EDA Airbnb Bookings Analysis - Capstone Project

Gaurav Jindal

Abstract:

This is the final technical report of our data analytic project titled "Airbnb Bookings Analysis - Capstone Project" as a part of our Data analytic course at Alma better. The goal is to analyze and predict the price and other variables in the New York Airbnb data. Also, a recommendation system will be built to recommend Airbnb listings according to the user preference.

Content:

Airbnb (ABNB) is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific locales. The company has come a long way since 2007, when its co-founders first came up with the idea to invite paying guests to sleep on an air mattress in their living room. According to Airbnb's latest data, it has in excess of six million listings, covering more than 100,000 cities and towns and 220-plus countries worldwide.

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became one of a kind service that is used and recognized by the whole world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data - data that can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.
This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

Problem Statement:

In this project we are analyzing Airbnb's New York City(NYC) data of 2019. Our main objective is to explore and analyze the data to discover key understandings about listing of properties on the platform. We will perform basic Exploratory Data Analysis(EDA). We will be find out key metrics that influence every Airbnb listing based on their location, different hosts and areas, prices, reviews, room type, listing name, traffic and other related factors.

Approach:

- 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 inorder to get a better result.
- Exploratory Data Analysis:
  After treating null values and duplicate columns we now performed basic EDA and observed the results through visualization and concluded the the results.
- Conclusion:
  After EDA we wrote the conclusions of our observations.

Dataset info:
Number of Columns:16 , Number of Samples:48895, Number of quantitative variables:10, Number of qualitative variables:6,

Attributes or Columns:
id, name,
host_id, host_name,
neighborhood-group,
neighborhood,
latitude, longitude,
room type, price,
minimum nights,
number_of_reviews,
last_review, reviews_per_month,
calculated_host_listings_count,
availability_365

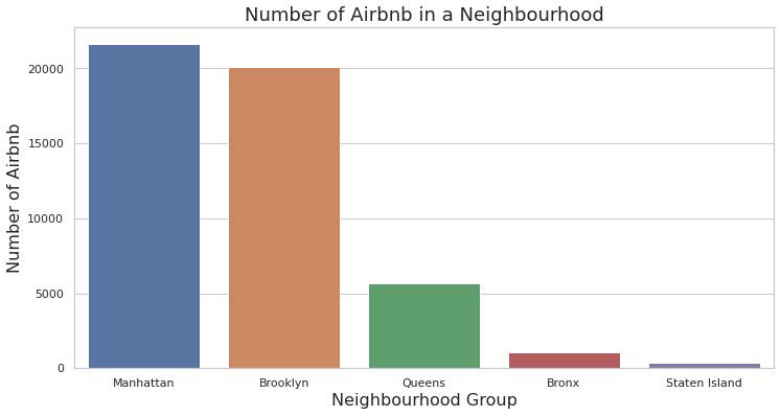
Since our dataset's contain several missing values preprocessing must be done. Missing values will either be deleted or replaced with the column mean or nan. based on how important the attribute is. Also, with respect to preprocessing the datatype of certain attributes like last_review must be changed to make processing easier. Our main goal is to analyze and find interesting patterns between the variables in our dataset's. Visualization is an important aspect of finding patterns. Hence several visualization techniques like bar graph, pie chart, Violin chart, correlation, etc. will be plotted to gain insights. We then plan to predict certain variables such as price by using predictive models. Several models will be explored and models with the best accuracy will be selected. We also got that

-Few columns like name, host name, last review had many missing values and then we replaced it with "missing". Importance for analysis, hence they were deleted.

-Reviews per month column had lot of missing rows but is important for analysis, hence missing values were replaced with the mean of that column.

