Code:

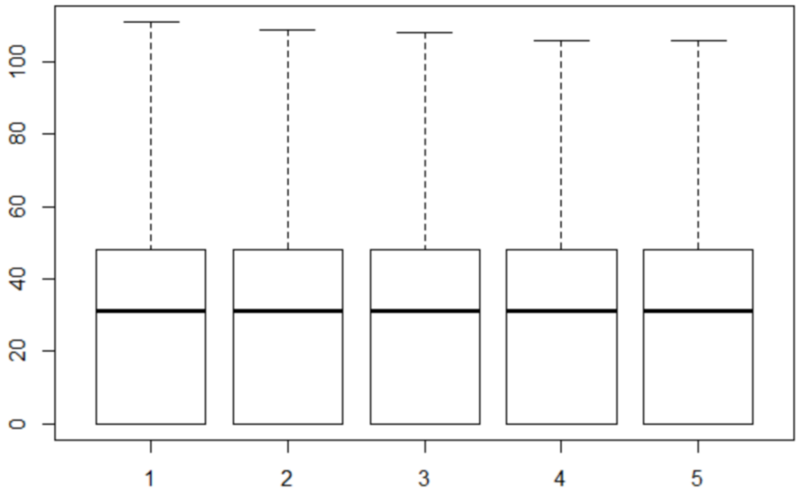
```
ny2 <- read.csv('C:/Users/dingj/Desktop/data_analytics/hw2/nyt2.csv')
attach(ny2)
ny3 <- read.csv('C:/Users/dingj/Desktop/data_analytics/hw2/nyt3.csv')
attach(ny3)
ny4 <- read.csv('C:/Users/dingj/Desktop/data_analytics/hw2/nyt4.csv')
attach(ny4)
ny5 <- read.csv('C:/Users/dingj/Desktop/data_analytics/hw2/nyt5.csv')
attach(ny5)
ny6 <- read.csv('C:/Users/dingj/Desktop/data_analytics/hw2/nyt6.csv')
attach(ny6)

boxplot(ny2$Age,ny3$Age,ny4$Age,ny5$Age,ny6$Age)
boxplot(ny2$Impressions,ny3$Impressions,ny4$Impressions,ny5$Impressions,ny6$Impressions)

hist(ny2$Age)
hist(ny3$Age)
hist(ny4$Age)
hist(ny5$Age)
hist(ny6$Age)
hist(ny2$Impressions)
hist(ny3$Impressions)
hist(ny4$Impressions)
hist(ny5$Impressions)
hist(ny6$Impressions)

plot(ecdf(ny2$Age),verticals=TRUE)
plot(ecdf(ny3$Age),verticals=TRUE)
plot(ecdf(ny4$Age),verticals=TRUE)
plot(ecdf(ny5$Age),verticals=TRUE)
plot(ecdf(ny6$Age),verticals=TRUE)
plot(ecdf(ny2$Impressions),verticals=TRUE)
plot(ecdf(ny3$Impressions),verticals=TRUE)
plot(ecdf(ny4$Impressions),verticals=TRUE)
plot(ecdf(ny5$Impressions),verticals=TRUE)
plot(ecdf(ny6$Impressions),verticals=TRUE)

shapiro.test(ny2$Age[1:5000])
shapiro.test(ny3$Age[1:5000])
shapiro.test(ny4$Age[1:5000])
shapiro.test(ny5$Age[1:5000])
shapiro.test(ny6$Age[1:5000])

shapiro.test(ny2$Impressions[1:5000])
shapiro.test(ny3$Impressions[1:5000])
shapiro.test(ny4$Impressions[1:5000])
shapiro.test(ny5$Impressions[1:5000])
shapiro.test(ny6$Impressions[1:5000])

boxplot(ny2$Clicks,ny3$Clicks)



hist(ny2$Clicks)
hist(ny3$Clicks)

plot(ecdf(ny2$Clicks),verticals=TRUE)
plot(ecdf(ny3$Clicks),verticals=TRUE)
```

shapiro.test(ny2$Clicks[1:5000])

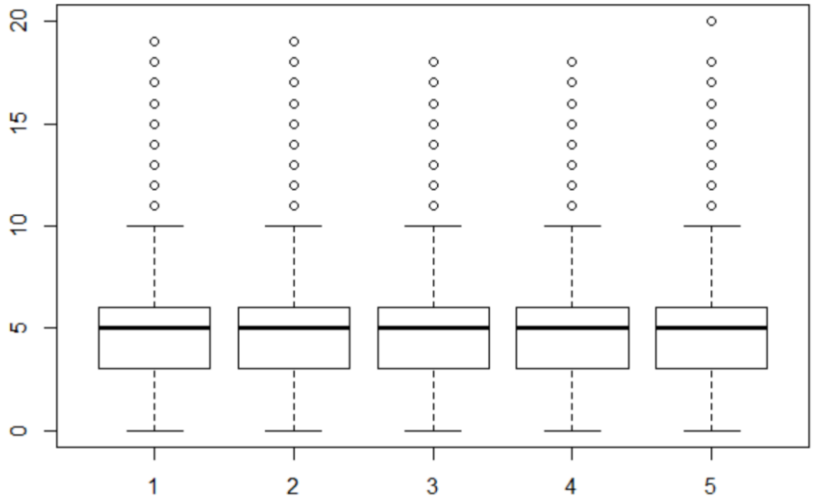
shapiro.test(ny3$Clicks[1:5000])

