Elevate your Airbnb: Secrets to becoming a top earner in Seattle

School of Mathematics, Computer
Science and Engineering
Department of Computer Science
City University of London

Abstract: In this paper, we combine and explore two open data datasets which contain the listings of Airbnb properties in Seattle and the reviews from users of the Airbnb platform. Our main goal is to use the total reviews variable as an indication of how popular neighbourhoods are but also do some research on the pricing strategies. We find that certain key factors and amenities influence the prices of Airbnb properties. Also, there seem to be some neighborhoods that are more popular than others and leave tourists with a higher level of satisfaction. Tourism in Seattle is at its peak during the Summer. Overall, the reviews suggest that tourists are satisfied with Seattle as a popular destination, but some key factors contribute to negative reviews.

I. Introduction

Airbnb is an online marketplace that allows people who want to rent out an apartment they own to connect with people who are looking for accommodations to stay. The company was founded in 2007 and today it has reached more than 7 million listings covering more than 200 countries worldwide. This has allowed hosts to list a variety of properties without any payment required [4]. On the other hand, customers can search the Airbnb listings not only by the price and location but also search properties by the type of property and some specific amenities that are included. At the same time, it allows customers to review the properties and give general feedback. This can help hosts improve their properties and even adjust their pricing and marketing strategies. But more importantly, this leads to higher customer satisfaction and revenue increase for the hosts.

Airbnb variables are categorized into two types. The host-controlled variables and the out-of-host-controlled variables. The main difference between the two types of variables is that host-controlled variables are variables that hosts can control and change like the prices of the accommodations they own and out of control host-controlled variables are variables that the hosts can't control, and the market determines their values like the number of reviews a property has [3]. In this paper, our goal is to identify hidden secrets that lead to high earnings in Airbnb by analyzing the features that hosts control such as pricing strategies and amenities of properties to optimize the performance of the out-of-host-controlled variables such as the increase in positive reviews and the overall improvement of their customer's experience. We have selected to analyze open-source Airbnb data in Seattle since it is considered one of the most popular tourist destinations in the US.

II. Data, Research Questions and Analytical Approach

A. Data

For this project, we gathered Seattle Airbnb open data from Kaggle. We used two of the three available datasets in the data source. The first dataset consists of the listings of available properties listed on the Airbnb platform. After removing all the irrelevant features that we won't be using for this analysis, we have a total of 19 features and 3818 different properties. The dataset contains different types of data like numerical, geographical and categorical data. The numerical variables are associated with the prices of accommodations, the average rating from reviews, the guests that the accommodation can host, and the number of different amenities that are included. The geographical data includes the zip code, the neighborhood, and the latitude and longitude of each listing. Finally, the categorical data includes features like the property type and the cancellation policy. Our aim for this project is to analyze the correlation between these variables mainly to identify what variables influence the popularity of properties in Seattle. This could be answered by using the number of reviews variable. We assume that properties that get a high number of reviews are used more often by tourists, which indicates that they are more preferred and more popular. We also aim to use geospatial data to create maps that help us identify similarities or dissimilarities between neighborhoods in Seattle.

The second dataset consists of the reviews that have been submitted to the platform between 2009-2016. This will help us analyze the sentiment of each review. The dataset also includes a date variable describing the review's submission date. This will be used in our analysis as an indication of when tourists tend to visit Seattle as we assume that someone usually reviews a property either during their trip or right after.

B. Analytical (Research) Questions

1. **What time of the year are the peak tourist seasons in Seattle as indicated by the timing and frequency of property reviews?**
   This question requires a time series analysis of the reviews. We aim to find out which time of the year tourists prefer to visit Seattle and identify potential seasonal trends.
2. **Which factors and amenities most influence the price of an Airbnb listing in Seattle?**
   This question requires a correlation analysis between certain amenities and property types and the prices. This will help us build a linear regression model to predict

the price of a listing. We aim to understand pricing trends and how the features contribute to higher or lower prices. This information can be especially useful for new Airbnb hosts to adjust their pricing strategies and maximize their profit.

