Data Analytics Assignment 3 Report

**Part 1.**

**a).** **Create boxplots for Age and Impressions**

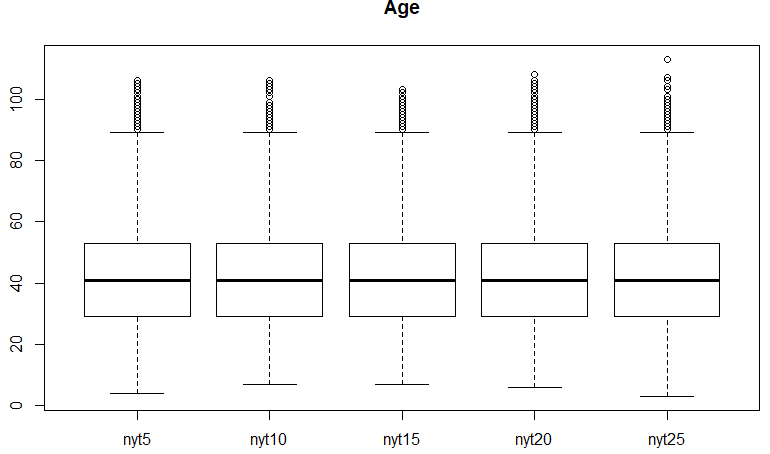


Figure . Age Boxplot

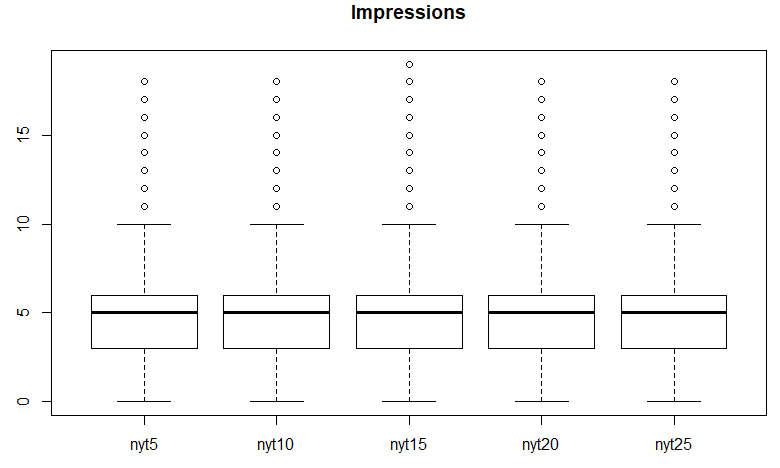


Figure . Impressions Boxplot

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**

