Assignment 1 - Data Wrangling

Dr. Reza Mohammadi

8/09/2020

The aim of this assignment is to use `exploratory data analysis` (EDA) techniques to explore the relationship among the real-world dataset `adult`. This dataset is available in the R package `liver`. You could find more information about this dataset at the following link on pages 2-3: https://cran.r-project.org/web/packages/liver/liver.pdf For this assignment use only the EDA techniques that we have learned so far (i.e. up till Chapter 3 of the book `Discovering knowledge in data`).

Your task is to answer the following questions and creating a report as a **R-markdown** (.Rmd file including R code). Please upload your **R-markdown** file with the **HTML** or **pdf** file on Canvas at the latest on **Wednesday 23 September 2020 at 23:59**. The total number of points assigned is 100.

1- Download R packages

Download the R packages `liver` and `ggplot2`. If it is needed, install the packages. (5 points)

2- Loading and understading the dataset

Upload the `adult` dataset which is available in the `liver` package. Report a summary of the dataset by using appropriate R functions. Are there any missing values? What is the number of variables? Which type of variables are they? (15 points)

3- Using EDA to analysis the dataset

Use the EDA techniques from week 2. For more information see Chapter 3 of the book `Discovering knowledge in data`. Indicate which variables have an association with the target variable `income` and which variables have no obvious association with the target variable. Explain why? (50 points)

4- Writing a summary

Summarize your EDA results from the previous question, just as if you were writing a report. (15 points)

5- Creating a report

Create a report as an R-markdown which should include the R code and the results of the code beside your interpretation. (15 points)