Capstone Project - 1

Airbnb Bookings Analysis

Name: Rudrajit Bhattacharyya

Cohort: Zanskar Pro

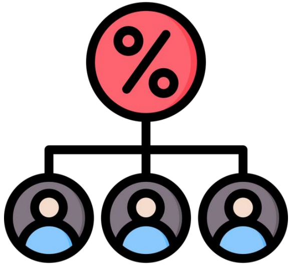
Let’s Explore:

1. Defining the problem statement
2. Defining the data
3. Data Cleaning
4. EDA
    a. Insights about hosts
    b. Insights about neighbourhood groups
    c. Insights about neighbourhoods
    d. Insights about room types
5. Conclusion

How Airbnb works?

Hosts list out their properties in Airbnb

Travellers search and book through Airbnb



Airbnb pays amount to the host after deducting their commission

Defining the problem statement

Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Based in San Francisco, California, the platform is accessible via website and mobile app.

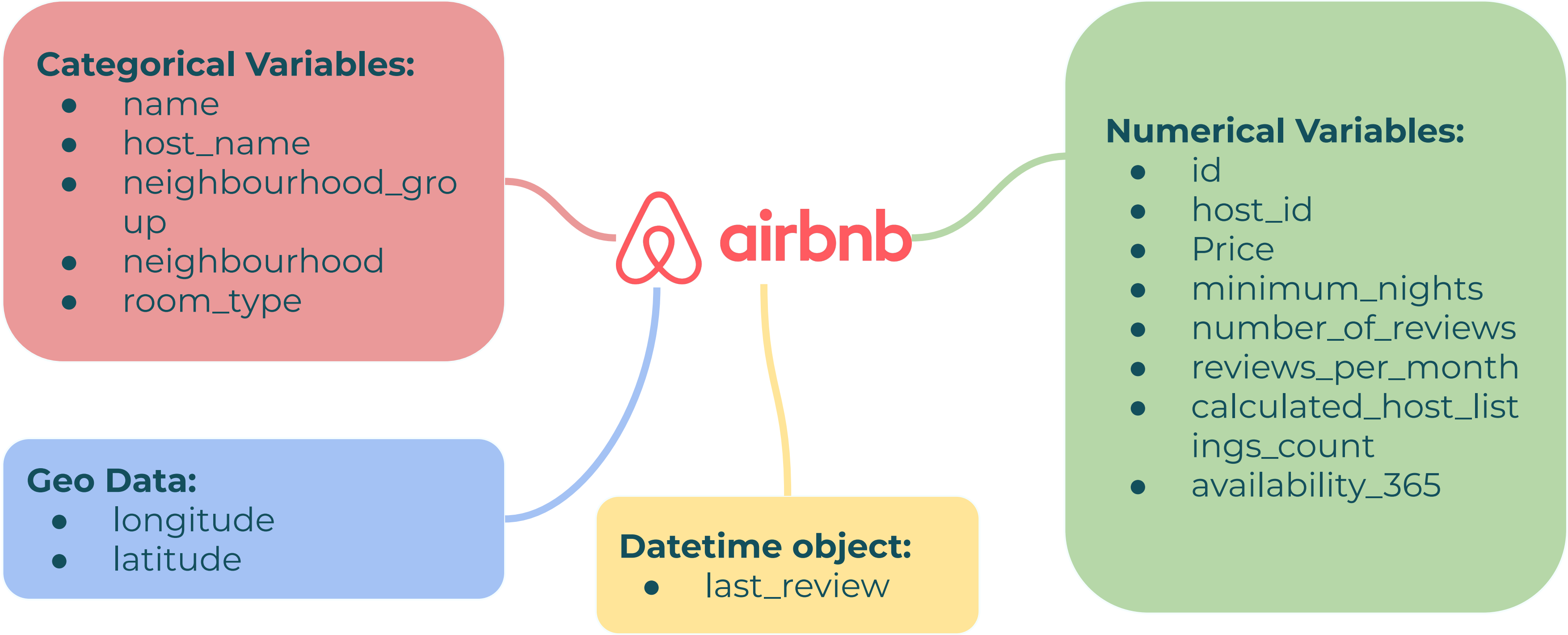
In this EDA project, a dataset of 49000 Airbnb listings in NYC is provided and my task is to explore and find out a few interesting insights which could be helpful in decision making. The main purpose of EDA is to detect any errors, outliers as well as to understand different patterns in the data. To accomplish this, the task was divided into 2 parts as follows:

Data Pipeline

- **Data pre-processing**: The first step consists of looking out for duplicates, missing values and outliers. There were a few missing values present which were dropped or imputed according to the variable at hand.

- **EDA**: Performed some univariate, bivariate and multivariate analysis to uncover insights about hosts, neighbourhood groups, neighbourhoods and room types in Airbnb NYC.

