Kevin Pan

Data Analytics (Level 6000)

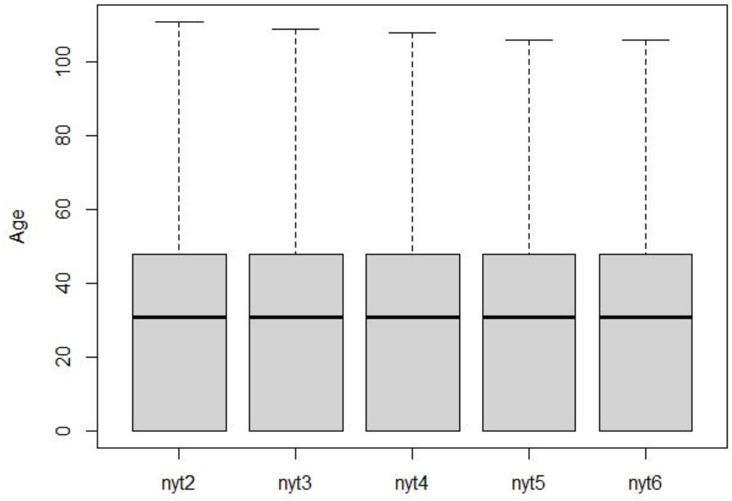
Assignment3

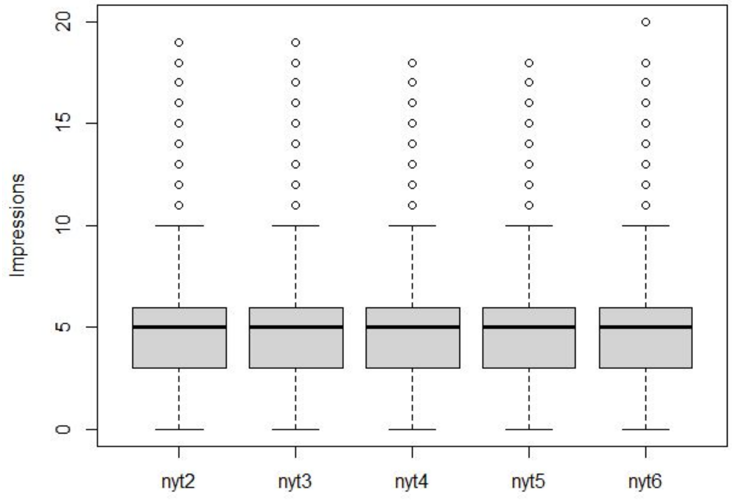
Due: March. 11, 2021

1. Choose any 5 of the nyt datasets except nyt1, perform the following:

a). Create boxplots for all 5 datasets for each of two key variables (you choose the variables), i.e. two figures (one for each variable) with 5 boxplots (for the 5 different datasets) in each. Describe/summarize the distributions. min. 3-4 sentences (3%)

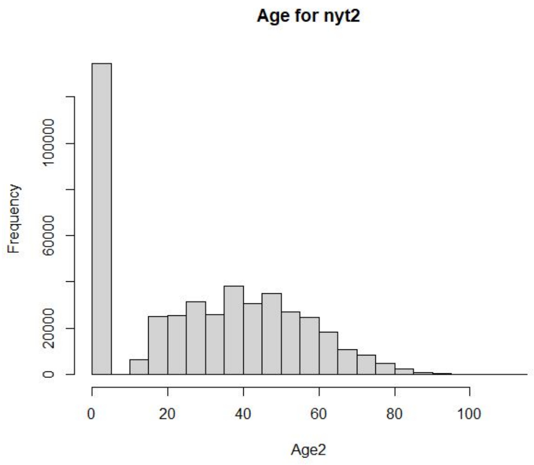
The five datasets I chose are nyt2, nyt3, nyt4, nyt5, and nyt6. And the two attributes are the Age and the Impression. Based on the Age boxplot, the median age for all datasets are around 31 years old. The min is 0 and the max is about 110. The max for nyt2 is slightly higher. The majority of people are from age of 0 to 45 so the data is skewed to the right across all datasets. For the impressions, the median for all datasets are around 5 and the majority of the data are in the range of 0 to 10 with >10 being outliers. The impression is strongly skewed to the right.

