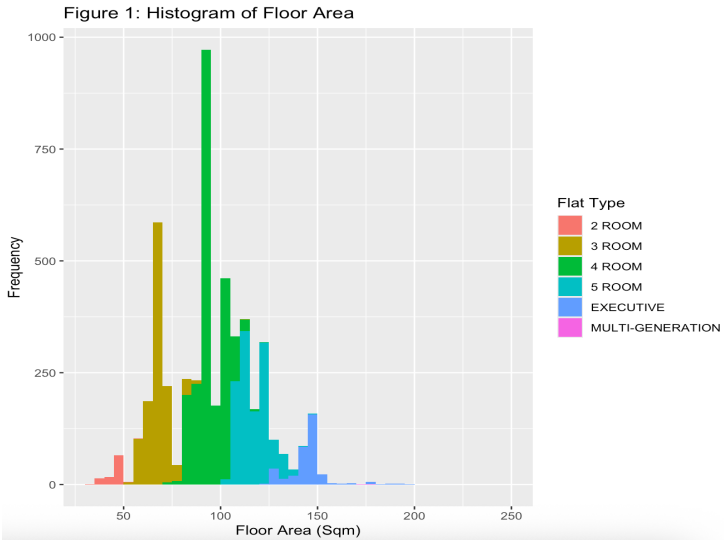
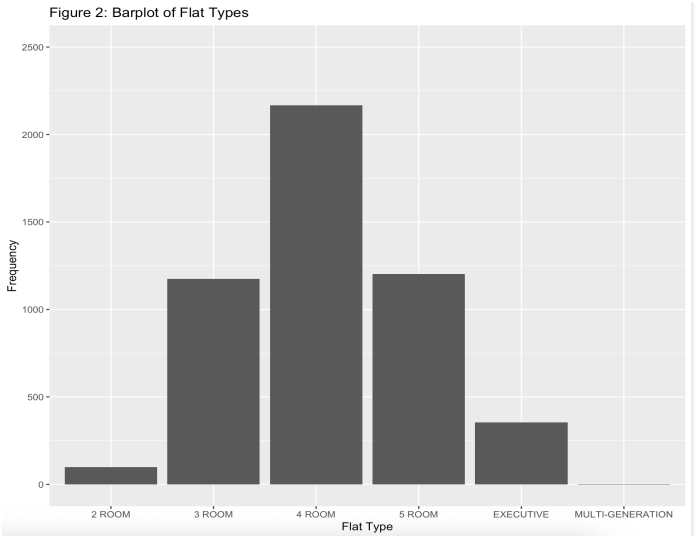
A0322981B Max Tan Kia Lok

Qn 1: From the dataset, the variables that are supposed to be treated as quantitative are **resale_price**, **remaining_lease** and **floor_area_sqm**. Remaining lease is written in both years and months, which we convert the months into years and express the variable in number of years.

