Capstone Project

AirBnb Data Analysis

By – RAHUL SINGH WALDIA

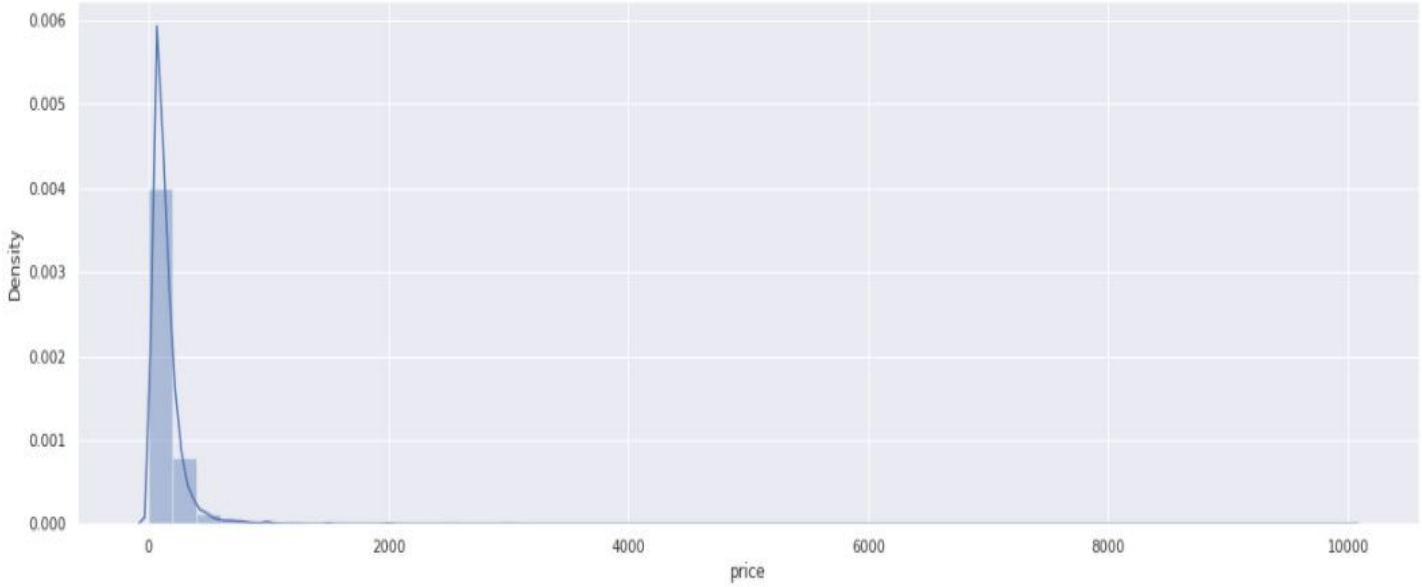
This San-Francisco based startup offers you someone's home as a place to stay instead of a hotel. You might be thinking of another unicorn in town as to OYO Hotels which has kind of a relatable business model but **Airbnb** allows you to be host for anyone anywhere with rooms/beds available in your personal space. OYO Rooms and Airbnb are by no means similar to each other, in fact, they are almost as opposite as the sky and sea. So, having much said let's just deep dive into our actuals on why are we basically here? Scroll below and have a feel!

This dataset has around **48,895** observations with 16 columns and it is a mix between categorical and numeric values. I have portrayed this detailed analysis as much simple as required to get a basic understanding even if someone is very new to this ;)

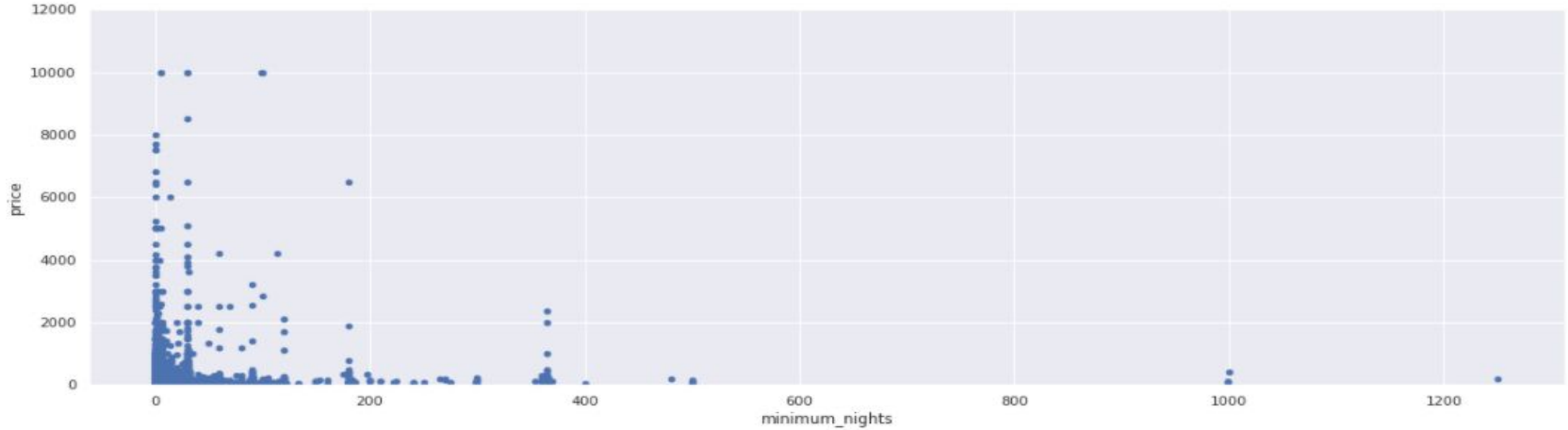
The very basic information about the dataset using df.info()
By basic inspection, a particular property name will have one particular host name hosted by that same individual but a particular host name can have multiple properties in an area. So, host_name is a **categorical variable** here. Also neighbourhood_group (comprising of Manhattan, Brooklyn, Queens, Bronx, Staten Island), neighbourhood and room_type (private,shared,Entire home/apt) fall into this category.
~id,latitude,longitude,price,minimum_nights,number_of_reviews,last_review, reviews_per_month, calculated_host_listings_count, availability_365 are **numerical variables**. By basic inspection, a particular property name will have one particular host name hosted by that same individual but a particular host name can have multiple properties in an area. So, host_name is a **categorical variable** here. Also neighbourhood_group (comprising of Manhattan, Brooklyn, Queens, Bronx, Staten Island), neighbourhood and room_type (private,shared,Entire home/apt) fall into this category.
~id,latitude,longitude,price,minimum_nights,number_of_reviews,last_review, reviews_per_month, calculated_host_listings_count, availability_365 are **numerical variables**.

```
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              48895 non-null  int64
 1   name                            48879 non-null  object
 2   host_id                         48895 non-null  int64
 3   host_name                       48874 non-null  object
 4   neighbourhood_group             48895 non-null  object
 5   neighbourhood                   48895 non-null  object
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object
 9   price                           48895 non-null  int64
 10  minimum_nights                  48895 non-null  int64
 11  number_of_reviews               48895 non-null  int64
 12  last_review                     38843 non-null  object
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64
 15  availability_365                48895 non-null  int64
dtypes: float64(3), int64(7), object(6)
```

