Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Coffee Shop in Sumaré, Brazil.

For coffee lovers, it's always welcome a nice place with a good location, especially when you want invite somebody to talk and share a good coffee.

Connection it's what we want between people, for some people it is practically synonymous of "Good morning."

Others prefer to drink it in the afternoon and many even confess that they drink several cups throughout the day, the fact is that, regardless of preferences, coffee is part of the lives of the majority of the population.

As the result, there a lot of coffee shop in the city of Sumaré and places being built, opening a new coffee shop there are a lot of considerations and the most important decision is the place that will be.

Business Problem

The target of this capstone project is identify and select the best locations in the city of Sumaré, Brazil to open a new coffee shop. Using tools of data science methodology and machine learning techniques like clustering, this project will provide results to check the best place in the city of Sumaré, Brazil.

Target Audience

This research is particular useful to new investors looking for open a new coffee shop in Sumaré, Brazil. According to Exame Magazine from brazil, the Coffee sector will grow in the next 10 years 25%.

And will be needed stores to sell coffee products and promote their brands to actual public and new ones.

Data

To solve the problem, we will need the following data:

1. List of neighborhoods in Sumaré, Brazil. This defines the scope of this project which is confined to the city of Sumaré in the state of São Paulo in Brazil.
2. Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
3. Venue data, particularly data related to Coffee shops. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

This Wikipedia page (https://pt.wikipedia.org/wiki/Lista_de_bairros_de_Sumaré) contains a list of neighborhoods in Sumaré, with a total of 188 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautiful soup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.
After that, we will use Foursquare API to get the venue data for those neighborhoods.
Foursquare API will provide many categories of the venue data, we are particularly interested in the Coffee Shop category in order to help us to solve the business problem.
This is a project that will make use of many data science skills, web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

Methodology

Firstly, we need to get the list of neighborhoods in the city of Sumaré and it's available in Wikipedia page (https://pt.wikipedia.org/wiki/Lista_de_bairros_de_Sumaré).

We will extract the list of neighborhoods data and we will need to get the geo location in the form of latitude and longitude to be able to use Foursquare. Geocoder package will allow us to convert address into geo location in the form of latitude and longitude. Next step, we will put the data into a Data Frame and check neighborhoods in map using Folium package.
Now we will List the top venues that are within a radius of 1500 meters using Foursquare API.
It's needed to register a Foursquare Developer Account to make API calls to Foursquare using the geo location of the neighborhoods.
Foursquare will return JSON format and we will extract the venue name, venue category, venue latitude and longitude. Now we will examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering.
And we will filter the "Coffee Shop" as venue category for the neighborhoods.
Lastly, we will execute k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is unsupervised machine learning algorithms and useful to solve the problem for this project. We will cluster into 3 clusters based on their frequency of occurrence for "Coffee Shop". The results will allow us to identify which neighborhoods have higher concentration of Coffee Shop while which neighborhoods have fewer number of coffee shop.
Based on the occurrence of coffee shop in different neighborhoods, it will help us to answer where is the best place to open a new coffee shop.

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for "Coffee Shop":

• Cluster 0: Neighborhoods with moderate number of coffee shop, red color.
• Cluster 1: Neighborhoods with low number to no existence of coffee shop, purple color.
 • Cluster 2: Neighborhoods with high concentration of coffee shop, green color
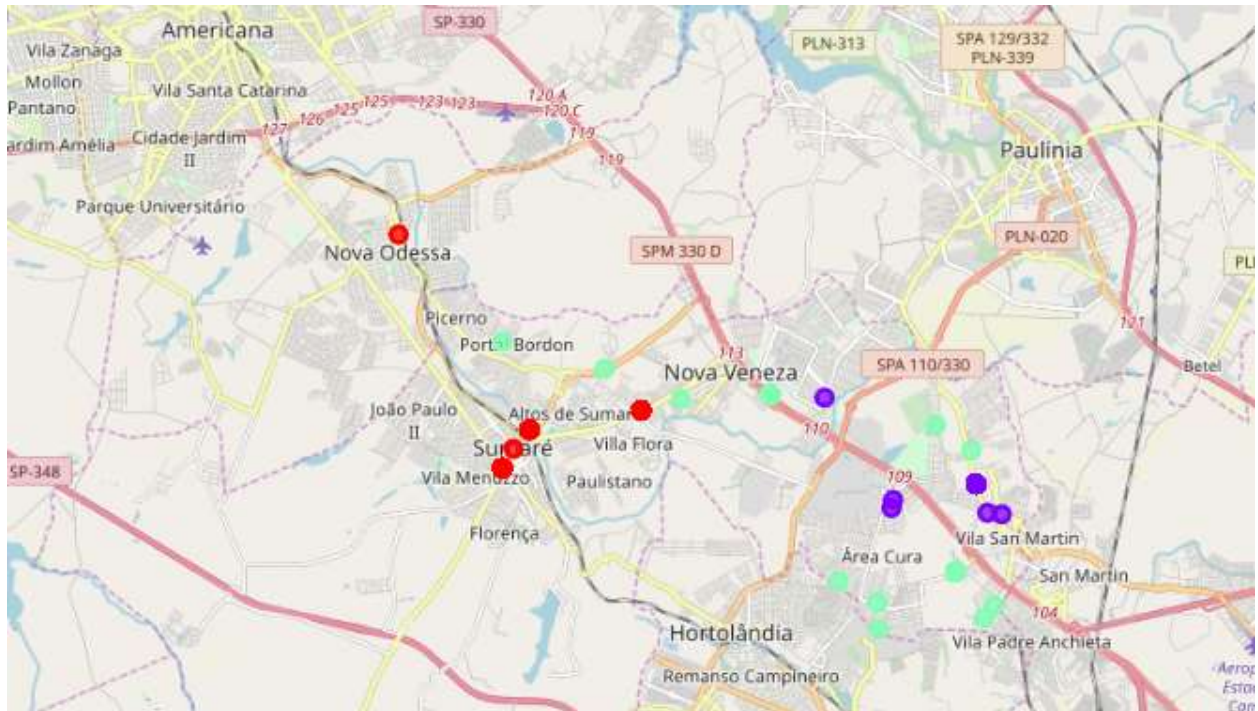
Discussion

As noted results, most of the coffee shops are high concentrated is in green color of Sumaré city, but also, purple color have very low number to no coffee shops in the neighborhoods of Sumaré city and represents a good business opportunity to open a coffee shop.
And coffee shops in green color there are too much competition and it's not a good place to open.

Limitations and Suggestions

In this project, we only consider frequency of occurrence of coffee shop, there are other factors such as traffic of people and cars, etc.

Conclusion

In this capstone we identified the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and provided recommendations to the investors about the best location to open a new coffee shop.

References

Café: um setor anticrise
https://exame.abril.com.br/economia/cafe-um-setor-anticrise/

Lista de bairros de Sumaré
https://pt.wikipedia.org/wiki/Lista_de_bairros_de_Sumaré

Foursquare Developers Documentation
https://developer.foursquare.com/docs