Shiwei Huang

1. Choose any 5 of the code and the plots you generate for the question below.

a). Create boxplots for all 5 datasets for each of key variables, i.e. two figures (one for each variable) with 5 boxplots (for the 5 different datasets) in each.

Describe/summarize the distributions.

For the age variable,

My nyt8 medium age is: 31

My nyt10 medium age is: 31

My ny15 medium age is: 20

My nyt12 medium age is: 32

My nyt25 medium age is: 22

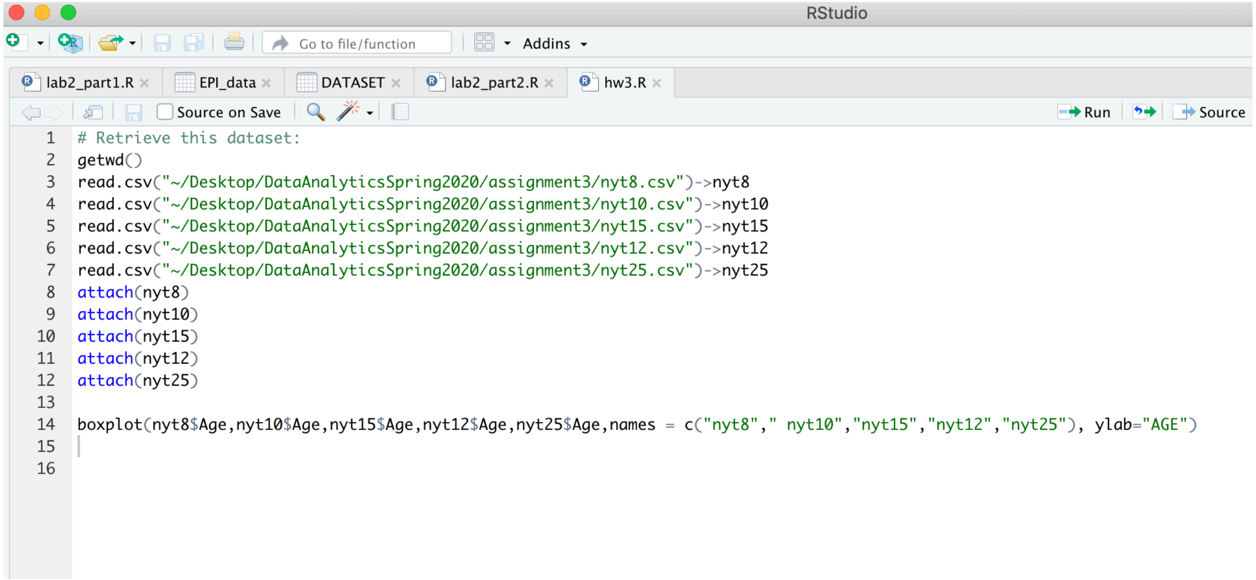
My nyt8 medium impressions is: 5

My nyt10 medium impressions is: 5

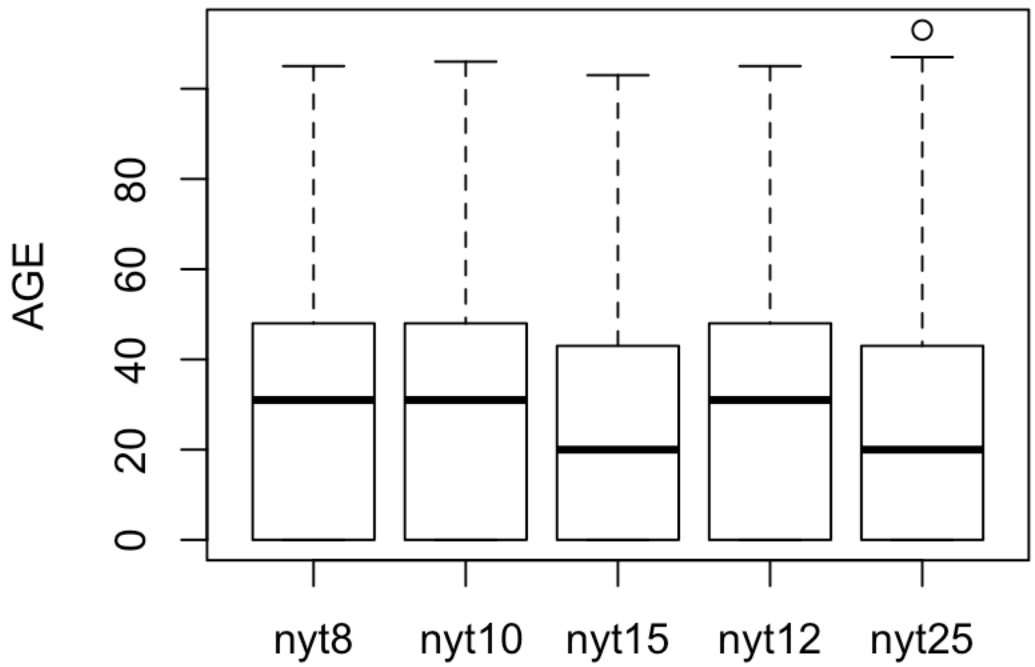
My ny15 medium impressions is: 5

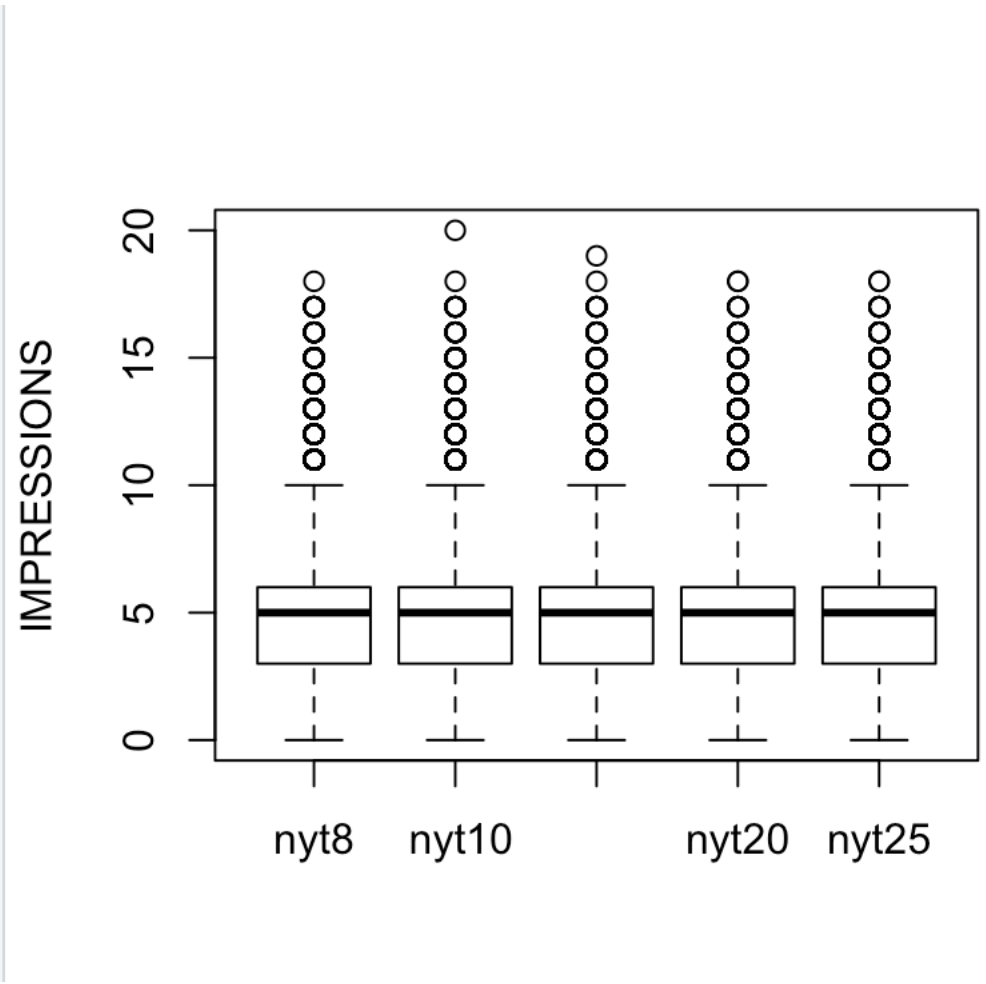
My nyt12 medium impressions is: 5

My nyt25 medium impressions is: 5

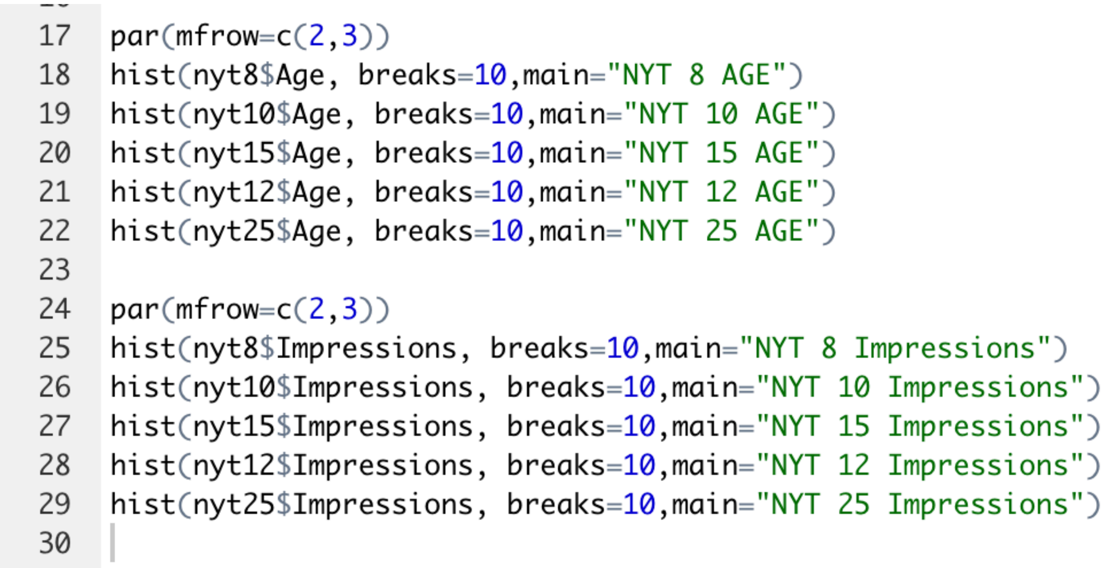


```
# Retrieve this dataset:
getwd()
read.csv("~/Desktop/DataAnalyticsSpring2020/assignment3/nyt8.csv")->nyt8
read.csv("~/Desktop/DataAnalyticsSpring2020/assignment3/nyt10.csv")->nyt10
read.csv("~/Desktop/DataAnalyticsSpring2020/assignment3/nyt15.csv")->nyt15
read.csv("~/Desktop/DataAnalyticsSpring2020/assignment3/nyt12.csv")->nyt12
read.csv("~/Desktop/DataAnalyticsSpring2020/assignment3/nyt25.csv")->nyt25
attach(nyt8)
attach(nyt10)
attach(nyt15)
attach(nyt12)
attach(nyt25)

boxplot(nyt8$Age,nyt10$Age,nyt15$Age,nyt12$Age,nyt25$Age,names = c("nyt8"," nyt10","nyt15","nyt12","nyt25"), ylab="AGE")
```


AGE
0
20
40
60
