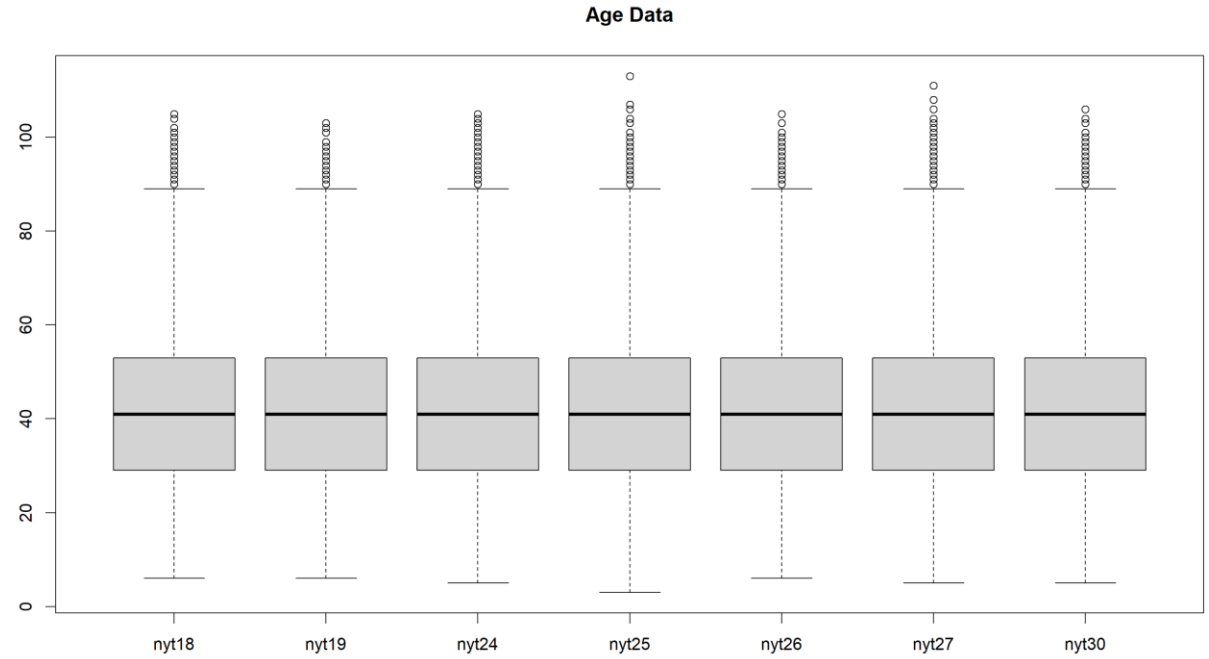
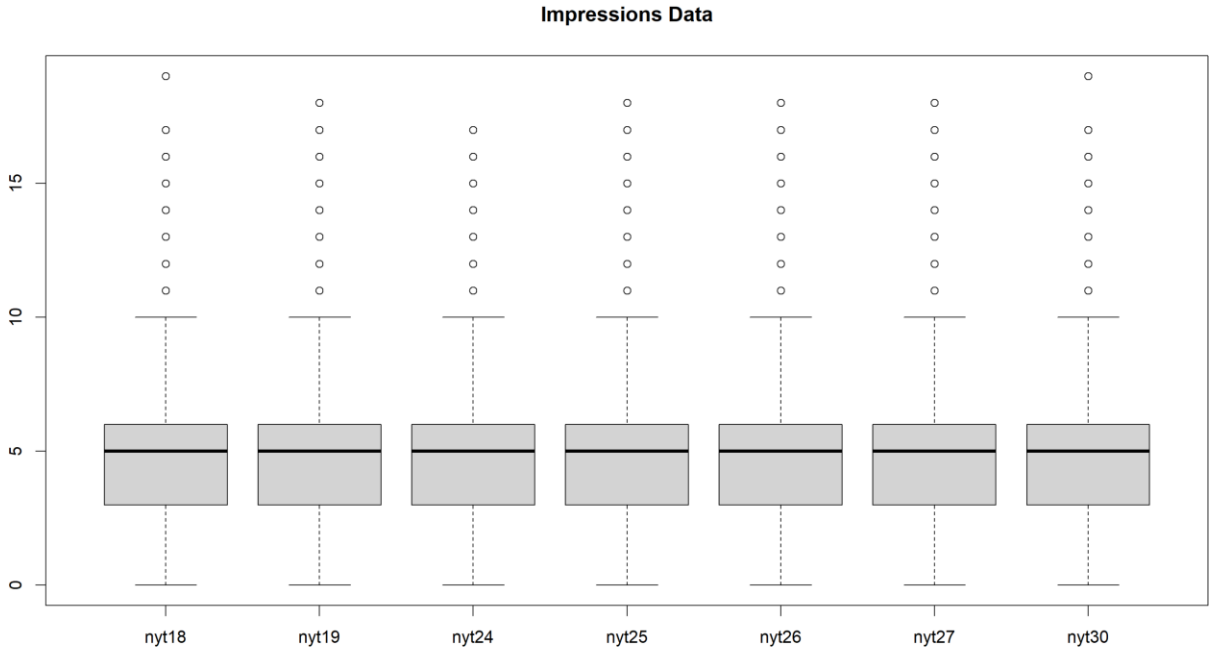
Part a

