6000 level

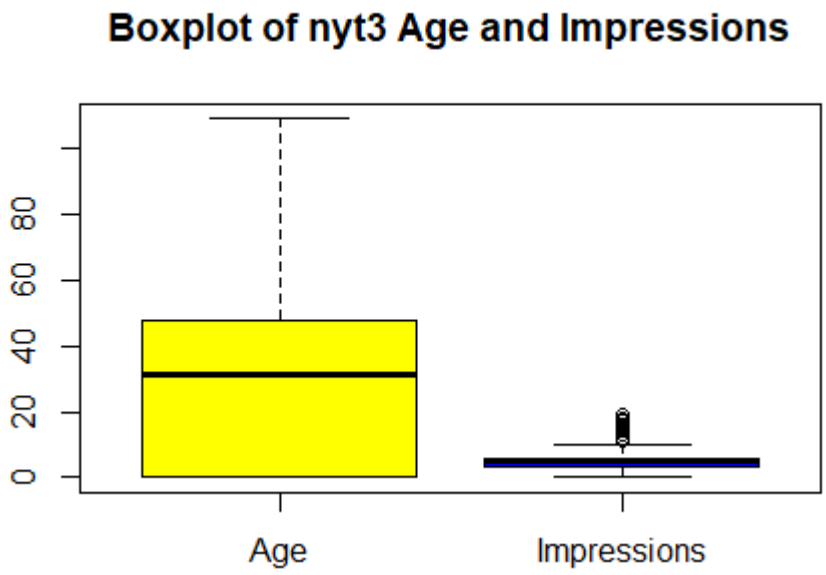
NAME: OGOCHUKWU NWACHUKWU GLORIA

RIN: 662058987

ASSIGNMENT 3

Question 1 a

Boxplot for nyt3, nyt4, nyt5, nyt6, nyt7, nyt8 and nyt9 Age and Impressions



Age:

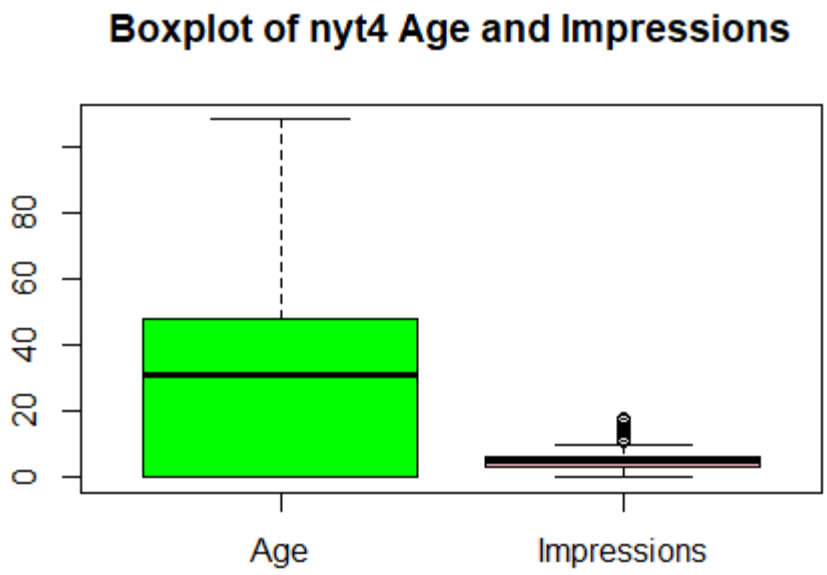
- Minimum Age: 0.00
- 1st Quartile (25th percentile): 0.00
- Median (50th percentile): 31.00
- Mean: 29.47
- 3rd Quartile (75th percentile): 48.00
- Maximum Age: 109.00

Impressions:

- Minimum Impressions: 0.000
- 1st Quartile (25th percentile): 3.000
- Median (50th percentile): 5.000
- Mean: 4.996
- 3rd Quartile (75th percentile): 6.000

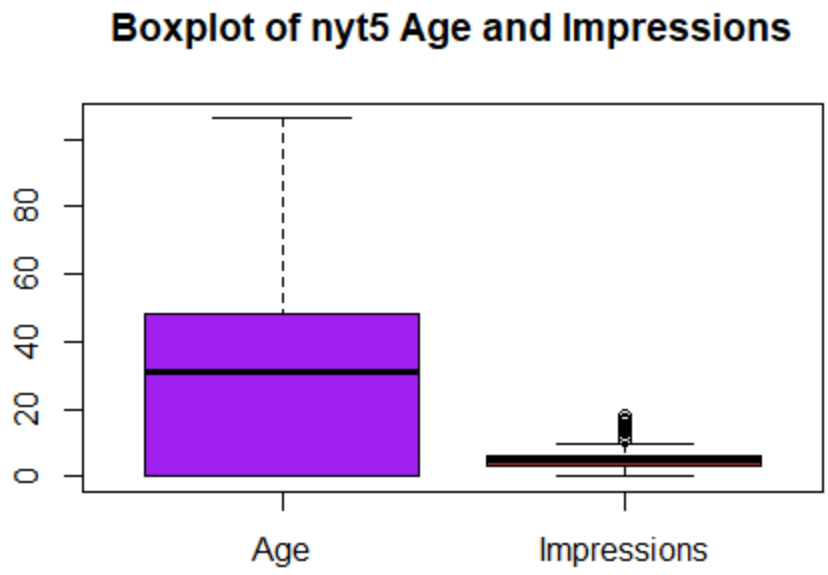
- Maximum Impressions: 19.000

In Age, majority of the data falls within the range of 0 to 48, as indicated by the first and third quartiles. The median age is 31, and the mean age is approximately 29.47. In Impressions, most of the data is clustered between 3 and 6 impressions.



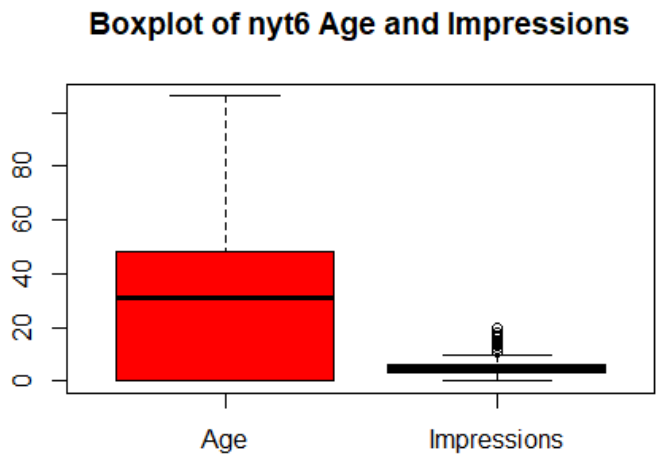
- **Age:**
  - Min: 0.00
  - 1st Qu.: 0.00
  - Median: 31.00
  - Mean: 29.43
  - 3rd Qu.: 48.00
  - Max: 108.00
- **Impressions:**
  - Min: 0.000
  - 1st Qu.: 3.000
  - Median: 5.000
  - Mean: 5.004
  - 3rd Qu.: 6.000
  - Max: 18.000

The age distribution has a minimum of 0, a maximum of 108, and is centered around the median and mean of 31 and 29.43, respectively. Impressions range from 0 to 18, with a median of 5 and a slightly higher mean of 5.004 compared to nyt3.



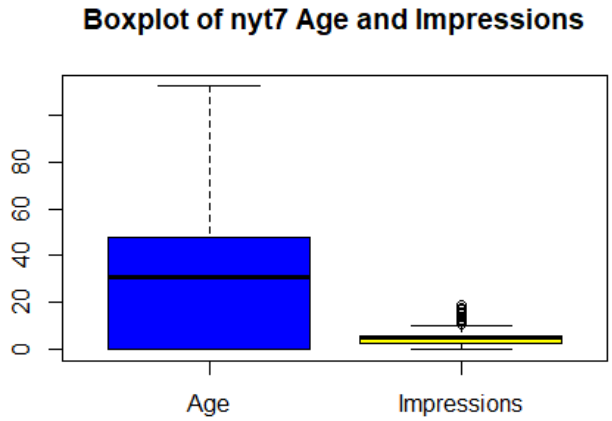
- **Age:**
  - Min: 0.00
  - 1st Qu.: 0.00
  - Median: 31.00
  - Mean: 29.43
  - 3rd Qu.: 48.00
  - Max: 106.00
- **Impressions:**
  - Min: 0.000
  - 1st Qu.: 3.000
  - Median: 5.000
  - Mean: 4.999
  - 3rd Qu.: 6.000
  - Max: 18.000

Age has a median of 31, mean of 29.43, and a range from 0 to 106. Impressions also follow a similar pattern, with a median of 5.



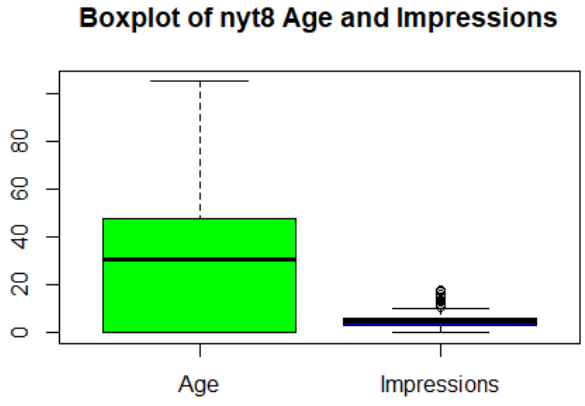
- **Age:**
  - Min: 0.00
  - 1st Qu.: 0.00
  - Median: 31.00
  - Mean: 29.46
  - 3rd Qu.: 48.00
  - Max: 106.00
- **Impressions:**
  - Min: 0.000
  - 1st Qu.: 3.000
  - Median: 5.000
  - Mean: 4.995
  - 3rd Qu.: 6.000
  - Max: 20.000

Age has a median of 31, a mean of 29.46, and a range from 0 to 106. Impressions have a median of 5 and a mean of 4.995, similar to previous datasets.



- **Age:**
  - Min: 0.00
  - 1st Qu.: 0.00
  - Median: 31.00
  - Mean: 29.52
  - 3rd Qu.: 48.00
  - Max: 112.00
- **Impressions:**
  - Min: 0
  - 1st Qu.: 3
  - Median: 5
  - Mean: 5
  - 3rd Qu.: 6
  - Max: 19

Age has a median of 31, a mean of 29.52, and a range from 0 to 112. Impressions has a median of 5 and a mean of 5.



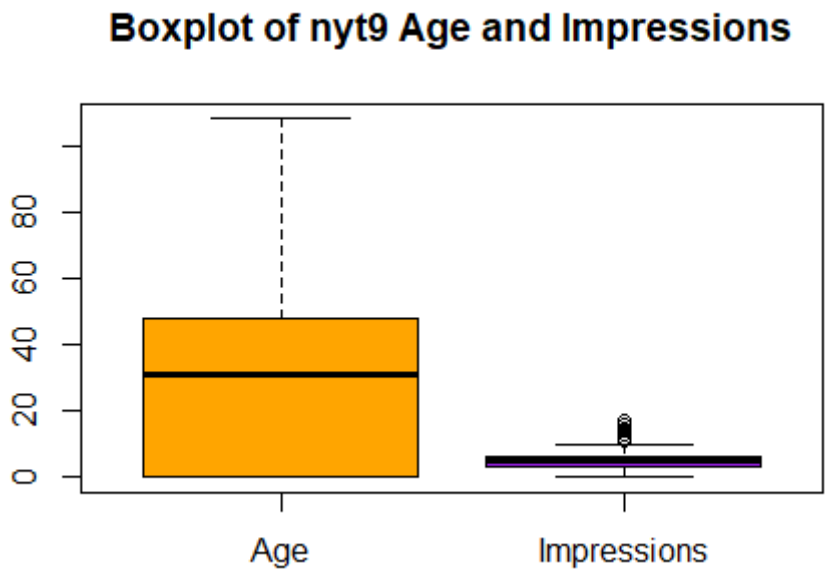
Age:

- Minimum (Min): 0.00
- 1st Quartile (1st Qu.): 0.00
- Median: 31.00
- Mean: 29.41
- 3rd Quartile (3rd Qu.): 48.00
- Maximum (Max): 105.00

Impressions:

- Minimum (Min): 0
- 1st Quartile (1st Qu.): 3
- Median: 5
- Mean: 5.001
- 3rd Quartile (3rd Qu.): 6
- Maximum (Max): 18

The Age variable the median (31.00) being slightly less than the mean (29.41). The 3rd quartile (48.00) is greater than the median, indicating the presence of potential outliers on the higher end of the data. In Impressions variable the median (5) is slightly less than the mean (5.001). The 3rd quartile (6) is greater than the median, suggesting possible outliers on the higher side of the data.



- **Age:**
  - Min: 0.00
  - 1st Qu.: 0.00

    - Median: 31.00
    - Mean: 29.45
    - 3rd Qu.: 48.00
    - Max: 108.00
- **Impressions:**
    - Min: 0.000
    - 1st Qu.: 3.000
    - Median: 5.000
    - Mean: 4.998
    - 3rd Qu.: 6.000
    - Max: 17.000

Age statistics are in line with the previous datasets, with a median of 31, a mean of 29.45, and a range from 0 to 108. Impressions has a median of 5 and a mean just below 5. In the boxplot for nyt9, the distribution of age and impressions aligns with the previous datasets, with most data clustered at the lower end of the scale.

Conclusively, the datasets nyt3, nyt4, nyt5, nyt6, nyt7, nyt8 and nyt9 share similar statistical characteristics for age and impressions. They all have a concentration of data points with low age values and low impressions, while having some variations in means and max values. Most of the data is centered around the lower values. The boxplots visually represent these distributions, showing where the majority of the data falls and the presence of outliers. These visualizations help identify data patterns and potential outliers.

Question b

normality test using Anderson Darling test

The Anderson-Darling normality test assesses whether a given sample comes from a normally distributed population. In your analysis, you performed this test on both Age and Impressions for multiple datasets (nyt3, nyt4, nyt5, nyt6, nyt7, nyt8 and nyt9) and found that the p-values for all tests were less than 2.2e-16. This extremely low p-value suggests that the data from all these datasets do not follow a normal distribution.

The results for nyt3, nyt4, nyt5, nyt6, nyt7, nyt8 and nyt9 Age and Impressions

For nyt3

data: nyt3$Age
A = 12507, p-value < 2.2e-16

data: nyt3$Impressions
A = 4652.5, p-value < 2.2e-16

For nyt4

data: nyt4$Age
A = 12645, p-value < 2.2e-16

data: nyt4$Impressions
A = 4614.8, p-value < 2.2e-16

For nyt5

data: nyt5$Age
A = 10541, p-value < 2.2e-16

data: nyt5$Impressions
A = 3843.1, p-value < 2.2e-16

For nyt6

data: nyt6$Age
A = 21752, p-value < 2.2e-16

data: nyt6$Impressions
A = 8007.3, p-value < 2.2e-16

For nyt7

data: nyt7$Age
A = 12842, p-value < 2.2e-16

data: nyt7$Impressions
A = 4705, p-value < 2.2e-16

For nyt8

data: nyt8$Age
A = 13297, p-value < 2.2e-16

data: nyt8$Impressions
A = 4832.9, p-value < 2.2e-16

For nyt9

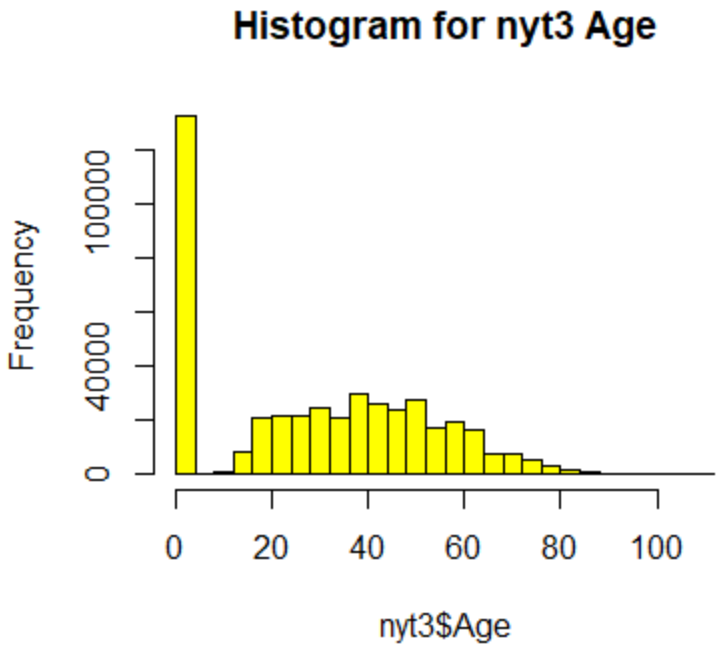
data: nyt9$Age
A = 13033, p-value < 2.2e-16

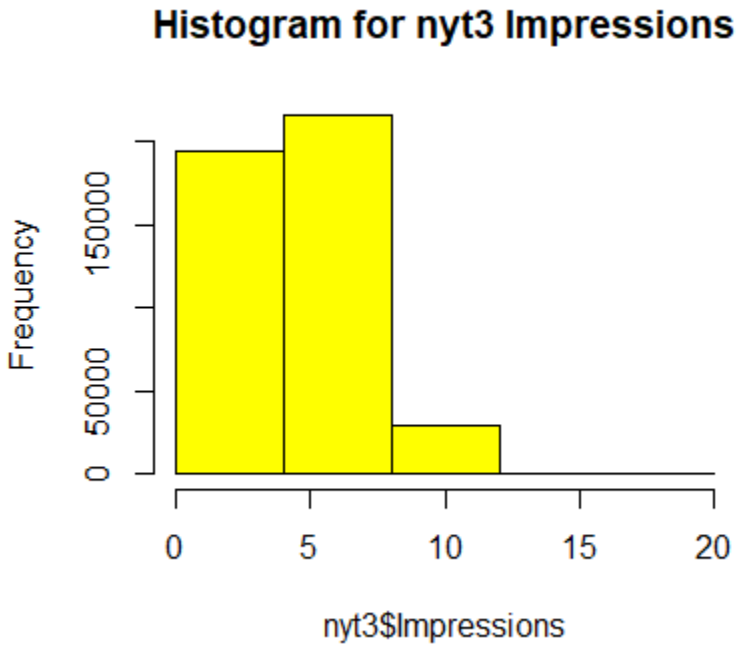
data: nyt9$Impressions
A = 4805.4, p-value < 2.2e-16

For both Age and Impressions in nyt3, nyt4, nyt5, nyt6, nyt7, nyt8, and nyt9, the data deviates significantly from a normal distribution as a result the null hypothesis (i.e., p-value > 0.05) was rejected as the data did not follow a normal distribution

