Introduction

The Problem: An individual in the New York City area is thinking about purchasing some property that they would then like to list on the mobile app Airbnb. They want to know the feasibility of such an idea in that does it make sense for them to list their property. What are the other types of properties in the New York City area that generally make money? Where are most of the properties in New York listed?

Ultimately the problem is this: Where should the individual buy their property, and what type, in the New York City area prior to listing?

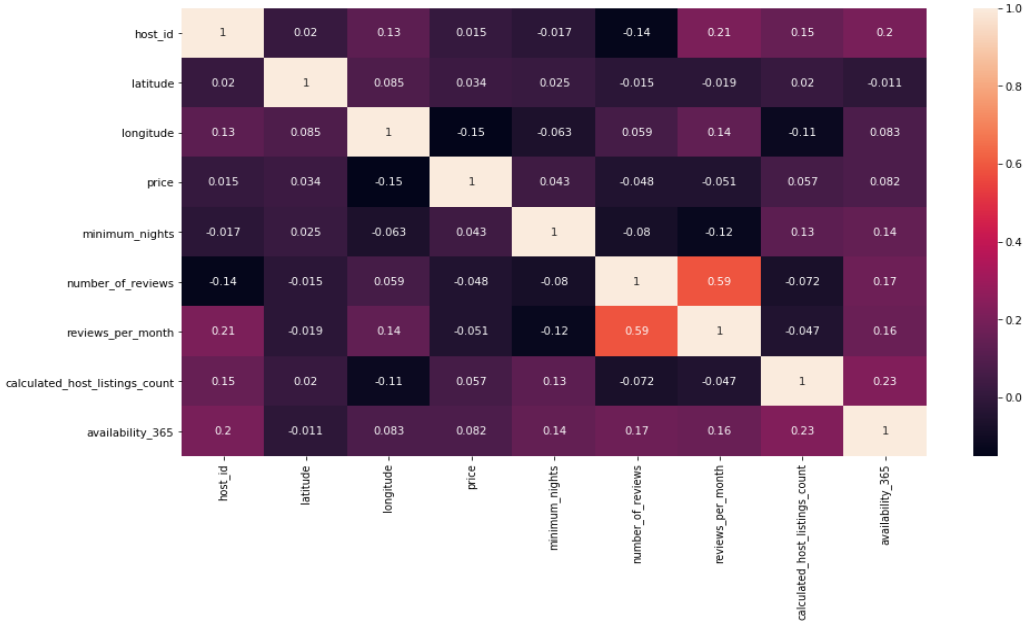
Data

The data used for this project will come from an Airbnb data set that was published to the data science website Kaggle (link to the data set will be posted below). The data set includes the following information: room type, location of property listings (broken in to neighborhood groupings and can be drilled down to individual neighborhoods as well), pricing, availability, and miscellaneous items such as number of reviews, reviews per month, and the date the last review was submitted. As part of our analysis, the miscellaneous columns from the data set will be removed as they add no value. Latitude and longitude data were also provided, which allows us to map exact locations of all the listings in the New York City area.

Methodology

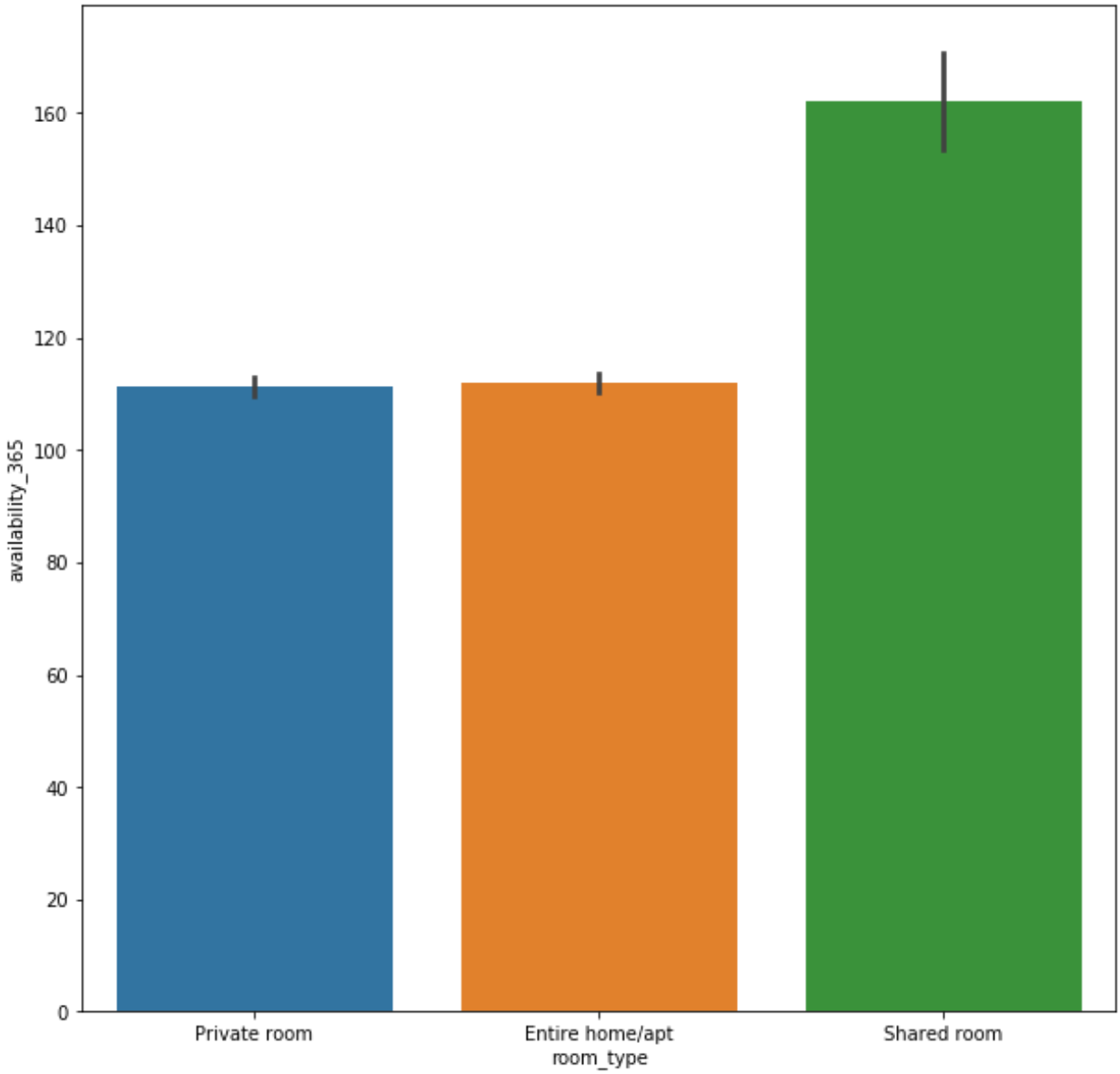
The first thing that was done was to import the Excel file in to a Pandas data frame. Once completed, the data cleaning process began. As mentioned above, there were a handful of miscellaneous elements that were irrelevant to the analysis, so they were removed from the data set. Those elements were: name, id, host_name, last_review. 'Name' provided a description of the property while id was the ID number of the property itself; host_name is self-explanatory. In terms of data cleaning, the final thing that was done was to remove all null values. The variable review_per_month contained thousands of null values which were replaced with 0;

Once the data was cleaned, I began an exploratory data analysis. A heat map indicated that there were very weak correlations between the variables in the data set, giving us an early indication that the process by which we were going to identify a location to purchase a property might not be sounds.



I then broke out the data set to find the neighborhood groups, of which there were 5. A box plot showed that 3 of the neighborhoods had a rental property priced at $10,000, while Manhattan seemed to house the priciest rentals on average; it also had the most rentals out of the 5 neighborhoods listed in the data set. Next, a scatter plot was created to check what the most popular type of room was, and it was clear that an entire home or an apartment was the rental of choice relative to a private room or shared room.

When checking the availability of the 3 room types a simple bar chart showed that shared rooms tended to be more available than a private room and entire homes, indicating that those room types did not seem to be in high demand.



All this exploratory analysis showed that the best place to purchase a property would be in Manhattan.

Methodology & Results

The next portion of this problem utilized regression to confirm that Manhattan was the best place to buy a property and to identify which neighborhood and type of property to purchase. The first regression run was a simple linear regression. Dummy variables were created for the 'neighbourhood_group' and 'room_type' variables in the data set. Generally, there should be k-1 dummies, but this analysis only used 1 for each column when there should have been 4 and 2, respectively. The coefficients of the linear regression showed that purchasing a property in Bronx would lead to a decrease in price by approximately $57, while purchasing an entire home to rent out would result in a price increase of about $121. The linear regression produced an r-squared score of .08, meaning that only 8% of the observed

variation can be explained by our model's inputs. Such a low score signals that this model would generally not be considered a good one.