It seems like a lot of the data is missing, and it's inconsistent on what's missing.

I chose age and impressions for the variables. I'm going to remove the rows

in which the age is 0, since that's not possible.



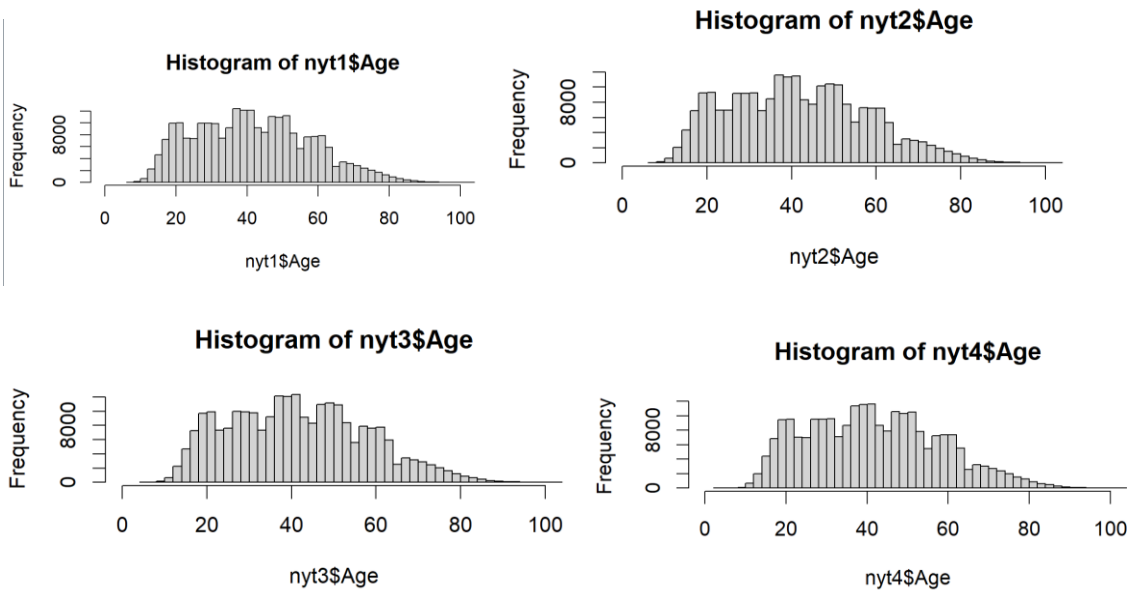



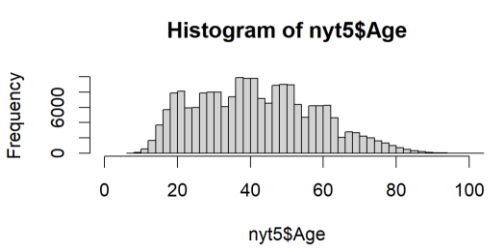
Discussion:

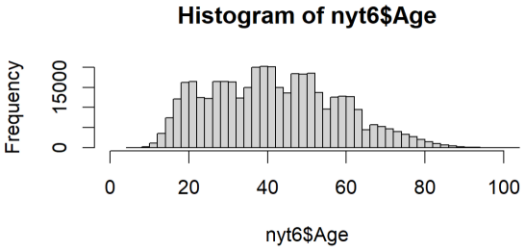
In both cases, it seems like the data has the exact same distribution across all nyt datasets. I actually thought that this might've been because I removed the rows with 0s, but I did it without removing those rows and they still all had the same distribution. Honestly it seems like all these nyt datasets came from one main dataset and each individual one was randomly sampled from the main one without replacement, they're so perfectly similar.

