Andrew Updegrove
661410127

Assignment 3

ITWS 6600

Question 1

a. The age distributions according to the boxplot were almost identical with a five number summary that looks like [0,0,31,48,108]. This is only different for NYT20 where the median age is 20 and the upper quartile is 43. NYT6 also had a maximum age of 106 as opposed to 108 of the others. These results are unsurprising considering how many datapoints are in each of these sets. The Impressions distributions only varied in the number and range of outliers. According to the 1.5 times inner-quartile range rule, there are quite a few outliers. This implies that lots of people have a few Impressions with those having more, going way beyond.

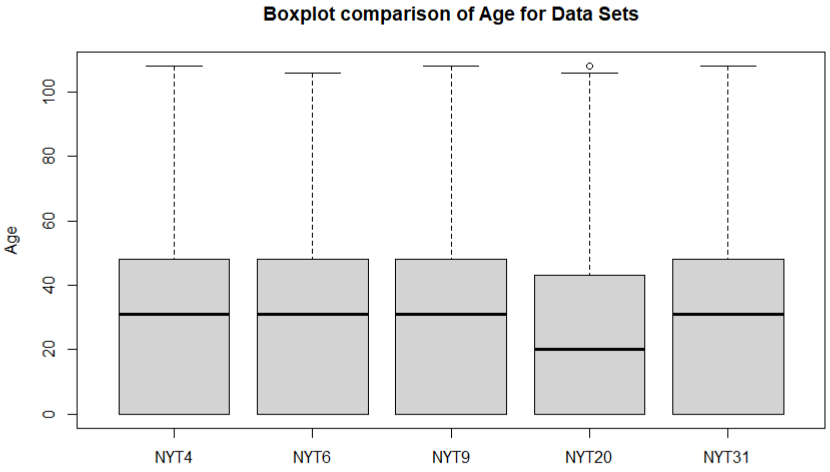

Figure 1: Box Plot for Age Variable

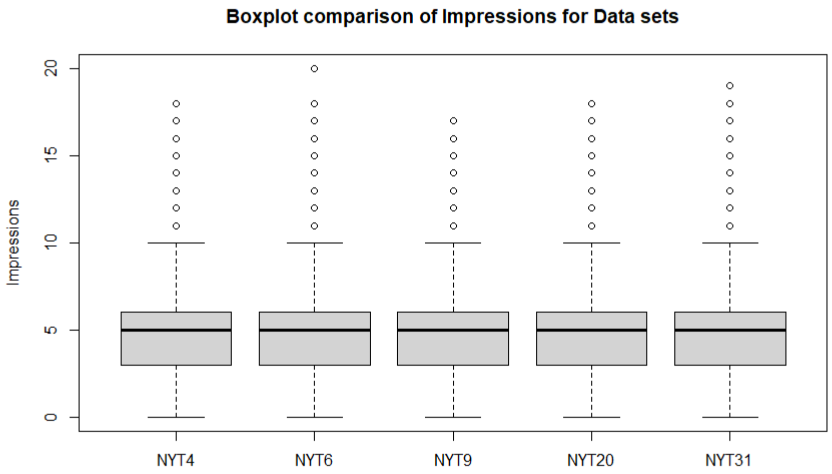

Figure 2: Box Plot for Impressions Variable

b. The histograms for both variables look like right-skewed normal distributions. The Age variable has a significant number of values that are in the 0 to 10 category. From the boxplots, it is implied that these are all values of zero. This makes sense as if an unregistered user visits the site, their age is not known. The Impressions variable looks fairly normal with some skew. This skew makes sense as Impressions is a a discrete value that is greater than or equal to zero and tends to be no greater than 10. This leaves little room for variation, so the data is right skewed.

