**Predicting best location for new café in Singapore**

Lucas Choy Wai Leong

June 30, 2020

**1. Introduction**

**1.1 Background**

Singapore is a small island city-state in Southeast Asia with a population of above 5 million and one of the highest population densities in the world. Café culture has been steadily growing in Singapore where the where youths and young adults visit these places more frequently. Many Singaporeans also treat cafes as a good place to study or get some work done because of the ambience and environment provided. Café hopping has become a norm where multiple cafes are visited within the same day with the sole purpose of experiencing as many of the cafes as possible.

**1.2 Problem**

Looking at the increase in demand for cafes, we want to set up a new line of cafes in Singapore. However, since there is already a good deal of competition in terms of locations similar to cafes, there is a need to identify where would be the best area for the first café.

**2. Data acquisition and cleaning**

**2.1 Data sources**

Data about similar venues such as latitude, longitude, venue name, likes and ratings are from foursquare where information is accessed from its API. Planning areas and the respective population data are taken from Wikipedia ([link](https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore)) where the Statistics Department of Singapore is referenced from 2018

**2.2 Data Cleaning**

The planning areas included separate smaller islands such as North-Eastern, Western and Southern Islands that were not realistic to set up a café at and where removed from the data. Another issue is that some of the venues found nearby planning areas did not have any ratings, these venues were excluded from the data in order to obtain a more accurate representation of the venue ratings in each planning area.

**2.3 Feature Selection**

Name, population and density were the only Features selected from the initial dataset from Wikipedia. Features such as translated names as well as area were dropped. Latitudes and longitudes of each area and venue were chosen as features as well as the average likes and average ratings from the foursquare data. Features such as planning area size and region were dropped since we already have population density and coordinates.

**3. Methodology**

**3.1 Definitions**

In order to ensure that competition is properly measured, venues that are similar to cafes that will impact its opening will have to be defined. These similar venue categories from Foursquare’s Category list would be bistro, coffee shop, dessert shop and cupcake shop as the either provide similar experiences to cafes or sell similar products. Additionally, Cafes are more likely to be successful if they are accessible, as a result public transportation stations such as MRT stations will be used as the location of each planning area.

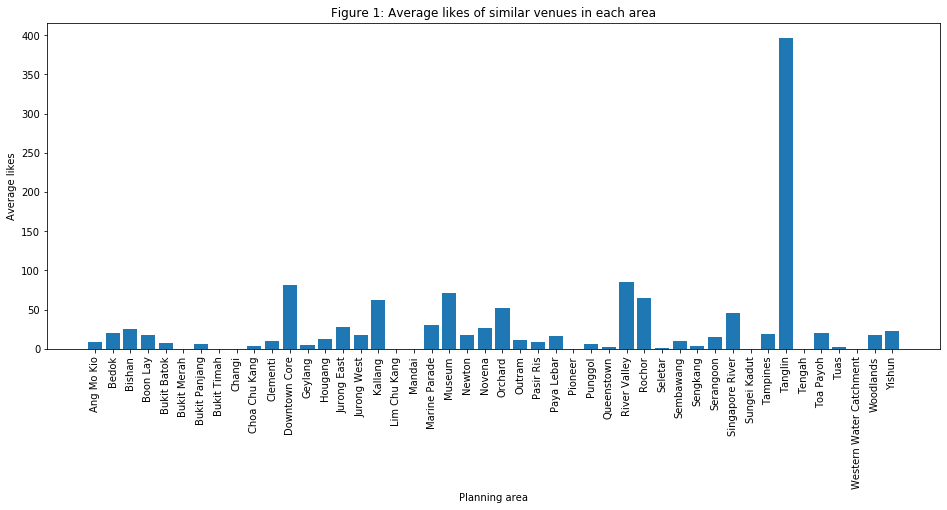
**3.2.1 Calculation of similar venue count, average likes and average ratings**

Venues data such as likes and ratings are pulled from the foursquare data based on a

