Applied Data Science Capstone Project

Going Green in The Garden City:

Veganism in Singapore

Prepared by:

Ryan Chan

1. Introduction

Veganism is seeing a growing movement in Singapore. In fact, the garden city is named the second most vegan-friendly city in Asia by PETA and the sixth most vegan-friendly city in the world by HappyCow. The customer base of HappyCow has even seen as increase of non-vegetarian customer base from 30% to 80%.

However, for an entrepreneur looking to tap into this growing market in Singapore, it is no easy feat to determine the best location to open a vegan restaurant without facing heavy competition or a low demand from the population surrounding the vicinity.

To tackle this issue, this report aims to address two questions:

1. Is there a correlation between the density population in each Singapore neighborhood and the number of vegan/vegetarian restaurants available in the location?
2. If there is a correlation, which opportunities/market space are available?

Given the limitations of data indicating the proportions of dietary preferences of residents in each neighbourhood in Singapore, this report assumes that the number of vegan/vegetarian restaurants opening in the vicinity is proportionate to the demand of vegan food in that vicinity.

2. Data

Data used in this project are collected from the various publicly available sources below:

1. Neighbourhoods in Singapore with population density collected from Wikipedia https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore
2. Coordinates of each neighborhood are obtained by running geocoding web service.
3. Foursquare API are used to get the 'Vegan/Vegetarian restaurants' with search radius for each neighborhood. Descriptive statistics and clustering are done to explore the data further.

3. Methodology

3.1 Data Acquisition and Cleaning

Pandas library is used to read the Wikipedia planning areas dataset into a dataframe using pandas.read_html() method. Below shows the initial dataframe extracted.

	Name (English)	Malay	Chinese	Pinyin	Tamil	Region	Area (km2)	Population[7]	Density (/km2)
0	Ang Mo Kio	NaN	宏茂桥	Hóng mào qiáo	ஆங் மோ கியோ	North-East	13.94	163950	13400
1	Bedok	*	勿洛	Wù luò	பிடோக்	East	21.69	279380	13000
2	Bishan	NaN	碧山	Bì shān	பீஷான்	Central	7.62	88010	12000
3	Boon Lay	NaN	文礼	Wén lǐ	பூன் லே	West	8.23	30	3.6
4	Bukit Batok	*	武吉巴督	Wǔjí bā dū	புக்கிட் பாத்தோக்	West	11.13	153740	14000

Columns unnecessary for this project, particularly Malay, Chinese, Pinyin and Tamil, were removed from the dataframe and missing values (denoted by *) were replaced with value 0. As population and density were object types, they were also changed to float64 types as shown below.

```
Name (English)      object
Region              object
Area (km2)         float64
Population[7]       object
Density (/km2)      object
dtype: object
```

```
Out[79]: Name (English)      object
         Region              object
         Area (km2)         float64
         Population[7]      float64
         Density (/km2)     float64
         dtype: object
```

3.2 Obtaining Coordinates for each neighborhood

Using geolocator API, a loop was run to obtain the coordinates (Lat/Long) for each neighborhood. The coordinates for each neighborhood are added as two new columns to the dataframe.

	Name (English)	Region	Area (km2)	Population[7]	Density (/km2)	Latitude	Longitude
0	Ang Mo Kio	North-East	13.94	163950.0	13400.0	1.37161	103.84546
1	Bedok	East	21.69	279380.0	13000.0	1.32425	103.95297
2	Bishan	Central	7.62	88010.0	12000.0	1.35079	103.85110
3	Boon Lay	West	8.23	30.0	3.6	1.33333	103.70000
4	Bukit Batok	West	11.13	153740.0	14000.0	1.34952	103.75277

Singapore's coordinates were obtained using geolocator and used as the starting coordinates for the map visualization as shown below.

3.3 Retrieving nearby venues for each neighborhood

Using a pre-registered Foursquare developer account, the Foursquare API was used to get the top 100 venues in each neighborhood with a search radius individualized to each neighborhood. The

venue name, coordinates, and category were obtained from the JSON response and appended to a list which was converted to a pandas dataframe as shown below.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Ang Mo Kio	30	30	30	30	30	30
Bedok	30	30	30	30	30	30
Bishan	30	30	30	30	30	30
Boon Lay	30	30	30	30	30	30
Bukit Batok	30	30	30	30	30	30
Bukit Merah	30	30	30	30	30	30
Bukit Panjang	30	30	30	30	30	30
Bukit Timah	30	30	30	30	30	30
Central Water Catchment	30	30	30	30	30	30
Changi	30	30	30	30	30	30
Changi Bay	12	12	12	12	12	12
Choa Chu Kang	30	30	30	30	30	30

