Housing Sale Prices and Venues Data Analysis of Jeddah City

Introduction

1.1 Background

Jeddah is second largest city of Saudi Arabia with 3.5 million people and growth rate of 3.5% per annum. Jeddah population represents 14% of the country population. The City is an economical hub with a seaport and spread over 1762 square km. The population density of Jeddah is 2.5 km.

The city of Jeddah is located on the west coast of the Kingdom (latitude 29.21 north & longitude 39.7 east), in the middle of the eastern shore of the Red Sea south of the Tropic of Cancer. To the east are the plains of Tihama, which are considered the low heights of the Hijaz region. To the west along the beach there are parallel chains of coral reefs.

Jeddah has grown during the last two decades of the 20th Century, which made the city a center for money and business in the Kingdom of Saudi Arabia and a major and important port for exporting non-oil related goods as well as importing domestic needs.

There are 140 districts in the city. Each disct consist of commercial and residential venues.

Here I want to develop a system that will help investors and residents for selecting an area for investments.

1.2 Problem

The city with high populations and 140 districts, investors and shop owners need a system that will help them to select a suitable place for their next store having low real state values and more business activities. As there is not such information available at any platform to guide the investors for a better and effective decision making.

we will map an information chart where the real estate index is placed on the city and each district is clustered according to the venue density.

1.3 Interest

The system has a great demand from fast food chain restaurants, saloons, beaty parlors and service centers. This project will be extremely helpful for them in their next move of opening a new business unit.

2. Data Acquisition and Cleaning

As the data is not available in readymade format, so we collect it from different sources spread on the internet and other data repositories.

2.1 Data Source

- As Jeddah have 140 administrator units so we need to have the name of each district and it location coordinates e.g. latitudes and longitudes.

We collect the list of Dist. from the url below and scrap it
Jeddah - Wikipedia

The scraped list dist is stored in into a data frame. As we just have the list of Dist. names, we also need the latitudes and longitudes of the those dist. So we collect latitude and longitude through google.

- Foursquare API is used to collect the venues in the nearby Borough.

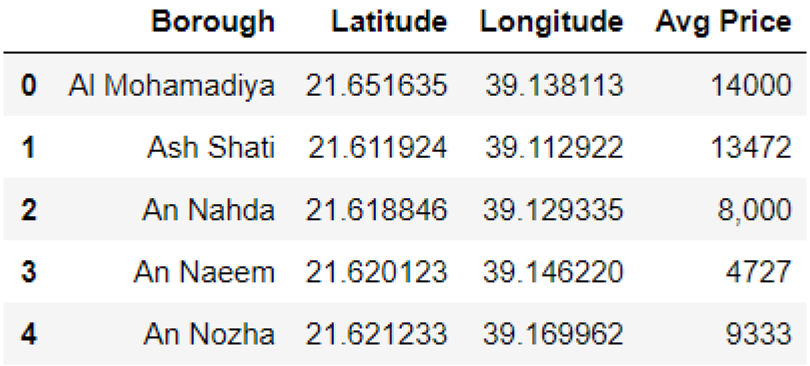
- We go through and manually search the house prices in each Borough

2.2 Data Cleaning

We have 140 Dist in the list. We short list this to 45.

2.3 Feature Selections

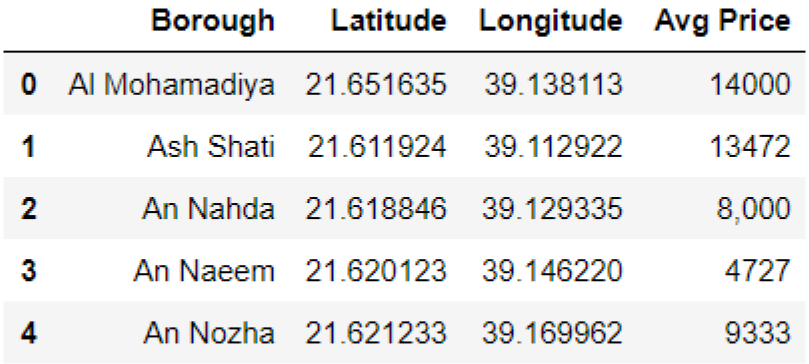
For our purpose we need name of Borough, Latitude, longitude, Avg Price(@ Sqr Meter). Here is the list.



	Borough	Latitude	Longitude	Avg Price
0	Al Mohamadiya	21.651635	39.138113	14000
1	Ash Shati	21.611924	39.112922	13472
2	An Nahda	21.618846	39.129335	8,000
3	An Naeem	21.620123	39.146220	4727
4	An Nozha	21.621233	39.169962	9333

3.Exploratory Data Analysis

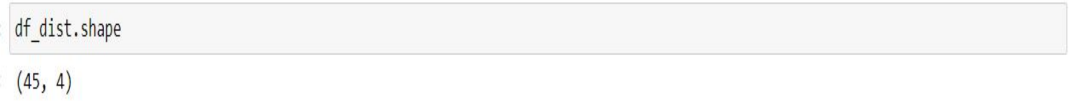
Our basic data frame name is “df_dist”.



	Borough	Latitude	Longitude	Avg Price
0	Al Mohamadiya	21.651635	39.138113	14000
1	Ash Shati	21.611924	39.112922	13472
2	An Nahda	21.618846	39.129335	8,000
3	An Naeem	21.620123	39.146220	4727
4	An Nozha	21.621233	39.169962	9333

The dataset has 4 columns Borough, Latitude, Longitude and Avg Prices.