80
nyt8
nyt10
nyt15
nyt12
nyt25

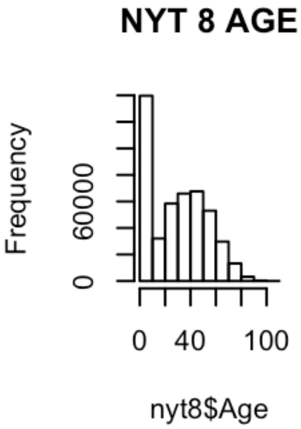


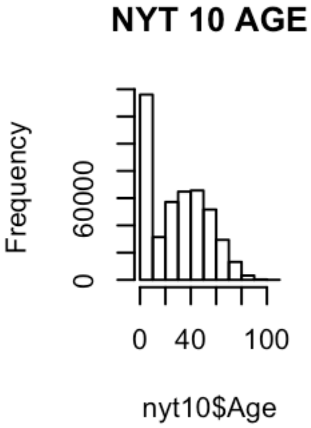
b). Create histograms for all 5 datasets for two key variables – can be the same variables in 1a or different. Describe the distributions in terms of known parametric distributions and similarities/ differences among them.

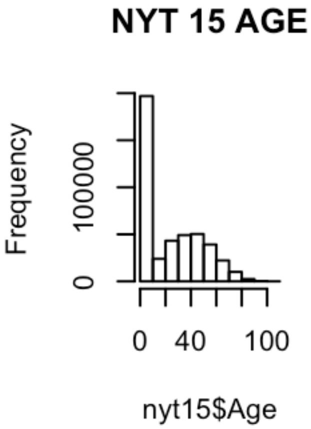


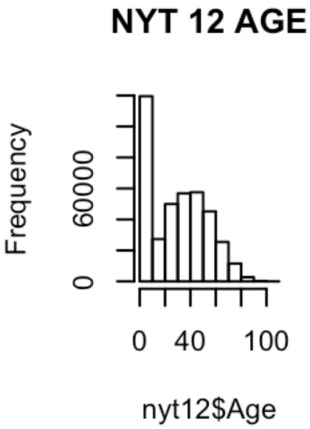
```r
par(mfrow=c(2,3))
hist(nyt8$Age, breaks=10,main="NYT 8 AGE")
hist(nyt10$Age, breaks=10,main="NYT 10 AGE")
hist(nyt15$Age, breaks=10,main="NYT 15 AGE")
hist(nyt12$Age, breaks=10,main="NYT 12 AGE")
hist(nyt25$Age, breaks=10,main="NYT 25 AGE")

par(mfrow=c(2,3))
hist(nyt8$Impressions, breaks=10,main="NYT 8 Impressions")
hist(nyt10$Impressions, breaks=10,main="NYT 10 Impressions")
hist(nyt15$Impressions, breaks=10,main="NYT 15 Impressions")
hist(nyt12$Impressions, breaks=10,main="NYT 12 Impressions")
hist(nyt25$Impressions, breaks=10,main="NYT 25 Impressions")

```

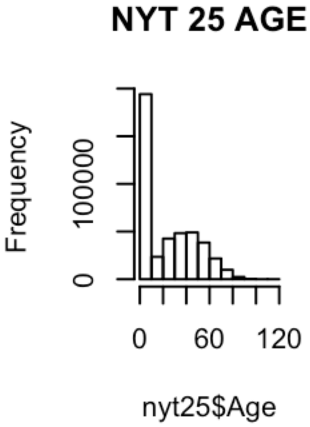
NYT 8 Age distribution: Most of the people are around of age 20-60.
NYT 10 Age distribution: Most of the people are around of age 40
NYT 15 Age distribution: Most of the people are around of age 40
NYT12 Age distribution: Most of the people are around of age 20-60
NYT 25 Age distribution: Most of the people are around of age 40. The ages are increasingly decreased after 60.



