Coursera Capstone Project

IBM Applied Data Science Capstone

Opening a Cafe in Lahore, Pakistan

By: Eshban Suleman

November 2019

Introduction

Lahore is the second largest city of Pakistan and is the capital of the province of Punjab. It houses around 11.13 million people and is among the most culturally and historically enriched cities of the subcontinent. Lahore is very famous around the globe for its authentic desi cuisine. Apart from being so historically important, Lahore has embraced the modernism quite well. One of it's modern facets includes the scintillating yet minimalist cafes. The recent wave of cafes has been very popular among the youth of Lahore. There are a lot of different cafes available for one to enjoy a nice cup of coffee with their favorite snack or just to consume the ambience for peace. Whatever the cause may be, the love for cafes is real.
Opening a new cafe is tricky task as their are a lot of parameters to be taken into account. Anything can go wrong but one of the most important parameters to consider is the location or the neighborhood in which one wants to open a cafe. Cafes represents a modern approach to living, so opening a cafe in a more advanced suburb may benefit a lot. Let's get into it more.

Business Problem

The objective of this capstone project is to analyze and select the best locations in the city of Lahore, Pakistan to open a new cafe. This project aims to provide solutions to the following business question:

- If an entrepreneur/property developer/business investor is looking to open a new cafe in the city of Lahore, Pakistan, where would you recommend them to open it?

This project implements various Data Science Methodologies and Machine Learning Techniques to answer that question.

Target Audience of this Project

This project is particularly useful to entrepreneurs/property developers/investors looking to open a cafe or to invest in one in the city of Lahore, Pakistan. This project is very timely as a rise in cafe culture has been observed in the youth of Lahore recently. Youth comprises of approximately 32% of the country's population and the cafe culture is on the rise in youth as mentioned earlier. The market is hot and ready for the taking.
Developed countries like the United States and Austria, consumes 5 kilograms and 10 kilograms of coffee per capita respectively. Whereas developing countries like Pakistan consumes less than 0.8 kilogram of coffee per capita. So, if you're a raw material provider, this might be a sector you should be interested in.

Data

To solve the problem, we will need the following data

- List of neighborhoods/towns/suburbs in Lahore. This will define the scope of the project which is confined to the city of Lahore, Pakistan.
- Coordinates of the neighborhoods i.e Latitude and Longitude coordinates of those suburbs. It's required to plot the map and also to get the venue data.
- Venue data, particularly data related to cafes in Lahore. We will use this data to perform clustering on the suburbs.

Data Source & Extraction Methodology

The wikipedia page (https://en.wikipedia.org/wiki/List_of_towns_in_Lahore) contains a table of all the neighborhoods in Lahore, with a total of 62 neighborhoods. We used web scraping techniques to extract the required data from the Wikipedia page. Web scraping involved the use of Python's requests module and BeautifulSoup library.
Then we extracted the geographical coordinates of the neighborhoods using the Python's geocoder package. It provided us with the latitude and longitude coordinates of the neighborhoods.
After that, we used the Foursquare API to get the venue data for all of those neighborhoods. Foursquare has one of the largest databases which consists of 105+ million places and is used by over 125,000 developers all around the globe. Foursquare API would provide us with a lot of different categories of venue data, but we are most interested in cafe category in order to solve the business problem at hand.
This project has made use of different Data Science methodologies which includes tasks like data extraction by web scraping, working with a 3rd party API (Foursquare), data wrangling, data cleansing, machine learning (k-Means Clustering) and intricate data visualization (Folium). In the next section, we would present the Methodology section in which we will discuss the steps taken to solve the business problem in detail.

Methodology

Business Problem

As described in the section of “Business Problem”, the problem we were trying to solve was that if an investor/entrepreneur/property developer etc wants to open a cafe in Lahore, Pakistan or wants to invest in one, which area should they choose to open/invest?
Next step was to obtain some data to build our hypothesis on.

