Predicting Resale Prices of HDB Flats in Singapore

IBM Professional Certificate in Data Science – Capstone Project (Coursera)

Pictures of the public housing in Singapore.

Source: istockphoto.con

Contents

1. Introduction

This is a project on predicting the resale prices of HDB flats.

The Housing and Development Board (“HDB”) is Singapore’s public housing authority. It plans and develops Singapore’s housing estates, building homes and transforming towns to create a quality living environment for the citizens. Within each town, HDB also provides various commercial, recreational and social amenities in the towns for the convenience of the residents.

HDB flats are home to over 80% of Singapore’s resident population, with about 90% of these resident households proudly owning their homes.

2. The Problem

Buying an HDB flat is one of the most talked about subject in Singapore among young couple starting their own families. Housing, by its very nature, is possibly one of the largest financial commitment for a young adult. Given a certain budget, where and what kind of resale flat should one buy? This problem was raised by a client who wishes to buy a flat in the town of Sengkang so that he can stay near his parents.

3. The Data Set

The data sources I use are publicly available sources and are easily found online. For HDB historical prices, I retrieved the resale transactions from Aug 2017 to date, retrieved from https://data.gov.sg/dataset/resale-flat-prices.

To extract the location data of the transaction, I extracted the coordinates (latitude and longitude) from onemap.sg using API and Python.

From the link, I will merge the information into a single file and used this merged file for analysis. The reason for separating the task is because of the long run time for the API data merge to happen.

Using predominantly linear regression, which has been proven to be a better predictor of property prices in other research paper, I shall develop a simple model to predict the resale prices of HDB flats.

4. Using API to link Data to One Map

As the data set does not have the geo locations, I used an API to One Map (https://www.onemap.sg/home/) to obtain the latitudes and longitudes of the various resale units. One Map is the authoritative national map of Singapore with the most detailed and timely updated information developed by the Singapore Land Authority.

5. Using API to link Data to Four Square

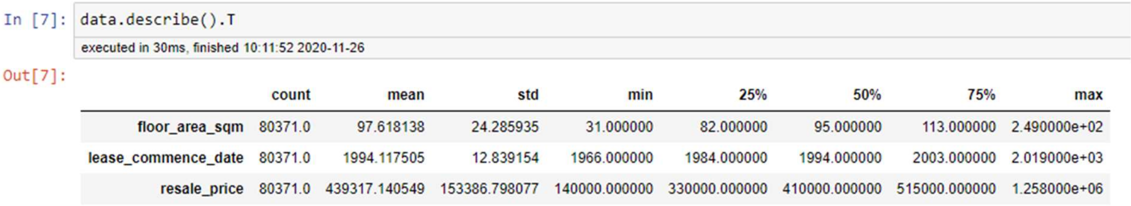
Notwithstanding that Singapore is a rather small city state and that the amenities such as schools and eateries are easily accessible, given that the HDB has a standard township plan, I shall utilise the API to Four Square to plot out the key venues for a typical HDB Town.

6. Exploratory Data Analysis

1) The data set

From the data set, between the period January 2017 – Sep 2020, there were 80,371 HDB resale transactions.

The key statistics, being the floor area (in square meters) and the resale price, is shown below.

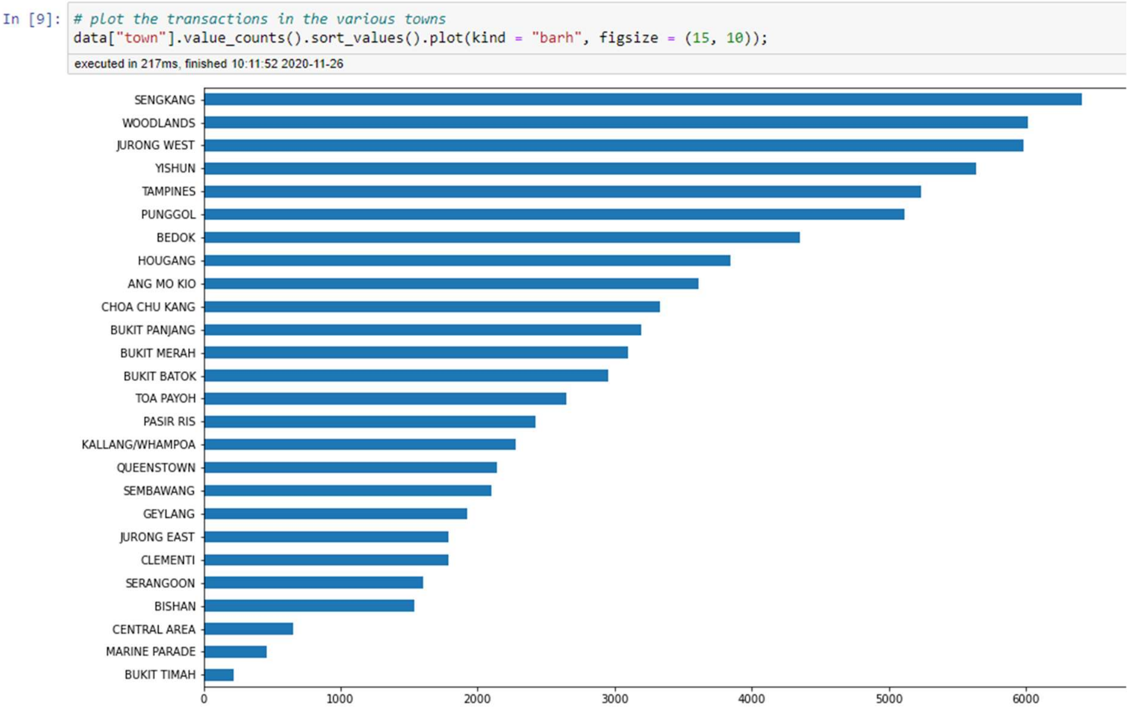


```
In [7]: data.describe().T
        executed in 30ms, finished 10:11:52 2020-11-26
Out[7]:
```

