Capstone Project

Applied Data Science Capstone by IBM

June 2021
IBM Data Science

Nicole Cheng

Table of Contents

1.Introduction

1.1. Background

Seattle is a seaport city on the West Coast of the United States. It is the seat of King County, Washington. With a 2019 population of 753,675, it is the largest city in both the state of Washington and the Pacific Northwest region of North America. The Seattle metropolitan area's population is 3.98 million, making it the 15th-largest in the United States. It is most diversified as a lot of people from different countries live there. It's over 41 million visitors have been travelled to Seattle from all over the world in 2019.

Therefore, it offers many business opportunities for people who wants to operate their own business. Comprehensive consideration of analysis for the business expansion is so much important as it is directly proportional to the cost of the business. The analysis from this report helps people strategically pick the suitable location to open a new restaurant.

1.2. Business Problem

The objective of this capstone project is to analyze and select the best location in Seattle to open a new Chinese restaurant. Using data science methodology and machine learning techniques like clustering. The assumption behind the analysis is that we can use unsupervised machine learning to create clusters of districts that will provide us a list of areas for consideration for the restaurant.

This capstone project aims to provide solutions to answer the business question: if a businessman is looking for a location to open a new Chinese restaurant, where would you recommend that they open it?

1.3. Target Audience

The target audience of this project include people who are interested in opening a new Chinese restaurant or any other types of restaurants in Seattle.

2.Data

To tackle the above-mentioned question, we need to have the dataset that contains:

List of the neighborhoods of Seattle -> this comes from the Wikipedia page
Geo-coordinates of the neighborhoods in Seattle -> this is obtained via geocoder
Top venues of neighborhoods -> Foursquare API is used to collect the venue data

3.Methodology

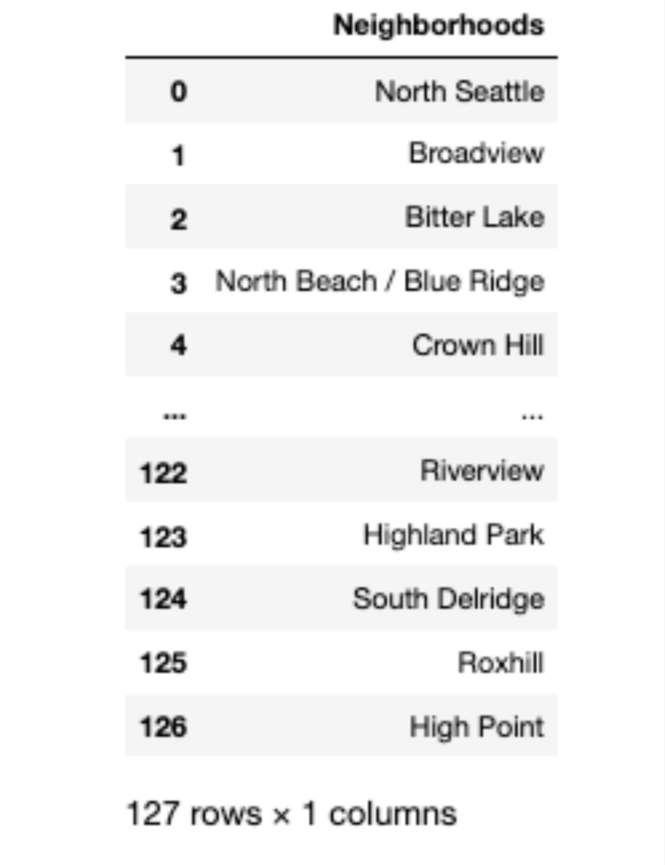
After scraping and exploring the data, we will get latitude and longitude coordinates by using Geocoder. We will use Foursquares API to get venue data. For clustering, K-means method will be applied. To be able to select the optimal number of clusters, the silhouette score will be used. We will also visualize the clusters in a map using Folium.

4.Source

The Wikipedia page https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Seattle is the major source that is being used to obtain all the neighborhoods of Seattle. We then use beautifulsoup4 package, a Python module that helps to scrape information from the web pages to extract all the tables from this Wikipedia page and convert it into a pandas dataframe. Then we use Python's geopy package to obtain the latitude and longitude of all the neighborhoods present in the dataframe.

