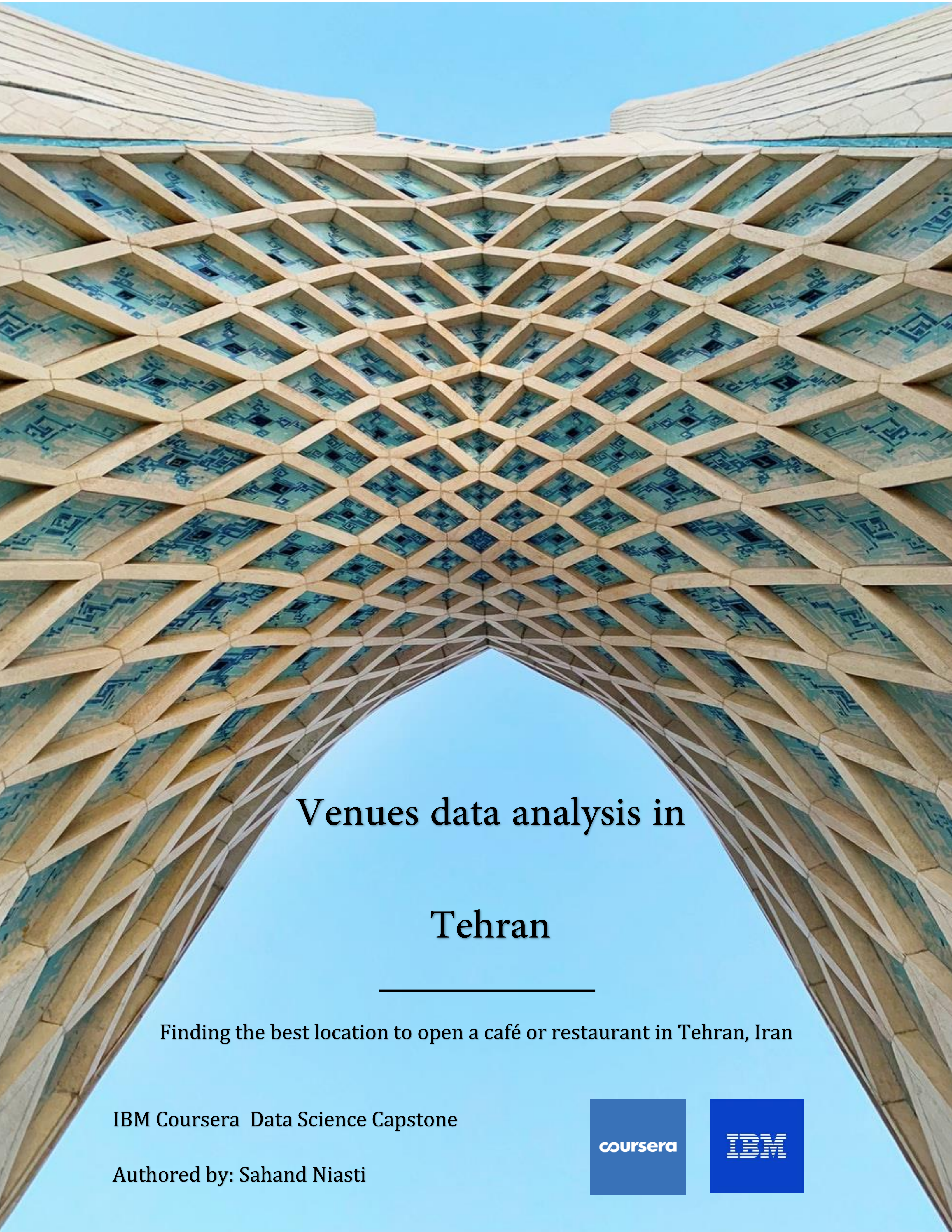


Venues data analysis in

Tehran

Finding the best location to open a café or restaurant in Tehran, Iran

IBM Coursera Data Science Capstone

Authored by: Sahand Niasti

Introduction

Tehran is one of the largest metropolises in the world where over **8.7 million** people live in city and around 15 million people in the larger metropolitan area of Greater Tehran. This city is the most populous city in Iran and Western Asia [1], and has the second-largest metropolitan area in the Middle East. So, it would not be shocking if you know that this city has a population density of **16,279** people per square kilometer [2]. As a resident of this city, I decided to use Tehran in my project. The city is divided up into 22 districts in total. All these information tells us that this city has quite an intertwined and mixed structure.

