Coursera Capstone Project

IBM Applied Data Science Capstone

TOURISM IN BAHRAIN

By: Maryam Almanea

August 2020

Table of Contents

1. Introduction

Bahrain is a small island country located in the Arabian Gulf between kingdom of Saudi Arabia and Qatar. Bahrain enjoys unique culture and tourism destination in the GCC region. It blends modern Arab culture with the history and tradition of the Middle East as the capital Manama been designated as Capital of Arab Culture in 2012.

Find right hotel to stay, restaurant to dine in or top places to visit might be a challenge for tourists. The main objectives of this project to explore Bahrain top big cities and venues surrounding the cities and recommend best places. Data science methodology and k-means clustering algorithm are used to carry out analysis. The result can help and satisfy visitor's needs, where best hotel to stay? Where best places to visit? And best restaurants to dine in.

2. Data

2.1 Data Sources

The data sources used in this project:

1. List of top largest cities in Bahrain and geographic location where data scraped from genomes website using BeautifulSoup python library.

2. Foursquare venue data

    1. List of venues surrounding the cities where basic details like: venue name, id, geolocation coordinates and category been extracted by regular Foursquare venue data API.

    2. The details of venues like total likes count been extracted by premium Foursquare Venue details API.

2.2 Data Cleaning

1. List of top largest cities were cleaned as follows:

    ✓ Split city column to get only city's name
    ✓ Split location into latitude and longitude coordinates.

2. Foursquare venue data cleaned as follows:

    ✓ Some of venue categories were removed as this project aims to benefit visitors; only 46 categories out of 115 were kept like hotels, coffee shops, restaurants, shopping malls.

3. Methodology

3.1 Exploratory Data Analysis

3.1.1 Bahrain Cities Data

First step is to get list of cities and geolocation coordinates, largest cities available on geoname website hence python's beautifulsoup library is used to extract data. Figure 2 show final list of largest cities after data cleaning.

	City	Location
0	Manama , Manama	26.228 / 50.586
1	Al Muharraq , Muharraq	26.257 / 50.612
2	Ar Rifā' , Southern Governorate	26.13 / 50.555
3	Dār Kulayb , Southern Governorate	26.069 / 50.504
4	Madīnat Ḩamad , Northern Governorate	26.115 / 50.507
5	Madīnat 'Īsá , Southern Governorate	26.174 / 50.548
6	Sitrah , Manama	26.155 / 50.621
7	Jidd Ḩafş , Manama	26.219 / 50.548
8	Al Ḩadd , Muharraq	26.246 / 50.654

Figure 1: Cities

	City	Latitude	Longitude
0	Manama	26.228	50.586
1	Al Muharraq	26.257	50.612
2	Ar Rifā'	26.13	50.555
3	Dār Kulayb	26.069	50.504
4	Madīnat Ḩamad	26.115	50.507
5	Madīnat 'Īsá	26.174	50.548
6	Sitrah	26.155	50.621
7	Jidd Ḩafş	26.219	50.548
8	Al Ḩadd	26.246	50.654

Figure 2: Final Cities

3.1.2 Foursquare Venue Data

The second step after extracting and clean list of cities is making API calls to Foursquare by passing cities and geographical coordinates. We got 392 venues and 115 unique categories. However, we limited to 46 categories as are not relevant for this project. Only some categories which could be useful for visitors were kept for instance, hotel, resort, park, shopping malls, market, grocery, restaurants, coffee shop, Movie Theater and historic site.

	index	City	Latitude	Longitude	Venue	id	Venue Latitude	Venue Longitude	Venue Category
0	0	Manama	26.228	50.586	Bahay Kubo Restaurant	4d07a2f18620224b7e34b940	26.228443	50.582647	Asian Restaurant
1	1	Manama	26.228	50.586	Sameeh pastries	4cb0419acbab236acbc29e73	26.226697	50.588958	Turkish Restaurant
2	2	Manama	26.228	50.586	The Lebanese Restaurant	4d288e6ff7a9224b630ef69f	26.228518	50.583612	Middle Eastern Restaurant
3	3	Manama	26.228	50.586	Al Wasmiya Village Restaurant	4e270708ae609b2f94ecc680	26.229518	50.581646	Restaurant
4	4	Manama	26.228	50.586	Jahan Grills	4dd7f705b0fb8af380937ea3	26.224671	50.587759	BBQ Joint
5	5	Manama	26.228	50.586	تكة ابل	4f26d629e4b02a9500c19159	26.229502	50.591782	Restaurant
6	6	Manama	26.228	50.586	تكة أمين	51d09d47498e2a039a702d75	26.228616	50.585451	BBQ Joint