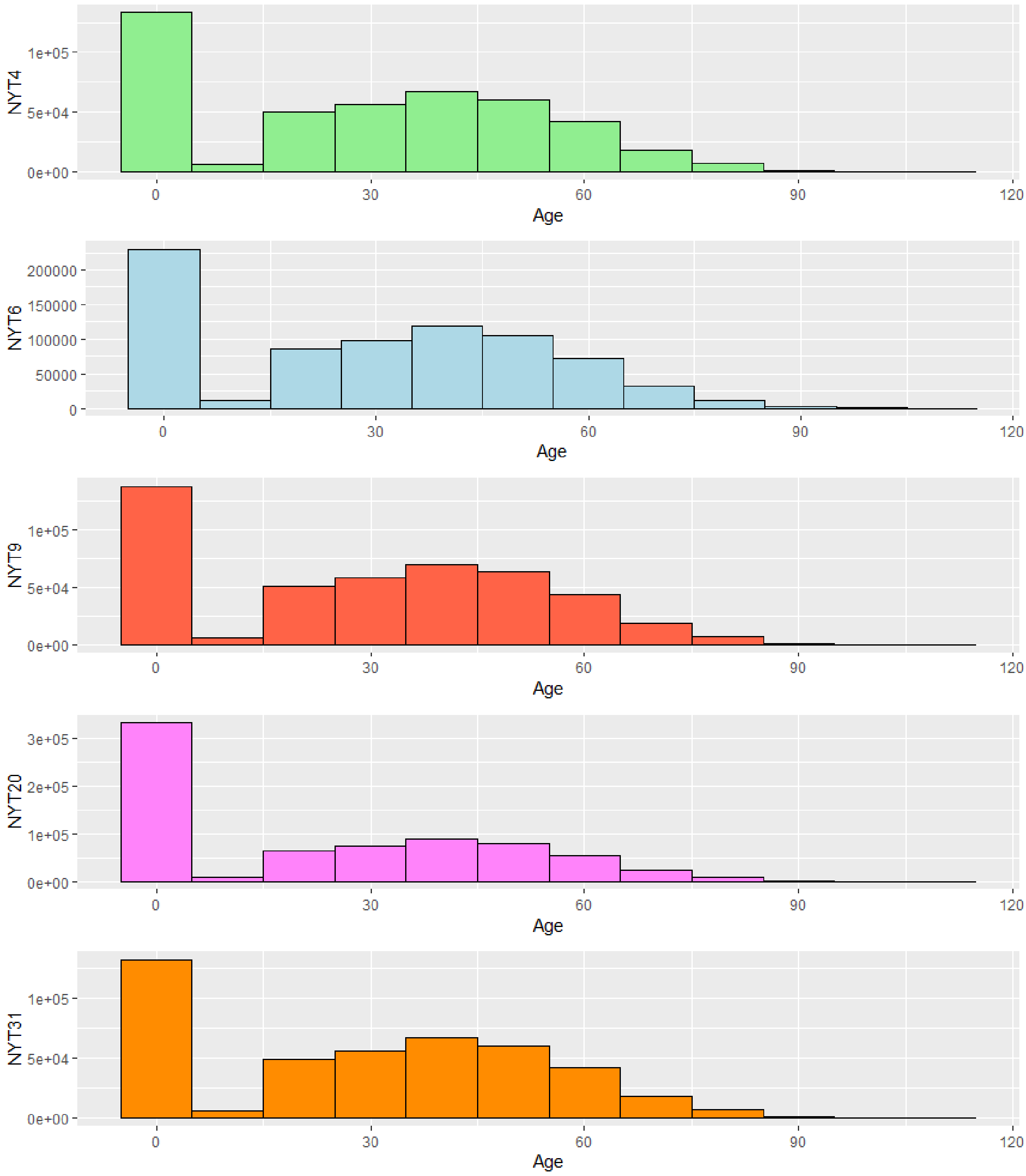

Figure 3: Histograms for Age Variable

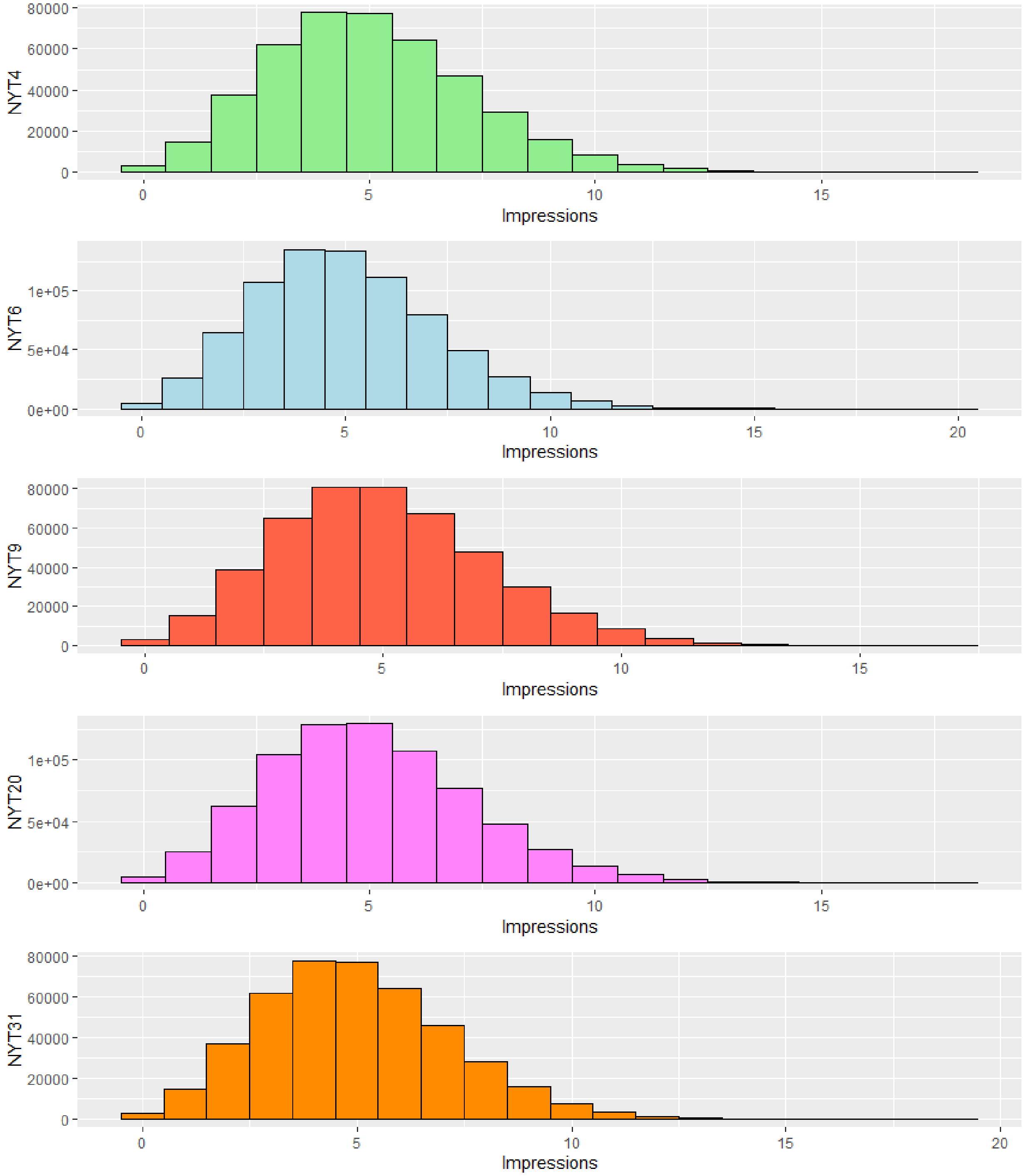

Figure 4: Histograms for Impressions Variable

c. The quantile-quantile plots were plotted against normal distributions. The quantile-quantile plot for Age are heavily biased towards the low end which was also agrees with what was seen in the previous plots. After the initial jump because of the large number of zeros, it is almost linear. The qq plot for Impressions is almost a straight line which would communicate a normal distribution, but it it is slightly concave up.

