IBM Capstone Project – Singapore Spa Analysis

By Warren Kearney

Table of contents

1. Introduction
2. Data
3. Methdologies
    a. Exploratory data analysis
    b. Inferential statistical testing
    c. Machine learning
4. Results
5. Discussion
6. Conclusion

1. Introduction

1.1. Business problem

My wife wants to open a spa business in Singapore, but is not sure what location would make most sense

Singapore very saturated commercially, has very dense population (5.6 million people on an island 50km East to West, and 27km North to South), and property is relatively expensive.

The question this exercise will look to answer is “where in Singapore represents a good opportunity to open a spa, either because of demand, supply, or cost”.

I would examine the current spa population island-wide, and assess patterns for factors that could either influence or explain presence or absence, and understand where a gap in the market might exist

The output of the analysis could also then be applied to other business types, and be relevant to any potential investor looking for statistical analysis to support their investment decision

1.2. Background discussion

The locations of spas in Singapore could be based on many factors: public transport links (train, bus), private transport links (motorways nearby, car parking availability), complimentary establishments nearby (e.g. restaurants, hairdressers), affordability of

commercial property leases, availability of commercial property space, the nearby target demographic population – either residential or workplaces.

It could also be the number of nearby competitors – is the market saturated already?

Is the type of area largely residential, or commercial? Are there malls or old-fashioned shop houses? Each would have a different dynamic.
The type of spa could be analysed based on segments, from budget to premium.

2. Data

2.1. Acquisition

First data source is Singapore location data, including neighbourhoods and coordinates. I used the government website *data.gov.sg* which had csv format files with coordinates, plus GeoSpatial data for a choloropleth map.

For spas plus nearby location data, I used Foursquare as this is the source covered by the IBM coursework, and connected via their API to obtain JSON files

For Singapore population and income data I used the government website *data.gov.sg* and connected via their API to obtain a JSON file

Note: I downloaded further csv format data on housing types, sex, age group, but would propose to use this for further analysis ("Resident Households by Monthly Household Income from Work, Ethnic Group and Sex of Head of Household, 2015")

2.2. Data cleaning

Columns needed to be removed from tables downloaded that weren't necessary for analysis.

The Geographic Map data format, downloaded from sg.gov. sg website for the chloropleth map needed to be changed from Shapefile to GeoJSON using "QGIS" application with the correct coordinate formatting.

Districts in the coordinate file needed to be made upper case to align to GeoJSON file district data

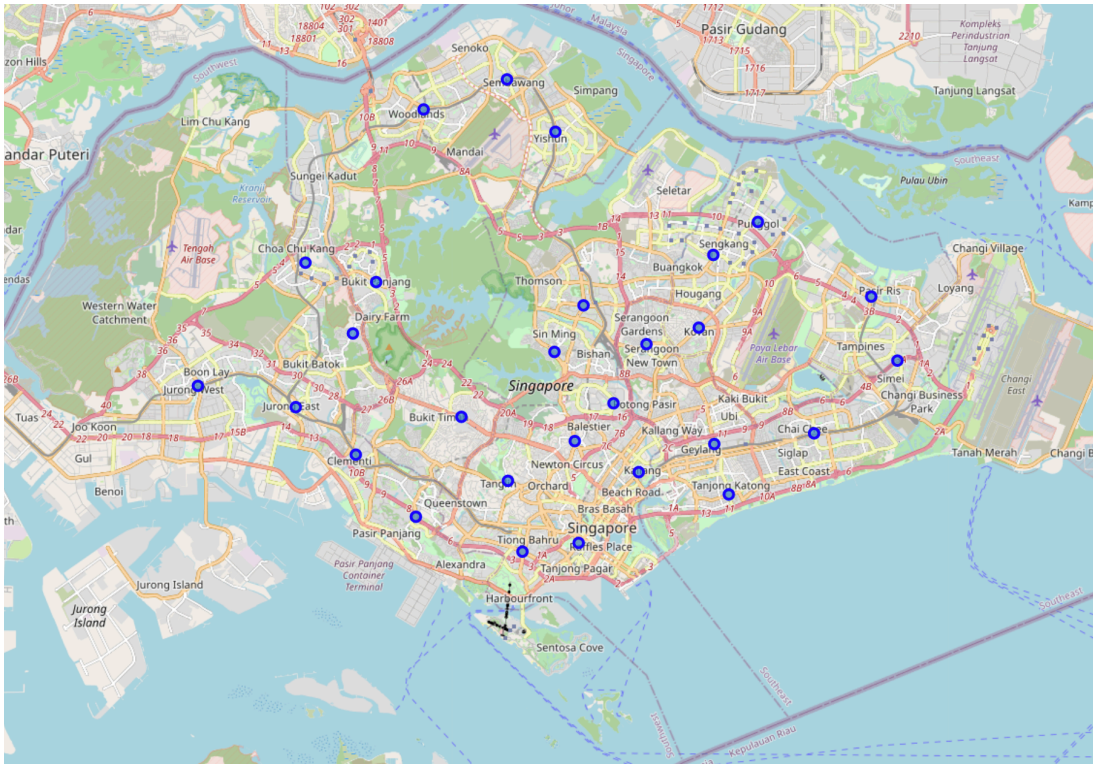
For the household income data, median values needed to be estimated for each category reported (mid-point of each group, e.g. for 1,000 – 2,000 SGD category, 1,500 was entered, except for the top category of 12,000 and above, for which 12,500 was used), then weighted averages calculated and totals removed from the data table

