Introduction

The primary objective of this study is to identify and analyze the factors that significantly influence the sales price of laptops. In collaboration with the product research team at Cityville Electronics, we have accessed a dataset containing over 1,000 entries, including detailed information on laptop brands, prices, processor types, core counts, and RAM memory. A deep understanding of how 'core i' processors (i.e., i3, i5, i7, etc.) are priced in the market will greatly assist in formulating the product pricing strategies of businesses.

The laptop market is fiercely competitive with various brands and technical specifications, and consumer choices are influenced by factors beyond price. This study aims to illuminate the crucial variables determining laptop prices and analyze how these interact, thereby enhancing competitive positioning in the market.

The primary variable analyzed in this study is the price of the laptops. The explanatory variables include brand, processor tier, RAM memory, and the number of processor cores. These factors are considered significant in influencing laptop sales prices, and our analysis will explore the relationships between them and their impact on price determination.

The research question we seek to answer is: "What are the most significant factors affecting the sales price of laptops, and how do these factors interact?" The target population encompasses users and manufacturers of laptops with various specifications, including 'core i' processors. This research will provide valuable insights to companies like Cityville Electronics, aiming to understand the price determination factors within this product range.

This study aims to provide comprehensive insights into the pricing dynamics of laptops with 'core i' processors, thereby aiding companies like Cityville Electronics in strategizing their market positioning and pricing mechanisms effectively.

ANALYSIS

Diagnostics

Checking the four main assumptions of the errors (independence, linearity, homoscedasticity, and normality) is necessary to ensure the validity of the multiple linear regression model and the accuracy of its result. We can check these assumptions by observing the residual plot, QQ plot, and box plot.

From the laptop dataset that we obtained from our client, our group focuses on four explanatory variables (brand, number of cores, RAM memory, and processor tier) and one response variable (price). Our group fits four separate linear models in R with each of the four different explanatory variables. From these linear models, we calculate the residual of each linear model. Our group used these residuals to plot the residual plot and QQ plot. The plots for our diagnostic are in the appendix section.

Independence

We can assess the independence assumption using a time series plot and residual vs fitted values plot. We need to observe the presence of a pattern from both plots. A pattern indicates the violation of the independence assumption. From the time series plot shown in Figure 1.1, we do not observe any trend for all variables. It shows that our data is independent. It is supported by the residual vs fitted plot shown in Figure 1.2, where there is no pattern in the residual plot. Therefore, we can assume that our data is independent of the evidence from observing both time series and residual vs fitted values plots.

Linearity and Homoscedasticity

We assess the linearity and homoscedasticity assumption by observing the residuals in the residual plot. In the residual vs predictor variable plot, we are observing whether there is a pattern in the residuals. If we observe a pattern or trend from the residuals, it means that there is a violation in the linearity assumption. On the other hand, if there is no obvious pattern, we can assume that the linearity assumption is reasonable. In addition, we are examining the vertical spread of the residuals to assess the homoscedasticity assumption. A constant vertical spread of the residuals indicates that the homoscedasticity assumption is acceptable.

In the residual vs. number of cores plot (Figure 1.3), we observe a positive correlation between residuals and the number of cores, explained by the increase in residuals with the greater number of cores. Therefore, there is a violation of the linearity assumption for the variable number of cores. A method to address this violation is by doing a log transformation to the data. When plotting the residual plot after doing the log transformation, we observe no pattern in the residual vs number of cores plot (Figure 1.4). Hence, the linearity assumption is acceptable. Figure 1.4 shows that the spread of the residuals is vertically constant, so we assume that the homoscedasticity assumption holds for the variable number of cores.

Next, for the RAM memory variable, we do not observe any pattern or trend in either of the residual plots before and after doing the log transformation (Figure 1.5-1.6). As a result, we can conclude that the linearity assumption is acceptable. Likewise, the vertical spread of the residuals remains constant, indicating that the homoscedasticity assumption is also acceptable.

Since brand and processor tier are categorical variables, we cannot check the linearity assumption using the residual plots. However, we can check the homoscedasticity assumption by creating a box plot of residuals to observe the residual spread.

The boxplot of residuals for the brand as the predictor variable (Figure 1.7) shows that the residuals for different laptop brands are not constant, indicating a violation of the homoscedasticity assumption. When comparing the boxplot in Figure 1.7 with the boxplot after doing a log transformation in Figure 1.8, we observe that the vertical spread of the residuals is more constant in Figure 1.8. Likewise, when checking the homoscedasticity assumption for processor_tier, we compare the boxplot of residuals before and after doing the log transformation (Figure 1.9-1.10). The residuals in Figure 1.10 have a more constant spread across different processor tiers.

Normality

We assess the normality assumption using a QQ plot. From observing the QQ plot for all variables (Figure 1.10-1.13), we find a significant deviation of residuals from the QQ line. The deviation of residuals from the line indicates a violation of the normality assumption. Our group handles this violation by doing a log transformation to the price variable.

