Where Are The Optimal Areas In Seattle For A New Restaurant?

1. Introduction

If some one is looking to open a new restaurant, Seattle might be a good choice. Why? According to Google, there are 3,358 restaurants in the Seattle city, a city with around 700000 population. That is about 1 restaurant for every 200 people, meaning there would be plenty of customers for a new restaurant.

But which areas would be optimal for the new restaurant? Intuitively, it should probably be some where already has lots of restaurants, that way when the customers of other dinners pass by, they would realize there is an new one opening, that could bring in the first batch of customers easily.

On the other hand, it might be wise to open a new restaurant in a area has fewer existing restaurants and higher population, so that the potential competition would be lower.

In this project, we will try to find out the answer with machine learning.

2. Data acquisition and cleaning

2.1 Data Source:

The data used in this project are from two sources:

1). Data that contains each zip codes, neighborhoods, and population within each zip code and neighborhoods. These data is downloaded and scraped from public websites.

2). Data that contains the venues within the areas of Seattle city. These data is acquired from Foursquare.com using its API.

2.2Data Cleaning:

1). The downloaded and scraped data sets.

Before the downloaded and scraped data sets are merged into one, some cleaning are done to each one of them.
The first one contains the zip codes, location(latitudes and longitudes), city,

population, population/sq.mile, and population rank in the nation. The data was quite clean, so I drop the unneeded columns(city and ranking).

The second data set was a bit messy, it contains the zip codes, cities, and neighborhoods, and sub regions in the Greater Seattle area, but they were sorted in three different manners, so there are 12 columns in total. And quite a few missing values.

To clean the data set, I kept the first 4 columns, and drop every row that the city name is not Seattle, then the city column is dropped as well. As for the missing neighborhood values, I used the sub Region value to fill out, and then the sub Region column is dropped too.

After the merged of the two data sets, there are still some missing neighborhood values, which is due to the error of the data source, luckily there is only 9 missing ones, and I manage to fill them in manually with google search results.

2). The venues data from Foursquare

The venues data acquired from Foursquare.com are categorical, there are 229 categories in total, such as “ATM”, “Book Store”, “Wine Bar” etc. It is worth mentioning that there are more than one type of restaurants, like “Vegetarian Restaurant”, “Italian Restaurant” and more, because we are interested in finding out the optimal location for general restaurant, thus all categories of restaurants are combined into one “Restaurant” category. And in order for them to be used in later modeling, they are transformed into one hot numerical data using the get_dummy function from Pandas.

For each zip code, there may be more than one row of venue data, to make it one data point for each zip code, the one hot venue data are grouped by each zip code, and using the mean value of individual categories to represent their frequency within the zip code area.

Combined the venues data with the location and population data from the previous step, the data set is ready for modeling.

3. Methodology

3.1Intuitive method.

As mentioned above, there are two potential solution to our question: Where is the

optimal location for a new restaurant?
The intuitive answer is where ever with the highest restaurant density (highest number of restaurants within the certain area), which makes a new restaurant noticeable to the customers of other restaurants, and people like trying out at new restaurant. The top 5 area with high restaurant frequency are as follow:

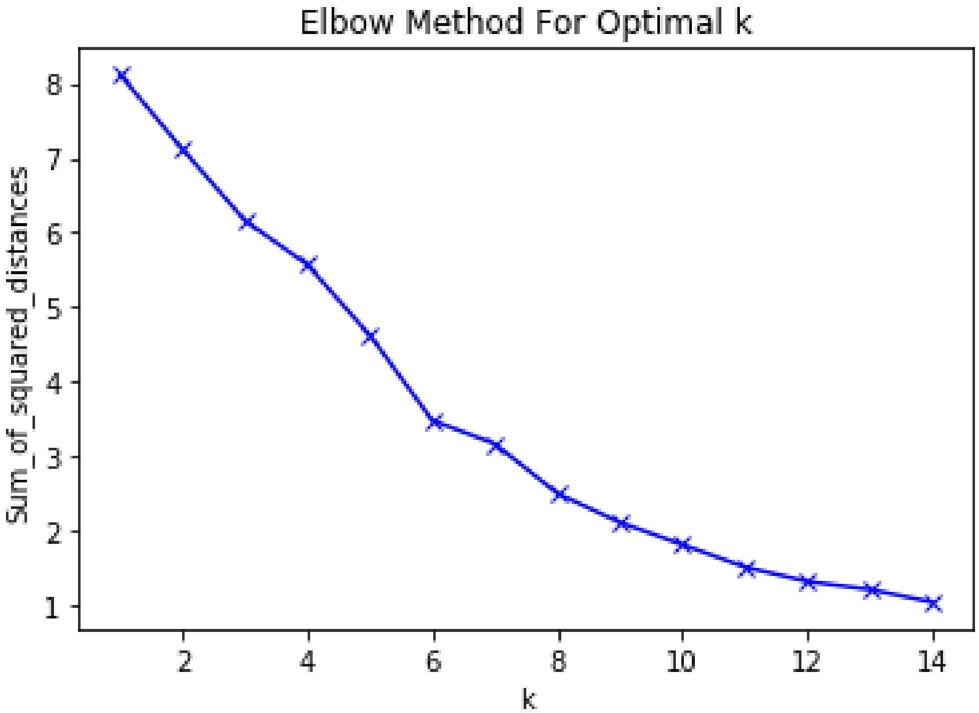
	Zip Code	Restaurant	Population	People / Sq. Mile	Neighborhood	Latitude	Longitude
0	98133	0.400000	42896	6041.66	Bitter Lake	47.740091	-122.342761
1	98118	0.333333	40791	6697.19	Southeast	47.539965	-122.274722
2	98121	0.260000	8558	17894.56	Downtown	47.614743	-122.345855
3	98154	0.260000	1	283.15	Downtown	47.606211	-122.333792
4	98115	0.250000	43567	6603.13	Northeast	47.685766	-122.292178

But there could be some problem with the above location. Places popular with customers tend to have lower land availability and higher cost. Hence we turn to the second method, find out places that are similar to these popular area.

