Finding best location to open a Japanese restaurant in Dubai using machine learning techniques and analysis of geo-spatial data.

by

Giridhar Reddy Kolan

Submitted for the requirements of

IBM Applied Data Science Capstone project - Dubai Neighborhood Analysis

Nov 2019

Table of Contents

1: INTRODUCTION

Dubai is the largest and most populous city in the United Arab Emirates, located on the Eastern coast of the Arabian Peninsula. As per Dubai statistics center estimates of 2018, there are approximately **3.2 million residents from over 200 nationalities** plus ever increasing tourists and traders on any given day.

Dubai is one of the fastest-growing cities in the world, increasing at a rate of 10.7% annually. The beauty and tolerance of the city make it a prime choice for expatriates, and the growing economy and availability of jobs makes it an appealing place to settle. Due to its warm hospitality, rich cultural heritage, best in class infrastructure, tax-free income and a strategic location in the center of the major trading continents Dubai has fast become one of the **world's most popular tourist destinations** and the best city to do business.

1.1 Business Problem

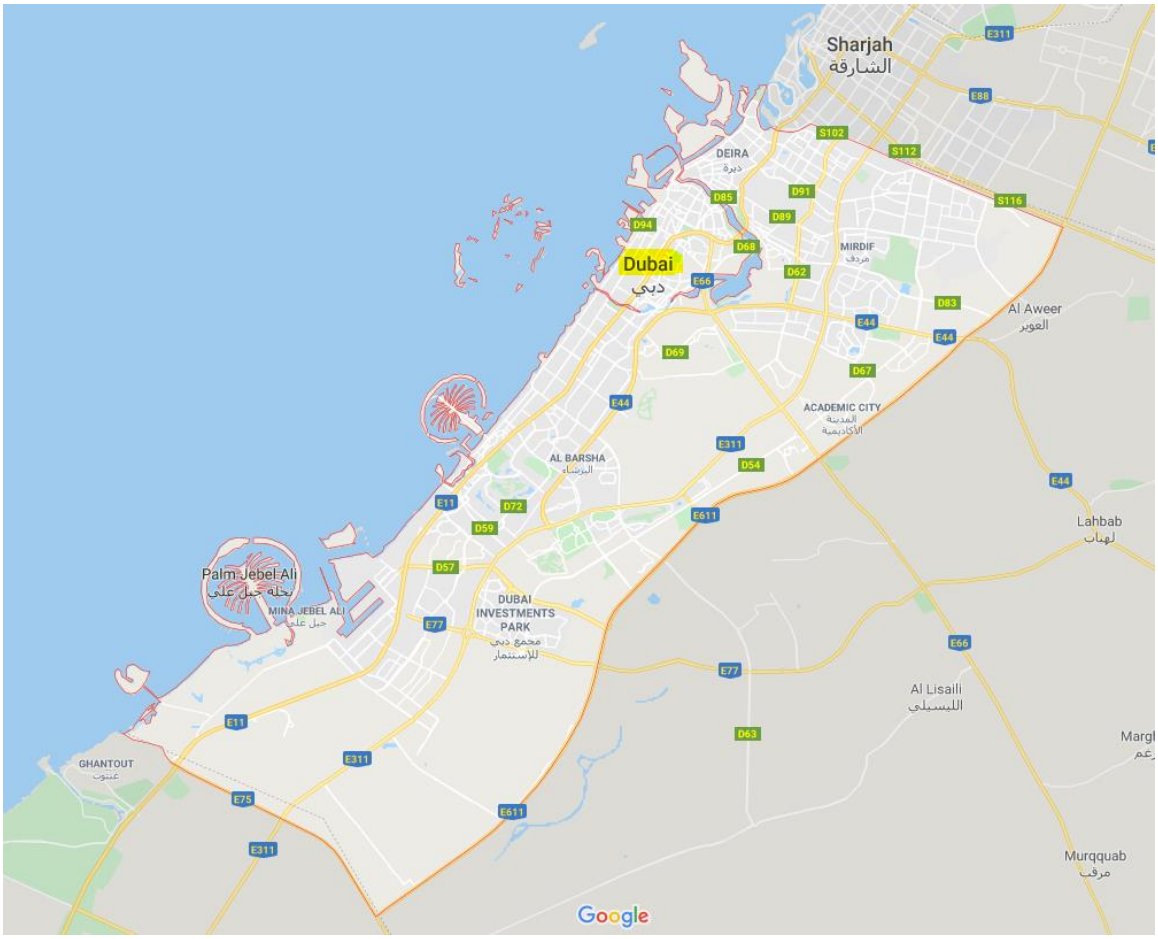
As the population of the Dubai city is growing rapidly with influx of working professionals and locals from just 1.3 million in 2005 to approximately 3.2 million in 2018 and inflow ever increasing tourists, which will be amplified with the approaching **World Expo 2020**, there is **best opportunity** for all prospective investors to venture into **Food and Beverage (F&B)** industry.

Upon quick checking of the Dubai Food and Beverage statistics data, it was observed that there is huge potential for **Japanese restaurants** serving authentic Japanese **Sushi Sashimi** and other healthy food options.

1.2 Objective

The main objective of this project to use machine learning techniques such as **K-Means** Clustering on geo-spatial / **location data from Foursquare API** and other sources to find out **best location for opening new Japanese restaurant**

1.3 Project Location



Dubai City and surroundings

.

2: DATA

2.1: Data sources

Based on our business problem requirements following parameters data sets were used:

1. **Tourists footfall:** What are top destinations of tourists in Dubai city with locations
2. **Population density:** Dubai neighborhoods and population counts / Densities
3. **Number of Venues** in the neighborhood to understand the commercial traffic
4. **Existing restaurants:** Availability of all types of restaurants in general and Japanese restaurants in particular around each neighborhood to measure the competition availability.

