Advanced Data Analytics Spring 2021



**New York City: Analysis of AirBnB Rentals Amidst a Pandemic**

Team 13: Ashton Wagner, David Hebert, Ethan Dark, Megan Fitzgerald, Renee Gagne

# **Table of Contents**

[Business Understanding](#_oqs1sijsmrcc) 2

[Data Understanding](#_mstoyrn5mg0z) 3

[Data Preparation](#_wedmeittw4jd) 3

[Modeling](#_ftwst7iecgkv) 6

[Evaluation](#_ha4cftlnlpwh) 9

[Deployment](#_h2docvi6u7r5) 10

[Appendix](#_gb5ygfq90skc) 12

# 

# 

# **Business Understanding**

# New York City is a top destination for people both in the United States and from around the world; the city welcomed a record-breaking 65.1 million visitors in 2018 alone[[1]](#footnote-0). Amidst all of this travel, one of the greatest costs of visiting ‘The Big Apple’ is accommodation. Since Airbnb’s inception in 2007, local listings have become an increasingly popular alternative to hotels; there are approximately 50,000 Airbnb listings in New York City today[[2]](#footnote-1). Despite this popularity, however, the emergence of a global pandemic has led to city-wide shutdowns and a stark decline in tourism---an essential part to the city’s economy. As a result, AirBnB hosts have experienced an average decline in profits by $4,036 since the beginning of Covid-19[[3]](#footnote-2). An important aspect of our project is to not only analyze the data, but to gain modern-day insights based on news articles that pertain to AirBnB rentals in NYC during the pandemic to determine methods to improve host profitability and the impact of the pandemic.

**Project Objective:** The objective of this project is to predict Airbnb rental prices based upon listing attributes using data collected in 2019. In doing so, NYC hosts will be empowered to make data-backed decisions regarding the cost of their rental space. Furthermore, by looking at data from both 2019 and 2020, we hope to gain modern-day insights based on news articles that pertain to AirBnB rentals in NYC during the pandemic.

**Business Question:** How can Airbnb hosts improve profitability of their listings? Additionally, what insights are made from comparing pre-COVID-19 data to data collected during the pandemic?

# **Data Understanding**

Guests and hosts have utilized the Airbnb platform to expand on traveling possibilities and present a more unique, personalized way to experience seeing the world. This dataset describes thousands of different listings based in the metro New York area, including all neighborhood boroughs: Manhattan, Brooklyn, Staten Island, Queens, and the Bronx. This dataset also includes metrics on minimum nights, availability throughout the year, price, frequency of listings, and review data. From here, we plan to implement multiple analyses tools, using R and Python, to find meaning and trends within the data. This dataset included all necessary information to find demographic information about hosts, geographical availability, and pricing metrics to make accurate predictions and infer conclusions about our data.

# **Data Preparation**

**Data Overview:** The 2019 and 2020 datasets began with 48895 and 36923 observations respectively, and both had 16 columns of mixed categorical and numerical values. Each of the datasets contain information pertaining to host identifier values, rental attributes, rental metrics based on guest experiences, and list prices. It is important to note that, conveniently, both the 2019 and 2020 datasets contain the same listing variables which allows for direct comparison between the years of record collection.

**Exploratory Analysis:** When beginning our analysis, we wanted to get a comprehensive understanding of how the data was distributed. We focused on creating different visualizations and running various descriptive statistics to get a better look at the data (*See Appendix Figures 1-11*). We looked specifically at the frequency and distribution of some of our columns with categorical data, like room type, neighborhood, and neighborhood group. While looking at these more in depth, we realized that the price was heavily influenced by room type and neighborhood. Listings that offer entire homes to their guests, as opposed to a private or shared home, typically reflect higher price points per night. Prices also were heavily influenced by where they were located, which references the neighborhood and neighborhood group. In *Appendix Figure 3,* we visualized neighborhood and price based on the Airbnbs’ location within the city, using the longitude and latitude columns from the dataset. Here we can see that Manhattan has the highest frequency of listings as well as the highest average prices throughout the neighborhood.

After further exploring the 2019 dataset, we found holes in the data that would affect our models significantly if we had not come to this realization. We noticed that there was an uneven distribution of listings across the five boroughs. Manhattan and Brooklyn had approximately 20,000 observations each accounting for 85% of total listings (*See Appendix Figure 1)*. In addition, when comparing the average prices of the listings separated by borough, we saw major price discrepancies; Manhattan & Brooklyn had a mean price range of close to $180, whereas other boroughs fell closer around the $100 range (*See Appendix Figure 8)*. As a result, we chose to only inspect the top two boroughs in depth and our statistical models on Manhattan and Brooklyn.

Along with building multiple visualizations and descriptive analytics to explore the data, we also used correlation to see how the different predictor variables are related to each other. We wanted to see how our most significant predictor variables, minimum nights, number of reviews, calculated host listings count, longitude, latitude, and availability, and how they are correlated with each other as well as price.

**Data Cleaning:** Our data exploration stage was critical in guiding the overall process of data cleansing. First, we removed columns determined as being unimportant to the predictive models overall, including: ‘id’, ‘name’, ‘host\_id’, ‘host\_name’, and ‘last\_review’. Although these do columns hold value in distinguishing between AirBnB listings, they do not specifically contribute to the models’ predictions. Additionally, the variables ‘latitude’ and ‘longitude’ were removed based upon the output of a correlation plot and the visualizations created in the exploratory stage.

After reducing the number of variables in the dataset, we further cleaned the data through specific alterations. Exploratory analysis revealed that approximately 20.5% of the data contained in the column ‘reviews\_per\_month’ were null; therefore, the median value of ‘reviews\_per\_month’ was imputed to replace the null values. The column ‘availiblity\_365’ underwent a log transformation due to its strong right skew. Outliers were found in the column ‘minimum\_nights’ with values of 999, 1000, and 1250 minimum nights. Therefore, rows corresponding to these values were removed from the dataset.

Finally, in order to take account for both the size and popularity of the 221 neighborhoods listed in the dataset, the ‘neighborhood’ column was replaced by a column that included all neighborhoods by name if they had 800 or more occurrences in the dataset. Thus, the smaller neighborhoods were classified as being ‘other’. This allowed for a reduction in the number of unique neighborhood values and an emphasis on rentals in the most popular neighborhoods of New York City. We also realized that there were columns with prices of zero. We kept them for the exploratory stage to notice the frequency on the map, but we removed them for our predictions, since it implied that the listing was unavailable.

**Data Organization:** The 2019 dataset was used for model building, as both the training and test sets. Since the 2020 year was one for the books, the data wasn’t as consistent with information collected prior to the pandemic; COVID-19 completely changed the vacation rental industry. For this reason, we chose to take it into consideration, but to not depend on, the 2020 data.

**Modeling**

**Regression Analysis**: After exploring the data and running descriptive analysis on the dataset, we utilized regression for further exploratory purposes. This analysis was useful for a general overview of the significance of predictors. To begin, we initially ran a multiple linear regression with all of the variables included in the dataset. The output from this regression was extremely difficult to interpret and minimal insight could be obtained. However, this insight aligned with our intuition about subsetting and modeling by borough in order to account for the differences within each borough. The team also wanted to see if the log transformed variables had any significant effect. Two additional models were run utilizing the non-correlated variables. The first model used the log transformed variables, ‘number\_of\_reviews’ and ‘availability\_365’, and the second utilized the non-transformed variables. The output and results were consistent.

**Decision Trees:** To explore the Manhattan and Brooklyn boroughs through decision trees, the data was partitioned by borough and maintained all variables remaining from the data cleansing process. The trees were built using the *rpart* library, with price as the predictor variable and the method set to *anova*. The decision trees crafted for both the Manhattan and Brooklyn subsets of the testing data consisted of three breaks: ‘room\_type’, ‘availiblity\_365’, and ‘calculated\_host\_listings\_count’. Regarding the Manhattan model, four rental price values were determined as $115, $240, $221, and $573, (*See Appendix Figure 12*). This model was then used to predict the NYC 2020 Manhattan subset of data. A similar process was performed for the Brooklyn subsets of training and test data, which had four prices determined as being $75, $174, $342, and $1572 (*See Appendix Figure 13*). Though this model is efficient and creates a simple visual output, it can be concluded that decision trees are not the most appropriate method of predicting the price of Airbnb rentals. This is because the models only produce four price values, which is not representative of the variety of values featured in the dataset.

