Clustering Districts of Istanbul

Ugur Savci

Introduction

Istanbul is a big city with huge population and great places to live. But with many options to open a restaurant, how we can determine the best places to open a restaurant in exactly true location that is supported with data science?

In this project, Our stakeholders want to open a restaurant but they do not have any idea where to open it . Their restaurant is a fast-food chain that is really famous over the world. They want to open their restaurants first in Istanbul,since Istanbul is the biggest city in Turkey with their capacity and population and other advantages.

We have been tasked as a data scientist to analyze and group boroughs with their similarity and create insight for the company.

At the end of the project. We will get an insight of each Borough and group them by their similarity.

1.Data

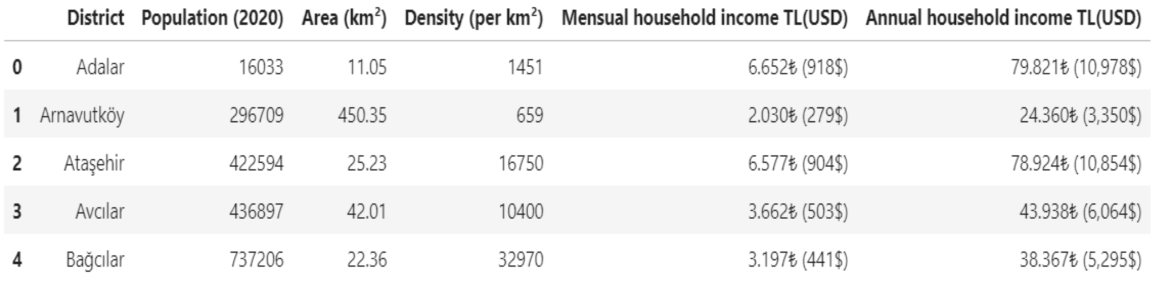
We will use following data for our project ;

1. Borough's general information that includes population,annual income etc. From Wikipedia(https://en.wikipedia.org/wiki/List_of_districts_of_Istanbul)
2. Geopy Library to get location information of boroughs.
3. Foursquare Api will be used to get information of venues.

2.Methodology

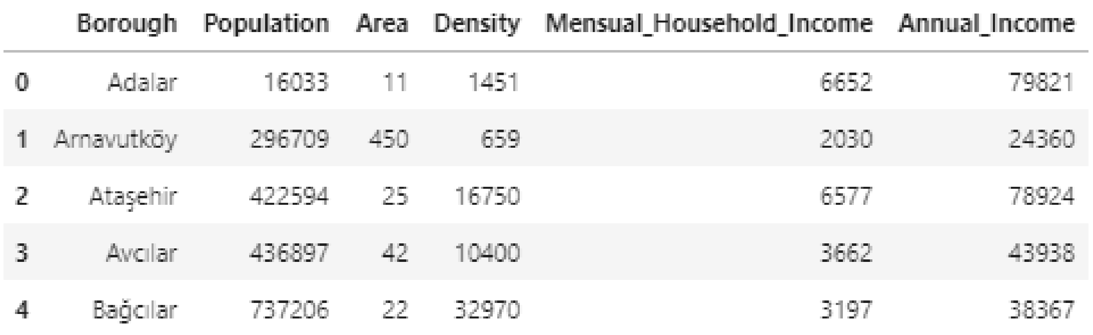
In this study I started my project by downloading and importing required libraries.After getting related libraries, I used web scrabing to get information of Boroughs (Population, Area, Density, Mensual Household Income and Annual Income).

After I utilized boroughs , I created dataframe to make the data tidy format.



