Capstone Project

Dubai's Apartments Sale Price and Venues Analysis

I. Introduction

I.a. Background and Description

Hosting the tallest tower in the world and more than 70 skyscrapers higher than 150 meters (1). With a population of more than 3.3 Million as of April 2020, Dubai's population density in 2018 was 716 persons per sq.km and a rate of 83.2 on Refined Economic Activity. (2)

Being a very attractive destination, choosing the **right apartment to purchase or invest** in **Dubai** could be achieved more advanced methods than the conventional approach. Depending on one's interests and preferences, different areas could have more or less value to the potential owner.

The goal of this project is to find the **optimal** location where a potential buyer of an apartment in Dubai is going to consider as a home. Compared to villas, apartments provide more convenience for a small family or a single person living alone. Criteria of convenience could differ depending on one's priorities. Some factors determining choice could be nearby venues like restaurants, malls, supermarkets and metro stations.

To overcome this challenge, **Data Science** methods and techniques like Machine Learning could help limit the best options and take in consideration all possible factors that will eventually contribute to one's choice and preference. In addition, these methods provide useful tools for Academia as well.

I.b. Data Acquisition

As we have defined the issue in the previous section, we can now list the data sources needed to complete the project:

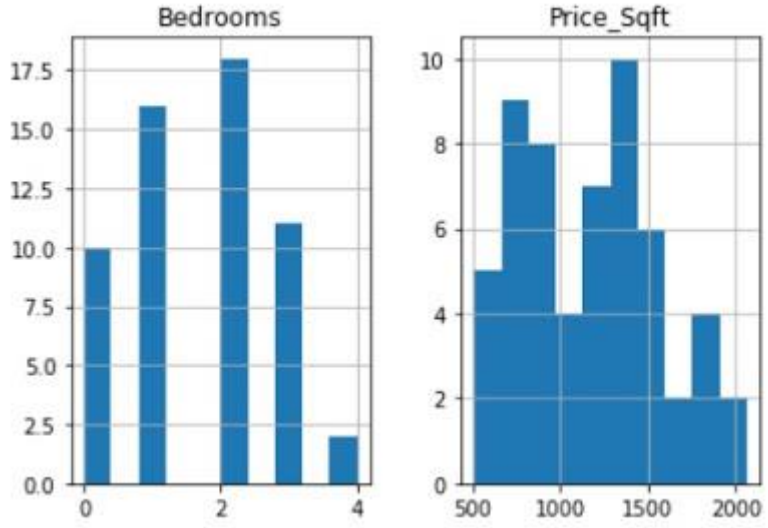
- **Bayut.com** (3)**:** One of the most popular website/apps which hosts the prices of properties including apartments in Dubai. It also gathers trends and indices which we can use to compare apartment sale prices.
- **Google Maps:** to scrape the latitudes and longitudes of locations
- **Foursquare API** (4): With the API, we can collect all necessary location data of nearby venues such as restaurants, pharmacies, mall..etc.
- **NYU Spacial Data Repository** (5): Large data which spacial information. Limited the use to finding Metro stations locations.

I.c. Data Filtration

Collected data needed cleaning and exploration to decided on the best set to carry on with the study. A of locations were specified to explore the types of apartments offered for sale in each location. Values were given to Bedrooms from **zero** for studios and maximum **4** bedrooms with the price per sqft. As the below description table shows there was a total of **57** offers with a mean of 1.6 for Bedrooms and a price of 1,178.16 per square feet.

	Bedrooms	Price_Sqft
count	57.000000	57.000000
mean	1.631579	1178.157895

The following histogram chart summarizes the data:



For the purpose of this project, we will only consider the prices of **One-Bedroom** type and will examine them in the **23** specified locations. As mentioned earlier, the target is small family or a single person living alone. In addition, following a more conservative approach, the mean of Bedrooms is 1.6 thus choosing the lower closest integer.

II. Methodology

II.a Locations Data and Sale Price

Google Maps was used for details of the locations and Bayut to find the average sale price per square feet for a one-bedroom apartment in specific locations in Dubai where one Bedrooms are available. Below a screenshot of the table's head:

	Location	Price_Sqft	Longitude	Latitude
0	Al Furjan	936	55.1459	25.0252
1	Al Quoz	876	55.2508	25.1514
2	Business Bay	1287	55.2729	25.1832
3	Cultural Village	1130	55.3379	25.2251
4	DIFC	1398	55.2770	25.2088

Folium in Python and Leaflet enables us to visualize the data and convert the above latitudes and longitudes to markers on a map as shown below. This is the code and a screenshot of the result interactive map.

```
# create map
map_df_data_0 = folium.Map(location=[25.2048, 55.2708], width=800, height=400, zoom_start=11)

# plot locations
for (index, row) in df_data_0.iterrows():
    folium.Marker(location=[row.loc['Latitude'], row.loc['Longitude']],
                  popup=row.loc['Location'],
                  tooltip='click').add_to(map_df_data_0)

# display map
map_df_data_0
```

II.b.Foursquare API

One of the largest technology company with most accurate locations' data. Foursquare also empowers the location data for Apple maps, Uber, Snapchat and many others. By implementing crowd sourcing technique, they enabled people to build their data while using the app.