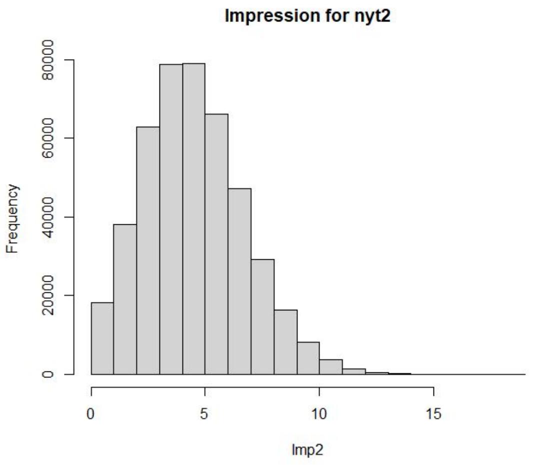


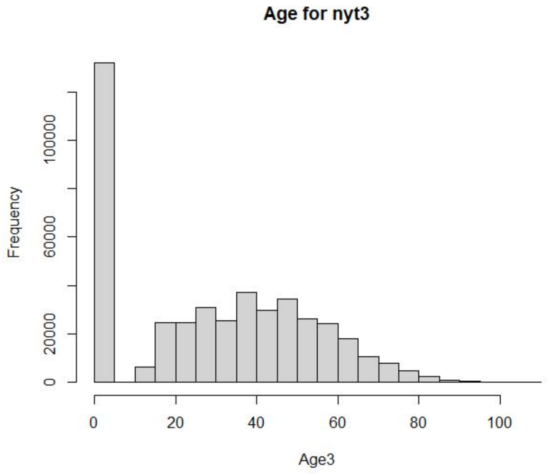


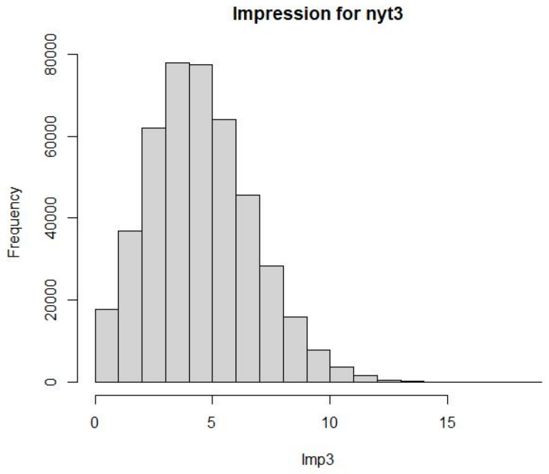
b). Create histograms for all 5 datasets for two key variables – can be the same variables in 1a or different (you choose the histogram bin width). Describe the distributions in terms of known parametric distributions and similarities/ differences among them. min. 3-4 sentences (3%)

For the Age histograms, the distributions are very similar. The largest age group is between the age of 5 to the age of 0. This is true for all of the datasets. The center of the distribution is around 45 years old, and the people over the age of 80 become increasingly small. The impression distribution for all 5 datasets is again, very similar. The modes of impressions are 4 and 5 times. The mean is around 5. The distribution roughly follows a normal distribution, with the exception that the distribution is slightly skewed to the right.

