Battle of Communities

Exploring Dubai - A Data Science Problem

Web-scrapping, Data Wrangling, Exploratory Data Analysis, Folium Map, Data Visualization, One Hot Encoding, Foursquare API, K-means Clustering & more

Author's LinkedIn: Kashif M. Sheikh

As a part of the final IBM Capstone Project, I was given this assignment to define a real-world problem that a typical data scientist goes through in his or her daily life and utilizing the knowledge gained throughout this course, solve the problem. This involved not only defining the problem but gathering the required data from web to perform the analysis and using Foursquare location data to perform the analysis and conclude the problem.

I don't have any data science background, only some programming experience having spent a few years dealing with control systems of power plants. However, this IBM data science course that I started just to learn a little about the "buzz" word, planning to spent a few weeks at the most, became a passion which ended up in me spending almost a year learning different data science tools and techniques that a data scientist would use often dealing with practical problems.

The objective of this report is to discuss a potential business opportunity of opening a new restaurant in Dubai, UAE based on the data available. The success prospects of the potential new restaurant would depend largely upon the genre/category of the restaurant and the locality/community the restaurant would be situated in.

1. Introduction: Business Case

Dubai, the city where I have been living in for five years now, is the most populous city in the UAE. 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. It is also a major global transport hub for passengers and cargo. According to government data, the population of Dubai is estimated at around 3.38 million as of January 2020.

A lot of international restaurant chains operate in Dubai and the business is booming. This promises tremendous opportunities to the potential new investors to try their fortune in this sector. This report would analyze restaurants situated in most populous communities of Dubai, their categories and would provide recommendations to potential investors on the genre of the restaurant they should think of opening in a certain locality.

Dubai is divided into various communities each having their individual community number. We will go through the steps that I have taken for creating this report; from data acquisition, wrangling and cleansing of data, performing exploratory data analysis, using Folium and Matplotlib libraries to create maps and graphs for visualization, and then finally coming up with results and conclusions.

Target Audience

Who would be interested in reading this report?

- Potential investors who would like to open a new restaurant in Dubai. This report would provide comprehensive guidance to such people.

- Customers who would like to see what dine out options they have available in a certain community close to them.

- Fellow data scientists who would like to use the analysis provided in this report to refine the results and further build models to support decisions and create stories out that data.

2. 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.

2.1. 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-Daghāyah [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.

2.2. 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-Ṣināʿiyah 2 [Al Quoz Industrial Area...	Al Quoz	25.130830	55.232730
2	Jabal ʿAlī aṣ-Ṣināʿiyah 1 [Jabal Ali Industria...	Jabel Ali Industrial	25.001900	55.126500
3	Warsān 1 [Warisan 1]	Warsan	25.162687	55.422592
4	Hūr al-ʿAnz	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

3. Methodology

3.1 Dubai Map Using Folium Library

Folium is a powerful Python library that helps create several types of Leaflet maps. The fact that the Folium results are interactive makes this library very useful for dashboard building.

Using the Nominatim module of Geopy library, I converted Dubai location into GPS coordinates returning the latitude and longitude values of Dubai. Then I used folium library and the GPS coordinates to create a map displaying all 18 communities of interest (using the dataframe that we created earlier) in Dubai.

[19]:
```
address = 'Dubai'

geolocator = Nominatim(user_agent="foursquare_agent")
location = geolocator.geocode(address)
dubai_latitude = location.latitude
dubai_longitude = location.longitude
print('The geographical coordinates of Dubai is are latitude : {} and longitude : {}'.format(dubai_latitude, dubai_longitude))
```

```
The geographical coordinates of Dubai is are latitude : 25.0750095 and longitude : 55.18876088183319
```

[20]:
```
map_dubai = folium.Map(location=[dubai_latitude, dubai_longitude],zoom_start=10)

for lat,lng, name, population in zip(df_final['Latitude'],df_final['Longitude'],df_final['Community Name'], df_final['Population (2018)']):
    label = '{}, {}'.format(name,population)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
    [lat,lng],
    radius=4,
    popup=label,
    color='blue',
    fill=True,
    fill_color='#3186cc',
    fill_opacity=0.7,
    parse_html=False).add_to(map_dubai)
map_dubai
```

See below the map of Dubai with communities marked.

