House Renting Price and Venues of Calgary Study

Yichen Zhang

August 13, 2020

1. Introduction

Calgary is the largest city in the western Canadian province of Alberta. With its abundant natural resources, Calgary is also recognized as a leader in the Canadian oil and gas industry. Calgary's economy expanded at a significantly higher rate than the overall Canadian economy (43% and 25%, respectively) over the ten-year period from 1999 to 2009 [1]. The highly developed oil and gas industry also makes Calgary one of the best Canadian cities for engineers.

As an engineering student in Canada, there are a great number of my friends who want to start their career in Calgary. The first challenge they will encounter is the house renting problem. The objective of this project is to help those who are facing house renting problems in Calgary establish a clear understanding of neighborhoods and regional average renting prices in Calgary. This project would contribute to assisting newcomers in choosing the best-fit Calgary neighborhood to live.

Based on the problem, by the end of the project, one should be able to easily access information such as the number of top five venues in the area, average house rental price, and more by clicking the marker on the map. This project would provide the information needed for newcomers who want to rent a house, and help them decide where to rent a house in Calgary.

2. Data collecting and cleaning

The following data has been used in order to complete the study:

- The list of postal codes of Canada-T, which contains information about neighborhoods in Calgary, available at Wikipedia [2]
- The nearby venues information of neighborhoods in Calgary, which is obtained using Foursquare API [3]
- Average renting price of a specific area in Calgary, which is found using Rentometer [4]

The detail of the data collecting and data cleaning would be explained as follows.

2.1 Scrap the data about neighborhoods in Calgary

The Calgary neighborhood data could be obtained on a Wikipedia page, shown in Figure 1. In order to perform further analysis, the first step is to transform the data table into a *Pandas* data frame.

Alberta - 157 FSAs [edit]

Postal Code	Borough	Neighborhood	Latitude	Longitude
T1A	Medicine Hat	Central Medicine Hat	50.036460	-110.679250
T2A	Calgary	Penbrooke Meadows, Marlborough	51.049680	-113.964320
T3A	Calgary	Dalhousie, Edgemont, Hamptons, Hidden Valley	51.126060	-114.143158
T4A	Airdrie	East Airdrie	51.272450	-113.986980
T5A	Edmonton	West Clareview, East Londonderry	53.5899	-113.4413
T6A	Edmonton	North Capilano	53.5483	-113.408
T7A	Drayton Valley	Not assigned	53.2165	-114.9893
T8A	Sherwood Park	West Sherwood Park	53.519	-113.3216
T9A	Wetaskiwin	Not assigned	52.9741	-113.3646
T1B	Medicine Hat	South Medicine Hat	50.0172	-110.651
T2B	Calgary	Forest Lawn, Dover, Erin Woods	51.0318	-113.9786
T3B	Calgary	Montgomery, Bowness, Silver Springs, Greenwood	51.0809	-114.1616
T4B	Airdrie	West Airdrie	51.2816	-114.0153
T5B	Edmonton	East North Central, West Beverly	53.5766	-113.4608
T6B	Edmonton	SE Capilano, West Southeast Industrial, East Bonnie Doon	53.5322	-113.4404
T7B	Not assigned	Not assigned	Not assigned	Not assigned
T8B	Sherwood Park	Outer Southwest	53.4482	-113.2706
T9B	Not assigned	Not assigned	Not assigned	Not assigned
T1C	Medicine Hat	North Medicine Hat	50.0556	-110.6822
T2C	Calgary	Lynnwood Ridge, Ogden, Foothills Industrial, Great Plains	50.9878	-114.0001
T3C	Calgary	Rosscarrock, Westgate, Wildwood, Shaganappi, Sunalta	51.0388	-114.098

Figure 1. Neighborhoods data available at Wikipedia

By using the *BeautifulSoup* and *Pandas* library, the data is scraped and transform into a data frame shown in Figure 2.

