Determining factors contributing to Resale Flat Prices in Singapore

1.0. Introduction

Being an island-state, Singapore has finite land to develop infrastructure. Being only 728km^2 in size or 13,673 times smaller than the United States [1], it is no wonder property is a major talking point in the country. The government of Singapore controls the pricing of houses to prevent inflation of house prices and to ensure that an average Singaporean has a roof over his head. The two types of government-controlled properties are Built-to-Order (BTO) flats or resale flats which once were BTO flats, but the owners opt to sell them. Most couple opt for BTO flats but there are some owners who would like to sell their flats or new families looking to move to a new property.

With such demand in property, future resale flat homeowners may be interested in figuring out what determines the prices of their new homes and whether the flat is overvalued or undervalued.

This report aims to develop a model to determine the relationship of prices of resale flats to a myriad of factors. Factors such as the town where the flat resides, distance to the nearest train station, size of the flat, list of venues around the flat and the number of years left in the lease will be considered.

2.0. Data Description

There are plenty of resources which can be obtained from the web to develop the model described.

- The coordinates of Mass Rapid Transit (MRT) stations can be found on data world [2], a community-driven repository for data. The dataset includes the station names, latitude and longitude of the stations which will be used to determine the proximity of houses to the stations.

- The list of prices of resale flats can be found in the Singapore's open data repository data.gov.sg [3]. The dataset includes month of sale of the resale flats, town where flat resides, storey range, size of flat, remaining lease and the resale price of the flat. All these factors will be used to develop the model for the price of resale flat.

- To obtain the list of amenities around the flat, the Foursquare API will be used.

3.0. Methodology

The locations of MRT stations are first plotted using *folium* to determine the general vicinity of MRT stations around Singapore which can be seen in Figure 1 below:

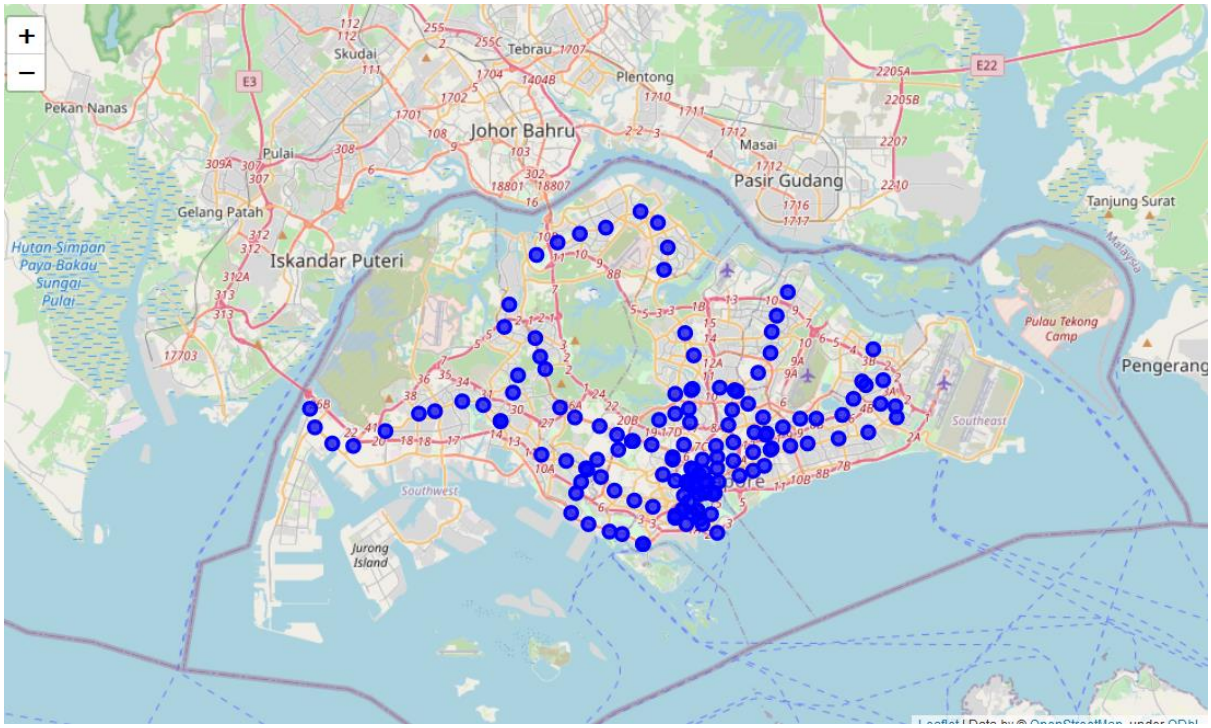

Figure 1 Location of MRT stations around Singapore

This data is plotted to determine the general vicinity of housing around Singapore as the common populace in Singapore relies on MRT to travel around.

We then use *street_name* field from the resale flat data to determine the latitude and longitude of flats of the dataset. The latitude and longitude are found using *geopy*. The latitude and longitude data are then used to determine the distance of each flat to every MRT station. The distance is calculated using *geopy* as well and we use the shortest distance to the nearest station. This distance is subsequently named as the *distance_to_mrt* field. The resultant dataframe looks like this:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 82590 entries, 0 to 82589
Data columns (total 13 columns):
town                82590 non-null object
flat_type           82590 non-null object
street_name         82590 non-null object
storey_range        82590 non-null object
floor_area_sqm      82590 non-null float64
flat_model          82590 non-null object
remaining_lease     82590 non-null object
latitude            82590 non-null float64
longitude           82590 non-null float64
STN_NAME            82590 non-null object
distance_to_mrt     82590 non-null float64
resale_price        82590 non-null float64
years_left          82590 non-null float64
dtypes: float64(6), object(7)
memory usage: 8.2+ MB
```

Foursquare API is used to determine the general venues around **1000m** from the flats based on the latitude and longitude information obtained early. The **top 10** common categories for venues are

then populated for each street address and from the top 10 per category, we get the top 20 most frequently appearing categories. The top 20 frequent categories appearing in Top 10 are:

```
['Coffee Shop',
 'Food Court',
 'Chinese Restaurant',
 'Asian Restaurant',
 'Fast Food Restaurant',
 'Café',
 'Bakery',
 'Supermarket',
 'Bus Station',
 'Noodle House',
 'Park',
 'Japanese Restaurant',
 'Indian Restaurant',
 'Dessert Shop',
 'Basketball Court',
 'Seafood Restaurant',
 'BBQ Joint',
 'Convenience Store',
 'Hotel',
 'Bus Stop']
```

Most are essentials such as Supermarket, Convenience Store and Food Court are in the Top 20 frequent categories.

To determine whether any of these categories play a part in influencing the prices of houses, we have to use get_dummies() to split them into individual columns, doing such results in 21 columns from just venue categories alone.

From the resale flat data, it seems like *town, flat_type, storey_range* and *flat_model* are categorical variables, I use the get_dummies() function to split them into dummy variables as well to be fed into the Linear Regression Model.

The resultant dataframe looks as such which will be fed into the linear regression model which is imported from statsmodels package as the statsmodels package provide more insights on the goodness of fit of the data and whether there is multi-collinearity between variables.

	ANG MO KIO	BEDOK	BISHAN	BUKIT BATOK	BUKIT MERAH	BUKIT PANJANG	CENTRAL AREA	CHOA CHU KANG	CLEMENTI	GEYLANG	...	Basketball Court	Seafood Restaurant	BBQ Joint	Convenience Store	Hote
0	1	0	0	0	0	0	0	0	0	0	...	0.0	0.0	1.0	1.0	0.
1	1	0	0	0	0	0	0	0	0	0	...	1.0	0.0	0.0	1.0	0.
2	1	0	0	0	0	0	0	0	0	0	...	0.0	0.0	0.0	1.0	0.
3	1	0	0	0	0	0	0	0	0	0	...	0.0	0.0	1.0	1.0	0.
4	1	0	0	0	0	0	0	0	0	0	...	0.0	0.0	0.0	1.0	0.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
82585	0	0	0	0	0	0	0	0	0	0	...	0.0	0.0	0.0	0.0	0.
82586	0	0	0	0	0	0	0	0	0	0	...	0.0	0.0	0.0	0.0	0.
82587	0	0	0	0	0	0	0	0	0	0	...	0.0	0.0	0.0	0.0	0.
82588	0	0	0	0	0	0	0	0	0	0	...	0.0	0.0	0.0	0.0	0.
82589	0	0	0	0	0	0	0	0	0	0	...	0.0	0.0	0.0	0.0	0.

82590 rows × 74 columns

Running model results in a R^2 = 0.859 but there are strong multi-collinearity as indicated by the statsmodel package as seen below.

```
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.95e-22. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Therefore, to reduce multi-collinearity, I use the Variance-Inflation Factor (VIF) method which is a good indicator on whether a variable is strongly correlated with other variables [4]. A VIF of > 5 shows that there is a strong correlation with other variables. I remove the variables with the highest VIF and iterate until there are no more variables with VIF > 5. As a result, the following remaining variables are obtained:

```
'ANG MO KIO', 'BEDOK', 'BISHAN', 'BUKIT BATOK', 'BUKIT MERAH',
'BUKIT PANJANG', 'CENTRAL AREA', 'CHOA CHU KANG', 'CLEMENTI', 'GEYLANG',
'HOUGANG', 'JURONG EAST', 'JURONG WEST', 'KALLANG/WHAMPOA',
'MARINE PARADE', 'PASIR RIS', 'PUNGGOL', 'QUEENSTOWN', 'SEMBAWANG',
'SENGKANG', 'SERANGOON', 'TAMPINES', 'TOA PAYOH', 'WOODLANDS', 'YISHUN',
'2 ROOM', '3 ROOM', '4 ROOM', '5 ROOM', 'EXECUTIVE', 'Adjoined flat',
'Apartment', 'DBSS', 'Improved', 'Improved-Maisonette', 'Maisonette',
'Model A', 'Model A-Maisonette', 'Model A2', 'Multi Generation',
'New Generation', 'Premium Apartment Loft', 'Premium Maisonette',
'Simplified', 'Standard', 'Terrace', 'Type S1', 'Type S2', 'Café',
'Bakery', 'Supermarket', 'Bus Station', 'Noodle House', 'Park', 'Japanese
Restaurant', 'Indian Restaurant', 'Dessert Shop', 'Basketball Court',
'Seafood Restaurant', 'BBQ Joint', 'Convenience Store', 'Hotel', 'Bus
Stop'
```

It seems like most of the categorical variables remain. The resultant model has a $R^2 = 0.794$ which seems to show relatively linear correlation.

4.0. Results of Linear Regression

4.0.1. Impact of Towns on Resale Prices

From the VIF iteration method, none of the towns are removed and therefore it plays a significant part in pricing of resale flats. The table below shows the coefficient for linear regression for the towns with SEMBAWANG town as a reference:

town	coef
BUKIT TIMAH	332800
QUEENSTOWN	291000
BUKIT MERAH	263900
BISHAN	255000
MARINE PARADE	237200
CLEMENTI	207300
CENTRAL AREA	188900
GEYLANG	174800
KALLANG/WHAMPOA	165300
TOA PAYOH	161900
SERANGOON	161400
ANG MO KIO	156300
BEDOK	130400
JURONG EAST	105600
TAMPINES	104900
PUNGGOL	101000
HOUGANG	90380
PASIR RIS	87820
SENGKANG	77130
BUKIT BATOK	76320
YISHUN	74410
BUKIT PANJANG	60380
JURONG WEST	30260
WOODLANDS	28560
CHOA CHU KANG	1117

Prime locations that are the most expensive are BUKIT TIMAH, QUEENSTOWN and BUKIT MERAH and the lowest ones are SEMBAWANG, CHOA CHU KANG and WOODLANDS.

4.0.2. Impact of Flat Type on Prices

flat type	coef
EXECUTIVE	358600
5 ROOM	348000
4 ROOM	200600
3 ROOM	72880

Executive type flats are the most expensive relative to other flat types relative to 1-room flats which is already included in the constant coefficient to prevent the dummy variable trap [5].

4.0.3. Impact of Model on Resale Prices

model	coef
Multi Generation	690600
Terrace	343700
Type S1	309900
Type S2	268200
Premium Apartment Loft	241400
DBSS	198700
Premium Maisonette	180100
Improved-Maisonette	89510
Model A-Maisonette	76990
Maisonette	68470
Apartment	47010
Model A	8318
Adjoined flat	7297
Model A2	-28610
New Generation	-67540
Simplified	-77170
Improved	-79760
Standard	-147700

Multi-generation flats cost the highest relative to the others models with Terrace flats being half of the multi-generation flats.

4.0.4. Impact of Nearby Venues to Prices of Flats

venue	coef
Hotel	63520
Japanese Restaurant	12550
Noodle House	11540
Café	10310
BBQ Joint	9890
Dessert Shop	9152
Bakery	9112
Convenience Store	1153
Park	-628
Basketball Court	-1809
Supermarket	-4317
Bus Station	-5965
Indian Restaurant	-8716
Seafood Restaurant	-9537
Bus Stop	-18400

Hotel contributes the highest to resale flat prices. This could be attributed to Hotels being in the city centre which is the prime location nearest to work.

5.0. Discussion

Variable Type	Average Cost	Median Cost
Town	$ 142,563	$ 130,400
Flat Type	$ 245,020	$ 274,300
Flat Model	$ 162,832	$ 84,635
Venues	$ 11,773	$ 9,152

With multiple variable types, the factor that contributes the highest to the prices of resale flats is **Flat Type.** As seen from the table above, the average and median cost coefficients of flat types is higher than the rest of the factors, with town coming in second and flat model coming in third.

While flat type on average seems like the highest contributing factor, one must not dismiss the impact of location of resale flats as most of the locations are relatively expensive.

While using statsmodels can help to determine the contributing factors to the prices of resale flats, this model should not be used to predict the prices of resale flats. To accurately predict the prices of resale flats, one can turn to optimization techniques such as Principle Component Analysis (PCA) which aims to reduce the variables to better achieve accuracy while losing out in interpretability of the model. In addition, different models such as Support Vector Machines (SVM) should be used for prediction as variables that are not linear can be better represented.

Such models are not used in this analysis as it is difficult to interpret the factors that contribute to resale prices. However, if one would like to focus on prediction, other techniques described above are more apt for the use case.

6.0. Conclusion

Several towns have higher prices for resale flats than others and the type of flat contribute the most to price of resale flats. While distance to MRT stations may be important, they fall short in comparison to location of flats and the type of flats which contribute more to the prices of flats.

With such analysis, buyers or sellers of resale flats get to understand better which factors contribute more to the prices of flats.

Bibliography

[1] M. L. Elsewhere, “Country Size Comparison,” [Online]. Available: https://www.mylifeelsewhere.com/country-size-comparison/singapore/united-states.

[2] H. X. Chia, “Train Stations in Singapore,” December 2020. [Online]. Available: https://data.world/hxchua/train-stations-in-singapore.

[3] Housing and Development Board, “Resale Flat Prices,” Housing and Development Board, Dec 2020 . [Online]. Available: https://data.gov.sg/dataset/resale-flat-prices.

[4] S. Saxena, “What is multicollinearity and how to remove it?,” Analytics Vidhya, Mar 2020. [Online]. Available: https://medium.com/analytics-vidhya/what-is-multicollinearity-and-how-to-remove-it-413c419de2f.

[5] S. Anand, “Dummy Variable Trap,” Medium, Apr 2019. [Online]. Available: https://www.algosome.com/articles/dummy-variable-trap-regression.html.