Analysis of BC's Lower Mainland Venue and Housing Price Data

Krishant Akella

May 31, 2021

Contents

1 Introduction

According to a recent study completed by Point 2 Homes, British Columbia (BC) contains some of the most expensive housing markets not only in Canada but all of North America. The buying craze may stem from a high demand along with a scarce supply, low interest rates and a financially strong middle class. Moreover, Vancouver is a city in BC that is surrounded by water which limits supply, consequently it continually ranks among top three most expensive cities to live in worldwide. Of course there are many possible explanations on why the lower mainland of BC is expensive and I intend on sharing some of these possible avenues in this report.

1.1 Business Problem and Interest

The objective of this project is to understand possible connections between common venues and housing prices of areas in the lower mainland of BC. This problem may interest investors looking to start new businesses where real estate costs are low but demand for a certain type of venue is high. Additionally, home buyers may want to know what cities have the least expensive housing prices, the different social locations, and options for transit. Likewise, real estate agents may use this information to help someone buy a house.

2 Data

2.1 Data Sources

The average housing price per area, area, and neighborhood data of BC was scraped from MoneySense which included 381 neighborhoods and 22 areas. An area is a city in the lower mainland of BC and each area contains a certain number of neighborhoods. Each neighborhood has an average housing price, however, the housing price is only related to the average of the whole area in which the neighborhood is located. To obtain the data I used the BeautifulSoup library and created a soup object to easily parse through the table and retrieve the relevant data.

Further, the area and neighborhood data were necessary to find the latitude and longitude of each neighborhood. I accomplished this by utilizing the GeoPy library and the Nominatim method. GeoPy is not a service, however, it is a library which implements certain API's. For example, the Nominatim method calls OpenStreetMap, it is free but has a low request limit. The average area price, area, neighborhood, latitude and longitude data was transformed into a pandas data frame. The neighborhoods of interest are apart of the lower mainland, the data set from MoneySense includes some neighborhoods outside of this region. The following list showcases the areas of interest in this report:

- West, North and East Vancouver
- Burnaby
- Richmond
- Coquitlam and Port Coquitlam
- New Westminster
- Maple Ridge
- Surrey
- Delta
- Abbotsford
- Langley City

A geoJSON file contains geographic features and non-spatial attributes, in this project, it is required to create a choropleth map in folium to visualize the average housing prices of different areas. I requested a geoJSON file of legally defined administrative areas of BC from DataBC catalogue and was able to attain the data. Moreover, only coordinates from the geoJSON file corresponding to the areas in my data frame are included.

Lastly, by utilizing the Foursquare API I was able to gather the coordinates of venues and types of venues. I was able to extract 3454 venues with 298 unique categories. In the API calls I requested up to 100 venues within a 500 meter radius of each neighborhood coordinate. I attempted to increase the radius parameter to retrieve more venue data, however, this resulted in poor fits when applied to the machine learning algorithm that I discuss in the Methodolgy section 3.

2.2 Data Cleaning

To remove neighborhoods that were contained in areas outside of the lower mainland of BC I set limits to how far the latitude and longitude coordinates can be located. I used the city of New Westminster as the center coordinate with latitude and longitude values 49.1737, -122.7604 respectively. By trial and error I restricted latitude and longitude values with the following constraints:

$$|latitude - 49.1737| > 0.3$$
$$|longitude - (-122.7604)| > 0.55$$

To correctly parse the geoJSON file to create the choropleth map I needed to ensure the areas in my data frame correctly matched the keys in the geoJSON file. I manually changed the strings of Vancouver to East Vancouver, North Vancouver - City to North Vancouver and Langley - City to Langley in the geoJSON file. This ensured that all the areas in the lower mainland were matched correctly with the areas in the data frame.

3 Methodology

In this section I discuss all the exploratory data analysis that was applied to visualize and segment the data. Furthermore, I discuss the K-means machine learning algorithm and the suitability of this algorithm to this problem. Finally, I provide information about the inferential statistical testing used to select the number of clusters for the machine learning algorithm.