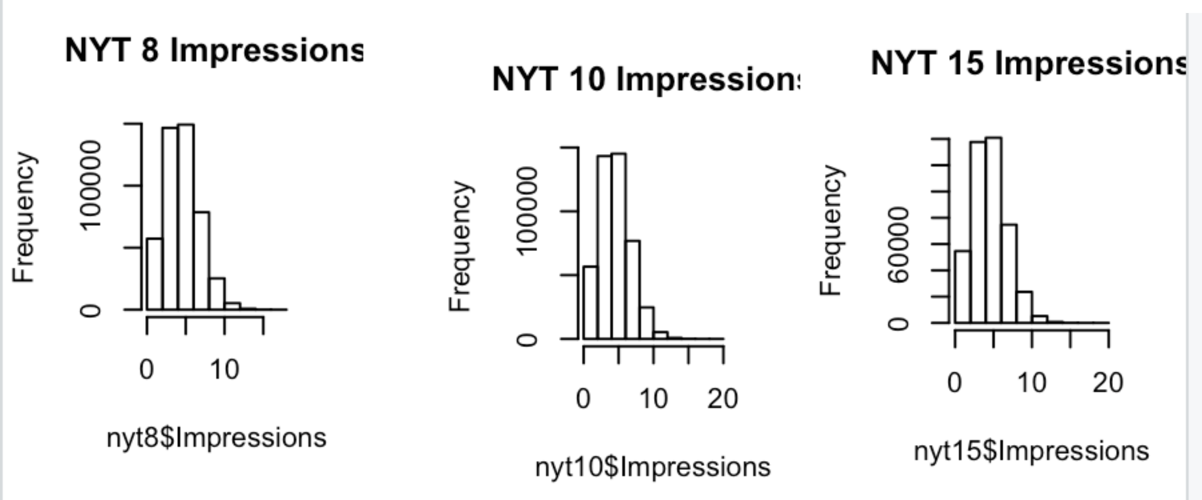


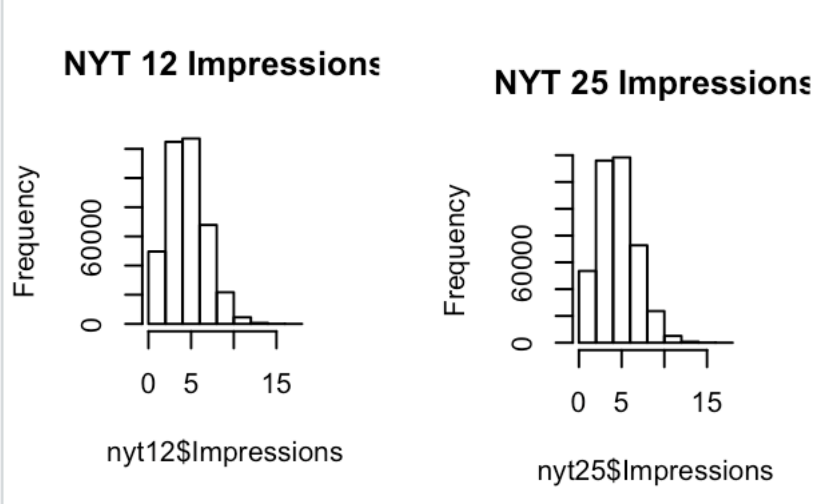






NYT 8 Impressions distribution: Most of the people are around of impressions 5.
NYT 10 Impressions distribution: Most of the people are around of Impressions 5.
NYT 15 Impressions distribution: Most of the people are around of Impressions 5.
NYT12 Impressions distribution: Most of the people are around of Impressions 5.
NYT 25 Impressions distribution: Most of the people are around of Impressions 5.

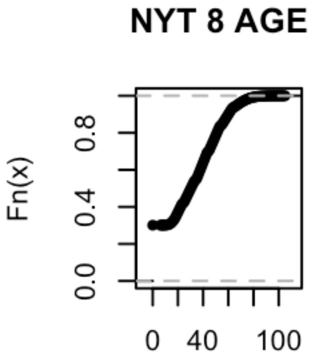


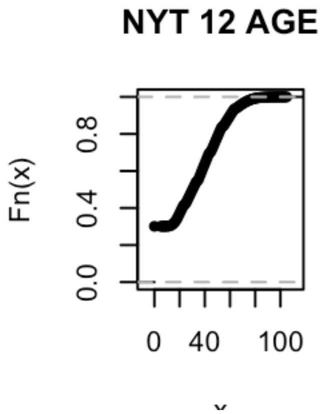


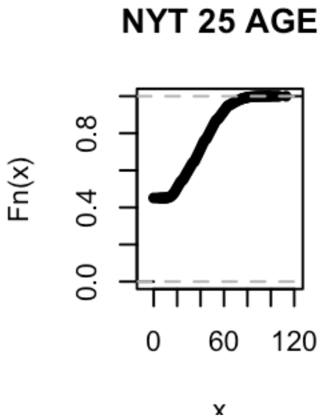
c. Plot the ECDFs. Plot the quantile-quantile distribution using a suitable parametric distribution you chose in 1b. Describe features of these plots.

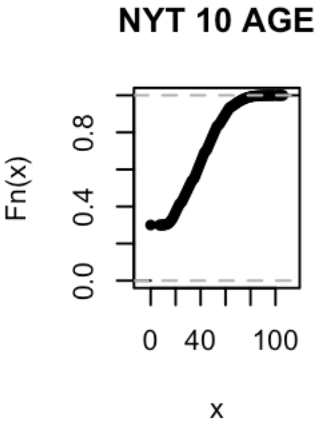
Empirical cumulative distribution curve(ECDF)

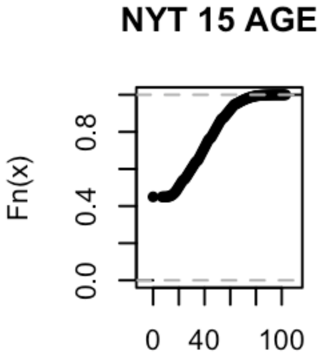
The data ranges from 0.3-1 for NYT8 Age
For NYT10 Age, the data ranges from 0.3 - 1
For NYT12 Age, the data ranges from 0.3 – 1 as well
For NYT15 Age, the data ranges from 0.42 - 1
For NYT25 Age, the data ranges from 0.5 – 1