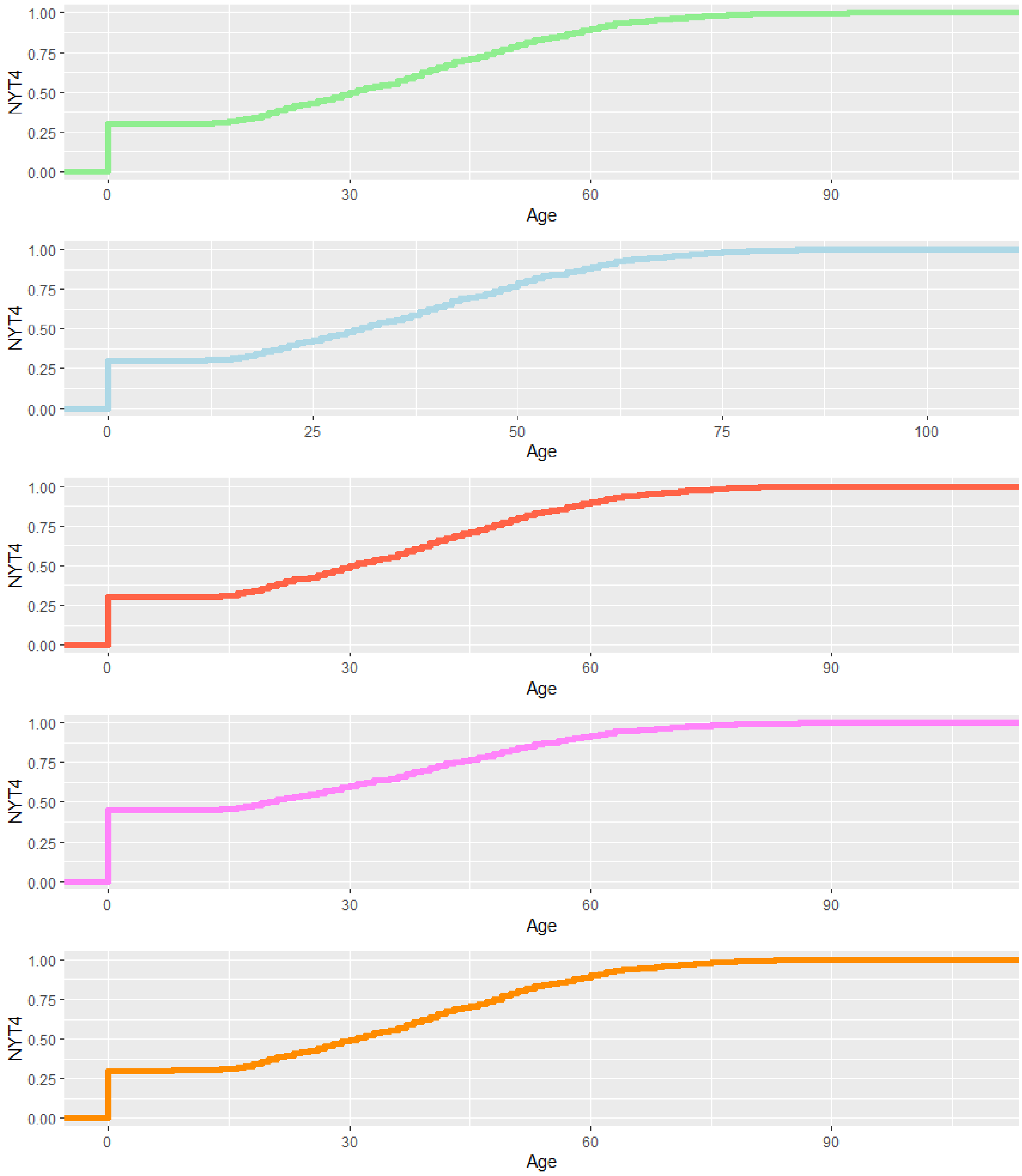

Figure 5: ECDFs for Age Variable

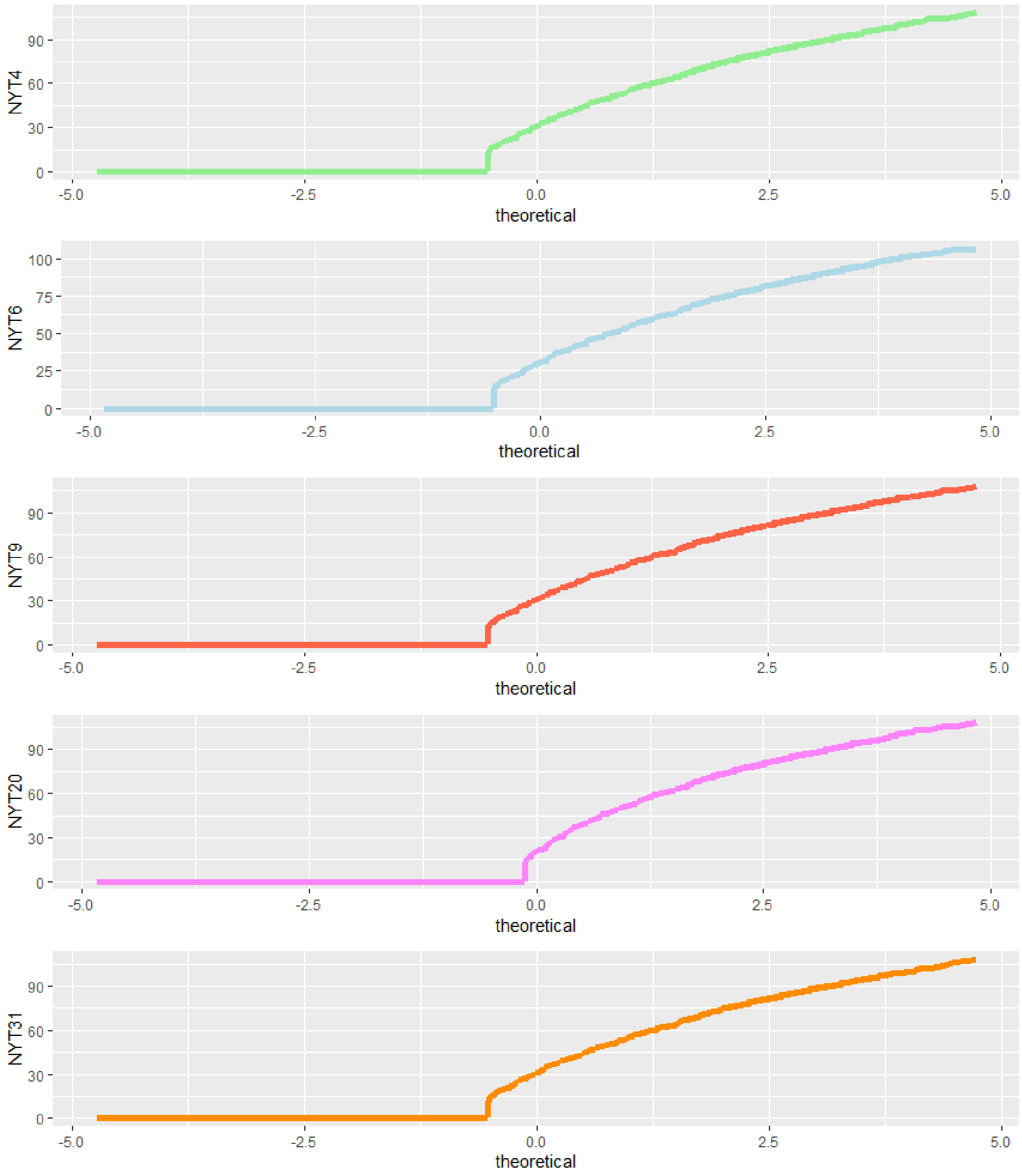

Figure 6: Quantile-Quantile Plots for Age Variable

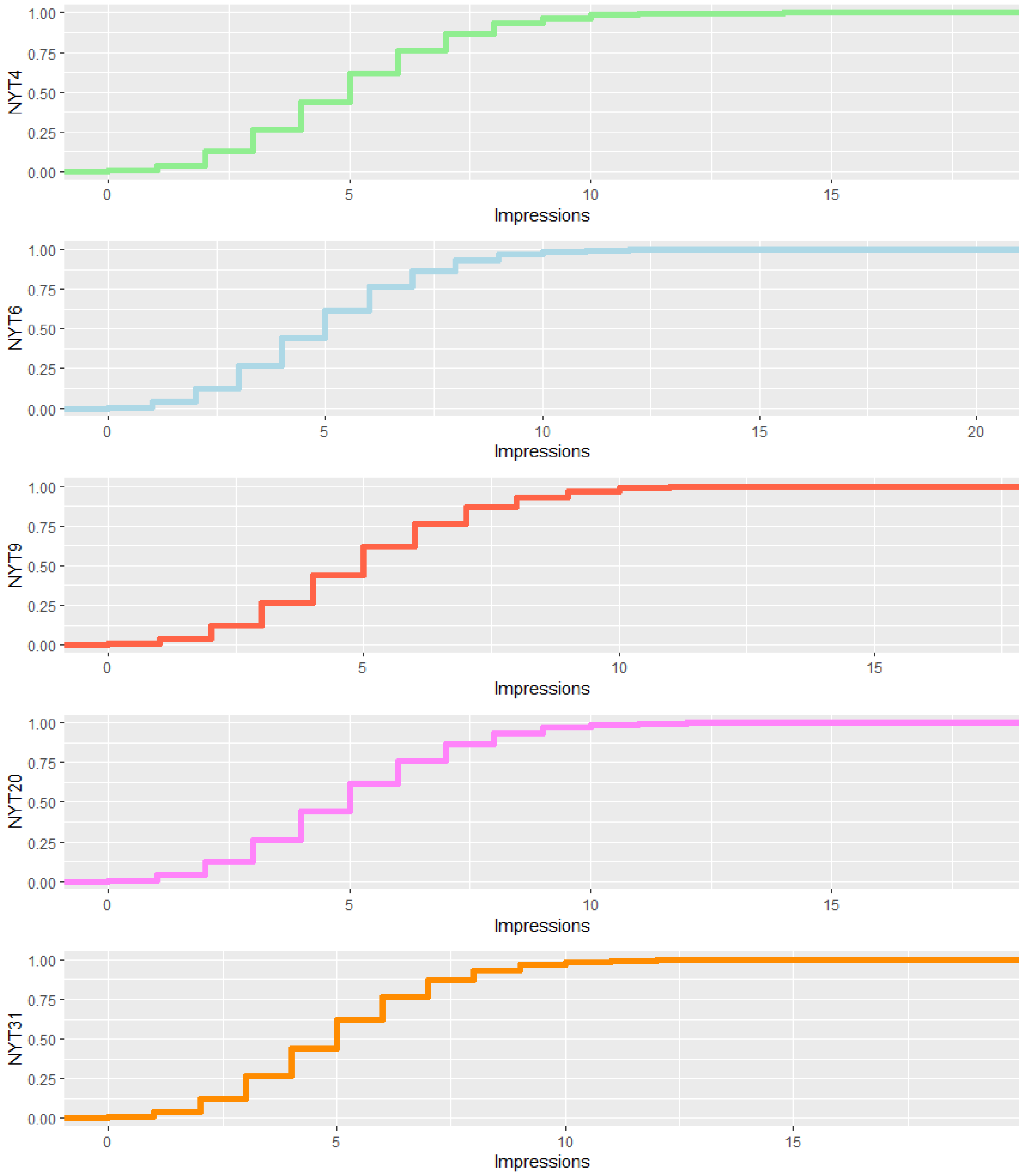

Figure 7: ECDFs for Impressions Variable

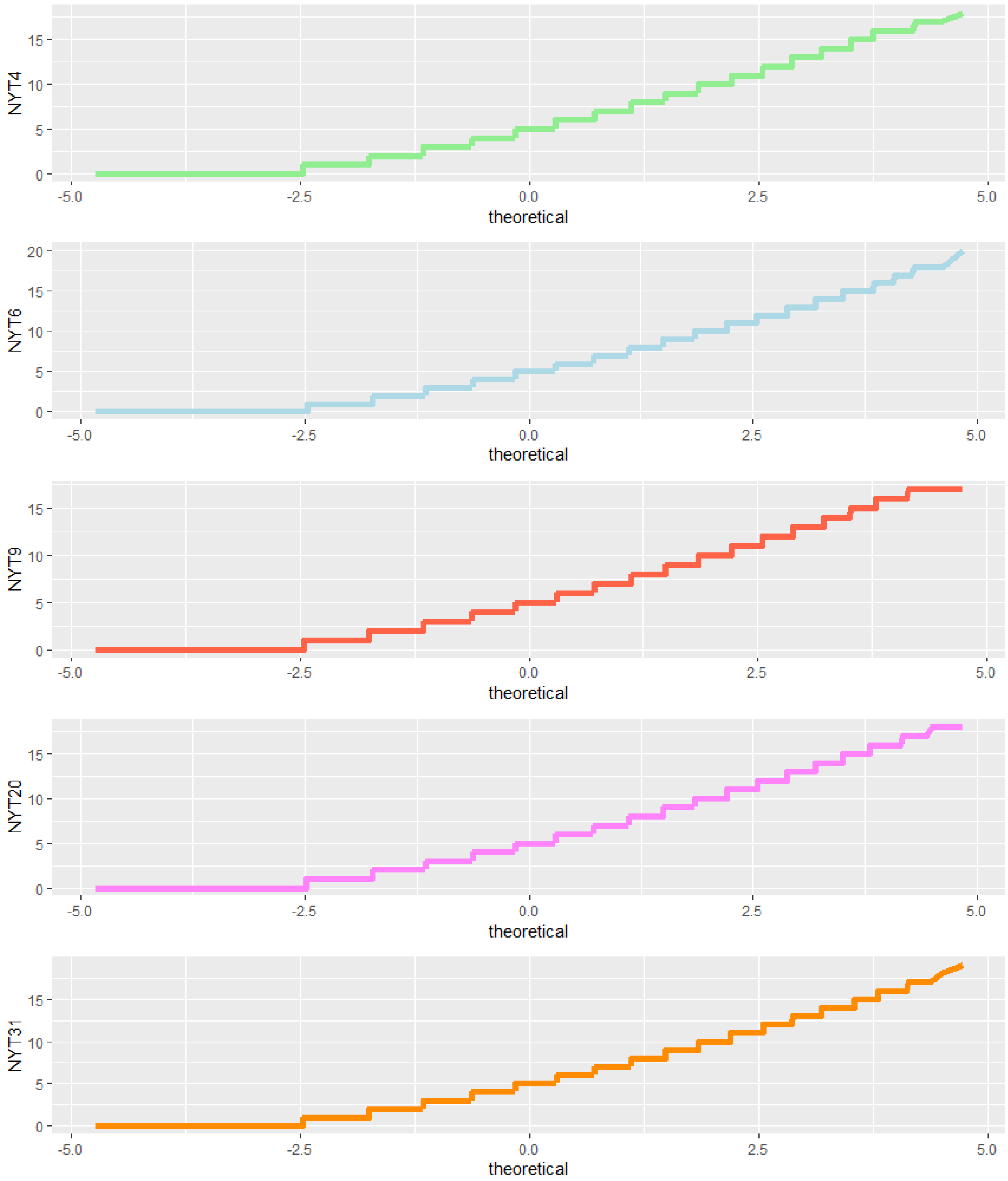

Figure 8: Quantile-Quantile Plots for Impressions Variable

d. The Shapiro-Wilks test for normality was performed on a random selection of 5000 data points for each variable from each data set. The p-value for all of them was less than $2.2e^{-16}$ meaning that the null hypothesis is rejected and the data was not from a normal distribution. It is surprising that the p-value is so low even for the Impressions variable even though it is relatively normal.

Dataset	W	p-value
NYT4	.913	2.2e-16
NYT6	.917	2.2e-16
NYT9	.908	2.2e-16
NYT20	.841	2.2e-16
NYT31	.915	2.2e-16

Table 1: Shapiro-Wilks Test for Age Variable

Dataset	W	p-value
NYT4	.970	2.2e-16
NYT6	.970	2.2e-16
NYT9	.971	2.2e-16
NYT20	.971	2.2e-16
NYT31	.972	2.2e-16

Table 2: Shapiro-Wilks Test for Impressions Variable

Question 2 (Required)

The data sets were then filtered by the Signed_In variable. The idea is that by removing the not signed in users it might fix the skew seen in the distributions while also getting rid of all the zero values for the Age variable. The assignement only asked for this to be done for two of the data sets but by the time I did two it was just as easy to do all five. So the filtering was done on all five data sets. Below is parts a through d with the data filtered to only use the data points where the user is signed in, indicated by the boolean value of 1 in the Signed_In column. The Age variable changed as expected however it still was found to be not normal by visual inspection and by the Shapiro-Wilks test. The Impression variable did not change significantly by only considering signed_in users. This implies that the Impressions of a User is independent of whether or not they are registered with the New York Time website. The p-value for every Shapiro-Wilks test was incredibly small which makes sense given the size of these data sets. At that number of data points, if the data was sample from a normally distributed population then it would be very evident.

