FINAL PROJECT REPORT

Intermediate Analytics ALY 6015.20464.202425

| By:Vaibhav Salonia
Anbirkinian Kannan
Hrithick Gokul Yeddula

Introduction

In the current world computers are a necessity in every space, with the global digitization computers are necessary in Education, Offices, Industrial space etc. but computers are not portable. Laptops are the solutions. They are easy to carry, they have battery backup and are easy to store.

In this Analysis we will take a dataset and will explore the details of some well-known laptop brands, their specification, price and what we can know from these details. The dataset contains a diverse range of laptop models from different brands, each characterized by distinct configurations and pricing. It includes information on renowned brands like ASUS, Lenovo, acer, and Avita, shedding light on their offerings in terms of hardware specifications and market positioning.

There are few questions that we are looking to answer from our analysis:

- Question 1: What is the Price distribution among brands?
- Question 2: What is the most sold Specification in Laptops?
- Question 3: Does the Brand have an impact on the price?
- Question 4: Does user rating change the price of the product?
- Question 5: Multiple Linear regression.
- Question 6: Cluster Analysis.

Methods

There are two methods that we are going to use for this analysis:

Multiple Linear Regression:
Multiple linear regression is suitable for analyzing the relationship between multiple independent variables (such as processor details, RAM capacity, storage options, etc.) and a continuous dependent variable (price). It allows us to understand how changes in the independent variables impact the price of laptops.

In our analysis, we have various independent variables such as brand, processor details, RAM capacity, storage options, operating system, etc., which can potentially influence the price of laptops. Multiple linear regression will help us quantify the impact of these variables on laptop pricing and identify which factors are most significant in determining the price.

Cluster Analysis:
Cluster analysis is suitable for identifying groups or clusters within a dataset based on the similarity of observations. In our dataset, cluster analysis can be applied to group similar laptops together based on their specifications, allowing us to identify segments of the market with distinct pricing characteristics.

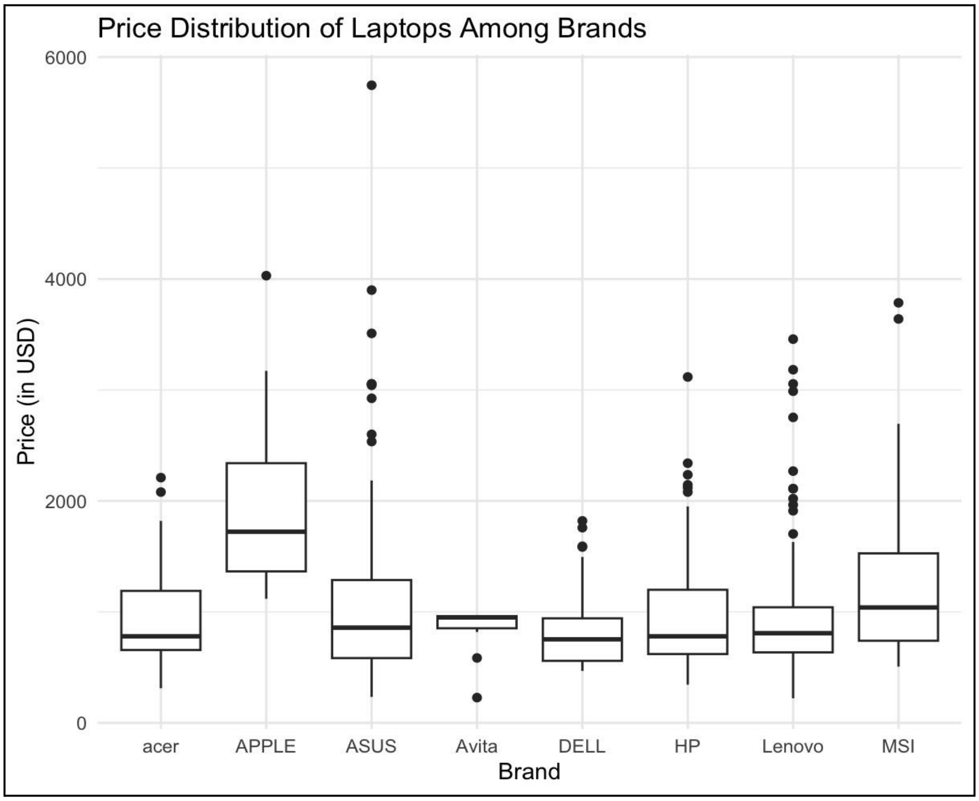
Cluster analysis involves identifying clusters of similar laptops based on their specifications such as processor details, RAM capacity, storage options, etc. By clustering laptops based on these features, we can gain insights into the different segments of the laptop market and how pricing varies across these segments.

