The Battle of the Neighbourhoods — Opening a Bubble Tea franchise in Singapore

A quest to find the best neighbourhood to start a Xing Fu Tang Bubble Tea franchise in Singapore.

Introduction: Business Problem

Originating from Taiwan in the early 1980s, **bubble tea** is a flavoured/milk tea beverage served with tapioca balls. In recent times, this tea drink has taken the world by storm. It's wildly popular among the youths of Asia, and more recently in the North America. This holds true for the Southeast-Asian city-state of **Singapore** as well.

When tightened circuit-breaker measures were announced in Singapore to curb the spread of COVID-19 on 21 April 2020, Singaporeans went out in droves to get their fix of bubble tea. Searches for the drink on Google spiked, and for the few bubble tea shops that were still allowed to operate during the lockdown, they find themselves running out of tapioca pearls within hours of opening.

As the retail sector gradually recovers from the COVID-19 pandemic in 2021, the mainstays of the Singaporean bubble tea scene should anticipate a return to normalcy and use the opportunity to scale mindfully. This includes one of Singapore's well-known bubble tea brand, **Xing Fu Tang**. At the writing of this report, Xing Fu Tang operates 10 outlets in Singapore and is garnering positive reviews.

For this business problem, we'll use data science tools to fetch, visualise and analyse geolocation, demographic, and commercial data.

We begin by locating bubble tea stores in Singapore using the Foursquare API, and checking if they are often located near shopping malls and Mass Rapid Transit (MRT) stations. We'll then import and visualise population, income and household dwelling data of Singapore's Planning Areas/Subzones[1].

With the above, we'll use a clustering model to find Subzones with similar characteristics, finding patterns in the data that'll show us where best to open a new Xing Fu Tang outlet. Promising areas should also be situated away from existing Xing Fu Tang outlets as well.

[1] Singapore's first-level and second-level census divisions, respectively.

Data

Based on the problem defined, we look to group Subzones based on their following features:

- Planning Area Boundaries – To define Planning Area
- Subzone Boundaries – To define Subzones
- Location of existing Xing Fu Tang outlets – To see where Xing Fu Tang already has a presence [1]

- Location of bubble tea shops from other franchises (Foursquare API) – To understand bubble tea shop location trends and to score clusters negatively
- Location of Mass Rapid Transit (MRT) Stations – To map MRT station locations and relate them to bubble tea shop location trends
- Location of Shopping Malls (Foursquare API) – To map shopping mall locations and relate them to bubble tea location trends
- Population of Target Demographic (20 - 44 years old) – To relate target demographic population in Subzones to bubble tea shop locations [2] Reference [3]
- Median Income of Residents – To relate median income and bubble tea shop locations [4]
- Aggregation of Dwelling Types – To relate household dwelling types and bubble tea shop locations [2][5]

[1] Locations scraped from official website

[2] Population Trends > Singapore Residents by Planning Area, Subzone, Age Group, Sex and Type of Dwelling, June 2011-2020

[3] Assumption: Share of bubble tea consumers in Singapore in 2021 is similar to that of China in 2019

[4] General Household Survey 2015 > Resident Working Persons Aged 15 Years and Over

[5] e.g. HDB Flats, Landed Property etc.

Python Modules

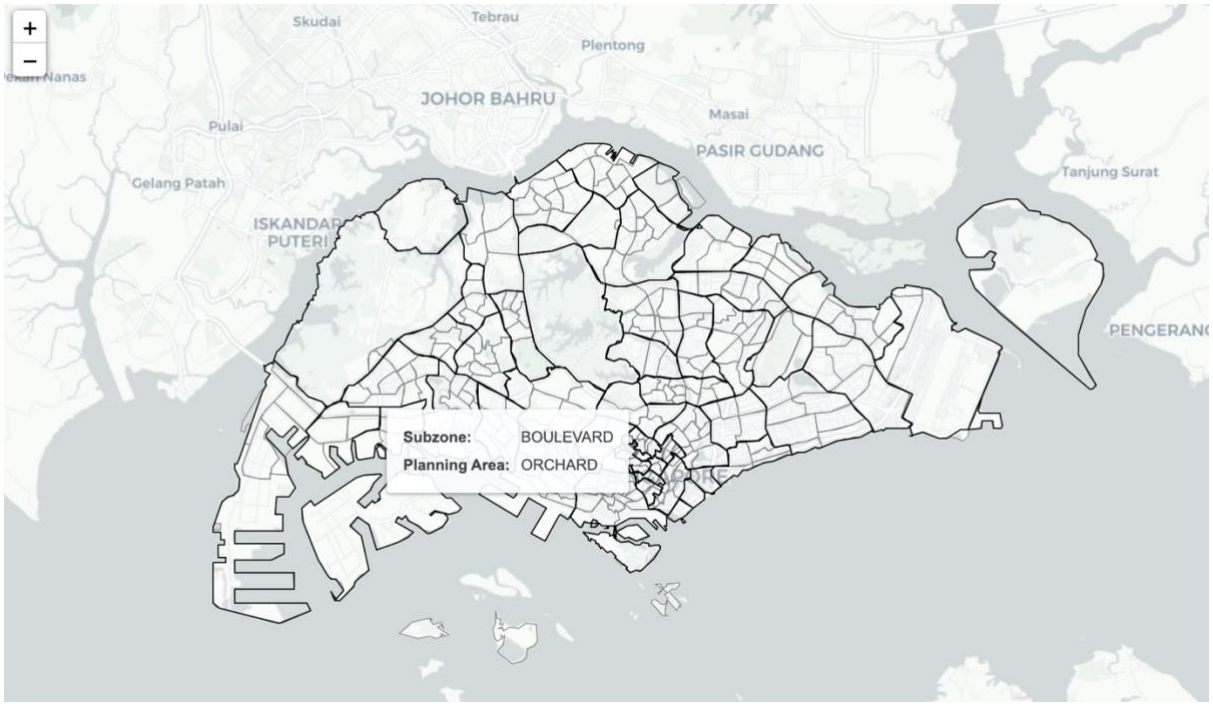
- Pandas — Data manipulation and analysis library
- Folium — Mapping module that visualises data on Leaflet.js maps
- Beautiful Soup — Used for web scraping
- Geopandas — Adds support for geographic data to pandas objects
- geopy.geocoders.Nominatim — Function that returns OpenStreetMap data when given an address
- Shapely — Geometric object manipulation
- sklearn — Machine learning library
- Matplotlib — Data visualisation library
- Branca — Folium element/colour manipulation
- requests — To send HTTP requests
- Miscellaneous: math, os

Boundary Data Preparations

Let's begin by preparing our canvas, namely a map of Singapore with her Planning Area and Subzone boundaries drawn.

We'll read the .geojson (an open standard format designed for representing simple geographical features) files into their geopandas dataframes, and plot our first map using Folium.

Folium makes it easy to visualize data that's been manipulated in Python on an interactive leaflet map.



Singapore with Planning Areas (black lined) and Subzones (grey lined)

We've plotted our first Folium graph. Note that the maps on the Jupyter Notebook preview here are interactive. This applies to all maps moving forward.

Bubble Tea Shop Locations

We'll follow up by getting the locations of Xing Fu Tang and other bubble tea brand outlets.

Scraping Xing Fu Tang Locations in Singapore

We could get Xing Fu Tang's outlet locations via the Foursquare API, but for the sake of variety and increased accuracy, we'll scrape their addresses from xingfutangsg.com using Beautiful Soup.

Beautiful Soup is a library that makes it easy to scrape information from web pages.

We'll use the Nominatim geolocator from OpenStreetMap. As the search criteria are somewhat strict, our input to the search function will be Mall or general area location accordingly.

Output: ['Ang Mo Kio ', 'Bukit Merah', 'Causeway Point', 'Century Square', 'Compass One', 'Hillion Mall', 'Paya Lebar Square', 'Plaza Singapura', 'Takashimaya', 'Northpoint City']

In the output, we see the names of the 10 Xing Fu Tang locations.

Surprisingly, searching for "Ang Mo Kio MRT" returns no results. However, searching for "Ang Mo Kio" returns the MRT station. We'll use that instead.

Other Bubble Tea Shop Locations in Singapore

Now we'll use the Foursquare API to retrieve the store location of some well-known bubble tea brands. I've referred to this top-10 list for this project.

The Foursquare Places API provides location based experiences with diverse information about venues, users, photos, and check-ins.

To access the Foursquare API, you can sign-up for a Foursquare Developer account here. The free plan is limited, but its features are enough for this purpose.

Note that we use the Category ID for "Bubble Tea Shop", which is "52e81612bcbc57f1066b7a0c". A list of Category IDs can be found here.

	id	name	categories	referralId	hasPerk	location.address	location.crossStreet	location.lat	location.lng
0	583509f3de0cbc3f332c52cd	KOI Café	[{'id': '52e81612bcbc57f1066b7a0c', 'name': 'B...	v-1619602441	False	#01-12	11 Collyer Quay	1.283848	103.851188
1	5155343de4b02587d19774e6	KOI Thé	[{'id': '52e81612bcbc57f1066b7a0c', 'name': 'B...	v-1619602441	False	#01-68 Plaza Singapura	68 Orchard Rd	1.299208	103.845523
2	4be3b0f0cbdbef3b609f60d8	KOI Café	[{'id': '52e81612bcbc57f1066b7a0c', 'name': 'B...	v-1619602441	False	#01-15 Bugis+	201 Victoria St	1.298976	103.854218
3	59f81740628c8321df3354e9	KOI Café	[{'id': '52e81612bcbc57f1066b7a0c', 'name': 'B...	v-1619602441	False	#B1-71, Raffles City Shopping Centre, 252 Nort...	NaN	1.293310	103.853674
4	50ccc614d63ee855763ee029	KOI Café	[{'id': '52e81612bcbc57f1066b7a0c', 'name': 'B...	v-1619602441	False	#01-85 Millenia Walk	9 Raffles Blvd	1.293392	103.859805
...	...	...	...	...	...	...	...	...	...
224	527f251e11d2140ef5361e52	LiHO	[{'id': '52e81612bcbc57f1066b7a0c', 'name': 'B...	v-1619602446	False	678A Woodlands Ave 6 #01-15	NaN	1.440565	103.801462
225	4e0a9c1b62e1c76af9ed0cec	LiHO	[{'id': '52e81612bcbc57f1066b7a0c', 'name': 'B...	v-1619602446	False	165, Bukit Merah Central, #01-3661	NaN	1.282981	103.817284
226	4e6c2a87b0fba3f50e33cf42	LiHO	[{'id': '52e81612bcbc57f1066b7a0c', 'name': 'B...	v-1619602446	False	Compass One B1-43	Sengkang Square	1.392401	103.895237
227	4d2da39a6e1eb1f76e750e5f	LiHO	[{'id': '52e81612bcbc57f1066b7a0c', 'name': 'B...	v-1619602446	False	#02-18A Junction 8	9 Bishan Pl	1.350387	103.848247
228	4f76d0c0e4b01af71245368a	LiHO	[{'id': '52e81612bcbc57f1066b7a0c', 'name': 'B...	v-1619602446	False	#02-66 AMK Hub	53 Ang Mo Kio Ave 3	1.369103	103.848241

