Coursera Capstone Project

Analyzing coffee shops in Portland, Oregon

Introduction & Business Problem

- Location is one of the most important thing to consider when opening a coffee shop.
- By looking on the plain map, it is hard to find the best place to setting up a shop.
- It will be great if we can have the information combined with a map of a city or a location of interest that can use to scope down the point of interest and made them make a decision easier.
- Useful to someone who considers opening a coffee shop.

Data sources

Main dataset - US Household Income Statistics from Golden Oak Research Group

- https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations
- Contains 32,536 rows and 19 columns.

Venue data - Foursquare API

- Venue identification
- Venue names
- Venue coordinates
- Distance from the defined coordinates
- Venue categories

Data cleaning

- Some columns that not necessary were dropped – id, state abbreviation, zip code, etc.
- The size of the unit of analysis must be considered, not too small and not too large.
- Removing duplicates by aggregate the column which has different values but the same coordinates into only one row per coordinate.
- Each coordinate was called as an area and was assigned a different number as a label to each coordinate.
- After cleaning the dataset, the dataset consists of 58 rows and 9 columns.

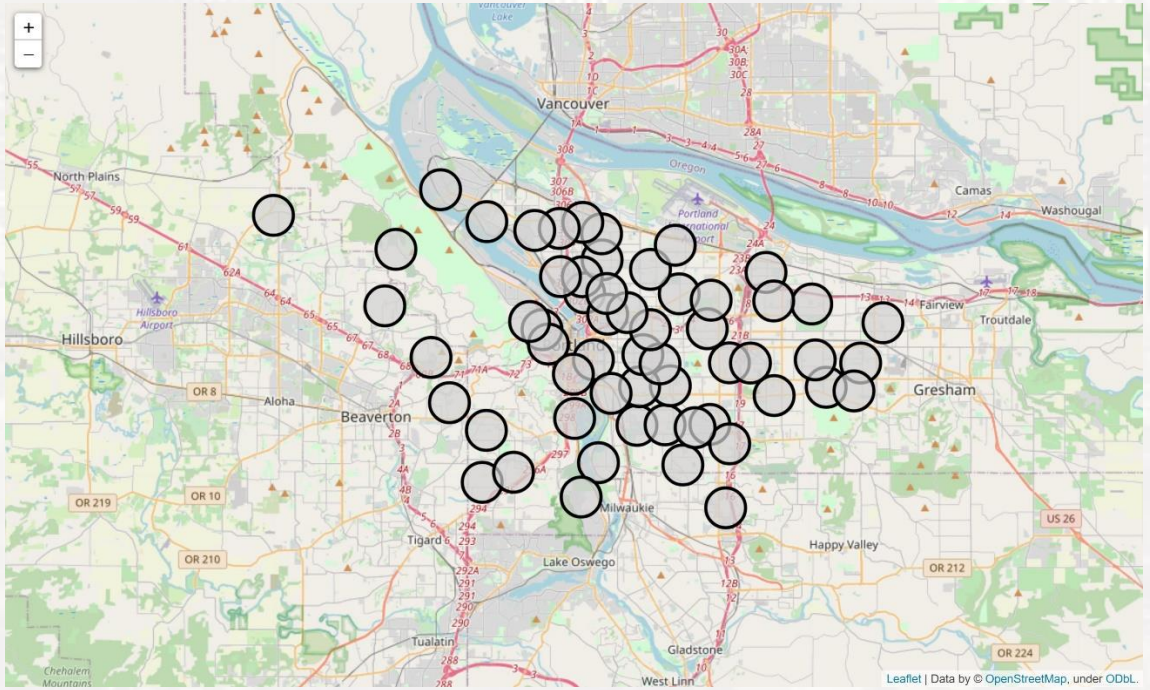
Feature selection

- 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.

	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

Methodology - Collecting a number of coffee shop data

- Radius parameter is the thing to consider.
- 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.



