Assignment 3

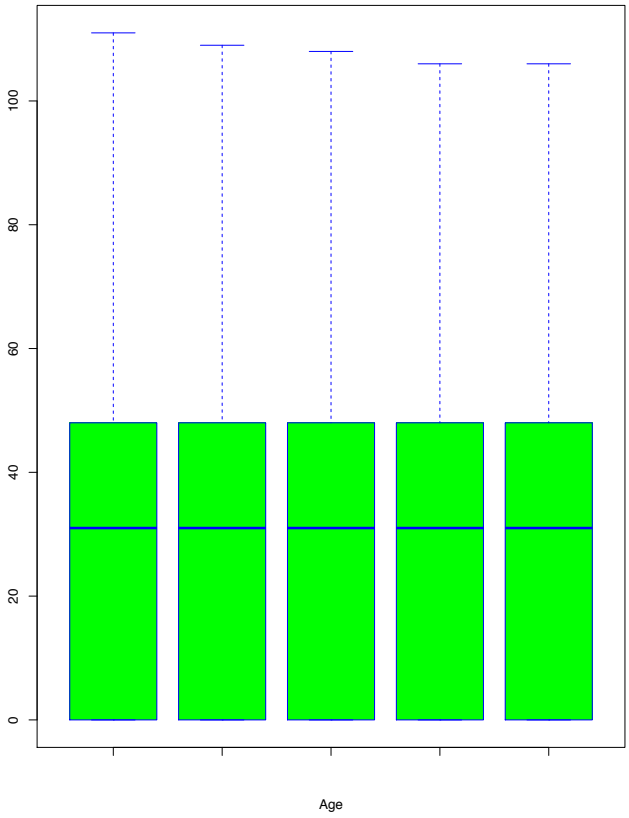
a). Create boxplots for all 5 datasets for each of two key variables (you choose the variables), i.e. two figures (one for each variable) with 5 boxplots (for the 5 different datasets) in each. Describe/summarize the distributions.

(i) Datasets: nyt2, nyt3, nyt4, nyt5, nyt6

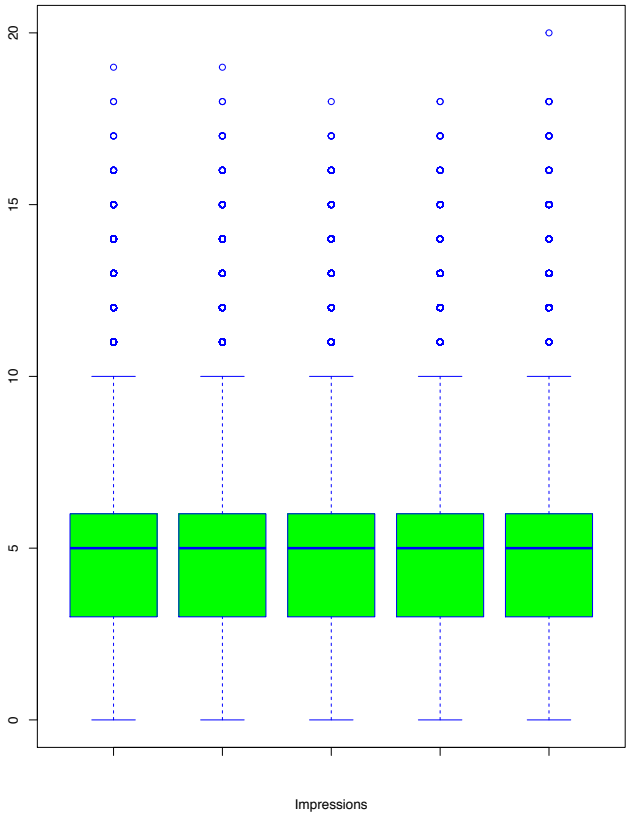
> boxplot(nyt2$Age, nyt3$Age, nyt4$Age, nyt5$Age, nyt6$Age, border = "blue", col = "green", xlab = “Age")

> boxplot(nyt2$Impressions, nyt3$Impressions, nyt4$Impressions, nyt5$Impressions, nyt6$Impressions, border = "blue", col = "green", xlab = “Impressions")

Variable: Age



Variable: Impressions



From the box-plots of Age variable of the five New York times dataset, the Age variable range from 0 to about 90. The median Age is around the mark of 30 years for all the five datasets. The interquartile range from all the five datasets for Age variable is around 50 years. The minimum value for Age also coincides with the Q1 mark.

