IBM Data Science
Applied Data Science Capstone

The Battle of Neighborhoods

OPENING NEW FOOD COURTS IN SINGAPORE'S PLANNING AREAS

BY
YONG GUAN JIE ANDREW

10 MAY, 2020

1. Introduction

Food courts/hawker centres are open-air complexes housing many stalls that sell a variety of cheap food. They are commonly found in South East Asian countries, and Singapore is no exception. Despite having an annual GDP of USD 372 billion in 2019, food courts in Singapore continue to remain the go-to place for many locals and foreigners looking for cheap and delicious food. There are even hawker stalls in Singapore awarded with Michelin stars, selling foods at affordable prices. It is no surprise that tourism is a major contributor to the Singaporean economy, attracting 18.5 million visitors to the country. Tourism contributed around 9.9% of the country's GDP, and around 8.6% of employment in 2016.

In the tiny state of 5.6 million inhabitants and total surface area of 725 km^2, Singapore is ranked second densest country in the world with population density of 7,804 inhabitants/km^2. With limited land resources and ever increasing population, the government has taken serious steps to ensure city planning in the country is sustainable for many generations. This has led to the establishment of Urban Redevelopment Authority, which has divided the country into 55 Planning Areas, which are the main urban planning and census divisions in Singapore. Development in these Planning Areas is guided by the Development Guide Plan.

With the growing demand in tourism and strict development guidelines set by Urban Redevelopment Authority, local municipalities/ entrepreneurs must carefully consider where to build food courts that can balance supply with demand. The results of this project are aimed at local municipalities/ entrepreneurs, to aid them in making decisions on where to open new food courts.

1.1 Business Problems

1. Does a relationship exist between population density and the number of food courts surrounding that population?
2. Where should local municipalities / entrepreneurs consider opening additional food courts that could meet population demands?

2. Data

Data that will be used to solve the above business problems are collected from various publicly-available sources.

1. The list of Planning Areas in Singapore are collected from a Wikipedia article entitled "Planning Areas of Singapore" found at https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore. This dataset lists the names of all Planning Areas, their areas (in km^2), and densities.
2. The coordinates of each Planning Area are obtained by running a geocoding web service (Nominatim, which is based on OpenStreetMap). The coordinates obtained for each Planning Area will be used to plot map markers on the map for visualisation.
3. The Foursquare API is used to get the top 100 venues in each Planning Area, with search radius individualised to each Planning Area. The results returned by the Foursquare API are extracted for relevant information, and descriptive statistics are done to further explore the data.

By combining the above data, further data analysis can be performed. Correlations between density and all venue categories are calculated. Bar charts, histograms, and scatter plots can be plotted to visualise the relationships between variables. Lastly, machine learning clustering algorithms are used to cluster data and find solutions to the business problems stated above. The clustered data are then plotted onto a map with map markers for better understanding.

3. Methodology

3.1 Data Acquisition and Wrangling

The datasets required for this project are publicly available online. The acquisition and wrangling of data can be divided into 3 stages:

1. Data acquisition and wrangling on Planning Areas
2. Data acquisition and wrangling for coordinates for each Planning Area
3. Data acquisition and wrangling on top venues in each Planning Area

3.1.1 Data acquisition on Planning Areas

The Pandas Library is used to read various datasets into a data frame. Using the pandas.read_html() method, the following dataframe is obtained from the Wikipedia article found at https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore:

	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

Table 1: Initial dataframe obtained from Wikipedia (first 5 rows shown)

This dataframe contains unnecessary columns and missing data (denoted by *). The “Malay”, “Chinese”, “Pinyin”, “Tamil” columns are dropped, and missing data are replaced with value of 0. Table 2 shows the dataframe after modification.

	Planning Area	Region	Area	Population	Density
0	Ang Mo Kio	North-East	13.94	163950	13400
1	Bedok	East	21.69	279380	13000
2	Bishan	Central	7.62	88010	12000
3	Boon Lay	West	8.23	30	3.6
4	Bukit Batok	West	11.13	153740	14000

Table 2: Dataframe after dropping columns and replacing missing values (first 5 rows shown)

Data formats of each column are checked and corrected using pandas.astype() method.

```
Planning Area     object
Region            object
Area             float64
Population        object
Density           object
dtype: object
```

Data formats before format correction

```
Planning Area     object
Region            object
Area             float64
Population       float64
Density          float64
dtype: object
```

Data formats after format correction

3.1.2 Data acquisition and wrangling for coordinates for each Planning Area

Using geopy 2.0 as a geocoding service, the coordinates of each Planning Area is obtained. The Nominatim geocoder is instantiated. Nominatim geocoder uses OpenStreetMap data, which is open data and is free to use. The coordinates of Singapore is obtained using the geocoder, and these will be used as the starting coordinates in map visualisation.

Coordinates of Singapore are 1.357107, 103.8194992

The coordinates of each Planning Area is obtained by geocoding a pandas dataframe with geopy, with rate-limiting taken into account. This is because a large number of dataframe rows might produce a significant amount of geocoding requests to a geocoding service, which might be throttled by the service. Thus, a RateLimiter() function is used to add delays (1 second) between geocoding calls to reduce the load on the geocoding service. A progress bar is shown using the tqdm() function.

A string suffix "suburb, Singapore" is concatenated to each Planning Area in the search query, to specify to the geocoding service that each search performed should return the coordinates of a suburb, not a landmark or train station.

Two new columns are added to the dataframe ("Latitude" and "Longitude"), which are filled with NaN values. These columns are then filled with coordinates obtained from the geocoding service using a for loop and pandas.at[] method. The first 5 results are shown in Table 3.

