Data Analysis of Airbnb Listing Prices and Venues in Seattle

Gulden Turgay
April 27, 2020

1. Introduction

1.1. Background

Airbnb is an online marketplace where users can arrange or offer accommodation or travel experiences since 2008. The properties in listings do not belong to the company. Instead it acts as a broker and receives commission for each booking. The price range of homestays depends on a range of variable such as the location, number of bedrooms and bathrooms, nearby transportation options and things to do around etc.

Seattle is one of the fastest-growing cities in United States and the largest one in the state of Washington with an estimated population of 744,955 as of 2018. Since the city is the heart of many major companies in U.S. and also has many attractions it has millions of visitors each year. It was reported that 38.1 million people visited the city of Seattle and the King County, which is the county Seattle is located in, in 2015 and 40 million visitors in 2018 with a groundbreaking record.

1.2. Problem and Interest

The goal of this project is to visualize neighborhoods of Seattle on map regarding the average nightly prices of Airbnb places and cluster the neighborhoods in order to see if there is any relation between the Airbnb prices and venues in the neighborhood, if so what kind of venues affect the most when it comes to pricing an Airbnb accommodation.

Anyone to rent their property as an Airbnb listing in Seattle would be interested in the problem in order to set an accurate price.

2. Data

The Airbnb data was sourced from Inside Airbnb website and the date that data was compiled on the website is 03/17/20. Foursquare API was used to get the top venues of given neighborhood in city of Seattle. Seattle neighborhood boundaries json data was utilized to get the neighborhood boundaries.

3. Exploratory Analysis

This project mainly consists of three steps which are analyzing Airbnb data, exploring Seattle neighborhoods and clustering Seattle neighborhoods.

3.1. Analyzing Airbnb data

In the first step, I have obtained the Airbnb data in order to get the average nightly Airbnb accommodation prices in each neighborhood in Seattle. For this purpose, I've used only the necessary attributes which are *neighbourhood, latitude, longitude* and *price*. After this step, I've checked if there is any null value to drop. Since there was not any null value, I've continued with renaming the columns as *Neighborhood, Latitude, Longitude* and *Price*. And then I've taken the mean of the price column to calculate the average price for each neighborhood and grouped them by neighborhood names (Figure 1).

	Neighborhood	Latitude	Longitude	Price
0	Adams	47.671661	-122.385505	150.292035
1	Alki	47.575465	-122.407382	151.793478
2	Arbor Heights	47.510568	-122.380087	117.450000
3	Atlantic	47.595194	-122.304142	217.438095
4	Belltown	47.615327	-122.345001	181.958025

Figure 1. Neighborhoods with latitude/longitude and average nightly prices in USD.

After obtaining desired information from the dataset, I've used the JSON file to get the neighborhood boundaries and modified it for only the city of Seattle and I've created a choropleth folium map to visualize the neighborhoods based on the average prices (Figure 2).