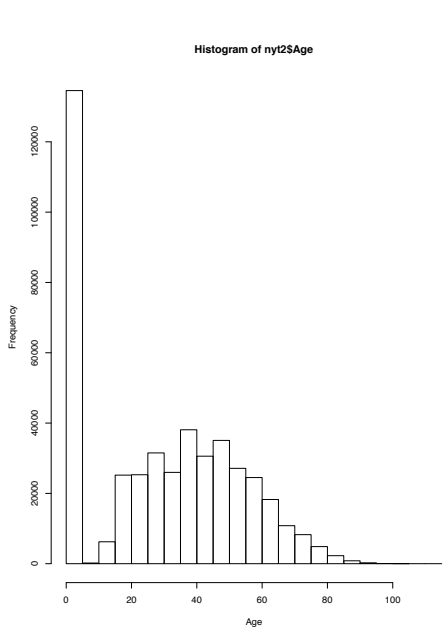
From the box-plots of Impressions variable it is evident that the Impressions variable ranges from 0 to 10 for all the five datasets(nyt2, nyt3, nyt4, nyt5, nyt6). With the maximum value of 10 and the minimum value of 0 the datasets have a median of 5. The interquartile range is around 3. The box-plot for “Impressions” variable is almost the same for all the five datasets taken above.

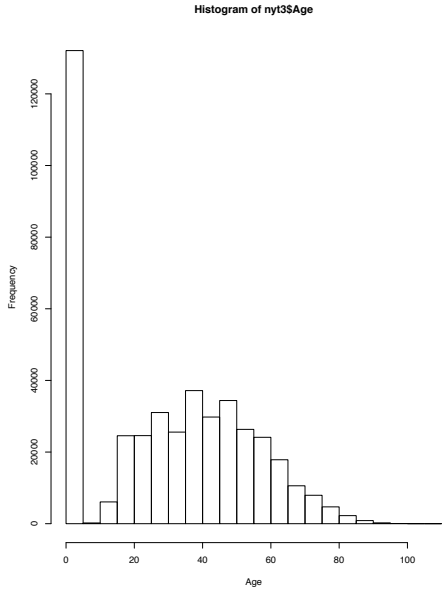
b). Create histograms for all 5 datasets for two key variables – can be the same variables in 1a or different (you choose the histogram bin width). Describe the distributions in terms of known parametric distributions and similarities/ differences among them.

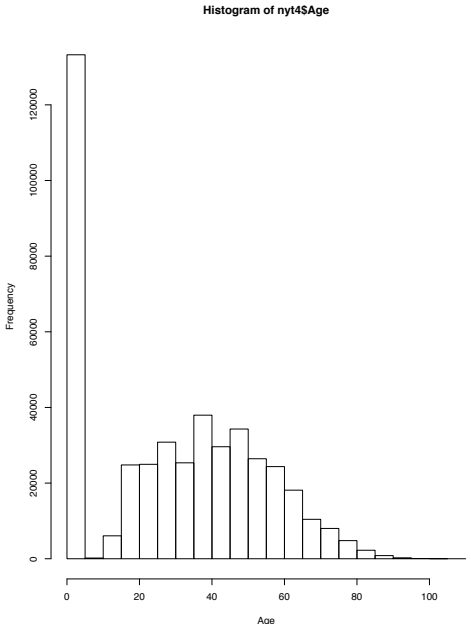
(i) Datasets: nyt2, nyt3, nyt4, nyt5, nyt6

