IST 707: Applied Machine Learning
Airbnb Project Report
Allison Deming, Nolan Mercado, Mikayla Scott, Alexender McCambridge
Dr. Ami Gates
09/22/2022

Contents

Allison Deming, Alexander McCambridge, Nolan Mercado, Mikayla Scott
ardeming@syr.edu, almccamb@syr.edu, nomercad@syr.edu, mscott18@syr.edu

Project Report
Data Visualization, Logistic Regression, Decision Trees, Correlation, Sentiment Analysis
September 22, 2022

EXTRA, EXTRA, READ ALL ABOUT IT!

1. Introduction

Traveling, whether for business or leisure, introduces individuals to new experiences, new markets, and unforgettable memories. However, if one's approach is unprepared, it can also be expensive and a tiring experience. Booking hotels, comparing costs, and being in an area that is unfamiliar can become quickly overwhelming leading to stress. This is the one thing that most individuals want to avoid when they are on holiday. There are many websites that offer what is called the "low price guarantee" on hotel rooms, but these are not the only options within a market. The option of non-traditional bookings such as shared rooms has only become available in the last decade or so. One answer to rising costs in housing markets and the hassles of a sterile hotel environment is booking an Airbnb. Airbnb is a service that allows property owners to rent out their unused spaces to travelers. It is popular not only because it is a less expensive alternative for those who cannot afford to stay in a hotel, but sometimes it can be difficult to even find a hotel room in a busy area; and the busier the area the more expensive those rooms that are available tend to become. People can rent out private rooms for individual guests, shared spaces for multiple people, or sometimes the entire property. This gives value to property owners who do not use the property year-round, and an opportunity for guests to have a more private as well as unique experience.

According to the official website, since the start of 2008, over 100 million people have booked rentals through Airbnb but how does it work? Airbnb works very similarly to a hotel, the difference is the house that is being rented out is owned by individuals, couples, or smaller real estate investment groups, but not by giant corporations such as Hilton Hotels, IGH (International Hotel Group), or Marriot (just to name a few). It is free to post a property for rent on Airbnb and for the host to post pictures of properties that help guests know exactly what they are renting. The host can set up the property listing to book anyone who meets certain criteria or require personal approval for each rental requested. This allows owners to exercise control over their investment.

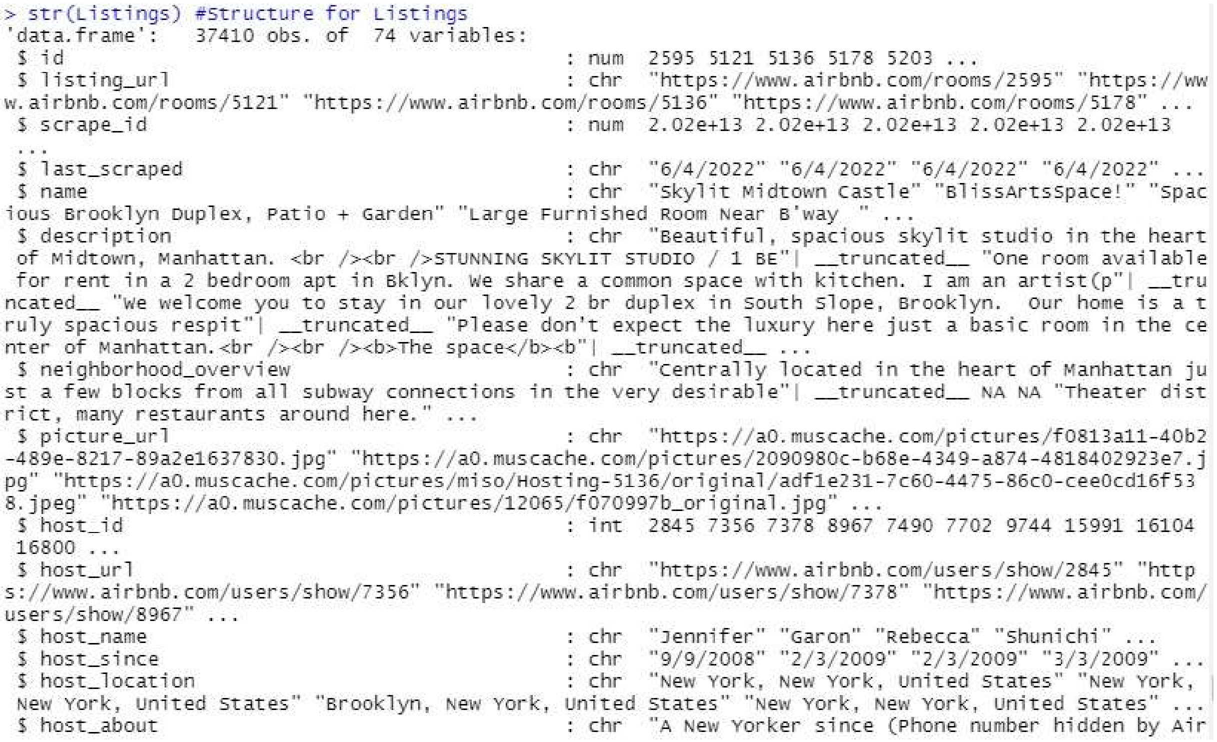
New York City is arguably the best city in the world. There is a certain excitement that completely encompasses someone there for the first time. Those who leave the city just to get away for a little while will eventually come back with that same certain love/hate attachment upon their return. The out-of-order controlled chaos, the nonstop blinking lights on Broadway, the concrete jungle of the five boroughs, even the heat radiating from the back of the MTA (Mass Transit Authority) bus with a younger gentleman selling the New York Times next to a bodega screaming, "Extra, Extra, Read all about it!" will scream, "Welcome Home Kid!" You can win anywhere if you can win here, same places, different faces.

New York City (NYC) is one of the most visited cities in the world and features many beautiful landmarks to visit and enjoy. From 1999 through the year 2020 the number of visitors to NYC has been massive, starting at around 29 million and gradually rising year over year to 66.5 million before the pandemic. Even during COVID the numbers were still large, in the range of 20-30 million visitors per year. The landmarks the people visit are plentiful but not limited to theaters, buildings, and a variety of department stores, but visiting famous landmarks such as the Statue of Liberty, Freedom Tower, Citi Field, Central Park, and the Empire State Building. Airbnb provides a fresh alternative to expensive and sterile hotels, making business travel and vacations affordable and comfortable for almost anyone.

Is there a way to help tourists make confident decisions for those who want to take a vacation in New York City? Is there a way to ensure business-minded people make calculated decisions that are on par with their own commitment to business decisiveness? Is there a way to summarize the five boroughs to find if something deeper is going on in New York City that is affecting Airbnb? How about Hosts and Guests? Will there be any availability? A deep dive is needed to find out what makes the best sense. Luckily for visitors and investors alike, Airbnb publishes vast amounts of data and provides it, for free, to the public. The collection and maintenance of this is part of their business strategy that relies heavily on data analytics.

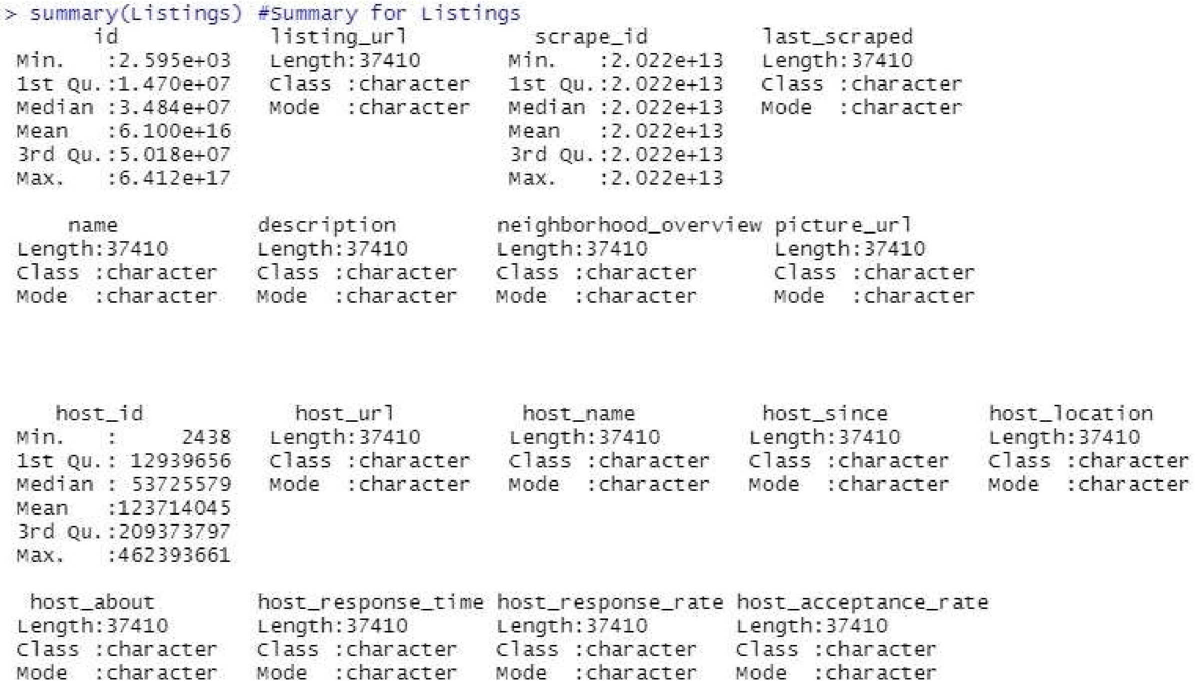
2. Data Preparation and Cleaning

The data set comes from a website named "Inside Airbnb." The company markets themselves as a mission-driven project that provides data and advocacy about Airbnb's impact on residential communities. As previously mentioned, NYC is a competitive real estate market, and the availability of affordable housing is in short supply. How a company like Airbnb impacts property values in addition to the neighborhood atmosphere would be important to all New Yorkers as well as city planning and leadership. In this dataset, there are a total of six CSV files that contain data relating to Airbnb rentals located in the five boroughs of New York. The data has one file named "listings.csv" which contains a summary of information and metrics for listings in New York City. Another data file named "reviews.csv" is a summary of review data and listing ID. Both listings and review have a truncated, or smaller version that is provided in CSV (comma separated value) form. The main difference between the larger and smaller reviews and listing files is the number of columns. The truncated or smaller data was used extensively in the data exploration in this report. A fifth CSV file with the top 300 tourist destinations was created along with their GPS Coordinates to provide information of the relationship between those destinations and the listings themselves. The sixth and last data file is named, "calendar.csv." This file contains detailed calendar data which shows those Airbnb rentals that are available on certain dates.



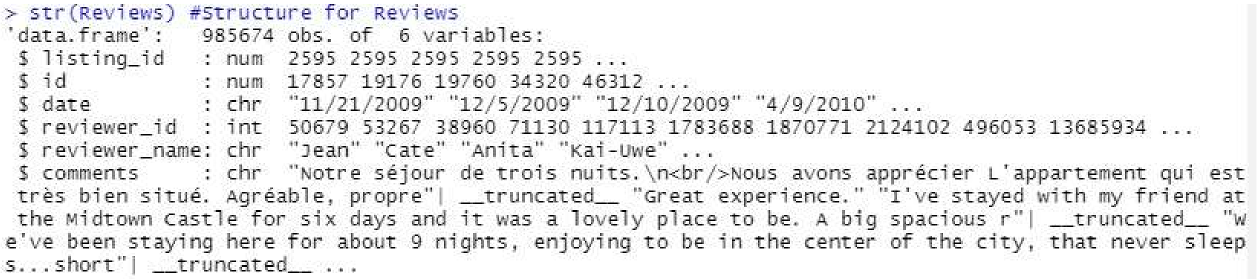
```
> str(Listings) #Structure for Listings
'data.frame':	37410 obs. of  74 variables:
 $ id                                          : num  2595 5121 5136 5178 5203 ...
 $ listing_url                                 : chr  "https://www.airbnb.com/rooms/2595" "https://ww
w.airbnb.com/rooms/5121" "https://www.airbnb.com/rooms/5136" "https://www.airbnb.com/rooms/5178" ...
 $ scrape_id                                   : num  2.02e+13 2.02e+13 2.02e+13 2.02e+13 2.02e+13
 ...
 $ last_scraped                                : chr  "6/4/2022" "6/4/2022" "6/4/2022" "6/4/2022" ...
 $ name                                        : chr  "Skylit Midtown Castle" "BlissArtsSpace!" "Spac
ious Brooklyn Duplex, Patio + Garden" "Large Furnished Room Near B'way  " ...
 $ description                                 : chr  "Beautiful, spacious skylit studio in the heart
 of Midtown, Manhattan. <br /><br />STUNNING SKYLIT STUDIO / 1 BE"| __truncated__ "One room available
 for rent in a 2 bedroom apt in Bklyn. We share a common space with kitchen. I am an artist(p"| __tru
ncated__ "We welcome you to stay in our lovely 2 br duplex in South Slope, Brooklyn.  Our home is a t
ruly spacious respit"| __truncated__ "Please don't expect the luxury here just a basic room in the ce
nter of Manhattan.<br /><br /><b>The space</b><b"| __truncated__ ...
 $ neighborhood_overview                       : chr  "Centrally located in the heart of Manhattan ju
st a few blocks from all subway connections in the very desirable"| __truncated__ NA NA "Theater dist
rict, many restaurants around here." ...
 $ picture_url                                 : chr  "https://a0.muscache.com/pictures/f0813a11-40b2
-489e-8217-89a2e1637830.jpg" "https://a0.muscache.com/pictures/2090980c-b68e-4349-a874-4818402923e7.j
pg" "https://a0.muscache.com/pictures/miso/Hosting-5136/original/adf1e231-7c60-4475-86c0-cee0cd16f53
8.jpeg" "https://a0.muscache.com/pictures/12065/f070997b_original.jpg" ...
 $ host_id                                     : int  2845 7356 7378 8967 7490 7702 9744 15991 16104
 16800 ...
 $ host_url                                    : chr  "https://www.airbnb.com/users/show/2845" "http
s://www.airbnb.com/users/show/7356" "https://www.airbnb.com/users/show/7378" "https://www.airbnb.com/
users/show/8967" ...
 $ host_name                                   : chr  "Jennifer" "Garon" "Rebecca" "Shunichi" ...
 $ host_since                                  : chr  "9/9/2008" "2/3/2009" "2/3/2009" "3/3/2009" ...
 $ host_location                               : chr  "New York, New York, United States" "New York,
 New York, United States" "Brooklyn, New York, United States" "New York, New York, United States" ...
 $ host_about                                  : chr  "A New Yorker since (Phone number hidden by Air
```