Figure 3: Nearby venues details

Then define 10 new categorical variables to better group venues based on similarities. For example:

- Hotel_sap (hotel , spa , lounge ,resort)
- Shopping (Men's Store, Boutique, Shopping Mall, Clothing Store)
- Café_snack (Café , Coffee Shop , Ice Cream Shop, Juice Bar ,Cafeteria , Sandwich)
- breakfast_bakery_diner (Breakfast Spot ,Bakery , Department Store ,Grocery Store)
- italian_cuisine (Italian Restaurant, Pizza Place)

	index	City	Latitude	Longitude	Venue	id	Venue Latitude	Venue Longitude	Venue Category	main_category
281	378	Al Hadd	26.246	50.654	Lulu Hypermarket	4ec7fe1430f8b09c0c9812bb	26.238721	50.651118	Shopping Mall	shopping
282	380	Al Hadd	26.246	50.654	Papa John's	4cb6043652edb1f703a86bfe	26.256143	50.648395	Pizza Place	italian_cuisine
283	381	Al Hadd	26.246	50.654	Hidd Co-Operative	4db3c8996e8179a91370f5d4	26.241519	50.656628	Shopping Mall	shopping
284	383	Al Hadd	26.246	50.654	KFC	59be5d22123a196ea4f35925	26.238708	50.651111	Fried Chicken Joint	fast_food
285	384	Al Hadd	26.246	50.654	Jasmis	4f0dd29fe4b040d4823328a8	26.235767	50.649196	Burger Joint	fast_food
286	385	Al Hadd	26.246	50.654	The Dragon	4ece88b9b8f7971d68774ca7	26.235728	50.649080	Hotel	hotel_spa
287	386	Al Hadd	26.246	50.654	King karak	4f26b21ae4b01dc94dd165f2	26.234891	50.656079	Café	cafe_snack