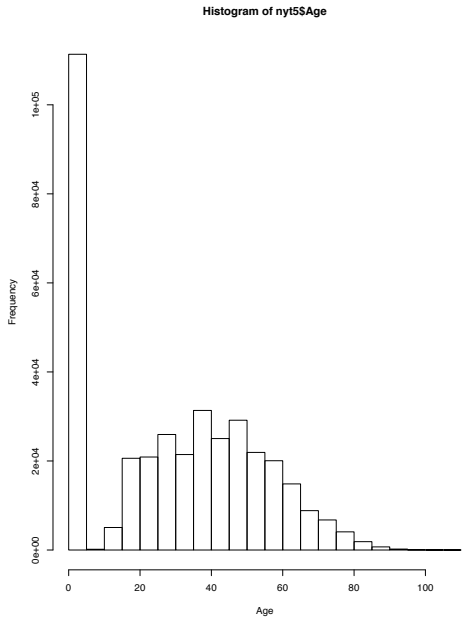
Variable: Age

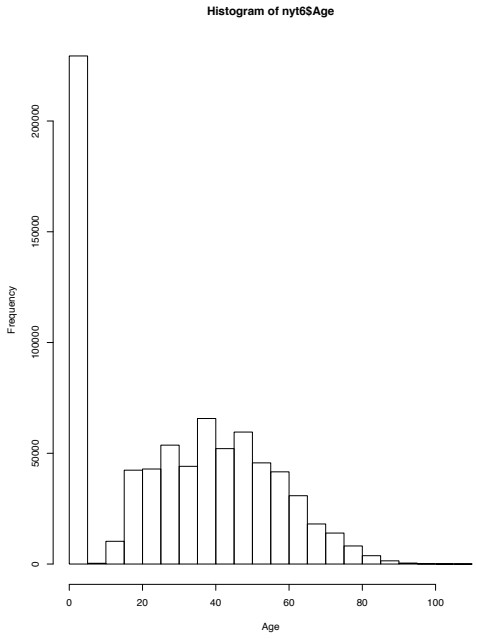
```
> hist(nyt2$Age, xlab="Age")
> hist(nyt3$Age, xlab="Age")
> hist(nyt4$Age, xlab="Age")
> hist(nyt5$Age, xlab="Age")
> hist(nyt6$Age, xlab="Age")
```







