IBM Data Science Capstone Project
The Battle of Neighborhoods

Jing Yang

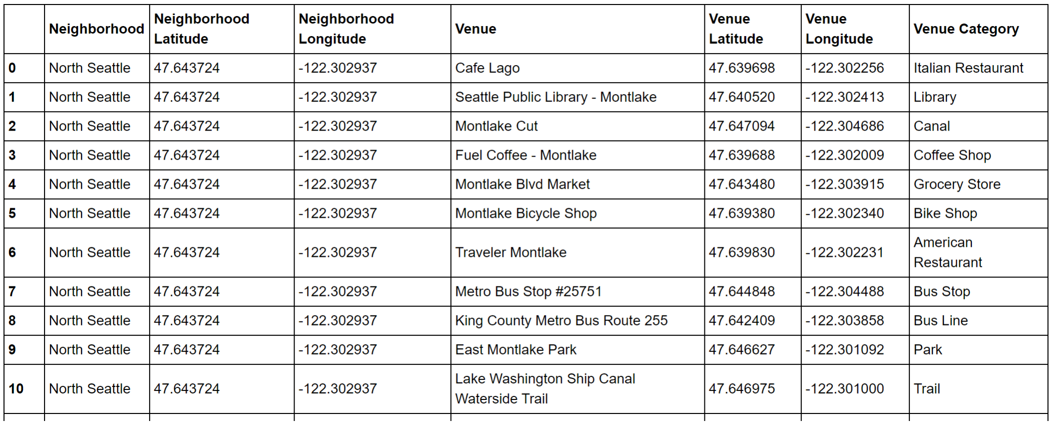
March 2021

1 Introduction

Seattle, as one of the biggest city of the west coast of U.S., is full of business opportunities. As a very divesed city, Seattle's residents are from all over the world. Moreover, the beautiful city always attracts many tourists. Therefore, there are many stakeholders who are interested in openning restaurants in Seattle. In this project, we consider a person who wants to open a Chinese restaurant in Seattle and help him/her to make a better decision of choosing the position. The factors we need to consider are small competition, good transportation, large flow of customers and good surrounding resources. To solve this problem, we segment the neighborhoods in Seattle by implementing the popular clustering technique. The final goal is to provide a list of most promising districts of Seattle to open a Chinese restaurant.

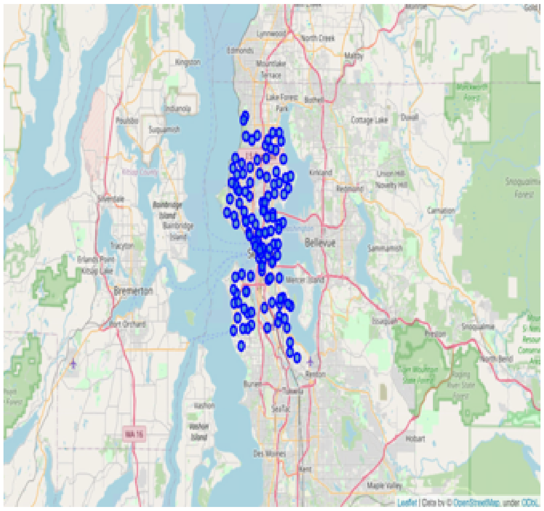
2 Data Preparation

The neighborhood data is scraped from Wikipedia (`https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Seattle`) and corresponding coordinates are obtained using geocoder, and all data of venues in each neighborhoods is obtained from Foursquare API. There are 127 official neighborhoods and 2970 venues in Seattle, which will be the main research objects. The dataframe that summarize the data is as follows:

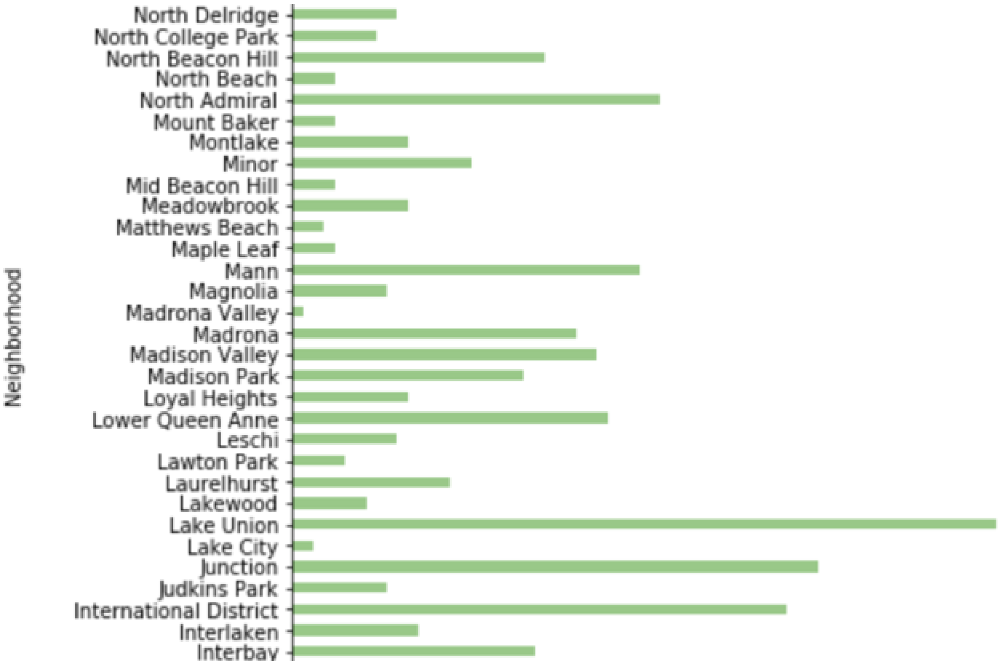


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North Seattle	47.643724	-122.302937	Cafe Lago	47.639698	-122.302256	Italian Restaurant
1	North Seattle	47.643724	-122.302937	Seattle Public Library - Montlake	47.640520	-122.302413	Library
2	North Seattle	47.643724	-122.302937	Montlake Cut	47.647094	-122.304686	Canal
3	North Seattle	47.643724	-122.302937	Fuel Coffee - Montlake	47.639688	-122.302009	Coffee Shop
4	North Seattle	47.643724	-122.302937	Montlake Blvd Market	47.643480	-122.303915	Grocery Store
5	North Seattle	47.643724	-122.302937	Montlake Bicycle Shop	47.639380	-122.302340	Bike Shop
6	North Seattle	47.643724	-122.302937	Traveler Montlake	47.639830	-122.302231	American Restaurant
7	North Seattle	47.643724	-122.302937	Metro Bus Stop #25751	47.644848	-122.304488	Bus Stop
8	North Seattle	47.643724	-122.302937	King County Metro Bus Route 255	47.642409	-122.303858	Bus Line
9	North Seattle	47.643724	-122.302937	East Montlake Park	47.646627	-122.301092	Park
10	North Seattle	47.643724	-122.302937	Lake Washington Ship Canal Waterside Trail	47.646975	-122.301000	Trail

We then get the map information of all the neighborhoods as follows.



The distribution of venues in different neighborhoods is as follows.



3 Methodology

In this section, we provide details of how we acquire the venue data based on Foursquare API platform. Moreover, with a detailed explanation of the feature selection process, we give a tailored clustering model for our problem.

3.1 Data Acquisition

The list of official neighborhoods, on which venue data hinges, is found on Wikipedia. The neighborhood data is scraped using BeautifulSoup and Requests packages with Python. Here we need to clarify that Seattle's data is summairized in a table on Wikipedia, which makes it easier for us to read the data. However, if this is not the case, we need to create the table manually. The corresponding coordinates of each neighborhoods are obtained using geocoder. All venue information is returned by Foursquare API.