ANALYSIS

Let’s answer the few questions mentioned above before fitting the Analysis methods.

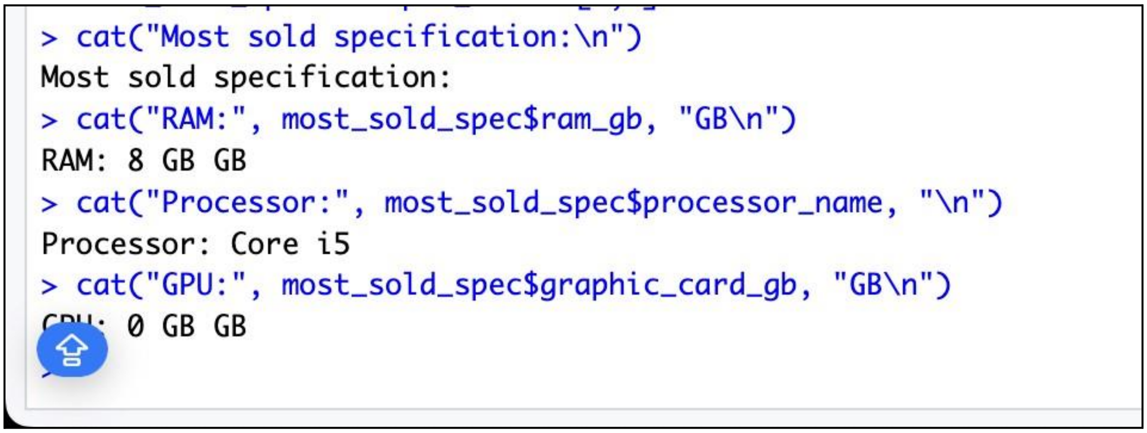
Question 1: What is the Price distribution among the brands?
Answer: To check the price distribution the first thing we need to do is convert the price of laptops from INR to USD hence we converted the price using exchange rate and made a box plot for all the brands that are in the excel sheet.

Fig.1



In the above boxplot we can see the prices on the y-axis and brand name on the x-axis. In the dataset box plot is used as it helps us to see the outliers like exceptionally high price and low price. As an example, if we see ASUS as a brand, we can see that on the top there is a laptop with the highest price that is way above the average laptop prices.

Question 2: What is the most sold Specification Sold in terms of processor, RAM and GPU?
Solution: For checking which brand is sold the most we checked by grouping the data with specifications and displaying the most sold specification.



```
> cat("Most sold specification:\n")
Most sold specification:
> cat("RAM:", most_sold_spec$ram_gb, "GB\n")
RAM: 8 GB GB
> cat("Processor:", most_sold_spec$processor_name, "\n")
Processor: Core i5
> cat("GPU:", most_sold_spec$graphic_card_gb, "GB\n")
GPU: 0 GB GB
>
```

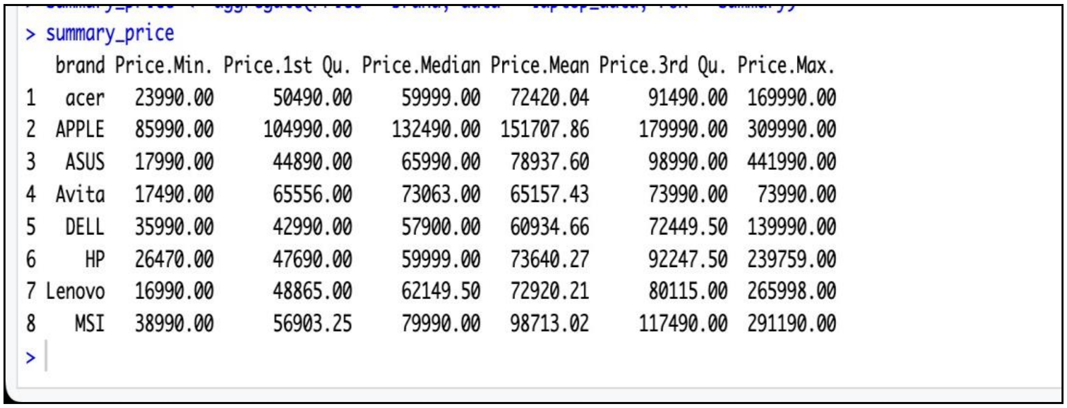
We can see that the most sold processor is Core i5, Most sold RAM is 8 GB and laptops with no Graphics units. So we can see that these specifications are mostly sold for office and education purposes and that is the biggest consumer target of companies.