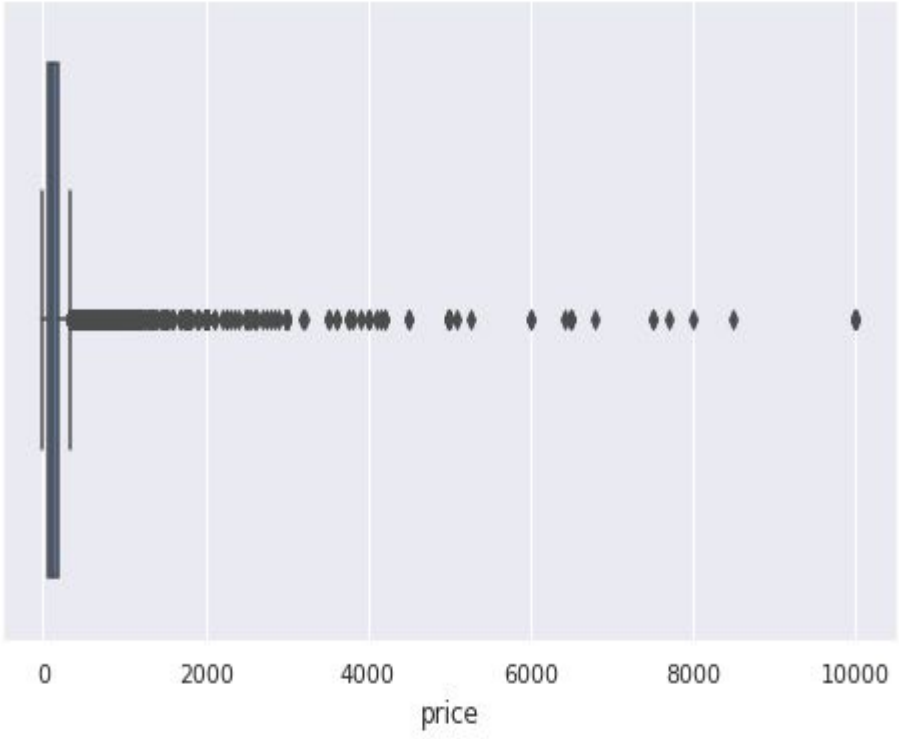
I have used seaborn **distplot** to plot this distribution curve. The distribution has a positively skewed tail at the very extreme as we can see. Also getting the **skewness** as 19.118939 and **kurtosis** to be 585.672879, *depicting the skewness value>1 and kurtosis is much high indicating presence of good amount of **outliers**, we will look later into this when we handle outliers!*



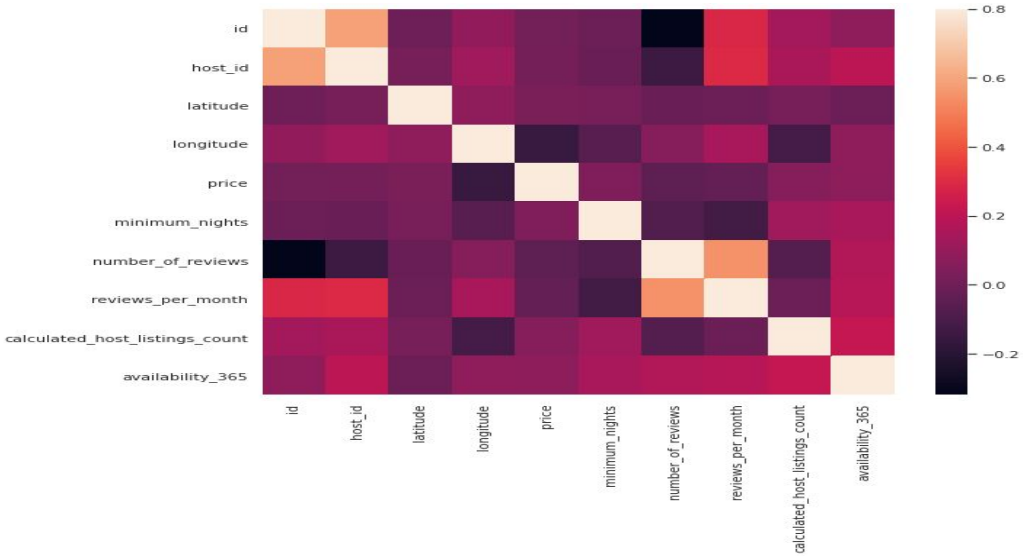
We'll be finding the relationship between these.There's an interesting lookout from the above scatter plot, what do you see?
many data points are clustured on 0 price range, few have min nights for stay but price is 0. looks like ***anomaly*** *in* ***price****.* two numerical variables using seaborn scatter plot as below:



Let’s see the boxplot of this **price** column to have a feel of the presence of outliers. Don’t worry, I’ll be handling these **outlier** values



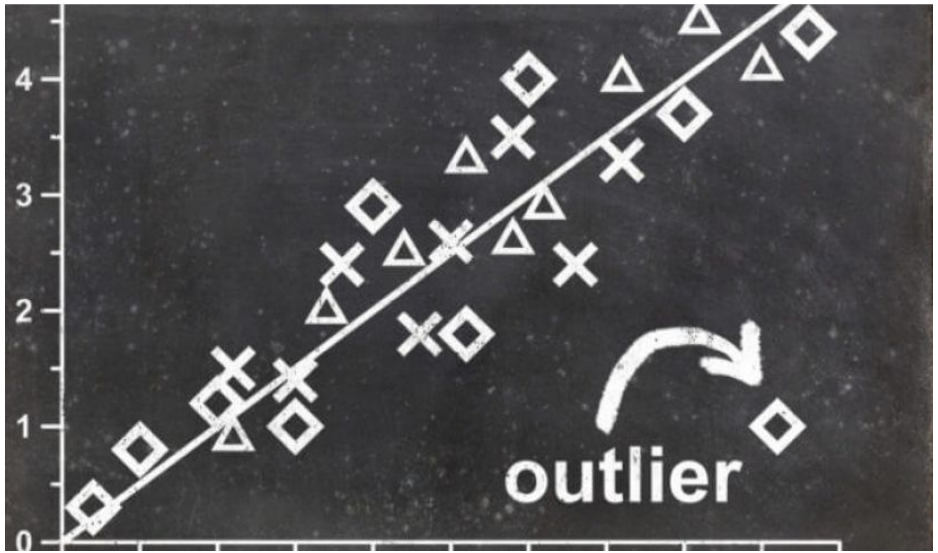
Also let's check the **correlation** matrix to understand how are the features interrelated with each other. I have plotted using seaborn heatmap to understand the *strength* between the variables used. this:**So, which are the most important features in this dataset from the heatmap?**
There's correlation among host_id to reveiws_per_month & availability_365 (sequential color bar is used between value and color). Also there's noticiable correlation between min_nights, no_of_listings_count & availability_365. Price also shows some correlation with availability_365 & host_listings_count.
no_of_reviews and reviews_per_month gives almost the same information. so we can carry out analysis with any of the two variable. Also, no_of_reviews is correlated to availability_365!



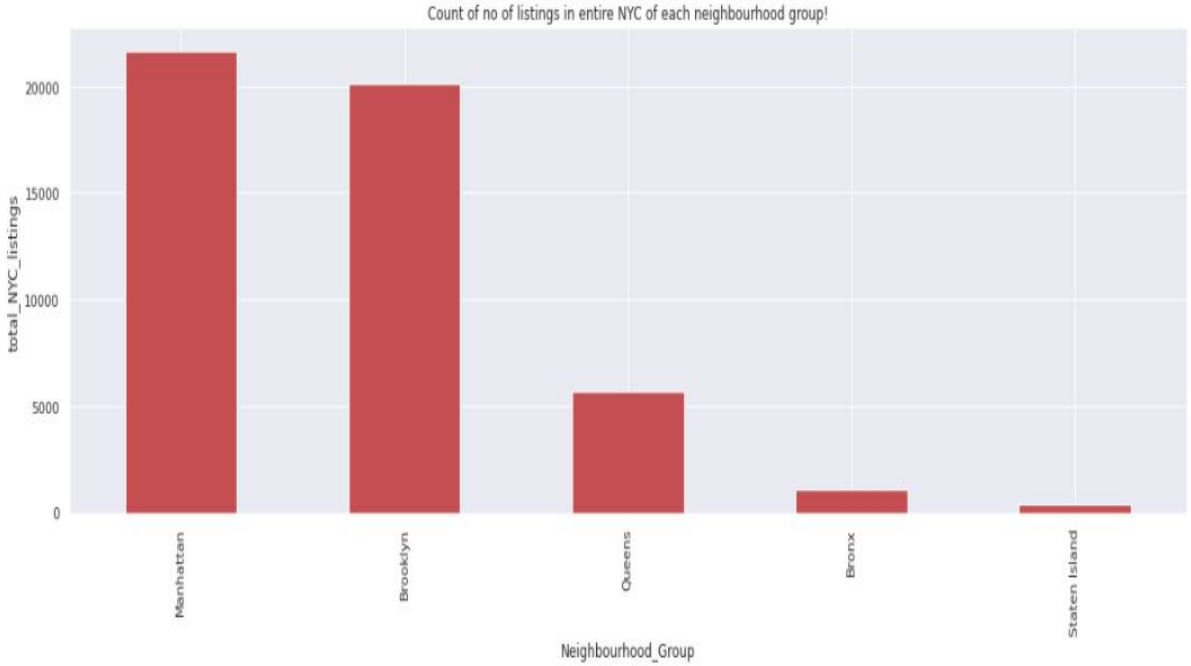
Let’s understand outliers and figure out some ways to deal with such anomalies in data

