Airbnb Price Prediction for New York City

Project Midterm Deliverable-Group 2

List of group members: Umme Samiha, Abha Ashapure, Shrishti Narle, and Jabir Mohammed.

❖ What has been completed so far

1. Project goal decided

We selected the type 2 project option as we want to enhance our understanding of data concepts and it allows us more hands-on experience with technology. The goal is to predict the prices based on various factors like room type, neighborhood, availability_365, number of reviews, minimum nights, etc.

2. Data collection

The data has been acquired from Kaggle. We selected Airbnb's New York City data. This data includes all the needed information to analyze relationships between price and various metrics of hosts, such as location, availability, room type, etc., to make predictions and draw conclusions.

3. Data Reduction

As instructed by the professor we reduced the rows to 27,124 as our original dataset was about 50,000 rows, which is too many for our computers to process.

4. Data Cleaning

The dataset which we selected had some missing and inappropriate values. For data cleaning, we used Python, Jupyter Notebook and Pandas library.

5. Data Manipulation

Data has been transformed into numeric by applying various statistical models.

6. 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.

7. Preprocessing

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

❖ **What outcomes have been produced, presenting preliminary results or content of the survey/research paper**

We have completed the data cleaning and data understanding portion of the project.
We have cleaned the data by eliminating unwanted values such as null values and missing values. We also normalized the dataset to remove outliers and extreme values.
To understand the dataset, we have used exploratory data analysis methods. The methods include Bar Chart, Scatterplot, Line Graphs, etc. We were able to detect some patterns and trends during this process which is explained below.

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	

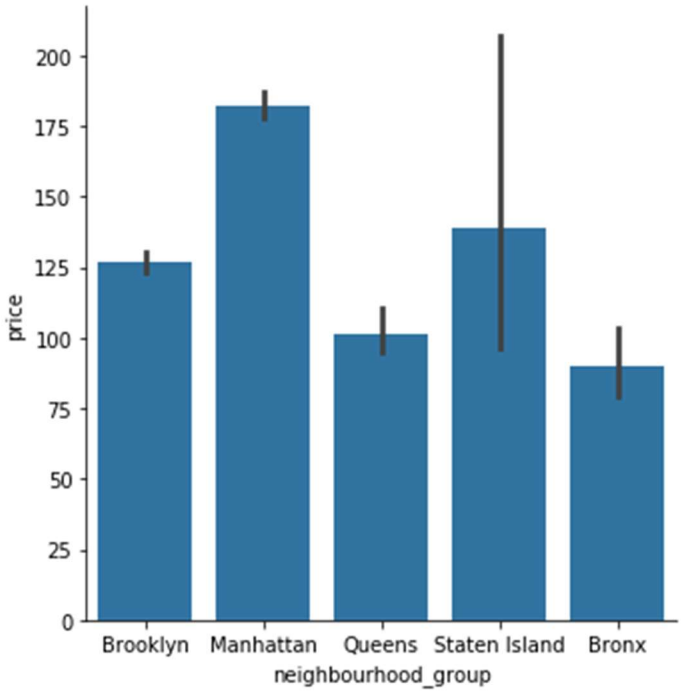
The above table displays the first five rows of the dataset. The dataset contains 27124 rows and 16 columns.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27124 entries, 0 to 27123
Data columns (total 16 columns):
id                                27124 non-null int64
name                              27109 non-null object
host_id                           27124 non-null int64
host_name                         27110 non-null object
neighbourhood_group               27124 non-null object
neighbourhood                     27124 non-null object
latitude                          27124 non-null float64
longitude                         27124 non-null float64
room_type                         27124 non-null object
price                             27124 non-null int64
minimum_nights                    27124 non-null int64
number_of_reviews                 27124 non-null int64
last_review                       22675 non-null object
reviews_per_month                 22675 non-null float64
calculated_host_listings_count    27124 non-null int64
availability_365                  27124 non-null int64
dtypes: float64(3), int64(7), object(6)
memory usage: 3.3+ MB
```

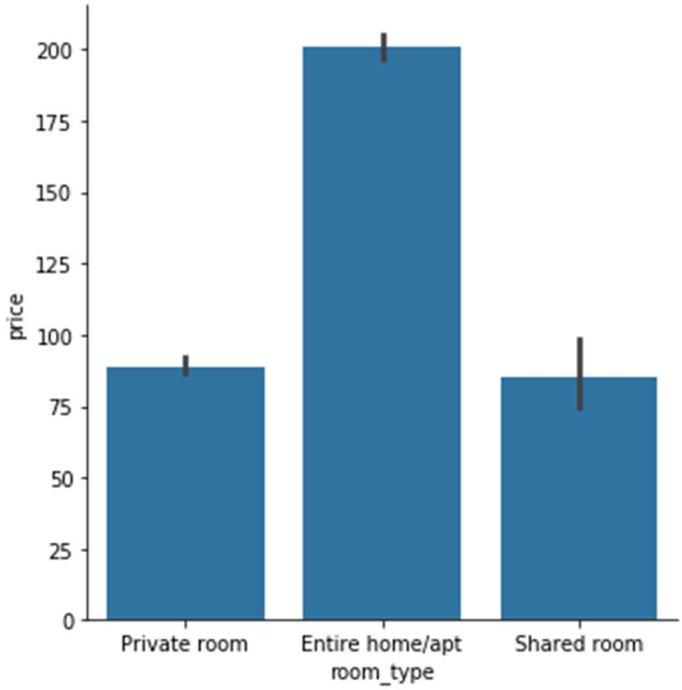
The above list displays the missing/null values in each row. The total number of rows is 27124. So if a column has rows less than 27124, then the difference between its number of rows and 27124 is the number of missing/null values. As we can see, there is a lot of missing/null data, which needs to be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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27124 entries, 0 to 27123
Data columns (total 15 columns):
latitude                          27124 non-null float64
longitude                         27124 non-null float64
price                             27124 non-null int64
minimum_nights                    27124 non-null int64
number_of_reviews                 27124 non-null int64
calculated_host_listings_count    27124 non-null int64
availability_365                  27124 non-null int64
Bronx                             27124 non-null uint8
Brooklyn                          27124 non-null uint8
Manhattan                         27124 non-null uint8
Queens                            27124 non-null uint8
Staten Island                     27124 non-null uint8
Entire home/apt                   27124 non-null uint8
Private room                      27124 non-null uint8
Shared room                       27124 non-null uint8
dtypes: float64(2), int64(5), uint8(8)
memory usage: 1.7 MB
```

