ST3131 Assignment

By Jay Tai Kin Heng, A0308787L

Introduction:

The resale price of Housing and Development Board (HDB) flats in Singapore is influenced by a wide range of factors, such as location, flat type, size, and remaining lease. In this report, we aim to develop an optimal linear regression model to predict HDB resale prices. The dataset used contains resale transactions from January to July 2021, comprising 11,527 records and 11 variables. We will begin by exploring the relationships between the response variable (resale price) and potential predictors, followed by constructing and evaluating various linear models. Finally, we will interpret the results and discuss the implications of the findings from the final model.

EDA - Response Variable:

Our response variable, *resale price (in SGD)*, is continuous and quantitative. Its distribution is unimodal and likely right-skewed, as observed from Figure 1 and supported by the right tail in the QQ plot. The mean resale price is 496,544.6, with a standard deviation of 161,965.1, ranging from 180,000 to 1,250,000, and 245 outliers identified. Since the resale price distribution deviates from normality, we use its log-transformed form as the response variable in subsequent analyses to stabilize the variance.

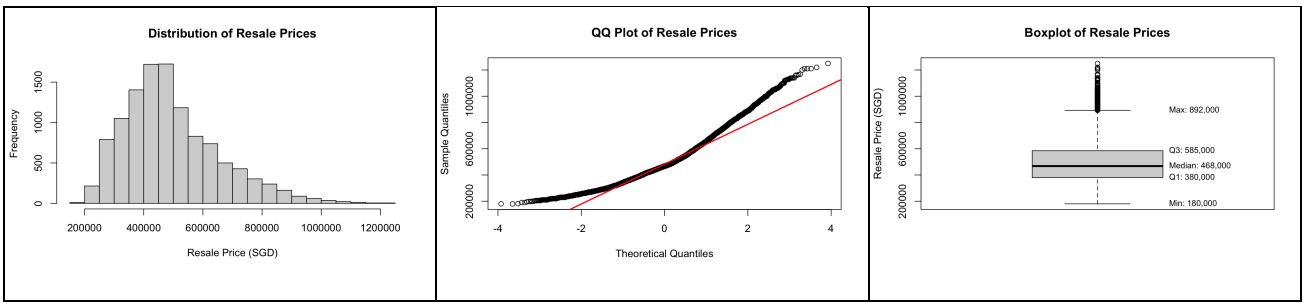


Figure 1: EDA plots of Resale Price

The dataset contains 10 additional predictors, which we classify into two categories — categorical and quantitative — to facilitate discussion.

EDA - Categorical Variables:

Month, Town, Flat type, Flat model, Block, Street name

The variable *Town* was grouped into five major regions — *Central*, *East*, *North*, *North-East*, and *West* — based on the Urban Redevelopment Authority's regional classification. This regional grouping reduces dimensionality while retaining spatial characteristics, as towns within each region share similar locational features and price dynamics.

Street name and *Block* were dropped since they capture micro-location details already reflected in *Town*. *Flat model* was also dropped to avoid multicollinearity, as it provides redundant information explained by *Flat type* and *Remaining lease*.

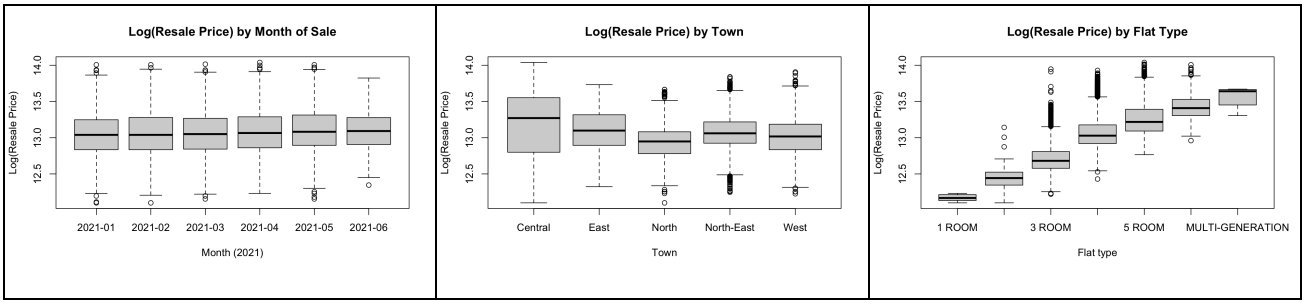


Figure 2: Boxplots of Log Resale Price against different Categorical Variables

The boxplots (Figure 2) show that:

- The month of sale has minimal impact on log resale price, consistent with the short time span of the dataset.
- Regional differences exist, with certain regions commanding higher prices, reflecting spatial variation.
- Flats with more rooms tend to have higher log resale prices.

