IBM – Coursera

Data Science Professional

Coursera Capstone project - Final report

Correlation between Singapore Housing Prices and its Popular Venues

Abhilash Divakaran

2019

Contents

```
----ANG MO KIO----
                 venue  freq
0           Food Court  0.16
1          Coffee Shop  0.10
2         Dessert Shop  0.04
3   Chinese Restaurant  0.04
4  Japanese Restaurant  0.03


----BEDOK----
                venue  freq
0         Coffee Shop  0.10
1  Chinese Restaurant  0.07
2          Food Court  0.06
3    Asian Restaurant  0.05
4                Café  0.05


----BISHAN----
                 venue  freq
0           Food Court  0.09
1          Coffee Shop  0.09
2   Chinese Restaurant  0.08
3      Thai Restaurant  0.06
4  Japanese Restaurant  0.05


----BUKIT BATOK----
                  venue  freq
0            Food Court  0.18
1           Coffee Shop  0.11
2    Chinese Restaurant  0.09
3  Fast Food Restaurant  0.07
4         Grocery Store  0.07


----BUKIT PANJANG----
                  venue  freq
0           Coffee Shop  0.08
1           Bus Station  0.08
2      Asian Restaurant  0.08
3  Fast Food Restaurant  0.06
4           Supermarket  0.04
```

1. INTRODUCTION

Singapore is a city-state 'small island country' in the South East Asia. Singapore is a modern affluent city combining of skyscrapers, industrial estates and broad transport routes. The population of Singapore sizes to more than five millions and makes the state a densely populated country.

Figure 1 - HDB flats - a common view in Singapore (Wikimedia Commons image)

Unlike the other densely populated countries, 50% the Singapore land is covered by greenery. With more than 50 nature parks and 4 reservoirs, it is a beautiful garden city.

2. TARGETED AUDIENCE

Prospective HDB flat buyers in Singapore are the primary targeted audience of this project. Singapore well connected to its nook and corners by roads and public transport system. As such, it will be highly unlikely that a buyer will consider the distance to workplace as a major factor for buying a house. This project will provide a meaningful insight of the other convenience factors that influence the pricing of HDB flats. This analysis will be also interesting for the people who aspiring to relocate to a different town as they gain insight of the most common venues of their interests.

The main goal will be exploring the neighborhoods of Singapore in order to extract the correlation between the real estate value and its surrounding venues.

The idea comes from the process of a normal family finding a place to stay after moving to another city. It's common that the owners or agents advertise their properties are closed to some kinds of venues like supermarkets, restaurants or coffee shops, etc.; showing the "convenience" of the location in order to raise their house's value.

So, can the surrounding venues affect the price of a house? If so, what types of venues have the most affect, both positively and negatively?

The target audience for this report are:

- Potential buyers who can roughly estimate the value of a house based on the surrounding venues and the average price.
- Real estate makers and planners who can decide what kind of venues to put around their products to maximize selling price.
- Houses sellers who can optimize their advertisements.

3. DATA DESCRIPTION

3.1 Background

Singapore has many clean, green and modern self-contained residential towns mushroomed all over the island.

Housing Development Board (HDB) is the major provider of the housing infrastructure in Singapore. A number of private residential developers too operates in every neighborhood. HDB provides subsidized residential flats for the citizen and the Permanent Residents of Singapore.

Singapore neighborhoods chosen as the observation target due to the following reasons:

- The availability of resale flat prices from the HDB.
- The diversity of prices between neighborhoods. For example, 3-bedrooms HDB flat in Central Area can cost S$2 million on average; while in woodlands it costs only S$400 thousands.
- The availability of geo data, which can used to visualize the dataset onto a map.

3.2 Data Source

The dataset will composed from the following two main sources:

- **HDB Resale Flat Prices data from ‘data.gov.sg’**: This dataset compiled by the Housing Development Board of Singapore and available for public use in data.gov.sg. This data comprises the price data from year 2017 January until 2019 August.
  We will make use of the year 2018 and 2019 data.
- **Foursquare API -** which provides the nearby venues of a given location, coordinates.
- **Google Map API** – for the location coordinates of the Singapore MRT stations

4. METHODOLOGY

The methodology section discuss about the following

- Data preparation and Cleaning
- Data visualization and
- Data Analysis

4.1 Data Preparation and Cleaning:

Our first data set is the resale flat prices from the Housing Development Board (HDB) of Singapore. The data, can accessed or downloaded from the link https://data.gov.sg/api/action/datastore_search?resource_id=42ff9cfe-abe5-4b54-beda-c88f9bb438ee/download.
However, for this analysis I have already downloaded the data set and placed in GitHub at https://raw.githubusercontent.com/abhi-divakaran/Coursera_Capstone/master/FINAL/resale-flat-prices-jan-2017-onwards.csv.

The data set contains resale price data from year 2017 until year 2019.

