**DS Report**

Word count: 747

**Problem 1**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Situation of the column** | **Cleaning actions** | **Justiﬁcation** |
| All | Duplicates (before/after cleaning) | Drop duplicates | Remove redundancy |
| Syntax errors, un-standardized, missing data | Extraction, strip, casefold, fill missing with \_MISSING\_ (categorical)/ 0 (numerical) | Make data easier to process, remove corrupted data |
| Numerical | Type inconsistencies | Convert string to float | Numerical should be numbers |
| color | Excess data cell | Split on commas, explode | Laptop with diﬀerent colors considered diﬀerent data |
| special\_features | Split into sorted tuples | Easier to query, ﬁnd duplicates |
| cpu | Split into brand, model | Easier to compare, query via brand |
| graphics\_coprocessor |
| ram | Un-standardized | Round | Maintains correctness |
| harddisk | Convert all to one unit (Techfident, n.d.; Lenovo, n.d.) |
| cpu\_speed |
| graphics | In-record errors, data from other column(s) | Move data to other column, backfill with data from graphics\_coprocessor | Other info can be gotten from respective columns. Maintain correctness |
| model | Add, remove extra data |
| brand |

|  |  |  |  |
| --- | --- | --- | --- |
| brand | Un-standardized, syntax errors | Pattern match, map semantically identical strings, fix grammar | Reduces unique values |
| model |
| color |
| cpu |
| os |
| special\_features |
| graphics\_coprocessor |
| special\_features | In-record errors, missing values | Extract features from “model” | Maintain correctness, reduce missing data |
| graphics\_coprocessor | Fill record from “graphics” |
| model  (before/after cleaning) | Missing values | Drop row | Cannot compare empty models |
| cpu\_speed | Drop column | Almost all data is missing (Fig.3) (Joan, 2022) |
| rating | None | Missing data shows rated laptops have been bought |
| graphics | Fill from observation (Fig.6) | Column should contain “dedicated” or “integrated”. Price of missing closely resembles “integrated” (Fig.7) |
| harddisk | Outliers | Remove outliers (Omar, 2018) | Remove anomaly (Fig.5) |
| screen\_size |
| ram |
| price |

|  |  |  |  |
| --- | --- | --- | --- |
| harddisk | Type inconsistency | Convert to Int64 | Cannot be floats |
| ram |
| graphics | Convert to bool | Easier to visualize as numerical |
| brand | Too many groups, missing data | Group less than 1% frequency into “others” (Raghuvansh, 2020) | Less groups helps visualization (Fig.1-2-4-6) |
| color |
| os |
| cpuBrand |
| gpuBrand |
| harddisk | Too many groups | Bin values (Seagate, n.d.) |
| All | Visually messy | Move unit to column name, rename, reorder | Easier to compare, read |

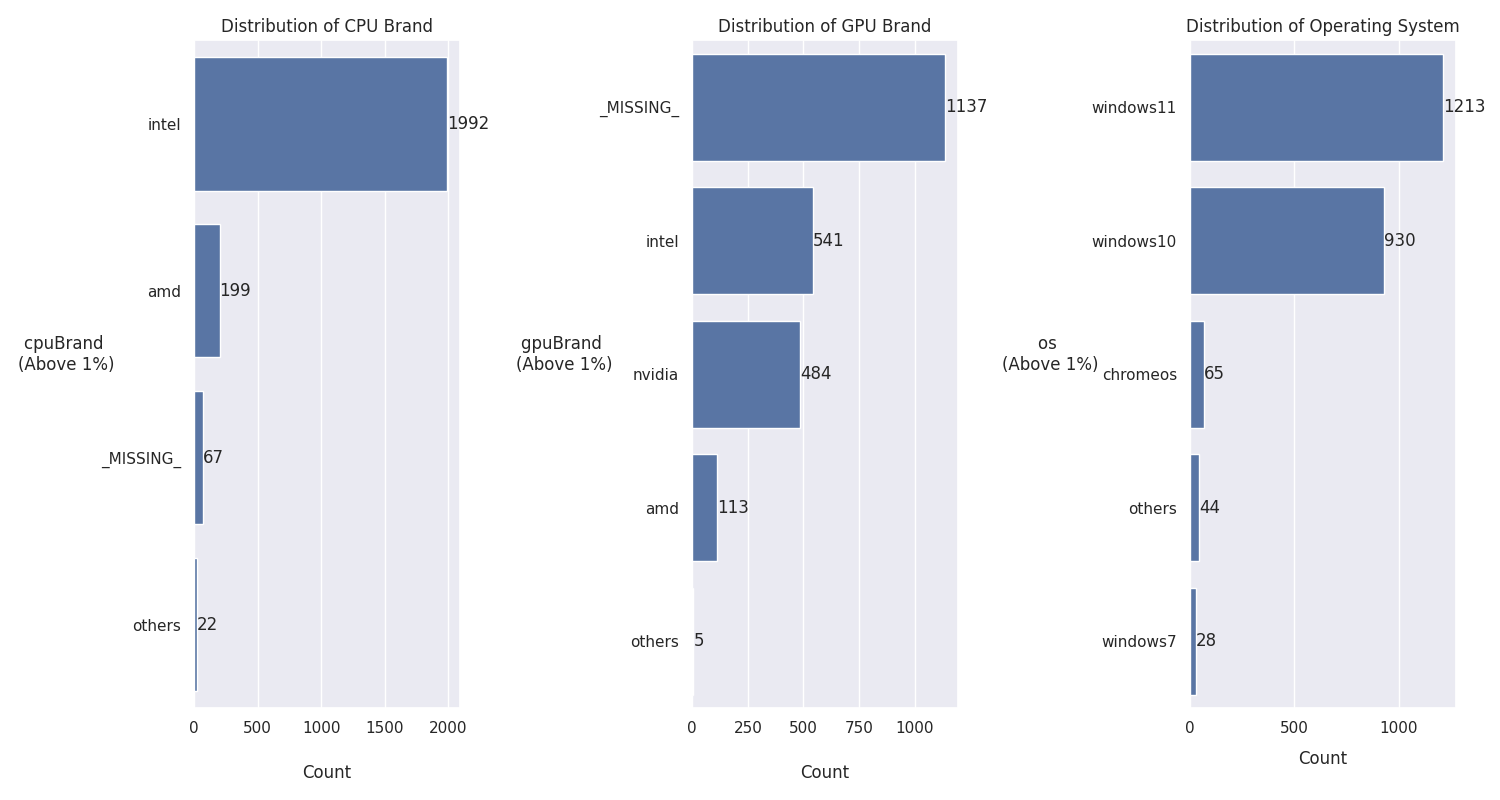
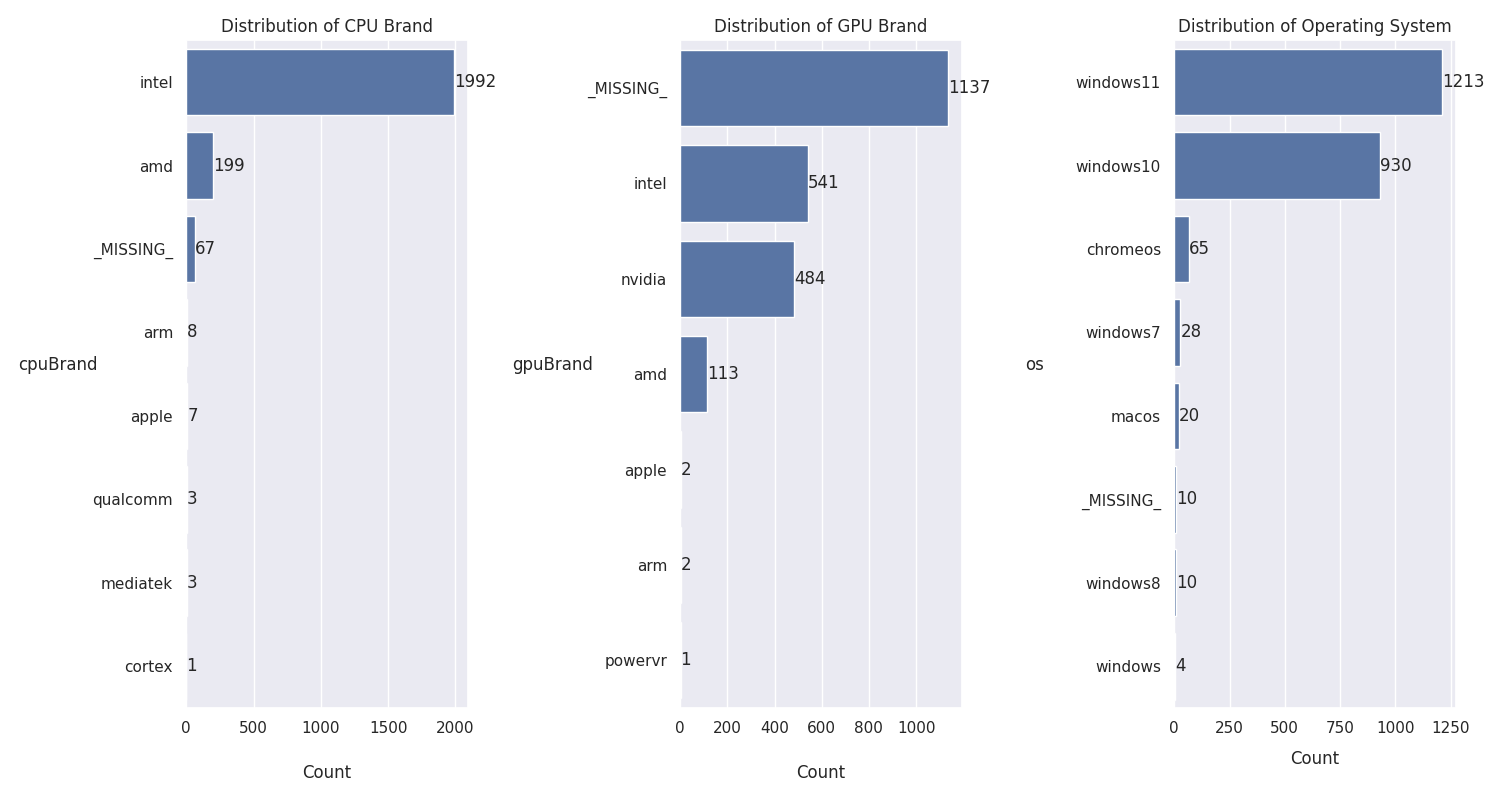
****

Figure-1: Before and after grouping cpu/gpu/os

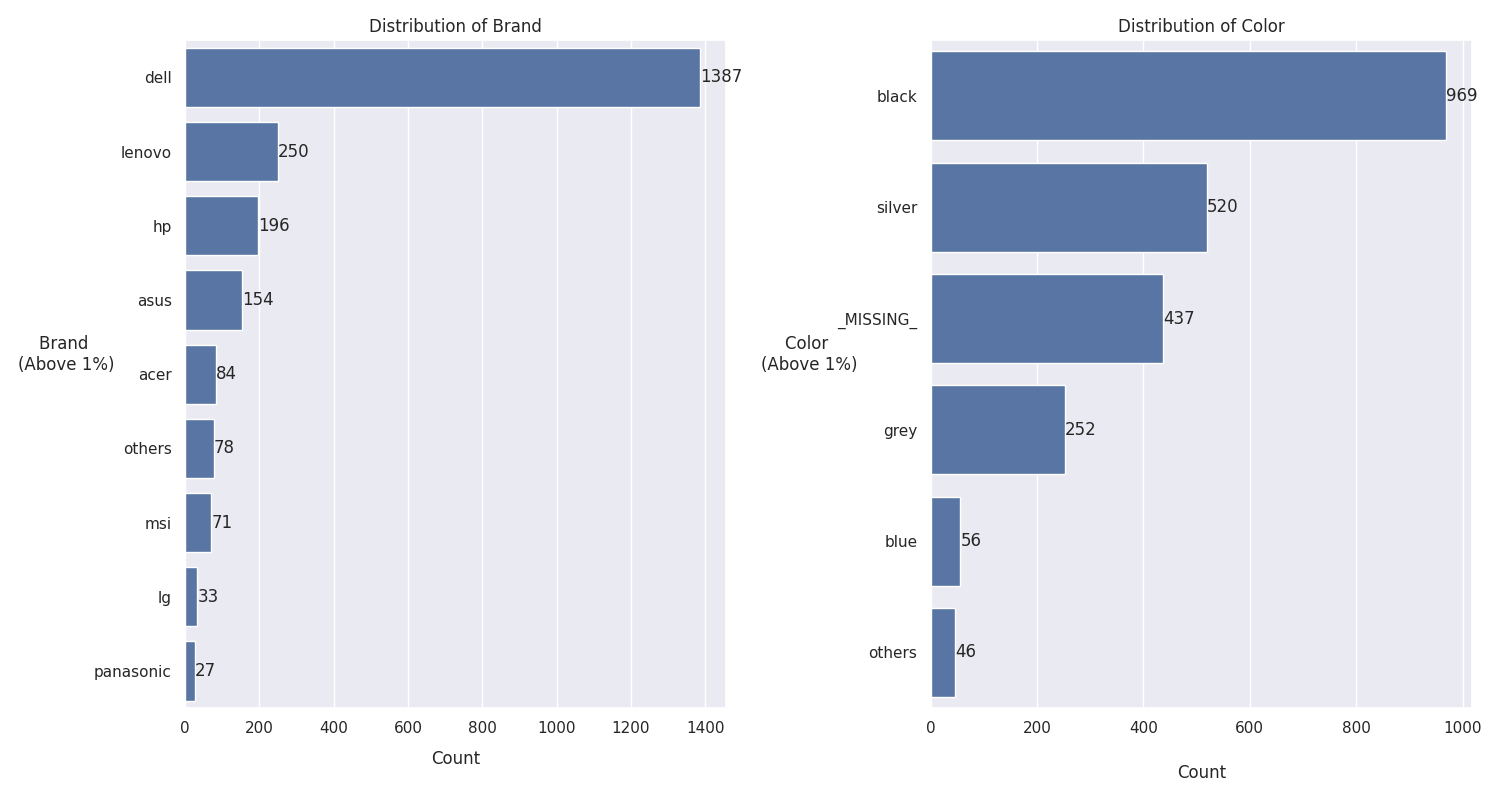
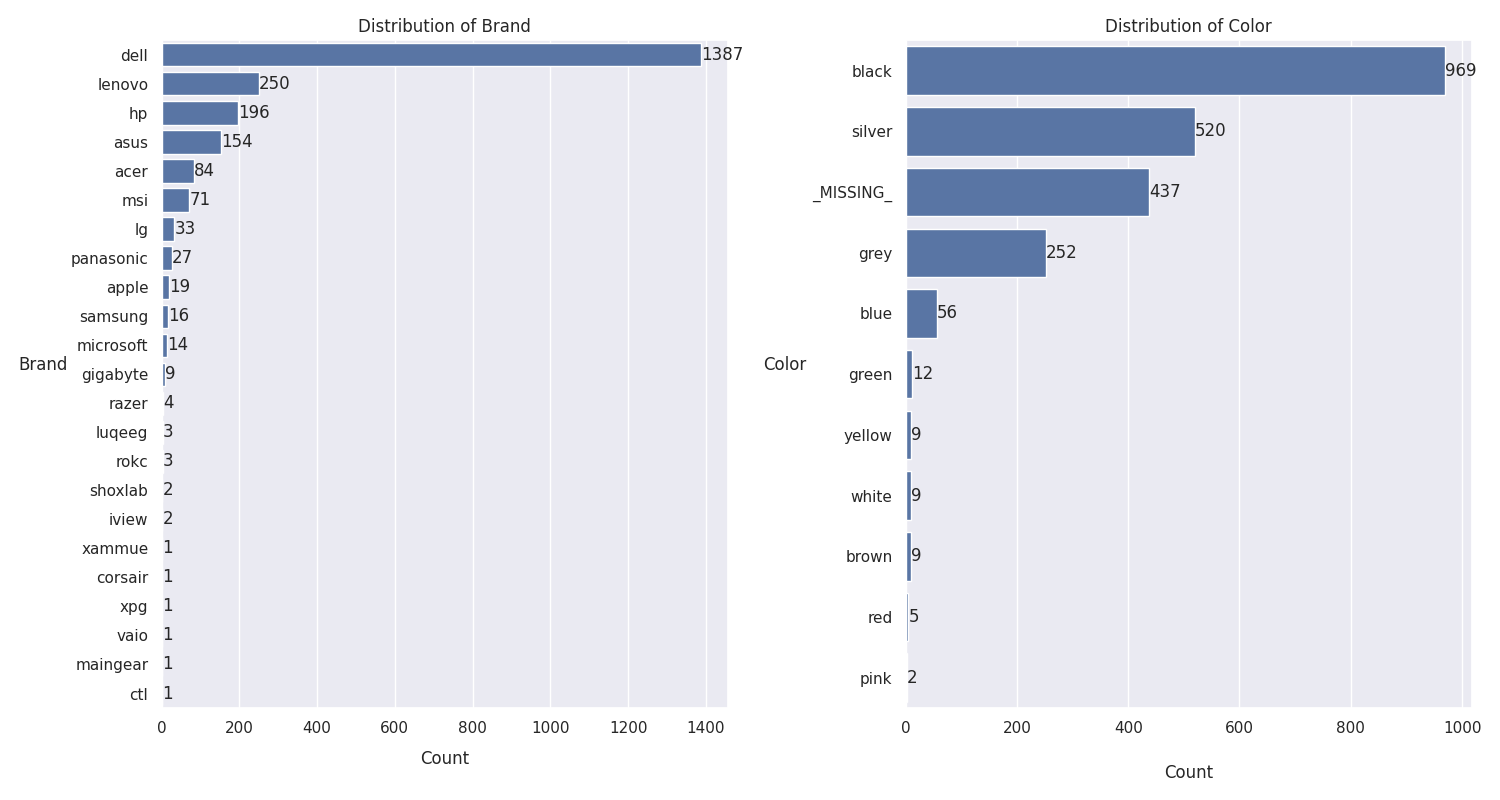
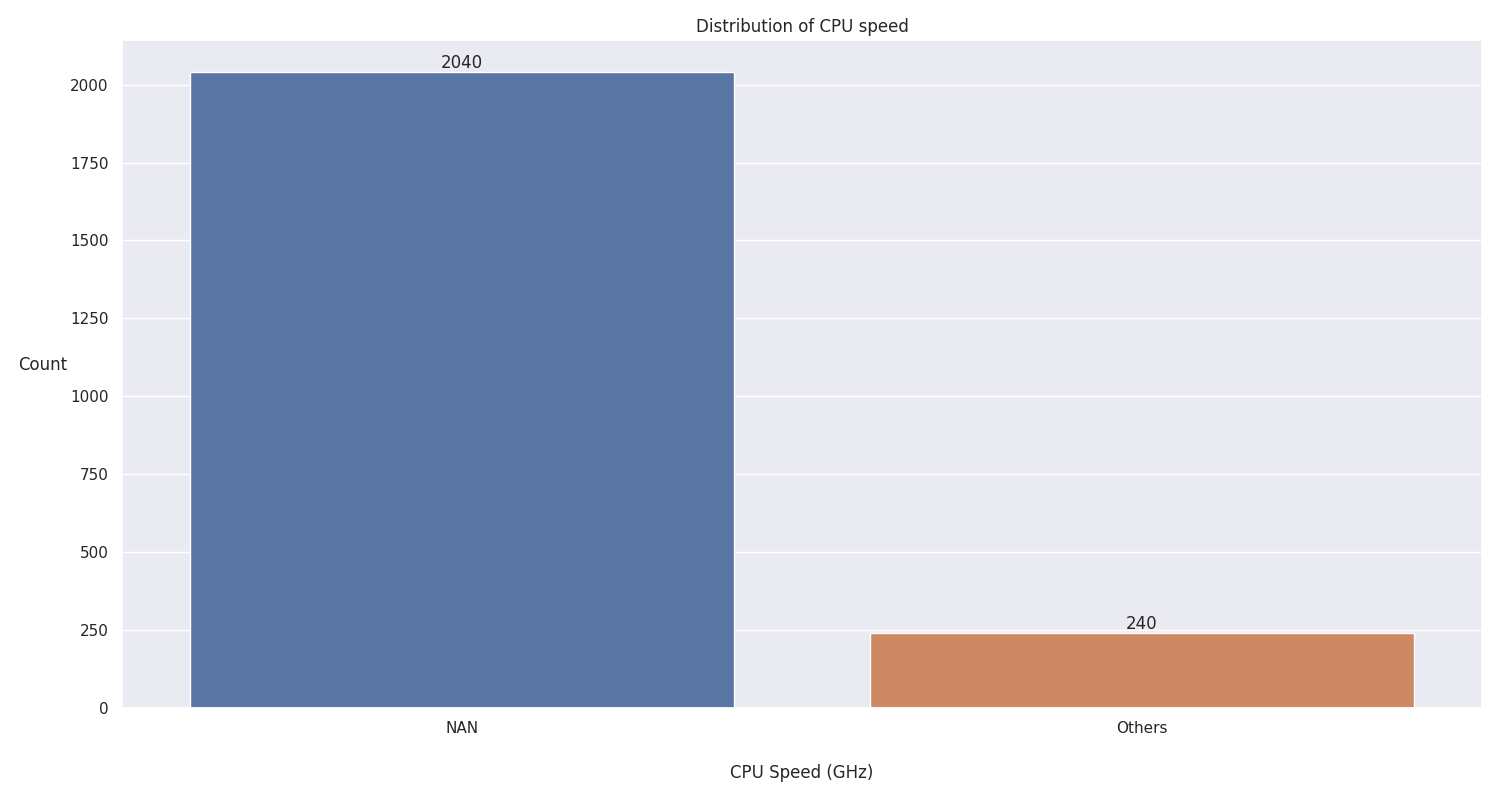
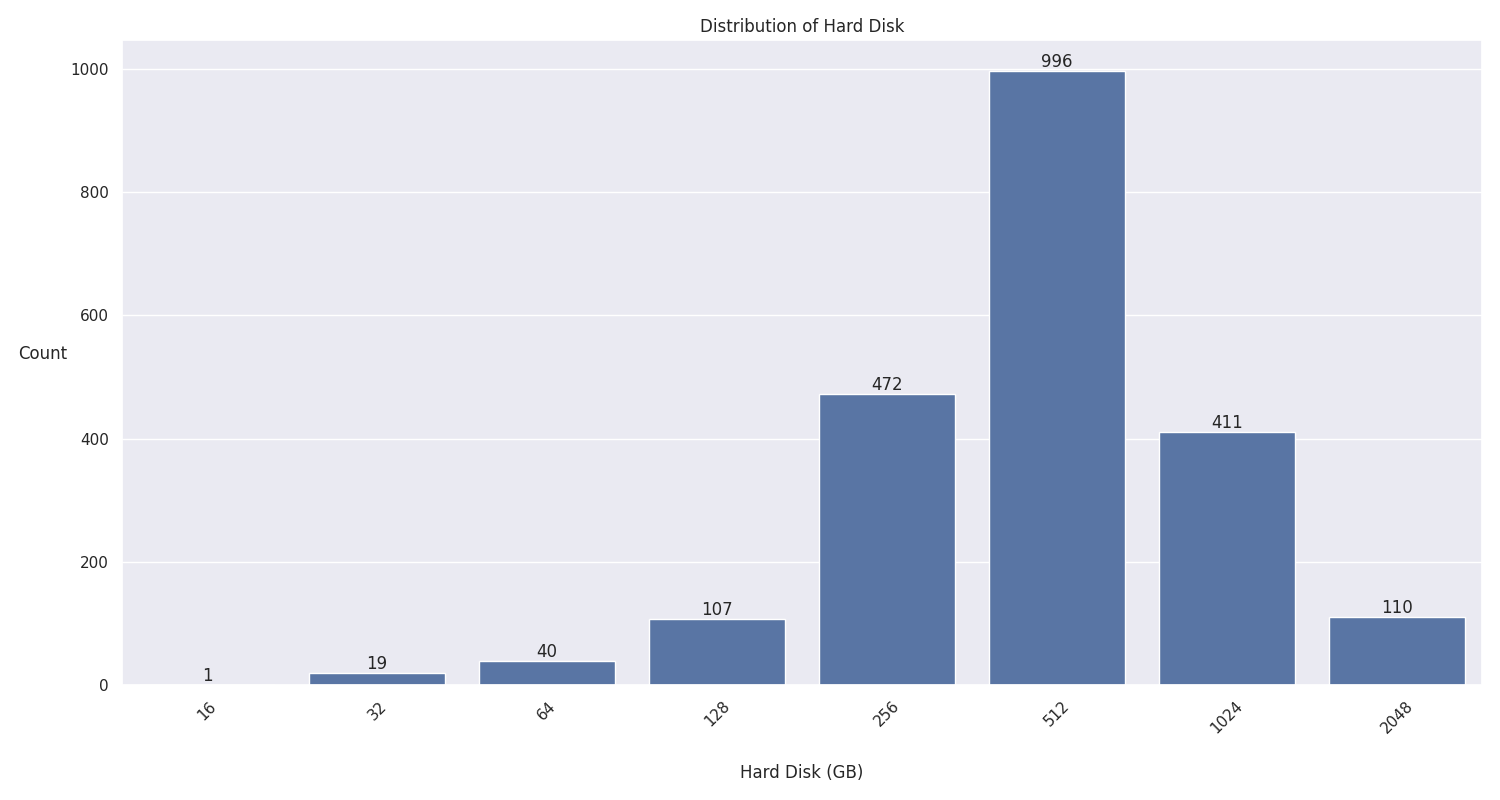
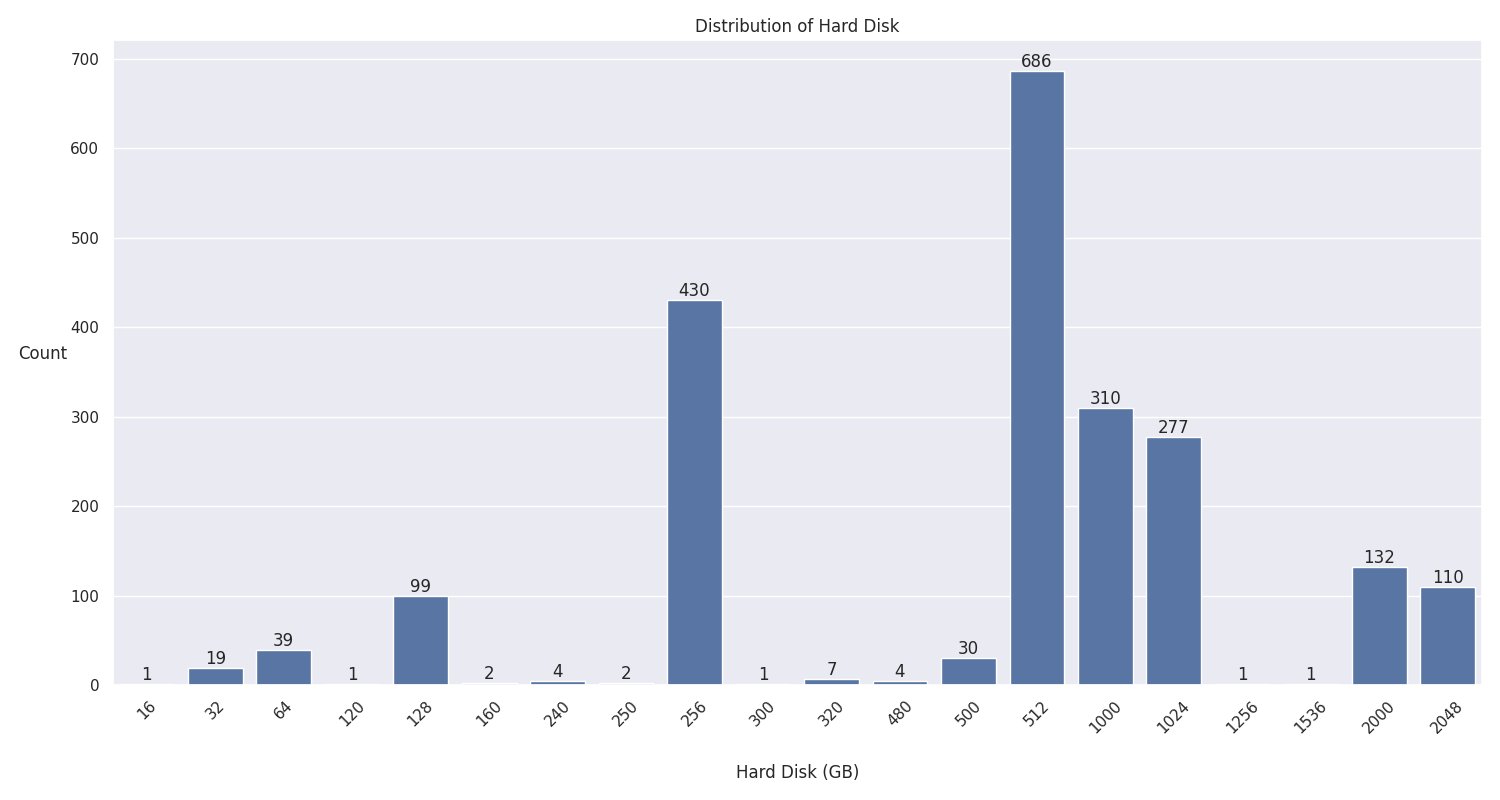
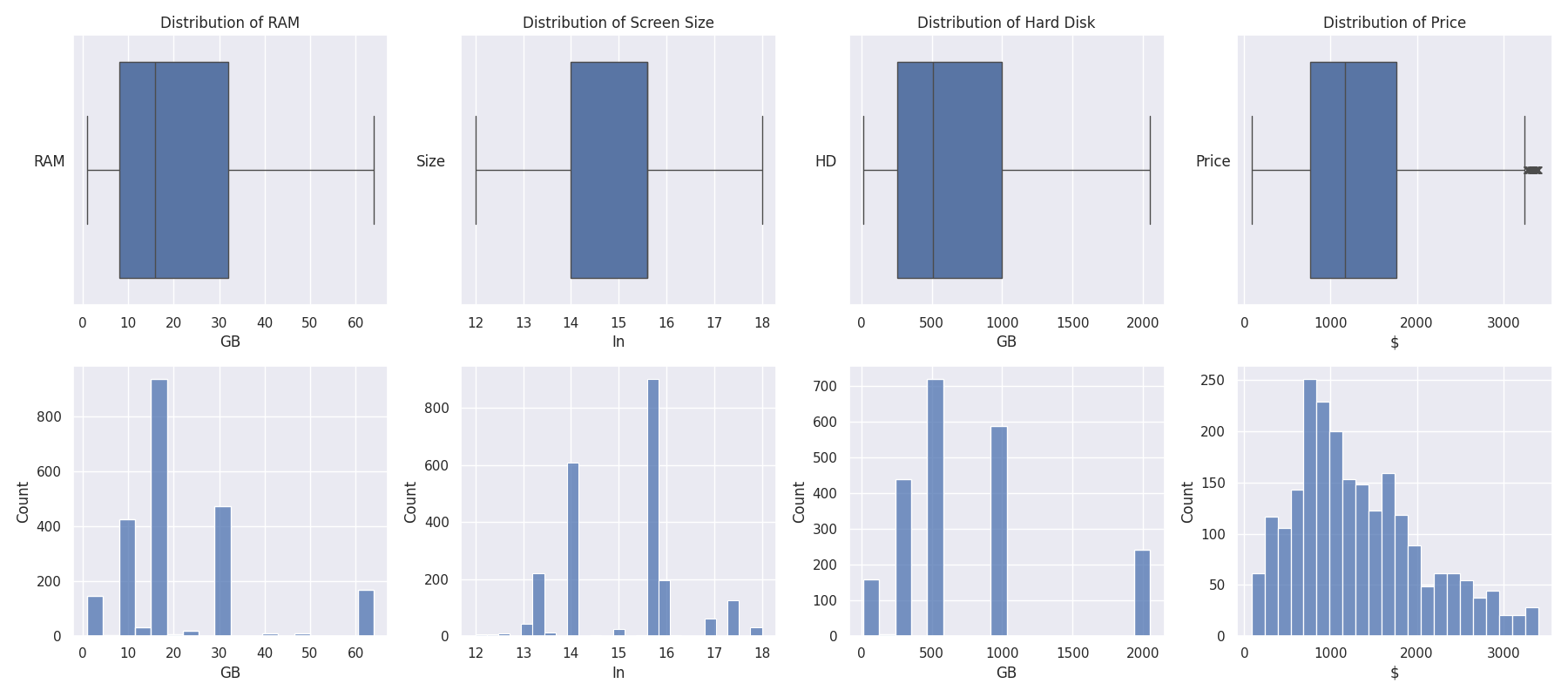
****

Figure-2: Before and after grouping brand/color

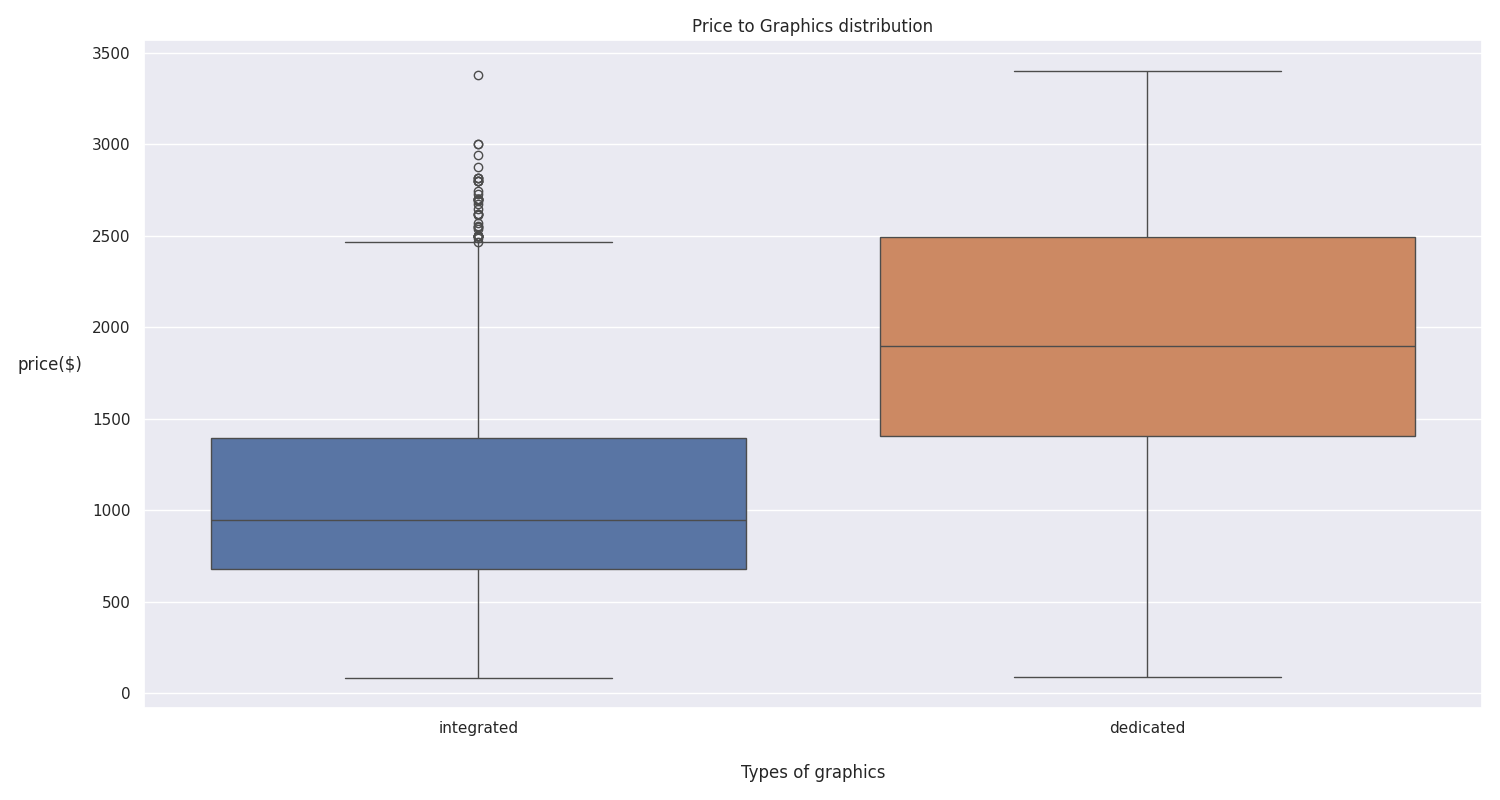
Figure-3: Sparse CPU speed

Figure-4: Before and after binning

A collage of graphs

Description automatically generatedFigure-5: Before and after removing outliers

A diagram of a graph

Description automatically generated with medium confidenceFigure-6: Before and after filling missing with integrated

**Problem 2**

**Choosing customer:**

* Gamer (Customer-1)
* College Student (Customer-2)

**Unique requirements:**

* Gamer: Strong laptop, large screen, windows os, good storage
* College Student: Popular trustworthy laptop, decent storage, lightweight

**Data Analysis**

Price

Fig-7 shows price having a medium-to-strong correlation with all variables besides rating, with the strongest being graphics and ram. As strong laptops usually cost more than $1000 (Laptopmedia, 2023), we assume price to be an indicator of performance, and so does graphics and ram. Customer-1 will likely want dedicated graphics and high ram.

A screenshot of a color chart

Description automatically generated Figure-7: Correlation price

Screen Size

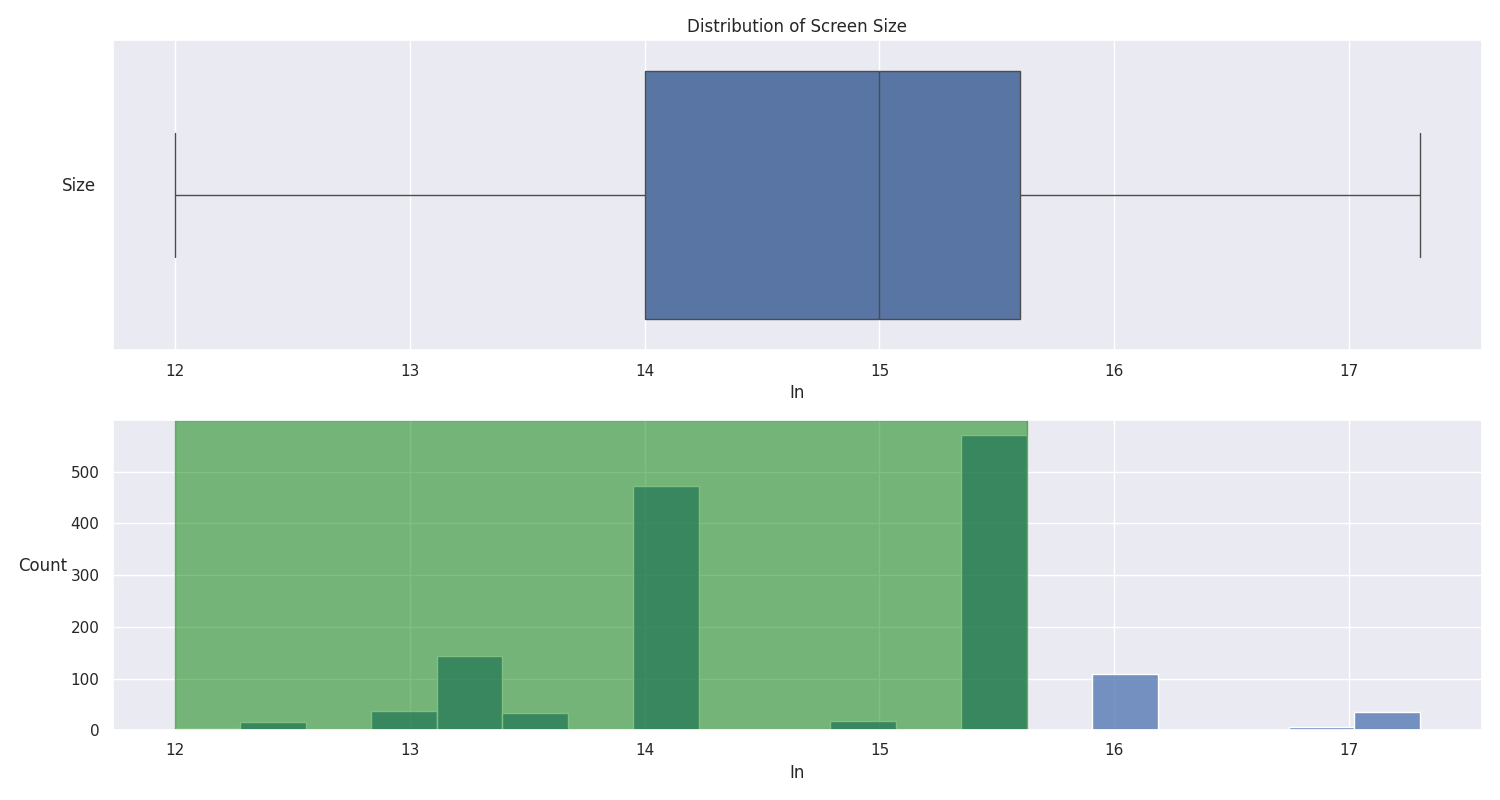
When discussing laptops, the trend is that the larger the laptop screen, the heavier it is (TheITBay, n.d). We classify large-screened laptops to be above 15.6in (StoneRefurb, n.d). Anything less than or equals to that (fig-8) will be for customer-2 as they want lightweight. 

Figure-8: Distribution of screen\_size

Fig-9 shows a strong positive correlation with screen\_size and graphics\_dedicated. This means if the graphics\_coprocessor is dedicated, the laptop is likely to be larger in screen size. Therefore, customer-2 would probably prefer integrated graphics.

A screenshot of a computer

Description automatically generatedFigure-9: Correlation screen\_size

GPU Brand

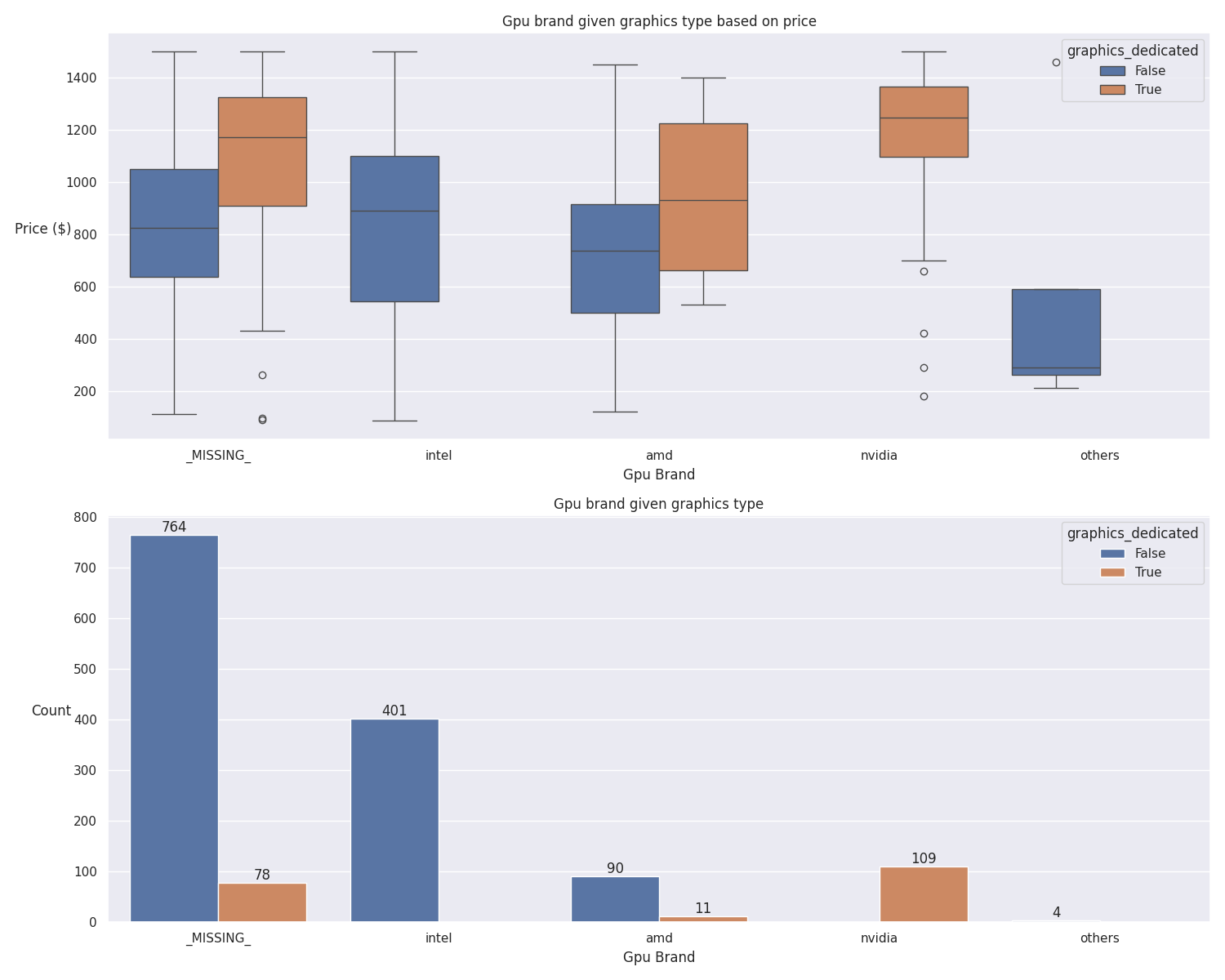
As rationalized above, customer-1 wants dedicated, and customer-2 wants integrated. Fig-10 shows that some GPU brands only sell one type of graphics, so the recommendation differs for each customer. We can also see how nvidia is higher priced than amd, suggesting higher performance. Customer-1 should get recommened nvidia or amd, and intel, amd or others for customer-2.

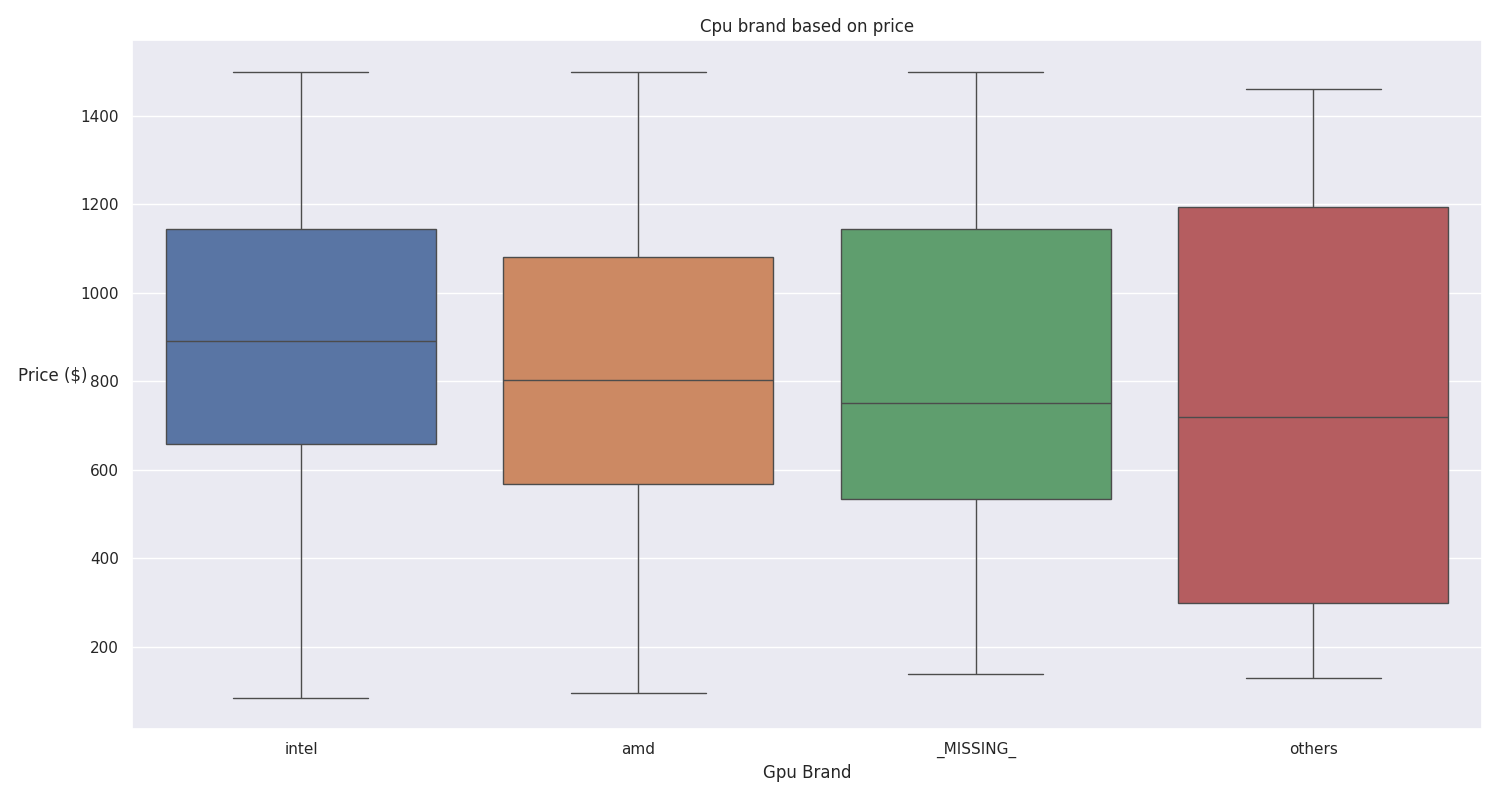
Figure-10: GPU brand/type to price

CPU Brand

From fig-11, we can see that CPU brand doesn’t affect other variables, so comparing it with other data won’t help us with recommendations. Fig-12 reinforces as it shows prices are approximately the same for each brand.

A screenshot of a computer

Description automatically generatedFigure-11: Correlation CPU brand

Figure-12: CPU brand price

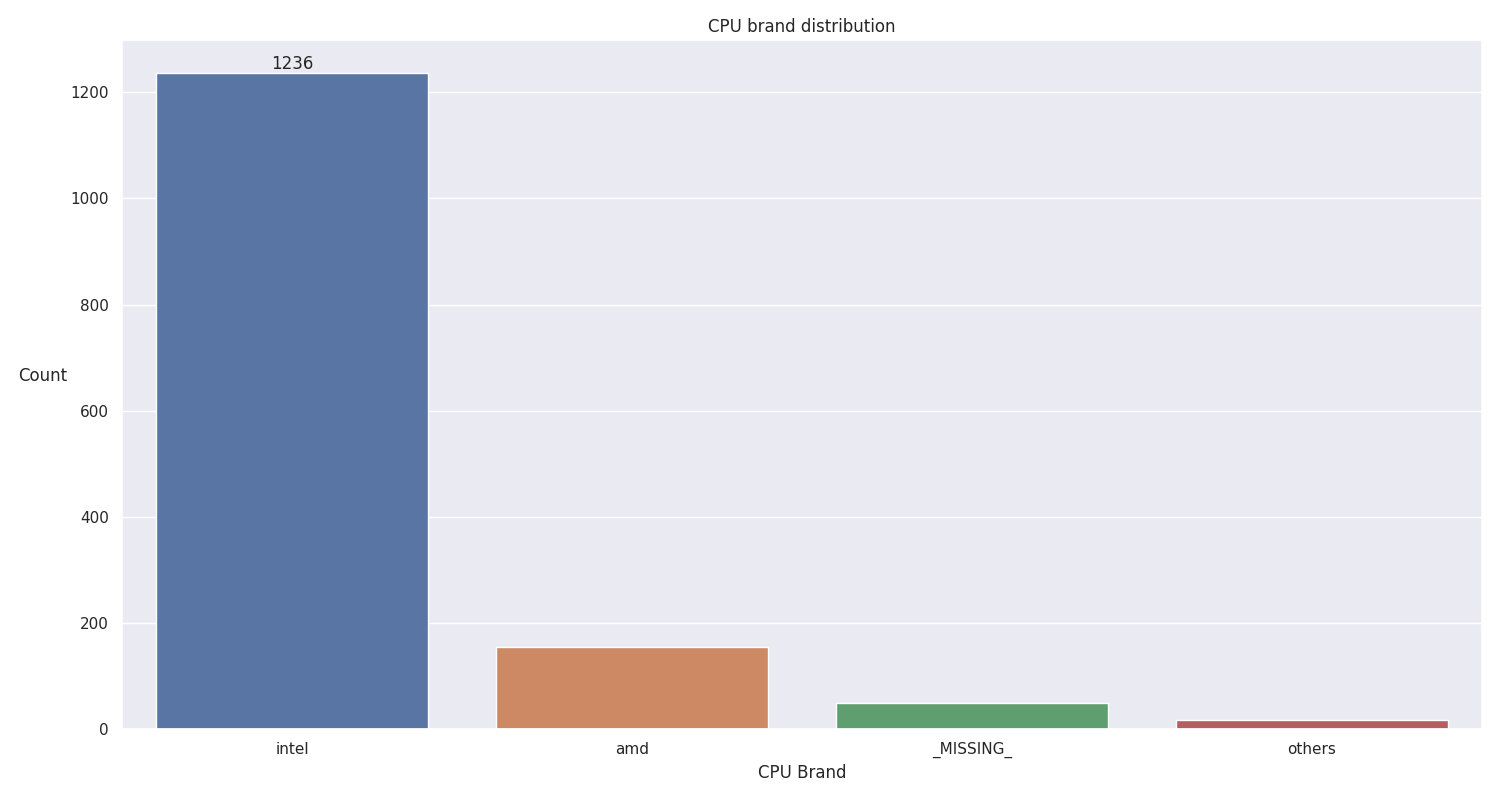
External sources say amd is better for gaming, whilst intel is better for work (Matthew, 2023). Fig-12 shows that intel is most popular which is what customer-2 wants.

