Tony Min
Data Analytics
October 17, 2023

Assignment 3

Choose NYT datasets: 7, 13, 14, 15, 17, 23, 29

Part (a)

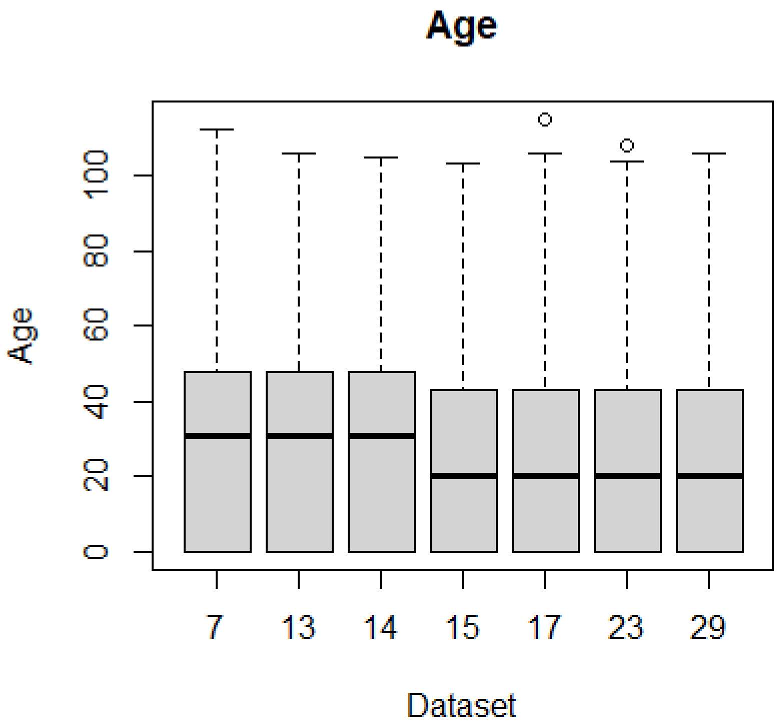


Figure 1: Age Boxplot

From the boxplot above we see that the ages within these data sets were very similar in terms of distributions. Datasets 7, 13, 14 had a higher average of age compared to datasets 15, 17, 23, 29. We had some outliers in dataset 17 and 23. We also see how the datasets were all skewed towards 0 this may be since people who did not fill out the age section in a form.

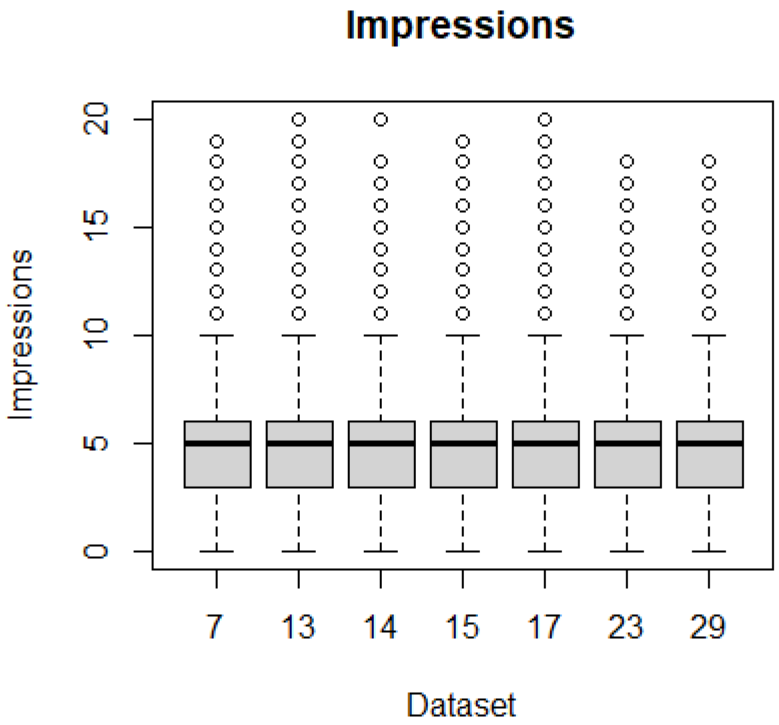


Figure 2: Impressions Boxplot

For the boxplots above we see how the distribution throughout all the datasets were all the same. It seems that all the datasets have many outliers and in dataset 13, 14 and 17 had an outlier at 20 impressions. The median for all the datasets were the same at 5 with also the same first and third quartile.

Part (b)

While conducting normality tests we used the Anderson Darling Test. Using this test for Age we use the test for all ages and then only ages greater than 0 then we also conducted the test on all the impression data. Our null hypothesis for this test would be that the data does follow a normal distribution. If our p-value was below 0.05 we would reject the null hypothesis and conclude that the data does not follow a normal distribution. The results of this test were that none of the p-values for any of the datasets came to be greater than 0.05 therefore we reject the null hypothesis. Meaning all the data sets, age, age without zeros, and impressions, did not follow a normal distribution.

