Teo Wei Lung

IBM Data Science Capstone: Battle of the neighborhoods

Contents

[Introduction / Business Problem 1](#_Toc46327132)

[Data 1](#_Toc46327133)

[Methodology 1](#_Toc46327134)

[Results 6](#_Toc46327135)

[Discussion 7](#_Toc46327136)

[Conclusion 8](#_Toc46327137)

Capstone Project Report

# Introduction / Business Problem

As a small city state with a population of close to 6 million, Singapore has grown from third to first world in the matter of a few decades and currently have the 7th highest GDP per capita in the world.

Knowing that Singapore is unique with its multi-cultural character where food is one of the defining features of the country, coupled with the fact that Singapore's is wealthy and has a stable government, a wealthy group of investors from Japan is thus looking to expand their presence Internationally and is looking at Singapore as their first overseas venture, to open an upscale Japanese restaurant that serves premium, authentic Japanese food where Japanese food is highly favored amongst diners.

As part of their efforts to expand, I have been tasked to determine the best location in Singapore to set up shop based on the criteria identified by the investors.

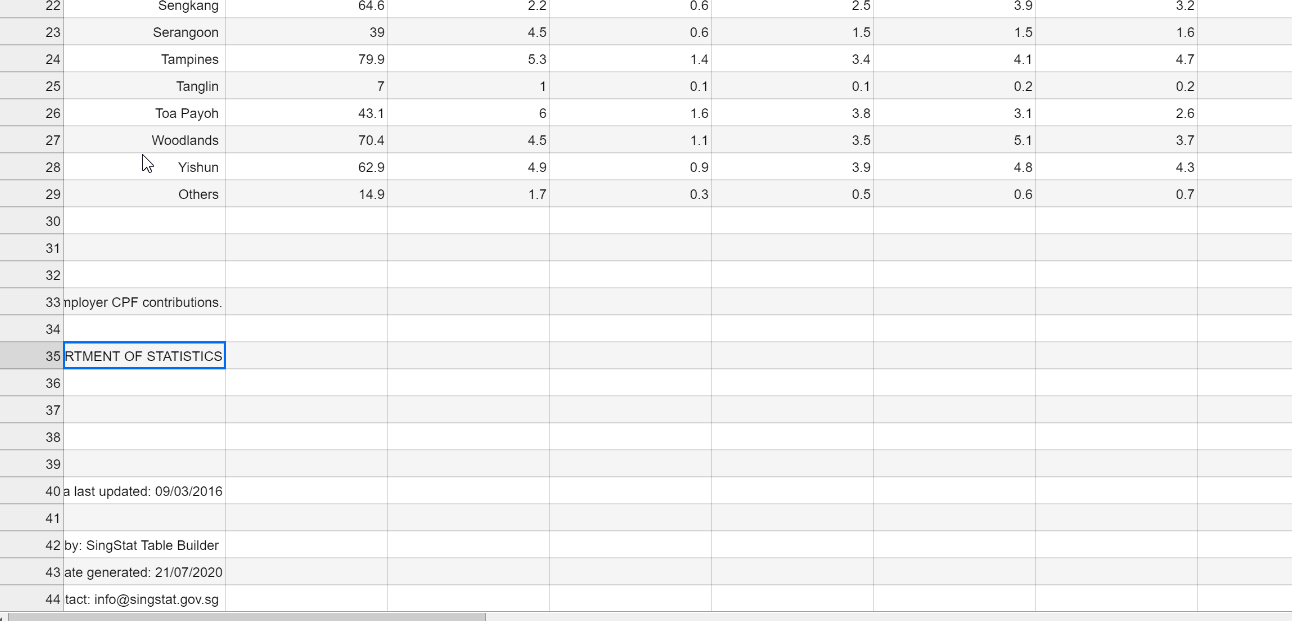
# Data

To accomplish my goal, I will first get a list of districts in Singapore as well as income data from **GenHouseholdSurvey2015.csv**, downloaded from the Department of Statistics Singapore with information about Resident Households by District and Monthly Household Income from Work. This will help us identify the districts in Singapore with higher income levels which is the target audience of the new restaurant.