The data frame has a shape :

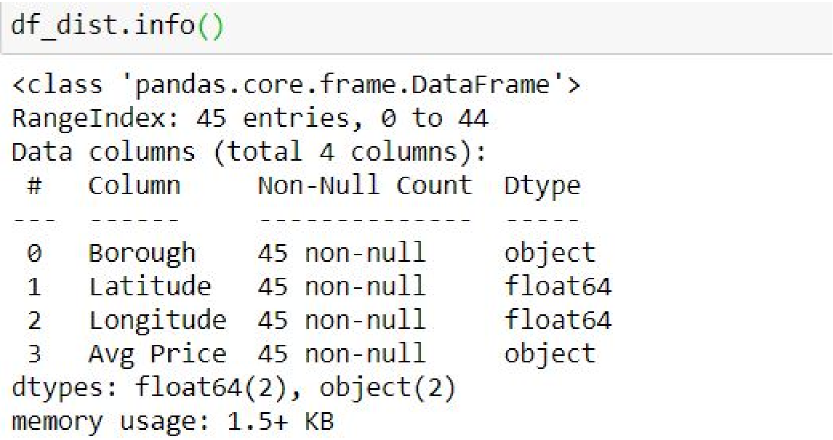


```
df_dist.shape
```

```
(45, 4)
```

It is means there are 4 columns and 45 rows.

The field type of data frame is as under. According to the below info, we don’t have any missing value in our data set.

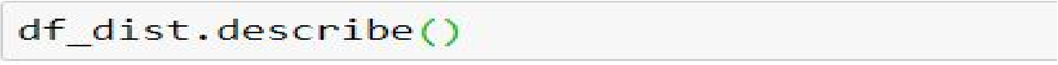


```
df_dist.info()
```

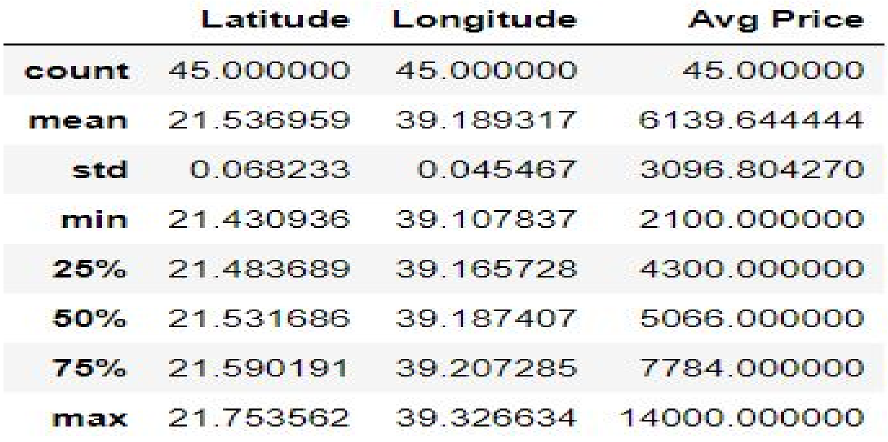
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45 entries, 0 to 44
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   Borough    45 non-null     object 
 1   Latitude   45 non-null     float64
 2   Longitude  45 non-null     float64
 3   Avg Price  45 non-null     object 
dtypes: float64(2), object(2)
memory usage: 1.5+ KB
```

We will convert the Avg data type to float type.

Here is the description of our data frame:



```
df_dist.describe()
```

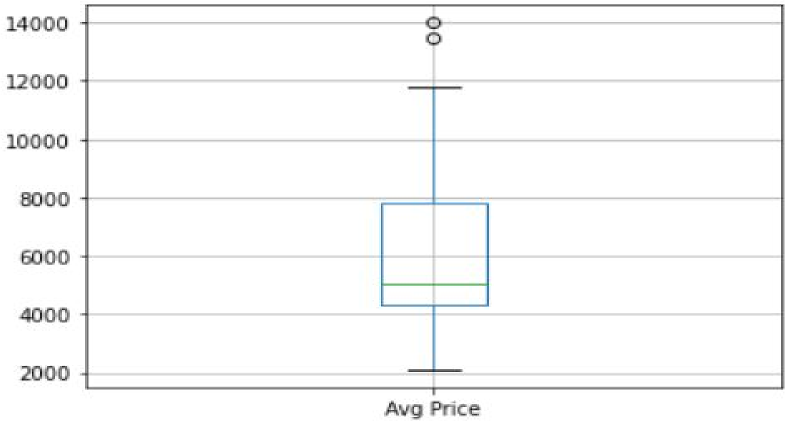


```
        Latitude   Longitude     Avg Price
count  45.000000   45.000000     45.000000
mean   21.536959   39.189317   6139.644444
std     0.068233    0.045467   3096.804270
min    21.430936   39.107837   2100.000000
25%    21.483689   39.165728   4300.000000
50%    21.531686   39.187407   5066.000000
75%    21.590191   39.207285   7784.000000
max    21.753562   39.326634  14000.000000
```

According the above description we have 45 rows, the mean of price is 6139 , the minimum price is 2100 and maximum value is 14000.

Outliers, being the most extreme observations, may include the sample maximum, sample minimum, or both, depending on whether they are extremely high or low.

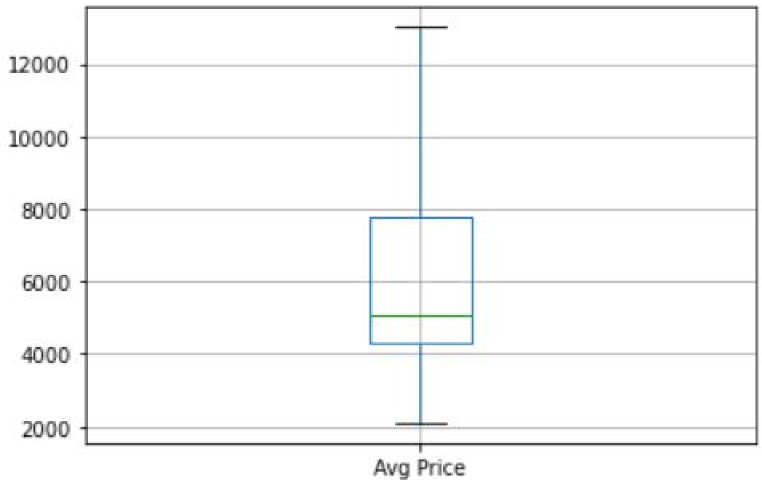
With the help of boxplot we will observe the outlier. Here is the observation for current data.



We have two solutions for outlies:

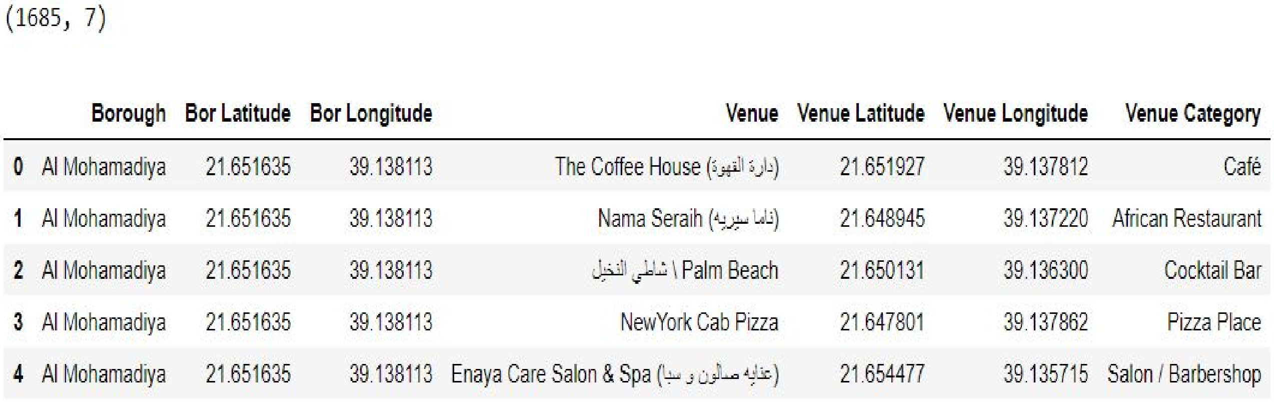
1. We drop the outlier values
2. We replace the outlier values with using IQR

Box Plot after treatment.



Venue Exploration

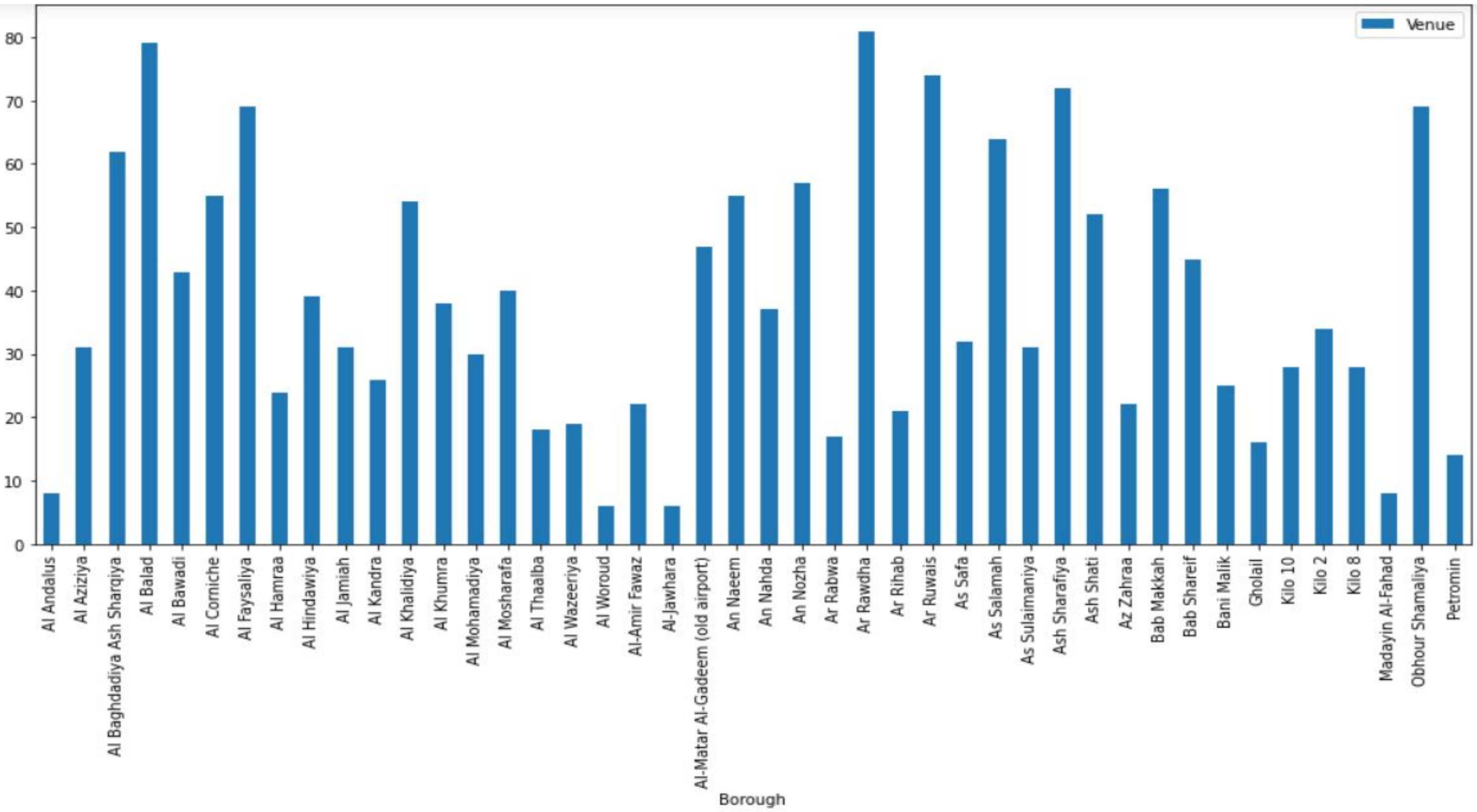
With the help of Foursquare API I collect the venues from all Borough (Dist). I try to collect 100 venues from each Borough within a radius of 800 meters. After the execution we get 1685 venues with their latitude, longitude and category. Here is snippet of the first 5 records.