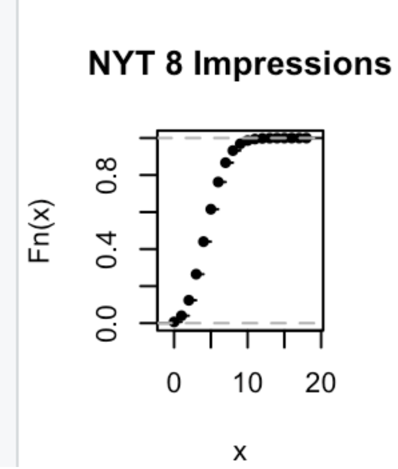


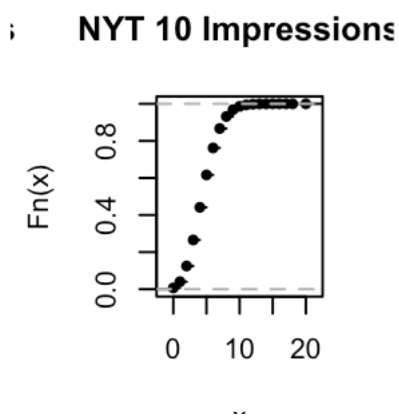


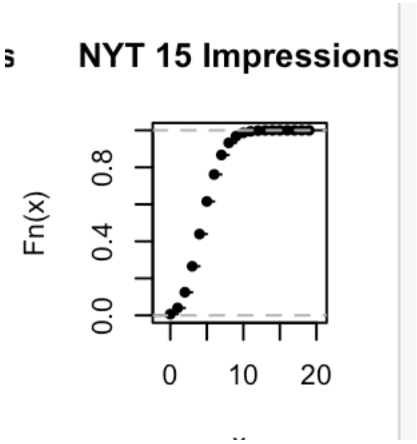


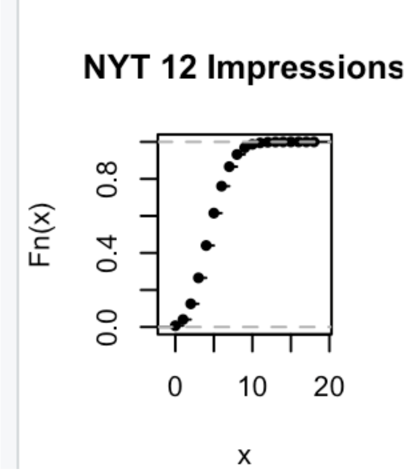


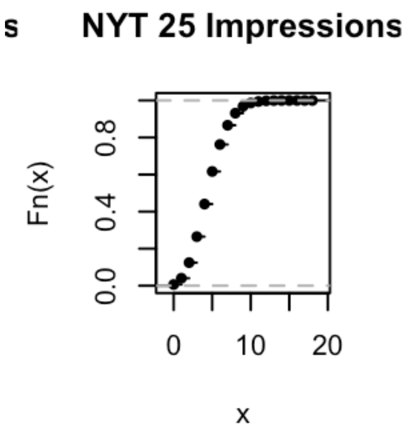
The data ranges from mostly 0.8 for NYT8 Impressions
For NYT10 Impressions, the data ranges from 0.8
For NYT12 Impressions, the data ranges from 0.8 as well
For NYT15 Impressions, the data ranges from 0.8
For NYT25 Impressions, the data ranges from 0.8