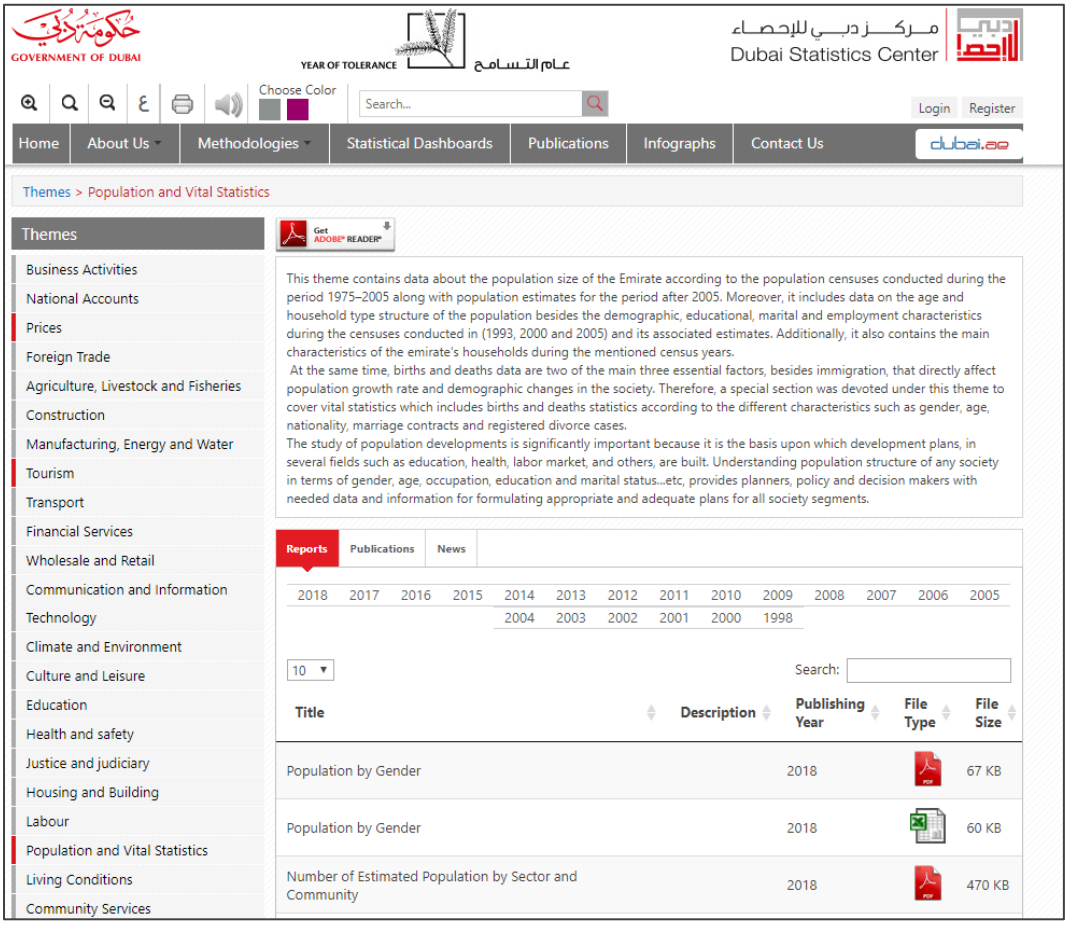
First two data sets were collected from the following internet sources:

- **Top Tourist Destinations in Dubai City:**

✓ https://www.globalmediainsight.com/blog/dubai-tourism-statistics/
✓ https://www.planetware.com/tourist-attractions-/dubai-uae-dub-dubai.htm

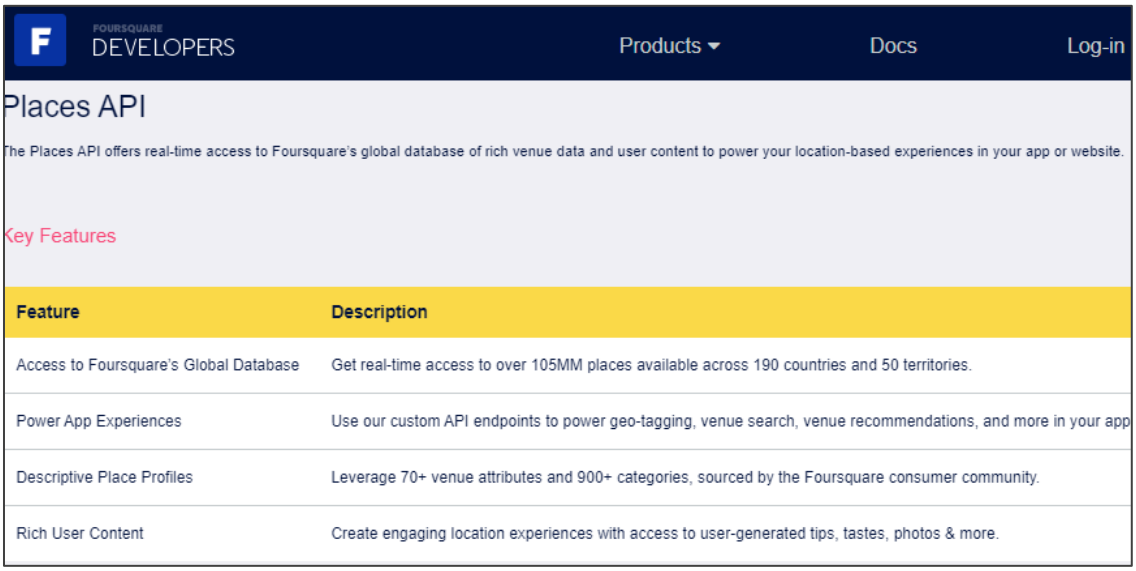
- **Number of Estimated Population in Dubai City:**

✓ https://www.dsc.gov.ae/en-us/Themes/Pages/Population-and-Vital-Statistics.aspx?Theme=42



This link contains data about the population size of the Dubai according to the population censuses conducted during the period 1975–2005 along with population estimates from 2006 to 2018. Moreover, it includes data on the age and household type structure of the population besides the demographic, educational, marital and employment characteristics during the censuses conducted in (1993, 2000 and 2005) and its associated estimates. Additionally, it also contains the main characteristics of the emirate's households during the mentioned census years.

Remaining data sets were collected / extracted and generated from internet source **https://foursquare.com/** using **FOURSQUARE API** credentials.