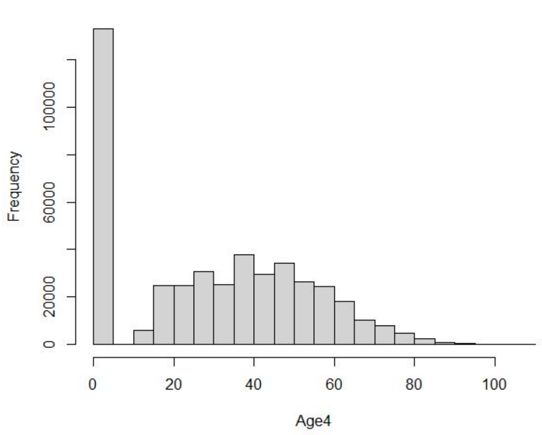




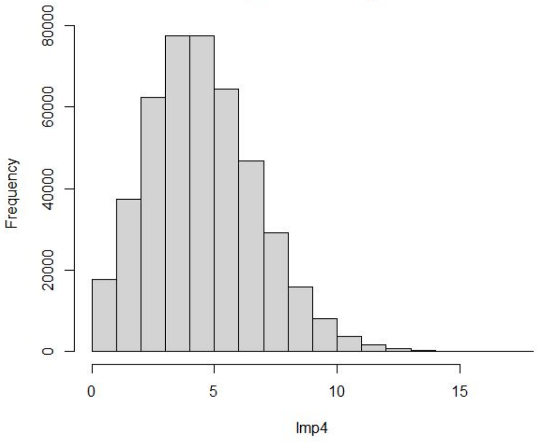




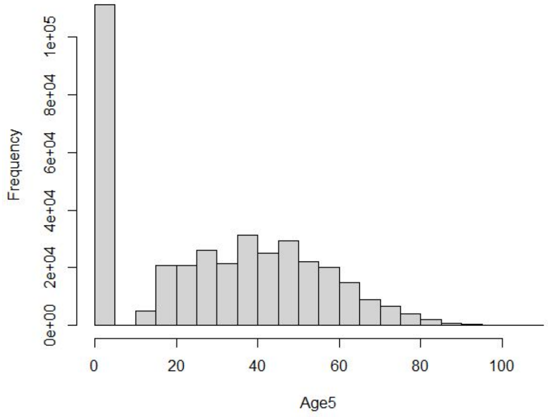
Age for nyt4



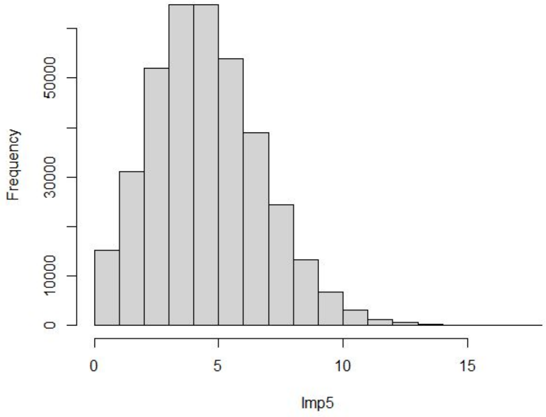
Impression for nyt4



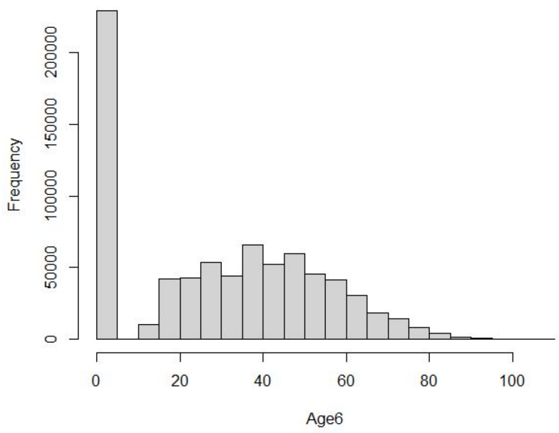
Age for nyt5



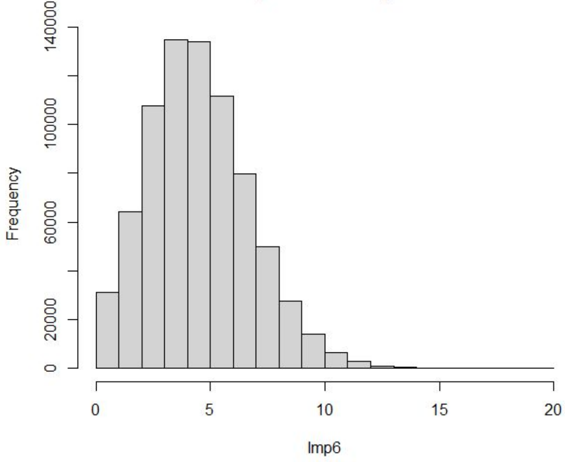
Impression for nyt5



Age for nyt6



Impression for nyt6



c). Plot the ECDFs (Empirical Cumulative Distribution Function for your two key variables. Plot the quantile-quantile distribution using a suitable parametric distribution you chose in 1b. Describe features of these plots. min. 3-4 sentences (4000-level 5%, 6000-level 3%)

All of the ECDFs plots appear to be “S” shaped. All of the qq plots seem to be “S” shaped as well.

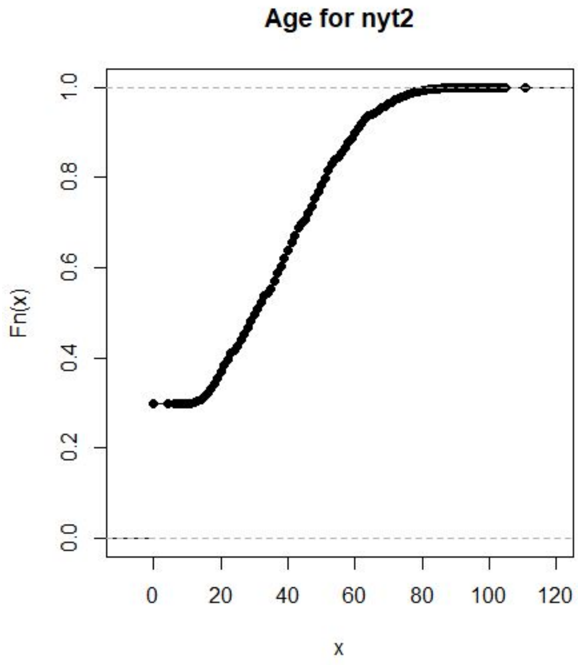
For nyt2 quantile-quantile distribution, the range for the age is from 0.3 to 1, the impression is from 0 to 1.

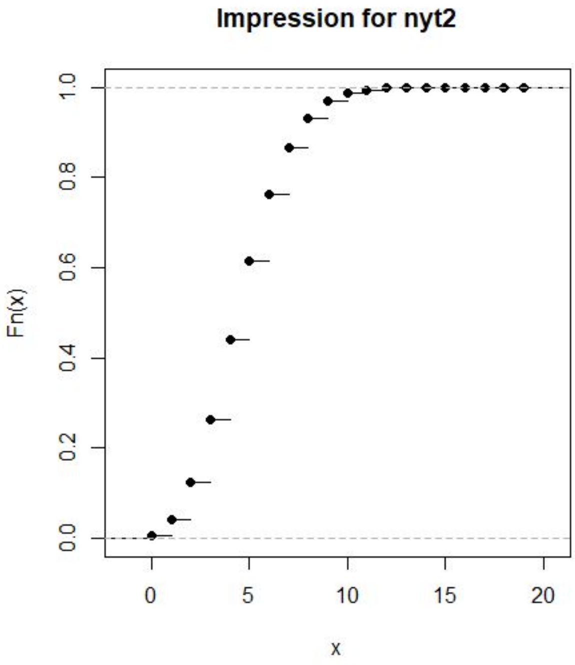
For nyt3 quantile-quantile distribution, the range for the age is from 0.3 to 1, the impression is from 0 to 1.

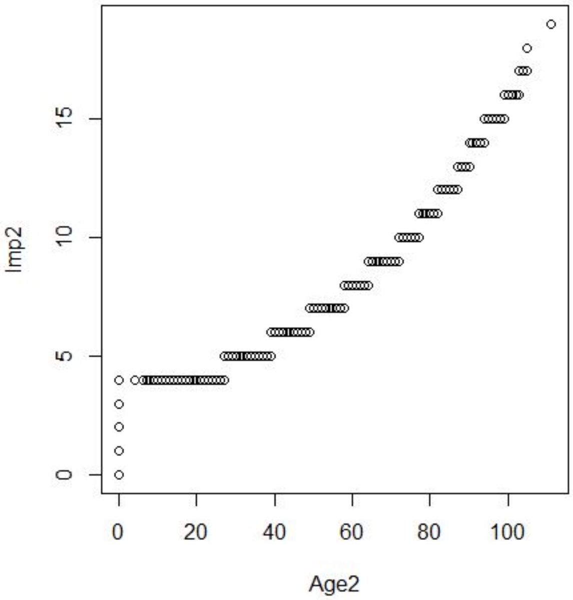
For nyt4 quantile-quantile distribution, the range for the age is from 0.25 to 1, the impression is from 0 to 1.

