DAI-101 Assignment-1 Report

By- Yash Jain

CSE

23125040

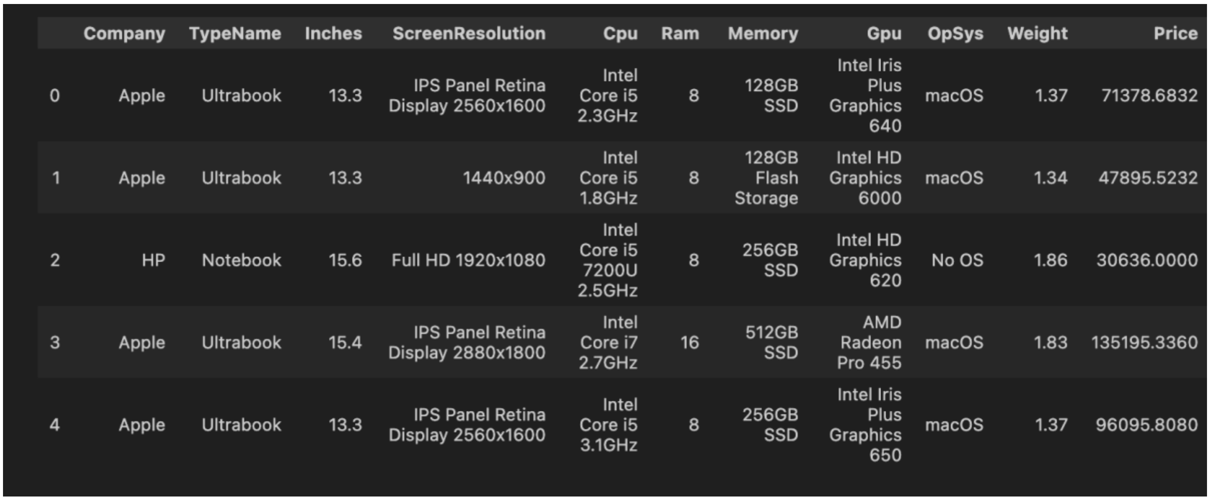
Dataset chosen: Laptop Specification and Pricing Dataset

About the dataset:

This dataset provides detailed information about various laptop models from multiple brands, including their specifications, hardware components, and pricing. It is useful for analysing trends in laptop configurations, comparing different brands, and understanding how various features influence the price. The dataset contains **1,303 records**, with each row representing a unique laptop model. It includes technical details such as screen size, resolution, processor type, RAM capacity, storage type, graphics card, and weight. Additionally, it specifies the operating system and the price of each laptop. The dataset can be used for exploratory data analysis (EDA), price prediction models, or customer preference studies.

Dataset Columns Description:

1. **Company** - The manufacturer or brand of the laptop (e.g., Apple, Dell, HP, Lenovo).
2. **TypeName** - The category/type of laptop, such as Notebook, Ultrabook, Gaming, or 2-in-1 Convertible.
3. **Inches** - The display size of the laptop, measured in inches.
4. **ScreenResolution** - The resolution of the screen and details about the display panel (e.g., Full HD, Retina, IPS panel).
5. **Cpu** - The processor model and speed, including details such as brand (Intel, AMD) and clock frequency.
6. **Ram** - The size of the system memory (RAM), typically in GB (e.g., 4GB, 8GB, 16GB).
7. **Memory** - The storage capacity and type (e.g., 1TB HDD, 256GB SSD, or hybrid configurations).
8. **Gpu** - The graphics processing unit (GPU), which may be integrated (Intel UHD Graphics) or dedicated (NVIDIA, AMD).
9. **OpSys** - The operating system installed on the laptop (e.g., Windows, macOS, Linux, No OS).
10. **Weight** - The weight of the laptop in kilograms, which helps understand portability.
11. **Price** - The price of the laptop in an unspecified currency (likely INR).
12. **Unnamed: 0** - An extra index column that may not be necessary for analysis.