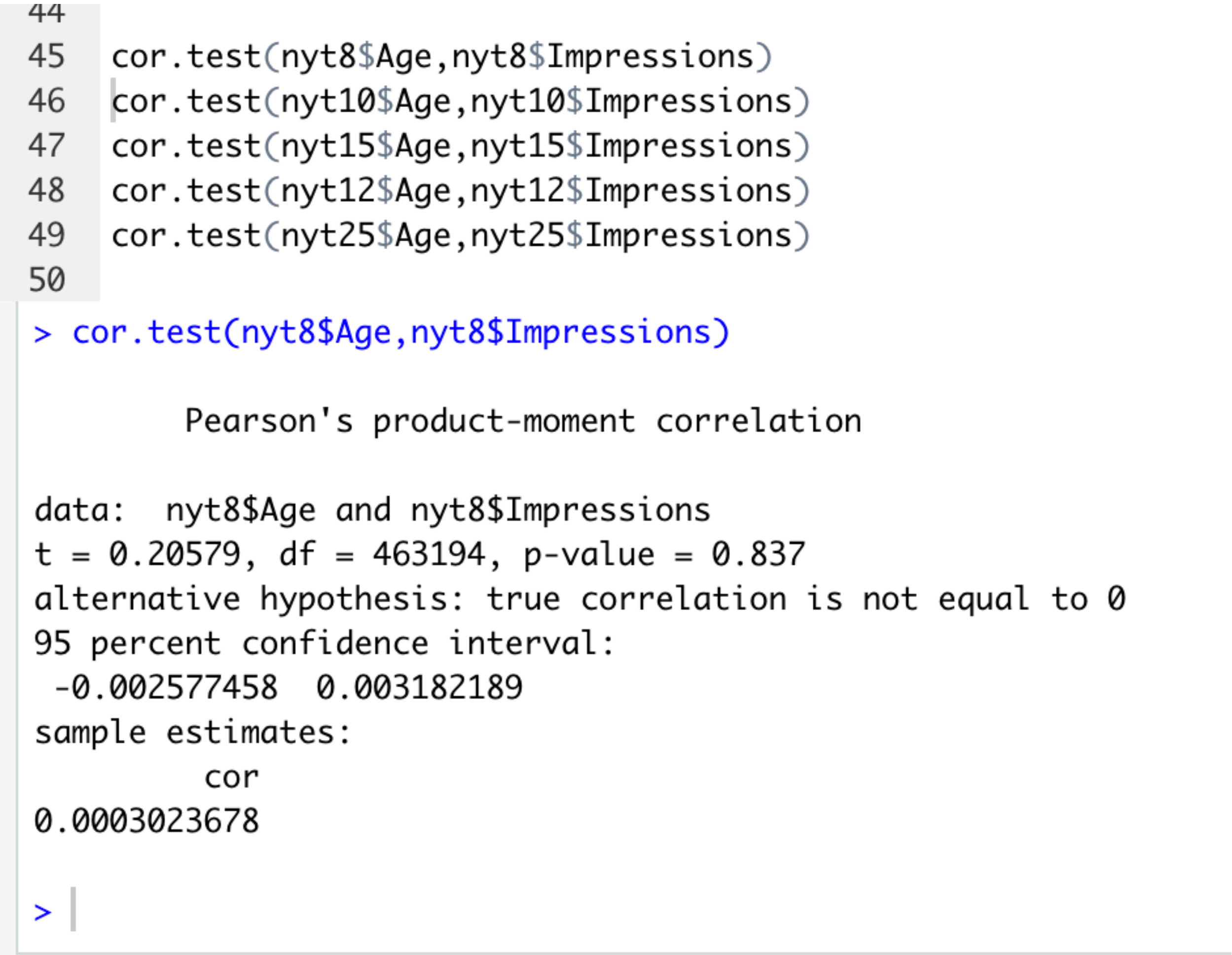






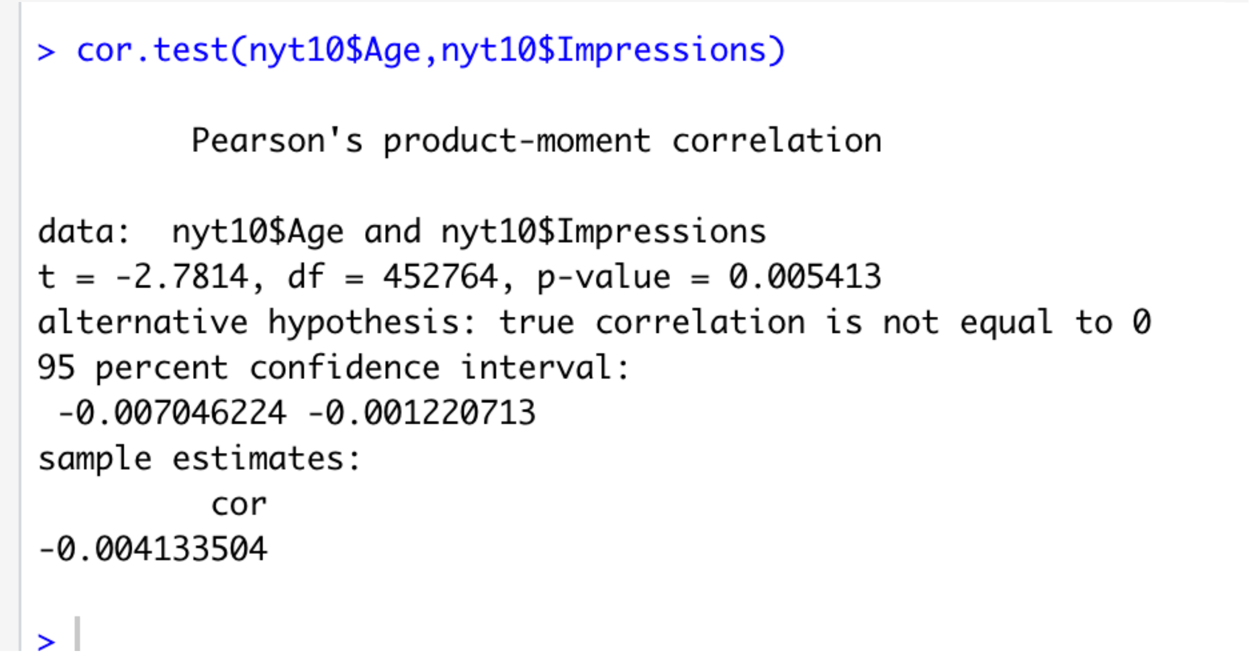


d. Perform a significance test that is suitable for the variables you are investigating. Discuss the test results and indicate whether the null hypothesis is valid.

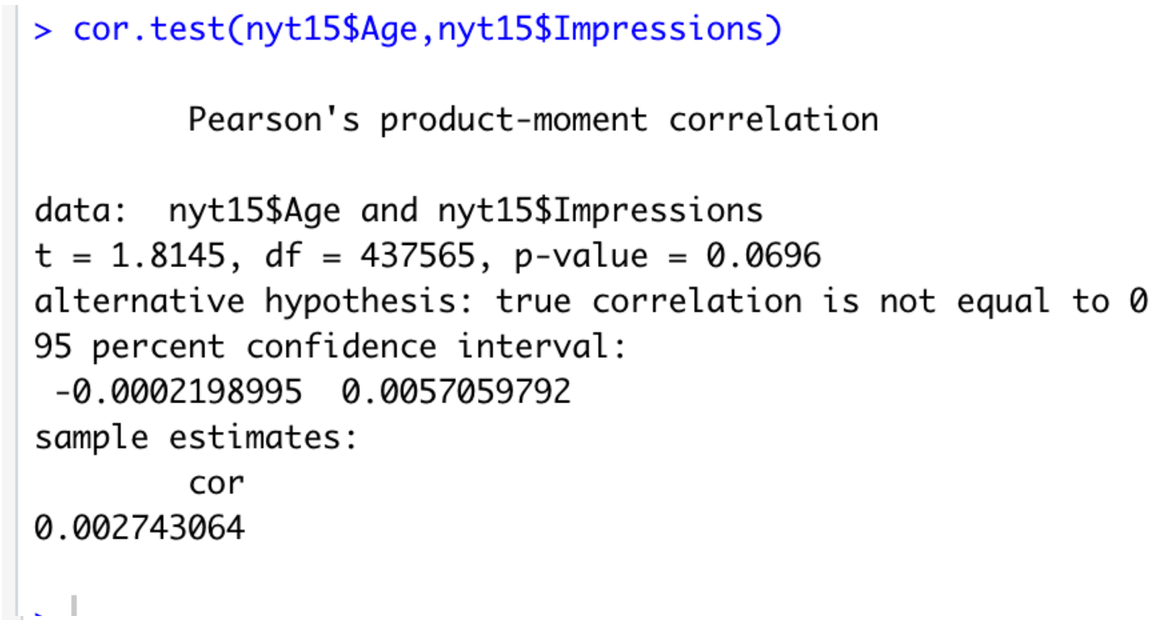


```
cor.test(nyt8$Age,nyt8$Impressions)
cor.test(nyt10$Age,nyt10$Impressions)
cor.test(nyt15$Age,nyt15$Impressions)
cor.test(nyt12$Age,nyt12$Impressions)
cor.test(nyt25$Age,nyt25$Impressions)
```

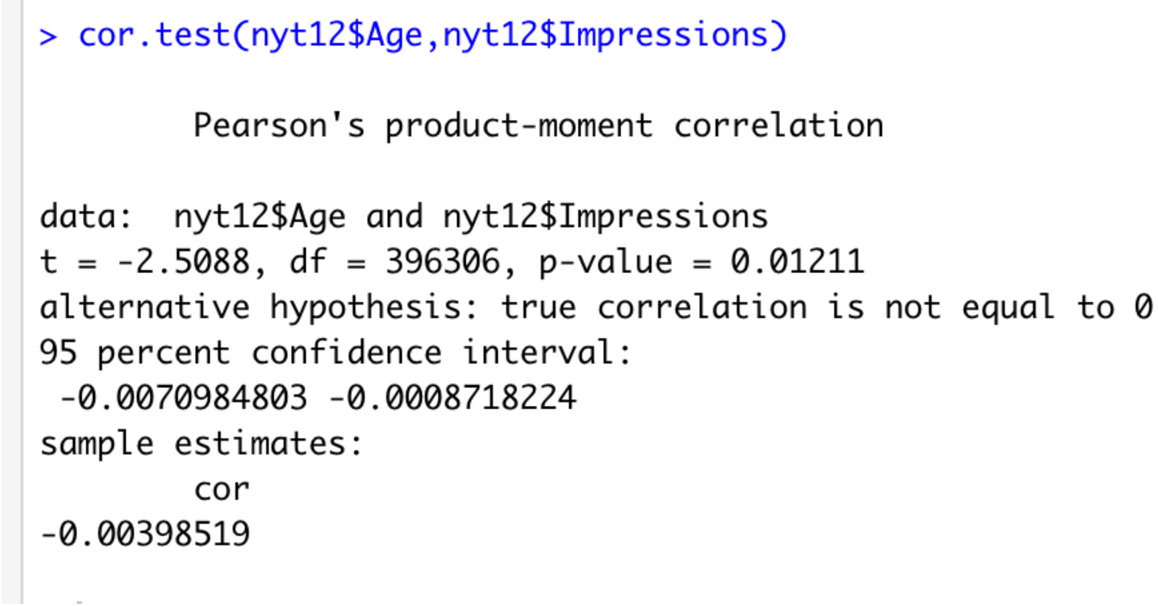
```
> cor.test(nyt8$Age,nyt8$Impressions)

	Pearson's product-moment correlation

data:  nyt8$Age and nyt8$Impressions
t = 0.20579, df = 463194, p-value = 0.837
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.002577458  0.003182189
sample estimates:
         cor 
0.0003023678 

> 
```