	Planning Area	Region	Area	Population	Density	Latitude	Longitude
0	Ang Mo Kio	North-East	13.94	163950.0	13400.00	1.369842	103.846609
1	Bedok	East	21.69	279380.0	13000.00	1.325670	103.931471
2	Bishan	Central	7.62	88010.0	12000.00	1.351912	103.848971
3	Boon Lay	West	8.23	30.0	3.60	1.345640	103.711802
4	Bukit Batok	West	11.13	153740.0	14000.00	1.348283	103.749019

Table 3: Latitude and Longitudes obtained using geocoding service (first 5 rows shown)

Using the coordinates from Table 3, a map of all Planning Areas are plotted.

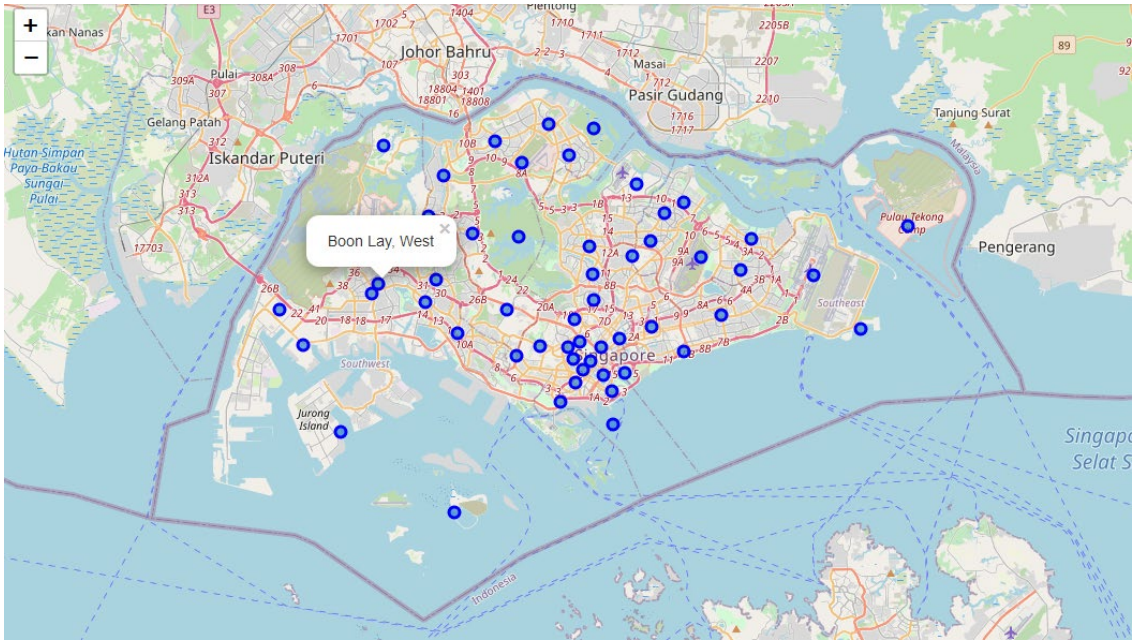

Figure 1: Map showing all Planning Areas in Singapore. The "Boon Lay, West" map marker is in close proximity to the "Jurong West, West" map marker.

Based on Figure 1 above, it appears that Boon Lay is closely situated to Jurong West. A manual web query on OpenStreetMap Nominatim: Search using the string "Boon Lay suburb, Singapore" returns two results with different coordinates. The first result points to Boon Lay neighbourhood, while the second result refers to Boon Lay Planning Area, which was verified as correct because it has polygonal coverage.

Returning to Jupyter Notebook, the geocoding service is run with the same search query, but with the argument *exactly_one=False*. The query returned two results with different coordinates. The second pair of coordinates is chosen, and re-assigned to the dataframe. The map is re-plotted to show the change in coordinates for Boon Lay Planning Area.

