**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**

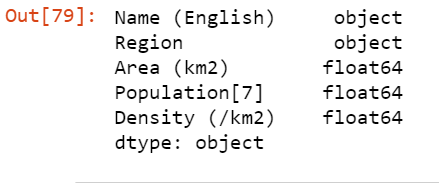
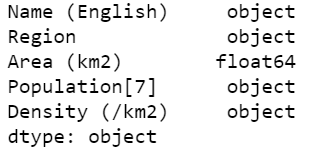
*3.1 Data Acquisition and Cleaning*

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.

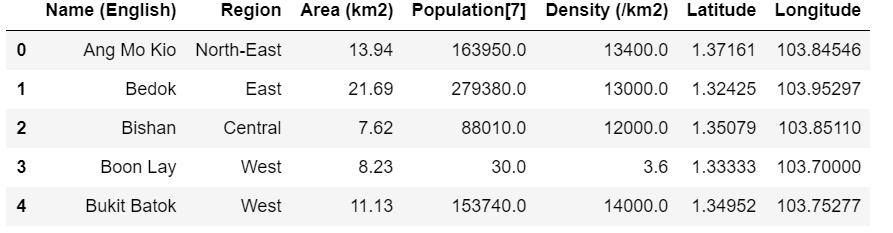
****

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.



*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. 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.