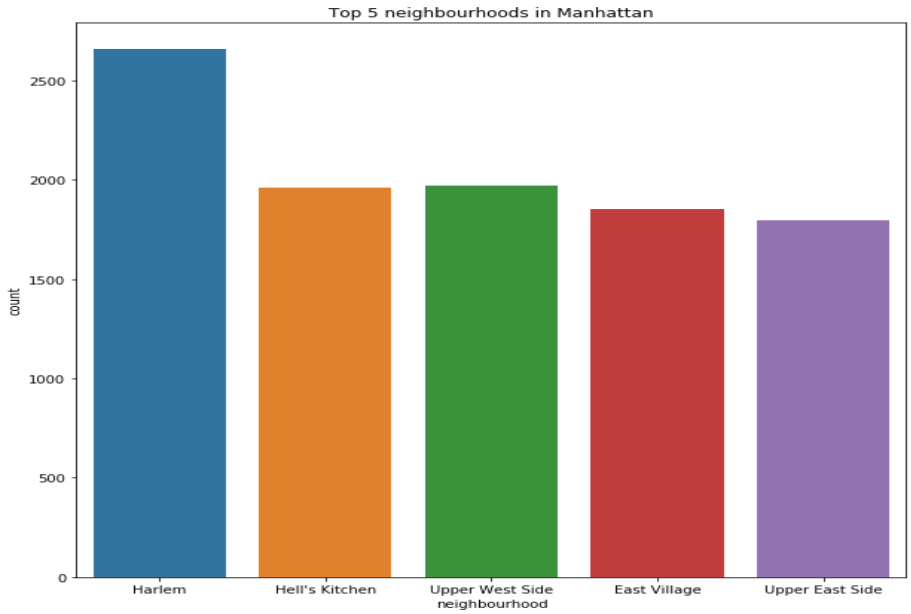
```
reg = LinearRegression()
reg.fit(x_train,y_train)
y_pred = reg.predict(x_test)
r_score = r2_score(y_test,y_pred)

print('Coefficients: ', reg.coef_)
print('R Squared: ', r_score)
adjusted_r_squared = 1 - (1-r_score)*(len(y)-1)/(len(y)-X.shape[1]-1)
print('Adjusted R Square: ', adjusted_r_squared)
```

```
Coefficients:  [-5.72286705e+01  1.21972370e+02  1.55098041e-01  6.00017281e-02
  1.46366618e-01]
R Squared:  0.08960224758415636
Adjusted R Square:  0.08950913893472434
```

The next regression run was a decision tree. I thought a decision tree regression would be worthwhile in this situation because it would break down the data in to nodes and look at the different possible combinations of price based on location, room type, etc. The resulting r-squared scored, however was even *lower* than that of the simple linear regression meaning that the model is essentially useless. Running a logistic regression resulted in something similar, as it produced a r-squared of .015.

Results

Out of the regressions run the linear regression proved to be the most viable, albeit with an extremely low r-squared of .08. The location of where to purchase the property was drilled down even further by focusing on the top 5 neighborhoods in Manhattan. After running another linear regression which focused on only the top 2 neighborhoods in Manhattan (Harlem and the Upper West Side), it was found that the best place to purchase a property to rent would be in Upper West Side of Manhattan. The exploratory analysis that was conducted said as much, but doing further analysis using regression confirmed the hypothesis. It was found that renting a property in Harlem as opposed to the Upper West Side would result in a price decrease of $72, while renting out a full home would result in a price increase of approximately $129.



```
manhattan_reg = LinearRegression()
manhattan_reg.fit(x_train, y_train)
y_pred = manhattan_reg.predict(x_test)
top_2_score = r2_score(y_test,y_pred)

print('R2 Score: ', top_2_score)
print('Coefficients: ', manhattan_reg.coef_)
```

```
R2 Score:  0.07456721496803431
Coefficients:  [-71.67992277 128.99411887  -0.13585938  -0.33836985   0.2443977 ]
```

Discussion

After review of the results, we would recommend to a potential buyer that they purchase an entire home or apartment in the Upper West Side of Manhattan to maximize the rental price. The buyer should be made aware the the suggestion does not come with a high degree of confidence because, as mentioned before, the resulting r-squared was fairly low. One potential analysis that could have been done would be a chi-squared goodness of fit test, which would have shown us how different the observed (y_hat) value was from the expected. There were also a couple of of variables used in the analysis that maybe should not have been. Namely 'calculated_host_listings_count' and 'availability_365'. These 2 variables seemed to not only have a marginal impact on the price, but a follow question would be to see if removing those to variables would result in a higher r-squared across all the regressions that were run. Ultimately, however, the recommendation remains that a potential buyer should look at purchasing a house in the Upper West Side of Manhattan.

Conclusion

The Airbnb data set downloaded from Kaggle allowed us to pose a simple question that many prospective New York home owners might ask themselves: If I am going to buy a property to rent out, where should I buy it and how much should I charge? The data set showed us that the best place to rent out property is in the Upper West Side of Manhattan; the type of rental should be either an entire home or apartment. This would result in maximizing the price point. When looking across the 5 neighborhoods, renting property in the Bronx would result in the largest price decrease while renting a shared room would result in weak demand. There were certain variables used in the regression analysis that, in hindsight, did not seem like adequate predictors of price. In future analysis, variables such as average rating and proximity to entertainment would help better predict price and allow us to more accurately selection a proper location to have a rental property.