Factors behind local rental properties pricing in New York City

Santiago Rodrigues Manica

27th, June 2020

Introduction

Background

Imagine you have the luck of owning real estate in New York City, which is available for renting. Since you live in the 8th most visited city in 2019 (https://edition.cnn.com/travel/article/most-visited-cities-euromonitor-2019/index.html), you know there is a high demand for accommodation. Since traditional hotels are especially expensive in the USA more are more tourist are looking for local and affordable accommodation. There is no surprise until 2019 companies like AirBNB kept growing (https://news.airbnb.com/airbnb-2019-business-update/). And there is no secret, for good and for bad, that renting to short-term tourism can bring higher revenues versus having income from a long-term tenant.

Problem

In this case you may be a clueless owner trying to guess what factors may influence your expected pricing. Since it may be, for example, due to the neighbourhood where the rental property is located, or may be affected by convenience factors, such as proximity to restaurants and entertainment.

Interest

In case you are a client curious about which factors may affect the price you could charge your guests, we will explore these data.

Methods

City of interest for this project: **New York City**, NY, USA

Datasets of interest

New York's AirBNBs (csv): https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data?select=AB_NYC_2019.csv a freely and publicly available dataset on Kaggle, which I stored as a csv file in my GitHub account. https://github.com/RM-Santiago/Coursera_Capstone/blob/master/AB_NYC_2019.csv

Foursaquare application programming interface (API): https://developer.foursquare.com/docs/places-api/ This API will be used to obtain the venues around the rental properties and will be useful for both exploratory data analysis (EDA) and inferential analysis.

Research questions and statistical methods:

After proper data management and EDA (including mapping and clustering), the project will try to answer to the following questions using libraries that allow data frame analysis and statistical testing (eg; Pandas):

1. Is the average price different between neighborhoods?

Using the **New York's AirBNBs dataframe** the mean and standard deviation of the price will be described across different neighborhoods and differences will be tested.

2. Is there an association between the average price and the neighborhood?

A simple linear regression will test the association between price (outcome/dependent variable) and the neighbourhood (categorical independent variable).

3. Is the average price different between the type of rental properties?

Using the **New York's AirBNBs dataframe** the mean and standard deviation of the price will be described across different types of accommodation (eg; whole apartment vs room only) and differences will be tested.

4. Is there an association between the average price and the type of rental properties?

A simple linear regression will test the association between price (outcome/dependent variable) and the type of rental property (categorical independent variable).

5. Considering the simultaneous effect of all candidate factors which may be associated with the price of a rental property?

Using the **New York's AirBNBs,** a multiple linear regression will test the association between price (outcome/dependent variable) and a set of dependent variables (neighborhood and type of rental property).

In order to test differences of price in USD (continuous variable) between groups the following statistical tests may be used;

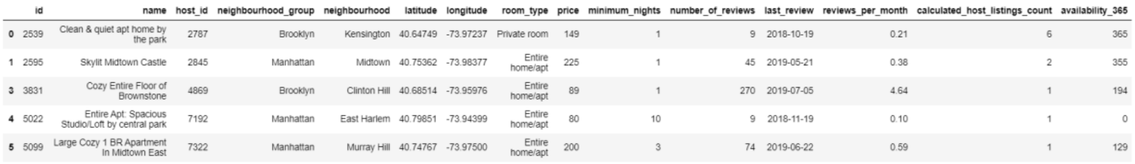
- Student's t-test; comparing two independent groups if prices have a normal distribution;
- Wilcoxon-Mann Whitney test; comparing two independent groups if prices have a non-normal distribution;
- Analysis of covariance (ANOVA); when comparing prices across more than 2 independent groups if prices have a normal distribution;
- Kruskal Wallis; when comparing prices across more than 2 independent groups if prices have a non-normal distribution.

