MH3511 DATA ANALYSIS WITH COMPUTER

GROUP PROJECT

Analysis of Singapore Resale Flat Prices

Sharvini D/O Veera Kumaran	**U2210544J**
Ivy Lui Xiao Qing	**U2340747F**
Lim Jia Le	**U2340936L**
Chew Xin Cong	**U2340215A**
Nguyen Minh Quang	**U2322454E**

Abstract:
With Singapore's thriving real estate market, HDB resale flats remain a key housing option for many residents. The resale prices of these flats fluctuate due to various factors, making it a subject of interest for home buyers, sellers and policymakers. Hence, we would like to examine the relationship between resale flat prices and the attributes of that flat and identify its significance.

Content Page

1. Introduction

In our project, a dataset containing the resale flat prices from 2023 to 2024 is used, with other variables such as floor area, flat type, town area, range of floors and remaining lease period. Based on this dataset, we seek to answer the following questions arising from the characteristics of reselling prices for HDB flats in Singapore:

1. Does HDB resale flat price depend on floor area (sqm)?
2. Does HDB resale flat price depend on the range of floors at which the flat is situated?
3. Does HDB resale flat price depend on the flat type (e.g., 3-room, 4-room, 5-room)?
4. Does HDB resale flat price depend on the town where the flat is located?
5. Does HDB resale flat price depend on the remaining lease of the flat?
6. Is there one factor that has a greater impact on salary compared to the others?

This report will cover the data descriptions and analysis using R language. For each of our research objectives, we performed statistical analysis and drew conclusions in the most appropriate approach, together with explanations and elaborations.

2. Data Description

The dataset, “sg-resale-flat-prices-2017-onwards.csv”, is obtained from Kaggle, an online data science community and contains 11 variables and more than 180000 observations. It includes records of resale flats in Singapore from 2017 onwards. The original data was collected from data.gov.sg. In this project, we choose the data only from 2023 - 6/2024 because between 2017 and 2022, due to the significant impact of COVID 19, HDB resale flat prices increased significantly. According to The Straits Times, the COVID-19 outbreak and its effects considerably increased the price of HDB resale properties. This increase was caused by supply shortages, low lending rates, and policy changes. We chose to analyze data starting from January 2023 to minimize this external influence and better understand how the factors in the dataset affect resale prices.

Before starting our analysis, we first cleaned our dataset to ensure that:

- ➔ Missing values & duplicated records were removed: There are no missing values, but we identified 576 duplicate rows, which were subsequently removed.
- ➔ Rows where the transaction date from before 2023 were removed.
- ➔ Convert storey range into a numerical category :
    - ◆ Ground Floor(1-3)
    - ◆ Ground-Low Floor(4-6)
    - ◆ Low-Medium Floor(7-12)
    - ◆ Medium-High Floor(12-50)
- ➔ Convert the values in ‘remaining lease’ to year only
- ➔ Columns such as "block", "street_name", “flat_model”, “lease_commence_date” and "month", as they are not needed for the analysis.

The resulting dataset contains 37845 observations of 6 variables:

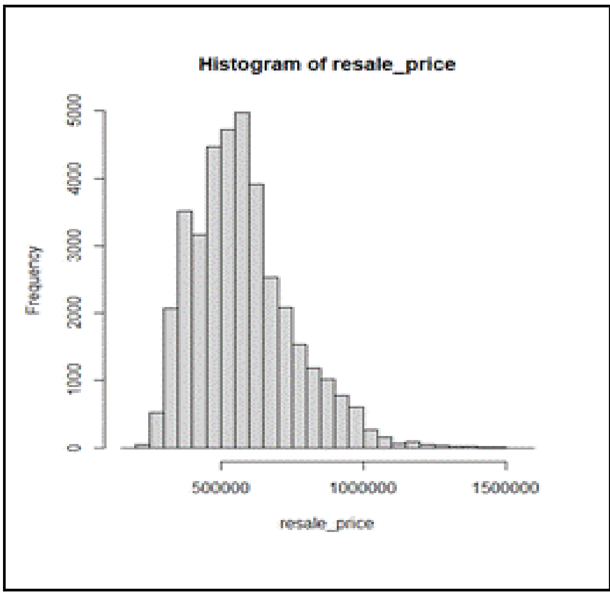
1. town : Location of HDB flats
2. flat_type : Indication of the number of rooms in a HDB flat
3. floor_area_sqm : Area of HDB flat
4. resale_price : Resale price of a HDB flat
5. floor_category : Indication of storey range from Ground to Medium-High Floor
6. remaining_lease_years : Remaining lease years for HDB flat

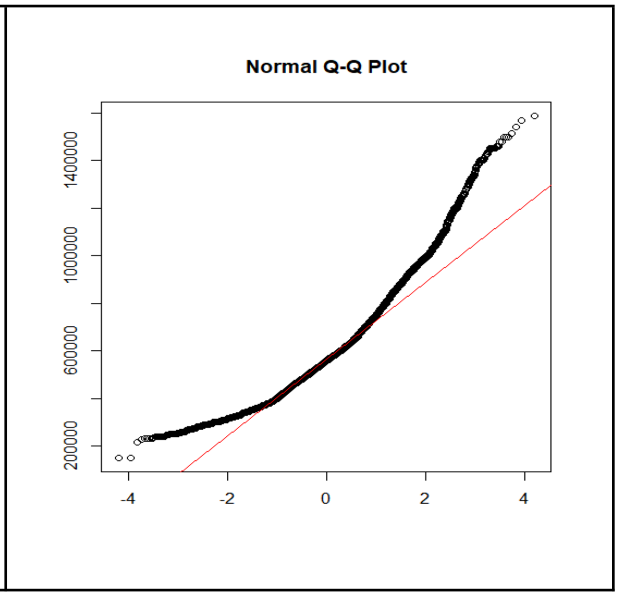
3. Description and Cleaning of Dataset

In this section, we shall look into the data in more detail. Each variable is investigated individually to look for possible outliers, and/or to perform a transformation to avoid highly skewed data.

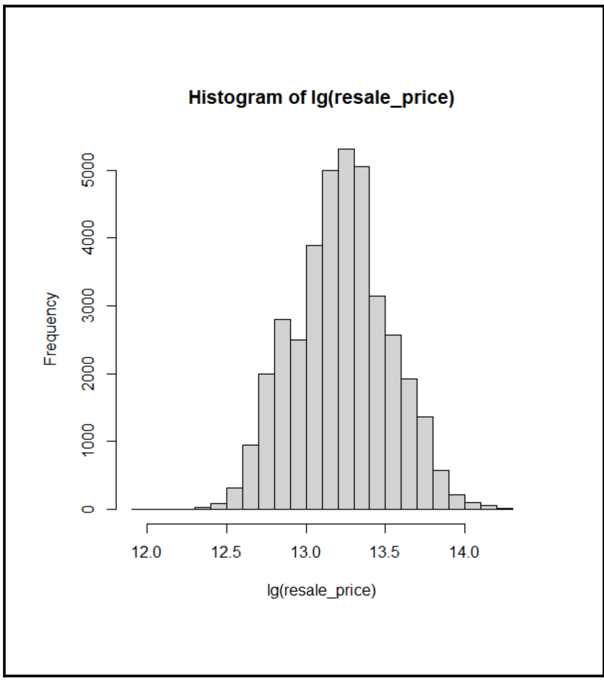
3.1 Summary Statistics For the Main Variable of Interest, *resale_price*

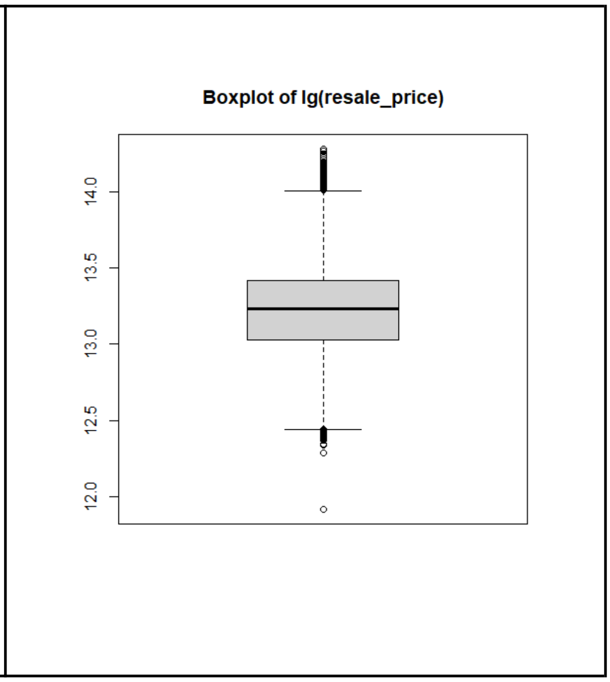
The following plots illustrate the overall distribution of the variable *resale_price*.



