Syed Sajjad Ali IBM Data Science Coursera

IBM Applied Data Science Capstone

Opening Shopping Malls for tourists in Lahore

July 2020

Introduction

Business Problem

The purpose of this capstone project is to find the neighborhoods in Lahore, Pakistan where there are not enough shopping malls close to the hotels so the tourists don't have to move across taking taxi or buses to do shopping and utilize the time they save exploring the landmarks of the city for which it is gaining popularity day by day. This initiative can definitely please the tourists and the more tourists are pleased with the place the more people will visit in the future and this will eventually help the tourism industry to grow in Lahore. Moreover, if the venture gets successful in Lahore this model can be replicated to the other major cities in Pakistan.

Target Audience

The project is being made for the tourism ministry of Punjab but it directly affects the tourists whether they have come from other cities of Pakistan or they travelled from other countries. For inter-city tourism the main targets are the people who travel to Lahore for business trips and they do not have a lot of time to travel far for shopping, the couples who stay at Lahore for a day or two before going to the northern areas of Pakistan to spend their Honeymoon, and the travellers who usually go for a complete Pakistan tour. The travellers from different countries come to visit to the traditional landmarks built during the Mughal Empire reign going back to the 15th century and they want to spend as much of their time, energy, and money to make that happen and shopping centers close to the places they are staying can easily let them shop, watch movies, eat, and have indoor gaming activities all at one place whenever they are free from their main purpose to visit.

Data

To solve our business problem we will use a combination of two datasets:

1. Lahore suburbs (neighborhoods) dataset from Mapcrow
2. Venues dataset from Foursquare location API

Lahore suburbs (neighborhoods) dataset from Mapcrow

Link: http://www.mapcrow.info/Lahore-PK-suburbs

The link contains the list of all the suburbs of Lahore District. There are a total of 159 unique suburbs in Lahore District but when we look at the Lahore City we limit the total suburbs to 64. First, BeautifulSoup and regex will be used to do web scraping to get the name and coordinates of all the neighborhoods in Lahore District. Python Geocoder Library will be used to get the coordinates of Lahore City. After that, Python Math Library will be used to calculate the distance between the center of Lahore City and all the suburbs. Using this method we will be able to filter all the neighborhoods in Lahore City.

Venues dataset from Foursquare location API

Foursquare is one of the most used location API in the world and it stores data for millions of places throughout along with details of each place. In this project foursquare API will be used to get the list of venues and their categories in 3km radius of each neighborhood. From there onwards, the list of venues which fall under hotels or shopping malls will be used and clustered to further investigate our business problem.

Methodology

Language: Python

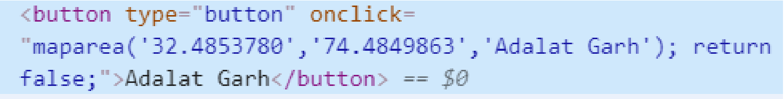
Libraries

- numpy to handle data in a vectorized manner
- pandas to work with dataframes
- json to handle JSON get requests from foursquare API
- Nominatim to convert an address into latitude and longitude values
- geocoder to get coordinates
- requests to handle GET requests
- BeautifulSoup to parse the HTML
- json_normalize to tranform JSON file into a pandas dataframe
- matplotlib to visualise data
- sklearn for Machine Learning Algorithm KMeans Clustering
- folium to render maps
- re to use regular expressions to manipulate string data
- math to calculate distance between two coordinates
- scipy spatial distance to form elbow graph for KMeans