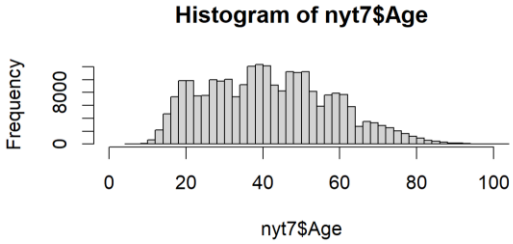
Part b

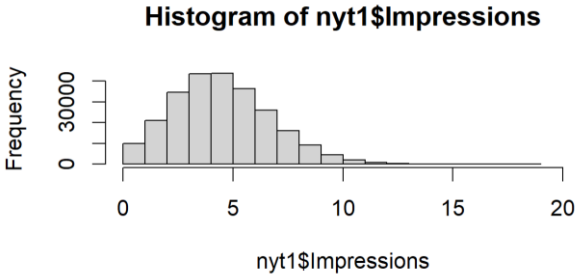
After experimenting, using 40 breaks seems to show a pattern without being too hard to see. I also forced the x and y axis so show all numbers.

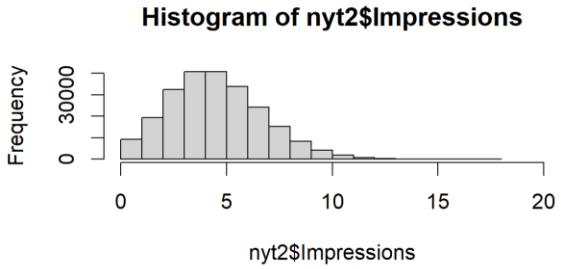


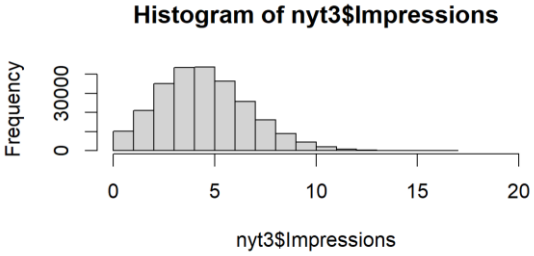

Histogram of nyt5$Age
Frequency
0
6000
0
20
40
60
80
100
nyt5$Age

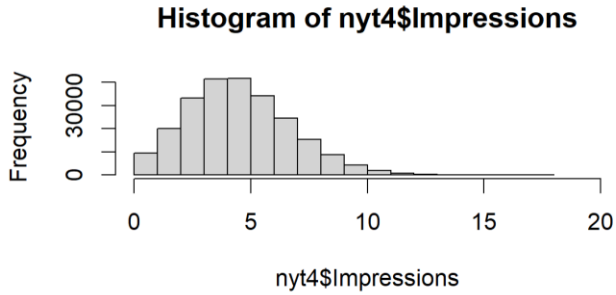

Histogram of nyt6$Age
Frequency
0
15000
0
20
40
60
80
100
nyt6$Age

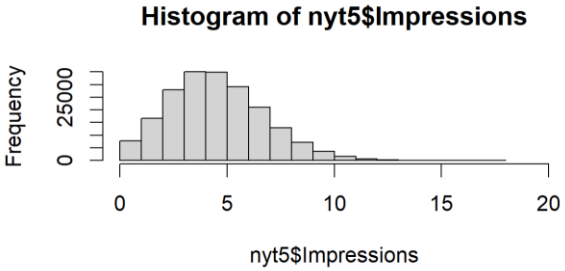

Histogram of nyt7$Age
Frequency
0
8000
0
20
40
60
80
100
nyt7$Age

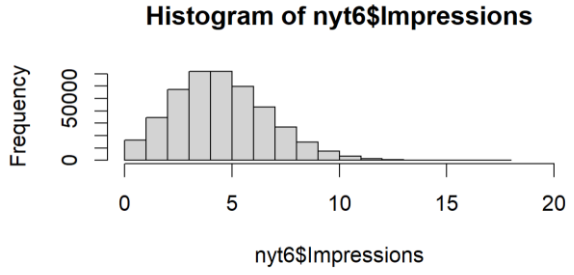

Histogram of nyt1$Impressions
Frequency
0
30000
0
5
10
15
20
nyt1$Impressions

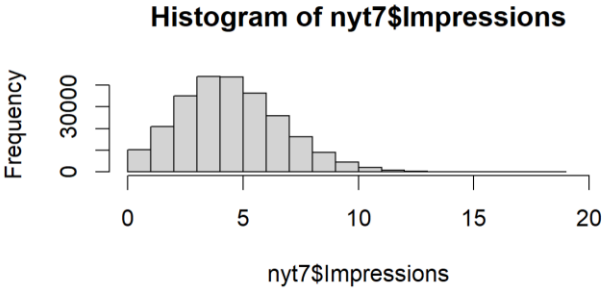

Histogram of nyt2$Impressions
Frequency
0
30000
0
5
10
15
20
nyt2$Impressions


Histogram of nyt3$Impressions
Frequency
0
30000
0
5
10
15
20
nyt3$Impressions


Histogram of nyt4$Impressions
Frequency
0
30000
0
5
10
15
20
nyt4$Impressions


Histogram of nyt5$Impressions
Frequency
0
25000
0
5
10
15
20
nyt5$Impressions


Histogram of nyt6$Impressions