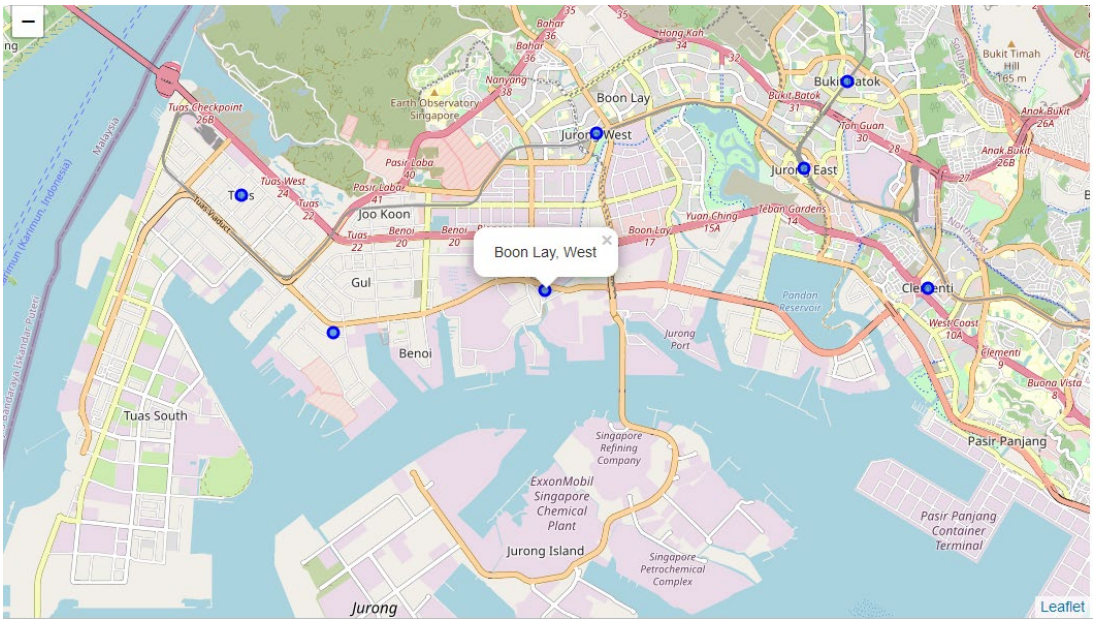

Figure 2: Map showing corrected coordinates of Boon Lay Planning Area

Before proceeding to data acquisition using FourSquare API, the "Search Radius" was calculated for each Planning Area. This changes the search radius in the FourSquare API request according to the size of each Planning Area, rather than setting a fixed search radius for all Planning Areas. The assumption here is that each Planning Area has a circular shape, which is a gross oversimplification. The "Search Radius" is calculated using the following equation:

$$Search\ Radius\ (m) = \sqrt{\frac{Area}{\pi}} \times 1000$$

Equation 1: Formula to calculate search radius

Using the pandas.apply(lambda x:) function, the "Search Radius" was calculated for each Planning Area, and added to the dataframe.

3.1.3 Data acquisition and wrangling on top venues in each Planning Area

A Foursquare developer account is registered by the author, and credentials are obtained. However, Foursquare API has a limit of 950 Regular API Calls per day for Sandbox Tier Accounts. By upgrading to a Personal Tier developer account, 99,500 Regular API Calls can me made per day. However, an hourly rate limit of 5,000 requests is still imposed.

The Foursquare API is used to get the top 100 venues in each Planning Area, with search radius individualised to each Planning Area. API calls are made to Foursquare by passing the coordinates and search radius of each Planning Area in a Python loop. Foursquare returns the venue data in JSON format. Using the json() function, the venue names, venue latitude, venue longitudes, and venue categories, are extracted and appended into a list. The list is then converted into a pandas dataframe using the pandas.DataFrame() method. The first 5 rows of the dataframe are shown below, which contains 3749 rows.

	Planning Area	PA Latitude	PA Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ang Mo Kio	1.369842	103.846609	Bishan - Ang Mo Kio Park	1.362219	103.846250	Park
1	Ang Mo Kio	1.369842	103.846609	Aramsa ~ The Garden Spa	1.362292	103.847602	Spa
2	Ang Mo Kio	1.369842	103.846609	Face Ban Mian 非板面 (Ang Mo Kio)	1.372031	103.847504	Noodle House
3	Ang Mo Kio	1.369842	103.846609	Old Chang Kee	1.369094	103.848389	Snack Place
4	Ang Mo Kio	1.369842	103.846609	FairPrice Xtra	1.369279	103.848886	Supermarket

Table 4: First 5 venues from all Planning Areas

This dataframe contains 324 unique venue categories. The pandas.drop_duplicates() method is used to remove duplicates due to overlapping search results from the Foursquare API request. The number of duplicated venues removed is 1014.

3.2 Exploratory Analysis and Normalising Data

After removing duplicates, the data frame is grouped by "Venue Category", and values for each row were counted using the pandas.count() method. The values are sorted by descending order. Table 5 shows the top 10 Venue Categories in all Planning Areas.

	Venue Category	Count
0	Chinese Restaurant	167
1	Food Court	166
2	Café	110
3	Coffee Shop	110
4	Asian Restaurant	96
5	Hotel	79
6	Japanese Restaurant	71
7	Park	65
8	Indian Restaurant	60
9	Bakery	59