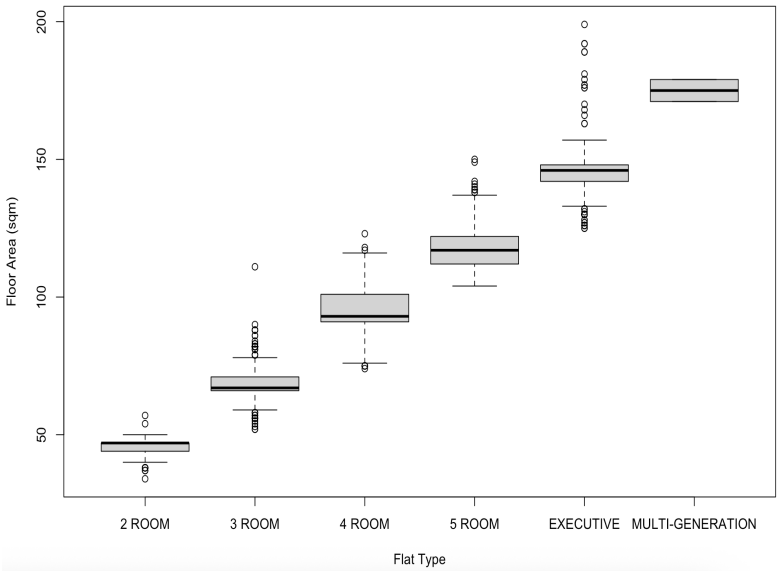


Qn 2: Figure 1 shows a histogram for floor area, which shows a multimodal distribution with multiple distinct peaks. The histogram is also right skewed. The highest peak is at the floor area of 90-95 sqm with a frequency of around 950. Other peaks include 65-70, 100-105, 120-125 and 145-150. Using colours the show the different flat types, we can observe that each flat type has floor area distribution at different parts of the histogram and are responsible for the different peaks in the histogram.

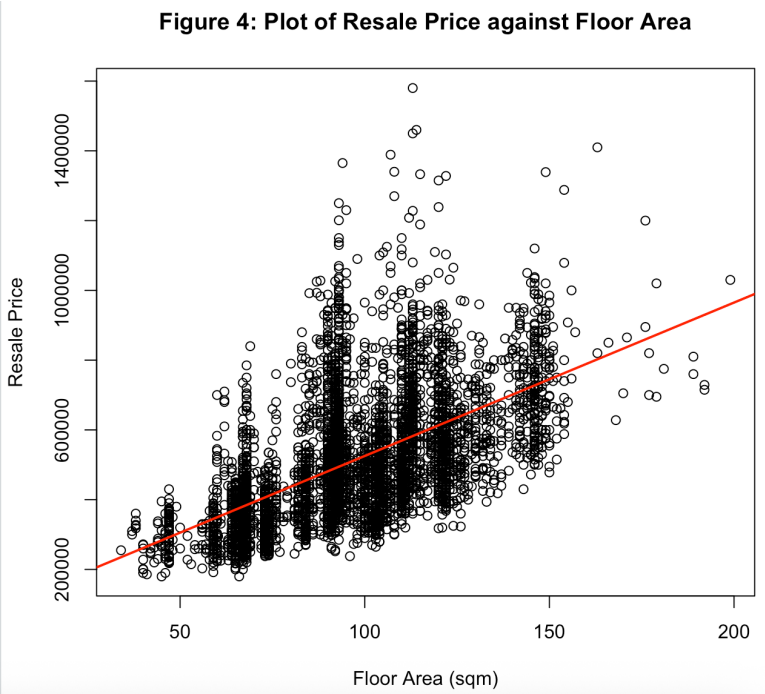


Qn 3: Figure 2 shows a bar plot for the different flat types. The modal category is 4 room flat, which accounts for about 2200 flats or 44% of all flats in the dataset. 3 Room flats are the next more frequent, followed by 5 room, executive, 2 room and finally multigeneration flat with a frequency of just 2.

Figure 3: Boxplot of Floor Area for each Flat Type



Qn 4: Figure 3 shows a boxplot of floor area for each flat type. The diagram shows a steady increase in the median from 2 room to multi-generation. 4 room and 5 room both have comparable and largest interquartile ranges (IQR) while 2 room has the smallest IQR. Every category except multi-generation contains outliers. 5 room contains only outliers that are greater than the max whisker reach while the other 4 have outliers on both sides of the median. 3 room has the most outliers. Executive has the largest range while multi-generation has the smallest range.

