NYC Airbnb Listings and Household Economics

By Alex Hingtgen, Allison Weber, Isabel Gomez, Collin Durbin, and Alec Gienger

1. Introduction

The online vacation rental marketplace Airbnb has become a widely popular platform that users rely upon when looking for a place to stay overnight during vacations. Airbnb offers a unique rental experience that enables customers to purchase overnight stays at homes, apartments, or hotel rooms. The price of renting an Airbnb place varies based on the location and time of when users book a reservation. Household income and the number of houses in an area can also be a key determinant in the availability and cost of a rental unit. The number of households with higher income can impact various aspects of an Airbnb listing and alter consumer pleasure.

In this project, we plan to use data on Airbnb rental locations throughout New York City in 2020 from Kaggle and analyze the impact that housing and economic measures in certain neighborhoods have on the cost and overall customer experience of overnight stays.

2. Data

For this project, we utilized two primary sources of data: Airbnb dataset containing a subset of information on New York City from Kaggle, and U.S. Census American Community Survey (ACS) data of 2019 New York City Profiles.

2.1 New York City Airbnb Listing Dataset

The first set of data we used for this project is an Airbnb dataset from Kaggle providing information on rental locations throughout the United States. The Kaggle dataset[1] that we retrieved initially contains 226,030 rows each describing a distinct Airbnb rental unit. However, we reduced this collection of data to only observations within New York City. We did this by filtering in Excel for observations located in New York City and only keeping listings from this metropolitan area. The dataset was largely composed of listings in New York City, so this data reduction still provided a substantial amount of observations to evaluate. This subset of the data regarding New York City consists of 45,756 rows about rental units within 5 unique neighborhood groups.

The dataset contains information from 2020 and has columns describing features such as host id, host name, listing id, listing name, latitude and longitude of listing, the neighbourhood and neighbourhood group (five boroughs), price, room type, minimum number of nights, number of reviews, last review date, reviews per month, availability, calculated host listings count, and city. We kept all of the fields in the original dataset as we believed they could provide value in all the identified observations. There were certain columns that we did not use, but they were considered throughout the course of our study. This data set was presented in a csv file format,

[1] https://www.kaggle.com/kritikseth/us-airbnb-open-data

allowing us to easily read it into a data frame. Table 1 describes the variables that we utilized from the Kaggle dataset in our project.

2.2 NYC ACS Housing Economics

To gain additional information on New York City and factors that impact Airbnb listings in the area, we seeked information from The William and Anita Newman Library website[2]. This page contains a component called the "NYC 2019 ACS Profiles" that lists links to data on social, economic, housing, and demographic data in New York City neighborhoods. The data in each of these hyperlinks comes from the U.S. Census American Community Survey (ACS) data collected in 2019. We focused on the economic hyperlinks[3] for each of the NYC boroughs as we found that page contained the most relevant information for our analysis.

Based on the hyperlinks describing economic information of the NYC neighborhoods, we wrote a crawling script that explored all five of the places of interest. Our crawling script contained data on the total households, households with income of $200,000 or more, and median and mean household income. The for loop that we utilized made it simple to gather these four data points on all of the boroughs we were evaluating. Being that households with income of $200,000 or more represented the highest income range for NYC neighborhoods, we reclassified this variable as "High_Income_Households" for the remainder of our study. Ultimately, this process produced a data frame composed of the 5 New York City boroughs as well as an additional comprehensive observation describing the total values for all the NYC neighborhoods.

2.3 Integrate New York City Airbnb Dataset and New York ACS

Being that our two datasets had a common column in the New York City borough, also described as the neighborhood in this project, we were able to horizontally merge our Kaggle dataset with the scraped ACS housing data. We did not have to use any unique packages in our horizontal integration process, as built-in functions merged the datasets. Thus, we merged the datasets by the neighborhood values that could be found in each data frame. To ensure that no observations were lost in our Kaggle Airbnb dataset, we coded to keep all of these observations in the merge function. The horizontal integration did not add any rows to the data, but did append four new columns to make our dataset. This resulted in a final data frame with 45,756 rows and 21 columns. A description of all variables can be found in Table 1. The reading in of our Kaggle dataset, crawling process, and data integration is included in the R script "FinalProject_Code.R."

