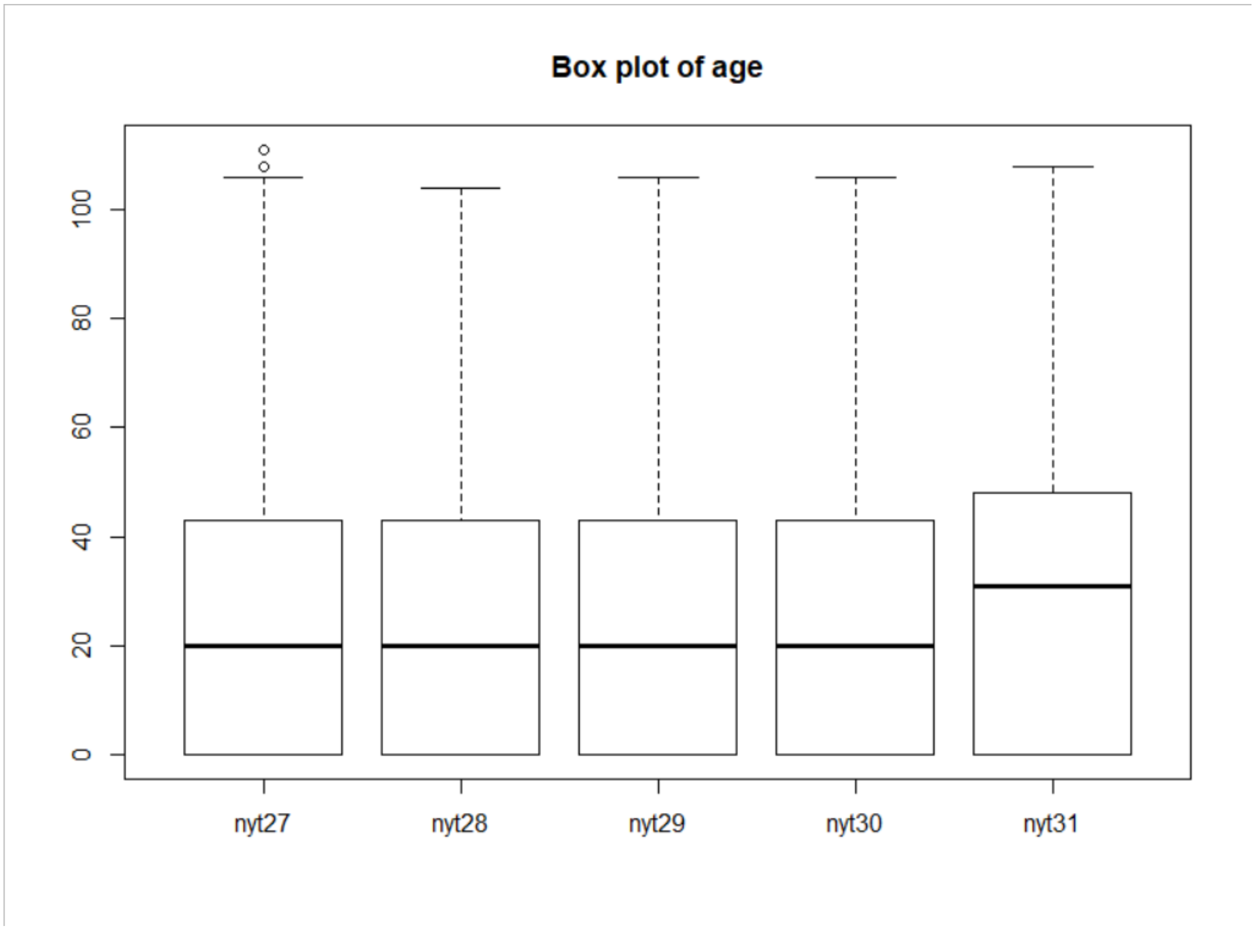
Donghuai Li
661991517

Assignment 3

Data Analytics

Question 1:

a. The distribution of Age among different nyt dataset is similar. These dataset have means near 20 and upper quantile near 40. The relative long span between upper quantile and max value implies datasets are skewed toward large number. The down quantile is overlaped by min value which implies there are many 0 in the data set.