	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
0	2017-01	ANG MO KIO	2 ROOM	406	ANG MO KIO AVE 10	10 TO 12	44.0	Improved	1979	61 years 04 months	232000.0
1	2017-01	ANG MO KIO	3 ROOM	108	ANG MO KIO AVE 4	01 TO 03	67.0	New Generation	1978	60 years 07 months	250000.0
2	2017-01	ANG MO KIO	3 ROOM	602	ANG MO KIO AVE 5	01 TO 03	67.0	New Generation	1980	62 years 05 months	262000.0
3	2017-01	ANG MO KIO	3 ROOM	465	ANG MO KIO AVE 10	04 TO 06	68.0	New Generation	1980	62 years 01 month	265000.0
4	2017-01	ANG MO KIO	3 ROOM	601	ANG MO KIO AVE 5	01 TO 03	67.0	New Generation	1980	62 years 05 months	265000.0

Figure 2 - Raw data as seen in a Pandas data frame

This dataset has 56,335 samples with 11 features. Target label in this dataset is the actual resale price of the flat, which is available as the feature ‘resale_price’.

A careful study has conducted on each of the features to define the weightage of each feature on our target label.

From the feature ‘month’, omitted all the data before 01 January 2018, since this study requires data, only from years 2018 and 2019.

The feature ‘town’ is the base label for which target is plotted

From the feature ‘flat_type’, the most popular types flats were identified and they were 4 Room, 5 Room and 3 Room in descending order. Analysis on most popular flat types provides a more realistic result.

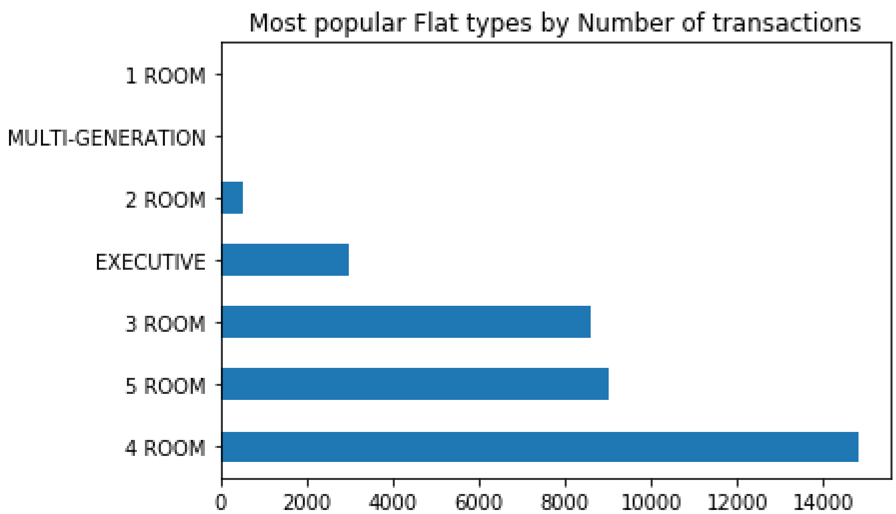

Figure 3 - Most Popular Flat Types

Features ‘block’ and ‘street_name’ are the location address of the entity, hence no corrections required.

Feature 6, 'storey_range' is a classification based on what level of the building is the flat located. This class has a minor effect in the flat pricing. For example, a flat in 9th floor might cost an additional 10 thousand than a flat in the 3rd floor. However, as we average out the prices by group and town, this becomes ignorable.

The next feature ‘floor_area_sqm’ is a numerical conversion of the feature ‘flat_type’. Since we have already tuned ‘flat_type’ no further actions is unnecessary.

Feature ‘flat_model’, definitely has some impact on the result. Certain models are too less in numbers, however costs as much as 3 times higher than another model with same flat type or floor area. As such, most common models has selected based on the number of transactions on each model.

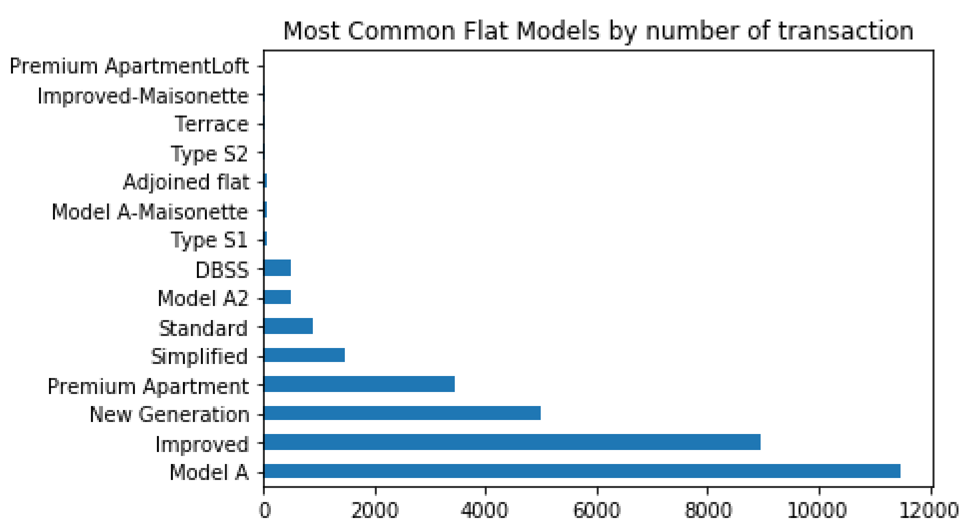

Figure 4 - Most Common Flat Models

After careful analysis it was evident that by number of transactions Model A, Improved, New Generation and Premium Apartment are the prominent models. This feature verifications were done with boxplot as well to confirm the result realistic. The box plot below visualizes the price distribution of the flats by flat-model.