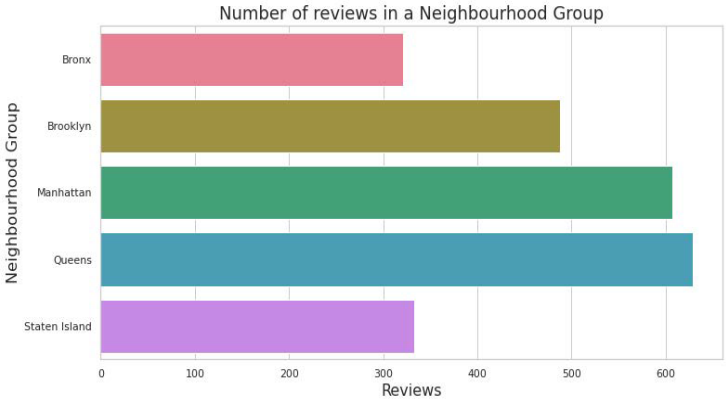
Number of listed Airbnb properties in different neighbourhood groups:

A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle and provides some indication of the uncertainty around that estimate using error bars. Bar plots include 0 in the quantitative axis range, and they are a good choice when 0 is a meaningful value for the quantitative variable, and you want to make comparisons against it.

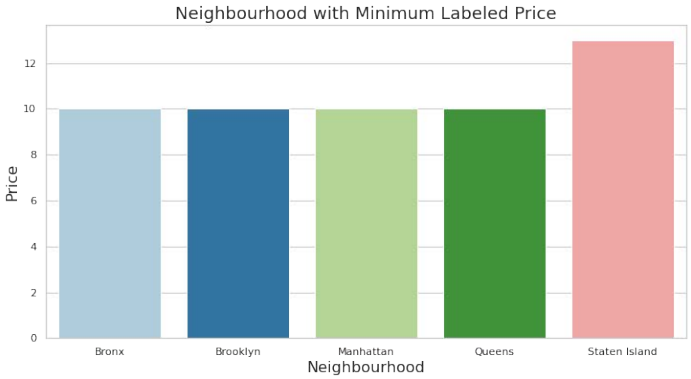


Plotting a bar plot for count of number of Airbnb in all Neighbourhood group areas

Number of reviews for listed Airbnb propertiesin different neighbourhood groups:

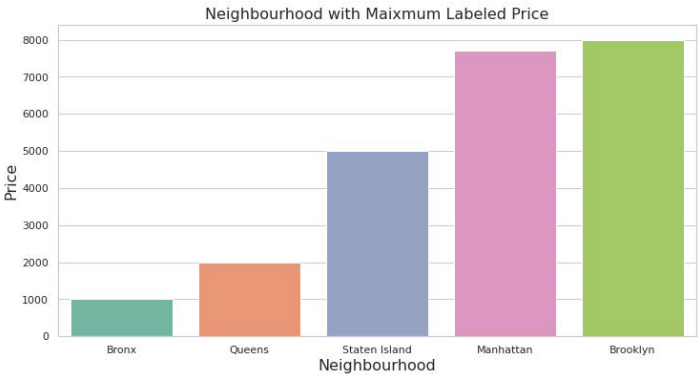


Plotting a bar plot for count of number of reviews in each Neighbourhood Group



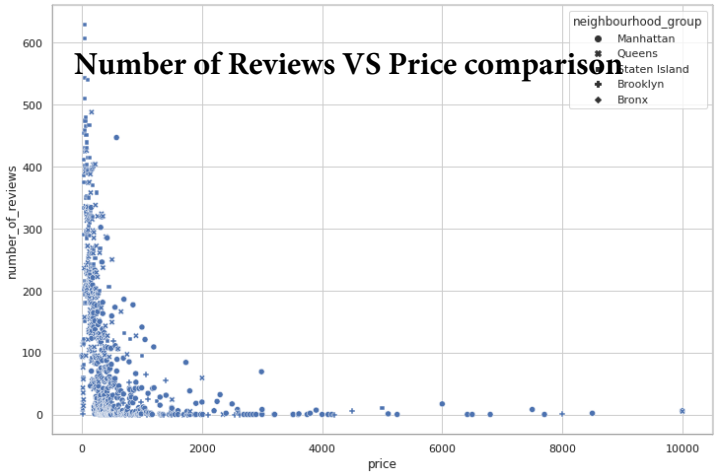
Plotting a bar plot for minimum price in each Neighbourhood Group

Maximum and Minimum Price of Airbnb properties in different neighbourhood groups



Plotting a bar plot for maximum price in each Neighbourhood Group

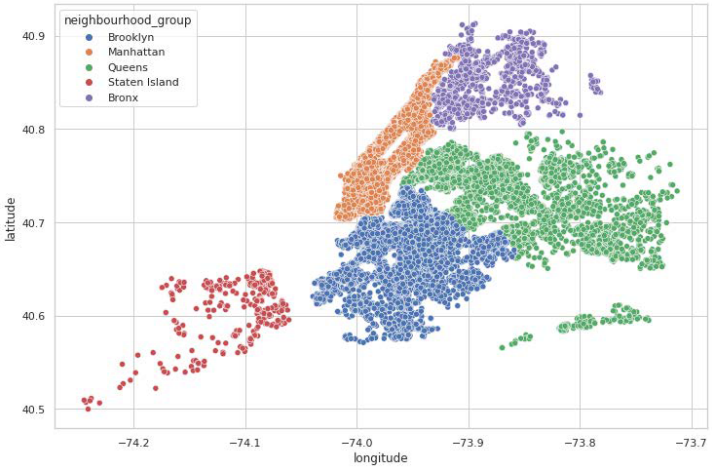
Number of Reviews VS Price comparison

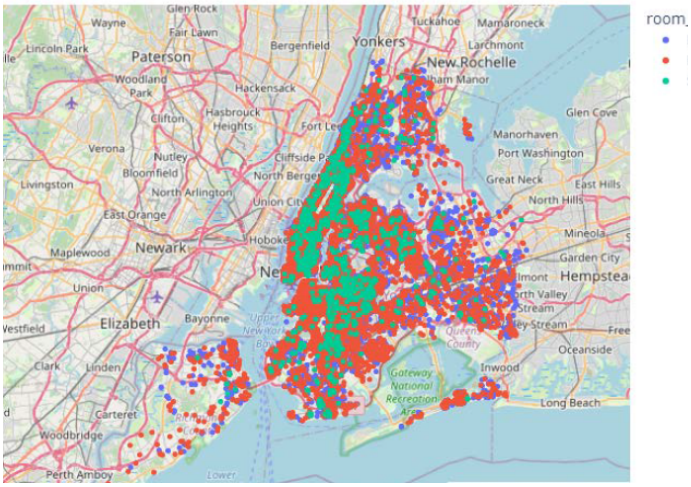


Plotting a scatter plot for comparison of number of reviews with the price

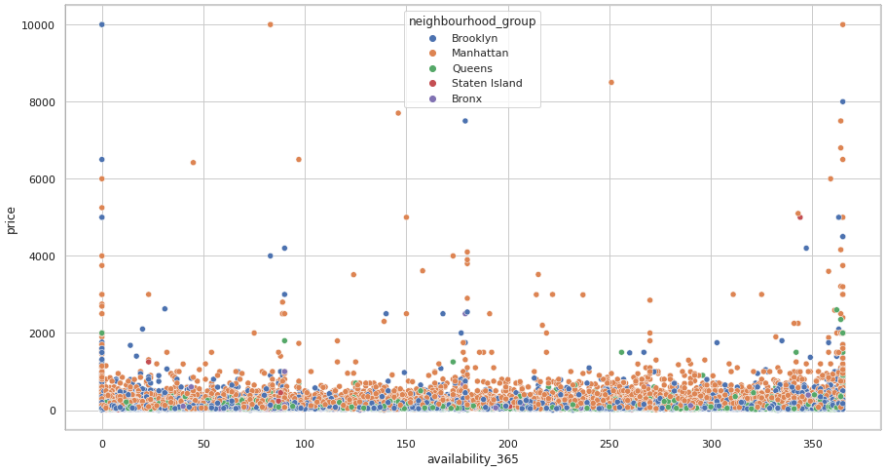
Locations of listed Airbnb properties in different neighbouhood groups of NYC:

Scatterplot can be used with several semantic groupings which can help to understand well in a graph. They can plot two-dimensional graphics that can be enhanced by mapping up to three additional variables while using the semantics of hue, size, and style parameters. All the parameter control visual semantic which are used to identify the different subsets. Using redundant semantics can be helpful for making graphics more accessible.



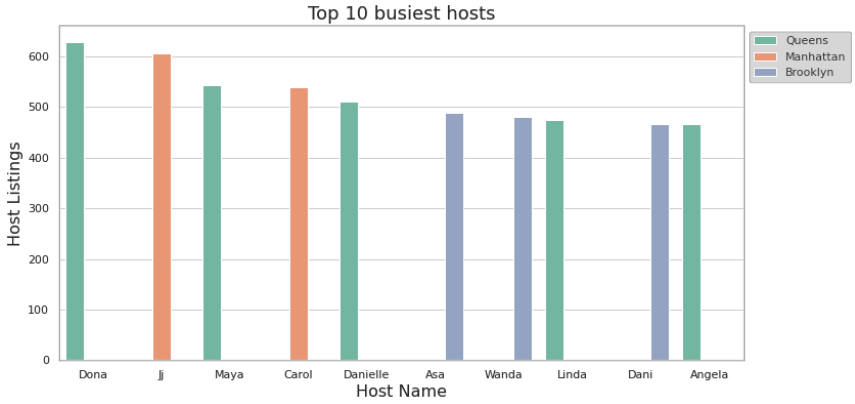


Price VS Availability in a year in different neighbourhood groups comparison



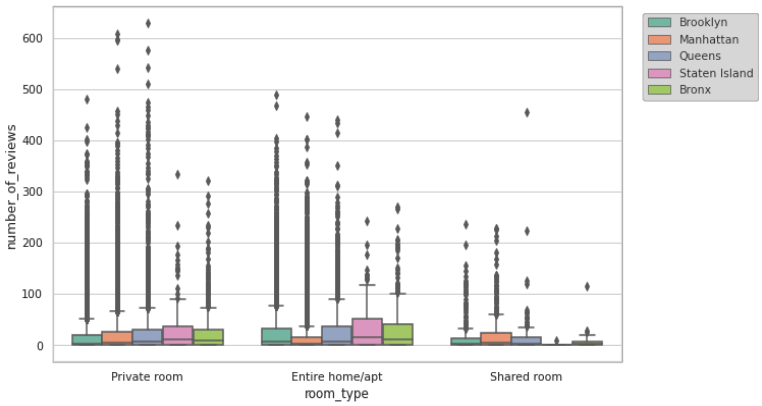
Plotting a scatter plot for Comparison of Availibity of Airbnb in a year with the Price for different neighbourhood groups

Busiest Hosts among the listed

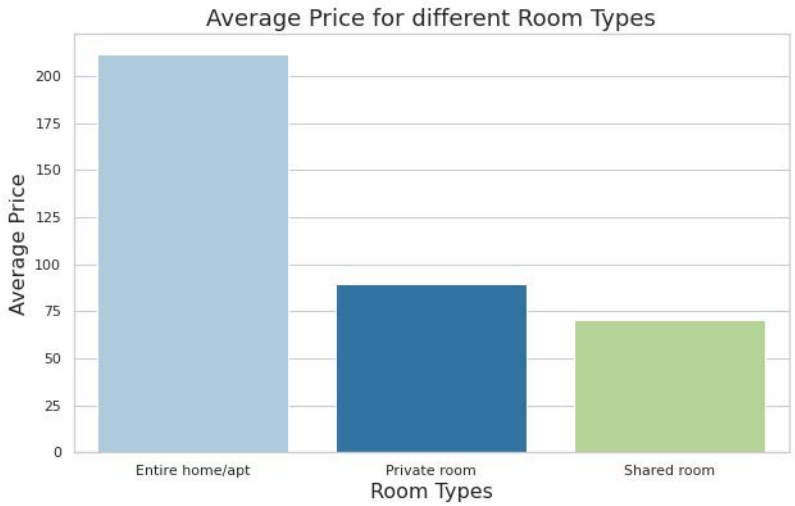


We plotted a bar chart for better understanding and visualization using top_busiest_host dataframe.

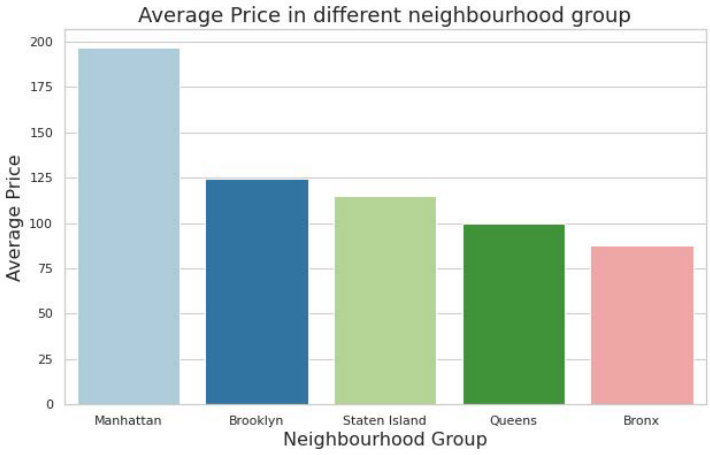
Number of Reviews VS Room Type Comparison



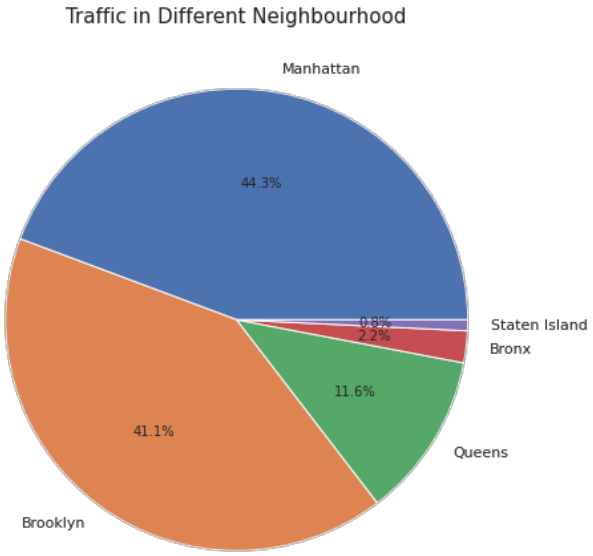
A box plot (or box-and-whisker plot) shows the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable. The box shows the quartiles of the dataset while the whiskers extend to show the rest of the distribution, except for points that are determined to be "outliers" using a method that is a function of the inter-quartile range.



Average Price of listed Airbnb properties in different neighbourhood groups and for different Room types

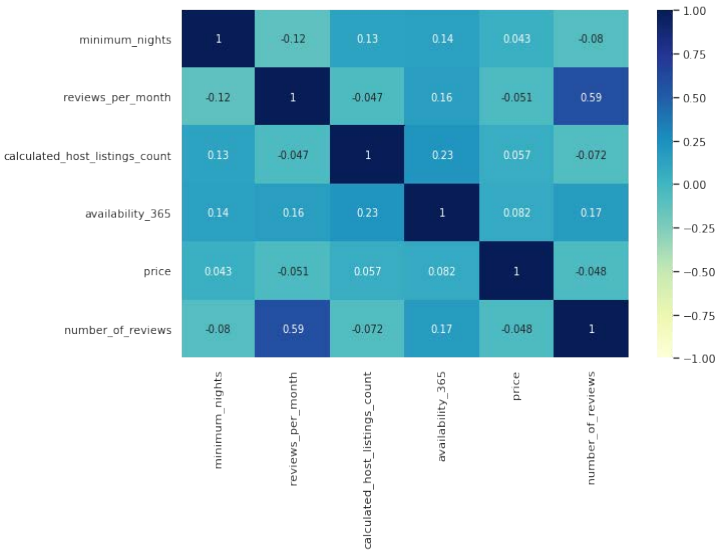


Traffic Details in different neighbourhood groups



We can observe that Manhattan is the most Traffic area since Manhattan has the most number of minimum night stays

Correlation check



A heatmap contains values representing various shades of the same colour for each value to be plotted. Usually the darker shades of the chart represent higher values than the lighter shade. For a very different value a completely different colour can also be used.

Conclusion

In this EDA project, different use cases are analysed for the
given dataset to make better business decisions and help analyse
customer trends and satisfaction, which can lead to new and
better products and services.
Few key points:

- Manhattan is the most focused place in New York for hosts to do their business
- Host Sonder(NYC) have the most listings and these listings are in Manhattan area.
- Manhattan has the most expensive Airbnb properties
- Dona is busiest Host and this is in Queens, followed by Jj in Manhattan and so on.
- Manhattan has the most number of minimumm night stayed, so Manhattan is the most Traffic area.

References-

Almabetter Notes
GeeksforGeeks