**Data Analytics Assignment 3**

1. a) The two variables I chose were ‘Impressions’ and ‘Clicks’. These variables both have a relatively low range, although impressions does have outliers in every data set. Impressions consistency has a first quartile value of 3 and a third quartile value of 6, as well as a median of 5. However, Clicks has first and third quartile values of 0, indicating that there are very few data points that are nonzero.

b) I completed Anderson-Darling normality tests for all of the datasets with the Impressions and Clicks variables. Every test I ran returned a p-value less than 0.05, indicating that the distributions were not normal. The distributions across data sets was very similar. The Impressions variable appears to have a Poisson distribution with a lambda value of 5. The Clicks variable appears to have a Poisson distribution with a lambda value between 0 and 0.5. Both of these variables are discrete in all of the nyt datasets.

c) The ECDFs of the Impressions and Clicks variables across datasets look similar in that they both begin at 0 and rise sharply. In other words, they look normally distributed but skewed right. This indicates similarity to a Poisson distribution with a lambda value less than 10. When plotted using a quantile-quantile plot against the Poisson distributions with lambda values of 5 and 0.25 respectively, both datasets seemed to correlate with the Poisson distributions very well, as the datapoints were very close to the diagonal line.

d) I chose to use the Kolmogorov-Smirnov test to compare the datasets with the Poisson distributions. The Impressions variable fit the Poisson distribution with a lambda of 5 nicely, as the p-value from that Kolmogorov-Smirnov test was approximately 0.9. However, the Kolmogorov-Smirnov test for the Clicks variable with the Poisson distribution did not result in a p-value above 0.05, which indicates that the null hypothesis is true and the distributions do not fit each other.

e) The data in this set was limited overall. I felt that age had the most broad range of values, but then comparing age to any other variable was difficult using a boxplot since the range of age was so much greater than the other three. Since impressions and clicks had similar central tendency measures, I chose to compare those two.

The graphs shown on the next page are examples of ECDF and QQ-plots I used, but since the variables were similar across datasets, I did not provide them all. I also did not provide the box plots since they were all very similar. However, I attached my code to the assignment which generates all of them.