I have done most of the common **data pre-processing steps** like *missing values treatment*, *checking duplicate records* but here comes the very important part in EDA which many overlook before fitting a ml model is removing the outliers values as many machine learning algorithms do not support **missing values** and also making ml models robust to **outliers.**

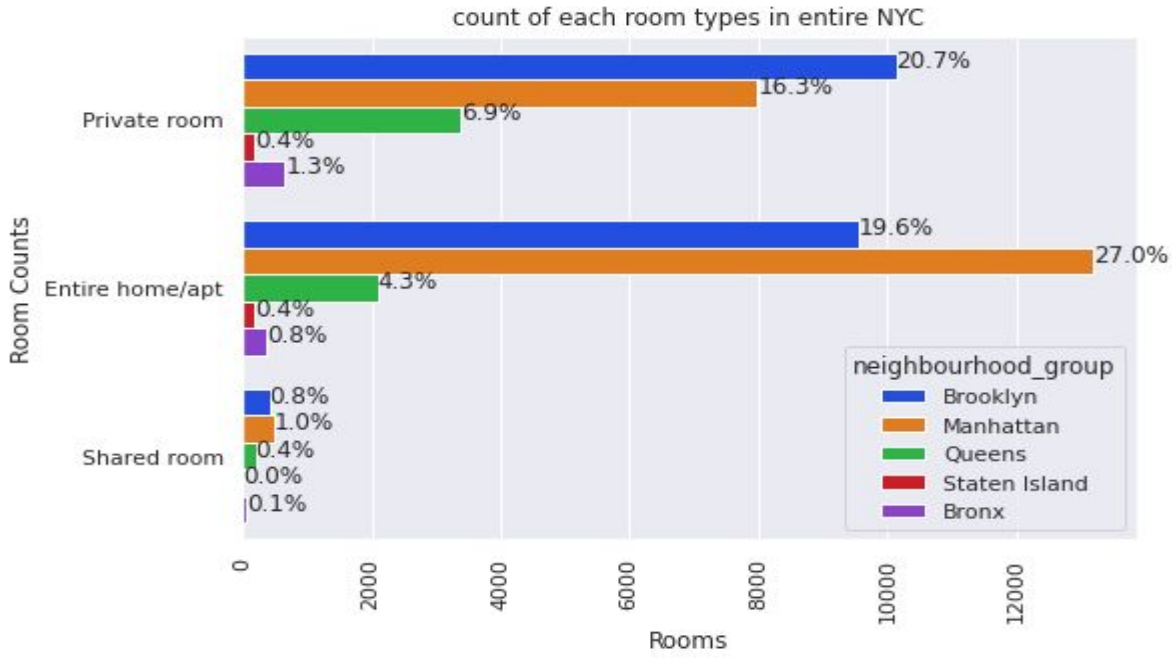
Well an **outlier** is a data point that lies outside the overall pattern in a distribution. Say we’re trying to understand the people’s income based on start and end of a project. We might measure income levels of our sample group at the start and the end. Imagine our results follow a linear distribution and looks like this:



Next, I'll be showing plots representing the count of Airbnb's in different *neighbourhood groups and neighbourhoods* of NewYork City. From the plot, we can easily visualize that maximum number of houses or apartments listed on Airbnb. Well, by now you have easily got to know the neighbourhood group having the highest no of listings in NYC.