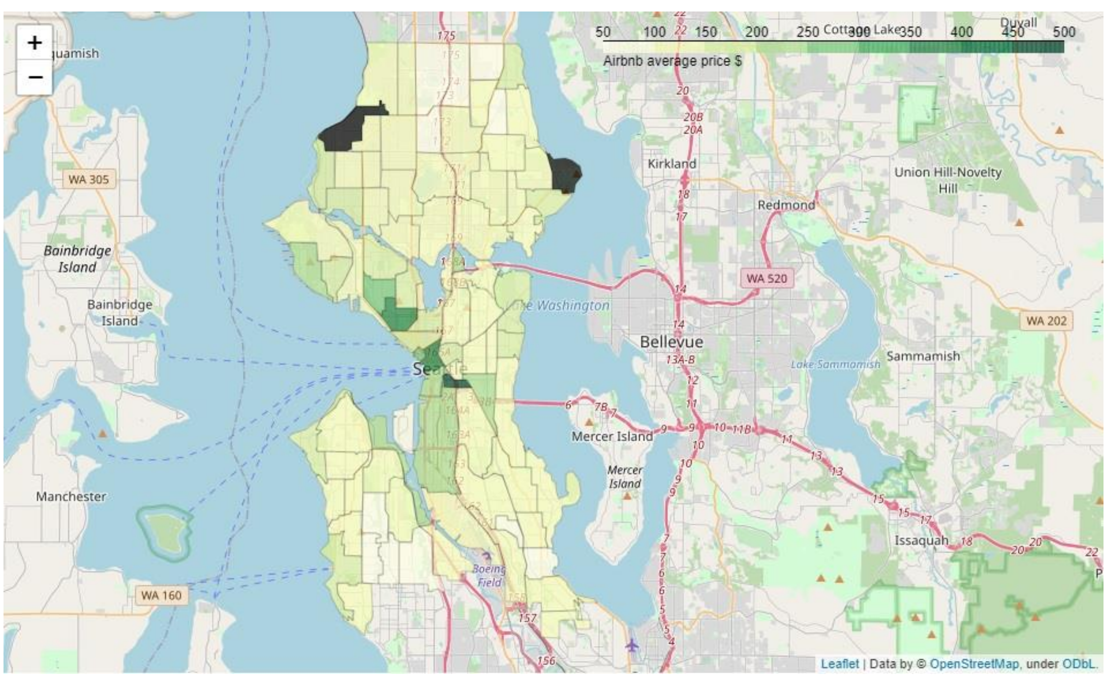

Figure 2. Screenshot of the choropleth map based on average price.

3.2. Exploring Seattle neighborhoods

In the second step, I've created a map for Seattle neighborhoods just to see which neighborhood is where. And then by using Foursquare API, I've had top 100 venues in each neighborhood within 500 meters along with their venue category and venue latitude/longitude. Then I've counted the venues in each neighborhood within the limit of 100. In some neighborhoods it already hit the limit. And additional calculation showed us there are 282 unique venue categories in Seattle within the limits I'd set earlier.

Following this, I've done one hot encoding and calculate the frequency of each venue category in each neighborhood. Using this information, I've created a new dataframe showing 10 most common venue type in each neighborhood (Figure 3).

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adams	Ice Cream Shop	Burger Joint	Thai Restaurant	Bakery	Coffee Shop	Sushi Restaurant	Mexican Restaurant	Dive Bar	Dessert Shop	Italian Restaurant
1	Alki	Coffee Shop	Park	Italian Restaurant	Ice Cream Shop	Trail	Sports Bar	Fish & Chips Shop	French Restaurant	Event Space	Beach Bar
2	Arbor Heights	Spa	Yoga Studio	Fish Market	Event Space	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Field	Fish & Chips Shop	Flea Market
3	Atlantic	Mediterranean Restaurant	Park	Seafood Restaurant	Sandwich Place	Bank	Bakery	Dry Cleaner	Residential Building (Apartment / Condo)	Trail	Skate Park
4	Belltown	Bar	Pizza Place	Breakfast Spot	Bakery	Italian Restaurant	Coffee Shop	Sushi Restaurant	Gym	Dive Bar	Marijuana Dispensary

Figure 3. Screenshot of the dataframe showing top 10 most common venue category.

3.3. Clustering the neighborhoods

In the third and last step, I've used K-means clustering algorithm to cluster the neighborhoods based on venue types and discover the certain similarities. To find the optimum k number of clusters I've applied elbow method first but there was no elbow occurred where the rest of the line flattened out (Figure 4).

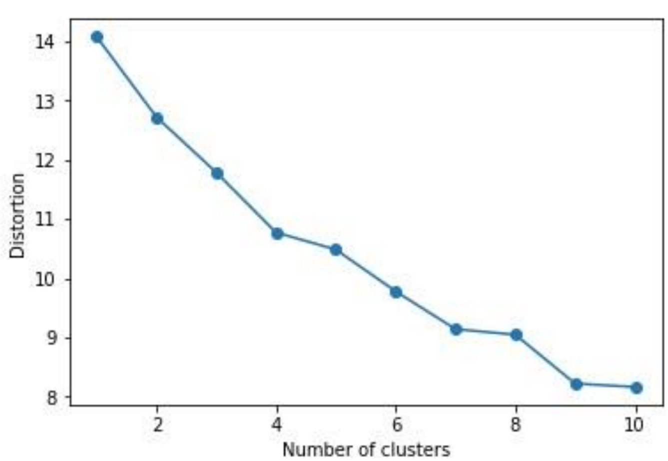

Figure 4. Elbow method line graph.

Since there was not any solid result from that, I've applied Silhouette Score method next. In this method, optimal k is where the silhouette score is closest to 1. When calculated it gave us k=3.

Based on the result of 3 clusters, I made bar charts to see what the similarities are in each cluster and labeled them accordingly.