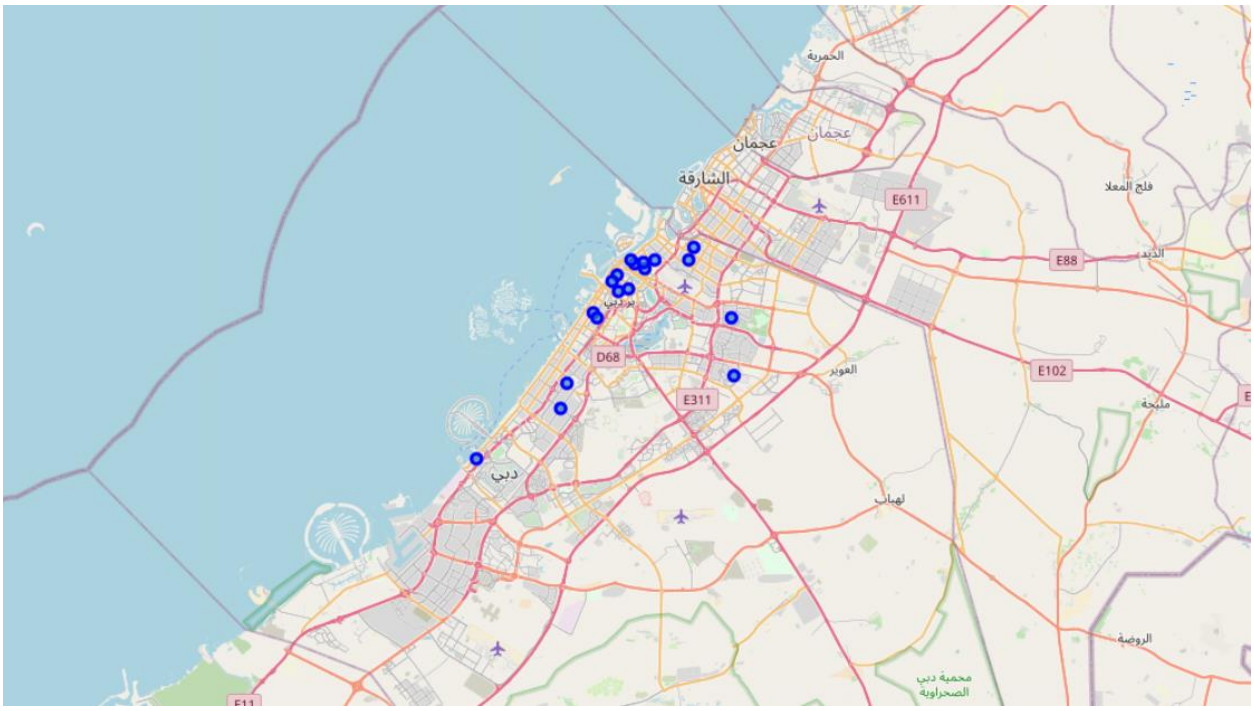

Figure 1: Map of Dubai superimposed with 18 communities represented as blue dots

3.2 Foursquare API

Foursquare data is very comprehensive, and it powers location data for companies like Apple, Uber etc. For this business problem I have used, as a part of the project, the Foursquare API to retrieve information about the venues around these 18 major communities of Dubai. The Foursquare API call returns a JSON file. Data will be extracted from JSON file to create a dataframe.

I used my Client ID and Client Secret key obtained from Foursquare when I had created my account.