When plotting the QQ plot with the variable log_price instead of price, the residuals are getting closer to the QQ line. The residuals in the middle of the QQ plot after doing log transformation (Figure 1.10-1.13) are along the QQ line. Although the residuals at both ends of the QQ plot do not perfectly match the line, we can still conclude that the normality assumption is acceptable after doing a log transformation to the price variable.

Model Validation

To answer our client's question of which factors influence the price, we have taken an adjusted fit with log price. For our analysis, we have used a backward elimination process.

Backward elimination process is the process by which we take an initial model with all the explanatory variables, brand, number of cores, Ram memory and laptops with Core i

processors that might affect our response variable, log Price. We take a p-value of 0.05 to check the significance of each variable on the model. Variables with p-value greater than 0.05 in the model are not seen as significant to our model. This is when we start the backward elimination and remove the variable with the highest p-value greater than 0.05. We rerun the model without the removed variable and repeat the process before until there are no variables with p-value greater than 0.05 left. This remaining model will be our final model containing all variables that are significant to the log Price of laptops.

Looking at our final model, we see that no variables are removed from the initial model. The lack of removal of any variables underscores the importance of each factor considered in influencing laptop prices within the dataset.

Coefficients

Looking at the summary of the model at significance level of 0.05, we see that the processor tier core i7 and i9 have more significance than i5 core.

Looking at brands, we see that Asus, Dell, HP, Lenovo, LG, Microsoft, MSI, Tecno, Ultimus, and Wings have significant positive coefficients, indicating higher prices compared to the reference brand.

Both the number of cores and RAM memory have positive coefficients, indicating that higher values of these variables are associated with higher laptop prices.

Significance Codes

The significance codes (typically denoted as asterisks) accompanying the coefficients indicate the level of statistical significance. Lower p-values correspond to greater significance. In this summary, asterisks denote the significance levels:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Looking ar Figure 1.14 in appendix, we can see that the brands mentioned above and the processors, num_cores and ram memory are significant at $p < 0.001$.

Model Fit

Multiple R-squared is 0.8216 which means that the predictors in the model explain approximately 82.16% of the variation in laptop prices.

The F-statistic is highly significant ($p < 2.2e-16$), suggesting that the overall model is a good fit for the data.

The backward elimination is not able to give us the best model as brand and core processors are categorical so we have also used ANOVA f-test.

ANOVA

Looking at the coefficients of brand, number of cores, ram memory, processor tiers from in figure 1.10, we can see that these variables have a significant effect on the log price. The p-values are less than 0.05 for all 4 variables so they are all significant. This corroborates our finding in the backward elimination and further emphasizes that all of these factors have a significant effect on log price.

Interpretation

The final multiple linear regression model along with ANOVA model suggests that processor tier (especially core i7 and core i9), brand, number of cores, and RAM memory significantly influence laptop prices. The positive coefficients for these predictor variables suggests that laptops with these specifications tend to have premium prices.

We will have to do sensitivity analysis and diagnostics checks in the future to maintain the reliability of the model's findings.

Results

This study has identified that the most significant factors impacting laptop sales prices are processor grade, brand, core count, and RAM capacity. Higher-tier processors, particularly i7 and i9, correlated with higher laptop prices, reflecting their premium market value. These processor grades have a profound impact on pricing, demonstrated by p-values less than 0.001, signifying a strong influence on price determination.

Regarding brands, Asus, Dell, HP, and Lenovo show a positive correlation with higher price points. This analysis confirms that these well-known brands command a premium compared to lesser-known brands. Additionally, laptops with higher core counts and greater RAM capacities are priced higher, mirroring the market's valuation of performance. Both variables showed statistically significant positive coefficients, underscoring their importance in price determination.

Bias Considerations

While the findings offer valuable insights into the factors affecting laptop prices, it's crucial to recognize potential biases within the dataset. The data, primarily sourced from Cityville Electronics, might reflect specific market segments or consumer preferences, potentially skewing the outcomes. Furthermore, the focus on 'core i' processors may limit the generalizability of the findings across the broader laptop market. Future analyses should consider incorporating a more diverse dataset to mitigate these biases and validate the study's conclusions.

The ANOVA analysis further validates the significance of the model. Processor grade, brand, core count, and RAM capacity all demonstrated significant p-values, indicating

a meaningful impact on laptop prices. This finding aligns with the multiple linear regression analysis, supporting that these selected factors play a crucial role in pricing.

The model's fit is evidenced by an R-squared value of 0.8216, indicating that it explains approximately 82% of the variance in laptop prices. The highly significant F-statistic ($p < 2.2e-16$) confirmed the overall model's adequacy for the data.

In conclusion, this research confirms that processor grade, brand, core count, and RAM are the primary drivers of pricing in the laptop market. The significant positive coefficients for these factors suggest that higher performance and brand recognition are associated with higher prices, aligning with market expectations. This underscores the importance of technical specifications and brand reputation in determining laptop prices. Future work will include sensitivity analysis to ensure the model's robustness and further explore the nuanced relationships between these variables and laptop pricing, while also addressing potential biases to strengthen the study's validity.

