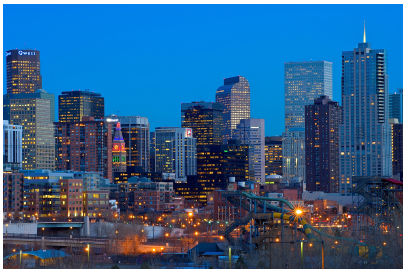
l

Capstone project

Auteur :
Warren LATA

Table des matières

1 Introduction

For many people, having some Asian restaurants near them is a great way to hang out and discover other specialties. They some diversity on neighborhoods and for the owners of these restaurants it provide consistent rental income. As a result, there are many shopping malls in the city of Denver and many more are being built.Of course, as with any business decision, opening a new restaurant requires serious consideration and is a lot more complicated than it seems. Particularly, the location is one of the most important decisions that will determine whether the mall will be a success or a failure.

2 Business Problem

The objective of this capstone project is to analyze and select the best locations in the city of Denver, Malaysia to open a new Asian Restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question : In the city of Denver, Colorado, if a property developer is looking to open a new restaurant, where would you recommend that they open it ?

3 Audience

This project is particularly useful to property developers and investors looking to open or invest in new Asian Restaurant in the city of Denver.

4 Data

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

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

— List of neighbourhoods in Denver. This defines the scope of this project which is confined to the city of Denver, Colorado.
— Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
— Venue data, particularly data related to asian restaurant . We will use this data to perform clustering on the neighbourhoods.

4.2 Sources of data and methods to extract them

This Wikipedia page (https ://en.wikipedia.org/wiki/Category :Neighborhoods_in_Denver) contains a list of neighbourhoods in Denver, with a total of 70 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods. After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the asian restaurant category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

5 Methodology

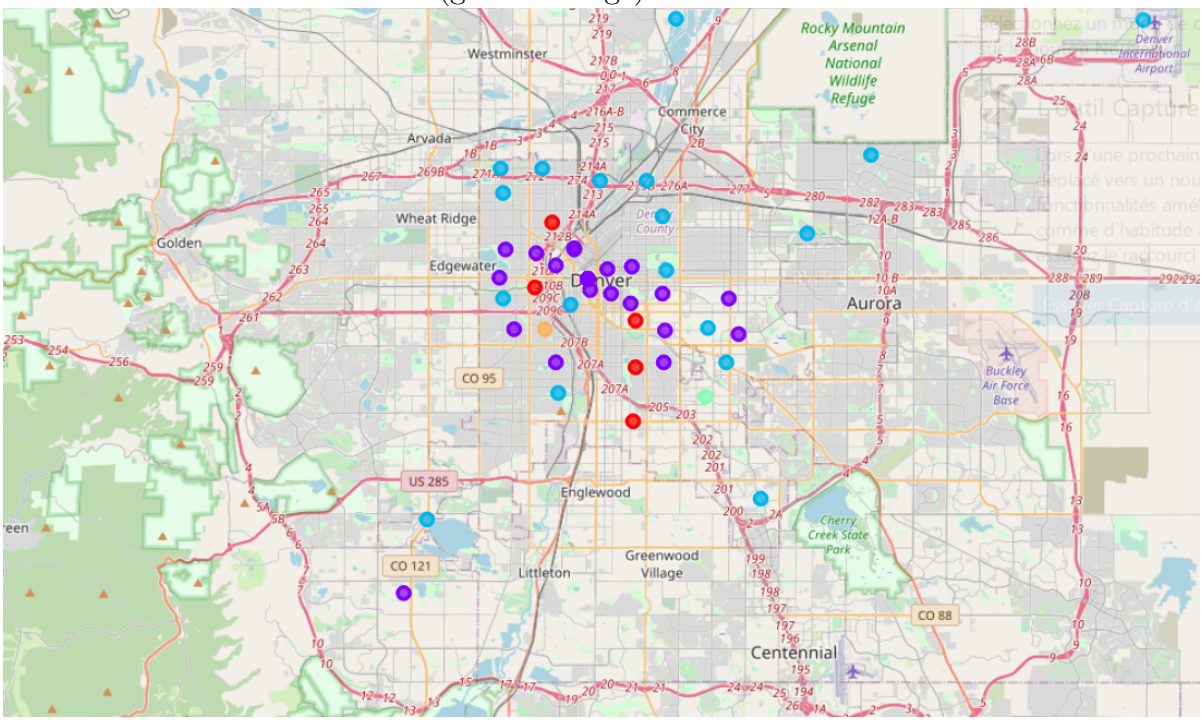
Firstly, we need to get the list of neighbourhoods in the city of Denver. Fortunately, the list is available in the Wikipedia page (https ://en.wikipedia.org/wiki/Category :Neighborhoods_in_Denver). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Denver.
Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2500 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood 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. Since we are analysing the “Asian Restaurant” data, we will filter the “Asian Restaurant” as venue category for the neighbourhoods.
Lastly, we will perform clustering on the data by using 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 one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for "Asian Restaurant". The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of Asian Restaurant different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Asian Restaurant.

6 Results

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

0 : Neighbourhoods with moderate number of restaurant and high concentration (red)Cluster 1 : Neighbourhoods with the highest number of existence of asian restaurant (blue) Cluster 2 : Neighbourhoods with no asian restaurant (purple) Cluster 3 4 : Neighbourhoods with low number of restaurant (green orange)



7 Discussion

As observations noted from the map in the Results section, Most of the Asian Restaurant are concentrated in the central area of Denver, with the highest number in cluster 3(villa park). On the other hand, cluster 2 has very no Asian Restaurant in the neighborhoods. This represents a great opportunity and high potential areas to open one as there is very no competition. In clusters 3 and 4 just one. Meanwhile, restaurant in cluster 1(central denver) are likely suffering from intense competition due to oversupply and high concentration of Asian Restaurant. Therefore, this project recommends property developers to capitalize on these findings to open new restau-

rant in neighborhoods in cluster 2 3 or 4 with little to no competition.

8 Conclusion

In this project, we have gone through the process of identifying 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 lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new asian restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is : The neighbourhoods in cluster 2 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new asian restaurant.