DSGA 1001 CAPSTONE REPORT

Analysis of NYC Airbnb Data

Team SEED

Sarvesh Patki (ssp6603)

Eugenia Fomitcheva (edf257)

Elizabeth Wheeler (eaw9166)

Dhruv Saxena (ds6802)

INTRODUCTION

Airbnb is a vacation rental company that operates an online marketplace and primarily deals with short-term, temporary rentals of privately owned properties. Since all of our capstone members have first-hand experience with Airbnb's services, we were interested in analyzing their data. For the purpose of this project, we found Airbnb's rentals in New York City and its 5 boroughs. The dataset was sourced from kaggle.com and describes listings in NYC in 2019. This is a cleaned version of the raw data that was released by Airbnb, and therefore can be vouched for its accuracy and relevance. We begin our project by first understanding the nuances of the data and how users respond to various listings. We then look into price predictability given a variety of numerical and categorical variables using regression and machine learning methods.

DATA

There are **48895 rows** in the dataset, each of which represents a unique listing. There are **16 columns**, each of which represents a unique feature, which has been described in the table below:

Column Name	Description	Type
id	Airbnb's unique identifier for the listing	int64
name	Name of the listing	string
host_id	Airbnb's unique identifier for the host/user	int64
host_name	Name of the host. Usually just the first name(s).	string
neighborhood_group	The neighborhood group geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.	string
neighborhood	The neighborhood is geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.	string
latitude	Uses the World Geodetic System (WGS84) projection for latitude and longitude.	float64
longitude	Uses the World Geodetic System (WGS84) projection for latitude and longitude.	float64
room_type	[Entire home/apt | Private room | Shared room | Hotel] All homes are grouped into the following three room types: Entire place/ Private room/ Shared room/ Entire place	string
price	daily price in local currency	int64
minimum_nights	minimum number of night stay for the listing (calendar rules may be different)	int64
number_of_reviews	The number of reviews the listing has	int64
last_review	The date of the last/newest review	object
reviews_per_month	The number of reviews the listing has over the lifetime of the listing	float64
calculated_host_listings_count	The number of listings the host has in the current scrape, in the city/region geography.	int64
availability_365	The availability of the listing 365 days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.	int64

HYPOTHESIS TESTING

We were mostly interested in analyzing the relationship between the price of a listing with factors such as the neighborhood where the listing is located, average monthly ratings for the listing, the minimum number of nights for the listing, and the number of listings by the host. This would help us better understand some of the dynamics of the dataset and set the stage for further exploratory analysis.

***For all the tests conducted, level of significance (α) is set to 0.005.**

Hypothesis Test 1: Does the price of the listings vary by neighborhood groups (i.e boroughs)?

First, we divided the listings borough-wise and removed listings where the price was not mentioned or was zero. Our **null hypothesis** is that the **price of listings does not vary across the boroughs**. In order to answer this question, we ran a two-sided Kruskal Wallis test. Given our group standard deviations were not equal, we decided it was not appropriate to use an ANOVA. Therefore, we instead chose to implement the non-parametric version of ANOVA, known as the Kruskal-Wallis test. It is worth noting that the KW test's non-parametric nature results in a loss of power and that the number of samples in our borough groups is not equally distributed. However, we felt the collective dataset size and comparable number of listings in Manhattan and Brooklyn was sufficiently large to mitigate this. In running the test, we obtain a p-value that is significant at the alpha level of 0.005. Hence, we **reject the null hypothesis**. Therefore, there is evidence that suggests **the price of listings varies by the boroughs they are located in**. This result is not surprising as we know some neighborhoods in New York tend to be more expensive given general desirability, location, better maintenance, and facilities/amenities.

Hypothesis Test 2: Do listings with more reviews have higher prices than listings with less reviews?

For this test we performed a median split and split the listings into two groups: listings with reviews more than and less than the media. Our **null hypothesis** is that **listings with more reviews do not have higher prices than listings with fewer reviews**. Our alternative hypothesis is that listings with more ratings have higher prices. We ran a one-tailed independent samples t-test on the groups and obtained a p-value < 0.005. Therefore, we **reject the null hypothesis**. Thus, the data suggests that **listings with more reviews have higher prices than listings with less reviews**. While the number of reviews per listing was available, the nature of these reviews (positive/negative/neutral) was not. Therefore, we are unable to make any claims regarding the sentiment of Airbnb reviews and price - perhaps some listings exceeded expectations and garnered (positive) reviews while average to below average listings received no attention. This is simply a conjecture, and a solid conclusion cannot be drawn without further information or actual reviews for the listings.