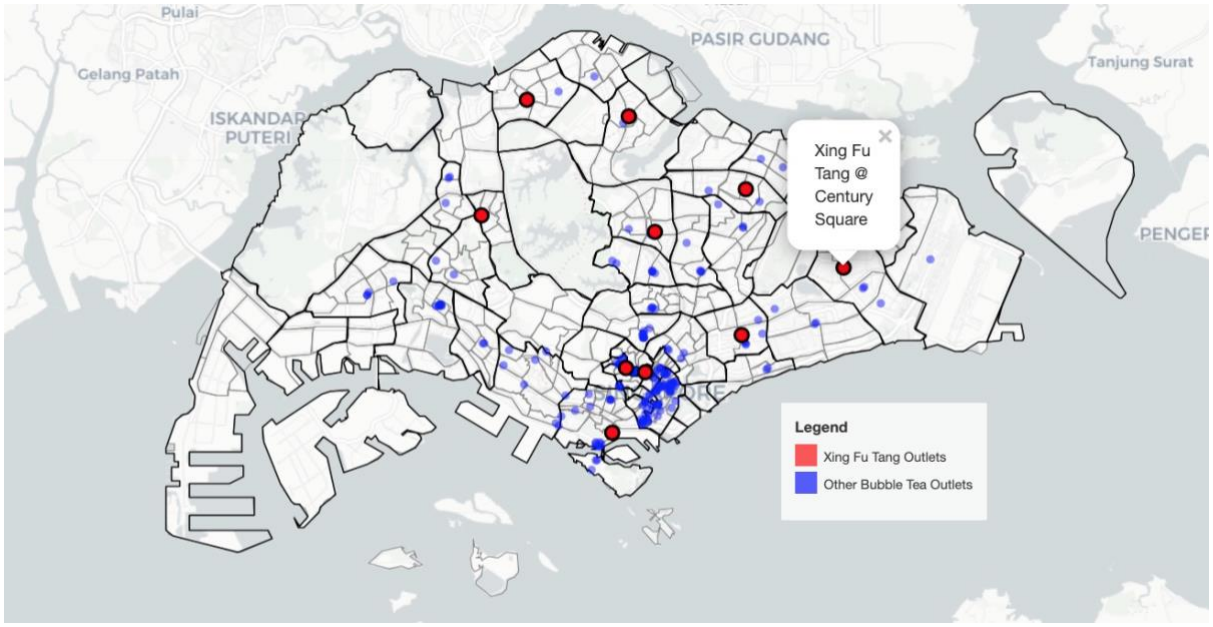
229 rows × 24 columns

Dataframe of most Bubble Tea shops location in Singapore

A dataframe is akin to a table, a 2D labelled data structure with columns, signifying data features (i.e. attributes) and rows, instances of data.

Plotting Locations of Xing Fu Tang and other Bubble Tea outlets

Now that we have location data for Xing Fu Tang outlets and stores of other bubble franchise, we can plot it onto our map.



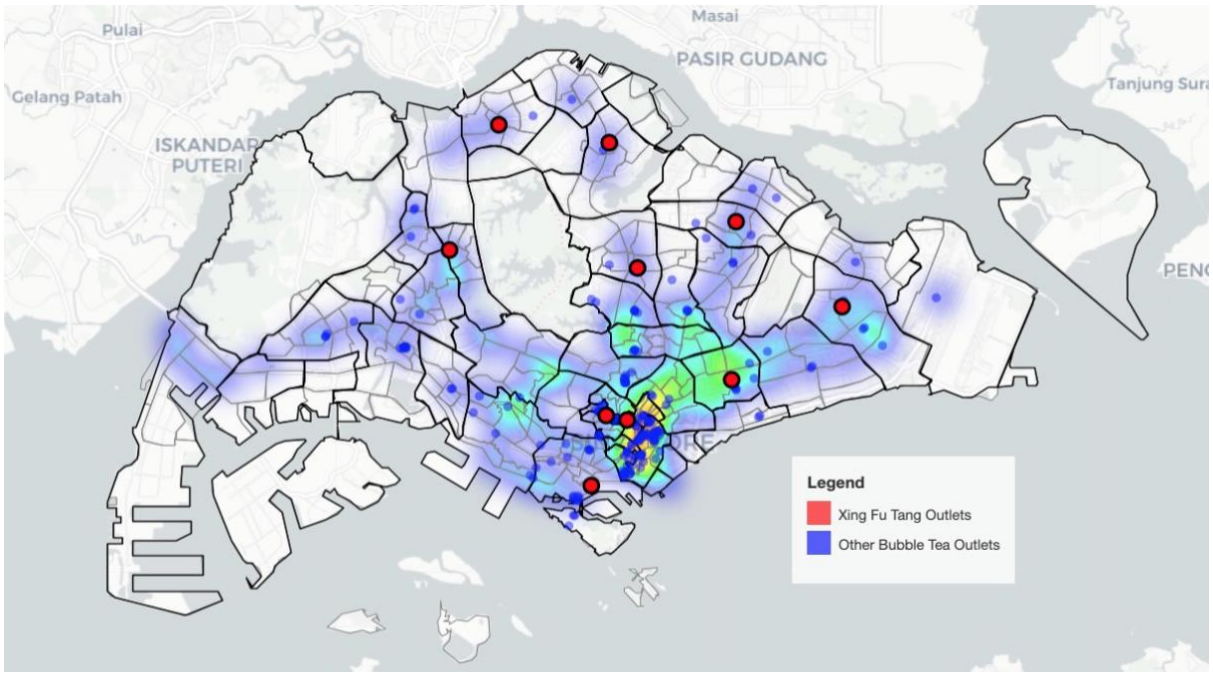
Perfect. We can see that bubble tea outlets are concentrated in the central region, branching mainly towards the north and east. Otherwise, we don't have much to go on about. Let's look at our next datasets.

MRT Stations and Malls Location

It's not uncommon to find a bubble tea store close to a mall or a mass transportation facility. Therefore, it's worth exploring the relationship between MRT/Mall locations with bubble tea outlets.

Let's retrieve, plot and compare these locations. To avoid having too many dots on the map, we'll use the Folium HeatMap function to plot an MRT HeatMap.

Credit goes to [yxlee245 on Kaggle] for compiling the coordinates of MRT stations in Singapore.



MRT Heatmap

Here we see that the location of bubble tea shops in Singapore has some correlations with the location of MRT stations.

There are some areas with a high density of MRT stations but with minimal/no bubble tea stores. Some examples include:

- Bukit Timah/Novena/Tanglin
- Kallang/Geylang boundary
- Tuas/Pioneer

One might assume that these are predominately residential areas with abundant MRT connections but minimal commercial zones, or industrial areas with basic commercial enterprises.

Nonetheless, the majority of bubble tea outlets are situated in the vicinity of an MRT station.

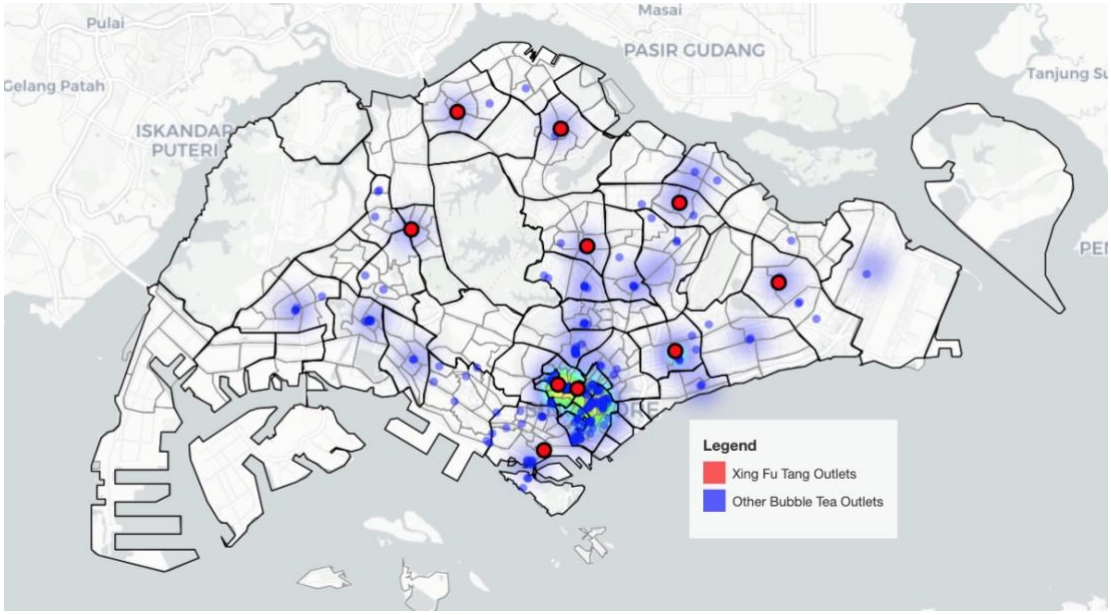
Shopping Mall Locations vs. Bubble Tea Shop Locations

We proceed by querying the locations of Shopping Malls from the Foursquare API. The Category ID for "Shopping Mall" is "4bf58dd8d48988d1fd941735".

	id	name	categories	referralId	hasPerk	location.address	location.lat	location.lng	location.labeledLatLn
0	5c73a4e61acf11002c69c41a	Funan	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	v-1619602452	False	107 North Bridge Road	1.291333	103.850121	[{'label': 'display', 'la 1.291332683974342
1	4c853176ee6fef3bd6893a5c	The Clementi Mall	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	v-1619602452	False	3155 Commonwealth Ave West	1.315036	103.764909	[{'label': 'display', 'la 1.315035580780165
2	4b058816f964a520f0b022e3	Tampines Mall	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	v-1619602452	False	4 Tampines Central 5	1.352567	103.944896	Na
3	4b058815f964a520a5b022e3	Ngee Ann City	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	v-1619602452	False	391 Orchard Rd.	1.302546	103.834566	[{'label': 'display', 'la 1.30254551168005
4	54813407498ef51e6987c675	Paya Lebar Square	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	v-1619602452	False	60 Paya Lebar Road	1.318632	103.892627	[{'label': 'display', 'la 1.318631705977262

First 5 rows of mall locations details dataframe

Similarly, we plot a heatmap of mall locations in Singapore.



Malls Heatmap

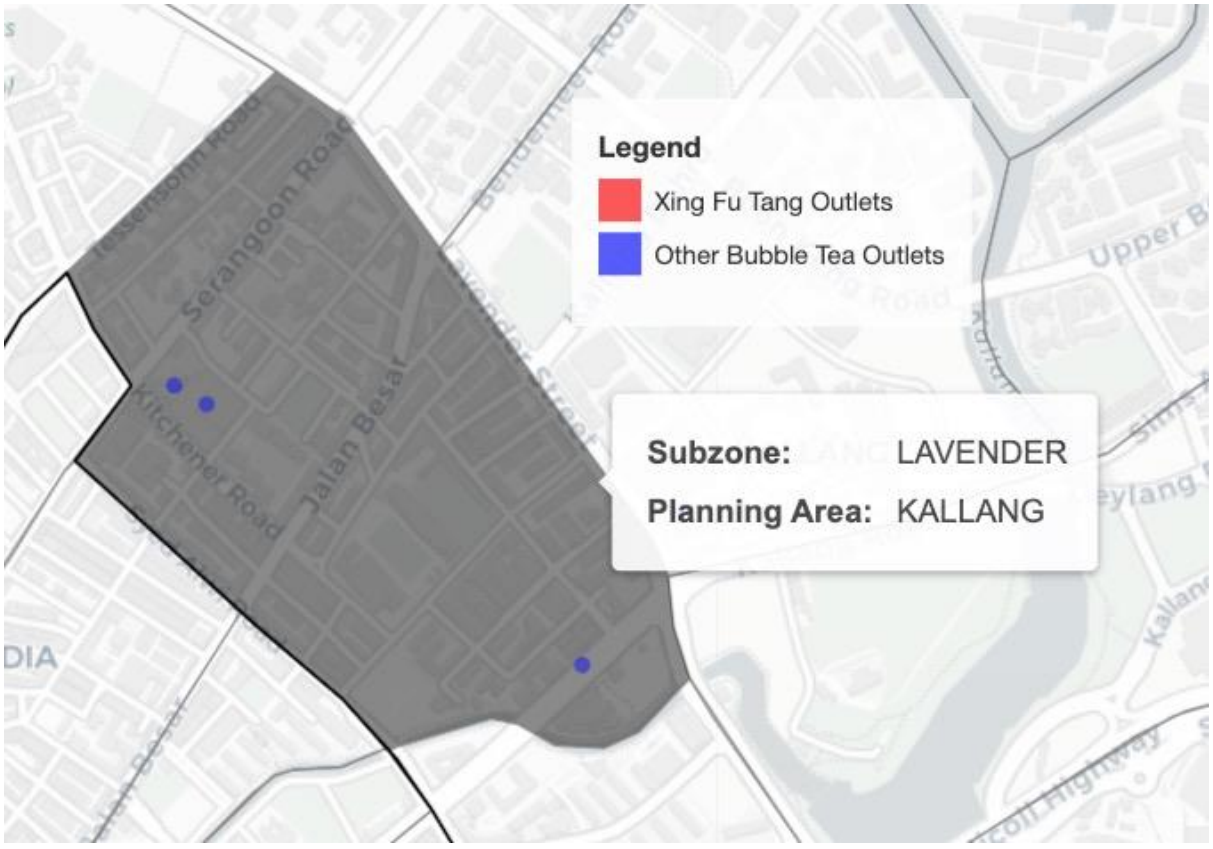
Here we also see a relationship between shopping mall and bubble tea shop locations. Some examples of this relationship include:

- Orchard (Planning Area)
- Downtown Core (Planning Area)
- Tampines East
- Geylang East
- Jurong Gateway.