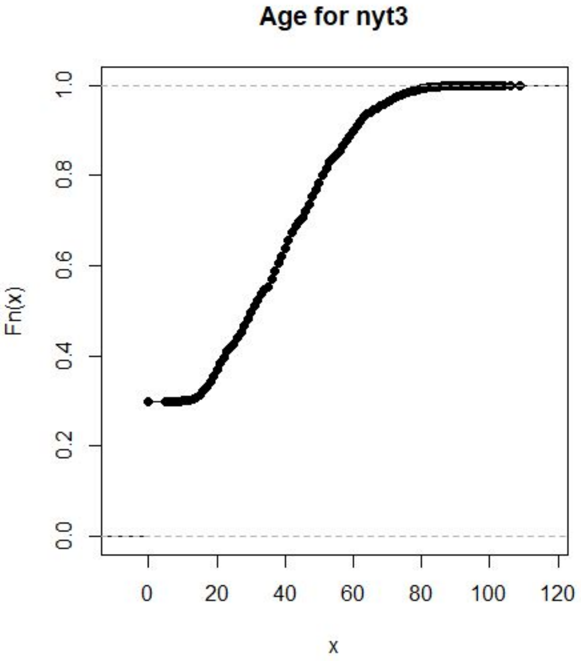
For nyt5 quantile-quantile distribution, the range for the age is from 0.3 to 1, the impression is from 0 to 1.

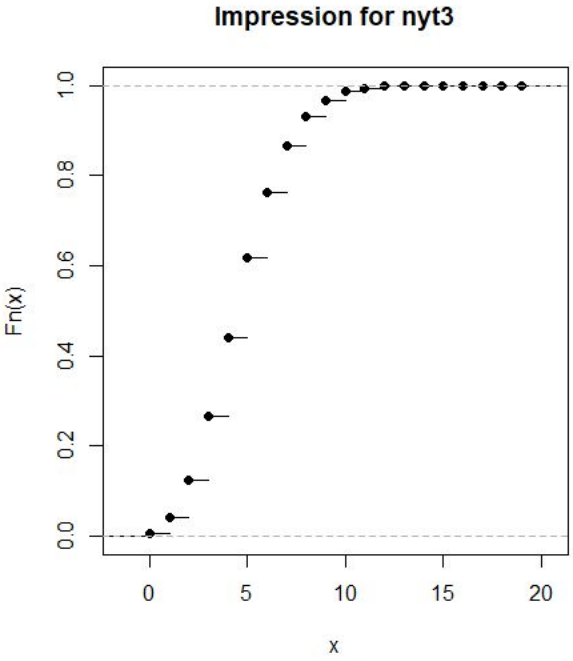
For nyt6 quantile-quantile distribution, the range for the age is from 0.3 to 1, the impression is from 0 to 1.