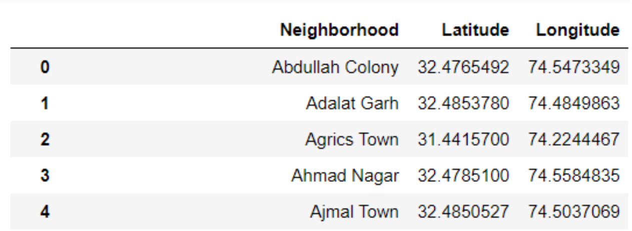
Algorithm: KMeans Clustering

Step by Step Process

- Importing the dataset from MapCrow using the link http://www.mapcrow.info/Lahore-PK-suburbs
- Parsing the HTML with BeautifulSoup to get neighborhoods with their coordinates. Now once we have one row each for one neighborhood we will clean the data and insert it into our neighborhoods dataframe. We will use regex to remove the unnecessary part from each row i.e. brackets, single quotes, semicolon, extra text. After that we will split the data with comma and insert the data into the dataframe

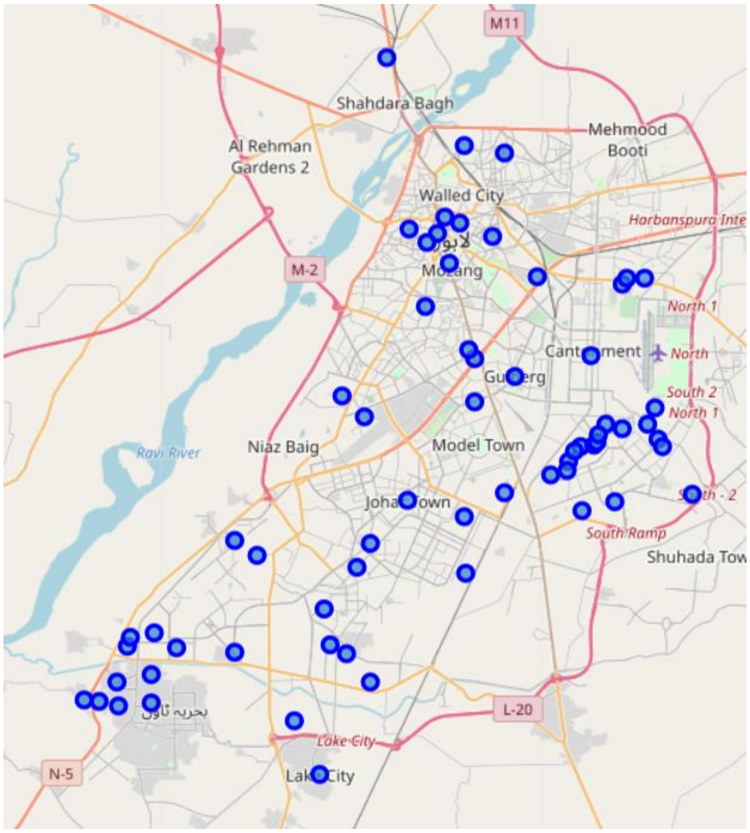


```
<button type="button" onclick=
"maparea('32.4853780','74.4849863','Adalat Garh'); return
false;">Adalat Garh</button> == $0
```

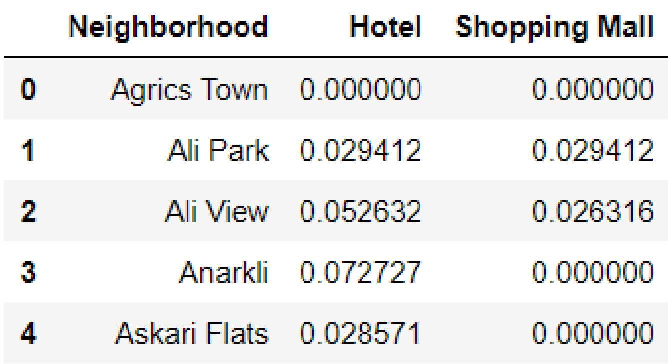


	Neighborhood	Latitude	Longitude
0	Abdullah Colony	32.4765492	74.5473349
1	Adalat Garh	32.4853780	74.4849863
2	Agrics Town	31.4415700	74.2244467
3	Ahmad Nagar	32.4785100	74.5584835
4	Ajmal Town	32.4850527	74.5037069

- Filtering the data from Lahore District to Lahore City limiting neighborhoods from 159 to 64 using a function that calculates distance between Lahore City and each neighborhood and limiting maximum distance to 50km. After that we visualize the neighborhoods on the map.

