Coursera Final Project: IBM Data Science

FARRUKH NAVEED ANJUM
RAWALPINDI, PAKISTAN
ANJUM.FARRUKH@GMAIL.COM

Overview

- Introduction
- Business Problem
- Data
  - Neighborhoods
  - Geocoding
  - Venue Data

Business Problem

- Rawalpindi is adjutant city of Islamabad (The Capital of Pakistan). Thousand of people and come here for , Upon arrival in new city Rawalpindi, they need to find hotels to stay in and restaurants for eating food. Also drink brewages in coffee shops.
- Goal of the exercise will be to find the ideal spot in the city where hotel can be built for maximizing the profits

Data

- Neighborhoods
  - Data for the neighborhoods in **Rawalpindi** can be scrapped from Wikipedia using BeautifulSoup library. We will use this data.
- Geocoding
  - We will use rawalpindi.csv data. Import into the Pandas Data Frame. Lat, Long (Geo Spatial Data) can be retrieve using **FourSquare API**.
  - It will be persisted into data frame and we will store it for later on use.
- Venue Data
  - We will find the venues using **Foursquare API**. Create another data frame, that will contain the venue details along with there respective neighborhoods.

Methodology

- Accuracy of GeoEncoding

  During first phase of development with Geocoder API, the number of erroneous results were of an great amount, which led to the development of an algorithm to analyze the accuracy of the Geocoding API used.

  In the algorithm developed, Geocoding API from various providers were tested, and in the end, Google Maps Geocoder API turned. out to have the least number of collisions (errors) in our analysis.

- Folium

  It has the data wrangling strengths of the **Python** ecosystem and the mapping strengths of the **leaflet.js** library. All cluster visualization are done with help of Folium which in turn generates a Leaflet map made using **OpenStreetMap** technology.

Methodology

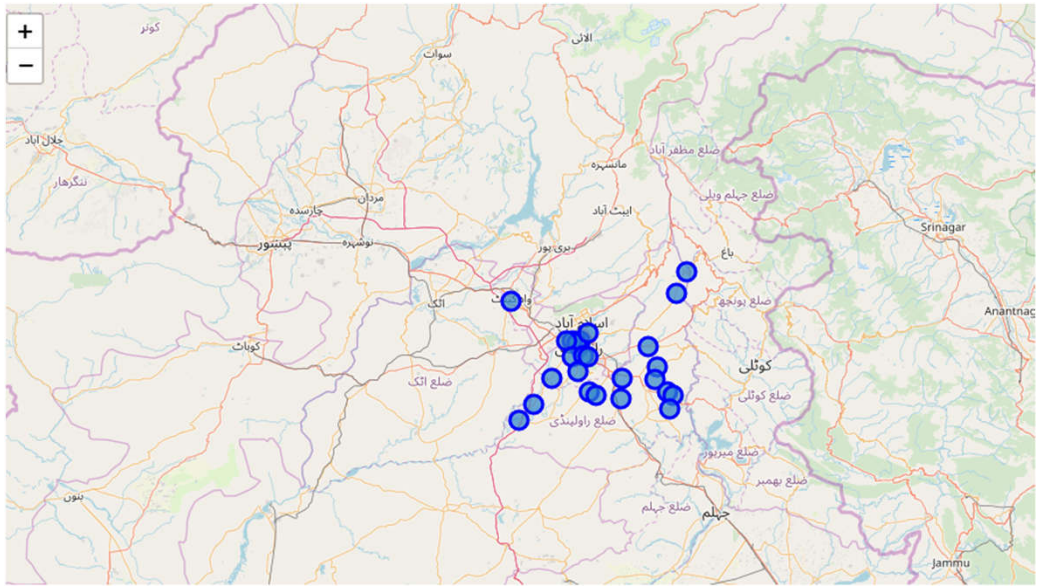

Figure: Neighborhoods of Rawalpindi.

Methodology

- One Hot Encoding

  One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.

  For the **K-means Clustering** algorithm. All unique items under Venue Category are one-hot encoded.