Conclusion

This study delved into the factors influencing laptop sales prices, focusing on processor grade, brand, core count, and RAM size. The analysis revealed significant findings:

Processor Grade: Higher-tier processors such as i7 and i9 were associated with increased laptop prices, indicating their premium market value. Statistical tests underscored the strong influence of these processor grades on price.

Brand Influence: We find that brands such as Asus, Dell, HP, and Lenovo have significant influence over pricing dynamics, commanding premium prices compared to lesser-known counterparts. In our regression model, positive coefficients associated with

these brands indicate a strong correlation between brand recognition and higher price points. This shows that consumers value brand reputation when making purchasing decisions in the laptop market.

Core Count and RAM Impact: Laptops with more cores and higher RAM were priced higher, emphasizing the market's valuation of performance. Both variables had positive coefficients and were statistically significant, further emphasizing their role in price determination.

The model fit was robust, explaining approximately 82% of the variance in laptop prices, with a high R-squared value and a highly significant F-statistic. These results imply that processor grade, brand reputation, core count, and RAM size are crucial drivers of laptop pricing.

Discussion

The findings of this study have important implications for businesses operating in the laptop market, such as Cityville Electronics. Understanding the key determinants of laptop prices enables companies to formulate effective pricing strategies and enhance their competitive positioning.

Processor grade is a pivotal factor, with higher-tier processors commanding premium prices. This highlights the importance of investing in advanced technologies to cater to consumers' demand for performance.

Companies can use brand reputation to raise the price of their products and differentiate themselves from their competitors.

Shortcomings

Data Limitations: This study relies on a single dataset for analysis, which may only represent specific periods and market segments. Future research could increase the generalizability and reliability of results by using more extensive and diverse datasets.

Selection Bias: The dataset used in this study might be subject to selection bias, wherein certain brands or models are overrepresented or underrepresented due to sampling methods or data collection processes. This bias could skew the analysis and lead to inaccurate conclusions about the factors influencing laptop prices.

Response Bias: There may be response bias in the data. As an example, respondents may provide biased information about their preferences, purchasing behavior, or perceptions of brand reputation. Factors such as brand loyalty, market influence, or social expectation bias may affect the relationship between variables and prices, thus affecting the reliability of the findings.

Variable Selection Constraints: This study ignores other potential factors such as product design, functionality, and marketing strategies. Future research could explore other factors that influence prices to gain a fuller understanding

Methodological Limitations: While the multiple linear regression analysis in this study provides valuable insights into the relationships between variables, it may ignore nonlinear relationships or interactions between factors. Additionally, the basic assumptions of regression analysis, such as linearity, mean square error, and normality, may not always apply to the actual data, which may affect the validity of the results.

Improvements for Future Research

Expand Sample Scope: Future studies can increase representativeness and credibility by using larger, more widely covered samples. This can be done by collecting data across different regions, periods, and market segments. Researchers can also improve the generality and representativeness of their findings and gain a more complete understanding of price determinants in the laptop market.

Include Additional Variables: To understand the multifaceted nature of laptop pricing, Future research could consider incorporating more variables such as product characteristics, market environments, and consumer preferences to comprehensively understand the multifactorial influences on price formation.

Method Optimization: Future research could explore different modeling methods and techniques to enhance the predictive power and interpretability of models.

Through these improvements, future research can gain a more comprehensive understanding of the mechanisms underlying laptop price formation and provide stronger support for developing more effective pricing strategies for businesses.