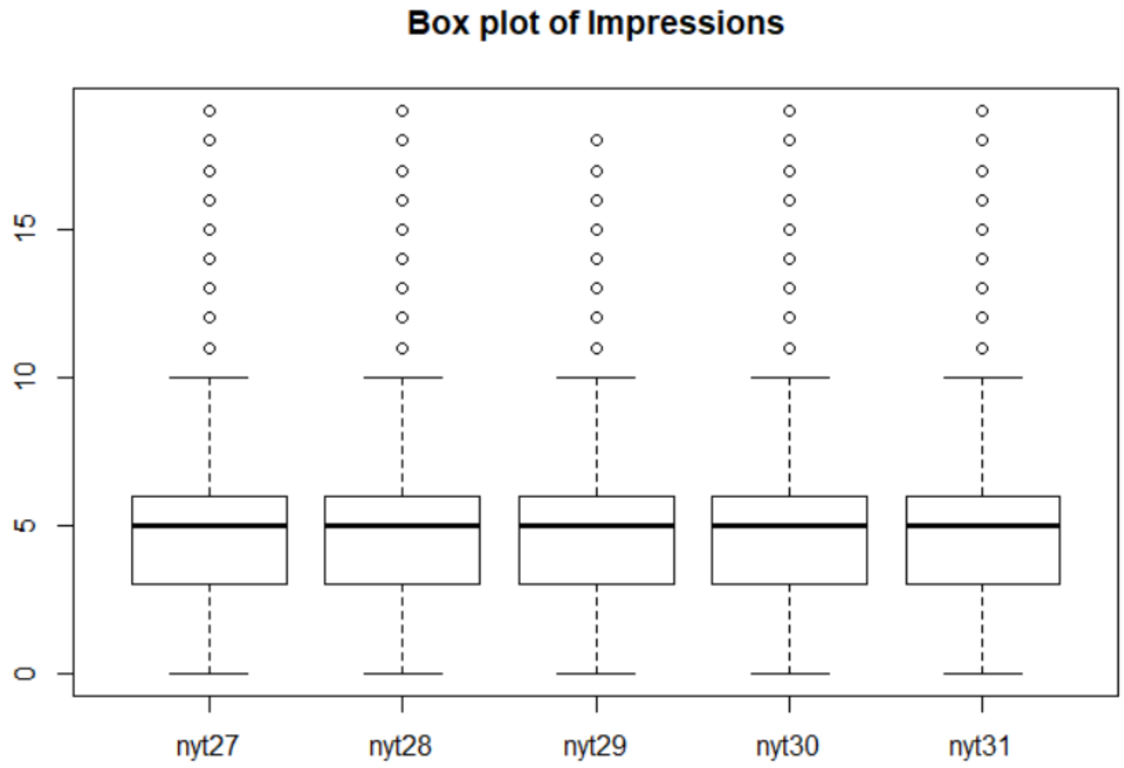


Donghuai Li
661991517

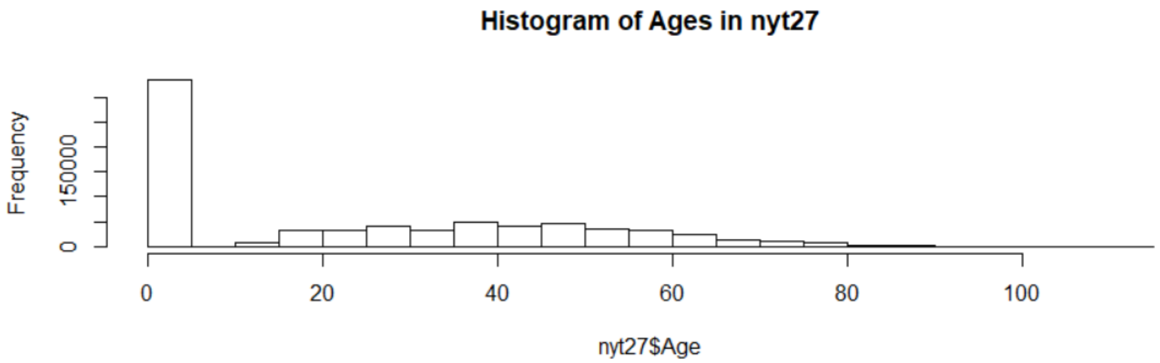
Assignment 3

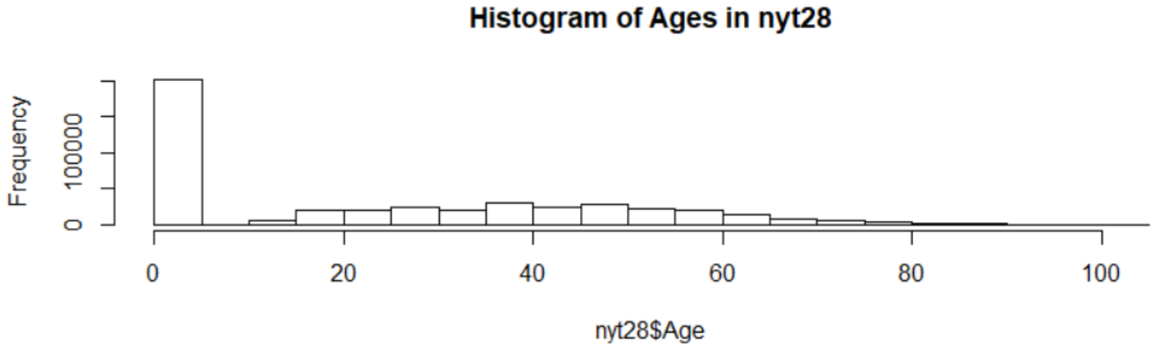
Data Analytics

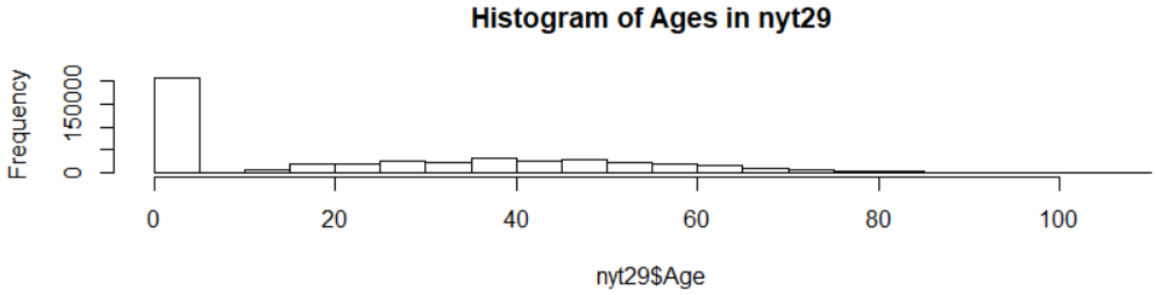
Boxplot of impressions shows average impressions is about 5. Lower quantile is farther from the average than upper quantile tells us more data is below 5. In each dataset there are many outliers.