Figure 2.1: Partial Structure of the Listings Data Frame in R Studio



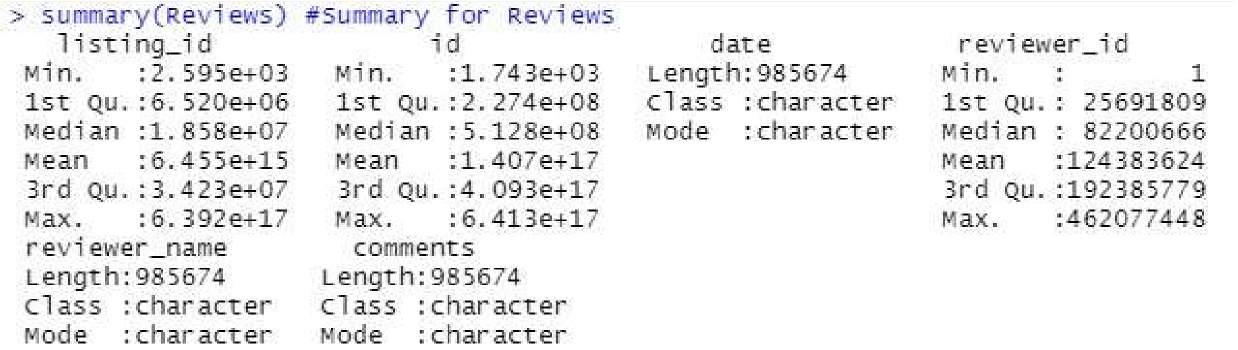
```
> summary(Listings) #Summary for Listings
       id              listing_url          scrape_id          last_scraped      
 Min.   :2.595e+03   Length:37410       Min.   :2.022e+13   Length:37410      
 1st Qu.:1.470e+07   Class :character   1st Qu.:2.022e+13   Class :character  
 Median :3.484e+07   Mode  :character   Median :2.022e+13   Mode  :character  
 Mean   :6.100e+16                      Mean   :2.022e+13                     
 3rd Qu.:5.018e+07                      3rd Qu.:2.022e+13                     
 Max.   :6.412e+17                      Max.   :2.022e+13                     

     name           description        neighborhood_overview picture_url       
 Length:37410       Length:37410       Length:37410          Length:37410      
 Class :character   Class :character   Class :character      Class :character  
 Mode  :character   Mode  :character   Mode  :character      Mode  :character  




    host_id            host_url          host_name          host_since         host_location     
 Min.   :     2438   Length:37410       Length:37410       Length:37410       Length:37410      
 1st Qu.: 12939656   Class :character   Class :character   Class :character   Class :character  
 Median : 53725579   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :123714045                                                                             
 3rd Qu.:209373797                                                                             
 Max.   :462393661                                                                             

  host_about        host_response_time host_response_rate host_acceptance_rate
 Length:37410       Length:37410       Length:37410       Length:37410        
 Class :character   Class :character   Class :character   Class :character    
 Mode  :character   Mode  :character   Mode  :character   Mode  :character    
```

Figure 2.2: Partial Summary of the Listings Data Frame in R Studio



```
> str(Reviews) #Structure for Reviews
'data.frame':	985674 obs. of  6 variables:
 $ listing_id   : num  2595 2595 2595 2595 2595 ...
 $ id           : num  17857 19176 19760 34320 46312 ...
 $ date         : chr  "11/21/2009" "12/5/2009" "12/10/2009" "4/9/2010" ...
 $ reviewer_id  : int  50679 53267 38960 71130 117113 1783688 1870771 2124102 496053 13685934 ...
 $ reviewer_name: chr  "Jean" "Cate" "Anita" "Kai-Uwe" ...
 $ comments     : chr  "Notre séjour de trois nuits.\n<br/>Nous avons apprécier L'appartement qui est
 très bien situé. Agréable, propre"| __truncated__ "Great experience." "I've stayed with my friend at
 the Midtown Castle for six days and it was a lovely place to be. A big spacious r"| __truncated__ "W
e've been staying here for about 9 nights, enjoying to be in the center of the city, that never sleep
s...short"| __truncated__ ...
```

Figure 2.3: Structure of the Reviews Data Frame in R Studio



```
> summary(Reviews) #Summary for Reviews
   listing_id              id                date            reviewer_id       
 Min.   :2.595e+03   Min.   :1.743e+03   Length:985674      Min.   :        1  
 1st Qu.:6.520e+06   1st Qu.:2.274e+08   Class :character   1st Qu.: 25691809  
 Median :1.858e+07   Median :5.128e+08   Mode  :character   Median : 82200666  
 Mean   :6.455e+15   Mean   :1.407e+17                      Mean   :124383624  
 3rd Qu.:3.423e+07   3rd Qu.:4.093e+17                      3rd Qu.:192385779  
 Max.   :6.392e+17   Max.   :6.413e+17                      Max.   :462077448  
 reviewer_name        comments        
 Length:985674      Length:985674     
 Class :character   Class :character  
 Mode  :character   Mode  :character  
```

Figure 2.4: Summary of the Reviews Data Frame in R Studio

Figures 2.2 - 2.4 show the structure of the data frames, and the summary values, respectively (at least in part). Only a partial data image is taken from RStudio here as the producing commands provide a very long output with 74 overall columns, and almost 40,000 observations or rows in the main data frame, and a much smaller value for the secondary companion. The two were combined later based upon a common identification number column between the two. For the next section, a truncated data set was used for exploratory data analysis (Section 2.2.2) due to the sheer size of the data set. This data was available on the Airbnb Data site where the zipped .gz full files were located, in addition to these smaller files for data exploration.

2.1 Data Exploration:

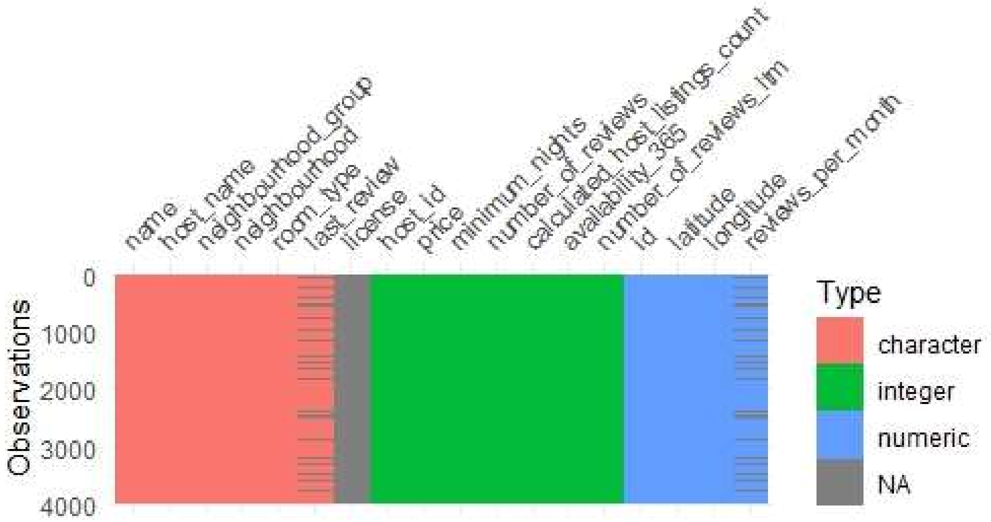


Figure 2.1.1: Missing Values (NAs) Indicated by Grey Coloring in Each Column for Small Data Set (Listings)

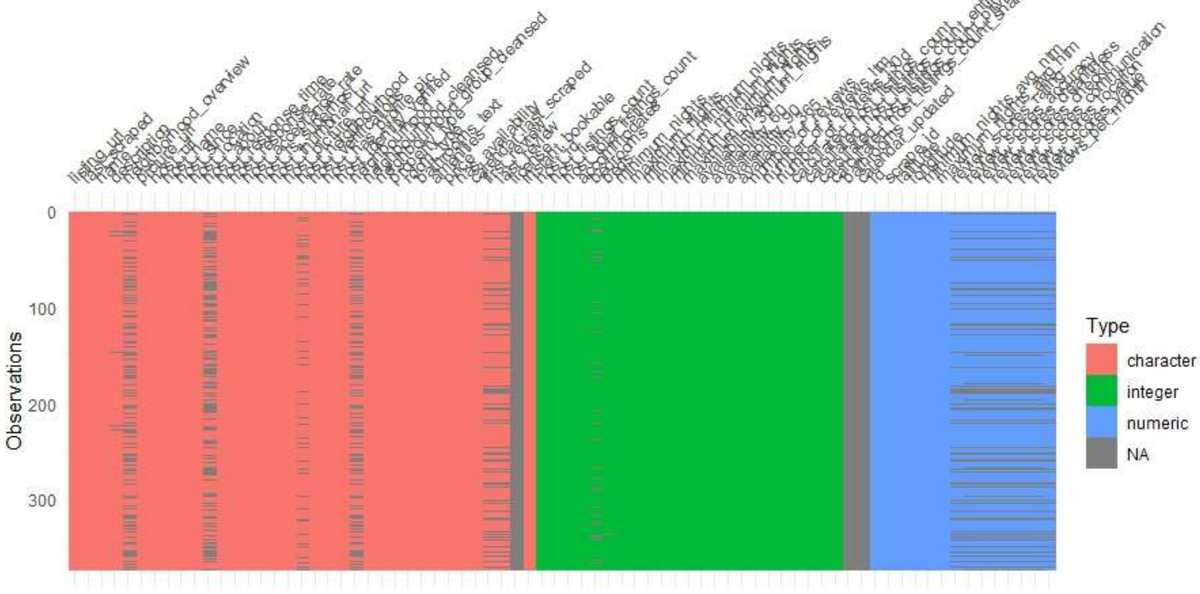


Figure 2.1.2: Missing Values (NAs) Indicated by Grey Coloring in Each Column for Large Data Set (Listings)

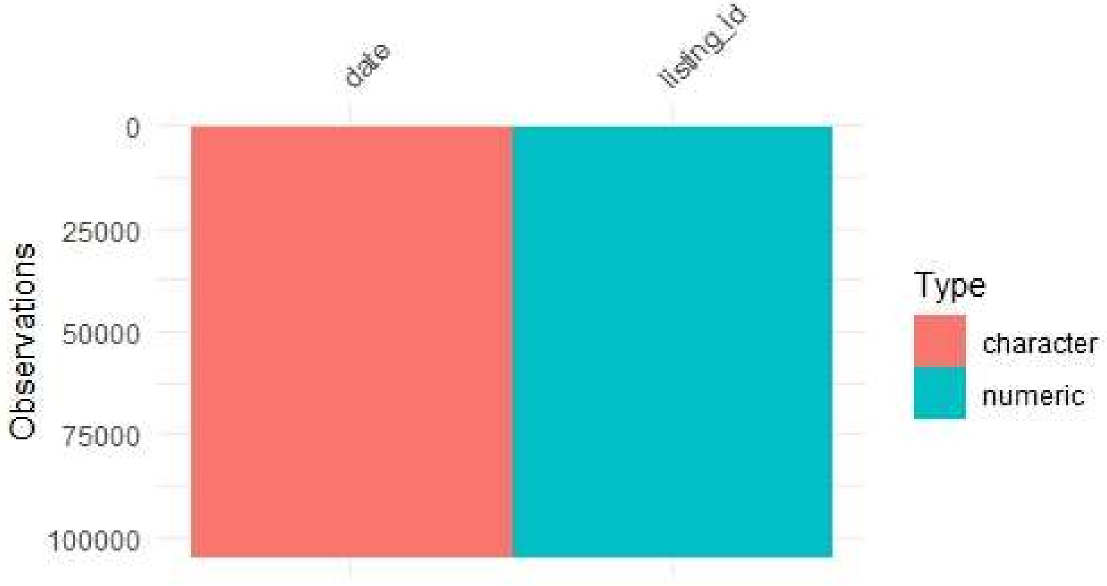


Figure 2.1.3: Reviews Data Truncated with Data Types (Truncated)

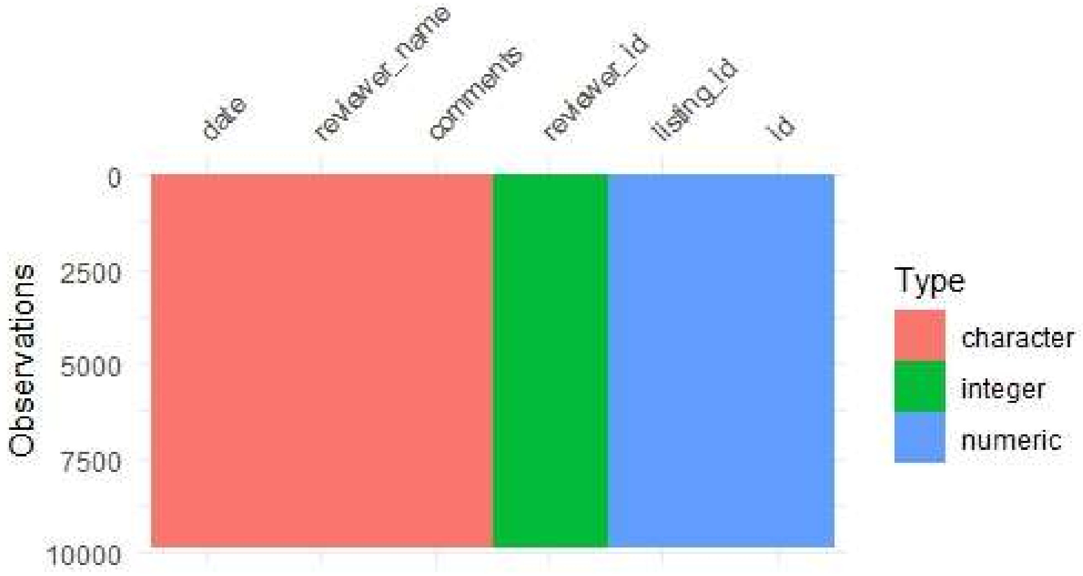


Figure 2.1.4: Reviews Data Large with Data Types (Large Data Set)

Exploring the listings shows some valuable introductory insights into what can be delved deeper into. Figure 2.1.1 shows the number of missing items from the data frame, as indicated by the color grey for NA. The same can be found in the subsequent Figures 2.1.2 through 2.1.4 This is using all four of the main data sets that were used during this project, two of which are truncated and used for many of the different exploratory visualizations (for ease and respective time of processing charts and graphs) whereas the other two data frames are much larger and contain the full values Looking at the bar plot below, Figure 2.1.5 contains Airbnb's counted across the five Boroughs of NYC. In fact, Manhattan and Brooklyn have more Airbnb availability than Queens, The Bronx, and Staten Island. This is enough variety for anyone to find a place, likely within their budget. This is true for leisure as well as work related travel.

