Between 1995 and 1998, students in an introductory statistics class taught by Professor John Eccleston and Dr. Richard Wilson at The University of Queensland in Australia took part in a simple exercise experiment. First, the students were randomly assigned to either a sitting group or a running group. Then, all the students took his/her own pulse. Depending on the pre-assigned group, each student either sat for a minute or ran in place for a minute. After that, each student took his/her pulse again. We are interested in determining which variables explain the largest proportion in variability in the student's first pulse measurement.

Variable	Description
Height	Height (cm)
Weight	Weight (kg)
Age	Age (years)
Sex	Sex (1 = male, 2 = female)
Smokes	Regular smoker? (1 = yes, 2 = no)
Alcohol	Regularly consumes alcohol? (1 = yes, 2 = no)
Exercise	Frequency of exercise (1 = high, 2 = moderate, 3 = low)
Ran	Did the student run or sit between the pulse measurements? (1 = ran, 2 = sat)
Pulse1	First pulse measurement (beats per minute)
Pulse2	Second pulse measurement (beats per minute)
Year	Year of class (95 – 98)

Run a multiple linear regression with *pulse1* as the outcome variable and *height*, *weight*, *age*, *sex*, *smokes*, *alcohol*, *exercise* as the predictor variables and Conduct a hypothesis test to examine whether pulse1 is linearly associated with *height*, *weight*, *age*, *sex*, *smokes*, *alcohol*, and *exercise*. Write a full report.

A multiple linear regression analysis was used to test whether a subject's first pulse measurement was linearly associated with 7 predictor variables: height, weight, age, sex, smoking frequency, alcohol consumption frequency, and exercise frequency. The F-statistic was 2.153 with 8 and 74 degrees of freedom, and the resulting p-value was 0.04106. With a p-value less than the α=0.05 significance level, the null hypothesis of there being no linear association between first pulse measurement and the 7 regressors was rejected. The coefficient of determination was 0.1888, meaning that the predictors accounted for 18.88% of the variability in the first pulse measurements in this dataset.

Check normality assumption. Explain your reasoning and provide formal statistical test and visual evidence. (For simplicity, we will assume that all of the other regression assumptions are met.)

Graphical diagnostics were performed on the residuals to check the normality assumption. On the Q-Q plot, all the residuals fall within the 95% confidence interval of the $y = x$ line except for one. The histogram of residuals shows the residuals to mostly follow a normal distribution except for one point. From visual examination, it appears that the normality assumption has been met except for one observation.

A Shapiro-Wilk normality test was used to numerically test the normality assumption. The test-statistic was 0.8732 and resulting p-value was <0.0001. With a p-value less than the α=0.05 significance level, the null hypothesis of the residuals following a normal distribution should be rejected. However, visual examination showed that only one point is causing deviation

from normality so the linear model should not be abandoned. Further analysis should continue with the normality assumption.

Check the outliers. Conduct the analysis to examine whether there are any potential outliers. If so, please list their ID. Explain your reasoning and provide the evidence for your analysis.

Studentized residuals were calculated to identify potential outliers in the dataset. The threshold for rejection was 3.5859. Only one point was larger than that threshold, which was observation 47 with a residual of 6.4945. Observation 47 had an ID of 73 and first pulse measurement of 145.

Check for influential points. Conduct the analysis to examine whether there are any potential influential points. If so, please list their ID. Explain your reasoning and provide the evidence for your analysis.

Cook's distances were calculated to identify potential influential points in the dataset. According to the rule of thumb, if an observation's Cook's distance is greater than $\frac{4}{n}$, that point should be examined further. Observation 76 had a $D_i = 0.5327$, which is slightly greater than $D_i > 0.5$ so it may be an influential point. An F-test was conducted to examine each point's percent influence on the linear model's fit, no observations had significant influence past the α=0.05 level. The Cook's distance plot confirmed that only one point, observation 76, had a $D_i > 0.5$. Observation 76 had an ID of 102 and first pulse measurement of 88.

Check for leverage points. Conduct the analysis to examine whether there are any potential leverage points. If so, please list their ID. Explain your reasoning and provide the evidence for your analysis.

Leverage distances were examined to identify potential leverage points. According to the rule of thumb, if an observation's leverage distance is greater than $\frac{2p'}{n}$, that point should be examined further. Three observations surpassed that threshold, observation 76 with a leverage distance of 0.66, observation 62 with a leverage distance of 0.4494, and observation 56 with a leverage distance of 0.3399. Observation 76 had an ID of 102 and first pulse measurement of 88. Observation 62 had an ID of 88 with a first pulse measurement of 74. Observation 56 had an ID of 82 with a first pulse measurement of 52.

Based on your analysis above, are there any data points concerning you? If so, what is your suggestion for the next step.

Based on my analysis, observation 47 is concerning to me because it's an outlier with such a high residual. Since only one outlier was identified, I would not worry too much about it because this is a large dataset so one observation wouldn't have a large influence the fit of the linear model.

The influential point identified, observation 76, did not have significant percentile influence on the linear model's fit, so I would not worry about any influential points.

Lastly, I would examine the three leverage points, observations 76, 62, and 56, more carefully to see how they affect the linear model's fit.