



Data Science Capstone Project 2020

The 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 ranges from mid-income to low income earners, hence small 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, results and conclusions.

The stakeholders are those who wish to setup any new business can refer to this report.

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. Potential for business setups will be within housing estates and areas surrounding it.

We will be using some Python libraries like pandas, geopy, folium, scikit-learn 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 longitude 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 library was used to create the data frame and several basic analysis is done.

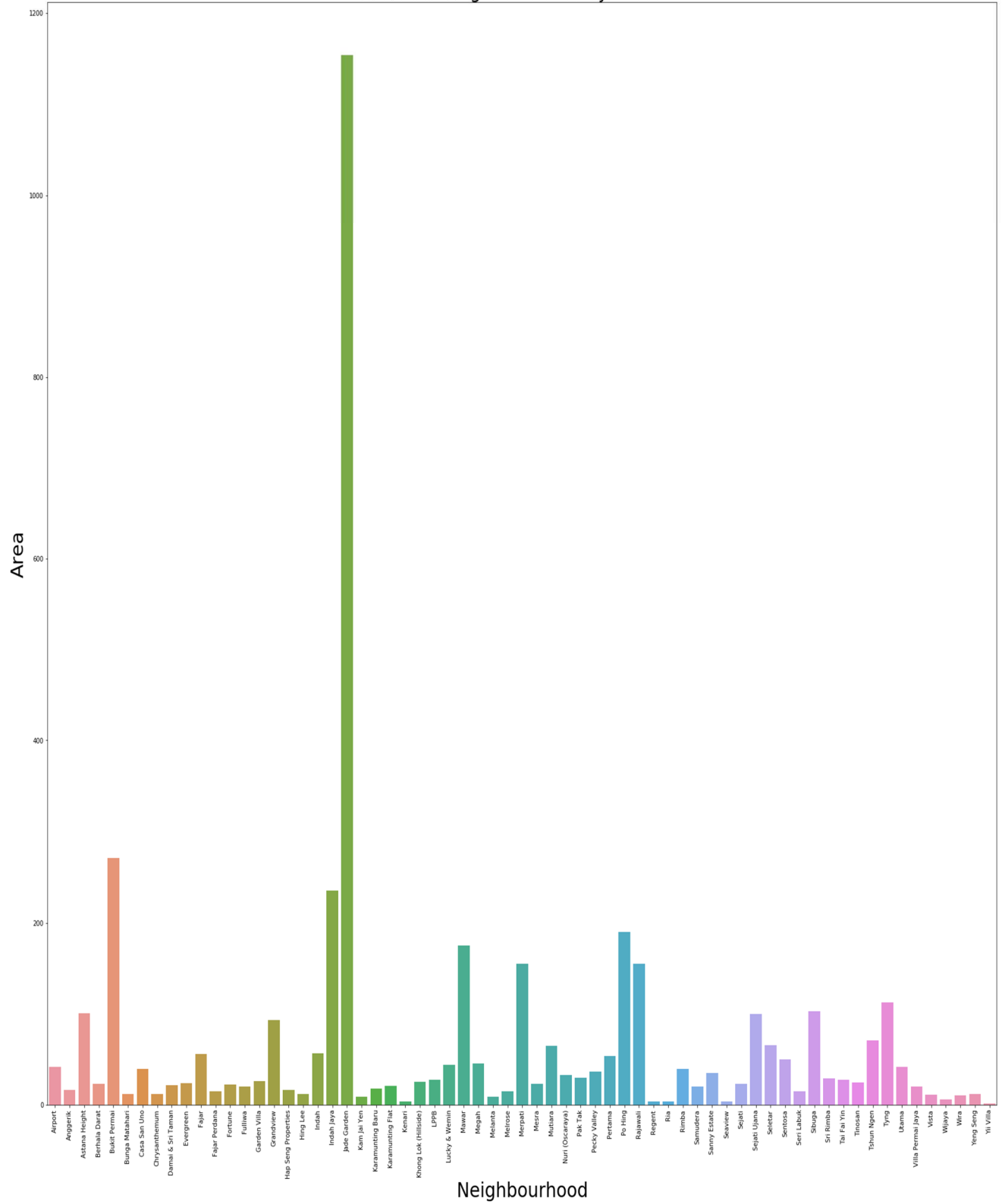
Below are several visualizations performed regarding the data:

- (a) Sandakan neighbourhoods by area size
- (b) Sandakan neighbourhoods by residential units
- (c) Heatmap to illustrate Correlation

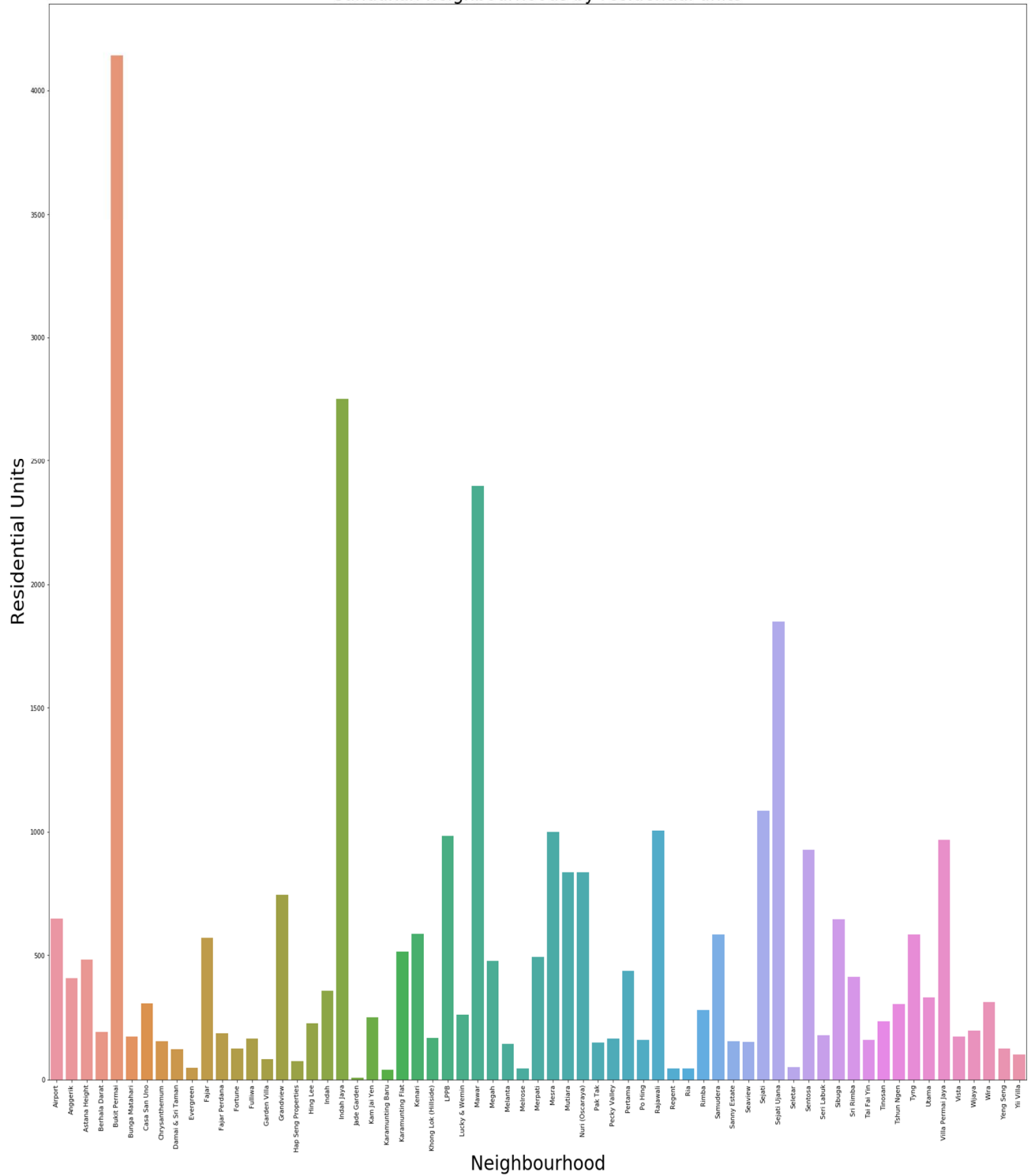
The findings are:

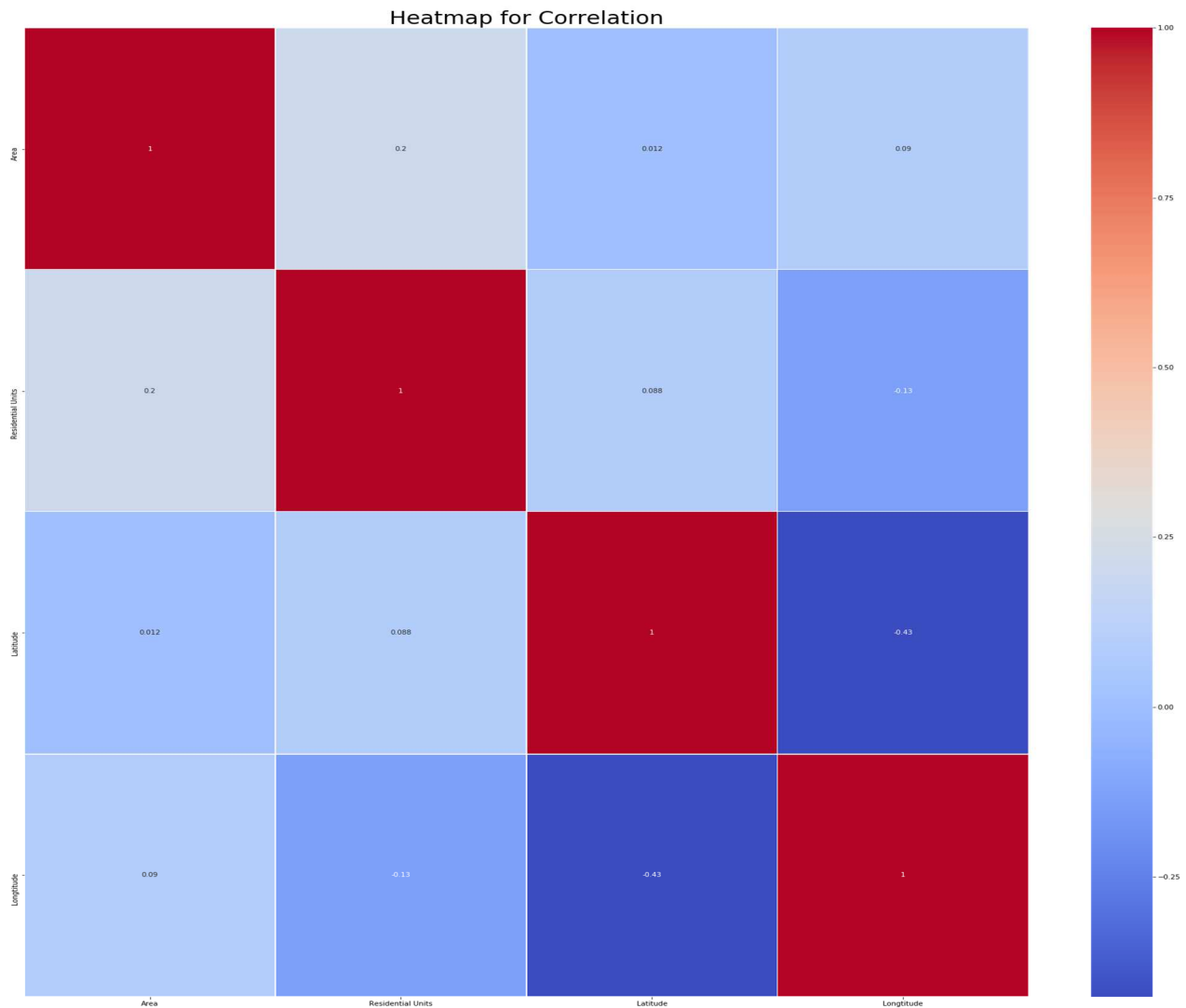
- (a) Jade Garden has largest area, 1154 acres
- (b) Bukit Permai has most number of residential units, 4142 house units
- (c) A small correlation of 0.2 between Area and Residential Units

Sandakan neighbourhoods by area size



Sandakan neighbourhoods by residential units





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

The information from csv file and foursquare venues are combined and arranged by neighbourhood. A new data frame 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 Jupyter notebook file is hosted in GitHub page.

DISCUSSION

From all the results, cluster 1 will be the most desirable place to set up businesses. It is recommended for new startups 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 identified 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 one or expand their businesses. Cluster 1 is the favorite one followed by cluster 4.

This report will enlighten and help business startups in Sandakan.