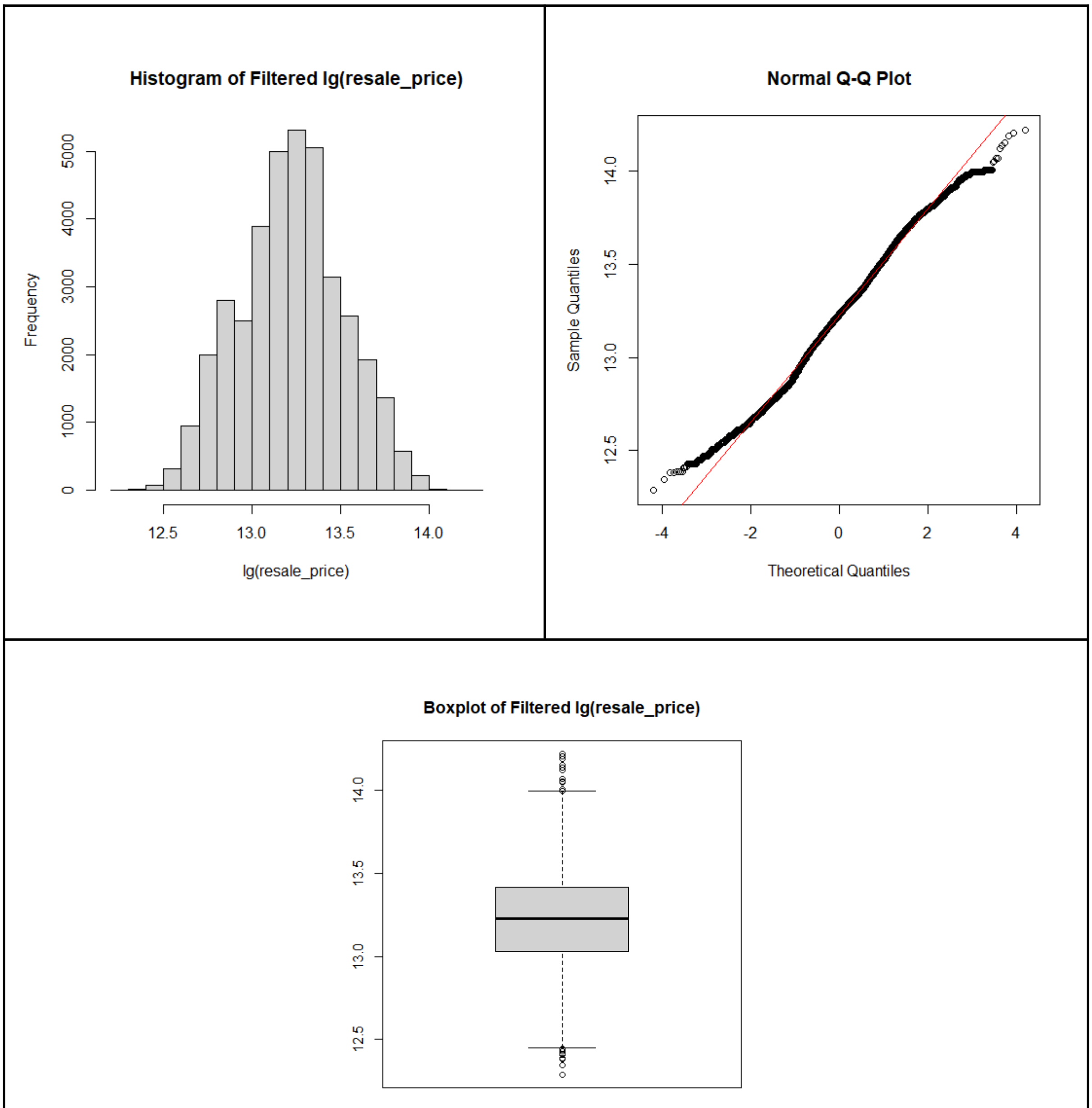


From both plots, it is evident that the main variable *resale_price* is right-skewed. Therefore, to normalize the variable, we applied log-transformation.





After applying log-transformation, the variable appears to be normally distributed. However, from the boxplot of *lg(resale_price)*, it appears that there are many outliers at both tails. Upon further investigation, the main reason for some data being on the right tail is that their *floor_type* are of the higher-ends (4-5 Room, Executive), large *floor_area_sqm* (>150 sqm) and high *remaining_lease_years* (>60 years). Similar explanation applies to the left tail. Therefore, we removed data that does not satisfy these conditions. For example, a 3 Room flat with 120 sqm and 58 years of remaining lease, lying on the right tail will be removed.



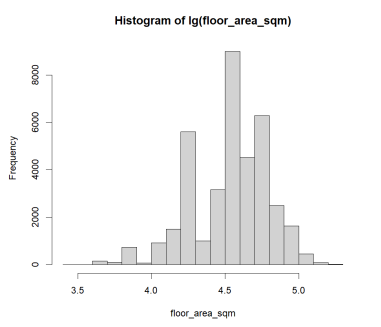
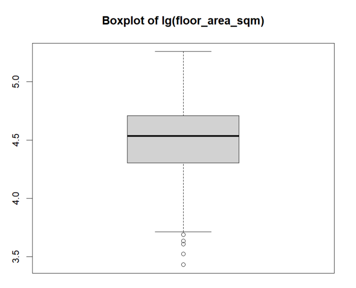

