IBM-Coursera

Data Science Professional Certificate

Capstone Project-Final Report

Potential Market Analysis of Istanbul Counties

Tuğkan Seçkin - 2020

TABLE OF CONTENTS

List of Figures ..... **i**

Introduction **1**

Data **2**

I. Quality of Life Index Dataset 2

II. Population Dataset 3

III. Location Dataset 4

IV. Unit Rent Price Dataset 5

V. Venues Dataset 5

Methodology **6**

I. Data Preparation 6

II. Modeling 7

Result **9**

Discussion **11**

Conclusion **12**

List of Figures

Figure 1-Quality of Life Index Dataset 3
Figure 2-Population Dataset 3
Figure 3-Population Dataset after manipulations 4
Figure 4-Location Dataset 4
Figure 5-Unit Rent Price Dataset 5
Figure 6-Foursquare API Venue Dataset 5
Figure 7-Merged Data frame 6
Figure 8-Merged data frame with Percentage column 7
Figure 9-Relation Between Unit Rent and Percentage 8
Figure 10-Relation Between Population and Percentage 8
Figure 11-Relation Between Index and Percentage 9
Figure 12-Result Dataset 9
Figure 13-Visualize the Cluster and Percentage Values on the Map of Istanbul 10

Introduction

First of all, this report and study was conducted within the scope of the IBM Data Science Capstone project. I aimed to find this answer: “If you decided to open a sea food restaurant in Istanbul where would it be?”. I tried to ask this question by thinking that I own a fish restaurant chain trying to break into the Turkish market from abroad. I used data science methods to answer this question. These methods are briefly as follows;

1. Business Understanding
2. Analytic Approach
3. Data Requirements
4. Data Collection
5. Data Understanding
6. Data Preparation
7. Modeling
8. Evaluation
9. Deployment
10. Feedback

All of these steps are called data science methodology.

Of course, it's not enough just to follow these steps. This methodology needs data to make sense. I'll explain this data in detail in the second section. But for now, let me tell you about the titles I've chosen and why I chose these titles.

Quality of Life Index: This data is provided by Istanbul University (IU) Faculty of Economics faculty member Assoc. Dr. Murat Şeker's quality of life research in Istanbul according to the districts. The reason I chose this data is because I want my potential customers to be high-quality ones.

Population: This data is obtained from the Wikipedia page of Istanbul. I chose this data because I thought the number of potential customers in the area where I was going to open the restaurant was important.

Unit Rent Price: I obtained this data from the a real estate site called www.emlakkulusi.com. I need this data because the cost of the shop I'm going to open will be significant.

Venues: Location information is one of the most important information for me. I used the Foursquare API to obtain this information. The aim of obtaining this data is to get the ratio of the total number of restaurants according to the sea food restaurants.

Now let's look at the data sets in more detail.

Data

I explained about the background of the data and why I chose these titles in the “Introduction” section, and now it's time to examine these data in more detail.

I. Quality of Life Index Dataset

This data set consists of two columns. The first column is the column where the counties are then called neighborhood. The second column is the column with index values . This dataset is retrieved from the "Istanbul_University_Quality_of_Life_Index_Research.csv" file. We can see a part of the data set in Figure 1.

	Ilce	Index
0	BEŞİKTAŞ	0.911
1	KADIKÖY	0.886
2	BAKIRKÖY	0.613
3	ŞİŞLİ	0.574
4	FATİH	0.49

Figure 1-Quality of Life Index Dataset

II. Population Dataset

The population data set consists of a total of eight columns. But I only used two columns. These eight columns are; the column where the counties are then called neighborhood , column with 2018 population data , column with 2019 population data , column with differences in 2018 and 2019 population data , column with percentage representation of population change , column with neighborhood numbers , column with the area owned by the district and column with population density. This dataset is retrieved from the "Istanbul_Census_2019.csv" file. We can see a part of the data set in Figure 2.