Data Acquisition

The Data Acquisition process was described in the subsection “Data Source & Extraction Methodology” fairly. The data was sourced from a Wikipedia page which enlisted all the neighborhoods/suburbs of Lahore, Pakistan. Web scraping techniques were implemented to extract all the neighborhoods into a pandas DataFrame for comfortable data wrangling. Python’s BeautifulSoup package was used for web scraping along with the requests module.
Python’s Geocoder was used to get the latitude and longitude coordinates of the neighborhoods which we extracted from the Wikipedia page.
Foursquare API is a database of over 105 million places and is used to get all the popular venues in the extracted neighborhoods of Lahore.

Data Cleaning

The process of cleaning data is often referred to as “Data Cleaning, Data Cleansing or Data Preprocessing”. It is a tedious task and takes upto 70% of the time in a Data Science project according to some studies. Fortunately for us, the data was not as bad as it usually is. So the time to clean the data was reduced a lot.
The steps taken to preprocess the data are as follows:

1. Parsing the HTML page into text using Python’s BeautifulSoup package
2. Extracting the HTML table containing the required data
3. Extracting the HTML list containing the required data from the table
4. Creating a Python list of all the neighborhoods from the HTML list
5. Discarding the city or country name from the neighborhoods/towns/suburbs by standard string replacement operation
6. Building a Pandas DataFrame to store the neighborhoods
7. Extracting the geographical coordinates for each neighborhood and storing them in the dataframe alongside each neighborhood
8. Using the Foursquare API to get all the venues for every neighborhood and storing them in a separate dataframe
9. Concatenating both dataframes to get a single comprehensive dataframe

The above mentioned steps resulted in following dataframe

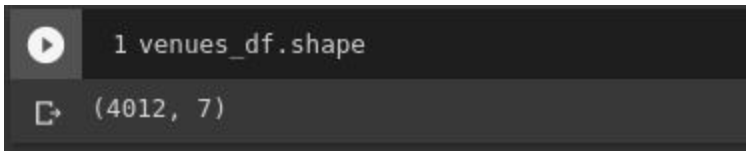
	Town	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Ravi Town	31.6149	74.2957	Ilyas Karahi	31.606977	74.306366	Pakistani Restaurant
1	Ravi Town	31.6149	74.2957	Minar-e-Pakistan	31.591604	74.309481	Monument / Landmark
2	Ravi Town	31.6149	74.2957	Fort Food Street	31.587092	74.311538	Food Court
3	Ravi Town	31.6149	74.2957	Badshahi Masjid	31.588195	74.311354	Mosque
4	Ravi Town	31.6149	74.2957	Fort View	31.587374	74.312010	Restaurant

The next step to perform some Exploratory Data Analysis on it.

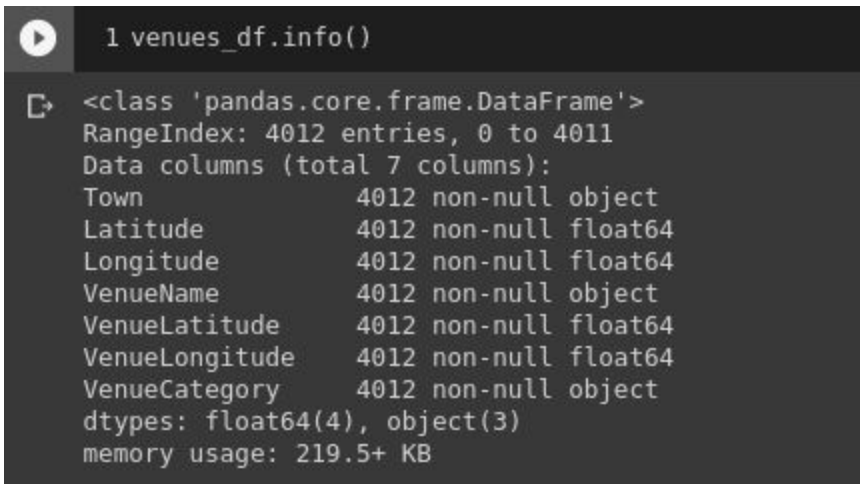
Exploratory Data Analysis

Next step in our process is to perform Exploratory Data Analysis or EDA for short. 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 with the help of summary statistics and graphical representations.
We used the following 3 steps to perform EDA on our dataset.