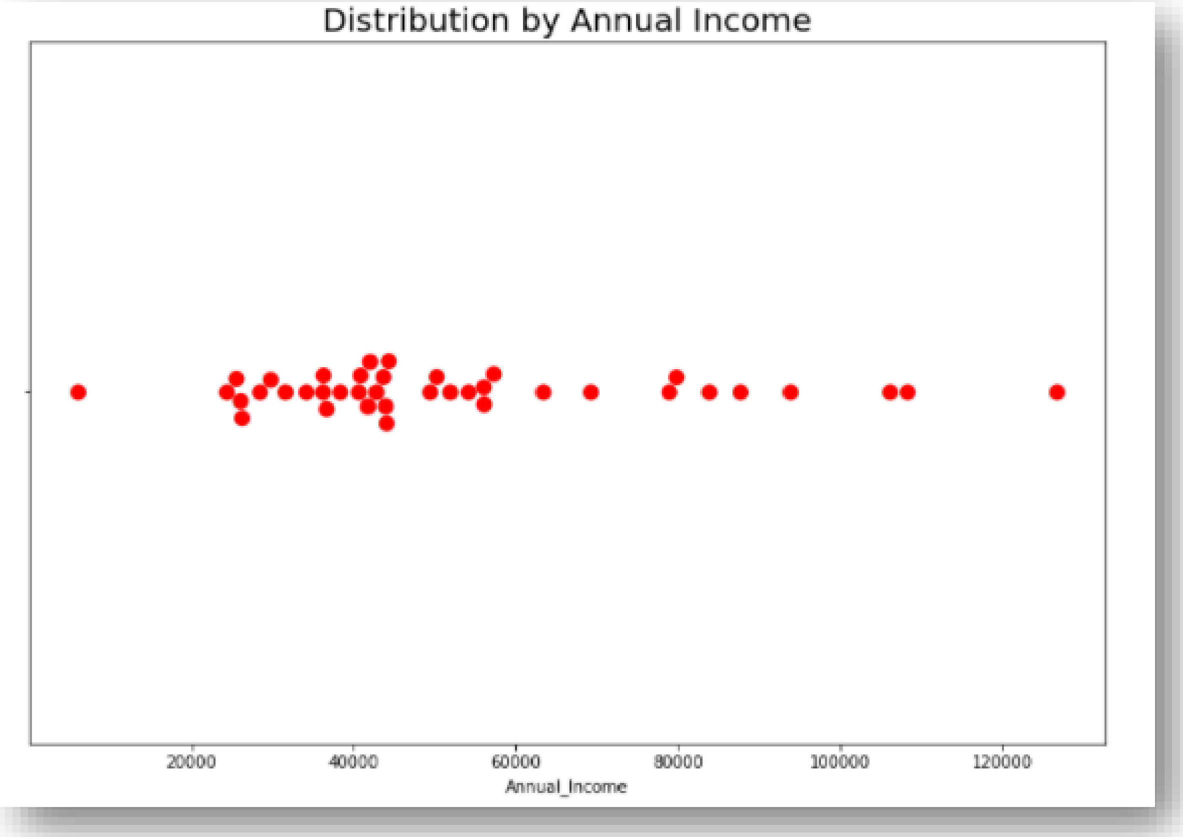
	District	Population (2020)	Area (km^2)	Density (per km^2)	Mensual household income TL(USD)	Annual household income TL(USD)
0	Adalar	16033	11.05	1451	6.652₺ (918$)	79.821₺ (10,978$)
1	Arnavutköy	296709	450.35	659	2.030₺ (279$)	24.360₺ (3,350$)
2	Ataşehir	422594	25.23	16750	6.577₺ (904$)	78.924₺ (10,854$)
3	Avcılar	436897	42.01	10400	3.662₺ (503$)	43.938₺ (6,064$)
4	Bağcılar	737206	22.36	32970	3.197₺ (441$)	38.367₺ (5,295$)

