BT4240 Machine Learning for Predictive Analytics

AY 2024/25 Semester 2

Predictive Modeling for HDB Resale Flat Pricing

Group 2 Final Report

Members:	Matriculation Number:
AARON TEO YUAN CAI	A0269646L
LAI HUI LING	A0254636Y
LEE WEI KIAT	A0273289N
LIANG SHI YIN, MARCUS	A0277877B

Table of Contents

1. Introduction

In recent years, Singapore's public housing resale market has seen a sharp price increase, sparking concerns over affordability and fair valuation. In 2024, resale flat prices climbed by 9.6%, nearly double the 4.9% rise recorded the year before (Lok, 2025). This rapid escalation has made it **increasingly difficult to assess property values fairly, especially in neighborhoods with sparse transaction data** (DollarBack Mortgage, n.d.).

This project proposes the development of a predictive model that utilizes information from comparable neighborhoods and their surrounding amenities to better estimate resale flat prices accurately. Through advanced data analytics, the model aims to deliver actionable pricing insights to both buyers and sellers, supporting more informed real estate decisions. Additionally, it seeks to serve as a reliable valuation tool in data-limited areas, helping to minimize uncertainty in property pricing.

1.1 Data Sources

The dataset was sourced from Singapore's official government data portal, published by the Housing & Development Board on February 25, 2025. It covers resale transactions from January 2014 to February 2025 with a total of 254,086 entries:

1. *Resale Flat Prices (Based on Registration Date), From Mar 2012 to Dec 2014*: Extracted only 2014 records (16,096 entries)
2. *Resale Flat Prices (Based on Registration Date), From Jan 2015 to Dec 2016*: 37,153 entries
3. *Resale flat prices based on registration date from Jan-2017 onwards*: 200,837 entries

(Refer to Table A1: Raw Data Columns & Definitions)

1.2 Existing Work & Related Studies

Many existing projects on HDB resale price prediction have incorporated geospatial and amenities-related data during their feature engineering stages. These projects often relied on relatively simple models such as Linear Regression, K-Nearest Neighbours (KNN), and Random Forests. Model performance was typically improved through standard techniques like feature selection and grid search. For example, one project reported an RMSE of approximately $46,000 using these methods (Loy, 2023), while another achieved a lower RMSE of around $35,000 by incorporating more refined features and tuning (Grrrnt, 2021).

However, many of these studies primarily focused on improving prediction accuracy and minimizing errors without delving into comprehensive analyses of model behaviour or interpretability. In contrast, our project not only revisits these classical approaches but also extends the methodology to include more complex models such as Support Vector Regression (SVR) with polynomial and radial basis function (RBF) kernels, as well as neural networks. These models can capture potential non-linear relationships between input features and resale prices.

Moreover, we place greater emphasis on in-depth evaluation of model performance and prediction errors to generate insights into model interpretability and trustworthiness. For an overview of our methodology, refer to Figure A1: Project Pipeline.

2. Preliminary Analysis

2.1 Data Cleaning

During the initial data exploration, it was found that 91 out of the total 254,086 entries contained missing values in several important columns, such as *max_floor_lvl*, *residential*, *commercial*, *market_hawker*, and *miscellaneous* (Figure B1). A closer examination traced these missing entries to areas undergoing en bloc redevelopment, particularly in the West Coast and Tanglin Halt regions, as reported by recent news sources (Heng, 2016; Channel News Asia, 2024). En bloc redevelopment involves the collective sale of buildings for reconstruction, often resulting in incomplete datasets. To preserve the integrity of our analysis, these 91 records were excluded from the dataset.

Furthermore, a duplication check identified 624 duplicate records within the dataset (Figure B2). Duplicate entries can distort statistical analyses, introduce biases, and lead to inaccurate insights. To ensure data integrity and reliability, these duplicate records have been removed from the dataset.

With the **removal of both missing and duplicate records**, the dataset now consists of **253,371** unique and complete observations. This refined dataset enhances the accuracy of subsequent analyses by eliminating inconsistencies and ensuring that each entry represents a distinct and valid observation.

2.2 Statistical Analysis

2.2.1 Resale Price

The *resale_price* of properties exhibits a right-skewed distribution, as shown in Figure B3. Most transactions fall within the $370,000 to $590,000 range, with a median price of $460,000. However, a small number of properties are transacted at significantly higher prices, creating a long tail towards the higher end. The presence of these high-value transactions is also evident in the boxplot, where multiple outliers extend beyond the upper whisker. Nonetheless, we decided not to remove the outliers as they represent genuine high-value transactions rather than data errors. Removing them could lead to a loss of important information, especially for understanding the factors influencing premium-priced properties.

To further investigate the trends in *resale_price* over time, we plotted a graph of average resale price against transaction month (Figure B4). The analysis revealed fluctuations in property prices over different periods, suggesting that inflation and market trends may have influenced these variations. To account for inflationary effects, we created a new column, *adjusted_price*, which represents the *resale_price* adjusted using the Resale Price Index (RPI) obtained from HDB Resale Statistics. This adjustment ensures that price comparisons across different periods are more meaningful by reflecting real purchasing power rather than nominal values.

We then analyzed the distribution of *adjusted_price* and observed that similar to the *resale_price*, it remains right-skewed, as depicted in the histogram (Figure B5). Given this skewness, we applied a log transformation to *adjusted_price* to normalize its distribution. This transformation helps reduce the impact of extreme values, making the data more suitable for statistical modeling and predictive analysis.

Thus, *adjusted_price* will be used as our response variable, and we will utilize ***log_adjusted_price*** **as the dependent variable in our predictive modeling**.

2.2.2 Town

The location of a property plays a crucial role in determining its *resale_price*, as factors like demand, housing supply, and local amenities vary across towns. Areas such as Sengkang, Woodlands, and Jurong West tend to have higher transaction volumes because of their larger supply of public housing. In contrast, premium locations like Bukit Timah, the Central Area, and Marine Parade generally see higher resale prices due to their attractive settings and limited housing availability (refer to Figure B6).

To better understand these trends, towns are grouped into **five *regions***: Central, North-East, North, West, and East. The Central region has the highest prices due to its proximity to commercial hubs, while the North and West regions provide more affordable housing options with high transaction volumes (Figure B7). The North-East offers a mix of mid-range and affordable flats, whereas the East has diverse pricing trends, with Marine Parade being a premium location.

Given that *bldg_contract_town* provides similar information to our town-based analysis, we have decided to **drop** the column. This approach ensures that the model remains efficient without compromising the predictive power of location-related factors.

2.2.3 Flat Type

Examining the distribution of *flat_type* provides valuable insights into the resale market. Among all categories, 4-room flats dominate the market with 106,117 transactions, while multi-generation flats are the rarest, with only 88 sales from 2014-2024 (Figure B8). Despite their low count, multi-generation flats have the highest average adjusted resale price at $1,133,595, likely due to their larger size.

2.2.4 Storey Range

Referring to Figure B9, mid-level storeys (04-06 and 07-09) are the most common among resale transactions. However, higher floors command higher adjusted resale prices, reflecting buyer preference for better views, ventilation, and reduced noise. This trend highlights the importance of *storey_range* in predicting resale prices.

2.2.5 Floor Area

The distribution of *floor_area_sqm* significantly influences *adjusted_price* in Singapore. Floor areas range from 50 sqm to 350 sqm, catering to diverse housing needs. In the Central Region, the adjusted price per square meter is highest due to its prime location and amenities (Figure B10). The East and North-East Regions offer a mix of floor areas, balancing affordability and space. Meanwhile, the West and North Regions provide larger units at lower *adjusted_price*, appealing to buyers who prioritize space over centrality.

2.2.6 Flat Model

The distribution of *flat_model* in the resale market shows that Model A and Improved flats dominate transactions, while premium models like Multi-Generation and Premium Maisonette

are the rarest, likely due to their high prices and limited supply (Figure B11). Price analysis reveals that premium models command significantly higher *adjusted_price*, reflecting their larger size and exclusivity. These variations highlight the need to account for *flat_model* differences when predicting resale prices, as both transaction volume and pricing trends vary widely across categories.

2.2.7 Remaining Lease

The *remaining_lease* on a property is essentially the duration for which the property can be utilized. As evident in Figure B12, we can see a positive correlation between the *remaining_lease* and *adjusted_price*. For properties with a shorter *remaining_lease*, there may be restrictions to housing loans and Central Provident Fund (CPF) usage, causing them to be less attractive (HomeGen, 2023). This results in lower demands and thus lower prices. This observed trend highlights the potential of *remaining_lease* as a predictor for resale prices.

2.2.8 Maximum Floor Level

The distribution of maximum floor levels in the resale market reveals that most transactions occur in mid-rise buildings, with the highest concentration around 10 to 15 floors (Figure B13). Taller buildings above 30 floors are significantly less common, suggesting that high-rise flats are either less frequently available or in lower demand. This variation in floor levels may influence pricing trends, as higher floors often command price premiums due to better views and reduced noise, highlighting the importance of considering *max_floor_lvl* in resale price predictions.

2.2.9 Total Dwelling Units

The distribution of *total_dwelling_units* indicates that most resale transactions occur in developments with a moderate number of units, while very small or extremely large developments are less common (Figure B14). Larger developments may offer more amenities and shared facilities, potentially driving demand, whereas smaller ones might provide exclusivity but limited communal features. Understanding this distribution is essential, as estate size can impact both buyer preferences and resale values.

2.2.10 Binary Variables

From Figure B15, we observe that *residential, market_hawker, multi-storey_carpark, and precinct_pavilion* exhibit highly skewed distributions. Given the lack of meaningful variation, these variables provide little predictive value and are therefore, **dropped** from our analysis.

However, *commercial* and *miscellaneous* display a more balanced distribution and a potential influence on *adjusted_price*. Properties in non-commercial areas have a higher average *adjusted_price* compared to those in commercial zones (Figure B16), suggesting that proximity to commercial spaces may negatively impact prices. Similarly, properties without miscellaneous facilities tend to have slightly higher prices than those with such facilities (Figure B17), indicating that additional amenities do not necessarily translate to higher resale values.

These preliminary insights will guide our modeling approach as we further analyze their impact on resale price prediction.

2.3 Feature Engineering

2.3.1 Proximity Based Features

In addition to *adjusted_price* and *region*, our model incorporates the proximity of key amenities such as hawker centers, schools, and MRT stations, which play a crucial role in influencing HDB resale prices by enhancing convenience. To integrate this information, we first retrieved the latitude and longitude of each resale location using *street_name* and *block* with the OneMap API. Afterward, we mapped the latitude and longitude of each resale flat against all identified amenities. Using the **Haversine formula**, we calculated and recorded the number of hawker centers, schools, and MRT stations located within 500m, 800m, and 1 km of each flat in separate columns. Additionally, we calculated the distance of each resale flat to the nearest hawker center, primary school, secondary school, junior college, and MRT. Figure B18 shows some examples of the proximity-based features engineered.

2.3.2 Ordinal and One Hot Encoding

Since the machine learning models we intend to use cannot process textual data, all categorical columns, such as *flat_model* and *storey_range*, need to be transformed into a numerical format. This conversion is achieved using either ordinal or one-hot encoding, depending on whether the column has a meaningful ranking. For columns such as *flat_model* with no inherent ranking, each unique category is converted into a separate binary column, with values 0 or 1 representing the absence or presence of that category. On the other hand, for columns like *storey_range*, which have a natural order, numeric values are assigned based on the floor range (e.g., “01 to 03” becomes 1, “4 to 6” becomes 2). This approach ensures equal treatment for each unique value by preventing unintended ranking for non-ordinal data, while preserving the order in ordinal data. Figure B19 shows some examples of the new way the data is being represented.

One-Hot Encoded features: *town, region, flat_type, flat_model, commercial, miscellaneous*
Ordinal Encoded features: *storey_range*

2.4 Correlation

Figure B20 shows the correlation between the numerical variables in the dataset. Based on the correlation heatmap, we can see that the proximity-based features for the number of proximities within the different distances are quite highly correlated, which is expected as the higher distances contain information about the lower distances. Therefore, we decided to drop the number of proximities within 800m as 500m and 1km give a clearer feature separation. Other features show no signs of high correlations and, thus, were not dropped.

3. Full Results from Analysis

The dataset was carefully curated to include only relevant features for predicting HDB resale prices, with the response variable set as *log_adjusted_price* to ensure a more normally distributed target for regression modelling (see Section 2.2). After applying ordinal and one-hot encoding to categorical variables (see Section 2.3), we developed several predictive models to estimate HDB resale prices. To assess their effectiveness, we utilized performance metrics such as **Mean Absolute Error (MAE) and Root Mean Square Error (RMSE),** providing insight into the accuracy of the predictions.

3.1 Linear Regression

Linear Regression is chosen for HDB resale price prediction because the model is **easy to interpret, providing clear insights** into how each factor influences resale value. The model is also **computationally efficient** and serves as a good baseline for comparison with more complex models. Building on Linear Regression, techniques like LASSO, Ridge, and Elastic Net can be applied to improve the model by handling multicollinearity and enhancing generalization.

3.1.1. Base Model

The base model for predicting the *log_adjusted_price* is a simple linear regression model using all encoded and scaled features. According to Table 1, the model's RMSE and MAE indicate that, on average, predicted prices deviate from actual values by $59,031.98 and $45,452.66, respectively. The fact that RMSE is higher than MAE suggests the model is affected by some large prediction errors that need to be addressed.

Model	RMSE	MAE
Base Linear Regression	59,031.98	45,452.66

Table 1: Linear Regression Base Model Result

3.1.2. Regularization

The base linear regression model may tend to overfit, especially when there are many features or multicollinearity introduced by the one-hot encoding. Regularization techniques would help mitigate this by penalizing large coefficients, reducing model complexity, and improving generalization.

The first regularization technique used is **LASSO (L1)**, which adds an absolute penalty to the regression coefficients. The purpose of LASSO is that it shrinks some of the coefficients to zero, **effectively performing feature selection**. This technique helps to **remove irrelevant features** and reduce the model's complexity. However, looking at the results in Table 2, LASSO regularization did worse than the base model of Linear Regression, suggesting that some of the important predictors are being set to zero.

Model	RMSE	MAE
LASSO	65,268.38	49,619.56

Table 2: Linear Regression with LASSO Result

The second regularization technique used is **Ridge (L2),** which differs from LASSO by shrinking coefficients without reducing them to zero. Ridge is useful for **addressing multicollinearity and enhancing model stability and accuracy**. As shown in Table 3, the results are very similar to those of the base model, indicating that multicollinearity is not a major concern and that all features, even those with smaller weights, contribute meaningfully to predicting resale prices.

Model	RMSE	MAE
Ridge	59,032.18	45,452.90

Table 3: Linear Regression with Ridge Result

The last regularization technique used is **Elastic Net**, which combines both LASSO and Ridge regularization by reducing some coefficients to zero while shrinking others. As shown in Table 4, its **performance lies between the results of LASSO and Ridge**, which is expected. This is likely because LASSO alone may have eliminated important predictors, making it less suitable for this model.

