Coursera Capstone IBM

Applied Data Science Capstone

"Opening a New Coffee Shop in Jakarta, Indonesia"

By: Aries Fitriawan

August 2019

Introduction

For many coffee lovers, visiting Coffee Shop is a best way to relax and enjoy themselves during their free time. Nowadays, Coffee Shop can be a best place for working, meeting client, or just to enjoy the ambience with a taste of signature local coffee. Coffee Shop are like a multipurpose place. For the business owner, the central location and the potential of the coffee shop provides a great distribution channel to market their products and services. The barista also more creative to provide the latest taste of coffee, combined with the other ingredients to attract the customer. Shopping Mall developers are also taking advantage of this trend to build more Coffee Shop tenant in their shopping mall to cater to the demand. As a result, there are many Coffee Shop in the city of Jakarta and many more are being built. Opening Coffee Shop also can allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new Coffee Shop requires serious consideration and is 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 store will be a success or not.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Jakarta, Indonesia to open a new Coffee Shop. Using data science methodology and machine learning techniques for clustering, this project aims to provide solutions to answer the business question:

In the city of Jakarta, Indonesia, if a business owner is looking to open a new Coffee Shop, where would you recommend that they open it?

Target Audience

This project is particularly useful to business owner and/or investors looking to open or invest in new Coffee Shop in the capital city of like Jakarta. This project is timely as the city is currently suffering from oversupply of Coffee Shop. Data from the Indonesia Jajak Pendapat (JAKPAT - https://blog.jakpat.net/indonesian-coffee-drinking-habit-survey-report/) released last year showed that an additional 15.84% percent of people who drink Coffee will choose to get their coffee from the Coffee Shop and it will be grow in 2019 and 5.94% considering the location of Coffee Shop as the most important factor to buy a coffee. And many of business owner are continued obsession with building more Coffee Shop.

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 confined to the city of Jakarta, the capital city of the country of Indonesia in South East Asia.
- Latitude and longitude coordinates of those districts based on Indonesian Census and Goggle Maps (manual) Calibration. This is required in order to plot the map and also to get the venue data.
- 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

Data from census in Jakarta were scraped and longitude-latitude data were manually calibrated using Google Maps. Kepulauan Seribu were excluded from the data, because it is not related to the business problem. A total of 42 districts (Kecamatan) were collected. The cleaned data was extracted in *jkt_district.csv* file http://data.jakarta.go.id/dataset/jumlahkecamatankelurahanrtrwdankkdkijakarta/resource/1d5b0bb0-3aa7-482a-9e65-fc03d466efac

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest databases 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 Coffee Shop 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 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 districts in Jakarta. Fortunately, the list of districts data have been collected from http://data.jakarta.go.id/dataset/jumlahkecamatankelurahanrtrwdankkdkijakarta/resource/1d5b0bb0-3aa7-482a-9e65-fc03d466efac and manually search centre of districts using Google Maps. Data can be seen in in *jkt_district.csv* file. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. The visualization can be seen in Image 1.

