|  |  |
| --- | --- |
| **Автор:** | **Потапова Елена Владимировна** |
| **Программа обучения:** | **Аналитика данных** |
| **Группа:** | **DA-16** |

Тема дипломного проекта:

**“Римские каникулы: анализ рынка краткосрочной аренды недвижимости в Риме по данным сервиса Airbnb**

**(исследование данных, составление рекомендаций)”**

Данное исследование в первую очередь может быть интересно для самостоятельных путешественников, пользователей сервиса Airbnb, планирующих поездку в Рим и желающих не прогадать с выбором наилучшего жилья для отпуска

**Источник данных:** <http://insideairbnb.com/get-the-data.html>

Об источнике: Inside Airbnb - это независимый некоммерческий набор инструментов и данных, который позволяет пользователям узнать, как сервис Airbnb на самом деле используется в 80 городах по всему миру. Проект изначально был представлен активистом Мюрреем Коксом, в рамках борьбы с рынком незаконной аренды в Нью-Йорке, но в дальнейшем вырос в международный проект. В данных используется только общедоступная информация, собранная с веб-сайта Airbnb. Информация о местоположении для объявлений анонимизируется Airbnb (на практике это означает, что местоположение объявления на карте или в данных будет находиться на расстоянии 0–450 футов (150 метров) от фактического адреса).

**Содержание датасета до обработки**

1. Файл **listings.csv** - основные данные для обработки и анализа, содержат информацию об опубликованных сервисом Airbnb объявлениях за период с 17.12.2020 по 05.01.2021 ([ссылка на код](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=MedJyXOlUq6c&line=2&uniqifier=1))

Количество строк: **28151**

Количество столбцов: **74**

[**Чтение данных:**](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=8RO76L0hfybJ&line=2&uniqifier=1)

Данные можно разделить на смысловые модули: **данные о хостах** ('host\_id', 'host\_url', 'host\_name', 'host\_since', 'host\_location', 'host\_about', 'host\_response\_time', 'host\_response\_rate', 'host\_acceptance\_rate', 'host\_is\_superhost', 'host\_thumbnail\_url', 'host\_picture\_url', 'host\_neighbourhood', 'host\_listings\_count', 'host\_total\_listings\_count', 'host\_verifications', 'host\_has\_profile\_pic', 'host\_identity\_verified'), **данные об объявлениях** ('id', 'listing\_url', 'scrape\_id', 'last\_scraped', 'name', 'description', 'picture\_url',), **расположение** ('neighborhood\_overview', 'neighbourhood', 'neighbourhood\_cleansed', 'neighbourhood\_group\_cleansed', 'latitude', 'longitude'), **описание недвижимости** ('property\_type', 'room\_type', 'accommodates', 'bathrooms', 'bathrooms\_text', 'bedrooms', 'beds', 'amenities', 'price'), **количество ночей для бронирования** ('minimum\_nights', 'maximum\_nights', 'minimum\_minimum\_nights', 'maximum\_minimum\_nights', 'minimum\_maximum\_nights', 'maximum\_maximum\_nights', 'minimum\_nights\_avg\_ntm', 'maximum\_nights\_avg\_ntm'), **доступность на определенные даты** ('calendar\_updated', 'has\_availability', 'availability\_30', 'availability\_60', 'availability\_90', 'availability\_365', 'calendar\_last\_scraped'), **отзывы и оценки** ('number\_of\_reviews', 'number\_of\_reviews\_ltm', 'number\_of\_reviews\_l30d', 'first\_review', 'last\_review', 'review\_scores\_rating', 'review\_scores\_accuracy', 'review\_scores\_cleanliness', 'review\_scores\_checkin', 'review\_scores\_communication', 'review\_scores\_location', 'review\_scores\_value', 'reviews\_per\_month'), **дополнительные данные и вычисляемые показатели** ('license', 'instant\_bookable', 'calculated\_host\_listings\_count', 'calculated\_host\_listings\_count\_entire\_homes', 'calculated\_host\_listings\_count\_private\_rooms', 'calculated\_host\_listings\_count\_shared\_rooms')

**Подходы, используемые для первоначальной оценки качества данных:**

* [считываем названия столбцов](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=YTDhyXFTfx22&line=1&uniqifier=1), в данном случае названия столбцов четко характеризуют содержимое данных и не требует поиска дополнительной расшифровки
* [оцениваем информацию](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=ZmaFL4HEhtiG&line=1&uniqifier=1) о датафрейме, в частности количество строк и тип данных
* [проверяем данные на пропуски](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=NPvYCJhRiHk8&line=2&uniqifier=1) и [визуализируем](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=eekc7JHQuZz8&line=4&uniqifier=1) для понимания общей структуры пропусков

Для дальнейшей обработки и отбора данных для анализа важно обозначить перечень вопросов, на которые в первую очередь сможет ответить это исследование:

**“Отзывы и оценки”:**

* Отзывы - действительно ли большое количество отзывов может являться показателем качества жилья?
* Какие критерии оценки оказывают наибольшее влияние на общее впечатление от проживания?