**KNN:** To see how accurately KNN model would predict prices, we segmented out the neighborhood groups and evaluated each. We looked at Manhattan and Brooklyn neighborhoods as we did in the other models, but for the KNN model we also added the three other neighborhood groups into one combined subset. To maintain measurement consistency across models and remove outliers, we only evaluated listings with a price under $500. To include categorical variables, we dummy coded each factor level into separate columns. We also manipulated the numeric variables by scaling each. With the model preparation complete, we were able to split and train the model.

For the KNN model we used a split of 80:20 for the train and test sets. In order to optimize the best value for the number of neighbors, we created a function that calculated the mean absolute error and ran a for loop with different k values to return a list of metrics. A graph visualization helped plot the MAE’s for each k value which we used to determine the best k for each neighborhood group (*See Appendix Figures 14-19*). We decided on k values of 20 for Manhattan, 30 for Brooklyn, and 55 for the others. With optimized k values, we were able to run ‘knn.reg’ functions on the data and return evaluation metrics. The MAE’s were 48.45 for Manhattan, 36.46 for Brooklyn, and 31.28 for the others *(See Table 1 below*).

**Random Forest:** We also wanted to explore how well a random forest model would perform on our 2019 NYC dataset. One of the primary reasons why we chose this model was because it can handle both categorical and numerical variables. Random forests can also handle regression or categorical problems so the flexibility of the model allows it to be used in a wide variety of applications and datasets. We first subsetted the data into two dataframes, Manhattan and Brooklyn. We then selected the features to be used that were suggested by our initial regression models and label encoded the ‘room\_type’ and ‘neighbourhood’ variables. Additionally, we filtered on listings that were $500 or less. Applying this threshold allowed us to account for only listings that the ‘average’ renter would be considering. Moreover, this would allow our models to have better predictive accuracy from the reduced influence from outliers.

We began by utilizing the scikit-learn library and built out a random forest for both data frames. Both models were trained and tested using a 85/15 train test split. After running the random forest models on the Manhattan and Brooklyn subsets, a 10-fold cross validation was utilized and returned MAE values 49.2 and 34.8 respectively (*See Table 1*).

|  |  |  |  |
| --- | --- | --- | --- |
| Model | MAE Manhattan | MAE Brooklyn | MAE Other |
| KNN | 48.4 | 36.5 | 31.3 |
| Random Forest (2019) | 49.2 | 34.8 | X |
| Random Forest (2020) | 54.2 | 36.4 | X |

*Table 1: Summary model metrics*

# 

# **Evaluation**

The random forest and KNN models performed better than the linear regression and the decision tree models. Both models’ MAE metrics were close in value, with a difference of approximately one unit. When considering both efficiency and performance, we recommend the deployment of the random forest model. The random forest model, while developed on pre-covid data, also performed well when tested on the 2020 COVID-19 data, though not as well when run with 2019 data. This implies that the random forest model performs best under normal vacation rental market conditions, and thus, will likely perform better when the market returns to normal conditions, or alternatively, sees a ‘new normal’. While we chose the random forest as the best predicting model, copious understanding of the data was established while making the other models. For example, the linear regression model informed further model building through revealing that borough, neighborhood, and room type were significant variables, and number of reviews was not. Furthermore, the decision tree model informed us that the majority of rental prices fall within a range of approximately $75 to $250 dollars. Additionally, the KNN model provided an in-depth understanding of both the most popular neighborhood subsets as well as smaller neighborhoods.

# 

# **Deployment**

Data analysis of the thousands of listings in this Airbnb dataset is a crucial factor for maximizing revenue for the company and the hosts alike. These thousands of listings contain copious amounts of data - data that can be used and analyzed for business decisions, security, and understanding the behaviors of Airbnb’s guests as well as it’s hosts. This dataset shows the performance of the platform, which can guide marketing initiatives, implementation of innovative services and much more.

Airbnb currently does not offer their hosts many noteworthy resources by way of personalized suggestions for how to price their offerings. They only show seasonal pricing trends for their multi-neighborhood area. By incorporating all of the 200+ neighborhoods in Manhattan into the random forest model, we capture the effect of attractions like Central Park and Times Square on rental prices. The model also captures the effect of the change in borough on price which can be significant especially when comparing Manhattan to the other borough. Other parameters like the type of room are important to consider when pricing an offering. This is because a shared room is a much different offering than an entire apartment, and has a very different perceived value to the consumer.

We recommend Airbnb implement a feature that offers hosts personalized suggestions taking the mentioned parameters into account. By making predictions based on the random forest model, we can help Airbnb suggest the optimal price for each host’s property, which either means more bookings by lowering price or more profit by raising price. Airbnb makes revenue on the 3% service fee they charge every time a property is rented out. This means if we can help Airbnb hosts make more profit, the organization will benefit along with them. On the other hand, if Airbnb suggests a price that is lower than what the host is currently charging, it may lead to a significant increase in bookings. While more guests means more revenue for both Airbnb and their hosts, it also means more work for the hosts. At the end of our presentation, we were asked if this posed an ethical dilemma for our hosts, since this pricing structure does benefit the company of Airbnb more than the host. Airbnb should take this into account, making their pricing suggestions non-obligatory, as it should still be up to the host what they feel comfortable charging for their listings. We were more interested in studying the patterns and trends associated with price in accordance with the data.

# 

# 

# **Appendix**

# **Team Contributions**

|  |  |
| --- | --- |
| Team Member | Contributions |
| Ashton Wagner | Descriptive visualizations, model selection using pycaret, implications & deployment, model evaluation, presentation slides |
| David Hebert | Descriptive analysis in R, Regression modeling in R, Random Forest Modeling in Python, presentation slides |
| Ethan Dark | Data preparation, KNN model in R, regression modeling, report writing, presentation slides |
| Megan Fitzgerald | Exploratory analysis, descriptive statistics, visualizations & regression in Python, final report writing, presentation slides |
| Renee Gagne | Data cleaning, decision tree modelling, presentation slide creation/formatting/preparation, report writing |

# 

# 

# 

# 

# 

# 

# **Appendix: Visualizations**

**Those below have been made in R using packages ggplot2 and ggmaps.**

Figure 1, Number of listings by borough:

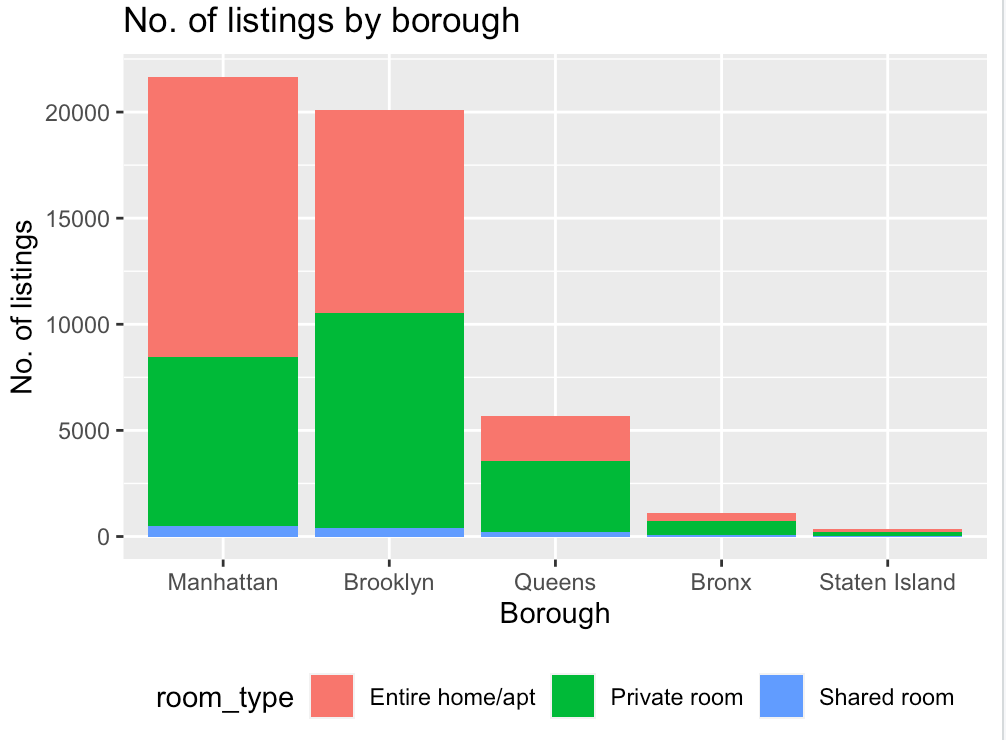


Figure 2, Price by room type:

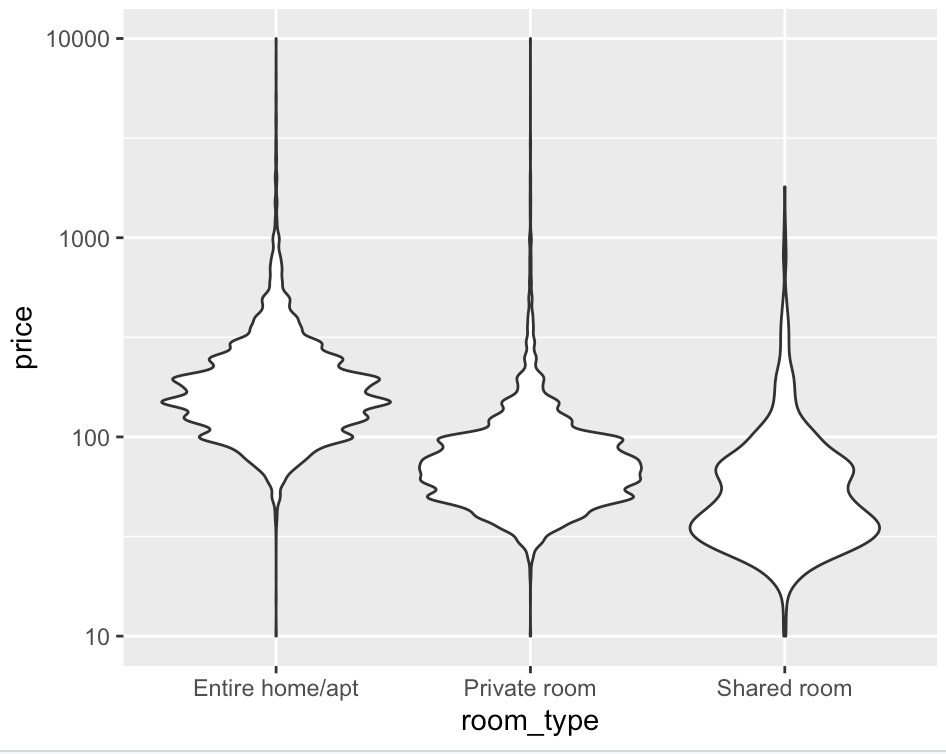
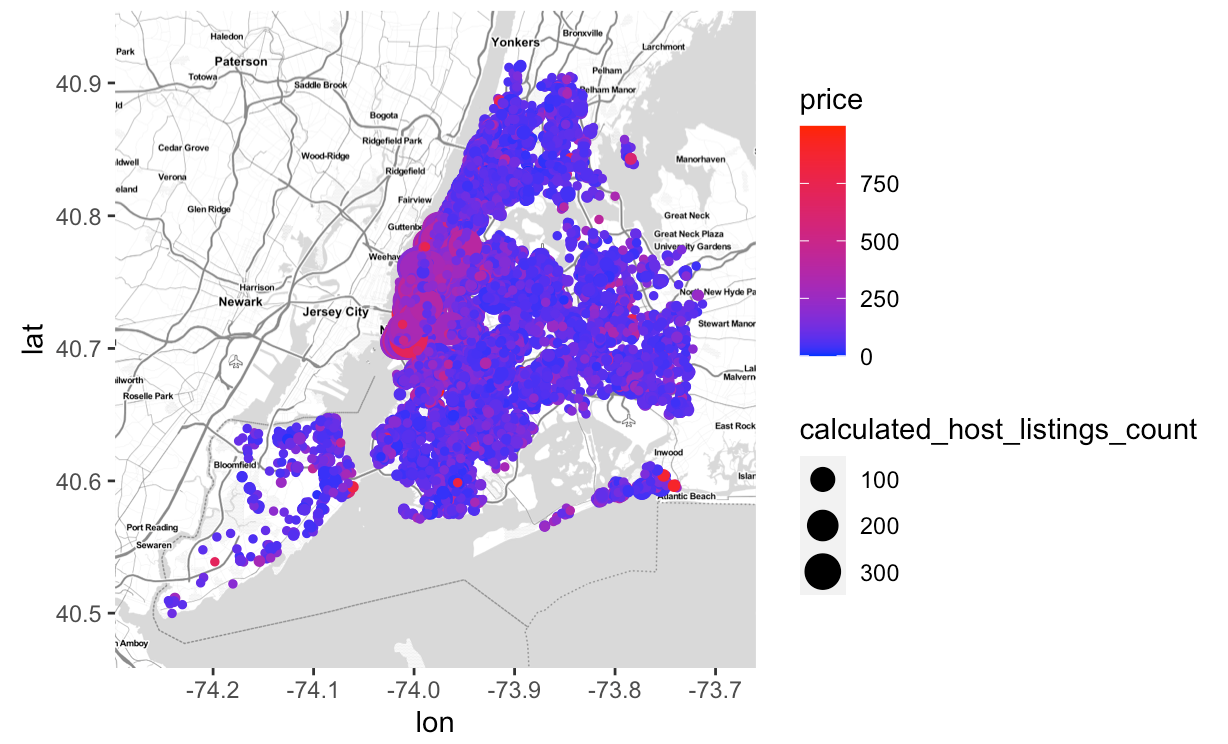


Figure 3, Price by Neighborhood using Longitude and Latitude Columns:



**Visualizations below were made in Python using plotly.express, Matplotlib, and seaborn packages.**

Figure 4, Listings in NYC Borough by Room Type:

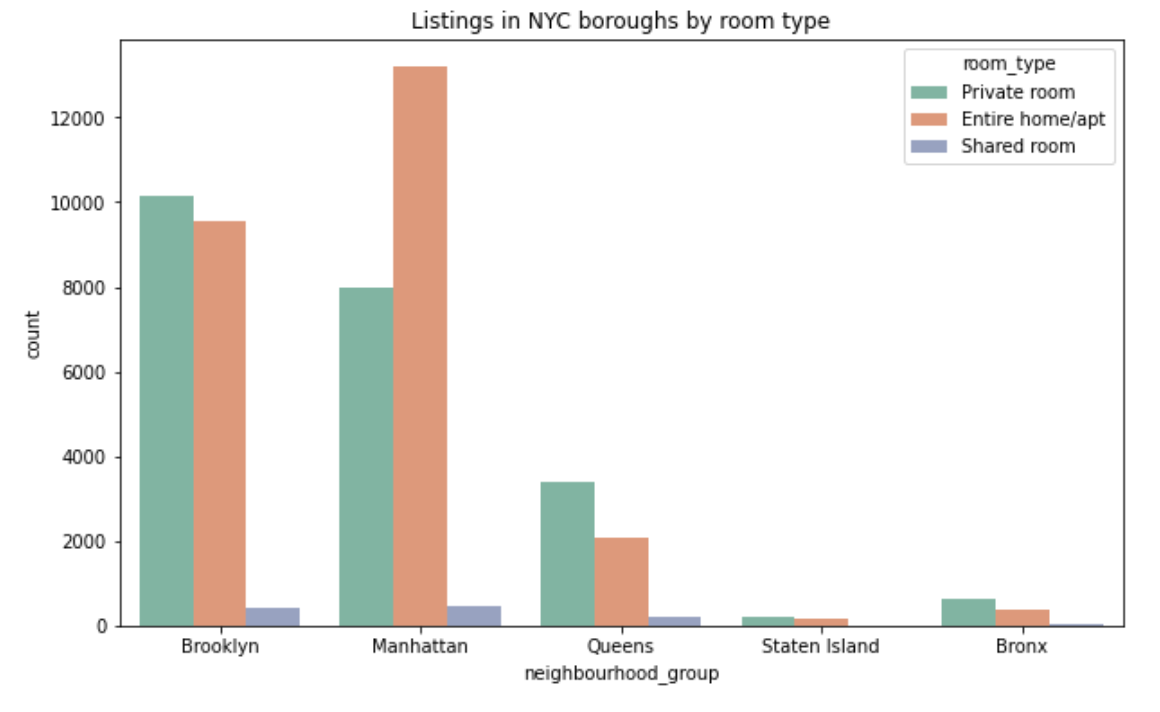


Figure 5, Distribution of Listings by Minimum Nights:

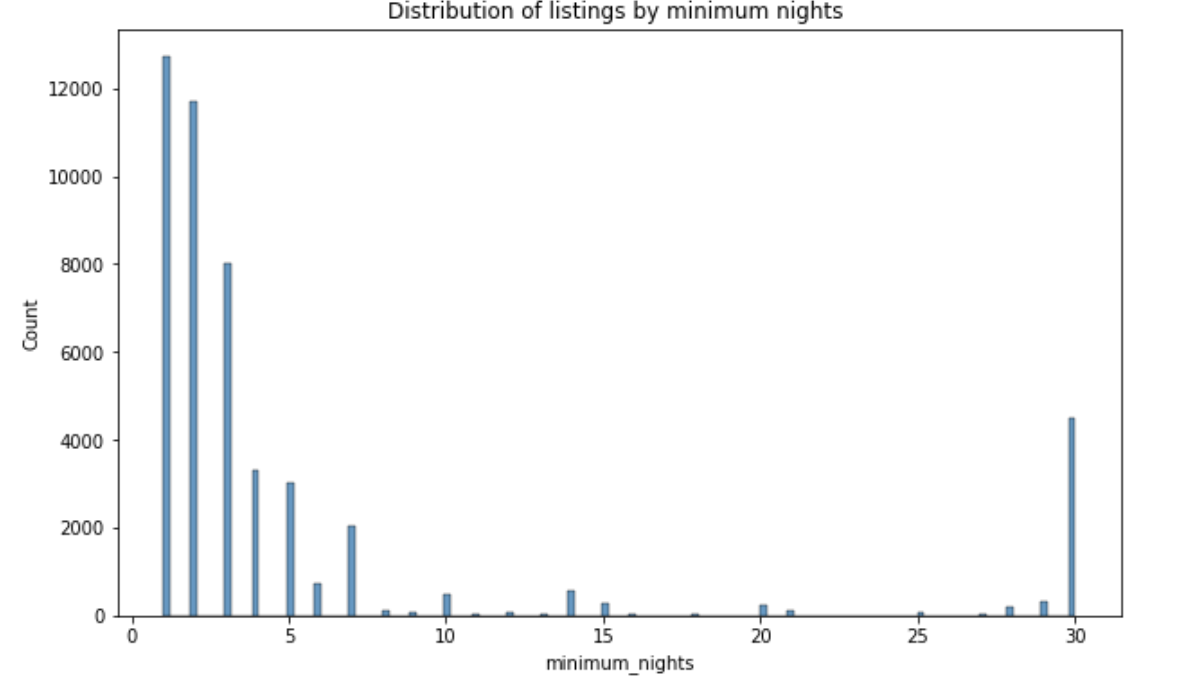


Figure 6, Top 50 neighborhoods where most Airbnb listings are available:

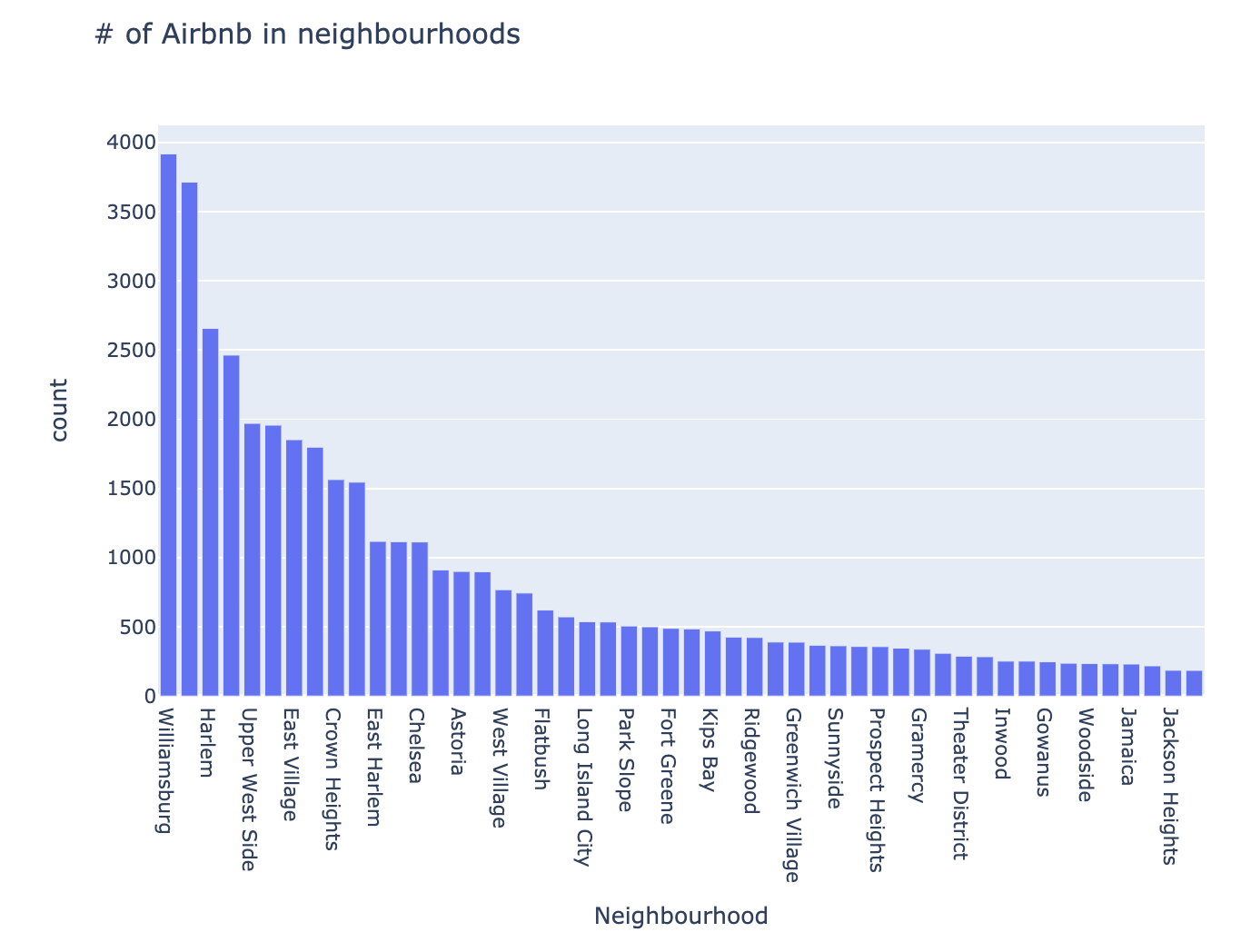


Figure 7, Minimum Number of nights distribution for listings:

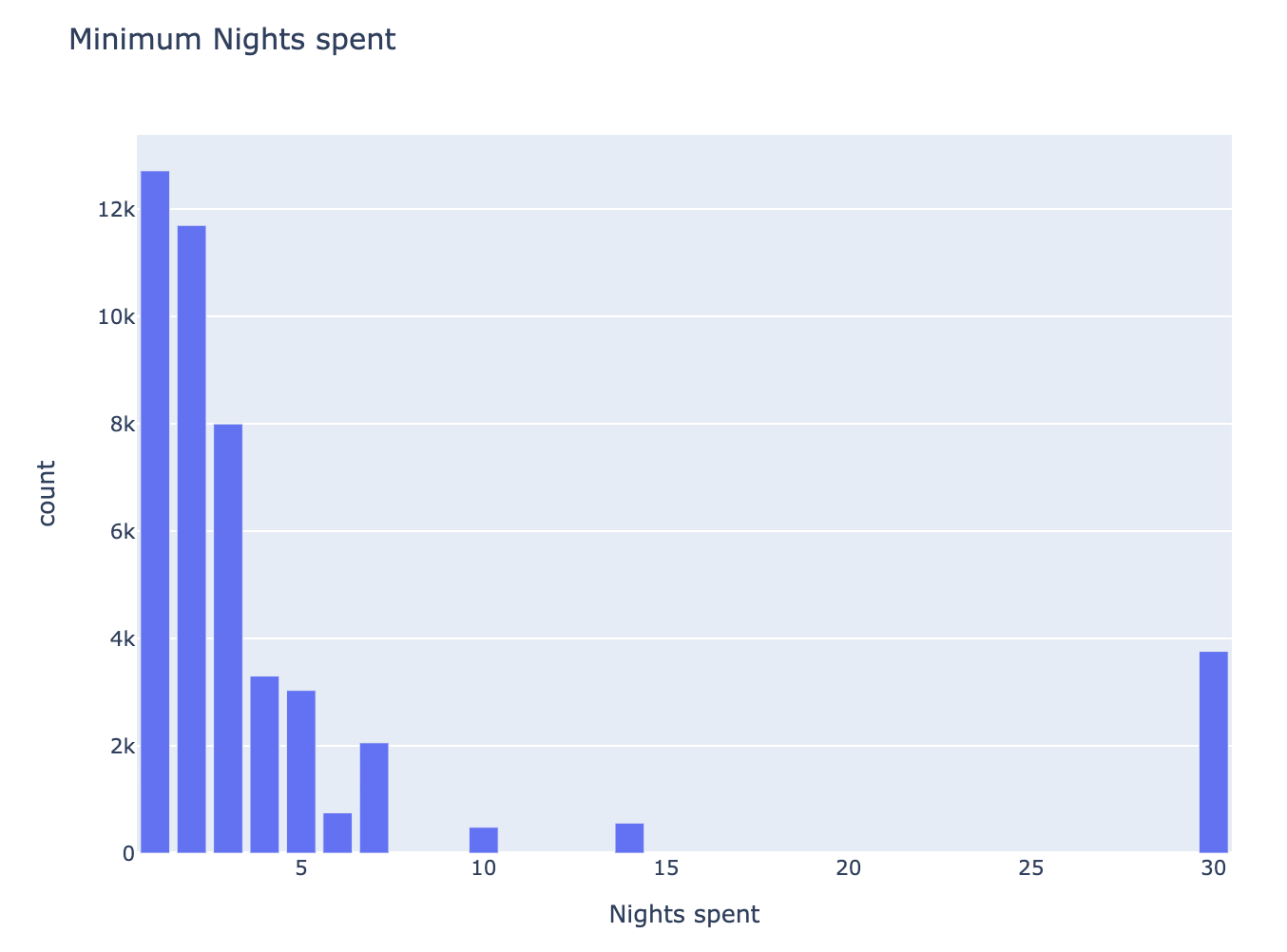


Figure 8, Average price of listings in each Neighborhood:

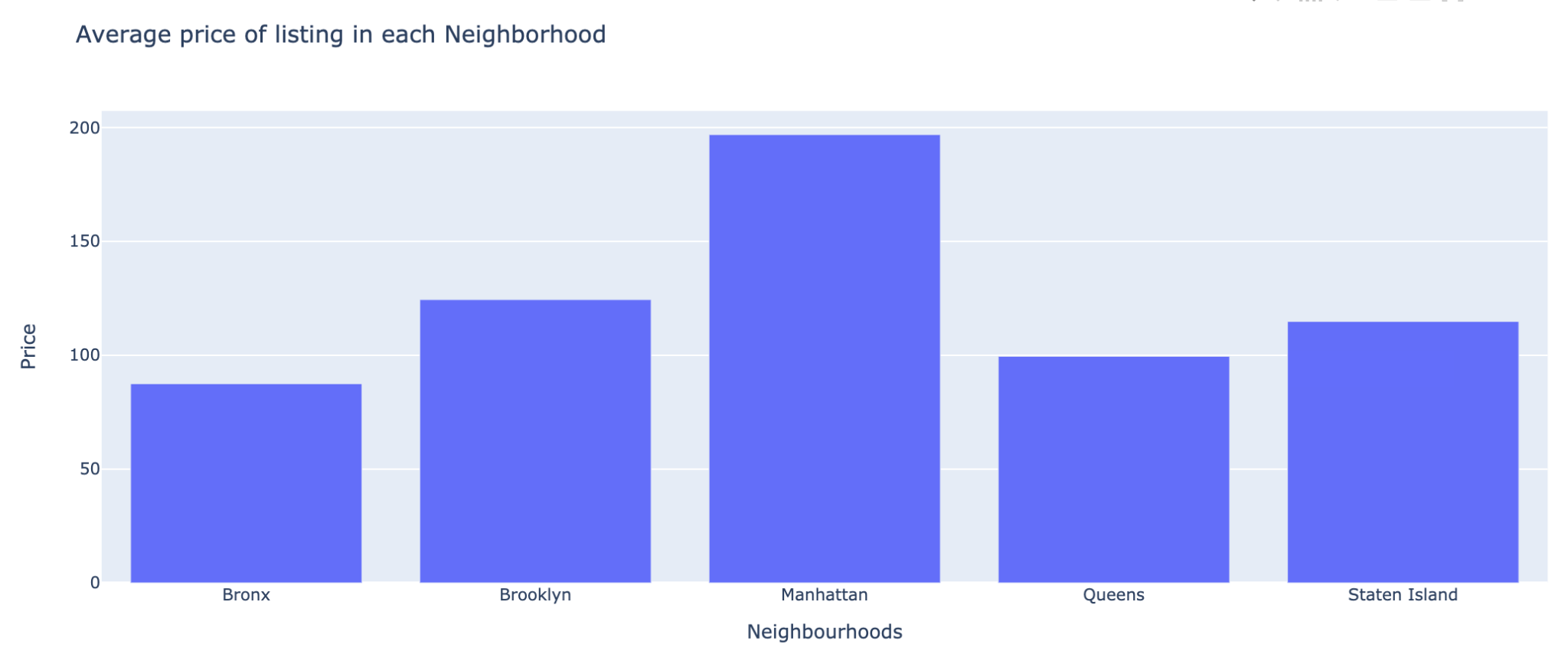


Figure 9, Average price per Room Type:

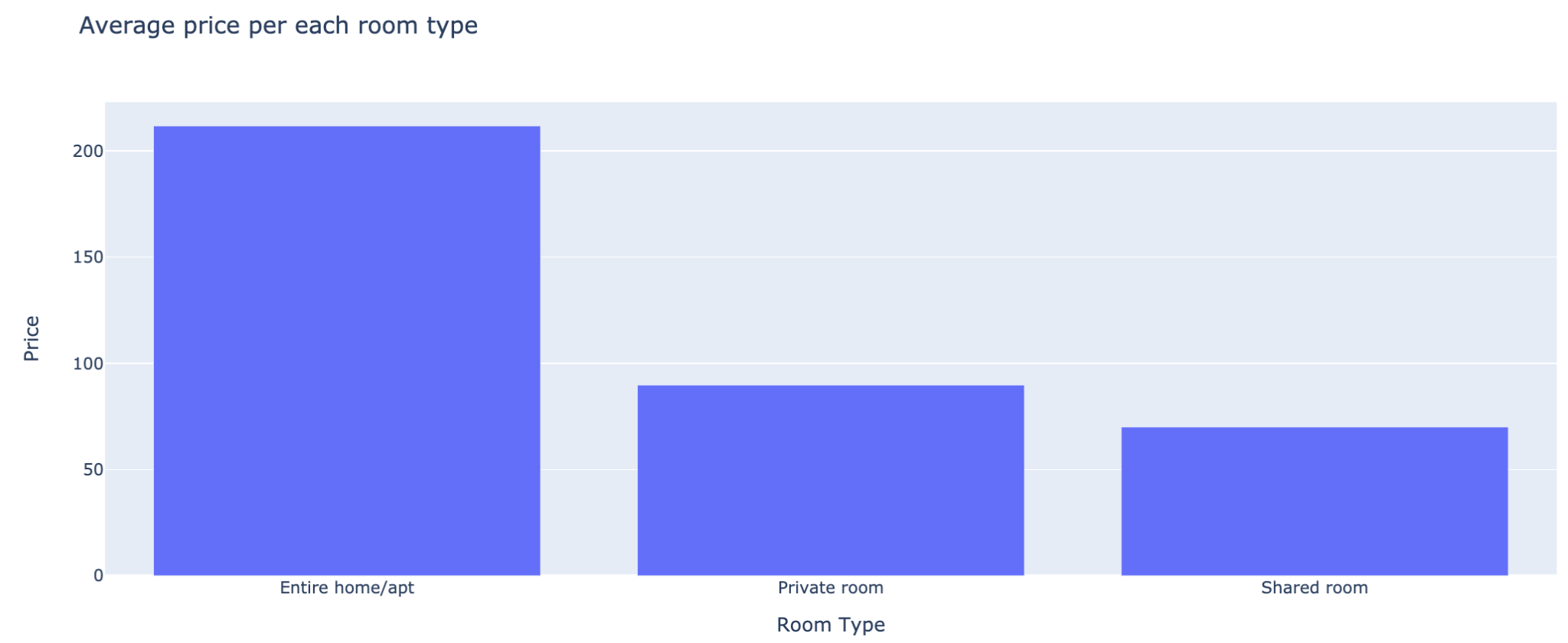


Figure 10, Visualization showing the number of reviews based off of neighborhood-group and price:

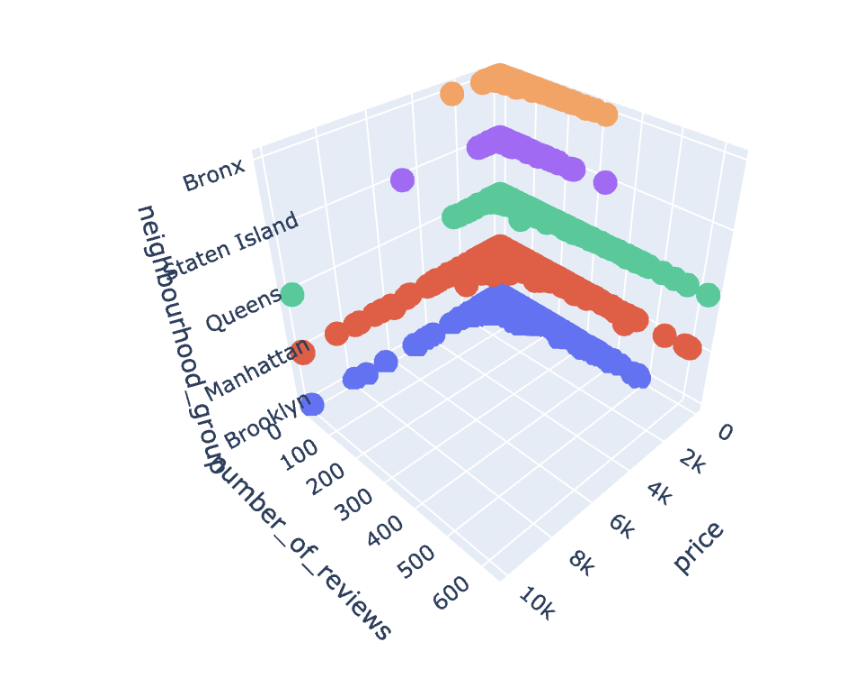
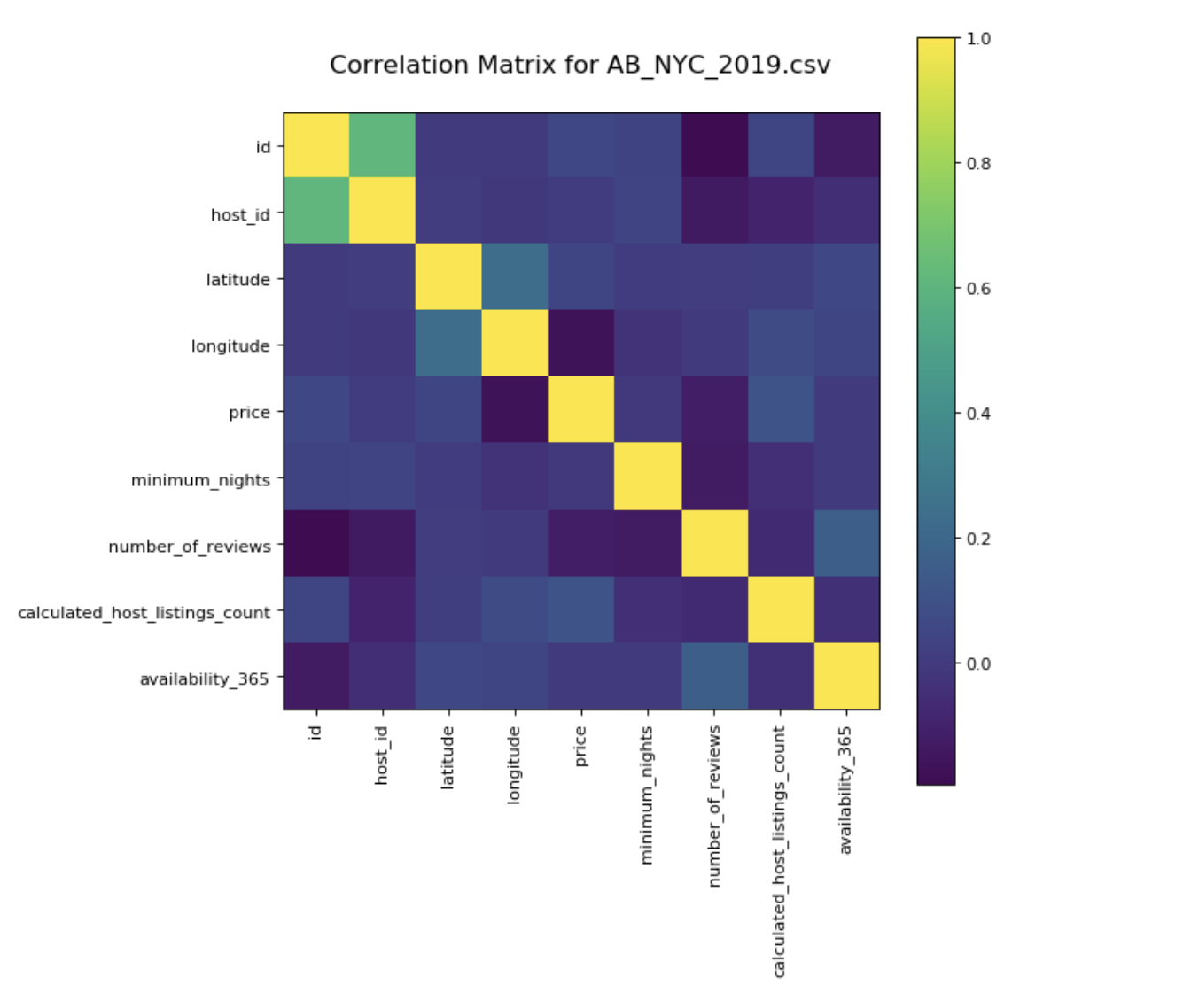