Data analysis

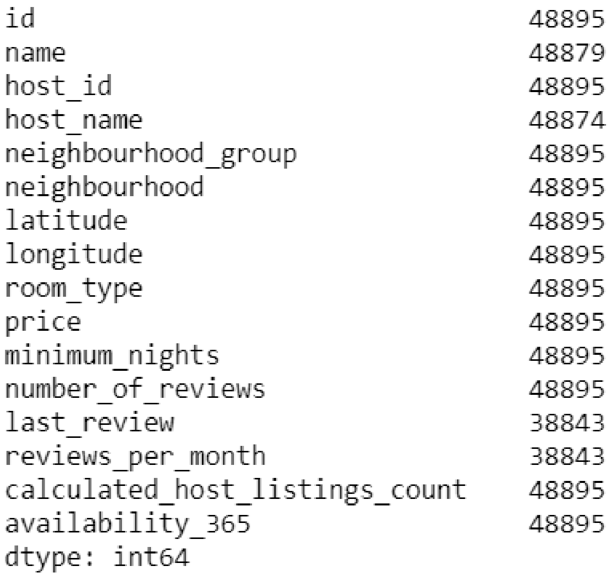
Now, moving to the data analysis that can be found on the Jupyter Notebook

https://github.com/RM-Santiago/Coursera_Capstone/blob/master/SANTIAGO_IBM_Battle%20of%20the%20Cities_Capstone.ipynb

After importing all relevant packages and obtaining the dataset from https://github.com/RM-Santiago/Coursera_Capstone/blob/master/AB_NYC_2019.csv . We can see there are 48,895 rental properties in New York.



	id	name	host_id	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
0	2539	Clean & quiet apt home by the park	2787	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
1	2595	Skylit Midtown Castle	2845	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355
3	3831	Cozy Entire Floor of Brownstone	4869	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	194
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	1	0
5	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Manhattan	Murray Hill	40.74767	-73.97500	Entire home/apt	200	3	74	2019-06-22	0.59	1	129

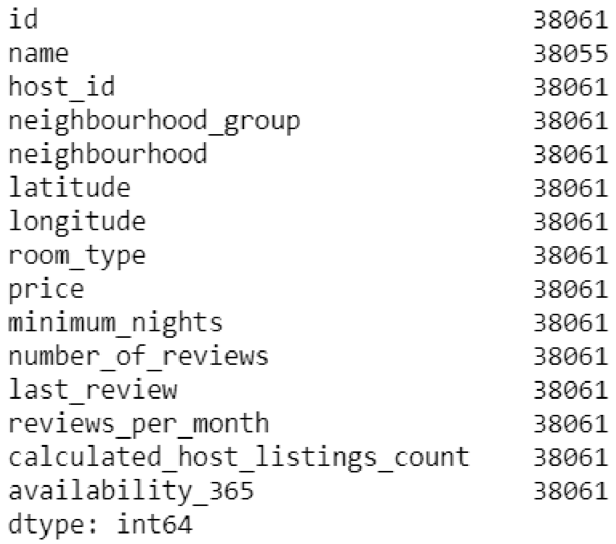


```
id                                48895
name                              48879
host_id                           48895
host_name                         48874
neighbourhood_group               48895
neighbourhood                     48895
latitude                          48895
longitude                         48895
room_type                         48895
price                             48895
minimum_nights                    48895
number_of_reviews                 48895
last_review                       38843
reviews_per_month                 38843
calculated_host_listings_count    48895
availability_365                  48895
dtype: int64
```

We see here there are only 38,843 properties with a “last review” but 48,895 properties with a given “number of reviews”. This happens because some properties have 0 reviews.

Since the name of the host is not relevant, it was dropped. Since we only want properties with a review, all these with a number of reviews equal to zero where dropped.

We have here as relevant variables; the property name and id, its price per night, the number of reviews, the neighbourhood where its located and the type of property.

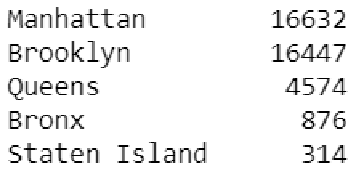


```
id                                38061
name                              38055
host_id                           38061
neighbourhood_group               38061
neighbourhood                     38061
latitude                          38061
longitude                         38061
room_type                         38061
price                             38061
minimum_nights                    38061
number_of_reviews                 38061
last_review                       38061
reviews_per_month                 38061
calculated_host_listings_count    38061
availability_365                  38061
dtype: int64
```

Now all number look the same. In fact there are only 38,055 rental properties with a name. However, since they all have an ID we won’t worry about it.

