Capstone Project

Battle of Neighborhoods (Week 2 Report)

By: Dennis Lam

INTRODUCTION

Sandakan is located in state of Sabah, East Malaysia. It is the second largest town. The town was founded in 1878 and today has approximate 160000 residents staying in there.

This Capstone project will try to find out where and which suburbs neighbourhood is suitable for setting up business.

BUSINESS PROBLEM

As a town, the people here are mainly mid-income to low income earners, hence small type of businesses are more suitable instead of building large shopping malls.

We shall do an analysis of Sandakan neighborhoods and present the report which will consists of methodology, findings and conclusions.

Any person who wished to setup any new business can refer to this report which will be published in later period.

DATA

The neighborhoods data is found in Sandakan Municipal website: http://www.mps.sabah.gov.my/isandakan.cfm#penduduk

The focus will be on these neighbourhoods (they are also called housing estates) around Sandakan. These people who stayed therein will be the main customer target for new businesses.

We will be using some Python libraries like pandas, geopy, folium, scikit-learn etc to explore and cluster these neighbourhoods to find out which one is suitable.

Foursquare APIs will be used to find out any interesting venues like food, offices, entertainment etc.

Google maps and Geopy library will be used to find latitude and longtitude for each neighbourhood as geolocation data is not available online.

METHODOLOGY

The project starts with data gathering from various sources like websites to create a csv file for data analysis.

Pandas was used to create the dataframe and several basic analysis is done.

Below are several visualizations regarding the data:

(a) Sandakan neighbourhoods by area size

(b) Sandakan neighbourhoods by residential units

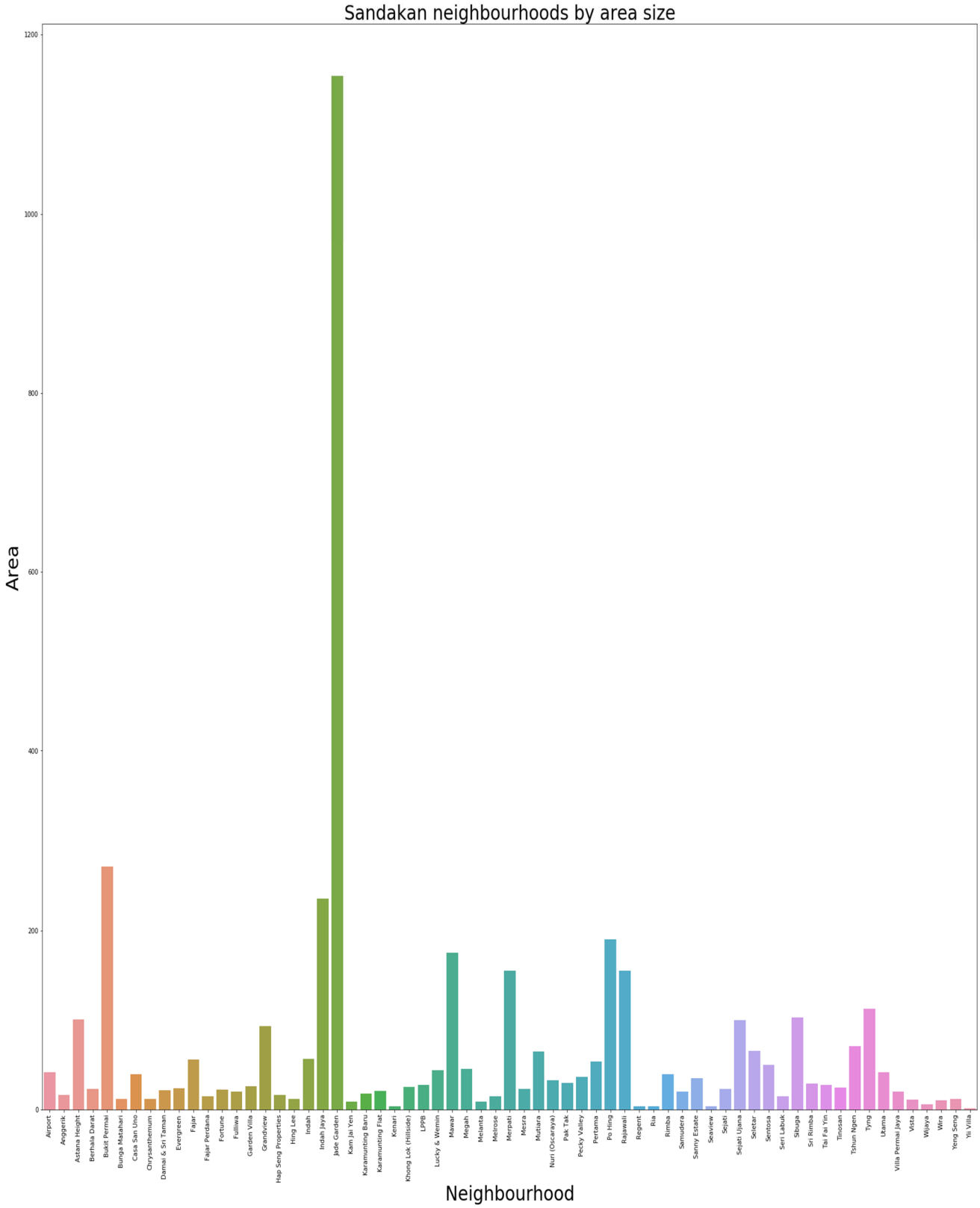
(c) Heatmap to illustrate Correlation

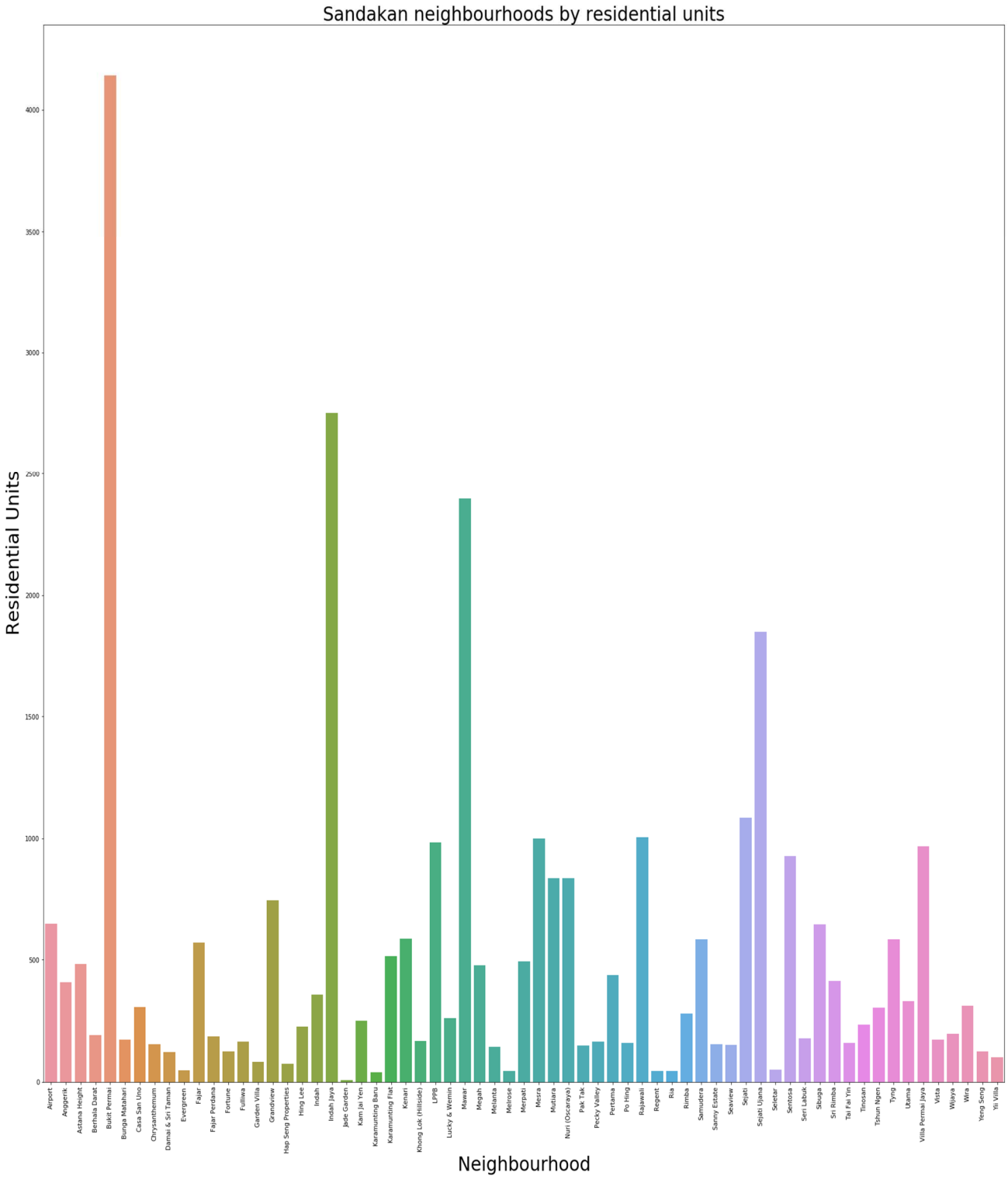
The findings are:

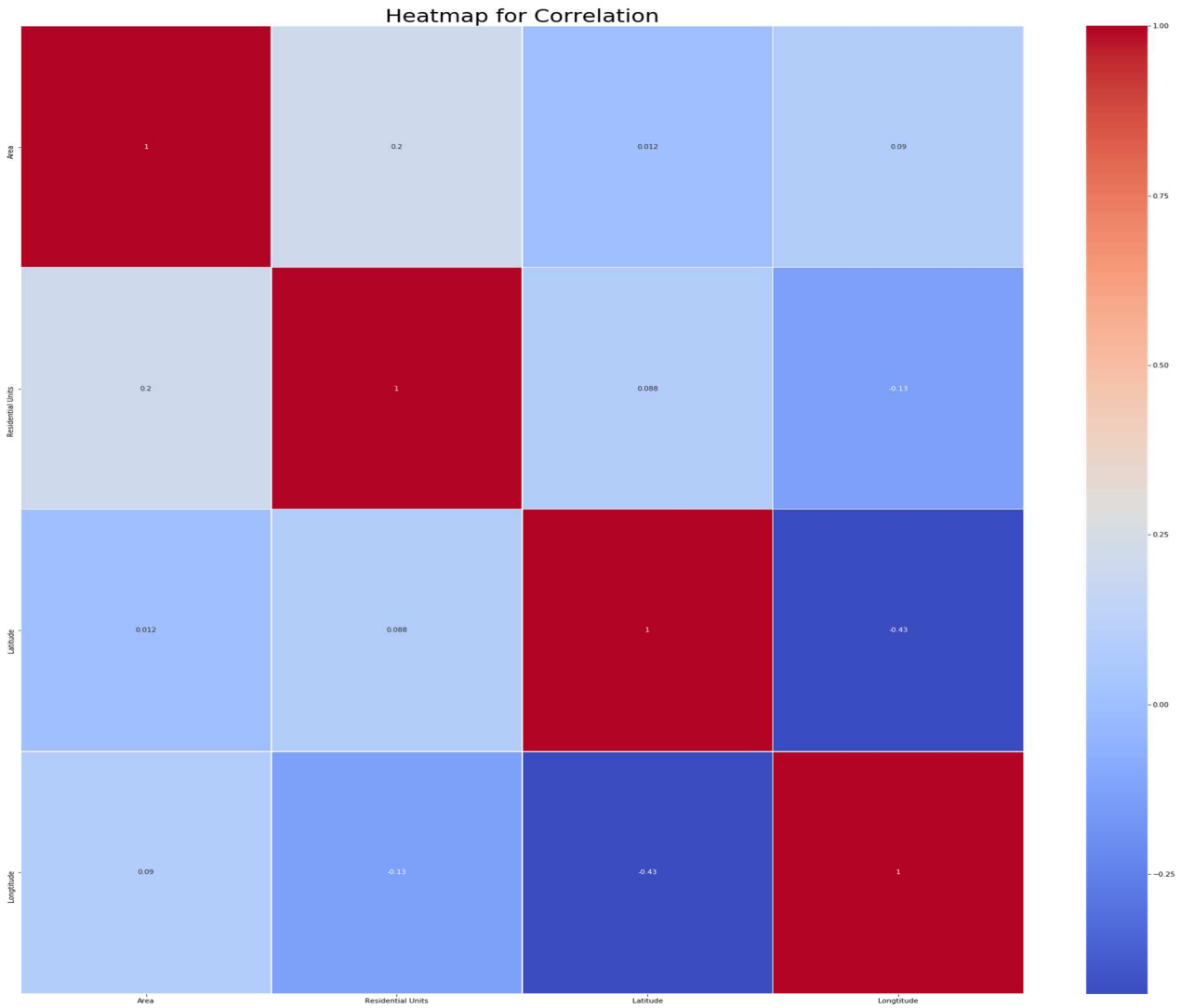
(a) Jade Garden has largest area, 1154 acres