3.1 Exploratory Data Analysis

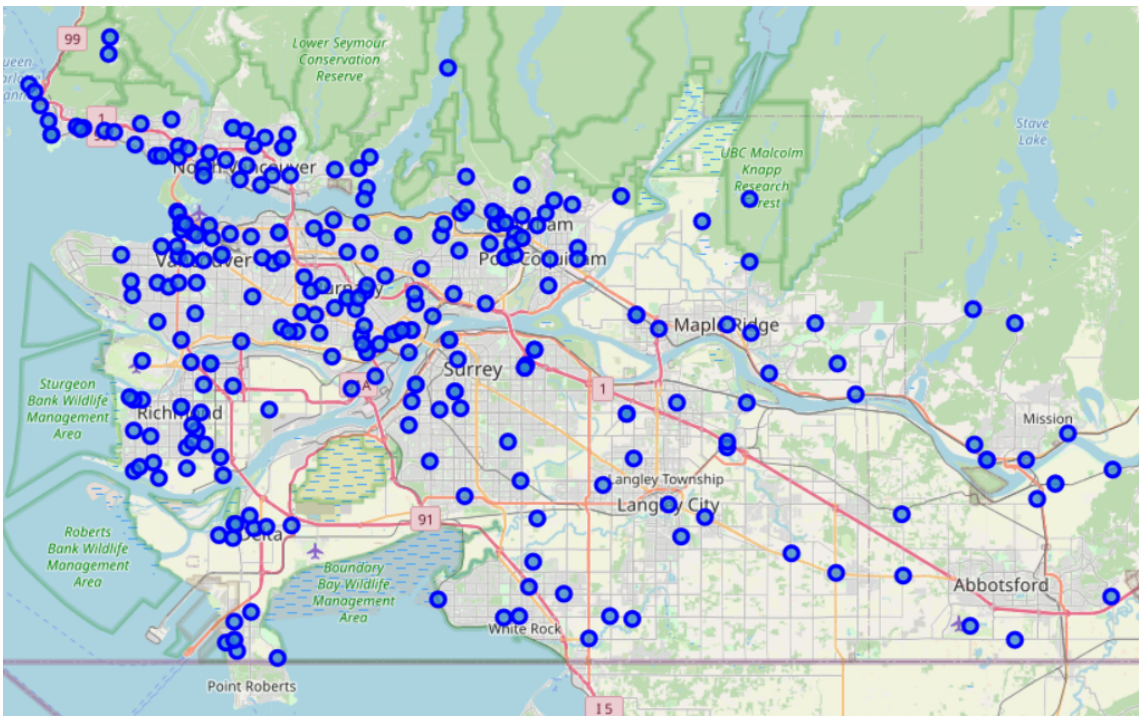

Figure 1: Location of neighborhoods in the lower mainland of BC.

To begin exploring the spatial data I visualized the locations of neighborhoods within each municipality of BC by creating a folium map utilizing the latitude and longitude data, see Figure 1. All municipalities contain more than one neighborhood, however, some have higher density of boroughs. For example, Vancouver and Burnaby have many boroughs clustered together, whereas Maple Ridge and Langley City have fewer boroughs clustered. Clearly, larger cities in terms of squared kilometers hold less clusters of neighborhoods.

To explore the average housing prices of each area I first plotted a histogram of the prices. I chose 3 bins each with its own color because it is clear to distinguish the different types of price ranges. As one can observe in Figure 2 there are three distinct price ranges: [670000, 1500000) - Low, [1500000, 2315000) - Medium, [2315000, 3150000) - High. Naturally, there are fewer houses with expensive average housing prices than affordable homes.