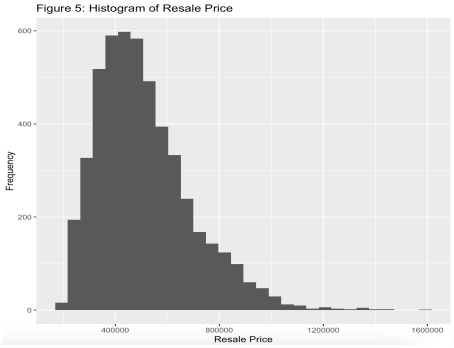


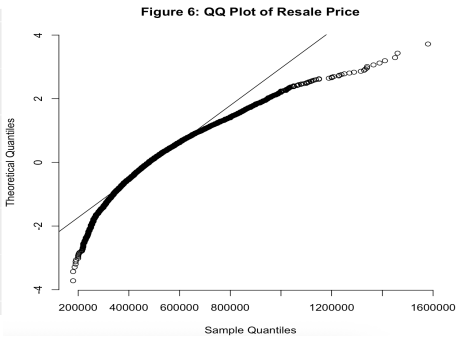
Qn 5: Figure 4 shows a scatterplot of resale price against floor area. There is a relationship with positive association between resale price and floor area. The correlation is 0.578 (3s.f.), which suggests a moderate positive relation between resale price and floor area. We note that the spread of points is unstable, with the middle having a greater spread compared to the left and right of the graph. Some points in the middle have larger resale prices than points on the right, which deviates from the trend that higher floor area will lead to higher resale prices.

Question 6

Exploratory Data Analysis

We first perform analysis on our response variable, which is resale price. Then, we want to find potential regressors by comparing them with resale price to see if there is a relationship.






```
        Shapiro-Wilk normality test

data:  df$resale_price
W = 0.94127, p-value < 2.2e-16
```

From figure 5, the histogram is right-skewed and not symmetrical, supported by the QQ Plot in figure 6 and the Shapiro-Wilk test. We know that **resale price** is not symmetrical and therefore likely not a good fit for a linear regression model. However, we will build our initial model with it first and verify whether transformation is needed after checking for the adequacy of the model.

For some of the variables, we perform data transformation to make the data more useable (Qn 1) and to group data categories together to reduce number of categories in a variable to avoid the model overfitting if the variable is used. For ordinal categorical variable "**month**" (which we note records the year and month of the data collection), we extract out the year and further group them into 3 groups (2017-2019, 2020-2022, 2023-2025) to reduce number of categories into a new variable "**year_group**" For ordinal categorical variable "**storey_range**", we note that the range for each category is the same, hence we decided to change it into a quantitative variable by taking the midpoint of the range and storing it as a number in "**average_storey**". For the nominal categorical variable "**town**", we adopt the division of Singapore into 3 main regions by the Urban

Redevelopment Authority (URA) under the new variable "region": Core Central Region (CCR), Rest of Central Region (RCR) and Outside Central Region (OCR).

Region	Towns
CCR	Central Area, Bukit Timah
RCR	Bishan, Queenstown, Toa Payoh, Geylang, Marine Parade, Kallang/Whampoa
OCR	Bedok, Bukit Batok, Bukit Merah, Hougang, Jurong West, Punggol, Sembawang, Sengkang, Woodlands, Yishun, Bukit Panjang, Pasir Ris, Serangoon, Ang Mo Kio, Chua Chu Kang, Clementi, Tampines, Jurong East

We decided to exclude "**block**" from the list of potential regressors since they are most likely noy going to yield any statistical association with the response variable. "**street name**", "**flat model**" and "**lease_commence_date**" also excluded since we felt that "**region**", "**flat type**" and "**remaining_lease_years**" already represented each variable well respectively. "**region**" and "**flat_type**" are broader categorizations while "**remaining_lease_years**" is the number of years left in the 99-year lease from its commencement date.

Categorical Variables

Figure 7: Resale Price by Year Group

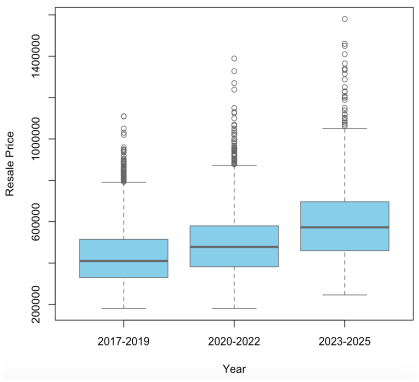
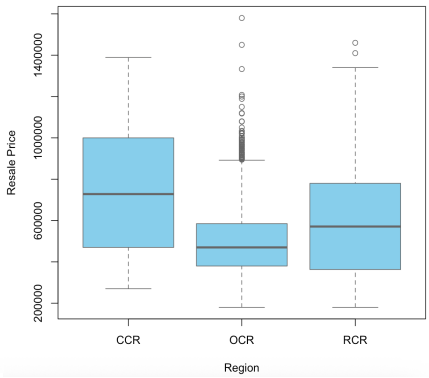


