Comparison and Clustering of Neighborhoods in Cairo, Istanbul, and Marrakesh

Table of Contents

1. Introduction

This study focuses on comparing neighborhoods across three major capitals: Cairo, Istanbul, and Marrakesh. The objective is to analyze neighborhood characteristics using online data sources, mainly to assist business owners in identifying similar or potentially successful locations for business expansion.

2. Data Sources

Web Scraping and API Usage

Wikipedia: Neighborhoods data is scraped from Wikipedia using BeautifulSoup for Cairo, Istanbul, and Marrakesh.

Geocoding: Geographical coordinates are obtained using the Geopy library, which interfaces with the Nominatim API.

Foursquare API: Venue data for each neighborhood is fetched using the Foursquare API, which provides information about various places within a specified radius.

Python Libraries

Pandas: For creating and manipulating dataframes.

NumPy: For numerical operations.

Requests and BeautifulSoup: For fetching and parsing HTML pages.

Geopy: For converting location names into latitude and longitude.

Folium: For map visualization.

Sklearn: For applying KMeans clustering.

3. Data Collection and Cleaning

Data collection involves fetching lists of neighborhoods from predefined Wikipedia pages for each city. This data is then cleaned to ensure consistency in neighborhood names and to prepare for geocoding.

Below links were used to fetch the 3 Capitals data which later sorted in datasets):

https://en.wikipedia.org/wiki/Category:Districts_of_Cairo

https://en.wikipedia.org/wiki/List_of_districts_of_Istanbul

https://en.wikipedia.org/wiki/Subdivisions_of_Marrakesh

4. Geolocation Data

Using Geopy's Nominatim service, each neighborhood's name is converted into latitude and longitude coordinates. Error handling is implemented to manage request timeouts and missing coordinates, with retries and alternative spellings used as necessary.

5. Visualization

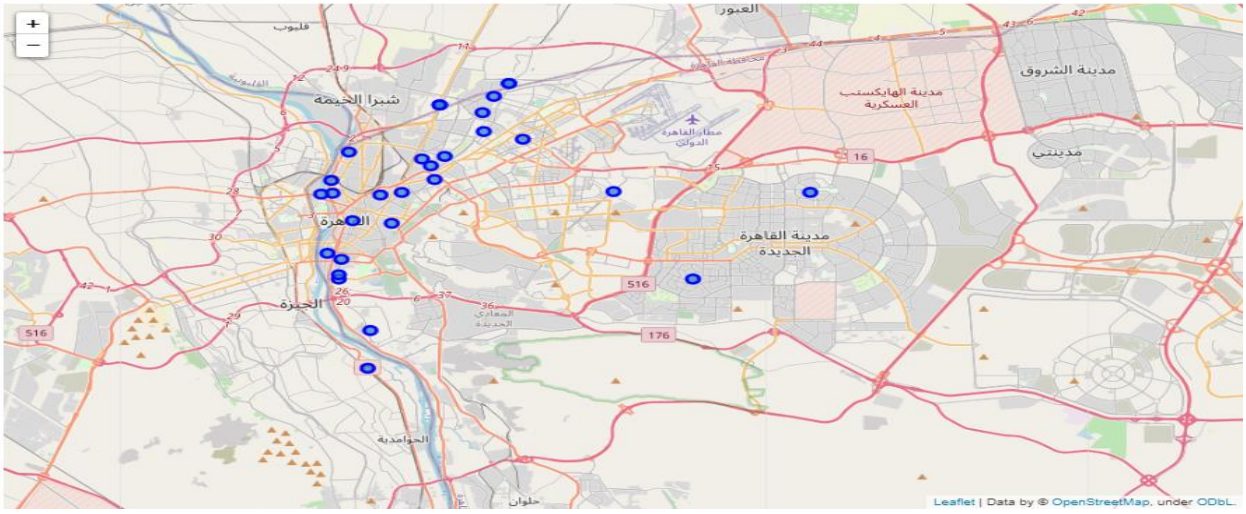
Maps are created using Folium to visualize the neighborhoods:

A global map displaying all three cities.

Individual maps for each city showing detailed neighborhood locations.

