IBM Data Science Capstone Project

Jahid Ahsan Khan
IBM Data Science Professional Certificate
8/18/20

Contents

IBM Data Science Capstone Project

1. Introduction & Business Problem

1.1 Introduction

The capital of Pakistan, Islamabad, considered as the second most beautiful capital of the world is home to some of the most favorite restaurants across the country. Every year millions of tourists both national and international visit here to enjoy the natural beauty of the city while feasting on the variety of foods it has to offer. But with many restaurants to choose from, it often fills the new tourists with overload of choices, and choosing a bad restaurant can lead to a miserable experience for the tourists.

1.2 Problem

For a travel advisor, when making profile for a city, it is important to know which areas of the city provide the best food experience for dining out so that they can guide their customers accordingly. This is what this project aims to find out for Islamabad.

2. Data Requirements & Acquisition

2.1 Data Requirements

The primary data requirement for the given problem will be the data containing the division of Islamabad into sectors. This data will then be utilized to get the information about the venues within each sector. The required information about each venue includes:

1. Name of the venue
2. Type of venue for categorizing
3. The location coordinates of venue
4. Popularity of the venue

2.2 Data Acquisition Approach

For this project, I have acquired the data of Sectors of Islamabad from Wikipedia here. The head of the data is shown here.

	Sectors of Islamabad
0	Diplomatic Enclave, Islamabad
1	A-17, Islamabad
2	A-18, Islamabad (page does not exist)
3	B-17, Islamabad
4	B-18, Islamabad (page does not exist)

The Nominatim API from GeoPy library was used to extract the coordinates of sectors of Islamabad. Some of those sectors were removed later whose detail is mentioned in data cleaning section.

	Sectors of Islamabad	address	latitude	longitude
0	Diplomatic Enclave, Islamabad	Diplomatic Enclave, اسلام آباد, وفاقی دارالحکومت اسلام آباد, 44000, پاکستان	33.7236	73.1118
1	A-17, Islamabad	Sector A, Jagiot, وفاقی دارالحکومت اسلام آباد, پاکستان	33.6919	73.2171
2	A-18, Islamabad	Sector A, Jagiot, وفاقی دارالحکومت اسلام آباد, پاکستان	33.6919	73.2171
3	B-17, Islamabad	B 17, وفاقی دارالحکومت اسلام آباد, پاکستان	33.6904	72.8286
4	B-18, Islamabad	18, Street 6, G-13/3, G-13, اسلام آباد, وفاقی دارالحکومت اسلام آباد, 44000, پاکستان	33.6568	72.9607

After that, Foursquare API was used to extract all the venues within 500-meter radius of sectors coordinates. The information extracted include the names of venues, category of venue, location coordinates of venue, and venue ID.

	Sectors of Islamabad	latitude	longitude	Venue	venue_id	Venue latitude	Venue longitude	Venue category
0	Diplomatic Enclave, Islamabad	33.723606	73.111830	Club 21 (French Club)	4db1479793a061576851381d	33.720966	73.112339	Lounge
1	Diplomatic Enclave, Islamabad	33.723606	73.111830	Canadian Club	4e5a1c7be4cd875e8eb7172d	33.721035	73.112292	Restaurant
2	Diplomatic Enclave, Islamabad	33.723606	73.111830	Gloria Jean's Cafe	5212de7d11d2af9b90ead237	33.721736	73.106869	Coffee Shop
3	Diplomatic Enclave, Islamabad	33.723606	73.111830	A Club	4cd9aad97bb06dcbf792a9b2	33.723374	73.117329	American Restaurant
4	D-12, Islamabad	33.701818	72.948619	D-12 Markaz	596b8fc1a22db76efcd87190	33.701010	72.950250	Shopping Plaza

Venue categories were used to filter our data to contain only restaurants in venue column and then venue IDs was used to find the count of likes for each restaurant which was used to

determine the popularity of that restaurant.

	Sectors of Islamabad	latitude	longitude	Venue	venue_id	Venue latitude	Venue longitude	Venue category	Total Likes
0	Diplomatic Enclave, Islamabad	33.723606	73.111830	Canadian Club	4e5a1c7be4cd875e8eb7172d	33.721035	73.112292	Restaurant	3
1	Diplomatic Enclave, Islamabad	33.723606	73.111830	Gloria Jean's Cafe	5212de7d11d2af9b90ead237	33.721736	73.106869	Coffee Shop	6
2	Diplomatic Enclave, Islamabad	33.723606	73.111830	A Club	4cd9aad97bb06dcbf792a9b2	33.723374	73.117329	American Restaurant	2
3	E-7, Islamabad	33.727498	73.051239	Texas Steak House	520a66b311d2aa759eac9baa	33.727476	73.047656	Steakhouse	20
4	E-7, Islamabad	33.727498	73.051239	Subway F-7	53341762498e11af2be50b71	33.719323	73.053343	Sandwich Place	3

3. Methodology

3.1 Data Preparation & Exploration

The data acquired from Wikipedia was cleaned by first removing the “page does not exist” statement from sector names where found. The shape of the data now is (80,1) meaning that the data scrapping from Wikipedia returned 80 names of sectors. The head of the data now looked like this.

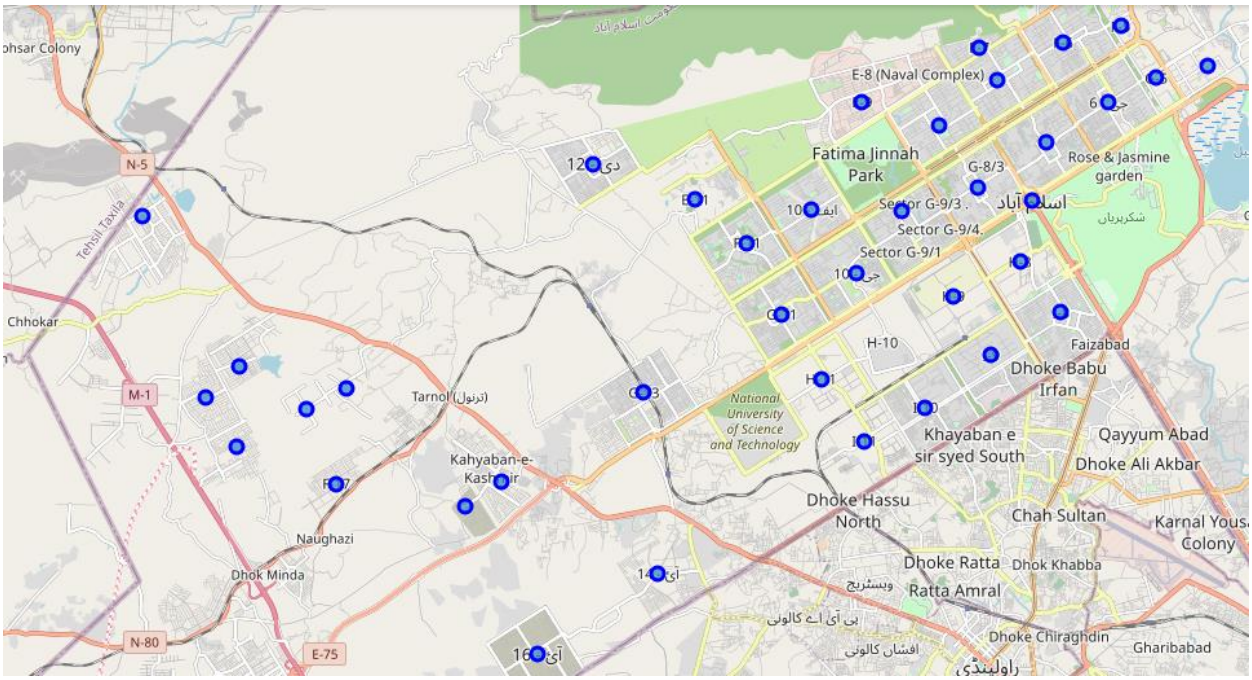
	Sectors of Islamabad
0	Diplomatic Enclave, Islamabad
1	A-17, Islamabad
2	A-18, Islamabad
3	B-17, Islamabad
4	B-18, Islamabad

A few names of sectors were wrongly mentioned and did not exist hence were removed manually. Apart from that, the Nominatim API could not locate the coordinates of few sectors which were identified after comparing them with returned addresses. Hence, they were removed as well. The final number of rows in the data is now 38 (instead of 80). The head of

the data after removing those sectors can be seen below.

	Sectors of Islamabad	address	latitude	longitude
0	Diplomatic Enclave, Islamabad	Diplomatic Enclave, اسلام آباد, وفاقی دارالحکومت اسلام آباد, 44000, پاکستان	33.7236	73.1118
1	B-17, Islamabad	B 17, وفاقی دارالحکومت اسلام آباد, پاکستان	33.6904	72.8286
2	D-12, Islamabad	دی - 12, وفاقی دارالحکومت اسلام آباد, 430000, پاکستان	33.7018	72.9486
3	D-17, Islamabad	D 17, F-17, اسلام آباد, وفاقی دارالحکومت اسلام آباد, 44000, پاکستان	33.6571	72.8546
4	D-18, Islamabad	D 18, F-17, Dhok Minda, وفاقی دارالحکومت اسلام آباد, پاکستان	33.6504	72.8456

Finally, the data was plotted on a map using Folium library to confirm that the coordinates received from API were correct.



The data obtained from foursquare API contained 369 venues with 91 unique categories. The venues were then filtered using category data to contain only the restaurants or other such categories for dining out. The code for this procedure is as follows.

```
islamabad_venues = islamabad_venues[islamabad_venues['Venue category'].str.contains('Restaurant|Joint|Steakhouse|Sandwich|Pizza|Tea Room|Place|Ice Cream Shop|Fish & Chips|Donut|Diner|Dessert|Coffee|Café|Breakfast')].reset_index(drop=True)
```

The number of venues were now down to 216 with 35 unique categories. Furthermore, due to difference in Sectors' Area, the radius of value 500 was used to extract all the venues of bigger

sectors which also resulted in some of the venues returning twice representing different sectors. These duplicates were removed while ensuring that those venues only represented their actual corresponding sectors. Now the final count of venues is 180.

Feature Engineering

The count of likes for a restaurant obtained from foursquare API was used to calculate the total number of likes of all the restaurants in a sector and was named “Total Likes per Sector”. Furthermore, restaurants for each sector were counted to make a new feature “No. of Venues per Sector”.

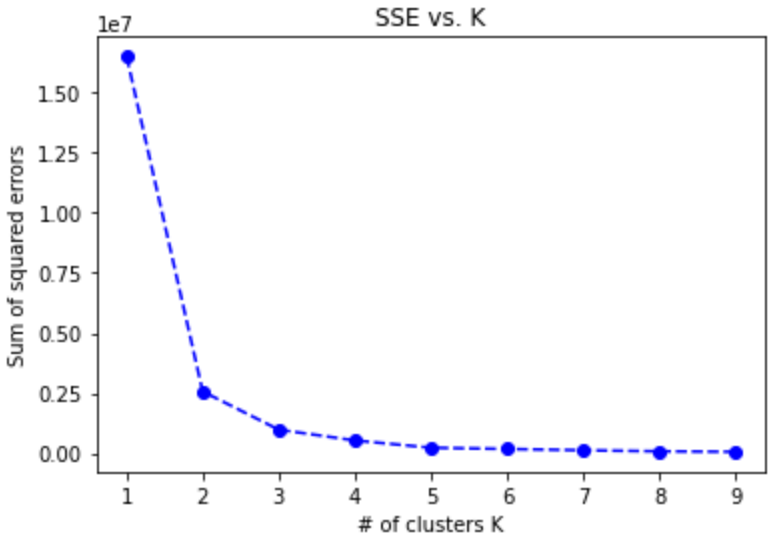
	Sectors of Islamabad	latitude	longitude	Venue	venue_id	Venue latitude	Venue longitude	Venue category	Venue Likes	Total Likes per Sector	No. of Venues per Sector
0	Diplomatic Enclave, Islamabad	33.723606	73.111830	Canadian Club	4e5a1c7be4cd875e8eb7172d	33.721035	73.112292	Restaurant	3	11	3
1	Diplomatic Enclave, Islamabad	33.723606	73.111830	Gloria Jean's Cafe	5212de7d11d2af9b90ead237	33.721736	73.106869	Coffee Shop	6	11	3
2	Diplomatic Enclave, Islamabad	33.723606	73.111830	A Club	4cd9aad97bb06dcbf792a9b2	33.723374	73.117329	American Restaurant	2	11	3
3	E-7, Islamabad	33.727498	73.051239	Texas Steak House	520a66b311d2aa759eac9baa	33.727476	73.047656	Steakhouse	20	27	4
4	E-7, Islamabad	33.727498	73.051239	Subway F-7	53341762498e11af2be50b71	33.719323	73.053343	Sandwich Place	3	27	4

Limitations of Foursquare API

The information extracted from venue IDs contained only Like Counts of the obtained venues. The Foursquare API also had other extractable details such as price tier, ratings, and tip count which could not be utilized in this project since majority of data in those requests were found to be missing. Moreover, the Foursquare API also had free usage limitations to conform to. Hence, only Like count parameter was used for determining popularity for this project even though ratings parameter could certainly have been useful.

3.2 Clustering

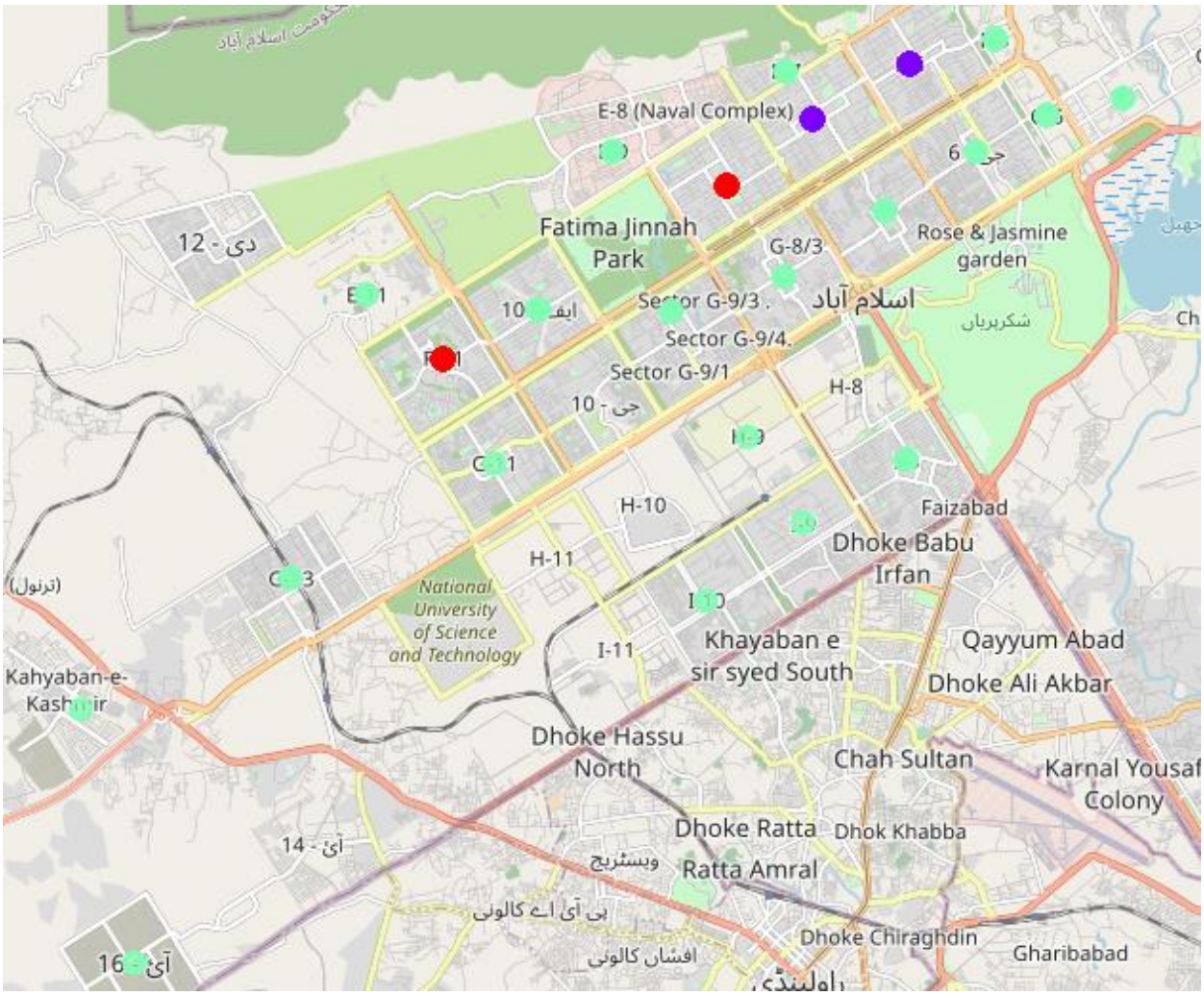
A good method to obtain valuable insights from data is by clustering the data into groups using unsupervised machine learning algorithms. In this project, I used K-means clustering technique to divide data into groups of k clusters. The data was prepared for this part after one-hot encoding the categorical variable ‘Venue Category’ and then adding the remaining numerical variables with it. After that, the optimal number of clusters ‘k’ was obtained after plotting the elbow plot given below.



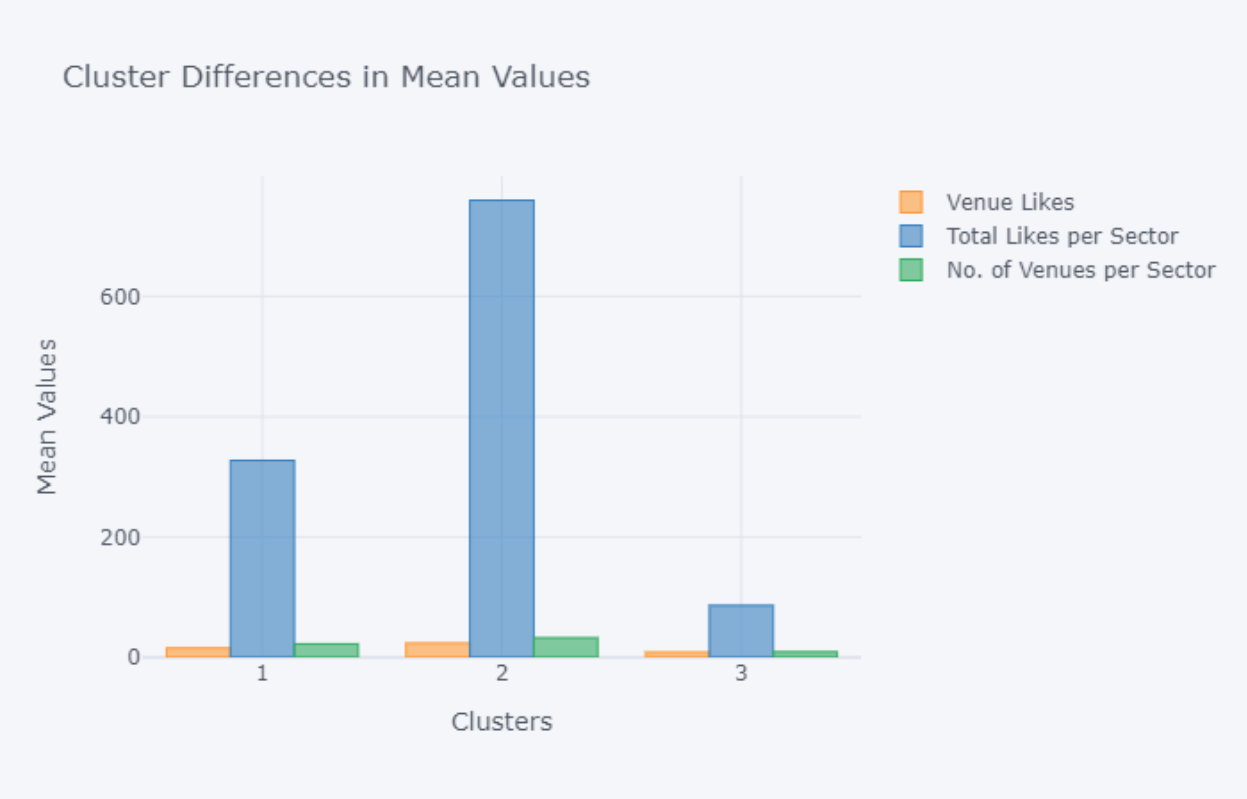
With this chart, one can see that the elbow was formed at k = 2. However, since 2 is very small number for analysis hence k = 3 was used for clustering to obtain more insights.

4. Results & Discussion

The assumption that goes with the results is that the higher the number of total likes and number of restaurants per sector, the more popular that sector will be. After the data was clustered into k = 3 groups, the results were visualized using folium library with which we can see that the Sectors F-8 and F-11 were grouped into first cluster while Sectors F-6 and F-7 were grouped into second cluster. All the remaining sectors were grouped into last cluster.

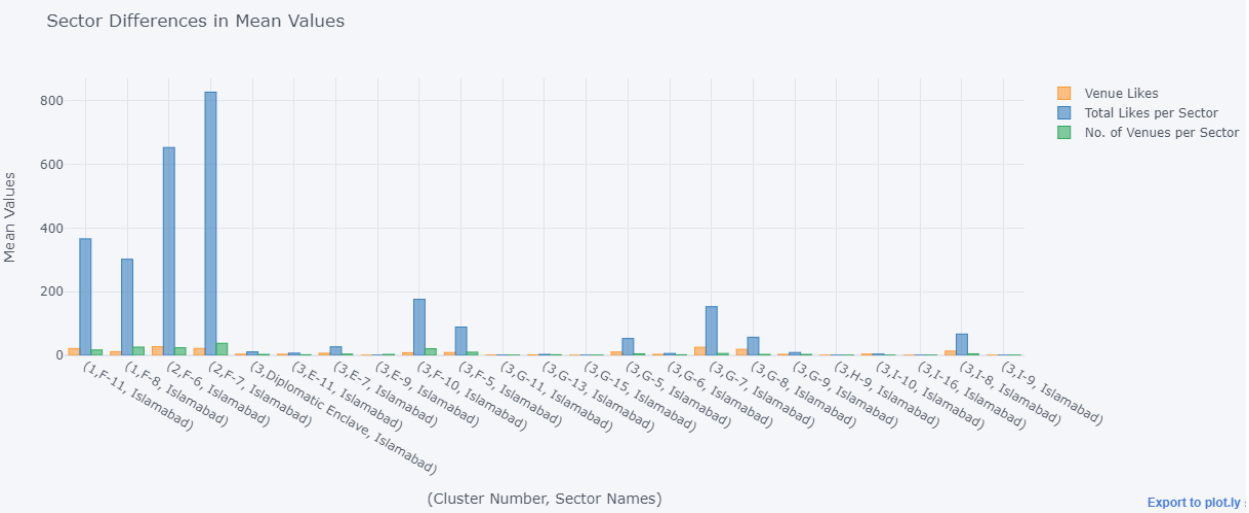


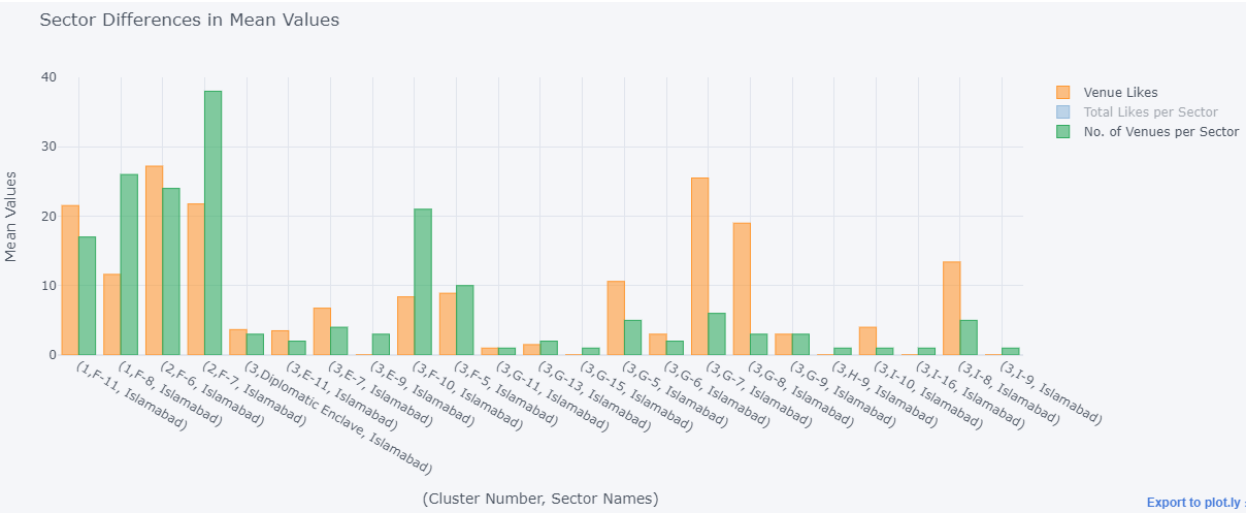
Upon visualizing the clusters, it was found that the cluster 2 had the highest mean likes followed by cluster 1 and cluster 3 whereas the same pattern followed when the mean of number of restaurants was calculated.


Cluster Differences in Mean Values
Venue Likes
Total Likes per Sector
No. of Venues per Sector
Mean Values
600
400
200
0
1
2
3
Clusters


Cluster Differences in Mean Values
Venue Likes
Total Likes per Sector
No. of Venues per Sector
Mean Values
30
25
20
15
10
5
0
1
2
3
Clusters

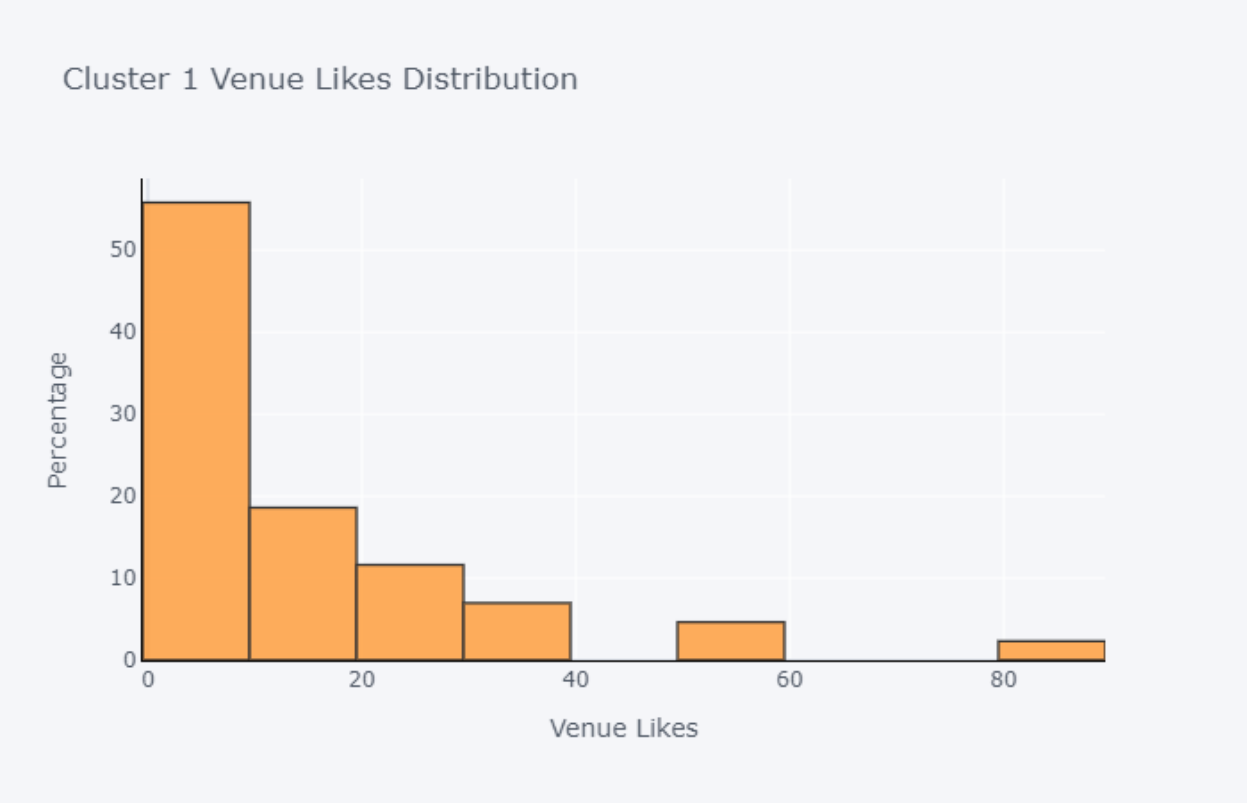
The mean differences in clusters can be further understood by visualizing them based on sectors.

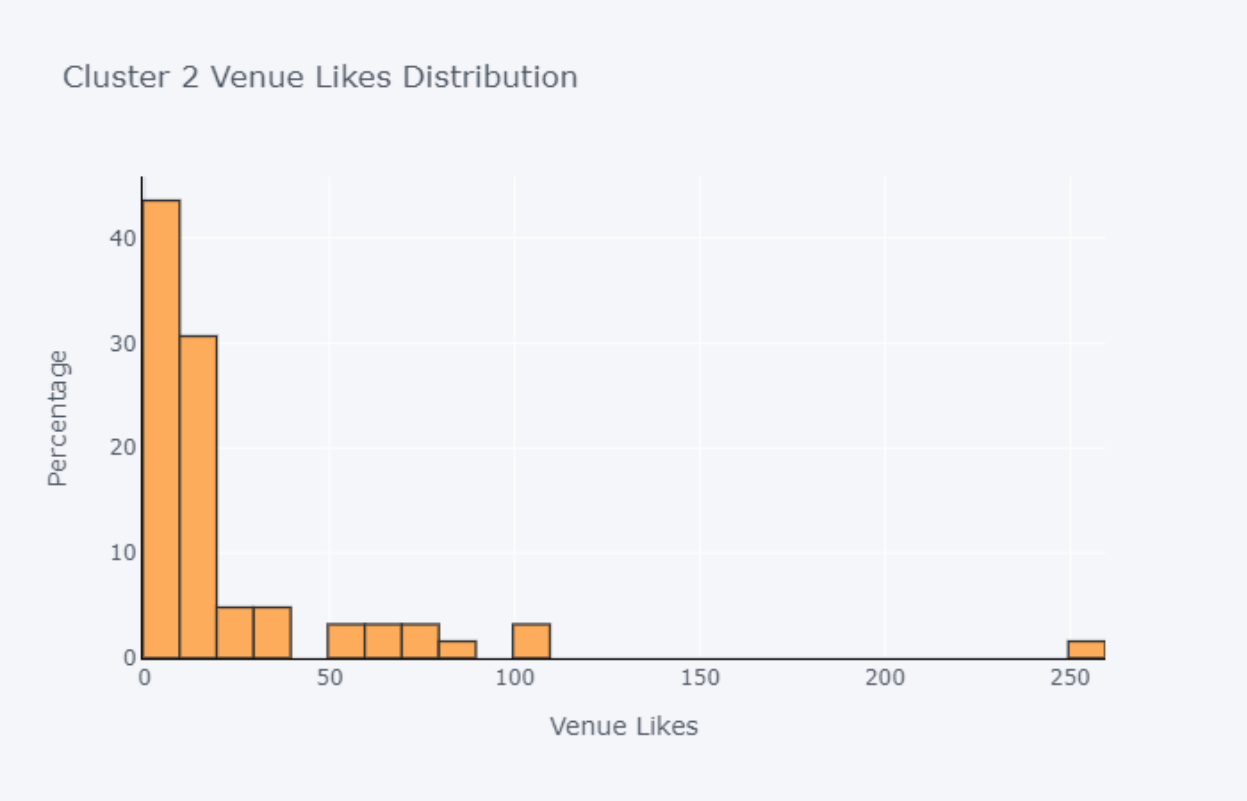


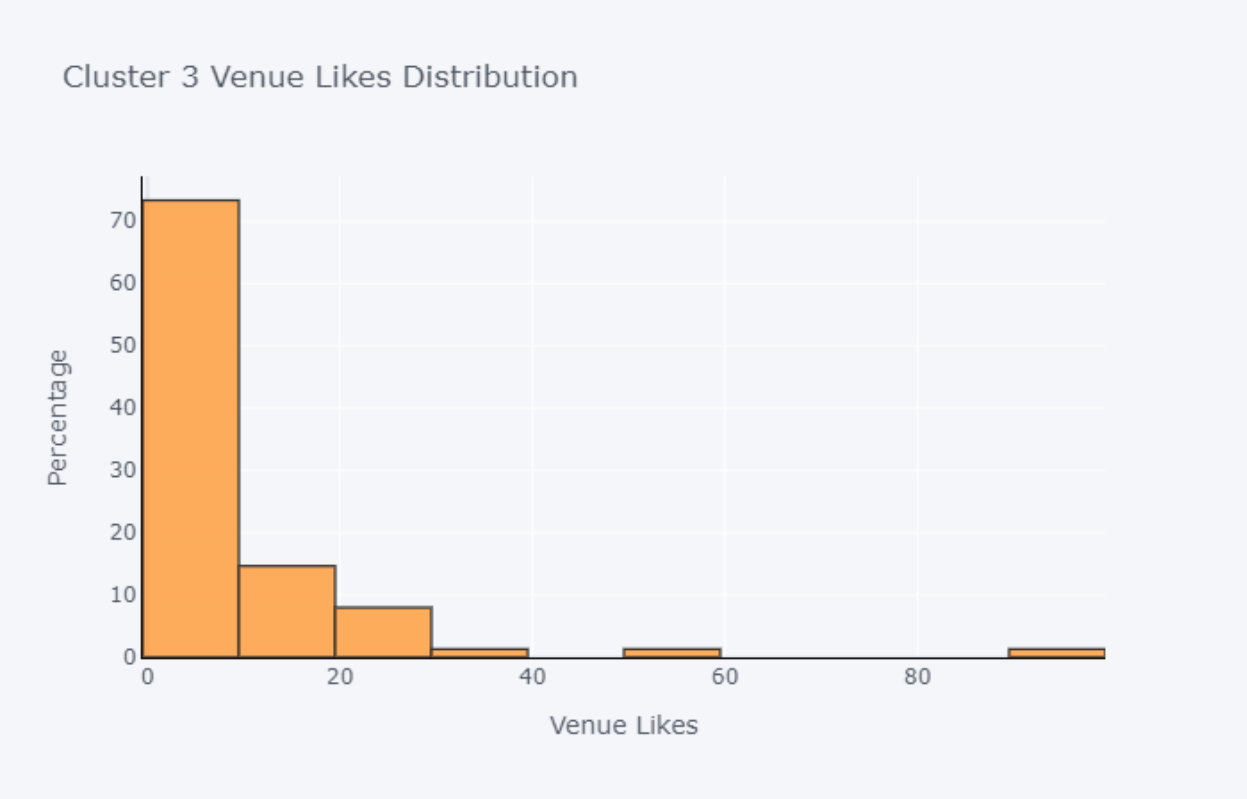


Here we can see that the Sector F-10 had above average number of restaurants but the mean value of likes was low whereas in Sectors G-5, G-7, G-8, and I-8 the mean value of likes was above average but the count of restaurants was low. The only sectors that stood out from the rest were the sectors of cluster 2 followed by cluster 1. This means that the sectors of cluster 2 i.e. Sectors F-6 and F-7 can be considered the most popular sectors for dining out followed by sectors of cluster 1 i.e. Sectors F-8 and F-11.

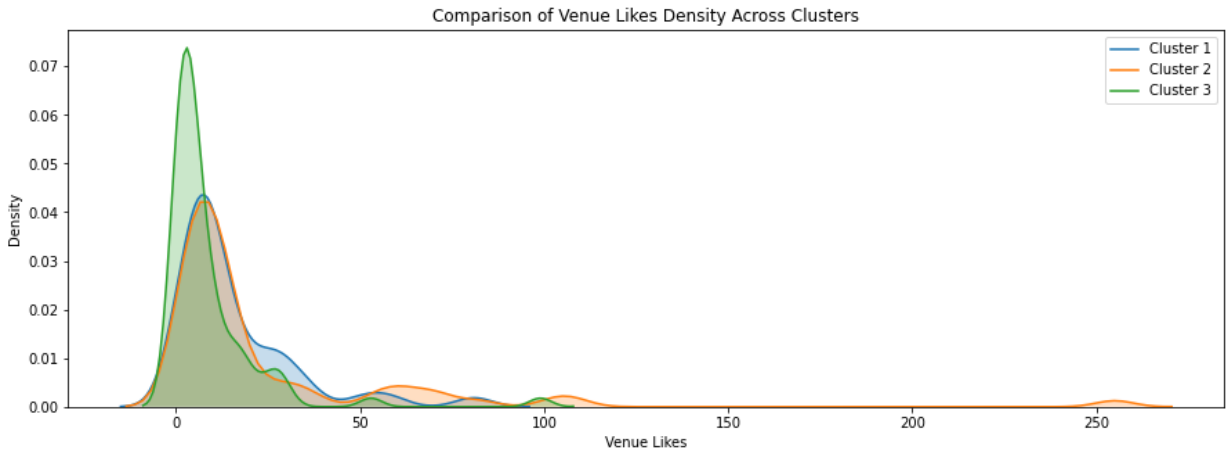
The distribution of likes in percentage for each cluster can be seen below.







With these graphs we can see that the share of venue likes between 0 to 9 is highest in cluster 3 i.e. >70% and lowest in cluster 2 i.e. above 40%. While the range of venue likes in cluster 2 is above 250 which shows that not only is cluster 2 most popular in terms of mean values but the most popular restaurants also fall in this cluster. The density distribution of venue likes with respect to clusters as seen below also validated these points.



Finally, the top 10 most popular venues across Islamabad are listed here.

	Sectors of Islamabad	Venue	Venue category	Venue Likes	Total Likes per Sector	No. of Venues per Sector	Cluster Labels
0	F-6, Islamabad	Chaaye Khana	Tea Room	255	653	24	2
1	F-7, Islamabad	Roasters Coffee House & Grill	Burger Joint	108	827	38	2
2	F-7, Islamabad	Hardee's	Fast Food Restaurant	103	827	38	2
3	G-7, Islamabad	Savour Foods	Pakistani Restaurant	99	153	6	3
4	F-7, Islamabad	Tuscany Courtyard	Italian Restaurant	82	827	38	2
5	F-11, Islamabad	Gloria Jean's	Coffee Shop	81	366	17	1
6	F-6, Islamabad	Street 1 Cafe	Italian Restaurant	71	653	24	2
7	F-6, Islamabad	Nando's	Portuguese Restaurant	70	653	24	2
8	F-7, Islamabad	Kabul Restaurant	Afghan Restaurant	63	827	38	2
9	F-6, Islamabad	Burning Brownie Cafe & Bake Shop	Coffee Shop	61	653	24	2

5. Conclusion

In conclusion, when it comes to having the best food experience in the city of Islamabad, based on the results of this project, a travel advisor can recommend the restaurants of Sectors F-6 and F-7 as a must visit while the restaurants of Sectors F-8 and F-11 can be considered highly recommended. The code requirements for this project were fulfilled using Google Colab. You can check out the Colab notebook for this project by clicking here.