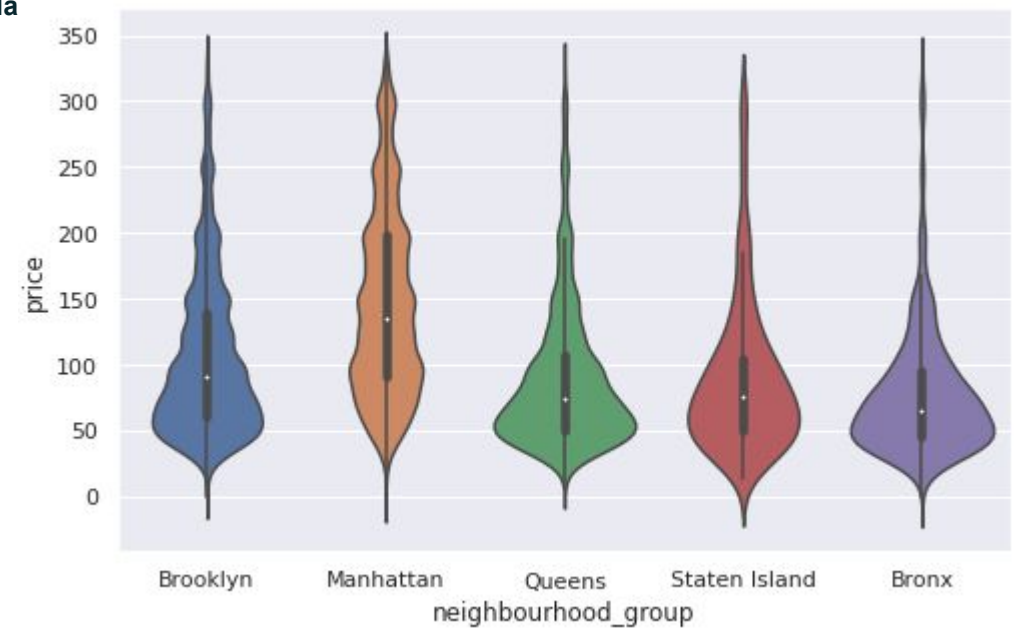


What to conclude from this?
Manhattan has more listed properties with **Entire home/apt** around 27% of total listed properties followed by **Brooklyn** with around 19.6%.
Private rooms are more in **Brooklyn** as in 20.7% of the total listed properties followed by Manhattan with 16.3% of them. While 6.9% of **private rooms** are from **Queens**.
Very few of the total listed have shared rooms listed on Airbnb where there's negligible or almost very rare **shared rooms** in **Staten Island** and **Bronx**.
We can infer that **Brooklyn, Queens, Bronx** has more **private room** types while **Manhattan** which has the highest no of listings in entire NYC has more **Entire home/apt room** types.

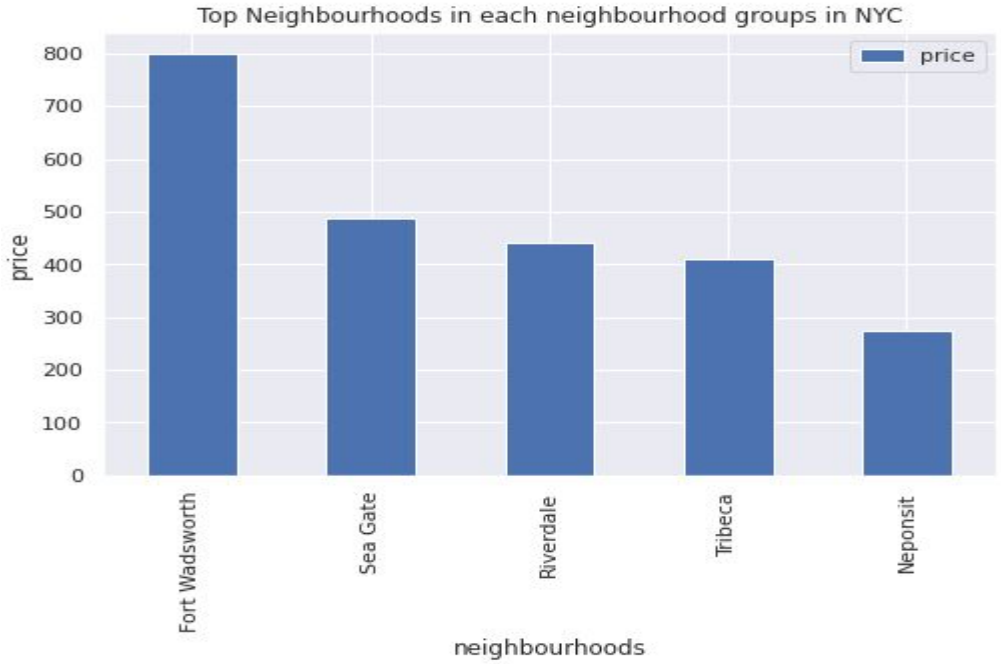


Lets now check for distribution of price across: Manhattan, Brooklyn, Queens, Bronx & Staten Island. Instead of checking distributions for each categories one by one we can simply do a violin plot for getting the overall statistics for each groups. But we'll get to know the median of price/neighbourhood group. As usual **Manhattan** being the most **costliest** place to live in, have price more than 140 USD followed by Brooklyn with around 80 USD on an average for the listings.

Queens, Staten Isla

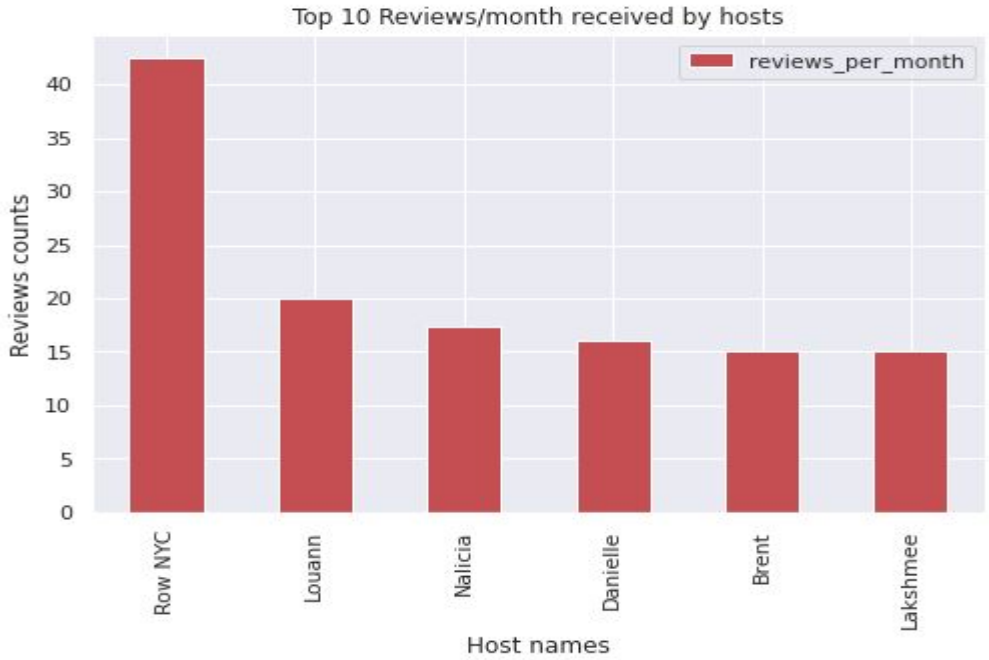


The bar plot above clearly depicts the neighbourhoods with listings having highest average price/day in each neighbourhood groups of NYC. Among the top neighbourhoods in each neighbourhood groups, top 2 of them namely: **Fort Wadsworth** & **Sea Gate**, origins from **Staten Island** & **Brooklyn** respectively.