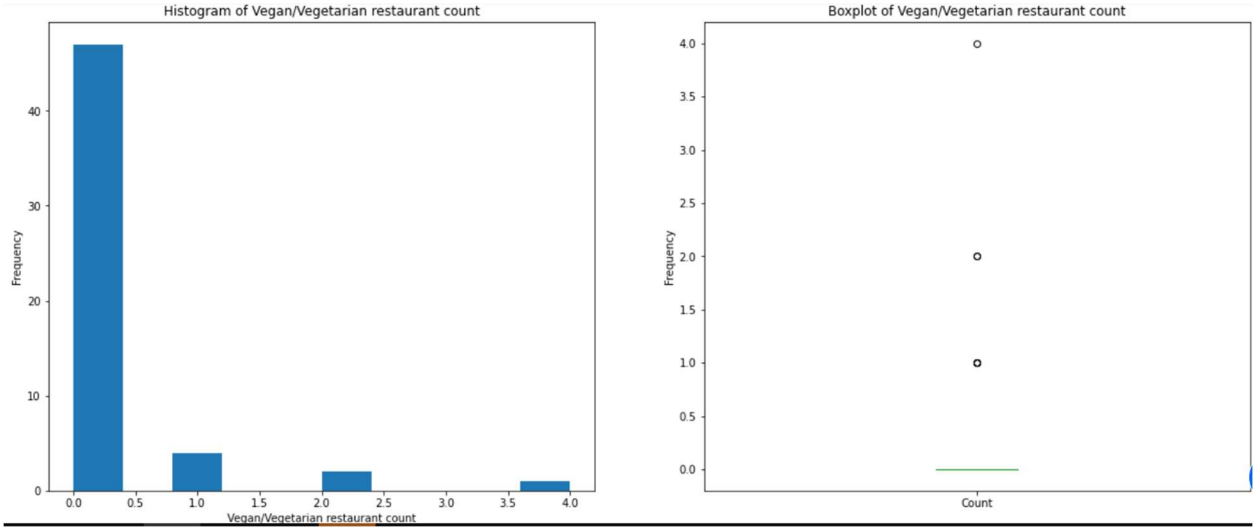
Due to overlapping search results from the Foursquare API, duplicates were removed with a total of 385 venue category duplicates being removed.

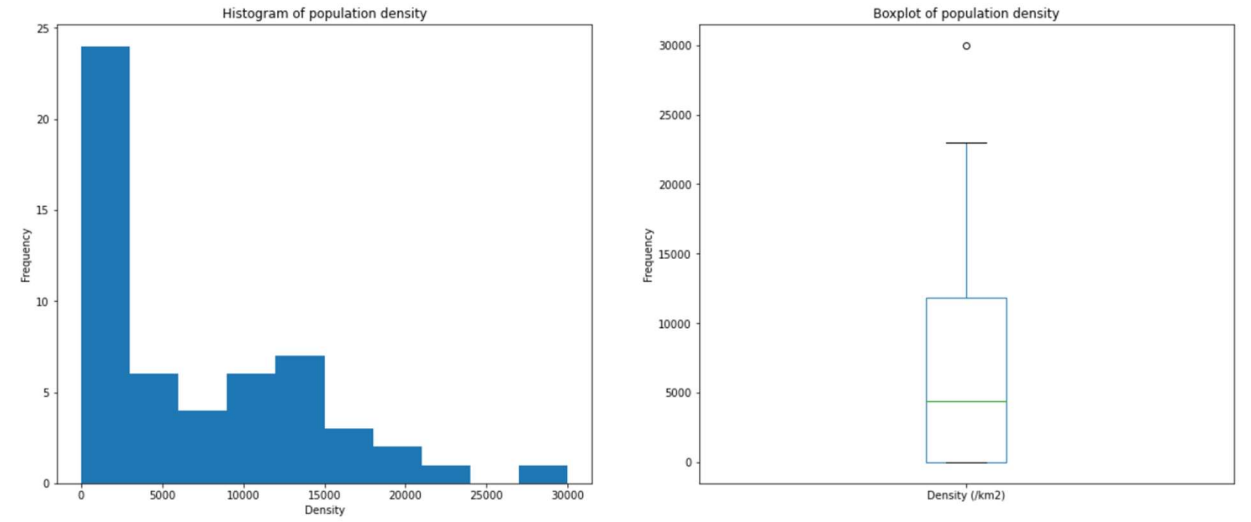
3.4 Normalizing and Exploratory Analysis of Data

After removing the duplicates, one-hot encoding was performed on the venue category. Categorical variables of the venue category were converted into dummy variables and these dummy variables were grouped by neighborhood to produce a total of 233 venue categories per neighborhood.