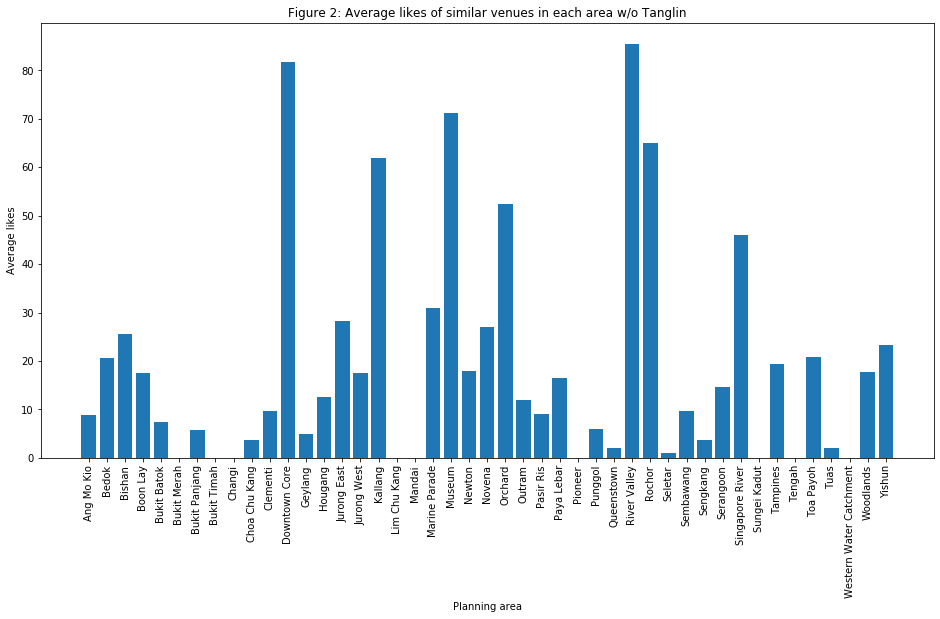
500-meter radius from each planning area. The venue categories are then compared with the defined similar venues and filtered out. The filtered-out venues are then counted, likes and ratings are averaged out based on each planning area.

**3.2.2 Average likes**

Based on the foursquare data, a bar chart is used to compare the average likes of each planning area. In figure 1, it can be seen that Tanglin seems to be an outlier with a very high number of average likes. After looking through the data, there is only 1 similar venue in the area which is PS café. This is a very famous and popular café in Singapore which along with it being the only similar venue in the area explains the unusually high amount of average likes.

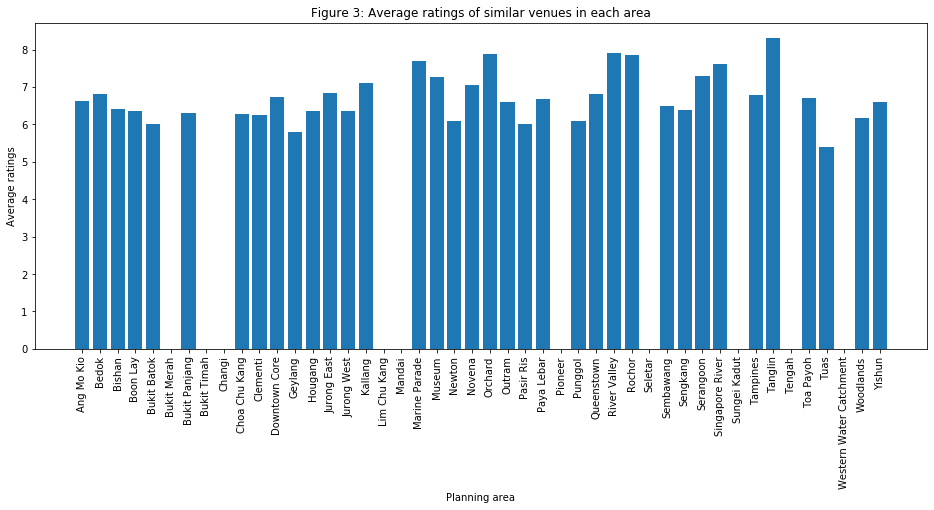


In figure 2, Tanglin is removed so the other areas can be compared more closely.



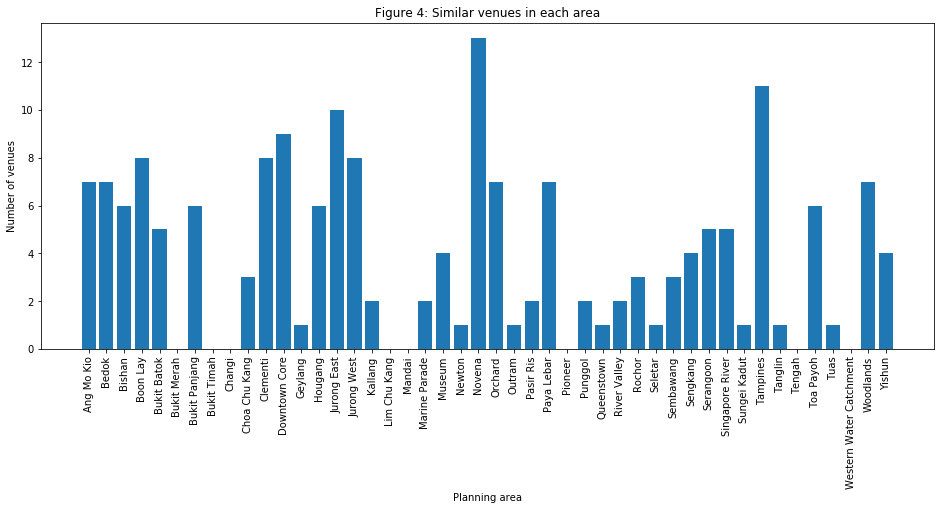
**3.2.3 Average ratings**

Similarly to average likes, a bar chart is also used to compare the average ratings. In Figure 3 it can be seen that all the average ratings are quite close together around the 5 to 8.5 range. Planning areas with 0 average ratings are due to the lack of similar venues in the planning area instead of an average rating of 0.