Visiting NYC is better when visiting for a short time and much easier than being there for a long time. This is likely why there are more Airbnb's available in Manhattan. Manhattan is the central point of all that happens in New York City. Broadway shows, visiting The Vessel, Grand Central Station, and exploring Times Square with all the beautiful lights, is in vast opposition to the somber feeling of reflection one can feel at the new Freedom Tower and 9/11 Museum. All these tourist attractions are in or near Manhattan and most of Airbnb rentals are in this location. The second top location is Brooklyn which holds several tourist attractions such as the Brooklyn Bridge, The Brooklyn Cyclones, Coney Island, MCU Park, and the Barclays center. All these tourist attractions affect the amount of Airbnb's available. Queens has a couple of attractions so there is the third highest, meanwhile, the Bronx and Staten Island which are mostly public housing and smaller facilities and will most likely not have as many Airbnb's available. This can be seen in the listings data frame in column Availability 365 and Neighborhood Group.

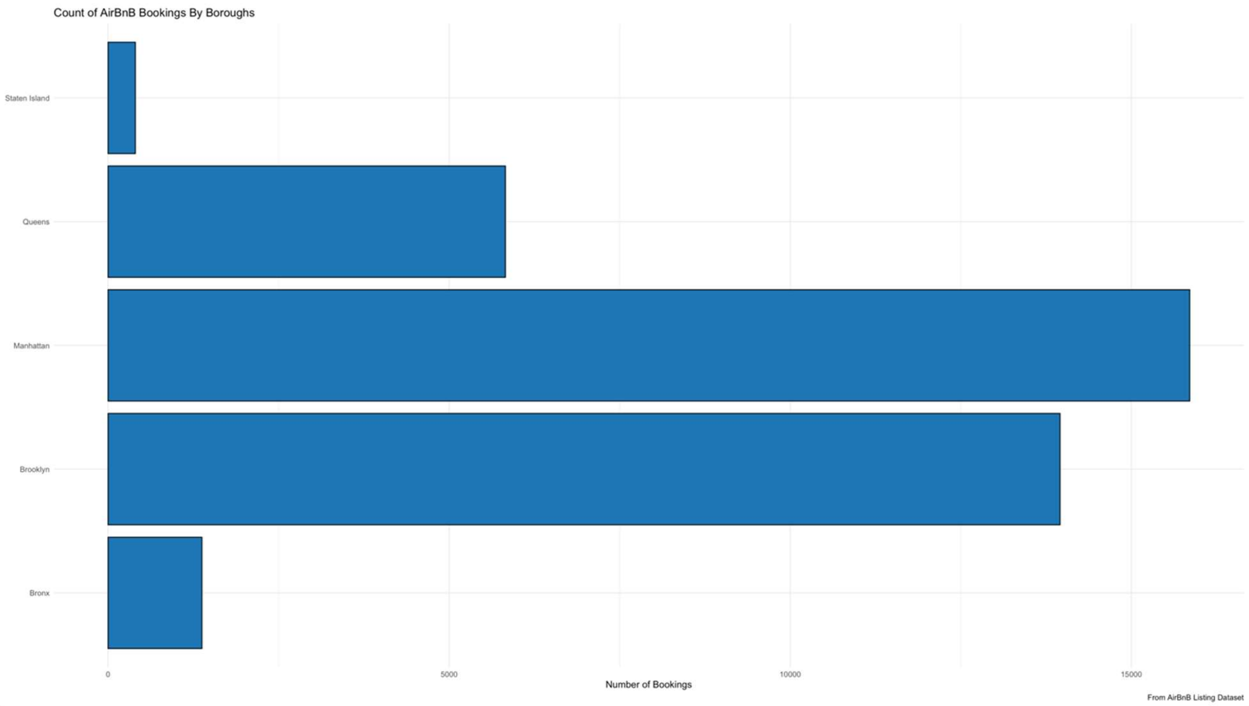


Figure 2.1.5: Horizontal Bar Plot of Price in NYC by Borough

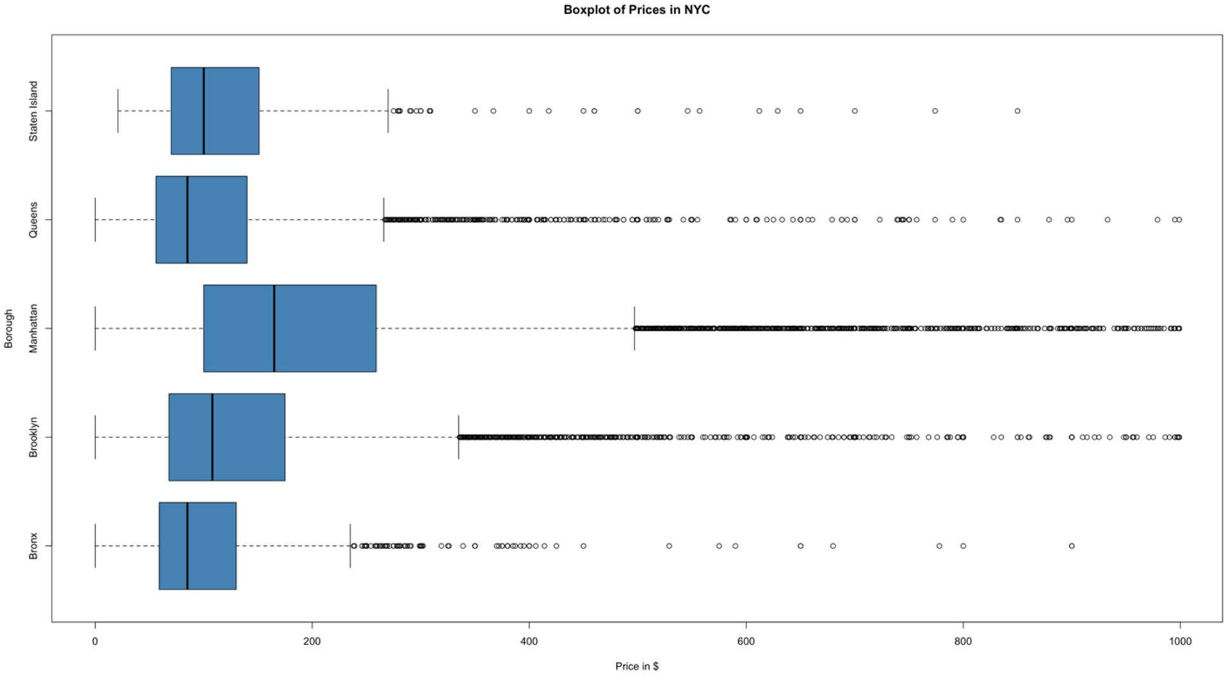


Figure 2.1.6: Horizontal Box Plot of Price in NYC with Outliers

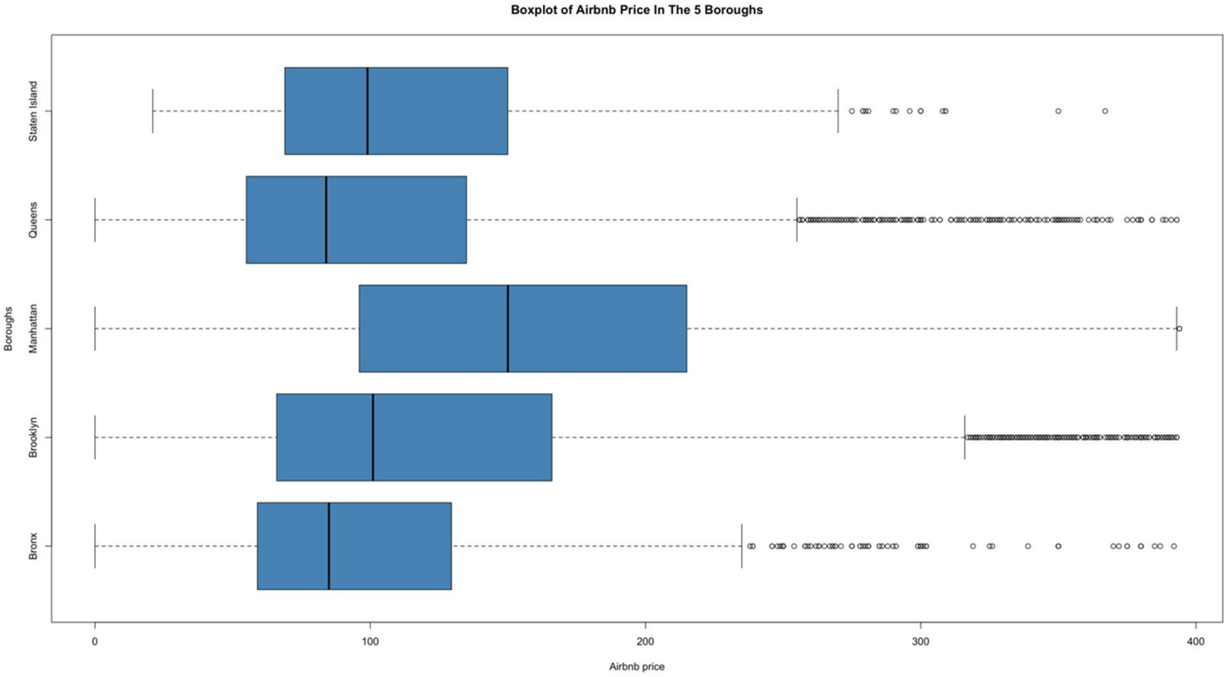


Figure 2.1.7: Boxplot of Price in NYC with Outliers

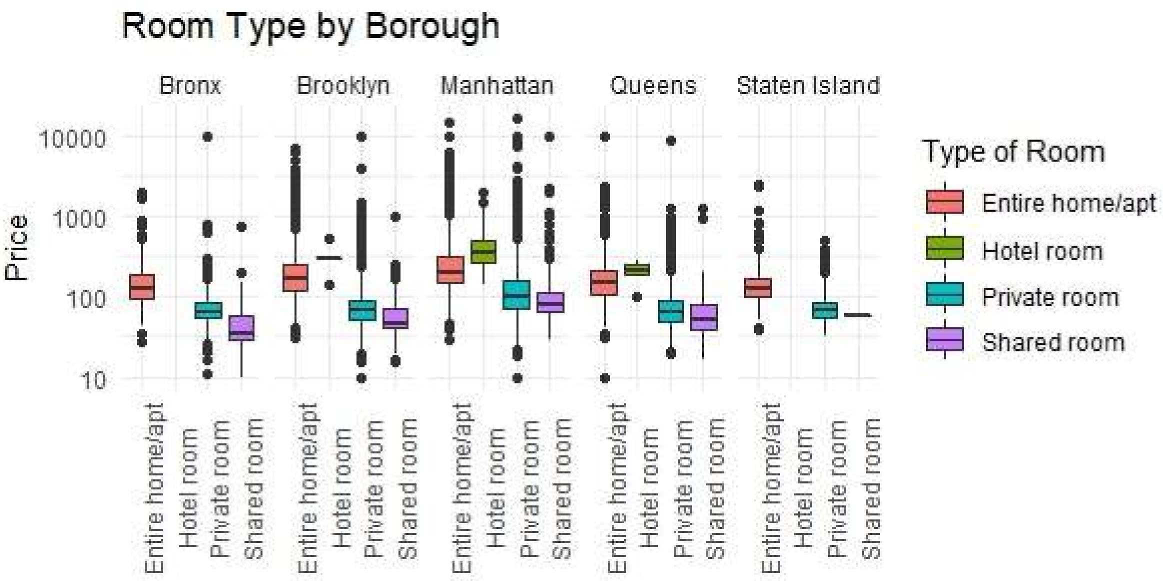


Figure 2.1.8: Room Type by Borough by Price Including Outliers

Looking at the box plots for the prices in New York City dispersed in the five boroughs, the prices are based on the location of Airbnb. A tourist is expected to rent out an Airbnb in Manhattan at a more expensive rate than any other borough in the city with the median price being $165 per night, with an average price of $212 per night. To put it into perspective, the median price of an Airbnb in the Bronx is $85, and the average price is $110 per night. There is an overabundance of Airbnb that are overpriced in all the boroughs, especially in Manhattan. This leads to speculation that some hosts are taking advantage of the high price of living in Manhattan for their Airbnbs. It also signifies that the rent is higher in Manhattan and setting their prices way over the maximum price in the Borough and the overall average of the Prices. The same can be said for Brooklyn Airbnb prices, however, the prices are way lower in Brooklyn than in Manhattan. The median price in Brooklyn is $108 and the average is $142. Compared to the median price on Staten Island is $100 for an Airbnb, and in Queens, the median is $85 while the average is $116. If a tourist is looking for an affordable price to go to New York City, the best place to go seems to be Brooklyn since there are nicer places, and more to see and do. This can be seen in Figures 2.1.3-2.1.4. Figure 2.1.5 shows a boxplot distribution by type of dwelling offered, by the price. In all cases, an entire home or apartment is more expensive than a shared space, or hotel room, which makes sense in the market. As hotel rooms are listed in numerous other places than just this site, they make up the fewest listing in all markets, and in fact none are shown to be available at this time in Staten Island and the Bronx.

Figure 2.1.9 shows a dot plot of room type by neighborhood, with the size of the circle of is representative of the number of rentals. Staten Island has the fewest rentals, with the most in Brooklyn and Manhattan, followed by Queens and the Bronx. The following two figures are representative of price, with the first being across all five boroughs, ranging from under $100 to well over $10,000. It is important to note that not only are shared rooms available, but the success of Airbnb has also led to more traditional hotel rooms listing available rooms on their platform.

While they are not most of the rooms available, this offers an easy comparison within the application for price and amenities.

The application also offers shared rooms for those by themselves or on a budget, all the way to private rooms, traditional hotels, and an entire home, apartment, or perhaps penthouse if one's budge permits. Both Manhattan and Brooklyn have over 7500 private rooms and full homes/apartments available to book; whereas Queens, the Bronx, and Staten Island have the same options but fewer individual spaces. This is likely because the areas are not as popular for tourists. Listing one's space on Airbnb is so popular, that the number of Airbnb rentals began to exceed the available number of open apartments in each borough in 2022. The value of property is being able to be maintained by the company or individuals that currently own it. NYC has always been a competitive market due to the limited space available, but this indicates a possible lack of affordable housing in the area, but that is outside the scope of this analysis. The price and number of overall listings can be seen in Figure 2.1.10 with a log transformation to help with the right skew and outliers in the data. Figure 2.1.9 shows the number of listings by Neighborhood and Room Type in a dot plot, with the smallest circular representations being less than 1000 in number of rentals, followed by four other sizes that range from 2500 to 10,000 listings.