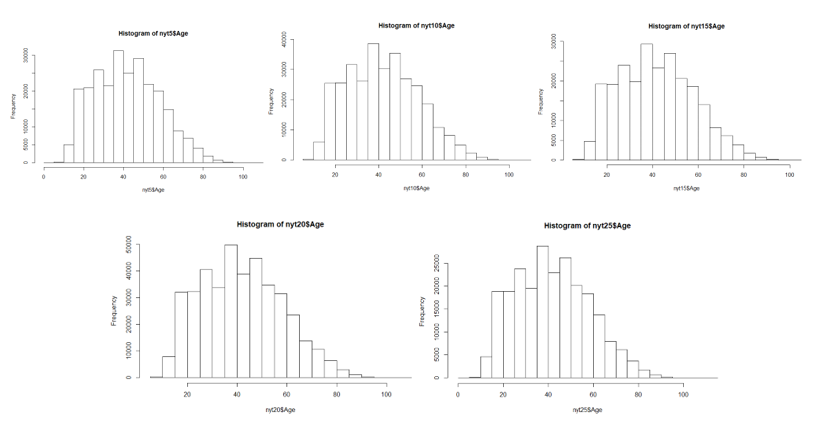


Figure .Histograms for Age

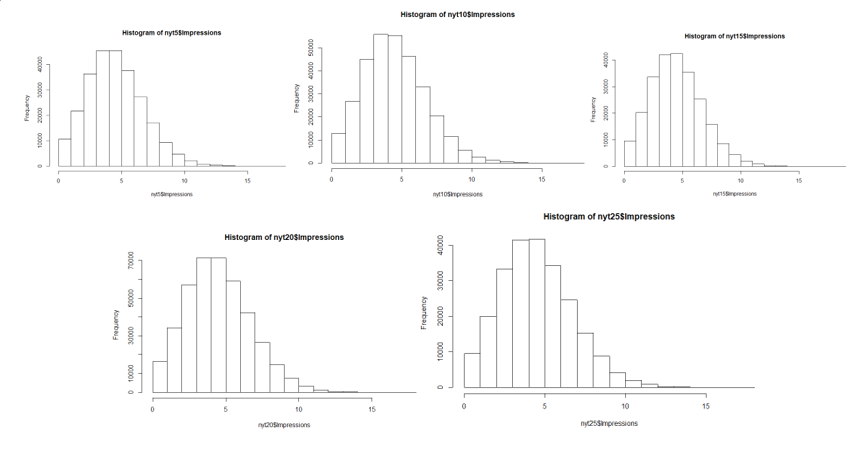


Figure . Histograms for Impressions

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 and .

**c). Plot the ECDFs and Q-Q plots**

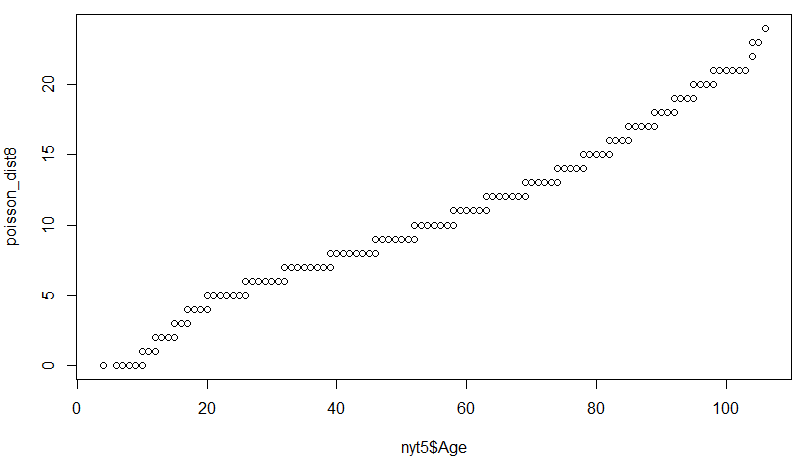
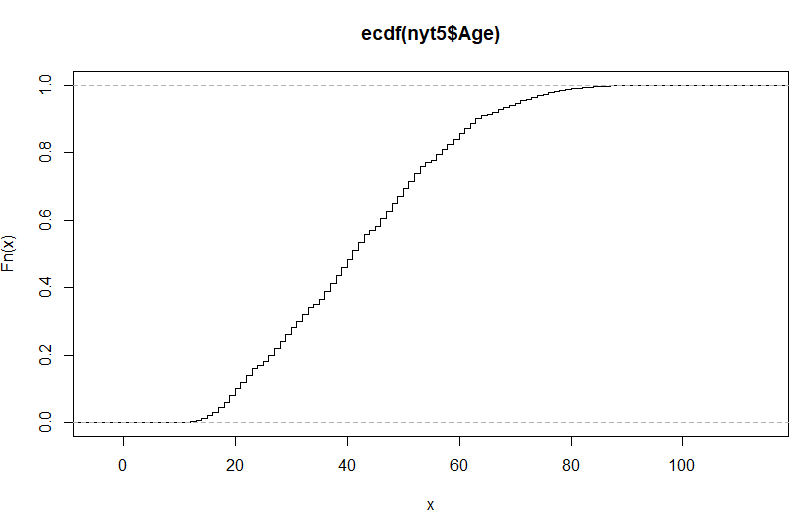