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.29   13.03   13.23   13.22   13.42   14.22
```

3.2 Summary Statistics For other Variables

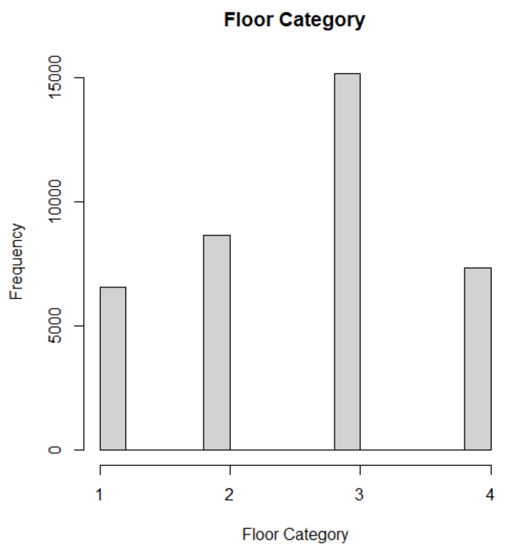
The histogram, the boxplot, the transformation applied and the outliers removed from the variables are tabulated in the following subsections.

- Floor Area, Floor Category, Flat Type, Town & Remaining Lease

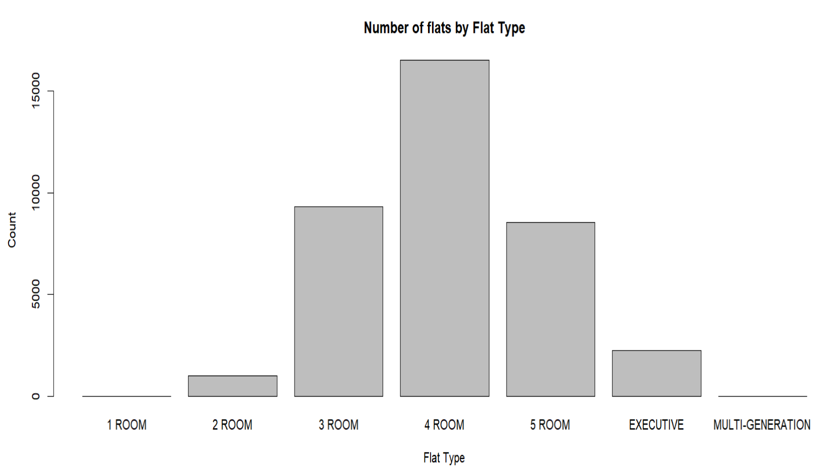
3.2.1 Floor Area

		• The log-transformation (base e) is applied • Appears to be well distributed, hence no data points removed.

3.2.2 Category of range of floors (*floor_category*)

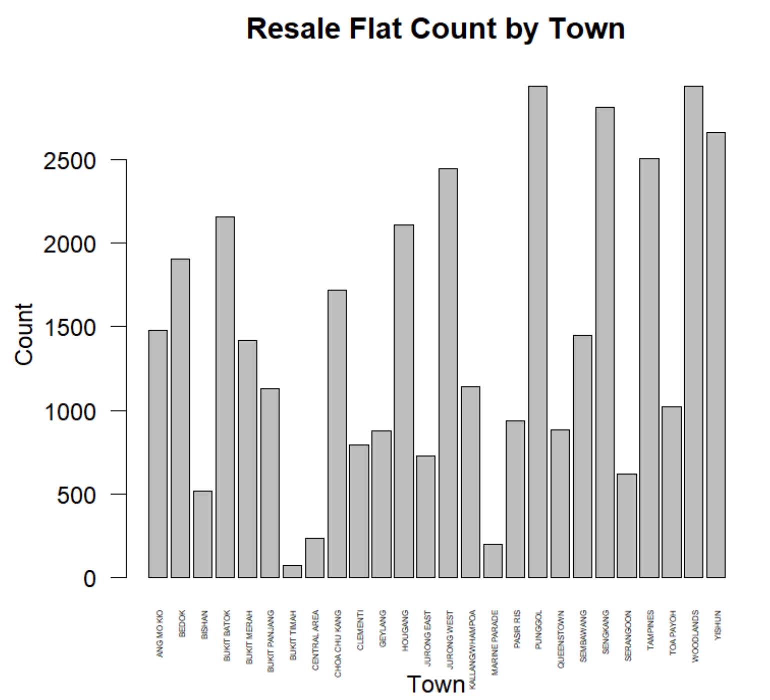
	• Appears to be well-distributed. • No changes required.

3.2.3 Primary flat_types in HDB flat



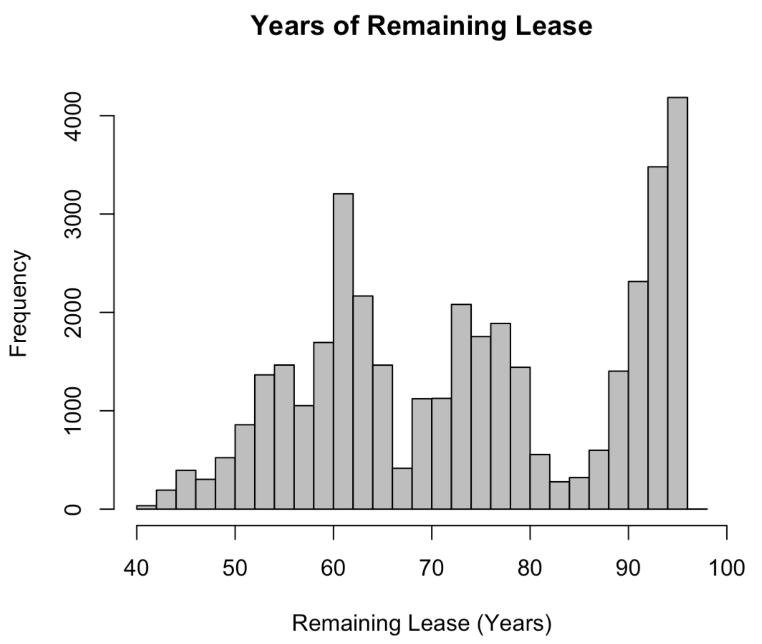
- The values from 1 Room and Multi-Generation will be removed as these flats have very low counts (1 and 7 respectively).
- There is a huge difference in the number of flats between flat types. The majority of flats are in the "4 ROOM", followed by 3, 5 ROOM. The count of 2 ROOM and EXECUTIVE are noticeably lower

3.2.4 Location of HDB flats (Town Area)



- No outlying value of *town* is removed.
- Punggol and Sengkang shows the highest number of resale flat count.
- Bukit Timah shows the lowest number of resale flat count.

3.2.5 Years of Remaining Lease

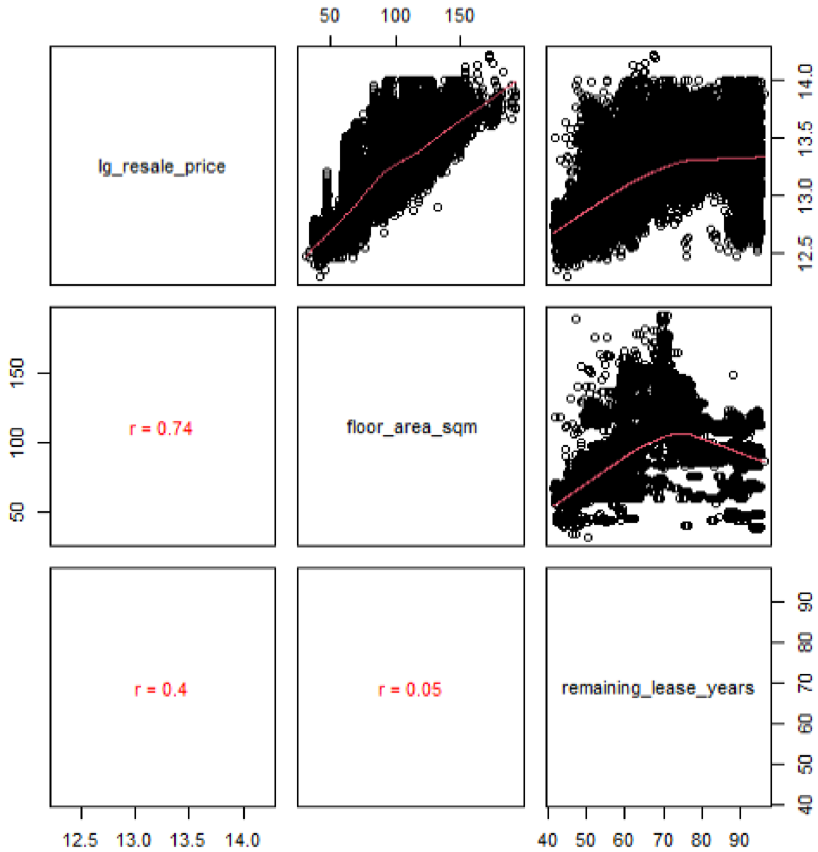
	● No outliers removed; all lease are retained ● Extracted years(integer part) and months(fractional part) ● Combined into decimal format (e.g., *"61 years 04 months" → 61.33*).

3.3 Final Dataset Analysis

Based on the above analysis, the final dataset for further testing will be stored in the variable "filtered_df$lg_resale_price" with all the unwanted variables and outliers for the variables of interest removed.

4. Statistical Analysis

4.1 Correlations between *lg_resale_price* and other Continuous Variables



From the correlation plot, we can observe that:

- *lg_resale_price* and *floor_area_sqm* are quite strongly and positively correlation (r = 0.74)
- *lg_resale_price* and *remaining_lease_years* are positively correlation (r = 0.40)
- *floor_area_sqm* and *remaining_lease_years* have almost no linear correlation (r = 0.05)

4.2 Statistical Tests

4.2.1 Relation between *resale_price* and *floor_area_sqm*

In this section we determine whether the resale price of a flat depends on the floor area of that flat. We perform a simple linear regression between *lg(resale_price)* and *floor_area_sqm.*

```
Call:
lm(formula = filtered_df$lg_resale_price ~ filtered_df$floor_area_sqm)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.65873 -0.13926 -0.03943  0.09922  0.84438 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                1.235e+01  4.227e-03  2922.4   <2e-16 ***
filtered_df$floor_area_sqm 9.124e-03  4.311e-05   211.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1981 on 37680 degrees of freedom
Multiple R-squared:  0.5431,	Adjusted R-squared:  0.5431 
F-statistic: 4.479e+04 on 1 and 37680 DF,  p-value: < 2.2e-16
```

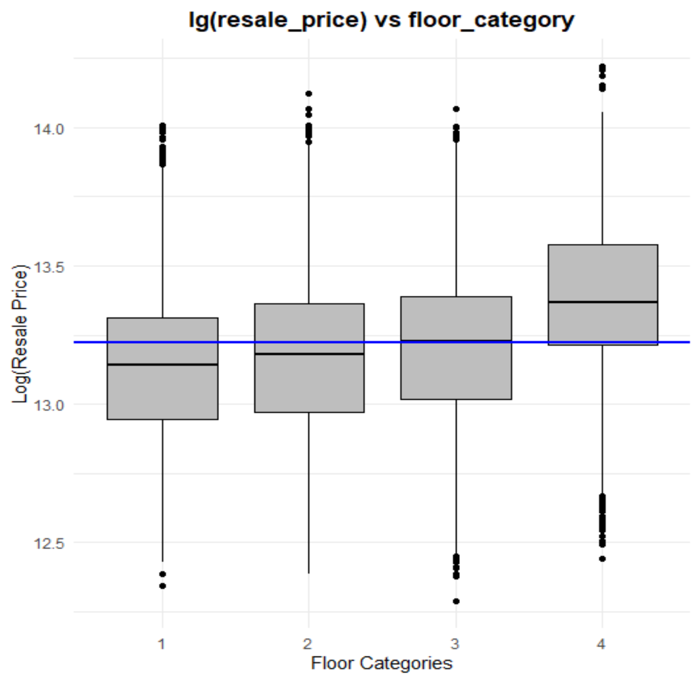
According to the linear regression model, the intercept(12.35) represents the predicted *lg_resale_price* when the *floor_area_sqm* is 0 and the slope(suggests that for each additional metre of floor area, the *lg_resale_price* increases by 0.009124.

The p-values (< 2.2e-16) for both intercept and the *floor_area_sqm* coefficient indicate that they are highly significant as it is lower than the significant level(0.05).

Furthermore, t-values (2922.4 and 211.6) are very large, confirming that *floor_area_sqm* has a strong influence on the *lg_resale_price*.

4.2.2 Relation between *resale_price* and *floor_category*

In this section, we aim to determine whether different floor categories (1-4) have a significant effect on the log-transformed flat resale price.



From the boxplot, The log(resale price) across the 4 floor categories are quite similar. Additionally, since *floor_category* is an independent categorical variable with multiple levels and the dependent variable *lg_resale_price* is continuous, an one-way ANOVA (Analysis of Variance) test will be conducted to determine the relation between *floor_category* and *resale_price*.

The mean of *lg_resale_price* of the 4 floor categories will be represented by μ_1, μ_2, μ_3 and μ_4 respectively. To carry out ANOVA, we carry out the following hypothesis test:

- **Null Hypothesis (H_0):** $\mu_1 = \mu_2 = \mu_3 = \mu_4$
- **Alternative Hypothesis (H_1):** Not all μ_i's are equal ($\mu_i \neq \mu_j$, for some i and j)