0	T1A	Medicine Hat	Central Medicine Hat	50.036460	-110.679250
1	T2A	Calgary	Penbrooke Meadows, Marlborough	51.049680	-113.964320
2	T3A	Calgary	Dalhousie, Edgemont, Hamptons, Hidden Valley	51.126060	-114.143158
3	T4A	Airdrie	East Airdrie	51.272450	-113.986980
4	T5A	Edmonton	West Clareview, East Londonderry	53.5899	-113.4413
...	...	...	...	...	...
175	T5Z	Edmonton	West Lake District	53.5966	-113.4882
176	T6Z	Not assigned	Not assigned	Not assigned	Not assigned
177	T7Z	Stony Plain	Not assigned	53.5202	-114.0135
178	T8Z	Not assigned	Not assigned	Not assigned	Not assigned
179	T9Z	Not assigned	Not assigned	Not assigned	Not assigned

180 rows × 5 columns

Figure 2. The dataframe

After transforming data into a data frame, data cleaning could proceed. Since this project only concerns neighborhoods in Calgary, boroughs with names other than Calgary should be removed from the data frame. In addition, rows with unknown latitude and longitude should be removed as well. Once the data cleaning process is finished, the following data frame contains all 34 areas in Calgary shown in Figure 3 could be obtained.

22	T3B	Calgary	Montgomery, Bowness, Silver Springs, Greenwood	51.0809	-114.1616
23	T3C	Calgary	Rosscarrock, Westgate, Wildwood, Shaganappi, S...	51.0388	-114.098
24	T3E	Calgary	Lakeview, Glendale, Killarney, Glamorgan	51.0227	-114.1342
25	T3G	Calgary	Hawkwood, Arbour Lake, Citadel, Ranchlands, Ro...	51.1147	-114.1796
26	T3H	Calgary	Discovery Ridge, Signal Hill, West Springs, Ch...	51.0566	-114.1815
27	T3J	Calgary	Martindale, Taradale, Falconridge, Saddle Ridge	51.0999	-113.9422
28	T3K	Calgary	Sandstone, MacEwan Glen, Beddington, Harvest H...	51.127	-114.0787
29	T3L	Calgary	Tuscany, Scenic Acres	51.1162	-114.2089
30	T3M	Calgary	Cranston, Auburn Bay, Mahogany	50.8902	-113.9892
31	T3N	Calgary	Northeast Calgary	51.1494	-114.0019
32	T3P	Calgary	Symons Valley	51.1793	-114.1333
33	T3R	Calgary	Northwest Calgary	51.1497	-114.2695

Figure 3. Cleaned data frame

At this point, the data frame is ready to be used in further analysis. Notice that the area in this project contains all neighborhoods with the same postal code.

2.2 Get nearby venues using Foursquare API

When using Foursquare API, it is a good practice to define your credential first.

```
CLIENT_ID = '' # your Foursquare ID
CLIENT_SECRET = '' # your Foursquare Secret
VERSION = '20200813' # Foursquare API version
Limit = 100

print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

Figure 4. Foursquare API credentials

Then I defined a Python function that could get nearby venue information using Foursquare API and transformed the data into a *Pandas* data frame.

```
(831, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rundle, Whitehorn, Monterey Park	51.0759	-114.0015	Shark Club Sports Bar & Grill	51.076126	-114.002560	Sports Bar
1	Rundle, Whitehorn, Monterey Park	51.0759	-114.0015	Toad 'n' Turtle Pubhouse	51.076554	-114.002768	Gastropub
2	Rundle, Whitehorn, Monterey Park	51.0759	-114.0015	Basha International Foods	51.072712	-114.000067	Grocery Store
3	Rundle, Whitehorn, Monterey Park	51.0759	-114.0015	Tony Roma's Ribs, Seafood, & Steaks	51.079903	-114.002528	BBQ Joint
4	Rundle, Whitehorn, Monterey Park	51.0759	-114.0015	JOEY Barlow	51.081126	-114.004019	New American Restaurant
5	Rundle, Whitehorn, Monterey Park	51.0759	-114.0015	Thai Boat	51.081273	-114.005602	Thai Restaurant
6	Rundle, Whitehorn, Monterey Park	51.0759	-114.0015	Chapters	51.075447	-113.991020	Bookstore
7	Rundle, Whitehorn, Monterey Park	51.0759	-114.0015	Noodle King	51.079928	-114.005548	Vietnamese Restaurant

Figure 5. Venues data frame

2.3 Collect average house rental prices for each area

In this project, the major concern is the house rental prices for two bedrooms and one washroom. The average house rental prices are collected using the Rentometer platform. The data has been saved into a CSV file first and then transformed into the data frame. The CSV file is available at https://github.com/Alred21/Coursera_Capstone/blob/master/Calgary_Average_Rental.csv. Notice that the average house rental price is for the area that might contain

one or more neighborhoods. Therefore, the average house rental price for a specific neighborhood might not be applicable.

