Acknowledgement

The dataset used in this project is obtained from data.gov.sg. We would like to acknowledge Housing and Development Board (HDB) for providing the tabulated dataset on resale flat price of Singapore from Jan 2017 to Feb 2022. We would also love to express our sincere gratitude to our Professor, Dr. Kelvin Yong, for the endless support and guidance to complete this project.

Background

The Housing & Development Board (HDB) was founded in 1960 (HDB, 2022) as an effort to develop housing estates, constructing homes, and reshaping cities to provide a high-quality living environment for all Singaporeans. HDB strives for exceptional outcomes in creating dynamic and sustainable communities. Statistics show that over 80% of Singaporeans are residing in public housing under HDB (R. Hirschmann, 2021). There are different flat types and flat models that provide flexibility to cater the living requirements of the residents.
Resale flat prices are influenced by many factors such as town, storey range, floor area, remaining lease and more. In this project, detailed analysis is performed to make statistical inference on the resale price of houses in Singapore. A predictor model is built to assist the buyers in finding the most preferable resale house according to their own budget.

Objectives

This project aims to:

1. Develop an overall picture of the flat resale prices using the dataset provided.
2. Determine the key factors that affect the house resale price.
3. Conduct regression models to predict the resale price.
4. Build a comprehensive pricing model to predict the selling price of resale flats.

Techniques Used

Data Acquisition and Data Cleaning: The dataset and all the basic libraries are imported. Data description is assessed, and the dataset is imported as Pandas DataFrame for analysis. The unimportant variables - “month”, “street name” and “block” are dropped. “month” only gives an indication on the date where the data is being extracted and hence does not affect resale flat price. “street name” and “block” are dropped as they may not provide any useful information. In this context, “town” could be a good variable as the proximity and accessibility of facilities at a particular town area could be further interpreted.

Part 1: Overall Analysis of Resale Price

An overall picture of resale price is performed by plotting an interactive histogram and an interactive boxplot to determine the distribution and spread of the data.

Part 2: Analysis of Numerical and Categorical Data

a. Numerical Data

Exploratory Analysis: Based on our interpretation, “floor area sqm”, “lease commence date” and “remaining lease” are identified as numerical variables. Since “lease commence date” and “remaining lease” are related, only “remaining lease” is used for numerical analysis. The data type of “remaining lease” is converted from object to integer.

The distribution of “floor area” and “remaining lease” is approaching normal distribution, and this can be justified by the skewness value of approximately zero for both variables. It is found that “floor area sqm” has a high number of outliers while “remaining lease” has no outliers. Next, ‘for’ loop is applied to draw the distribution of these variables. The correlation between the variables with “resale price” is represented by a heatmap and a pair plot is used for the visualisation of the pairwise relationship.

Linear Regression: To predict the “resale price” using “floor area sqm”, the dataset is split randomly with 80% as train set and remaining 20% as test set. The goodness of fit of the model is measured by explained variance (R^2 score).

b. Categorical Data

Exploratory Analysis: Categorical variables are extracted from the original dataset. The categorical variables are concatenated with “resale price” to examine the relationship of each variable with “resale price”. The boxplots are individually plotted to summarize the distributions of the variables and are rearranged in an increasing trend of median values to facilitate the visualization.

Part 3: Final Analysis of Data

“town”, “flat type”, “storey range”, “flat model” and “floor area sqm” are the important variables used to build a mathematical model for prediction of house resale price. The variables are encoded into corresponding numerical values by [LabelEncoder()]. For example, for “flat type”, 1 room flat is labelled as 0.

Multi-Variate Regression: The dataset is split randomly into a train set and test set in the ratio of 4:1.
Other Regression Models: Gradient boosting regression and random forest regression are used to minimize the prediction error. The R^2 scores for each regression model are computed and compared.

Part 4: Building Mathematical Model

The data for each variable is sorted in an ascending order. For example, “storey range” is sorted numerically from ‘01 to 03’ to ‘13 to 15’. A pricing model is defined and built to predict resale price according to the different inputs of the variables.

Results

Part 1

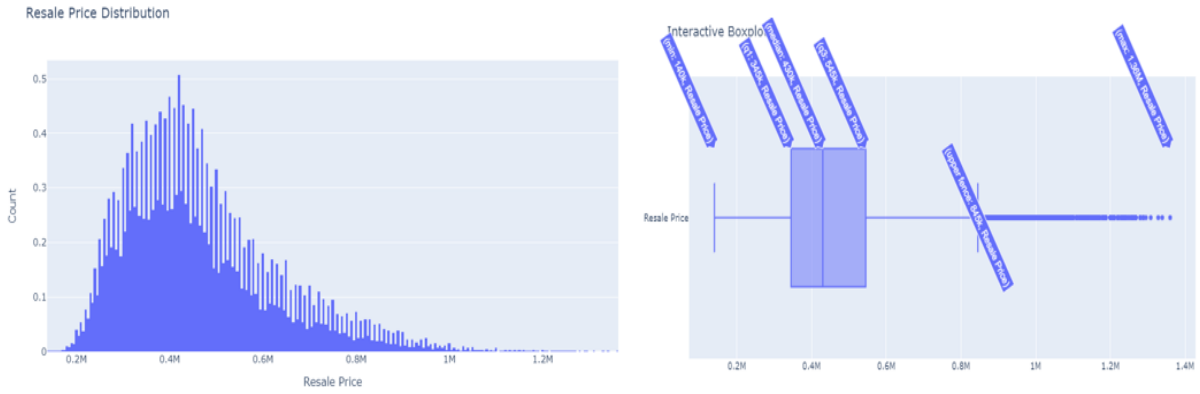

Figure 1: Interactive histogram and interactive boxplot of resale price

The percentage of outliers is 3.68% (<5%), and the skewness is 1.022, which is within the acceptable range. This implies that the resale price can be considered as a good dependent variable for statistical analysis.

Part 2

Numerical Data

Figure 2: Heatmap of correlation matrix for numerical variables

Based on Figure 2, both numerical variables show a positive correlation with “resale price”. “floor area sqm” has a higher correlation of 0.62 and linearity with “resale price" while “remaining lease” shows a low correlation of 0.34 with “resale price”. Hence, “floor area” is significant in estimating resale price.

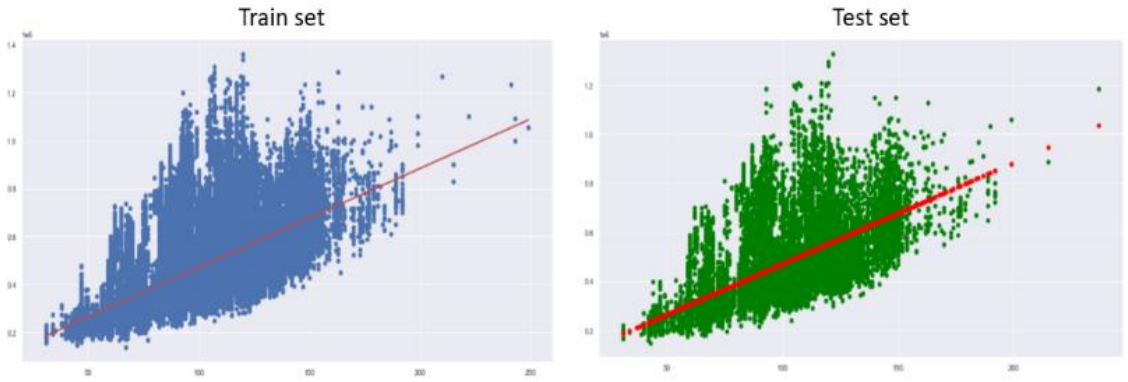

Figure 3: Uni-variate regression of resale price on floor area

From Figure 3, it is seen that the data points are far away from the linear regression line. The explained variance (R^2) was found to be about 0.39, which is relatively small. Thus, it may not be the best model to predict the selling price.

Categorical Data

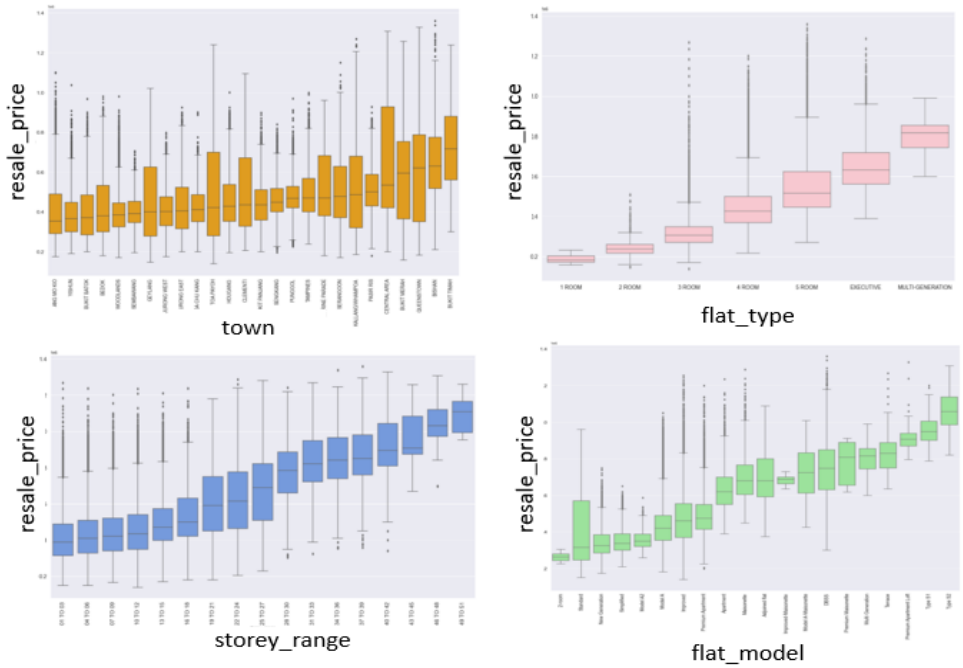

Figure 4: Boxplot for each categorical variable with resale price

Based on the boxplots, “town”, “flat type”, “storey range”, “flat model” are important variables as the boxplots show high variations in resale price across the levels.

Part 3

Regression Model

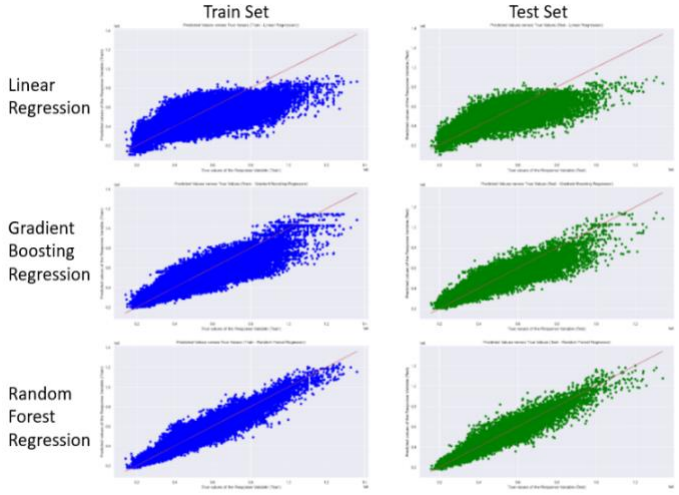

Figure 5: Regression models

	Model	R2 (Train)	R2 (Test)
0	Linear Regression	0.544613	0.542435
1	Gradient Boosting Regressor	0.783942	0.780345
2	Random Forest Regressor	0.931430	0.909987

Figure 6: R^2 of the regression models

As shown in Figure 5, the data points using the random forest regression model are the closest to the linear regression line. Based on Figure 6, random forest regression has the highest R^2 score among the three regression models. This indicates that it has a smaller error and is better at handling categorical variables. Therefore, it is the most preferred model for resale flat price prediction.

Part 4

Pricing Model

Figure 7: Predicted "resale price"

The model is used to compute the estimated resale price of a house. From the figure above, the predicted resale price is $219311 and the actual resale price is $232000 (from data point 1 in csv file).

$$\text{Percentage deviation} = \frac{\$(232000 - 219311)}{\$232000} \times 100\% = 5.47\%$$

Since the percentage deviation is relatively small, the mathematical model is valid in predicting the price of a resale flat.

Limitations

The data types for each variable are based on our own judgement, which may be inaccurate. Also, the presence of outliers in the data may affect the results. One example is that the resale flats in Bukit Timah are much more expensive than those in Ang Mo Kio.

In addition, the transformation of categorical data into numerical data by label encoding may not be the best approach. When label encoding is performed, the town names are ranked based on the alphabets although it is not an ordinal variable. To illustrate this, Ang Mo Kio is encoded with 0 and Bedok is encoded with 1. In this scenario, the model may capture the relationship between "town" in such a way that Ang Mo Kio < Bedok, which may be incorrect.

Besides, the "resale price" can be affected by other factors which are not accounted for in this analysis. For instance, economy, maturity of the estate, buyer's preference and more.

The percentage difference between the true resale price and predicted resale price computed by the model is inconsistent. An example is shown below:

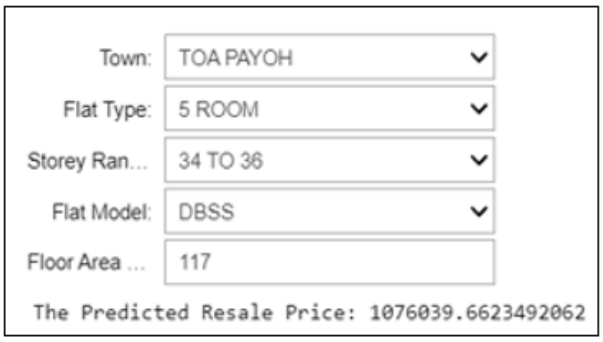

Figure 8: Predicted "resale price"

The predicted "resale price" is $1076039 while the actual "resale price" is $1220000 (from data point 116348 in csv file). The percentage deviation is calculated to be 11.80%, which is considerably high.

Recommendations

Various options for encoding categorical variables such as one-hot encoding can be considered. More factors can be included in the analysis such as proximity of the flat to facilities, accessibility, or interest rate of resale flat. Mean squared error (MSE) of different machine learning models can be compared to determine the quality of the model. The lower the MSE value, the better the model.

Conclusion

The objectives of the report are achieved. “town”, “flat type”, “storey range”, “flat model” and “floor area sqm” are used as the predictors for the resale flat prices. It is found that the random forest regression model yields the highest R^2 score and is the best learning algorithm in this case. Therefore, it is used to develop a model for the estimation of HDB resale price that provides the buyers an insight to the resale flat market in Singapore.

References

Housing and Development Board, Singapore (2022). *HDB History and Town.* Retrieved from: https://www.hdb.gov.sg/aboutus/history#:~:text=The%20Housing%20%26%20Development%20Board%20was,to%20solve%20Singapore's%20housing%20crisis [Accessed 12 March 2022]

Housing and Development Board, Singapore (2022). *Resale Flat Prices*. Retrieved from: https://data.gov.sg/dataset/resale-flat-prices [Accessed 10 March 2022]

Statista, R. Hirschmann (2021). Singapore: population living in public housing 2020. Retrieved from: https://www.statista.com/statistics/966747/population-living-in-public-housing-singapore/ [Accessed 15 March 2022]