# Project_(631)

**Affect of RAM and Storage on Laptop prices**

**Mariyam Anna Jose, Akhil Sachar**

## Table of contents

# 1 Introduction

Laptops vary widely in price based on their technical specifications. Two of the most influential hardware components that contribute to price are *RAM size* and *storage capacity*. This study investigates the impact of these two features on laptop pricing. The primary research question is,

*"How do Storage Capacity and RAM Size affect laptop price , and is there an interaction between them". Hypothesis:*

- *Null Hypothesis ($H_0$) **:* There is no significant effect of RAM size or storage capacity on laptop price.
- *Alternative Hypothesis ($H_1$) **:* There is a significant effect of RAM size and/or storage capacity on laptop price.

### *Dataset Description*

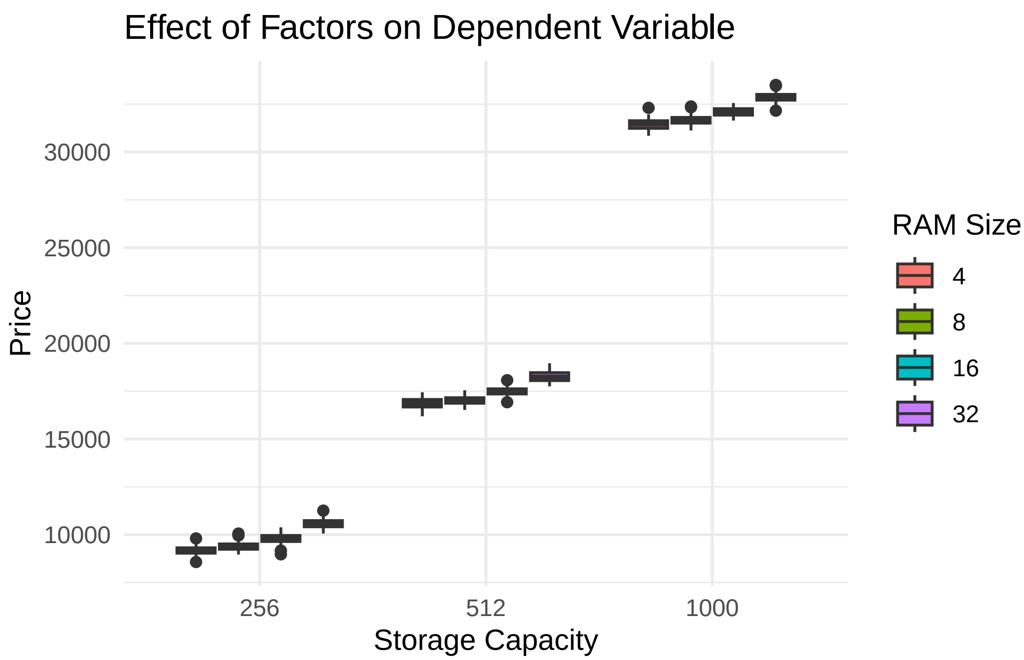
The dataset, named Laptop_price.csv, was obtained from Kaggle.com. It contains *1,000 observations* and includes the following variables:

- *Brand* (5 levels)
- *RAM_Size* (4 levels: 4GB, 8GB, 16GB, 32GB)
- *Storage_Capacity* (3 levels: 256GB, 512GB, 1000GB)
- *Price* (continuous response variable)

