IBM Applied Data Science Capstone: Singapore

By: Tan Jing Jie

Introduction:

Singapore has some of the best infrastructure facilities in South East Asia. Making it one of the most attractive places for people to set up businesses or start a company. Individuals who are interested in living here are also motivated by the various job opportunities available here.

However, it is almost impossible to ensure that a city is saturated with world-class infrastructure facilities. In this project, I will analyse Singapore infrastructure and try to draw meaningful conclusions from it.

Business Problem:

1. List and visualise all major areas of Singapore with top existing infrastructure.
2. What are the best areas in Singapore as per infrastructure?
3. Which areas have the potential for the development of infrastructure of various kinds?
4. Which areas lack infrastructure facilities?
5. Where is the best place to stay within Singapore for all vital infrastructure facilities?

Data Description:

For this project we need the following data:

- Data containing postal codes of various areas in Singapore, along with their latitude and longitude.
    - Data source: data.gov.sg
    - Description: With the postal codes of areas in Singapore, we can convert it into GPS Coordinates.
- Data on the different types of infrastructure in the different areas of Singapore.
    - Data source: Foursquare API
    - Description: With the API we can get information on the venues in the area.

Methodology:

- Data Gathering
  We need the list of areas in Singapore with geographical coordinates so that we can make use of the Foursquare API. Fortunately, we can obtain a list of government offices in different areas of Singapore from https://data.gov.sg/dataset/sgo-satellite-offices . Using Google Maps, we can obtain the geographical coordinates from postal codes.

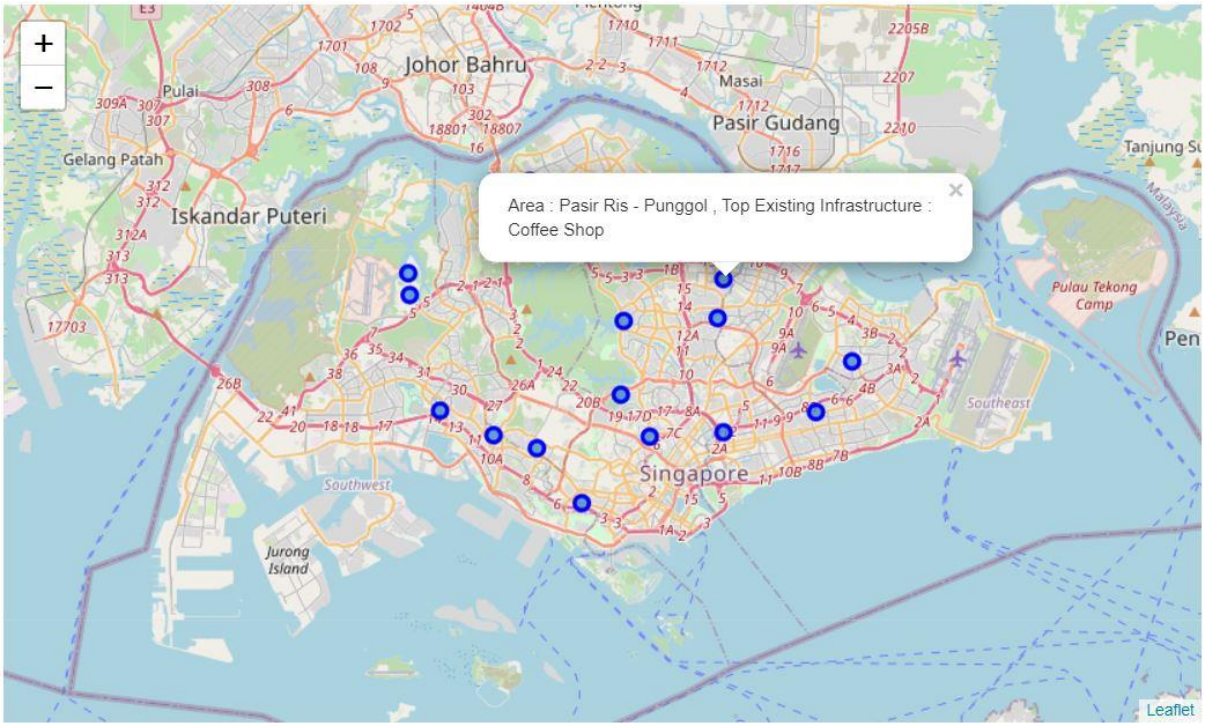
- Data Visualization
  Visualize the areas in a map using Folium package. This allows us to perform a check to make sure that the geographical coordinates are accurate.

- Data Clustering
  We will perform clustering on the data using k-means clustering. K-means clustering identifies k number of centroids and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. We will cluster the areas into 3 clusters based on the number of existing infrastructures in that area. The results will allow us

to identify which areas have highest, medium and lowest concentration of infrastructures. This will help us answer the questions mentioned in the Business Problem.

