A. Introduction

One of the biggest factors when you're trying to realize your dream of opening a coffee shop business is the location that you choose. The location of a coffee shop, or just about any type of business for that matter, plays a huge role in the success of the shop. Choosing the right location is key to any good business endeavor.

For this project I choose Dubai city, Dubai one of the world's fastest growing economies with a population is estimated at around 3,400,800, Business startups take advantage of the enormous commercial prospects and innovative business modules offered by the city.

Thus, the project goal is to figure out the best locations for opening up a new coffee shop in Dubai city using the clustering technique, namely *k-Means*.

To ensure this project's success we need to look for:

- **Demographics**

  We need to look into who is in the area we're considering. There are certain types of people that likely will be looking for coffee more than others. So because of this, we should look for a location that has higher populations of the kind of people we're talking about. Examples would be college students, parents, shoppers and touristists.

  For this purpose I use *Foursquare API* to explore the popular venues around the location of each Community or Neighborhood of Dubai city and filter the result to get just these following category:

  1. College and Universities.
  2. Shopping malls.
  3. Historical sites.

- **Competition**

  What kind of competition there is in the area you're looking at is very important for you to know. If you open a coffee shop or other type of coffee business in an overly saturated market then the likelihood that your business will succeed will be much lower.

  So using *Foursquare API* I get the coffee shop each Neighborhood of Dubai city have.

- **Neighboring venues**

  Now let's talk about businesses and venues that if you're close to, it can actually help drive your sales up more, Such as Gyms, or venues related to Foot and car traffic and require that customers have to wait for something and venues that cater to families and having fun.

  For this study I use *Foursquare API* to explore Dubai Neighborhoods to find these venues which will be harmonious with the coffee shop:

  1. Banks.
  2. Gym.
  3. Parks.

B. Data acquisition and cleaning

The followings are data sources that I use for this project:

1st Data:

Dubai Communities and there coordinates as Geojson file from **Kaggle website** [Dubai Neighborhoods](#) .

2nd Data:

I use the coordinates from the previous dataset and **foursquare credentials** through its API and retrieve the popular venues along with their details and filter the result to get only the following venues that I use in clustering:

1. College and University.
2. Shopping mall.
3. Historical site.
4. Coffee shop.
5. Park.
6. Bank.
7. Gym.

C. **Methodology**

C.1. Analytic Approach:

We approach the problem using the clustering technique, namely *k-Means*. This approach enables the audience to see how similar neighborhoods about their demographics. We can then examine each cluster and determine the venue categories that distinguish each cluster, Based on that we can determine the best potential location for our coffee shop.

C.2. Neighborhoods Analysis

Initially, I get the data of Dubai Community from a Geojson file from **Kaggle website** containing all the communities and its coordinates and other features, the result DataFrame have **226 community**.

	Community	Community Code	Sector	Population	Area Sq Km	Latitude	Longitude
0	HEFAIR	991	9	0	143.565131	24.693237	55.249140
1	AL QUSAIS IND. SECOND	243	2	8834	1.728331	25.281654	55.393184
2	AL JAFILIYA	323	3	23963	1.669479	25.237635	55.286851
3	JABAL ALI SECOND	592	5	1570	5.059545	25.053979	55.115282
4	WADI AL SAFA 5	648	6	21241	16.318810	25.077814	55.352154

Next I plotted the coordinates on a map using the **folium library** as shown in the Figure.

- Given the coordinates information, I uses the Foursquare API to explore the neighborhoods, and get the **top 500 venues** within a **radius of 1 km** for each.

	Community	Community Latitude	Community Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	AL QUSAIS IND. SECOND	25.281654	55.393184	Tim Hortons	25.282066	55.397946	Coffee Shop
1	AL QUSAIS IND. SECOND	25.281654	55.393184	Madina Mall مدينة مول	25.282042	55.397978	Shopping Mall
2	AL JAFILIYA	25.237635	55.286851	Tea Star Cafeteria	25.234618	55.286414	Coffee Shop
3	WADI AL SAFA 5	25.077814	55.352154	Costa Coffee - The Villa	25.081673	55.351222	Coffee Shop
4	AL MURAQQABAT	25.265541	55.323152	Caribou Coffee	25.264900	55.320942	Coffee Shop

- Next I Filter the DataFrame to get the categories of interest only which are :

    1. Coffee Shop
    2. Historic Site
    3. Park
    4. Gym

As a result, it returns **1091 venues** with the previous category.