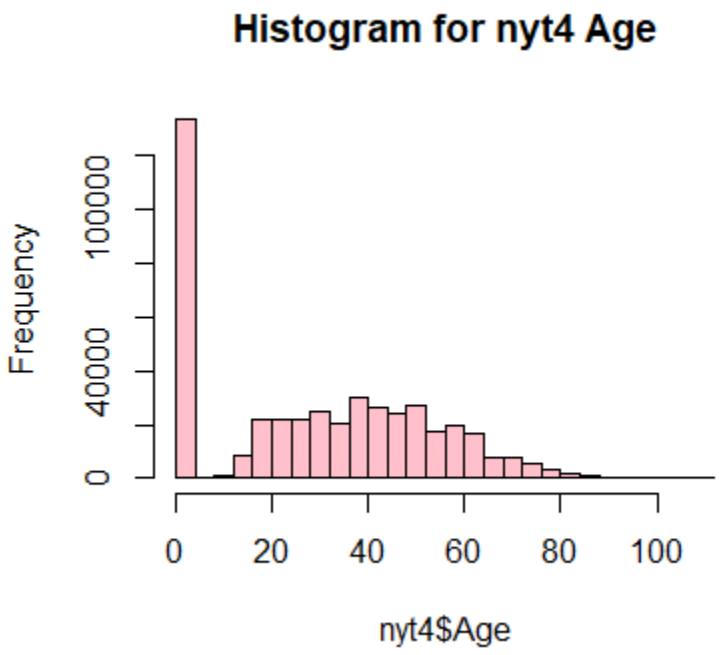
Histogram for nyt3, nyt4, nyt5, nyt6, nyt7, nyt8 and nyt9 "Age and Impressions"

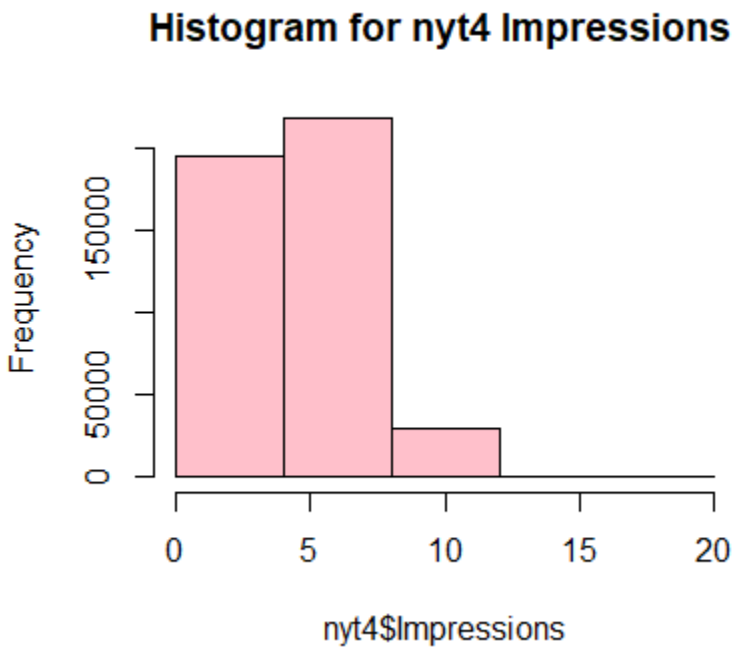
for nyt3,



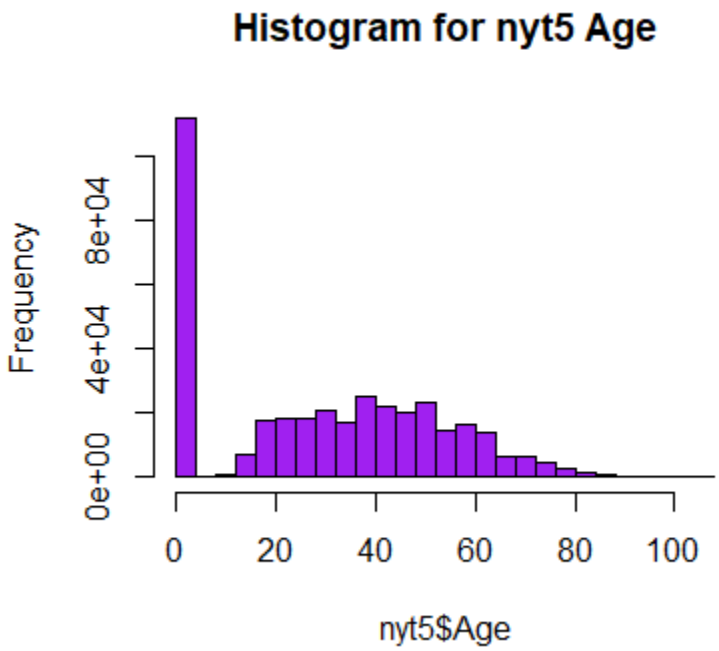


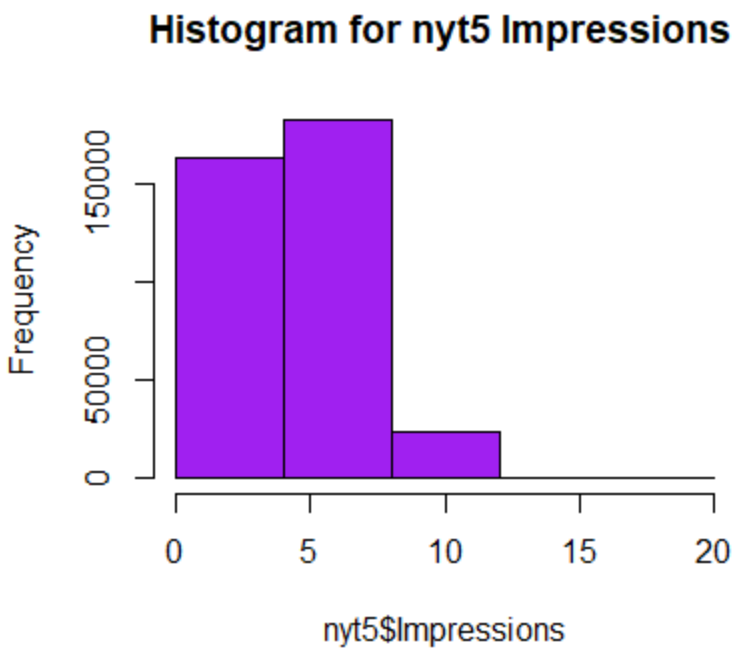
For nyt4,


Histogram for nyt4 Age
Frequency
0
40000
100000
0
20
40
60
80
100
nyt4$Age

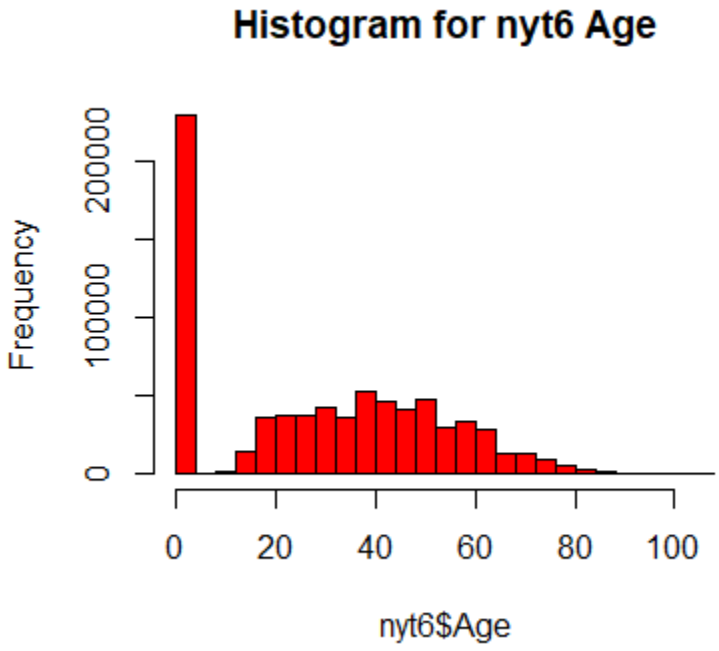

Histogram for nyt4 Impressions
Frequency
0
50000
150000
0
5
10
15
20
nyt4$Impressions

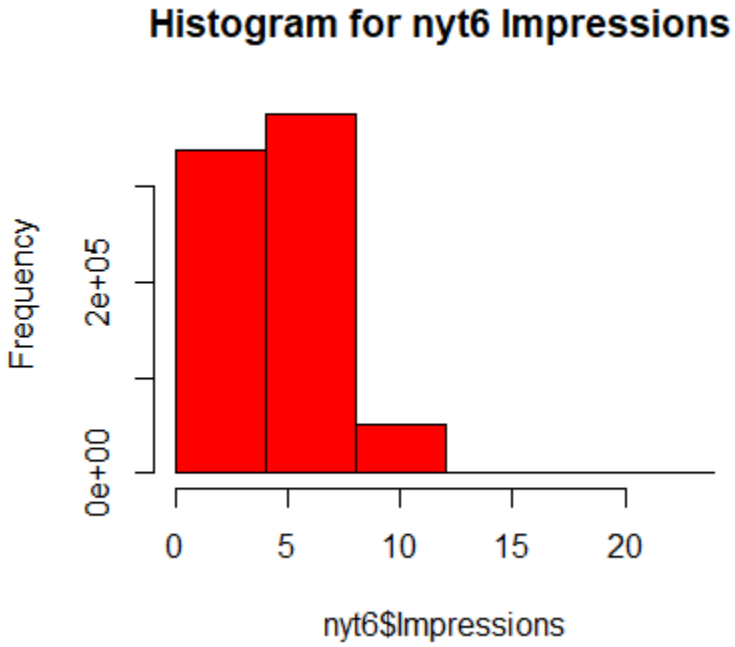
For nyt5,


Histogram for nyt5 Age
Frequency
0e+00
4e+04
8e+04
0
20
40
60
80
100
nyt5$Age

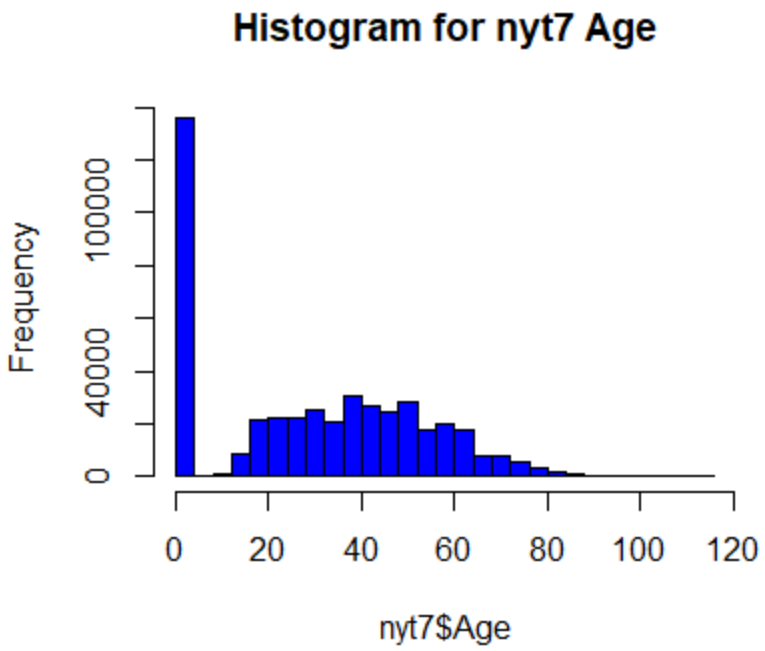

Histogram for nyt5 Impressions
Frequency
0
50000
150000
0
5
10
15
20
nyt5$Impressions

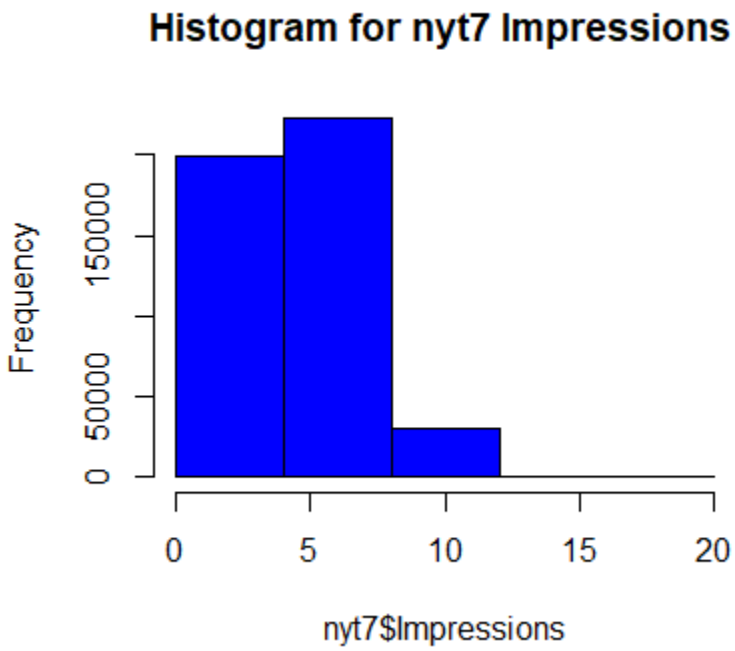
For nyt6,


Histogram for nyt6 Age
Frequency
0
100000
200000
0
20
40
60
80
100
nyt6$Age

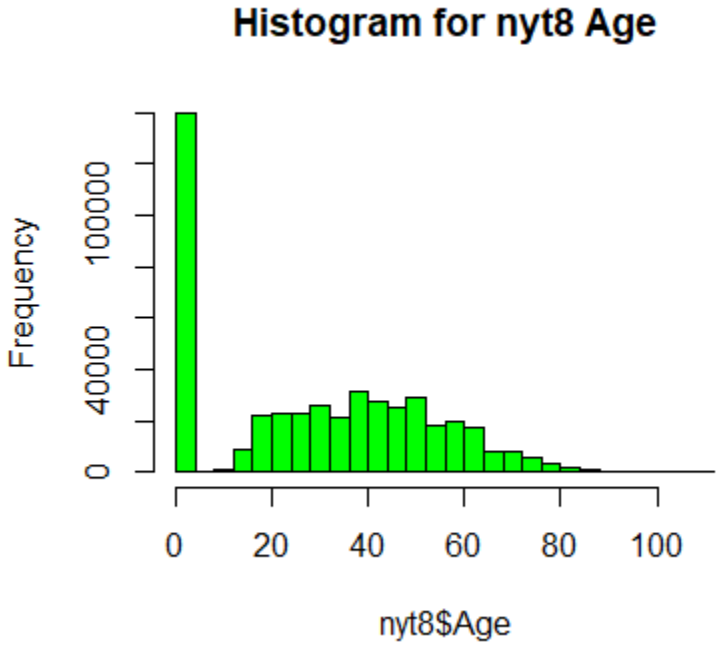

Histogram for nyt6 Impressions
Frequency
0e+00
2e+05
0
5
10
15
20
nyt6$Impressions

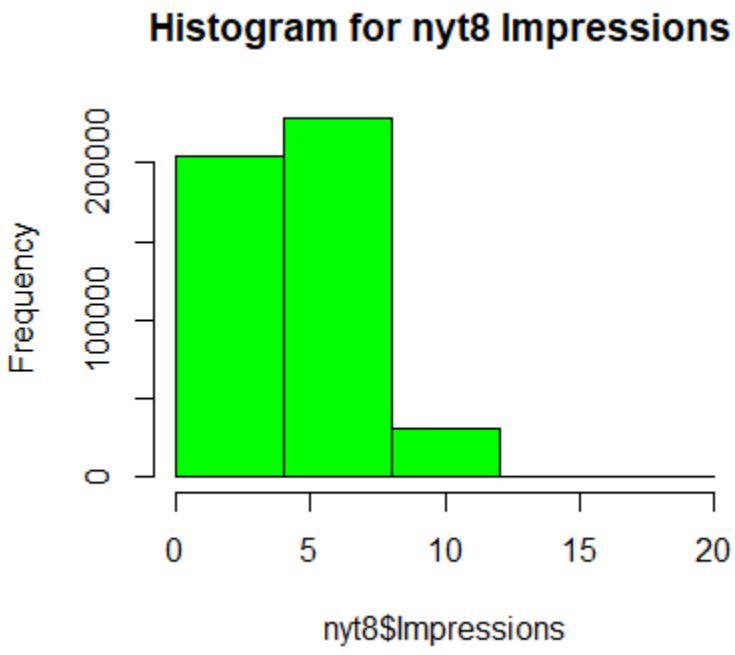
For nyt7,


Histogram for nyt7 Age
Frequency
0
40000
100000
0
20
40
60
80
100
120
nyt7$Age

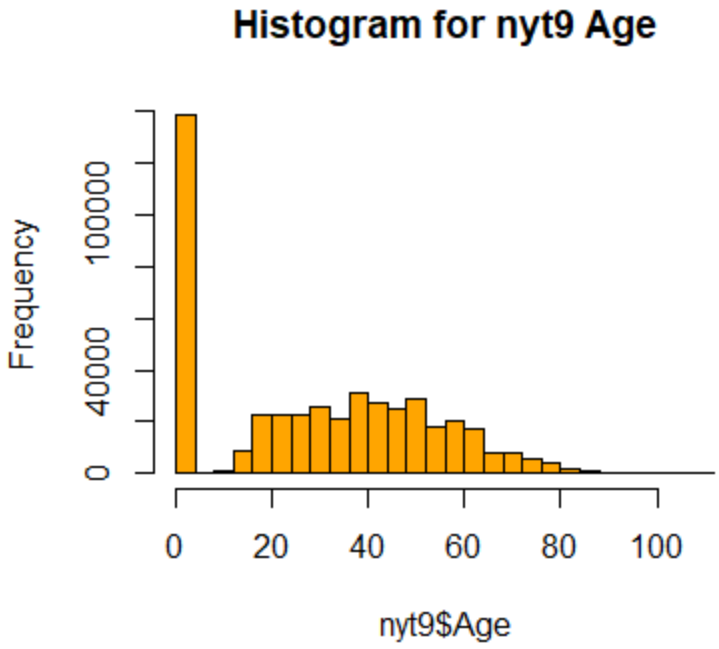

Histogram for nyt7 Impressions
Frequency
0
50000
150000
0
5
10
15
20
nyt7$Impressions

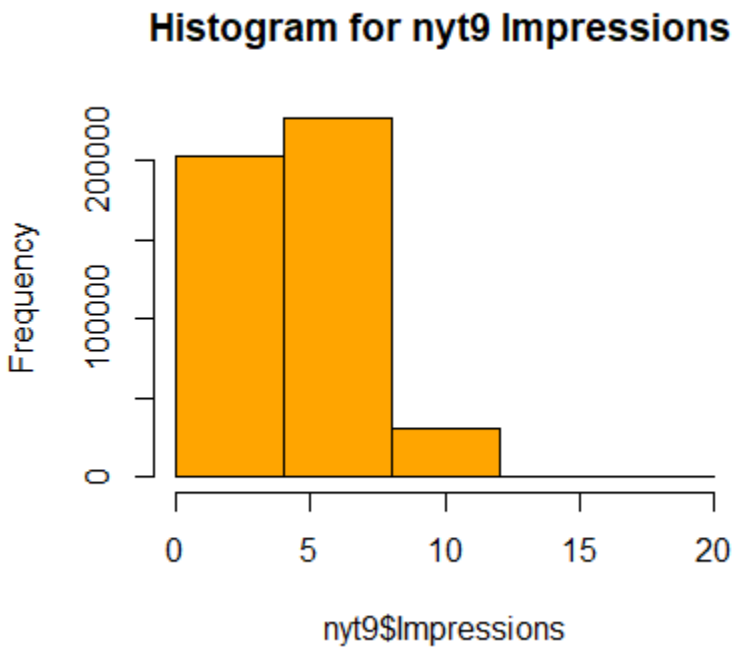
For nyt8,


Histogram for nyt8 Age
Frequency
0
40000
100000
0
20
40
60
80
100
nyt8$Age


Histogram for nyt8 lmpressions
Frequency
0
100000
200000
0
5
10
15
20
nyt8$lmpressions

For nyt9


Histogram for nyt9 Age
Frequency
0
40000
100000
0
20
40
60
80
100
nyt9$Age



Age Distribution (for all datasets):

- The "Age" variable for all datasets appears to be positively skewed, as the mean is less than the median, and the distribution extends to higher ages.
- The minimum age is 0, and the maximum age is around 109, with a median age around 31.
- The distribution is right-skewed, indicating that the majority of observations are concentrated towards the younger age group.
- The boxplots show similar trends across datasets, with some variations in the range and quartiles.

