ORIE 3120 Milestone 1
Matan Auerbach, Sydney Ho, Serena Huang, Owen Rector

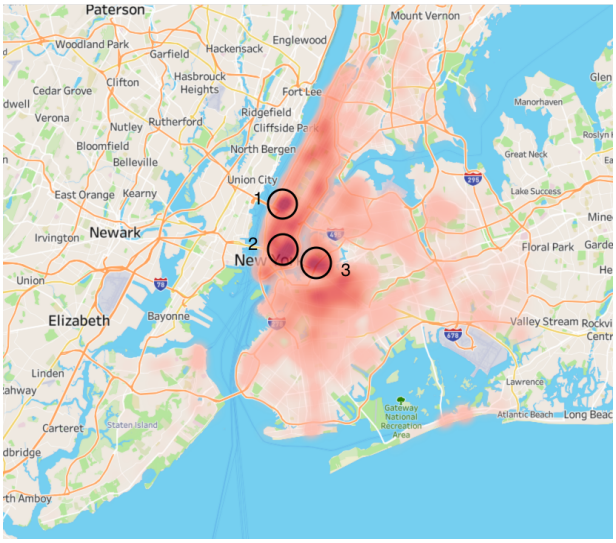
Dataset: **NYC Airbnb Data**

We chose to analyze a dataset consisting of various Airbnb listings in New York City from the year 2019. This dataset has 16 features and includes information such as borough, neighborhood, coordinates, price, review metrics, and more—each of which can be used for predictions, inference, and analysis. There is data for over 45,000 Airbnb listings across each of the five New York City boroughs. Using this dataset, we hope to explore and answer the following questions:

1. Are there certain neighborhoods that are more popular for Airbnb listings?
2. How do the prices vary across different neighborhoods/room types in NYC?
3. Can we predict the daily price of an Airbnb listing based on its location, room type, review metrics, and other features?

To initiate the process of answering these questions, we have conducted preliminary data cleansing and produced the following visualizations:

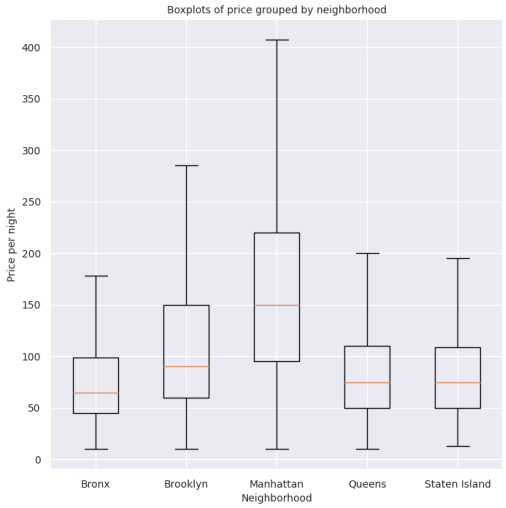
Question 1: Are there certain neighborhoods that are more popular for Airbnb listings?



The heat map was created via Tableau by plotting all of Airbnb listings on the map by their coordinates and mapping more specifically by density. Darker areas on the map correspond to areas with a higher number of listings and lighter areas correspond to areas with a lower number of listings. There are several areas on the heat map that have a higher density of listings, however there are three noticeable high-density neighborhoods that we will focus on: Hell's Kitchen (1), East Village (2), and Williamsburg (3). These neighborhoods are the most popular for Airbnb listings and this could be due to a number of factors such as proximity to public transportation and tourist attractions. All three of these neighborhoods have various options for public transportation, having easy access to MTA subway lines and buses. These neighborhoods are also close to popular tourist attractions such as Times Square with Hell's Kitchen, SoHo with the East Village, and the Brooklyn Bridge with

