|  |
| --- |
| Airbnb Price Prediction for New York City |

|  |
| --- |
| Naveen Jindal School of Management  University of Texas at Dallas  OPRE 6346.501  Spring 2020 |

|  |  |
| --- | --- |
| Group 2: Shrishti Narle, Jabir Mohammed, Abha Ashapure, Umme Samiha  Guided By: Hakki Cankaya |  |



Acknowledgements

We have put our sincere efforts into this project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend our sincere thanks to all of them.

We delightedly wish to express our respect and sincere gratitude to the Faculties of JSOM for their support and encouragement and lending us all the facilities required to proceed with our study.

Our deepest thanks to our Professor, Dr. Hakki Canakaya, for providing his valuable guidance at all stages of the project including his advice, constructive suggestions, supportive attitude, and continuous encouragement. He has keenly observed the project and guided us to make necessary corrections as and when needed.

Table of Contents

[Problem Statement……………………………3](#_Toc37701486)

[Data Exploration and Visualization…….4](#_Toc37701487)

[Data Cleaning…………………………………….6](#_Toc37701488)

[Data Analysis……………………………………..9](#_Toc37701489)

[Conclusion ………………………………………….20](#_Toc37701490)

[Potential Future Work ………………………..26](#_Toc37701491)

[Implementation Environment …………….27](#_Toc37701492)

[References ………………………………………….28](#_Toc37701493)

[Appendix …………………………………………….29](#_Toc37701494)

|  |
| --- |
| Problem Statement |

|  |  |  |
| --- | --- | --- |
| Airbnb is an online marketplace which redefines how people live and travel. This platform has had a positive economic impact in communities around the world as hosts tend to use listings as an additional source of income and allows travelers to choose unique places to stay outside of the generic prospect of hotels. Furthermore, Airbnb has made traveling more affordable for the common person, which makes price a major decision factor when deciding between listings. Because price is such an important component of a listing, we propose an analysis of various factors’ effect on the price of an Airbnb in order to develop a model that will predict the price. Price prediction will utilize data and statistics to provide insight on future prices based on what we know of past behaviors of these factors. | | |
| Questions to answer:   1. What relationship exists between each attribute and Price? 2. Is it possible to create an accurate prediction model which predicts Price based on these factors?   How To Rob An Airbnb | Hacker Noon | | |
|  |  |  |

Data Set

The data has been acquired from Kaggle. We selected Airbnb’s New York City data. Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019.

Dataset Description

This public dataset is from Kaggle and can be found on <https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data> . There are 16 columns and 48,896 rows. Each row consists of a distinct Airbnb listing. The size of data is 6.75 MB. Attributes are as follows:

1. ID

Listing ID.

1. Name

Name of the listing.

1. Host\_id

Host ID.

1. Host\_name

Name of the host.

1. Neighbourhood\_group

New York City is comprised of 5 boroughs including Bronx, Brooklyn, Manhattan, Staten Island, and Queens. This attribute specifies which borough the listing resides in.

1. Neighbourhood

Specifies the neighborhood/area of a borough in which the listing resides in.

1. Latitude

Latitude coordinates of the listing.

|  |
| --- |
| Data Exploration and Visualization |

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

Dataset Description

1. Longitude

Longitude coordinates of the listing.

1. Room\_type

Hosts must list the type of their room as entire home/apartment, private room, or shared room.

1. Price

Price of the listing in dollars (USD).

1. Minimum\_nights

Minimum length of stay requirement set by a host for a listing.

1. Number\_of\_reviews

Number of reviews a listing has.

1. Last\_review

Gives the date of when the last guest review was written.

1. Reviews\_per\_month

Guests may write reviews after their stay. This attribute takes into account an average of the reviews per month a listing receives.

1. Calculated\_host\_listings\_count

Hosts may have multiple listings. An example of this is creating separate listings for multiple rooms within one home/apartment or listing different homes.

1. Availability\_365

Hosts are able to specify which days they want their listing to be available and unavailable. This attribute is the amount of days in a year the listing is available.

|  |
| --- |
| Data Cleaning |

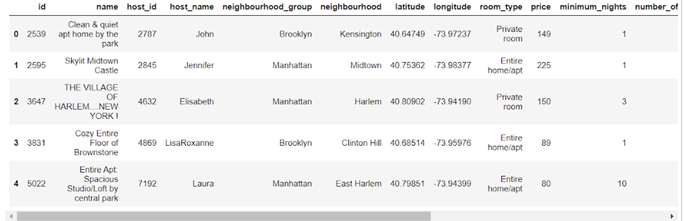
Data cleaning is the process of identifying and removing (or correcting) inaccurate records from a dataset, table, or database and refers to recognizing unfinished, unreliable, inaccurate or non-relevant parts of the data and then restoring, remodeling or removing the dirty or crude data.