Figure 11, Correlation matrix for all included columns before data cleaning



**Visualizations below were made in R**

Figure 12, Decision tree based off of NYC 2019 Manhattan subset of data

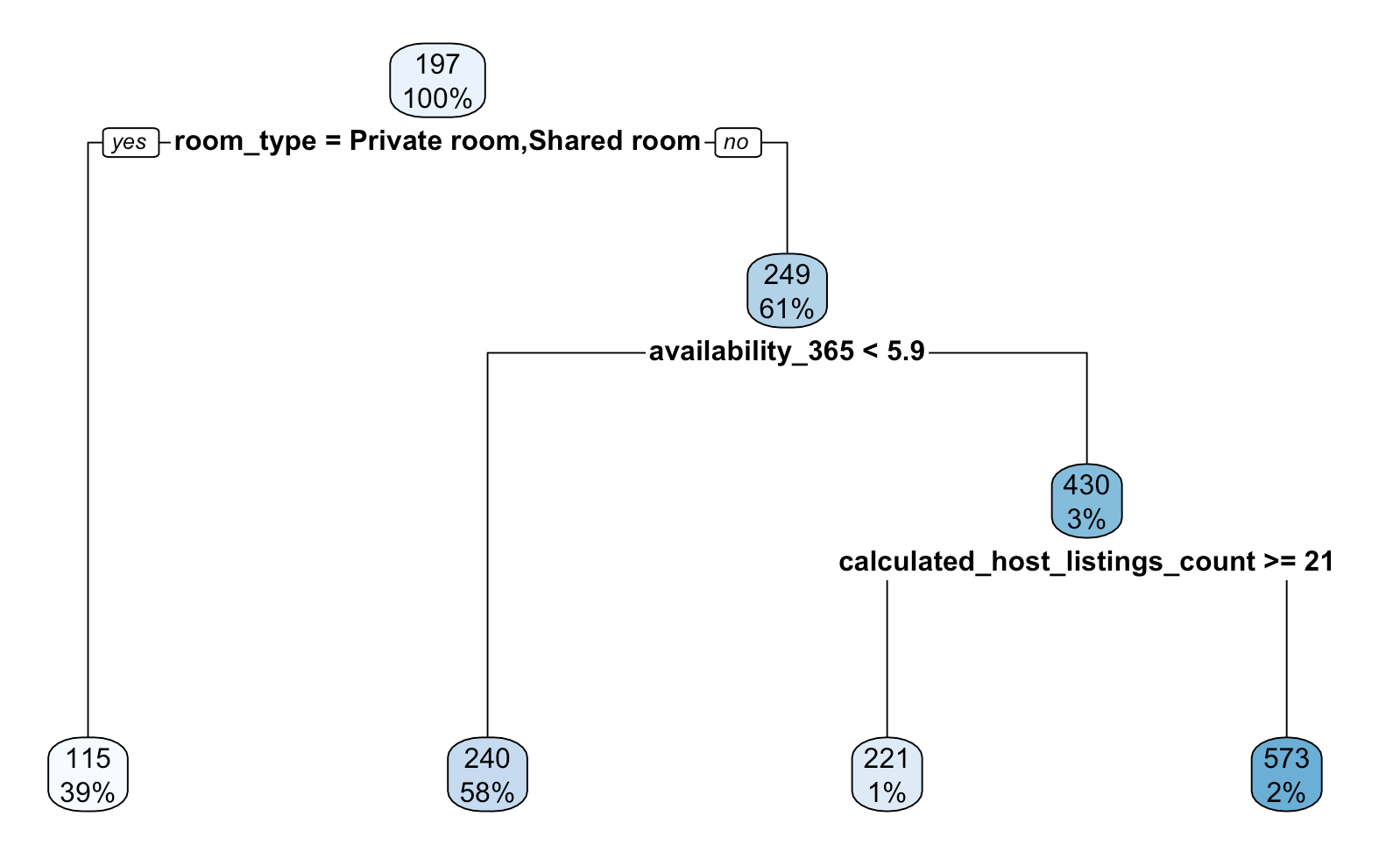


Figure 13, Decision tree based off of NYC 2019 Brooklyn subset of data

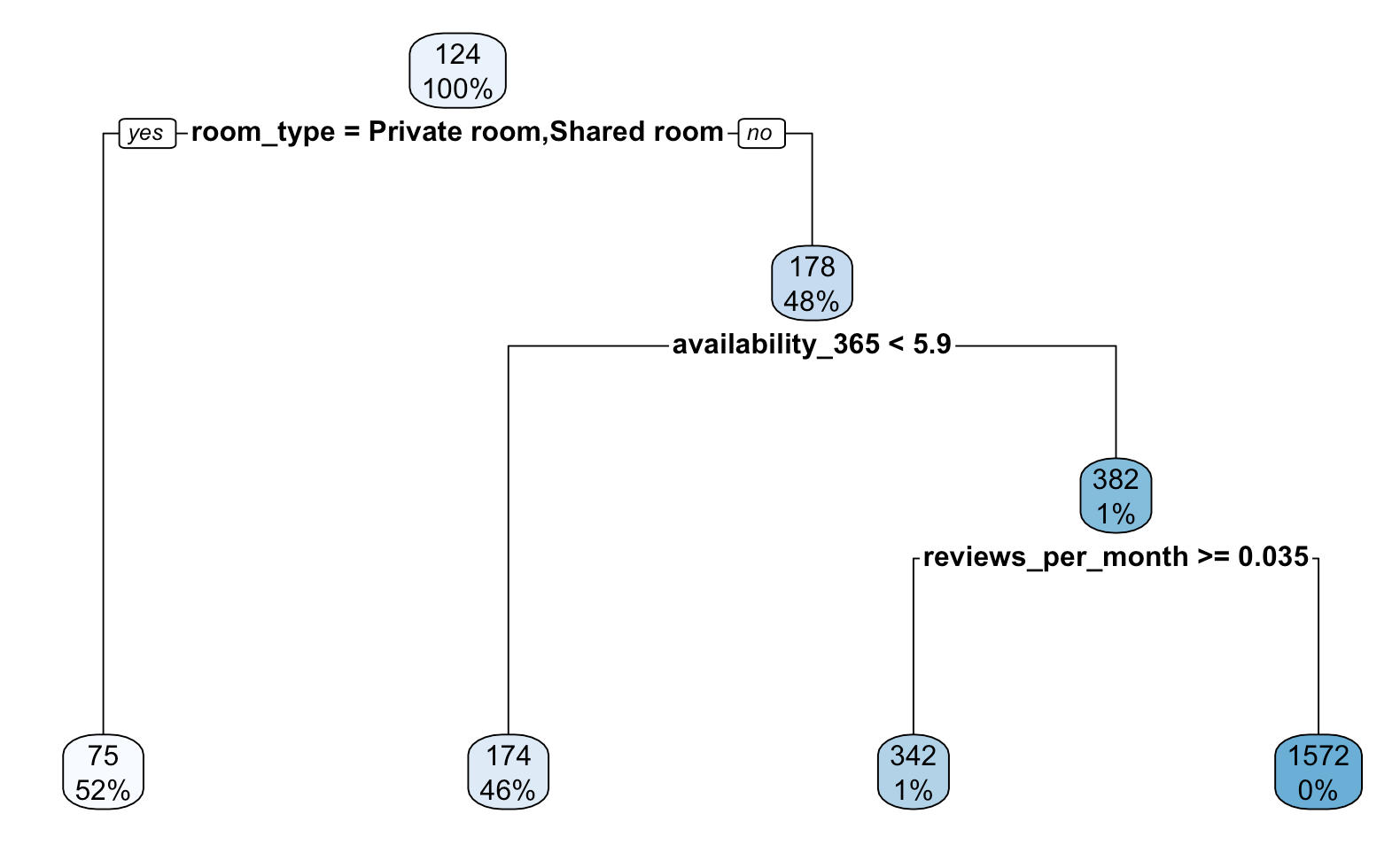


Figure 14-15, KNN model k-optimization and evaluation plot for Manhattan

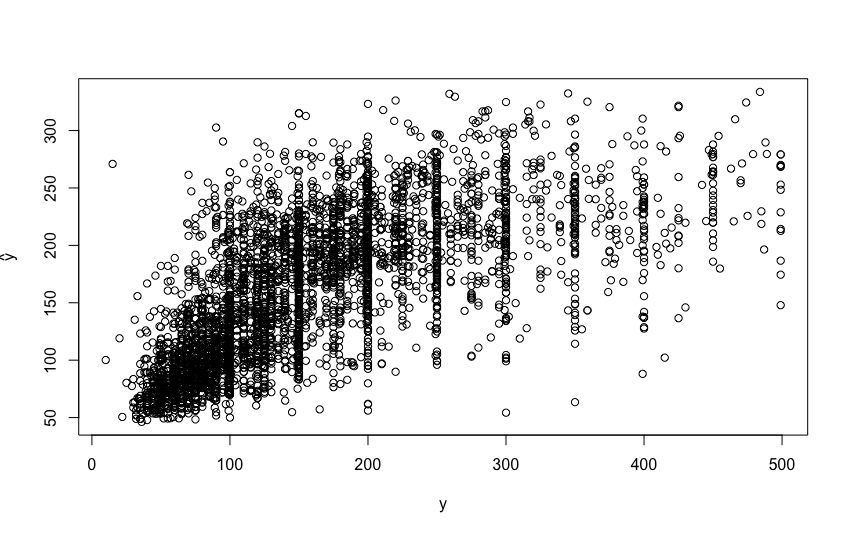
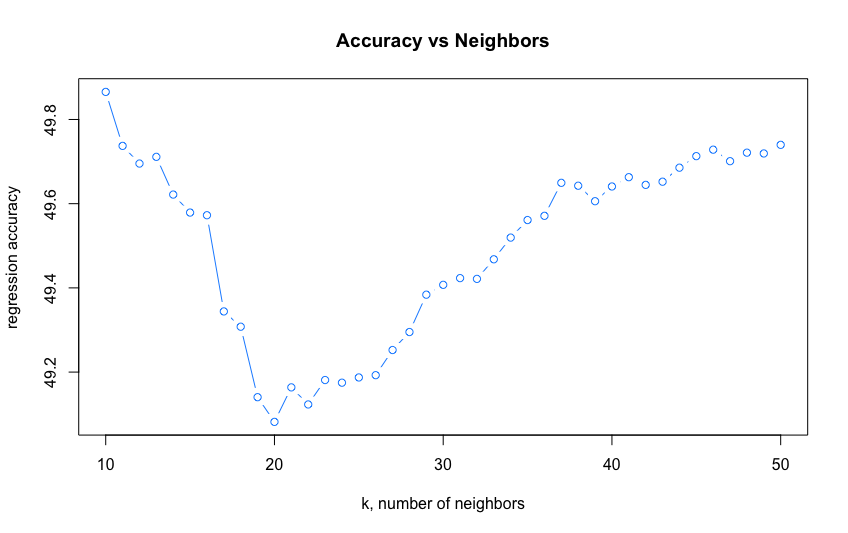


