**Exercise 1**

Open the file **World Happiness 2019.csv** again. This dataset has a “happiness score” which was measured by simply asking the people in the country how happy they are on a scale of 0 to 10 and then calculating the average per country. It also contains many other variables which might be related to happiness. The dataset was compiled by the Sustainable Development Solutions Network (2020).

1. Run simple regressions with Happiness\_Score as the dependent variable and only independent variable at a time. Summarize the results.
2. Check if there is any multicollinearity between the variables (except for Happiness\_Score).
3. Run a multiple regression with Happiness\_Score as the dependent variable and all of the independent variables in a single model. Are there any differences compared to when you ran the simple regressions?

**Exercise 2**

Open the file **winequality-red.csv**. You don’t have to do any data cleaning. This dataset contains data of the Portuguese “Vinho Verde” wine which also used in the study by Cortez et al (2009).

1. Check if there is multicollinearity and deal with it in case it is a problem.
2. Carry out a regression analysis in which you try to predict the quality of the wine using all appropriate variables. What do you see?
3. Run a new model but this time exclude the variables which were not significant in the first model. Are there any changes to the coefficients?

**Exercise 3 (more advanced)**

Open the file **GBvideos.csv**. This dataset contains a large amount of YouTube videos which have been trending in the UK. You want to see if you can predict how much people like a video by looking at the number of views and the number of comments.

1. Why does it not make sense to simply predict the number of likes based on the number of views and comments?
2. Filter the data so that you only include videos with at least 1 like and 1 dislike. Also only use videos which have not blocked their rating, not blocked comments, and have not been removed. If you don’t know how to do this, try googling it.
3. Create a variable, like\_percentage that measures the total number of likes divided by the total number of likes and dislikes
4. Check if there is multicollinearity between the variables.
5. Run the regression model where you predict the like\_percentage using views and comments. What do you think these results could mean?

**References**

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Sustainable Solutions Development Network (2020. *World Happiness Report 2020*. Retrieved from: https://worldhappiness.report/ed/2020/