# Sinduja Sekar

# 7 November 2019

AirBNB accomodations in NewYork city

## Introduction

The purpose of this project is to find the best airBNB accomodation around New York city.The dataset for this project is carefully chosen so that it includes the right number of features and enough data to perform exploration and modelling on the data. The aim of finding the best accommodation depends on various features like the neighbourhood , price, proximity to famous sites and the city and reviews from previous visitors. The distance of the accommodation from the city’s centre is not part of the given dataset and would be obtained from the foursquare API and merged into the existing dataset. The latitude and longitude data in the existing dataset would be used to merge the dataset with the foursquare API results. The foursquare API can also be used to further explore the venues for example, the API can be used to check for restaurants, cinema and parks around for recreation during the stay. The insights and results from this project would benefit travellers in making better decisions in choosing the airbnb accomodation and would aid in getting the accomodation that fits all their criteria.

## Data Acquisition and cleaning

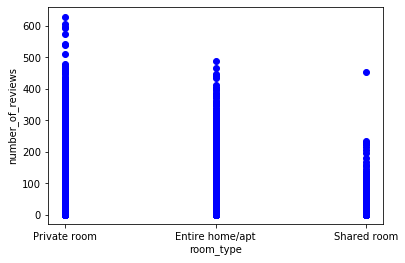
The dataset for this project is obtained from the kaggle dataset and contains 48000 samples and 16 features. After examining the shape and description of the data, and dropping unwanted features, I have checked the dataset for null values. The neighbourhood features contains some null values which can be handled by imputing with the right values that can be obtained from another dataset using the latitude and longitude values. The feature neighbourhood\_group is renamed as borough for better understanding and comprehension. The data for the reviews per month column is ceiled and converted to a integer value from a floating point value and the missing data for the review column are less than 50% of the rows and deleted if they are updated recently. Further exploratory analysis can be conducted own this cleaned dataset.

## Feature Selection

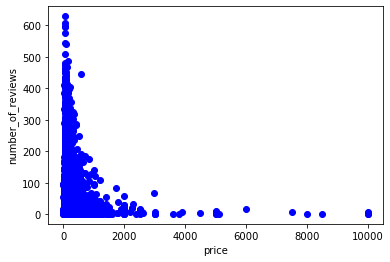
After handling the null values and cleaning up the data, the dataset is carefully examined to select the features on which we depend to find insights about our data. In this case unnecessary and duplicate columns like last\_reviews and reviews\_per\_month are dropped. The neighbourhood group is renamed to borough to improve the readability of the data. The reviews\_per\_month data is rounded and converted to int to make it more meaningful. The altitude and longitude of the city centre is obtained using the four square API and the difference between the latitude and longitude of the BNB locations and the city centre is calculated.

Exploratory Data Analysis

Further exploratory data analysis is performed by studying the relationship between the various features in the data set.

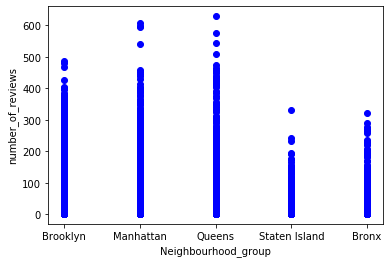


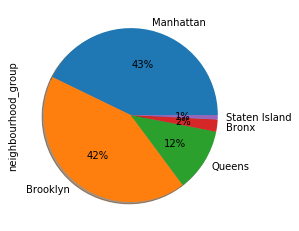
Relationship between the number of reviews and the type of rooms



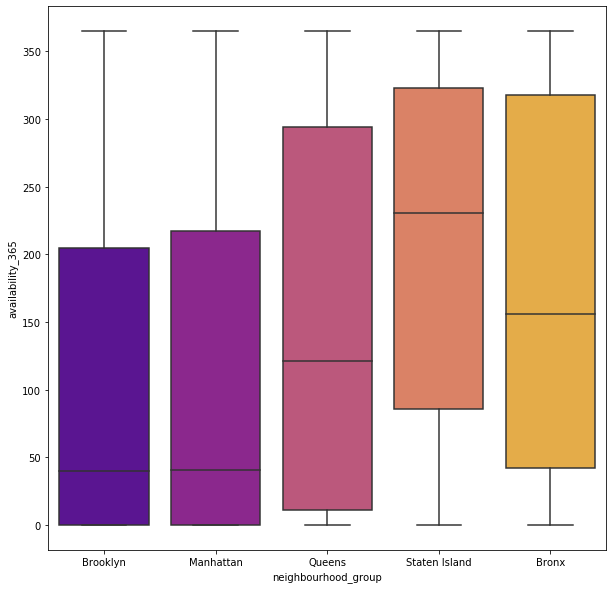
The relationship between the price and number of reviews helps us predict the price range preferred by consumers.

The following scatter plot and pie digram enable us to better represent the distribution of the Airbnb accomodations in the neighbourhood groups.

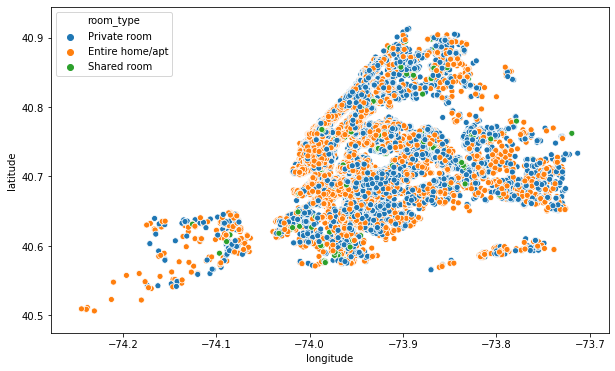




Room availability in each of the neighbourhood groups can be represented as a box plot



The following scatter plot describes the room types based on the neighbourhood group.



## Applying regression models and conclusion

I applied linear regression to calculate and predict the r squared mean and the decision tree to predict the price for the airbnb and also compared and correlated relationships between various factors that constitute towards selecting the best option for the BNB accommodation.