Appendix

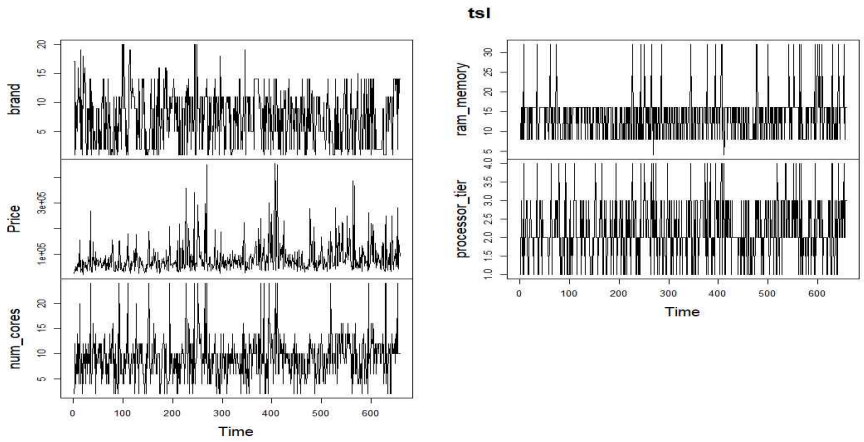

Figure 1.1 Time Series Plot of All Variables

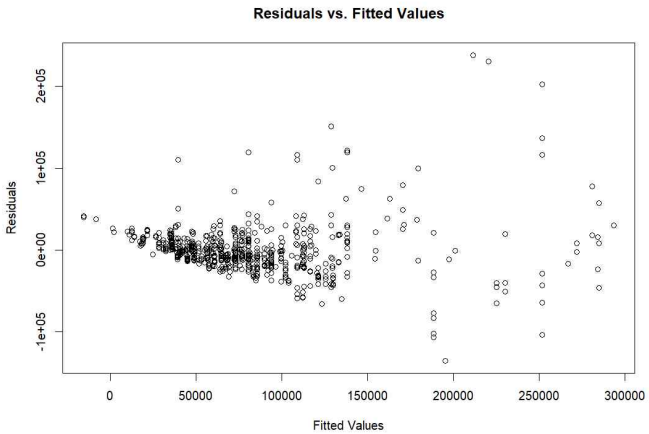

Figure 1.2 Residual vs Fitted Values Plot

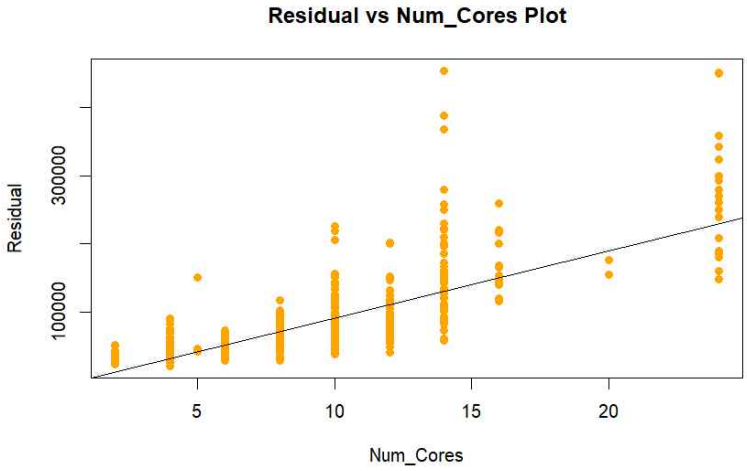

Figure 1.3 Residual vs Num_Cores Plot

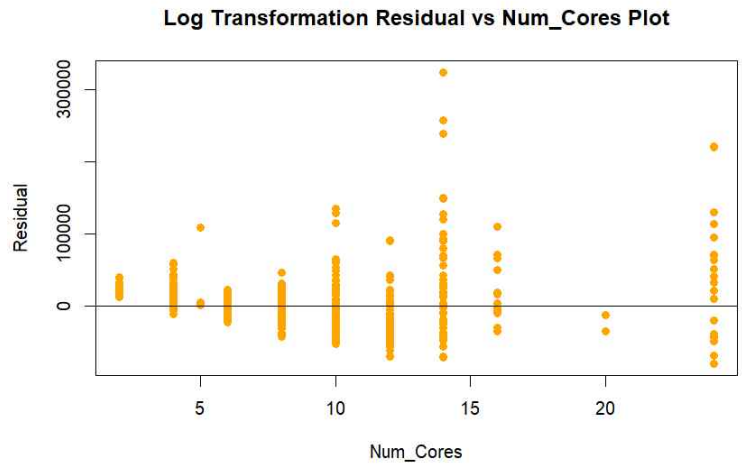

Figure 1.4 Log Transformation Residual vs Num_Cores Plot

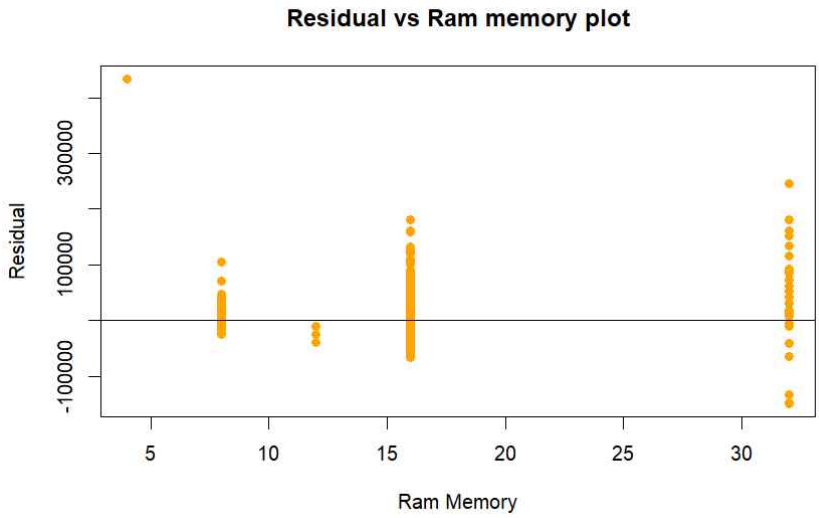

Figure 1.5 Residual vs Ram memory Plot

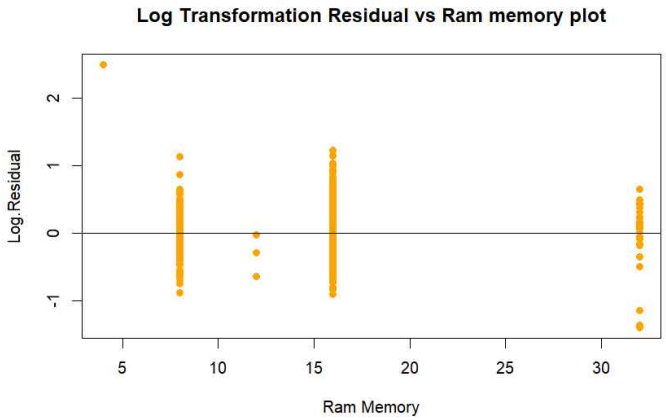

Figure 1.6 Log Transformation Residual vs Ram memory plot

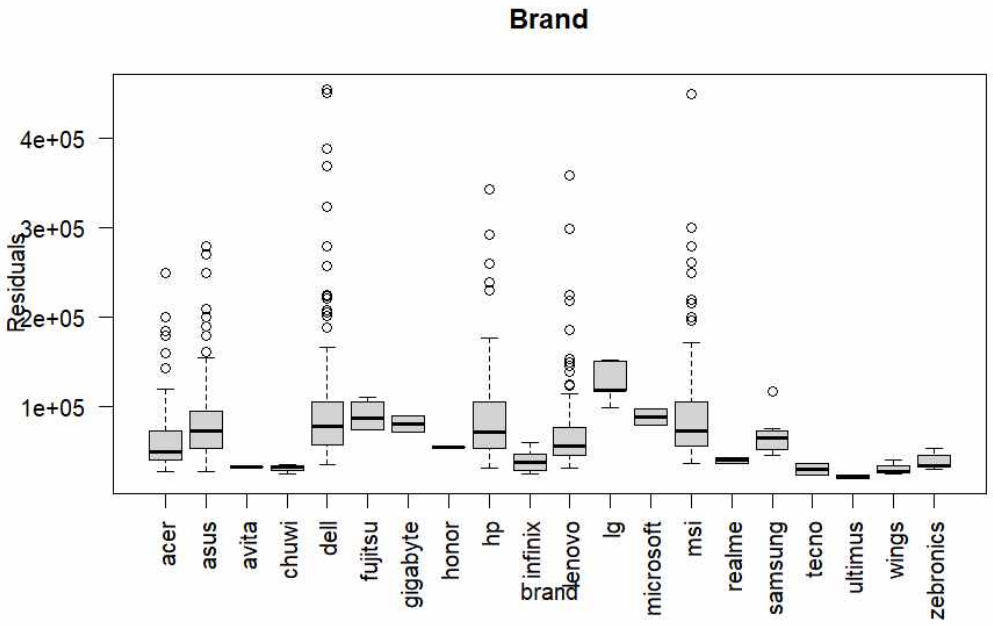

Figure 1.7 Residual vs Brand Boxplot

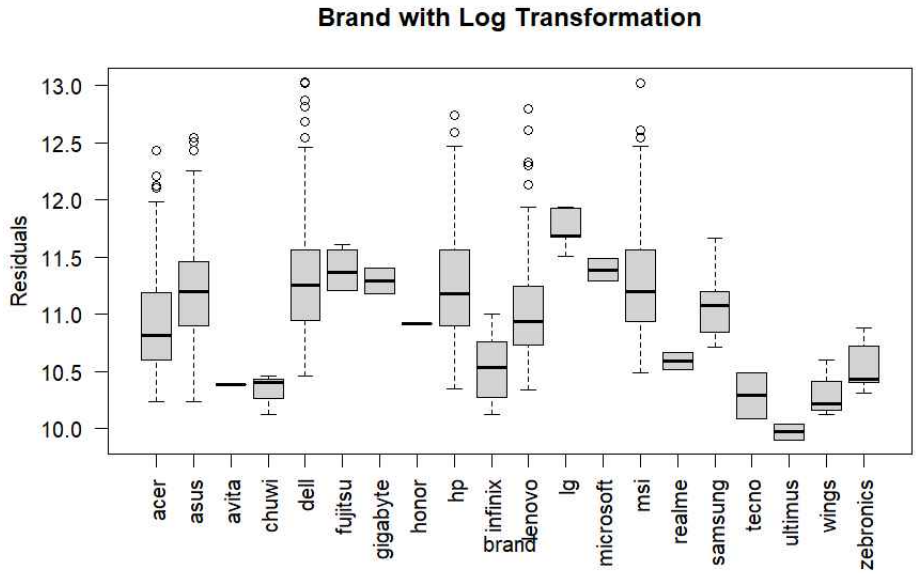

Figure 1.8 Residual vs Brand Boxplot after Log Transformation

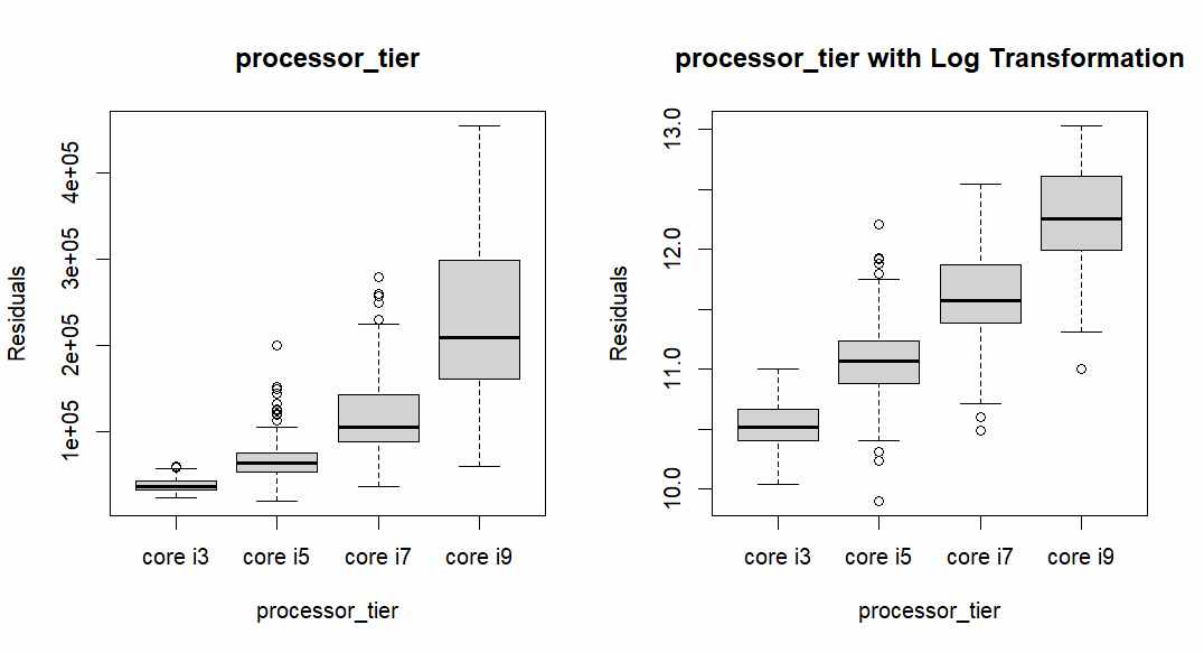

Figure 1.9 Residual vs Processor tier Boxplots

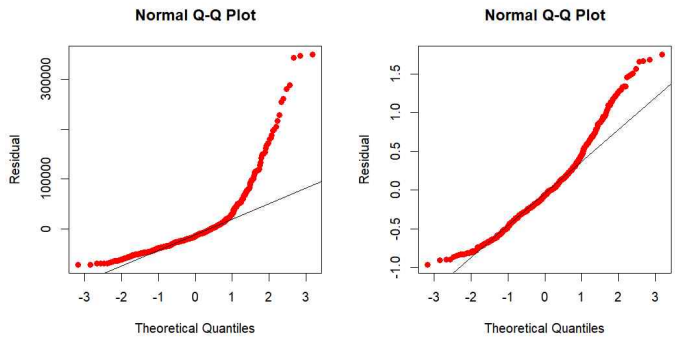

Figure 1.10 Brand QQ plot

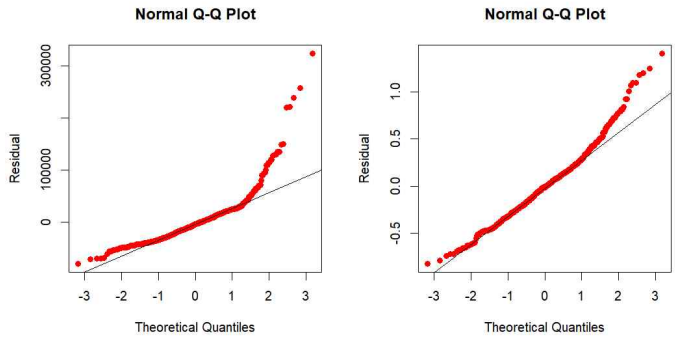

Figure 1.11 Num_Cores QQ plot

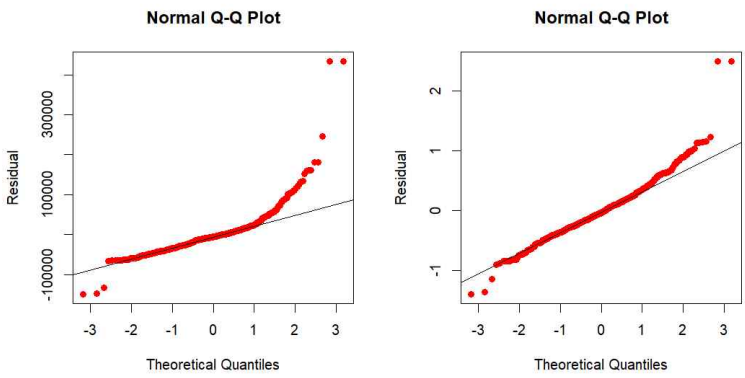

Figure 1.12 Ram Memory QQ plot