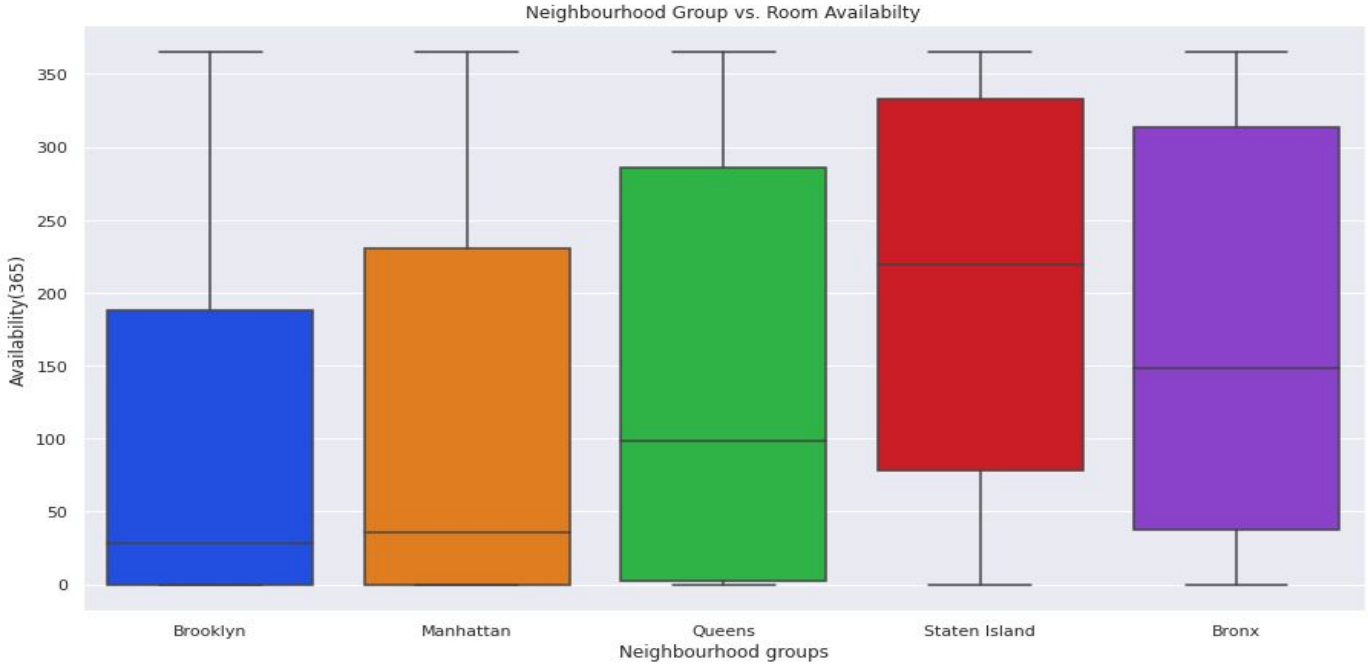


It clearly indicates that people mostly prefer living in an **entire home/apt** on an average of more than **8 nights** followed by guests who stayed in **shared room** where average stay is **6–7 nights**.

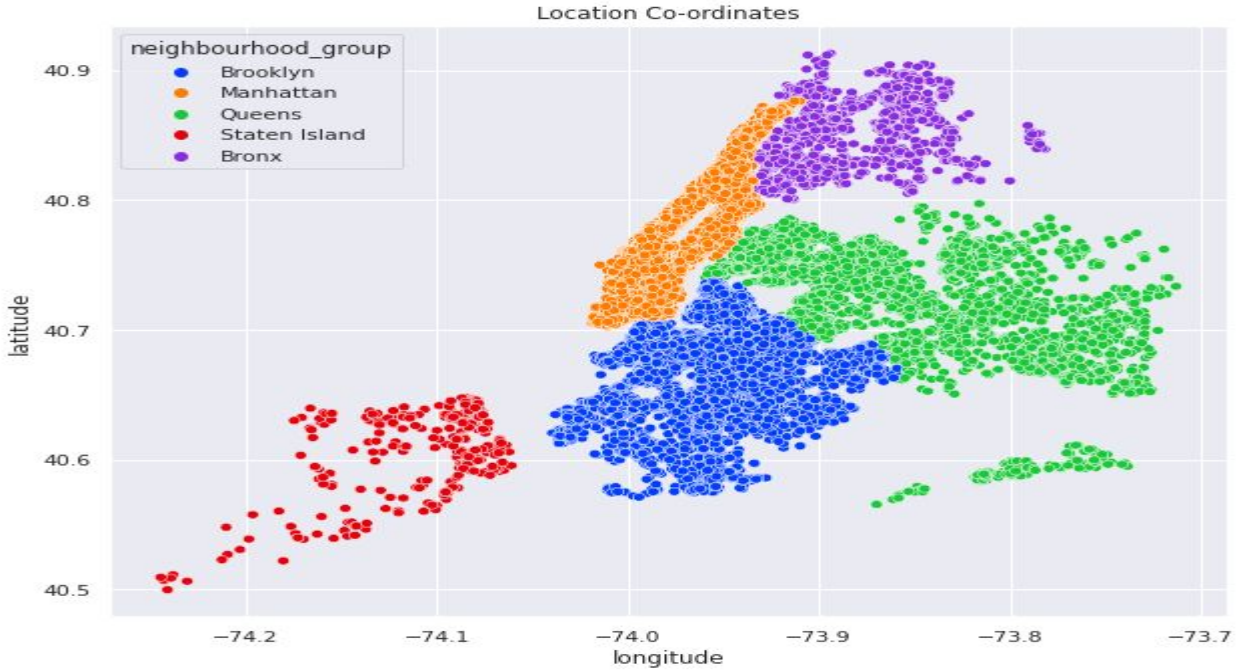
Row NYC holds the title as the most reviewed host with more than 40 reviews/month on average.