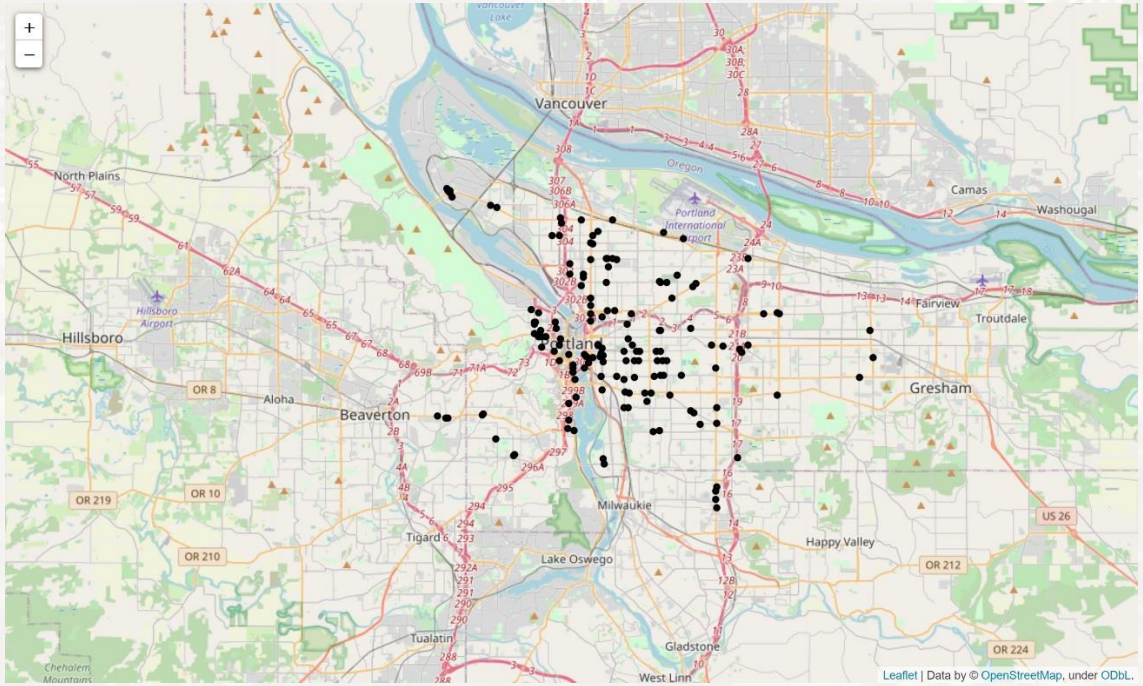
Methodology - Collecting a number of coffee shop data

- 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.

	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

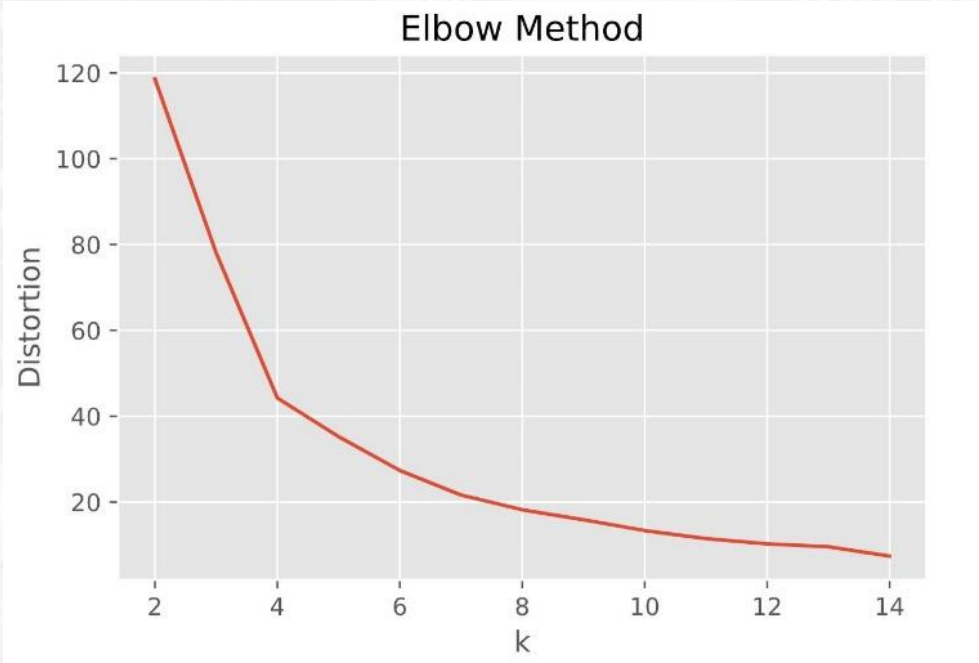
Methodology - Collecting a number of coffee shop data

- The cleaned dataframe consists of 2,404 unique venues.
- Then, only the coffee shop category is required so other categories are filtered out.
- The location of coffee shops around Portland is visualized by a map using a folium library, with the center of the map around Portland.
- There is a dense coffee shop at the center of the city and some coffee shops in the suburb.