```
> aov(filtered_df$lg_resale_price~factor(filtered_df$floor_category))
Call:
   aov(formula = filtered_df$lg_resale_price ~ factor(filtered_df$floor_category))

Terms:
                factor(filtered_df$floor_category) Residuals
Sum of Squares                            248.9938 2988.2371
Deg. of Freedom                                  3     37678
```

```
> summary(aov(filtered_df$lg_resale_price~factor(filtered_df$floor_category)))
                                      Df Sum Sq Mean Sq F value Pr(>F)
factor(filtered_df$floor_category)     3    249   83.00    1047 <2e-16 ***
Residuals                          37678   2988    0.08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a high F-value of **1047** and a p-value of **~ 0** , which is lesser than α (level of significance) **= 0.05**, H_0 is rejected. This indicates that *floor_category* has a statistically significant effect on *lg_resale_price*.

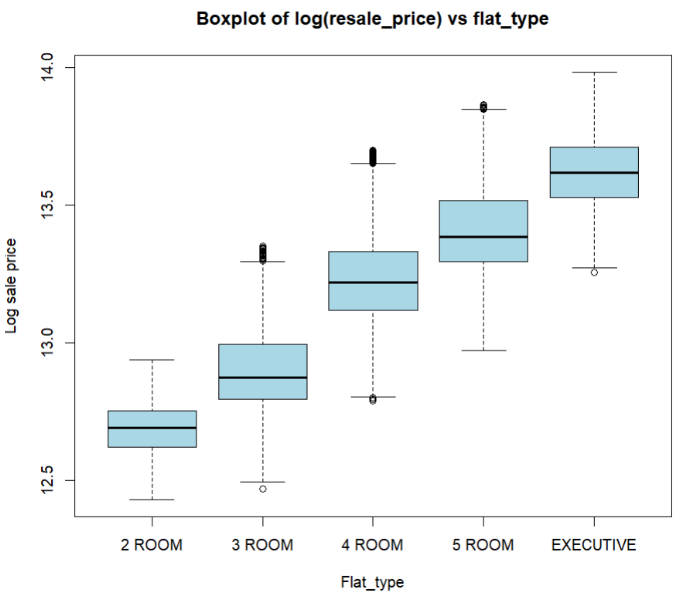
Following the rejection of H_0, further analysis is done using Pairwise T-test to compare the log-transformed resale prices between the 4 different floor categories and identify which specific pairs of floor categories differ from each other.

```
> pairwise.t.test(filtered_df$lg_resale_price, filtered_df$floor_category, p.adjust.method='none')

        Pairwise comparisons using t tests with pooled SD

data:  filtered_df$lg_resale_price and filtered_df$floor_category

  1      2      3
2 <2e-16 -      -
3 <2e-16 <2e-16 -
4 <2e-16 <2e-16 <2e-16
```

Based on the p-values obtained from the Pairwise T-test, it is evident that every pair of floor categories has a statistically significant difference in their log-transformed resale price. This suggests that *floor_category* has a clear and strong influence on *lg_resale_price*.

4.2.3 Relation between *resale_price* and *flat_type*

In this section, we investigate whether the resale price is influenced by the flat type of HDB.



Looking at the graph, we can see that when the number of rooms increases, the median log resale price also grows. This suggests that, on average, larger flats (with more rooms) command higher resale prices.

The 3-, 4-, and 5-room categories exhibit broader distributions. This could reflect a larger demand in those types of flat with prime locations, renovations or unique features.

```
> summary(aov(filtered_df$lg_resale_price ~ factor(filtered_df$flat_type)))
                                 Df Sum Sq Mean Sq F value Pr(>F)    
factor(filtered_df$flat_type)     4   1888     472   13208 <2e-16 ***
Residuals                     37669   1346       0                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The result from the ANOVA test with F value(13208) and p value of ~ 0 < 0.05 confirms that Flat type has a significant effect on resale price

```
> t.test(filtered_df$lg_resale_price[filtered_df$flat_type == "3 ROOM"], filtered_df$lg_resal
e_price[filtered_df$flat_type == "2 ROOM"], alternative = "greater")

	Welch Two Sample t-test

data:  filtered_df$lg_resale_price[filtered_df$flat_type == "3 ROOM"] and filtered_df$lg_resa
le_price[filtered_df$flat_type == "2 ROOM"]
t = 56.409, df = 1729.5, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.2263311       Inf
sample estimates:
mean of x mean of y 
 12.92047  12.68734 
```

```
> t.test(filtered_df$lg_resale_price[filtered_df$flat_type == "4 ROOM"], filtered_df$lg_resal
e_price[filtered_df$flat_type == "3 ROOM"], alternative = "greater")

	Welch Two Sample t-test

data:  filtered_df$lg_resale_price[filtered_df$flat_type == "4 ROOM"] and filtered_df$lg_resa
le_price[filtered_df$flat_type == "3 ROOM"]
t = 135.51, df = 20199, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.340642      Inf
sample estimates:
mean of x mean of y 
 13.26530  12.92047 
```

```
> t.test(filtered_df$lg_resale_price[filtered_df$flat_type == "5 ROOM"], filtered_df$lg_resal
e_price[filtered_df$flat_type == "4 ROOM"], alternative = "greater")

	Welch Two Sample t-test

data:  filtered_df$lg_resale_price[filtered_df$flat_type == "5 ROOM"] and filtered_df$lg_resa
le_price[filtered_df$flat_type == "4 ROOM"]
t = 64.233, df = 19703, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.1541496       Inf
sample estimates:
mean of x mean of y 
  13.4235   13.2653 
```

```
> t.test(filtered_df$lg_resale_price[filtered_df$flat_type == "EXECUTIVE"], filtered_df$lg_re
sale_price[filtered_df$flat_type == "5 ROOM"], alternative = "greater")

