COMP2271 REPORT(750words)

Problem 1

Column name	Situation of the column	Cleaning actions/steps	Justification/Explanation
brand	Non-brand names, missing values, actual laptop names	Fill missing and non-brand values with 'unknown', replace models with company names.	Ensures brand column consistency and accuracy, aiding in clear brand identification
model	Non-standardized names, brand and model names mixed	Use regex to standardize and remove brand names, unnecessary words/symbols.	Ensures consistency and clarity in model names, simplifying identification and comparison of laptop models
color	non-standardized colors, unseparated colors	Standardized color names using regex to map to generic colors. Separated entries with multiple colors into distinct, standardized color	Enhances the ability to search for and analyze color data making it easier to filter and maintain consistency, in color-based analysis
special_features	Missing values, unseparated features.	Filled missing values using forward and backward fill within the same model group. Standardized features using regex and separated features with a '/'.	Helps in readability, feature-based comparison, and analysis.

Cpu	Non-standardized text, incorrect CPU names	Standardized CPU names using regex patterns, removed irrelevant or incorrect data	Improves data accuracy, ensures user is not misguided by incorrect data. Easier categorization and CPU bases analyses
Cpu_speed_GHz	missing values, speed with units (GHz) as strings, outliers	Filled missing values within CPU groups, extract numerical part, convert to column	Grouping CPU ensures the imputed speeds are consistent with the type of CPU, numerical value of CPU speed is needed for analysis, units are consistent across the dataset.
graphics	Empty values, incorrect entries	Filled missing values, remove specific graphic card models, replace with 'integrated' or 'dedicated'	Clarifies graphics information, distinguishing it from detailed graphics coprocessor data.
graphics_coprocessor	missing values and brand names.	Filled missing values using forward and backward fill within the same CPU group. Created a new column processor_brand.	CPUs often come with associated graphics coprocessors, so this method ensures consistency. Enhances data clarity and make filtering easier.
OS	Non-OS names, unstandardized values, short form (e.g., “win 10”)	Mapped non-OS names to 'unknown', standardize short forms and names using regex	Provides clear and standardized operating system data.
ram	size with unit as a string (e.g., "8 GB").	Extracted the numerical part using regex and converted to a separate numerical column.	The numerical value of RAM is needed for analysis, and units are consistent across the dataset.
harddisk	size with units (GB or TB) as strings.	Extract numerical part, convert to float, standardize sizes to GB.	Maintains unit consistency for comparison and analysis
screen_size	size with unit as a string (e.g., "14 Inches").	Extracted the numerical part using regex, converted to a	Focuses on relevant numerical value for analysis.

		separate screen_size_inches.	
price	Currency symbols, stored as strings	Removed currency symbols, commas using regex, converted to float	Currency symbols and commas are non-numeric characters that need to be removed for mathematical operations. Conversion to float allows for numerical analysis.
Duplicates	1818 duplicate rows.	Removed duplicates.	Prevents skewed analysis and inaccurate conclusions

Problem 2

Customer A (Travel): This customer requires a laptop suitable for traveling. The criteria for filtering include:

- Price under $1500 for affordability.
- Screen size smaller than 15 inches for portability.
- Hard disk space greater than 512 GB for ample storage.

Screen Size Distribution: Shows laptop screen size variety by budget.

RAM Distribution: Represents RAM size options, linking performance and price.

Price Distribution: Focuses on laptop prices, mainly under $1500

	brand	model	color	special_features	cpu	cpu_speed_GHz	graphics	processor_brand	graphics_coprocessor	OS	ram_GB	harddisk_GB	screen_size_inches	price	rating
532	Dell	latitude rugged 5430	Silver	Backlit keyboard	Intel core i5	2.0	Integrated	Intel	Unknown	Windows 11 pro	16.0	1000.0	14.0	1499.99	5.0
519	Hp	envy	Silver	Backlit keyboard / Fingerprint reader	Intel core i7116-5g7	NaN	Integrated	Intel	iris xe	Windows 10 pro	16.0	1024.0	13.3	1499.00	3.8
510	Dell	xps 9320	Gray	Bluetooth / WiFi	Intel core i7	1.0	Integrated	Intel	Unknown	Windows 11 pro	16.0	1000.0	13.4	1494.62	NaN
508	Dell	latitude 7420	Black	Bluetooth / WiFi	Intel core i7	1.0	Integrated	Intel	Unknown	Windows 11 home	32.0	1000.0	14.0	1493.99	NaN
509	Dell	xps 9320	Gray	Bluetooth / WiFi	Intel core i7	1.0	Integrated	Intel	Unknown	Windows 11 home	16.0	1000.0	13.4	1493.99	NaN

Fig 2. For Customer with travelling preferences recommended laptops are shown.

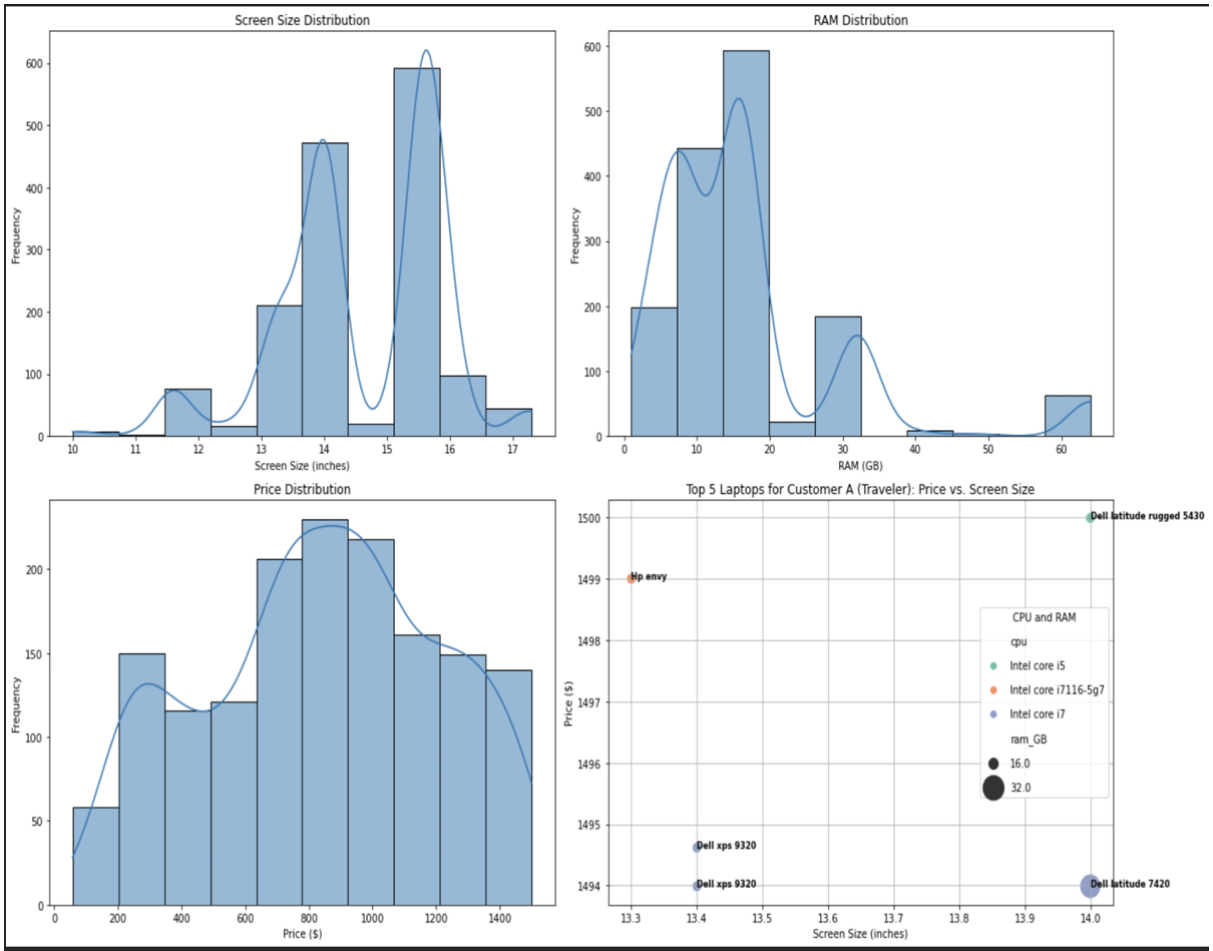


Fig 3. For Customer with travelling preference.

Customer B (Graphic Designer): This customer requires a laptop suitable for graphic designing. The criteria for filtering include:

- **Graphics:** Dedicated graphics card because it is better graphics than integrated.
- **RAM:** Minimum of 16GB as for graphics designing minimum 16GB is recommended.
- **Screen Size:** Preferably larger screens for design work.
- **CPU speed**: Faster processor is required in graphic designing.
- **Harddisk:** Minimum of 512 GB for ample storage as graphic designing task can take much storage.

laptops were sorted by CPU speed, RAM, and user ratings.

	brand	model	color	special_features	cpu	cpu_speed_GHz	graphics	processor_brand	graphics_coprocessor	OS	ram_GB	harddisk_GB	screen_size_inches	price	rating
32	Msi	gf63 thin 10scxr	Black	Anti-glare / Backlit keyboard / HD audio	Intel core family	3200.0	Dedicated	Intel	uhd 620	Windows 10 home	64.0	2048.0	15.6	1029.99	NaN
1712	Dell	precision 7780	Silver	Bluetooth / WiFi	Intel core i9	5.0	Dedicated	Nvidia	quadro rtx 3000	Windows 11 pro	128.0	4000.0	17.0	4736.68	NaN
1542	Dell	x17 r2	Blue	Bluetooth / WiFi	Intel core i9	5.0	Dedicated	Nvidia	geforce rtx 3080	Windows 11 home	64.0	2000.0	17.3	3208.03	5.0
1605	Dell	xps	Silver	Unknown	Intel core family	5.0	Dedicated	Nvidia	geforce rtx 2080 super	Windows 11 pro	64.0	4096.0	17.0	3785.54	5.0
1607	Dell	xps9720	Silver	Unknown	Intel core family	5.0	Dedicated	Nvidia	geforce rtx 2080 super	Windows 10 pro	64.0	4096.0	17.0	3799.00	5.0

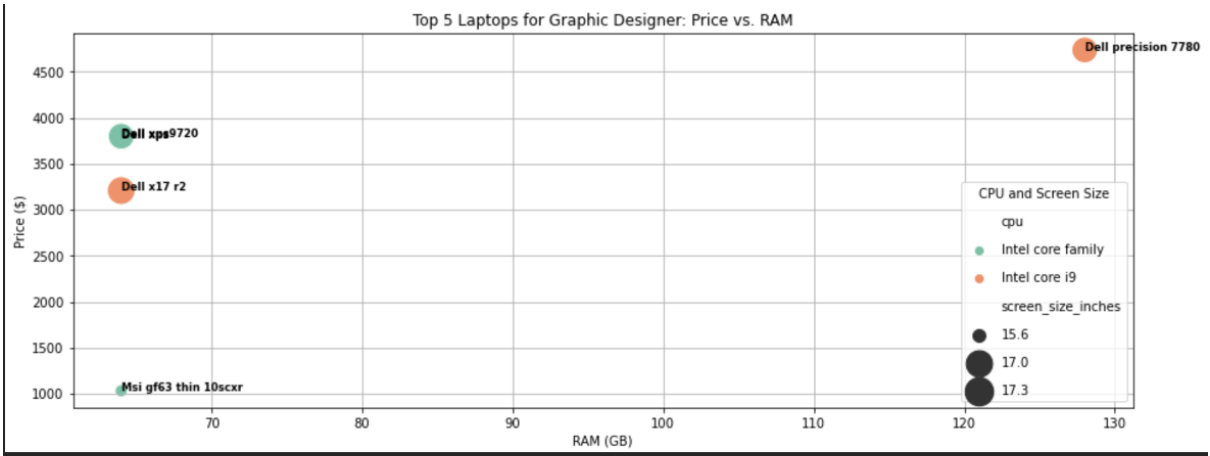


Fig 5. Top 5 laptops for Customer (Graphic designer)

Top 5 Laptops for Graphic Designers:

- This scatter plot correlates RAM with price among the top 5 laptops, color-coded by CPU speed and sized by screen size.
- Indicates a positive correlation between RAM and price, suggests higher performance comes at a higher cost.