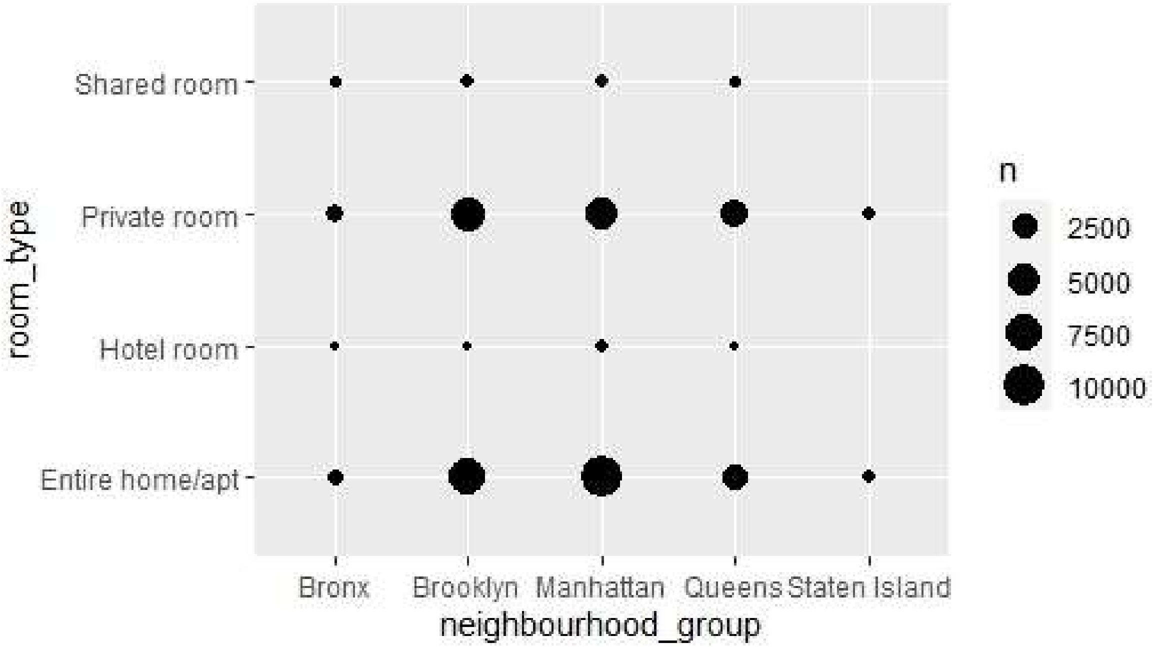


Figure 2.1.9: Number of Listings by Neighborhood and Room Type

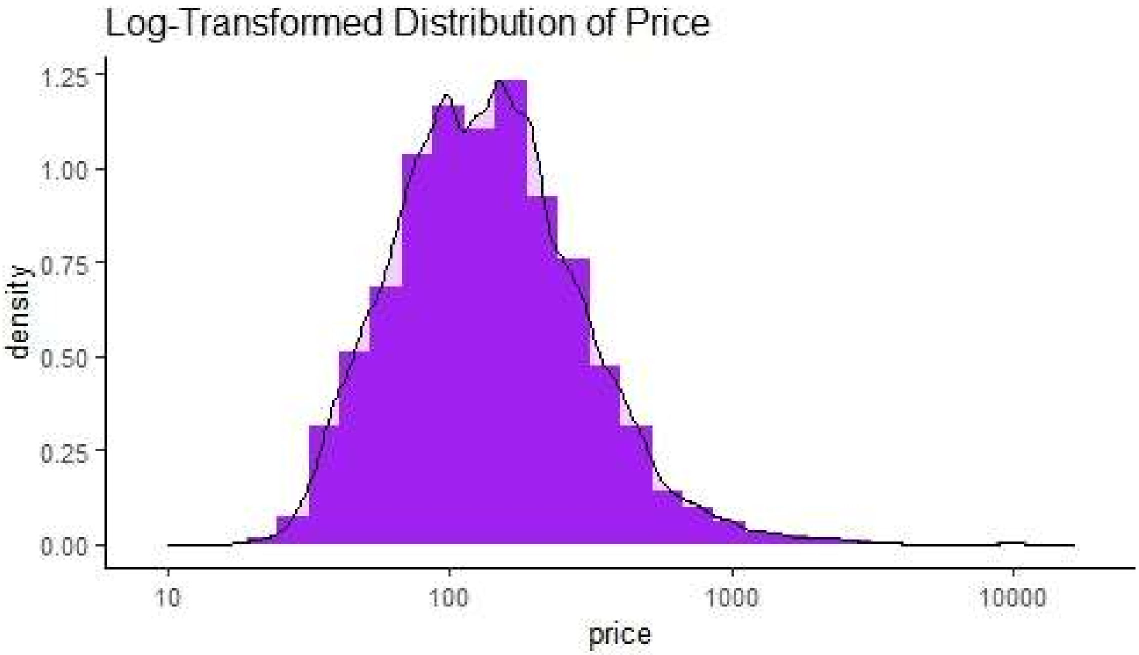


Figure 2.1.10: Log Transformed Distribution of Price by Density

Figure 2.1.12 is a line graph of broken down by the five boroughs, including the Bronx, Brooklyn, Manhattan, Queens, and Staten Island. In this graphical representation, the Bronx is illustrated by a pink line, Brooklyn a mustard-colored line, Manhattan a green line, Queens a blue line and Staten Island is Purple. The X axis of the graph shows prices, which range from zero to about $1000. The Y axis represents the number of available rentals. While each distribution is different, the peak price is in roughly the same location, which is around $100, with a right skew that extends to anywhere between $500 and $1000. The largest number of listings appear to be in Brooklyn with just short of 8000 unique points, Manhattan having almost 7000, Queens having around 3100, the Bronx showing less than 1000, and Staten Island having the fewest points available. This can also be seen in Figure 2.1.17 which shows points between zero and $1000, or what would be considered the bottom tenth of the data available, which is most affordable in each borough. This chart was made using the truncated data, whereas Figure 2.1.11 was made using the larger data set.

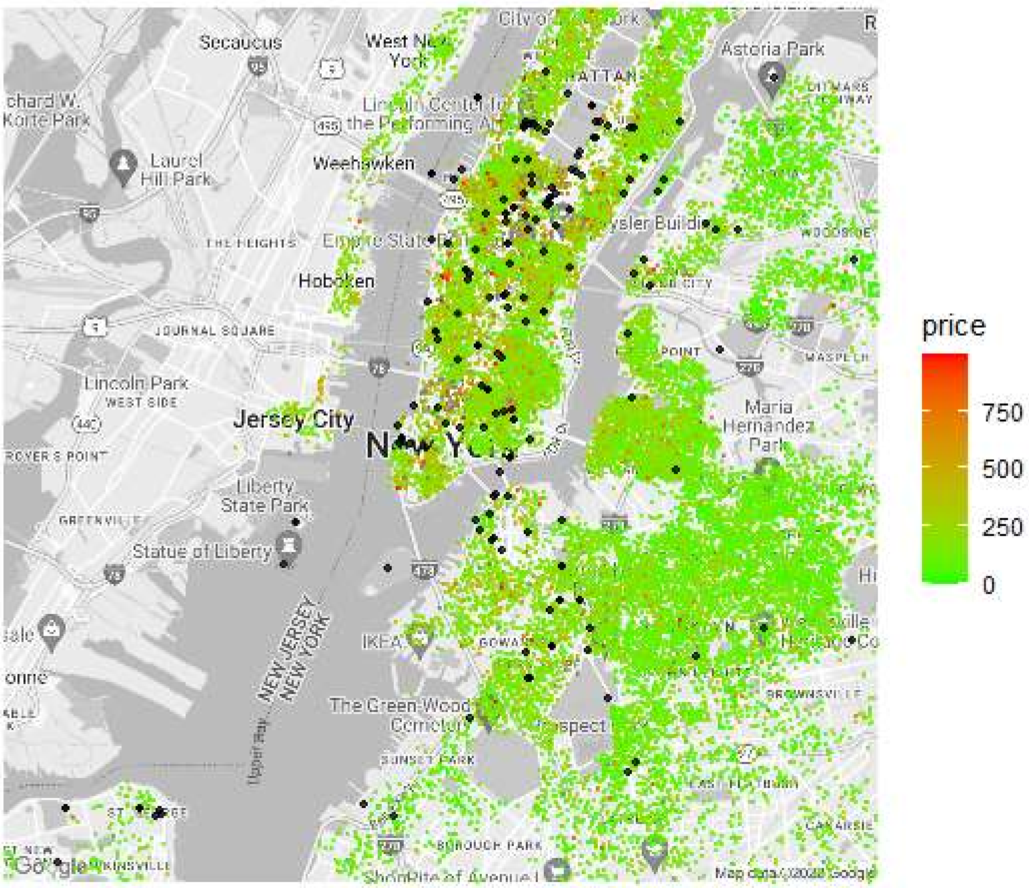


Figure 2.1.11: Map of NYC Latitude/Longitude by Price and Relative Location
Green = Low, Red = High; Black = Top 300 Tourist Destinations Within the Neighborhoods (made exclusively for this project).

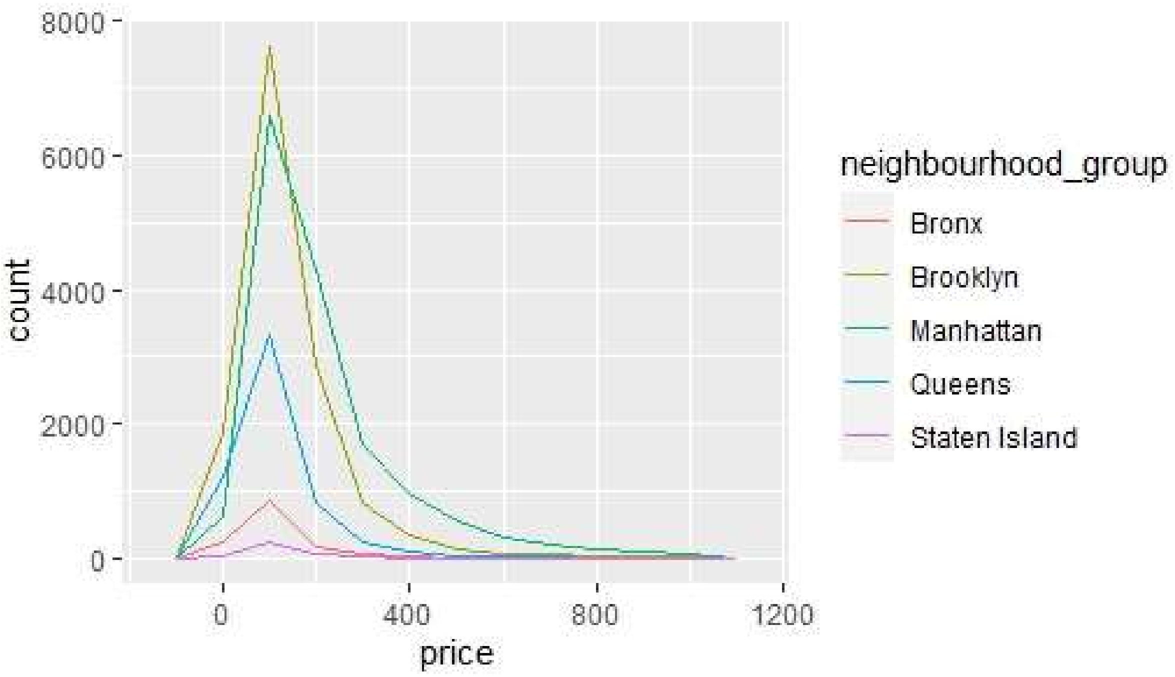


Figure 2.1.12: Price by Neighborhood by Number of Listings

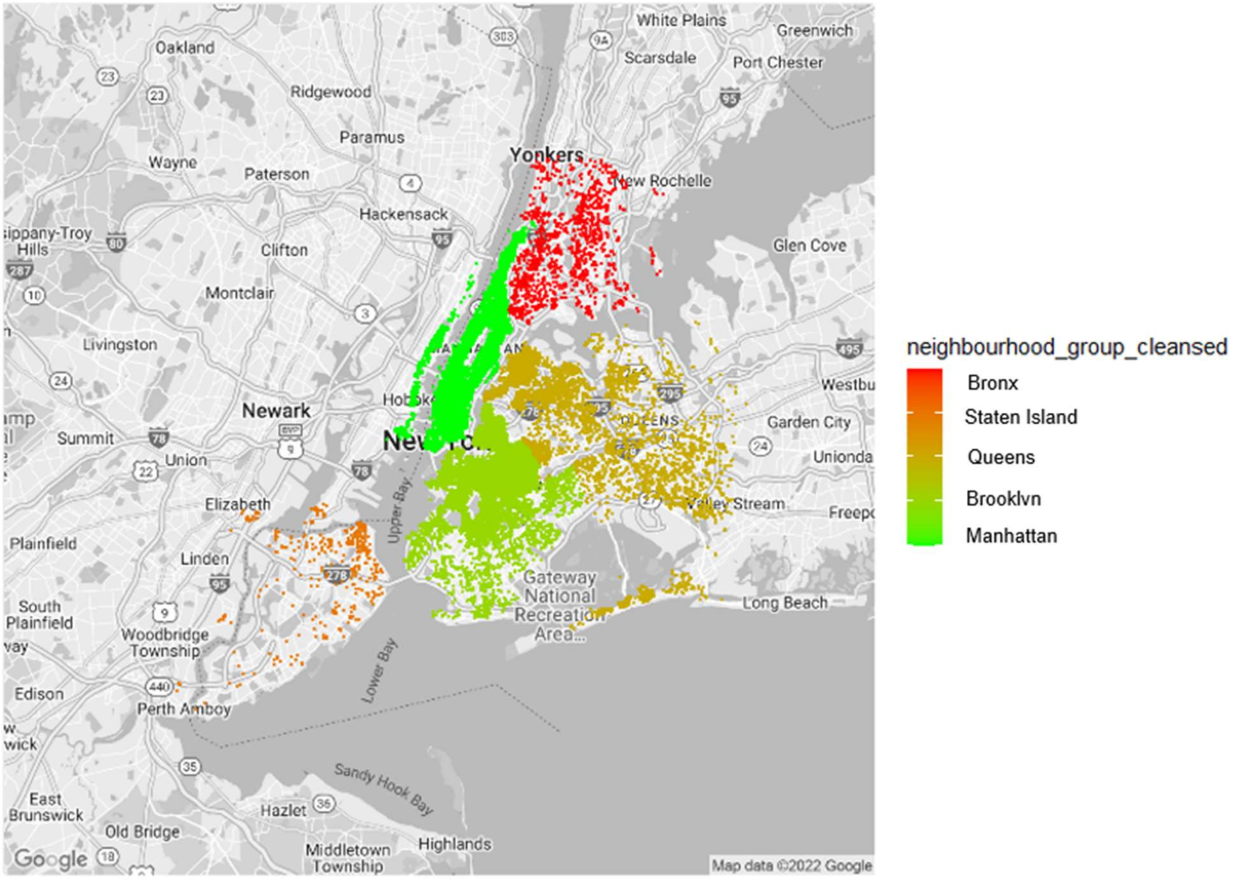


Figure 2.1.13: Listings by Borough Rentals of Airbnb Locations

Figure 2.1.13 is a listing of available rentals plotted on a map by latitude and longitude and divided by borough by the color on the legend color gradient. The legend is not categorical as some others are, but continual in terms of shading from red to green. For example, The Bronx and Staten Island are Red and Orange respectively, with Queens being in the middle and showing a slight greenish-red tint, and Brooklyn and Manhattan following with a yellow green color, and a bright green in turn.