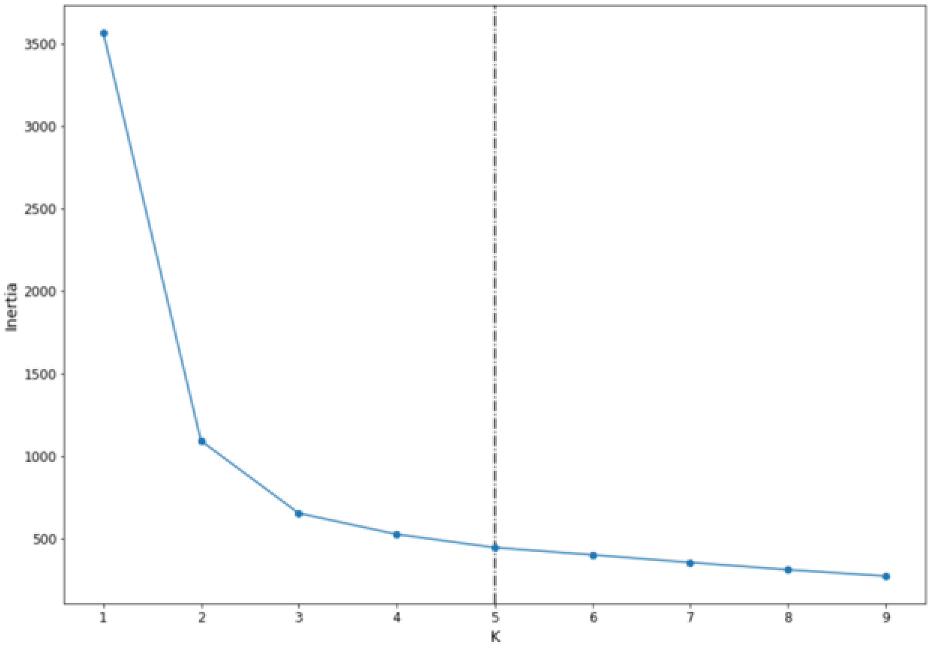
Given that we are interested in the venues information of Seattle, we first take a detailed look at all the venues in different neighborhoods. The Venue Category variable returned by Foursquare API, a categorical feature is transformed to numeric features by using one-hot encoding. The frequency of each venue category in each neighborhood is calculated to be used as key features in neighborhood segmentation.

3.2 Clustering

Recall that our business problem is to choose good positions for a Chinese restaurant, it is more reasonable to focus on some key features that affect the decision making process rather than considering all venues in a neighborhood.

Notice that Seattle is a city full of tourism spots and education-directed city. Moreover, it is obvious that customer flow is very dense in tourism places and shopping malls. Whereas, we need to avoid the places with high competition. Thus in this project, we only consider the number of Chinese restaurants, restaurants, shops, colleges, museums (and related tourism spots) and theaters.

Then, we implement k-means clustering model to segment neighborhoods in Seattle. We explore the best choice of the number of clusters by trying different options from 1 to 10. The input feature of the clustering model is as aforementioned. Using the Elbow method, we find the best number of clusters is 5. Then, we finalize the clustering model. The result from Elbow method is as follows.

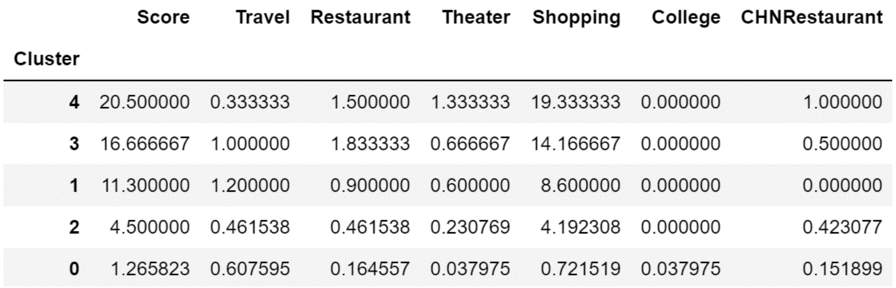


4 Results

Through the clustering model, we label all the neighborhoods. Then, we score the neighborhoods with the following formula:

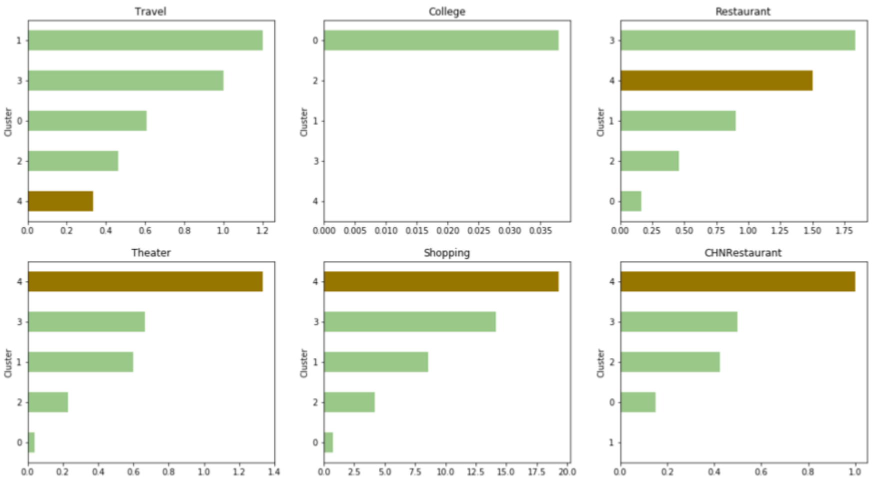
$Score = \#rest + \#tourism + \#college + \#shop + \#theater - 2 * \#Chineserest.$

The sorted clusters are as follows.

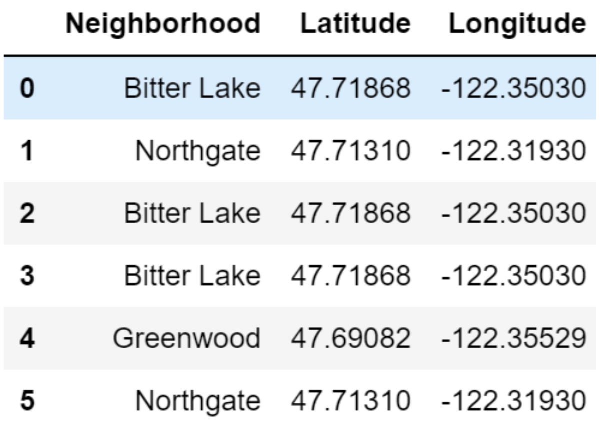


Cluster	Score	Travel	Restaurant	Theater	Shopping	College	CHNRestaurant
4	20.500000	0.333333	1.500000	1.333333	19.333333	0.000000	1.000000
3	16.666667	1.000000	1.833333	0.666667	14.166667	0.000000	0.500000
1	11.300000	1.200000	0.900000	0.600000	8.600000	0.000000	0.000000
2	4.500000	0.461538	0.461538	0.230769	4.192308	0.000000	0.423077
0	1.265823	0.607595	0.164557	0.037975	0.721519	0.037975	0.151899

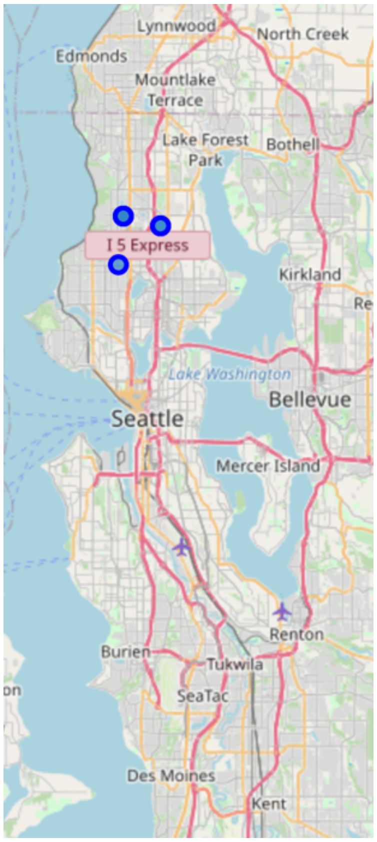
As in the above table, we can see that the highest scored cluster is cluster 4. Then, we check the distribution of selected type of venues in cluster 4. As shown in the following graphs, we can see that cluster 4 has large amount of restaurants, theaters and shops. Although the number of Chinese restaurants is also large, given that Chinese restaurants are not enough in the whole area, the result is reasonable.



We then select 6 highest scored neighborhoods in the highest scored cluster as our potential places for opening a Chinese restaurant.



	Neighborhood	Latitude	Longitude
0	Bitter Lake	47.71868	-122.35030
1	Northgate	47.71310	-122.31930
2	Bitter Lake	47.71868	-122.35030
3	Bitter Lake	47.71868	-122.35030
4	Greenwood	47.69082	-122.35529
5	Northgate	47.71310	-122.31930



5 Conclusion and Discussion

The goal is to find positions of neighborhoods to open a Chinese restaurant in Seattle. We acquire the detailed data of neighborhoods and venues of Seattle. We select 6 key features that are more directly related to a Chinese restaurant's position. Trough k-means clustering with selected features, we segment the neighborhoods into 5 parts. We score the clusters and choose the highest scored cluster.