- Shape of dataset



- Information about dataset



- Descriptive statistics

```
1 venues_df.describe()
```

	Latitude	Longitude	VenueLatitude	VenueLongitude
count	4012.000000	4012.000000	4012.000000	4012.000000
mean	31.190866	73.848889	31.176979	73.850702
std	1.641052	1.765605	1.639055	1.767588
min	24.871700	67.005500	24.831101	66.991945
25%	31.549720	74.334200	31.508792	74.322060
50%	31.549720	74.343610	31.520656	74.345933
75%	31.549720	74.343610	31.552246	74.351830
max	33.644100	74.567760	33.686826	74.607632

Since most of the data is geographical coordinates, descriptive statistics doesn't make sense. So, we used some other techniques as well.

- Grouping Towns/Suburbs/Neighborhoods by most venues

```
1 venues_df.groupby(["Town"]).count()
```

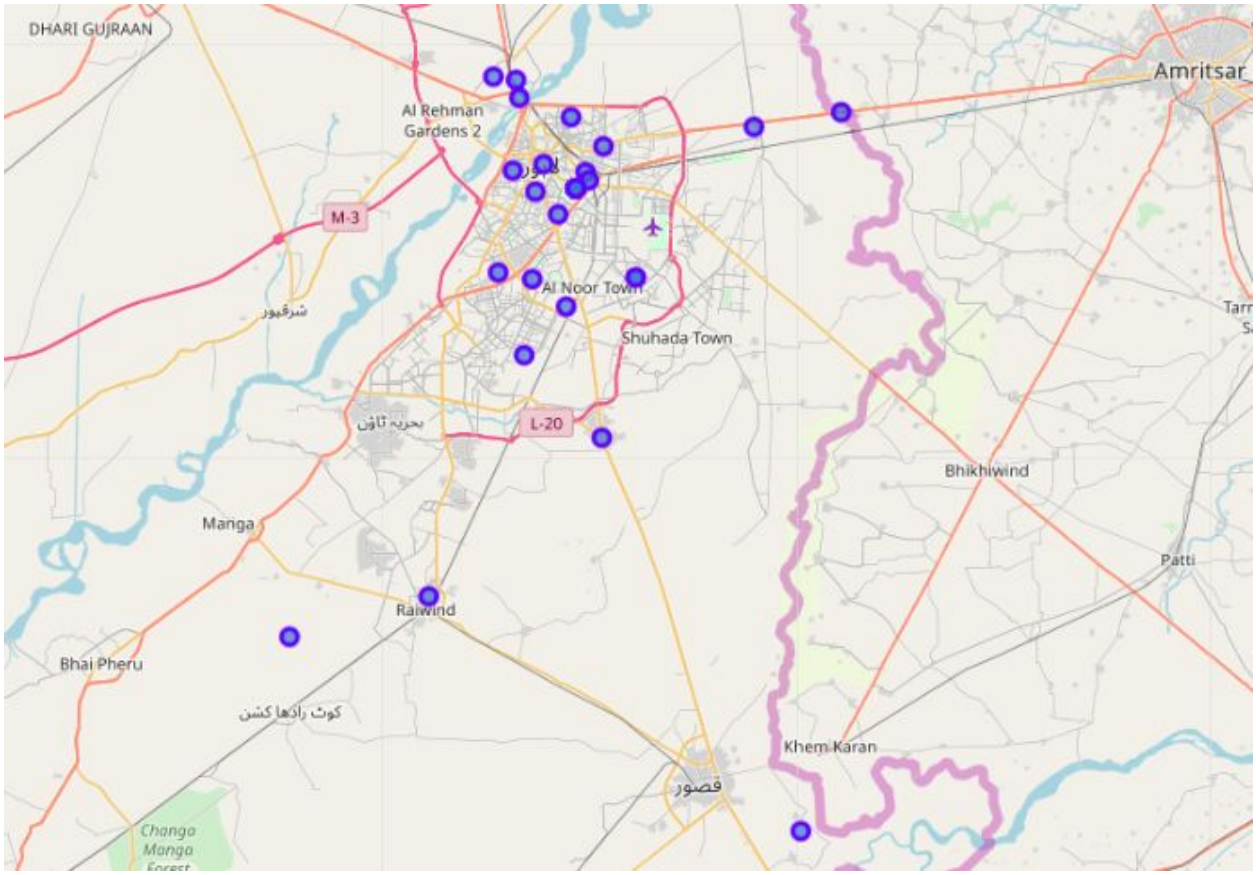
Town	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
Abdalian Cooperative Housing Society	100	100	100	100	100	100
Anarkali	68	68	68	68	68	68
Awan Town	8	8	8	8	8	8
Aziz Bhatti Town	100	100	100	100	100	100
Baba Shah Jamal	100	100	100	100	100	100
Barki	100	100	100	100	100	100
Batapur	6	6	6	6	6	6
Begampura	59	59	59	59	59	59
Cavalry Ground	100	100	100	100	100	100
Data Gunj Buksh Town	100	100	100	100	100	100
Defence	100	100	100	100	100	100
Education Town	100	100	100	100	100	100
Faisal Town	100	100	100	100	100	100
Garden Town	100	100	100	100	100	100