Table 5: Top 10 Venue Categories in all Planning Areas

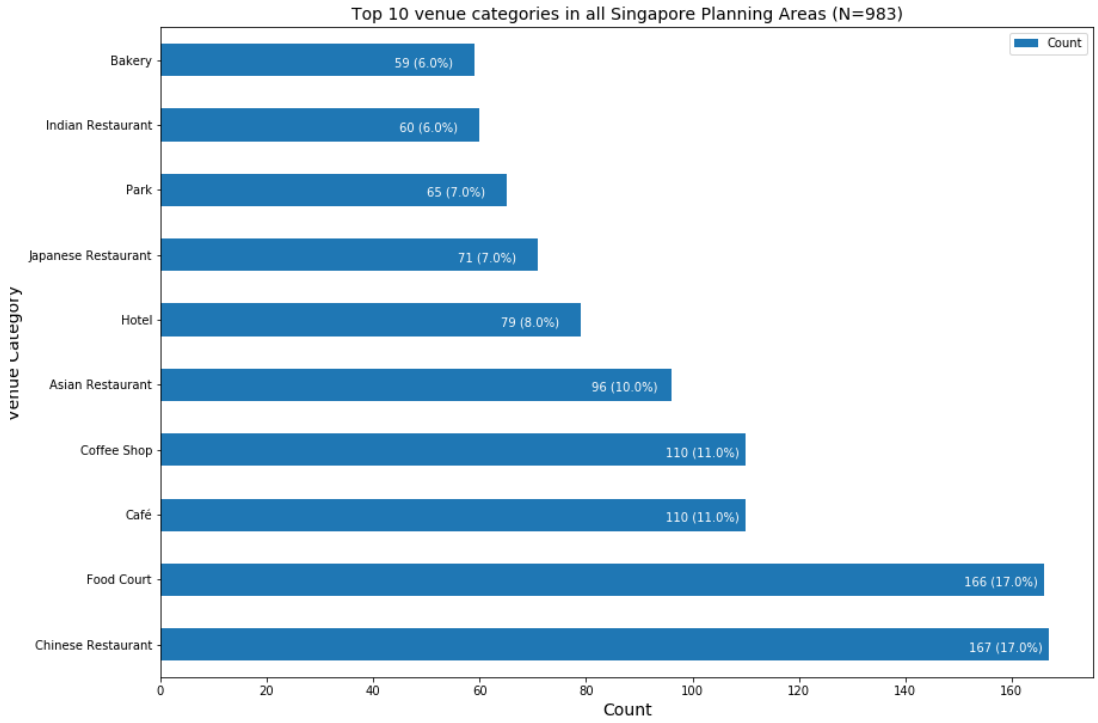

Figure 3: Horizontal bar chart of top 10 Venue Categories

For further analysis, one-hot encoding is performed on "Venue Category" using pandas.get_dummies() method on the dataframe in Table 4. This converts categorical variables into dummy/indicator variables. The resulting dataframe contains 323 columns. These dummy variables are grouped by "Planning Area" and are summed to produce the total number of all 323 Venue Categories in each Planning Area.

Pearson correlations are computed for each "Venue Category" against "Density" using pandas.corrwith() method. There correlations were sorted in descending order, and the results are shown Table 6 below.

The "Venue Category" of interest, Food Court, is selected from the summed one-hot encoded dataframe. The columns "Planning Area" and "Density" are added, and visualisations are performed to analyse data distribution.

```
Pearson correlation of top 5 venue categories with population density:
Food Court            0.641849
Coffee Shop           0.532433
Pool                  0.521503
Chinese Restaurant    0.472719
Market                0.448904
dtype: float64

Pearson correlation of bottom 5 venue categories with population density:
Airport           -0.218587
Theater           -0.232275
Harbor / Marina   -0.274835
Boat or Ferry     -0.289389
Smoke Shop        -0.323860
dtype: float64
```

Table 6: Pearson correlations of top 5 and bottom 5 venue categories with population density

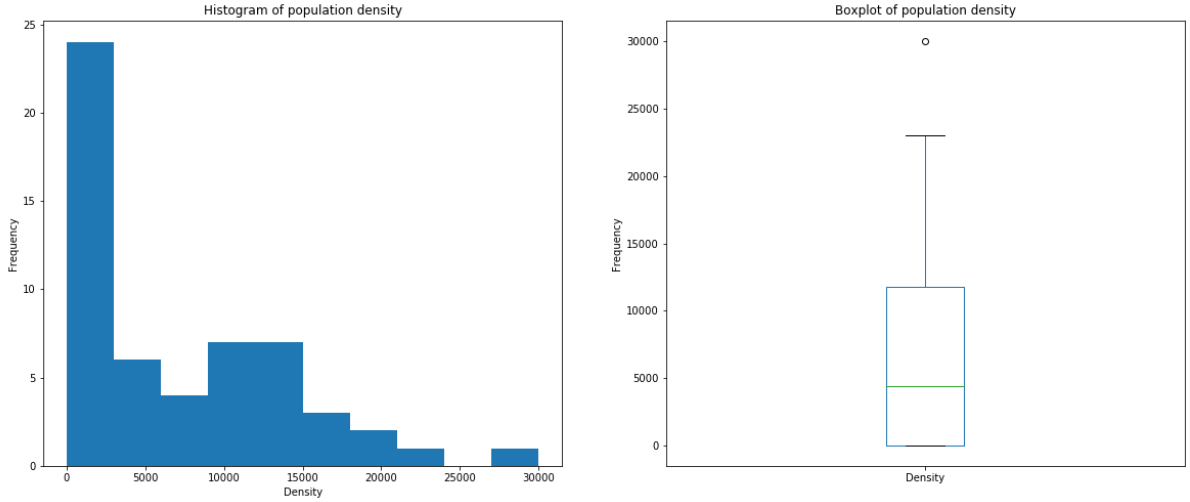

Figure 4: Histogram and Boxplot of Density

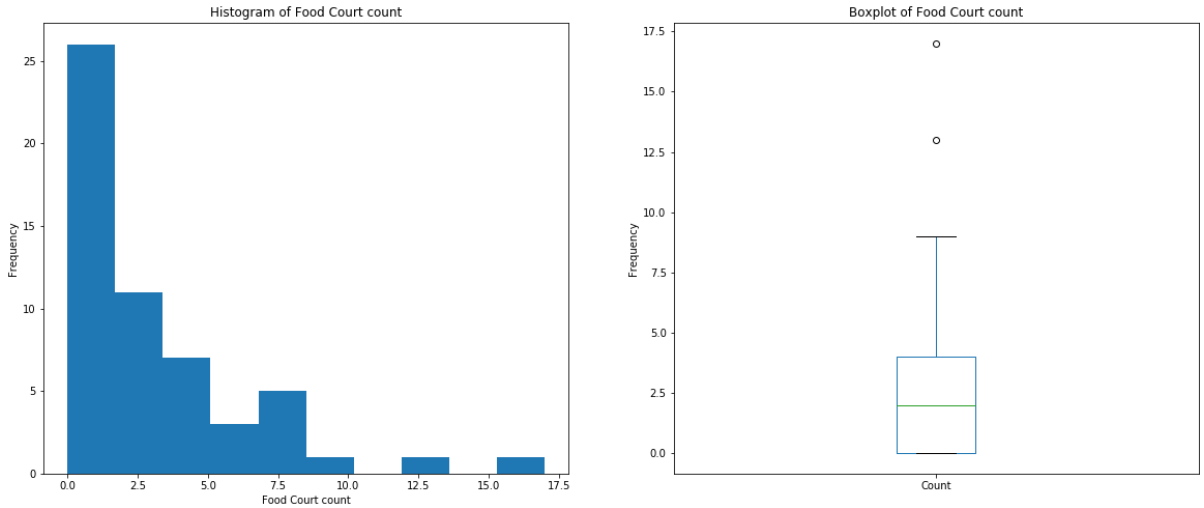

Figure 5: Histogram and Boxplot of Food Court count

Based on Figure 4 and Figure 5, the distributions of "Density" and "Food Court" are right-skewed. Before proceeding with further analysis, data standardisation needs to be done to produce meaningful comparisons between these two variables. Using the StandardScaler() function from