5.Analysis

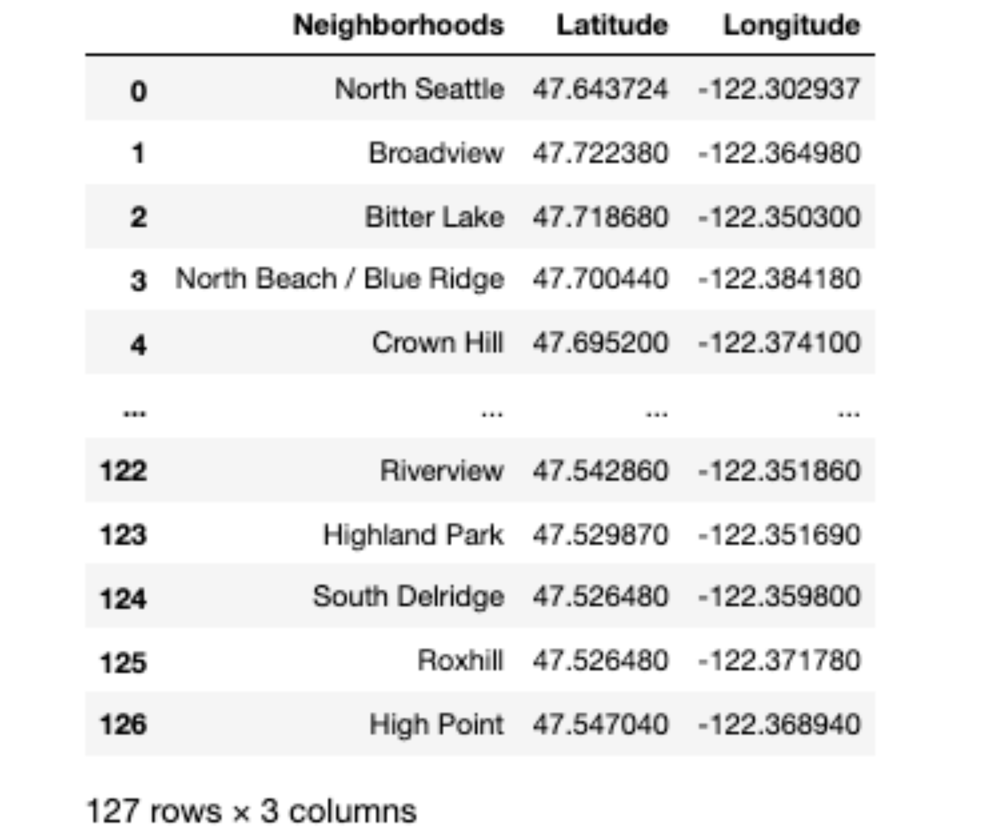
We scrap data from Wikipedia page into a DataFrame. The html table is converted to Pandas DataFrame for cleaning and preprocessing.



	Neighborhoods
0	North Seattle
1	Broadview
2	Bitter Lake
3	North Beach / Blue Ridge
4	Crown Hill
...	...
122	Riverview
123	Highland Park
124	South Delridge
125	Roxhill
126	High Point

127 rows × 1 columns

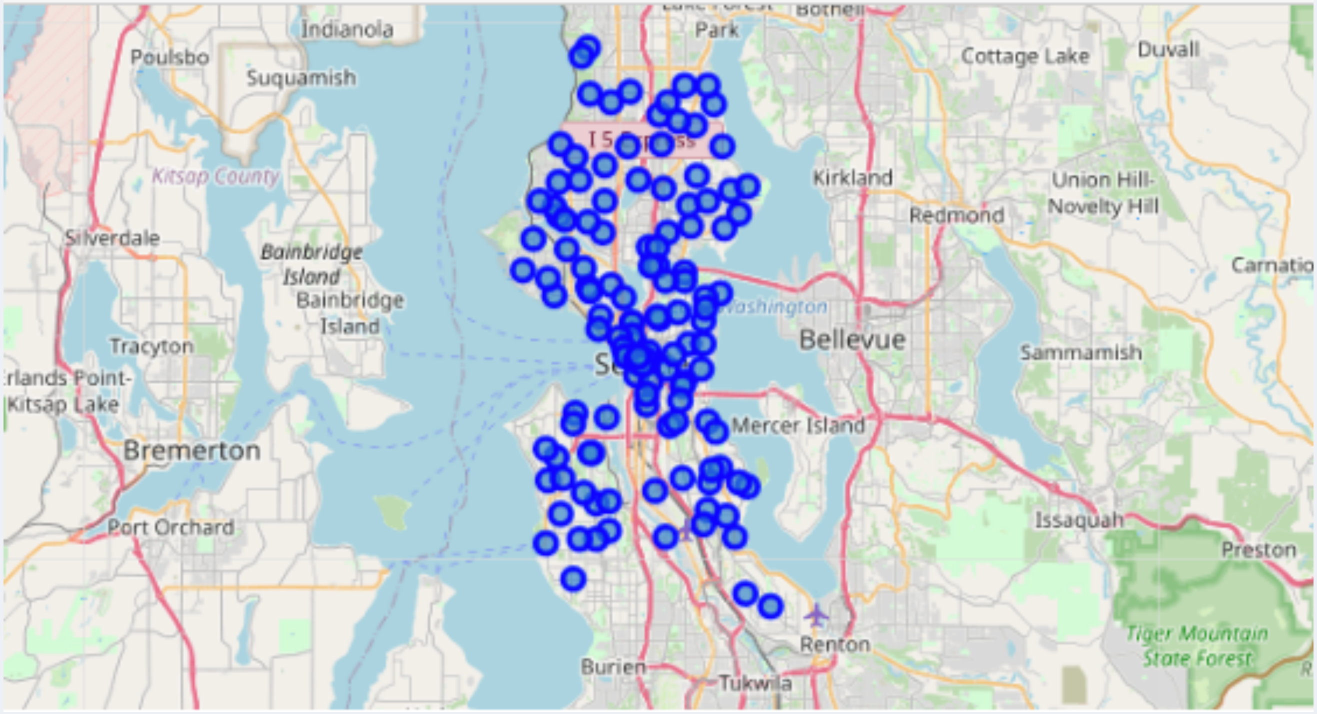
We get latitude and longitude coordinates by using Geocoder. We create a temporary DataFrame to populate the coordinates into Latitude and Longitude. We merge the coordinates into the original DataFrame and save the DataFrame as CSV file.



