DSC 423: Data Analysis and Regression
Assignment 07: Regression Pitfalls

Your submission must include your name and student ID. Your submission must include the honor statement: "I have completed this work independently. The solutions given are entirely my own work." Your submission must be submitted as a PDF.

1) (50 points) Download the Pisa2009 Dataset from the D2L. The Programme for International Student Assessment (PISA) is a test given every three years to 15-year-old students from around the world to evaluate their performance in mathematics, reading, and science. This test provides a quantitative way to compare the performance of students from different parts of the world. In this homework assignment, we will predict the reading scores of students from the United States of America on the 2009 PISA exam. The dataset contains information about the demographics and schools for American students taking the exam, derived from 2009 PISA Public-Use Data Files distributed by the United States National Center for Education Statistics (NCES). Each row in the dataset represents one student taking the exam. The datasets have the following variables:

- grade: The grade in school of the student (most 15-year-olds in America are in 10th grade)
- male: Whether the student is male (1/0)
- raceeth: The race/ethnicity composite of the student
- preschool: Whether the student attended preschool (1/0)
- expectBachelors: Whether the student expects to obtain a bachelor's degree (1/0)
- motherHS: Whether the student's mother completed high school (1/0)
- motherBachelors: Whether the student's mother obtained a bachelor's degree (1/0)
- motherWork: Whether the student's mother has part-time or full-time work (1/0)
- fatherHS: Whether the student's father completed high school (1/0)
- fatherBachelors: Whether the student's father obtained a bachelor's degree (1/0)
- fatherWork: Whether the student's father has part-time or full-time work (1/0)
- selfBornUS: Whether the student was born in the United States of America (1/0)
- motherBornUS: Whether the student's mother was born in the United States of America (1/0)
- fatherBornUS: Whether the student's father was born in the United States of America (1/0)
- englishAtHome: Whether the student speaks English at home (1/0)
- computerForSchoolwork: Whether the student has access to a computer for schoolwork (1/0)
- read30MinsADay: Whether the student reads for pleasure for 30 minutes/day (1/0)
- minutesPerWeekEnglish: The number of minutes per week the student spend in English class
- studentsInEnglish: The number of students in this student's English class at school
- schoolHasLibrary: Whether this student's school has a library (1/0)
- publicSchool: Whether this student attends a public school (1/0)
- urban: Whether this student's school is in an urban area (1/0)
- schoolSize: The number of students in this student's school
- readingScore: The student's reading score, on a 1000-point scale

Write a professional report detailing your analysis of the dataset including your efforts to…

a. Create a training and testing set using n-fold cross validation.
b. Perform appropriate univariate and bivariate analysis on the data.
c. Check for multicolinearity.
d. Create appropriate dummy variables.
e. Perform feature selection.
f. Check for appropriate second order terms.
g. Check for appropriate interaction terms.
h. Transform variables as needed.
i. Evaluate your final model as if for a data scientist.
j. Write a summary as if for a layman.