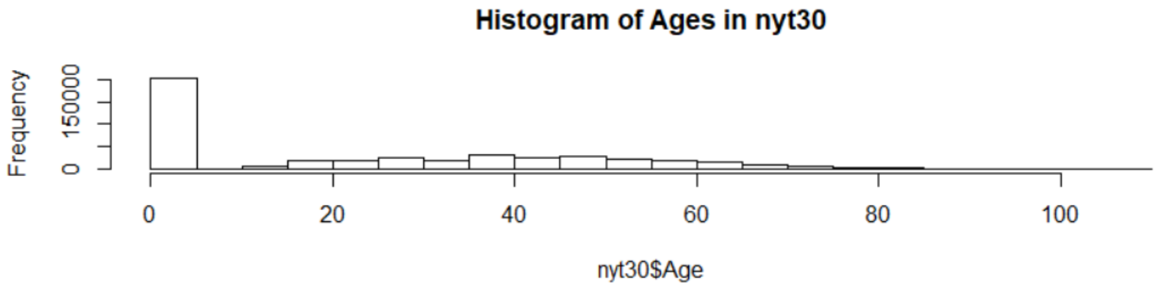


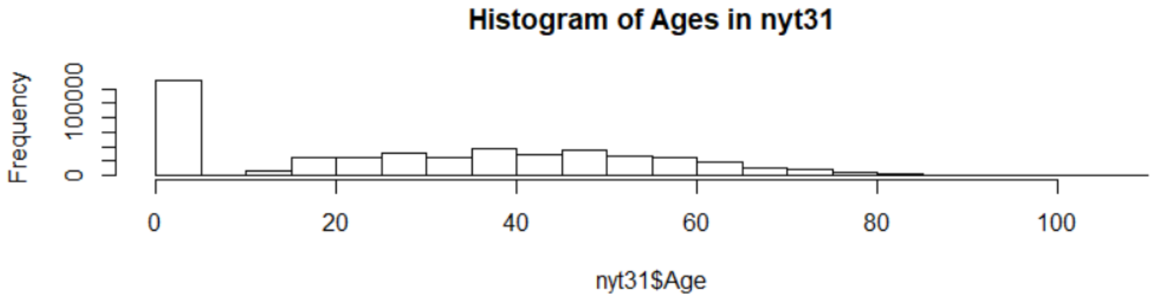
b. Every dataset have a huge proportion of 0 in age. The 0 is too many that the distribution of other age is hardly distinguish. We can only see a relative normal distribution on age from 10 to 90. The nyt31 have least 0 so its age distributiom is more conspicious than other data set



