Where to open a bakery in São Paulo, Brazil?

IBM Capstone Project

Paulo Thadeu Soares Chavarelli

1. INTRODUCTION:

Location is the key factor when opening a venue. The pandemic caused by covid-19 caused the population to move less and to use places near their homes such bakeries and drugstores. Despite the lockdown and virus consequences, it is possible to find opportunities in the middle of the pandemic and one of them is to open the right business in the right neighborhood.

The objective of this project for Capstone is to analyze and select the best locations in São Paulo neighborhoods to open a new bakery. Having data Science tools and machine learning algorithms such as k-means and clustering, the goal of this project is to answer the question: Where to open a bakery in São Paulo, Brazil?

2. DATA:

To answer the question, the following data is needed:

- List of neighborhoods in São Paulo, Brazil.
- Latitude and longitude coordinates of the neighborhoods. Required in order to plot the map and get venue data.
- Venue data related to bakeries in São Paulo neighborhoods. Used to perform clustering in neighborhoods data frame.

The data was acquired from the Wikipedia page (https://pt.wikipedia.org/wiki/Lista_dos_distritos_de_S%C3%A3o_Paulo_por_popula%C3%A7%C3%A3o) that contains the a list of 92 neighborhoods in São Paulo. Web scraping algorithms and techniques will be used to extract the list from Wikipedia to a pandas data frame.

3. METHODOLOGY:

First of all, the data with the list of neighborhoods in São Paulo city is required, the data is available in a Wikipedia page in the form of a list. With web scraping using python request library, the data is extracted and transformed in a pandas data frame. The geographical location is needed to later on use Foursquare API, using the Geocoder libraries, we can convert addresses into latitudes and longitude that are added to the data frame. After the pandas data frame is populate with the neighborhoods and their geographical coordinates, the Folium that is a map visualization library is used to visualize the data frame with a map.

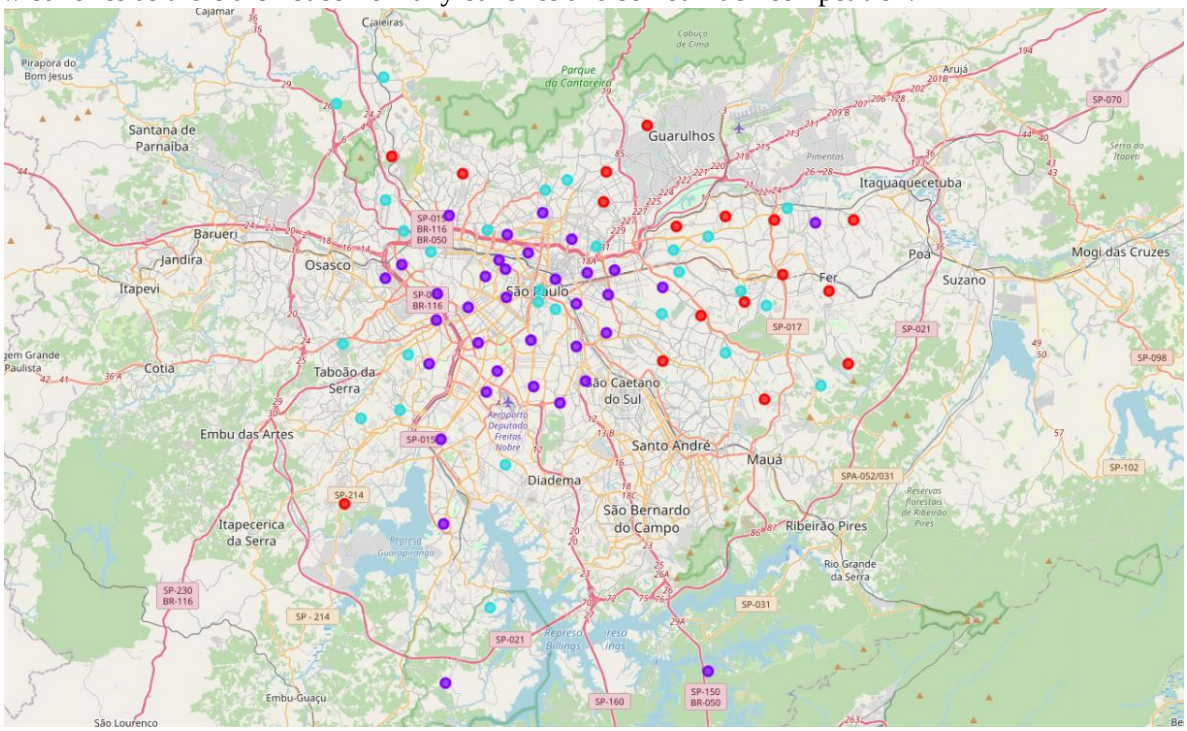
In sequence, the Foursquare API will be used to get the top 100 venues within a radius of 2000 meters in the first neighborhood called “Grajau”, to use the API credentials like client_id and client_secret is needed. Then API calls are made to Foursquare that will

return in a JSON archive the venue data containing venue name, venue category, venue latitude and venue longitude. With the Foursquare data acquired, it is possible to analyze the number of venues in each neighborhood and to know which of them are unique. Them the neighborhoods will be grouped to get the mean of frequency of occurrence of each venue category, that is needed to prepare the data for clustering.

To finish, the data is clustered using k-means clustering unsupervised algorithm that identifies k number of centroids and then allocates the data point to the closest cluster .The neighborhood is clustered into 4 clusters based on the frequency occurrence for “bakeries”, the result will let us answer the question “ where to open a bakery in São Paulo, Brazil?”

4. RESULTS:

Most of the bakeries are agglomerate in Cluster 3(purple points). Cluster 1(red points) has the fewest bakeries representing a great opportunity and high potential areas to open new bakeries as there are not some many bakeries and so not much competition.



5. DISCUSSION:

As shown in the result map, most of the bakeries are concentrated in the central area of São Paulo, the most populated part of the city represents by Cluster 1 is one with the fewest bakeries showing a great potential in building a bakery and getting great results and profits.

6. CONCLUSION:

The project gone through the processes of identifying a problem, specifying the required data, extract and prepare the data for analysis, perform machine learning algorithms to provide a recommendation based on result gathered from the k-means clustering. We must keep in mind that we only considered the frequency of occurrence factor and for further iterations is we should use other factors to have a more precise recommendation. The results of this project will help who wants to open a bakery in São Paulo city.