Data Analytics - Fall 2023
Assignment 1

DUE IN: Tuesday, 26.09.2023 in class,

FORMAT: You are strongly encouraged to work on the assignments and to provide a solution for them. For each week a different team is in charge, see file **Teams.pdf** on Microsoft Teams. The team in charge shall prepare a 15 minutes presentation showing and explaining their solution approach to the assignment.

The assignments will be created with R in mind as tool for using. For each assignment two teams will be in charge: one preparing a solution in R, the other one in Python.

Insurance data set

Download the insurance dataset from the following link `https://www.kaggle.com/mirichoi0218/insurance`.

1. Load the CSV by doing the following:

   ```
   setwd("C:/your\_filepath")
   insurance <- read.csv("insurance.csv", stringsAsFactors = TRUE)
   ```

2. Check the data structure by using the function `str()`

3. Check the statistics and check for normality of the data before a regression task by using the function `summary()` on the dependent variable `charges`.

4. We can see that the mean value is higher than the median, which indicates a skewed distribution. Take a guess about how it is skewed. Then, visualize it by using a histogram and see the skewed distribution. (Use the function `hist()`.)

5. Regression tasks are straightforward for numeric features. If there are categorical features, special care needs to be taken. We will see later how R handles them by default it if we are going straight to the Regression Model part. Now we are going to look closer to the correlations between those numeric features by using the `cor()` function.

   [Sidenote: You can extract the numeric variables by subsetting using `insurance[c("age", "bmi", "children", "charges")]`]

6. Visualizing the correlations with the `pairs()` function. And look at the scatterplot matrix. What do you see? How would you interpret it?

7. With this scatterplot matrix just now, it is quite difficult to find some trends. Try out the `pairs.panels()` function to plot on `insurance[c("age", "bmi", "children", "charges")]` and see the difference.

   [Sidenote: Install the package `psych` in order to use the `pairs.panels()` function.]

8. What does the `loess-curve` for `age` and `bmi` indicate?

9. Split the data now in training and test data using 80% for the training data and the rest for test data. How many observations are in each data set?

10. Create a linear regression model for the training data with the function `lm()`. You can either write it like:

    ```
    ins_model <- lm(charges ~ age+children+bmi+sex+smoker+region, data = train)
    ```

    or like the following, which is shorter because of the dot. The dot can be used to specify all features:

    ```
    ins_model <- lm(charges ~ ., data = train)
    ```

11. Look at the regression coefficients by typing the name of the model: `ins_model`. Try to interpret the values. How did R handle the categorical features?

12. Now, evaluate the regression model by using the `summary()` function on `ins_model`. What does it tell us? What is the p-value, what does the Multiple R-squared indicate?

13. Evaluate your model using the test data and compute the mean square prediction errors.

14. Repeat the steps in questions 9 to 13 ten times and compare training and test error for the different samples. What can you say about bias and variance of these models?