NYC Airbnb Project

David Mao[1, a] and Sergio Jimenez[2, b]

[1] *Student ID: 914766922*

[2] *Student ID: 915205181*

Airbnb is an online service that sets up a marketplace that links travelers seeking affordable lodging with hosts who provide them. However, in such a market, not only are there many diverse options being offered by different hosts, but guests also all have different personal needs and desire different services. In such a diverse and volatile market, it becomes important for Airbnb to understand how both hosts and guests behave. The vast amount of data at Airbnb's disposal can be instrumental in doing so. Through careful data analysis of past listing data, Airbnb can better understand the rental marketplace characteristics and make better strategic decisions, such as understanding where high concentrations of listings are, where hot spots of high prices are, and the behavior of super hosts.

[a] dlmao@ucdavis.edu

[b] ssjimenez@ucdavis.edu

I. INTRODUCTION

The dataset we will use for this analysis contains information of Airbnb listings in New York City for the year 2019. New York City is by far the most populous city in the United States, and is a hotspot for tourism. In our report, we seek to determine what trends exist between location and listings. We additionally attempt to find some important behaviors Airbnb hosts and tenants exhibit. First, we will make several geographical visualizations of the listing data and observe important trends. We then use spatial data analysis with use of Moran's I to find local patterns of spatial autocorrelation. Lastly, we systematically analyze all the variables given in the dataset to see if there are any insights on host and tenant behavior. Some important insights we found were that hot spots of expensive listings and neighborhoods are found in Manhattan and Northern Brooklyn. Finally, we find several insightful trends about AirBnb hosts and consumers, such as listings with high availability come from hosts with more than one listing, the most popular listings are private, affordable, and close to airports.

II. DATA ANALYSIS

A. Data Preparation

Our dataset include 16 variables from 2019 Airbnb listings in NYC, with each row corresponding to a unique listing. Tab. (I) shows the variables that we need to further deal with. We first observe that last_review and reviews_per_month both include 10052 null values. These null entries directly correspond to the same rows and rows that also have a number_of_reviews value of zero. With this insight, we can deal with null last_review values by replacing it with the earliest last review date in the whole dataset for the sake of our analysis, and null reviews_per_month by replacing it with zero. Additionally, last_review comes in a string format, so we can replace it with a date time format. However, date time format is also not very insightful so we will further alter this variable by changing it to an integer based on the difference in days between the latest review in the whole dataset and each latest review value. Finally, there are some null values for name and host_name, but we can ignore these values, as they do not impact our analysis.

B. Geovisualization

We start our analysis by trying to find any important insights based on the location of each rental unit. Our dataset includes four variables that have to do with the location of each rental. Specifically, neighbourhood_group denotes which New York City Borough, specifically Bronx, Brooklyn, Manhattan, Queens, or Staten Island, each listing belongs to, neighbourhood denotes the city neighborhood each listing belongs to, and longitude and latitude combine to denote the exact coordinates of each listing.

We first start by analyzing what is characteristic of each of the five boroughs. Fig. (1) shows a map of New York City's five boroughs along with a point denoting the noting the location and the color of each point denoting the price. With this map, we can see some obvious characteristics of each borough. Manhattan seems to be the most expensive borough, and has a large concentration of listings. Brooklyn looks to be the next most expensive borough, with a large concentration of listings among the north side near the border with Manhattan. Queens seems to have certain sections that are expensive but has much more concentrated sections of cheaper listings. Staten Island has a very small amount and sparsely located listings. Bronx seems to have the cheapest listings overall. A boxplot of the prices for each borough in Fig. (2) confirms what we observed about the prices. Manhattan is the most expensive, followed by Brooklyn. Bronx is the least expensive.

Looking at just the boroughs may be too general though, and we can see that there exists different behaviors within each boroughs. We can narrow our focus by looking at the neighborhoods next. As a special note, since the neighborhood maps of our dataset do not match the official list of neighborhoods provided by most maps, for the rest of our analysis, we will be using the map's neighborhood names and assigning listings to these neighborhoods based on their coordinates. Fig. (3) shows the five neighborhoods that have the most listings, with most of the top number of listings occurring in Manhattan and Northern Brooklyn. Since Airbnb is a tourist catered product, we attempt to see if tourism is a main indicator of why these neighborhoods have high listings. Fig. (4) shows ten of the most common tourist locations (TripAdvisor, 2021). Nearly all of the locations are located in Manhattan, which is in tune with the large amount of listings and the price of listings in this area. We also would like to know what neighborhoods are on average expensive. A map of this is shown in Fig. (5) while a barplot of both the most expensive and highest listing neighborhoods is shown in Fig. (6). Ignoring the clear outlier (which is explained in Tab. (II)), Manhattan clearly has the most expensive listing prices, likely due to its proximity to the cultural centers and tourist locations as well as the nature of the borough. However, only two of the most expensive neighborhoods are also present on the list of neighborhoods with the most listings. In fact, a comparison of Fig. (3) and Fig. (5) shows that the neighborhoods with the most listings are also relatively cheaper than the surrounding areas. This perhaps shows us that there is a high demand for relatively cheaper housing that are still within proximity of iconic locations and cultural centers.

Another geographical point of interest is the location of the New York City Subway system. The subway system is a cheap tool that allows tourists to easily travel across the city. Fig. (7) is a map of all the New York City subway lines and stops. There are many lines and stops centered around downtown Manhattan, but there are also many lines that reach all over the city, with the exception of Staten Island. A rental with easy access to a subway may be more desirable for tourists, as with quick access means easy access to the rest of the city. To analyze this, we will look at all the listings that are within 100 meters of a subway stop. Fig. (8) shows the visualization of the listings within 100 meters of a subway station and Fig. (9) shows boxplots of the distribution of prices of listings within 100 meters and further than 100 meters from a subway station for listings under 500 dollars. Comparing the two distributions, the mean of the listings within 100 meters is around 30 dollars more than those further than 100 meters. Our distributions are not normal, but we can use a one sided Mann-Whitney U Test to test if the sample distributions are equal. We obtain a p-value less than 0.001, so prices of listings near subway stations is significantly higher.

C. Exploratory Spatial Data Analysis

From our visualizations, it seems clear that is likely there exists spacial autocorrelation, as similar priced neighborhoods and listings are located close to each other. In this section, we will quantify spacial autocorrelation using Moran's I. Knowing this will help us understand precisely where spatial clusters exist. Moran's I statistic is given by:

$$I = \frac{n}{S_0} \frac{\sum_i \sum_j w_{i,j} z_i z_j}{\sum_i z_i^2}, \tag{1}$$

Here, $z_i = x_i - \bar{x}$, which denotes how far a feature is from its mean, $w_{i,j}$ is the spatial weight between i and j, and S_0 is the sum of all spatial weights $w_{i,j}$. When i and j are related and are relatively large or small, then the resulting value will be positive. If i is small and j large, then the resulting value is negative. This value is normalized to be between +1 and -1. A high negative value would indicate dispersion, a value close to zero would indicate no relationship, and a high positive value would indicate clustering (Getis and Ord, 1992). To test the significance of Moran's I, with the null hypothesis that the data is randomly distributed, the z-value can be calculated as:

$$z_I = \frac{I - \mathbf{E}(I)}{\sqrt{\mathbf{E}(I^2) - \mathbf{E}(I)^2}}, \tag{2}$$

where $\mathbf{E}(I) = \frac{-1}{n-1}$ (Getis and Ord, 1992).

Moran's I helps us quantify global autocorrelation, but does not tell us where locations are clustering. We would like to know about local autocorrelation, which given a location tells us how much of the surrounding locations are similar. Local Indicators of Spatial Associaion (LISA) uses a measure of local spatial autocorrelation and determines four things: where clusters of high values are found (HH), where clusters of low values are found (LL), where high outliers are found (HL), and where low outliers are found (LH) (Anselin, 1995). For our measure of local spatial autocorrelation, we will use Local Moran's I, which is given by:

$$I_i = \frac{x_i - \bar{x}}{S_i^2} \sum_{j,j \neq i} w_{i,j}(x_j - \bar{x}) \tag{3}$$

where $w_{i,j}$, x_ij, are defined the same way as Eq. (1). Additionally, the z-value for Local Moran's I is also computed the same way as Eq. (2), except using $\mathbf{E(I_i)}$. When a location has a positive Local Moran's I, it is part of a cluster. When it has negative Local Moran's I, it is an outlier (Anselin, 1995).

We can calculate Moran's I to see if there is spatial autocorrelation in the price of neighborhoods. First, we need to define how to find $w_{i,j}$ for different neighborhoods. Since neighborhoods represented as shapes, we can let $w_{i,j} = 1$ if the shape of two neighborhoods share a vertex and 0 otherwise (Queen contiguity distance). We calculate Moran's I to be 0.41 for average neighborhood price, showing that similarly priced neighborhoods in New York City are often clustered together. We also find that the p-value for the Moran's statistic is $p_I < 0.001$, indicating that the clusterings shown by the neighborhoods are significant. The Moran Scatter plot in Fig. (10) visualizes this autocorrelation. We next plot the local Moran statistic to see where the clusters occur and look for any outliers. Fig. (11) shows the results of LISA at a significance level of 0.05. We can clearly see that there is a large cluster of high-priced neighborhoods in Manhattan and northern Brooklyn, while Bronx has a large cluster of low-priced neighborhoods. Several high priced outliers exist across the whole city, with a good amount of them in Queens.

We repeat this process for individual rentals based on their coordinates. Since we are not dealing with shapes anymore, we will calculate weights by taking the one hundred nearest points as neighbors (KNN). We similarly get a positive Moran's I of 0.36, which has a significant p-value $p_I < 0.001$. This is visualized in Fig. (12). LISA results at a significance of 0.05 are shown in Fig. (13). High price clusters are concentrated in Manhattan and Northern Brooklyn while there are low price clusters scattered everywhere else. High price clusters also occur along the Rockaway Peninsula (island southside of Queens), which can be explained by the abundance of beach houses there. Additionally, we can quantify what we observed in Fig. (3), that there exists low high clusters in the places that also have the highest number of listings.