Impressions Distribution (for all datasets):

- The "Impressions" variable also exhibits positive skewness in all datasets. The mean is very close to the median, indicating a relatively symmetric distribution.
- Impressions have a minimum value of 0, with the maximum value around 19, and a median value of 5.
- The distribution of Impressions shows fewer extreme values and is more concentrated around the median compared to the Age distribution.

Differences Among Datasets:

- In terms of Age and Impressions, the seven datasets (nyt3, nyt4, nyt5, nyt6, nyt7, nyt8, nyt9) exhibit similar patterns and have nearly identical summary statistics.

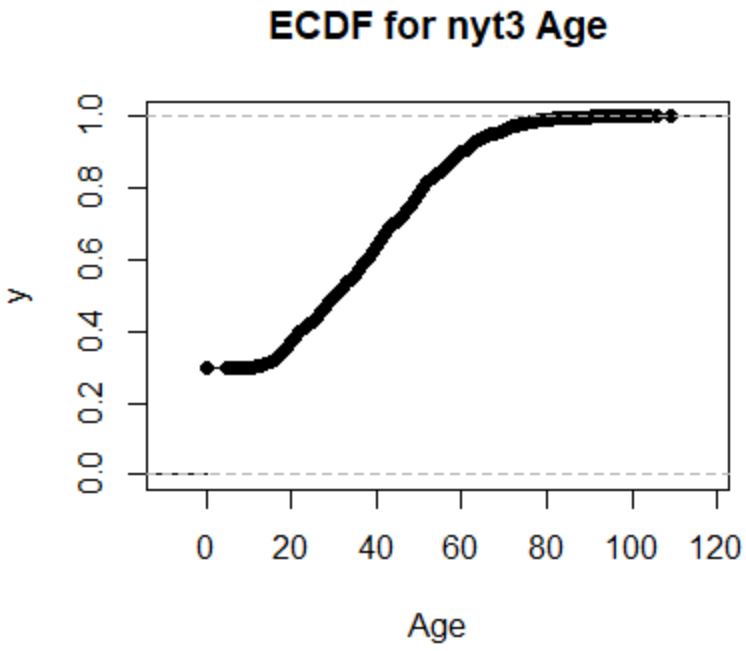
- The distributions are all right-skewed for Age and have a more symmetric distribution for Impressions.
- The distributions are consistent across datasets, with similar central tendencies and ranges.
- The boxplots show consistent characteristics across datasets for both Age and Impressions.

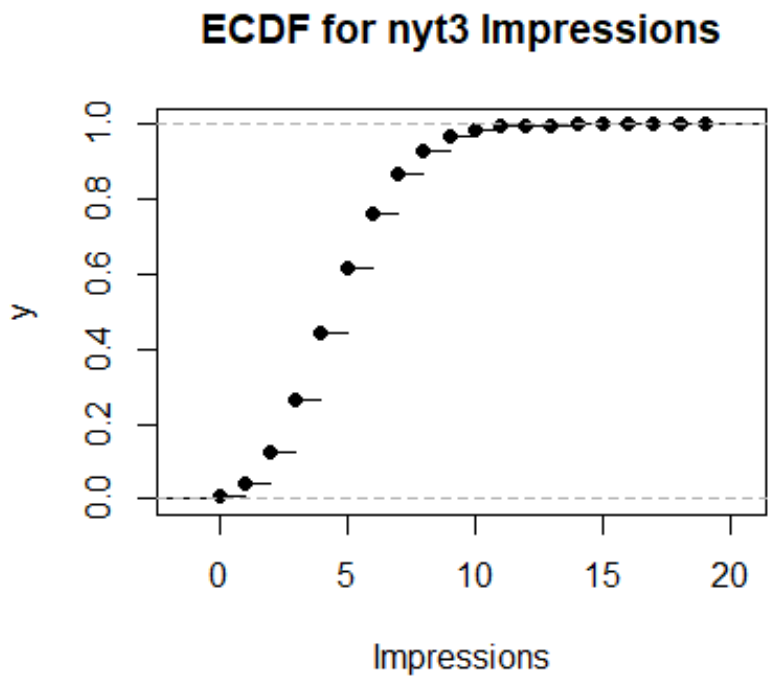
Conclusively, it can be seen that the Age and Impressions variables in these datasets exhibit similar statistical properties and right-skewed, indicating that the distributions are largely consistent among these datasets.

Question c

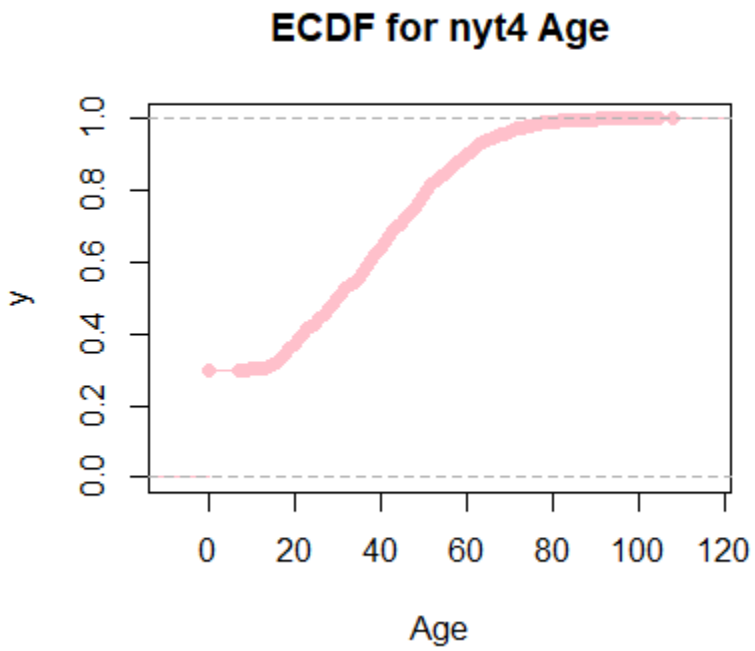
Empirical Cumulative Distribution Function (ECDFs) for nyt3, nyt4, nyt5, nyt6, nyt7, nyt8 and nyt9 Age and Impressions

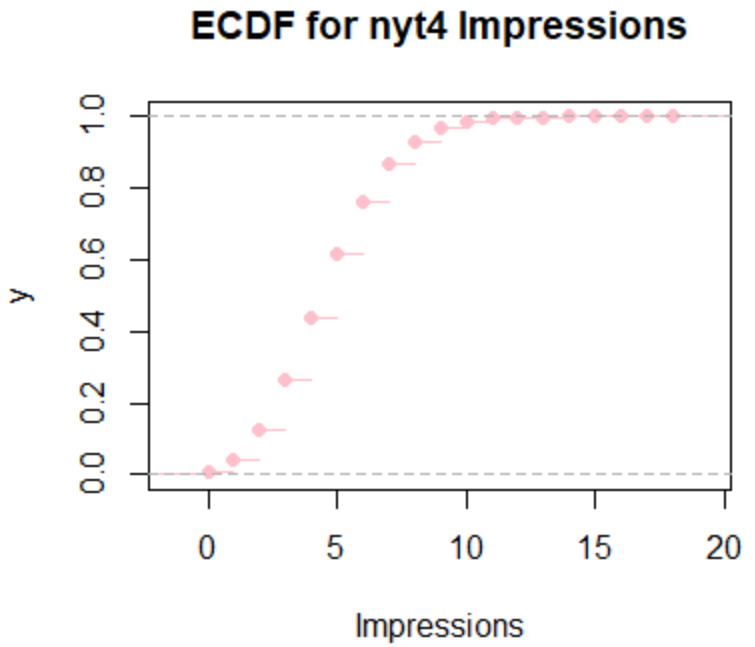
for nyt3



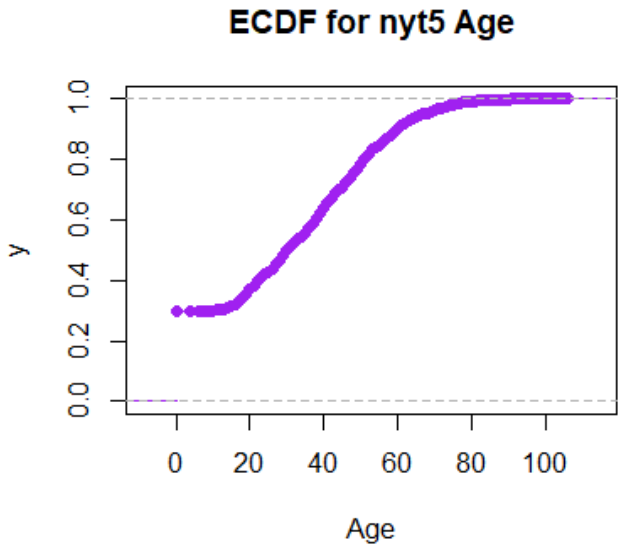

ECDF for nyt3 Impressions
y
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Impressions

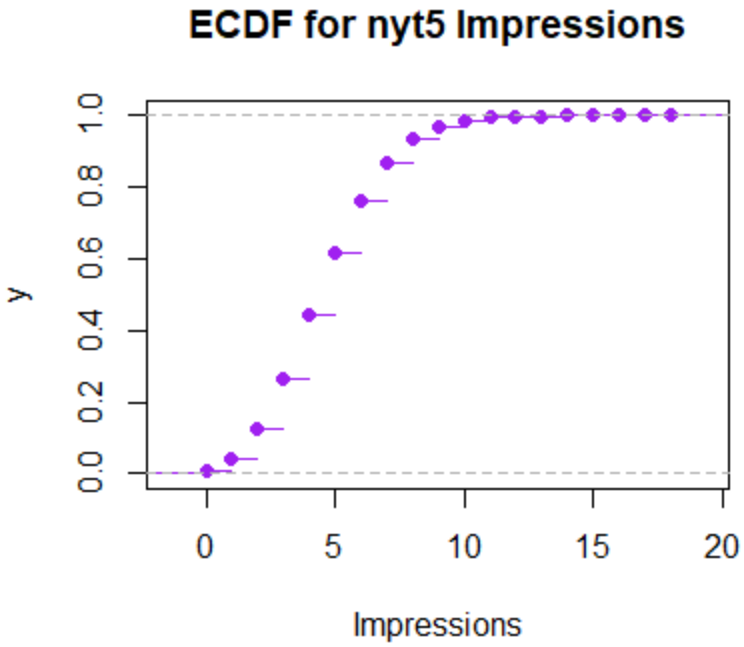
for nyt4


ECDF for nyt4 Age
y
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
Age

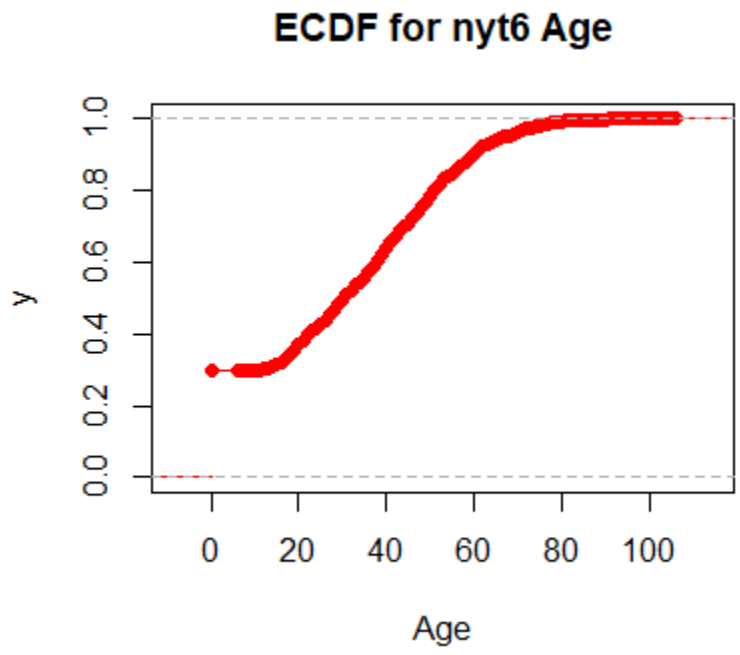

ECDF for nyt4 Impressions
y
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Impressions

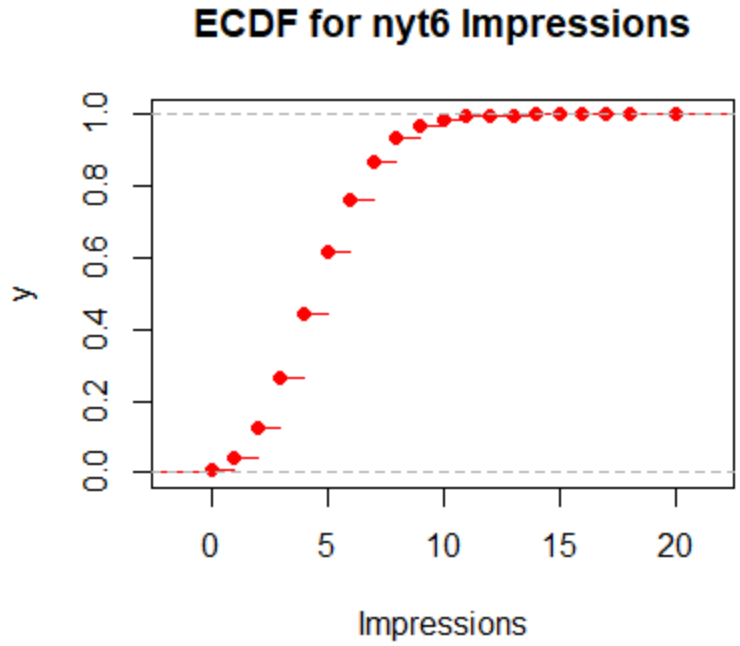
for nyt5


ECDF for nyt5 Age
y
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Age

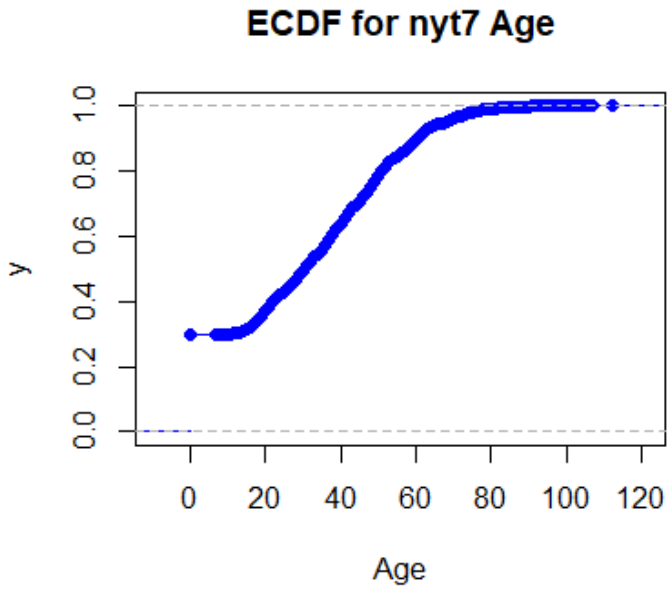

ECDF for nyt5 Impressions
y
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Impressions

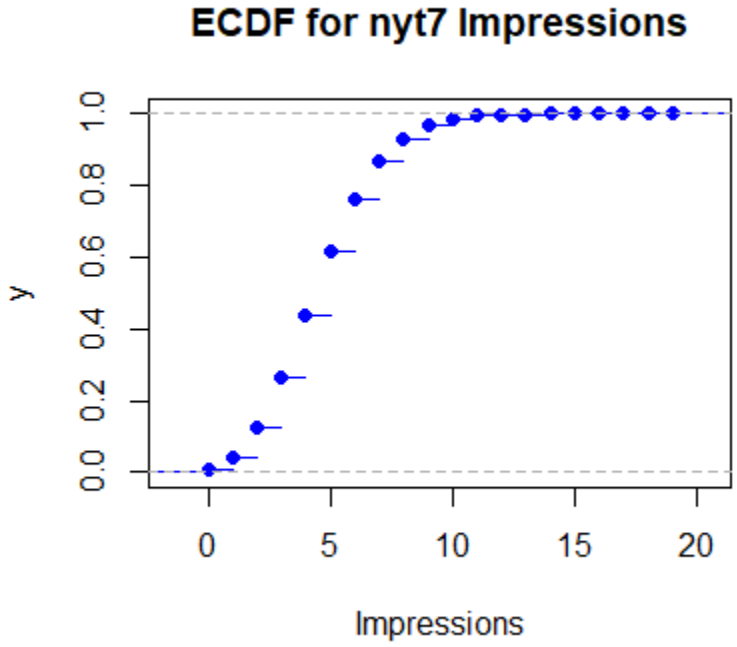
for nyt6


ECDF for nyt6 Age
y
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Age

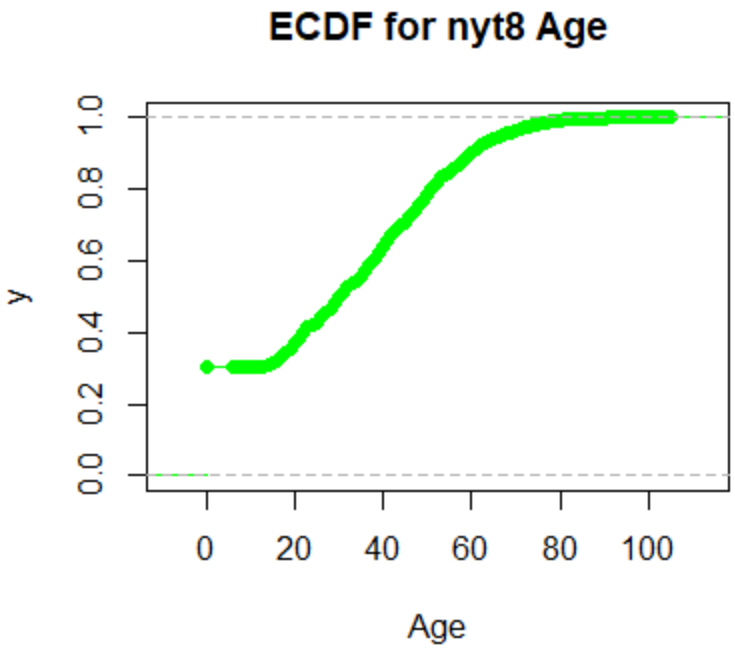

ECDF for nyt6 Impressions
y
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Impressions