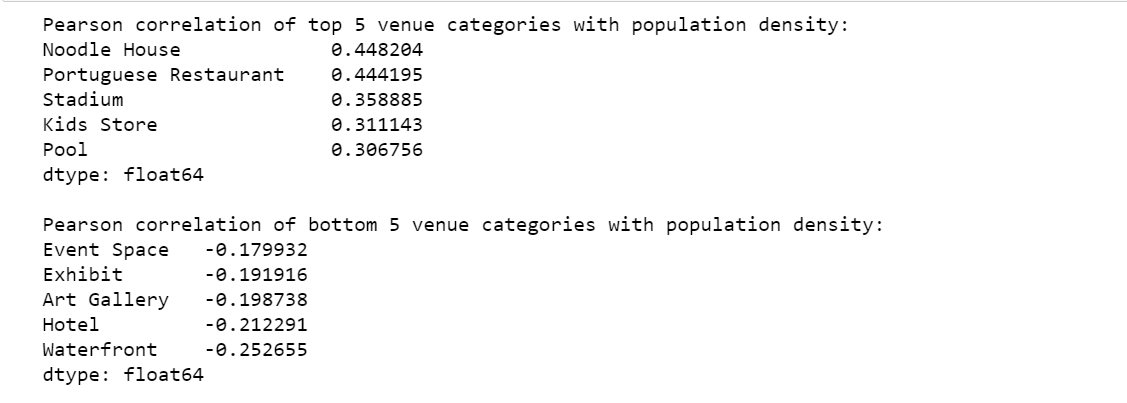


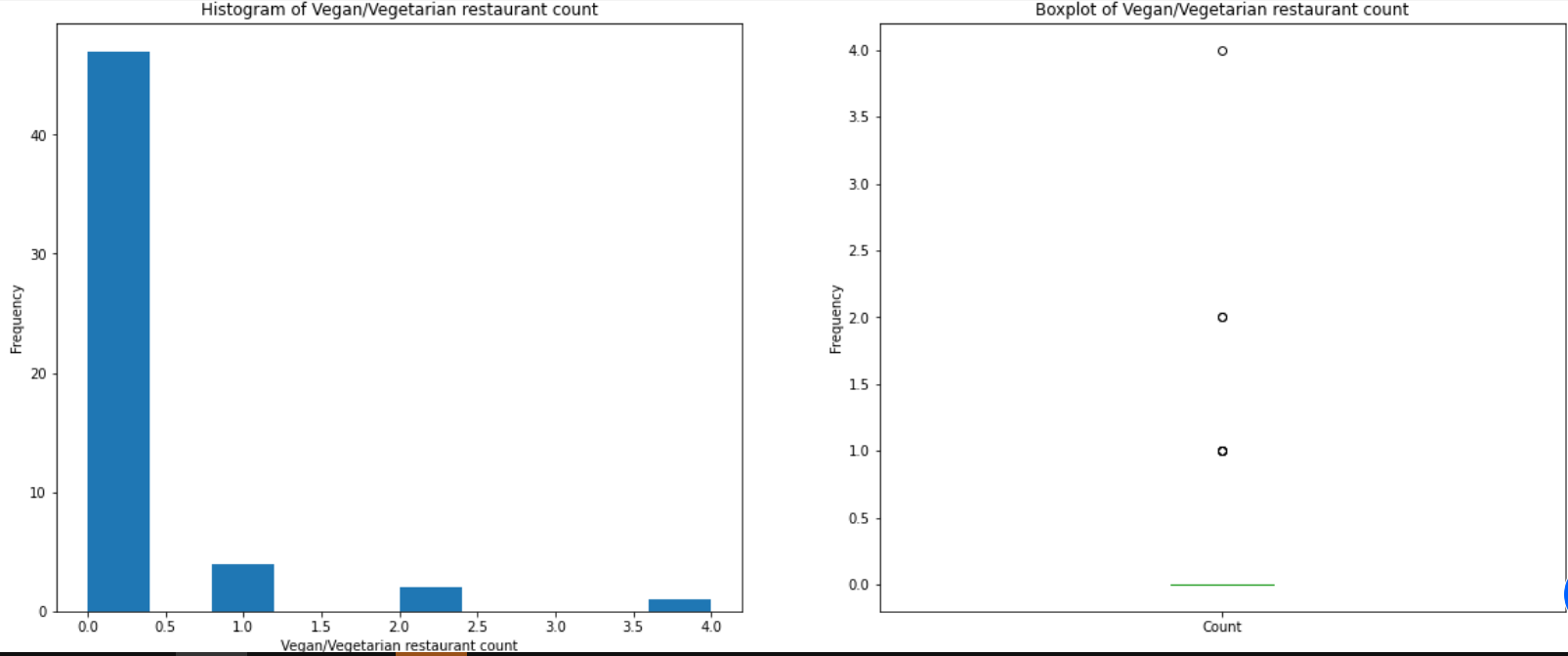
Due to overlapping search results from the Foursquare API, duplicates were removed with a total of 385 venue category duplicates being removed.

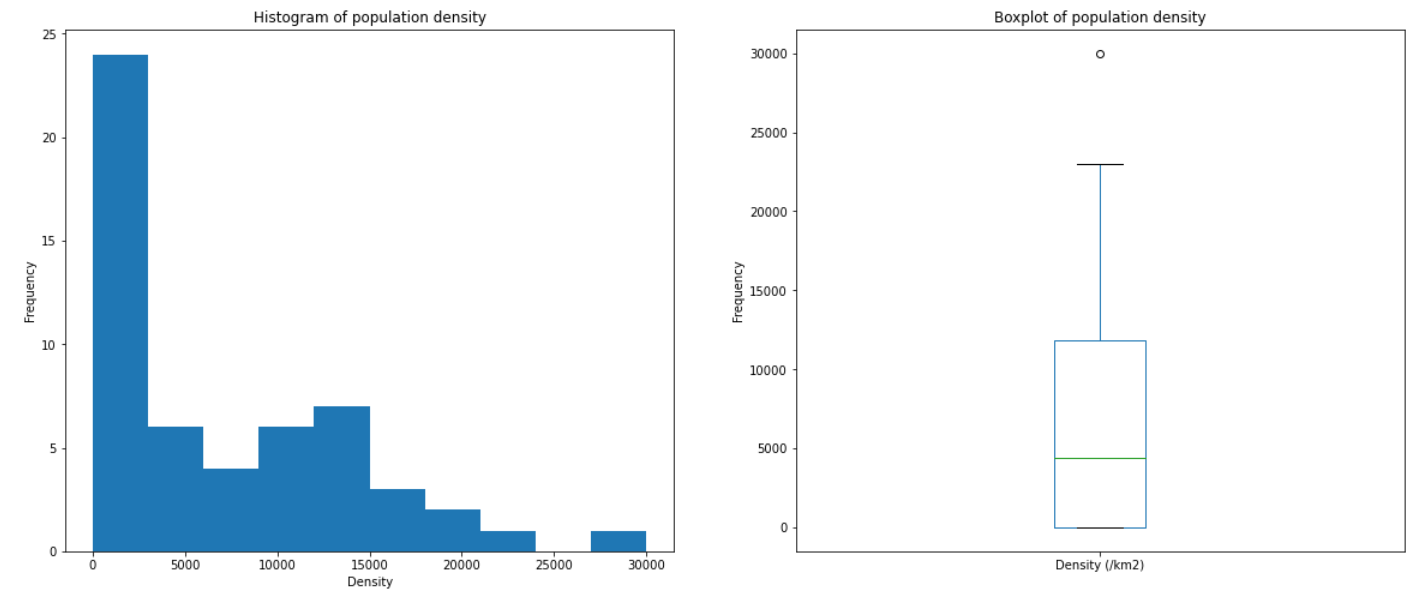
*3.4 Normalizing and Exploratory Analysis of Data*