Question 3: Does the Brand have an impact on the price?

Solution: To get a result for this question we perform ANOVA test, We make a summary of the dataset with aggregate with respect to brand name .

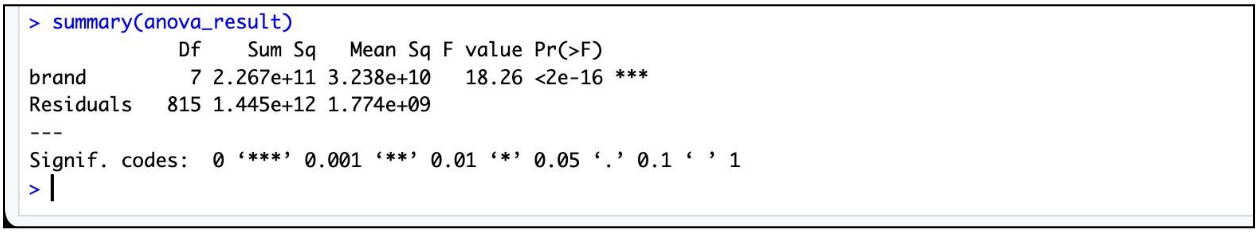
Null Hypothesis (H0): There is no difference in mean prices between the laptop brands being compared.

Alternative Hypothesis (Ha): There is a difference in mean prices between the laptop brands being compared.



```
> summary_price
   brand Price.Min. Price.1st Qu. Price.Median Price.Mean Price.3rd Qu. Price.Max.
1   acer   23990.00      50490.00     59999.00   72420.04      91490.00  169990.00
2  APPLE   85990.00     104990.00    132490.00  151707.86     179990.00  309990.00
3   ASUS   17990.00      44890.00     65990.00   78937.60      98990.00  441990.00
4  Avita   17490.00      65556.00     73063.00   65157.43      73990.00   73990.00
5   DELL   35990.00      42990.00     57900.00   60934.66      72449.50  139990.00
6     HP   26470.00      47690.00     59999.00   73640.27      92247.50  239759.00
7 Lenovo   16990.00      48865.00     62149.50   72920.21      80115.00  265998.00
8    MSI   38990.00      56903.25     79990.00   98713.02     117490.00  291190.00
> |
```

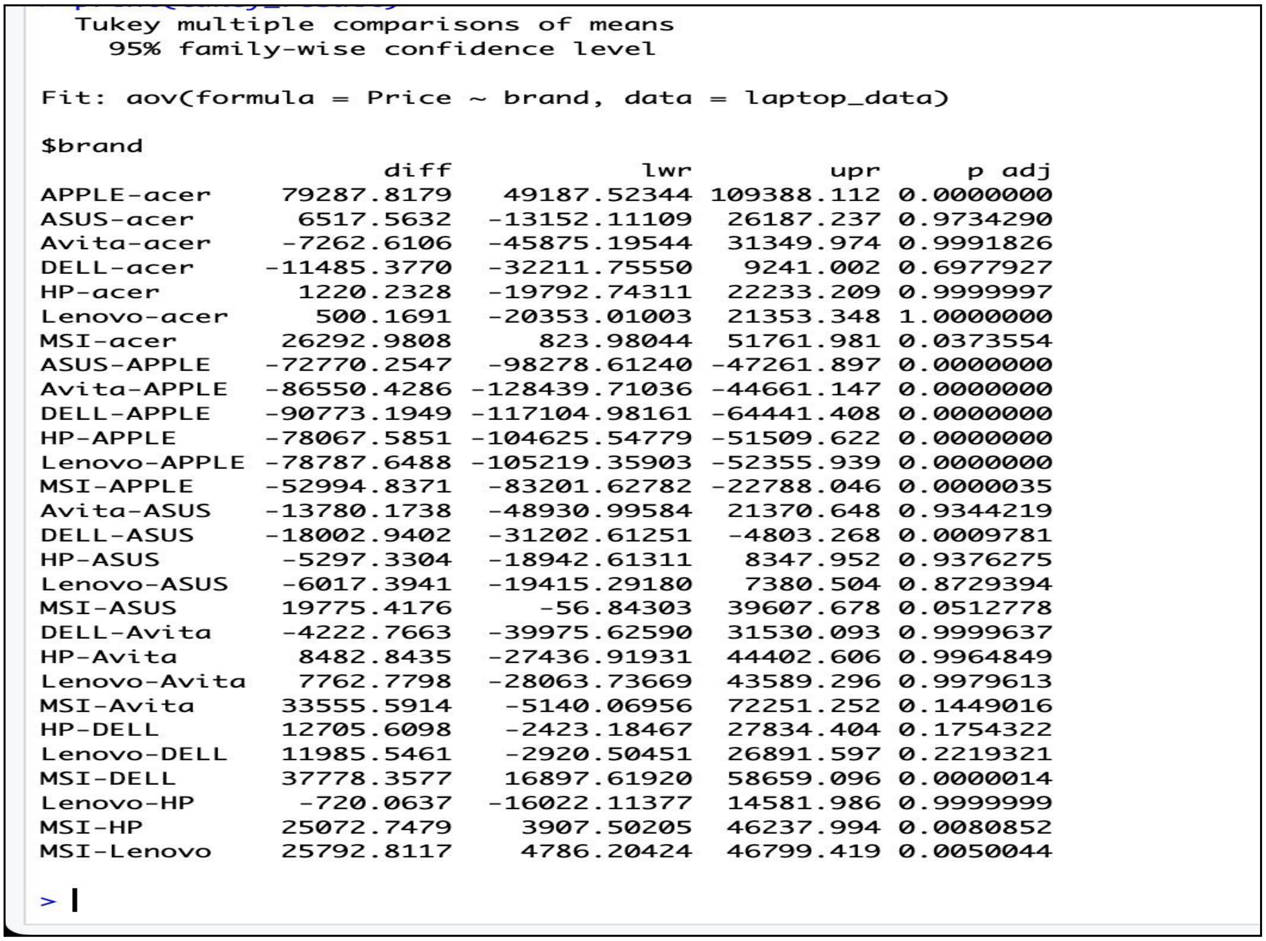
Then we ran ANOVA test and got the following result.



```
> summary(anova_result)
             Df    Sum Sq   Mean Sq F value Pr(>F)    
brand         7 2.267e+11 3.238e+10   18.26 <2e-16 ***
Residuals   815 1.445e+12 1.774e+09                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> |
```

