Data Science Assignment 1b

Timothy Fischer

Task Three: Data formatting and validation

Data formatting and validation involves making sure data is correct and suitable for analysis (Tablue). Without data formatting and validation, data analysis would most definitely give us inconsistent and skewed results. For example, if we don't normalize numeric variables with different scales, biases are introduced when comparing these variables since larger values have a greater effect on statistical and machine learning methods.

Common data formatting and validation methods:

- Data Cleaning: involves identifying errors and filling in inconsistencies in a dataset. It is crucial to make sure data is reliable and correct.
- Normalisation: involves changing the scale of variables to be the same, usually between 1 and 0.
- Standardisation: transforms data so that it has a mean of one and a standard deviation of zero.
- Data Transformation: Involves applying mathematical operations to data in order to change the distributions of data to better fit statistical models.
- Data Aggregation: involves combining data into stronger variables, often so that we can extract more meaningful data. Also, to perform dimensionality reduction in special cases.