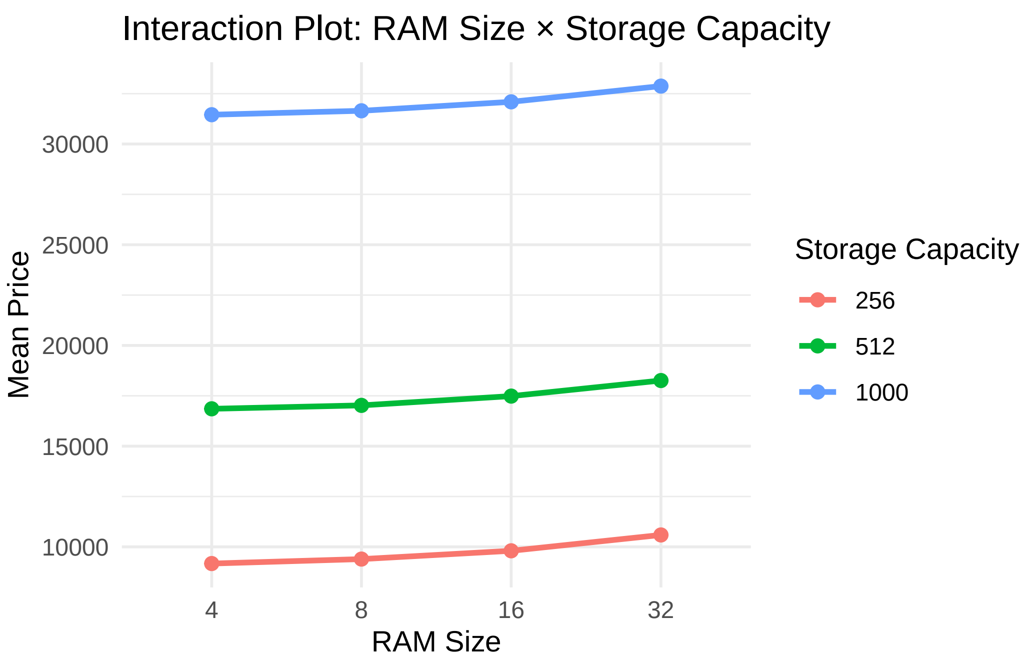
The dataset is *unbalanced*, That means all combinations of RAM and storage levels have not the same number of observations.

## 3. Results and Interpretation

### 3.1 Visualizations and Interpretation




- *Boxplot:* Displayed how laptop prices increase with both RAM size and storage capacity. Within each storage level, higher RAM was associated with higher price. The distinct separation between boxplots suggests statistically meaningful differences.




- *Interaction Plot:* Showed parallel lines for each storage level as RAM increases, confirming that there is no significant interaction. RAM size and storage capacity each contribute to the price of a laptop independently. The effect of one does not depend on the level of the other — both independently raise the price.

### 3.2 Two-Way Type Two ANOVA

We performed a Two-Way Type 2 ANOVA to assess the effects of RAM size and storage capacity on laptop price. Additionally we are also looking to find out the interaction effect between RAM size and storage capacity. We have decided to use type two ANOVA because the data we are using is unbalanced. Moreover we have evaluated the model assumptions and performed post-hoc analysis using Tukey's Honest Significant Difference (HSD) test to determine which group differences were statistically significant.

### 3.3 Main Effects and Interaction

From the ANOVA output, we observed the following:

| Factor | F-value | P-value |
|---|---|---|
| Storage Capacity | 722,500 | $<$ 2e-16 |
| RAM Size | 1,692 | $<$ 2e-16 |

| Factor | F-value | P-value |
|---|---|---|
| Storage × RAM Interaction | 0.204 | 0.9756 |

- The p-values for both RAM size and storage capacity are well below 0.05, indicating that each has a statistically significant and strong effect on laptop price.
- The interaction term, however, has a p-value of 0.9756, which is much higher than the 0.05 threshold. This clearly shows that the combined effect of RAM and storage on price is not statistically significant.
- As a result of this finding, we updated our model to an additive model, which assumes that the effects of RAM and storage are independent and additive, not multiplicative.

From the updated ANOVA model output, we observed the following:

| Factor | F-value | P-value |
|---|---|---|
| Storage Capacity | 723,923 | $< 2e\text{-}16$ |
| RAM Size | 1,700.3 | $< 2e\text{-}16$ |

### 3.4 Model Assumptions

To ensure the validity of the ANOVA results, we conducted diagnostic checks on model assumptions:

- The Shapiro-Wilk test produced a p-value greater than 0.05. This means we fail to reject the null hypothesis and conclude that the residuals are approximately normally distributed.
- Levene's test also resulted in a p-value greater than 0.05, supporting the assumption of equal variances across groups.
- A QQ plot and Residuals vs Fitted Values plot confirmed that there were no significant deviations from normality or homoscedasticity.

These checks validate the reliability of our ANOVA findings.

