Федеральное государственное образовательное бюджетное  
 учреждение высшего образования

**«Финансовый университет при Правительстве Российской Федерации»**

**Московский филиал Финуниверситета**

**Кафедра** «Информационные технологии и анализ больших данных»

**ДОМАШНЕЕ ТВОРЧЕСКОЕ ЗАДАНИЕ**

**по дисциплине** «\_\_\_\_\_\_\_\_\_\_\_\_Эконометрика\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_»

**на тему:** «\_\_Эконометрический\_анализ\_рынка\_жилой\_недвижимости\_\_»

**Выполнила студентка** \_\_2\_\_ курса,

группы \_\_\_\_ПМ22-3\_\_\_\_\_\_\_,

формы обучения\_\_очная\_\_\_\_\_\_\_

\_\_\_Перминова\_Мария\_Александровна\_\_

**Проверил преподаватель:**

\_\_\_Михайлолва\_Светлана\_Сергеевна\_\_\_\_\_

\_\_\_\_профессор ДАДиМО\_\_\_\_\_\_\_

Москва 2023 г.

Содержание:

[**1) Актуальность, цель, задачи, гипотезы работы**](#_cxgicg1d9kpl) **3**

[**2) Данные**](#_zdymtickqqu0) **3**

[**3) Алгоритм автоматизированного создания спецификации модели**](#_s8gr39eny8dy) **5**

[a) Автоматизированный отбор значимых факторов с помощью метода включения](#_vbhit9z3jge4) 5

[b) Подсчёт основных статистических показателей для всех моделей: статистики Стьюдента для коэффициентов регрессии, тест Фишера для всей модели](#_oq0hgv6iq5tg) 6

[c) Проверка на наличие гетероскедастичности](#_lmvwrdfjreqf) 6

[d) Проверка на наличие автокорреляции](#_o23cdq1stbee) 7

[e) Проверка на наличие мультиколлинеарности](#_h01gbqpqk0sh) 7

[**4) Оценка качества модели**](#_s5yxg72mjcd6) **7**

[**5) Заключение**](#_fl7zoo4fz536) **8**

[Вывод: как результат мы получили таблицу с помощью которой можно делать предсказания о стоимости квартиры на основе её характеристик. (см файл df\_factors2.csv в папке решения)](#_eyupbfjix61u) 8

[Способы применения работы:](#_73jpkx2n49wf) 8

[Что пробовала, но не получилось:](#_y01nt4p522hx) 9

[Полезные ресурсы:](#_fms5zr9zprdw) 9

## Актуальность, цель, задачи, гипотезы работы

Актуальность: анализ рынка жилой недвижимости и дальнейшее построение моделей для прогноза стоимости важная задача, в качественной реализации которой заинтересованы многие группы лиц: риэлторы, люди, планирующие приобретение жилья, инвесторы, экономисты и пр.

Цель работы: Создать программу, которая автоматизирует процесс построения спецификаций эконометрических моделей. С помощью полученного алгоритма на основе данных по регионам России за 2018-2021 года создать для каждого из регионов качественную модель.

Задачи:

1. Найти готовый датасет, в котором будут представлены данные о российском рынке недвижимости (мы будем рассматривать только квартиры).
2. Предобработать данные, посмотреть статистики, сделать выводы
3. Написать алгоритм автоматизированного создания спецификации модели. Данный алгоритм должен включать в себя: отбор факторных переменных, проверку на значимость коэффициентов и уравнения в целом, проверку на наличие гетероскедастичности, автокорреляции и мультиколлинеарности.
4. Выбрать из датафрейма несколько наблюдений, сделать для них предсказание, рассчитать ошибку, сделать вывод о качестве моделей.

Гипотеза: у каждого региона в силу экономических, географических и прочих особенностей цены на квартиры формируются по-разному. Определенные значения факторов могут сильно увеличивать стоимость квартиры.

## 

## Данные

В качестве данных для работы были выбраны данные с платформы kaggle о российской недвижимости по 84 регионам за 2014-2021 года.