For this project, the API was utilized to collect nearby locations of the specified locations. This service provides the names and categories of the venues. The limit was set to 100 venues in **600 meters** Radius. This is a head table of the returned data:

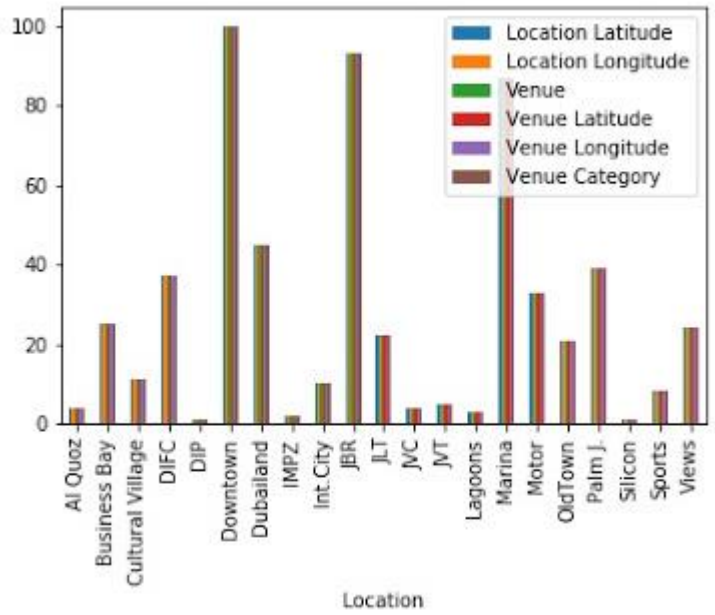
	name	categories	cc
0	The club masakin pool	Pool	AE
1	Al Furjan Villas	Housing Development	AE
2	Al Furjan Club	Gym / Fitness Center	AE
3	Al Furjan Pavilion	Shopping Mall	AE
4	Carrefour Market	Supermarket	AE

II.c. Merge Datasets

Now that we have gathered all nearby venues from the specified locations, the two tables could be merged to assess which Location had the most returned values.

	Location	Location Latitude	Location Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Al Quoz	25.1514	55.2508	Emirates Post - Al Quoz Fourth Post Office	25.150564	55.248757	Post Office
1	Al Quoz	25.1514	55.2508	chai wala cafe	25.151540	55.250066	Cafeteria
2	Al Quoz	25.1514	55.2508	Al Khail Gym	25.148910	55.249430	Gym
3	Al Quoz	25.1514	55.2508	Golden Myanmar	25.148683	55.246970	Asian Restaurant
4	Business Bay	25.1832	55.2729	Gulf Court Hotel Business Bay	25.182244	55.274908	Hotel

The bar chart shows that **Downtown Dubai** reached the **100** limit of venues which we set in the code. **JBR** and **Marina** returned with **93** and **87** venues respectively.

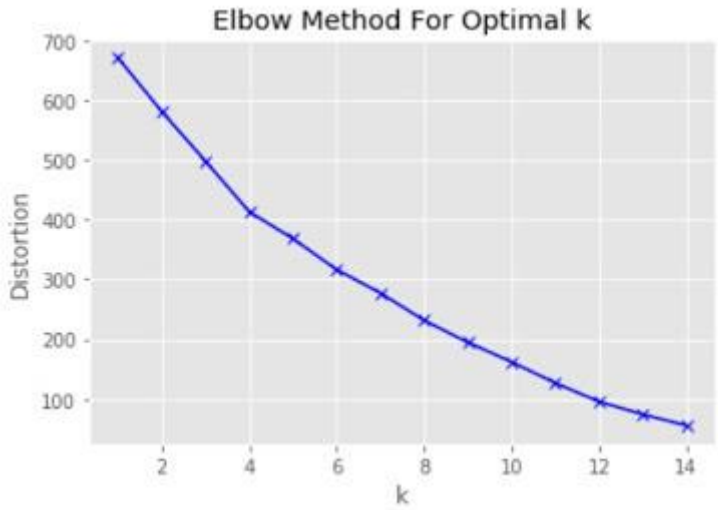


This analysis shows that there are **148** unique categories based on Foursquare data. The following table summarizes the top 6 most common venues per location.