	Company	TypeName	Inches	ScreenResolution	Cpu	Ram	Memory	Gpu	OpSys	Weight	Price
0	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 2.3GHz	8	128GB SSD	Intel Iris Plus Graphics 640	macOS	1.37	71378.6832
1	Apple	Ultrabook	13.3	1440x900	Intel Core i5 1.8GHz	8	128GB Flash Storage	Intel HD Graphics 6000	macOS	1.34	47895.5232
2	HP	Notebook	15.6	Full HD 1920x1080	Intel Core i5 7200U 2.5GHz	8	256GB SSD	Intel HD Graphics 620	No OS	1.86	30636.0000
3	Apple	Ultrabook	15.4	IPS Panel Retina Display 2880x1800	Intel Core i7 2.7GHz	16	512GB SSD	AMD Radeon Pro 455	macOS	1.83	135195.3360
4	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 3.1GHz	8	256GB SSD	Intel Iris Plus Graphics 650	macOS	1.37	96095.8080

This dataset is ideal for studying how different laptop features correlate with pricing and can be used for tasks such as data cleaning, Exploratory Data Analysis (EDA), feature engineering, and machine learning model development for price prediction.

DATA CLEANING:

1. **Inspecting the structure:**
   The dataset contained 1303 rows and 12 columns.
   It contained a column name Unnamed: 0 which was serial numbers in the dataset and was irrelevant for our analysis and I removed it.

2. **Handling Null values:**
   There were 30 rows out of 1303 rows which had null values in all the columns thus they should be removed as we should not generate the null values if all the columns are null in a row.
   So now there are 1273 rows remaining.

3. **Handling Duplicate values:**
   There were 29 duplicate rows and thus were removed and thus leave us with 1244 rows.

4. **Fixing the categorical values:**
   There were 2 rows in which the value of Inches and Weight was '?' and thus were removed leaving us with 1242 rows finally.
   The columns Inches, Weights and RAM which should though be numerical but were present as object datatypes in the dataset so I converted them as:
   Inches were converted into float,
   Weights were stripped with kg at the end and converted into float,
   RAM was stripped with GB at the end and converted into int datatype.

   Thus we have now the following categorical and numerical features:-

Categorical features:-

- Company
- TypeName
- ScreenResolution
- Cpu
- Memory
- Gpu
- OpSys

Numerical features:-

- Inches
- RAM
- Weight
- Price
-

5. **Detecting and treating outliers:**

In order to detect the outliers in the dataset, I calculated the Z-score for each numerical feature and set the threshold value as 3. Whichever row had absolute value of z-score greater than this threshold were removed. There were about 45 such rows and finally we have 1199 rows remaining on which we will be performing the EDA.

Exploratory Data Analysis (EDA):

1. **Univariate Analysis:**

For univariate analysis, we will separately analyse each column.

- Summary Statistics:

For numerical features, I calculated the statistical values like mean, median, mode and variance.

- Frequency distributions:

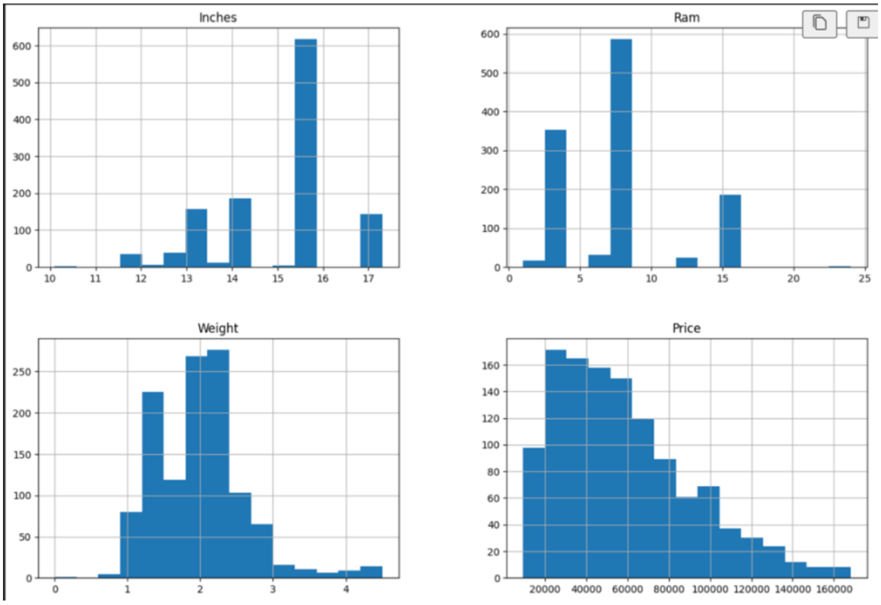
For categorical features, I made the frequency distributions for each of them on the basis of which I made the following observations:-

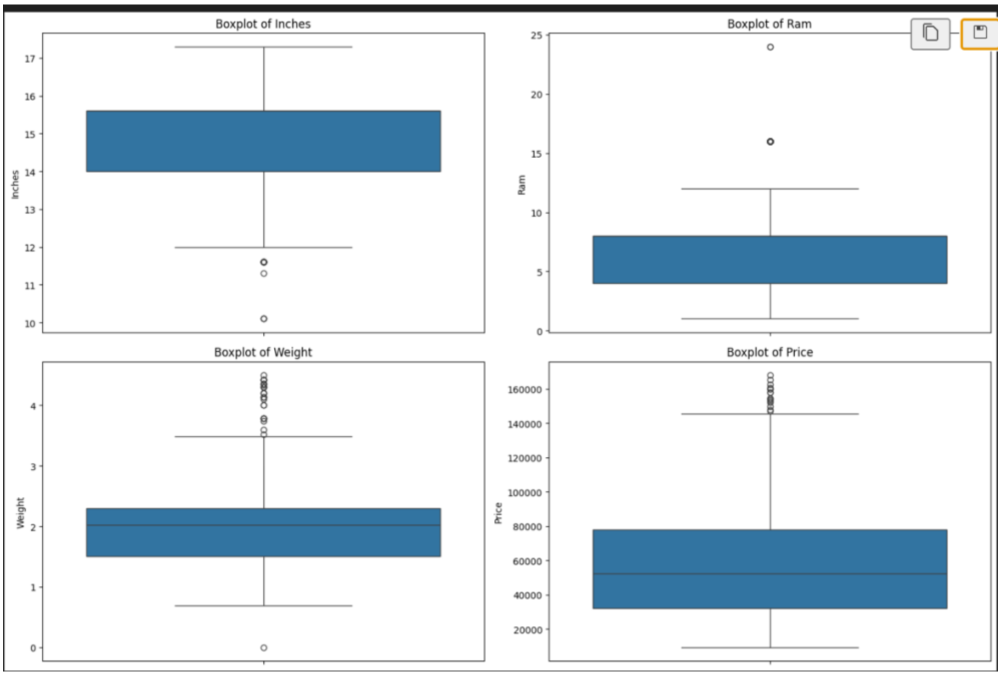
- The highest amount of laptops were sold by Dell (268), Lenovo (267) and HP (258), whereas other companies sold less than half than these values.
- The highest sold laptop type was Notebook (676) type which constituted about 50% of the dataset, followed by Gaming (184).
- Most laptops sold were of Full HD 1920x1080 Resolution.
- Around 264 laptops had Intel HD Graphics 620 card.
- Almost all the laptops (982) has Windows 10 as their Operating System.

- Histograms and Boxplots:

Further, I plotted histograms and boxplots for the numerical features for data visualization and observed the following points:-

- The size of laptops was mostly 15.6 inches.
- Almost half of the laptops had 8GB RAM.
- Most of the laptops ranged from 1.2-2.7 Kg.
- Price range varied from Rs. 20,000 to Rs. 1,60,000 but most were in the range of Rs. 30,000 to Rs 80,000.

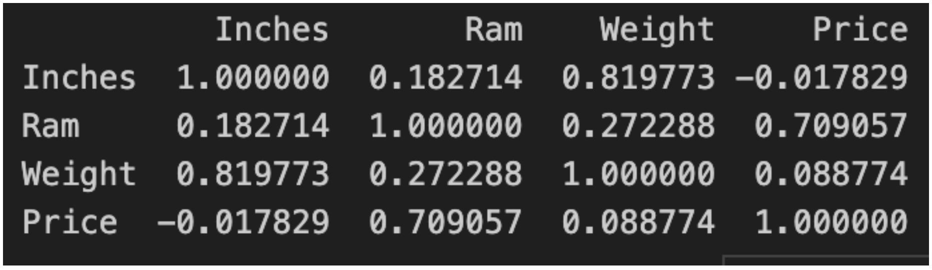




2. **Bivariate Analysis:**

- Correlation Matrix:

The correlation matrix revealed that there was a high correlation of 0.82 between Inches and Weight which is because bigger laptops have a larger weight. Further the correlated between RAM and Price is also high (0.71) which is because laptops having higher RAM are generally more expensive.



```
          Inches       Ram    Weight     Price
Inches  1.000000  0.182714  0.819773 -0.017829
Ram     0.182714  1.000000  0.272288  0.709057
Weight  0.819773  0.272288  1.000000  0.088774
Price  -0.017829  0.709057  0.088774  1.000000
```

- Scatterplots:

Pairwise scatterplots for numerical features vs Prices reveal the relationship seen in the correlation matrix graphically. The Inches and Ram features have non-continuous values and thus their scatterplots are either vertical or horizontal.

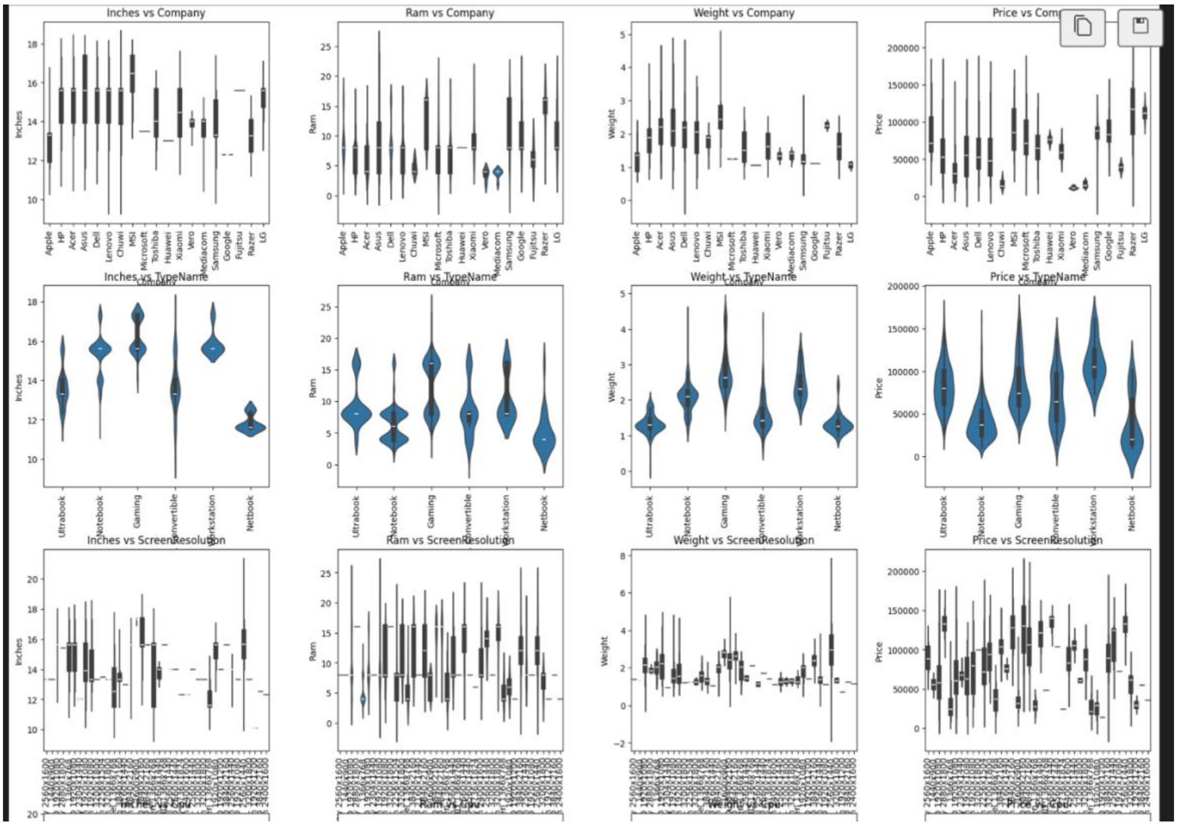
The Weight and Price columns have continuous values and thus provide a better representation in scatterplots.

- ## Analyzing Categorical columns with Numerical ones:

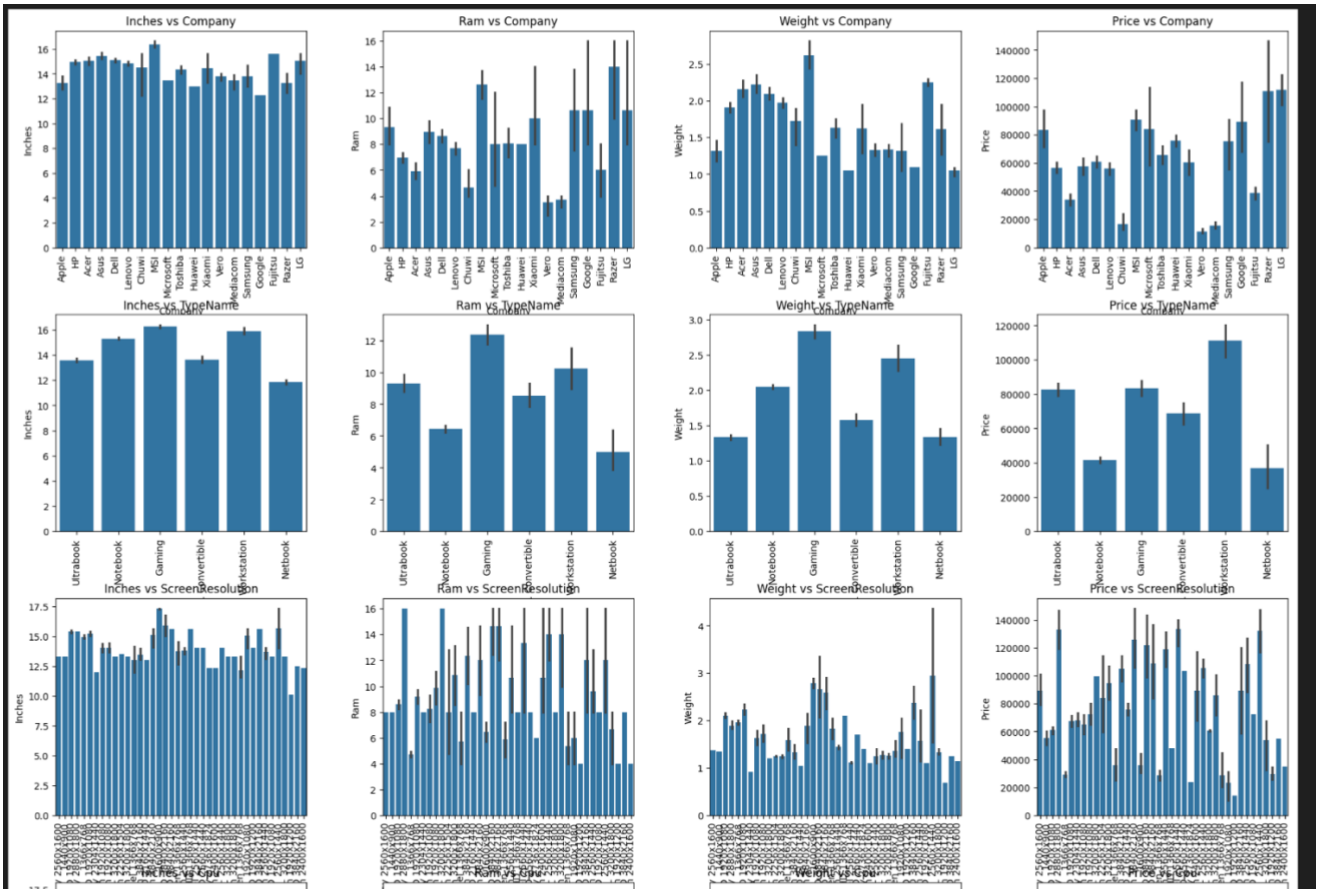
In order to study the relationship between the categorical and numerical columns, we have several methods like barplots, violinplots and boxplots.

The columns like CPU, GPU, and Memory have a lot of categories due to which studying their graphs is harder than for the other columns as seen in the graphs.

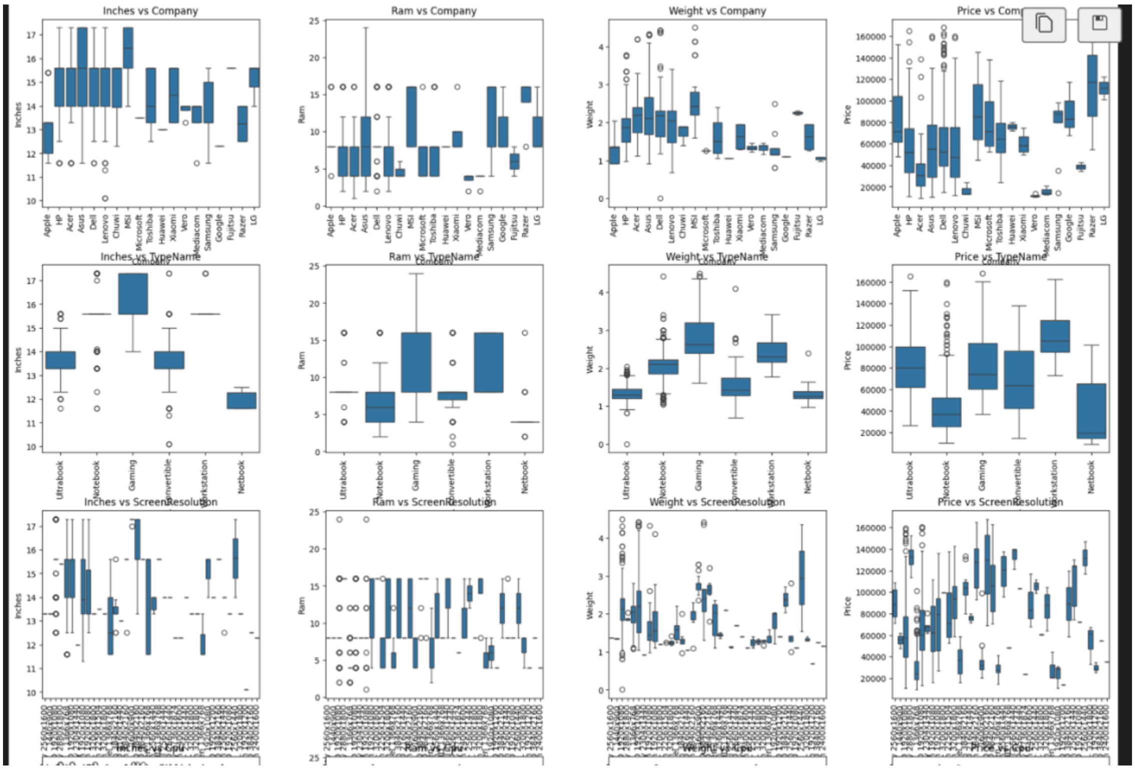
Violinplots:-



Barplots:-



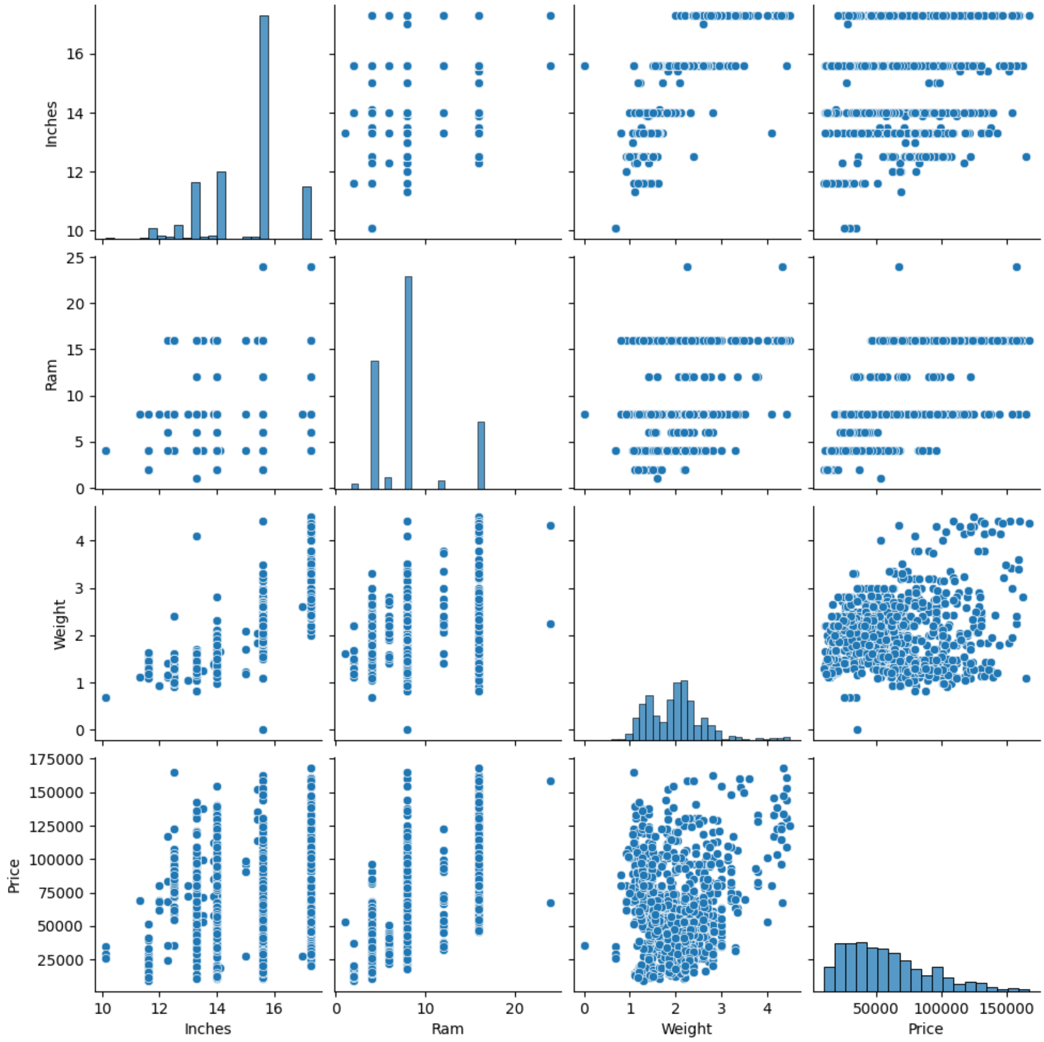
Boxplots:-



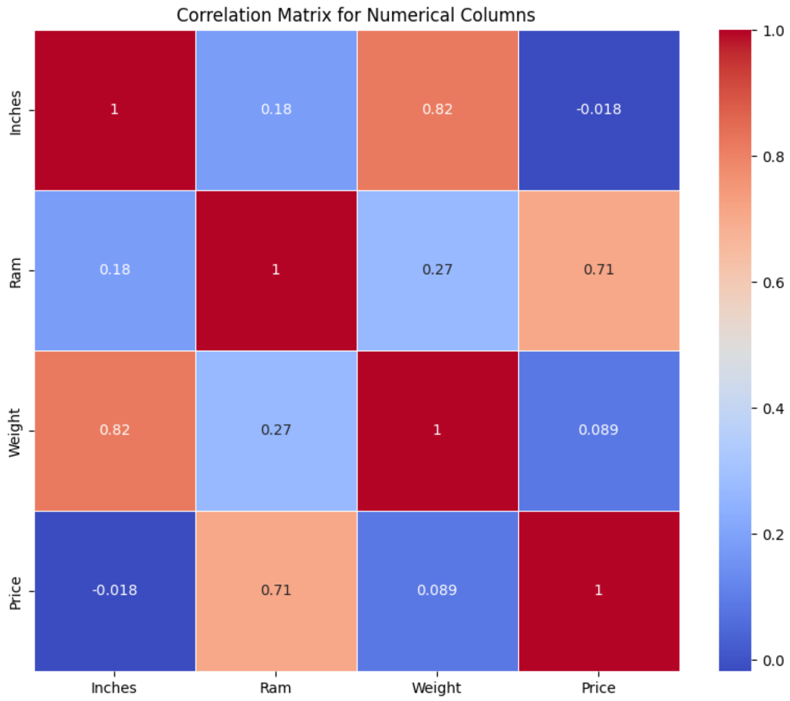
3. Multivariate Analysis:

- Pairplots:

Pairplots are used to plot pairwise scatterplots for the numerical columns. Analyzing these gives us information about the variation of one numerical feature with other and as we saw in correlation matrix as well that RAM and Price increase together and Inches and Weights increase together as well.

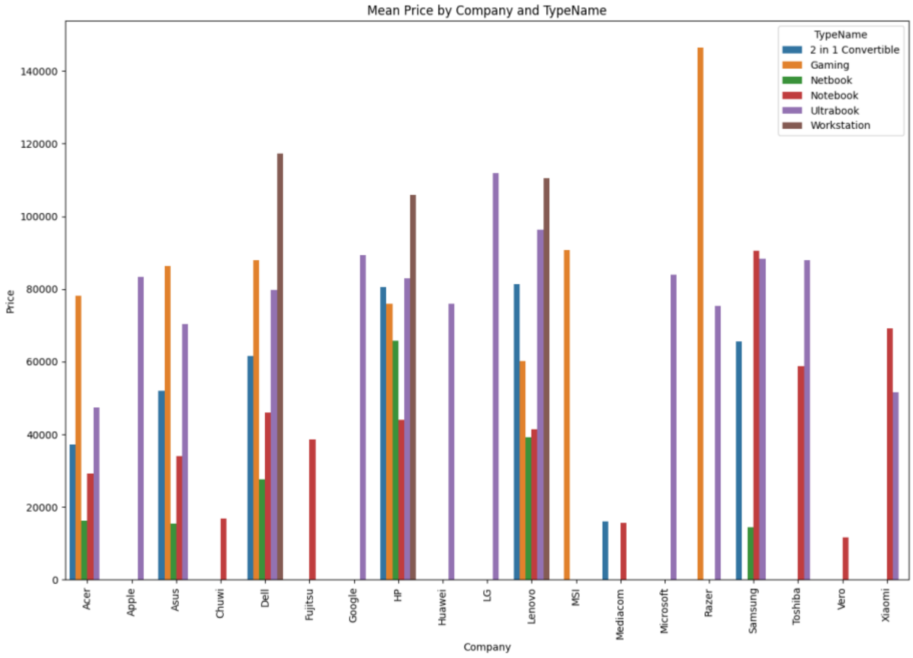


- Heatmap: Heatmap is used to vizualize the correlation matrix. Darker colour means higher correlation and lighter means less correlation. It is already explained under the correlation matrix above.



- Grouping Company and TypeName columns to study relationship with Pricing:

I plotted the graph of the mean price of laptops based on a company and laptop type and it shows that companies like Razer and Dell have relatively costly laptops and companies like Chuwi, Mediacom and Vero sell cheaper laptops.



Conclusion:

Through this exploratory data analysis (EDA) on the **Laptop Specification and Pricing Dataset**, we have uncovered key insights into the factors influencing laptop prices and their relationships with various specifications. The dataset, consisting of 1,303 laptop records (later cleaned to 1,199 after preprocessing), provided valuable information about brands, hardware configurations, and pricing trends.

Our **univariate analysis** revealed that Dell, Lenovo, and HP dominate the market, with **Notebook-type laptops being the most common**. We also found that most laptops come with **Full HD resolution, 8GB RAM, and weigh between 1.2-2.7 kg**.

In **bivariate analysis**, we observed a **strong correlation between RAM and Price (0.71)**, indicating that laptops with higher RAM tend to be more expensive. Similarly, **larger screens are associated with heavier laptops**. Scatterplots and barplots further confirmed these trends.

Lastly, in **multivariate analysis**, pairplots and heatmaps effectively visualized relationships between numerical features. We also examined how brand and laptop type affect pricing, with premium brands like **Razer and Dell offering high-end, expensive models**, while budget brands like **Chuwi and Mediacom target affordability**.

This study provides a foundation for **predictive modelling, customer segmentation, and price optimisation strategies**. Future work could include advanced machine learning models to predict laptop prices or a deeper study into consumer preferences.