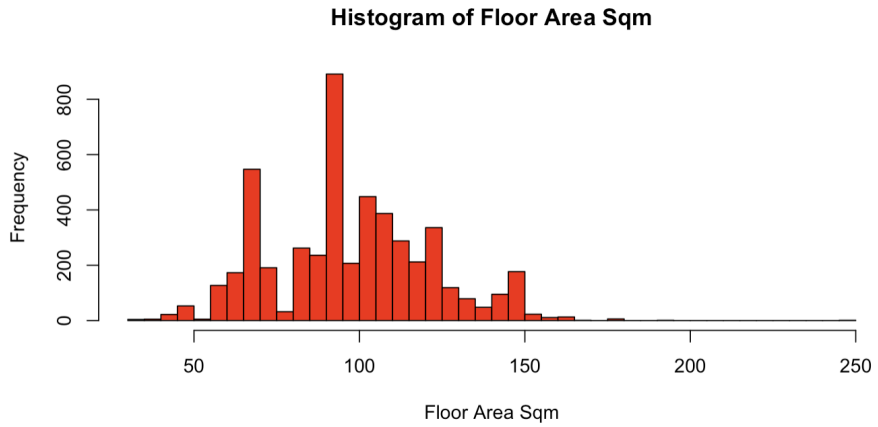
**Neo Zi Ning A0286410J**

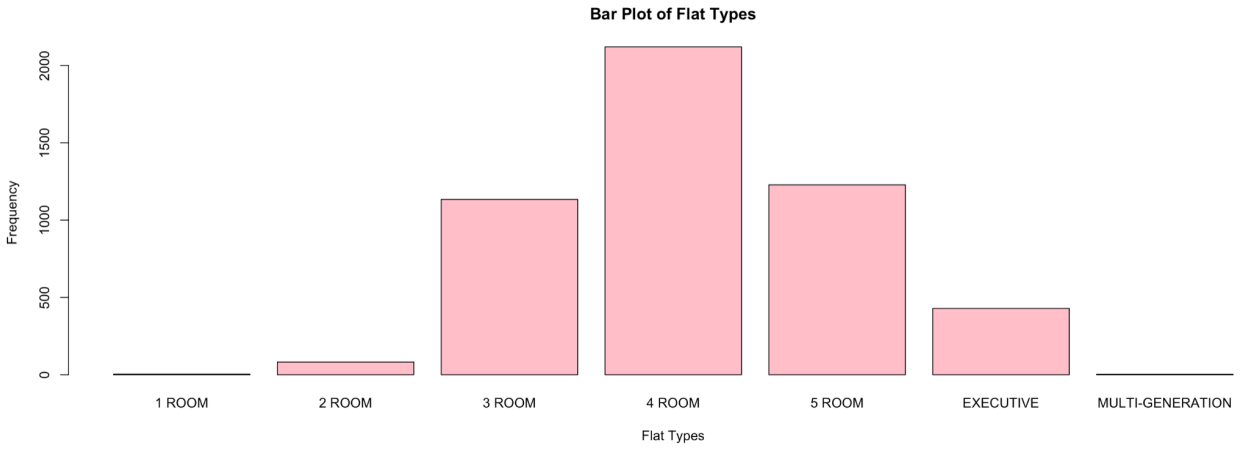
Q1.
The variables that are treated as quantitative variables are floor_area_sqm and resale_price.
I transformed resale_price by dividing by 10000 and transformed remaining_lease from years and months into the number of months.

Q2.