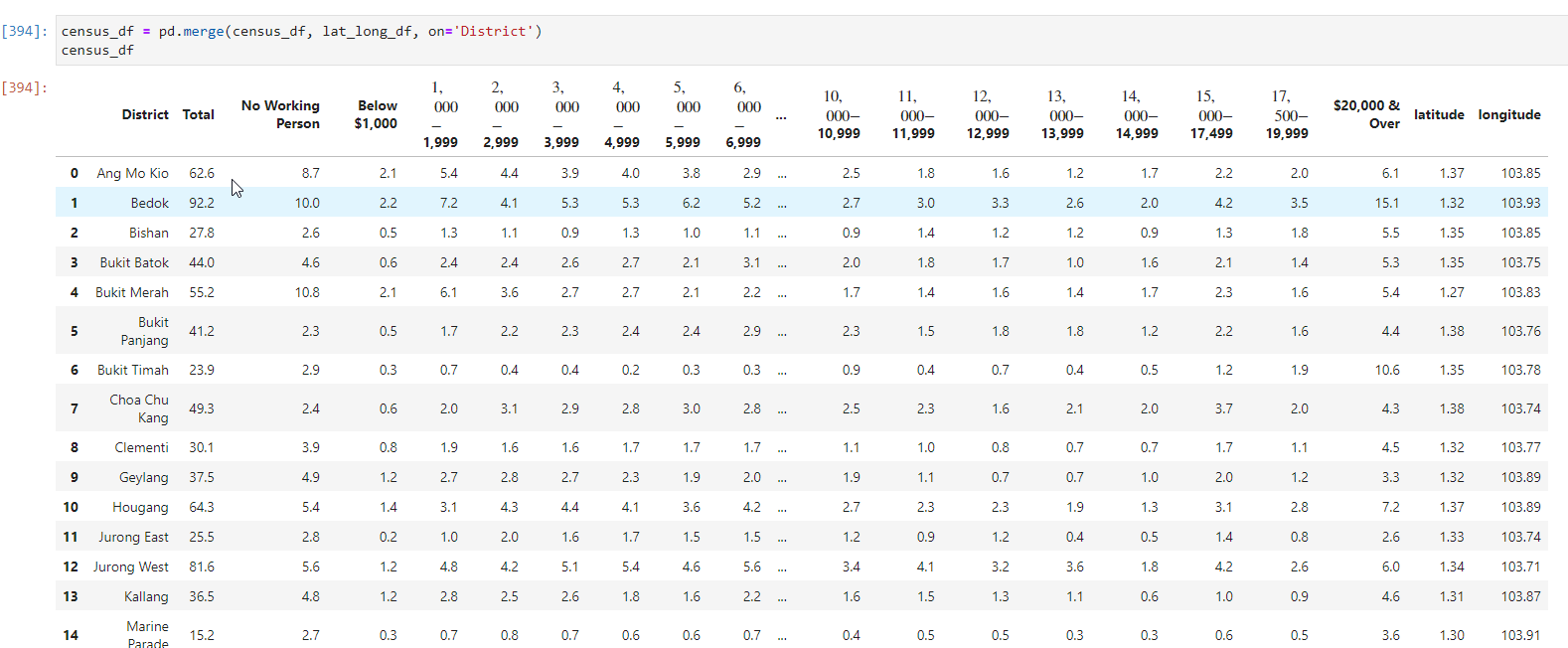
In addition, I will leverage the Foursquare API to find out information about the popularity of different food venues in Singapore. This will assist me to discover each district's preferences of venues, so that I can cross refer income information as well as food venue preference information to determine the best place to set up the restaurant.

