Coursera Capstone Project

Analyzing coffee shops in Portland, Oregon

Akkawatch Thouchamongkol

July 2020

1. Introduction

Opening a coffee shop is a dream business for many people because it is a business that can make highly profitable, just making a great coffee. While a great coffee is the heart of the business, the location of the shop is also the most important thing to consider too. Because if choosing the wrong location, in other words, choose a location that is not the target group, or a location in an area with a lot of competitors, this may cause the sales will not meet the target.

So, if someone considers opening a coffee shop, what type of place will be a good location to start a business? By looking on the plain map, it is hard to say where is the good place to be setting up a shop. On the other hand, it will be great if we can have this information combined with a map of a city or a location of interest. The main goal of this project will be exploring the coffee shops in Portland, Oregon, and combine with key indicators of each area, to extract some insights so people that want to open a new coffee shop can scope down the point of interest and made them make a decision easier.

This project will be useful to someone who considers opening a coffee shop. People who like to visit the coffee shops may also interest, or someone that would like to open the other type of venue like a restaurant, just adapt a few feature or process to the related field.

2. Data

2.1 Data sources

The main dataset used in the project is the US Household Income Statistics from Golden Oak Research Group from Kaggle's website [1]. The database contains about 32,000 records on US Household Income Statistics & Geo Locations. A full description of the dataset can be found on Kaggle's website [1].

Information about all venues including venue identification, names, their coordinates, distance from the defined coordinates, and categories obtained by passing the required parameters to the Foursquare API including the coordinates from the main dataset, credentials, API version, radius around the defined coordinates and limit of the results.

2.2 Data cleaning

After data was downloaded, the data was read and transformed into a table by pandas dataframe function. First, the data consist of 32,536 rows and 19 columns. The location summary of data was created by grouping all data by state, county, and city and counting all samples in each group. There are 52 states, 1,133 counties, and 8,128 cities in the dataset.

Some columns that not necessary were dropped. For example, id which is the label for each row, state abbreviation which is the same meaning as state name, type of city, zip code, area code, etc.

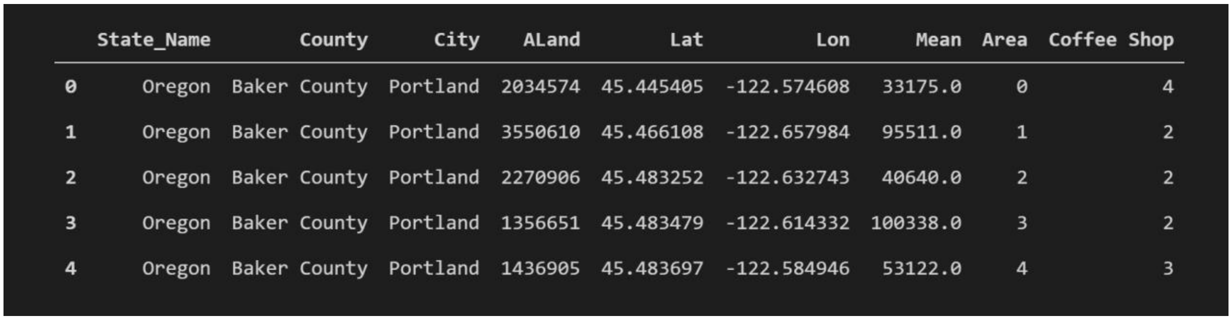
Then, the size of the unit of analysis must be considered. The unit of analysis must be not too small to cause a very small number of venues after getting the result from Foursquare API, and not too big to reach the limit number of venues after getting the result, even if the limit is set more than 100, the result still not more than that. After considered, selecting the city as a unit of analysis is the best choice, which will be segmented into each area.

After that, the data was extracted into a new dataset which has a state name equal to Oregon, and a city equal to Portland. The extracted data contain 64 areas in Portland and 58 unique areas. This means the dataset has 6 duplicates area. Duplicates were removed by aggregate the column which has different values but the same coordinates into only one row per coordinate.

Another problem is total coordinates in the Portland dataset was not equal to the total real city's neighborhood, and since the dataset most specific location label is the same place but a different coordinate, each coordinates in the dataset cannot assume as a neighborhood. So, each coordinate was called as an area and assigned a different number as a label to each coordinate.