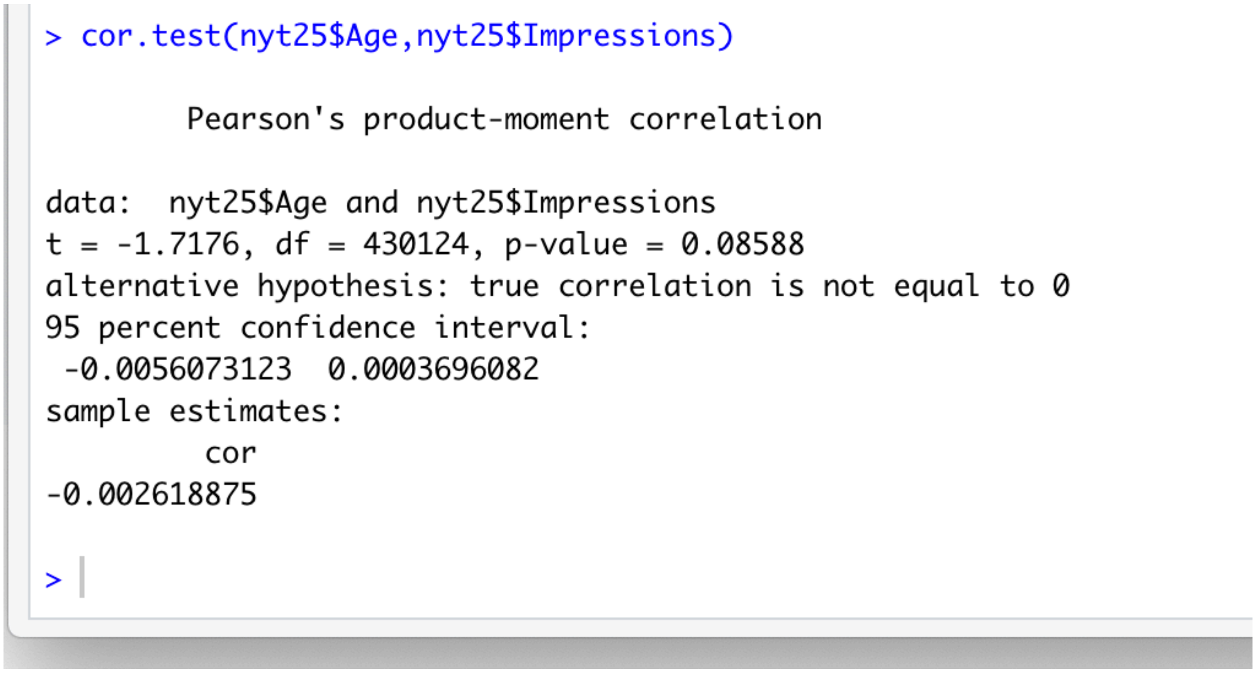
```
> cor.test(nyt10$Age,nyt10$Impressions)

	Pearson's product-moment correlation

data:  nyt10$Age and nyt10$Impressions
t = -2.7814, df = 452764, p-value = 0.005413
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.007046224 -0.001220713
sample estimates:
         cor 
-0.004133504 

> 
```



```
> cor.test(nyt15$Age,nyt15$Impressions)

	Pearson's product-moment correlation

data:  nyt15$Age and nyt15$Impressions
t = 1.8145, df = 437565, p-value = 0.0696
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0002198995  0.0057059792
sample estimates:
        cor 
0.002743064 
```



```
> cor.test(nyt12$Age,nyt12$Impressions)

	Pearson's product-moment correlation

data:  nyt12$Age and nyt12$Impressions
t = -2.5088, df = 396306, p-value = 0.01211
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0070984803 -0.0008718224
sample estimates:
        cor 
-0.00398519 
```



```
> cor.test(nyt25$Age,nyt25$Impressions)

	Pearson's product-moment correlation

data:  nyt25$Age and nyt25$Impressions
t = -1.7176, df = 430124, p-value = 0.08588
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0056073123  0.0003696082
sample estimates:
         cor 
-0.002618875 

> 
```

```
> cor.test(nyt8$Age,nyt8$Impressions)

	Pearson's product-moment correlation

data:  nyt8$Age and nyt8$Impressions
t = 0.20579, df = 463194, p-value = 0.837
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.002577458  0.003182189
sample estimates:
	cor 
0.0003023678 

> cor.test(nyt10$Age,nyt10$Impressions)

	Pearson's product-moment correlation

data:  nyt10$Age and nyt10$Impressions
t = -2.7814, df = 452764, p-value = 0.005413
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.007046224 -0.001220713
sample estimates:
	cor 
-0.004133504 
```

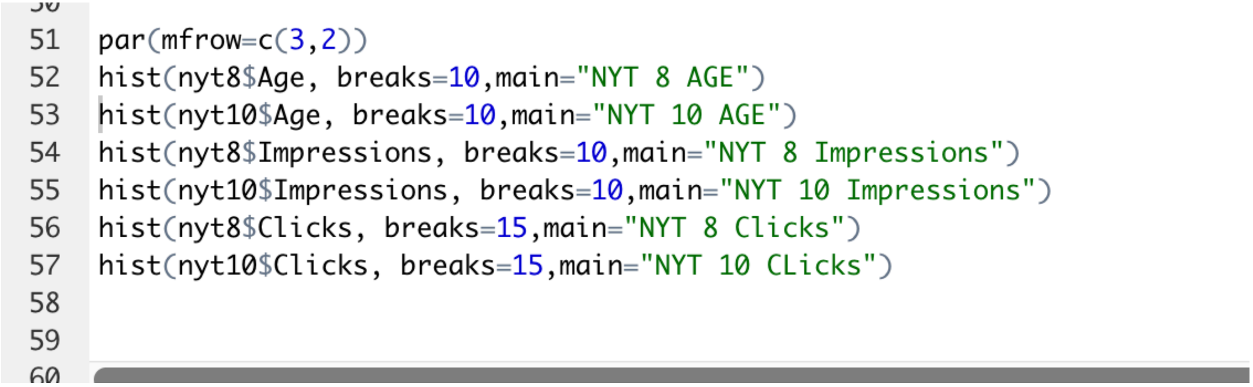
```
> cor.test(nyt15$Age,nyt15$Impressions)

        Pearson's product-moment correlation

data:  nyt15$Age and nyt15$Impressions
t = 1.8145, df = 437565, p-value = 0.0696
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0002198995  0.0057059792
sample estimates:
       cor
0.002743064


> cor.test(nyt12$Age,nyt12$Impressions)

        Pearson's product-moment correlation

data:  nyt12$Age and nyt12$Impressions
t = -2.5088, df = 396306, p-value = 0.01211
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0070984803 -0.0008718224
sample estimates:
       cor
-0.00398519


> cor.test(nyt25$Age,nyt25$Impressions)

        Pearson's product-moment correlation

data:  nyt25$Age and nyt25$Impressions
t = -1.7176, df = 430124, p-value = 0.08588
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0056073123  0.0003696082
sample estimates:
        cor
-0.002618875
```

e). Discuss any observations you had about the datasets/ variables, other than the data in the dataset.