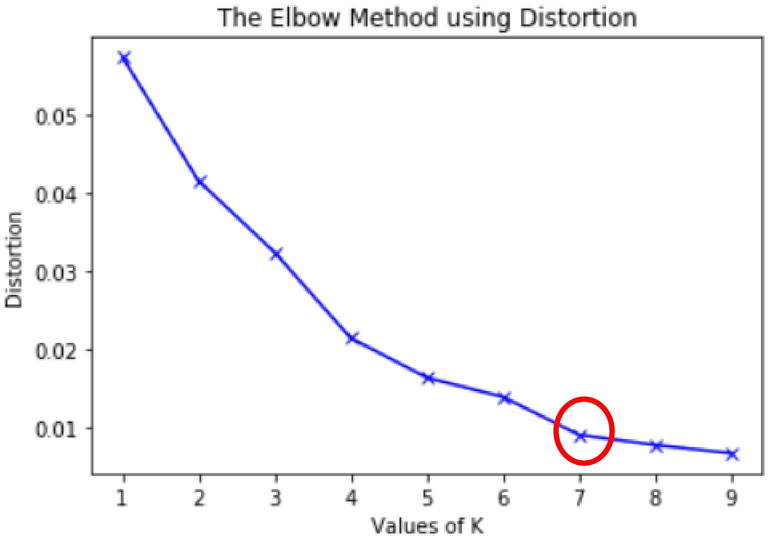


- Get venues for each neighborhood using Foursquare API, group the categories by neighborhoods, apply one hot encoding on the categories and group the resultant data by Neighborhoods with mean values for each category.
- Next step is to filter the data as we only need the Neighborhoods, Hotels, and Shopping Mall columns.



	Neighborhood	Hotel	Shopping Mall
0	Agrics Town	0.000000	0.000000
1	Ali Park	0.029412	0.029412
2	Ali View	0.052632	0.026316
3	Anarkli	0.072727	0.000000
4	Askari Flats	0.028571	0.000000

- Then we used the elbow method to find the optimal number of clusters for our algorithm. This is what we got.



- As shown in the diagram, the optimal clusters were 7 so we trained our model with the filtered data and 7 clusters.
- After that we examined the results by visualizing the clusters and looking at the underlying data in each cluster.

