Tutorial 2

Alloys dataset: Import the Alloys_needs_data_cleaning dataset into your Python statistical software. This dataset comprises mechanical property information on American National Standards Institute (ANSI) steel and copper alloys utilized in machine design (**170** alloys in total). The data was sourced from the Autodesk Material Library and obtained from Kaggle.

The dataset includes key properties such as:

- Ultimate Tensile Strength (UTS) measured in MPa
- Yield Strength (YS) measured in MPa
- Elastic Modulus (E) measured in MPa
- Shear Modulus (G) measured in MPa
- Poisson's Ratio (mu) measured in units of length
- Density (Ro) measured in Kg/m^3
- The final column distinguishes between the alloys (steel → Iron = 1, copper → Iron = 0).

The variables that we will focus on is:

- UTS (MPa)
- YS (MPa)

Review the dataset and inspect the data for data cleaning/data preprocessing.

1) Let's first look for completeness. For each column of data of interest, how many data values and data rows contain missing data? Briefly explain what you could do with these data values. Now go ahead and delete the rows that have any missing values.

2) Next, look for data-type validity. Since we care about UTS and YS, let's check that the data type matches what you would want the data to be used for in an analysis of mechanical properties. What are the data types for UTS and YS in the dataset? If the data type does not match what you want it to be, discuss what the data type is that you want and why, and the reasons that you find in the dataset for any datatype mismatch, and correct this. (Hint: There are a lot of way to do this: If you need to remove any values that do not match your desired data type, you can delete the row of data based on the datatype mismatch once you identify, or you can change mismatches to NaNs and then deal with those missing values by deleting by row). In the end, you want the data types to be what you want specified.

3) Now look at data validity. What do you think the limits of plausible data would be for UTS and YS? Go ahead and clean the data based on deleting any nonsense values. Again, delete entire rows to end up with a clean dataset of complete data on UTS and YS.

4) Choose a graph to represent UTS, YS, or UTS and YS and why you chose the visualization that you did. Show and discuss your graph(s).