**“Расположение недвижимости”:**

* В каких районах города нам чаще всего готовы предложить жилье? Как сделать правильный выбор, чтобы расположиться близко к основным туристическим местам и меньше ударить по карману

**“Ценообразование и цена”:**

* Какие ценовые сегменты можно выделить?
* На какую среднюю стоимость аренды стоит рассчитывать?
* Какие ценообразующие критерии показывают наибольшую взаимосвязь с ценой аренды?

**“Недвижимость”:**

* Обзор предлагаемого жилья, какие варианты недвижимости нам готовы предложить и что они будут в себя включать
* Какие удобства в жилье чаще всего нам готовы предоставить, какие необычные виды удобств могут предложить?

**“Хосты”:**

* Вопрос безопасности: какие способы верификации хостов использует сервис Airbnb?
* Что делает хоста суперхостом? Стоит ли специально подбирать жилье ориентируясь на эту категорию

**“Объявления”:**

* Какие слова чаще всего используют хосты в объявлениях, чтобы привлечь наше внимание?

**“Срок аренды”:**

* На какой срок нам готовы предоставить недвижимость для сдачи? Действительно ли Airbnb является системой только для краткосрочной аренды жилья?

**Отбор данных для анализа:**

* [удаляем данные](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=3D6Fsw7PR9He&line=2&uniqifier=1), имеющие 100% пропуски
* отбираем часть актуальных данных из смысловых модулей датасета, исходя из сформулированных выше вопросов
* удаляем данные, которые не будут использованы при дальнейшем анализе
* [составляем список](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=Y0f5KKhF4kg9) данных и их интерпретацию (выбрано 34 столбца)
* смотрим основные [статистические показатели](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=7EvqA8b5PICv&line=3&uniqifier=1) данных
* так как цены представлены в долларах и типом данных ‘object’ для актуализации исследования [переводим](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=if8ND-4jYLzW&line=2&uniqifier=1) их в числовой формат и конвертируем в евро по курсу на момент публикации данных (10.02.2021)

**Очистка и подготовка данных:**

Проведем разделение данных по типу:

* **категориальные и текстовые** (‘name’, ‘host\_verifications’, ‘neighbourhood\_cleansed’, ‘property\_type’, ‘room\_type’, ‘bathrooms\_text’, ‘amenities’)

Подход к очистке и подготовке данных для этой категории:

*оценка количества пропусков* **→** ‘host\_verifications’, 'neighbourhood\_cleansed', ‘property\_type’, ‘room\_type’ - пропусков не имеют, пропуски столбцов ‘name’, ‘amenities’ - заполнены ‘unknown’, ‘bathrooms\_text’ - заполнение модой.

*подготовка* **→** ‘name’, ‘host\_verifications’, ‘property\_type’ - использование [формулы](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=KA3bgd2SRXFP&line=1&uniqifier=1) для комплексной обработки текстовых данных (включает в себя устранение лишней пунктуации, пробелов, стоп слов, приведение к нижнему регистру и лемматизацию), для ‘amenities’ - формула имеет немного другую [структуру](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=huOPaev4-v0y&line=2&uniqifier=1), это сделано для получения смысловых блоков словосочетаний, что позволило в дальнейшем получить более точную визуализацию признака. ‘bathrooms\_text’ - данные разделены на количественную (количество ванных комнат - 'bathrooms') и булевую часть (является ли ванная отдельной или общей - 'bathrooms\_private'), количество извлечено с помощью [регулярного выражения](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=OtDIGXF3GmUI&line=2&uniqifier=1), булевая часть - согласно смысловой составляющей (наличие слова ‘shared’). Для ‘property\_type’ проведено [обобщение данных](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=VrN-2DesgwYw&line=1&uniqifier=1), так как описание типа были слишком разрозненные и трудно анализируемые.

*переименованы для упрощения восприятия* **→** ‘name’ в 'description', 'neighbourhood\_cleansed' в 'neighbourhood'

*оценка структуры полученных данных* **→** для всех столбцов, с помощью функции value\_counts

* **булевые** (‘host\_is\_superhost’, ‘host\_identity\_verified’)

Подход к очистке и подготовке данных для этой категории:

*замена данных* **→** t\f на 1\0, для упрощения анализа и визуализации

*оценка количества пропусков -* для каждой категории кол-во пропусков минимально (по 11) - проведено заполнение модой

*оценка структуры полученных данных* **→** для всех столбцов, с помощью функции value\_counts и построения гистограммы распределения

* **даты** (‘host\_since’,’first\_review’, ‘last\_review’)

Подход к очистке и подготовке данных для этой категории:

*изменение структуры данных* **→** ‘host\_since’ - получим из даты [год](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=oAWcySeTJlZ4&line=2&uniqifier=1), ’first\_review’, ‘last\_review’ - [подсчитаем дни](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=oJZ0kwWdlXT9&line=1&uniqifier=1) от указанной даты до момента публикации данных и переведем данные в [категориальные](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=8W-uz9I6mKuY&line=2&uniqifier=1) с указанием временных промежутков

*оценка количества пропусков* **→** ‘host\_since’ - кол-во пропусков минимально (также 11) - проведено заполнение модой, ’first\_review’, ‘last\_review’ - пропусков много (6382 - это связано с тем, что не все объявления в принципе имеют отзывы), чтобы избежать искажений при анализе, заполним пропуски значением 'no\_reviews'