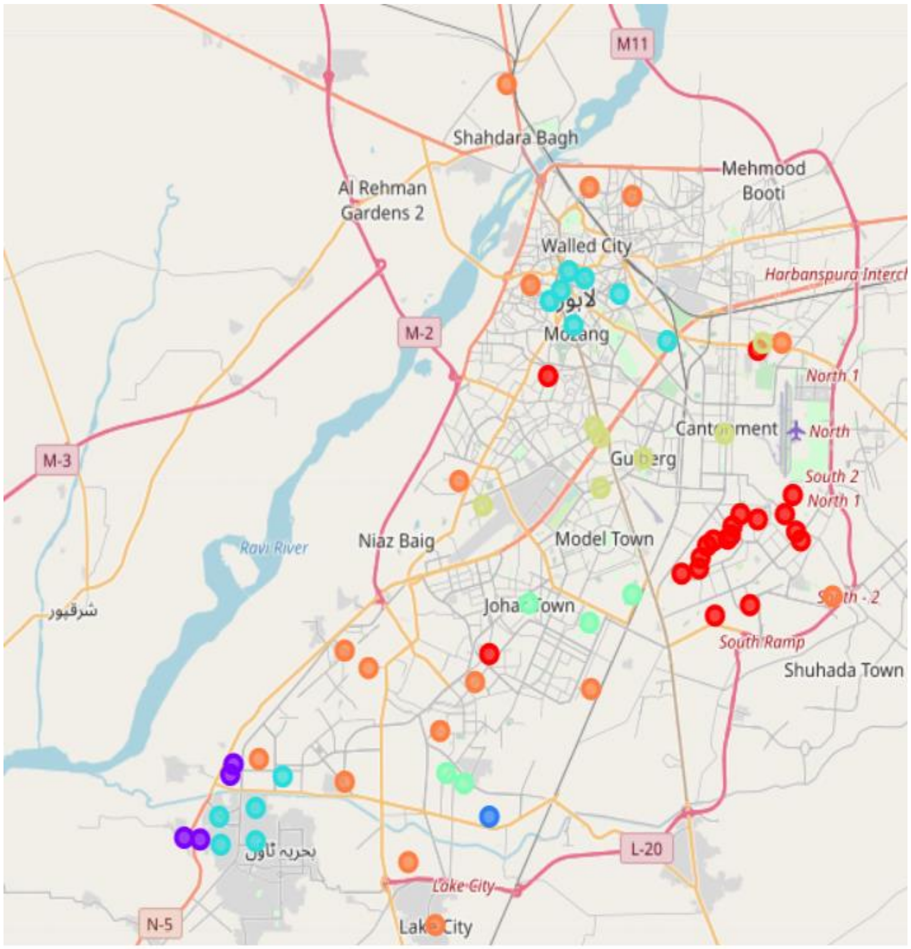
Results

The results from the k-means clustering show that we can categorize the neighborhoods into 7 clusters based on the relationship between Hotels and Shopping Malls in the neighborhoods.

The results are as follows:

Cluster 0 (19 neighborhoods) : Medium/High hotels and Medium/High shopping malls
Cluster 1 (4 neighborhoods): High hotels and no shopping malls
Cluster 2 (1 neighborhood): No hotels and High shopping malls
Cluster 3 (12 neighborhoods): Medium/High hotels and low shopping malls
Cluster 4 (5 neighborhoods): Low hotels and Medium/High shopping malls
Cluster 5 (7 neighborhoods): No hotels and low shopping malls
Cluster 6 (16 neighborhoods): No hotels and no shopping malls

The results can be visualized on the following map:



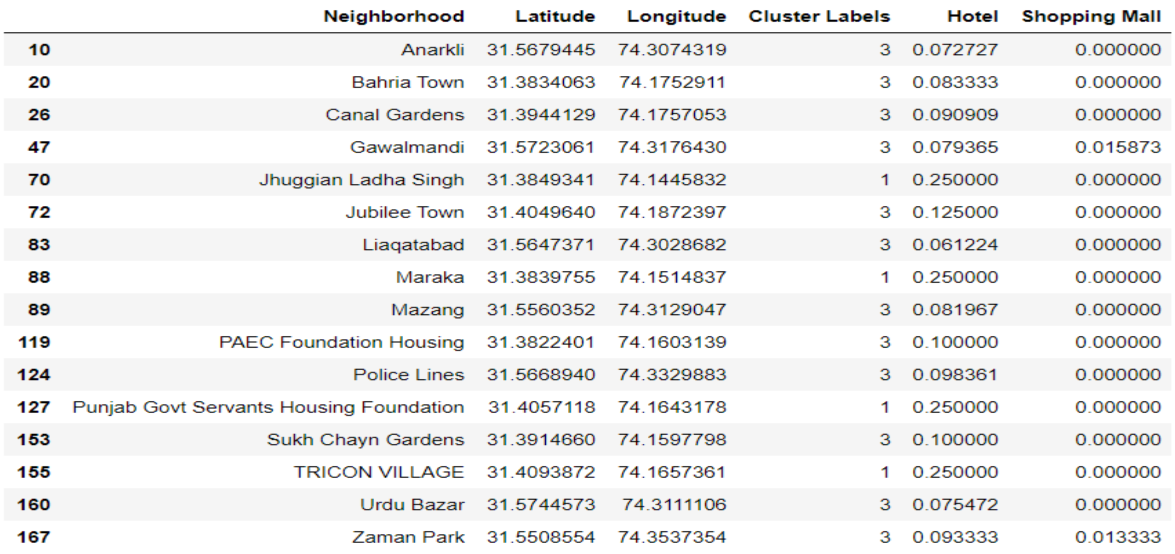
0: Red
1: Purple
2: Dark Blue
3: Light Blue
4: Light Green
5: Olive Green
6: Orange

