STA437H1S / 2005 HS Winter 2020 Project

Due: In Crowdmark by 10pm on Friday, April 3, 2020. **Email submissions or late projects are disallowed.**

1. Grading / Instructions

- Use R (or R Studio) to do the data manipulations and analyses.
- Compile your solution as a PDF document (Word, LaTeXor Rmarkdown can be your base).
- Presentation of solutions is very important. Your project should have two main sections- Report and Appendix. Include relevant plots and quote relevant numbers from your R output for your report. In the Appendix, include your R codes; your codes must be readable and reproducible. Five marks will be awarded for excellent presentation and your appendix.
- You may partner with at most one person for this project. Crowdmark will allow for group submission. For academic integrity, personalized your code as much as possible, using your first name(s). **All graphical plots produced must be given a title with the last 4 digits of your student number or your partner's**.
- Your data report should not exceed 10 pages in length (excluding plots and R codes). However, use at most 4 plots in your report.

2. The Data

The data to be considered for this project was downloaded from the World Happiness report of 2017; see

Statistical Appendix for "The social foundation of world happiness", John F. Helliwell, Hailing Huang and Shun Wang, Chapter 2, *World Happiness Report 2017*, March 21, 2017

The file `happiness2017.csv` in Quercus contains the data. The main variables in the dataset are:

- `Ladder`- the happiness score or subjective well-being of a country. Values are in the range 0 to 10, where 10 represent the best possible life.
- `LogGDP`- the logarithm of a country's 2016 GDP.
- `Social`- the national average of the binary responses to the question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"
- `HLE`- Healthy life expectancy at birth based on data reported from the World Health Organization (WHO)

- `Freedom` - the national average to the question “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?”
- `Generosity` - the residual of regressing the national average of the response to the question “Have you donated money to charity in the past month?” on GDP per capita
- `Corruption` - the national average to the two questions “Is corruption widespread throughout the government or not?” and “Is corruption widespread within businesses or not?”
- `Positive` - a measure of positive affect as the average of three positive affect measures - happiness, laugh and enjoyment in the Gallup World Poll.
- `Negative` - a measure of negative affect as the average of three negative affect measures - worry, sadness and anger in the Gallup World Poll.
- `gini`- the gini of household income in international dollars as reported in Gallup

The aim is to predict the happiness score of a country using the other nine (9) variables. Your report should include the following four sections, along with an Appendix.

3. The Project Report

I. **Summary/Abstract** (5 marks)- Write an executive summary of your analysis by using between 100 and 300 words. Do not include any plots in this part.

II. **Data Manipulation and Summary** (10 marks)- Set the seed of your randomization to be the last four digits of your student number or your partner’s student number. From the raw data, which is in alphabetical order, take a random sample of 100 countries for your data. Carry out the natural first step to data analysis by displaying multivariate data graphically and to obtain summary statistics. Note that, you should take into account any interesting features in the data that you come across. This implies making suitable transformations, handling outliers, removing countries with missing values or making suitable imputations for missing values. As a next step, you should check multivariate normality assumption for your data as most of the techniques we cover require normality.

III. **Multiple Linear Model using Original variables** (10 marks)- Fit a linear model for the response variable- happiness score using all the original explanatory variables. Check whether the assumptions of this model are satisfied. Report the ‘Adjusted R-squared’ value from the R `summary(lm(...))` output

IV. **Multiple Linear Model using Principal Components** (10 marks) - Fit a linear model for the response variable- happiness score using less than 9 principal components. Discuss whether you found the principal components from the covariance matrix or the correlation matrix. Discuss how you decided on the number of principal components that you retained. Produce a scree plot. Report the ‘Adjusted R-squared’ value from the R `summary(lm(...))` output for your ‘final’ model. Interpret the principal components retained.