3: METHODOLOGY

Choosing a **good location** for **restaurant business** is the single most effective thing for its success. Having a good menu and professional staff is important to restaurant success, but having a good location will give your business another push toward success. While choosing a location on low commercial traffic areas might save you on rent, it won't allow your business much visibility. Setting up the restaurant in area with a lot of **commercial activities and foot traffic** puts your restaurant business out to a lot more people and offers you the opportunity to flourish.

Therefore, to achieve our objective of finding the best site for a **new Japanese restaurant** I have followed below **step by step methodology**.

1. Collect **top tourist destinations data set** from above web links / pages and create an .CSV file of these places. Update the tourist destinations file for missing values, such as **Latitude and Longitude values**, collected manually from **google earth.**
2. Collect and clean 2018 Population data of Dubai from the **Dubai Statistics Center** website as an excel sheet and transform into .CSV file.

3. Collect Dubai sectors and community **shape file** from Open sources and clean the geometries.
4. Join the 2018 community level population data with Dubai shape file and calculate **population density** per square kilometer per community based on the total population and geographic extent of the community.
5. **Explore and understand** the population distribution of Dubai neighborhoods.
6. Collect **venues** data from **FOURSQUARE API** for top tourist destination locations to understand commercial activity traffic around those locations.
7. **Spatially Compare** tourist destination locations with the population data and venues data to understand which locations are inside the high population density and are in highly commercial activity traffic areas.
8. **Eliminate** tourist destinations which are located in very low population density with very low commercial activity traffic (venue counts) from further analysis.
9. Collect **existing Japanese restaurants** counts around shortlisted tourist destinations from using FOURSQUARE API ***and remove*** tourist destinations from the list, which have **Japanese restaurants** in their neighbourhood.
10. Perform **k-means cluster analysis** on shortlisted locations and check the competition availability (existing restaurants) and choose at least two locations with less competition for further in-suite exploration to open new Japanese restaurant.

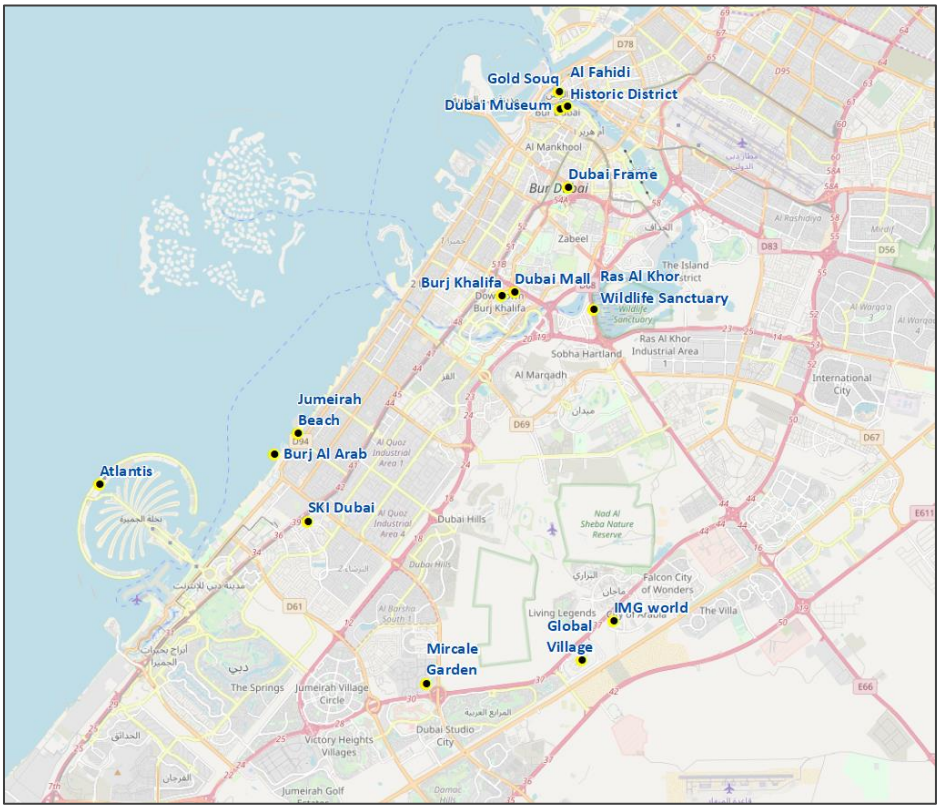
4: DATA EXPLORATION

As described in the Data section top tourist destinations data sets were collected, cleaned and transformed into .CSV format.

[2]:

	Sno	Destination	Longitude	Latitude
0	1	Burj Khalifa	55.274376	25.197242
1	2	Dubai Mall	55.279440	25.198540
2	3	Mircale Garden	55.244704	25.059860
3	4	Dubai Museum	55.297246	25.263557
4	5	Burj Al Arab	55.185343	25.141303
5	6	Global Village	55.305683	25.068157
6	7	Atlantis	55.116978	25.130439
7	8	SKI Dubai	55.198304	25.117329
8	9	IMG world	55.318116	25.082081
9	10	Al Fahidi Historic District	55.299905	25.264324

Up on overlaying these top tourist destinations on a map, it was discovered that these destinations are spatially well distributed covering most part of Dubai city.



Similarly, **Dubai Population 2018** data is collected from Dubai statistics Center, cleaned and transformed into .CSV format.

```
[4]: df_pop_2018 = pd.read_csv('Dubai_Population_2018.csv')
     df_pop_2018.head(10)
```

[4]:

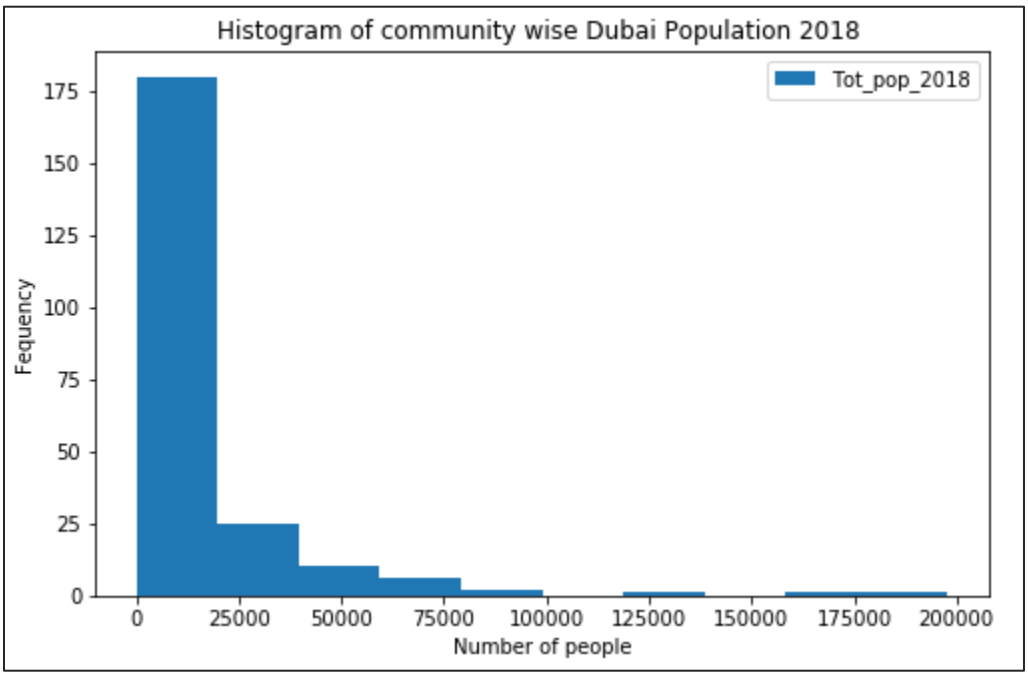
	Sno	Sector	CommunityName	Tot_pop_2018	Area_SqKM
0	1	Sector 1	AYAL NASIR	18925	0.2126
1	2	Sector 1	AL MURAR	38294	0.4762
2	3	Sector 1	AL DHAGAYA	15453	0.2176
3	4	Sector 1	NAIF	48804	0.9177
4	5	Sector 3	AL SUQ AL KABEER	46929	1.0744
5	6	Sector 1	AL SABKHA	3861	0.0898
6	7	Sector 1	HOR AL ANZ	81741	2.1501
7	8	Sector 1	AL MURQABAT	68717	1.8487
8	9	Sector 3	AL HAMRIYA	33421	1.0165
9	10	Sector 1	AL MUTEENA	43473	1.3670

Quick review of the population data gives the following summary table.

[8]:

	Sno	Tot_pop_2018	Area_SqKM
count	226.000000	226.000000	226.000000
mean	113.500000	14125.110619	24.112101
std	65.384759	24297.850367	40.936536
min	1.000000	0.000000	0.089800
25%	57.250000	419.250000	2.903325
50%	113.500000	5860.500000	7.592450
75%	169.750000	16723.500000	22.650825
max	226.000000	197838.000000	253.107300

There are in total 226 communities are there in Dubai. The population of communities ranges from 0 to 197,838. The frequency distribution of this population data can be visualized by following histogram plot.



The Community with highest Total population is: **MUHAISANAH SECOND**

```
[6]: df_pop_2018.loc[df_pop_2018['Tot_pop_2018'].idxmax()]

[6]: Sno                            11
     Sector                     Sector 2
     CommunityName    MUHAISANAH SECOND
     Tot_pop_2018                197838
     Area_SqKM                    6.849
     Name: 10, dtype: object
```

Community with highest area is : **AL FAGAA'**

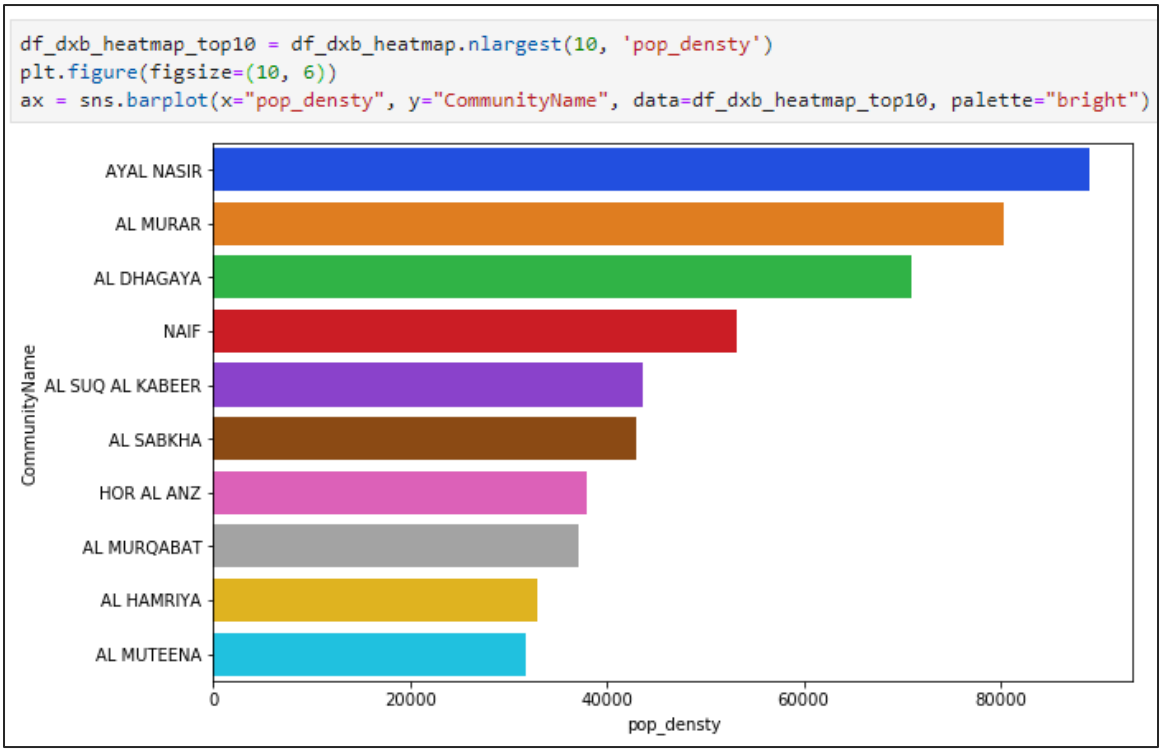
```
[7]: df_pop_2018.loc[df_pop_2018['Area_SqKM'].idxmax()]

[7]: Sno                  191
     Sector             Sector 9
     CommunityName     AL FAGAA'
     Tot_pop_2018            395
     Area_SqKM           253.107
     Name: 190, dtype: object
```

It was observed that the community with highest area is not the community with highest population. Population distribution of the city can be better understood by the population density of each community than the total population. Population density of each community is calculated using total population and area of community as shown below.

