**Analysis / Model Methods**

* Picked variables based on relevance and fewest null values
* Pulled picked variables from full dataset
* Plotted variables on a histogram to check for normality
  + Indicated that two variables needed normalizing (SUM\_GISArc, SUM\_TotalV)
* Reduced the data set to complete cases (no nulls)
* Logged the variables and created new columns for them
* Split data for sampling (CE\_y vs CE\_n)
* Randomly pull out 50-100 CE and 500-1000 NCE for test set
* Remaining CE\_y for the training set
* Randomly pull 10x NCE for the remainder of the training set
* Fit model on training set
  + Completed a likelihood ratio to test to test univariate models against the null (does adding variable x improve model fit?)
  + Completed a variance inflation factor to detect any collinearity
    - Age and children present score were highly correlated
    - removed children present to remove collinearity
  + Completed a backwards stepwise model selection by Akaike Information Criterion (AIC)
  + Selected final variables from backwards stepwise results
* Added fitted values to train set data to plot