[18]:
```
CLIENT_ID = 'CLIENT ID' # your Foursquare ID
CLIENT_SECRET = 'CLINET SECRET' # your Foursquare Secret
VERSION = '20180604'

print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

Next, I decided to retrieve 100 top venues for each of the 18 communities in Dubai within a radius of 2 kilometers around the latitude and longitude coordinates. The venues returned in the JSON file contained all sorts of venues, not only the restaurants. Total number of venues returned by Foursquare was 1,745 with 197 unique categories.

However, since we are interested in only the restaurants, so we will select only the venues where “Venue Category” is a restaurant. However, I found out that some of the “venue category” don’t

contain restaurant despite it being one. After exploration of data set, I decided to include Joint, Pizza, Burrito and Steakhouse as well while creating a dataframe for restaurants.

With this filter applied to venues returned by Foursquare, a total of 670 restaurants were returned. Below is the dataframe extracted from the JSON file and snippet of the code.

[56]:
```
# Create a dataframe out of it to concentrate only on restaurants (restaurants, joints shops, pizza, burrito and steakhouses)

dubai_venues_restaurant = dubai_venues[dubai_venues['Venue Category'].str.contains('Restaurant|Joint|Pizza|Burrito|Steakhouse')].reset_index(drop=True)

dubai_venues_restaurant.index = np.arange(1, len(dubai_venues_restaurant)+1)

print ("Shape of the dataframe with Venue Category only Restaurant: ", dubai_venues_restaurant.shape)

dubai_venues_restaurant.head()
```

```
Shape of the dataframe with Venue Category only Restaurant:  (670, 7)
```

[56]:

	Community Name	Community Latitude	Community Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	Al Quoz	25.13083	55.23273	Texas Roadhouse	25.130050	55.243130	Steakhouse
2	Al Quoz	25.13083	55.23273	Bertin Bistro	25.141190	55.217942	French Restaurant
3	Al Quoz	25.13083	55.23273	Jones The Grocer (جونز ذا جروسر)	25.139828	55.216732	Australian Restaurant
4	Al Quoz	25.13083	55.23273	Crumbs Elysee	25.139569	55.217273	French Restaurant
5	Al Quoz	25.13083	55.23273	Olive Garden	25.140091	55.217412	Italian Restaurant

Using Folium, Dubai map was superimposed by restaurant locations highlighted in different colors for each community.

The code that I used is as follows.

[60]:
```
## Show in Map the Top Rated Restaruants in the Top 18 Populous Communities in Dubai

map_restaurants = folium.Map(location=[dubai_latitude, dubai_longitude], zoom_start=11, tiles="openstreetmap",
                             attr="<a href=https://github.com/python-visualization/folium/>Folium</a>")

# set color scheme for the Venues based on the Major Communities

Communities = ['Al Quoz','Warsan','Hor Al Anz','Al Karama','Al Muraqqabat','Mirdif','Al Nahda 2','Dubai Marina','Al Badaa','Naif',\
               'Al Souq Al Kabeer','Al Muteena', 'Al Raffa', 'Al Qusais 1', 'Al Satwa', 'Al Quoz Third', 'Al Murar', 'Al Mankhool']

x = np.arange(len(Communities))

rainbow = ['#00ff00', '#ff00ff','#0000ff','#ffa500' ,'#ff0000',
           '#00ff00', '#ff00ff','#0000ff','#ffa500' ,'#ff0000',
           '#00ff00', '#ff00ff','#0000ff','#ffa500' ,'#ff0000',
           '#00ff00', '#ff00ff','#0000ff']