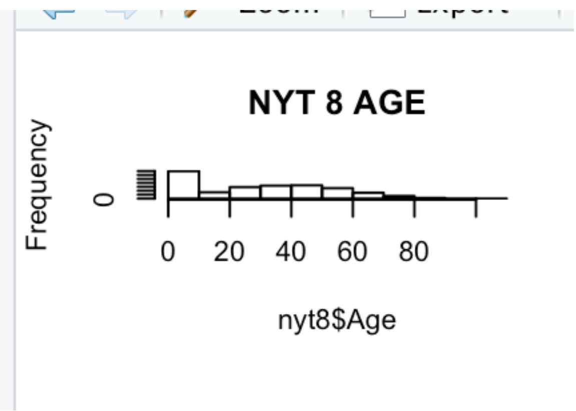
I figured out sign_in is a 0 or 1. So I figured out sign_in 0 means not signed in. And sign_in 1 means meaning signed in already.
Similar with the gender 0 and 1. They could be either male or female.

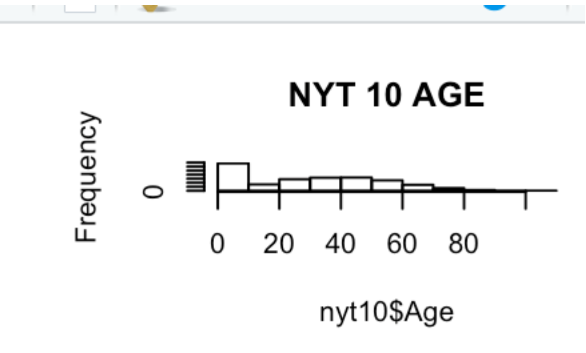
2. 6600 level question. Filter the distributions you explored in Q1 using one or more of other variables for only 2(not 5) of the nyt datasets. Repeat Q1b, Q1c and Q1d and draw any conclusions from this study.

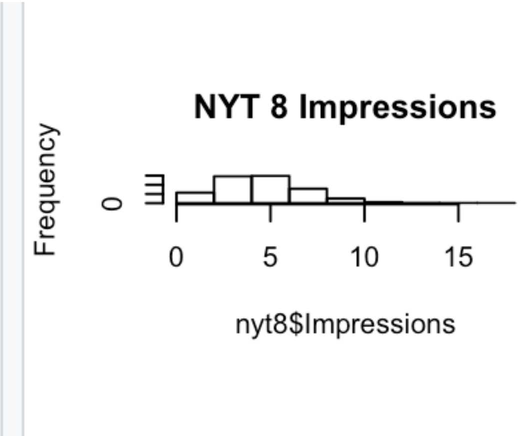
Repeat Q1b

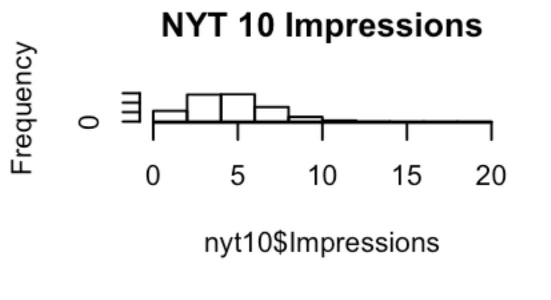


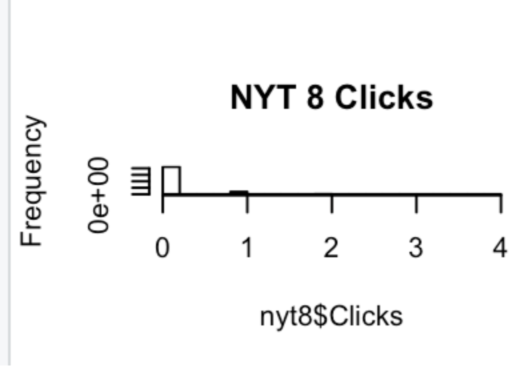
```
par(mfrow=c(3,2))
hist(nyt8$Age, breaks=10,main="NYT 8 AGE")
hist(nyt10$Age, breaks=10,main="NYT 10 AGE")
hist(nyt8$Impressions, breaks=10,main="NYT 8 Impressions")
hist(nyt10$Impressions, breaks=10,main="NYT 10 Impressions")
hist(nyt8$Clicks, breaks=15,main="NYT 8 Clicks")
hist(nyt10$Clicks, breaks=15,main="NYT 10 CLicks")
```