(b) Bukit Permai has most number of residential units, 4142 house units

(c) A small correlation of 0.2 between Area and Residential Units


Sandakan neighbourhoods by area size
Area
Neighbourhood
0
200
400
600
800
1000
1200
Airport
Anggerik
Astana Height
Berhala Darat
Bukit Permai
Bunga Matahari
Casa San Uno
Chrysanthemum
Damai & Sri Taman
Evergreen
Fajar
Fajar Perdana
Fortune
Fulliwa
Garden Villa
Grandview
Hap Seng Properties
Hing Lee
Indah
Indah Jaya
Jade Garden
Kan Jai Yen
Karamunting Baru
Karamunting Flat
Kenari
Khong Lok (Hillside)
LPPB
Lucky & Wemin
Mawar
Megah
Melanta
Melrose
Merpati
Mesra
Mutiara
Nuri (Oscaraya)
Pak Tak
Pecky Valley
Pertama
Po Hing
Rajawali
Regent
Ria
Rimba
Samudera
Sanny Estate
Seaview
Sejati
Sejati Ujana
Seletar
Sentosa
Seri Labuk
Sibuga
Sri Rimba
Tai Fai Yin
Tinosan
Tshun Ngen
Tyng
Utama
Villa Permai Jaya
Vista
Wijaya
Wira
Yeng Seng
Yi Villa