D. Price

Being a peer-to-peer market, Airbnb supposedly creates a marketplace where small time renters can compete with more traditional vacation rental providers, and an important aspect of this is price. Observing a graph of the listing prices in Fig. (14), we can see that the median of the prices is around 90 dollars (See Tab. (III)), which is around the median hotel price of New York City, showing that Airbnb has accomplished creating such a marketplace. We also observe a few outliers in our dataset however, namely very high priced listings. Taking only the most expensive listings (those that cost more than 500 dollars), we look at the most common description words and room types in Fig. (15). We can conclude that these expensive listings are mainly luxury apartments located around Manhattan and Brooklyn.

E. Reviews

There are three variables in our dataset pertaining to reviews, and we attempt to understand their relationship with each other and the rest of the dataset. Fig. (16) shows a pair plot for the three review variables and price. We observe a strong correlation between the three review variables, which is not a surprising observation given that the the more reviews a listing has the more reviews per month the listing likely also has, etc. However, the correlation between reviews and price is observed to be low, which since reviews is an indicator usage, shows that Airbnb consumers are unique in what they value in a rental. However, the highest reviewed values are also some of the cheapest. Looking at the price of listings with at least 400 reviews in Tab. (IV), the median price is a mere 55 dollars. We also look at the common words in the description, like we did the previous section. Fig. (17) shows that these listings are popular due to them being next to JFK or LGA airport, and other factors such as privacy and no cleaning fee.

F. Minimum Number of Nights, Availability, Room Type, and Total Host Listings

These final four variables also provide some important insight in determining both consumer and renter behavior. Examining the minimum number of nights, we observe that while most listings have a low number of day requirements, akin to what a vacation rental is. However, there are noticeable spikes in listings that have 30 day requirements, as shown in Fig. (18). This shows that while the majority of Airbnb consumers are vacation renters, a small portion of them might be using Airbnb to find lodging for longer-term places. Next, given that Airbnb is targeted towards home owners trying to make some money with their home while on vacation, one would expect most listings to have a low year-round availability. However, as Fig. (19) shows, while a high number of listings have low year-round availability, it is not as much as one would think. There is a high amount of full year round available listings, indicating there are full-time renters using Airbnb. We can see next if the availability of the listing is related to whether or not the host has more than one listing. Fig. (20) shows that indeed, the percentage of listings available for lower amounts of time are from hosts with only one listing. Looking at room types, as shown in Fig. (21), shared rooms have the lowest amount offered, indicating that people usually prefer their privacy while renting places, and there are likely less renters willing to have strangers living with them. The price of each room type is not surprising as well, since whole homes should cost more than a single room. Combining both availability and room type gives us some interesting insight as well. Fig. (22) shows us that while whole homes and private rooms are most commonly available for shorter terms, shared rooms are most commonly available year-round, which makes sense given the nature of shared rooms. Also, specifically for whole homes, ones that are available for lower amounts of time on average usually go for a lower price then those year round. This might be explained by people going on vacation have less flexibility and therefor cannot charge as high of a price.

III. CONCLUSION

In conclusion, we found there were several important trends in the locations of our dataset. We were able to find that the densest and most expensive listings and neighborhoods were located in Manhattan and Northern Brooklyn, likely due to New York City's iconic tourist locations and cultural centers being located mainly in Manhattan. Additionally, listings within 100 meters of a subway station were significantly more expensive that those further away. Spatial data analysis concluded that hot spots for pricy listings were found in Manhattan and Northern Brooklyn, with a few low outliers in Northern Brooklyn coinciding with popular areas to live. In terms of host and consumer behavior, based on reviews, we noticed that both affordable and expensive listings had listings with high number and frequency of reviews, a testimony to the diverse tastes of Airbnb users, although the highest reviews did mostly belong to lower priced listings. We found that although most of Airbnb hosts are part-time hosts, a significant amount of hosts rent their properties full time. Most listings are targeted towards short vacation stays,but a small amount listings require long stays, indicating that the users of Airbnb extend beyond Airbnb's intended vacation renters. Additionally, these listings available for longer times also usually had hosts that had more than one listing. We also found that there was a significant difference in prices of the different room types. Finally, shared rooms tend to be available year-round while private rooms and houses are likely available for only small parts of the year.

IV. APPENDIX A: PLOTS

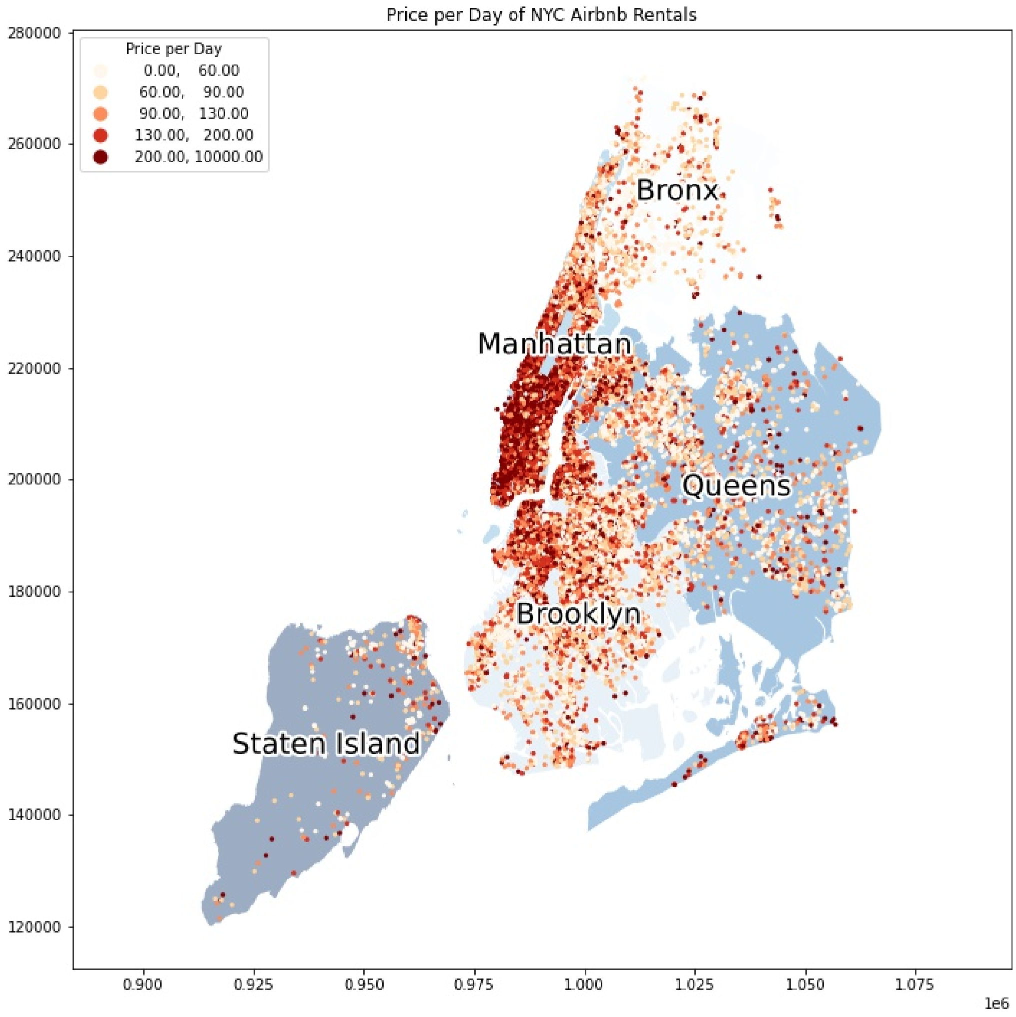
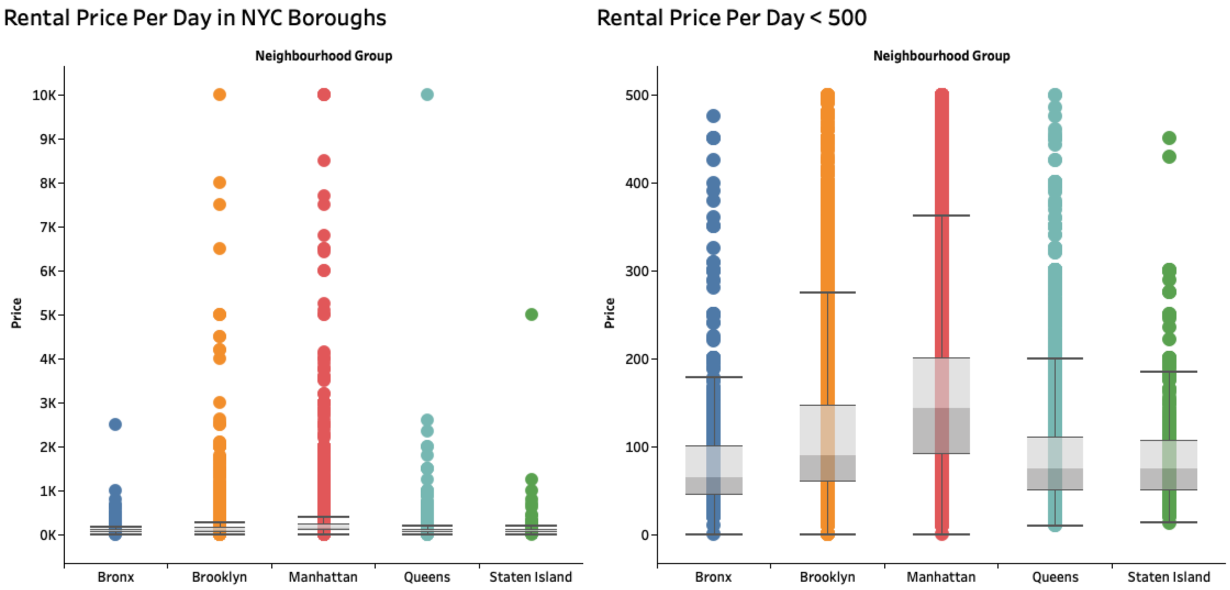


FIG. 1. Neighbourhood Group Price Map

There is both a large concentration of rentals in Manhattan that are also very expensive. Next in both concentration and pricing is Brooklyn, particularly the northern section near Manhattan. Staten Island has the sparsest distribution of rentals. Bronx seems to have the lowest rental cost overall.



Rental Price Per Day in NYC Boroughs

Rental Price Per Day < 500

FIG. 2. Neighbourhood Group Price Box Plots

Manhattan by far has the most expensive Airbnb listings in comparison to the other 4 NYC Boroughs. Brooklyn comes in second, Queens third, Staten Island fourth, and Bronx last.