Table 1 Data Dictionary

Column	Type	Description
id	Numeric	Unique ID identifying each

[2] https://guides.newman.baruch.cuny.edu/nyc_data/nbhoods
[3] https://mcdc.missouri.edu/applications/acs/profiles/report.php?p=38&g=16000US3651000&s=Economic

		Airbnb unit
name	Text	Name of the Airbnb listing on the online platform
host_id	Numeric	Unique ID identifying the host of the Airbnb
host_name	Text	Name of the host for the rental unit
neighbourhood	Numeric	The neighbourhood that the Airbnb is located in
neighbourhood_group	Text	Borough that the Airbnb is located in
latitude	Numeric	Latitudinal coordinate of the Airbnb's location
longitude	Numeric	Longitudinal coordinate of the Airbnb's location
room_type	Text	Style of rental unit listing (e.g., Entire home/apt, Private room)
price	Numeric	Price of the listing per night
minimum_nights	Numeric	Minimum number of nights required to book
number_of_reviews	Numeric	Total number of reviews for the listing
last_review	Date	The date that the listing was last reviewed
reviews_per_month	Numeric	Average amount of reviews the listing received per month
calulated_host_listing_count	Numeric	Number of listings by the host
availability_365	Numeric	Number of days in year the listing is available for rent
city	Text	Name of the city the listing is located in

Total_Households	Numeric	Number of total households in each of the 5 neighborhood groups
High_Income_Households	Numeric	Number of total households in each of the 5 neighborhood groups with annual income of $200,000 or more
Median_Income	Numeric	Median annual income for the 5 neighborhood groups
Mean_Income	Numeric	Mean annual income for the 5 neighborhood groups

3. Analysis

The goal of this project is to analyze the role of total households and household income in Airbnb listings located in different New York City boroughs. A central focus was also put on understanding the different aspects of an Airbnb listing and the differences in price, availability, and customer experience in the five NYC areas.

3.1 Listing Customer Experience Reviews

How greatly does the price impact the number of reviews a listing received? We assumed that Airbnb listings with a very high price would likely receive more reviews as customers would like to rave about their positive experience. In contrast, a room that is a very low price would also receive a high number of reviews since customers would be displeased with the quality provided. By running summary statistics, we found the average price of an Airbnb listing was $149.60. The mean number of reviews received was 22.56, which seemed slightly skewed by certain listings receiving high amounts of reviews yet we thought that was important to consider. Furthermore, the correlation test that we ran to analyze the linear correlation between price and number of reviews also showed there was no correlation. The p-value below .05 helped us to reject the null hypothesis for this test.

Is there a substantial difference in the number of reviews a rental unit receives across the different New York City neighborhoods? Some neighborhoods may receive more reviews for their Airbnb listings because the quality of stay may be especially poor or high in rental units. By using the barplot function with tapply we were able to efficiently find the average number of reviews for each neighborhood group. The graph below indicates that Staten Island received the most average listing reviews by a wide margin. On the other hand, Manhattan received a considerable amount less in average reviews. Both of these insights can be used to consider the consumer experience at the listing.

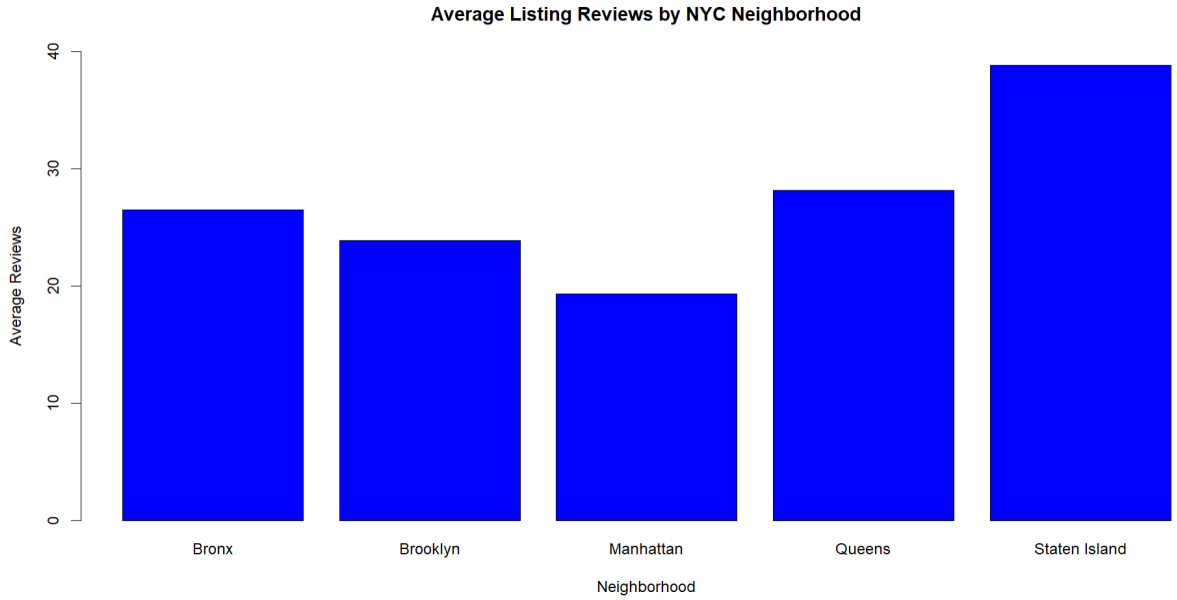

Figure 1 Histogram of average listing reviews by NYC neighborhood

How does the average reviews vary among the different neighborhoods when considering room type? Being that we already know which neighborhoods receive the most reviews, we looked to dig deeper and see the breakdown on which rooms in these neighborhoods received more reviews.