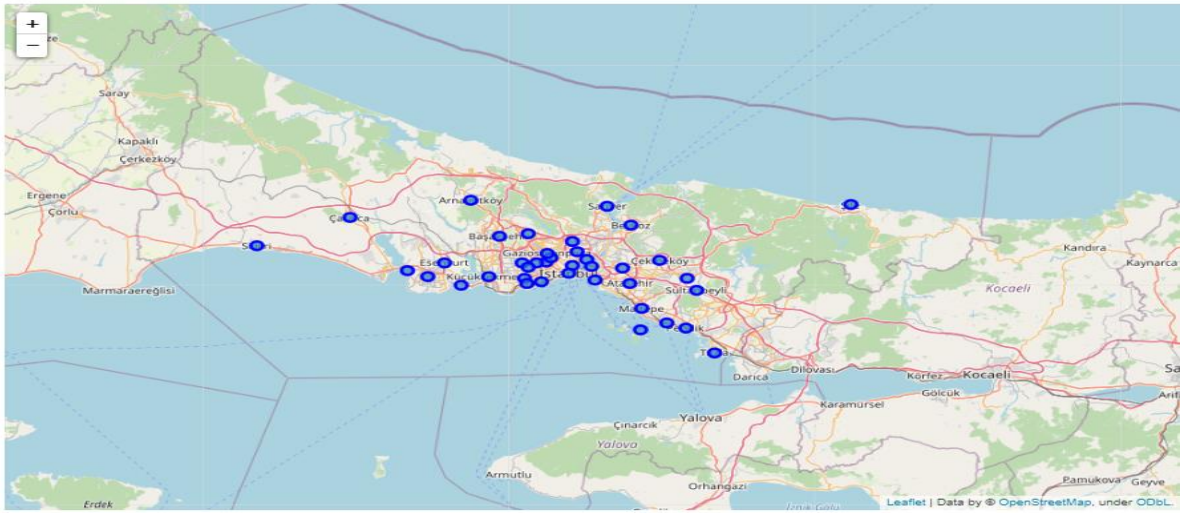
Cairo Map:

```
map_Cairo = generate_map('Cairo', 11)
map_Cairo
```



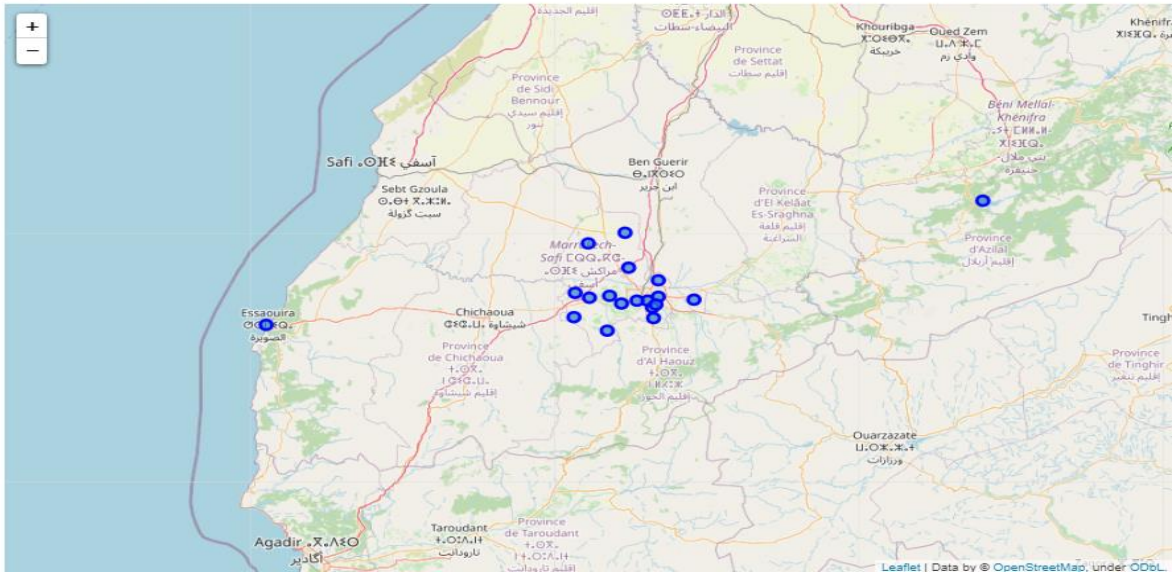
Istanbul Map:

```
map_Istanbul = generate_map('Istanbul', 11)
map_Istanbul
```



Marakesh Map:

```
map_Marrakesh = generate_map('Marrakesh', 11)
map_Marrakesh
```



6. Venue Data Collection

Using the Foursquare API, venue data within a 1000-meter radius of each neighborhood center is collected. The data includes venue names, categories, and locations, which are crucial for the clustering analysis.

