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 Shopping Mall. 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_2.csv file (http://data.jakarta.go.id/dataset/jumlahkecamatankelurahanrtrwdankkdkijakarta/resource/1d5b0b b0-3aa7-482a-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 Shopping Malls 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).