Table 2 Average Reviews by Neighborhood and Room Type

	neighbourhood_group	Entire home/apt	Hotel room	Private room	Shared room
1	Bronx	32.32374	NA	23.95700	13.29787
2	Brooklyn	28.51354	41.70370	19.61484	13.53211
3	Manhattan	16.43642	22.13429	23.55557	22.11311
4	Queens	32.43559	7.50000	26.34468	15.27684
5	Staten Island	40.82759	NA	37.34459	6.75000

We created a dplyr summary table that shows the average number of reviews in each neighborhood and for each room type. Essentially, this table used two categorical variables to study the numeric variable of average customer reviews. In order to enhance the visual of the summary table, we employed the reshape2 package and dcast function to build a pivot table. Finally, the table showed that entire home and private room were the main reasons for Staten Island having so many customer reviews. However, in Manhattan the entire home or apartment

room type resulted in the lowest number of consumer reviews for a rental unit in that neighborhood.

3.2 Price of Rental by Neighborhood Group

Which neighborhood group has the highest average prices of rental unit listings? We aggregated the data by neighborhood group and was able to find that the maximum average price was $193.0506 and the neighborhood group was Manhattan. We used the built-in max function to find the highest value in the aggregated data.

Is there substantial difference in price among the different NYC neighborhoods? To show the difference between the average price among the different groups we created a qplot.

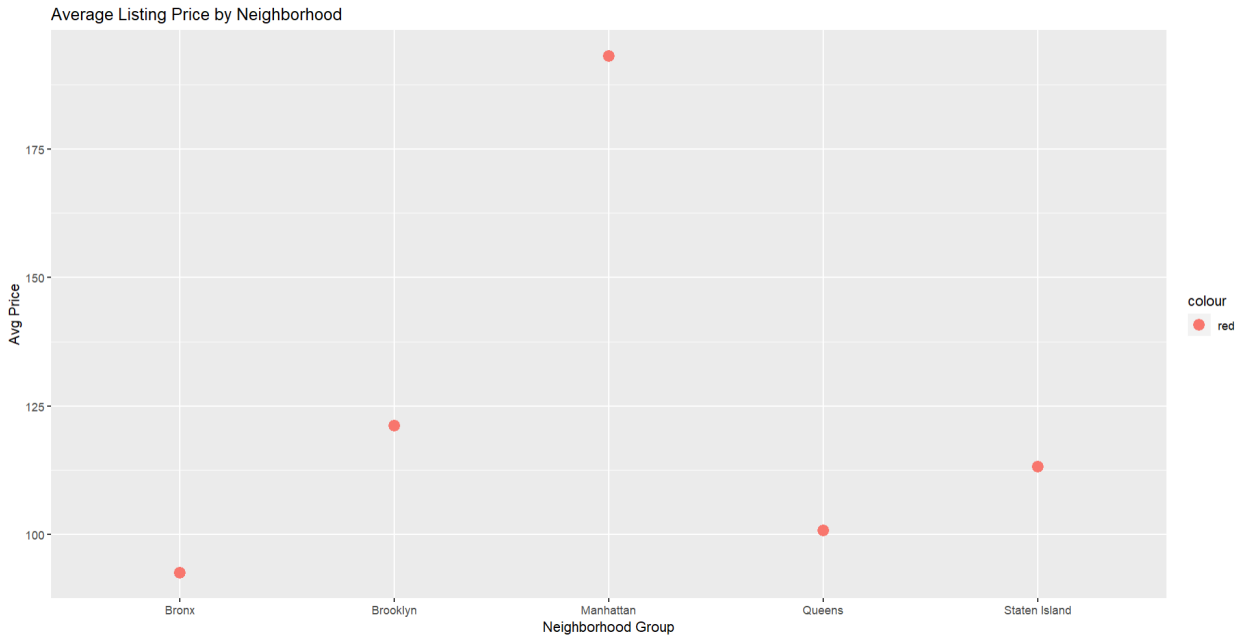

Figure 2 Average Listings Price by Neighborhood

The plot shows that Manhattan has a substantially higher average price than the other four groups, which makes sense since it is where the majority of tourist attractions in New York are located. In addition, the plot also shows that the other remaining neighborhoods are all relatively close in terms of average price.