Problem1:

a.The box plot for Ages

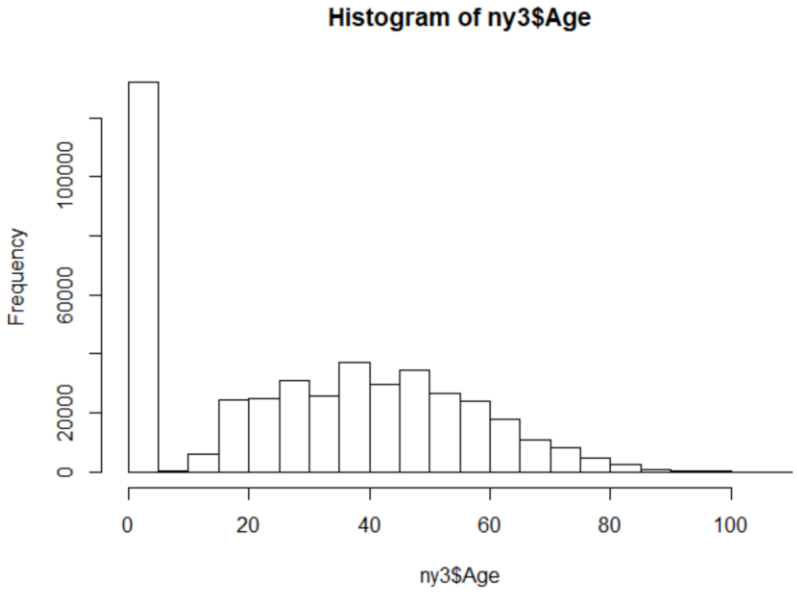


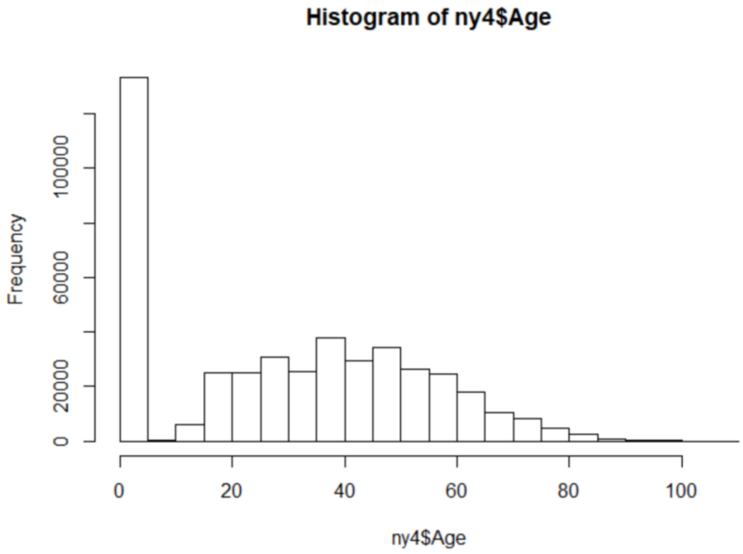
The box plot for Impressions

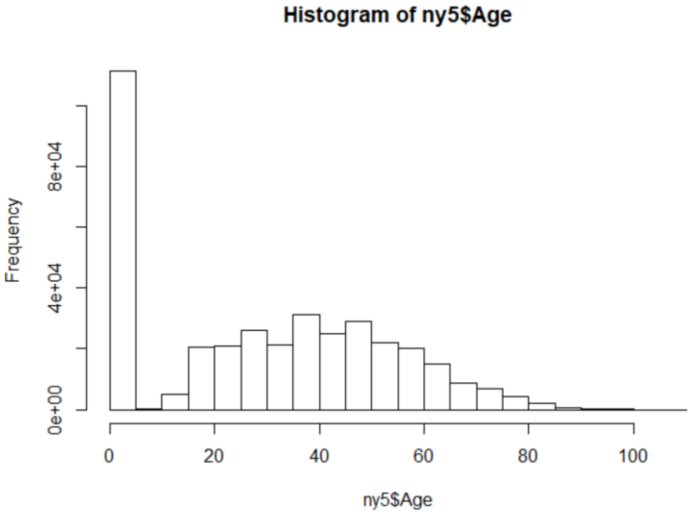


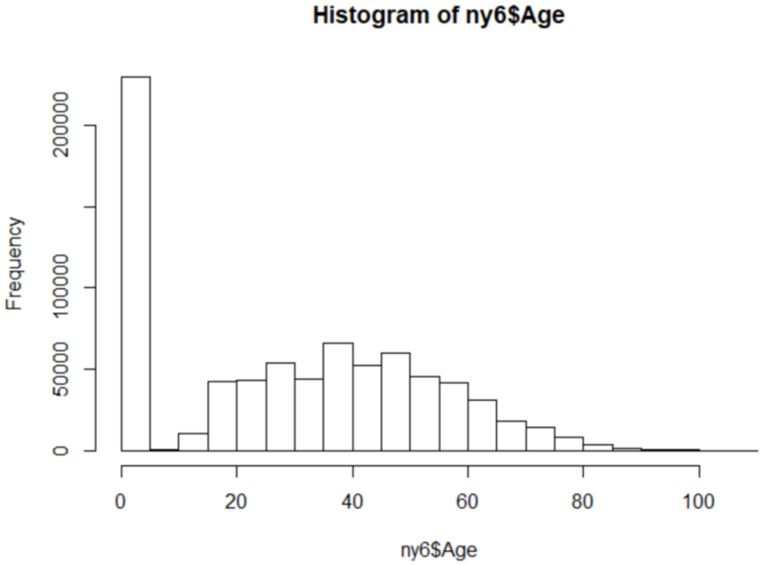
From the boxplot for ages, the average ages