PSYC 560
Due: January 26, 2023

Assignment #1: Linear Regression

[Q1] The data file **druguse_training.csv** contains the following variables measured on 977 participants from American northwest urban areas (Duncan, Duncan, Alpert, Hops, Stoolmiller, & Muthén, 1994).

- Cig1 (Cigarette consumption: 1 = lifetime abstainers, 2 = 6-month abstainers, 3 = current use of fewer than four times a month, 4 = current use of between 4 and 29 times a month, 5 = current use of 30 or more times a month)
- Gender (male = 0, female = 1)
- Age
- SES (average of parental income and education level)
- Marital (Parental marital status: 0 = single, 1 = married or living in a committed relationship)
- Family (Family status: 1 = step or foster families, 2 = others)

Let us assume that cig1 is a continuous variable.

(1) Apply a multiple linear regression to examine the effects of gender, age, SES, marital, and family on cig1. Report and interpret the results. **(4 points)**
[1. Report and interpret R^2; 2. Report and interpret the coefficient estimates and their statistical significance at $\alpha = .05$]

(2) Add the quadratic term of age to the linear regression model considered in (1) and examine whether the quadratic term is statistically significantly associated with the dependent variable. (**2 points**)

(3) Apply a KNN regression to the same dependent variable, using gender, age, SES, marital, and family as predictors with $K = 5$. Calculate and report R^2. (**2 points**)

[Q2] Using the coefficients estimated from the linear regression models (1) and (2) and KNN regression (3), examine how well these methods predict new 227 participants' cigarette consumption levels in the data file **druguse_test.csv**. Report which method performs best in terms of the mean squared error (MSE) for the test data. (**2 points**)