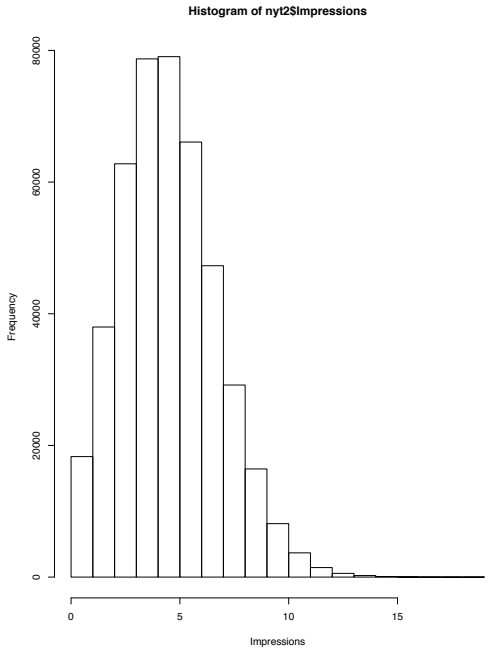


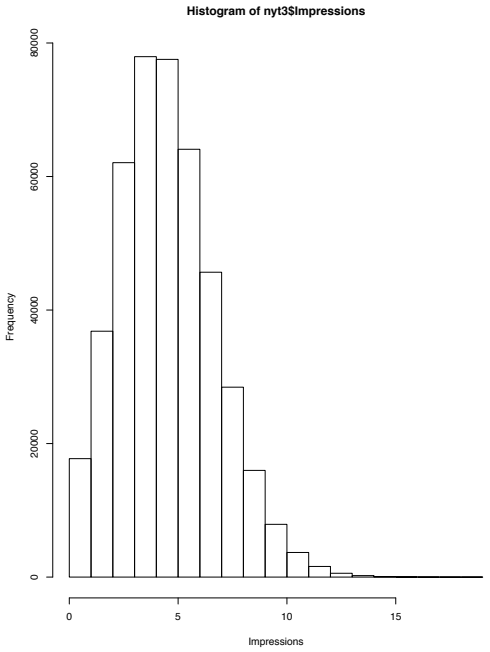


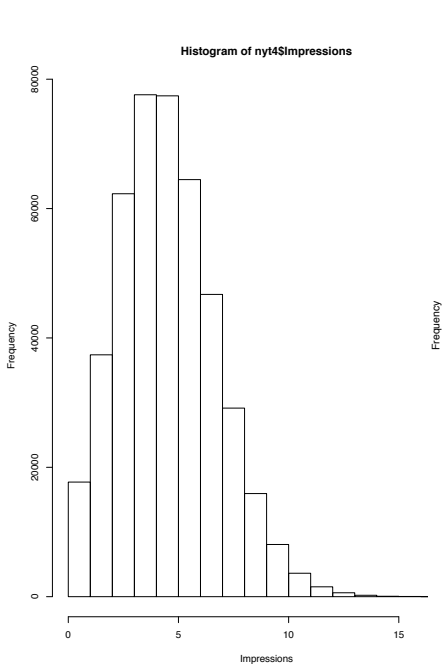
From the distributions of variable Age for the above five datasets we can see that the variable age first increases to a certain point and the decreases. The frequency of people with age of around 40 years is the highest. There are a large number of Age with values 0 in all five of the datasets.

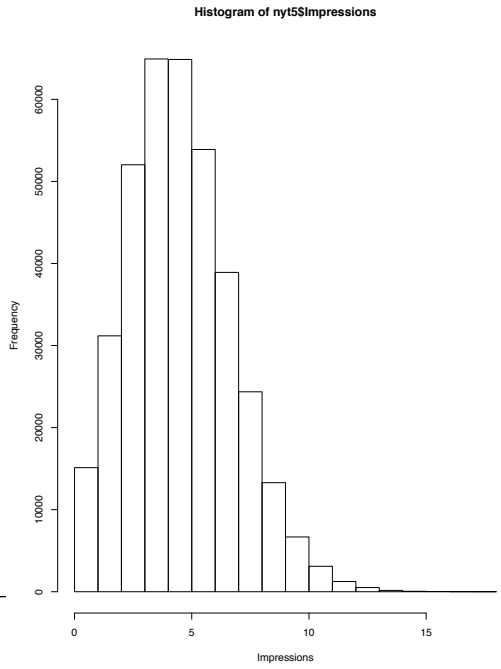
(ii) Datasets: nyt2, nyt3, nyt4, nyt5, nyt6
Variable: Impressions

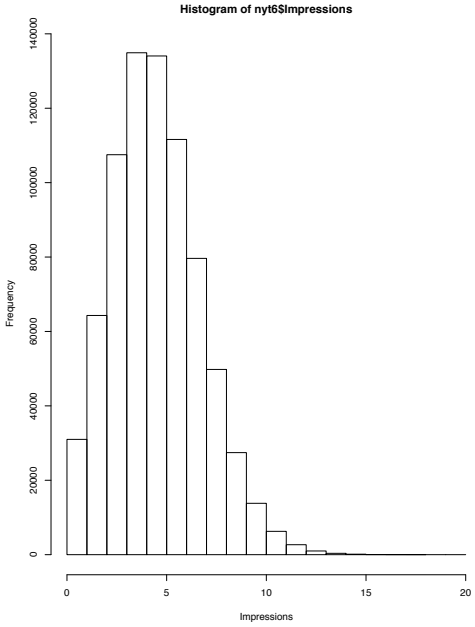
```
> hist(nyt2$Impressions,xlab="Impressions")
> hist(nyt3$Impressions,xlab="Impressions")
> hist(nyt4$Impressions,xlab="Impressions")
> hist(nyt5$Impressions,xlab="Impressions")
> hist(nyt6$Impressions,xlab="Impressions")
```