Other more standard cleaning tasks were renaming columns, values limited to top 10, tables sorted in ascending or descending orders for lists or graphs, tables combined for analysis and grouped with counts or aggregation, or filtered for e.g. Spas.

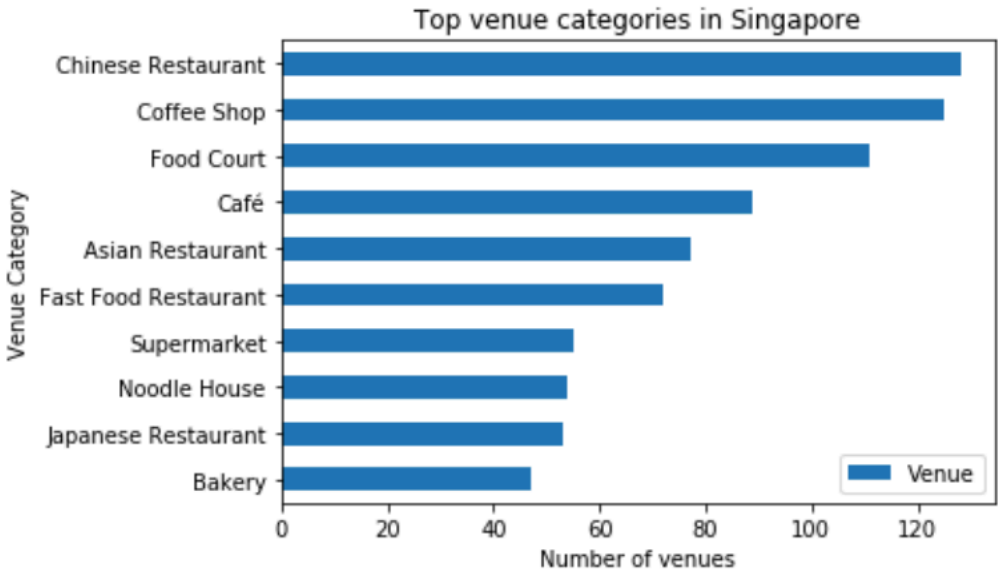
Methodologies

2.3. Exploratory Data analysis

First to examine what the government assigned standard Singapore 'districts' are – in this case called 'planning areas' – there are 29 of them.