	Welch Two Sample t-test

data:  filtered_df$lg_resale_price[filtered_df$flat_type == "EXECUTIVE"] and filtered_df$lg_r
esale_price[filtered_df$flat_type == "5 ROOM"]
t = 57.55, df = 4285.5, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.1963349       Inf
sample estimates:
mean of x mean of y 
 13.62561  13.42350 
```

All four p-values in those pairwise comparisons are close to zero, indicating strong evidence that the mean resale price follows the order: Executive > 5-room > 4-room > 3-room > 2-room. This supports the conclusion that larger flat types generally command higher resale prices.

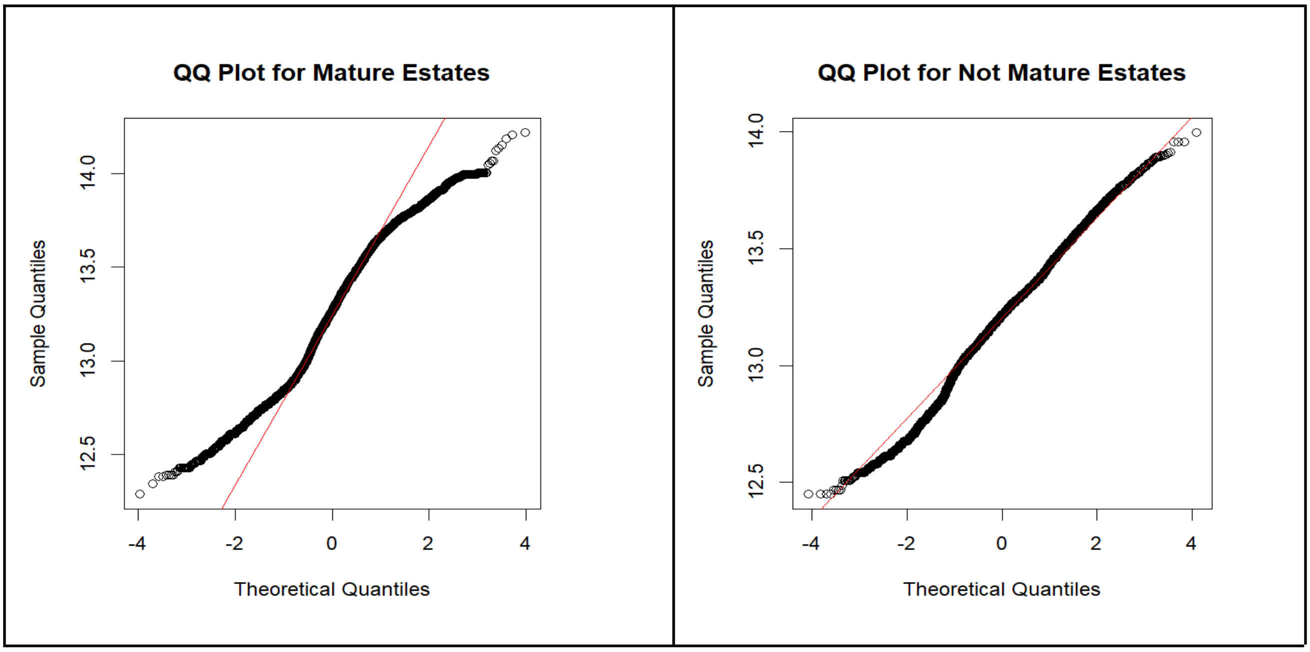
4.2.4 Relation between *resale_price* and *town*

In this section, we determine if the town where the flat is located affects HDB resale flat prices.

To simplify the analysis, we categorised HDB towns into Mature and Non-Mature Estates based on HDB's official classification. The table below shows how the towns are classified.

Mature Estate	Non-Mature Estate
'ANG MO KIO', 'BEDOK', 'BISHAN', 'BUKIT MERAH', 'BUKIT TIMAH', 'KALLANG/WHAMPOA', 'CLEMENTI', 'CENTRAL AREA', 'GEYLANG', 'MARINE PARADE', 'PASIR RIS', 'QUEENSTOWN', 'SERANGOON', 'TAMPINES', 'TOA PAYOH'	'BUKIT BATOK', 'BUKIT PANJANG', 'CHOA CHU KANG', 'HOUGANG', 'JURONG EAST', 'JURONG WEST', 'PUNGGOL', 'SEMBAWANG', 'SENGKANG', 'WOODLANDS', 'YISHUN'

From the Boxplot of Log Resale Price by Estate Maturity, we can see that the median resale price between the mature and non-mature group differs. The median resale price for the mature group is higher than the median resale price for the non-mature group.



Referring to the QQ Plot for Mature Estates, the data does not appear to follow a normal distribution. The points deviate significantly from the diagonal red line.
Referring to the QQ Plot for Not Mature Estates, the data appears to follow a normal distribution. The points are closely aligned with the diagonal line, indicating that the distribution is approximately normal.

Therefore, since one of the groups does not follow a normal distribution, the **Mann-Whitney U test** can be used to compare the median resale prices between the two groups as normality is not assumed.

Null Hypothesis (H_0): There is no difference in the median resale price between mature and non-mature estates.

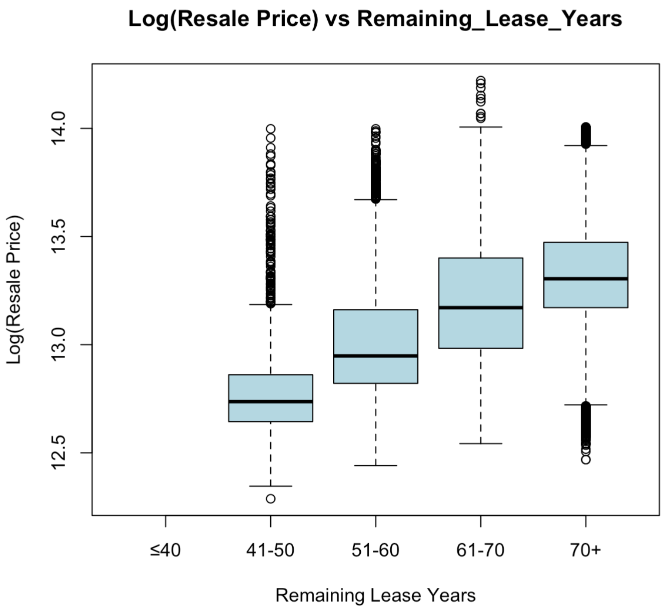
Alternative Hypothesis (H_1): There is a difference in the median resale price between mature and non-mature estates.

```
> wilcox.test(lg_resale_price ~ estate_maturity, data = filtered_df)

	Wilcoxon rank sum test with continuity correction

data:  lg_resale_price by estate_maturity
W = 185787712, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Based on the **Mann-Whitney U test**, at a significance level of 0.05, p-value = 2.2e-16 is less than 0.05. Hence, we reject the null hypothesis and conclude that there is a difference in the median resale price between mature and non-mature estates. This tells us that the town locations affect HDB resale flat prices.

4.2.5 Relation between *resale_price* and *remaining_lease_years*

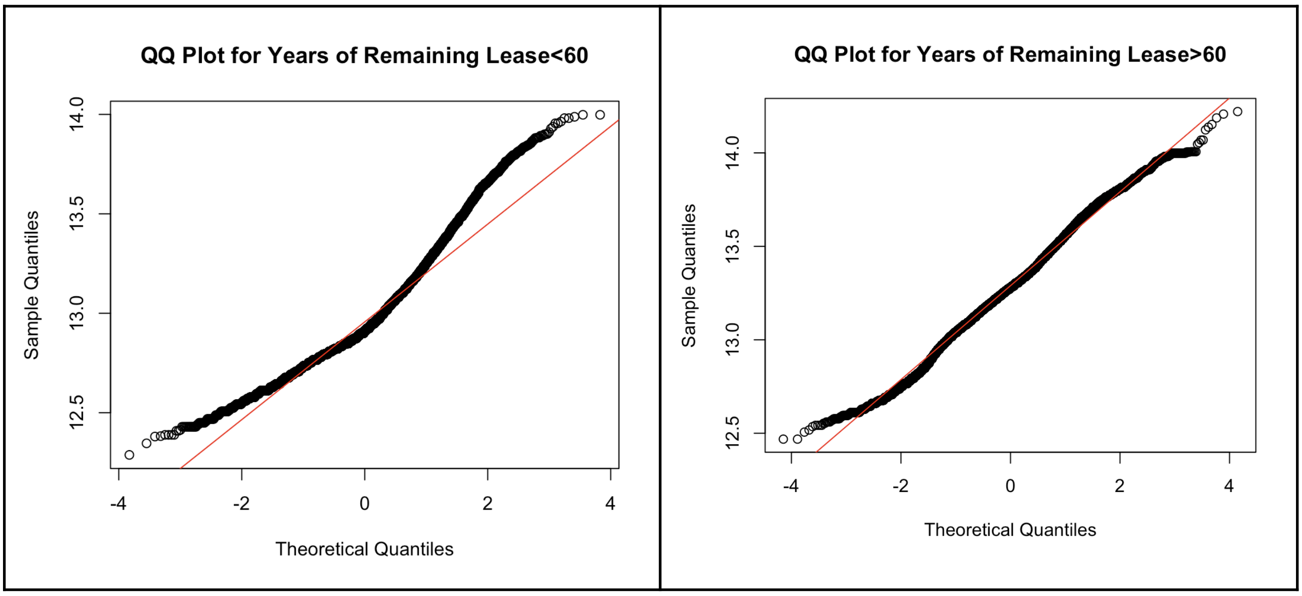
This section examines whether the years of remaining lease have a significant influence on the log-transformed resale price of flats.