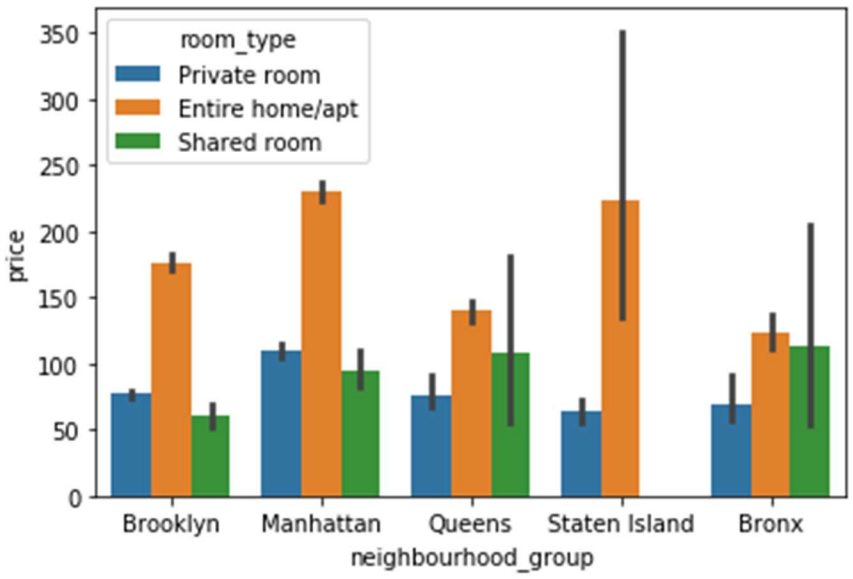
The above list was generated after cleaning the dataset. As we can see, none of the columns contain any missing/null values. During this process, we also eliminated columns like id, name, host_id, host_name, neighborhood, number_of_reviews, last_review, reviews_per_month because these columns were not contributing to the dataset in the sense they held no analytical value. We also converted qualitative variables such as neighbourhood_group and room_type into qualitative 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.