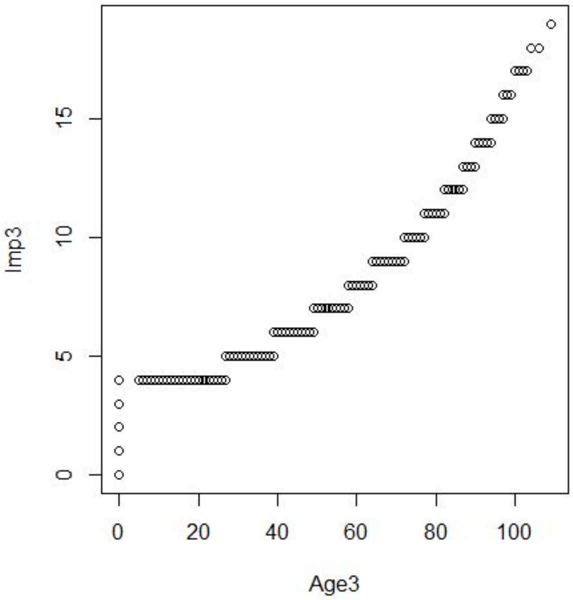


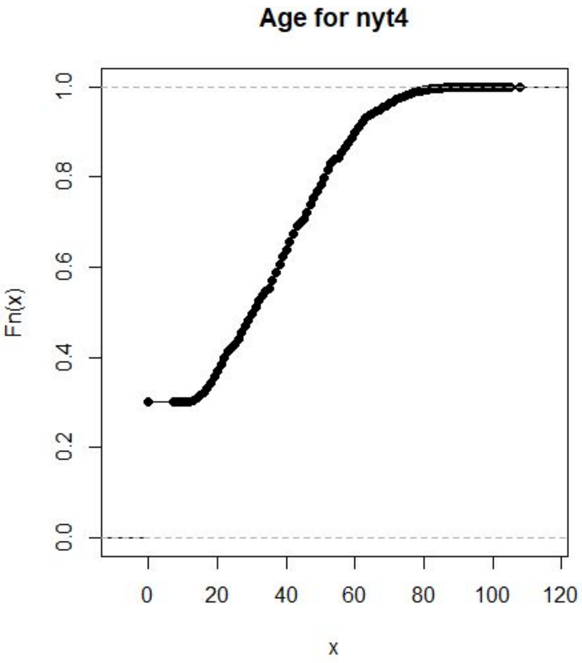


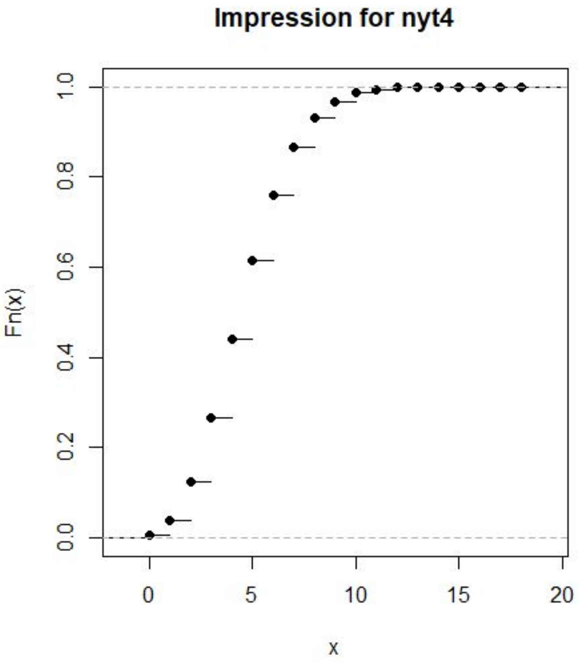


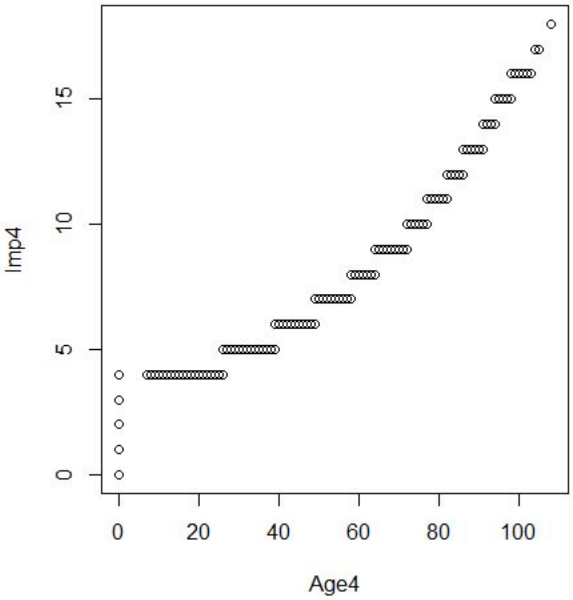

Age for nyt3

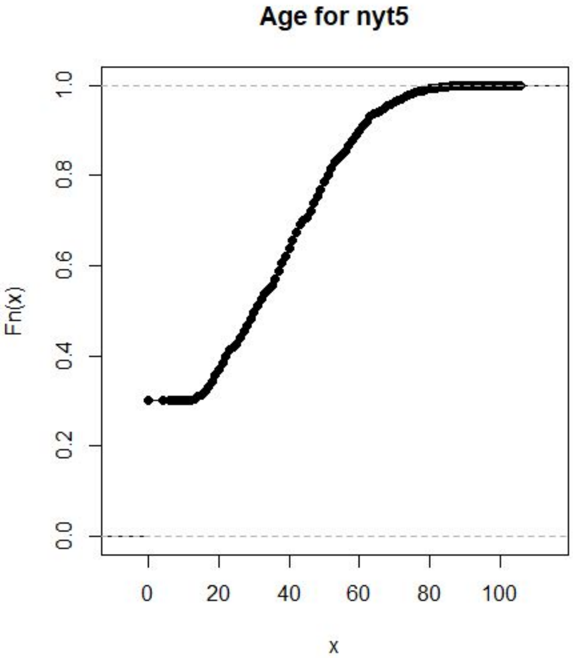

Impression for nyt3

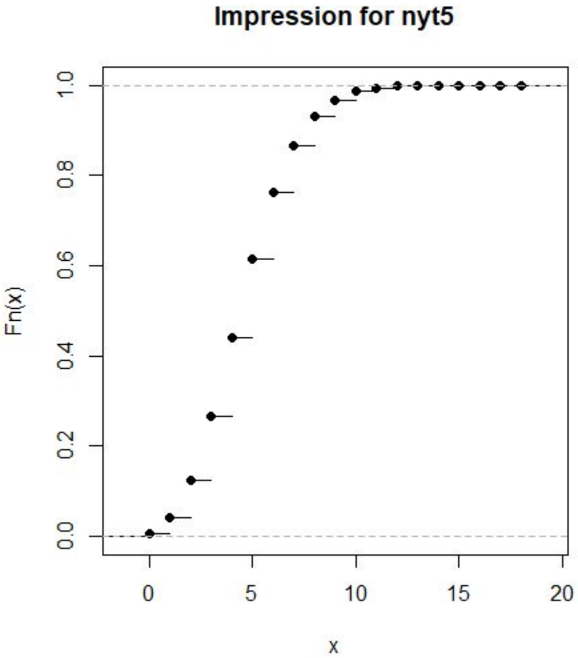

Imp3
0
5
10
15
Age3
0
20
40
60
80
100

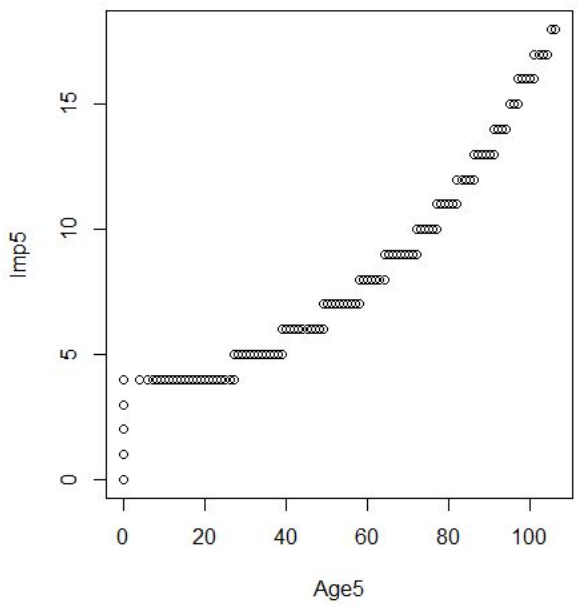

Age for nyt4
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
x
0
20
40
60
80
100
120