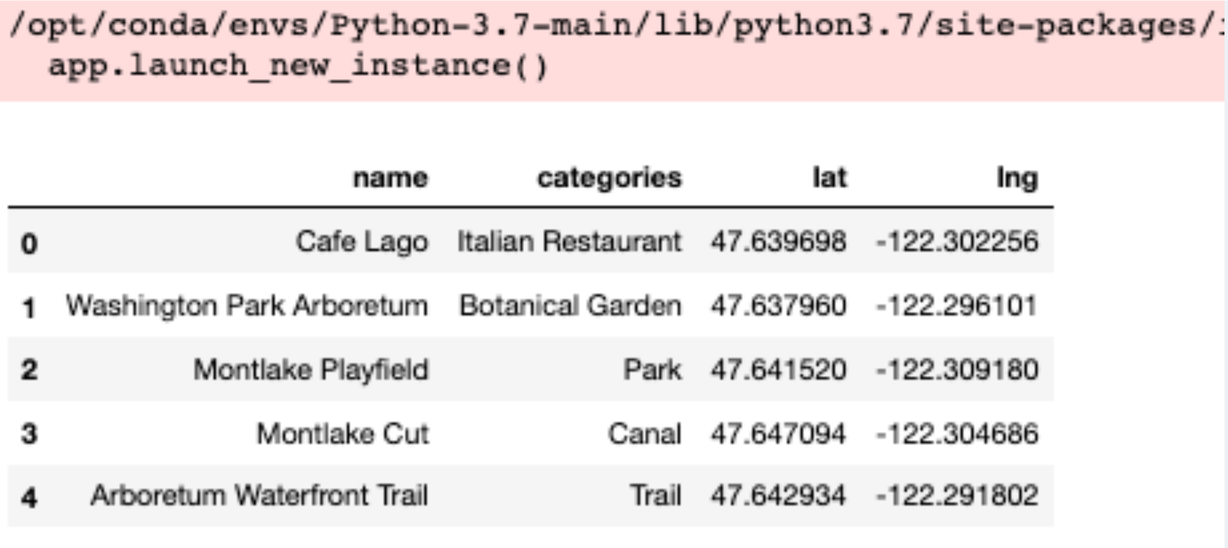
	Neighborhoods	Latitude	Longitude
0	North Seattle	47.643724	-122.302937
1	Broadview	47.722380	-122.364980
2	Bitter Lake	47.718680	-122.350300
3	North Beach / Blue Ridge	47.700440	-122.384180
4	Crown Hill	47.695200	-122.374100
...	...	...	...
122	Riverview	47.542860	-122.351860
123	Highland Park	47.529870	-122.351690
124	South Delridge	47.526480	-122.359800
125	Roxhill	47.526480	-122.371780
126	High Point	47.547040	-122.368940

127 rows × 3 columns

We create a map of Seattle using latitude and longitude values and add markers to the map.