3.3 Impact from Availability of Listings

How does the room type and availability impact the minimum number of nights that a customer must book for an Airbnb listing? To determine how the minimum nights stayed impacts availability we ran a correlation test to determine if there is any linear correlation. This test outputted a low correlation value of 0.1515449, resulting in none or very minimal correlation between these two observations. In addition, for quality assurance we analyzed the p value of

2.2e-16, which is less than 0.05 therefore we reject our null hypothesis that availability and minimum nights stayed are linearly correlated.

When analyzing the impact and relationship between room type and availability we utilized an anova test to compare numeric and categorical features. After analyzing the anova model summary we reject the null hypothesis that availability and room type. As shown in the boxplot, figure 3, we reject the null hypothesis because the features do not all have the same mean and have no impact on one another.

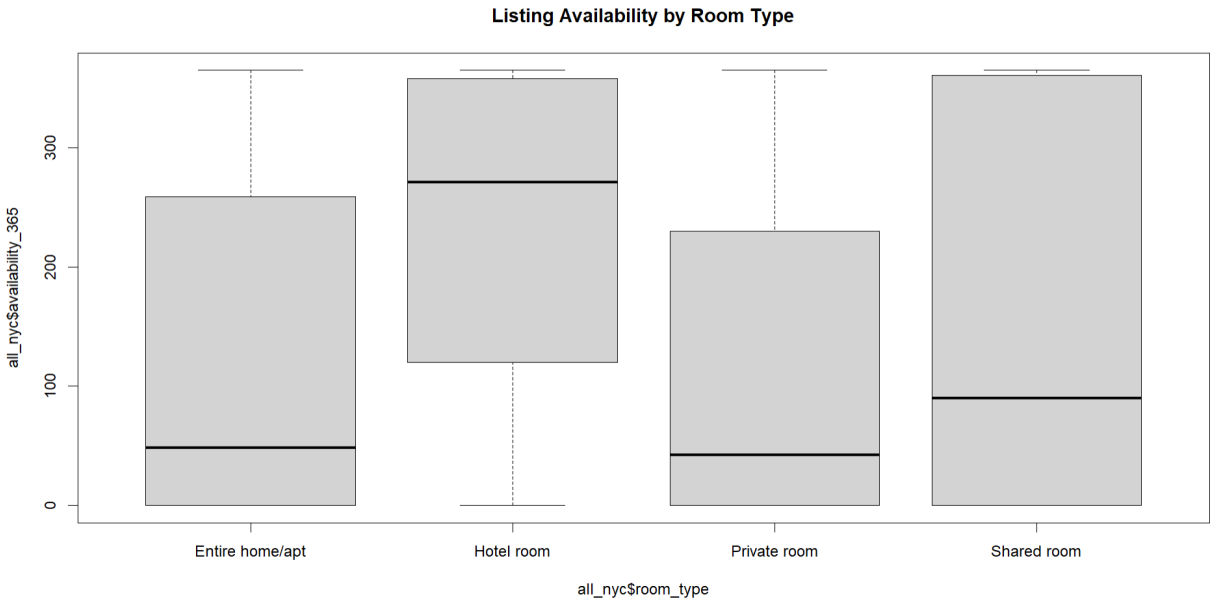

Figure 3 Listing Availability by Room Type

In addition we can analyze the availability of the different listing types in comparison to one another in the pie chart, figure 3.

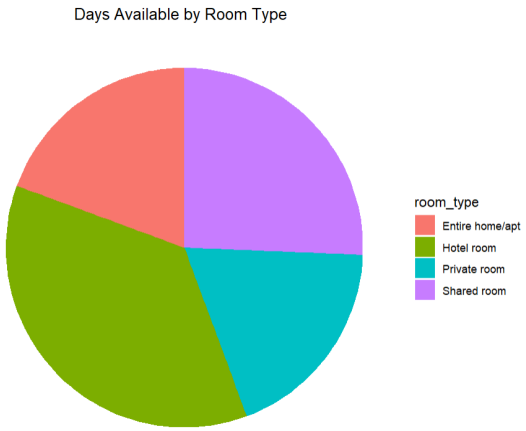

Figure 4 Listing Days Available by Room Type

For example, we see that hotel rooms typically have the highest availability at 36.27%, followed by shared room types with 25.64% availability. Entire home/apt have a listing availability of 19.41% and private rooms have the lowest percent availability at 18.68%. These data figures can also be seen in Table 3 below. These insights are valuable for customers to help plan their Airbnb booking based on their preference of room type and minimum nights stayed.