Finally, to obtain the number of coffee shops in each area, all venues around each area coordinate and it's categories need to obtain from Foursquare API first, then filter out categories that are not the coffee shops then group and count the data by area. The thing to consider is the radius parameter, in this case, the radius is set to 1,000 meters, which can get enough venues, but caused an overlapping. To fix this, a venue id and venue distance from the coordinate need to be obtained from the Foursquare API, then sort the venue dataframe by venue distance ascending, then use panda's drop duplicate function and keep the first, that is, the minimum distance to the coordinate among the duplicates. This will make sure that all the venues are unique.

After cleaning the dataset, the dataset consists of 58 rows and 9 columns. The cleaned dataset with a first 5 rows is visualized as shown in figure 1.



	State_Name	County	City	ALand	Lat	Lon	Mean	Area	Coffee Shop
0	Oregon	Baker County	Portland	2034574	45.445405	-122.574608	33175.0	0	4
1	Oregon	Baker County	Portland	3550610	45.466108	-122.657984	95511.0	1	2
2	Oregon	Baker County	Portland	2270906	45.483252	-122.632743	40640.0	2	2
3	Oregon	Baker County	Portland	1356651	45.483479	-122.614332	100338.0	3	2
4	Oregon	Baker County	Portland	1436905	45.483697	-122.584946	53122.0	4	3

Figure 1. A cleaned dataset with first five rows

2.3 Feature selection

Selected features will consist of some variables about key indicators of the economic and the number of coffee shops in each area. The selected features name and its description are described as follows:

- ALand: The square area of land at the geographic or track location.
- Mean: The mean household income of the specified geographic location.
- Coffee Shop: Total number of coffee shops per area.

3. Methodology

This section describes in detail the method of collecting a number of coffee shop data and the algorithm used on the data.

3.1 Method of collecting a number of coffee shop data

As stated in section 2, the radius parameter is the thing to consider. The problem is some area is very close, some area is far from each other. Which cause a problem in selecting the appropriate number of radius to cover most area in the city. At first, a radius of about 500 meters was selected, but it causes a very small number of venues. Then, the radius was changed to 1,000 meters. The result is caused by overlapping in many areas, which about 500 duplicate venues in the dataset. The area selected by defined radius and coordinates, to get the result from the Foursquare API is visualized as shown in figure 2.

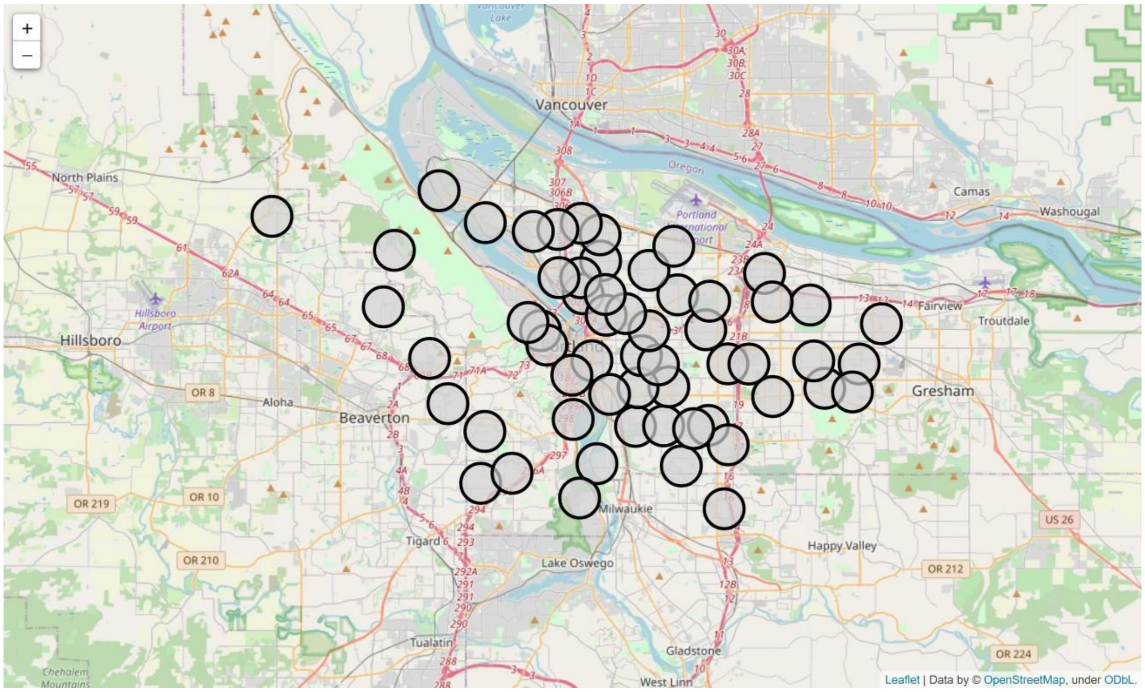
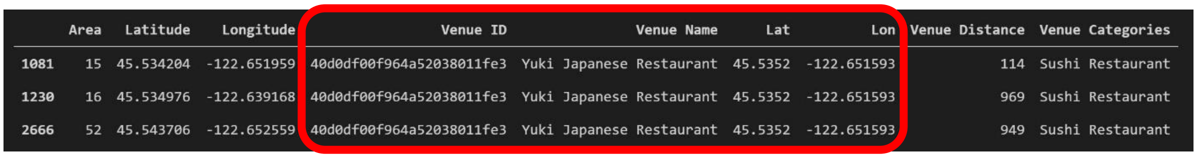