Figure-13: CPU brand distribution

Ram and Hard Disk

Fig-14 shows a large jump in price between graphics type, suggesting better performance. 512GB is most popular, but customer-2 wants higher storage, so 1024GB is recommended.

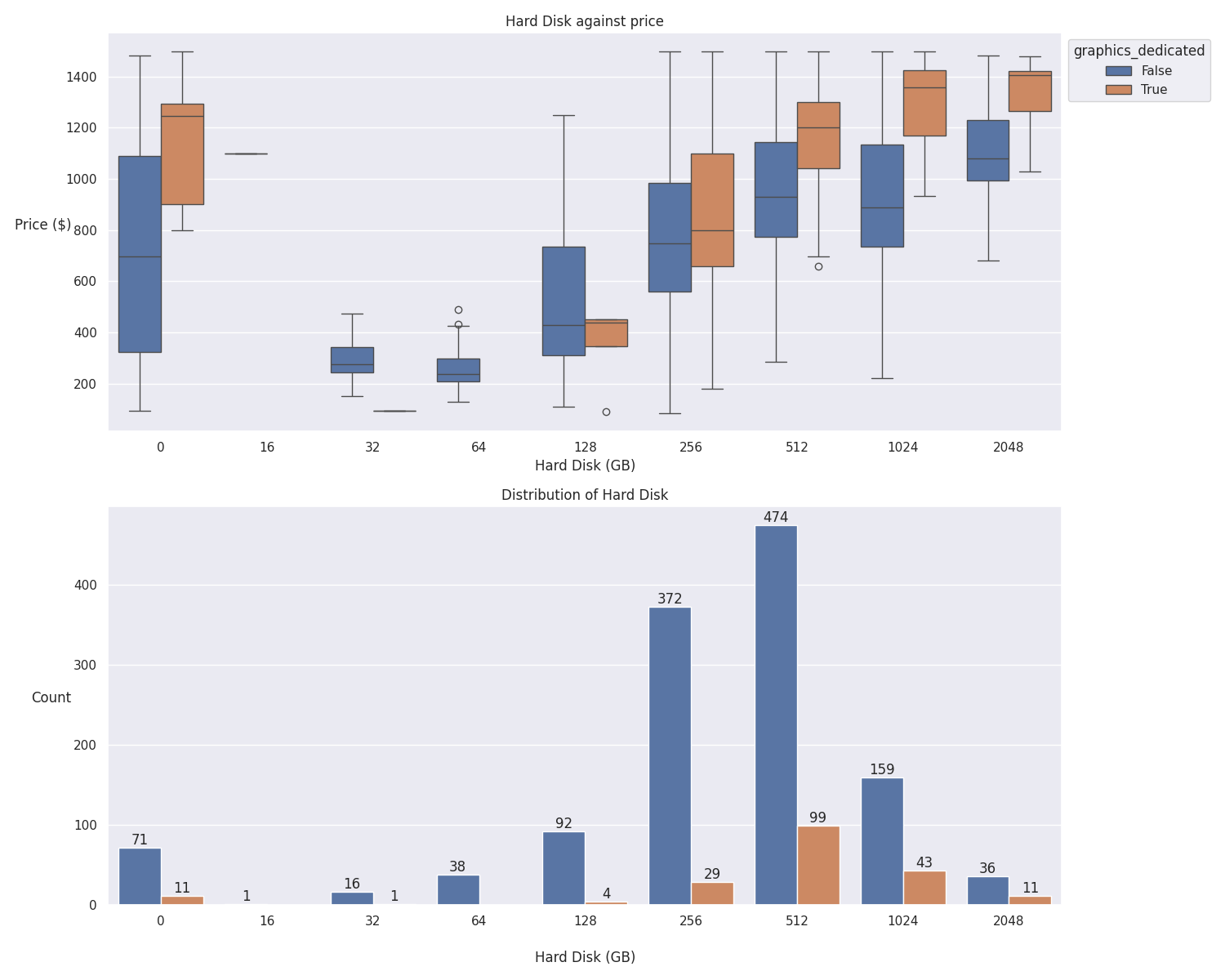
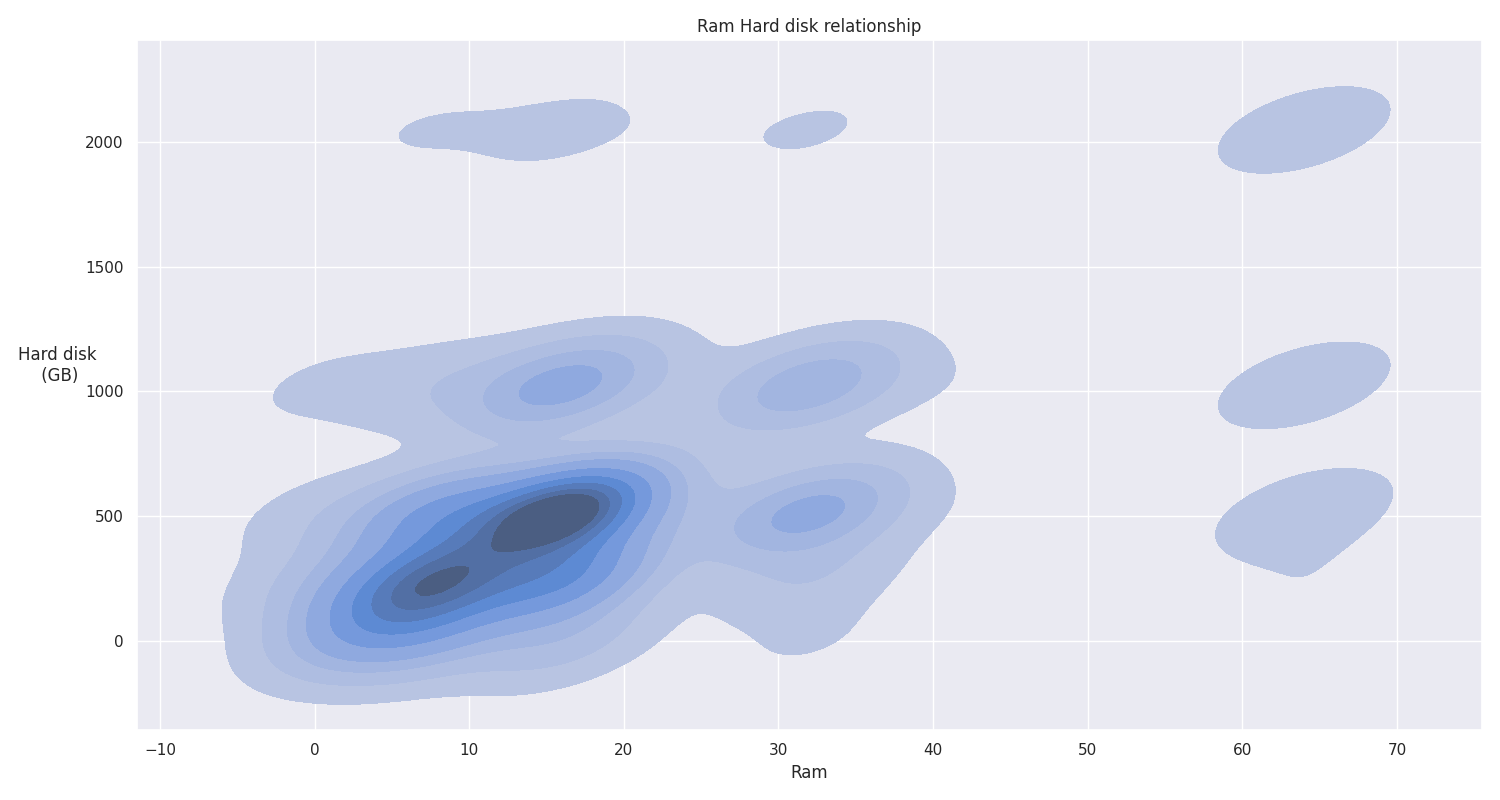
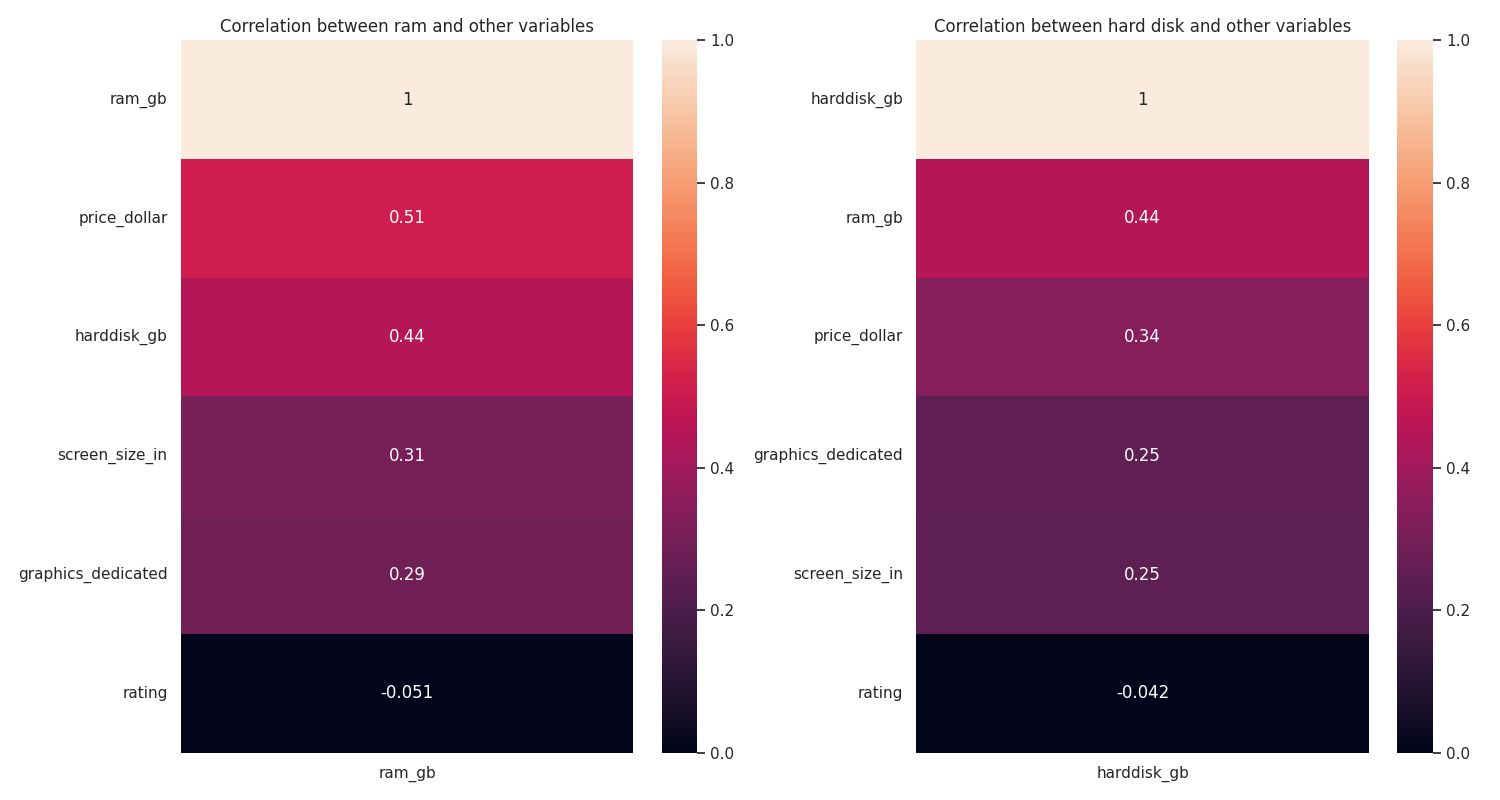
Figure-14: HDD distribution and Price

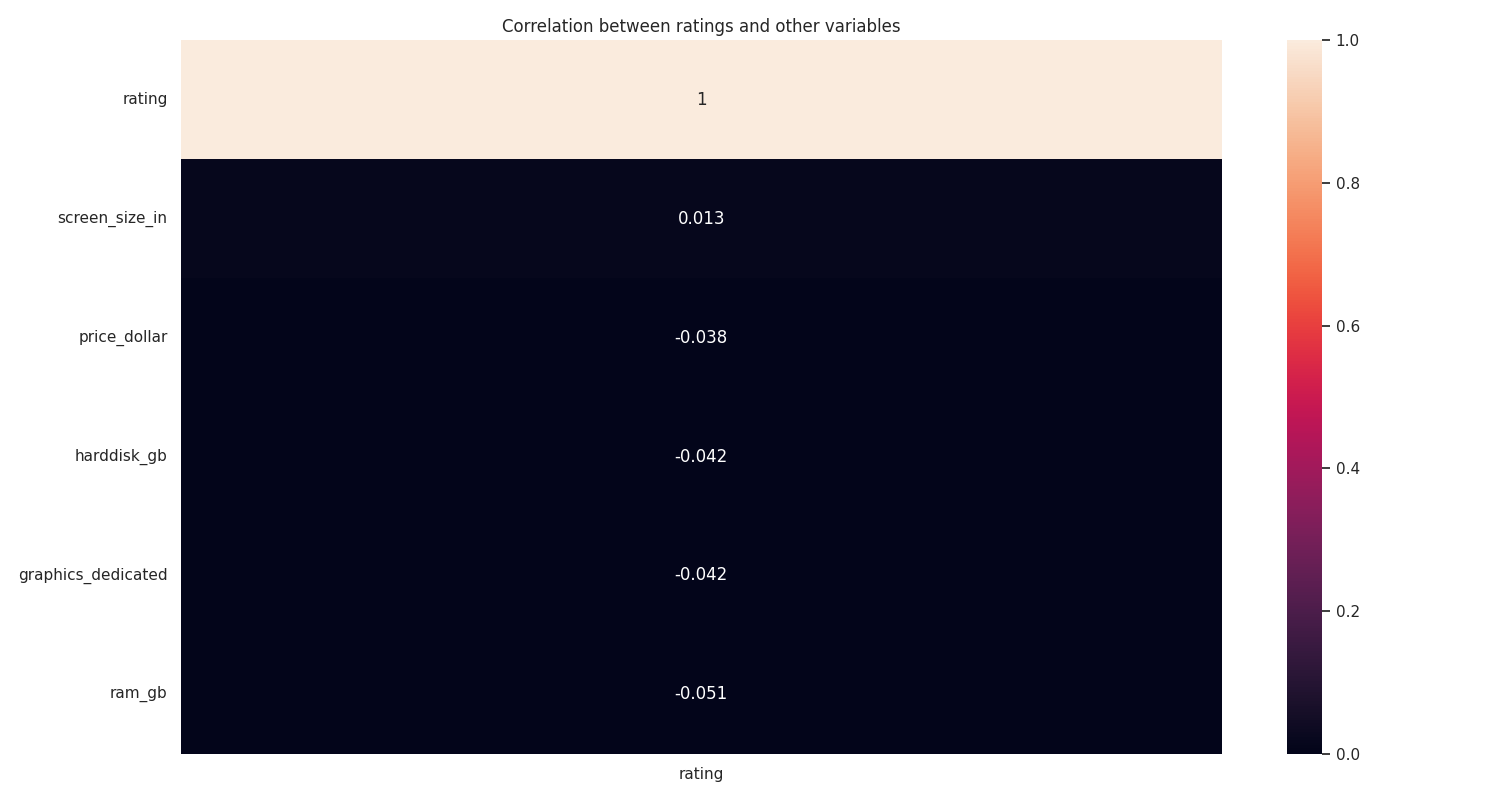
Fig-15 shows 8GB or 16GB ram is most popular with 512GB storage.

Figure-15: HDD vs Ram

 Figure-16: HDD Ram correlation

Ratings

Fig-17 shows ratings has no correlation with anything. However, no ratings suggest an unpopular laptop (Amazon, n.d), so a popular laptop should have ratings (above 4.5).

 Figure-17: Ratings correlation

**Recommendation**

Customer-1: A screenshot of a computer

Description automatically generated

Customer-2:

A screenshot of a computer

Description automatically generated

**Sources**  
Raghuvansh, T. (2020, November 2*). DATA PREPROCESSING: Decreasing Categories in Categorical Data,* MEDIUM. <https://medium.com/analytics-vidhya/data-preprocessing-decreasing-categories-in-categorical-data-132e8b4a4fd>

Omar, E. (2018, Februrary 28) *The Ultimate Guide to Data Cleaning.* MEDIUM. <https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4>

Joan, N (2022, May 21) *Handling Missing Values — Data Science*, MEDIUM. <https://medium.com/mlearning-ai/handling-missing-values-data-science-7b8e302264ee>

Seagate. (n.d.) *Storage capacity measurement standards*. <https://www.seagate.com/gb/en/support/kb/storage-capacity-measurement-standards-194563en/> (Last Accessed: 08/12/2023)

Techfident (n.d.) *How Much Storage Do I Need on My Laptop?* <https://techfident.co.uk/how-much-storage-do-i-need-on-my-laptop/> (Last Accessed: 08/12/2023)

Lenovo (n.d.) *What is processor speed?* <https://www.lenovo.com/gb/en/glossary/what-is-processor-speed/> (Last Accessed: 08/12/2023)

StoneRefurb (n.d.) *Advantages of Small and Large Laptop Screen.* <https://www.stonerefurb.co.uk/which-screen-size-should-i-choose> (Last Accessed: 08/12/2023)

TheITBay (n.d.) *Large Screen Laptops.* <https://www.theitbay.com/laptops-tablets/popular-laptop-searches/large-screen-laptops> (Last Accessed: 08/12/2023)

Laptopmedia (2023, December) *Top 100 Best Gaming Laptop Deals (Price/Performance)*. <https://laptopmedia.com/gb/top-100-best-gaming-laptop-deals-price-performance/>

Matthew S. S. (2023, November 1) *AMD vs. Intel: Which CPUs Are Better for Gaming?*, IGN. <https://www.ign.com/articles/amd-vs-intel-cpu-comparison>

Amazon (n.d) *Understanding Customer Reviews and Ratings.* <https://www.amazon.co.uk/gp/help/customer/display.html?nodeId=G8UYX7LALQC8V9KA> (Last Accessed: 09/12/2023)