*оценка структуры полученных данных* **→** для всех столбцов, с помощью функции value\_counts и построения гистограммы распределения

* **количественные** (‘id’, ‘host\_id’, ‘host\_acceptance\_rate’, ‘host\_response\_rate’, ‘host\_listings\_count’, ‘latitude’, ‘longitude’, ‘accommodates’, ‘bedrooms’, ‘beds’, ‘minimum\_nights’, ‘maximum\_nights’, ‘availability\_365’, ‘number\_of\_reviews’, ‘price\_euro’, 'review\_scores\_rating', 'review\_scores\_accuracy', 'review\_scores\_cleanliness', 'review\_scores\_checkin', 'review\_scores\_communication', 'review\_scores\_location', 'review\_scores\_value')

Подход к очистке и подготовке данных для этой категории:

*оценка количества пропусков* **→** ‘id’, ‘host\_id’, ‘latitude’, ‘longitude’, ‘accommodates’, ‘minimum\_nights’, ‘maximum\_nights’, ‘availability\_365’, ‘number\_of\_reviews’, ‘price\_euro’ - пропусков нет, ‘host\_acceptance\_rate’, ‘host\_response\_rate’, ‘bedrooms’, ‘beds’, ‘host\_listings\_count’ - есть пропуски в данных, а также присутствуют данные равные ‘0’, что противоречит логике данных для ‘bedrooms’, ‘beds’ и ‘host\_listings\_count’, поэтому они были также приравнены к ‘nan’. Столбцы оценки отзывов - 'review\_scores\_rating', 'review\_scores\_accuracy', 'review\_scores\_cleanliness', 'review\_scores\_checkin', 'review\_scores\_communication', 'review\_scores\_location', 'review\_scores\_value' - множественные пропуски в данных, как выше упоминалось, не все объявления имеют отзывы. .Для заполнения пропусков применены различные подходы: ‘host\_acceptance\_rate’, ‘host\_response\_rate’ и все “review\_scores...” - пропуски заполнены ‘0’, ‘host\_listings\_count’ - пропуски заполнены [вычисляемым показателем](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=knVwaM-CFUFu&line=1&uniqifier=1), ‘bedrooms’, ‘beds’ - заполнены медианой показателей, существенных искажений статистики после заполнения не выявлено.

*изменение структуры данных* **→** ‘minimum\_nights’, ‘maximum\_nights’, ‘availability\_365’ - данные переведены в категориальные с разделением на временные промежутки

*оценка выбросов в данных* **→** для столбцов 'accommodates', 'beds', 'bedrooms', а также полученного ранее столбца 'bathrooms', произведена оценка наличия выбросов, далее использован ряд логических [проверок](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=itil2FEuAhtk&line=1&uniqifier=1) для нахождения искаженных данных.Для столбцов ‘latitude’, ‘longitude’ все значения соответствуют географическим координатам Рима и в корректировке не нуждаются. Столбец ‘price\_euro’ показал существенные выбросы в данных и для него отдельно также проведен ряд [проверок](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=jHQXKVzQQWHv&line=2&uniqifier=1) на логику данных

*нормализация данных* **→** для столбцов‘host\_acceptance\_rate’, ‘host\_response\_rate’, 'review\_scores\_rating', 'review\_scores\_accuracy', 'review\_scores\_cleanliness', 'review\_scores\_checkin', 'review\_scores\_communication', 'review\_scores\_location', 'review\_scores\_value - изменен масштаб данных с помощью preprocessing.[MinMaxScaler()](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html) (масштабируем данные в диапазон от 0 до 1)

*оценка структуры полученных данных* **→** для всех столбцов, с помощью функции value\_counts и построения гистограммы распределения.

После очистки, качество данных [повышено](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=COwYHIwHItWm&line=2&uniqifier=1), устранены пропуски, минимизированы выбросы отдельных показателей и искажения данных, без существенной потери количества строк, [статистические показатели](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=Q-6itm9wqA20&line=1&uniqifier=1) не получили существенных искажений.

1. Файл **neighbourhoods.geojson -** данные для построения графических карт, содержат информацию о расположении муниципальных районов Рима

Количество строк: **15**

Количество столбцов: **3 (‘**neighbourhood’, ‘neighbourhood\_group’, ‘geometry’)

[**Чтение данных**](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=1VybZsYq4XxM&line=2&uniqifier=1)**:**

**Обработка данных:** ‘neighbourhood\_group’ содержит 100% пропусков данных и был удален

**Выявленные взаимосвязи между показателями** ([матрица](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=oyWyIegV6vQ3&line=3&uniqifier=1))