Figure 2. Each Portland area in a black circle defined by radius and coordinates

After getting the result, many duplicates must clean up. In some specific venue, there are two or more duplicates from a different area which come from an overlapping. The method to clean up the duplicate is to find which area has the least distance to the venue, then keep the venue in that area and drop the others. Duplicate venues in the dataframe are visualized as shown in figure 3.



```
      Area   Latitude    Longitude                  Venue ID               Venue Name      Lat          Lon  Venue Distance  Venue Categories
1081    15  45.534204  -122.651959  40d0df00f964a52038011fe3  Yuki Japanese Restaurant  45.5352  -122.651593             114  Sushi Restaurant
1230    16  45.534976  -122.639168  40d0df00f964a52038011fe3  Yuki Japanese Restaurant  45.5352  -122.651593             969  Sushi Restaurant
2666    52  45.543706  -122.652559  40d0df00f964a52038011fe3  Yuki Japanese Restaurant  45.5352  -122.651593             949  Sushi Restaurant
```

Figure 3. Example of duplicate venues in the dataframe

Finally, the venue dataframe was cleaned up. The dataframe consists of 2,404 unique venues. Then, only the coffee shop category is required so other categories are filtered out. Now the dataset contains only the venue that category is the coffee shop. The location of coffee shops around Portland is visualized by a map using a folium library, with the center of the map around Portland. Each coffee shop data in the dataset has a latitude and longitude values. Those values were passed in as a location parameter in folium’s circle marker function and added to the map. The coffee shop around Portland is visualized as a filled black circle as shown in figure 4.

