Manhattan more expensive due to 36% more Entire homes/apts

Jeeva Ramasamy

Data Science Capstone Project

Data Driven Report

https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

Background

Airbnb, founded in 2008, revolutionized the lodging industry by introducing a groundbreaking peer-to-peer platform that connects travelers with unique accommodations around the world. Serving as a global community marketplace, Airbnb allows hosts to rent out their properties, whether it be entire homes, private rooms, or shared spaces, offering travelers a diverse range of lodging options beyond traditional hotels. Over the years, the platform has grown into a hospitality giant, disrupting traditional travel norms and shaping the way people explore and connect with new destinations.

In the vibrant landscape of New York City, Airbnb plays a crucial role in shaping travelers' experiences and generating revenue for hosts. Using data-driven insights, we can gain a comprehensive view of trends, pricing dynamics, and host strategies to understand how Airbnb operates in New York City. Analyzing the vast dataset of NYC listings provides valuable information on occupancy rates, popular neighborhoods, and the factors influencing pricing variations. Hosts can leverage data to optimize their property listings, ensuring a competitive edge in a bustling market. Travelers, equipped with insights into average rental prices, seasonal demand fluctuations, and user reviews, can make informed decisions that align with their preferences and budget. The data of Airbnb in NYC thus becomes a powerful tool, enhancing transparency and facilitating a seamless connection between hosts and guests in this dynamic urban landscape.

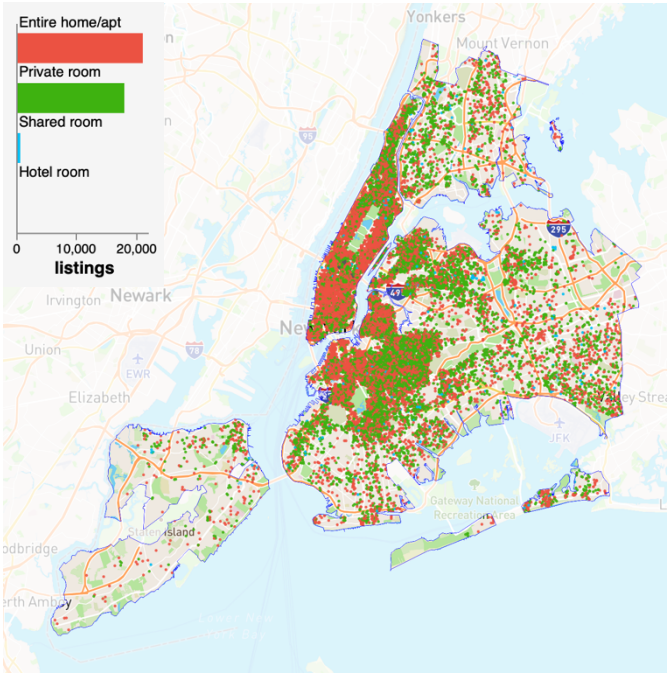
Datasets

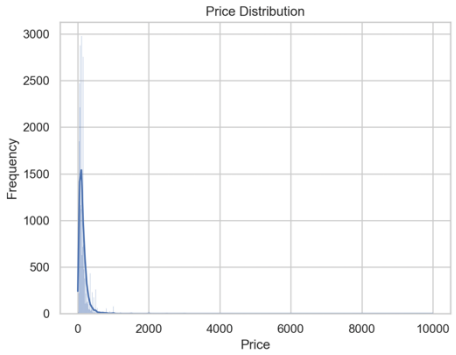
The NYC Airbnb listings dataset is from Kaggle, but the original data (which was cleaned before uploading to Kaggle) is from insideairbnb.com. This dataset has variables such as neighborhood, latitude, longitude, room type, reviews, availability, price, etc. The data does not

include all listings in New York City; it only has listings from 2019, which means it could potentially be outdated. However, the patterns we observe based on neighborhood, room type, etc. would most likely be generalizable to NYC. Additionally, I have included NYC Car Accident data as well as NYC Shooting Incident data (both from Kaggle) to gain a better understanding of correlations we may observe. Using these datasets, we might be able to find some predictors of an Airbnb listing price in New York City.

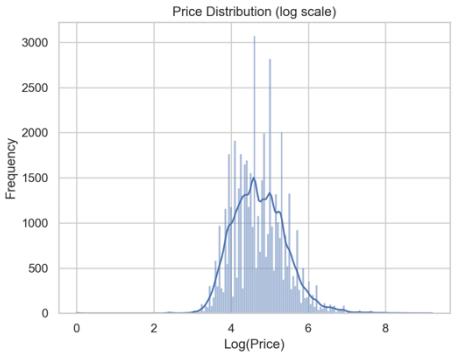
Exploring the Data

Before looking at the variables and their relations, it is best to start with a visualization of the data we are working with. The map to the right is from insideairbnb.com, the origin of this dataset. As we can see from this map, most listings are in Manhattan and Brooklyn. Since there is an obvious bias toward some neighborhoods in the data, we need to keep this in mind while making conclusions about relationships between different variables.

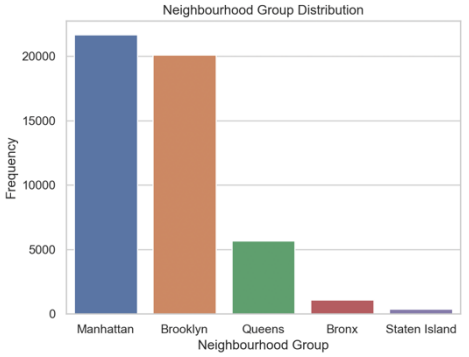


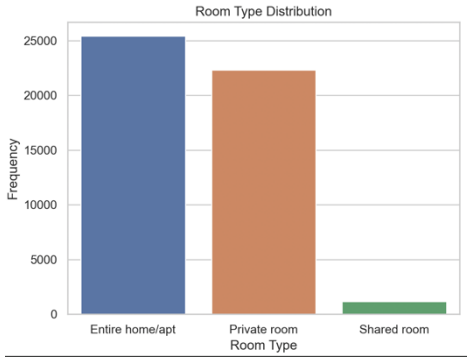


As shown in the first histogram to the left, the distribution of Airbnb listing price is heavily right skewed. To better visualize the distribution of price, we can use a logarithmic transformation on price. This results in an approximately normal distribution on the log scale. This is



important to keep in mind, since any regression to predict price would theoretically perform better using the log transformation.

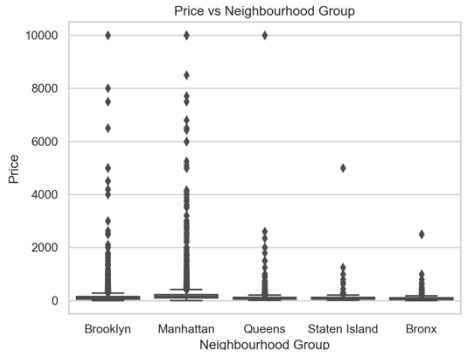


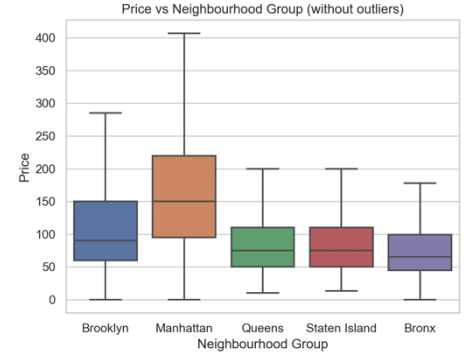


Now, let's look at some bar charts for categorical variables. The bar chart for Neighborhood Group confirms that Manhattan and Brooklyn comprise of most listings in the dataset. There are not many listings in Bronx and Staten Island which might lead to not enough statistical power to detect differences correctly. The same applies to the Room Type variable. Most of the listings are Entire home/apt or Private room with very little Shared rooms.

Hypothesis #1: Manhattan listings cost more than Brooklyn listings on average

The boxplots on the right show the distribution of listing price across all neighborhood groups. Since there are many outliers, which make it hard to distinguish differences, we can plot the same data to only include the box and whiskers. The second boxplot shows a visible difference between Manhattan and the other neighborhood groups. However, let us check this assumption using a t-test. Since Manhattan and Brooklyn have approximately equal sample

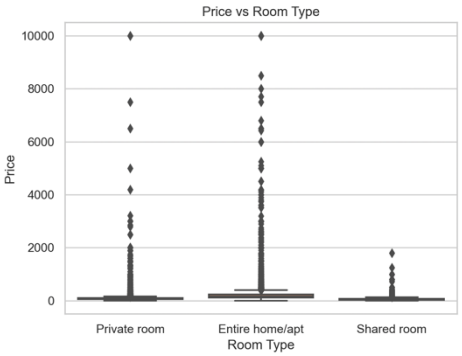


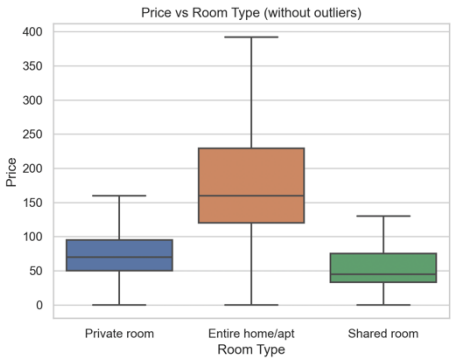


sizes, it would be best to compare these two. This eliminates small sample bias and ensures that results are reliable.

An assumption for the t-test is that the distribution is approximately normal, which requires the log transformation. Since some price values are 0 in the listings, the log of price is not being calculated as intended for these cases. Since there are very few cases, we can remove them from the calculation. The null hypothesis is that there is no difference in price between Manhattan and Brooklyn. The significance level for this test is 0.05. All assumptions for the t-test such as independence, normality, and equal sample sizes are met. The result of the t-test is t-stat: 67.208 and p-value: 0.0. As the p-value is less than the significance level, we reject the null hypothesis. The average listing price in Manhattan is higher than the average listing price in Brooklyn on Airbnb.

Hypothesis #2: Entire homes/apts costs more than Private rooms on average





The boxplots on the right show the distribution of listing price across the different room types. As with the previous boxplots, it is difficult to distinguish differences when there are so many outliers. The second boxplot shows Entire home/apt having noticeably higher prices than both other options. We will test this using a t-test. Since Entire home/apt and Private room have approximately equal sample sizes, it would be best to compare these two. This eliminates small sample bias and ensures that results are reliable.

As with the previous test, we need a log transformation and removal of entries with a price of 0. The null hypothesis is that there is no difference in price between Entire home/apt and Private room. The significance level for this test is 0.05. All assumptions for the t-test such as independence, normality, and equal sample sizes are met. The result of the t-test is t-stat: 170.156 and p-value: 0.0. As the p-value is less than the significance level, we reject the null hypothesis. The average listing price for Entire homes/apts is higher than the average listing price for Private rooms on Airbnb.

Checking Independence using Bayesian Reasoning

We have seen that Manhattan has higher prices on average and Entire homes/apts have higher prices on average. However, how do we know if this is not due to collinearity? To check whether an Airbnb listing in Manhattan is more likely to have Entire home/apt, we can use Bayesian Odds.

room_type	Entire home/apt	Private room	Shared room
neighbourhood_group			
Bronx	379	652	60
Brooklyn	9559	10132	413
Manhattan	13199	7982	480
Queens	2096	3372	198
Staten Island	176	188	9

The contingency table above shows the respective number of Airbnb listings for each category.

WHAT ARE THE ODDS THAT ENTIRE HOME/APT IS IN MANHATTAN?

Belief: Neighborhood group is Manhattan

Observation: Room type is Entire home/apt

Prior Probability = len(df[df['neighbourhood_group'] == 'Manhattan'])/len(df) = 0.44301

Prior Odds = prior/(1-prior) = 0.795

True Positive rate = len(df[(df['room_type'] == 'Entire home/apt') & (df['neighbourhood_group'] == 'Manhattan')])/len(df[df['neighbourhood_group'] == 'Manhattan']) = 0.609

False Positive rate = len(df[(df['room_type'] == 'Entire home/apt') & (df['neighbourhood_group'] != 'Manhattan')])/len(df[df['neighbourhood_group'] != 'Manhattan']) = 0.448

Likelihood Ratio = true_positive/false_positive = 1.359

Posterior Odds = prior_odds * likelihood_ratio = 1.081

Posterior Probability = posterior_odds/(1+posterior_odds) = 0.519

The increase in probability between posterior and prior means that observing that Room Type is Entire home/apt increases the chances of Neighborhood Group being Manhattan. This means that Entire homes/apts are more common in Manhattan than in other neighborhood groups in New York City.

WHAT ARE THE ODDS THAT PRIVATE ROOM IS IN BROOKLYN?

Belief: Neighborhood group is Brooklyn

Observation: Room type is Private Room

Prior Probability = len(df[df['neighbourhood_group'] == 'Brooklyn'])/len(df) = 0.411

Prior Odds = prior/(1-prior) = 0.698

True Positive rate = len(df[(df['room_type'] == 'Private room') & (df['neighbourhood_group'] == 'Brooklyn')])/len(df[df['neighbourhood_group'] == 'Brooklyn']) = 0.504

False Positive rate = len(df[(df['room_type'] == 'Private room') & (df['neighbourhood_group'] != 'Brooklyn')])/len(df[df['neighbourhood_group'] != 'Brooklyn']) = 0.424

Likelihood Ratio = true_positive/false_positive = 1.190

Posterior Odds = prior_odds * likelihood_ratio = 0.831

Posterior Probability = posterior_odds/(1+posterior_odds) = 0.454

The increase in probability between posterior and prior means that observing that Room Type is Private room increases the chances of Neighborhood Group being Brooklyn. This means that Private rooms are more common in Brooklyn than in other neighborhood groups in New York City.

Conclusion

Based on the Statistical testing and Bayesian reasoning (as well as common sense), we can infer that Entire homes/apts cost more on Airbnb than Private rooms on average. This is because the rented area will most likely be larger, hence the higher price. However, we also noticed that Manhattan has higher prices than Brooklyn. This could partially be attributed to the correlation to room type as we saw in the previous section.

I have included additional datasets on NYC boroughs to get better insights into this phenomenon. According to the Car accident dataset, 22.3% accidents occur in Brooklyn, 18.9% accidents in Queens, 18.4% accidents in Manhattan, 9.6% accidents in Bronx, and 3.4% accidents in Staten Island. According to the Shootings dataset, 41.2% shootings occur in Brooklyn, 28.6% shootings in Bronx, 14.9% shootings in Queens, 12.2% shootings in Manhattan, and 3.0% shootings in Staten Island. Based on this data, Brooklyn seems to also be a more dangerous place to visit in New York City compared to the other boroughs. This could also attribute to the lower prices in the listings as less people would be interested in staying there.

Since this data contains only information from New York City, the findings can only be generalized to NYC listings. Moreover, it is important to acknowledge that there may be other variables not considered in this study that could have an impact on price. The understanding of the factors that influence Airbnb listing prices can help both hosts and customers make data-driven decisions.

References

https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

https://www.kaggle.com/datasets/pavetr/nypdcollisions

https://www.kaggle.com/datasets/thaddeussegura/new-york-city-shooting-dataset

http://insideairbnb.com/new-york-city

https://news.airbnb.com/about-us/