* Отрицательная направленность корреляционной связи - 'host\_days\_active' и 'id', 'host\_days\_active' и 'host\_id', 'host\_days\_active' и 'host\_since', 'id' и 'number\_of\_reviews'. Связь между всеми этими критериями основана на давности активации - пример возможной интерпретации: чем больше прошло дней с момента активации, тем меньшим числом будет обозначен id пользователя, но можно обозначить и наоборот: чем большим числом будет обозначен id пользователя, тем меньше дней прошло с его активации в сервисе.
* Положительная направленность корреляционной связи - можно разделить на подблоки:
  + связи также построенные на времени активации: 'id' - 'host\_id', 'id' - 'host\_since', 'host\_since' - 'host\_id'
  + подблок оценок недвижимости на основании отзывов: 'review\_scores\_rating', review\_scores\_accuracy', 'review\_scores\_cleanliness', 'review\_scores\_checkin', 'review\_scores\_communication', 'review\_scores\_location' - имеют достаточно высокую корреляцию между собой (мультиколлинеарны), эти критерии в дальнейшем рассмотрим подробнее, а также их взаимосвязь с 'number\_of\_reviews'
  + взаимосвязи ценообразования: 'price\_euro' - 'accommodates', 'price\_euro' -'bedrooms', 'price\_euro' - 'beds', 'price\_euro' - 'bathrooms', 'price\_euro' - 'bathrooms\_private'
  + взаимосвязи с категорией superhost: здесь все показатели оценок на основании отзывов ('review\_scores\_rating', review\_scores\_accuracy', 'review\_scores\_cleanliness', 'review\_scores\_checkin', 'review\_scores\_communication', 'review\_scores\_location'), а также 'superhost' -'number\_of\_reviews', 'superhost' - 'host\_response\_rate', 'superhost' - 'host\_acceptance\_rate', также рассмотрим подробнее
  + взаимосвязи, определяющие облик предлагаемой недвижимости: 'accommodates'- 'bathrooms', 'accommodates'- 'bedrooms', 'accommodates'- 'beds', 'bathrooms' - 'bedrooms', 'bathrooms' - 'beds', 'bathrooms' - 'beds', 'bedrooms' - 'beds'

**Исследование полученных данных:**

**Отзывы и оценки**

23% объявлений не имеют отзывов - то есть эту недвижимость или ещё никто ни разу не снимал, или никто из снимающих жилье не стал оставлять отзыв.

Не каждая сдача недвижимости на Airbnb равна полученному отзыву, но сделаем допущение, что отзывы после сдачи жилья все-таки чаще оставляют - этому способствует в том числе система автоматических напоминаний от сервиса, что в течение 14 дней после отъезда есть возможность оставить отзыв, а также двусторонняя система отзывов (хост также оценивает гостя и отзыв можно увидеть только после того, как оставишь свой или прошло 14 дней) - [как устроена система отзывов на жильё](https://www.airbnb.ru/help/article/13/%D0%BA%D0%B0%D0%BA-%D1%83%D1%81%D1%82%D1%80%D0%BE%D0%B5%D0%BD%D0%B0-%D1%81%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B0-%D0%BE%D1%82%D0%B7%D1%8B%D0%B2%D0%BE%D0%B2-%D0%BD%D0%B0-%D0%B6%D0%B8%D0%BB%D1%8C%D0%B5).

Рассмотрим подробнее, сколько времени прошло с момента первого и с момента последнего отзыва для опубликованных объявлений:

Для этого построены [гистограммы](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=NBl1tl4XmPzq&line=6&uniqifier=1) распределения на основе признаков 'time\_since\_first\_review' и 'time\_since\_last\_review', получены результаты:

* от 0 до 6 месяцев :
  + первый отзыв - у 2% объявлений
  + последний отзыв - 22% объявлений
* 6-12 месяцев -
  + первый отзыв - 2%
  + последний отзыв - 17%
* 1-2 года -
  + первый отзыв - 18%
  + последний отзыв - 27%
* 2-3 года -
  + первый отзыв - 25%
  + последний отзыв - 8%
* более 4 лет -
  + первый отзыв - 30%
  + последний отзыв - 3%

**Сформулирован вывод:** на сервисе Airbnb представлено достаточно большое количество недвижимости в Риме, которая сдается в аренду уже давно (точно более 4 лет - 30% объявлений и более 2-3 лет еще 25%), за последний же год есть всего 4% объявлений, получивших свой первый отзыв (по 2% за каждые полгода) - сказывается в том числе падение спроса, связанное с существенным сокращением туристического потока из-за карантинных мер. Из всей недвижимости более активно сдается 22% недвижимости (есть отзывы за последние полгода) и есть 3% объявлений, отзывов по которым нет более 4 лет.

Следующий рассмотренный вопрос - действительно ли большое количество отзывов может являться показателем качества жилья?

Для ответа сформировано два датафрейма с выборкой из основного столбца ‘number\_of\_reviews’ по 3000 наблюдений: первый с количеством отзывов ниже [среднего](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=8QVqkJyScxk2&line=1&uniqifier=1) ([less\_reviewed](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=nX67uATshNdv&line=1&uniqifier=1)), второй - выше среднего ([more\_reviewed](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=iZWeCP0Kc2YQ&line=2&uniqifier=1)). Нижняя граница показателей первого датасета исключает количество отзывов менее 10, так как небольшое количество отзывов может быть субъективным.