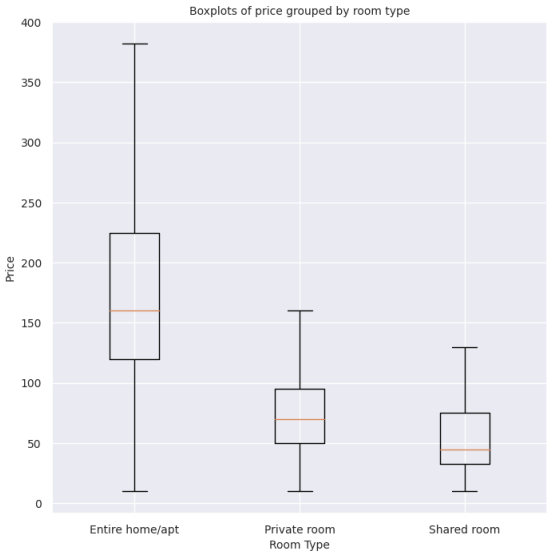
Williamsburg. The information found through this heat-plot will be beneficial later on when attempting to make further inferences and predictions based on the locations of Airbnbs in NYC.

Question 2: How do the prices vary across different neighborhoods(boroughs)/room types in NYC?



neighbourhood_group	count	mean	std	min	25%	50%	75%	max
Bronx	1089.0	85.361800	77.761144	10.0	45.0	65.0	99.0	1000.0
Brooklyn	20050.0	118.303890	96.670552	10.0	60.0	90.0	150.0	1200.0
Manhattan	21526.0	180.638066	139.723956	10.0	95.0	150.0	220.0	1200.0
Queens	5656.0	95.008310	74.527596	10.0	50.0	75.0	110.0	1000.0
Staten Island	371.0	98.584906	96.138752	13.0	50.0	75.0	109.0	1000.0

To help answer this question, we first created a boxplot by grouping data by borough to represent the variation in prices of Airbnb's across the differing boroughs. I also printed out the summary statistics to interpret the plot. This plot helps us see clearly which boroughs have the highest average prices, most variation, and overall trends. Each box on the plot represents the interquartile range (IQR) of prices for each borough, with the whiskers extending to the minimum and maximum values that are within 1.5 times the IQR. In this particular plot, outliers of this whisker range have been excluded. The box plot allows for a visual comparison of the median prices and spread of prices across the different neighborhoods. In this graph, we see that the highest median (the yellow line on the plots) is found in properties in Manhattan (150), with Brooklyn (90), Queens(75), Staten Island (75), and the Bronx (65) following in that order. Additionally, we see that to go along with the highest median, Manhattan also has a standard deviation (139) that exceeds all other boroughs, with a decent proportion of their properties above the 75th percentile range of all the other boroughs individually. This finding is generally consistent with the popularity of Manhattan as a tourist destination.



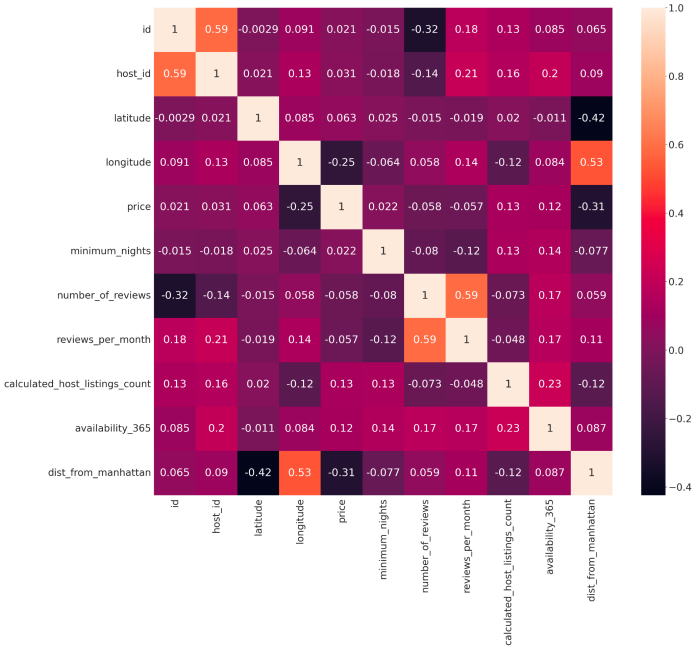
room_type	count	mean	std	min	25%	50%	75%	max
Entire home/apt	25259.0	196.212914	131.844884	10.0	120.0	160.0	225.0	1200.0
Private room	22277.0	84.969655	70.064162	10.0	50.0	70.0	95.0	1200.0
Shared room	1156.0	67.731834	81.076816	10.0	33.0	45.0	75.0	1000.0

To supplement the first boxplot, we also grouped by room type to create an alternative visualization which depicts the distributions of price across the three different room styles: Entire Home/Apt,

Private Room, and Shared Room. From the result, we see a clear disparity between room type and price. Airbnb properties offering up the Entire home/apt have a significantly higher mean (196) and median(160) than Private Rooms and shared rooms as seen in the summary stats. This makes sense, because likely properties that offer the entire room not only offer the chance for a significantly larger number of guests than one private or shared room, but are more luxurious and private—which would drive up the price per night.

Overall, these results are helpful in identifying how the different room types and apartment locations contribute to the price, and will come in handy when later performing predictive analysis to attempt to predict the prices based on variable values. It is important to note here that the room_type variable and borough variables are strictly categorical, and will need to be converted into numerical values of some sort before running them on a machine learning model.

Question 3: Can we predict the daily price of an Airbnb listing based on its location, room type, review metrics, and other features?



The first visualization we created to help answer this question was a correlation matrix. We did this to see the ways that the different numerical variables interact with one another, with the goal of eventually building a machine learning model to predict prices of Airbnb properties. The resulting heatmap shows the correlation coefficients between all pairs of numeric variables in the DataFrame. Interestingly enough from this matrix, we see that there are very few variables that have any strong interactions with one another (R^2 values close to 1). One of the strongest correlations we see is the relationship between dist_from_manhattan and price. Dist_from_manhattan was a variable that we added to the dataset (using the Times Square Lat/Long coordinates) in order to visualize how properties closer and further away from Times Square might vary in price. Its relationship with price shows a negative correlation coefficient of -0.31. All other interactions are mostly insignificant. To visualize this interaction a bit more, the next visualization shows a scatter plot mapping price vs. dist_from_manhattan.

In this visualization, similar to how we predicted, we see a negative correlation between the variables. Generally, the most expensive properties (restricted at 1200 dollars) appear to be in Manhattan close to Times Square, and as you get further and further away their price drops. This is an interesting observation, and will be instrumental in the future as we attempt to answer our overarching question about being able to predict prices based on the variables we have at hand in this dataframe. Using this scatter plot paired with the correlation matrix and box plots we generated above, we have a solid starting point for attempting to construct a model that has a chance at being able to predict prices of Airbnb properties in New York City.

	price	number_of_reviews	reviews_per_month
count	38337.000000	38337.000000	38337.000000
mean	129.497457	29.464590	1.377972
std	87.616212	48.385119	1.685298
min	10.000000	1.000000	0.010000
25%	68.000000	3.000000	0.190000
50%	100.000000	10.000000	0.720000
75%	165.000000	34.000000	2.030000
max	599.000000	629.000000	58.500000

```
Text(0.5, 1.0, 'Price vs. Number of Reviews')
```

In this visualization, we see a visible yet slight negative relationship between the price and number of reviews. Prior to running any visualizations, we had two major hypotheses about what this relationship would look like. Either a higher number of reviews would drive the price of the Airbnb higher, or the properties with the greatest number of reviews would reflect the most affordable prices. Moving forward, we can perform regression and causality analyses, in conjunction with other key variables to avoid the effects of confounding variables, in order to evaluate whether the number of reviews has a statistically significant impact on the price of an Airbnb. This analysis and information is important because it determines how impactful the feedback system on the Airbnb platform is. Not only may the quantity of reviews be interesting to look at, but this also poses the opportunity to work with text mining and see if the actual content in the reviews also matters. With future exploration of reviews data, we can develop insights regarding the veracity of review data and which factors have the most significant impact on an Airbnb's listing price.