3. **Which areas have the most expensive properties? How do the prices correlate with the number of reviews?**
   We aim to compare the spatial distribution between the prices and reviews to see how they correlate across the neighborhoods of Seattle.
4. **What is the defining relationship between pricing strategies and guest satisfaction across different neighbourhood clusters?** In this question, we group the neighbourhoods of Seattle based on their prices, total reviews, and average scores from reviews. We aim to see how the neighbourhoods are grouped and observe the similarities or dissimilarities of the clusters.
5. **Are there any particular areas in Seattle with a high number of positive or negative reviews?**
   This question requires a geospatial analysis of the distribution of positive and negative reviews across all of Seattle.
6. **Are there any certain amenities or features that consistently appear in positive or negative reviews?**
   To answer this question, we extracted a keyword for each review and created word clouds that show the most often-used keywords.

C. Analytical Approach (Analysis)

Since we are dealing with open-source Airbnb data, data preparation, data deviation and data cleaning techniques were applied to the datasets to reduce the number of features, remove null values and overall improve their quality. Data preparation was concluded with the application of data derivation techniques to create and transform several attributes. Finally, after the listings and reviews were merged, the final dataset included a total of 25 features and 83722 observations. The reason we combined the 2 datasets is to have all of the reviews sorted by each property.
For each research question listed above, we devised the following plan to lead us to our findings:

1. In this research question, the date feature was used from the reviews as an indication of the datetime a tourist visits Seattle. Data preparation included transforming the date variable to a datetime object to extract the months and years separately. Three separate graphs were created that show the monthly and yearly frequency of reviews between 2009-2016. The goal of this analysis is to identify potential seasonal trends in tourism that indicate the season that tourists prefer to visit Seattle.

2. To answer this research question, a price prediction model was created using multiple linear regression. Before building the mode, a correlation analysis was applied between the host-controlled variables such as bedrooms, bathrooms, amenities, guests included, accommodates, minimum and maximum nights as well as property type. The same features were chosen in a previous study by [3]. The reason that only these variables were chosen is to understand what features influence higher prices. This will help hosts to prioritize certain amenities and adjust their overall pricing strategies.
Extra data preparation and preprocessing techniques to derive new data were included in this section, such as grouping the property types and one-hot encoding to convert categorical data to numerical. Since the property types are not considered ordinal data, the binary variables were transformed to dummy variables as explained by [5]. Finally, we performed a 70:30 test train split to our dataset.
The evaluations of the multiple linear regression model were subject to the OLS method. The evaluation metrics that were included are the (1) mean absolute error which is known as the average of the absolute difference between the actual values and the predicted values, (2) the R-square which describes the proportion of variance that can be explained by the independent variables and (3) the adjusted R-square which a metric that is adjusted for the total predictors for the model [1].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \quad (1)$$

MAE- Mean Absolute Error

n –Number of predictions

Y_i– Observed values

$\hat{Y}_i$– Predicted values

$$R^2 = 1 - \frac{SSR}{TSS} \quad (2)$$

SSR–Sum of Squares of residuals

TSS – Total Sum of Squares

$$R^2{}_{adj} = 1 - \frac{(1 - R^2)(n-1)}{n-p-1} \quad (3)$$

R^2 – R-Squared

p – Number of predictors

N – Total number of samples

3. This research question mainly focuses on data preparation and derivation techniques used to create 2 maps using the folium library to show and compare the spatial distribution of the total reviews listings get and their prices. The data was grouped by the latitude and longitude variables to visualize the average price and number of reviews by longitude and latitude group. We want to see if there are any areas with high demand and analyse whether

these areas also have properties with high prices. This is an indication to the willingness of visitors to pay more for properties in certain areas.