City	No. of Neighborhoods	Avg Neighborhood Radius Considered (m)
Cairo	32	1000
Istanbul	38	1000
Marrakesh	33	1000

Venue Analysis

Venues are categorized, and one-hot encoding is performed to facilitate the clustering process. Dataframes are created to summarize the number of venues per neighborhood and the most common venue types.

The venues for each location are stored in separate dataframes.

To study the venues, the dataframes containing the venues are grouped by neighborhoods and summed up.

The number of venues per each neighborhood depends on the Capital Geographical, Taken radius.

As per above table: A 1000 m Radius was used for all 3 Capitals, which can be changed later on to enhance our results and get more venues for every Neighborhood.

For Cairo and Istanbul, we see many venues for ever neighborhood, but Marrakesh venues are less. Realistically this is not true, and this can be considered as Foursquare not having detailed venues for all neighborhoods. My assumption is that venues in Foursquare are more recognizable and an international restaurant franchisee will like to be in neighborhoods with more recognizable venues.

Clustering

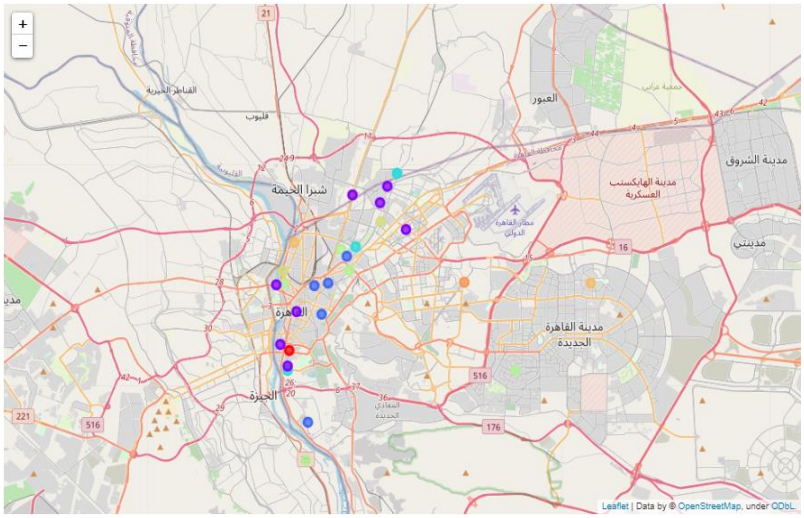
Neighborhoods are clustered based on venue similarity using the KMeans algorithm. Two types of clustering are performed:

Individual Clustering will help understand how the individual locations can be clustered. To be consistent with all the individual location clustering and the complete clustering, there are going to be 8 clusters.

It must be noted that cluster labels are not the same across different locations.

Below are the summary of 3 Capital created (Individual Clustering):

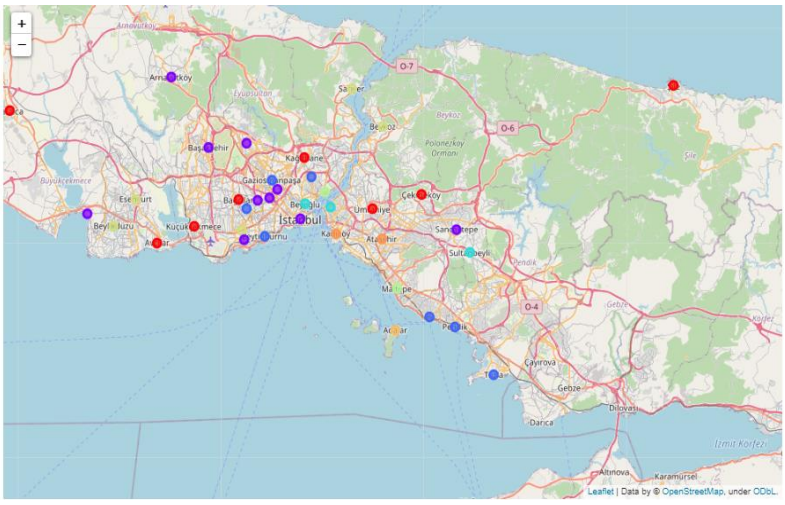
Cairo Individual Clustering:



Cluster Labels	Neighbourhood
0	1
1	9
2	5
3	2
4	3
5	1
6	4
7	6

8 rows × 25 columns

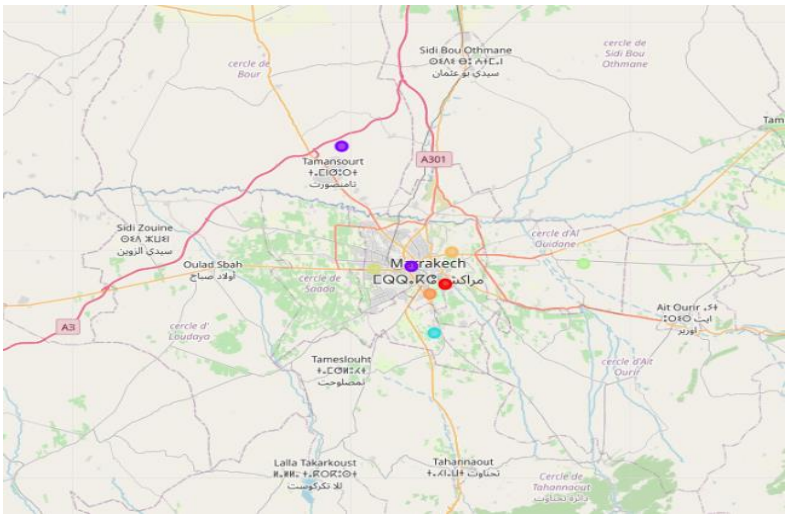
Istanbul Individual Clustering:



Cluster Labels	Neighbourhood
0	9
1	10
2	7
3	5
4	1
5	2
6	3
7	2

8 rows × 25 columns

Marakesh Individual Clustering:



Cluster Labels	Neighbourhood
0	1
1	2
2	1
3	1
4	1
5	1
6	1
7	1

8 rows × 25 columns

Individual Clustering Summary:

I. In Cairo: There are 30 Clusters, (1,2,7) having most of the neighborhoods.
II. In Istanbul: There are 39 Clusters, (0,1,2) having most of the neighborhoods.
III. In Marrakesh and as mentioned in the clustering map, very few neighborhoods available, and it's difficult to judge if cluster with 1 neighbor is outlier since most on them having 1 or 2 neighbors.

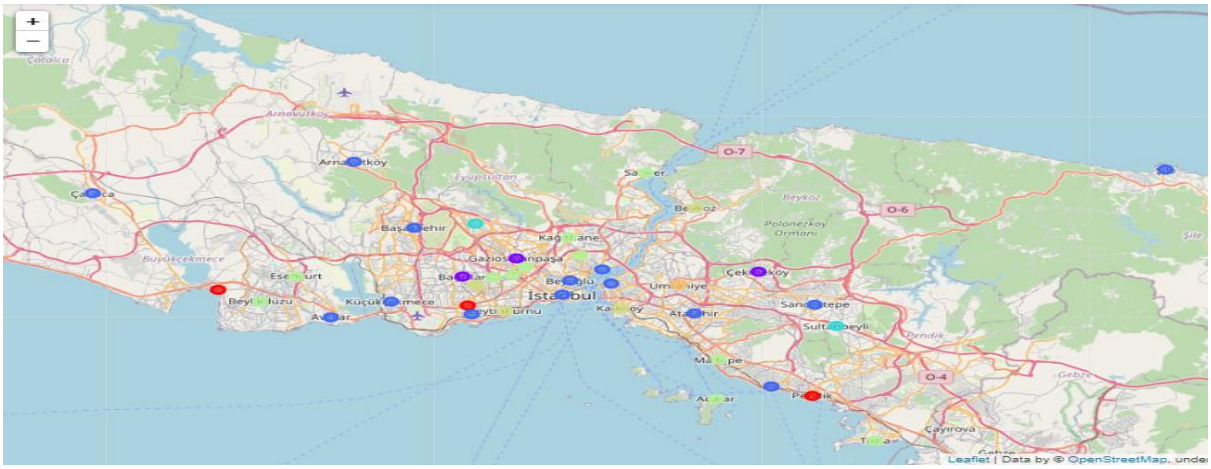
Note: for Marrakesh (10 out of 19 neighborhoods returned zero values by foursquare) and were not appended in the dataframe, its recommended to regenerate the data of neighborhoods again to get more venues, and that can be done by increasing the radius as mentioned above also by correcting some correcting names of neighborhoods which were rejected by foursquare .

