Capstone Project — The Battle of Neighborhoods in Erbil: Restaurants

Suzanne Zakaria Hussein

September 2, 2021

As part of the IBM Data Science Professional certification [1], you will find an overview of my final project in this post.

As prepared for the task, we proceed to the problem's description, data preparation, and the final step-by-step analysis. Detailed codes are provided on Github, a link can be found at the end of the post.

1- Introduction/Business problem

Erbil, also called Hawler, is the capital and most populated city in the Kurdistan Region of Iraq. It has around 1.5 million inhabitants. [2]

The city of Erbil is a historical, tourist and commercial city, and tourists and traders from various cities of Iraq and neighboring countries flock to it annually, and these numbers are gradually increasing.

It is also famous for its various restaurants of oriental, western and other foods. In this exercise, we will support different visitors to list and visualize Erbil districts that fit their needs in terms of culinary/ food venues.

2- Description of the data

For the data, I used two resources.
First Wikipedia to get all districts of Erbil city Erbil Governorate - Wikipedia [3]. And second, Foursquare API to get the most common venues of Erbil restaurants.
Food Erbil (foursquare.com)[4].

3- Methodology

3.1 Data Preparation

My main dataset, By creating dataframe for Districts in Erbil city which imported from Wikipedia page and cleaning data using pandas library as shown in the following figure.

	District	Population (2009)[9]	Number of households
0	Erbil	792981	152899
1	Dashty Hawler	186346	34264
2	Makhmur	173801	30678
3	Soran	154945	27707
4	Shaqlawa	124628	23420
5	Koya	95246	18727
6	Kabat	93442	16015
7	Mergasor	44661	9711
8	Choman	23730	4749
9	Rawanduz	21280	4235

Figure 1: Dataframe of Erbil Districts

We get 10 districts. Then by using geopy.geocoders library get latitude and longitude for each district.

	District	Latitude	Longitude
0	Erbil	36.191162	44.009465
1	Dashty Hawler	36.207872	44.133571
2	Makhmur	35.776121	43.580317
3	Soran	39.496457	43.517432
4	Shaqlawa	36.411803	44.309316
5	Koya	26.373185	91.565149
6	Kabat	-8.300700	114.315896
7	Mergasor	36.979367	44.214966
8	Choman	35.971203	45.578947
9	Rawanduz	36.614596	44.526519

Figure 2: Dataframe with latitude and longitude

I used python **folium** library to visualize geographic details of Erbil and its boroughs and I created a map of Erbil with boroughs superimposed on top. I used latitude and longitude values to get the visual as below:

Figure 3: The figure for latitude longitude before fixing

As shown on top figure they gives wrong latitude and longitude for two districts (Soran and Koya) because of duplicate name with other districts in Turkey and India Countries. I get the right latitude and longitude for them by Latlong [5] and revisualize geographic details of Erbil city.

	District	Latitude	Longitude
0	Erbil	36.191162	44.009465
1	Dashty Hawler	36.207872	44.133571
2	Makhmur	35.776121	43.580317
3	Soran	36.654700	44.537800
4	Shaqlawa	36.411803	44.309316
5	Koya	36.079072	44.634040
6	Khabat	36.276362	43.668284
7	Mergasor	36.979367	44.214966
8	Choman	35.971203	45.578947
9	Rawanduz	36.614596	44.526519