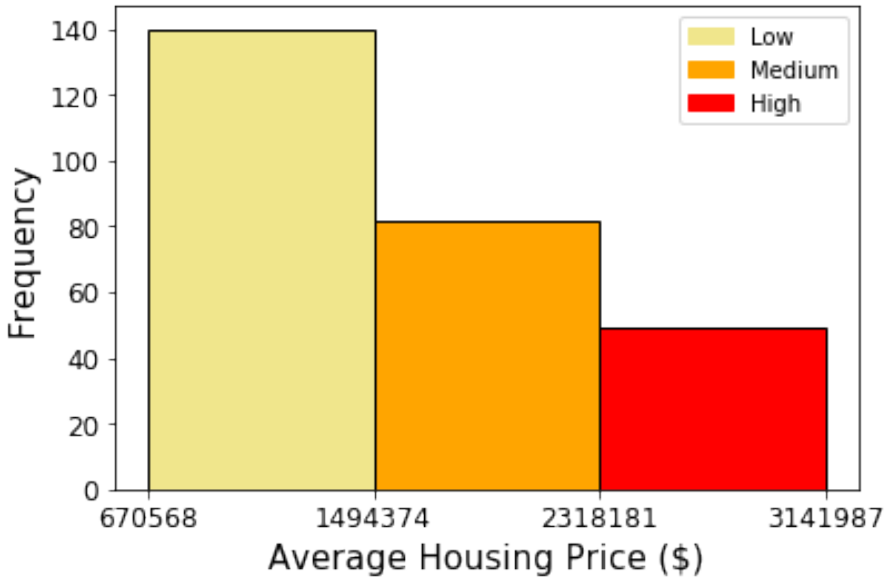

Figure 2: Histogram of average housing prices per area.

Furthermore, to visualize the average housing prices for each area I created a choropleth map in folium using the three price ranges. There is one city in the high average price range which is West Vancouver. The majority of the densely populated areas are categorized in the medium price range (North Vancouver, Burnaby, Richmond). The rest of the areas are classified as low.

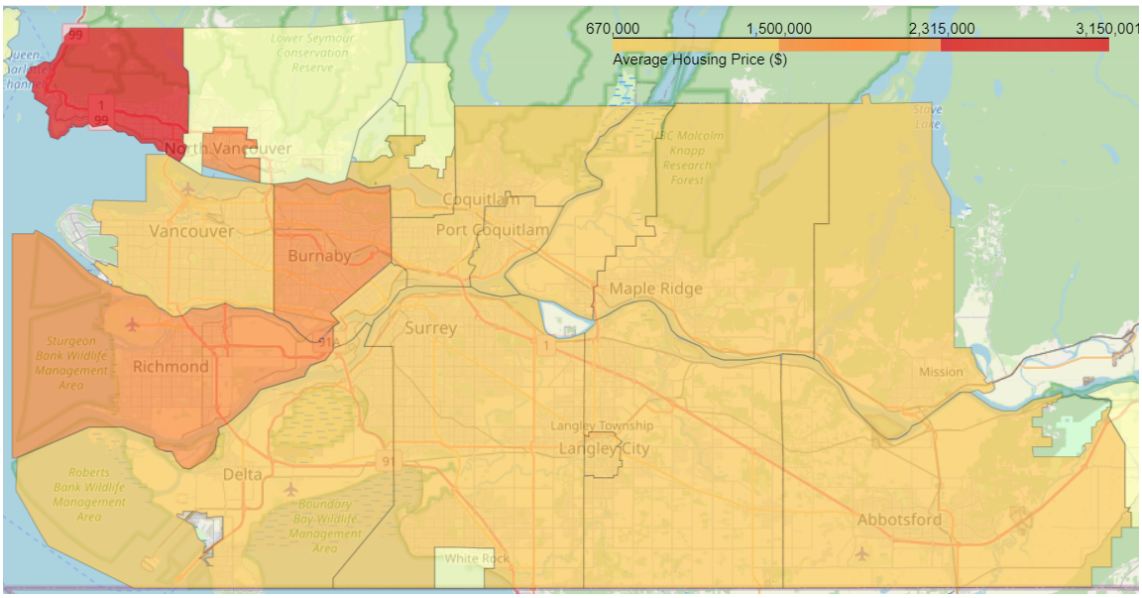

Figure 3: Choropleth map of the average housing price in each area of the lower mainland of BC.

At this point, the main goal is to segment the venue data around the neighborhoods in each area, find clusters and superimpose the clusters onto the choropleth map. I applied the Foursquare API to retrieve venue data for each neighborhood. Furthermore, the venue data is non-specific other than being common venues within a neighborhood. I then segmented the data to find the top 10 common venues around each neighborhood. The data frame below shows the first 5 neighborhoods with their top 10 common venues.