Frequency
0
50000
0
5
10
15
20
nyt6$Impressions

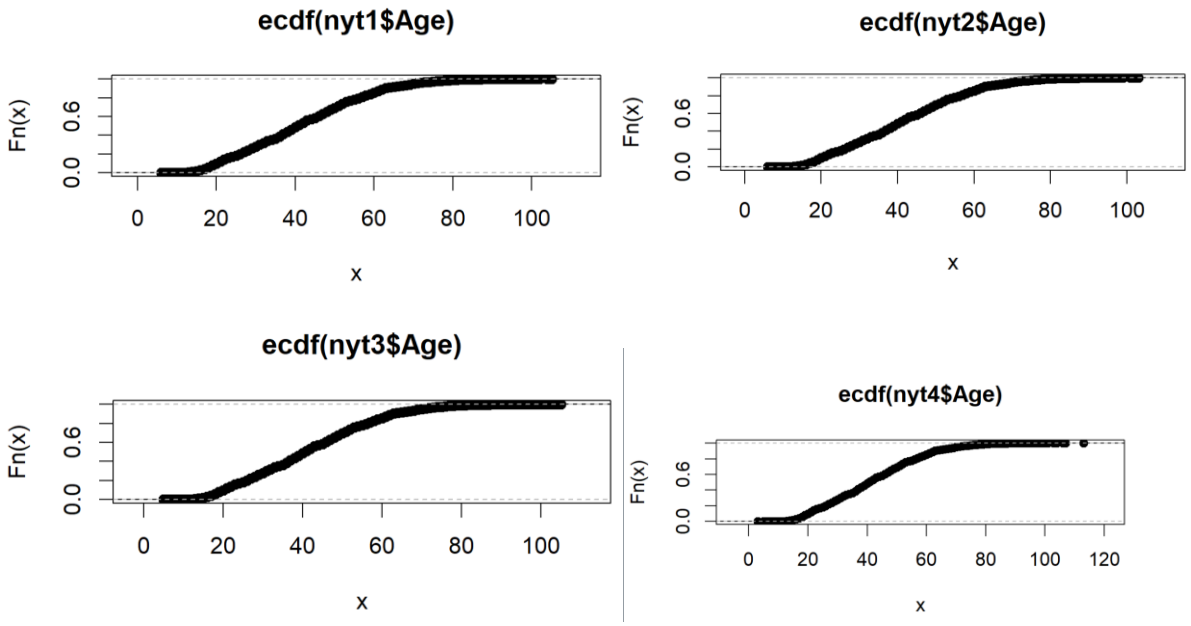

Histogram of nyt7$Impressions
Frequency
0
30000
0
5
10
15
20
nyt7$Impressions

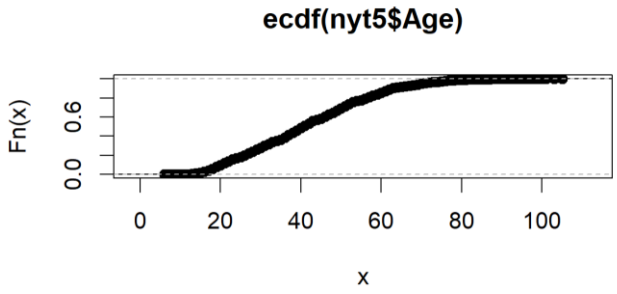
Discussion:

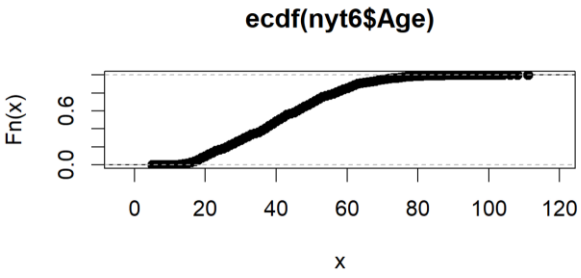
Regarding the age, yet again we see the exact same pattern emerge. All the datasets have essentially the exact same distribution. However, it's worth noting that for some reason nyt27 (my 6) has a little less than double the amount of data of the other data sets. This is true even without removing the 0 rows, so I don't know why it's so much bigger. The distribution skews right slightly and has a spiked pattern. Skewing right makes sense as the older you are the you'll probably use nyt less, but I can't think of a reason that ages 25ish, 35ish, and 45ish would use it less than other neighbors.

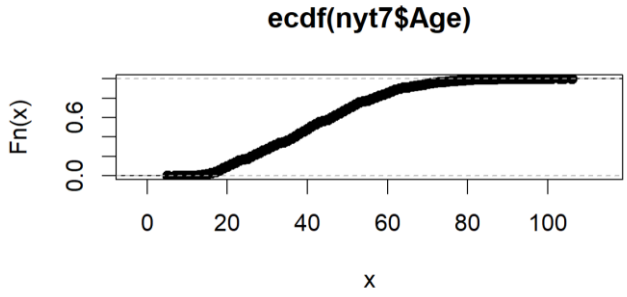
Regarding the Impressions, they all yet again have essentially the same distribution. Ignoring the total num row differences, it seems to be a relatively gaussian curve, however a little skewed to the right.