of the New York times readers are around 30 years old and average impressions are 5. The average ages and impressions are almost the same across different datasets. The max age for New York times readers are over 100 years old.

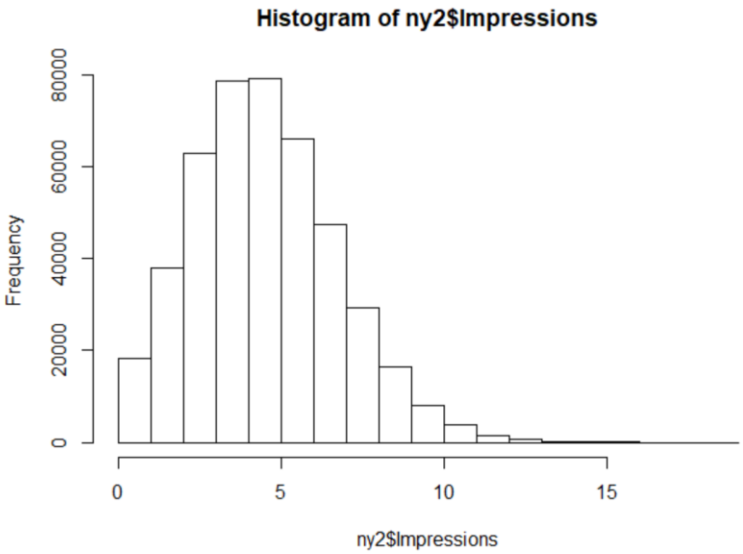
b.