Для базового сравнения признаков построен [боксплот](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=I5Y17QVlq1vd&line=3&uniqifier=1) - наблюдаем, что медиана второй выборки лежит примерно на границе 25 квартиля первой выборки. Далее оцениваем распределения выборок, тестируем на [нормальность](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=u_DRteV5nijF&line=5&uniqifier=1) с помощью [stats.normaltest()](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html), а также [визуально](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=Xa-f2wNsw0S6&line=1&uniqifier=1). Тест отвергает нормальность распределений, но из-за большого количества наблюдений распределения визуально близки к нормальным, имеют колоколообразную форму, но слева прослеживаются удлиненные “хвосты”.

Сформулируем гипотезы:

* H0 - Общая оценка рейтинга недвижимости одинакова для объявлений с количеством отзывов ниже среднего и с количеством отзывов выше среднего
* H1 - Общая оценка рейтинга недвижимости различна для объявлений с количеством отзывов ниже среднего и с количеством отзывов выше среднего

Для проверки гипотез проведена оценка выборок с помощью следующих критериев:

* Т-критерий с использованием поправки Уэлча (поправка используется, так как дисперсии выборок [различны](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=CI4pNsMTv6HW&line=2&uniqifier=1))
* непараметрический U-критерий Манна - Уитни

Принимаем α равное 0.05

Результат для Т-критерия pvalue=0.000000017, для Манна-Уитни pvalue=0.002913581182052803. В обоих случаях результат позволяет отвергнуть H0, то есть общая оценка рейтинга недвижимости различна для объявлений с количеством отзывов ниже среднего и с количеством отзывов выше среднего.

**Сформулирован вывод:** при поиске жилья в том числе можно ориентироваться на большое количество отзывов у объявления

Но как же формируется сама общая пользовательская оценка предлагаемого жилья? Какие из других имеющихся критериев оценки имеют большую с ней взаимосвязь?

Для ответа сформирован отдельный [датафрейм](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=Z7RSwpbT4XZL&line=4&uniqifier=1), содержащий критерии оценки недвижимости и общую оценку. Далее в него отобраны показатели отзывов выше нуля и построена [матрица корреляции](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=xrYK35FTe7ve&line=1&uniqifier=1) по отношению к показателю общей оценки. Выстроены [regplot](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=P8_BuYhYy5_B&line=2&uniqifier=1) для визуальной оценки распределения критериев оценки и общей оценки недвижимости. Исходя из полученных данных сделан **вывод:** среди всех имеющихся критериев оценки недвижимости пользователями, можно наблюдать наибольшую взаимосвязь между общей оценкой и оценкой показателя "цена-качество". Далее по убывающей следует "точность описания", "чистота", "коммуникация с хостом", "процедура заселения" и самую наименьшую взаимосвязь показывает "расположение недвижимости".

**Расположение недвижимости**

Предлагаемая недвижимость распределена по 15 муниципальным округам Рима неравномерно - 54% всего жилья расположено в историческом центре.Но в каких районах города лучше выбрать жилье, работает ли в Риме правило "дальше от центра - зато дешевле"?

Для ответа на вопрос для начала был произведен подсчет медианы цены недвижимости и количества объявлений для каждого муниципального района ([ссылка](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=xpDFVfRIaIKp&line=1&uniqifier=1)) и на основании полученных данных построен [барплот](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=mgcbH4Y_RWWJ&line=4&uniqifier=1). Медиана цены аренды в историческом центре составила 71 евро и эта цена объективно выше, чем в других районах города (например, на втором месте район XIII Aurelia с медианой 55 евро). Казалось бы, можно сделать вывод, что жить в центре дороже и лучше выбрать более отдаленные районы. Но действительно ли это так? Если посмотреть какое количество объявлений приходится на каждый район, то в историческом центре недвижимости сдается в 8,6 раз больше, чем в Aurelia. Цена аренды в центре имеет существенный разброс (от 7 до 3735 евро за ночь аренды) и даже медиана не дала адекватной картины ценообразования в районе. Поэтому важно было построить карту города с распределением предложения относительно цены, а также диаграмму рассеяния цен относительно каждого муниципального района и количества гостей. На карте были отмечены некоторые из главных достопримечательностей города, с которыми чаще всего приезжают ознакомиться туристы, а также два основных аэропорта. Для наглядности визуализация разделена по двум ценовым сегментам [до 200 евро](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=2EBoY-XvbceZ&line=2&uniqifier=1) и [более 200](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=uyx5t3RPjV0r&line=3&uniqifier=1). Для диаграммы рассеяния в сегменте до 200 евро для наглядности также применялась выборка из 3000 наблюдений.