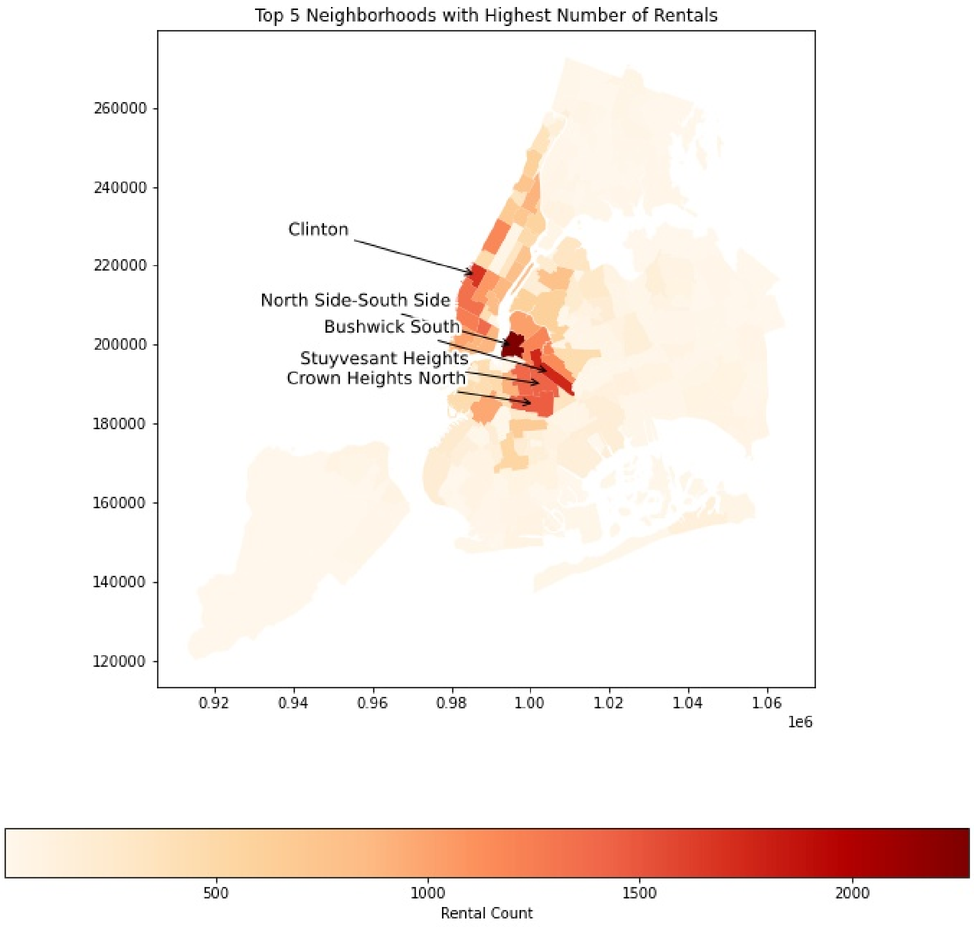


FIG. 3. Neighbourhood Price Map

Neighborhoods with the highest amount of rentals are close to the popular downtown Manhattan, but also from neighborhoods that are relatively cheaper than the surrounding areas.

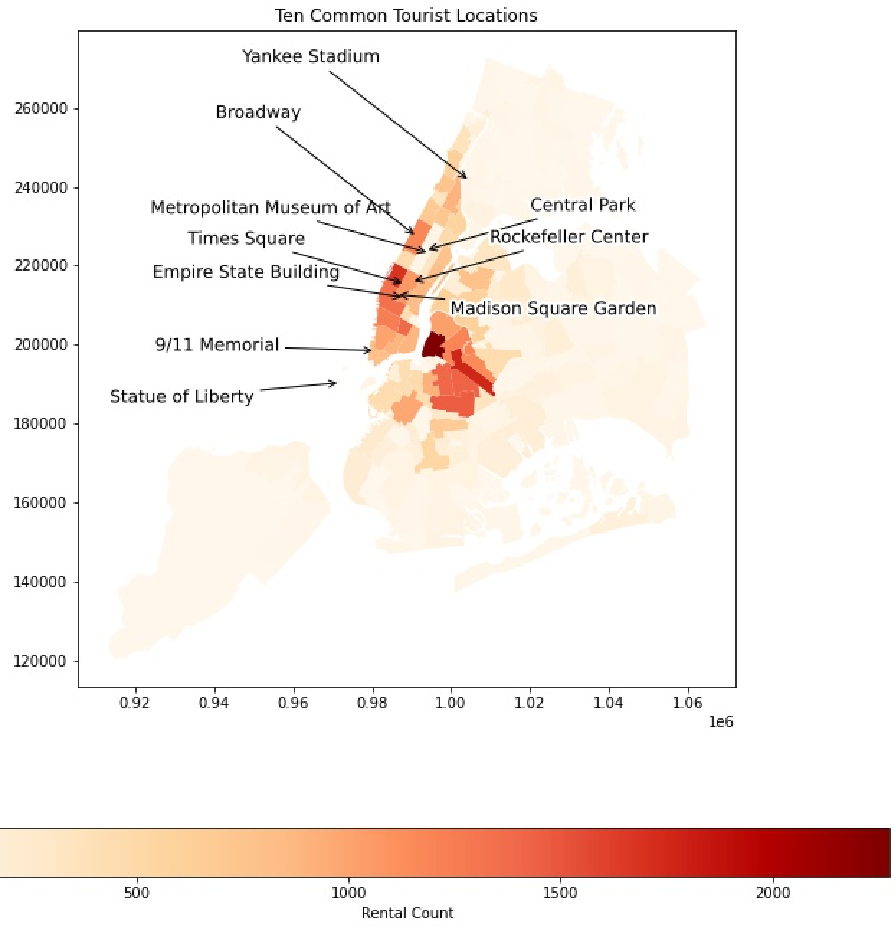


FIG. 4. Neighborhood Rental Counts with Common Tourist

As expected, most of the popular tourist locations are located in Manhattan, with the exception of Yankee Stadium. The neighborhoods offering the most rentals are close to these locations.

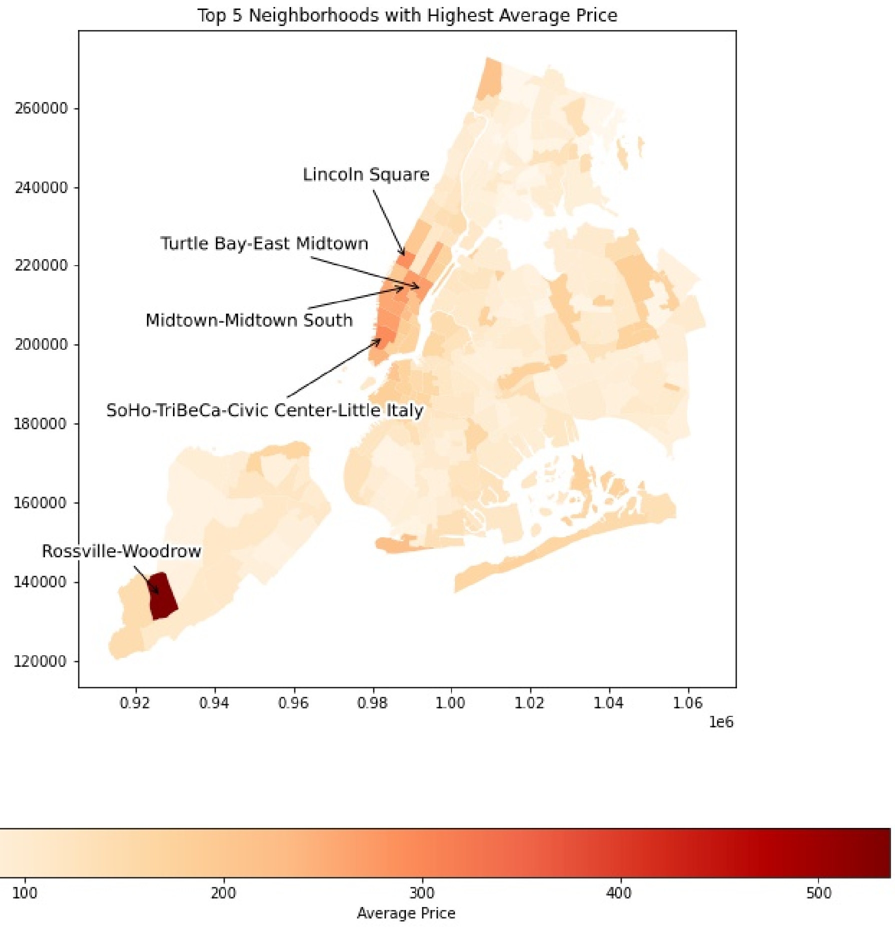


FIG. 5. Neighborhood Locations with the Highest Average Price

The most expensive neighborhoods are found near Downtown Manhattan, with the exception of Rossville-Woodrow, which is an obvious outlier.