Figure 8: Resale Price by Region



From figure 3, 7 and 8, we can see that all data do show some sort of trend. From the ordinal categorical variable "**year**", we can see that the later the year, the higher the resale price. This may be attributed to factors like inflation that naturally causes prices to increase and higher demand for housing in general. From nominal categorical variable "**region**", we can see that CCR has the highest median resale price while OCR has the lowest. It's likely because CCR is considered 'prime' region for housing location hence houses can fetch a higher price there.

Numerical Variables

Figure 9: Resale Price against Remaining Lease Years

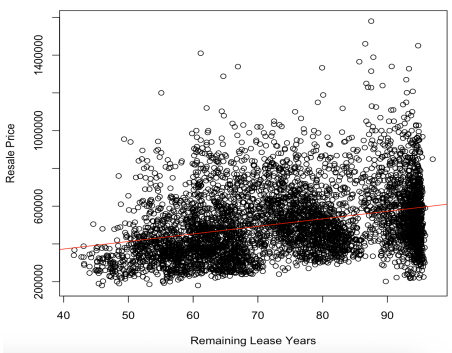
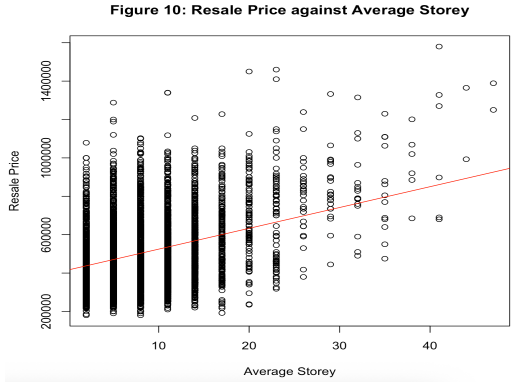


Figure 10: Resale Price against Average Storey



	resale_price	avg_storey	floor_area_sqm	remaining_lease_year
resale_price	1.0000000	0.349984048	0.578270944	0.3128522
avg_storey	0.3499840	1.000000000	-0.005528206	0.2597987
floor_area_sqm	0.5782709	-0.005528206	1.000000000	0.1207313
remaining_lease_year	0.3128522	0.259798653	0.120731258	1.0000000

From figure 4, 9 and 10, we can see clear trends of **resale price** having a positive relationship with **average storey**, **lease year** and **floor area**. However, we note that the correlation between resale price and the numerical variables are rather weak from the coefficient matrix, with floor area having the highest correlation at 0.578 (3s.f.). This suggests that transformations may be needed for the variables to have a stronger linear correlation with resale price.

The correlation matrix also shows that there are little to no correlation between the regressor variables, hence the regressors can be deemed to be linearly independent and we do not need to consider adding any interaction term between 2 regressors.

Initial Model M_i

Initial model M_i built in this report will be the **response variable resale price** with regressors **region**, **year**, **flat type**, **average storey**, **remaining lease years** and **floor area**. The reason for these regressors is because they show some trend with the response variable and some regressors can be considered as broader classifications of other variables (which are excluded due to reasons above) so that our model has less indicator variables.