- Category like **Colleges and Banks** didn't show up using explore with Foursquare API URL request, so I had to use search in the request instead, as a result I get two Dataframes one for Universities and the other for Banks.
- After merging these Dataframes, it return a Dataframes got **2224 venues** with these category:

    1. College and University.
    2. Shopping mall.
    3. Historical site.
    4. Coffee shop.
    5. Park.
    6. Bank
    7. Gym

- **One-hot encoding** is done on the venue data to transform the categorical venue data into numerical data for clustering purposes and **mean of occurrence of each category** is calculated.

	Community	Bank	Coffee Shop	Gym	Historic Site	Park	Shopping Mall	University
0	ABU HAIL	0.333333	0.000000	0.0	0.0	0.666667	0.0	0.0
1	AL BADA'	0.000000	0.333333	0.0	0.0	0.666667	0.0	0.0
2	AL BARAHA	0.666667	0.333333	0.0	0.0	0.000000	0.0	0.0
3	AL BARSHA FIRST	0.333333	0.666667	0.0	0.0	0.000000	0.0	0.0
4	AL BARSHA SECOND	0.000000	0.500000	0.0	0.0	0.500000	0.0	0.0

- To determine the number of centroids for the k-means algorithm, **Silhouette Score Elbow method** are employed to choose the centroids to minimize the inertia or sum-of-squares error in the data.