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.

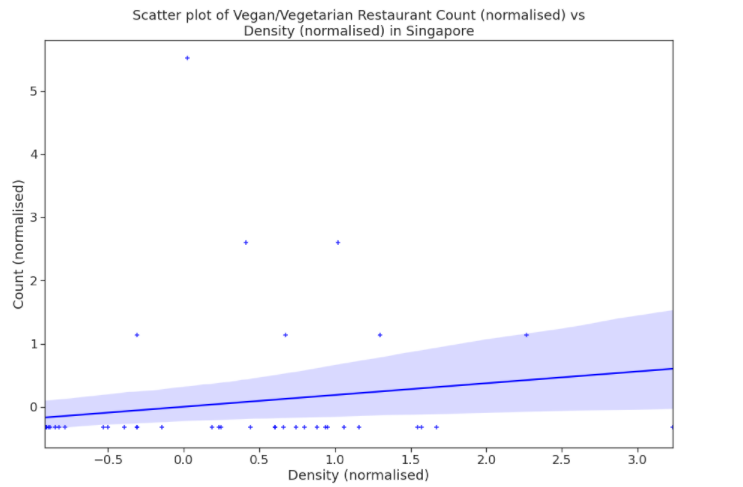






The distributions of density and vegan/vegetarian restaurant count seems to be right-skewed. However, further data analysis and standardization is required before meaningful comparisons can be discerned from the 2 variables.

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.

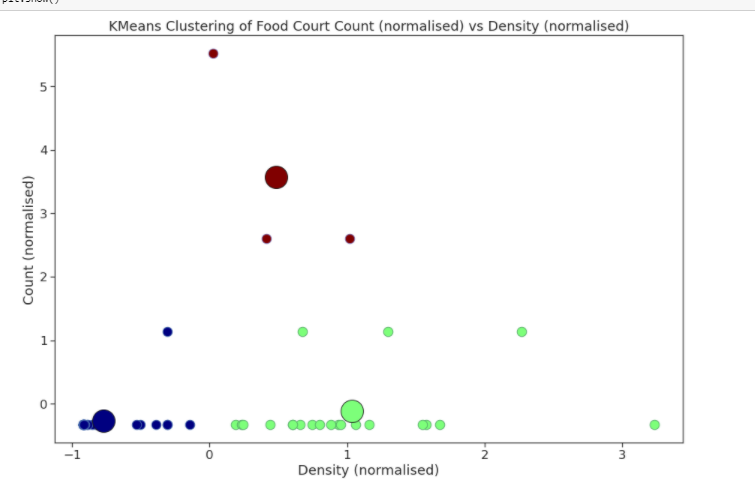


There seems to be a positive but very weak correlation between density and vegan/vegetarian restaurant count.

*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.

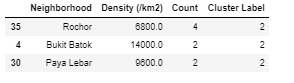
The number of clusters, or K, was selected to be 3. The normalized values were fitted and the generated K-means clusters are as shown below.

**

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.





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.