	Location	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Al Quoz	Asian Restaurant	Post Office	Cafeteria	Gym	Yoga Studio	Frozen Yogurt Shop
1	Business Bay	Hotel	Middle Eastern Restaurant	Italian Restaurant	Coffee Shop	Lounge	Japanese Restaurant
2	Cultural Village	Hotel Pool	Resort	Spa	Molecular Gastronomy Restaurant	Metro Station	Gym / Fitness Center
3	DIFC	Middle Eastern Restaurant	Coffee Shop	Hotel	Italian Restaurant	Gym / Fitness Center	Steakhouse
4	DIP	Rock Climbing Spot	Shipping Store	Yoga Studio	French Restaurant	Cupcake Shop	Diner

III.d. K-Mean Clustering

One of the most used methods of unsupervised Machine Learning, which means the system self-learns based on features, is K-Mean Clustering. However, to identify the most fitting number of cluster k could be challenging. To estimate K, an approach known as Elbow Method could be used.



Based on the above plot, estimated that K=10

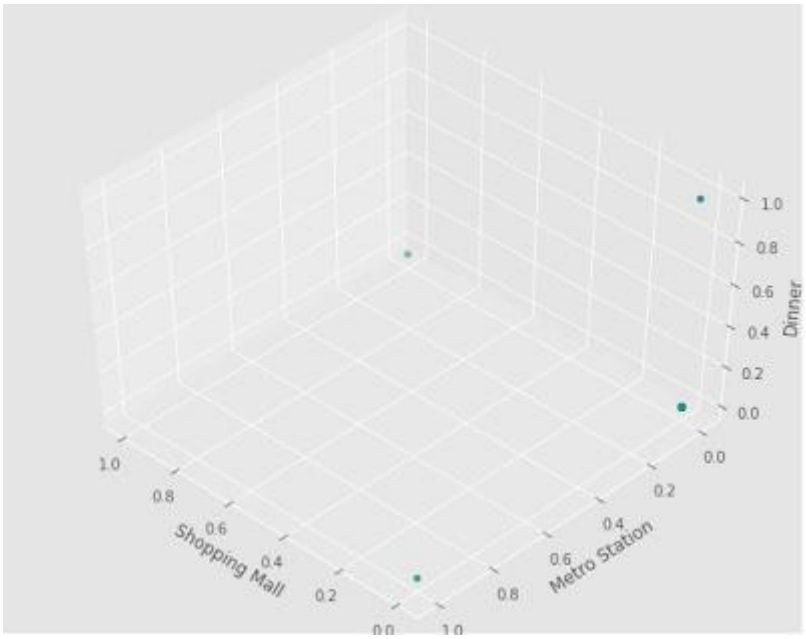
IV. Results

Once K has been estimated, the data is then updated with the clusters as explained in the table below.

	Location	Price_Sqft	Longitude	Latitude	Cluster_Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
1	Al Quoz	876	55.2508	25.1514	5.0	Post Office	Cafeteria	Asian Restaurant	Gym	Yoga Studio	Falafel Restaurant
2	Business Bay	1287	55.2729	25.1832	1.0	Hotel	Italian Restaurant	Coffee Shop	Lounge	Middle Eastern Restaurant	Japanese Restaurant
3	Cultural Village	1130	55.3379	25.2251	1.0	Hotel Pool	Harbor / Marina	Molecular Gastronomy Restaurant	Metro Station	Bar	Gym / Fitness Center
4	DIFC	1398	55.2770	25.2088	1.0	Coffee Shop	Middle Eastern Restaurant	Hotel	Gym / Fitness Center	Italian Restaurant	Café

A map of clusters shows the results

Another way to look at the results in through 3D visualization as shown in the following:



V. Discussion

A much clearer vision is achieved with the explained methodology. Segmenting the city of Dubai based on a variety of attributes provides a broader perspective on the **23** specified locations. Each location has more or less concentration or mixture of venues. While some locations had 100 venues, others had fewer than 10.

VI. Conclusion

With the advances in technology, accuracy is increasing. Data Science provided powerful tools and approaches combined with best and complex methodologies which were previously not possible or extremely difficult.

VII. References:

(1) Dubai Statistic Centre
(2) Dubai-Wikipedia
(3) Bayut.com
(4) Foursquare API
(5) NYU Repository