# Methodology

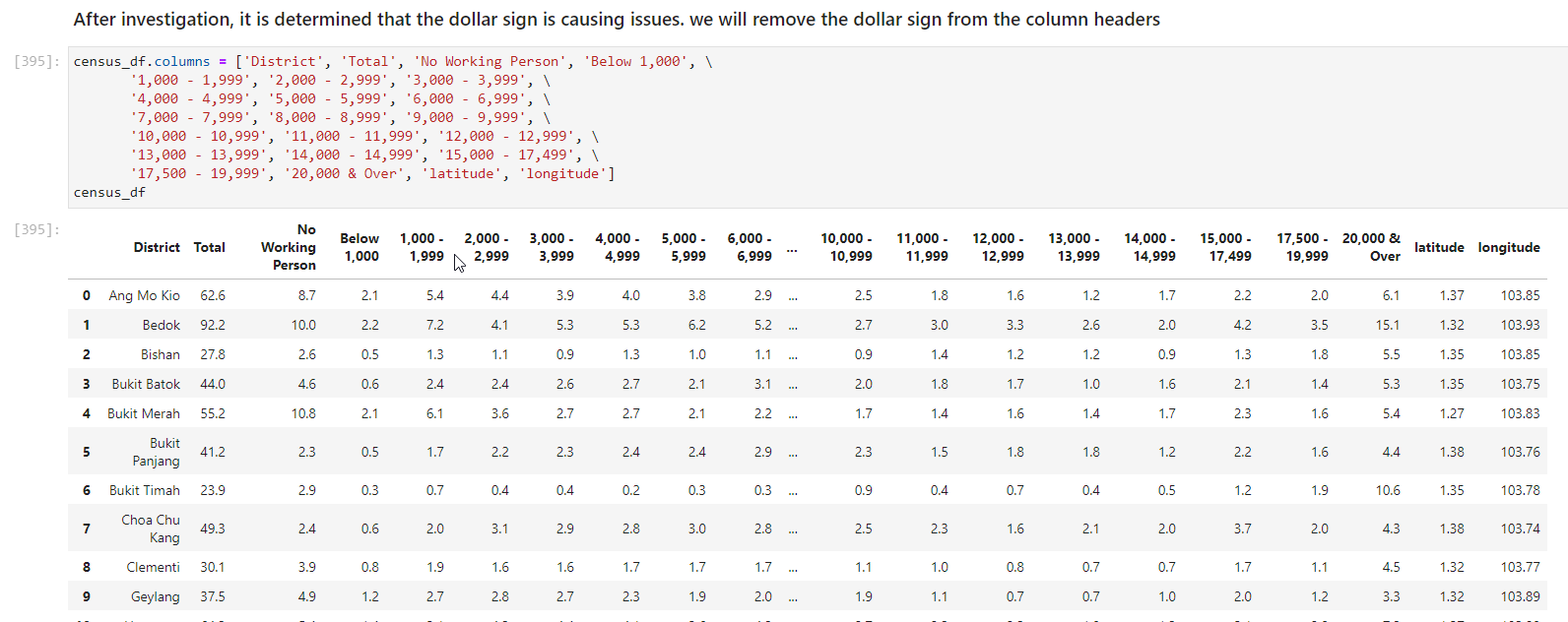
After downloading the CSV with district income level data from the Singapore Department of Statistics website, I first read the data into a pandas dataframe so that I can do data cleaning and other related tasks.