Figure 4:Dataframe with right latitude and longitude

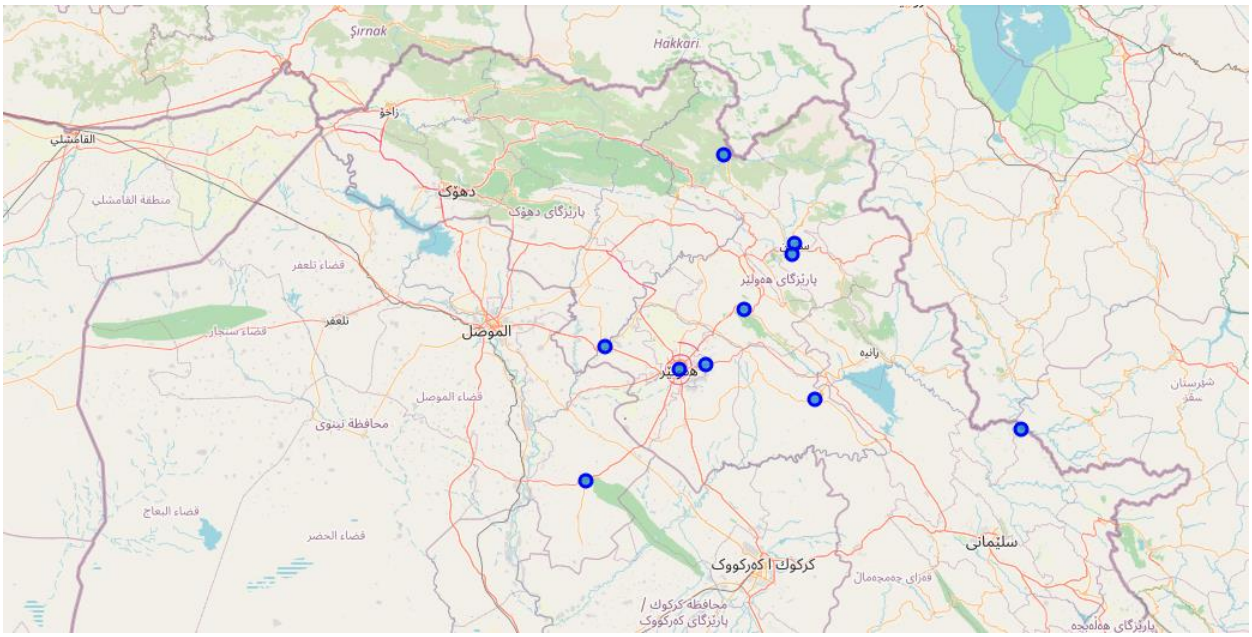

Figure 5:The figure for latitude longitude after fixing

3.2 Exploratory Data Analysis

Firstly, I will utilize exploratory data analysis(EDA) to reveal covered-up properties of information and give valuable bits of knowledge to the peruser, both future travelers and financial specialists.

By utilizing the Foursquare API and get the best 100 settings that are in Erbil inside a radius of 500 meters.

```
Restaurant                  11
Middle Eastern Restaurant    6
Asian Restaurant             1
Turkish Restaurant           1
Mediterranean Restaurant     1
Italian Restaurant           1
Japanese Restaurant          1
Kebab Restaurant             1
Persian Restaurant           1
Afghan Restaurant            1
Fast Food Restaurant         1
Arepa Restaurant             1
Doner Restaurant             1
Seafood Restaurant           1
Name: Venue Category, dtype: int64
```

```
] print('There are {} unique categories.'.format(len(Venues_only_restaurant['Venue Category'].unique())))

There are 14 unique categories.
```

Figure 6: Restaurants Categories in Erbil

We take note that 14 unique venue categories were returned by Foursquare and Regular Local Restaurants within the best of the list as we can see above.