### 3.5 Post-Hoc Analysis

To explore specific group differences, we applied Tukey's HSD test:

- All pairwise comparisons between different storage capacities showed highly significant differences. This confirms that increasing storage, such as moving from 256GB to 512GB or from 512GB to 1TB, is strongly associated with a noticeable increase in laptop price.

- Comparisons across different RAM sizes also revealed statistically significant differences, although the magnitude of the effect was somewhat less pronounced compared to storage capacity.

## 4. Discussion

### 4.1 Conclusion

The results of our study provide compelling evidence that both RAM size and storage capacity have statistically significant and independent effects on laptop pricing. This means that each of these components individually contributes to the price of a laptop, and the pricing impact of one does not vary based on the level of the other. In practical terms, upgrading RAM or storage will result in an increase in price, regardless of the configuration of the other component. For instance, increasing RAM will raise the price irrespective of whether the device has a small or large storage capacity, and vice versa.

This finding simplifies the modeling of laptop pricing. It implies that manufacturers and retailers likely apply an additive strategy to these hardware specifications. Each feature contributes a separate and measurable cost to the final price, without modifying the impact of the other.

From a business perspective, this insight is especially valuable for new manufacturers entering the market. It allows them to adopt a straightforward pricing approach for their product lines. Understanding that consumers are generally willing to pay more for increased RAM or storage capacity—independently—enables these manufacturers to design pricing structures that reflect clear hardware value propositions. For example, offering configurations with 8GB, 16GB, and 32GB of RAM can each carry a distinct price premium, without needing to account for variations in storage when determining that premium.

Likewise, the results are beneficial from a consumer standpoint. The independent, additive pricing model enables consumers to better predict how much more they would need to pay for specific hardware upgrades. This transparency aids in making more informed purchasing decisions, as buyers can clearly see the cost implications of opting for higher storage or additional RAM. In turn, this predictability fosters a more competitive and consumer-friendly market environment.

Overall, the independence of RAM and storage effects on pricing reinforces the practicality of linear modeling in hardware pricing and opens pathways for both manufacturers and consumers to make data-driven decisions.

### 4.2 Limitations and Future Scope

We did not take into account several other variables that could also impact laptop prices:

- Processor performance (e.g., Intel Core i5 vs. i7, AMD Ryzen series) is a major contributor to a laptop's overall performance and market price.
- Graphics capability, particularly discrete GPUs for gaming or professional use (e.g., NVIDIA, AMD), can significantly raise a laptop's cost.
- Release date is another critical factor, as newer models typically come with a premium price due to updated features and newer technology.

Future Work:

- We can use a balanced dataset. A balanced dataset ensures that all categories are equally represented.This reduces bias in the results and improves the findings across different types of laptops.
- We can explore the effect of brand on price. The brand-related factors can cause significant price variation, even among laptops with similar technical specifications.

## 5. References

1. Dataset used was from kaggle and the link is https://www.kaggle.com/code/msjahid/exploring-laptop-price-trends/input
2. Notes from STAT-631 (Dr. Jiyoun Myung) at CSU East Bay.

### 6. Appendix

The presentation and R code with plots, diagnostic tests and our resulting output is given at https://github.com/akhil-sachar/Laptop-price-trends-STAT-631-.

The R code we used is given below:-

Our research question is - **"How do Storage Capacity and RAM Size affect laptop price, and is there an interaction between them?"**

Loading the dataset and necessary libraries

```
laptop <- read.table("Laptop_price.csv", header = TRUE, sep = ",")

library(dplyr)
library(ggplot2)
library(car)
```