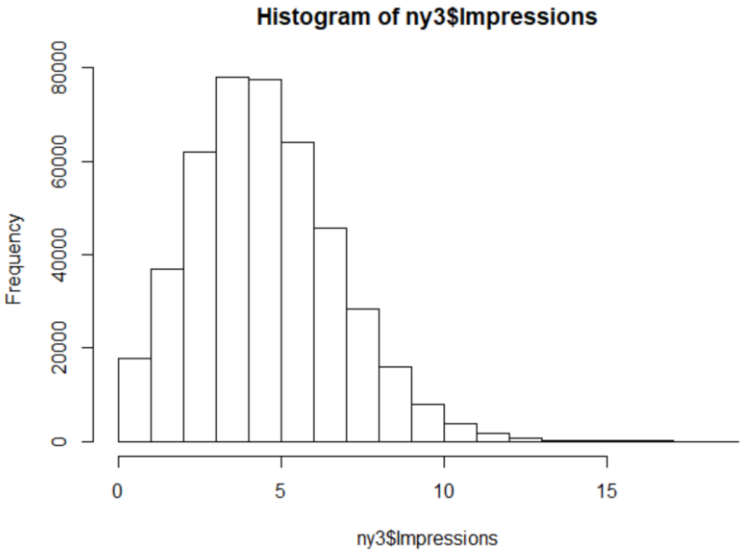


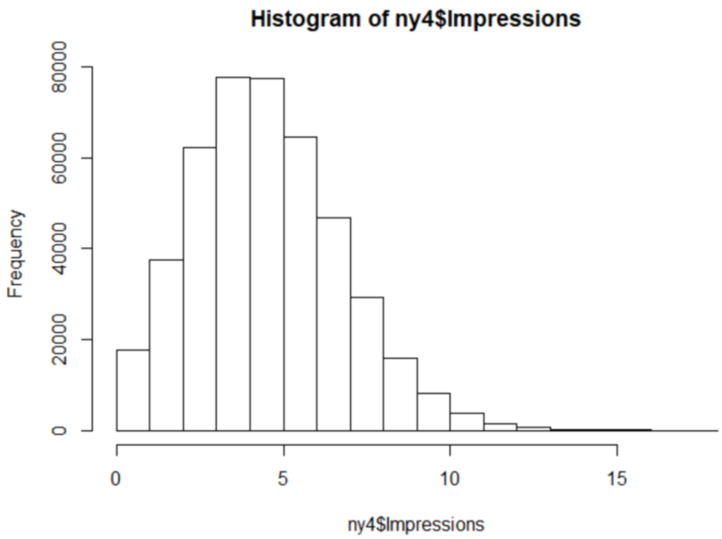


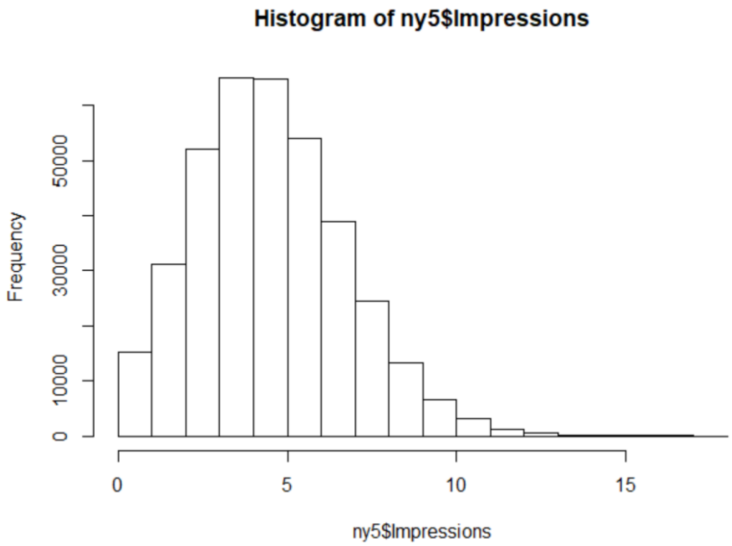

Histogram of ny5$Age
Frequency
0e+00
4e+04
8e+04
0
20
40
60
80
100
ny5$Age

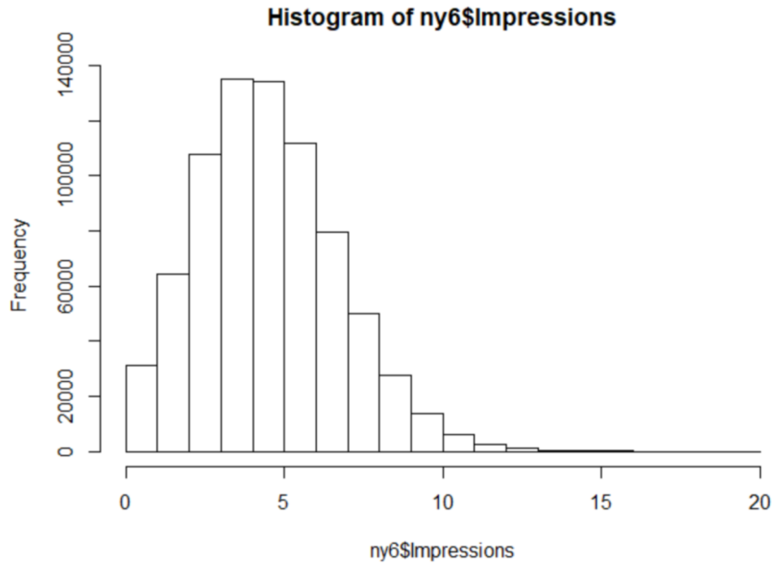

Histogram of ny6$Age
Frequency
0
50000
100000
200000
0
20
40
60
80
100
ny6$Age


Histogram of ny2$Impressions
Frequency
0
20000
40000
60000
80000
0
5
10
15
ny2$Impressions


Histogram of ny3$Impressions
Frequency
0
20000
40000
60000
80000
0
5
10
15
ny3$Impressions