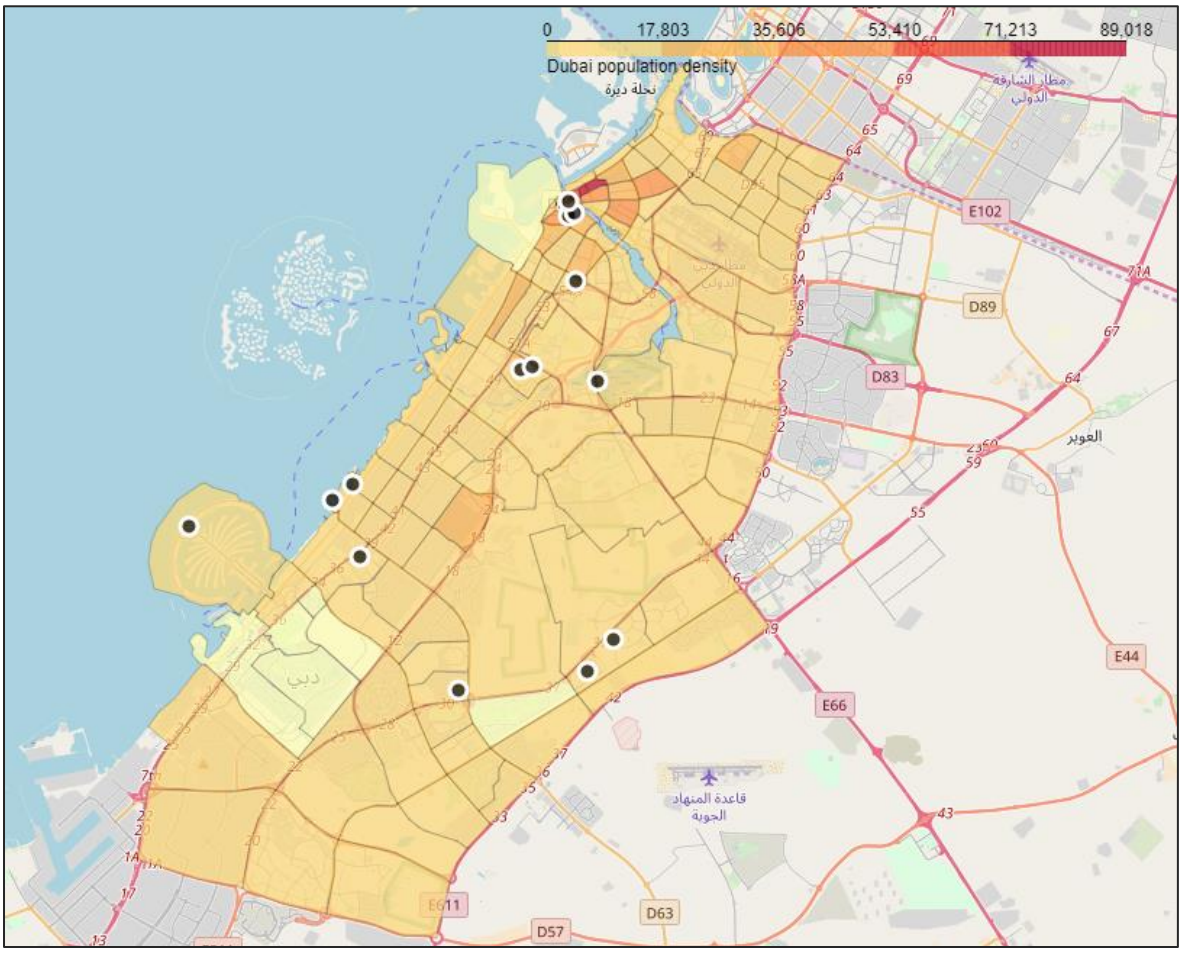
```
df_pop_2018['pop_densty'] = df_pop_2018['Tot_pop_2018']/ df_pop_2018['Area_SqKM']
df_pop_2018.pop_densty = df_pop_2018.pop_densty.round(decimals=0)
df_pop_2018.head(10)
```

Following bar graph shows the top ten population density communities of Dubai.

```
df_dxb_heatmap_top10 = df_dxb_heatmap.nlargest(10, 'pop_densty')
plt.figure(figsize=(10, 6))
ax = sns.barplot(x="pop_densty", y="CommunityName", data=df_dxb_heatmap_top10, palette="bright")
```



Same population density data was mapped using **Folium map** and Dubai GeoJSON file, centered around Dubai latitude and longitude values (55.274376, 25.197242), with an initial zoom level of 11 and superimposed by locations of top tourist destinations represented by black to white circles.



With careful observation of the above map we can infer that the **some of the top tourist destinations** are in **low population density** areas.

5: GEO-SPATIAL ANALYSIS AND MACHINE LEARNING

To understand venues counts (**commercial activity**) around 1 kilometer from each of the location I have used **Foursquare API** and extracted the data as Pandas data frame. In total 985 venues were extracted with 7 columns and 171 unique **venue categories** were curated.

```
print(Dubai_venues.shape)
Dubai_venues.head()
```

```
(985, 7)
```

After trying to understand commercial activity (venue count) around each tourist destination, it was identified that there **are some destinations** with **low commercial activities (Venue counts)** as illustrated in the following table.