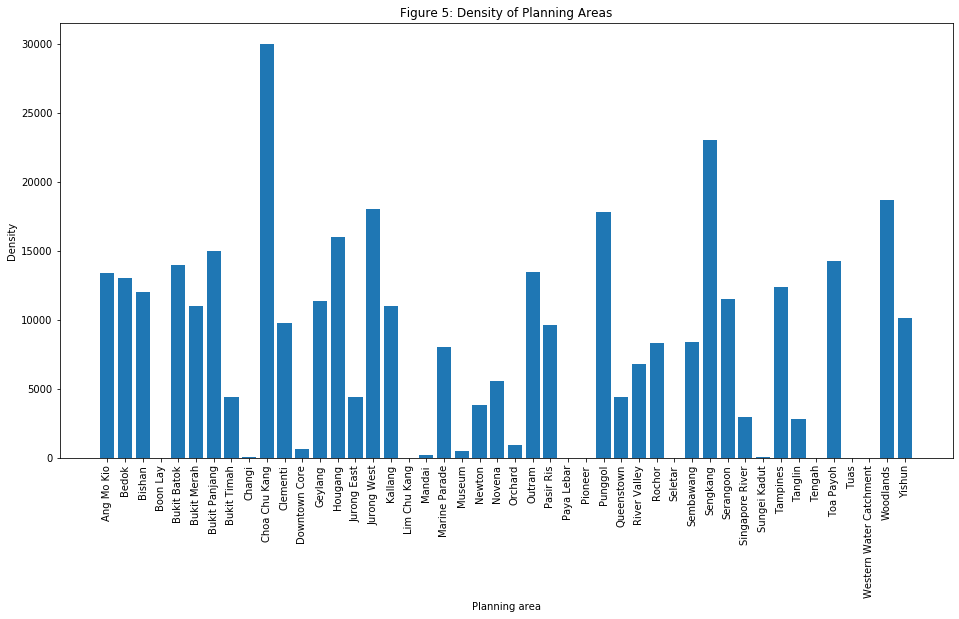
**3.2.4 Counting similar venues**

A bar chart is used to compare the number of similar venues in each planning area. From figure 4, number of similar venues vary greatly from 0 to 13.



**3.3 Population Density**

A bar chart is used to compare the population density of each planning area. From figure 5, it can be seen that there are some planning areas with a very low population density. This is likely due to the fact that these are not residential areas where high rise housing is located since majority of the population live in such buildings.



**3.4 Clustering Areas using K-means**

The planning areas are assigned to different clusters based on the different categories of nearby venues. From the original nearby venue data pulled from Foursquare before the filtering of similar venues, the different venues are grouped by their planning area and are changed to dummy variables. K-means is an unsupervised machine learning algorithm that compares each of the venue’s categories in each planning area and clusters them based on how similar their categories are. The number of clusters used is 5 for this algorithm.

Based on the clusters assigned, Folium is used to visualise the planning area and their different clusters on the Singapore map.

**4. Results**

**4.1 Average likes**

Based on figure 1, the planning area with the most amount of average likes is Tanglin as previously mentioned. There are several planning areas with 0 average likes which is Bukit Merah, Bukit Timah, Changi, Lim Chu Kang, Mandai, Pioneer, Seletar, Sungei Kadut, Tengah and Western Water Catchment.

**4.2 Average ratings**

From figure 3, Tanglin has the best average ratings of 8.3. Out of all the planning area it is the only one to have an average rating of above 8 whereas the others are around 5 to 8. The planning area with the lowest average ratings is Tuas with average rating of 5.4. As mentioned before planning areas with a rating of 0 is due to no reviews made by users. Planning areas without any reviews are Bukit Merah, Bukit Timah, Changi, Lim Chu Kang, Mandai, Pioneer, Seletar, Sungei Kadut, Tengah and Western Water Catchment.

**4.3 Similar venue count**

From figure 4, it shows that Novena has the most amount of similar venues with a count of 13 whereas planning areas such as Bukit Merah, Bukit Timah, Changi, Lim Chu Kang, Mandai, Pioneer, Seletar, Sungei Kadut and Western Water Catchment have no similar venues nearby their location.