Data Reduction

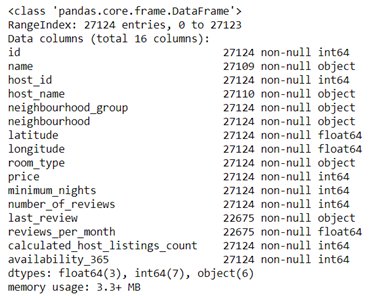
As instructed by the professor we reduced the rows to 27,124 as our original dataset had 48,896 rows, which was too many for our computers to process.

Data Cleaning

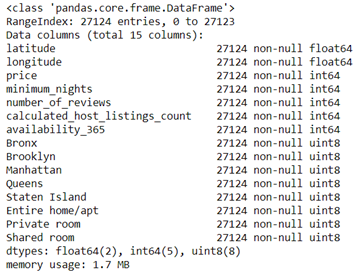
The dataset had some missing and incorrect values. For data cleaning, we used Python, Jupyter Notebook and Pandas library. We have also used NumPy for removing all the NA values. We also normalized the dataset to remove outliers and extreme values.



The above table displays the first five rows of the dataset. The dataset contains 27124 rows and 16 columns.



The above list displays the missing/null values in each row. The total number of rows is 27124. So if a column has rows less than 27124, then the difference between its number of rows and 27124 is the number of missing/null values or incorrect data. As we can see, there is a lot of missing/null data, which needs to be cleaned.



The above list was generated after cleaning the dataset. As we can see, none of the columns contain any missing/null values. During this process, we also eliminated columns like id, name, host\_id, host\_name, neighborhood, last\_review, reviews\_per\_month because these columns were not contributing to the dataset in the sense they held no analytical value as per our understanding. We also converted qualitative variables such as neighbourhood\_group and room\_type into quantitative variables for simplicity of execution.

Lastly, we split the dataset into two parts, one part will be used as the training dataset and the other part will be used as the validation dataset.

|  |
| --- |
| Data Analysis |

Data analysis is the process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions and supporting decisions. To analyze the data after cleaning, there have been a few steps performed such as data manipulation, data visualization and pre-processing.

Data Manipulation

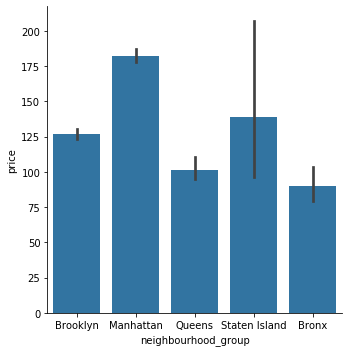
Data has been transformed into numeric by applying various statistical models and pre-processing steps. We have used MinMaxScalar for scaling the data and have transformed the object type variables into numeric.

Data Visualization

Few bar graphs and scatter plots have been created to analyze the data using matplotlib, seaborn and pandas in python and in R as well.

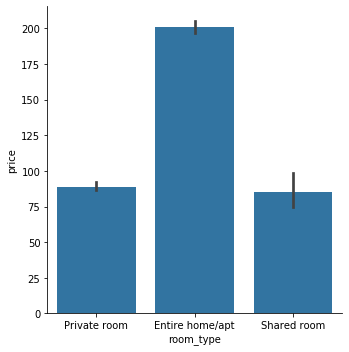
To understand the dataset, we have used exploratory data analysis methods. To get a better understanding of how each attribute affects the Price, we created graphs for each variable against Price. The graphing methods include Bar Charts, Scatterplots, Line Graphs, Correlation Matrices, and Heat Maps. We were able to detect some patterns and trends during this process which is explained below.

1. **Neighborhood\_Group vs. Price**



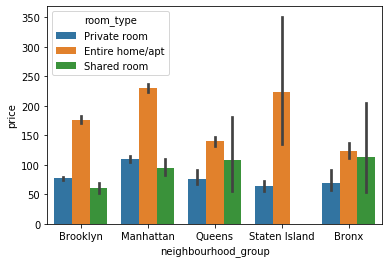
The above graph is plotted between the neighbourhood\_group and price. We can notice that the price for Airbnb in the Manhattan area is the highest. Whereas, the price is lowest in the Bronx area. This is expected as Manhattan has the highest cost of living among boroughs in New York City, which coincides with this neighborhood having the most expensive Airbnbs.