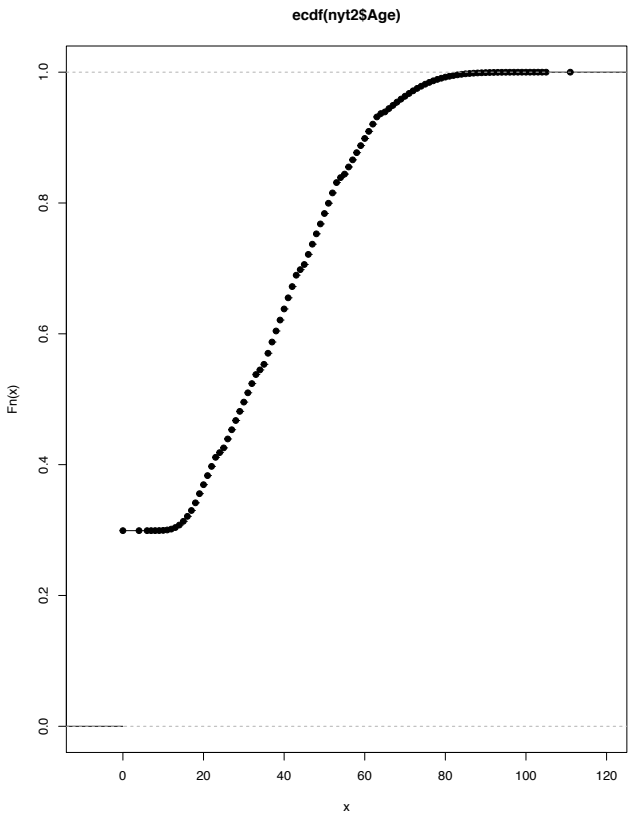


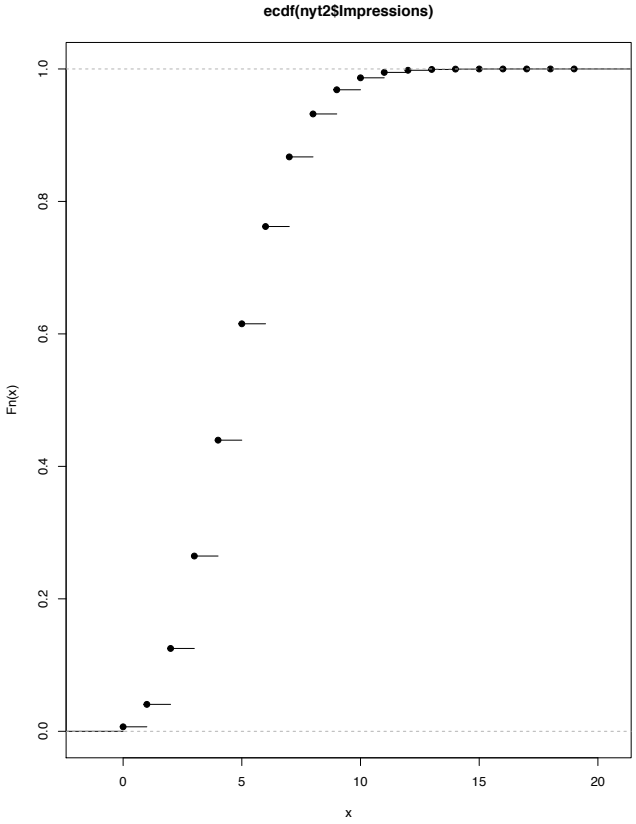


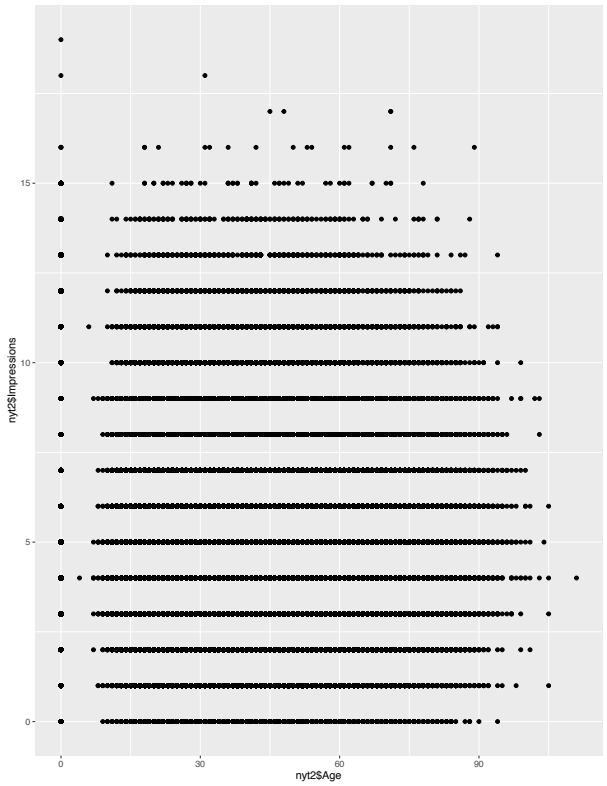


The variable Impressions for all five of the datasets considered above ranges from 0 to 13. The maxima in all five of the datasets occurs around the 4 to 5 mark. The frequency of 4 is the highest in general in all five of the datasets. The distribution of the Impressions variable forms a bell shaped sort of distribution with the value increasing up to a certain limit and then decreases back to zero.

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.







The e.c.d.f. (empirical cumulative distribution function) Fn is a step function with jumps i/n at observation values, where i is the number of tied observations at that value. Missing values are ignored. For observations x=(x1,x2, … xn), Fn is the fraction of observations less or equal to t,

i.e., $Fn(t)=\#\{xi \leq t\} /n=1nn\sum i=11[xi \leq t]$.

Reference: https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/ecdf

So, we end up plotting the function fn , plot a feature of your data in order from least to greatest and see the whole feature as if is distributed across the data set. This suggests that there are lesser number of tied observations in the variable.

So, we can see that in case of Age variable, it is mostly continuous with small steps.

For the impressions variable ecdf comes out to be somewhat non-continuous function with large steps. This suggests us that there are large number of tied observations in the Impressions variable.

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.

(i) cor.test(nyt2$Age, nyt2$Impressions)