EDA - Quantitative Variables:

Floor area (sqm), Lease commence date, Remaining lease, Storey Range

To enable quantitative analysis, *Storey range* was converted into a numerical variable representing the midpoint of each range (e.g., "04 TO 06" → 5). This allows the model to capture the effect of floor height as a continuous predictor.

Remaining lease was first converted into total months to ensure consistency as a numeric variable. However, since *Lease commence date* and *Remaining lease* represent the same underlying concept — the age of the flat — we retained *Lease commencement date* for interpretability and dropped *Remaining lease* to prevent multicollinearity.

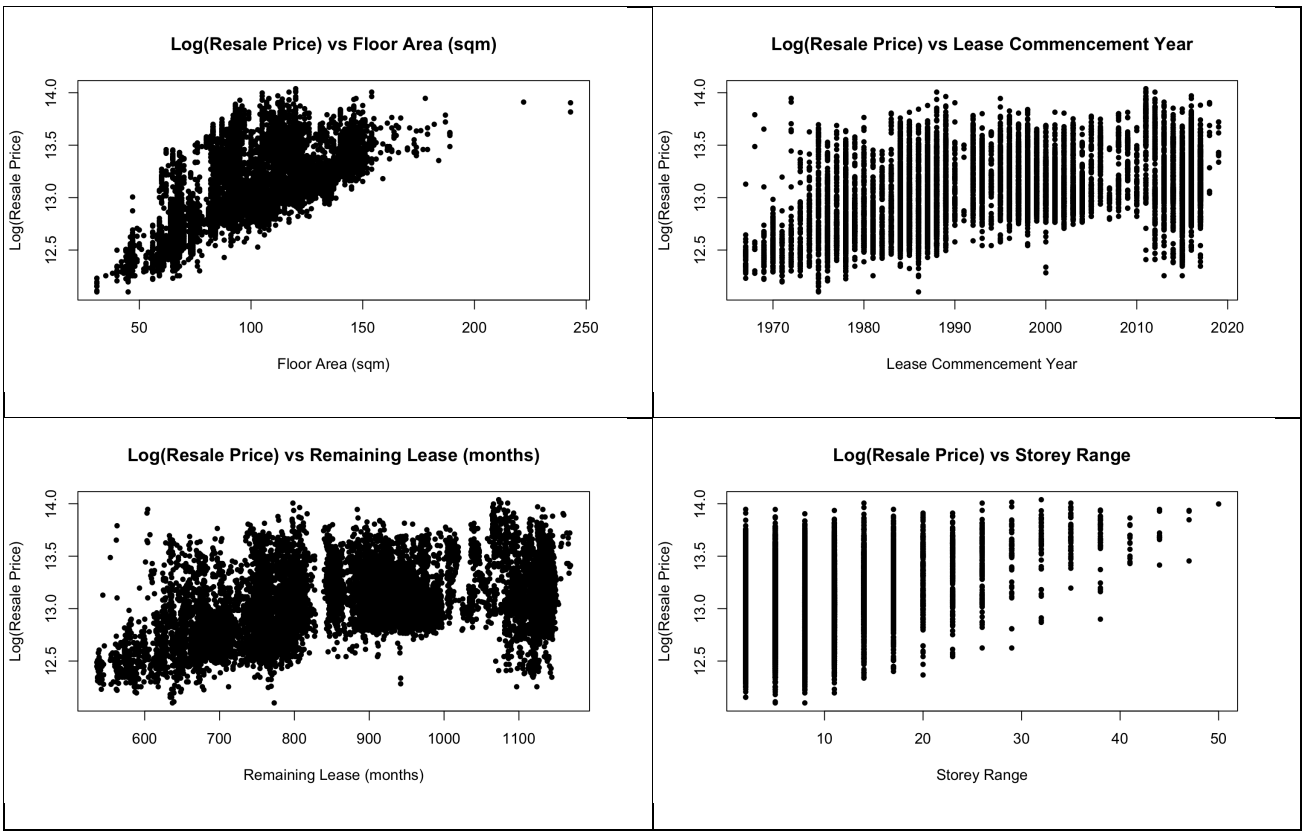


Figure 3: Scatterplots of Log Resale Price against different Quantitative Variables

Scatterplots (Figure 3) show that:

- *Floor area* is strongly positively correlated with log resale price (r = 0.6848), indicating that larger flats generally command higher prices.
- *Lease commence date* shows a weaker positive correlation (r = 0.4098), suggesting that newer flats tend to have slightly higher resale prices, though other factors such as location and flat type also influence the price.

- *Storey range* has a relatively weak positive correlation with log resale price (r = 0.3488). The scatterplot shows an “inverted triangle” pattern: flats on higher floors typically have higher prices, but there are fewer flats at these storeys, while lower floors are more common and exhibit a wider spread of resale prices.

To summarize, for the initial linear model, we use *log resale price* as the response variable, with the following predictors: *Month, Town, Flat type, Storey range, Floor area (sqm), and Lease commence date*.