# add markers to the map
# markers_colors = []
for lat, lon, poi, comm in zip(dubai_venues_restaurant['Venue Latitude'],
                               dubai_venues_restaurant['Venue Longitude'],
                               dubai_venues_restaurant['Venue Category'],
                               dubai_venues_restaurant['Community Name']):
    label = folium.Popup(str(poi) + ' ' + str(comm), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=7,
        popup=label,
        color=rainbow[Communities.index(comm)-1],
        fill=True,
        fill_color=rainbow[Communities.index(comm)-1],
        fill_opacity=0.3).add_to(map_restaurants)
```

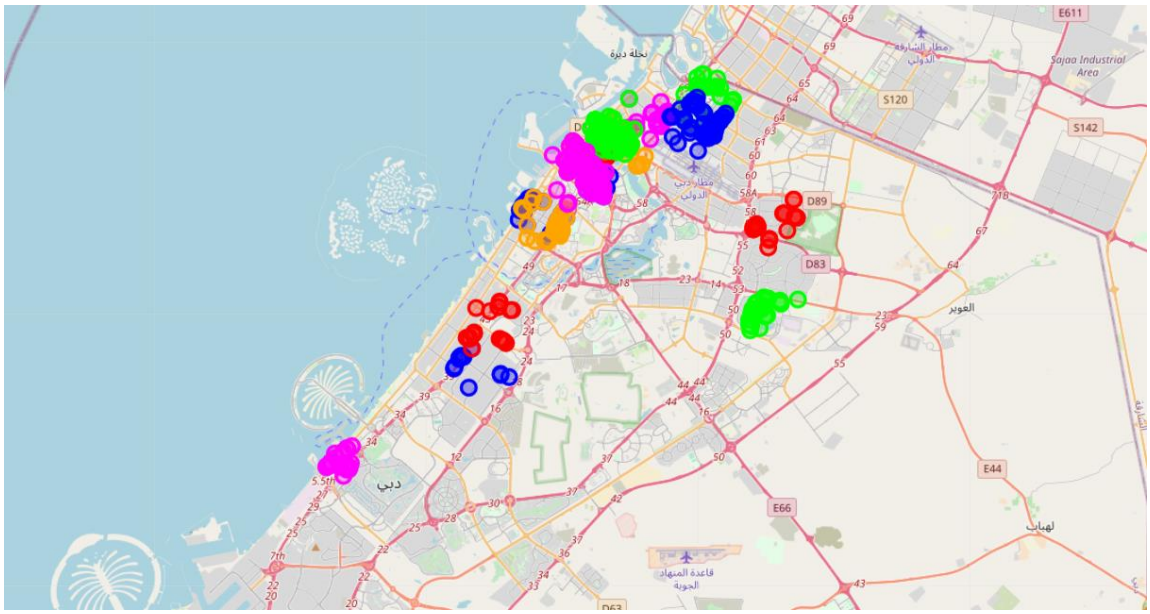

Figure 2: Map of Dubai displaying restaurants in 18 communities of interest

3.3 Exploratory Data Analysis (EDA)

Several restaurants in each community were returned, however we are interested in exploring the most common category of restaurants in a specific community. Hence, I created a new dataframe containing top 10 restaurants in each community.