Complete Clustering: Across all cities to identify similar neighborhoods across borders.

8. Clustering Results

Cairo Complete Clustering:

Istanbul Complete Clustering:



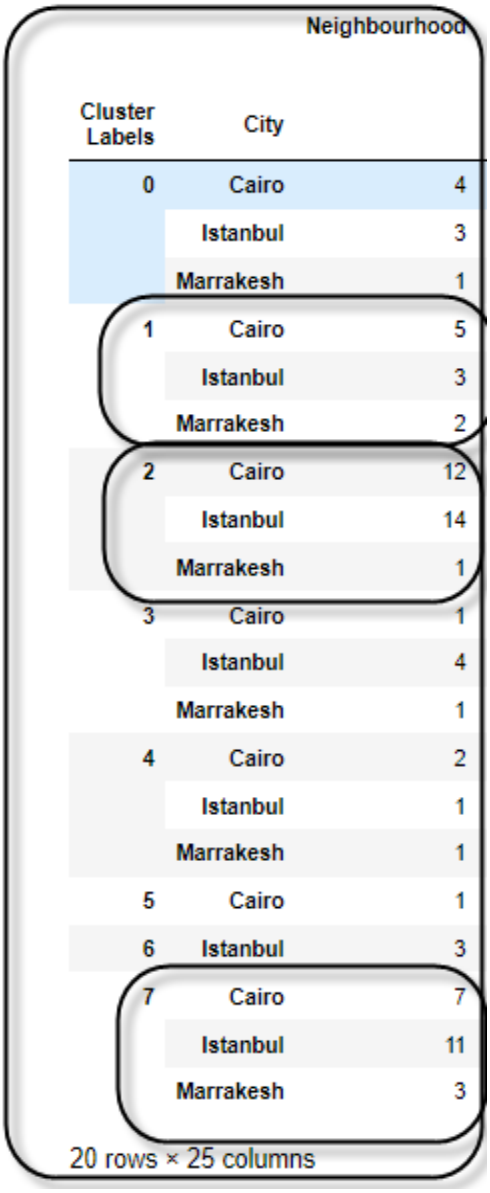
Marakesh Complete Clustering:

The results of the clustering are visualized on the city maps, with each cluster represented by a different color. Summaries of the clustering include:

The most common venues in each cluster.

Potential similarities between neighborhoods in different cities based on their cluster groupings.

9. Conclusion and Recommendations



Cluster Labels	City	Neighbourhood
0	Cairo	4
	Istanbul	3
	Marrakesh	1
1	Cairo	5
	Istanbul	3
	Marrakesh	2
2	Cairo	12
	Istanbul	14
	Marrakesh	1
3	Cairo	1
	Istanbul	4
	Marrakesh	1
4	Cairo	2
	Istanbul	1
	Marrakesh	1
5	Cairo	1
6	Istanbul	3
7	Cairo	7
	Istanbul	11
	Marrakesh	3

20 rows × 25 columns

The analysis identifies clusters of neighborhoods across Cairo, Istanbul, and Marrakesh that share similar characteristics, particularly in terms of venue types. These insights can guide business owners in making informed decisions about where to expand based on existing successful locations.

- Cluster 2 and 7 having highest Neighboring number for both Cairo and Istanbul

The big takeaway from this is that there are three clusters with Neighbourhoods from all the locations (2, 5, 6). So, these Neighbourhoods can be considered similar based on the venues present in them.

Discussion

The objective of this analysis was that if there is a restaurant in both Cairo and Istanbul and we want to open Marrakesh then in which Neighbourhoods of the cities they should open. Based on Complete Clustering Neighbourhoods in clusters 1,2 and 7 are similar Neighbourhoods. So, if the restaurant in the Neighbourhoods of these clusters in Cairo and Istanbul then a new franchise can be opened in the Neighbourhoods of the same clusters in Marrakesh and . Since majority of the Neighbourhoods of all the locations are in these three clusters then there is a good probability of finding a match.