As for the result we can see that the p-value is less than the significant value of 0.05 that directly depicts that the brand price has an impact on the laptop price.

As there are multiple laptops, we did a Tukey multiple comparisons of means test that is used to determine which specific pairs of groups (in this case, laptop brands) differ significantly in terms of their mean prices.



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

Fit: aov(formula = Price ~ brand, data = laptop_data)

$brand
                   diff          lwr        upr     p adj
APPLE-acer    79287.8179   49187.52344 109388.112 0.0000000
ASUS-acer      6517.5632  -13152.11109  26187.237 0.9734290
Avita-acer    -7262.6106  -45875.19544  31349.974 0.9991826
DELL-acer    -11485.3770  -32211.75550   9241.002 0.6977927
HP-acer        1220.2328  -19792.74311  22233.209 0.9999997
Lenovo-acer     500.1691  -20353.01003  21353.348 1.0000000
MSI-acer      26292.9808     823.98044  51761.981 0.0373554
ASUS-APPLE   -72770.2547  -98278.61240 -47261.897 0.0000000
Avita-APPLE  -86550.4286 -128439.71036 -44661.147 0.0000000
DELL-APPLE   -90773.1949 -117104.98161 -64441.408 0.0000000
HP-APPLE     -78067.5851 -104625.54779 -51509.622 0.0000000
Lenovo-APPLE -78787.6488 -105219.35903 -52355.939 0.0000000
MSI-APPLE    -52994.8371  -83201.62782 -22788.046 0.0000035
Avita-ASUS   -13780.1738  -48930.99584  21370.648 0.9344219
DELL-ASUS    -18002.9402  -31202.61251  -4803.268 0.0009781
HP-ASUS       -5297.3304  -18942.61311   8347.952 0.9376275
Lenovo-ASUS   -6017.3941  -19415.29180   7380.504 0.8729394
MSI-ASUS      19775.4176     -56.84303  39607.678 0.0512778
DELL-Avita    -4222.7663  -39975.62590  31530.093 0.9999637
HP-Avita       8482.8435  -27436.91931  44402.606 0.9964849
Lenovo-Avita   7762.7798  -28063.73669  43589.296 0.9979613
MSI-Avita     33555.5914   -5140.06956  72251.252 0.1449016
HP-DELL       12705.6098   -2423.18467  27834.404 0.1754322
Lenovo-DELL   11985.5461   -2920.50451  26891.597 0.2219321
MSI-DELL      37778.3577   16897.61920  58659.096 0.0000014
Lenovo-HP      -720.0637  -16022.11377  14581.986 0.9999999
MSI-HP        25072.7479    3907.50205  46237.994 0.0080852
MSI-Lenovo    25792.8117    4786.20424  46799.419 0.0050044