Displaying summary statistics and first 6 rows of the dataset

```r
summary(laptop)
```

```
    Brand           Processor_Speed    RAM_Size     Storage_Capacity
 Length:1000        Min.   :1.512   Min.   : 4.0   Min.   : 256.0
 Class :character   1st Qu.:2.089   1st Qu.: 8.0   1st Qu.: 256.0
 Mode  :character   Median :2.761   Median :16.0   Median : 512.0
                    Mean   :2.751   Mean   :15.5   Mean   : 584.6
                    3rd Qu.:3.363   3rd Qu.:32.0   3rd Qu.:1000.0
                    Max.   :3.999   Max.   :32.0   Max.   :1000.0
  Screen_Size        Weight           Price
 Min.   :11.01   Min.   :2.001   Min.   : 8570
 1st Qu.:12.64   1st Qu.:2.717   1st Qu.:10114
 Median :14.10   Median :3.465   Median :17287
 Mean   :14.06   Mean   :3.467   Mean   :19604
 3rd Qu.:15.53   3rd Qu.:4.213   3rd Qu.:31566
 Max.   :16.99   Max.   :4.991   Max.   :33504
```

```r
head(laptop)
```

```
   Brand Processor_Speed RAM_Size Storage_Capacity Screen_Size   Weight
1   Asus        3.830296       16              512    11.18515 2.641094
2   Acer        2.912833        4             1000    11.31137 3.260012
3 Lenovo        3.241627        4              256    11.85302 2.029061
4   Acer        3.806248       16              512    12.28036 4.573865
5   Acer        3.268097       32             1000    14.99088 4.193472
6     HP        1.881348       16              256    11.94396 4.840268
      Price
1 17395.093
2 31607.606
3  9291.024
4 17436.728
5 32917.991
6  9543.720
```

Converting columns to factors for categorical analysis

```r
laptop <- laptop %>%
  mutate(across(c(Storage_Capacity, RAM_Size, Brand), as.factor))
```

Checking whether our dataset is balanced or not

```r
tapply(laptop[-1,]$Price,
laptop[-1,]$Storage_Capacity:laptop[-1,]$RAM_Size, length)
```

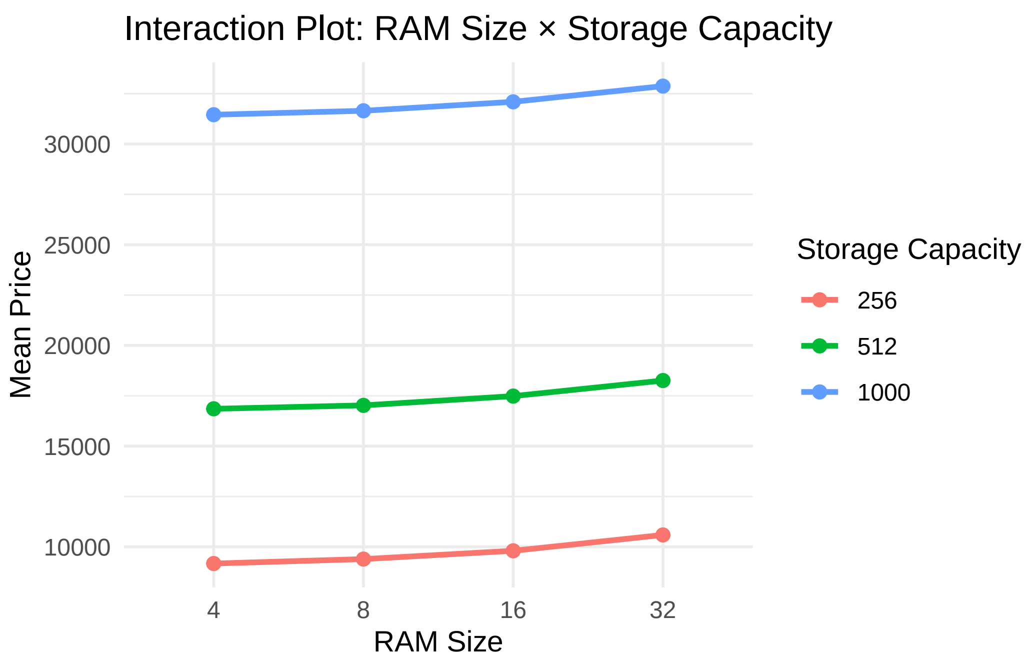
```
  256:4   256:8  256:16  256:32   512:4   512:8  512:16  512:32  1000:4  1000:8
     88      87      84      98      80      64      80      82      75      95
1000:16 1000:32
     72      94
```