Figure 4: venue details – main category

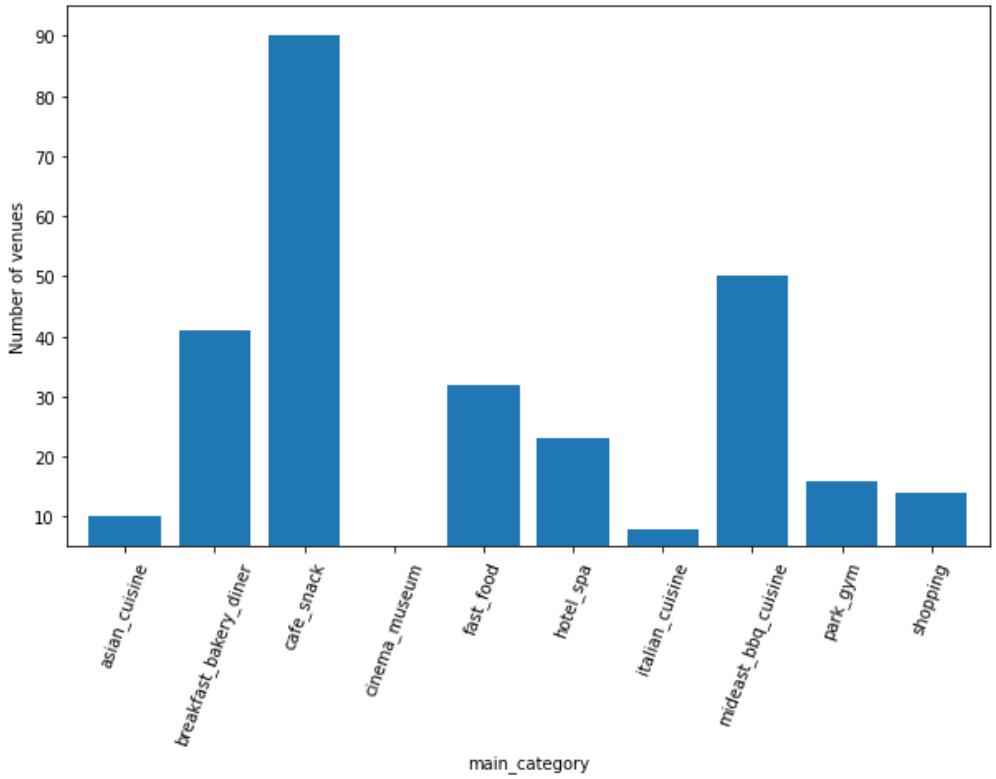

Figure 5: Number of venues per defined main category

3.1.3 Foursquare Venues Details Data

Make API calls to premium foursquare endpoint by passing venue's id https://api.foursquare.com/v2/venues/**{Venue_id}**/likes, the response shows venues details but only count which represent total venue likes is used in this project.

	total likes
count	290.000000
mean	33.510345
std	75.758048
min	0.000000
25%	6.000000
50%	11.000000
75%	27.000000
max	787.000000

Figure 6: Descriptive Statics of total likes

The above figure shows that 25% of venues got 6 likes or less, 50% got 11 or less and 75% got 27 or less likes. Based on descriptive statistic of total likes, new categorical variables are created as follows:

- 1 star : total likes 6 or less
- 2 star : total likes more than 6 and less or equal to 11
- 3 star : total likes more than 11 and less or equal to 27
- 4 star : total likes more than 27

ndex	City	Latitude	Longitude	Venue	id	Venue Latitude	Venue Longitude	Venue Category	main_category	total likes	venue rate
0	Manama	26.228	50.586	Bahay Kubo Restaurant	4d07a2f18620224b7e34b940	26.228443	50.582647	Asian Restaurant	asian_cuisine	48	star4
1	Manama	26.228	50.586	Sameeh pastries	4cb0419acbab236acbc29e73	26.226697	50.588958	Turkish Restaurant	mideast_bbq_cuisine	38	star4
2	Manama	26.228	50.586	The Lebanese Restaurant	4d288e6ff7a9224b630ef69f	26.228518	50.583612	Middle Eastern Restaurant	mideast_bbq_cuisine	45	star4
3	Manama	26.228	50.586	Al Wasmiya Village Restaurant	4e270708ae609b2f94ecc680	26.229518	50.581646	Restaurant	mideast_bbq_cuisine	11	start2
4	Manama	26.228	50.586	Jahan Grills	4dd7f705b0fb8af380937ea3	26.224671	50.587759	BBQ Joint	mideast_bbq_cuisine	55	star4

Figure 7: venue details -venue rate

3.2 Data Preparation

As shown above that list of venue data by Foursquare is not relevant to this project therefore, many categories were removed from dataset which result in 290 venues and 46 unique categories. Moreover, re categorize venue's category based on similarity for better grouping into 10 groups and categorize venue based on total like into 4 groups.

In order to prepare dataset for clustering and deal with categorical variables (main category and venue rate), one hot encoding is used to convert categorical data to integer data as machine leaning algorithms can't process categorical data directly.

	City	asian_cuisine	breakfast_bakery_diner	cafe_snack	cinema_museum	fast_food	hotel_spa	italian_cuisine	mideast_bbq_cuisine	p
0	Manama	1	0	0	0	0	0	0	0	
1	Manama	0	0	0	0	0	0	0	1	
2	Manama	0	0	0	0	0	0	0	1	
3	Manama	0	0	0	0	0	0	0	1	
4	Manama	0	0	0	0	0	0	0	1	
5	Manama	0	0	0	0	0	0	0	1	
6	Manama	0	0	0	0	0	0	0	1	
7	Manama	0	0	0	0	0	0	0	1	
8	Manama	0	0	0	0	0	0	0	0	
9	Manama	0	0	0	0	0	1	0	0	

