INDEX:

Airbnb Exploratory Data Analysis

VINAYAK ABHINAV
Data science trainees,
AlmaBetter, Bangalore

Abstract:

Airbnb has revolutionized the hospitality industry. Prior to 2008, travelers would have likely booked a hotel or hostel for their trip to another town. Nowadays, many of these same people are opting for Airbnb.

The idea behind Airbnb is simple: Find a way for local people to make some extra money renting out their spare home or room to people visiting the area. Hosts using this platform get to advertise their rentals to millions of people worldwide, with the reassurance that a big company will handle payments and offer support when needed. And for guests, Airbnb can offer a homey place to stay that has more character, perhaps even with a kitchen to avoid dining out, often at a lower price than what hotels charge.

Keywords:machine learning,surge pricing,dynamic pricing,classified labels

1.Problem Statement

Airbnb is an online marketplace since 2008, which connects people who want to rent their homes with people who are looking for accommodations in a particular location.

I. Number of Neighbourhood in each Neighbourhood Groups
II. Distribution of rooms according to latitude and longitude and room density in their regions
III. Room types and their percentage in each neighborhood groups
IV. Relationship between room type and Price
V. Price and Review Relationship
VI. Minimum, Maximum and Average Price according to the neighborhood group
VII. Most popular neighborhood group(region)
VIII. The Most and Least Expensive Neighborhoods

2. Introduction

2.1. What is Airbnb?

Airbnb is an online marketplace since 2008, which connects people who want to rent their homes with people who are looking for accommodations in a particular location. It covers more than 81,000 cities and 191 countries worldwide. The company, which is based in San Francisco, California, does not own any of the property listings, but it receives commissions from each booking like a broker. The name "Airbnb" comes from "air mattress Bed and Breakfast." The Airbnb logo is called the Bélo, which is a short version for saying 'Belong Anywhere'.

Airbnb hosts list many different kinds of properties such as private rooms, apartments, shared rooms, houseboats, entire houses, etc.

2.2. Airbnb Dataset:

This dataset describes the listing activity and metrics in NYC for 2019. It includes all the necessary information in order to find out more about hosts, prices, geographical availability, and necessary information to make predictions and draw conclusions for NYC. The explanation of the variables in our data, which consists of 16 columns and 48,895 rows, will be made in the next part. The data used in this assignment is called New York City Airbnb Open Data which is downloaded from AlmaBetter. This public dataset is a part of Airbnb, and the original source can be found on this website.

2.3. Objectives:

In this project, we will perform an exploratory data analysis (EDA) in order to investigate each of the variables and also come up with a conclusion for the relationship between variables. The main purpose is to identify which variables affect the price mostly. In addition to these, we will explore which neighborhood groups and room types are the most popular ones among the guests, and which hosts are the most preferred ones. The processes during the eda can be listed as below:

- ➔ Data Cleaning
- ➔ Data Preprocessing
- ➔ Data Manipulation
- ➔ Data Visualization
- ➔ Exploring the information

2.4 Questions that can be answered by this analysis

- Number of Neighbourhood in each Neighbourhood Groups
- Distribution of rooms according to latitude and longitude and room density in their regions
- Room types and their percentage in each neighbourhood hroups
- Relationship between room type and Price
- Price and Review Relationship
- Minimum, Maximum and Average Price according to the neighbourhood group
- Most popular neighbourhood group
- The Most and Least Expensive Neighborhoods
- The Most and Least Available Neighborhoods
- Availability of Room Types According to the Neighborhood Groups
- Minimum Nights and Neighborhood Relationship
- The Most Popular Hosts in NYC Airbnb
- Variation in price according to the last review year
- The Average Number of Reviews in Each Neighbourhood
- Most common words used in name column
- Price Group Analysis of Neighborhood Groups

3. Data Explanation:

3.1 Used Libraries:

I have used several packages during the analysis of the historical data of Airbnb in NYC in order to make data manipulation and visualization. The list of packages used in this EDA can be seen below:

- **numpy** => To perform mathematical operations
- **pandas** ===> For dataframe
- **matplotlib** ==> For Visualization of Data
- **seaborn** ====> For Visualization of Data
- **plotly.express** => For Visualization of Data
- **plotnine** ====> For Visualization of Data
- **warnings** ===> For ignorance of any kind of unnecessary warnings
- **dataprep** => Describing and understanding the data
- **wordcloud** ===> for preparing worldcolud
- **PIL** ====> opening, manipulating different image file formats
- **re** =====> Manipulate all kinds of text and data

3.2. Features:

This dataset contains 16 features/variables about Airbnb listings within New York City. Below are the features with their descriptions:

1. **id**: Listing ID (numeric variable)
2. **name**: Listing Title (categorical variable)
3. **host**_id: ID of Host (numeric variable)
4. **host**_name: Name of Host (categorical Variable)
5. **neighbourhood_group**: Neighbourhood group that contains listing (categorical variable)
6. **neighbourhood**: Neighbourhood group that contains listing (categorical variable)
7. **latitude**: Latitude of listing (numeric variable)
8. **longitude**: Longitude of listing (numeric variable)
9. **room_type**: Type of the offered property (categorical variable)
10. **price:** Price per night in USD (numeric variable)
11. **minimum_nights**: Minimum number of nights required to book listing (numeric variable)
12. **number_of_reviews:** Total number of reviews that listing has (numeric variable)
13. **last_review:** Last rent date of the listing (date variable)
14. **reviews_per**_month: Total number of reviews divided by the number of months that the listing is active (numeric variable)
15. **calculated_host_listings_count**: Amount of listing per host (numeric variable)
16. **availability_365**: Number of days per year the listing is active (numeric variable)

3.3. Dataset Insights:

Dataset Insights

1. last_review has 10052 (20.56%) missing values -- **Missing**
2. reviews_per_month has 10052 (20.56%) missing values --- **Missing**
3. host_id is skewed -- **Skewed**
4. longitude is skewed -- **Skewed**
5. price is skewed -- **Skewed**
6. minimum_nights is skewed -- **Skewed**
7. number_of_reviews is skewed -- **Skewed**
8. calculated_host_listings_count is skewed -- **Skewed**
9. availability_365 is -- **Skewed**
10. name has a high cardinality:
    47905 distinct values **High Cardinality**
11. host_name has a high cardinality:
    11452 distinct values **High Cardinality**
12. neighbourhood has a high cardinality:
    221 distinct values **High Cardinality**
13. last_review has a high cardinality:
    1764 distinct values **High Cardinality**
14. last_review has constant length 10
    Constant Length
15. longitude has 48895 (100.0%) negatives **Negatives**
16. number_of_reviews has 10052 (20.56%) zeros **Zeros**
17. availability_365 has 17533 (35.86%) zeros **Zeros**

3.4. Data Statistics

Dataset Statistics

- Number of Variables 16
- Number of Rows 48895
- Missing Cells 20141
- Missing Cells (%) 2.6%
- Duplicate Rows 0
- Duplicate Rows (%) 0.0%
- Total Size in Memory 23.5 MB
- Average Row Size in Memory 504.1 B
- Variable Types
  - Numerical: 10
  - Categorical: 6

4. Steps involved:

Exploratory Data Analysis

After loading the dataset we performed this method by comparing the features available in the dataset Airbnb with other features. This process helped me figuring out various aspects and relationships among the categorical and numerical variables.

Null values Treatment

Our dataset contains a large number of null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result.

Data Cleaning :

In this step we applied a function to cleaning the missing values. I used word missing to the place of missing values.

New features:

On the basis of quantiles I categorize the price in to five groups.
I used 20% , 40% , 60% , 80% and 100% quantile.