Figure 16-17, KNN model k-optimization and evaluation plot for Brooklyn

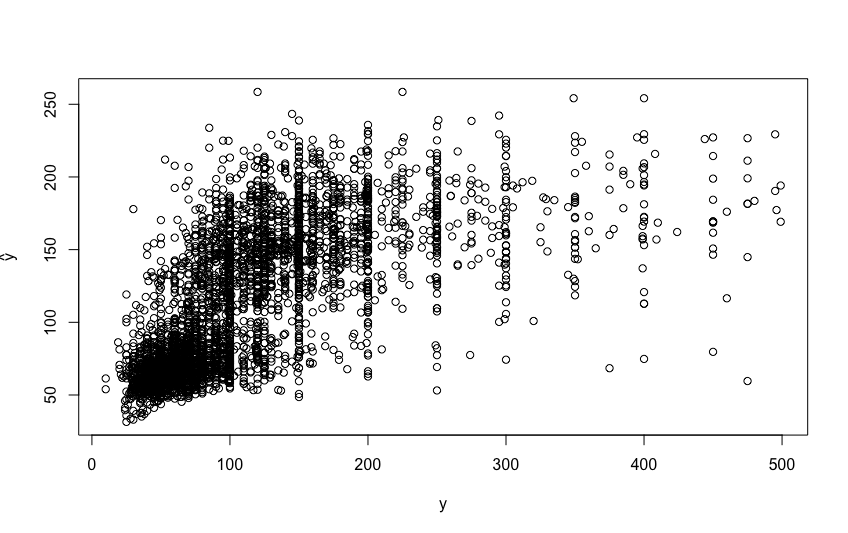
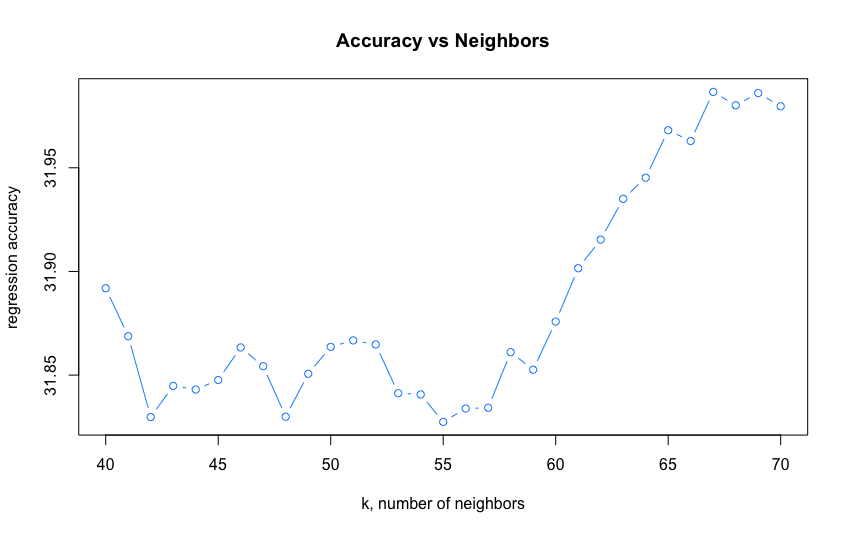
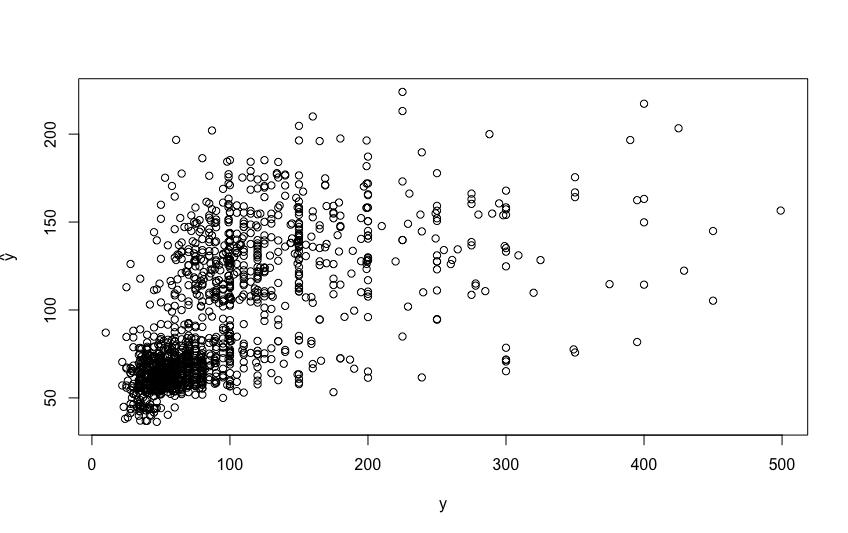
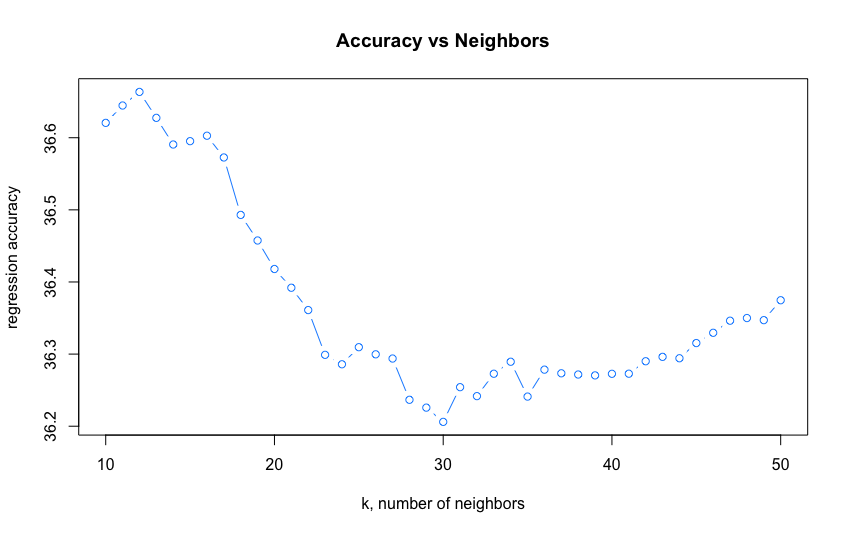


Figure 18-19, KNN model k-optimization and evaluation plot for Queens, Staten Island and Bronx



1. Source: [*https://www.baruch.cuny.edu/nycdata/tourism/index.html*](https://www.baruch.cuny.edu/nycdata/tourism/index.html) [↑](#footnote-ref-0)
2. Source: [*https://www.investopedia.com/articles/personal-finance/010215/hotels-vs-airbnb-new-york-city-visitors.asp*](https://www.investopedia.com/articles/personal-finance/010215/hotels-vs-airbnb-new-york-city-visitors.asp) [↑](#footnote-ref-1)
3. Source: [*https://www.forbes.com/sites/lealane/2020/06/09/how-bad-are-covid-19-pandemic-effects-on-airbnb-guests-hosts/?sh=6bf00b837432*](https://www.forbes.com/sites/lealane/2020/06/09/how-bad-are-covid-19-pandemic-effects-on-airbnb-guests-hosts/?sh=6bf00b837432) [↑](#footnote-ref-2)