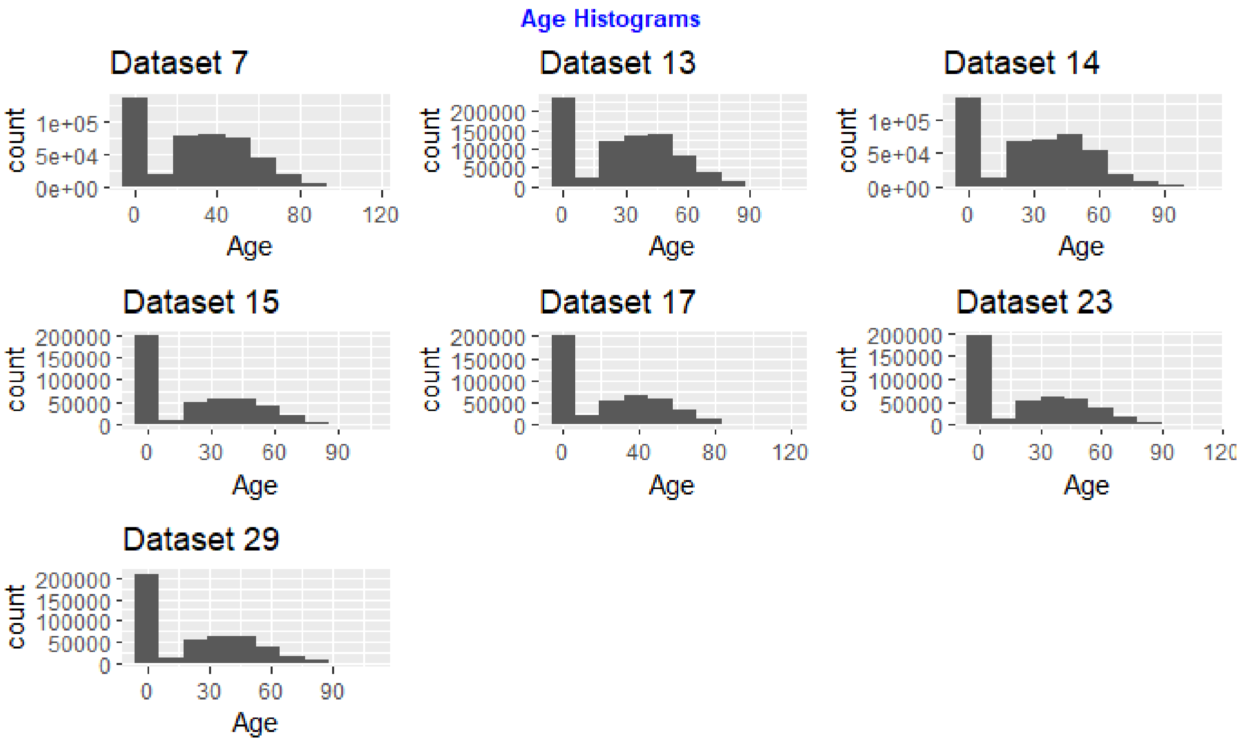


Figure 3: Age Histogram

As we see in Figure 3, the histograms above are all skewed towards 0, this again is probably due to the users who didn't fill out their age. It seems like the average age was around 40 to 50 years old throughout all datasets. We see in dataset 14 that we had some ages that were around or above 90. Below we will get rid of the data where the age was 0 and keep only data points where the age is greater than 0. As shown below we see more clearly that average age is in the 40-50 range.

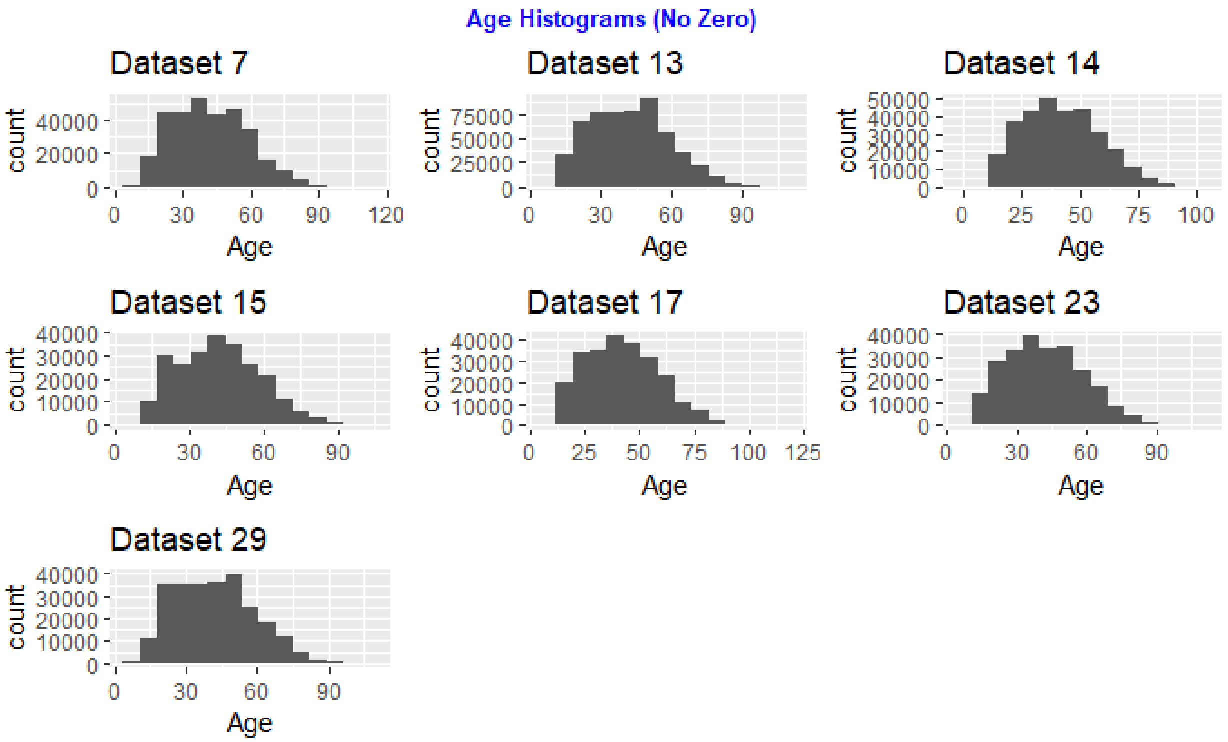


Figure 4: Age Histogram (Without Age 0)

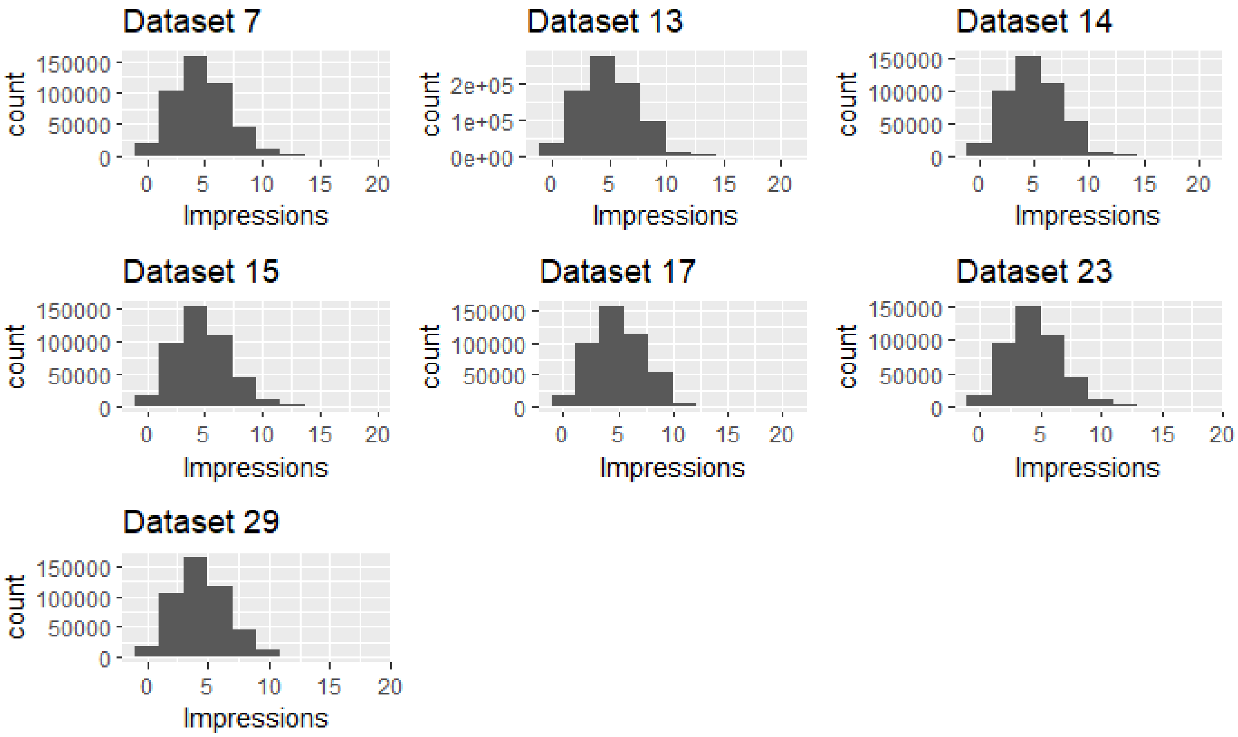


Figure 5: Impressions Histogram

As shown above in Figure 5, the number of impressions seemed to have averaged was 5. All datasets seemed to have to have similar distributions. It seemed like on average that no datasets went over 15 impressions so it was skewed towards the right. In the comparison from our histogram to our boxplot we see that it makes sense how there were impressions over 15 since the maximum value on the boxplot was 10 with a couple outliers above that point.

Part (c)

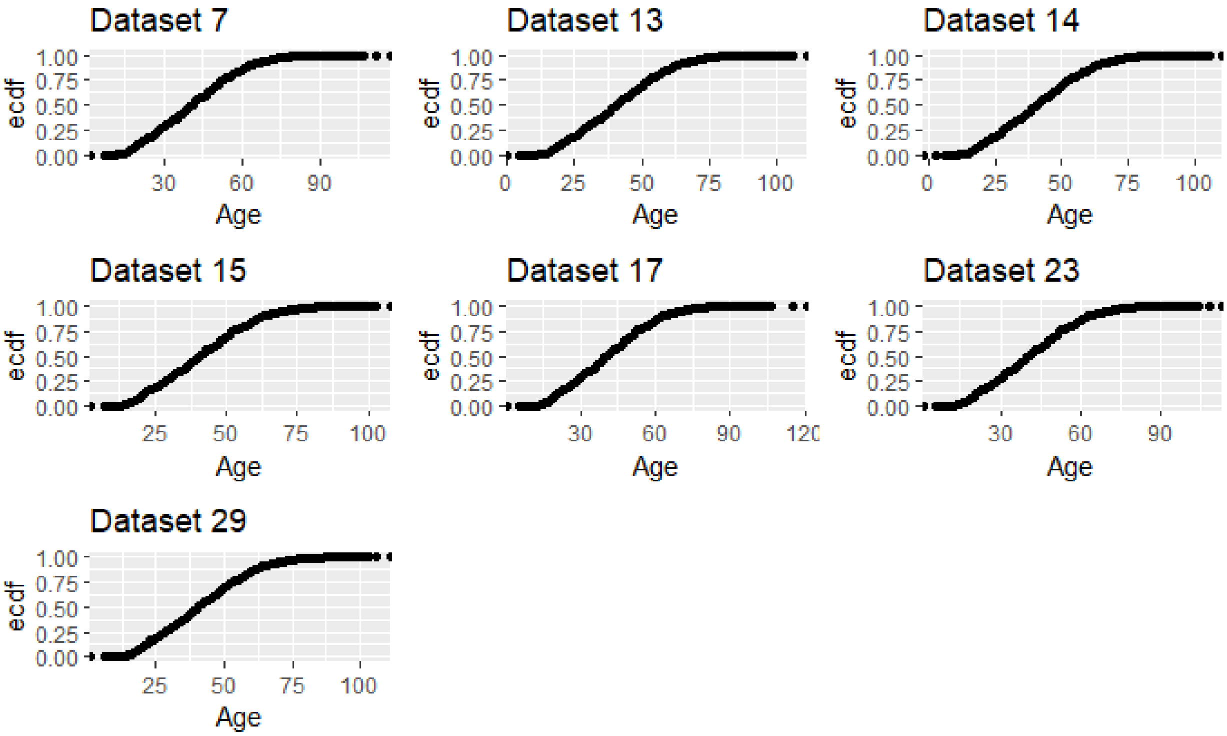


Figure 6: Age ECDF

From Figure 6 above, it seems that the curve is close to perfect. Additionally to being almost perfect we see there are few gaps in between points except for data set 17 there is a big gap. This may have happened since looking at the plot the x-axis goes all the way to up to 120 while most of the other datasets go only to 100 years. From looking at dataset 17 Figure 1 we see it had an outlier above 100 which may be the cause of this gap between points for the ecdf graph.

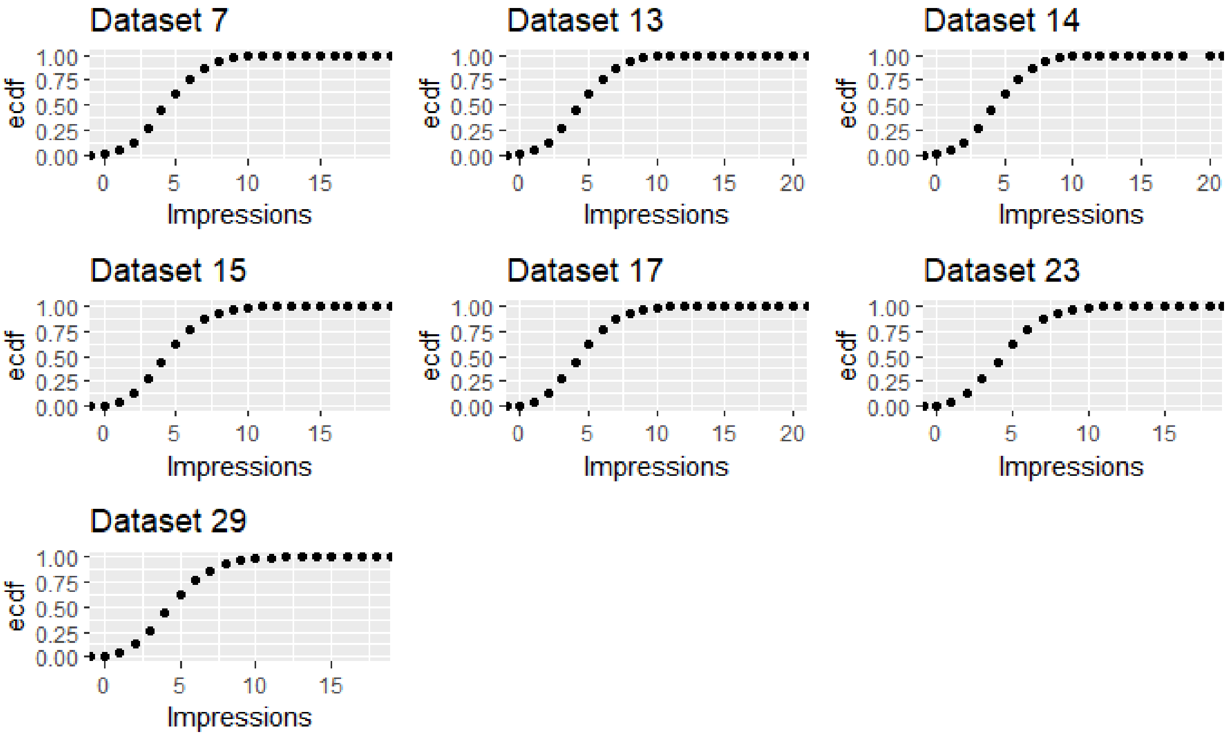


Figure 7: Impressions ECDFs

From Figure 7 the impression ECDFs for each datasets we see how the tail of the graph is longer on the right side. This is because we are reaching 10+ impressions which were not as prevalent in the datasets. The longer tail on the right side led to the graph to be skewed to the left. This makes sense since we learned before that there weren't many data points after a threshold of 10 impressions so at 10 impressions nearly all the data points have less than 10 impressions.

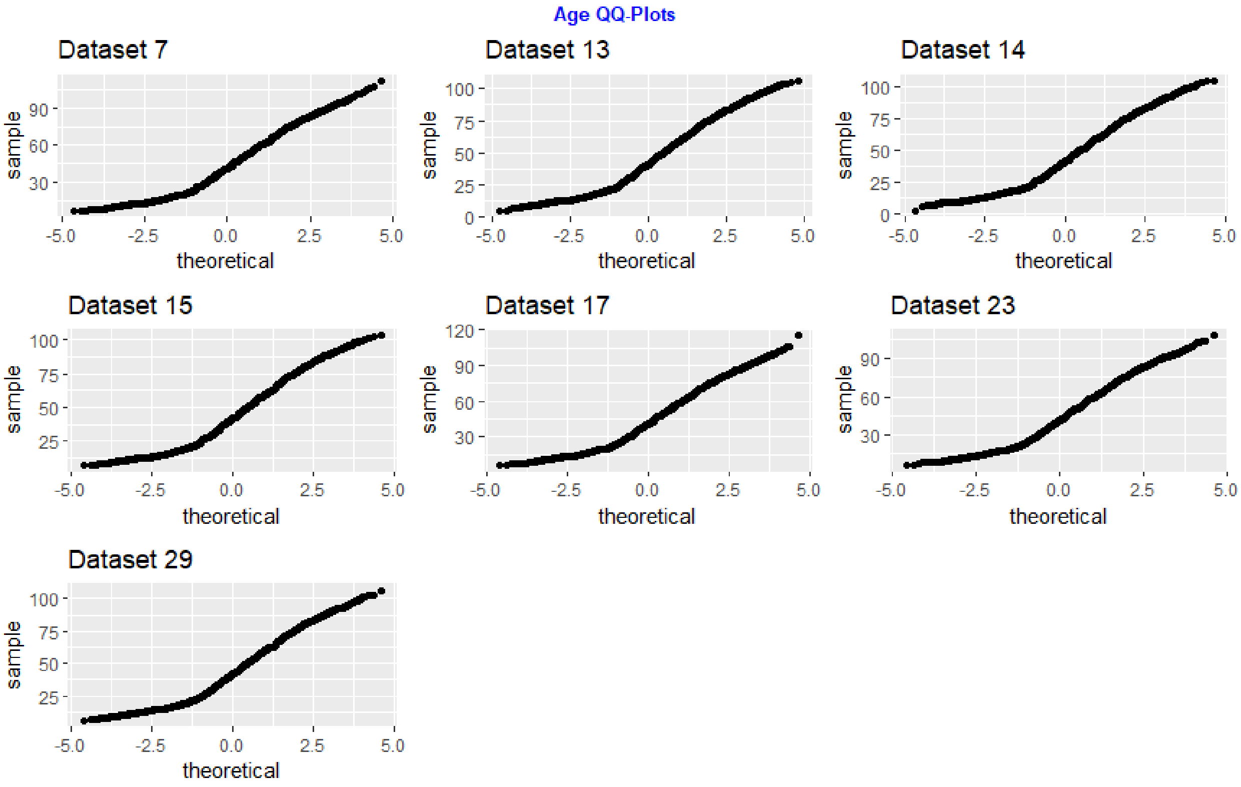


Figure 7: QQ-Plot for Ages

As we see from these qq plots it does not fit a normal distribution. It seemed to have an elbow at around an age under 25 on all datasets. This makes sense since in the figure above from Figure 1 we see that the median age is around 20-30 therefore most of the data is there.

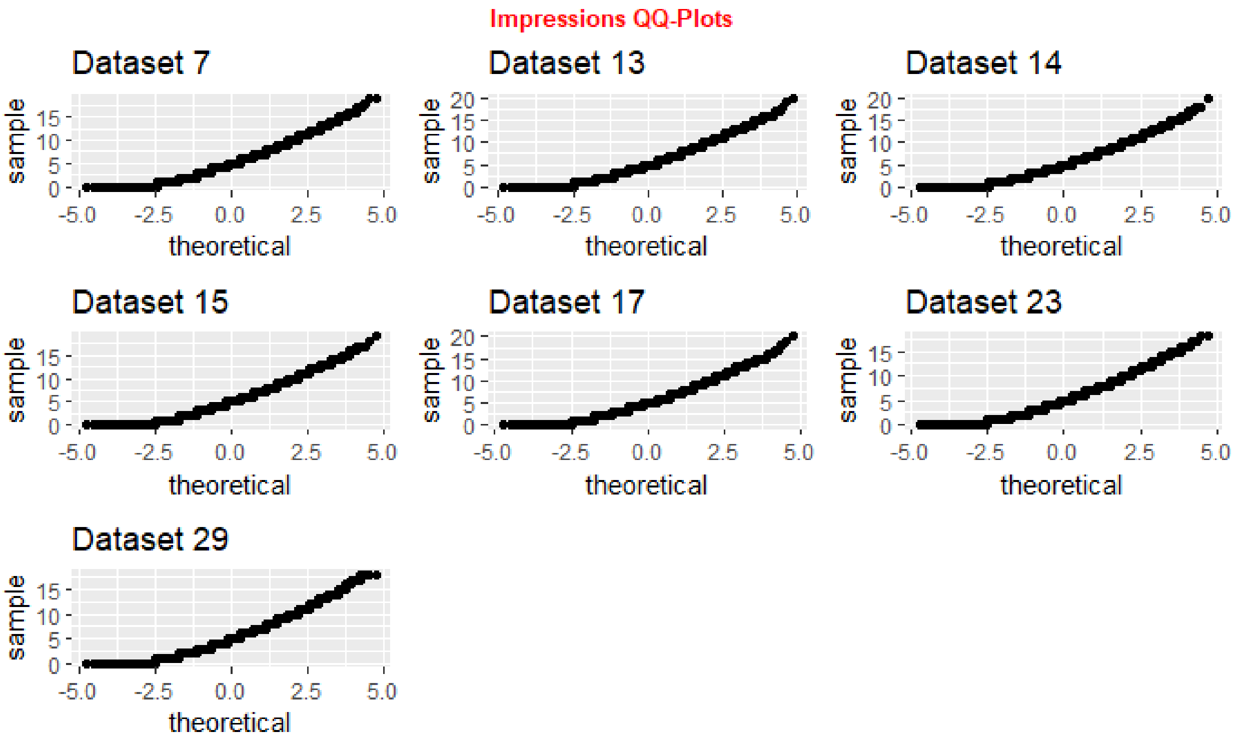


Figure 9: QQ-Plot for Impressions

From these qq-plots we also see that the data does not follow a normal distribution line. We don't see as big of an elbow as the qq-plots of age but we have a similar type of curve. We see how it starts going up at 5 which matches the histogram in Figure 2 as the median was also 5.