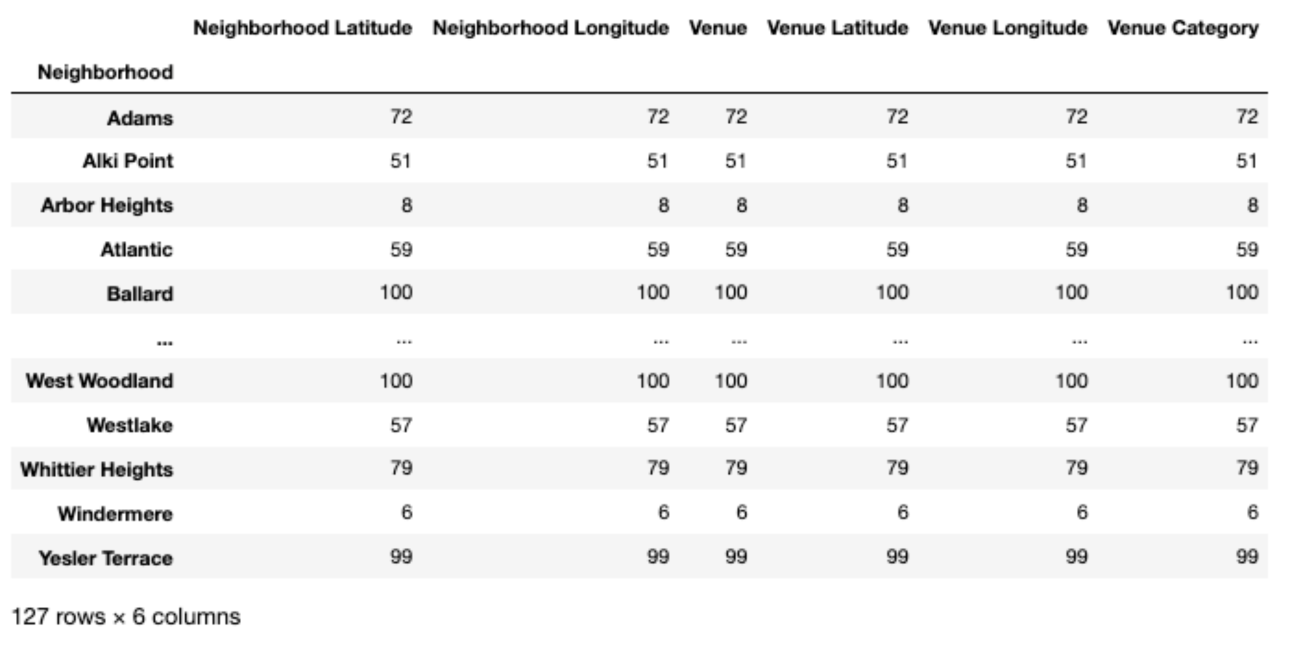
Getting the top 100 venues that are in North Seattle within a radius of 1000 meters. This will be obtained from Foursquare. Defining a function to get categories. Now we are ready to clean the json and structure it into a pandas DataFrame.



```
/opt/conda/envs/Python-3.7-main/lib/python3.7/site-packages/:
  app.launch_new_instance()
```

	name	categories	lat	lng
0	Cafe Lago	Italian Restaurant	47.639698	-122.302256
1	Washington Park Arboretum	Botanical Garden	47.637960	-122.296101
2	Montlake Playfield	Park	47.641520	-122.309180
3	Montlake Cut	Canal	47.647094	-122.304686
4	Arboretum Waterfront Trail	Trail	47.642934	-122.291802

We could explore other venue as well. Now, checking how many venues were collected for other districts as well.

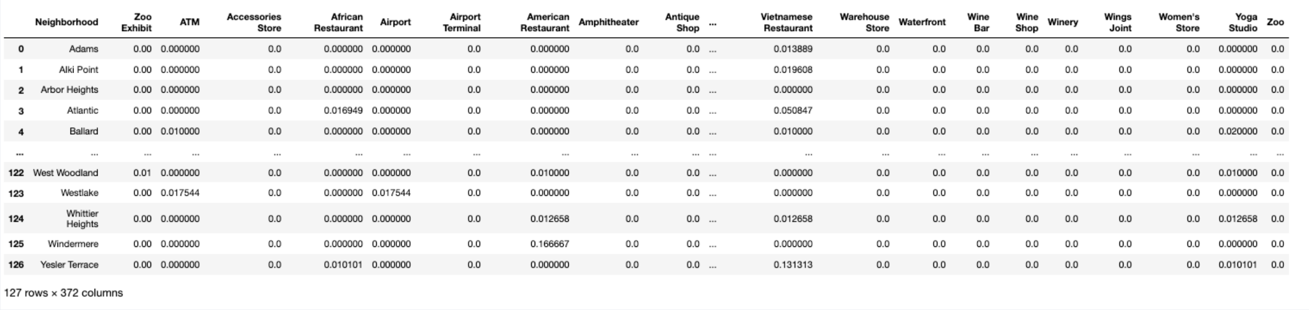


	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Adams	72	72	72	72	72	72
Alki Point	51	51	51	51	51	51
Arbor Heights	8	8	8	8	8	8
Atlantic	59	59	59	59	59	59
Ballard	100	100	100	100	100	100
...	...	...	...	...	...	...
West Woodland	100	100	100	100	100	100
Westlake	57	57	57	57	57	57
Whittier Heights	79	79	79	79	79	79
Windermere	6	6	6	6	6	6
Yesler Terrace	99	99	99	99	99	99

127 rows × 6 columns

There are 372 unique categories.