[29]:

	Destination	Venue_count
11	Mircale Garden	5
12	Ras Al Khor Wildlife Sanctuary	6
9	IMG world	33
7	Global Village	36
1	Atlantis	55
10	Jumeirah Beach	75
2	Burj Al Arab	82
8	Gold Souq	93
0	Al Fahidi Historic District	100
3	Burj Khalifa	100
4	Dubai Frame	100
5	Dubai Mall	100
6	Dubai Museum	100
13	SKI Dubai	100

Also, I have tried to understand exact population density value of each tourist destination. To get this information, I have used **Geo-pandas spatial join** capabilities. First I have converted tourist destinations data and Dubai City geo-json file into **Geo-pandas** data frames.

```
# convert tourists destinations to Geopandas Geodataframe
gdf_top_destins = gpd.GeoDataFrame(df_top_destins,   geometry=gpd.points_from_xy(df_top_desti
gdf_top_destins.crs = {'init' :'epsg:4326'}  # assignes geographic coordinate system - wgs84
gdf_top_destins.crs
```

```
{'init': 'epsg:4326'}
```

```
# get the Dubai City geojson file into geodataframe
#gdf_dxb_city = "Dubai_City_GCS.json"
gdf_dxb_city = gpd.read_file('Dubai_City_GCS.json')
gdf_dxb_city.crs
```

And then used the **spatial join** to determine which tourist point spatially locates in which community area and then obtained the corresponding population density for each point.

```
sjoin_tourist_pts = gpd.sjoin(gdf_top_destins, gdf_dxb_city, how="inner", op='intersects')
```

```
sjoin_tourist_pts.head()
```

	OBJECTID	Sno	Destination	Longitude	Latitude	COMM_NUM	Community_Code	Tot_pop_2018	Sector	CommunityName	Area_SqKM	Pop_Density_2018
0	1	1	Burj Khalifa	55.274376	25.197242	345	345	18698	Sector 3	BURJ KHALIFA	3.262775	5730.704059
1	2	2	Dubai Mall	55.279440	25.198540	345	345	18698	Sector 3	BURJ KHALIFA	3.262775	5730.704059
2	3	3	Mircale Garden	55.244704	25.059860	673	673	4566	Sector 6	AL BARSHA SOUTH THIRD	4.581933	996.522624
3	4	4	Dubai Museum	55.297246	25.263557	312	312	46929	Sector 3	AL SUQ AL KABEER	1.074399	43679.287034
4	5	5	Burj Al Arab	55.185343	25.141303	366	366	7021	Sector 3	UM SUQAIM THIRD	3.112740	2255.568922

Once I have tourist destinations locations with exact population densities, I have merged already extracted venues data with this data frame to get the venue counts and population density at each tourist location.

```
[35]: df_merged = pd.merge(sjoin_tourist_pts, dxb_venues1, on='Destination')
      df_merged.head()
```

Sno	Destination	Longitude	Latitude	CommunityName	Area_SqKM	Pop_Density_2018	Venue_count
1	Burj Khalifa	55.274376	25.197242	BURJ KHALIFA	3.262775	5730.704059	100
2	Dubai Mall	55.279440	25.198540	BURJ KHALIFA	3.262775	5730.704059	100
3	Mircale Garden	55.244704	25.059860	AL BARSHA SOUTH THIRD	4.581933	996.522624	5
4	Dubai Museum	55.297246	25.263557	AL SUQ AL KABEER	1.074399	43679.287034	100
5	Burj Al Arab	55.185343	25.141303	UM SUQAIM THIRD	3.112740	2255.568922	82

Once I have the both values, as explained methodology sections I have eliminated tourist destinations which are located in low population density or with low commercial activity traffic (venue counts) from further analysis.

For this study, based on stakeholders and Subject Matter Expert's (SME) opinions, I have eliminated destinations with **less than 700 people** per Square Kilometer or **Venues count is less than 70** within one Kilometer range.

```
df_shortlist1 = df_merged.loc[(df_merged.Pop_Density_2018 > 700 ) & (df_merged.Venue_count > 70)]
print(df_shortlist1.shape)
df_shortlist1.head()
```

```
(8, 13)
```

Now, I have ended up with following **8 shortlisted** tourist destinations for further analysis.

```
[53]: df_shortlist1[['Destination','CommunityName','Pop_Density_2018','Venue_count']]
```

[53]:

	Destination	CommunityName	Pop_Density_2018	Venue_count
0	Burj Khalifa	BURJ KHALIFA	5730.704059	100
1	Dubai Mall	BURJ KHALIFA	5730.704059	100
3	Dubai Museum	AL SUQ AL KABEER	43679.287034	100
4	Burj Al Arab	UM SUQAIM THIRD	2255.568922	82
7	SKI Dubai	AL BARSHAA FIRST	7322.909633	100
9	Al Fahidi Historic District	AL SUQ AL KABEER	43679.287034	100
10	Jumeirah Beach	UM SUQAIM SECOND	3353.225474	75
11	Gold Souq	AL RASS	19781.519420	89

As per the established methodology, to understand existing competition around these locations, I have collected count of existing all types of restaurants in general and Japanese restaurants in particular.

```
df_restaurant = Dubai_venues[Dubai_venues['Venue Category'].str.contains('Restaurant')]
df_jap_restrnt = Dubai_venues[Dubai_venues['Venue Category'].str.contains('Japanese Restaurant')]
```

```
print(Dubai_venues.shape)
print(df_restaurant.shape)
print(df_jap_restrnt.shape)
```

```
(985, 7)
(292, 7)
(6, 7)
```

In total there were 292 restaurants and 6 Japanese restaurants.