Pearson correlations were computed for each venue category against population density and sorted in descending order as shown below.

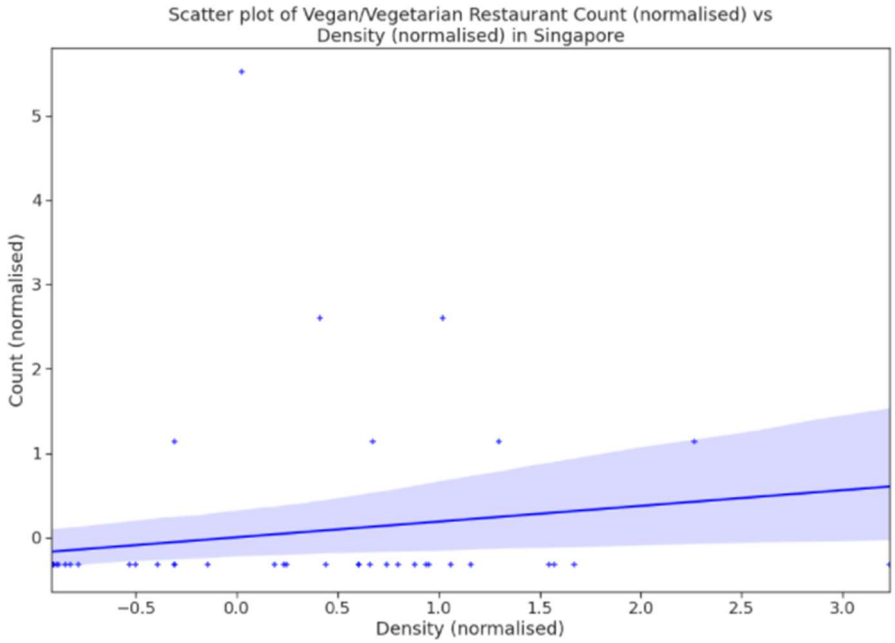
```
Pearson correlation of top 5 venue categories with population density:
Noodle House             0.448204
Portuguese Restaurant    0.444195
Stadium                  0.358885
Kids Store               0.311143
Pool                     0.306756
dtype: float64

Pearson correlation of bottom 5 venue categories with population density:
Event Space   -0.179932
Exhibit       -0.191916
Art Gallery   -0.198738
Hotel         -0.212291
Waterfront    -0.252655
dtype: float64
```





The distributions of density and vegan/vegetarian restaurant count seems to be right-skewed. However, further data analysis and standardization is required before meaningful comparisons can be discerned from the 2 variables.

Thus, these variables were standardized using the StandardScaler() function and a scatterplot with regression line is plotted using seaborn library. Below shows the following result.

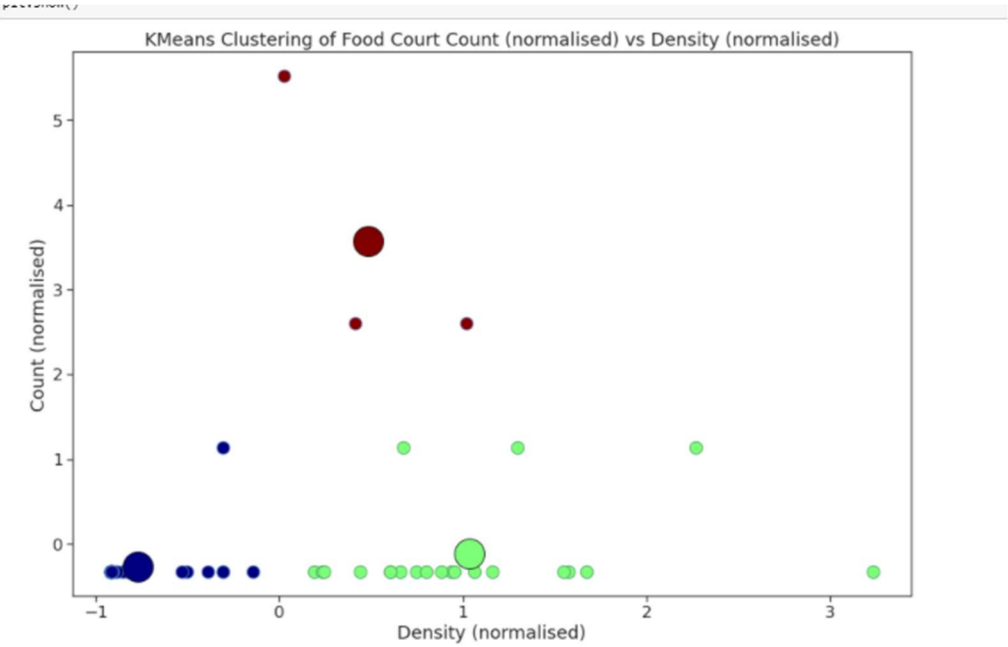


There seems to be a positive but very weak correlation between density and vegan/vegetarian restaurant count.