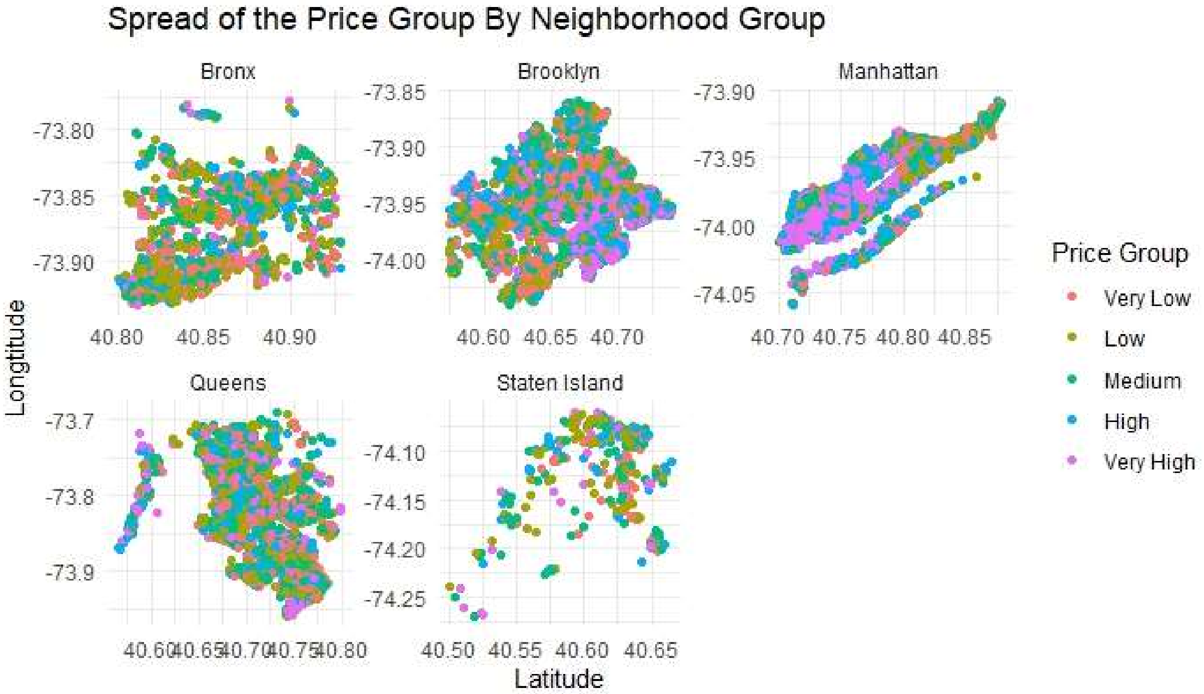


Figure 2.1.14: Spread of Price Group by Neighborhood in the Five Boroughs

Figure 2.1.14 shows the discretization of each of the main Neighborhoods in NYC, with price separated into quintiles, where each represented 20% of the whole. The bottom 20% is pink and in the very low group, next is low which is shown in chartreuse, the middle 20% of data is the medium quintile, which is shown in green. High, the next 20% is represented by blue, and the very high values, or the top 20% are shown in purple. Each of the boroughs are shown mapped via their latitude and longitudinal coordinates.

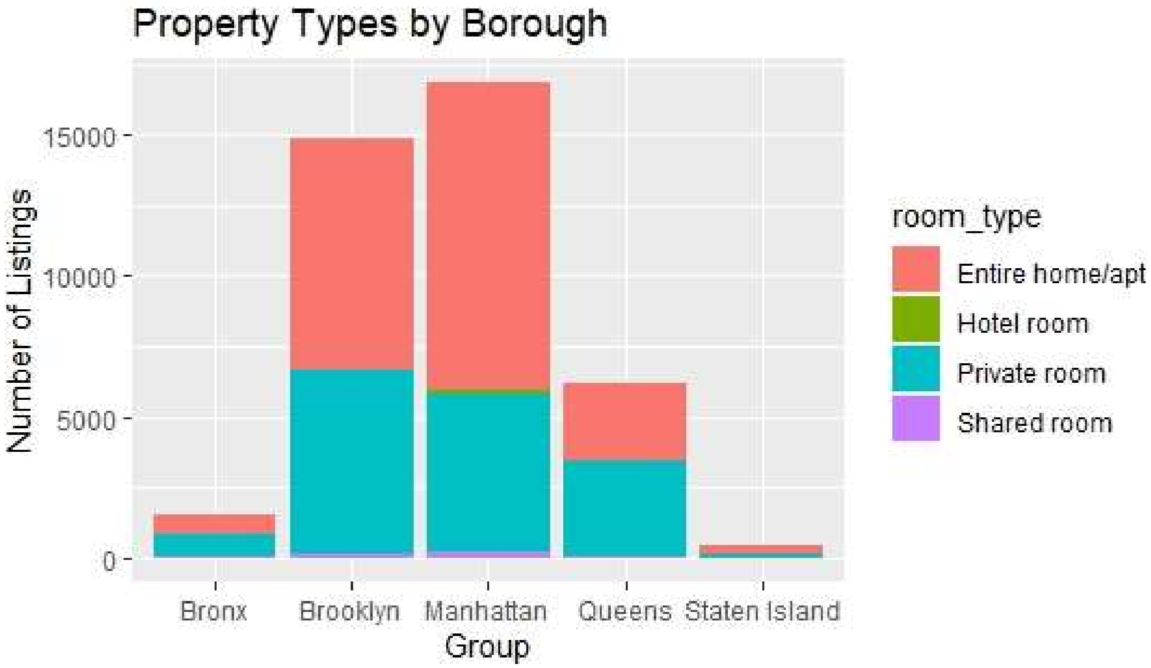


Figure 2.1.15: Stacked Bar Chart by Borough, Room Type and Number of Listings

Figure 2.1.15 is a stacked bar chart that shows the number of listings on the Y Axis that range anywhere from less the 100 to over 15,000. The X Axis is done categorically, with the Bronx, Brooklyn, Manhattan, Queens, and Staten Island listed left to right. The legend shows the different options of room type, with an entire home or apartment in pink, a traditional hotel room in green, a private room within a shared residence as blue, and a shared room within a residence a purple. In this Figure one can also see that the fewest listings exist in Staten Island, followed by the Bronx and Queens. Brooklyn and Manhattan have the most listings of all types with Brooklyn coming in just under 15,000 and Manhattan just over 15,000. Additionally, entire homes or apartments tend to make up about half of the listings available to patrons of Airbnb, likely being the more popular options. Shared rooms and homes, though more budget friendly, tend to be less socially acceptable to those renting as well as the owners of the dwelling itself.

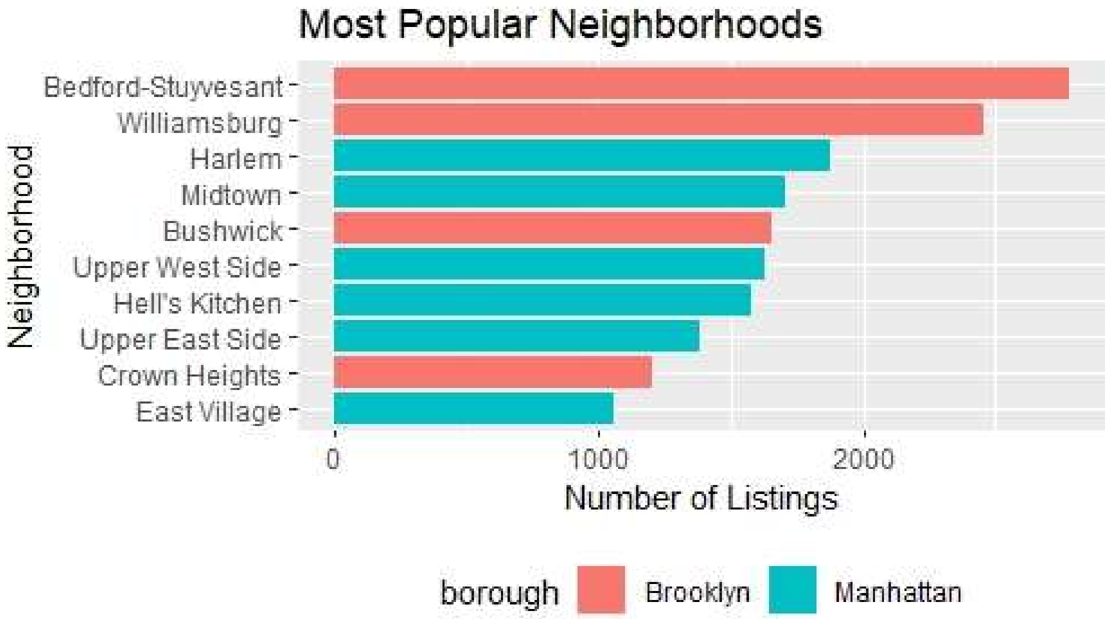


Figure 2.1.16: Most Popular Neighborhoods by Number of Listings by Borough

Figure 2.1.16 is a horizontal bar chart that shows the most popular neighborhoods within the two most popular and populous neighborhoods of Brooklyn and Manhattan. Brooklyn has four different neighborhoods on this list, which are Bedford-Stuyvesant, Williamsburg, Bushwick, and Crown Heights in descending order of available listings. Manhattan is home to six different neighborhoods on this list, with Harlem, Midtown, the Upper West Side, Hell's Kitchen, the Upper East Side, and the East Village being shown in descending order of availability. The neighborhood with the most listings is Bedford-Stuyvesant clocking in at just under 3000 available Airbnb listings, followed by Williamsburg at just under 2500 options. Harlem is third on the list with around 1800 listings, followed by the rest.

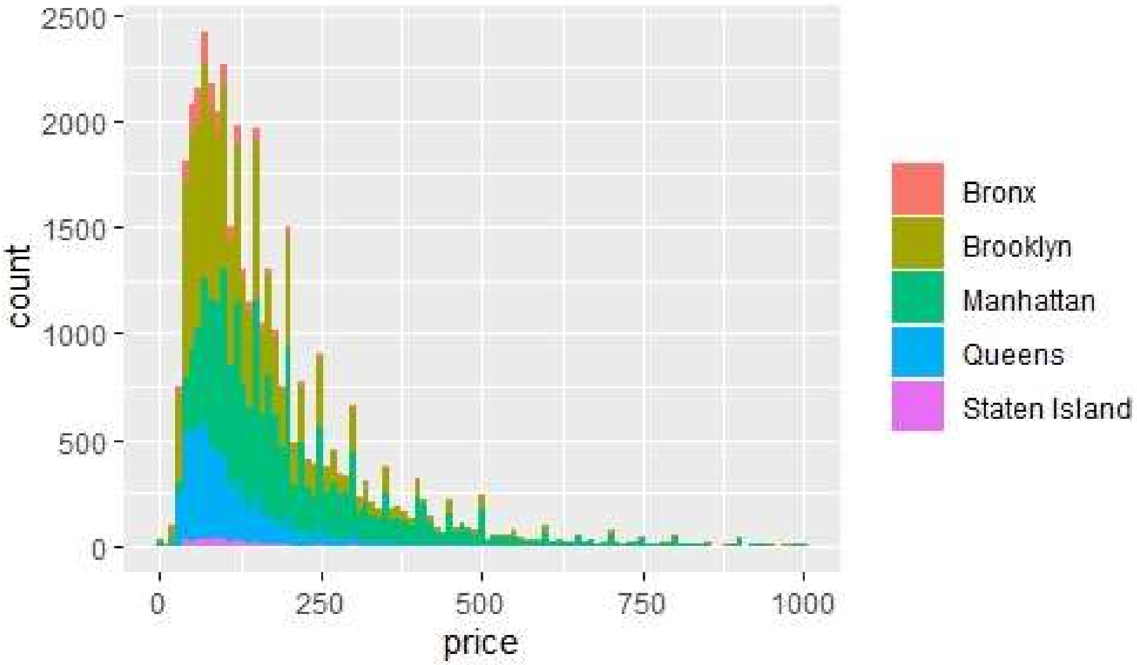


Figure 2.1.17: Listings Under $1000 by Borough and Price and Count

Figure 2.1.17 narrows down the listings that are available by price. Theis shows prices less than $1000, and as some listings are as high as $10,000 per rental, this is only a small percentage of the available listings but considered the more affordable ones. The histogram shows that many of the listings are right skewed, with about 40% of the listings falling between zero and $100, and another 30% to 40% falling between $100 and $300. The right tail goes all the way to $1000, but everything to the right of that was excluded from this analysis and graphical representation. Each borough was represented by a different color, with peaks in different areas price wise. Staten Island has the fewest listings and is represented by the color purple. Queens peaks around $85 and has the next highest number of listings represented by blue in this figure. Manhattan has the middle number of listings, and the peak price is around $100 (although there are listings much higher than that, in the tens of thousands of dollars) and is represented by green in this graph. Brooklyn has the next highest number of Airbnb listings with a peak price lower than that of Manhattan, but higher than that of Queens. It is represented by a moss green color on the graph. The Bronx has the highest number of listings, coming in just under 2500 Airbnb locations. Queens is represented by pink on this Figure.

Next, the data was discretized by price into five groups. The first, and lowest was labeled "Very Low", the next up was labeled "Low", the middle group and center of the arrangement was labeled "Medium" in price, the fourth group was labeled "High", and the fifth group was labeled "Very High" in terms of price. These are mapped by the coordinates provided by each Airbnb rental, giving a rough outline of each individual borough itself, aside from Staten Island which has the fewest rental available. Within each may is the latitude on the X Axis, and the Longitude on the Y Axis, with Very low represented by a pink color, Low priced rentals represented by green, Medium represented by a teal color, High priced units are represented by a cobalt blue, and very high-priced rentals being represented by purple. This can be seen in Figure 2.1.14.

The reviews data set provided the reviews for the Airbnb listings. Extensive sentiment analysis was done to find correlations between different variables in the listings data set and the reviews written on those same Airbnbs. The first task was to identify the top words in the data set. The word cloud that can be seen below (Figure 2.1.18) was generated from a random sample of 10,000 reviews. Words such as “great”, “place”, and “stay” were among the top recurring words in the random sample.

