Singapore New Business Location

IBM Data Science Capstone

Warren Kearney

Agenda

Business Problem

Discuss business problem and who would be interested

- My wife wants to open a spa business in Singapore, but is not sure what location would make most sense
- Singapore very saturated commercially, very dense population (5.6 million people on an island 50km East to West, and 27km North to South)
- 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 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

Background Discussion

Discuss business problem and who would be interested

- Set-up of a spa could be dependent on several factors:
  - Transport links – train, bus, motorway
  - Car parking availability
  - Complimentary establishments nearby – eg restaurants, hairdressers
  - Affordability of commercial property leases
  - Availability of commercial property space
  - Nearby target demographic population – either residential or workplaces
  - Number of nearby competitors
  - Type of area – is it largely residential? Commercial? Mall? Each would have a different dynamic
  - Type of spa – should be analysed based on segments, from budget to premium?

Data types and sources - 1

Describe the data that will be used to solve the problem and the source of the data

- SG existing neighbourhoods and districts – first cluster view, without Machine Learning
  - Source: use official list from Data.gov.sg
- Coordinates of the neighbourhoods – for pictorial representation on a map
  - Use Geopy to look up above list
- Demographic data by population, income and neighbourhood – to assess target locations / audience
  - Data.gov.sg – "Resident Households by Monthly Household Income from Work, Ethnic Group and Sex of Head of Household, 2015"
- Attributes of existing spas – to see distribution and location density, and viewing by attribute
  - Foursquare, using API set up for Capstone, ideally attributes of Categories (filtering for 'spa') ratings, hours, latitude, longitude

Data types and sources - 2

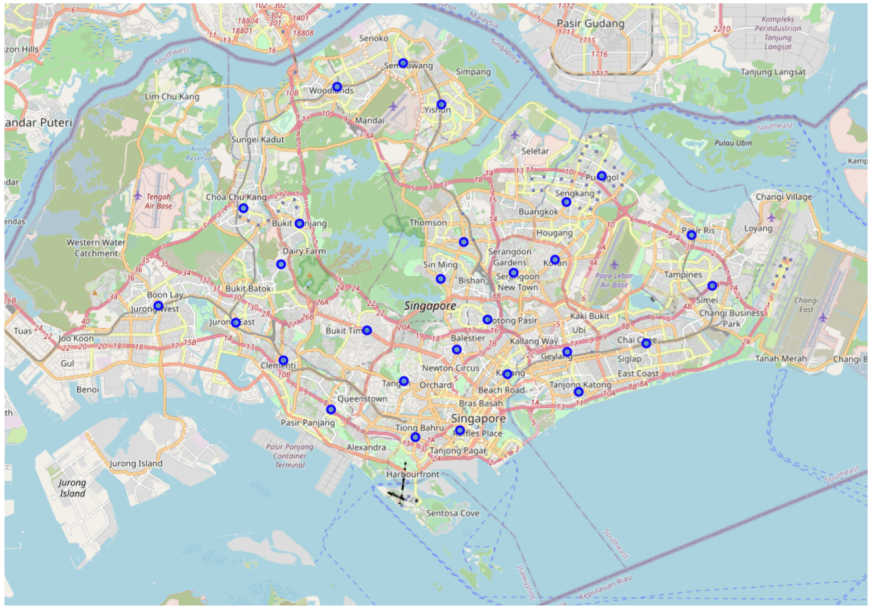
Describe the data that will be used to solve the problem and the source of the data

- Locations of potential other complimentary businesses (e.g. hotel) – to see if attractive locations based on what is nearby
  - Foursquare, ideally attributes of Categories (primary, secondary, etc.), assessing holistic list of categories first then, creating list of likely complimentary values – e.g. restaurants, hairdresser… - and their locations
- Housing and demographic data – to see location of target demographic
  - Data.gov.sg - "Resident Households by Monthly Household Income from Work, Ethnic Group and Sex of Head of Household, 2015"
  - Data.gov.sg – "Singapore Residents by Planning Area, Subzone, Age Group, Sex and Type of Dwelling, June 2011-2019
- Nearby conveniences – e.g. supermarkets, car parks, MRT stations – to assess additional benefits
  - Foursquare, using categories, with supplementary data sets where necessary (e.g. Kaggle)