  We can inspect that the **best k value** for this task is **4.**

Silhouette Score Elbow for KMeans Clustering
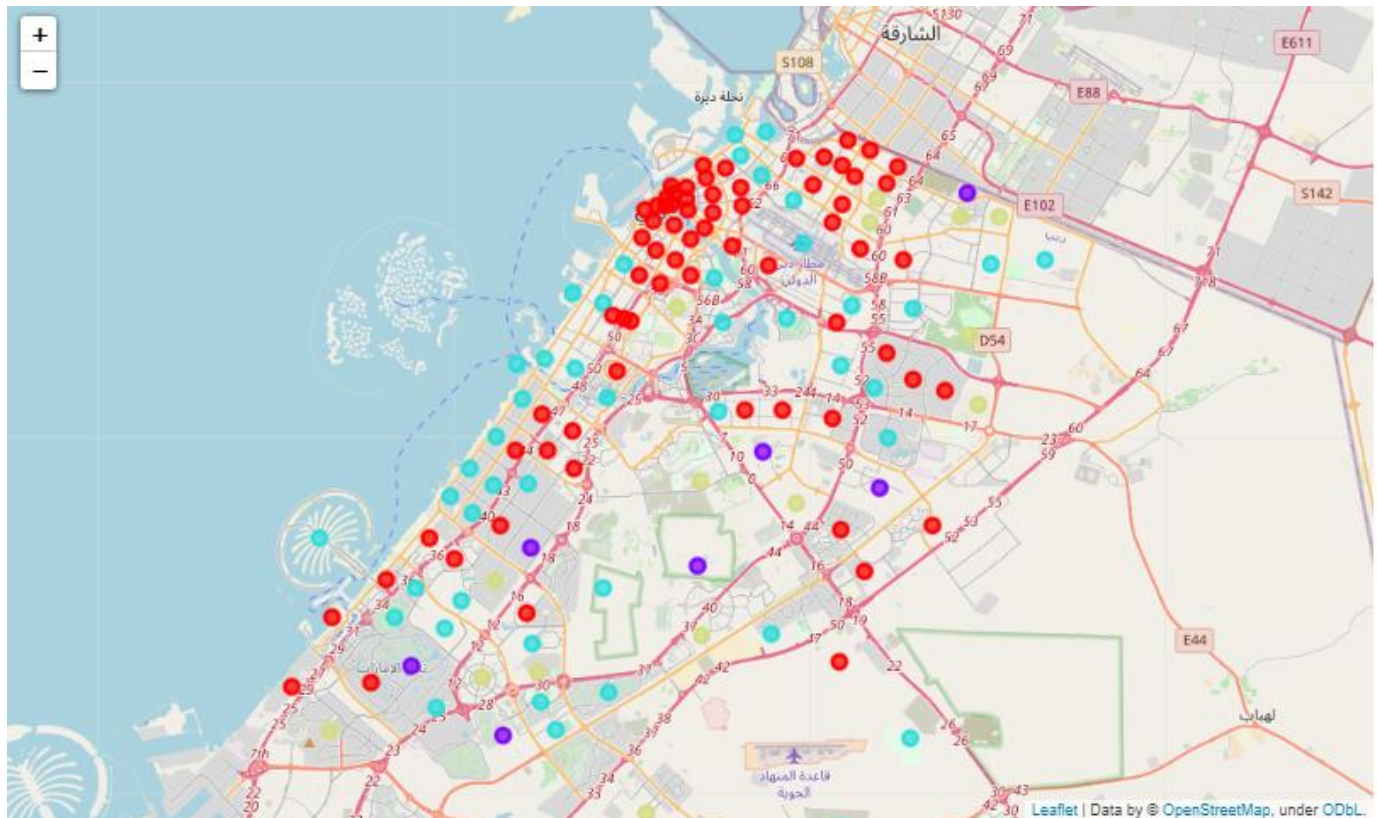
elbow at $k = 4$, $score = 0.409$
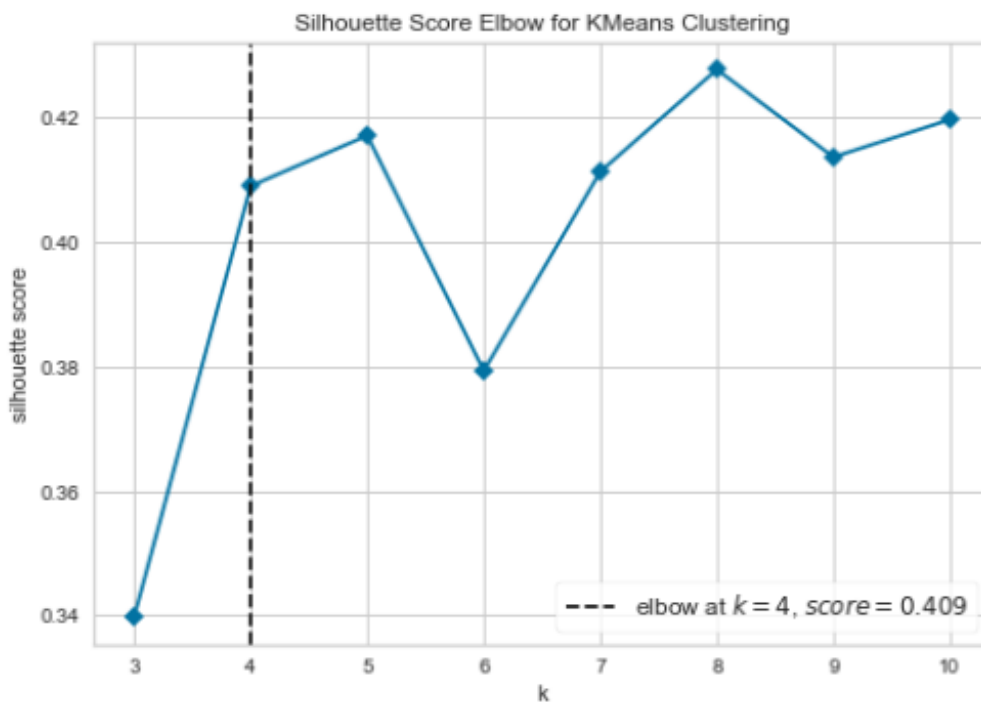
C. Results

Finally, let's visualize the resulting clusters!

