**Stat291 Spring 2023: Group Project**

**Due 04/28/2023**

In homework 1, you have created a data set that contains the life expectancy in 2009 for 100+ countries, together with other characteristics like population, income, continent etc.

Develop a multiple regression model to find explanatory variables that can help explain and predict the life expectancy in 2009

What you may try

* You can include both numerical and categorical variables in the model
* You can create categorical variables based on the continuous variables, for example different income levels
* You can try interaction models
* You can try second order models
* You can even try interaction models that involve second order terms

What you should be careful

* Are there any transformations needed?
* Does multi-collinearity exist?
* Are there influential points or outliers?

What you should include in the report

* Exploratory data analysis
* Description of how you reach your final model.
* Residual analysis
* Anything that you think worth specifying, like outliers.

You may start from graphical investigation of the relationship between the dependent variable and predictors, develop the first order model, and then gradually include the second order, or interaction terms into your model. Try all the possible two way interaction terms to see if any of them is significant.

In your report, you do not need to include full details of all the regression work that you tried. But you should list things that were attempted. Your report should provide enough details to justify your final selection of the best model and to show the major steps that lead to your conclusion. Make sure to include the residual plot for the final best fit model to verify the model assumptions and explain the effect of the predictor variables on the life expectancy

It is possible that the best model you tried still has this or that kind of problems, but the basic idea is to let you have a sense how to develop regression models.