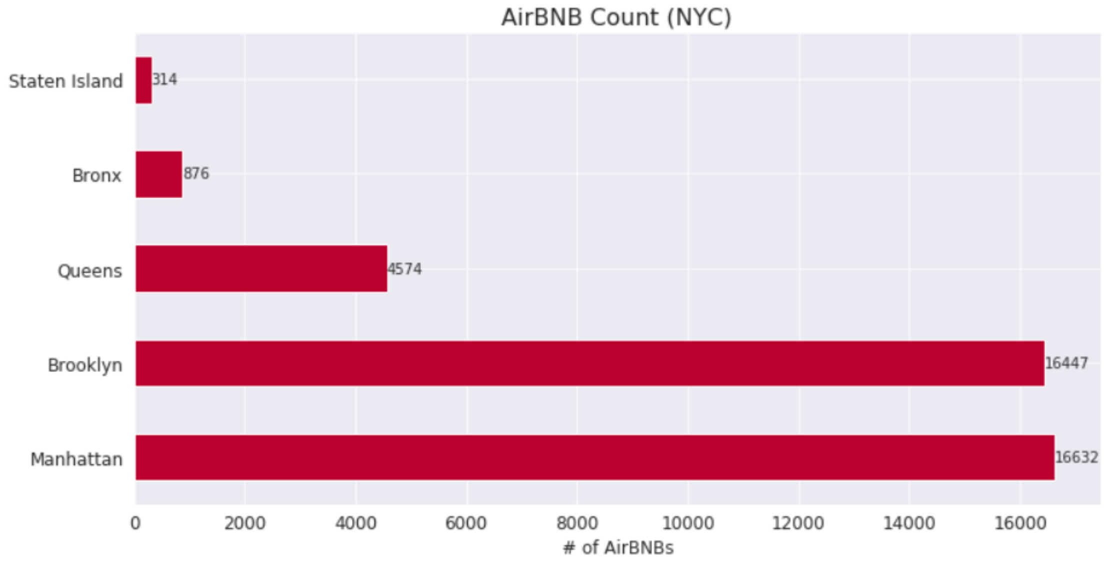
Now, after integrating the API. Let’s explore some data.

How many properties are there in any major Neighborhood?



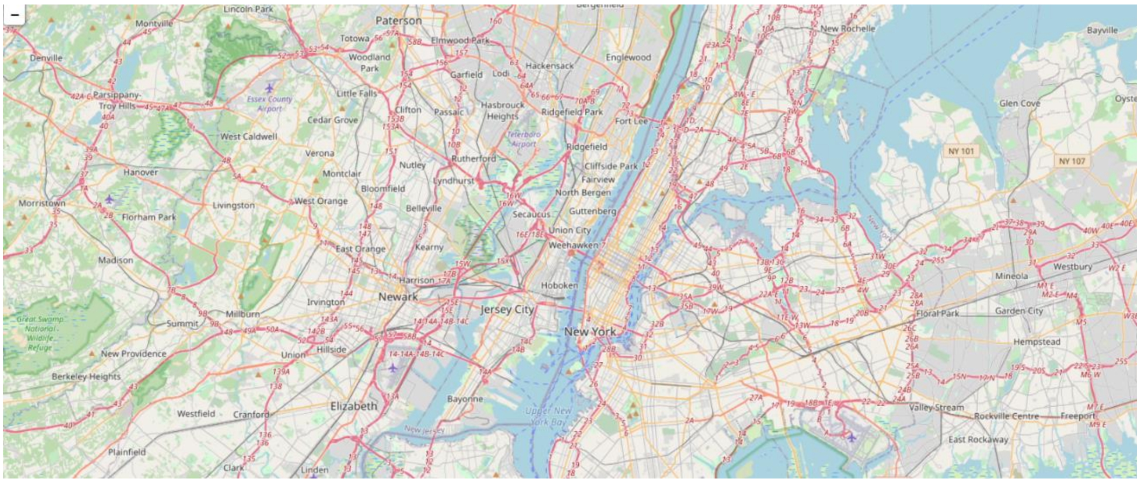
```
Manhattan        16632
Brooklyn         16447
Queens            4574
Bronx              876
Staten Island      314
```

What about a graphical version?