1. **Very Low:** If the price is between 0 USD to 61 USD
2. **Low:** If price is between 60 USD and 91 USD
3. **Medium:** If price is between 90 USD and 131 USD
4. **High:** If price is between 131 USD and 201 USD
5. **Very High;** If price is between 200 USD and 1000 USD

In these steps I used seaborn library to find

Standardization of features

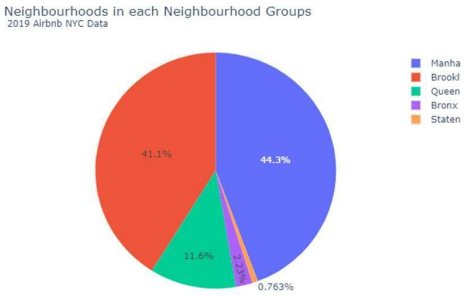
Our main motive through this step was to scale our data into a uniform format that would allow us to utilize

the data in a better way while performing numerical operation and applying different library and function to it. Manhattan has maximum number of neighbourhood which is followed by Brooklyn.

5. EDA and Visualization:

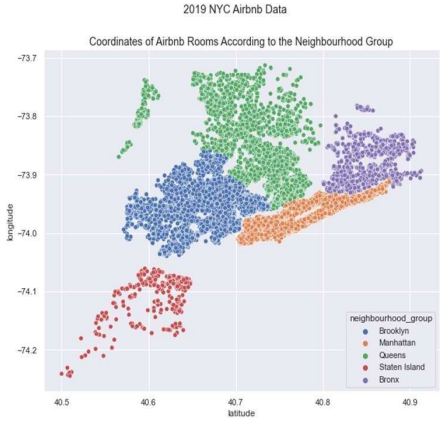
5.1. Number of neighbourhood in each neighbourhood_group:

- Manhattan has maximum number of neighbourhood which is followed by Brooklyn.
- Numbers of neighbourhoods in Manhattan is 21.661K (21661) which is 44.3%.
- Bronx and Staten Island has least Numbers of neighbourhoods.



5.2. Coordinates of Neighborhood Groups:

I used plotly and seaborn library for visualization of latitude and longitude and conclude that the distribution of rooms in Bronx and Staten Island is very less then others. Brooklyn and Manhattan are distributed balanced in their regions.



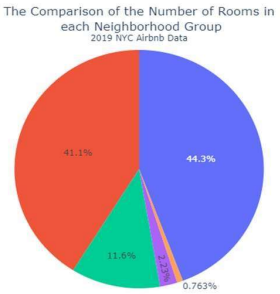
5.3. Count Of Rooms in every Neighbourhood Group

- Manhattan and Brooklyn have 21.66K(21661) and 20.104K(20104) rooms in compare to others.
- Manhattan and Brooklyn have less area than other neighborhood groups but number of rooms are maximum. It means that density of rooms is higher in Manhattan and Brooklyn.

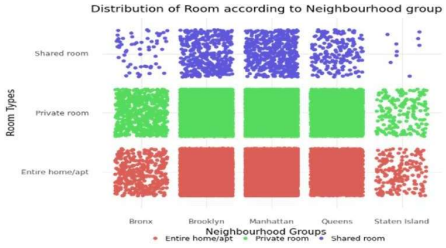
5.2. Percentage of Rooms in Each Neighborhood Group:

There are almost 50000 rooms in our data set. As we want to find the number of rooms

and compare with each other, first we draw a pie chart and then we summarize in the table to provide clear difference. The results illustrate that the rooms in Manhattan and Brooklyn constitute the huge majority, i.e., the sum of these two percentage is equal to 85.4%.



We can infer there's is very less shared room throughout NYC as compared to private and Entire home/apt.



5.4. Relationship between room type and Price:

Before starting my analysis, I checked for the outlier points in this dataset and I take the quantile 1 and 3 as references.

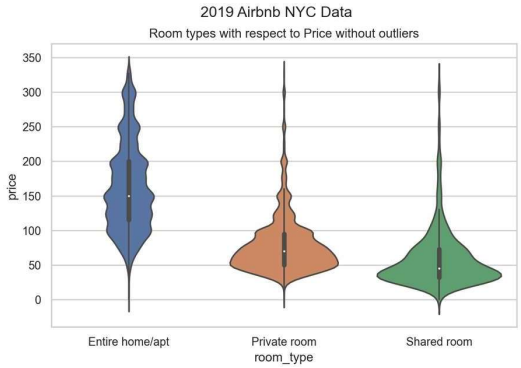
```
qtl1 = airbnb_df.price.quantile(0.25)
qtl3 = airbnb_df.price.quantile(0.75)
iqr = qtl3 - qtl1

lower = qtl1 - iqr * 1.5
upper = qtl3 + iqr * 1.5

lower,upper
```

```
(-90.0, 334.0)
```