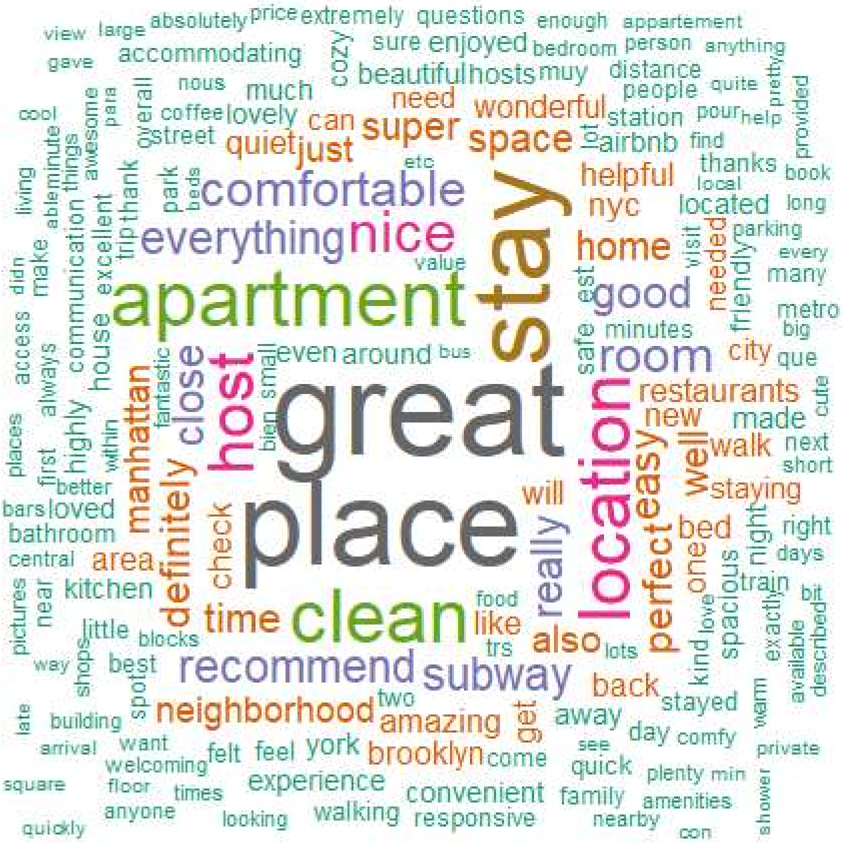


Figure 2.1.18: Word Cloud of the Reviews Data Frame with Size Representing Totals

In addition to the random sample of 10,000 of the reviews. The Table (2.1.18) shown below shows words that occurred more than 50,000 times words across all 985,674 revies. There were 77 words that occurred more than 50,000 times across the entirety of all the reviews data frame combined.

Experience	Great	Everything	Just	Like	Location	lovely
Much	New	Perfect	Place	Really	Room	spacious
Staying	Time	York	City	Friendly	Good	Hosts
Kitchen	Manhattan	Subway	Walk	Apartment	Definitely	Helpful

Recommend	Stay	Wonderful	Around	Can	Highly	host
Located	Need	One	Quiet	Will	Need	One
Convenient	Area	Away	Bathroom	Comfortable	Night	Safe
Street	Close	Clean	Nice	Airbnb	Space	Also
Cozy	Amazing	Check	Communication	Made	Well	Felt
Home	Nyc	Back	Easy	Needed	Super	Enjoyed
Loved	Beautiful	Bed	Neighborhood	Brooklyn	Restaurants	House

Table 2.1.19: Most Popular Neighborhoods by Number of Listings by Borough

Word associations were also produced using some of the top recurring words from the random sample of 10,000 reviews. Some of those results can be seen in the table (2.1.20) below:

	Top Recurring Words								
Associated Words (AW)	**Stay**	**Apartment**	**location**	**clean**	**host**	**Manhattan**	**comfortable**	**perfect**	**wonderful**
AW 1	Definitely	Subway	slide	quiet	blankets	jersey	bed	bundled	gilding
AW 2	Enjoyed	Walk	tendency	bathroom	abnormally	boarders	beds	miserable	opulent
AW 3	Mad	Restaurants	UberEATS	room	balled	companies	felt	pants	overstate
AW 4	Throughout		delivery	everything	blood	shore	kitchen	prove	specimen
AW 5	Night	NYC	frustrating	kitchen	corrected	skyscraper	well	purchased	verbally
AW 6	Home	Kitchen	tent	towels	dried	SuperShuttle	room	sweaters	wickedly
AW 7	City	Neighborhood	fence	sparkling	heats	tunnels	subway	trekking	hexagonal
AW 8	Extremely	Large	parties	safe	lectured	minutes	home	city	linsey

Table 2.1.20: Top Recurring Words in the Review Data Frame

A lot of the most recurring words in the random sample were associated with positive reviews, words like wonderful, nice, and perfect were often associated with other positive words. Other words that are typically neutral, like location, clean and Manhattan showed attributes of that specific word. Host was an interesting case, most of the words that were associated with host are words that are typically not found in a positive review.

Sentiment analysis was also run on the entire dataset. Within the dataset (985,674 reviews) there were 26,643 negative reviews, 73,006 neutral revies and, 886,025 positive reviews. Negative reviews make up 2.7 %, neutral reviews make up 7.4% and positive reviews make up 89.9% of the dataset. Where the sentiment of a review is based on a scale from -1 to 1 where -1 is defined as extremely negative, 1 is extremely positive and 0 is neutral. For this analysis negative is defined as a compound score of in between -1 and -0.05, neutral is defined as a compound score of in between -0.05 and 0.05, and positive is any compound score between 0.05 and 1.

Reviews were labeled and then merged with the listing's dataset on the "id" column from the listings data set and the id column in the review's dataset. The original location of the Airbnb listing was combined with the sentiment tag of corresponding review from the review's dataset. In the map below green depicts an Airbnb with a corresponding review that was predicted to be positive by the sentiment analysis model, orange is neutral, and red is a negative review.

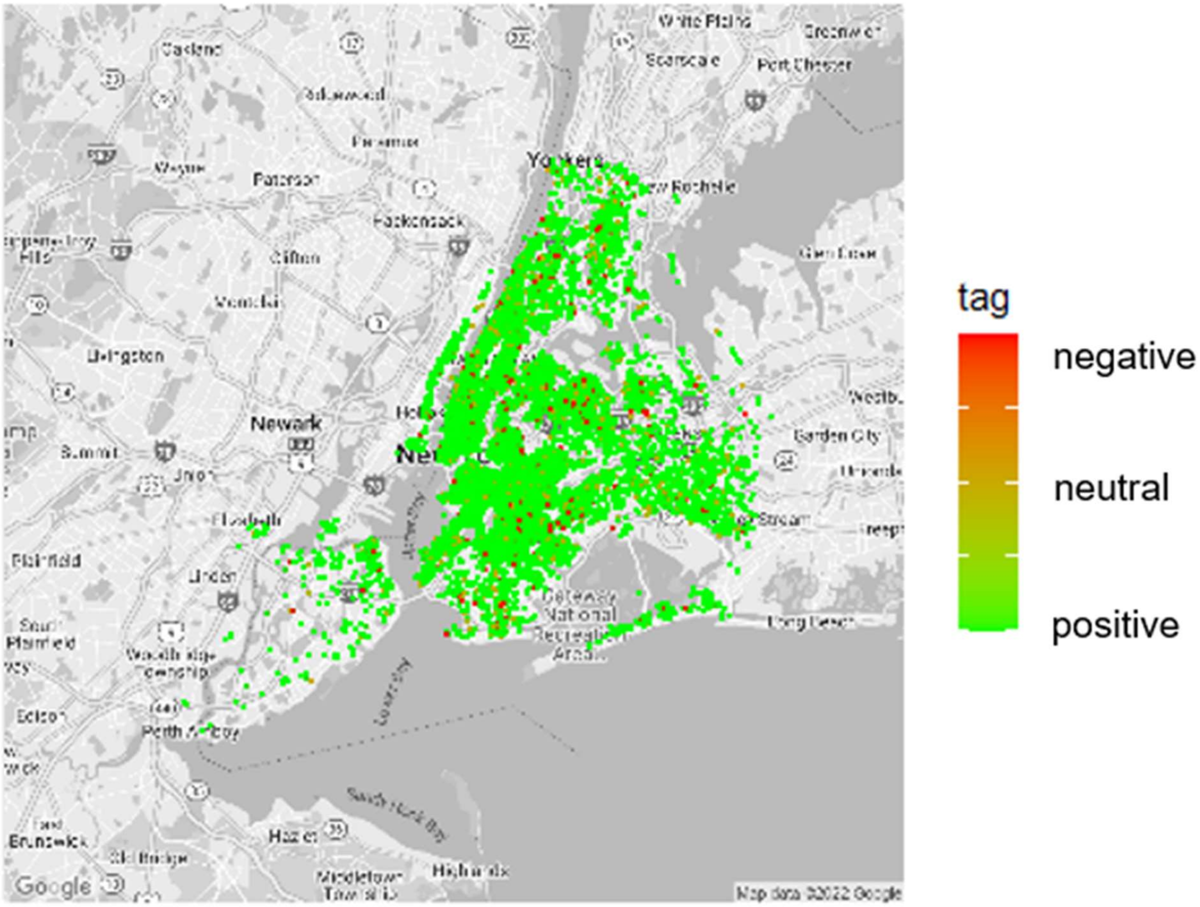


Figure 2.1.21: Reviews from Negative to Positive (Red to Green) by Location

2.2 Models and Methods:

It is important to model the factors for renters and compare those Airbnbs to each other. For example, a renter may want to know if the price is relative to the location or number of Airbnbs hosted by the same person. They may want to know if the number of days that Airbnb will be available for 365 days a year, or the number Airbnb's host will post is related. A correlation matrix will help indicate which variables move in tandem with each other. If the two variables move in the same direction, then those variables are positively correlated. If they move in the opposite direction, they are negatively correlated. It is also possible that there is no correlation at all, but that is not what is seen here in some of the pairwise cases. According to the correlation matrix below, an Airbnb which is available 365 days out of the year, is positively correlated with room type, location, price, and minimum nights. So as one increases or decreases, the other does the same. These regions are marked with a faded blue circle. What is interesting is that the price is negatively correlated with the location and room type. It is a very strong negative correlation. This can be seen in Figure 2.2.1.

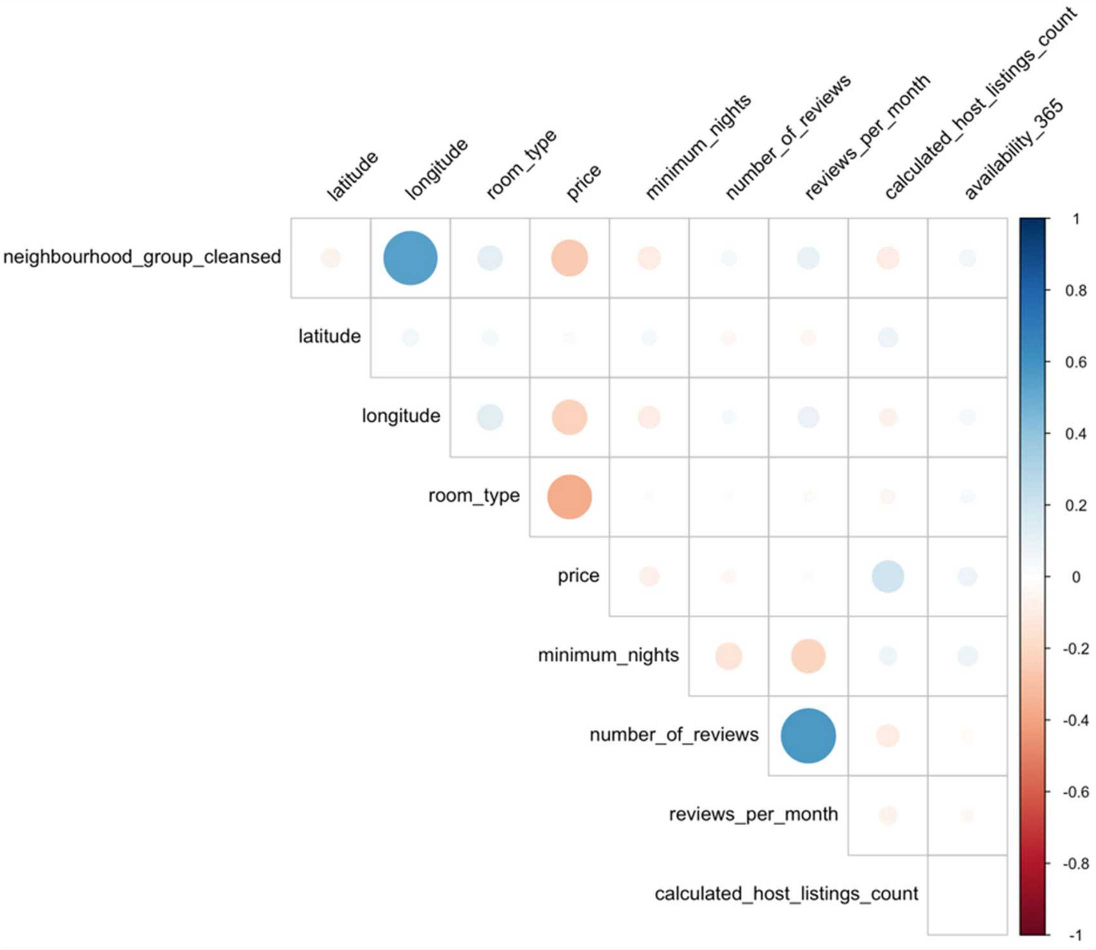
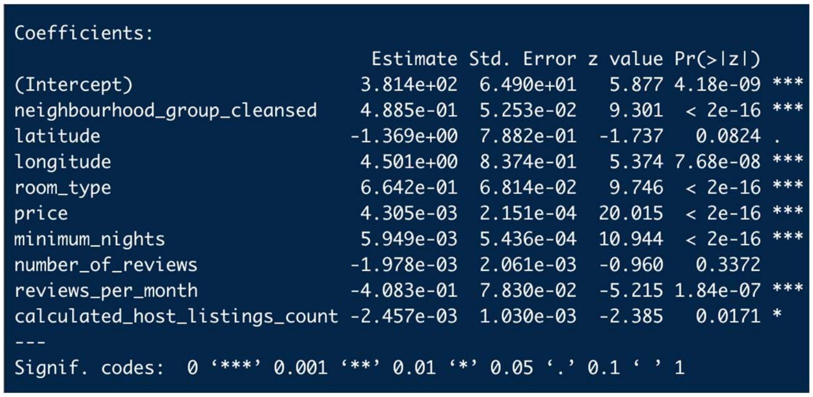