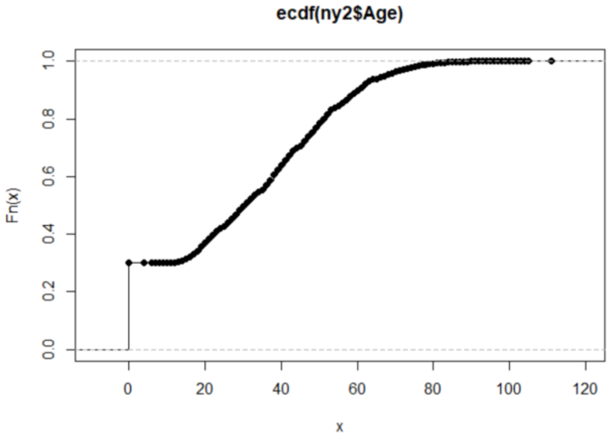

Histogram of ny4$Impressions
Frequency
0
20000
40000
60000
80000
0
5
10
15
ny4$Impressions

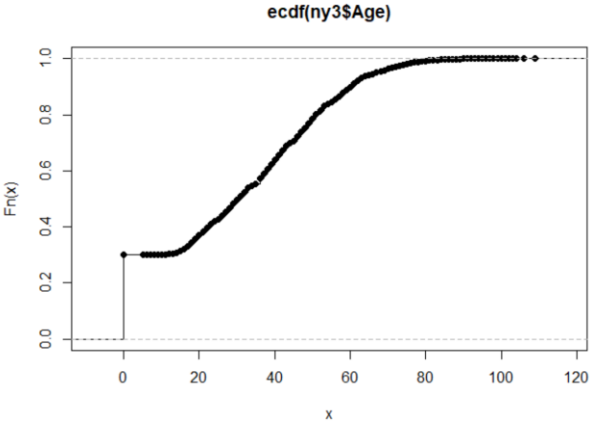

Histogram of ny5$Impressions
Frequency
0
10000
30000
50000
0
5
10
15
ny5$Impressions

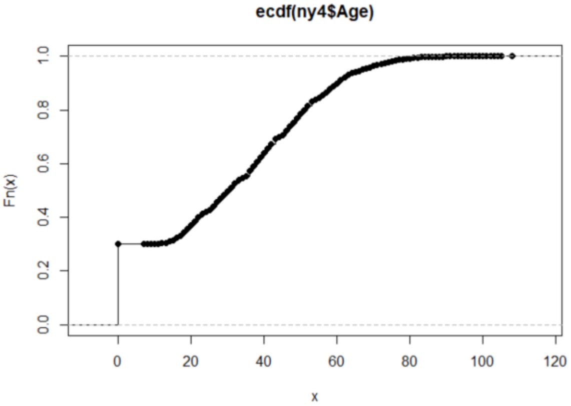


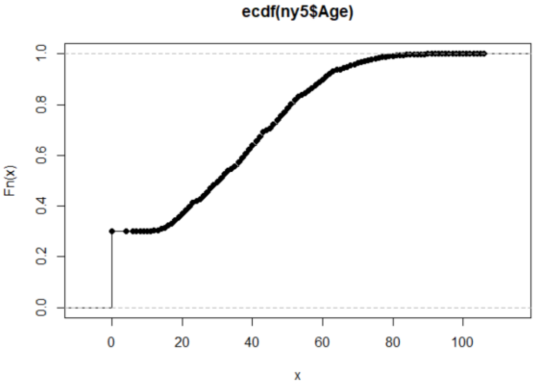
From the histograms, most New York Times readers ages from 20- 60. Most new York times readers have 3 – 6 impressions. The distribution of both age and impression are normal distribution.

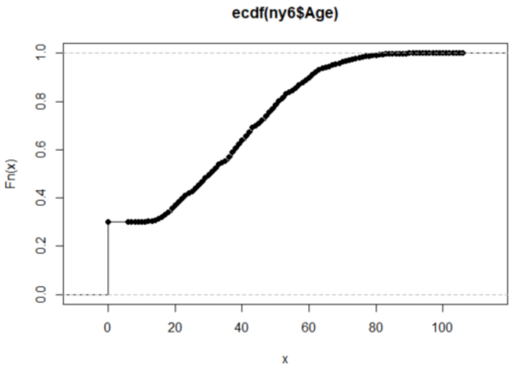
c.