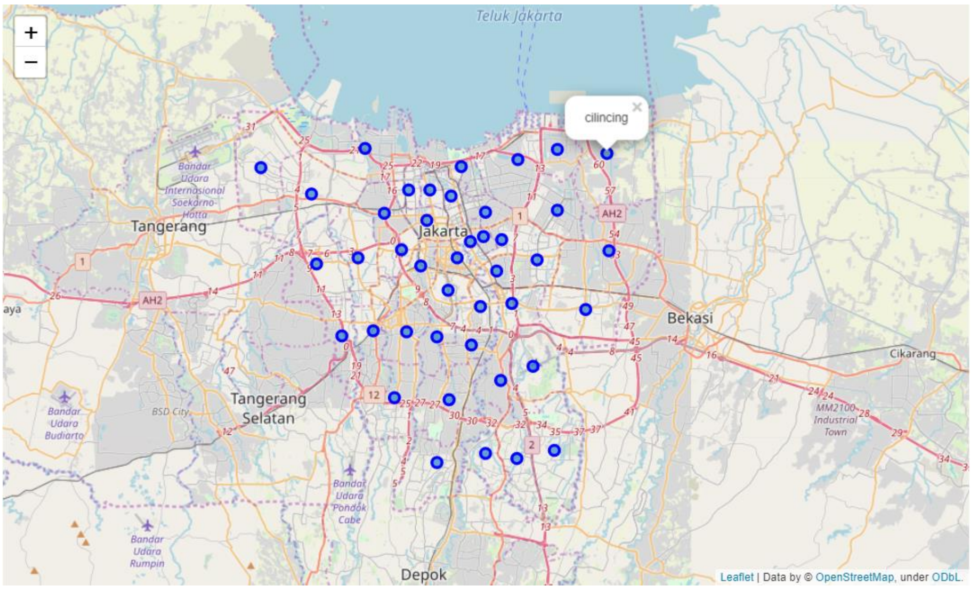

Image 1. Visualization of Jakarta Map, the blue dot represent the centre of each district (Kecamatan).

This allows us to perform a sanity check to make sure that the geographical coordinates data collected 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 Foursquare secret key. Then we make API calls to Foursquare using the geographical coordinates of the districts in a loop function. Foursquare will return the venue data in JSON format and the JSON data will be extracted by venue name, venue category, venue latitude and longitude.

With the data, we can check how many venues were returned for each districts 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 frequency mean of occurrence for each venue categories. Then, the data will be prepared for clustering processes. Since we will analyse “Coffee Shop” data, we will filter the “Coffee Shop” 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. The first step of K-means clustering is defining the number of k. In this case, the elbow method will be done for defining the best number of k. Elbow method is one method to validate the number of clusters is the elbow method. The idea of the elbow method is to run K-means clustering on the dataset for a range of values of k (say, k from 1 to 10), and for each value of k calculate the sum of squared errors (SSE). Then, plot a line chart of the SSE for each value of k. If the line chart looks like an arm, then the "elbow" on the arm is the value of k that is the best. The idea is that we want a small SSE, but that the SSE tends to decrease toward 0 as we increase k (the SSE is 0 when k is equal to the number of data points in the dataset, because then each data point is its own cluster, and there is no error between it and the center of its cluster). So our goal is to choose a small value of k that still has a low SSE, and the elbow usually represents where we start to have diminishing returns by increasing k and avoid k=2 due to very general result.

We will cluster the neighbourhoods based on their frequency of occurrence for “Coffee Shop”. The results will allow us to identify which neighbourhoods have higher concentration of coffee shop while which neighbourhoods have fewer number of coffee shop. Based on the occurrence of coffee shop in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new coffee shop.

Results

The result of elbow methods can be seen in Image 2.

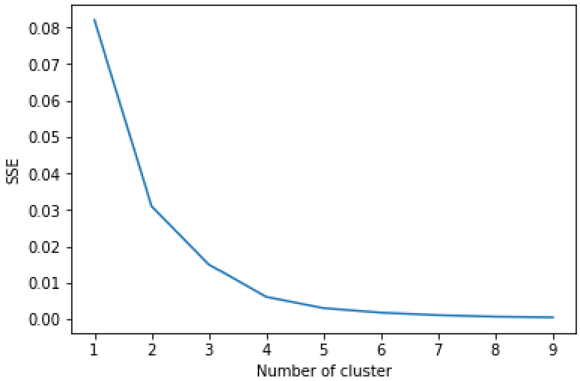

Image 2 The result of elbow method can be seen the elbow is began to appear and slightly decreasing in k=3 (we don’t use k=2, due to very general result).

The results from the elbow method for k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Coffee Shop”:

- ✓ Cluster 1: Neighbourhoods with lowest number to no existence of coffee shop

- ✓ Cluster 2: Neighbourhoods with high concentration of coffee shop
- ✓ Cluster 3: Neighbourhoods with moderate number of coffee shop

The results of the clustering are visualized in the map below with cluster 1 in purple colour, cluster 2 in green colour, and cluster 3 in red colour.

