A Visualization Study on New York City Airbnb Pricing

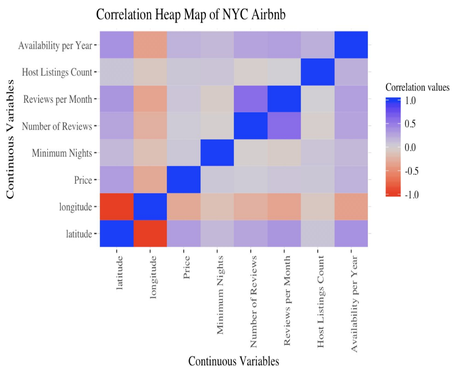
Author: Xi Chen, Serena Gao, Junning Gu, Jingyuan Xing **Project Supervisor:** Larry Wasserman

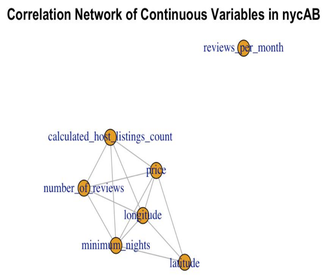
Background and Introduction

Since 2008, guests and hosts have used Airbnb to explore more travel possibilities, unique cultural experience and convenient way to earn additional income. As one of the most popular tourist destinations, a good understanding of New York City Airbnb pricing can provide people with better housing choices. 2019 New York City airbnb pricing data includes following variables: host id, listing name, host name, neighbourhood group, neighbourhood, latitude, longitude, toom type, minimum nights, number of reviews, last review date, reviews per month, host listings count, availability per year and price. **The goal of this project is to investigate how different factors affect daily listing prices.**

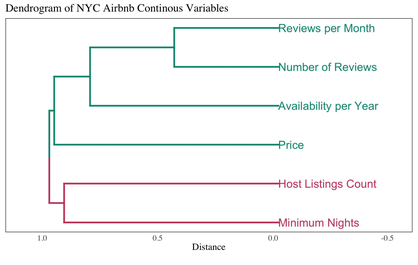
Data Preprocessing

The distribution for the price variable is unimodal, with a huge peak at the lower end and a positive skew with a long tail extending to the right hand side. The distribution has a mean of 152 and a median of 106. It has multiple outliers, ranging from 2500 dollars to 10000 dollars per day.





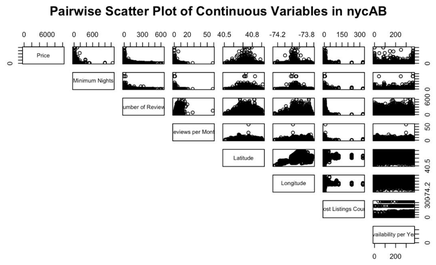
The Correlation Heatmap shows that price and latitude and availability per year are positively correlated. Latitude and longitude, longitude , review per month and price are negatively correlated. And price and minimum nights, price and number of reviews are almost not correlated. The results agree with the correlation network. Review per month is left out in correlation network because our significance level is high.



The dendrogram of continuous variables in the dataset shows 2 clusters of the variables. Reviews per Month, Number of Reviews and Availability per Year are more correlated with Price.

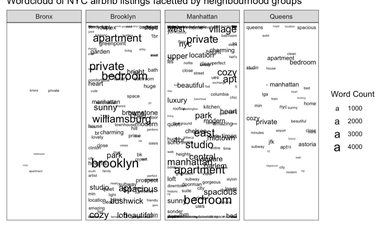
Analysis and Results

- **Pairwise Scatter Plots**



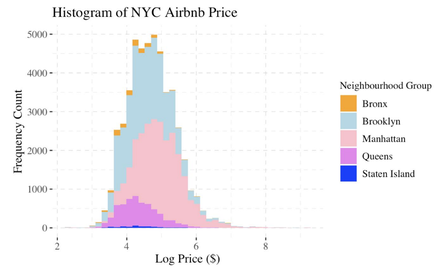
In the pairwise scatter plot, we mainly focus on first row of plots. We can see that minimum nights, number of reviews, reviews per month, host listing counts all have a negative correlation with price(when price are higher, positive when price are lower). While availability per year, longitude and latitude seem to be randomly distributed.

- **Word Cloud of the listings**



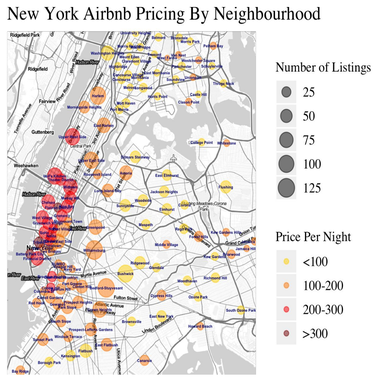
From the word cloud, we can deduce that listings in NYC, with the majority coming from Manhattan, Brooklyn and Queens, are mostly private bedrooms in apartments. The descriptions use the word “Cozy” a lot.

- **Stacked Histogram of Price Distribution by Neighbourhood**



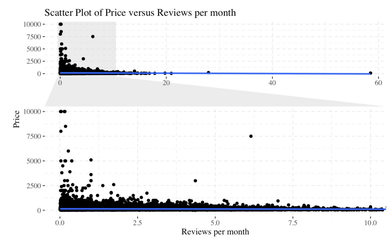
The stacked histogram of log price by neighbourhood shows us that Bronx, Staten Island and Queens has the lowest prices; Brooklyn is in between; and Manhattan is the most expensive neighbourhood.

- **Relationships between Price and Neighbourhood**



Comparing the dot sizes, Harlem, Upper West Side, Upper East Side, Hell’s Kitchen, Bedford Stuyvesant and Williamsburg have most listings. Comparing the color, Tribeca, Battery Park City and Flatiron District are the most expensive neighbourhoods.

- **Relationship between Price and Number of Reviews Per Month**



From this plot, for reviews per month from 0 to 10, the price is roughly negatively related to number of reviews per month. For reviews per month more than 10, price stays stable at low values. The linear regression does not capture the relationship well for lower values of reviews per month (0-10).

- **Stacked Histogram of Price Distribution by Room Type**

The stacked histogram of log price by room type shows that shared room is usually the cheapest type; entire home/apartment is usually the most expensive; private room is priced in between.

Conclusion

There are mainly two perspectives from which guests and hosts can gauge the price of a listing - categorical features and numerical features. From categorical perspective, location has a big impact on prices that lower manhattan and river side brooklyn area have highest price range. Room type also affect listing prices in that share rooms are cheaper while private rooms are more expensive. As for numerical feature, the most pronounced correlation is the negative correlation between price and reviews per month.