```
(1685, 7)
```

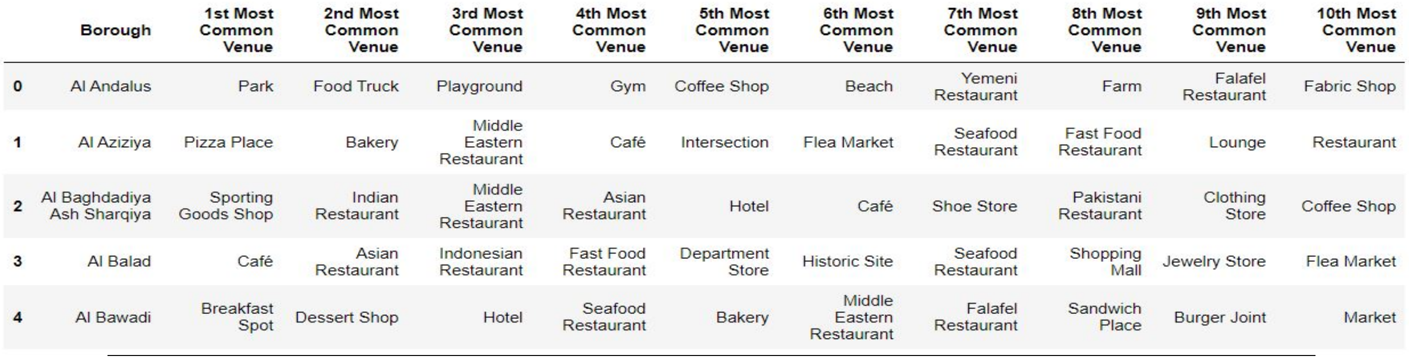
	Borough	Bor Latitude	Bor Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Al Mohamadiya	21.651635	39.138113	The Coffee House (دارة القهوة)	21.651927	39.137812	Café
1	Al Mohamadiya	21.651635	39.138113	Nama Seraih (ناما سيريه)	21.648945	39.137220	African Restaurant
2	Al Mohamadiya	21.651635	39.138113	شاطي النخيل \ Palm Beach	21.650131	39.136300	Cocktail Bar
3	Al Mohamadiya	21.651635	39.138113	NewYork Cab Pizza	21.647801	39.137862	Pizza Place
4	Al Mohamadiya	21.651635	39.138113	Enaya Care Salon & Spa (عنايه صالون و سبا)	21.654477	39.135715	Salon / Barbershop

which are more than 70 and in Al Andulas, Al Waroud, Al Jowhara and Madian Al Fahad is less than 10 venus.



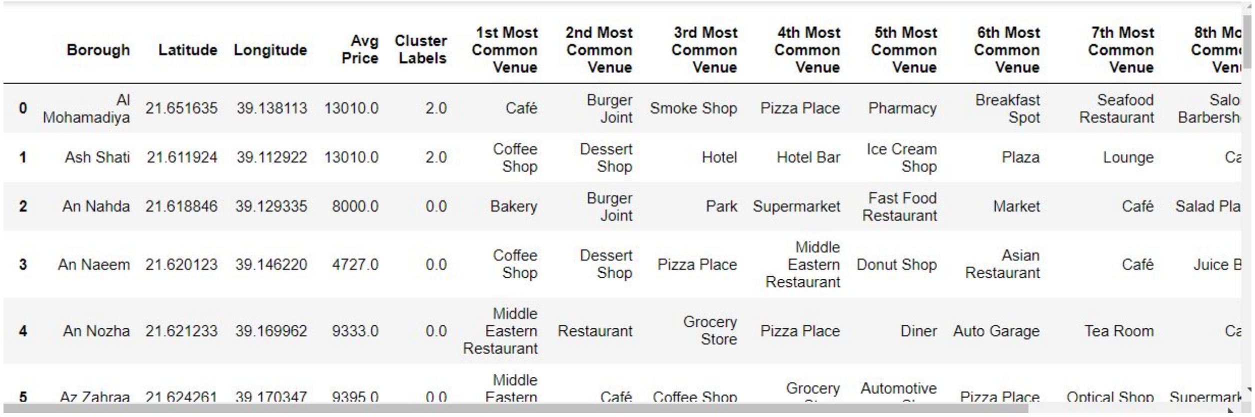
From the explored data we also find out that there 197 unique categories in the data set. For better underrating we further manipulate our data set and get the top 10 common venues from the data set.

Here is snapshot of that data set.



	Borough	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 Andalus	Park	Food Truck	Playground	Gym	Coffee Shop	Beach	Yemeni Restaurant	Farm	Falafel Restaurant	Fabric Shop
1	Al Aziziya	Pizza Place	Bakery	Middle Eastern Restaurant	Café	Intersection	Flea Market	Seafood Restaurant	Fast Food Restaurant	Lounge	Restaurant
2	Al Baghdadiya Ash Sharqiya	Sporting Goods Shop	Indian Restaurant	Middle Eastern Restaurant	Asian Restaurant	Hotel	Café	Shoe Store	Pakistani Restaurant	Clothing Store	Coffee Shop
3	Al Balad	Café	Asian Restaurant	Indonesian Restaurant	Fast Food Restaurant	Department Store	Historic Site	Seafood Restaurant	Shopping Mall	Jewelry Store	Flea Market
4	Al Bawadi	Breakfast Spot	Dessert Shop	Hotel	Seafood Restaurant	Bakery	Middle Eastern Restaurant	Falafel Restaurant	Sandwich Place	Burger Joint	Market