Discussion

- Cluster 0 gives neighborhoods with a healthy relationship between the hotels and shopping malls since the more hotels are present in the area the more shopping malls are there.
- Cluster 1 gives neighborhoods with a really unhealthy relationship between the hotels and shopping malls since there are no shopping malls in the area although there are hotels
- Cluster 2 gives neighborhoods with no hotels but a presence of shopping malls
- Cluster 3 gives neighborhoods with a really unhealthy relationship between the hotels and shopping malls since there are almost no shopping malls in the area although there are hotels but the number of hotels are lower than that in Cluster 1
- Cluster 4 gives neighborhoods with low hotels but a presence of shopping malls but the presence is low
- Cluster 5 gives neighborhoods with no hotels but a presence of shopping malls but the presence is low
- Cluster 6 gives neighborhoods with no hotels and no shopping malls

Recommendations

- The neighborhoods with the clusters that are "High hotels and no shopping malls" and "Medium/High hotels and low shopping malls" because they satisfy our business problem because in these neighborhoods there is an opportunity to build shopping malls for the tourists. Those neighborhoods are shown below.



	Neighborhood	Latitude	Longitude	Cluster Labels	Hotel	Shopping Mall
10	Anarkli	31.5679445	74.3074319	3	0.072727	0.000000
20	Bahria Town	31.3834063	74.1752911	3	0.083333	0.000000
26	Canal Gardens	31.3944129	74.1757053	3	0.090909	0.000000
47	Gawalmandi	31.5723061	74.3176430	3	0.079365	0.015873
70	Jhuggian Ladha Singh	31.3849341	74.1445832	1	0.250000	0.000000
72	Jubilee Town	31.4049640	74.1872397	3	0.125000	0.000000
83	Liaqatabad	31.5647371	74.3028682	3	0.061224	0.000000
88	Maraka	31.3839755	74.1514837	1	0.250000	0.000000
89	Mazang	31.5560352	74.3129047	3	0.081967	0.000000
119	PAEC Foundation Housing	31.3822401	74.1603139	3	0.100000	0.000000
124	Police Lines	31.5668940	74.3329883	3	0.098361	0.000000
127	Punjab Govt Servants Housing Foundation	31.4057118	74.1643178	1	0.250000	0.000000
153	Sukh Chayn Gardens	31.3914660	74.1597798	3	0.100000	0.000000
155	TRICON VILLAGE	31.4093872	74.1657361	1	0.250000	0.000000
160	Urdu Bazar	31.5744573	74.3111106	3	0.075472	0.000000
167	Zaman Park	31.5508554	74.3537354	3	0.093333	0.013333

- There are also neighborhoods with no hotels and shopping malls so they can also be targeted in the future if the government has any plans of building hotels in the area.

Conclusion

In this project, we have gone through the process of identifying the business problem, using analytical approach to select how data will answer our business problem, specifying the data required, extracting data, understanding it, preparing it, train our model on the data, examining results, and lastly providing recommendations to the relevant stakeholder i.e. Punjab tourism industry. The business problem that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 and 3 out of the 7 clusters formed have the most need to open shopping malls as there are no or low shopping malls in the areas as compared to the hotels there. The findings of this project will let the government take active measures which will definitely please the tourists and the more tourists are pleased with the place the more people will visit in the future and this will eventually help the tourism industry to grow in Lahore.

References

1. http://www.mapcrow.info/Lahore-PK-suburbs
2. https://developer.foursquare.com/
3. https://github.com/syedsajjad-ali/Coursera_Capstone/blob/master/IBM%20Applied%20Data%20Science%20Capstone%20Project.ipynb
4. https://github.com/syedsajjad-ali/Coursera_Capstone/blob/master/IBM%20Applied%20Data%20Science%20Capstone.pdf