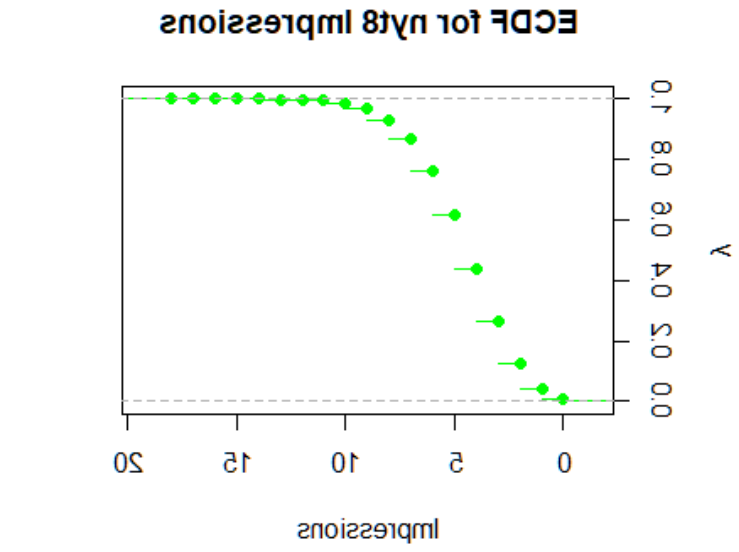
for nyt7


ECDF for nyt7 Age
y
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
Age

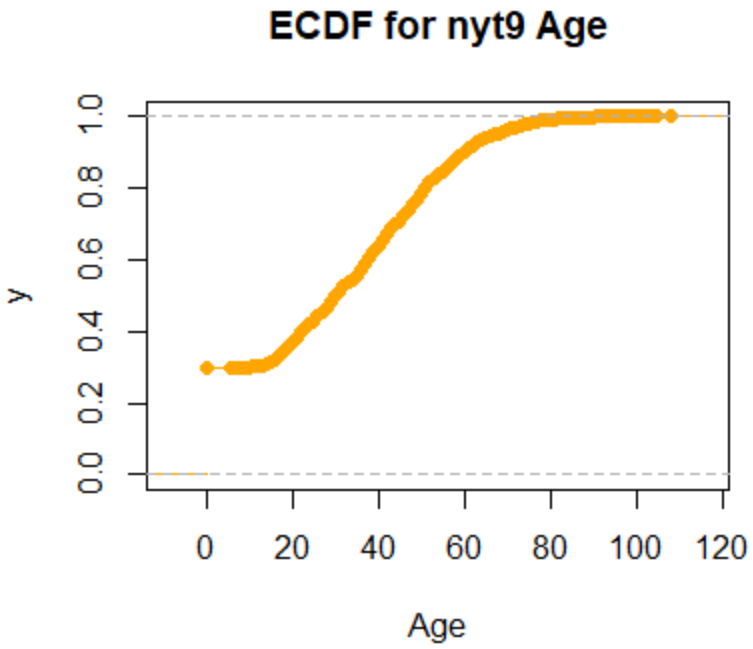

ECDF for nyt7 Impressions
y
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Impressions

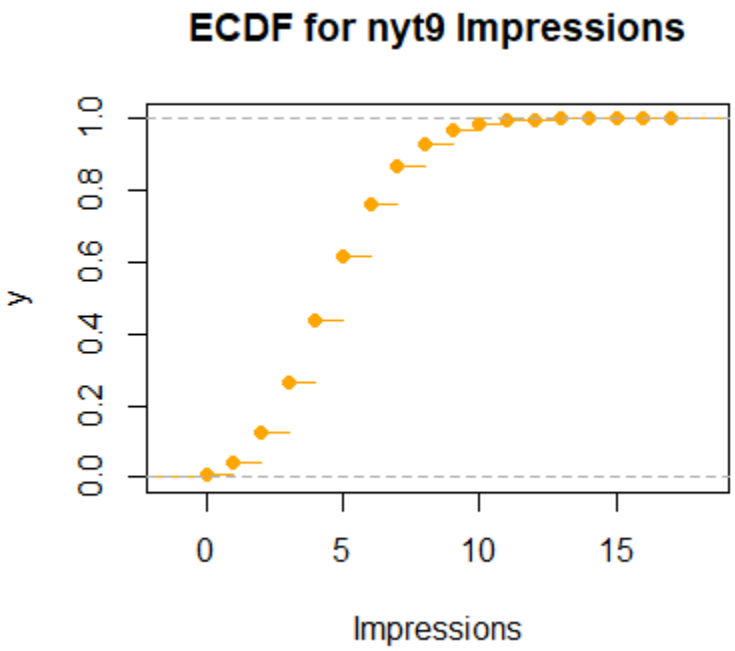
for nyt8


ECDF for nyt8 Age
y
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Age


ECDF for nyt8 Impressions
y
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Impressions

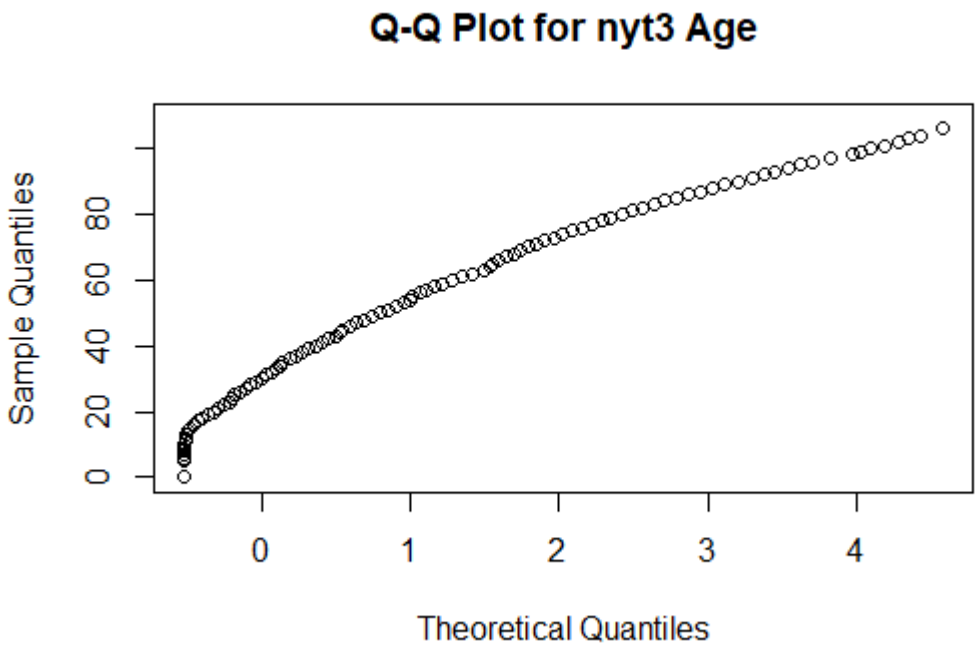
for nyt9

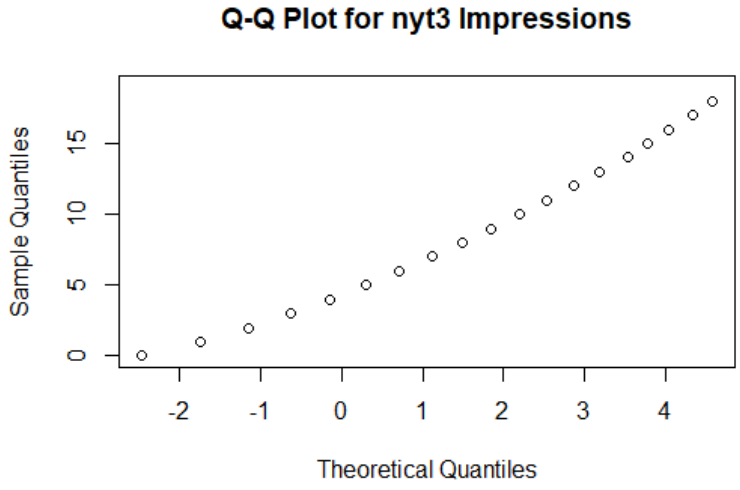




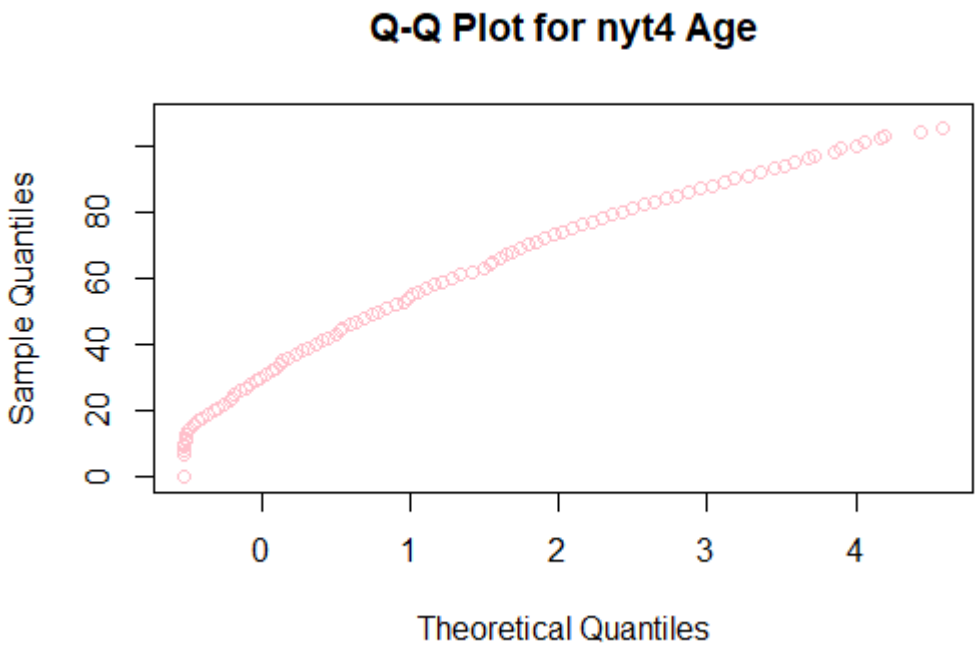
Q-Q plot for nyt3, nyt4, nyt5, nyt6, nyt7, nyt8 and nyt9 Age and Impressions

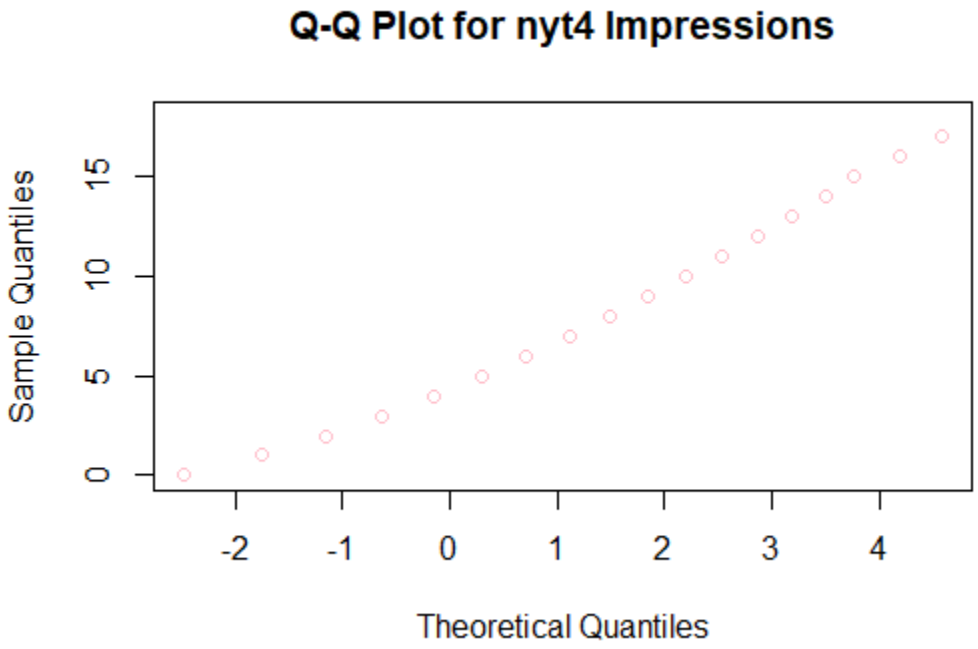
For nyt3


Q-Q Plot for nyt3 Age
Sample Quantiles
0
20
40
60
80
0
1
2
3
4
Theoretical Quantiles

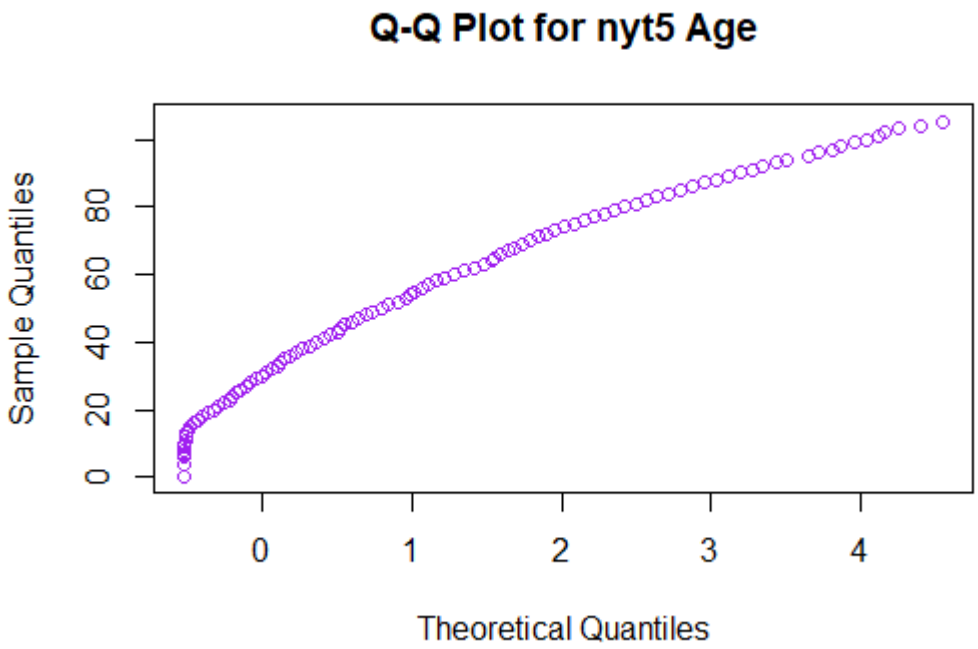

Q-Q Plot for nyt3 Impressions
Sample Quantiles
0
5
10
15
-2
-1
0
1
2
3
4
Theoretical Quantiles

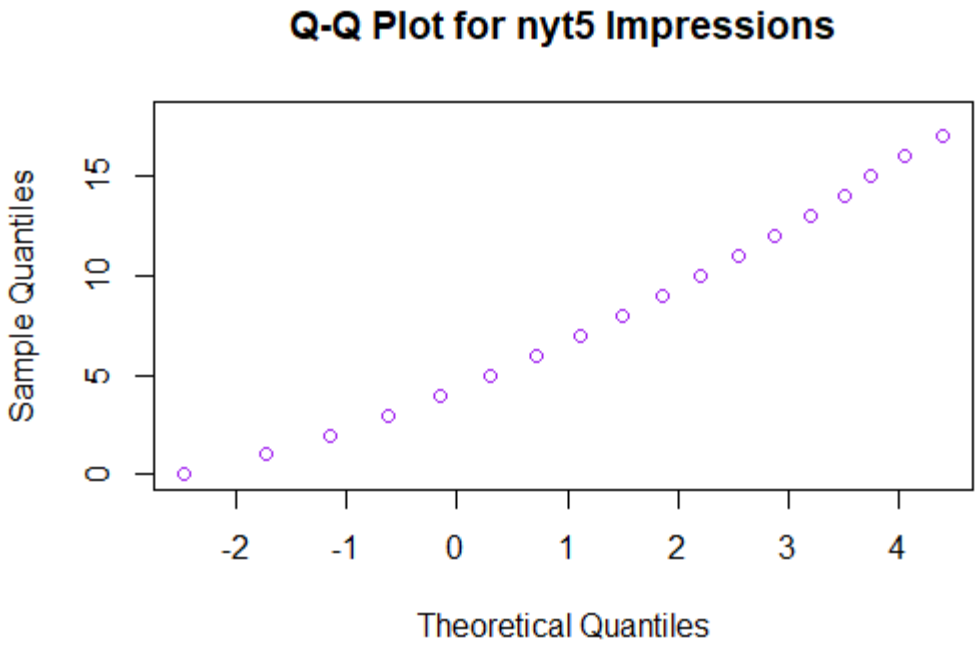
For nyt4


Q-Q Plot for nyt4 Age
Sample Quantiles
0
20
40
60
80
0
1
2
3
4
Theoretical Quantiles