- Counting all the unique venues

```
1 print('There are {} uniques categories.'.format(len(venues_df['VenueCategory'].unique())))
```

```
There are 127 uniques categories.
```

- Plotting some neighborhoods on the map of Lahore



This helped us hypothesized that most venues are in the towns in Northern/North-Western Lahore. So, we might want to open a new cafe with less competition.

Model Development

EDA helped us hypothesize about our data, in the model development stage we will test our hypothesis using Machine Learning techniques. Let's go over the problem once again. We want list of neighborhoods where opening a new cafe would be most beneficial. To gather different

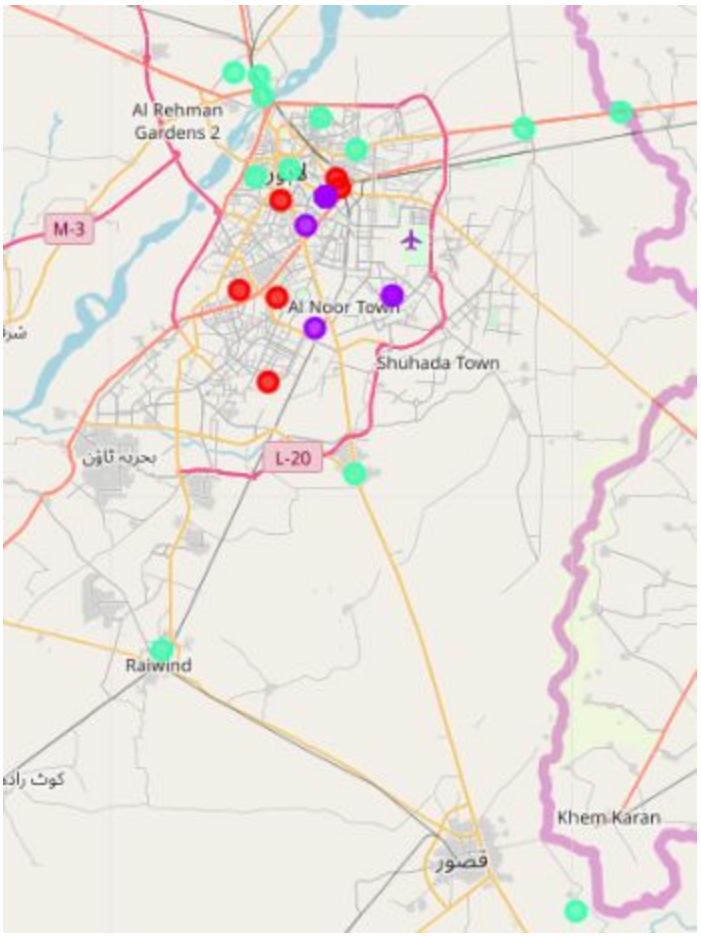
neighborhoods based on occurrences of cafes would require a Machine Learning technique called “Clustering”, which falls under the category of “Unsupervised Learning”.
The particular algorithm we will use for this problem is called “k-Means Clustering Algorithm”. This is a pretty straightforward and one of the most commonly used algorithms. We will cluster the neighborhoods/suburbs into 3 categories based on the frequency of occurrence of “Cafes”. The results will allow us to choose a neighborhood with fewer number of cafes in it. So it will help us formulate an answer for our business problem.