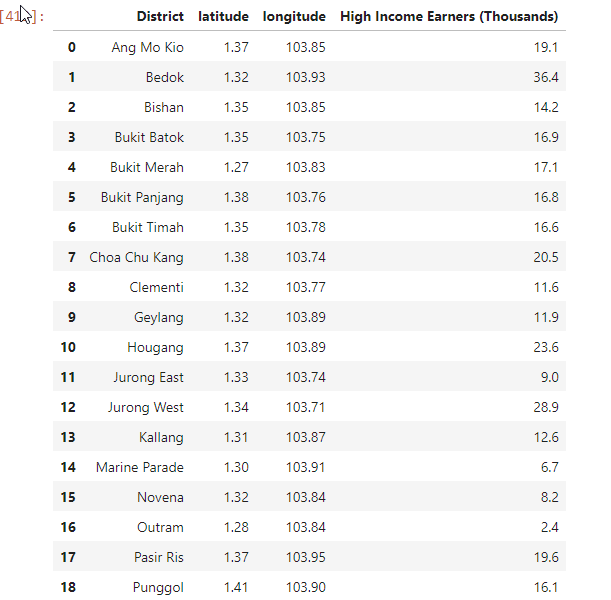
We are going to analyze data from residential districts in Singapore, therefore, we can safely remove “Others” location from this analysis. Hence, from the above screenshot, we notice that rows 29 and onwards are not required. We will proceed to drop these rows.



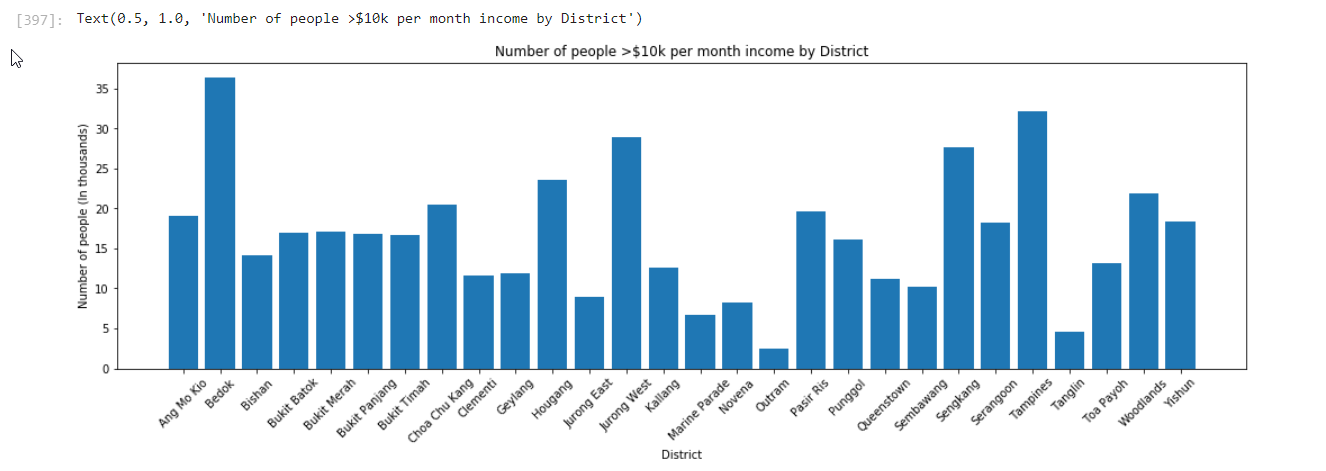
This is part of the dataframe of the information we want. The headers are displaying formatted characters, we want to fix this. Turns out that the dollar signs are causing the header characters to be formatted.