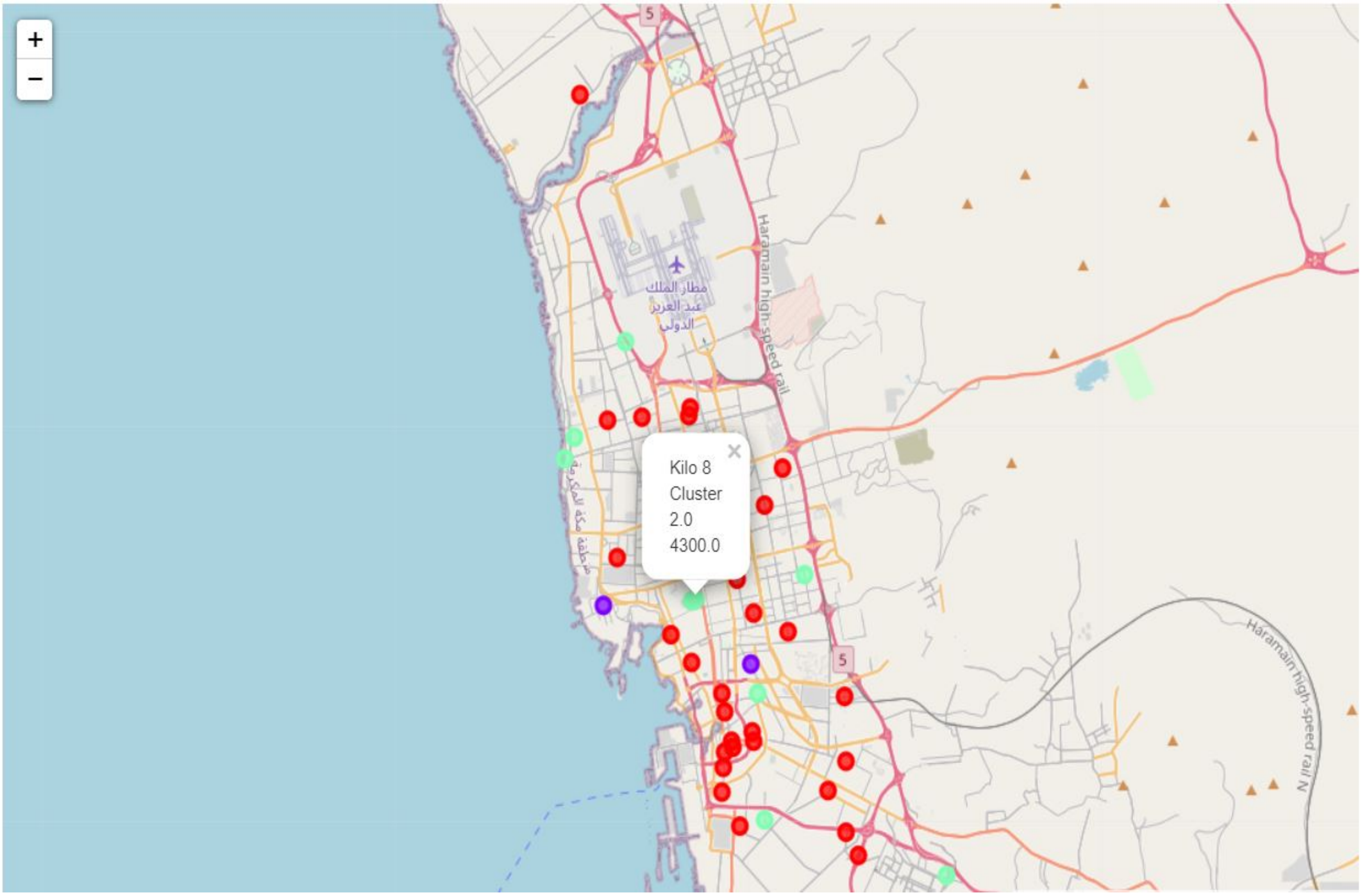
KMeans is used for data modeling and clustering. First K optimum value is calculated which is 3. After the data is fit and model the resulting clusters are merged with basic data set. Here is the merged data set.



	Borough	Latitude	Longitude	Avg Price	Cluster Labels	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 Mo Comm Ven
0	Al Mohamadiya	21.651635	39.138113	13010.0	2.0	Café	Burger Joint	Smoke Shop	Pizza Place	Pharmacy	Breakfast Spot	Seafood Restaurant	Salo Barbersh
1	Ash Shati	21.611924	39.112922	13010.0	2.0	Coffee Shop	Dessert Shop	Hotel	Hotel Bar	Ice Cream Shop	Plaza	Lounge	Ca
2	An Nahda	21.618846	39.129335	8000.0	0.0	Bakery	Burger Joint	Park	Supermarket	Fast Food Restaurant	Market	Café	Salad Pla
3	An Naeem	21.620123	39.146220	4727.0	0.0	Coffee Shop	Dessert Shop	Pizza Place	Middle Eastern Restaurant	Donut Shop	Asian Restaurant	Café	Juice B
4	An Nozha	21.621233	39.169962	9333.0	0.0	Middle Eastern Restaurant	Restaurant	Grocery Store	Pizza Place	Diner	Auto Garage	Tea Room	Ca
5	Az Zahraa	21.624261	39.170347	9395.0	0.0	Middle Eastern	Café	Coffee Shop	Grocery	Automotive	Pizza Place	Optical Shop	Supermark

The final merged data set has the Brough, latitudes, longitudes and average prices of properties with top then venues.

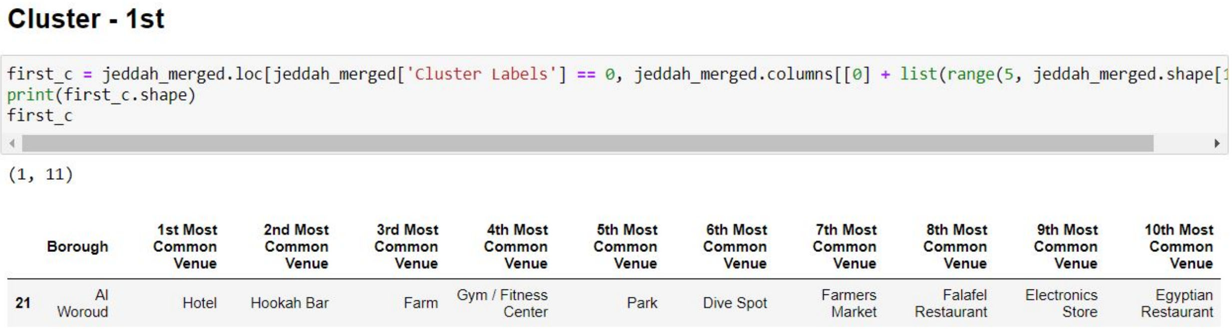
Here is the representation on the map of clustered data in different



3rd Cluster

1st Cluster

Frist Cluster has one 1 row that means there are 10 venues. Here is the shape of the 1st cluster.



Cluster - 1st

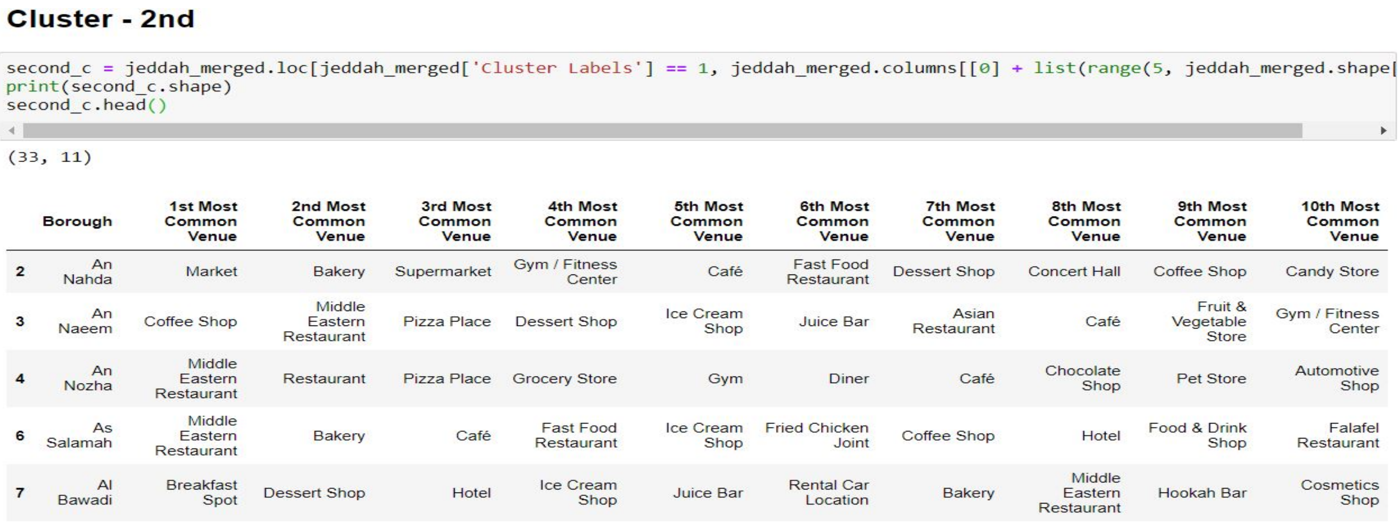
```
first_c = jeddah_merged.loc[jeddah_merged['Cluster Labels'] == 0, jeddah_merged.columns[[0] + list(range(5, jeddah_merged.shape[1
print(first_c.shape)
first_c
```

