**Instructions:**

**There are a total of four (4) multi-part questions, with point values noted for each question.**

**Please show your calculations, or the details of your program(s) for each problem. You must supply the SAS program, and the program should be commented so that each step is clearly explained.**

**Combine all your answers/files into a single zipped file and post the zipped file to “HW\_Midterm” in CANVAS.**

1. Using the “Baseball” dataset in CANVAS, we are trying to predict home runs based on the other numerical variables in the data set. Perform PCA analysis on other numerical variables. How many components should be extracted? Why? (25 points)
2. Using the “Cereals” data set, we are interested in predicting nutritional ratings based on sodium content. (10 points)
3. Construct the appropriate scatter plot.
4. Perform the appropriate regression analysis.

1. For the “Depression” dataset, in CANVAS, predict the reported level of depression as given by “Cat\_total”, using:
2. “Income”, “Sex” and “Age” as independent variables. Analyze the residuals and decide whether or not it is reasonable to assume that they follow a normal distribution.
3. Identify 15 observations that appear to be most influential. State your Criteria. Delete these points and repeat the regression.
4. Identify 15 observations that have the most leverage on the model. State your Criteria. Delete these points and repeat the regression.

(45 points)

1. Using distributed SQL, calculate the average “Age” of the “Spanish Bank” customers. (20 points)