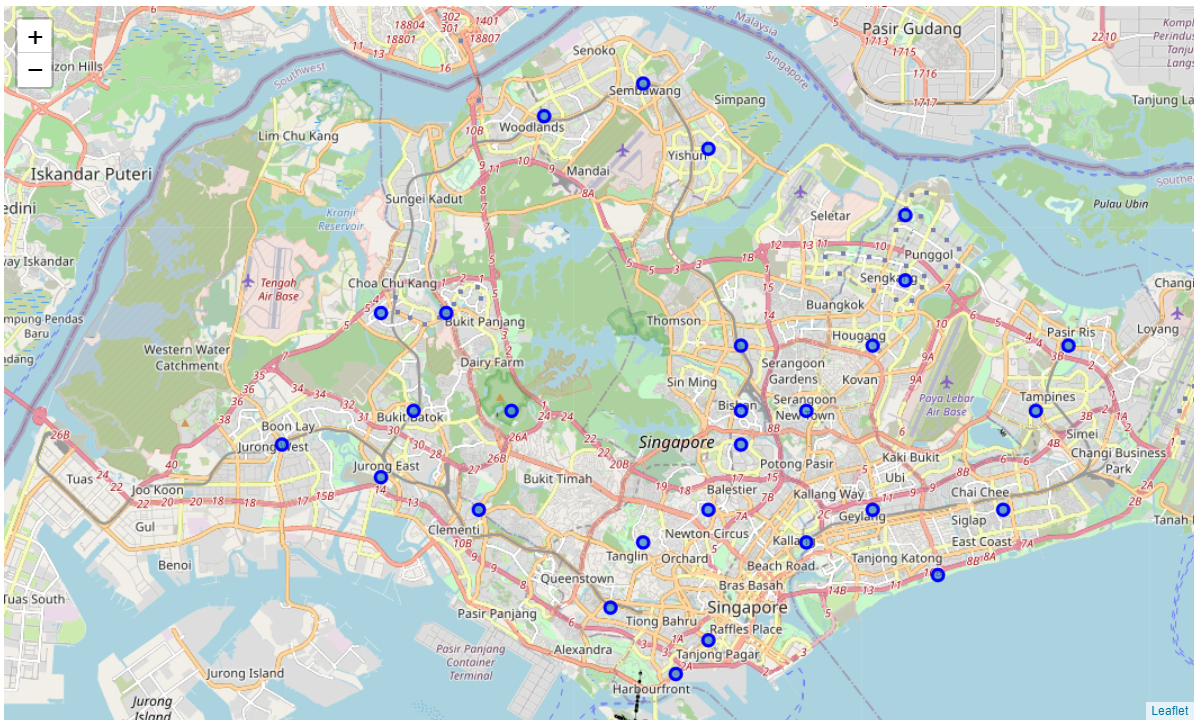
Dataframe is now displaying correctly. Since the client is intending to open a restaurant that caters to the upper class, we are going to focus on data columns for people earning >$10k per month. We will add the numbers up and create a new column called ‘High Income Earners’. We will then proceed to drop the rest of the columns that will not be required



Dataframe after data is aggregated and cleaned up. Lets have a look at the histogram representation of the data



Bedok has the highest number of high income earners, followed by Tampines, then Jurong West.

