Assignment (worth 20% of your final grade)

Deadline for completion is the 30th of October 2022

On life satisfaction, children in Ireland rated themselves as having one of the lowest rates in the OECD/EU (72%), with 28% marking a score of 5 or under on a scale of 10 (UNICEF Office of Research). Among the issues contributing to these low scores, body image, pressure to succeed in school and bullying.

The Health Behaviour in School-aged Children (HBSC) survey is a WHO collaborative cross-national study that monitors the health behaviours, health outcomes and social environments of school-aged children every four years. The HBSC Ireland team, based at the Health Promotion Research Centre, University of Galway conducted the nationally representative survey of Irish school children in 2006, 2010, 2014 and 2018. This data set contains:

- **P_Content** The percentage of children that reported being happy with they way they are,
- **Age** the Age of the child,
- **Sex** the Sex of the child Male or Female
- **Year** the Year the information was collected,

Read the data set available on Brightspace contained in a Assignment.csv file into R.

Exploratory Data Analysis (24 marks):

For each question in the EDA section please provide the lines of R code required to produce your results and the tables and figures produced by R.

1. Using a boxplot, the density and the descriptive statistics (mean, min, max, median, and quantiles). Describe the distributions and the difference in the distributions for the percentage of school aged children that reported being happy with they way they are in 2006 with respect to sex (i.e. female vs male). (6 marks)
2. Using a boxplot, the density and the descriptive statistics (mean, min, max, median, and quantiles). Describe the difference in the distributions for the percentage of female school aged children that reported being happy with they way they are with respect to year (2006; 2010; 2014; 2018). (4 marks)
3. Using a boxplot, the density and the descriptive statistics (mean, min, max, median, and quantiles). Describe the difference in the distributions for the percentage of male school aged children that reported being happy with they way they are with respect to year (2006; 2010; 2014; 2018). (4 marks)
4. Convert the categorical variable **Sex** to a factor. Describe and illustrate the frequency of the categorical variable **Sex** with respect to year (2006; 2010; 2014; 2018) (4 marks)
5. Using the correlation and scatter plots discuss the relationship between **P_Content** and **Year** for males and females separately (6 marks)

Regression Model (62 marks):

1. Using R fit a simple linear regression model to the data with **P__Content** as the response variable and **Year** as a numeric predictor variable for females. Define and describe the terms in your mathematical equation for the model. (Also provide you R code) (4 marks)
2. Interpret the estimate of the intercept term (2 marks).
3. Interpret the estimate of the slope (2 marks).
4. What is the standard error of a parameter? Calculate and comment on the standard error of the estimate of the intercept and slope term. (4 marks)
5. Calculate and interpret the confidence intervals for β_0 (Provide you R code) (5 marks)
6. Calculate and interpret the confidence intervals for β_1 (Provide you R code) (5 marks)
7. What does the confidence interval of a parameter measure? (5 marks)
8. Does a 95% confidence interval always contain the population parameter? (5 marks)
9. Compute and interpret the hypothesis test $H_0 : \beta_0 = 0$ vs $H_a : \beta_0 \neq 0$. State the test statistic. Compare the test statistic to the correct distribution value and state your conclusion. Also, report the p-value and the conclusion in the context of the problem. (8 marks)
10. Compute and interpret the hypothesis test $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$. State the test statistic. Compare the test statistic to the correct distribution value and state your conclusion. Also, report the p-value and the conclusion in the context of the problem. (8 marks)
11. Interpret the F-statistic in the output in the summary of the regression model. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem. (6 marks).
12. Interpret the R-squared value. (2 marks)
13. Interpret the residual standard error of the simple linear regression model. (2 marks)
14. Calculate, plot and comment on the shape of the prediction intervals for the estimated values of Y (Provide you R code) (4 marks)

Regression Model Diagnostics (14 marks):

1. Are there any influential observations? (2 marks)
2. Examine the residuals of the regression model and comment on whether you think the residuals satisfy the assumptions required for small sample inference. Provide the rationale for your answer (10 marks).
3. Based on the information in Q2. How could you use the other information in the dataset to potentially improve your simple linear regression model? (2 marks)