Data Analytics Assignment 3 Report

Part 1.

a). Create boxplots for Age and Impressions

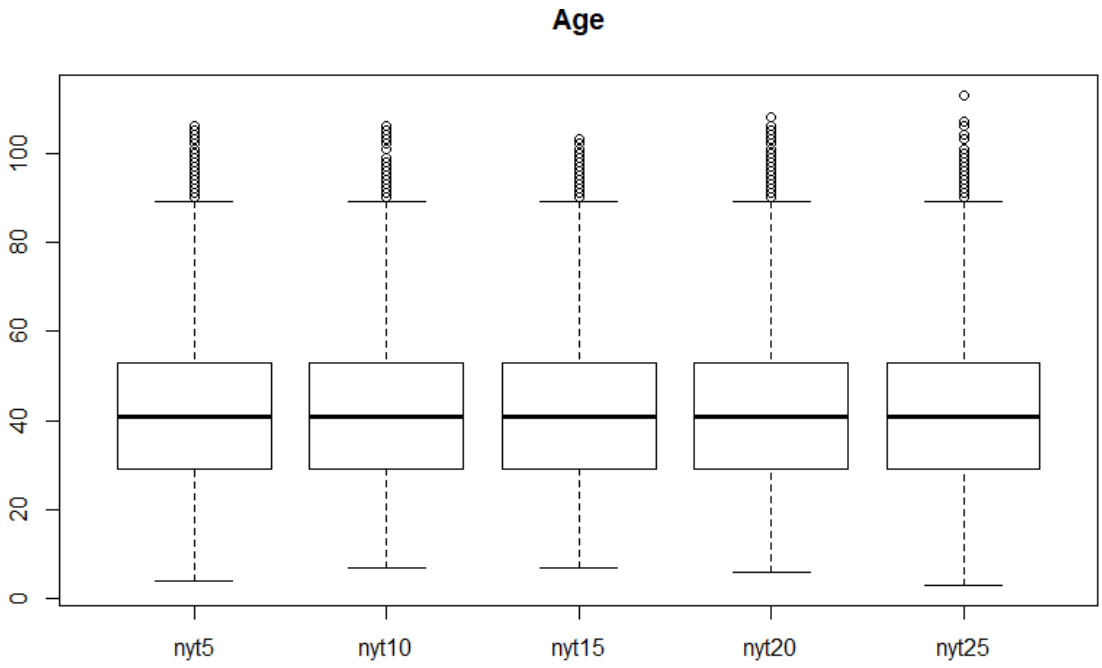

Figure 1. Age Boxplot

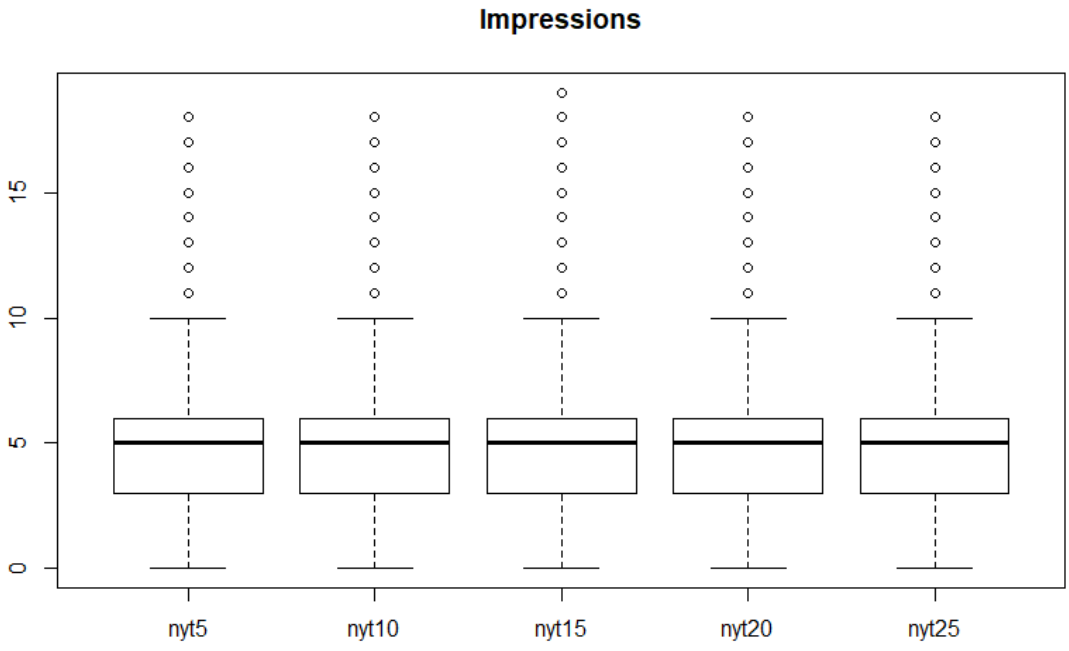

Figure 2. Impressions Boxplot

I chose to analyze variable *"Age"* and *"Impressions"*. As shown in the figures above, the age distributions and the impressions' distributions across different New York Times data files are similar. Figure 1 suggests that the median age of NYT subscribers is around 40 with multiple outliers at above their age of 90. Figure 2 indicates that the median number of impressions for a reader is 5, while there are also outliers having more than 10 impressions.