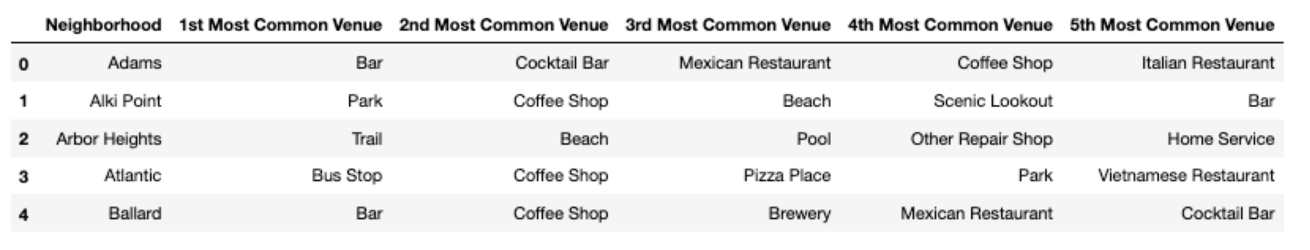
We use one hot encoding and add neighborhood column back to DataFrame. Grouping rows by neighborhood and by taking the mean of the frequency of occurrence of each category.



	Neighborhood	Zoo Exhibit	ATM	Accessories Store	African Restaurant	Airport	Airport Terminal	American Restaurant	Amphitheater	Antique Shop	...	Vietnamese Restaurant	Warehouse Store	Waterfront	Wine Bar	Wine Shop	Winery	Wings Joint	Women's Store	Yoga Studio	Zoo
0	Adams	0.00	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.0	...	0.013889	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
1	Alki Point	0.00	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.0	...	0.019608	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
2	Arbor Heights	0.00	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
3	Atlantic	0.00	0.000000	0.0	0.016949	0.000000	0.0	0.000000	0.0	0.0	...	0.050847	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
4	Ballard	0.00	0.010000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.0	...	0.010000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.020000	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
122	West Woodland	0.01	0.000000	0.0	0.000000	0.000000	0.0	0.010000	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.010000	0.0
123	Westlake	0.00	0.017544	0.0	0.000000	0.017544	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
124	Whittier Heights	0.00	0.000000	0.0	0.000000	0.000000	0.0	0.012658	0.0	0.0	...	0.012658	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.012658	0.0
125	Windermere	0.00	0.000000	0.0	0.000000	0.000000	0.0	0.166667	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
126	Yesler Terrace	0.00	0.000000	0.0	0.010101	0.000000	0.0	0.000000	0.0	0.0	...	0.131313	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.010101	0.0

127 rows × 372 columns

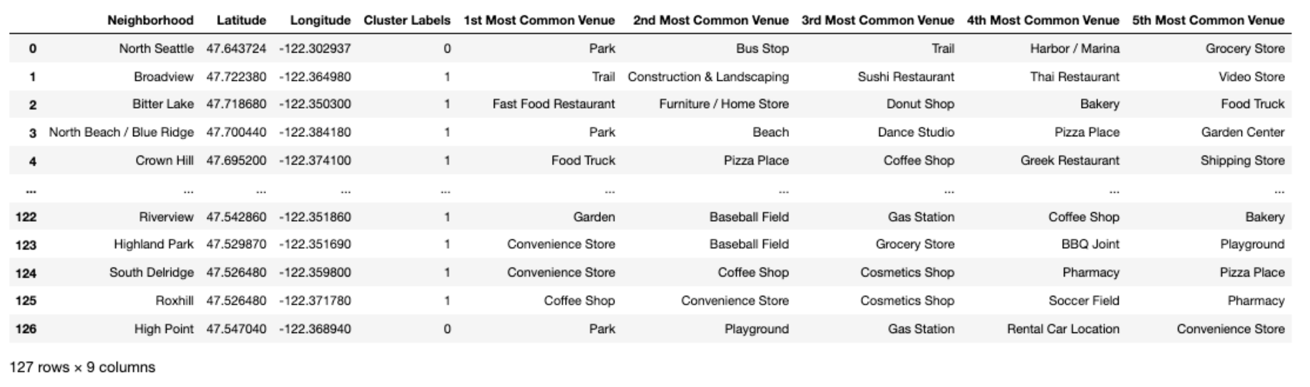
Let's see each neighborhood along with the top 5 most common venues.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Adams	Bar	Cocktail Bar	Mexican Restaurant	Coffee Shop	Italian Restaurant
1	Alki Point	Park	Coffee Shop	Beach	Scenic Lookout	Bar
2	Arbor Heights	Trail	Beach	Pool	Other Repair Shop	Home Service
3	Atlantic	Bus Stop	Coffee Shop	Pizza Place	Park	Vietnamese Restaurant
4	Ballard	Bar	Coffee Shop	Brewery	Mexican Restaurant	Cocktail Bar

For clustering, K-means method will be applied. To be able to select the optimal number of clusters, the silhouette score will be used. First, let's find out the optimal number of clusters. For that, we will create a graph of the silhouette scores.