	count	mean	std	min	25%	50%	75%	max
floor_area_sqm	80371.0	97.618138	24.285935	31.000000	82.000000	95.000000	113.000000	2.490000e+02
lease_commence_date	80371.0	1994.117505	12.839154	1966.000000	1984.000000	1994.000000	2003.000000	2.019000e+03
resale_price	80371.0	439317.140549	153386.798077	140000.000000	330000.000000	410000.000000	515000.000000	1.258000e+06

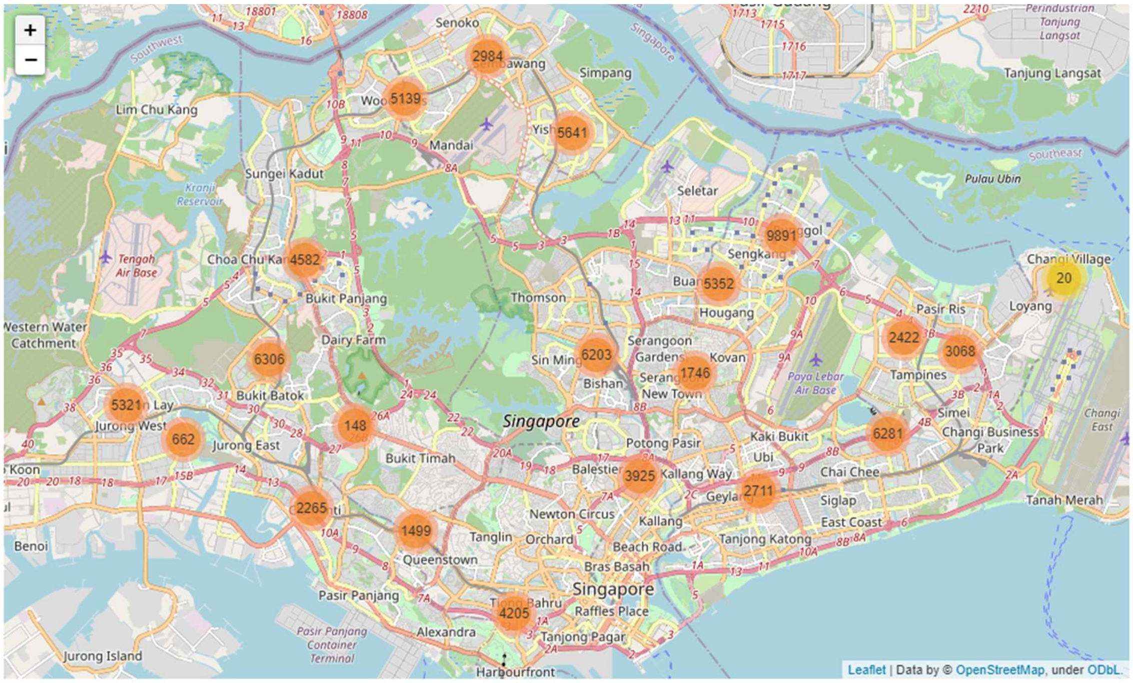
2) Overview of Sales by Geographical Location

A data count and bar chart visualisation show that there are higher number of transactions in certain younger towns. Noteworthy is that under the HDB policy, home owners that have purchased new flats from the HDB is has to live in the purchase premises for at least 5 years before the home owner can resell the flat. This is because the flats are heavily subsidised by the government in an effort to ensure that its citizens have a permanent roof over their heads. Being heavily subsidised, any re-selling usually means that the seller would have made some capital gains from the sales.



```
In [9]: # plot the transactions in the various towns
        data["town"].value_counts().sort_values().plot(kind = "barh", figsize = (15, 10));
        executed in 217ms, finished 10:11:52 2020-11-26
```

A pictorial overlay of the number of sales on folium map gives a clear visual on the spread of the resale of the HDB flat. Singapore is a very small city state and as mentioned in my introduction, 80% of the population stays in HDB flats. The flats are built across the whole island and not centred in a specific zone.