From the Boxplot of Log Resale Price by Remaining Lease Years, we observe a clear upward trend in the median resale price as the remaining lease increases. Flats with more than 70 years of lease remaining have the highest median log resale price, followed by those in the 61–70 and 51–60 year categories. Conversely, flats with ≤40 years remaining exhibit the lowest median resale prices. This suggests that HDB flats with longer remaining leases generally command higher resale prices, likely due to greater long-term value and financing eligibility. The interquartile range also appears to widen in the 61–70 and 70+ groups, indicating more price variability among newer flats.

```
> model_lease <- lm(lg_resale_price ~ remaining_lease_years, data = filtered_df)
> summary(model_lease)

Call:
lm(formula = lg_resale_price ~ remaining_lease_years, data = filtered_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.87682 -0.17351 -0.02468  0.16130  1.05022 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           1.264e+01  6.878e-03 1838.40   <2e-16 ***
remaining_lease_years 7.813e-03  9.122e-05   85.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2682 on 37680 degrees of freedom
Multiple R-squared:  0.163,	Adjusted R-squared:  0.1629 
F-statistic:  7336 on 1 and 37680 DF,  p-value: < 2.2e-16
```

To quantify the effect of remaining lease years on log resale price, a simple linear regression was performed. The model shows a raw correlation between lease length and log resale price ($p < 2e-16$). For each additional year of remaining lease, the log resale price increases by approximately 0.0078 units. This suggests that HDB flats with longer leases command higher prices. This corresponds to an approximate 0.78% increase in resale price per additional lease year, reinforcing the observed trend in the boxplot.



Both Q-Q plots reveal systematic deviations in their tails, indicating that the data does not follow a normal distribution. The deviation in the upper and lower tails suggests skewness or excess kurtosis. Given this non-normality, parametric tests are unsuitable. We proceed with the non-parametric **Mann-Whitney U test** to evaluate differences between groups.

Null Hypothesis (H_0): There is no difference in the median log resale prices between the different lease year groups of HDB flats.

Alternative Hypothesis (H_1): There is a difference in the median log resale prices between the different lease year groups of HDB flats.

```
> filtered_df$lease_group <- ifelse(filtered_df$remaining_lease_years > 60, "Above 60", "60 and Below")
> filtered_df$lease_group <- as.factor(filtered_df$lease_group)
> wilcox.test(lg_resale_price ~ lease_group, data = filtered_df)

	Wilcoxon rank sum test with continuity correction

data:  lg_resale_price by lease_group
W = 47144348, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Based on the **Mann-Whitney U test**, at a significance level of 0.05, the p-value (< 2.2e-16) is less than 0.05. We reject the null hypothesis and conclude that there is a statistically significant difference in the median log resale prices across the different lease groups. The number of years of remaining lease has a significant impact on the resale value of HDB flats.

5. Multiple Linear Regression

In this part of our report, we aim to figure out which factor (Estate Maturity, Flat Type, Floor Area, Floor Category & Remaining Lease Years) contributes significantly to the log-transformed resale price. To achieve our aim, we built a Multiple Linear Regression model for lg(*resale_price*) based on the 5 factors.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.49277 -0.07803 -0.01326  0.06883  0.83955

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)
(Intercept)                1.173e+01  1.130e-01  103.748  < 2e-16 ***
flat_type2 ROOM           -6.360e-03  1.130e-01   -0.056 0.955122
flat_type3 ROOM            1.763e-01  1.130e-01    1.560 0.118745
flat_type4 ROOM            2.710e-01  1.131e-01    2.396 0.016578 *
flat_type5 ROOM            2.955e-01  1.132e-01    2.610 0.009066 **
flat_typeEXECUTIVE         3.793e-01  1.134e-01    3.344 0.000827 ***
flat_typeMULTI-GENERATION  5.432e-01  1.213e-01    4.477 7.59e-06 ***
floor_area_sqm             6.860e-03  8.873e-05   77.311  < 2e-16 ***
floor_category2            3.223e-02  1.852e-03   17.400  < 2e-16 ***
floor_category3            6.435e-02  1.674e-03   38.431  < 2e-16 ***
floor_category4            1.441e-01  1.996e-03   72.195  < 2e-16 ***
estate_maturityNot Mature -2.322e-01  1.329e-03 -174.794  < 2e-16 ***
remaining_lease_years      9.101e-03  4.697e-05  193.781  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1129 on 37669 degrees of freedom
Multiple R-squared:  0.8516,    Adjusted R-squared:  0.8515
F-statistic: 1.801e+04 on 12 and 37669 DF,  p-value: < 2.2e-16
```

Based on the model, we can conclude that *estate_maturity, floor_area_sqm, floor_category, remaining_lease_years* and *flat_type*, other than 2 and 3 room flats are statistically significant features. Additionally, the R-squared value of **0.8516** suggests that **85.16%** of the variation in lg(*resale_price*) can be explained by the predictors. The small difference of **0.0001** between Multiple R-squared and Adjusted R-squared indicates that the model is not overfitted. The fitted model is:

lg(*resale_price*) ≈ 11.730 + 0.0069 * *floor_area_sqm* + 0.0091 * *remaining_lease_years* + β_1 + β_2 + β_3

where:
β_1 = the coefficient for respective estate maturity (e.g. -0.2322 if the town is not mature)
β_2 = the coefficient for respective flat type (e.g. 0.1763 if the flat type is 3 Room)
β_3 = the coefficient for respective floor category (e.g. 0.1441 if the flat category is 4)

6. Conclusion

This report aims to address the initial research questions by analyzing the dataset containing the resale flat prices from 2023 to 2024 with rigorous statistical methods.

The findings revealed several key insights:

- Floor area(sqm) is a highly significant factor when setting the price of resale flat due to its low p-value. Thus we can conclude that larger floor areas are associated with higher resale price for flats.
- Our findings reveal clear price variations across different floor categories, demonstrating that floorheight directly influences property values.
- Our analysis also shows that types of flat impact resale prices significantly. The median price grows when the number of rooms increases.
- Resale prices in mature estates systematically differ and are generally higher from those in non-mature areas, reinforcing the role of location in Singapore's HDB market.
- Analysis shows a clear link between remaining lease duration and HDB resale value-every additional year raises prices by ~0.78%. This highlights how lease length directly impacts buyer willingness to purchase.
- Based on the coefficient magnitudes and statistical significance in our multiple linear regression model, remaining_lease_years(t=193.8) appears to be the most influential factor in affecting HDB resale prices, followed by estate_maturity(t=-174.8), floor_category4 (72.2), floor_area_sqm (t=77.3), and flat_type(multi generation)(t=4.5).
  This ranking is based on t-values, which measure each variable's statistical significance in the model. While floor_area_sqm has a higher t-value (77.3) than floor_category4 (72.2), the latter reflects a specific categorical effect (top-floor premium) that stands out more sharply in the analysis. Lease years dominate due to their precise, linear impact, whereas estate maturity's binary nature inflates its t-value. The t-values highlight reliability of effects, though practical importance may differ.

While the findings of our report are able to provide us with some insights on how various factors affect the HDB resale flat prices in Singapore, our analysis relies on data from a single time period found online. In retrospect, we initially assumed that factors like location(mature estates) or flat type would dominate HDB resale prices. However, our findings show otherwise that the years of remaining lease and floor area had far stronger statistical inference-an insight we might have missed without quantitative analysis. While these findings offer concrete insights, we recognise that our single-period dataset cannot capture market volatility or policy shifts over time. Future work could strengthen these conclusions by incorporating longitudinal data and additional variables such as proximity to MRT stations.

7. Appendix

Cleaning of Dataset (2. Data Description)

```r
library(dplyr)
library(zoo)

#####################################################################################################################
uncleanedhdb <- read.csv("sg-resale-flat-prices-2017-onwards.csv", header=TRUE)

# Check for missing values
colSums(is.na(uncleanedhdb))

#Check for duplicates and remove duplicates
duplicates <- uncleanedhdb %>%
  group_by(resale_price, month, town, flat_type, block, street_name, storey_range, floor_area_sqm, flat_model, lease_commence_date, remaining_lease) %>%
  filter(n() > 1)
cat("Number of duplicated rows:", nrow(duplicates), "\n")

removeduplicates_hdb_dt <- uncleanedhdb %>%
  distinct(resale_price, month, town, flat_type, block, street_name, storey_range, floor_area_sqm, flat_model, lease_commence_date, remaining_lease, .keep_all = TRUE)

cat("Number of rows after removing duplicates:", nrow(removeduplicates_hdb_dt), "\n")

head(removeduplicates_hdb_dt)

#keep only rows from 2023 to 2024
cleaned_df <- removeduplicates_hdb_dt[
  as.yearmon(removeduplicates_hdb_dt$month, "%Y-%m") >= as.yearmon("2023-01") &
    as.yearmon(removeduplicates_hdb_dt$month, "%Y-%m") <= as.yearmon("2024-12"),
]

head(cleaned_df)
```

```r
#Add new column to categorize storey range
categorize_storey <- function(storey_range) {
  lower_bound <- as.numeric(substr(storey_range, 1, 2))
  if (lower_bound >= 1 && lower_bound <= 3) {
    return(1)  # Ground Floor
  } else if (lower_bound >= 4  && lower_bound <= 6) {
    return(2)  # Ground-Low Floor
  } else if (lower_bound >= 7 && lower_bound <= 12) {
    return(3)  # Low-Medium Floor
  } else {
    return(4)  # High Floor
  }
}
cleaned_df$floor_category <- sapply(cleaned_df$storey_range, categorize_storey)
floor_counts <- table(cleaned_df$floor_category)
print(floor_counts)
```

```r
#Convert the values in 'remaining lease' in to year only, remove the months
convert_remaining_lease <- function(lease) {
  # Extract years
  years <- as.numeric(sub(" years.*", "", lease))
  # Extract months (set to 0 if not present)
  months <- ifelse(grepl("month", lease), as.numeric(sub(".*?([0-9]+) month.*", "\\1", lease)), 0)
  # Calculate numeric years
  years + (months / 12)
}
cleaned_df <- cleaned_df %>%
  mutate(remaining_lease_years = sapply(remaining_lease, convert_remaining_lease))

#Remove variables "month", "block", "street_name", "storey_range", "flat_model", and "lease_commence_date" ++
cleaned_df <- cleaned_df %>%
  select(-month, -block, -street_name, -storey_range, -flat_model, -lease_commence_date, -remaining_lease )
names(cleaned_df)
```