По визуальным данным сделаны следующие выводы:

* Большая часть всех предложений сконцентрирована в историческом центре города (I Centro Storico), также существенное количество предложений расположено в прилегающему к нему району II Parioli/Nomentano (третий по кол-ву объявлений). Основная часть часто посещаемых достопримечательностей, также сконцентрирована в историческом центре или расположена максимально к нему близко.
* Предложения в других районах имеют основную тенденцию "притяжения" и увеличения концентрации в направлении к центральному району (I Centro Storico)
* Район VII San Giovanni/Cinecittà (второй по кол-ву объявлений) помимо центрально-направленной тенденции имеет концентрацию предложений вблизи аэропорта Чампино, поэтому предложения распределены достаточно равномерно по территории района
* Районы X Ostia/Acilia и XI Arvalia/Portuense обладают в свою очередь дополнительным "центром притяжения" - международным аэропортом Фьюмичино. Район X Ostia/Acilia также имеет сконцентрированные предложения у приморских территорий.
* Ценовая политика предложений различна, в любом районе города можно выбрать на любой кошелек. Из-за высокого разброса, реально цена слабо зависит от расположения, то есть даже в самом центре (где медиана цены составила 71 евро) можно найти весьма недорогое жилье (в том числе, в котором можно разместиться компанией в несколько человек) и таких предложений достаточно много, а ценовая политика строится, по большей части, на основании других факторов.
* В сегменте жилья дороже 200 евро за ночь также наблюдается максимальная концентрация предложений в историческом центре города, в остальных районах можно отметить только единичные объявления. Также в этом сегменте довольно часто встречаются предложения для размещения больших компаний гостей.

Так какой район города выбрать? Если цель посещения города именно туризм и посещение достопримечательностей, то нет смысла искать жильё далеко от муниципального района I Centro Storico и тратить драгоценное время путешествия на перемещения по городу из отдаленных районов в центр и обратно (а также деньги на билеты или такси), так как существенной разницы в цене между центром и отдаленными районами города не прослеживается. Если же самый центр города покажется слишком шумным и полным суеты, то есть множество предложений в II Parioli/Nomentano, компактному району близко расположенному к центру города, но намного более размеренному и спокойному.

Но, если цена слабо зависит от расположения, то какие же факторы ее формируют?

**Ценообразование и цена**

Есть часть предложений с экстремально высокой ценой за ночь - цена выглядит умышленно завышенной, как объяснение, есть несколько предположений: или эти объявления временно 'заморожены', или хосты проверяют забронирует ли кто-то предложенную недвижимость за такую цену, или цена указана не за ночь, а за минимальный период аренды.

Условно, все ценовые предложения можно декомпозировать на [три ценовых сегмента](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=zXWs-T7nhQ85&line=8&uniqifier=1):

1. Сегмент до 200 евро - 26596 наблюдений (94.8%)
2. от 201 до 1000 евро - 1395 наблюдений (4.9%)
3. от 1001 евро - 51 наблюдение (0,3%)

Далее рассмотрена связь цены и ценообразующих критериев - построена матрица [корреляции](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=6au0r4Fg4LJ_&line=1&uniqifier=1), [regplot](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=BDUrB9yCLLQS&line=4&uniqifier=1) и [barplot](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=No7WFAyP5NSz&line=16&uniqifier=1) для каждого критерия. На основе полученных данных, сформулирован вывод, что наибольшая взаимосвязь прослеживается между ценой и количеством размещенных гостей, далее по убывающей взаимосвязь с количеством спален, кроватей и ванных комнат. Более четкая линейность характерна для распределения цены и количества гостей - чем большее количество гостей готовы принять, тем выше цена. Для пары цена - количество спальных комнат - цена линейно растет до кол-ва в 8 комнат, далее резко растет на значении в 9 комнат и после распределяется хаотично. Схожая картина в количестве спальных мест (также после значения в 9 мест) и в количестве ванных комнат (после 4,5 комнат). Возможно, разброс связан с тем, что большое количество спальных, ванных комнат и кроватей могут быть как в дорогом доме, где их может быть много, но будет проживать несколько гостей, так и недорогой хостел, который можно снять большой компанией. Для анализа влияния на цену критерия “личная ванная или общая” построена [3D](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=LAMg63uK7x9A&line=4&uniqifier=1) визуализация с дополнительным разделением на кол-во гостей, из которой следует, что общая ванная предлагается при небольшом кол-ве гостей (ориентировочно 2) и цена на такие предложения несколько ниже, чем когда ванная комната личная.

Имеющиеся данные явно не могут полностью объяснить формирование цены аренды и существует ряд ценообразующих критериев не представленных в этом датасете.

Далее посмотрим, что же чаще всего включает в себя предлагаемая для аренды недвижимость?

* Большая часть недвижимости вмещает от 1 до 5 [гостей](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=QAb7uNbAsHzf&line=2&uniqifier=1), но есть предложения и для компаний до 15 человек
* [По типу недвижимости](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=-p2fecdfgbTO&line=2&uniqifier=1) 89% предложений занято квартирами, но при желании можно снять и домик на дереве, иглу или комнату в замке
* В 63% предложений можно получить комнату или дом целиком
* 68% включают в себя только одну спальню и одну ванную комнату (в 92% объявлений она предполагается личная)
* Чаще всего предлагаемое количество спальных мест от 1 до 3
* [Предложения](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=GYsas_W6RW9q&line=1&uniqifier=1) апартаментов более характерны для центра и приморских территорий, а дома в более отдаленных районах города.

**Отсюда вывод:** собирательный образ недвижимости, которую предложат гостю в Риме выглядит так - это квартира целиком в историческом центре города с одной спальной комнатой и одной кроватью, с одной отдельной ванной комнатой для двух гостей.