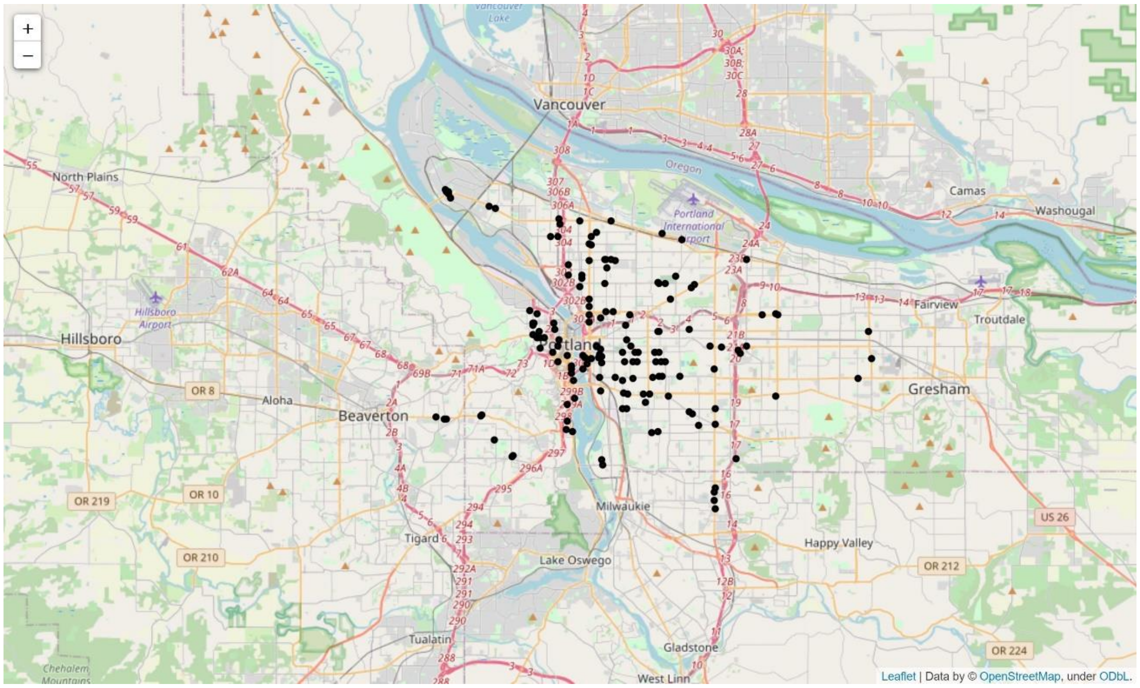


Figure 4. Location of the coffee shop around Portland, Oregon

From the map, there is a dense coffee shop at the center of the city and some coffee shops in the suburb. In the next section, the coffee shop will be segment and cluster into a group based on the selected indicator in the dataset.

3.2 Clustering data

The target of analysis is to see the relationship of the coffee shop using the select indicator in the dataset. So, k-means clustering was used. To do this, the features must be selected first. The detail of the selected feature is explained in section 2.3. Then, the data were normalized by removing the mean and scaling to unit variance. The k-means algorithm required k number of clusters as a parameter. The number of optimal k is found by running the algorithm and calculate the minimized within-cluster sum of squares distances of samples to their closest cluster center for a range of values for k called elbow method. The result of an elbow method is visualized as shown in figure 5.

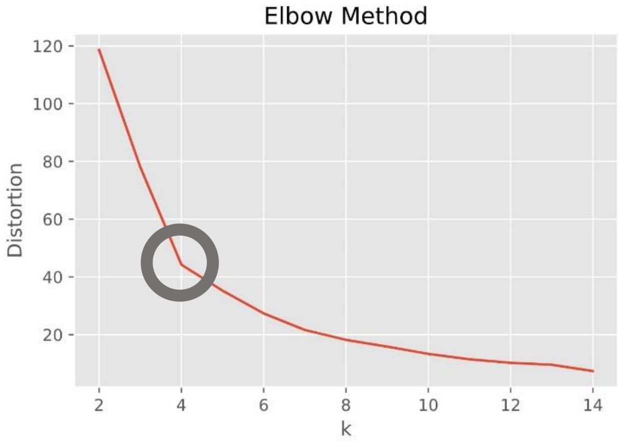


Figure 5. Elbow method represents the optimal k in k-means clustering

The number of optimal k is defined by the elbow point in an elbow method. From the result, the elbow point is at the circle which k equal to 4. So, the data will segment into 4 groups. The result of the algorithm is the cluster label for each area. Then, these labels were inserted into a new column on the dataframe. The cluster label will be used to make a summary of each cluster in the next section.

4. Results

This section describes the summary of the clustering algorithm. First, a map center around Portland with a marker of each area was created to visualize the area and their belong clusters as shown in figure 6. The cluster is separated by a different color. Most of the area in the cluster with an orange marker located around the center of the city while the area with a yellow marker and black marker located further from the center of the city, and area with a purple marker is located far from the city.