Then using the seaborn and matplotlib libraries, I created a bar chart to show that with 93 restaurants, Indian cuisine stands at the top followed closely by Middle Eastern Cuisine. Whereas, fast food chains, Asian and Chinese restaurants are also popular in Dubai.

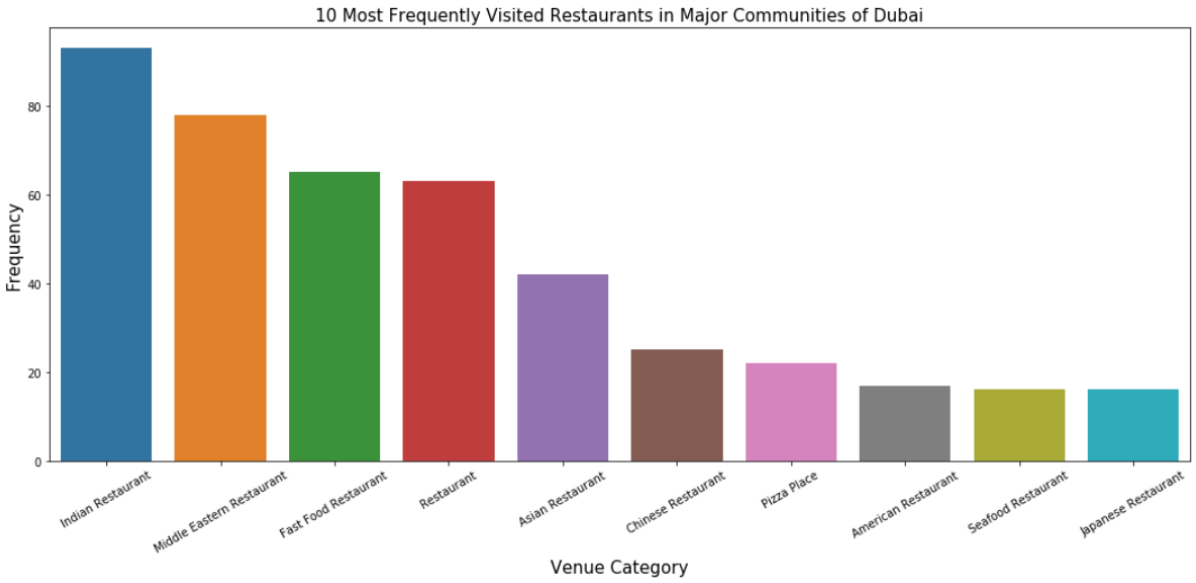

Figure 3: Most Common Restaurants in Dubai Communities

To provide more powerful visualization and give a comparison of different type of restaurants located in Dubai, I created a pie-chart as shown below.

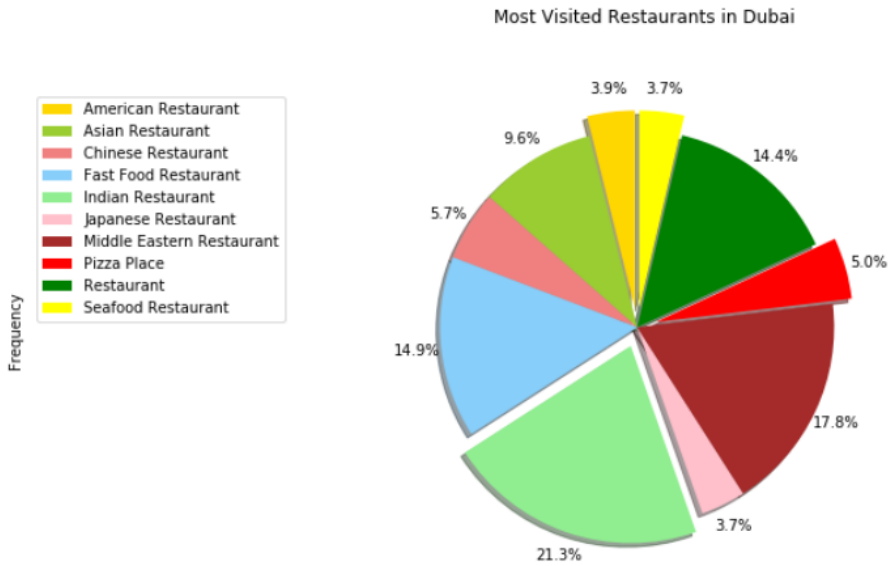

Figure 4: Pie-chart showing popular restaurants in Dubai

The results show that 21.3% of restaurants in Dubai are Indian Food restaurants while Middle Eastern restaurants constitute 17.8% of total restaurants in Dubai.

To continue with further exploratory data analysis, I created a new dataframe using Pandas onehot encoding of restaurants in Dubai for the "Venue Category" and then added Community Name back into the dataframe. The code snippet for onehot encoding is below.