3. Data analysis

3.1 Locate the areas and create the map

Before creating the map, the latitude and longitude of Calgary should be found. *Geocoders* has been used to get the latitude and longitude as shown in Figure 6.

```
address = 'Calgary, AB'

geolocator = Nominatim(user_agent="YYC")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Calgary are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinate of Calgary are 51.0534234, -114.0625892.
```

Figure 6. The geographical coordinate of Calgary

Then using the data frame shown in Figure 3 and the *folium* map, the areas can be shown in the map.

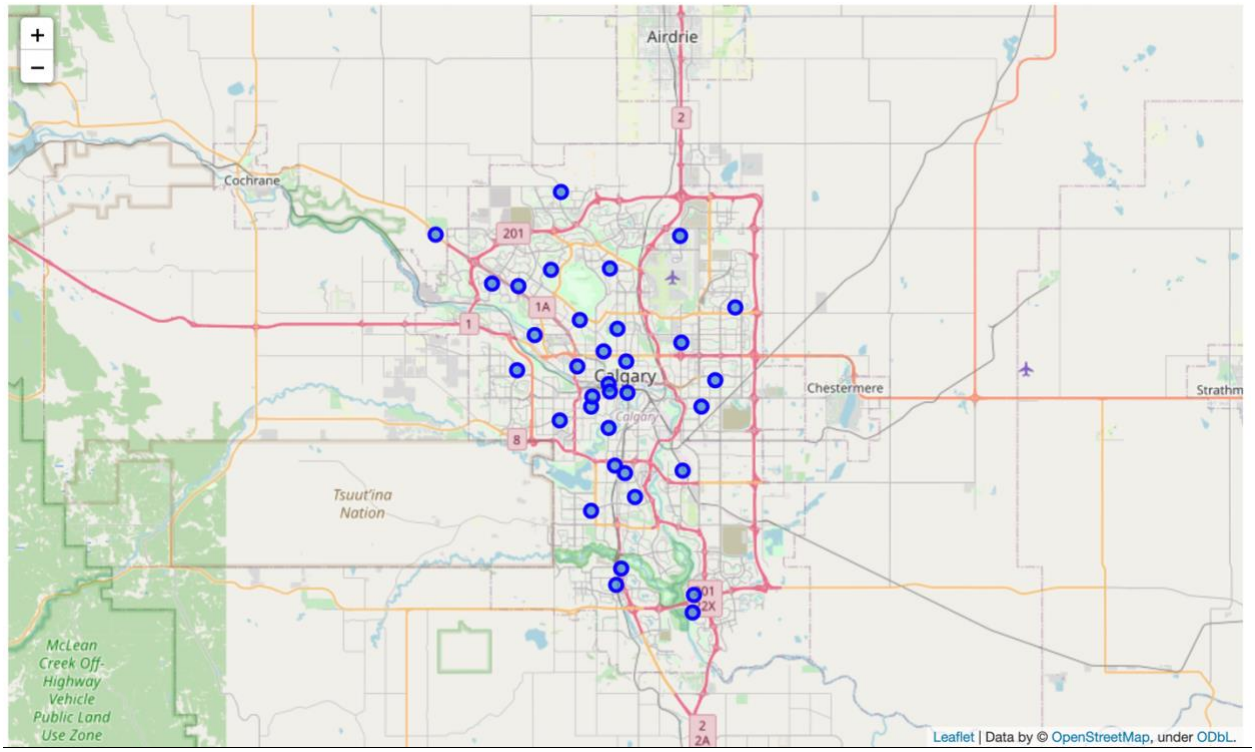

Figure 7. Thirty-four areas in Calgary

3.2 Preparing for clustering

This part aims to obtain the top ten most common venues for each area, which will be used to do clustering. As shown in Figure 5, the venue data has already been obtained. By using the one-hot encoding method, the most common venues could be found, as shown in Figure 8.

```
----Braeside, Cedarbrae, Woodbine----
                venue  freq
0     Hockey Arena  0.12
1             Pool  0.12
2              Pub  0.12
3  Ice Cream Shop  0.12
4         Pharmacy  0.12
```

Figure 8. Top 5 most common venues in the area

Now the data is ready to be transformed into a data frame, which is needed for clustering.