Checking structure and data types of the dataset

```r
str(laptop)
```

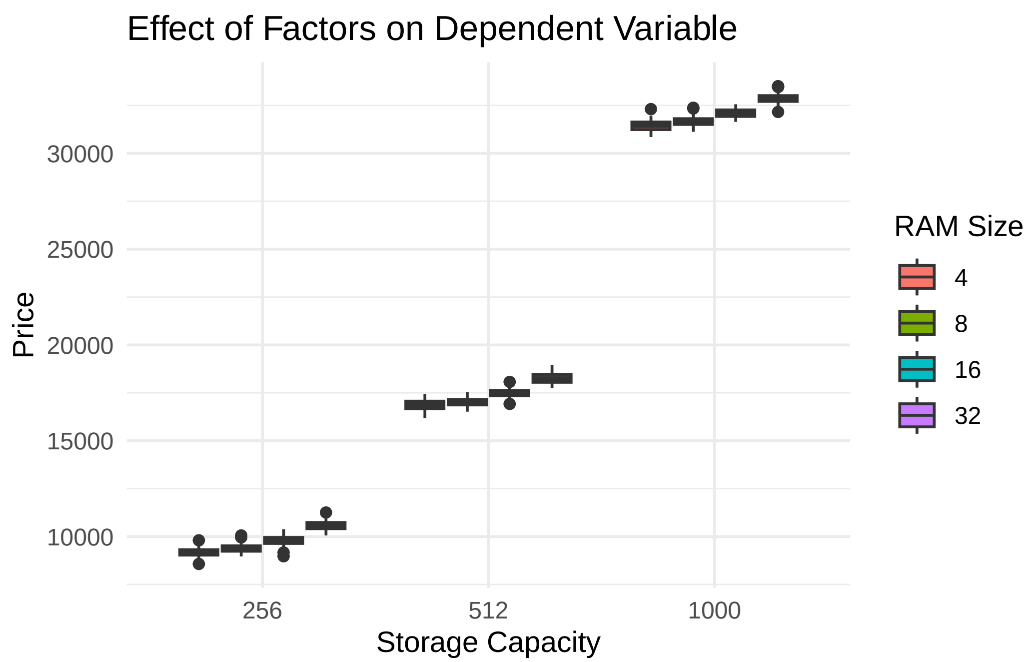
```
'data.frame':	1000 obs. of  7 variables:
 $ Brand           : Factor w/ 5 levels "Acer","Asus",..: 2 1 5 1 1 4 5 5 5 1 ...
 $ Processor_Speed : num  3.83 2.91 3.24 3.81 3.27 ...
 $ RAM_Size        : Factor w/ 4 levels "4","8","16","32": 3 1 1 3 4 3 4 1 1 4 ...
 $ Storage_Capacity: Factor w/ 3 levels "256","512","1000": 2 3 1 2 3 1 1 1 2 2 ...
 $ Screen_Size     : num  11.2 11.3 11.9 12.3 15 ...
 $ Weight          : num  2.64 3.26 2.03 4.57 4.19 ...
 $ Price           : num  17395 31608 9291 17437 32918 ...
```

```r
# Summarize mean price by group
summary_data <- laptop %>%
  group_by(RAM_Size, Storage_Capacity) %>%
  summarise(mean_price = mean(Price, na.rm = TRUE), .groups = "drop")

# Interaction plot
ggplot(summary_data, aes(x = RAM_Size, y = mean_price, color = Storage_Capacity, group = Sto
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  theme_minimal() +
  labs(title = "Interaction Plot: RAM Size × Storage Capacity",
       x = "RAM Size",
       y = "Mean Price",
       color = "Storage Capacity")
```




Boxplot to visualize distribution of Price by Storage_Capacity and RAM_Size

```
ggplot(laptop, aes(x = Storage_Capacity, y = Price, fill = RAM_Size)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Effect of Factors on Dependent Variable",
       x = "Storage Capacity",
       y = "Price",
       fill = "RAM Size")
```