<https://www.kaggle.com/datasets/mrdaniilak/russia-real-estate-20182021>

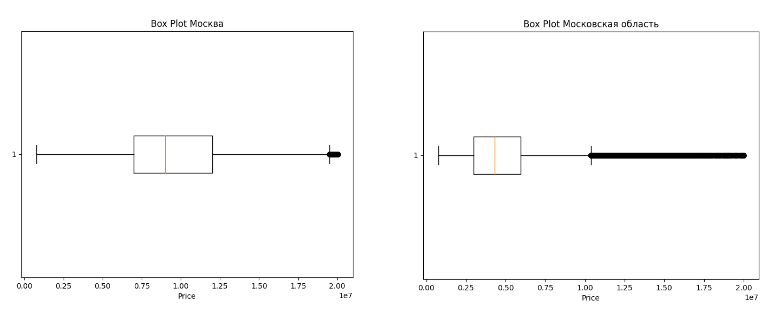
Данные имеют следующие поля:

* `date` - дата публикации объявления;
* `time` - время публикаци объявления;
* `geo\_lat` - значение координаты (широта);
* `geo\_lo` - значение координаты (долгота);
* `region` - код региона РФ;
* `building\_type` - Тип здания. 0 - Прочее. 1 - Панельный дом. 2 - Монолит. 3 - Кирпичный. 4 - Блочный. 5 - Деревянный;
* `object\_type` - Тип квартиры. 1 - Вторичное жилье; 11 - Новая квартира в новостройке;
* `level` - Этаж, на котором находится квартира;
* `levels` - Количество этажей;
* `rooms` - Количество жилых комнат. Если значение -1 - это значит, что квартира является "студией";
* `area` - Совокупная площадь квартиры;
* `kitchen\_area` - Площадь кухни;
* `price` - Цена в рублях РФ.

Была проведена предобработка данных. Были убраны выбросы, добавлен столбец region\_name, путём получения с помощью сервиса (<https://www.geonames.org/>) по значениям широты и высоты названий регионов России. Среди регионов было решено оставить те, по которым есть хотя бы 500 наблюдений, т.к. например для Сахалинской области в датасете было только 8 наблюдений, что крайне мало.

Также данные были проверены на наличие пустых значений, таковых не оказалось. После предобработки для каждого из регионов были построены гистограммы распределения цен и графики box plot. Данные графики уже на этапе анализа данных могут многое сказать о рынке недвижимости того или иного региона.

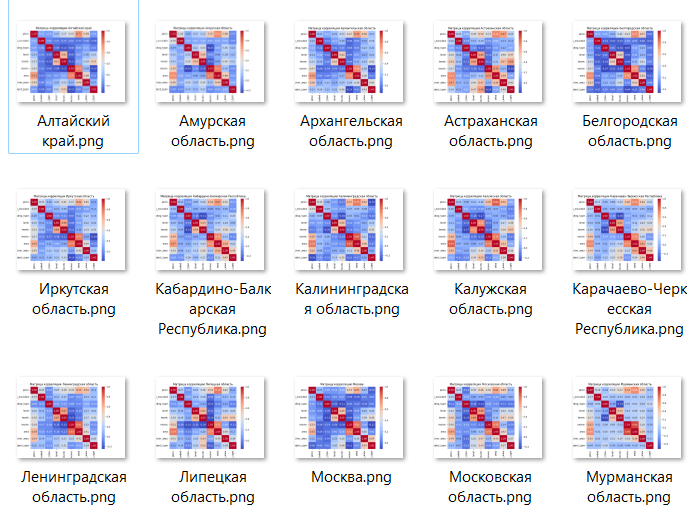
Например для Москвы средняя стоимость квартиры примерно 8 млн. руб. В то время как для Московской области средняя цена приблизительно 5 млн. руб.





А если посмотреть на графики распределения цен на квартиры, будет видно, что многие из них имеют правостороннюю асимметрию.

Также были построены матрицы корреляции и матрицы частных коэффициентов корреляции для удобства ручного анализа.



