# IBM Data Science Coursera Capstone Project

## Introduction: Business Problem

İstanbul is the most crowded city in Turkey and almost 20million people live there. Also it is very important destination for tourism. And last decade mostly Arabic,Persian people come İstanbul and they like shopping especially at the mall and go to restaurant.

That is the reason, location of the shopping mall and restaurant is the most important decision that will support that will be a success or not.

This project will attempt to explore patterns of subdistricts within İstanbul by categorizing them into clusters in order to identify existing trends within neighborhoods of İstanbul. From there on, recommendations can be made on which category of neighborhood will be most suitable for a certain type of venue to be opened.

The result of this project is aimed at general entrepreneur but may be most useful for entrepreneurs on the food and beverage sector given that location can be the deciding factor for a success.

## Data

To analyze trends in İstanbul’s subdistrict, the list of subdistrict is obtained from below link;

<https://en.wikipedia.org/wiki/List_of_districts_of_Istanbul>

Venue queries will then be made by subdistricts using FourSquare APIs. The resulting data regarding venue category will be used to observe commonality between subdistricts. The commonality clusters can then provide insight on which type of venue will thrive better on which cluster.

In summary, the following data is required to meet the objective:

* Subdistricts of İstanbul
* Coordinates of these subdistricts
* Trending Venues on the area
* Venue categories

## Methodology

Given that our objective is to generally categorize the subdistricts, we will use K-means clustering algorithm to categorize each of the subdistricts within İstanbul.

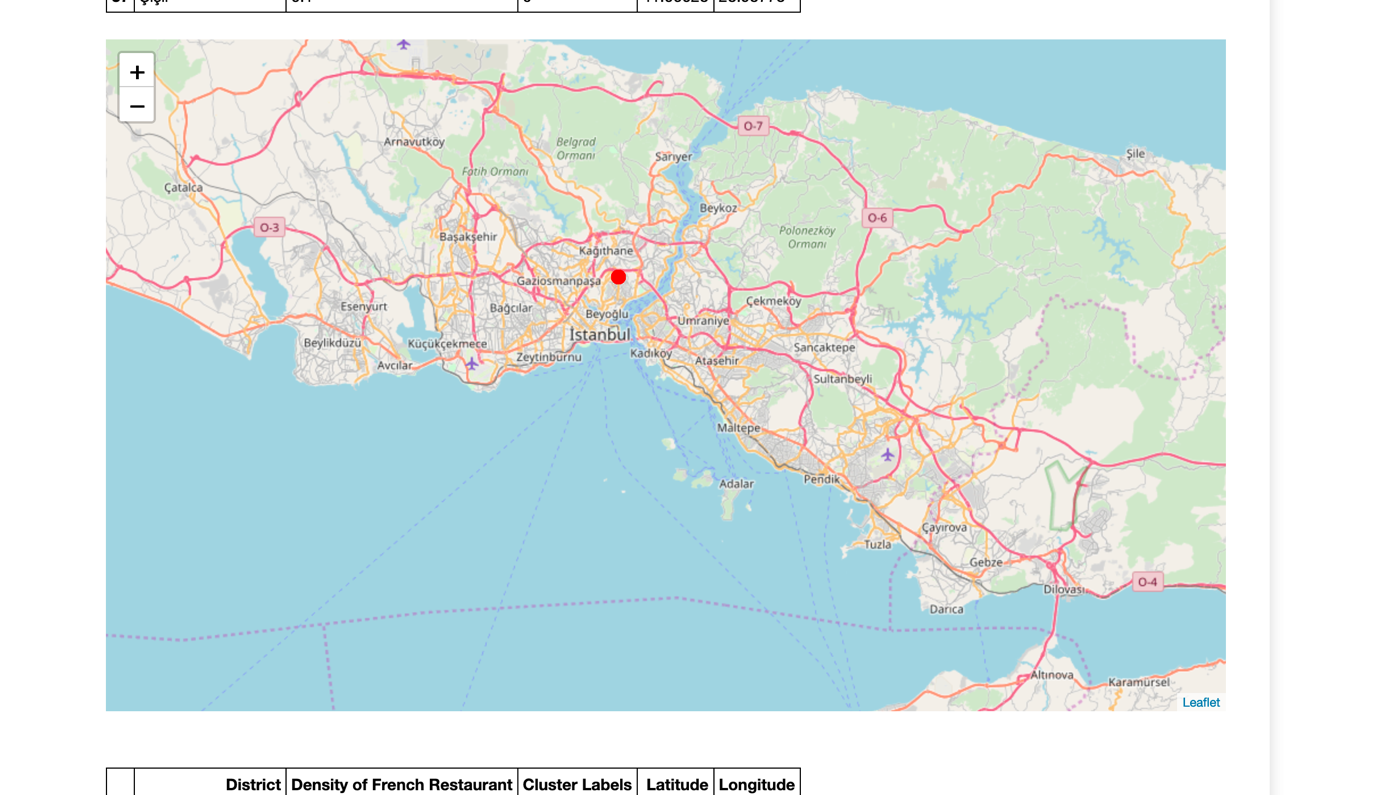
A one-hot encoding will be done on the venue dataframe and it will be grouped by subdistrict. The encoding will return venue categories as column per subdistrict, which will then be grouped to provide weighting of venue type occurence on each subdistrict.