Part (d)

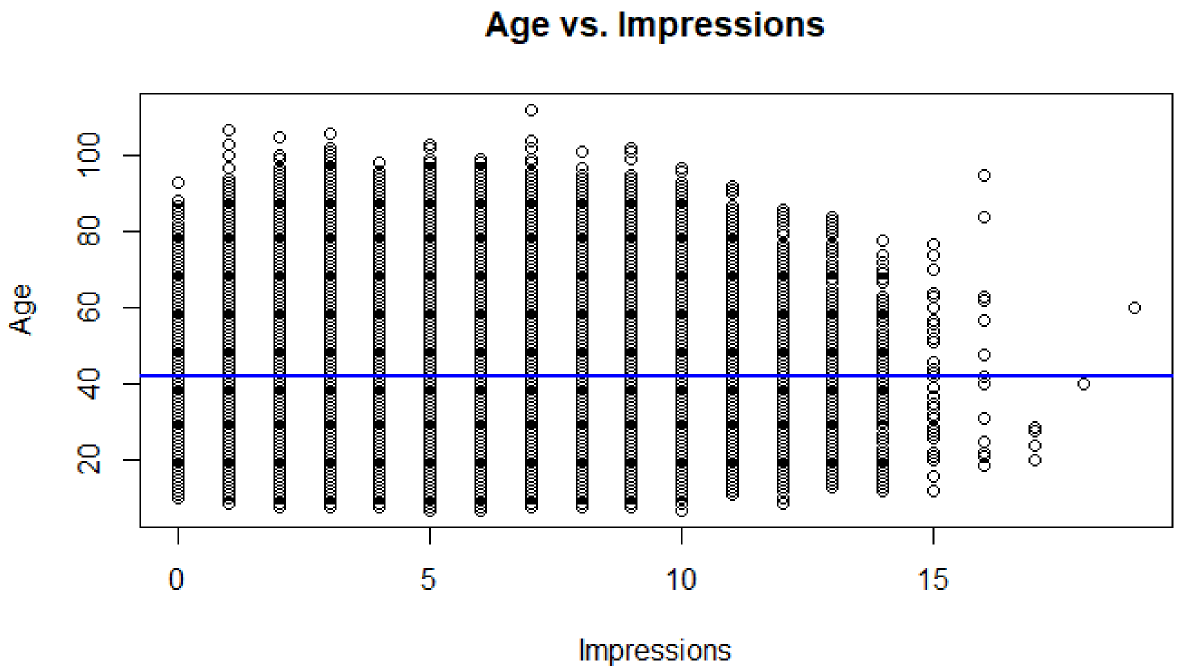


Figure 10: Linear Regression Age vs. Impressions

The linear regression equation was
Age = 42.08 + .016 (Impressions)
By using a linear regression model with age as the independent variable and impressions as the dependent variable. From the graph we see that there was no correlation between age and impressions a user has. As the slope of the line (0.016) was so low we see that the number of impressions does not affect the line that much, it keeps it quite straight. From the figure we see the data points plotted and that the data does not follow the linear regression line.

Part (e)

Observation

In looking at these datasets we see how similar each dataset that were chosen was. Which make sense since the data came from the same region. The datasets all had five primary variables: age, gender, impressions, clicks and signed-in status. One thing we saw was the amount of 0s in the age variable, we can assume that these values came from individuals that did not fill in the age section. Another variable that was slightly unclear was that the gender values were 0 and 1 which doesn't help us in signifying if someone was female or male.

Question 2: (6000 Level)

Part (b)

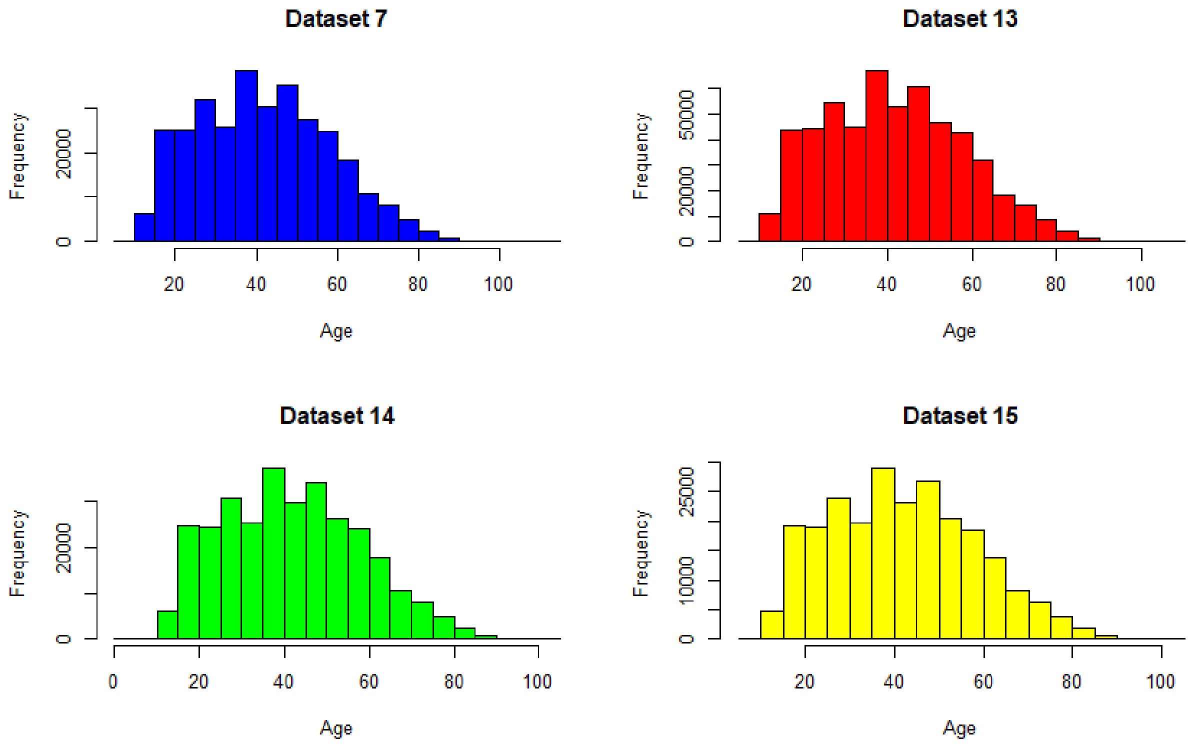
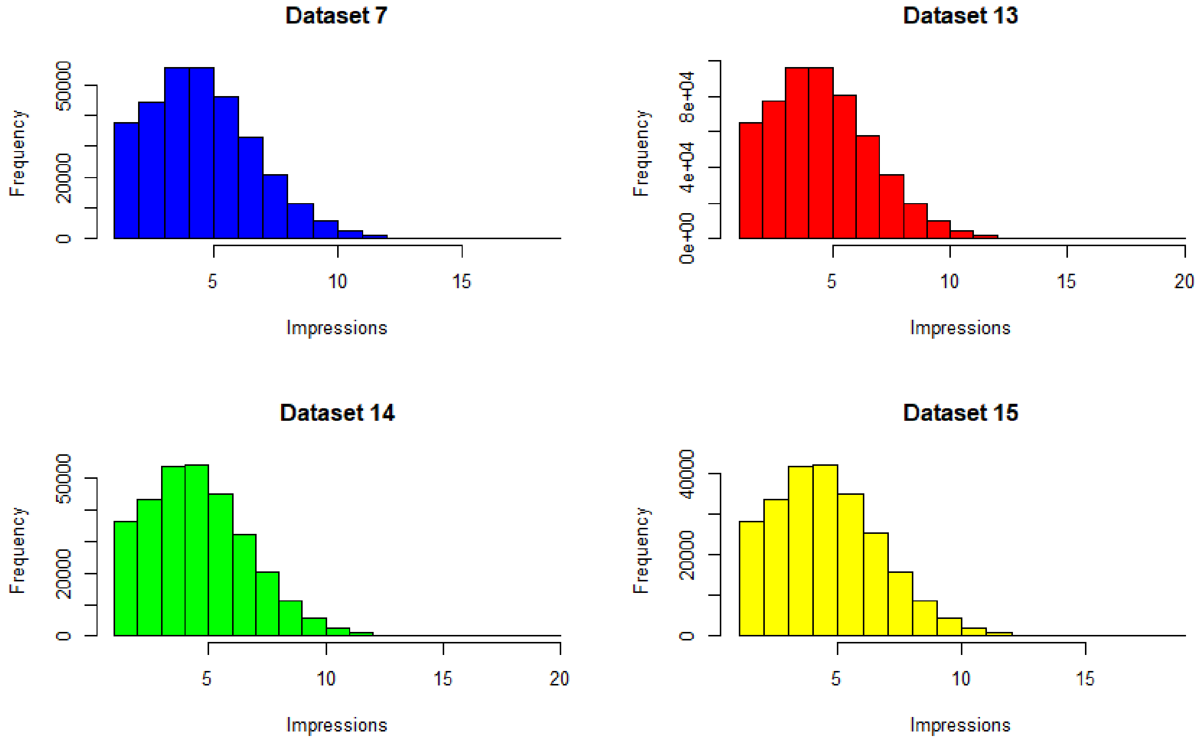


Figure 11: Histogram of Age (Filtered)

We took datasets 7, 13, 14 and 15 for this part of the assignment. We have filtered the data into that only had users whose age was greater than 0 and their impressions were greater than 0. From the histograms we see that it is not normally distributed and more skewed towards the left. This makes sense in the scope of things since users who would be filling out this form would be on the younger side instead of the older adults. Therefore we see more data points towards the left of the age range, also looking at the graphs where the highest frequencies occur at ages of 35-40, 45-50 and 25-30 in that order. I find it interesting that people aged 80+ could have filled out the form as well and the youngest age is around 10-15 years old.


Dataset 7

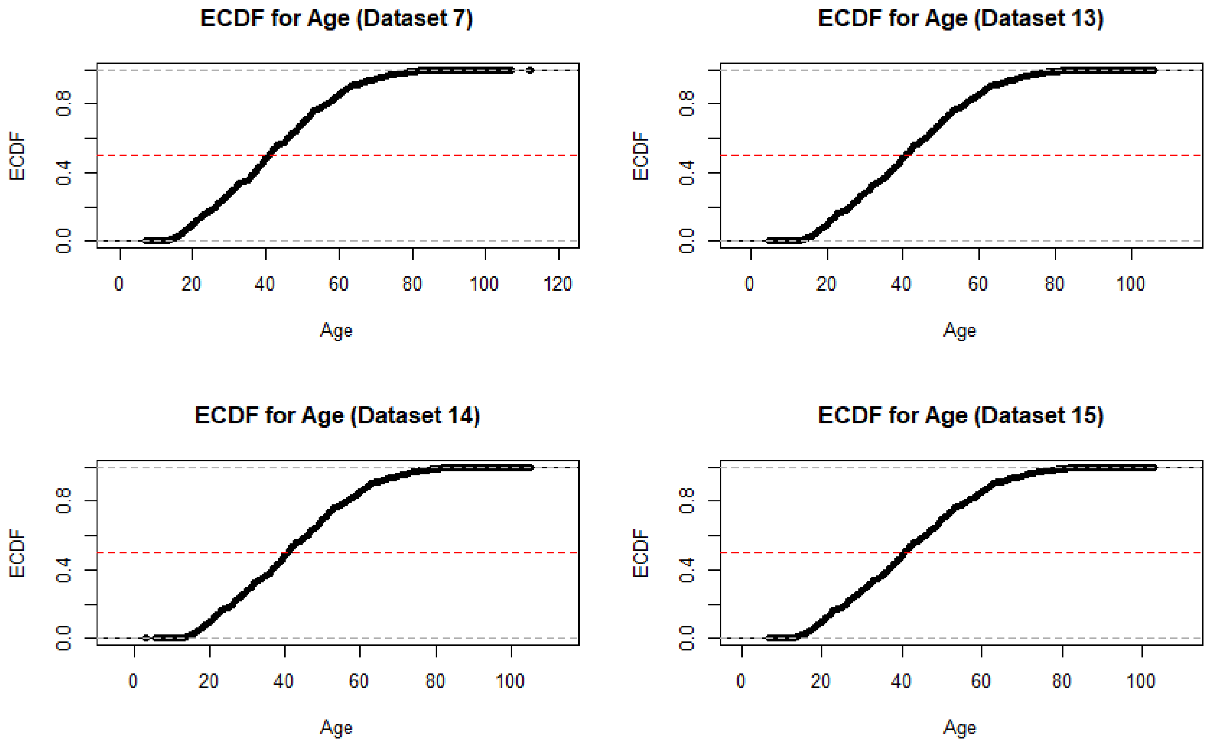
Dataset 13

Dataset 14

Dataset 15

Figure 12: Histogram of Impressions (Filtered)

Now with our new dataset we can clearly see from Figure 12 that the data is not normally distributed. We see that the data is very skewed to the left with both the mean and the median being around 5 impressions. We see that there is a minimum of 1, since the dataset only has impressions and then a maximum of 12 impressions in this dataset.


ECDF for Age (Dataset 7)

ECDF for Age (Dataset 13)

ECDF for Age (Dataset 14)

ECDF for Age (Dataset 15)

Figure 13: ECDF of Age (Filtered)

Using the Anderson Darling test we found p-values less than 0.05 for all the datasets for Age and Impressions. Therefore we are rejecting the null hypothesis such that we have come to the conclusion that the data does not follow a normal distribution. From Figure 13 we see that it does not follow a normal distribution. Looking at the red line (50 percentile) we see that where it meets the black points / line it skews towards the left instead of the middle.

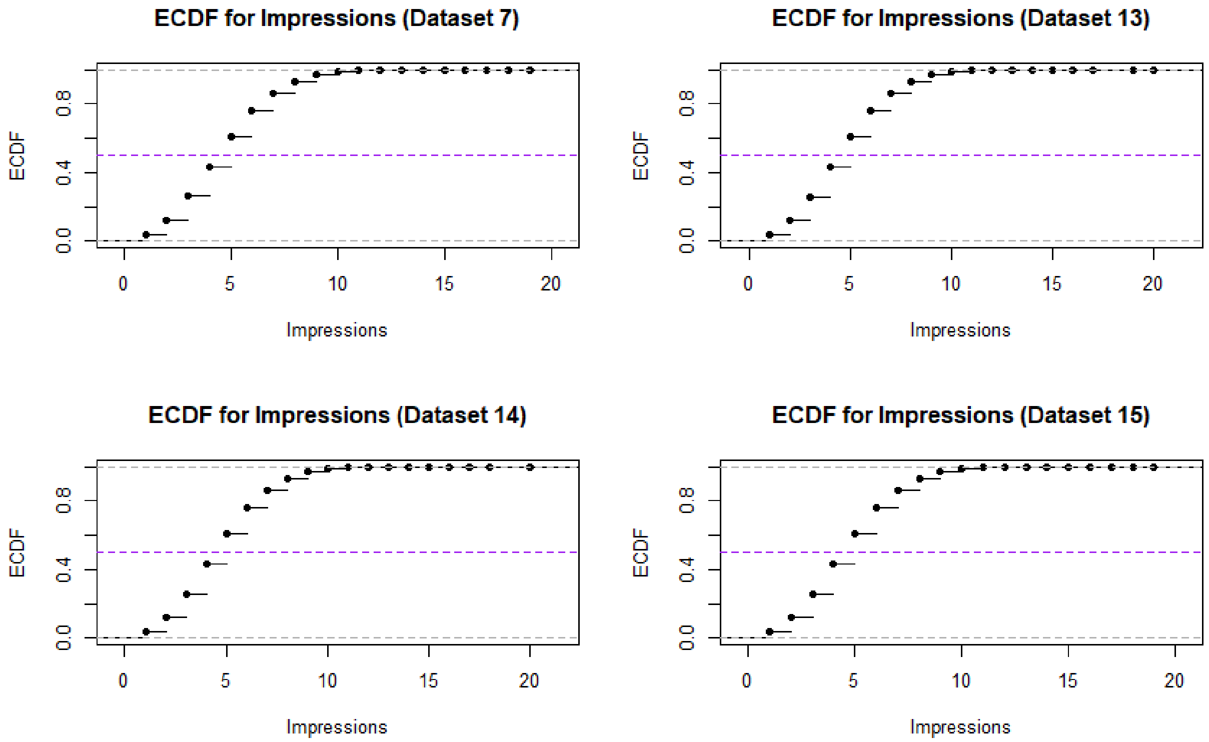
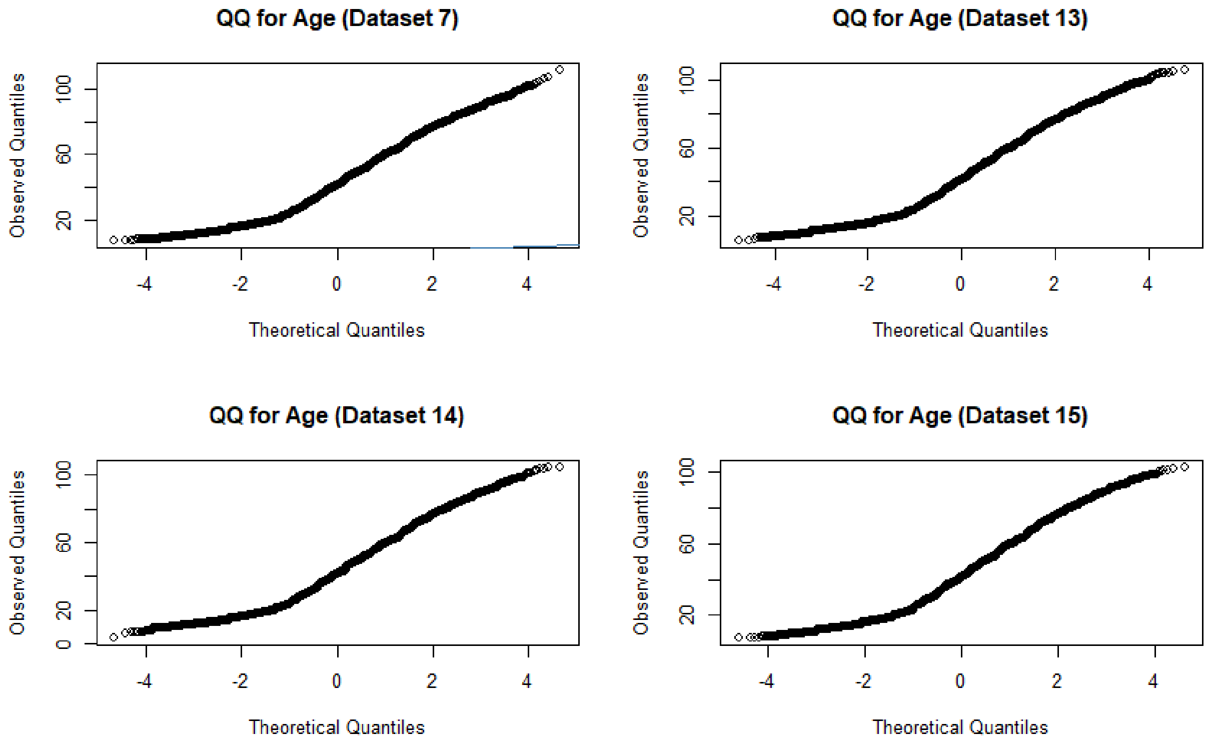


Figure 14: ECDF of Impressions (Filtered)

From Figure 14 we see how the datasets are even more skewed towards the left. Looking at the purple line (50 percentile) we see how it ‘hits’ the data points around 5 instead of 10, towards the middle. From this we see that the impressions from all datasets do not follow normal distributions. Most of the data has been explained before 10 impressions as seen in Figure 14 due to the data points hitting the ceiling of the graph at 10 impressions.


QQ for Age (Dataset 7)

QQ for Age (Dataset 13)

QQ for Age (Dataset 14)

QQ for Age (Dataset 15)

Figure 15: QQ-Plot for Age (Filtered)

To generate the QQ-Plot for Age we first normalize the data and then we plot it against the observed values. With the normalized data we see that the elbow of the data points were skewed to the left of 0 and at around 20 for the observed quantiles. We see that in Figure 15 we do not follow a straight line hence proving that the Age from all the datasets did not follow a normal distribution.

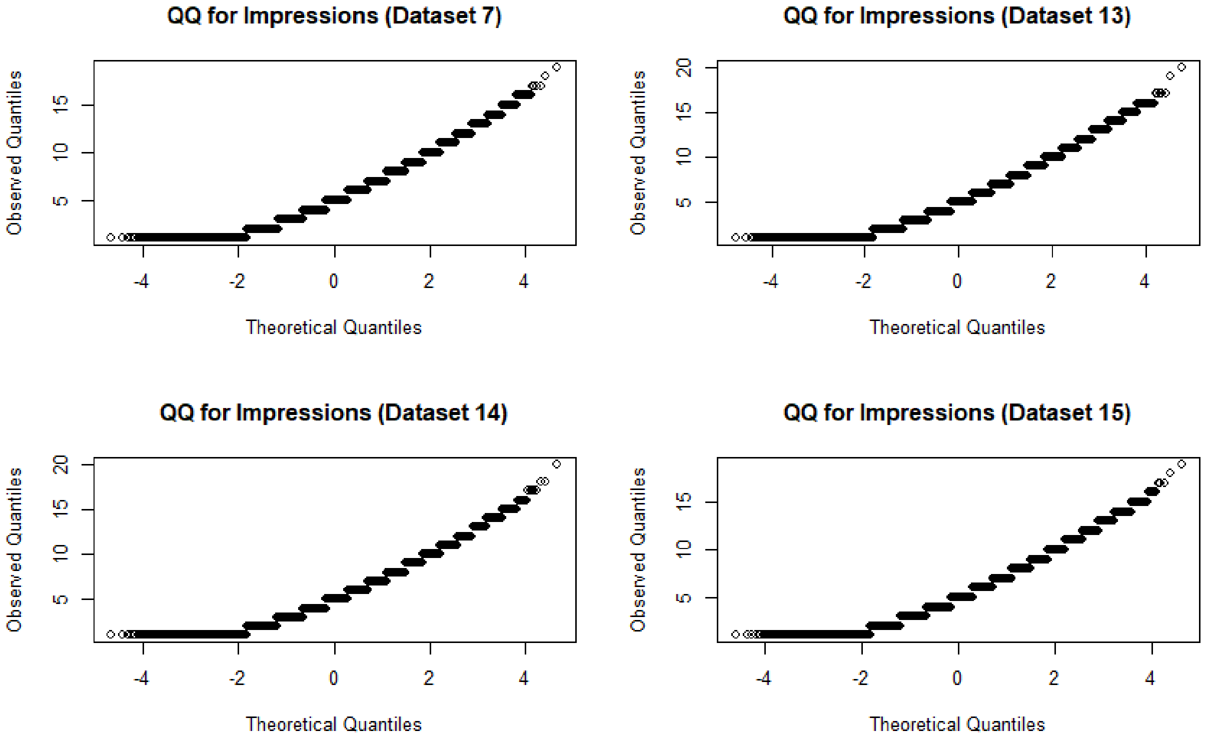


Figure 16: QQ-Plot for Impressions (Filtered)

Looking at Figure 16, we see that the impressions from all the datasets also do not follow a normal distribution. Comparing the QQ-plot for impressions and the QQ-plot for ages we see how they both are not straight lines but impressions have a flatter part at the beginning compared to ages but they all have elbows in the plots. Alongside these elbows they all seem to be skewed to the left from the midpoint of the plots. From these observations and the normality test we have proven that these datasets for Age and Impressions do not follow a normal distribution.

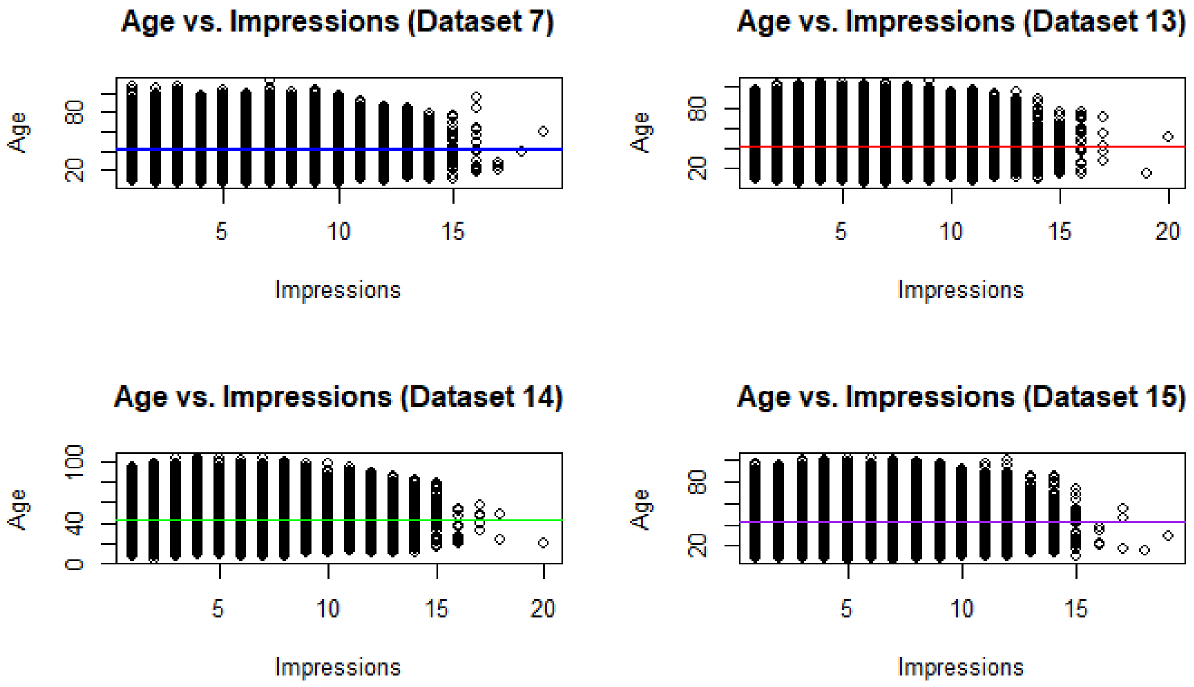


Figure 17: Linear model for Datasets

Creating a linear model of Age = m(impressions) + b we got different y-intercept (b) values and slope (m) values but they were quite similar. The resulting values for the linear regression equation was Age = 42 + 0.1 (Impressions). The y-intercept was quite consistent with being around 42 and the slopes were very low for all datasets. From the plots we see that the linear regression line shows no correlations between age and impressions since the slope is so small the line is basically a straight line from the y-intercept. From the values we got from the linear regression model and plots we have shown that Age and Impressions have no correlation with each other in these datasets.