	Ilce	Nüfus 2018	Nüfus 2019	Fark	Nüfus Artışı %	Mah.Say.	Alanı km2	Yoğunluk
0	ADALAR	16.119	15.238	-881.000	-5,47	5	11	1.385
1	ARNAVUTKÖY	270.549	282.488	11.939	4,41	38	453	624.000
2	ATAŞEHİR	416.318	425.094	8.776	2,11	17	25	17.004
3	AVCILAR	435.625	448.882	13.257	3,04	10	50	8.978
4	BAĞCILAR	734.369	745.125	10.756	1,46	22	23	32.397

Figure 2-Population Dataset

After the manipulations, all that remained were the columns with county data and 2019 population data. We can see a part of the data set in Figure 3.

	Ilce	Nüfus 2019
0	ADALAR	15.238
1	ARNAVUTKÖY	282.488
2	ATAŞEHİR	425.094
3	AVCILAR	448.882
4	BAĞCILAR	745.125

Figure 3-Population Dataset after manipulations

III. Location Dataset

This data set was created with geocoder. With this method, I received the central coordinates of all the counties in Istanbul. I printed this data to "Istanbul_Lat_Long.csv". The data set consists of three columns with the names of the counties, the latitude and longitude data. We can see a part of the data set in Figure 4.

	Ilce	Latitude	Longitude
0	BEŞİKTAŞ	41.042847	29.007528
1	KADIKÖY	40.991572	29.027017
2	BAKIRKÖY	40.983541	28.867973
3	ŞİŞLİ	41.061273	28.985020
4	FATİH	41.009633	28.965165

Figure 4-Location Dataset

IV. Unit Rent Price Dataset

This dataset consists of a total of two columns. Column one is the column with county names. The second column consists of average unit rent prices. This dataset is retrieved from the "Istanbul_UnitRentPrice2020.csv" file. We can see part of the data set in Figure 5.

	Ilce	Birim Kira (TL/m2)
0	ADALAR	19
1	BAKIRKÖY	24
2	BEŞİKTAŞ	27
3	BEYKOZ	15
4	BEYOĞLU	27

Figure 5-Unit Rent Price Dataset

V. Venues Dataset

This data set was created using the Foursquare API. The data set consists of a total of 7 columns. These columns are; column with county names , column with county latitude coordinates , column with county longitude coordinates , column with the name of the venue , column with venue latitude coordinates , column with county longitude coordinates and column with category of venue. We can see a part of the data set in Figure 6.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	BEŞİKTAŞ	41.042847	29.007528	Four Seasons Hotel Bosphorus	41.042101	29.012102	Hotel
1	BEŞİKTAŞ	41.042847	29.007528	Terrace Cafe At Four Season	41.042168	29.012391	Café
2	BEŞİKTAŞ	41.042847	29.007528	Conrad İstanbul Executive Lounge	41.047035	29.008963	Roof Deck
3	BEŞİKTAŞ	41.042847	29.007528	Beşiktaş Kahvesi Hookah Lounge	41.044550	29.001968	Hookah Bar
4	BEŞİKTAŞ	41.042847	29.007528	Şairler Parkı	41.042328	29.000285	Park

Figure 6-Foursquare API Venue Dataset

Methodology

The idea behind my project was to conduct a market analysis based on certain factors. I've talked about these factors in previous sections. In this section, I'm going to talk about how I prepared data sets, model selection, and the reasons of model that I selected.

I. Data Preparation

First of all, we have a four datasets. So we must merge them. But before merge I change some column names. The column with the counties in all four data sets is called “Ilce”. I changed that name by the name “Neighborhood”. Then I combine four datasets according to the Neighborhood column. We can see a part of the data set in Figure 7.

	Neighborhood	Index	Population	Unit Rent(TL/m2)	Latitude	Longitude
0	BEŞİKTAŞ	0.911	182.649	27	41.042847	29.007528
1	KADIKÖY	0.886	482.713	20	40.991572	29.027017
2	BAKIRKÖY	0.613	229.239	24	40.983541	28.867973
3	ŞİŞLİ	0.574	279.817	22	41.061273	28.985020
4	FATİH	0.49	443.090	18	41.009633	28.965165

Figure 7-Merged Data frame

Then, as I said above I used Foursquare API for venue data. After that I averaged the number of venue categories by county, filtered with the word “restaurant” in it and I created another column named “Sum” which includes the average total number of the restaurants in each counties.

After that I create a column named “Ratio” which includes the ratio of total number of the restaurants and sea food restaurants in each counties. Then I merged column named “Ratio” and the data frame that I merged before according to the Neighborhood column. After that I renamed column named “Ratio” to “Percentage” and I minted with the values in the column with the 100. We can see a part of the data set in Figure 8.

	Neighborhood	Latitude	Longitude	Index	Population	Unit Rent(TL/m2)	Percentage
0	ADALAR	40.876259	29.091027	-0.142	15.238	19.0	26.666667
2	ATAŞEHİR	40.984749	29.106720	0.046	425.094	16.0	23.076923
3	AVCILAR	40.980135	28.717547	-0.161	448.882	12.0	36.363636
4	BAHÇELİEVLER	40.997700	28.850600	0.053	611.059	14.0	21.428571
5	BAKIRKÖY	40.983541	28.867973	0.613	229.239	24.0	23.529412

Figure 8-Merged data frame with Percentage column

II. Modeling

In this section I'm going to talk about the model I chose and why I chose this model. First of all there are many machine learning models. There are areas where each has disadvantages and advantages. These models vary depending on the usage areas and most importantly the characteristics of the data you have. Since I don't have any marked data, I have to use the method of unsupervised learning. I'm going to use the Clustering method, one of the unsupervised learning methods. The clustering method is often used in customer grouping. As a result of my research, I decided to use the K-Means model, the most widely used model in the Clustering method. At the same time, the lack of linear relationship between the data has been a factor in this decision.

We can see the charts of the relationship between the data in Figure 9, Figure 10, and Figure 11.

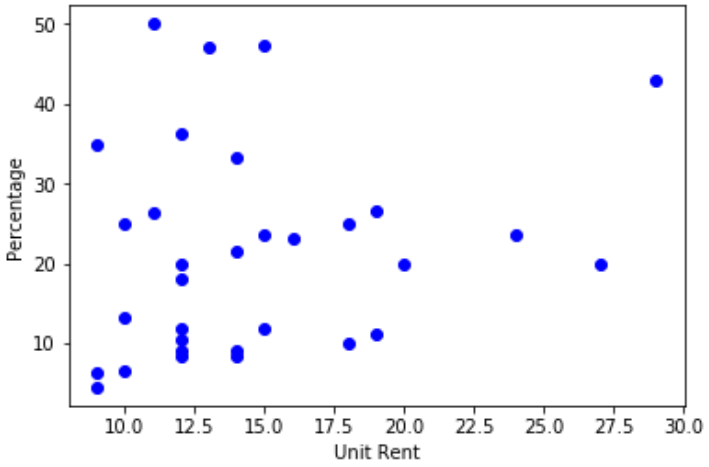

Figure 9-Relation Between Unit Rent and Percentage

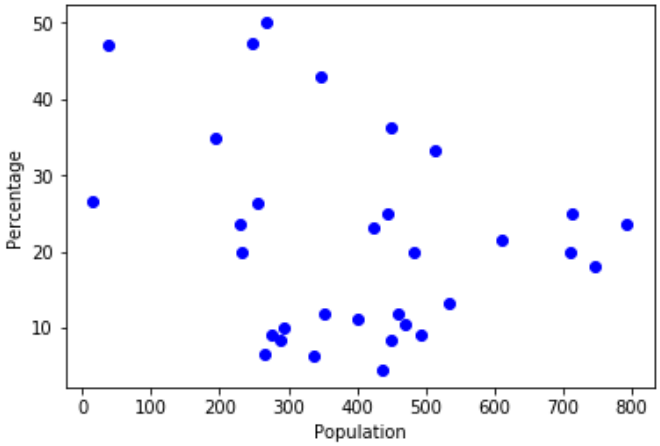

Figure 10-Relation Between Population and Percentage

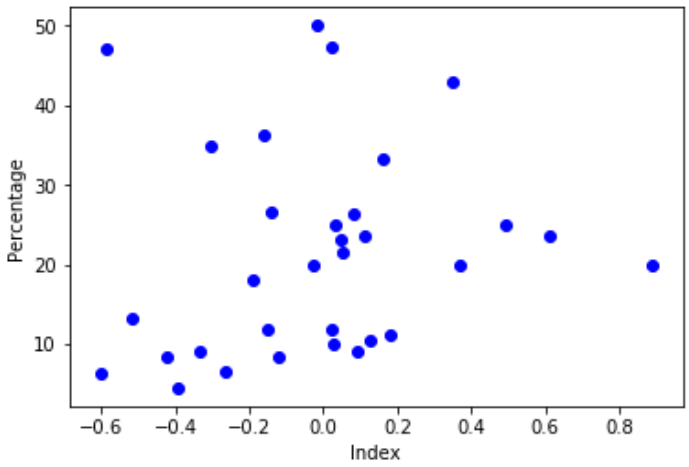

Figure 11-Relation Between Index and Percentage

Result

After I decided on the model, I ran the model on the data set I created. I have set a "k" value to 5. The value of "k" represents the number of clusters. At the end of this process, I added the cluster numbers assigned to the data set to the column called cluster layers which are created. We can see a part of the result data set in Figure 12.

	Cluster Labels	Neighborhood	Latitude	Longitude	Index	Population	Unit_Rent	Percentage
0	1	ADALAR	40.876259	29.091027	-0.142	15.238	19.0	26.666667
2	0	ATAŞEHİR	40.984749	29.106720	0.046	425.094	16.0	23.076923
3	0	AVCILAR	40.980135	28.717547	-0.161	448.882	12.0	36.363636
4	3	BAHÇELİEVLER	40.997700	28.850600	0.053	611.059	14.0	21.428571
5	2	BAKIRKÖY	40.983541	28.867973	0.613	229.239	24.0	23.529412

Figure 12-Result Dataset

At the end of this process, I used the folium library to show the cluster and percentage values on the map of Istanbul.(Figure 13)

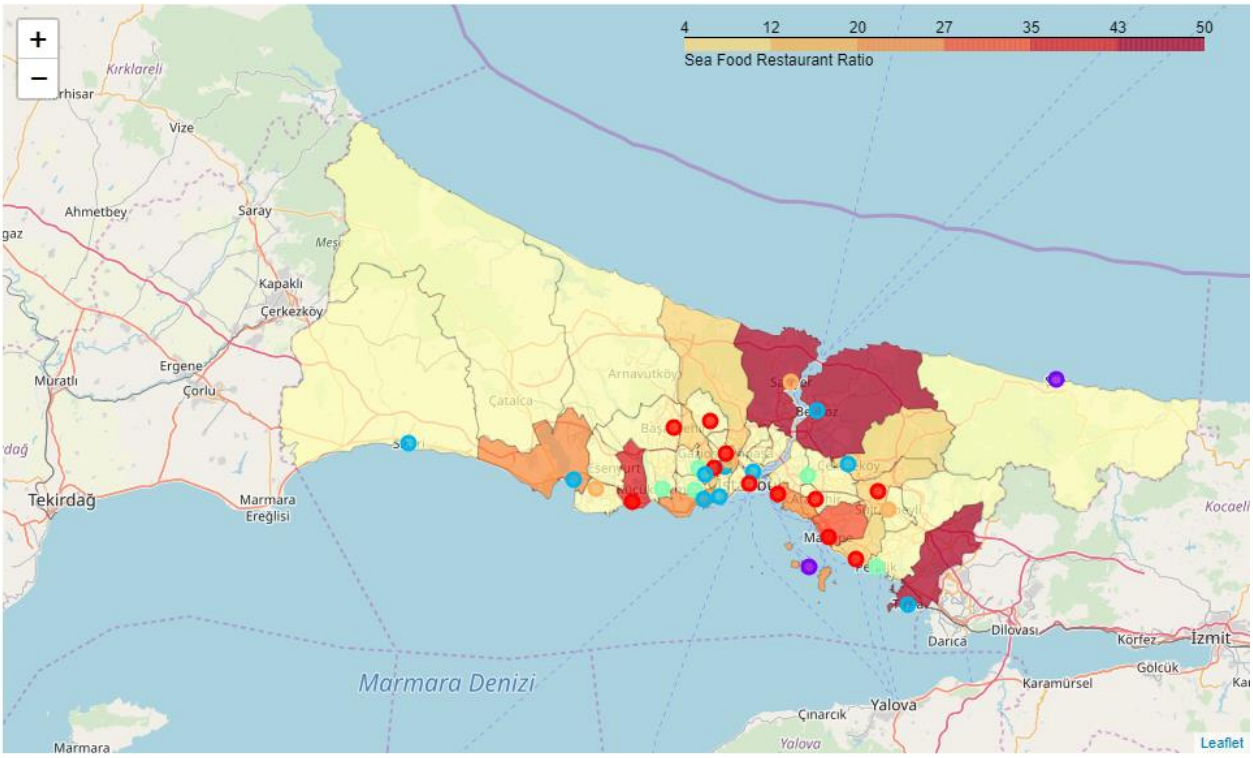

Figure 13-Visualize the Cluster and Percentage Values on the Map of Istanbul

- Red dots show Cluster_0.
- Purple dots show Cluster_1.
- Blue dots show Cluster_2.
- Green dots show Cluster_3.
- Orange dots show Cluster_4.

Discussion

Before we start, let's decide what the decisive factor is for us. If you are about to enter the market and your goal is to provide quality service, it would be important for you to have a good quality of life in the district where you will open the restaurant. So the Index will be the most decisive factor for you.

Cluster_0:In this cluster there are two counties that stand out according to the Index: Kadıköy and Fatih. Kadıköy has 0.886 Index number, 482.713 Population , 20.0 Turkish Liras Unit Rent Price and 20% Sea Food Restaurant ratios over other restaurants respectively. Fatih has 0.490 Index number, 443.090 Population, 18.0 Turkish Liras Unit Rent Price and 25% Sea Food Restaurant ratios over other restaurants respectively.

Cluster_1:In this cluster we have only two counties. Those counties are Adalar and Şile. Adalar has -0.142 Index number, 15.238 Population, 19.0 Turkish Liras Unit Rent Price and 26% Sea Food Restaurant ratios over other restaurants respectively. As for Şile, it has -0.587 Index number, 37.692 Population, 13.0 Turkish Liras Unit Rent Price and 47% Sea Food Restaurant ratios over other restaurants respectively.

Cluster_2:In this cluster there are two counties that stand out according to the Index: Bakırköy and Beyoğlu. Bakırköy has 0.613 Index number, 229.238 Population, 24.0 Turkish Liras Unit Rent Price and 23% Sea Food Restaurant ratios over other restaurants respectively. Beyoğlu has 0.367 Index number, 233.323 Population 27.0 Turkish Liras Unit Rent Price and 20% Sea Food Restaurant ratios over other restaurants respectively.

Cluster_3:In this cluster there is one county that stands out according to the Index: Küçükçekmece. Küçükçekmece has 0.114 Index number, 792.821 Population, 15.0 Turkish Liras Unit Rent Price and 23% Sea Food Restaurant ratios over other restaurants respectively.

Cluster_4:In this cluster there are two counties that stand out according to the Index: Sarıyer and Eyüp. Sarıyer has 0.347 Index Number, 347.214 Population, 29.0 Turkish Liras Unit Rent Price and 42% Sea Food Restaurant ratios over other restaurants respectively. Eyüp has 0.183 Index number, 400.513 Population, 29.0 Turkish Liras Unit Rent Price and 11% Sea Food Restaurant ratios over other restaurants respectively.

Conclusion

With the help of map and clusters, it can be said that the most remarkable options are Bakırköy from cluster_2 and Kadıköy from cluster_0. But if we go deeper, we can see there is some major difference between these two counties. First, Kadıköy has greater Index Number and Population. However Bakırköy has 3% more sea food restaurant percentage than Kadıköy and Unit Rent Price is 4.0 Turkish Liras more. But for percentage we should not forget the restaurants that are not in Foursquare. Second is the causes of non-mathematical effects. These effects are transportation, trustic venues, intensity of cultural and artstic activities. Of these three factors, trustic venues, cultural and artistic activities stand out for Kadikoy. But transportation is equal for both. Because they have access to both Kadıköy and Bakırköy metro lines, bus lines, ferry lines and metrobus lines.

This project was prepared under the “Applied Capstone Project” at the end of Coursera and IBM's Data Science Professional Certificate program.