Теперь о хостах:

* Хостами выступают не только частные лица, также есть представители риэлторских агентств, гостевых домов и отелей
* Меньше четверти хостов являются суперхостами
* 78% хостов прошла верификацию
* Большая часть хостов зарегистрирована более 5 лет назад
* Половина хостов принимает всех обратившихся за арендой
* 45% хостов всегда отвечают на поступающие обращения
* Только 33% сдают только один объект недвижимости

Так что же на счет безопасности аренды?

Как мы могли заметить, большая часть хостов прошла процедуру верификации, что безусловно делает процедуру сдачи недвижимости более безопасной. Какие же основные методы верификации хостов использует сервис Airbnb?

Для анализа показателя был использован [счетчик слов](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=lWNGfihV21av&line=4&uniqifier=1), а далее построен [барплот](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=lUAxuGYiS5PF&line=2&uniqifier=1) с полученным результатом. Процедура верификации чаще всего состоит из нескольких компонентов (например, подтвержден номер телефона, электронная почта и удостоверение личности и сделано селфи).

Далее рассмотрено подробнее, что стоит за категорией "суперхост". Для поиска ответа на этот вопрос сформирован отдельный датафрейм для поиска взаимосвязей между признаками, которые характеризуют [хоста](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=sMlgHaAsJtIQ&line=3&uniqifier=1) и построена [корреляционную матрицу](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=EkK6IuccKsDf&line=2&uniqifier=1).

Исходя из данных, размещенных на сайте [airbnb](https://www.airbnb.ru/d/superhost), на получение значка "Суперхост" влияют такие показатели как:

* общий рейтинг
* частота приема гостей
* частота отмены бронирований
* частота ответов на сообщения

Для полноценной оценки частоты приема гостей необходимы данные за более длительный период (1 год), но остальные указанные выше показатели действительно имеют некоторую положительную корреляцию с исследуемым признаком.

Далее построено несколько различных визуализаций, которые помогают ответить на несколько вопросов и понять в чем же на самом деле суперхост лучше:

* [Проходил ли хост верификацию аккаунта](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=Jo_LK_CebmY9&line=4&uniqifier=1)? Из прошедших верификацию хостов и суперхостов примерно равное количество, из тех, кто не прошёл - хостов немного больше
* [В каком году зарегистрировался хост](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=EflgZTjMMBwl&line=5&uniqifier=1)? Особых различий по этому показателю между хостом и суперхостом нет
* [На какую часть сообщений хост отвечает](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=Zbxm-98Le6d_&line=7&uniqifier=1)? Суперхосты чаще отвечают на все сообщения, так как этот показатель является одним из требований для получения значка "superhost"
* [Какую долю заинтересовавшихся жильем хост согласился принять](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=zSZQdFGQmYZp&line=3&uniqifier=1)? Суперхосты также реже отказывают гостям в приеме, так как частый отказ без уважительной причины может способствовать потере значка "superhost"
* [Как распределяется цена аренды в сегменте до 200 евро](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=-H1xtxvdUsVO&line=4&uniqifier=1)? Значительных различий в цене также не обнаружено, но у суперхостов немного реже встречаются предложения бюджетного жилья до 50 евро за ночь и чуть больше предложений от 100 до 150 евро.
* [Как распределяется общий рейтинг?](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=n1C_MYVIPFiK&line=3&uniqifier=1) Суперхосты значительно чаще получают максимальные оценки в отзывах гостей (высокая оценка также является одним из критериев получения значка "superhost")

**Вывод:** суперхост будет просто более отзывчивым хозяином с более стабильным сервисом (так как ему необходимо "держать марку"), с меньшей вероятностью неприятных сюрпризов по основным показателям его оценки. Если жилье необходимо на короткий срок и только на переночевать, то особой необходимости искать суперхоста нет.

Каждый хост, размещая свое объявление, скорее всего стремится, чтобы его заметили и выбрали именно его предложение о сдаче жилья. Далее было рассмотрено, какие слова чаще всего используют хосты в объявлениях, чтобы привлечь внимание гостей. Для решения этой задачи был использован [счетчик слов](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=Pocuyshr0gjB&line=3&uniqifier=1) и визуализация [wordcloud](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=5GdB-67601-B&line=3&uniqifier=1). Текст объявлений оказался в основном далёк от художественного стиля и больше построен на описаниях. Хосты предпочитают, в первую очередь, подчеркнуть выгодность расположения и кратко дать описание по виду недвижимости и что она содержит (количество комнат, ванных, кроватей и тд.).

С помощью тех же инструментов ([счетчик слов](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=4WDQJ7PGqN7M&line=3&uniqifier=1) и [wordcloud](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=Qi4zaKL0idAd&line=4&uniqifier=1)) было рассмотрено, какими удобствами оборудовано предлагаемое для аренды жилье.

Чаще всего для удобства гостей есть вешалки, утюг, вай-фай, фен, обогреватель, предметы первой необходимости и набор первой помощи, телевизор, кондиционер, кухня, стиральная машина, отдельное пространство для работы, горячая вода, гидромассажная ванна, кофемашина, лифт и даже парковочное место. Из более редкого и необычного предлагают, например, детскую ванночку и кроватку, фортепиано, пуровер для кофе, рисоварку и хлебопечку, сейф, спортзал и бассейн, возможность курения в помещении и пребывания с домашними животными.

