

Opening a new Coffee Shop in Dubai City, UAE

1. Discussion and Background of the Business Problem:

Introduction Section

This final project explores the best locations for coffee shop throughout the Dubai city. Dubai Located in the eastern part of the Arabian Peninsula on the coast of the Persian Gulf, Dubai aims to be the business hub of Western Asia. A center for regional and international trade since the early 20th century, Dubai's economy relies on revenues from trade, tourism, aviation, real estate, and financial services. With its diverse culture, come diverse food items. In another hand, the coffee shop has become a promising business, especially in an emergent business city like Dubai. The objective of this project is to find the best neighborhood to open a coffee shop by avoiding highly competitive market.

In this project, two criteria will be considered:

- 1- Minimum coffee shop in the area to minimize competition
- 2- Population density to maximise profit

Target Audience

- Business investor who wants to open a coffee shop.
- Budding Data Scientists, who want to implement some of the most used.
- Exploratory Data Analysis techniques to obtain necessary data, analyze it and, finally be able to tell a story out of it.

Data Selection

1. For the above objective, we will be using open-data acquired from the Dubai Statistics Center. The data is available in the form of an Excel sheet, which will require a considerable amount of refinement. The data source is accessible at below location:

Data Source: https://www.dsc.gov.ae/Report/DSC_SYB_2019_01%20_%2002.xlsx

- Description: This data set contains the required information. And we will use this data set to explore various neighborhoods of Dubai city.
- <https://en.wikipedia.org/wiki/Dubai>

2. Coffee shop in neighborhoods of Dubai city.

- Data Source: Foursquare API
- Description: By using this API we will get all the venues in the neighborhoods. We can filter these venues to get only Coffee shop.

Approach

- Collect the Dubai city data
from https://www.dsc.gov.ae/Report/DSC_SYB_2019_01%20_%2002.xlsx.
- Using Foursquare API we will get all venues for each neighborhood.
- Filter out all venues which are coffee shop.
- Analyzing using Clustering (Specially K-Means):
 1. Find the best value of K
 2. Visualize the neighborhood with a number of coffee shop.
- Compare the Neighborhoods to Find the Best Place for Starting up a coffee shop

Problem Statement

1. What is the best location for a coffee shop in Dubai city?
2. In what Neighborhood should I open a coffee shop?

2. Data Preparation:

After data extraction, first data cleaning is done. Dubai city data where extracted and cleaned following these steps:

- 1- A considerable amount of header and footer are present in the report of data extracted, so it is necessary to remove it.

- 2- The report also contains additional columns which we do not require as they represent the same information in Arabic.
- 3- if you look at the report, the communities are split in Sectors which are showed in the report as splitter rows. First, we will get the index of the row having text “sector”, then dropping the row based on the found index.
- 4- Order of column is changed in our data frame, so community is put first and the second column is population. Then, data are arranged with taking population descending order

	community	population
56	MUHAISANAH 2	196316
163	WARSAN 1	106072
23	HOR AL ANZ	83187
147	JABAL ALI 1	75287
77	AL KARAMA	75066

Figure 1 : Dubai city data after cleaning

Second data set which is coordinates extracted using GeoPy by leveraging Google Maps or some other data source provider are also cleaned or prepared. It was noticed that area names have a suffix like FIRST, SECOND, THIRD, ETC, while the same areas were marked with number 1, 2, 3 in Google Maps. Consequently, GeoPy will not find the coordinates of a neighborhood. As a solution, we will replace the suffix (FIRST, SECOND...) with numerical values (1,2,...).

Since we are interested only interested in commercial and residential areas for our coffee shop, all industrial areas will be removed.

Some of the names of locality in this dataset were not as they are represented in map providers. For example, 'Al Quoz' is named as 'Al Goze.' This can cause inconsistency and may leave us excluding the populated areas from our analysis. 35 naming corrections were done.

Now that we have our desired dataframe, we will proceed to stage 2 of our work. In stage 2, we will extract each community's coordinates and append it to our data frame. To minimize

the time required to extract such information, we will be obtaining the coordinates of the top 100 communities with the highest population. In this data set, NaN values are dropped.

3. Folium Mapping:

Folium will be used for mapping. The communities' locations are shown based on our dataset. With Folium, we will map out Dubai and then place markers for each community we have in our dataframe. Communities are presented on the figure below.