The red circles for cluster 1.
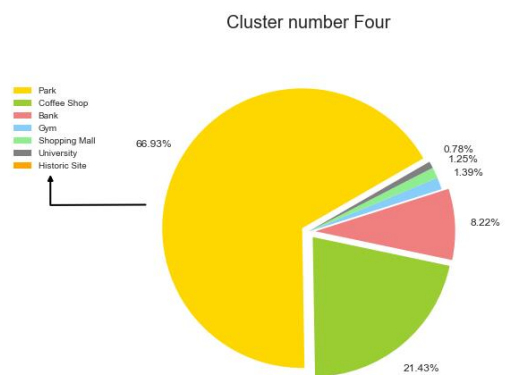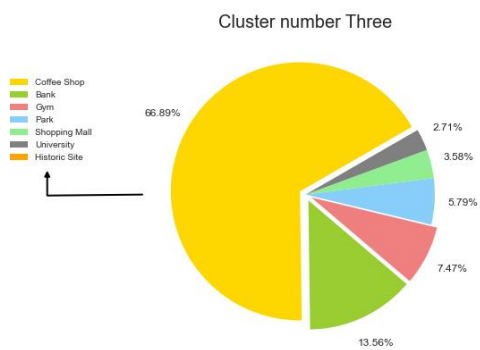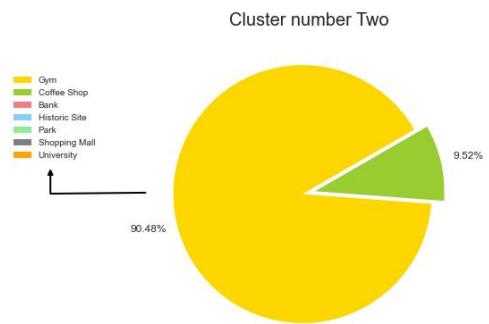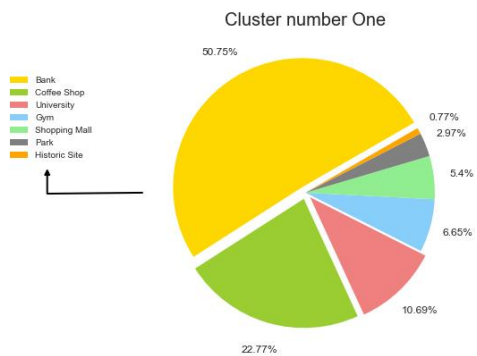
The light blue circles for cluster 3.

The green circles for cluster 4.

The purple circles for cluster 2.

As a result, we can examine venues listed inside each cluster and define the discriminating venue categories that distinguish them.