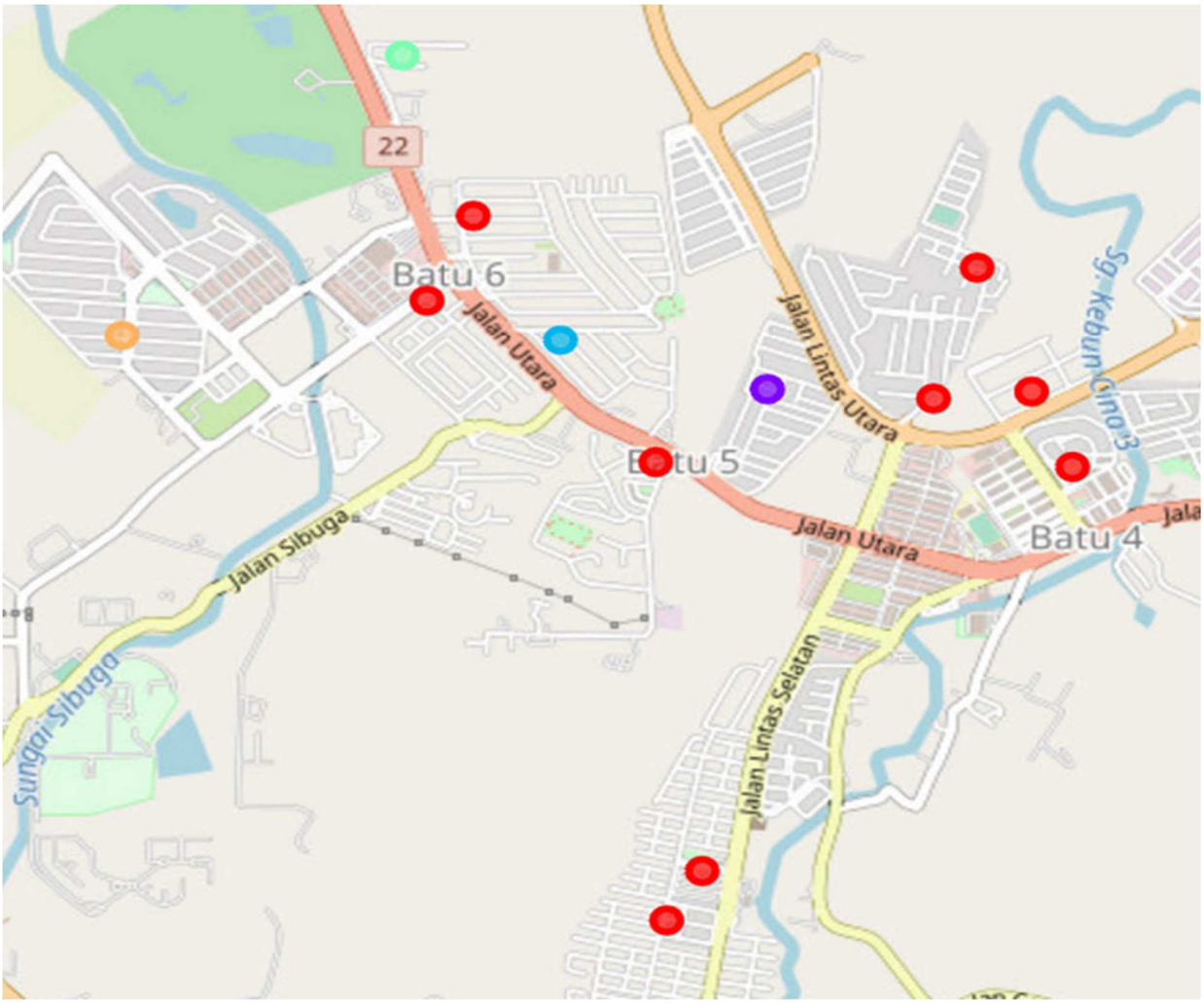

Sandakan neighbourhoods by residential units
Residential Units
Neighbourhood
0
500
1000
1500
2000
2500
3000
3500
4000
Airport
Anggerik
Astana Height
Berhala Darat
Bukit Permai
Bunga Matahari
Casa San Uno
Chrysanthemum
Damai & Sri Teman
Evergreen
Fajar
Fajar Perdana
Fortune
Fulliva
Garden Villa
Grandview
Hap Seng Properties
Hing Lee
Indah
Indah Jaya
Jade Garden
Kam Jai Yen
Karamunting Baru
Karamunting Flat
Kenari
Khong Lok (Hillside)
LPPB
Lucky & Wemin
Mawar
Megah
Melanta
Melrose
Merpati
Mesra
Mutiara
Nuri (Oscaraya)
Pak Tak
Pecky Valley
Pertama
Po Hing
Rajawali
Regent
Ria
Rimba
Samudera
Sanny Estate
Seaview
Sejati
Sejati Ujana
Selatar
Sentosa
Seri Labuk
Sibuga
Sri Rimba
Tai Fai Yin
Timoan
Tshun Ngen
Tyng
Utama
Villa Permai Jaya
Vista
Wijaya
Wira
Yong Seng
Yi Villa



The area of interest will be from mile 4 to mile 6 since majority of neighbourhoods concentrated there. Using folium package, we map the exact locations. The Foursquare to get common venues mentioned by users.

The information from csv file and foursquare venues are combined and arranged by neighbourhood. A new dataframe is created with top 10 venues for each neighbourhood.

Using K-means clustering to segment them up to five clusters resulting in the following image in next page:



The results will help us to identity which neighbourhood clusters are more concentrated as a unit.

RESULTS

Cluster 0: Has a total of 9 neighbourhoods

Cluster 1: Has 1 neighbourhood

Cluster 2: Has 1 neighbourhood

Cluster 3: Has 1 neighbourhood

Cluster 4: Has 1 neighbourhood

The Jupyter notebook will provide more details in a separate file.

DISCUSSION

From all the results, cluster 0 will be the most desirable place to set up businesses. It is recommended for new entrants to check the places as these areas are competitive and requires new ideas of product and services to thrive.

Cluster 4 has good potential as there will be new shop buildings constructed to house new businesses. With proper planning and marketing, businesses can pull in customers from other nearby clusters.

The remaining clusters will be low to poor visibility due to stand alone neighbourhoods.

In future, this study can be extended to miles 7 and 8 as there are significant neighbourhoods. Businesses can also consider these areas for expansion.

CONCLUSION

In short, we have identify a business problem, collected and prepared relevant data, performed machine learning to cluster the neighbourhoods and finally provided recommendations to business owners who wish to set up new or expand their businesses. Cluster 0 is the favourite one followed by cluster 4.

It is hoped that this report can enlighten and help business owners.