Predicting Ideal Haidilao Franchise Location in Singapore

Matthew Loke

May 24, 2020

# 

Table of Contents

[**Introduction** 3](#_Toc41344869)

[Background 3](#_Toc41344870)

[Problem 3](#_Toc41344871)

[Interest 4](#_Toc41344872)

[**Data Acquisition and Cleaning** 4](#_Toc41344873)

[Data Sources 4](#_Toc41344874)

[Feature Selection 4](#_Toc41344875)

[Data Cleaning 5](#_Toc41344876)

[**Data Visualisation** 5](#_Toc41344877)

[Area of Service 5](#_Toc41344878)

[Proximity to Population Clusters 6](#_Toc41344879)

[**Exploratory Data Analysis** 6](#_Toc41344880)

[Clustering 6](#_Toc41344881)

[Relationships 7](#_Toc41344882)

[**Predictive Modelling** 8](#_Toc41344883)

[Criteria Testing 8](#_Toc41344884)

[Classification Models 8](#_Toc41344885)

[Accuracy Score 8](#_Toc41344886)

[**Conclusion** 8](#_Toc41344887)

[**Future** **directions** 8](#_Toc41344888)

# **Introduction**

## Background

The brand Haidilao was founded in 1994. After over 20 years of development, Haidilao International Holding Ltd. has become a world-renowned catering enterprise. Over the years, Haidilao has withstood the challenges of the market as well as customers, and has successfully forged a quality hot pot brand which has earned a reputation for itself. As a large-scale chain catering enterprise with operations all over the world, Haidilao adheres to integrity in business. It gives the highest priority to continuously improving the quality and safety of its food products, providing more thoughtful services to its customers while delivering healthier, safer and more nutritious food.

Haidilao also faces close competition from the Paradise Group, with its franchise Beauty in the Pot. Beauty in the Pot offers a high quality hot pot experience and has also built its reputation. The ingredients are artfully made by their team of chefs and promises an amazing hot pot experience for its customers.

## Problem

My client is looking to open a Haidilao franchise in Singapore and requests that I recommend a location for them to open in. The following are information regarding specifications of business my client is intending to run.

1. Chinese Establishment
2. Mid range Pricing
3. Known brand franchise
4. Family oriented establishment

#### As such, the following details are required in order to which is the ideal location for the new restaurant

1. Locations of current Haidilao franchises
2. Population Density in each region
3. Location of Competitor, Beauty in the Pot
4. Must be in a shopping mall

## Interest

Investors intending to open franchises will be interested to know the data collected regarding ideal locations for starting Haidilao or Beauty in the Pot franchises. Changes in the query to different types of food franchises can yield the ideal location for those franchises and investors for those franchises will be interested in this data.

# **Data Acquisition and Cleaning**

## Data Sources

The data required was mentioned above and the relevant sources for each respective detail was obtained as follows.

1. The data of Haidilao franchises was scraped using Foursquare.
2. The data about the districts in Singapore can be found from [Resident Population by Planning Area/Subzone and Type of Dwelling, 2015](https://data.gov.sg/dataset/resident-population-by-planning-area-subzone-and-type-of-dwelling-2015?view_id=4f094c60-64ec-4b87-8e21-458fb90eab78&resource_id=5fe7f164-14d7-4dae-927c-46fe72b4822f). The districts information will be used to find population clusters in order to determine customer flow rates.
3. The data of competitor franchises was scraped using Foursquare.
4. The data of mall locations available for rent was scraped using Foursquare.

## Feature Selection

After obtaining each dataset, the features which are required for examining which location is ideal was selected. Below is a summary of the feature which was kept and the reason behind retaining the information.

|  |  |  |
| --- | --- | --- |
| Feature | Reason | Dataset |
| ID | To identify the franchise | Haidilao, Beauty in the Pot |
| Formatted Address | To check the exact location of the business | Haidilao, Beauty in the Pot |
| lat | To obtain the exact point on folium maps | Haidilao, Beauty in the Pot |
| lng | To obtain the exact point on folium maps | Haidilao, Beauty in the Pot |
| level\_2 | The district names | Residential population data |
| value | District population | Residential population data |

*Table 1: Selected Features from Data Sources*

The other features extracted from web scraping either contained similar information, redundant information or information which is not consistent across the rows. As such, the information was dropped from the dataset. The data were processed separately to fulfil the different criteria and pieced together for predictive modelling,

## Data Cleaning

Data was downloaded or scraped from multiple sources. The data was cleaned and arranged into useful datasets to be used for data visualisation, testing or prediction. The data is processed as follows.

The data of Haidilao franchises were extracted and the information of the ID, formatted address, latitude and longitude were kept in order to have the coordinates of each Haidilao store. Similarly, the data on the competitor was extracted and visualised on the map of Singapore using the folium package. The markers had a radius of 40 to visualise the approximate area serviced.