Data Summary



Categorical Variables:

- name
- host_name
- neighbourhood_group
- neighbourhood
- room_type

Numerical Variables:

- id
- host_id
- Price
- minimum_nights
- number_of_reviews
- reviews_per_month
- calculated_host_listings_count
- availability_365

Geo Data:

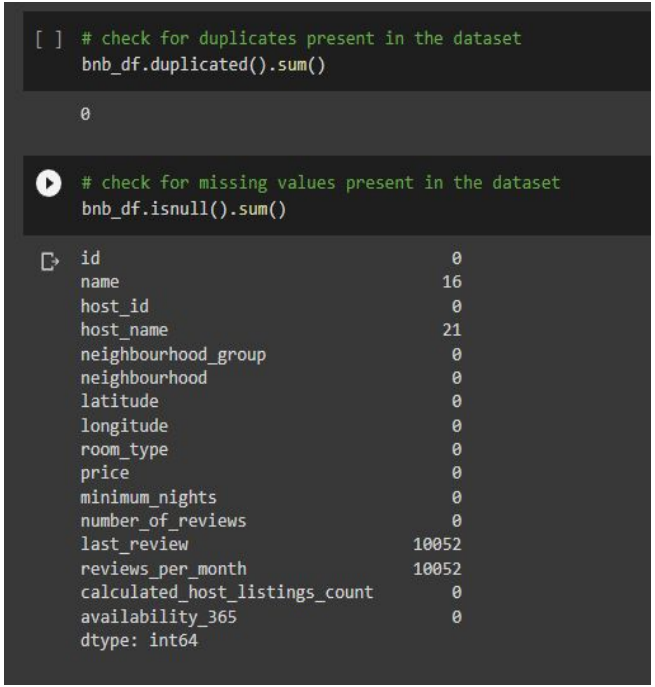
- longitude
- latitude

Datetime object:

- last_review

Data Cleaning

- The first step is to check for duplicate values. As we can observe that there were no duplicate values present in the dataset.
- The second step was to find out missing values in the dataset. We found 4 columns in which there were missing values present.
- We dropped id, name and last_review as these were irrelevant for our analysis.
- We imputed all NaN values in host_name by ‘No Name’ and reviews_per_month by ‘0’.



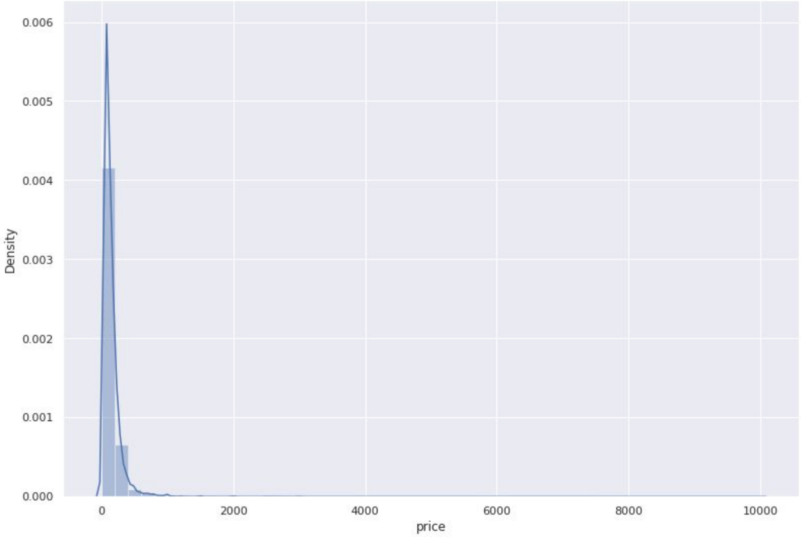
```
[ ] # check for duplicates present in the dataset
    bnb_df.duplicated().sum()

    0
```

```
# check for missing values present in the dataset
bnb_df.isnull().sum()
```

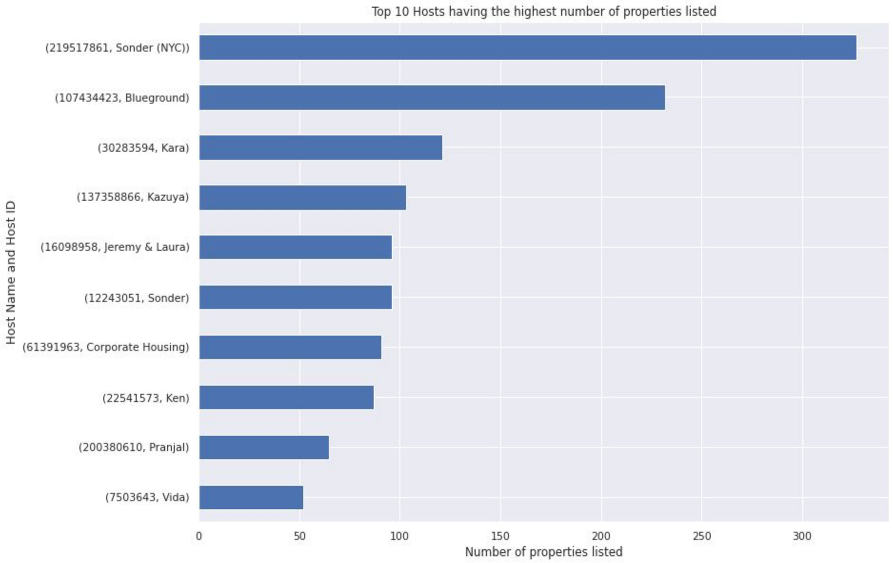
```
id                                    0
name                                 16
host_id                               0
host_name                            21
neighbourhood_group                   0
neighbourhood                         0
latitude                              0
longitude                             0
room_type                             0
price                                 0
minimum_nights                        0
number_of_reviews                     0
last_review                       10052
reviews_per_month                 10052
calculated_host_listings_count        0
availability_365                      0
dtype: int64
```