```
[106]: # one hot encoding
dubai_onehot = pd.get_dummies(dubai_venues_restaurant[['Venue Category']], prefix="", prefix_sep="")

# add community column back to dataframe
dubai_onehot['Community Name'] = dubai_venues_restaurant['Community Name']

# move community name column to the first column

fixed_columns = [dubai_onehot.columns[-1]] + list(dubai_onehot.columns[:-1])
dubai_onehot = dubai_onehot[fixed_columns]

dubai_onehot
```

Next, I grouped rows by community names and by taking the mean of the frequency of occurrence of each category, followed by returning top 5 restaurants in each community. An example of top 5 restaurants that two of the communities "Al Badaa" and "Al Karama" returned are shown below along with the code.

[110]:
```
num_top_venues = 5

for hood in dubai_grouped['Community Name']:
    print("++++"+hood+"++++")
    temp = dubai_grouped[dubai_grouped['Community Name'] == hood].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')
```

```
++++Al Badaa++++
                      venue  freq
0          French Restaurant  0.13
1  Middle Eastern Restaurant  0.10
2                 Restaurant  0.10
3           Asian Restaurant  0.06
4               Burger Joint  0.06


++++Al Karama++++
                  venue  freq
0     Indian Restaurant  0.23
1            Restaurant  0.08
2    Seafood Restaurant  0.08
3  Fast Food Restaurant  0.06
4      Asian Restaurant  0.06
```

Later, I created a function to return the names of top 10 restaurants in each of the eighteen communities in Dubai that we are exploring. This would give us an idea of what the popular type of restaurant is in each of these communities and would form the basis of our investment decision.

[113]:

	Community Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Al Badaa	French Restaurant	Middle Eastern Restaurant	Restaurant	Greek Restaurant	Asian Restaurant	Burger Joint	Gluten-free Restaurant	Indian Restaurant	Italian Restaurant	Vegetarian / Vegan Restaurant
1	Al Karama	Indian Restaurant	Restaurant	Seafood Restaurant	North Indian Restaurant	Vegetarian / Vegan Restaurant	Asian Restaurant	Fast Food Restaurant	Korean Restaurant	Dim Sum Restaurant	Middle Eastern Restaurant
2	Al Mankhool	Indian Restaurant	North Indian Restaurant	Asian Restaurant	Seafood Restaurant	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Middle Eastern Restaurant	American Restaurant	Filipino Restaurant	Restaurant
3	Al Muraqqabat	Middle Eastern Restaurant	Restaurant	Indian Restaurant	Japanese Restaurant	Fast Food Restaurant	American Restaurant	Chinese Restaurant	Iraqi Restaurant	Asian Restaurant	Pizza Place
4	Al Murar	Middle Eastern Restaurant	Fast Food Restaurant	Restaurant	Japanese Restaurant	Asian Restaurant	Fried Chicken Joint	Indian Restaurant	Iraqi Restaurant	Pizza Place	African Restaurant

As we can see from the dataframe returned by our function that in Al Badaa community, French restaurants are mostly common whereas in Al Karama, most restaurant venues offer Indian cuisine. Similar inferences can be made for each of the eighteen communities in Dubai.

3.4 Clustering the Communities

Finally, I tried to cluster these 18 communities based on the restaurant categories and using K-Means clustering algorithm. Hence, our expectation would be that communities shall be clustered based on the similarities of the restaurant categories.

Snapshot of the code used to perform K-means clustering of these communities and a snapshot of the returned dataframe is as below.

[115]:
```
# set number of clusters
kclusters = 5

dubai_grouped_clustering = dubai_grouped.drop('Community Name', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(dubai_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

[115]:
```
array([2, 1, 1, 3, 3, 3, 1, 2, 4, 1], dtype=int32)
```

[118]:
```
# add clustering labels
community_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

dubai_merged = df_final

# merge dubai_grouped with df_final to add latitude/longitude for each community
dubai_merged = dubai_merged.join(community_venues_sorted.set_index('Community Name'), on='Community Name')

dubai_merged.head() # check the last columns!
```

[118]:

	Community Name	Population (2018)	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th M Comm Ve
1	Al Quoz	158543	25.130830	55.232730	2	Italian Restaurant	French Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Indian Restaurant	A Restau
3	Warsan	97159	25.162687	55.422592	0	Chinese Restaurant	Fast Food Restaurant	Middle Eastern Restaurant	Restaurant	Fried Chicken Joint	Restau
4	Hor Al Anz	81741	25.276680	55.335560	3	Middle Eastern Restaurant	Indian Restaurant	Fast Food Restaurant	Iraqi Restaurant	Restaurant	Moroc Restau
5	Al Karama	70558	25.248900	55.306100	1	Indian Restaurant	Restaurant	Seafood Restaurant	North Indian Restaurant	Vegetarian / Vegan Restaurant	A Restau
8	Al Muraqqabat	68717	25.268040	55.324920	3	Middle Eastern Restaurant	Restaurant	Indian Restaurant	Japanese Restaurant	Fast Food Restaurant	Amer Restau

Based on the clustering information, I created a map using Folium showing different communities belonging to their respective clusters.

