LAPTOP SALES DATA ANALYSIS IN EXCEL

Submitted by: Pall Laura

In this project, we explore the relationships between various laptop specifications and their prices using data analysis and regression modeling in Microsoft Excel.

By examining factors such as screen size, CPU speed, RAM, and GPU type, we aim to uncover insights into how these attributes influence laptop pricing.

Through visualizations and statistical analysis, we seek to develop a predictive model that can help us better understand and predict laptop prices based on their key features.

Data source: https://www.kaggle.com/datasets/ehtishamsadiq/uncleaned-laptop-price-dataset

CLEANING DATA:

1. **Remove Duplicates**: I've got duplicate rows in my dataset, so I'll start by removing them. To do this, I'll select the entire data range, go to the "Data" tab, and click on "Remove Duplicates." Then, I'll choose the columns I want to check for duplicates (likely all columns in my case) and click "OK."

2. **Fix Column Headers**: I see that my column headers need adjustment. I'll simply click on the cell containing the header I want to change, type in the new header, and press Enter.

3. **Data Type Conversion**: Some columns, like "Inches," have data with commas instead of periods for decimal places (e.g., "13,3" instead of "13.3"). To fix this, I'll create a new column next to it and use the formula `=SUBSTITUTE(A2, ",", ".")` (assuming my data starts from the second row). I'll then drag this formula down to apply it to all rows and copy-paste the values back into the "Inches" column.

4. **Cleaning Text Data**: I'll inspect columns like "ScreenResolution," "Cpu," "Ram," "Memory," "Gpu," or "OpSys" for inconsistent or messy data. For instance, I can split the "ScreenResolution" column into two by using Excel's Text to Columns feature.

5. **Remove Special Characters**: If there are unwanted special characters or symbols in my data, I can use the "Find and Replace" feature (Ctrl + H) to locate and replace them with nothing or the appropriate value.

6. **Check for Missing Data**: I'll look for missing values (empty cells) in my dataset and decide how to handle them, whether by filling in the missing data or removing rows with missing information, based on their importance.

7. **Formatting**: I'll ensure consistent formatting, such as date formats, currency symbols, and number formatting across my dataset.

8. **Check for Outliers**: I'll examine numerical columns, such as "Weight" and "Price," for outliers using Excel functions like IQR (Interquartile Range) or Z-score. I'll consider whether to remove or adjust these outliers.

9. **Data Validation**: I'll apply data validation rules to certain columns, like "Type Name" or "OpSys," to ensure that data adheres to specific categories or values.

10. **Save the Cleaned Data**: After cleaning my data, I'll save it as a new file or overwrite the existing one, depending on my preference.

DATA ANALYSIS WITH VISUALIZATION

Here are 5 questions that can be answered using the provided data:

1. What is the average screen size (in inches) of the laptops in the dataset?

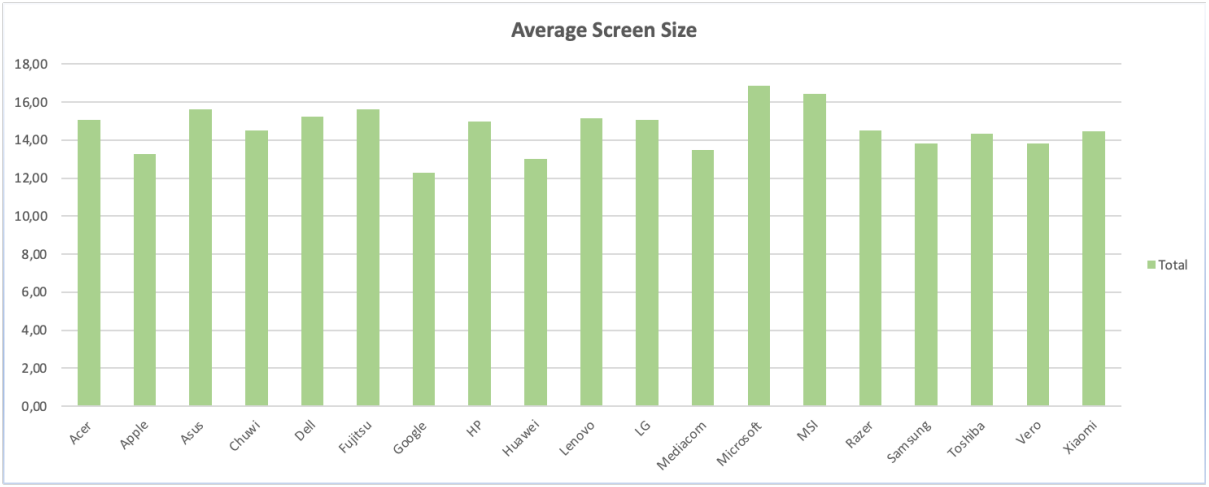

Figure 1: Average screen size of the laptops

2. Which company has the highest average laptop price?

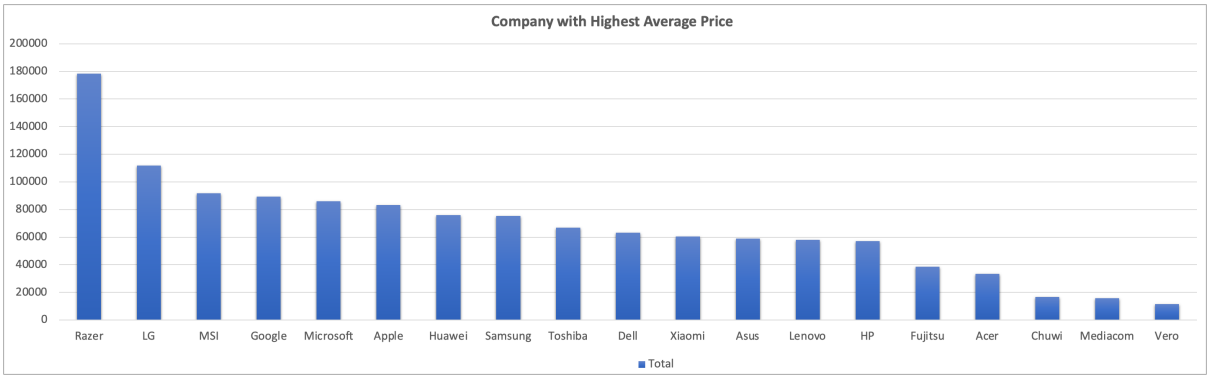

Figure 2: Highest average laptop price

3. What is the most common operating system (OpSys) among the laptops in the dataset?

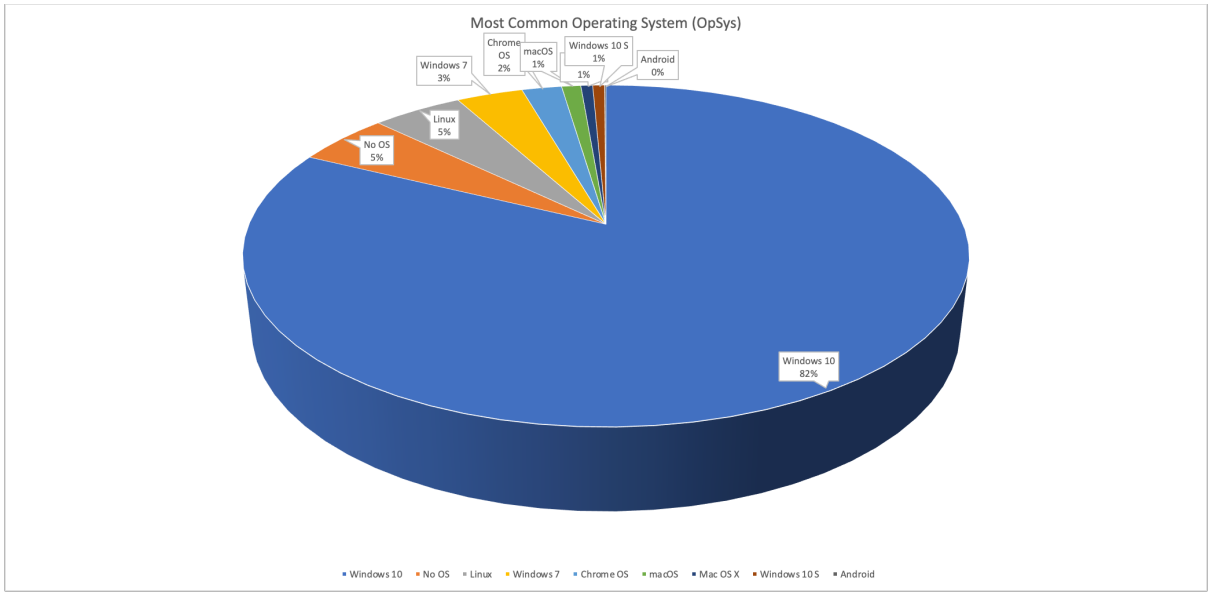

Figure 3: Most common operating system

4. Which laptop has the highest amount of RAM (in GB)?

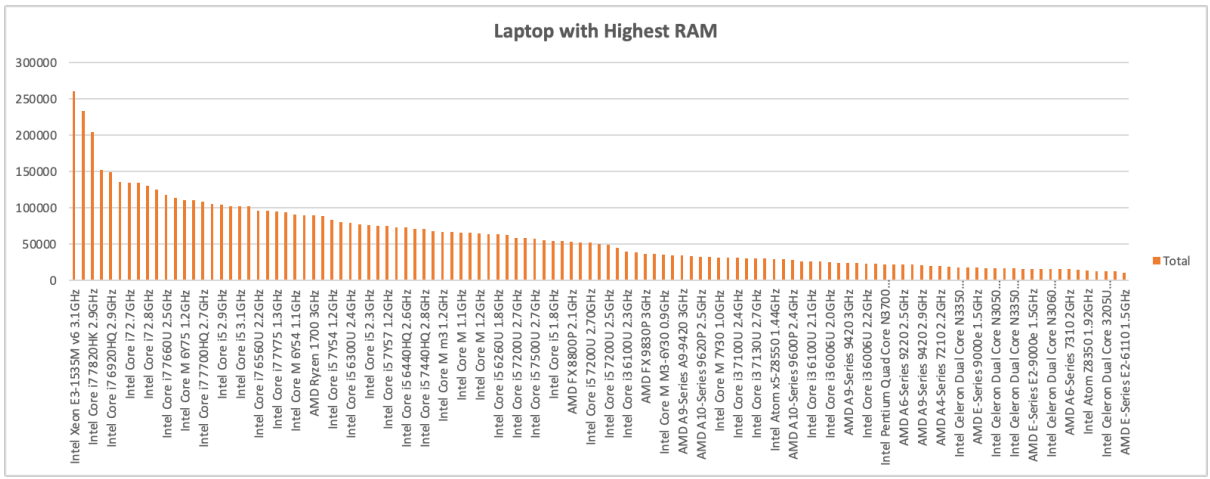

Figure 4: Highest amount of RAM

5. What is the range of laptop prices in the dataset (i.e., the difference between the highest and lowest prices)?

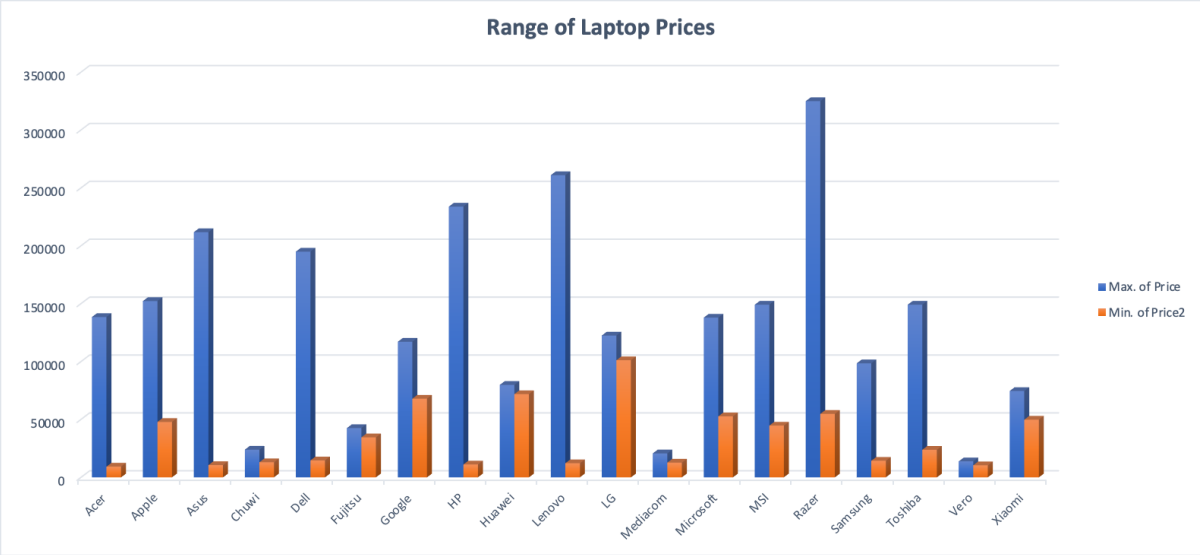

Figure 5: Range of laptop price

STATISTICAL ANALYSIS

The provided output is from a regression analysis, and it contains various statistics and coefficients that help you interpret the relationship between your predictor variable (X Variable RAM) and the target variable (Price).

SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0,68564772								
R Square	0,4701128								
Adjusted R S	0,4696959								
Standard Err	27186,085								
Observations	1273								
ANOVA									
	df	*SS*	*MS*	*F*	*Significance F*				
Regression	1	8,3341E+11	8,3341E+11	1127,62372	1,706E-177				
Residual	1271	9,3937E+11	739083220						
Total	1272	1,7728E+12							
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*	
Intercept	21026,7033	1387,27753	15,156811	7,941E-48	18305,0976	23748,309	18305,0976	23748,309	
X Variable 1	4600,08892	136,988505	33,5801089	1,706E-177	4331,34047	4868,83738	4331,34047	4868,83738	

Tabel 1: Linear regression in Excel