1. **Room\_Type vs. Price**



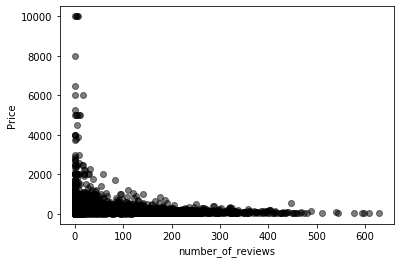
The above graph is plotted between the room\_type and price. We can notice that an entire home is most expensive out of the three. We also examine that there is not much difference between the price of a private room and a shared room. Therefore to further investigate this point we plotted a graph to check if the neighborhood group has an effect on the price of room type.

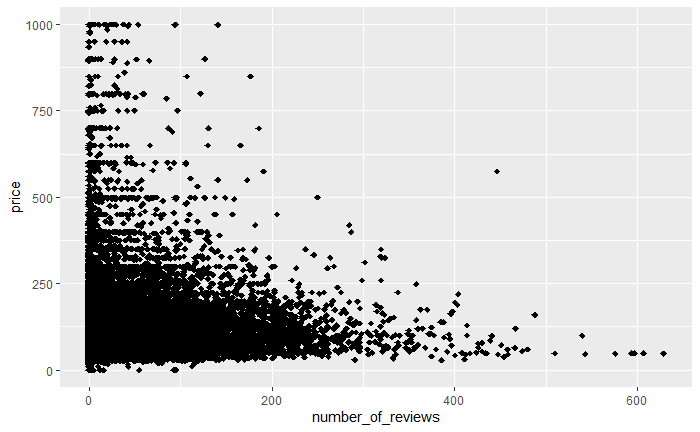
1. **Room\_Type-Wise Graph of Neighborhood\_Group vs. Price**



The above graph is plotted between neighborhood\_group, price and room\_type. In Brooklyn and Manhattan, we observed the expected outcome that private rooms are more expensive than shared rooms. However, we also notice that in Queens and Bronx, the price of a shared room is higher than the price of a private room. We can observe that in these two areas the difference between the prices for all the room types is not quite significant. So, we can deduce that in Queens and Bronx, the Airbnb’s are priced at almost the same amount.

1. **Number\_of\_Reviews vs. Price**





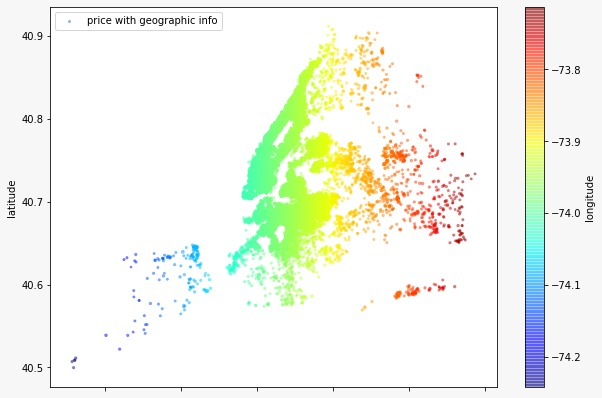
The above two scatterplots are plotted between a number\_of\_reviews and price. We notice a negative relationship between these variables, which means, as the price increases, the number of reviews decreases. We see that there are less reviews for more expensive Airbnbs and more reviews for substantially lower priced Airbnbs. This is expected as most observations seem to remain in the more affordable priced Airbnbs rather than on either end of extremes.

1. **Latitude vs Longitude**



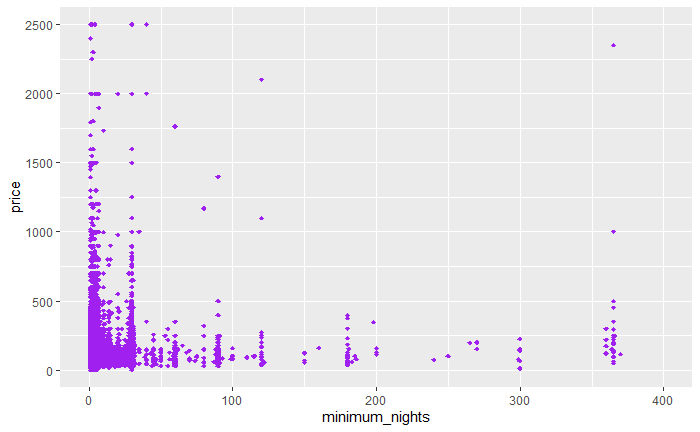
The above scatterplot is plotted between the longitude and latitude of the Airbnb. We cannot observe any trend or pattern in this scatterplot. So to investigate further, we added price as a factor.