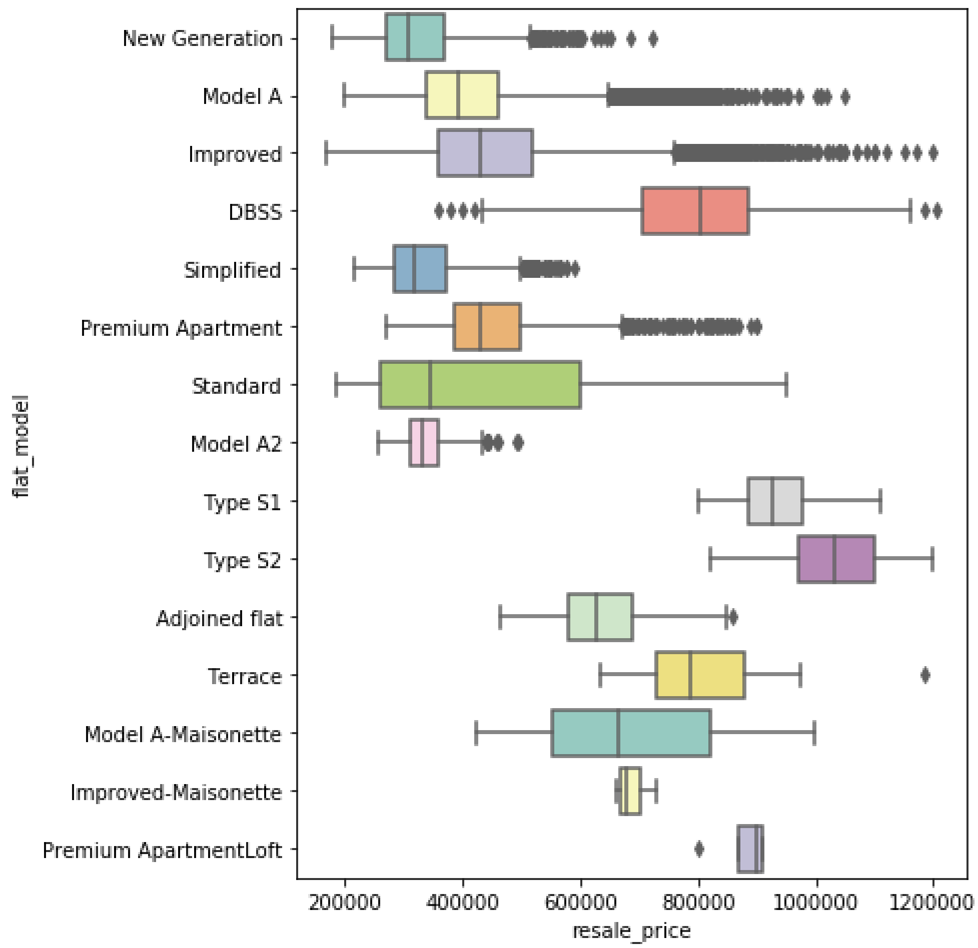

Figure 5 - price distribution by flat model

It was found that, the 7 models in bottom of the chart and Model DBSS belongs to expensive category and they are not popular. Model A2, Standard and Simplified are considerable but not necessary to do so. Considering these models in our modelling may lead to data inaccuracy. There for all these models were dropped out.

The next two features are pertaining to a single reference. In Singapore when an HDB flat is bought, the transaction is not really a sale - it's a lease of 99 years. That means the buyer owns the property for 99 years. The government has set the life of a flat as 99 years.

Once the age of the flat reaches 35 years, the maximum loan amount a bank can provide falls to 65% from the original 85%. This turns away most of the buyers. Hence there is price drop for old flats. Certain towns in Singapore are considered as old towns and many flats there might be older ones.
Considering this. segregate the flats older than 35 years out and choose the remaining for our study.

For the ease of calculation, a new feature has added ‘Age’, which indicates the current age of the flat. The records then filtered by feature ‘age’, where flats with age over 35 years were dropped from the data frame.

	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price	Age
9	2018-01	ANG MO KIO	3 ROOM	466	ANG MO KIO AVE 10	04 TO 06	67.0	New Generation	1984	65 years 08 months	255000.0	35
16	2018-01	ANG MO KIO	3 ROOM	473	ANG MO KIO AVE 10	07 TO 09	67.0	New Generation	1984	65 years 06 months	328000.0	35

Figure 6 - Dataframe after dropping records for old flats

The integrity of this new data frame has then compared with the original dataset by plotting the resale price grouped to flat models as “Before and After”.