Hypothesis Test 3: Do listings with a higher number of minimum nights and listings with a lower number of minimum nights have different prices?

Again, we applied a median split to our data in order to create two groups: listings with a higher number of minimum nights and listings with a lower number of minimum nights. Our **null hypothesis** is that **listings with a higher number of minimum nights and a lower number of minimum nights do not have different prices**. We ran a 2-tailed independent samples t-test on the two groups and obtained a p-value > 0.005, hence leading us to **fail to reject the null hypothesis**. Thus, **we cannot conclude that listings with a higher number of minimum nights and those with a lower number of minimum nights have a difference in prices.**

Hypothesis Test 4: Do hosts with more listings charge higher prices than hosts with less listings?

First, we performed a median split of the data, dividing into two groups: those with low host listing counts and those with high host listing counts. Then we removed the duplicate entries in our data so that only unique host ids are considered. Our **null hypothesis** is that **listings by hosts with more listings are not priced higher than listings by hosts with less listings.** We ran a one-tailed independent samples t-test on the groups and obtained a p-value < 0.005, leading us to **reject the null hypothesis.** Thus, **listings by hosts with more listings are priced higher than listings by hosts with less listings.** This may be due to the reason that hosts with more listings are affluent and are able to maintain their properties and hence charge higher prices. It is also possible that hosts with more listings earn a major chunk of their income from airbnb rentals and treat it as a business. They may have purchased properties in prime locations where prices are expected to be higher and given the demand. These are just some potential driving factors that could explain a result that is unlikely due to chance alone.

Hypothesis Test 5: Are listings that contain the word 'luxury' priced higher than listings that contain the word 'modern'?

While we were perusing the data, we realized that a lot of listings mentioned the words 'luxury' and 'modern'. So, we were curious as to whether there is a significant difference in prices between listings that contain these two words. Furthermore, from a significance testing standpoint, the number of listings with each of these words was in the context of 2,000, making our two testing groups even in magnitude. Our **null hypothesis** is that **listings that contain the word 'luxury' are not priced higher than listings that contain the word 'modern'.** Running a one-tailed Mann Whitney U-test we got a p-value < 0.005 and hence we **reject the null hypothesis**. Thus, this suggests **listings that contain the word 'luxury' are priced higher than listings that contain the word 'modern'**. In running a two-tailed Mann Whitney U-test, we found that the prices of these two types of listings are different. It seems that people are willing to pay more for properties described as luxurious versus those described as modern. In fact, the mean price for the 'luxury' group was in the context of $238 a night versus $150 for the 'modern' group.

REGRESSION

***For all of our models we used a standard train:test split of 0.8:0.2.**

Following our hypothesis testing, we were interested in implementing a **multivariable regression** model to predict the price of listings based on the following 7 parameters: 'minimum_nights', 'number_of_reviews', 'neighborhood', 'neighborhood_group', 'accomodation', 'availability_365', 'calculated_host_listings_count'. In order to do this, we first needed to perform data pre-processing in order to encode categorical variables such as 'neighborhood', 'neighborhood_group', and 'accomodation' numerically. At this stage, we also ensured that we had no missing data, and then trimmed the data for all future modeling use by setting a minimum 5% and upper 95% threshold to reduce outliers. The resulting distribution of listing price is given in **Figure 0**.

The multivariable linear regression model for this catered dataset resulted in an R^2 score of 0.3491.

We also ran **ridge and lasso regression** for alpha = [1e-5, 0.1, 1, 10, 20] and [1e-3, 1e-2, 1e-1, 1] respectively, and found that for **ridge** we obtained the lowest test error at a high alpha value of 20 while for **lasso** the lowest test error was obtained for alpha = 1e-1. In future explorations, it may be worthwhile considering a log transformation of the data to normalize the distribution and help remove some of the skewness we observe here.

MACHINE LEARNING

After performing regression analysis, we decided to perform a principal component analysis and run the k-means clustering algorithm. While predicting price given our independent variables has proven to be a challenging task, we felt compelled to better understand what variables were contributing to the variance of each principal component and whether the k-means algorithm could provide us with interesting visual insights.

Principal Component Analysis

For the PCA, we considered only the columns which are numerical in nature: 'minimum_nights', 'number_of_reviews', 'availability_365', 'calculated_host_listings_count', and 'listing_price'. First, we normalized the data as a prerequisite for running the PCA. Next, we found the eigenvalues and got the principal components. We plot a scree plot, and based on the **Kaiser rule we opted to select only the first two principal components (Figure 1)**. The loadings matrix also helped us semantically interpret the correlation between the five columns that we had chosen and the principal components that we selected. We can see that 'calculated_host_listings_count' (followed closely by 'availability_365') contributed most significantly to Component 1 while 'number_of_reviews' contributed most to Component 2 **(Figure 2)**. In **Figure 3** we can see this shown visually with vectors.

Clustering

We then ran an unsupervised machine learning algorithm called k-means which is used for clustering data to examine if our data has an inherent structure. To find the optimal number of clusters for our data we first plotted the **'total within-cluster distance'** versus the number of clusters. We can see that there is **no pronounced elbow**, although by 'eyeballing', we can estimate that the optimal number of clusters is around three **(Figure 4)**. Given the lack of a definite optimal value, we also used the **silhouette method**. We can see from the plot that the peak of the silhouette scores on a range of 2 to 10 occurs at k = 3 **(Figure 5)**. Therefore, we chose to run k-means on our two chosen principal components with 3 clusters, the result of which is reflected in **Figure 6**. While the result certainly divides our datapoints with limited overlap, overall it seems that this data could also be represented as one cluster.

Further Machine Learning Methods

As we previously mentioned, our goal in this project was to try to predict the price of listings based on the various numerical and categorical parameters available to us. On the basis of our

dimensionality reduction results obtained with the PCA, we created five models and compared the results using R^2 methodology. The first of these was a return to the **multivariable regression** we looked at previously, but this time **using just two principal components**, i.e. two predictor variables. The result was frankly bad, as demonstrated in **Table 1**. Both our train and test errors were very high with an extremely low R^2 value of 0.026. This indicated to us that while the variables we chose to run this regression with explained the greatest proportion of variance in the PCA, they have little predictive power when it comes to predicting listing price.

In order to achieve a regression model similar to our original which had 7 predictors, we chose to **include the categorical variables along with those from the PCA, resulting in a 5-predictor model**. This proved to drastically improve the 2-variable model, achieving an R^2 much like we had seen before. From these two linear regression models we concluded that our categorical variables - 'neighborhood', 'neighborhood_group', and 'accommodation' would be needed in further price prediction models.

As a final step in seeing whether we could improve upon our model, we implemented supervised learning models best suited for regression and working with categoricals - **a neural network architecture**, **a decision tree, and a random forest**. The results of the neural network with (5,5) hidden layers were marginally better than those achieved for multivariable regression using categorical variables although the simple decision tree with a max depth of 10 yielded the best R^2 of our algorithms. The results for all these implementations have been tabulated in **Table 2**.

Table 1

Algorithm	R^2 score
MLR with PCs Only	0.0260
MLR with PCs and Categorical variables	0.3414

Table 2

Algorithm	R^2 score on training data	R^2 score on test data
Neural Network	0.3613	0.3573
Decision Tree	0.4978	0.4443
Random Forest	0.4231	0.4088

CONCLUSION

We began with hypothesis testing where we posed certain questions about the relationship between price and various factors to gain an understanding as to the relationships between variables in our dataset. While we answered some interesting questions, there are certainly opportunities for expanding this dataset to gain more insights. For example, having actual reviews for some listings would be interesting as we could conduct sentiment analysis and ask questions like 'do high rated listings or listings with positive reviews tend to have higher listing prices?'.

Coming to the regression analysis, as we saw, the features that we had were not optimal for predicting Airbnb listing prices. There are certainly factors with potential to influence the prices that the dataset did not include as columns such as the safety of the neighborhood, proximity to tourist spots or public transport facilities/ grocery stores, whether there were any overhead charges to be borne by the tenant, and likewise. Furthermore, even with this information, predicting Airbnb listing prices could prove to be a challenge as ultimately the host sets the price, and not a predetermined algorithm.

Finally, in the machine learning section, we analyzed the data by finding the principal components and using them along with categorical data to try and enhance our predictive model. While we found that the categorical variables are valuable in the model and were able to achieve some improvement in combining these with variables found in our PCA, with the best model being the decision tree, ultimately we concluded that predicting Airbnb listing prices is simply a hard task.

The future scope of this project would include *i)* expanding our dataset if possible with listing reviews and ratings to see whether these have better price prediction capabilities and *ii)* broadening our analysis to multiple cities in the US in order to draw a comparison between Airbnb listings by region or seeing if other cities have more consistency in their listings.

APPENDIX

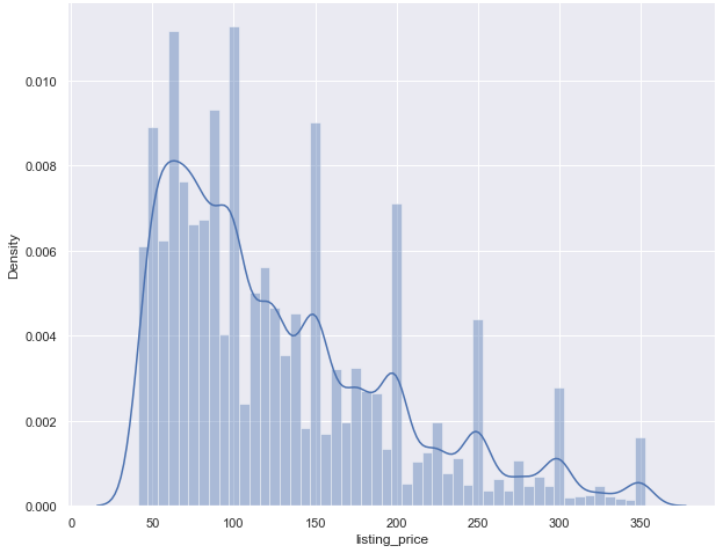

Figure 0 - Distribution of listing price following outlier removal

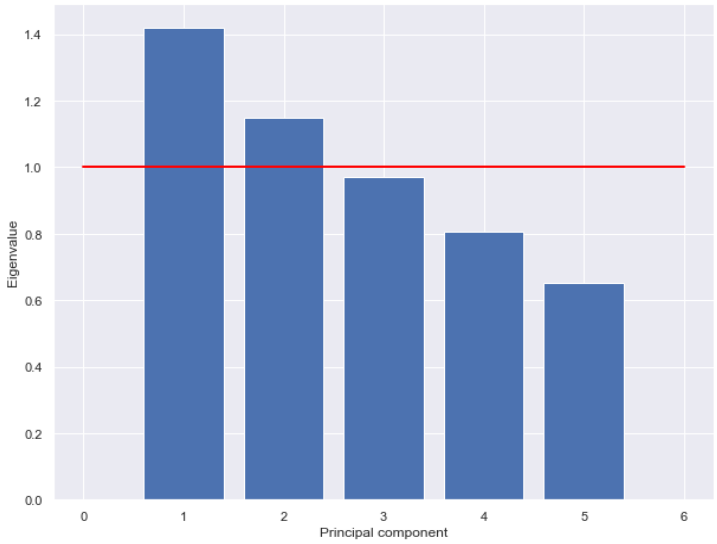

Figure 1 - Eigenvalues vs Principal Components along with Kaiser Line

	Component1	Component2
minimum_nights	0.500295	-0.197054
number_of_reviews	0.023653	0.877011
availability_365	0.665416	0.487773
calculated_host_listings_count	0.713207	-0.167110
listing_price	0.465562	-0.273963

Figure 2 - Loadings Matrix

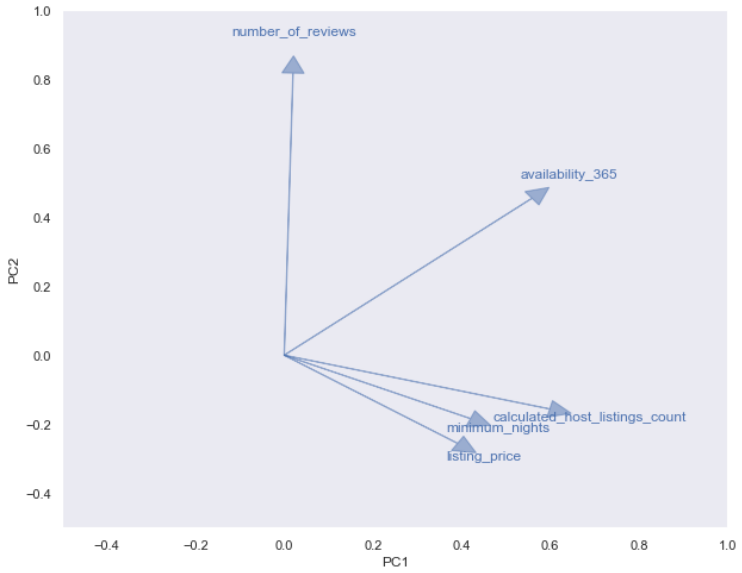

Figure 3 - Loadings plot for PCA components

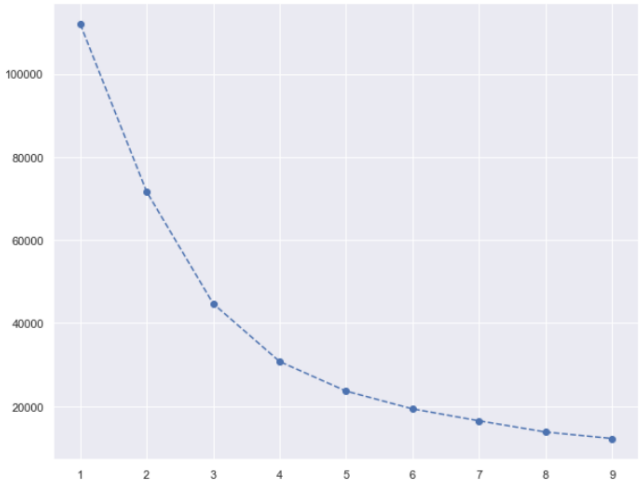

Figure 4 - Total within cluster distance vs k

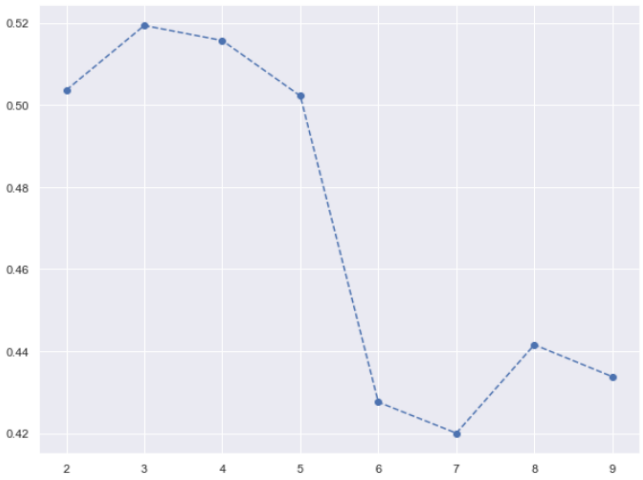

Figure 5 - Sum of silhouette scores vs k

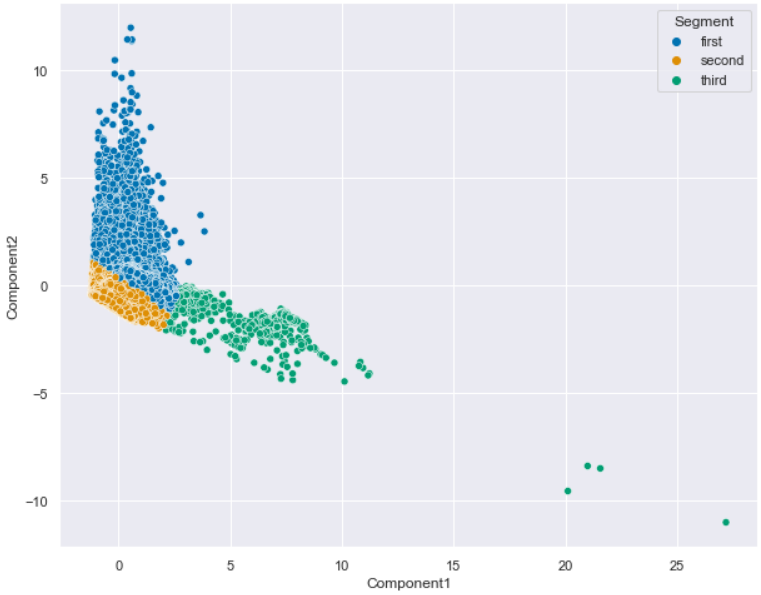

Figure 6 - Clusters obtained using k-means with k=3