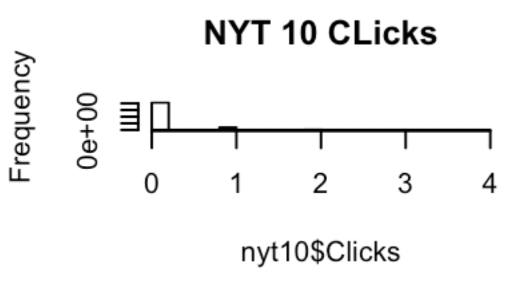

NYT 8 AGE
Frequency
0
0
20
40
60
80
nyt8$Age


NYT 10 AGE
Frequency
0
0
20
40
60
80
nyt10$Age

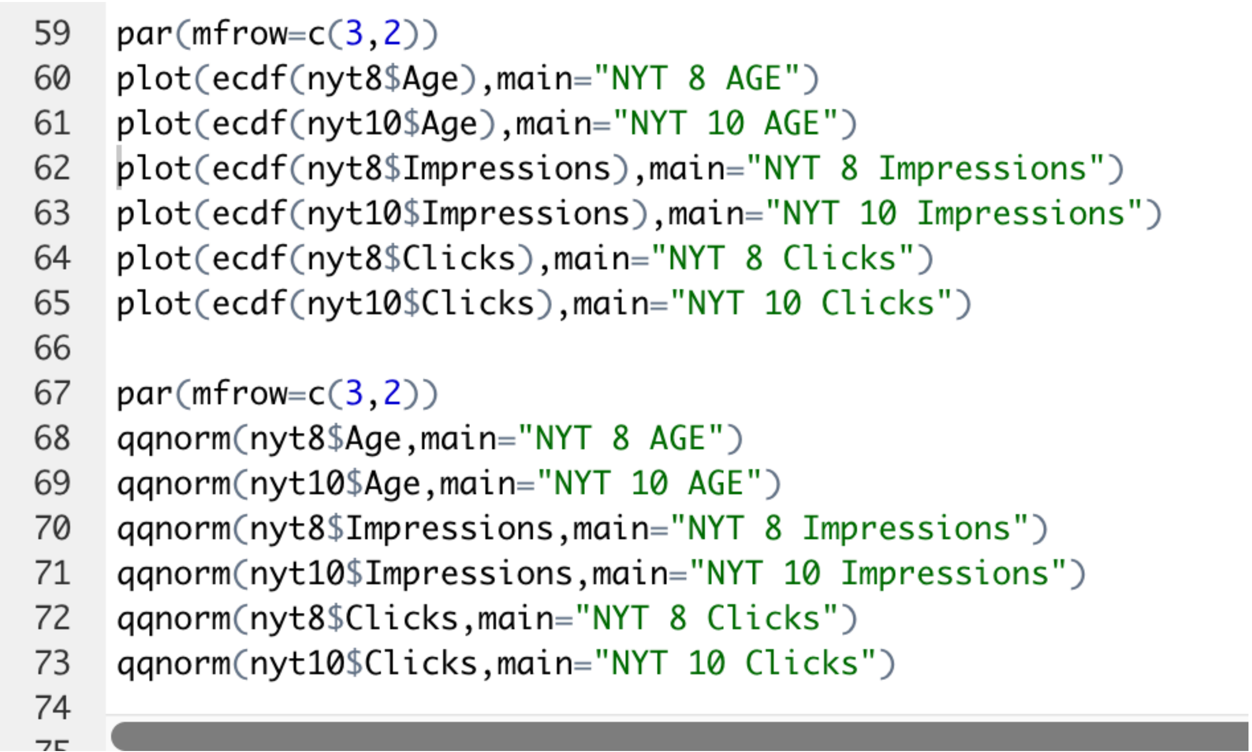

NYT 8 Impressions
Frequency
0
0
5
10
15
nyt8$Impressions


NYT 10 Impressions
Frequency
0
0
5
10
15
20
nyt10$Impressions

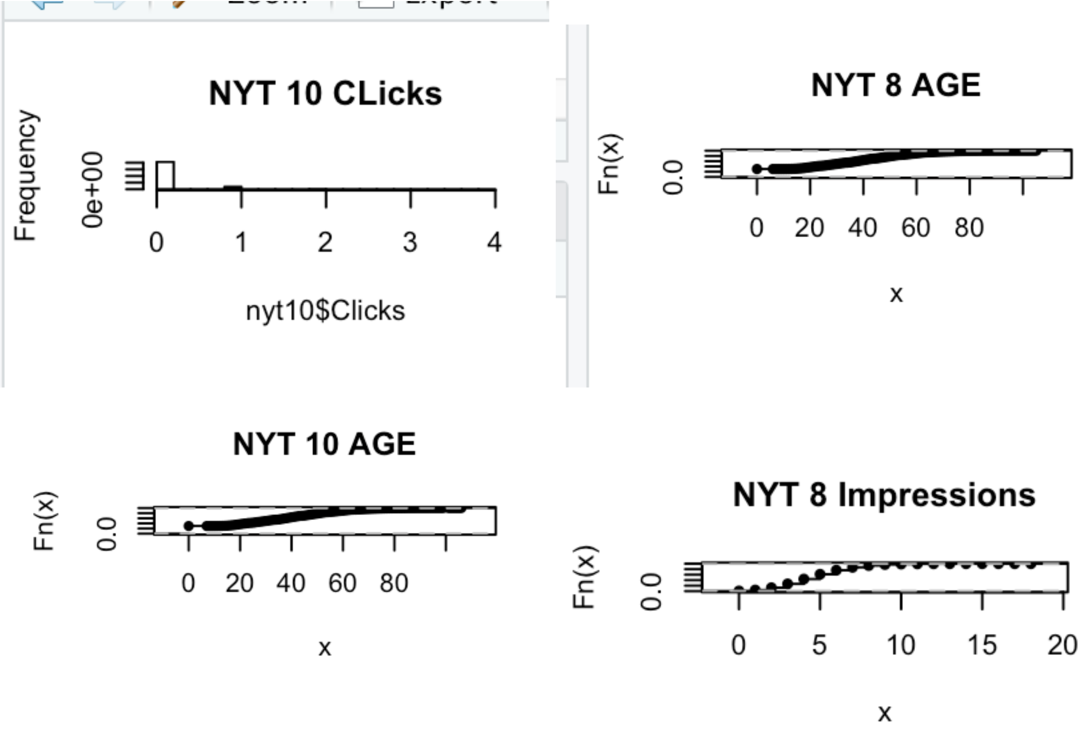

NYT 8 Clicks
Frequency
0e+00
0
1
2
3
4
nyt8$Clicks

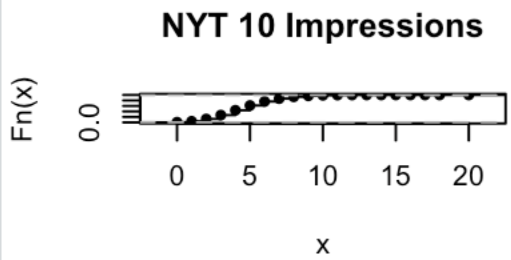

NYT 10 CLicks
Frequency
0e+00
0
1
2
3
4
nyt10$Clicks

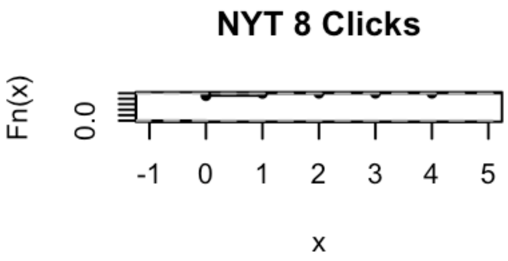
Q1c

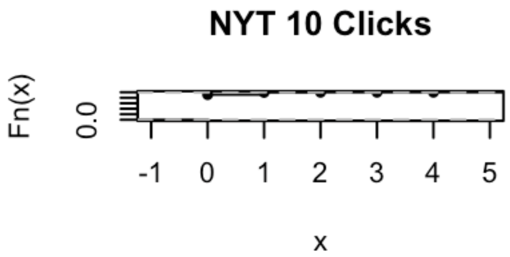


```
par(mfrow=c(3,2))
plot(ecdf(nyt8$Age),main="NYT 8 AGE")
plot(ecdf(nyt10$Age),main="NYT 10 AGE")
plot(ecdf(nyt8$Impressions),main="NYT 8 Impressions")
plot(ecdf(nyt10$Impressions),main="NYT 10 Impressions")
plot(ecdf(nyt8$Clicks),main="NYT 8 Clicks")
plot(ecdf(nyt10$Clicks),main="NYT 10 Clicks")

par(mfrow=c(3,2))
qqnorm(nyt8$Age,main="NYT 8 AGE")
qqnorm(nyt10$Age,main="NYT 10 AGE")
qqnorm(nyt8$Impressions,main="NYT 8 Impressions")
qqnorm(nyt10$Impressions,main="NYT 10 Impressions")
qqnorm(nyt8$Clicks,main="NYT 8 Clicks")
qqnorm(nyt10$Clicks,main="NYT 10 Clicks")

```

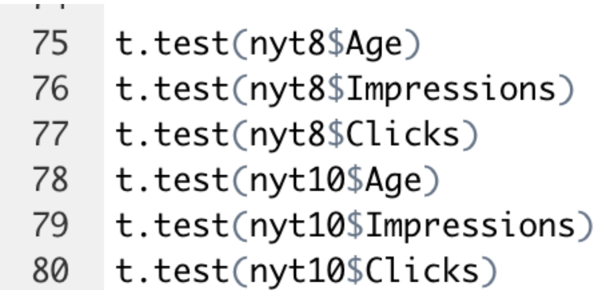








Q1d



```
t.test(nyt8$Age)
t.test(nyt8$Impressions)
t.test(nyt8$Clicks)
t.test(nyt10$Age)
t.test(nyt10$Impressions)
t.test(nyt10$Clicks)
```