When I analyzed the lower and upper bound of the non-outliers data, the lower bound was obtain as minus 90. In the given data set, as I considered the price of the Airbnb room, there is no negative price. For this reason, I only consider the upper bound. The upper bound address the 334. This means that, if the price value is greater than 334, it becomes an outlier values. In this data set, there are 2972 outliers.

RoomType vs Price:

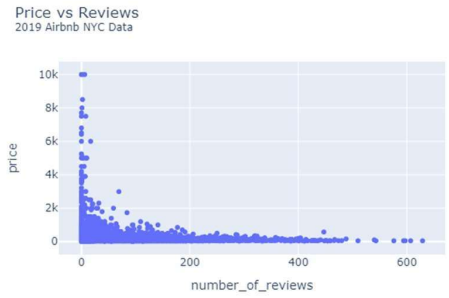
- Average cost of stay in Entire home/apt is near about 153 USD.
- Average cost of Private room type is near about 60 USD.
- Average cost of Shared room type is very less which is 50 USD.



5.5. Price According to the reviews

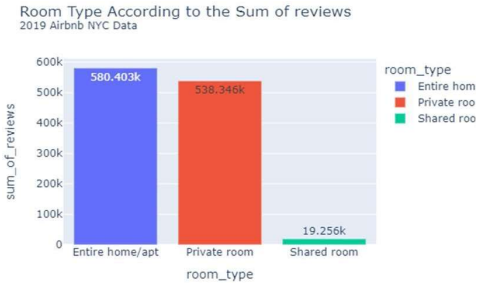
I plot a scatter plot by using seaborn library to know the distribution of price according to the review. I conclude that first choice of customers is low price because spread of the

points is very dense in lower region.



5.6. Room Type according to the reviews

- Number of reviews in Entire home/apt is 580.403K
- Number of reviews in Private room is 538.346K(538346)
- We can observe why the number of Entire home/apt and Private room is maximum because first preference of customers is Entire home/apt and second is Private room.



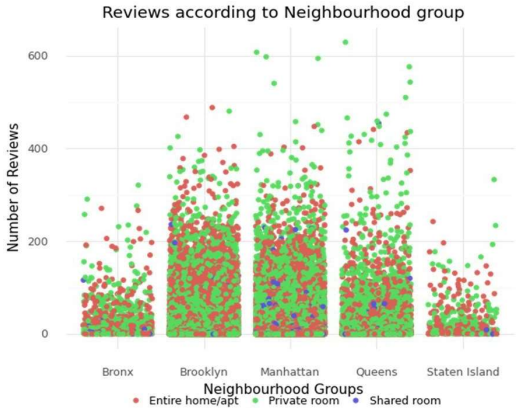
5.7. Minimum, Maximum and Average Price:

I use seaborn library and plot violins to know the average , minimum and maximum price for a particular region.

	neighbourhood_group	price		
		mean	min	max
0	Bronx	77.365421	0	325
1	Brooklyn	105.699614	0	333
2	Manhattan	145.952835	0	334
3	Queens	88.904437	10	325
4	Staten Island	89.235616	13	300

5.8. Most popular neighbour hood group :

In this section I establish a relation between neighborhood group and review per month. Using plotnine library's ggplot geom_jitter() function I plotted a jitter graph.



- From above plot we can observe that in each neighborhood groups , Entire home/apt and Private room have maximum numbers of reviews.

