Taylor Stevenson
MSDS 422
Assignment 1: Allstate Claims Severity EDA

1. Provide appropriate descriptive statistics and visualizations to help understand the marginal distribution of the dependent variable.

-The first graph displayed in Section 1, is a histogram of the dependent variable values. This distribution has a right-hand tail. This means it is positively skewed and this is also known based on the calculation for its skewedness, which was 3.79. There is a notable difference between the mean and median. The mean will fluctuate depending on outliers present.

2. Investigate missing data and outliers.

-There is no missing data. Using the function .isnull(), the data set was checked for any null values in any of the columns.

-To check for outliers, a boxplot graph was generated. It appeared that there were some outliers existing for every independent variable.

3. Investigate at least three potential predictors of the dependent variable and provide appropriate graphs / statistics to demonstrate the relationships.

-Choosing the best predictors initially seemed simple in that they would be the most strongly correlated with the dependent variable. This was not the case in this dataset, as it is still unclear as to what the cont# label truly represents.

4. Engage in feature creation by splitting, merging, or otherwise generating a new predictor.

-By generating scatterplots of every independent variable against the dependent variable, it appeared that cont7 and cont11 were the best predictors to merge. This was done by using the concatenating the scatterplots together. These predictors were chosen based on their strong positive correlations.

5. Using the dependent variable, perform both min-max and standard scaling in Python.

-The dependent variable was scaled to min-max and standard criteria. Above both of these newly scaled sets, is the original trainDf.