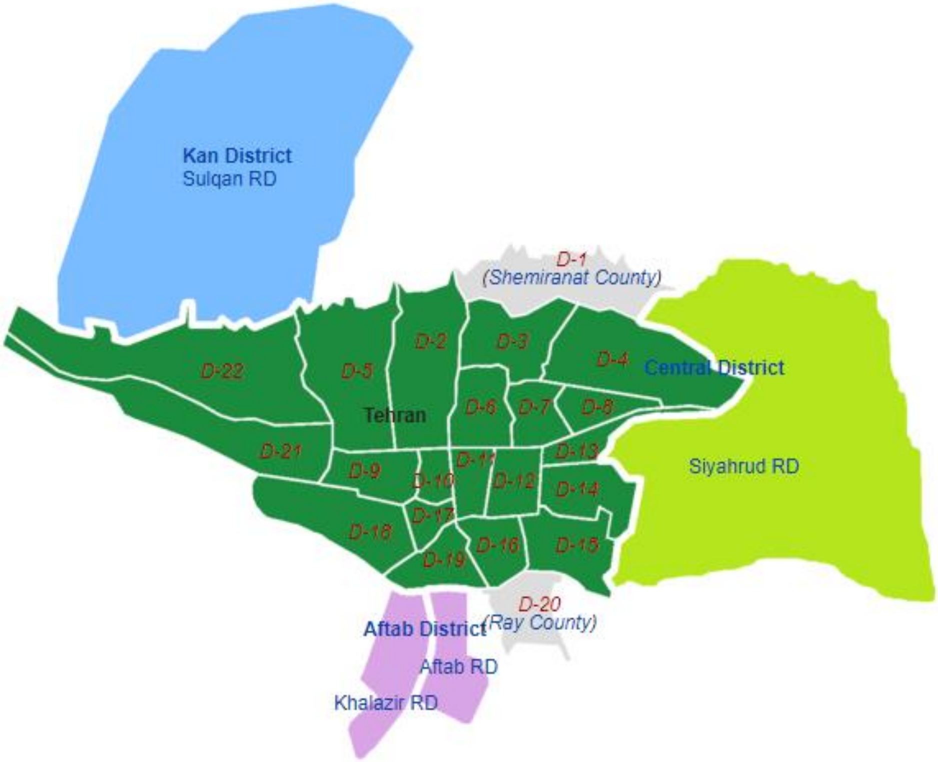


Figure 1. Tehran boroughs distribution in a map

A business manager who have never lived in Tehran sees this populated city as a great opportunity to earn money. If This person tends to invest in a form of café or restaurant, where or which boroughs should be selected to open such facility. In order to answer this question, we have to build a model to get some recommendations where our stakeholders prefer to start business.

Business problem

The objective of this capstone project is to analyze and select the best locations in the city of Tehran, Iran to open a café or restaurant. Using data science methodology and machine learning techniques like clustering, and taking into account the fact that our business owner prefer to choose a district according to the social places density, leads us to the aim to provide solutions to this business question: In the city of Tehran , if a property developer is looking to open a new café or restaurant, where would you recommend that they open it?

Data Description

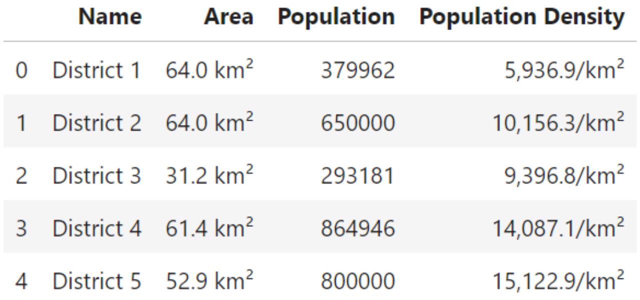
To solve the problem, we need the list of boroughs in Tehran accompanied with their Latitude and Longitude coordinates. This defines the scope of the project which is confined to the city of Tehran, help to plot the map, and eventually aid to get the venue data of the capital city of Iran. Sources of data and methods to extract them would be:

- Foursquare: It is a local search-and-discovery service which provides information on different types of entertainment, drinking and dining venues. Foursquare has an API that can be used to query their database and find information related to the venues, such as location, overall category, reviews and tips. [3]
- Wikipedia: There are not too many public data available and related to demographic and social parameters for the city of Tehran. Therefore, we decided to trust on this page and extract required information, with the help of Python requests and beautifulsoup packages. [4]

- Google Map: I also used ‘Search Nearby’ option to get the center coordinates of each Borough. [5]

Methodology

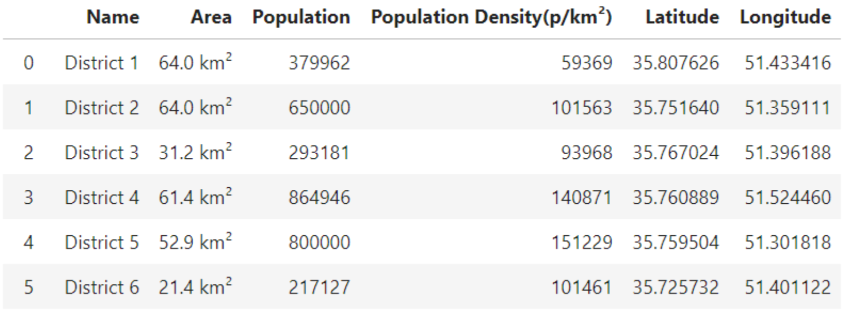
Firstly, we need to get the list of boroughs in the city of Tehran. This list was available in the Wikipedia page, but luckily, we managed to find a csv file on Github related to different boroughs located in Tehran. This data has been downloaded and will be read for further analysis.



	Name	Area	Population	Population Density
0	District 1	64.0 km^2	379962	5,936.9/km^2
1	District 2	64.0 km^2	650000	10,156.3/km^2
2	District 3	31.2 km^2	293181	9,396.8/km^2
3	District 4	61.4 km^2	864946	14,087.1/km^2
4	District 5	52.9 km^2	800000	15,122.9/km^2

Figure 2. The list of boroughs in Tehran

However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geolocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude.



	Name	Area	Population	Population Density(p/km^2)	Latitude	Longitude
0	District 1	64.0 km^2	379962	59369	35.807626	51.433416
1	District 2	64.0 km^2	650000	101563	35.751640	51.359111
2	District 3	31.2 km^2	293181	93968	35.767024	51.396188
3	District 4	61.4 km^2	864946	140871	35.760889	51.524460
4	District 5	52.9 km^2	800000	151229	35.759504	51.301818
5	District 6	21.4 km^2	217127	101461	35.725732	51.401122

Figure 3. the list of boroughs with latitude and longitude coordination

After gathering the data, we will populate the data into a pandas DataFrame and then visualize the boroughs in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Tehran.

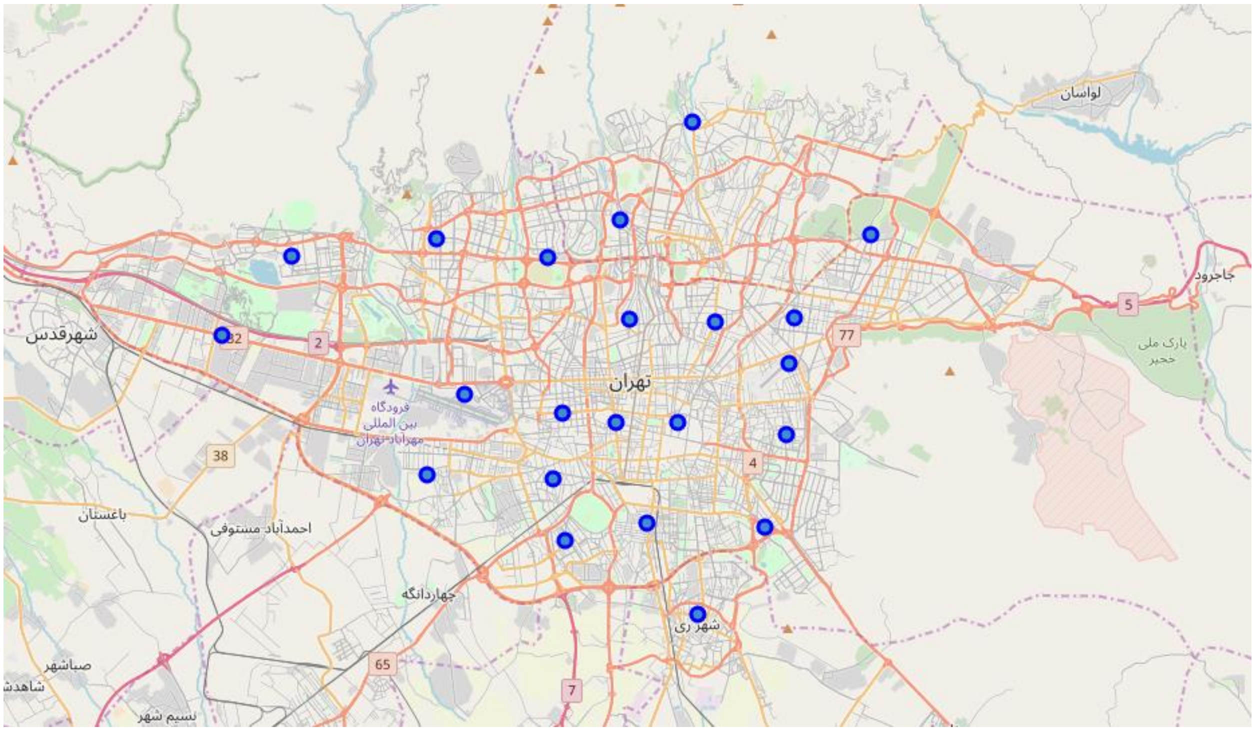
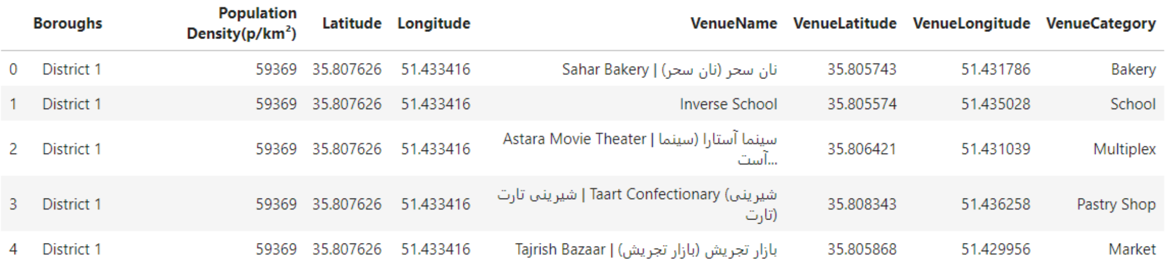


Figure 4. map of Tehran with boroughs superimposed on top

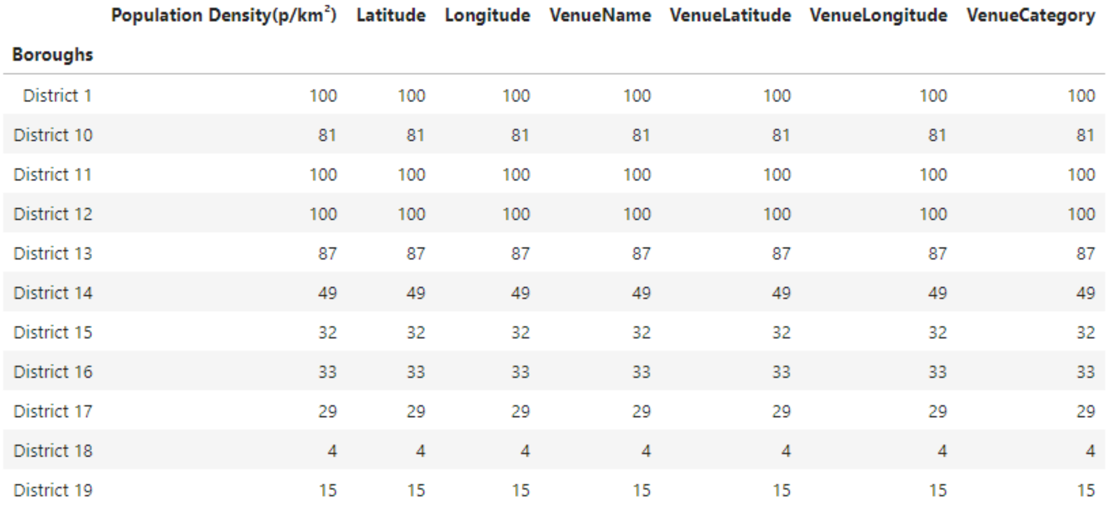
Next, we will use Foursquare API to get the top 100 venues that are within a radius of 3000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the boroughs in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude.



	Boroughs	Population Density(p/km^2)	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	District 1	59369	35.807626	51.433416	Sahar Bakery | (نان سحر) نان سحر	35.805743	51.431786	Bakery
1	District 1	59369	35.807626	51.433416	Inverse School	35.805574	51.435028	School
2	District 1	59369	35.807626	51.433416	Astara Movie Theater | (سینما) آستارا سینما آست...	35.806421	51.431039	Multiplex
3	District 1	59369	35.807626	51.433416	شیرینی تارت | Taart Confectionary (شیرینی تارت)	35.808343	51.436258	Pastry Shop
4	District 1	59369	35.807626	51.433416	Tajrish Bazaar | (بازار تجریش) بازار تجریش	35.805868	51.429956	Market

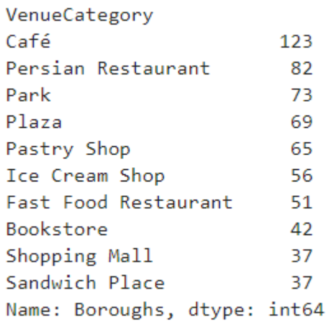
Figure 5. the extracted data from the Foursquare APIs

With the data, we can check how many venues were returned for each boroughs and examine how many unique categories can be curated from all the returned venues.



	Population Density(p/km^2)	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
Boroughs							
District 1	100	100	100	100	100	100	100
District 10	81	81	81	81	81	81	81
District 11	100	100	100	100	100	100	100
District 12	100	100	100	100	100	100	100
District 13	87	87	87	87	87	87	87
District 14	49	49	49	49	49	49	49
District 15	32	32	32	32	32	32	32
District 16	33	33	33	33	33	33	33
District 17	29	29	29	29	29	29	29
District 18	4	4	4	4	4	4	4
District 19	15	15	15	15	15	15	15

Figure 6. the number of venues returned for each borough



```
VenueCategory
Café                    123
Persian Restaurant       82
Park                     73
Plaza                    69
Pastry Shop              65
Ice Cream Shop           56
Fast Food Restaurant     51
Bookstore                42
Shopping Mall            37
Sandwich Place           37
Name: Boroughs, dtype: int64
```

Figure 7. The number of locations per Venue Category in Tehran

We can see that how District 1, 2, 3, 5, 6, 7, 8, 11 and 12 reached the **100** limit of venues. On the other hand, District 18, 19, 20 and 21 are below **25** venues in our given coordinates with Latitude and Longitude, in below graph.

The result doesn't mean that inquiry run all the possible results in boroughs. Actually, it depends on given Latitude and Longitude information and here is we just run single Latitude and Longitude pair for each borough. We can increase the possibilities with boroughs information with more Latitude and Longitude coordinates.

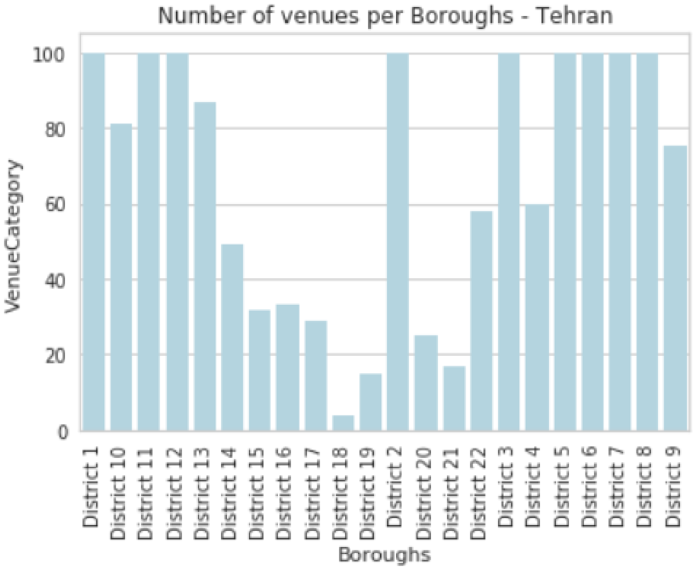
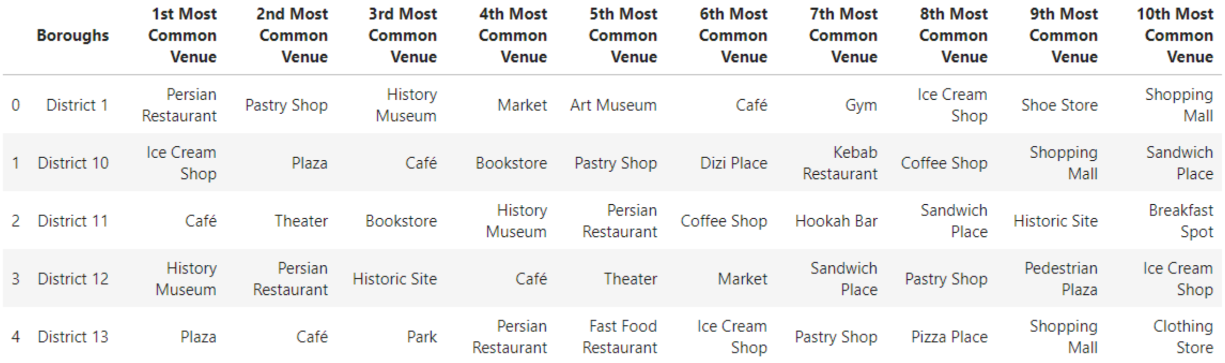


Figure 8. bar chart of the number of venues per borough in Tehran

Then, we will analyze each borough by grouping the rows by boroughs and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data to be used in clustering. This table gives us information about the list of top 10 venues in each district, which shows the popularity of venues in different boroughs of Tehran.



	Boroughs	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	9th Most Common Venue	10th Most Common Venue
0	District 1	Persian Restaurant	Pastry Shop	History Museum	Market	Art Museum	Café	Gym	Ice Cream Shop	Shoe Store	Shopping Mall
1	District 10	Ice Cream Shop	Plaza	Café	Bookstore	Pastry Shop	Dizi Place	Kebab Restaurant	Coffee Shop	Shopping Mall	Sandwich Place
2	District 11	Café	Theater	Bookstore	History Museum	Persian Restaurant	Coffee Shop	Hookah Bar	Sandwich Place	Historic Site	Breakfast Spot
3	District 12	History Museum	Persian Restaurant	Historic Site	Café	Theater	Market	Sandwich Place	Pastry Shop	Pedestrian Plaza	Ice Cream Shop
4	District 13	Plaza	Café	Park	Persian Restaurant	Fast Food Restaurant	Ice Cream Shop	Pastry Shop	Pizza Place	Shopping Mall	Clothing Store

Figure 9. the popularity of venues in each borough

Lastly, we will perform clustering on the data by using k-means clustering. This algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. First, we will run K-Means to cluster the boroughs into **8** clusters because when we analyze the K-Means with elbow method it ensured us that the 8 degree for optimum k of the K-Means.

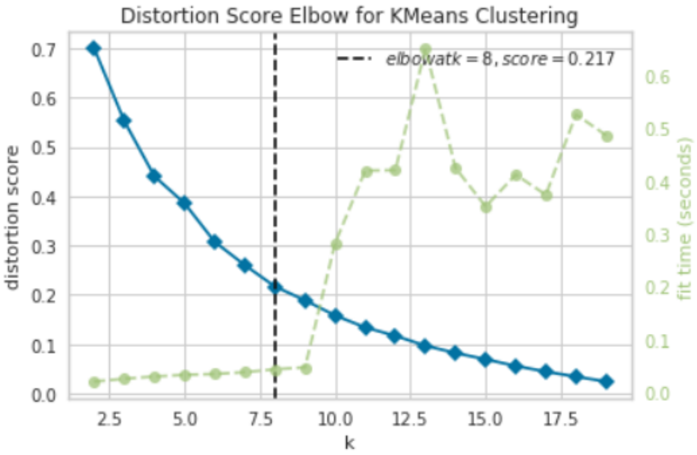
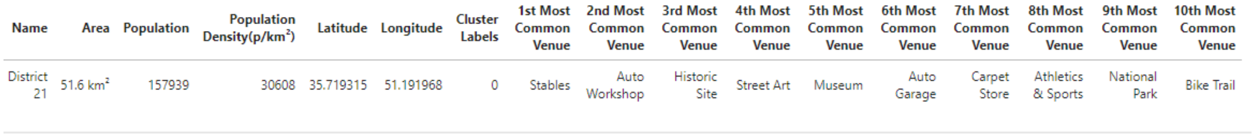


Figure 10. data analyzing with Elbow method to figure out the optimum K number

We will cluster the boroughs into 8 clusters based on their frequency of occurrence. The results will allow us to identify which boroughs have higher concentration on café and restaurant while which boroughs have fewer number of them. Based on the occurrence of such venues in different boroughs, it will help us to answer the question as to which districts are most suitable to open new café or restaurant.

Results

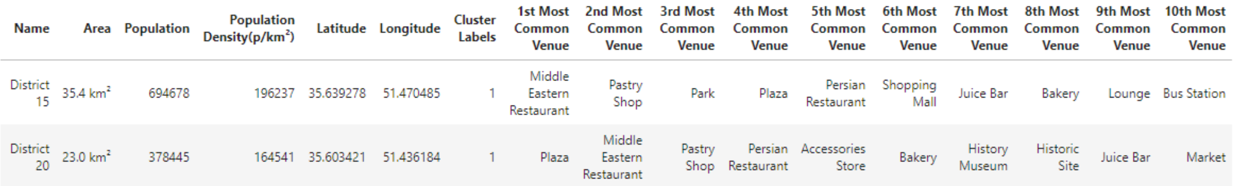
The results from the k-means clustering show that we can categorize the boroughs into 8 clusters based on the frequency of occurrence for venues:



Name	Area	Population	Population Density(p/km^2)	Latitude	Longitude	Cluster Labels	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	9th Most Common Venue	10th Most Common Venue
District 21	51.6 km^2	157939	30608	35.719315	51.191968	0	Stables	Auto Workshop	Historic Site	Street Art	Museum	Auto Garage	Carpet Store	Athletics & Sports	National Park	Bike Trail

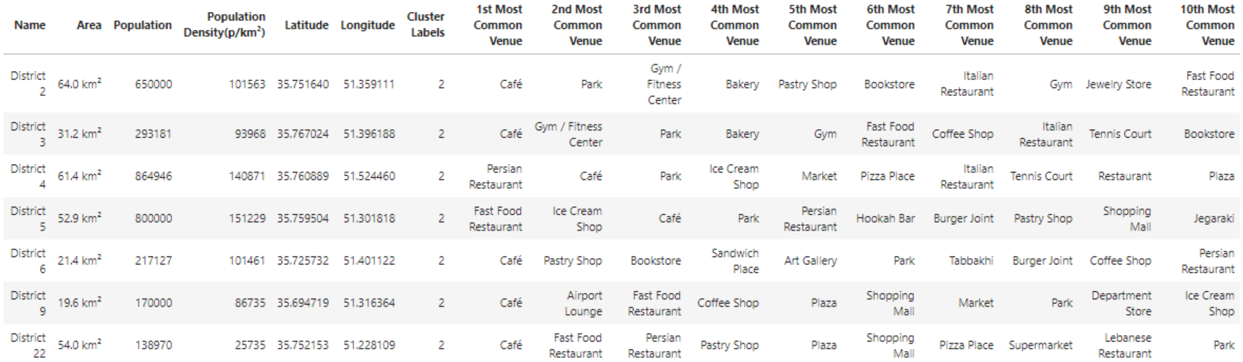
Figure 11. cluster 0 - Red

As we can see, the most common venue in this category is stables, and most auto workshops are placed here. So, we call it "stables and auto workshops". The next clusters are subsequently named based on frequency of facilities and venues.



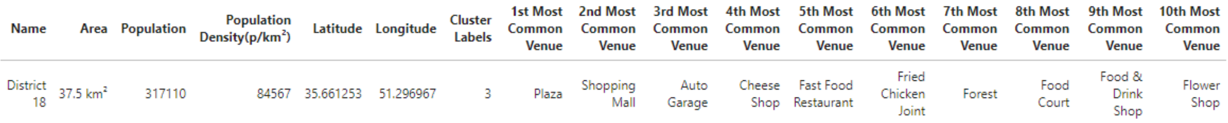
Name	Area	Population	Population Density(p/km^2)	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
District 15	35.4 km^2	694678	196237	35.639278	51.470485	1	Middle Eastern Restaurant	Pastry Shop	Park	Plaza	Persian Restaurant	Shopping Mall	Juice Bar	Bakery	Lounge	Bus Station
District 20	23.0 km^2	378445	164541	35.603421	51.436184	1	Plaza	Middle Eastern Restaurant	Pastry Shop	Persian Restaurant	Accessories Store	Bakery	History Museum	Historic Site	Juice Bar	Market

Figure 12. Cluster 1 – Purple – "middle east restaurants & pastry shops"



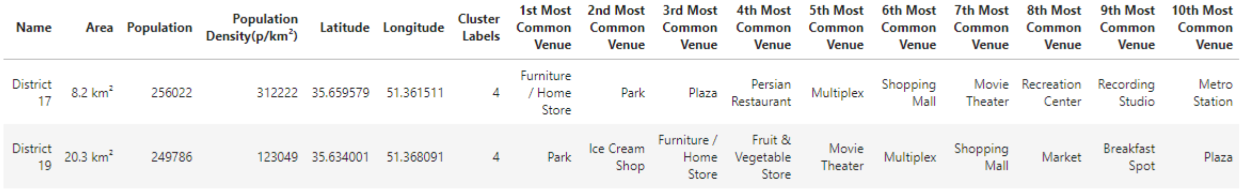
Name	Area	Population	Population Density(p/km^2)	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
District 2	64.0 km^2	650000	101563	35.751640	51.359111	2	Café	Park	Gym / Fitness Center	Bakery	Pastry Shop	Bookstore	Italian Restaurant	Gym	Jewelry Store	Fast Food Restaurant
District 3	31.2 km^2	293181	93968	35.767024	51.396188	2	Café	Gym / Fitness Center	Park	Bakery	Gym	Fast Food Restaurant	Coffee Shop	Italian Restaurant	Tennis Court	Bookstore
District 4	61.4 km^2	864946	140871	35.760889	51.524460	2	Persian Restaurant	Café	Park	Ice Cream Shop	Market	Pizza Place	Italian Restaurant	Tennis Court	Restaurant	Plaza
District 5	52.9 km^2	800000	151229	35.759504	51.301818	2	Fast Food Restaurant	Ice Cream Shop	Café	Park	Persian Restaurant	Hookah Bar	Burger Joint	Pastry Shop	Shopping Mall	Jegaraki
District 6	21.4 km^2	217127	101461	35.725732	51.401122	2	Café	Pastry Shop	Bookstore	Sandwich Place	Art Gallery	Park	Tabbakhi	Burger Joint	Coffee Shop	Persian Restaurant
District 9	19.6 km^2	170000	86735	35.694719	51.316364	2	Café	Airport Lounge	Fast Food Restaurant	Coffee Shop	Plaza	Shopping Mall	Market	Park	Department Store	Ice Cream Shop
District 22	54.0 km^2	138970	25735	35.752153	51.228109	2	Café	Fast Food Restaurant	Persian Restaurant	Pastry Shop	Plaza	Shopping Mall	Pizza Place	Supermarket	Lebanese Restaurant	Park

Figure 13. Cluster 2 – Navy Blue – "Café & restaurant"



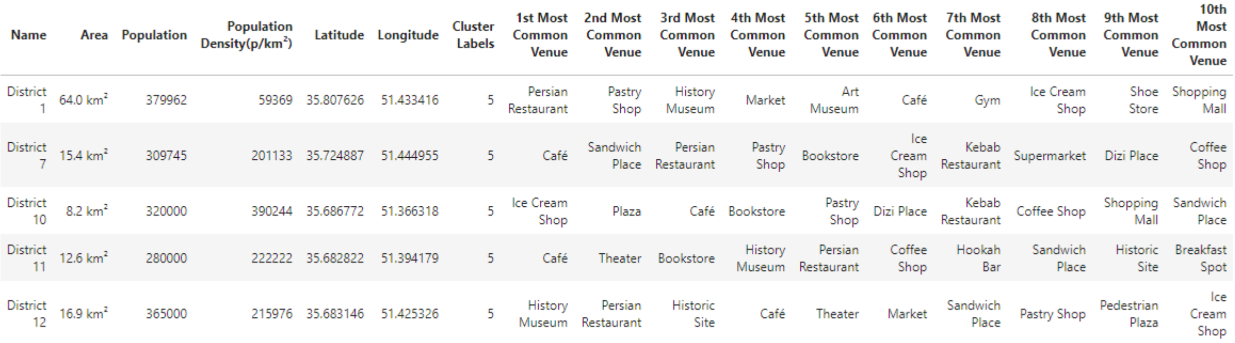
Name	Area	Population	Population Density(p/km^2)	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
District 18	37.5 km^2	317110	84567	35.661253	51.296967	3	Plaza	Shopping Mall	Auto Garage	Cheese Shop	Fast Food Restaurant	Fried Chicken Joint	Forest	Food Court	Food & Drink Shop	Flower Shop

Figure 14. Cluster 3 – Sky Blue - “Plaza & shopping mall”



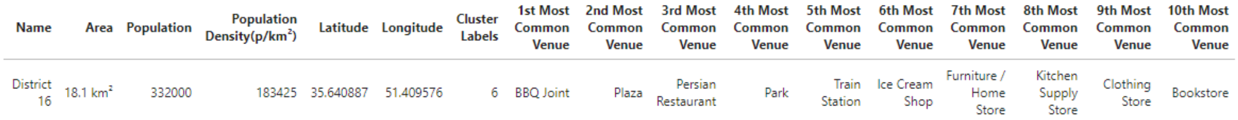
Name	Area	Population	Population Density(p/km^2)	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
District 17	8.2 km^2	256022	312222	35.659579	51.361511	4	Furniture / Home Store	Park	Plaza	Persian Restaurant	Multiplex	Shopping Mall	Movie Theater	Recreation Center	Recording Studio	Metro Station
District 19	20.3 km^2	249786	123049	35.634001	51.368091	4	Park	Ice Cream Shop	Furniture / Home Store	Fruit & Vegetable Store	Movie Theater	Multiplex	Shopping Mall	Market	Breakfast Spot	Plaza

Figure 15. Cluster 4 – Cyan – “furniture & park”



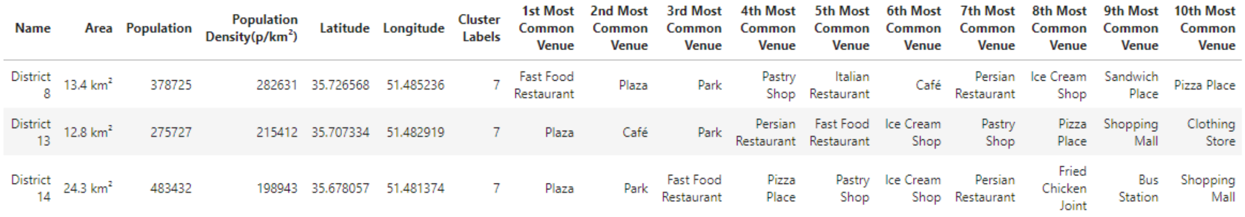
Name	Area	Population	Population Density(p/km^2)	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
District 1	64.0 km^2	379962	59369	35.807626	51.433416	5	Persian Restaurant	Pastry Shop	History Museum	Market	Art Museum	Café	Gym	Ice Cream Shop	Shoe Store	Shopping Mall
District 7	15.4 km^2	309745	201133	35.724887	51.444955	5	Café	Sandwich Place	Persian Restaurant	Pastry Shop	Bookstore	Ice Cream Shop	Kebab Restaurant	Supermarket	Dizi Place	Coffee Shop
District 10	8.2 km^2	320000	390244	35.686772	51.366318	5	Ice Cream Shop	Plaza	Café	Bookstore	Pastry Shop	Dizi Place	Kebab Restaurant	Coffee Shop	Shopping Mall	Sandwich Place
District 11	12.6 km^2	280000	222222	35.682822	51.394179	5	Café	Theater	Bookstore	History Museum	Persian Restaurant	Coffee Shop	Hookah Bar	Sandwich Place	Historic Site	Breakfast Spot
District 12	16.9 km^2	365000	215976	35.683146	51.425326	5	History Museum	Persian Restaurant	Historic Site	Café	Theater	Market	Sandwich Place	Pastry Shop	Pedestrian Plaza	Ice Cream Shop

Figure 16. Cluster 5 – Green – “café with entertaining venues “



Name	Area	Population	Population Density(p/km^2)	Latitude	Longitude	Cluster Labels	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	9th Most Common Venue	10th Most Common Venue
District 16	18.1 km^2	332000	183425	35.640887	51.409576	6	BBQ Joint	Plaza	Persian Restaurant	Park	Train Station	Ice Cream Shop	Furniture / Home Store	Kitchen Supply Store	Clothing Store	Bookstore

Figure 17. Cluster 6 – Yellow – “BBQ Joint “



Name	Area	Population	Population Density(p/km^2)	Latitude	Longitude	Cluster Labels	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	9th Most Common Venue	10th Most Common Venue
District 8	13.4 km^2	378725	282631	35.726568	51.485236	7	Fast Food Restaurant	Plaza	Park	Pastry Shop	Italian Restaurant	Café	Persian Restaurant	Ice Cream Shop	Sandwich Place	Pizza Place
District 13	12.8 km^2	275727	215412	35.707334	51.482919	7	Plaza	Café	Park	Persian Restaurant	Fast Food Restaurant	Ice Cream Shop	Pastry Shop	Pizza Place	Shopping Mall	Clothing Store
District 14	24.3 km^2	483432	198943	35.678057	51.481374	7	Plaza	Park	Fast Food Restaurant	Pizza Place	Pastry Shop	Ice Cream Shop	Persian Restaurant	Fried Chicken Joint	Bus Station	Shopping Mall

Figure 18. Cluster 7 - Orange – “Plaza “

The results of the clustering are visualized in the map below with colors mentioned in each cluster.

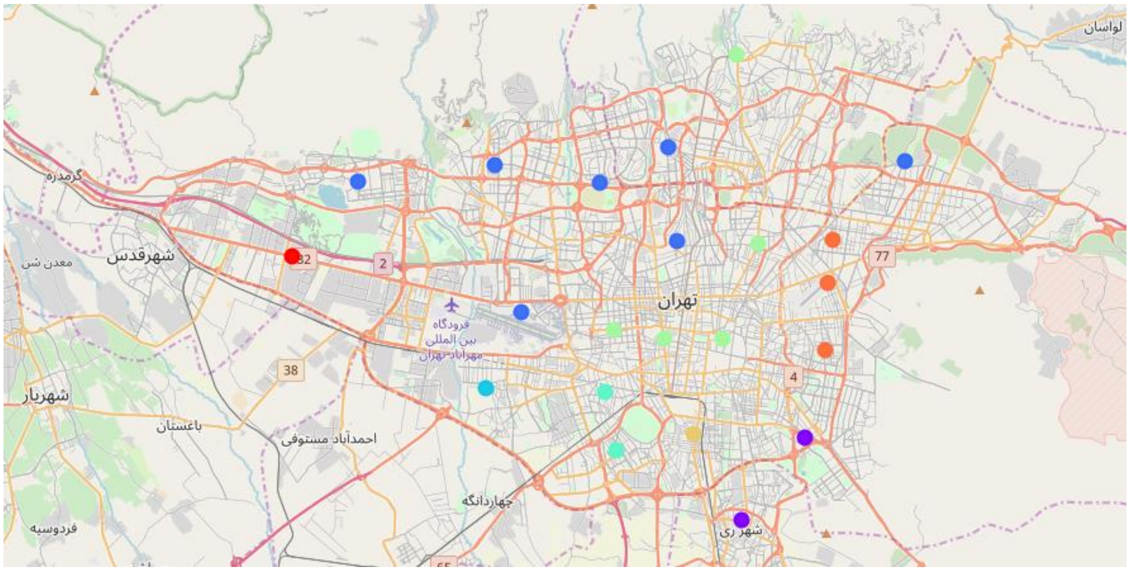


Figure 19. Clustered map of Tehran

Finally, our aim was to illustrate which boroughs or neighborhood is proper if a business owner tend to invest in a café or restaurant. So, we decide to compare the population of each borough with boroughs clustered in this city. As a result, a choropleth map was created to show such information.

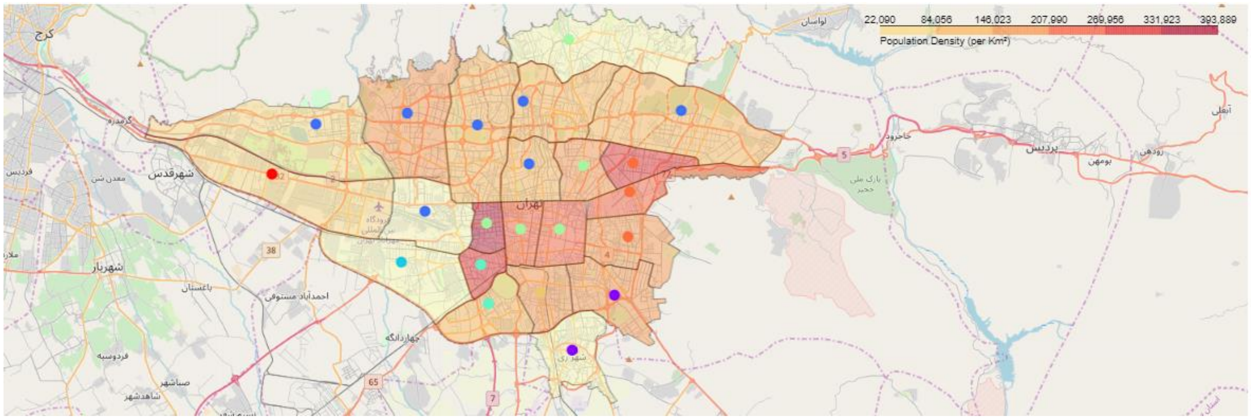


Figure 20. Clustered map merged with population density

Discussion

As I mentioned before, Tehran is a huge city with high population density in a narrow area. The total number of measurements and population densities of the 22 districts in total can vary. Due to such complexity, numerous approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high quality results for this city. When we tested the Elbow method, the optimum k value was set to 8, as it was shown on the diagram.

However, only 22 district coordinates were used. For more detailed and accurate guidance, such as investigating all neighborhoods within the districts, a different dataset could be implemented to reach a more detailed result.

In the end, the clustered data was merged with the population density of each borough in order to achieve a better result. For example, the data clustered as “café & restaurant “was in 7 district. By taking into account the population, which was shown in figure 20, business men or women can make a better decision.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 8 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new café or restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: boroughs in cluster 2 seem to be the optimum option. However, if the stakeholders tend to choose a place based on availability of other venues, cluster 5 can be a viable option as well.

The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations with regard to their personal preferences. All we could do was to enlighten them with a new perspective.

Limitations and Suggestions for Future Research

In this project, we only consider two factors i.e. frequency of occurrence of venues and the population density. There are other factors such as income of residents and average house price that surely have impact on decision making of a new café or restaurant. However, to the best knowledge of this researcher, such data were not available or it could take considerable amount of time to be achieved. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine a preferred location. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

Sahand Niasti

References

[1] https://en.wikipedia.org/wiki/List_of_metropolitan_areas_in_Asia

[2] https://en.wikipedia.org/wiki/Tehran

[3] https://developer.foursquare.com/

[4] https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Tehran

[5] https://www.google.com/maps