sklearn.preprocessing library, these variables are fitted and transformed into new values, with mean = 0 and standard deviation = 1. These standardised values are stored in a new data frame using pandas.DataFrame() method, and a scatter plot with regression line is plotted using seaborn.regplot() function. The scatter plot is shown in Figure 6.

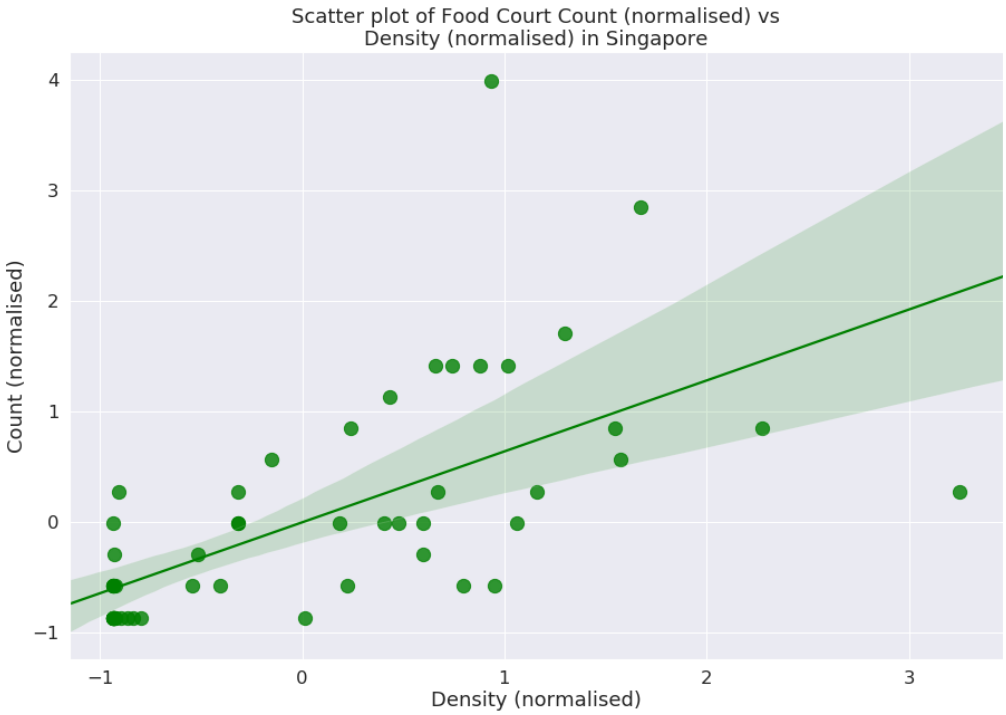

Figure 6: Scatter plot of Food Court Count (normalised) vs Density (normalised) with R=0.64

3.3 Clustering with Machine Learning

Two types of clustering machine learning algorithms are used in this project, DBSCAN algorithm and k-nearest neighbour algorithm. The algorithms are run on the standardised Food Court dataset, and the results are plotted and compared on a Cartesian plane to determine which clustering algorithm produces better results.

3.3.1 DBSCAN clustering

DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. This is a common clustering algorithm technique, which works based on density of object. This algorithm can find any arbitrary shape cluster without getting affected by noise.

The standardised values were fitted using the DBSCAN() function from sklearn.cluster library, with epsilon = 0.35 and minimum samples =3. The generated labels are stored in a variable. An array of Booleans is created using the labels. Using a for loop, Boolean array and generated labels, the DBSCAN clusters are visualised in Figure 7 below.