	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
1	Abbotsford West	Platform	Women's Store	Food & Drink Shop	Filipino Restaurant	Fish & Chips Shop	Fish Market	Flea Market	Flower Shop	Food	Food Court
2	Aberdeen	ATM	Flower Shop	Women's Store	Food Court	Filipino Restaurant	Fish & Chips Shop	Fish Market	Flea Market	Food	Food & Drink Shop
3	Albion	Grocery Store	Coffee Shop	Food & Drink Shop	Filipino Restaurant	Fish & Chips Shop	Fish Market	Flea Market	Flower Shop	Food	Women's Store
4	Aldergrove Langley	Fast Food Restaurant	Liquor Store	Coffee Shop	Pizza Place	Inn	Burger Joint	Business Service	Juice Bar	Pharmacy	Sandwich Place
5	Ambleside	Coffee Shop	Sushi Restaurant	Café	Italian Restaurant	Park	Asian Restaurant	Restaurant	Bank	Sandwich Place	Fruit & Vegetable Store

Figure 4: Data frame consisting of neighborhoods and their top 10 common venues.

I utilized the venue data to explore the most common venues within specific areas by creating a function to return a bar chart that describes the most common venues. According to the choropleth map, West Vancouver is in the high price range, Burnaby is in the medium price range with a dense cluster of neighborhoods, Maple Ridge is in the low price range with neighborhoods that are spread apart, Port Coquitlam is also in the low price range however, the neighborhoods are much more clustered together. I analyzed these four distinct areas to examine similarities and differences in venues.

(a) (b)

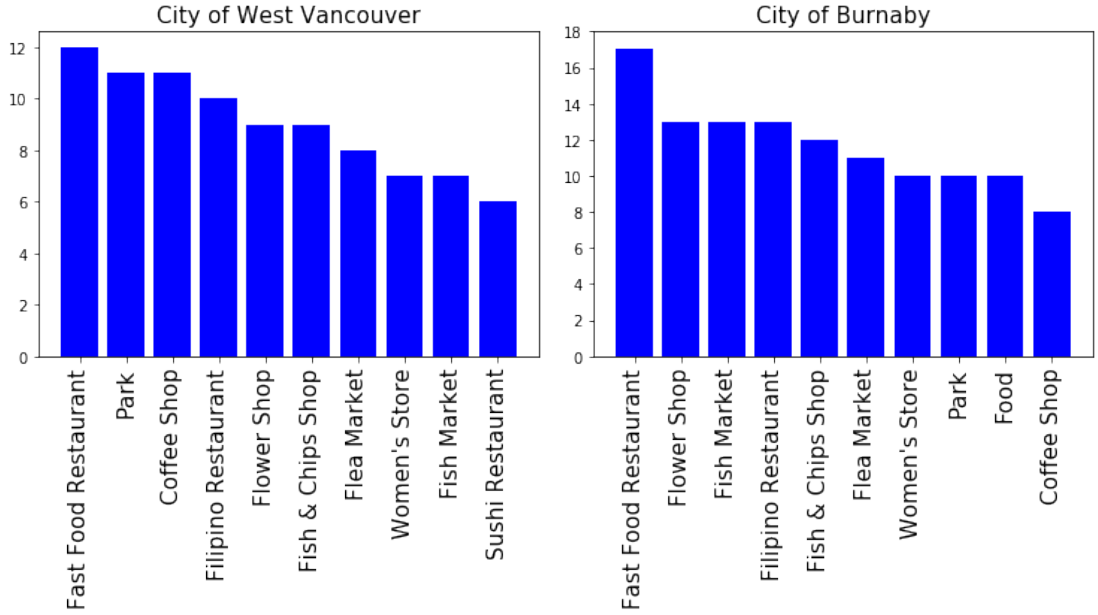

Figure 5: Figures a) and b) are bar plots displaying the 10 most common venues in West Vancouver and Burnaby respectively.

