# Segmenting and Clustering Suburbs – Gold Coast City

Tyla Ono

1st July 2020

## 1. Introduction

Gold Coast City is the sixth largest city in Australia and is one of the country's fastest growing cities, averaging a 4.4 per cent growth in Gross Region Product over the past five years. Gold Coast's highly multicultural population currently sits at 590,000 people and is forecasted to reach over one-million people by the year 2050. Given the region's strong economic growth and substantial expected increase in population, the Gold Coast will likely attract many future businesses within the hospitality industry including food and beverages, tourism, lodging and recreation. The location of the venue is a critical factor for the success of a business within hospitality. This project aims to segment and cluster the suburbs in Gold Coast city based on the most common venues within each suburb. The results of segmentation and clustering, as well as population data of each suburb on the Gold Coast will be used to recommend optimal locations for future businesses in the hospitality sector.

## 2. Data Acquisition and Processing

The investigation required collecting data on the most common venues of given suburbs on the Gold Coast and the population in each suburb.

The most common venues of each suburb were found using the Foursquare API which provides location-based data on information such as venues, users and photos. The API requires the latitudinal and longitudinal coordinates of the area of interest, the number of venues returned by Foursquare API and the radius of interest as inputs. Therefore, the geographical coordinates of each Gold Coast suburb had to be found to use the Foursquare API. This was achieved by scraping the list of suburbs of the Gold Coast, as well as their corresponding postal codes and population from Wikipedia [1]. The table also contained localities in addition to suburbs and were filtered out such that only Gold Coast suburbs remained as shown below in Figure 1.

|   | Suburb | Type | Postcode | 2016 Population |   |   | Suburb | Type | Postcode | 2016 Population |
|---|--------|------|----------|-----------------|---|---|--------|------|----------|-----------------|
| 0 | Advancetown | Locality | 4211 | 482 |   | 0 | Arundel | Suburb | 4214 | 10246 |
| 1 | Alberton | Locality | 4207 | 590 |   | 1 | Ashmore | Suburb | 4214 | 11910 |
| 2 | Arundel | Suburb | 4214 | 10246 |   | 2 | Benowa | Suburb | 4217 | 8741 |
| 3 | Ashmore | Suburb | 4214 | 11910 |   | 3 | Biggera Waters | Suburb | 4216 | 8534 |

*Figure 1 Removing localities from the table containing data on the name of suburbs, postcode and 2016 population for Gold Coast city*

It can be observed that there are suburbs with duplicate postcodes such as Arundel and Ashmore. These will be grouped together and considered as one neighbourhood, since the geographical coordinates will be found based on the postcode and will result in identical locations. Therefore, suburbs with identical postcodes were grouped together and the population of grouped suburbs were summed together as shown in Figure 2.

| | Suburb | Postcode | 2016 Population |
|---|---|---|---|
| 0 | Arundel, Ashmore, Molendinar, Parkwood | 4214 | 37233 |
| 1 | Benowa, Bundall, Main Beach, Surfers Paradise | 4217 | 40836 |
| 2 | Biggera Waters, Coombabah, Hollywell, Paradise... | 4216 | 37391 |
| 3 | Bilinga, Coolangatta | 4225 | 7752 |
| 4 | Broadbeach, Broadbeach Waters, Mermaid Beach, ... | 4218 | 31871 |

*Figure 2 Dataframe of Gold Coast suburbs merged by identical postcodes (first 5 rows)*

The format of the dataframe shown in Figure 2 enabled the computation of the geographical coordinates for each neighbourhood using ArcGIS in conjunction with Geocoder. The geographical coordinates were computed and stored into the dataframe by looping through each neighbourhood and using the respective postcode as the input, as shown in Figure 3 below.

| | Suburb | Postcode | 2016 Population | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | Arundel, Ashmore, Molendinar, Parkwood | 4214 | 37233 | -27.961448 | 153.369775 |
| 1 | Benowa, Bundall, Main Beach, Surfers Paradise | 4217 | 40836 | -28.010320 | 153.388423 |
| 2 | Biggera Waters, Coombabah, Hollywell, Paradise... | 4216 | 37391 | -27.916455 | 153.376175 |
| 3 | Bilinga, Coolangatta | 4225 | 7752 | -28.164080 | 153.513242 |
| 4 | Broadbeach, Broadbeach Waters, Mermaid Beach, ... | 4218 | 31871 | -28.045100 | 153.412372 |
| 5 | Burleigh Heads, Burleigh Waters, Miami | 4220 | 31230 | -28.090163 | 153.431955 |
| 6 | Carrara, Highland Park, Nerang, Pacific Pines | 4211 | 52255 | -28.003917 | 153.273565 |
| 7 | Clear Island Waters, Merrimac, Robina | 4226 | 34297 | -28.074371 | 153.396415 |
| 8 | Coomera, Pimpama, Upper Coomera | 4209 | 47977 | -27.850325 | 153.294098 |
| 9 | Currumbin, Currumbin Waters | 4223 | 12041 | -28.187680 | 153.420190 |
| 10 | Elanora, Palm Beach | 4221 | 26799 | -28.118788 | 153.455825 |
| 11 | Helensvale, Hope Island | 4212 | 28048 | -27.909250 | 153.330702 |
| 12 | Jacobs Well, Ormeau | 4208 | 16299 | -27.785815 | 153.344017 |
| 13 | Labrador, Southport | 4215 | 50169 | -27.964902 | 153.398770 |
| 14 | Mudgeeraba, Neranwood, Tallai, Worongary | 4213 | 23454 | -28.124356 | 153.319140 |
| 15 | Oxenford | 4210 | 11842 | -27.938874 | 153.257630 |
| 16 | Reedy Creek, Varsity Lakes | 4227 | 21685 | -28.115023 | 153.403525 |
| 17 | Tallebudgera | 4228 | 3667 | -28.174760 | 153.377182 |
| 18 | Tugun | 4224 | 6588 | -28.148840 | 153.491323 |
| 19 | Yatala | 4207 | 1312 | -27.813417 | 153.203113 |

*Figure 3 Dataframe of each Gold Coast neighbourhood and corresponding geographical coordinates (complete dataframe)*

The location of each Gold Coast neighbourhood was visualized onto a map based on the computed geographical coordinates using the Folium library on Python, as shown below in *Figure 4*.
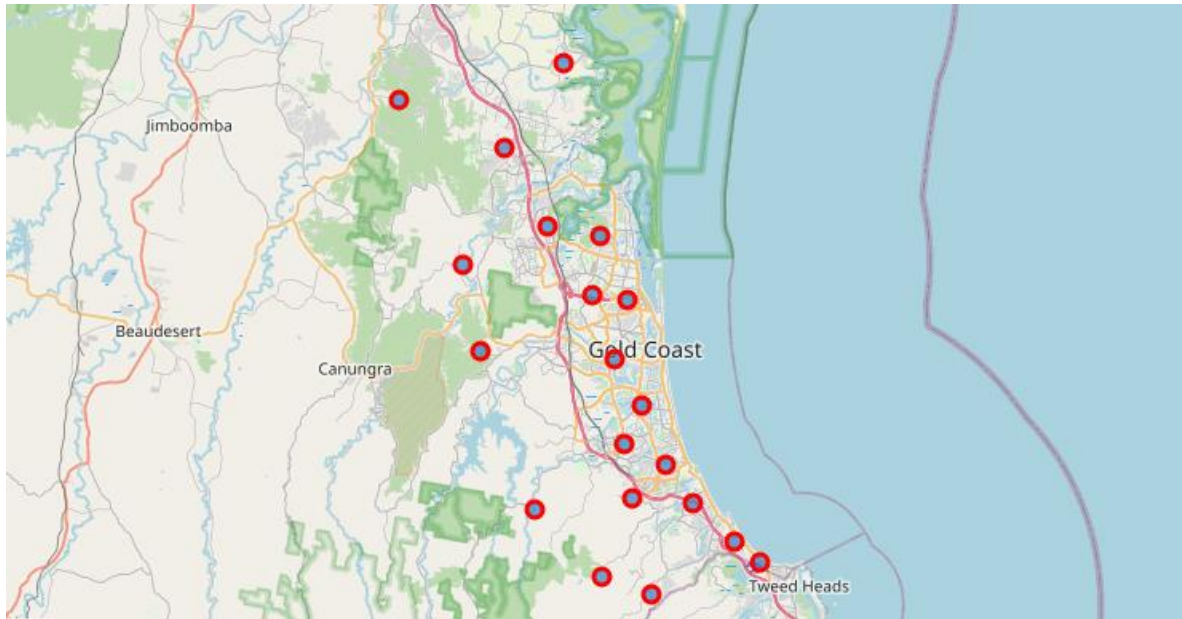
Figure 4 Mapping the geographical location of each neighbourhood on the Gold Coast based on the computed coordinates

The Foursquare API was used to find the top 100 most common venues in each neighbourhood within a one-kilometre radius. One-hot encoding and grouping was implemented to obtain the frequency of each venue. The top 10 most common venues were summarised in a dataframe in order for each neighbourhood as shown below in *Figure 5*.

| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Arundel, Ashmore, Molendinar, Parkwood | Harbor / Marina | Bakery | Video Game Store | Chinese Restaurant | Gift Shop | Fishing Spot | Fast Food Restaurant | Dive Shop | Discount Store | Construction & Landscaping |
| 1 | Benowa, Bundall, Main Beach, Surfers Paradise | Café | Video Game Store | Bakery | Sushi Restaurant | Fast Food Restaurant | Coffee Shop | Chinese Restaurant | Botanical Garden | Burger Joint | Sandwich Place |
| 2 | Biggera Waters, Coombabah, Hollywell, Paradise... | Airport | Park | Paper / Office Supplies Store | Café | Fishing Spot | Fast Food Restaurant | Dive Shop | Discount Store | Construction & Landscaping | Coffee Shop |
| 3 | Bilinga, Coolangatta | Café | Airport | Hostel | Airport Lounge | Airport Terminal | Fast Food Restaurant | Sandwich Place | Resort | Burger Joint | Coffee Shop |
| 4 | Broadbeach, Broadbeach Waters, Mermaid Beach, ... | Supermarket | Shopping Mall | Bakery | Grocery Store | Caribbean Restaurant | Smoke Shop | Fast Food Restaurant | Discount Store | Breakfast Spot | Malay Restaurant |
| 5 | Clear Island Waters, Merrimac, Robina | Gift Shop | Business Service | Video Game Store | Caribbean Restaurant | Fishing Spot | Fast Food Restaurant | Dive Shop | Discount Store | Construction & Landscaping | Coffee Shop |
| 6 | Coomera, Pimpama, Upper Coomera | Supermarket | Gym | Gym / Fitness Center | Health & Beauty Service | Video Game Store | Caribbean Restaurant | Fast Food Restaurant | Dive Shop | Discount Store | Construction & Landscaping |
| 7 | Currumbin, Currumbin Waters | Café | Caribbean Restaurant | Gift Shop | Fishing Spot | Fast Food Restaurant | Dive Shop | Discount Store | Construction & Landscaping | Coffee Shop | Chinese Restaurant |
| 8 | Elanora, Palm Beach | Gourmet Shop | Supermarket | Shopping Mall | Pizza Place | Coffee Shop | Café | Fishing Spot | Fast Food Restaurant | Dive Shop | Discount Store |
| 9 | Helensvale, Hope Island | Golf Course | Liquor Store | Gourmet Shop | Gift Shop | Fishing Spot | Fast Food Restaurant | Dive Shop | Discount Store | Construction & Landscaping | Coffee Shop |
| 10 | Labrador, Southport | Fishing Spot | Dive Shop | Construction & Landscaping | Business Service | Video Game Store | Caribbean Restaurant | Gift Shop | Fast Food Restaurant | Discount Store | Coffee Shop |
| 11 | Reedy Creek, Varsity Lakes | Airport | Home Service | Playground | Caribbean Restaurant | Fishing Spot | Fast Food Restaurant | Dive Shop | Discount Store | Construction & Landscaping | Coffee Shop |
| 12 | Tugun | Athletics & Sports | Video Game Store | Caribbean Restaurant | Gift Shop | Fishing Spot | Fast Food Restaurant | Dive Shop | Discount Store | Construction & Landscaping | Coffee Shop |

Figure 5 Dataframe summarising the top 10 most common venues in each Gold Coast neighbourhood (ordered)

# 3. Exploratory Data Analysis

Clustering is a common machine learning technique used to gain an intuition with respect to the structure of the dataset and groups observations with similar features into homogenous subsets. Therefore, Gold Coast neighbourhoods that have similar types of venues can be identified and segmented through the use of clustering algorithms.

## 3.1 K-means clustering

K-means is a simple clustering technique that employs a centroid-based algorithm to minimize the distance between the points within a cluster. One of the main challenges of implementing K-means clustering is the selection of the optimal number of clusters, *K*. A common technique to determine the *K* for the K-means algorithm is to plot an 'elbow curve', where the horizonal axis represents the number of clusters and the vertical axis plots an evaluation metric. It is an iterative process of using different values of *K* and computing an evaluation metric such as the inertia. The optimal cluster value is where a sharp decrease in inertia can be observed on the elbow curve and the rate of reduction in inertia becomes constant as the number of cluster increases. An elbow curve was plotted using the K-means algorithm on the dataframe containing the top ten most common venues in each Gold Coast neighbourhood as shown below in Figure 6.
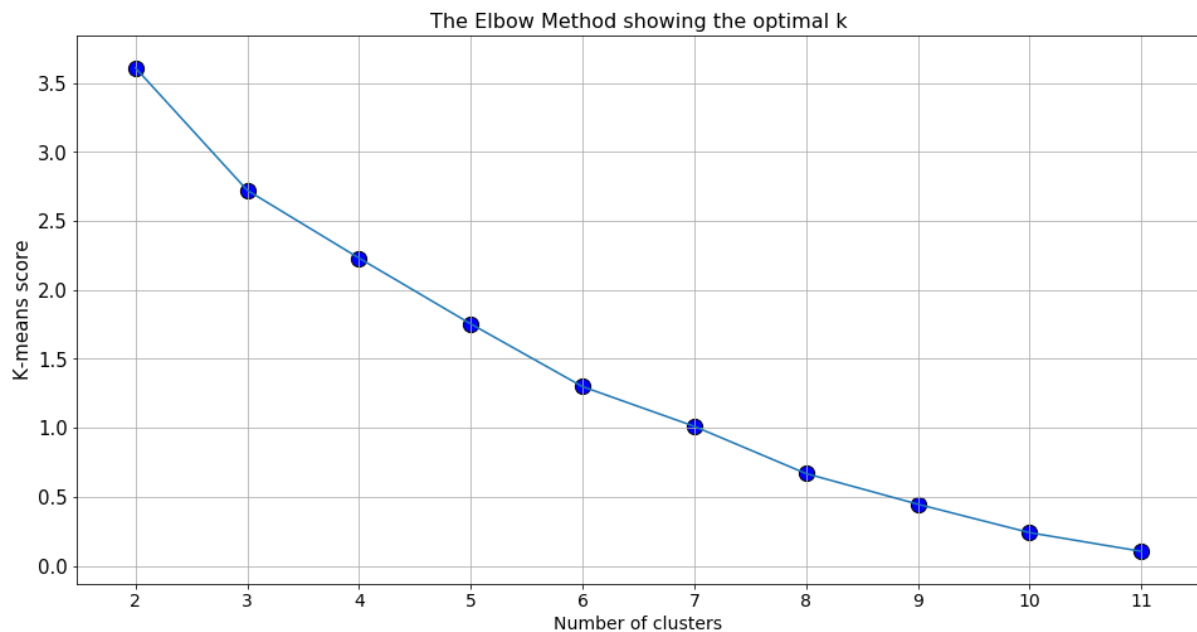


Figure 6 Elbow curves illustrating the K-means score based on inertia for differnt number of clusters K

It can be observed that a sharp decrease in inertia can be observed between *K*=2 and *K*=3 followed by a constant decrease. Although this is likely the optimal value for *K*, the method produced an ambiguous result, making it difficult to select the optimal value. Therefore, an alternative method called the Silhouette Coefficient Method was used to verify the result obtained from the Elbow Curve Method. The Silhouette Coefficient is calculated using the mean intra-cluster distance (*a*) and the mean nearest-cluster distance (*b)* for each sample. It is calculated using the following equation:

$$Silhouette\ Coefficient = \frac{b - a}{\max(a, b)}$$

The optimal Silhouette Coefficient is 1 and the worst value is -1 indicating that a sample has been assigned to the wrong cluster, while values near 0 suggest overlapping clusters that could be

optimised. A scatter plot of the Silhouette Coefficient for various numbers of clusters was generated as shown in Figure 7 below.
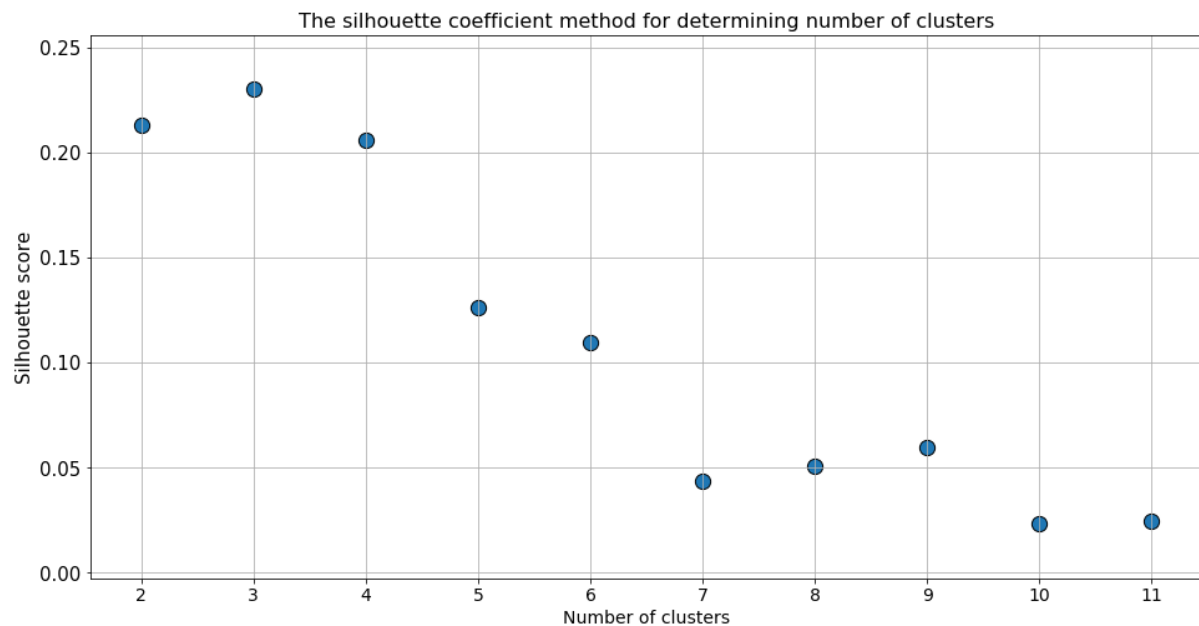


The silhouette coefficient method for determining number of clusters

Figure 7 Plot of Silhouette Coefficients for various numbers of clusters $K$

It can be observed that the Silhouette Coefficient is a maximised for $K$=3, which supports the ambiguous result found using the Elbow Curve Method. Thus, K-means clustering was performed using three clusters for this investigation.

## 4. Results

The clustered neighbourhoods on the Gold Coast using the K-means algorithm were separated into four clusters as follows:

| | |
|---|---|
| Cluster 1 | Dining/Grocery/Shopping Malls (Hospitality) |
| Cluster 2 | Outdoor Recreation/Sports Stores |
| Cluster 3 | Cafes/Clothing Stores |
| Cluster 4 | No recorded venues |

The resulting clusters of each neighbourhood on the Gold Coast based on the most common venues was illustrated on a clustered map. Furthermore, a choropleth map was superimposed in the clustered map to visualize the population of each suburb in 2016. The combined map is illustrated below in
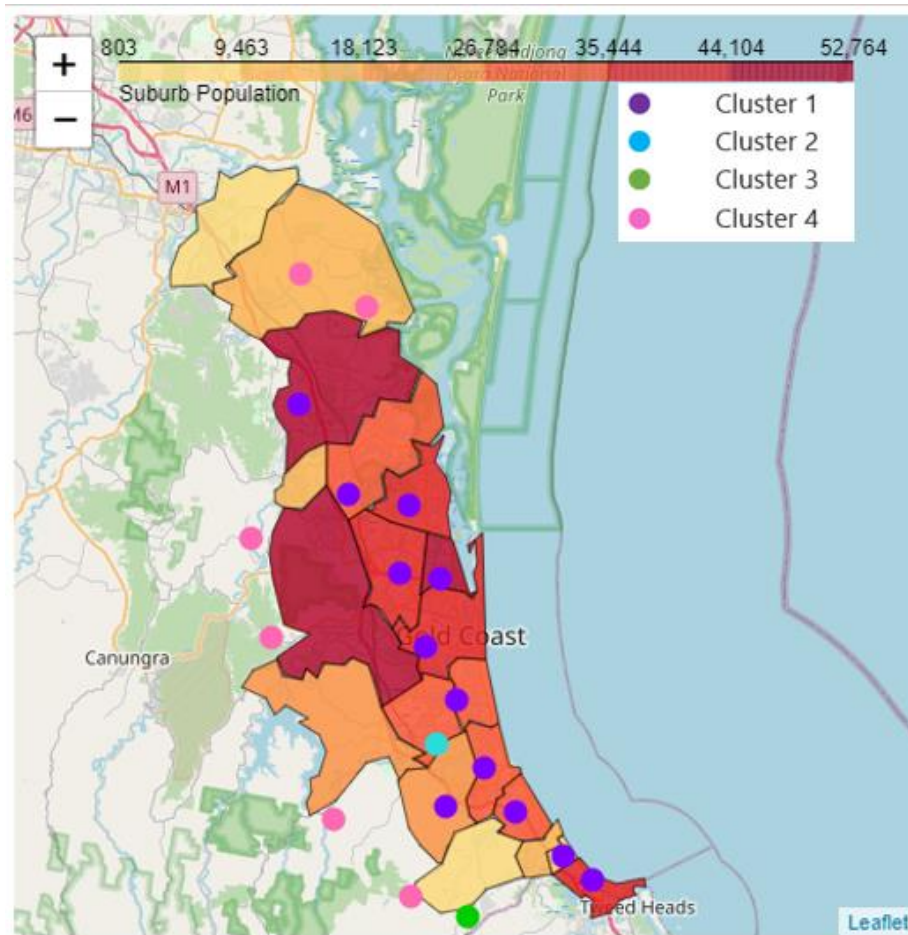
Figure 8 Gold coast map illustrating the clustered neighbourhoods based on common venues and a superimposed choropleth map illustrating the population in each suburb in 2016

# 5. Discussion

It can be observed from the generated clusters that cluster 1 neighbourhoods are predominantly located along the coastline. Conversely, cluster 4 neighbourhoods whereby no venues were recorded by the Foursquare API tend to appear further inland. Visual comparison of the population data and the location of each cluster does not suggest a direct correlation between the two features. Despite Western Gold Coast suburbs (Nerang/Carrara/Highland Park) having one of the largest populations in the city, it is still labelled as cluster 4. Based on the presented data alone, business owners in the hospitality sector have chosen to setup venues near the coastline – irrespective of population. Future businesses in hospitality may be able to leverage the high population in neighbourhoods labelled as cluster 4 and acquire first mover advantage. However, it is also critical to consider various other factors that have been omitted from this investigation such as property rental rates, demographics and average household income for each suburb.

# 6. Conclusion and Future Work

Given the projected growth in Gold Coast population, there is significant potential for future businesses especially in accommodation and food services industry since it is one of Gold Coast's major industries. This project has investigated how similar venues are distributed across the Gold Coast through K-means clustering and the effect that suburb population has on the location of venues. Future work will involve assessing various other factors that have an effect on the success of a business when it comes to deciding the venue location.