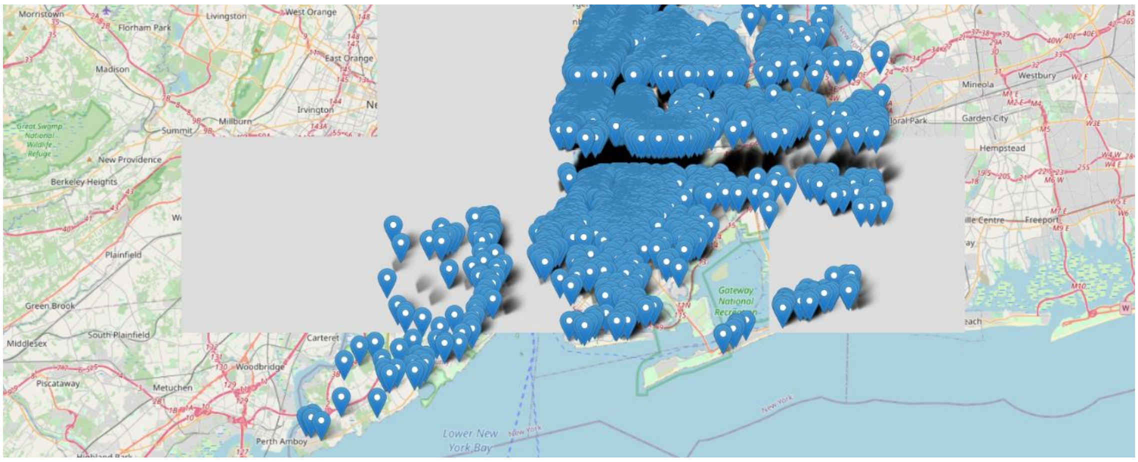


Let’s take a look at the charts

First, New York, New York…

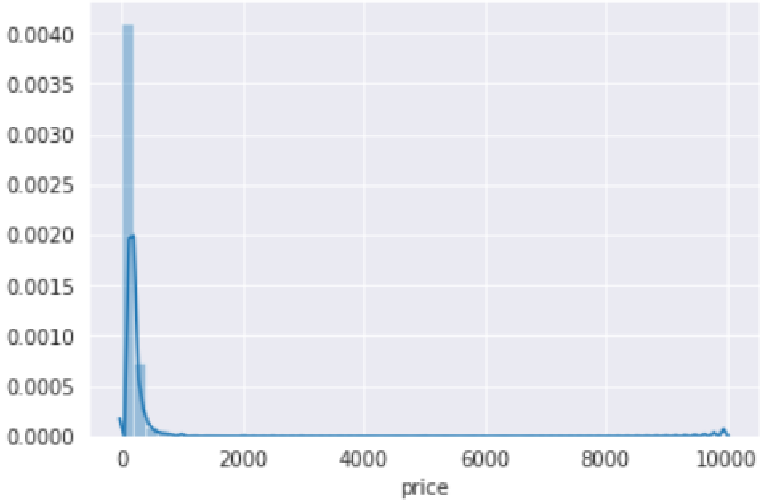


And now with the Rental Properties



(sorry for the grays areas – too much info for my computer’s memory)

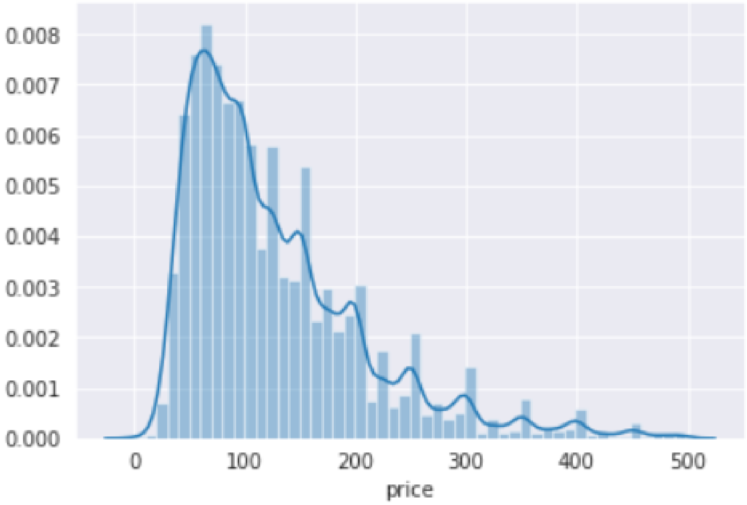
Let’s look at our variable of interest price, which is continuous (USD, $).