Table 3 Airbnb Listing Availability by Room Type

	room_type	availability
Entire home/apt	Entire home/apt	121.1068
Hotel room	Hotel room	226.3715
Private room	Private room	116.5552
Shared room	Shared room	160.0318

3.4 Total Households by Neighborhood

How does the total households and number of households with income above $200,000 in parts of New York City affect the availability of Airbnb rental units? We were curious if areas with more households would result in having more availability as there are more places to stay. This analysis also shows if areas with high income households are related to the number of households. In Figure 5 we see the breakdown of average availability of an Airbnb based on the 5 different boroughs.

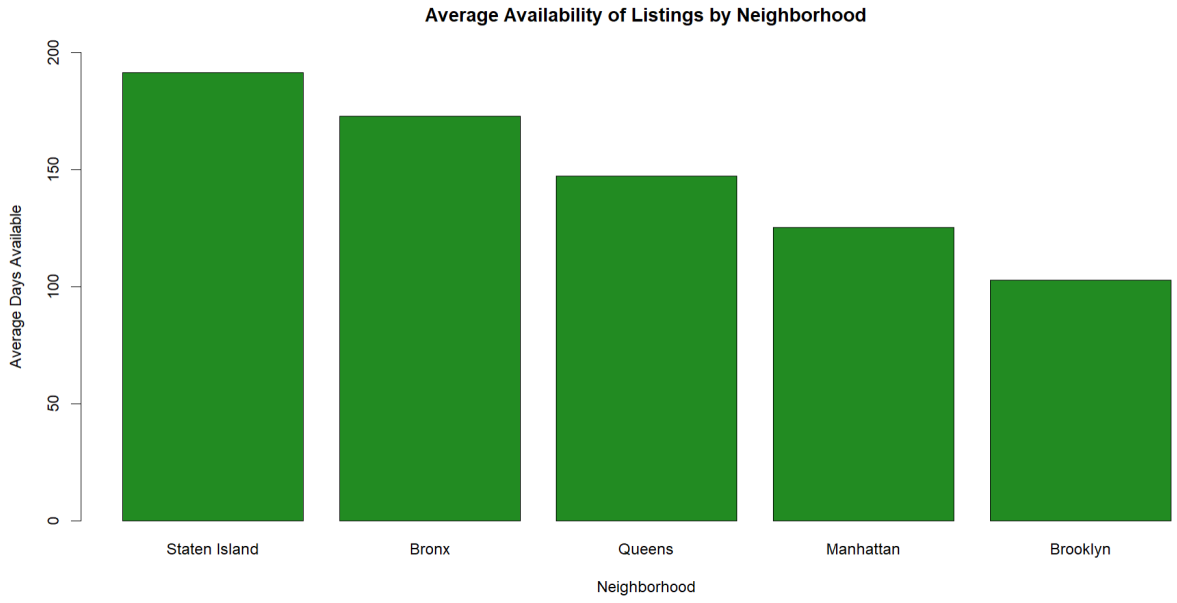

Figure 5 Average Availability of Listings by Neighborhood

For making this bar-plot I used the function “tapply” to get the aggregate of the availability and neighborhood_group. Then I used that variable for the making of the barplot. From this visual we can see that Staten Island has the most availability which is interesting because it is not the richest or most populated borough. Inversely, Manhattan and Brooklyn are low on the average availability.

Next, we created a table utilizing the dplyr package that shows the nieghborhood_group, total households, high income households, and percent of high income households. In Table 4 below we see the percent of high-income houses and Staten Island isn’t the highest at 13.79%. However, Staten Island has the lowest number of total households and second lowest number of high-income households.

Table 4 Percent of High Income Households by Neighborhood

	neighbourhood_group	Total_Households	High_Income_Households	Percent_HighIncome_Houses
1	Bronx	513890	16758	3.261009
2	Brooklyn	978091	111304	11.379718
3	Manhattan	768203	192633	25.075794
4	Queens	784552	72801	9.279308
5	Staten Island	166297	22941	13.795198