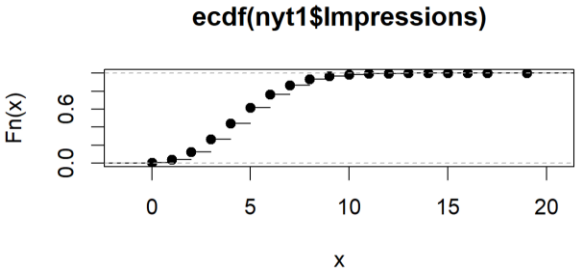
Part c

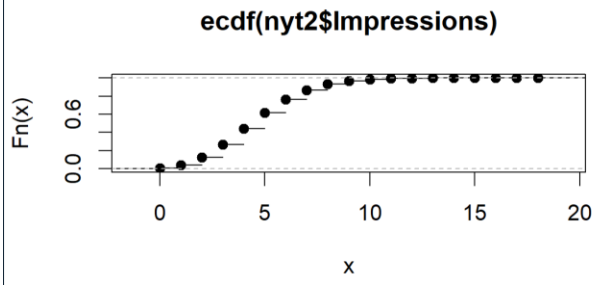


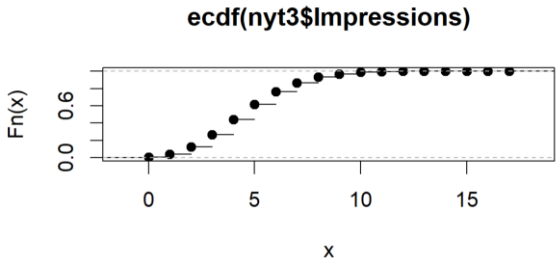

ecdf(nyt5$Age)
Fn(x)
0.0
0.6
0
20
40
60
80
100
x

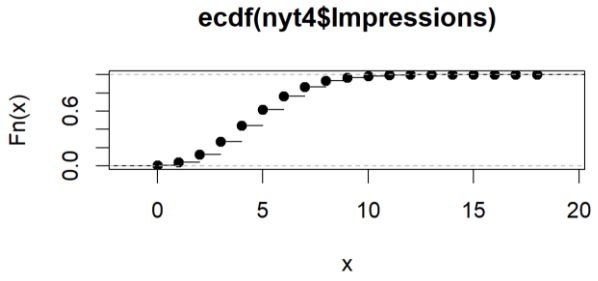

ecdf(nyt6$Age)
Fn(x)
0.0
0.6
0
20
40
60
80
100
120
x

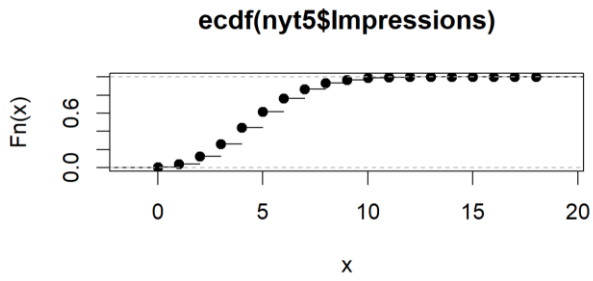

ecdf(nyt7$Age)
Fn(x)
0.0
0.6
0
20
40
60
80
100
x

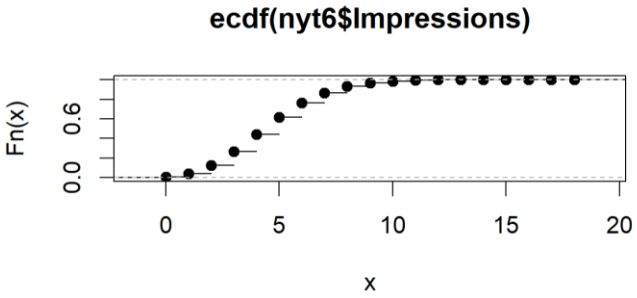

ecdf(nyt1$Impressions)
Fn(x)
0.0
0.6
0
5
10
15
20
x