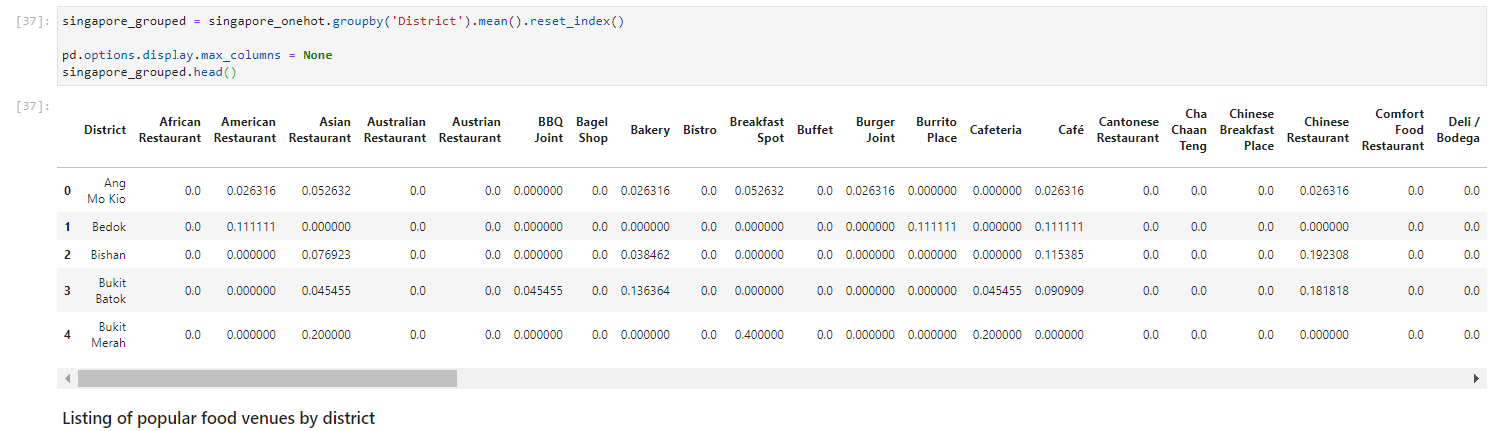


Next, we plot all the different district location on the Singapore map using the Folium map visualization library. We will then attempt to use Foursqaure’s Explore API to find out nearby food venues by district.



After calling the API, we have a total of 773 food venues across all districts in Singapore. The above dataframe is a sample of the data returned.

To perform further analysis, we need to perform one hot encoding, followed by grouping the data by District. This is important for K-means machine learning later. This data transformation also allows us to find the mean frequency of food venues for each district as well (as shown in image below).



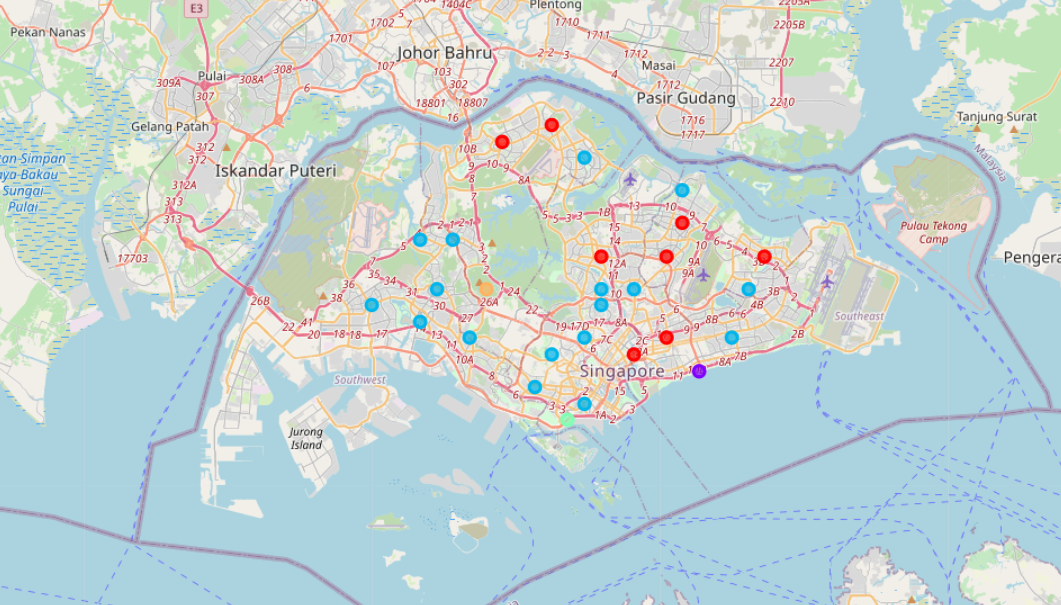
Now we can rank top 5 food venues by district. Bedok is the district with highest frequency allocated to “Japanese restaurant.



We can now finally perform machine learning on our data. We will use K-means learning with K = 5. With the cluster labels generated, we attach it to the dataframe.