4. To answer this question, the dataset was transformed and grouped to show the summary statistics for each neighbourhood. In particular, the dataset was prepared and transformed to show the average number of reviews, prices and scores from reviews. We then used the K Means clustering algorithm to group the neighbourhoods based on those statistics. After trying with a different number of clusters, the neighbourhoods were finally clustered to 4 groups. Since our dataset does not contain a variable with the true clustering labels, we used the elbow method to evaluate the model (the optimal number of clusters). The elbow method is a graphical representation of the sum of the square distance between the points of a cluster and the cluster centroid. This helps us understand how similar or dissimilar the clusters are [6]. The elbow method is an iterative process whereas we keep increasing the number of clusters, we calculate the distance of the points from the nearest centroid and assign the points to the cluster of the nearest centroid. We keep repeating this process until the position of the centroid does not change. This is called the elbow point where the graph starts to flatten out after several clusters generated. That is the optimized number of clusters which will give us very dissimilar clusters [2].
To evaluate the differences between the clusters, we have also created a Multi-Dimensional Scaling visualization scatterplot which is a 2-dimension graphical representation of the data and can help us identify neighbourhoods that are similar and dissimilar.

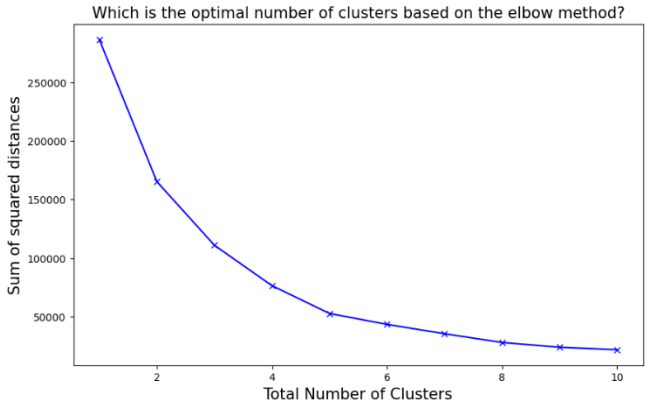


Figure 1. A graphical representation of the elbow method applied to the neighborhoods

Based on the graph above, the optimal number of clusters seems to be 4 because that is when the line starts to flatten out.

5. For this research question, one new column was created with the use of sentiment analysis. Sentiment analysis is considered a natural language processing technique that predicts the reviews to be either positive, neutral, or negative based on the polarity score of each sentiment. Then we performed a geospatial analysis of the reviews to visualize the spatial distribution of positive and negative reviews. This approach can be useful to identify areas with a higher concentration of positive or negative reviews.

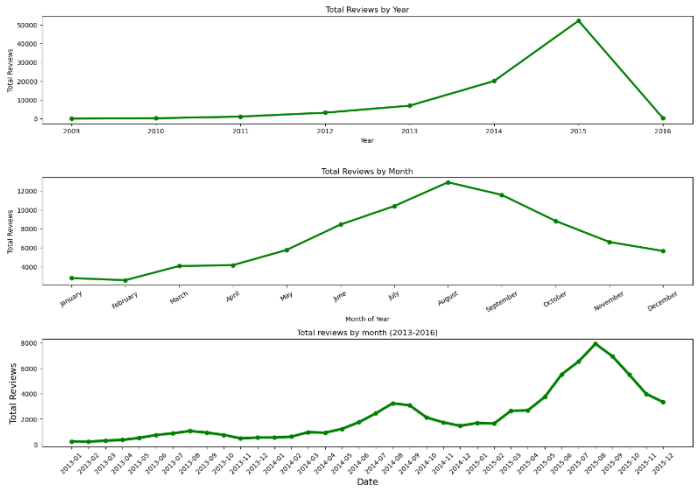
6. To answer this research question, a new variable was derived from the reviews with the use of natural language processing techniques. This variable consists of the main. The Rake algorithm was used from the nltk library in Python to extract the main keyword from the review that shows sentiment. We then constructed two separate word clouds for positive and negative reviews to see which keywords are used more often in positive and negative reviews. This can give us some valuable information about the property types, amenities, and overall characteristics that customers value and don't value.

III. FINDINGS, REFLECTIONS AND FUTURE WORK

➢ Findings

A. Seasonal trends in Seattle

To identify the seasonal trends, we produced 3 graphical representations of the distribution of reviews by month, and year.



