Problem 1

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Situation of the column** | **Cleaning action/steps** | **Justification/Explanation** |
| All columns | Mixed uppercase & lowercase | Apply casefold to convert to lowercase | Easier to standardise and find duplicates |
| model | Useless marketing terminology present | Remove all instances from each model value | Does not provide meaningful information to identify laptop |
| Extra data belonging to other columns | Extract into correct column |  |
| Models contain the brand | Remove instances of brand in the model | Data duplication, can be found in other column |
| Missing values | Drop rows with model missing | Laptop with no model cannot be identified easily. |
| brand | Impute brand based on model |  |
| graphics | Values other than ‘integrated’ and ‘dedicated’ | Move other value into graphics\_coprocessor |  |
| cpu | In-record errors, | Extract into cpu\_brand and cpu\_series, cpu\_model | Granular data is easier to analyse |
| graphics | Extract into graphics\_brand, graphics\_details |
| ram | Unstandardised | Round & convert all values to a consistent unit | Easier to compare and read |
| harddisk |
| cpu\_speed | Convert all values to a consistent unit. |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| brand | Unstandardised, Syntax Errors, Extra trailing & leading whitespace |  |  |
| model |  |  |  |
| color |  |  |  |
| OS |  |  |  |
| cpu\_brand |  |  |  |
| cpu\_series |  |  |  |
| cpu\_model |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| screen\_size | Non-standard column names with no units | Re-name screen\_size\_inches | Standardised meaningful names with units are clearerref |
| color | Re-name colour |
| harddisk | Re-name harddisk\_gb |
| ram | Re-name ram\_gb |
| cpu\_speed | Re-name cpu\_speed\_ghz |
| price | Re-name price\_usd |
| OS | Re-name os |
| cpu | Empty column | Drop columns | Data extracted into other columns |
| grpahics\_coprocessor |
| brand | Type inconsistency | Change type to **str** | All columns were Object type and contained various datatypes. This is not consistent or accurate. |
| model |
| colour |
| cpu\_series |
| cpu\_model |
| os |
| special\_features |
| graphics |
| graphics\_brand |
| graphics\_details |
| harddisk\_gb | Change type to **Int64** |
| ram\_gb |
| screen\_size\_inches | Change type to **float** |
| rating |
| price\_usd |
| brand | Strip rows |  |  |
| model |  |  |
| colour |  |  |
| cpu\_series |  |  |
| cpu\_model |  |  |
| os |  |  |
| special\_features |  |  |
| graphics |  |  |
| graphics\_brand |  |  |
| graphics\_details |  |  |
| 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 all duplicates in the data | Duplicates skew the dataref |
| brand | Too many groups | Less frequent values go into ‘OTHER’ category | Less groups improves visualisation readability |
| colour |
| cpu\_brand |
| os |
| graphics\_brand |
| cpu\_series | Bin values by cpu\_brand |
| ram\_gb | Bin values |
| 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 | Current 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>