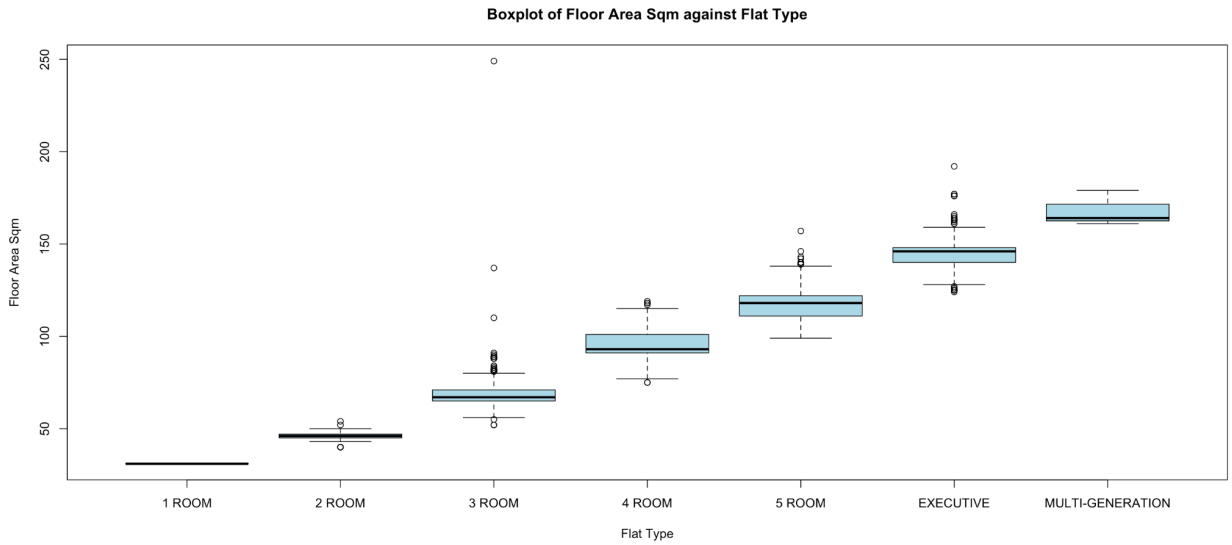

The histogram of Floor Area Sqm is slightly right-skewed since it has a longer right tail.
It is unimodal, and the data is mainly clustered between 70sqm and 130sqm. There are suspected outliers from 170sqm to 190sqm.

Q3.



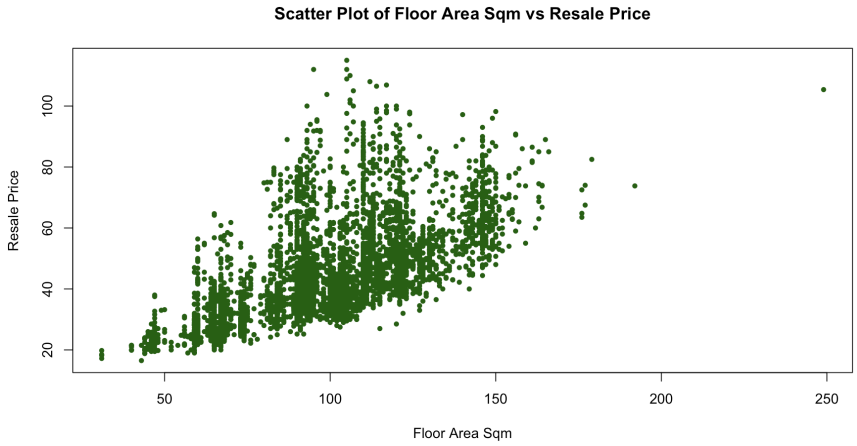

4 Room Flat Types have the highest frequency and are the most common flat type. Meanwhile, 1 Room Flat Types and Multi-Generation Flat Types have relatively low frequency and are much less common flat types.

Q4.




As the flat type changes from 1 Room to 2 Room to 3 Room to 4 Room to 5 Room to Executive and to Multi-Generation, the median of the floor area sqm increases.
The 1-room flat type and the Multi-Generation flat type show no outliers.
There are 2 outliers above the median and 1 outlier below the median for the 2-room flat type.
There are 68 outliers above the median and 3 outliers below the median for the 3-room flat type.
There are 4 outliers above the median and 2 outliers below the median for the 4-room flat type.
There are 10 outliers above the median and none below the median for the 5-room flat type.
There are 18 outliers above the median and 15 below the median for the Executive flat type.
1 Room flat type is symmetric, while 2 Room, 3 Room, 4 Room, 5 Room, Executive, and Multi-generation flat types appear to be right-skewed.
As the flat type changes from 1 Room to 2 Room, the interquartile range (IQR) increases.
With a further change to 3 Room, the IQR increases again. Moving to 4 Room, the IQR continues to increase. However, when the flat type changes to 5 Room, the IQR decreases, suggesting slightly more consistency in floor area compared to 4 Room flats. This decreasing trend continues with a further change to Executive, and again to Multi-Generation, where the IQR becomes even smaller.

Q5.



There is a positive relationship between Floor Area Sqm and Resale Price.
The trend of the scatter plot can be approximated with a straight line, especially for floor areas between 50sqm and 130 sqm. However, there is wider spread at floor areas above 130sqm.
The correlation value = 0.6292177, indicating that the relationship between Floor Area Sqm and Resale Price is positive and is moderately strong.

