Exercise 2

31.10.2024

In this exercise, we will be working with the dirty_iris dataset. You can download this dataset from: https://raw.github.com/edwindj/datacleaning/master/data/dirty_iris.csv

Data Cleaning:

EX2.1

a. Determine the type of different variables
b. Calculate the percentage of missing observations for each variable ('visdat' library, refer to Lecture 3).
c. Identify any outliers for each variable within each species using box plots. Determine the approach for handling these outliers, as discussed in Lecture 3.
d. Employ mean or median imputation to handle the NA values in the dataset, according to your choice.

Data visualization and exploration

EX2.2

a. Visualize the distribution of Species in cleaned iris data set (barplot).
b. Analyze the relationship between the sepal length and sepal width and between petal length and petal width (scatterplot, geom_point), use different color for each species. According to analysis answer the following questions:
   i. Which species has smaller sepal lengths but larger sepal widths?
   ii. Which species has larger sepal lengths but smaller sepal widths?

EX2.3

a. Compare Petal Length and Petal Width in different species by visualization techniques (scatterplot, geom_point)
b. Use histogram to visualize the distributions of different variables (petal length, petal width, sepal length, sepal with)

EX2.4

a. Create a new variable, sepal_to_petal_ratio, defined as the ratio of sepal length to petal length for each observation and visualize its associations with sepal length and petal length.
b. Add new variable to the main dirt_iris dataset.

EX2.5

a. Compute a suitable measure of central tendency and measure of dispersion for each variable, use a barplot to visualize it.
b. Create a new matrix to store statistical summaries of each variable in the dirty_iris dataset. For each variable, calculate the **mean**, **median**, **variance**, and **standard deviation**. Set up the matrix with columns named "Mean", "Median", "Variance", and "Standard Deviation", and rows corresponding to each variable name in the dataset.