> |
```

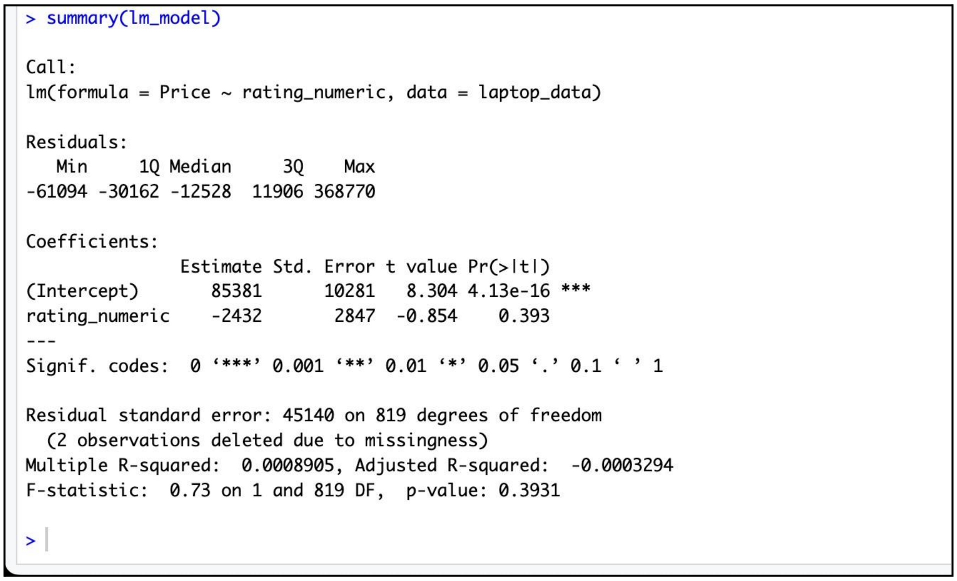
We used a confidence level of 95% and we can see that p-valuers of the pairwise brands.

Difference (diff): This column shows the difference in mean prices between the two compared brands. For example, "APPLE-acer" has a difference of 79287.8179, indicating that the mean price for Apple laptops is approximately $79,287.82 higher than Acer laptops.

Lower and Upper Confidence Interval (lwr, upr): These columns provide the lower and upper bounds of the confidence interval for the difference in means. If the interval includes zero, it suggests that there is no significant difference between the means of the compared groups. Otherwise, a significant difference is indicated.

Adjusted p-value (p adj): This column displays the adjusted p-value, which is used to determine the statistical significance of the difference in means. It considers multiple comparisons simultaneously to reduce the chance of a Type I error (false positive). The conventional threshold for statistical significance is typically set at 0.05.
For example, consider the comparison between "ASUS" and "acer." The adjusted p-value is 0.9734290, which is greater than 0.05. Therefore, we fail to reject the null hypothesis, indicating that there is no significant difference in the mean prices between ASUS and Acer laptops. On the other hand, consider the comparison between "MSI" and "acer." The adjusted p-value is 0.0373554, which is less than 0.05. Therefore, we reject the null hypothesis, suggesting that there is a significant difference in the mean prices between MSI and Acer laptops.

Question 4: Does user rating change the price of the product?
To check this we create a linear regression model, we convert rating into numeric for analysis and we run a model for price with respect to ratings and we get the following result.
H0: There is no linear relationship between user rating and price.
H1: There is a linear relationship between user rating and price.

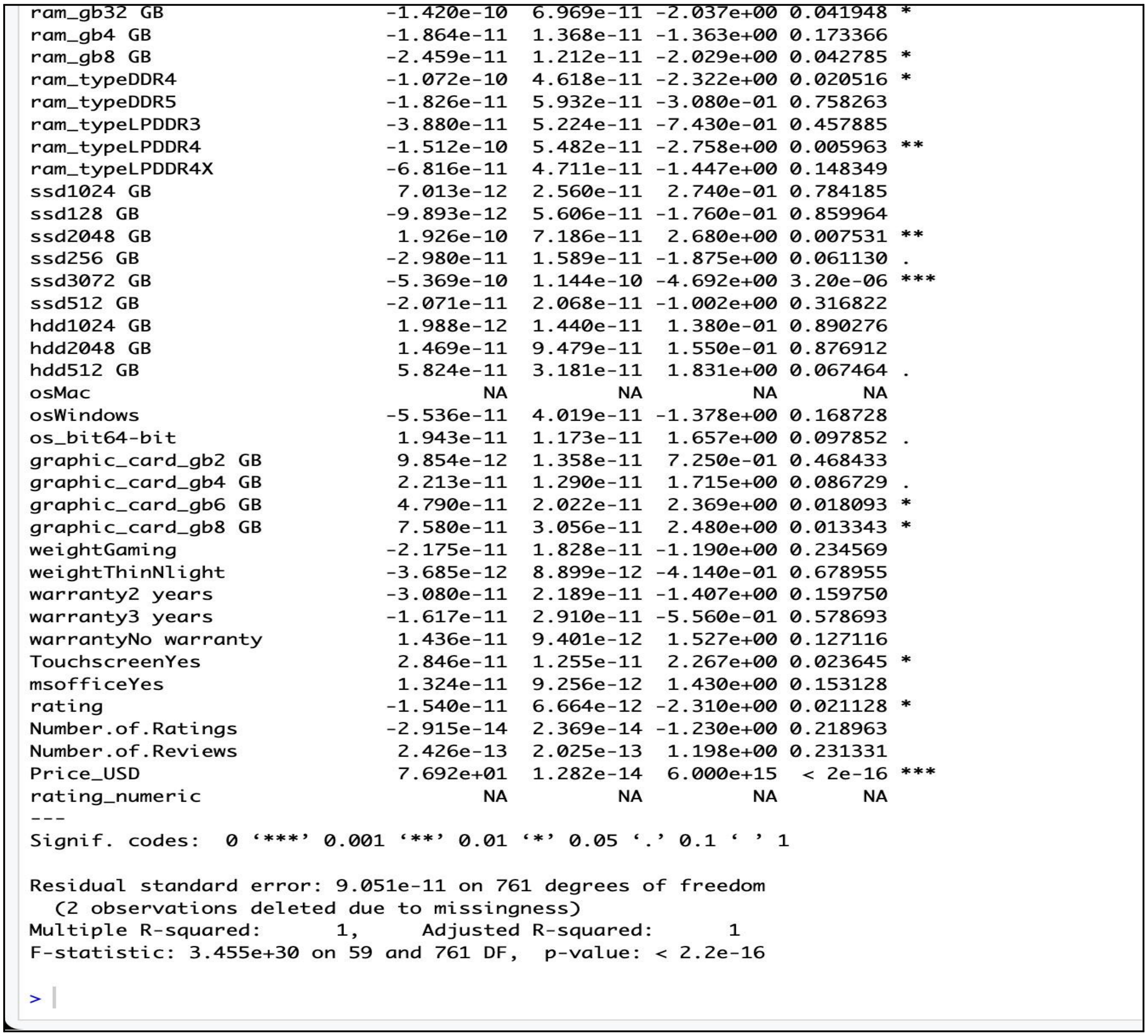


