1. INTRODUCTION

In 2022, 77% of the resident population in Singapore lived in public housing under the Housing and Development Board (HDB). With such a huge market for HDB flats, creating a model that can accurately predict the value of a HDB flat based on its intrinsic attributes and neighbourhood characteristics may be useful in real-world scenarios. By investigating factors that explain the resale prices of HDB properties as well as the significance of such variables, this report aims to achieve the above using various machine learning methods and compare them to a benchmark model to evaluate their accuracy. This report concludes with Decision Trees being the best model used for predicting HDB resale prices, followed by K-nearest Neighbours and Multiple Linear Regression. The report is organized in the following order: data discussion in Section 2, followed by data visualization in Section 3. Section 4 reports the results of my finding from the methods used, with Section 5 ending with concluding statements.

---

2. DATA DISCUSSION

The data used in this report consists of HDB flat resale transactions available for 2021 as well as geospatial information taken from open sources such Data.gov.sg and OneMap.gov.sg. The data consists of 6000 observations and 230 variables. First, I checked for missing values within the data and removed any outliers, giving me a data set left with 5865 observations. I then applied a 50/50 split to the data, with 2933 observations in the training set and another 2932 in the test set before using them in my supervised machine learning methods. Since the resale prices of HDB are large, I normalized the data by dividing them by 1000 before running the regression models.

3. DATA VISUALIZATION

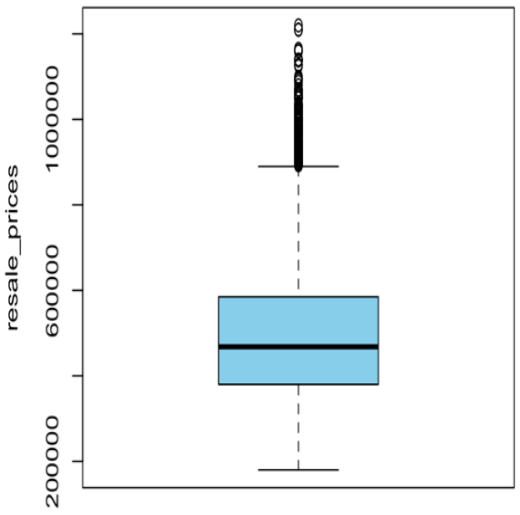


Fig. 1 Box plot of resale_prices

When plotting the original data in a box plot, many of the observations fall outside of 1.5 * IQR, making them outliers. It is interesting to note that all outliers fall above the upper bound and none under the lower bound as shown in Fig.1. This could be due to Singaporeans highly overestimating the actual value of certain HDB flats because of reasons such as misinformation or poor judgement, causing their resale prices to fall way beyond 1.5 * IQR. Therefore, they have been excluded them from our analysis so that they do not skew the results of our regression models.

A correlation matrix is created to examine the correlation between certain predictors and the HDB resale prices. From the correlation matrix in Fig. 2, we observe that *floor_area_sqm* have the highest positive correlation (0.6310), followed by *max_floor_lvl* (0.5071) and *remaining_lease* (0.3510). These positive correlations indicate that resale prices are likely higher when such intrinsic attributes of the properties increase, which makes sense as they are associated with a larger or newer housing. Continuous variables such as *Dist_nearest_mall, Dist_nearest_Ghawker, Dist_nearest_station, Dist_nearest_CC, no_primary_schools_1km* and *Dist_CBD* are negatively correlated to resale prices, which is within expectations as access to such amenities improves one's standard of living. Hence living closer to them would likely mean higher prices for apartments in the area. It is interesting to note that the presence of postal codes containing '44' is positively correlated to resale prices. This is contrary to the context provided, which states that there is an aversion towards such numbers due to superstitious beliefs. One possible reason could be that such beliefs no longer hold much weight into the consideration of buying HDBs among the younger generation of homeowners.