b). Create histograms for Age and Impressions

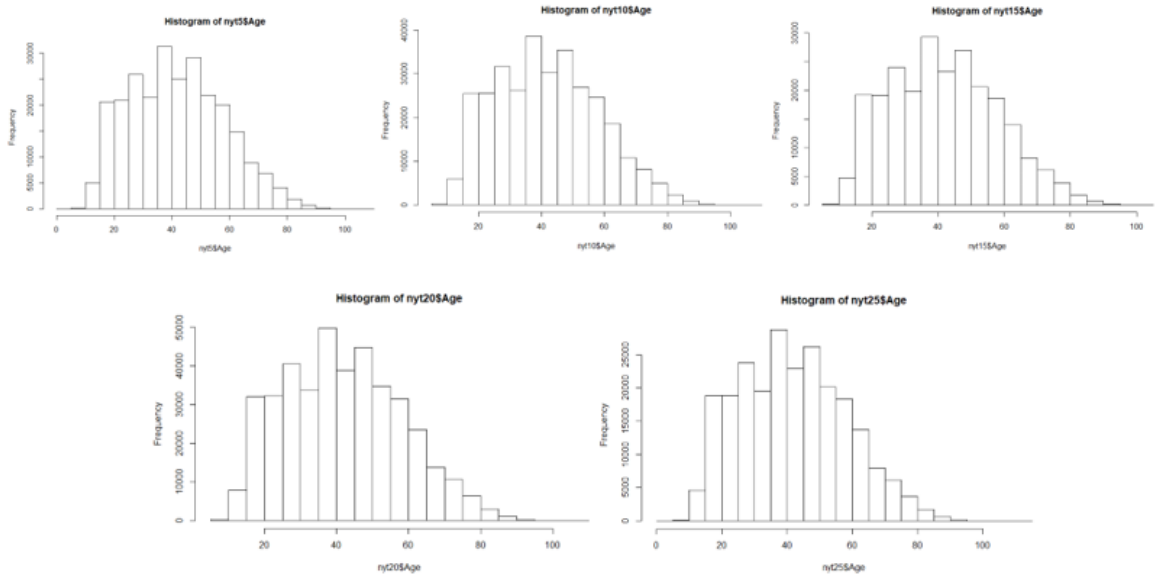

Figure 3.Histograms for Age

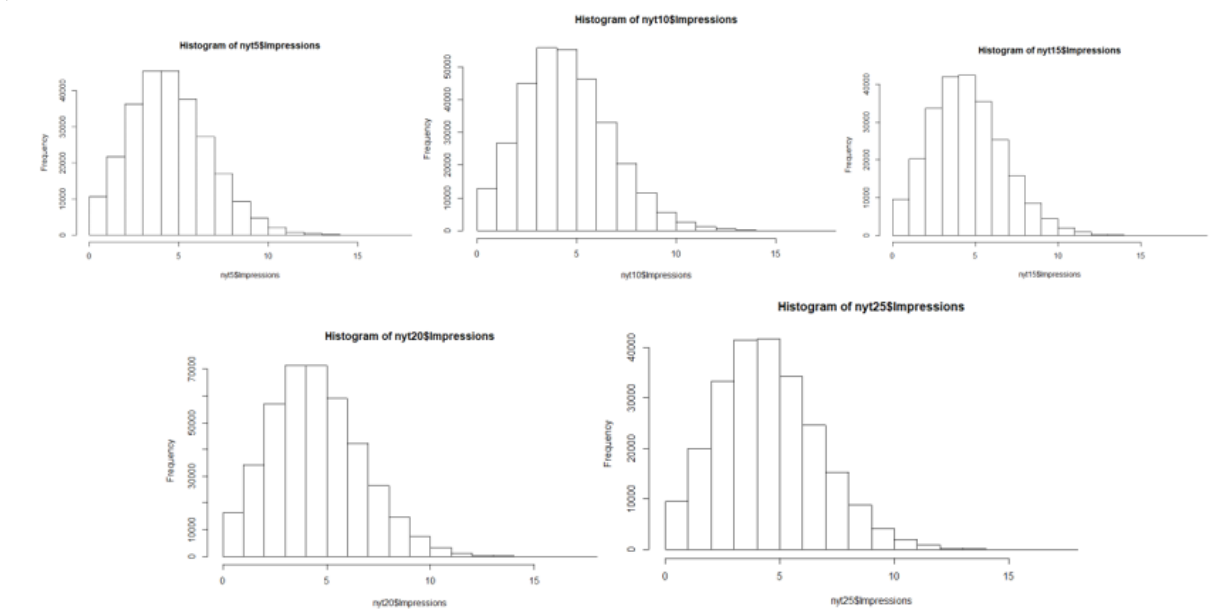

Figure 4. Histograms for Impressions

As observed in Figure 3 and 4 above, all the distributions are skewed right by a little bit. The histograms clearly indicates that both *"Age"* and *"Impressions"* are discrete right-tailed distributions similar to Poisson distributions. However, *"Age"* seems to have wider bodies than *"Impressions"* on average, it should be fitted to a higher lambda than *"Impressions"*. In this case, I suggest that *"Age"* is a Poisson distribution with lambda of 8, and *"Impressions"* is also a Poisson distribution but with a lambda of 4, denoted $Pois(8)$ and $Pois(4)$.