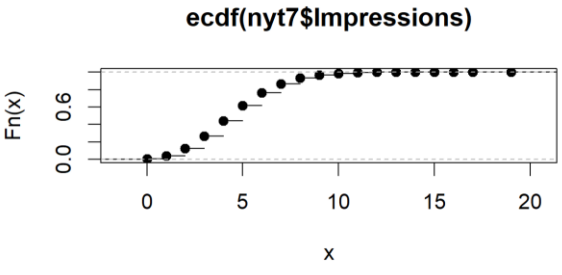

ecdf(nyt2$Impressions)
Fn(x)
0.0
0.6
0
5
10
15
20
x


ecdf(nyt3$Impressions)
Fn(x)
0.0
0.6
0
5
10
15
x

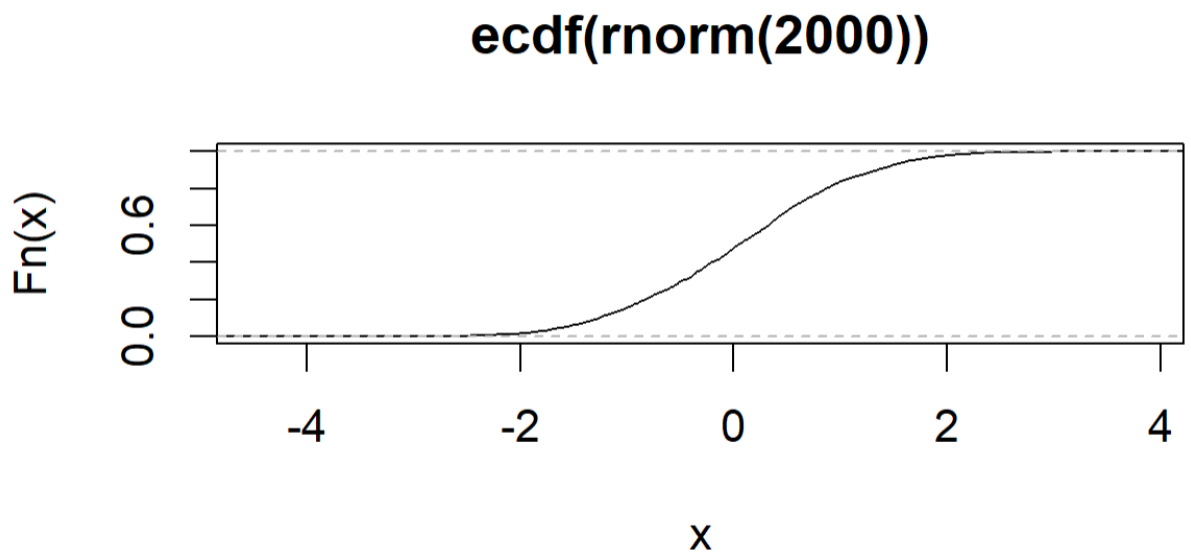

ecdf(nyt4$Impressions)
Fn(x)
0.0
0.6
0
5
10
15
20
x


ecdf(nyt5$Impressions)
Fn(x)
0.0
0.6
0
5
10
15
20
x


ecdf(nyt6$Impressions)
Fn(x)
0.0
0.6
0
5
10
15
20
x


ecdf(nyt7$Impressions)
Fn(x)
0.0
0.6
0
5
10
15
20
x

I also plotted the ecdf of a normal distribution for visual comparison:

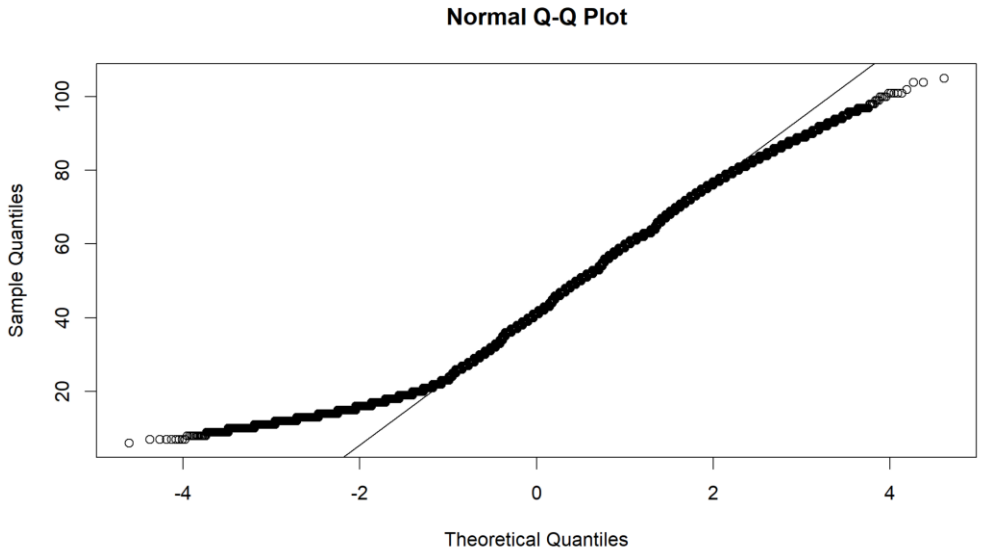



Get a normal distribution of each and qqplot against that

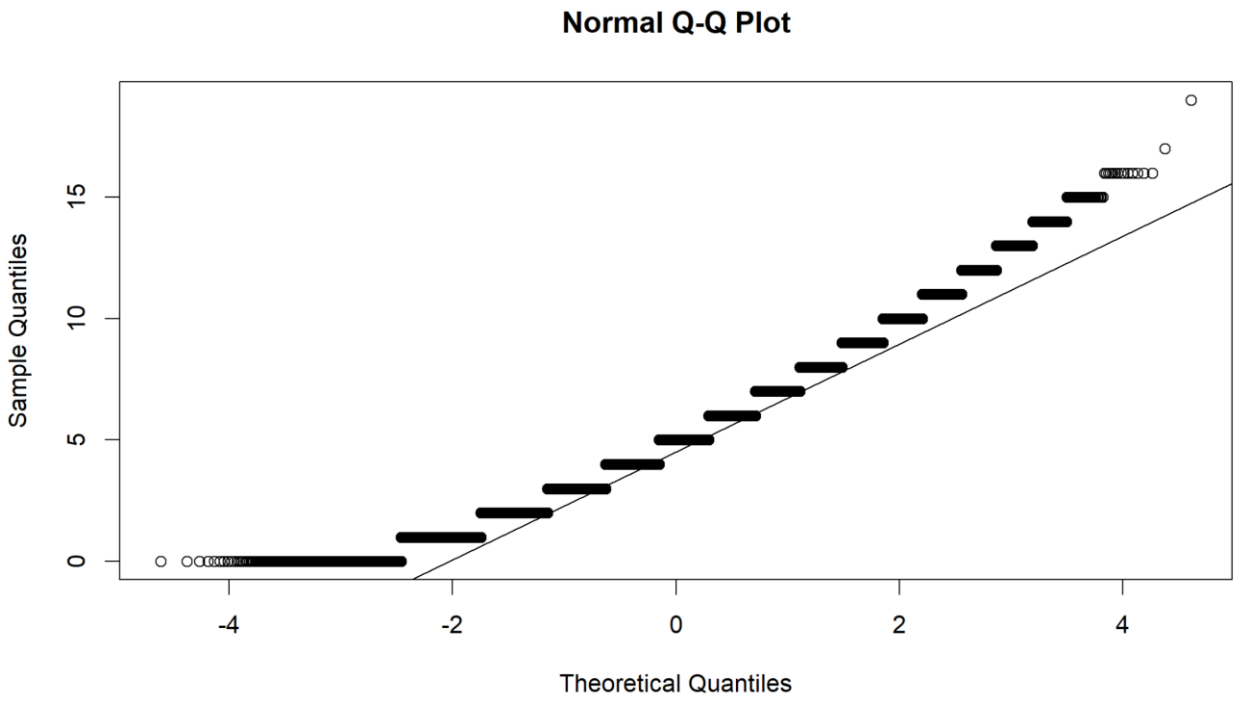
As this takes a while to compute and I wasn't asked to do this for each dataset

I'm just going to use one of the datasets.

For Age:



For Impressions:




Discussion:

Of course, when doing the ecdf all the graphs look the same. Moving past that, I've realized that this data is almost certainly not normal. From looking at the ecdf of the variables, it's close but is slightly more linear. This is made even more clear when we make the qqplot. The middle is almost perfectly normal, however once we leave around 2 quantiles in either direction it gets decisively less normal. It seems that in both variables the data gets more spread out. This reminds me now more of a Cauchy distribution.

Part d

I want to use a Shapiro-wilks test for this, but that only works on sample sizes up to 5000.

Start of discussion: While doing research for this, I found out that most normality testing is useless when it comes to very large sample sizes like this one since once you get large it's basically guaranteed that our data will deviate from the idealistic normal distribution.

To demonstrate this, I'm going to do the Shapiro-wilks test on increasingly larger samples from the data.

Sample size = 50

```
        Shapiro-Wilk normality test

data:  impressionsSmall
W = 0.93613, p-value = 0.0095
```

Sample size = 500

```
        Shapiro-Wilk normality test

data:  impressionsSmall
W = 0.96821, p-value = 6.119e-09
```

Sample size = 5000

```
        Shapiro-Wilk normality test

data:  impressionsSmall
W = 0.97303, p-value < 2.2e-16
```

Discussion Continued:

As we can see, the test gets increasingly confident that the data does not follow a normal distribution as the number of samples we pull from it goes up.

Therefore, I don't believe that we can really say for sure whether or not

the data follows a normal distribution. Or rather, we should say that since the data set is large it does not follow a normal distribution.

It is worth noting that it fails to have a normal distribution even on a small sample, but I still believe that using normally tests on data of this size isn't very effective. Visualizing with the qqplot is probably the best we'll get, and there I saw that it likely wasn't normal (at least to my eye), so that's my conclusion.

This seems to not be normally distributed and we fail to reject the null hypothesis.