Presented by Meshrif Alruily

Clustering the neighborhoods in the city of Almadinah using K-means

Introduction

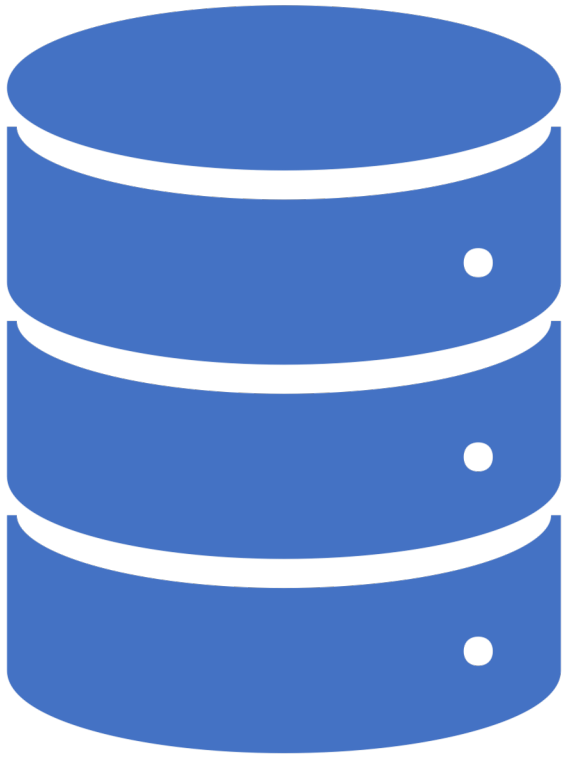
- Almadinah city in Saudi Arabia has the second holey mosque for Muslims around the world. Because of that it is visited during a whole year. Therefore, it is good for visitors to know the neighborhoods of the city and its characteristics. This will lead them to book their hotels and stay during their visits in their favorite neighborhoods. As a result, in this project I am aiming to cluster the neighborhoods in the city of Almadinah using k-means clustering technique.

Problem Description

The goal of this project is to cluster the neighborhoods in the city of Almadinah using k-means clustering technique based on similarity venues categories located in neighborhoods being studied.

Objective

- The main goal of this project is to cluster the neighborhoods in the city of Almadinah in Saudi Arabia. For achieving that the neighborhoods of Almadinah city will be studied and analyzed first. Then, K-means is implemented. This will assist visitors of the city to know where to stay in the city during their visit.



Data Description & Preparation

To consider the objective stated, we can list the below data sources used for the analysis.

a) KSA Neighborhood Data

- The following github project (https://github.com/homaily/Saudi-Arabia-Regions-Cities-and-Districts/tree/master/json) contains Saudi Arabia Regions, Cities and Districts. The data is public data collected from https://maps.address.gov.sa/ as ,all coordinates in (Lat, Lon) This data divided into three parts:
  - regions.json
  - cities.json
  - districts.json

Sample Data

	region_id	name_ar	name_en	population	capital_city_id	Latitude	Longitude
0	1	منطقة الرياض	Riyadh	6777146	3	24.70000	46.73333
1	2	منطقة مكة المكرمة	Makkah	6915006	6	21.42718	39.84349
2	3	منطقة المدينة المنوره	Madinah	1777933	14	24.47058	39.60781
3	4	منطقة القصيم	Qassim	1215858	11	26.33034	43.97436
4	5	المنطقة الشرقية	Eastern Province	4105780	13	26.44199	50.10920

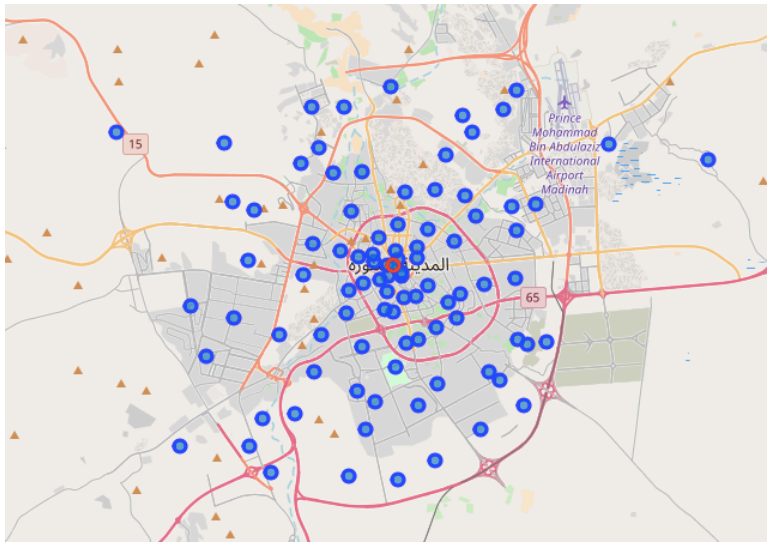
	city_id	region_id	name_ar	name_en	Latitude	Longitude
0	1	7	تبوك	Tabuk	28.41464	36.53387
1	2	7	نعمي	Na'mi	28.30508	35.74931
2	3	1	الرياض	Riyadh	24.70000	46.73333
3	4	7	حميظ	Humayt	28.65152	35.38013
4	5	2	الطائف	At Taif	21.26848	40.41667

	district_id	city_id	region_id	name_ar	Neighborhood	Latitude	Longitude	boundaries
0	10100003001	3	1	حي العمل	Al Amal Dist.	24.644966	46.723598	[[24.64900056, 46.7248791], [24.64742521, 46.7...
1	10100003002	3	1	حي النموذجية	Al Namudhajiyah Dist.	24.655615	46.696523	[[24.65018372, 46.70227584], [24.64939455, 46....
2	10100003003	3	1	حي الجرادية	Al Jarradiyah Dist.	24.618815	46.696110	[[24.61729504, 46.70655201], [24.61727163, 46....
3	10100003004	3	1	حي الصناعية	Al Sinaiyah Dist.	24.646447	46.741675	[[24.64061956, 46.75999871], [24.6342299, 46.7...
4	10100003005	3	1	حي منفوحة الجديده	Manfuha Al Jadidah Dist.	24.613354	46.716463	[[24.61343234, 46.72718798], [24.61090205, 46....

Data Description & Preparation

b) Coordinate data for each Neighborhood in KSA Cities:

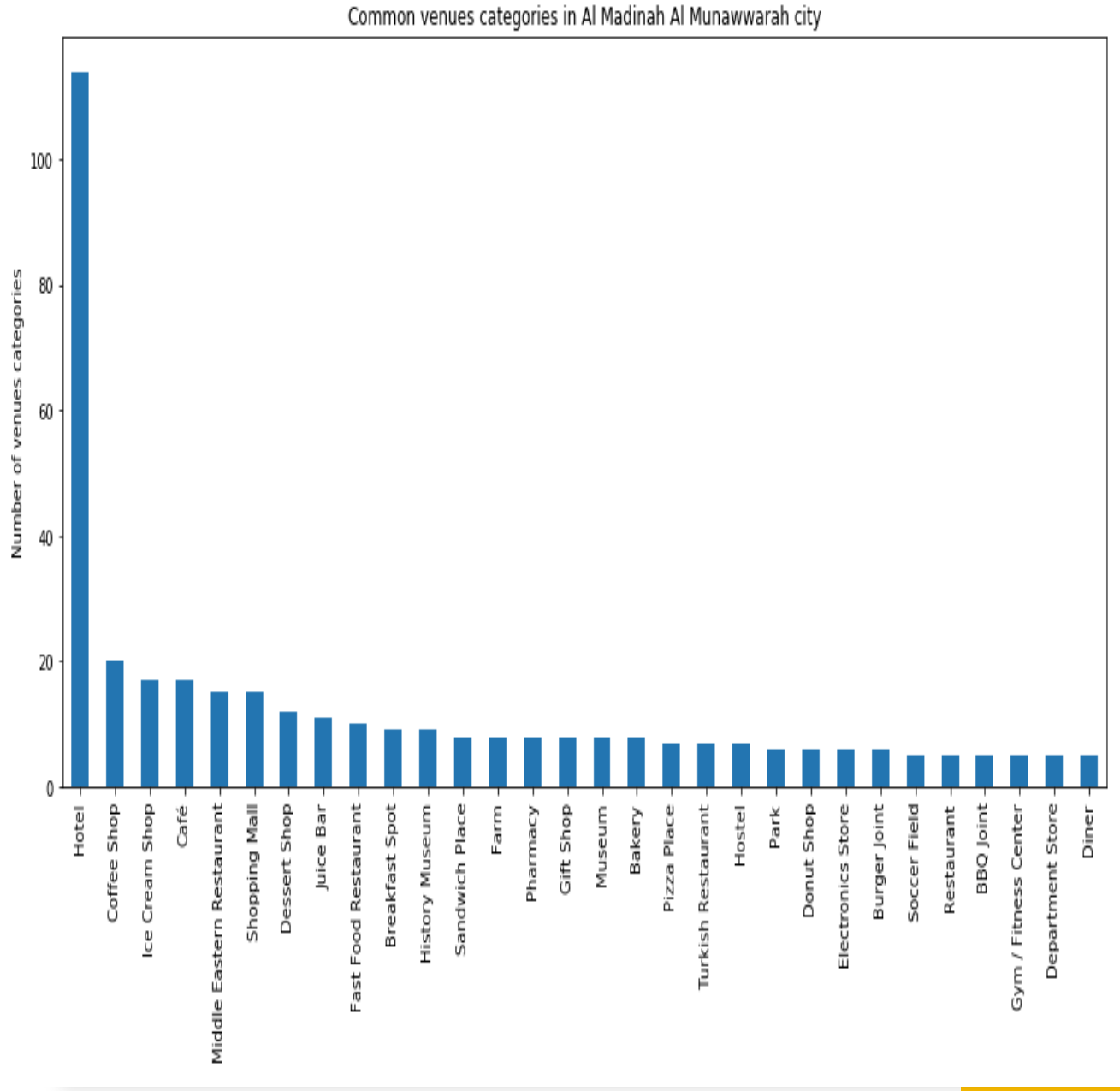
- All the venues in each neighborhood in the KSA cities will be collected from Fousquare using Fousquare API By using this api we will get all the venues in each neighborhood. We can filter these venues to get only two cities will be anlyzed.

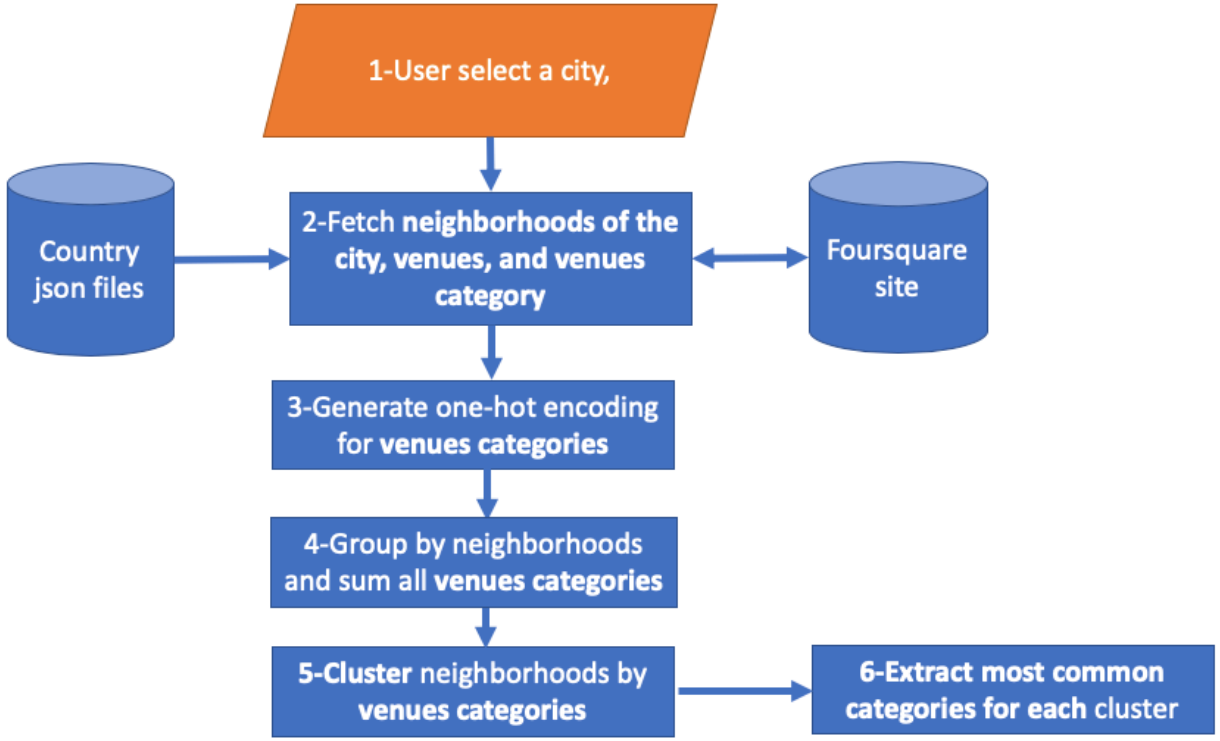


Data Analysis

It can be seen most common categories in the city are:

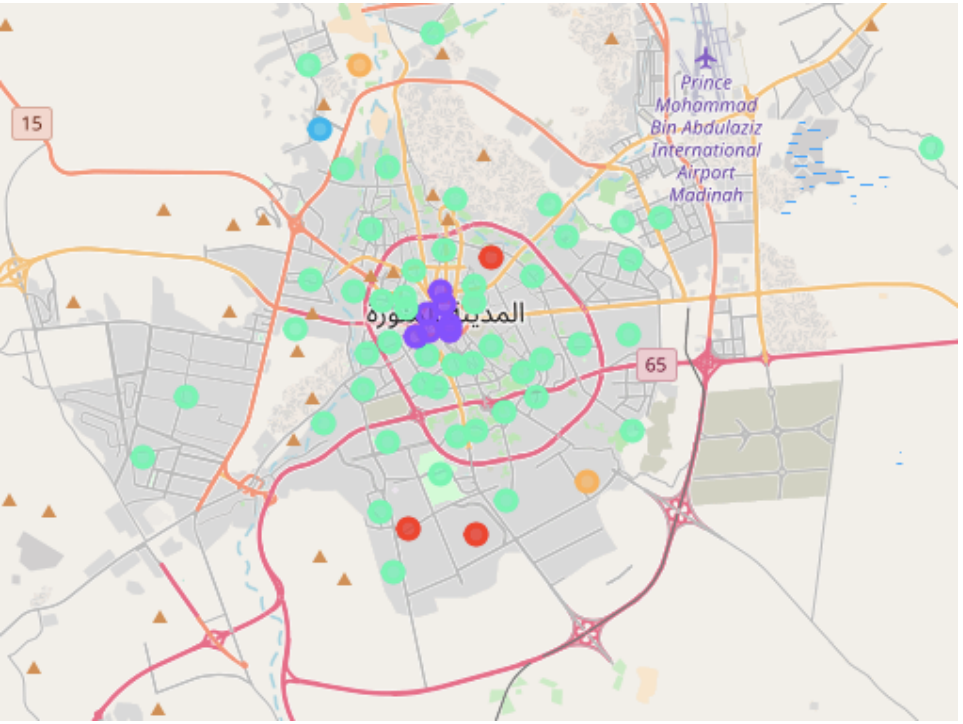
- - Hotels
- - Coffee shops
- - Ice cream shops
- - Café
- - Shopping mall





As shown, there are 6 steps for achieving clustering.

Methodology



Results

Cluster 1

```
[8]: almadinah_merged.loc[almadinah_merged['Cluster Labels'] == 0 , almadinah_merged.columns[[1] + list(range(5,
```

[8]:

Neighborhood	city_id	Longitude	boundaries	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
14_Ar Ranuna Dist.	14	39.597323	[[24.41683728, 39.60076966], [24.41679279, 39....	Hookah Bar	Café	Yemeni Restaurant	Flower Shop	Cricket Ground	Department Store	Dessert Shop
14_Bani Harithah Dist.	14	39.627848	[[24.48564547, 39.61575111], [24.48662775, 39....	Persian Restaurant	Café	Diner	Indian Restaurant	Yemeni Restaurant	Fast Food Restaurant	Cricket Ground
14_Shuran Dist.	14	39.622276	[[24.4122708, 39.62312234], [24.412227, 39.623...	Café	Yemeni Restaurant	Flower Shop	Cricket Ground	Department Store	Dessert Shop	Diner

Cluster 2

```
[9]: almadinah_merged.loc[almadinah_merged['Cluster Labels'] == 1, almadinah_merged.columns[[1] + list(range(5, a
```

[9]:

Neighborhood	city_id	Longitude	boundaries	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue

Results

- **Name each cluster based on most common venues in this cluster**

```
Cluster_No: 0
Diner                   3
Department Store        3
Dog Run                 3
Yemeni Restaurant       3
Dessert Shop            3
Name: Venue, dtype: int64

Cluster_No: 1
Hotel                          7
History Museum                 4
Hostel                         4
Middle Eastern Restaurant      4
Farmers Market                 3
Name: Venue, dtype: int64

Cluster_No: 2
Diner                  1
Cricket Ground         1
Department Store       1
Donut Shop             1
Flower Shop            1
Name: Venue, dtype: int64

Cluster_No: 3
Dessert Shop             29
Yemeni Restaurant        27
Diner                    24
Doner Restaurant         23
Falafel Restaurant       23
Name: Venue, dtype: int64

Cluster_No: 4
Dessert Shop           2
Cosmetics Shop         2
Department Store       2
Flower Shop            2
Diner                  2
```

Conclusion

- We can name cluster from most two/three Venues as following:
  - - Cluster_0: Cricket Ground area,
  - - Cluster_1: Hotel
  - - Cluster_2: Diner
  - - Cluster_3: Yemeni Restaurant
  - - Cluster_4: Cosmetics Shop, Flower Shop