Model	RMSE	MAE
Elastic Net	62,107.88	47,257.61

Table 4: Linear Regression with Elastic Net Result

3.1.3. Summary of Base Model

In summary, linear regression serves as a strong baseline for predicting HDB resale prices due to its simplicity and low computational cost. After applying regularization techniques to enhance generalization, **Ridge regression** indicates that all features contribute to predicting resale prices, even if their individual impact is minimal.

3.1.4. Interaction Term

To enhance the model, **interaction terms** were introduced to capture non-linear relationships that basic linear regression cannot effectively model. Specifically, interactions between *town/region* and features such as distance, *storey_range*, and *remaining_lease* were created to reflect area-specific effects. Ridge regression, which previously showed the best performance, was used to train the model and help prevent overfitting. As shown in Table 5, the inclusion of interaction terms improved the model by reducing the MAE by approximately $5,000. Therefore, this model represents the most effective linear regression model.

Model	RMSE	MAE
Ridge	52,976.99	40,583.27

Table 5: Linear Regression with Ridge on Interaction Term Result

3.1.5. Prediction Error by Town

The interaction model was used to generate predictions on the test dataset. As shown in Figure C1, the model performs better in heartland areas compared to prime and high-value locations,

where prediction errors are more significant. This is likely due to the higher demand and limited supply in prime areas, where prices are influenced more by prestige than by measurable factors like amenities or *flat_type*. For instance, a recent HDB resale transaction in Bishan reached $1.568 million, which is a significant outlier compared to the surrounding average of approximately $1 million (EdgeProp, 2024). In contrast, prices in heartland areas are more strongly influenced by features like *floor_area_sqm* and *flat_type*. However, individual buyer preferences still contribute to pricing variability, underscoring the need for non-linear models to better capture these complex relationships.

3.2 Support Vector Regressor (SVR)

SVR was selected as one of the predictive models to use due to its ability to **capture non-linear relationships** between the independent and target variables. Specifically, it helps identify complex relationships between features such as proximities to amenities or remaining lease against resale price. To model these non-linear relationships, SVR employs kernel functions such as RBF and polynomial kernels, which simulate the **transformation of the data into higher-dimensional spaces**. It is important to note that due to computational constraints, cross-validation was not performed during the model training process.

3.2.1 Base Model

Before training the models, all numerical features were scaled to ensure that larger features do not disproportionately influence the model. For better comparison purposes, a base model was first trained using the linear, RBF, and polynomial kernels, with the same hyperparameters C = 1 and epsilon = 0.1. As shown in Table 6, the linear kernel performed significantly worse than the RBF and polynomial kernels. This finding suggests a non-linear relationship between the independent and target variables. Therefore, further improvements placed more emphasis on the non-linear kernels, specifically the polynomial and RBF kernels.

Model	RMSE	MAE
Base Linear	91,897.27	47,220.32
Base RBF	41,622.16	30,827.07
Base Polynomial Degree 2	46,360.49	34,350.42

Table 6: SVR Base Model Results

3.2.2 Permutation Importance

Since SVR does not provide feature importance scores, an alternative method called permutation importances was used. This approach randomly shuffles the values of each feature and observes how the model performs to find the importances. By identifying and utilizing only important features, it helps to reduce model complexity, prevent overfitting, and potentially enhance model performance. As such, to assess performance based on feature relevance, models were trained using the top 10, 30, and 50 most important features, as identified in Figure C2, and evaluated using both polynomial and RBF kernels for comparison.

Based on the results in Table 7, it is evident that the **RBF kernel consistently outperformed the polynomial kernel**. Further analysis revealed that the performance difference was due to outliers

or noise in the dataset. The RBF kernel transformation diminished the impact of the outliers, while the polynomial kernel amplified them, leading to less accurate results. Hence, subsequent efforts placed more emphasis on improving the model with the RBF kernel.

Furthermore, as performance remained similar between the models using the top 50 and all features, only the top 50 features were retained for subsequent models. This reduction in the number of features helped lower model complexity and reduce the risk of overfitting, thereby improving generalization. While other feature selection techniques such as Recursive Feature Elimination (RFE) were considered to select the optimal number of features, they were ultimately not pursued due to computational limitations.

	Polynomial Degree 2		**RBF**	
No of Features	**RMSE**	**MAE**	**RMSE**	**MAE**
10	95,808.15	63,226.29	81,142.98	53,341.96
30	71,193.47	44,362.88	48,061.08	34,830.03
50	49,114.42	35,465.02	41,686.01	30,918.16

Table 7: Polynomial vs RBF Results (Top 10, 30, 50 Features)

3.2.3 Grid Search

To improve the performance of the model using the RBF kernel, the grid search algorithm was used to find the best hyperparameters. The hyperparameters tuned included *C*, to regulate the model's flexibility and error, and *epsilon* to set the tolerance for the margin of error allowed for predictions. The grid search evaluated combinations of C = [0.1, 1, 10] and epsilon = [0.01, 0.1, 0.5], and achieved the best performance when C = 1 and epsilon = 0.01. With this new set of hyperparameters, the errors were reduced by around 2,500, as shown in Table 8.

Best Hyperparameters	**RMSE**	**MAE**
RBF: C=1, epsilon=0.01	39,155.10	28,361.84

Table 8: SVR Results for Grid Search

3.2.4 Bagging Regressor

To prevent overfitting, the **bagging technique** was also explored. This approach involves averaging the predictions across several SVR models trained on bootstrapped samples to reduce variance. However, as shown in Table 9, the bagged model resulted in slightly worse RMSE and MAE scores. This is likely because the current SVR model already exhibits low variance and strong stability. Therefore, the group decided not to use the bagged model.

Bagging Model	**RMSE**	**MAE**
5 estimators used	39,277.31	28,454.55

Table 9: SVR Results for Bagging Model

3.2.5 Evaluation on Test Set

Based on the RMSE and MAEs of all the models evaluated, the best-performing SVR model is the one described in section 3.2.3, which uses 50 features with the RBF kernel, C = 1, and epsilon = 0.01. For a fair assessment of the model, the model is evaluated against the test dataset. As shown in Table 10, the RMSE and MAE of the validation and test sets were similar, indicating that the model did not overfit and **generalized well** against unseen data.

Final Model	**RMSE**	**MAE**
Validation Set	39,155.10	28,361.84
Test Set	39,417.81	28,314.07

Table 10: SVR Best Model Evaluation on Test Set

However, from the residual plot shown in Figure C3, it is evident that the residuals have a very slight funnel shape, suggesting that **more expensive resale flats tend to have slightly higher errors**. The higher errors are likely due to noise in the dataset and the smaller number of transactions for higher-priced units. However, the team has decided that this error is acceptable as the model still performs reasonably well, especially for areas with higher volumes of data.

Further analysis of the errors for each town also revealed that areas with lower transaction volumes also tend to have higher errors (Figure C4). While there were attempts to resolve this problem through techniques like frequency encoding, the prediction errors for each town, and the overall RMSE (39762.41) and MAE (28607.36) remained roughly the same. Therefore, the variables created through frequency encoding were dropped and not included in the model.

3.3 Neural Network

To accurately predict HDB resale prices, a neural network model was implemented due to its ability to **capture complex, nonlinear relationships among multiple interacting factors,** such as *town*, *flat_type*, *storey_range*, and *remaining_lease*. Unlike traditional regression models that assume fixed linear dependencies, neural networks adaptively learn intricate patterns within the data, making them highly effective for real estate pricing, where feature interactions are difficult to model explicitly. Additionally, neural networks excel at **handling high-dimensional data**, including both numerical and categorical features, without requiring extensive manual feature engineering.

3.3.1. Base Model

The initial model was built using **5-fold cross-validation** to evaluate its performance and ensure generalizability. It consists of an input layer for the preprocessed features, followed by three hidden layers with 256, 128, and 64 neurons, respectively, each utilizing **ReLU activation** functions to capture complex nonlinear relationships. To enhance model robustness and mitigate overfitting, batch normalization and dropout layers (with rates of 0.3 and 0.2, respectively) were incorporated. The output layer features a single neuron with a linear activation function, predicting the *log_adjusted_price* of HDB resale units. Compiled with the **Adam optimizer**, Table 11 and Figure C5 shows the average results of the model across the five folds.

Model	RMSE	MAE
Base Model	45,531.74 ± 895.68	31,719.75 ± 642.15

Table 11: Neural Network Base Model Results

The results indicate that the base model performed reasonably well, showing its capability to model the pricing complexities of HDB units. However, while the performance is commendable, there remains room for improvement.

3.3.2. Improved Model

To improve the model's predictive performance, the architecture was expanded to include **larger hidden layers** consisting of 512, 256, and 128 neurons. This deeper structure enables the model to better learn complex, nonlinear patterns influencing HDB resale prices. Moreover, advanced training callbacks were incorporated, such as **EarlyStopping** to prevent overfitting by halting training when validation loss stops decreasing, **ReduceLROnPlateau** to automatically fine-tune the learning rate, and **ModelCheckpoint** to save the best-performing model based on validation loss. Together, these strategies help the model learn efficiently while adapting to the dataset's intricacies.

After training the improved model using the same 5-fold cross-validation technique, performance metrics were evaluated to assess the impact of these enhancements. The results in Table 12 and Figure C6 indicate a noticeable improvement in predictive accuracy, with average metrics showing a reduction in both RMSE and MAE compared to the base model.

Model	RMSE	MAE
Improved Model	40,715.70 ± 622.71	29,325.84 ± 409.45

Table 12: Improved Neural Network Model Results

3.3.3. Feature Importance

To better understand the factors influencing HDB resale prices, SHapley Additive exPlanations (SHAP) values were used to analyze feature importance. Figure C7 highlights that *floor_area_sqm* is the most influential factor, followed by *remaining_lease*. This aligns with real-world expectations, as **larger flats and those with longer lease durations typically command higher prices**.

To assess the effect of feature selection, we retrained the neural network model using only the top 20 features identified by SHAP. While the performance remains reasonable in Table 13 and Figure C8, it does not outperform the Improved Model. This suggests that while the top features identified via SHAP are highly relevant, **excluding the remaining features may still lead to a loss of valuable information contributing to predictive accuracy.**

Model	RMSE	MAE
Feature Importance Model	51,275.84 ± 2768.36	35,689.60 ± 1760.85

Table 13: Neural Network with Feature Importance Results

3.3.4. Addressing Data Scarcity Across Towns

To address our core problem statement regarding valuation fairness and accuracy, especially in towns with limited transaction data, we extended our error analysis to examine model performance across different towns. As shown in Figure C9, the model performed well in high-volume towns like Sengkang and Jurong West but struggled in data-scarce towns such as Bukit Timah and Central Area.

In an attempt to mitigate the underperformance in data-scarce towns, we experimented with a **town-weighted loss function**, giving higher importance to underrepresented towns during training. We also introduced L2 regularization to reduce overfitting by penalizing large weight values, with the goal of improving generalization. However, the results did not yield the desired improvements (Table 14).

Model	RMSE	MAE
Town-weighted Model	80,071.22 ± 17815.86	59,924.09 ± 13091.35
Town-weighted Model (with Regularization)	78,719.58 ± 4204.70	56,352.65 ± 3262.16

Table 14: Neural Network Town-Weighted Model With & Without Regularization Results

Although the results with regularization are better than without, these values represent a significant deterioration compared to our baseline and best-performing models. One plausible reason is that **overweighting towns with limited data introduces noise**, making the model overly sensitive to small, potentially unrepresentative patterns. This not only reduces the model's ability to generalize but also compromises its performance in well-represented areas. This outcome reveals a key limitation that simply adjusting model weights cannot compensate for the lack of diverse and sufficient training data in underrepresented towns. In areas with very few historical transactions, the model lacks enough examples to learn accurate relationships between features and prices.

To improve performance in the long run, **collecting more transaction data in these sparse regions** is essential. This would provide a richer, more balanced dataset and enable fairer, more reliable valuations across all towns. In the meantime, as a practical workaround, we address this limitation by **leveraging neighborhood-level amenities**. By incorporating features such as proximity to schools, shopping centers, and public transport, the model can use external contextual signals to make more informed predictions.

Model	RMSE	MAE
Without Amenities Model	44,188.19 ± 2,142.50	31,821.02 ± 1,534.93
Best Model (with Amenities)	40,715.70 ± 622.71	29,325.84 ± 409.45

Table 15: Neural Network Model With and Without Amenities Results

The results in Table 15 clearly show that the use of amenity-based features leads to the most accurate and consistent model, particularly benefiting towns with fewer transactions. Given these findings, the **Best Model (with Amenities)**, which is also known as the Improved Mode, was chosen as the final model. While increasing the number of neurons beyond 512 (e.g., to 1024) could have further expanded the model's capacity to learn complex patterns, it was not pursued. A larger model would significantly **increase computational cost and the risk of overfitting**, especially given the dataset's size and characteristics. Instead, a balanced architecture was prioritized to optimize performance without excessive parameterization.

3.3.5. Evaluation on Test Set

To assess the model's generalization capability, final testing was conducted on the test dataset using the best model. The final **test results** closely aligned with the validation performance (Table 16), with the model achieving:

Model	RMSE	MAE
Test Results on Final Model	41,732.70	29,406.96

Table 16: Final Neural Network Model Test Results

The relatively low MAE suggests that, on average, the model's predictions deviate from actual prices by approximately $29,407, which is an acceptable range given the variability of real estate prices. The residual plot (Figure C10) provides further insight into model performance. Residuals are mostly centered around zero, suggesting no strong systematic bias. However, a slight funnel shape is observed, with higher variance at the upper price range. This trend is most likely due to the **limited availability of higher-priced data in the training set**, leading to reduced model accuracy for premium properties. Despite these limitations, the model remains a reliable tool for price estimation, with room for refinement through expanding the dataset with more high-value transactions.

3.4 Random Forest

Random Forest is an ensemble of decision trees and was used for HDB resale data, as it **allows complex non-linear relationships between features** and **handles mixed data feature types** without requiring heavy feature engineering. Random Forest is robust to overfitting, especially with noisy data, and performs well even when the dataset has many features. Additionally, Random Forest provides insights into feature importance, enabling a better interpretation of which factors most influence resale prices. Its ensemble approach reduces variance and improves generalization, making it a reliable and accurate method for predicting HDB resale prices.

3.4.1 Base Model

The base model was built using the default parameters fitted using the training data and validated using the validation data. This will be used to extract the feature importance to fine-tune the model.

