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 is a 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 |
| graphics | Trailing & Leading whitespace | Strip whitespace |
| special\_features | Identical but shuffled rows | Convert to set, then sort | Improves duplicate detection |
| cpu | Unstructured data | Extract into cpu\_brand, cpu\_series, cpu\_model | Granular data is easier to analyse |
| graphics | Extract into graphics\_brand, graphics\_details |
| cpu | Empty | Drop columns | Data was extracted |
| grpahics\_coprocessor |
| screen\_size | Non-standard column names (no units) | screen\_size\_inches | Standardised, meaningful names are clearerref |
| color | colour |
| harddisk | harddisk\_gb |
| ram | ram\_gb |
| cpu\_speed | cpu\_speed\_ghz |
| price | price\_usd |
| OS | os |
| brand | Type inconsistency | Change to **str** | All 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 | Change to **Int64** |
| ram\_gb |
| screen\_size\_inches | Change to **float** |
| rating |
| price\_usd |
| cpu\_speed\_ghz | Mostly empty | Drop column | Over 88% missingref |
| rating | None | Only drop if over 80% missing, and there’s a valid reasonref |
| special\_features |
| colour |
| ALL COLUMNS | Duplicates | Drop duplicate rows | Duplicates skew dataref |
| brand | Too many groups | Less-frequent become ‘OTHER’ | Less groups improves visualisation readability |
| colour |
| cpu\_brand |
| os |
| graphics\_brand |
| cpu\_series | Bin using cpu\_brand |
| ram\_gb | Bin by ranges |
| harddisk\_gb |
| harddisk\_gb | Outliers | None | Outlier values do not imply erroneous/false valuesref |
| ram\_gb |
| rating |
| price |
| screen\_size\_inches | Remove screen-sizes above 21inches | Largest laptop screen-size is 21inchesref |

<https://encord.com/blog/data-cleaning-data-preprocessing/#:~:text=These%20duplicates%20can%20skew%20the,a%20particular%20analysis%20or%20model>.

<https://medium.com/mlearning-ai/handling-missing-values-data-science-7b8e302264ee>

<https://largest.org/technology/laptops-by-screen-size/>

<https://www.practicalpythonfordatascience.com/03_cleaning_data>

<https://swaathi317.medium.com/five-golden-rules-for-cleaning-data-in-power-bi-a50ed37dda54>