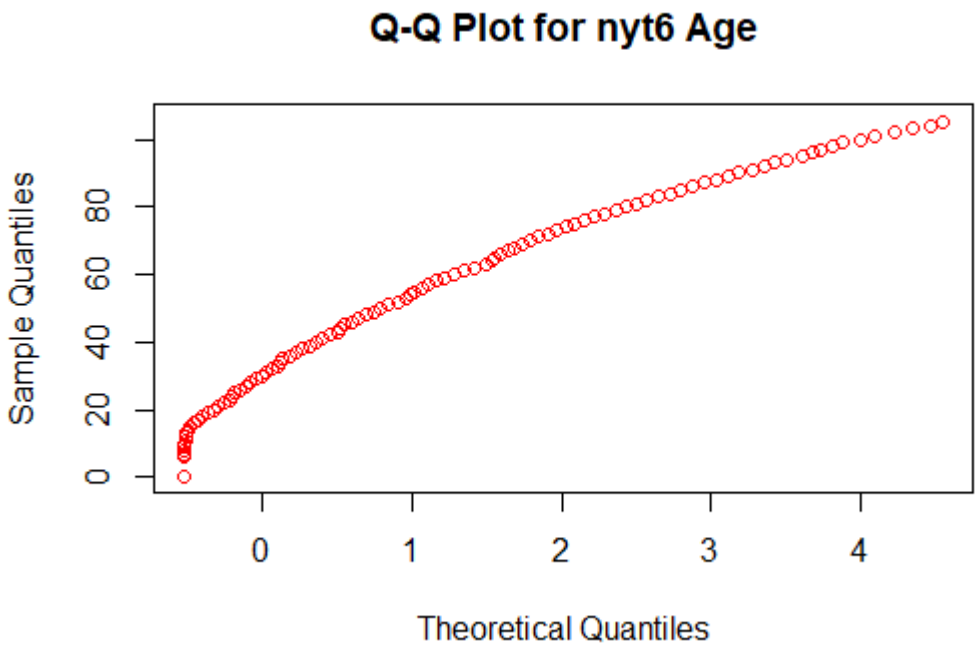

Q-Q Plot for nyt4 Impressions
Sample Quantiles
0
5
10
15
-2
-1
0
1
2
3
4
Theoretical Quantiles

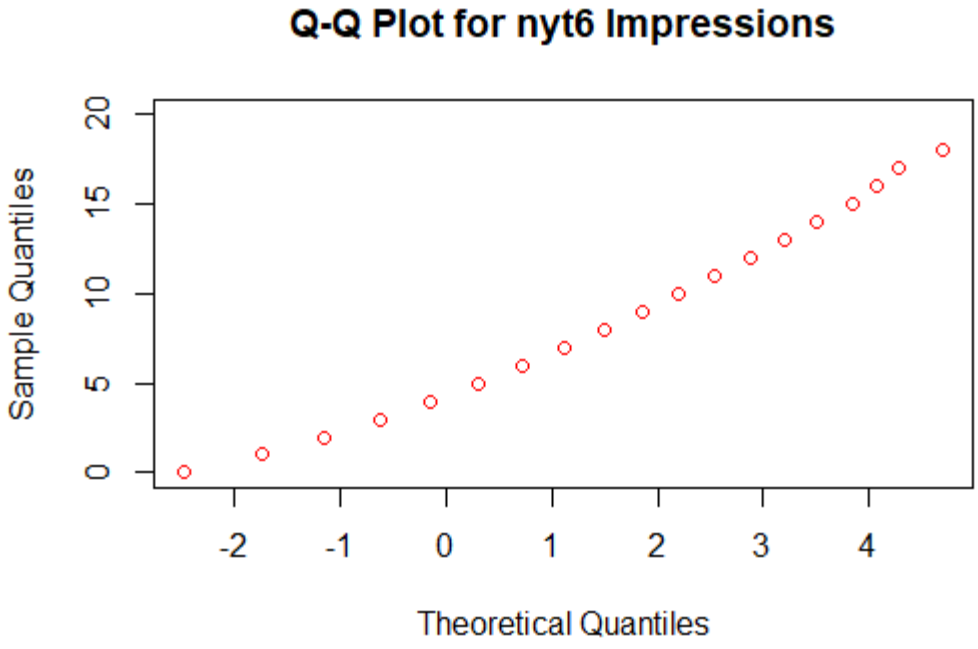
For nyt5


Q-Q Plot for nyt5 Age
Sample Quantiles
0
20
40
60
80
0
1
2
3
4
Theoretical Quantiles

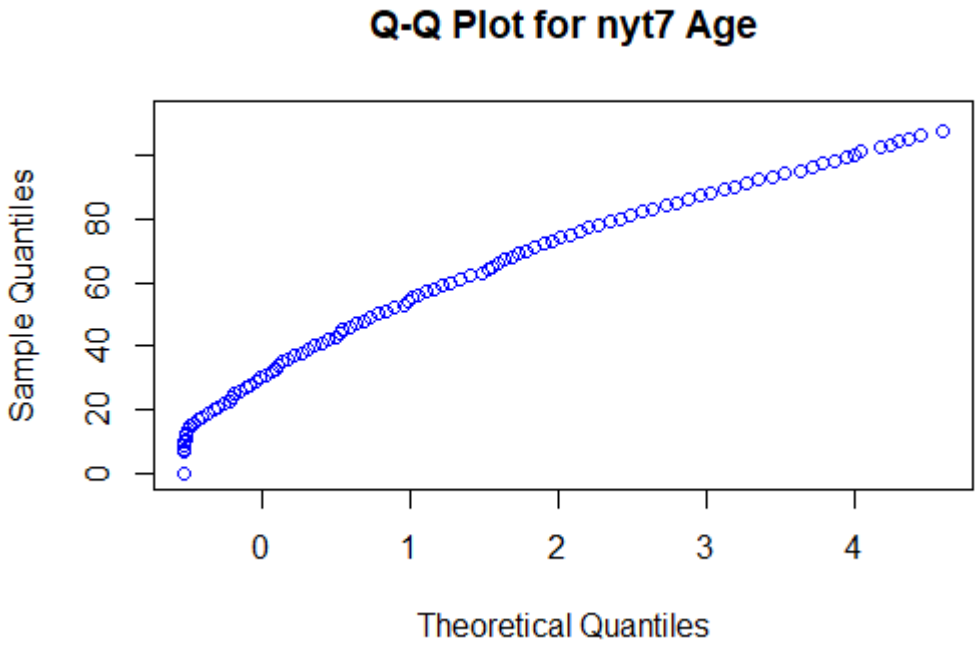

Q-Q Plot for nyt5 Impressions
Sample Quantiles
0
5
10
15
-2
-1
0
1
2
3
4
Theoretical Quantiles

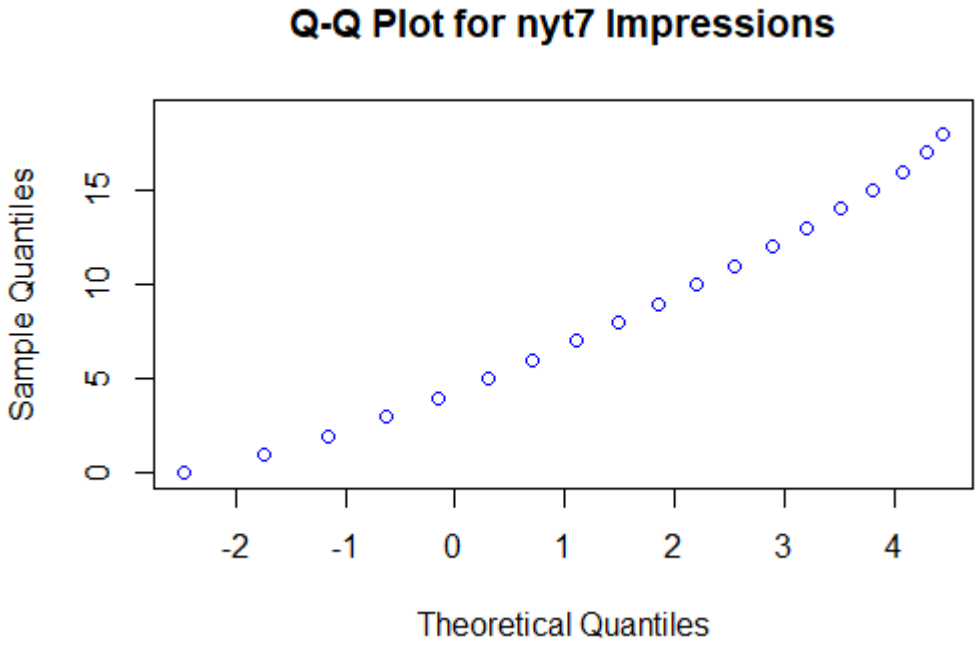
For nyt6


Q-Q Plot for nyt6 Age
Sample Quantiles
0
20
40
60
80
0
1
2
3
4
Theoretical Quantiles

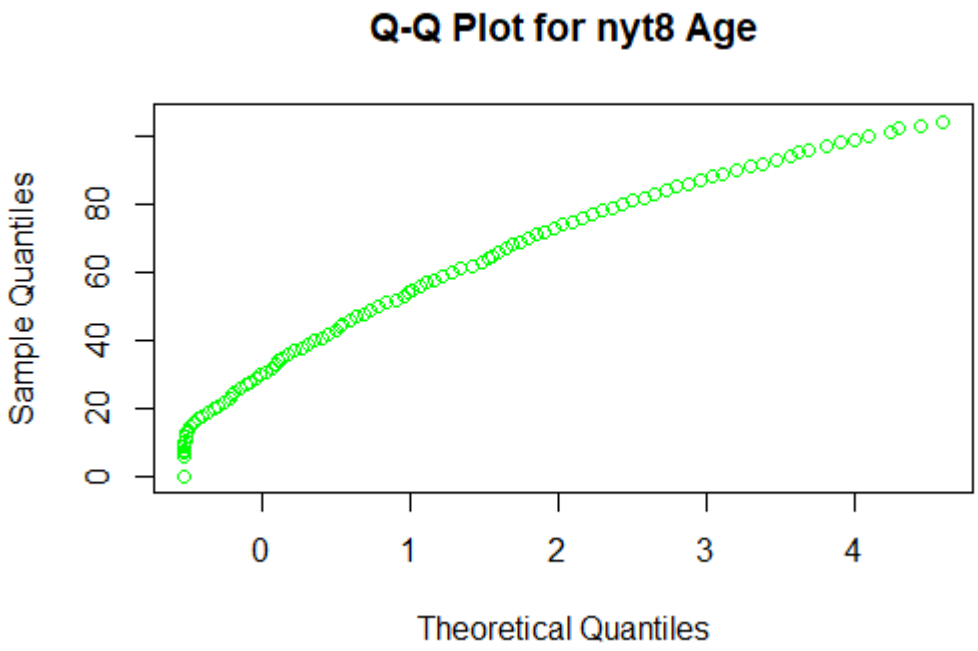

Q-Q Plot for nyt6 Impressions
Sample Quantiles
0
5
10
15
20
-2
-1
0
1
2
3
4
Theoretical Quantiles

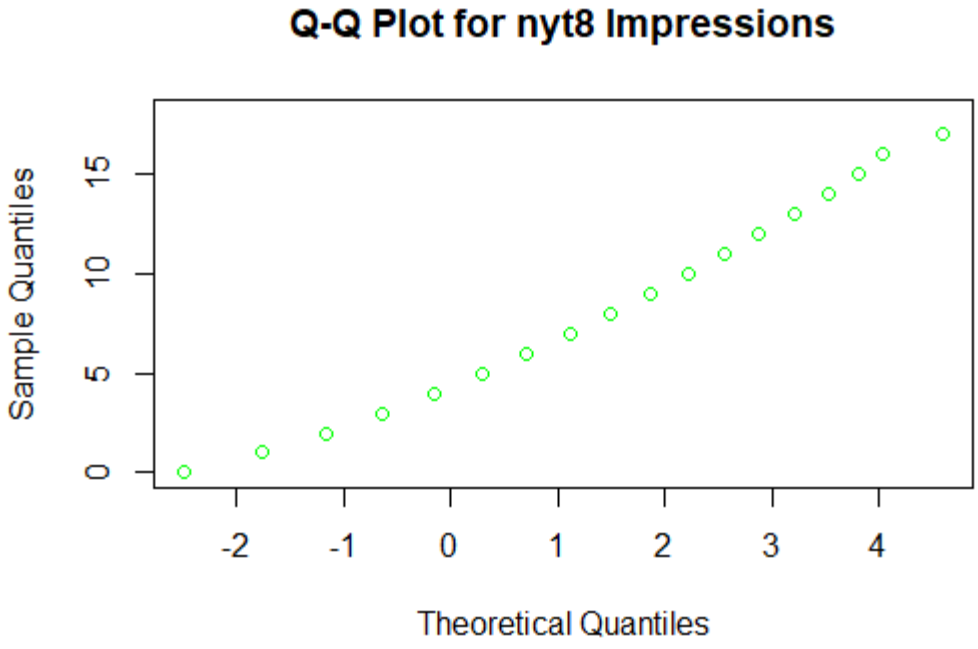
For nyt7



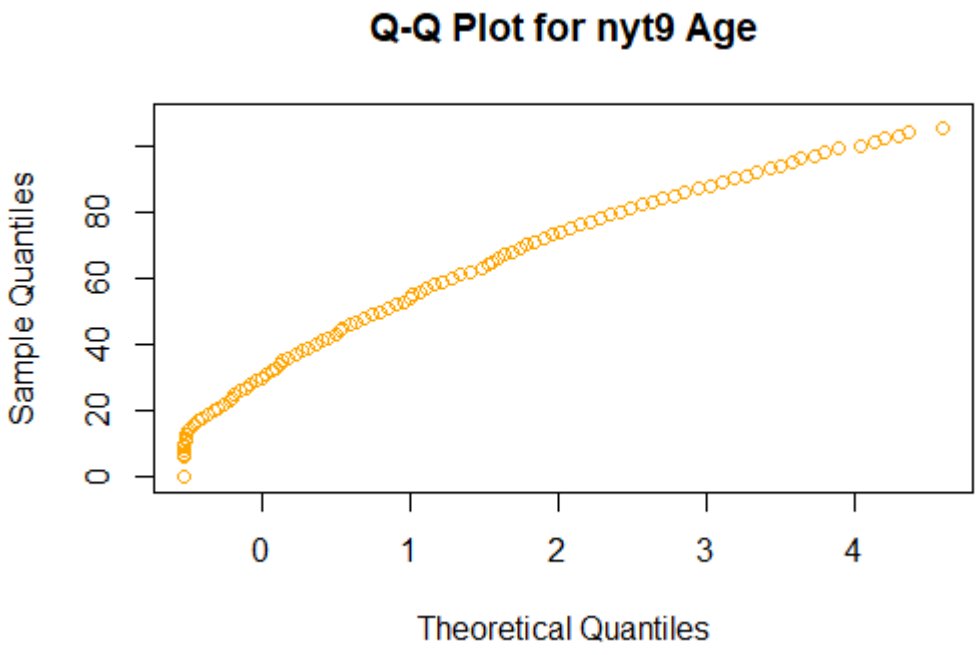


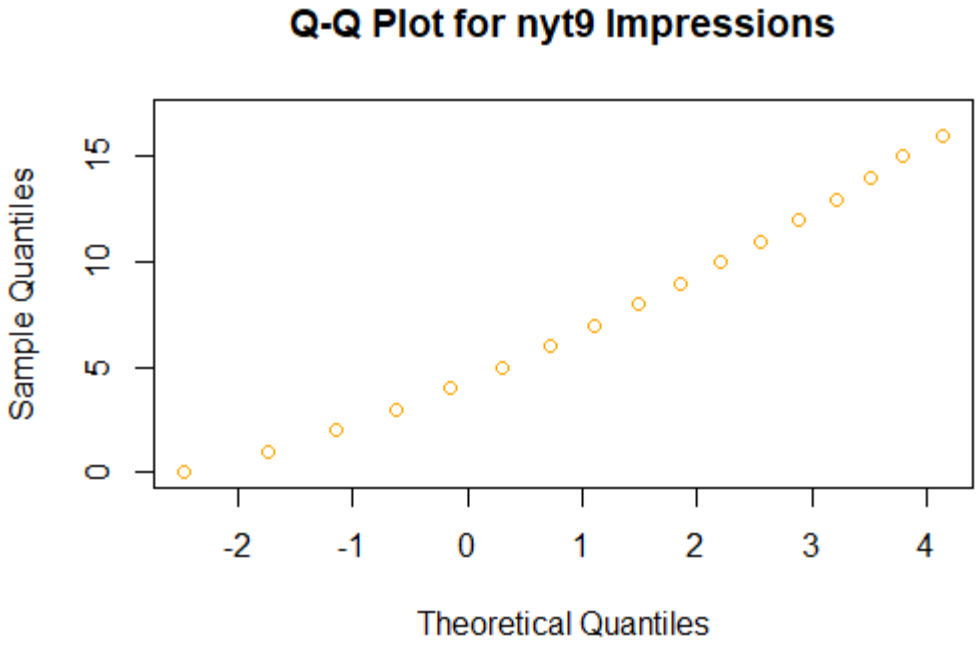
For nyt8


Q-Q Plot for nyt8 Age
Sample Quantiles
0
20
40
60
80
0
1
2
3
4
Theoretical Quantiles


Q-Q Plot for nyt8 Impressions
Sample Quantiles
0
5
10
15
-2
-1
0
1
2
3
4
Theoretical Quantiles

For nyt9


Q-Q Plot for nyt9 Age
Sample Quantiles
0
20
40
60
80
0
1
2
3
4
Theoretical Quantiles


Q-Q Plot for nyt9 Impressions
Sample Quantiles
0
5
10
15
-2
-1
0
1
2
3
4
Theoretical Quantiles

From the ECDFs of Impressions all the values are centered around 5 which is the expected the median. The quantile-quantile plots also follow a near straight line and hence we can conclude that the distribution is normal between 0 - 10. The same can be said about the Age variable, which follows an almost normal distribution between 20 to 60.

Question d

Significance Testing for nyt3, nyt4, nyt5, nyt6, nyt7, nyt8 and nyt9 Age and Impressions

I used T- test for my Significance Testing.T-test, is a statistical hypothesis test used to determine whether there is a significant difference between the means of two groups or conditions.
T-Test Results for Age and Impressions Data:

nyt3:

- t-value: 683.93
- Degrees of Freedom: 448,241
- p-value: < 2.2e-16 (essentially zero)

A highly significant difference between the means of "nyt3$Age" and "nyt3$Impressions" was observed. The 95% confidence interval for the difference in means (24.40155, 24.54181) does not contain zero, indicating a substantial difference.

nyt4:

- t-value: 684.26
- Degrees of Freedom: 450,778
- p-value: < 2.2e-16 (essentially zero)

A highly significant difference between the means of "nyt4$Age" and "nyt4$Impressions" was observed. The 95% confidence interval for the difference in means (24.35846, 24.49841) does not contain zero, indicating a substantial difference.

nyt5:

- t-value: 626.03
- Degrees of Freedom: 376,957
- p-value: < 2.2e-16 (essentially zero)

A highly significant difference between the means of "nyt5$Age" and "nyt5$Impressions" was observed. The 95% confidence interval for the difference in means (24.35811, 24.51111) does not contain zero, indicating a substantial difference.

nyt6:

- t-value: 901.74

- Degrees of Freedom: 778,197
- p-value: < 2.2e-16 (essentially zero)

A very substantial difference between the means of "nyt6$Age" and "nyt6$Impressions" was observed. The 95% confidence interval for the difference in means (24.40973, 24.51607) does not contain zero, indicating a substantial difference.

nyt7:

- t-value: 693.47
- Degrees of Freedom: 460,565
- p-value: < 2.2e-16 (essentially zero)

A substantial difference between the means of "nyt7$Age" and "nyt7$Impressions" was observed. The 95% confidence interval for the difference in means (24.4504, 24.5890) does not contain zero, indicating a substantial difference.

nyt8:

- t-value: 699.93
- Degrees of Freedom: 471,444
- p-value: < 2.2e-16 (essentially zero)

A substantial difference between the means of "nyt8$Age" and "nyt8$Impressions" was observed. The 95% confidence interval for the difference in means (24.34483, 24.48156) does not contain zero, indicating a substantial difference.

nyt9:

- t-value: 697.77
- Degrees of Freedom: 467,674
- p-value: < 2.2e-16 (essentially zero)

A substantial difference between the means of "nyt9$Age" and "nyt9$Impressions" was observed. The 95% confidence interval for the difference in means (24.38298, 24.52034) does not contain zero, indicating a substantial difference.