With the data downloaded from FourSquare, the most popular venue categories are explored – mostly Food & Beverage establishments



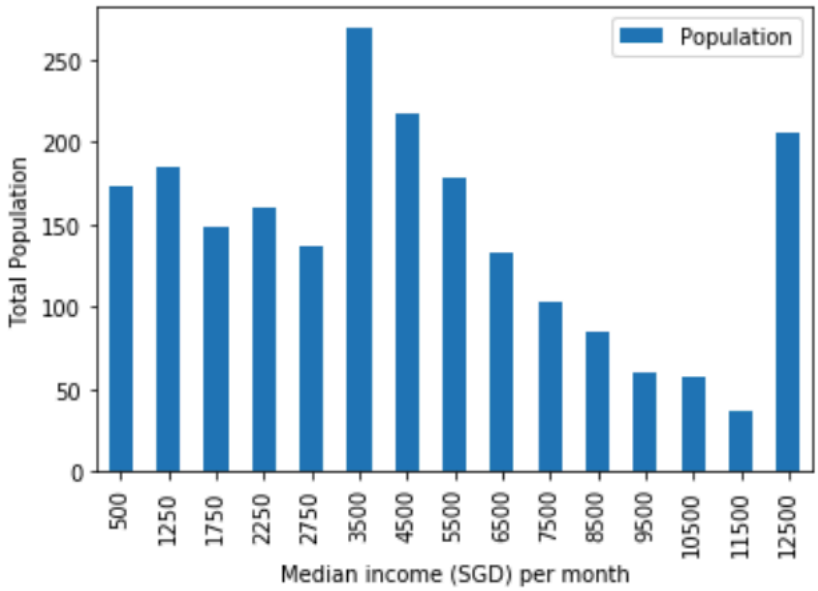
Next to examine the Spa distribution, given the subject of the exercise. Only 13 were reported by Foursquare – much less than expected.

Count, and the locations is as below

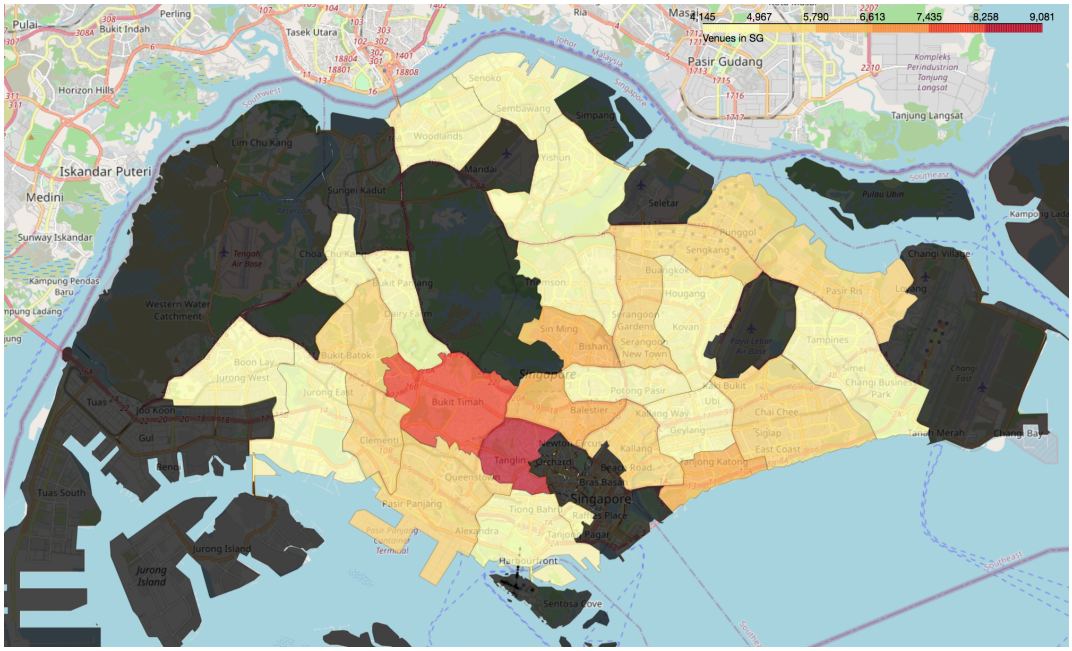
	Venue
District	
Bishan	3
Tanglin	3
Outram	2
Ang Mo Kio	1
Bukit Timah	1
Kallang	1
Sengkang	1

Next point to explore is the wealth distribution in Singapore – what the spread of monthly income is per person, from the data.gov.sg website.