1. **Price-Wise Latitude vs Longitude**



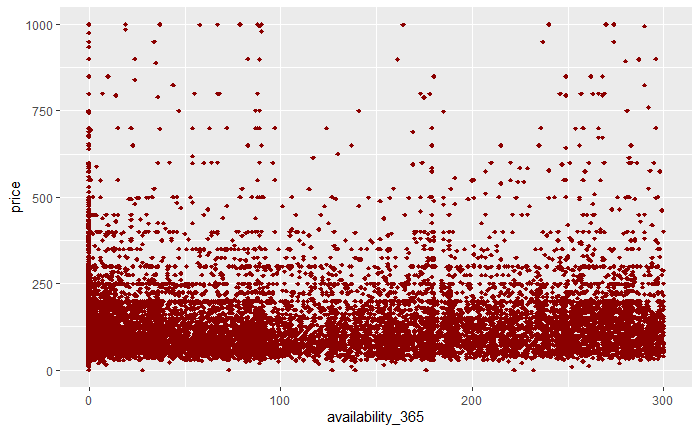
The above scatterplot is plotted between longitude and latitude with respect to price. We can notice a sort of pattern in the scatterplot. The heat bar on the left hand side shows the variation in price according to the color.

1. **Minimum\_Nights vs. Price**



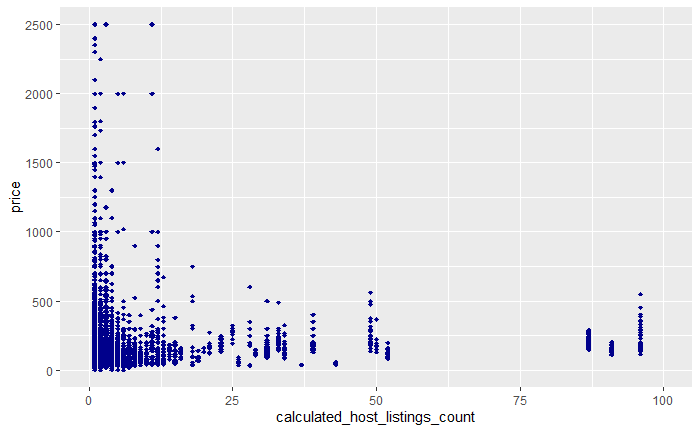
The above scatterplot is plotted between minimum\_nights and price. Although not quite evident, we can notice that both these variables have an inverse relationship. As the minimum\_nights increases, the price decreases. An explanation for this is that, according to Airbnb, renters encourage longer stays by offering weekly and monthly discount rates to have a lower guest turnover rate. However, most guests do not rent Airbnb’s for long term stays as shown by the lower amount of observations for longer minimum stays in the scatterplot.

1. **Availability\_365 vs Price**



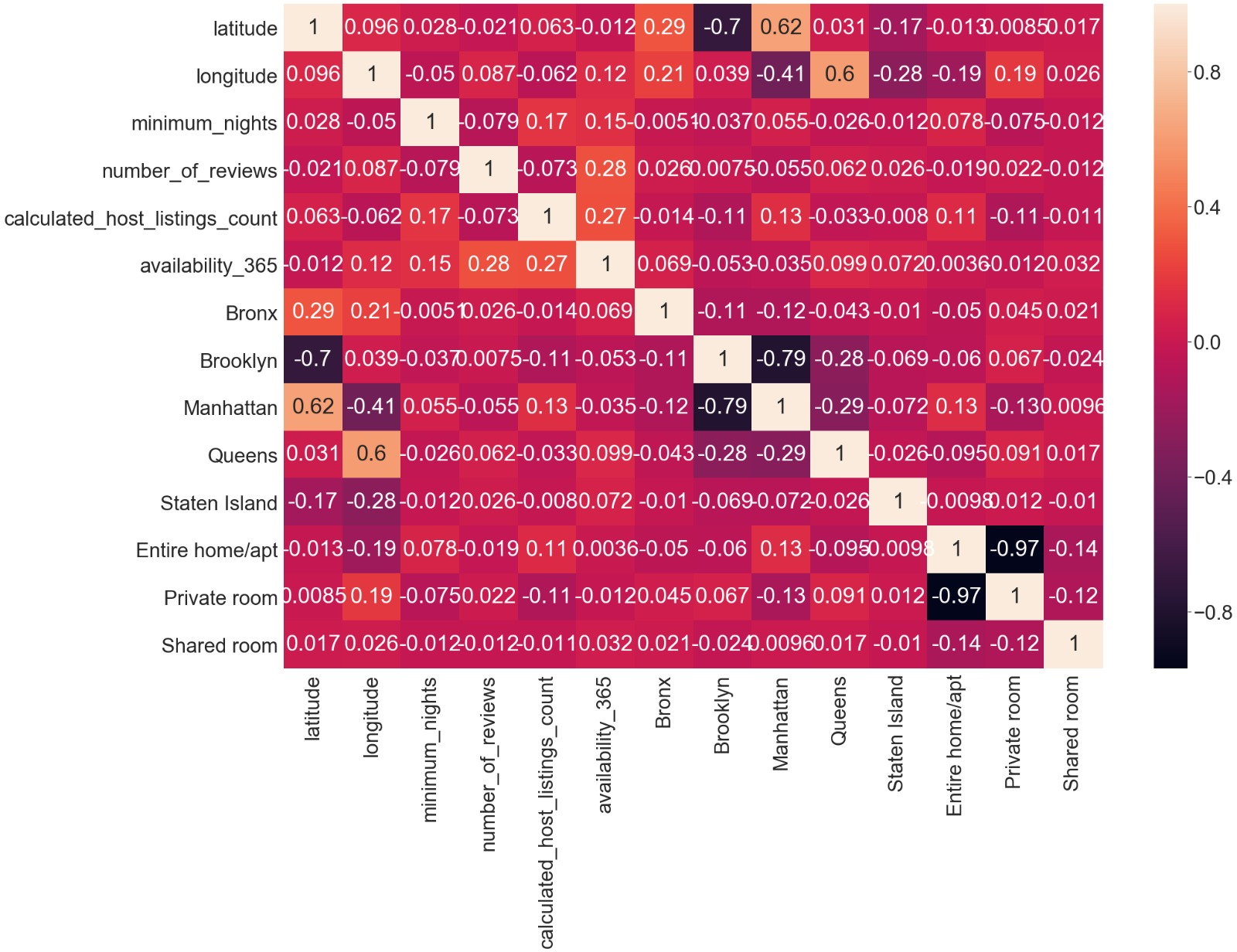
The above scatterplot is plotted between availability\_365 and price. There is no noticeable pattern between these two variables. It seems that availability\_365 has a minimal effect on the price of an Airbnb based on the graph.

1. **Calculated\_Host\_Listings\_Count vs Price**



The above scatterplot is plotted between calculated\_host\_listings\_count and price. After examining the plot, we can observe that there is an inverse relationship between the two variables. The inverse relationship demonstrates that as calculated\_host\_listing\_count decreases, the price of the Airbnb increases.

1. **Correlation Matrix**



This is a heat map showing the collinearity between all of the independent variables among each other. We can see that Manhattan and Brooklyn along with Entire home/apt and Private room both have high collinearity.

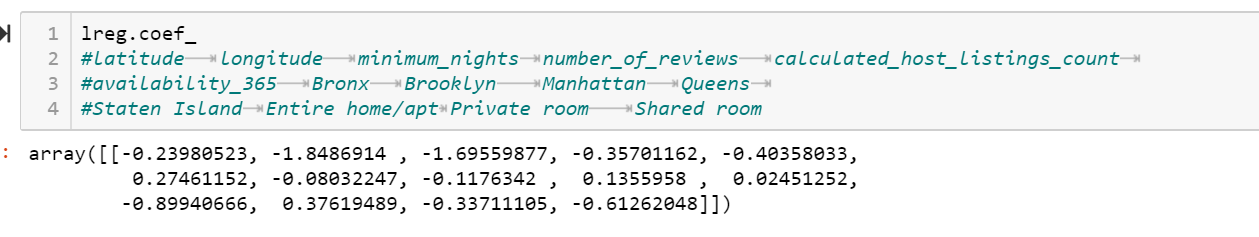
Pre-processing

Preprocessing includes splitting the dataset into training and test datasets and their scaling.

|  |
| --- |
| Conclusion |

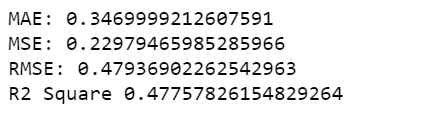
From the screenshot below of the code we can see that we got a linear regression equation as:

*Price =  β0+(-0.24)\*latitude +(-1.85)\*longitude + (-1.69)\*minimum\_nights + (-0.36)\*number\_of\_reviews + (-0.40)\*calculated\_host\_listings\_count + ( 0.27)\* availability\_365 + ( -0.08)\* Bronx + ( -0.12)\* Brooklyn + (  0.14)\* Manhattan + ( 0.0245)\* Queens + ( -0.90)\* Staten Island + ( 0.38)\* Entire home/apt + (-0.337)\* Private room + (-0.612)\* Shared room + μ0*

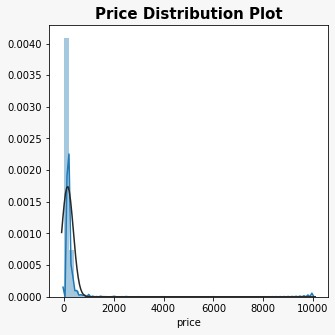
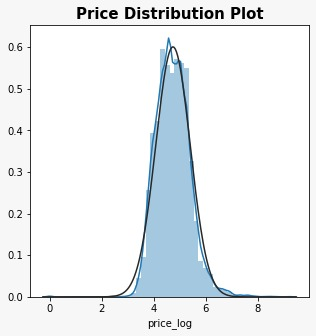


We can see that the most effect we have is of the locality where the Airbnb resides. The major effect on price is on Staten Island where the price decreases by 90% if keeping other factors constant. The costliest Airbnb’s are located in Manhattan followed by Queens. As we have seen before in the graph the entire home or an apartment costs 38% higher than private room or shared room. The price of a shared room is very low at 61%.