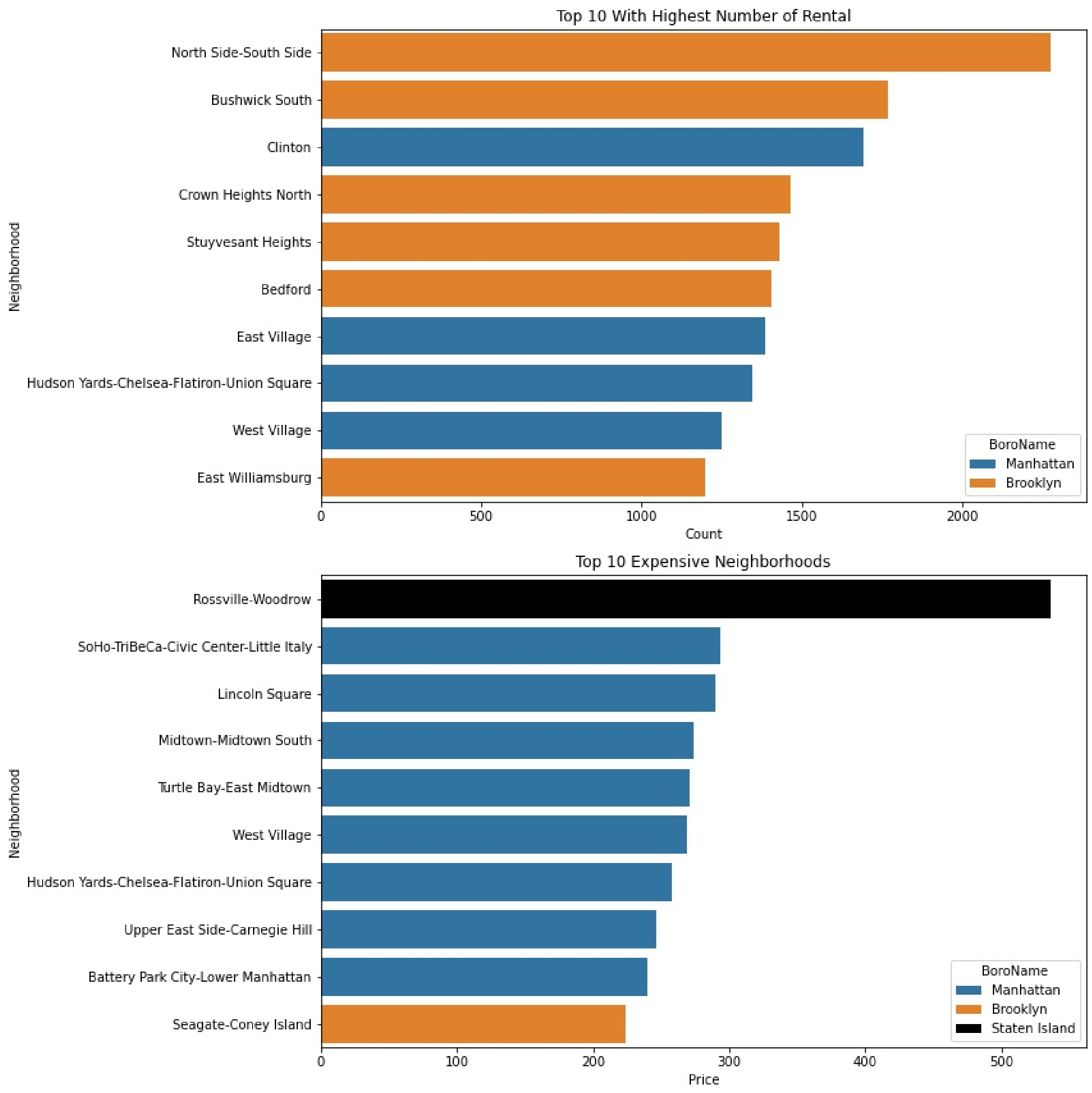


FIG. 6. Most Expensive Neighbourhoods and Highest Number of Listings by Neighbourhoods

It's evident that Manhattan has the most expensive neighbourhoods to rent, barring the one outlier. However, Brooklyn includes many neighborhoods that have a large number of rentals. Coupled with Fig. (3) and Fig. (5), it is apparent that expensive places do not usually have a high number of listings available.

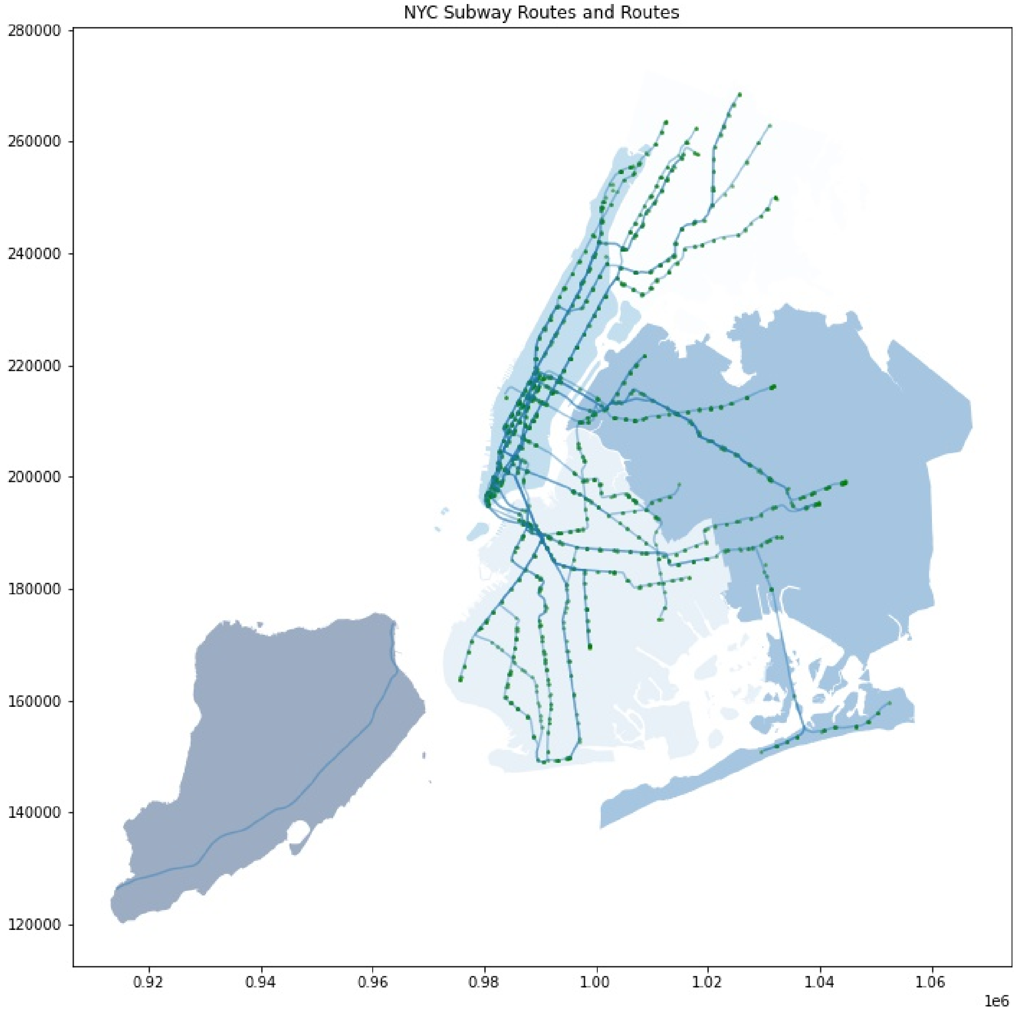


FIG. 7. Map of New York City Subways and Subway Stations

We can see that the subways are concentrated in downtown Manhattan. Other than Staten Island, the subway system seems to comprehensively cover most areas of New York City.

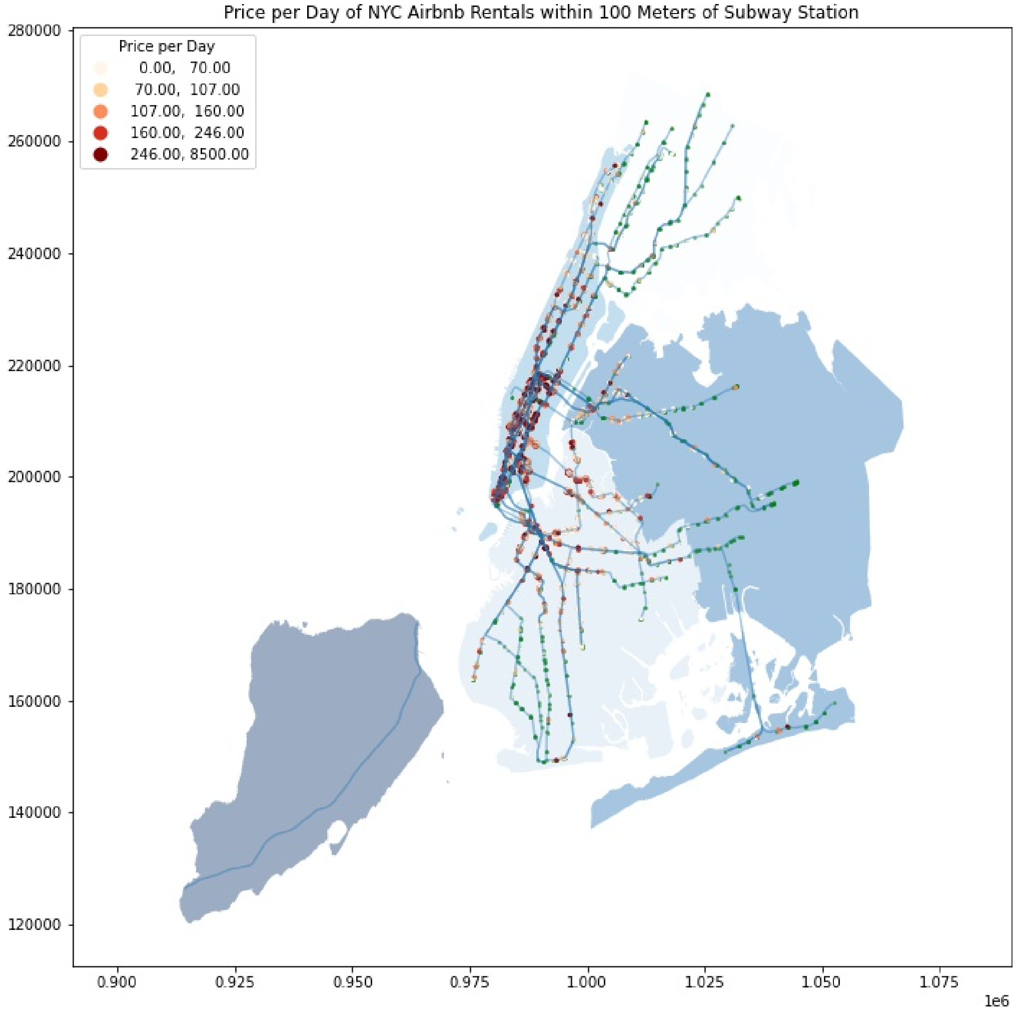


FIG. 8. Map of Listings close to subway stations

A visualization of the listings within 100 meters of a subway station seems to indicate these listings are expensive.

