IBM Applied Data Science Capstone

Venues data analysis in Tehran

Author : Sahand Niasti

Date : September 2019

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.
- Question : , if a property developer is looking to open a new café or restaurant, where would you recommend that they open it?

Data Description

- **Foursquare**: It is a local search-and-discovery service which provides information on different types of entertainment, drinking and dining venues.
- **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.
- **Google Map**: I also used 'Search Nearby' option to get the center coordinates of each Borough

Methodology

1. 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.
2. 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.
3. 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.

Methodology - 1

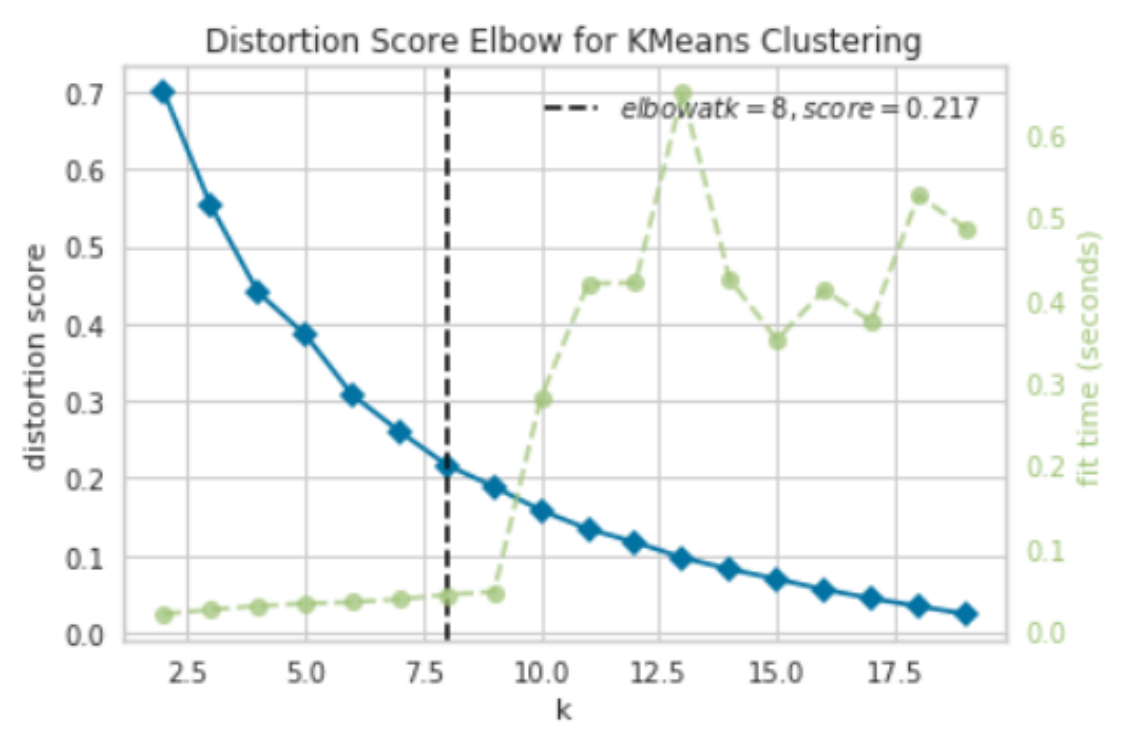
map of Tehran with boroughs superimposed on top