Fitting Two-Way ANOVA: Storage_Capacity and RAM_Size

```r
# Fitting Two-Way ANOVA: Storage_Capacity and RAM_Size
model <- aov(Price ~ Storage_Capacity * RAM_Size, data = laptop)
Anova(model,type=2)
```

```
Anova Table (Type II tests)

Response: Price
                               Sum Sq  Df    F value Pr(>F)
Storage_Capacity           8.7763e+10   2 720444.501 <2e-16 ***
RAM_Size                   3.0919e+08   3   1692.108 <2e-16 ***
Storage_Capacity:RAM_Size  7.4541e+04   6      0.204 0.9756
Residuals                  6.0178e+07 988
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A **p-value $<$ 0.05** indicates a **statistically significant** effect.

For both Storage_Capacity and RAM_Size have p value less than 0.05.

- **Storage_Capacity** has a strong and significant impact on laptop price.

- **RAM_Size** also has a strong and significant impact on laptop price.

Fitting updated Two-Way ANOVA: Storage_Capacity and RAM_Size

```
# Fitting updated Two-Way ANOVA: Storage_Capacity and RAM_Size
redmodel <- aov(Price ~ Storage_Capacity + RAM_Size, data = laptop)
Anova(redmodel,type=2)
```

```
Anova Table (Type II tests)

Response: Price
                     Sum Sq  Df  F value    Pr(>F)    
Storage_Capacity 8.7763e+10   2 723923.0 < 2.2e-16 ***
RAM_Size         3.0919e+08   3   1700.3 < 2.2e-16 ***
Residuals        6.0252e+07 994                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

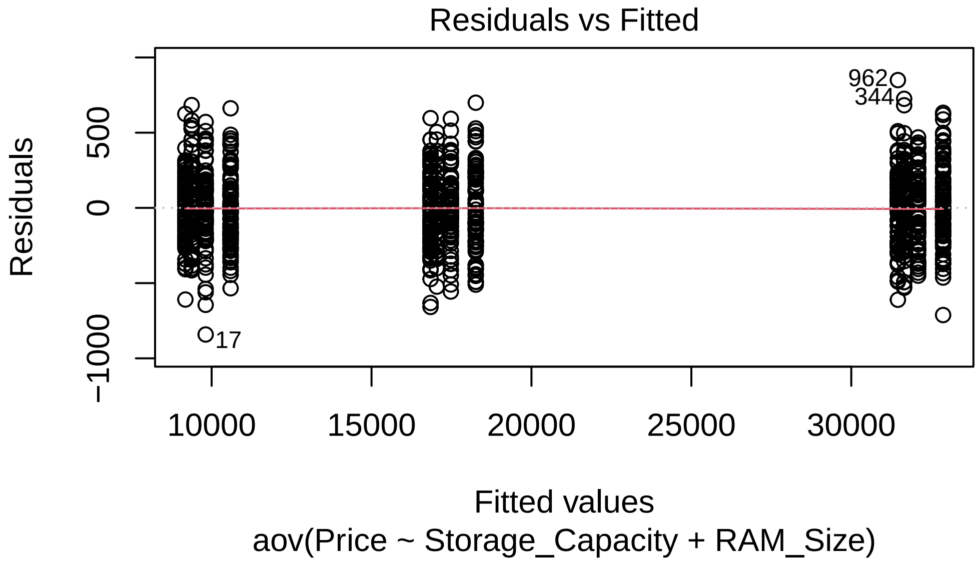
Shapiro-Wilk test for normality of residuals

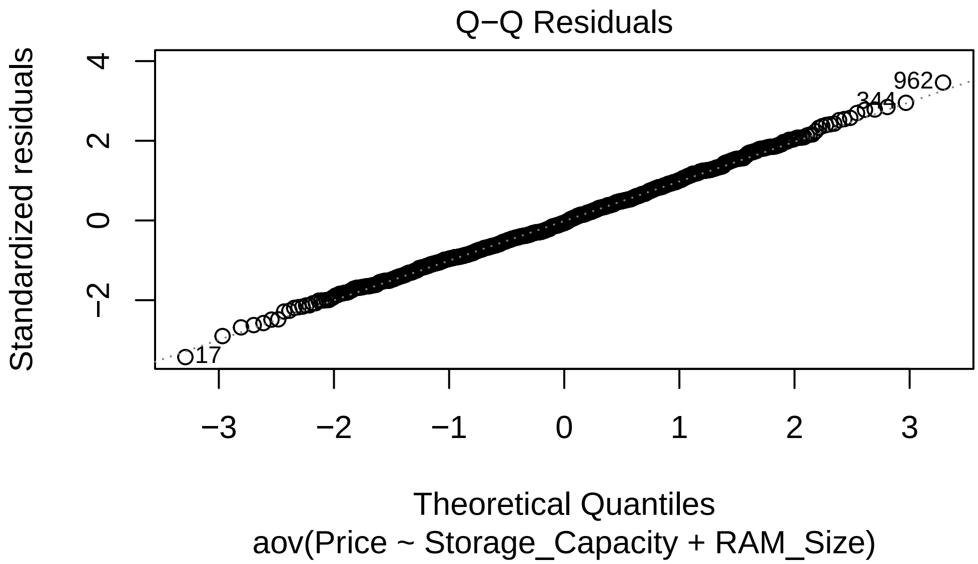
```
shapiro.test(residuals(redmodel))
```