Model	RMSE	MAE
Base Model	36823.20	26219.06

Table 17: Random Forest Base Model Results

3.4.2. Feature Importance

Feature importance can be derived from the Random Forest, which uses the metric called Gini Importance. It measures the total reduction of the Gini impurity of the dataset when a particular feature is used for splitting. To use the metric, the higher the Gini Importance, the higher the importance of the feature. Figure C11 shows the top 10 most important features according to Gini Importance; ***floor_area_sqm*** has the highest importance by a large margin, followed by *max_floor_lvl* and *remaining_lease*.

Subsequently, we plotted the partial dependence plot to see how each feature would affect the prediction of the HDB resale price. Figure C12 shows the partial dependence of the top 6 most important features. Based on the plots, it is evident that ***floor_area_sqm*** **emerges as the most influential factor** with a positive relationship, where larger floor areas result in higher resale prices. Similarly, this trend can be observed for *max_floor_lvl* and *remaining_lease*. From the *1km_hawkers* plot, we can also conclude that there is a premium price for having a hawker within 1 km's proximity. Conversely, variables like *nearest_MRT_Distance_km* and *Nearest_JC_km* exhibited minimal impact on resale price, suggesting that beyond a certain accessibility threshold, proximity to MRT stations and Junior Colleges has a limited effect.

To further refine feature selection, three models using different numbers of features, specifically, the top 5, 10, and 95% features, were used for hyperparameter tuning. This comparison aims to identify the truly impactful variables while removing irrelevant or noisy features, ultimately improving model performance and interpretability.

3.4.3. Grid Search with Feature Importance

Hyperparameter tuning was performed using Grid Search, exploring combinations such as n_estimators [100, 200], max_depth [None, 10, 20], min_samples_split [2, 5], and min_samples_leaf [1, 2]. This process was applied to all three models, and the one achieving the best performance was selected for further use.

Model	RMSE	MAE
Top 5 Features	51063.65	35461.36
Top 10 Features	38093.92	26739.56
95% Importance	37232.82	26309.43

Table 18: Random Forest with Grid Search and Feature Importance Results

3.4.4. Grid Search without Feature Importance

Since the models did not show improvement over the baseline model after applying Grid Search, it suggests that all features could be contributing meaningfully to the prediction of HDB resale

prices. Therefore, the baseline model was subjected to Grid Search to determine its optimal hyperparameters.

Best Hyperparameters	RMSE	MAE
min_samples_split=5, Min_samples_leaf:1, n_estimators=200, max_depth: None	36515.98	25972.46

Table 19: Random Forest with Grid Search and without Feature Importance Results

3.4.5. Evaluation on Test Set

The model without removing any features performed the best. Therefore, to assess the model's generalization capability, final testing was conducted on the test dataset using the best model.

Best Hyperparameters	RMSE	MAE
min_samples_split=5, Min_samples_leaf:1, n_estimators=200, max_depth: None	36544.37	25919.59

Table 20: Random Forest Best Model Evaluation

The MAE suggests that, on average, the model's prediction deviates from the actual prices by approximately $25,919.59. A closer analysis of the residual plot in Figure C13 shows the points mainly centring around zero with a relatively constant variance; there is no obvious sign of heteroskedasticity. However, there are some outliers, but there are no curves and trends, which indicates the random forest model is **capturing non-linearities well**.

To understand the model better, the first three decision trees out of 200 n_estimators are analysed, shown in Figures C14, C15 & C16. The decision trees shown are **binary trees** that have strong repetitions of **key splits from *floor_area_sqm, max_floor_lvl, and remaining_lease*** at top levels and are consistently dominant, which indicates those features have **high predictive power.** It is also observed that it will split with lesser important features observed in the initial interpretation of feature importance, in Figure C11, such as schools and hawkers before splitting by the key features again then finally arriving at the value which indicates these **weak features are signals** that contribute to predicting the final resale HDB price.

3.5 Ensemble Modelling

Ensemble modeling **combines the outputs of several individual models** to create a stronger and more accurate overall model. By utilizing the unique strengths of each model, this technique often outperforms any single model on its own. It generally works by averaging the predictions from the models, which helps minimize errors and reduce bias, resulting in improved prediction accuracy.

Four ensemble models were developed by adjusting the number of models used and the averaging method **to minimize MAE**. The first two ensemble models incorporate all four models (Random Forest, SVR, Neural Network, and Linear Regression), using flat averaging and weighted averaging, respectively. While the third and fourth models use only the top two performers (Random Forest & SVR), with predictions based on the flat and weighted average

methods, respectively. As shown in Table 21, the weighted average of all models produced the best results. Although the weights are rounded to 4 significant figures, the inclusion of linear regression in All Models (Weighted Average) indicates it was included in the process but had minimal influence on the final prediction.

Ensemble Model	Weights	RMSE	MAE
All Models (Flat Average)	All: 0.25	39609.1168	27928.4275
Random Forest + SVR (Flat Average)	All: 0.5	36051.1918	25732.6363
All Models (Weighted Average)	RF: 0.6878 SVR: 0.1862 NN: 0.1261 LG: 0.0000000007	35645.0116	25384.5024
Random Forest + SVR (Weighted Average)	RF: 0.7170 SVR: 0.2830	35771.8482	25453.6145

Table 21: Ensemble Model Results

4. Discussion and Summary

4.1 Final Model Evaluation

In this project, we have explored multiple models, including Linear Regression, Support Vector Regression, Neural Network, Random Forest, and Ensemble Modelling, to predict the target variable: resale price. The models with the best performance concerning the test set in order are:

Position	Model	RMSE	MAE
1	Ensemble	35645.01	25384.50
2	Random Forest	36544.37	25919.59
3	SVR	39,417.81	28,314.07
4	Neural Network	41,732.70	29,406.96
5	Linear Regression	52,072.81	40,644.30

Table 22: Best Results for Each Model Type

Since the “best” model was selected based on its performance on the test set, relying solely on that for final evaluation could introduce bias. To properly assess the model’s generalisation ability, we took an additional step by retrieving fresh resale data from the official website. After processing the new dataset and generating predictions, we compared the predicted values against the actual resale prices, as shown in Table 23. The **results reveal only a minimal difference** in

RMSE and MAE between the test and new datasets, indicating that the ensemble model is not overfitted and performs consistently well on unseen data.

Model	RMSE	MAE
Test Set	35645.01	25384.50
New Data	35508.67	25731.09

Table 23: Final Model Evaluation Against Latest Retrieved Data

Upon further analysis, the top two features are *remaining_lease* and *floor_area_sqm*. Thus, to evaluate the models further, the MAE is plotted against both of these features with the test data and new data.

Firstly, the *remaining_lease* is grouped into 10-year intervals to facilitate analysis. Figure C17 shows the MAE across different lease bins for the test dataset, while Figure C18 presents the MAE for the new dataset. In both datasets, the final model displays inconsistent performance across the various lease bins. On the new data, the model performs best for flats with 50–60 years of remaining lease. For the test data, the lowest MAE is observed in the 40–50 and 90–100 year bins. However, the **model consistently struggles with mid-aged leases between 60–90 years, especially in the 70–80 year range, where errors are highest**. One possible reason for this difficulty is that flats in this lease range may exhibit greater variation in other influencing factors such as renovation status, location-specific demand, or pricing anomalies, which the model may not fully capture. Additionally, there is no clear increasing or decreasing trend in MAE, indicating that the use of the RPI helped mitigate time-series effects.

Next, the *floor_area_sqm* feature is binned into ranges: <50 sqm, followed by 50 sqm intervals, and >350 sqm, to simplify analysis. Figure C19 displays the MAE across floor area bins for the test data, while Figure C20 presents the MAE for the new data. In the test set, MAE increases with floor area, indicating the **model struggles more with predicting prices of larger flats.** Notably, the 200–250, 250–300, and >300 sqm bins show significantly higher MAEs compared to the overall average. This is likely due to outliers, as each of these bins contains only one data point, as illustrated in Figure C21. Similarly, the new data shows a rising trend in MAE as floor area increases, reinforcing the model's difficulty with larger flats. Bins above 200 sqm in the new data have zero MAE due to a lack of records. Overall, the increased prediction error for larger flats is primarily caused by limited data availability in those ranges.

4.2 Improvements

While the final model's performance is reasonably good, there are still several techniques and models that could have been tested to improve the model's performance further.

More data, especially for lower transacted areas, could be obtained to expose the model to a wider range of cases and improve its generalisation capabilities. However, real-world data depends on market activity and the frequency of such transactions, which is beyond our control. We could use techniques like oversampling to expose the model to higher-priced resale flats.

However, this technique may not accurately reflect real-world market behaviour, potentially leading to an unrealistic model.

More features that may be useful for predicting the resale price can also be **engineered**. Some examples include the number of MRT lines the closest MRT is on, school names as some parents may be interested in sending their child to certain schools, and policy-related columns such as the number of days it has been since the last cooling measure. However, doing so would increase the dimensionality and model complexity, taking up more computational resources, and may even introduce multicollinearity. Therefore, it is best to do this with caution to ensure that the added complexity does not outweigh the insights brought by the newly created features.

Feature selection techniques like RFE can also be tested to find the optimal number of features for each model. Additionally, more hyperparameters can be included in our GridSearch to explore a wider range of models within each model type. For instance, degree 3 can be used for SVR with the polynomial kernel. Other models, such as KNN, can also be tested. However, do note that these techniques are computationally extensive.

An **alternative target variable** could also be used for the prediction task. For instance, instead of predicting the adjusted price, we could predict the adjusted price per square meter. Using this variable allows for better and easier comparison across flats of different sizes. Additionally, doing so reduces the influence of floor area, allowing the model to focus more on relevant features. This adjustment could be particularly useful as our analysis showed that floor area consistently ranked top for the most important features. Doing so would allow the model to focus on other features, such as the remaining lease, region, and the newly engineered features.

Pruning can also be done to improve neural network generalisation capabilities. Unlike feature importances, which remove the feature entirely after selection, pruning removes only certain weights or neurons within the network. Doing so allows the model to retain all features while only using necessary weights when required. As a result, the model becomes more efficient and less prone to overfitting, potentially improving its ability to generalise to unseen data.

4.3 Summary

In summary, this project aims to predict HDB resale prices, emphasising more on analysis and lower transacted areas. Based on our results, the overall better performances in the non-linear models suggest that the relationship between the independent and target variables is non-linear. Among all the models trained, the ensemble model performed the best, most likely due to its ability to leverage the strengths of each model. Throughout the project, it was also observed that areas with lower transaction volumes tend to have higher prediction errors. While various techniques were used in an attempt to resolve the underlying issues and improve the performance, they did not help or were insufficient. However, with the techniques and possible changes mentioned in section 4.2, the model's performance may improve, or more insights could be gained. Regardless, our team believes that our final model can perform reasonably well, especially for areas that have high transaction volumes.

5. References

Channel News Asia. (2024, February 5). *Vacated Tanglin Halt flats to be refurbished as temporary homes for families*. CNA. Retrieved March 2, 2025, from https://www.channelnewsasia.com/singapore/tanglin-halt-flats-hdb-sers-temporary-pphs-family-housing-4099456

DollarBack Mortgage. (n.d.). *Cash Over Valuation (COV) HDB 2024: 3 Ways To Avoid It!* DollarBack Mortgage. https://dollarbackmortgage.com/blog/cash-over-valuation-hdb/

EdgeProp. (2024, 07 09). *Most expensive HDB in Bishan sold for $1.568M, 4th highest in Singapore*. EdgeProp. https://www.edgeprop.sg/property-news/most-expensive-hdb-bishan-sold-1568m-4th-most-expensive-singapore

Grrrnt. (2021). *Predicting Future Prices of HDB Resale Flats in Singapore*. Github. https://github.com/grrrrnt/hdb-resale-price-prediction

Heng, J. (2016, August 3). *Eight West Coast Road blocks up for Sers, replacement flats in Clementi and West Coast*. The Straits Times. Retrieved March 2, 2025, from https://www.straitstimes.com/singapore/housing/eight-west-coast-road-blocks-up-for-sers-replacement-flats-in-clementi-and-west

HomeGen. (2023, 09 22). *Lease Decay and Its Impact on HDB Value: A Comprehensive Analysis*. HOMEGEN. https://www.homegen.sg/post/lease-decay-and-its-impact-on-hdb-value-a-comprehensive-analysis

Housing & Development Board. (n.d.). *Resale Statistics*. Housing & Development Board. https://www.hdb.gov.sg/residential/selling-a-flat/overview/resale-statistics

Lok, B. H. (2025, Jan 2). *Singapore public housing resale prices rise 9.6% in 2024*. Reuters. https://www.reuters.com/markets/asia/singapore-public-housing-resale-prices-rise-96-2024-2025-01-02/

Loy, E. (2023). *Project 2 - Singapore Housing Data and Kaggle Challenge*. Github. https://github.com/enochloy/hdb-resale-predictor

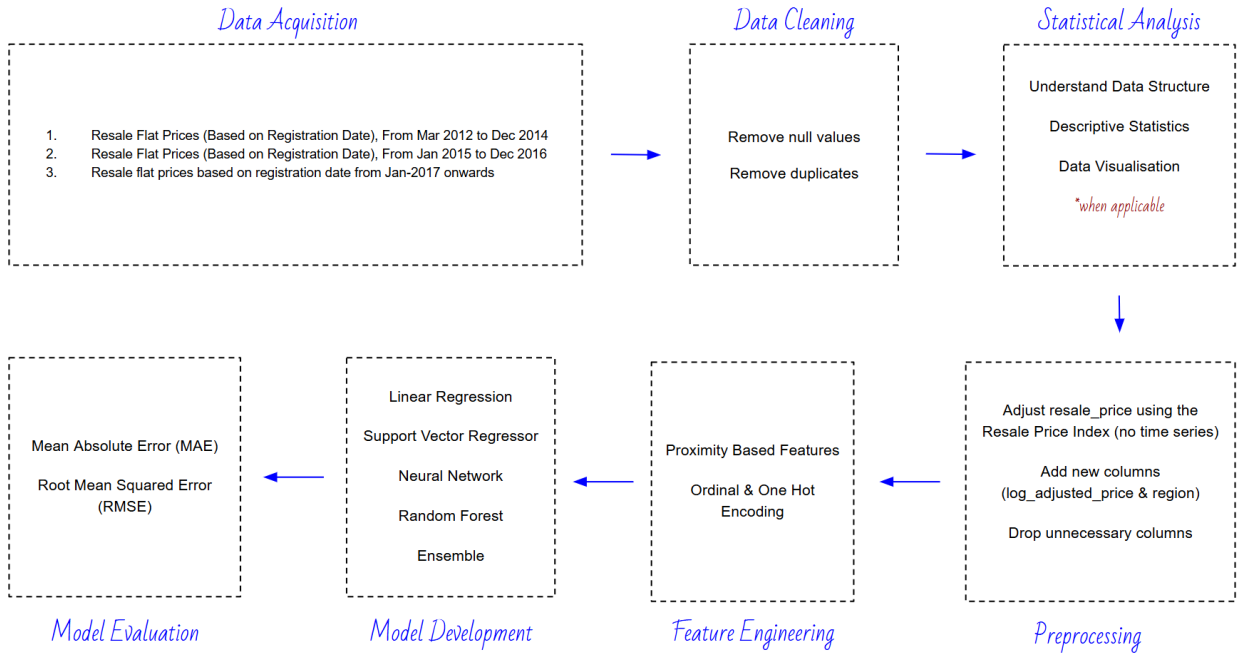
6. Appendix

Appendix A: Overview

Table A1: Raw Data Columns & Definitions

S/N	Column Name	Data Type	Description
1	month	Datetime (YYYY-MM)	Month of sale
2	town	Categorical Text	Designated residential area
3	flat_type	Categorical Text	Classification of units by room size
4	block	Text	HDB building containing multiple flats
5	street_name	Text	Road name where the flat is located
6	storey_range	Categorical Text	Approximate floor range of the sold unit
7	floor_area_sqm	Numeric	Interior space of the unit (square meters)
8	flat_model	Categorical Text	Classification based on construction generation
9	lease_commence_date	Datetime (YYYY)	Start year of the lease agreement
10	resale_price	Numeric	Transaction price of the flat
11	residential	Binary Text	Indicates if the building is primarily for residential use
12	commercial	Binary Text	Indicates if commercial facilities are present within the building
13	miscellaneous	Binary Text	Includes admin offices, childcare centers, education centers, Residents' Committee centers
14	multistorey_carpark	Binary Text	Indicates if a multi-story car park is attached to the building
15	precienct_pavilion	Binary Text	Indicates the presence of a pavilion for community gatherings
16	bldg_contract_town	Text	Legend for town abbreviations
17	total_dwelling_units	Numeric	Total number of residential units within the building