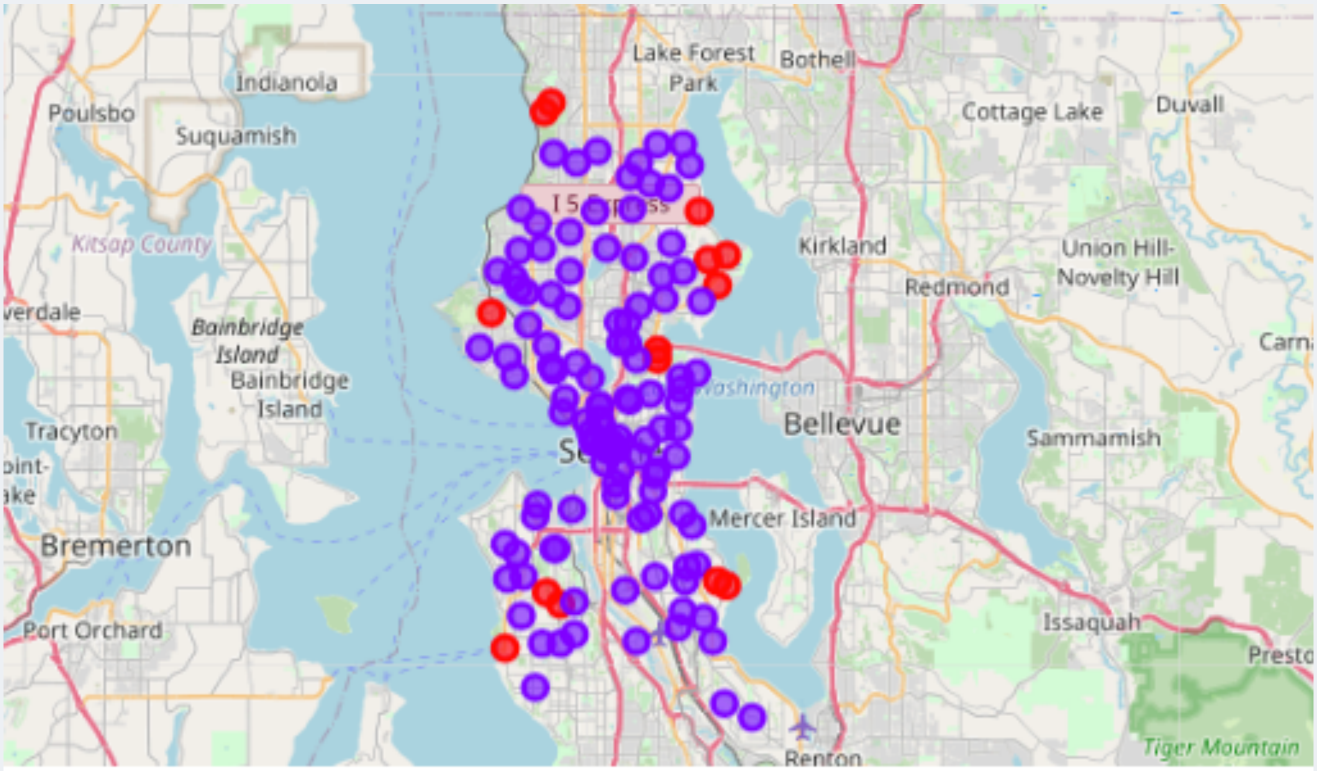
Based on this graph, we can see that the optimal number of clusters is 2.
Now let's run the K-means clustering with the optimal number of clusters, which is 2.



	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	North Seattle	47.643724	-122.302937	0	Park	Bus Stop	Trail	Harbor / Marina	Grocery Store
1	Broadview	47.722380	-122.364980	1	Trail	Construction & Landscaping	Sushi Restaurant	Thai Restaurant	Video Store
2	Bitter Lake	47.718680	-122.350300	1	Fast Food Restaurant	Furniture / Home Store	Donut Shop	Bakery	Food Truck
3	North Beach / Blue Ridge	47.700440	-122.384180	1	Park	Beach	Dance Studio	Pizza Place	Garden Center
4	Crown Hill	47.695200	-122.374100	1	Food Truck	Pizza Place	Coffee Shop	Greek Restaurant	Shipping Store
...	...	...	...	...	...	...	...	...	...
122	Riverview	47.542860	-122.351860	1	Garden	Baseball Field	Gas Station	Coffee Shop	Bakery
123	Highland Park	47.529870	-122.351690	1	Convenience Store	Baseball Field	Grocery Store	BBQ Joint	Playground
124	South Delridge	47.526480	-122.359800	1	Convenience Store	Coffee Shop	Cosmetics Shop	Pharmacy	Pizza Place
125	Roxhill	47.526480	-122.371780	1	Coffee Shop	Convenience Store	Cosmetics Shop	Soccer Field	Pharmacy
126	High Point	47.547040	-122.368940	0	Park	Playground	Gas Station	Rental Car Location	Convenience Store

