Word Count: 750

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-5)
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-6)	Less frequent become 'OTHER'	Less groups = better visualisation readability (Figure-7)
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-8)	Largest laptop screen-size is 21inches (Levin, 2022)

Problem 2

Customer 1: Video Editor

- Large screen for multitasking
- Sufficient RAM and Storage
- Powerful CPU
- Dedicated GPU Preferred

Customer 2: Travel Photographer

- Lightweight
- Lots of storage
- Long battery life

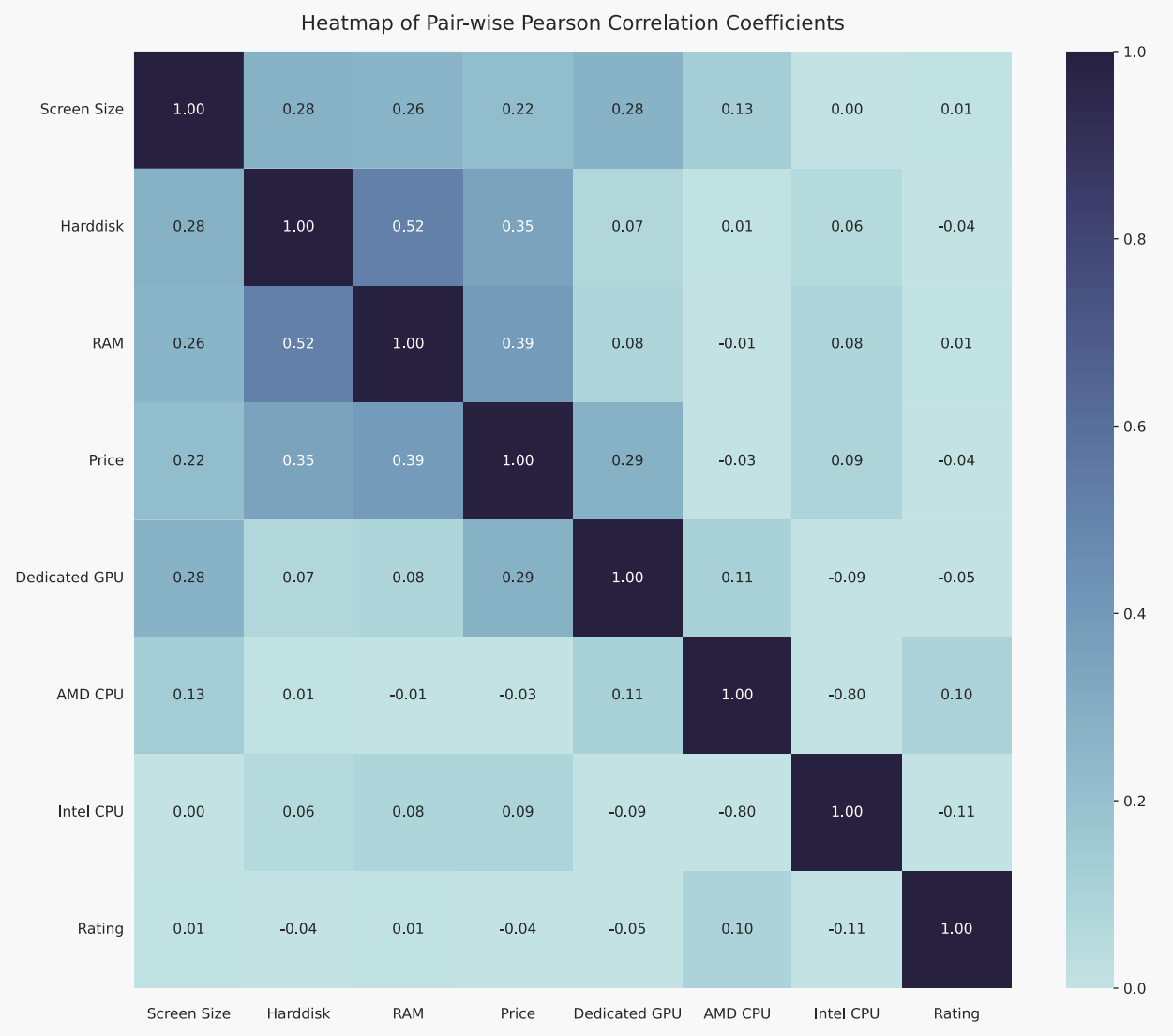

Figure-1: Pearson Correlation Coefficient Matrix

From *Figure-1*, it is clear there are many correlations within the data.