RAM Distribution :

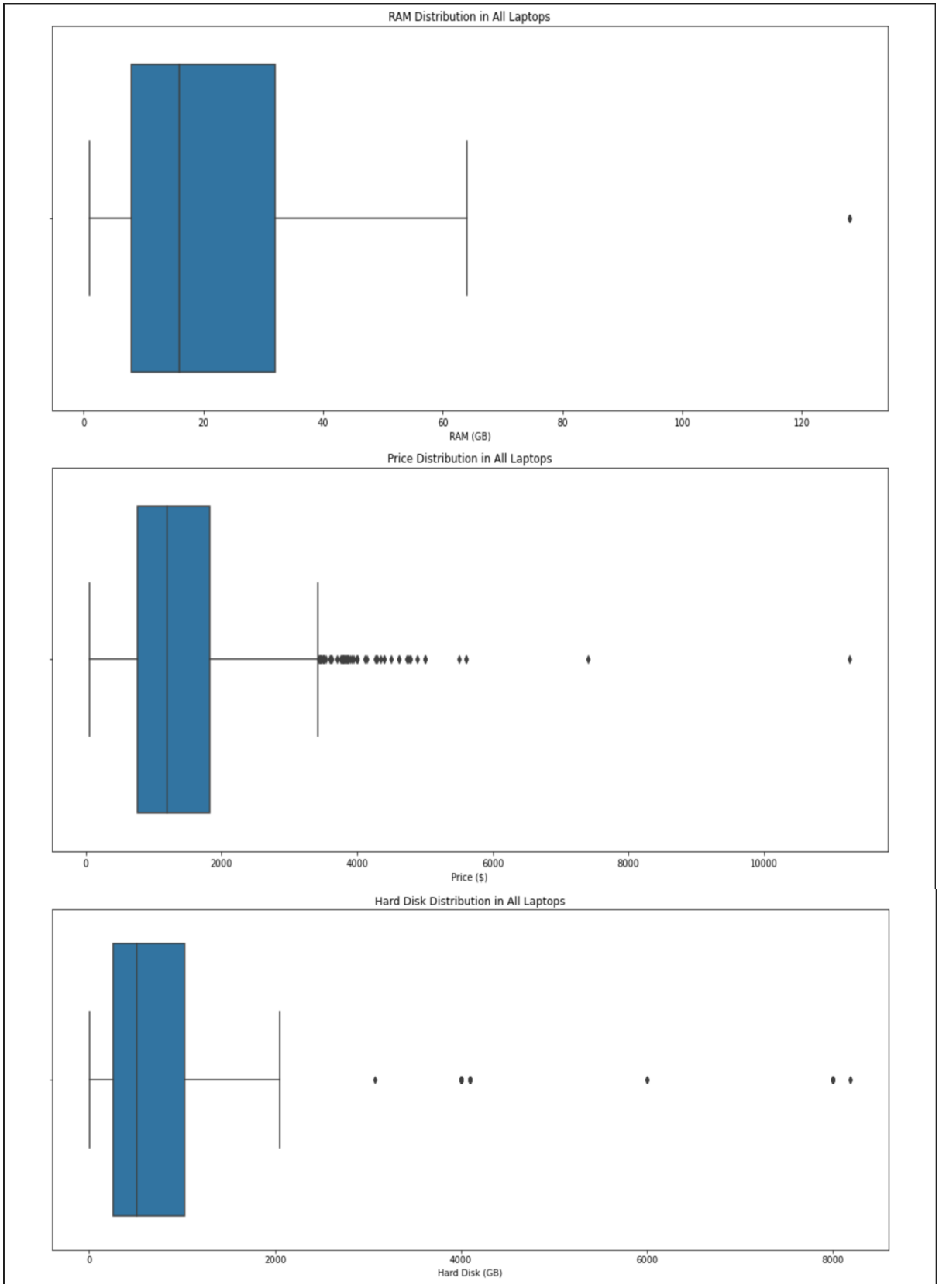
- Box plot all laptops shows a majority clustering around a median value.
- For graphic design, laptops with higher RAM are preferred, as indicated by upper quartile of the distribution.

Price Distribution in:

- The price distribution graph shows a wide range in laptop prices.
- Higher prices often correlate with better specifications, there are affordable options that meet our criteria.

Hard Disk Distribution :

- Shows laptops with larger storage are available, which is crucial for storing large graphic design files.


RAM Distribution in All Laptops
0
20
40
60
80
100
120
RAM (GB)
Price Distribution in All Laptops
0
2000
4000
6000
8000
10000
Price ($)
Hard Disk Distribution in All Laptops
0
2000
4000
6000
8000
Hard Disk (GB)