Q6.
model_0 = resale_price ~ floor_area_sqm + remaining_lease

| Regressor | P-Value | Included/Not Included in Final Model |
|---|---|---|
| floor_area_sqm | <2e-16 | Included |
| remaining_lease | <2e-16 | Included |

Since both regressors are significant to the model, as seen from the very small p-value, both regressors are included in the final model.

model_1 = resale_price ~ town + floor_area_sqm + remaining_lease
model_2 = resale_price ~ town + flat_type + floor_area_sqm + remaining_lease
model_3 = resale_price ~ town + flat_type + storey_range + floor_area_sqm + remaining_lease
model_4 = resale_price ~ town + flat_type + storey_range + floor_area_sqm + flat_model + remaining_lease

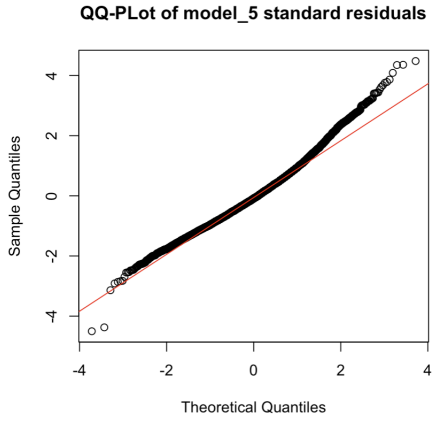
| Model | Regressor added to the model | Adjusted R-Squared | Change in Adjusted R-Squared compared to previous model | Included/Not Included in Final Model |
|---|---|---|---|---|
| model_1 | town | 0.8081 | 0.3602 | Included |
| model_2 | flat_type | 0.8124 | 0.0043 | Not Included |
| model_3 | storey_range | 0.8409 | 0.0285 | Included |
| model_4 | flat_model | 0.861 | 0.0201 | Included |

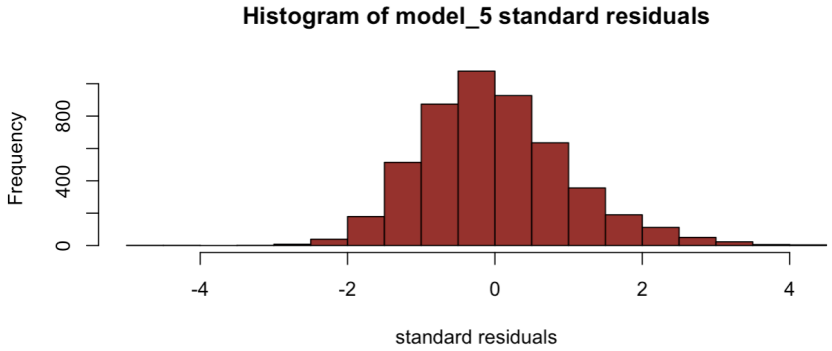
Reasons why some factors are not used in the models:

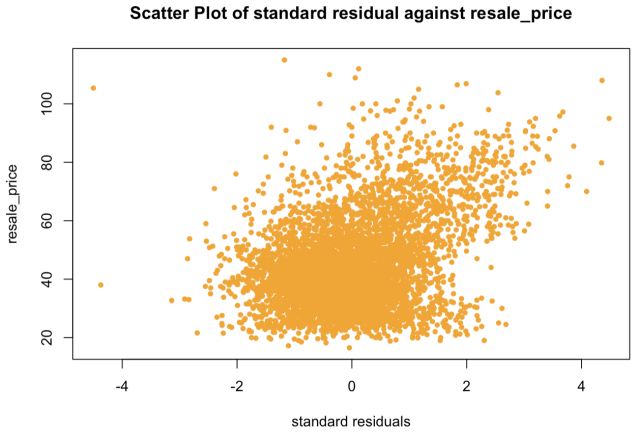
| Factor | Reason |
|---|---|
| month | it may introduce too many dummy variables (e.g. one for each month), leading to model complexity.<br>If month is highly correlated with other time-based variables (e.g. remaining lease), it may cause multicollinearity. |
| block | the area of the block is represented by the town variable and is more of an identifier rather than a meaningful predictor. |

| | |
|---|---|
| street_name | the area of the street is also represented by the town variable is also more of an identifier rather than a meaningful predictor. |
| lease_commencement_date | it is already represented by remaining_lease and adding this factor might lead to multicollinearity. |

model_5 = resale_price ~ town + storey_range + floor_area_sqm + flat_model + remaining_lease