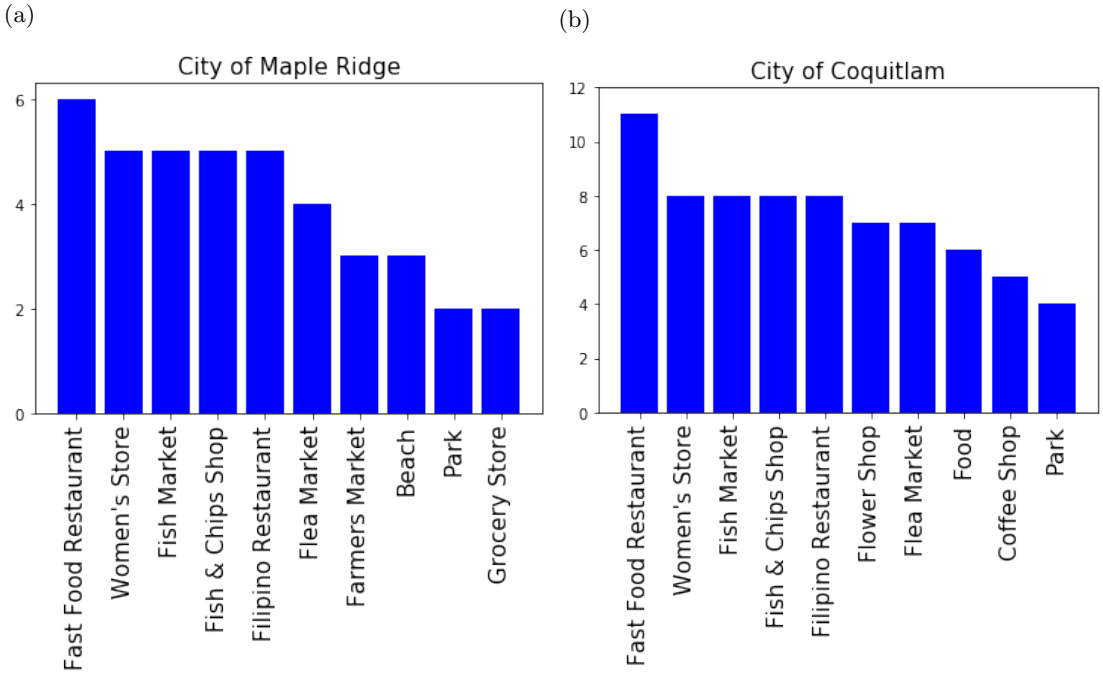

Figure 6: Figures a) and b) are bar plots displaying the 10 most common venues in Maple Ridge and Coquitlam respectively.

By inspecting the bar plots it is evident that there are many similarities in the types of venues between the four cities. For example, fast food restaurants are the most common venue in all four cities, with Burnaby having the most. Coquitlam and Maple Ridge share the same top 5 common venues, however, Coquitlam has twice as many locations. Another observation is that parks are a common venue in all cities although West Vancouver has the most park locations. In general, the medium and high price ranged cities have more restaurants and parks than low price ranged cities.

3.2 K-Means Algorithm

Due to the observations concluded from investigating the top 10 common venues of different cities, I decided to use the K-means clustering algorithm to locate clusters with similar venues. The K-means clustering method is an unsupervised learning algorithm that attempts to minimize the distance between points within a cluster with their centroid. It can easily be implemented and is intuitive. To prepare the data for the learning algorithm I preprocessed the data by transforming the categorical data into numerical data by the process of one hot encoding. Some problems that were encountered was When I doubled the size of my data set via the Foursquare API I was unable to fit the model well. Specifically, I was receiving inconsistent results when I examined the fit results.

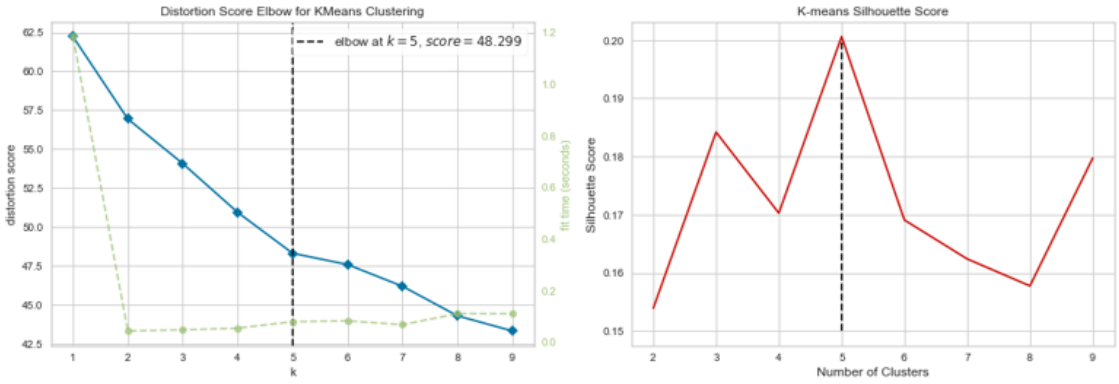

Figure 7: First figure displays results from the elbow method and the second figure shows the results from the silhouette score method. Both methods indicate k=5 is the optimal amount of clusters.