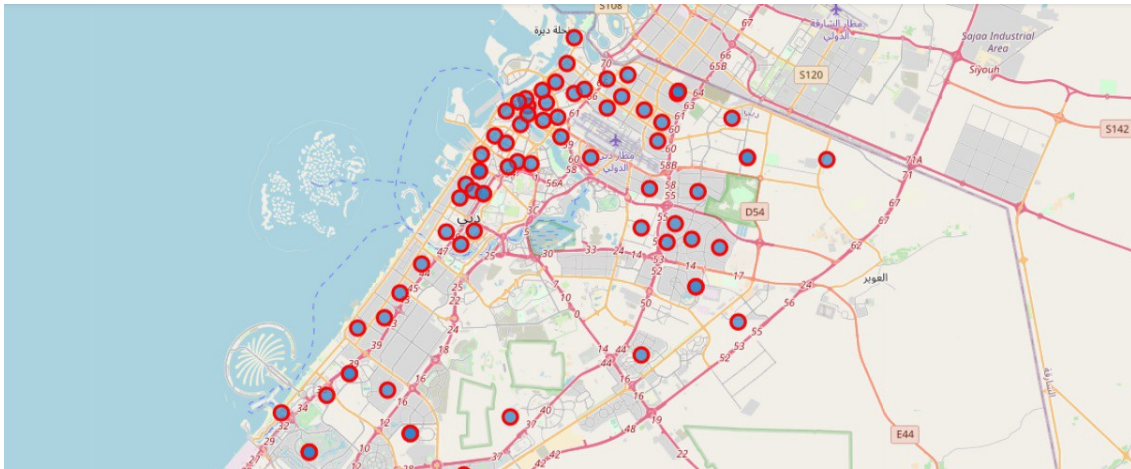


Figure 2 : Communities' locations with Folium mapping

Because we sorted our dataframe based on population and we picked top 100 communities out of the complete dataset, we are able to cover most of the residential/commercial communities.

4. Foursquare:

We will start with a single community. First, we will get coordinates and then generate GET URL for Foursquare API call. We will be requesting for Top 100 venues in the locality.

6 venues were returned by Foursquare for MUHAISANAH 2 (figure 3).

	Community	Neighborhood	Latitude	Neighborhood	Longitude	Venue	Venue Latitude	Venue Longitude	Category
0	MUHAISANAH 2		25.280555		55.410502	McDonald's (مكتوناندز)	25.281423	55.411649	Fast Food Restaurant
1	MUHAISANAH 2		25.280555		55.410502	LuLu Center	25.280531	55.410506	Grocery Store
2	MUHAISANAH 2		25.280555		55.410502	McDonald's LuLu Village	25.281949	55.411271	Fast Food Restaurant
3	MUHAISANAH 2		25.280555		55.410502	Al Ansari Exchange	25.281536	55.411369	Currency Exchange
4	MUHAISANAH 2		25.280555		55.410502	Lion Gym	25.278308	55.412009	Gym

Figure 3 : Venues returned by Foursquare for MUHAISANAH 2.

Now, we will create a function which will loop through all communities and will compile the list venues. We obtained 198 unique categories.

5. Prepare & Analyze

Data will be prepared so it can conform to ML standards. Categories provided by Foursquare are in label form. The machine learning algorithms cannot operate on label data directly. They require all input variables and output variables to be numeric. To transform our data to numerical form, we will perform One Hot Encoding. This will transpose out Category labels in to Features/Columns in our dataframe with value as 0 or 1. the category mean of for each community was calculated.

In our project, we are only interested by coffee shop. Hence, we will filter our data frame since it includes above 200 categories. At the end, we will get the list of coffee shop.

6. Clustering & Analysis

At this stage, we will create clusters of communities based on where the Coffee Shop are situated. Once we have a visual of the cluster, we can start breaking those clusters and see how many are in each cluster. In our case, we decided to cluster our community into 5. Second step, is to create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood. Finally, we merged our dubai_grouped data with dubai_venues_data to add latitude/longitude for each neighbourhood.

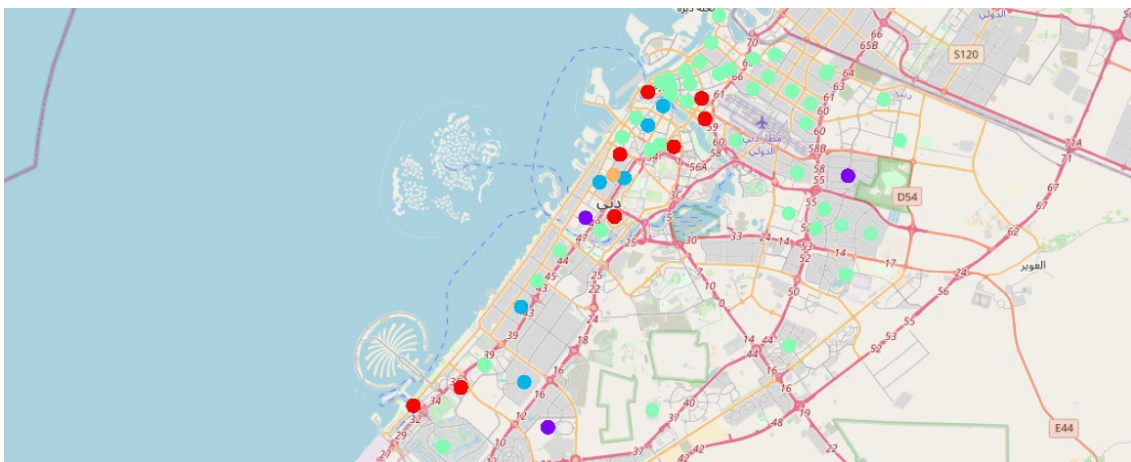


Figure 4 : Dubai city clusters

7. Conclusion

As you can see from the above Folium map, Communities in cluster 3, marked as Green, do not have any 'Coffee Shop,' Which gives us a lot of choices to select where we want to start our business. There are certainly fewer Coffee Shop in communities in cluster 2, 4 and there is a potential to have a successful business.