Summary of M_i

```
Call:
lm(formula = df$resale_price ~ df$floor_area_sqm + df$remaining_lease_years +
    df$avg_storey + df$region + df$year_group + df$flat_type)

Residuals:
    Min      1Q  Median      3Q     Max
-267870  -65222  -10251   49458  648529

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                -66354.7    21110.5  -3.143  0.00168 **
df$floor_area_sqm            3647.6      204.2  17.860  < 2e-16 ***
df$remaining_lease_years     3413.2      113.1  30.179  < 2e-16 ***
df$avg_storey                6948.2      248.3  27.984  < 2e-16 ***
df$regionOCR              -278963.4    12622.7 -22.100  < 2e-16 ***
df$regionRCR              -123570.9    12986.7  -9.515  < 2e-16 ***
df$year_group2020-2022      64247.1     3242.1  19.816  < 2e-16 ***
df$year_group2023-2025     176539.9     3471.4  50.856  < 2e-16 ***
df$flat_type3 ROOM          63004.3    10971.0   5.743 9.86e-09 ***
df$flat_type4 ROOM          83467.1    14049.5   5.941 3.03e-09 ***
df$flat_type5 ROOM         104542.5    17711.8   5.902 3.82e-09 ***
df$flat_typeEXECUTIVE      154155.5    22828.6   6.753 1.62e-11 ***
df$flat_typeMULTI-GENERATION 191161.4   72939.6   2.621  0.00880 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 95400 on 4987 degrees of freedom
Multiple R-squared:  0.7219,    Adjusted R-squared:  0.7213
F-statistic:  1079 on 12 and 4987 DF,  p-value: < 2.2e-16
```

F-Test: Model is statistically significant due to extremely low p -value
T-Test: All variables have low p-values for their T-tests, showing that they are all significant on their own in the model.
R-squared: 0.7219 and 0.7213 for multiple and adjusted respectively. This means the model can explain about 72.2% of the response value in the dataset. It’s possible that the higher number of indicators caused the model to overfit, so we should aim to make later models simpler.

Figure 11: Standard Residual against Fitted Response

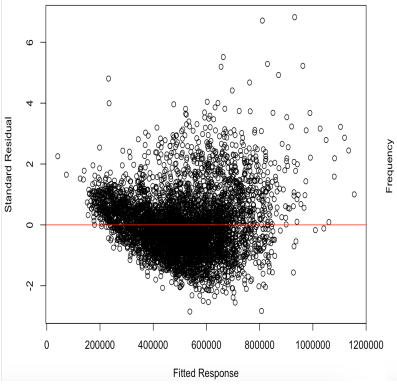


Figure 12: Histogram of Standard Residual

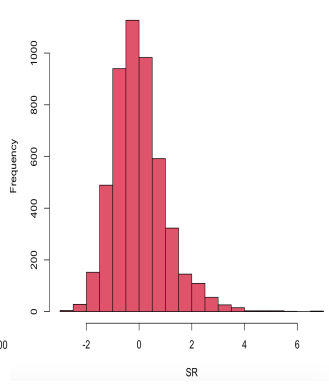


Figure 13: Q-Q Plot of Initial Model

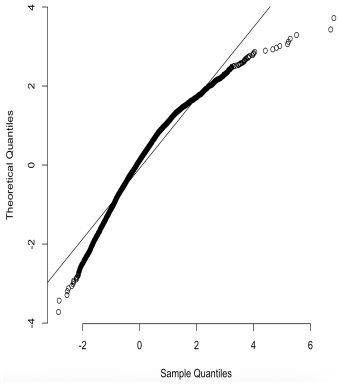
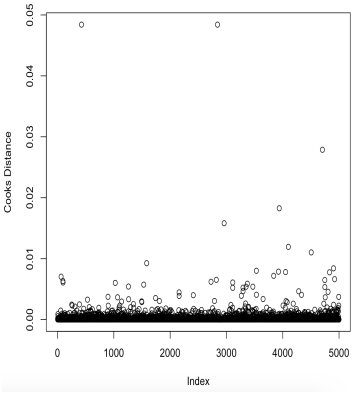


Figure 14: Cooks test for Initial Model



Residual plot: There is a funnel shape observed in figure 11, showing that the constant variance assumption is violated and suggests that the response variable should be transformed. Linearity is also violated since U-shaped is observed.

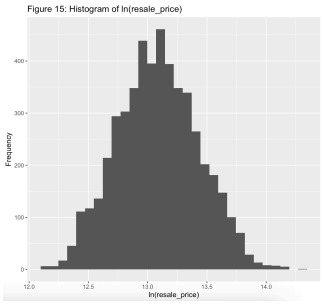
Normality: Model is heavily right-skewed evident from figure 13 QQ-Plot showing the model has a longer right tail and shorter left tail than normal, supported by figure 12 right-skewed histogram.

