Data Cleaning Report

Column Name	Situation of the column	Cleaning actions / steps	Justification / explanation
brand	Spelling errors, incorrect data	1) Reduce number of categories.	1) Easier comparison.
model	Missing data	1) Remove missing data.	1) Vital information for research question.
screen_size (in)		1) Discretize into bins, all in inches.	1) Easier comparison.
color	Missing data, spelling errors	1) Fix spelling errors.	
harddisk (GB)	Missing data, inconsistent units, not numerical	1) Numerical, all in GB.	1) Easier comparison.
ram (GB)			
cpu	Missing data, inconsistent naming cpu speeds	1) Include manufacturer names (consistency). 2) Move CPU speeds to column.	
cpu_speed (GHz)	Missing data, not numerical	1) Numerical, all in GHz. 2) Import external speeds[1].	2) Reduce missing data.
OS	Missing data, inconsistent naming, spelling errors	1) Clean to OS name and version number 2) Move misplaced data to correct column.	1) Easier grouping and comparison
graphics		1) Limit to 4 categories (integrated, dedicated, both and NA), moving coprocessor names.	1) Columns defined more clearly, removing crossover between them. Finding information and comparison becomes easier.
graphics_coprocessor		1) Contains only coprocessor name	

[1] (Intel, 2023) (AMD, 2023)

		or NA if integrated.	
special_features		1) Group similar feature names under one.	1) Easier comparison.
rating	Missing data, values between 0-5	1) Discretize into star rating.	1) More intuitive rating.
price ($)	Not strictly numerical	1) Make numerical. 2) Remove outliers (top 0.25%)	1) Allows numerical processing. 2) Very sparse and skewed mean a lot, so best to remove unless we get more data around that price range.

Data Analysis Report

Customer 1 works outdoors and needs at least 500GB harddisk and 8GB RAM. Ideally water-resistant and anti-glare.

After meeting minimum requirements, only the Lenovo 'ThinkPad' has both water-resistance and anti-glare. For other recommendations, *Figure 1* shows price variation does not significantly differ based on anti-glare screens, with the most laptops in each case being between $700-$1000, so it is worth buying one. Anti-glare screens are old technology (pre-2000), so customers expect not to pay high prices for it, which the data reflects.

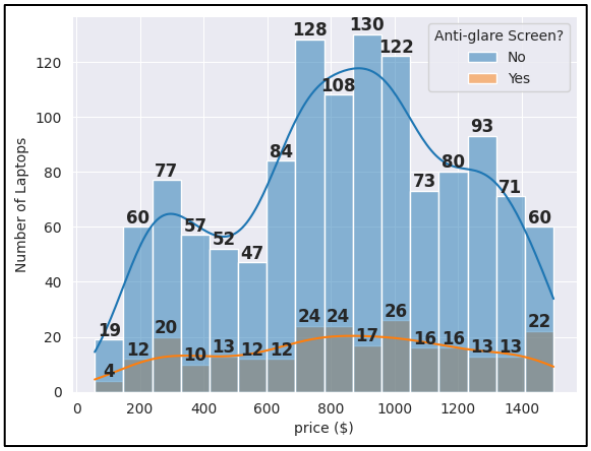

Figure 1: Price variation with anti-glare screens

Figure 2: Heatmap showing Pearson correlation coefficients of dataset

Figure 2 shows that RAM and harddisk have a Peason correlation coefficient of 0.52, meaning that they are strongly correlated, so the following metric is useful to maximise both at once:

$$RAM \times harddisk / RAM + harddisk$$

This reason for this correlation is that computers tend to have large RAM if they need to handle large quantities of data, and so a large harddisk is useful for storage.