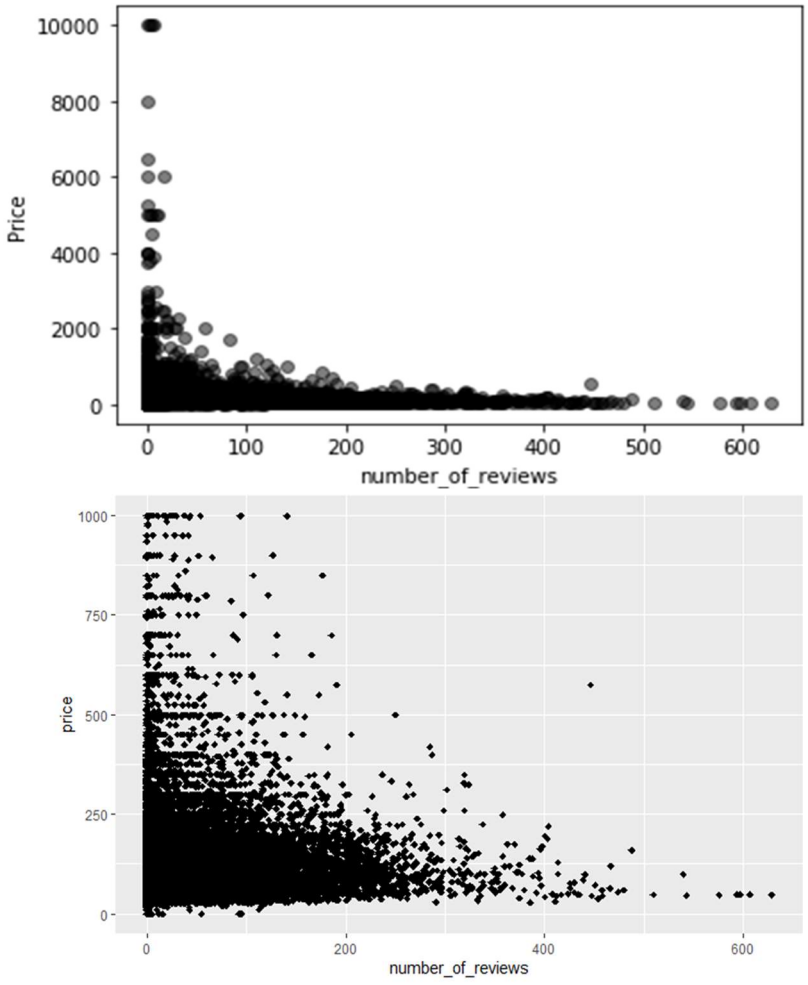
The above graph is plotted between the neighbourhood_group and price. We can notice that the price for Airbnb in the Manhattan area is the highest. Whereas, the price is lowest in the Bronx area.



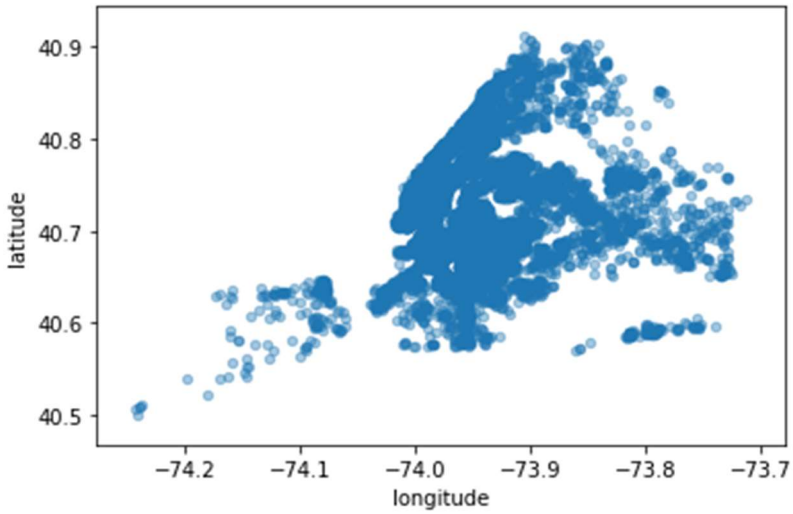
The above graph is plotted between the room_type and price. We can notice that an entire home is most expensive out of the three. We can also notice that there is not much difference between the price of a private room and a shared room. Therefore to further investigate this point we plotted a graph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