Methodology - 2

the popularity of venues in each borough

	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

Methodology - 3



Data analyzing with Elbow method to figure out the optimum K number

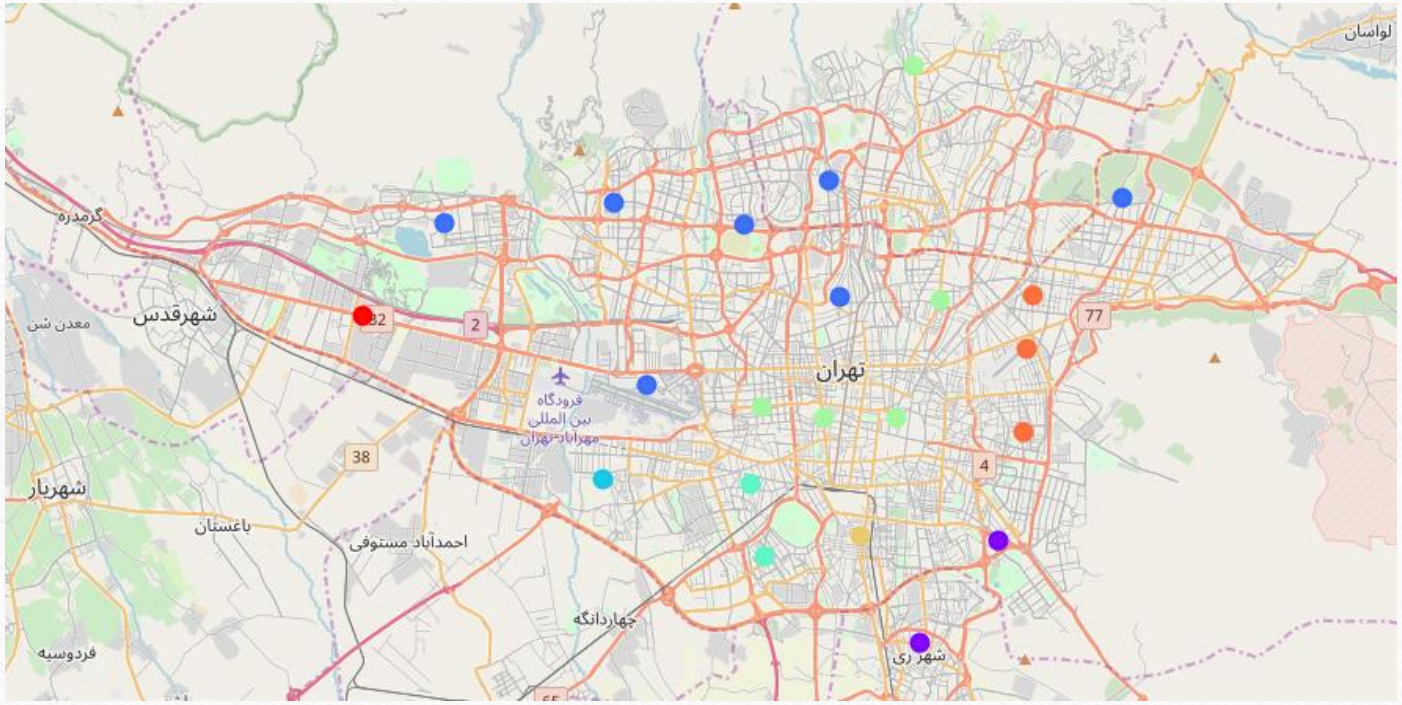
Results

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 :

1. cluster 0 – Red - “stables and auto workshops”
2. Cluster 1 – Purple – “middle east restaurants & pastry shops”
3. Cluster 2 – Navy Blue – “Café & restaurant”
4. Cluster 3 – Sky Blue - “Plaza & shopping mall”
5. Cluster 4 – Cyan – “furniture & park”
6. Cluster 5 – Green – “café with entertaining venues “
7. Cluster 6 – Yellow – “BBQ Joint “
8. Cluster 7 - Orange – “Plaza “

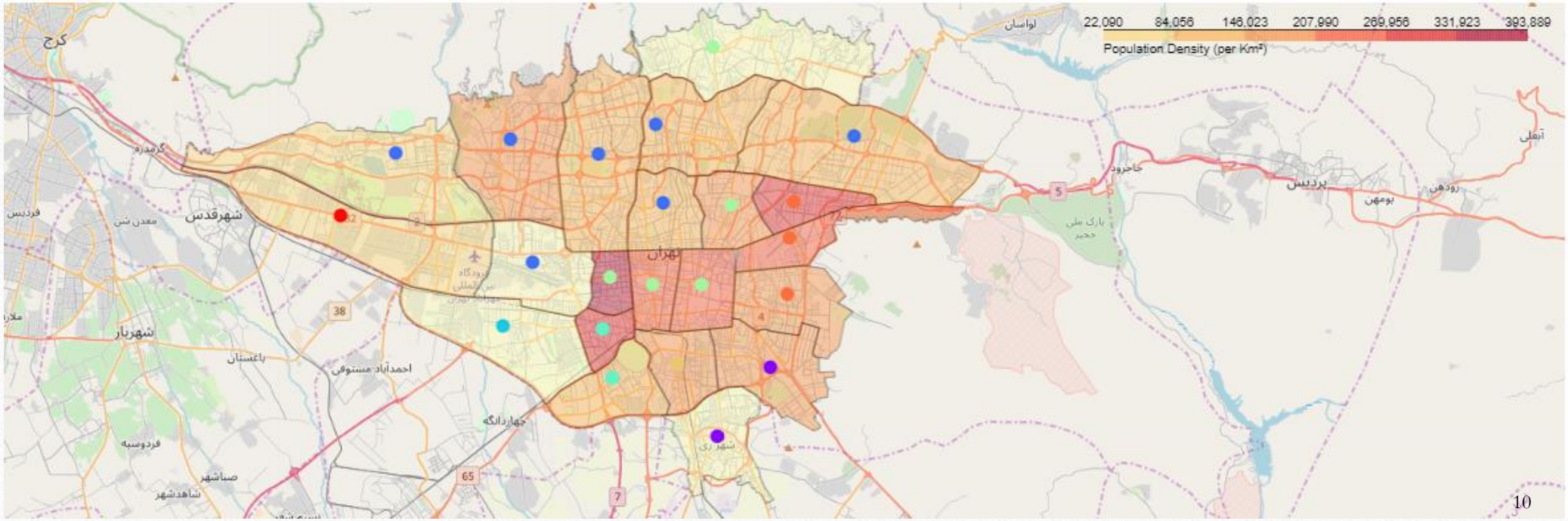
Results

Clustered map of Tehran



Results

Clustered map merged with population density



Discussion

- 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.
- 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.

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.
- 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.
- This project also 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.

The end