Figure 8: One hot encoding

3.3 Clustering

3.3.1 Finding the optimal number of clusters

In order to determine optimal number of clusters K, elbow method and silhouette score method were used and tested on range of k between 2 and 5

3.3.1.1 Elbow Method

Below is a plot for sum of squared errors (SSE) of k in range of 2 to 5. As k increase the SSE decrease . The location of bend is generally considered as indicator for optimal number of clusters hence k=4 could be optimal number of clusters

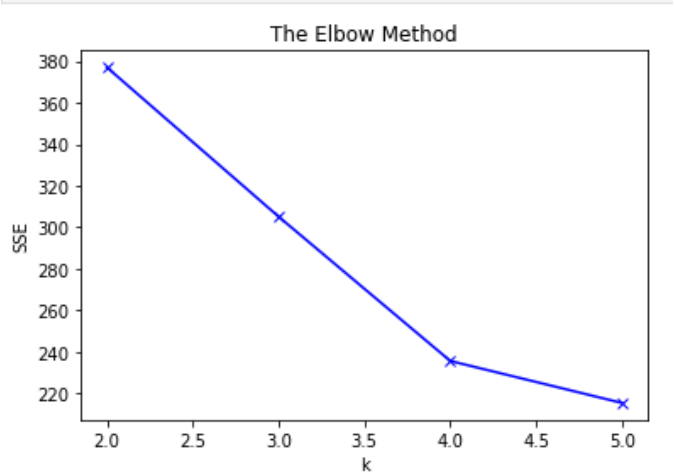

Figure 9: Elbow Method

3.3.1.2 Silhouette Score Method

Below is a plot for average silhouette scores of k in range of 2 to 5. The maximum silhouette score is considered as optimal number of clusters hence dataset can be clustered into 4 clusters.

	K	Score
0	2	0.2015
1	3	0.2882
2	4	0.3807
3	5	0.3485

Figure 10: Silhouette Scores

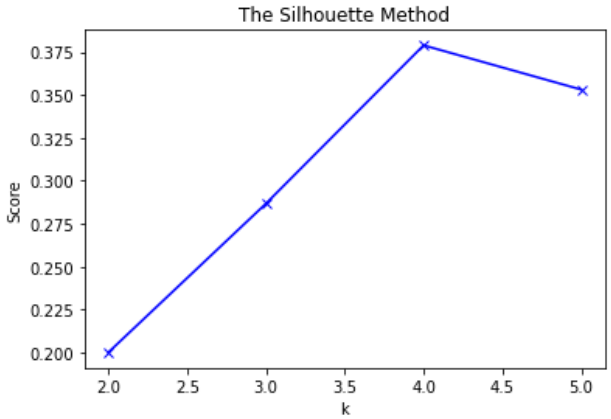

Figure 11: Silhouette Score Method

3.3.2 K-Means clustering

K-means algorithm is used to perform data clustering with 4 clusters. Its centroid – based algorithm in which it identifies k number of centroids and calculate distance of data points from centroid of cluster. Data points allocated to one cluster where distance from centroid in minimum of all clusters.
As a result of clustering, the generated labels were added to final dataset containing all information as shown below.

	index	City	Latitude	Longitude	Venue	id	Venue Latitude	Venue Longitude	Venue Category	main_category	total likes	venue rate	Cluster label
0	0	Manama	26.228	50.586	Bahay Kubo Restaurant	4d07a2f18620224b7e34b940	26.228443	50.582647	Asian Restaurant	asian_cuisine	48	star4	3
1	1	Manama	26.228	50.586	Sameeh pastries	4cb0419acbab236acbc29e73	26.226697	50.588958	Turkish Restaurant	mideast_bbq_cuisine	38	star4	3
2	2	Manama	26.228	50.586	The Lebanese Restaurant	4d288e6ff7a9224b630ef69f	26.228518	50.583612	Middle Eastern Restaurant	mideast_bbq_cuisine	45	star4	3
3	3	Manama	26.228	50.586	Al Wasmiya Village Restaurant	4e270708ae609b2f94ecc680	26.229518	50.581646	Restaurant	mideast_bbq_cuisine	11	start2	2
4	4	Manama	26.228	50.586	Jahan Grills	4dd7f705b0fb8af380937ea3	26.224671	50.587759	BBQ Joint	mideast_bbq_cuisine	55	star4	3

Figure 12: Venue details - cluster label

4. Results

The results of K-means clustering for 289 venues are visualized on below map where the yellow markers represent cluster 0, blue markers – cluster 1, green markers- cluster 2 and red markers – cluster 3.

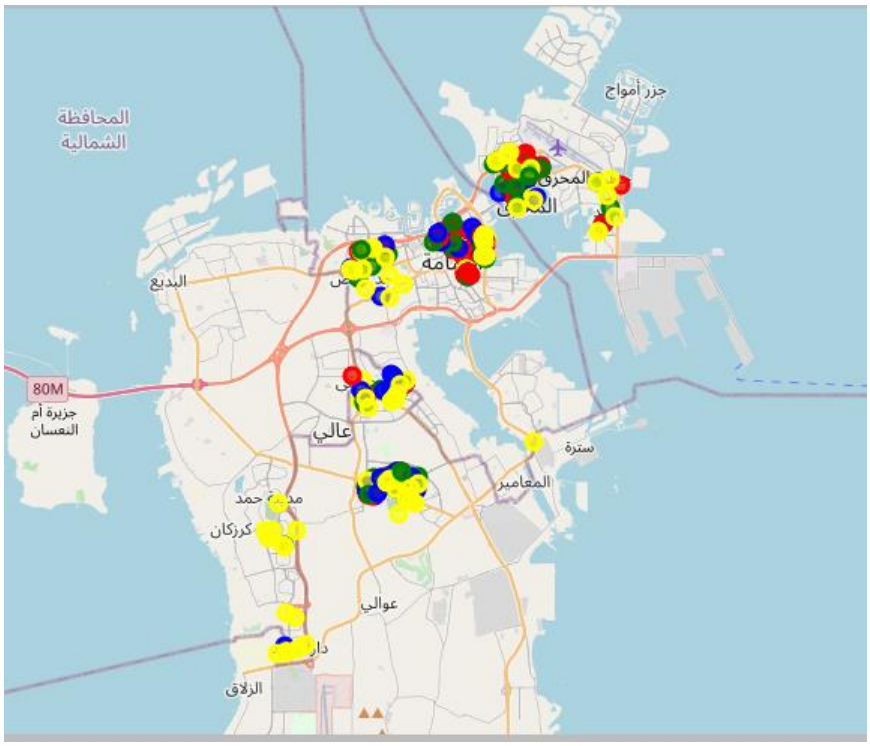

Figure 13: visualization of clusters on map

Figures 14 and 15 show that total number of venues is largest in cluster 0 followed by cluster 3 and 1 with 72,71 venues respectively and lowest in cluster 2. Despite convergence in number of venues, the highest user recommended venues are in cluster 3 then cluster 1,2 while least in cluster 0 based on user total likes .

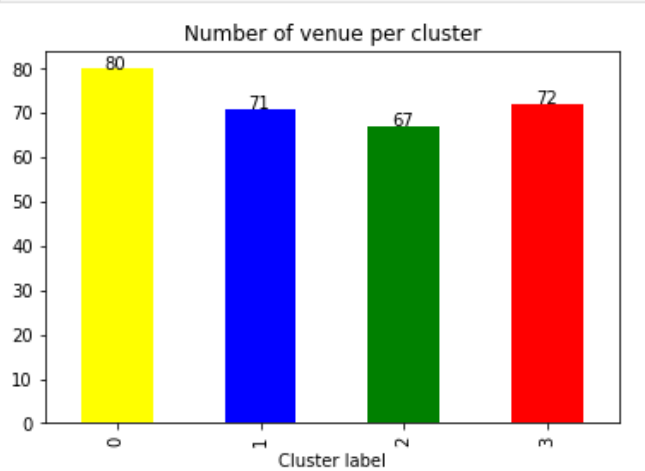

Figure 14: Number of venues per cluster

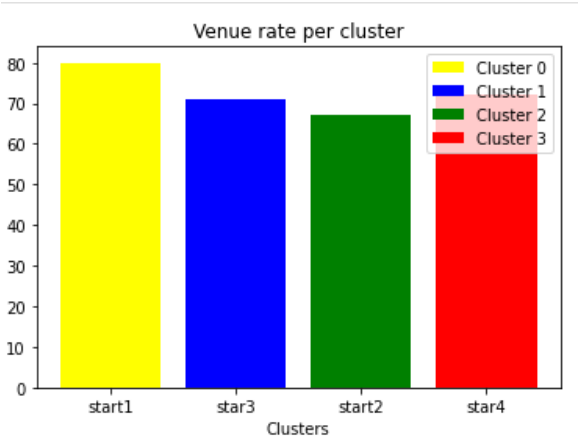

Figure 15: Venue rate per cluster

5. Discussion

The main characteristics of clusters can be described as follows:

- **Cluster 0**, has 80 venues with lowest rating (1 star) where mainly located at Ar Rifa city. The most common categories are food, hotels, park and shopping. There are many coffee shops, breakfast and different cuisines like Middle East, Italian and Asian.

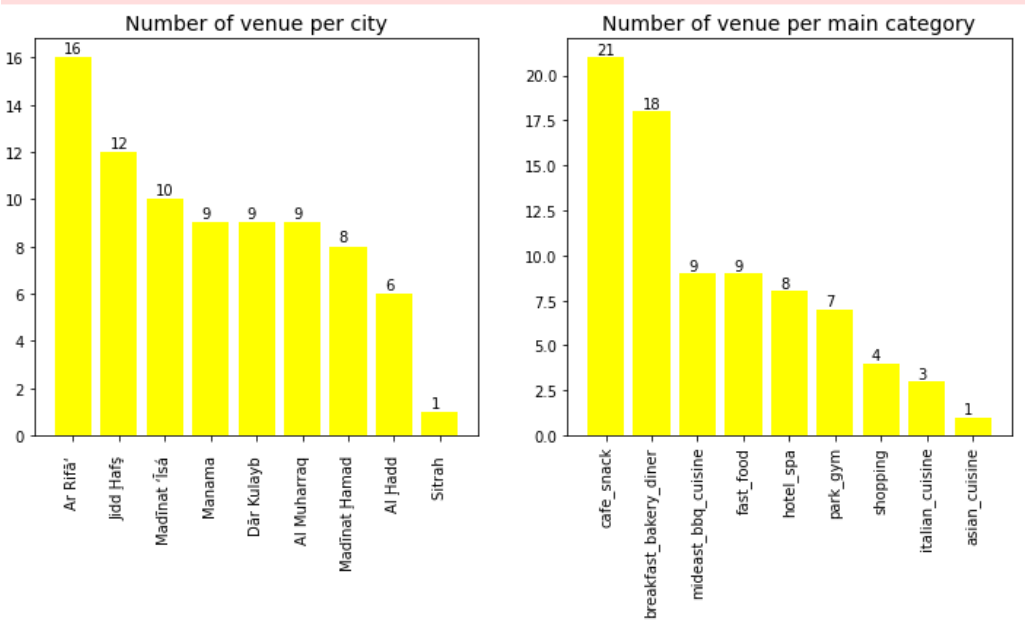

Figure 16: Num of venues per city and main category – cluster 0

- **Cluster 1**, has 71 venues of 3 star rating where mainly located at Ar Rifa city. The most common categories are food, shopping, hotel, cinema, museum and park. There are many coffee shops, breakfast and different cuisines like Middle East, fast food and Asian.

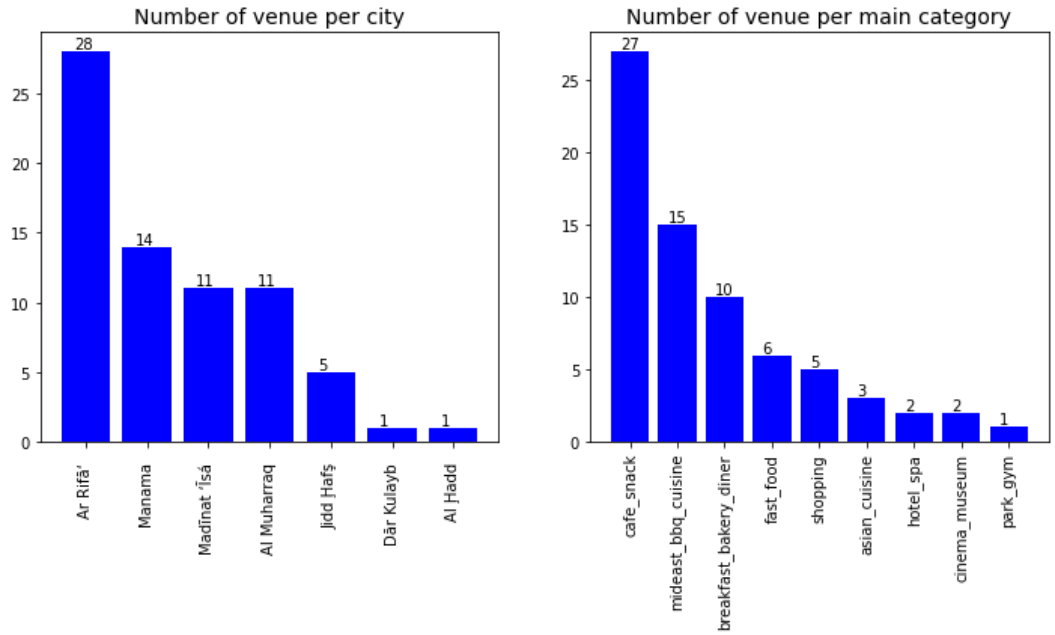

Figure 17: Num of venues per city and main category – cluster 1

- **Cluster 2** , has 67 venues of 2 star rating where mainly located at Riffa , Muharraq , Manama and Madinat Isa. The most common categories are food, ,hotel, shopping ,park and cinema. There are many coffee shops , breakfast and different cuisines like middle east ,fast food , Italian and Asian.

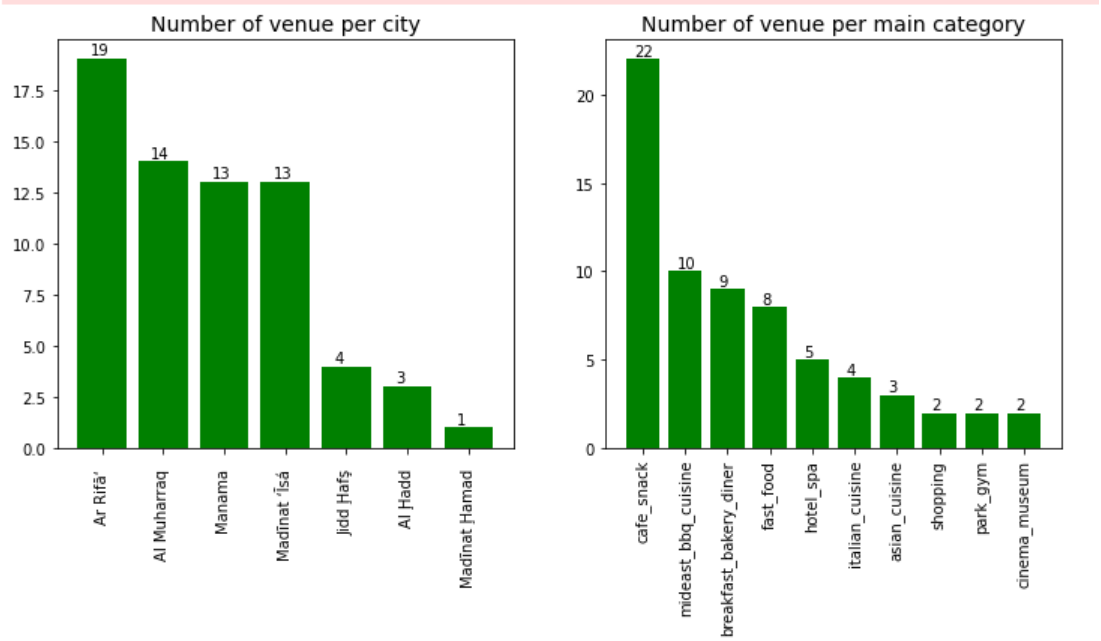

Figure 18: Num of venues per city and main category – cluster 2

- Cluster 3, has 72 venues and best rating among other clusters and mainly located at Manama city. The most common categories are food, hotel, park, shopping, cinema and museum. There are many coffee shops, breakfast and different cuisines like Middle East, fast food and Asian.

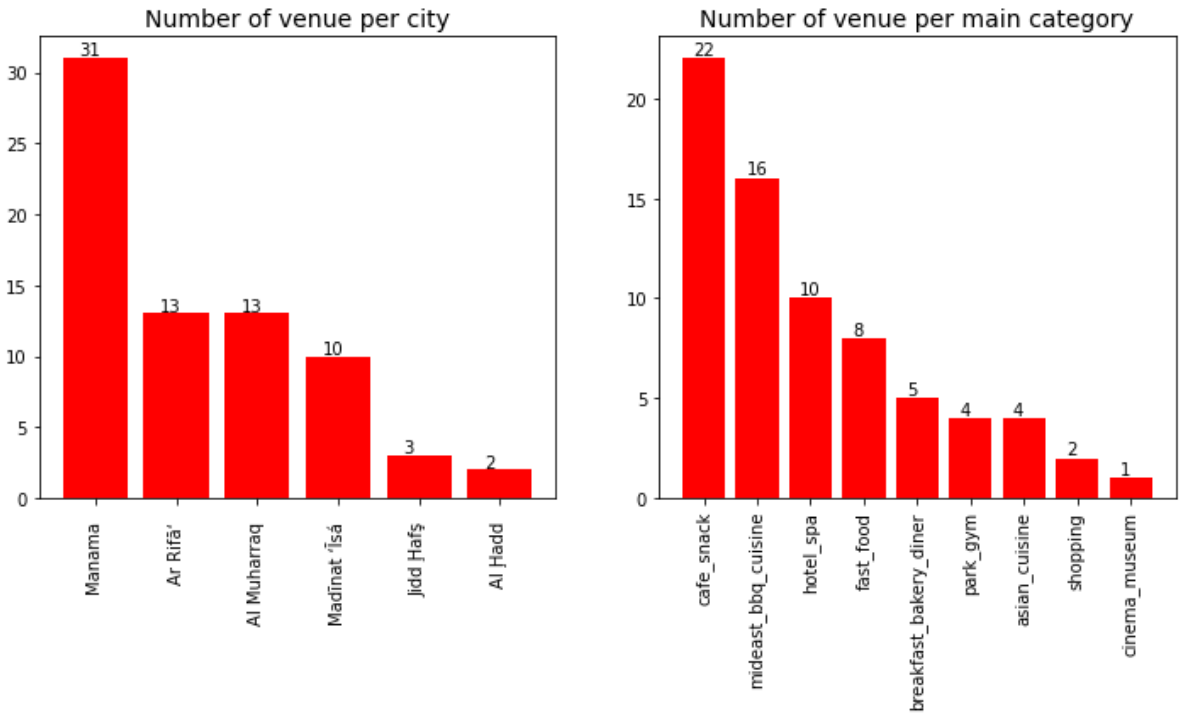

Figure 18: Num of venues per city and main category – cluster 3

6. Limitation and Future Research

This project focused only on largest cities in Bahrain since other cities data couldn't be found easily online .Further research could include other cities than just 9 cities; it would be valuable to add more cities to analysis. Moreover, Free Foursquare account is used to get venue data but it has limitation like number of requests per day and return results, paid account can be used to overcome such limitations.

7. Conclusion

This project started with identifying business problem, extracting and cleaning data then data clustering using k-means into 4 clusters. The findings show venues of cluster 3 are highly recommended where highly rated places, starting from hotels , park ,gym ,historical site, shopping malls , coffee shops and cuisine from different cultures .