В списке удобств можно также заметить пункт "возможность долговременного проживания". Чаще всего Airbnb воспринимают как сервис именно для краткосрочного проживания, но так ли это?

Посмотрим на [распределение максимального и минимального количество дней пребывания](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=QXpeSRskmyUx&line=3&uniqifier=1).

* 1-7 дней - примерно в 97% размещенных объявлений рассматривается именно этот период, как минимально возможное время пребывания, и только 3% объявлений, в которых максимальное время пребывания также ограничено этим периодом
* 8-14 дней - около 1% указали этот период как минимальное время проживания и 5% как максимальное
* 15-21 день - практически не указано, как минимальное время, как максимальное около 2%
* 22-31 день - есть около 2% объявлений, в которых минимальное время аренды 3-4 недели, этим же периодом ограничивают максимальное время пребывания около 22% владельцев недвижимости (здесь скорее всего имеет место юридический аспект - согласно местному законодательству, аренда недвижимости свыше 30 дней должна сопровождаться заключением договора, зарегистрированного в налоговых органах, и как результат, уплатой налога на доход с аренды [https://www.agenziaentrate.gov.it](https://www.agenziaentrate.gov.it/))

В результате, минимальное время аренды, за исключением единичных объявлений (часть из которых похожи на временно неиспользуемые), ограничено периодом от 1 до 31 дней. Оставшиеся 68% ограничений на максимальное время аренды распределились так:

* 32-60 дней - для 3% объявлений
* 61-180 дней - для 7% объявлений
* 181-365 дней - для 5% объявлений
* 366+ дней - и около 53% объявлений предлагают максимальный срок от года и выше

**Вывод:** Airbnb - это не только про краткосрочную аренду, если есть время и возможность, с помощью сервиса можно найти жилье и на достаточно длительный период.

Но так ли легко забронировать недвижимость? Много ли [свободных дней для аренды могут предоставить хосты на ближайший год](https://colab.research.google.com/drive/1-8xeM4WwKi6Yfft3ueQoNVEZn9Dq2fo2#scrollTo=mqFYEAmNnJnd&line=2&uniqifier=1)?

* 0 свободных дней - 13% недвижимости не имеют свободных дней бронирования или временно прекратили её сдачу, закрыв календарь
* 1-14 свободных дней - 2% недвижимости
* 15-30 свободных дня - 1% недвижимости
* 31-60 свободных дней - 2% недвижимости
* 61-120 дней - 9% недвижимости
* 121-180 дней - 11% недвижимости
* 181-240 дней - 4% недвижимости
* 241-300 дней - 8% недвижимости
* 301-364 дней - 34% недвижимости
* 365 дней - 16% недвижимости, то есть недвижимость готова к сдаче, но не имеет ни одного бронирования на ближайший год

**Вывод:** Из-за вариативности интерпретации, возникает сложность с адекватной оценкой этого критерия - в случае, малого доступного количества свободных дней для аренды, есть как минимум два варианта: или хост открывает бронирование постепенно на небольшое количество дней, или бронь действительно практически полная. Но в любом случае, есть 16% недвижимости полностью доступной для бронирования в течение года

**Выводы по исследованию:**

* Читайте отзывы и смотрите на их количество - так проще подобрать лучшее предложение .
* Изучите критерии оценки объявления в отзывах, посмотрите, из каких критериев получилась общая оценка и ориентируйтесь на то, что для вас более важно.
* Стоит выбрать жилье в центре или близких к нему районах, если цель посещения города туризм и знакомство с достопримечательностями. Проще всего в центре будет найти отдельную квартиру с одной спальной и ванной комнатой для пары гостей. Большинство хостов предоставит базовый набор удобств для проживания, а кто-то возможно даже удивит, предложив что-то необычное.
* Большая часть предложений лежит в сегменте до 200 евро за ночь, но даже в центре есть объявления на любой кошелек (от 7€ до 3735€) .Из известных по данным факторов, цена будет строится исходя из количества гостей, спальных комнат и кроватей, а также количества ванных и является ли она личной или общей.
* Хост - это может быть не только частное лицо, если это необходимо, можно найти и представителей отелей и гостевых домов.
* Не стоит искать недвижимость только среди объявлений с категорией суперхост, особенно, если вам нужно просто переночевать - разница между хостом и суперхостом невелика. Но, если вы гостите у суперхоста, то скорее всего получите более стабильный сервис и радушный прием.
* Стоит обратить внимание на то, чтобы аккаунт хоста был верифицирован сервисом airbnb, тем более, таких хостов большинство - это вопрос спокойствия и безопасности проживания.
* Если же это необходимо, то можно найти и жилье для долговременной аренды, сервис это также позволяет.

**Возможные пути развития и улучшения данной работы:**

* использовать данные за более длительный период, для изучения динамики исследуемых показателей.
* более детальная проработка отсеивания неправдоподобных цен, возможно связать с анализом информации в тексте объявления
* анализ текста объявлений для выявления потенциально “проблемного” хоста
* добавить предсказательную модель для цены аренды недвижимости

|  |  |
| --- | --- |
|  |  |