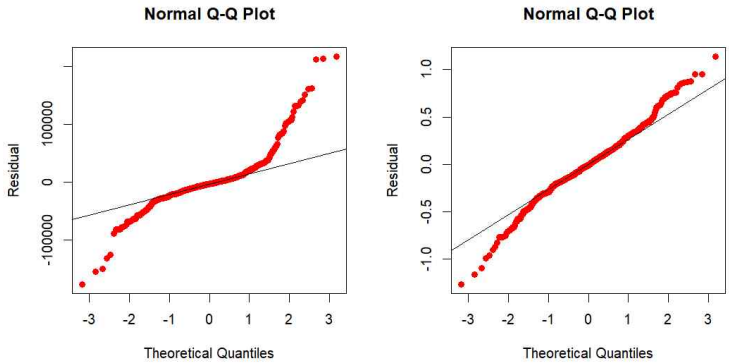

Figure 1.13 Processor_tier QQ plot

```
lm(formula = paste(response, "~", paste(c(include, cterms), collapse = " + ")),
    data = l)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68362 -0.13945 -0.01962  0.11780  1.11354

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           10.023915   0.041417 242.027  < 2e-16 ***
brandasus              0.145283   0.035050   4.145 3.86e-05 ***
brandavita            -0.065625   0.236930  -0.277 0.781887
brandchuwi            -0.206768   0.139108  -1.486 0.137673
branddell              0.264653   0.037857   6.991 6.96e-12 ***
brandfujitsu           0.075055   0.121451   0.618 0.536808
brandgigabyte          0.347441   0.168560   2.061 0.039688 *
brandhonor            -0.037244   0.236563  -0.157 0.874948
brandhp                0.229904   0.035851   6.413 2.80e-10 ***
brandinfinix          -0.418161   0.074237  -5.633 2.67e-08 ***
brandlenovo            0.157501   0.035891   4.388 1.34e-05 ***
brandlg                0.375671   0.109678   3.425 0.000654 ***