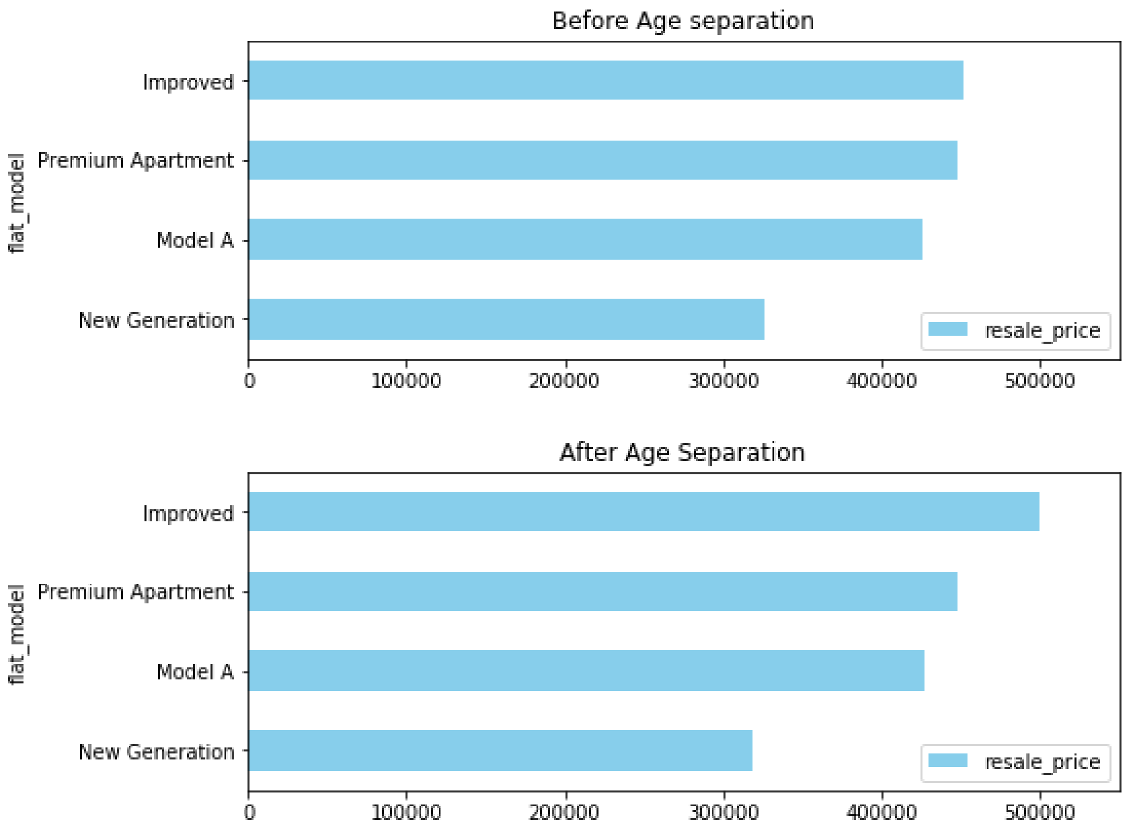

Figure 7 - flat model price comparison with and without older blocks.

An integrity check has then performed on the feature ‘town’. This was to verify that data from how many towns are still available.
The out put is 25 towns instead of 26.

```
Name of the towns:
 ['ANG MO KIO' 'BEDOK' 'BISHAN' 'BUKIT BATOK' 'BUKIT MERAH' 'BUKIT PANJANG'
 'BUKIT TIMAH' 'CENTRAL AREA' 'CHOA CHU KANG' 'CLEMENTI' 'GEYLANG'
 'HOUGANG' 'JURONG EAST' 'JURONG WEST' 'KALLANG/WHAMPOA' 'PASIR RIS'
 'PUNGGOL' 'QUEENSTOWN' 'SEMBAWANG' 'SENGKANG' 'SERANGOON' 'TAMPINES'
 'TOA PAYOH' 'WOODLANDS' 'YISHUN']
Number of towns in the dataset:  25
```

Figure 8 - result window with the names of 25 Towns in the dataset

Singapore has in fact 26 towns. Our original dataset as well have data from 26 towns.
But Marine Parade Town is not really a residential area with any subsidized housing.
The HDB flats there are mostly very old and categorized as conservative.
It's a prime area for private housing projects.
Therefore, records from Marine Parade will be naturally dropped by our earlier data cleaning.

Feature optimization was completed here and proceeded to extract the label and the target out of the data. Town names and the flat price were brought into a new data frame

	town	resale_price
9	ANG MO KIO	255000.0
16	ANG MO KIO	328000.0
34	ANG MO KIO	340000.0
35	ANG MO KIO	635000.0
42	ANG MO KIO	448000.0

Figure 9 - New data frame with town names and resale price from each transaction

This data frame was then optimized using 'groupby' feature. Together with grouping the resale price was averaged into median of the resale prices for that town.
The price distribution of the selected towns was plotted in order to get a detailed look.