Figure 2.2.1 Correlation Matrix of NYC Airbnb

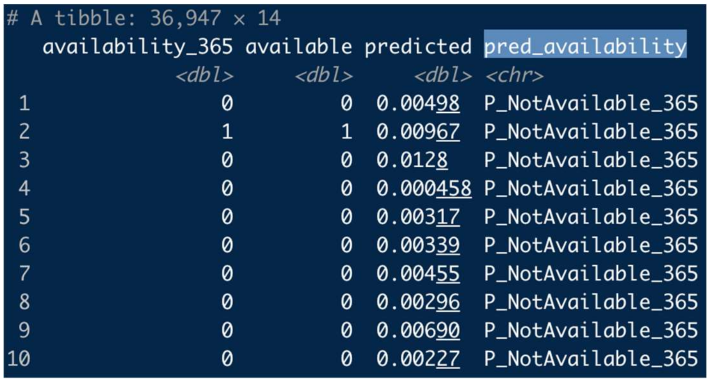
Knowing what is correlated in this data naturally sets the tone for performing a binomial logistic regression model. Logistic regression is a statistical method to predict a certain outcome with limited options. In this case, what effects the probability that an Airbnb will be available 365 days? Prediction is based upon the location, room type, price, minimum nights, and reviews per month. A logistic model predicts a dependent variable by analyzing the relationship between one or more existing independent variables (predictors). Computing the model gives the following results. According to the model, location, room type, price, minimum nights, and reviews per month are statistically significant (α < .0001) and are likely predictors for an existing availability. It is indicated by the stars at the end of these rows. Three stars indicates that α < .0001, and one star indicates that α < .05. This shows that the result is accurate more than 99.9% and 95% of the time, respectively.



```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)    
(Intercept)                   3.814e+02  6.490e+01   5.877 4.18e-09 ***
neighbourhood_group_cleansed  4.885e-01  5.253e-02   9.301  < 2e-16 ***
latitude                     -1.369e+00  7.882e-01  -1.737   0.0824 .  
longitude                     4.501e+00  8.374e-01   5.374 7.68e-08 ***
room_type                     6.642e-01  6.814e-02   9.746  < 2e-16 ***
price                         4.305e-03  2.151e-04  20.015  < 2e-16 ***
minimum_nights                5.949e-03  5.436e-04  10.944  < 2e-16 ***
number_of_reviews            -1.978e-03  2.061e-03  -0.960   0.3372    
reviews_per_month            -4.083e-01  7.830e-02  -5.215 1.84e-07 ***
calculated_host_listings_count -2.457e-03  1.030e-03  -2.385   0.0171 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 2.2.2 Model Summary

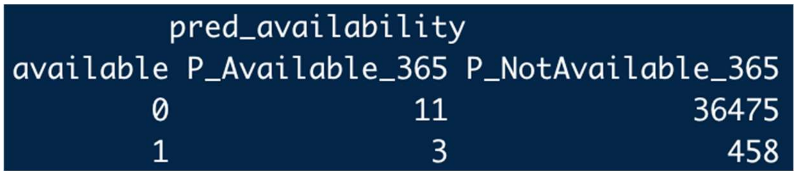
After constructing the logistic regression model, it was best to choose the predicted percentage of the outcome based on other variables in the data table. In the table below, the zeros and ones in the “availability_365” variable are whether the Airbnb will be available for the year. One indicates yes, and zero indicates no. The “predicted_percentage” variable was created to show the percentage that the algorithm calculated to predict whether this particular Airbnb rental will be available for the year and the “pred_availability” variable is the result of the calculation. Sometimes these numbers will not line up since they are predicted values and will show a probability of whether an Airbnb will be available 365 days out of the year or not. Anything that is over .50 will be predicted as yes and anything below .50 will be predicted as no.



```
# A tibble: 36,947 × 14
   availability_365 available predicted pred_availability
              <dbl>     <dbl>     <dbl> <chr>
 1                0         0  0.00498  P_NotAvailable_365
 2                1         1  0.00967  P_NotAvailable_365
 3                0         0  0.0128   P_NotAvailable_365
 4                0         0  0.000458 P_NotAvailable_365
 5                0         0  0.00317  P_NotAvailable_365
 6                0         0  0.00339  P_NotAvailable_365
 7                0         0  0.00455  P_NotAvailable_365
 8                0         0  0.00296  P_NotAvailable_365
 9                0         0  0.00690  P_NotAvailable_365
10                0         0  0.00227  P_NotAvailable_365
```

Table 2.2.3 Prediction Summary

According to the table below, the algorithm for the prediction correctly predicted that an Airbnb was “not available” for 36,475 places. It incorrectly predicted “not available” for 11 of those Airbnb's were available. The algorithm incorrectly predicted “not available” for 458 Airbnb's that were available, and it correctly predicted three Airbnb's that were available for 365 days. This can be seen graphically in Table 2.2.2.