There are some problems faced with the datasets. First, the data extracted from Foursquare API does not have a standardised format for the address and name, resulting in some manual work in cleaning the data. In addition, whether the location is in a mall is not explicitly stated in the formatted address. From my prior knowledge of malls in Singapore as well as inferring from the # in the address, i manually filled up the columns for whether it is in a mall or not.

Second, the extraction of information from the Resident population does not have a location which i can tag the latitude and longitude to show in the folium map. As such, I queried the location individually and generated another dataset to obtain the approximate latitude and longitude of each district.

Third, when the query was searched, there were entries which were similar but not exactly the same. This would result in inaccuracies when using the data. Hence, I edited the script to delete the entries which did not belong to any of the franchises.

# **Exploratory Data Analysis**

## Clustering

Clustering was done for the different areas of Singapore to determine the population clusters around Singapore. As the data points are relatively close to each other and clusters may overlap, it was decided that Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was to be used.

Each cluster was colour coded to see the probable locations to set up the business. Outliers in the clustering were colour coded to be black. Outliers in this clustering were then deemed to be datapoints (districts) where setting up the franchise would not serve a lot of purpose. However, each datapoint would also be tagged with different population size. Hence, these clusters were then grouped up to obtain the final cluster population. As such, the following information was derived.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster Label | Colour | Population size (avg) | Latitude (Centroid) | Longitude (Centroid) |
| 0 | Green | 227588 | 1.364981 | 103.743794 |
| 1 | Yellow | 160964 | 1.312217 | 103.854052 |
| 2 | Purple | 383552 | 1.337624 | 103.942858 |
| 3 | Orange | 271286 | 1.432985 | 103.832206 |
| 4 | Blue | 310810 | 1.396437 | 103.886605 |
| -1 | Black | 235811 | 1.376222 | 103.785225 |

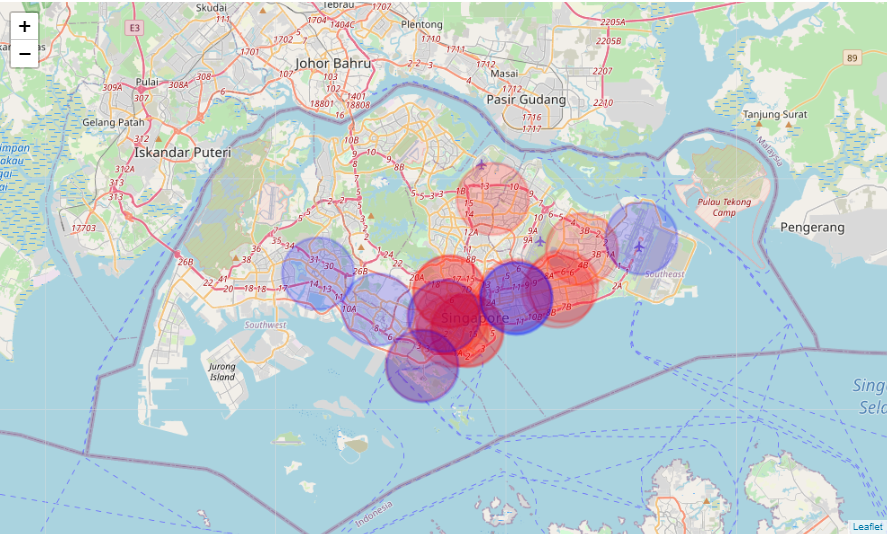
*Table 2: Data regarding clusters*

# **Data Visualisation**

For the bulk of this project, a lot of data visualisation is required in order to develop a solution. This is due to the fact that the extracted data are unsupervised data and the visualisation would help in coming up with a way to use these unlabeled data to generate a workable solution.

## Area of Service

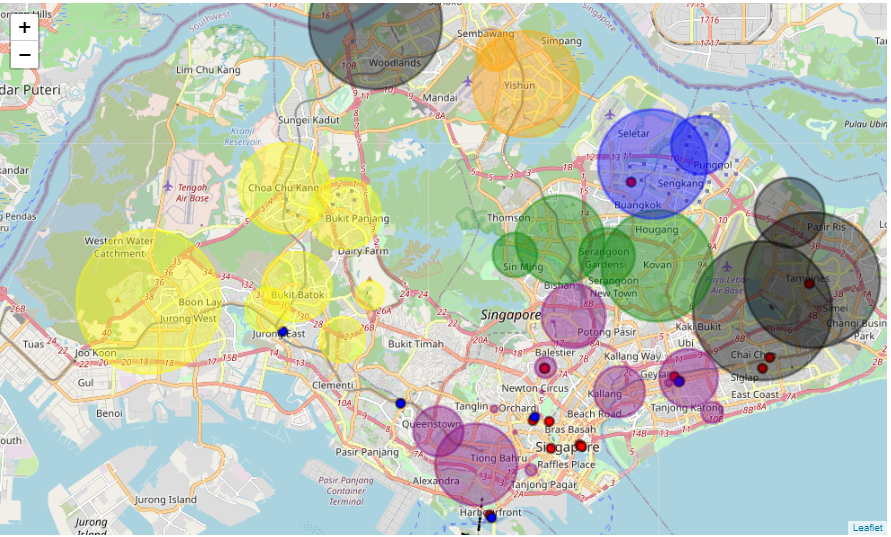
Using the mined data for Haidilao and Beauty in the Pot franchise locations, the area of service can be determined. The area of service refers to the approximate area where locals or nearby residents would visit that franchise if given an opportunity. From the markers, it can be noted that the west region and northern regions of Singapore are not covered by these franchises. This feature was noted for further development for the ideal location selection.