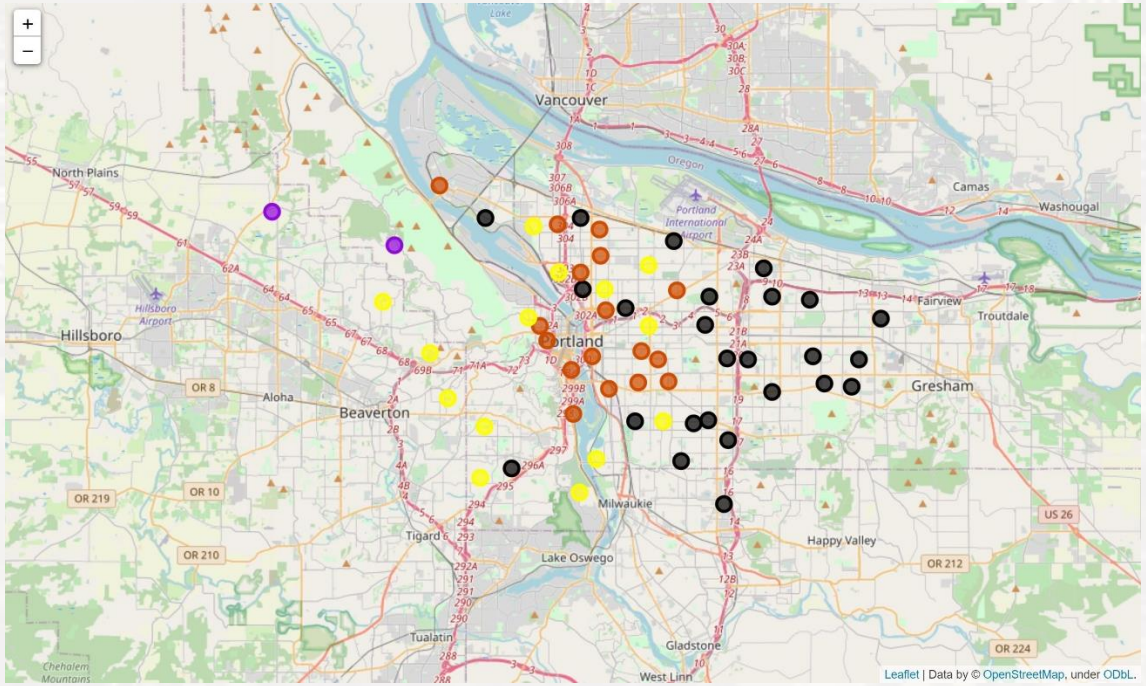
Methodology - Clustering data

- 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.
- Data were normalized by removing the mean and scaling to unit variance.
- The number of optimal k is found by the elbow point in an elbow method.
- So, the data will segment into 4 groups.



Results & Discussion

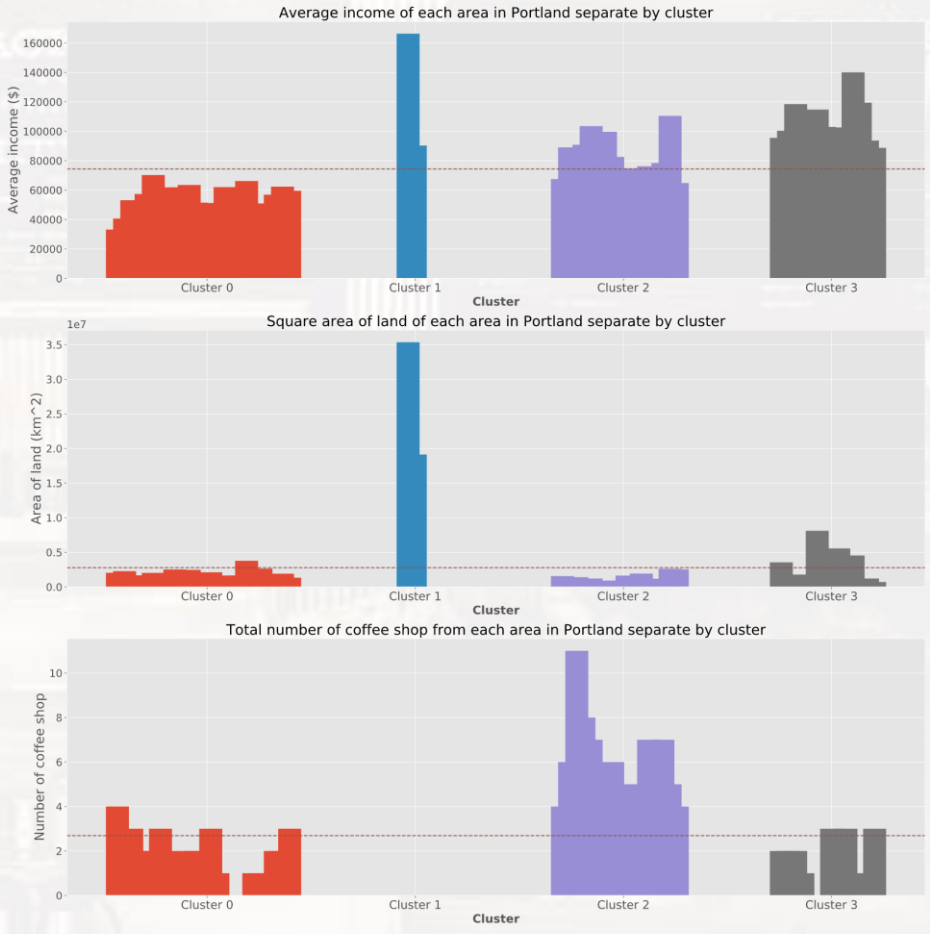
- The cluster is separated by a different color.
- Most of the area in the cluster with an orange marker located around the center of the city.
- The area with a yellow marker and black marker located further from the center of the city.
- Area with a purple marker is located far from the city.