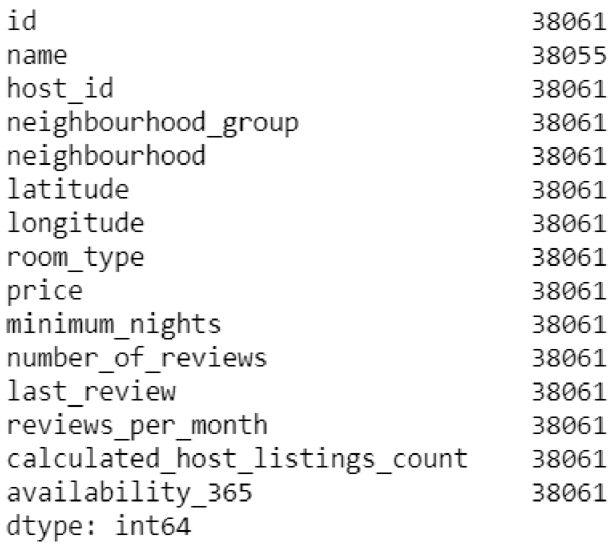


Even though data looks fairly normal there is a low % of outliers that push the price as high as $10,000 USD a night. We can eliminate the <1% outliers (most probably luxury places) since we are busy dealing with the average tourist

As we can see now we have something closer to a normal distribution with a shorter tail What if we repeat the process?

After excluding the 1% outliers we end up with





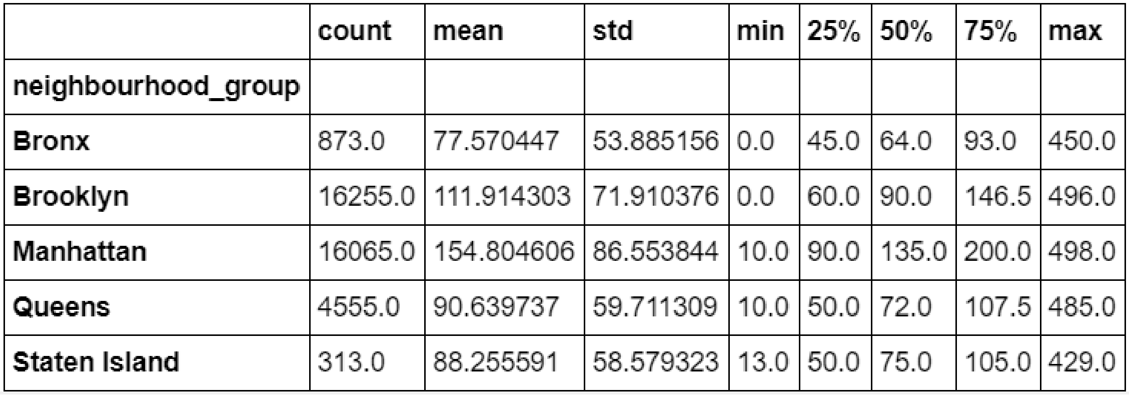
```
id                                38061
name                              38055
host_id                           38061
neighbourhood_group               38061
neighbourhood                     38061
latitude                          38061
longitude                         38061
room_type                         38061
price                             38061
minimum_nights                    38061
number_of_reviews                 38061
last_review                       38061
reviews_per_month                 38061
calculated_host_listings_count    38061
availability_365                  38061
dtype: int64
```

Now we have only 38061 (from an initial 48843). But it is a high number. We are sacrificing some outliers in exchange of better power statistical techniques (assuming a normal distribution)

Results - Now let’s try to answer to our questions:

1) **Is the average price different between neighborhoods?**

Table – Distribution of price across neighborhoods (in USD)



	count	mean	std	min	25%	50%	75%	max
neighbourhood_group								
Bronx	873.0	77.570447	53.885156	0.0	45.0	64.0	93.0	450.0
Brooklyn	16255.0	111.914303	71.910376	0.0	60.0	90.0	146.5	496.0
Manhattan	16065.0	154.804606	86.553844	10.0	90.0	135.0	200.0	498.0
Queens	4555.0	90.639737	59.711309	10.0	50.0	72.0	107.5	485.0
Staten Island	313.0	88.255591	58.579323	13.0	50.0	75.0	105.0	429.0