Building Models – Initial Model (M0):

```
Call:
lm(formula = log(resale_price) ~ month + town + flat_type + storey_range +
    floor_area_sqm + lease_commence_date, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63630 -0.08762 -0.01244  0.07746  0.71643

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -5.0685191  0.2160316 -23.462  < 2e-16 ***
month2021-02              0.0109463  0.0038917   2.813  0.00492 **
month2021-03              0.0185823  0.0037711   4.928 8.44e-07 ***
month2021-04              0.0278929  0.0038111   7.319 2.67e-13 ***
month2021-05              0.0399064  0.0039944   9.990  < 2e-16 ***
month2021-06              0.0553566  0.0124180   4.458 8.36e-06 ***
townEast                 -0.1969752  0.0042704 -46.125  < 2e-16 ***
townNorth                -0.3749700  0.0046397 -80.817  < 2e-16 ***
townNorth-East           -0.2928621  0.0038683 -75.709  < 2e-16 ***
townWest                 -0.3205321  0.0040002 -80.128  < 2e-16 ***
flat_type2 ROOM           0.1356094  0.0481335   2.817  0.00485 **
flat_type3 ROOM           0.3290025  0.0473856   6.943 4.04e-12 ***
flat_type4 ROOM           0.4009178  0.0483694   8.289  < 2e-16 ***
flat_type5 ROOM           0.4301175  0.0495009   8.689  < 2e-16 ***
flat_typeEXECUTIVE        0.4951049  0.0512553   9.660  < 2e-16 ***
flat_typeMULTI-GENERATION 0.5662775  0.0695418   8.143 4.25e-16 ***
storey_range              0.0097574  0.0002144  45.513  < 2e-16 ***
floor_area_sqm            0.0075915  0.0001764  43.046  < 2e-16 ***
lease_commence_date       0.0085703  0.0001063  80.602  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1323 on 11508 degrees of freedom
Multiple R-squared:  0.8261,    Adjusted R-squared:  0.8259
F-statistic:  3038 on 18 and 11508 DF,  p-value: < 2.2e-16
```