3.5 Clustering with K-Means Clustering

This method of partitioning clustering method divides data into non-overlapping subsets without labels or cluster internal structure while trying to minimize intra-cluster distances and maximize inter-cluster distances.

The number of clusters, or K, was selected to be 3. The normalized values were fitted and the generated K-means clusters are as shown below.



All datapoints are given cluster labels in the algorithm including outliers. The cluster labels generated were inserted into the dataframe, with the merged dataframe now containing neighborhood, population density, count of vegan/vegetarian restaurants, coordinates and cluster label.

4. Results

A summary of each cluster is as shown below.

	Neighborhood	Density (/km2)	Count	Cluster Label
35	Rochor	6800.0	4	2
4	Bukit Batok	14000.0	2	2
30	Paya Lebar	9600.0	2	2

	Neighborhood	Density (/km2)	Count	Cluster Label
40	Simpang	11500.0	1	1
39	Serangoon	23000.0	1	1
15	Hougang	16000.0	1	1
0	Ang Mo Kio	13400.0	0	1
23	Marine Parade	8000.0	0	1
49	Tuas	14300.0	0	1
46	Tanglin	12400.0	0	1
38	Sengkang	8400.0	0	1
36	Seletar	8300.0	0	1
33	Queenstown	17800.0	0	1
29	Pasir Ris	13500.0	0	1
18	Kallang	11000.0	0	1
1	Bedok	13000.0	0	1
17	Jurong West	18000.0	0	1
14	Geylang	11400.0	0	1
12	Clementi	9800.0	0	1
11	Choa Chu Kang	30000.0	0	1
6	Bukit Panjang	15000.0	0	1
5	Bukit Merah	11000.0	0	1
2	Bishan	12000.0	0	1
53	Yishun	18700.0	0	1

	Neighborhood	Density (/km2)	Count	Cluster Label
16	Jurong East	4400.00	1	0
3	Boon Lay	3.60	0	0
31	Pioneer	3.40	0	0
51	Western Water Catchment	0.25	0	0
50	Western Islands	2.30	0	0
48	Toa Payoh	1.40	0	0
47	Tengah	2800.00	0	0
45	Tampines	53.20	0	0
44	Sungei Kadut	0.00	0	0
43	Straits View	244.00	0	0
42	Southern Islands	3000.00	0	0
41	Singapore River	0.00	0	0
37	Sembawang	26.30	0	0
34	River Valley	4400.00	0	0
32	Punggol	8.30	0	0
28	Outram	960.30	0	0
7	Bukit Timah	4400.00	0	0
27	Orchard	5600.00	0	0
26	Novena	1.20	0	0
25	Newton	3800.00	0	0
24	Museum	480.00	0	0
22	Marina South	0.00	0	0
21	Marina East	0.00	0	0
20	Mandai	180.20	0	0
19	Lim Chu Kang	5.20	0	0
13	Downtown Core	680.00	0	0
10	Changi Bay	0.00	0	0
9	Changi	80.62	0	0
8	Central Water Catchment	0.00	0	0
52	Woodlands	13.00	0	0

Cluster 2 (n=0) contain neighborhoods with the highest number of vegan/vegetarian restaurant compared to cluster 1 (n=1), with mean density = and cluster 0 (n=2).

5. Discussion

The accuracy of the K means algorithm is sufficiently high given the frequency of overlapping density between cluster 1 and 2. However, based on our observations we observe that majority of vegan/vegetarian restaurants are present in cluster 2, with the highest count being in Rocher. While there is a need for more data to safely conclude that opening a vegan restaurant is advisable for a vegan restaurant entrepreneur given the weak correlation and sufficiently accurate clustering algorithm, this finding seems to point towards the direction of Bukit Batok and Paya Lebar being potential areas.

6. Conclusion

In this capstone project, data was extracted from Wikipedia with information pertaining the density population of each neighborhood in Singapore in which each neighborhood's coordinates were obtained using geolocator. Foursquare API was also used to get venue for each neighborhood within a specified radius for each neighborhood and K-means clustering machine learning algorithm was finally used to cluster the data.

Pertaining to the business problem at hand on opening a vegan restaurant in Singapore, there is a weak positive correlation between population density and the number of vegan/vegetarian restaurants surrounding the neighborhood. This evidence seems to indicate that opening a new vegan restaurant in cluster 3 should be considered, but further analysis and a more accurate model alongside more quantity of data and features such as average household income should be explored before a more convincing statement can be presented.