## Алгоритм автоматизированного создания спецификации модели

### Автоматизированный отбор значимых факторов с помощью метода включения

В качестве подхода к отбору факторов для модели я использовала пошаговый метод.

1) На первом шаге мы строим модель парной регрессии с фактором, коэффициент корреляции у которого с целевой переменной максимальный.

2.1) Далее на каждом шаге цикла выбирается тот фактор, добавление которого в модель увеличивает коэффициент детерминации наибольшим образом.

2.2) Важно, что если рассматриваемый фактор имеет сильную корреляционную связь, (более 0.8) с другими, уже отобранными факторами, такой регрессор в модель не включаем.

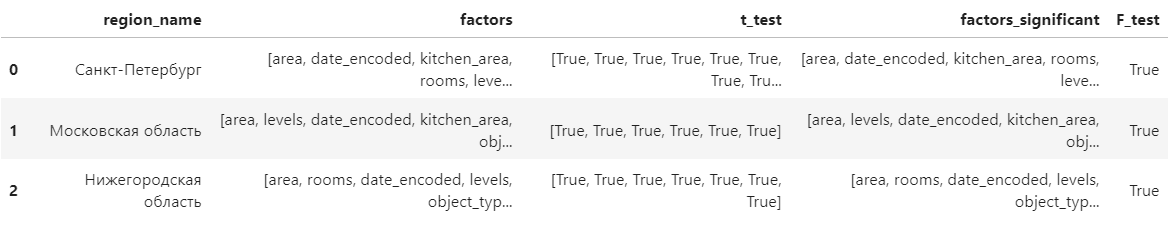
3) В случае малого количества факторных переменных, мы можем менять значение переменной level, которая отвечает за минимальный прирост коэффициента детерминации при добавлении нового фактора в модель

Как результат мы получаем датафрейм, где каждому региону соответствует список факторных переменных.



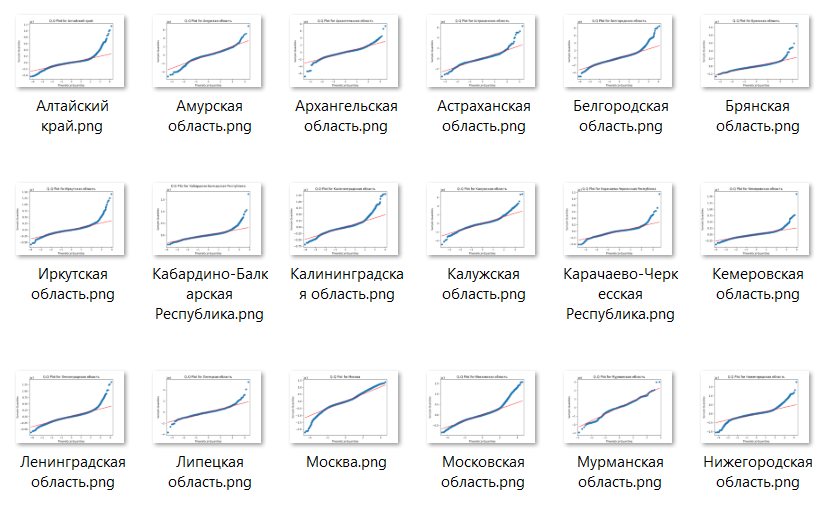
### Подсчёт основных статистических показателей для всех моделей: статистики Стьюдента для коэффициентов регрессии, тест Фишера для всей модели

Далее для каждого региона строим модели множественной линейной регрессии с помощью пакета statsmodels, рассчитываем t статистики Стьюдента для каждого из факторов. Сохраняем списки статистически значимых факторов для каждого из регионов. Далее опять строим модели множественной линейной регрессии, но уже беря только статистически значимые факторы, рассчитываем F статистику Фишера для всего уравнения в целом. Результат данного этапа выглядит следующим образом:



### Проверка на наличие гетероскедастичности

Для проверки на наличие гетероскедастичности, было решено использовать графический метод (Q-Q plot), а также тест Шапиро-Уилка. Данный тест хорошо себя показывает как на малых, так и на больших выборках, в случае наших данных. Тест не выполняется ни для одного из регионов. Это показалось подозрительным, поэтому я провела ещё тест Хетера-Бройша-Пагана, результат тот же. Наличие гетероскедастичности я могу объяснить тем, что в наших данных присутствуют наблюдения для разных квартир, и, условно, нашу выборку можно разбить на более мелкие подвыборки. Посмотрим как это скажется на предсказательной способности модели.