We see most shopping malls have at least 1 bubble tea shop, and a sizable group of bubble tea shops aren't close to shopping malls.

Compiling Location Data

To give some features (characteristics) to our Subzones so that we may group them later on, we'll need to count how many bubble tea shops, Xing Fu Tang outlets, MRT stations, and malls are in a given Subzone.



We'd like to have a function that returns a 3 given Lavender's polygon, and the coordinates of the 3 outlets in these Subzones

For this, we define a function that counts the number of locations that are in a given area. Inputs include the coordinates of a point location (e.g. a mall) and the area boundary (e.g. Boulevard subzone).

Great, we loops through all Subzones and counts the number of Bubble Tea Stores in each. This is repeated for Xing Fu Tang outlets, MRT stations, and malls.

	Subzone	Planning Area	geometry	other_boba_count	xft_boba_count	mrt_count	mall_count
0	MARINA EAST	MARINA EAST	POLYGON Z ((103.88025 1.28386 0.00000, 103.880...	0.0	0.0	0.0	0.0
1	INSTITUTION HILL	RIVER VALLEY	POLYGON Z ((103.83764 1.29560 0.00000, 103.837...	0.0	0.0	0.0	0.0
2	ROBERTSON QUAY	SINGAPORE RIVER	POLYGON Z ((103.83410 1.29248 0.00000, 103.834...	3.0	0.0	0.0	1.0
3	JURONG ISLAND AND BUKOM	WESTERN ISLANDS	MULTIPOLYGON Z (((103.71253 1.29163 0.00000, 1...	0.0	0.0	0.0	0.0
4	FORT CANNING	MUSEUM	POLYGON Z ((103.84718 1.29700 0.00000, 103.847...	0.0	0.0	0.0	0.0

First 5 rows of the Subzone dataframe

In the end, we'll get something like the above.

Demographic Data

Besides location data, we'd like to also use some demographic data to cluster the Subzones.

The metrics we'll look at are:

- Population of 20–44 year olds by Subzones
- Dwelling Type by Subzones
- Median Income by Planning Area

Let's start with Population data.

Data of Population of 20–44 year olds by Subzones

To have a more accurate clustering of Subzones, it's worth identifying the age segment(s) that are more relevant to our studies.

A market study conducted in China in 2019 shows that the majority of bubble tea consumers are born between 1980 and 1999, which translates to an age range of 22 to 41.

Let's assume that:

- Bubble tea consumption patterns are similar in both China and Singapore.
- The age range remains similar in both 2019 and 2021.

We'll select the most relevant ages in the data below and sum them by age group and Subzone.

The data source from the Singapore Department of Statistics (DOS) is segmented by Subzone/Planning Area, Age Group, Sex, and Type of Dwelling.

For population data, let's create a new dataframe for Subzone and age segmentation. The age brackets relevant to us are '20_to_24','25_to_29', '30_to_34', '35_to_39' and '40_to_44'. We'll create a column called "pop_total20_44" to sum up the numbers in these ranges by Subzones.

	Subzone	Planning Area	geometry	other_boba_count	xft_boba_count	mrt_count	mall_count
0	MARINA EAST	MARINA EAST	POLYGON Z ((103.88025 1.28386 0.00000, 103.880...	0.0	0.0	0.0	0.0
1	INSTITUTION HILL	RIVER VALLEY	POLYGON Z ((103.83764 1.29560 0.00000, 103.837...	0.0	0.0	0.0	0.0
2	ROBERTSON QUAY	SINGAPORE RIVER	POLYGON Z ((103.83410 1.29248 0.00000, 103.834...	3.0	0.0	0.0	1.0
3	JURONG ISLAND AND BUKOM	WESTERN ISLANDS	MULTIPOLYGON Z (((103.71253 1.29163 0.00000, 1...	0.0	0.0	0.0	0.0
4	FORT CANNING	MUSEUM	POLYGON Z ((103.84718 1.29700 0.00000, 103.847...	0.0	0.0	0.0	0.0

First 10 rows of age dataframe

We've now gotten a dataframe that contains population segmented by Subzone and age, as well as generated a column with the total population of Subzone residents aged between 20 and 44 years.

Data of Type of Dwelling by Subzones

Another data feature that may help us cluster similar Subzones is the Type of Dwelling. Information about the Types of Dwellings can be found here.

This data property may give insights into the type of residents, real estate value, and general wealth of a Subzone.

Let's give some weights to the type of dwellings found in Singapore:

- Others: 1
- HDB 1- and 2-Room Flats: 2
- HDB 3-Room Flats: 3
- HDB 4-Room Flats: 4
- HDB 5-Room and Executive Flats: 5
- HUDC Flats (excluding those privatised): 6
- Condominiums and Other Apartments: 7
- Landed Properties: 8

These dwelling weights are arbitrary (A 4-room HDB flat might not cost twice as much as a 1 or 2-rooms HDB flat) but they are unequally weighted to represent relative values. An area of improvement can be to adjust the weights so they better represent the property values.

We'll calculate a Dwelling Index weighted by the population of residents living in a type of dwelling. Not to be confused by the dwelling weight above.

The formula for calculating the weighted average for the Dwelling Index is:

$$\bar{D} = \frac{\sum_{i=1}^{n} P_i Dw_i}{\sum_{i=1}^{n} P_i}$$

Where:

- $\bar{D}$ = Dwelling Index
- P_i = Population of Residents
- Dw_i = Dwelling Weight

	Subzone	Planning Area	geometry	other_boba_count	xft_boba_count	mrt_count	mall_count	dwell_idx	pop_total	pop_total20_44
0	MARINA EAST	MARINA EAST	POLYGON Z ((103.88025 1.28386 0.00000, 103.880...	0.0	0.0	0.0	0.0	NaN	0	0
1	INSTITUTION HILL	RIVER VALLEY	POLYGON Z ((103.83764 1.29560 0.00000, 103.837...	0.0	0.0	0.0	0.0	6.980892	3140	1190
2	ROBERTSON QUAY	SINGAPORE RIVER	POLYGON Z ((103.83410 1.29248 0.00000, 103.834...	3.0	0.0	0.0	1.0	6.979933	2990	1090
3	JURONG ISLAND AND BUKOM	WESTERN ISLANDS	MULTIPOLYGON Z (((103.71253 1.29163 0.00000, 1...	0.0	0.0	0.0	0.0	NaN	0	0
4	FORT CANNING	MUSEUM	POLYGON Z ((103.84718 1.29700 0.00000, 103.847...	0.0	0.0	0.0	0.0	7.000000	180	50

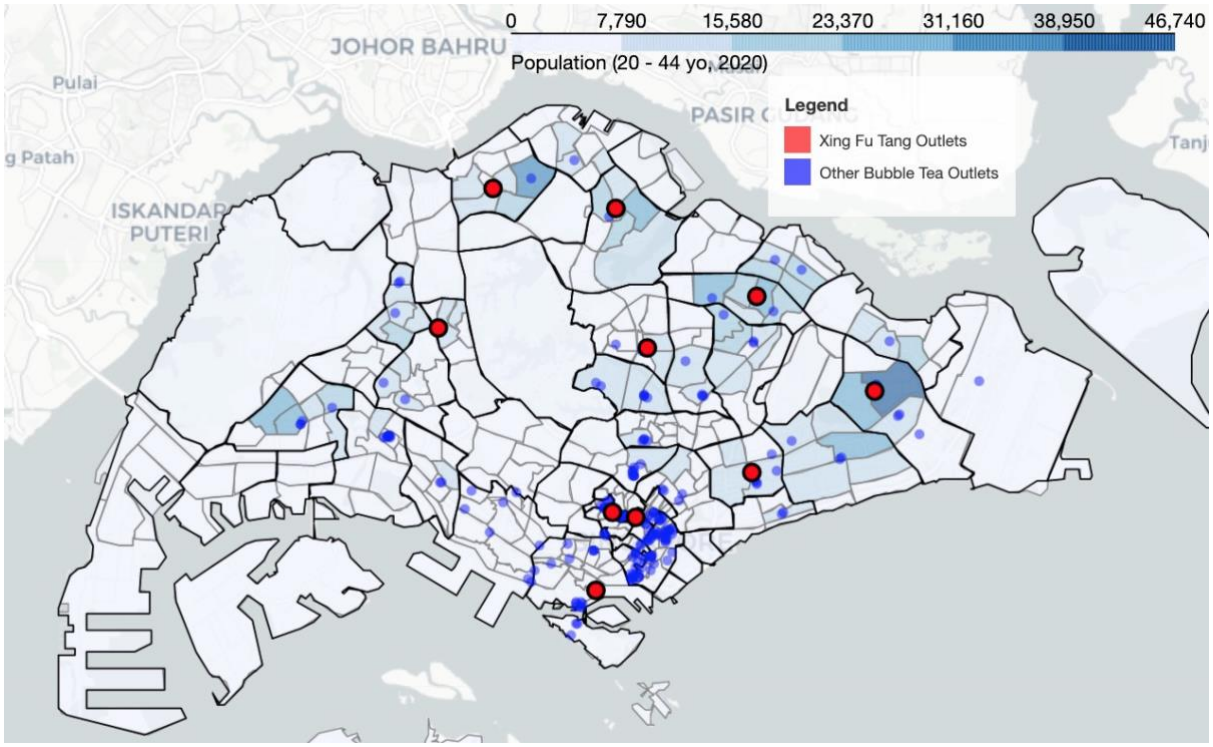
First 5 rows of updated Subzone dataframe

Here we see that Fort Canning has a Dwelling Index of 7, hinting at it being a region of relatively higher wealth.

Plotting of Population Data

Now that we've gotten our Population (20–44 yo) and Dwelling Index data, let's plot them onto maps. We'll use the Folium choropleth function to generate the following maps

Choropleth Maps display divided geographical areas or regions that are coloured, shaded or patterned in relation to a data variable.