3.2Clustering method:

The idea of this solution is that: if area with high restaurant frequency are good choices but not available, then area similar to them should be good choices as well. So the K Means clustering is used to find out those similar area.

The elbow method is applied to decide the optimal K for the model, and the result is shown below:



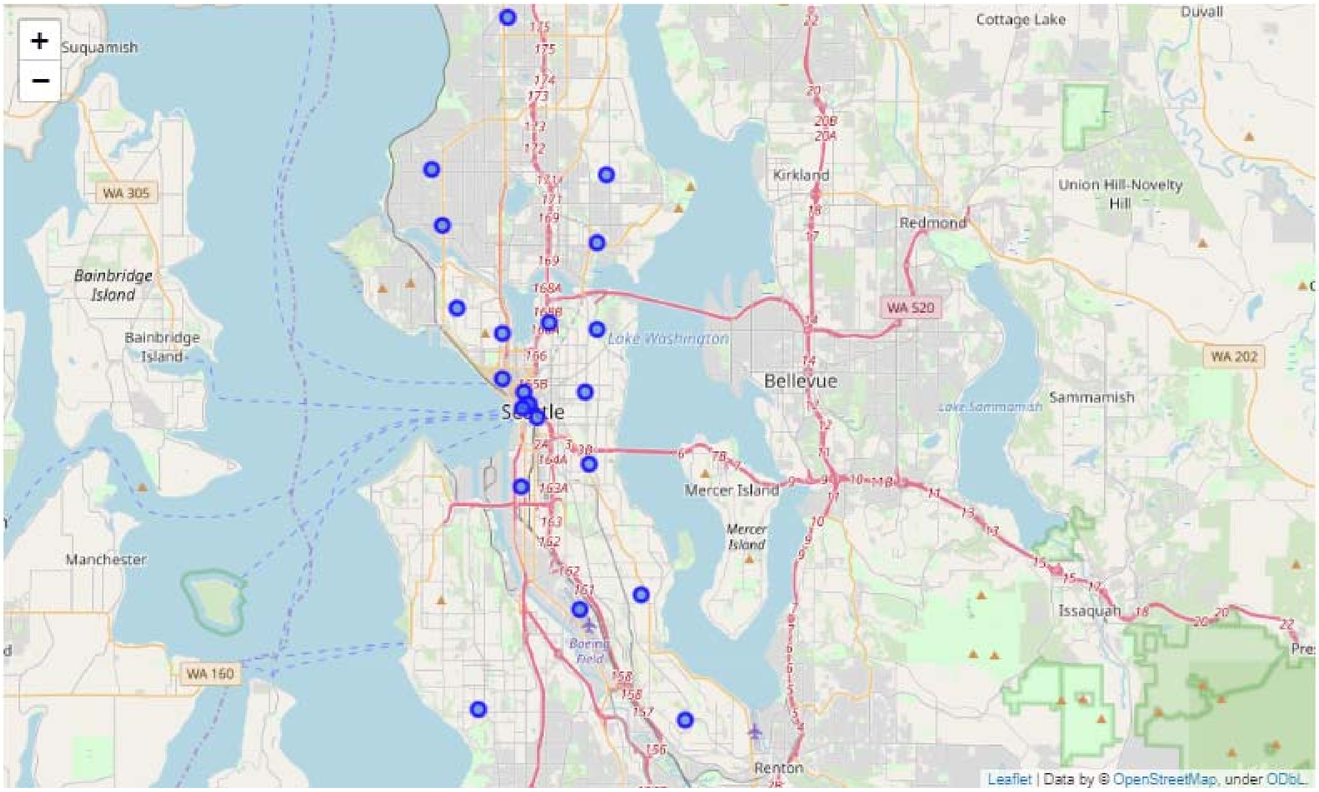
It seems that K=8 is the 'elbow point'.

4. Results:

There are 22 areas that fall in the same category with those restaurant-popular areas, here is some of them in the table:

Zip Code	Population	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
98101	9010	Downtown	Restaurant	Coffee Shop	Hotel
98102	19424	Capitol Hill	Restaurant	Trail	Gym
98104	13095	Downtown	Restaurant	Coffee Shop	Cocktail Bar
98105	38963	Northeast	Restaurant	Clothing Store	Furniture / Home Store

As it is shown in the above table, most of the areas of this cluster has restaurant as their most common venue. It suggests that they are all restaurant-popular areas. And below is all of these areas marked on map:



5. Discussion:

It is not surprising that there are 22 areas in Seattle city that are suitable for opening an restaurant, after all, it is one of the largest city in the U.S, but if a type of restaurant is specified, such as a Mexican restaurant or French restaurant, it should be able to narrow down.