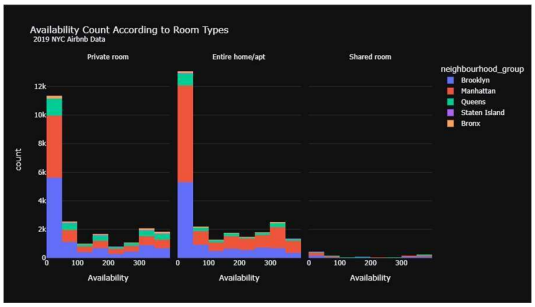
- Brooklyn and Manhattan have maximum numbers of reviews which are followed by the Queens.
- It seems that the first choice of customers is Entire home/apt and Private room.
- Brooklyn and Manhattan are most popular neighborhood group

5.9. The Most and Least Expensive Neighborhoods:

- Staten Island is the most available neighborhood group in the top 15.
- Manhattan, on the other hand, does not have any neighborhood in the top 15.
- There are many neighborhoods in the data set with zero availability.
- Woodrow(Staten Island) , (Bay Terrace, Staten Island(Staten Island)) and New Dorp (Staten Island) do not have availability.

5.10. Availability of Room Types According to the Neighborhood Groups:

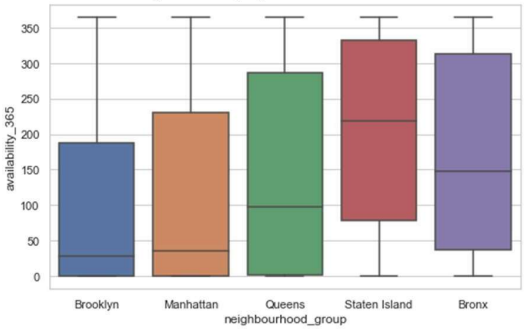
I want to analyze the availability of room types according to the neighborhood groups. By using plotly library I draw a histogram. It can be said that entire home/apt and private room can be reached every day in a year, whereas, shared room is not always accessible.



5.11. The most and least availability of Rooms:

For this relation I plot boxes to know the price in every region.
I conclude that:

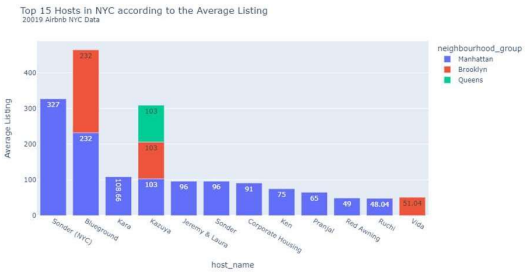
- Staten Island has most has most availability of room all over the year.
- Brooklyn has the least availability of room all over the year.
- Looking at the sideward categorical box plot we can infer that the listings in Staten Island seems to be more available throughout the year to more than 300 days.
- On an average, these listings are available to around 210 days every year followed by Bronx where every listings are available for 150 on an average every year.



5.12. The Most Popular Hosts in NYC Airbnb:

Like we find the most popular neighborhoods, we can also determine the most popular host in NYC according to listing counts.

- In top 15 listings, most of the neighborhoods are from Manhattan.
- Blueground has equal average listings (232) through the Manhattan and Brooklyn



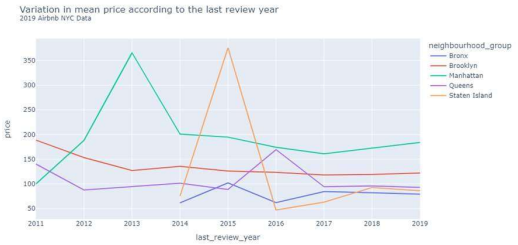
5.13. Variation in price according to the last review year:

- With 10052 null values establish a relationship between last review year and price

I found that:

- All over the year in 2011 average price was lowest in Manhattan.
- In 2013 Manhattan had highest price.
- In 2015 Queens had very highest price.

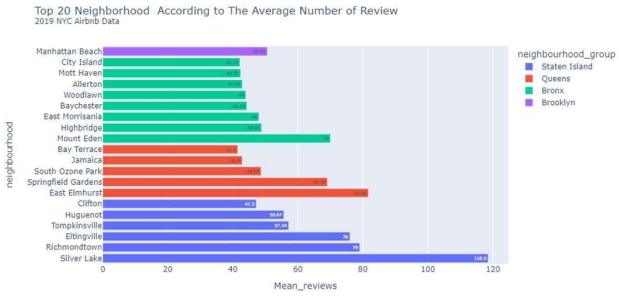
Another analysis can be made by using number of reviews:



5.14. The Average Number of Reviews to the Each Neighborhood:

Another analysis can be made by using number of reviews:

- The results show that Bronx and Staten Island take the most of the reviews.
- On the other hand, there is no neighborhood from Manhattan in the top 20.