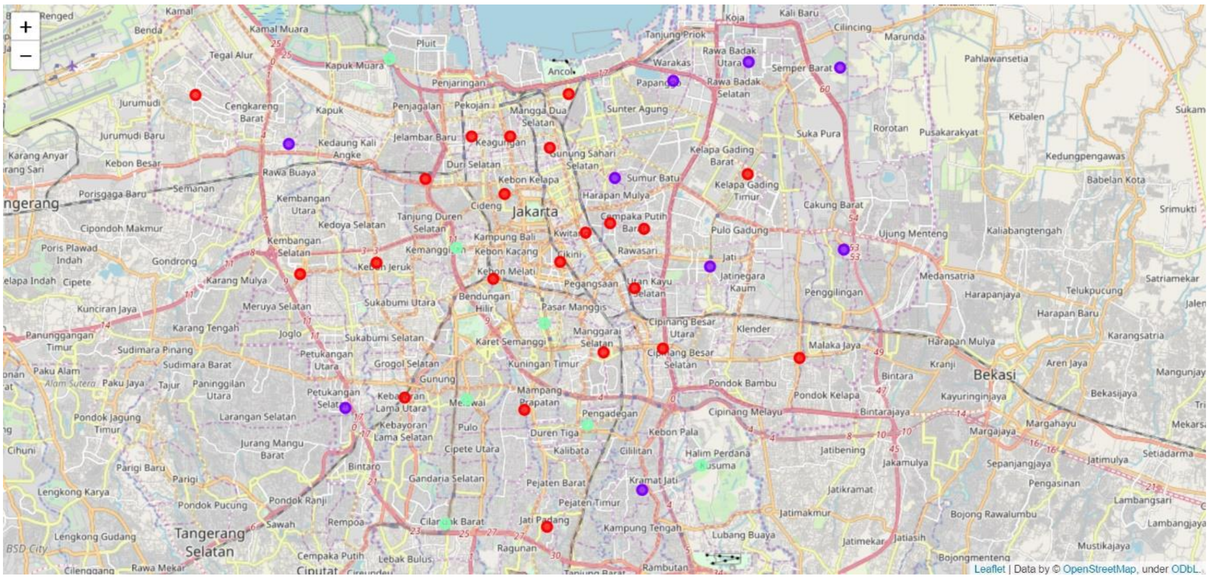

Image 3 Cluster result mapping

Discussion

As observations noted from the map in the Results section, most of the coffee shop are concentrated in the southern area of Jakarta, with the highest number in cluster 2 and moderate number in cluster 3. On the other hand, cluster 1 has very low number to no Coffee Shop in the neighbourhoods. This represents a great opportunity and high potential areas to open new Coffee Shop as there is very little to no competition from existing coffee shop. Meanwhile, coffee shop in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of coffee shop. From another perspective, this also shows that the oversupply of coffee shop mostly happened in the southern area, with the central and northern area still have very few coffee shops. Therefore, this project recommends coffee shop inventors on these findings to open new coffee shop in neighbourhoods in cluster 1 with little to no competition. Meanwhile, coffee shop inventors with unique selling propositions to stand out from the competition can also open new coffee shop in neighbourhoods in cluster 3 with moderate competition. Lastly, Coffee Shop inventors to avoid neighbourhoods in cluster 2 which already have high concentration of Coffee Shop and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of coffee shop, there are other factors such as distance between coffee shop, population and income of residents that could influence the location decision of a new coffee shop. However, to the best knowledge of this researcher such data are not available to the districts level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new coffee shop. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results and the detail of each neighbourhood.

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. coffee shop business owner, inventor and investors regarding the best locations to open a new coffee shop.

To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 are the most preferred locations to open a new coffee shop. 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 coffee shop.