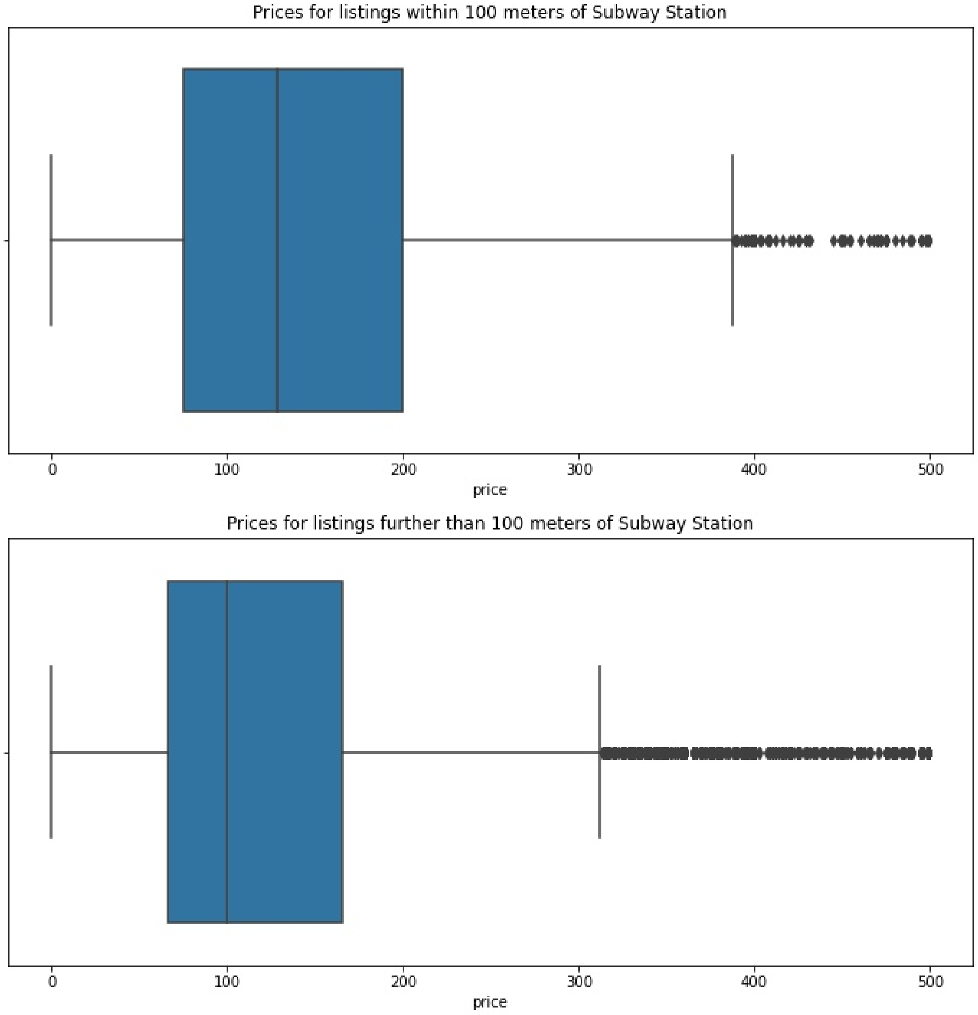


FIG. 9. Box plots of listing prices within and further than 100 meters from a subway station

On average, it seems that listings within 100 meters are more expensive than outside.

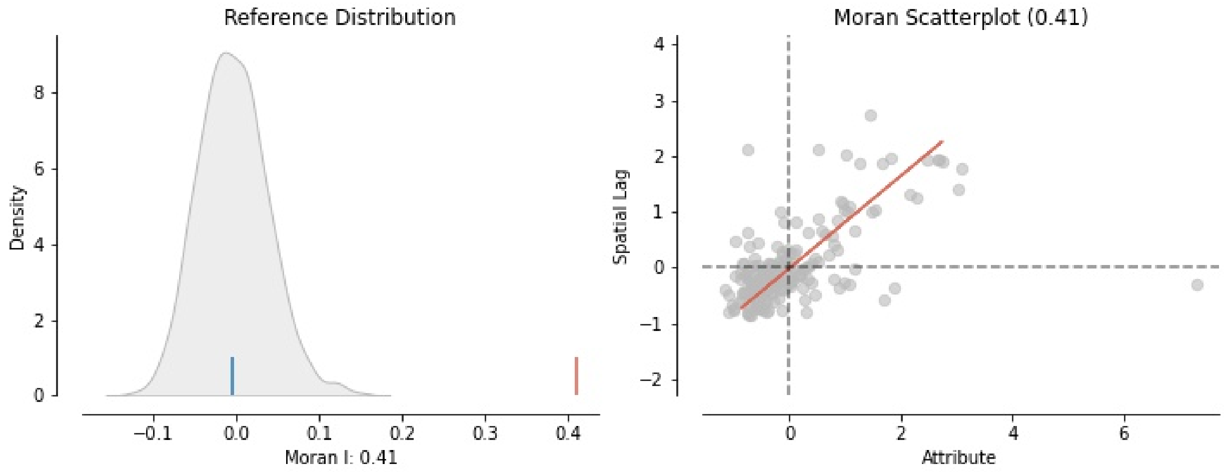


FIG. 10. Moran's I for Average Neighbourhood Price

A positive Moran's I shows that high price and low price neighborhoods tend to cluster together.

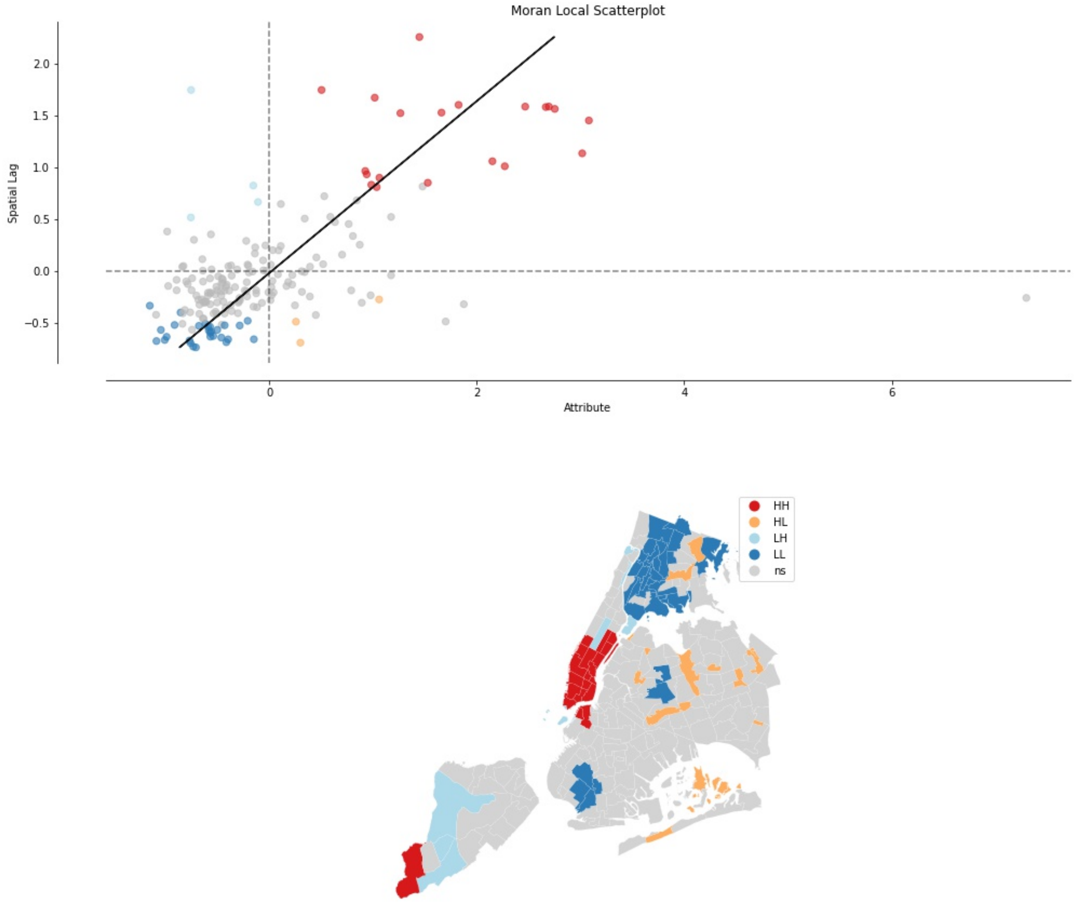


FIG. 11. LISA plot for Neighborhood Average Prices

The local spatial autocovariances shows cluster of high prices in Manhattan and Northern Brooklyn and clusters of low prices in Bronx. Few high price outliers are found mainly in Queens.

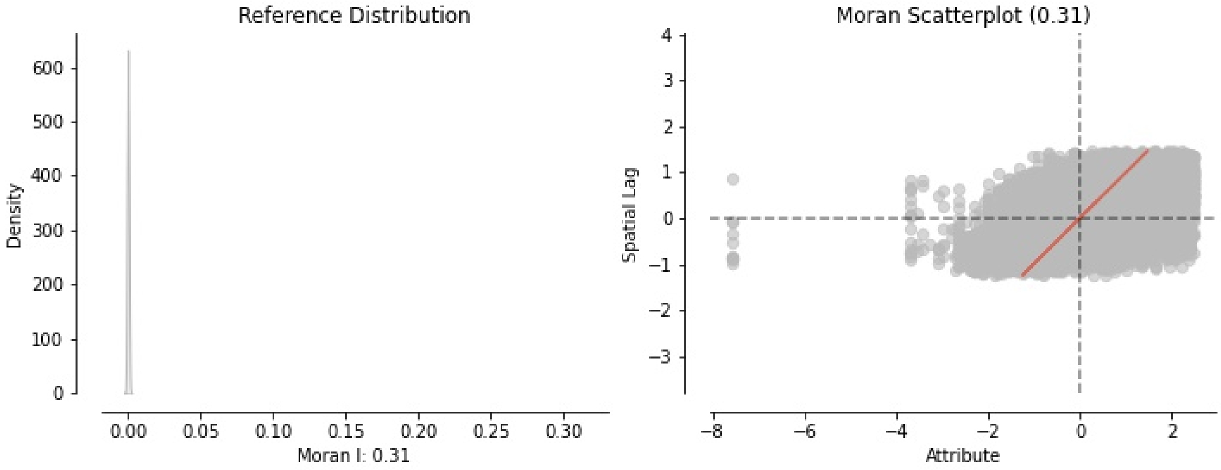


FIG. 12. Moran's I for Rental Location and Price

Moran's I calculated on the 100 closest nearest neighbors shows high spatial autocorrelation among rentals close to each other.

