An IBM Data Science Capstone Project

Classification Of Venues Around Istanbul Metro Stations

Submitted by: Karla Marie I. De Jesus

Introduction

Istanbul – A bustling metropolis of 13 million people. The city of Byzantium or Constantinople, as this city was previously called, sits in the center of the world. It is a city located on the boundaries of Europe and Asia, with the Bosphorus Strait dividing the European and Asian side. Istanbul is never shy of stories to tell when it comes to history, as it has been the subject of many, many wars and rebellions just to get control of this strategic and wealthy city. From then until today, it hosts a mix of people of different cultural backgrounds, religion, even economic status.

But each of these 15 million people has places to go to, every single day. So how does Istanbul manage to move 15 million people every day? The answer mainly lies on its very effective and convenient public transportation system. Although Istanbul is also proud of its Metrobus, Bus, Ferry, Tramway and Funicular systems, this paper will only focus on Metro Lines around the city. There are, in total, 84 Metro stations around the city. There are currently six working metro lines across the metropolis, called M1, M2, M3, M4, M5 and M6. However, constructions are ongoing for M7, M8 and M9, all envisioned to be completed by 2023.

The Metro operates daily from 06:00- 00:00, while M1, M2, M4, M5 and M6 lines were declared to be operating at 24 hours during weekends and holidays. However, due to the Corona Virus, 24 hour operations were temporarily discontinued by the authorities. A flat fare of 3.50 TL is charged from the passenger by paying with a reloadable fare card called Istanbul Kart or by purchasing tokens from machines in metro stations. Istanbul Kart is the universal fare card in Istanbul which can also be used in busses, ferries, funiculars and trams.

I chose to do this study because it will be very helpful for entrepreneurs who are interested in finding the perfect spot for their intended business. By understanding the types of venues that are in a certain area, business people can easily evaluate the potentials and disadvantages of each depending on their requirements. Likewise, individuals who are looking for a good area to live or work can also find helpful insights from the clustering and classification that I will conduct throughout this study.

Data

This paper will focus on the 89 stations of 6 metro lines. By using Foursquare location data, I will explore the different establishments located on a 750-meter radius around each station. Each establishment is categorized by Foursquare accordingly. By identifying the types of establishments and its category, we will be able to identify which areas are excellent for a specific type of business. Some of the main categories to be used in this research are the following: Arts and Entertainment, College and University, Event, Food, Nightlife Spot, Outdoors and Recreation, Professional and Other Places, Residence, Shop and Service and Travel and Transport. These are the Top 10 categories according to Foursquare.

I obtained the coordinates of each metro station from Foursquare and Wikipedia as well as the establishment types around it. From this information, we can cluster each according to category and identify if they are in a desirable location for a new business. Because Istanbul is a very populous city, there is really no distinct classification on areas that are “commercial” or “residential”. What you would normally find are commercial establishments on the ground floor, and above it are residential flats.

Methodology

The main source of data for this study is from Foursquare API, Wikipedia and the official website of Istanbul Metro, www.metro.istanbul. Mainly, I gathered the names of stations and its coordinates. To get a better representation of how the different lines are in relation to the city, I plotted the stations on the map of Istanbul and I assigned them the actual colors of each line.

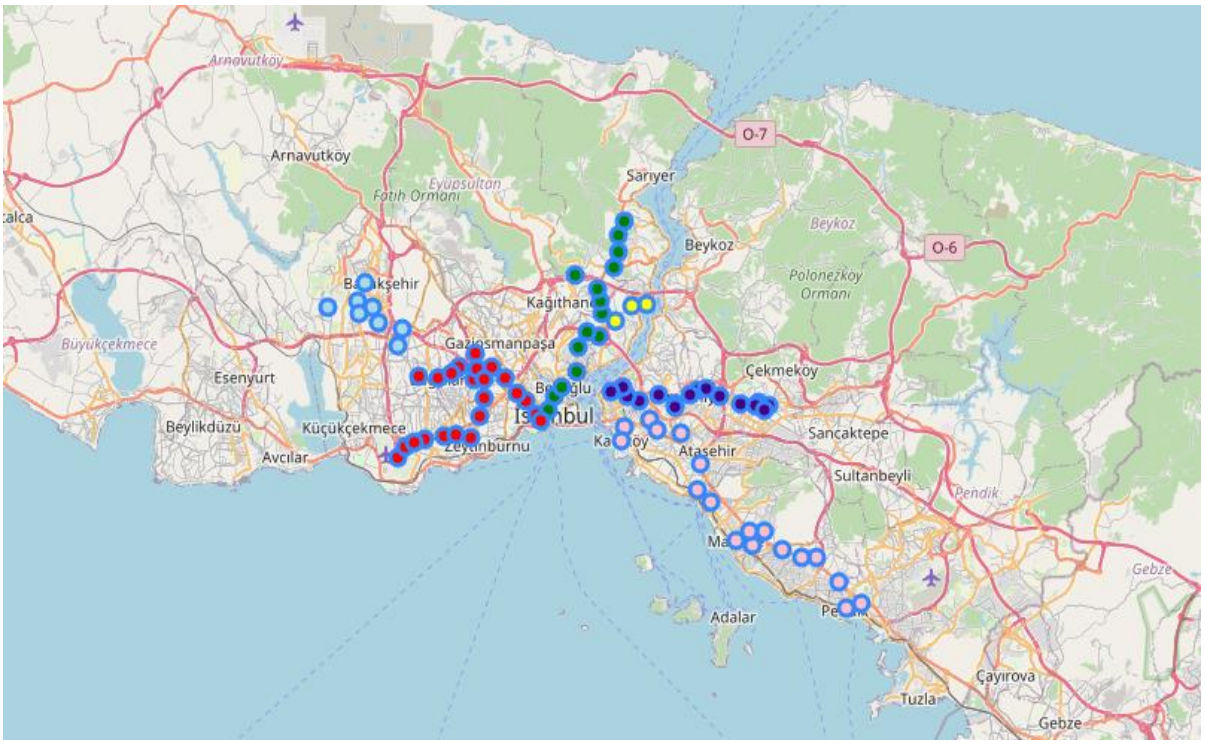

Figure 1. Istanbul Metro Map

Foursquare has many venues categories and even hundreds more sub categories. Although we will only be working with the Top 10 categories mentioned above. I was able to determine the Top 10 categories and its corresponding Category ID. Figure 1 shows the dataframe containing the categories data that I extracted from Foursquare.

	Category	categoryId
0	Arts & Entertainment	4d4b7104d754a06370d81259
1	College & University	4d4b7105d754a06372d81259
2	Event	4d4b7105d754a06373d81259
3	Food	4d4b7105d754a06374d81259
4	Nightlife Spot	4d4b7105d754a06376d81259
5	Outdoors & Recreation	4d4b7105d754a06377d81259
6	Professional & Other Places	4d4b7105d754a06375d81259
7	Residence	4e67e38e036454776db1fb3a
8	Shop & Service	4d4b7105d754a06378d81259
9	Travel & Transport	4d4b7105d754a06379d81259

Figure 2. Categories Dataframe

To determine the count of each venue around each station, a radius limit of 750 meters was set. 750 meters is considered walking distance and not too far from the station anymore. From the data I have, I was able to query from Foursquare the count of each venue type around every station. A JSON file is returned by Foursquare, and by inspecting the JSON, I was able to extract the data that I need specifically. However, when I perused the raw data that we have, it showed that some rows have 0 or missing data. Having 0's and NaN's in our dataset is not advisable because we want to have a clean dataset to be able to come up with the most accurate and precise results. By wrangling and cleaning the data, I managed to fill up the cells with NaN's and zeros. Now that the data is clean and all empty cells have been filled up, the data is ready to be analyzed.

Standard Scaler is the module I used from sklearn to normalize the data and we can compare them with each other logically. K-means clustering is the algorithm used to classify each

into groups with similar features. By using k=4, I categorized each cluster according to the scores we obtained from standardizing and clustering our data.

To visualize the 4 clusters better, I plotted them on the map and assigned colors to represent each cluster.

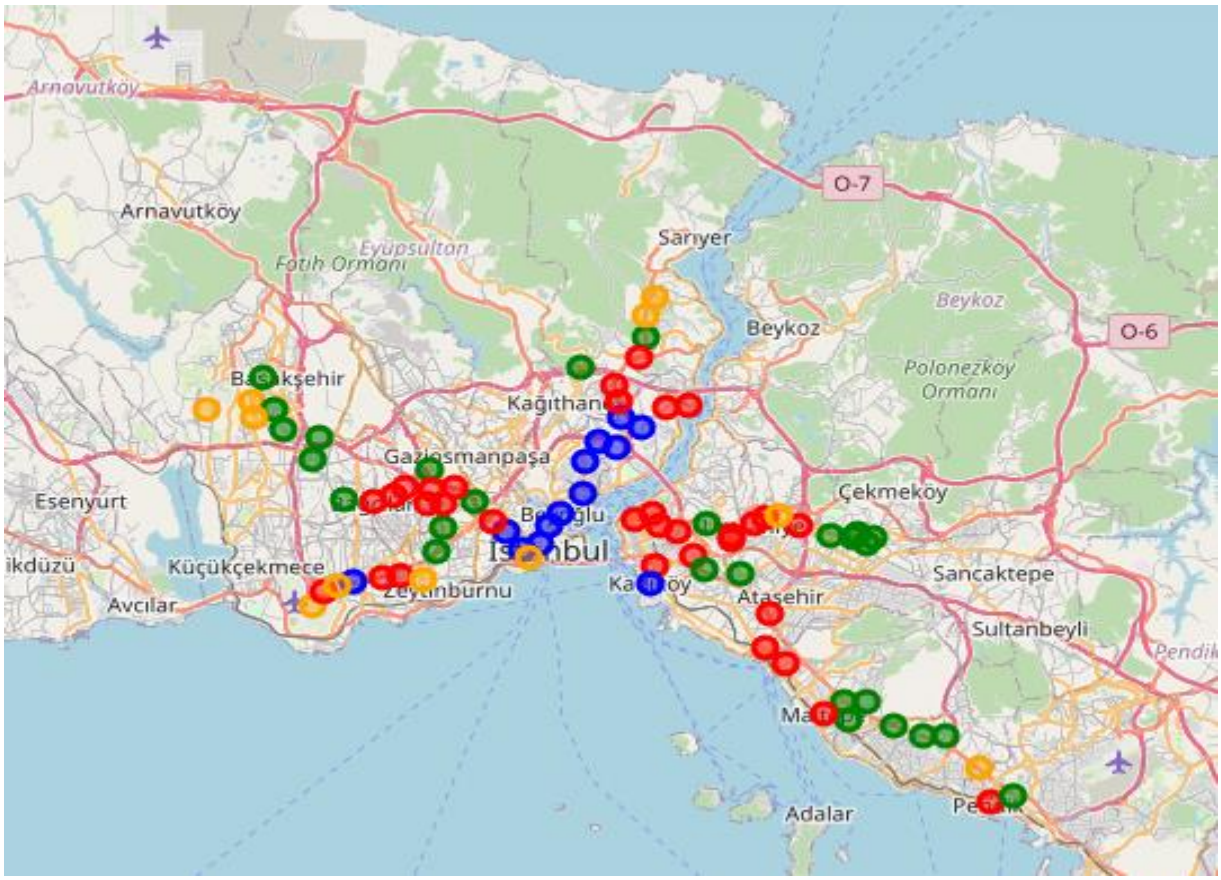

Figure 3. Visualization of Clusters

Results

Four colors represent our clusters, from highest to lowest score.

1. Blue - Areas with high venues count
2. Green - Areas with quite some venues
3. Red- Areas with not so much venues
4. Orange - Areas with very little number of venues

Discussion

By analyzing the results, we can identify that "Food" and "Professional and Other Places" are the most common type of venue category, while there seems to be very little under the "Event" category. It is true that Istanbul is surrounded by restaurants, cafes, pastry shops and similar areas. This is not surprising because Turks are naturally very sociable and eating and drinking tea or coffee with friends and family is their favorite pasttime. Although Istanbul is not the capital of Turkey, Istanbul is still the biggest metropolis in the country, so it is also why offices and commercial buildings is second of the most common venues.

From clustering the venues data, we are able to visualize the city according to the markers on the map above. Blue markers have the highest numbers of venues. This is logical because those parts of the city marked in Blue are the busiest and considered the heart of the city. Green and Red areas are somewhat scattered just outside the perimeters of Blue markers. These green and red areas are mostly residential areas, but they still have their own share of local venues around them. The orange markers on the map are the areas with very little number of venues.

This is because they are located on the outskirts of the city where not so many people live and majority of the land is not yet developed as of this time.

Conclusion

As conclusion to this study, it was found that Blue areas around the city are the best option for business people to establish their business. High venue counts would mean more foot traffic around the area and therefore, more possible customers for a business. Another advantage is that these venues are located walking distance from every Metro station, that would make it very accessible to people coming from all over the city.