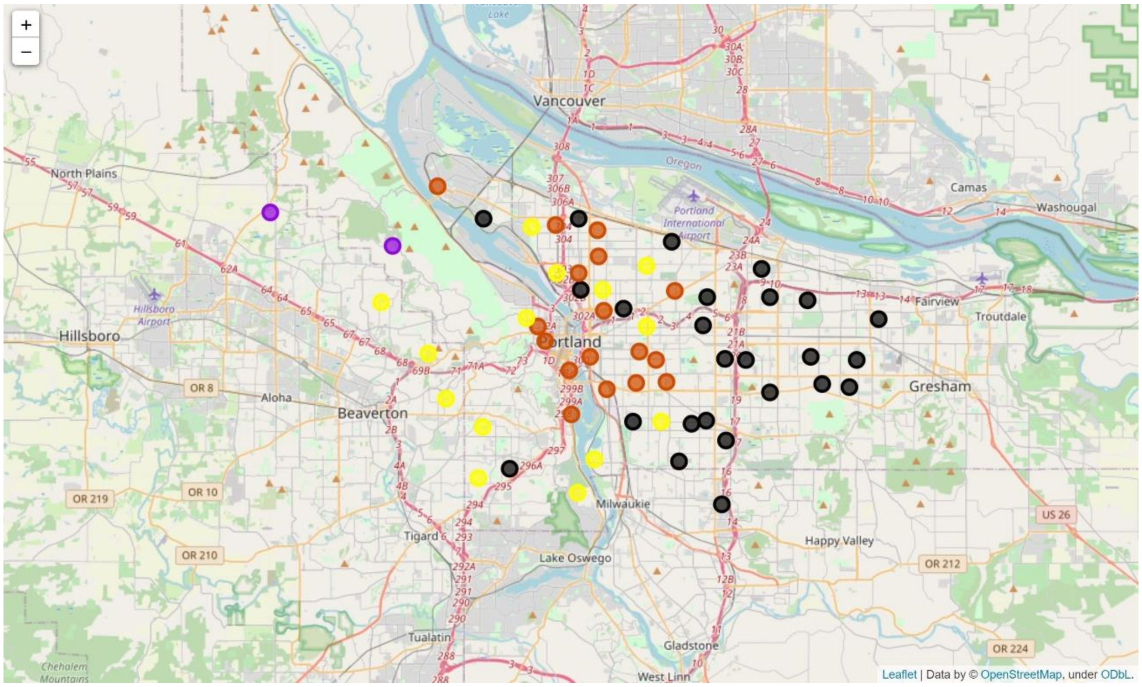


Figure 6. Map of the area in Portland with color-separated by a cluster

Next, the values of each area on each feature were summarized as visualized in figure 7 and figure 8. Barplot was used to describe the summary. From the figure, the plot has 3 features including average income, a square area of land, and the total number of coffee shops. These features are divided into 3 parts as each subplot. Each subplot is separated by x-ticks as each cluster and each cluster including the value of each area in the cluster as a bar. On the y-axis is the values of each feature. And brown dashes line in each subplot is an average value of each feature.

