**PROJECT FOR MODULE 3 (REGRESSION)**

**[75 Total Points]**

**MARINE BIOLOGY**

You have been hired for a project as a statistical consultant for a renowned national marine research center to help analyze data related to dolphin longevity and health. The research center has been tracking a number of dolphins in the wild for several years and has provided you with the attached data set. This data set contains two variables that have been assigned to dolphins based on their most recent physical examinations: longevity (a health index based on a physical examination of the dolphin, 100 = maximum possible value) and IQ (measured via a series of validated tests performed by a scientist). The set also contains two additional variables: a close relationships index (higher scores indicate the dolphin has formed one or more close relationships with other dolphins) and a social integration score (higher scores indicate a dolphin is more socially active within its pod). These variables have been calculated using analysis of the GPS tracking data of the dolphins over the same time period. **You have been asked to investigate creating a model for the longevity score based on the other variables provided**. If a reliable model could be created, this could help marine biologists use GPS data to identify dolphins with higher health risks, and could also help extend the reach of research dollars by potentially helping scientists reduce their experimental costs of performing physical examinations of dolphins in the wild.

1. **(8 Points)** Provide appropriate data visualizations showing the relationship between longevity and each of the other variables in the analysis. Characterize the relationships using appropriate statistics.
2. **(10 Points)** Create a linear regression model of longevity based on the independent variable that has the greatest correlation with longevity.

Write a paragraph explaining the model (be sure to include an appropriate data visualization, the model equation, and coefficient of determination for the model).

1. **(8 Points)** How confident can we be in this model? Asked another way, perform a hypothesis test at an appropriate level of confidence to confirm that the coefficient of the predictor variable is nonzero.
2. **(9 Points)** Confirm that any assumptions required for a good linear regression model hold true in this case. Indicate if there are any leverage points or outliers in the data set that should be concerning?
3. **(12 Points)** Confirm that a linear model was a good choice for this variable by providing data visualizations of a quadratic and a cubic regression model based on this one predictor variable. Which model seems to have the best fit statistically?
4. **(10 Points)** Now, create a multivariate linear regression model of dolphin longevity using all of the data available to you. Characterize the fit of this model to the data using the appropriate statistic(s). Does it improve on the model you created for question 2?
5. **(9 points)** For each of the three predictor variables, indicate if you can be 95% confident that the variable’s coefficient is nonzero (i.e. that it should in fact be included in the model). Provide the p-value for each test/variable.
6. **(9 points)** Based on your answer to the previous question, create one final multivariate model of longevity, including only the predictor variable(s) that you believe should be in the model. Provide the model equation below, provide relevant statistics about the model fit, and briefly discuss the fit in comparison to the other models you’ve created.