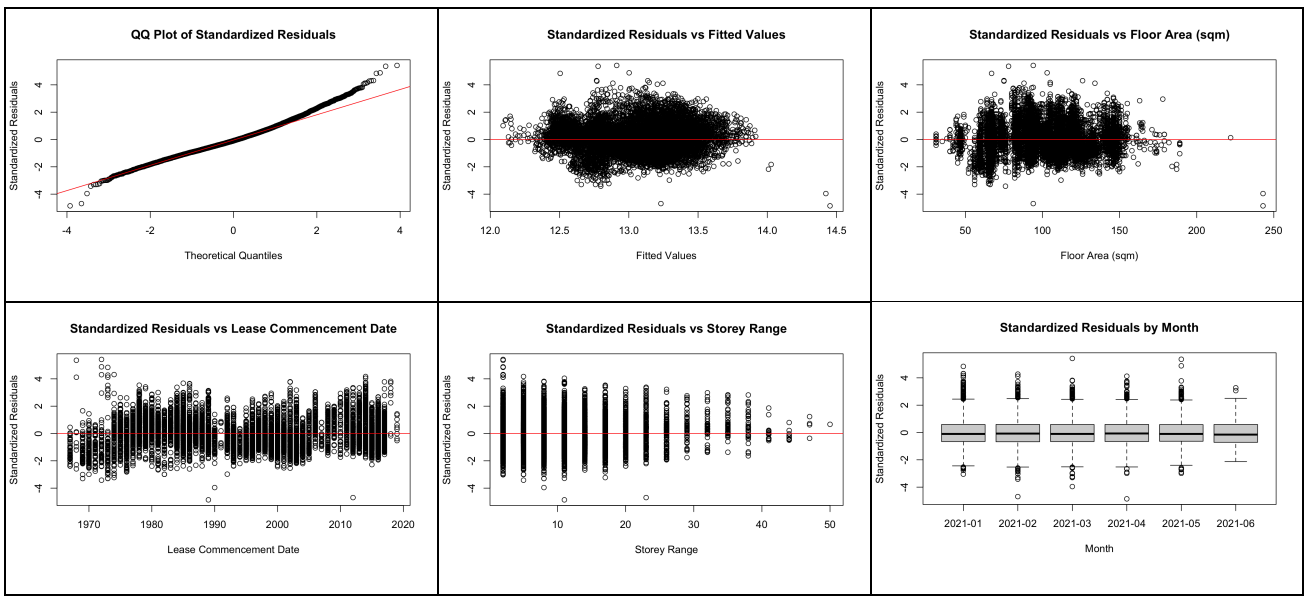
Figure 4: Summary Table for Initial Model M0