data: nyt2$Age and nyt2$Impressions
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

(ii) cor.test(nyt3$Age, nyt3$Impressions)
data: nyt3$Age and nyt3$Impressions
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

(iii) cor.test(nyt4$Age, nyt4$Impressions)
data: nyt4$Age and nyt4$Impressions
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

(iv) cor.test(nyt5$Age, nyt5$Impressions)
data: nyt5$Age and nyt5$Impressions

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

(v) cor.test(nyt6$Age, nyt6$Impressions)
data: nyt6$Age and nyt6$Impressions
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

For all the five datasets taken, the true correlation between the variables Age and Impressions is not equal to zero. All the five datasets have a large p-value (> 0.05) which indicates weak evidence against the null hypothesis, so we cannot reject the null hypothesis. The p-value is the least for nyt4 dataset and highest for nyt2 dataset. The lower the p-value, the lesser the chance that this much correlation happened as a matter of chance. The p-value should be very low for us in order for us to trust the calculated metric. The lower the p-value (< 0.01 or 0.05 typically), stronger is the significance of the relationship. Here, for all the five datasets the p-value is on the higher side. Higher positive correlation means higher relation between the variables of the data. Negative correlation means that an increase in one variable reliably predicts a decrease in the other variable. Here the two variable Age and Impressions in all the five datasets are very slightly correlated as their correlation value is close to zero.