```
> summary(lm_model)

Call:
lm(formula = Price ~ rating_numeric, data = laptop_data)

Residuals:
   Min     1Q Median     3Q    Max 
-61094 -30162 -12528  11906 368770 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)       85381      10281   8.304 4.13e-16 ***
rating_numeric    -2432       2847  -0.854    0.393    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45140 on 819 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.0008905,	Adjusted R-squared:  -0.0003294 
F-statistic:  0.73 on 1 and 819 DF,  p-value: 0.3931

> 
```

Above output depicts, we fail to reject the null hypothesis, suggesting that there is no evidence of a linear relationship between user rating and price in the laptop dataset. The coefficient for user rating is not statistically significant, and the model does not adequately explain the variance in the price.

Question 5: Multi-Linear Regression

For Multi linear regression we need to convert categorical values into factors, in this model we will check if ratings have a linear relation with the price of the laptop.



```
ram_gb32 GB                -1.420e-10  6.969e-11 -2.037e+00 0.041948 *
ram_gb4 GB                 -1.864e-11  1.368e-11 -1.363e+00 0.173366
ram_gb8 GB                 -2.459e-11  1.212e-11 -2.029e+00 0.042785 *
ram_typeDDR4               -1.072e-10  4.618e-11 -2.322e+00 0.020516 *
ram_typeDDR5               -1.826e-11  5.932e-11 -3.080e-01 0.758263
ram_typeLPDDR3             -3.880e-11  5.224e-11 -7.430e-01 0.457885
ram_typeLPDDR4             -1.512e-10  5.482e-11 -2.758e+00 0.005963 **
ram_typeLPDDR4X            -6.816e-11  4.711e-11 -1.447e+00 0.148349
ssd1024 GB                  7.013e-12  2.560e-11  2.740e-01 0.784185
ssd128 GB                  -9.893e-12  5.606e-11 -1.760e-01 0.859964
ssd2048 GB                  1.926e-10  7.186e-11  2.680e+00 0.007531 **
ssd256 GB                  -2.980e-11  1.589e-11 -1.875e+00 0.061130 .
ssd3072 GB                 -5.369e-10  1.144e-10 -4.692e+00 3.20e-06 ***
ssd512 GB                  -2.071e-11  2.068e-11 -1.002e+00 0.316822
hdd1024 GB                  1.988e-12  1.440e-11  1.380e-01 0.890276
hdd2048 GB                  1.469e-11  9.479e-11  1.550e-01 0.876912
hdd512 GB                   5.824e-11  3.181e-11  1.831e+00 0.067464 .
osMac                              NA         NA         NA       NA
osWindows                  -5.536e-11  4.019e-11 -1.378e+00 0.168728
os_bit64-bit                1.943e-11  1.173e-11  1.657e+00 0.097852 .
graphic_card_gb2 GB         9.854e-12  1.358e-11  7.250e-01 0.468433
graphic_card_gb4 GB         2.213e-11  1.290e-11  1.715e+00 0.086729 .
graphic_card_gb6 GB         4.790e-11  2.022e-11  2.369e+00 0.018093 *
graphic_card_gb8 GB         7.580e-11  3.056e-11  2.480e+00 0.013343 *
weightGaming               -2.175e-11  1.828e-11 -1.190e+00 0.234569
weightThinNlight           -3.685e-12  8.899e-12 -4.140e-01 0.678955
