для регрессии были выбраны простые классические модели машинного обучения. Их название, выбранные гиперпараметры, время работы и эффективность представлена в таблице ниже. Ожидаемо, что ансамблевые модели показали лучший результат.

Regressor	Time, ms	cv_mae	cv_mse	cv_rmse	mae	mse	rmse
RandomForestRegressor	4.93	8.870000e+01	1.863500e+04	1.356000e+02	9.042000e+01	1.858924e+04	1.363400e+02
XGBRegressor	1.39	8.950000e+01	1.864520e+04	1.357000e+02	9.235000e+01	1.925382e+04	1.387600e+02
BayesianRidge	0.23	9.470000e+01	1.870080e+04	1.359000e+02	9.935000e+01	1.929003e+04	1.388900e+02
Lasso	0.06	9.470000e+01	1.874060e+04	1.361000e+02	9.799000e+01	1.861594e+04	1.364400e+02
SVR linear	0.62	9.070000e+01	2.099900e+04	1.433000e+02	9.513000e+01	2.234496e+04	1.494800e+02
KNeighborsRegressor 6 distance	0.10	1.205000e+02	3.358880e+04	1.823000e+02	1.199600e+02	3.371083e+04	1.836100e+02
KNeighborsRegressor 6 uniform	0.11	1.219000e+02	3.429010e+04	1.841000e+02	1.205900e+02	3.462852e+04	1.860900e+02
SVR sigmoid	0.74	1.498000e+02	5.403780e+04	2.302000e+02	1.530200e+02	5.784715e+04	2.405100e+02
SVR rbf	0.88	1.577000e+02	5.829140e+04	2.392000e+02	1.614600e+02	6.259566e+04	2.501900e+02
SVR polynomial	0.53	1.622000e+02	6.043480e+04	2.435000e+02	1.633000e+02	6.388533e+04	2.527600e+02
LinearRegression	0.17	9.338028e+10	1.264878e+25	1.783580e+12	3.820287e+11	6.406919e+25	8.004323e+12



Распределение реальных и предсказанных результатов можно увидеть на этом графике. По нему видно, что большие значения модель предсказывает крайне плохо, но неплохо справляется с основной массой цен. Средняя абсолютная ошибка составляет 89$, средняя среднеквадратичная 135$ (в этом параметре сильно влияют ошибки на больших значениях).

Последний график иллюстрирует, как сама модель оценивает важность признаков. На графике изображено только первые 10 по важности, важность остальных не так важна. Как и предполагалось, наиболее важными признаками являлись площадь и количество фото.



ВЫВОД

В проекте был реализован сбор данных о аренде квартир в Минске с сайта Куфар, покрытый тестами, очищен и подготовлен датасет и выполнен анализ данных. Анализ данных был представлен в виде наглядных графиков с комментариями по поду закономерностей и предположением о влиянии признаков друг на друга и цену аренды и в дальнейшем их подтверждением. Дополнительно были обучены модели для предсказания цены, наиболее точной оказался случайный лес с ошибкой MAE 89$ и RMSE 135$.