The optimal amount of clusters to use in the K-means algorithm may sometimes be arbitrary, however, the elbow method and silhouette score may determine the optimal number of clusters. To clearly demonstrate the inflection point at which the model becomes the best fit I utilized the k-elbow visualizer package from yellowbrick. The distortion score is the sum of squared error which is the difference between each coordinate and its centroid, squared and then summed. Similarly, the silhouette score is a metric to test the best fit, it ranges from -1 to 1 and from the figure below the peak is at k=5, which agrees with the elbow method. The silhouette score at k=5 is 0.2 which does not indicate the best fit as a score closer to 1 would indicate the best fit.

After fitting the data using 5 clusters I retrieved the labels for each neighborhood and appended the cluster labels to the original data frame along with latitude and longitude data see Figure 8. I also checked how many elements each cluster label contains, see the table below. The final result will include clusters from the K-Means algorithm superimposed onto the choropleth map describing the housing prices of each area.

	Area	Neighborhood	Area Average Price	Latitude	Longitude	Cluster Labels	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 Mos Commo Venu
0	Burnaby	Brentwood Park	1526503.0	49.275226	-122.992929	3.0	Bus Stop	Café	Park	Women's Store	Filipino Restaurant	Fish & Chips Shop	Fish Market	Fle Mark
1	Coquitlam	Coquitlam West	1206271.0	49.262768	-122.816954	0.0	Park	Bus Stop	Women's Store	Flower Shop	Fast Food Restaurant	Filipino Restaurant	Fish & Chips Shop	Fish Mark
2	Burnaby	Metrotown	1526503.0	49.225852	-123.003894	1.0	Bakery	Hotel	Toy / Game Store	Cosmetics Shop	Coffee Shop	Fast Food Restaurant	Sporting Goods Shop	Furniture Hom Stor
3	Vancouver East	Mount Pleasant VE	1455124.0	49.264048	-123.096249	1.0	Diner	Coffee Shop	Sushi Restaurant	Brewery	Bakery	Thrift / Vintage Store	Breakfast Spot	Vietnames Restaurar
4	Burnaby	Edmonds BE	1526503.0	49.212129	-122.959234	0.0	Park	Café	Playground	Coffee Shop	Gym / Fitness Center	Food & Drink Shop	Food	Flowe Sho

Figure 8: Final data frame consisting of all data collected including results from K-Means algorithm.

Cluster Label	Count
0	21
1	184
2	4
3	25
4	7

4 Results

To analyze the results from the K-Means algorithm I explored the 5 different clusters labels. Furthermore, I assembled a single plot containing 5 bar plots (one for each cluster) to visualize the types and frequency of venues allocated to each cluster.

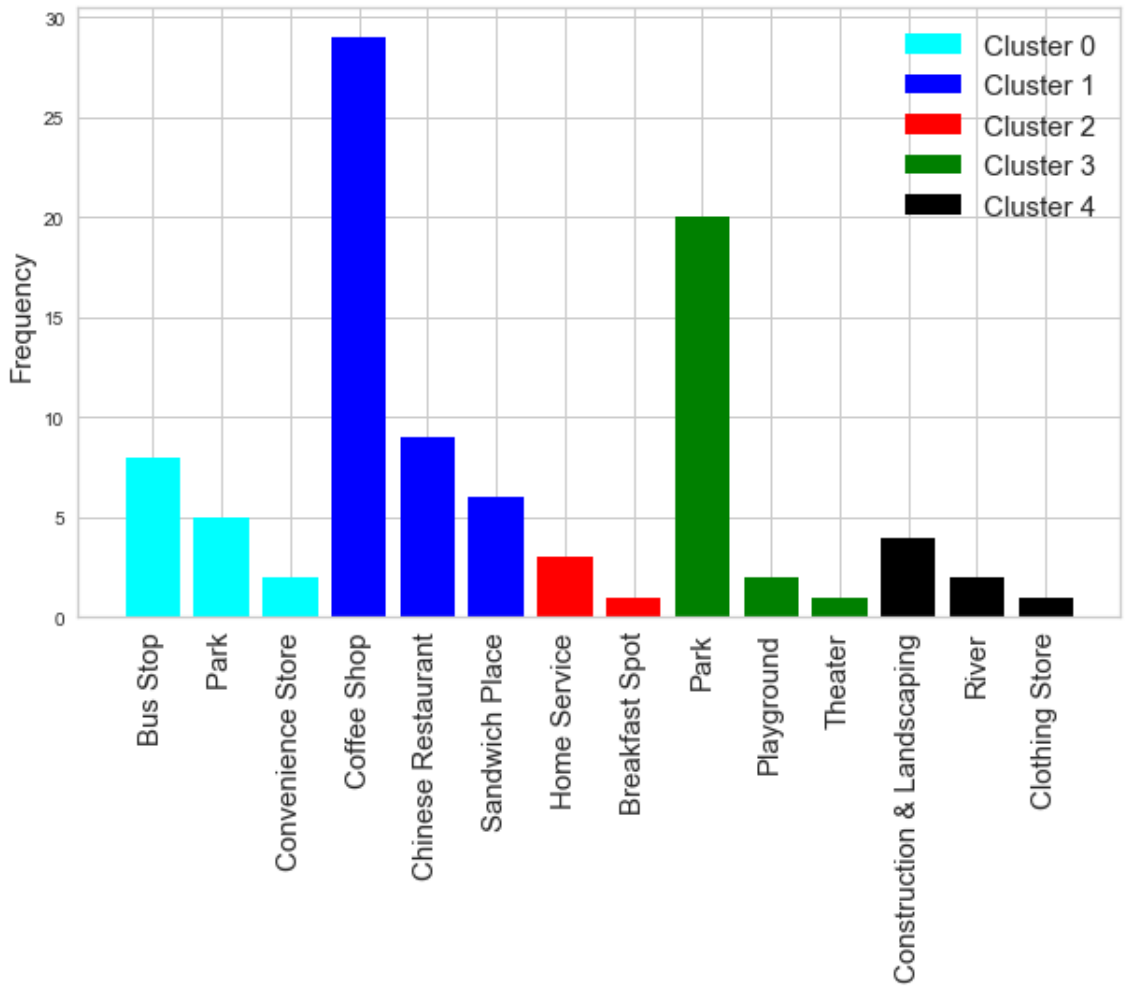

Figure 9: A single plot containing 5 individual bar plots describing the 5 cluster labels.

A summary of the cluster analysis is as follows. Cluster 0 primarily consists of bus stops and parks. Cluster 1 contains plenty of coffee shops and a variety of restaurants. Cluster 2 does not have many venues in this cluster, less than 5 home services and breakfast spots hence, it may not reveal many insights. Cluster 3 has many outdoor facilities like parks and playground with a few theatres. Cluster 4 is a composition of construction/landscaping, rivers and clothing stores and similar to cluster 2 does not contain many venues.

Note that when I analyzed these clusters I only retrieved the most common venue belonging to that cluster. For example, cluster 1 indicates that the most common venue is coffee shops but this number can be higher if I were to take into account the top 2 most common venues for each cluster label. I only included the most common venue for each cluster label assigned to each neighborhood due to simplicity and to generalize the cluster.

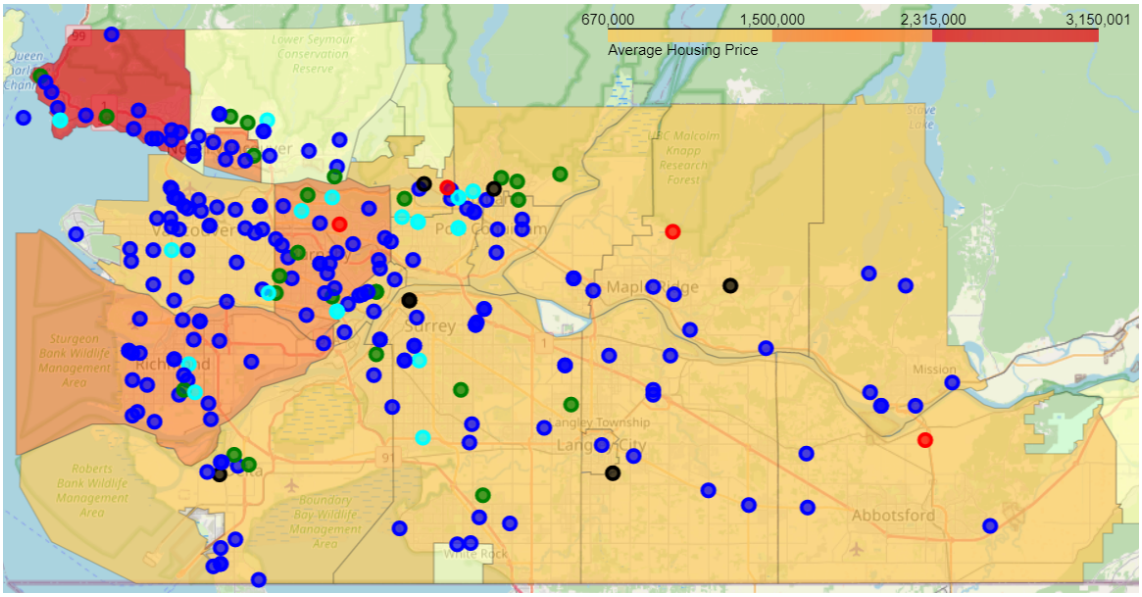

Figure 10: A single plot containing 5 individual bar plots describing the 5 cluster labels.

The final result is the chorpleth map superimposed with markers indicating cluster labels. In Figure 10 one can visualize the general center of the clusters color coded according to Figure 9. The K-Means algorithm was run many times with sometimes different results. However, the algorithm always resulted in a cluster describing coffee shops and restaurants, in addition to clusters describing outdoor facilities, bus stops and construction/landscaping.In the following section I will discuss any observations from these results.

5 Discussion

With the final map in hand many observations can be made. The high and medium areas which include West and North Vancouver, Burnaby and Richmond have many dense clusters of venues related to coffee shops and a variety of restaurants. Additionally, the parks and playground cluster is centered around Burnaby extending out to Richmond, West Vancouver, Port Coquitlam and Surrey. It is evident that the high and medium areas contain many restaurants and venues for leisure. Conversely, the low housing priced areas contain less densely populated clusters of restaurants and leisure venues.

Various home services and construction/landscaping venues are fairly distant from the high and medium areas except for a few outliers. Moreover, transit appears to be more common in the areas where neighborhoods are densely clustered. Consequently, travelling around these cities may be easier than Abbotsford, Delta, Langley and Maple Ridge.

The results from the K-means algorithm provides useful insights to make connections between housing prices and common venues throughout the lower mainland of BC. Further, I can suggest a few recommendations on where to buy a house. For example, New Westminster and Port Coquitlam seem like reasonable areas. Due to the fact that both of these cities are in the low housing price range, yet, still include densely clustered venues like coffee shops, restaurants, parks and transit. Moreover, if someone is looking to open a restaurant or coffee shop they should look to open one near a neighborhood in a remote area like Maple Ridge, Surrey, and Langley instead of Vancouver, Burnaby or Richmond.

6 Conclusion

Ultimately, expensive areas tend to have many restaurants, coffee shops, parks and outdoor facilities that are all nearby together. In addition, it is easier to traverse these cities because

transit is more available. By contrast, remote areas that have lower housing prices include a handful of venues that are spread apart with less options for transit. One may be able to use this type of information to make decisions about opening restaurants or buying a real estate property near desirable venues.

This project can be extended by exploring the Fraser Valley in addition to the lower mainland, including apartments and condo pricing data. This may yield potential new insights about the remote and populated cities. Furthermore, if housing price data could be extracted about neighborhoods along with a corresponding geoJSON file, analysis similar to this project may provide detailed analysis about the neighborhoods, instead of the areas that encompass the boroughs. Finally, research in this project alongside data about future infrastructure developments could result in predicting housing prices in certain areas.