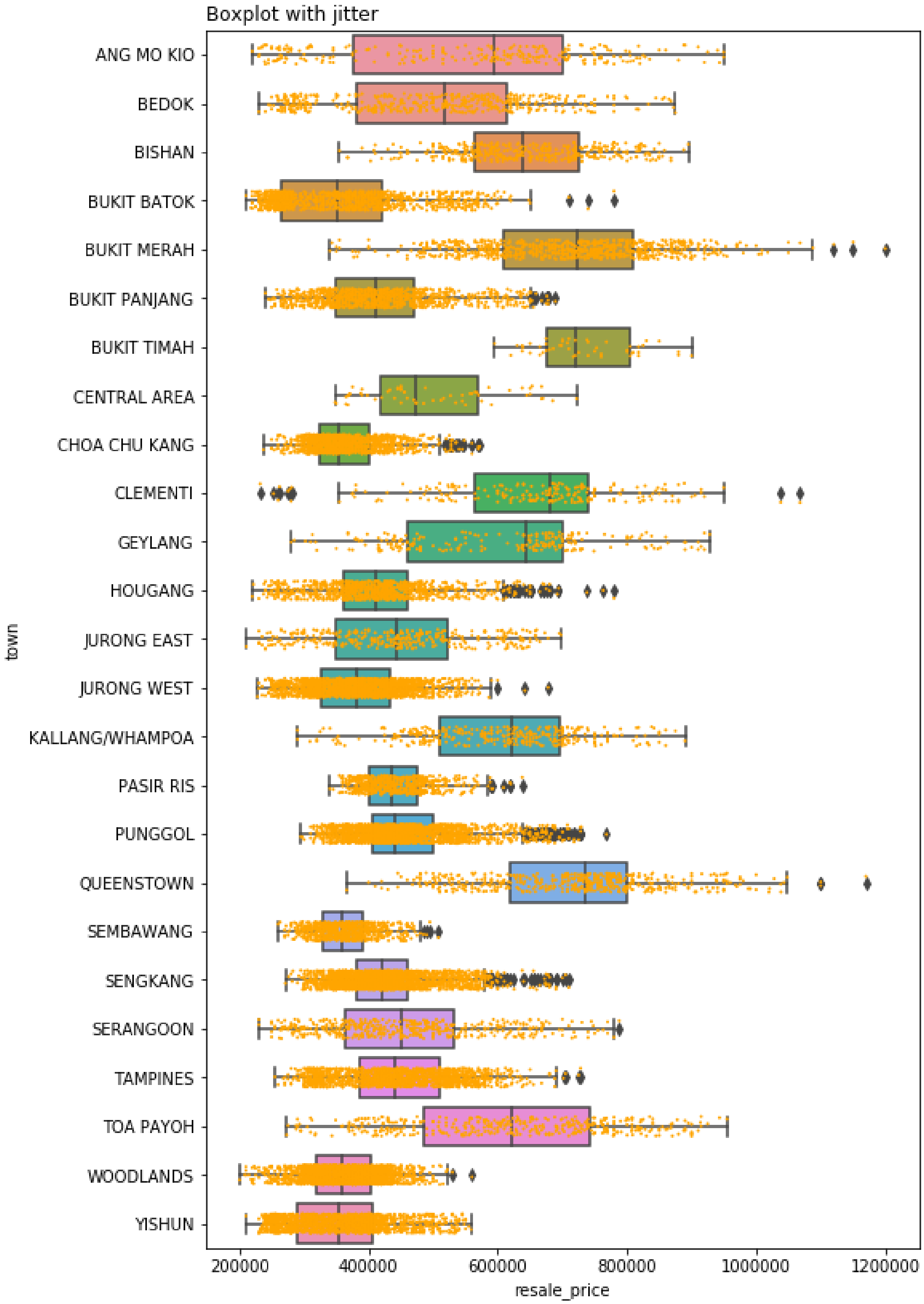

Figure 10 - Price distribution of flats by town

Pandas median 'groupby' and 'median' features were used to find the median of the selling price by town.

	Town	Resale_Price
0	BUKIT BATOK	352000
1	YISHUN	355000
2	CHOA CHU KANG	355000
3	SEMBAWANG	360000
4	WOODLANDS	360000

Figure 11 - Median resale flat price grouped to town

A price comparison was then carried out.

4.2 Data Visualization

In order to find the insight of Singapore real estate average price between neighborhoods, the best way is visualization. In order to plot the town into a geographical map we need to append the coordinates to each town.

For this project We consider the metro railway station (MRT), of each town as their center reference. The latitude and longitude for each town extracted using Google API and saved as mrt_locdada.csv in Github.

This data then merged with the HDB flat resale data excerpt.

	Town	Resale_Price	Latitude	Longitude
0	ANG MO KIO	595000	1.370025	103.849588
1	BEDOK	516500	1.324043	103.930205
2	BISHAN	638000	1.350920	103.848206
3	BUKIT BATOK	352000	1.349069	103.749596
4	BUKIT PANJANG	410000	1.378340	103.762452

Figure 12 - Towns with location coordinates

This dataset was then plotted on a folium map with towns as markers.