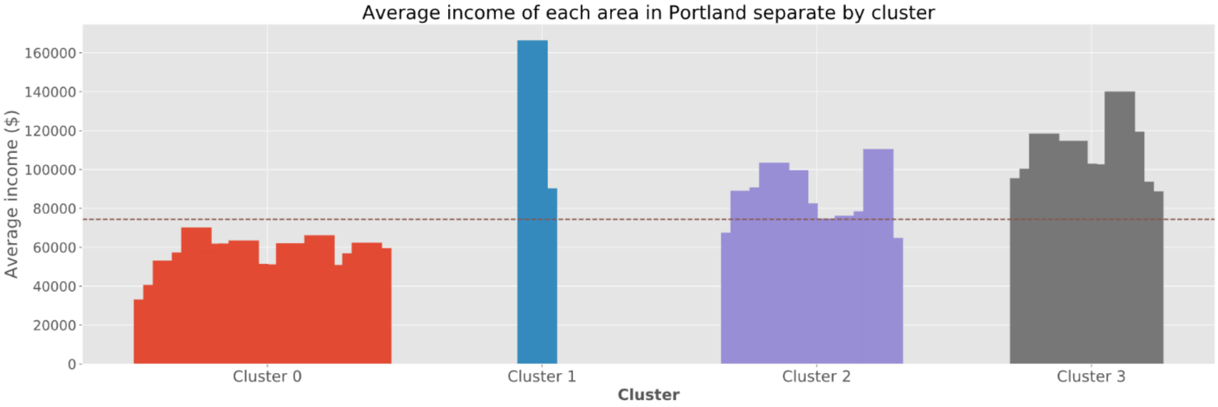


Figure 7. Barplot summarizing each cluster statistics on each feature

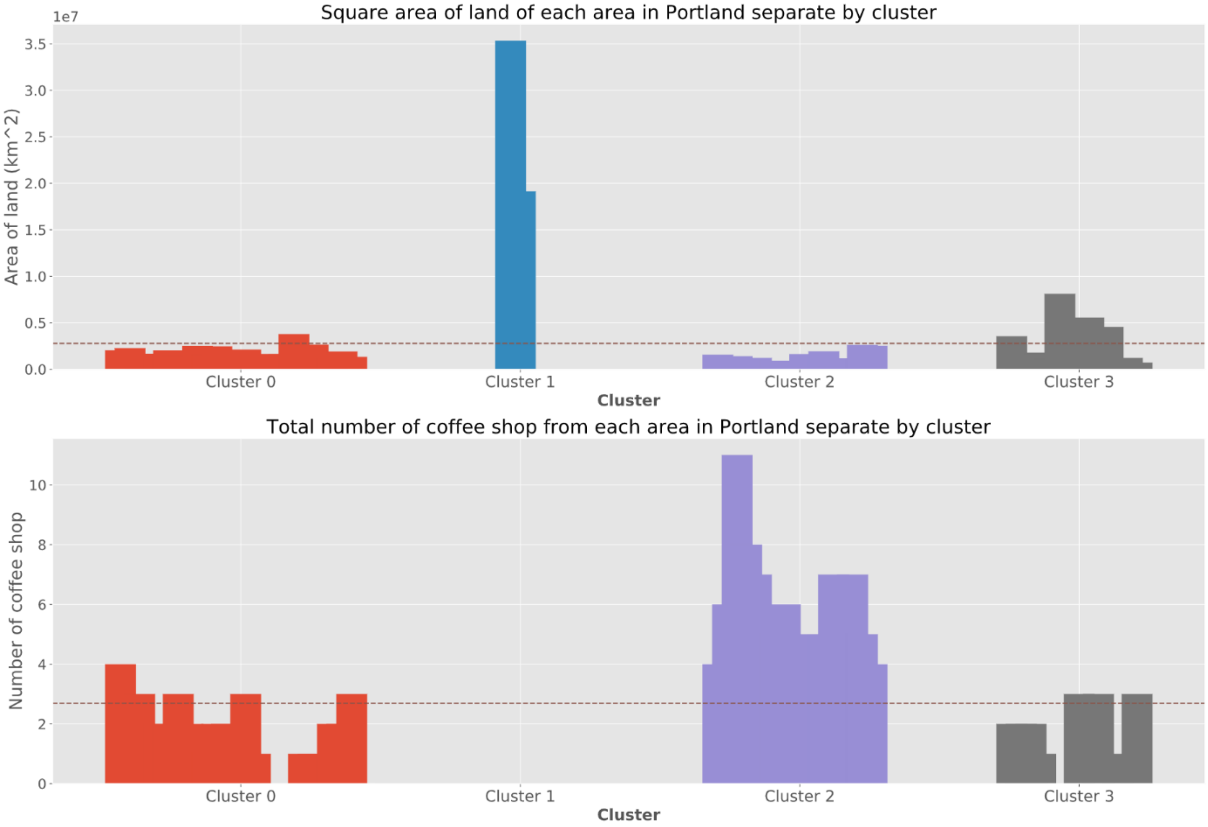


Figure 8. Barplot summarizing each cluster statistics on each feature (Continue)

Finally, the values of the area from each cluster on the plot can describe the relationship as follows:

- Cluster 0: Below average household income, small area of land, and several coffee shops.
- Cluster 1: High household income, large area of land but no coffee shops.
- Cluster 2: Around average household income, small area of land, and many coffee shops.
- Cluster 3: High household income, medium area of land, and several coffee shops.

