Exploratory Analysis Of Laptop Prices And Their Determinants

SOHINI MANDAL

Project Overview

The objective of this project is to analyze the Laptop Pricing dataset to understand patterns, relationships, and factors influencing laptop prices. The dataset consists of 238 laptop observations with 12 features, including hardware specifications, manufacturer, operating system, and category. Exploratory Data Analysis (EDA) helps in identifying feature distributions, relationships, and potential biases that may influence predictive modeling.

Data In Hand

Dataset name: Laptop Pricing

Dataset link: https://www.kaggle.com/datasets/huzdaria/laptop-pricing

Dataset description: This dataset is a collection of 12 features related to various laptops, such as brand, processor type, RAM, storage capacity, and other specifications. It contains 238 observations initially. The dataset also includes the corresponding prices of these laptops. This dataset can be used for regression analysis to predict the prices of laptops based on their features. The dataset is suitable for data scientists, machine learning enthusiasts, and researchers who are interested in building regression models to predict the prices of laptops based on various features.

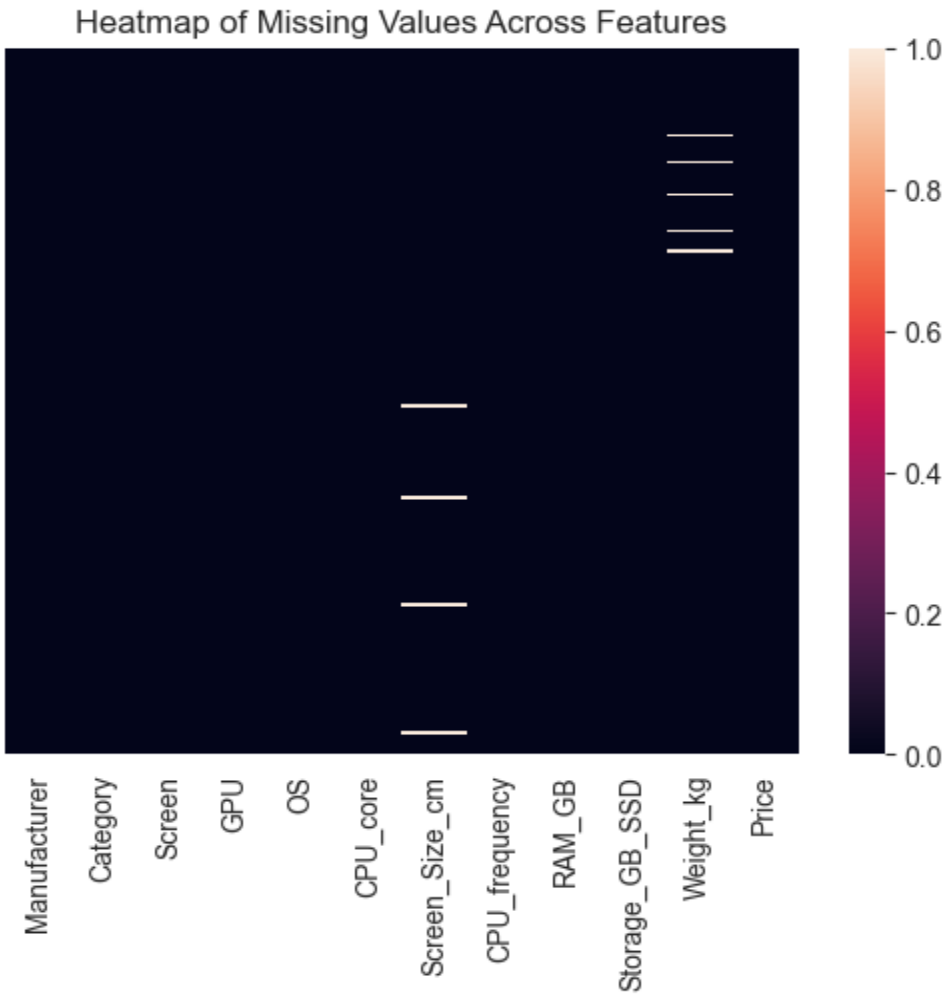
Feature	***Description***
Manufacturer	Company producing the laptop (Dell, HP, Lenovo, etc.)
Category	Laptop category (Gaming, Netbook, Notebook, Ultrabook, Workstation)
GPU	Graphics card manufacturer (AMD, Intel, NVidia)
OS	Operating system (Windows, Linux)
CPU_core	Processor type (Intel i3, i5, i7)
Screen_Size_cm	Laptop screen size in cm
CPU_frequency	CPU operating frequency in GHz
RAM_GB	Installed RAM in GB
Storage_GB_SSD	SSD storage capacity in GB
Weight_kg	Laptop weight in kilograms
Price	Laptop price in USD
Screen	Screen panel type (IPS, Full HD)

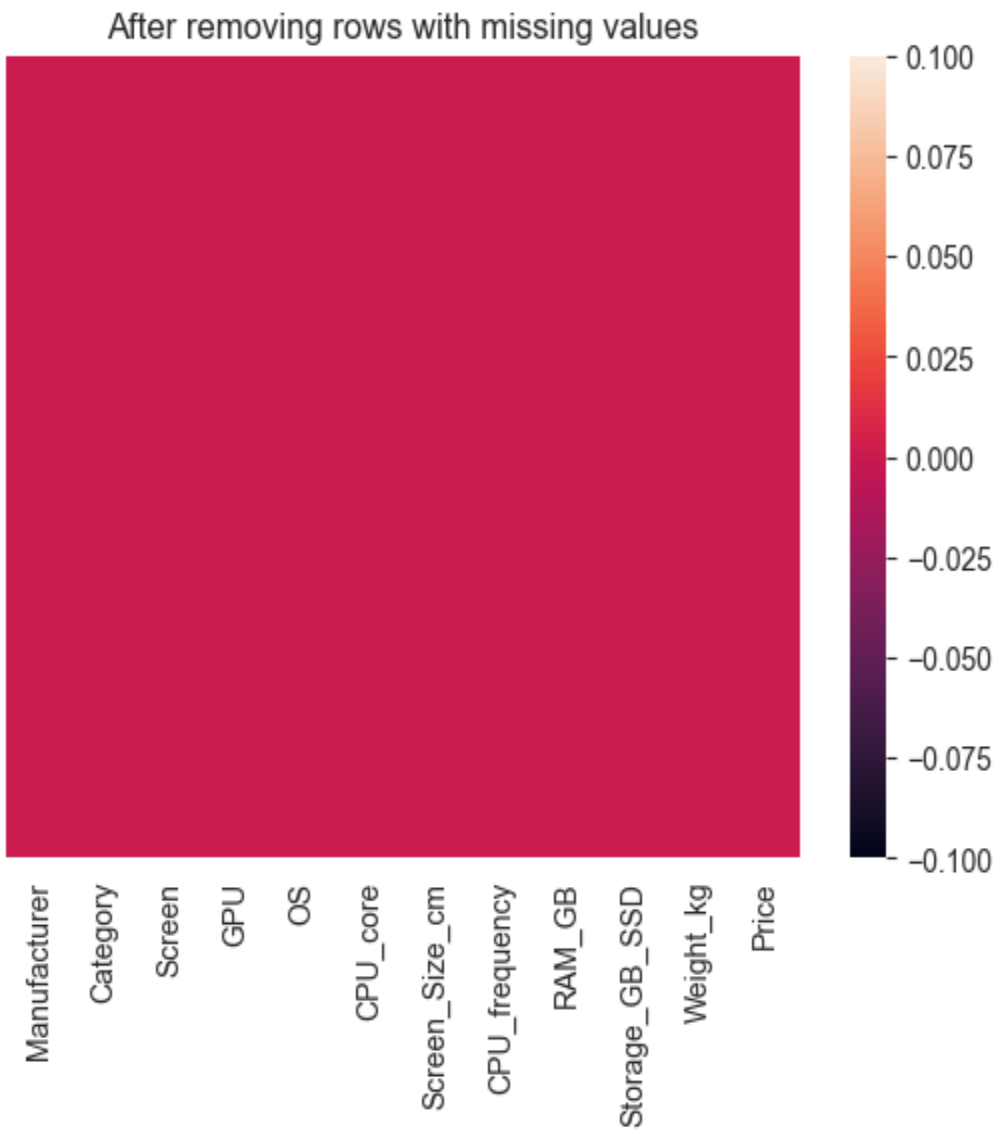
Data Cleaning and Preprocessing

Before performing exploratory data analysis and modeling, the dataset underwent a thorough cleaning and preprocessing phase to ensure consistency, completeness, and usability.

1. Handling Missing Values:

- Missing values were identified in the following columns:
    - Screen_Size_cm → 4 missing values
    - Weight_kg → 5 missing values
- These rows were removed, resulting in a final dataset of **229 complete records**.
- Removing missing rows was deemed appropriate due to the small proportion of missing data (~3.8% of the original dataset), ensuring minimal impact on overall analysis.
- To visualize the presence of missing values in the dataset, heatmaps were plotted before and after removing the incomplete rows.





2. Duplicate Records:

- The dataset was checked for duplicate entries.
- No duplicates were present, confirming that all records are unique.

3. Data Type Verification:

Prior to exploratory analysis, all variables were examined to verify that their data types accurately reflected their underlying measurement scales and analytical roles. The initial dataset contained a mix of object, integer, and floating-point data types.

Although several variables were stored as integers, not all integer-valued variables represent true numerical measurements. Consequently, variables were reclassified based on their conceptual meaning rather than their raw storage type.

Numerical Variables

The following features represent quantitative measurements where arithmetic operations and distance-based comparisons are meaningful:

- **Screen_Size_cm** – laptop screen size measured in centimeters
- **CPU_frequency** – CPU operating frequency measured in GHz
- **RAM_GB** – installed memory capacity in GB
- **Weight_kg** – laptop weight measured in kilograms
- **Price** – laptop price measured in USD

These variables were retained as numeric (float or int) to support distributional analysis, correlation analysis, and regression modeling.

Categorical Variables

The following features represent qualitative attributes or discretized specifications and were therefore treated as categorical variables:

- **Manufacturer** – laptop brand (Dell, HP, Lenovo, etc.)
- **Category** – laptop category (Gaming, Netbook, Notebook, Ultrabook, Workstation)
- **Screen** – screen panel type (IPS Panel, Full HD)
- **GPU** – GPU manufacturer (AMD, Intel, NVidia)
- **OS** – operating system (Windows, Linux)
- **CPU_core** – processor tier (Intel i3, i5, i7)
- **Storage_GB_SSD** – SSD storage capacity (128 GB, 256 GB)

Although *Category*, *GPU*, *OS*, *CPU_core*, and *Storage_GB_SSD* were originally stored as integer values, these integers represent labels or tiers, not continuous numerical measurements. Treating them as numerical would incorrectly imply equal spacing and arithmetic meaning.

No structural inconsistencies or invalid data types were detected, confirming the integrity of the dataset prior to encoding.

4. Categorical Encoding:

For exploratory data analysis and subsequent modeling, categorical variables were explicitly converted to pandas category data types. This conversion improves memory efficiency, ensures correct statistical interpretation, and prevents inappropriate numerical operations on non-numeric attributes.

After encoding, the dataset contained a clear separation between categorical and numerical variables, as reflected in the reduced memory footprint and updated data types.

Feature	*Description*	*Encoded Type*

Manufacturer	Laptop manufacturer	Nominal categorical
Category	Laptop category	Nominal categorical
Screen	Screen panel type	Nominal categorical
GPU	GPU manufacturer	Nominal categorical
OS	Operating system	Nominal categorical
Storage_GB_SSD	SSD storage capacity (128 GB, 256 GB)	Nominal categorical
CPU_core	Processor tier (Intel i3 < i5 < i7)	Ordinal categorical
Screen_Size_cm	Screen size in centimeters	Float
CPU_frequency	CPU operating frequency in GHz	Float
RAM_GB	Installed RAM in GB	Integer
Weight_kg	Laptop weight in kilograms	Float
Price	Laptop price in USD	Integer

Special Considerations:

- *CPU_core* was treated as an ordinal categorical variable because it represents ordered processor tiers (i3 < i5 < i7), not actual physical core counts. Preserving this order avoids misleading numeric assumptions while retaining hierarchical information.
- *Storage_GB_SSD*, although numeric in nature, was treated as categorical because it contains only two discrete values. Treating it as continuous would add no analytical benefit and could imply unsupported intermediate values.

Overall, this encoding strategy ensures that each variable is analyzed in a manner consistent with its conceptual meaning, supporting valid exploratory insights and reliable downstream modeling.

Descriptive Statistics of Numerical Variables

Statistic	*Screen_Size_cm*	*CPU_frequency*	*RAM_GB*	*Weight_kg*	*Price*
Count	229	229	229	229	229
Mean	37.29	2.36	7.87	1.87	1458.42
Std Dev	2.97	0.41	2.46	0.50	574.23
Min	30.48	1.20	4.00	0.81	527
25%	35.56	2.00	8.00	1.48	1068
Median (50%)	38.10	2.50	8.00	1.88	1333
75%	39.62	2.70	8.00	2.20	1763
Max	43.94	2.90	16.00	3.60	3810

The descriptive summary provides an overview of the central tendency, dispersion, and range of the numerical features in the dataset.

Screen_Size_cm has a mean of 37.29 cm and a median of 38.10 cm, with values ranging from 30.48 cm to 43.94 cm. The narrow interquartile range indicates that most laptops cluster around standard screen sizes, reflecting industry standardization.

CPU_frequency shows moderate variation, with values between 1.2 GHz and 2.9 GHz. The mean (2.36 GHz) and median (2.5 GHz) suggest that most laptops operate at higher clock speeds, with fewer low-frequency processors.

RAM_GB is highly concentrated, with a median and interquartile range fixed at 8 GB, indicating limited variability in memory configurations. The presence of higher values up to 16 GB suggests a smaller number of high-performance models.

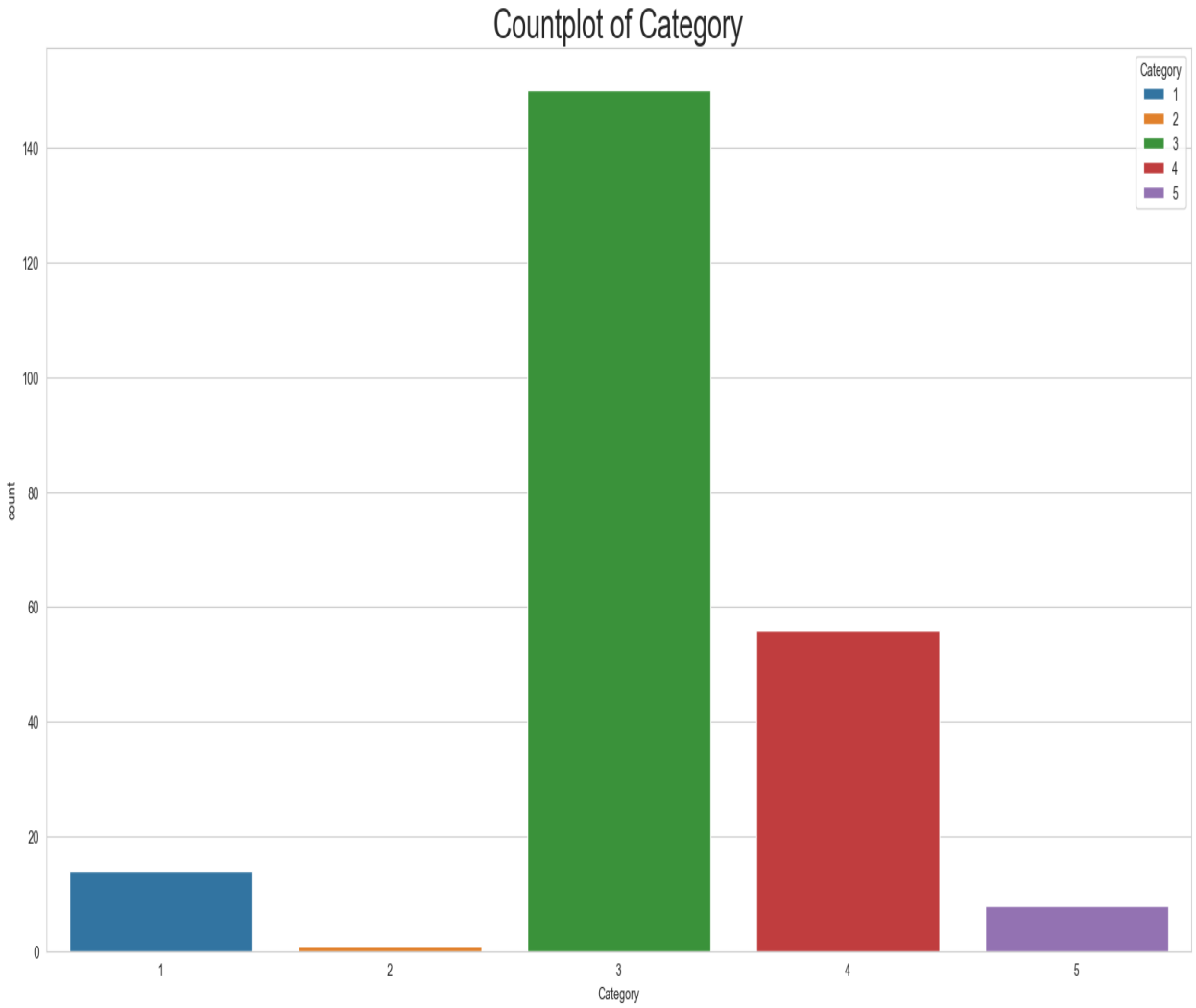
Weight_kg ranges from 0.81 kg to 3.6 kg, with a mean of 1.87 kg and a median of 1.88 kg. This indicates that most laptops fall within a moderate weight range, while heavier models form the upper tail.

Price exhibits substantial variability, ranging from USD 527 to USD 3810. The mean price (USD 1458.42) is higher than the median (USD 1333), indicating a positively skewed distribution driven by a relatively small number of high-priced laptops. The large standard deviation reflects wide price dispersion across different configurations.

Overall, the descriptive statistics highlight limited variability in hardware specifications such as RAM and CPU cores, contrasted with substantial variation in price, suggesting that pricing is influenced by combinations of features rather than any single specification alone.

Univariate Analysis

1



The dataset is heavily imbalanced across laptop categories.

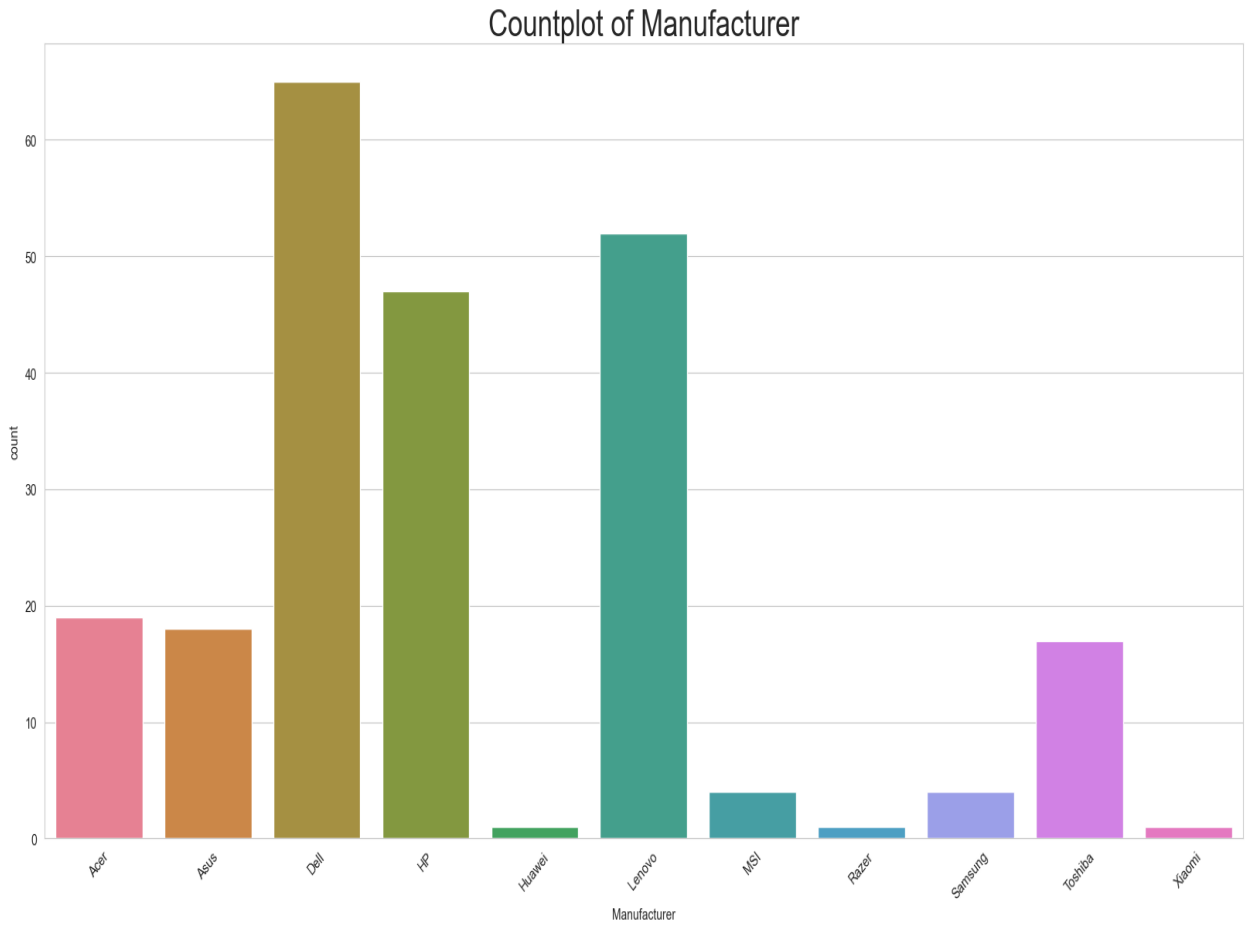
Notebooks (Category 3) dominate the data, followed by Ultrabooks (Category 4), while Gaming, Workstation, and especially Netbooks are sparsely represented.

- The data reflects market availability, not balanced design.

- Any model trained on this data will be biased toward Notebook-type laptops.

- Performance metrics averaged over all data will be misleading for minority categories.

2



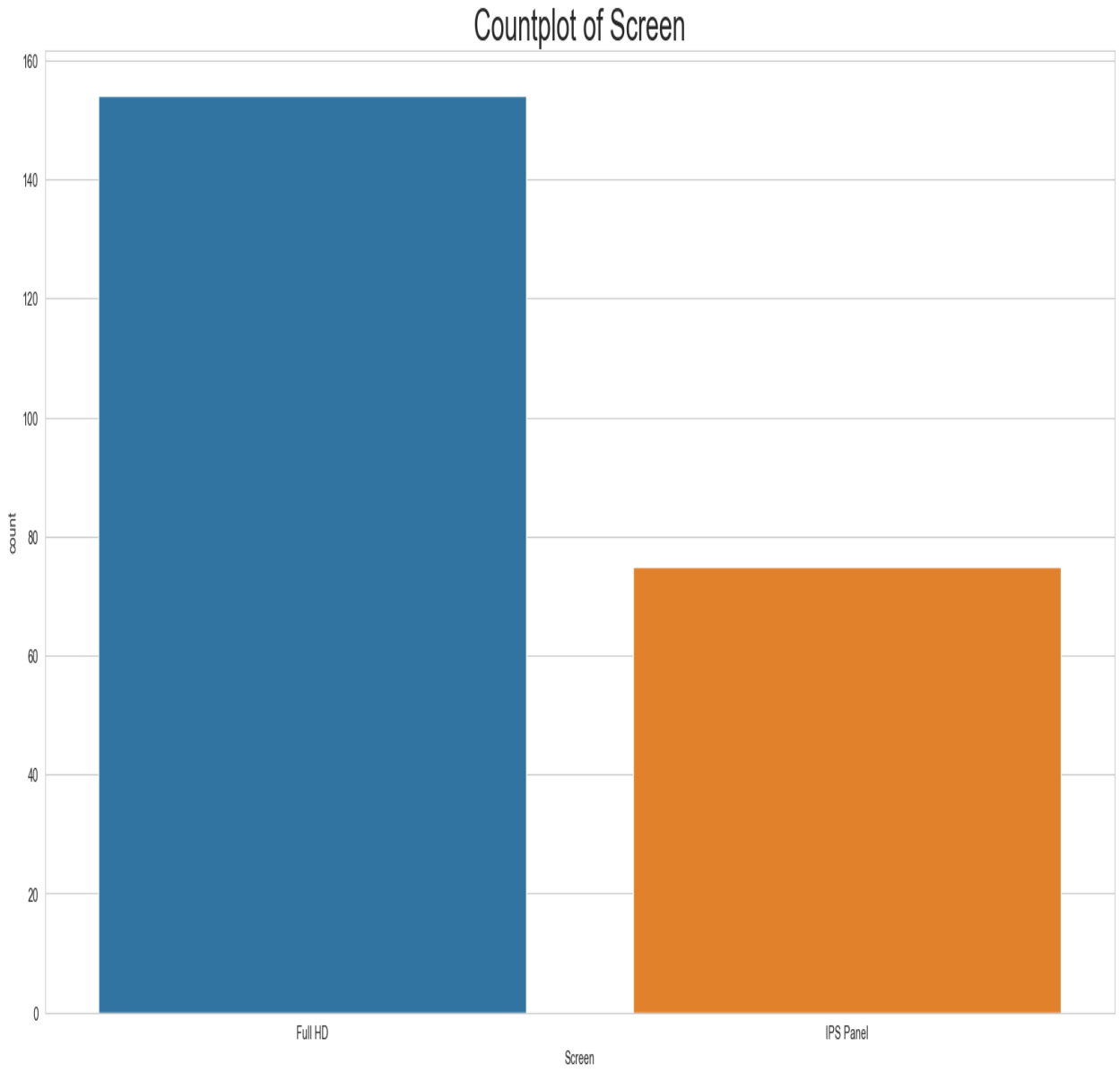
The dataset is dominated by a few major manufacturers, with Dell, Lenovo, and HP contributing the majority of observations. Other brands such as Huawei, Razer, Xiaomi, and MSI are sparsely represented.

- The data reflects market concentration, not equal brand representation.

- Any price or performance model trained on this data will be implicitly biased toward dominant brands.

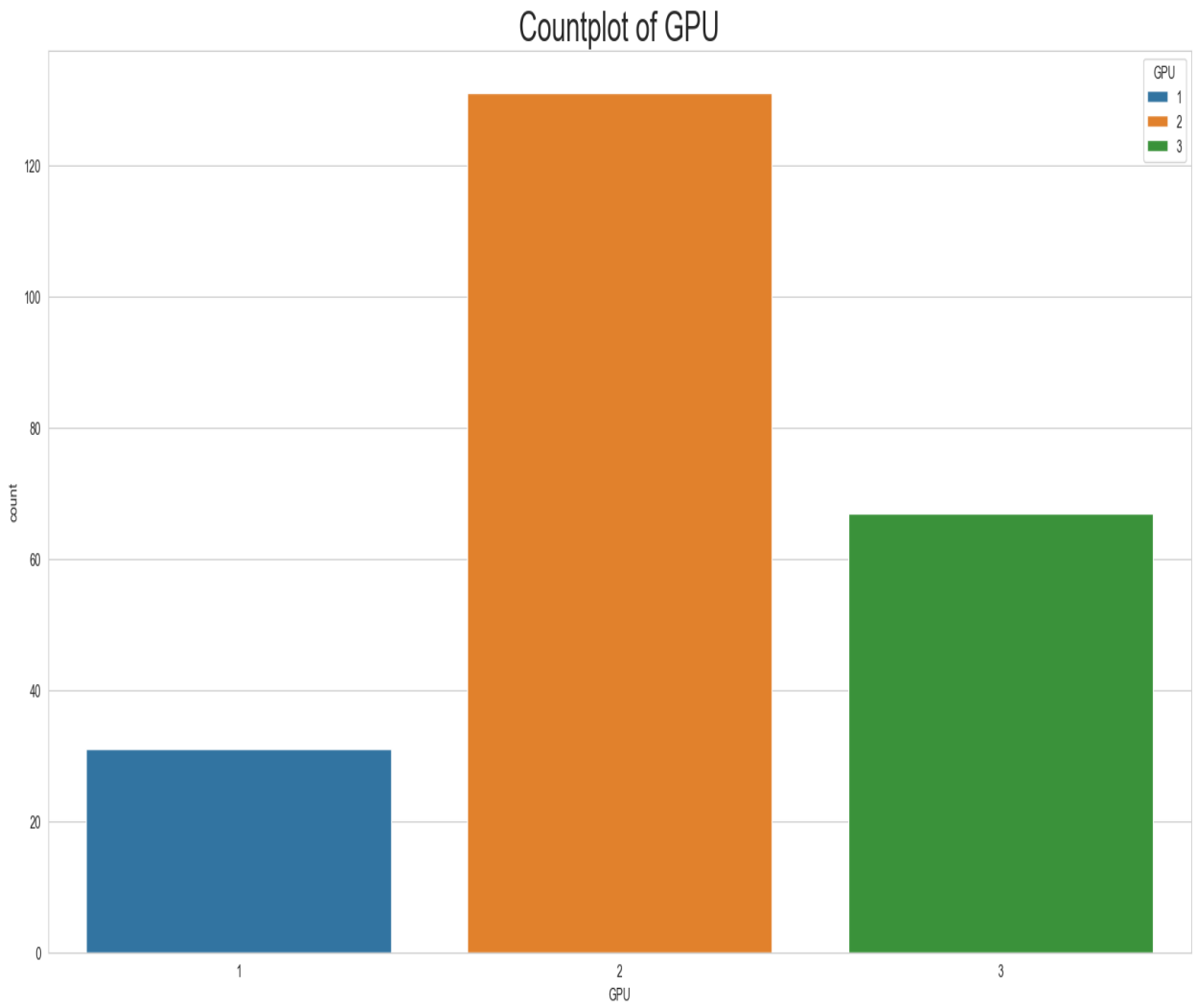
- Effects estimated for rare manufacturers will be unstable and unreliable

3



The count plot indicates that Full HD screens occur far more frequently than IPS panels in the dataset, demonstrating a clear imbalance in the distribution of screen types. This suggests that observations with IPS panels constitute a smaller portion of the data and may have limited influence in analyses involving the Screen variable.

4

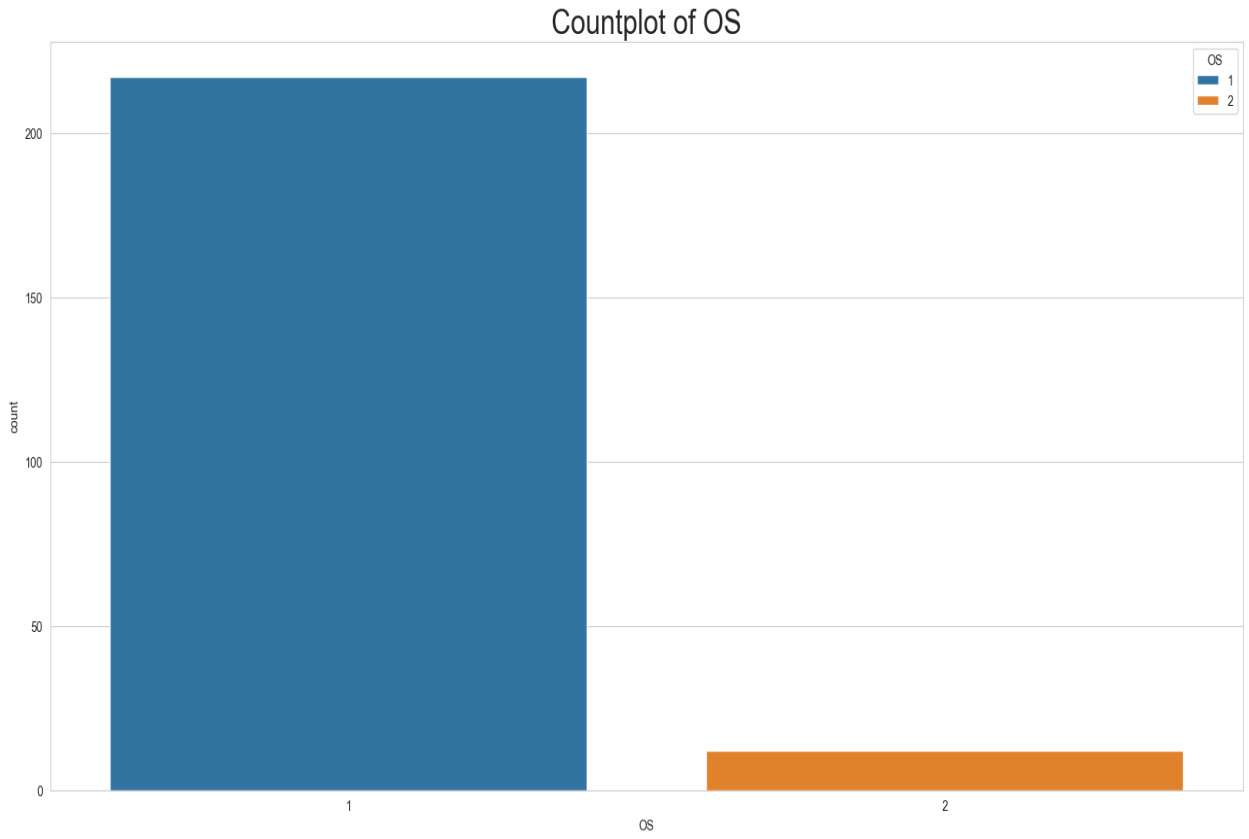


The count plot of the GPU variable shows that Intel GPUs (coded as 2) dominate the dataset, while NVIDIA GPUs (coded as 3) are moderately represented and AMD GPUs (coded as 1) appear least frequently. This indicates a non-uniform distribution of GPU manufacturers, with a clear concentration of observations in the Intel category.

- The dataset is heavily skewed toward Intel-based graphics, meaning that any analysis involving the GPU variable will be influenced primarily by Intel observations.

- Comparisons involving AMD GPUs may be statistically unstable due to their smaller sample size.

5

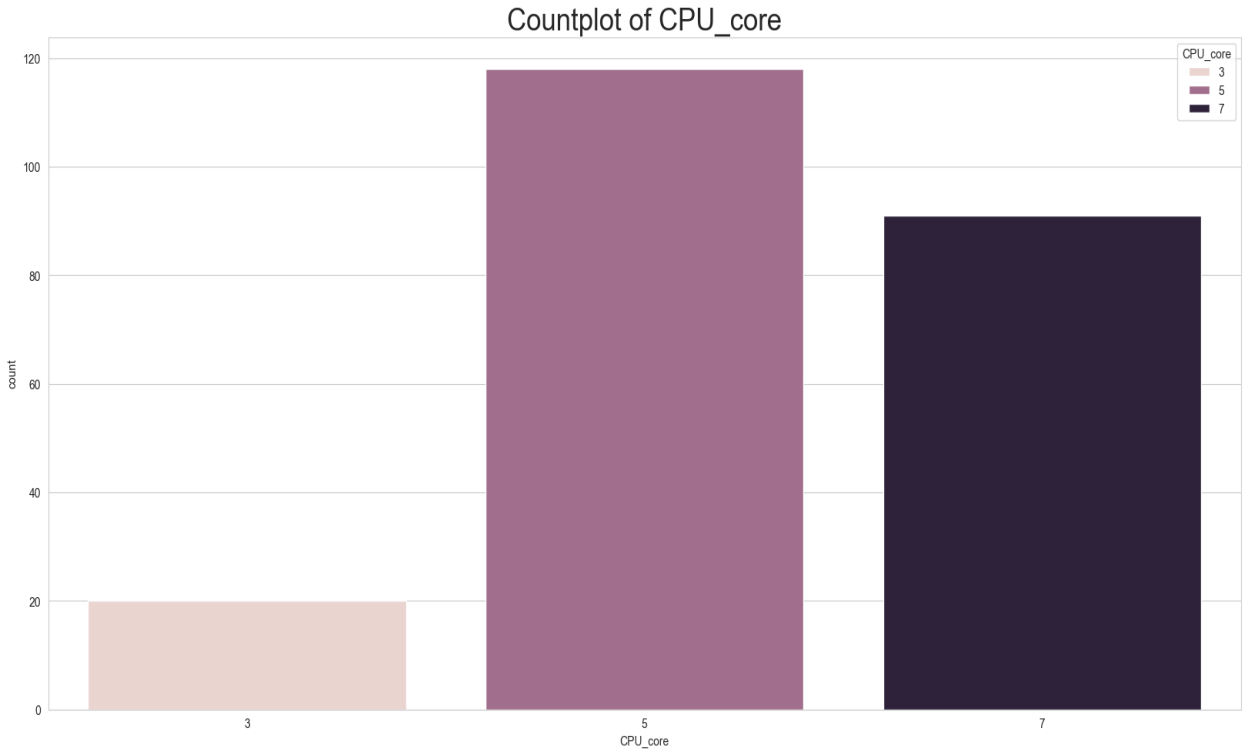


The count plot of the operating system variable shows an overwhelming dominance of Windows (coded as 1), with Linux (coded as 2) appearing in only a small fraction of observations. This indicates a severe class imbalance in the OS variable within the dataset.

- The dataset primarily reflects Windows-based laptops, and Linux systems are poorly represented.

- Any analysis involving the OS variable is likely to be biased toward Windows, potentially masking patterns or insights associated with Linux.

6



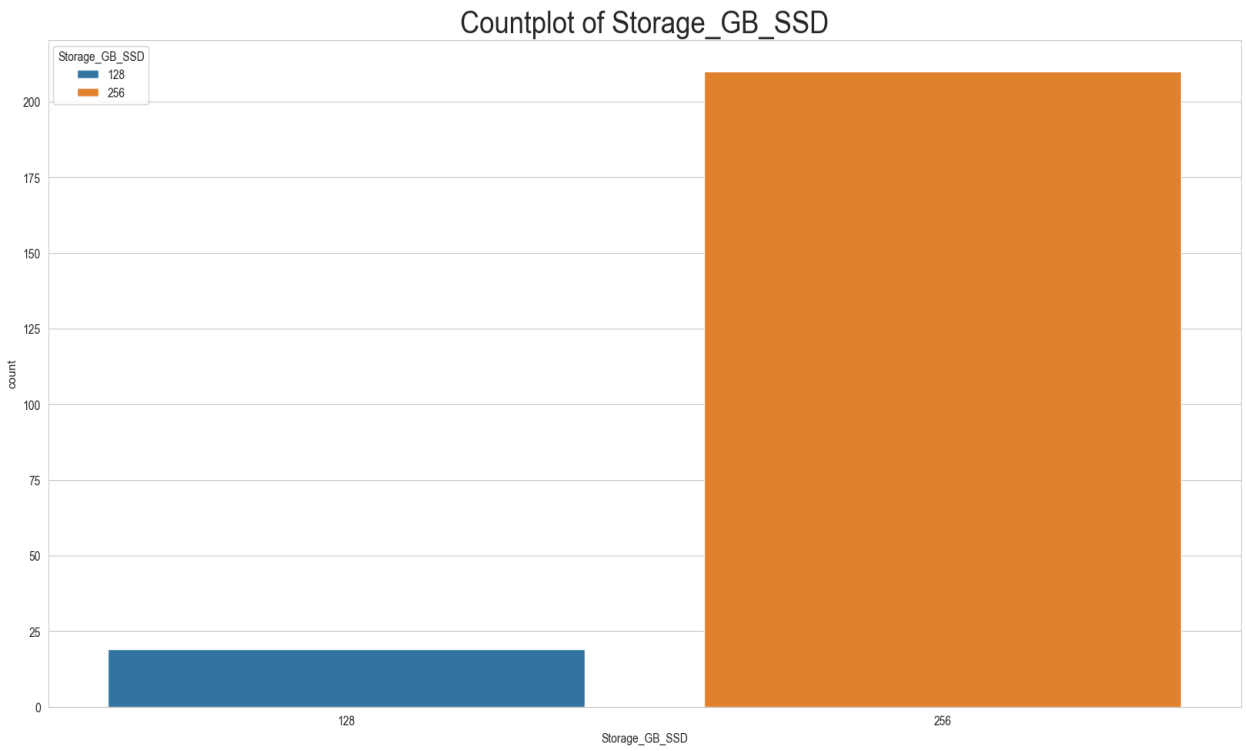
The countplot illustrates the distribution of processor tiers within the dataset, where CPU cores are encoded as Intel i3 (3), i5 (5), and i7 (7). The results indicate a clear imbalance in processor representation.

Intel i5 processors constitute the largest proportion of the dataset, making them the dominant processor tier. This suggests that the majority of systems included are equipped with mid-range processors, which are commonly used due to their balance between performance and cost.

Intel i7 processors form the second-largest group, indicating a substantial presence of high-performance systems, though they are less common than i5-based systems. In contrast, Intel i3 processors are minimally represented, showing that entry-level or low-performance systems are comparatively rare in the dataset.

Overall, the distribution implies that the dataset primarily reflects mid- to high-performance computing devices, with a strong emphasis on Intel i5 processors. This bias should be considered when analyzing performance-related outcomes or building predictive models, as results may be more representative of mid-range systems than low-end configurations.

7



The countplot shows the distribution of SSD storage capacity (Storage_GB_SSD) among the laptops in the dataset. Two storage configurations are present: 128 GB and 256 GB.

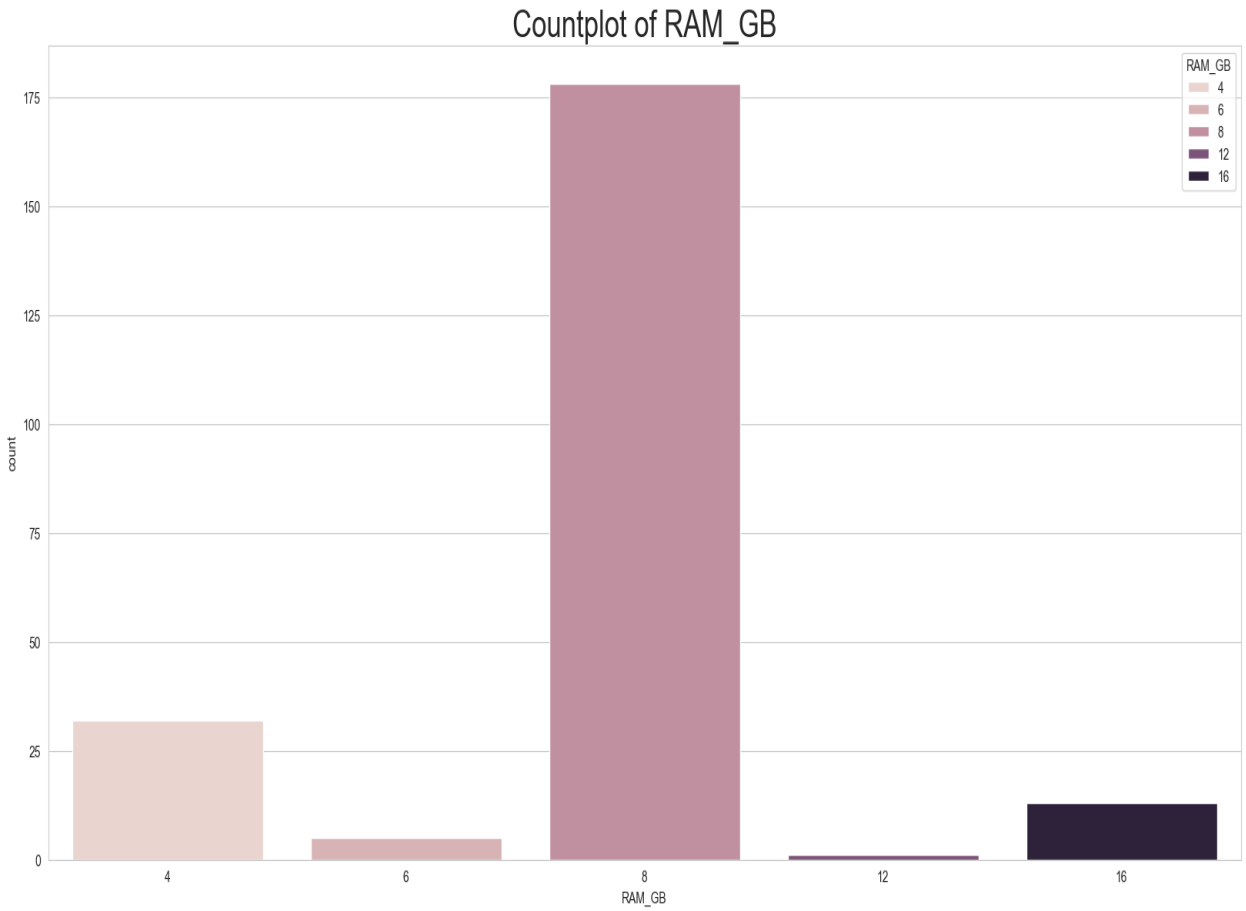
From the plot and summary statistics, it is evident that 256 GB SSDs dominate the dataset, with 210 laptops, while 128 GB SSDs are significantly underrepresented, appearing in only 19 laptops. This indicates a strong skew toward higher storage capacity systems.

The dominance of 256 GB SSD storage suggests that the dataset primarily represents modern, mid-range laptops, where larger SSD capacities have become standard. The limited number of 128 GB configurations implies that entry-level or budget laptops are less prevalent in the data.

For price prediction and regression modeling, this imbalance may influence the model's learning process, as SSD storage size is likely to be a strong predictor of laptop price. The model may therefore perform better when predicting prices for laptops with 256 GB SSDs, while predictions for 128 GB SSD configurations may be less reliable due to their limited representation.

Overall, the storage distribution highlights a dataset bias toward higher-capacity SSDs, which should be considered during feature analysis, model training, and evaluation.

8

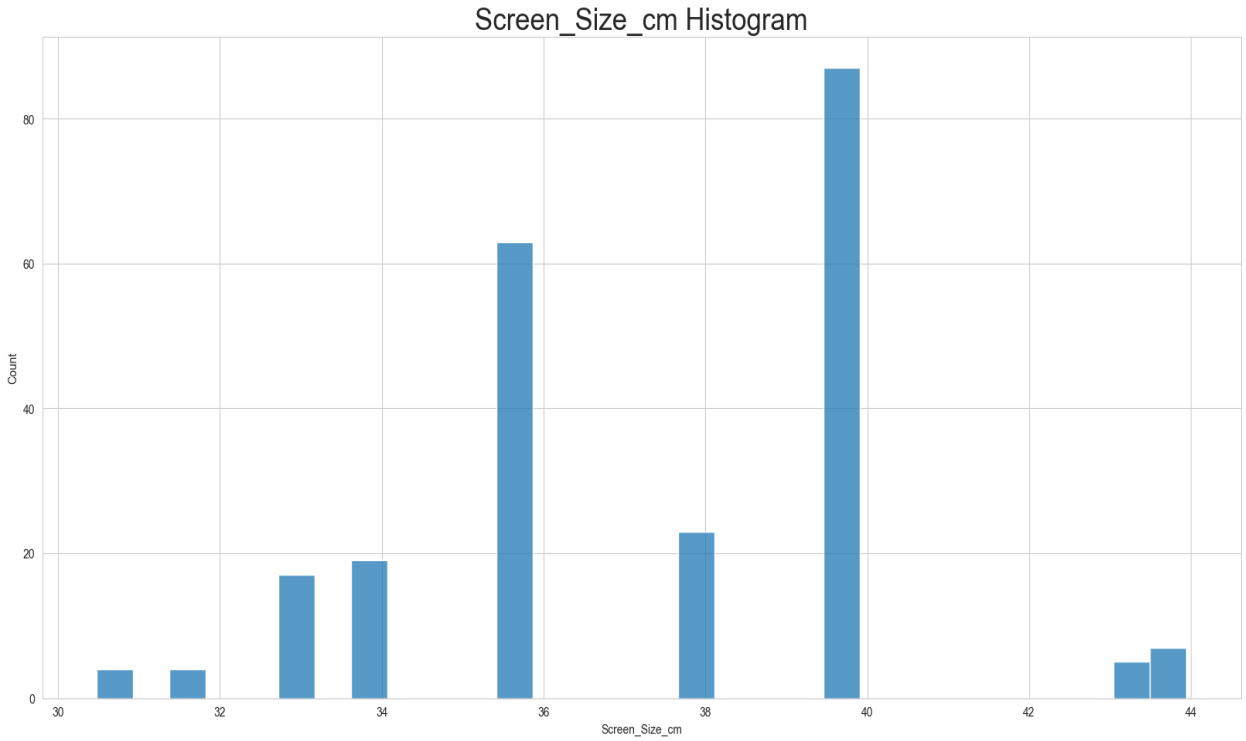


The RAM_GB variable represents the installed memory capacity of laptops in the dataset. The observed RAM values are discrete and include 4, 6, 8, 12, and 16 GB.

The distribution is highly concentrated at 8 GB RAM, which appears 178 times, making it the most frequent configuration in the dataset. In comparison, 4 GB RAM occurs 32 times, 6 GB RAM occurs 5 times, 16 GB RAM occurs 13 times, and 12 GB RAM occurs only once.

This indicates that the dataset is imbalanced, with a large proportion of observations clustered at a single RAM value (8 GB), while other RAM configurations are sparsely represented. The dataset therefore provides unequal coverage across different RAM capacities, which should be noted during analysis and modeling.

9



Although Screen_Size_cm is numeric, in this dataset it takes only a few distinct, repeated values

(e.g., 30.48, 31.75, 33.02, 33.782, 35.56, 38.10, 39.624, 43.18, 43.942).

These values represent fixed screen size standards, not a smoothly varying continuous measurement.

The objective here is to see how many laptops exist for each exact screen size, not to study a smooth distribution.

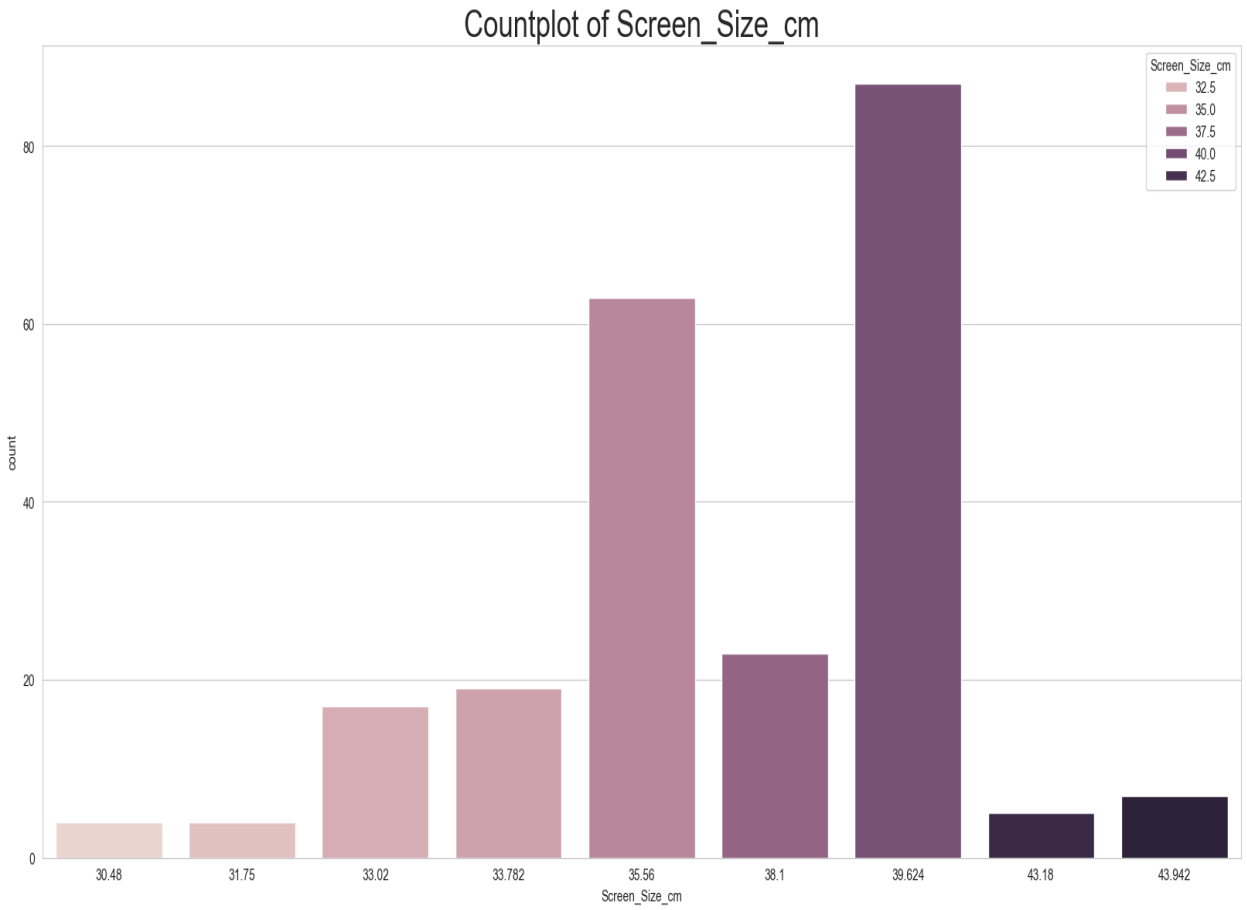
The histogram groups nearby screen sizes into bins, which:

- Hides the fact that sizes occur at exact discrete values.

- Can give a misleading impression of continuity.

Since Screen_Size_cm takes a limited number of discrete values with repeated observations, a countplot is more appropriate than a histogram for visualizing the frequency of each screen size in the dataset.

10



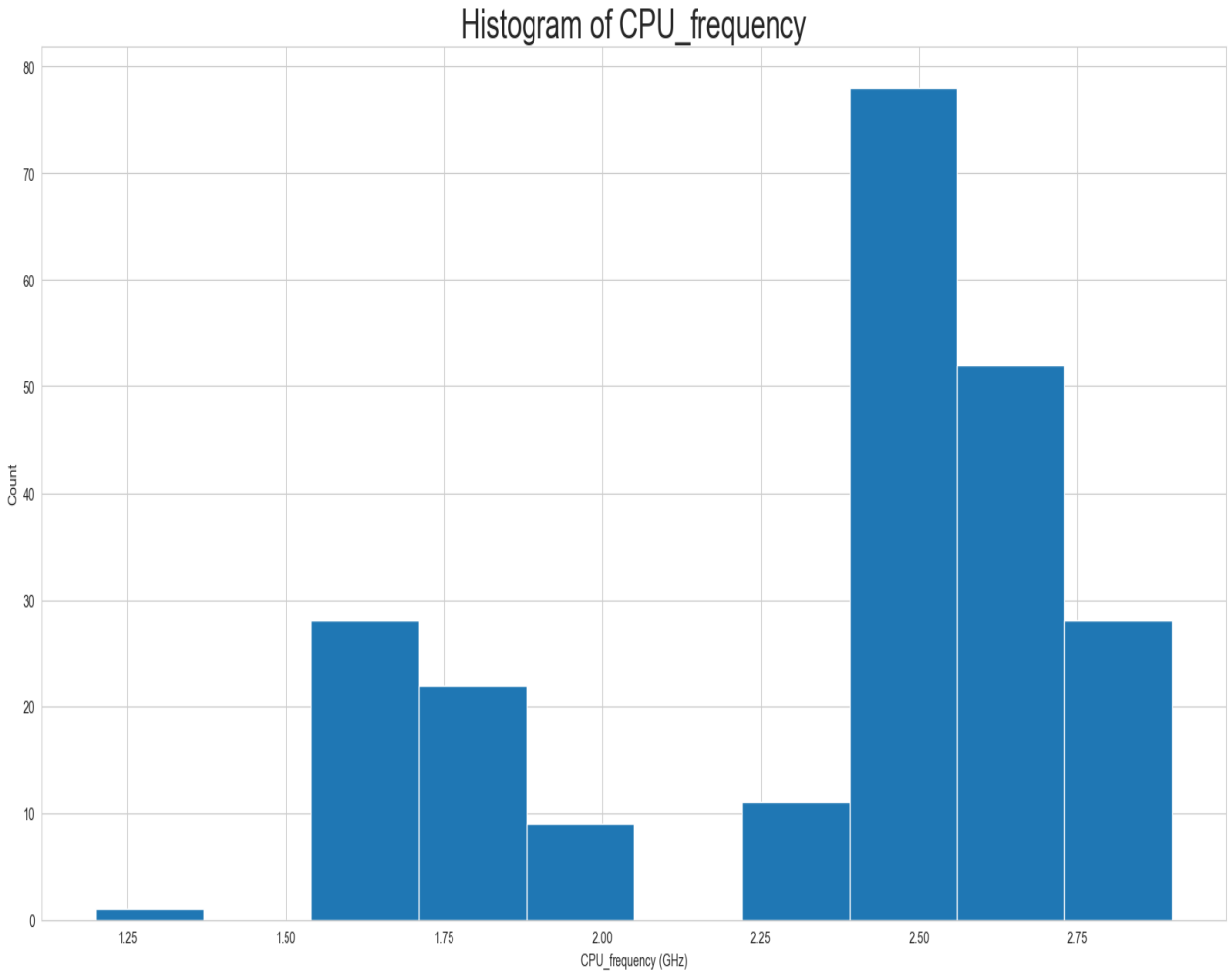
The Screen_Size_cm variable represents the laptop screen size measured in centimeters. The values are discrete and take a limited number of distinct measurements.

The distribution shows that 39.624 cm screens are the most frequent, appearing 87 times, followed by 35.560 cm screens with 63 observations. These two screen sizes together account for a large proportion of the dataset. Other screen sizes such as 38.100 cm (23), 33.782 cm (19), and 33.020 cm (17) occur with moderate frequency.

Larger screen sizes, 43.942 cm (7) and 43.180 cm (5), are relatively rare, while smaller screen sizes such as 31.750 cm (4) and 30.480 cm (4) have the lowest counts.

Overall, the dataset is unevenly distributed across screen sizes, with a strong concentration around a few specific screen measurements and limited representation at the smaller and larger extremes.

11

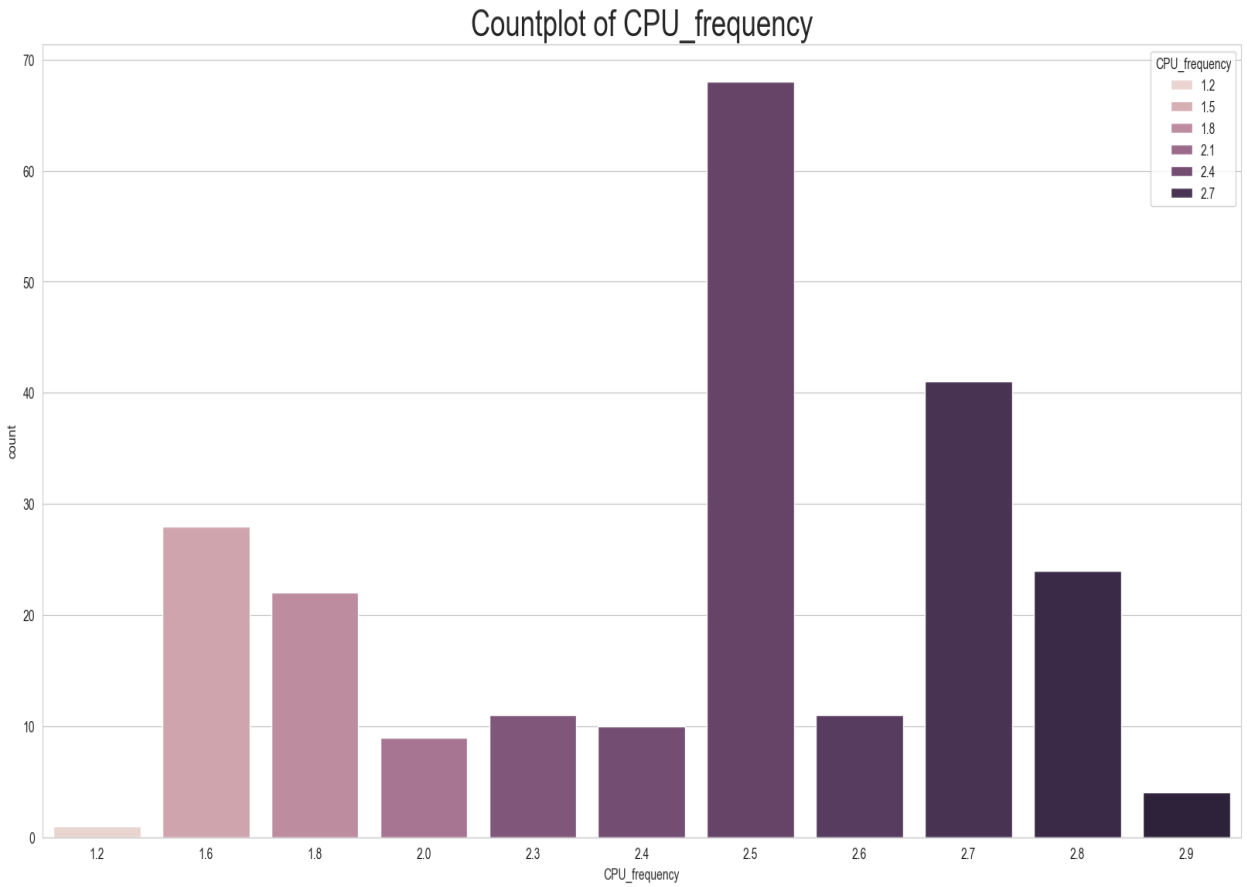


The histogram of CPU_frequency shows how the CPU operating frequencies are distributed across the dataset. The values range from approximately 1.2 GHz to 2.9 GHz.

The distribution is not uniform. A large concentration of observations lies between 2.4 GHz and 2.8 GHz, with the highest frequency around 2.5 GHz, indicating that this range contains the majority of laptops in the dataset. Lower CPU frequencies, particularly those below 2.0 GHz, occur much less frequently and form a sparse left tail. Very high frequencies close to 2.9 GHz also appear infrequently.

Overall, the histogram indicates a clustered distribution with most observations concentrated in the higher frequency range and relatively few observations at the lower and extreme upper ends.

12



Although CPU_frequency is a continuous numerical variable, in this dataset it is recorded at a limited number of discrete values (for example, 1.6, 1.8, 2.5, 2.7). Because of this:

A countplot is suitable for showing how many laptops correspond to each exact recorded frequency.

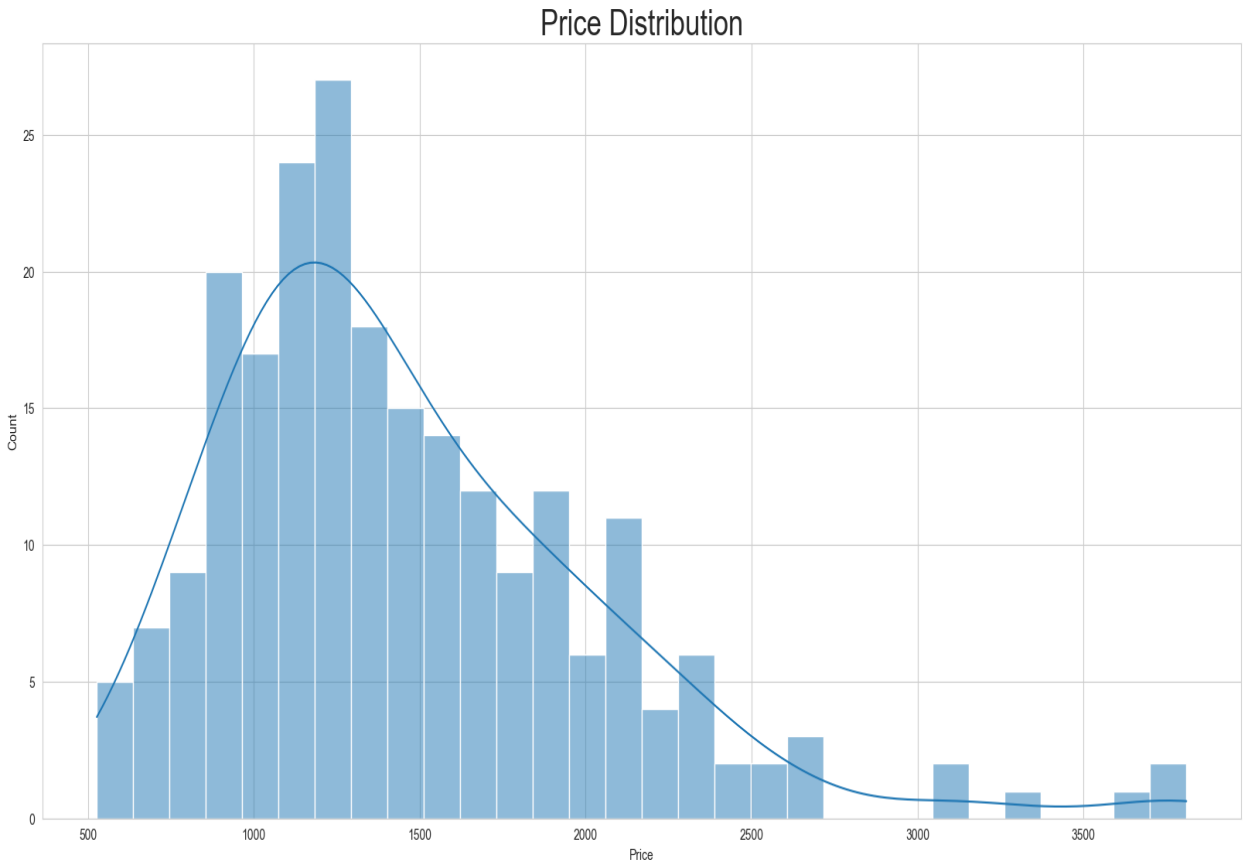
It helps identify which specific CPU frequencies are most common in the dataset.

The plot is useful for frequency comparison across observed values, not for analyzing overall distribution shape.

Therefore, while a histogram is more suitable for studying the distribution, a countplot is also valid for examining the frequency of individual CPU frequency values present in the data.

The most frequently occurring CPU frequency is 2.5 GHz, with 68 observations, followed by 2.7 GHz (41 observations) and 2.8 GHz (24 observations). Lower frequencies such as 1.6 GHz (28) and 1.8 GHz (22) also appear multiple times, while very low (1.2 GHz) and very high (2.9 GHz) frequencies occur only a few times.

13



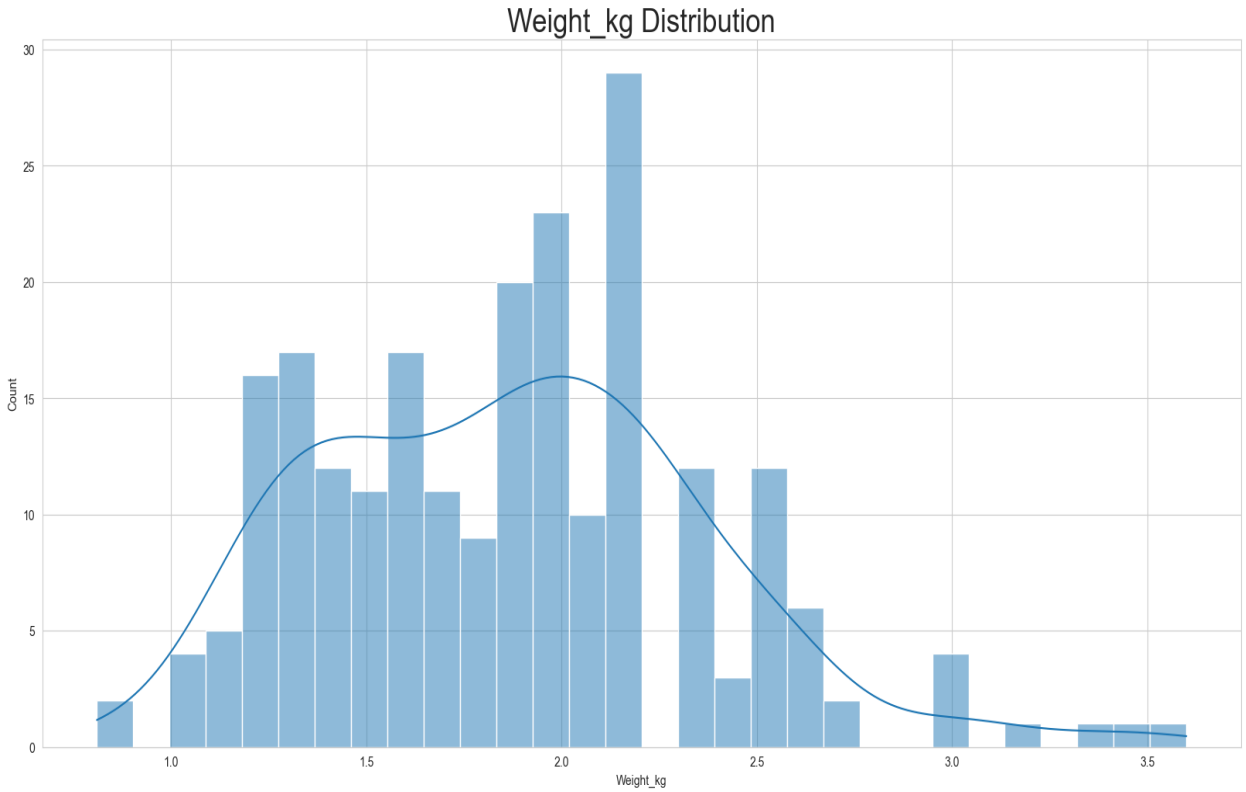
The histogram of Price shows the distribution of laptop prices across the dataset. Prices span a wide range, approximately from 500 USD to 3800 USD.

The distribution is positively skewed (right-skewed). A large number of observations are concentrated in the lower to mid price range, with the highest frequency occurring roughly between 900 USD and 1400 USD. As price increases beyond this range, the number of laptops decreases steadily.

A long right tail is observed, indicating the presence of a small number of high-priced laptops above 2500 USD, extending up to around 3800 USD. These high values occur infrequently compared to lower price points.

Overall, the price distribution is asymmetric, with most laptops priced toward the lower end of the scale and relatively few laptops at very high prices.

14



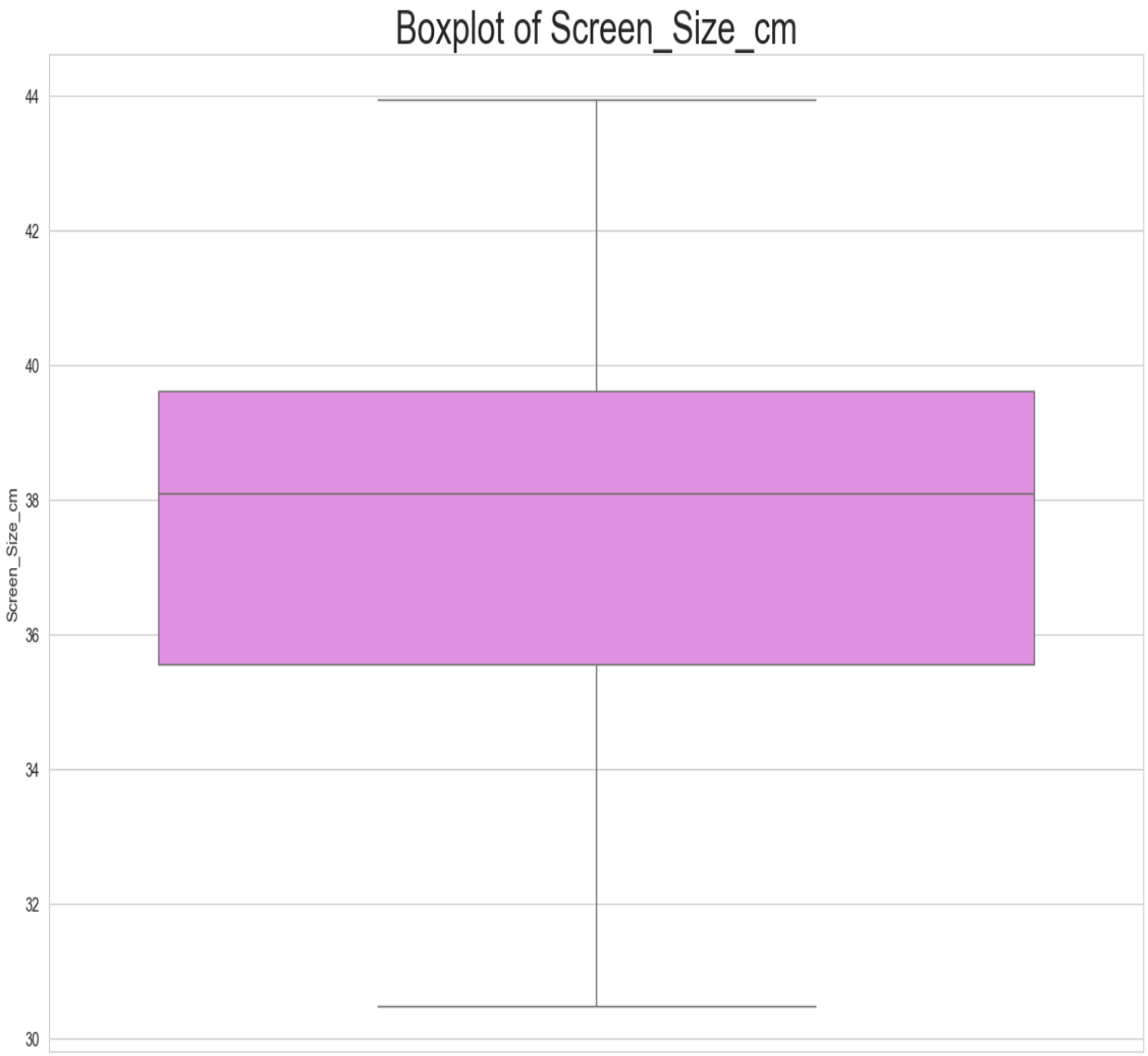
The histogram of Weight_kg shows the distribution of laptop weights in the dataset. The weights range approximately from 0.8 kg to 3.6 kg.

Most observations are concentrated between 1.2 kg and 2.3 kg, indicating that the majority of laptops fall within this weight range. The highest density appears around 1.8–2.1 kg, where the count of laptops is greatest.

The distribution is positively skewed, with a right tail extending toward heavier laptops above 2.5 kg. These heavier laptops occur less frequently, with only a small number of observations beyond 3.0 kg.

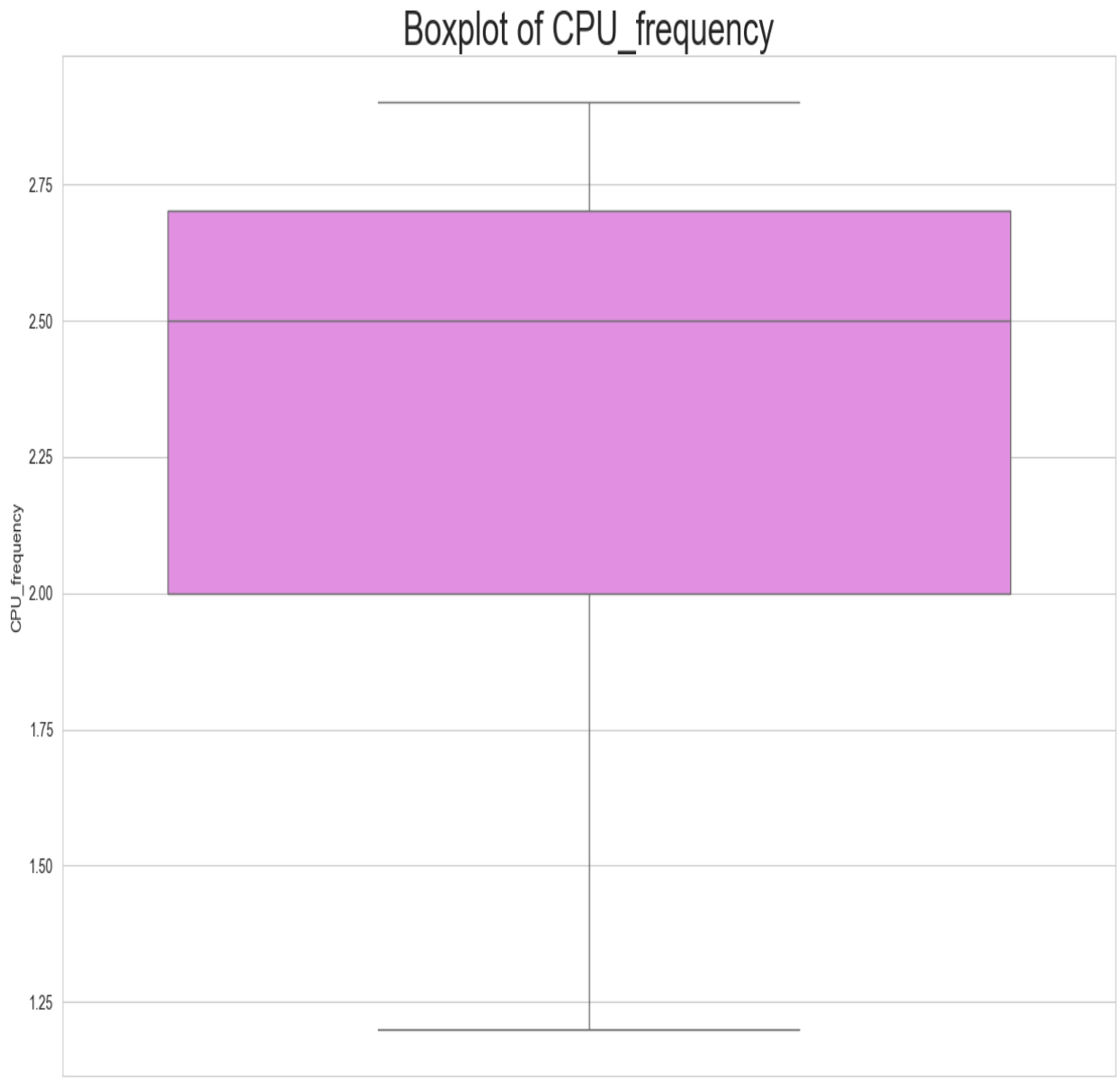
Overall, the weight distribution is asymmetric, with most laptops clustered in the lower-to-middle weight range and relatively few very heavy laptops.

15



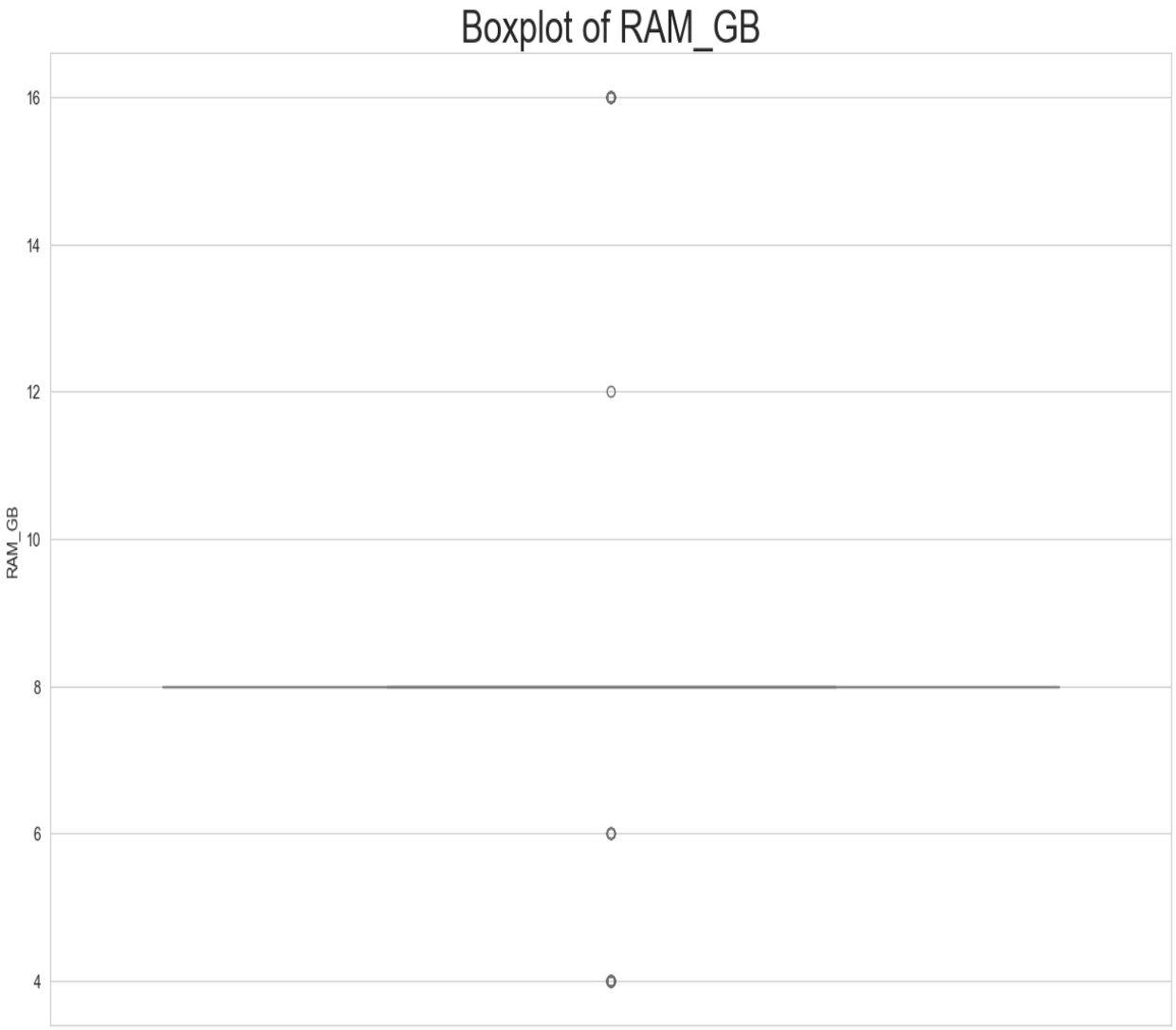
Screen size shows a moderate spread, with most laptops clustered around the median. The interquartile range is relatively small, indicating limited variation, and there are no significant outliers. This suggests screen sizes are fairly standardized across laptops.

16



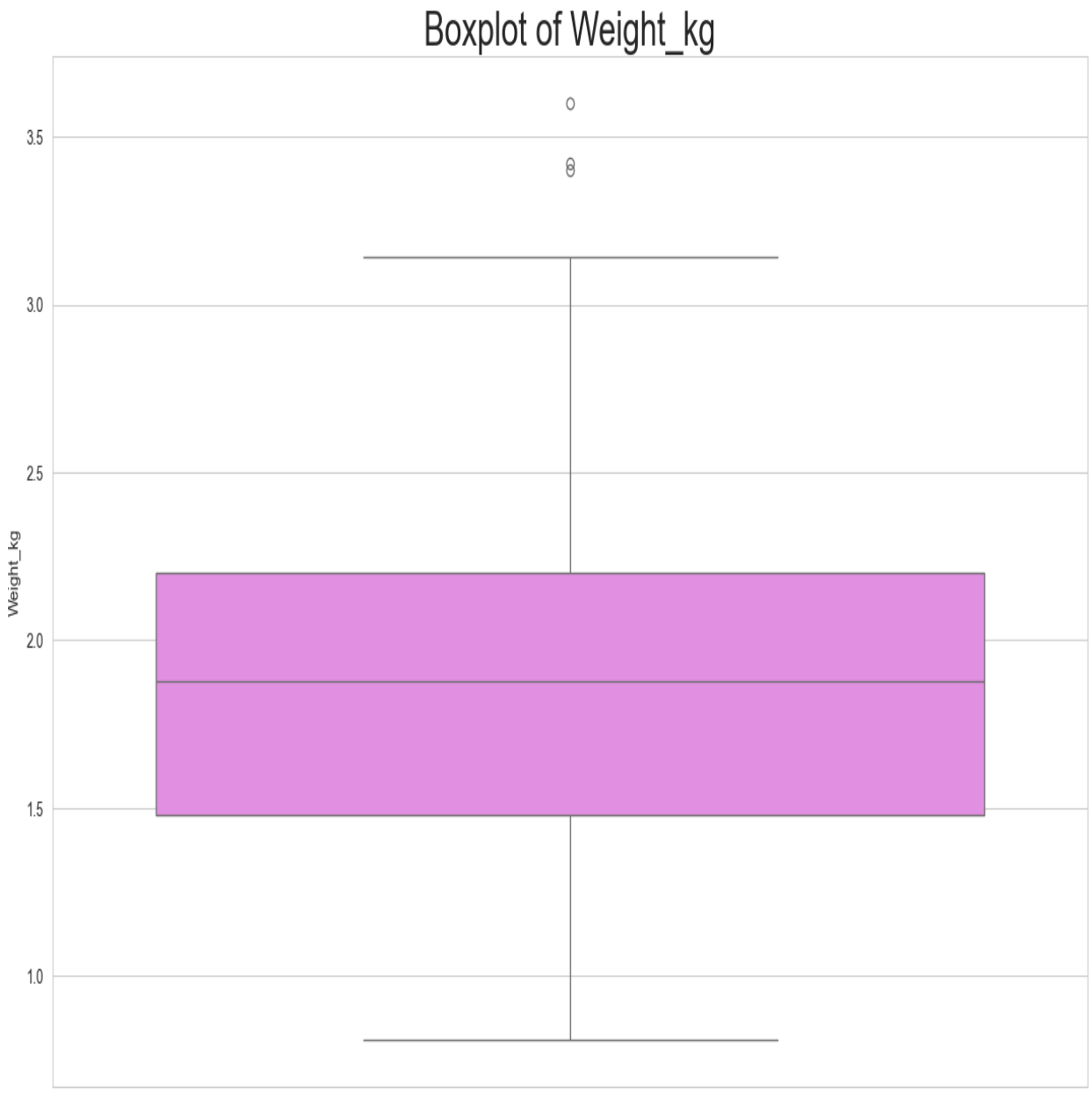
CPU frequency is concentrated within a narrow range, with a small interquartile range and no extreme outliers. This indicates that most laptops operate at similar clock speeds, reflecting limited variability in processor frequency.

17



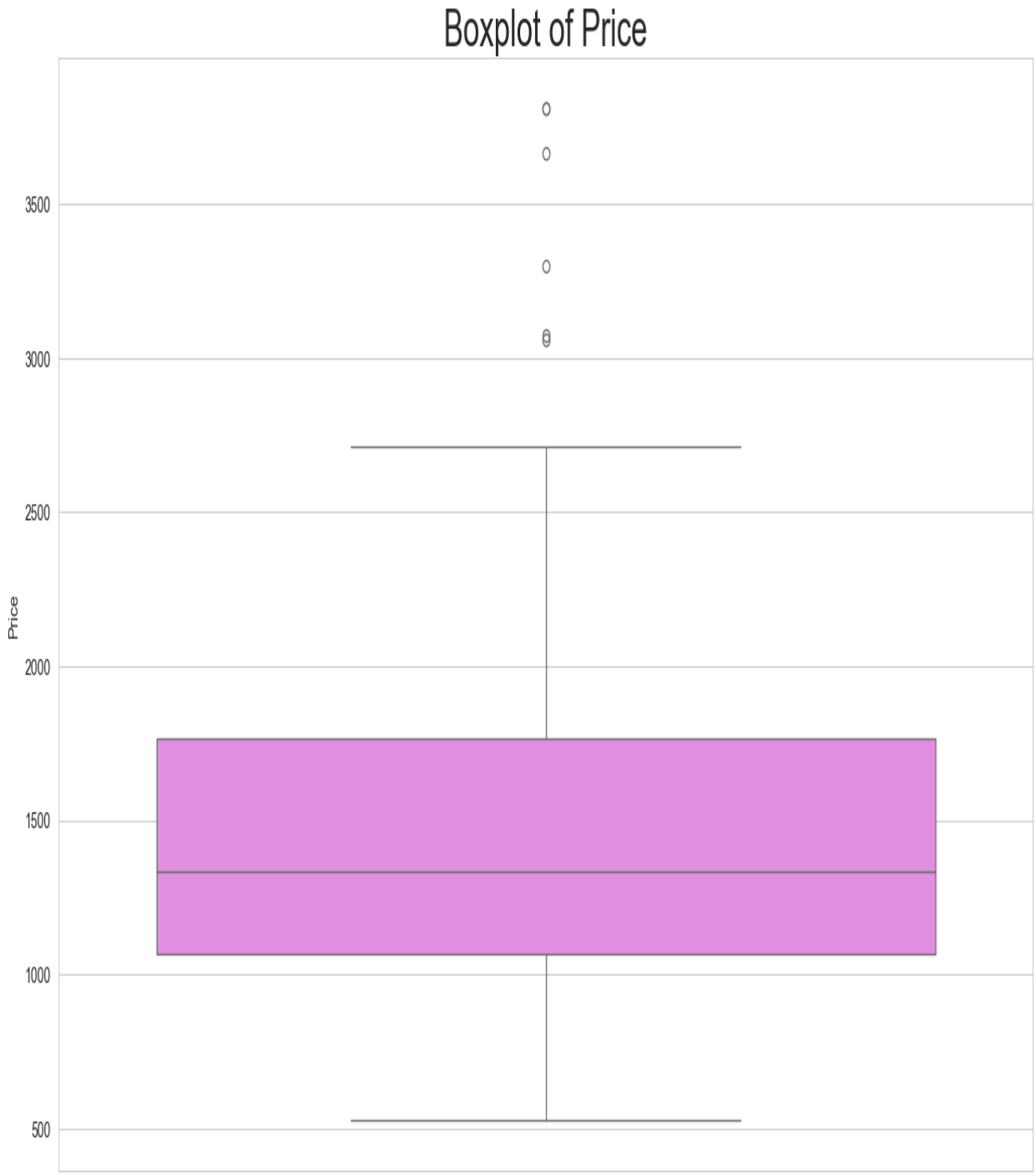
RAM values are highly concentrated around a single value (approximately 8 GB), resulting in an extremely small interquartile range. A few lower and higher outliers (e.g., 4 GB, 6 GB, 12 GB, and 16 GB) indicate less common configurations. Overall, RAM is largely standardized.

18



Weight shows a relatively narrow spread, with most laptops clustered around the median. A few upper outliers indicate heavier models, which are less common. The distribution is slightly right-skewed, suggesting that most laptops are lightweight to moderately heavy.

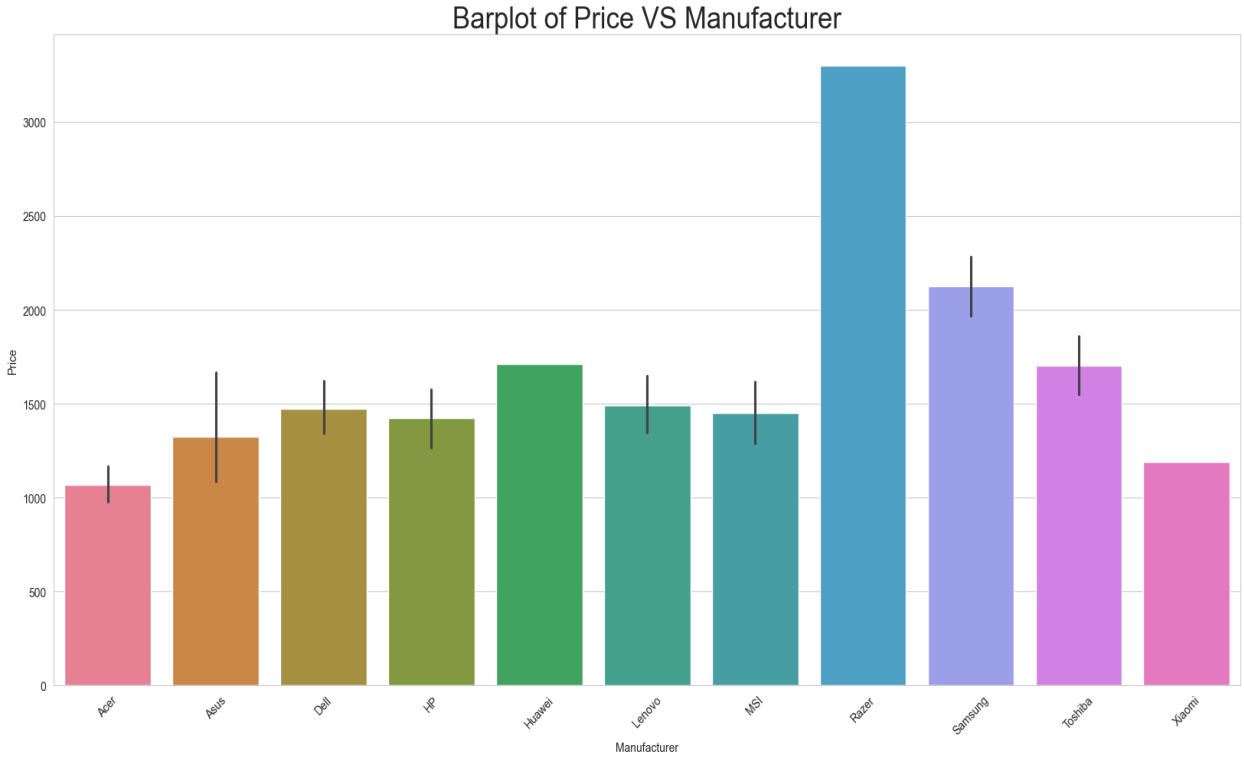
19



Price exhibits a wide spread with a large interquartile range and several high-value outliers. The longer upper whisker indicates a right-skewed distribution, suggesting the presence of a small number of high-end, premium laptops compared to the majority of moderately priced models.

Bivariate Analysis

1



The bar plot illustrates the average laptop price across different manufacturers, along with variability indicated by the error bars.

Razer stands out with the highest average price, suggesting that its laptops are positioned in the premium segment, likely due to high-end specifications and brand positioning.

Samsung and Toshiba also show relatively higher average prices, indicating a focus on mid-to-high range laptops.

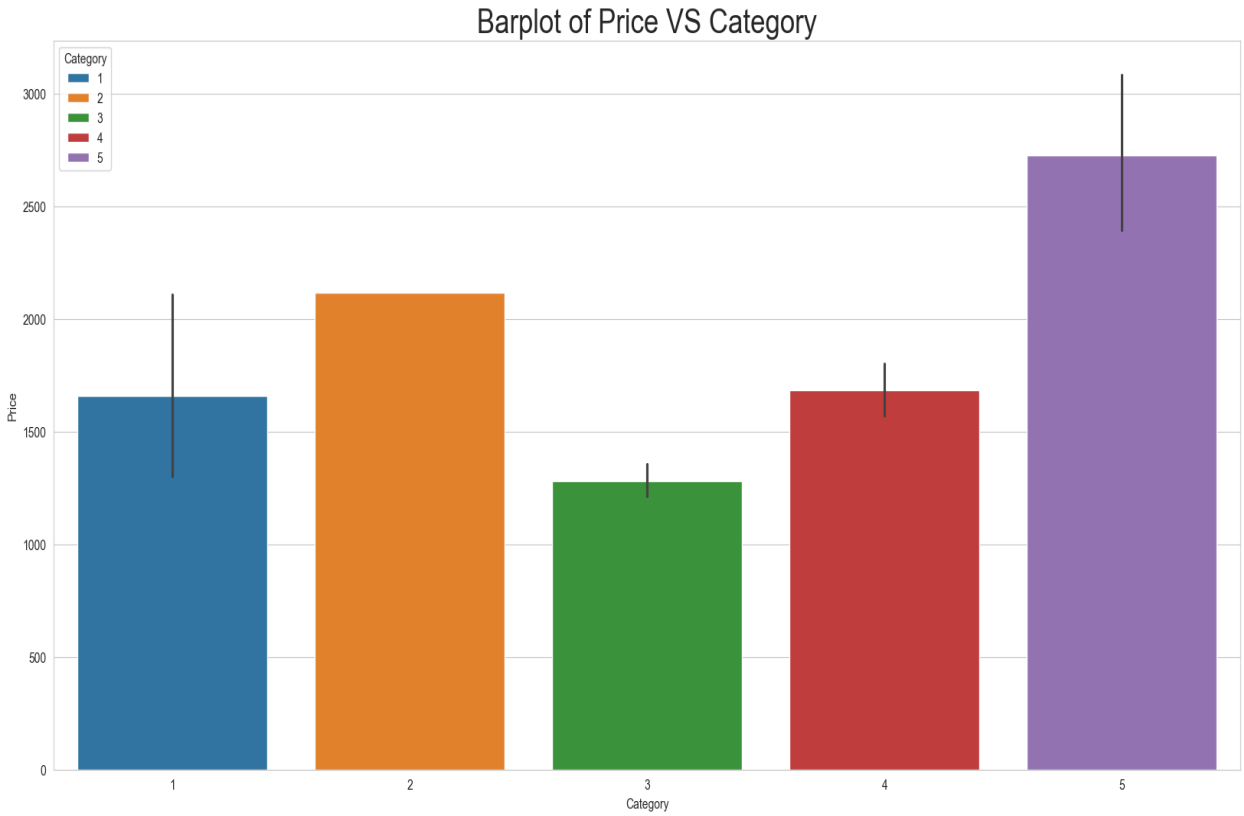
Huawei, Dell, Lenovo, MSI, and HP fall within a moderate price range, reflecting a balance between performance and affordability.

Acer and Xiaomi have the lowest average prices, suggesting that these brands primarily target the budget or value-oriented segment.

The error bars reveal that some manufacturers (such as Asus and Dell) exhibit greater price variability, indicating a wider range of configurations and models, whereas brands like Acer and Xiaomi show more consistent pricing.

Overall, the plot highlights clear brand-level pricing differences, showing that manufacturer identity plays a significant role in laptop pricing. This suggests that price is influenced not only by technical specifications but also by brand strategy and market positioning.

2

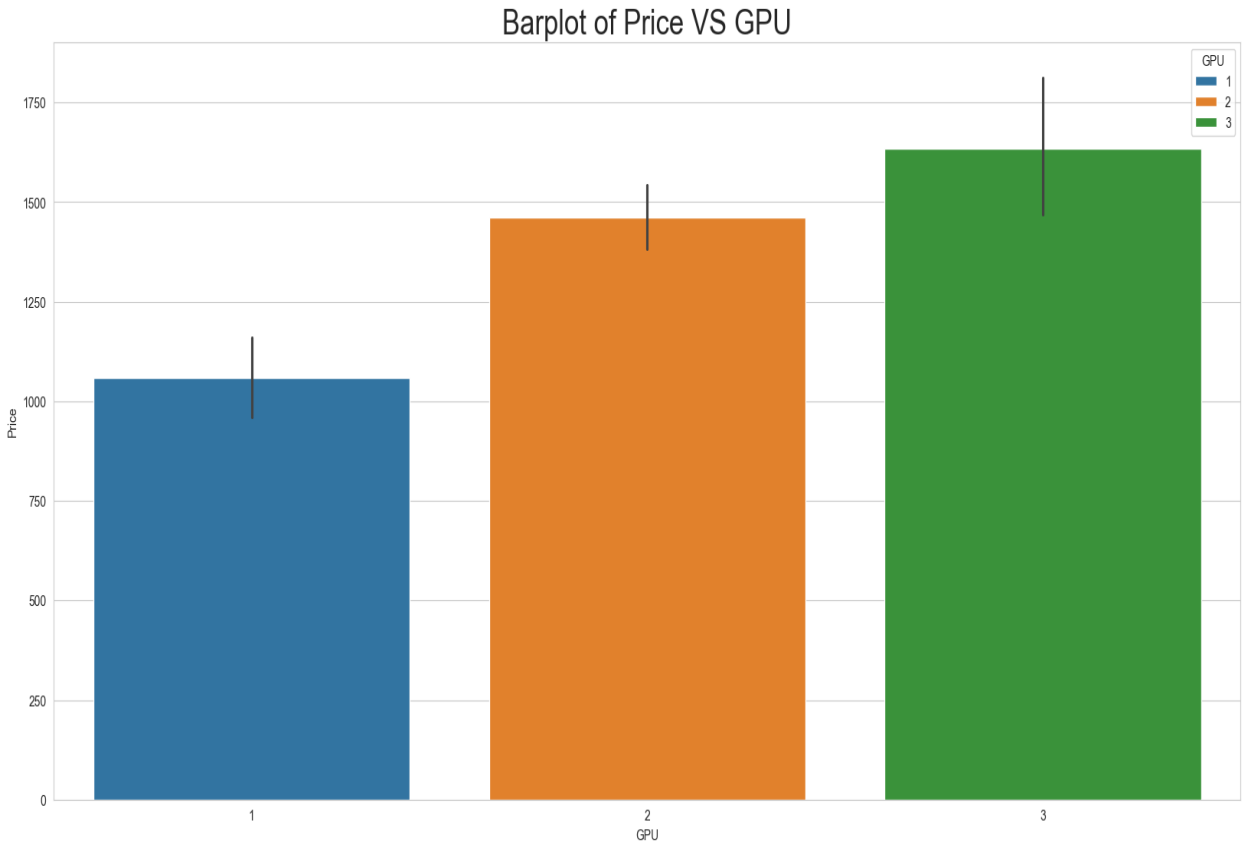


The bar plot shows the average laptop price for each category along with price variability. Notebook laptops (Category 3) have the lowest average price, indicating they are the most affordable category in the dataset. Gaming laptops (Category 1) and Ultrabooks (Category 4) have similar average prices, both higher than Notebooks but lower than Workstations.

Netbooks (Category 2) show a relatively high average price with very little variation, suggesting that prices within this category are consistent in the dataset. Workstations (Category 5) have the highest average price among all categories and also exhibit high variability, indicating a wide range of prices for professional-grade systems.

The error bars reveal that Gaming and Workstation categories have the greatest price dispersion, while Notebook and Netbook categories show more stable pricing. Overall, the plot indicates that laptop price increases with category complexity, with professional and high-performance categories costing more.

3

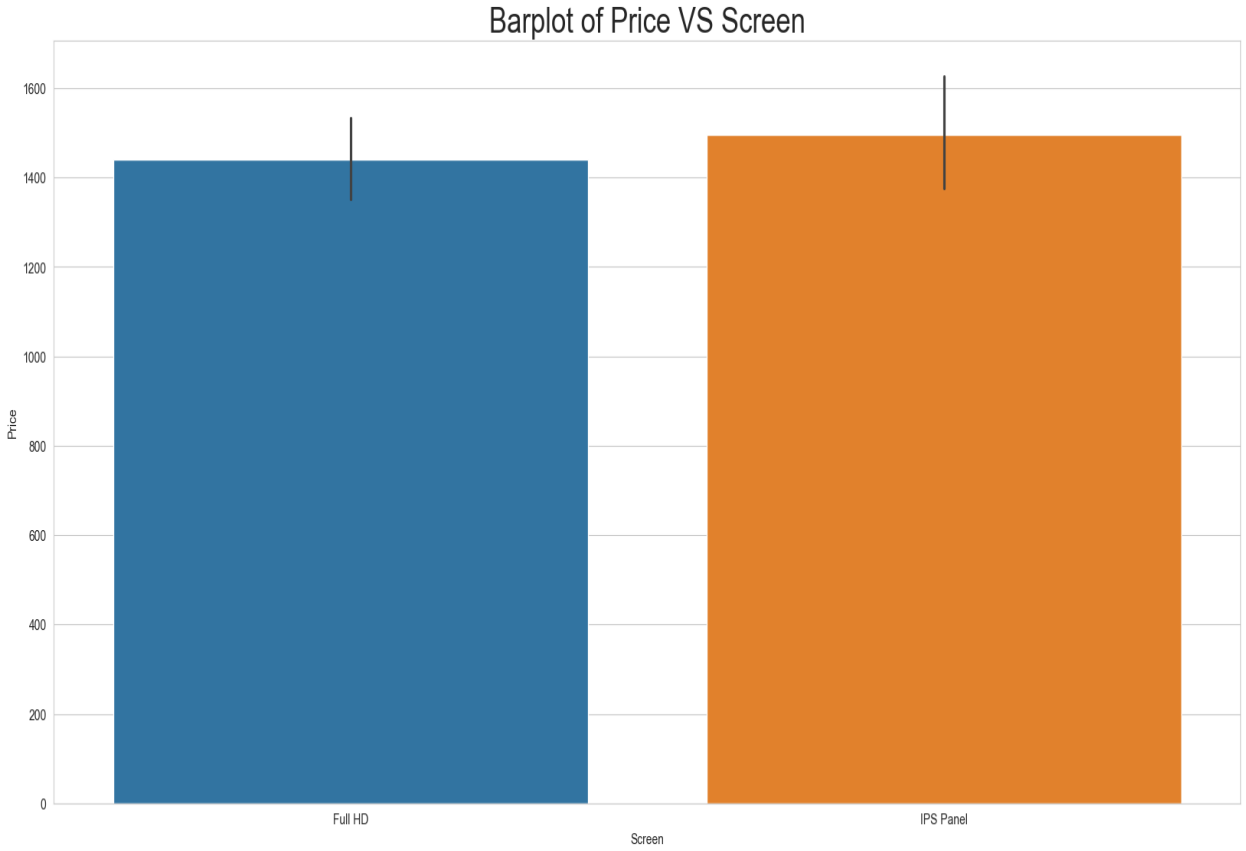


The bar plot illustrates the relationship between laptop price and GPU type. Laptops equipped with AMD GPUs (1) have the lowest average price, indicating their association with more budget-oriented systems. Intel GPUs (2) show a higher average price than AMD, suggesting their use in mid-range laptops. NVidia GPUs (3) have the highest average price, reflecting their presence in high-performance and premium laptops.

The error bars represent price variability within each GPU category. NVidia GPUs (3) exhibit the greatest variability, indicating a wide range of prices across different laptop configurations. AMD GPUs (1) show greater variability than Intel GPUs, despite having a lower average price, suggesting a broader spread of pricing in AMD-based systems. Intel GPUs (2) have the smallest error bar, indicating the most consistent pricing among the three GPU types.

Overall, the plot demonstrates that both average laptop price and price variability increase with GPU capability, confirming that GPU type is a significant factor influencing laptop pricing.

4

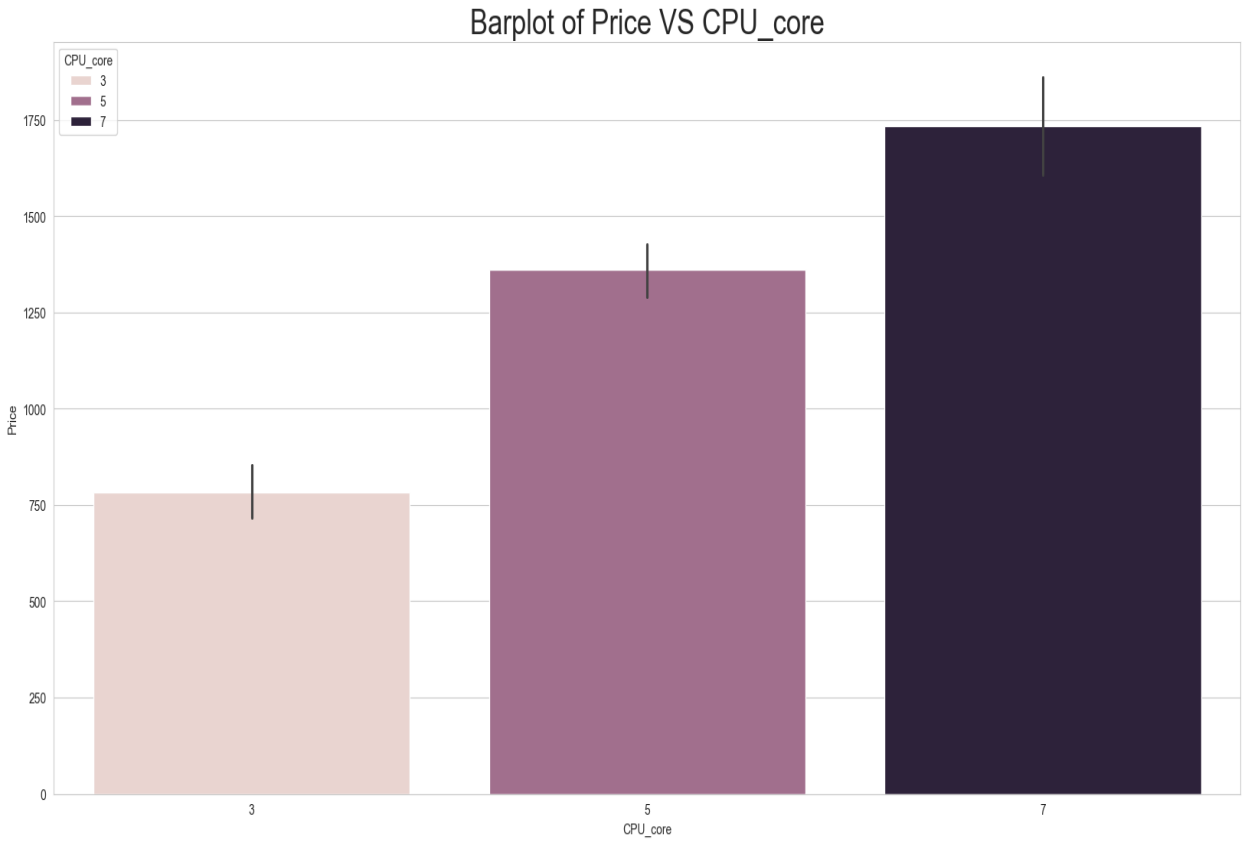


The bar plot shows the relationship between laptop price and screen type. Laptops with IPS Panel screens have a slightly higher average price compared to those with Full HD screens, indicating that IPS displays are generally associated with more premium laptops.

The error bars represent price variability within each screen category. Both screen types exhibit a similar level of variability, although IPS Panel laptops show a marginally higher spread in prices. This suggests that IPS screens are used across a broader range of laptop configurations.

Overall, the plot indicates that screen type has a modest influence on laptop pricing, with IPS Panel displays contributing to a slightly higher average price compared to Full HD screens.

5

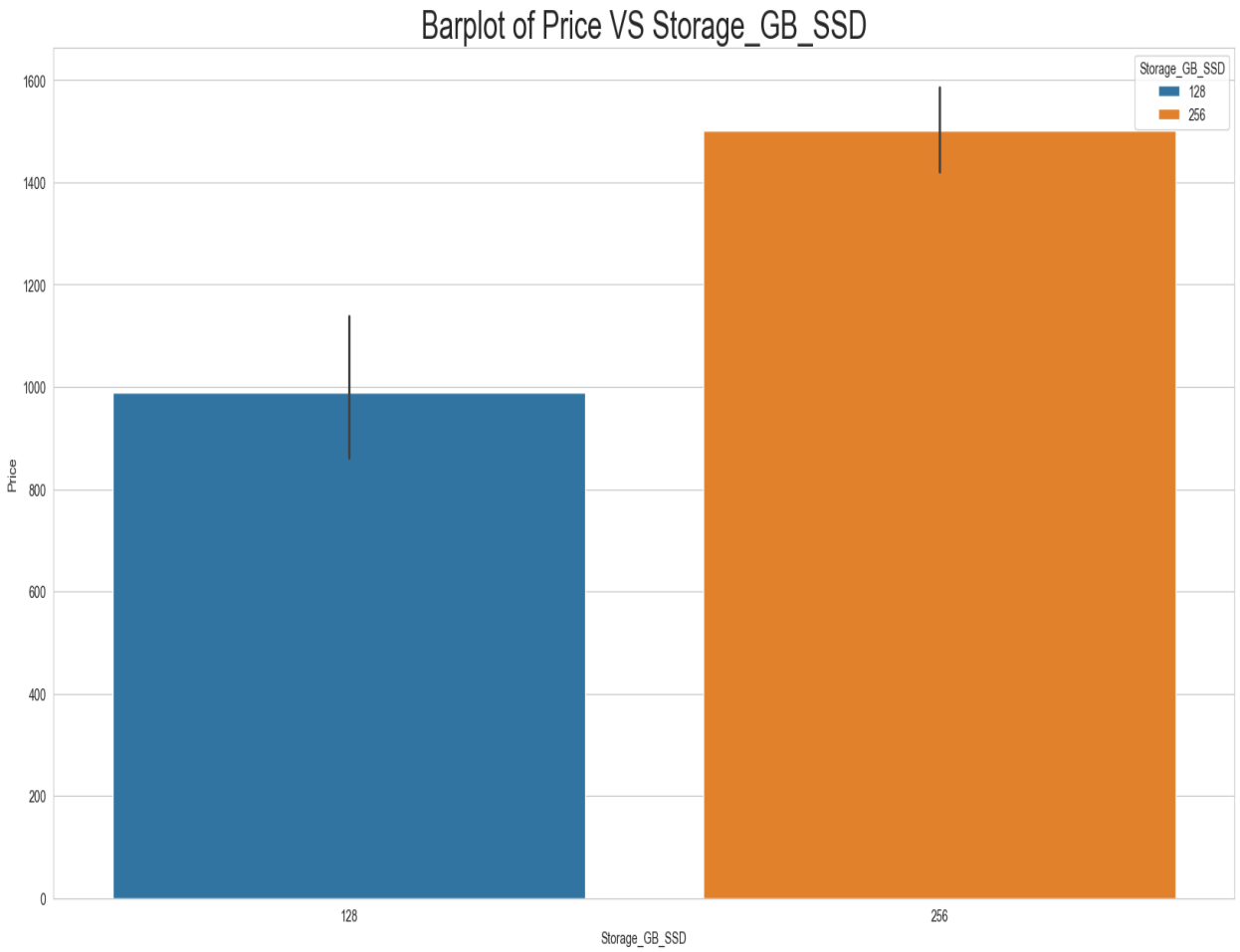


The bar plot illustrates the relationship between laptop price and CPU core type. Laptops with Intel i3 processors (CPU_core = 3) have the lowest average price, indicating their use in entry-level or budget systems. Laptops with Intel i5 processors (CPU_core = 5) show a higher average price, reflecting their position in mid-range laptops. Intel i7 processors (CPU_core = 7) have the highest average price, indicating their use in high-performance and premium laptops.

The error bars represent price variability within each CPU category. Price variability increases with CPU capability, with i7-based laptops showing the largest variation, followed by i5, while i3 laptops exhibit the least variability. This suggests that higher-end processors are used across a wider range of laptop configurations.

Overall, the plot shows a clear positive relationship between CPU performance and laptop price, confirming that CPU core type is a strong factor influencing laptop pricing.

6

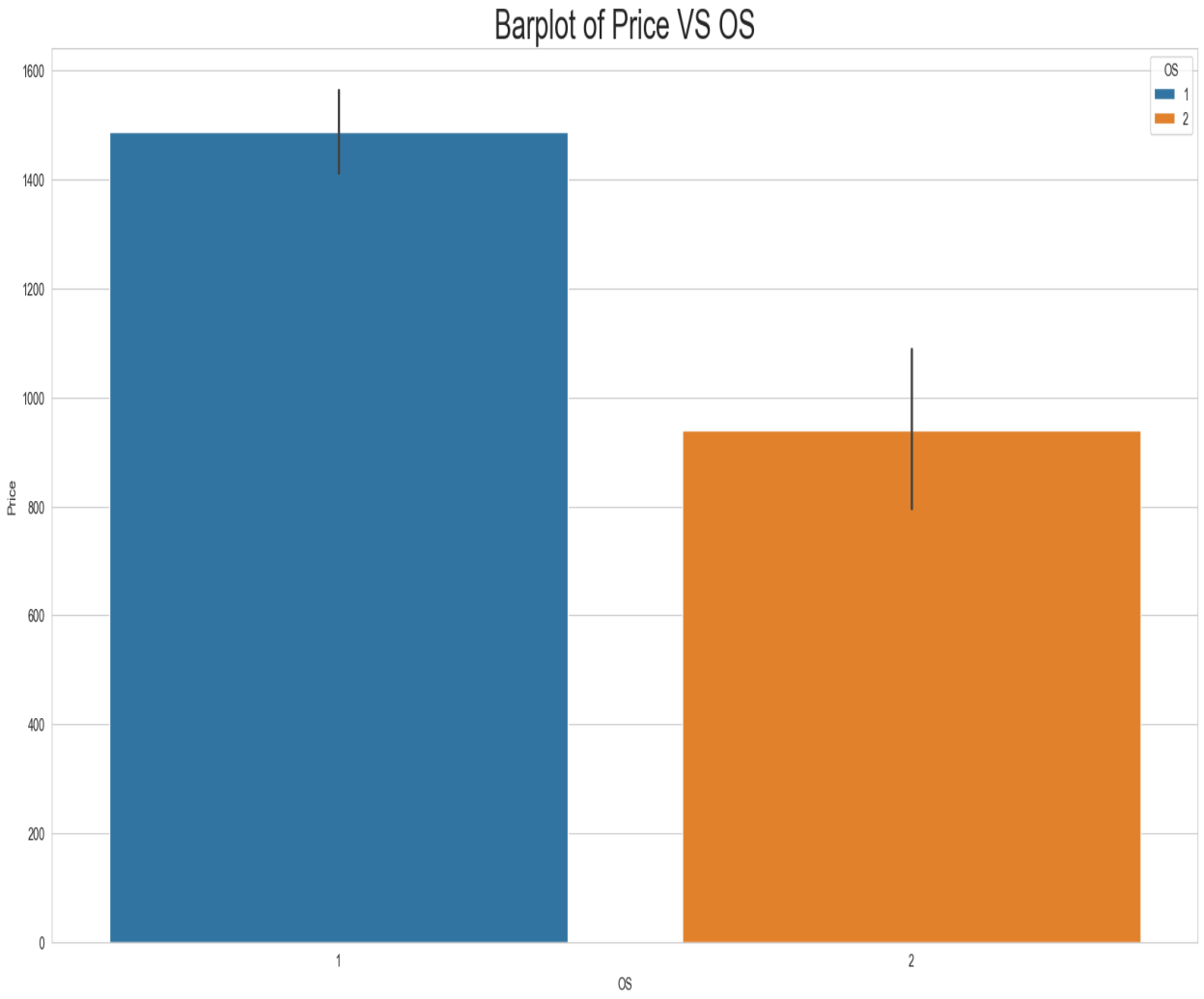


The barplot shows a clear positive association between SSD storage capacity and laptop price. Laptops with 256 GB SSDs have a substantially higher average price than those with 128 GB SSDs, indicating that increased storage capacity is associated with higher cost.

The error bars (variability around the mean) suggest moderate price dispersion within each storage category. While both groups show some variation, the 256 GB category remains consistently more expensive, implying that storage capacity is an important pricing factor rather than the difference being driven by a few extreme values.

Overall, the plot indicates that SSD storage size is a strong categorical determinant of price, with higher-capacity configurations commanding a premium.

7



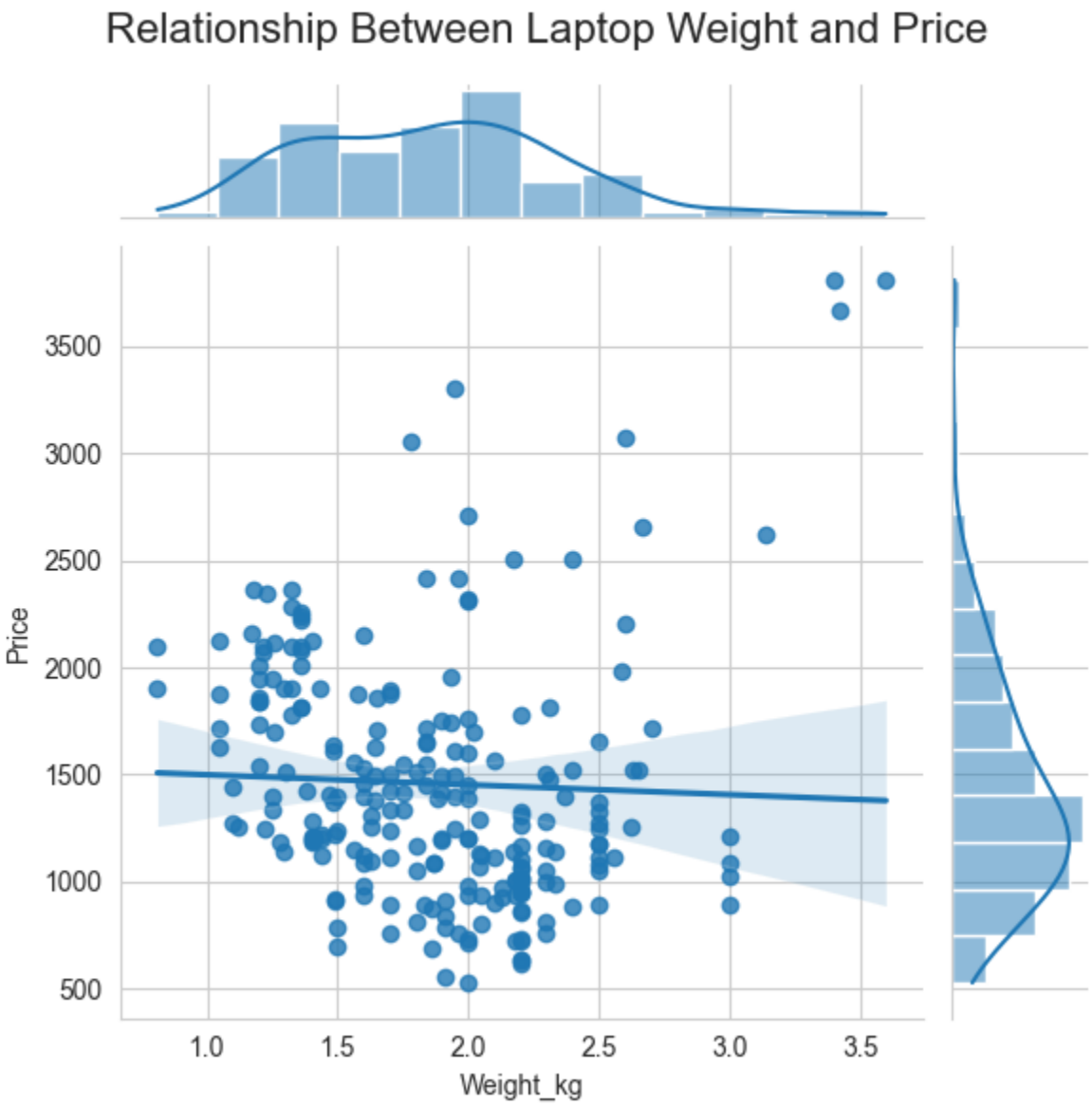
The bar plot shows average laptop prices across two operating system categories, with error bars indicating price variability.

Laptops with OS Windows (category 1) have a noticeably higher mean price than those with OS Linux (category 2), indicating a clear association between operating system and pricing. Prices for OS 1 are more tightly clustered, suggesting consistent premium positioning, while OS 2 shows greater variability, reflecting a wider range of mostly lower- to mid-priced laptops.

The limited overlap of error bars and clear separation between means suggest a meaningful difference in average prices. However, this plot does not imply causality, as pricing differences may also be influenced by hardware specifications, brand strategies, or target markets.

Overall, the operating system appears to be a strong categorical indicator of laptop price, with OS 1 associated with higher-priced models and OS 2 with more affordable options.

8

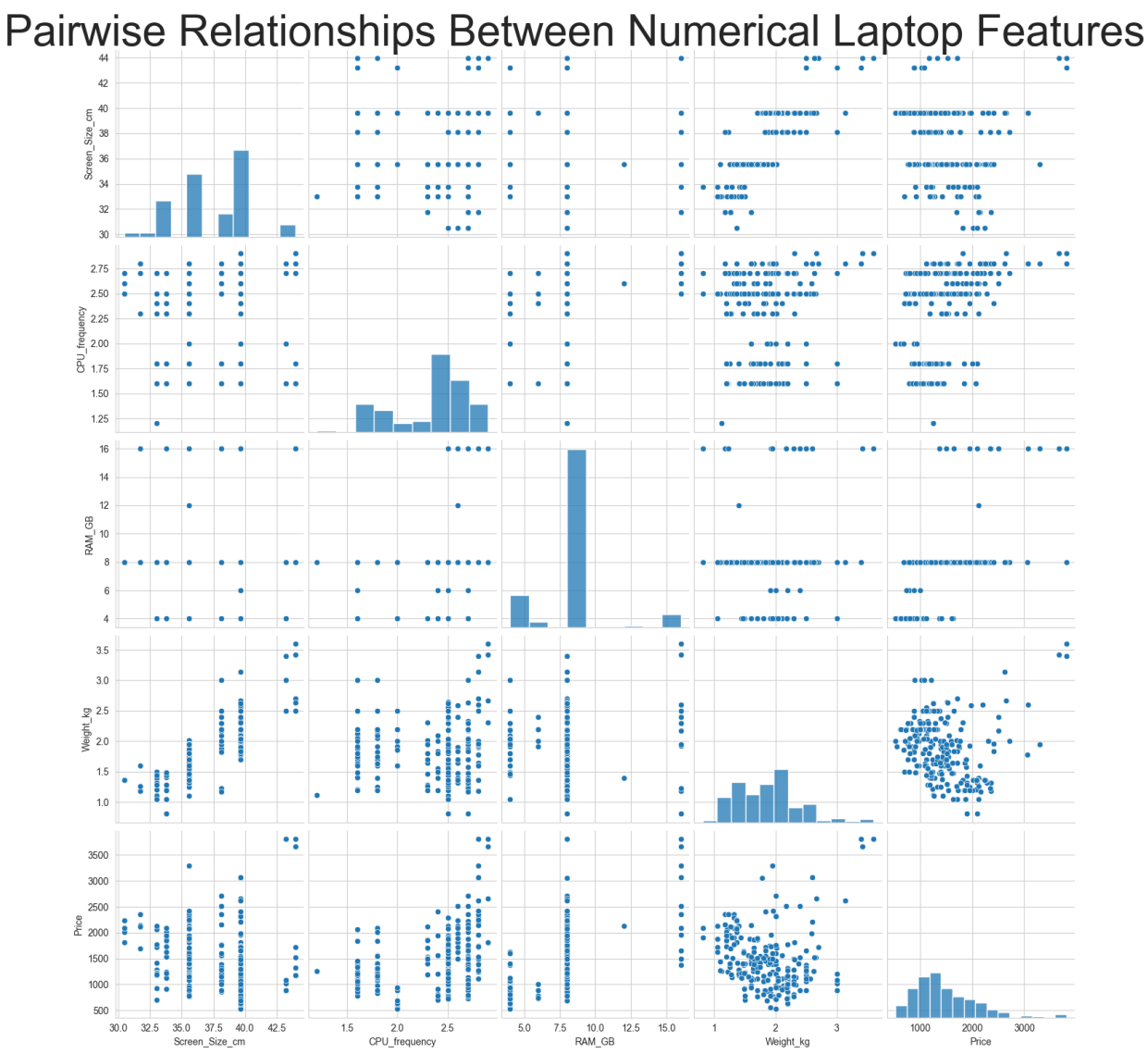


The joint plot correctly displays the relationship between Weight_kg and Price, combining scatter, marginal distributions, and a fitted regression line. The scatter shows a weak negative association, indicating that heavier laptops tend to be slightly cheaper on average, though the relationship is not strong. The wide dispersion of points around the regression line suggests low explanatory power of weight alone for predicting price.

The marginal distributions indicate that weight is moderately right-skewed, with most laptops concentrated at lower weights, while price is positively skewed, with a long right tail corresponding to premium models. The regression line and confidence band are consistent with the data pattern and do not indicate model misspecification.

Overall, the plot is correctly constructed and supports the conclusion that laptop weight has only a weak and noisy relationship with price.

9



The pairplot visualizes relationships among the numerical variables Screen_Size_cm, CPU_frequency, RAM_GB, Weight_kg, and Price. Diagonal panels show univariate distributions, while off-diagonal panels illustrate bivariate associations in terms of direction, strength, and structure.

Univariate Distributions

Screen_Size_cm is concentrated at specific values, reflecting standardized laptop display dimensions.

CPU_frequency is clustered around higher frequencies, with fewer observations at lower values.

RAM_GB is highly concentrated at a single value, indicating limited variability in memory configurations.

Weight_kg exhibits a right-skewed distribution, with most laptops in the lower-weight range and a small number of heavier models.

Price shows a positively skewed distribution, with most laptops priced in the lower to mid range and a long upper tail representing premium models.

Bivariate Relationships

CPU_frequency vs Price exhibits a moderate positive association with increasing price variability at higher frequencies, indicating heteroscedasticity.

CPU_frequency vs RAM_GB shows weak to moderate association with clear clustering, reflecting fixed CPU–RAM combinations: CPU frequency and RAM size are not strongly correlated—an increase in CPU frequency does not consistently lead to an increase in RAM. However, there is some relationship, which is why the association is described as weak to moderate rather than completely absent.

The clear clustering indicates that the data points form distinct groups instead of being spread smoothly. These clusters arise because devices are typically manufactured and sold in standardized configurations (for example:

2.4 GHz with 8 GB RAM, 3.0 GHz with 16 GB RAM, etc.), rather than arbitrary combinations.

As RAM increases, laptop prices generally increase.

However, RAM only takes a few fixed values (for example, 4, 8, 16 GB), so the relationship appears in steps or horizontal bands rather than a smooth trend.

Within each RAM level, prices vary widely, meaning RAM alone does not fully determine price.

RAM_GB vs Weight_kg and Screen_Size_cm show weak or negligible associations, suggesting relative independence.

Screen_Size_cm vs Weight_kg exhibits a moderate positive relationship, indicating that larger screens are generally associated with heavier laptops.

Screen_Size_cm vs Price shows a weak positive trend with substantial dispersion.

The Weight_kg vs Price scatter plot shows a weak negative association between laptop weight and price. As weight increases, there is a slight tendency for prices to decrease, indicating that lighter laptops are generally priced higher. However, the relationship is not linear, as the data points are widely dispersed rather than aligned along a straight line. The substantial variability in price for laptops with similar weights highlights a noisy pattern, suggesting that weight alone does not strongly explain price differences and that other factors such as internal specifications, build quality, and brand value play a more significant role

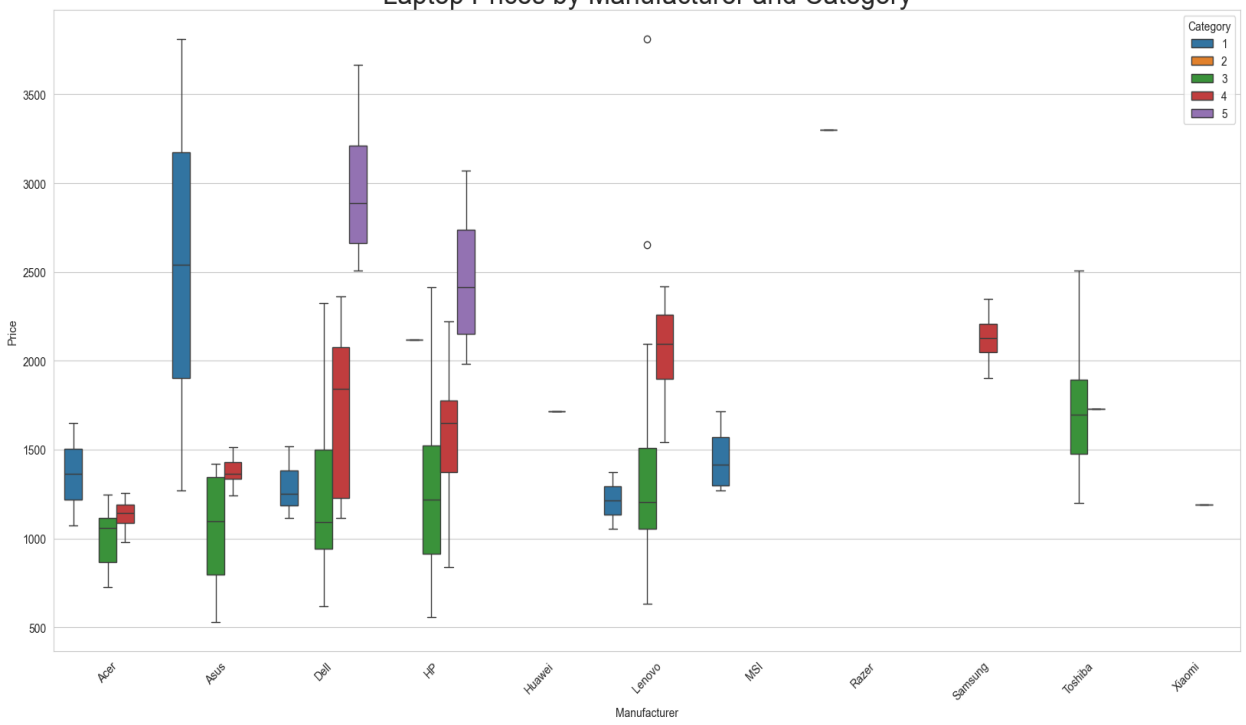
Plots of Weight_kg against CPU-related variables reveal no evident linear trends, with data points appearing scattered across the range, suggesting little to no direct association.

Structural Patterns and Implications

Many scatterplots exhibit vertical and horizontal banding due to discrete predictors and standardized configurations, resulting in clustered rather than continuous point clouds. Several relationships show increasing price variability at higher performance levels, indicating heteroscedasticity. These patterns suggest that linear correlations may understate associations and that group-based comparisons, interaction terms, or robust modeling approaches may be more appropriate.

Multivariate Analysis

1



The box plot presents the distribution of laptop prices across different brands, with prices further differentiated by laptop categories. Overall, both brand and category strongly influence laptop pricing, and substantial variation is observed within and across manufacturers.

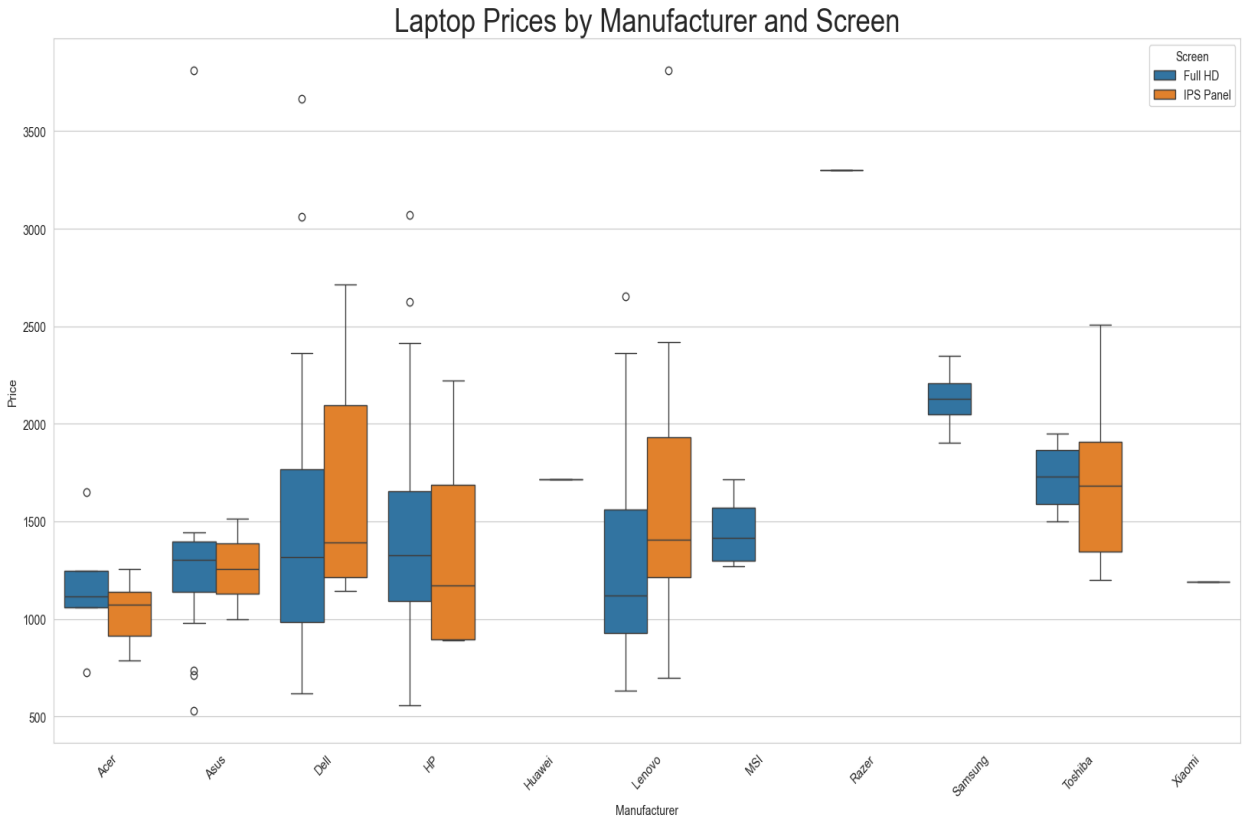
Across most brands, higher categories (Ultrabook and Workstation) consistently show higher median prices, while Notebook and Netbook categories tend to occupy the lower price range. This indicates that, regardless of brand, premium categories command higher prices due to better hardware and build quality.

Dell and HP display wide price distributions across multiple categories, suggesting that these brands offer laptops ranging from budget notebooks to high-end workstations. Their higher-category models show significantly higher medians, highlighting strong segmentation by category. Asus and Acer also cover multiple categories but with generally lower median prices, indicating a stronger presence in the mid-range and budget segments.

Lenovo shows considerable variability across categories, with some high-priced models appearing as outliers, reflecting a mix of affordable laptops and premium-category offerings. Razer stands out with consistently high prices and minimal variability, implying a focus on premium gaming and workstation-class laptops. MSI and Samsung tend to cluster in the upper-mid to premium range, particularly in higher categories, with relatively less spread.

Brands such as Huawei and Xiaomi show limited variability and fewer category representations, suggesting a narrower product range in the dataset.Overall, the plot demonstrates that laptop category largely determines the price level, while brand influences the price range and variability within each category.

2



The box plot illustrates how laptop prices vary across manufacturers and screen types (Full HD vs IPS Panel). Overall, laptops with IPS Panel screens tend to have higher median prices than those with Full HD screens for most brands, indicating that IPS displays are generally associated with more premium models.

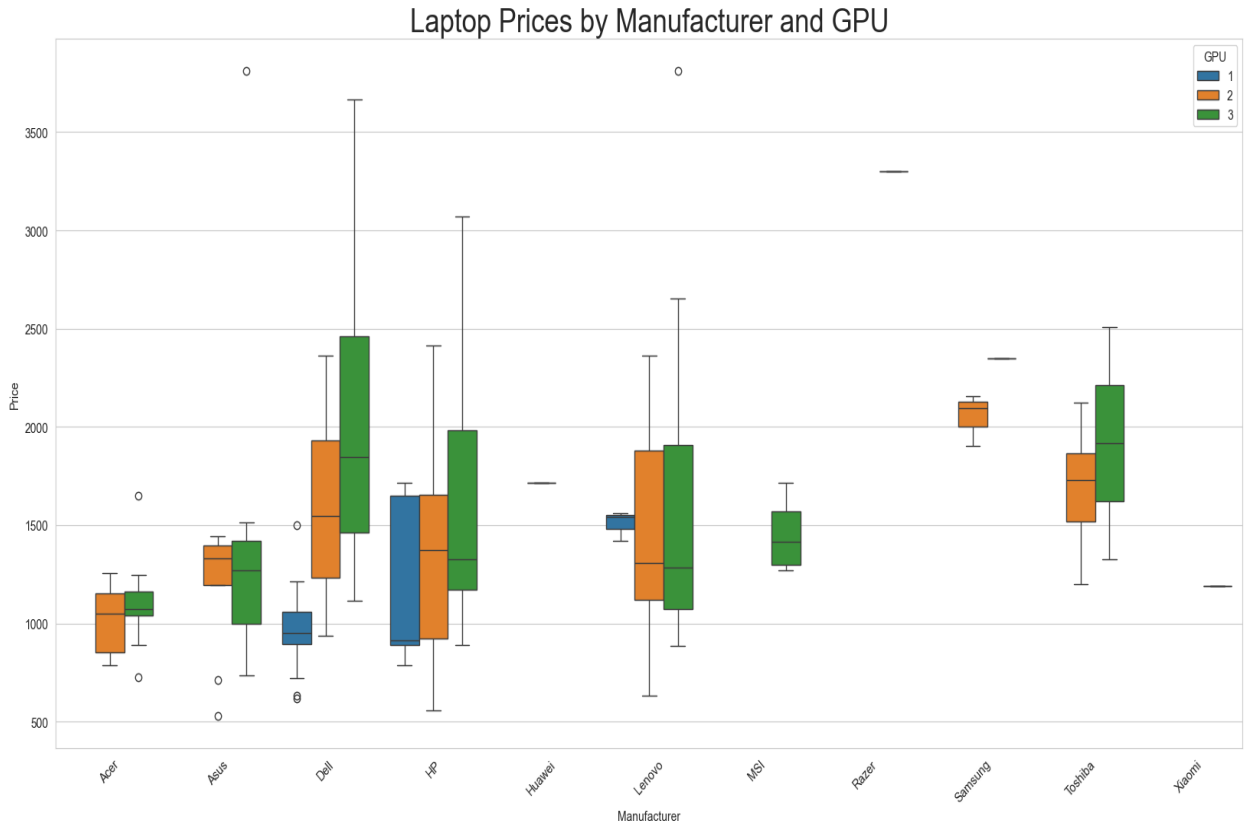
For major manufacturers such as Dell, HP, Lenovo, and Asus, IPS Panel laptops show both higher median prices and wider price ranges, suggesting that IPS screens are commonly used in mid-to-high-end configurations. Full HD models from these brands are typically more affordable and exhibit comparatively lower medians.

Razer and Samsung stand out with consistently high prices, particularly for Full HD models, reflecting their premium market positioning. These brands show limited variability, implying a focused product lineup. MSI shows moderate pricing with relatively tight distributions, indicating consistent pricing across screen types.

Brands such as Acer and Toshiba display moderate price ranges, with IPS Panel models generally priced higher than Full HD counterparts. Huawei and Xiaomi show minimal variation and fewer observations, suggesting limited product representation in the dataset.

Overall, the plot demonstrates that screen type influences laptop pricing within each brand, with IPS Panel displays typically commanding a price premium. However, the manufacturer effect remains strong, as premium brands maintain higher prices regardless of screen type.

3



The box plot illustrates how laptop prices vary across manufacturers and GPU types. Overall, both brand and GPU type have a strong influence on pricing, with higher-end GPUs consistently associated with higher prices within the same brand.

Across most manufacturers, laptops with GPU type 3 (NVidia) have the highest median prices, followed by GPU type 2 (Intel), while GPU type 1 (AMD) generally shows the lowest prices. This pattern indicates that stronger GPU capability leads to higher laptop prices regardless of brand.

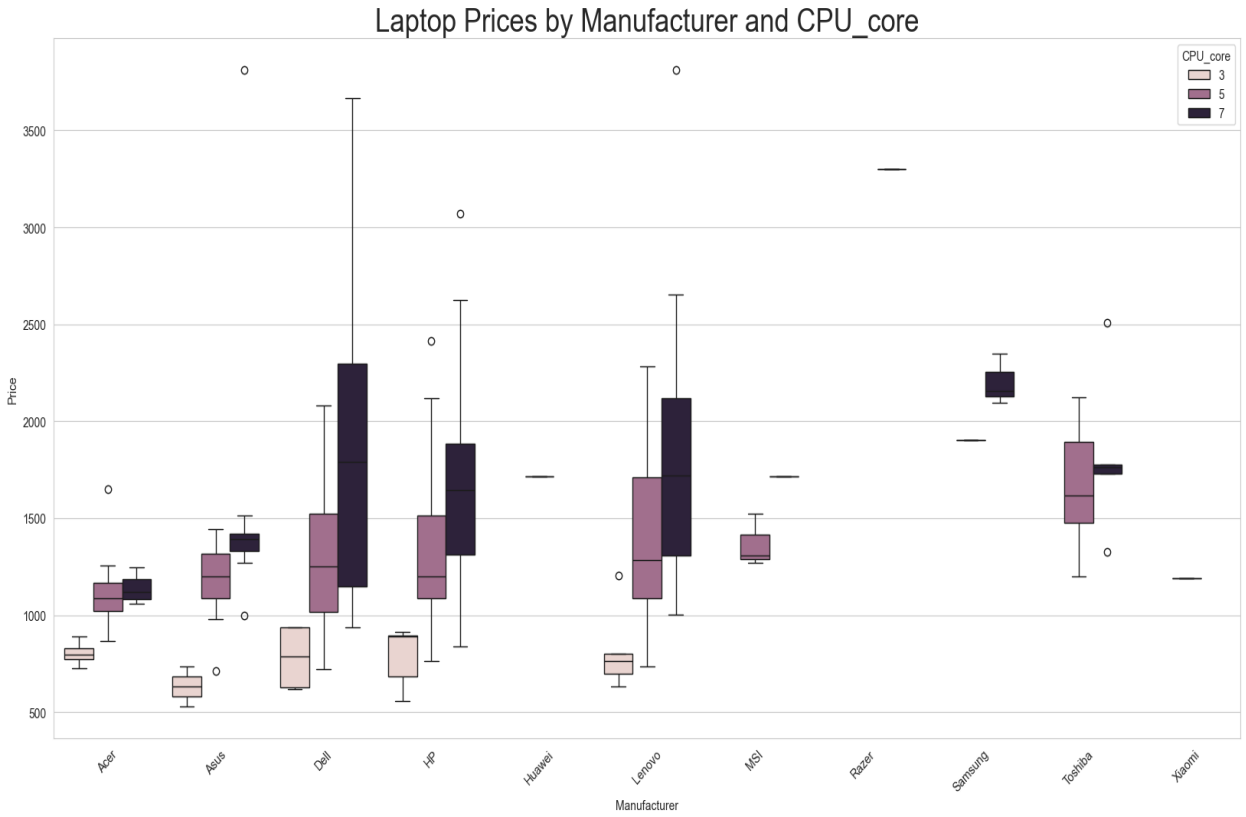
Dell, HP, and Lenovo show wide price distributions across all GPU types, suggesting a broad product portfolio ranging from budget to premium systems. In these brands, the price gap between AMD/Intel GPUs and NVidia GPUs is particularly pronounced, highlighting the premium placed on dedicated GPUs.

Asus and Acer exhibit moderate price levels with noticeable variability, especially for NVidia GPUs, indicating a mix of mid-range and high-performance laptops. MSI shows relatively tighter price distributions, implying more consistent pricing across its GPU offerings.

Premium brands such as Razer and Samsung show high prices with limited variability, reflecting a focused high-end product strategy. Huawei and Xiaomi show minimal variation, likely due to fewer models represented in the dataset.

Overall, the plot demonstrates that GPU type significantly affects laptop pricing within each manufacturer, while brand positioning determines the overall price range and variability.

4



The box plot shows how laptop prices vary across manufacturers and CPU core types. Across all brands, there is a clear upward trend in price as CPU capability increases from Intel i3 (CPU_core = 3) to i5 (5) and i7 (7).

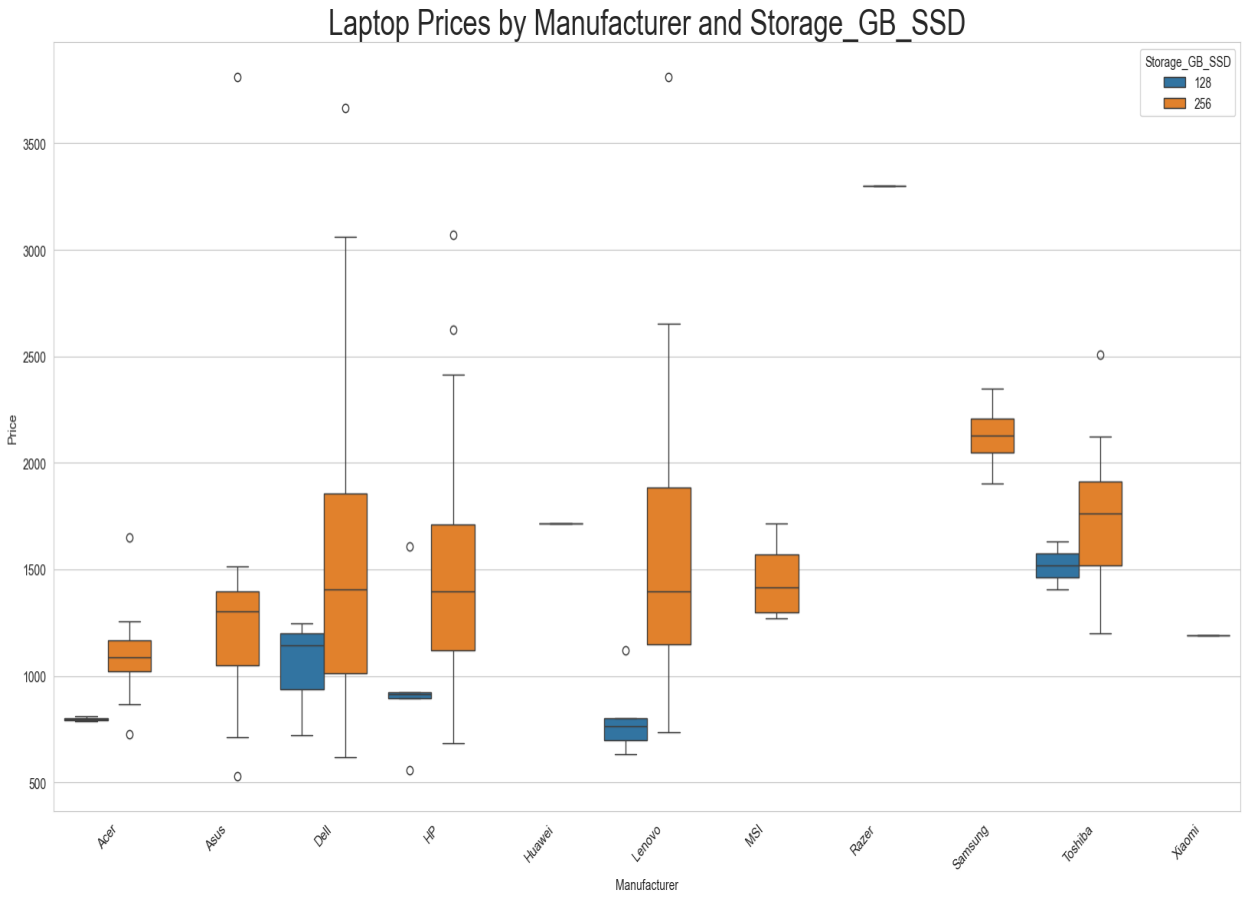
For most manufacturers, i3-based laptops have the lowest median prices, reflecting their use in entry-level systems. i5-based models occupy the mid-price range and show moderate variability, indicating a balance between performance and cost. i7-based laptops consistently have the highest median prices and the widest price ranges, highlighting their presence in premium and high-performance configurations.

Brands such as Dell, HP, and Lenovo exhibit wide price distributions across all CPU types, suggesting a broad product lineup spanning budget to high-end models. In these brands, the price gap between i3 and i7 laptops is particularly pronounced. Asus and Acer show similar trends but at generally lower price levels, indicating a stronger focus on budget and mid-range segments.

Razer and Samsung stand out with high-priced i7 models and limited variability, reflecting a premium market positioning. MSI shows relatively consistent pricing, mainly concentrated around higher-performance CPUs. Huawei and Xiaomi display limited variation, likely due to fewer models in the dataset.

Overall, the plot demonstrates that CPU core type is a strong determinant of laptop price within each brand, while the manufacturer influences the overall price range and variability.

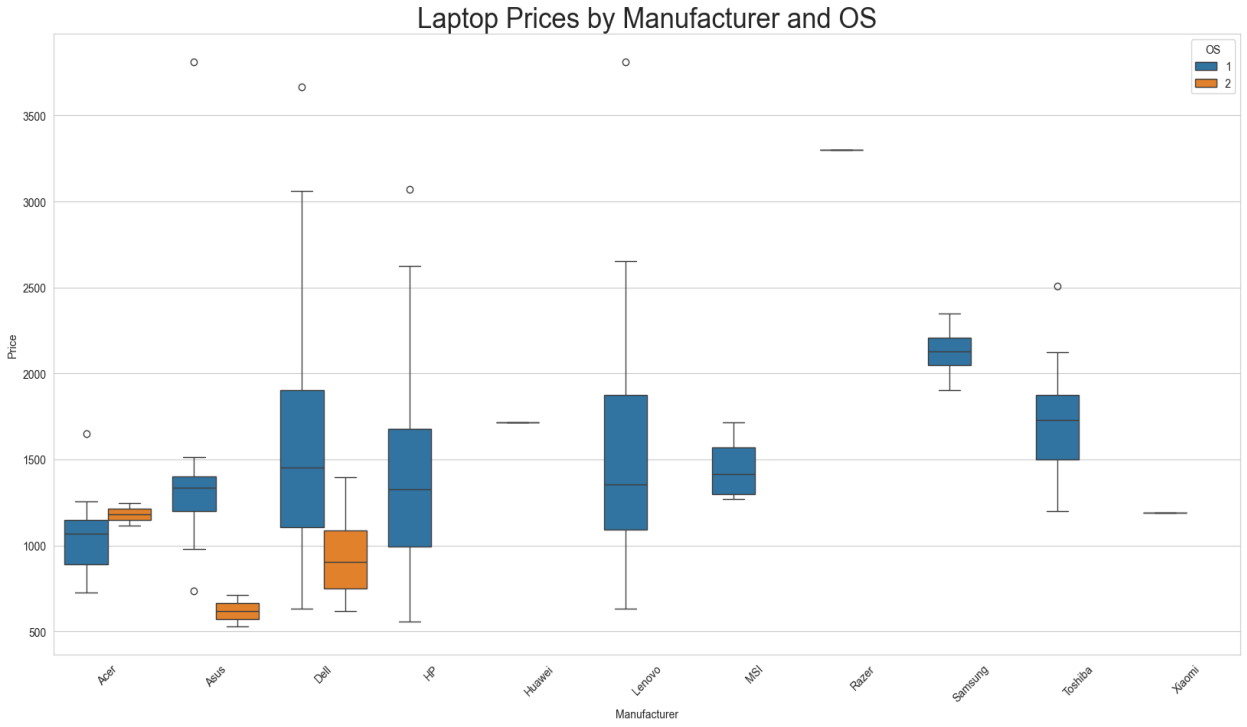
5



The boxplot analysis indicates that laptop prices vary systematically across manufacturers and SSD capacities. Across all brands, models equipped with 256 GB SSDs consistently exhibit higher median prices than their 128 GB counterparts, demonstrating a clear and uniform price premium associated with increased storage capacity. Premium manufacturers such as Razer and Samsung occupy the highest price ranges, while Acer and Asus are concentrated in the lower to mid-price segments. Brands such as Dell, HP, and Lenovo show substantially wider price dispersion and multiple high-end outliers, reflecting broader and more diverse product portfolios.

Within nearly every brand, higher SSD capacity is associated with higher prices, indicating that SSD capacity positively influences price within brand categories. At the same time, brand identity moderates the magnitude of this effect, with premium brands offering higher-priced configurations even at comparable storage levels. Overall, laptop pricing is shaped by the combined influence of manufacturer positioning and SSD capacity, with additional variability arising from other hardware specifications.

6



This analysis examines laptop prices across different manufacturers, comparing devices running Windows (OS 1) and Linux (OS 2).

Overall, laptops running Windows are consistently priced higher than their Linux counterparts across most manufacturers. Windows models also show greater price variability, indicating a wider range of configurations from entry-level to premium devices. In contrast, Linux-based laptops are generally positioned in the lower price segment with more compact price distributions, suggesting fewer high-end offerings.

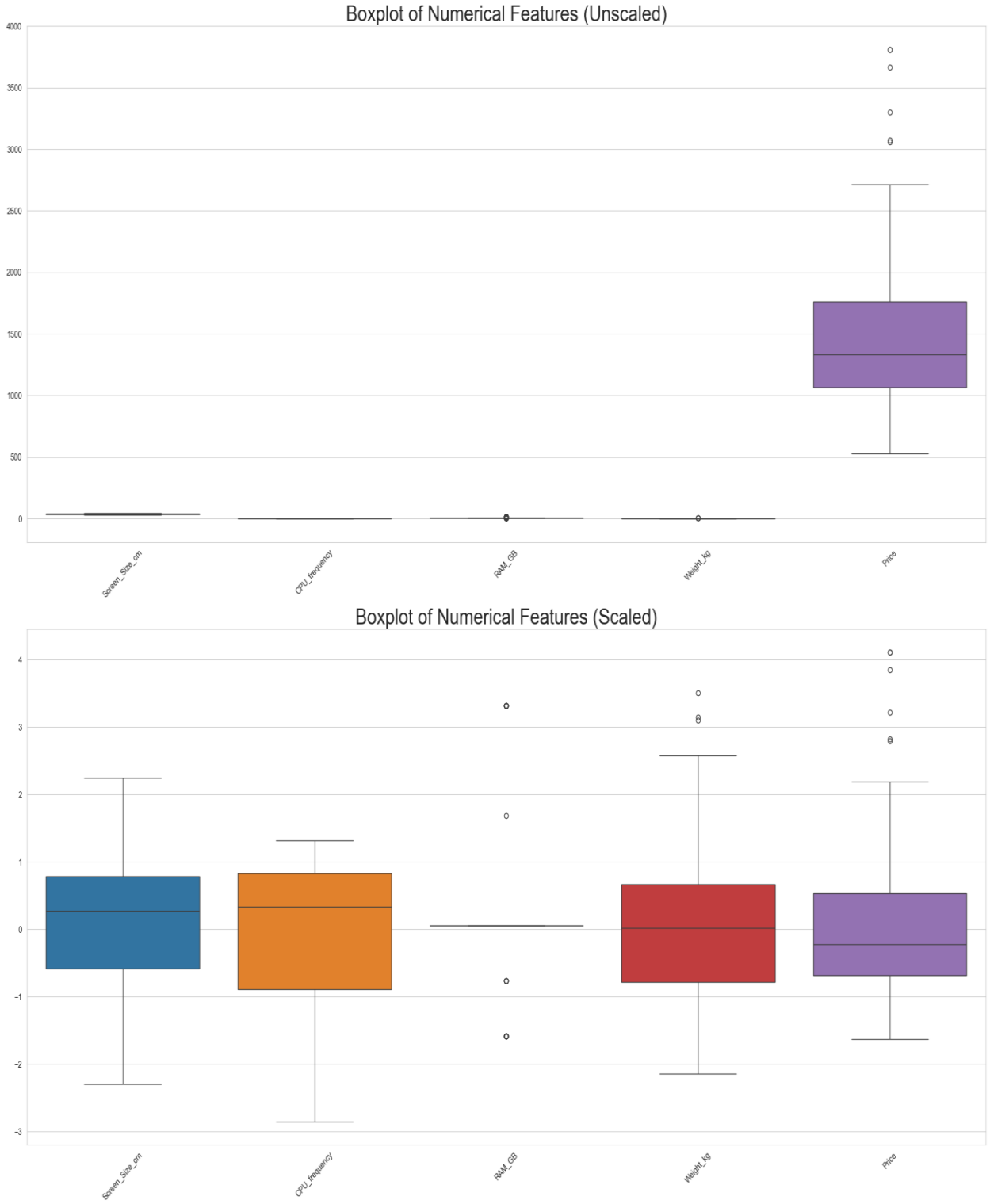
Manufacturer-wise observations show that Acer has relatively similar pricing for both operating systems, although Windows devices still command a slightly higher median price. Asus demonstrates a clear distinction, where Windows laptops occupy a higher price range while Linux laptops are clustered at the lower end, reflecting budget-focused Linux offerings.

Dell and HP exhibit wide price spreads for Windows laptops, including several high-price outliers, highlighting their diverse product portfolios. Their Linux models, however, are consistently cheaper and show limited variability. Lenovo follows a similar trend, with Windows laptops spanning from mid-range to premium prices, while Linux models remain fewer and lower priced.

Some manufacturers, such as Huawei, MSI, Razer, Samsung, and Xiaomi, predominantly or exclusively offer Windows-based laptops in this dataset. Among these, Razer stands out with very high and tightly clustered prices, confirming its premium market positioning, while MSI and Samsung focus on mid-to-high priced Windows systems.

In summary, the plot clearly indicates that Windows laptops dominate the higher price segments across brands, offering a broad spectrum of configurations, whereas Linux laptops are primarily positioned as cost-effective alternatives, with limited presence in the premium category. This reflects current market strategies where Windows is associated with mainstream and high-end consumer demand, while Linux is more commonly offered in budget or specialized models.

7



Boxplots were used to examine the distribution, spread, and outliers of numerical variables in both unscaled and scaled forms. This allows interpretation in original units as well as comparison of relative variability across features

Unscaled Boxplot

In the unscaled boxplot, variables are shown in their original measurement units.

Price exhibits the largest spread, with a wide interquartile range and several extreme upper outliers. This indicates substantial variation in laptop prices and the presence of a small number of high-end models. The longer upper whisker confirms a right-skewed distribution.

Weight (kg) shows moderate dispersion with a few upper outliers, indicating that heavier laptops exist but are relatively uncommon.

RAM (GB) is highly concentrated at lower values, with a very small interquartile range. One or two visible outliers correspond to high-memory laptops, but most observations lie close together.

CPU frequency displays very limited variability, suggesting most laptops operate within a narrow clock-speed range.

Screen size (cm) also shows a tight distribution, reflecting standardization in display sizes.

Due to the much larger numerical scale of price, other variables appear compressed, which reflects scale differences rather than low variability.

Scaled Boxplot

To enable comparison across variables, features were standardized and visualized again.

After scaling, all variables are centered around zero, allowing comparison of relative dispersion.

Price continues to show high variability, with a wide spread and multiple extreme outliers, indicating strong relative variation even after normalization.

Weight also shows considerable spread and several outliers, suggesting meaningful relative variability.

RAM appears highly compressed around zero, with an extremely small interquartile range. This indicates that most laptops have very similar RAM values, and only a few observations deviate substantially from the mean.

Screen size and CPU frequency remain tightly clustered, confirming low relative variability across the dataset.

The scaled boxplot highlights that RAM, screen size, and CPU frequency vary little relative to their means, while price and weight exhibit greater relative dispersion.

Key Observations

Unscaled boxplots are effective for identifying outliers and interpreting values in real-world units.

Scaled boxplots reveal relative variability across features.

Price shows the highest variability in both absolute and relative terms.

RAM is highly standardized across most laptops, with only a few high-end deviations.

Screen size and CPU frequency are the most consistent features in the dataset.

The unscaled and scaled boxplots provide complementary perspectives. The unscaled visualization preserves real-world interpretability, while the scaled visualization clarifies relative variability across features. Together, they indicate that laptop prices and weights vary substantially, whereas RAM, screen size, and CPU frequency are largely standardized across models.

Covariance Matrix

Variable	*CPU_core*	*Screen_Size_cm*	*CPU_frequency*	*RAM_GB*	*Weight_kg*	*Price*
Screen_Size_cm	0.0876	8.8282	−0.0086	−0.0352	1.2142	−224.6427
CPU_frequency	0.1225	−0.0086	0.1674	0.2328	0.0132	91.3989
RAM_GB	1.4500	−0.0352	0.2328	6.0354	0.0703	809.6785
Weight_kg	0.0429	1.2142	0.0132	0.0703	0.2451	−11.3722
Price	338.8151	−224.6427	91.3989	809.6785	−11.3722	329743.1400

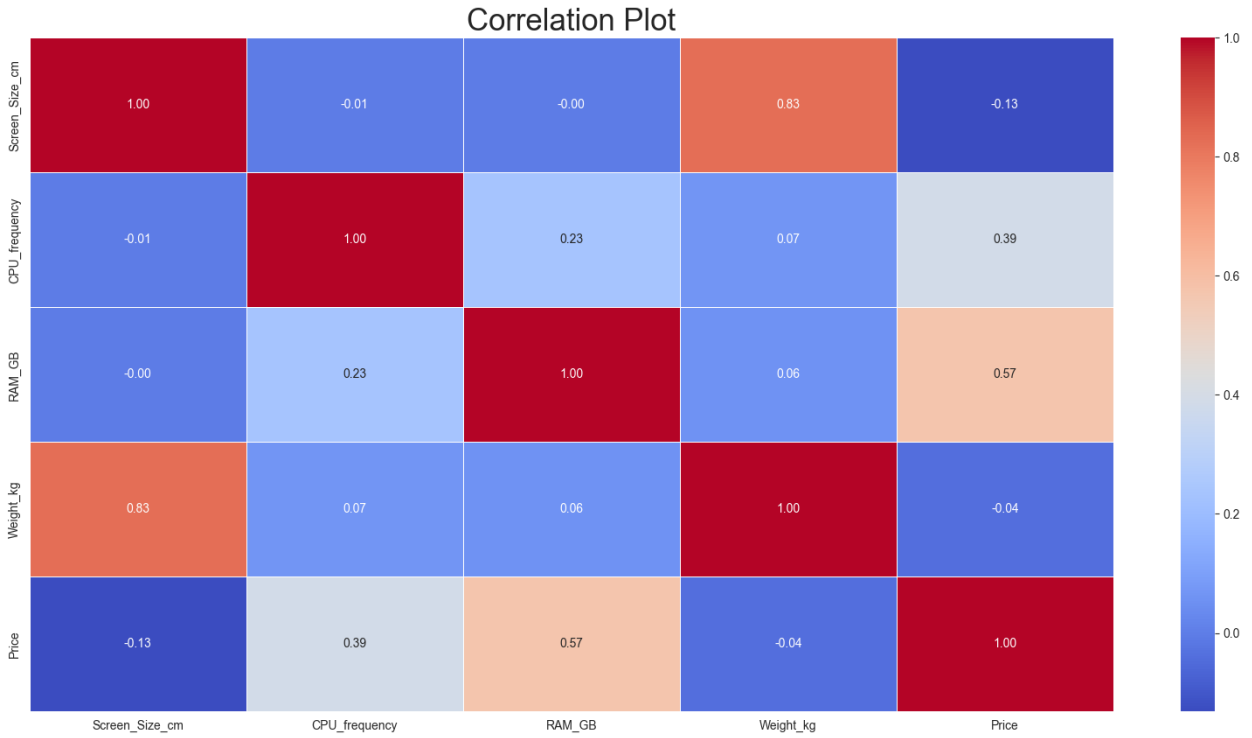
The covariance matrix summarizes the joint variability among the numerical variables: *CPU_core, Screen_Size_cm, CPU_frequency, RAM_GB, Weight_kg,* and *Price*. The diagonal elements represent the variances of individual variables, which differ substantially across features, reflecting the fact that these variables are measured on different scales and units.

The off-diagonal elements indicate the direction and magnitude of pairwise covariation. Positive covariance values suggest that two variables tend to increase together, while negative values indicate that one variable tends to increase as the other decreases. For example, *RAM_GB* and *Price* exhibit a strong positive covariance, indicating that higher memory configurations are generally associated with higher prices. In contrast, *Screen_Size_cm* and *Price* show a negative covariance, suggesting that larger screen sizes are not necessarily associated with higher prices in this dataset.

However, because covariance is scale-dependent, its magnitude is influenced by the units and variability of the underlying variables. As a result, covariance values cannot be directly compared across different variable pairs to assess relative strength of association. For this reason, while the covariance matrix is useful for understanding the direction of joint variability, it is limited in its interpretability.

To enable meaningful comparison of linear relationships across variables, the correlation matrix is subsequently examined. By standardizing variables to a common scale, the correlation matrix provides a clearer and more interpretable measure of association strength, making it more suitable for feature analysis and selection.

Correlation Analysis



The correlation heatmap summarizes the pairwise linear relationships among the numerical variables: The correlation heatmap summarizes the pairwise linear relationships among the numerical variables: *Screen_Size_cm*, *CPU_frequency*, *RAM_GB*, *Weight_kg*, and *Price*. By standardizing covariances, the correlation matrix allows direct comparison of relationship strengths across variables.

Laptop price exhibits its strongest positive correlation with RAM_GB ($r \approx 0.57$), indicating that higher memory capacity is consistently associated with higher-priced laptops. A moderate positive correlation is also observed between Price and CPU_frequency ($r \approx 0.39$), suggesting that processor performance plays a meaningful, though not exclusive, role in determining price. In contrast, Price shows a near-zero correlation with Weight_kg ($r \approx -0.04$), implying that laptop weight has little to no direct linear influence on pricing.

Among the physical attributes, Screen_Size_cm and Weight_kg demonstrate a strong positive correlation ($r \approx 0.83$), reflecting an intuitive design relationship: laptops with larger screens tend to be heavier. However, screen size itself shows negligible correlation with price and with most other performance-related features, indicating a limited direct role in price determination.

Overall, the correlation structure suggests that performance-related specifications, particularly memory capacity and processor speed, are the primary quantitative drivers of laptop price in this dataset. Physical characteristics such as screen size and weight contribute less directly to pricing. Additionally, the absence of extremely high correlations among most predictors indicates limited multicollinearity, supporting the suitability of these variables for joint inclusion in regression or predictive modeling frameworks.

Conclusion

This project conducted a comprehensive exploratory data analysis of the Laptop Pricing dataset to understand its structure, feature behavior, and the key factors influencing laptop prices. The analysis revealed that the dataset reflects a highly standardized and imbalanced market, dominated by mid-range laptops with similar hardware configurations, alongside a smaller number of premium and entry-level systems.

Price exhibits substantial variability and a pronounced right-skewed distribution, driven by a limited number of high-end laptops with premium specifications. In contrast, several hardware features—particularly RAM capacity, CPU frequency, and screen size—show limited variability, indicating strong standardization in commonly sold laptop configurations. This suggests that price differences arise not from gradual changes in single specifications but from discrete jumps between predefined hardware tiers.

Performance-related attributes emerge as the most influential determinants of price. RAM_GB shows the strongest positive association with price, followed by CPU_frequency and CPU_core, confirming that computational capability plays a central role in pricing. GPU type and SSD storage capacity also demonstrate clear price stratification, with higher-end configurations consistently associated with higher prices across manufacturers. In contrast, physical characteristics such as screen size and weight exhibit weak or negligible direct relationships with price, despite showing strong interrelationships with each other.

Categorical factors—including manufacturer, laptop category, GPU type, operating system, and CPU tier—introduce substantial systematic price differences. Premium brands and categories consistently occupy higher price ranges, while budget-oriented brands cluster at lower price levels. At the same time, wide price dispersion within several brands highlights the combined influence of multiple specifications and market positioning rather than any single feature.

The dataset is characterized by notable class imbalance across categories, brands, operating systems, and hardware tiers. Notebooks, Windows-based systems, Intel GPUs, 8 GB RAM, and 256 GB SSD configurations dominate the data. As a result, statistical summaries and predictive models derived from this dataset are expected to perform best for these common configurations, while inferences for underrepresented groups—such as Linux systems, low-RAM devices, or rare brands—should be interpreted with caution.

Overall, the exploratory analysis provides a clear understanding of the dataset's pricing dynamics, structural limitations, and feature relationships. The findings establish a strong foundation for subsequent regression or machine learning models, while emphasizing the importance of accounting for class imbalance, discrete feature structures, heteroscedasticity, and non-uniform variability in any further statistical or predictive analysis.