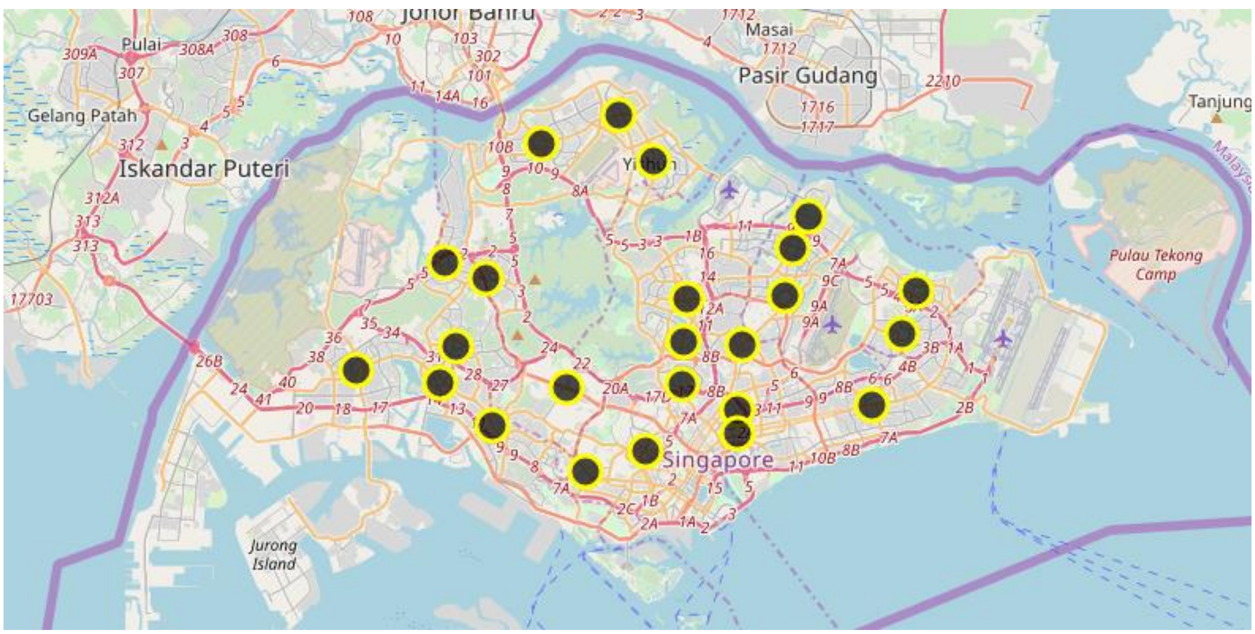

Figure 13 - Map of Singapore with Towns as markers

4.3 Exploratory Analysis

In this section describes the exploratory analysis on each towns. Foursquare API used to gather information on the most popular venues of each town.

Using Foursquare API function "getNearbyVenues" extracted the following information for the data frame: Venue ID, Venue Name, Coordinates: Latitude and Longitude and Category Name.

Using Foursquare API function "getVenuesByCategory" performed category based venue search to simulate user venue searches based on certain places of interest.

This search extracts the following information: Venue ID, Venue Name, Coordinates: Latitude and Longitude and Category Name

For each retrieved venueID, retrieved the venues category rating as well. The result is a "json" structure.


```
Out[36]: {'meta': {'code': 200, 'requestId': '5dcc24f3618f43001b08e873'},
          'response': {'suggestedFilters': {'header': 'Tap to show:',
            'filters': [{'name': 'Open now', 'key': 'openNow'}]},
           'headerLocation': 'Ang Mo Kio',
           'headerFullLocation': 'Ang Mo Kio, Singapore',
           'headerLocationGranularity': 'neighborhood',
           'totalResults': 91,
           'suggestedBounds': {'ne': {'lat': 1.3790250090000091,
             'lng': 103.85857377367182},
            'sw': {'lat': 1.361024990999991, 'lng': 103.84060222632817}},
           'groups': [{'type': 'Recommended Places',
             'name': 'recommended',
             'items': [{'reasons': {'count': 0,
                'items': [{'summary': 'This spot is popular',
                  'type': 'general',
                  'reasonName': 'globalInteractionReason'}]},
               'venue': {'id': '4bc30123abf495215220c393',
                'name': 'Old Chang Kee',
                'location': {'address': 'AMK Hub',
                 'crossStreet': '53 Ang Mo Kio Avenue 3',
                 'lat': 1.369093882325005,
                 'lng': 103.84838865753792,
                 'labeledLatLngs': [{'label': 'display',
                   'lat': 1.369093882325005,
                   'lng': 103.84838865753792}],
```


Figure 14 - Foursquare API query result as json

The json result then cleaned and loaded as a pandas dataframe.

	name	categories	lat	lng
0	Old Chang Kee	Snack Place	1.369094	103.848389
1	FairPrice Xtra	Supermarket	1.369279	103.848886
2	Subway	Sandwich Place	1.369136	103.847612
3	PLAYe	Hobby Shop	1.369109	103.848225
4	Face Ban Mian 非板面 (Ang Mo Kio)	Noodle House	1.372031	103.847504

Figure 15 - Foursquare result json converted to dataframe

Exploratory analysis has carried out for all Singapore towns and grouped by neighborhood, taking the mean of the frequency of occurrence of each category. Towns with top five most common venues were listed.

```
----ANG MO KIO----
                 venue  freq
0           Food Court  0.16
1          Coffee Shop  0.10
2         Dessert Shop  0.04
3   Chinese Restaurant  0.04
4  Japanese Restaurant  0.03


----BEDOK----
                venue  freq
0         Coffee Shop  0.10
1  Chinese Restaurant  0.07
2          Food Court  0.06
3     Asian Restaurant  0.05
4                Café  0.05


----BISHAN----
                 venue  freq
0           Food Court  0.09
1          Coffee Shop  0.09
2   Chinese Restaurant  0.08
3      Thai Restaurant  0.06
4  Japanese Restaurant  0.05


----BUKIT BATOK----
                  venue  freq
0            Food Court  0.18
1           Coffee Shop  0.11
2    Chinese Restaurant  0.09
3  Fast Food Restaurant  0.07
4         Grocery Store  0.07


----BUKIT PANJANG----
                  venue  freq
0           Coffee Shop  0.08
1           Bus Station  0.08
2      Asian Restaurant  0.08
3  Fast Food Restaurant  0.06
4           Supermarket  0.04
```

Figure 16 - Towns with top five most common venues

This data was put into a dataframe and venues sorted in descending order. A new dtaframe created which displays the top ten venues for each town.