As the number of nights increases the prices go down by 169% which is extremely high.



As this is the linear regression model, it explains around 48% of the variation in the data and remaining remains unexplained. We tried performing analysis by other models too like ridge, lasso, SVR but the most variation is explained by Linear Regression only.



Normalized

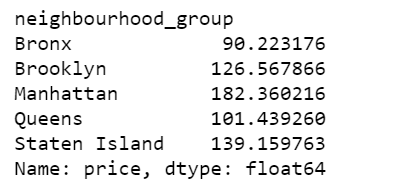
With Positive Skew

The Price variable has positive skewness which we reduced by converting the values of the price column into log values. Then we scaled the data using MinMaxScalar and we applied linear regression on the model so far created with price as our dependent variable and other variables as independent. The above results were obtained.

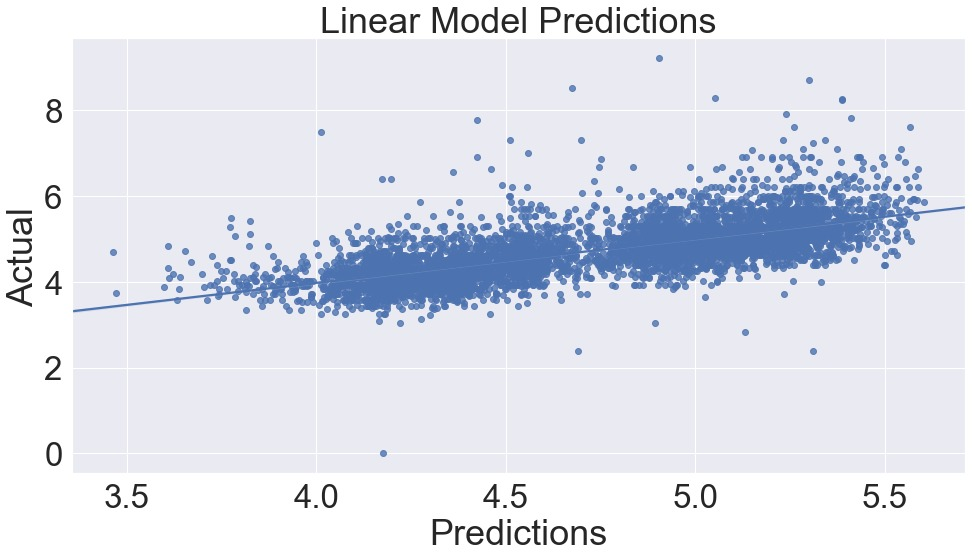
Mean Absolute Error (MAE) shows the difference between predictions and actual values.

Root Mean Square Error (RMSE) shows how accurately the model predicts the response.

R^2  Is calculated to find the goodness of fit measure.



The above picture shows the average price of the Airbnb according to the neighbourhood\_group. As explained by the linear regression model, the costliest area is Manhattan with $182.36 as the average price or about 14% higher price according to the equation. The cheapest place is the Bronx with average price as $90.22.



The above graph that we see is a representation of the final model and the straight line is our equation which we have described above. The actual values are the original prices we have in our dataset and the predicted values are the ones that have been predicted by our model as represented by the line.

1. Neighbourhood\_Group

Prices for Airbnb are highest in the neighborhood group Manhattan, second and third highest in Staten Island and Brooklyn while lowest in Queens and Bronx respectively.

1. Room\_Type

Price for Airbnb’s entire home/apartment are higher than private and shared rooms. From the analysis we can see that the shared rooms cost extremely low compared to entire home/apt and private rooms.

1. Room\_Type wise in different Neighbourhood\_Groups

Comparing the prices room\_type wise in different neighborhood\_groups shows that Brooklyn and Manhattan have higher prices for private rooms than shared while Queens and Bronx have relatively similar prices for private room than shared rooms. All five locations have entire home/apt prices as higher compared to private and shared rooms.