Exploratory Data Analysis

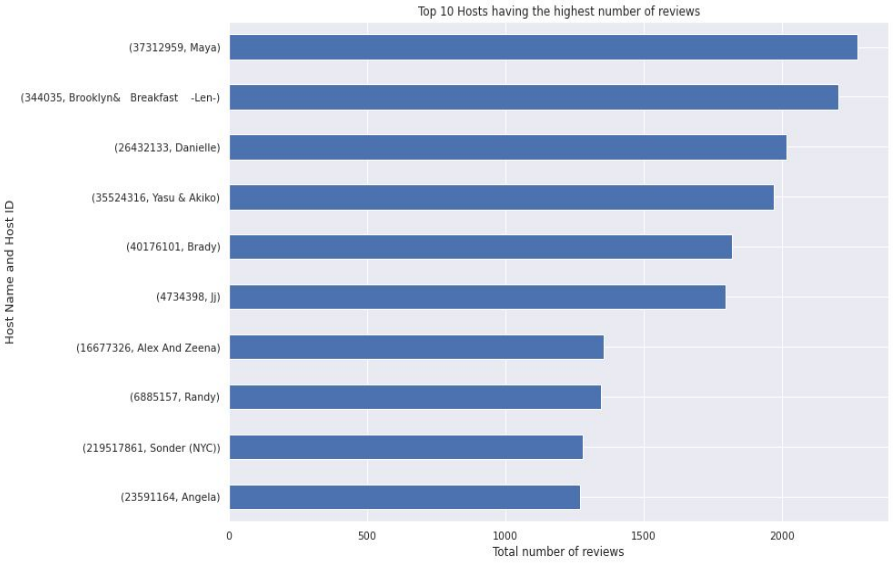
- The price column was heavily skewed.
- Most of the prices was in between 10 to 200$.
- We didn’t treat the high values as outliers because there are few observations in minimum_nights which are high as 1250 and for those few observations the price can naturally go high upto 10000$.



Exploratory Data Analysis


Top 10 Hosts having the highest number of properties listed

Sonder (NYC) has the most properties listed (327)


Top 10 Hosts having the highest number of reviews

Maya has the most number of reviews (2273)

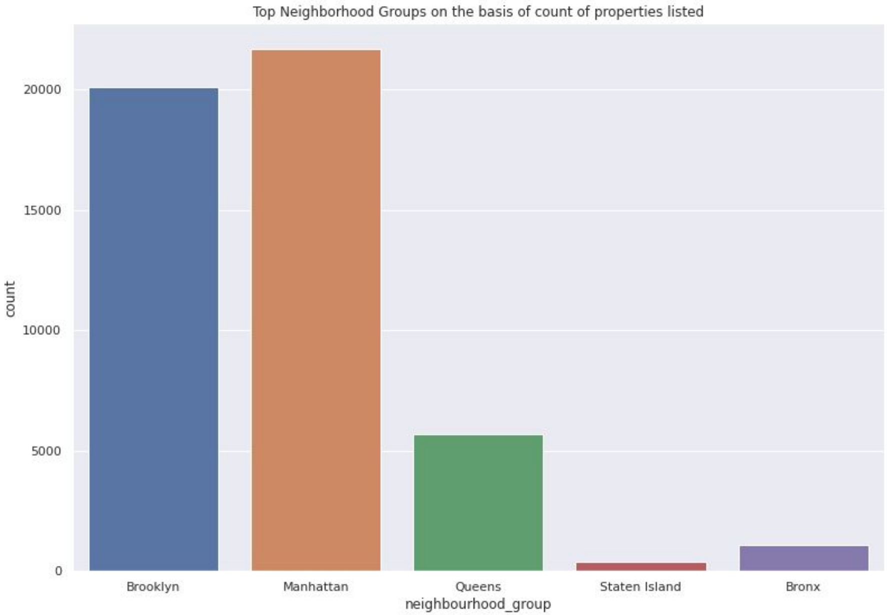
Exploratory Data Analysis

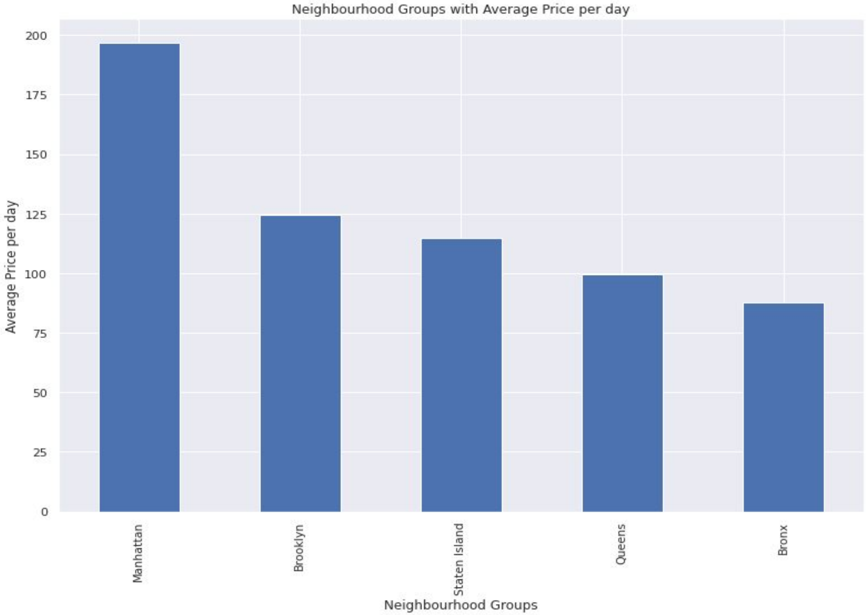
Maya receives the most number of customers and there are multiple reasons behind it:

- The price at which she offers her properties is less than the average of all neighbourhood groups
- The condition for minimum nights is 1 which is way less than many other properties
- She is available for quite a healthy number of days
- As she receives the most reviews, customers will naturally turn their heads towards her properties

Hence, we can conclude that host Maya is the busiest in NYC.

Exploratory Data Analysis

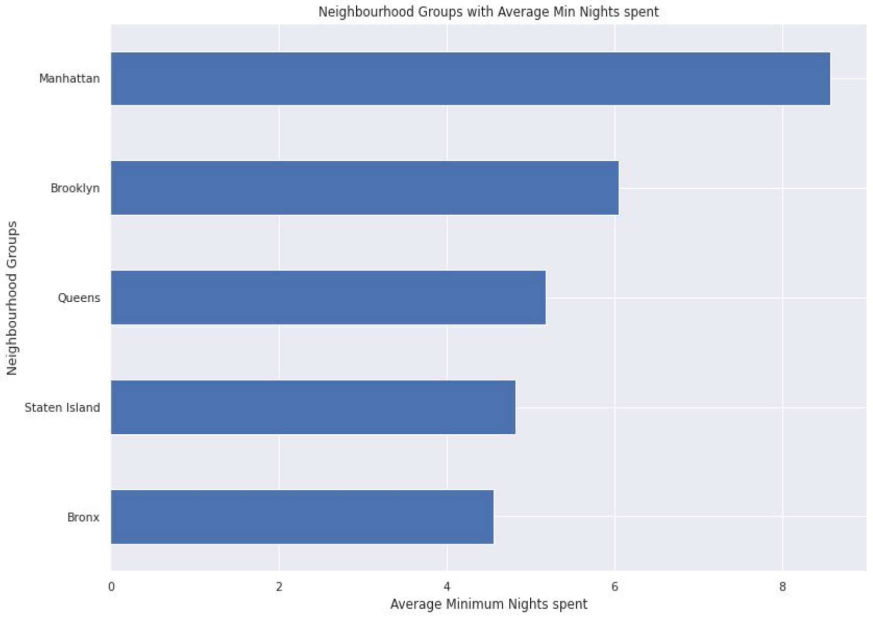




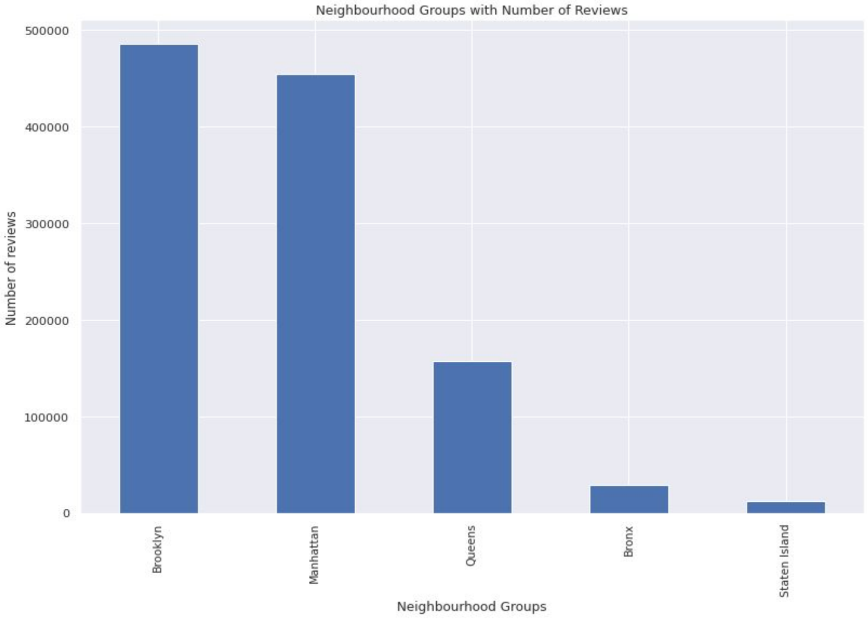
Manhattan has the most number of listings followed by Brooklyn

Manhattan is the most expensive and Bronx, the least expensive neighbourhood

Exploratory Data Analysis

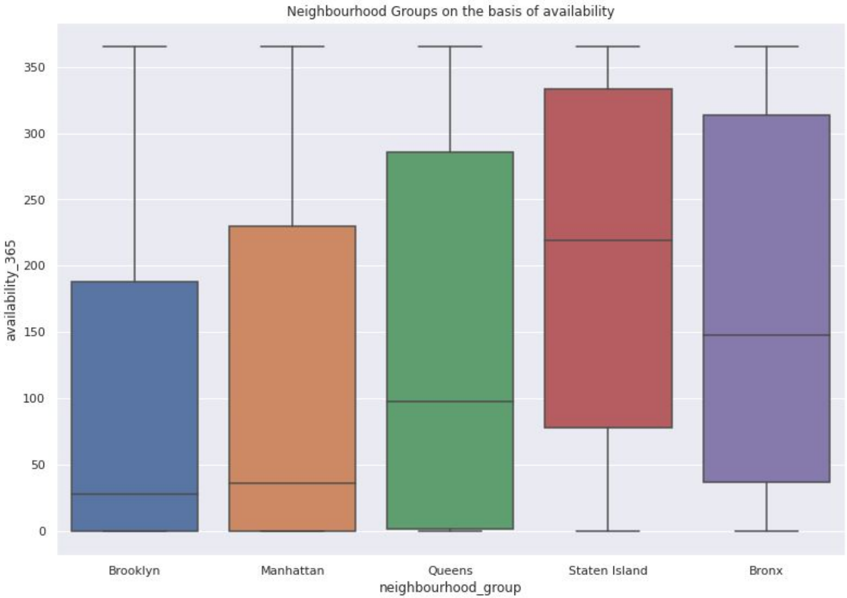


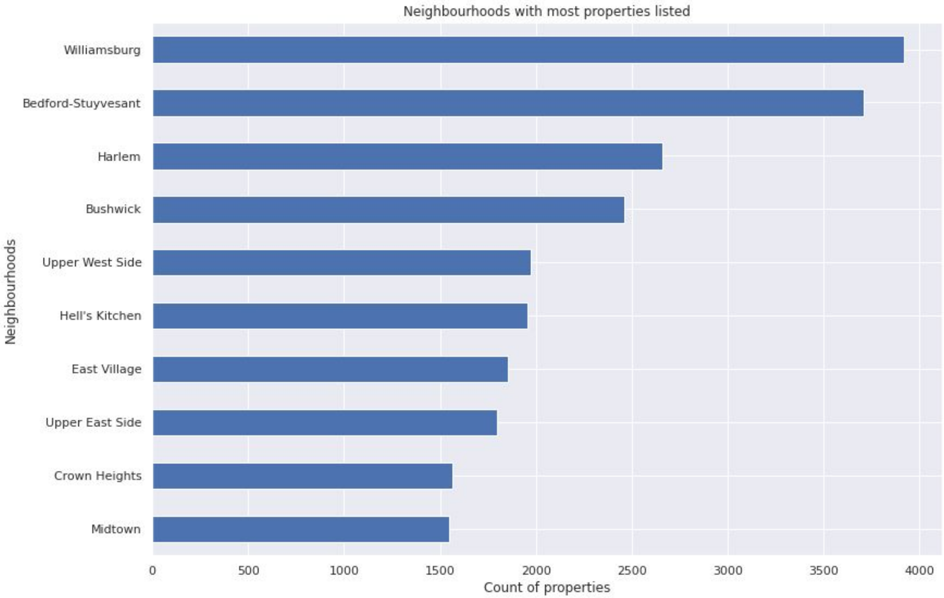
Customers prefer to spend more nights in Manhattan



Brooklyn and Manhattan are the busiest areas

Exploratory Data Analysis

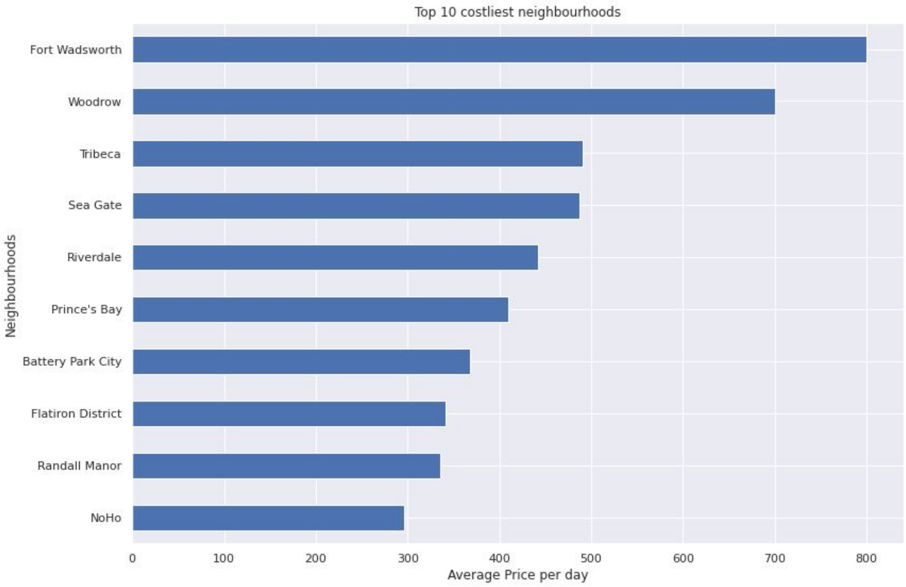


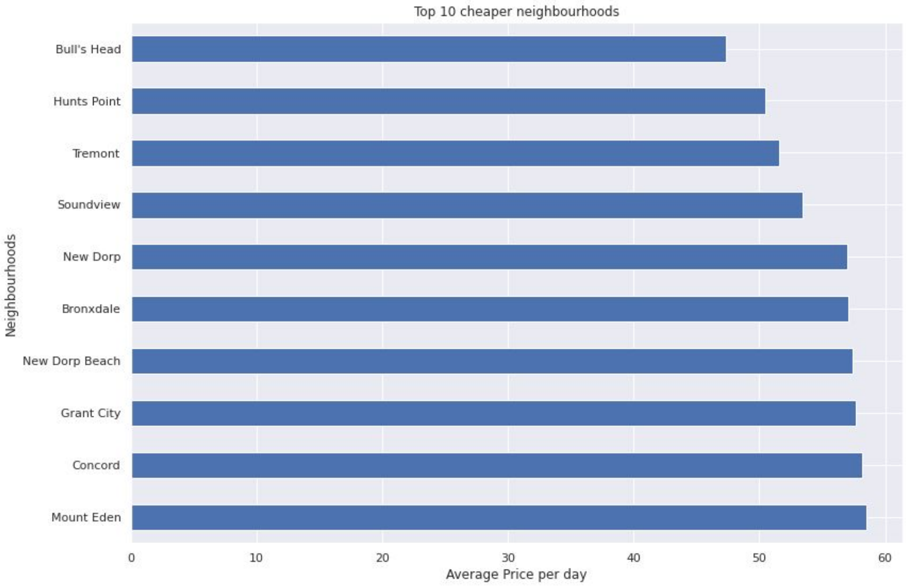


Staten Island has the most available properties out of 365 days

Williamsburg has the most number of properties listed (3919)

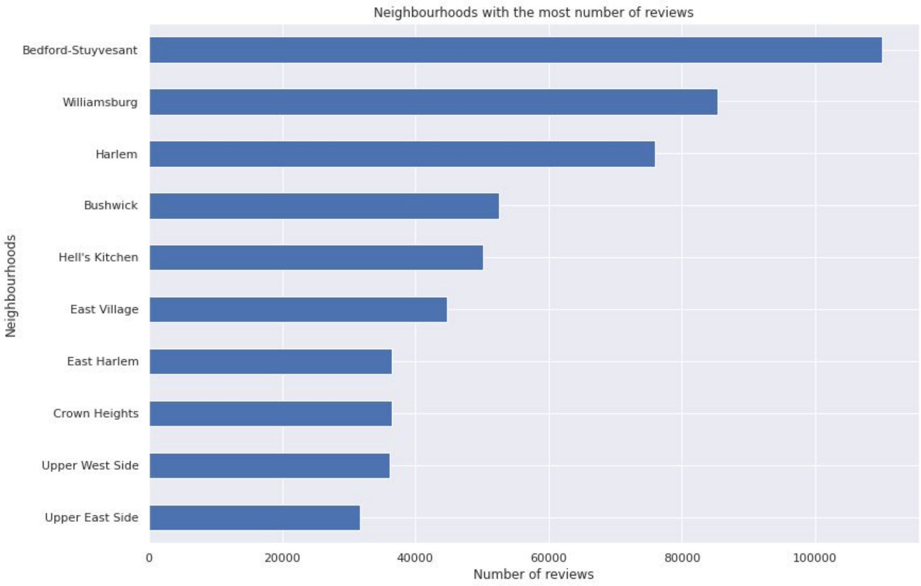
Exploratory Data Analysis

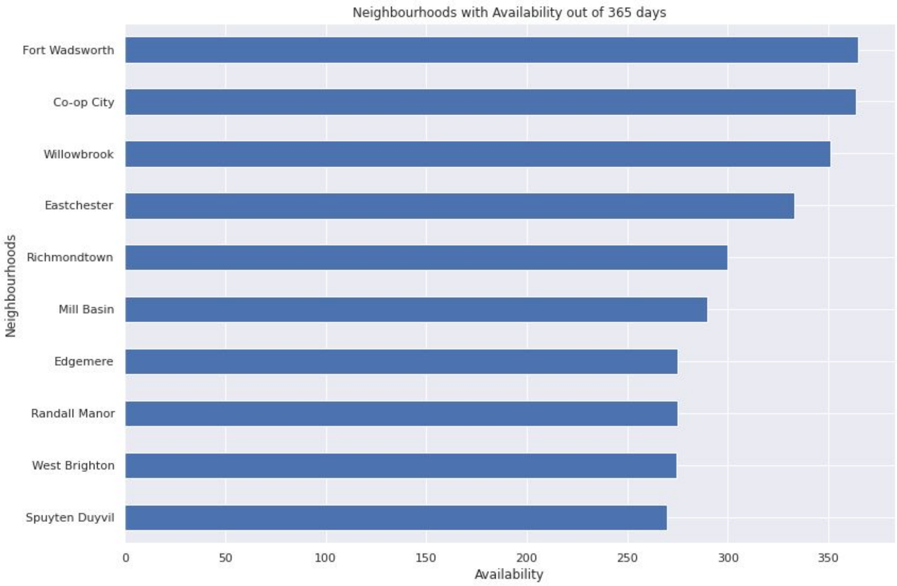




Fort Wadsworth is the most expensive neighbourhood and Bull’s Head is the least expensive. Interestingly both the neighbourhoods fall under Staten Island which is neither the expensive nor the cheapest neighbourhood group.

Exploratory Data Analysis

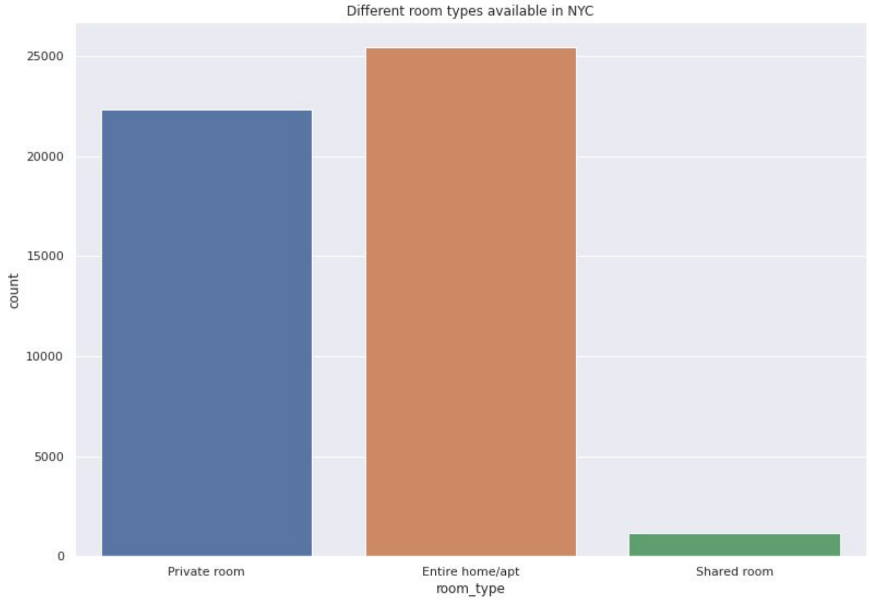


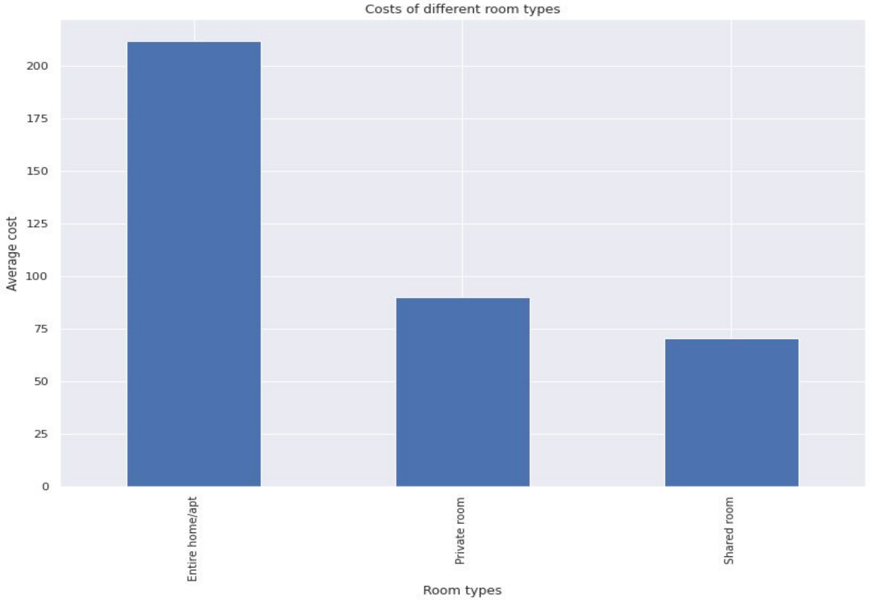


Bedford-Stuyvesant and Williamsburg in Brooklyn is the neighbourhood with the most number of reviews

Fort Wadsworth is the most available neighbourhood which might be due to the high costs

Exploratory Data Analysis

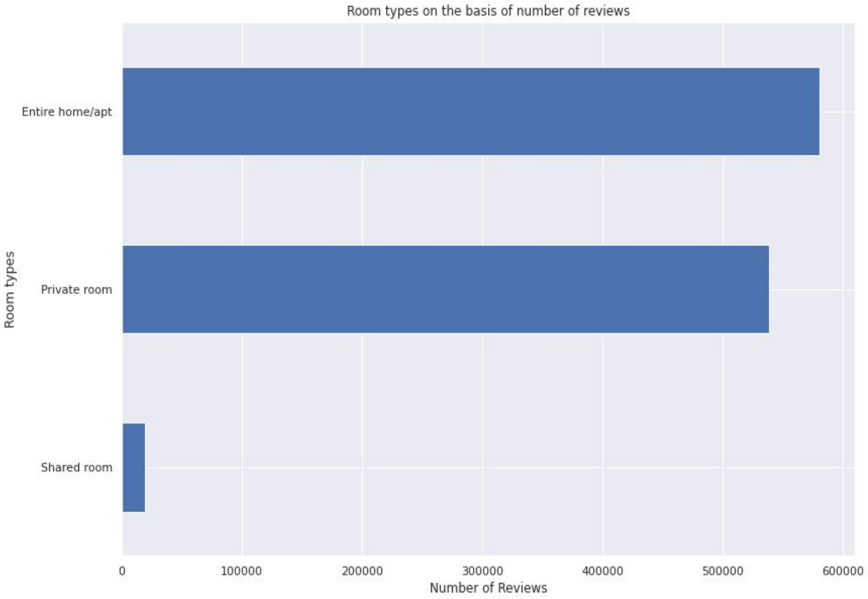




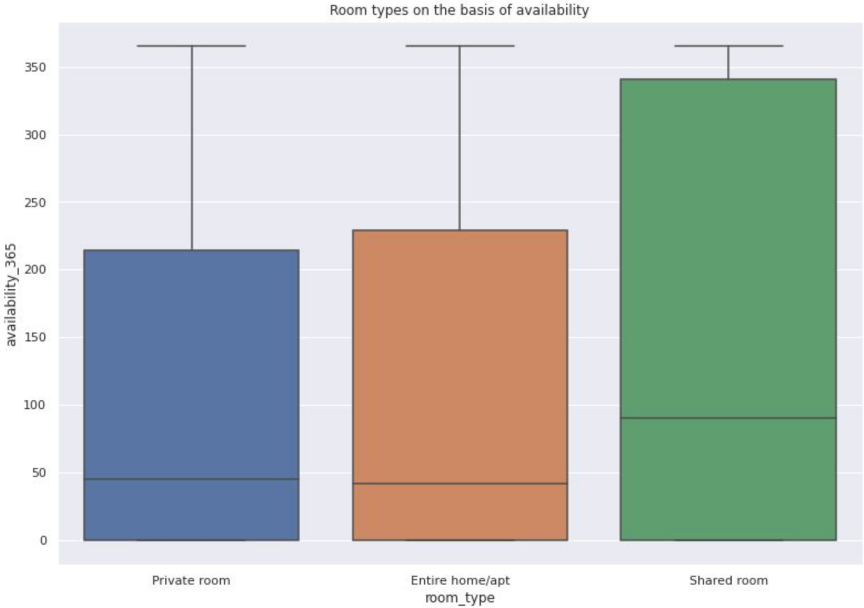
Entire Home/Apt is the most available room type in NYC followed by Private Rooms

Entire Home/Apt costs way more than other room types

Exploratory Data Analysis



People prefer Entire Home or Private Rooms way more than Shared Rooms



Naturally, the shared rooms stay the most available out of 365 days

Conclusion

We have reached the end of our analysis of Airbnb listings in NYC. Let us now summarize few of the important insights we gathered:

Host Maya is the busiest host in NYC and there are multiple reasons in favor of it like price, min_nights, availability, reviews etc. She has a total of 5 properties listed in the same neighbourhood.

Manhattan and Brooklyn are the most expensive neighbourhoods and they receive the most traffic as well. Due to many tourist attractions and number of properties available people tend to visit these two areas more.

Entire Home/Apt is the costliest room type available but still the most preferred room type for customers. There is a significant difference in traffic between Entire Home or Private Rooms than Shared Rooms.

Thank You !