STAT 210
Applied Statistics and Data Analysis:
Homework 2

Due on Sept. 18/2022

Question 1

You will need the file `Human_data.txt`. Place this file on your working environment.

(a) Read the file `Human_data.txt` and store this in an object called `human`. Before reading the data, check whether the file has a header. If it does, use the appropriate argument in the read function to include the header. Look at the structure of `human` using the function `str`.

(b) The body mass index (BMI) is defined as a person's weight in kilograms divided by the square of height in meters. Add a column named `bmi` to the data frame with the value of this index for each subject. Count how many subjects have BMI above 30.

(c) Calculate mean and standard deviation for `bmi` according to `Gender`. Compare these results and comment. Plot `bmi` against `age`, color the dots by `Gender`, and comment.

(d) Using `subset`, create a new data frame from `human` with the variables `Head_size, Height_cm, Weight_kg` for subjects with age between 30 and 50 (both inclusive) and head size bigger than 26. Call this new data frame `human1`.

(e) Use the function `apply` twice to calculate the mean and standard deviation for each of the three variables in `human1`. Call the vectors you obtain `human.mean` and `human.sd`.

(f) Use the function `sweep` twice, first to subtract the mean for each variable to the values in `human1` and then to divide by the standard deviation. Store the result in a data frame named `human.std`.

(g) The previous procedure is known as *standardization*. The resulting columns in the `human.std` should now have mean zero and variance equal to one. Verify this using `apply`.

Question 2

For this question you will use again the file `human` that you created in the first question.

(a) Use the function `split` on the file `human` with second argument `Gender` and store the result in an object called `human2`. Describe this object.

(b) Using the data in `human2` obtain a numerical summary (`summary`) for the variable `Salary` for males and females and compare.

(c) Use again the function `split` on `human` but now you want to use two variables for splitting the data, `Gender` and `Work`. Look at the help for this function to find out how to do this. Call the resulting object `human3`. Describe the file `human3`.

(d) Using the data in `human3` obtain numerical summaries for the variable `Salary` for males and females that work and compare.

(e) The function `cut` divides the range of values of a continuous variable into intervals and creates a factor according to which interval the values fall. You have to use this function to divide the range of salaries in the file `human` into three intervals, according to the following scheme: below 8000 is `low`, between 8000 and 18000 is `medium`, and more than 18000 is `high`. Call the resulting factor `sal`. Use the function `table` to count how many subjects fall in each category.

(f) Using the factor `sal` and the variable `Gender`, split the file `human` and call the resulting file `human4`. Using this file, obtain numerical summaries for the variable `Salary` for males and females that have a high salary and compare.