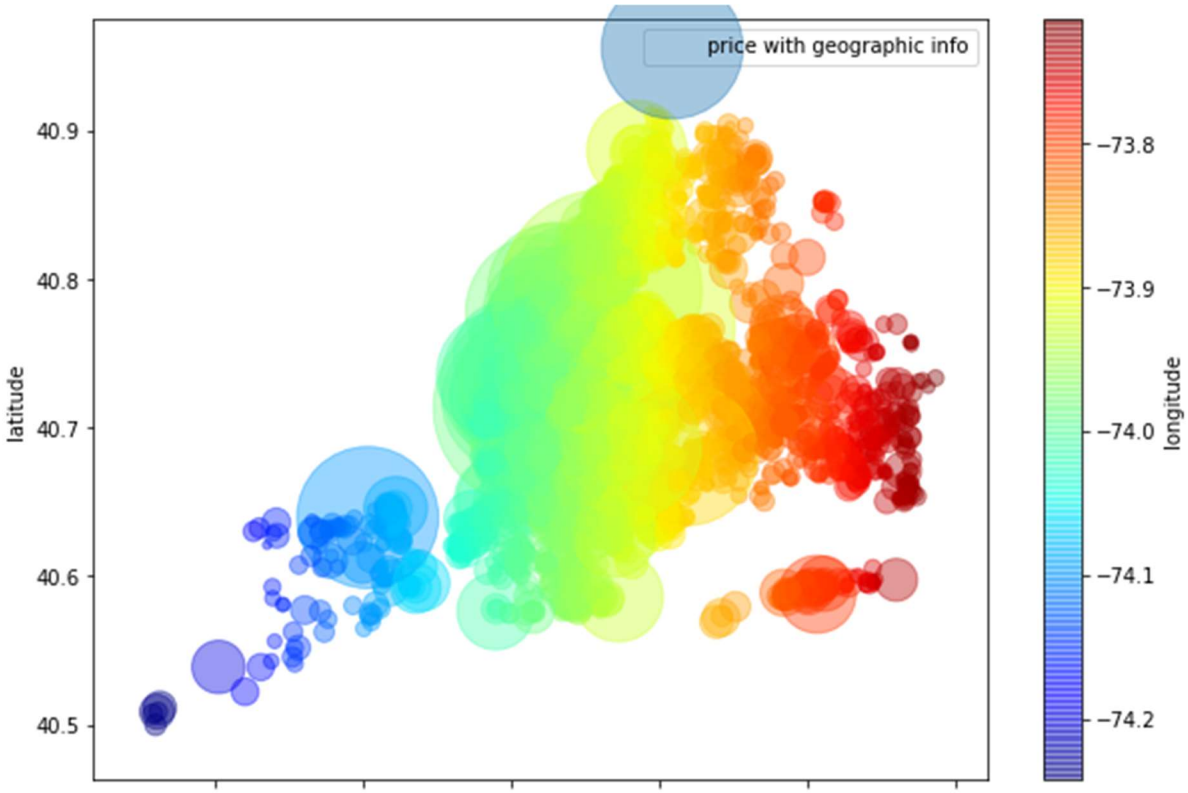
The above graph is plotted between neighborhood group, price and room type. We can notice that in Queens and Bronx, the price of a shared room is higher than the price of a private room. We can also notice that in these two areas the difference between the prices for all the room types is not quite significant. So, we believe that in Queens and Bronx, the Airbnbs are priced at almost the same amount.



The above two scatterplots are plotted between a number of reviews and price. We can notice a negative relationship between these variables, which means, as the price increases, the number of reviews decreases. This shows that the majority of people prefer to stay in low priced Airbnbs.



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.



The above scatterplot is plotted between longitude and latitude with respect to price. We can notice a sort of pattern in the scatterplot. But this pattern is ambiguous.

❖ **What are the next steps**

1. In the next steps, we are going to apply various algorithms like Support Vector Regression, Linear Regression, Decision Tree, Logistic Regression, Ridge and Lasso to train the data and test the accuracy of the model.
2. Explore the data more.
3. Based on the results we will decide which one predicts the results with more accuracy.
4. Interpretation of results with test data.
5. Final report writing.

❖ **Were there any changes since you have started the project; if so, what are these changes and why they needed to happen.**

We have had to make a few changes to the project since we first started. Firstly, the original dataset had about 50,000 rows but we reduced it down to 27,124 rows as per the professor's recommendation to cut the dataset down to a manageable amount for our computers to handle. Accordingly, we planned on using all 16 columns that came with the original dataset, but deemed some columns as unnecessary for our prediction model or contained too many null values to be useful. This includes dropping columns host_id, name, id, host_name, last_review, and reviews_per_month, leaving 11 columns rather than the original 16 columns.

Secondly, our initial plan included utilizing Hadoop, Spark, Flume, and Hive in this project if we felt comfortable doing so, however, due to the COVID-19 circumstances we are unable to work hands-on in our course to practice and learn how to apply these concepts on our project. Due to this, we have decided not to pursue the Hadoop, Spark, Flume, and Hive routes but instead focus mainly on using Python and R as we are most comfortable with these languages.

Overall, there have not been too many changes made to our initial plan other than following through with the professor's guidance to reduce the dataset and focusing our project onto two technology platforms rather than six.