IBM Data Science Capstone Blog Post

# The Case

Istanbul is one of the largest and most touristic cities of the World. Surely, a restaurant opened here will attract many customers and be successful. In this project, I am going to analyze the venues in Istanbul by district and try to find the optimal to place to open a restaurant looking at the already existing number of restaurants in each district.

# The Data

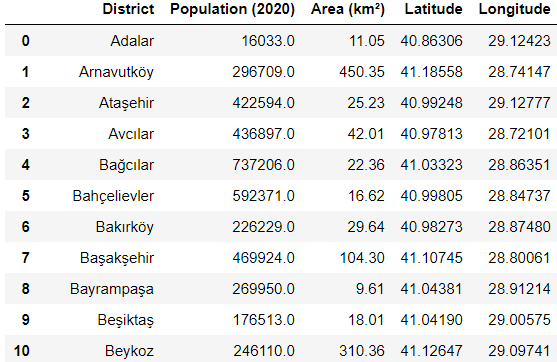
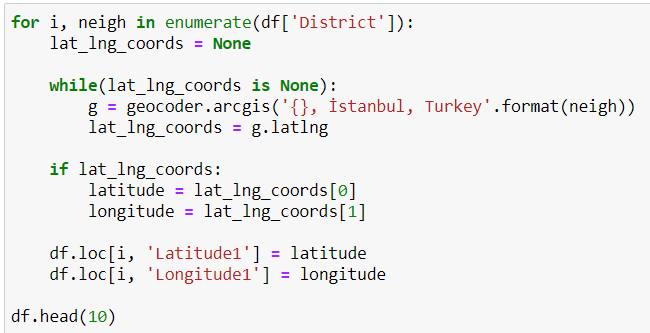
## Dataset

Although there were many datasets where I could get the data from, I decided to use the data from Wikipedia for each district for 2 reasons: First, Wikipedia is a reliable data source with a minimum amount of missing values. Second, this particular data set contained information about household income and population density which will be useful for the second part of the project. The dataset contained non- important data such as the population and household income which were all removed before starting the analysis.