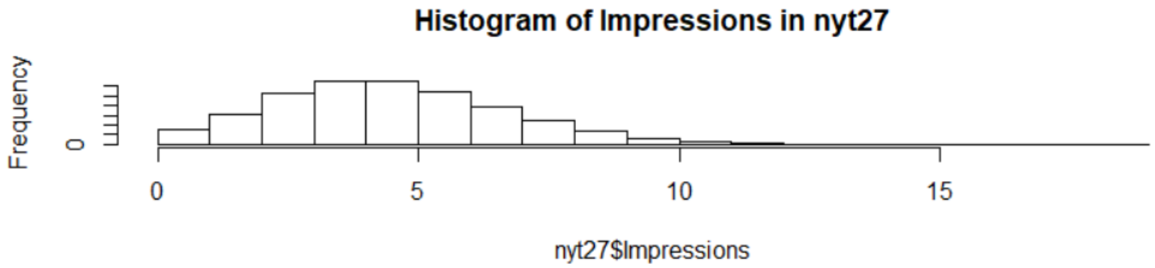


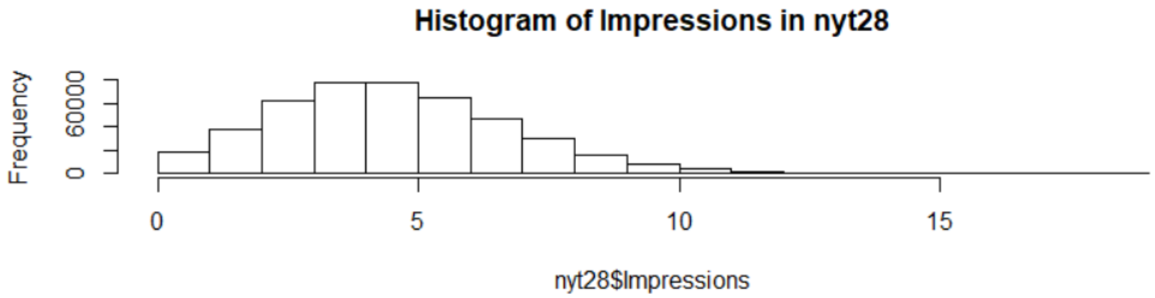


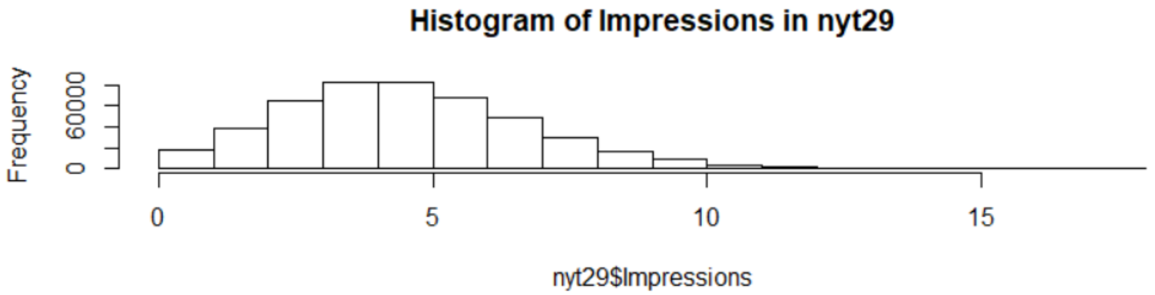

Histogram of Ages in nyt30
Frequency
0
150000
0
20
40
60
80
100
nyt30$Age

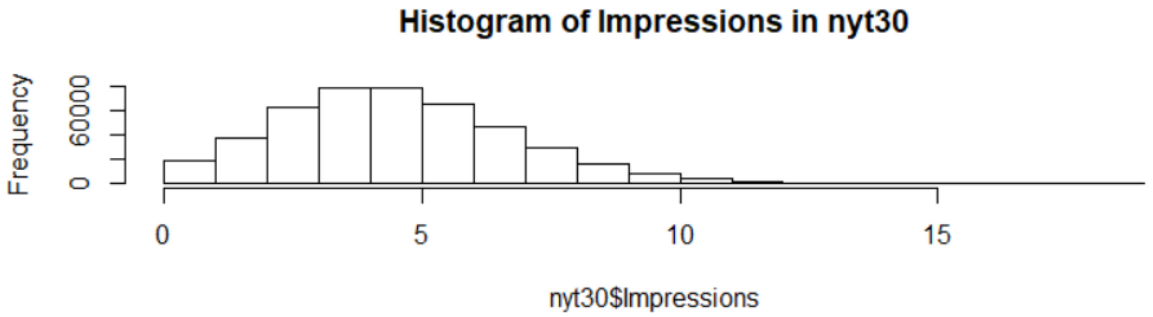

Histogram of Ages in nyt31
Frequency
0
100000
0
20
40
60
80
100
nyt31$Age

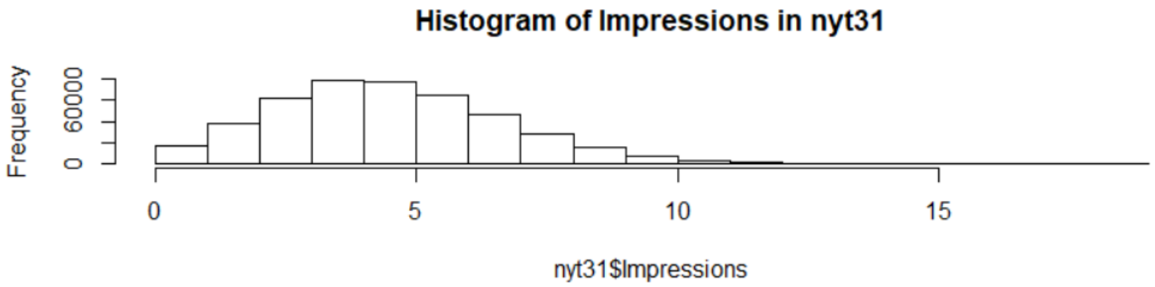
The histograms for impression are right-skewed distribution, which means there are some outlier impression have value over 10.





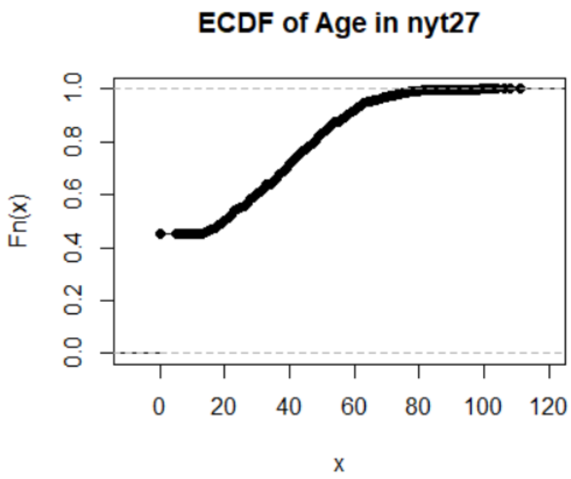


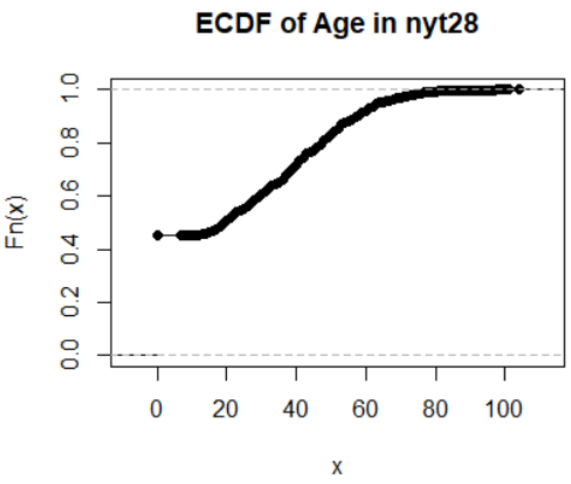

Histogram of Impressions in nyt30
Frequency
0
60000
0
5
10
15
nyt30$Impressions

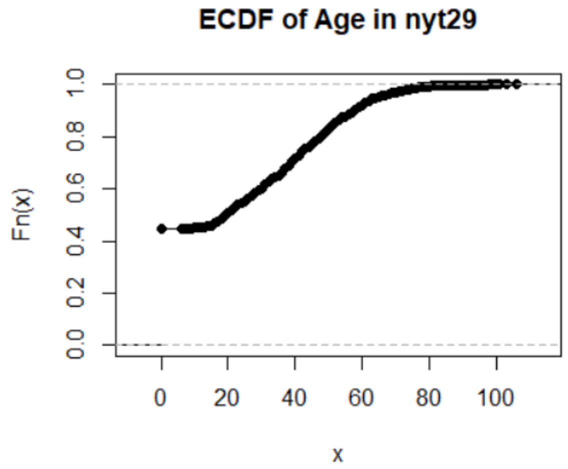

Histogram of Impressions in nyt31
Frequency
0
60000
0
5
10
15
nyt31$Impressions

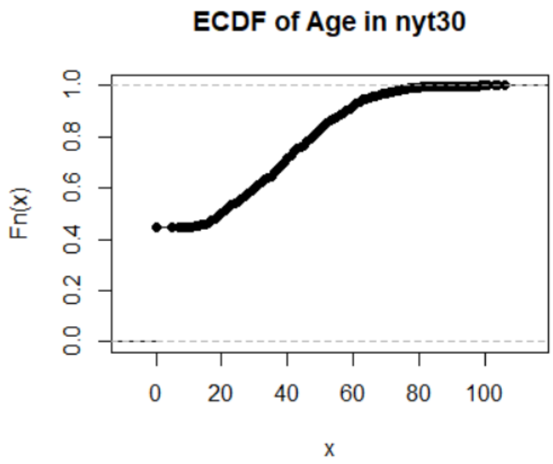
C. The quantile-quantile plots are test the normal distribution. The qqplot for age has a plain part before 0 in norm quantiles, the reason is we have found before: Large amount of 0 in data set.

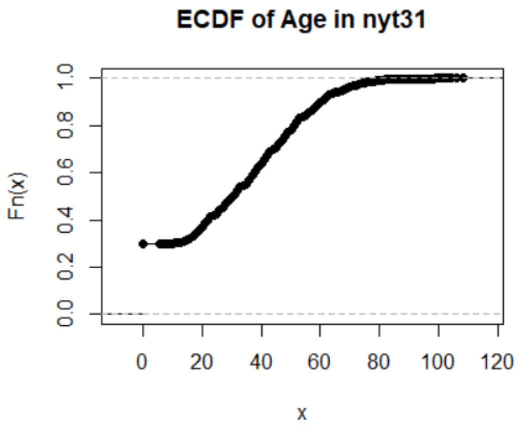
The qqplots of impression start their match with a straight line from -2 in norm quantiles which equals a similarity of normal distribution, though they are concave up when compared to a restrict straight line.