Here's how to interpret each part of the summary output:

Regression Statistics:

- **Multiple R**: This is the multiple correlation coefficient. It measures the strength and direction of the linear relationship between the predictor variable(s) and the target variable. In this case, it's approximately 0.686, indicating a moderate positive correlation.

- **R Square (R^2)**: This is the coefficient of determination. It represents the proportion of the variance in the target variable (Price) that can be explained by the predictor variable(s) (X Variable RAM). An R^2 of 0.47 means that 47% of the variance in Price can be explained by X Variable RAM.

- **Adjusted R Square**: This adjusts the R^2 value for the number of predictor variables in the model. It's similar to R^2 but accounts for model complexity.

- **Standard Error**: This is a measure of the variability or dispersion of the residuals (the differences between the predicted values and the actual values). A lower standard error indicates a better fit.

- **Observations**: The number of data points in your dataset used for the regression analysis.

ANOVA (Analysis of Variance):

- **df**: Degrees of freedom. It represents the number of values in the final calculation of a statistic that are free to vary.

- **SS**: Sum of squares. It measures the total variability in the dependent variable (Price).

- **MS**: Mean square. It's the sum of squares divided by its degrees of freedom.

- **F**: The F-statistic is used to test the overall significance of the regression model. A high F-statistic and a low p-value (below the significance level) indicate that the model is statistically significant.

- **Significance F**: This is the p-value associated with the F-statistic. In this case, it's very close to zero (1.7056E-177), indicating strong evidence against the null hypothesis.

Coefficients:

- **Intercept**: This is the y-intercept of the regression equation. It represents the estimated value of Price when X Variable RAM is zero. In this case, it's approximately 21,026.70.

- **X Variable RAM**: This is the coefficient for your predictor variable (X Variable RAM). It represents the change in the estimated value of Price for a one-unit change in X Variable RAM. In this case, for every one-unit increase in X Variable RAM, Price is estimated to increase by approximately 4,600.09.

Standard Error, t Stat, P-value, Lower 95%, Upper 95%:

- **Standard Error:** This is the standard error associated with each coefficient. It measures the precision of the estimate.

- **t Stat:** The t-statistic measures how many standard errors the coefficient is away from zero. A higher absolute t-statistic indicates that the coefficient is more statistically significant.

- **P-value:** This tests the null hypothesis that the coefficient is equal to zero. A low p-value (typically below the significance level, such as 0.05) suggests that the coefficient is statistically significant.

- **Lower 95% and Upper 95%:** These represent the 95% confidence interval for the coefficients. It provides a range in which we can be 95% confident that the true coefficient lies.

CONCLUSION

In summary, based on this regression analysis, you can conclude that there is a statistically significant positive relationship between X Variable RAM and the Price of the items. For every one-unit increase in X Variable RAM, Price is estimated to increase by approximately 4,600.09, and the model explains about 47% of the variance in Price.