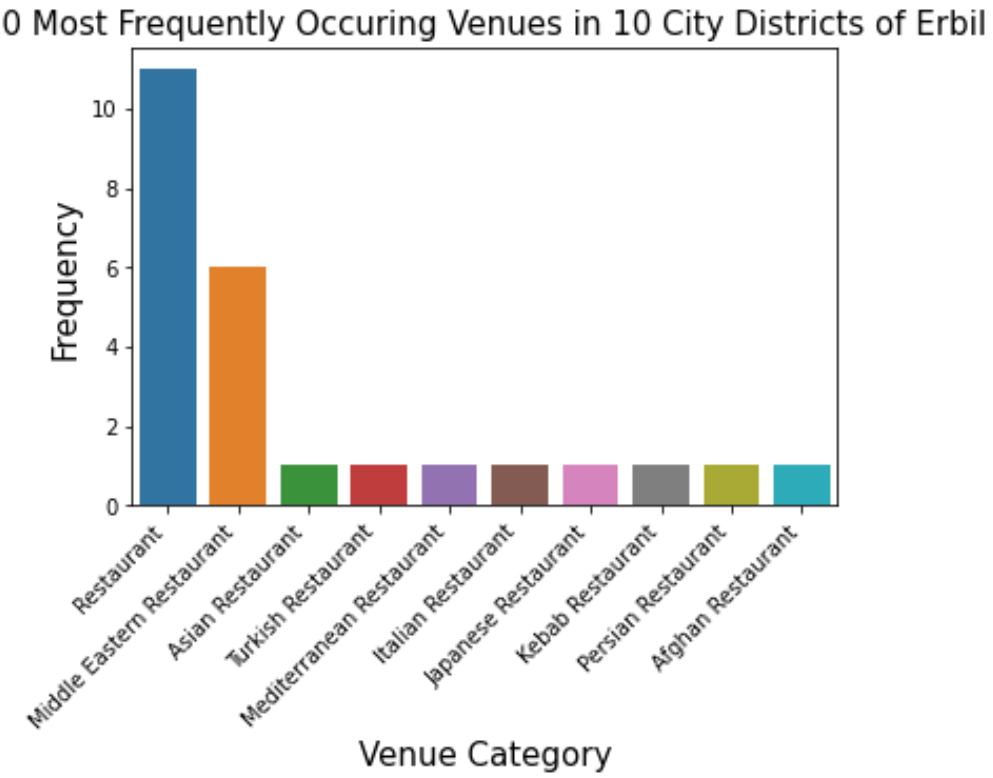

Figure 7: 10 most frequently occurring venues in 10 city districts of Erbil

To find clusters of restaurant types for each city districts, I first transformed the data frame with the restaurant venues, associated to city districts, by one-hot encoding (0/1), as seen in the picture below.

	Neighborhood	Afghan Restaurant	Arepa Restaurant	Asian Restaurant	Doner Restaurant	Fast Food Restaurant	Italian Restaurant	Japanese Restaurant	Kebab Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Persian Restaurant
1	Erbil	0	0	0	0	0	0	0	1	0	0	0
2	Erbil	0	0	0	0	0	0	0	0	0	0	0
3	Erbil	0	0	0	0	0	0	0	0	0	1	0
4	Erbil	0	0	1	0	0	0	0	0	0	0	0
5	Erbil	0	0	0	0	0	0	0	0	0	1	0

Figure 8:one-hot encoding

Then I grouped the one hot encoding dataframe by using Grouping by neighborhoods and showing the mean of the frequency of occurrence for each category of restaurants.

	Neighborhood	Afghan Restaurant	Arepa Restaurant	Asian Restaurant	Doner Restaurant	Fast Food Restaurant	Italian Restaurant	Japanese Restaurant	Kebab Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Persian Restaurant	Restaurant
0	Dashty Hawler	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.0	1.00000
1	Erbil	0.043478	0.043478	0.043478	0.043478	0.043478	0.043478	0.043478	0.043478	0.043478	0.26087	0.0	0.26087
2	Shaqlawa	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.0	1.00000
3	Soran	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.5	0.50000

Figure 9: Groupby one-hot encoding Dataframe

And results of each neighborhood along with the top 10 most common venues:

```
----Dashty Hawler----
                      venue  freq
0                Restaurant   1.0
1          Afghan Restaurant   0.0
2           Arepa Restaurant   0.0
3           Asian Restaurant   0.0
4           Doner Restaurant   0.0
5       Fast Food Restaurant   0.0
6         Italian Restaurant   0.0
7        Japanese Restaurant   0.0
8           Kebab Restaurant   0.0
9  Mediterranean Restaurant   0.0

----Shaqlawa----
                      venue  freq
0                Restaurant   1.0
1          Afghan Restaurant   0.0
2           Arepa Restaurant   0.0
3           Asian Restaurant   0.0
4           Doner Restaurant   0.0
5       Fast Food Restaurant   0.0
6         Italian Restaurant   0.0
7        Japanese Restaurant   0.0
8           Kebab Restaurant   0.0
9  Mediterranean Restaurant   0.0

----Erbil----
                       venue  freq
0  Middle Eastern Restaurant  0.26
1                 Restaurant  0.26
2          Afghan Restaurant  0.04
3           Arepa Restaurant  0.04
4           Asian Restaurant  0.04
5           Doner Restaurant  0.04
6       Fast Food Restaurant  0.04
7         Italian Restaurant  0.04
8        Japanese Restaurant  0.04
9           Kebab Restaurant  0.04

----Soran----
                      venue  freq
0    Persian Restaurant   0.5
1              Restaurant   0.5
2     Afghan Restaurant   0.0
3      Arepa Restaurant   0.0
4      Asian Restaurant   0.0
5      Doner Restaurant   0.0
6  Fast Food Restaurant   0.0
7    Italian Restaurant   0.0
8   Japanese Restaurant   0.0
9      Kebab Restaurant   0.0
```

Figure 10: The top 10 most common venues for each neighborhood

We will use prescriptive analytics to assist a traveler in choosing an area to go for a restaurant. In addition, I will utilize clustering (KMeans). Finally, we attempt to cluster these 10 districts based on the venue categories and using K-Means clustering. Our expectation would be based on the similitudes of venue categories; these districts will be clustered.

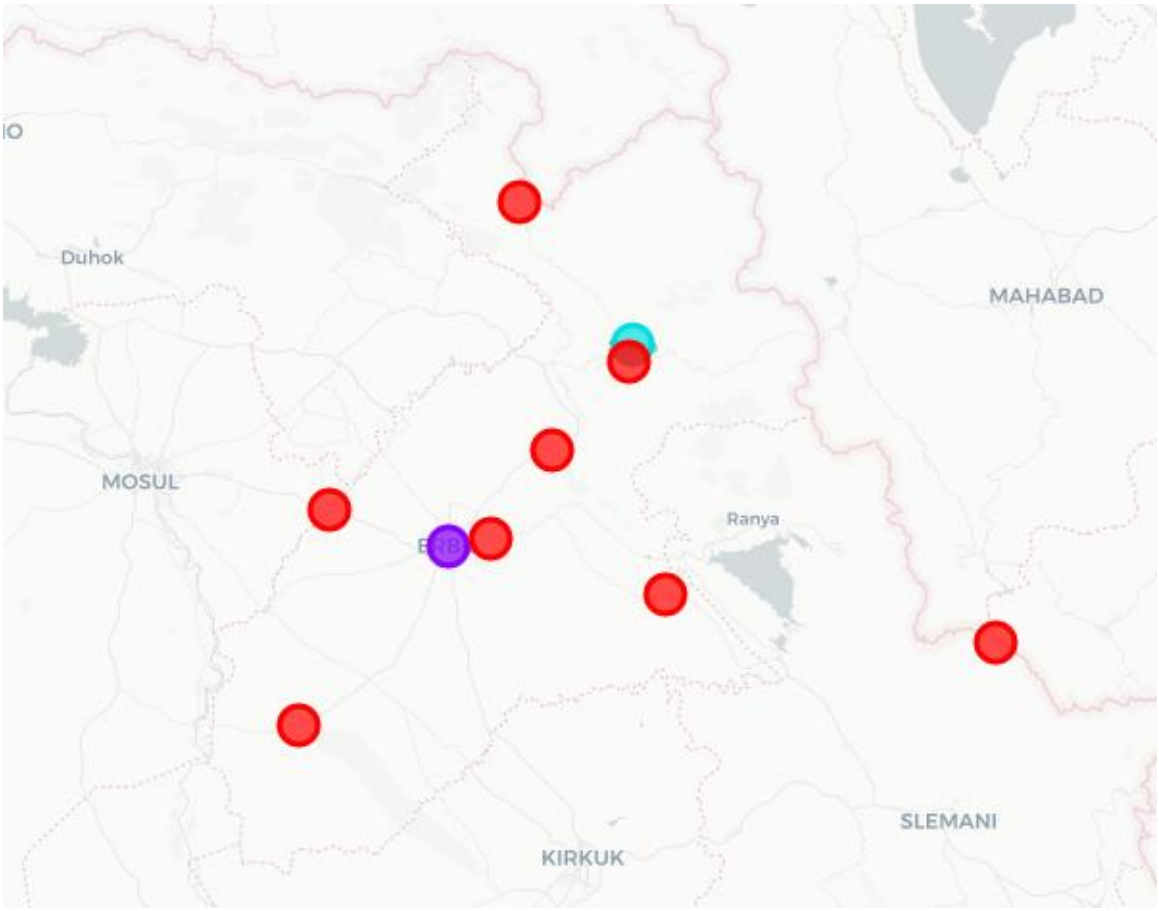

Figure 11: Three clusters

4- Results & discussion

We looked at the Restaurants in Erbil and were able to discover some interesting insights that might be useful to travelers and people with business interests. We founding out that Regular Local restaurants top the charts of most common venues in the 10 districts. And Erbil district has a maximum number of restaurants.

The clustering is completely based on the most common venues obtained from Foursquare data.

In our analysis, we have ignored other factors like the distance of the venues from closest stations, range of prices of restaurants, and so on, since we don't have such data and it would be difficult to farm it for a small exploratory study like ours. Hence, our analysis only helps travelers to get an overview of Restaurants distribution by categories in the 10 major districts of Erbil.

5- Conclusion

With this exercise the visitors to Erbil City can have a review on most restaurant types, thus, they can taste different types of desired cuisines.

6- References

[1]- https://www.coursera.org/professional-certificates/ibm-data-science

[2]- Erbil History | Salahaddin University-Erbil (su.edu.krd)

[3]- Erbil Governorate - Wikipedia

[4]- https://foursquare.com/explore

[5]- https://www.latlong.net/