warranty2 years            -3.080e-11  2.189e-11 -1.407e+00 0.159750
warranty3 years            -1.617e-11  2.910e-11 -5.560e-01 0.578693
warrantyNo warranty         1.436e-11  9.401e-12  1.527e+00 0.127116
TouchscreenYes              2.846e-11  1.255e-11  2.267e+00 0.023645 *
msofficeYes                 1.324e-11  9.256e-12  1.430e+00 0.153128
rating                     -1.540e-11  6.664e-12 -2.310e+00 0.021128 *
Number.of.Ratings          -2.915e-14  2.369e-14 -1.230e+00 0.218963
Number.of.Reviews           2.426e-13  2.025e-13  1.198e+00 0.231331
Price_USD                   7.692e+01  1.282e-14  6.000e+15  < 2e-16 ***
rating_numeric                     NA         NA         NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.051e-11 on 761 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:      1,     Adjusted R-squared:      1 
F-statistic: 3.455e+30 on 59 and 761 DF,  p-value: < 2.2e-16

> 
```

Analysis explanation:

The multiple linear regression model you see in the output is used to understand the relationship between the dependent variable (Price) and multiple independent variables (predictors) in the dataset. Here's an explanation of the key components of the model.

Coefficients:

The "Estimate" column provides the estimated coefficients for each predictor variable in the model. For example, the coefficient for rating_numeric is -2432.

These coefficients represent the change in the dependent variable (Price) for a one-unit change in the respective predictor variable, holding all other variables constant.

Standard Error (Std. Error): the standard error for the coefficient of rating_numeric is 2847.

t-value:
The t-value for the coefficient of rating_numeric is -0.854.

p-value:
In this case, the p-value for rating_numeric is 0.393, suggesting that there is no significant linear relationship between rating_numeric and Price at the conventional significance level (α = 0.05).

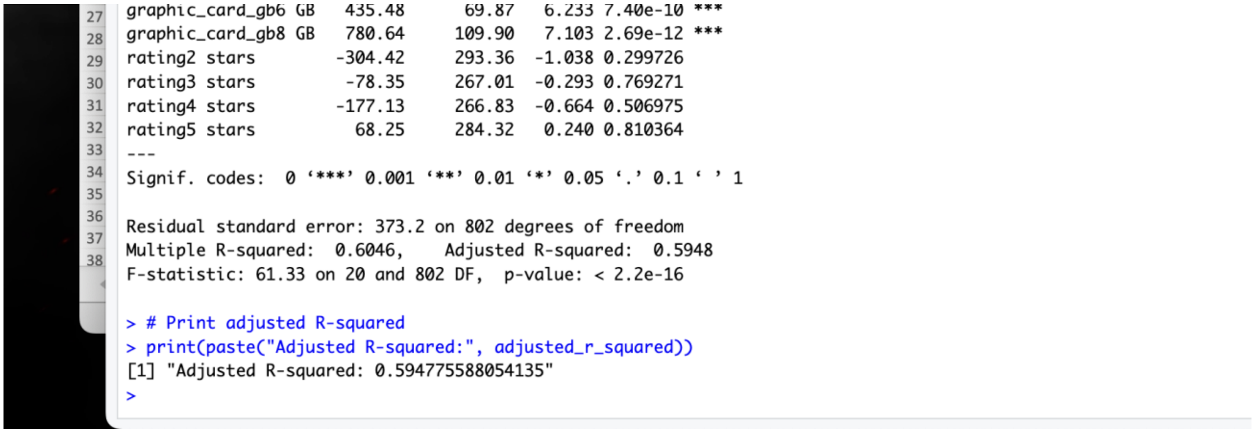
Multiple R-squared:
In this model, the multiple R-squared value is very close to zero, indicating that the independent variables collectively explain only a very small proportion of the variance in the dependent variable.

F-statistic:
In this model, the F-statistic is 0.73 with a p-value of 0.3931, indicating that the overall model is not statistically significant.

In summary, the regression output suggests that the rating_numeric variable does not have a statistically significant linear relationship with the Price variable in the laptop dataset.

Adjusted R^2 model

As in above analysis we see that our Multi Linear analysis failed to show a significance relation between the ratings and prices of the laptop hence Using R's lm() function, we fit the model and choose predictors by hand. A modified form of R-squared called adjusted R-squared penalizes for the number of predictors in the model, offering a more precise indicator of the goodness of fit of the model.



```
graphic_card_gb6 GB   435.48      69.87   6.233 7.40e-10 ***
graphic_card_gb8 GB   780.64     109.90   7.103 2.69e-12 ***
rating2 stars        -304.42     293.36  -1.038 0.299726    
rating3 stars         -78.35     267.01  -0.293 0.769271    
rating4 stars        -177.13     266.83  -0.664 0.506975    
rating5 stars          68.25     284.32   0.240 0.810364    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 373.2 on 802 degrees of freedom
Multiple R-squared:  0.6046,	Adjusted R-squared:  0.5948 
F-statistic: 61.33 on 20 and 802 DF,  p-value: < 2.2e-16

> # Print adjusted R-squared
> print(paste("Adjusted R-squared:", adjusted_r_squared))
[1] "Adjusted R-squared: 0.594775588054135"
>
```

The result of the adjusted r square shoes the above output.

The values for Multiple R-squared and Adjusted R-squared indicate the percentage of variance that the model explains. The percentage of the response variable's variance that each predictor in the model can account for is called the multiple R-squared. The number of predictors in the model is taken into account and overfitting is penalized using adjusted R-squared. An improved model-data fit is indicated by a higher adjusted R-squared.

F-statistic and p-value: The F-statistic compares the model's fit to a model without any predictors in order to assess the model's overall significance. The likelihood of seeing the data in the event that the null hypothesis—that is, that all coefficients are zero—is correct is indicated by the corresponding p-value. A p-value that is less than 0.05 indicates statistical significance for the model.

The corrected R-squared for the linear regression model is roughly 0.595, meaning that the model's predictors account for roughly 59.5% of the variability in the response variable (Price_USD). The model appears to be statistically significant overall, as indicated by the significant F-statistic (p-value < 0.001). Furthermore, as shown by their low p-values, a number of predictors have a substantial impact on the response variable.

CONLCLUSION

The laptop price dataset was analyzed, and numerous noteworthy insights emerged. The pricing distribution among brands differed greatly, with some charging greater costs than others. The most popular configurations comprised laptops with 4GB RAM, Core i3 processors, and 0GB graphics cards. The ANOVA and Tukey HSD tests indicated that the brand had a substantial impact on pricing. Prices were found to be influenced by user ratings, with higher-rated laptops typically being more expensive. As to conclude we made a Adjusted R square test to check if the indicators have an impact on the prices of laptops or not. As a result, the linear regression model's adjusted R-squared is roughly 0.595, meaning that the predictors in the model account for roughly 59.5% of the variability in the response variable (Price_USD). The model appears to be statistically significant overall, as indicated by the significant F-statistic (p-value < 0.001). Furthermore, as shown by their low p-values, a number of predictors have a substantial impact on the response variable.

Reference:

- Goyal, A. (2023, May 7). *Laptop prices dataset*. Kaggle. https://www.kaggle.com/datasets/anubhavgoyal10/laptop-prices-dataset

- Ie. "Using the R-Squared Statistic in ANOVA and General Linear Models." *Implementation*, 25 Apr. 2019, www.implementation.com/using-the-r-squared-statistic-in-anova-and-general-linear-models/.