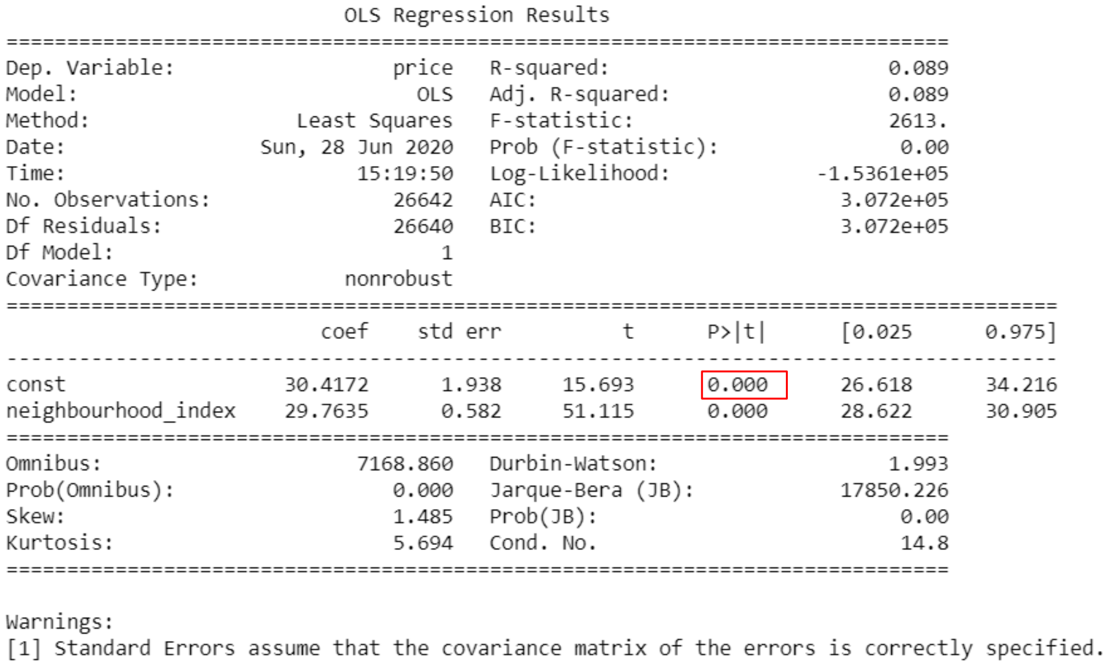
ANOVA test: $p<0.05$

As we can see there is a difference in the average price across different neighborhoods, in increasing order; Bronx (78), Staten Island (88), Queens (60), Brooklyn (111), and Manhattan (155). With values in USD ($). This difference is statistically relevant ($p<0.05$), after running an ANOVA test.

2) Is there an association between the average price an the neighborhood?

Here we have to run a simple linear regression, using Price as dependent variable and type of Neighborhood as independent variable.

- Null hypothesis: There is no association between price and Neighborhood (p>0.05)
- Alternative hypothesis: There is an association between price and Neighborhood (p<0.05).