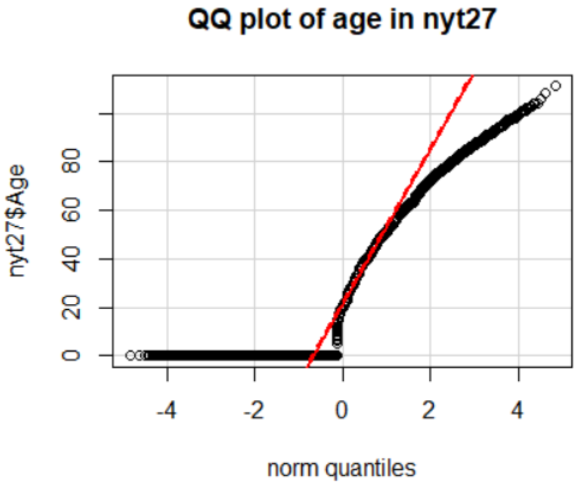


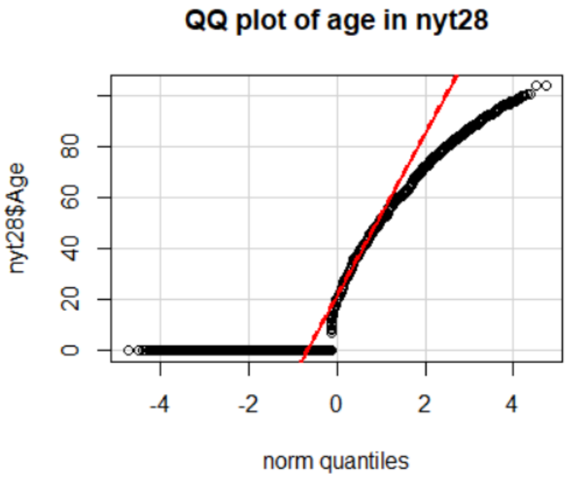


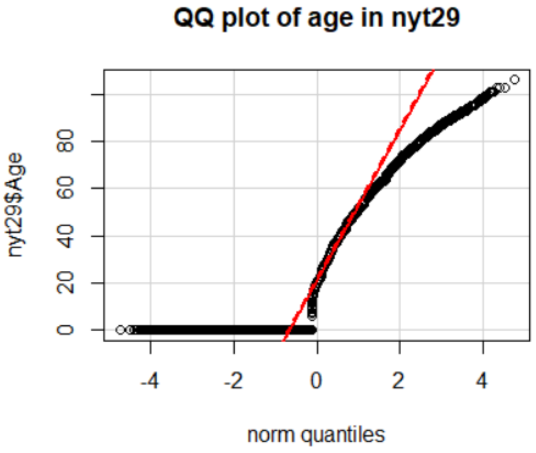


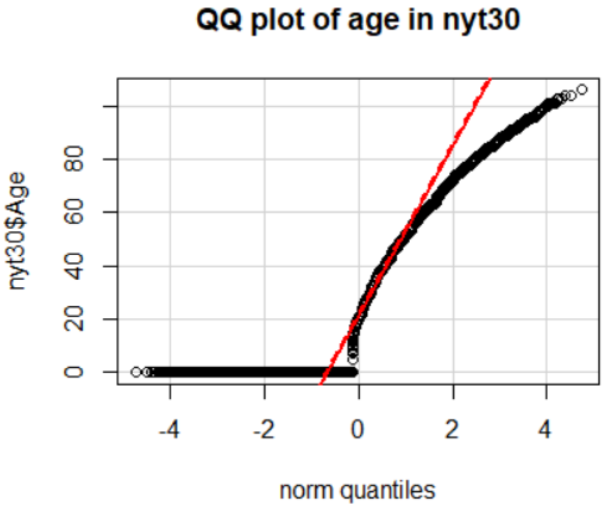


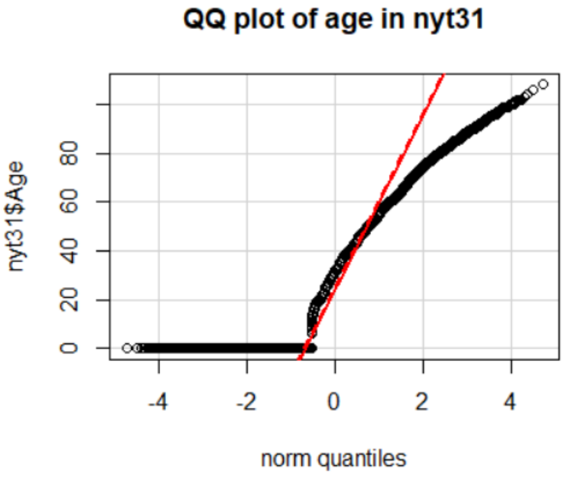



QQ plot of age in nyt27
nyt27$Age
0
20
40
60
80
-4
-2
0
2
4
norm quantiles


QQ plot of age in nyt28
nyt28$Age
0
20
40
60
80
-4
-2
0
2
4
norm quantiles


QQ plot of age in nyt29
nyt29$Age
0
20
40
60
80
-4
-2
0
2
4
norm quantiles


QQ plot of age in nyt30
nyt30$Age
0
20
40
60
80
-4
-2
0
2
4
norm quantiles

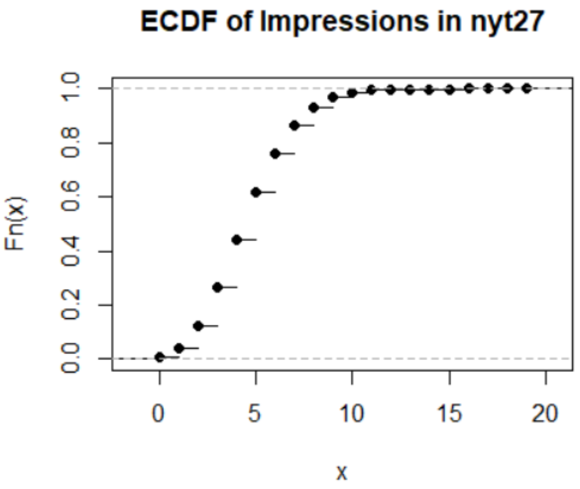

QQ plot of age in nyt31
nyt31$Age
0
20
40
60
80
-4
-2
0
2
4
norm quantiles

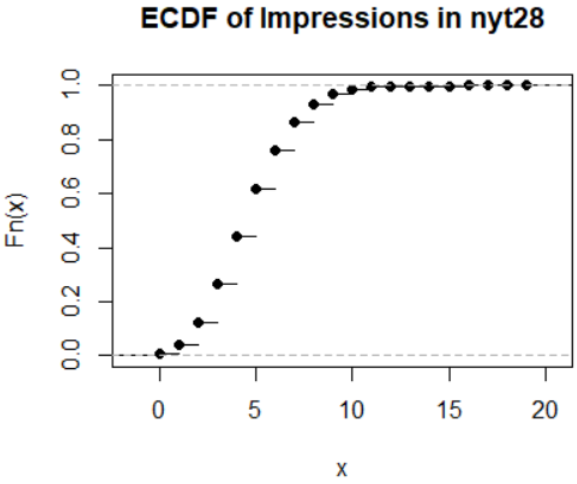
Donghuai Li
661991517

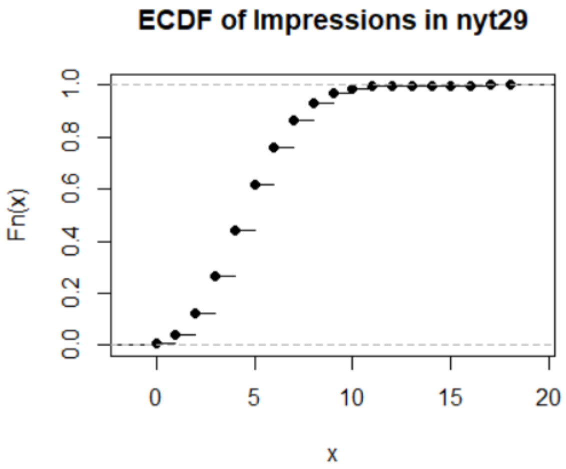
Assignment 3

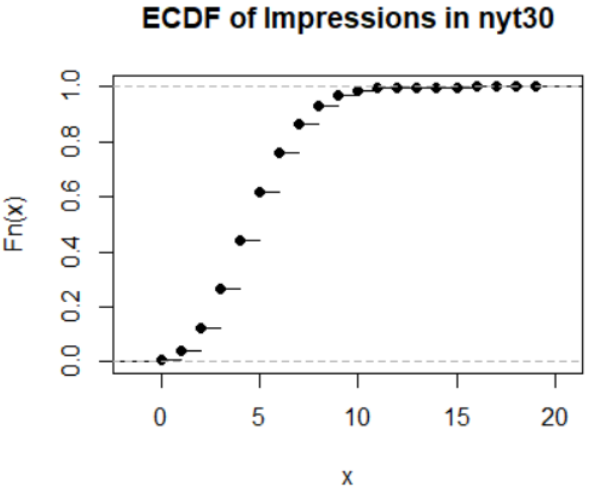
Data Analytics

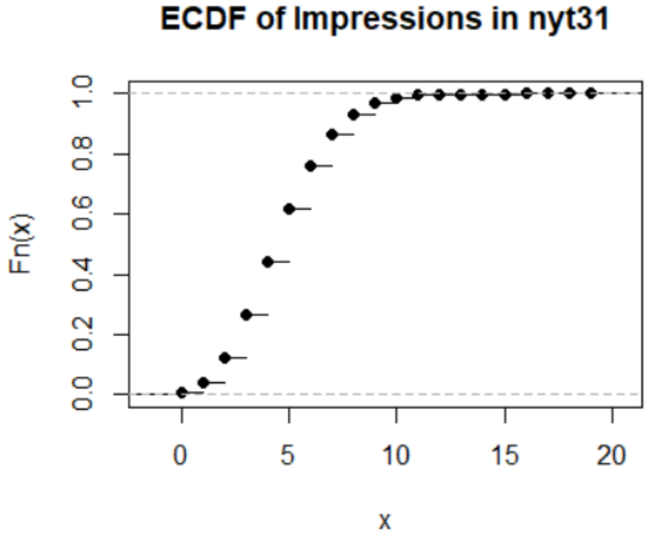
ECDF of impressions



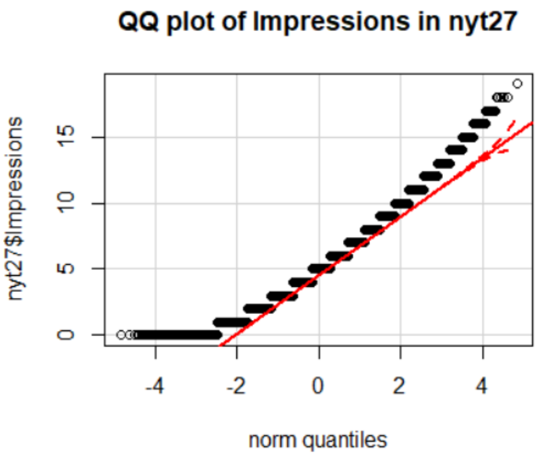