- Top 10 most common venues

  As we have vast variety in the venues, only the **top 10 common venues** are selected and a new Data Frame is made, which is used to train the **K-means Clustering Algorithm.**

Methodology

- Optimal Numbers of Clusters

  Silhouette Score is a measure of how similar an object is to its own cluster **(cohesion)** compared to other clusters **(separation).** The silhouette ranges from **-1 to +1**, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

  As per the **Silhouette Score** of various clusters below **20**, the optimal cluster size.

Methodology

- KMeans Clustering

  The venue data is then trained using KMeans Clustering Algorithm to get the desired clusters to base the analysis on. KMeans was chosen as the variables (Venue Categories) are huge, and in such situations KMeans will be computationally faster than other clustering algorithms.

Results

- The neighborhoods are divided into N clusters where N is the number of clusters found using the optimal approach. The clustered neighborhoods are visualized using different colors so as to make them distinguishable.

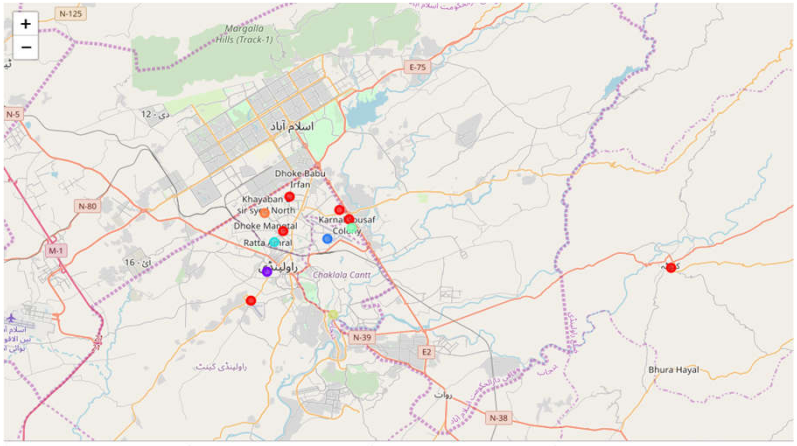

Figure: Neighborhoods of Rawalpindi (Clustered)

Discussion

- Five places namely Bhall, Dhamial Camp, Kahula, Raja Town, Satellite Town and Urdu Bazar falls near market or either bus stations. Hence are in same cluster.

	Neighborhood	PostalCode	Borough	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 Most Common Venue	7th Most Common Venue	8th Most Common Venue	9 C
7	Bhall	47570	Islamabad	33.624901	73.113793	0.0	Bakery	Burger Joint	Bus Stop	Shopping Mall	Breakfast Spot	Bus Station	Café	Clothing Store	
17	Dhamial Camp Rawalpindi	46500	Islamabad	33.565709	73.030530	0.0	Fast Food Restaurant	Gas Station	Cricket Ground	Breakfast Spot	Burger Joint	Bus Station	Bus Stop	Café	
27	Kahuta	47330	Islamabad	33.589614	73.388553	0.0	Shopping Mall	Train	Breakfast Spot	Burger Joint	Bus Station	Bus Stop	Café	Clothing Store	
60	Rawalpindi Raja Town	46320	Islamabad	33.631516	73.106109	0.0	Outlet Mall	Gas Station	Train	Cricket Ground	Breakfast Spot	Burger Joint	Bus Station	Bus Stop	
61	Rawalpindi Satellite Town	46300	Islamabad	33.641235	73.063475	0.0	Bakery	Cricket Ground	Frozen Yogurt Shop	Clothing Store	Shopping Mall	Breakfast Spot	Burger Joint	Bus Station	
62	Rawalpindi Urdu Bazar	46020	Islamabad	33.616461	73.057487	0.0	Jewelry Store	Flea Market	Train	Cricket Ground	Breakfast Spot	Burger Joint	Bus Station	Bus Stop	

Figure: Cluster having Train as most common venue

Conclusion

- The middle class in Pakistan can loosely be defined as the section of society that comprises households with a minimum monthly income of $320. A household on average consists of six members. If this categorization is correct in a broad sense, the size of the middle class in our country has grown to nearly 50 million of Pakistan's total population of 200 million. This estimate is not based on any scientific survey but on anecdotal evidence and social observations. However, one can argue that the size of Pakistan's upper middle class is smaller, not exceeding 20 million at best.
- Hence opening the a Hotel (Along with Restaurant) near railway stations area can get one $1250 per day profit in case average of 50 people stay there.