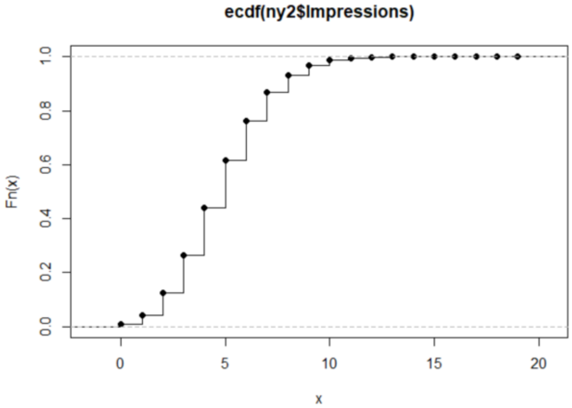


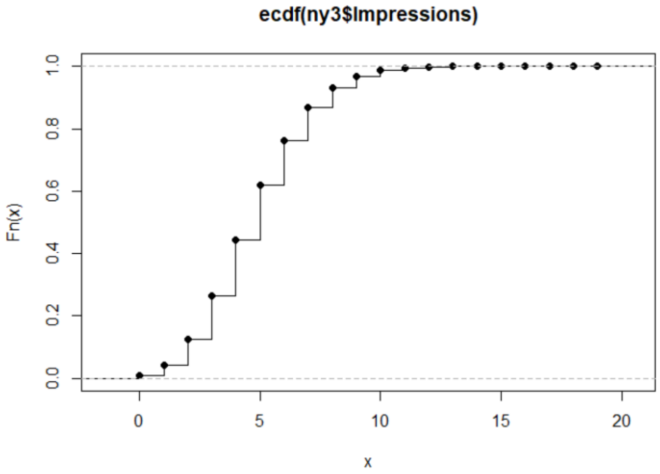

ecdf(ny3$Age)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
x

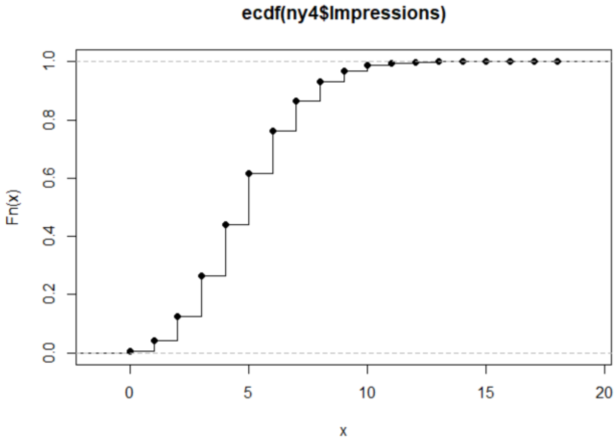

ecdf(ny4$Age)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
x

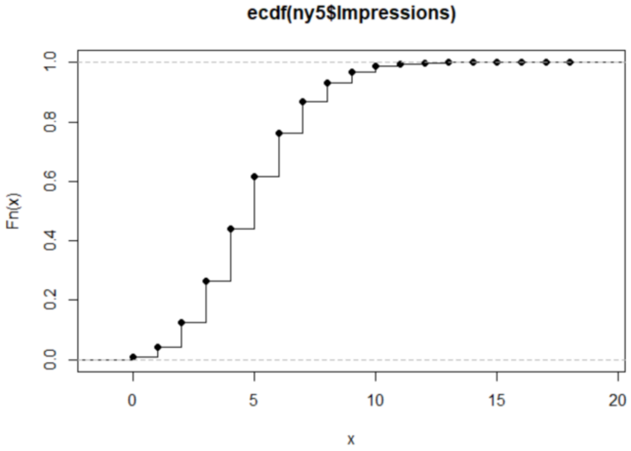

ecdf(ny5$Age)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
x

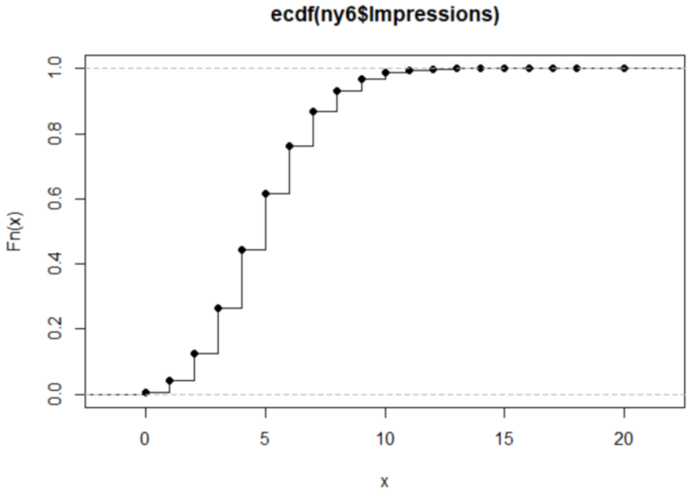

ecdf(ny6$Age)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
x


ecdf(ny2$Impressions)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
x


ecdf(ny3$Impressions)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
x

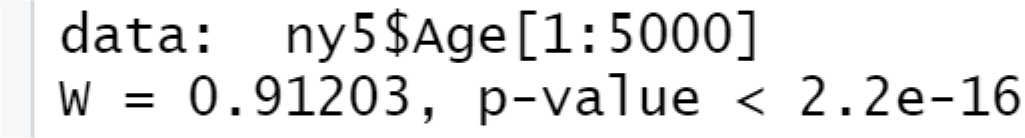






From the plot for ECDF of ages, we can find out that for most datasets, there are around 30% of people missing ages(age 0). There are also 60% of people less than age 40. From the plot for ECDF of impressions, we can find out that for most datasets, 50% of people have less than 5 impressions and less than 5 % of people have more than 10 impressions.

d.