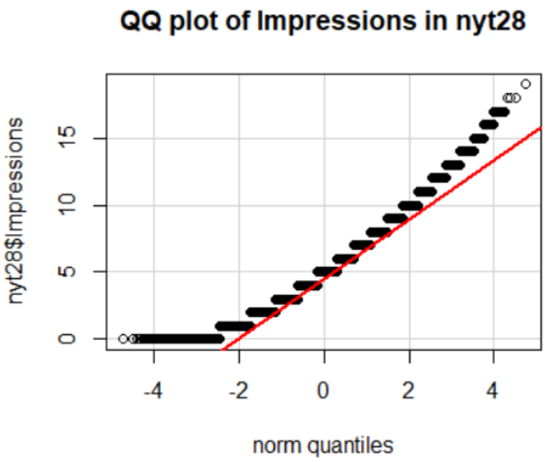


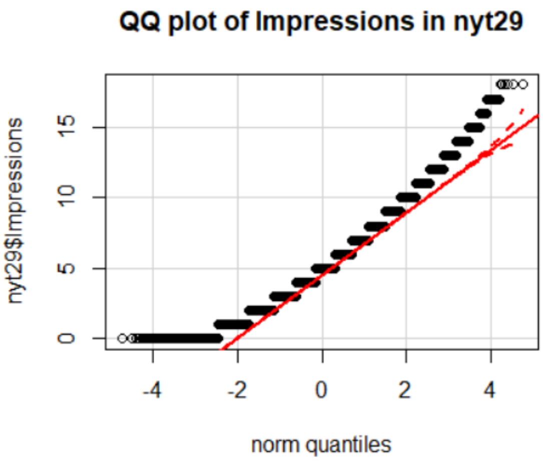


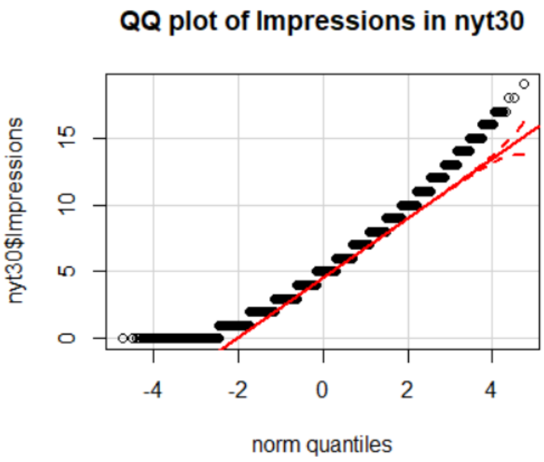


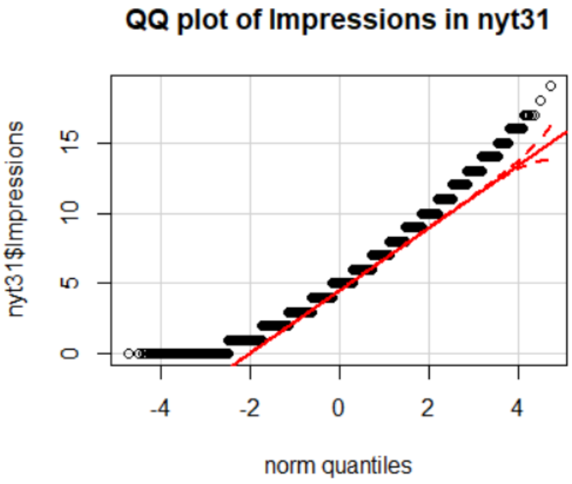
qqplot of impressions







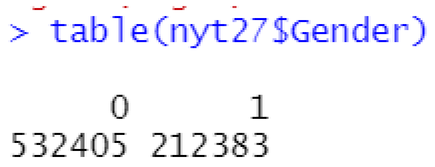




D. I apply the Shapiro-Wilk test for datasets. The Shapiro-Wilk test only accept 3 to 5000 inputs, so I check the sample of variable in 5000, 500 and 30 of data points. Given random 30 sample, all the p-value far less than 0.05. So we can reject null hypothesis and accept these data sets are normal distribution.

Dataset	W	p-value
NYT27	.747	8.1e-06
NYT28	.834	2.8e-04
NYT29	.867	1.4e-03
NYT30	.828	2.2e-04
NYT31	.903	9.9e-03

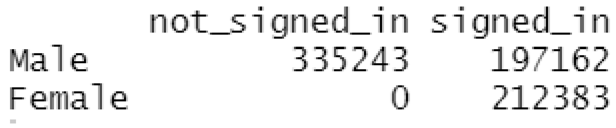
E. I found a intresting fact. When I count the gender data, the 0 is out number the 1. For example, in nyt27, we have 532k 0, but we only have 212k 1.



```
> table(nyt27$Gender)

     0      1 
532405 212383 
```