	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	Braeside, Cedarbrae, Woodbine	Ice Cream Shop	Pub	Coffee Shop	Gym	Convenience Store	Hockey Arena	Pharmacy	Pool	Food Court	Frozen Yogurt Shop
1	Brentwood, Collingwood, Nose Hill	Elementary School	Building	Hockey Rink	Sporting Goods Shop	Yoga Studio	Food Truck	Gas Station	Furniture / Home Store	Frozen Yogurt Shop	Fried Chicken Joint
2	Bridgeland, Greenview, Zoo, YYC	Noodle House	Chinese Restaurant	Scenic Lookout	Vietnamese Restaurant	Sandwich Place	Japanese Restaurant	Pub	Middle Eastern Restaurant	Bank	Coffee Shop
3	City Centre, Calgary Tower	Coffee Shop	Pub	Restaurant	Sushi Restaurant	Hotel	French Restaurant	Steakhouse	Italian Restaurant	Bakery	Bar
4	Connaught, West Victoria Park	Restaurant	Coffee Shop	Pub	Italian Restaurant	Pizza Place	Bar	Vietnamese Restaurant	Café	Brewery	Burger Joint

Figure 9. Data frame showing the top 10 most common venues in each area

3.3 Clustering the neighborhoods

With venue data fully prepared, the clustering can proceed. The unsupervised learning K-means algorithm is used in this project to cluster the neighborhoods. In order to get the optimal K value, the elbow method can be applied. According to Figure 10, the optimal value for K, seven, can be found at the elbow.

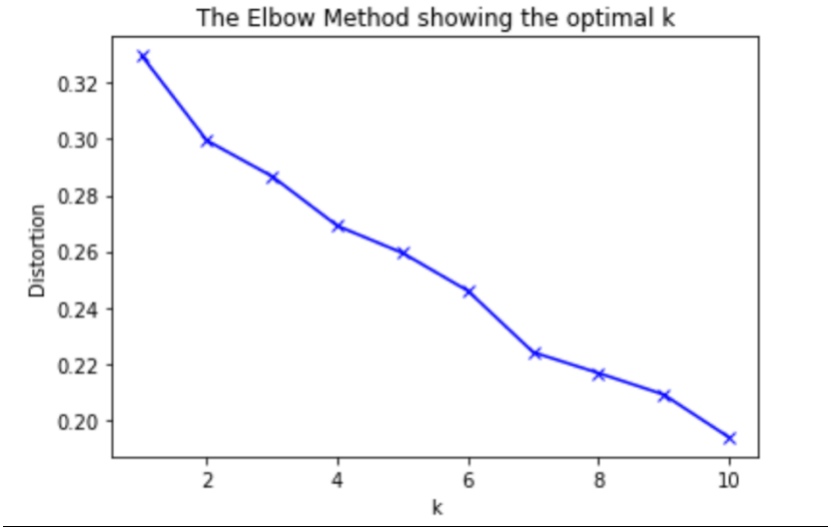

Figure 10. Get the optimal K using elbow method

Then by using *sklearn*, we can give cluster labels to all thirty-four areas. Combining all the results, we get the following data frame.

	PostalCode	Borough	Neighborhood	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	8 C
0	T1Y	Calgary	Rundle, Whitehorn, Monterey Park	51.0759	-114.0015	0	Vietnamese Restaurant	Hotel	American Restaurant	Burger Joint	Health & Beauty Service	Rental Car Location	Automotive Shop	Re
1	T2A	Calgary	Penbrooke Meadows, Marlborough	51.049680	-113.964320	6	Convenience Store	Vietnamese Restaurant	Fast Food Restaurant	Noodle House	Bank	Yoga Studio	Food Court	
2	T2B	Calgary	Forest Lawn, Dover, Erin Woods	51.0318	-113.9786	0	Convenience Store	Smoke Shop	Sandwich Place	Bar	Bank	Falafel Restaurant	Fried Chicken Joint	
3	T2C	Calgary	Lynnwood Ridge, Ogden, Foothills Industrial, G...	50.9878	-114.0001	0	Convenience Store	Diner	Clothing Store	Theater	Pizza Place	Falafel Restaurant	Food Truck	
4	T2E	Calgary	Bridgeland, Greenview, Zoo, YYC	51.0632	-114.0614	0	Noodle House	Chinese Restaurant	Scenic Lookout	Vietnamese Restaurant	Sandwich Place	Japanese Restaurant	Pub	Re

Figure 11. Data frame with cluster labels

Now we can visualize the clustering results based on map shown in Figure 7.