### Проверка на наличие автокорреляции

В качестве теста на автокорреляцию был выбран тест Дарбина-Уотсона, все предпосылки для его применения выполняются: гетероскедастичность отсутствует, нет ошибок в спецификации моделей. Все регионы прошли тест, автокорреляция отсутствует.

### Проверка на наличие мультиколлинеарности

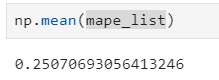
Учитывая, что при отборе факторов для каждой модели при прямом проходе мы добавляли регрессор в модель только в том случае, если он не коррелировал с остальными, мы сильно снизили вероятность возникновения мультиколлинеарности. Будем использовать VIF тест. Высокий VIF для одного из регрессоров означает, что этот регрессор сильно коррелирован с другими переменными в модели. Факторы с высоким значением VIF мы убираем.

Также в самом конце добавим в нашу единую таблицу столбец с коэффициентами регрессии. В результате мы получили таблицу, в которой содержится вся необходимая информация для построения моделей.

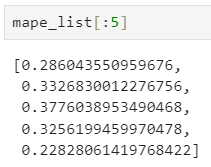
## Оценка качества модели

Выберем для каждого региона рандомные 300 строк, сделаем предсказания и оценим качество модели с помощью метрики MAPE.

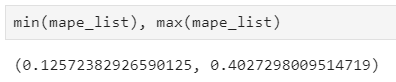
Средняя ошибка модели 0.25, что является хорошим результатом.



Вот значения MAPE для первых 5 регионов:



Минимальные и максимальные значения ошибки. В наилучшем случае модель ошибается в 10% наблюдений, в наихудшем в 40%



## Заключение

### Вывод: как результат мы получили таблицу с помощью которой можно делать предсказания о стоимости квартиры на основе её характеристик. (см файл df\_factors2.csv в папке решения)



### Способы применения работы:

1. Если к имеющимся данным добавить различные экономические факторы с росстата, например рождаемость в регионе, уровень развития строительного бизнеса и пр., можно будет выявить нетривиальные закономерности между регионом и факторами, которые влияют на стоимость квартир.
2. Если спарсить данные с 2021 по 2023 года и добавить к нашим данным, то можно использовать алгоритм для предсказания стоимости квартир в регионах на следующий год. Это может быть полезно риелторам, а также людям, планирующим покупку жилья в будущем.
3. Несмотря на то, что наличие гетероскедастичности не сильно испортило качество модели, в будущем будет полезно научиться её смягчать. Для этого можно использовать как классические подходы, так и пытаться строить модели на данных в разрезе региона и других факторов.

### Что пробовала, но не получилось:

1. Парсить данные. За основу было взято готовое решение (<https://github.com/lenarsaitov/cianparser>), однако данный парсер извлекал из данных не все доступные поля (метро, например), а также использование сторонних прокси, не решало проблему появления капчи. В будущем планируется написание собственного парсера, способного извлекать информацию о расстоянии до метро, название района, где находится квартира и пр.

### Полезные ресурсы:

1. <https://www.kaggle.com/datasets/mrdaniilak/russia-real-estate-20182021>
2. <https://studfile.net/preview/2829140/page:8/>
3. <https://github.com/romanakentev/real-estate/blob/main/Real_Estate__ENG.ipynb>
4. <https://www.hse.ru/mirror/pubs/share/423684138.pdf>
5. https://github.com/lenarsaitov/cianparser