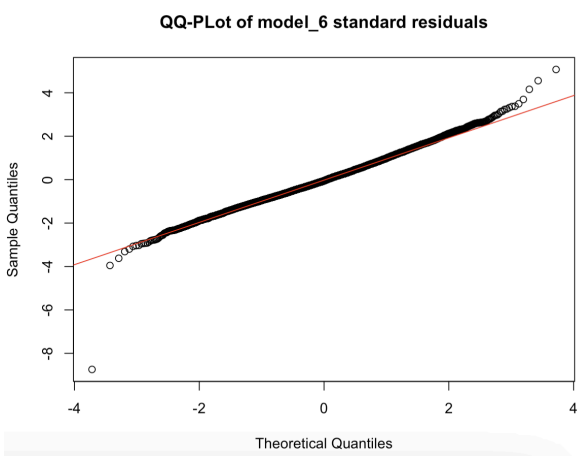



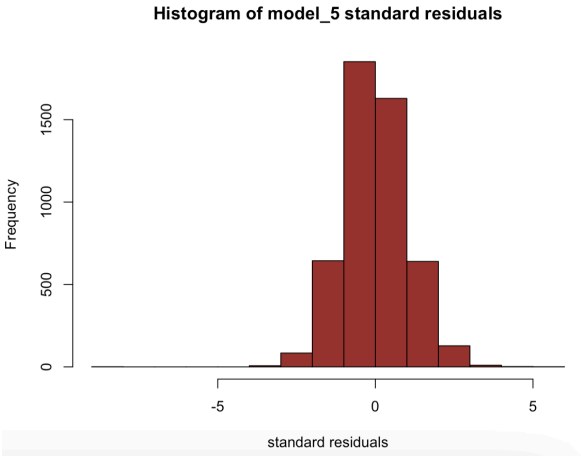



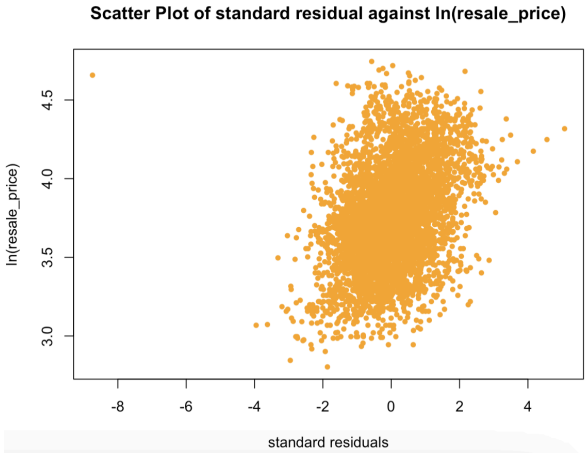

F-statistic p-value: < 2.2e-16

As we can see from the QQ-plot and histogram, the model violates the normality assumption based on the heavy right-tails. The scatter plot also has a funnel shape, and hence the model violates the constant variance assumption. I checked for outliers which has SR < -3 or SR > 3 and influential points using Cook's distance, and there are 36 outliers but 0 influential points. As seen from the F-test's small p-value, the model is significant. Hence, I transformed the response variable, resale_price, to fix the constant variance and normality assumption without removing any data points.

model_6 = ln(resale_price) ~ town + storey_range + floor_area_sqm + flat_model + remaining_lease