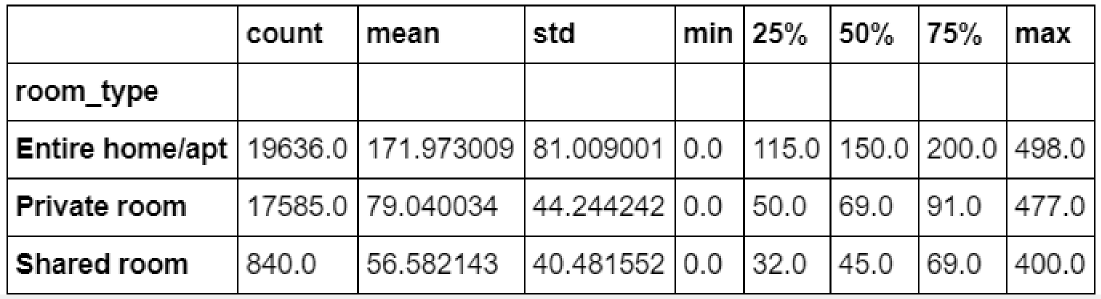
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.089
Model:                            OLS   Adj. R-squared:                  0.089
Method:                 Least Squares   F-statistic:                     2613.
Date:                Sun, 28 Jun 2020   Prob (F-statistic):               0.00
Time:                        15:19:50   Log-Likelihood:            -1.5361e+05
No. Observations:               26642   AIC:                         3.072e+05
Df Residuals:                   26640   BIC:                         3.072e+05
Df Model:                           1
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                  30.4172      1.938     15.693      0.000      26.618      34.216
neighbourhood_index    29.7635      0.582     51.115      0.000      28.622      30.905
==============================================================================
Omnibus:                     7168.860   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            17850.226
Skew:                           1.485   Prob(JB):                         0.00
Kurtosis:                       5.694   Cond. No.                         14.8
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

*Neighborhood is a categorical variable with dummy levels (each Neighborhood is a level).

After running a simple linear regression, there is an association between Price and the Neighborhood where the rental property is located (p<0.05)

3) Is the average price different between whole apartments and rooms?

Table – Distribution of price across type of building (in USD)



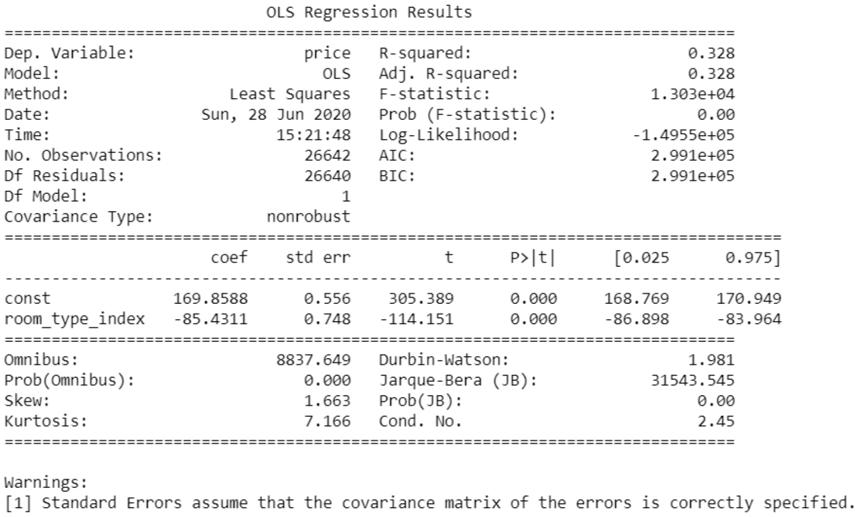
	count	mean	std	min	25%	50%	75%	max
room_type								
Entire home/apt	19636.0	171.973009	81.009001	0.0	115.0	150.0	200.0	498.0
Private room	17585.0	79.040034	44.244242	0.0	50.0	69.0	91.0	477.0
Shared room	840.0	56.582143	40.481552	0.0	32.0	45.0	69.0	400.0

There average price of a rental property is different according to the room type; in increasing order; Shared room (57), Private room (79), and entire home/apartment (172). All prices are in USD ($). There is an statistically relevant difference (p<0.05), after running the ANOVA test.

4) Is there an association between the average price and the type of apartment?

Here we have to run a simple linear regression, using Price as dependent variable and type of rental prperty as independent variable.

- Null hypothesis: There is no association between price and type of rental property (p>0.05)
- Alternative hypothesis: There is an association between price and type of rental property (p<0.05)