1. Number\_of\_Reviews

There is negative relationship between the number of reviews and prices. Price decreases with the increase in number of reviews. We can assume that elite customers who stays in expensive rooms do not give as many reviews about their stay, which is represented by lower amount of observations.

1. Latitude & Longitude

The relationship between Latitude, Longitude, and Price is based on the areas of New York. We have not thoroughly analyzed the relationship between these three as area is described by neighbourhood and neighbourhood\_groups already.

1. Minimum\_Nights

Prices are lower for long duration while higher for short duration in terms of minimum nights. This is represented as an inverse relationship between minimum\_nights and price.

1. Availability\_365

If the availability is more, then the prices are increasing based on our model.

|  |
| --- |
| Potential Future Work |

Studying the economic effects price of Airbnb has on surrounding neighborhoods.

A more methodological model can be created by expanding out machine learning model with more attribute types, which were not given by our current dataset, to give a more accurate prediction.

A price prediction simulator can be created on the basis of proposed model which could be used to obtain the Airbnb options after providing required preferences and budget.

Constructing the same research in different locations around the world to examine how different cultures and locations can affect the price rather than limiting the study to one city.

Examine how different holidays, seasons, or events may affect the prices of Airbnbs to assess the extent of price gouging.

|  |
| --- |
| Implementation Environment |

Computing Environment

Personal computing environment has been used to analyze the data and prepare the report.

Hardware

Processor- i7

RAM- 16 GB

Software

Operating System: 64- bit operating system, Windows-10

Software Tools Utilized in Analysis

1. Python (For Statistical Analysis)

Version- 3.7

Libraries- Pandas, Numpy, Seaborn, SciPy, Matplotlib, Scikit

1. R (For visualization)

Version- 3.6.2

Libraries- ggplot, dplyr

1. Anaconda
2. Jupyter Notebook
3. R-Studio
4. Microsoft Excel
5. Microsoft Word

|  |
| --- |
| References |

Airbnb. (n.d.). How do I set custom weekly or monthly pricing?

Retrieved from <https://www.airbnb.com/help/article/1233/how-do-i-set-custom-weekly-or-monthly-pricing>

CBS News. (2013, April 16). 3 New York Boroughs Make Top 10 Most Expensive

Places to Live.

Retrieved from <https://newyork.cbslocal.com/2013/04/16/3-new-york-boroughs-make-top-10-most-expensive-places-to-live/>

Dgomonov, & Airbnb. (2019, August 12). New York City Airbnb Open Data.

Retrieved from <https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data>

Sklearn.linear\_model.LinearRegression. (n.d.).

Retrieved from <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>

The Python Standard Library. (n.d.).

Retrieved from <https://docs.python.org/3/library/>

|  |
| --- |
| Appendix |

R Code for Visualization

1. Number\_of\_Reviews vs Price

ggplot(data, aes(x=number\_of\_reviews, y=price)) +

geom\_point(size=2, shape=18) +

xlim(0,600) +

ylim(0,1000)

1. Minimum\_Nights vs Price

ggplot(data, aes(x=minimum\_nights, y=price)) +

geom\_point(color="purple", shape=18) +

xlim(0,400) +

ylim(0,2500)

1. Availability\_365 vs Price

ggplot(data, aes(x=availability\_365, y=price)) +

geom\_point(color="darkred", shape=18) +

xlim(0,300) +

ylim(0,1000)

1. Calculated\_Host\_Listings\_Count vs Price

ggplot(data, aes(x=calculated\_host\_listings\_count, y=price)) +

geom\_point(color="darkblue", shape=18) +

xlim(0,100) +

ylim(0,2500)

Python Code for Visualization and Analysis

import pandas as pd

df= pd.read\_csv('NYC\_airbnb.csv')

df.head()

df.shape

df.info()

df['name'].unique()

df.drop(['name'], axis=1, inplace=True)

df['id'].unique()

df.drop(['id'], axis=1, inplace=True)

df.drop(['host\_name'], axis=1, inplace=True)

df.head()

df.drop(['last\_review'], axis=1, inplace=True)

df.isnull().sum()

df.groupby('neighbourhood\_group')['price'].mean()

df.drop(['reviews\_per\_month'], axis=1, inplace=True)

df.isnull().sum()

df.tail()

import matplotlib.pyplot as plt

import seaborn as sns

sns.barplot("neighbourhood\_group", "price",data=df,hue="room\_type");