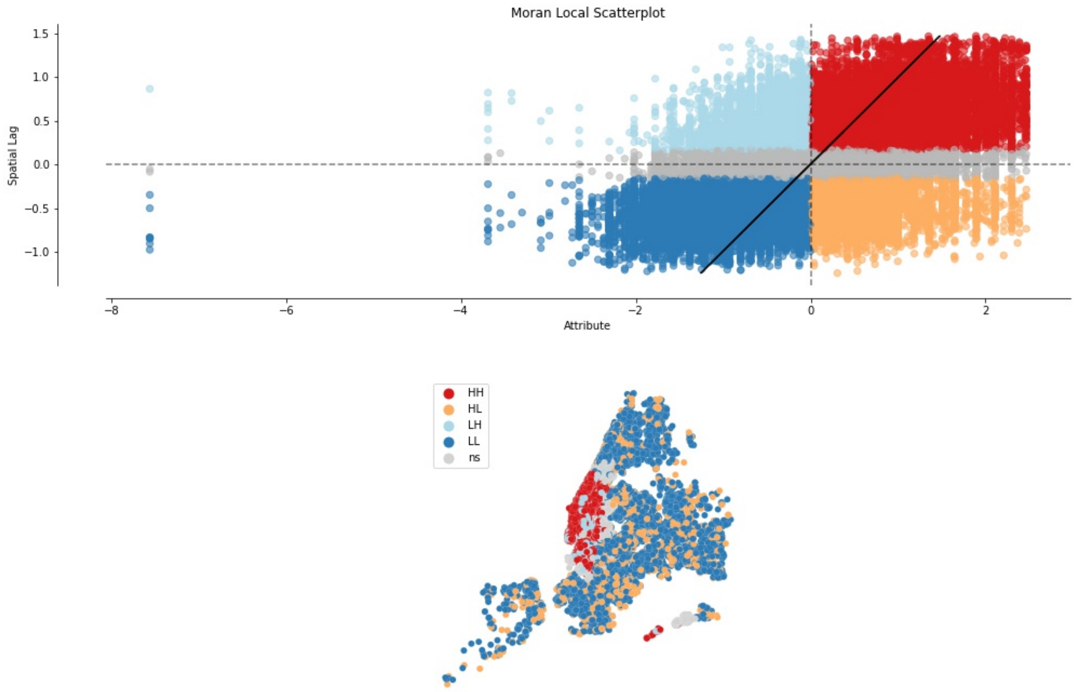


FIG. 13. LISA plot for Rental Location and Price

Clusters of high priced rentals are found mainly in Manhattan and Brooklyn while low priced clusters are found in many places around the city. High clusters are found in Turtle Bay, which correspond to beach vacation homes. Additionally there are noticeable high low clusters in Northern Brooklyn that correspond to the same areas we observe in Fig. (3) that have high listing counts.

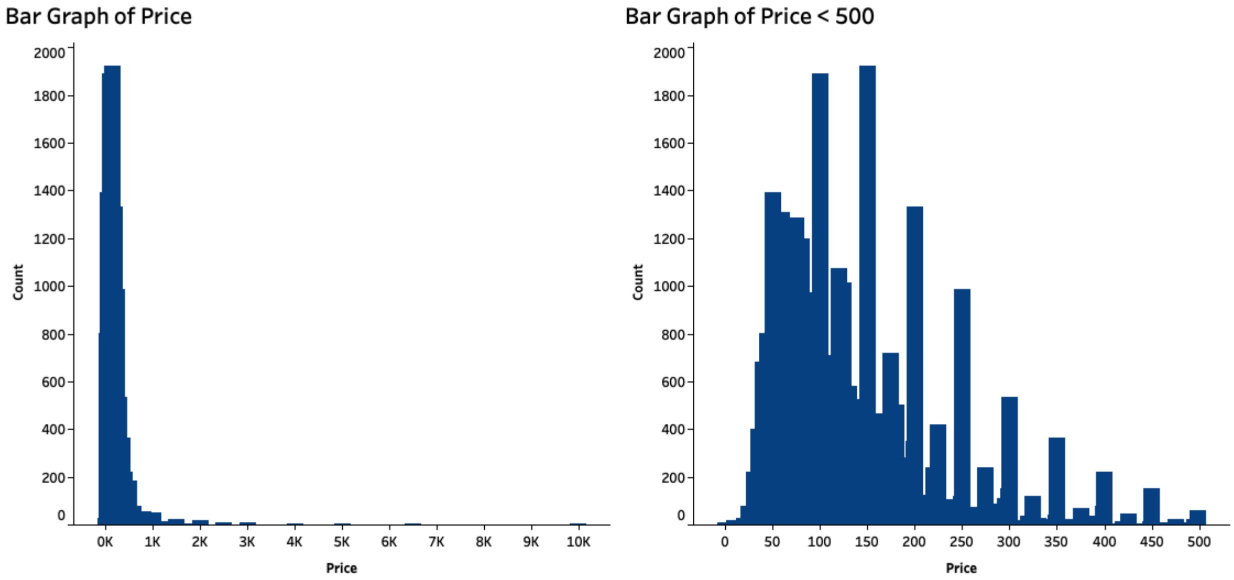


FIG. 14. Bar Graph for Price

Listing Prices are highly skewed right, with the majority of listings falling around 100 dollars a night, which is around the typical rate for hotels.

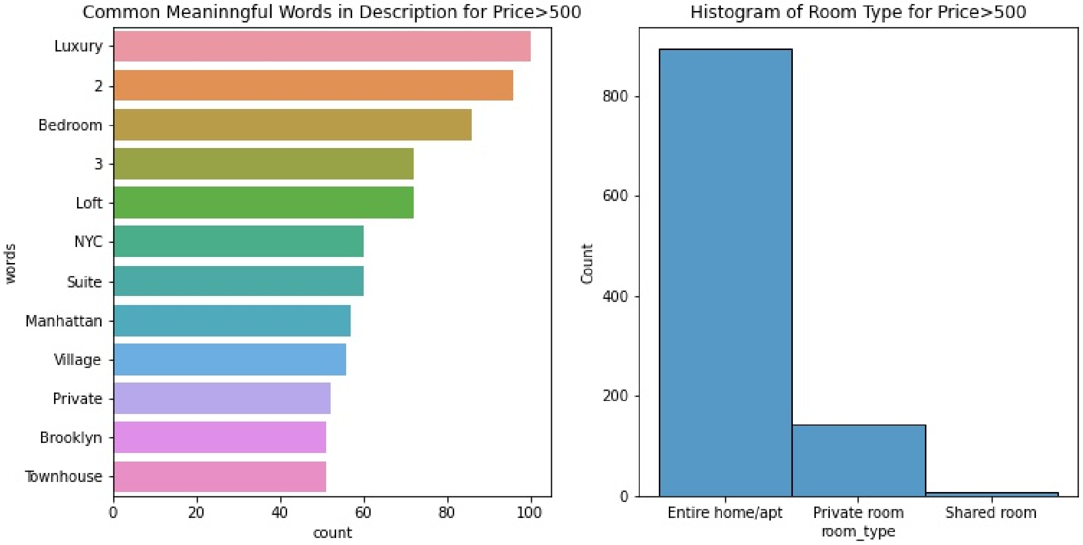


FIG. 15. Histogram of Common Description Words and Room Type for Expensive Listings

The common high words and histogram of the common room types for listings above 500 dollars show that a good portion of expensive listings are made up of luxury suites in Manhattan and Brooklyn.