Figure . ECDF of Age Figure . Q-Q plots of Age and Poisson

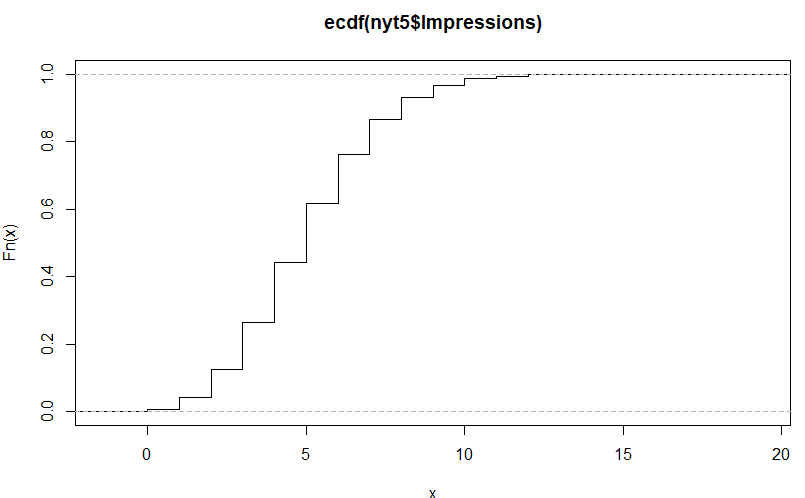
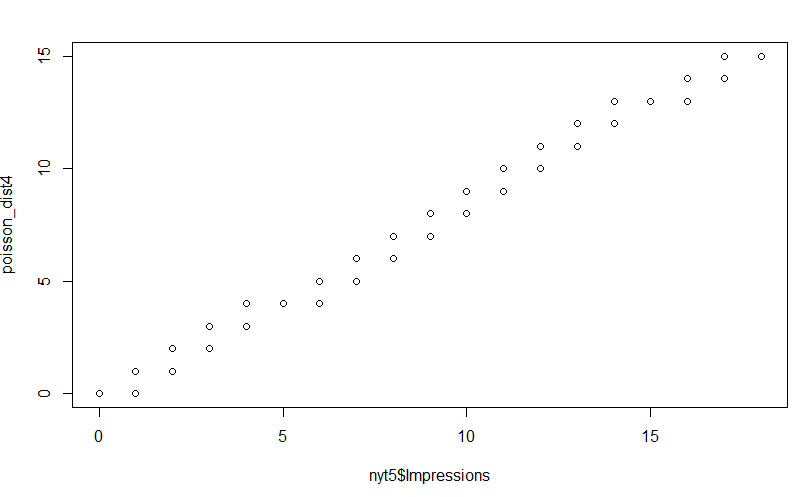
 

Figure . ECDF of Impressions Figure . Q-Q plots of Impressions and Poisson

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 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 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.