Then I made a table of gender and signed_in. The intresting part is: the combination of not signed in and female is zero. I assume 0 represent male and 1 represent female here. So I guess the record have male as default when viewer don't sign in.

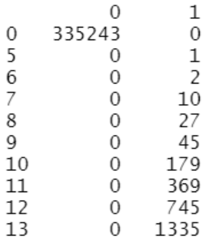


```
       not_signed_in signed_in
Male          335243    197162
Female             0    212383
```

This result also shown in other dataset.

Then I continue explore whether other variables have “default” value.

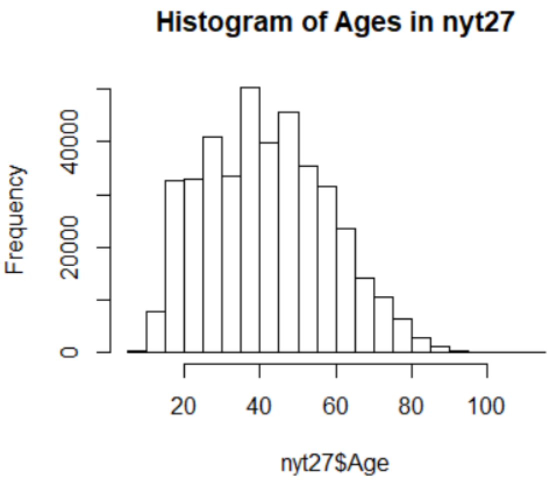
I combined the age and signed_in. Result shows all the age 0 data have 0 in signed_in column. So age 0 is another “default” value.

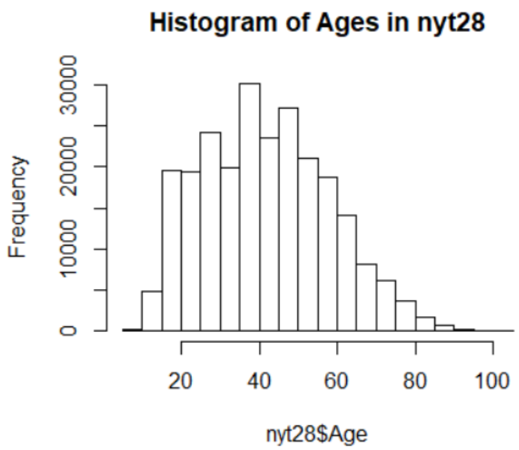


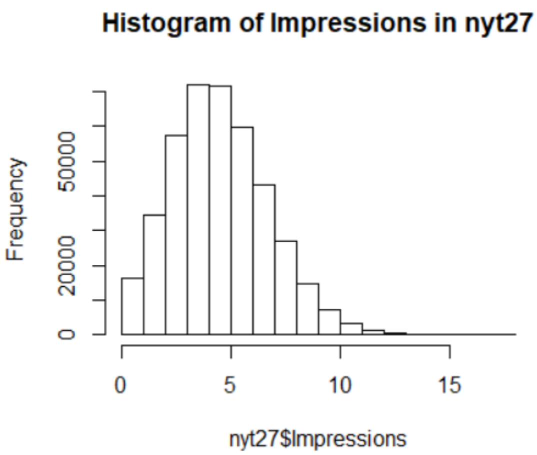
```
          0      1
0    335243      0
5         0      1
6         0      2
7         0     10
8         0     27
9         0     45
10        0    179
11        0    369
12        0    745
13        0   1335
```