c). Plot the ECDFs and Q-Q plots

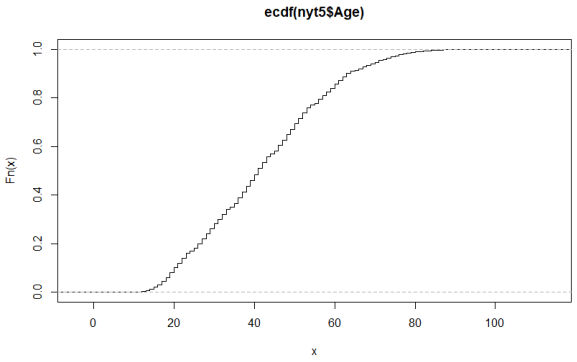

Figure 5. ECDF of Age

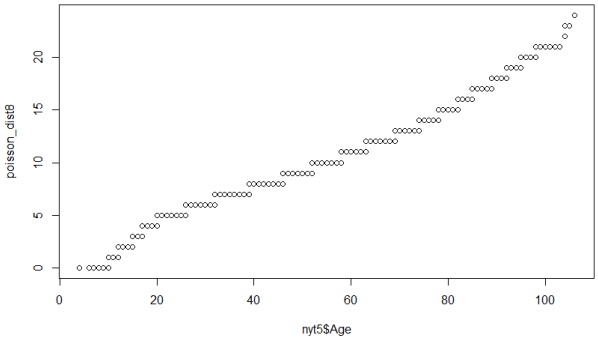

Figure 6. Q-Q plots of Age and Poisson

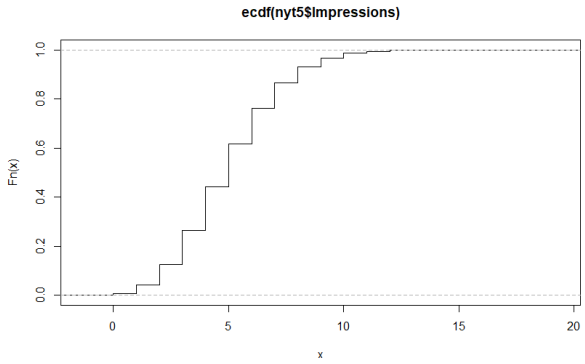

Figure 7. ECDF of Impressions

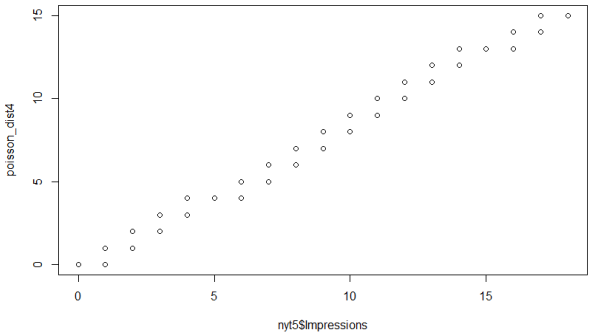

Figure 8. Q-Q plots of Impressions and Poisson

As shown in the ECDF plots above, the curves look like staircases as expected since both *"Age"* and *"Impressions"* are discrete numeric variables. The Q-Q plot between *"Impressions"* and $Pois(4)$ is very straight, indicating that "Impressions" is indeed very similar to a Poisson distribution with a mean of 4. However, the Q-Q plots of *"Age"* is curved at both tails, meaning that the extreme values in *"Age"* cannot be well described as in $Pois(8)$ and it could be a hybrid of several more distributions.

d). Perform significance test on Impressions

Null Hypothesis: The number of impressions has no effect on the number of clicks.

Alternative: There exist some correlation between the number of impressions and the number of clicks.

```
        F test to compare two variances

data:  lm(nyt5$Clicks ~ nyt5$Impressions) and lm(nyt5$Clicks ~ 1)
F = 0.98615, num df = 258975, denom df = 258976, p-value = 0.0003868
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9785824 0.9937748
sample estimates:
ratio of variances
         0.9861493
```

Figure 9. F-test between Impressions and Clicks

According to Central Limit Theorem, we can assume the distributions of variables are normal as the data size being large. So, in this question, I can perform a F-test between the linear model “Clicks” vs “Impressions” and a model of “Clicks” that only consists of intercept. The resulted p-value is smaller than 0.01, meaning that the variances of the two models are significantly different. Hence, we have sufficient evidence to reject the null hypothesis, and claim that there indeed exist some linear correlation between the number of impressions and the number of clicks.

Part 2.

For this part, I used Gender variable to filter out only the female user data.