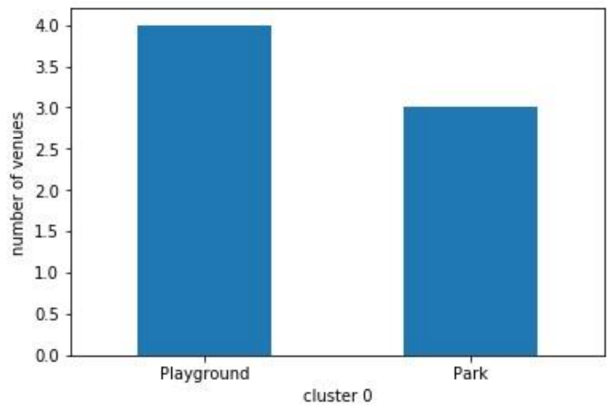

Figure 5. Bar chart of the first cluster.

In the first cluster (Figure 5), which has a cluster label of '0', there are 4 playgrounds and 3 parks so I label this cluster as 'Outdoor Activities'.

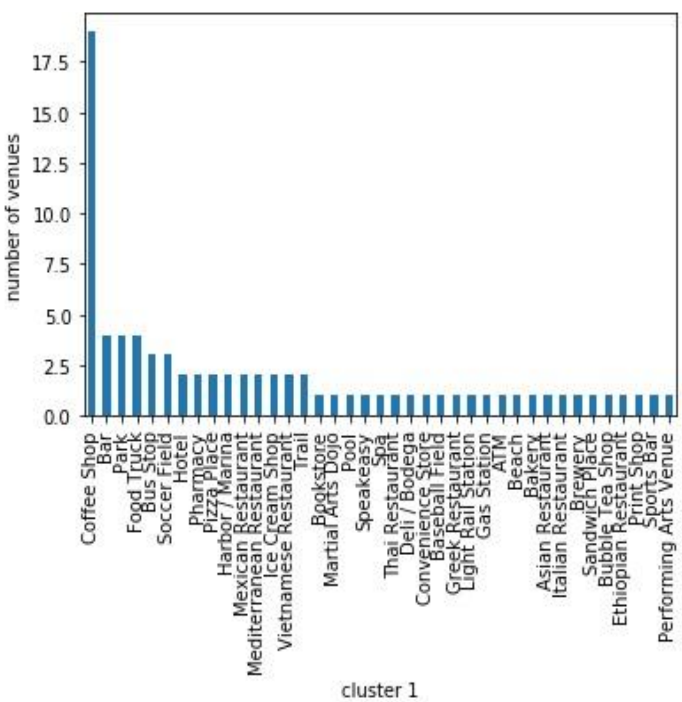

Figure 6. Bar chart of the second cluster.

In the second cluster (Figure 6), which has a cluster label of '1', there are 19 coffee shops by far more than any other type of venue. 4 bars, 4 parks and 4 food trucks follow coffee shops. It's appropriate to label this cluster as 'Coffee Shops'.

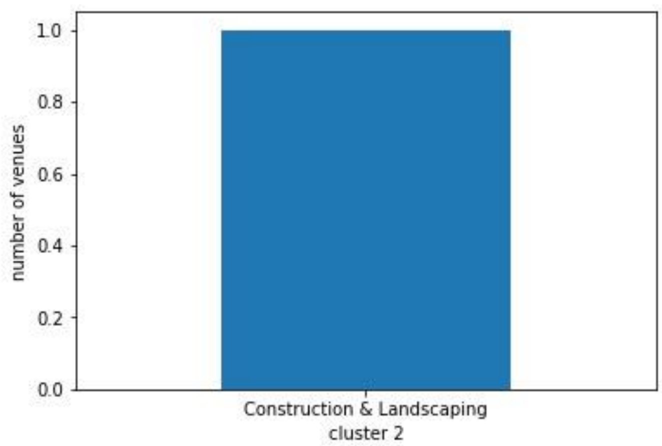

Figure 7. Bar chart of the third cluster.

In the last cluster (Figure 7), which has a cluster label of '2', there is only 1 venue which is a Construction & Landscaping type. So, I label this cluster as 'Landscaping'.

When finalizing the project, I've created a new attribute for top 3 venue category with their counts for each neighborhood and merged the attribute to the dataframe, where 10 most common venue types displayed for each neighborhood, along with Labels attribute which has the cluster labels and Price attribute which has average Airbnb price for each neighborhood.

