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.