brandmicrosoft         0.440154   0.168560   2.611 0.009234 **
brandmsi               0.163154   0.040174   4.061 5.49e-05 ***
brandrealme            0.148009   0.169050   0.876 0.381616
brandsamsung           0.029900   0.093442   0.320 0.749084
brandtecno            -0.400337   0.169729  -2.359 0.018643 *
brandultimus          -0.474330   0.169050  -2.806 0.005173 **
brandwings            -0.318162   0.121632  -2.616 0.009115 **
brandzebronics        -0.271658   0.093463  -2.907 0.003782 **
num_cores              0.042479   0.003363  12.632  < 2e-16 ***
ram_memory             0.022171   0.002202  10.069  < 2e-16 ***
processor_tiercore i5  0.233669   0.028834   8.104 2.75e-15 ***
processor_tiercore i7  0.542257   0.038842  13.961  < 2e-16 ***
processor_tiercore i9  0.753791   0.069105  10.908  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2345 on 633 degrees of freedom
Multiple R-squared:  0.8216,    Adjusted R-squared:  0.8148
F-statistic: 121.4 on 24 and 633 DF,  p-value: < 2.2e-16
```

Figure 1.14 - Summary of Backward Elimination

```
Analysis of Variance Table

Response: Log_Price
                Df  Sum Sq Mean Sq  F value    Pr(>F)
brand           19  31.812   1.674   30.436 < 2.2e-16 ***
num_cores        1 101.833 101.833 1851.176 < 2.2e-16 ***
ram_memory       1  14.863  14.863  270.188 < 2.2e-16 ***
processor_tier   3  11.827   3.942   71.668 < 2.2e-16 ***
Residuals      633  34.821   0.055
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1.15 - Summary of ANOVA

R Code

```r
install.packages("dplyr")
library(dplyr)
install.packages("vcd")
library(vcd)

################ Vincent Putra
################ Checking the 4 assumptions on the error
########## load the data and fit a lm function
laptop<-read.csv("laptops.csv")
laptop <- subset(laptop, select = c(brand, Price,num_cores,ram_memory,processor_tier))
laptop <- filter(laptop,processor_tier %in% c("core i3","core i5","core i7","core i9"))
laptop <- laptop %>% mutate(Log_Price=log(Price))
fit<- lm(Price~brand+num_cores+ram_memory+processor_tier,data=laptop)
residuals<- resid(fit)