*Figure 1: Area of Service Visualisation*

## Proximity to Population Clusters

The population clusters visualised with a colour code in order to distinguish between clusters. The cluster markers sized were also varied according to the population size in the area to allow a visualisation of the amount of people in the area. This is to help visualise the size of the customer base in each individual location. In addition, the franchise markers were added in, using small blue markers for Beauty in the Pot and red markers for Haidilao franchise locations. This allows the visualisation of proximity of existing franchises to population clusters.



*Figure 2: Visualisation of population clusters and franchise locations*

# **Results**

From the cluster population sizes, the biggest cluster would be the purple cluster, which represents the location of the Central Business District in Singapore. In addition, from the folium map generated above, we can see that there are many hotpot franchises around the purple clusters. This means that there is a relationship between the hotpot franchise location as well as the cluster size. Hence, this will be used in the criteria testing for the locations.

It is also noticed that there are big population clusters in North, West and Central Singapore. As such, locations like Yishun in the north, Boon Lay in the west and Serangoon in Central Singapore can be deemed as ideal locations to set up the business. However, in the west, there is a Beauty in the Pot at Jurong East and there will be some form of competition. Serangoon is also close to the Haidilao in Seletar and there will be some competition from that outlet as well. Hence, we shall select Yishun as the ideal location for starting up a Haidilao franchise. The next step is to predict if it is a good selection with the relevant criteria testing.

# **Predictive Modelling**

## Criteria Testing

From the problem statement above, the information regarding location of Haidilao and Beauty in the Pot franchises, population density and clusters as well as proximity to shopping malls was obtained. The hypothesis of Yishun being an ideal location for setting up a franchise was determined and a new dataset was created. The dataset contained the criteria for testing, which involved whether there is a mall in the area, the proximity of the business to cluster centroids as well as the distance from current franchises. The dataset was generated using prior knowledge which was mined. The first criteria was determined from the formatted address. It was noticed that all franchise locations were in malls. The second criteria ensures that the franchise is near a broad customer base. The last criteria ensures that competition between similar franchises or outlets is minimised.

## Classification Models

Decision Tree Classifier was used in this instance. From the previous section of data mining, there is relevant data to map an input to an output, allowing for supervised learning. The mapping is based on example input and output pairings. Therefore, the first step is to use the data regarding the current Haidilao and Beauty in the Pot franchises and split into train and test sets. Then the data was fitted into the decision tree classifier model and tested accordingly. In this instance, the decision tree is a classification tree, whereby the decision variable is categorical, providing Boolean values.

## Accuracy Score

After fitting model to the Decision Tree Classifier, the accuracy score, imported from the metrics package, was determined to be 1.0, indicating perfect prediction with this prediction tree as well as this train test set. This means that the prediction of the decision tree that Yishun is a good location to set up the business, according to the current criteria.

# **Future** **directions**

The size of the data set can be seen to be very small. As such, the scoring for the decision tree classifier is 1.0, saying that the prediction is 100% accurate. In addition, other features should be identified to add in more information and criteria to provide a more accurate prediction for where the new franchise should be set up.

In addition, since the dataset for the train and test set were all considered as true, it is hard to prove when the data should be false and the prediction may not be as reliable. Hence, one solution may be to add in more dataset using different kinds of businesses to fill up the gap for the dependent variable, giving the prediction model the ability to show True and False prediction statements.

Additional testing can be done to test the similarity of hotpot businesses. A query of hotpot businesses can be done to extract the data of hotpot businesses and another classification tree can be used with the relevant data of ratings, price etc can be developed to give more accurate group of competitors in the hotpot industry.

Lastly, an additional step can be added in. The prediction of where the ideal location should be can be added into the code. The ideal location would test the various districts and clusters and more information regarding whether there are available stores for rent in the mall would have to be scraped.

# **Conclusion**

From the data available from the data sources mentioned above, I believe that my client should open his franchise at Northpoint Shopping Mall, Yishun, due to the proximity to a large population which is not served by the franchise and its competitor. However, There are a few changes which can improve the analysis to create a more holistic analysis. These changes can be made in future projects by parties interested in finding an ideal location for their franchise.