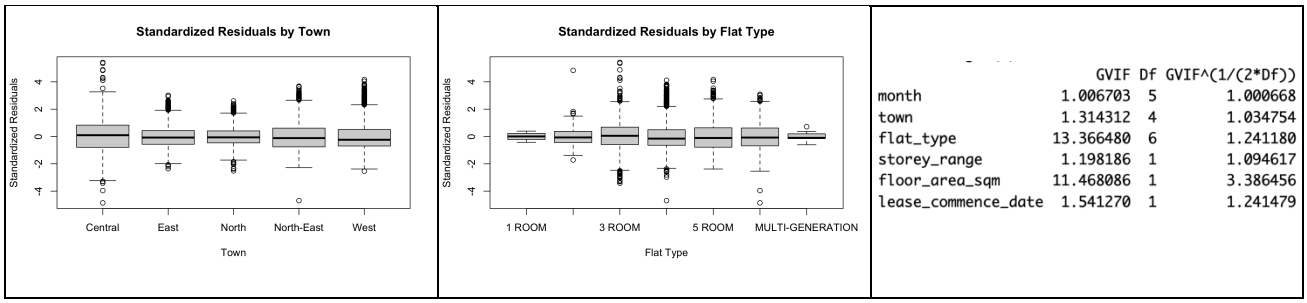
```
Analysis of Variance Table

Response: log(resale_price)
                    Df Sum Sq Mean Sq  F value    Pr(>F)
month                5   3.44   0.687   39.274 < 2.2e-16 ***
town                 4  61.71  15.427  881.287 < 2.2e-16 ***
flat_type            6 684.85 114.142 6520.724 < 2.2e-16 ***
storey_range         1  82.11  82.109 4690.712 < 2.2e-16 ***
floor_area_sqm       1  11.39  11.393  650.856 < 2.2e-16 ***
lease_commence_date  1 113.72 113.722 6496.732 < 2.2e-16 ***
Residuals        11508 201.44   0.018
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: ANOVA Table for Initial Model M0

The initial linear model predicted *log resale price using month, town, flat type, storey range, floor area (sqm), and lease commence date*. This model explained a substantial portion of variation in resale prices (Multiple R-squared = 0.8261, Adjusted R-squared = 0.8259) with a residual standard error of 0.1323 on the log scale. The overall F-statistic was highly significant (3038, p < 2.2e-16), indicating that the predictors collectively had a strong effect.





	GVIF	Df	GVIF^(1/(2*Df))
month	1.006703	5	1.000668
town	1.314312	4	1.034754
flat_type	13.366480	6	1.241180
storey_range	1.198186	1	1.094617
floor_area_sqm	11.468086	1	3.386456
lease_commence_date	1.541270	1	1.241479

Figure 6: MAC plots for Initial Model M0

Model Assumption Checks (Initial Observations) (Figure 6):

- Residual patterns: Standardized residuals showed minor deviations from normality. Residuals vs. fitted values indicated no strong heteroscedasticity, though extreme residuals were observed.
- Numerical predictors: Floor area residuals showed some extreme points; storey range displayed a reverse-cone shape, suggesting variance might benefit from transformation.
- Categorical predictors: Flat type residuals showed high variability across categories, indicating grouping might improve stability. Month had minimal practical effect.
- Outliers / Influential points: 240 outliers were detected, but no influential points were present.
- Multicollinearity: GVIF values indicated weak to moderate correlation between floor area and flat type, which is expected; other predictors had low multicollinearity.

Decisions to improve the model:

1. Drop month – minimal practical impact.
2. Group flat type – reduce residual variability.
3. Group or Transform storey range – address reverse-cone variance pattern.
4. Log-transform floor area – improve linearity and stabilize residuals.

<u>Building Model – Final Model (M5):</u>

The final model predicts log resale price using the transformed and grouped predictors:

$$\begin{aligned} log\left(resale_{price}\right)_{hat} \\ &= -7.6604 + 0.8766 \cdot log\left(floor_{area_{sqm}}\right) + 0.0085 \cdot lease_{commence_{date}} \\ &- 0.2049 \cdot I(town = East) - 0.3806 \cdot I(town = North) - 0.3009 \\ &\cdot I(town = North - East) - 0.3277 \cdot I(town = West) - 0.0279 \\ &\cdot I\left(flat_{type} = Small - Medium\right) + 0.0444 \cdot I(storey = Mid) + 0.0871 \\ &\cdot I(storey = High) + 0.2596 \cdot I(storey = Very\,High), \end{aligned}$$

$Where\ I(.)\ is\ an\ indicator\ function\ equal\ to\ 1\ if\ the\ condition\ is\ true\ and\ 0\ otherwise.$
$The\ baseline\ categories\ are\ Central\ for\ town, Large\ for\ flat\ type\ and\ Low\ for\ storey\ group.$

```
Call:
lm(formula = log(resale_price) ~ log(floor_area_sqm) + lease_commence_date +
    town + flat_type_grouped + storey_grouped, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.71115 -0.09138 -0.01284  0.08084  0.65985

