Coursera IBM Data Science Capstone Project

Opening New Cafe in Jakarta, Indonesia

by : Toton Dwi Antoko
February 2021

Introduction :

Nowadays, a coffee shop is not only a place to drink coffee but also a place to hang out, work or even hold meetings. Being a forum is not only for adults but also for the millennial generation. Drinking coffee has become a trend that occurs in all circles of society in Indonesia, especially in Jakarta. Coffee is not only seen as a stress reliever, but has also become a hereditary culture. However, the culture of drinking coffee has been around for a long time and is not something that has just emerged in Indonesia recently. It is also a part of everyday life that attracts many entrepreneurs to open their own coffee shop businesses, because the prospects are good. However, before opening a new business, it is necessary to know the strategic place where you want to open the business.

Business Problem :

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Cafe in Jakarta, Indonesia. By using data science methods and machine learning techniques such as clustering, this project aims to provide solutions to answer the business question: In the city of Jakarta, if an enterpreneur wants to open a Cafe, where should they consider opening it?

Target Audience :

The enterpreneur who wants to find the location to open new Cafe.

Data Description:

The data set that I have used for solving the problem is:

- A complete list of neighborhoods in Jakarta, Indonesia. Source of the data is Wikipedia.org
- Geographical coordinates (latitude and longitude) of those neighborhoods. Source of the data will be FourSquare.
- FourSquare provided Venue data which is related to Cafe. We will use this data to perform clustering on the neigbourhoods.

Data Sources

This wikipedia page https://en.wikipedia.org/wiki/Central_Jakarta contains a list of neighborhoods in Jakarta, with a total of 44 neighborhoods. 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 neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbors. After that, we will use the Foursquare API to get the venue data for those neighborhoods.

Now that we know the data we need, we use the Foursquare API to get the venue data for the environment. Foursquare is one of the largest databases with 105+ million places and is used by more than 125,000 developers. Foursquare provides many categories of venue data, and what I use here is the cafe venue data. This is a project that will take advantage of many data science skills, from web scraping (Wikipedia), working with APIs (Foursquare), data cleaning, data disputing, 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 we performed and the machine learning techniques used.

Methodology

First, we need to get a list of neighborhoods in Jakarta. The list is available on the Wikipedia page https://en.wikipedia.org/wiki/Central_Jakarta. We will do web scraping using the Python programming language and the beautifulsoup library to extract the desired data. However, this is just a list of names, so we need to know or get geographic coordinates in the form of latitude and longitude available in the Foursquare API. To do this, we need the Geocoder library which is useful for converting addresses to geographic coordinates in latitude and longitude form. After the data has been collected, we will save the data into the Pandas DataFrame and then visualize the data in a map using the Folium library. This allows us to carry out checks to ensure that the geographic coordinate data displayed by the Geocoder is visualized correctly in the city of Jakarta.

Next, we will utilize the Foursquare API to obtain the top 100 venues within a 2000 meter radius. For that, we first need to register a Foursquare Developer Account to get a Foursquare ID and Foursquare secret key. Then we'll make an API call to Foursquare that passes the geographic coordinates of the environment in a Python loop. Foursquare will return the place data in JSON format and we will extract the place name, place category, latitude and longitude of the place. With that data, we can check how many places were returned for each environment and check how many unique categories can be curated of all the returned places. Then, we will analyze each environment by grouping the rows by environment and taking the mean of the frequency at which each place category occurs. That way, we also prepare data for use in clustering. Since we are analyzing "Cafe" data, we will filter "Cafe" as the place category for neighborhoods.