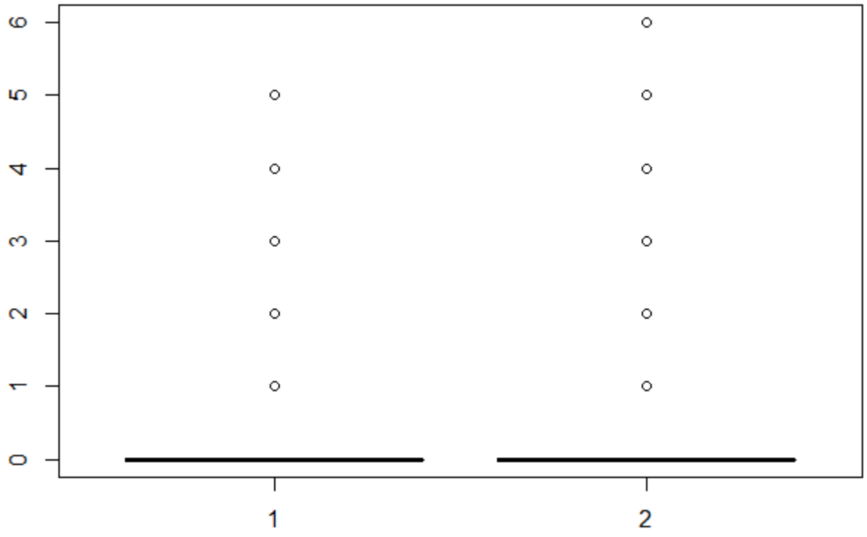


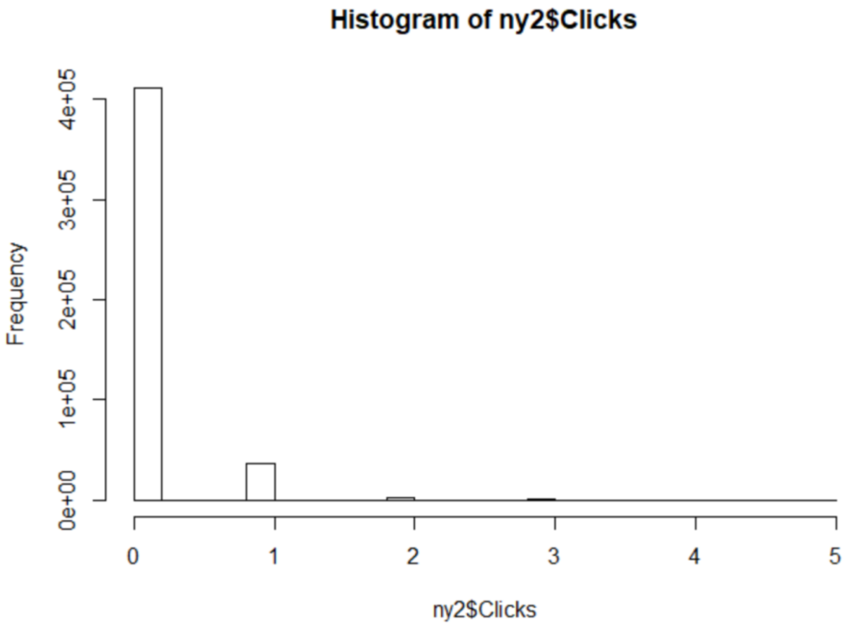
```
data:  ny5$Age[1:5000]
W = 0.91203, p-value < 2.2e-16
```

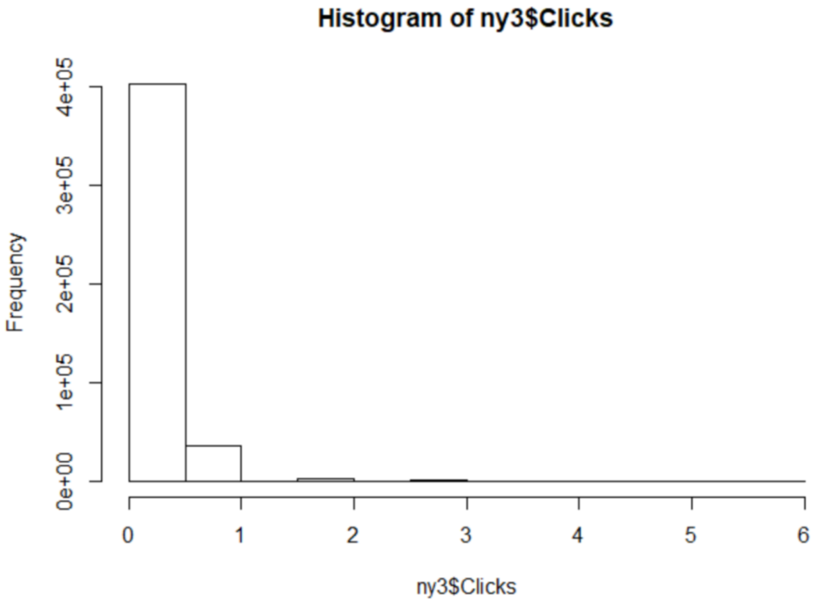
For Age variable in all 5 datasets, p-value is less than 0.05 which means the null hypothesis is invalid.