Results & Discussion

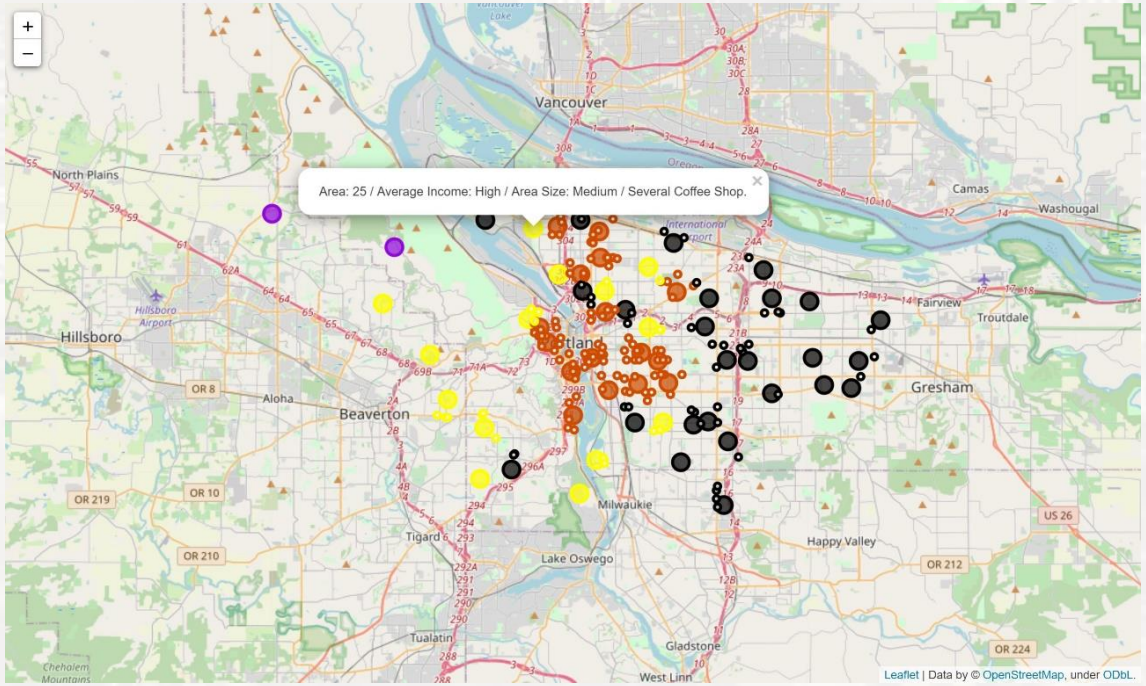
- The relationship can be described as follows:
  - Cluster 0: Below average household income, small area of land, and several coffee shops.
  - Cluster 1: High household income, large area of land but no coffee shops.
  - Cluster 2: Around average household income, small area of land, and many coffee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

- These characteristics are distinguished as the difference can be seen on each cluster.



Results & Discussion

- 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.
- Cluster of the purple marker is considered as the error - not in the city, near the hill.
- The best answer could be cluster 3 of the yellow markers.
- Other clusters can be considered depending on individual circumstances.



Conclusion

- 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.
- 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.
- 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.