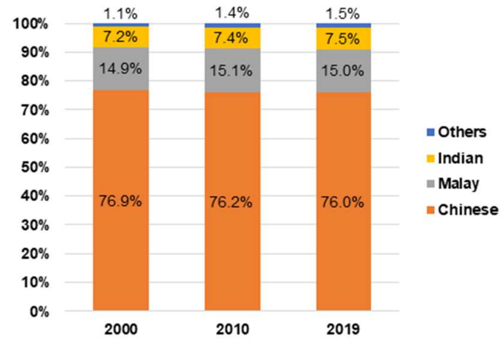


The reasons are multi-fold and I will highlight only 2 key reasons:

- There is a deliberate government policy to ensure that there are no under-developed zones in the country as this can result in social-economic issues as seen in some other countries. In some countries, there are ghettos and similar less attended areas due to neglect and discrimination; and

- Zooming down to the micro level, each HDB block has a certain racial restriction that is representative of the Singapore racial profile. The proportion of each race in the citizen population as at June 2019 (source: https://www.gov.sg/article/what-are-the-racial-proportions-among-singapore-citizens) is shown in Chart 1 below:

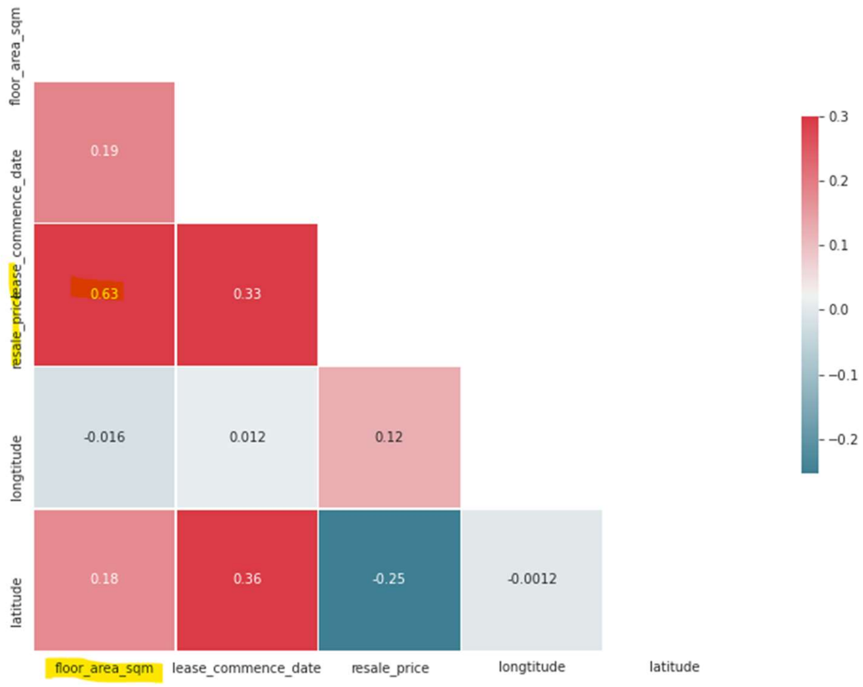
Chart 1: Proportions of each race in the citizen population, as of June



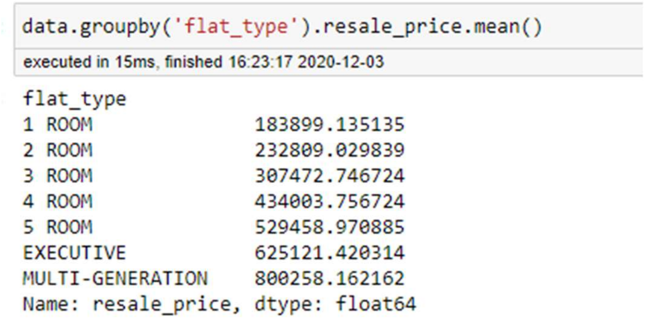
As such, each block of flats can only have around 76% of Chinese, 15% of Malay and the balance going to Indian and other races. The critical rationale is that this is also a proactive intervention to ensure that there is cultural and racial diversity within each town. This is unlike some other countries where there are zones for each race and there is no intermingling.

3) Correlation

From the correlation heatmap shown below, the most prominent correlation is between the resale price and the floor area.



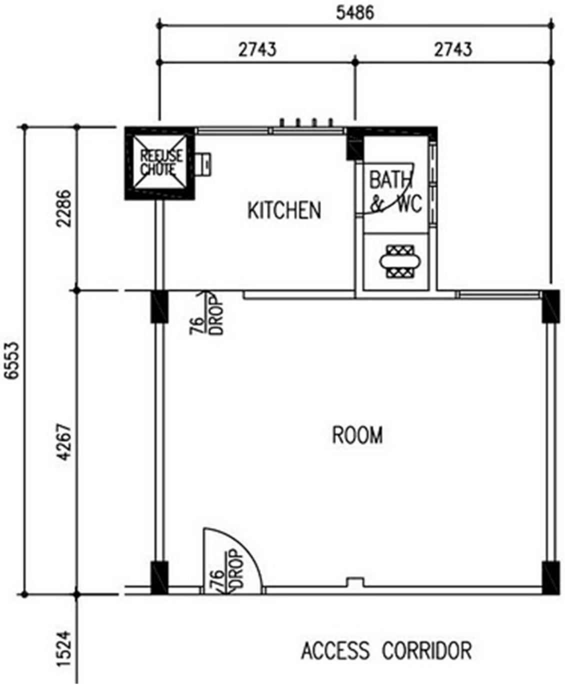
This is further collaborated by the resale mean price as follows:



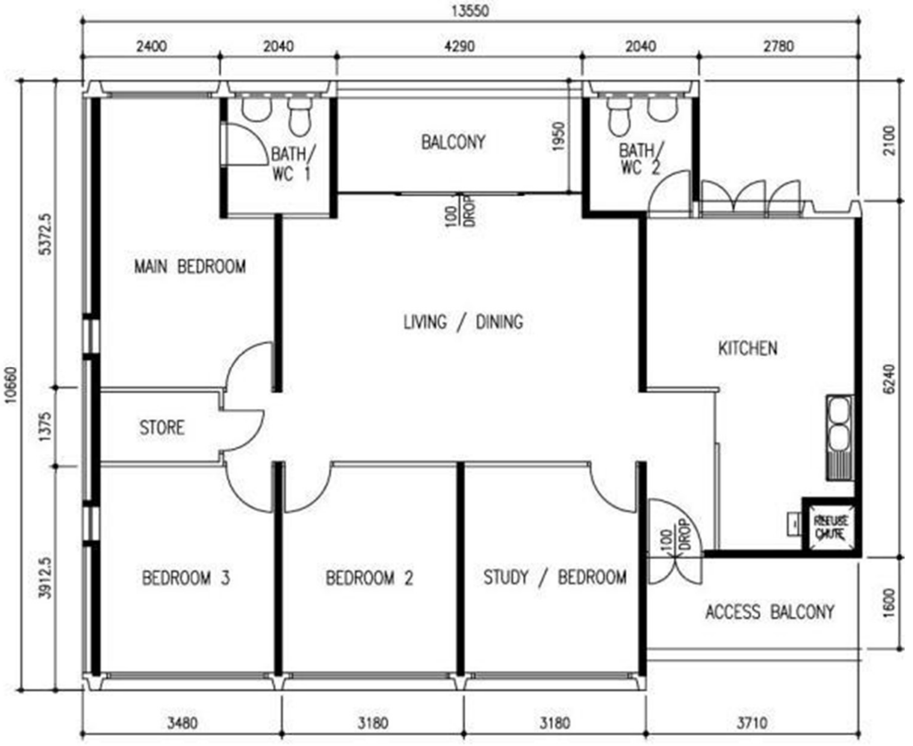
```
data.groupby('flat_type').resale_price.mean()
executed in 15ms, finished 16:23:17 2020-12-03

flat_type
1 ROOM              183899.135135
2 ROOM              232809.029839
3 ROOM              307472.746724
4 ROOM              434003.756724
5 ROOM              529458.970885
EXECUTIVE           625121.420314
MULTI-GENERATION    800258.162162
Name: resale_price, dtype: float64
```

In the Singapore context, 1 room flat is typically the smallest (in terms of floor area) and the multi-generation is the largest. As multi-generation flats are rarer, I have shown here only the floor plan of a typical 1 room flat and 5 room flat to give a context to the descriptions.

<u>Picture of a 1-room flat</u>

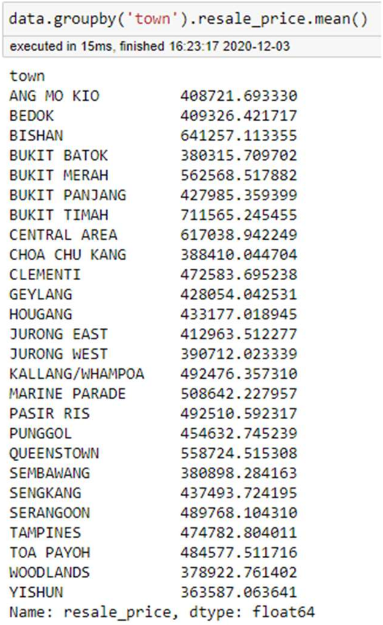


Picture of a 5-Room Flat

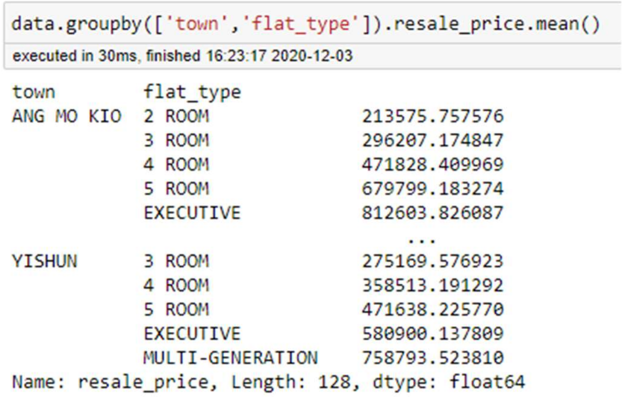


4) Differences in Prices Amongst Towns

Due to the popularity of certain estates, there are some price differences in various parts of Singapore. For example, prices in more matured estates (i.e. estates that are more developed and has more amenities) will cost more as compared to younger estates (i.e. estates that are only being developed and may lack certain amenities). This is shown by two tables below:



```
data.groupby('town').resale_price.mean()
executed in 15ms, finished 16:23:17 2020-12-03

town
ANG MO KIO         408721.693330
BEDOK              409326.421717
BISHAN             641257.113355
BUKIT BATOK        380315.709702
BUKIT MERAH        562568.517882
BUKIT PANJANG      427985.359399
BUKIT TIMAH        711565.245455
CENTRAL AREA       617038.942249
CHOA CHU KANG      388410.044704
CLEMENTI           472583.695238
GEYLANG            428054.042531
HOUGANG            433177.018945
JURONG EAST        412963.512277
JURONG WEST        390712.023339
KALLANG/WHAMPOA    492476.357310
MARINE PARADE      508642.227957
PASIR RIS          492510.592317
PUNGGOL            454632.745239
QUEENSTOWN         558724.515308
SEMBAWANG          380898.284163
SENGKANG           437493.724195
SERANGOON          489768.104310
TAMPINES           474782.804011
TOA PAYOH          484577.511716
WOODLANDS          378922.761402
YISHUN             363587.063641
Name: resale_price, dtype: float64
```



```
data.groupby(['town','flat_type']).resale_price.mean()
executed in 30ms, finished 16:23:17 2020-12-03

town        flat_type
ANG MO KIO  2 ROOM              213575.757576
            3 ROOM              296207.174847
            4 ROOM              471828.409969
            5 ROOM              679799.183274
            EXECUTIVE           812603.826087
                                     ...
YISHUN      3 ROOM              275169.576923
            4 ROOM              358513.191292
            5 ROOM              471638.225770
            EXECUTIVE           580900.137809
            MULTI-GENERATION    758793.523810
Name: resale_price, Length: 128, dtype: float64
```

For example, for a comparable sized flat (e.g. a 4-room flat) in Ang Mo Kio has a mean price of SGD472k but in Yishun, it is only at SGD359k, representing a difference of approximately 24%. Ang Mo Kio is a closer to the Central Business District (approximately 20 minutes by the Mass Rapid Transit, the local train system) as compared to 40 minutes for a train commute from Yishun.

5) Overview of Sengkang

The problem statement herein is to predict the price of HDB flat for a client. Given that this client is interested in the Sengkang estate as he prefers to live closer to his parents, we have focused the prediction using only data of resale for Sengkang estate only. This is to increase the accuracy of the prediction given that there are some town related price related adjustments that have to be performed otherwise.

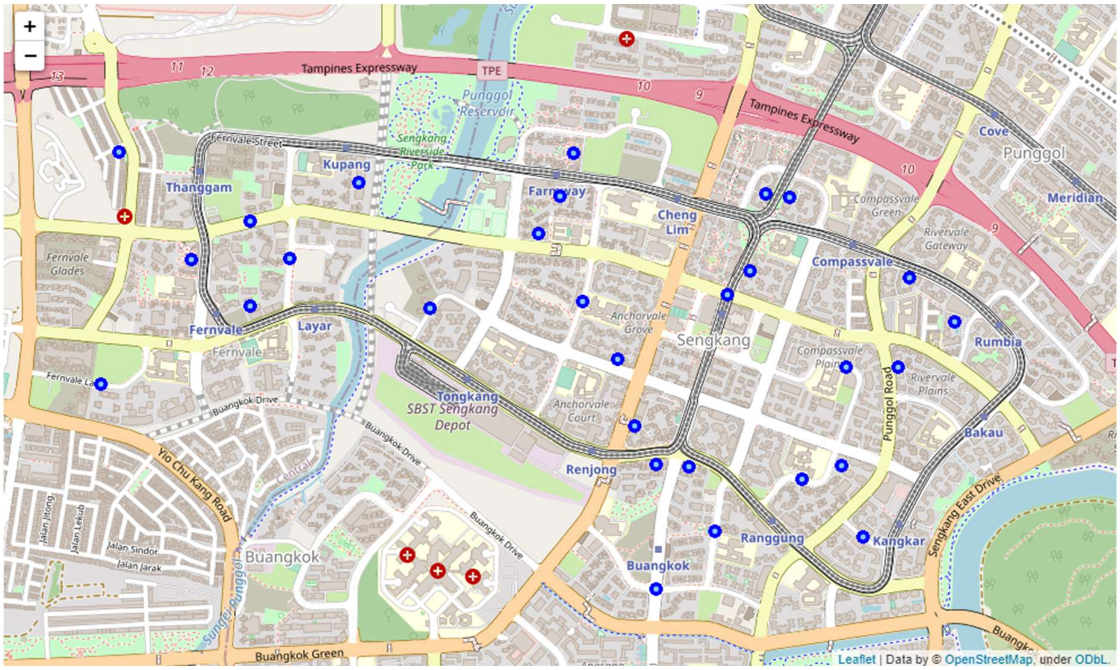
Sengkang is a primarily residential town situated to the north of Hougang New Town in the north-eastern part of Singapore,[6] under the North-East Region as defined by the Urban Redevelopment Authority (URA).

The town is bordered to the north by the Tampines Expressway (TPE), to the east by the Kallang-Paya Lebar Expressway (KPE), Yio Chu Kang Road and Buangkok Drive to the south and the Central Expressway (CTE) to the west. A new industrial area, 'Sengkang West Industrial Area', is to be built to the west of Sengkang West Road in the near future.

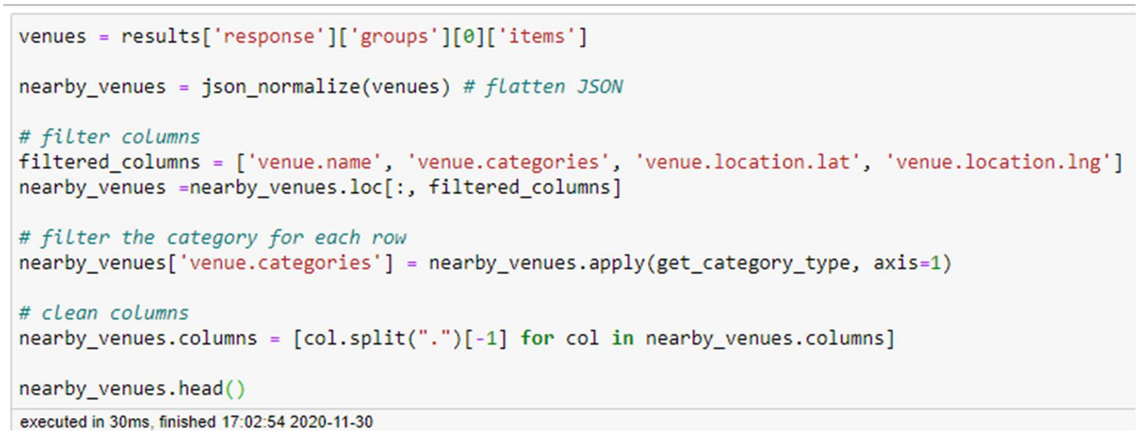
As of 1 March 2020, Sengkang has a population of 240,640, most of whom are part of the working population. The most populous subzone is Rivervale with 61,400 residents, closely followed by Sengkang Town Centre with 60,800 residents. Sengkang West, however, has just ten residents, while Lorong Halus North is completely unpopulated. Packed into an area of 10.59 km^2 (4.09 sq mi), of which just 3.97 km (1.53 sq mi) are designated as residential areas, Sengkang has a population density of 22,000 people per km^2 (57,000 per mi).

(source: https://en.wikipedia.org/wiki/Sengkang#:~:text=Sengkang%20is%20a%20planning%20area,to%20244%2C600%20residents%20in%202019.)

Using Folium, I have mapped out the various parts of Sengkang.

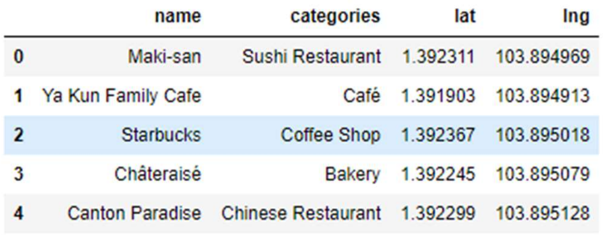


The various main venues as extracted by Four Squares are as follows:

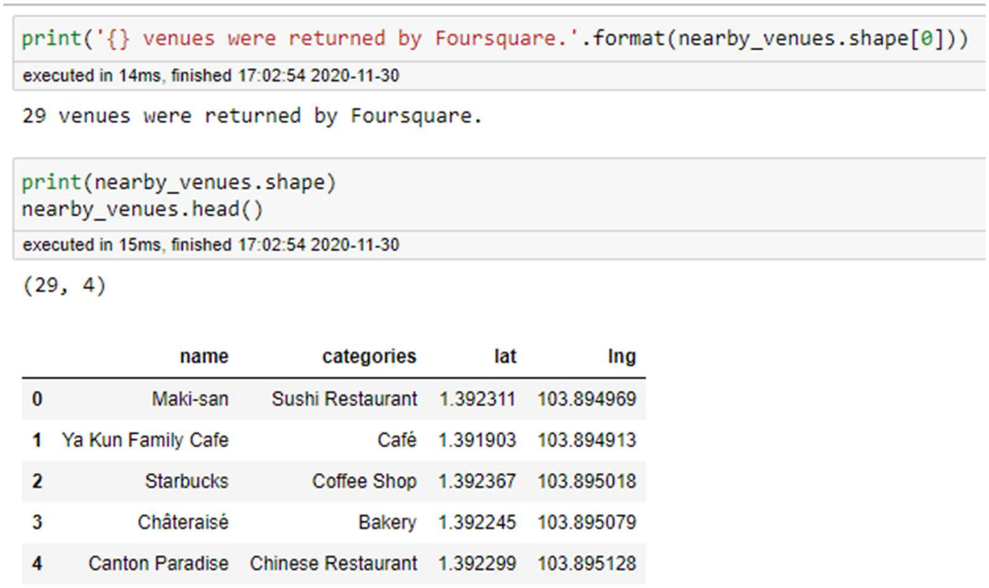


```
venues = results['response']['groups'][0]['items']

nearby_venues = json_normalize(venues) # flatten JSON

# filter columns
filtered_columns = ['venue.name', 'venue.categories', 'venue.location.lat', 'venue.location.lng']
nearby_venues =nearby_venues.loc[:, filtered_columns]

# filter the category for each row
nearby_venues['venue.categories'] = nearby_venues.apply(get_category_type, axis=1)

# clean columns
nearby_venues.columns = [col.split(".")[-1] for col in nearby_venues.columns]

nearby_venues.head()
```

executed in 30ms, finished 17:02:54 2020-11-30



	name	categories	lat	lng
0	Maki-san	Sushi Restaurant	1.392311	103.894969
1	Ya Kun Family Cafe	Café	1.391903	103.894913
2	Starbucks	Coffee Shop	1.392367	103.895018
3	Châteraisé	Bakery	1.392245	103.895079
4	Canton Paradise	Chinese Restaurant	1.392299	103.895128



```
print('{} venues were returned by Foursquare.'.format(nearby_venues.shape[0]))
```

executed in 14ms, finished 17:02:54 2020-11-30

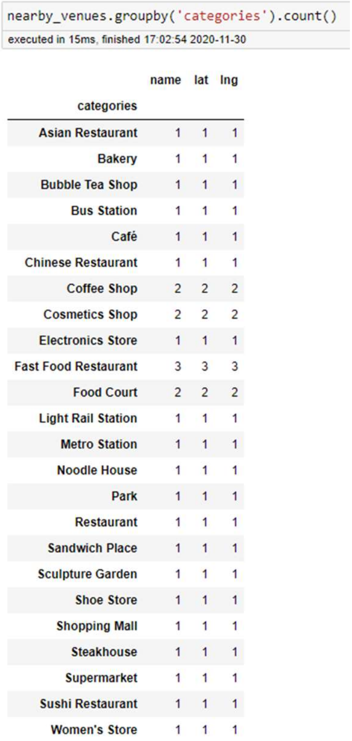
```
29 venues were returned by Foursquare.
```

```
print(nearby_venues.shape)
nearby_venues.head()
```

executed in 15ms, finished 17:02:54 2020-11-30

```
(29, 4)
```

	name	categories	lat	lng
0	Maki-san	Sushi Restaurant	1.392311	103.894969
1	Ya Kun Family Cafe	Café	1.391903	103.894913
2	Starbucks	Coffee Shop	1.392367	103.895018
3	Châteraisé	Bakery	1.392245	103.895079
4	Canton Paradise	Chinese Restaurant	1.392299	103.895128



```
nearby_venues.groupby('categories').count()
```

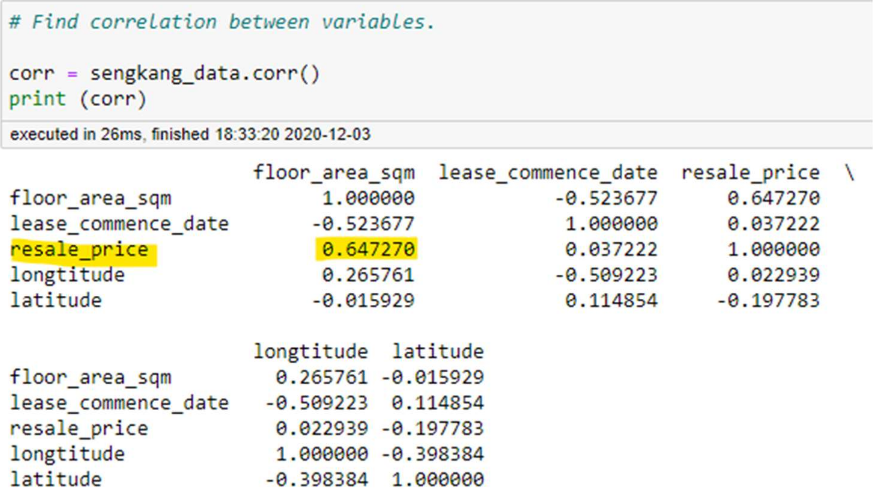
executed in 15ms, finished 17:02:54 2020-11-30

	name	lat	lng
categories			
Asian Restaurant	1	1	1
Bakery	1	1	1
Bubble Tea Shop	1	1	1
Bus Station	1	1	1
Café	1	1	1
Chinese Restaurant	1	1	1
Coffee Shop	2	2	2
Cosmetics Shop	2	2	2
Electronics Store	1	1	1
Fast Food Restaurant	3	3	3
Food Court	2	2	2
Light Rail Station	1	1	1
Metro Station	1	1	1
Noodle House	1	1	1
Park	1	1	1
Restaurant	1	1	1
Sandwich Place	1	1	1
Sculpture Garden	1	1	1
Shoe Store	1	1	1
Shopping Mall	1	1	1
Steakhouse	1	1	1
Supermarket	1	1	1
Sushi Restaurant	1	1	1
Women's Store	1	1	1

7. Price Prediction using Train-Test Split (Linear Regression)

1) Confirmation of Correlation

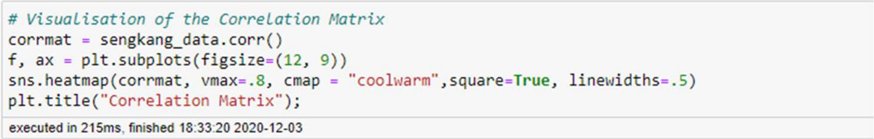
In the earlier section, the primary correlation for resale price was floor area. For the Sengkang data, I shall perform a simple re-validation before proceeding with the train test.



```
# Find correlation between variables.

corr = sengkang_data.corr()
print (corr)
```

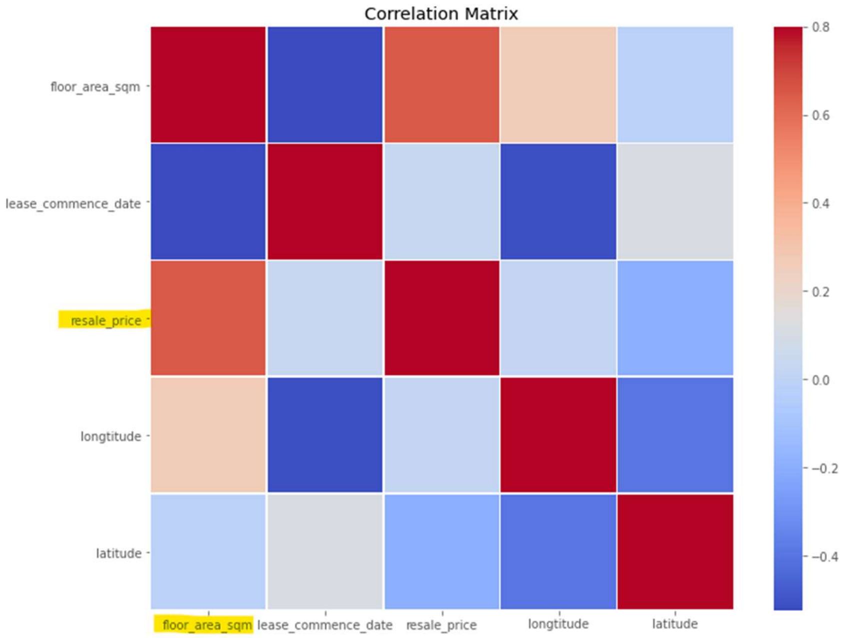
executed in 26ms, finished 18:33:20 2020-12-03

```
                   floor_area_sqm  lease_commence_date  resale_price  \
floor_area_sqm           1.000000            -0.523677      0.647270
lease_commence_date     -0.523677             1.000000      0.037222
resale_price             0.647270             0.037222      1.000000
longtitude               0.265761            -0.509223      0.022939
latitude                -0.015929             0.114854     -0.197783

                   longtitude  latitude
floor_area_sqm       0.265761 -0.015929
lease_commence_date -0.509223  0.114854
resale_price         0.022939 -0.197783
longtitude           1.000000 -0.398384
latitude            -0.398384  1.000000
```



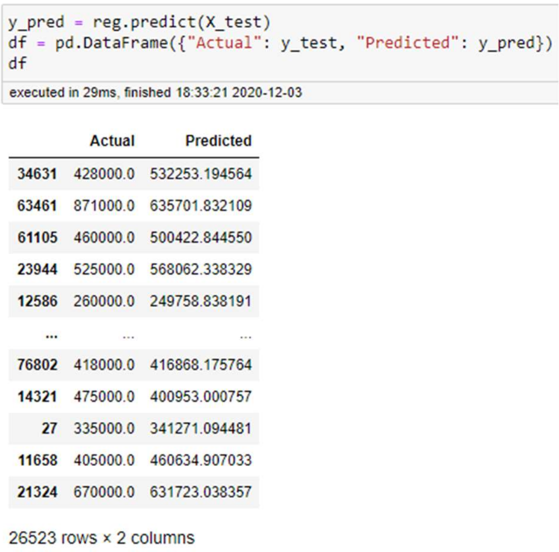
```
# Visualisation of the Correlation Matrix
corrmat = sengkang_data.corr()
f, ax = plt.subplots(figsize=(12, 9))
sns.heatmap(corrmat, vmax=.8, cmap = "coolwarm",square=True, linewidths=.5)
plt.title("Correlation Matrix");
```

executed in 215ms, finished 18:33:20 2020-12-03



2) Results of the Train-Test-Split Model

I shall not elaborate on the steps involved in the above model as this is shown in the coding. Below is actual versus predicted resale after performing the running of the model. The result shows that the model fits fairly closely to the actual result and hence, may be used for the price prediction requested by the client.



```
y_pred = reg.predict(X_test)
df = pd.DataFrame({"Actual": y_test, "Predicted": y_pred})
df
```

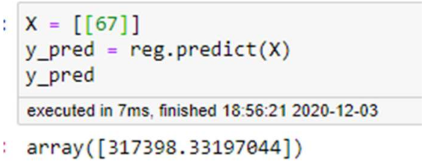
executed in 29ms, finished 18:33:21 2020-12-03

	Actual	Predicted
34631	428000.0	532253.194564
63461	871000.0	635701.832109
61105	460000.0	500422.844550
23944	525000.0	568062.338329
12586	260000.0	249758.838191
...	...	...
76802	418000.0	416868.175764
14321	475000.0	400953.000757
27	335000.0	341271.094481
11658	405000.0	460634.907033
21324	670000.0	631723.038357

26523 rows × 2 columns

8. The Prediction

The client intends to purchase a flat of approximately 67 sq.m. Fitting the floor area into the model, we get a predicted price of SGD317,398. The final contracted price will still be subject to negotiation between buyer and seller but this price will provide a good guide to the seller.



```
X = [[67]]
y_pred = reg.predict(X)
y_pred
```

executed in 7ms, finished 18:56:21 2020-12-03

```
array([317398.33197044])
```

9. Conclusion

The linear regression price prediction shows close proximity to the actual selling price, which will help the client make a more informed decision on the price that he should pay for his choice unit.