```
(1, 11)
```

	Borough	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
21	Al Woroud	Hotel	Hookah Bar	Farm	Gym / Fitness Center	Park	Dive Spot	Farmers Market	Falafel Restaurant	Electronics Store	Egyptian Restaurant

2nd Cluster

Second cluster has 33 rows that means it has 330 venues. Here is the shape and summary of the 2nd clusters.



Cluster - 2nd

```
second_c = jeddah_merged.loc[jeddah_merged['Cluster Labels'] == 1, jeddah_merged.columns[[0] + list(range(5, jeddah_merged.shape[
print(second_c.shape)
second_c.head()
```

```
(33, 11)
```

	Borough	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
2	An Nahda	Market	Bakery	Supermarket	Gym / Fitness Center	Café	Fast Food Restaurant	Dessert Shop	Concert Hall	Coffee Shop	Candy Store
3	An Naeem	Coffee Shop	Middle Eastern Restaurant	Pizza Place	Dessert Shop	Ice Cream Shop	Juice Bar	Asian Restaurant	Café	Fruit & Vegetable Store	Gym / Fitness Center
4	An Nozha	Middle Eastern Restaurant	Restaurant	Pizza Place	Grocery Store	Gym	Diner	Café	Chocolate Shop	Pet Store	Automotive Shop
6	As Salamah	Middle Eastern Restaurant	Bakery	Café	Fast Food Restaurant	Ice Cream Shop	Fried Chicken Joint	Coffee Shop	Hotel	Food & Drink Shop	Falafel Restaurant
7	Al Bawadi	Breakfast Spot	Dessert Shop	Hotel	Ice Cream Shop	Juice Bar	Rental Car Location	Bakery	Middle Eastern Restaurant	Hookah Bar	Cosmetics Shop

3rd Clusters

Third clusters has 10 rows that means it has 330 venues. Here is the summary and out of the model.

Cluster - 3rd

```
third_c = jeddah_merged.loc[jeddah_merged['Cluster Labels'] == 2, jeddah_merged.columns[[0] + list(range(5, jeddah_merged.shape[
print(third_c.shape)
third_c.head()
```