Conclusively, in all datasets, the extremely low p-values and non-overlapping confidence intervals provide strong evidence against the null hypothesis.

Question e

observations about the datasets/ variables

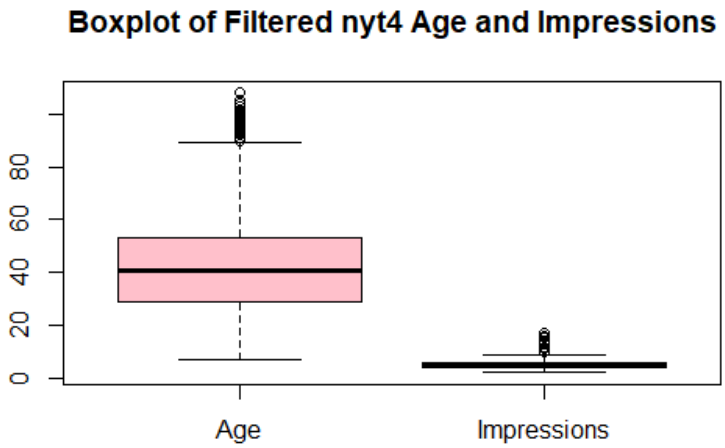
I observed the graphical representations of the data for both boxplots and histograms, when generated, exhibited strikingly similar patterns across all four datasets (nyt3, nyt4, nyt5, and nyt6). Also, the right-skewed pattern was consistent among the age and impressions distributions graphical representations. I noticed that that the datasets contain a significant number of zero values

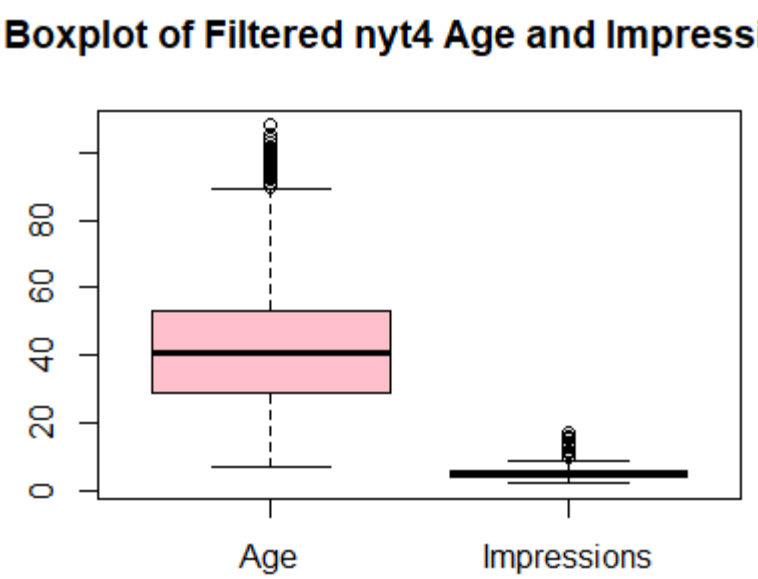
In addition to this, the results of the Anderson-Darling normality tests; obtained p-values were consistently less than 0.05 for both age and impressions data, signaling a clear departure from normal distribution. This reinforces the understanding that these datasets do not conform to a typical bell-shaped curve.These findings collectively underscore the uniqueness of the datasets and their non-normally distributed nature.

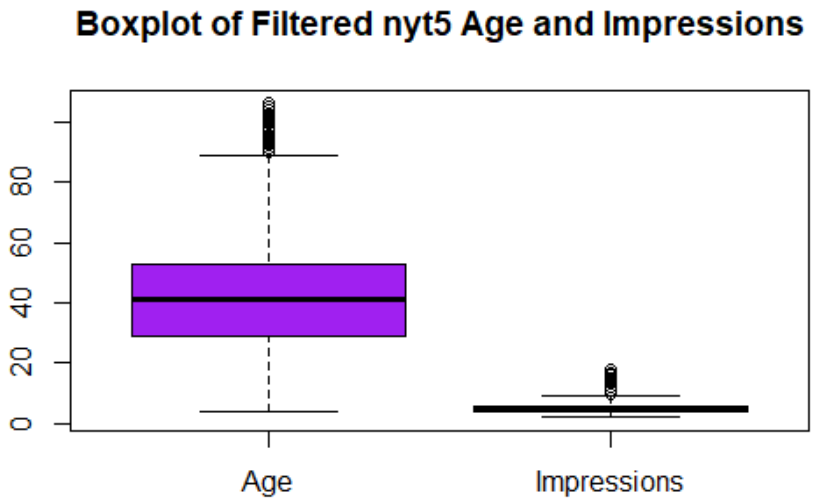
Question 2

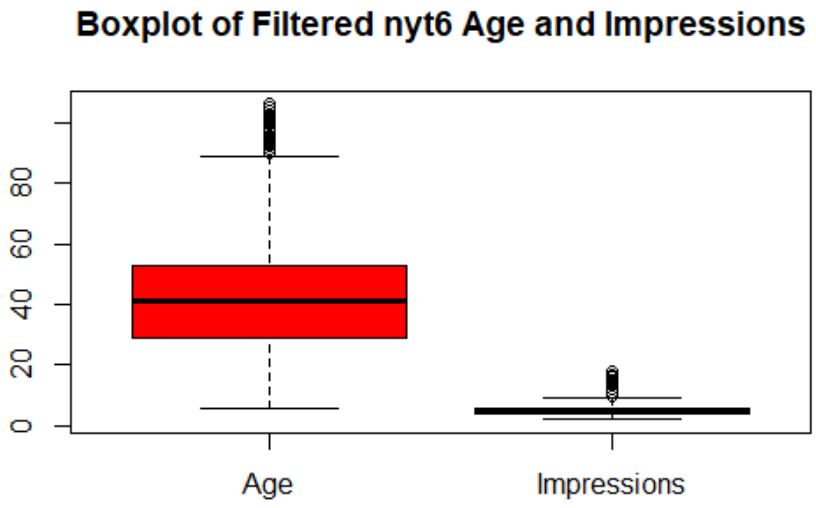
Filtered the distributions for Age and Impressions for nyt3, nyt4, nyt5 and nyt6

Box plot

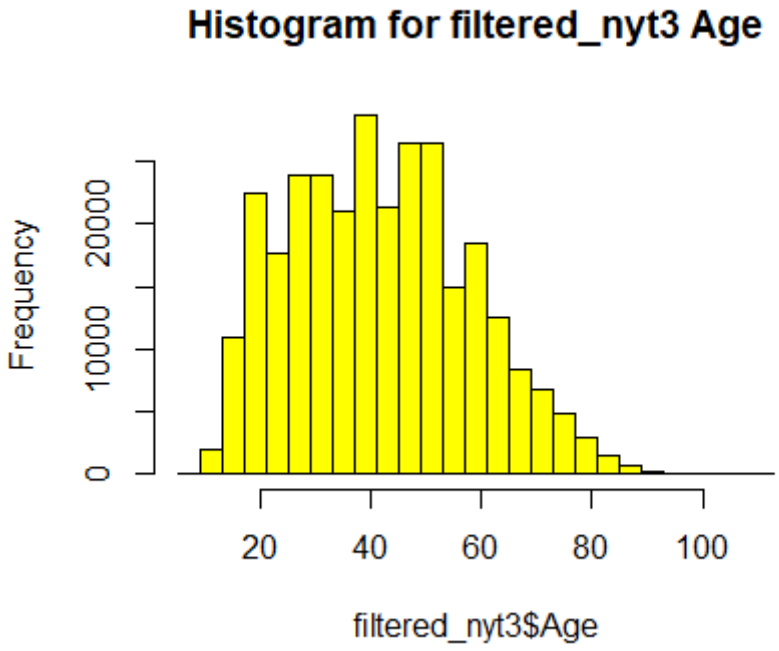


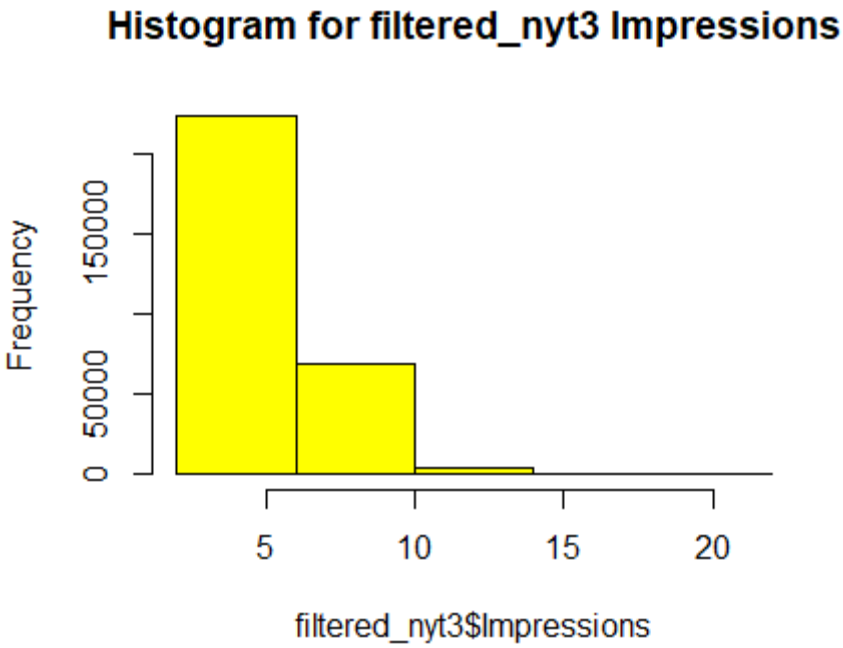

Boxplot of Filtered nyt4 Age and Impressions
0
20
40
60
80
Age
Impressions

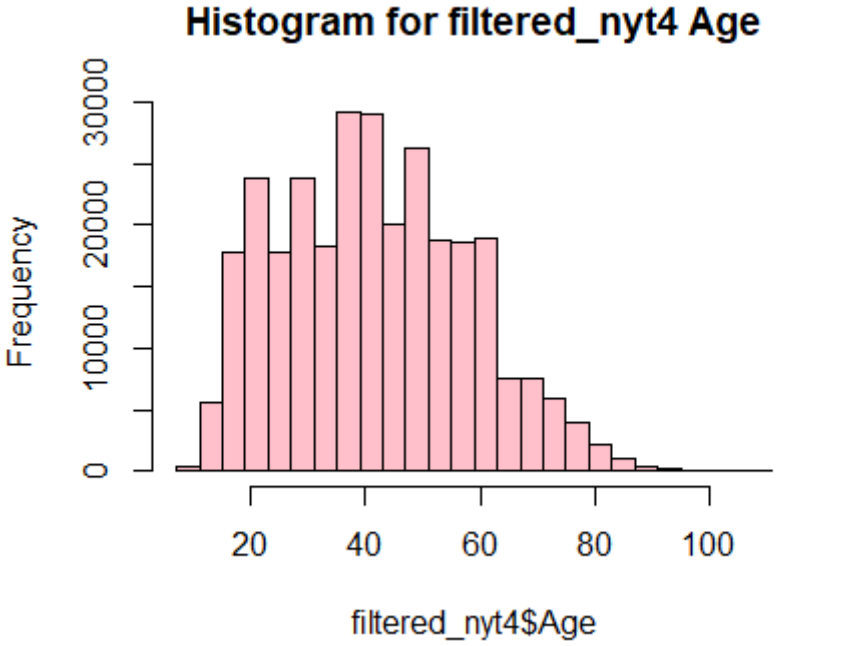

Boxplot of Filtered nyt5 Age and Impressions
0
20
40
60
80
Age
Impressions

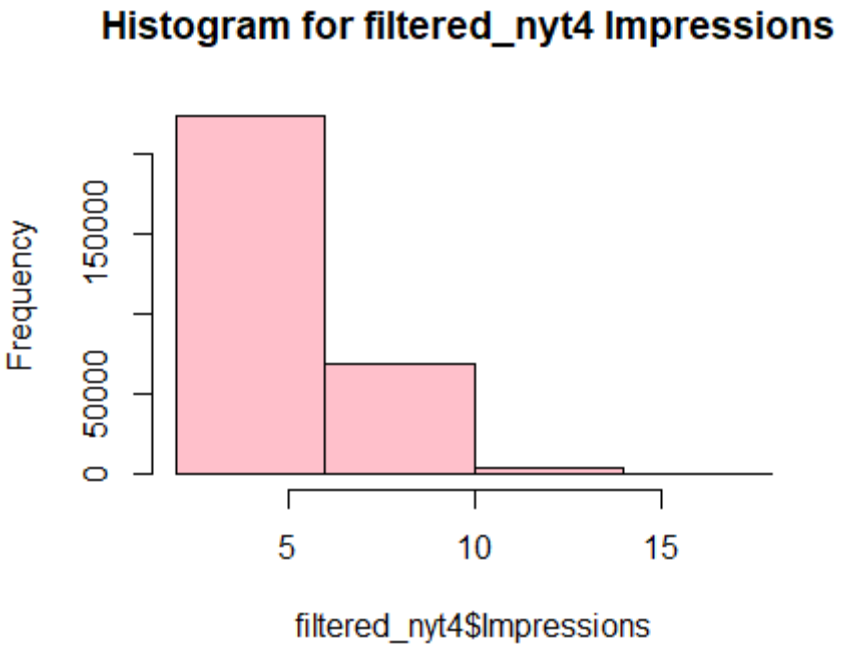

Boxplot of Filtered nyt6 Age and Impressions
0
20
40
60
80
Age
Impressions

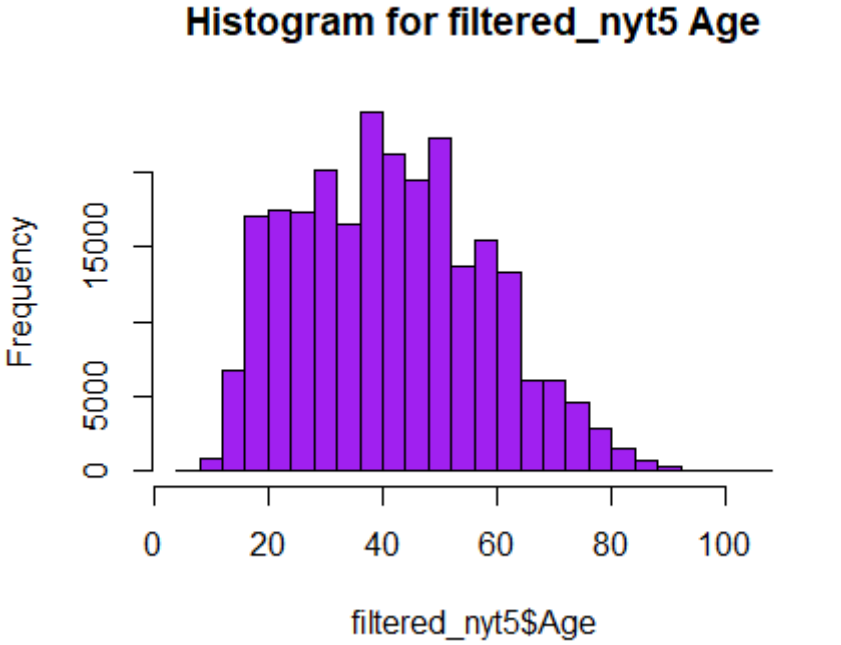
Histogram

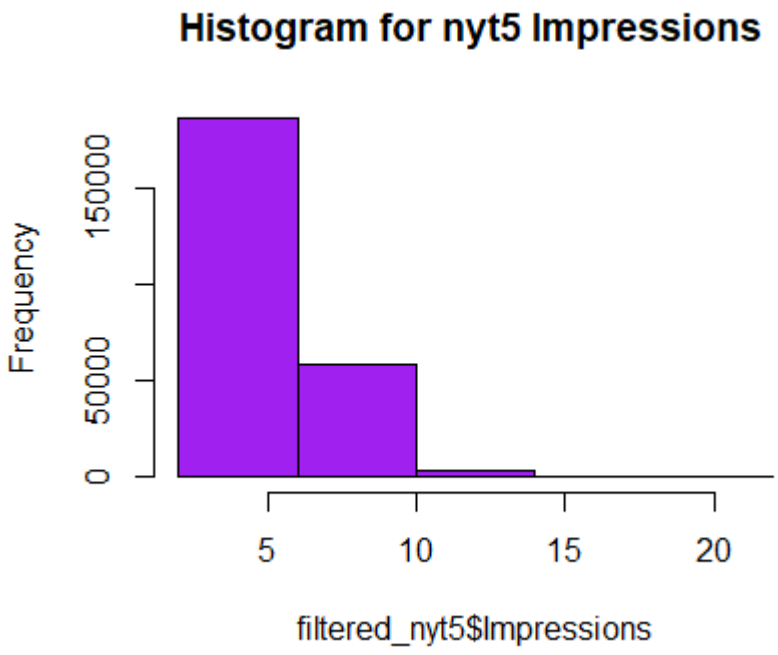

Histogram for filtered_nyt3 Age
Frequency
0
10000
20000
20
40
60
80
100
filtered_nyt3$Age