	Town	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	ANG MO KIO	Food Court	Coffee Shop	Dessert Shop	Chinese Restaurant	Fast Food Restaurant	Japanese Restaurant	Grocery Store	Basketball Court	Snack Place	Bus Line
1	BEDOK	Coffee Shop	Chinese Restaurant	Food Court	Café	Asian Restaurant	Supermarket	Sandwich Place	Bakery	Fast Food Restaurant	Japanese Restaurant
2	BISHAN	Food Court	Coffee Shop	Chinese Restaurant	Thai Restaurant	Japanese Restaurant	Seafood Restaurant	Café	Bubble Tea Shop	Stadium	Cosmetics Shop
3	BUKIT BATOK	Food Court	Coffee Shop	Chinese Restaurant	Grocery Store	Fast Food Restaurant	Bus Station	Malay Restaurant	College Cafeteria	Sandwich Place	Convenience Store
4	BUKIT PANJANG	Bus Station	Coffee Shop	Asian Restaurant	Fast Food Restaurant	Gym	Noodle House	Food Court	Shopping Mall	Light Rail Station	Supermarket

Figure 17 - Top ten venues for each town

4.4 Clustering by machine learning algorithm k-means

Using k-means identified the similarities between towns based on the most common venues. Using clustering identical towns then plotted into clusters.

This data ported as a new pandas data frame.

	Town	Resale_Price	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	
0	ANG MO KIO	595000	1.370025	103.849588	1	Food Court	Coffee Shop	Dessert Shop	Chinese Restaurant	Fast Food Restaurant	Japanese Restaurant	G S
1	BEDOK	516500	1.324043	103.930205	0	Coffee Shop	Chinese Restaurant	Food Court	Café	Asian Restaurant	Supermarket	S P
2	BISHAN	638000	1.350920	103.848206	0	Food Court	Coffee Shop	Chinese Restaurant	Thai Restaurant	Japanese Restaurant	Seafood Restaurant	C
3	BUKIT BATOK	352000	1.349069	103.749596	1	Food Court	Coffee Shop	Chinese Restaurant	Grocery Store	Fast Food Restaurant	Bus Station	M R
4	BUKIT PANJANG	410000	1.378340	103.762452	0	Bus Station	Coffee Shop	Asian Restaurant	Fast Food Restaurant	Gym	Noodle House	F C

Figure 18 - New dataframe with clustering

Using Folium the clusters then visualized, which uses differences in shading or coloring to indicate a property's values or quantity within predefined areas.

5. Results:

The price modelling was a very fine work and the result turned out very realistic. The most economic towns in Singapore with reference to the resale price of housing was identified as Bukit Batok, Yishun, Choa Chu Kang, Sembawang and Woodlands.

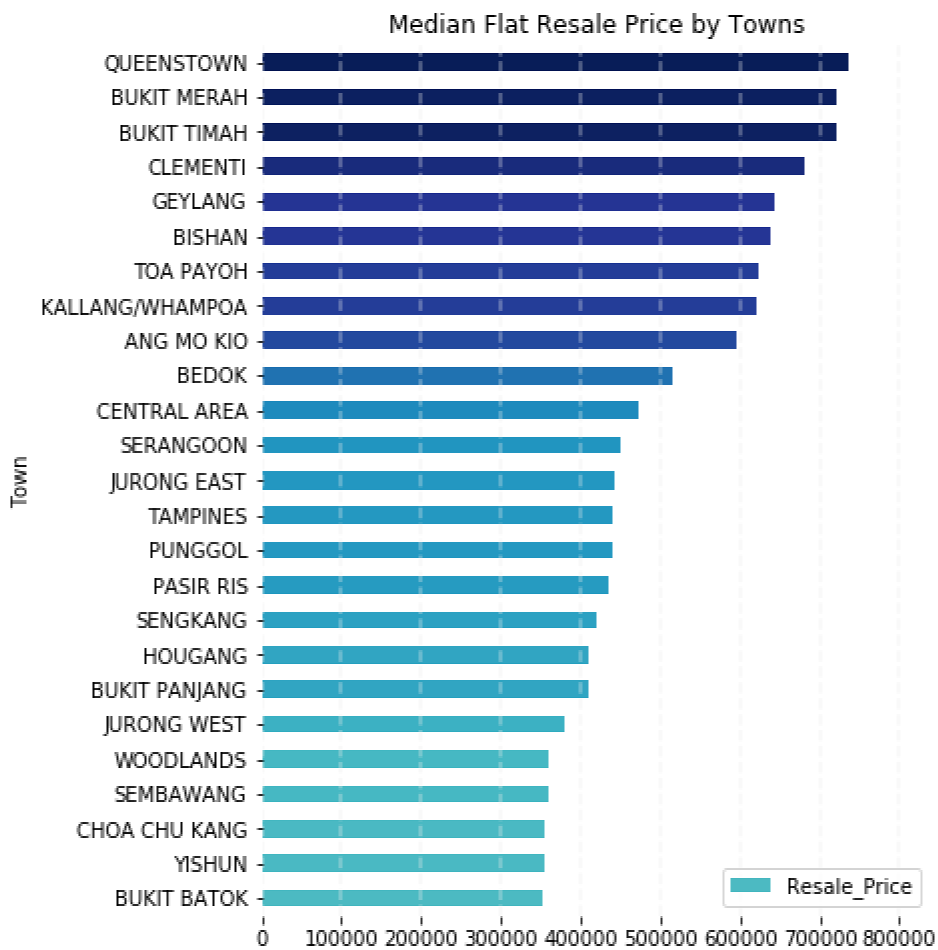

Figure 19 - Resale Pricing model of HDB flats

This is indeed a realistic model. If a direct query on the raw data was conducted for the lowest pricing, the town Ang Mo Kio, Bishan, Pasir Ris and Punggol would be in the top list.