```
(10, 11)
```

	Borough	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 Mohamadiya	Café	Burger Joint	Pizza Place	Smoke Shop	Gym	Bakery	Convenience Store	Pharmacy	Cocktail Bar	Restaurant
1	Ash Shati	Coffee Shop	Café	Dessert Shop	Chocolate Shop	Hotel Bar	Hotel	Plaza	Lounge	Ice Cream Shop	Park
5	Az Zahraa	Café	Automotive Shop	Gym	Grocery Store	Middle Eastern Restaurant	Warehouse Store	Coffee Shop	Mobile Phone Shop	Cosmetics Shop	African Restaurant
13	Al Andalus	Playground	Coffee Shop	Park	Gym	Beach	Baseball Stadium	Doner Restaurant	Farmers Market	Farm	Falafel Restaurant
15	Ar Rihab	Café	Market	Ice Cream Shop	Intersection	Fruit & Vegetable Store	Soccer Field	Food Truck	Italian Restaurant	Indian Restaurant	Car Wash

5. Conclusion & Future Directions

- Building of a useful model that will help to decide the location for a store based on the available data.
- Accuracy of the model can be improved
- More real state can be collected for better accuracy of different categories.
- Ideas (Physical Data, Financial Data and stack holder involvement.)

6. Methodology

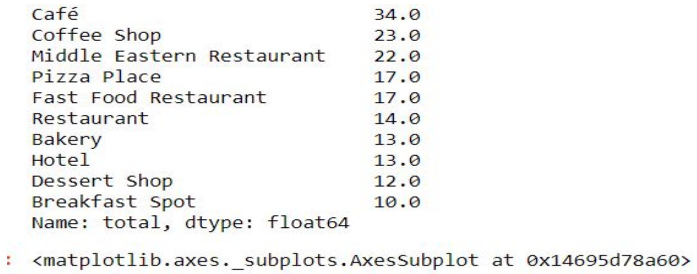
- Collection of data from different location
- Analysis of the missing data
- Cleaning of data
- Data Modeling
- Conclusion.

7. Results

We have three clusters each cluster have top 10 venue list. Here the summary of the clusters.

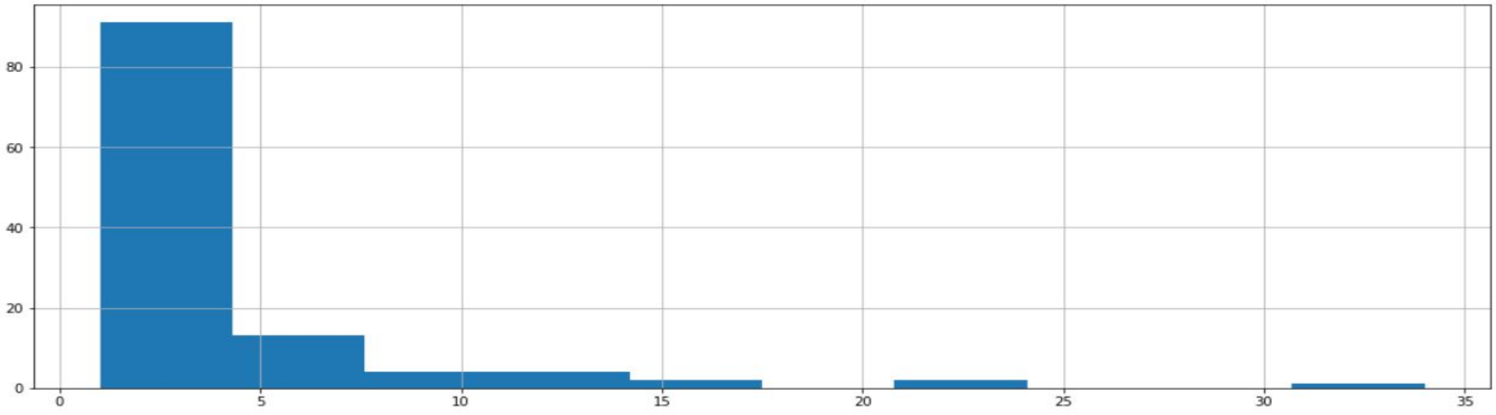
	1st Cluster	2nd Cluster	3rd Cluster
Venues	10	330	100

From the drawn result set we figure out that the Café exist in highest number, which is 34 and coffee shops are 23, please find blow the