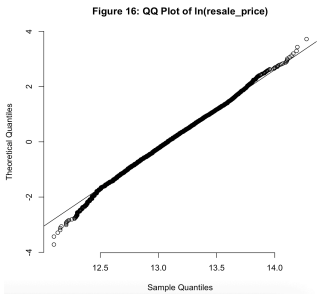
Outliers: There are numerous points in Standard Residual in figure 11 that are above 3, but figure 14 cook's distance test show no influential points, thus we do not need to investigate the outlier points.

Transformation of Variables

The evidence from the plots above strongly suggests transforming the response variable **resale price**. We have tried inverse, logarithm and square root and found that **ln(resale_price)** gives the best overall improvement in linearity between variables, with the main improvement being with floor area. We tried transformations on the numerical variables, but they either yielded no effects or led to worse correlation with the response variable. Lastly, we note from figure 3 that there is a relation between **floor area** and **flat type**. We decided to exclude **flat type** from the final model since floor area already captures the information flat type contains and excluding it will reduce the number of indicator variables, reducing the chance of the linear model overfitting. Below shows the correlation table after transforming **resale price** and its plots:

	resale_price	avg_storey	floor_area_sqm	remaining_lease_year
resale_price	1.0000000	0.318188874	0.625751369	0.3482373
avg_storey	0.3181889	1.000000000	-0.005528206	0.2597987
floor_area_sqm	0.6257514	-0.005528206	1.000000000	0.1207313
remaining_lease_year	0.3482373	0.259798653	0.120731258	1.0000000





Final Model M_f

In our final model, we use ln(resale_price) as the response variable which we see from figure 15 is symmetrical and supported by the QQ-plot in figure 16. For the final regressors, we decided to use **region**, **year**, **average storey**, **remaining lease years** and **floor area**.

Summary of M_f

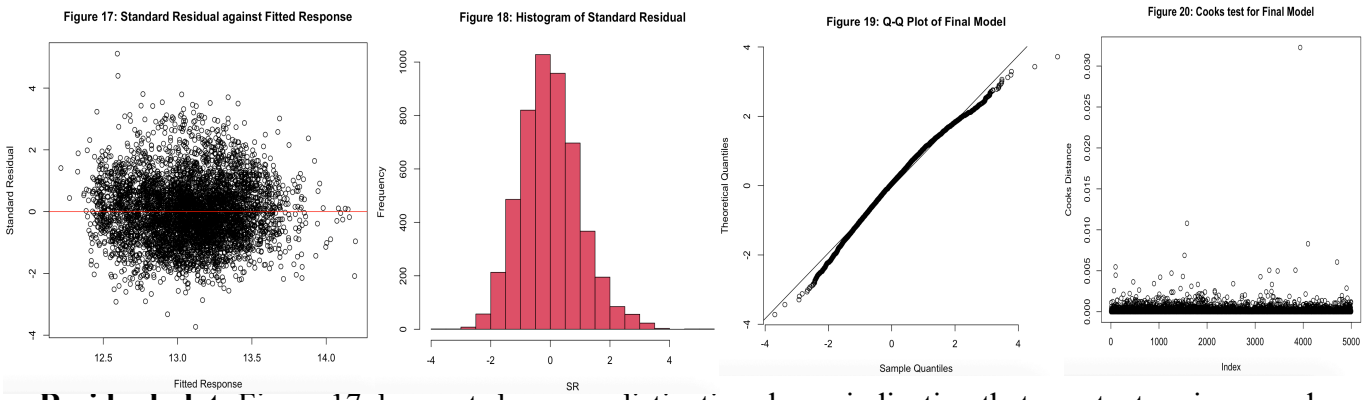
```
Call:
lm(formula = log(df$resale_price) ~ df$floor_area_sqm + df$remaining_lease_years +
    df$avg_storey + df$region + df$year_group)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60991 -0.11370 -0.00860  0.09949  0.87280

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              11.7883248  0.0265772 443.550  < 2e-16 ***
df$floor_area_sqm         0.0093707  0.0001009  92.851  < 2e-16 ***
df$remaining_lease_years  0.0072023  0.0001805  39.908  < 2e-16 ***
df$avg_storey             0.0112786  0.0004317  26.129  < 2e-16 ***
df$regionOCR             -0.4425789  0.0220439 -20.077  < 2e-16 ***
df$regionRCR             -0.1803572  0.0226892  -7.949 2.31e-15 ***
df$year_group2020-2022    0.1379934  0.0056588  24.386  < 2e-16 ***
df$year_group2023-2025    0.3518208  0.0060521  58.132  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1667 on 4992 degrees of freedom
Multiple R-squared:  0.7618,    Adjusted R-squared:  0.7614
F-statistic:  2281 on 7 and 4992 DF,  p-value: < 2.2e-16
```