Results:

1. Display top existing infrastructure for each area in Singapore.



2. What are the best locations in Singapore as per infrastructure?

Gym / Fitness Center	0
Hotel	3
Movie Theater	0
Park	0
Pharmacy	1
Playground	1
Restaurant	2
Shopping Mall	0
Supermarket	2
Theater	0
Total infrastructure	16

	6
Area	Jalan Besar
Postal Code	328127
City	Singapore
Bus Station	1
Business Service	0
Café	5
Electronics Store	0
Farmers Market	0
Garden	0
Gym	1

3. Which of your choice areas have the potential for the development of infrastructure of various kinds?

```
These are infrastructures with highest potential in Jalan Besar area :
Business Service
Electronics Store
Farmers Market
Garden
Gym / Fitness Center
Movie Theater
Park
Shopping Mall
Theater
```

4. Which areas lack infrastructure facilities?

	Area	Postal Code	Total infrastructure
9	Marsiling - Yew Tee	689286	1
10	Nee Soon	760234	1

5. Where is the best place to stay within Singapore for all vital infrastructure facilities?

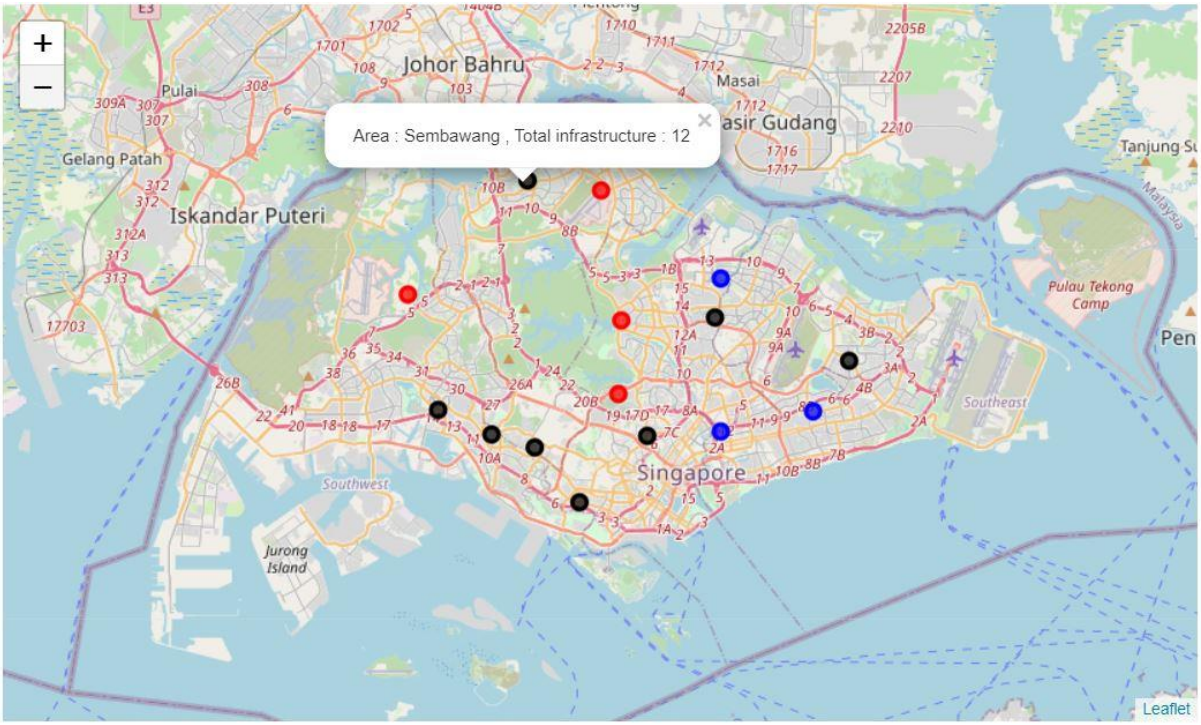
	Area	Total infrastructure
6	Jalan Besar	14

Clustering Based on Total Vital Infrastructure:

The results from the k-means clustering show that we can categorize the areas into 3 clusters based on the number of vital infrastructure facilities.

- Cluster 1: Areas with a low number of vital infrastructures.
- Cluster 2: Areas with a moderate number of vital infrastructures.
- Cluster 0: Areas with a high number of vital infrastructures.

The results of the clustering a visualized in the map below with cluster 0 in black, cluster 1 in red and cluster 2 in blue.



Discussion:

- Jalan Besar is the best location in Singapore as per infrastructure
- Most of the areas with high number of infrastructures are located in the South of Singapore.
- This could be due to the more developed and “matured” areas located in the South.
- If a company is interested in building new vital infrastructures, they may look for areas in the North and West of Singapore.
- If an individual wants to stay in areas with relatively more infrastructures, he might consider areas in the South of Singapore.

Conclusion:

In this project, I have gone through the process of identifying business problems, data extraction, visualizing of data and performing machine learning on the data. With the results, we are able to provide solutions to the business problems. Business developers are able to identify suitable areas to build new infrastructure. Visitors and immigrants can obtain insights on the areas of Singapore that may bring about a higher quality of life.