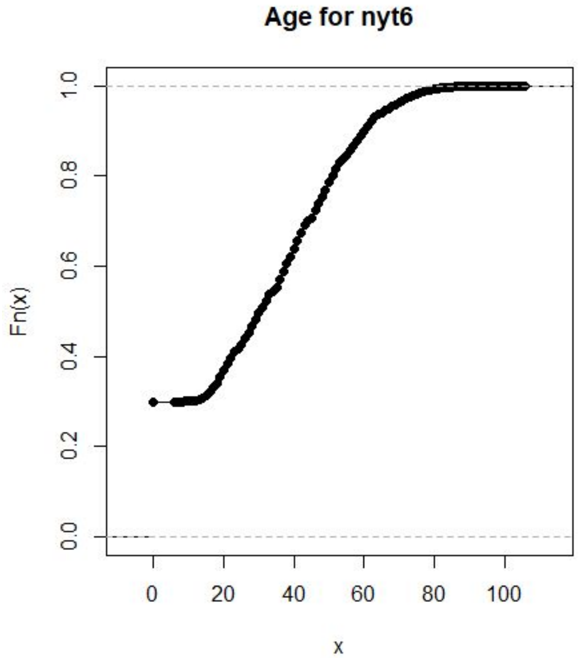

Impression for nyt4
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
x
0
5
10
15
20

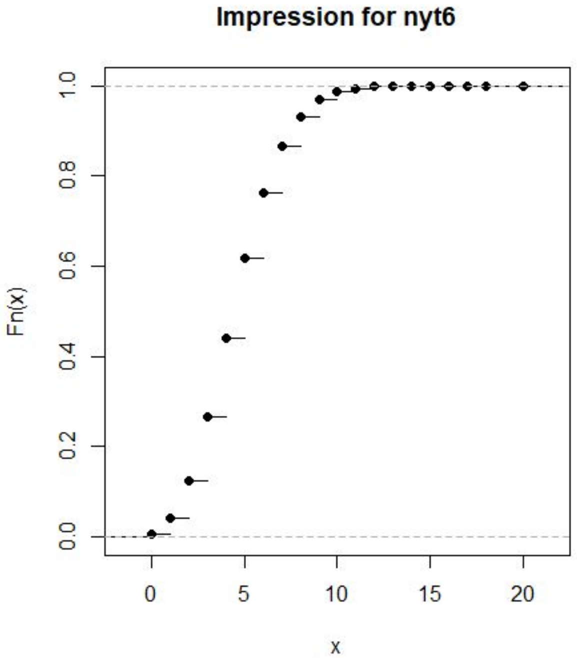

Imp4
0
5
10
15
0
20
40
60
80
100
Age4

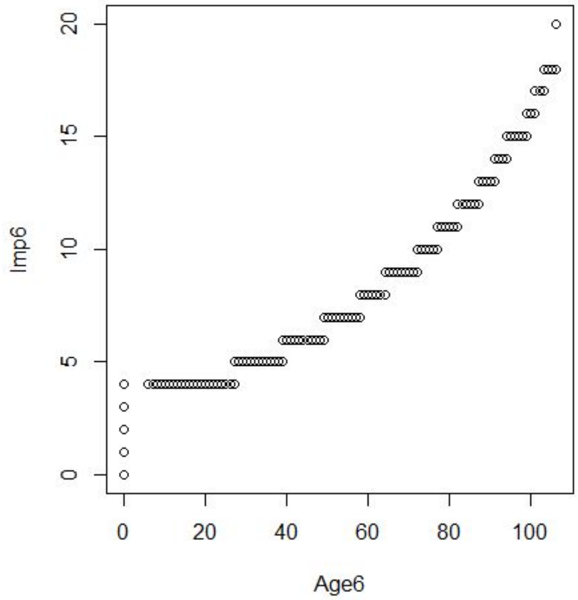

Age for nyt5
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
x


Impression for nyt5
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
x




Age for nyt6


Impression for nyt6



d). Perform a significance test that is suitable for the variables you are investigating. Discuss the test results and indicate whether the null hypothesis is valid. min. 3-4 sentences (4000-level 4%, 6000-level 3%)

For all of the datasets, the p-values are all above 0.5 which is a strong indicator that the correlation is low. From the sample estimated cor, the values are also very low. This means that the datasets do not have enough evidence to reject the null hypothesis. There is no to very little to no evidence that the two variables are correlated.

```
> cor.test(Age2, Imp2)
data:  Age2 and Imp2
t = 0.058745, df = 449933, p-value = 0.9532
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.002834377  0.003009532
sample estimates:
        cor
8.757832e-05

> cor.test(Age3, Imp3)
data:  Age3 and Imp3
```

```
t = -0.25811, df = 440368, p-value = 0.7963
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.003342459  0.002564573
sample estimates:
          cor
-0.0003889463

> cor.test(Age4, Imp4)
data:  Age4 and Imp4
t = -0.48762, df = 442855, p-value = 0.6258
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.003677950  0.002212471
sample estimates:
          cor
-0.0007327455

> cor.test(Age5, Imp5)
data:  Age5 and Imp5
t = -0.2012, df = 370326, p-value = 0.8405
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.003551360  0.002890111
sample estimates:
          cor
-0.0003306278

> cor.test(Age6, Imp6)
data:  Age6 and Imp6
t = 0.38002, df = 764508, p-value = 0.7039
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.001806966  0.002676217
```

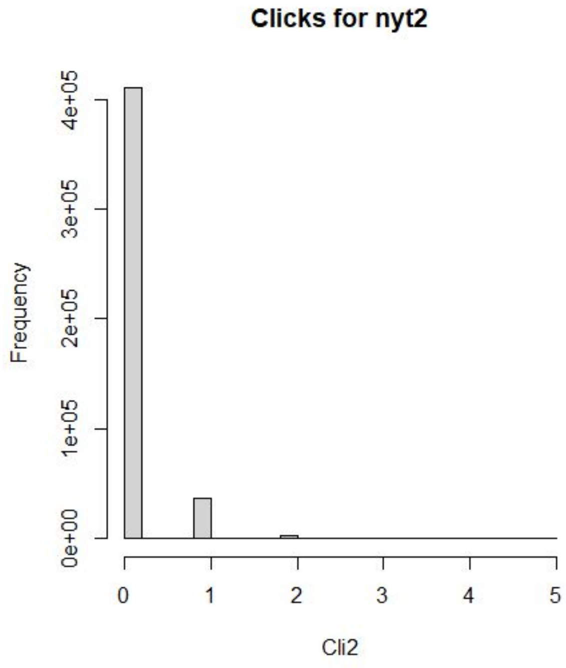
```
sample estimates:
        cor
0.0004346276
```

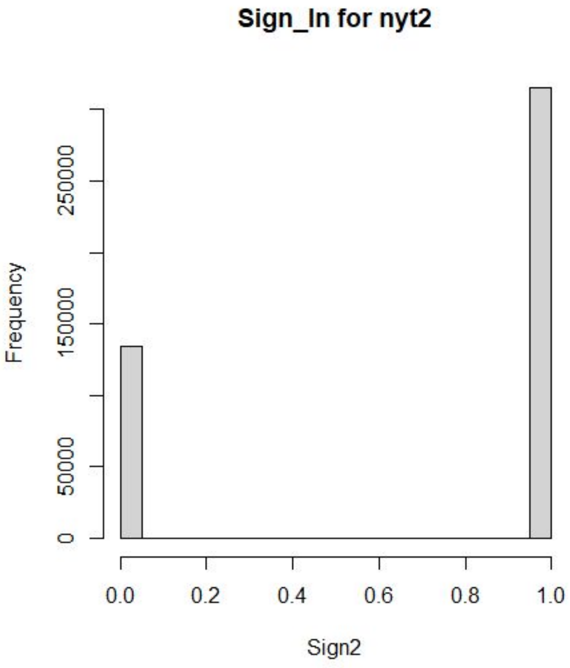
e). Discuss any observations you had about the datasets/ variables, other data in the dataset (0% ;-))

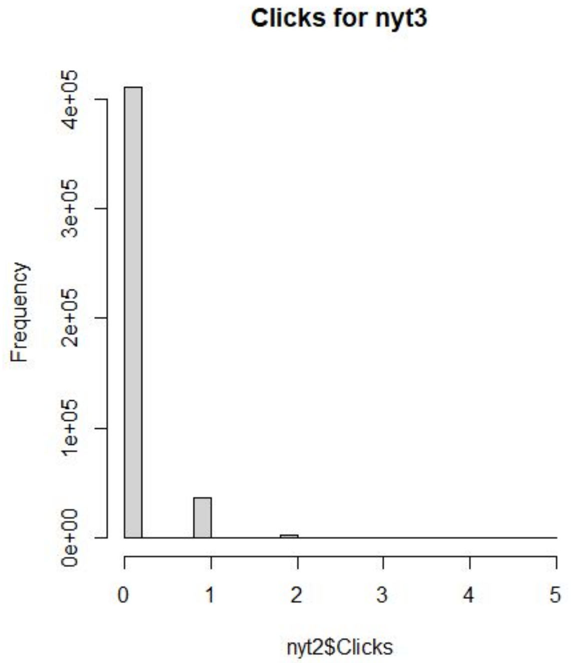
There is very little correlation between age and impression, which means people of all age groups tend to click at the same rate. Largely, this observation makes sense, but for very children and old people, I expect the impression to be low.

2. 6600-level question (3%). Filter the distributions you explored in Q1 using one or more of the other variables for only 2 (not 5) of the nyt datasets. Repeat Q1b, Q1c and Q1d and draw any conclusions from this study. min. 3-4 sentences

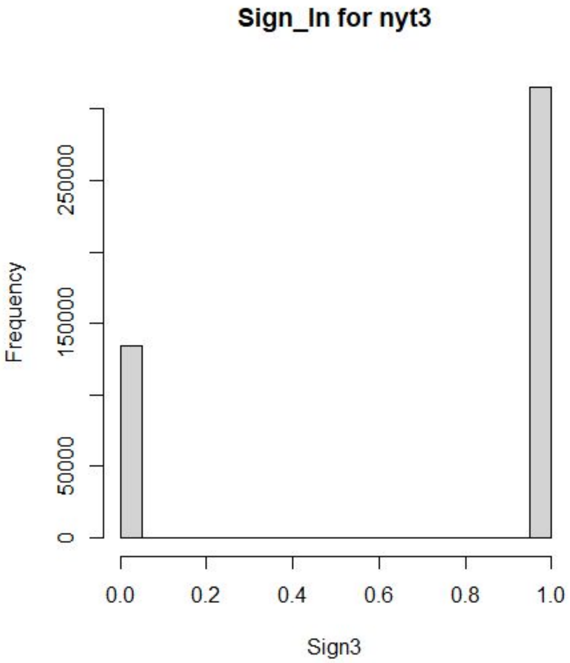
For part 2, clicks and sign_in are chosen for the variables. Sign_in are values between 0 and 1, 0 being not signed in and vise versa. The majority of people did not click and a few people clicked once. Based on the p-value, this isn't any correlation between clicks and sign_in.