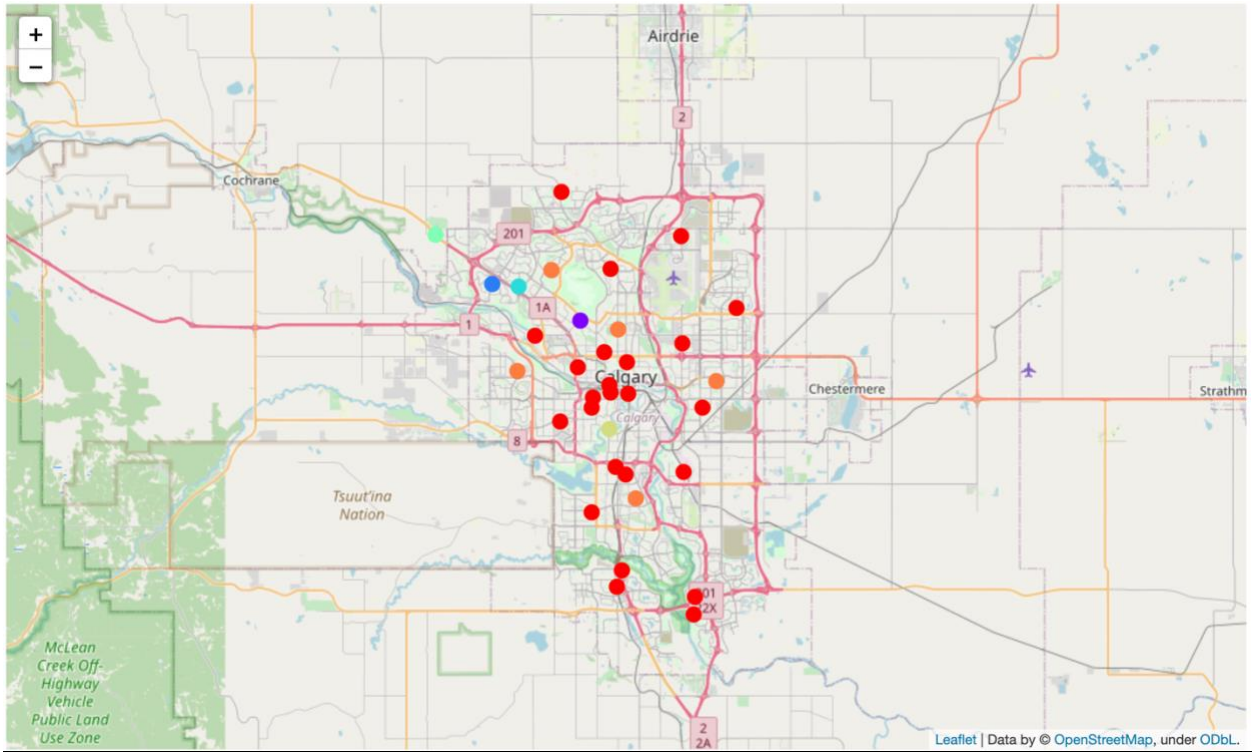

Figure 12. Clustering results

Next, we can rename the cluster labels based on the main feature of each area in order to make people understand the cluster labels. We can summarize the first most common venue in each area, and the result is shown below in Figure 13.