Exploratory Data Analysis - 1

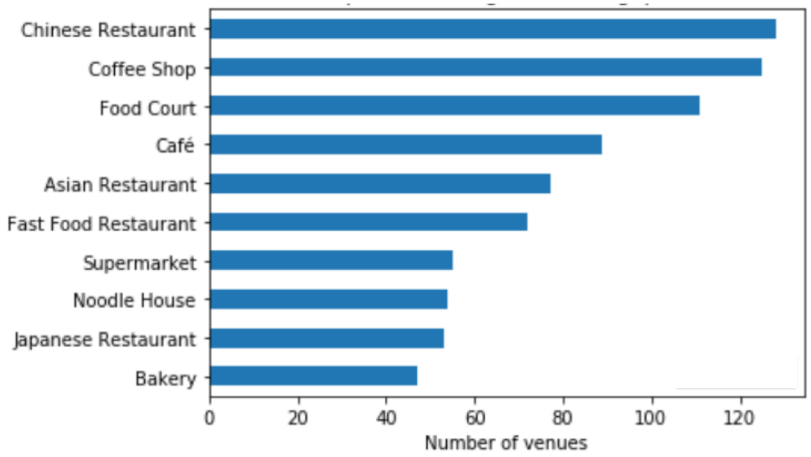
Singapore districts

Can see slightly denser clustering in the centre of the island



Foursquare popular venue categories

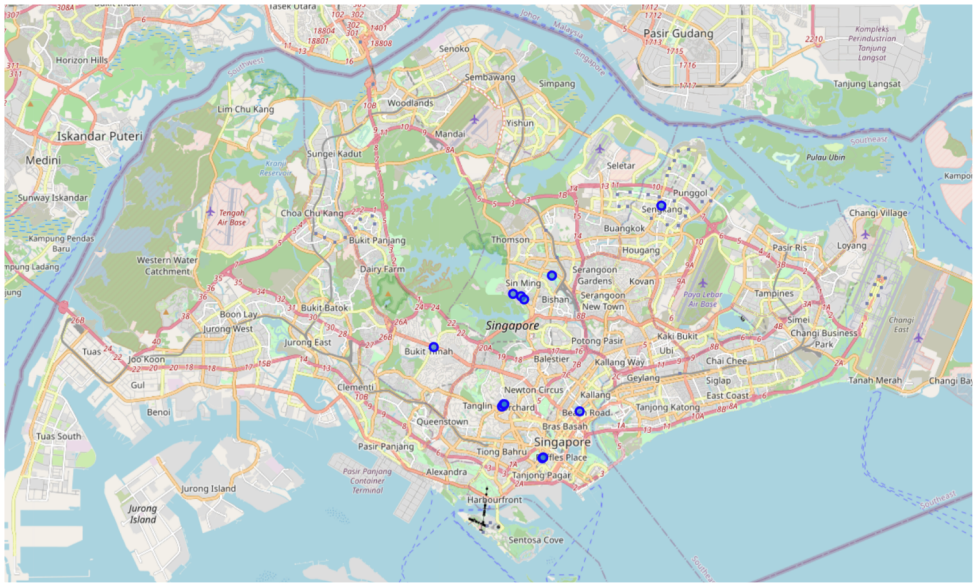
Food & beverage dominates the categories!



Exploratory Data Analysis - 2

Spa distribution

Very few spas are showing vs expectations - limited usage of Foursquare data in Singapore means limitation in statistical analysis

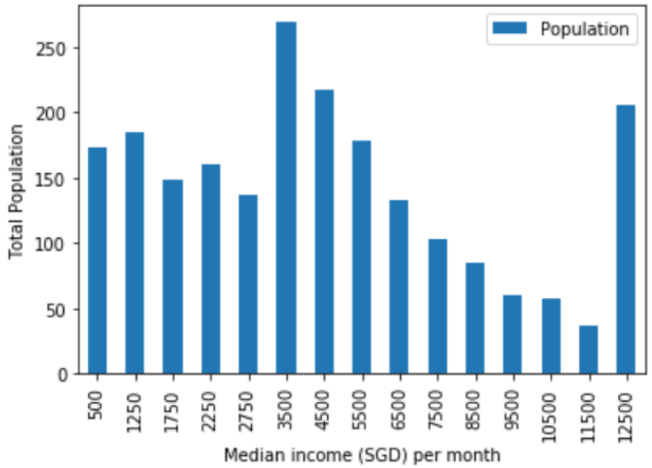


District	Venue
Bishan	3
Tanglin	3
Outram	2
Ang Mo Kio	1
Bukit Timah	1
Kallang	1
Sengkang	1

Exploratory Data Analysis - 3

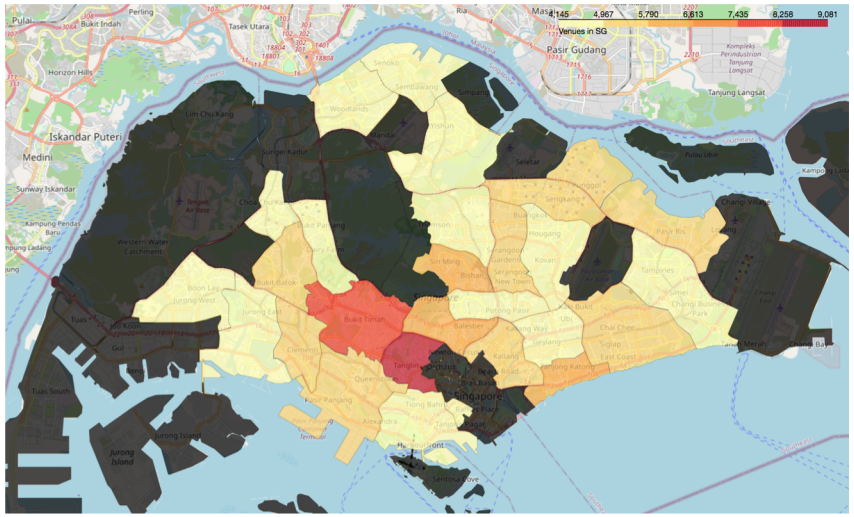
Wealth distribution in Singapore

Can see a bell shape with more people in middle-income bracket



Wealth per District

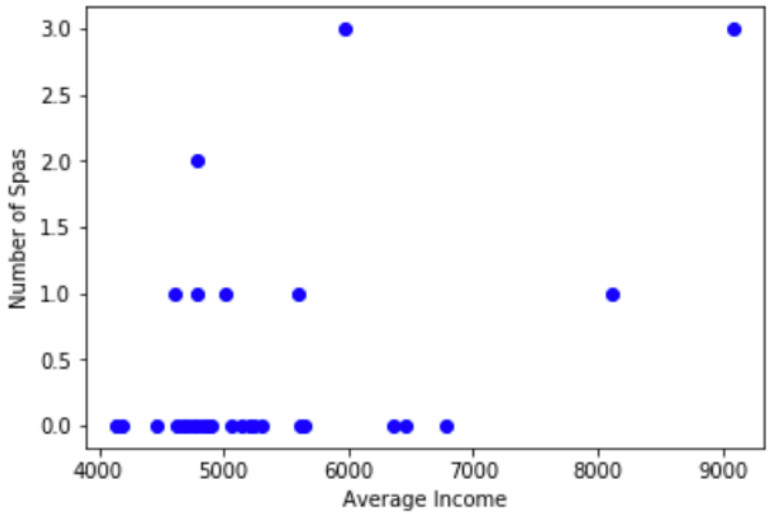
Can see the highest median income per month is central areas



Inferential statistical testing

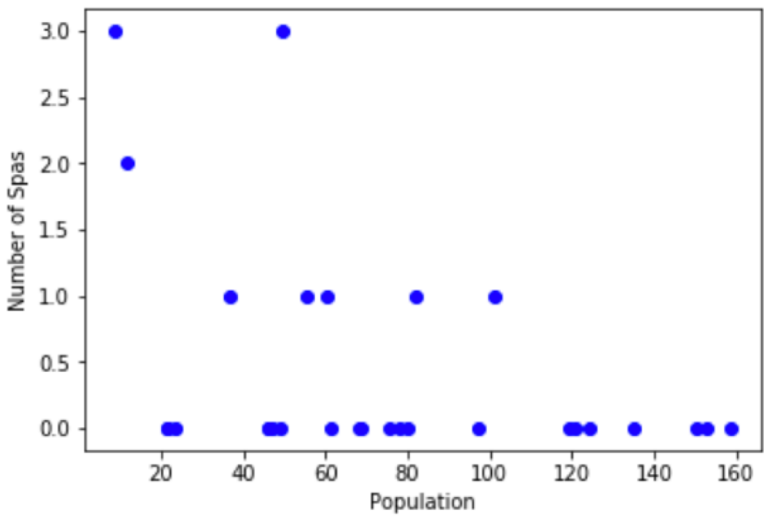
Relationship between Spa count and income

No perceived pattern



Relationship between Spa count and population

Seems lower population = more spas



Machine Learning

Linear regression models

- Llook for correlation between #spas in neighbourhood and other factors, eg ethnic group, demographic, employment, hotels,

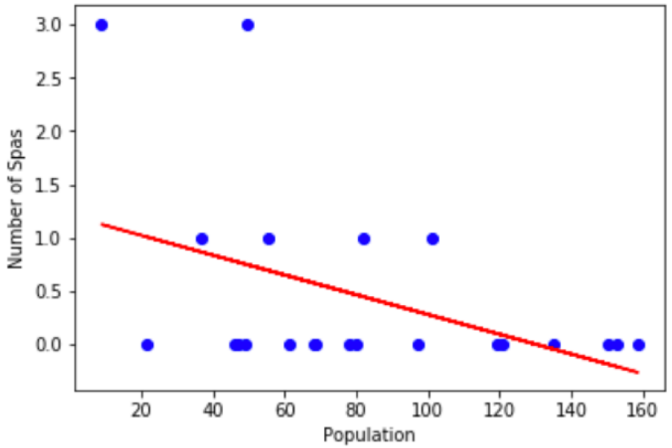
K-means clustering applied to location features

- Linear regression – look for correlation between #spas in neighbourhood and other factors, eg ethnic group, demographic, employment, hotels,

Linear Regression Results

Linear regression applied to explore relationship with spas and population and income

A clear inverse corellation exists between population and number of spas




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

Mean absolute error

Given lower scores are desired, and our outcomes are 0.3, 0.66 could be considered high

Mean squared error

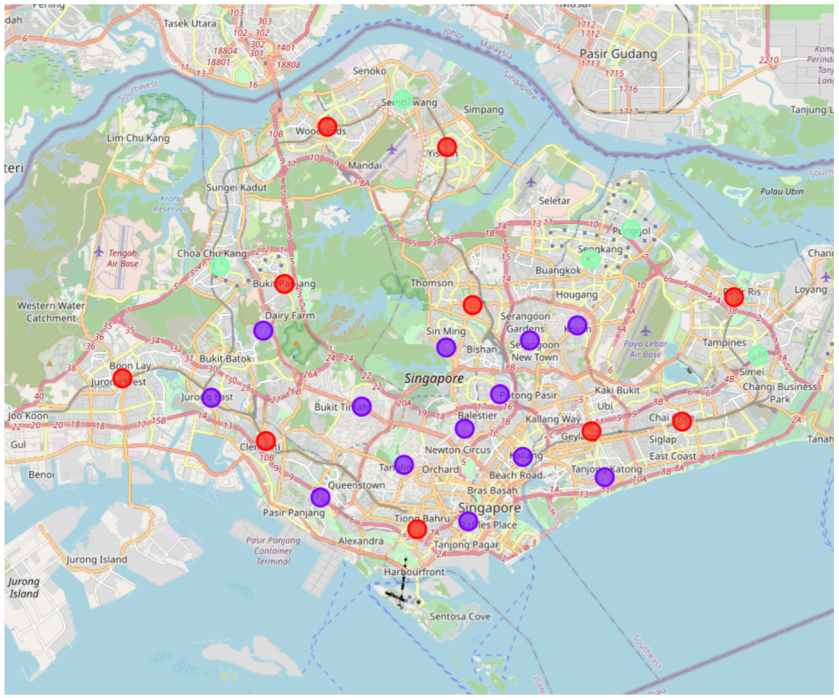
Given lower scores are desired, and our outcomes are 0.3, 0.66 could be considered high