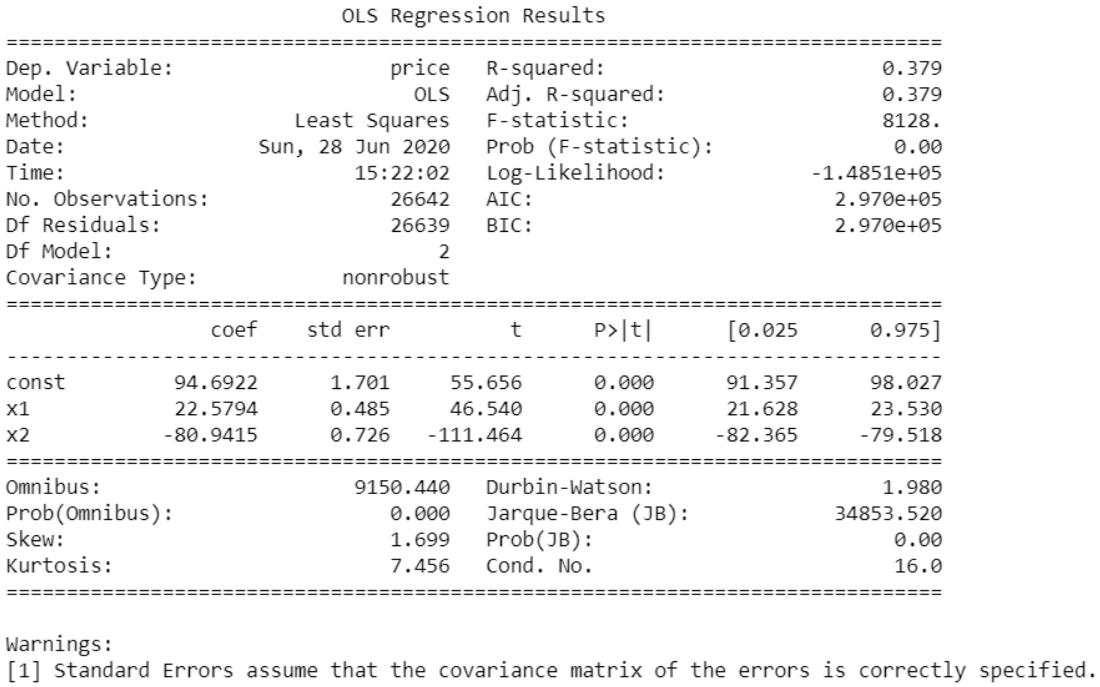
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.328
Model:                            OLS   Adj. R-squared:                  0.328
Method:                 Least Squares   F-statistic:                 1.303e+04
Date:                Sun, 28 Jun 2020   Prob (F-statistic):               0.00
Time:                        15:21:48   Log-Likelihood:            -1.4955e+05
No. Observations:               26642   AIC:                         2.991e+05
Df Residuals:                   26640   BIC:                         2.991e+05
Df Model:                           1
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const             169.8588      0.556    305.389      0.000     168.769     170.949
room_type_index   -85.4311      0.748   -114.151      0.000     -86.898     -83.964
==============================================================================
Omnibus:                     8837.649   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            31543.545
Skew:                           1.663   Prob(JB):                         0.00
Kurtosis:                       7.166   Cond. No.                         2.45
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

*Rental property is a categorical variable with dummy levels (each type of rental property is a level).

After running a simple linear regression, there is an association between Price and the rental property where the rental property is located (p<0.05).

5) Which of these factor is associated with the price when considering all of them?

Here we have to run a multivariable linear regression, using Price as dependent variable and both neighborhood and type of rental property as independent variables.



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  price   R-squared:                       0.379
Model:                            OLS   Adj. R-squared:                  0.379
Method:                 Least Squares   F-statistic:                     8128.
Date:                Sun, 28 Jun 2020   Prob (F-statistic):               0.00
Time:                        15:22:02   Log-Likelihood:            -1.4851e+05
No. Observations:               26642   AIC:                         2.970e+05
Df Residuals:                   26639   BIC:                         2.970e+05
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         94.6922      1.701     55.656      0.000      91.357      98.027
x1            22.5794      0.485     46.540      0.000      21.628      23.530
x2           -80.9415      0.726   -111.464      0.000     -82.365     -79.518
==============================================================================
Omnibus:                     9150.440   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            34853.520
Skew:                           1.699   Prob(JB):                         0.00
Kurtosis:                       7.456   Cond. No.                         16.0
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Here we can see that both neighborhood and type of rental property are important for the pricing of the rental property, with a p-value<0.05

Discussion:

As we can see the local renting property business if a crowded market. However, it has an average return from $78 to $155, according to the neighbourhood, or between $40 to $172 according to the type of property.

An investor can charge higher fees in a whole home/apartment in Manhattan.