######## plot a time series and residual vs fitted plot
tsl <- ts(laptop)
plot(fitted(fit), residuals, main = "Residuals vs. Fitted Values",
     xlab = "Fitted Values", ylab = "Residuals")

##### box plot and qqplot for brand

brandF <- factor(laptop$brand)
cor(laptop$Price,as.numeric(brandF))
fit_brand <- lm(Price~brand,laptop)
resid_brand <- resid(fit_brand)
fit_LogBrand <- lm(Log_Price~brand,laptop)
resid_LogBrand <- resid(fit_LogBrand)
qqnorm(resid_brand,pch=16,col="red",ylab="Residual")
qqline(resid_brand)
qqnorm(resid_LogBrand,pch=16,col="red",ylab="Residual")
qqline(resid_LogBrand)

par(mfrow = c(1, 2))  # Arrange plots in one row with two columns
boxplot(Price~brand,data=laptop, ylab = "Residuals",main="Brand")
boxplot(Log_Price~brand,data=laptop, ylab = "Residuals",main="Brand with Log Transformation")

```

```r
######### residual and qqplot for num_cores

fit_NC <- lm(Price~num_cores,laptop)
resid_NC <- resid(fit_NC)
fit_LogNC <-lm(Log_Price~num_cores,laptop)
resid_LogNC <- resid(fit_LogNC)
summary(fit_NC)
options(scipen=10)

plot(laptop$num_cores,laptop$Price,col="orange",pch=16,xlab="Num_Cores",ylab="Residual",
     main="Residual vs Num_Cores Plot")
abline(fit_NC)

plot(laptop$num_cores,resid_NC,col="orange",pch=16,xlab="Num_Cores",ylab="Residual",
     main="Log Transformation Residual vs Num_Cores Plot")
abline(fit_LogNC)

abline(0,0)
qqnorm(resid_NC,pch=16,col="red",ylab="Residual")
qqline(resid_NC)
qqnorm(resid_LogNC,pch=16,col="red",ylab="Residual")
qqline(resid_LogNC)


######### residual and qqplot for ram_memory
cor(laptop$Price,laptop$ram_memory)
fit_RM <- lm(Price~ram_memory,laptop)
resid_RM <- resid(fit_RM)
fit_LogRM <- lm(Log_Price~ram_memory,laptop)
resid_LogRM <- resid(fit_LogRM)
plot(laptop$ram_memory,resid_RM,col="orange",pch=16,xlab="Ram Memory",ylab="Residual",
     main="Residual vs Ram memory plot")
plot(laptop$ram_memory,resid_LogRM,col="orange",pch=16,xlab="Ram Memory",ylab="Log.Residual",
     main="Log Transformation Residual vs Ram memory plot")
abline(0,0)
qqnorm(resid_RM,pch=16,col="red",ylab="Residual")
qqline(resid_RM)
qqnorm(resid_LogRM,pch=16,col="red",ylab="Residual")
qqline(resid_LogRM)
```

```r
#################### boxplot and qqplot for processor tier
PT_F <- factor(laptop$processor_tier)
cor(laptop$Price,as.numeric(PT_F))
fit_PT <- lm(Price~processor_tier,laptop)
resid_PT <- resid(fit_PT)
fit_LogPT <- lm(Log_Price~processor_tier,laptop)
resid_LogPT <- resid(fit_LogPT)

abline(0,0)
qqnorm(resid_PT,pch=16,col="red",ylab="Residual")
qqline(resid_PT)
qqnorm(resid_LogPT,pch=16,col="red",ylab="Residual")
qqline(resid_LogPT)
boxplot(Price~processor_tier,data=laptop, ylab = "Residuals",main="processor_tier")
boxplot(Log_Price~processor_tier,data=laptop, ylab = "Residuals",
        main="processor_tier with Log Transformation")
```

```r
#BHAVANA
#MODEL VARIATION

# Initial model with all explanatory variables
initial_model <- lm(Log_Price ~ brand + num_cores + ram_memory + processor_tier, data = laptop)

# Perform backward elimination
back <- step(initial_model, direction = "backward")

summary(back)
anova(back)
```

Contribution by Team Members

JooHyeok Seo

- **Section**: Introduction and Results
- **Contributions**: Defined the primary objectives of the study, analyzed and interpreted the results, and examined the impact of key factors derived from the dataset on laptop pricing.

Vincent Putra

- **Section**: Analysis and Diagnostics
- **Contributions**: Performed diagnostic procedures for the multiple linear regression model, evaluated assumptions of independence, linearity, homoscedasticity, and normality, conducted data transformations, and prepared regression diagnostic plots.

BHAVANA

- **Section**: Model Validation and Interpretation
- **Contributions**: Applied backward elimination for model validation, identified significant variables, performed ANOVA analysis, interpreted the results, and assessed the model fit.

Patrick

- **Section**: Conclusion and Discussion
- **Contributions**: Summarized the study's findings, provided in-depth analysis of the factors influencing market pricing, discussed the limitations of the study, and suggested directions for future research.