F-Test: Model is statistically significant due to extremely low p -value
T-Test: All variables have low p-values for their T-tests, showing that they are all significant on their own in the model.
R-squared: 0.7618 and 0.7614 for multiple and adjusted respectively. This means the model can explain about 76.2% of the response value in the dataset. This is an improvement of 4% from our initial model, making our final model a better fit for the dataset given.



Residual plot: Figure 17 does not show any distinctive shape, indicating that constant variance and linearity assumptions are not violated, a big improvement compared to our initial model.

Normality: Model is slightly right-skewed as shown in the QQ-plot in figure 19 with a slighter longer right tail and slightly shorter left tail, supported by figure 18 slightly right-skewed histogram. It's a better outcome than our initial model.

Outliers: There are less outlier points above 3 from figure 17, and Cook's test also showed that they are not influential, hence nothing further needs to be done.

Evaluation and Interpretation of M_f

The indicator functions below define how different values of **region** and **year_group** influence the response variable, ln(resale_price), when the baseline values are **region** = **"CCR"** and **year_group** = **"2017-2019"** respectively.

$$I_1(Region = \text{"OCR"}) = \begin{cases} 1 \; if \; \text{"}OCR\text{"} \\ 0 \; if \; otherwise \end{cases} \qquad I_2(Region = \text{"RCR"}) = \begin{cases} 1 \; if \; \text{"}RCR\text{"} \\ 0 \; if \; otherwise \end{cases}$$

$$I_3(year\_group = \text{"2020} - \text{2022"}) = \begin{cases} 1 \; if \; \text{"2020} - \text{2022"} \\ 0 \; if \; otherwise \end{cases} \qquad I_4(year\_group = \text{"2023} - \text{2025"}) = \begin{cases} 1 \; if \; \text{"2023} - \text{2025"} \\ 0 \; if \; otherwise \end{cases}$$

From the summary table from R, we can evaluate our model's equation as such

$$\begin{aligned} \ln(resale\_price) &= 11.8 + 0.00937 \times floor\_area + 0.00721 \times remaining\_lease \\ &+ 0.0113 \times avg\_storey - 0.443 \times I_1(Region = \text{"OCR"}) \\ &- 0.180 \times I_2(Region = \text{"}RCR\text{"}) + 0.138 \times I_3(year\_group = \text{"2020} - \text{2022"}) \\ &+ 0.352 \times I_4(year\_group = \text{"2023} - \text{2025"}) \end{aligned}$$

We acknowledge some limitations with our linear model. Since **year_group** is one of the regressors of the final model and is a categorical variable, we note that this model will not be suitable to be used to predict resale prices using data that falls outside of the year range of 2017-2025. We also note that there are **towns** not represented in the dataset for all 3 regions, so using a data point from a town not contained in the dataset to predict may also not yield accurate results. The model seems to show that the further the house is from the CCR (OCR is further away compared to RCR), the lower the resale price. It also shows that the later the year, the higher the resale price as well. Lastly, the greater the size of the flat and the longer the remaining lease, the higher the resale price.

Overall, M_f is a rather simple, adequate and well-fitted for our linear regression model. Note that when using this final model to predict resale price, the actual **resale price** value will be calculated as such: $predicted\ resale\ price = e^{model\_output}$ since our response variable is **ln(resale_price)** in the final model.