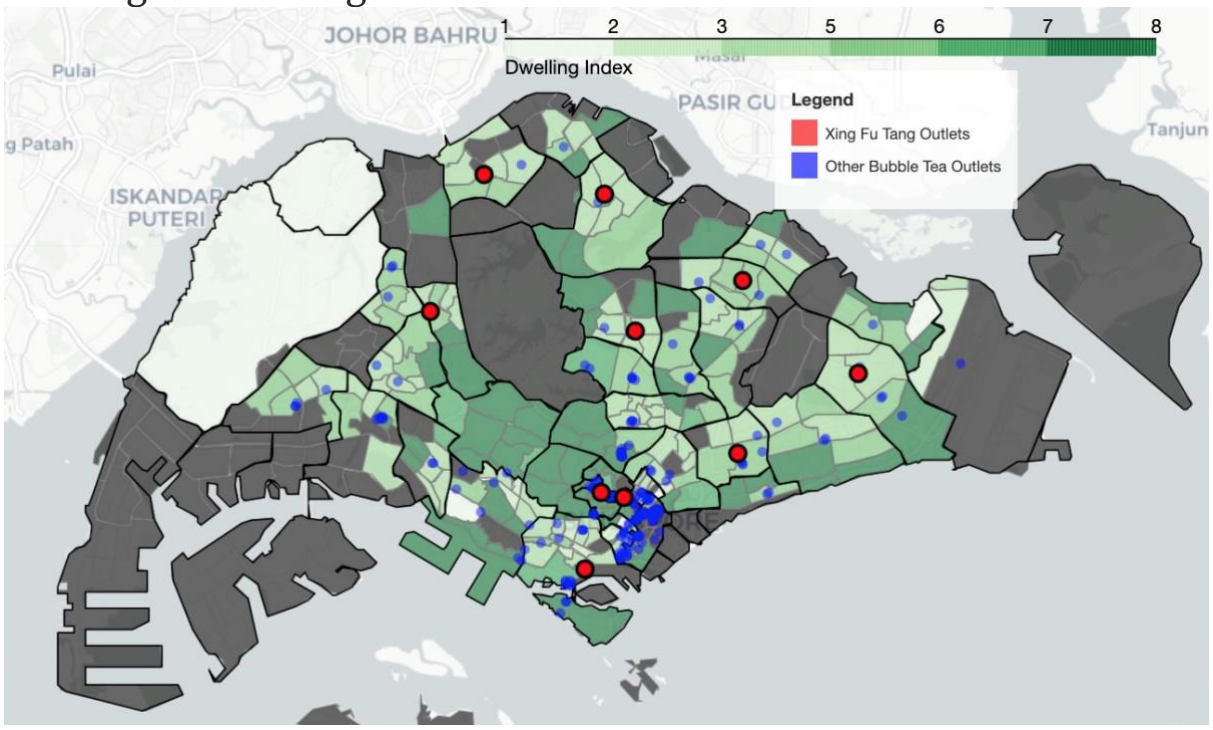
Choropleth of population (20–44 yo)

On the map, we see that Singaporeans aged between 20 to 44 years live in Subzones away from the central region, mainly in Jurong West, Woodlands, Yishun, Sengkang/Punggol and Tampines/Bedok, among other Subzones.

We also see that current Xing Fu Tang outlets are close to the majority of these population hotspots, suggesting that our methodology is in line with Xing Fu Tang's Singapore strategy.

Let's now plot the Dwelling Index choropleth.

Plotting of Dwelling Index Data



Choropleth of dwelling index

On the map above, we see tells of 3 trends.

Firstly, we see that 7 of the 10 Xing Fu Tang's current outlets are located in areas of medium dwelling index (the higher the index, the "wealthier" the Subzones). These are also generally areas with large populations of 20–44 year olds.

Secondly, we see that 3 of the 10 Xing Fu Tang's current outlets are within the central region, in areas of medium-high dwelling index.

Thirdly, we rarely see any bubble tea outlets in areas of high dwelling index. These are also generally areas of lower population of 20–44 year olds.

These trends apply to other bubble tea brand outlets as well.

We can hypothesise that most bubble tea brands cater towards areas of medium wealth, with the exception being in the city centre (where they potentially cater to white-collar workers or retain a presence for brand relevance). Areas of high dwelling index may also be less densely populated as landed properties don't house as many people as flats.

Median Income by Planning Area

To add substance to our analysis on Dwelling Types, we'll also analyse the median income of residents by Planning Area.

The income data segments residents into monthly income (SGD) brackets ranging from “Below 1,000” to “5,000–5,999” to “12,000 & Over”. The data source from the Singapore Department of Statistics (DOS).

Planning Area	Below $1,000	$1,000 - $1,4	$1,500 - $1,9	$2,000 - $2,4	$2,500 - $2,9	$3,000 - $3,9	$4,000 - $4,9	$5,000 - $5,9	$6,000 - $6,9	$7,000 - $7,9	$8,000 - $8,9	$9,000 - $9,9	$10,000 - $10	$11,000 - $11	$12,000 & Over
ANG MO KIO	9.7	12.1	7.9	7.4	6.8	11.5	9.8	7.9	6	4	3.1	2.2	2.6	1.7	8.6
BEDOK	12.2	13.6	12.1	9.7	9.6	17.2	13.4	12.2	9.2	5.4	5.1	4	4.5	2.7	19.4
BISHAN	3.9	3.7	2.2	2.9	2.4	4.6	4.7	3.8	2.7	2.8	3.1	1.8	1.8	1.6	7.6
BUKIT BATOI	6.1	6.5	5.2	5.8	4.9	8.6	7.6	6.7	4.4	3.5	2.6	2.4	2.5	1.5	7.4
BUKIT MERA	8.5	9.2	6.9	6.6	4.4	8.3	7	5.5	4.7	4.2	3	2.1	2.2	1.5	8

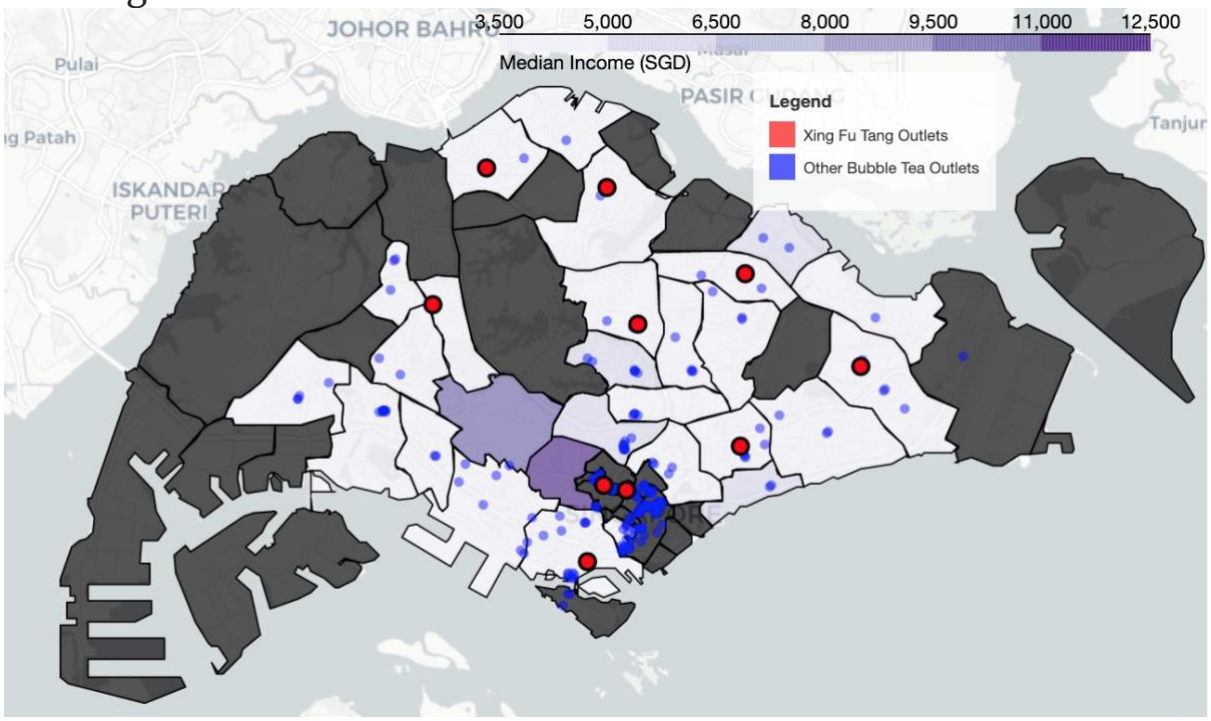
Screenshot of .csv file

To get the median income, we sum across the rows cumulatively, divide the total by 2 (getting the median population number). The median income bracket will be the bracket where the median population number lies.

Planning Are	Below $1,00	$1,000 - $1,4	$1,500 - $1,9	$2,000 - $2,4	$2,500 - $2,9	$3,000 - $3,9	$4,000 - $4,9	$5,000 - $5,9	$6,000 - $6,9	$7,000 - $7,9	$8,000 - $8,9	$9,000 - $9,9	$10,000 - $1(	$11,000 - $11	$12,000 & O	cum_med_pc	med_inc_bra	median_inc
ANG MO KIO	9.7	21.8	29.7	37.1	43.9	55.4	65.2	73.1	79.1	83.1	86.2	88.4	91	92.7	101.3	50.65	$3,000 - $3,9	3500
BEDOK	12.2	25.8	37.9	47.6	57.2	74.4	87.8	100	109.2	114.6	119.7	123.7	128.2	130.9	150.3	75.15	$4,000 - $4,9	4500
BISHAN	3.9	7.6	9.8	12.7	15.1	19.7	24.4	28.2	30.9	33.7	36.8	38.6	40.4	42	49.6	24.8	$5,000 - $5,9	5500
BUKIT BATC	6.1	12.6	17.8	23.6	28.5	37.1	44.7	51.4	55.8	59.3	61.9	64.3	66.8	68.3	75.7	37.85	$4,000 - $4,9	4500
BUKIT MERA	8.5	17.7	24.6	31.2	35.6	43.9	50.9	56.4	61.1	65.3	68.3	70.4	72.6	74.1	82.1	41.05	$3,000 - $3,9	3500
BUKIT PANJ	5.1	11.6	16.8	23.1	29.6	40.1	48.4	55.6	61.1	64.9	68.3	70.9	72.9	73.9	80	40	$3,000 - $3,9	3500
BUKIT TIMAI	2.1	3.5	4.7	5.9	6.6	8.7	10.7	13.6	14.9	16.8	18.4	19.9	22.1	23.2	36.9	18.45	$9,000 - $9,9	9500
CHOA CHU I	7.3	15.1	21.4	28.9	35.4	49.9	60.5	68.5	74.4	79.8	84.8	87.4	89.6	90.9	97.4	48.7	$3,000 - $3,9	3500
CLEMENTI	3.8	7.4	9.8	12.1	14.7	19.7	24	27.6	30.7	33.9	35.6	37.4	39.1	40.1	46.2	23.1	$4,000 - $4,9	4500
GEYLANG	6.5	12.4	16.8	22.1	25.6	33.4	39.1	43.5	47	49.4	51.8	53.8	55.3	56.3	61.2	30.6	$3,000 - $3,9	3500
HOUGANG	10.1	20.9	29.1	39.5	47.2	64.2	76	87.1	94.5	99.5	103.6	106.4	109.6	111.5	120.7	60.35	$3,000 - $3,9	3500
JURONG EA	4.3	8.4	11.7	15.3	19.1	24.3	29.7	33.3	36.1	38.3	40	41	42.2	42.7	46.1	23.05	$3,000 - $3,9	3500
JURONG WE	13.4	27.2	38.8	51.1	62.2	85	104.3	117.9	126.6	136.2	142	146	149	150.8	158.5	79.25	$3,000 - $3,9	3500

Taking Ang Mo Kio as an example, we see that the total population is 101,300, half of that is 50,650, which falls into the income bracket of 3,000–3,999 (i.e. median income of SGD3,500).

Plotting of Median Income Data



Choropleth of median income

Plotting the median income onto the map, we see that the trend agrees with the trends we see in the Dwelling Index, namely that bubble tea outlets are mainly situated in middle-income areas.

So far, we've retrieved all the data required and visualised them to explore their contents. In the Analysis chapter, we'll prepare the dataframes to be used in the machine learning algorithm.

