School of Chemistry, Chemical and Biomedical Engineering

NANYANG TECHNOLOGICAL UNIVERSITY

SINGAPORE

CB0494 Mini Project:

Analysing resale HDB flat prices in Singapore

Name: **Ong Keng Yap (U2220617L)**

Jaden Chua (U2220767J)

Team: **CBE21 Team 4**

Contribution:

	Keng Yap	Jaden
Programming	50%	50%
Report	50%	50%
Presentation	50%	50%

Content Page

Background 2

Objectives 2

Techniques used 2

Results 4

Limitations and Recommendations 5

Conclusion 6

References 7

Background

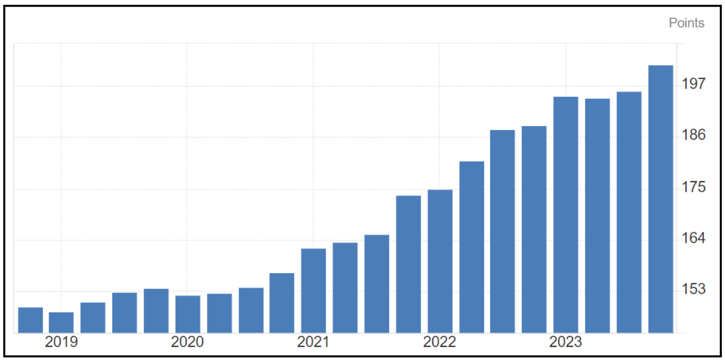


Figure 1: Graph of property price index from 2019 to 2024 [1]

In recent years, property prices have been rising rapidly. As such, new couples find it difficult to afford new homes. Young couples in Singapore face the increasingly pertinent problem of finding suitable and affordable housing, and HDB flats remain the best bang for their buck. With the supply of new BTO flats struggling to keep up with the ever-increasing demand, more and more couples have turned to the resale market in hopes of finding a new home [2]. Resale flat prices are influenced by various factors including storey range, flat type, town and floor area.

In our project, we have chosen to focus on a newly engaged couple, John and his fiance Jane. They are looking for a resale flat near Jane's parents who live in Bishan to qualify for the HDB proximity grant [3] to offset some of the costs. They also plan to have children in the near future and have the intention of staying long term. As such, the flats in consideration will be in Bishan and have either 4 or 5 rooms .

Objectives

Through this project, we aim to:

1) Give an overview of flat resale prices based on the data provided
2) Establish the important factors that impact resale price
3) Utilise machine learning techniques, namely univariate linear regression, to devise a model which predicts resale price

Techniques used

1. Data cleaning

We decided to drop irrelevant variables including 'month', 'street name' and 'block' as these data have little effect on resale price. We then converted the data type of 'remaining_lease' from object to integer and assigned a new variable 'remaining_lease_months'. Following that, we assessed the numerical variables 'remaining_lease_months' and 'floor_area_sqm' and their relationship with the resale price. There was a decent correlation between 'floor_area_sqm' and resale price, while that of 'remaining_lease_months' was low, with a correlation coefficient value of 0.13 (Figure 2). As such, we have decided to remove 'remaining_lease_months' as part of our consideration.

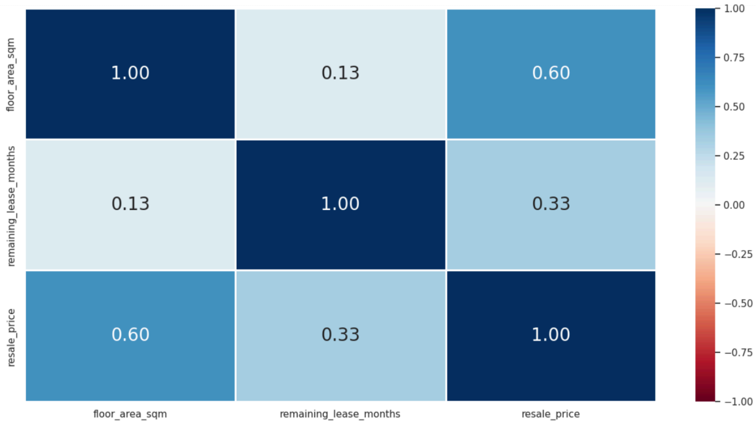


Figure 2: Heatmap of numerical variables

We then analysed the categorical data, namely 'town', 'flat_type', 'storey_range' and 'flat_model' by plotting individual boxplots of each variable against resale price. The subcategories are arranged by order of increasing median to better visualise the distribution.

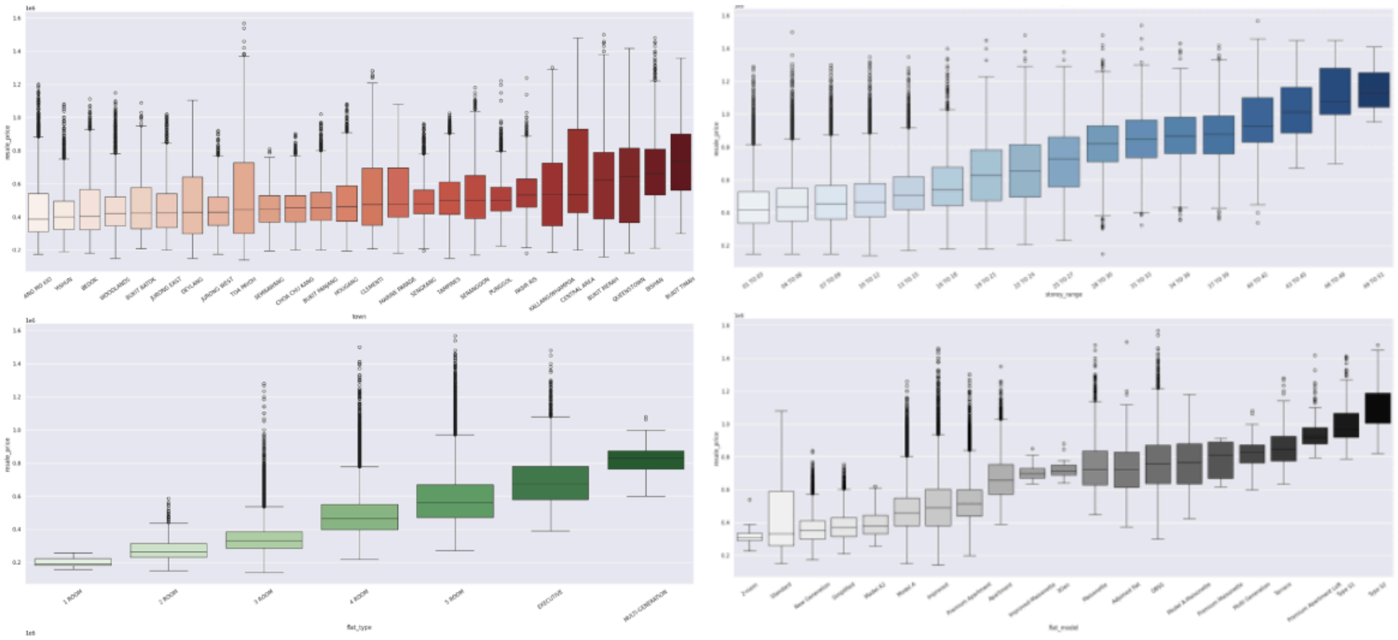


Figure 3: Boxplots for each categorical variable plotted against Resale Price

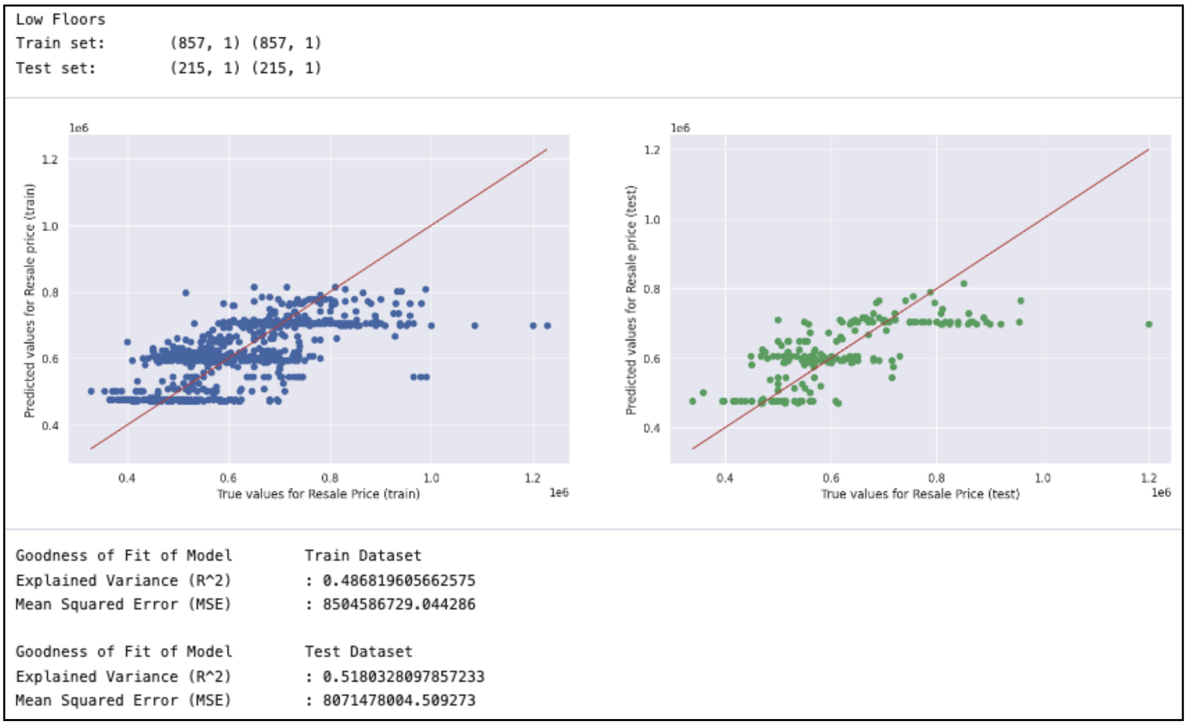
Based on the boxplots above, it is clear that all 4 categorical variables have a clear increasing trend relating to resale price and should be considered.

Drawing back to our problem statement, we have decided to focus on 4 and 5-room flats for further analysis. The couple is also looking for flats in the town of Bishan and we will isolate the data for this particular town. Additionally, there is little information regarding the flat models and new home buyers are less likely to be concerned with this factor when purchasing a flat. As such, it will be dropped from our dataset. Lastly, as ‘storey_range’ is an important factor, we will split the dataset into low, medium and high ranges and iterate over each range for 4 and 5-room resale flats in Bishan.

2. Linear Regression

Through linear regression, we have obtained a model that predicts resale price given the floor area of the flat. We first segregated the data according to low, medium and high floors, where low floors range from storey 1 to 6, medium floors from 7 to 21 and high floors from 22 to 51. We then used the LinearRegression class imported from the sklearn.linear_model library to conduct linear regression for data points which lie within the storey range. We randomly split the data into 80% for training and 20% for testing. This was done for the three defined floor ranges, and three linear regression models and results were obtained for each of low, medium and high floor ranges for 4 and 5-room resale flats in Bishan.

Results



```
Low Floors
Train set:       (857, 1) (857, 1)
Test set:        (215, 1) (215, 1)
```

1e6
1.2
1.0
0.8
0.6
0.4
Predicted values for Resale price (train)
0.4
0.6
0.8
1.0
1.2
1e6
True values for Resale Price (train)
1e6
1.2
1.0
0.8
0.6
0.4
Predicted values for Resale price (test)
0.4
0.6
0.8
1.0
1.2
1e6
True values for Resale Price (test)

```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.486819605662575
Mean Squared Error (MSE)        : 8504586729.044286

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.5180328097857233
Mean Squared Error (MSE)        : 8071478004.509273
```

Figure 3: predictor for low floors

Figure 3 shows an example of the predictor for the low floors. For a low-storey flat that the couple has found, they can use this model to predict a resale price. If the actual price is greater than the predicted price, this flat is likely overpriced in the area. The couple should consider flats with resale prices lower than the predicted prices using our model.

Our predictors are generally reliable, with good accuracy. The R^2 values show a good correlation for our train and test set. The low and medium-storey predictor has an R^2 value of about 0.5. However, the high-storey predictor has an R^2 value of only 0.1 and is less accurate. A possible reason is the lack of data points for high storey range. This will be further discussed in the following section.

Limitations and Recommendations

1. Renovation works were not accounted for

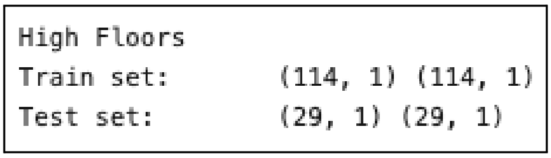
Renovation works heavily influence a flat’s price because a well-renovated flat is likely to appeal to potential buyers, which may increase the flat’s price. Renovations that were done long ago are likely to be a “liability, rather than a feature of the house” [4], and such flats are likely to be priced lower. We should include this factor in our analysis in future works to create a more accurate predictor.

2. Proximity to MRT stations was not considered

It is observed that 4-room flats located in the vicinity of an MRT station have resale prices “10% higher than those further away from the MRT station” [5]. Flats nearer to MRT stations improve convenience for travel and are likely to be more sought after. Clearly, a flat’s proximity to MRT stations has an impact on the accuracy of the prediction and should be included in our analysis for future works.

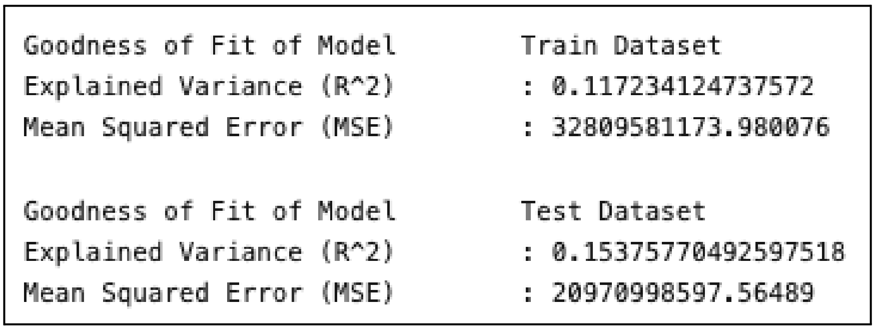
3. Lack of data points

Referring back to our problem statement, we focused solely on flat types of 4 and 5-room and the town of Bishan. This decreases the number of data points used for regression which hurts the reliability of the prediction. This problem is especially pertinent in the data regarding high floors, where there are only 114 data points used in the training set for regression.



```
High Floors
Train set:      (114, 1) (114, 1)
Test set:       (29, 1) (29, 1)
```

Figure 4: Number of data points used in train and test set for high-storey resale flats in Bishan



```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.117234124737572
Mean Squared Error (MSE)        : 32809581173.980076

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.15375770492597518
Mean Squared Error (MSE)        : 20970998597.56489
```

Figure 5: Linear Regression results for high-storey resale flats in Bishan

This lack of data widens the confidence interval and increases the variability of data, which reduces the accuracy of prediction. This is especially so for high-floor flats in Bishan, as proven by the low R^2 value and high mean squared error (Figure 5). One way to mitigate this would intuitively be to add more data points upon each sale of resale flats. This would not only provide a better representation of the population of high-storey flats, but also reduce the variability and increase the accuracy of the prediction.

4. Inability to include categorical variables for regression

Due to the nature of the available data, there are only 2 numerical variables which can be used for linear regression: 'remaining_lease_months' and 'floor_area_sqm'. Since our problem statement was defined to require resale price which is a numerical output, only linear regression can be used out of the machine learning techniques taught in this course. As we have stated in the "Techniques Used" section, the categorical variables provided in the dataset are important and cannot be ignored when building a model. However, linear regression by nature allows only numerical variables as inputs. In our analysis, we narrowed our scope in an attempt to account for the variables 'town', 'flat_type' and 'storey_range' when performing linear regression. Despite this, our method has limited effectiveness in including the categorical variables in our prediction model, causing less than ideal R^2 value and mean squared errors.

For future work, we should directly include the categorical data in our analysis. Some tools that we can utilise include one-hot encoding and random forest generators. Such tools help represent the categorical variables in our analysis more accurately and are likely to produce more accurate models.

Conclusion

In this project, we used linear regression to create models to predict the expected prices for flats in the Bishan area which fit John and Jane's expectations, which aids in their assessment of whether a flat is worth buying. We have considered various factors in our analysis and narrowed down our variables to town, floor area and storey range. Given these variables, a predicted resale price can be produced using our model. The couple should refrain from purchasing a flat with prices that are higher than the predicted price and instead opt for those that are lower.

Our models, though decently accurate, can be further improved. We have identified several key factors including renovation works, proximity to MRT stations, lack of data points and inability to include categorical data for regression, which should be worked on for future works.

References

[1]: Trading Economics (n.d.). Singapore Residential Property Price Index. https://tradingeconomics.com/singapore/housing-index

[2]: Cheng, I. (2023, November 8). *Some younger home owners prefer HDB resale flats that offer bigger living spaces, more amenities*. The Straits Times. https://www.straitstimes.com/business/property/some-younger-home-owners-prefer-hdb-resale-flats-that-offer-bigger-living-spaces-more-amenities#:~:text=Figures%20from%20HDB%20show%20that,in%202020%20and%202021%20respectively

[3]: Number One Property (2023, June 12). *HDB Proximity Grant: Unlocking Benefits of the Proximity Housing Grant (PHG) for Resale Flats.* https://numberoneproperty.com/hdb-proximity-grant/#:~:text=The%20HDB%20Proximity%20Grant%20(PHG)%20is%20a%20one%2Dtime,flat%20near%20their%20parents%2Fchild.&text=The%20grant%20provides%20up%20to,amount%20varies%20based%20on%20proximity.

[4] PropertyGuru Editorial Team (2020, November 24). *Can Renovations Affect Your Property Resale Value? Here's Our List of Do's and Don'ts.* PropertyGuru. https://www.propertyguru.com.sg/property-guides/qanvast-can-renovation-affect-property-resale-value-37428

[5] Poh, J. (2021, October 29). *HDB flats near MRT stations in Singapore*. PropertyGuru. https://www.propertyguru.com.sg/property-guides/hdb-near-mrt-how-much-more-56969