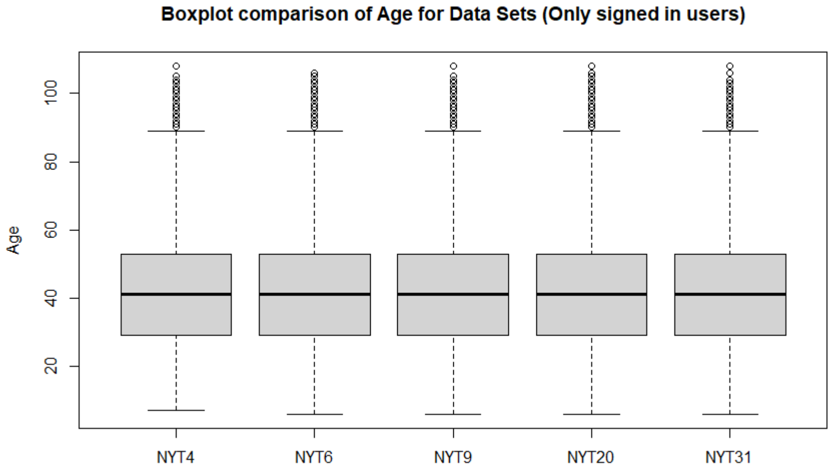

Figure 9: Box Plot for Age Variable with only Signed In Users

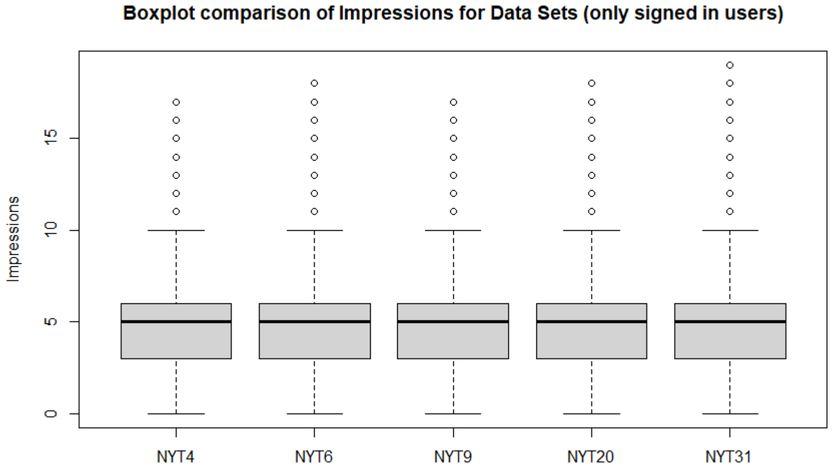

Figure 10: Box Plot for Impressions Variable with only Signed In Users

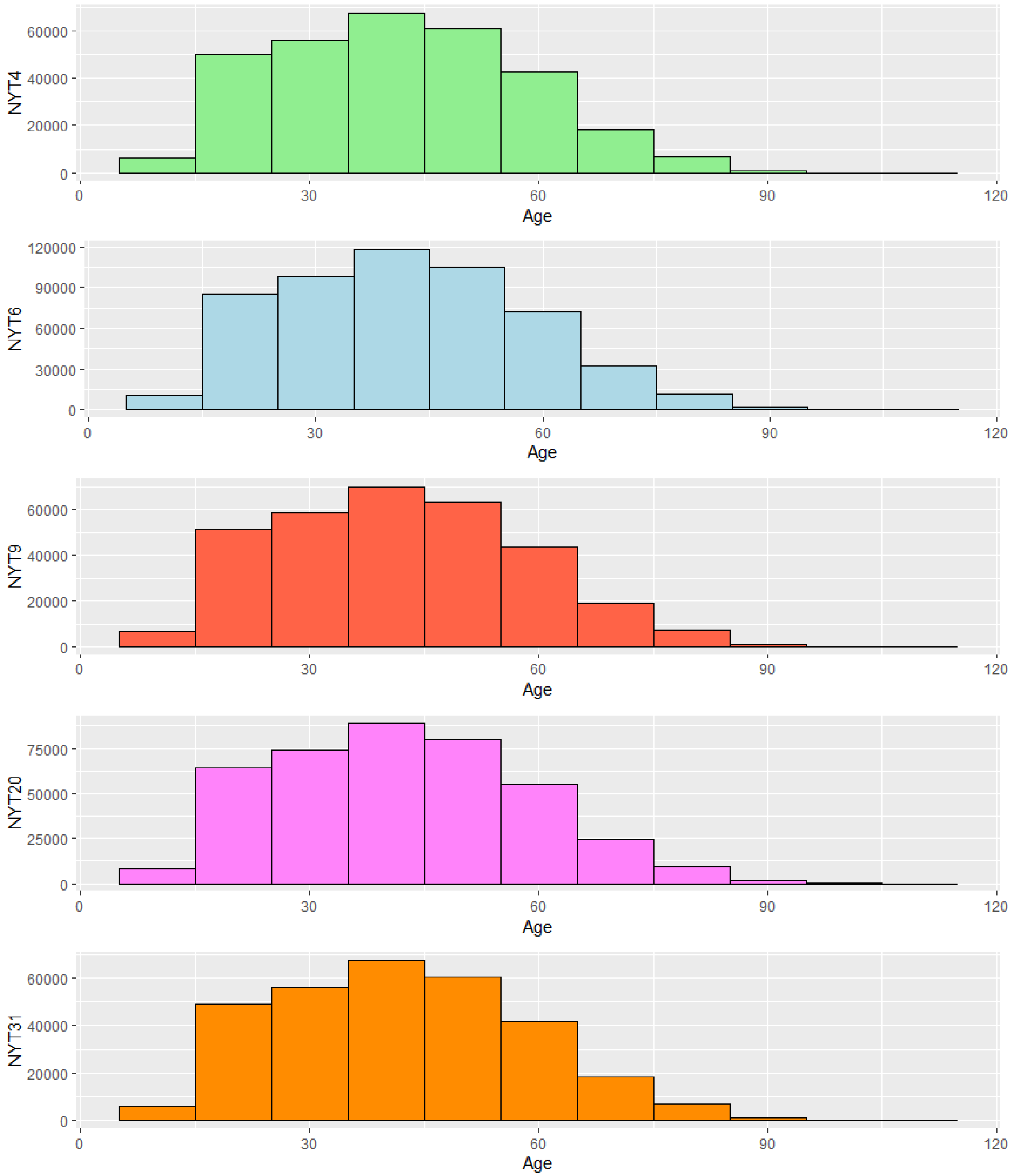

Figure 11: Histograms for Age Variable with only Signed in Users

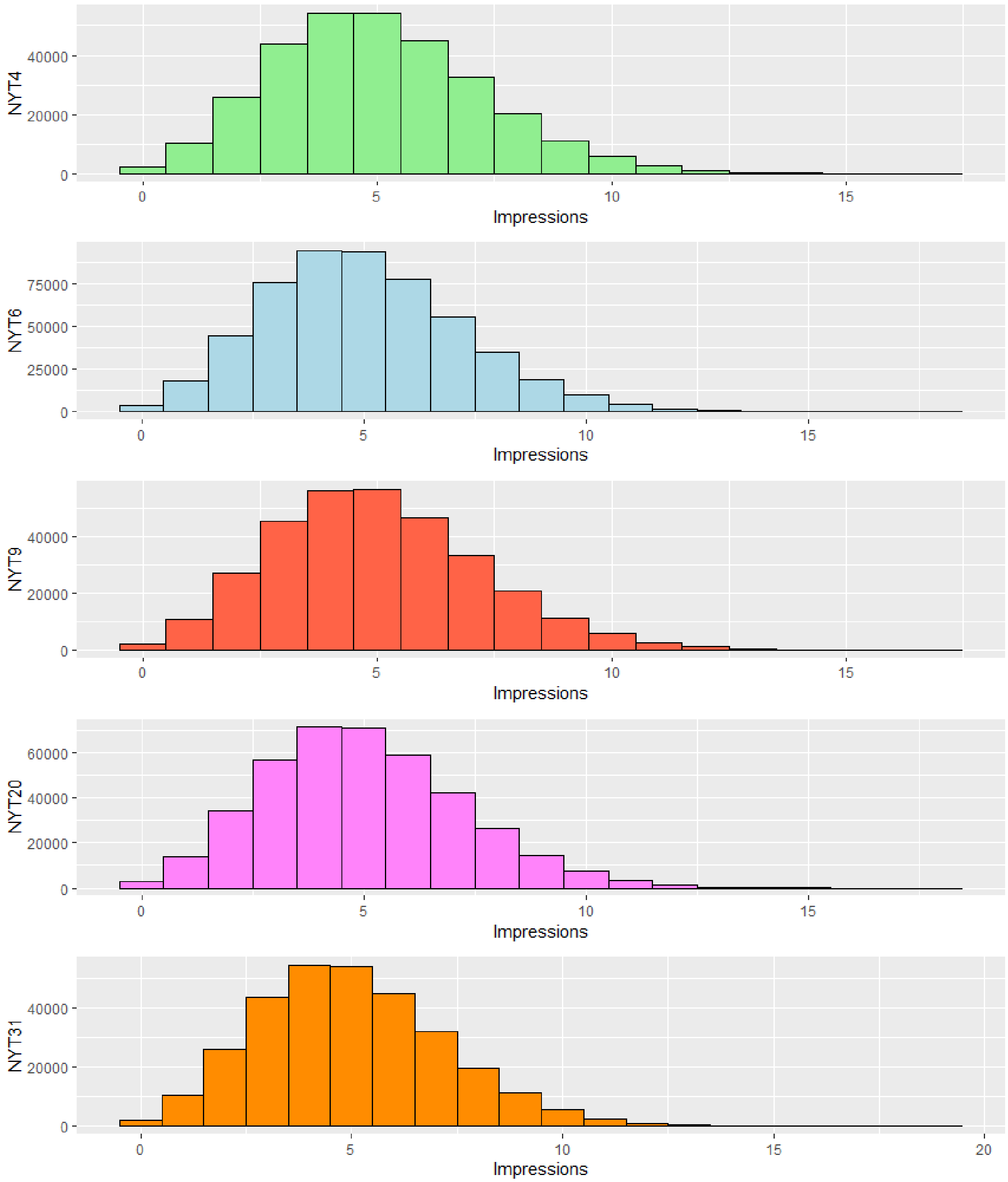

Figure 12: Histograms for Impressions Variable with only Signed in Users

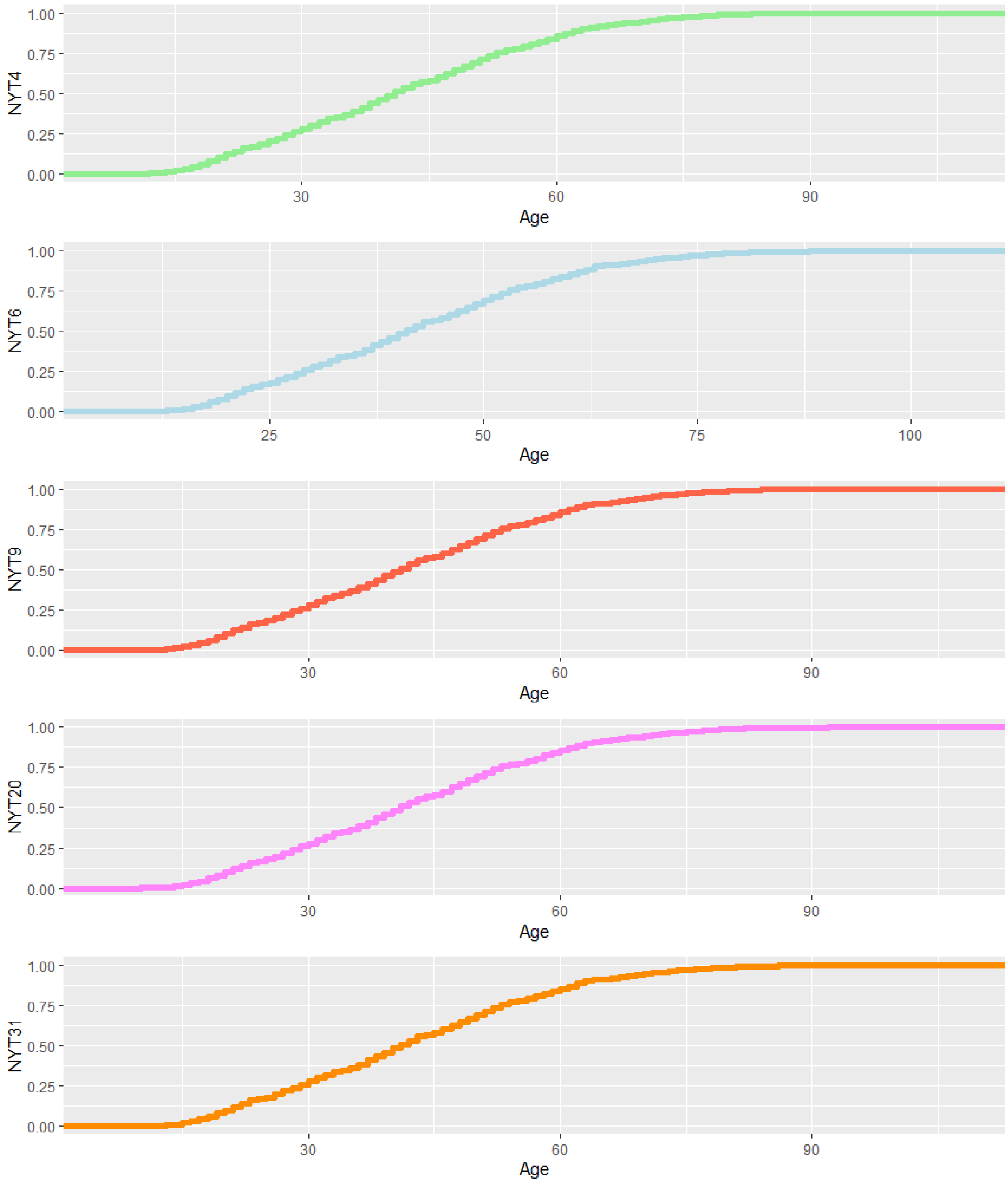

Figure 13: ECDFs for Age Variable with only Signed in Users

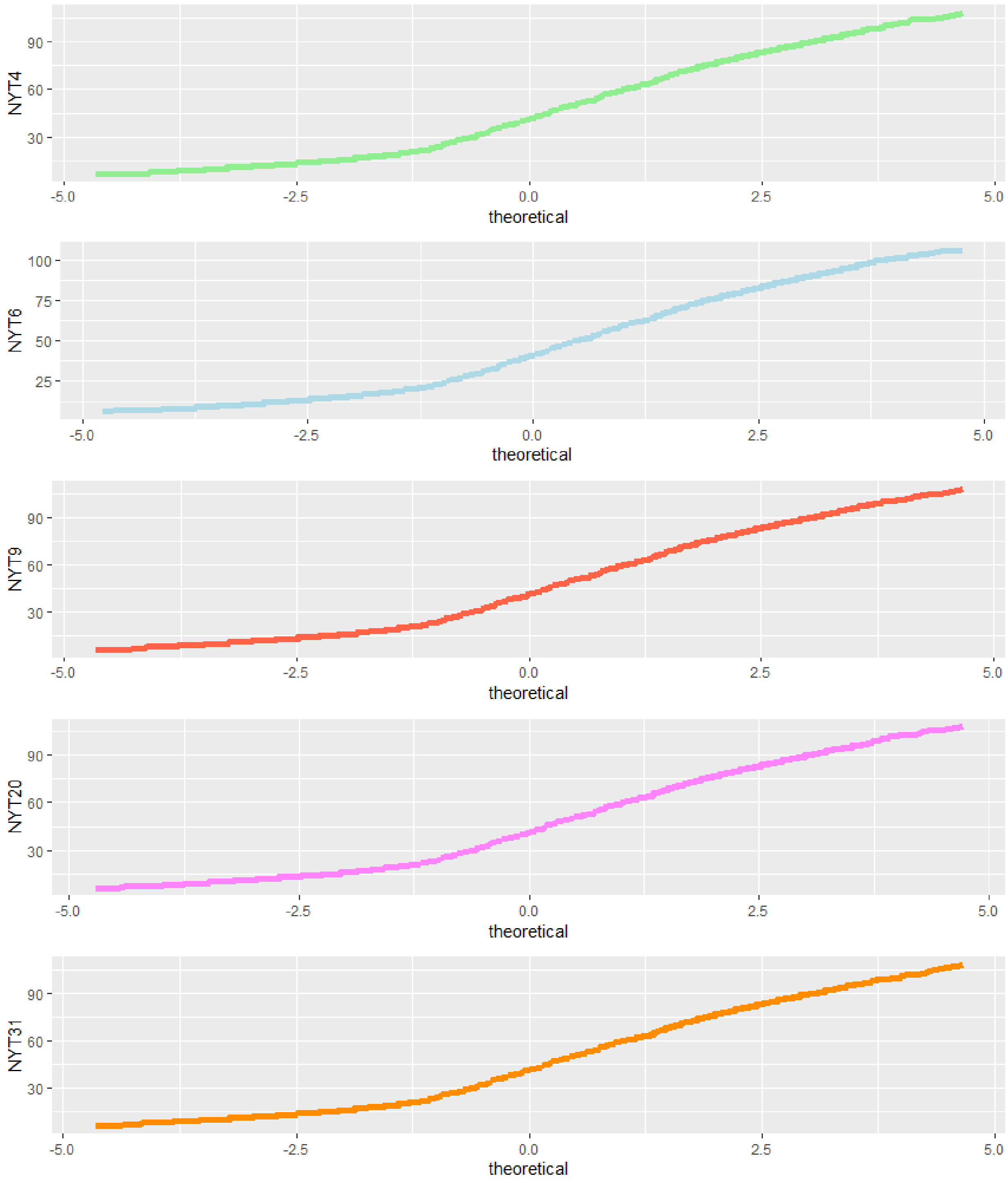

Figure 14: Quantile-Quantile Plots for Age Variable with only Signed in Users

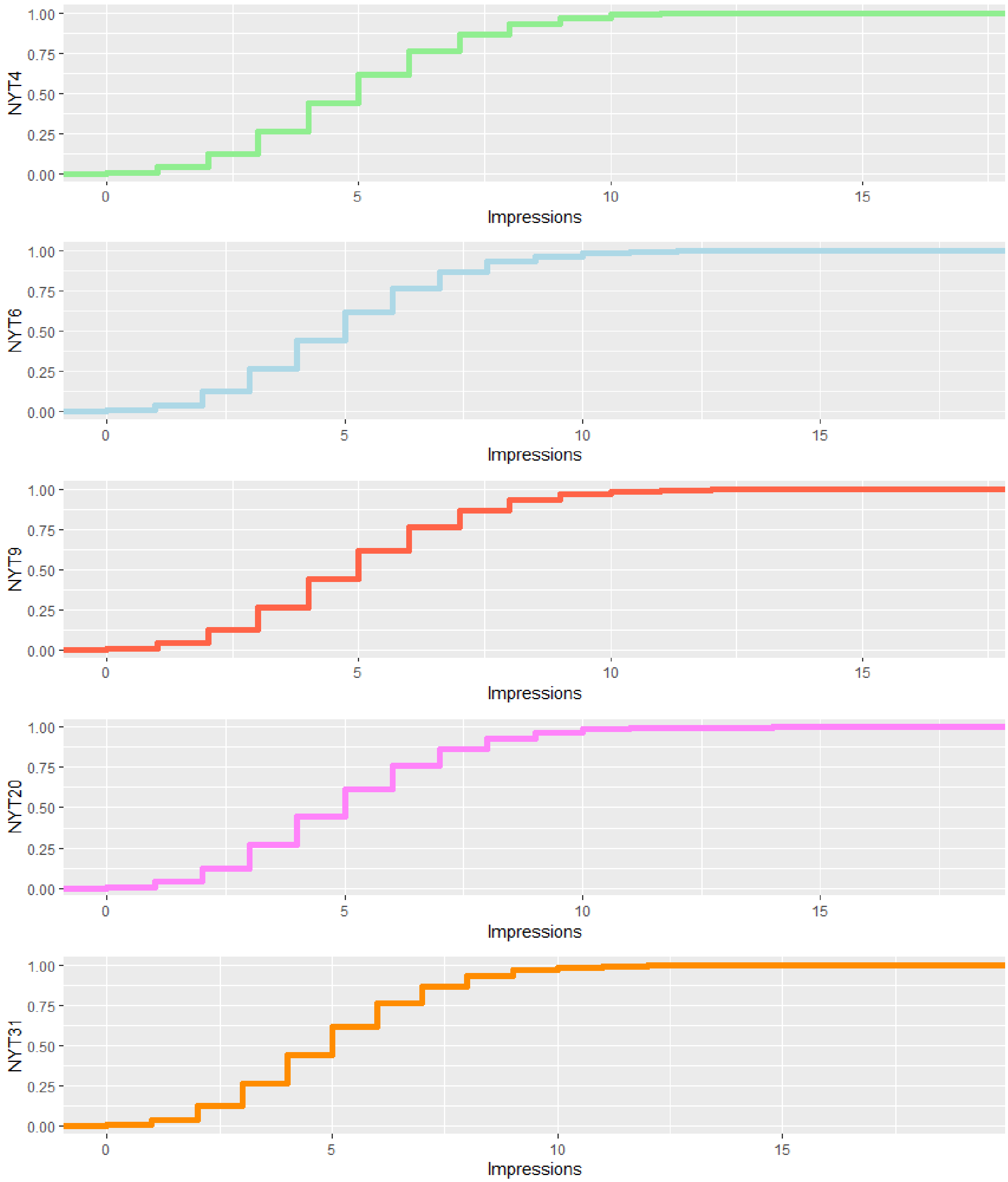

Figure 15: ECDFs for Impressions Variable with only Signed in Users

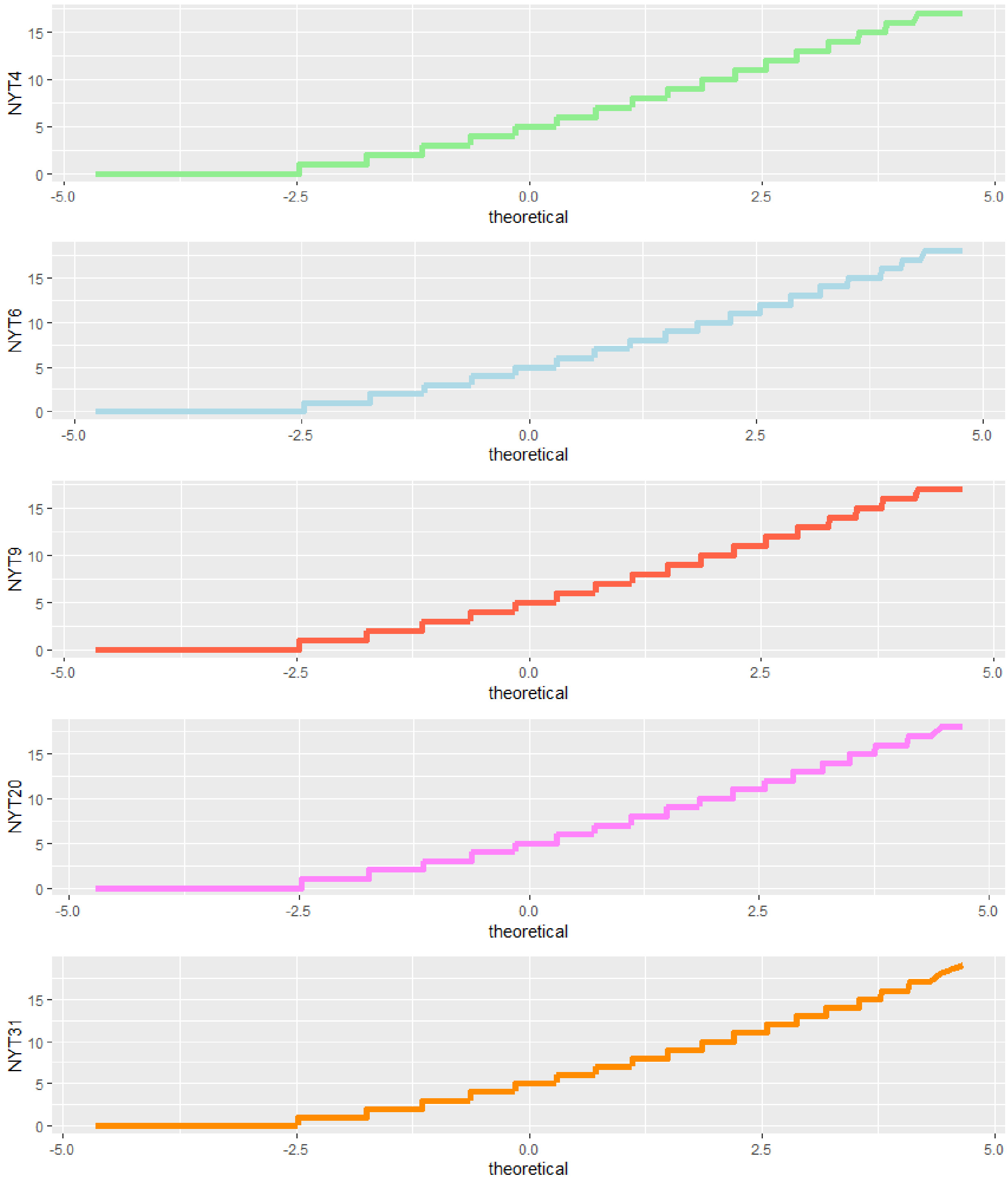

Figure 16: Quantile-Quantile Plots for Impressions Variable with only Signed in Users

Dataset	W	p-value
NYT4	.978	2.2e-16
NYT6	.981	2.2e-16
NYT9	.981	2.2e-16
NYT20	.980	2.2e-16
NYT31	.978	2.2e-16

Table 3: Shapiro-Wilks Test for Age Variable

Dataset	W	p-value
NYT4	.971	2.2e-16
NYT6	.968	2.2e-16
NYT9	.971	2.2e-16
NYT20	.967	2.2e-16
NYT31	.971	2.2e-16

Table 4: Shapiro-Wilks Test for Impressions Variable