New York City Airbnb Data Analysis

Instructor: Prof. Ying Lin
Team Members: Aniruddh Garge, Jeet Ganatra, Tanushree Shetty

1. Problem and Objective

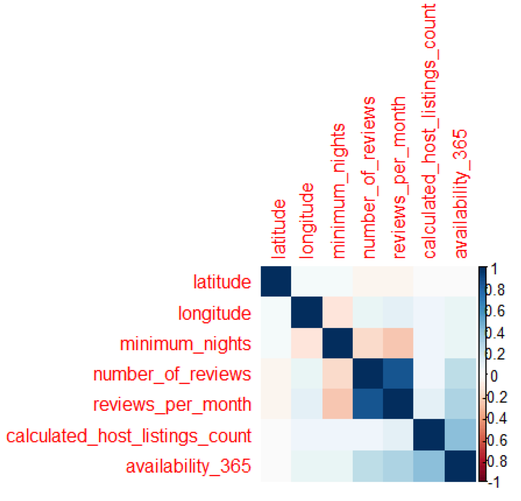
Since 2008, guests and hosts have used Airbnb to expand on travelling possibilities and present more unique, personalized way of experiencing the world. Airbnb is an online marketplace for arranging or offering lodging, primarily homestays or tourism experiences. Today, Airbnb became one of a kind service 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.

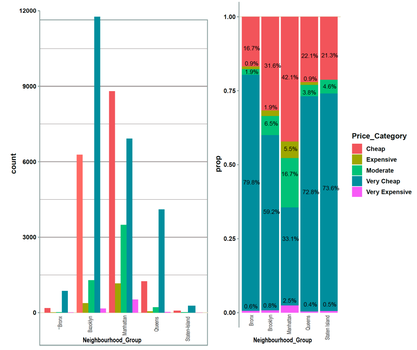
2. Goals

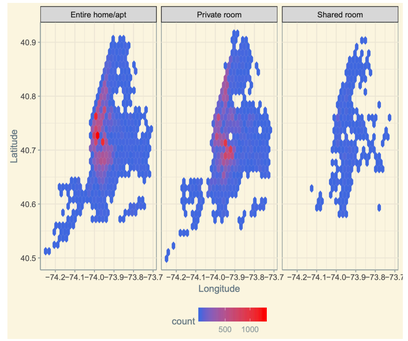
To predict the price-category each Airbnb room falls under, based on the NYC neighborhood (boroughs), location, room-type and reviews so that customers can have a better experience in selecting the type of rooms they want and fits their budget.

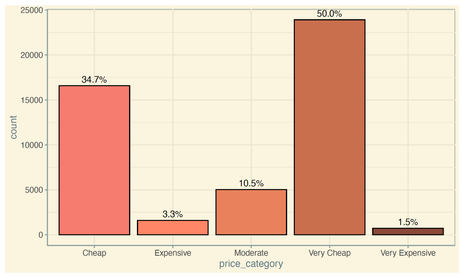
3. Data Description

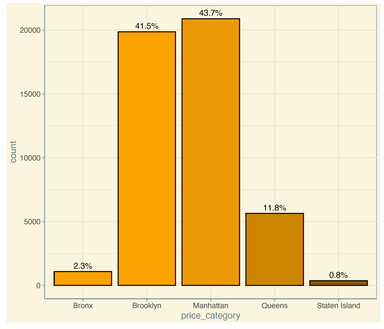
This dataset describes the listing activities and metrics in NYC, NY for 2019. The data has 48,895 entries and 16 total columns and is a mix of numerical and categorical values. The dataset has information on the borough, arears within the borough, location, type of room and reviews of Airbnb listings.











4. Model Description

Random Forrest:
It is an aggregation of multiple decision trees which are cumulatively used to classify the data we used. It generates several decision trees and performs majority voting in order to predict the price categories of the listings.

Decision Trees:
This algorithm is derived from the traditional tree structure. It generates trees based on a specific information measure and decides of predicting the price category.

SVM:
Forms support vector from the data in order to separate all the listed classes from each other. Uses decision boundaries in order to make the final prediction.

We have tried to accommodate the upper listed models in our data analysis and compared the respective accuracies which are listed in the table.

Model	Features	Accuracy
Decision Tree	Neighborhood, Location,room-type, Reviews, Availability	70%
Random Forest	Neighborhood, Location,room-type, Reviews, Availability	60.84%
SVM	Neighborhood, Location,room-type, Reviews, Availability	69.04%

7. Conclusion

This Airbnb dataset for the 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. First, we have found hosts that take good advantage of the Airbnb platform and provide the most listings. Machine learning is used extensively by Airbnb to enhance user experience, increase customer satisfaction and improve their product. Our analysis on the Airbnb data using ML models like SVM, Decision Trees and Random Forrest, predicts The price-category (Very cheap, Cheap, Moderate, Expensive., Very expensive) in which the listing will fall, based on the borough, neighborhood., location, reviews and availability. Overall, we discovered a very good number of interesting relationships between features and explained each step of the process. This data analytics is very much mimicked on a higher level on Airbnb Data/Machine Learning team for better business decisions, control over the platform, marketing initiatives, implementation of new features and much more. From our test run, the insights we found are:

1. Manhattan and Brooklyn boroughs have the most expensive room listings in NYC
2. Almost all rooms in the Financial District, Manhattan have expensive rooms.

7. Reference

https://community.rstudio.com/c/shiny
https://shiny.rstudio.com/images/shiny-cheatsheet.pdf
https://www.kaggle.com/jvpeluso/nyc-airbnb-open-data

Data Source: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data