Figures 2,3 and 4. Total number of reviews aggregated by month, year and total number of reviews by month between 2013-2016

The number of reviews tend to increase from January with a rapid rise during the summer months and a slow decrease during the Autumn and winter Months. This seems to the be the case between the years 2013-2016 with 2015 being the peak year. This indicates that Seattle is a popular destination during the Summer but not so much during the Winter.

B. *Price Prediction Model*

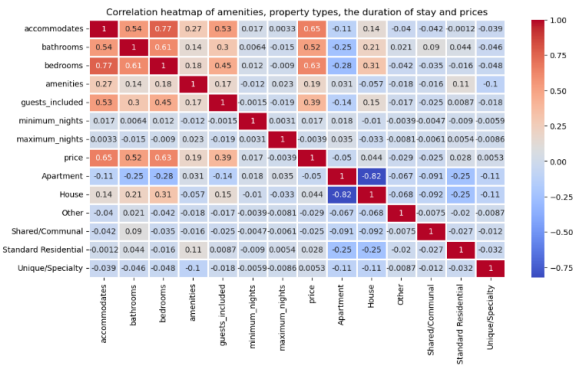


Figure 5. Correlation heatmap of airbnb prices with the amenities, property types and duration of stay

From the above figure, we summarize that the main factors that seem to influence the prices of Airbnb listings are the amenities that correlate with the size of accommodations like total bathrooms, bedrooms and accommodates. The length of stay doesn't have much of an impact on price.

Figure 6. Comparison of model predictions and actual prices

In the above figure we can see a graphical representation of how close our price predictions were to the actual prices of the test set. Although the evaluation metrics suggest that our model overall performed well, we can see that it struggles to identify very expensive properties. The test of normality indicates that our model might be influenced by outliers. In future work this would be addressed, but for the purpose of this project, we wanted to keep the outliers to see how they influence our model.

C. *Geospatial distribution of prices and reviews*

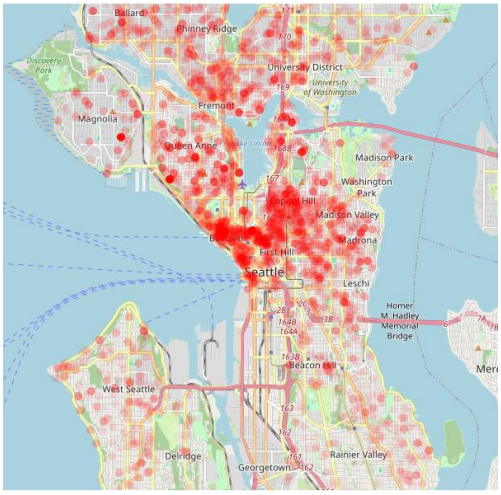


Figure 7. Spatial Distribution of Prices

It seems that the most expensive properties can be mainly found in the center of the city and in certain neighborhoods across the north part.

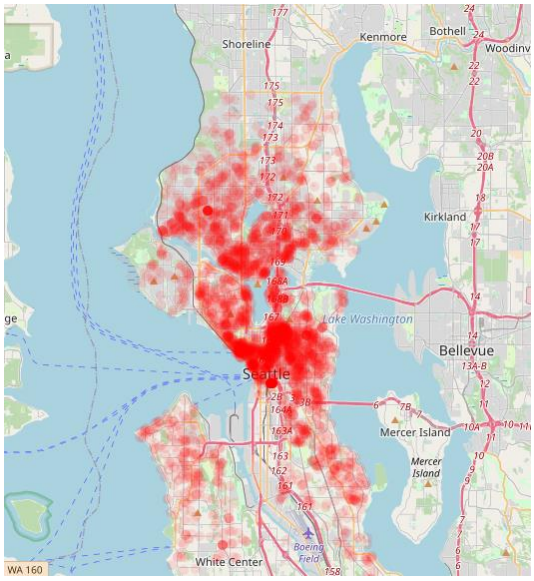


Figure 8. Spatial Distribution of Total Reviews

Similarly, to the price distribution, it seems that the highest reviewed properties are mainly in the center of the city and in the north areas. Perhaps this could be due to the proximity to attractions and recreation areas. Unfortunately, we do not have geospatial data regarding the most popular tourists attractions to make the comparisons. This map though are a good indication of which areas attract more tourists and can be useful information for Airbnb hosts.