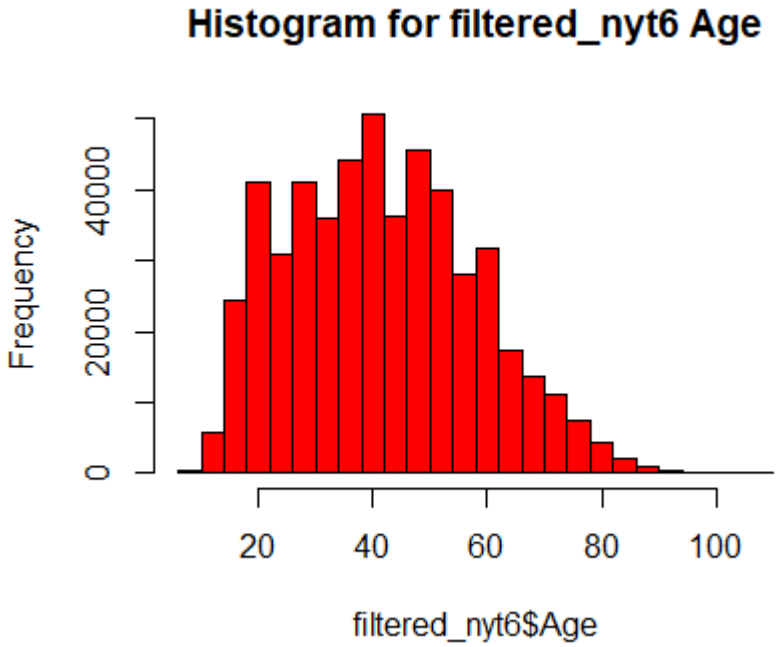

Histogram for filtered_nyt3 Impressions
Frequency
0
50000
150000
5
10
15
20
filtered_nyt3$Impressions

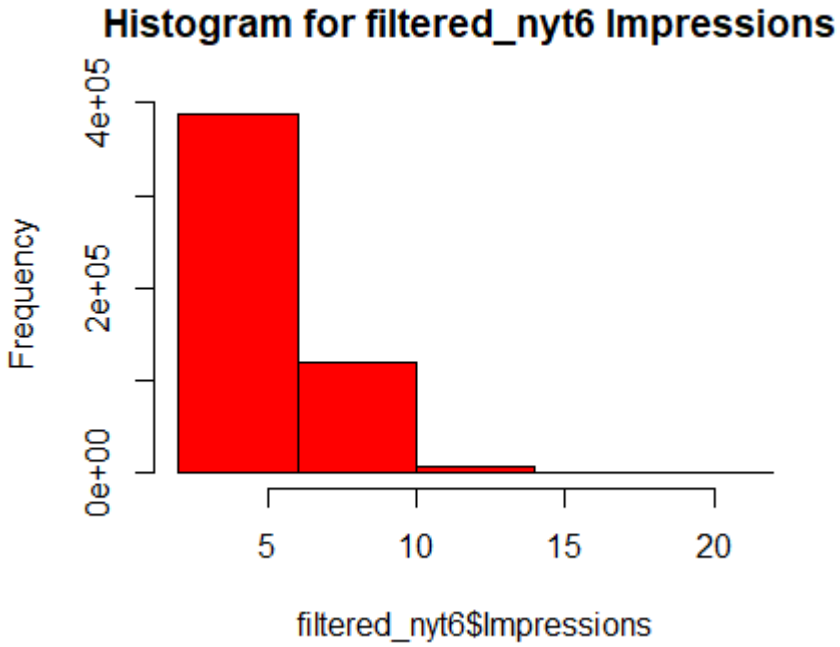

Histogram for filtered_nyt4 Age
Frequency
0
10000
20000
30000
20
40
60
80
100
filtered_nyt4$Age


Histogram for filtered_nyt4 Impressions
Frequency
0
50000
150000
5
10
15
filtered_nyt4$Impressions

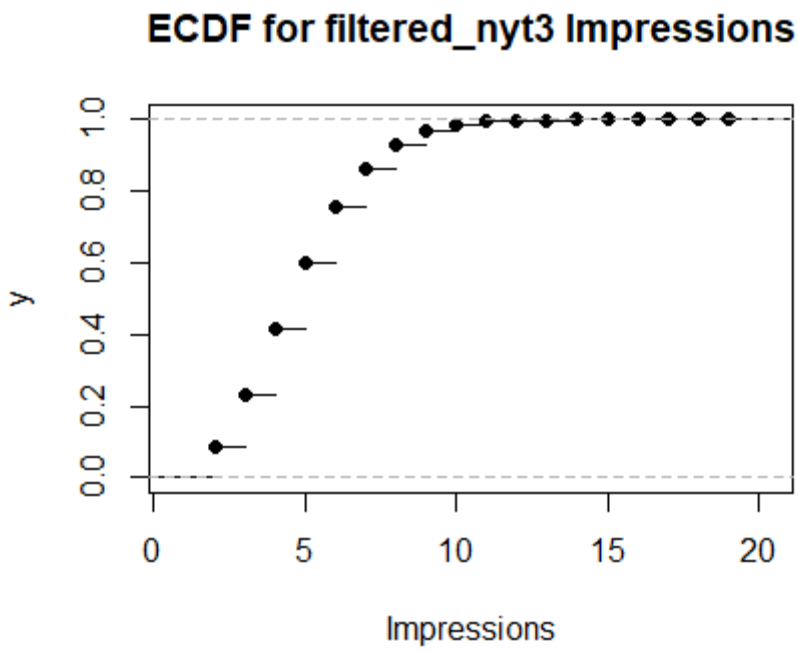

Histogram for filtered_nyt5 Age
Frequency
0
5000
15000
0
20
40
60
80
100
filtered_nyt5$Age


Histogram for nyt5 Impressions
Frequency
0
50000
150000
5
10
15
20
filtered_nyt5$Impressions

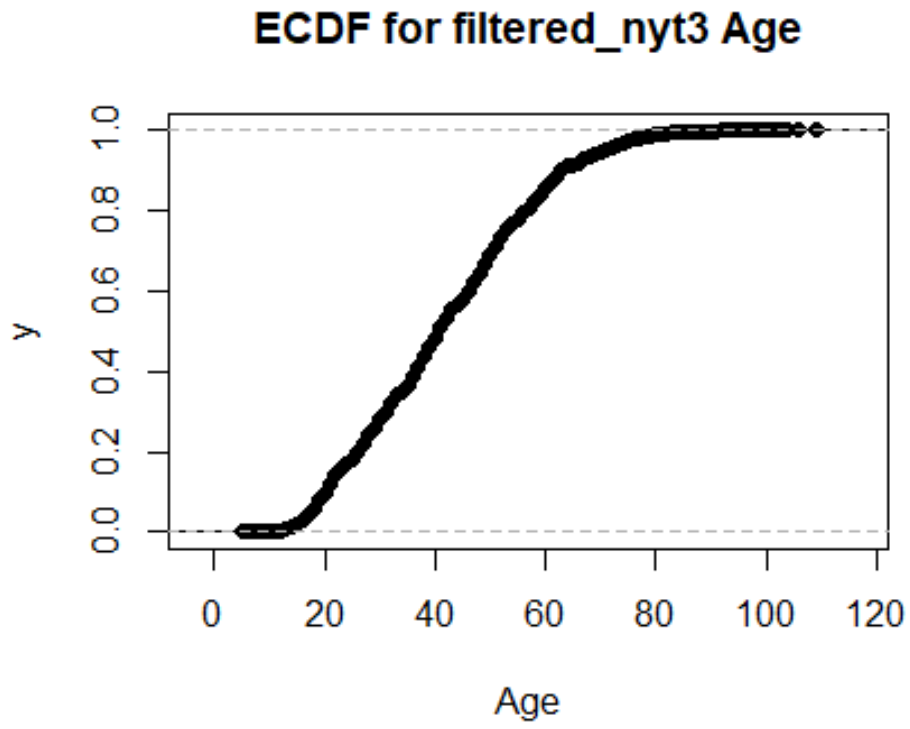

Histogram for filtered_nyt6 Age
Frequency
0
20000
40000
20
40
60
80
100
filtered_nyt6$Age

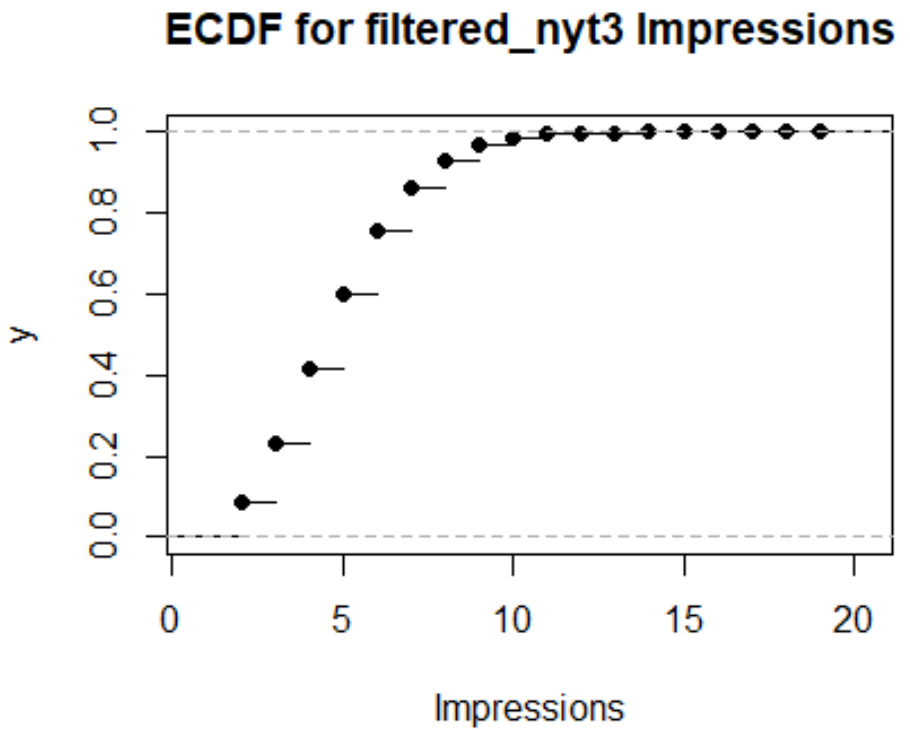

Histogram for filtered_nyt6 Impressions
Frequency
0e+00
2e+05
4e+05
5
10
15
20
filtered_nyt6$Impressions

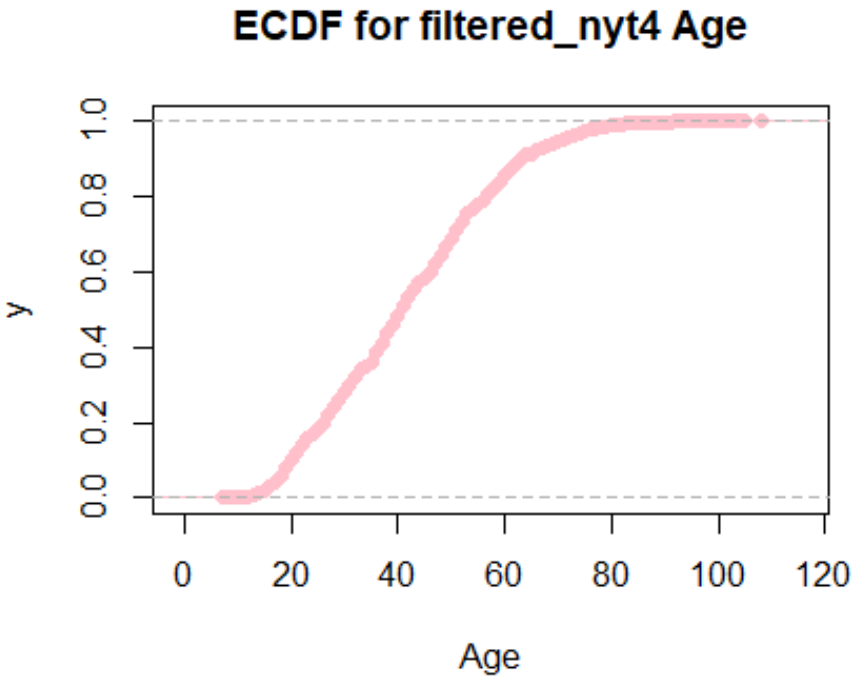
ECDF

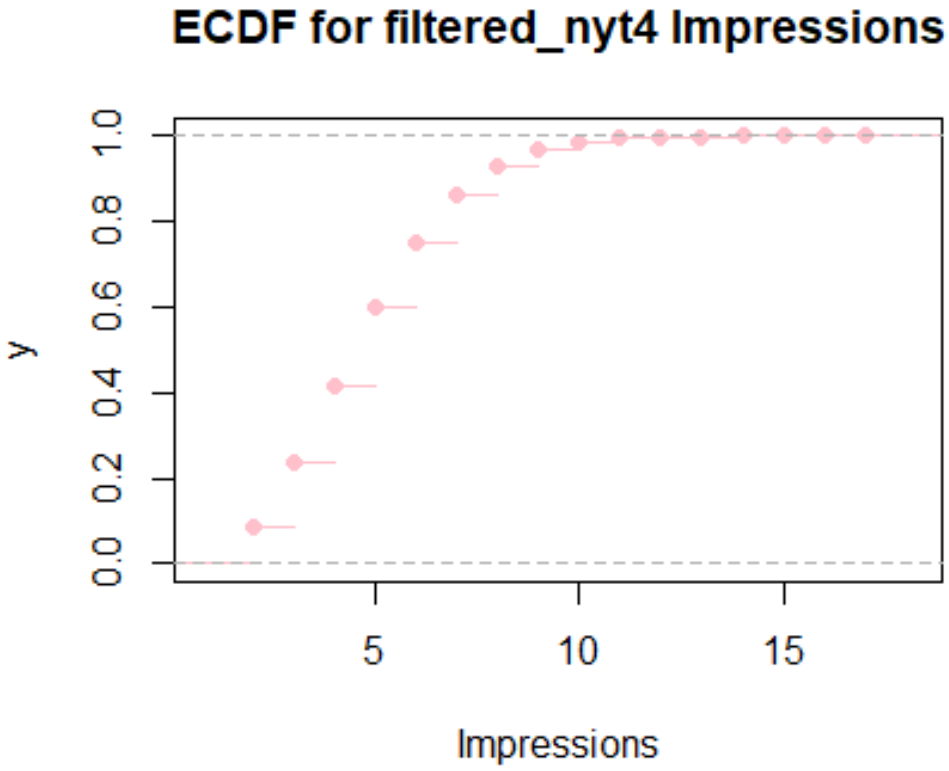

ECDF for filtered_nyt3 Impressions
y
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Impressions

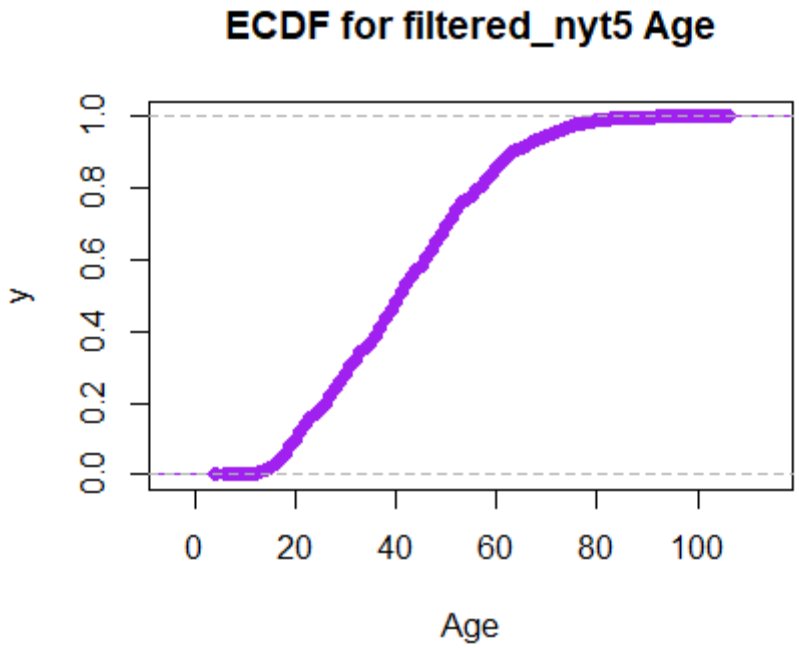
ECDF

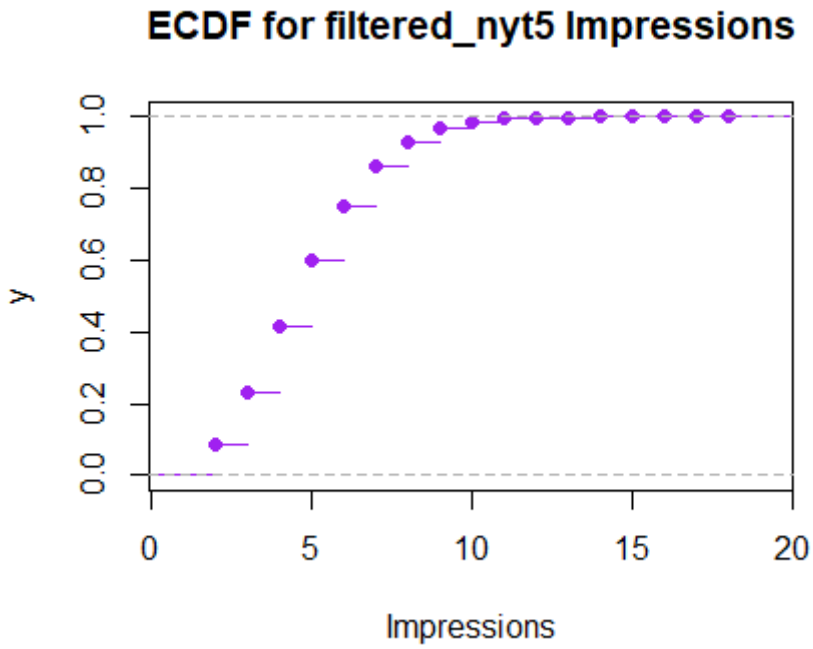

ECDF for filtered_nyt3 Age
y
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
Age

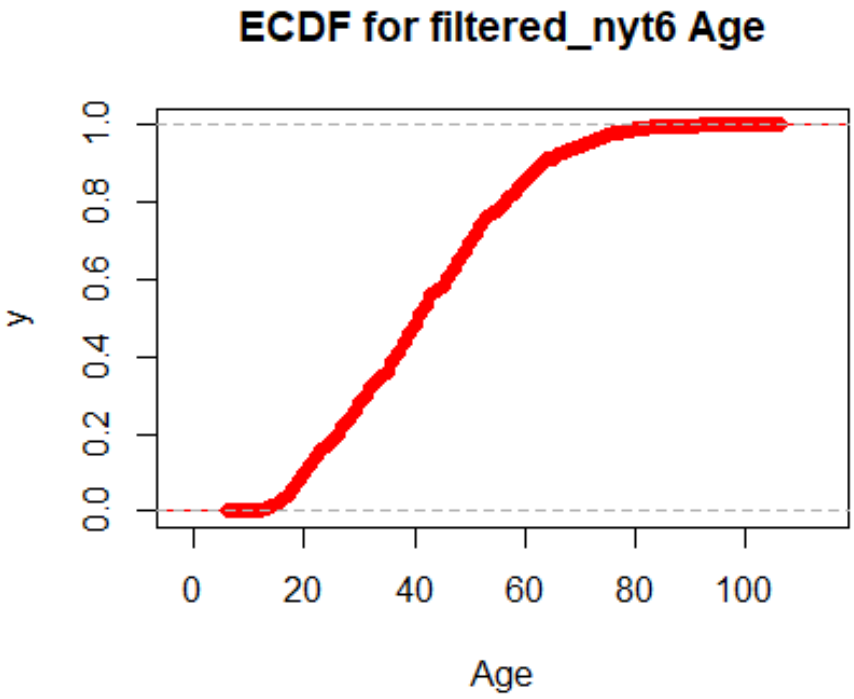

ECDF for filtered_nyt3 Impressions
y
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Impressions