It follows a bell curve shape, so ‘middle income’ most common, with a spike for 12,500, given there was no detail over 12,000 SGD.



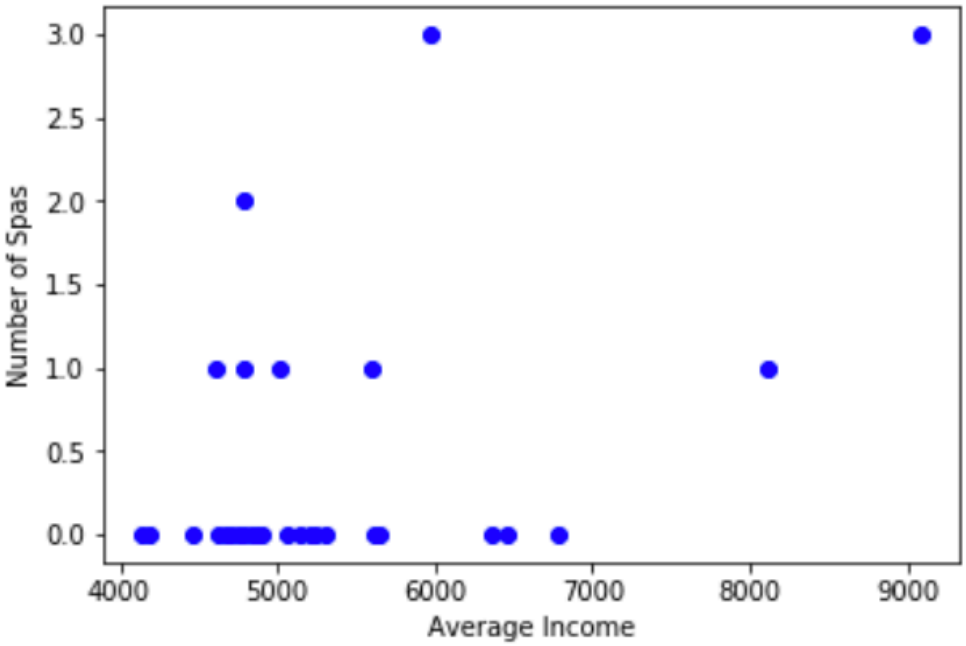
The average median income was calculated for each district, and plotted on a choropleth map as below – it shows clearly the higher income is in the centre of the island in Bukit Timah and Tanglin areas



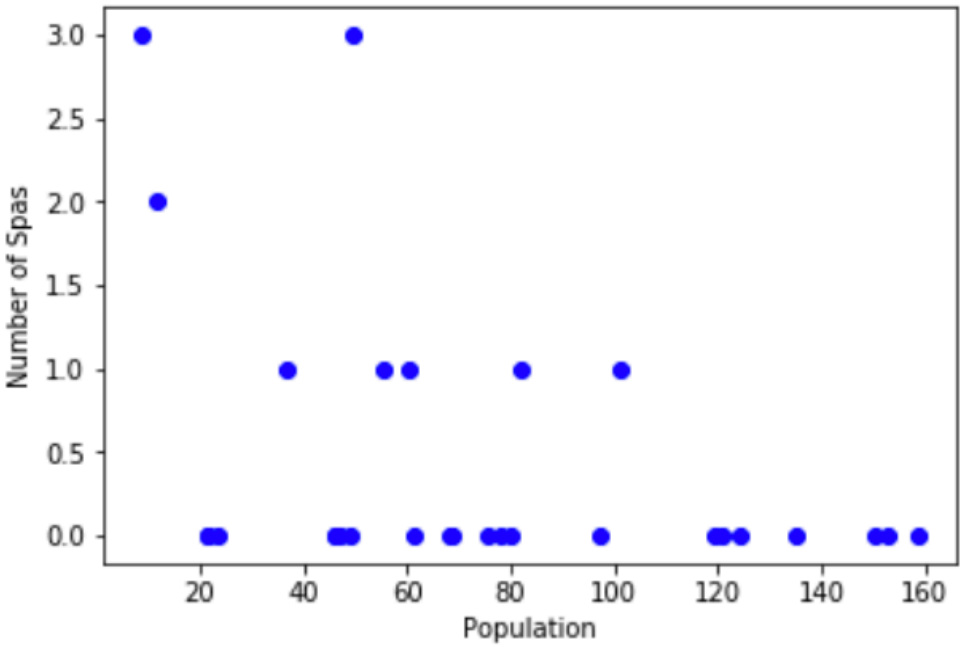
The black areas are largely commercial, park, or military zones, so match expectations

2.4. Inferential Statistical Testing

Firstly, with the idea of selecting targets for any linear regression models, I explore the relationship between number of spas and average income, but there is little visual pattern seen below



Next the relationship between number of spas and population per district. This does show some correlation at least, with higher number of spas seen in lesser populated areas.



3.3. Machine Learning

Firstly I select linear regression model to project the number of spas in a district based on population, given the observed pattern in the above statistical testing tasks.

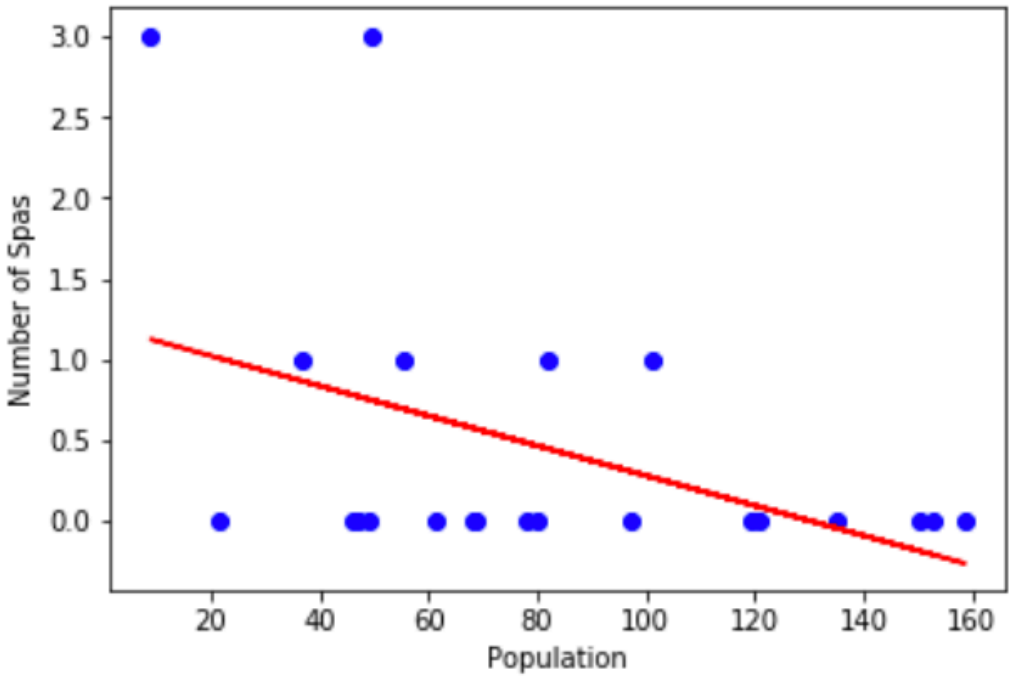
Next, I select K-means clustering applied to venue categories in each district based on the Foursquare data, to see if the clustered groups resemble the spa distribution.

4 Results

4.1 Linear Regression Results

Linear regression applied to number of spas as determined by population, given optically there seemed to be a weak correlation.

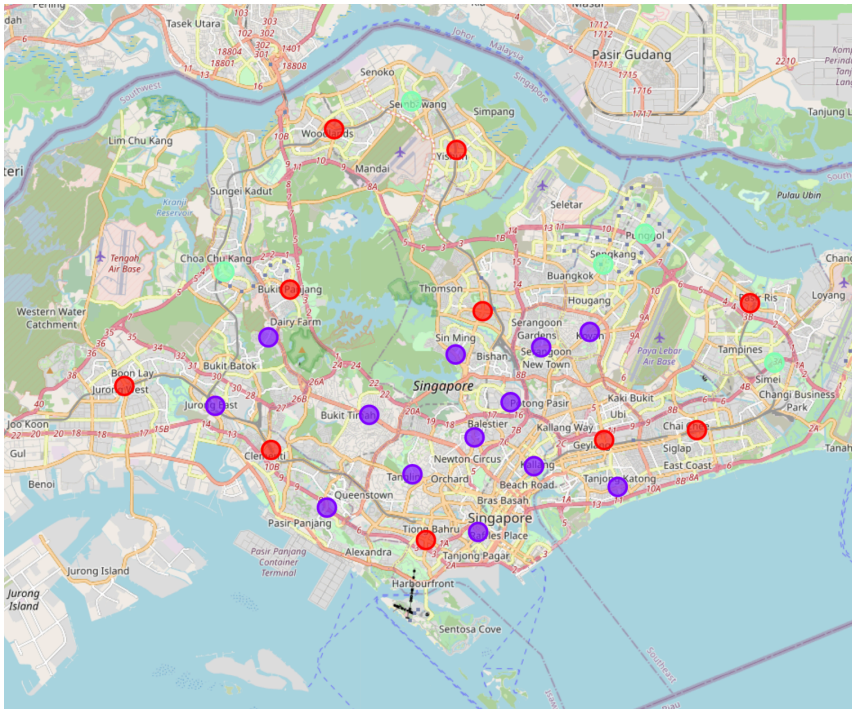
Model scoring shows relatively high error rates for Mean Absolute Error and Residual Mean Squared Error, given ideally the scores should be as low as possible, and an R2 score that implies the mean would be better used.



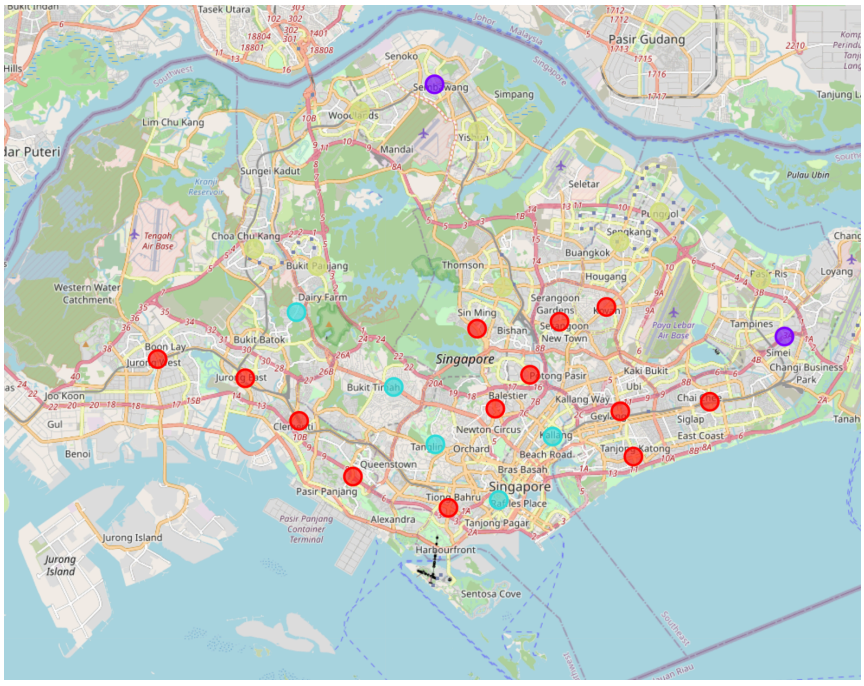

```
Mean absolute error: 0.66
Residual sum of squares (MSE): 0.54
R2-score: -3.75
```