R squared score

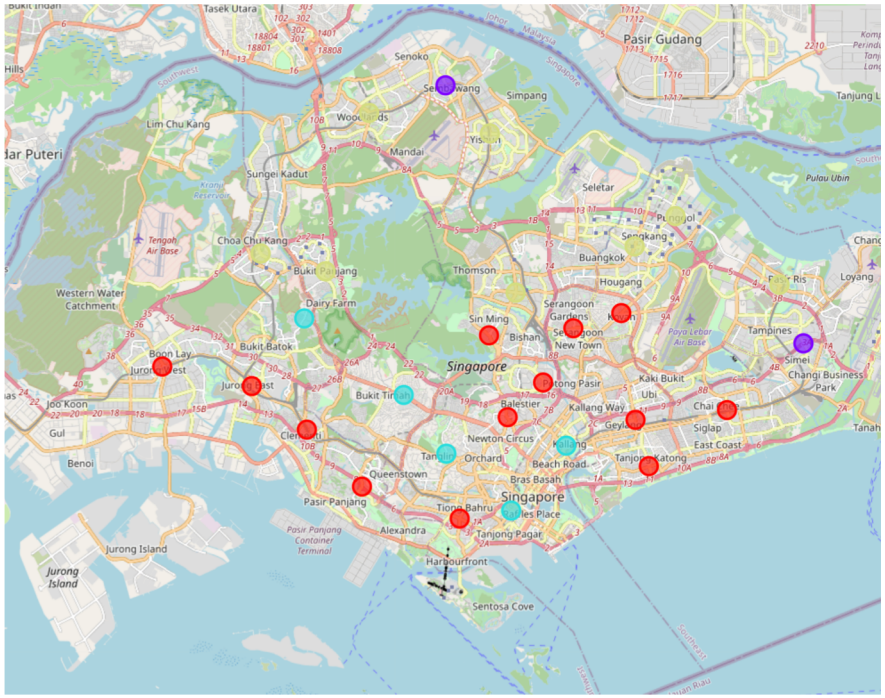
A negative score means doing worse than the mean value…