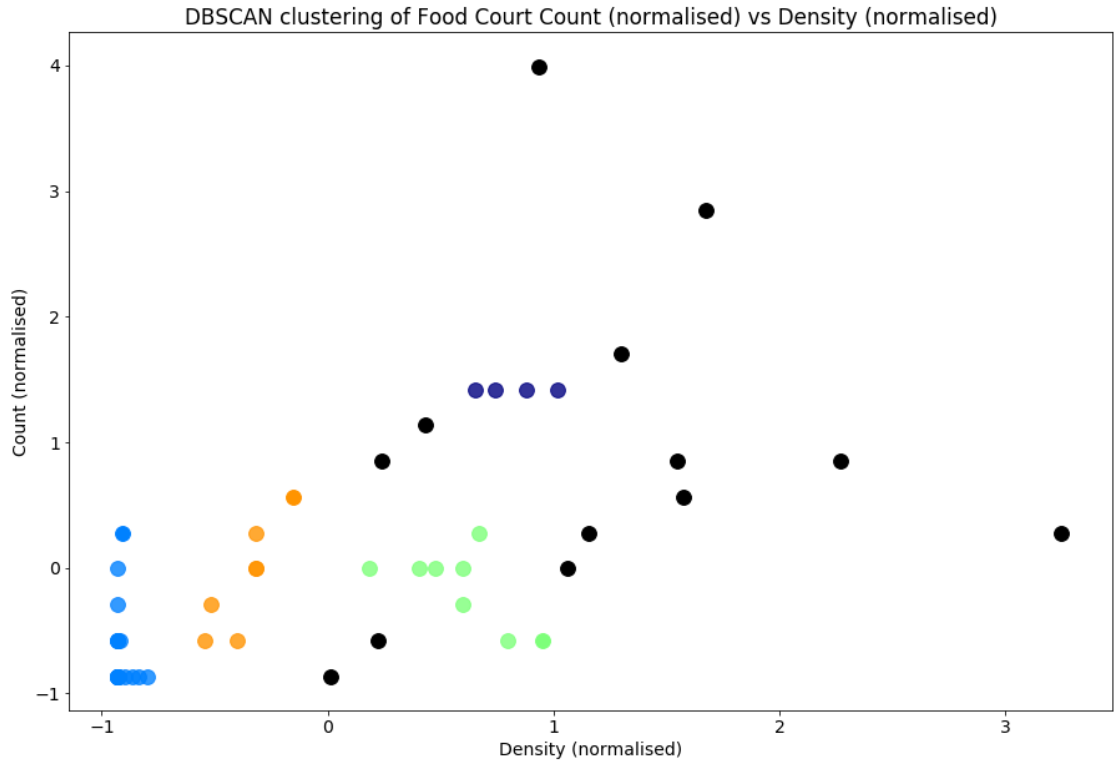

Figure 7: DBSCAN clusters

Based on Figure 7, the DBSCAN algorithm generated 4 clusters (highlighted by the blue, orange, light green, and purple dots), while highlighting the remaining dots as outliers. While the coloured dots are appropriately clustered based on density, too much information (13 points of data) has been left out and labelled as outliers. A different clustering algorithm needs to be considered to cluster the above data.

3.3.2 K-Means clustering algorithm

K-Means clustering is a type of partition clustering that divides data into K non-overlapping subsets or clusters without any cluster internal structure or labels. It is an unsupervised algorithm. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar. K-Means tries to minimize the intra-cluster distances and maximize the inter-cluster distances.

The standardised values were fitted using the KMeans() function from sklearn.cluster library, with number of clusters = 4. The generated labels and cluster centres are stored in variables. Using a for loop, generated labels, and cluster centres, the K-Means clusters are visualised in Figure 8 above.

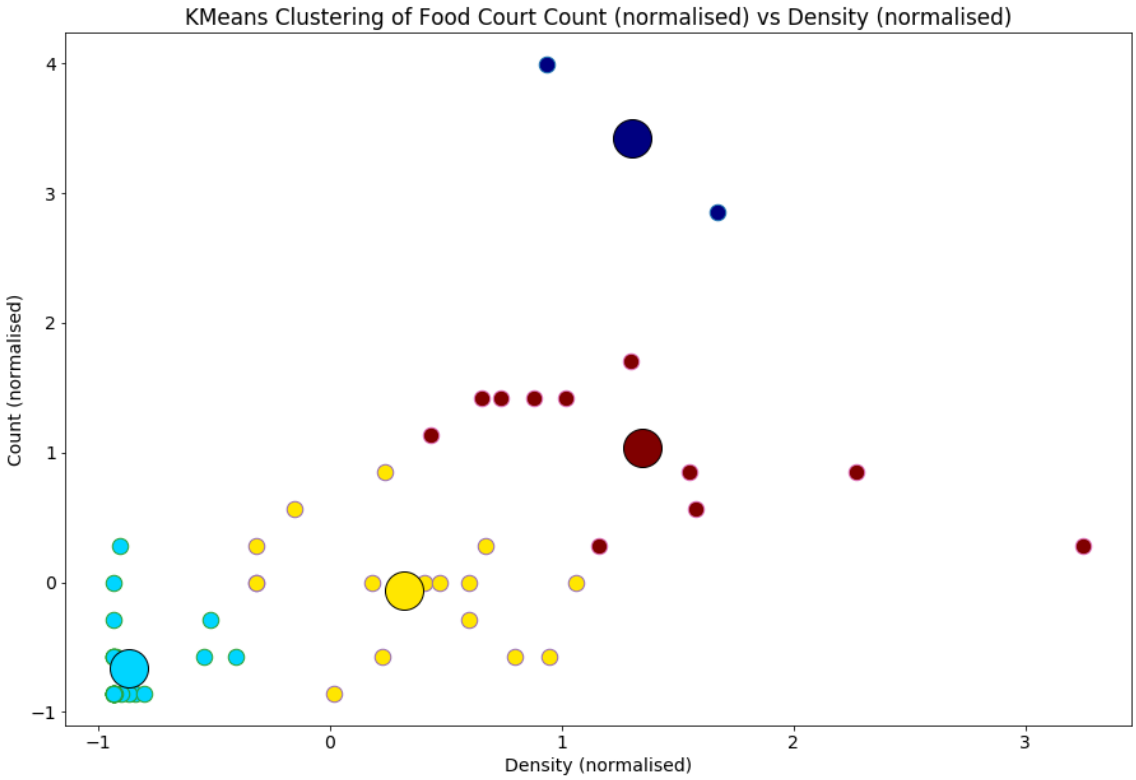

Figure 8: K-Means clustering. Cluster 0: dark blue circles, Cluster 1: light blue circles, Cluster 2: yellow circles, Cluster 3: maroon circles. Large circles represent centroids of their respective clusters.

Based on Figure 8, all data points are given cluster labels, including outliers. The K-Means clustering algorithm generates better clustering than DBSCAN clustering algorithm. Table 7 below shows the clusters generated by K-Means algorithm.

```
1    26
2    16
3    11
0     2
Name: Cluster Label, dtype: int64
```

Table 7: Cluster labels generated by K-Means algorithm

We will proceed by inserting the cluster labels generated from K-Means algorithm into the dataframe. The combined dataframe now contains the columns: “Planning Area”, “Density”, “Count”, “Latitude”, “Longitude”, and “Cluster Label”.