Most notably, RAM and Harddisk have a strong correlation (0.52), and both have a moderate correlation (0.39 and 0.35 respectively) with price. Unsurprisingly, screen size and dedicated GPU have a moderate correlation (0.22 and 0.29 respectively) with price.

Most interesting is the correlation between Dedicated GPU and screen size. This is likely because dedicated GPUs require more cooling and space in the laptop, and therefore have a bigger screen. For customer 2, this suggests they may need an integrated GPU as it would likely have a smaller screen, and hence weight less. This, together with the fact that integrated GPUs consume less power (HP, 2023), implies a bigger battery life.

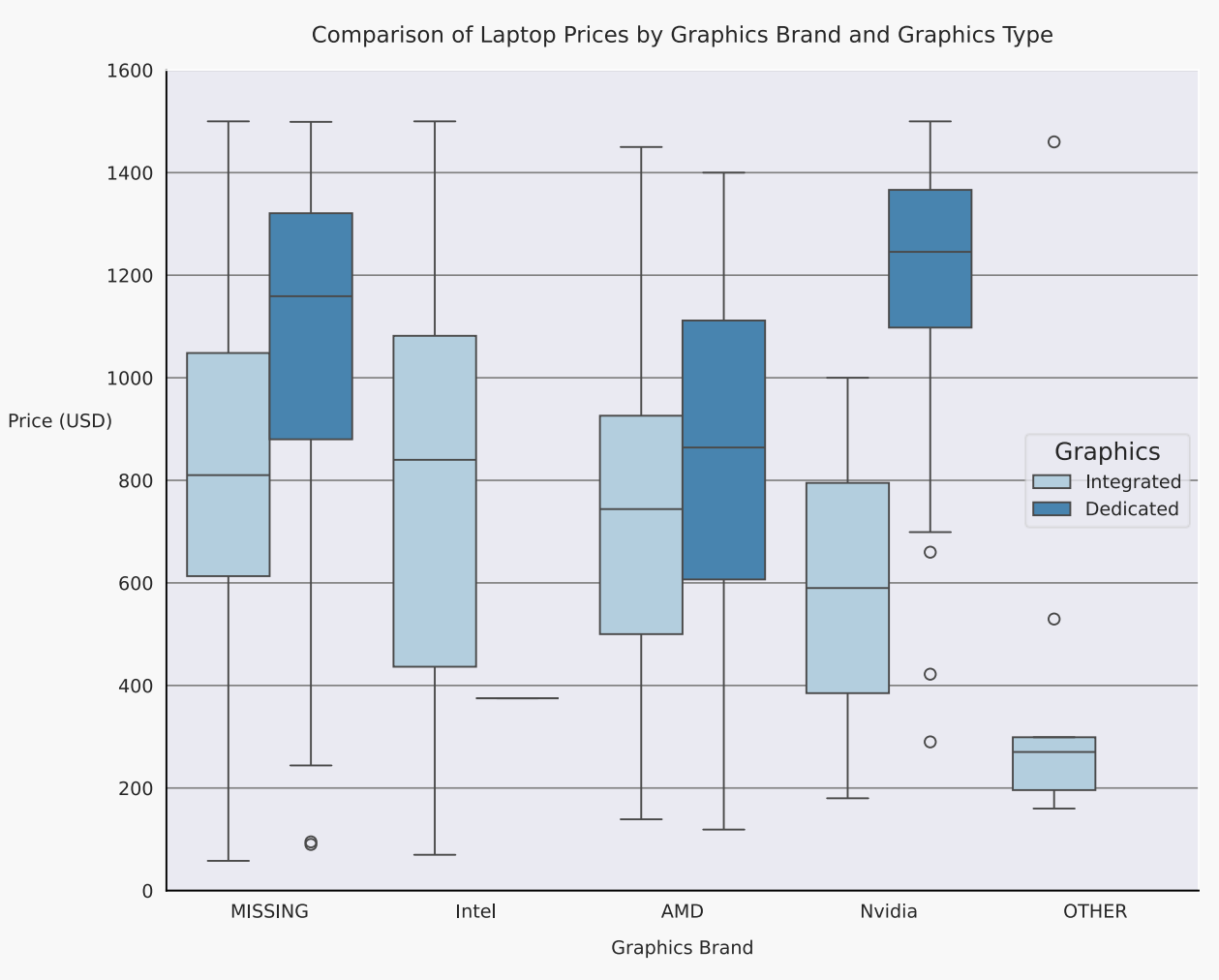

Figure-2: Laptop Prices grouped by Graphics Brand & Type

While not all brands make both dedicated and integrated, the general trend is that laptops with dedicated GPUs are more expensive than integrated. Additionally, Figure-2 shows that laptops with dedicated Nvidia GPUs are more expensive than laptops with dedicated AMD GPUs, and all other GPUs.

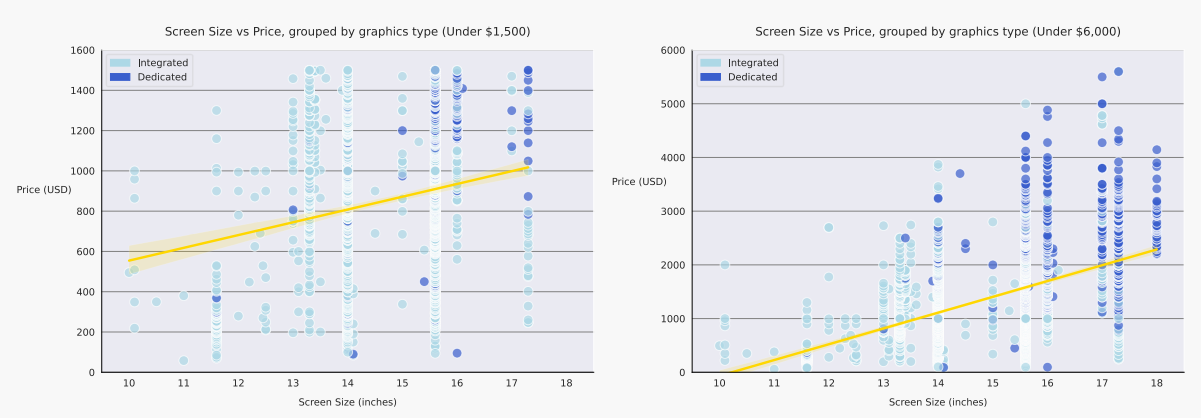

Figure-3: Screen Size vs Price grouped by graphics.

Figure-3 also shows that dedicated GPUs tend to have larger screens. Within the budget, most dedicated GPUs occur above 15.6inches.

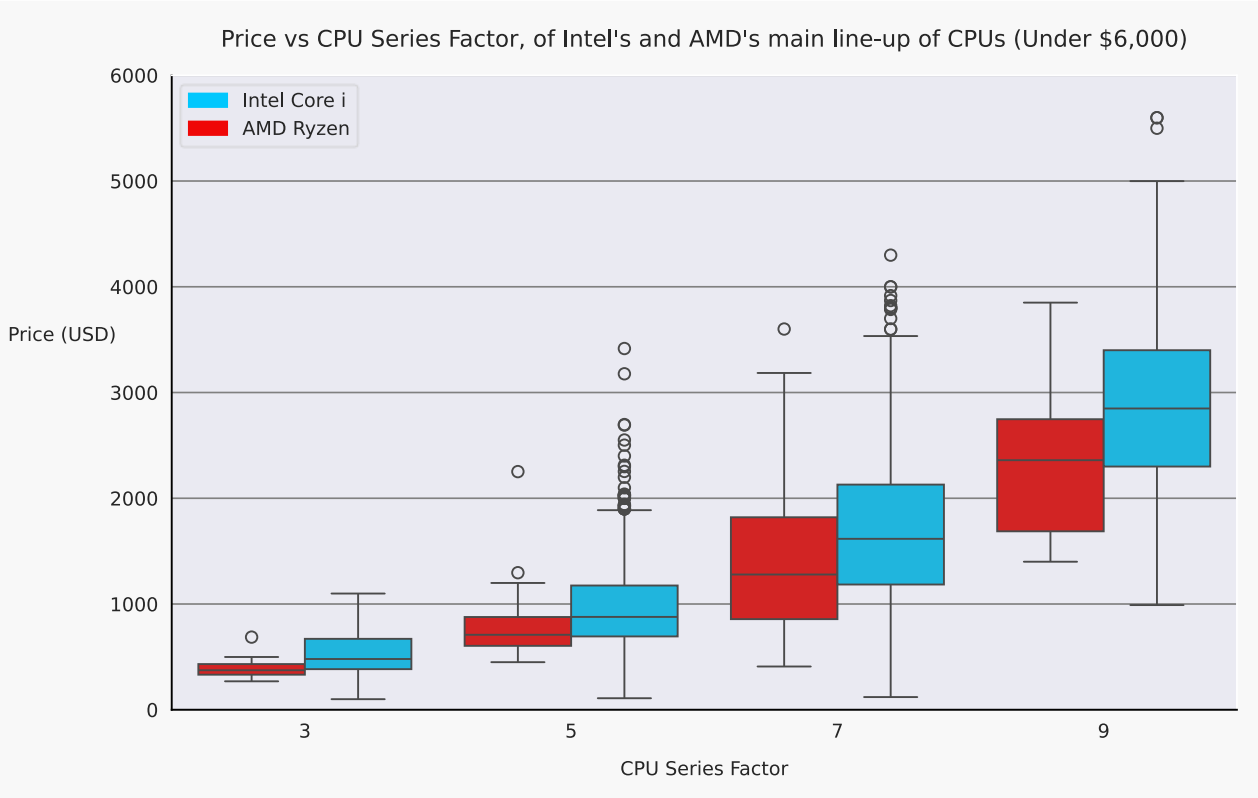

Figure-4: Price vs CPU Series Factor

If we assume CPU series across Intel and AMD are equivalent in performance, an immediate trend is visible, a higher series factor means a higher price. Additionally, at every series factor, the median price of Intel CPUs is higher than AMD CPUs – this matches with the *Figure-1* correlation.

AMD CPUs also have a weak correlation (0.10) with rating (Figure-1). This suggests AMD CPUs are cheaper and higher rated.

Using this analysis and other sources, the following criteria are chosen:

Customer 1:

- Large screen (>15inches) (Microsoft, 2023)
- >32GB RAM (Kingston Technology, 2021),
- >1TB Harddisk
- AMD CPU
- Dedicated GPU

brand	model	screen_size_inches	colour	harddisk_gb	cpu_brand	cpu_series	cpu_model	ram_gb	os	special_features	graphics	graphics_brand	graphics_details	rating	price_usd
dell	g15 gaming 5525	15.6	grey	1000.0	amd	ryzen 7	_MISSING_	32.0	windows 11 home	anti-glare,backlit keyboard,hd audio,numeric k...	dedicated	_MISSING_	_MISSING_	NaN	1259.00
asus	rog strix g15	15.6	grey	1000.0	amd	ryzen 7	_MISSING_	32.0	windows 11 pro	anti-glare,backlit keyboard,hd audio,numeric k...	dedicated	_MISSING_	_MISSING_	NaN	1279.00
asus	vivobook pro 15	15.6	blue	1000.0	amd	ryzen 9	_MISSING_	32.0	windows 11 home	_MISSING_	dedicated	nvidia	rtx 4060	4.6	1399.99
dell	g15 5525	15.6	grey	2000.0	amd	ryzen 7	_MISSING_	32.0	windows 11 home	_MISSING_	dedicated	nvidia	rtx 3050 ti	NaN	1409.99
dell	g15	15.6	grey	1000.0	amd	r series	_MISSING_	32.0	windows 10 home	backlit keyboard	dedicated	_MISSING_	_MISSING_	4.1	1429.99

Customer 2:

- > 2TB storage (Donadi, 2022)
- Integrated graphics
- > 16GB RAM (Crucial, 2023)
- Maximum Medium Screen (<15inches) (Microsoft, 2023)

brand	model	screen_size_inches	colour	harddisk_gb	cpu_brand	cpu_series	cpu_model	ram_gb	os	special_features	graphics	graphics_brand	graphics_details	rating	price_usd
dell	latitude 7420	14.0	black	2000.0	intel	i7	_MISSING_	16.0	windows 11 home	bluetooth	integrated	_MISSING_	_MISSING_	NaN	880.99
dell	latitude 7320	13.3	black	2000.0	intel	i7	_MISSING_	16.0	windows 11 home	bluetooth	integrated	_MISSING_	_MISSING_	NaN	880.99
dell	latitude 7320	13.3	black	2000.0	intel	i7	_MISSING_	16.0	windows 11 pro	_MISSING_	integrated	_MISSING_	_MISSING_	NaN	881.74
dell	latitude 7420	14.0	black	2000.0	intel	i5	_MISSING_	16.0	windows 10 home	bluetooth	integrated	_MISSING_	_MISSING_	NaN	886.99
dell	latitude 7420	14.0	black	2000.0	intel	i7	_MISSING_	16.0	windows 10 home	bluetooth	integrated	_MISSING_	_MISSING_	NaN	886.99

Figures

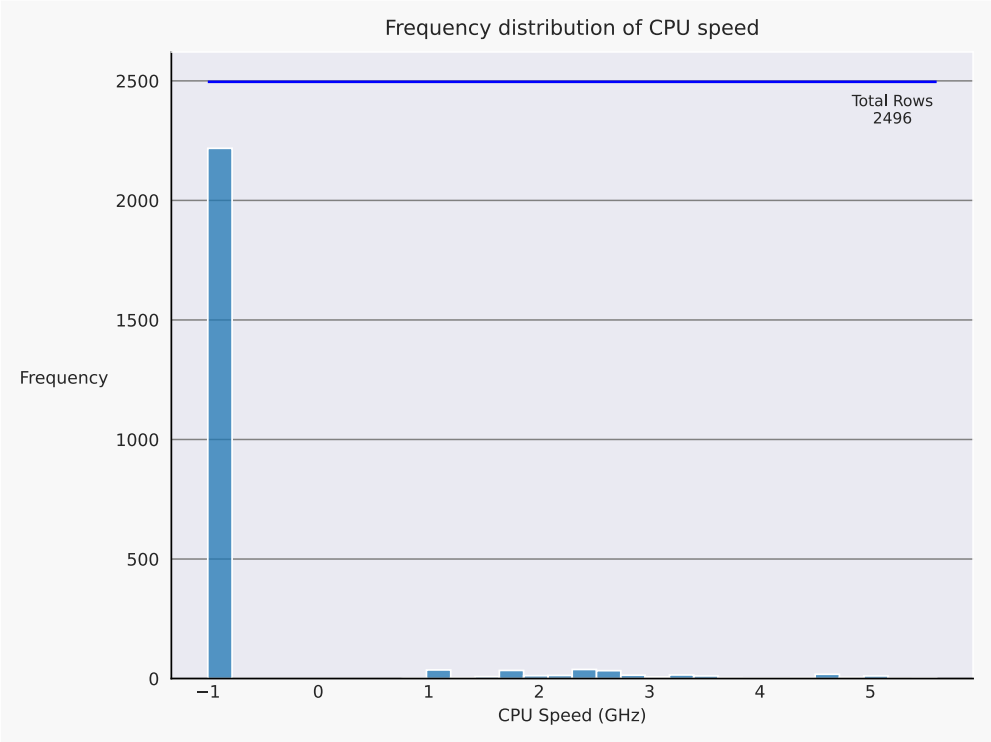

Figure-5: CPU Speed Distribution

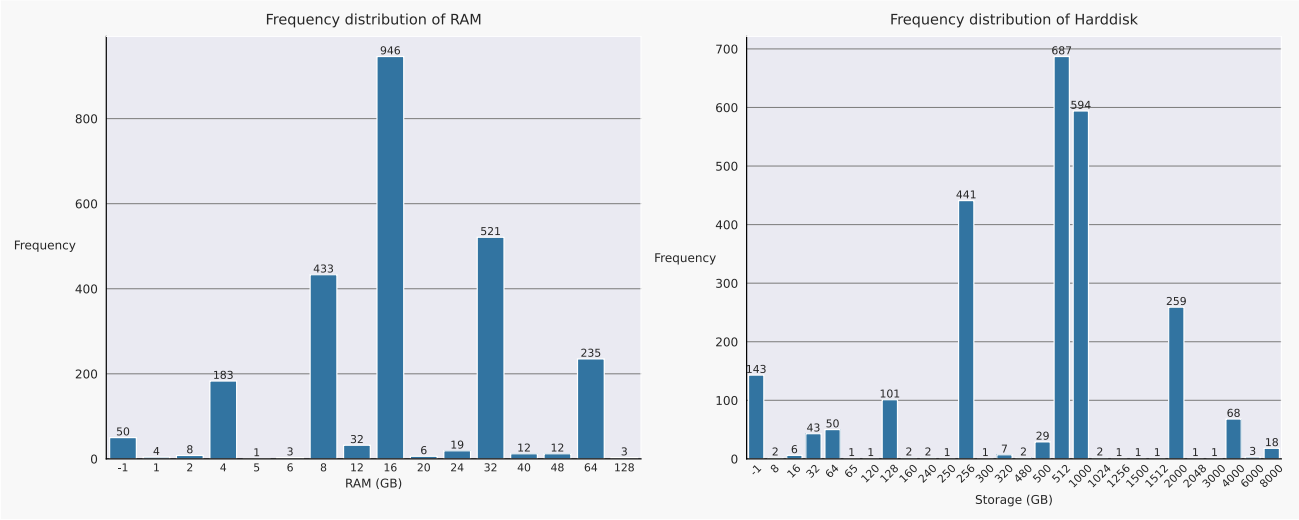

Figure-6: RAM and Harddisk Frequency Distributions

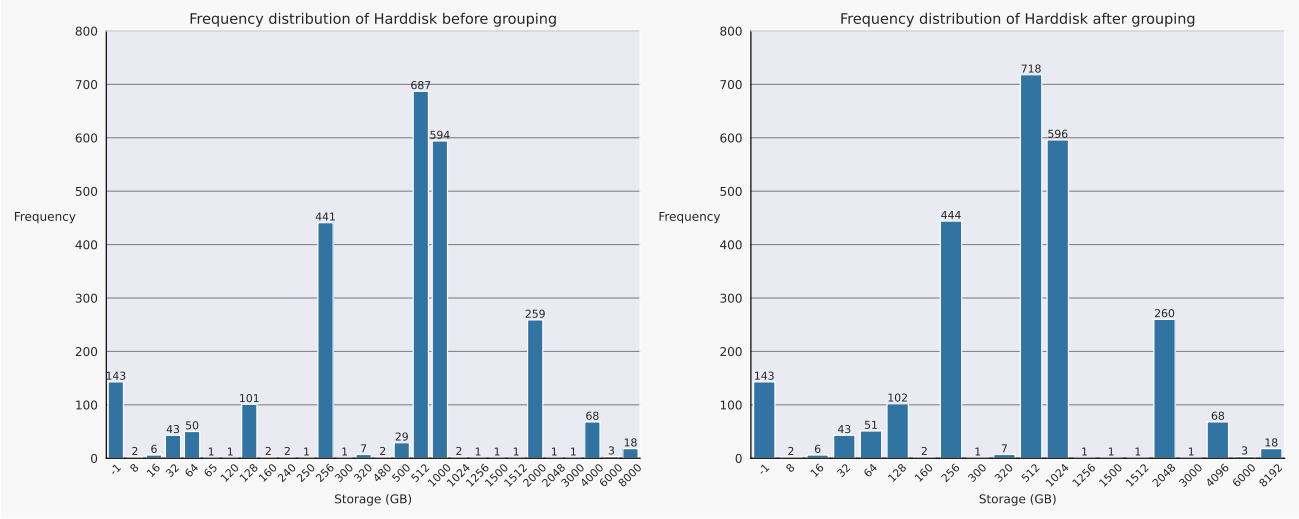

Figure-7: Effects of grouping Harddisk capacity

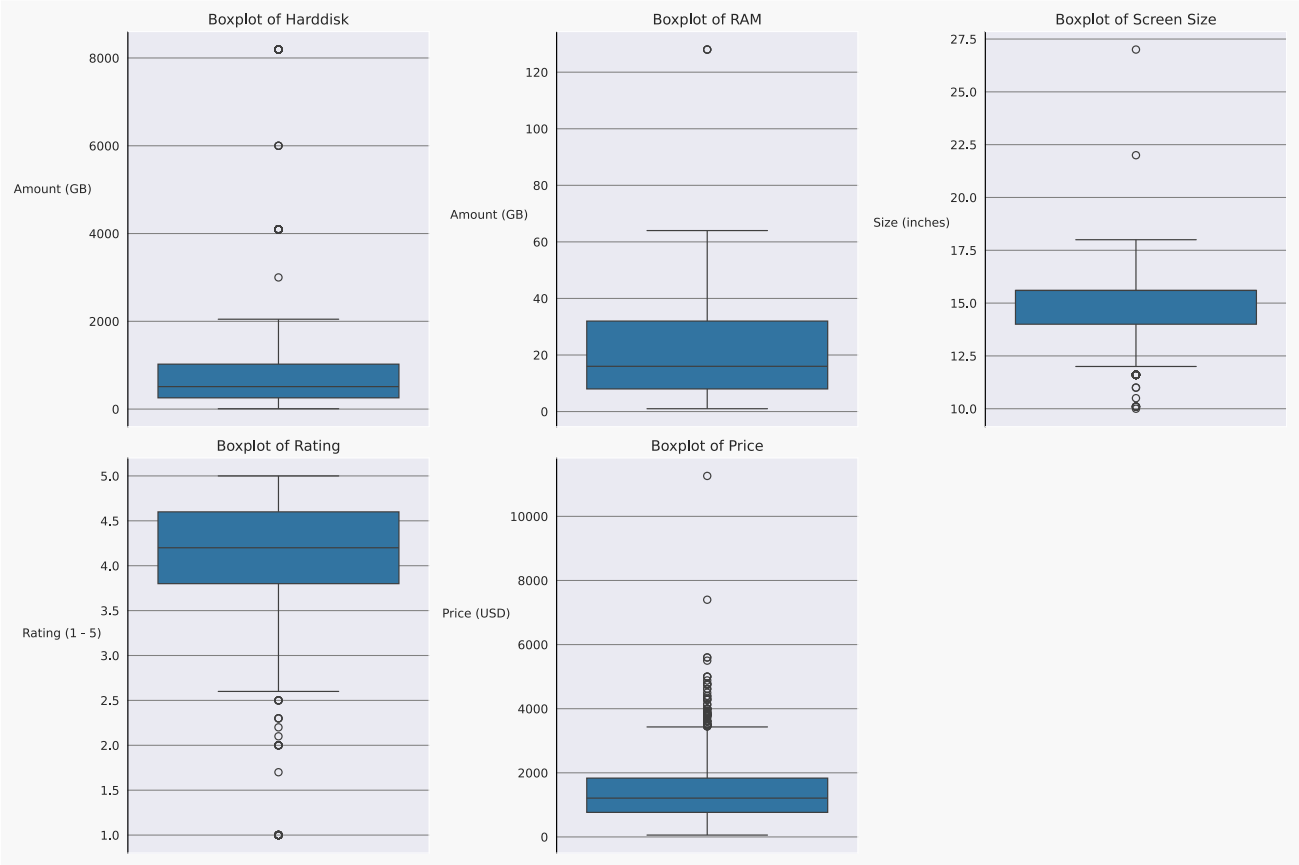

Figure-8: Boxplots of all numerical columns

Bibliography

Crucial. (2023). *How Much RAM Do You Need For Your Computer Memory?* Retrieved from Crucial by Micron: https://uk.crucial.com/articles/about-memory/how-much-ram-does-my-computer-need

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