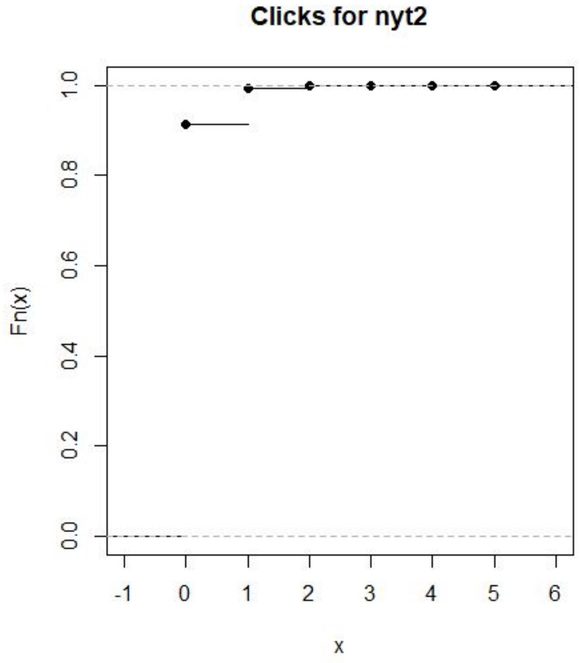




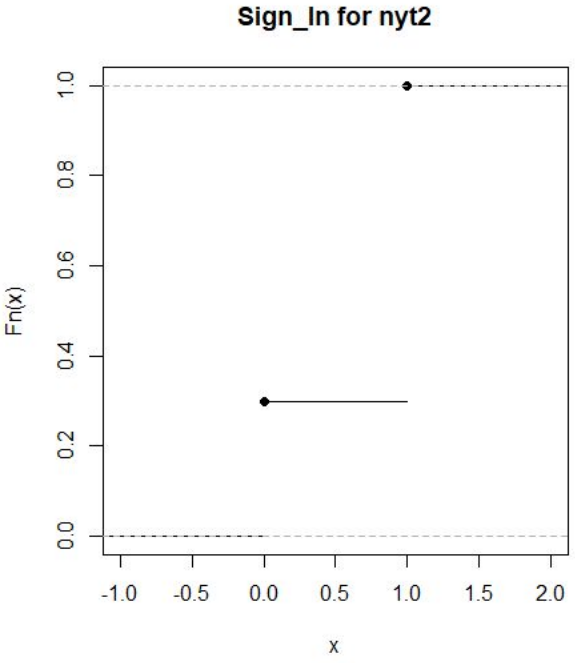
Clicks for nyt3



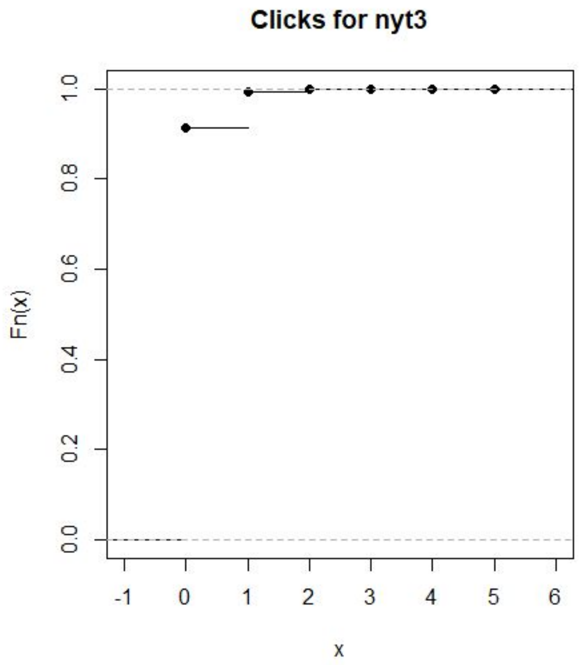
Sign_In for nyt3



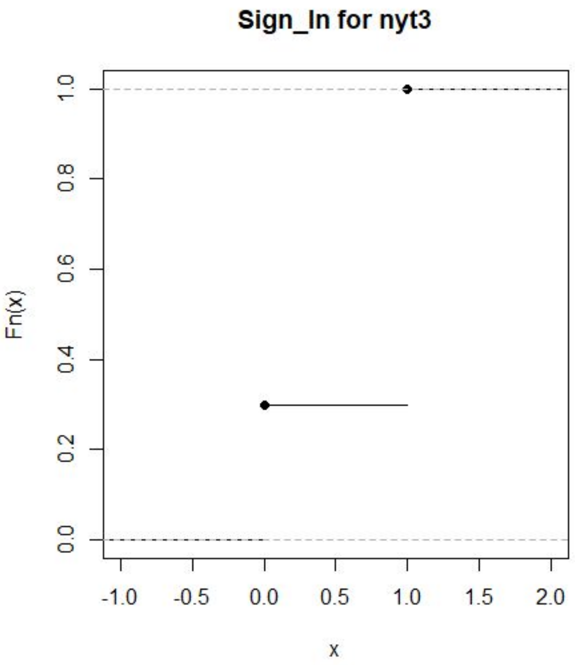
Clicks for nyt2



Sign_In for nyt2



Clicks for nyt3



Sign_In for nyt3

```
> cor.test(Cli2, Sign3)
data:  Cli2 and Sign3
t = -70.017, df = 449933, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1067088 -0.1009278
sample estimates:
       cor
-0.1038192
```

```
> cor.test(Cli2, Sign3)
data:  Cli2 and Sign3
t = -70.017, df = 449933, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1067088 -0.1009278
sample estimates:
       cor
-0.1038192
```