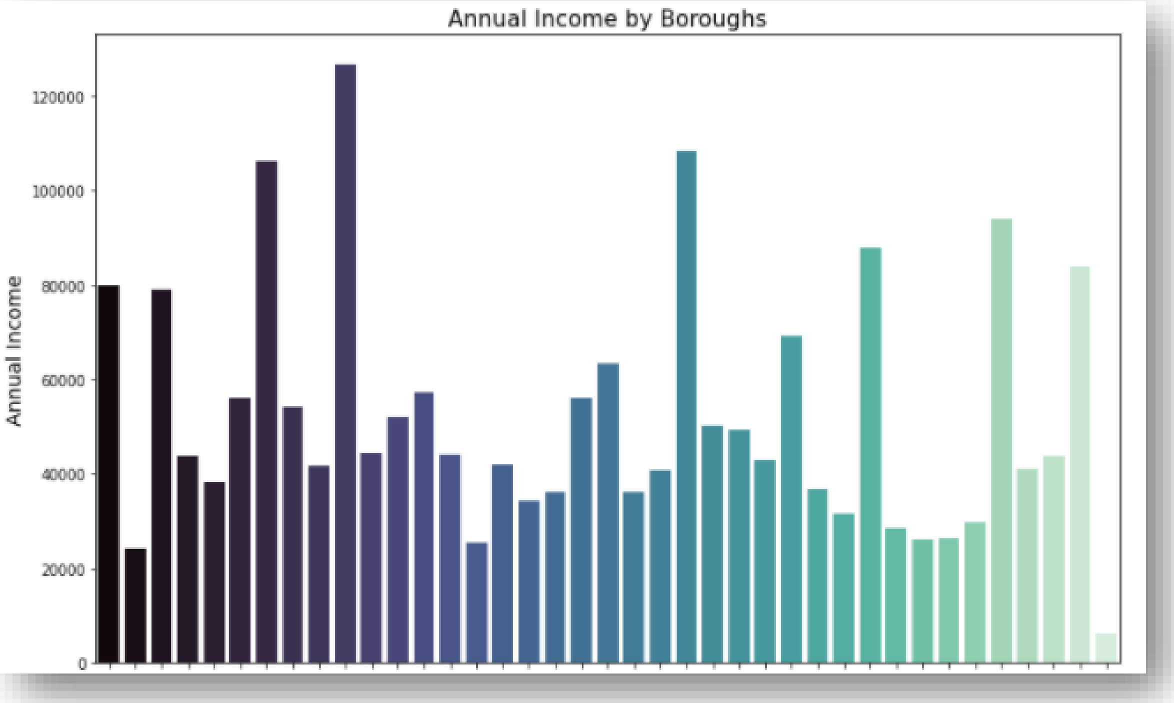
As we can see from dataframe, we need to clean the “Mensual household income” and “Annual Income” column, I changed the column name to be more clear and drop unnecessary rows.



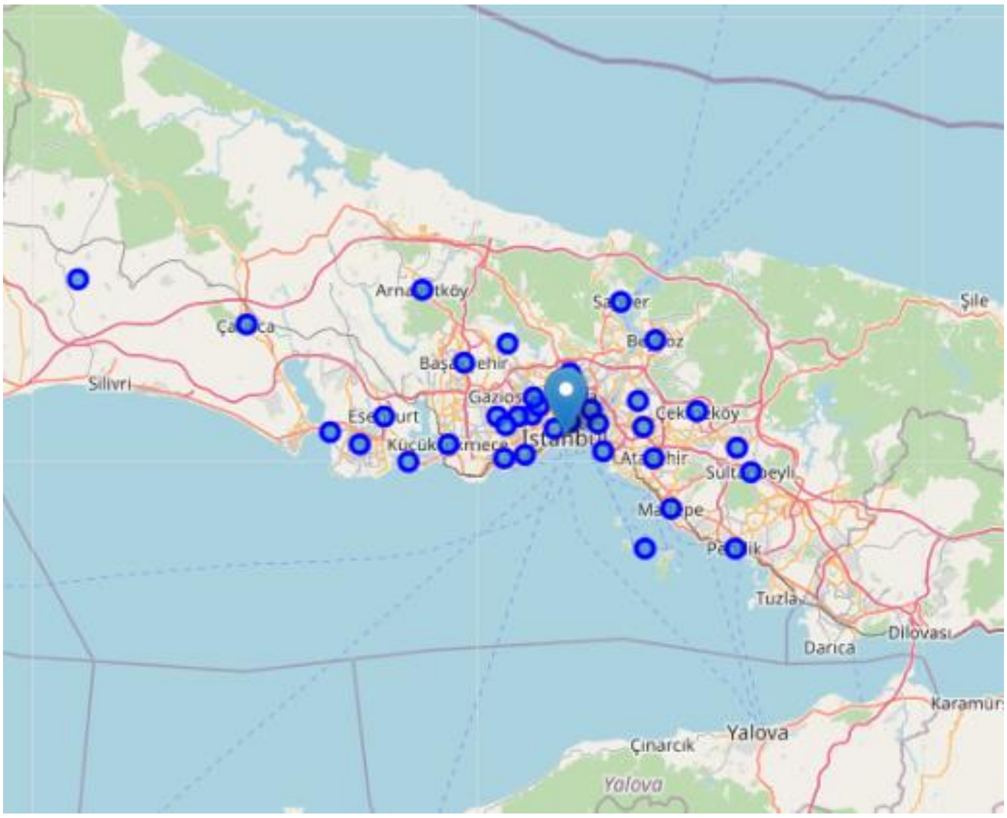
	Borough	Population	Area	Density	Mensual_Household_Income	Annual_Income
0	Adalar	16033	11	1451	6652	79821
1	Arnavutköy	296709	450	659	2030	24360
2	Ataşehir	422594	25	16750	6577	78924
3	Avcılar	436897	42	10400	3662	43938
4	Bağcılar	737206	22	32970	3197	38367

After dataframe has been cleaned and organized. Some exploratory data analysis has been performed.



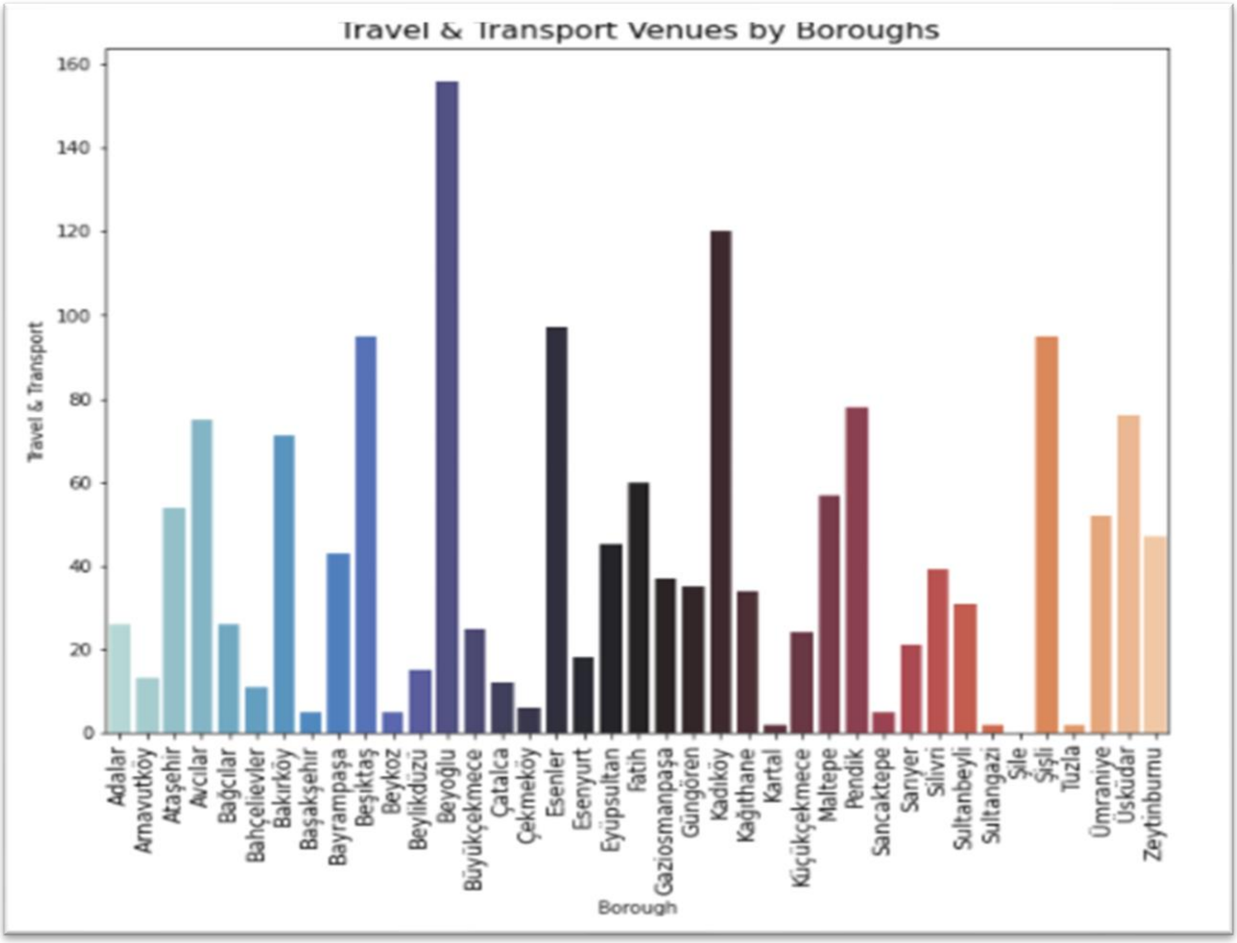


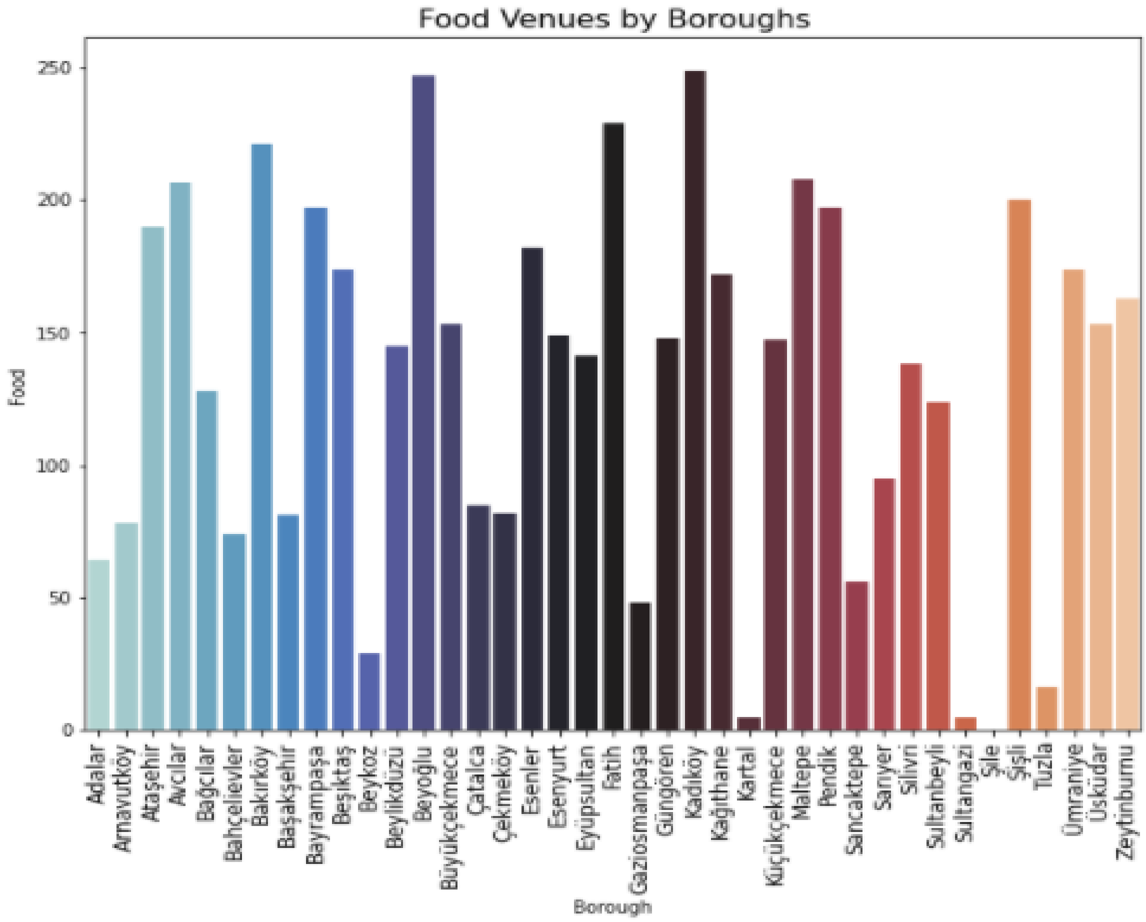
I used python geopy library to get the location of each borough folium library to visualize the locations.



After visualization, I connected the Foursquare API to obtain venues information by categories

I added this information to my data and saved as a new dataset.To understand our venues i did some exploratory data analysis.With this information I was able to infer something about venues and boroughs together.


Travel & Transport Venues by Boroughs
Travel & Transport
0
20
40
60
80
100
120
140
160
Adalar
Arnavutköy
Ataşehir
Avcılar
Bağcılar
Bahçelievler
Bakırköy
Başakşehir
Bayrampaşa
Beşiktaş
Beykoz
Beylikdüzü
Beyoğlu
Büyükçekmece
Çatalca
Çekmeköy
Esenler
Esenyurt
Eyüpsultan
Fatih
Gaziosmanpaşa
Güngören
Kadıköy
Kağıthane
Kartal
Küçükçekmece
Maltepe
Pendik
Sancaktepe
Sarıyer
Silivri
Sultanbeyli
Sultangazi
Şile
Şişli
Tuzla
Ümraniye
Üsküdar
Zeytinburnu
Borough

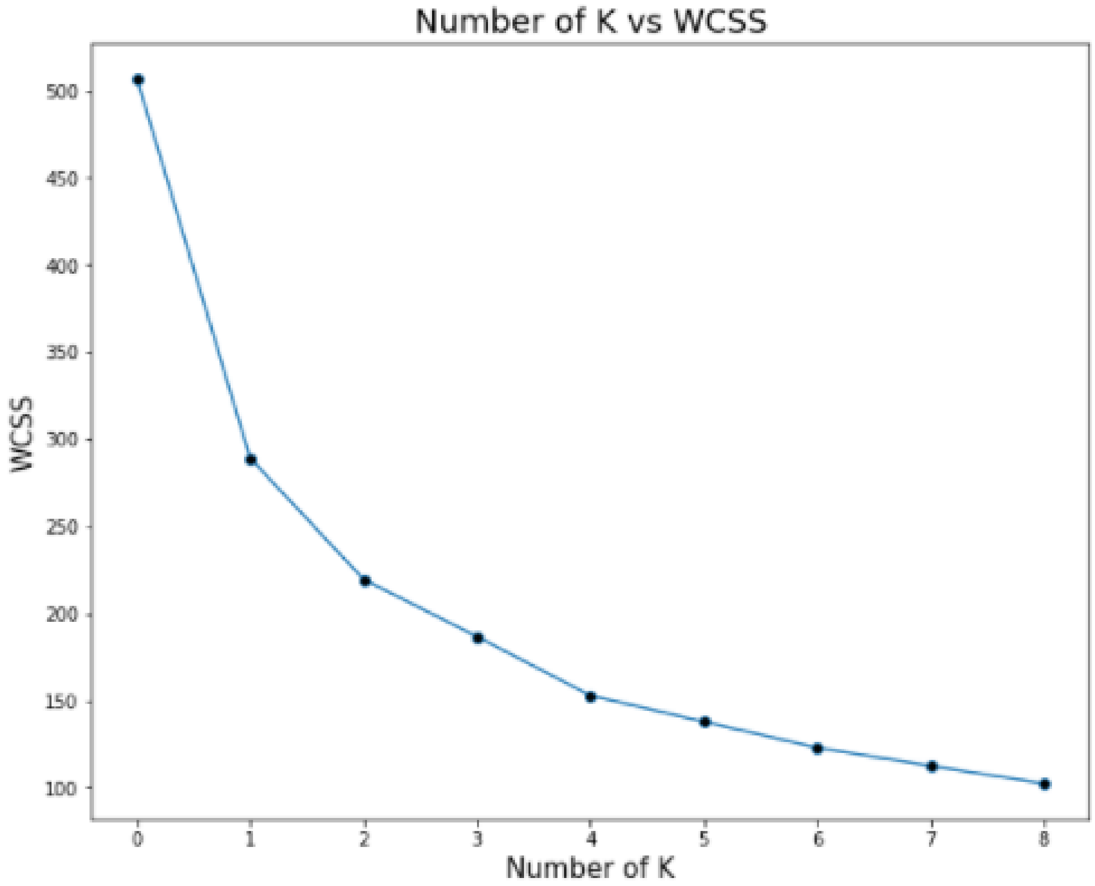


We have some common venue categories in boroughs. In this reason I used unsupervised learning algorithm. K-Means algorithm will help us group boroughs with their similar features.

Before We run the model. I dropped some variable and scale our new dataset.

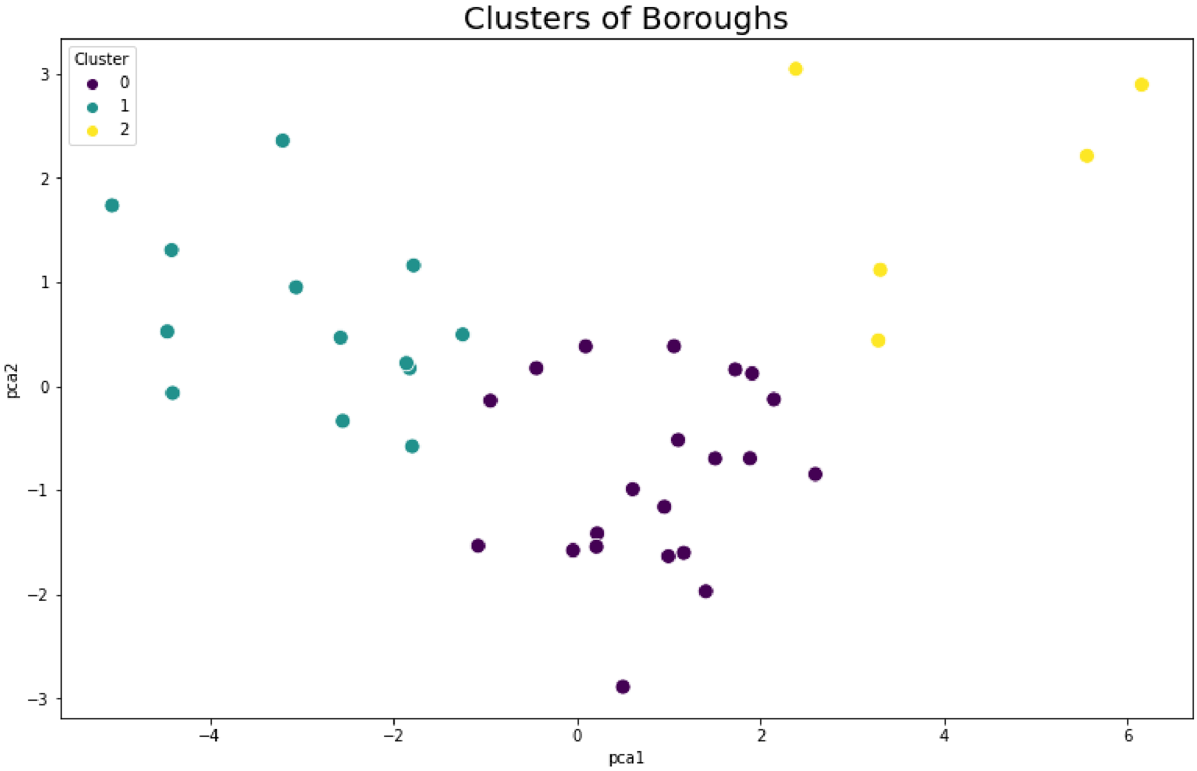
To select our best group , Elbow Method is used to choose optimal group for our dataset.

In this project I used k = 3 although elbow method says that number = 2 is the best. I wanted to group our boroughs more and there is no significantly difference between 2 and 3.



After chosing the k value , I run our model and get the labels. We can now create a new dataset with our label.

Our dataset is ready to be visualized . We can also filter dataset by their group.



3.Results

Cluster 0 : This group has higher number of venues and annual income. There are many Transportation ,Food Venues,Nightlife Spot in this group.If our menu is expensive. We can choose to start with this group.

Cluster 1 : Boroughs in this group has higher population compared to cluster 2. There are many food venue. Annual income of this group is average of entire boroughs.It is a middle segment group of our analysis.

Cluster 2 : This group has the lowest venue and people who live in these boroughs has average annual.Their population is also lower than others in general.

4.Discussion

In this project , We used K-Means algorithm to cluster our dataset.K-Means clustering algorithm is moslty used algorithm amongst Unsupervised Learning.

After we explore our dataset with venues information. With this study , We can get important information based on Boroughs.

To get business decision to open a restaurant , We could also expand our dataset with venues name,population details etc. to target more specifically.

This project also can be used for people who have never been to Istanbul but want to know basic information about venues, type of boroughs.

5.Conclusion

In this study, I analyzed districts of Istanbul and clustered boroughs based on their similarity. This algorithm can help us understand each group of boroughs and their similarity.With this study,we can understand pattern of each group and get insight about each district.In terms of people who consider open a restaurant in Istanbul , this study helps them understand district and their similarity. For example If we want to open a restaurant in Kadıköy, we can also compare Kadıköy to other districts in the same cluster.