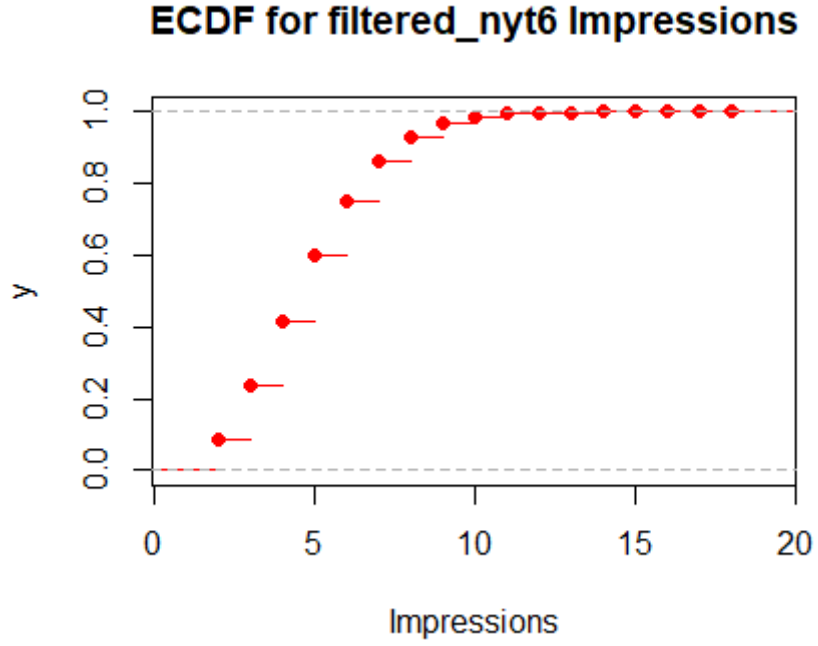

ECDF for filtered_nyt4 Age
y
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
Age


ECDF for filtered_nyt4 Impressions
y
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
Impressions

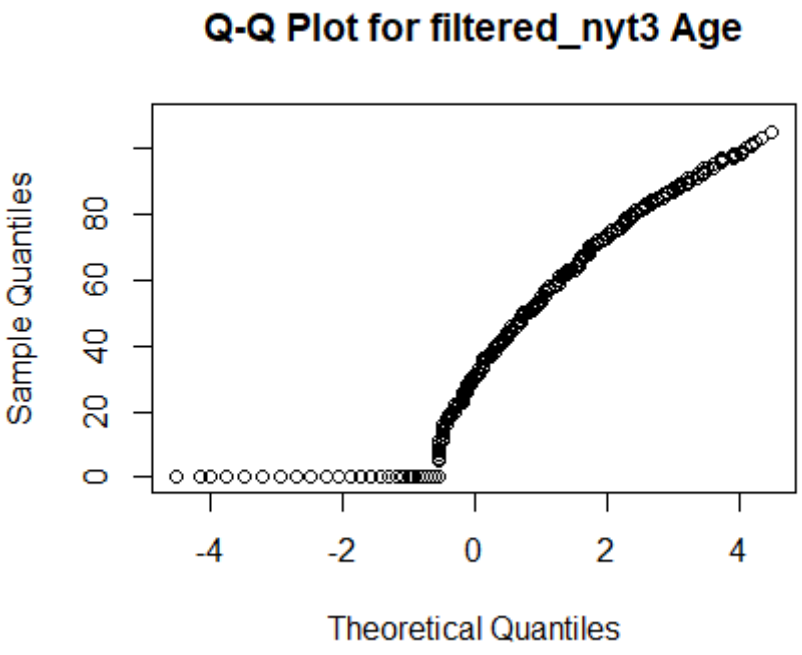

ECDF for filtered_nyt5 Age
y
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Age

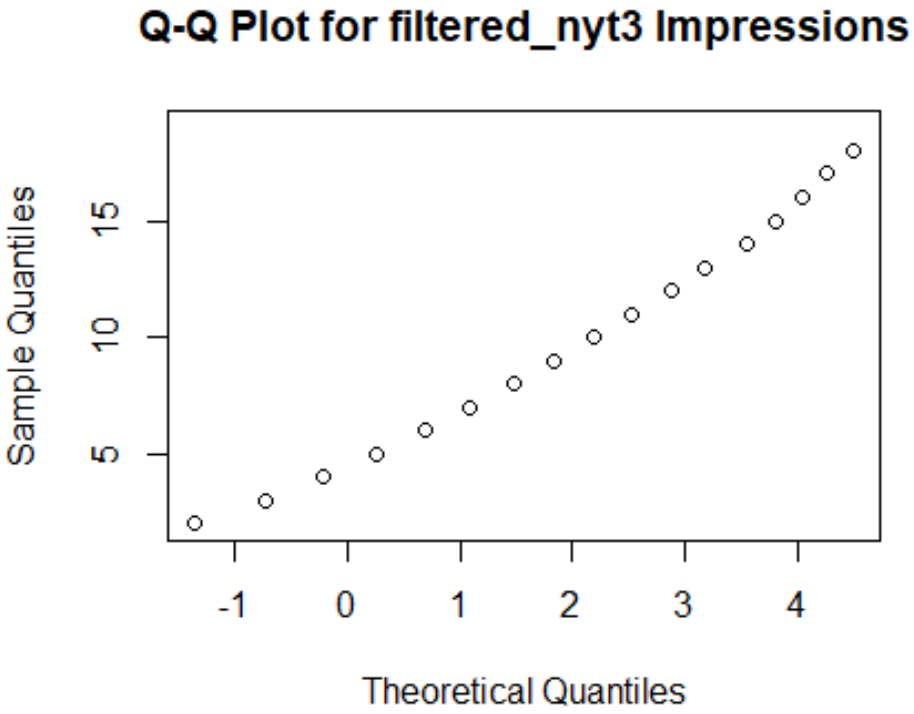

ECDF for filtered_nyt5 Impressions
y
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Impressions

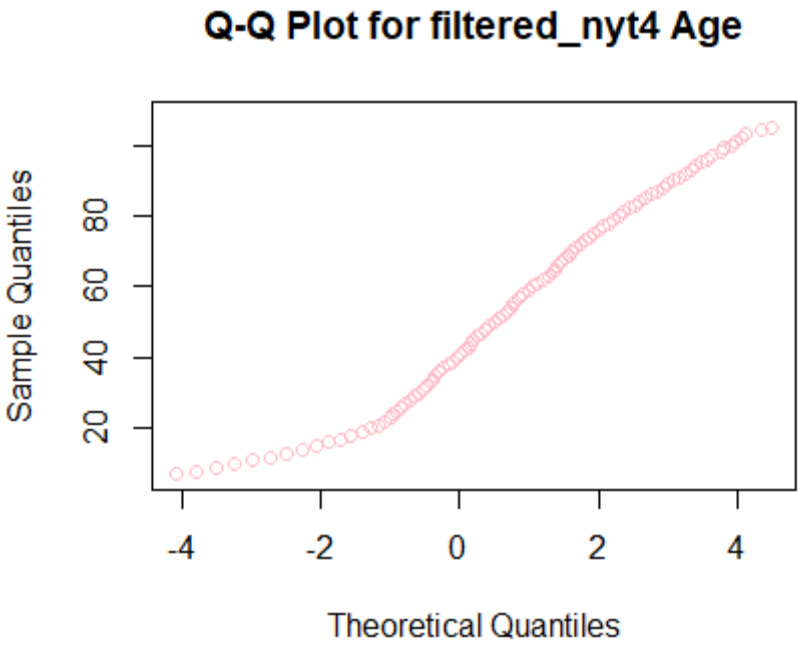

ECDF for filtered_nyt6 Age
y
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Age

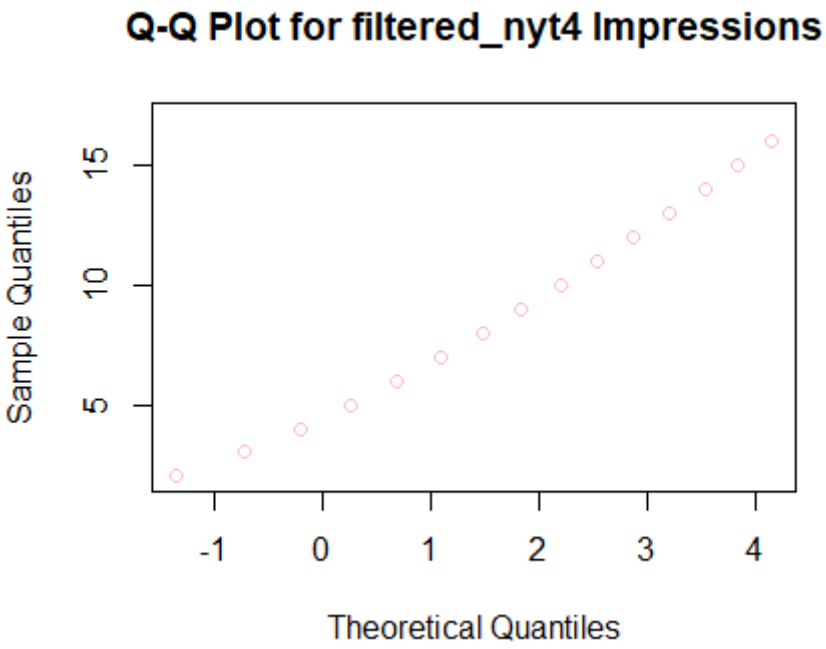

ECDF for filtered_nyt6 Impressions
y
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Impressions

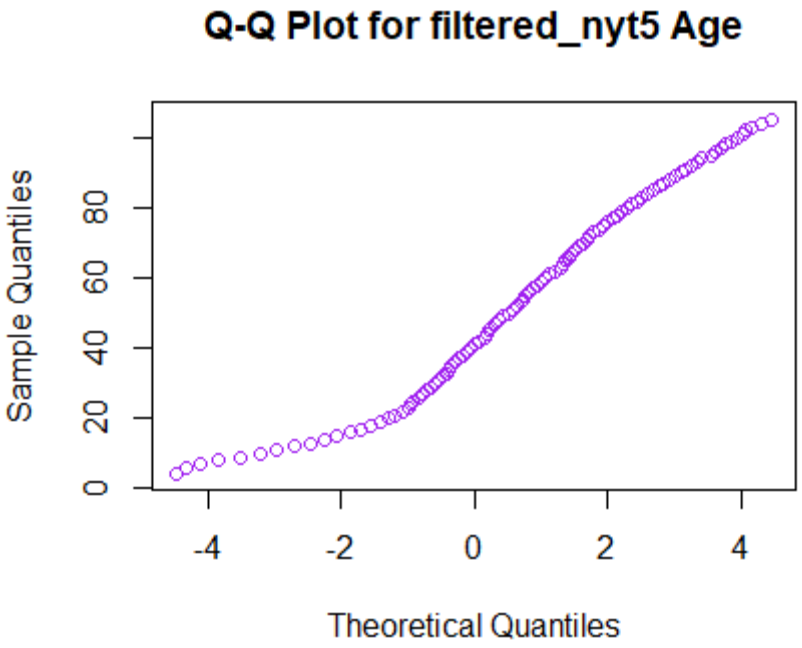
Q-Q of filtered nyt3, nyt4, nyt5, nyt6

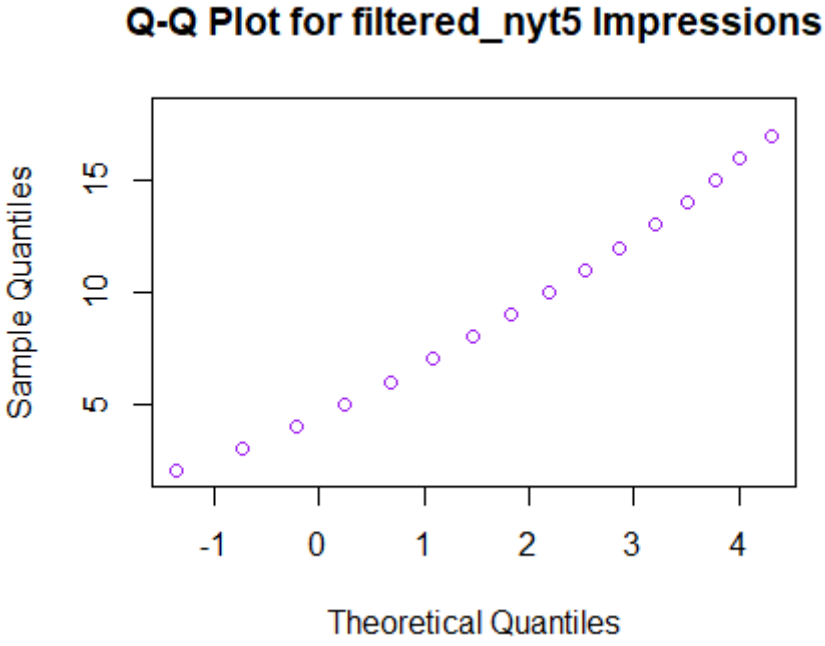


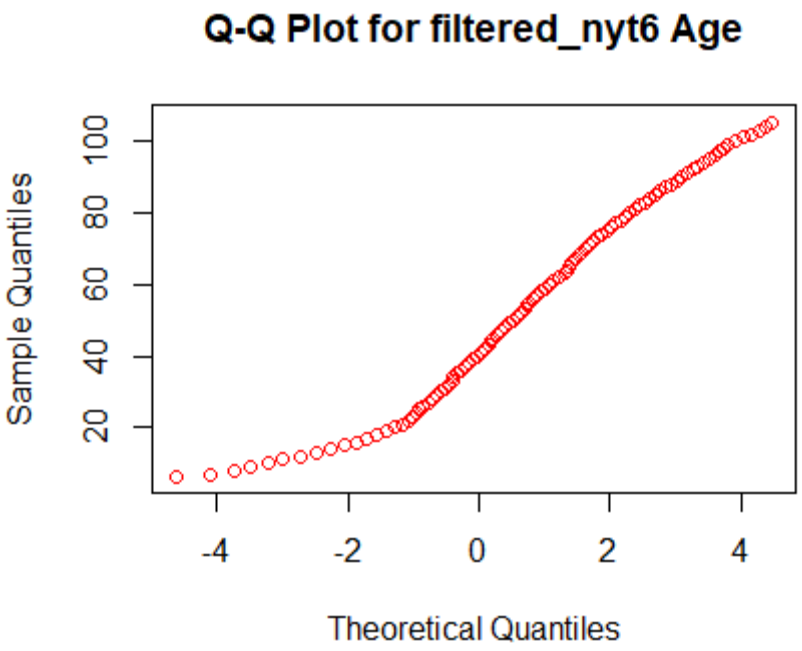


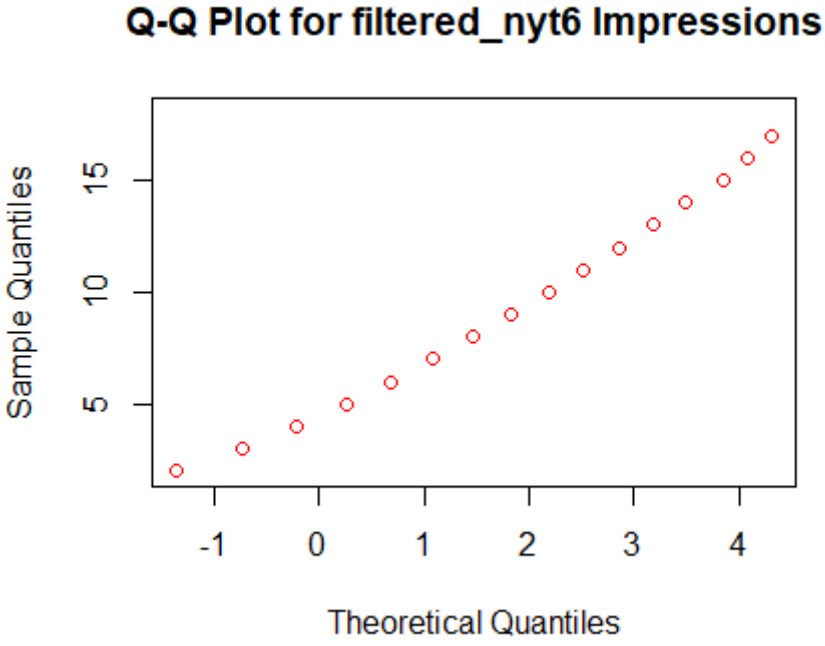

Q-Q Plot for filtered_nyt4 Age
Sample Quantiles
20
40
60
80
-4
-2
0
2
4
Theoretical Quantiles


Q-Q Plot for filtered_nyt4 Impressions
Sample Quantiles
5
10
15
-1
0
1
2
3
4
Theoretical Quantiles


Q-Q Plot for filtered_nyt5 Age
Sample Quantiles
0
20
40
60
80
-4
-2
0
2
4
Theoretical Quantiles


Q-Q Plot for filtered_nyt5 Impressions
Sample Quantiles
5
10
15
-1
0
1
2
3
4
Theoretical Quantiles


Q-Q Plot for filtered_nyt6 Age
Sample Quantiles
20
40
60
80
100
-4
-2
0
2
4
Theoretical Quantiles


Q-Q Plot for filtered_nyt6 Impressions
Sample Quantiles
5
10
15
-1
0
1
2
3
4
Theoretical Quantiles

In summary, the age distributions within all four datasets (nyt3, nyt4, nyt5, and nyt6) exhibit right-skew, with a central tendency indicated by a mean age of 42-43 years. Similarly, the impressions distributions in these datasets display a right-skewed pattern, with a median value hovering around 5 across the board.

Also, observed from the histograms, most of the individuals around the age of forty had greater number of impressions compared to individuals of other age groups

Furthermore, the results of Anderson-Darling normality tests yield low p-values (less than 0.05) for both age and impressions, providing evidence that these datasets do not adhere to a normal distribution. Consequently, I confidently reject the null hypothesis.