D. Cluster analysis of neighbourhoods by the total reviews, prices and review scores

In the table below, we can see the summary statistics of each cluster.

Cluster	Average Price	Average Number of Reviews	Average Score from Reviews	Total Neighbourhoods
1	94.08	37.95	94.36	36
2	96.21	203.02	95.72	6
3	164.03	41.11	94.27	13
4	101.25	84.69	95.17	32

Figure 9. Summary Statistics of Neighborhood Clusters

Neighbourhoods in cluster 2 have the most visited properties with the highest scores from reviews. This indicates that these neighbourhoods are more attractive to tourists and tend to have the best reviews. What is interesting though is that these neighbourhoods also have moderate prices in comparison to other clusters. Cluster 3 represents the most expensive neighbourhoods with low scores and slightly lower satisfaction from tourists. These neighbourhoods seem to be performing the worst among all. Cluster 1 represents most neighbourhoods in Seattle with the least expensive properties that tend to be visited rarely and have lower scores from reviews. Finally, cluster 4 neighbourhoods are similar to cluster 1, but are a bit more expensive and have a lot more reviews and higher scores.

From the analysis of the clusters, it seems that neighbourhoods with properties that have cheaper or moderate prices tend to be more favourable for tourists and leave them with higher satisfaction.

Below, we can see a graphical representation of the similarities between the clusters and neighbourhoods.

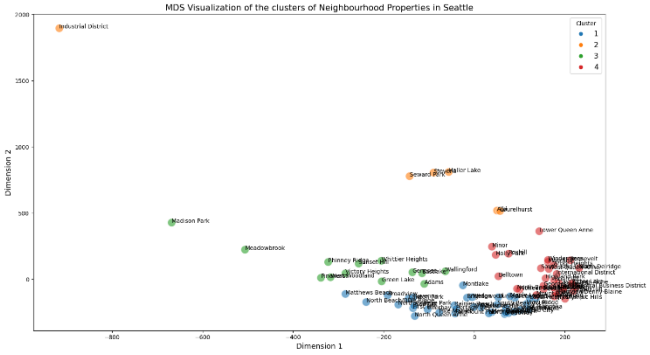


Figure 10. MDS Visualization of Seattle Neighbourhoods

Based on the above graph, it seems that the clusters are dissimilar between them. Based on the summary statistics of each cluster, Industrial District, Seward Park, Haller Lake, Stevens, Alki and Laurelhurst are the most popular tourist areas.

E. Geospatial analysis of positive and negative reviews

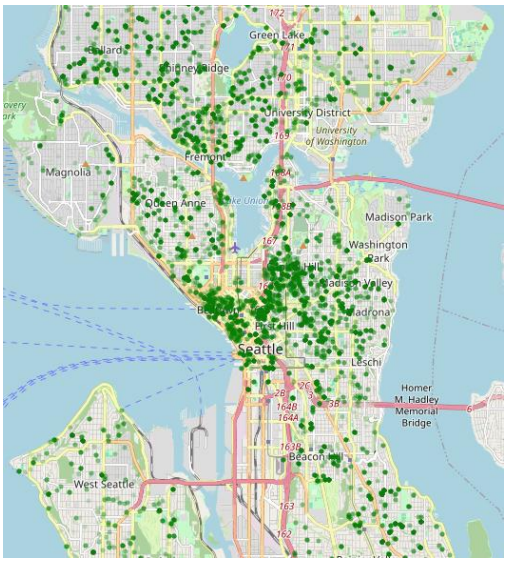


Figure 11. Spatial Distribution of Positive Reviews

Based on the map above, it seems that most positive reviews are concentrated in the center of the city and in some areas of the north part.

