Milestone Report

Problem Statement

Airbnb is an online marketplace that allows owners to offer their properties, primarily homestays, to others at a price. This marketplace is used in over 65,000 cities and generates 500,000 stays per night. The main inspiration for the dataset that will be used for this project was to see what hosts generate the most traffic and why. This correlates with the problem that I would like to address through this capstone project. One problem that can come up with Airbnb is for new users attempting to list their properties. By being able to predict a price for certain accommodations based on their attributes, users will be able to list their properties at a reasonable price as well as maximize the profits of both the owner and Airbnb.

Target Client

The target client for this project would be those who are looking to list their properties on the Airbnb app or website. By predicting the prices of certain accommodations based on their attributes, the owner will be able to know what price he or she should list their property at. Airbnb can also use this study to see which owners will do well and come up with how much the company can make.

Dataset

The data that will be used for this project will be the New York City Airbnb Open Data dataset that is publicly accessible through the Kaggle website. The information the data provides ranges from location of accommodation, general neighborhood, as well as number of reviews. Though only three things were stated, the data has more attributes that can help aid in solving the problem that is being studied for this project. Predicting the pricing of an Airbnb given its attributes can be easily accomplished through this data as it has many attributes that can be observed. The data can be found through this URL: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data.

Data Wrangling

Through Data Wrangling we were able to clean up the dataset (if it was needed) to prepare it for subsequent step of exploratory data analysis

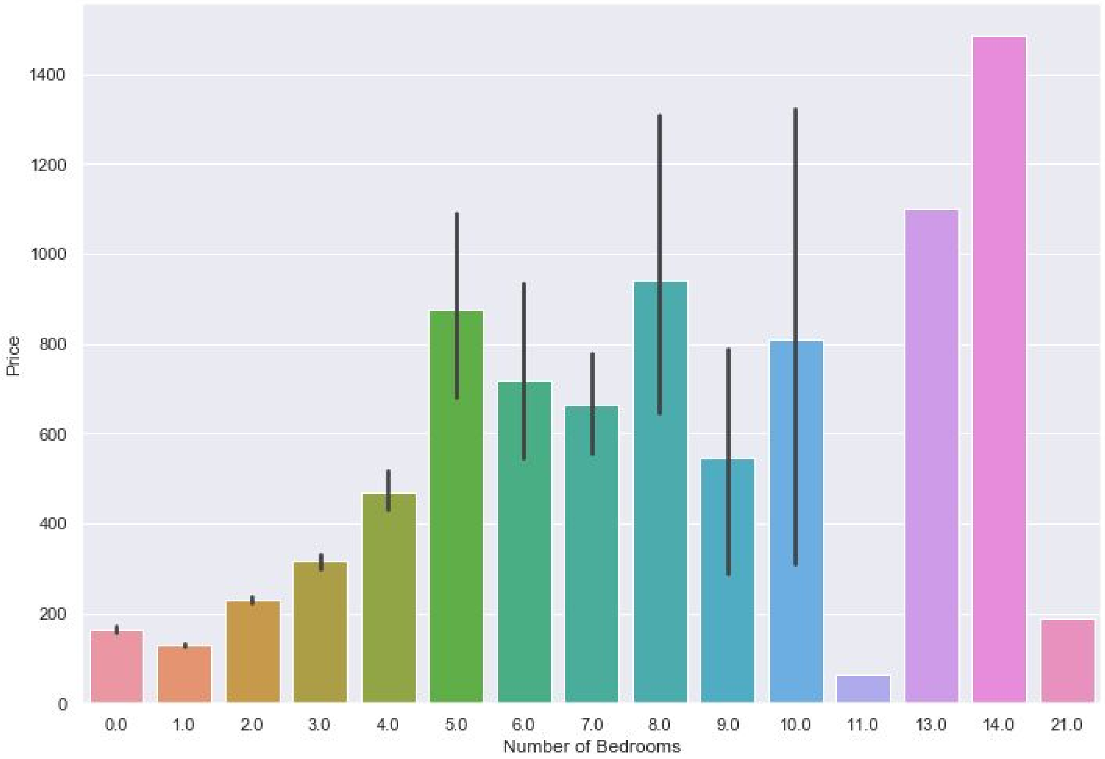
1. First Step of Data Wrangling would be to make sure the relevant columns are used from the data step, getting rid of columns that will provide no value to our research.
2. To further clean up the columns we make sure variables and cases are separated so that variables (number of rooms, neighborhood, etc.) are in the columns and cases (accommodation listings) are in the rows.
3. Next step is to make sure the case names are consistent and are not affected by any extra spaces or punctuation other than those that are needed for the name of the listings and other entries within the data set.
4. Dealing with null or empty data holes will be the following step. The way I dealt with these holes was filling in empty cells with NA.
5. Fixing Data Types such as time and fixing data with special characters is the next important step after dealing with null entries.

Initial Findings

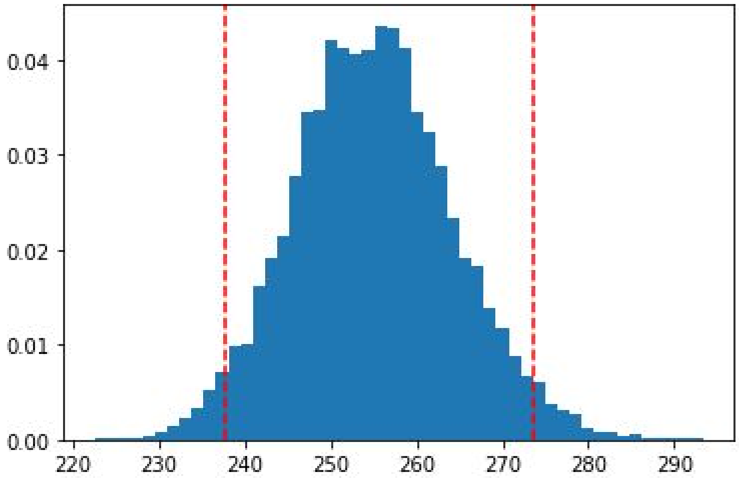
Through exploratory data analysis a lot can be observed and answered about the relationships between some of the columns or variables. Some important questions we ask are:

1. What variable affects the listing price the most?
2. Does the type of the listing or location of the listing matter?
3. If location matters, what locations are the most popular and why?

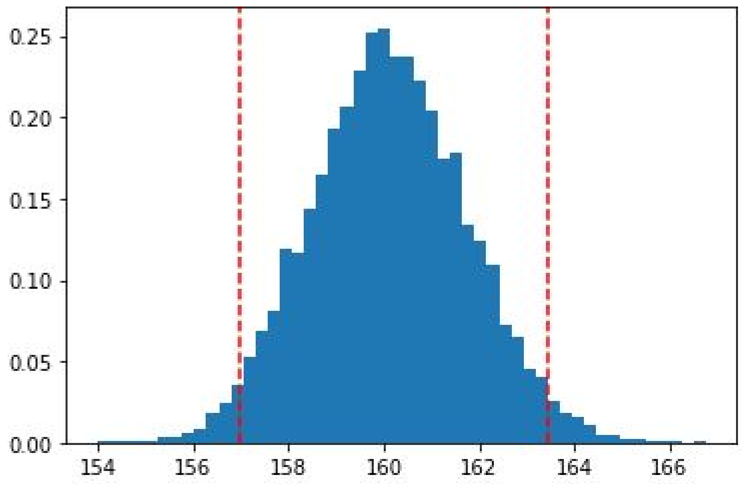
For the first question asked, I took a look into how bathrooms, bedrooms, and beds, as well as the neighborhood of the listing all affected the price of a listing. The one constant would be the fact that with the more bedrooms a place has the pricier the establishment becomes. One graph from the data would be the following:



As seen there may be some outliers that affect the data; however, the fact that the more bedrooms a listing has, the pricer it becomes is true. Looking at the above graph one can observe that the price increases as the number of bedrooms increase.



The above graph shows the mean(257~) and 97.5 and 2.5 percentiles for accommodations that have more than 3 rooms. The graph below shows the same values for accommodations that have 3 or less rooms. As seen the more rooms and accomodation has the pricier it will become.



Now when looking at location of a listing and property_type of accomodation it is seen that these two have significant effects on the price. Accommodations that are in the Manhattan neighborhood have a significant price increase than those of the other neighborhoods. Though this is no surprise given Manhattan's surroundings and location, two key attributes that have an effect on pricing of neighborhoods. When comparing neighborhoods that are suburban to those that are near or are cities, those that are near or are cities will obviously have a higher price point.