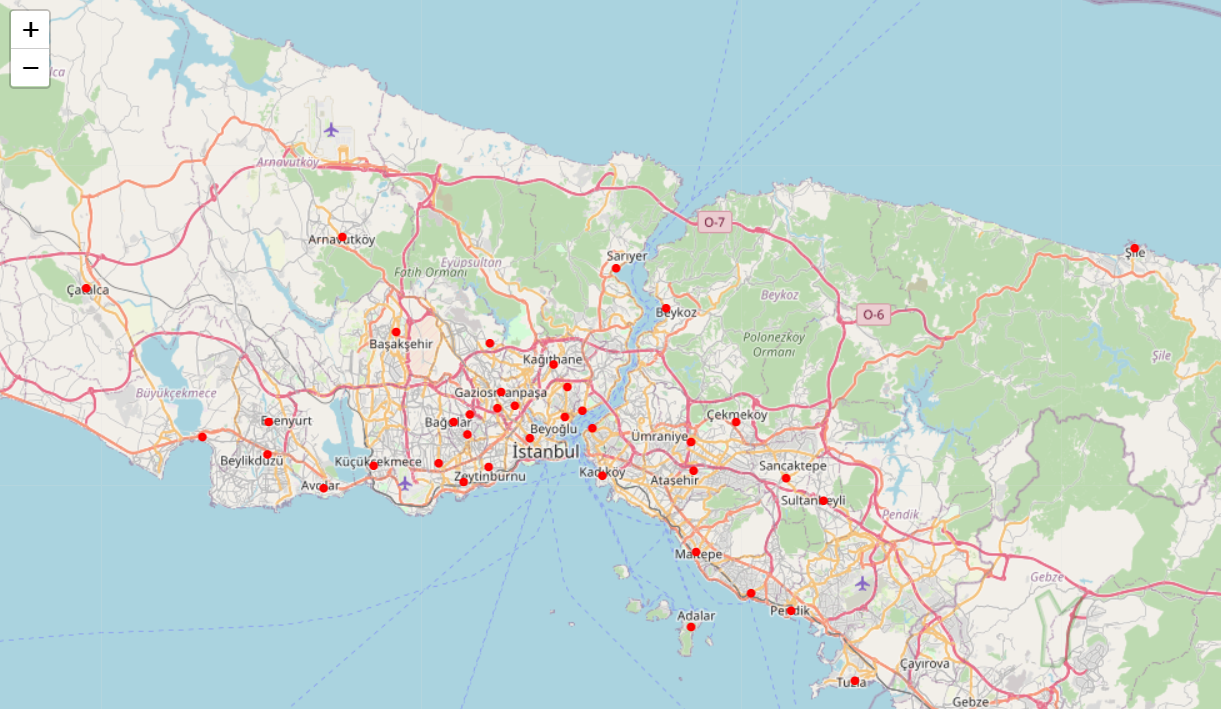
## Venues Data

Regarding the location and venues, data was extracted from Foursquare API, one of the most reliable data sources for venues all around the world. From Foursquare, I extracted the top 200 venues of each district in Istanbul. There are a total of 3659 venues in 36 districts in Istanbul. The function used to return nearby venues from foursquare API can be seen below.

## Coordinate Data

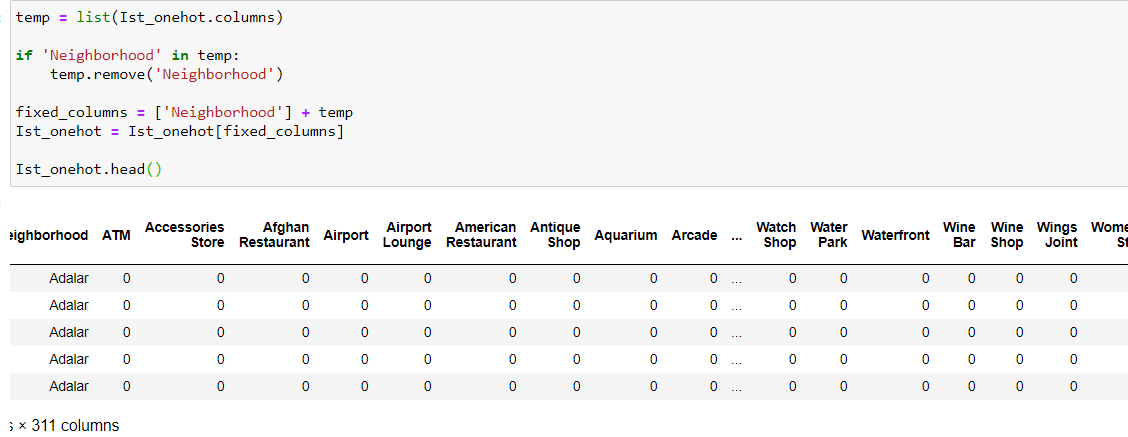
The latitude and longitude coordinates are necessary to be able to redeem data from Foursquare API and to draw maps using folium. In the dataframe extracted from Folium, The coordinates of the districts were absent, hence the library Geocoder was used to extract the coordinates of each district in Istanbul.

