Coursera Capstone

IBM Applied Data Science Capstone

“Opening a new Cafe in Lahore, Pakistan”

Eshban Suleman
November 2019

Business Problem

As cafe culture is on the rise in Pakistan and new cafes are being launched frequently, what is the best location in the city of Lahore to open a new cafe for maximum profit

The objective is to get a list of neighborhoods with least competition

1. Data

Data was sourced from following sources

- ➔ **Wikipedia Page**
  Provided the list of all the neighborhoods of Lahore
- ➔ **Geocoder Package**
  Geographical coordinates of those suburbs
- ➔ **Foursquare API**
  Provided with the top venues in those neighborhoods/suburbs

Wikipedia Page.

Web scraping was used to scrape the list of neighborhoods of Lahore from the following URL

[https://en.wikipedia.org/wiki/List_of_towns_in_Lahore]

BeautifulSoup was used for this process

Geographical Coordinates.

Geographical Coordinates were also needed for getting venues in those neighborhoods

Geocoder package was used to get latitude and longitude coordinates for all the neighborhoods

Foursquare API.

Foursquare API enabled us to use the geographical coordinates to get all the top venues in vicinity

This helped us in creating a well structured DataFrame for further processing

2. Methodology

Following steps were involved in our process

- ➔ **Data Acquisition**
  This involved collecting relevant data
- ➔ **Data Cleaning**
  Cleaning the data as required
- ➔ **Exploratory Data Analysis**
  To formulate a hypothesis
- ➔ **Model Development**
  To validate our hypothesis and gathering results

Data Acquisition.

This process involved obtaining the required data from different sources

The sources are mentioned in the Data section

Data Cleaning.

It is a tedious task and takes upto 70% of the time in a Data Science project according to some studies

We used different techniques of web scraping to extract just the needed data from the HTML

So, we had pretty clean data to work with

Exploratory Data Analysis.

It refers to the critical process of

- performing initial investigations on data so as to discover patterns
- to spot anomalies
- to test hypothesis and to check assumptions

We used Graphical Visualization

Model Development.

Using Machine Learning to test our hypothesis or hypothesize fresh

We used Unsupervised Learning technique knowns as Clustering

The clustering algorithm we used was k-Means Clustering

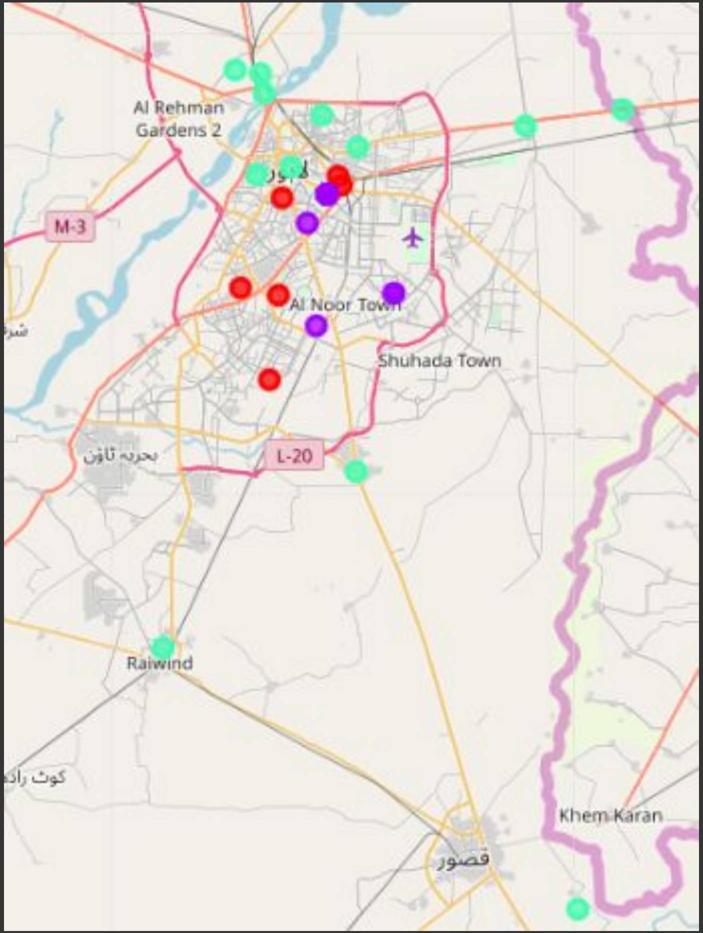
3. Results

The results from the k-Means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrences of cafes.

➔ **Cluster-0 (Red)**
Low frequency of cafes

➔ **Cluster-1 (Purple)**
High frequency of cafes

➔ **Cluster-3 (Cyan)**
Moderate frequency of cafes


Al Rehman
Gardens 2
M-3
Al Noor Town
Shuhada Town
بحریہ ٹاؤن
L-20
Raiwind
کوٹ راد
Khem Karan
قصور

4. Discussion

According to the observations made on the data visualization map in section “Results”, following assumption can be made:

- ➔ **Cluster-0**
  Low number of cafes present a much less competition in business
- ➔ **Cluster-1**
  Highest number of cafes are in these neighborhoods
- ➔ **Cluster-2**
  Presents some areas which are on the outskirts of city, also include roadside cafes on highways

5. Recommendation

According to our results, it is proved that the neighborhoods/suburbs which lie in the cluster-0 have the lowest number of cafes.

So opening a cafe in these area would provide much less competition than the areas which lies in the cluster-1, which happens to have the highest number of cafes.

Areas in cluster-2 provides an interesting insight as well. Travelling people would also need cafes along their journeys to freshen up, so it also provides a great opportunity.

Thank You.