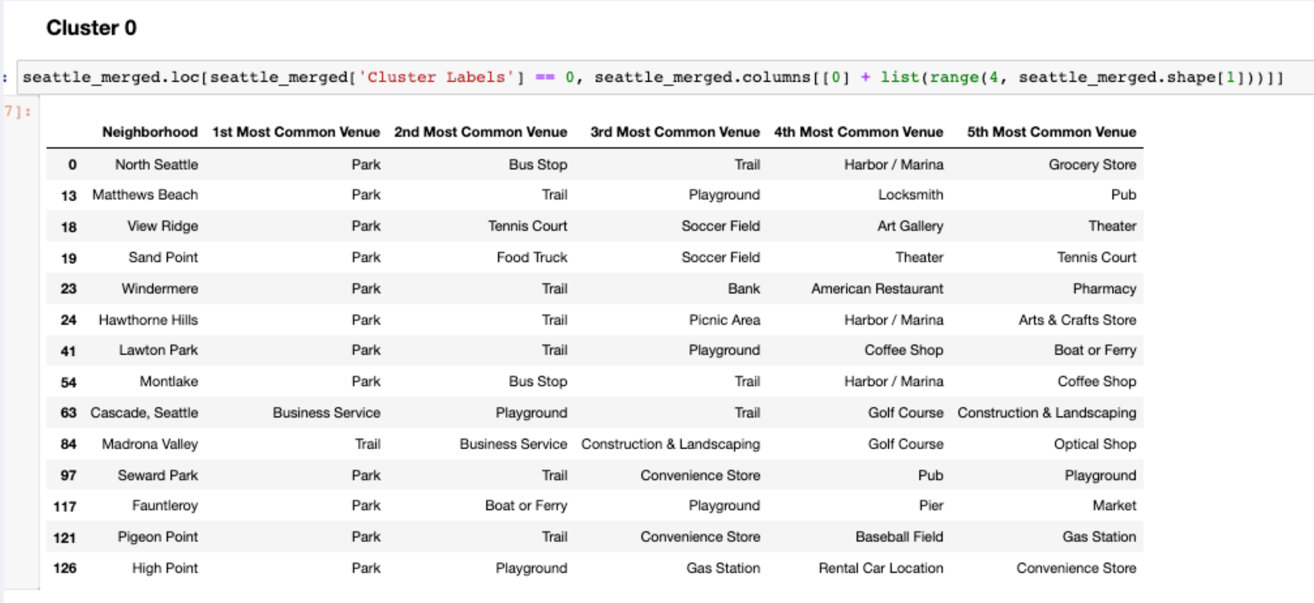
127 rows × 9 columns

We create a map, set the color scheme for the clusters and add markers to the map.



6.Results

Let’s examine the clusters.