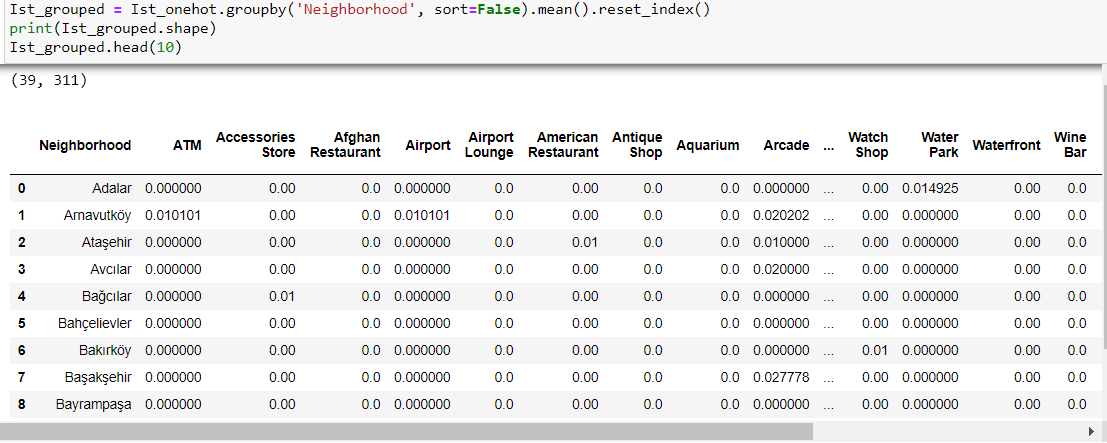
## Folium Mapping

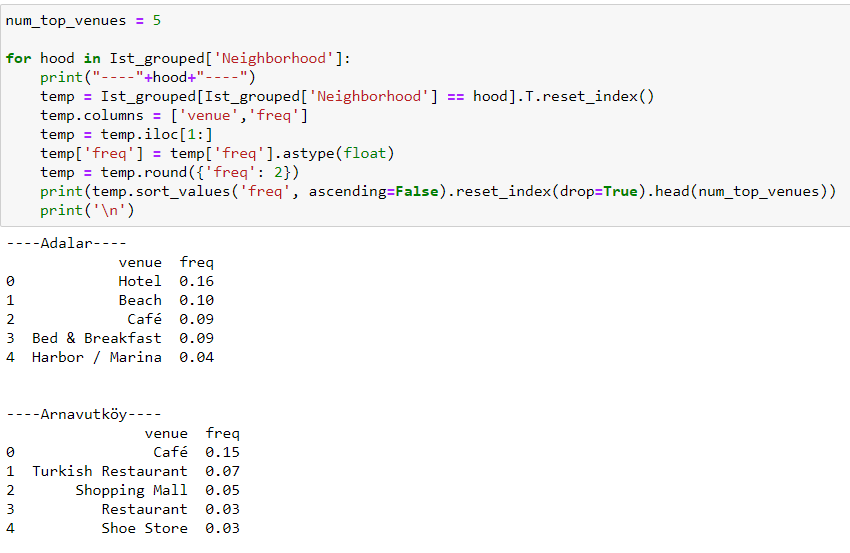
 The map below is plotted using the folium mapmaker library in Python. The Districts of Istanbul are marked in red in the map below.

## One-hot Encoding Process

In machine learning, one-hot encoding is a frequently used method to deal with categorical data. Because many machine learning models need their input variables to be numeric, categorical variables need to be transformed in the pre-processing part. In our project, we apply one-hot encoding to our dataset in order to prepare it for the K-means clustering algorithm that we are going to use in order to get insights about the optimal district to open up our restaurant in İstanbul.





As seen above, the dataset is mostly filled with zero since most of the venues aren’t contained in districts, hence we sort the dataframe to find the most common attractions within each district. We do this by the following code: 

Converting the data into a pandas data frame, we obtain the following table that contains the most common venue in each district in Istanbul: 

Looking at the table, we obtain insights on the most popular venues by each region, this data will help us select a district in which restaurants are more popular so that we have a higher chance of the restaurant we opened to succeed.

