IBM capstone Project Course Week 4 Assignment 2

2020/05/05 Christianus Frederick Hotama

Project Title: Predict the Best District for the House Investment Based on the infrastructure facility by using the Regression Technique

Data:

• List of districts in South Korea (in Korea, it is called -gu) the district. In this study, we will include every district in every city in South Korea

• Latitude and longitude coordinates of those districts. This is required to plot the map and also to get the venue data.

• Infrastructure data for each district in the last 10 years, particularly data related to the subway station, bus stop, bus terminal, hospital, restaurant, supermarket, and traditional market. We believe the existence of those facilities may increase the value of the property. We sample this information from 1 January 2010 until 1 January 2020 to see the infrastructure for each district in the last 10 years. Later, we will use this data to perform the data prediction about which districts have more development compared to others in the last 10 years.

Methodology:

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_districts_in_South_Korea) contains a list of districts in South Korea, with a total of 131 districts from 22 largest cities in Korea. 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 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 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 Public Infrastructure 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, machine learning (regression), and map visualization (Folium).