Methodology

The goal of this report is to identify optimal Subzones for Xing Fu Tang's next branch in Singapore.

We've retrieved/calculated the following data:

- Number of existing Xing Fu Tang outlets by Subzones
- Number of bubble tea shops from other franchises by Subzones
- Number of Mass Rapid Transit (MRT) Stations by Subzones
- Number of Shopping Malls by Subzones
- Population of Target Demographic (20–44 years old) by Subzones
- Median Income of Residents by Planning Area
- Aggregation of Dwelling Types (Dwelling Index) by Subzones
- Boundary Data for Subzones and Planning Area

Next, we'll use:

- K-nearest Neighbour algorithm to fill some of the missing values.
- StandardScalar to scale the data accordingly (to avoid biasing the model towards datasets large values).
- K-means method to cluster Subzones into similar groups of Subzones (clusters).
- A range of K values, Silhouette Score and the Elbow method to determine the best K to use for K-means.
- We'll present these findings graphically and come to a final decision.

Analysis

Data Review

We begin by reviewing the data we have and determine whether they are fit for processing (garbage in = garbage out).

	Subzone	Planning Area	geometry	other_boba_count	xft_boba_count	mrt_count	mall_count	dwell_idx	pop_total	pop_total20_44	median_inc
0	MARINA EAST	MARINA EAST	POLYGON Z ((103.88025 1.28386 0.00000, 103.880...	0.0	0.0	0.0	0.0	NaN	0	0	NaN
1	INSTITUTION HILL	RIVER VALLEY	POLYGON Z ((103.83764 1.29560 0.00000, 103.837...	0.0	0.0	0.0	0.0	6.980892	3140	1190	NaN

Our work dataframe shows missing values

A preview of what we have now shows that:

We have a mixture of columns with information (area names, geodata) and numeric values. We will have to split this dataframe into cluster_info and cluster_val as our machine learning algorithm takes only numeric values.

We have some missing values for Dwelling Index and Median Income, we'll have to plug those holes.

Data Cleansing

First, we split the columns.

We'll also have to remove the "pop_total" column as that data is linearly related to "pop_total20_44". Keeping "pop_total" in would result in multicollinearity, which gives unwanted additional weight to the population data.

Multicollinearity refers to a situation in which more than two explanatory variables in a multiple regression model are highly linearly related.

Second, let's clean up some missing values.

	other_boba_count	xft_boba_count	mrt_count	mall_count	dwell_idx	pop_total20_44	median_inc
0	0.0	0.0	0.0	0.0	NaN	0	NaN
1	0.0	0.0	0.0	0.0	6.980892	1190	NaN

For rows with 0 or NaN values only (e.g. row 1), we'll fill the NaN values with 0.

For rows with NaN and other values, we'll perform missing data imputation using the K-nearest Neighbour algorithm.

In short, K-nearest Neighbour works by approximating the value of a missing datapoint based on the values of neighbouring datapoints.

"Birds of the same feather flock together"

```
Anymore NaN values? False
```

	other_boba_count	xft_boba_count	mrt_count	mall_count	dwell_idx	pop_total20_44	median_inc
0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
1	0.0	0.0	0.0	0.0	6.980892	1190.0	5700.0

Great, let's move on.

Data Scaling

We start with scaling the data to avoid biases towards data points with larger numbers (e.g. population, median income).

StandardScaler works by removing the mean of a dataset from its datapoints and scaling the datapoints to unit variance.

Finding the Best K value for K-means

Despite the similar name, K-nearest Neighbour isn't the same as K-means.

- "K" in K-nearest Neighbour refers to number of nearest neighbours to an unlabeled datapoint the algorithm checks to label the said datapoint.

- “K” in K-means refers to the number of clusters the algorithm will attempt to “group” a dataset into. K = 5 means you’ll get 5 clusters in the end.

For our case, we’ll be using the Elbow method (plotted in blue below) and the Silhouette Score method (plotted in red below) to find the best K value for K-means.

- Elbow method, also known as the Sum of Squared Distance measures the error between a cluster’s centre and its datapoints. The smaller the value the better.
- Silhouette Score method studies the separation distance between clusters. The larger the value the better.

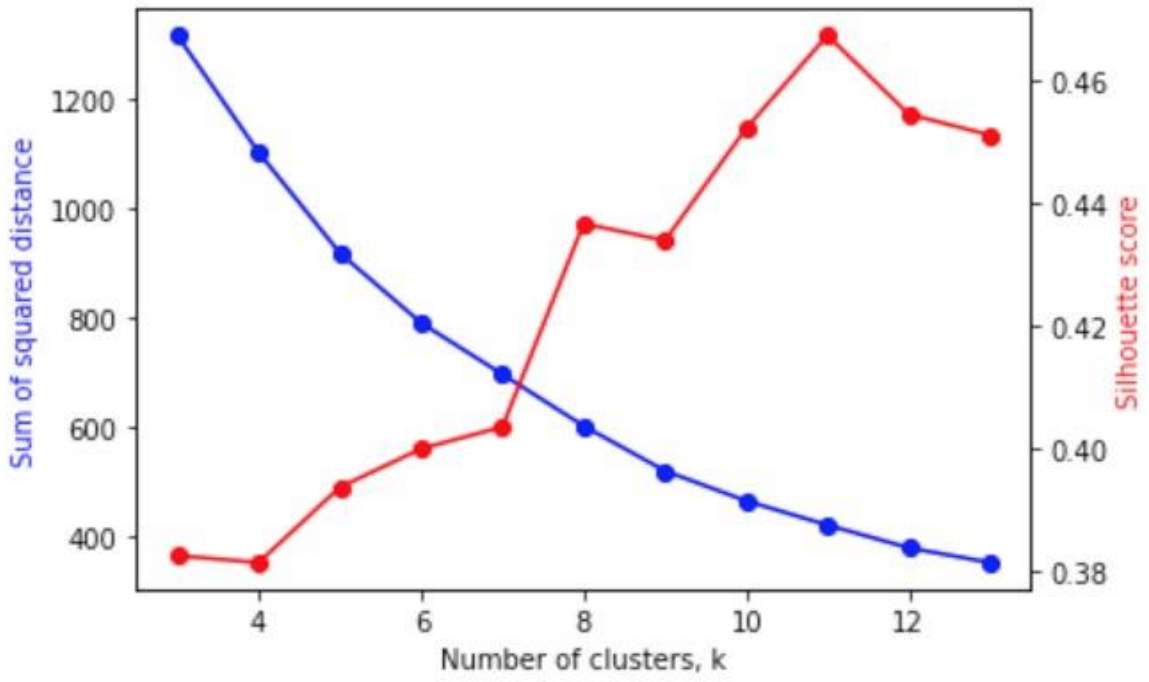
Potential Silhouette scores range from -1 to 1:

-1: the point is in the wrong cluster

0: the point is on the decision boundary between 2 neighbouring clusters

+1: the point is far away from neighbouring clusters. (Great!)

Credits: Tony Xu’s materials on clustering similar neighbourhoods have been a large help




```
Optimal k value: 11
```

In the graph above, we see a peak in the Silhouette score and the optimal K value. Our ideal number of clusters for our K-means clustering is 11.

K-means Clustering with Optimal K

Let's run the algorithm with the optimal K.

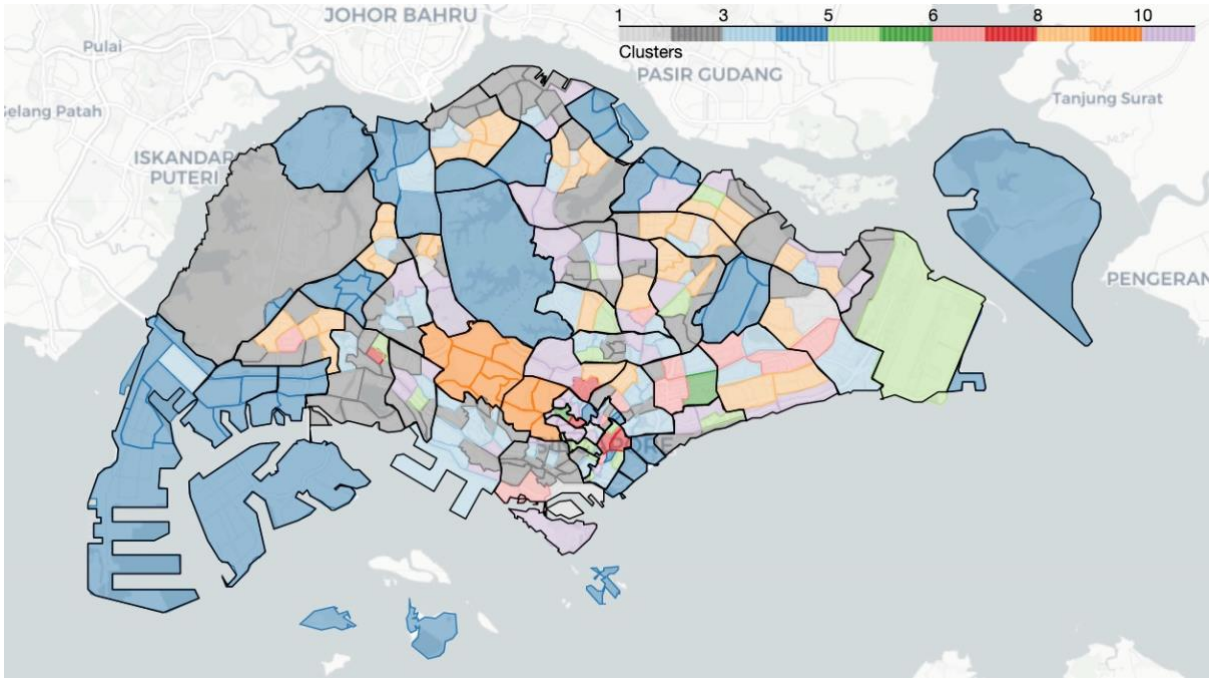
	Subzone	Planning Area	geometry	other_boba_count	xft_boba_count	mrt_count	mall_count	dwell_idx	pop_total20_44	median_inc	cluster
0	MARINA EAST	MARINA EAST	POLYGON Z ((103.88025 1.28386 0.00000, 103.880...	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	4
1	INSTITUTION HILL	RIVER VALLEY	POLYGON Z ((103.83764 1.29560 0.00000, 103.837...	0.0	0.0	0.0	0.0	6.980892	1190.0	5700.0	11
2	ROBERTSON QUAY	SINGAPORE RIVER	POLYGON Z ((103.83410 1.29248 0.00000, 103.834...	3.0	0.0	0.0	1.0	6.979933	1090.0	4100.0	5
3	JURONG ISLAND AND BUKOM	WESTERN ISLANDS	MULTIPOLYGON Z (((103.71253 1.29163 0.00000, 1...	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	4
4	FORT CANNING	MUSEUM	POLYGON Z ((103.84718 1.29700 0.00000, 103.847...	0.0	0.0	0.0	0.0	7.000000	50.0	2700.0	11

First 5 rows of clustered Subzone dataframe

In the dataframe above, we see that each Subzone has been assigned to a cluster.

Plotting Clusters onto Map

Let's colour-code the clusters and plot them onto a map. We'll use branca.colormap (imported here as cmp), a utility module to define the cluster colours.

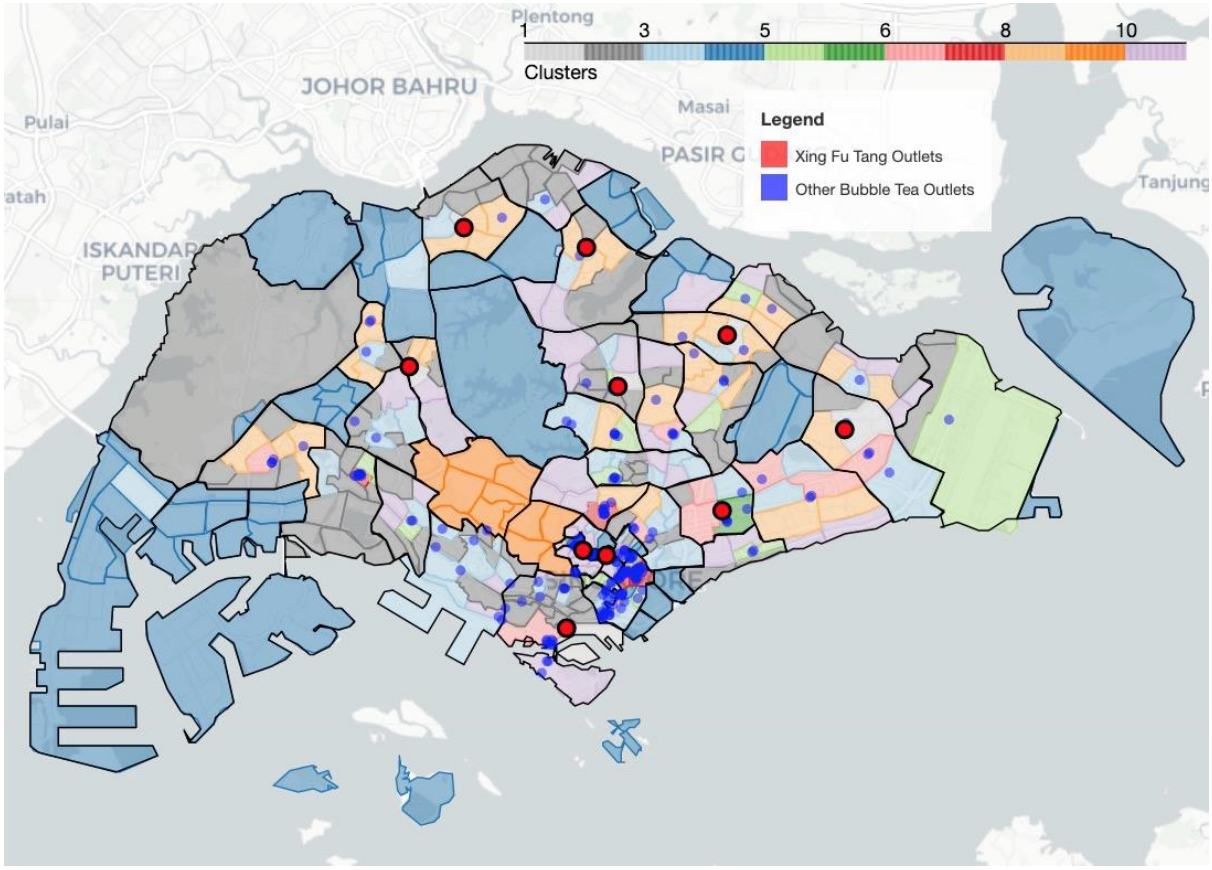


Map of 332 Subzones clustered into 11 clusters

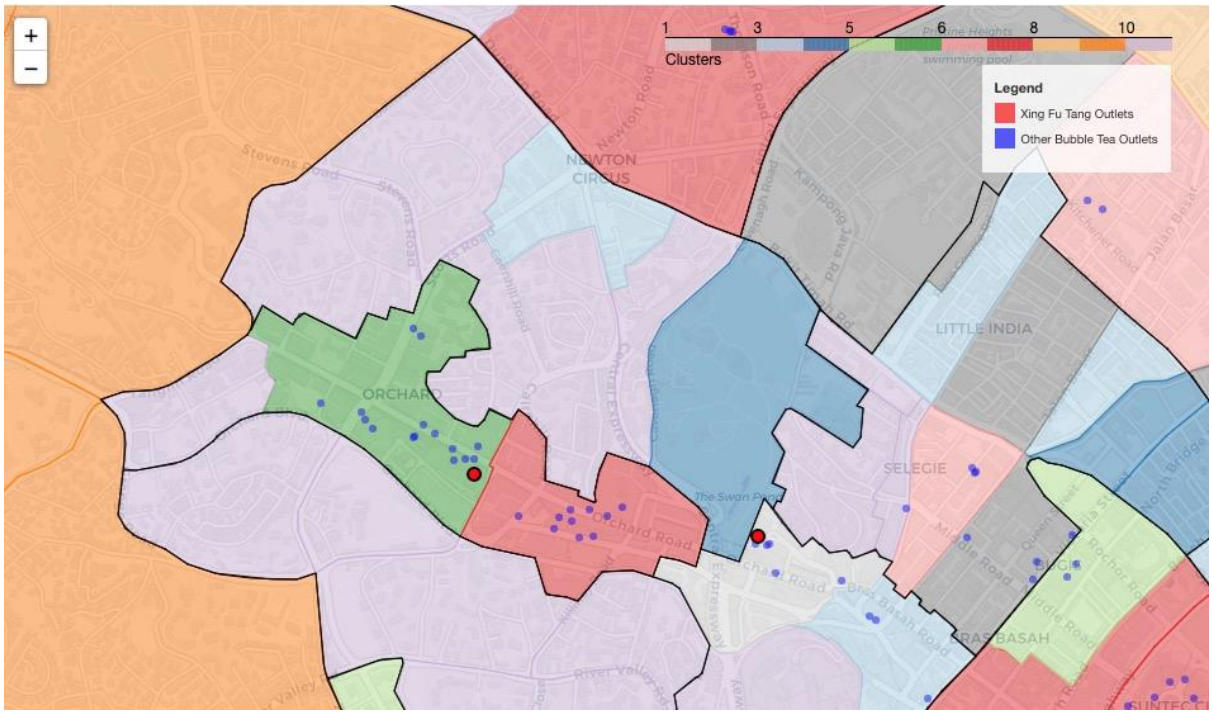
Voila, that's Singapore's Subzones clustered based on the traits we specified above.

Comparing Clusters with Bubble Tea Locations

Let's look at the clusters based on bubble tea locations.



Map of clusters vs. bubble tea outlet locations



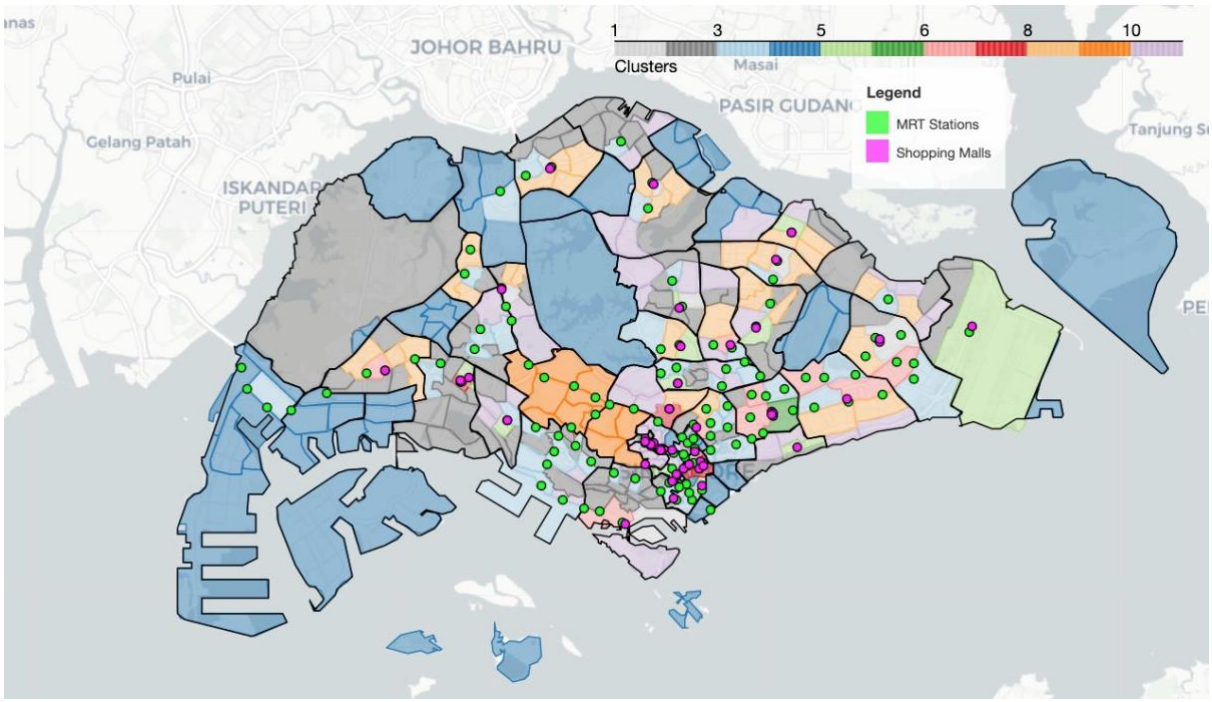
Here, we see that Subzones containing Xing Fu Tang outlets are clusters 1 (light grey) and 6 (dark green), with cluster 6 being very competitive for Xing Fu Tang.

Cluster 8 (dark red) contains Subzones with many bubble tea outlets but no Xing Fu Tang outlets.

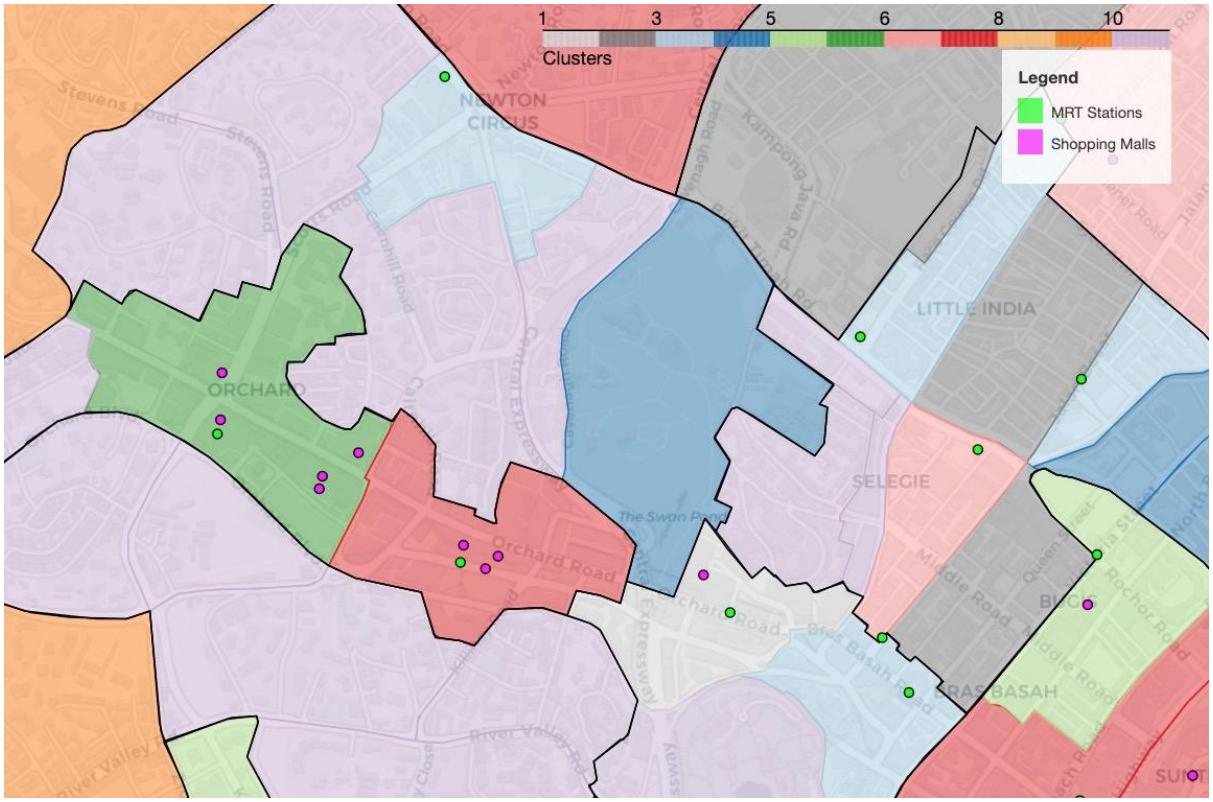
Clusters 2 (dark grey), 4 (dark blue), 10 (dark orange) and 11 (lilac) have a distinctively low number of bubble tea outlets.

Comparing Clusters with MRT/Mall Locations

Let's now look at clusters vs. MRT/Mall locations.



Map of clusters vs. MRT/mall locations



Since we've established that bubble tea shops are often found near MRT stations and Malls, it's no surprise to see similar trends here.

Scoring Clusters to find Best Clusters

We're not done. Currently, we do not have a metric to rank the clusters from best (good locations for Xing Fu Tang's new branch) to worst (locations Xing Fu Tang should avoid).

We can give each Subzone in a cluster a score based on the number of MRT/Malls (positive scores) and the number of bubble tea shops (negative scores) in a Subzone.

I've elected to go with the following scores:

- MRT: +3 (I assume 1 MRT station isn't fully saturated until it has 3 bubble tea shops)
- Mall: +5 (I assume 1 Mall isn't fully saturated until it has 5 bubble tea shops)
- Bubble tea shop from other brands: -1
- Existing Xing Fu Tang outlet: -10

Subzone	Planning Area	geometry	other_boba_c	xft_boba_cou	mrt_count	mall_count	dwell_idx	pop_total20_	median_inc	cluster	mrt_mall_sco	bubble_score	subzone_score
BEDOK NOR	BEDOK	POLYGON Z	2	0	2	1	4.04231191	26480	4500	7	11	2	9
ALJUNIED	GEYLANG	POLYGON Z	0	0	3	0	4.70793809	13740	3500	7	9	0	9
LORONG AH	HOUGANG	POLYGON Z	0	0	1	1	4.81448382	10550	3500	5	8	0	8
SOMERSET	ORCHARD	POLYGON Z	10	0	1	3	7	30	1800	8	18	10	8
JURONG WE	JURONG WE	POLYGON Z	3	0	2	1	4.64540896	24220	3500	7	11	3	8
LAVENDER	KALLANG	POLYGON Z	3	0	2	1	4.38888889	3100	3500	7	11	3	8
SERANGOO	SERANGOO	POLYGON Z	3	0	2	1	4.72466694	8430	4500	7	11	3	8
CHANGI AIR	CHANGI	POLYGON Z	1	0	1	1	0	0	3000	5	8	1	7

Both Bedok North and Aljunied scored highly as they have relatively few bubble tea shops for the number of MRT/Malls within their boundaries.

However tempting to conclude our report using just the individual Subzone scores, we need to remember to look at clusters as a whole as clusters take into account other features such as Dwelling Index, Population (20 to 44 yo) and Median Income.

Clusters with Best Average Scores

To get the best cluster, we “merge” all Subzones into their clusters by averaging their feature values. We pick the best cluster by the largest average Subzone score.

```
Cluster 7 has the highest subzone score of 6.27
```

cluster	other_boba_count	xft_boba_count	mrt_count	mall_count	dwell_idx	pop_total20_44	median_inc	mrt_mall_score	bubble_score	subzone_score
7	2.545455	0.0	2.181818	0.454545	4.226077	10360.000000	3954.545455	8.818182	2.545455	6.272727
8	10.200000	0.0	1.200000	2.200000	4.206674	696.000000	3640.000000	14.600000	10.200000	4.400000
5	2.333333	0.0	0.533333	1.000000	3.921935	4047.333333	3853.333333	6.600000	2.333333	4.266667
6	13.500000	1.0	1.500000	4.500000	5.822515	5105.000000	4000.000000	27.000000	23.500000	3.500000
3	0.604167	0.0	1.062500	0.000000	4.686203	3762.708333	4056.250000	3.187500	0.604167	2.583333
10	0.000000	0.0	0.666667	0.000000	7.179964	2575.833333	10500.000000	2.000000	0.000000	2.000000
9	0.285714	0.0	0.171429	0.000000	4.679342	16765.428571	4014.285714	0.514286	0.285714	0.228571
4	0.017544	0.0	0.070175	0.000000	0.035088	0.526316	177.192982	0.210526	0.017544	0.192982
2	0.146067	0.0	0.000000	0.000000	3.524181	3088.426966	3959.550562	0.000000	0.146067	-0.146067
11	0.160000	0.0	0.000000	0.000000	7.191772	1536.200000	4430.000000	0.000000	0.160000	-0.160000
1	3.375000	1.0	1.125000	0.875000	4.504840	11513.750000	3750.000000	7.750000	13.375000	-5.625000

Dataframe of clusters scored

	other_boba_count	xft_boba_count	mrt_count	mall_count	dwell_idx	pop_total20_44	median_inc	cluster	mrt_mall_score	bubble_score	subzone_s
count	11.000000	11.0	11.000000	11.000000	11.000000	11.00000	11.000000	11.0	11.000000	11.000000	11.00
mean	2.545455	0.0	2.181818	0.454545	4.226077	10360.00000	3954.545455	7.0	8.818182	2.545455	6.27
std	1.967925	0.0	0.404520	0.522233	1.595862	9119.15018	633.030231	0.0	2.926369	1.967925	2.45
min	0.000000	0.0	2.000000	0.000000	0.000000	0.00000	3500.000000	7.0	6.000000	0.000000	2.00
25%	1.500000	0.0	2.000000	0.000000	4.004896	2140.00000	3500.000000	7.0	6.000000	1.500000	4.50
50%	3.000000	0.0	2.000000	0.000000	4.388889	9410.00000	3500.000000	7.0	9.000000	3.000000	7.00
75%	3.000000	0.0	2.000000	1.000000	4.716303	13830.00000	4500.000000	7.0	11.000000	3.000000	8.00
max	7.000000	0.0	3.000000	1.000000	6.386555	26480.00000	5300.000000	7.0	14.000000	7.000000	9.00

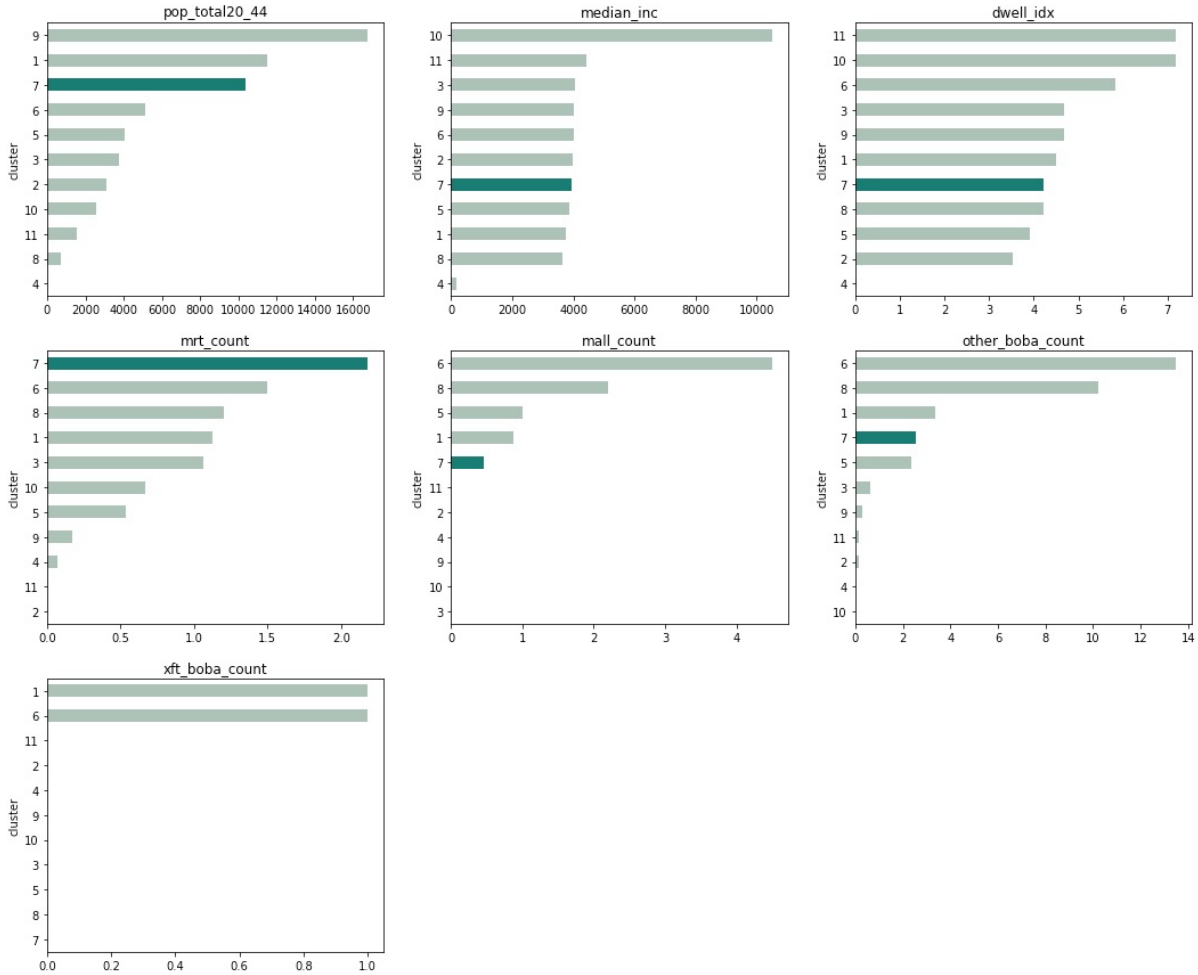
Description of data in cluster 7

We’ve found our top cluster (cluster 7). It has 11 Subzones in it (count = 11) with an average Subzone score of 6.3.

Feature Analysis of Best Cluster

Let’s explore our best cluster even further. First, we’ll compare the best cluster with its counterparts on all 7 features (e.g. Median Income, Dwelling Index etc.)

Credits: Tony Xu for a great bar chart function



Feature analysis of cluster 7 against other clusters

In the bar charts above, we see Cluster 7 ranked 3rd in terms of the average Population of 20–44 year olds.

In terms of average median income and dwelling index, Cluster 7 is consistent with our previous considerations that bubble tea shops are usually located in areas of medium wealth.

Cluster 7 has high connectivity owing to it having the highest average MRT count of 2+.

Xing Fu Tang can expect moderate competition as Subzones in Cluster 7 having on average 2+ bubble tea outlets.

Cluster 9 is worth discussing. It has the highest average Population of 20–44 year olds and has a relatively low number of bubble tea outlets. However, it doesn't rank highly due to a lack of malls and MRT stations, which indicates it being largely residential with low commercial foot traffic.

Creme de la Creme (Top 5 Promising Subzones)

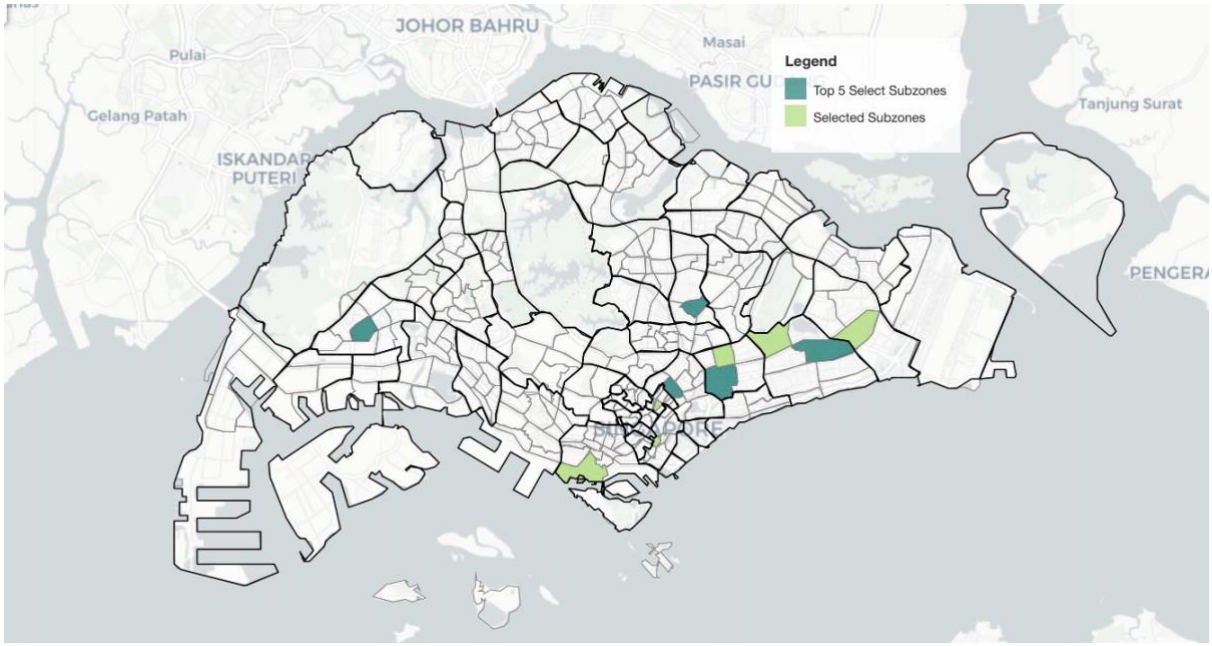
Let's further narrow down Subzones in Cluster 7 to the Top 5 best candidates for Xing Fu Tang's new outlet.

Subzone	Planning Area	geometry	other_boba_c	xft_boba_cou	mrt_count	mall_count	dwell_idx	pop_total20_4	median_inc	cluster	mrt_mall_sco	bubble_score	subzone_score
BEDOK NOR	BEDOK	POLYGON Z	2	0	2	1	4.04231191	26480	4500	7	11	2	9
ALJUNIED	GEYLANG	POLYGON Z	0	0	3	0	4.70793809	13740	3500	7	9	0	9
LAVENDER	KALLANG	POLYGON Z	3	0	2	1	4.38888889	3100	3500	7	11	3	8
SERANGOOI	SERANGOOI	POLYGON Z	3	0	2	1	4.72466694	8430	4500	7	11	3	8
JURONG WE	JURONG WE	POLYGON Z	3	0	2	1	4.64540896	24220	3500	7	11	3	8

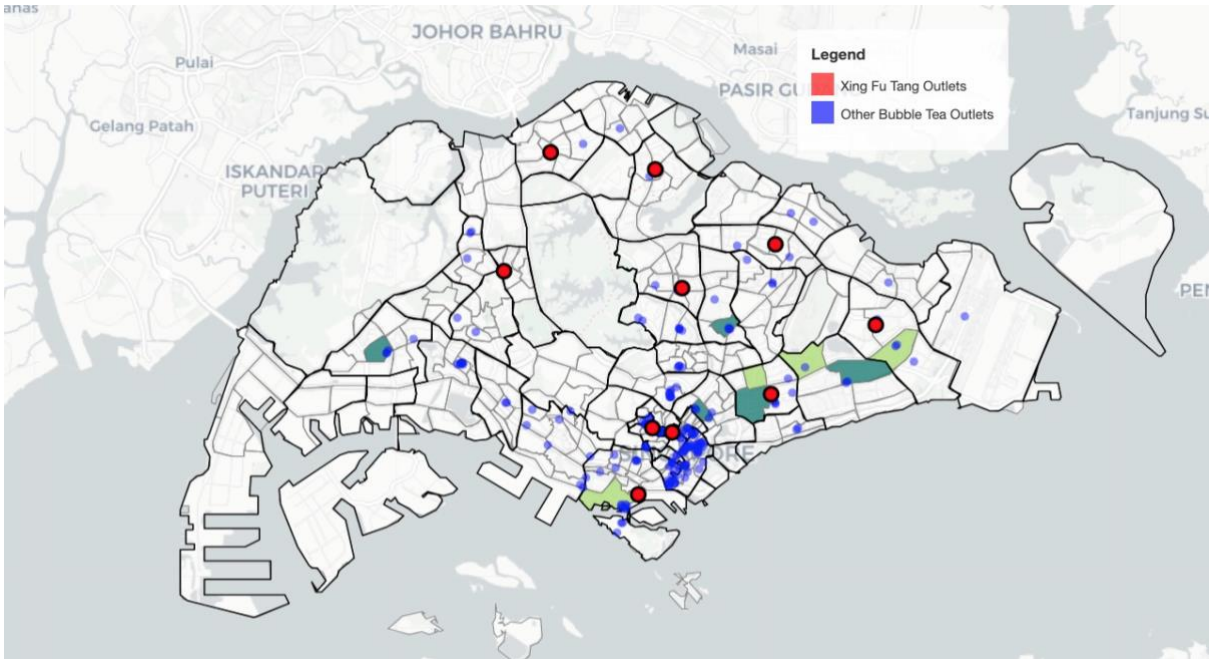
Tied for first place are Bedok North and Aljunied with Subzone scores of 9 each. In second place we have Lavender, Serangoon Central and Jurong West Central each with 8 points.

Plotting Best Subzones

Let's plot these Subzones on a map, along with the rest of Cluster 7.



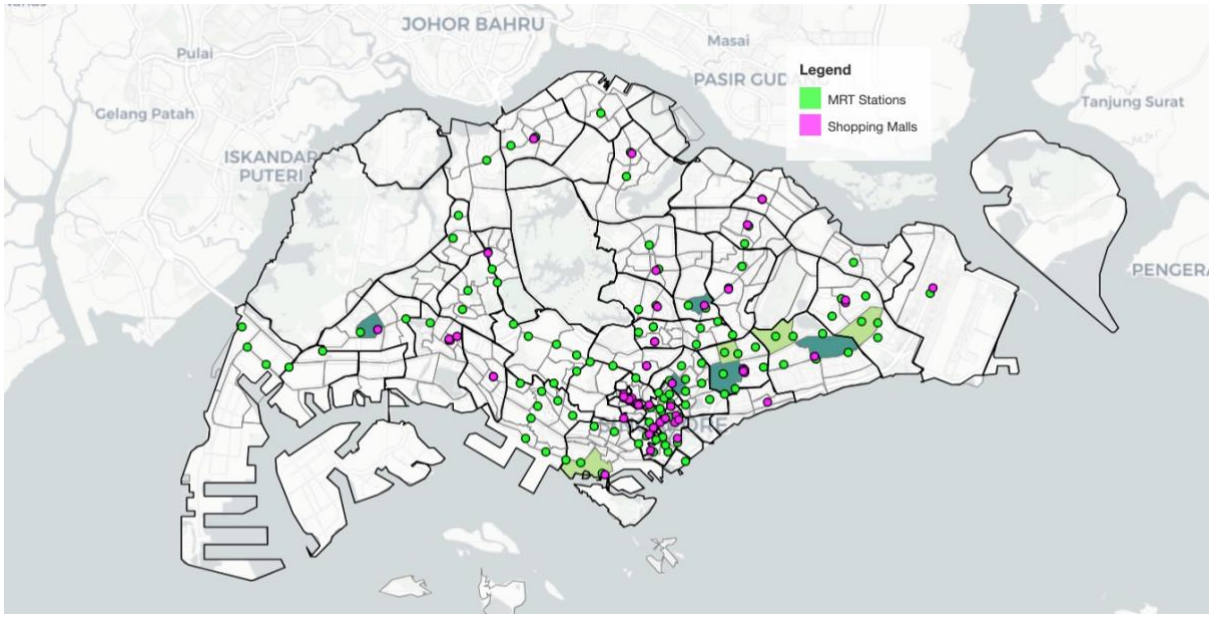
Let's plot the bubble tea shops locations onto the map to compare further.



Except for Aljunied, the Top 5 Subzones are located in midpoints between Xing Fu Tang Outlets, or for the case of Jurong West Central, in a potential region to expand into.

We can also see that Aljunied is selected due to a relative lack of Bubble Tea Outlets in the Subzone.

Let's now plot the MRT/Mall locations onto the map.



Consistent with the notion that bubble tea shops are usually situated close to MRT/Malls, we see most selected Subzones having an abundance of these 2 locations.

This concludes our analysis on the best Subzone to locate a new Xing Fu Tang.

Results and Discussion

In this report, we've clustered Singapore's 332 Subzones into 11 clusters using the K-means method. The clustering is based on 7 features, namely the following:

- Number of existing Xing Fu Tang outlets by Subzones
- Number of bubble tea shops from other franchises by Subzones
- Number of Mass Rapid Transit (MRT) Stations by Subzones
- Number of Shopping Malls by Subzones
- Population of Target Demographic (20–44 years old) by Subzones
- Median Income of Residents by Planning Area
- Aggregation of Dwelling Types (Dwelling Index) by Subzones

Data visualisation and initial exploratory data analysis show a potential relationship between locations of Bubble Tea Shops with MRT Stations and Shopping Malls.

With additional demographic and types of dwelling data, we can infer that Subzones with a high population of 20–44 year olds and a medium amount of wealth (medium income and moderate dwelling types) have a higher number of MRT stations and Malls. We can postulate that these Subzones are more densely populated with medium-cost flats, and see larger commercial foot traffic.

This, in turn, attracts a larger number of Bubble Tea shops.

By scoring each Subzone positively by the number of MRT Stations/Mall it contains, and scoring negatively when Bubble Tea Shops are present, we can determine promising Subzones for Xing Fu Tang's next branch.

Based on the data given, we've found that the 5 most promising Subzones are (in descending order) Bedok North, Aljunied, Lavender, Serangoon Central and Jurong West Central.

This study only serves as an initial recommendation for further insights gathering. As the study only looks at the 7 aforementioned features, the accuracy of the model is limited. Other factors such as real-time commercial traffic data, education levels, occupation data, strongly-correlating venue types (e.g. large overlap of bubble tea — hotpot restaurants consumer trends) can be considered to improve the accuracy of the model.

Additionally, while the Subzone division is the smallest census division for Singapore, other methods to divide Singapore into smaller units of division (e.g. hexagonal grid) may offer insights on a better level of resolution, giving accuracy to the level of metres.

Conclusion

The goal of this report is to identify promising areas to locate a Xing Fu Tang outlet in Singapore. Dividing Singapore into its 332 subzones, we fetched Subzone data for the relevant features and clustered similar Subzones into 11 clusters.

We scored the clusters by the number of MRT stations, malls, and bubble tea outlet it contains. The promising Subzones identified had higher counts of MRT stations and malls and a lower count of bubble tea outlets. The 5 most promising Subzones are (in descending order) Bedok North, Aljunied, Lavender, Serangoon Central and Jurong West Central.

The next steps include further scrutiny of these 5 Subzones for their suitability and additional considerations for other factors that might improve the model's accuracy.

Code

https://github.com/BryanJian/Coursera_Capstone/blob/master/Battle%20of%20the%20Neighbourhoods%20(Singapore).ipynb

References

A similar project in Montreal, Canada for Movie Theaters

Relationship between bubble tea outlets and MRT stations

Obtaining median income by Planning Area

Plotting custom legend onto Folium maps

Data Sources

Planning Area Boundaries

Subzone Boundaries

Location of existing Xing Fu Tang outlets

Location of bubble tea shops from other franchises (Foursquare API)

Location of Mass Rapid Transit (MRT) Stations

Location of Shopping Malls (Foursquare API)

Population of Target Demographic (20–44 years old)

Median Income of Residents

Aggregation of Dwelling Types