Data Science for HDB Flats

1. Introduction

1.1 In Singapore, most families live in flats subsidised by the state. These flats are managed by the Housing Development Board ("HDB") and are colloquially known as "HDB flats". HDB flats (which are usually contained in high-rise blocks) are scatted across the country and are mostly made available to married couples and families. Because of the state subsidy, HDB flats are an attractive option to young couples who are strapped for cash. This is especially so since land in Singapore is scarce and expensive.

1.2 But how should people go about deciding which HDB flat to purchase? The orthodox and boring advice is to look for a location that is close to their working spaces. Flats closer to Singapore's central business district tends to be more expensive, and young couples are thus often advised to strike a balance between distance and cost.

1.3 The orthodox advice, however, has always been a little suspect. Compared to most developed countries, Singapore is very small and has a well-developed public transportation system. Work ravel times are usually not the bane of existence that is the case in many countries. Further, in today's remote working world, it is arguably silly to pick one's residence based on how proximate it is to the office.

1.4 Instead, it is probably more useful to base one's search on how proximate or closely associated a region is to certain amenities (e.g., yoga studies and gyms). Purchasers could simply pay a premium for a HDB flat in central Singapore to guarantee finding a flat that is physically proximate to many amenities, but that is not necessary because Singapore has been designed such that each township would have access to various amenities (link: Shaping the future of Singapore's heartlands - TODAY (todayonline.com)). In other words, a less pricey HDB flat in the suburbs could be equally proximate to a centrally located HDB flat.

1.5 However, because there is no readily navigable data that sets out the amenities surrounding each HDB flats, the search for the perfect flat can be a real pain, especially since there are over a million HDB flat in Singapore. This project seeks to ameliorate the problem by clustering HDB flats around the country based on their characteristics, thereby allowing a couple to select flats based on the popular venues associated with each flat.

2. Data

2.1 Sources

2.1.1 Two sources of data will be used for this project.

(A) The first source of data is from Kaggle (link: Singapore HDB Postal Code Mapper (2018) | Kaggle). Prepared by user MyleeSG, the data in the CSV file sets out the coordinates and postal codes for all HDB blocks as of 2019.

(B) The second source of data is from Foursquare. I will be using the Foursquare API to extract the top venues associated with each HDB flat.

2.2 Data Cleaning

2.2.1 The Kaggle dataset (which I named "hdb") was massive and had 25,293 rows

(according to `hdb.shape()`), each of which represented a block of HDB flats. There was also a lot of extraneous data, such as the full address of each block.

2.2.2 Calling `hdb.head()` gave the following result:

	postal	latitude	longtitude	searchval	blk_no	road_name	building	address	postal.1
0	398614	1.312763	103.883519	# 1 LOFT	1	LORONG 24 GEYLANG	# 1 LOFT	1 LORONG 24 GEYLANG # 1 LOFT SINGAPORE 398614	398614
1	398721	1.312390	103.881504	# 1 SUITES	1	LORONG 20 GEYLANG	# 1 SUITES	1 LORONG 20 GEYLANG # 1 SUITES SINGAPORE 398721	398721
2	629875	1.309135	103.679463	1 BENOI ROAD SINGAPORE 629875	1	BENOI ROAD	NIL	1 BENOI ROAD SINGAPORE 629875	629875
3	439731	1.305466	103.895674	1 BOSCOMBE ROAD SINGAPORE 439731	1	BOSCOMBE ROAD	NIL	1 BOSCOMBE ROAD SINGAPORE 439731	439731
4	659592	1.344619	103.749789	1 BUKIT BATOK STREET 22 SINGAPORE 659592	1	BUKIT BATOK STREET 22	NIL	1 BUKIT BATOK STREET 22 SINGAPORE 659592	659592

2.2.3 Having 25,293 rows was great because it assured me that the dataset was sufficiently robust (every block comprises multiple HDB flats, which means the dataset approximated over 1m HDB flats). However, Foursquare only allows 950 Regular API Calls per day for Sandbox Tier Accounts (and I cannot afford to upgrade my account). It was therefore necessary to reduce the number of datapoints, which is unfortunate for accuracy but necessary for my wallet.

2.2.4 There were a few ways that I could have cut down on the data points. For example, I could have simply removed every 100th row of the dataframe. But that seemed too arbitrary and would have comprised the results. I needed a more objective way to reduce the amount of data, and it occurred to me that using postal codes might be useful.

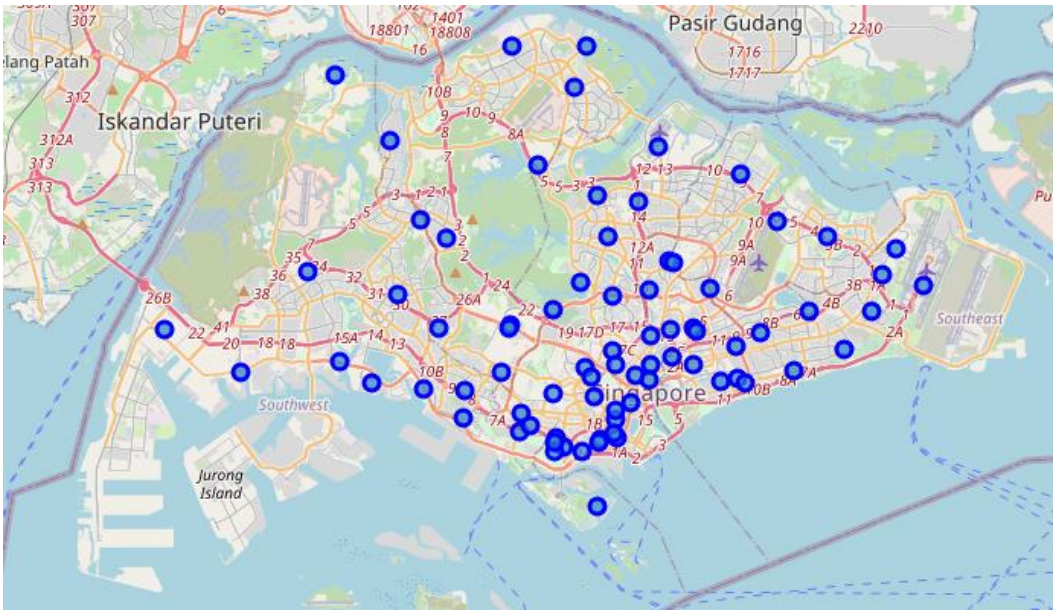
2.2.5 A search on Wikipedia revealed that Singapore's postal codes are based on postal sectors (i.e., as denoted by the first two digits of each postal code), which are in turn based on the geographical territory that the HDB block is located in (link: Postal codes in Singapore - Wikipedia). With this insight, I created a dataframe that removed some columns (e.g. "`searchval`") and added a new column setting out the postal sector for each block. I then removed duplicates under the new column, leaving only unique postal sectors.

2.2.6 After cleaning up the data, `hdb` looked like this:

	postal	latitude	longtitude	road_name	address	postal sector
0	398614	1.312763	103.883519	LORONG 24 GEYLANG	1 LORONG 24 GEYLANG # 1 LOFT SINGAPORE 398614	39
2	629875	1.309135	103.679463	BENOI ROAD	1 BENOI ROAD SINGAPORE 629875	62
3	439731	1.305466	103.895674	BOSCOMBE ROAD	1 BOSCOMBE ROAD SINGAPORE 439731	43
4	659592	1.344619	103.749789	BUKIT BATOK STREET 22	1 BUKIT BATOK STREET 22 SINGAPORE 659592	65
5	618292	1.314283	103.723913	BUROH LANE	1 BUROH LANE SINGAPORE 618292	61

3. Methodology

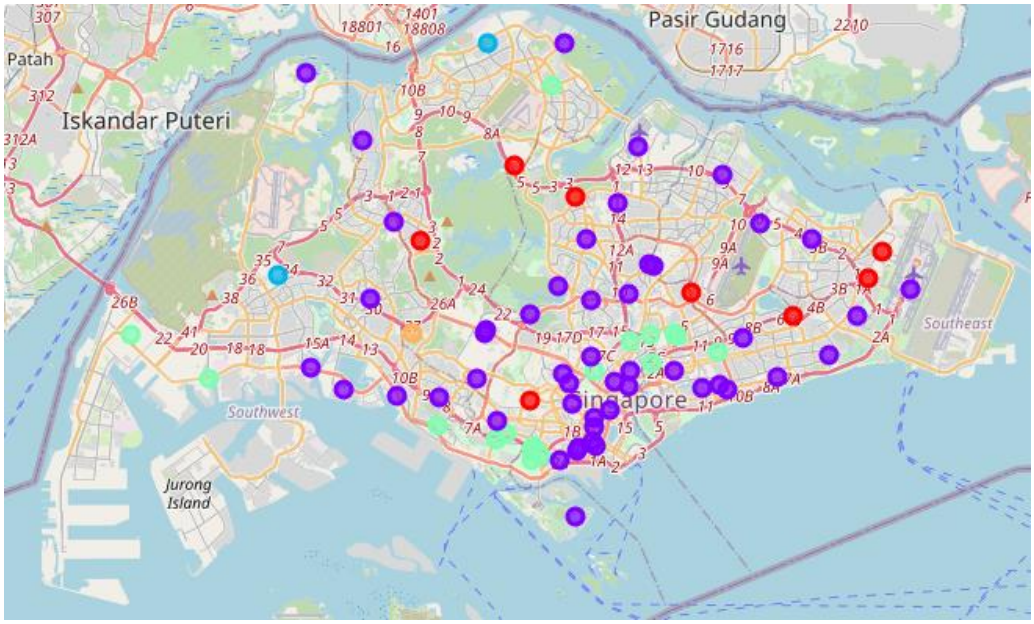
3.1 The first step was to map out the postal sectors using Folium.



3.2 The map showed a good distribution of postal sectors, as expected. As part of the exploratory process, I then used the Foursquare API to figure out what the top venues were for each postal sector. The results for the first few hits were as follows:

Road Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
AIRPORT BOULEVARD	Road	Airport	Office	Coffee Shop	Food
AMOY STREET	Korean Restaurant	Japanese Restaurant	Food Court	Coffee Shop	Italian Restaurant
ANG MO KIO AVENUE 6	Food Court	Fast Food Restaurant	Dessert Shop	Bubble Tea Shop	Coffee Shop
BEACH ROAD	Hotel	Café	Japanese Restaurant	Shopping Mall	Dessert Shop
BEDOK RESERVOIR VIEW	Bus Station	Indian Restaurant	Steakhouse	Playground	Pizza Place

3.3 Using K-means clustering, I clustered the HDB blocks into 5 clusters based on the top venues associated with each block. Mapped, the result was as follows.

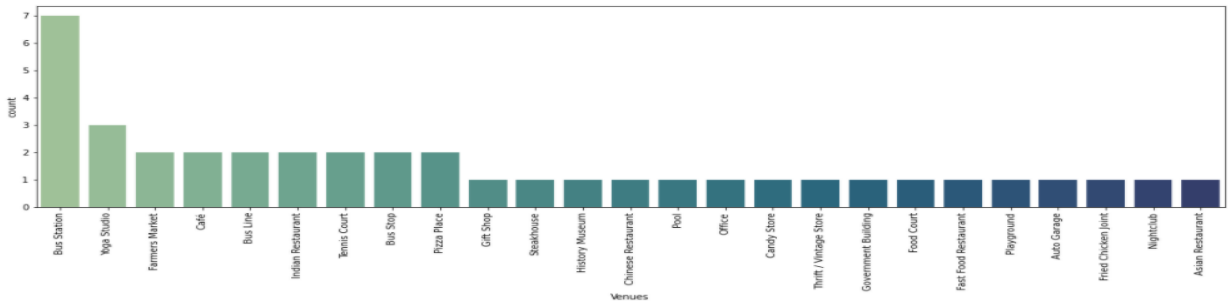


4. Results and Discussion

4.1 Cluster 1

4.1.1 Cluster 1 (depicted as red dots in the map above) is a fairly small cluster.

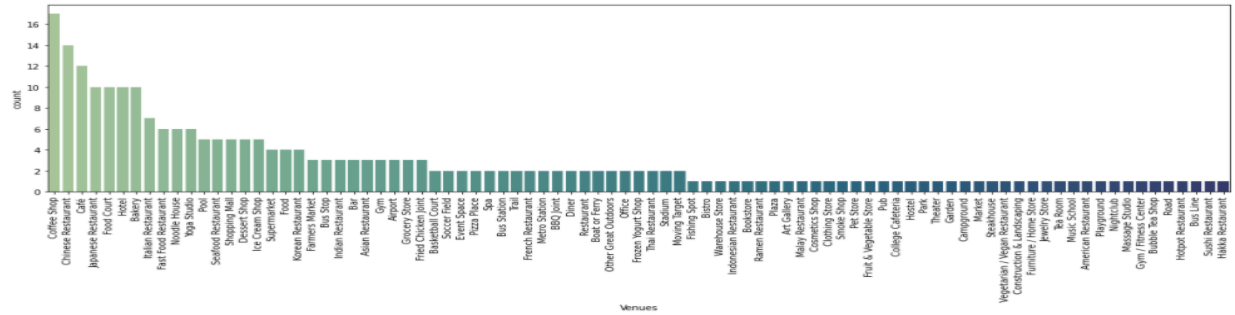
4.1.2 People who are interested in living in areas most closely associated with bus stations, yoga studios and farmers market should consider purchasing a HDB flat in this cluster.



4.2 Cluster 2

4.2.1 Cluster 2 (depicted as purple dots in the map above) is significantly larger.

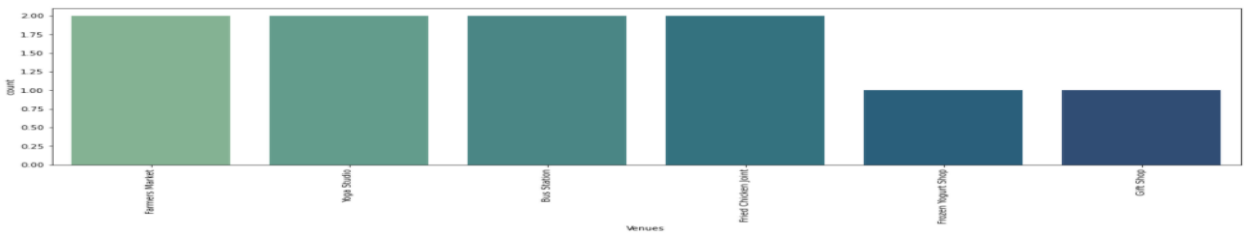
4.2.2 People who want to live in sectors associated with food options should consider HDB flats from this cluster. This cluster represents a great option for foodies.



4.3 Cluster 3

4.3.1 Cluster 3 (depicted as orange dots in the map above) is very small, comprising only two HDB blocks.

4.3.2 This cluster is similar to Cluster 1, insofar as it should also be attractive to people who are drawn to farmers markets, yoga studios and bus stations.

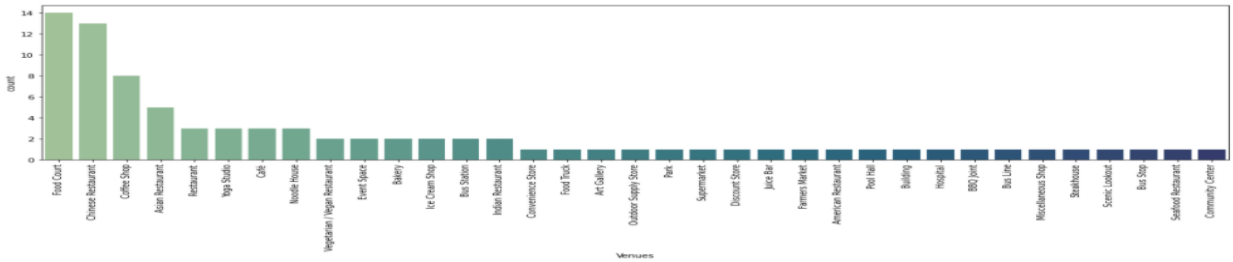


4.4 Cluster 4

4.4.1 Cluster 4 (depicted as neon green dots in the map above) is about the same size

as Cluster 3.

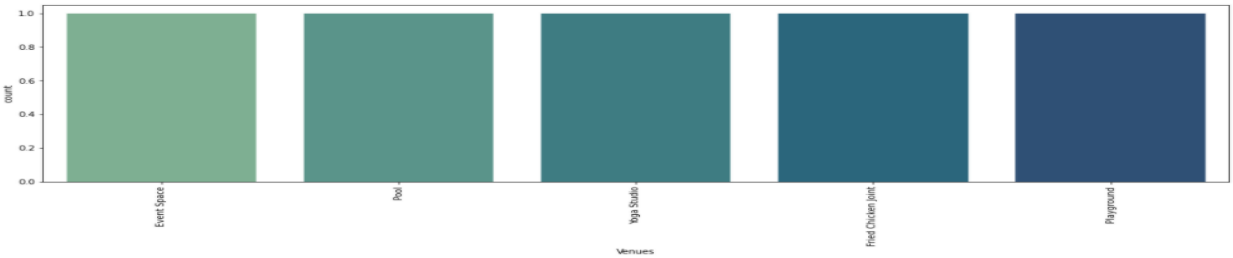
4.4.2 Like Cluster 2, Cluster 4 is most appropriate for people who are foodies. The difference is that Cluster 4 is more ideal for foodies that like Chinese food and variety in the form of food courts.



4.5 **Cluster 5**

4.5.1 Cluster 5 (depicted as an orange dot in the map above) has only 1 HDB block.

4.5.2 Cluster 5 is good for purchasers who wish to live in places most commonly associated with pools, playgrounds and event spaces.



5. Conclusion

5.1 In conclusion, this project has demonstrated that it is possible to cluster HDB flats according to their top venues. This would be useful for purchasers looking for HDB flats based on parameters that are distinct from the question of how far the flat is from Singapore's central business district.