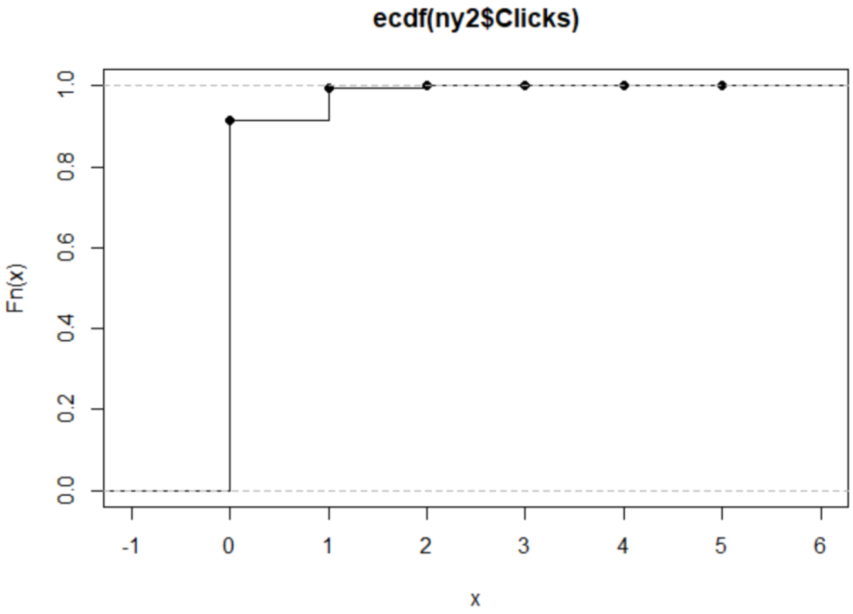
For Impression variable in all 5 datasets, p-value is less than 0.05 which means the null hypothesis is invalid.

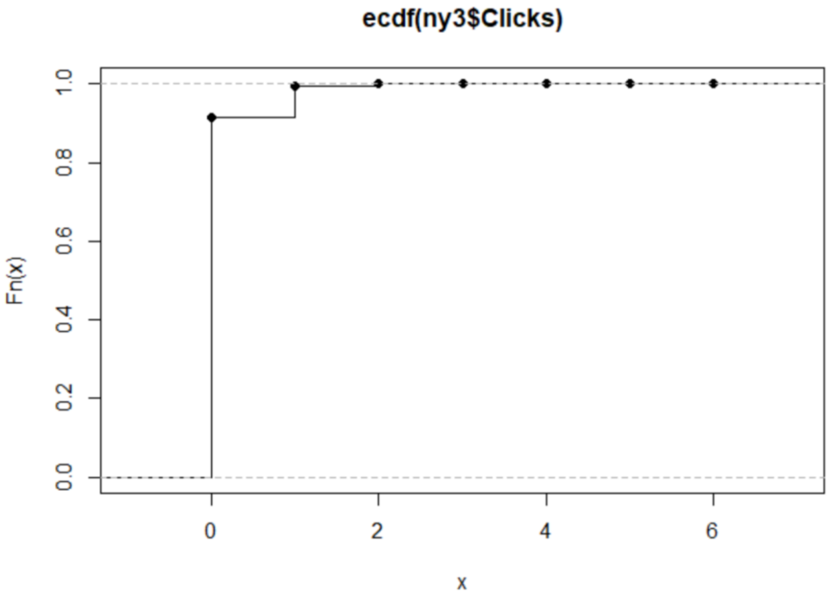
Problem 2.




Histogram of ny2$Clicks
Frequency
0e+00
1e+05
2e+05
3e+05
4e+05
0
1
2
3
4
5
ny2$Clicks


Histogram of ny3$Clicks
Frequency
0e+00
1e+05
2e+05
3e+05
4e+05
0
1
2
3
4
5
6
ny3$Clicks





For the clicks variable, most people have 0 clicks. There are almost less than 1% of people who have more than 1 clicks. The max number of clicks is 6. The distribution of the clicks is right skewed distribution.