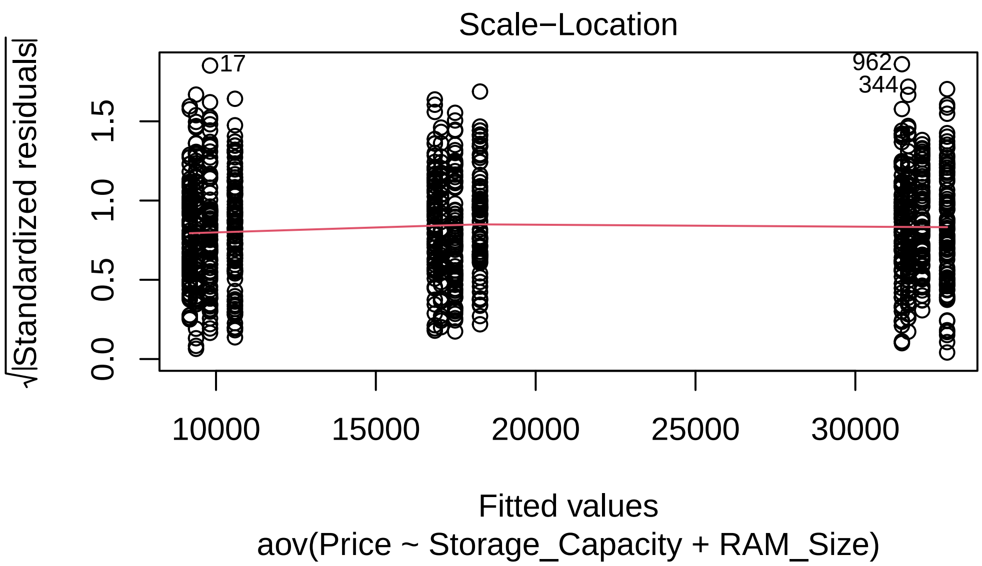
```

	Shapiro-Wilk normality test

data:  residuals(redmodel)
W = 0.99825, p-value = 0.4054
```

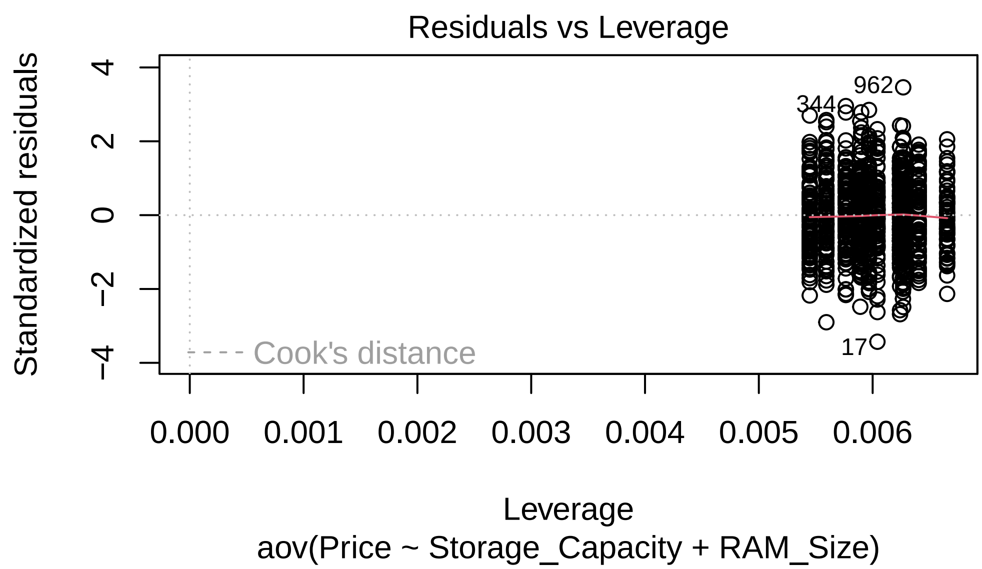
Plotting to check for assumptions

```
plot(redmodel)
```


