# Finding The Best Place to Establishing Office in Tehran

Alireza Siami

July 4, 2021

## 1. Introduction

### 1.1 Background

This Project is a partial fulfillment of the Coursera IBM Data Science certification course. The project requirements are to leverage the "Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve." I decided to work on my home town, Tehran, Iran as my project, though the amount of data about Tehran in the internet is low in compare to other cities. Even the Foursquare API may have a limited applicability because the app may not be widely used in my city.

Tehran is the capital of Iran and Tehran Province. With a population of around 8.7 million in the city and 15 million in the larger metropolitan area of Greater Tehran, Tehran is the most populous city in Iran and Western Asia, and has the second-largest metropolitan area in the Middle East (after Cairo). It is ranked 24th in the world by the population of its metropolitan area. [https://en.wikipedia.org/wiki/Tehran].

### 1.2 Problem

Nowadays, international oil companies are developing interest in investing in Iran for a number of reasons. The main one is that the sanctions against Iran, especially Iranian oil industry are decreasing these days. Besides, the work force in Iran is much cheaper than other oil-rich countries. Assuming these, some of oil companies are inclined to establish an office in the capital of Iran (Tehran).

From a foreign company view, there are some factors that contribute to select suitable location for installing the office. The important ones are:

- The location of office should be in business area. (Being close to other government and private offices)
- The location should have high rate of green environment so the neighborhood would be pleasant for the clerks and workers.
- Being close to hotels and restaurants with foreign cuisine.
- Easy access to public transports like metro.

In this project the different potential neighborhoods based on the factors that are important to choose better location, have been compared and finally the optimum neighborhood which satisfy most of the factors has been selected.

## 2. Data Acquisition and Cleaning

### 2.1 Data Sources

In this section, I describe the data that has been used in the project.

- Data Source 1 – Neighborhood Data
  Because of the shortage in data about Greater Tehran in the internet, I decided to carry out my project on Tehran that includes 22 districts. I first needed to obtain a list of all the locations or neighborhoods in Tehran. This information has been obtained from the following web address:
  https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Tehran.
- Data Source 2 – Geographical Coordinates
  Geographical coordinates for each neighborhood have been obtained with the aid of GEOPY Library.
- Data Source 3 – Venue categories
  I used the Foursquare API to retrieve venues, using the coordinates obtained in Data Source 2 above.

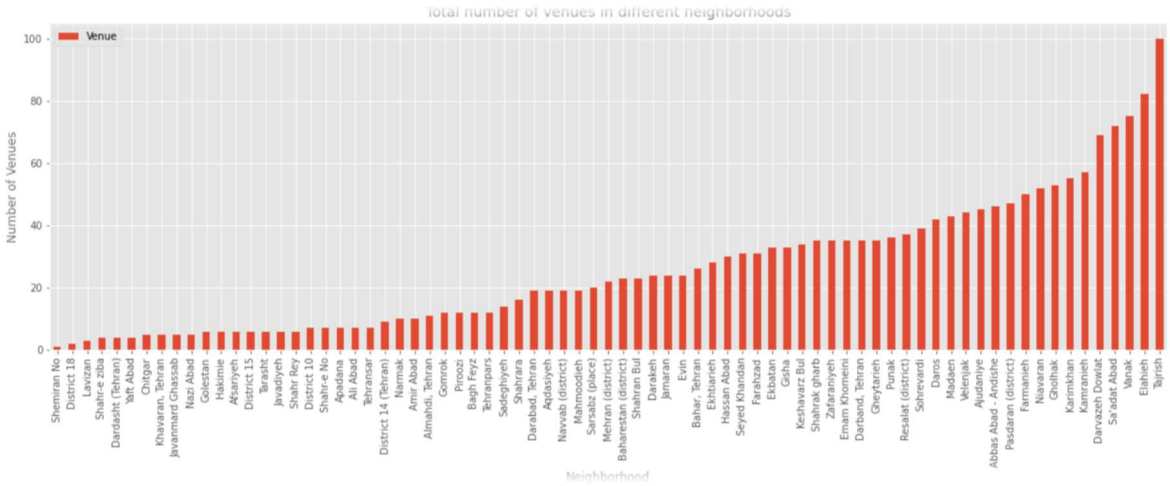
### 2.2 Data Cleaning

I used Beautiful Soup library to parse the html file and then to get my required information from that. By exploring the Wikipedia page and also the content of the file that returned by Beautiful Soup object, it can be noticed that the first 20 of 'URL' tags contain the list of neighborhoods. Therefore, by examining the 'URL' tags I extracted the list of neighborhoods. Then I converted the list of neighborhoods to the Pandas Data Frame. By looking at the Data Frame it is clear that it does not include all the neighborhoods and some of them missed. Therefore, I inserted the missed neighborhoods into the new Data Frame and then concatenate both of them.

By examining the neighborhoods data frame with Geopy API it can be seen that some of the neighborhood names are not as same as what they are in Geopy API. So, I changed them into matched spelling in the Geopy API. I used Geopy API to obtain longitude and latitude of the neighborhoods and then I added them in the data frame. As some neighborhoods does not have longitude and latitude in Geopy API and their latitude and longitude are zero in the data frame, I deleted these rows from the Data Frame. Also, the location of 'Surena Street' is out of this project scope so I deleted that too.

## 3. Methodology and Discussion

After saving the neighborhood names in the data frame with their corresponding latitude and longitude, Foursquare API was used to obtain the list of venues for each neighborhood and then they were inserted into the data frame under new column named “Venue”.
At this point, by grouping the data frame on neighborhoods and sorting that we could find the total number of venues each neighborhood has. Then Matplotlib library could be used to visualize the number of venues for each neighborhood.



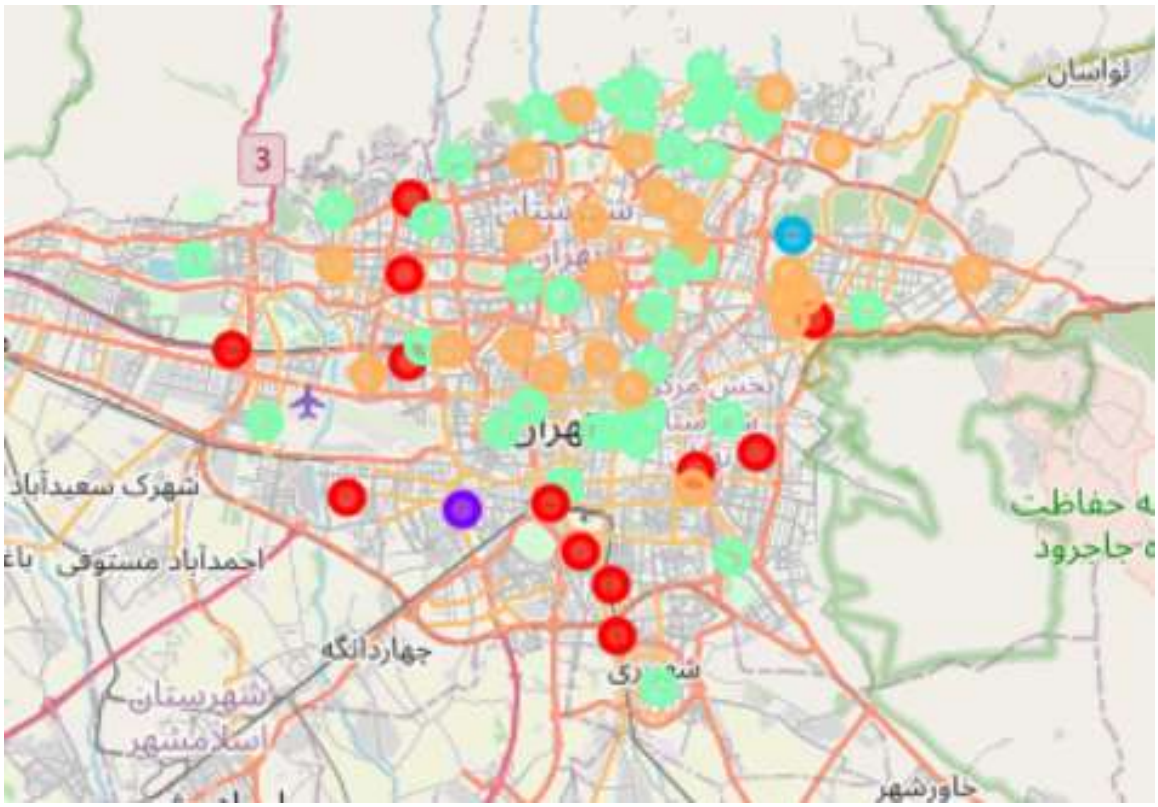

From this figure the following results can be extracted:

- There are 77 neighborhoods.
- Shemiran No, District 18, Lavizan, Shahr-e Ziba, and Dardasht are the neighborhoods with the lowest number of venues.
- Tajrish, Elahieh, Vanak, Sa’adat Abad, and Darvazeh Dowlat have the highest number of venues.
- Tajrish is the only neighborhood that reached the maximum limit number of venues. (100)

The venues which including but not excluding parks, restaurants, hotels, ATMs, metro stations, bus stops were converted to one hot system since they are categorical and cannot be used. Now, there is a new data frame which has neighborhoods as the index and list of venues as the columns.
After that, by grouping the previous Data Frame by neighborhood and using mean function we could get the mean frequency of each venue for each neighborhood. And then I used a function to sort the venues in descending order. Then, I created the new data frame and display the top 10 venues for each neighborhood. The result is shown here:

| | 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 - Andishe | Café | Fast Food Restaurant | Persian Restaurant | Sandwich Place | Pastry Shop | IT Services | Auto Garage | Art Gallery | Paper / Office Supplies Store | Comfort Food Restaurant |
| 1 | Afsariyeh | Sports Club | Amphitheater | Plaza | Bookstore | Diner | Kebab Restaurant | Notary | Moroccan Restaurant | Mosque | Mountain |
| 2 | Ajudaniye | Jewelry Store | Gym / Fitness Center | Café | Italian Restaurant | Fast Food Restaurant | Ice Cream Shop | Market | Coffee Shop | Persian Restaurant | Flower Shop |
| 3 | Ali Abad | Park | Plaza | Metro Station | Restaurant | Supermarket | Taxi Stand | Nail Salon | Monument / Landmark | Moroccan Restaurant | Mosque |
| 4 | Almahdi, Tehran | Park | French Restaurant | Supermarket | Flea Market | Gym | Pastry Shop | Persian Restaurant | Ice Cream Shop | Dance Studio | Market |

Then I applied k-means model to cluster the neighborhood. By examining different numbers of cluster, I have reached to 5 as it shows more better the difference between Tehran neighborhoods.
By the help of Folium library we can see the different clusters on the Tehran map:




For noticing the especial characteristics of each cluster, I inserted cluster numbers to the data frame which has neighborhood list as index and venues as columns and grouped them by cluster number while I imposed mean function. Then I sorted venues by their high frequency in each cluster. So, I obtained new data frame with cluster number as index and top common venues as columns. The resulting data frame can be seen here:

| | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | Park | Plaza | Shopping Mall | Metro Station | Amphitheater | Italian Restaurant | Clothing Store | Market | Persian Restaurant | Cosmetics Shop |
| 1 | 1 | Furniture / Home Store | American Restaurant | Mobile Phone Shop | Moroccan Restaurant | Mosque | Mountain | Movie Theater | Multiplex | Museum | Music Store |
| 2 | 2 | Persian Restaurant | American Restaurant | Notary | Monument / Landmark | Moroccan Restaurant | Mosque | Mountain | Movie Theater | Multiplex | Museum |
| 3 | 3 | Persian Restaurant | Café | Fast Food Restaurant | Pastry Shop | Plaza | Shopping Mall | Bakery | Gym / Fitness Center | Park | Restaurant |
| 4 | 4 | Café | Fast Food Restaurant | Park | Persian Restaurant | Coffee Shop | Pizza Place | Plaza | Ice Cream Shop | Pastry Shop | Shopping Mall |

From exploring the map and also from examining clusters, cluster 4 is the best candidate for establishing the office. This cluster shows high frequency of green environment and different restaurants and also from the map it is clear that the most of neighborhoods of this cluster are close to downtown.

## 4. Conclusion

Now, by exploring the neighborhood of cluster 4 and assuming the main factors which have mentioned earlier, Resalat (district) is the best place for the installing of office. In this neighborhood, venues like Park, Hotel, Bus stop and Restaurant have high frequency. Therefore, the office would be close to hotels, restaurants and also public transports as well as parks would provide green environment.

It is important to consider that amount of data about Tehran is lower than other cities in the internet. So the result can be more accurate with more data. For future studies, one can consider other features such as population density, the rate of crime, and the price of offices for each neighborhood.