Ranking the four best ranked laptops by this metric, **the recommendations** are:

	A	B	C	D	E	F	G
1	**brand**	**model**	**screen_size (in)**	**color**	**harddisk (GB)**	**ram (GB)**	**cpu**
2	dell	latitude 5440	14	NA	512	32	intel core i7
3	dell	inspiron 14	14	platinum	2000	64	intel core i5
4	lenovo	thinkpad	14	black	512	16	intel core i7
5	dell	vostro	14	NA	2000	64	intel core i5
6	dell	5000	14	platinum	2000	64	intel core i7
7	**OS**	**graphics**	**special_features**	**rating**	**price ($)**	**ram*harddisk/ram+harddisk**	
8	windows	integrated	anti glare / backlit keyboard	NA	996.99	30.11764706	
9	windows	integrated	anti glare / hd audio / memory c	NA	1209.99	62.01550388	
10	windows	integrated	anti glare / water resistant / fing	4 stars	1249	15.51515152	
11	windows	integrated	anti glare	3 stars	1381.99	62.01550388	
12	windows	dedicated	anti glare / hd audio / memory c	NA	1406.04	62.01550388	

<u>Customer 2</u> is an animator needing a large screen (at least 15in) and RAM (at least 16GB, ideally more). Ideally an included stylus and dedicated graphics.

After accounting for minimum requirements, *Figure 3* shows that no laptops in the data have both dedicated graphics and a stylus. Moreover, the median price (integrated: $805.99, dedicated: $1107.30) vastly differs based upon graphics since GPUs are expensive. Therefore, to give a range of choice, recommendations will include a laptop with a stylus, despite having integrated graphics. The one with largest screen size was chosen since this customer values a large screen.

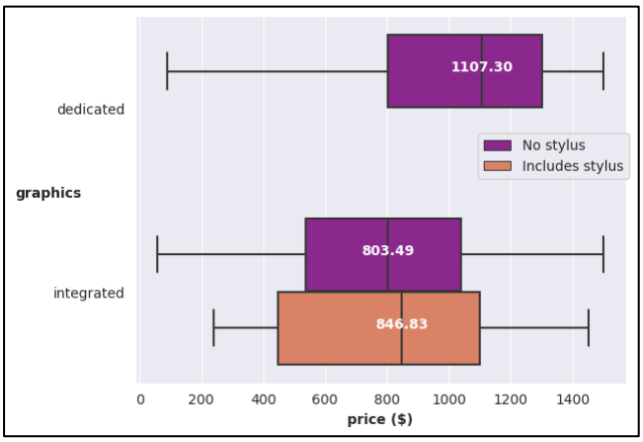

Figure 3: Range of prices for graphics and stylus options

To maximise RAM, *Figure 4* shows there are laptops with 64GB RAM less than $700 with integrated graphics. For both 64GB RAM and dedicated graphics, laptops are at least $1000. In either case, most 64GB RAM laptops have 15.6in screens. Therefore, if screen size and/or dedicated graphics are more important than maximum RAM, laptops will be more expensive. For choice range, the cheapest integrated and dedicated laptops with 64GB are recommended. *Figure 4* also shows there is one with dedicated graphics for less than $800, so this is recommended as good value. The last recommendation has dedicated graphics, 32GB RAM but maximum screen size of 17.3in.

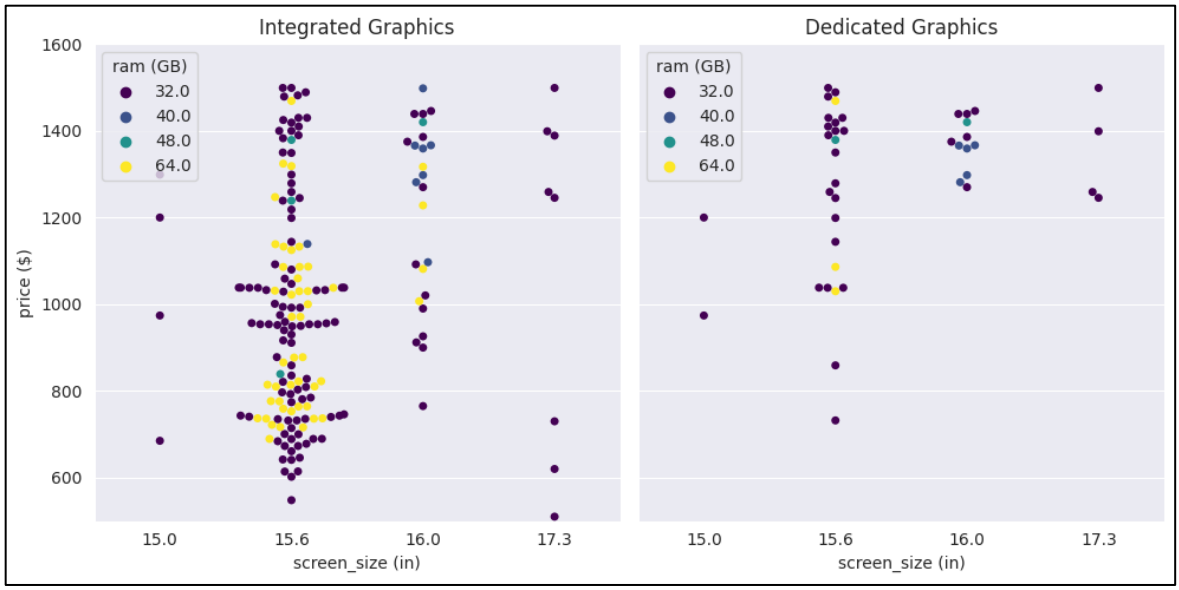

Figure 4: Price variation according to screen size and graphics: integrated (left) or dedicated (right).

The recommendations:

	A	B	C	D	E	F	G
1	**brand**	**model**	**screen_size (in)**	**color**	**harddisk (GB)**	**ram (GB)**	**cpu**
2	dell	inspiron 3511	15.6	black	256	64	intel core i5
3	hp	pavilion	15.6	horizon blue	1000	32	amd ryzen 7
4	msi	gf63 thin 10scxr	15.6	black	2000	64	intel core i5
5	dell	inspiron 7000	16	platinum silver	512	16	intel core i5
6	dell	precision 7000	17.3	grey	512	32	intel core i7
7	**OS**	**graphics**	**graphics_coprocessor**	**special_features**	**rating**	**price ($)**	
8	windows 10	integrated	NA	wifi / bluetooth	NA	689.33	
9	windows 11	dedicated	amd radeon graphics	NA	4 stars	732	
10	windows 10	dedicated	NA	anti glare / hd audi	NA	1029.99	
11	windows 10	integrated	NA	hd audio / stylus / r	NA	1099.99	
12	windows 10	dedicated	rtx a3000	anti glare screen	NA	1245.49	

General Observations

Both sets of recommendations are heavily populated by Dell. *Figure 5* shows more than half the data is Dell, explaining the skew towards them. These proportions are unreflective of the most popular laptop brands by ownership[2], limiting the usefulness of the dataset and the reliability of the recommendations.

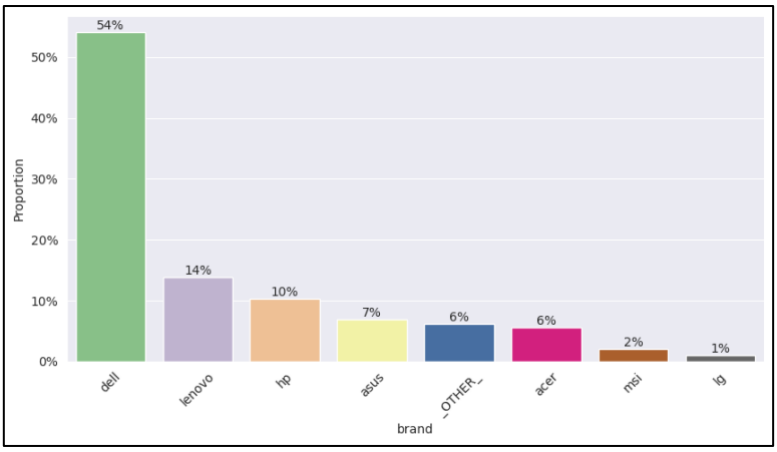

Figure 5: Proportions of laptops by brand

[2] (Statista, 2021)

Bibliography

AMD. (2023, December 10). *Processor Specifications*. Retrieved from https://www.amd.com/en/products/specifications/processors?s_platform%5B%5D=23291

Intel. (2023, May 1). *Comparison Chart for Intel® Core™ Laptop Processor Family*. Retrieved from https://www.intel.com/content/www/us/en/support/articles/000028083/processors.html

Statista. (2021, October 21). *Most Popular Laptop Brands in US*. Retrieved from https://www.statista.com/chart/26039/most-popular-laptop-brands-us/