Word Count: 549

PERRY, LUKE (Student)

Data Cleaning and Analysis Report

# Problem 1

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Issue | Cleaning Actions | Justifications |
| all | Missing entries | Unless other course of action is identified, replace entry with “unknown” | This will remove missing entries that might affect programmatic data analysis and make the data more readable |
| all | Inconsistent character capitalisation and leading/ trailing whitespace | Make all characters lower case and remove leading/ trailing whitespace | This will reduce the chance of two equivalent values incorrectly be confused as different values |
| model | Missing entries | Delete rows with missing model | With no way to uniquely identify the laptop, there’s no use considering the record |
| model | Duplicate entries | Delete all records bar 1, with the mode averages of the duplicates | Duplicate records provide no extra value. In the case duplicate values have conflicting information a new record must be produced by taking a modal average to estimate the correct information |
| Model/brand | Unnecessary division of columns | Merge columns | The brand can be directly inferred from the model, there’s no need for separate columns. If the model value isn’t empty then merge columns. |
| screen\_size | unnecessary inclusion of units in entries | remove string "Inches" from all screen\_size entries | Since all screen sizes are in inches there's no need to uniquely specify. Furthermore, removing strings can allow values to be treated as floats and operated on numerically |
| screen\_size | Missing entries | Replace missing entries with average screen size | By replacing with the average it means that for any future calculations made with a record with missing screen\_size value, it will be assumed to be average to not unnecessarily discredit or overcredit it |
| color | Missing entries | Replace with “unknown” | This will remove missing entries that might affect programmatic data analysis and make the data more readable |
| color | Entries have non-standard names and misspellings | * Create a dictionary of allowed colors * Use pattern matching to match unusual entries to those in the dictionary | This will make all entries in the color column legible by both humans and any programs I write |
| harddisk | Missing entries | Replace missing entries with average disk size | By replacing with the average it means that for any future calculations made with a record with missing harddisk value, it will be assumed to be average to not unnecessarily discredit or overcredit it |
| harddisk | Inconsistent units | Convert all entries to use the GB and remove all string units declarations | By converting all entries to GB it allows quick and easy comparison between values. This is due to a lack of a need to convert values and the entries can be treated as integers immediately due to a lack of string inclusion. |
| cpu | Missing entries | Replace with “unknown” | This will remove missing entries that might affect programmatic data analysis and make the data more readable |
| cpu | Inconsistent entries | Enforce all values take the form (core i\*, rhyzen \*, apple \*, ), by using pattern matching | As an overwhelming majority of laptops use either an intel or AMD processor, this will ensure most entries have consistent referencing to ensure equivalence. |
| ram | Missing entries | Replace missing entries with average ram size | By replacing with the average it means that for any future calculations made with a record with missing ram value, it will be assumed to be average to not unnecessarily discredit or overcredit it |
| ram | unnecessary inclusion of units in entries | remove string "GB" from all ram entries | Since all ram sizes are in GB there's no need to uniquely specify. Furthermore, removing strings can allow values to be treated as floats and operated on numerically |
| Graphics/ graphics\_coprocessor | Redundant columns | Merge columns | Only one column is needed for processors, if there isn’t a card specified that isn’t an integrated card then it can be accurately represented with just the “integrated” keyword |
| Cpu\_speed | Incredibly sparsely populated | Remove the entire column. | Since the column is so sparsely populated and the cpu column is filled mostly by cpu familys rather than specific models, so the clock speed cannot be inferred with any accuracy. |

# Problem 2