K-means clustering results - 1

Basic clustering using most popular places - 3

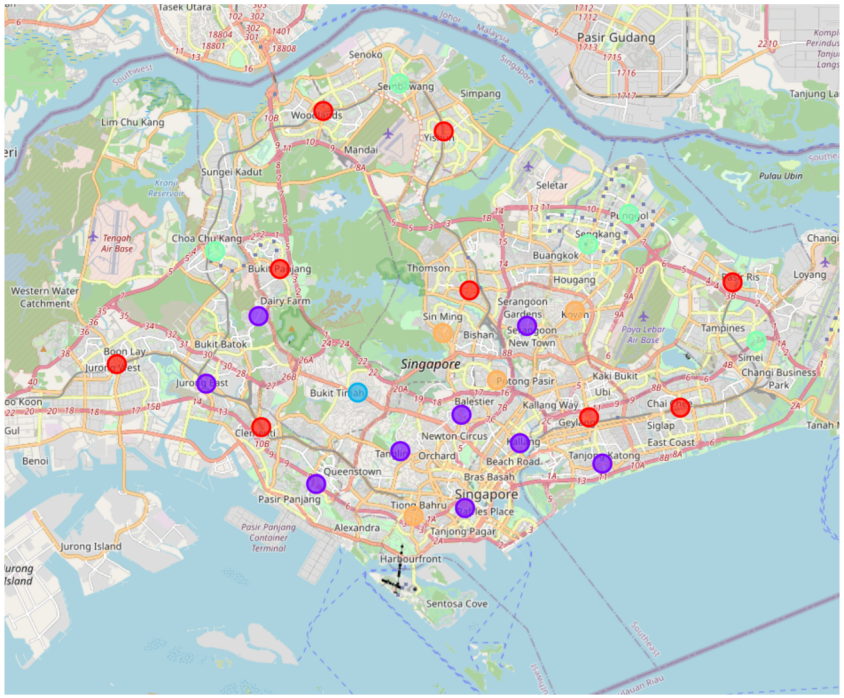


Basic clustering using most popular places - 4

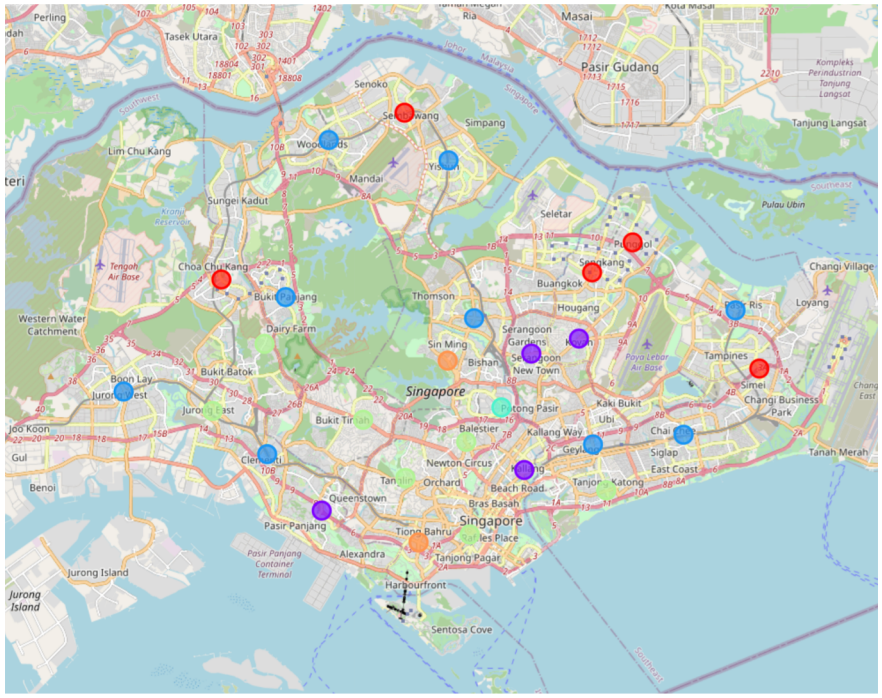


K-means clustering results - 2

Basic clustering using most popular places - 5

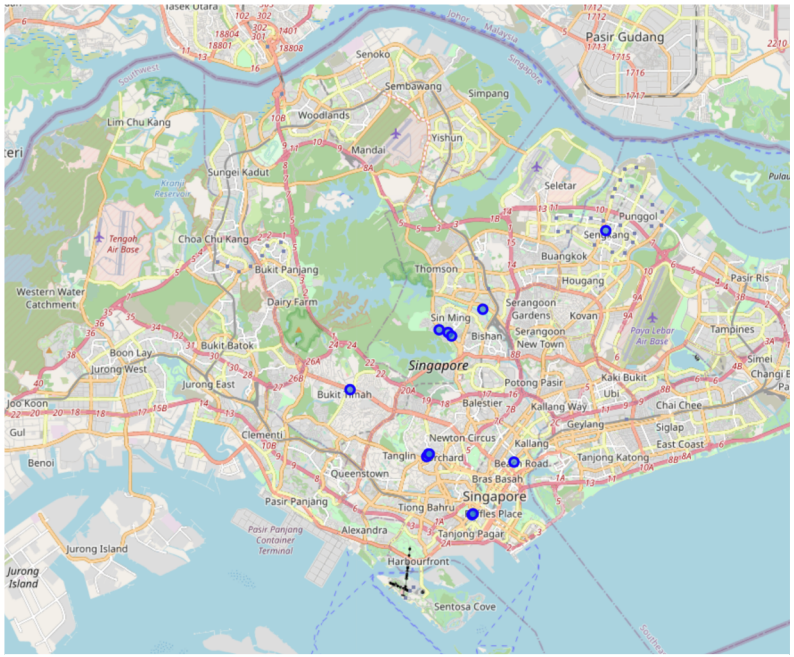


Basic clustering using most popular places - 6

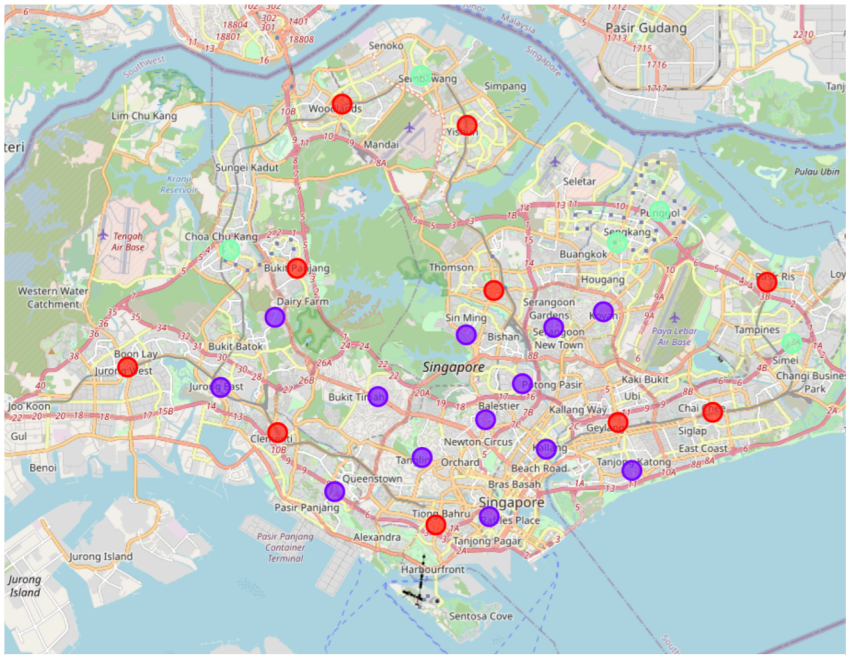


K-means clustering results – ‘best’

Spa locations



Clustering with the most overlaps with spas



Key Observations

Data availability

Much less spa and location data from Foursquare than anticipated – not a widely used platform in Asia

Largely tourist-type venues that have been recorded

Correlation of population to number of spas

There is an inverse relationship – the higher the resident population, the less spas are seen

Clustering of locations overlapping with spas

The clustering pattern that correlates most to the spa distribution is central and across the south coast

Conclusion

1) Spa Business Opportunities

- Spas more likely to be in commercial hub locations
- Overlapping statistical clusters of districts by popular venue categories highlighted **Jurong, Dairy Farm , Pasir Panjang, Tanjong Katong, and Serangoon** as districts that could be considered for the site of a new spa that would align with existing ones

2) Further analysis to be carried out

- More Singapore-centric source of location data as basis for analysis
- More multi-source regression analysis of demographic drivers, such as Ethnicity of spa, of population, age, distance to transport links, etc.

Agenda

Appemdix

Data links

Data used in the analysis

- Demographic data by sex and race and neighbourhood
  - Import urllib url = 'https://data.gov.sg/api/action/datastore_search?resource_id=d683afc9-d1e6-45a1-8d51-073710b7daca&limit=5&q=title:jones' fileobj = urllib.urlopen(url) print fileobj.read()
- Singapore Residents by Planning Area, Subzone, Age Group, Sex and Type of Dwelling, June 2011-2019
  - import urllib url = 'https://data.gov.sg/api/action/datastore_search?resource_id=6d799e1c-ab06-4fad-bee3-2f28606bc971&limit=5&q=title:jones' fileobj = urllib.urlopen(url) print fileobj.read()
- Neighbourhoods
  - https://docs.onemap.sg/#search API for Singapore coordinates
- Foursquare