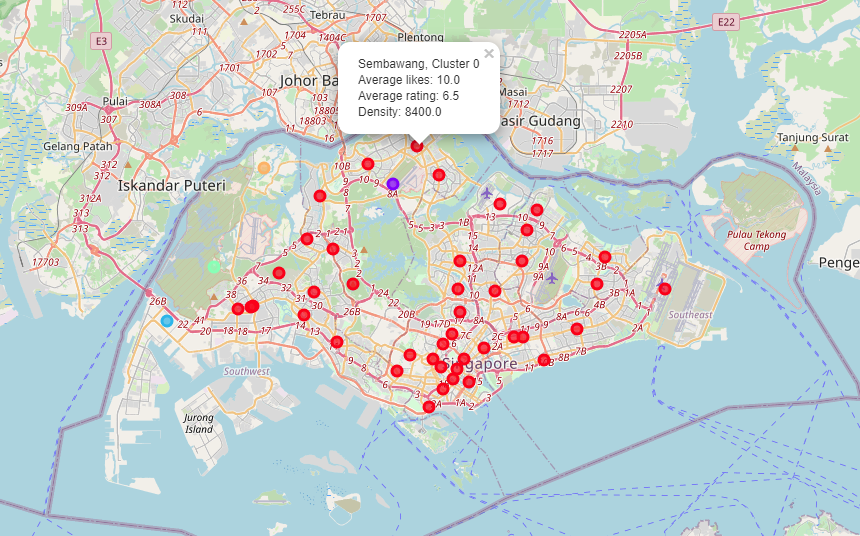
**4.4 Population density**

From figure 5, Chua Chu Kang has the highest population density of 30,000 per square kilometre. In particular, Boon Lay, Changi, Lim Chu Kang, Paya Lebar, Pioneer, Selear, Sungei Kadut, Tuas and Western Water Catchment have a population density very close to 0.

**4.5 Cluster of planning areas**

From figure 6, most of the planning areas are very similar and belong to cluster 0. Cluster 1 has a single area, Mandai, in it. Exploring the data shows that the only venue in Mandai is an Asian Restaurant. Tuas is the only planning area in cluster 2 since it only has an outdoor supply shop and a food court as its venues. Cluster 3 contains Western Water Catchment as the only area since it has a gun range as it’s only venue and gun ranges are very rare in Singapore Lim Chu Kang belongs to cluster 4 where it has a farm and zoo exhibit nearby. Farms are very rare in Singapore and there is only 1 zoo in the country.

Figure 6: clustering of planning areas in Singapore



**5 Discussion**

**5.1 Limitations**

Even though Foursquare has a large amount of location data in Singapore, it seems that the data is lacking in terms of users. This is reflected by the relatively low number of reviews and hence likes and ratings of each venue. This consequently results in likes and ratings not being the best indicator of how well a venue is doing and its influence, a better indicator might be the number of check-ins at each venue but that is only available to the owner of the venues.

Next, even though population density gives a good indication of the amount of people in an area, it only represents the people who live in that area. Popular shopping areas such as Orchard would have a very low population density since not many people live there. Foot traffic data would be more ideal but it is very difficult to collect.

Lastly, choosing MRT stations as the coordinates for each planning area does not fully explore all the options in each planning area as there may be more viable café locations deeper into the planning areas away from the MRT stations.

**5.2 Recommendations**

Using the similar venue counts as the first consideration, areas such as Bukit Merah, Bukit Timah, Changi, Lim Chu Kang, Mandai, Pioneer, Tengah and Western Water Catchment would be the best choice to start a café since there are no similar venues. Next, we consider the Population density as well resulting in Bukit Merah being the best choice by far with a population density of 11,000 per square kilometre.

However, another question is what a good area would be if a small level of competition is acceptable. Firstly, we consider the number of potential customers using the population density data. Chua Chu Kang, Hougang, Jurong West, Punggol, Sengkang and Woodlands have high population densities of above 16,000 per square kilometre. Next, these areas with the least competition are Chua Chu Kang and Punggol with 3 and 2 similar venues respectively. The average rating and average likes of these 2 areas are close but since Chua Chu Kang has a population density of 30,000 as compared to Punggol’s 17,800, Chua Chu Kang would be the best area for a café with minimum competition.

**6. Conclusion**

In this study, the best planning area to setup a new café chain was explored using population as well as venue location data. Based on the limitation present with the data, a concrete solution to the best planning area cannot be chosen. However, there are some planning areas that already look a lot more promising based on this study, with a clear choice for some conditions. It is also important to note that the best area to set up a new café chain depends on the different levels of risk that the owner is willing to accept. Moving forward, more ideal data can be collected such as venue earnings, venue patronage and foot traffic can be obtained to make comparisons more accurate and realistic. Moreover, additional locations can be explored within each planning area to provide a more comprehensive comparison.