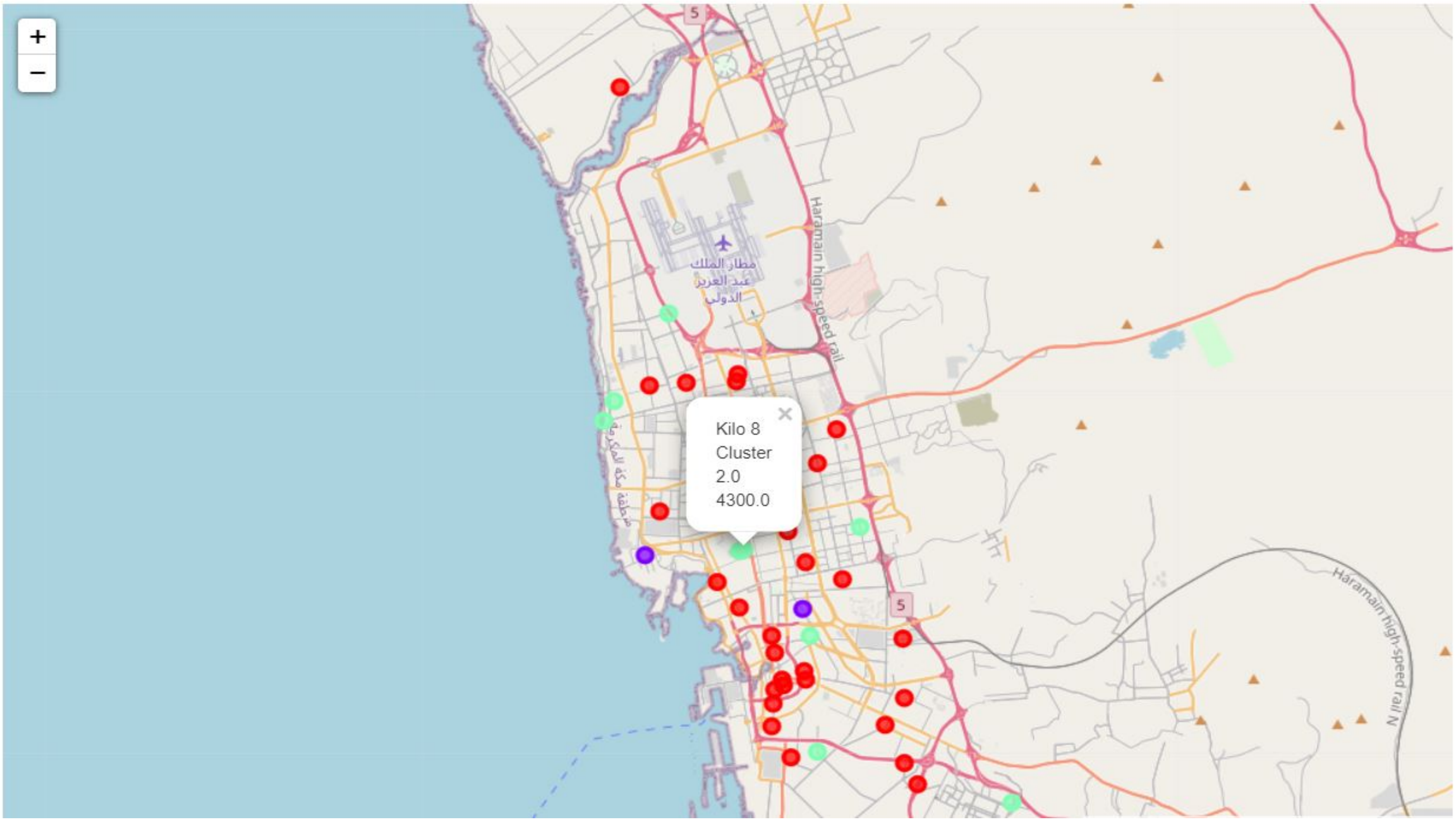


```
Café                           34.0
Coffee Shop                    23.0
Middle Eastern Restaurant      22.0
Pizza Place                    17.0
Fast Food Restaurant           17.0
Restaurant                     14.0
Bakery                         13.0
Hotel                          13.0
Dessert Shop                   12.0
Breakfast Spot                 10.0
Name: total, dtype: float64

: <matplotlib.axes._subplots.AxesSubplot at 0x14695d78a60>
```



We also draw a map that shows the average price of real state in each cluster.



8. Discussion

Jeddah is a second largest city with a high population. There are 140 districts in the city that divide the city in administrative units.

Kmeans algorithm is used for clustring. I set the optimum k value to 3. However, I only used 40 districts and collect their average real-estate prices per square meter. For better out comes data is cleaned and processed.

I add the data and my code on git hub, so that I will help users in future, if they have some thing like to work on.

I ended the study by visualizing the data and cluster it on the Jeddah map.

9. Conclusion

My conclusion is, if anyone investor, existing business expansion manager want to decide to establish a new branch or a new unit. This study will provide them a solid base to move forward.

Thanks

Anwar

References

1. https://www.jeddah.gov.sa/
2. Jeddah - Wikipedia
3. https://www.propertyfinder.sa/
4. https://www.bayut.sa/en/
5. https://www.zaahib.com/