Figure A1: Project Pipeline



Appendix B: Exploratory Data Analysis

Figure B1: Null values

	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	resale_price	remaining_lease	max_floor_lvl	residential	commercial	market_hawker	miscellaneous	multistorey_carpark	precinct_pavilion	bldg_contract_town
313	2014-01-01	CLEMENTI	3 ROOM	513	WEST COAST RD	01 TO 03	68.0	New Generation	322000.0	65	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
318	2014-01-01	CLEMENTI	3 ROOM	516	WEST COAST RD	01 TO 03	68.0	New Generation	335000.0	72	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
319	2014-01-01	CLEMENTI	3 ROOM	514	WEST COAST RD	04 TO 06	68.0	New Generation	339000.0	65	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
324	2014-01-01	CLEMENTI	3 ROOM	513	WEST COAST RD	07 TO 09	68.0	New Generation	350000.0	65	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
670	2014-01-01	QUEENSTOWN	2 ROOM	27	TANGLIN HALT RD	07 TO 09	46.0	Standard	265000.0	58	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
42188	2016-06-01	CLEMENTI	3 ROOM	518	WEST COAST RD	01 TO 03	68.0	New Generation	260000.0	63	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
42190	2016-06-01	CLEMENTI	3 ROOM	514	WEST COAST RD	04 TO 06	68.0	New Generation	278000.0	63	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
42197	2016-06-01	CLEMENTI	3 ROOM	513	WEST COAST RD	13 TO 15	68.0	New Generation	290000.0	63	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
43964	2016-07-01	CLEMENTI	3 ROOM	513	WEST COAST RD	04 TO 06	68.0	New Generation	295000.0	63	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
45638	2016-08-01	CLEMENTI	3 ROOM	518	WEST COAST RD	07 TO 09	68.0	New Generation	269000.0	63	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

91 rows × 19 columns

Showing null values for several key columns in the raw data

Figure B2: Duplicates

```
df.duplicated().sum()
✓ 0.6s                                                    Python
624
```

Showing the number of duplicates in the raw data

Figure B3: Plots for Resale Price

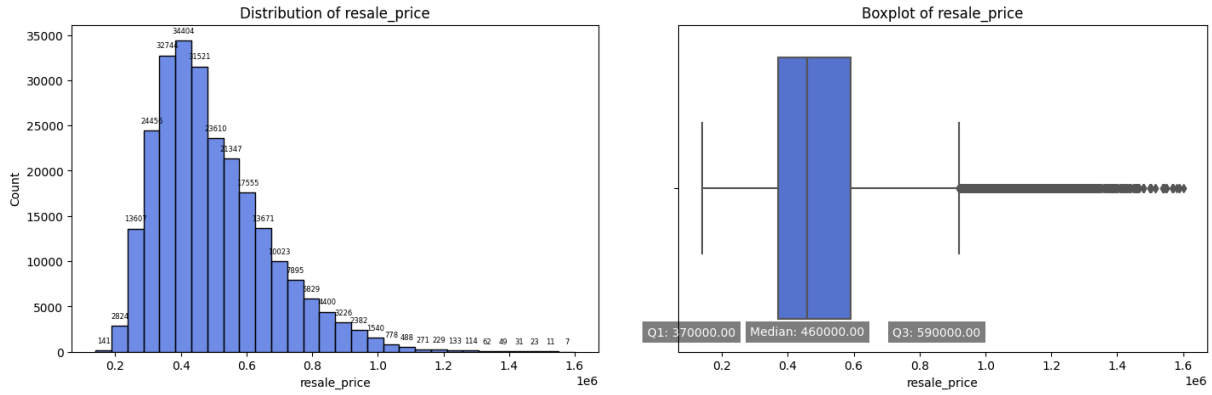


Figure B4: Average Resale Price by Year

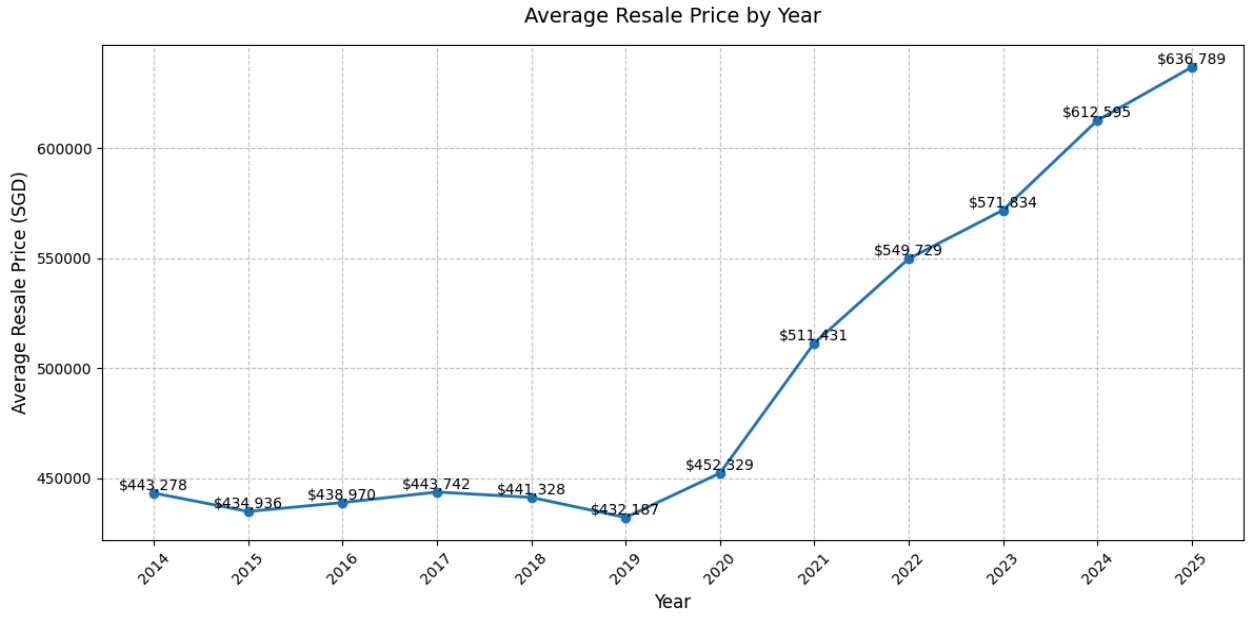


Figure B5: Original Adjusted Price Distribution vs Log-Transformed Adjusted Price Distribution

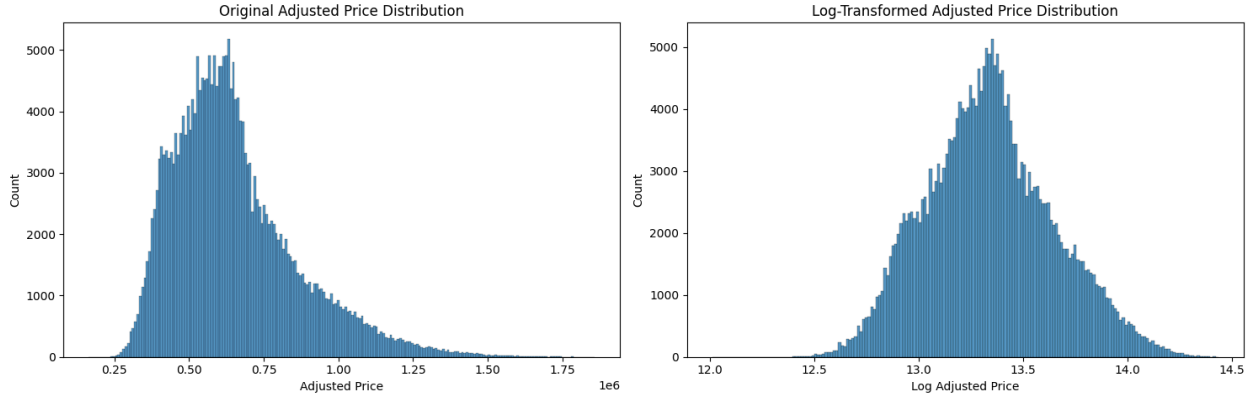
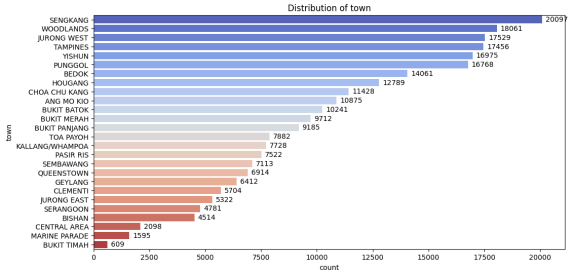


Figure B6: Plots for Town



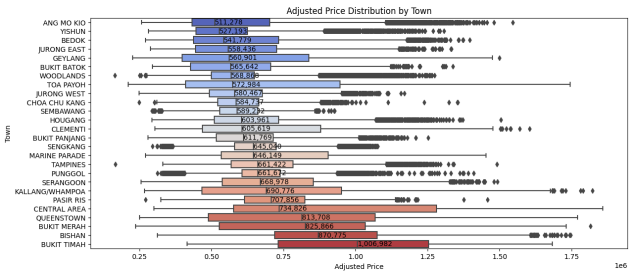
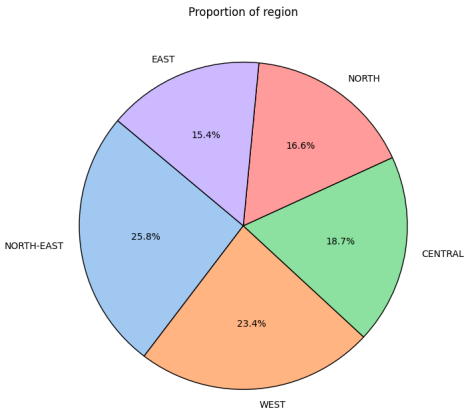


Figure B7: Plots for Region



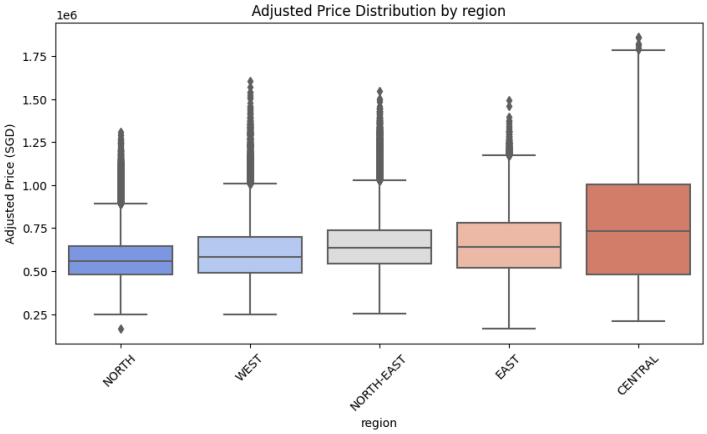
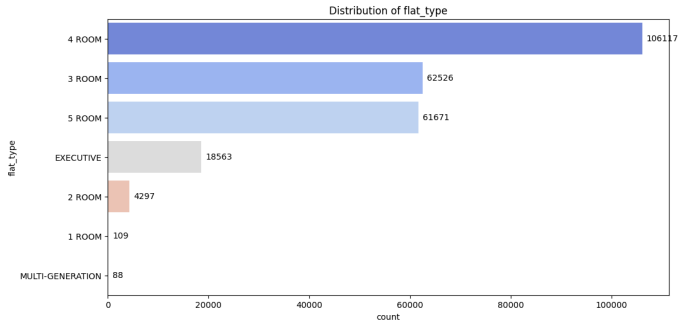


Figure B8: Plots for Flat Type



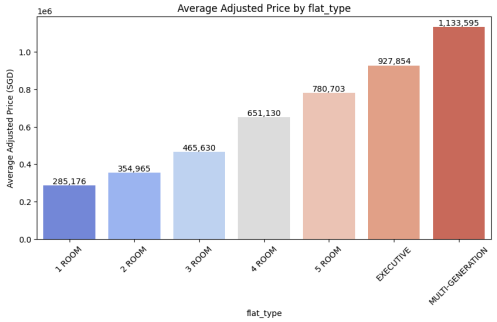
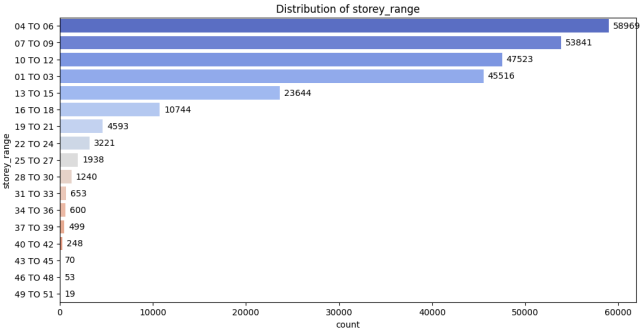