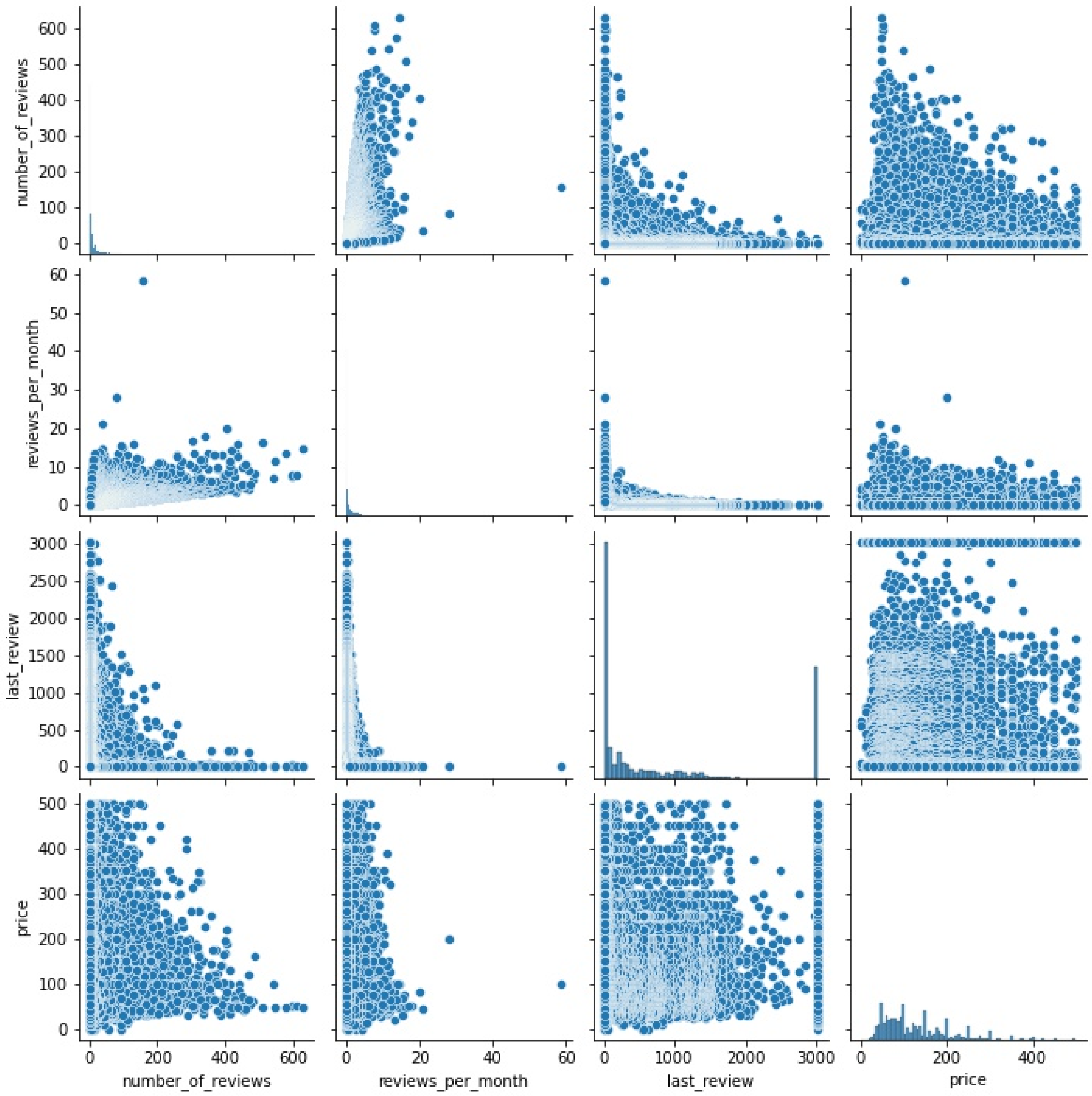


FIG. 16. Pair Plot for Review Variables

Strong correlation is shown between the number of reviews, reviews per month, and last review, which is not surprising since all three variables are based on reviews. Overall correlation of these variables compared to price is weak. However, the listings with the highest amount of reviews and reviews per month have prices at around 70 dollars, which is around the expected price of a vacation rental.

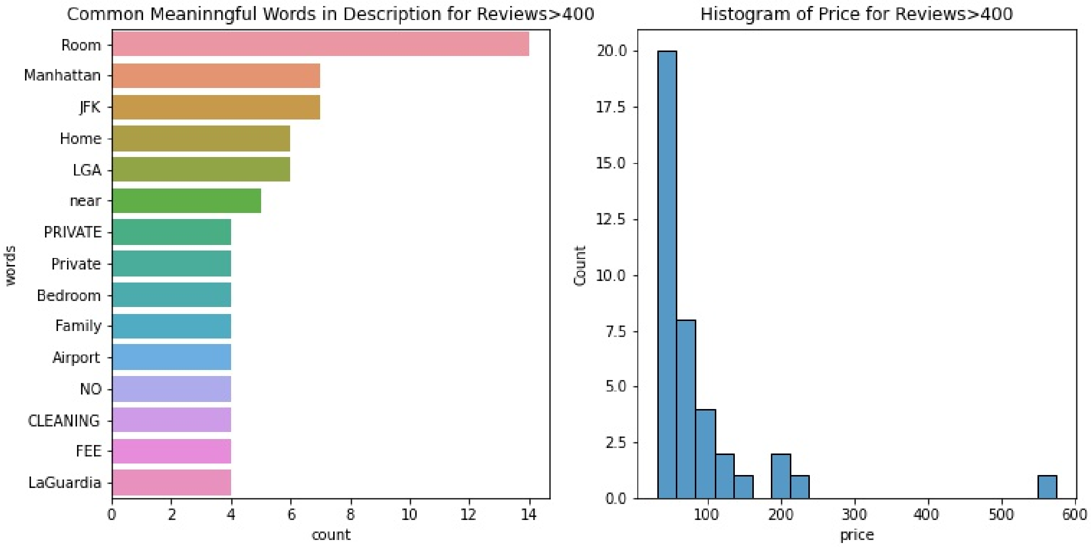


FIG. 17. Common Words of High Review Listings with Histogram of Prices

Common words from listings with high reviews shows that proximity to the airport, with mentions of JFK and LGA, seems to be highly prioritized by Airbnb users. We also observe that privacy and no clearning fees are also popular with Airbnb users. A histogram of the prices shows that although there are some expensive listings with high reviews, the majority are low priced.

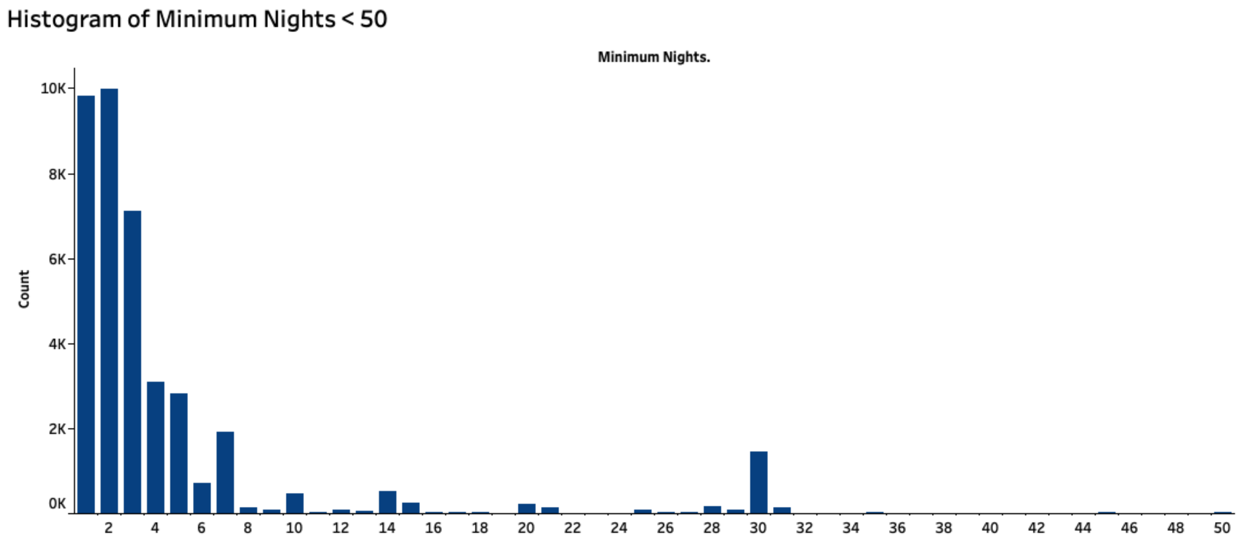


FIG. 18. Histogram of Minimum Number of Nights less than 50

The vast majority of listing have a low minimum number of nights requirement, as akin to Airbnb being a vacation rental service. However, there is a noticeable spike at 30 days here, which might indicate listings targeted towards people staying on longer business trips.

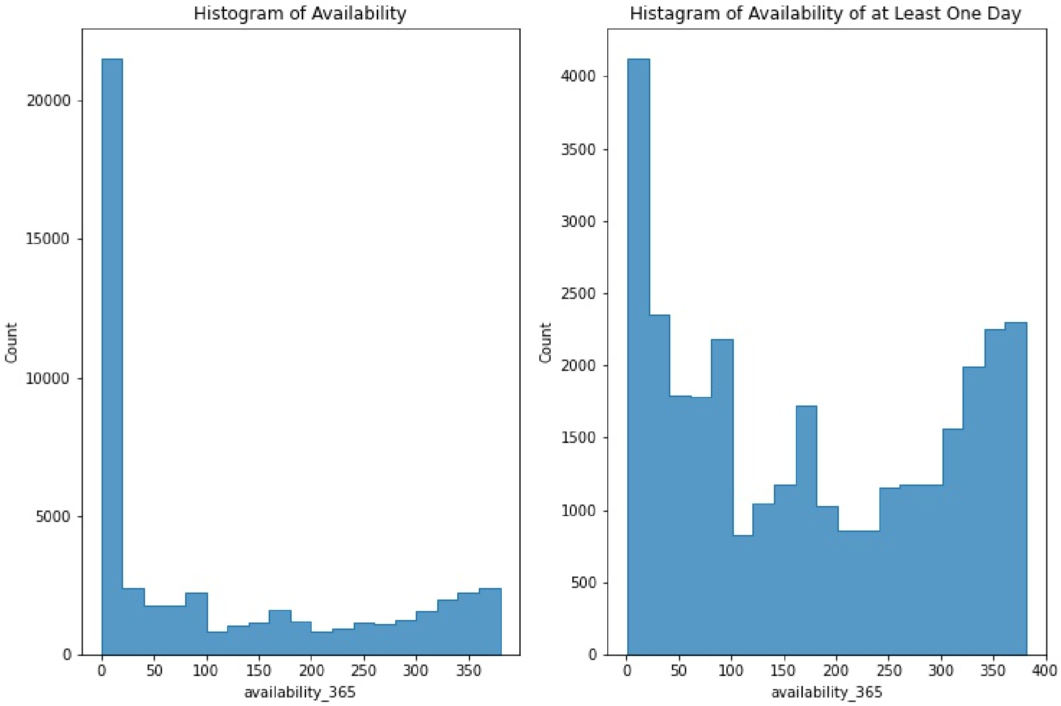


FIG. 19. Histograms of Availability

At a first glance it might seem like low year-round availability makes up most of the listings but most of the listings are not available in 2019. The right plot shows perhaps that high year-round availability might be more common than anticipated.