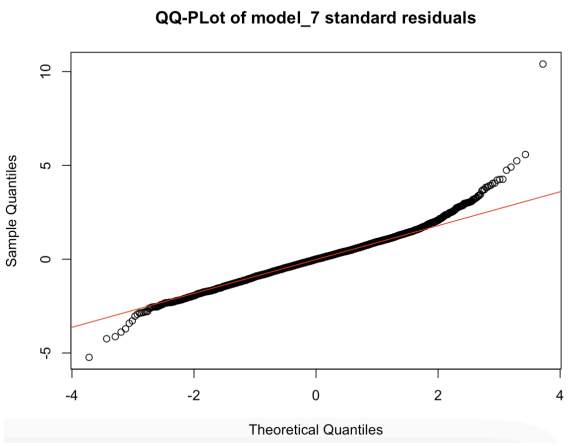


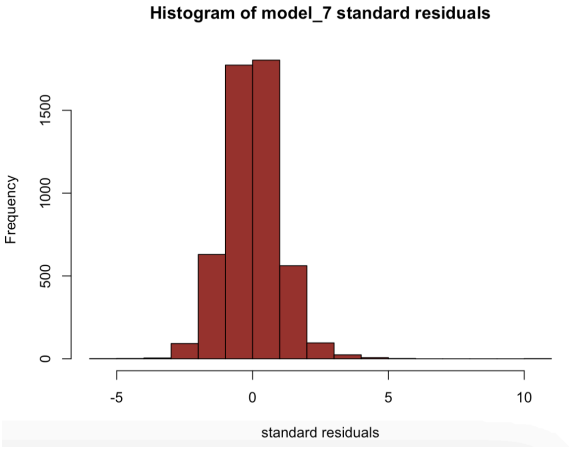



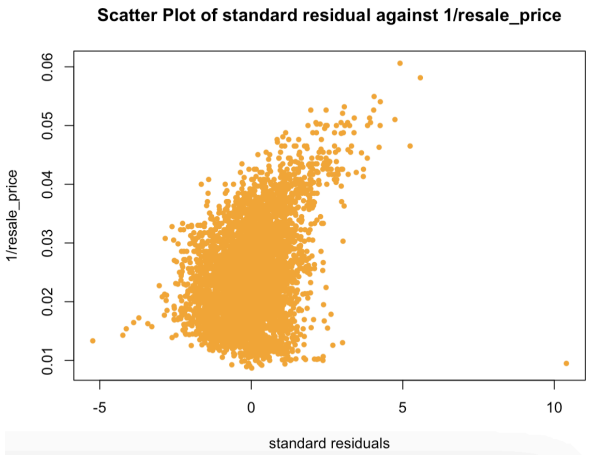




F-statistic: < 2.2e-16

model_7 = 1/resale_price ~ town + storey_range + floor_area_sqm + flat_model + remaining_lease

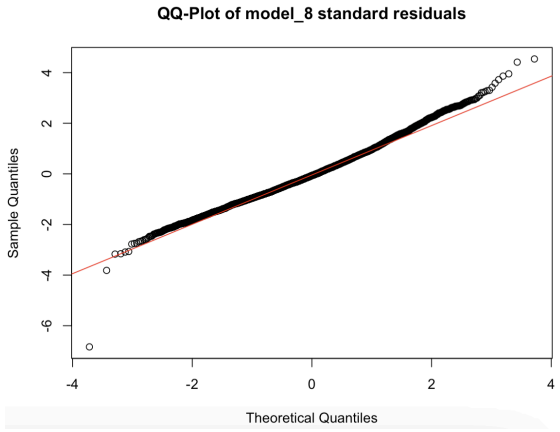


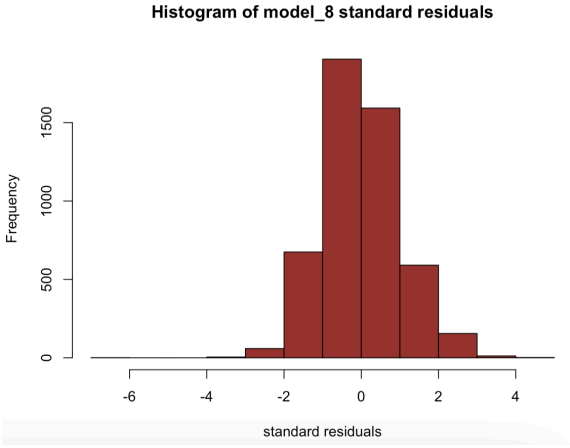



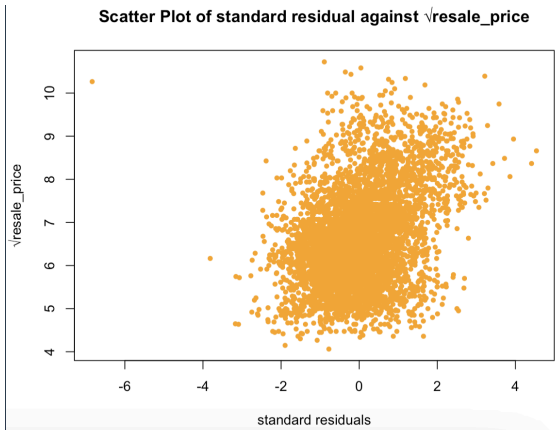



F-statistic: < 2.2e-16

model_8 = √resale_price ~ town + storey_range + floor_area_sqm + flat_model + remaining_lease




F-statistic: < 2.2e-16

Based on the QQ-plots and histograms of model_6, model_7, and model_8, the standardized residuals of model_8 most closely follow a normal distribution. This is evident from its QQ-plot, which shows the lightest tails, and its histogram, which is the most bell-shaped among the three models.
Additionally, from the scatter plots of model_6, model_7, and model_8, the scatter plot of model_8 is the least funnel-shaped.
Since all 3 models are statistically significant as seen from the small p-values of the F-statistic, the chosen final model is model_8 as it follows the assumptions of constant variance, normality, and linear relationship most closely. model_8 also has a value of 0.8724 for multiple R-squared, showing that the regressors explain 87.24% of the resale_price.