Histogram and Q-Q plot of *resale_price* (3.1 Summary Statistics For the Main Variable of Interest, *resale_price)*

```
# Histogram of resale_price
hist(cleaned_df$resale_price, main="Histogram of resale_price", breaks=30, xlab="resale_price")

# QQplot of resale_price
qqnorm(cleaned_df$resale_price)
qqline(cleaned_df$resale_price, col="red")
```

Log-Transformation of *resale_price*, Histogram and Boxplot of *lg_resale_price* (3.1 Summary Statistics For the Main Variable of Interest, *resale_price)*

```
# Perform Log-Transformation on resale_price
cleaned_df$lg_resale_price <- log(cleaned_df$resale_price)

# Histogram of lg(resale_price)
hist(cleaned_df$lg_resale_price, main="Histogram of lg(resale_price)", breaks=30, xlab="lg(resale_price)")

# Boxplot of lg(resale_price)
boxplot(cleaned_df$lg_resale_price, main="Boxplot of lg(resale_price)")
```

Identification, Investigation and Removal of Outliers (3.1 Summary Statistics For the Main Variable of Interest, *resale_price)*

```
# Identify and Investigate Outliers
IQR_value <- IQR(cleaned_df$lg_resale_price)
lower_bound <- quantile(cleaned_df$lg_resale_price, 0.25) - 1.5 * IQR_value
upper_bound <- quantile(cleaned_df$lg_resale_price, 0.75) + 1.5 * IQR_value
outliers <- cleaned_df$lg_resale_price < lower_bound | cleaned_df$lg_resale_price > upper_bound
outliers <- cleaned_df[outliers,]

# Remove Outliers
## Filter out right-tail outliers based on conditions
filtered_rt_df <- cleaned_df %>%
  filter(
    lg_resale_price > upper_bound,
    flat_type %in% c("EXECUTIVE", "5 ROOM", "4 ROOM"), # Floor types
    floor_area_sqm > 150, # Large floor area
    remaining_lease_years > 60 # High remaining lease years
  )
## Filter out left-tail outliers based on conditions
filtered_lt_df <- cleaned_df %>%
  filter(
    lg_resale_price < lower_bound,
    flat_type %in% c("1 ROOM", "2 ROOM", "3 ROOM"), # Floor types
    floor_area_sqm < 50, # Small floor area
    remaining_lease_years < 50 # Low remaining lease years
  )
## Combine the filtered right-tail and left-tail data into one dataframe
filtered_df <- cleaned_df %>%
  filter(lg_resale_price >= lower_bound & lg_resale_price <= upper_bound) %>% # Remove All Outliers from both tails
  bind_rows(filtered_rt_df, filtered_lt_df) # Add back data which satisfies the conditions
```

Histogram, Boxplot, Q-Q plot and Summary of Filtered *Dataframe* (3.1 Summary Statistics For the Main Variable of Interest, *resale_price)*

```
# Histogram of Filtered lg(resale_price)
hist(filtered_df$lg_resale_price, main="Histogram of Filtered lg(resale_price)",xlab="lg(resale_price)")

# Boxplot of Filtered lg(resale_price)
boxplot(filtered_df$lg_resale_price, main="Boxplot of Filtered lg(resale_price)")

# QQplot of Filtered lg(resale_price)
qqnorm(filtered_df$lg_resale_price)
qqline(filtered_df$lg_resale_price, col="red")

# Summary of Filtered lg(resale_price)
summary(filtered_df$lg_resale_price)
summary(filtered_df$lg_resale_price)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
12.29   13.03   13.23   13.22   13.42   14.22
```

Histogram and Boxplot of *floor_area_sqm* (3.2.1 Summary Statistics of *floor_area_sqm)*

```
# Histogram of Floor Area
hist(filtered_df$floor_area_sqm, main = 'Histogram of floor_area_sqm', xlab = 'floor_area_sqm')
hist(log(filtered_df$floor_area_sqm), main = 'Histogram of lg(floor_area_sqm)',xlab = 'floor_area_sqm')

# Boxplot of Floor Area
boxplot(filtered_df$floor_area_sqm, main = 'Boxplot of floor_area_sqm')
boxplot(log(filtered_df$floor_area_sqm), main ='Boxplot of lg(floor_area_sqm)')
summary(filtered_df$floor_area_sqm)
```

Histogram and Boxplot of *floor_category* (3.2.2 Summary Statistics of *floor_category)*

```
# Histogram of floor_category
hist(filtered_df$floor_category, main="Floor Category", xlab="Floor Category", xaxt="n")
axis(1, at=seq(min(filtered_df$floor_category), max(filtered_df$floor_category), by=1))
```

Barplot of Number of Flats by Flat Type (3.2.3 Summary Statistics for Flat Types)

```
> barplot(flat_types_count, main = "Number of flats by Flat Type", xlab = "Flat Type", ylab = "Count")
>
```

Barplot of Resale Flat Count by Town (3.2.4 Location of HDB flats (Town Area))

```
>
> town_counts <- table(filtered_df$town)
> town_counts

        ANG MO KIO            BEDOK           BISHAN      BUKIT BATOK      BUKIT MERAH   BUKIT PANJANG
              1477             1904              517             2157             1420            1128
       BUKIT TIMAH     CENTRAL AREA    CHOA CHU KANG         CLEMENTI          GEYLANG         HOUGANG
                68              231             1717              794              875            2112
       JURONG EAST      JURONG WEST KALLANG/WHAMPOA    MARINE PARADE        PASIR RIS         PUNGGOL
               727             2445             1140              198              937            2938
        QUEENSTOWN        SEMBAWANG         SENGKANG        SERANGOON         TAMPINES       TOA PAYOH
               881             1449             2813              619             2505            1024
         WOODLANDS           YISHUN
              2943             2663
> barplot(
+   town_counts,
+   las = 2,
+   main = "Resale Flat Count by Town",
+   xlab = "Town",
+   ylab = "Count",
+   cex.names = 0.35
+ )
```

Histogram of remaining_lease_years (3.2.5 Years of Remaining Lease)

```
hist(filtered_df$remaining_lease_years,
     breaks = 30,
     col = "grey",
     border = "black",
     xlab = "Remaining Lease (Years)",
     ylab = "Frequency",
     main = "Years of Remaining Lease")
```

Correlation plot of *lg_resale_price*, *floor_area_sqm* and *remaining_lease_years* (4.1 Correlations between *lg_resale_price* and other Continuous Variables)

```
# Create a custom panel function to display correlation
panel.cor <- function(x, y, digits = 2, prefix = "r = ", cex.cor = 1) {
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- cor(x, y, use = "complete.obs")
  txt <- paste0(prefix, round(r, digits))
  col <- ifelse(r < 0, "blue", "red")
  text(0.5, 0.5, txt, cex = cex.cor, col = col)
}

# Select only the relevant columns
corr_data <- filtered_df[, c("lg_resale_price", "floor_area_sqm", "remaining_lease_years")]

# Generate the correlation plot
pairs(corr_data,
      lower.panel = panel.cor,
      upper.panel = panel.smooth,
      main = "Scatterplot Matrix with Correlation Coefficients")
```

Linear Regression Model of *floor_area_sqm* against *lg_resale_price* (4.2.1 Relation Between *lg_resale_price* and *floor_area_sqm)*

```
# Linear Regression Model
lr.model = lm(formula = filtered_df$lg_resale_price ~ filtered_df$floor_area_sqm)
summary(lr.model)
```