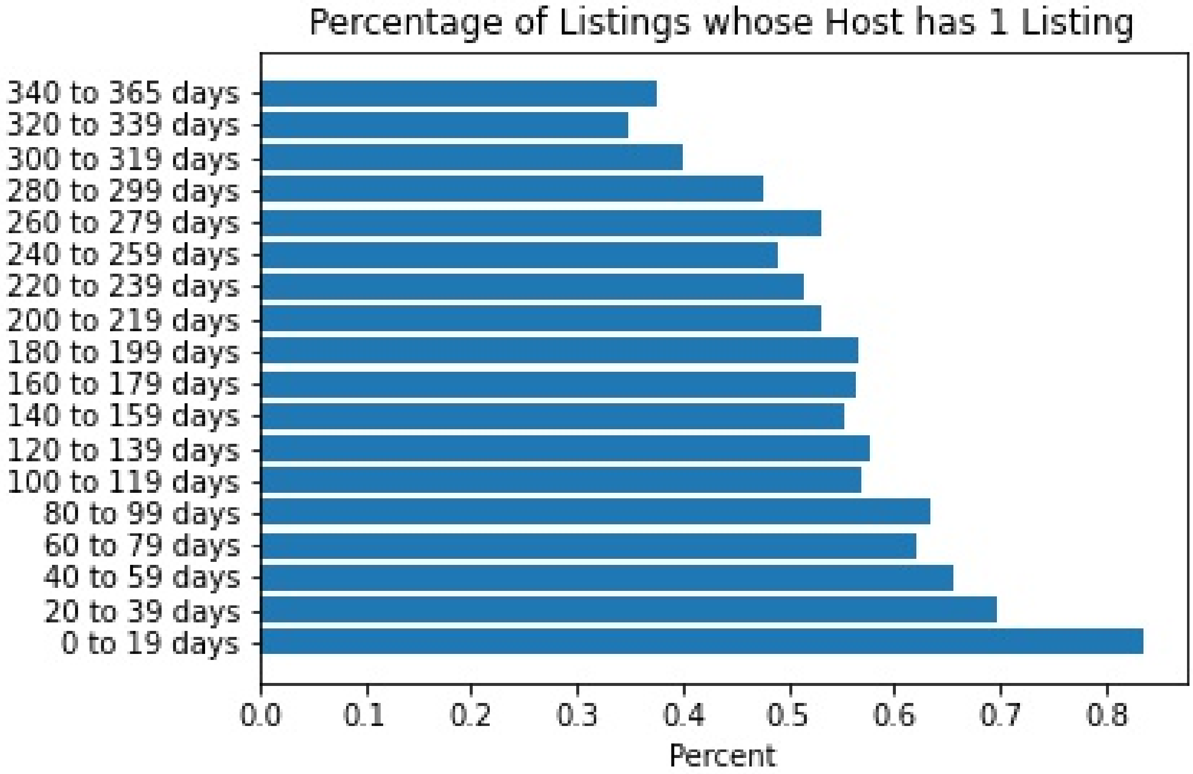


FIG. 20. Bar plot percentage of Hosts who has 1 Listing

Generally, the lower the availability per year, the more likely the host only has one listing. It would seem that super hosts are more likely to have their listings available for longer times.

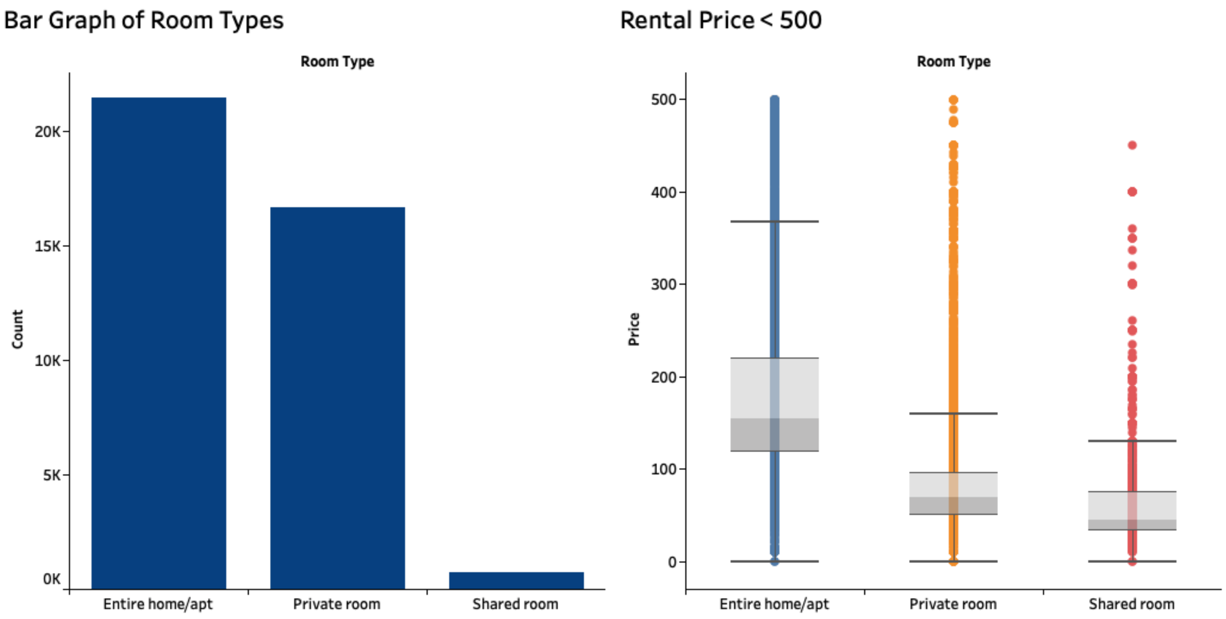


FIG. 21. Bar Graph of Room Type with Box Plot of Room Type Price

The low amount of shared room shows that consumers are much more likely to to want private rooms. The prices of each room type reflects the obvious: entire homes are worth more than a single room, and privacy is worth more.

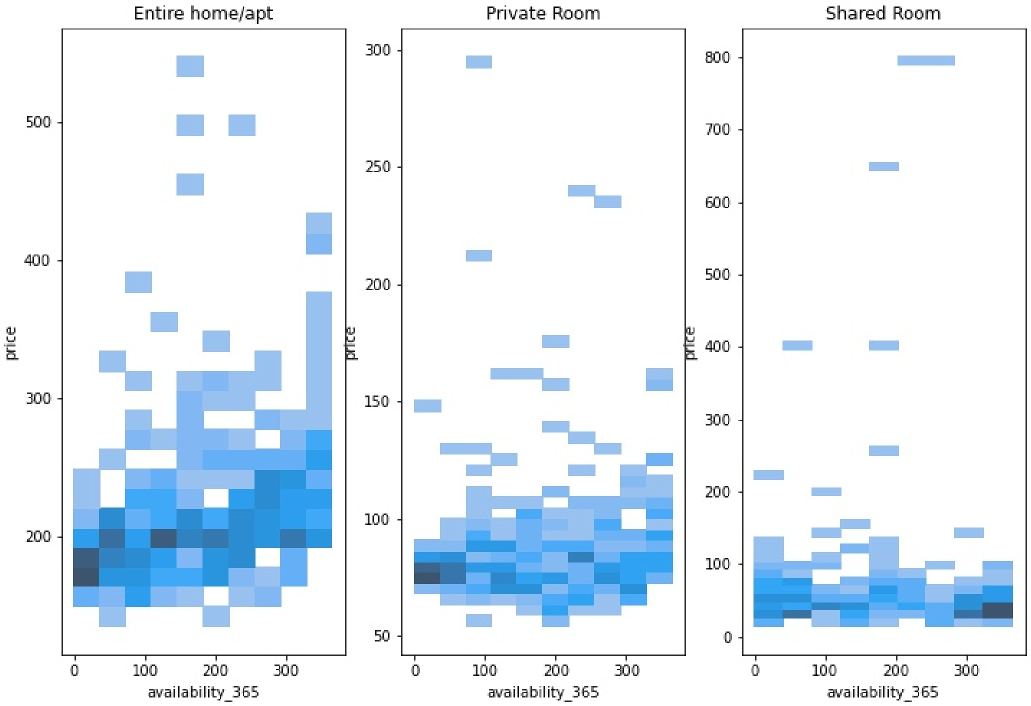


FIG. 22. 2D Histogram of Average Price versus Availability for Different Room Types

Entire homes and private rooms are more likely available for short-term while shared rooms are more likely to be available year round. For specifically entire houses, it seems like a low availability implies a lower price.

V. APPENDIX B: TABLES

TABLE I. A table of problematic variables

Variable Name	Variable Type	Python dtype	NULL Value Count
name	categorical	object	16
host_name	categorical	object	21
last_review	discrete numeric	object	10052
reviews_per_month	continuous numeric	float64	10052

TABLE II. Listing in Rossville-Woodrow

ID	Name	Room Type	Price
27086249	Anna's place bed and breakfast	Private room	118
34835762	Central Hall Colonial with Free Parking Bus EX...	Entire home/apt	1250
1798271	Spacious center hall colonial	Entire home/apt	700
26258351	Escape NYC in the Borough of Parks!	Entire home/apt	75

There are only four listings in this neighborhood, with two abnormally high priced colonial homes. One of them has a description that describes bus access, which implies its not a typical single family stayover.

TABLE III. Summary of Price Statistic

mean	std	min	25%	50%	75%	max
152.720687	240.154170	0	69	106	175	10000

A quantitative measure of price statistics shows that price is heavily skewed. The standard deviation of the prices is larger than the 75% quantile. Oddly enough, there were a small number of listings that were priced at zero, but this is due to bad data, and the these listings today do not have a price of zero.

TABLE IV. Summary of Price Statistic for Reviews Greater than 400

mean	std	min	25%	50%	75%	max
88.871795	92.041232	32	44	55	92	575

The price statistics for high reviews shows that the most popular listings, the vast majority of them have relatively low prices. In fact, the 75% quantile shows a lower price than the median of prices.

VI. SUPPLEMENTARY MATERIAL

The Python code, as well as an in-depth explanation of the methods and packages used in the analysis, is provided in 'AirbnbAnalysis.ipynb'.

Interactive versions the visualizations were created with the help of Tableau Public. Link to Tableau work: https://public.tableau.com/profile/sergio.jimenez#!/vizhome/STA160Project1/Dashboard9?publish=yes

The dataset used in the analysis is provided in 'Airbnb_NYC_2019.csv'.

The shapemap of New York City boroughs is found in 'nybb_21a'.

The shapemap of New York City neighborhoods is found in 'nynta_19d'.

These two maps are courtesy of https://www1.nyc.gov/.

The shapemap of New York City subway lines is found in 'nyu_2451_34758'

The shapemap of New York City subway stops is found in 'nyu_2451_34760'

These two maps are courtesy of https://geo.nyu.edu/.

VII. REFERENCES

Anselin, L. (**1995**). “Local indicators of spatial association—lisa,” Geographical Analysis **27**(2), 93–115.

Getis, A., and Ord, J. K. (**1992**). “The analysis of spatial association by use of distance statistics,” Geographical Analysis **24**(3), 936–941.

TripAdvisor (**2021**). “The 15 best things to do in nyc - 2021,” https://www.tripadvisor.com/Attractions-g60763-Activities-New_York_City_New_York.html (Last viewed April 18, 2021).