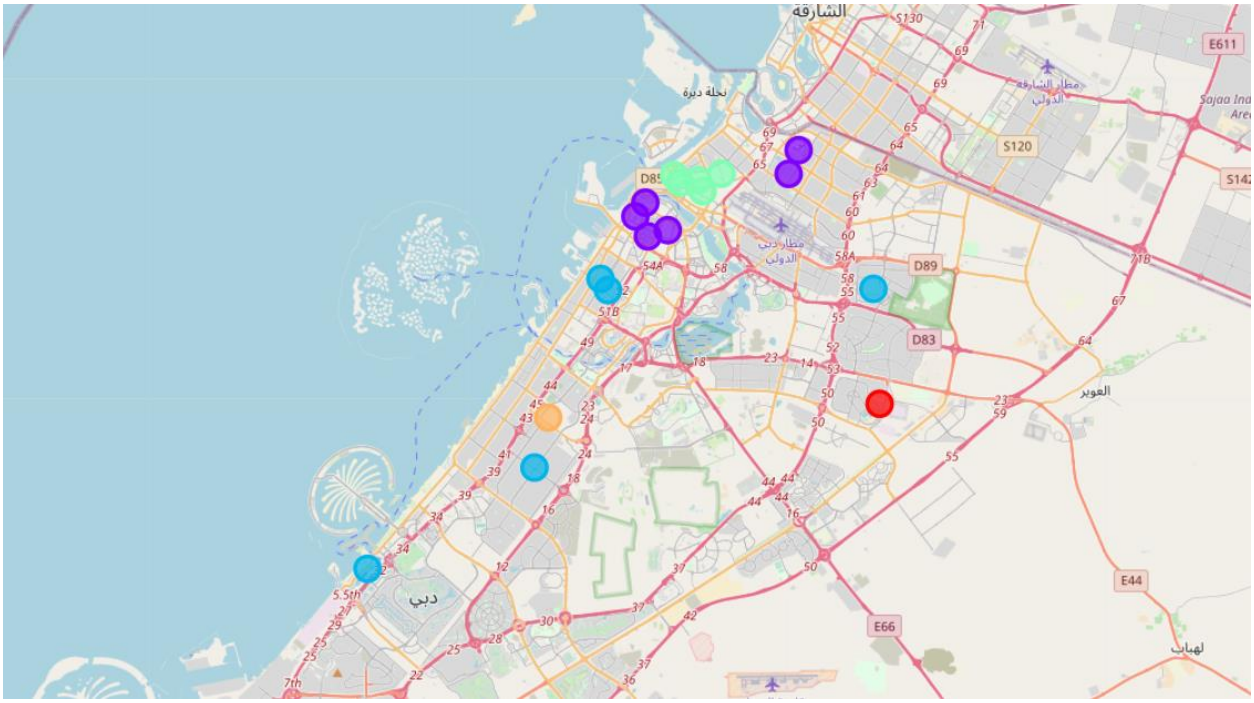

Figure 5: Five clusters of Dubai communities

Each of the 5 clusters of communities is represented by a separate color code. The communities are clustered based on similarities in restaurant category located in that specific community.

Each cluster was then individually examined to see the most common type of restaurant in each cluster. Ten most common venues in each cluster were returned.

Snippet of the code and the results for one out of five clusters is shown below.