Boxplot of *lg_resale_price and floor_category* (4.2.2 Relation Between *lg_resale_price* and *floor_category)*

```
# Boxplot of lg_resale_price and floor_category
ggplot(filtered_df, aes(x = floor_category, y = lg_resale_price)) +
  geom_boxplot(fill = "grey", color = "black") +  # Set box color to grey and outline color to black
  geom_hline(aes(yintercept = mean(lg_resale_price), color = mean_line), linetype = "solid", size = 1) +
  scale_color_manual(values = c("blue")) +  # Manually set color for mean line
  labs(
    title = "lg(resale_price) vs floor_category",
    x = "Floor Categories",
    y = "Log(Resale Price)",
    color = "Legend"
  ) +
  theme_minimal() +
  theme(
    legend.position = "none",  # Remove the legend
    plot.title = element_text(hjust = 0.5, face = "bold")  # Center-align and bold the title
  )
```

1-Way ANOVA and Pairwise t-test of *lg_resale_price and floor_category* (4.2.2 Relation Between *lg_resale_price* and *floor_category)*

```
# 1-Way ANOVA test
aov(filtered_df$lg_resale_price~factor(filtered_df$floor_category))
summary(aov(filtered_df$lg_resale_price~factor(filtered_df$floor_category)))

# Pairwise t-test
pairwise.t.test(filtered_df$lg_resale_price, filtered_df$floor_category, p.adjust.method='none')
```

Remove records of 1 Room and Multi-generation Type and boxplot of lg_resale_price vs flat type (4.2.3 Relation between lg_resale_price vs flat_type)

```
filtered_df1 = filtered_df[filtered_df$flat_type != '1 ROOM' & filtered_df$flat_type != 'MULTI-GENERATION',]
boxplot(lg_resale_price ~ flat_type, data = filtered_df1,
        main = "Boxplot of log(resale_price) vs flat_type", xlab = "Flat_type", ylab = "Log sale price", col = "lightblue", border = "black")
abline(h = mean(filtered_df1$lg_resale_price), col = "red")
```

1-Way ANOVA and Welch's t-test of lg-resale_price and flat_type (4.2.3 Relation between lg_resale_price and flat_type)

```
#1-Way ANNOVA
summary(aov(filtered_df1$lg_resale_price ~ factor(filtered_df1$flat_type)))

#Welch's t-test
t.test(filtered_df$lg_resale_price[filtered_df$flat_type == "3 ROOM"],
       filtered_df$lg_resale_price[filtered_df$flat_type == "2 ROOM"],
       alternative = "greater")
t.test(filtered_df$lg_resale_price[filtered_df$flat_type == "4 ROOM"],
       filtered_df$lg_resale_price[filtered_df$flat_type == "3 ROOM"],
       alternative = "greater")
t.test(filtered_df$lg_resale_price[filtered_df$flat_type == "5 ROOM"],
       filtered_df$lg_resale_price[filtered_df$flat_type == "4 ROOM"],
       alternative = "greater")
t.test(filtered_df$lg_resale_price[filtered_df$flat_type == "EXECUTIVE"],
       filtered_df$lg_resale_price[filtered_df$flat_type == "5 ROOM"]
       , alternative = "greater")
```

Boxplot of Log Resale Price by Estate Maturity (4.2.4 Relation between *resale_price* and *town*)

```
> mature_estates = c("ANG MO KIO", "BEDOK", "BISHAN", "BUKIT MERAH","BUKIT TIMAH", "KALLANG/WHAMPOA", "CLEMENTI", "CENTRAL AREA"
"GEYLANG", "MARINE PARADE", "PASIR RIS", "QUEENSTOWN", "SERANGOON", "TAMPINES", "TOA PAYOH")
>
> filtered_df$estate_maturity = ifelse(filtered_df$town %in% mature_estates, "Mature", "Not Mature")
>
> View(filtered_df)
> head(filtered_df)
        town flat_type floor_area_sqm resale_price floor_category remaining_lease_years lg_resale_price estate_maturity
1 ANG MO KIO    2 ROOM             44       267000              1              55.41667        12.49500          Mature
2 ANG MO KIO    2 ROOM             49       300000              2              53.50000        12.61154          Mature
3 ANG MO KIO    2 ROOM             44       280000              2              54.08333        12.54254          Mature
4 ANG MO KIO    2 ROOM             44       282000              3              54.08333        12.54966          Mature
5 ANG MO KIO    2 ROOM             45       289800              1              62.08333        12.57695          Mature
6 ANG MO KIO    3 ROOM             67       380000              2              54.08333        12.84793          Mature

# Creating the boxplot
boxplot(lg_resale_price ~ estate_maturity, data = filtered_df,
        main = "Boxplot of Log Resale Price by Estate Maturity",
        xlab = "Estate Maturity", ylab = "Log Resale Price")

# Adding red dashed horizontal lines at the medians for each estate maturity group
medians <- tapply(filtered_df$lg_resale_price, filtered_df$estate_maturity, median)

for (i in 1:length(medians)) {
  abline(h = medians[i], col = "red", lwd = 2, lty = 2)
}
```

QQ Plot for Mature Estates, QQ Plot for Not Mature Estates (4.2.4 Relation between *resale_price* and *town*)

```
mature_data = subset(filtered_df, estate_maturity == "Mature")
not_mature_data = subset(filtered_df, estate_maturity == "Not Mature")

qqnorm(mature_data$lg_resale_price, main="QQ Plot for Mature Estates")
qqline(mature_data$lg_resale_price, col="red")

qqnorm(not_mature_data$lg_resale_price, main="QQ Plot for Not Mature Estates")
qqline(not_mature_data$lg_resale_price, col="red")
```

Mann-Whitney U test (town)(4.2.4 Relation between *resale_price* and *town*)

```
wilcox.test(lg_resale_price ~ estate_maturity, data = filtered_df)
```

Boxplot of Log Resale Price by remaining_lease_years (4.2.5 Relation between *resale_price* and *remaining_lease_years*)

```
filtered_df$lease_bin <- cut(filtered_df$remaining_lease_years,
                             breaks = c(0, 40, 50, 60, 70, 99),
                             labels = c("≤40", "41-50", "51-60", "61-70", "70+"),
                             right = FALSE)

boxplot(lg_resale_price ~ lease_bin, data = filtered_df,
        main = "Log(Resale Price) vs Remaining_Lease_Years",
        xlab = "Remaining Lease Years",
        ylab = "Log(Resale Price)",
        col = "lightblue")
```

Linear regression of log resale price ~ remaining lease years(4.2.5 Relation between *resale_price* and *remaining_lease_years*)

```
model_lease <- lm(lg_resale_price ~ remaining_lease_years, data = filtered_df)
summary(model_lease)
```

QQPlot for *remaining_lease_years*<60, QQPlot for *remaining_lease_years*>60

```
remaining_lease_years1<-subset(filtered_df,remaining_lease_years>60)
qqnorm(remaining_lease_years1$lg_resale_price, main="QQ Plot for Years of Remaining Lease>60")
qqline(remaining_lease_years1$lg_resale_price, col = "red")

remaining_lease_years2<-subset(filtered_df,remaining_lease_years<60)
qqnorm(remaining_lease_years2$lg_resale_price, main="QQ Plot for Years of Remaining Lease<60")
qqline(remaining_lease_years2$lg_resale_price, col = "red")
```

Mann-Whitney U test (remaining_lease_years) (4.2.5 Relation between *resale_price* and *remaining_lease_years*)

```
#Create a grouping variable: Lease > 60 years vs ≤ 60 years
filtered_df$lease_group <- ifelse(filtered_df$remaining_lease_years > 60, "Above 60", "60 and Below")
filtered_df$lease_group <- as.factor(filtered_df$lease_group)

# Mann-Whitney U test (Wilcoxon rank-sum test)
wilcox.test(lg_resale_price ~ lease_group, data = filtered_df)
```

Multiple Linear Regression Model of *lg_resale_price* (5. Multiple Linear Regression)

```
# Fit Linear Regression Model
linear_model <- lm(lg_resale_price ~ .-resale_price-town, data = filtered_df)
summary(linear_model)
```

8. References

Lim, D. Singapore HDB Resale Flat Prices (2017–2024). *Kaggle.* Available: https://www.kaggle.com/datasets/darrylljk/singapore-hdb-resale-flat-prices-2017-2024.

Ng, M. (2021). HDB resale prices climb for 4th consecutive quarter amid covid-19 vaccine optimism. *The Straits Times*. Available: https://www.straitstimes.com/singapore/housing/hdb-resale-prices-climb-for-4th-consecutive-quarter-volume-dips.