Cluster number One

Bank
Coffee Shop
University
Gym
Shopping Mall
Park
Historic Site

50.75%
0.77%
2.97%
5.4%
6.65%
10.69%
22.77%

Cluster number Two

Gym
Coffee Shop
Bank
Historic Site
Park
Shopping Mall
University

90.48%
9.52%

Cluster number Three

Coffee Shop
Bank
Gym
Park
Shopping Mall
University
Historic Site

66.89%
2.71%
3.58%
5.79%
7.47%
13.56%

Cluster number Four

Park
Coffee Shop
Bank
Gym
Shopping Mall
University
Historic Site

66.93%
0.78%
1.25%
1.39%
8.22%
21.43%

1. *Cluster 1: "Banks Venues"*

   The first cluster contains 71 communities, with the Banks as the first most common venue and coffee shop, university as the 2nd and 3rd.
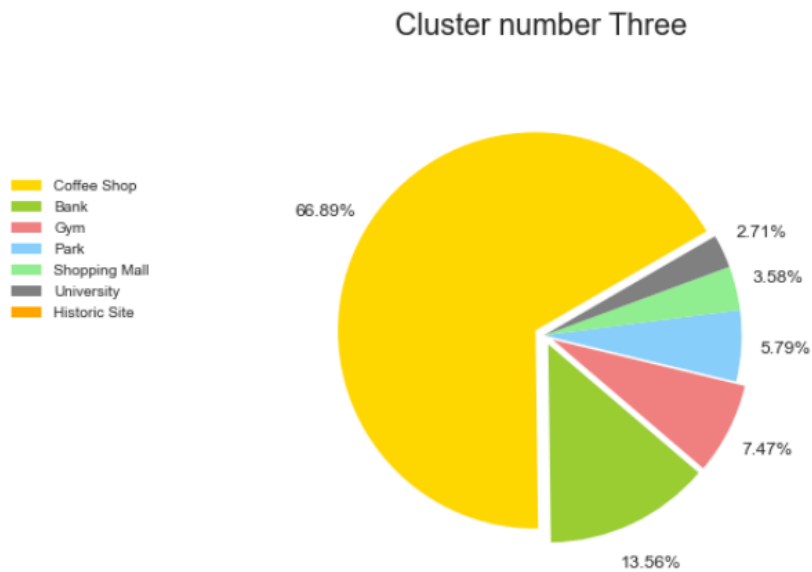
   ### Cluster number One
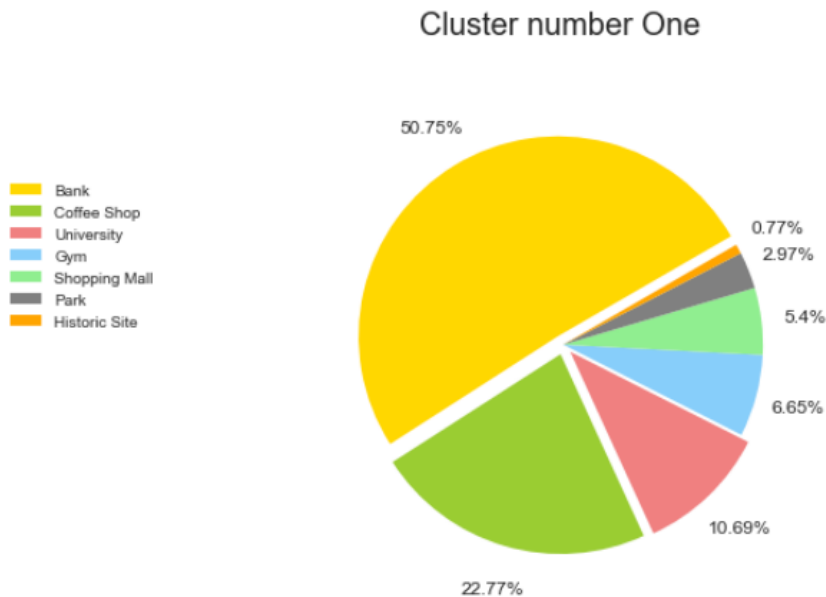
2. *Cluster 2: "Gym Venues"*

