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.

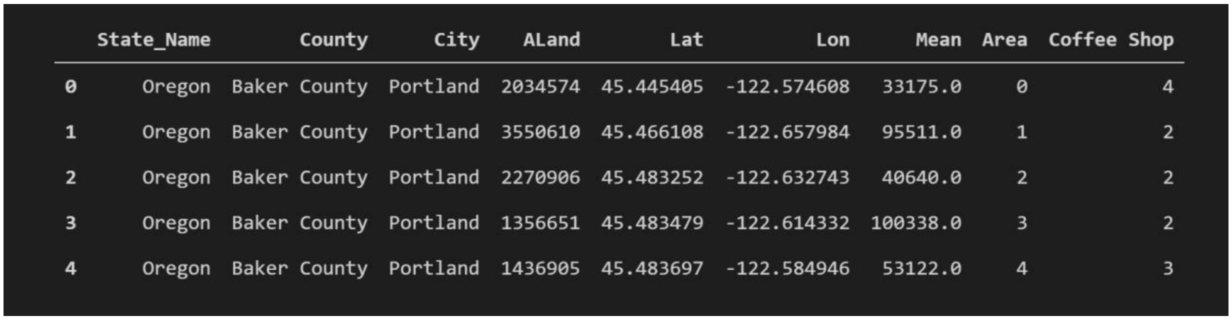
Then, the size of the unit of analysis must be considered. The unit of analysis must be not too small to cause a very small number of venues after getting the result from Foursquare API, and not too big to reach the limit number of venues after getting the result, even if the limit is set more than 100, 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.