```
         pred_availability
available P_Available_365 P_NotAvailable_365
        0              11              36475
        1               3                458
```

Table 2.2.4 Logistic Regression Model Confusion Matrix

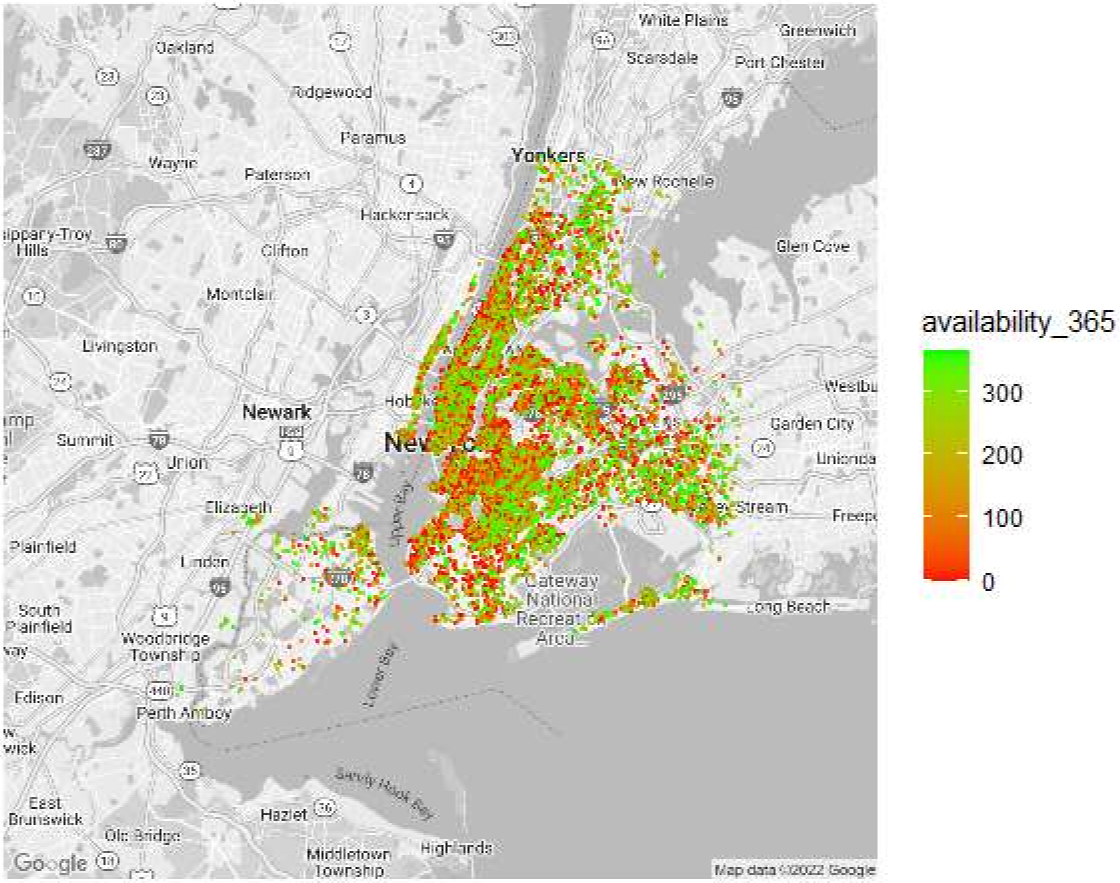


Figure 2.2.5: Map of Available Locations 365 Days of the Year

3. Analysis

3.1 Decision Trees

Decisions Trees are used for binary classification with supervised machine learning. This method is used to show how the machine is making decisions by creating a graphical tree representation that shows where splits are made, and at what values. It uses an inverted tree-like structure to model the relationship between independent variables and a dependent variable. Decision Trees are used to predict future outcomes from past data. The visual representation is literally a root tree which is the starting decision node and is connected to leaves by branches, holding a collection of nodes that are connected to other nodes. These trees are interpreted by the structure of the decision tree as a set of rules or guiding principles.

Below lies a series of decision trees for labeling prices for each borough of NYC Airbnb. According to the calculations of the Decision tree, the probability of the price depends on the room type, the minimum nights that the renter needs to stay, and what the exact location of the Airbnb is. All the nodes at the bottom of the tree represent whether an Airbnb will be inexpensive, whether it is a great price in terms of the market, or expensive in comparison to other rentals. These values were chosen based on the number that represents the quartiles in the earlier box plots.

For example, in the Bronx, if the Airbnb room type is either an entire home, private room, hotel room (traditional hotel rooms are often listed on Airbnb when there is excess stock), or shared room, and the minimum nights is greater than two and a half nights, then it will most likely be inexpensive, if the minimum nights are two and a half nights or less, it will be a much larger price in the Bronx. If none of these conditions are met, then the Airbnb's that will be available will most likely be expensive to rent.

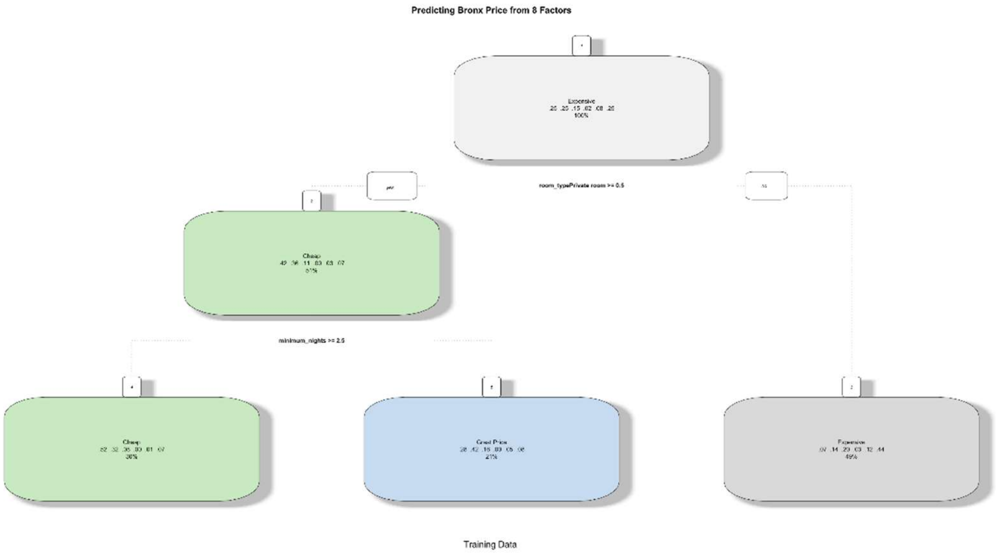


Figure 3.1.1: Decision Tree Predicting Price in the Bronx

In Brooklyn, if the Airbnb room type is either an entire home, private room, hotel room, or shared room, and the minimum nights is greater than five and a half nights, then it will most likely be cheap, if the minimum nights are five and a half nights or less, it will be a great price in Brooklyn. If none of these conditions are met, then the Airbnb's that will be available will most likely be expensive to rent.

In Manhattan, if the Airbnb room type is either an entire home, private room, hotel room, or shared room, then it will most likely be a cheap rental. If the room type is not any of those, and the longitude is less than -74 nights or less, which is toward Times Square (a very popular tourist destination in addition to being considered the heart of NYC), the rental will be expensive.

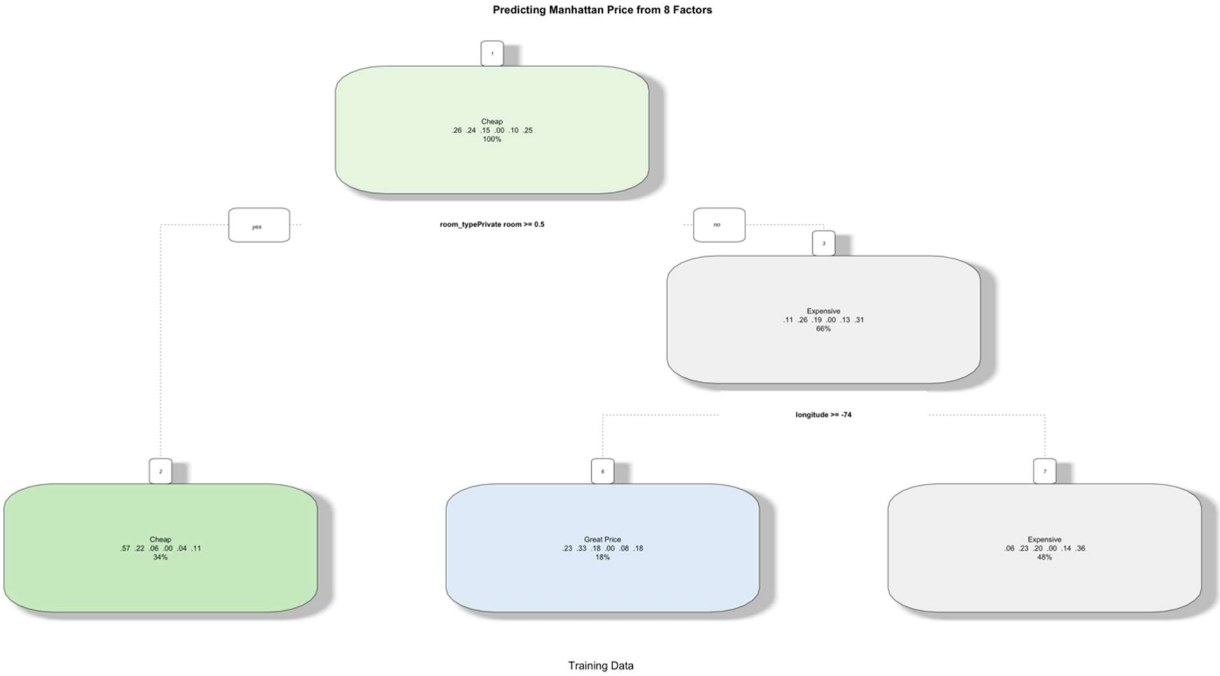


Figure 3.1.2: Decision Tree Predicting Price in Manhattan

In Queens, if the Airbnb room type is either an entire home, private room, hotel room, or shared room, then it will most likely be an inexpensive rental. If the room type is not any of those, and the longitude is less than -74 nights or less, which is toward Times Square, the rental will be expensive, like those found in Manhattan.

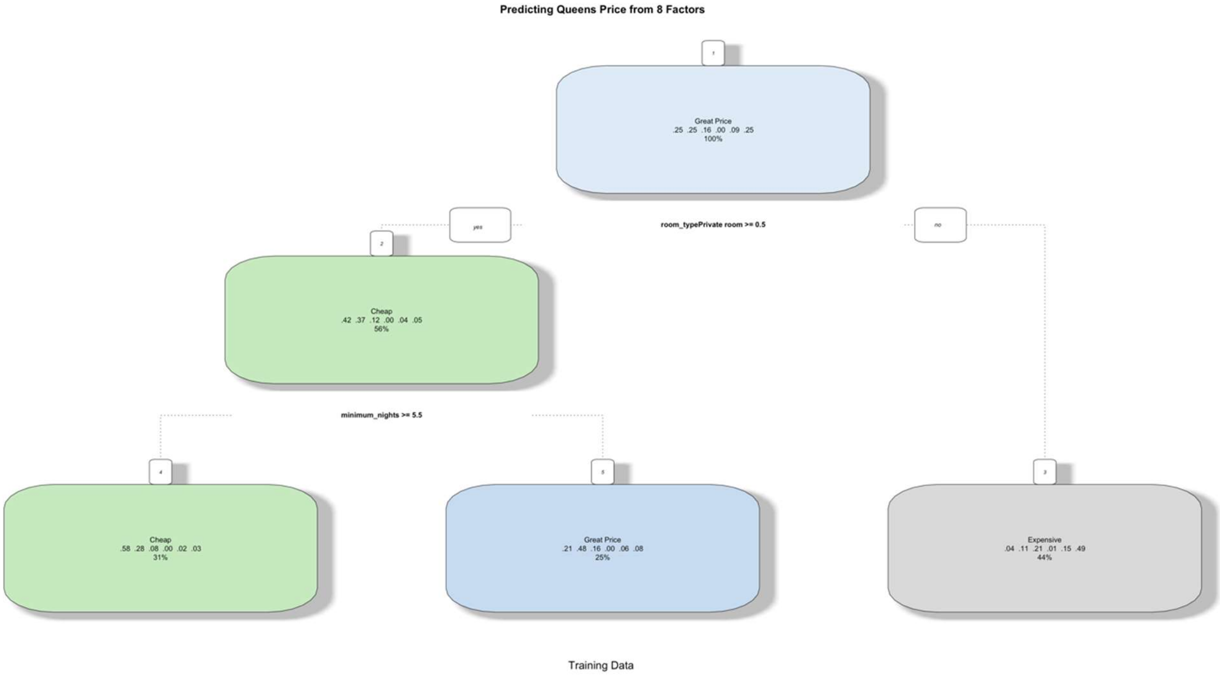


Figure 3.1.3: Decision Tree Predicting Price in Queens

Finally in Staten Island if the Airbnb room type is either an entire home, private room, hotel room, or shared room, then it will most likely be an inexpensive rental, otherwise, the rental will be expensive. As seen in previous illustrations, Staten Island has the fewest options in terms of rentals on Airbnb. This could be because there are fewer options available for tourist type destinations. Additionally, it is known as a borough that many people who work in the city use to live in daily and commute from. As a more working-class area one is less likely to have a rental property which is not occupied.

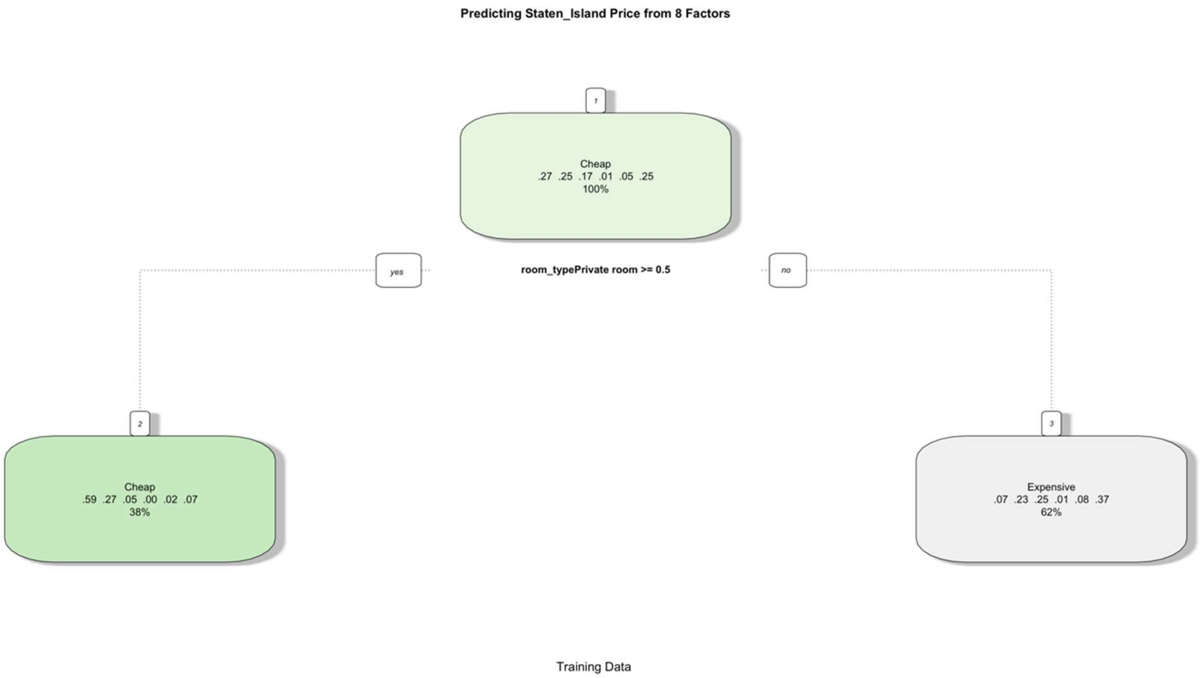


Figure 3.1.4: Decision Tree Predicting Price in Staten Island

The following decision trees were based on the merged reviews and listings dataset. The first decision tree is based on the sentiment of the review that was written on a specific listing. There were three possible different outcomes that a review could be determined, either negative, positive, or neutral. There were four different variables that the model was trained on, the positive sentiment score, the negative sentiment score, the neutral score and the compound score, all numeric values that fall between -1 and 1. The decision tree shows the classification of a review as either a positive, negative or neutral review, where positive is represented as 1 (red), neutral is represented as 2 (gray), and negative is represented as 3 (gray).

The logic of the decision tree is based primarily on the compound score that is associated with each review. If the compound score is greater than 0.05 then the review is automatically characterized as a positive review. If the compound score is not greater than 0.05, then the review is either negative or neutral. The review is neutral if the compound score falls between -0.05 and 0.05. The review is characterized as negative if the compound score falls below -0.05. The result of the model shows that 89% of the testing set were determined to be positive reviews. 8% of the testing set were neutral and 3% of all the testing set were determined to be negative. The classified results returned an accuracy of 1 for this model.

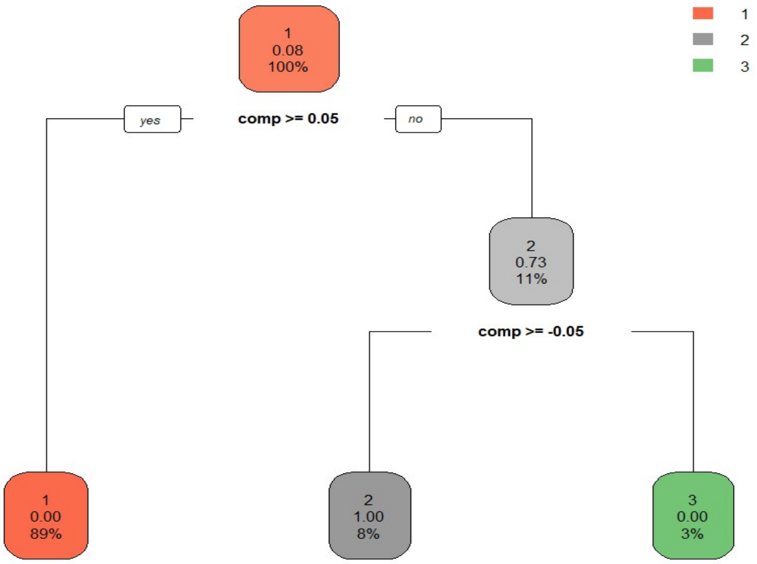


Figure 3.1.5: Decision Tree Predicting Sentiment of Review based on Sentiment

3.2 Kmeans Airbnb Pricing Groups

The price of Airbnb is usually one of several factors in deciding to book or not. Grouping Airbnb by price ranges can help streamline the decision process. Using a Kmeans algorithm the Airbnb prices are grouped into three categories (k=3): Expensive, Average, and Inexpensive. Visualizing the pricing groups by the borough, the Manhattan borough has the biggest groups of Inexpensive and Average priced Airbnbs'. With many of the attractions of New York City found in Manhattan, it makes sense that the customers could find large numbers of affordable Airbnb's in this borough. Further supporting this model is the lowest number of Airbnb's for all three categories is found in the Staten Island borough, which is the furthermost borough to Manhattan.

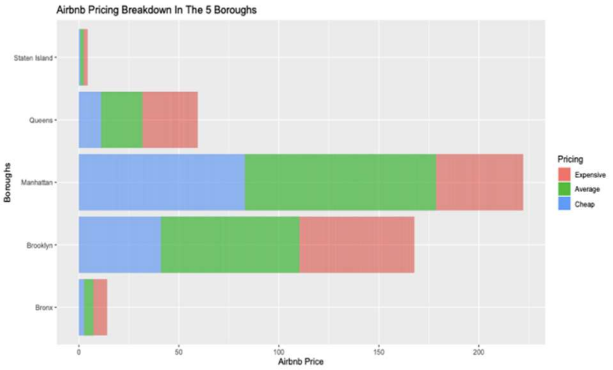


Figure 3.2.1: Airbnb Pricing Groups by Kmeans (k=3)

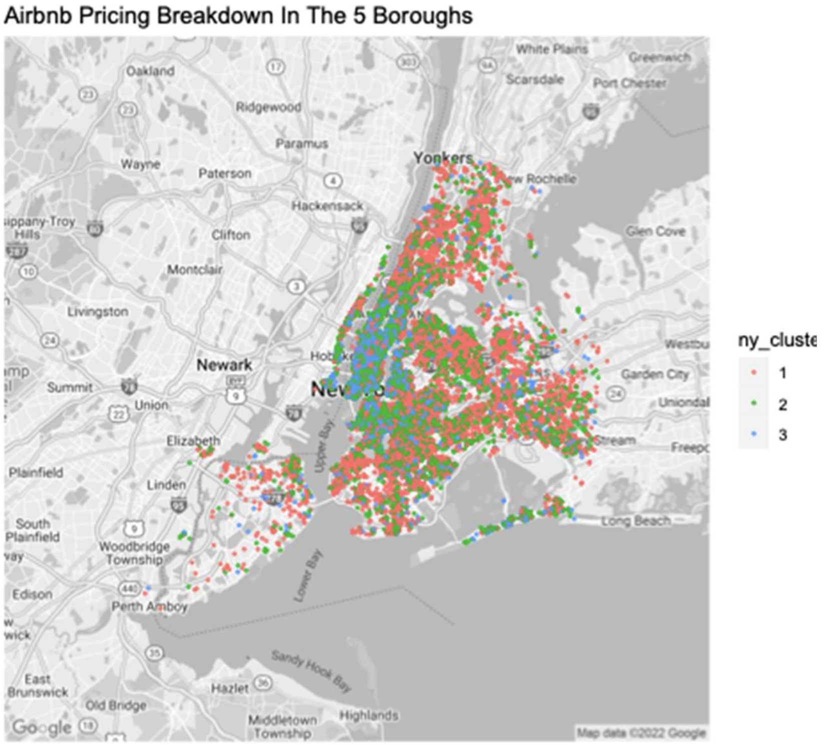


Figure 3.2.2: Airbnb NYC Map Cluster Kmeans=3 Groups (1=Expensive, 2=Average, 3=Cheap)

4. Conclusion

Airbnb has come a long way since its beginnings in 2007, where its name comes from a comparison of smaller Bed and Breakfast boutique hotels in combination with their founders renting out an air mattress to travelers to help them cover their rent. This idea from its humble beginnings has sprung into over six million individual listings in over 220 countries and 100,000 cities in the United States alone. Airbnbs offer a much more “home-like” environment through their unique listings that include everything from standard, cookie-cutter hotel rooms to tree houses. One can choose anything from an entire home to just a couch, and homeowners can offer part if not all of their home to people they approve of in exchange for prices agreeable to both parties. This makes travel more affordable for many who may not have had the opportunity to visit a place like NYC in the past, especially if they are willing to be flexible and share a space.

The initial data analysis showed that, as suspected, shared space is less expensive than non-shared space in NYC (and likely other places as well). The more private the area or the building, the higher the price will be. Amenities tend to add to the price tag, in addition to popular locations and being in the more tourist-oriented boroughs. Manhattan, ripe with opportunities for tourism, is the highest-cost borough, as shown by the analysis of this Airbnb data. That is followed by Brooklyn and Queens being slightly less expensive.

NYC is one of the more highly regulated areas of Airbnb, with more laws affecting who can and who cannot lease out a space, and which spaces cannot be rented. The owner is required to be on premises a certain number of days a year, and during the stay of most renters. This helps to deter problems with owners renting to people who might use the space as a true “party-house” and not be around to deal with the consequences of those choices. These regulations may make it harder for homeowners to make a profit but help maintain the community atmosphere as well as the turnover rate which can be disruptive to full-time tenants. The fact that more Airbnb rentals are available than affordable apartments for rent makes it likely that more regulation will be seen in the future compared to less.

Through the multiple types of data visualization, analyses, and machine learning algorithms many different things make up the price of an Airbnb rental. It is not just the borough in which it resides, but also the amenities and location in relation to different tourist destinations that can affect the price at any given time of rental. For example, those rentals close to the center of business and trade in the city as well as popular tourist destinations will carry a steeper price tag than those in out-of-the-way locations that are a hop, skip, and jump from local attractions. The willingness to give up one’s privacy and share space will not only be easier on the wallet but possibly lead to a more unique and genuine experience in NYC compared to more traditional hotel rooms.

As NYC continues to impose restrictions on where Airbnb can be located, the company will need to adjust and change its marketing strategy to deal with these changes. With consumers having more choices than ever when visiting the city, it is important to be aware of what retains customers, truly wows guests, and keeps them coming back for more in the future. Since the relative location of local tourists' attractions is correlated with the likelihood of an Airbnb space being rented, a program that gives discounts to tourists who use Airbnb to visit these locations could be a good marketing strategy to match tourists with their desired destinations during their trip. Discounts on local transportation, food delivery, and an insider’s guide to the area could keep the reviews coming in as positive for those that make money from their rentals, and with the advent and investiture in data from Airbnb the company allows for more informed purchasing and maintenance of rental properties. Investing in data is investing in the company's future, and right now that future is truly bright.