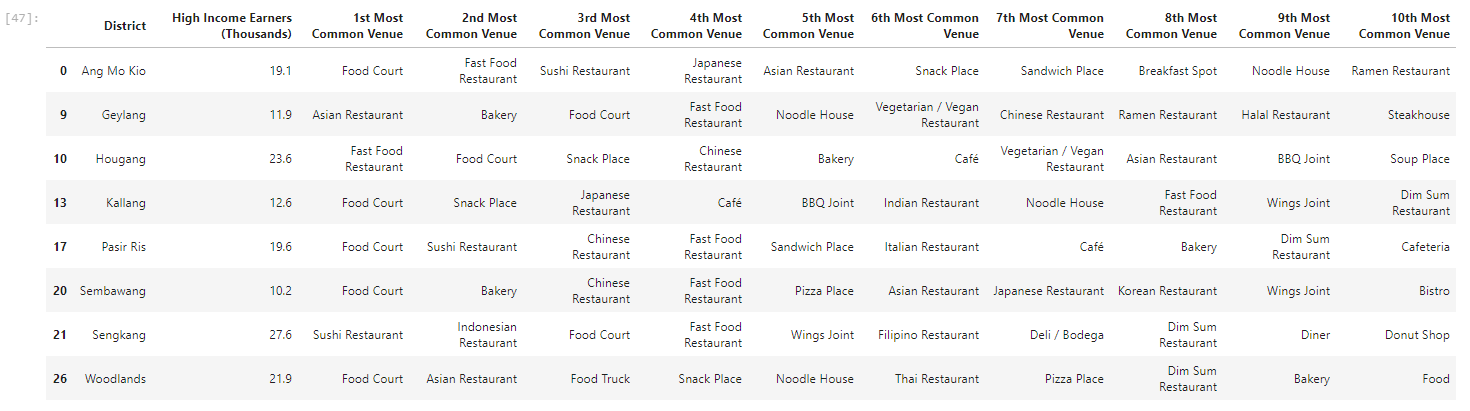


The following shows the cluster plot on the Singapore map.



# Results

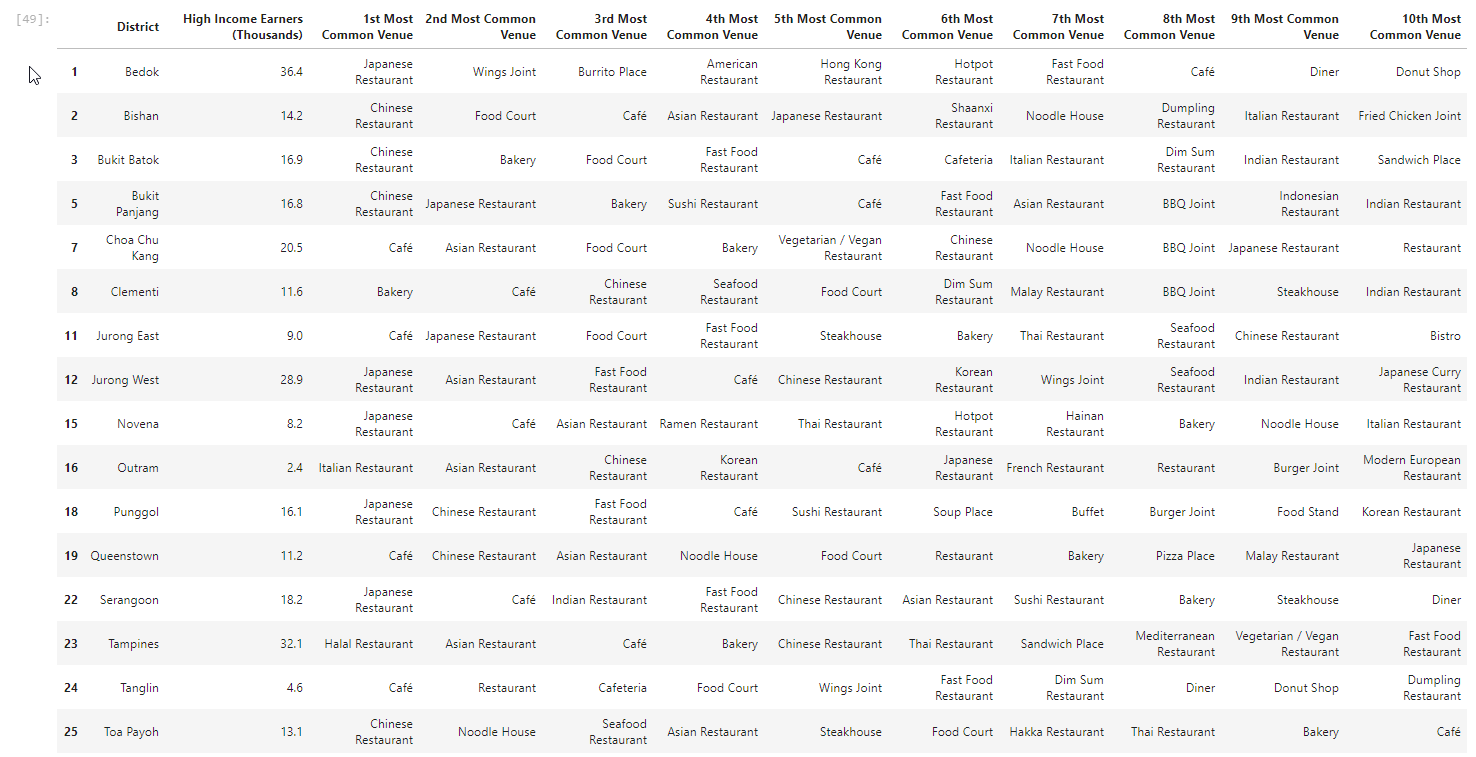
The following images shows the details of each of the 5 clusters:



**Cluster 0**



**Cluster 1**



**Cluster 2**



**Cluster 3**



**Cluster 4**

From the results above:

* Cluster 0 (Most prefer eating at food courts / low cost sushi chains / fast food)
* Cluster 1 (Marine Parade is situated in the city centre, most visited venue is the café probably because it is a popular place to do work / have meetings / lunch)
* **Cluster 2 (Most promising cluster with many districts preferring Japanese restaurants)**
* Cluster 3 (This cluster at Bukit Merah prefer “Breakfast Spot”, referring to hawker centers and affordable, neighborhood food fare)
* Cluster 4 (Bukit Timah has got high preference for Halal cuisine. Not what the client is looking for)

# Discussion

Other important factors that were not examined in this exercise that can influence decision of location selection is the relative space rental costs as well as presence of competitors serving similar Japanese cuisine at each identified district.

Another potential weakness in this exercise is that the income data was from 2015 (latest data available as of 2020) and thus may have deviations from the real data today.

Finally, we primarily considered recommending setting up the restaurant in a district with the highest income earners. In reality, Singapore is a small country where people oftentimes like to travel to the city centre to meet their friends & families to have meals together. Hence, the central part of Singapore may potentially be more advantageous in terms of maximising target audience of the entire country in general (e.g. Jurong West located at the Western end of Singapore has got a high number of high-income earners as well, so if the shop is in Bedok, which is situated East of the country, the client potentially misses out on reaching these people). Bearing that in mind, we could consider performing further analysis and entertaining the possibility of setting up the shop somewhere in the central part of Singapore instead of Bedok.

# Conclusion

Taking from the results, we determined that Cluster 2 (Light Blue) has got the highest preference for Japanese cuisine, especially the districts Serangoon, Jurong West and Bedok, Punggol, Novena. Thus, it is fairly safe to conclude that the new shop should be set up somewhere in Cluster 2. Let us now sort the Cluster 2 Districts by number of high-income earners to decide the best district within the cluster to set up shop



The district winner is Bedok! Jurong West is a strong runner-up to set up a Japanese restaurant as well.