Figure B9: Plots for Storey Range



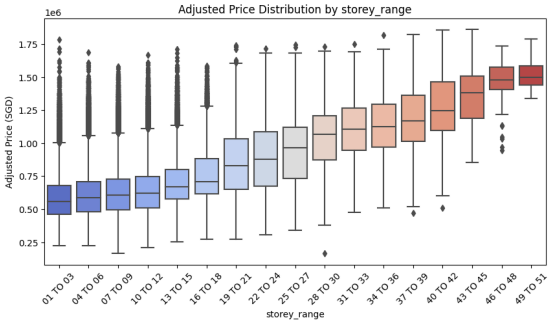


Figure B10: Scatterplot of Adjusted Price against Floor Area by Region

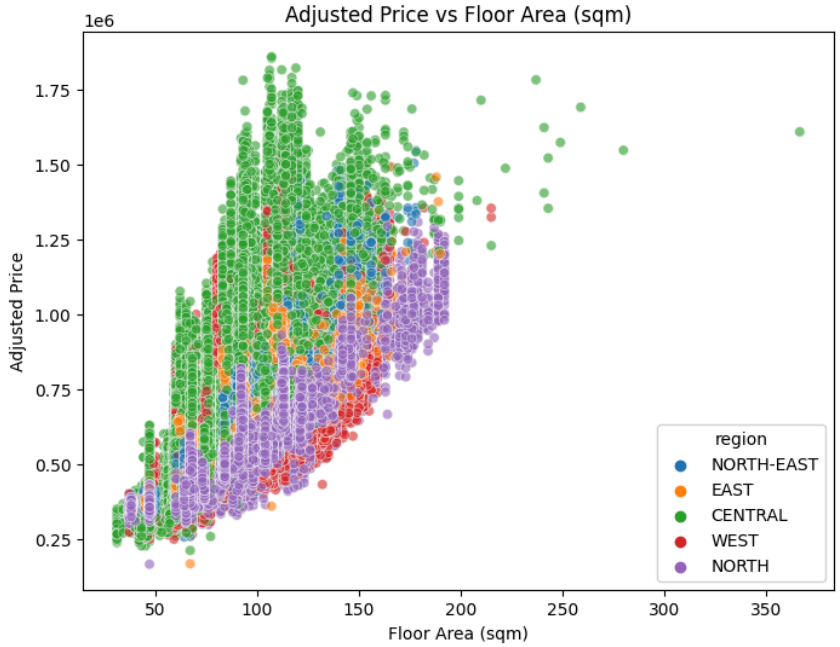


Figure B11: Plots for Flat Model

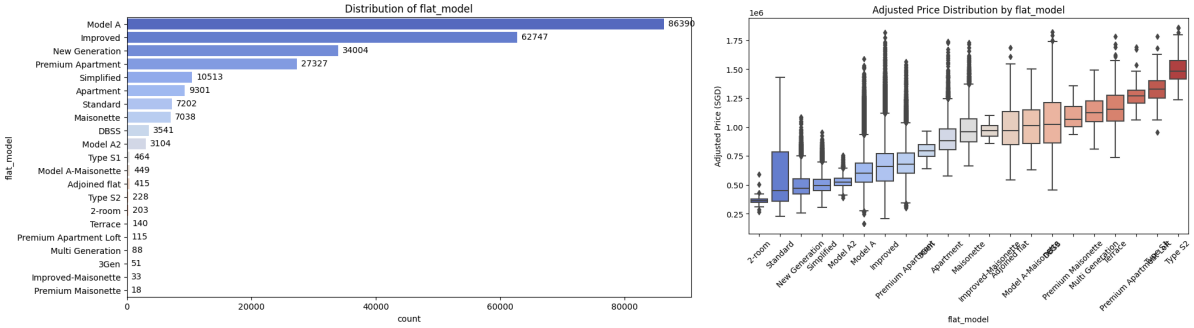


Figure B12: Scatterplot of Remaining Lease against Adjusted Price

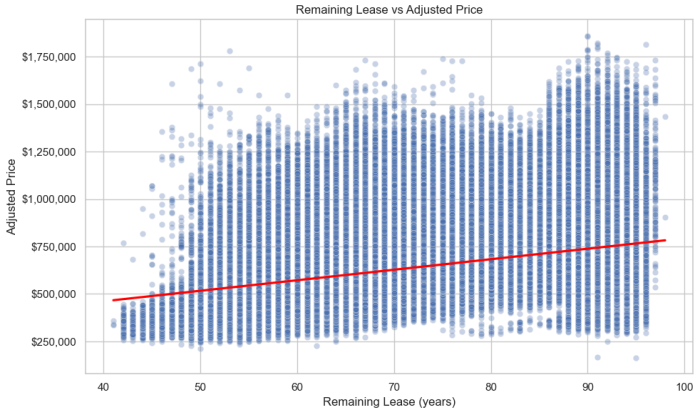


Figure B13: Plots for Maximum Floor Level

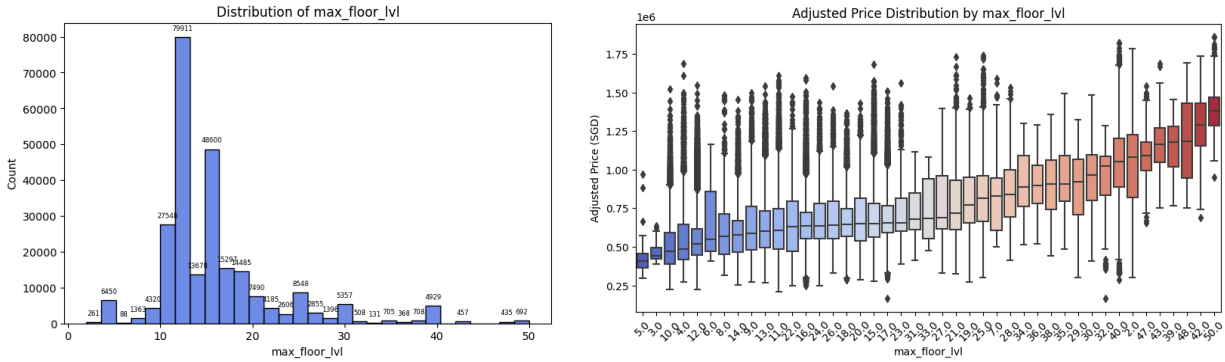


Figure B14: Adjusted Price vs Total Dwelling Units

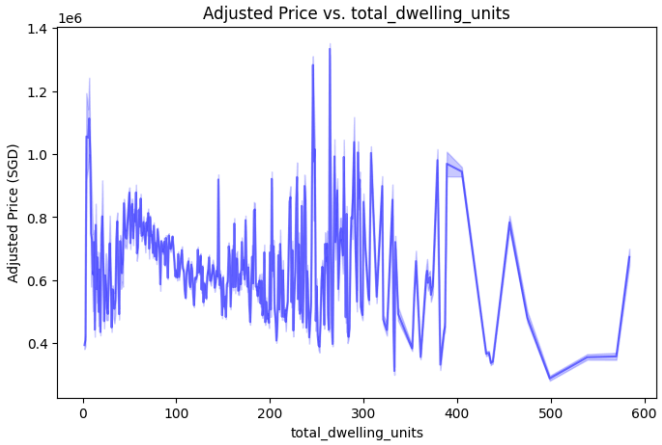


Figure B15: Statistics for Binary Variables

	residential	commercial	market_hawker	miscellaneous	multistorey_carpark	precinct_pavilion
N	0 (0.0%)	209883 (82.84%)	253348 (99.99%)	176870 (69.81%)	253313 (99.98%)	252872 (99.8%)
Y	253371 (100.0%)	43488 (17.16%)	23 (0.01%)	76501 (30.19%)	58 (0.02%)	499 (0.2%)

Figure B16: Plots for Commercial

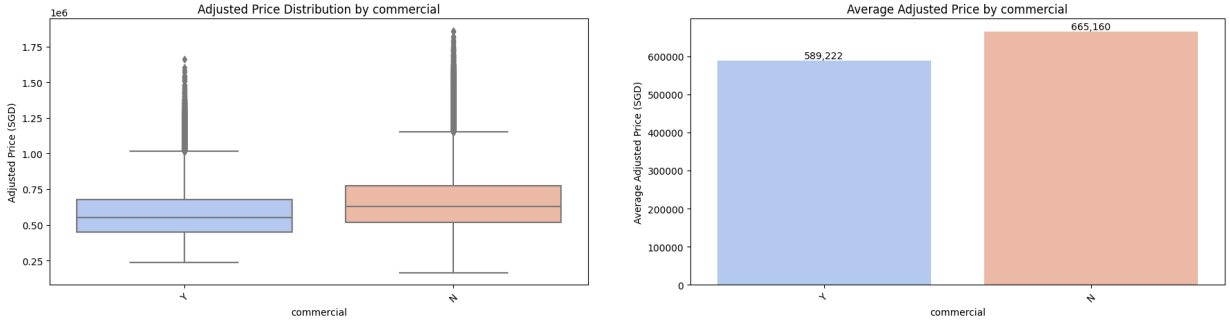


Figure B17: Plots for Miscellaneous

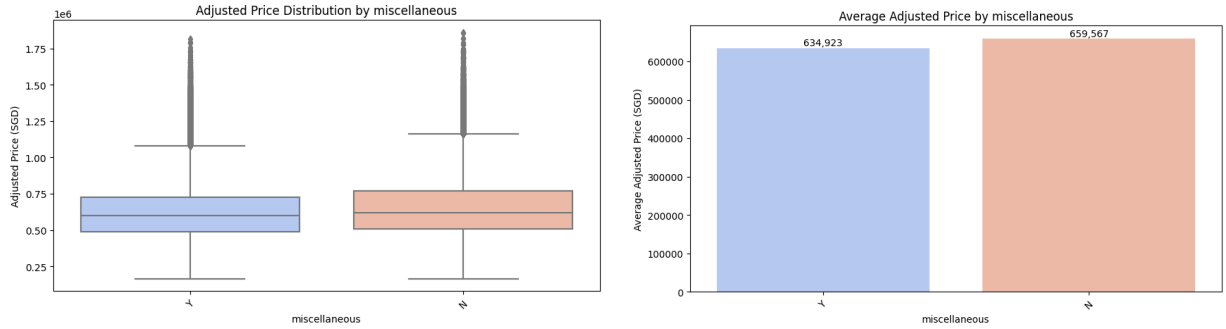


Figure B18: Proximity-Based Features

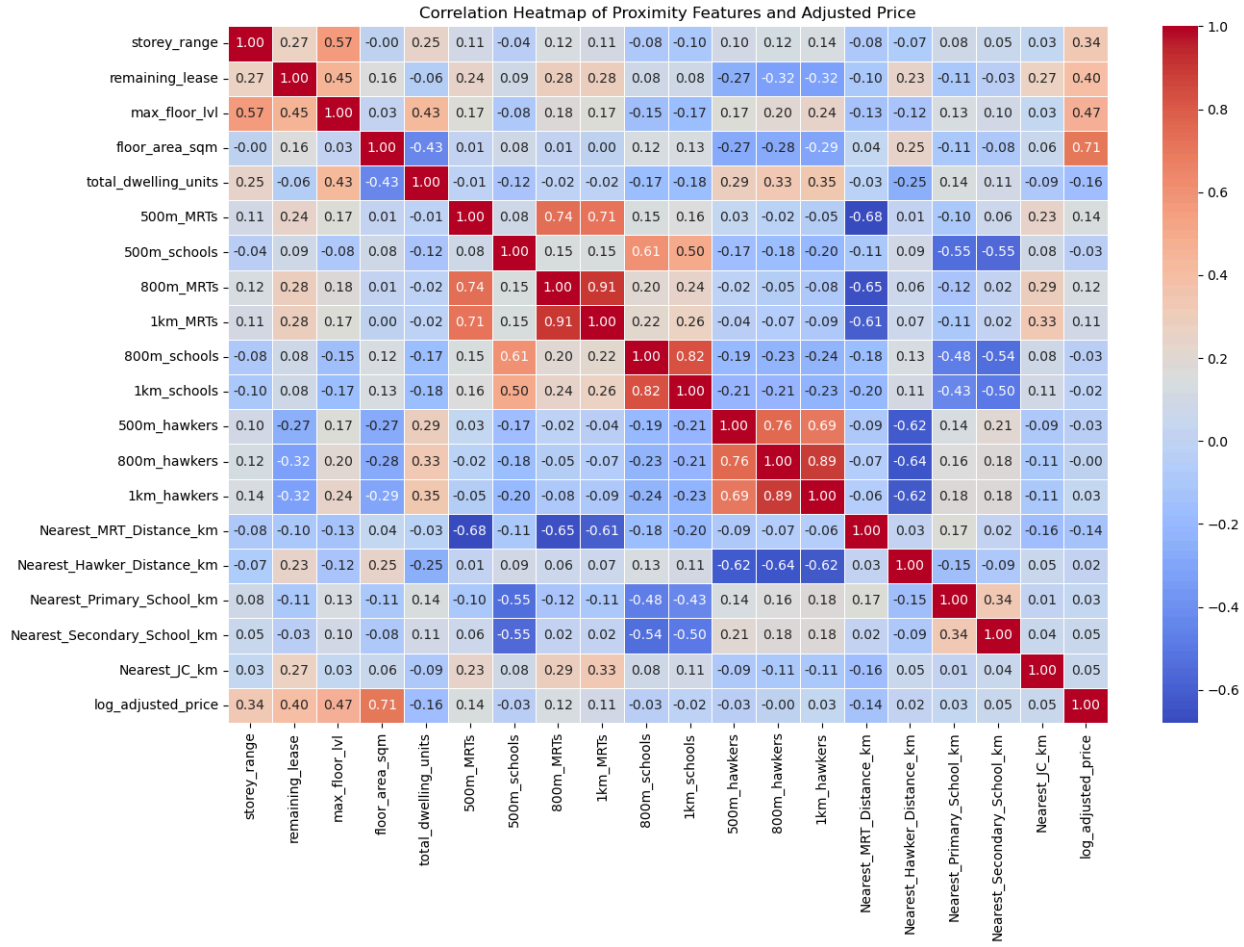
	street_name	block	latitude	longitude	500m_MRTs	800m_MRTs	1km_MRTs	Nearest_MRT_Distance_km
0	ANG MO KIO AVE 8	510	1.373401	103.849073	1	1	1	0.389381
1	ANG MO KIO AVE 5	603	1.380201	103.835756	0	1	2	0.589978
2	ANG MO KIO AVE 10	476	1.362388	103.857881	0	0	0	1.248878
3	ANG MO KIO AVE 10	558	1.371236	103.859192	0	0	0	1.080696
4	ANG MO KIO AVE 5	605	1.379672	103.836487	0	1	3	0.654067
...	...	...	...	...	...	...	...	...
253366	YISHUN RING RD	328	1.429780	103.843057	0	0	1	0.895828
253367	YISHUN ST 61	614	1.419829	103.836033	1	1	1	0.434857
253368	YISHUN ST 71	723	1.426020	103.829939	0	1	1	0.679695
253369	YISHUN ST 81	836	1.415452	103.833091	1	1	1	0.215093
253370	YISHUN ST 81	824	1.413745	103.833303	1	1	1	0.406126

Figure B19: One Hot & Ordinal Encoding

flat_model_Premium Apartment	flat_model_Premium Apartment Loft	flat_model_Premium Maisonette	flat_model_Simplified	flat_model_Standard	flat_model_Terrace	flat_model_Type S1	flat_model_Type S2
False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False

storey_range	storey_range_encoded
01 TO 03	1
04 TO 06	2
07 TO 09	3
10 TO 12	4
13 TO 15	5
16 TO 18	6
19 TO 21	7
22 TO 24	8
25 TO 27	9
28 TO 30	10
31 TO 33	11
34 TO 36	12
37 TO 39	13
40 TO 42	14
43 TO 45	15
46 TO 48	16
49 TO 51	17

Figure B20: Correlation Heatmap



Appendix C: Modelling

Figure C1: Linear Regression Prediction Error by Town

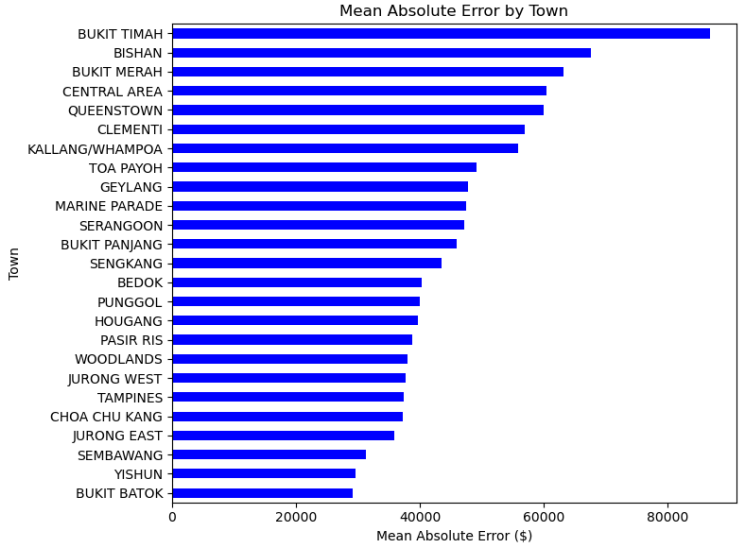


Figure C2: SVR Permutation Importance

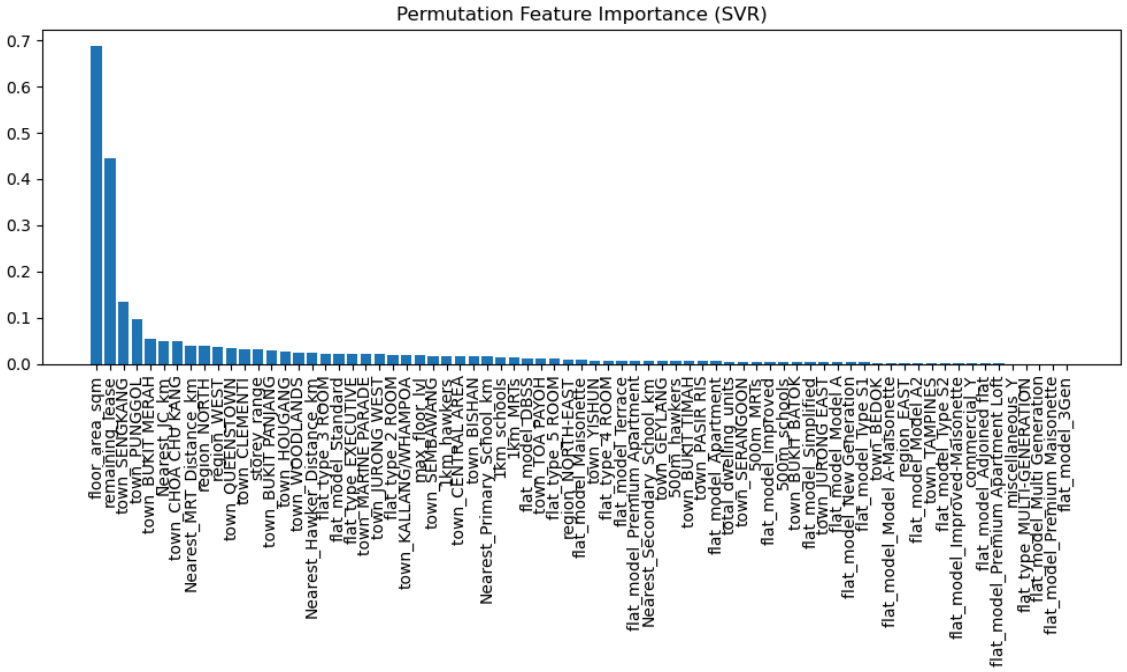


Figure C3: SVR Best Model Residuals and Prediction vs Actual Plot

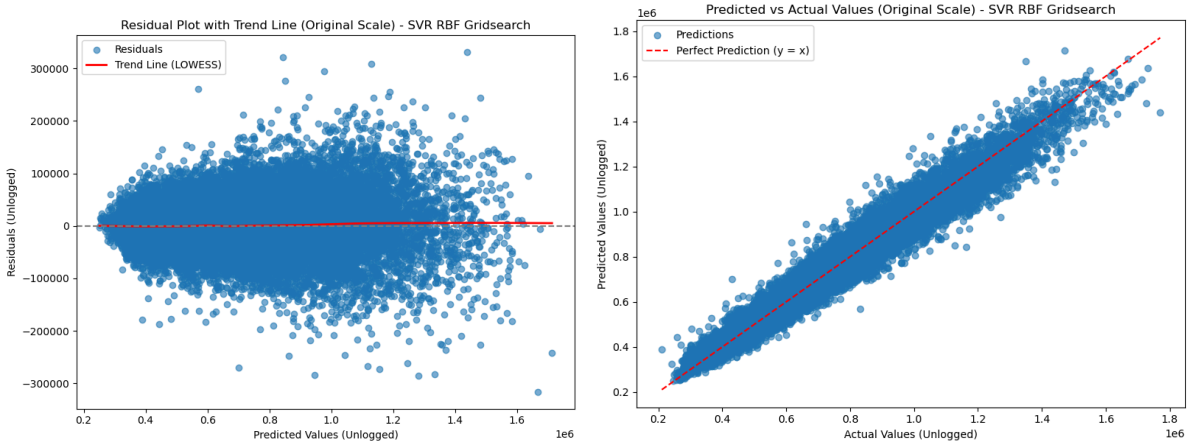


Figure C4: SVR Prediction Error by Town

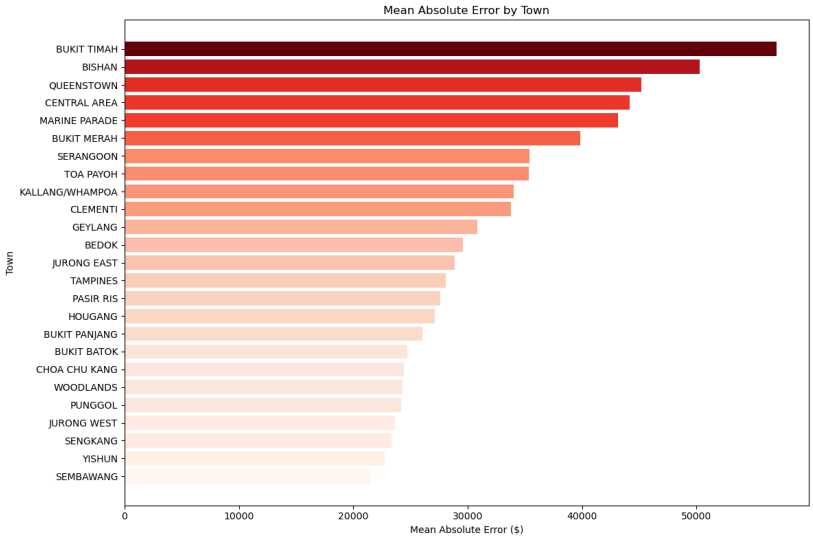


Figure C5: Neural Network Base Model

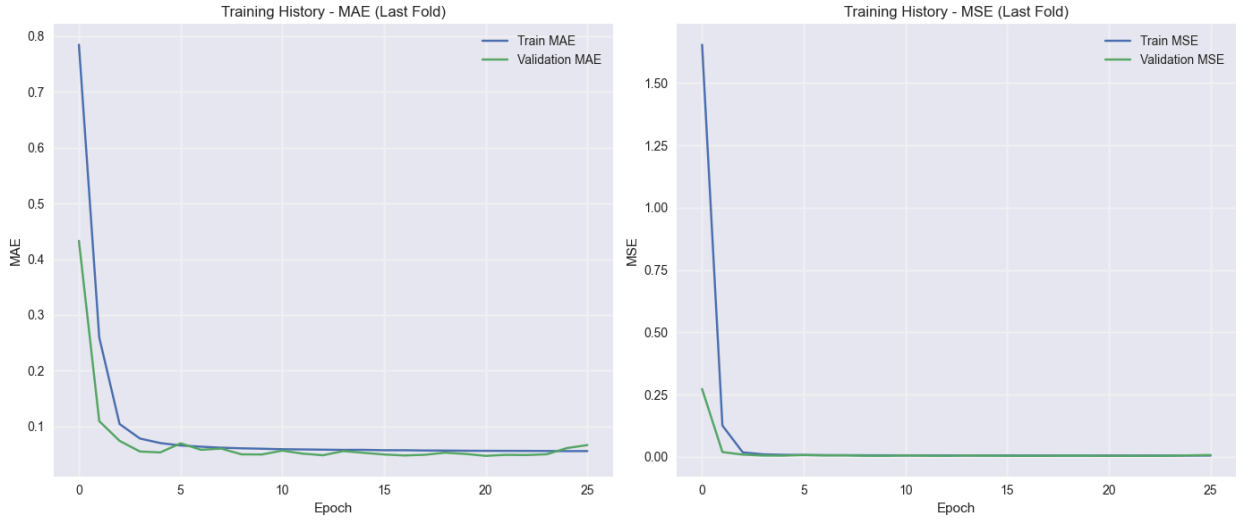


Figure C6: Neural Network Improved Model

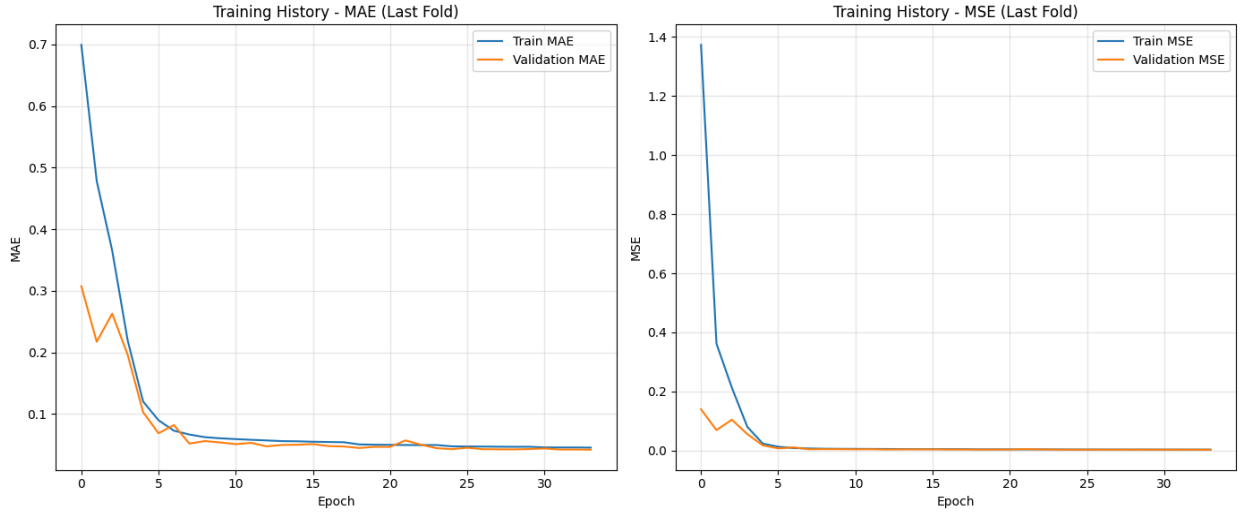


Figure C7: Neural Network Top 10 Feature Importance

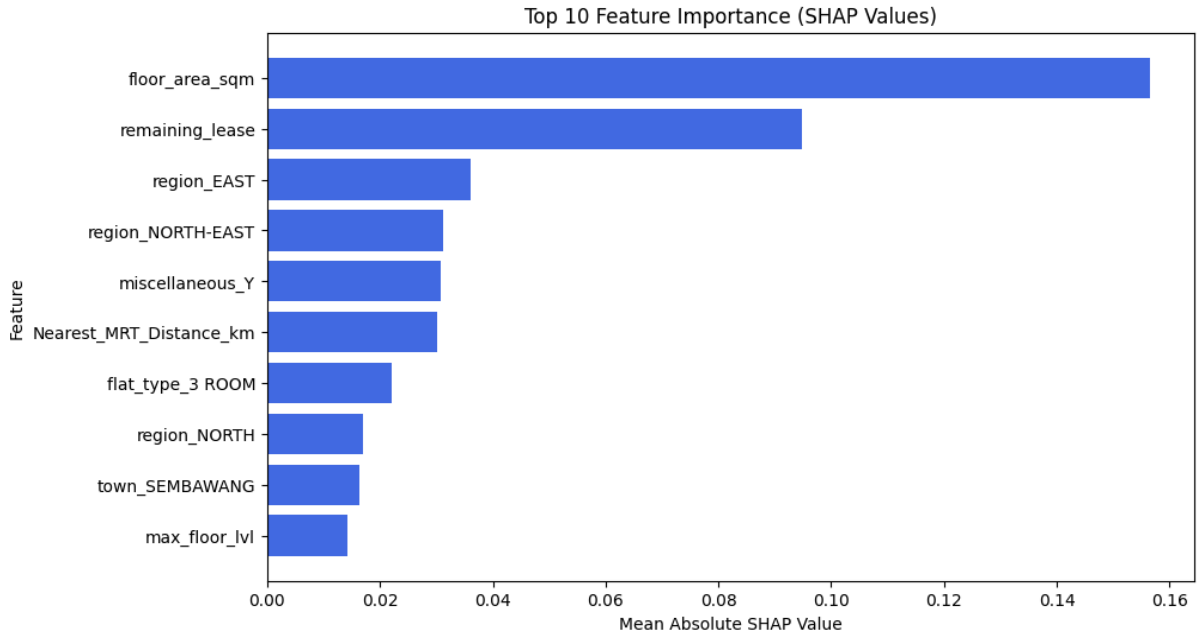


Figure C8: Neural Network Feature Importance Model

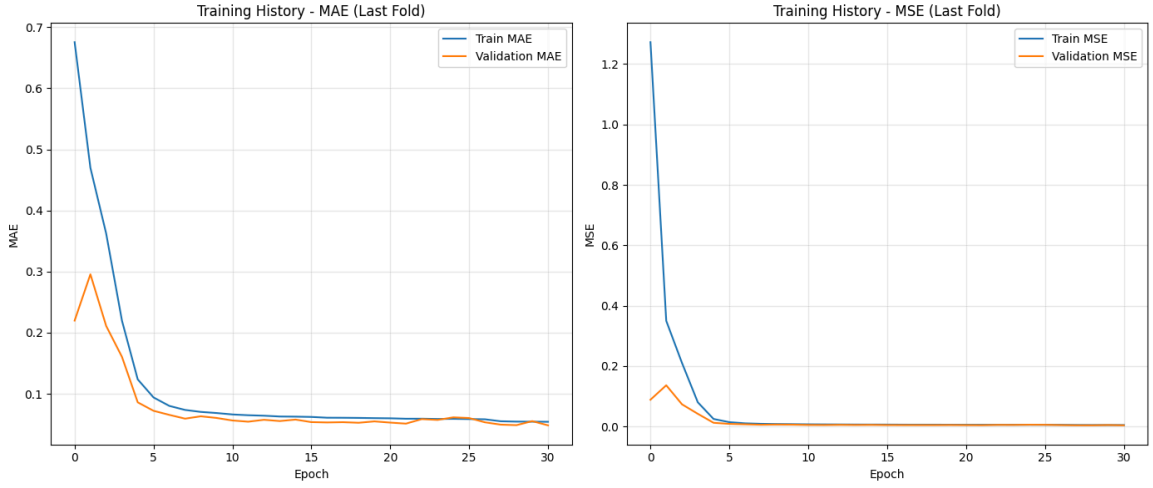


Figure C9: Neural Network Prediction Error by Town

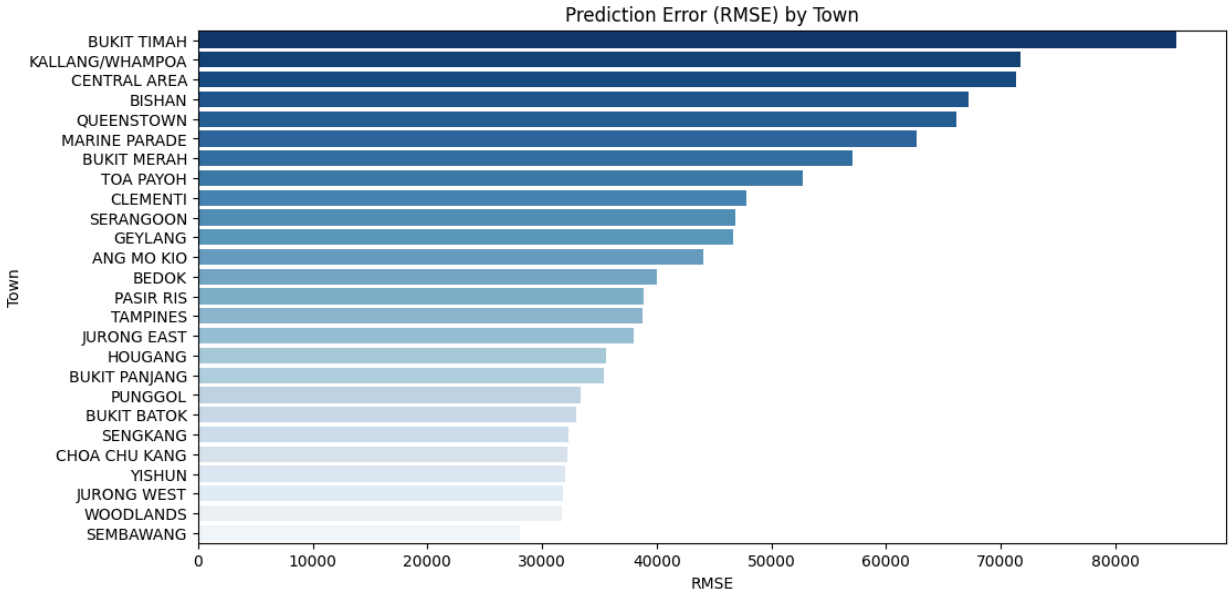


Figure C10: Neural Network Evaluation on Test Set

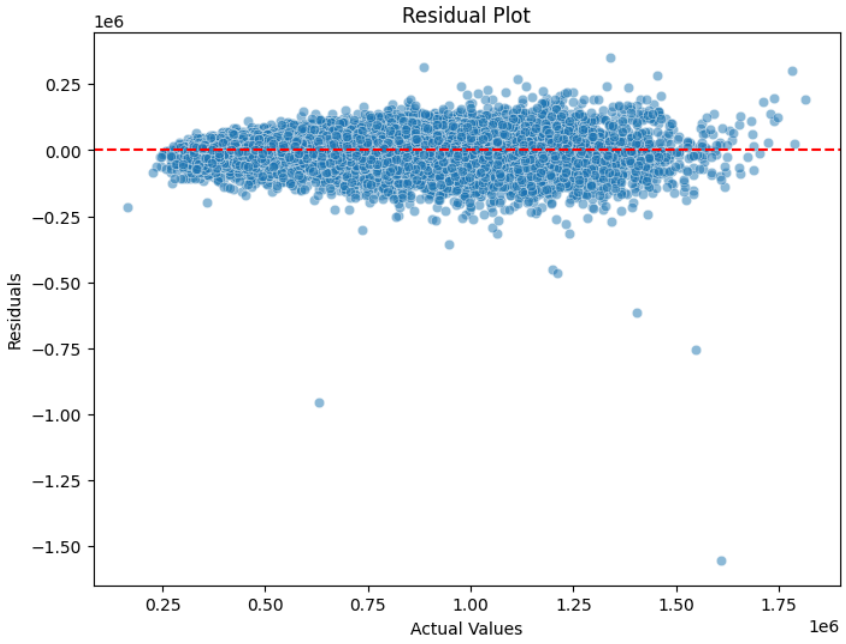


Figure C11: Random Forest Top 10 Most Important Features

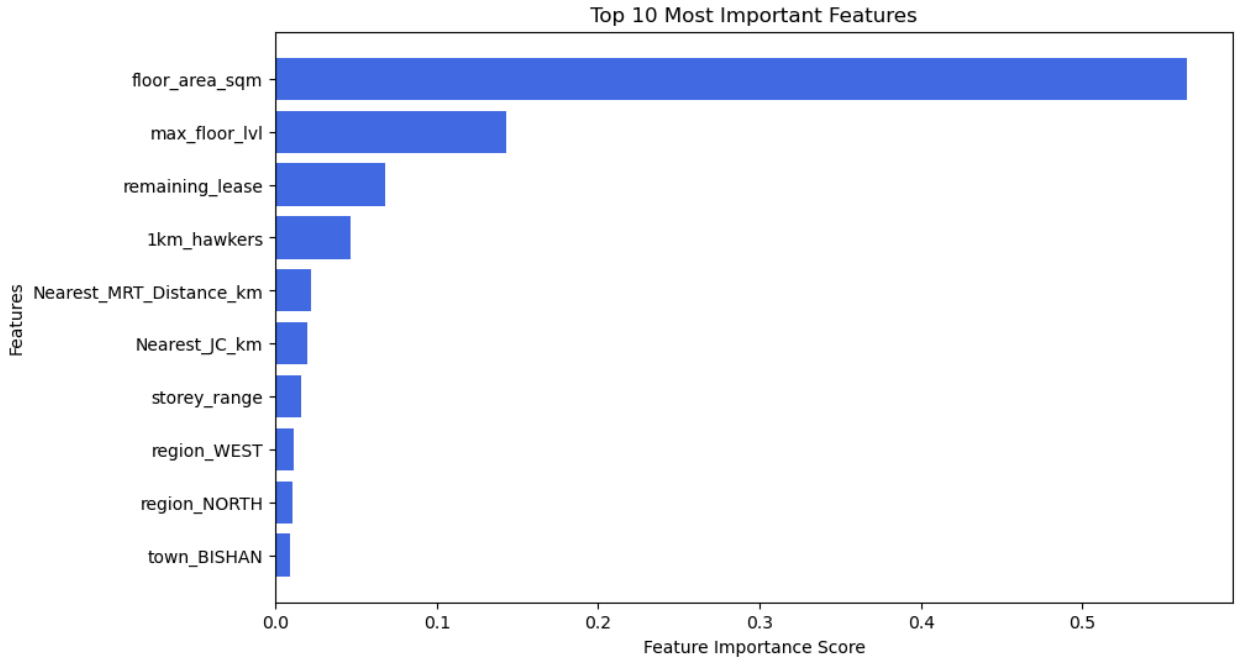


Figure C12: Random Forest Partial Dependence Plot

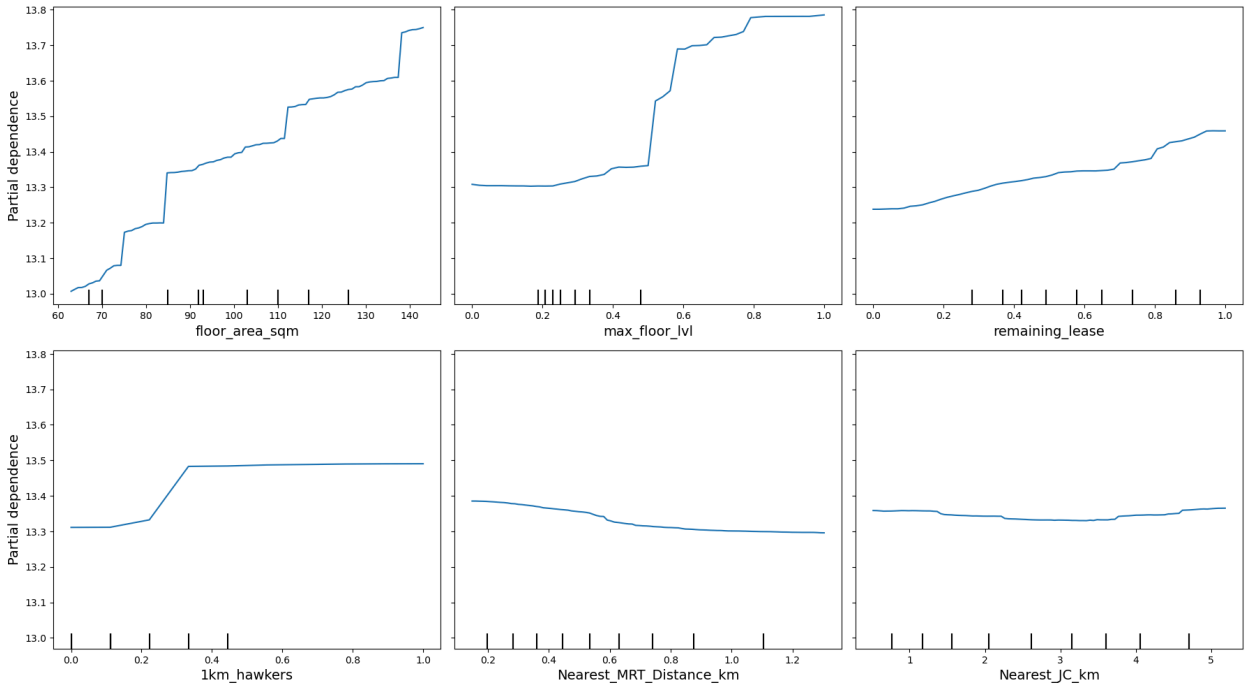


Figure C13: Random Forest Residual Plot

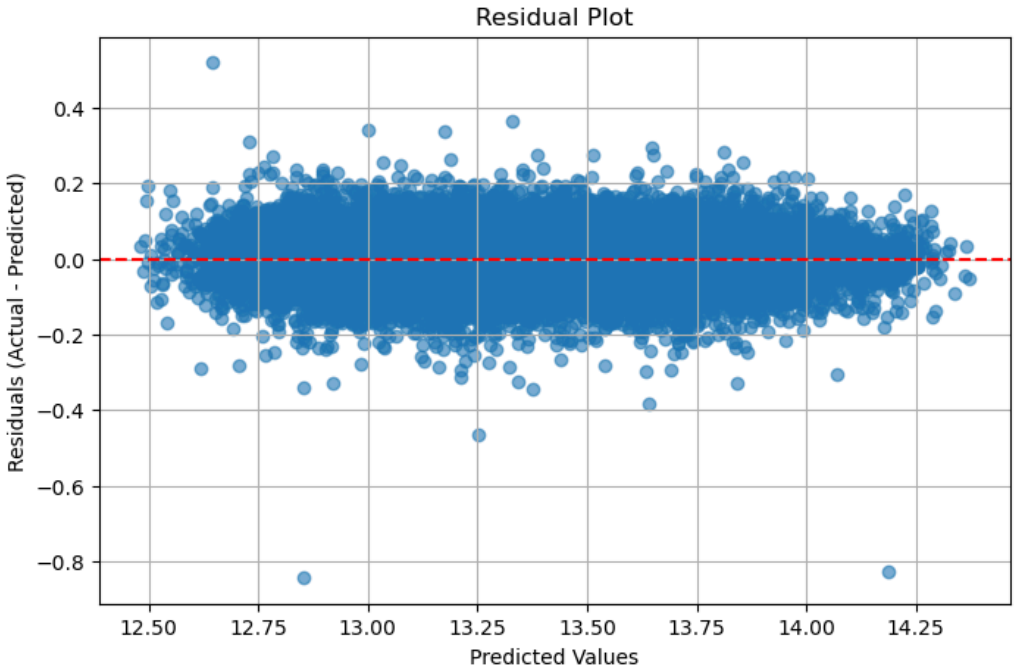


Figure C14: Random Forest: First Tree

```
|--- floor_area_sqm <= 84.50
|   |--- max_floor_lvl <= 0.57
|   |   |--- floor_area_sqm <= 59.50
|   |   |   |--- Nearest_Primary_School_km <= 1.80
|   |   |   |   |--- floor_area_sqm <= 45.50
|   |   |   |   |   |--- remaining_lease <= 0.32
|   |   |   |   |   |   |--- remaining_lease <= 0.15
|   |   |   |   |   |   |   |--- Nearest_JC_km <= 1.26
|   |   |   |   |   |   |   |   |--- storey_range <= 3.50
|   |   |   |   |   |   |   |   |   |--- floor_area_sqm <= 44.00
|   |   |   |   |   |   |   |   |   |   |--- remaining_lease <= 0.13
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 5
|   |   |   |   |   |   |   |   |   |   |--- remaining_lease >  0.13
|   |   |   |   |   |   |   |   |   |   |   |--- value: [12.64]
|   |   |   |   |   |   |   |   |   |--- floor_area_sqm >  44.00
|   |   |   |   |   |   |   |   |   |   |--- value: [12.64]
|   |   |   |   |   |   |   |   |--- storey_range >  3.50
|   |   |   |   |   |   |   |   |   |--- value: [12.70]
|   |   |   |   |   |   |   |--- Nearest_JC_km >  1.26
|   |   |   |   |   |   |   |   |--- Nearest_Secondary_School_km <= 1.06
|   |   |   |   |   |   |   |   |   |--- remaining_lease <= 0.10
|   |   |   |   |   |   |   |   |   |   |--- Nearest_MRT_Distance_km <= 0.58
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 5
|   |   |   |   |   |   |   |   |   |   |--- Nearest_MRT_Distance_km >  0.58
|   |   |   |   |   |   |   |   |   |   |   |--- value: [12.57]
...
|   |   |   |   |   |   |   |   |   |--- value: [13.93]
|   |   |   |   |   |   |   |--- total_dwelling_units >  0.17
|   |   |   |   |   |   |   |   |--- value: [13.90]
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Figure C15: Random Tree: Second Tree

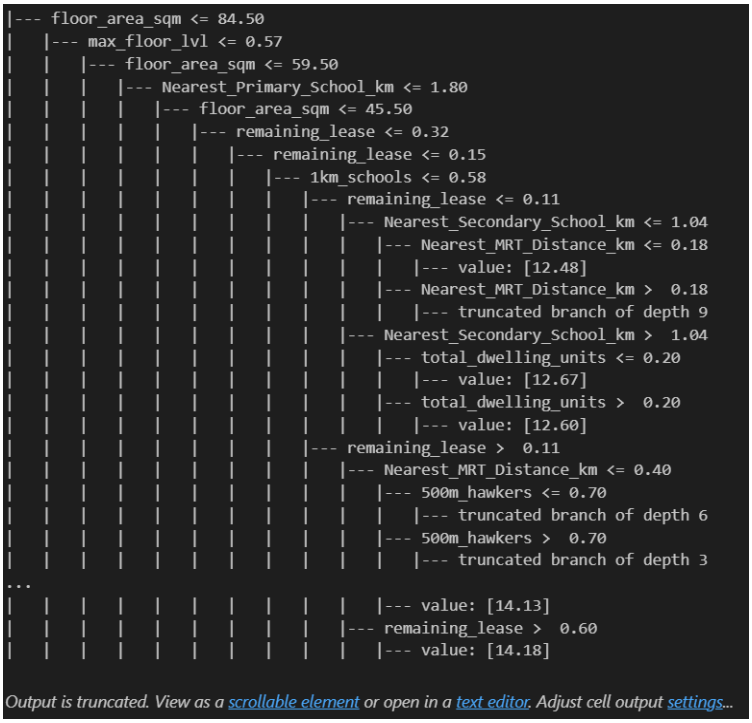


Figure C16: Random Tree: Third Tree

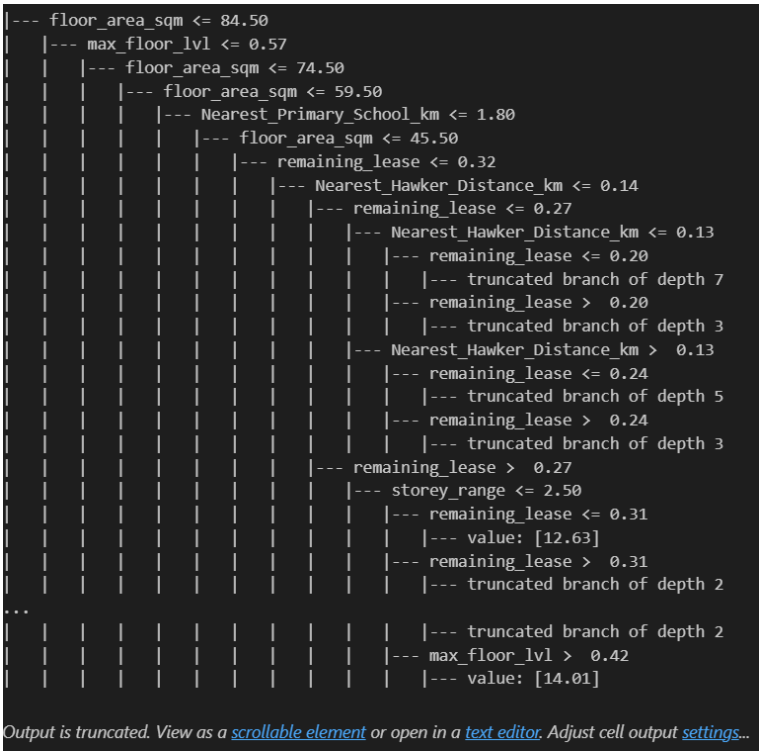


Figure C17: MAE by Remaining Lease Bins (Test Data)

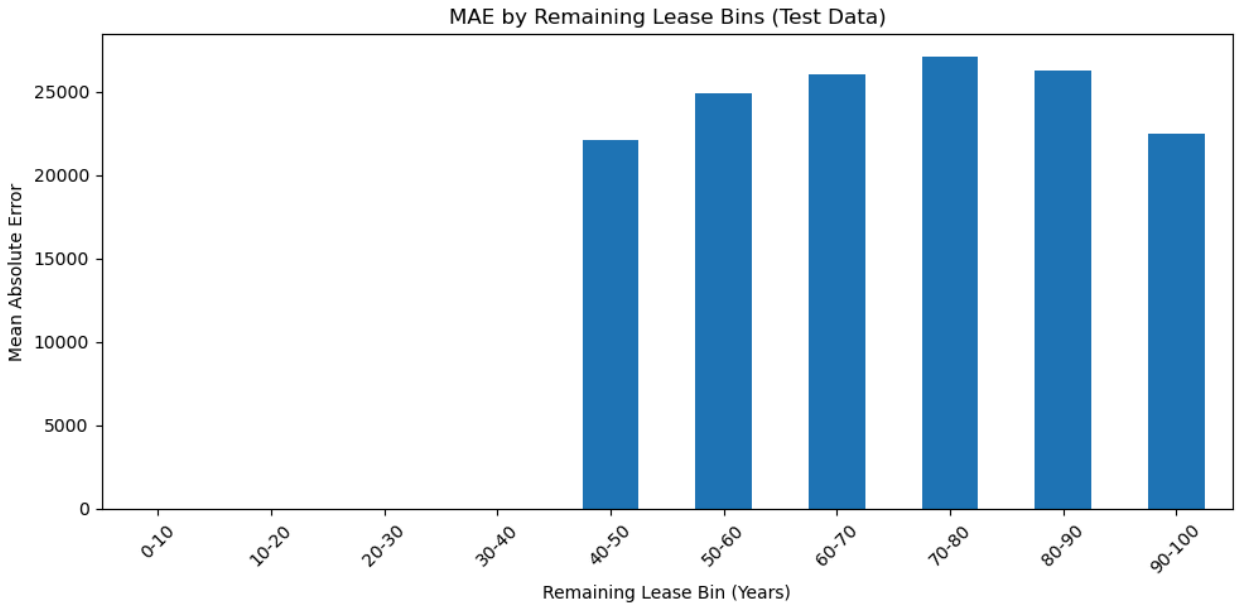


Figure C18: MAE by Remaining Lease Bins (New Data)

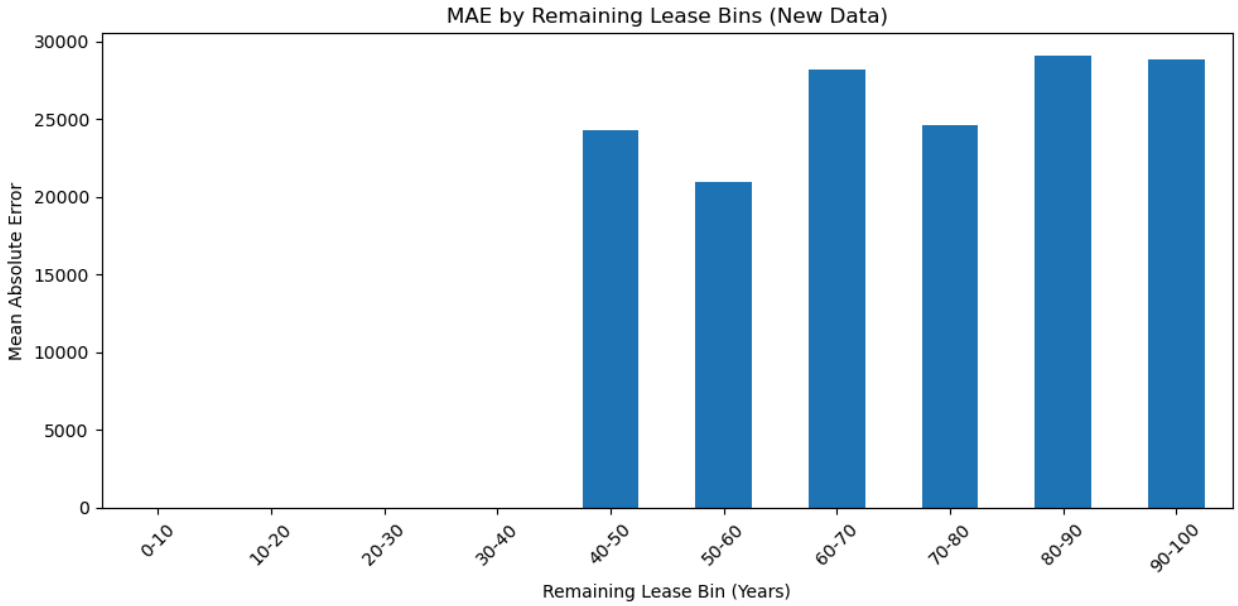


Figure C19: MAE by Floor Area Bins (Test Data)

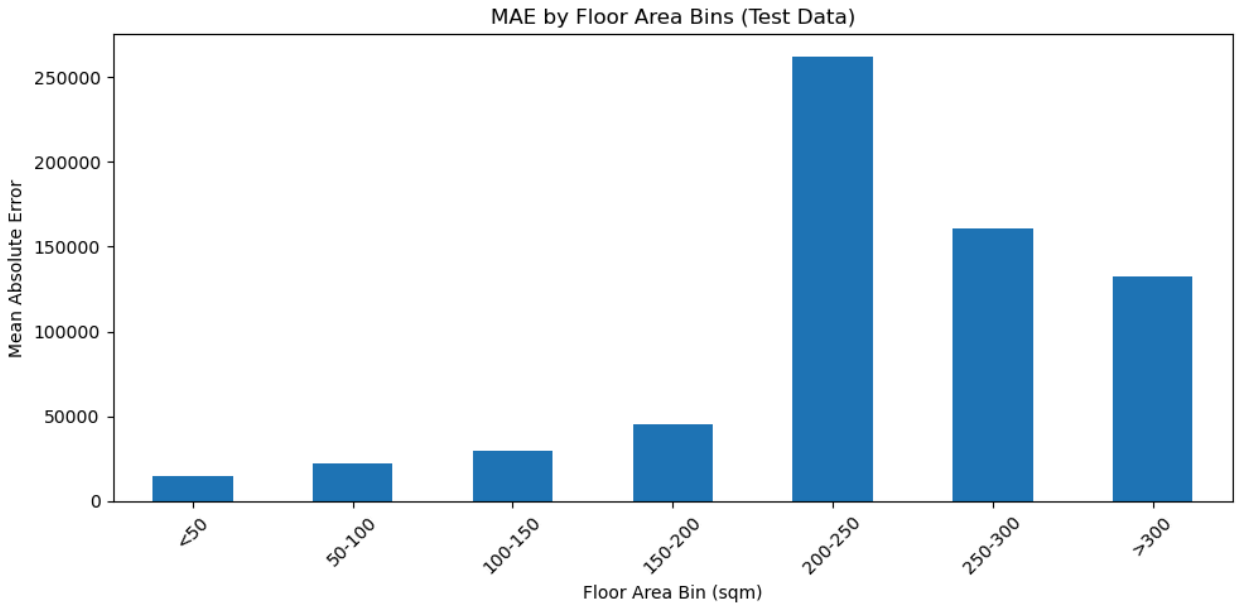


Figure C20: MAE by Floor Area Bins (New Data)

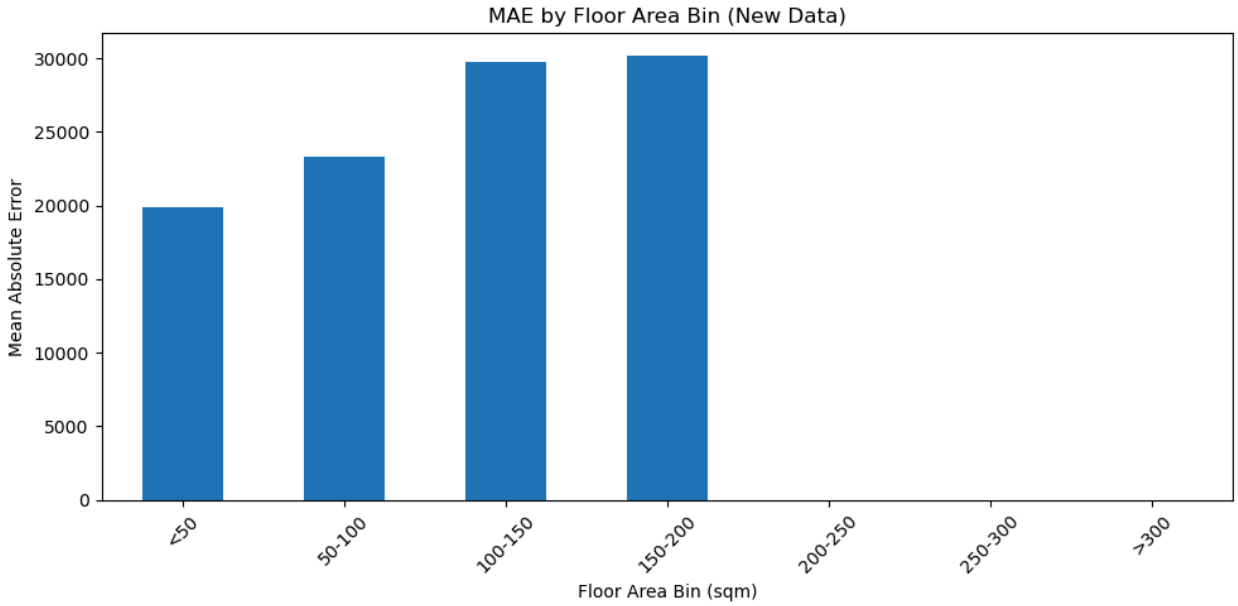


Figure C21: MAE by Floor Area Bins Count (Test Data)

```
Count of rows per area_bin:
area_bin
<50          802
50-100     26844
100-150    22300
150-200      726
200-250        1
250-300        1
>300           1
Name: count, dtype: int64
```