The encoded dataframe will be further filtered into top venues before the K-means clustering algorithm will be run over it. This will return cluster labels over the subdistricts. The clusters will be observed one by one manually to determine its content.

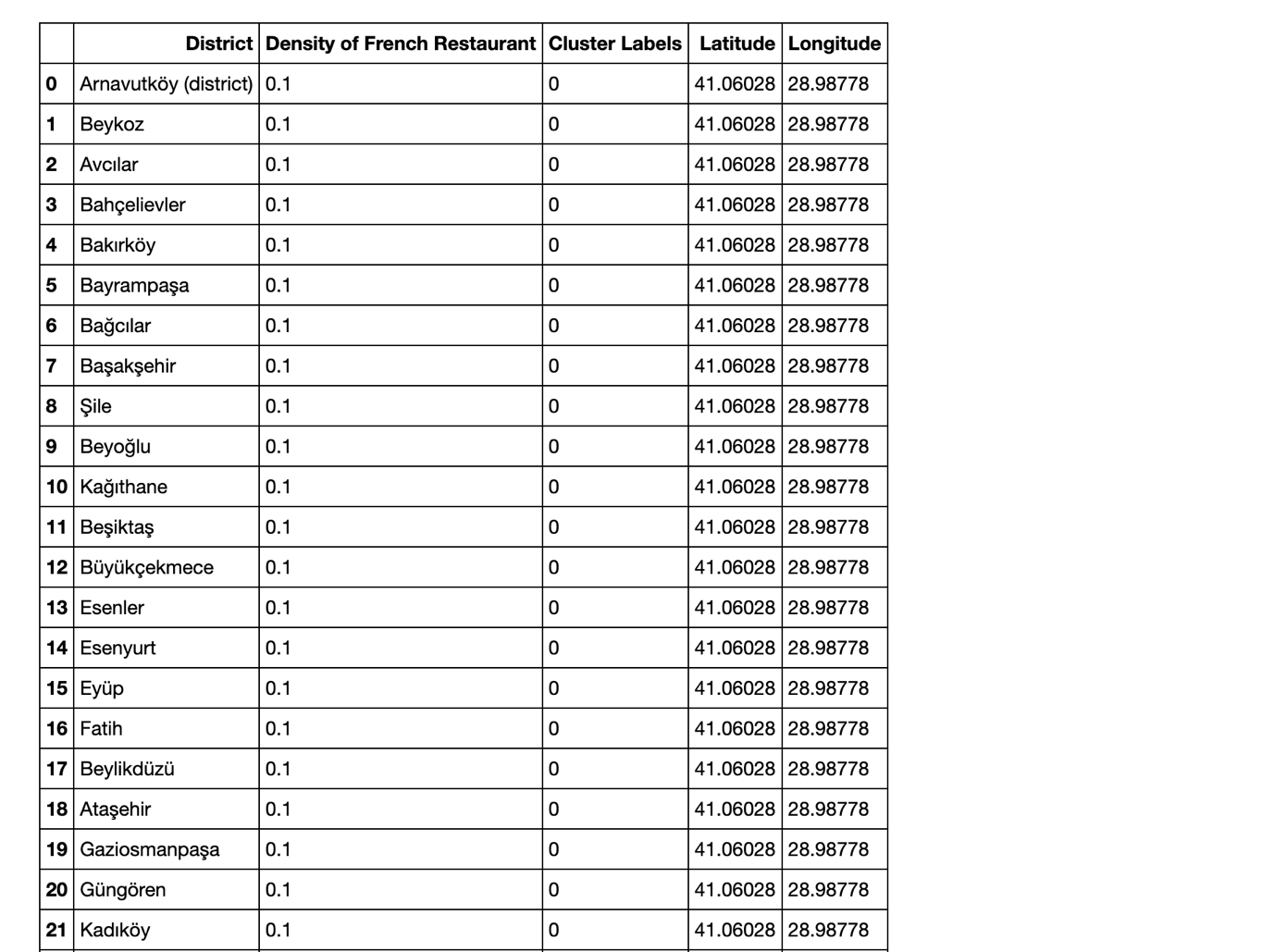
Recommendation will be made based on the clusterring.

## Analysis

### Map of Clusters

****

**Cluster 0**



**Cluster 1**

## Result and Discussion

Groupings as a result of K-means clustering algorithm tallies with how İstanbul historically develops. Having most of cluster 1 does not have French restaurant and we can invest cluster 1 for that.

## Conclusion

Opening of new French restaurant may be best done in cluster 1 where there are less of such restaurant to compete.