Residuals vs Fitted
Residuals
-1000
0
500
962
344
17
10000
15000
20000
25000
30000
Fitted values
aov(Price ~ Storage_Capacity + RAM_Size)


Q−Q Residuals
Standardized residuals
-2
0
2
4
962
344
17
−3
−2
−1
0
1
2
3
Theoretical Quantiles
aov(Price ~ Storage_Capacity + RAM_Size)




- **Residuals vs Fitted** - linearity and equal variance (homoscedasticity) assumptions are met
- **Q-Q plot** - Normality assumption is satisfied.

- **Scale-Location** - Homoscedasticity assumption satisfied.
- **Residuals vs Leverage** - No concerning outliers.
- **All assumptions are satisfied**.

Levene's test to check equality of variances

```
leveneTest(Price ~ Storage_Capacity * RAM_Size, data = laptop)
```

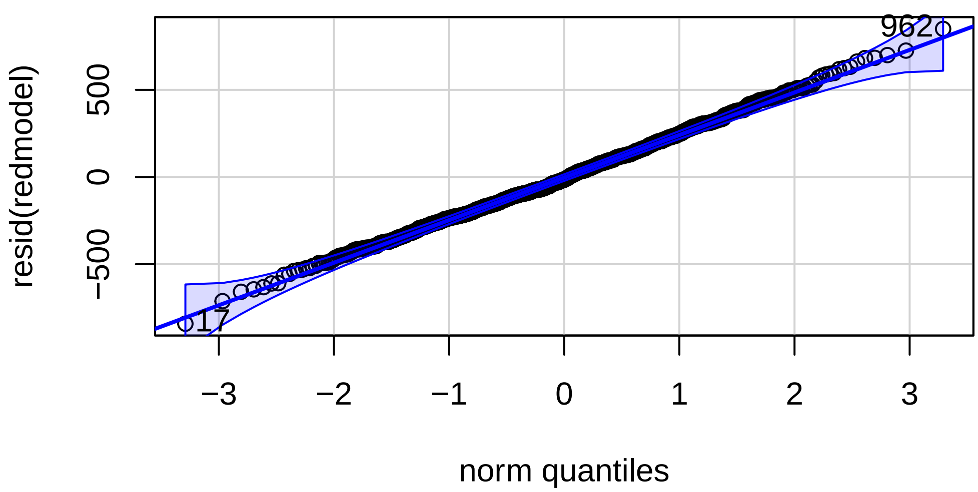
```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group  11  0.9494 0.4917
      988
```

```
#Equal variances are satisfied
```

**Residuals are normal**

Q-Q plot for reduced model residuals

```
qqPlot(resid(redmodel))
```



```
[1] 962  17
```

Tukey HSD test for pairwise comparisons among groups

```r
tukey <- TukeyHSD(redmodel)
tukey
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Price ~ Storage_Capacity + RAM_Size, data = laptop)

$Storage_Capacity
             diff       lwr       upr p adj
512-256   7666.62  7621.639  7711.601     0
1000-256 22280.25 22236.320 22324.170     0
1000-512 14613.63 14567.999 14659.252     0

$RAM_Size
           diff       lwr       upr p adj
8-4    196.2716  138.9680  253.5753     0
16-4   633.9371  576.0953  691.7789     0
32-4  1413.9737 1358.1440 1469.8034     0
16-8   437.6655  379.9981  495.3329     0
32-8  1217.7021 1162.0531 1273.3511     0
32-16  780.0366  723.8336  836.2396     0
```