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -7.6604026  0.1990357  -38.49  < 2e-16 ***
log(floor_area_sqm)             0.8766372  0.0075960  115.41  < 2e-16 ***
lease_commence_date             0.0084839  0.0001007   84.25  < 2e-16 ***
townEast                       -0.2049308  0.0043834  -46.75  < 2e-16 ***
townNorth                      -0.3806238  0.0047814  -79.60  < 2e-16 ***
townNorth-East                 -0.3009224  0.0040067  -75.10  < 2e-16 ***
townWest                       -0.3276645  0.0041134  -79.66  < 2e-16 ***
flat_type_groupedSmall-Medium -0.0279434  0.0039358   -7.10 1.32e-12 ***
storey_groupedMid               0.0443847  0.0028757   15.44  < 2e-16 ***
storey_groupedHigh              0.0870690  0.0038013   22.91  < 2e-16 ***
storey_groupedVery High         0.2595887  0.0067333   38.55  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1359 on 11516 degrees of freedom
Multiple R-squared:  0.8163,    Adjusted R-squared:  0.8162
F-statistic:  5118 on 10 and 11516 DF,  p-value: < 2.2e-16
```

Figure 7: Summary Table for Final Model M5

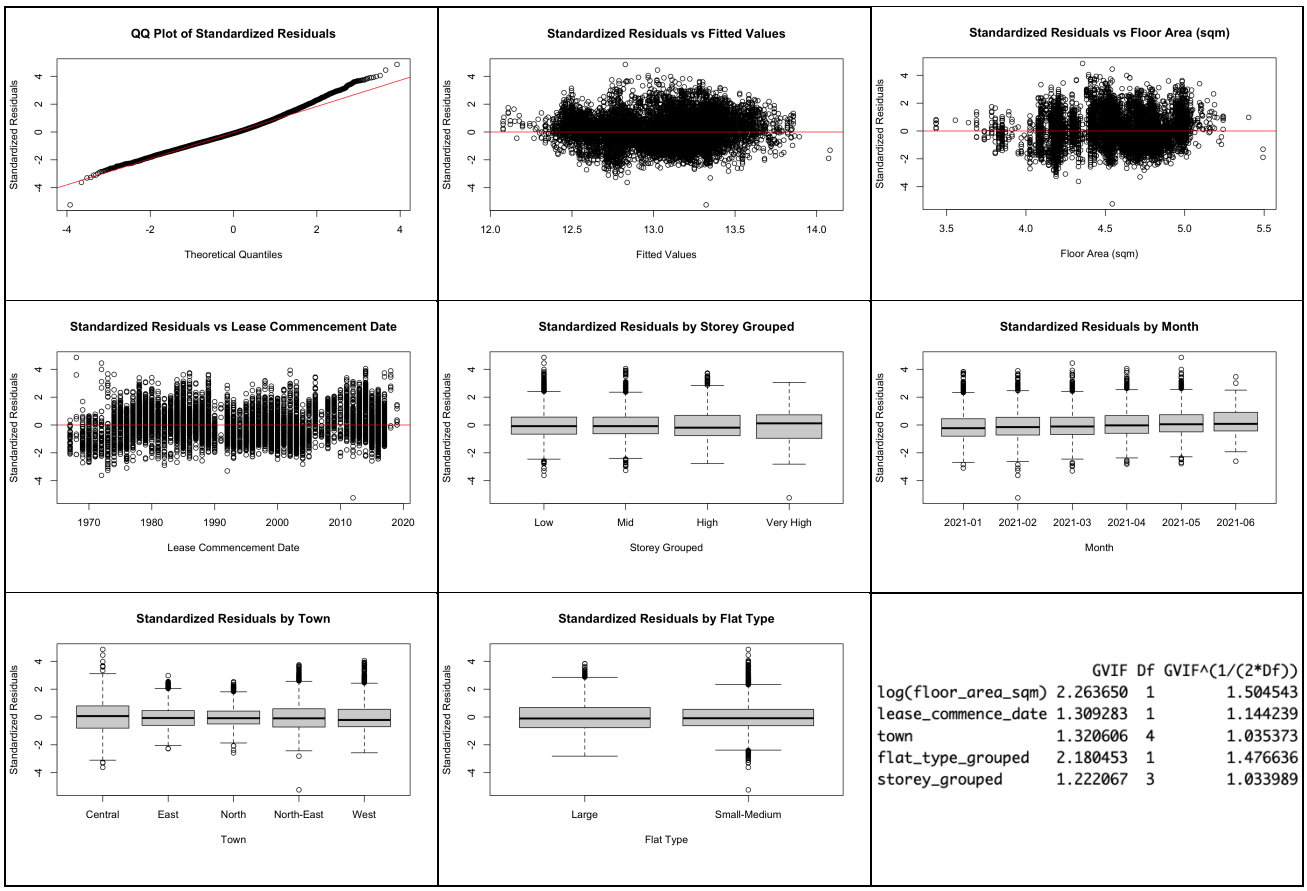
```
Analysis of Variance Table

Response: log(resale_price)
                     Df Sum Sq Mean Sq   F value    Pr(>F)
log(floor_area_sqm)   1 571.77  571.77 30937.409 < 2.2e-16 ***
lease_commence_date   1 120.24  120.24  6505.963 < 2.2e-16 ***
town                  4 221.34   55.33  2993.999 < 2.2e-16 ***
flat_type_grouped     1   1.22    1.22    65.801 5.494e-16 ***
storey_grouped        3  31.26   10.42   563.838 < 2.2e-16 ***
Residuals         11516 212.83    0.02
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 8: ANOVA Table for Final Model M5

The model explains a substantial proportion of the variation in resale prices. The multiple R-squared is 0.8163, while the adjusted R-squared is 0.8162, indicating a strong overall fit. The residual standard error on the log scale is 0.1359, and the overall F-statistic of 5118 ($p < 2.2e\text{-}16$).

These confirms that the predictors collectively have a highly significant effect on resale prices. These performance metrics demonstrate that the model captures most of the systematic variation in resale prices while leaving relatively small unexplained residuals.




```
                        GVIF Df GVIF^(1/(2*Df))
log(floor_area_sqm) 2.263650  1        1.504543
lease_commence_date 1.309283  1        1.144239
town                1.320606  4        1.035373
flat_type_grouped   2.180453  1        1.476636
storey_grouped      1.222067  3        1.033989
```

Figure 9: MAC plots for Final Model M5

Model Assumption Checks (Final Model) (Figure 9):

- Residual patterns: Standardized residuals showed reasonable normality, with no major deviations observed. Residuals versus fitted values indicated no strong heteroscedasticity, though a few extreme residuals were present.
- Numerical predictors: Residuals for log-transformed floor area and lease commencement date showed mostly random scatter, supporting linearity

assumptions. Occasional standardized residuals exceeded ±3 but were limited in number and did not affect overall model stability.

- Categorical predictors: Grouping flat type and storey range improved residual stability. Boxplots by town, flat type, and storey group indicated that the reverse-cone pattern previously observed for storey range was mitigated. Median residuals were close to zero across all categories.
- Outliers / Influential points: 209 outliers were identified. No influential points were detected, indicating that extreme observations did not unduly affect coefficient estimates.
- Multicollinearity: GVIF values adjusted for degrees of freedom were all below 1.55, indicating low multicollinearity. Weak to moderate correlation between floor area and flat type was expected and did not compromise coefficient interpretation.

Overall, the final model provides a better fit than the initial model, as it more closely satisfies the underlying OLS assumptions and captures key determinants of resale price more accurately.

Interpreting the Final Model (M5):

Although the model predicts the logarithm of resale price, the coefficients can still be interpreted in terms of their impact on actual resale prices since the logarithm is a monotonic function. An increase in the predicted log resale price corresponds to a proportional increase in the resale price on the original scale.

Floor area has the strongest positive effect: larger flats command higher prices, with a 1% increase in floor area associated with an approximate 0.88% rise in resale price. Lease commencement date also has a positive and significant effect, implying that newer flats sell for higher prices—each additional year of age contributes to roughly a 0.85% increase in log resale price.

Location plays a substantial role. Relative to the Central region, resale prices are lower in the East (-0.205), North (-0.381), North-East (-0.301), and West (-0.328), reflecting regional price disparities consistent with locational accessibility and demand differences. Flat type has a smaller but significant effect, with Small-Medium flats priced slightly below Large flats (-0.028). Storey level also shows a clear gradient: resale prices increase steadily with floor height, and Very High floors enjoy the largest premium (0.260), suggesting that buyers value elevation for privacy and views.

Overall, the model confirms intuitive relationships—flats that are larger, newer, on higher floors, and centrally located achieve higher resale values. Transformations and grouping improved model interpretability while retaining key economic relationships.

Implications of the Results:

The findings from the final model (M5) offer practical insights for homeowners, policymakers, and property analysts. The strong positive relationship between floor area and resale price highlights buyers' continued preference for larger flats, suggesting that maintaining a diverse supply of unit sizes remains important for meeting housing needs. The positive effect of lease commence date confirms that newer flats command higher resale values, reflecting market sensitivity to lease decay. Regional and storey-level effects reveal how accessibility and flat attributes shape demand. Flats in central regions and on higher floors command price premiums, underscoring the influence of location, connectivity, and vertical comfort on housing values.