And then I've added cluster markers containing neighborhood names, top 3 venues and average nightly Airbnb prices in US dollars to the folium map (Figure 8) that I'd created in the first step. This gives us a chance to see if there's any relation between venues and Airbnb prices in neighborhoods.

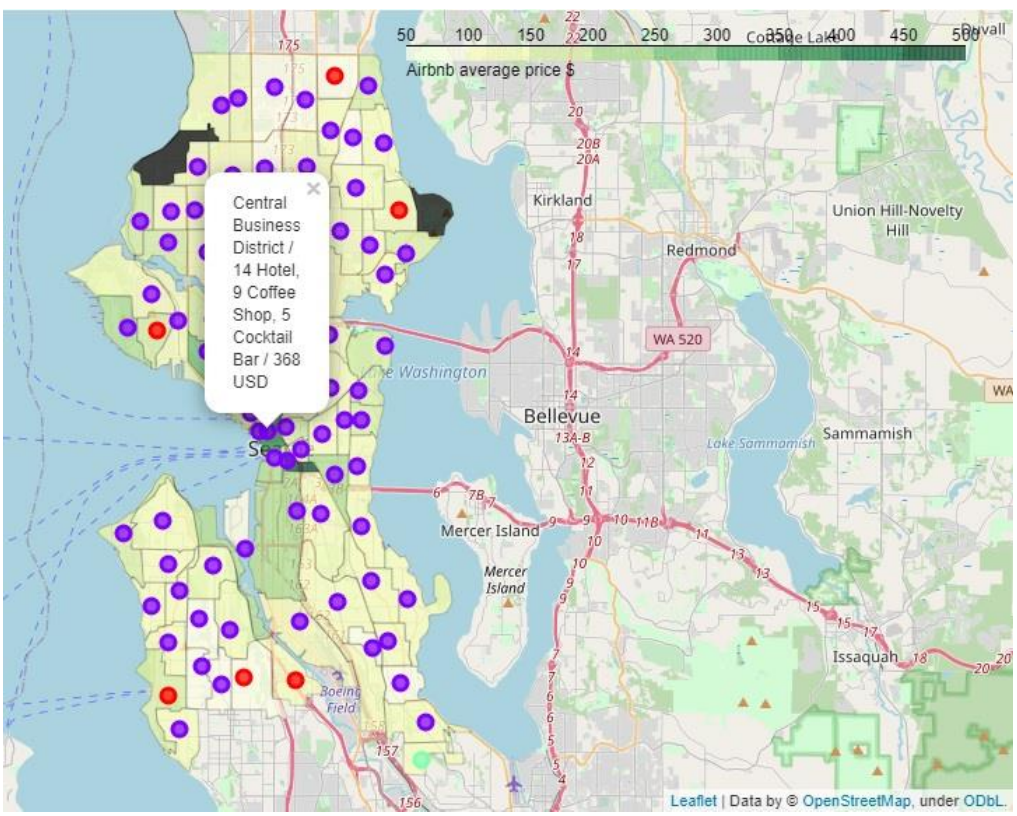

Figure 8. Screenshot of the interactive folium map.

4. Results and Discussion

The analysis shows that the Airbnb accommodations with higher ranges (between 250 USD-500 USD) of nightly average prices are mostly located in central neighborhoods than the rest of the city of Seattle. It seems like the average price range is the same in big part of the city which is between 100 and 150 USD while it decreases in northern and southern outskirts of the city.

The purple colored cluster shows us there are mostly coffee shops all around the city and considering the fact that Seattle is known for its outstanding coffee culture and many of the city's residents are coffee admirers it is not surprising. When we turn our focus to red colored markers, since they are labeled as 'Outdoor Activities' they might be residential areas. And in the last cluster there is only one type of venue which is Construction & Landscaping. The neighborhood that belongs to this cluster is located in the southernmost of the city.

When we take a look at darker green colored areas there are hotels, restaurants, theaters and opera centers along with coffee shops. Especially the number of hotels can give us an idea that these neighborhoods mostly attract people for touristic or business purposes and it makes accommodation prices higher in these neighborhoods. But in general, it doesn't seem like the venue types in neighborhoods affect the Airbnb prices except for central Seattle.

5. Conclusion

The result shows us that the prices mostly depend on the location rather than venues. On the other hand, it indicates that the neighborhoods with the greatest number of hotels and their surrounding neighborhoods would be high price range areas to rent a property as an Airbnb accommodation. Since the location and surrounding venues are not the only criteria for pricing, this project can be improved with further analysis.