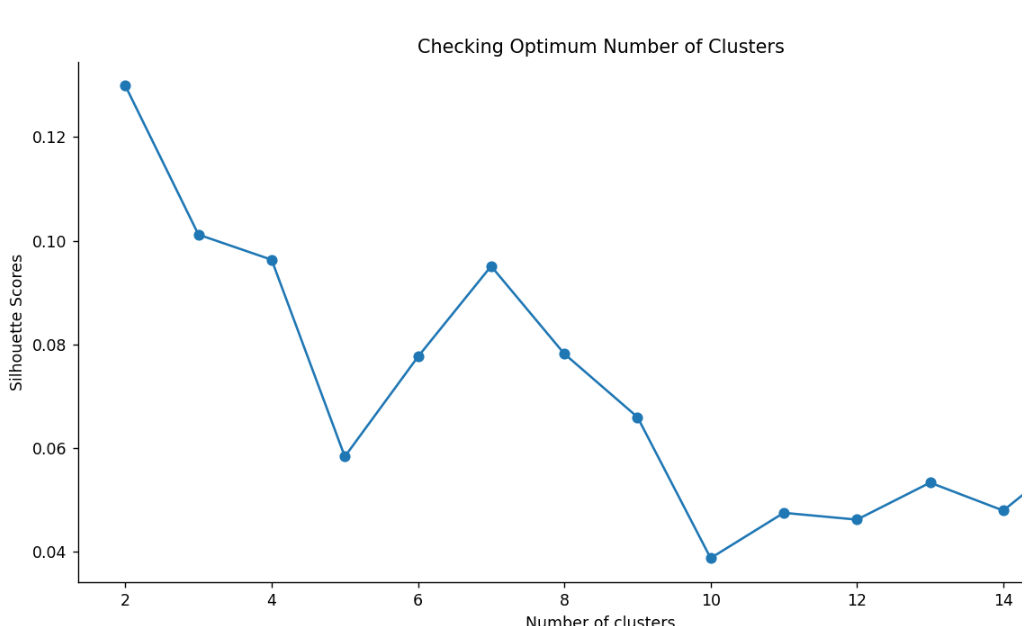
## Clustering Neighborhoods

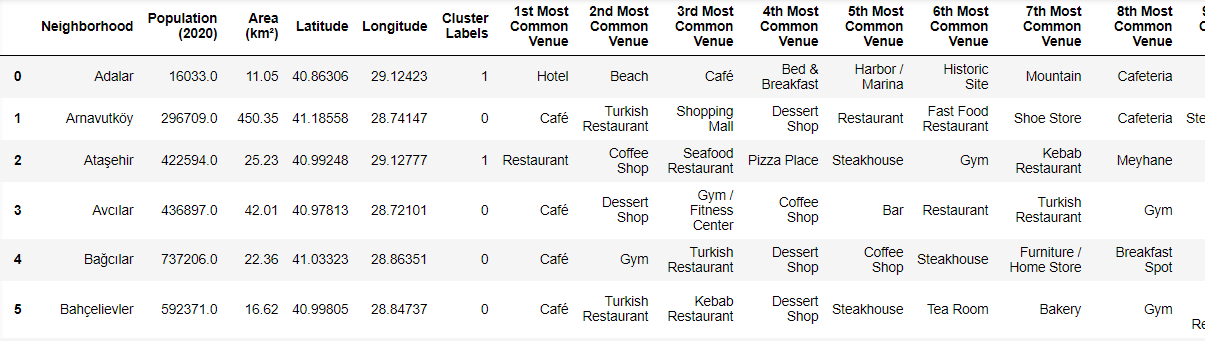
After wrangling the data and making it suitable to use in a clustering machine learning algorithm, Now we can use KMeans clustering method to cluster the neighborhoods.

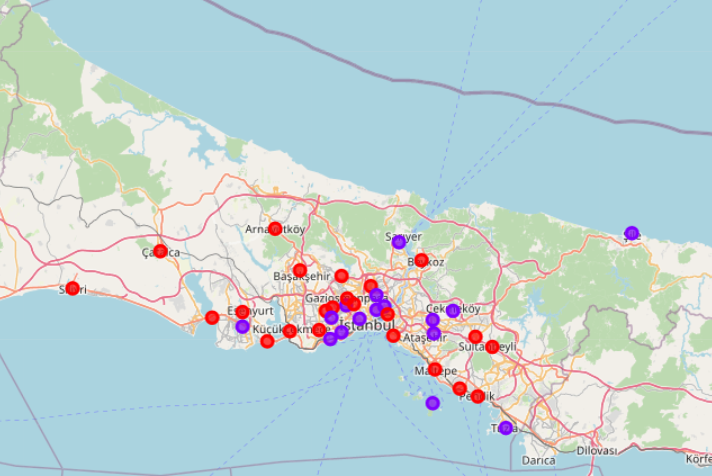
First, we need to determine how many clusters to use. This will be done using the Silhouette Score. We will define a function to plot the Silhouette Score that will be calculated using different number of clusters.

We run the following code to obtain the optimal number of clusters to use in our clustering analysis:



The output graph shows us that the optimal number of clusters to use in our clustering analysis of 36 districts are 2. So we split the data into 2 clusters, one in which opening restaurants is optimal, and the other in which it isn’t.



The column ‘Cluster Labels’ shows the districts with a higher probability of success when a restaurant is opened (denoted with 1) and the others with zero. On the map, we can visually see which areas are optimal to open a restaurant: 

From our analysis, we conclude that the areas marked with purple are the districts in which opening a restaurant will have a higher chance of success compared to the red areas. However, it is important to note that there are many other factors that come into play such as cost efficiency, transportation ways of transportation and so on.