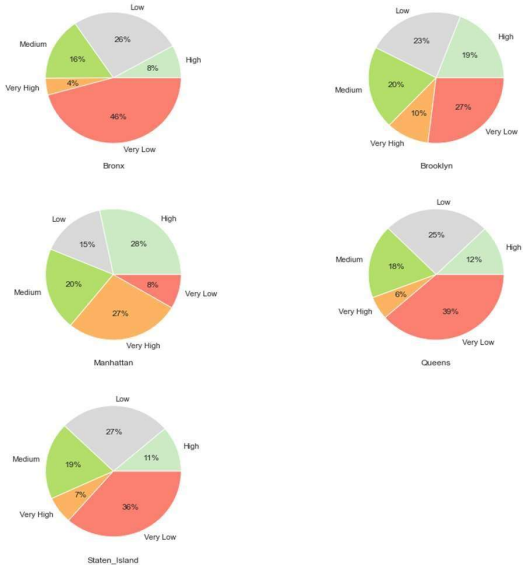


3.15. Word Cloud:

- Like the numerical values, Airbnb data includes verbal information such as name. By using this information, we can obtain the most used words in “name” column which describes the room features.
- With the WordCloud library, I tried to make the wordcloud process reproducible. After geting the data frame of the frequencies, it shows the wordcloud plot.
- As we can see in the plot,"private","room","heart","nyc" and "apartment" words are the most common words in the name column. It means that most of the customers of Airbnb looks for the private rooms, so that these listings have these words in their names.
- As we can see from the plot that “brooklyn” and “manhattan” words are common in the name of the listings. We can infer that Brooklyn and Manhattan would have more listing than the others.
- bedroom comes with different words like "one bedroom","private bedroom","bedroom apt" and "bedroom apartment". These are the most common words in the name column.

5.16. Price Group Analysis of Neighborhood Groups:

By using quantile function, I divide price interval into five. Then, we define values in this intervals as "Very low", "Low", "Medium", "High" , and "Very high". Then by using these categorical values, I prepare pie chart and for each neighborhood group.



We summarize the results as follow:

- ✓ The most of the rooms in Bronx, Queens and Staten Island has very low price.
- ✓ Bronx -- 46% (499)
- ✓ Queens -- 39% (2185)
- ✓ Staten Island -- 36% (136)
- ✓ The rooms with very low, low and medium prices in the Brooklyn are almost distributed equal percentage.
- ✓ The very high price group in Manhattan has higher percentage than the other price groups.
- ✓ Manhattan -- 27%

8. Conclusion:

That's it! We reached the end of our exercise.
Starting with loading the data so far we have done EDA , null values treatment, encoding of categorical columns, feature selection and then In this study, we address the explanatory analysis of the Airbnb data with several key features such as price, neighborhood, neighborhood group, room type, number of reviews, etc. By using these data .

- ➢ We obtain price and neighborhood relationship, i.e., Manhattan is the most expensive Airbnb region when we compare the other neighborhood groups. On the other hand, the least expensive region is Bronx.
- ➢ Another analysis is conducted by using room type. The results show that the entire home/apt type is more preferable and the others are private room and shared room, respectively.
- ➢ In terms of listings Manhattan is on the top.
- ➢ In terms of reviews Staten Island is on the top.
- ➢ In terms of availability Staten Island has most has most availability of room all over the year
- ➢ To make a different analysis instead of numerical analysis, we use WordCloud which makes text mining.

- Number of reviews are also investigated to find which neighborhoods take the most review according to the neighborhood group.

References-

1. Kaggle.com
2. GeeksforGeeks
3. Analytics Vidhya