[43]:

	Destination	japan_restaurant_count
0	Atlantis	1
1	Burj Al Arab	1
2	Burj Khalifa	3
3	Dubai Frame	1

After merging these two datasets with the shortlisted destinations data frame. I have eliminated tourist destinations which have Japanese restaurants in their neighborhoods.

```
df_shortlist2 = pd.merge(df_shortlist1, df_restaurant_grpby, how ='left', on='Destination')
```

```
df_shortlist3 = pd.merge(df_shortlist2, df_jap_restrnt_grpby, how ='left', on='Destination')
```

```
df_shortlist4 = df_shortlist3[df_shortlist3.japan_restaurant_count < 1]
print(df_shortlist4.shape)
df_shortlist4[['Destination','CommunityName','Pop_Density_2018','Venue_count', 'restaurant_count', 'japa
```

```
(6, 15)
```

	Destination	CommunityName	Pop_Density_2018	Venue_count	restaurant_count	japan_restaurant_count
1	Dubai Mall	BURJ KHALIFA	5730.704059	100	11	0.0
2	Dubai Museum	AL SUQ AL KABEER	43679.287034	100	35	0.0
4	SKI Dubai	AL BARSHAA FIRST	7322.909633	100	21	0.0
5	Al Fahidi Historic District	AL SUQ AL KABEER	43679.287034	100	42	0.0
6	Jumeirah Beach	UM SUQAIM SECOND	3353.225474	75	21	0.0
7	Gold Souq	AL RASS	19781.519420	89	28	0.0

Now we have **6 shortlisted tourist destinations** within high population density, high commercial activities and with no Japanese restaurants nearby.

To this dataset, I have applied K-means clustering algorithm, which is unsupervised machine learning technique that clusters given dataset into defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.

I have clustered the shortlisted tourist destinations dataset set into 3 distinct subgroups.

```
Dxb_center = [25.197242, 55.274376]
poi_latlons = [[lat, lng] for lat, lng in zip(df_top_destins['Latitude'], df_top_destins['Longitude'])]

from sklearn.cluster import KMeans
number_of_clusters = 3

good_xys = [[lat, lng] for lat, lng in zip(df_shortlist4['Latitude'], df_shortlist4['Longitude'])]
kmeans = KMeans(n_clusters=number_of_clusters, random_state=0).fit(good_xys)
cluster_centers = kmeans.cluster_centers_

dxb_map1 = folium.Map(location=Dxb_center, zoom_start=11)
folium.TileLayer('cartodbpositron').add_to(dxb_map1)
HeatMap(restaurant_latlons, radius = 22).add_to(dxb_map1)
```

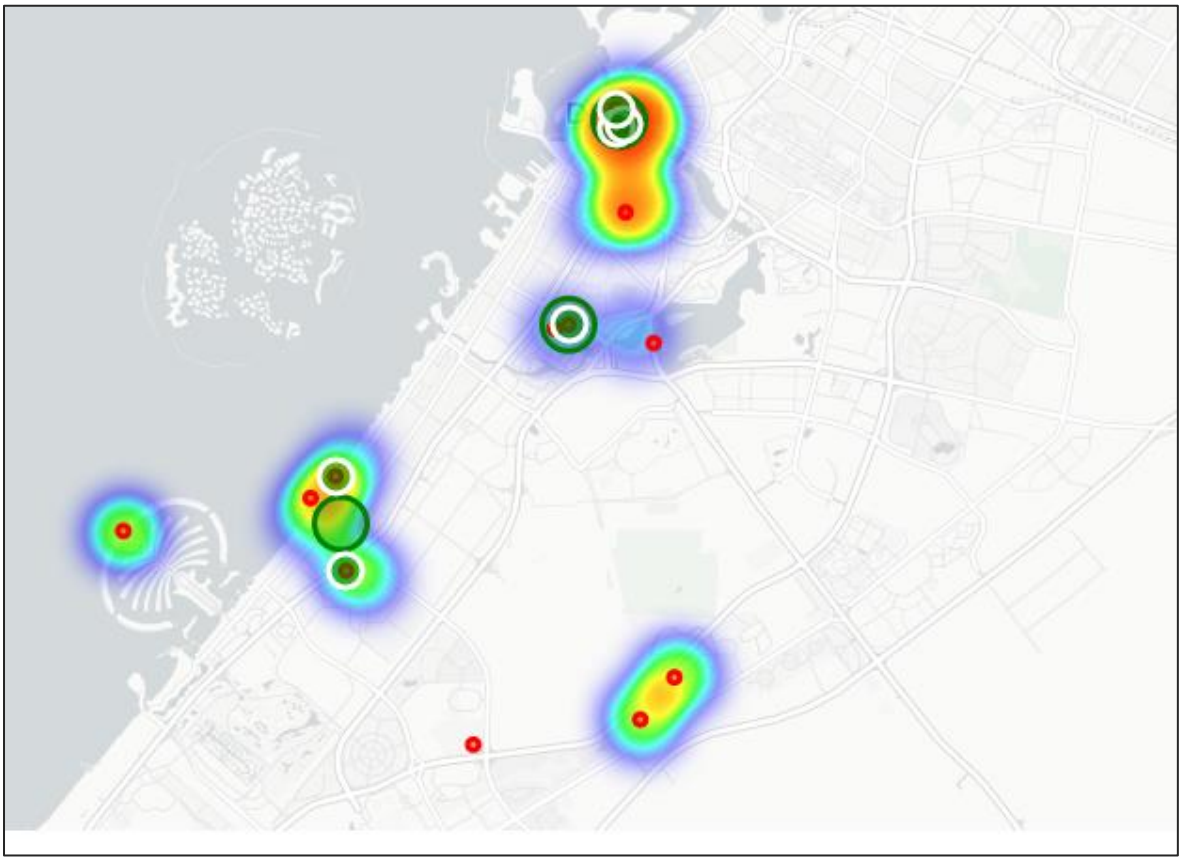
6: RESULTS AND CONCLUSIONS

6.1: Results

Based on the above analysis we ended up with 4 different resultants datasets.

1. Shortlisted tourist destinations
2. The restaurants densities around each tourist destination
3. K-means clusters with their centroids
4. Eliminated tourist destinations

All these datasets were mapped using the folium map application along with open street map as a background map.



Map showing the heat map of restaurants, shortlisted tourist destinations (white rings) along with their cluster centers (green rings) and eliminated tourist destinations (red rings).

When we examine the resultant map it was observed that, clusters near **Dubai mall** and **SKI Dubai** have **less competition** (existing restaurants density) compared with cluster near **Dubai museum**.