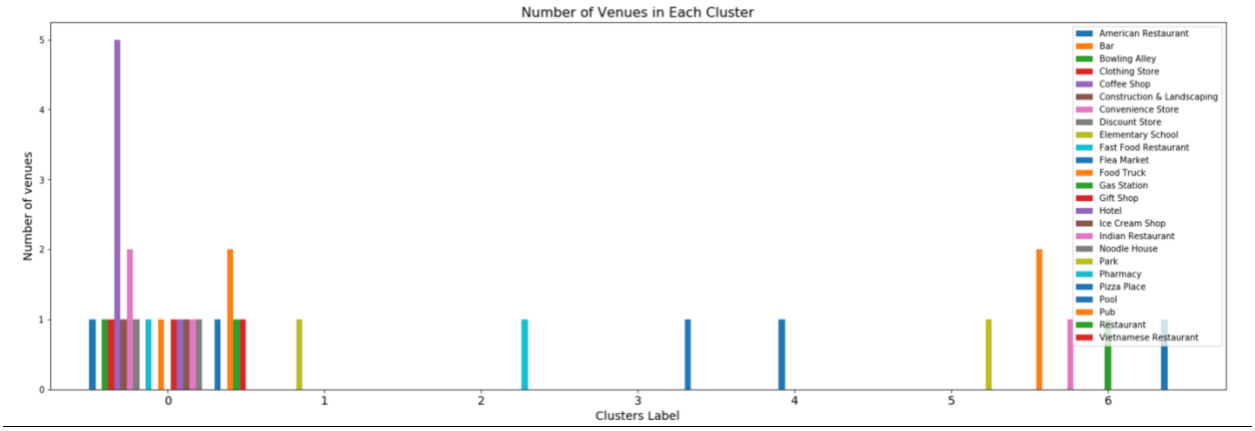

Figure 13. The 1st most common venue in each area

We can now rename the cluster labels and transform the data into a data frame.

	Clusters	Labels
0	0	Hockey Rink Venue
1	1	Playground and Liquor Store Venue
2	2	Pool and Gas Station Venue
3	3	Mixed Social Venue
4	4	Pizza Place Venue
5	5	Flea Market Venue
6	6	Skating Rink Venue

Figure 14. The renamed cluster labels

After completing this step, one can easily identify the main feature of an area in Calgary based on its cluster label.

3.4 Classify each area based on the average house rental prices

Using the data frame mentioned in section 2.3, we can get the following data frame showing the average house rental prices for each area from low to high.

	PostalCode	Neighborhood	Cluster Labels	Average_Renting_Price
0	T3J	Martindale, Taradale, Falconridge, Saddle Ridge	0	660
1	T2E	Bridgeland, Greenview, Zoo, YYC	0	988
2	T3R	Northwest Calgary	4	1027
3	T3P	Symons Valley	0	1042
4	T3L	Tuscany, Scenic Acres	2	1045
5	T3K	Sandstone, MacEwan Glen, Beddington, Harvest H...	0	1054
6	T3A	Dalhousie, Edgemont, Hamptons, Hidden Valley	6	1073
7	T3G	Hawkwood, Arbour Lake, Citadel, Ranchlands, Ro...	3	1090

Figure 15. Data frame contains average house rental prices

In order to make people quickly understand the average house rental prices in an area, we can classify and label the area based on its average house rental price. In this project, I classify them into five categories, which are low level, below average, average level, above average, and high level. The histogram showing the

average house rental price distribution and the data frame with the labels is shown in Figure 16 and Figure 17, respectively.

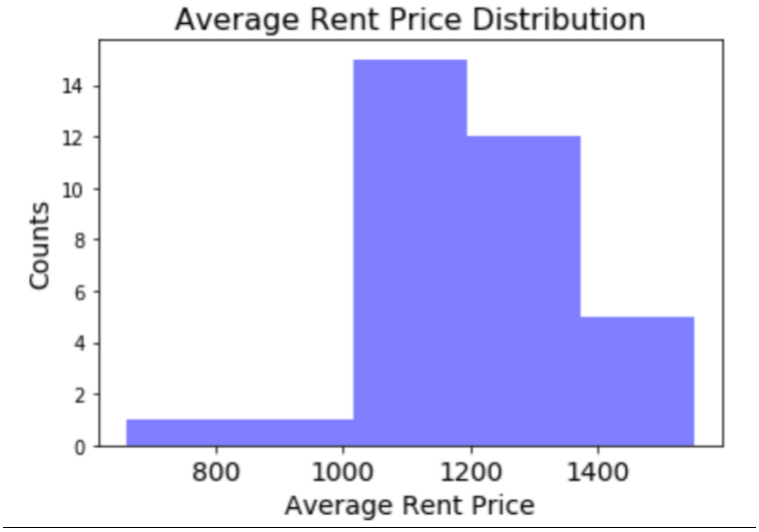

Figure 16. Histogram showing the house rental prices distribution

	Average_Renting_Price	Price-Categories
0	1273	Above Average
1	1258	Above Average
2	1192	Average level
3	1348	Above Average
4	988	Below Average
5	1352	Above Average
6	1215	Above Average

Figure 17. Data frame with the rent price labels

3.5 Get the distance from the city center

Similar to section 3.4, we can get the distance from the city center for each area and group them into three categories, close range, medium-range, and relatively far. The distance from the city center can be calculated using the latitude and longitude. In this project, I used a function to calculate the distance, which is available at https://www.geeksforgeeks.org/program-distance-two-points-earth/.

The histogram showing the average distance distribution and the data frame with the labels are shown in Figure 18 and Figure 19.

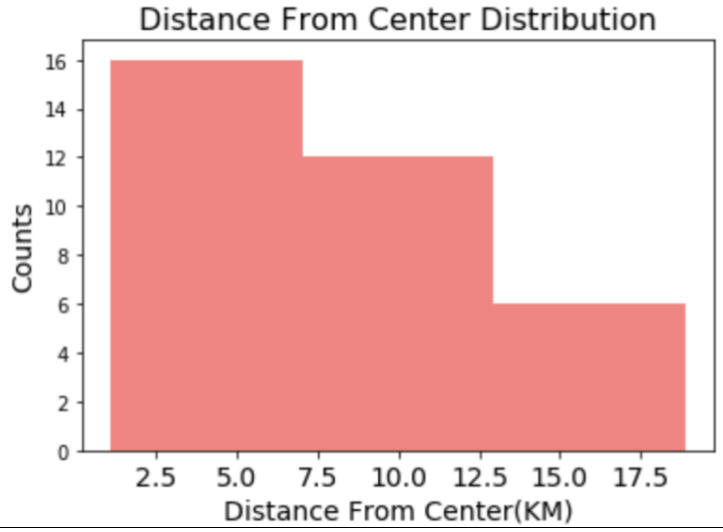

Figure 18. Histogram showing the distance from center distribution

	Distance From Center(KM)	DFC-Categories
0	4.946702	Close range
1	6.881563	Close range
2	6.345139	Close range
3	8.505917	Medium range
4	1.090281	Close range
5	1.339083	Close range
6	7.530583	Medium range

Figure 19. Data frame with the distance labels

3.6 Get the top five venues in the specific area

The last step of data analysis is to get the top five most common venues in each area. The results can be stored in a data frame before proceeding to the final visualization.

	Neighborhood	Top5
0	Braeside, Cedarbrae, Woodbine	1 Coffee Shop, 1 Convenience Store, 1 Gym, 1 H...
1	Brentwood, Collingwood, Nose Hill	1 Building, 1 Elementary School, 1 Hockey Rink...
2	Bridgeland, Greenview, Zoo, YYC	3 Chinese Restaurant, 3 Noodle House, 2 Bank, ...
3	City Centre, Calgary Tower	7 Coffee Shop, 5 Pub, 4 Restaurant, 3 Bakery, ...
4	Connaught, West Victoria Park	8 Restaurant, 4 Coffee Shop, 4 Italian Restaur...

Figure 20. Top 5 most common venues in each area

4. Results and Discussion

Before proceeding to the final visualization, we can first merge all the data needed and combine the data into one data frame.

Average_Renting_Price	Price-Categories	Distance From Center(KM)	DFC-Categories	Clusters	Labels	Top5
1273	Above Average	4.946702	Close range	0	Hockey Rink Venue	5 Vietnamese Restaurant, 4 Hotel, 3 American R...
1258	Above Average	6.881563	Close range	6	Skating Rink Venue	2 Convenience Store, 1 Bank, 1 Fast Food Resta...
1192	Average level	6.345139	Close range	0	Hockey Rink Venue	3 Convenience Store, 2 Bar, 2 Sandwich Place, ...
						1 Clothing Store, 1

Figure 21. Part of the final data frame

By using *folium* map, we are now able to complete the final visualization. The results are shown in Figure 22.

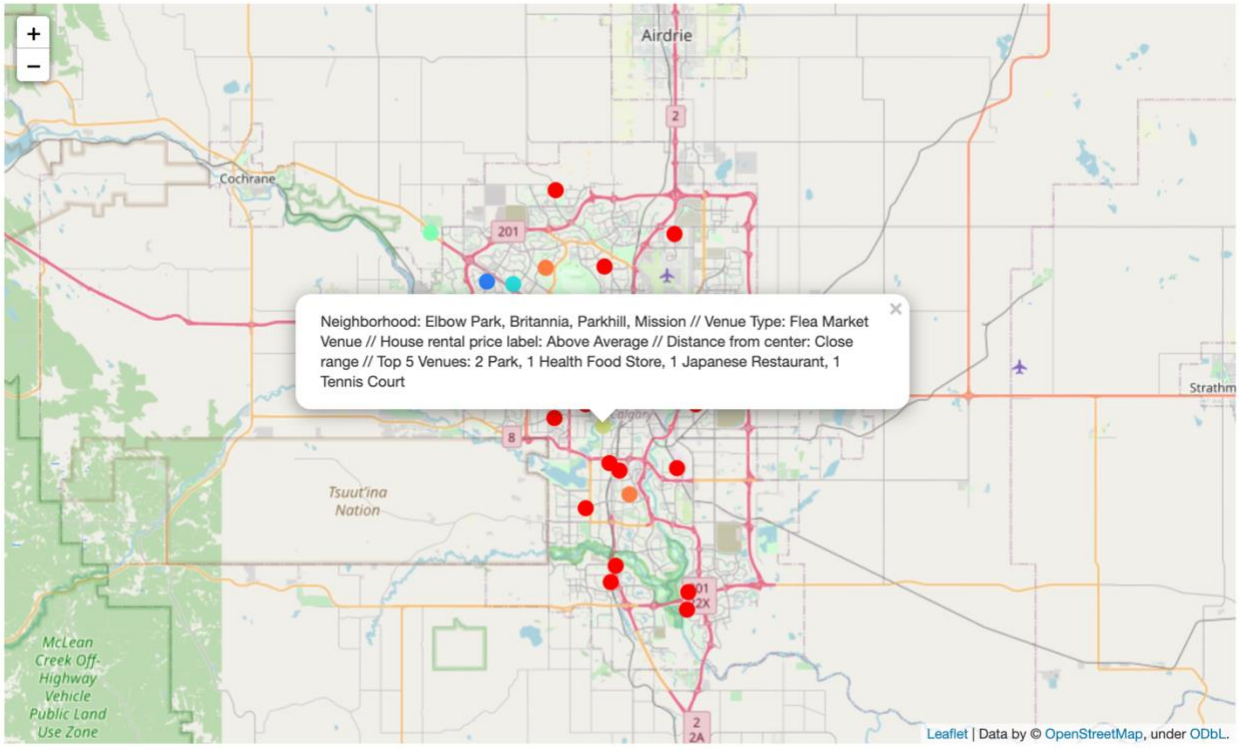

Figure 22. Final Visualization

As shown in Figure 22, people can quickly obtain information, including neighborhood names, venue type, house rental, distance from the city center, and the top five most common venues. One can quickly establish a general understanding of different Calgary areas by simply clicking the markers on the map. The information provided by this project will have a positive impact in assisting newcomers in deciding where to rent a house in Calgary. However, there are some limitations exist, which will be explained as follows.

First of all, I collected the neighborhood data based on postal codes. Neighborhoods with the same postal code are combined into an area. Therefore, one marker on the map might represent one or more neighborhoods. Although the complexity of collecting and cleaning data has been greatly reduced using this method, the accuracy of data could be limited. It could provide a general understanding of different areas in Calgary to people, but further analysis and more data would be needed if one requires information for a specific neighborhood in Calgary.

Besides, when finding the optimal K value for the K-means algorithm, the elbow method has been used. However, only 34 areas are included, which leads to

difficulties locating the elbow. With the need for higher accuracy, the current data set could be expanded.

For the average house rental prices, data was collected using the Rentometer platform and transformed into a CSV file. However, the average house renting prices in an area could vary every month. Therefore, in future studies, the data could be directly obtained from the platform to remain up to date.

5. Conclusion

In this study, I analyzed the average house rental prices in Calgary. I have successfully completed this project and visualized the results on the map. This project would be helpful for people who want to rent a house in Calgary. One can quickly establish a general understanding of different Calgary areas by simply clicking the markers on the map. People can quickly obtain information such as neighborhood names, venue type, house rental, distance from the city center, and the top five most common venues, which helps them decide where to rent a house in Calgary.

6. Reference

[1] Calgary

[2] List of postal codes of Canada: T

[3] Foursquare API

[4] Rentometer

The project Jupyter Notebook is available at https://github.com/Alred21/Coursera_Capstone/blob/master/Calgary_Neighborhoods_and_House_Rental_Prices_Study.ipynb