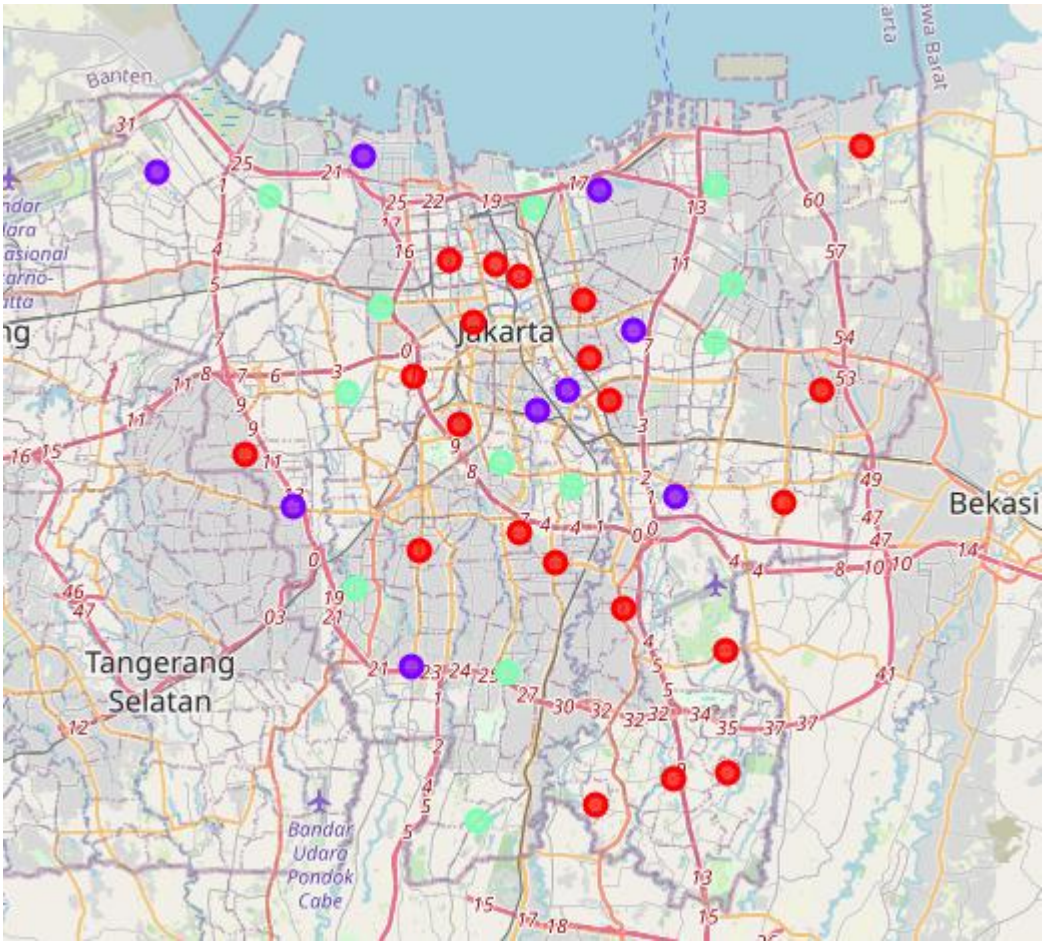
Finally, we will cluster the data using k-means clustering. The K-mean algorithm identifies the k number of centroids, and then allocates each data point to the closest cluster, keeping the centroids as small as possible. It is one of the simplest and most popular unsupervised machine learning algorithms and is quite suitable for solving the problem in this project. We will group the neighborhoods into 3 clusters based on the frequency at which they appear for the "Cafe" venue category. The results will make it easier for us to identify which areas have the most cafes and which areas have the least. Based on the presence and density of cafes in the area, it will help us answer the question which area is most suitable for opening a new cafe.

Results

The cluster results by the k-means algorithm show that regions can be categorized into 3 clusters based on the frequency of occurrence of "Cafe":

- Cluster 0: Areas with the least number of cafe frequencies
- Cluster 1: Areas with a medium number of cafe frequencies
- Cluster 2: Areas with the most number of cafe frequencies

The cluster results are visualized on the map below, with cluster 0 colored red, cluster 1 purple, and cluster 2 mint green.



Discussion

As shown from the map in the Results section, the distribution of cafes in Jakarta is not centralized but scattered in each region, with the highest number of distribution in cluster 2 and the number being in cluster 1. Whereas for cluster 0 it has a very low number or cafe in the environment can still be said to be insufficient. This is a great opportunity and high potential to open a new cafe because there is very little or no competition. Meanwhile, the cafes in cluster 2 tend to have fierce competition because even though they are not very crowded. On the other hand, the results also show that the existing cafes are not centralized so it is still possible to open cafe businesses in any cluster. But even so, it will be a step to open a new cafe in cluster 0, because there is clearly minimal competition.

Limitation and Suggestion for Future Research

In this project we only consider one factor, namely the frequency of cafe appearances, other factors such as population and income as well as people's interest in hanging out which can influence the decision to locate a new cafe. So that further research can design a methodology to use this data to be used in a grouping algorithm to determine a good location to open a new cafe.

Conclusion

In this project, we have gone through the process of identifying business problems, determining the required data, extracting and preparing data, conducting machine learning by grouping data into 3 clusters based on their similarities, and finally providing recommendations to relevant stakeholders. namely entrepreneurs regarding the best location to open a new cafe. To answer the business questions that arise in the introduction, the answers proposed by this project are: The environment in cluster 0 is the best location to open a new cafe. It is hoped that the findings from this project will help relevant stakeholders to take advantage of opportunities in high potential locations while avoiding overcrowded areas in their decisions to open new cafes.

References

List of neighborhoods in Jakarta : https://en.wikipedia.org/wiki/Central_Jakarta
Foursquare Developer Documentation : https://developer.foursquare.com/docs
Dhisasmito, P. P. (2020, February 4).Understanding customer loyalty in the coffee shop industry (A survey in Jakarta, Indonesia).British Food Journal