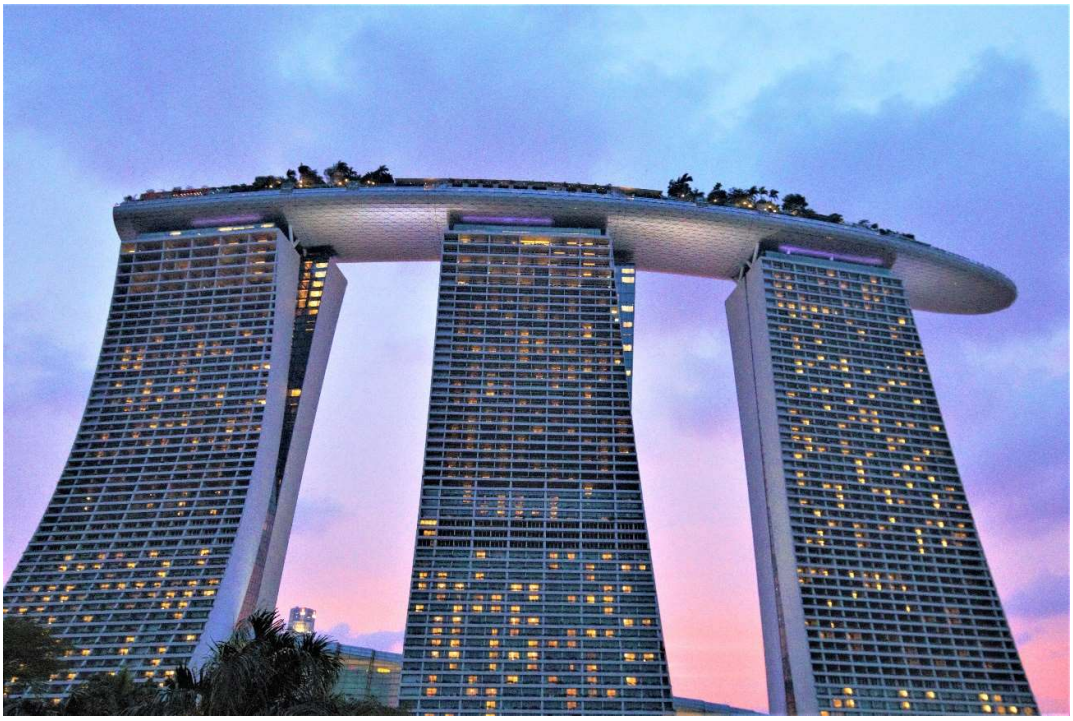
IBM Applied Data Science Project
Coursera Capstone

Potential Places to Establish Hospitals in Singapore



Vincentius Indra Sulaiman

March 2020

Data Requirements, Collection, and Preparation

Data Requirements

To fulfil the objectives of the current study, publicly available data is of upmost importance to provide valuable insights. As a whole, the required data are as follows:

1. List of neighbourhoods (towns) in Singapore. This dictates the scope of the present study which is the country of Singapore.
2. Location coordinates (latitude and longitude) of the neighbourhoods. This will be important to map the data.
3. Population of Singaporean residents in the neighbourhoods. This will be used to cluster Singapore towns.
4. Location of large hospitals and medical centres in Singapore. This will be used to cluster Singapore towns. Hospital size is determined by the presence of wards.

Methods of Data Collection

Based on the aforementioned required data, the following methods will be employed to collect them:

1. Scraping: https://en.wikipedia.org/wiki/New_towns_of_Singapore to obtain list of neighbourhoods with the total population and residential areas.
2. Utilising the geopy library to obtain geographical coordinates of Singapore towns.
3. Accessing Foursquare data through API to obtain crucial data on the locations of hospitals and medical centres in Singapore. Hospital wards fall under the foursquare category id: 58daa1558bbb0b01f18ec1f7.

Preparation of Data

After the data collection step has been done, the following measures will be undertaken to process the raw data:

1. Cleaning: to remove unnecessary parts of raw data. Cleaning is an iterative process: it is done in parts across the whole process, not only in the beginning section of the project.
    a. Removing unnecessary columns from wikipedia table: name Chinese, Romanised Chinese name in Pinyin, name in Tamil, number of dwelling units, and projected number of ultimate dwelling units.

    b. Modifying problematic data: town of Yishun that does not return geographical coordinates. The coordinates are searched on google and appended manually on the dataframe.
    c. Renaming column names to ease analysis step.
    d. Removing unnecessary duplicate columns that arise from hospital ward counting step.
    e. Removing unnecessary columns that may be a hindrance to training algorithm.
    f. Removing unnecessary data columns when analysing clusters.
2. Creating dataframes: to further simplify the processing step especially after extracting Foursquare data.
3. Scaling the data using MinMaxScaler from Sklearn to provide better machine learning algorithm.
4. Merging various information to selected dataframes: to provide the most useful information in the least amount of dataframes to further ease the analysis step.

Analysis of The Problem

The prepared data will be analysed in accordance with the following steps:

1. Data manipulation: arithmetic calculation of population density by dividing the total town population with town area.
2. Applying machine learning method: clustering. This is done by first initiating the KMeans training model from Sklearn, finding the optimal k-number for clustering the towns based on population density and amount of hospitals.
3. Mapping the data: hospital locations and the clusters in a Folium map.

By the implementation of clustering method of machine learning, the analysis step is expected to provide sufficient information and description on the clusters of Singapore towns. These findings will be the basis for the readers to further comprehend the healthcare economic opportunity in Singapore.

References:

- https://www.statista.com/statistics/378424/average-age-of-the-population-in-singapore/
- https://www.statista.com/statistics/378566/age-structure-in-singapore/
- https://hospitals.webometrics.info/en/Asia_Pacifico/South%20East%20Asia
- https://mothership.sg/2016/10/nifty-map-shows-which-parts-of-spore-has-densest-population-guess-which-is-no-1/
- https://en.wikipedia.org/wiki/New_towns_of_Singapore