4. Results

Using the Map() function from folium library, all Planning Areas are plotted on the map, with different coloured map markers according to their cluster labels.

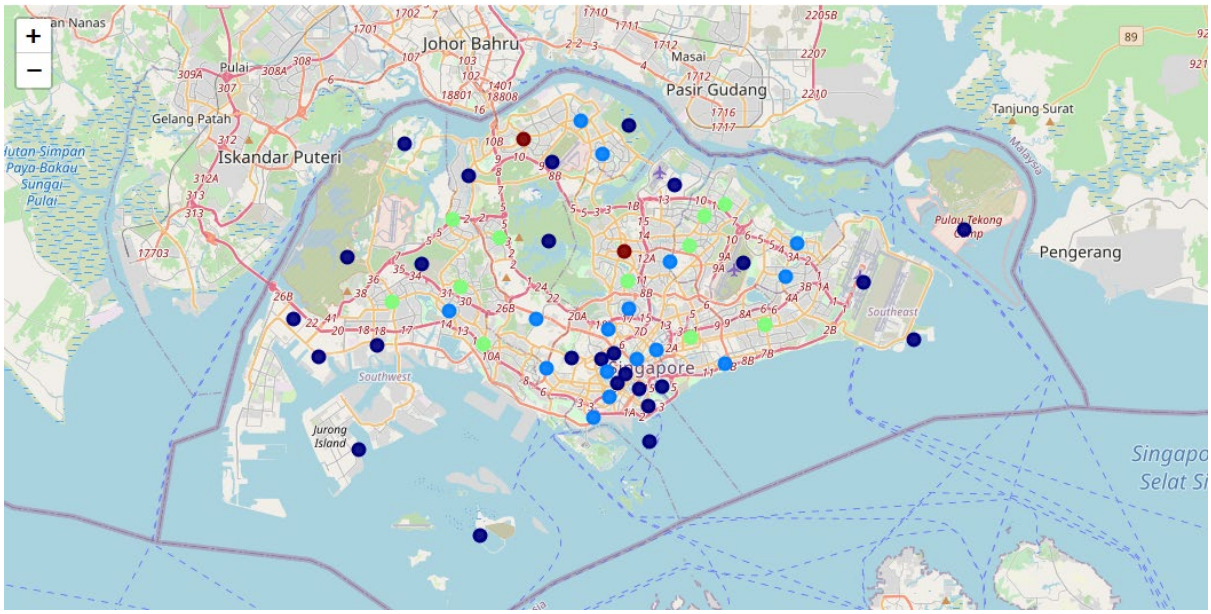

Figure 9: Map of all Planning Areas with colour-coded clusters. Brown map markers (cluster 0), dark blue map markers (cluster 1), light blue map markers (cluster 2), light green map markers (cluster 3).

The merged dataframe can be grouped by clusters, and a summary of each cluster is obtained for further analysis.

Cluster 0 (n=2) are areas with high numbers of Food Courts per Planning Area (range: 13 - 17) with moderate Population Density (median density=16050.0).

	Planning Area	Density	Count	Cluster Label
0	Ang Mo Kio	13400.0	17	0
53	Woodlands	18700.0	13	0

Table 8: Cluster 0 attributes, sorted by Count in descending order

Cluster 1 (n=26) are places with low numbers of Food Courts per Planning Area (range: 0 - 4) with low Population Density (median density=6.75).

	Planning Area	Density	Count	Cluster Label
20	Mandai	180.20	4	1
50	Tuas	2.30	3	1
42	Singapore River	3000.00	2	1
32	Pioneer	8.30	2	1
3	Boon Lay	3.60	1	1

Table 9:Cluster 1 attributes (first 5 rows shown), sorted by Count in descending order

Cluster 2 (n=16) are areas with low numbers of Food Courts per Planning Area (range: 0 - 6) with moderate Population Density (median density=9000.0).

	Planning Area	Density	Count	Cluster Label
38	Sembawang	8400.0	6	2
27	Novena	5600.0	5	2
16	Jurong East	4400.0	4	2
40	Serangoon	11500.0	4	2
7	Bukit Timah	4400.0	3	2

Table 10: Cluster 2 attributes (first 5 rows shown), sorted by Count in descending order

Cluster 3 (n=11) are areas with medium number of Food Courts per Planning Area (range: 4 - 9) with high Population Density (median density=15000.0).

	Planning Area	Density	Count	Cluster Label
15	Hougang	16000.0	9	3
1	Bedok	13000.0	8	3
2	Bishan	12000.0	8	3
4	Bukit Batok	14000.0	8	3
14	Geylang	11400.0	8	3

Table 11: Cluster 3 attributes (first 5 rows shown), sorted by Count in descending order

5. Discussion

When acquiring data from Wikipedia, the author noticed that most places have low Density values or 0 (Table 2). These values are verified with the original source material. This can be explained that these Planning Areas are not habitable/ not gazetted for residential living. For example, the Central Water Catchment Planning Area is one of the main water catchment areas of the country, and acts as a natural reservoir. Nearby islands, such as Western Islands, Southern Islands, and North-Eastern Islands are not as populous as the mainland. Orchard Planning Area, which has a population of 990 persons, is a tourist attraction area, and is not expected to have many residential dwellings. The Downtown Core Planning Area is the main commercial area of the country, and only has a population density of 680 persons/km^2, compared with the country's average population density of 7,804 persons/km^2. It must be borne in mind that the reported population demographics refers to residential population, and were based on a survey conducted in 2015. These figures do not do not include approximately 1.6 million non-permanent residents of Singapore.

The non-uniform distribution of residents throughout the country has resulted in a right-skewed histogram, as seen in Figure 4. Most places have low density, as described in the paragraph above. The boxplot highlights a single outlier, which refers to the Choa Chu Kang Planning Area, with a density of 30,000 persons/km^2. Because of the skewness, values must be standardised prior to further analysis/ machine learning to produce meaningful results.

Table 5 shows the top 10 venue categories returned by the Foursquare API in all Planning Areas throughout Singapore. It is noted that 8 out of 10 categories are food-related categories, while the remaining two are Hotel and Park. It can be concluded that the top businesses in all Planning Areas combined in Singapore are food-related.

The values in Table 6 show the top 5 highest and bottom 5 lowest Pearson correlations of all venue categories with population density. The highest value in the list is Food Court, with R value of 0.64. This shows a moderately positive relationship between Food Court and Density. This is confirmed by the regression line drawn in the scatter plot in Figure 6. As the population density increases, the Food Court number increases.

Other venue categories with moderate positive correlation with density include Coffee Shops, Pools, Chinese Restaurants, and Markets. Conversely, venue categories with negative correlation with density include Smoke Shop, Boat or Ferry, Harbor/Marina, Theater, and Airport. It is understandable that places with Boat or Ferry, Harbor/Marina, and Airport have lower population densities, as these businesses/landmarks require larger land areas for operations.

Based on the results of the study and the map plotted in Figure 9, local municipalities/ entrepreneurs should consider opening Food Courts in Cluster 3, i.e. areas with high population density and medium number of Food Courts. For example, the two outliers with highest density in Cluster 3 in Figure 8, refer to Choa Chu Kang and Sengkang. These places with high population density may benefit the population from more strategically placed Food Courts in those areas.

	Planning Area	Density	Count	Cluster Label
11	Choa Chu Kang	30000.0	4	3
39	Sengkang	23000.0	6	3

Table 12: Planning Areas with top 2 highest densities in Cluster 3, with moderate numbers of Food Courts, sorted by Density in descending order

5.1 Limitations of the study

1. Population and density data obtained for each Planning Area only refer to residential density. It does not reflect urban density in non-residential areas. The data obtained also do not reflect the latest population figures, as the demographic survey was conducted in 2015. These figures also do not include approximately 1.6 million non-permanent residents of Singapore.
2. Getting correct coordinates from for each Planning Area. By using the Nominatim geocoding service, we are relying on the accuracy and precision of results returned by the service. One challenge is the coordinates obtained for the Boon Lay Planning Area. According to the Wikipedia article Boon Lay Planning Area, the Boon Lay Planning Area is different from Boon Lay, which is a residential neighbourhood located in the adjacent Jurong West Planning Area.
3. In order to calculate the search radius for each Planning Area, an assumption is made that the each Planning Area is circular in shape. This is done to individualise a search radius for each Planning Area, based on their size.
4. Results returned by the Foursquare API could contain duplicates.

5.2 Strategies taken to overcome limitations

1. Demographic data are manually verified with source material from Singapore Department of Statistics website.
2. Search queries using the Nominatim geocoder were suffixed with the string “suburb, Singapore” to specify to the geocoding service that each search performed should return the coordinates of a suburb, not a landmark or train station.
3. Duplicate results were dropped using the pandas.drop_duplicates() method.
4. Non-normally distributed data were normalised prior to further analysis.

5.3 Recommendations for future studies

Future studies should consider including non-residential population density from urban areas, such as Downtown Core and Orchard. However, this recommendation has many challenges, as demographic values in non-residential urban areas suffer from high population turnovers from local and foreign visitors/tourists, and cannot be assigned to a single value. Another suggestion is using anonymised foot-traffic data as an indicator of density in urban areas.

Future studies should consider replicating the results of this study in other countries/ cities, especially the correlation of population density with the types of businesses surrounding that population.

6. Conclusion

In this Capstone project, we have extracted information from the internet, get coordinates for each area, and used Foursquare API to get venues surrounding every area. Data is wrangled, correctly formatted, and normalised before further data analysis was done. Exploratory analysis and visualisations are done to gain a better understanding of the data. Finally, machine learning algorithms are used to cluster data.

With reference to the business problems listed under Introduction, a moderately positive relationship exists between population density and food courts that surround that population. Local municipalities / entrepreneurs should consider opening additional food courts in Cluster 3, i.e. areas with high population density and medium number of Food Courts.

References

GeoPy. (2018). *Welcome to GeoPy's documentation! — GeoPy 1.21.0 documentation*. Retrieved from GeoPy: https://geopy.readthedocs.io/en/stable/

OpenStreetMap. (10 May, 2020). *OpenStreetMap Nominatim: Search*. Retrieved from OpenStreetMap: https://nominatim.openstreetmap.org/

pandas development team. (2014). *pandas documentation — pandas 1.0.3 documentation*. Retrieved from pandas: https://pandas.pydata.org/docs/

Singapore Department of Statistics. (25 September, 2019). *Geographic Distribution - Latest Data*. Retrieved from Singapore Department of Statistics: https://www.singstat.gov.sg/find-data/search-by-theme/population/geographic-distribution/latest-data

Wikipedia. (17 January, 2020). *Boon Lay Planning Area*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Boon_Lay_Planning_Area

Wikipedia. (7 May, 2020). *Planning Areas of Singapore*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore

Wikipedia. (10 May, 2020). *Singapore*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Singapore