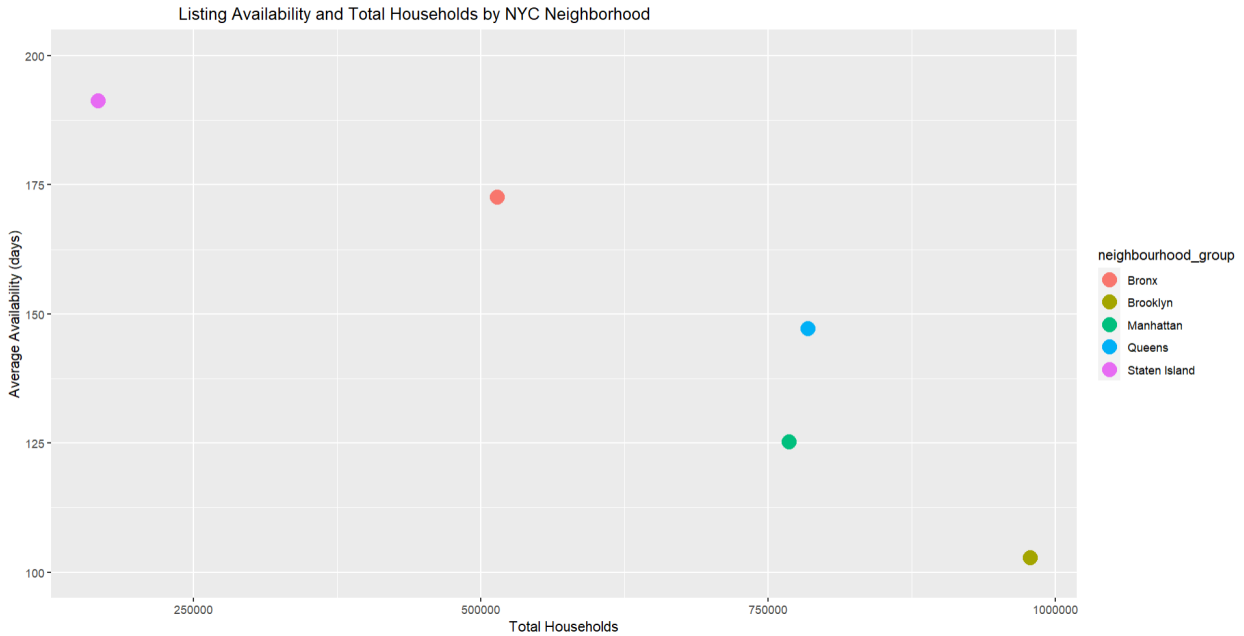

Figure 6 Listing Availability and Total Households by NYC Neighborhood

For Figure 6, we used the scales, ggplot2, and dplyr packages to create a qplot showing average availability and total households. Looking at Figure 6 it is evident that there is a trend that the

more total households there are in a borough, the less average availability. This is interesting because intuitively one would think the more households, the more available Airbnbs, but not according to our visualization.

3.5 Impact of Household Income on Price and Customer Reviews
Is there a relationship between the median or mean income in an area and the number of customer reviews on the listing describing the customer experience? Specifically, do areas with higher median or mean income cause more customer reviews for listings or for average price to be higher also?

In order to determine if a relationship exists between household income and number of customer reviews and also household income and listing price, we used a correlation test. First, we ran a correlation test for Median Income and Number of Customer Reviews. This yielded a correlation coefficient of -0.05. The small negative coefficient allows us to determine that these two variables are slightly negatively correlated. Next, we ran another correlation test between Median Income and Number of Customer Reviews. This yielded a correlation coefficient of -0.06. Another small negative coefficient allowed us to determine that again the two variables are slightly negatively correlated. We decided to utilize both Median Income and Mean Income to determine if there were any major discrepancies between the two. Specifically, to determine if one impacted the test far more than the other.

Another correlation test was used for Median Income and Price. This correlation test resulted in a correlation coefficient of 0.11. This allows us to determine that the two variables are positively correlated. Finally, we ran a correlation test for Mean Income and Price. The correlation coefficient from this test was 0.12, again demonstrating a positive correlation.

Based on our 4 tests, we can determine that Mean Income impacted the test slightly more in the direction of the correlation compared to Median Income. To conclude our findings from the correlation tests, both Mean and Median Income are positively correlated with Price of the listing. This indicates that as Mean or Median Income increases that Price will rise as well. We can determine from this information that listing prices are higher in neighborhoods where people possess higher Mean or Median Incomes. Number of customer reviews will slowly go down as average income gets higher. This indicates that people with higher average incomes may not always be willing to leave a review.

The chart ‘Mean Income and Price’ illustrates that listing prices vary more in the neighborhoods of Queens, Brooklyn, and Manhattan. Data for the remaining two neighborhoods indicate that listing prices vary far less and are lower than $1,500. The chart ‘Median Income and Price’ illustrates similar findings. Prices for each of the five neighborhoods and their respective average incomes are mostly clustered below the price point of $3,750.