Clustering visualization results are insightful in identifying the identical towns based on most popular venues.

Towns binned into five clusters based on their 'most popular venues' similarity.

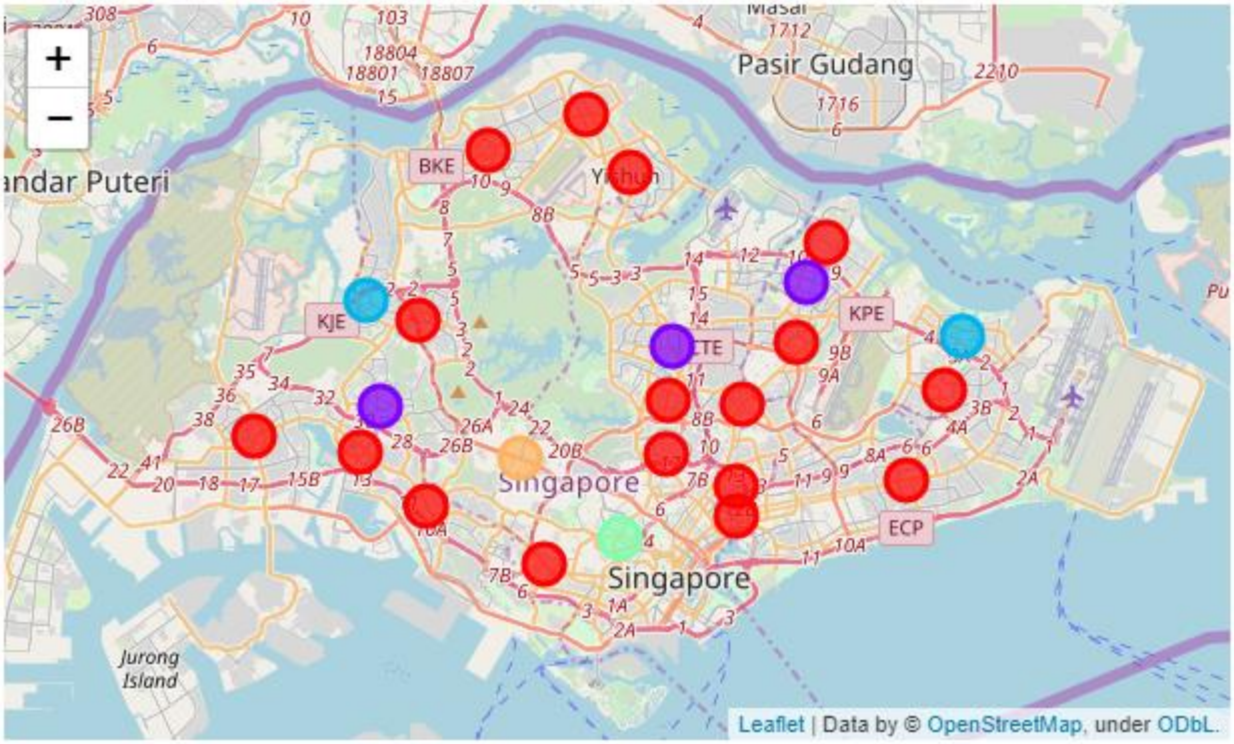

Figure 20 - Clustering of Singapore Towns

The result on clustering is dynamic with reference to the time and date of Foursquare API querying.

6. Discussion:

The real challenge is constructing the dataset:

- Usually the needed data isn't publicly available.
- When combining data from multiple sources, inconsistent can happen. And lots of efforts are required to check, research and change the data before merge.
- For data obtained through API calls, different results are returned with different set of parameters and different point of time. Multiple trial and error runs are required to get the optimal result.
- Even after the dataset has been constructed, lots of research and analysis are required to decide if the data should be kept as is or be transform by normalization or standardization.

It can be considered the most important process in the whole data science pipeline. Which can affect the most on the result.

On the other hand, choosing the suitable technique to construct the model is also a worthwhile process. As this report shows that, by applying a different method, the result can be improved.

7. Conclusion:

It’s unfortunately that the analysis couldn’t produce a precise model or showing any strong coefficient correlation for any venue type. But we can still get some meaningful and logical insights from the result.

Doing this project helps practicing every topic in the specialization, and thus, equipping learners with Data Science methodology and tools using Python libraries. Also doing a real project certainly helps one learns so much more outside the curriculum, as well as realizes what more to research into after completing the program. And as this report shows, there are surely a lot of things to dig into.

Toward the person that went through this project, many thanks for the time and thoughts.

References:

https://wiki2.org/en/K-means_clustering.

Table of Figures: