Problem 1

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Situation of the column** | **Cleaning action/steps** | **Justification/Explanation** |
| ALL COLUMNS | Mixed case values. | Convert to lowercase | More standardised. Improves duplicate detection |
| model | Marketing terminology | Remove instances from model | Not meaningful.  Doesn’t help identify laptops |
| Extra data from other columns | Extract into correct column | Enhances clarity of data |
| Rows contain brand | Remove brand from model | Data duplication |
| Missing values | Drop rows | Cannot identify laptop with absolute certainty without model. |
| brand | Impute brand based on model | Brand & model are intrinsically related – Less missing data |
| graphics | Values other than integrated, dedicated | Move other values into graphics\_coprocessor column | graphics column = binary classification |
| ram | Unstandardised numeric values | Round & convert to consistent unit | Easier to compare & read |
| harddisk |
| cpu\_speed | Convert to consistent unit. |
| brand | Unstandardised, Syntax Errors, Trailing & Leading whitespace | Pattern match (RegEx)  Map semantically identical values to common format  Strip whitespace | Reduces number of unique values.  Improves comparisons, duplicate detection. |
| model |
| color |
| OS |
| cpu |
| special\_features |
| graphics\_coprocessor |
| Wgraphics | Trailing & Leading whitespace | Strip whitespace |
| special\_features | Identical but shuffled rows | Convert to set, then sort | Improves duplicate row detection |
| cpu | Unstructured, complex data | Extract into cpu\_brand, cpu\_series, cpu\_model | Granular data easier to analyse |
| graphics | Extract into graphics\_brand, graphics\_details |
| Missing values | Backfill/Impute values from graphics\_details | Less missing data |
| cpu | Empty | Drop columns | Data was extracted |
| grpahics\_coprocessor |
| screen\_size | Non-standard column names (no units) | screen\_size\_inches | Standardised, meaningful names are clearer (Sundaramurugan, 2022) |
| color | colour |
| harddisk | harddisk\_gb |
| ram | ram\_gb |
| cpu\_speed | cpu\_speed\_ghz |
| price | price\_usd |
| OS | os |
| brand | Type inconsistency | New type = **str** | Columns were Objects, containing various datatypes - Not consistent or accurate. |
| model |
| colour |
| cpu\_series |
| cpu\_model |
| os |
| special\_features |
| graphics |
| graphics\_brand |
| graphics\_details |
| harddisk\_gb | New type = **Int64** |
| ram\_gb |
| screen\_size\_inches | New type = **float** |
| rating |
| price\_usd |
| cpu\_speed\_ghz | Mostly empty | Drop column (Ngugi, 2022) | Over 88% missing (Figure 1) |
| rating | None | Not over 80% missing, and no valid reason (Ngugi, 2022) |
| special\_features |
| colour |
| ALL COLUMNS | Duplicates | Drop duplicate rows | Duplicates skew data (Dhar, 2023) |
| brand | Too many groups  (Figure 2) | Less frequent become ‘OTHER’ | Less groups = better visualisation readability  (Figure 3) |
| colour |
| cpu\_brand |
| os |
| graphics\_brand |
| cpu\_series | Bin using cpu\_brand |
| harddisk\_gb | Bin using ranges |
| ram\_gb | None | Values too small |
| harddisk\_gb | Outliers | None | Outlier values don’t imply erroneous/false values (Elgabry, 2019) |
| ram\_gb |
| rating |
| price |
| screen\_size\_inches | Remove screen-sizes above 21inches  (Figure 4) | Largest laptop screen-size is 21inches (Levin, 2022) |

Problem 2

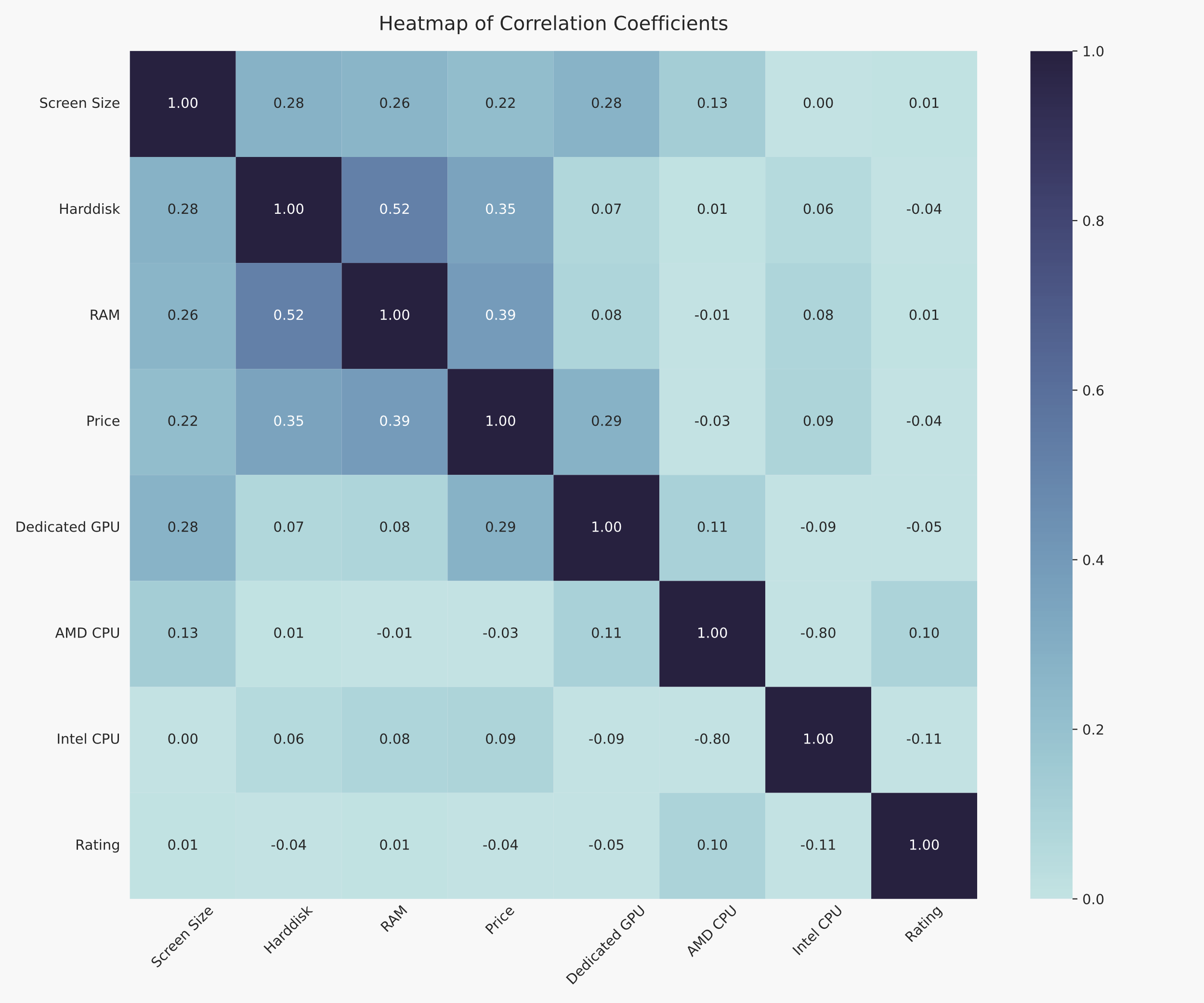
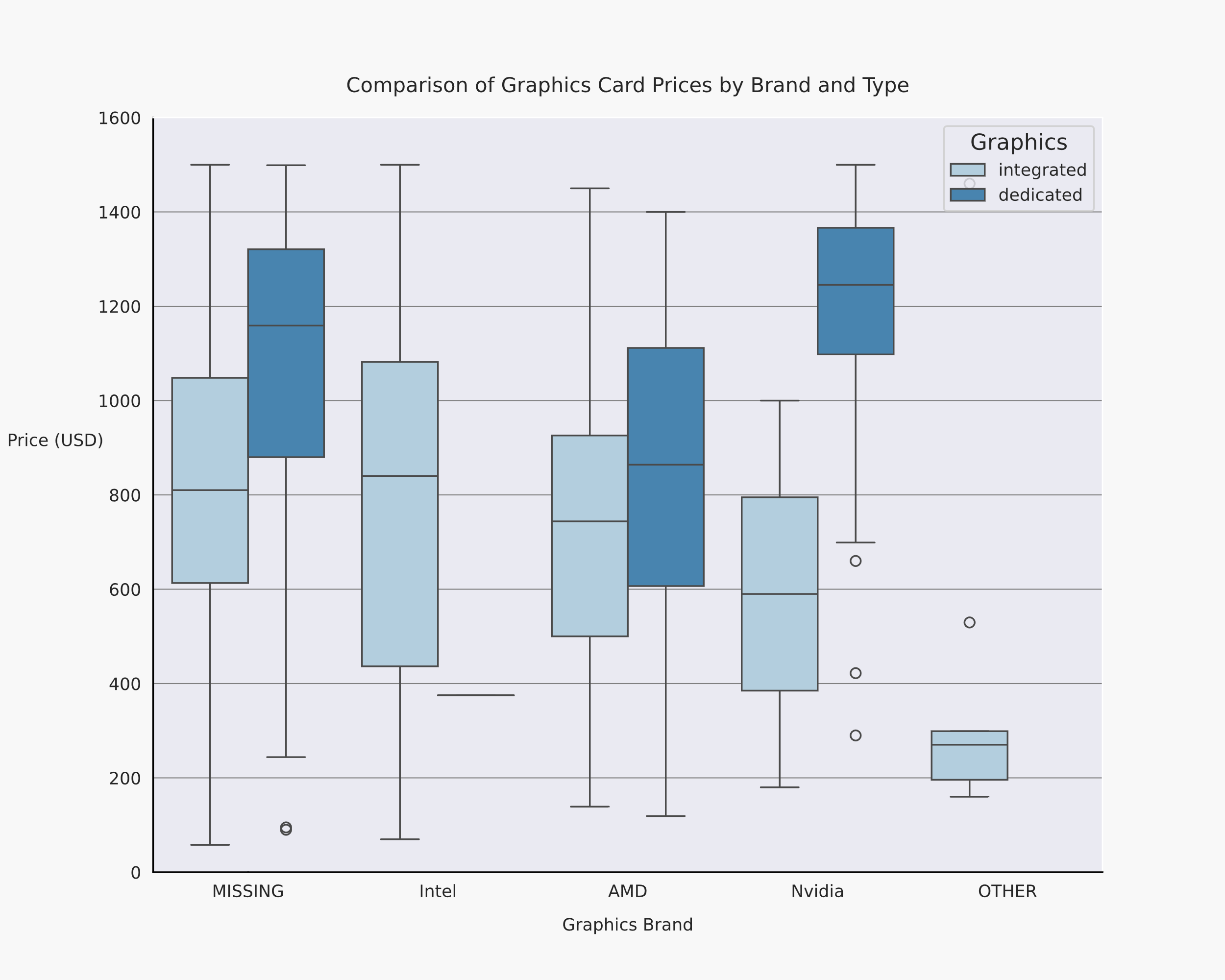


Figure 5: Pair-wise Pearson Correlation Coefficient Matrix



Customer 1: Video Editor

* Large screen for multitasking
* Sufficient RAM and Storage
* Powerful GPU

Filtering down to screen sizes greater than 15” (Microsoft, 2023)

Customer 2: Travel Photographer

* Lightweight
* Lots of storage
* Long battery life

>2TB, 13” to 15”, integrated only

# Figures

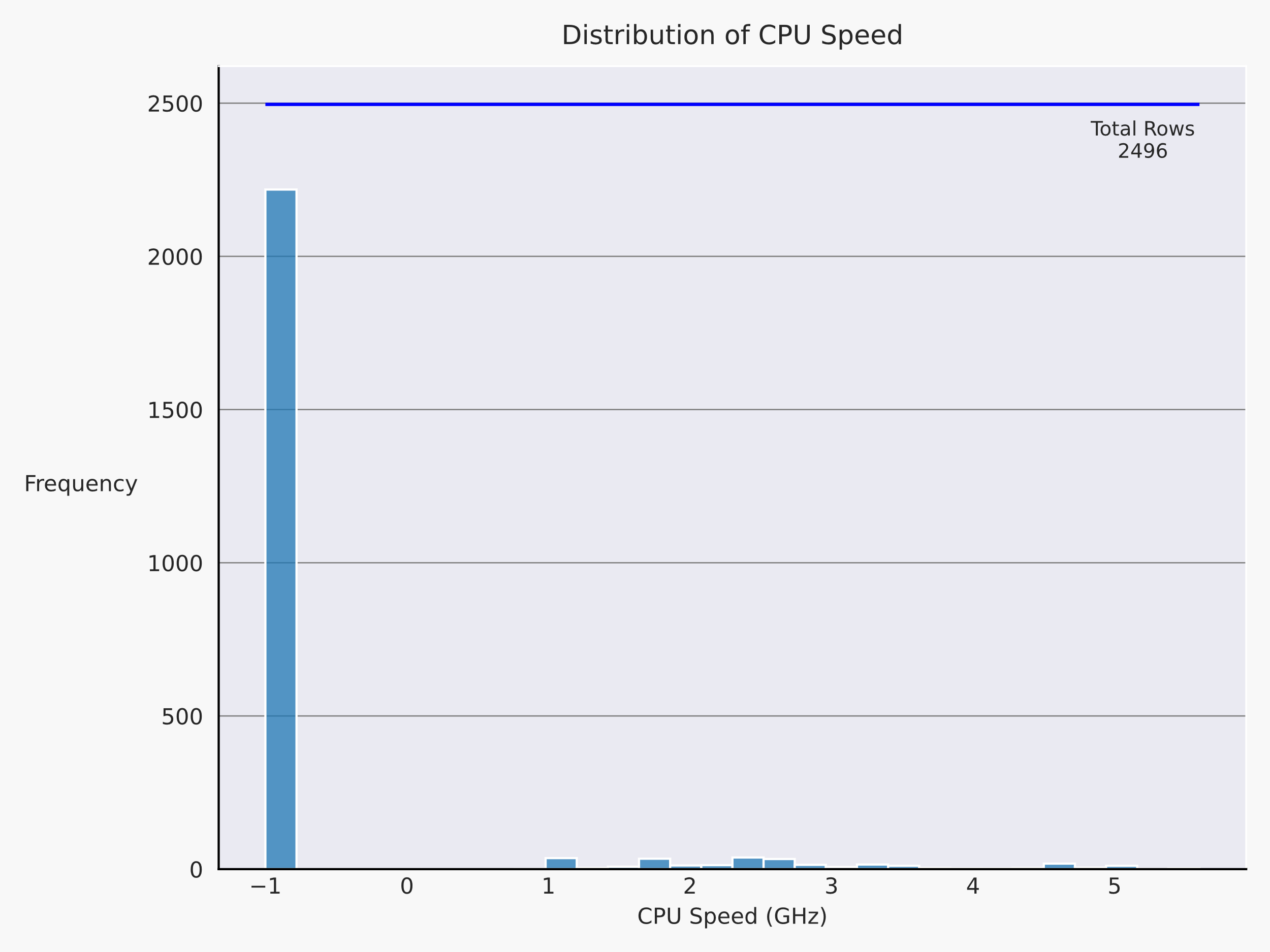


Figure 1: Distribution of CPU Speed

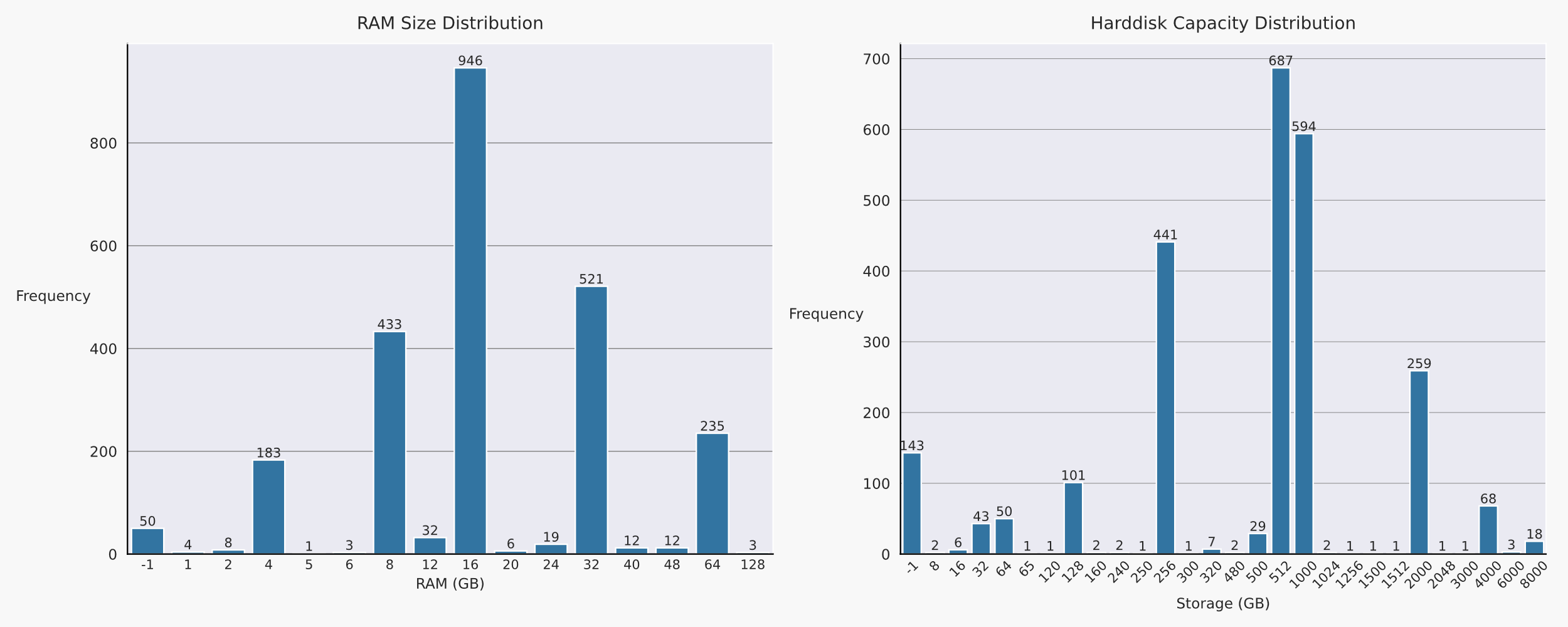


Figure 2: RAM and Harddisk value distributions

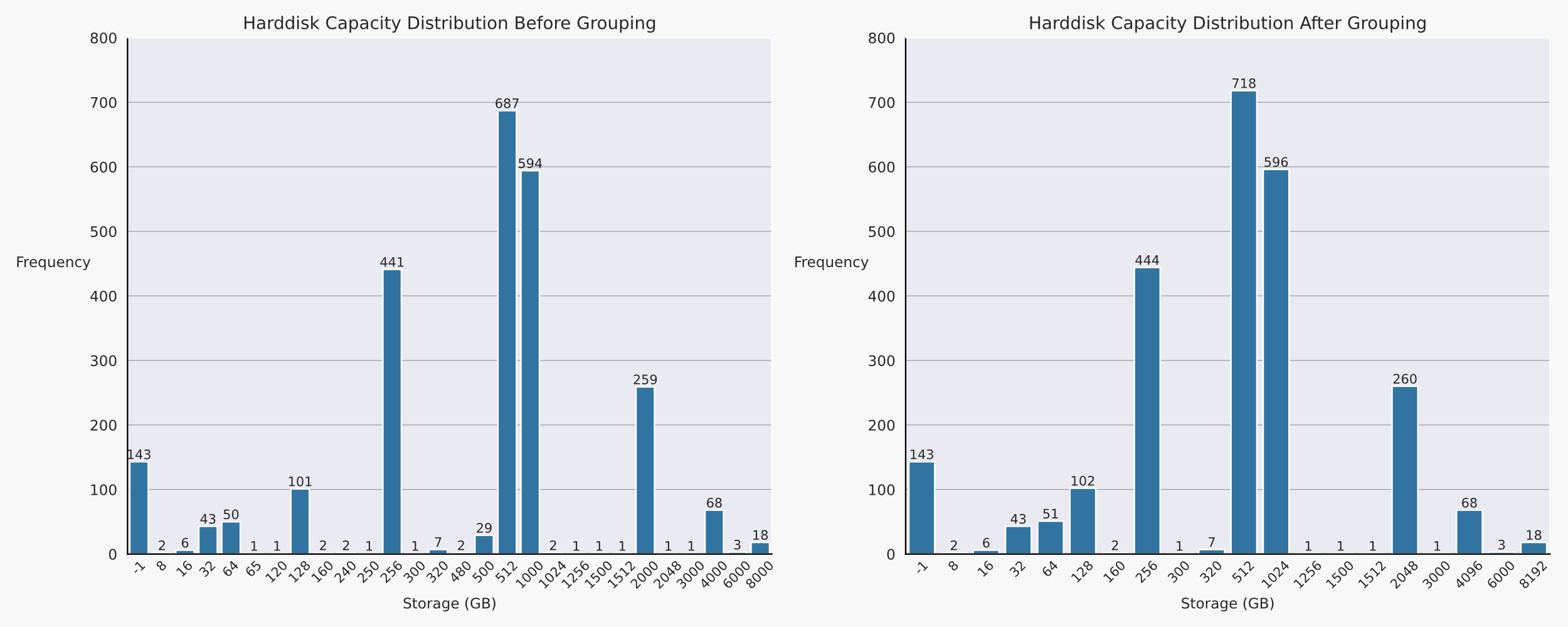


Figure 3: Effects of grouping Harddisk capacity

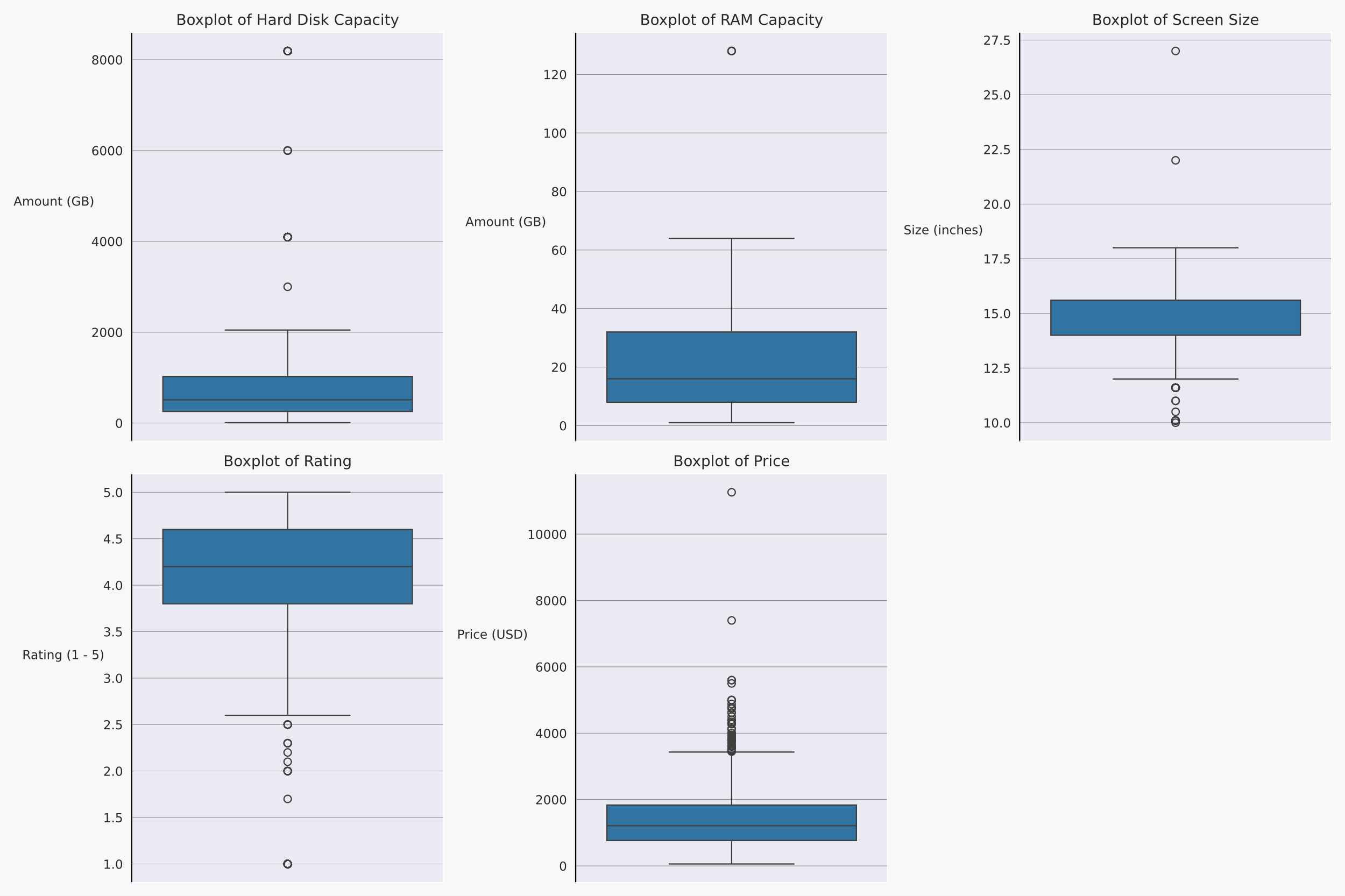


Figure 4: Boxplots of all numerical columns

# Bibliography

Dhar, A. (2023, October 6). *A Comprehensive Guide to Data Cleaning Techniques.* Retrieved from Medium: https://medium.com/@abhishiktadhar111/a-comprehensive-guide-to-data-cleaning-techniques-ebb2659c89a7

Donadi, C. (2022, July 18). *How Much Photo Storage Do You Need for Photography*. Retrieved from chrissydonadi: https://chrissydonadi.com/how-much-photo-storage-do-you-need-for-photography/

Elgabry, O. (2019, February 28). *The Ultimate Guide to Data Cleaning.* Retrieved from Medium: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

HP. (2023). *HP Tech Takes: Integrated vs Dedicated Graphics Cards; How to Choose the Best GPU*. (D. Horowitz, Editor) Retrieved from hp.com: https://www.hp.com/ca-en/shop/offer.aspx?p=integrated-vs-dedicated-graphics-cards

Kingston Technology. (2021, March). *How much memory do you need for video editing?* Retrieved from Kingston: https://www.kingston.com/unitedkingdom/en/blog/pc-performance/how-much-memory-needed-for-video-editing

Levin, N. (2022, November 23). *8 Largest Laptops By Screen Size.* Retrieved from Largest.org: https://largest.org/technology/laptops-by-screen-size/

Microsoft. (2023, May 30). *How to choose the best laptop screen size*. Retrieved from Microsoft: https://www.microsoft.com/en-us/surface/do-more-with-surface/how-to-choose-the-best-laptop-screen-size

Ngugi, J. (2022, May 21). *Handling Missing Values - Data Science.* Retrieved from Medium: https://medium.com/mlearning-ai/handling-missing-values-data-science-7b8e302264ee

Sundaramurugan, S. (2022, May 25). *Five Golden Rules for Cleaning Data in Power BI.* Retrieved from Medium: https://swaathi317.medium.com/five-golden-rules-for-cleaning-data-in-power-bi-a50ed37dda54