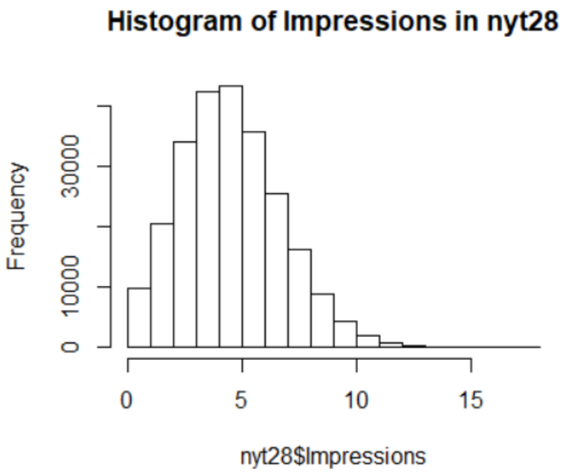
The impressions and clicks have similar proportion of data points in each gender.

Question2: As the pattern I found before, I only choose signed_in equals to 1 this time to plot histogram and qq plot.

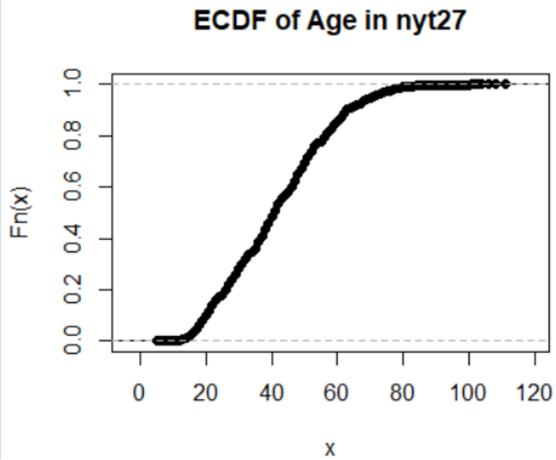


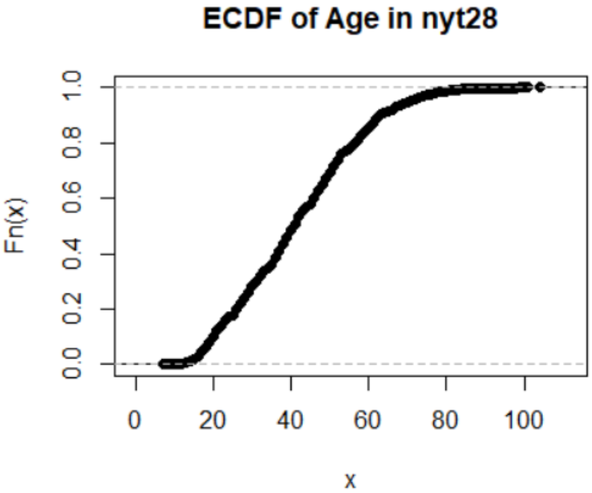


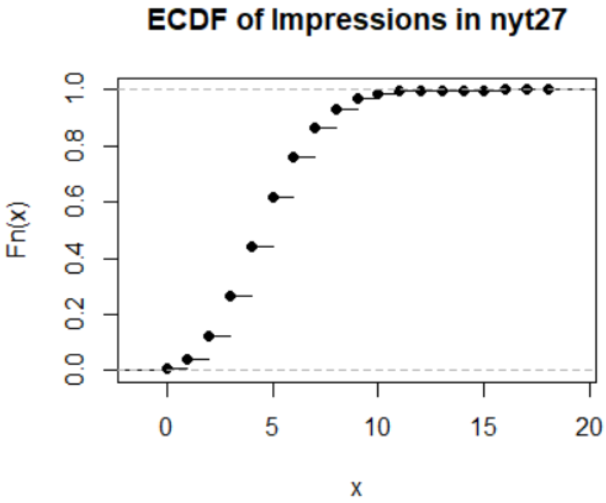


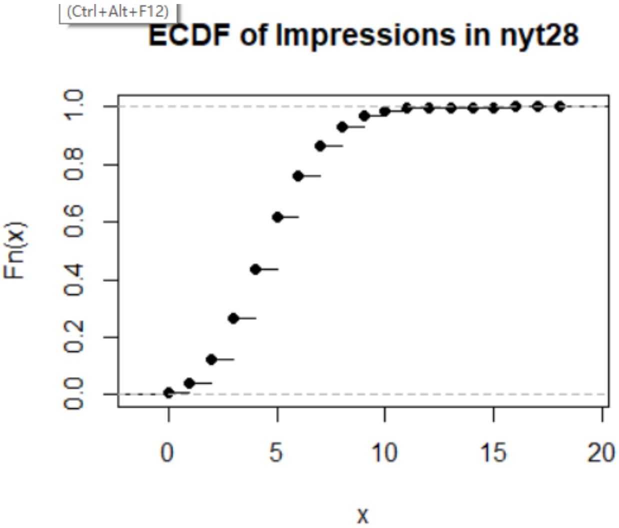


Histograms plot the right skew clearly than before.









The ECDF graphs of impressions and age give us a steady increase repectively.

Dataset	W	p-value
NYT27	.943	0.115
NYT28	.960	0.3278

The Shapiro tests have p-vale all above 0.05, along with histogram before, we believe the age distribution is not normal distributed given confidence level of 5%.