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: Neighborhoods with low to no existence of cafes
- Cluster-1: Neighborhoods with high number of cafes
- Cluster-2: Neighborhoods with moderate number of cafes

The results are visualized below on a map of Lahore



Discussion

According to the observations made on the data visualization map in section “Results”, highest number of cafes are saturated in the Cluster-1 (Purple) which is comprised of neighborhoods of Central Lahore. Cluster-0 (Red) is a cluster of low occurrences of cafes and includes the neighborhoods of North and North-Wester Lahore i.e Walled City of Lahore and around, it makes sense since it is culturally oldest part of the city and is very close to its roots. Cluster-2 (Cyan) is a moderate level cluster but an interesting property that this cluster exhibits is that it has many data points outside the Lahore city, mainly on highways and nearby districts.
So, we can recommend that a new cluster should open in a neighborhood that falls under the Cluster-0 as it has the lowest number of cafes and the competition is not as tough as in Cluster-1 but this also means that people are not very much attracted towards cafes in these areas/neighborhoods. But this is just a logical deduction and there’s neither any data nor any evidence to support this claim.
Cluster-2 also provide a list of good neighborhoods to open a new cafe if an investor/property developer is interested in a roadside cafe which is a must for any travelers travelling through Cluster-2 highways etc.

Limitations & Future Work

In this project, we only took one factor into account i.e. frequency of occurrence of cafes but we are fully aware of the fact that there are a lot of other factors that come into play in determining the success of a cafe. Some of these factors include the average income of neighborhoods, literacy rate in those neighborhoods/suburbs, political and economic standing of that part of town etc. However, to the best knowledge of the researcher such data wasn’t available to begin with. So an alternate strategy was devised to just consider the frequency of cafes in different suburbs. Foursquare API was used in their Sandbox Tier Account which reduces the results greatly. Using their paid services would have provided a more rich data to work with.
The future work will be focused on leveraging this current frequency based feature to create an API endpoint which will receive the city and desired venue as input and return the optimal cluster enlisting all the neighborhoods in it.

Conclusion

We traversed the entire Data Science pipeline in this project and completed all the required steps to get a solution for our business problem. From specifying the problem, obtaining the required data, extracting and wrangling the data, preprocessing data, performing Machine Learning on that data to test our hypothesis, to providing recommendations to the stakeholders, everything was covered in this project.

The solution provided by this project for the business problem discussed in the Introduction section is as follows:

"The neighborhoods/suburbs in Cluster-0 are the most optimal and preferred locations to open a new cafe in the city of Lahore, Pakistan"

The findings of this project will help the stakeholders to decide which neighborhood might be the best option for opening a new cafe. Which in turn increase their revenue or strengthen their investment choices.

References

- List of Towns in Lahore
  [https://en.wikipedia.org/wiki/List_of_towns_in_Lahore]

- Foursquare Developers Documentation
  [https://developer.foursquare.com/docs]

- Growing Popularity of Cafe Culture in Pakistan
  [https://horeca-world.com/growing-popularity-of-coffee-cafes-in-pakistan/]

Appendix

Cluster-0

- Mayo Gardens
- Green Town
- Garhi Shahu
- Faisal Town
- Islamia Park
- Mustafa Town
- Iqbal Town

Cluster-1

- NFC Employees Cooperative Housing Society
- Qila Gujar Singh
- Mozang Chungi
- Model Town
- Sanda
- Shalimar
- Mansoorah
- Township
- Lahore Cantonment
- Lahore Cantonment
- Valencia
- Kot Lakhpat
- Johar Town
- WAPDA Town
- Nishtar Town
- Abdalian Cooperative Housing Society
- Jati Umra
- Hassan Town
- Aziz Bhatti Town
- Baba Shah Jamal
- Barki
- Cavalry Ground
- Ichhra
- Data Gunj Buksh Town
- Defence
- Education Town

- Garden Town
- Wagha
- Ghurki

Cluster-2

- Anarkali
- Awan Town
- Islampura
- Shahdara Bagh
- Shad Bagh
- Harbanspura
- Sabzazar
- Gawalmandi
- Ravi Town
- Gulberg
- Batapur
- Begampura
- Kahna Nau
- Gulberg
- Krishan Nagar
- Ladheke
- Raiwind
- Muslim Town