The same findings were confirmed by calculating the ratio of restaurants to total number of venues around these shortlisted locations.

```
df_shortlist4['restaurantratio'] = df_shortlist4.restaurant_count / df_shortlist4.Venue_count
df_shortlist4.head()
```

```
final_locations = df_shortlist4.sort_values('restaurantratio')
final_locations[['Destination', 'restaurantratio']]
```

	Destination	restaurantratio
1	Dubai Mall	0.110000
4	SKI Dubai	0.210000
6	Jumeirah Beach	0.280000
7	Gold Souq	0.322581
2	Dubai Museum	0.350000
5	Al Fahidi Historic District	0.420000

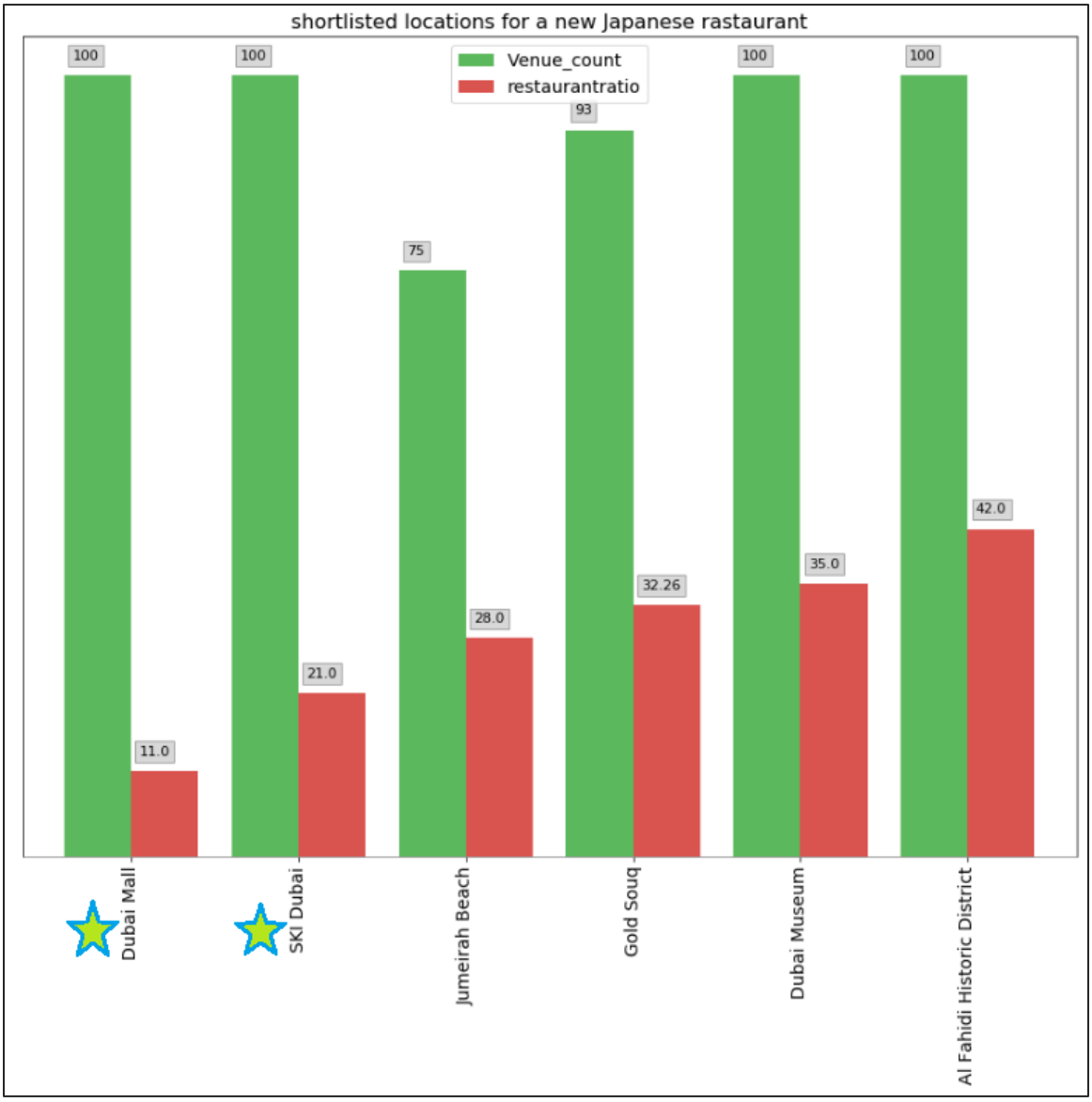
The final results data frame is re-arranged as below for plotting purpose.

```
final_locations_grph = final_locations[['Destination', 'Venue_count', 'restaurantratio' ]]
final_locations_grph.set_index('Destination', inplace=True)
final_locations_grph.restaurantratio= final_locations_grph.restaurantratio*100
final_locations_grph.restaurantratio = final_locations_grph.restaurantratio.round(decimals=2)
final_locations_grph
```

Destination	Venue_count	restaurantratio
Dubai Mall	100	11.00
SKI Dubai	100	21.00
Jumeirah Beach	75	28.00
Gold Souq	93	32.26
Dubai Museum	100	35.00
Al Fahidi Historic District	100	42.00

We can clearly see that **Dubai Mall** and **SKI Dubai** less competition of existing restaurants compared to other shortlisted tourist destinations.

Below bar graph illustrates the final results of this analysis for selecting the best locations to open a Japanese restaurant in Dubai



6.2: Conclusions

This concludes our analysis study of finding best location to open a Japanese restaurant in Dubai using machine learning techniques and analysis of geo-spatial data. By using a combination of different datasets of Dubai city from DSC, Foursquare API and other sources we were able to analyze, discover and describe tourist destinations neighborhoods with their population density statistically describe commercial activities (venue counts) and understand the existing competition by measuring the existing restaurants availability.

Through this study we have identified 2 prospective locations, **1. Dubai Mall** and **2. SKI Dubai** for opening new Japanese restaurant. These areas have **low number of restaurants** and **no Japanese restaurants nearby**, and are located within the **high population density** areas. These locations are very popular with tourists, fairly close to city center and well connected by public transport.

These final locations need be considered only as a starting point for exploring area neighborhoods in search for exact restaurant sites / vacant places / plots based on the investor investment risk profile.

Finally, we are able to use learned data science knowledge, Database models, Visualization techniques, python tools and IBM applications to successfully complete the project.