Battle of Communities

Exploring Dubai - A Data Science Problem

1. Data Acquisition

Data collection stands at the core of understanding and resolving a data science problem. Collection of authentic and reliable data is imperative to come up with a conclusion that is accurate and consistent. Hence, verifying the authenticity of the source of data becomes very important for a data scientist before the actual analysis should be started.

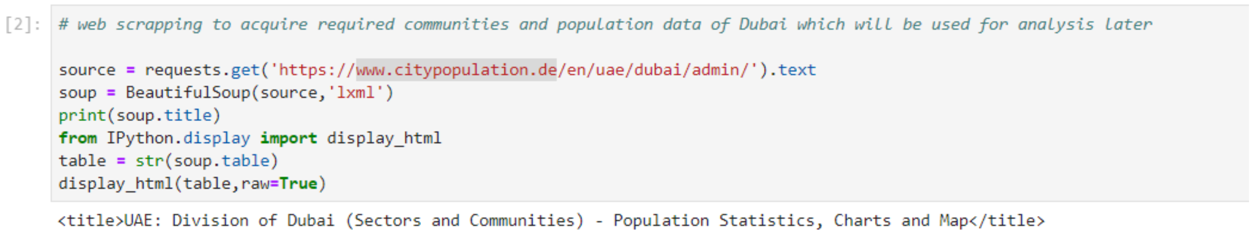
For my data science problem, I had two options available for collection of data related to communities in Dubai. I decided to acquire population data of communities and then perform restaurant analysis only for the communities which are most populated since an investor would like to open a restaurant in a community which offers more potential customers.

- Wikipedia – This contained information containing names of communities in Dubai but population data for those communities was not available for all the communities, rather it was missing for most of them. Hence, I decided to explore web more to find the statistical information somewhere else that I could use for my project.

- Fortunately, there is always a way out. I managed to find the population data of various communities of Dubai on a Dutch website (www.citypopulation.de) containing population data for major cities of the world. This data contained population estimate in 2010, 2015 and 2018. I would have liked to have 2020 population data for a more recent analysis, but I guess we can live with 2018 population estimates for now.

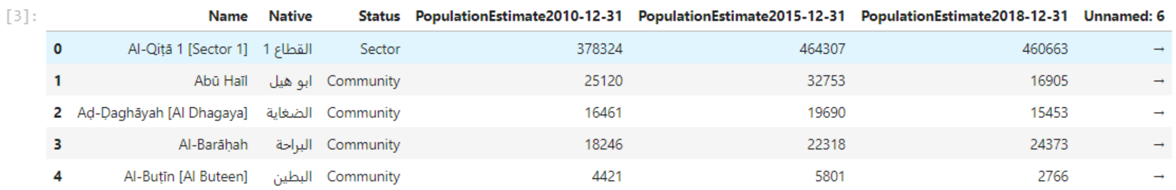
1.1. Web Scrapping

I made use of the communities' information available on the website and using Beautifulsoup4 and requests libraries in Python, managed to scrap the required data from the website.



```
[2]: # web scrapping to acquire required communities and population data of Dubai which will be used for analysis later

     source = requests.get('https://www.citypopulation.de/en/uae/dubai/admin/').text
     soup = BeautifulSoup(source,'lxml')
     print(soup.title)
     from IPython.display import display_html
     table = str(soup.table)
     display_html(table,raw=True)

     <title>UAE: Division of Dubai (Sectors and Communities) - Population Statistics, Charts and Map</title>
```

After some manipulation, I created a dataframe that looked like this.



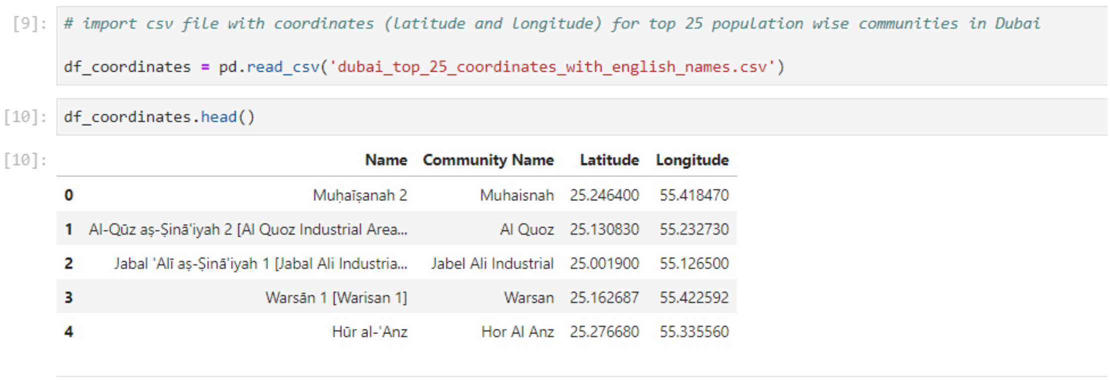
[3]:

	Name	Native	Status	PopulationEstimate2010-12-31	PopulationEstimate2015-12-31	PopulationEstimate2018-12-31	Unnamed: 6
0	Al-Qiṭā 1 [Sector 1]	القطاع 1	Sector	378324	464307	460663	→
1	Abū Hail	ابو هيل	Community	25120	32753	16905	→
2	Aḍ-Ḍaghāyah [Al Dhagaya]	الضغاية	Community	16461	19690	15453	→
3	Al-Barāḥah	البراحة	Community	18246	22318	24373	→
4	Al-Buṭīn [Al Buteen]	البطين	Community	4421	5801	2766	→

Since I am not interested in population estimates of 2010 and 2015, I decided to get rid of these columns. Similarly, community name in native language (Arabic) and the Sector column are not required for my data analysis, hence I decided to get rid of these columns as well. Finally, the column "Unnamed: 6" didn't offer ant value, hence that would also be deleted.