Looking at the 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.



Let’s also see what can be done with latitude and longitude?

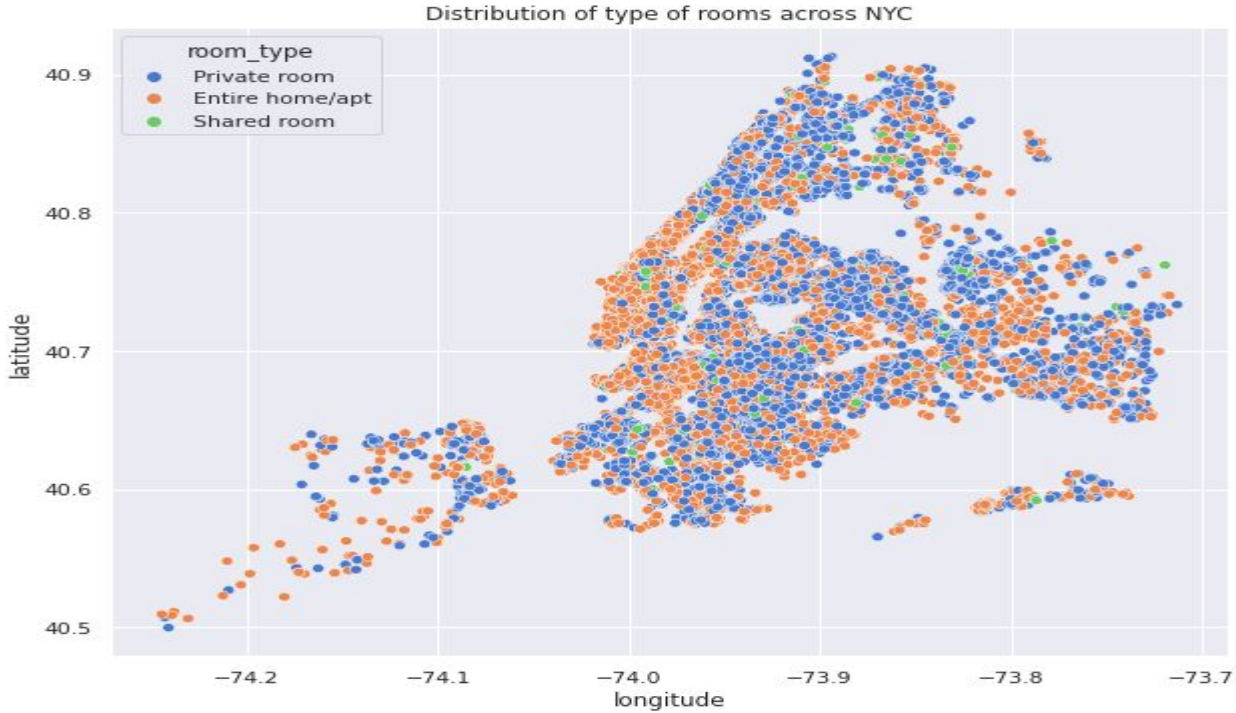


Now, let's check for the distribution of types of rooms across all neighbourhood groups of NYC!

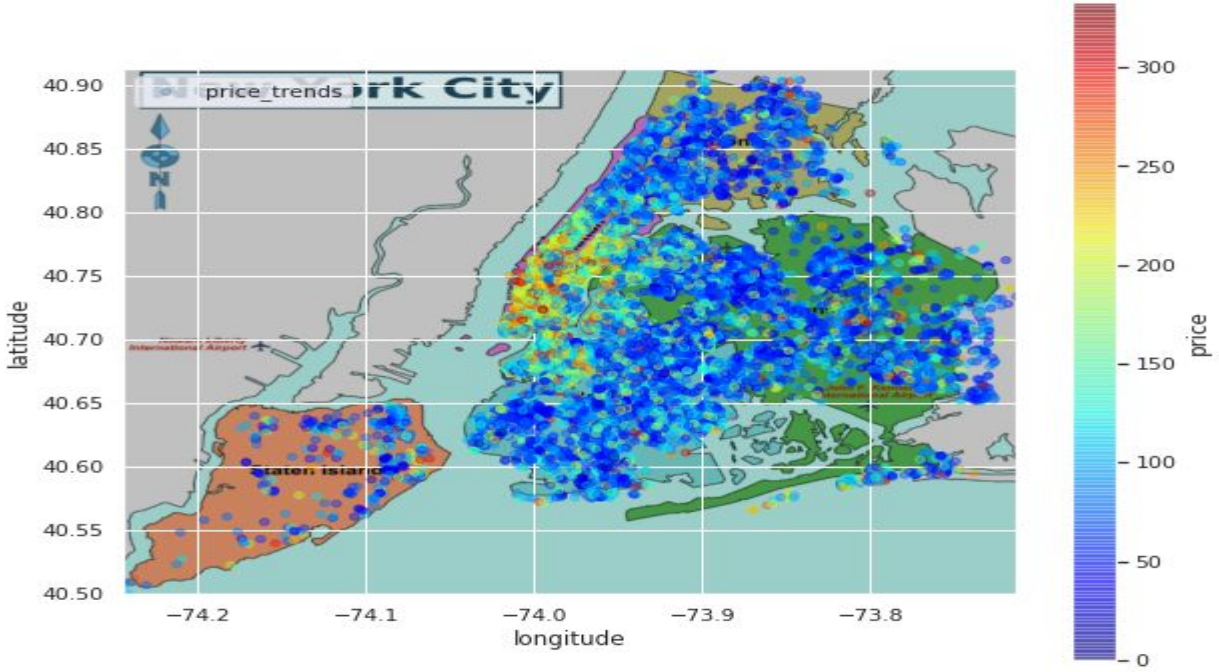
By the two scatterplots of latitude vs longitude we can infer there's is very less **shared room** throughout NYC as compared to private and Entire home/apt.

