COURSERA CAPSTONE

IBM Applied Data Science Capstone

Opening a New Coffee Shop in Jakarta, Indonesia

By : Muhammad Sabastian Riva'i

June 2020

Introduction

Coffee shop is the most visited place to hang out with friends and release stress. Beside of that, coffee is Indonesian's most favourite beverage. The citizen of Indonesia have a huge desire to visit coffee shop to chit chat with friends and enjoy coffee. For those who wants to make a coffee shop, location is an important aspects. We have to know if we have any competitor in the area, the crowd of the area, rent cost of the area etc. Of course, as with many business decision, opening a new coffee shop requires serious consideration and a lot more complicated than it seems. Particularly, the location of the coffee shop is one of the most important decisions that will determine whether the coffee shop will be a success or a failure.

Business Problem

The objective of this capstone project is to analyze and select the best location in Jakarta, Indonesia to open a new coffee shop. Using data scince methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Kuala Lumpur, Malaysia, if an investor is looking to open a new coffee shop, where would you recommend that they open it?

Target Audience of This Project

This project is particularly useful to property developers and investors looking to open or invest in new coffee shop in Jakarta, Indonesia. This project is timely as the city is currently suffering from oversupply of coffee shops. Recently, there are a lot of coffee shops are opened in Jakarta. Chairman Specialty Coffee Association of Indonesia predict the number of coffee shops in Jakarta will increase to 20% by the end of the year. So this project will be useful for anyone who want to open coffee shop in Jakarta.

Data

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

- List of districts in Jakarta. This defines the scope of this project which is to confined to the city of Jakarta, the capital city of Indonesia.
- Latitude and longitude coordinates of those districts. This is required in order to plot the map and get the venue data from foursquare.
- Venue data, particularly data related to coffee shop. We will use this data to perform clustering on the districts.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Districts_of_Jakarta) contains a list of districts in Jakarta, with a total of 44 districts. 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 coordiinates of the districts using Python Geocoder package which will give us the latitude and longitude coordinates of the districts.

After that, we will use Foursquare API to get the venue data for those districts. Foursquare has one of the largest database of 105+ million places and is used by over 125000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Coffee Shop category in order to help us solve the business problem described above. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), machine learning (K-means clustering) and map visualization (Folium).

Methodology

First, we need to get the list of districts in the city of Jakarta, Indonesia. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Districts_of_Jakarta). We will do web scraping using Pyton requests and beautifulsoup packages to extract the list of districts 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 plot it and get venue data from foursquare API. To do so, we use the Python 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 pandas DataFrame and then visualize the districts 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 Jakarta.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare developer account in order to obtain the Foursquare ID and secret key. We then make API calls to Foursquare passing in the geographical coordinates of the districts in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, and venue coordinate. With the data, we can check how many venues were returned for each district and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each district by grouping the rows by district and taking the mean of frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the Coffee shop data, we will filter the Coffee Shop and Café as venue category for the districts.

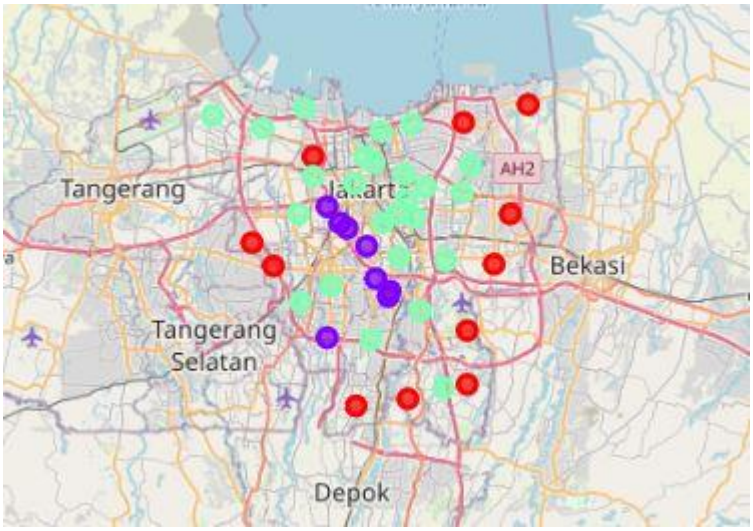
Last, we will perform clustering on the data by using K-Means clustering. K-Means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the districts into 3 cluster based on their frequency of occurrence for Coffee Shop and Café. The results will allow us to identify which districts have higher concentration of coffee shop while and which districts have fewer number of coffee shop. Bsed on the occurrence of coffee shops in different districts, it will help us to answer the question as to which districts are most suitable to open new coffee shop.

Results

The results from the k-means clustering show that we can categorize the districts into 3 clusters based on the frequency of occurrence for coffee shop and café:

- Cluster 0: Districts with low number to no existence of coffee shop/café
- Cluster 1: Districts with moderate number of coffee shop/café
- Cluster 2: Districts with high concentration of coffee shop/café

The results of the clustering are visualized in the map below with cluster 0 in red, cluster 1 in purple, and cluster 2 in mint green.



Discussion

As observations noted from the map in the results section, most of the coffee shop are concentrated in the central area of Jakarta, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to no coffee shop in the districts. This represents a great opportunity and high potential areas to open new coffee shop/café in cluster as there is very little to no competition from existing coffee shop/café.

Limitations and Suggestions for Future Research

In this project, we only consider the number of coffee shop and café while there are many other factors such as population and income of residents that could influence the location decision of a new coffee shop, the density of the districts, the environment around the districts, etc. In addition, this project made by the free sandbox tier account of foursquare API that came with limitations as to the numer of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more optimal results.