4.2 K-Means Clustering Results

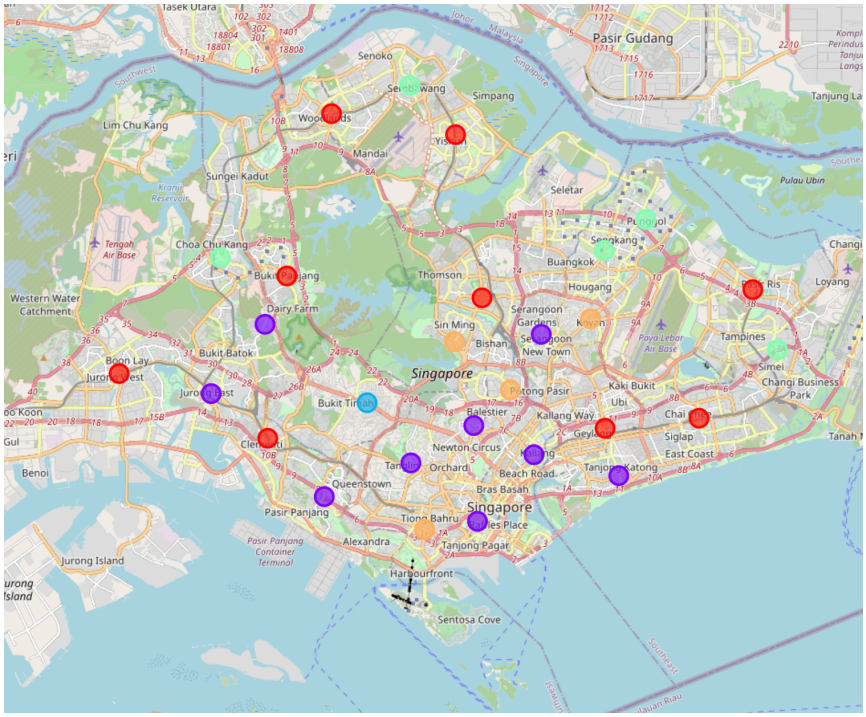
Below is for K equals ‘3’. The distribution is realtively basic, but a patter is observable of central locations versus the outer locations.



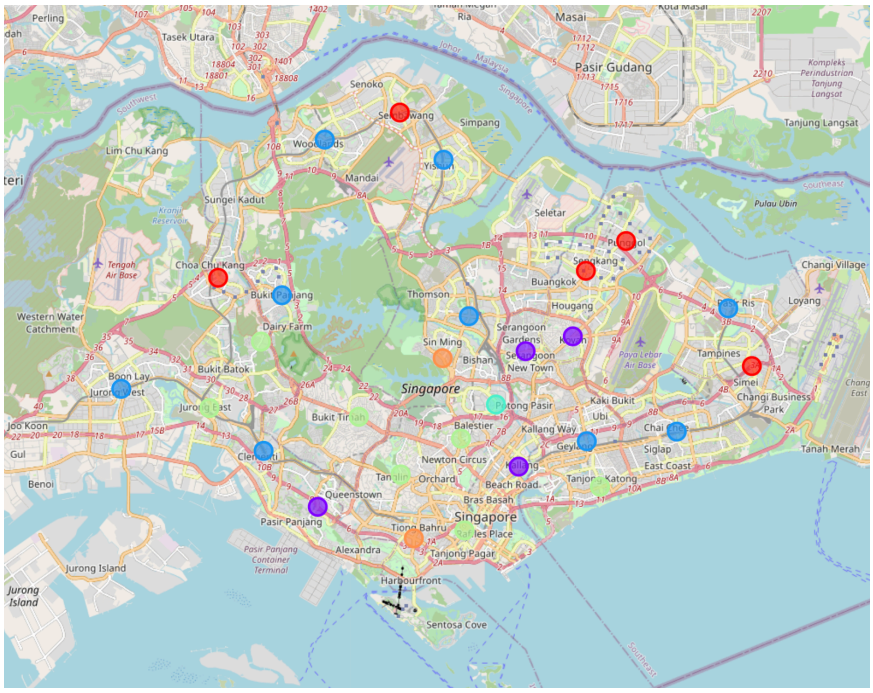
Below is for K equals ’4’:



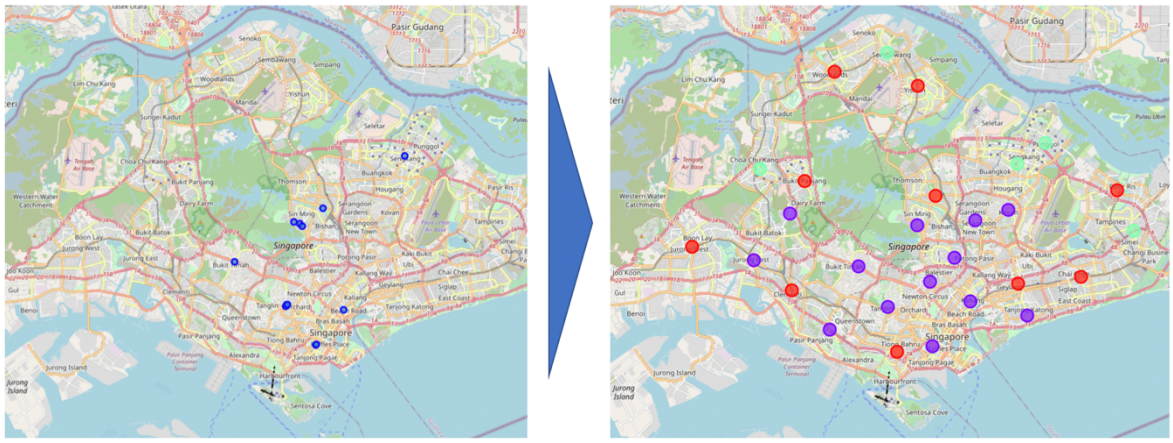
Below is for K equal to ‘5’. There is some resemblance to the spa distribution, but no cluster that seems to have much overlap.



And below shows K equal to ‘6’. At this point, the data is starting to look too fragmented.



Optimal result appears to be K equal to 3, in terms of coverage of spa locations by the purple cluster, and offering some opportunity that could be explored for new spas in the Jurong, Dairy Farm , Pasir Panjang, Tanjong Katong, and Serangoon districts.



5. Discussion

First point to observe is that there is much much less Foursquare data than expect for Spas, which was a significant limiting factor for analysis. This is likely because Foursquare is not a widely used application in Asia. It explains the frequent identification of eating and drinking establishments, rather than spas, as more tourist-accessible. Possibly also why more venues identified in central areas (more commonly frequented by tourists) than on the outskirts of the city.

Second is the inverse correlation of number of spas to population. This is understandable as commercial activity is hubbed in Singapore around transport links (train stations etc.) so would make good locations.

Thirdly, the clustering analysis showed K3 provided a reasonable degree of overlap with existing spa locations, plus some cluster locations with no spas, which is what this study was looking to identify.

6. Conclusion

This study identified that primarily less populated, so more commercial hub locations are the more common choice for a spas, plus examining the clusters of districts by popular venue categories highlighted several districts that could be considered for the site of a new spa that would align with existing ones.

I treated this exercise as an opportunity to demonstrate capability with statistical and machine learning tools. For a genuine project, I would firstly obtain data from a more Asia-established tool, such as Google, plus for statistical analysis would explore further determining attributes, such as ethnic groups, age, proximity to transport hubs, etc. It would

also be useful to see a timeline of spas – where have they been set up and failed, and explore why.

Some segmentation of spas themselves could also be useful – are they budget, premium, etc. and what products or services do they offer, opening hours, type (e.g. Swedish, Thai, Javanese, etc.)?