1.2. Location Data (GPS Coordinates; Latitude & Longitude)

This data set did not contain information for GPS coordinates for the communities. After some due diligence, I decided to work around this problem by exporting the dataframe in a .csv file and then manually adding GPS coordinates; latitude and longitude (from Wikipedia) for all Dubai communities. Also, I included a new column containing the names of the communities only in English excluding special characters that were present in earlier community names.

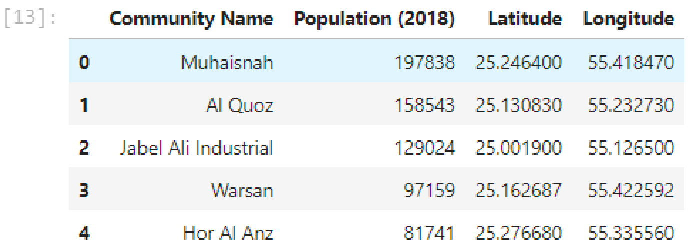


```
[9]: # import csv file with coordinates (latitude and longitude) for top 25 population wise communities in Dubai

     df_coordinates = pd.read_csv('dubai_top_25_coordinates_with_english_names.csv')

[10]: df_coordinates.head()
```

[10]:

	Name	Community Name	Latitude	Longitude
0	Muḥaīṣanah 2	Muhaisnah	25.246400	55.418470
1	Al-Qūz aṣ-Ṣinā'iyah 2 [Al Quoz Industrial Area...	Al Quoz	25.130830	55.232730
2	Jabal 'Alī aṣ-Ṣinā'iyah 1 [Jabal Ali Industria...	Jabel Ali Industrial	25.001900	55.126500
3	Warsān 1 [Warisan 1]	Warsan	25.162687	55.422592
4	Hūr al-'Anz	Hor Al Anz	25.276680	55.335560

Later, I merged two dataframe and did some manipulation to finally come up with a dataframe that I could finally use for further analysis of locations and venues especially restaurants. This dataframe appeared like this.

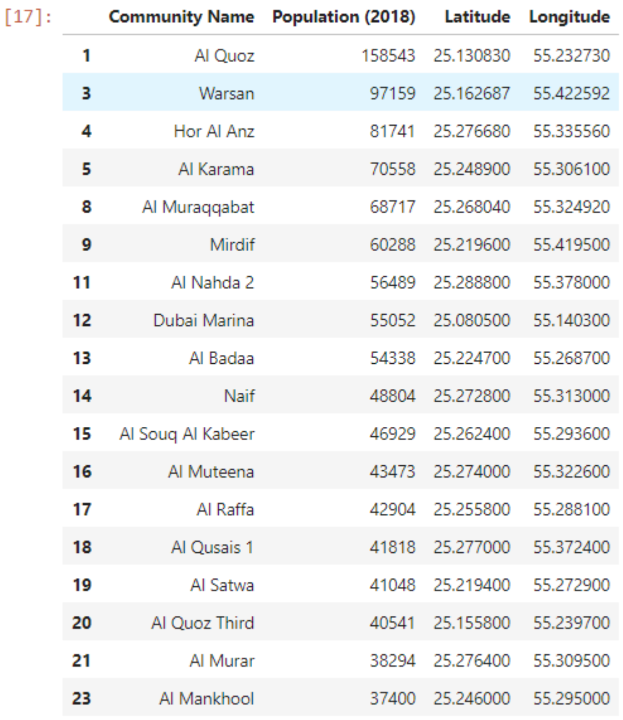


[13]:

	Community Name	Population (2018)	Latitude	Longitude
0	Muhaisnah	197838	25.246400	55.418470
1	Al Quoz	158543	25.130830	55.232730
2	Jabel Ali Industrial	129024	25.001900	55.126500
3	Warsan	97159	25.162687	55.422592
4	Hor Al Anz	81741	25.276680	55.335560

As a starting point, I decided to use top 24 highly populated communities of Dubai for further analysis since choice of opening a restaurant would be in a busy community rather than a deserted one.

We are interested in exploring venues of Dubai's populous communities, we found out during analysis using Foursquare API (explained later in this report) that six out of twenty four communities of interest have returned very less venues, hence we should not explore those communities any further as our new restaurant would definitely not be in one of those communities. After deleting the entries related to those communities, the dataframe for remaining communities appeared like this.



[17]:

	Community Name	Population (2018)	Latitude	Longitude
1	Al Quoz	158543	25.130830	55.232730
3	Warsan	97159	25.162687	55.422592
4	Hor Al Anz	81741	25.276680	55.335560
5	Al Karama	70558	25.248900	55.306100
8	Al Muraqqabat	68717	25.268040	55.324920
9	Mirdif	60288	25.219600	55.419500
11	Al Nahda 2	56489	25.288800	55.378000
12	Dubai Marina	55052	25.080500	55.140300
13	Al Badaa	54338	25.224700	55.268700
14	Naif	48804	25.272800	55.313000
15	Al Souq Al Kabeer	46929	25.262400	55.293600
16	Al Muteena	43473	25.274000	55.322600
17	Al Raffa	42904	25.255800	55.288100
18	Al Qusais 1	41818	25.277000	55.372400
19	Al Satwa	41048	25.219400	55.272900
20	Al Quoz Third	40541	25.155800	55.239700
21	Al Murar	38294	25.276400	55.309500
23	Al Mankhool	37400	25.246000	55.295000