   The second cluster holds 7 communities, with Gyms the most common ones.
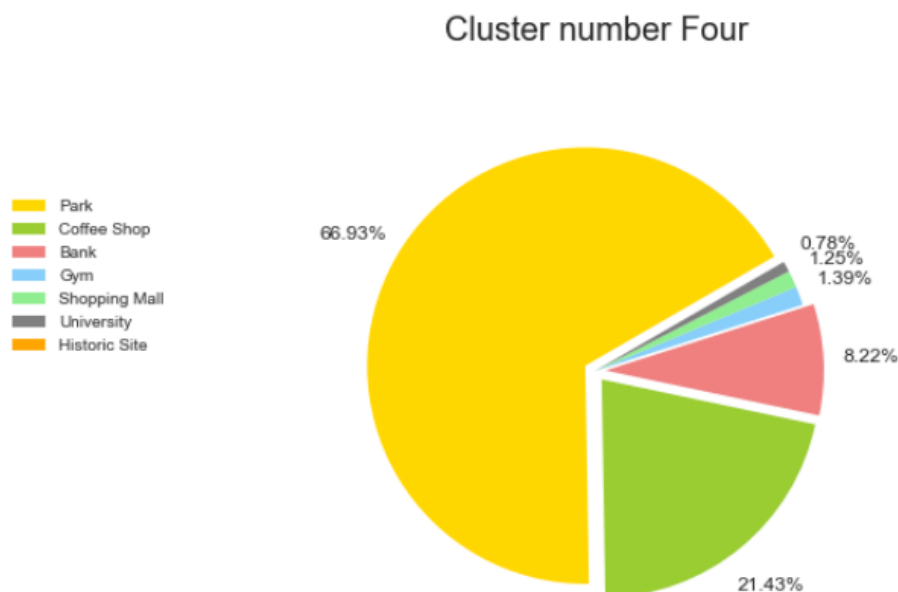
   ### Cluster number Two
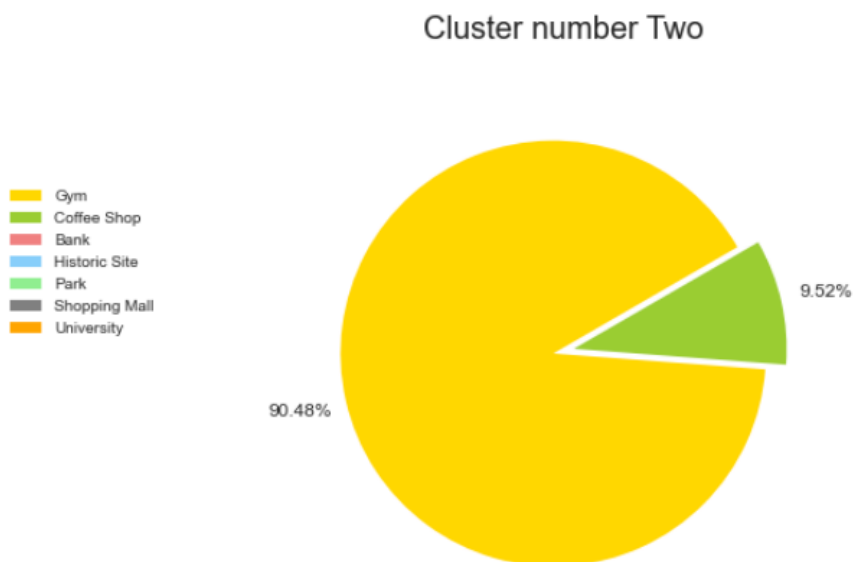
3. *Cluster 3: "Coffee shop Venues"*

The third cluster includes 45 communities with coffee shop as the most occurrence venue category.

Cluster number Three

Legend:
- Coffee Shop
- Bank
- Gym
- Park
- Shopping Mall
- University
- Historic Site

66.89%
2.71%
3.58%
5.79%
7.47%
13.56%

4. *Cluster 4: "Park Venues"*

The fourth cluster has 16 communities with the park, coffee shop, and Banks as the majority venues.

Cluster number Four

Legend:
- Park
- Coffee Shop
- Bank
- Gym
- Shopping Mall
- University
- Historic Site

66.93%
0.78%
1.25%
1.39%
8.22%
21.43%

D. Discussion

The project's main goal is to determine the best location for opening a coffee shop business in Dubai. Discussing what locations can be considered "the best" may vary, but we can equate it as the most conducive ones by considering the following criteria:

- **Demographics & Neighboring businesses:**

  Cluster 3 has the most coffee shops in their neighborhoods and could contend with the products you serve. Therefore, **cluster 3 is not recommended**

- **Cluster 1 are recommended neighborhoods** to inspect further, if you target business men or venues related to car and foot traffic and require people to wait like banks or college students, shoppers, cluster 1 might be a good choice since it has Banks as the most common venue and a variety of other venues that complete your business as universities, shopping malls and Gyms and a low percentage of competitors.

- **Cluster 4 neighborhoods'** could be a good choice if you target people who spend their morning or afternoon outside, cluster 3 might be a good choice since it has "park" as the most common venue but it has a less variety compared to cluster 1.

- **Cluster 2 neighborhoods'**
  Since Health & fitness fanatics are obsessed with coffee, juices and smoothies so if you target people who go to Gym cluster 2 could be a good choice for your coffee shop.

E. Conclusion

Finding the best location to start a business can be challenging and quite frustrating due to many uncertainties. However, we can quickly gain meaningful insights into the city and its neighborhoods with data available today. This helps everyone, including entrepreneurs, business owners, and stakeholders, to make solid decisions based on facts.