# correlation matrix

sns.set(font\_scale=3)

plt.figure(figsize=(30, 20))

sns.heatmap(df.corr(), annot=True)

g = sns.FacetGrid(df, height=5, aspect=.99)

g.map(sns.barplot, "neighbourhood\_group", "price");

g = sns.FacetGrid(df, height=5, aspect=.99)

g.map(sns.barplot, "room\_type", "price");

colors = (0,0,0)

plt.scatter(df['number\_of\_reviews'] ,df['price'] , c=colors, alpha=0.5)

plt.xlabel('number\_of\_reviews')

plt.ylabel('Price')

plt.show()

df['neighbourhood\_group'].unique()

f=pd.get\_dummies(df['neighbourhood\_group'])

df.drop('neighbourhood\_group', axis=1, inplace=True)

df=pd.concat([df,f], axis=1)

from scipy import stats

plt.figure(figsize=(5,5))

sns.distplot(df['price'],fit= stats.norm)

plt.title("Price Distribution Plot",size=15, weight='bold')

Here the Data of price is right skewed so we need to normalix=ze this data. A log transforation would be better to remove the skeweness and normalize the price

df['price\_log'] = np.log(df.price+1)

plt.figure(figsize=(5,5))

sns.distplot(df['price\_log'],fit= stats.norm)

plt.title("Price Distribution Plot",size=15, weight='bold')

df.head()

df.drop('host\_id', axis=1, inplace=True)

df.drop('neighbourhood', axis=1, inplace=True)

df.drop('price',axis=1,inplace=True)

df.head()

df.room\_type.unique()

f=pd.get\_dummies(df['room\_type'])

df.drop('room\_type', axis=1, inplace=True)

df=pd.concat([df,f], axis=1)

df.head()

df.info()

df.plot(kind="scatter", x="longitude", y="latitude", alpha=0.4)

plt.show()

df.plot(kind="scatter", x="longitude", y="latitude",

s=df['price\_log'], label="price with geographic info"

,c='longitude' , cmap=plt.get\_cmap("jet"),

colorbar=True, alpha=0.4, figsize=(10,7),

)

plt.legend()

plt.show()

y= pd.DataFrame(df['price\_log'])

df.drop('price\_log', axis=1, inplace=True)

X= pd.DataFrame(df)

y.head()

X.head()

#preprocessing steps: spliting the dataset

#scaling train and test datasets

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import MinMaxScaler

from sklearn.metrics import accuracy\_score

X\_train\_org, X\_test\_org, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=101)

scaler = MinMaxScaler()

X\_train = scaler.fit\_transform(X\_train\_org)

X\_test = scaler.transform(X\_test\_org)

from sklearn.linear\_model import LinearRegression

lreg = LinearRegression(normalize=True)

lreg.fit(X\_train, y\_train)

predictions = lreg.predict(X\_test)

pred=predictions.reshape(-1)

pred

lreg.coef\_

#Price = β0+(-0.24)\*latitude +(-1.85)\*longitude + (-1.69)\*minimum\_nights + (-0.36)\*number\_of\_reviews + (-0.40)\*calculated\_host\_listings\_count +

# ( 0.27)\* availability\_365 + ( -0.08)\* Bronx + ( -0.12)\* Brooklyn + ( 0.14)\* Manhattan + ( 0.0245)\* Queens +

# ( -0.90)\* Staten Island + ( 0.38)\* Entire home/apt + (-0.337)\* Private room + (-0.612)\* Shared room

#latitude longitude minimum\_nights number\_of\_reviews calculated\_host\_listings\_count

#availability\_365 Bronx Brooklyn Manhattan Queens

#Staten Island Entire home/apt Private room Shared room

y\_test\_np=np.array(y\_test)

y\_test\_shape=y\_test\_np.reshape(-1)

from sklearn import metrics

import numpy as np

def print\_evaluate(true, predicted):

mae = metrics.mean\_absolute\_error(true, predicted)

mse = metrics.mean\_squared\_error(true, predicted)

rmse = np.sqrt(metrics.mean\_squared\_error(true, predicted))

r2\_square = metrics.r2\_score(true, predicted)

print('MAE:', mae)

print('MSE:', mse)

print('RMSE:', rmse)

print('R2 Square', r2\_square)

print\_evaluate(y\_test, lreg.predict(X\_test))

error\_airbnb = pd.DataFrame({

'Actual Values': np.array(y\_test).flatten(),

'Predicted Values': predictions.flatten()}).head(20)

error\_airbnb.head(5)

df2 = pd.DataFrame({'Actual': np.array(y\_test).flatten(), 'Predicted': predictions.flatten()})

df2.head()

df1=df2.head(25)

df1.plot(kind='bar',figsize=(8,8))

plt.grid( linestyle='-', linewidth='0.25', color='green')

plt.grid(linestyle=':', linewidth='0.25', color='black')

plt.show()

plt.figure(figsize=(16,8))

sns.regplot(pred,y\_test\_shape)

plt.xlabel('Predictions')

plt.ylabel('Actual')

plt.title("Linear Model Predictions")

plt.show()