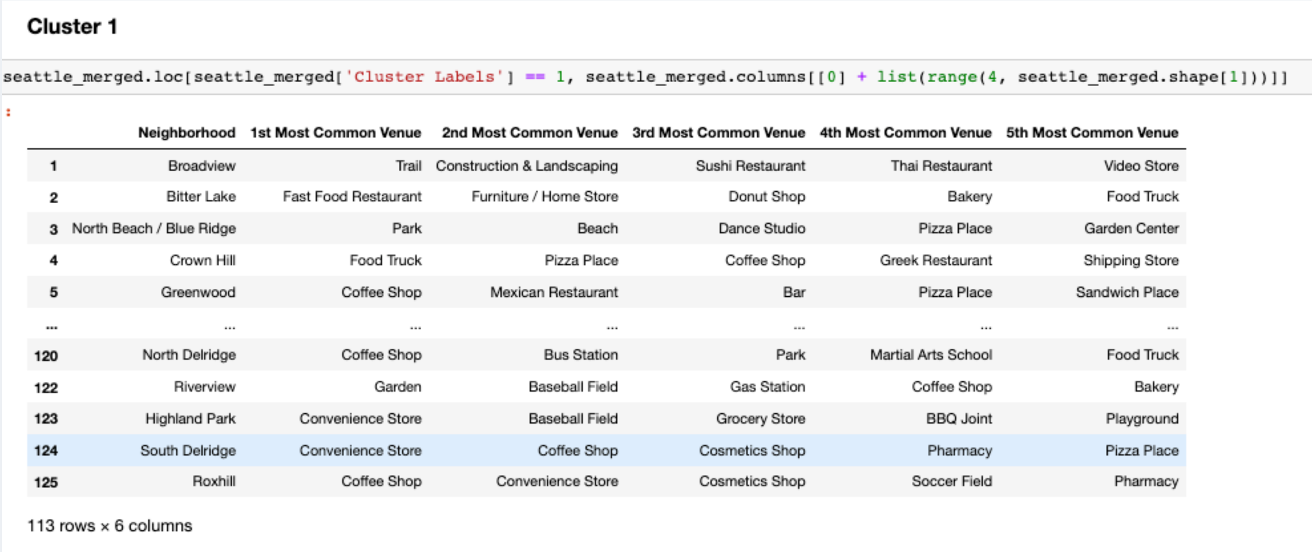


Cluster 0

```
: seattle_merged.loc[seattle_merged['Cluster Labels'] == 0, seattle_merged.columns[[0] + list(range(4, seattle_merged.shape[1]))]]
```

7]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	North Seattle	Park	Bus Stop	Trail	Harbor / Marina	Grocery Store
13	Matthews Beach	Park	Trail	Playground	Locksmith	Pub
18	View Ridge	Park	Tennis Court	Soccer Field	Art Gallery	Theater
19	Sand Point	Park	Food Truck	Soccer Field	Theater	Tennis Court
23	Windermere	Park	Trail	Bank	American Restaurant	Pharmacy
24	Hawthorne Hills	Park	Trail	Picnic Area	Harbor / Marina	Arts & Crafts Store
41	Lawton Park	Park	Trail	Playground	Coffee Shop	Boat or Ferry
54	Montlake	Park	Bus Stop	Trail	Harbor / Marina	Coffee Shop
63	Cascade, Seattle	Business Service	Playground	Trail	Golf Course	Construction & Landscaping
84	Madrona Valley	Trail	Business Service	Construction & Landscaping	Golf Course	Optical Shop
97	Seward Park	Park	Trail	Convenience Store	Pub	Playground
117	Fauntleroy	Park	Boat or Ferry	Playground	Pier	Market
121	Pigeon Point	Park	Trail	Convenience Store	Baseball Field	Gas Station
126	High Point	Park	Playground	Gas Station	Rental Car Location	Convenience Store



Cluster 1

```
seattle_merged.loc[seattle_merged['Cluster Labels'] == 1, seattle_merged.columns[[0] + list(range(4, seattle_merged.shape[1]))]]
```

:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Broadview	Trail	Construction & Landscaping	Sushi Restaurant	Thai Restaurant	Video Store
2	Bitter Lake	Fast Food Restaurant	Furniture / Home Store	Donut Shop	Bakery	Food Truck
3	North Beach / Blue Ridge	Park	Beach	Dance Studio	Pizza Place	Garden Center
4	Crown Hill	Food Truck	Pizza Place	Coffee Shop	Greek Restaurant	Shipping Store
5	Greenwood	Coffee Shop	Mexican Restaurant	Bar	Pizza Place	Sandwich Place
...	...	...	...	...	...	...
120	North Delridge	Coffee Shop	Bus Station	Park	Martial Arts School	Food Truck
122	Riverview	Garden	Baseball Field	Gas Station	Coffee Shop	Bakery
123	Highland Park	Convenience Store	Baseball Field	Grocery Store	BBQ Joint	Playground
124	South Delridge	Convenience Store	Coffee Shop	Cosmetics Shop	Pharmacy	Pizza Place
125	Roxhill	Coffee Shop	Convenience Store	Cosmetics Shop	Soccer Field	Pharmacy

113 rows × 6 columns

Cluster 0 is districts where Park rated at top, but behind that trail, Business Service, playground area is also present. They are mainly outdoor sport place, but not really the vibrant, lively part of the city.

Cluster 1 is the biggest cluster, but this is where we see lots of gastronomy related venues (coffee shop, fast food restaurant, food truck, pizza place, sushi restaurant, Mexican restaurant, etc..).

7.Discussion and Recommendation

By looking at the cluster data, we can see that cluster 1 is the one that we are the most interested in. There are not so many Chinese restaurants at top 5 most common venues. It's a great opportunity for the restaurant owner to consider the districts from cluster 1 as a potential location for the new Chinese restaurant.

These are the districts where gastronomy is well represented. All kinds of delicacies are concentrated in these areas. Therefore, it will attract more and more people to come here.

8.Conclusion

The project was analyzed based on the toolset of data science and relied on the use of Python and Python libraries including Pandas, Scikit, Folium.

Data was collected from a different type of sources and in different formats. For analysis, machine learning technique was applied. The output of the analysis provided a thorough base for the recommendation for the business problem in question.