[134]:
```
#cluster 1

dubai_cluster_1 = dubai_merged.loc[dubai_merged['Cluster Labels'] == 1, dubai_merged.columns[[1] + list(range(5, dubai_merged.shape[1]))]]

print("no of communities in cluster 1 is {}" .format(dubai_cluster_1.shape[0]))


dubai_cluster_1
```

```
no of communities in cluster 1 is 6
```

[134]:

	Population (2018)	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
5	70558	Indian Restaurant	Restaurant	Seafood Restaurant	North Indian Restaurant	Vegetarian / Vegan Restaurant	Asian Restaurant	Fast Food Restaurant	Korean Restaurant	Dim Sum Restaurant	Middle Eastern Restaurant
11	56489	Indian Restaurant	Fast Food Restaurant	Restaurant	Middle Eastern Restaurant	Italian Restaurant	Fried Chicken Joint	American Restaurant	Asian Restaurant	Chinese Restaurant	Pizza Place
15	46929	Indian Restaurant	Fast Food Restaurant	Middle Eastern Restaurant	Asian Restaurant	Chinese Restaurant	Persian Restaurant	Dim Sum Restaurant	Seafood Restaurant	Restaurant	Japanese Restaurant
17	42904	Indian Restaurant	Fast Food Restaurant	Pizza Place	Chinese Restaurant	Middle Eastern Restaurant	North Indian Restaurant	Seafood Restaurant	Indonesian Restaurant	French Restaurant	Filipino Restaurant
18	41818	Indian Restaurant	Fast Food Restaurant	Middle Eastern Restaurant	Restaurant	Fried Chicken Joint	Asian Restaurant	Pakistani Restaurant	South Indian Restaurant	Seafood Restaurant	BBQ Joint
23	37400	Indian Restaurant	North Indian Restaurant	Asian Restaurant	Seafood Restaurant	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Middle Eastern Restaurant	American Restaurant	Filipino Restaurant	Restaurant

4. Results

I have used data from web resources like Wikipedia and www.citypopulation.de, python libraries like Geopy, Seaborn and Matplotlib, Foursquare API, and finally using clustering to set up a very realistic data analysis scenario.

Results of the data analysis can be summarized as follows:

- Indian restaurants are largely popular in 18 most populous communities of Dubai. 21% restaurants in these communities are offering Indian food. In 6 out of 18 communities we explored, Indian restaurants are the most common.

- The genre of a potential new restaurant to be established varies between communities and is not constant across them. For example, in Al Quoz community, Italian restaurants are common whereas in Warsan community, most commonly cuisine is Chinese. Hence, the choice of the category would depend upon the locality where a potential new restaurant would be opened.

- The communities nearing each other are not similar in terms of restaurants choice. This is evident when we look at clustering plot. Communities geographically close to each other are dissimilar in choice of restaurant, hence falling in different clusters.

5. Conclusion

We have completed our analysis of "Battle of Communities" in Dubai; the virtual battle between communities competing for offering the best location for establishing a new restaurant business to the potential investors. The analysis initially started with exploring 24 most populated communities in Dubai, eventually concluded by analyzing 18 of them as 6 communities were later not considered for further analysis because venues returned by Foursquare for those communities were very few compared to other communities and we wanted to maintain our focus on communities of more interest to refine the analysis and get the results as accurate as possible.

As a final remark, I would like to mention that the accuracy of the analysis depends largely on the dataset we use to perform our analysis with. Hence, a data scientist should spend some time evaluating the authenticity of the data he or she is acquiring from various sources and intend to use in his project. Having lived in Dubai for almost five years now, I can confidently state that the results I got are largely correct as the typical occurrence of restaurants in those communities matches my analysis.

The analysis, however, does not and should not end here. In order to refine the results further and suggest a potential choice of restaurant category and community to make an investment decision, the commonality of a genre of restaurant in a specific community or cluster should not be the only

criteria. For example, our results would be more accurate if we consider the average income and hence buying/spending power of residents of a certain community in our analysis. Similarly, we could perform regression analysis to predict the possible revenues, number of customers and income. However, we would need much more data to perform such a statistical analysis which is beyond the scope of this report.