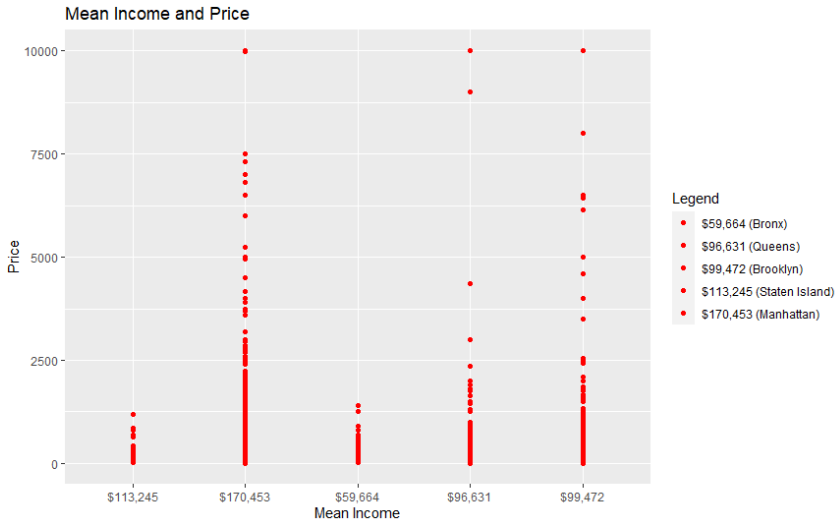

Figure 7 Neighborhood Mean Income and Price

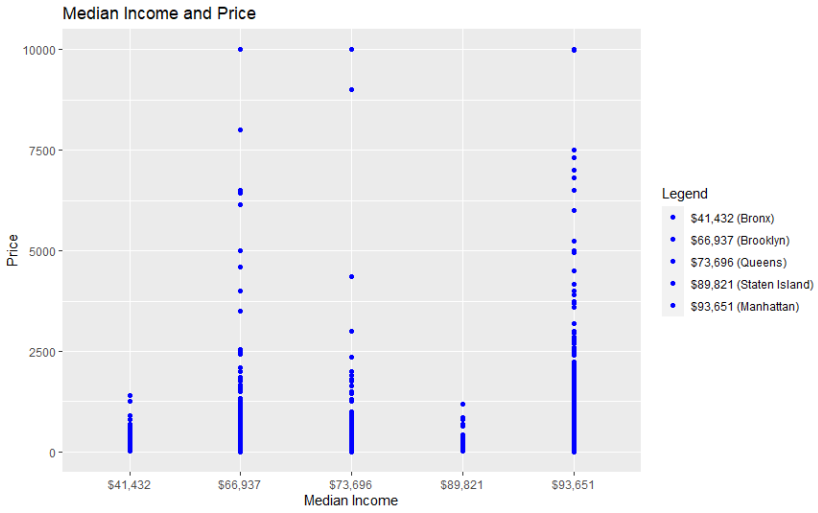

Figure 8 Neighborhood Median Income and Price

The chart ‘Mean Income and Number of Reviews’ illustrates that again, the variance is in the Queens, Brookly, and Manhattan neighborhoods. The number of reviews for the five neighborhoods are clustered mostly below 500. ‘Median Income and Number of Reviews’ demonstrates similar findings.