From the result, these characteristics are distinguished as the difference can be seen on each cluster. These are the insight that can serve the needs of anyone who related to the problem as defined.

5. Discussion

As an observation from the result section, notice that the coordinate did not use as a feature in the algorithm, but each area in the same cluster seems to be close to each other in some side of a city and a little mixed in some part significantly as visualized in figure 9. From the figure, there is a cluster of the purple marker is considered as the error. The purple marker on the far left is not in the Portland area, and another purple marker is near the hill. The cluster can be removed and done the clustering again.

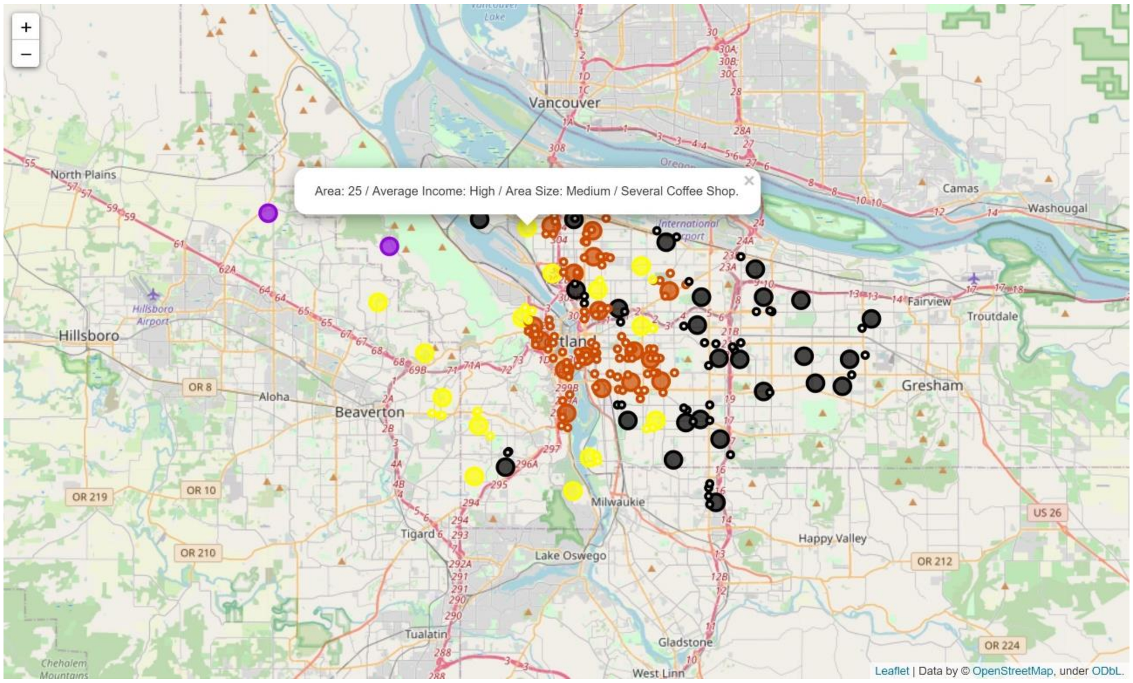


Figure 9. Map of the area and coffee shops in Portland with color-separated as a cluster

On the other hand, after comparing the characteristics among the cluster, the best answer for the business problem, in this case, could be cluster 3 of the yellow markers in term of high household income, medium land area size, and several coffee shops for not much competitors as it should be. However, other clusters can be considered depending on individual circumstances.

6. Conclusion

In this project, we have gone through the process of defining business problem, identify and gather the data required, explore and prepare data for modeling, segmenting and clustering, and examine the results. The area on Portland was clustered into 4 groups, each have their unique characteristics. Therefore, these are just primary analysis with the purpose to scope down the point of interest and make people who want to open a coffee shop make a decision easier. There are much more additional factors to consider of, for example, age targeting, budget and/or environment around. Final decision can be made upon considering all the factors and select the best place based on the requirement, situation, and constraint.

Reference

[1] https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations