

# Clustering Tehran venues based on each district & neighborhoods

By Ali Yamini

29 September 2019

## 1. Introduction

### 1.1 Background

Tehran is a very big city in terms of the number of districts and population. Tehran has over 22 districts with a population of 12 million people(2018) which makes it the second most populated city in the middle east after Istanbul. A big city like Tehran has lot's of neighborhoods and venues.

### 1.2 Problem

Unfortunately, a big city like Tehran doesn't have a good analysis of their venues. For example, a tourist doesn't know the best places in Tehran based on each neighborhood so In this project we are going to cluster each neighborhood in 22 districts of Tehran based on their top 10 venues in each neighborhood.

### 1.3 Interest

The output of this project can be very helpful for tourists or anyone who is interested in finding the best venues of Tehran based on their neighborhoods.

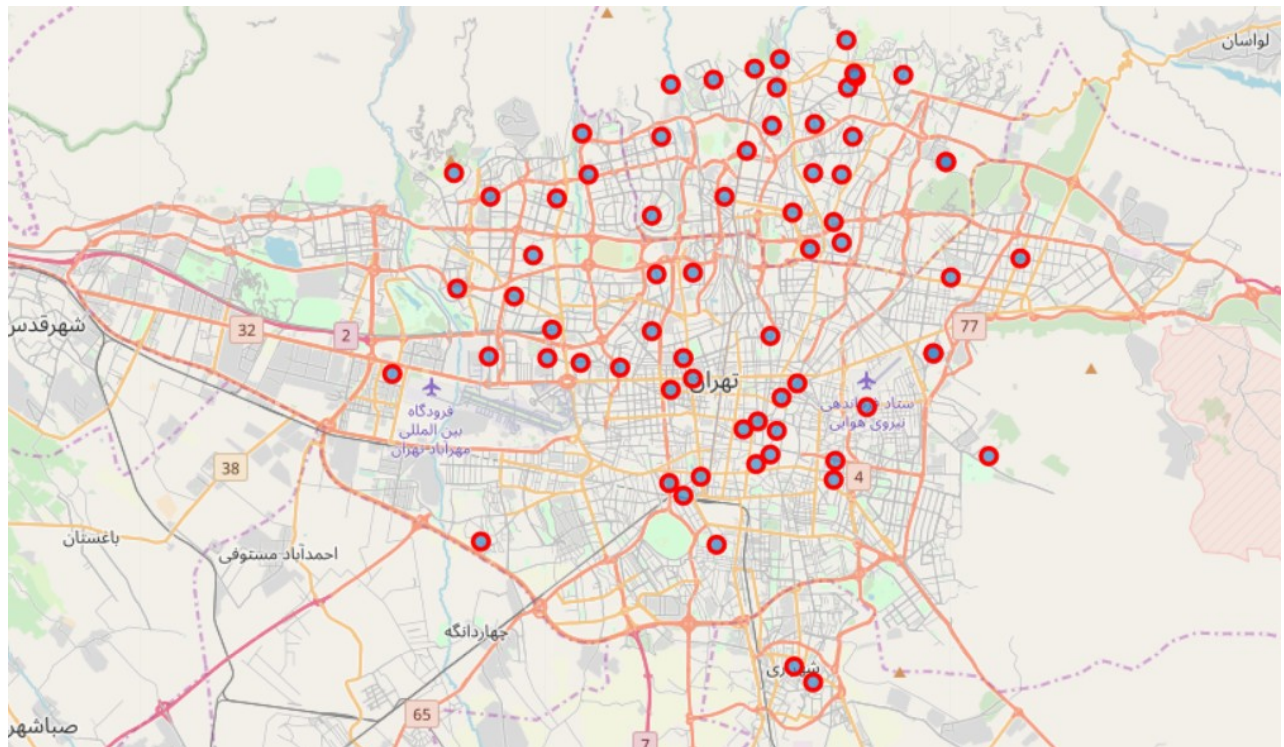
## 2. Data acquisition

Surprisingly, Tehran doesn't have a clean database for data scientists. As a data scientist you have to pretty much collect everything you need yourself. So my first step would be to collect districts and neighborhoods data from Wikipedia and put each one in a row of my Dataframe. After this step I will collect the coordinates for my neighborhoods to plot my data on a map. Coordinates will be collected by Geopy library. For the Final step, I will get my venues data based on each neighborhood using [Foursquare](#) API which surprisingly has a valuable database for the city of Tehran.

We gathered our data, thorough multiple steps. First we manually collected our data from the source that we had previously mentioned. In this step we collected name and geographical position of each neighborhood in the city of Tehran. By collecting

this data we now have 2 columns named City side and Neighborhood which we are going to iterate through them. After this step we collected coordinates of each of our neighborhoods by using geopy library. For some reasons, few neighborhoods had been falsely coordinated which we corrected them at the end. The reason for this was because geopy uses open street maps to collect geographical data for each neighborhood, and sometimes our neighborhood is not correctly mentioned in the open street map database So we handled it after that geopy's work was done.

Now we have 4 columns in our database including City side, Neighborhood, Latitude and Longitude. So now we can plot our data into a map using folium.



### 3. Methodology

The methodology used to approach this problem includes some statistical exploration of the data and some visualizations. The main machine learning technique involved in the development of this project is clustering, in concrete the K-Means algorithm was used, implemented with Python. At a first moment, the main problem was how to obtain the necessary data to build a constructive approach to the problem to be tackled. Usually, to solve these kinds of optimal business location problems, a lot of consumer's data are needed, but for this example and for the sake of simplicity, the focus was put mainly on neighborhoods. With all this being considered, it was decided that the main goal to efficiently solve this problem, was firstly to define what our target population is, and secondly, find the areas where this population is living, and finally, examine the venues and restaurants in this area to see if our product could work. Here is an example of the data used:

	City side	Neighborhood	latitude	longitude
0	North	Aghdasieh	35.593837	51.444406
1	North	Lavizan	35.777055	51.502150
2	North	Ajodanieh	35.807800	51.483600
3	North	Darakeh	35.804346	51.382710
4	North	Darband	35.813517	51.429482

The First map is plotted after that the table above gathered through some processes.

In the next part we obtained the nearest venue to each neighborhood. This data was obtained through latitude and longitude of each location using Foursquare database.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Aghdasieh	35.593837	51.444406	Fina Food   فینا فود	35.593667	51.442405	Fast Food Restaurant
1	Aghdasieh	35.593837	51.444406	زمین چمن مطهری	35.591102	51.446235	Football Stadium
2	Aghdasieh	35.593837	51.444406	فست فود آرادArad Fastfood	35.593630	51.449789	Fast Food Restaurant
3	Ajodanieh	35.807800	51.483600	R&A coffee shop	35.805885	51.484921	Coffee Shop
4	Ajodanieh	35.807800	51.483600	Royal Aghdasiye Fitness Center	35.805618	51.479701	Gym / Fitness Center

Looking at this sample, it is possible to see the names of the venues, their coordinates, and the category of each venue. The results are ordered by boroughs. This is a vital step in the segmentation process, since all the important data about the venues is obtained from here. Once the venues per boroughs were obtained, it was then needed to look at the mean occurrence of each venue by neighborhood:

```

----Abbas Abad----
      venue  freq
0 Persian Restaurant 0.25
1           Plaza    0.17
2       Women's Store 0.08
3       Astrologer    0.08
4           Market    0.08

----Afsariyeh----
      venue  freq
0           Plaza    0.17
1       Sports Club    0.17
2           Bookstore 0.17
3           Diner     0.17
4 Kebab Restaurant    0.17

----Aghdasieh----
      venue  freq
0 Fast Food Restaurant 0.67
1   Football Stadium    0.33
2           Yoga Studio 0.00
3           Nail Salon  0.00
4 Moroccan Restaurant  0.00

----Ajodanieh----
      venue  freq
0           Gym     0.18
1       Pizza Place  0.09
2       Coffee Shop  0.09
3           Café     0.09
4 Fast Food Restaurant 0.09

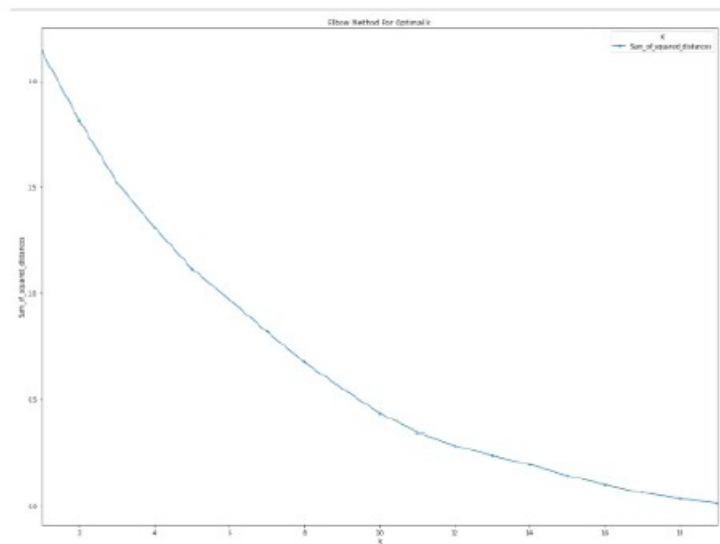
```

This what the frequencies of occurrence looks like. With this data, it is possible to know which the most common venues are:

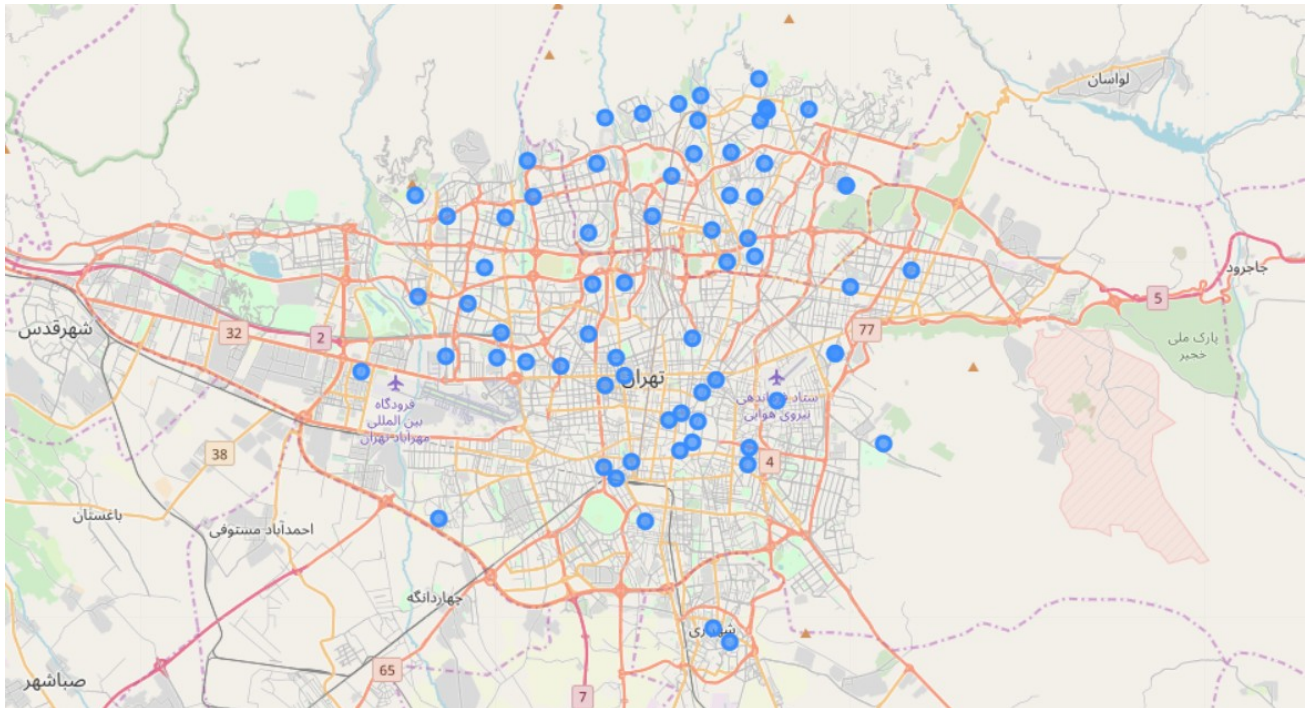
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abbas Abad	Persian Restaurant	Plaza	Tailor Shop	Costume Shop	Carpet Store	Metro Station	Market	Astrologer	Women's Store	Furniture Store
1	Afsariyeh	Bookstore	Plaza	Sports Club	Diner	Kebab Restaurant	Supermarket	Drugstore	Fast Food Restaurant	Farm	Farm
2	Aghdasieh	Fast Food Restaurant	Football Stadium	Women's Store	Donut Shop	Farm	Falafel Restaurant	Fabric Shop	Electronics Store	Dry Cleaner	Dry Cleaner
3	Ajodanieh	Gym	Fast Food Restaurant	Park	Gym / Fitness Center	Pizza Place	Coffee Shop	Japanese Restaurant	Café	Market	Agriculture
4	Amir Abad	Sports Club	Soccer Field	IT Services	Burger Joint	Athletics & Sports	Health & Beauty Service	Cultural Center	Drugstore	Farm	Culture

This process is progressive, once a piece of information is obtained, it is possible to go for the next one. With this data in hand, now the segmentation can be made, and the

clusters created. But first it is necessary to determine somehow, what the appropriate number of clusters is. To perform this task, the elbow method was used. This method consists in plotting a hypothetical and usually large number of clusters in our data, and draw a curve representing the squared distances between each cluster. At some point, the distances will descend to a point where there is no need to keep increasing them. This means that creating more divisions in the data (clusters) is pointless as the difference between groups starts being highly difficult to appreciate:



This is our curve. The distances start reducing importantly from cluster 6 on. So, it was determined that the optimal number of clusters for this problem was 6. With this being done, it is possible to build the clusters now and have a look at them:



These are the 6 clusters on the map of Tehran, it is possible to see how many neighborhoods belong to each cluster, which is also important information. Now it is possible to examine the data of each cluster:

	City side	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	6th Most Common Venue
0	North	Aghdasieh	35.593837	51.444406	4.0	Fast Food Restaurant	Football Stadium	Women's Store	Donut Shop	Farm	Falafel Restaurant
1	North	Lavizan	35.777055	51.502150	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	North	Ajodanieh	35.807800	51.483600	0.0	Gym	Fast Food Restaurant	Park	Gym / Fitness Center	Pizza Place	Coffee Shop
3	North	Darakeh	35.804346	51.382710	2.0	Persian Restaurant	Café	Restaurant	Tennis Court	Athletics & Sports	Hookah Bar
4	North	Darband	35.813517	51.429482	2.0	History Museum	Garden	Café	Burger Joint	Supermarket	Steakhouse