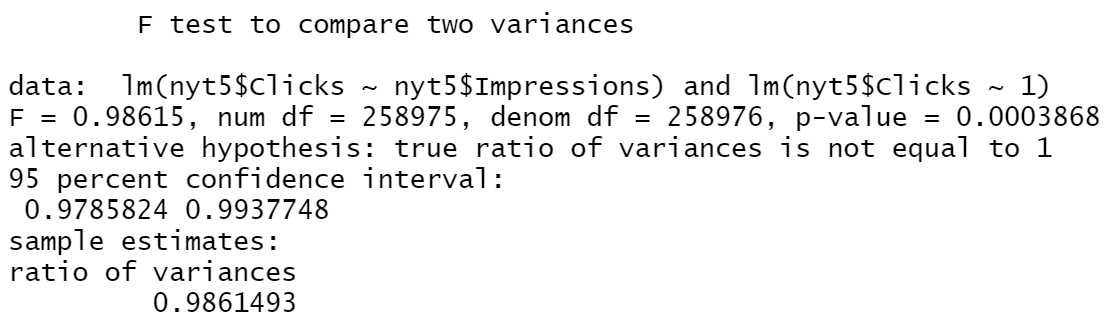


Figure . F-test between Impressions and Clicks

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**

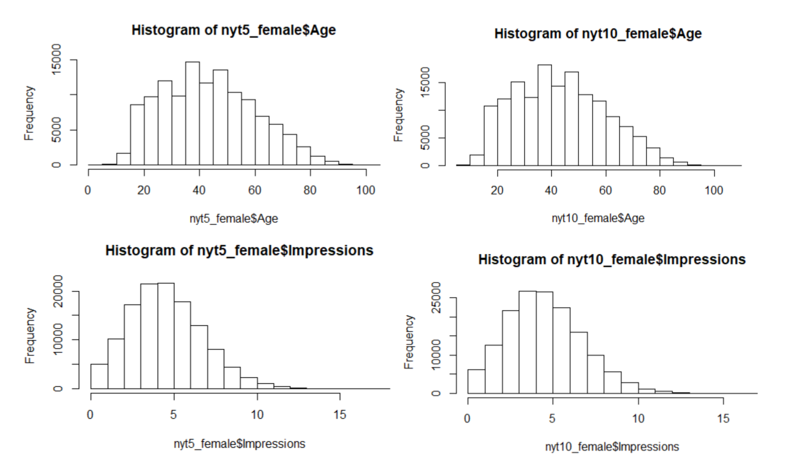


Figure . Histogram of female

**c). Plot the ECDFs and Q-Q plots**

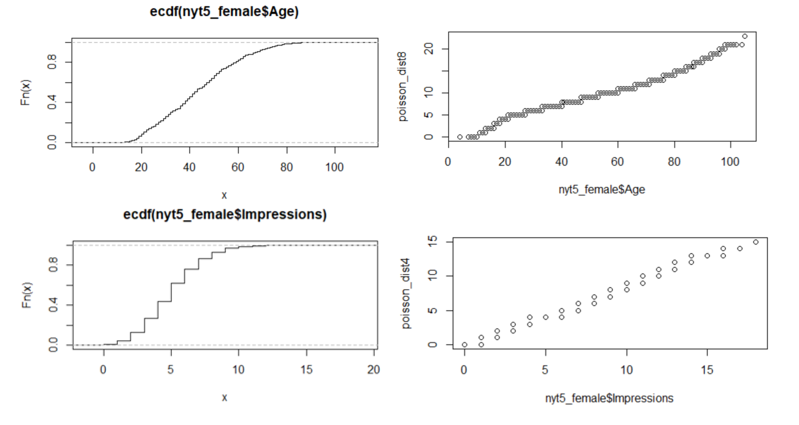


Figure . ECDF and QQ plots for female

**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.

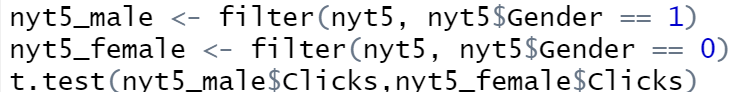


Figure . Code snippet for performing t-test

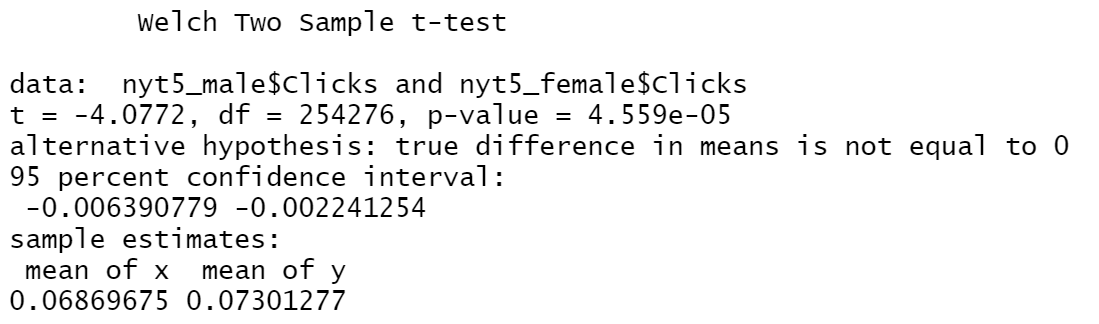


Figure . t-test result

**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.