b). Create histograms for Age and Impressions

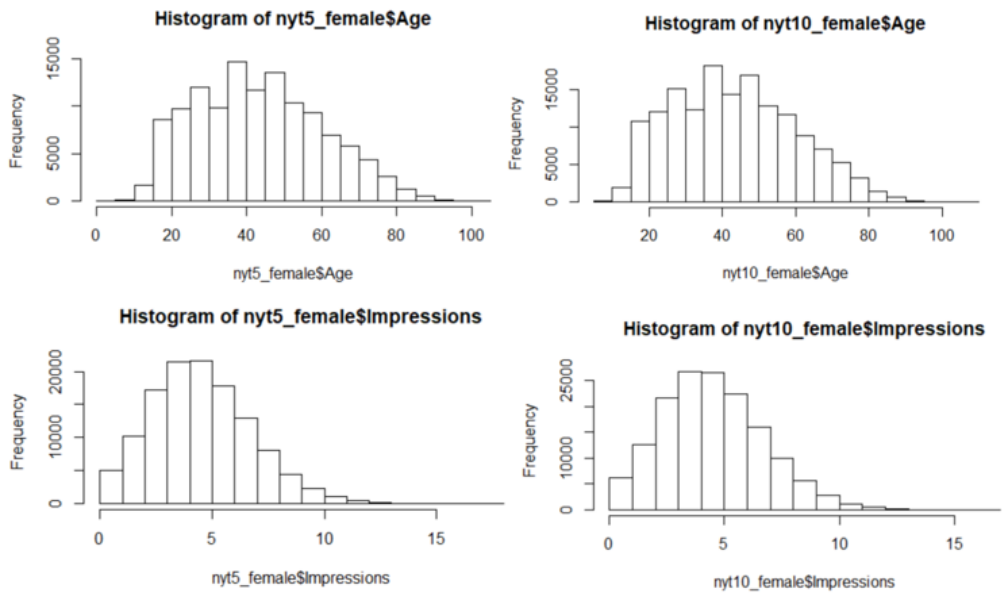

Figure 10. Histogram of female

c). Plot the ECDFs and Q-Q plots

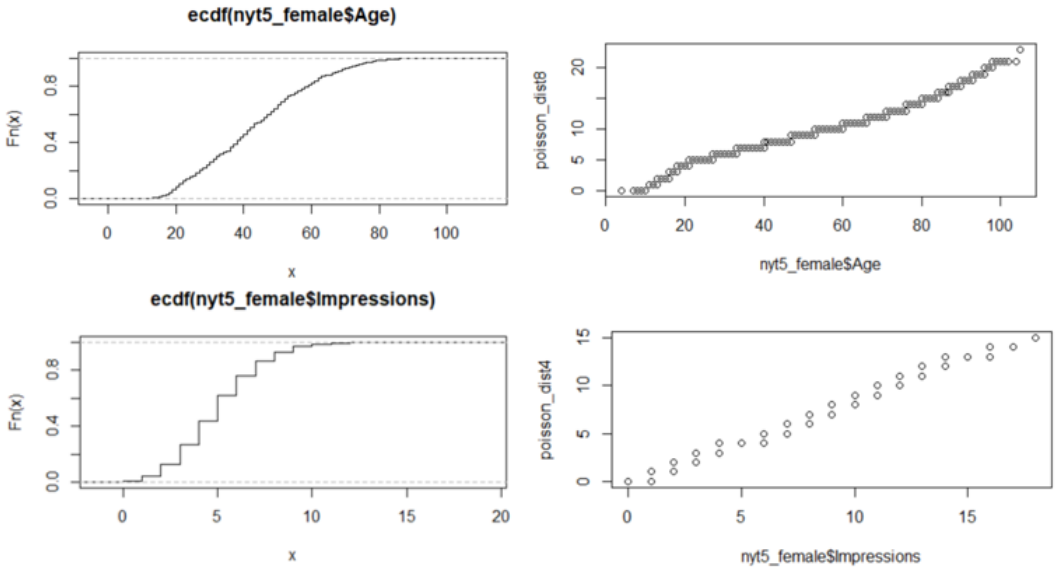

Figure 11. ECDF and QQ plots for female

d). Perform t-test between

For this question, I will be testing whether the distributions of the number of clicks are different for male and female.

Null hypothesis: the distributions of the number of clicks are the same for male and female.

Alternative: the distributions of the number of clicks between male and female are different.

```
nyt5_male <- filter(nyt5, nyt5$Gender == 1)
nyt5_female <- filter(nyt5, nyt5$Gender == 0)
t.test(nyt5_male$Clicks,nyt5_female$Clicks)
```

Figure 12. Code snippet for performing t-test

```
        Welch Two Sample t-test

data:  nyt5_male$Clicks and nyt5_female$Clicks
t = -4.0772, df = 254276, p-value = 4.559e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.006390779 -0.002241254
sample estimates:
 mean of x  mean of y
0.06869675 0.07301277
```

Figure 13. t-test result

e). Conclusion:

As shown in Figure 10 and 11, the distributions of "Age" and "Impressions" of female are very similar to that of male. The *"Impressions"* variable follows Poisson distribution while *"Age"* comes from a combination of several distributions.

Even though the variation in *"Age"* and *"Impressions"* between female and male are similar, the result from a t-test shown in Figure 13, with a p-value smaller than 0.01, suggests that we are more than 99% confident that the average number of clicks between men and women are different. On average, holding the number of impressions constant, a female tends to click 0.0043 more times than a man, meaning that New York Times should invest more on adding impressions to female to achieve higher click rates.