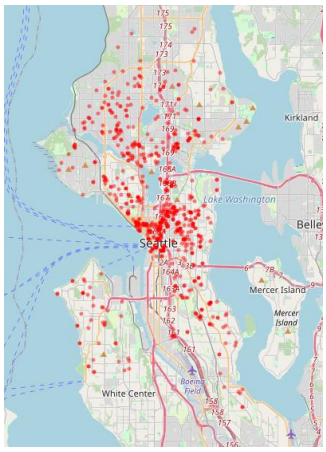


Figure 12. Spatial Distribution of Negative Reviews

There seem to be a lot fewer negative reviews. Similar to the positive reviews, most negative reviews can be found in the center of the city and in some areas in the north part of the city.

F. Analysis of keywords from positive and negative reviews

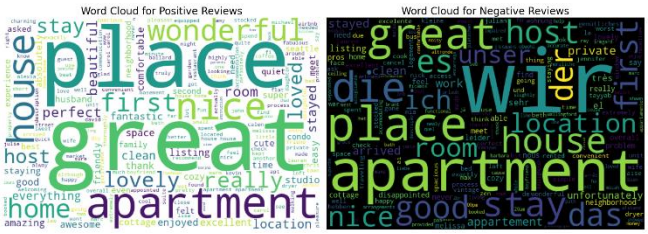


Figure 13. Word clouds for positive and negative reviews

Figure 13 shows us that the keywords like 'lovely', 'apartment', 'great' and 'place' in positive reviews show that guests feel comfortable during their stays. Words like 'clean' and 'convenient' shown in both positive and negative reviews indicate that the visitors value the physical attributes of the properties. The word 'location' seems to be frequently used in both positive and negative reviews. This indicates that the location of the properties is a key factor that correlates with a high level of satisfaction.

- **Reflections and Future Work**

The purpose of this project was to uncover trends in the data that could help hosts improve the quality of their properties and improve their marking and pricing strategies. It seems that moderate or cheaper prices tend to be more favourable with tourists. Hosts should adjust their pricing strategies based on the seasonality that was uncovered. During the Summer where there is higher demand, Airbnb hosts can slightly increase the prices but during the Winter they should reduce them to try to attract visitors who might be looking for deals. Hosts should look to rent apartments in the centre of the city rather than the suburbs but especially in certain neighborhoods that seem to be way more attractive than others.

In future work, the analysis of a combination of various airbnb datasets for different cities would give us the opportunity to compare Seattle with other popular tourist destinations. This will help us understand how Seattle could improve overall as a tourist destination.

Section	Word Count
Abstract	127
Introduction	291
Data	296
Analytical (Research) Questions	196
Analysis (Analytical Approach)	970
Findings, Reflections and Further Work	780
Total	**2660**

REFERENCES

[1] A. V. Tatachar, 'Comparative Assessment of Regression Models Based On Model Evaluation Metrics', vol. 08, no. 09, 2021.

[2] R. Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan, 'Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method', *J. Phys.: Conf. Ser.*, vol. 1361, no. 1, p. 012015, Nov. 2019, doi: 10.1088/1742-6596/1361/1/012015.

[3] B. McNeil, 'Price Prediction in the Sharing Economy: A Case Study with Airbnb data'.

[4] J. Folger, 'How Airbnb Works-for Hosts,Guests, and the Company Itself'. Accessed: Dec 23, 2023. [online]. Available at: https://www.investopedia.com/articles/personal-finance/032814/pros-and-cons-using-airbnb.asp#:~:text=Airbnb's%20Impact&text=Airbnb%20has%20made%20travel%20more,available%20space%20to%20these%20travelers

[5] J. Brownlee, 'Why One-Hot Encode Data in Machine Learning'. Accessed: Dec 23, 2023. [online]. Available at: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-

[6] B. Saji, 'Elbow Method for Finding the Optimal Number of Clusters in K-Means'. Accessed: Dec 23, 2023. [online]. Available at: https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/?fbclid=IwAR3ROkQAu0xWLJ95MDeyI9WRq-myoqQ_b2WIWwMMF9FFtpt5ekZCzi7_lPw

DATASETS

Seattle Airbnb Open Data -

https://www.kaggle.com/datasets/airbnb/seattle