	resale_price
resale_price	1.0000
floor_area_sqm	0.6310
Remaining_lease	0.3510
Dist_nearest_mall	-0.0563
Dist_nearest_waterbody	0.0924
max_floor_lvl	0.5071
postal_2digits_44	0.0126
Dist_nearest_CC	-0.0203
Dist_nearest_GHawker	-0.0520
Dist_nearest_station	-0.1056
Dist_nearest_ADF	0.0934
no_primary_schools_1km	-0.1186
Dist_CBD	-0.2693
mature	0.2194

Fig. 2 Correlation Matrix Summary

---

4. FINDINGS REPORT

Principal Component Analysis / Regression

Principal Component Analysis (PCA) is used here for dimension reduction. However, as seen from Fig. 3, the proportion of variance explained by each component is small, needing 6 PCs to reach 85.0% of variance explained. The resultant scree plot produced does not show any obvious "elbow" with a significant portion of variance explained, hence it is difficult select the optimal number of PCs used. This tells us that the variables used are not highly correlated, possibly driven by many different economic forces. Thus, dimension reduction is unsuccessful. K-fold cross validation was used to select the number of PCs used in Principal Component Regression (PCR). However, it is interesting to note that the validation plot produced does not have a minimum point as seen in Fig. 4, making choice of number of components difficult. When using Principal Component Regression to predict resale prices, the model produced a high Mean Squared Error (MSE) of 36772.25. This could be because variance was not greatly reduced during PCA. Furthermore, useful signals may be within the low variance components that are discarded, worsening model predictions.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.744028	1.415076	1.172409	1.089881	1.007356	0.8546625	0.8038748	0.6387848
Proportion of Variance	0.276510	0.182040	0.124960	0.107990	0.092250	0.0664000	0.0587500	0.0371000
Cumulative Proportion	0.276510	0.458550	0.583510	0.691500	0.783750	0.8501500	0.9089000	0.9459900
	PC9	PC10	PC11					
Standard deviation	0.535989	0.4783315	0.279252					
Proportion of Variance	0.026120	0.0208000	0.007090					
Cumulative Proportion	0.972110	0.9929100	1.000000					

Fig. 3 Summary of PCs

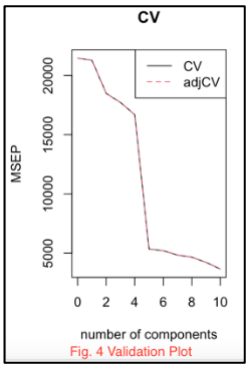


Fig. 4 Validation Plot

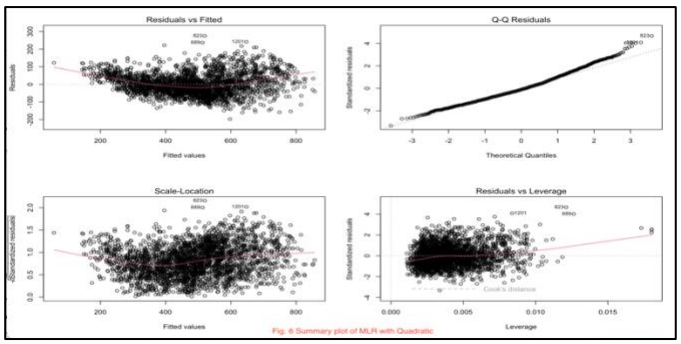
Multiple Linear Regression

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -223.66562   10.83125 -20.650  < 2e-16 ***
Remaining_lease        5.41805    0.12013  45.102  < 2e-16 ***
floor_area_sqm         4.55212    0.04857  93.714  < 2e-16 ***
Dist_nearest_mall    -22.70599    3.48637  -6.513 8.65e-11 ***
mature                56.97659    3.44538  16.537  < 2e-16 ***
max_floor_lvl          3.37739    0.22611  14.937  < 2e-16 ***
Dist_nearest_GHawker  -17.71203    0.99796 -17.748  < 2e-16 ***
Dist_nearest_CC       -46.09574    4.07844 -11.302  < 2e-16 ***
Dist_nearest_station  -36.91673    3.34993 -11.020  < 2e-16 ***
no_primary_schools_1km  -3.37814    0.86219  -3.918 9.13e-05 ***
Dist_CBD                -9.10560    0.42486 -21.432  < 2e-16 ***
Dist_nearest_ADF        -0.04583    0.85325  -0.054    0.957
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59.39 on 2921 degrees of freedom
Multiple R-squared:  0.836,     Adjusted R-squared:  0.8354
F-statistic:  1354 on 11 and 2921 DF,  p-value: < 2.2e-16
```

Fig. 5 Linear Model Summary

Fig. 5 shows the summary produced when a multiple linear regression model is fitted with a few self-chosen predictors based on intuition and domain knowledge, producing an adjusted R-squared value of 0.8354 and a relatively high MSE of 3761.893. This is because the linear regression model assumes a linear fit, while the true relationship may not be linear. From the coefficients, we observe that *Dist_nearest_ADF* is not a significant variable in the model since t–value (0.832) < 1.96, so we can remove it. We can also improve on this model by removing the additive assumption and including a quadratic term. In this case, I have chosen to introduce a quadratic term for *Remaining_lease*, improving model fit as demonstrated in Fig.6 and improving adjusted R-squared value up to 8.364. MSE is reduced to 3732.551 as well.



Decision Trees

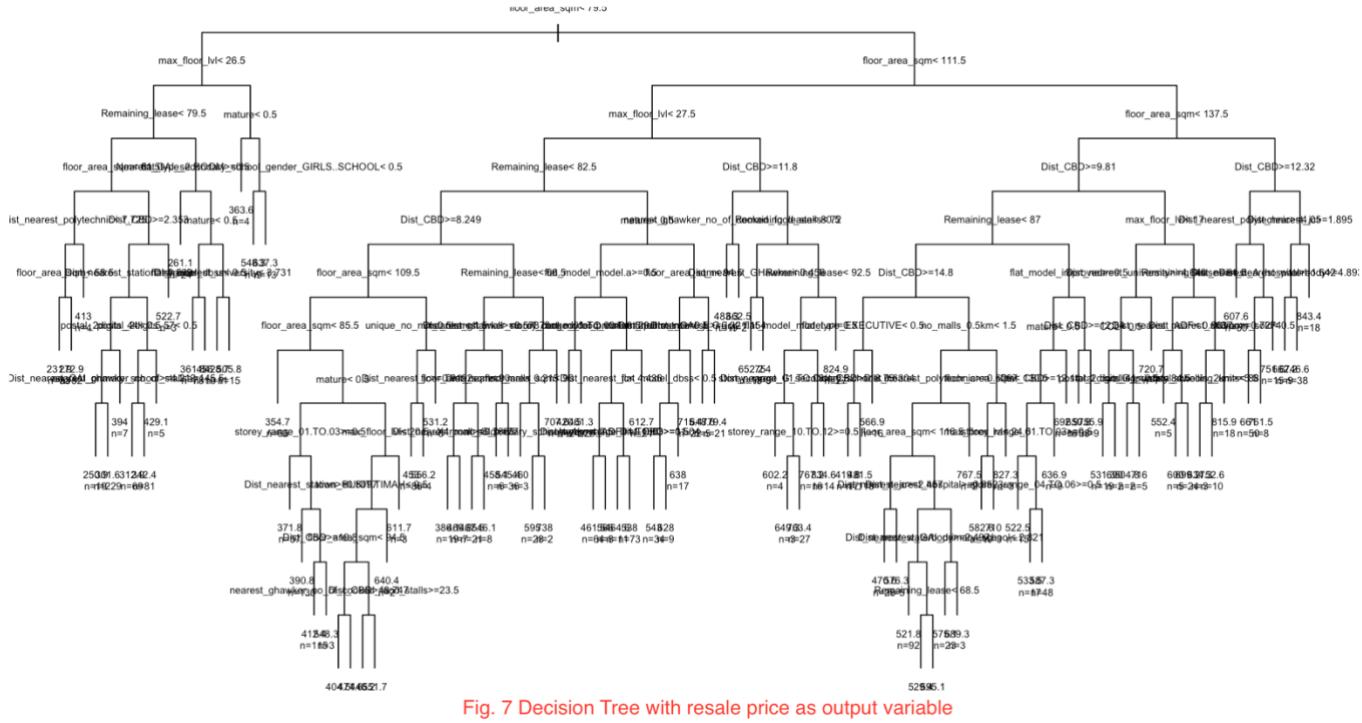


Fig. 7 Decision Tree with resale price as output variable

Decision tree is used as a supervised machine learning technique here. The best complexity parameter derived from K-fold cross validation is very small at 5.05*10^-4, meaning that there is a small penalty on complexity and thus resulting in the pruned tree being very large (Fig. 5). Evidently, it is hard to interpret the tree due to overlapping texts in later splits of the tree. Nonetheless, valuable information

can be determined from the initial splits such as *floor_area_sqm, max_floor_lvl, Remaining_lease and Dist_CBD* being the core predictors for this model. This model produces fairly accurate predictions, with an MSE of 2828.732.

K-Nearest Neighbours

This non-parametric machine learning method finds the K most "similar" training points in the predictor space and lets them vote on the class of test observations. Using cross-validation for model selection and regularization parameter tuning, we find that kbest = 3. As seen from Fig. 8, predicted resale prices largely matches the actual resale prices of HDBs in the test data. This tells us that using KNN produces decently accurate results, yielding an MSE of 3410.46. However as predicted resale prices increase, it starts to deviate from the true value. This could be due to KNN method being highly sensitive to the scale of observed variables.

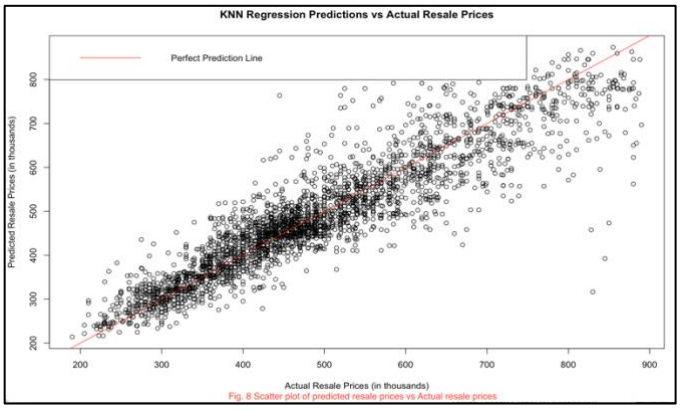


Fig. 8 Scatter plot of predicted resale prices vs Actual resale prices

5. Conclusion

Model Used	Mean Squared Error
Principal Component Analysis/Regression	36772.25
Multiple Linear Regression (MLR)	3732.55
Decision Trees	2828.73
K-nearest Neighbours (KNN)	3410.46

Table 1 Report Summary

In conclusion, the methods used in this report have proven that HDB resale prices are driven by both structural and environmental characteristics, proving the hedonic pricing theory to be true. This study also tells us that it is indeed worth to go beyond baseline methods and utilize more advanced machine learning methods to predict HDB prices. As seen in Table 1, MLR produced an MSE of 3732.55 and serves as a good basis of comparison. Both Decision Trees and KNN achieved better results, with a smaller MSE of 2828.73 and 3410.46. PCA/PCR produced the worst results in predicting resale prices, with an MSE of 36773.25. Thus, we conclude that decision trees is the best predictive model. The methods used here seem to agree that intrinsic attributes such as *floor_area_sqm* and *Remaining_lease* are the main predictors in HDB resale prices, with extrinsic attributes contributing to a smaller extent. As input variables were chosen based on intuition for certain models, there is a possibility of omitting other important variables that could have had a significant impact on HDB resale prices. Also, the dataset used may not entirely capture all possible predictors that affect HDB resale prices, even with 230 variables. One example would be construction works in the area, which could decrease housing value due to noise pollution or future government projects which would increase current housing prices. Nevertheless, the Decision Tree model would still be useful in real-world predictions as it is able to provide valuable insights into HDB resale prices.