95% of the listings on Airbnb are either **Private room** or **Entire/home apt**. Very few guests had opted for shared rooms on Airbnb.

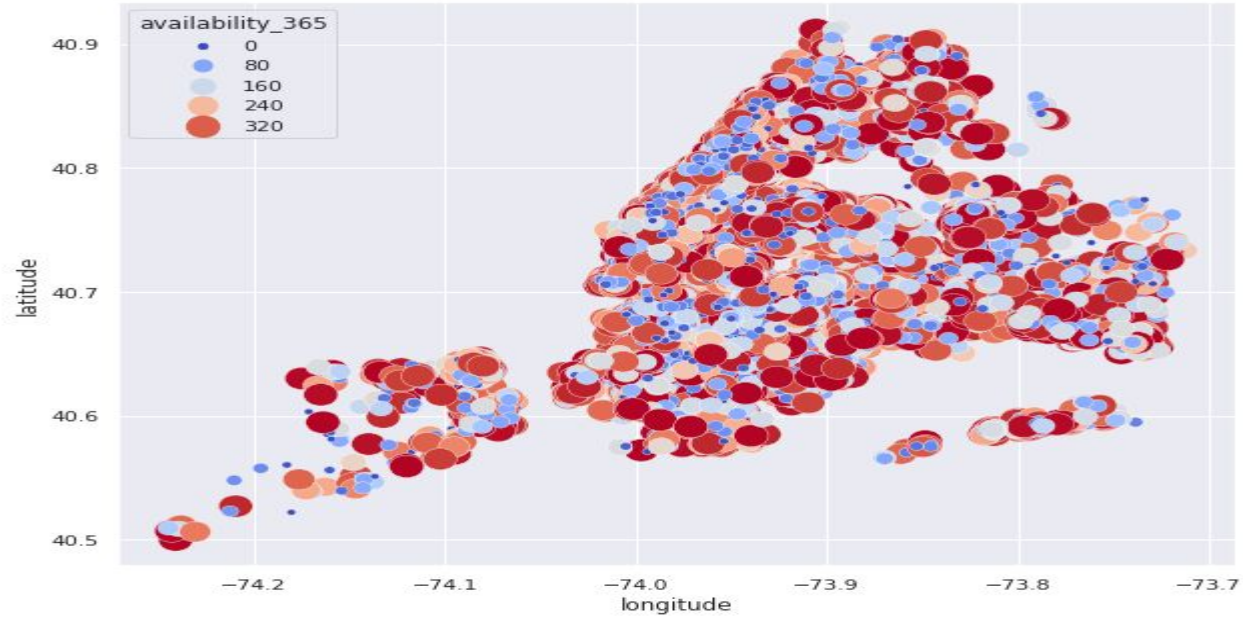
Also, guests mostly prefer this room types when they are looking for a rent on Airbnb as we found out previously in our analysis.



The scatterplot showing the price variables across these co-ordinates in a more authentic way using the original NYC boroughs map by saving the original map image in my local directory and then reading the image using cv2 imread function.
We can infer that there are high range of prices across **Manhattan** followed by **Brooklyn and Queens** being the most costliest place to stay in NYC.



I’ve plotted the scatterplot depicting the availability of listings available throughout NYC in a year. I have used hues with different sizes based on the availability ranges.
Bronx & **Staten Island** has listings which are **mostly available** throughout the year, might be the case as they are not much costlier as compared to other boroughs as in Manhattan, Brooklyn & Queens.



End notes

Through this exploratory data analysis and visualization, we gained several interesting insights into the Airbnb rental market. This Airbnb dataset for 2019 year appeared to be a very rich dataset with a variety of columns that allowed us to do deep data exploration on each significant column presented. After that, we proceeded with analyzing boroughs and neighborhood listing densities and what areas were more popular than another, their price variations, their availability as per room types. Also we emphasized on key findings like room types and their preferred stays by guests, the top reviewed hosts and their listings.

Next, we put good use of latitude and longitude columns to create a geographical heatmap color-coded by the price of listings

I have used Seaborn and Matplotlib for creating all the visualizations. This is just a glimpse of eda on the airbnb dataset and there's no any predictions involved.