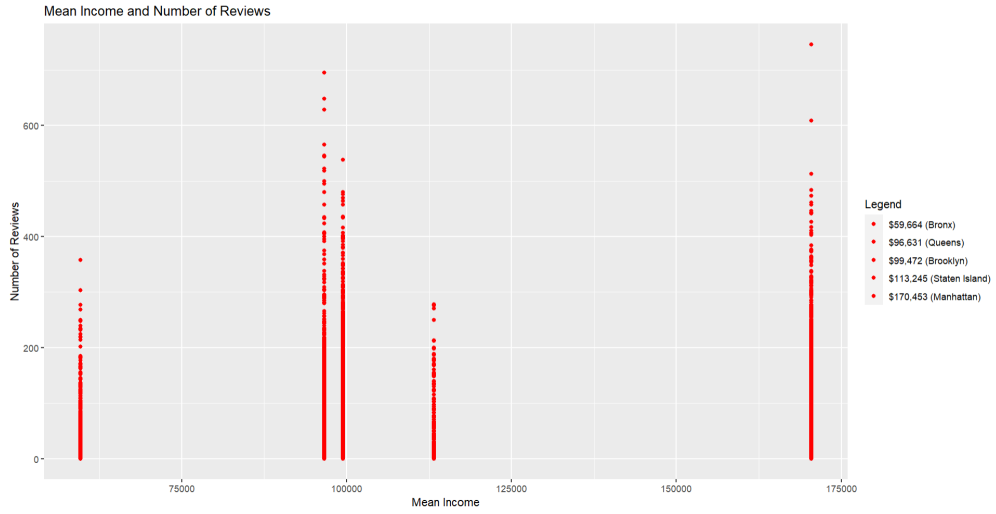

Figure 9 Neighborhood Mean Income and Number of Reviews

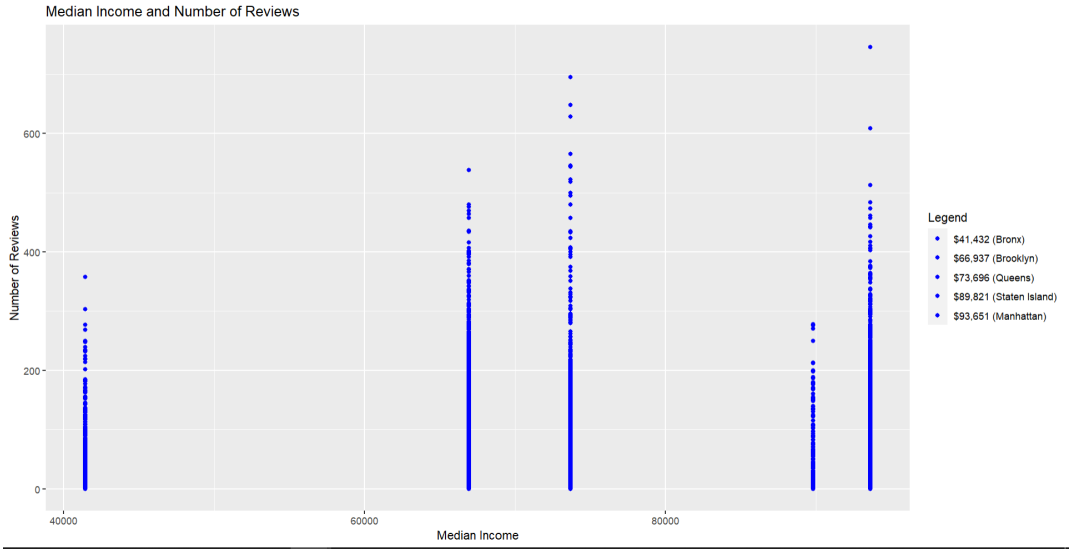

Figure 10 Neighborhood Median Income and Number of Reviews

All code for analysis and visualization is also included in our R script named "FinalProject_Code.R."

4. Conclusion

This project integrates a Kaggle dataset for characteristics of Airbnb listings in New York City with information on New York City household economic data that was collected by a U.S. census survey. We scraped data about the total households and their economic status for the five

unique boroughs located within New York City and horizontally integrated that with our original dataset. This allowed us to study how certain Airbnb listing features are related for the different neighborhoods, as well as analyze how household economics impact the quality and pricing of Airbnb rental units. Specifically, we performed 5 analyses to consider the relationship between Airbnb listing features and neighborhood household data. Our project analysis considered customer reviews for listings, the price of listings by neighborhood, the impact of listing availability, total households by neighborhoods, and the impact of household income on price and customer reviews. Through our summary tables, visualizations, and correlation tests, we found evidence showing the relationship between different Airbnb listings variables. Initially, we discovered that purchasing an Airbnb listing costs around $150, while there is an option to rent a $10,000 unit. The price of listings surprisingly do not impact the number of reviews that much, as on average all listings receive some reviews. Listings located in Staten Island typically receive the most reviews while Manhattan rental units receive the least amount. This was an interesting insight considering Manhattan is the most wealthy area and has the highest average prices for Airbnb listings. Moreover, we considered how different room types impact the number of reviews given for an Airbnb unit. Entire homes and apartments typically obtain the most reviews, as was the case for the neighborhood with the most reviews on average of Staten Island. We concluded that although Manhattan had few reviews for the entire home rental unit type, it may be because customers enjoy the stay and are not inclined to submit reviews. As we continued our study, we found that Bronx listings were typically the cheapest below $100 while the other neighborhood average prices were far below that of Manhattan. The rental type of hotel room has the highest availability while private rooms have the least. By neighborhood, the most availability for a listing is found in Staten Island even though it has the least amount of total houses. We observed a consistent trend of more total households decreasing availability. Next, our group learned that listing prices are higher in neighborhoods with higher household incomes. This insight makes sense, as higher income areas likely have nicer homes and rental offerings for customers. The total customer reviews also decrease as average household income rises.

Overall, we felt our project provided us with significant insights on Airbnb listings in the New York City area and enabled us to utilize our coding skills. Still, a limitation included the dataset we used as well as the data we scraped; those were from over a year ago so we could obtain more current data. The quality of the data also was not the best and could include less N /A values. Another limitation we faced was using data from only NYC, as our conclusions are limited only to this unique area. It would be interesting to see how these conclusions relate to other metropolitan areas. In terms of future work, our team could pursue a similar analysis for other cities. We could tailor our analysis to cities or areas that are popular vacation destinations. Completing this analysis could allow consumers to possess strong information regarding Airbnbs in their destination areas. The information could also be productive to use in order to choose between renting an Airbnb or a hotel room. We believe our final analysis shows the important factors involved in New York City Airbnb listings and would impact customers' interests.