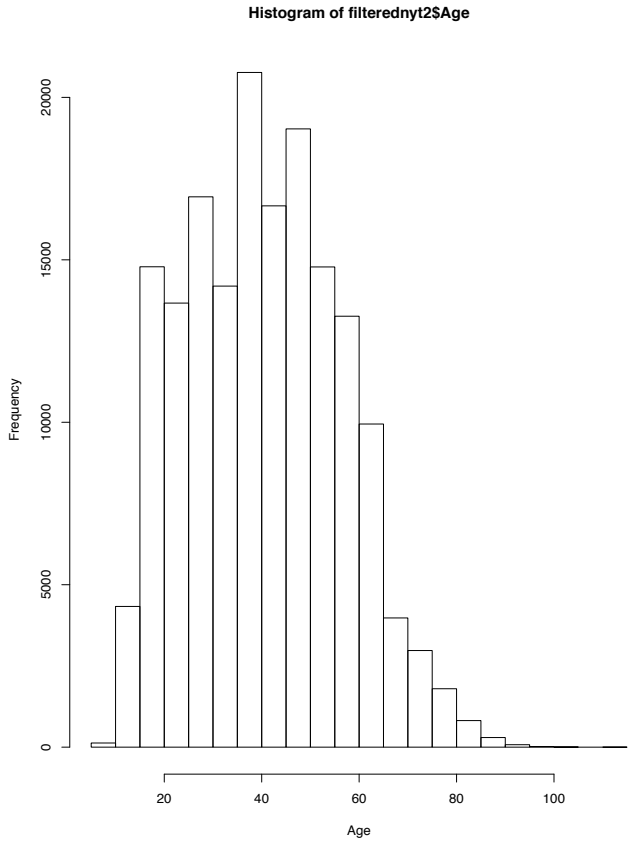
2. Filter the distributions you explored in Q1 using one or more of the other variables for only 2 (not 5) of the nyt datasets. Repeat Q1b, Q1c and Q1d and draw any conclusions from this study.

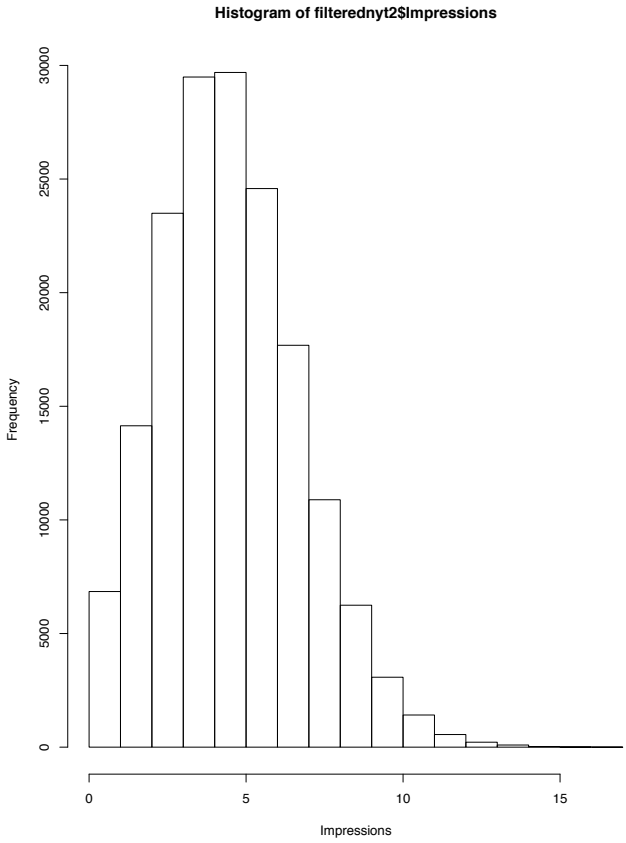
Dataset: nyt2
I have filtered nyt2 on the basis fo Gender such that Gender ==1
> filterednyt2 = subset(nyt2, Gender == 1)
> filterednyt2

	Age	Gender	Impressions	Clicks	Signed_In
1	48	1	3	0	1
3	15	1	4	0	1
9	24	1	2	0	1
11	31	1	5	0	1
13	56	1	5	0	1
14	52	1	6	0	1
15	48	1	2	0	1
17	47	1	4	0	1
25	25	1	8	0	1
28	42	1	3	0	1
29	30	1	7	0	1
35	47	1	4	0	1
37	39	1	6	0	1
38	24	1	6	0	1
40	55	1	2	0	1

Q1b) Plotting the histograms for the two filtered variables

> hist(filterednyt2$Age,xlab="Age")
> hist(filterednyt2$Impressions,xlab="Impressions")



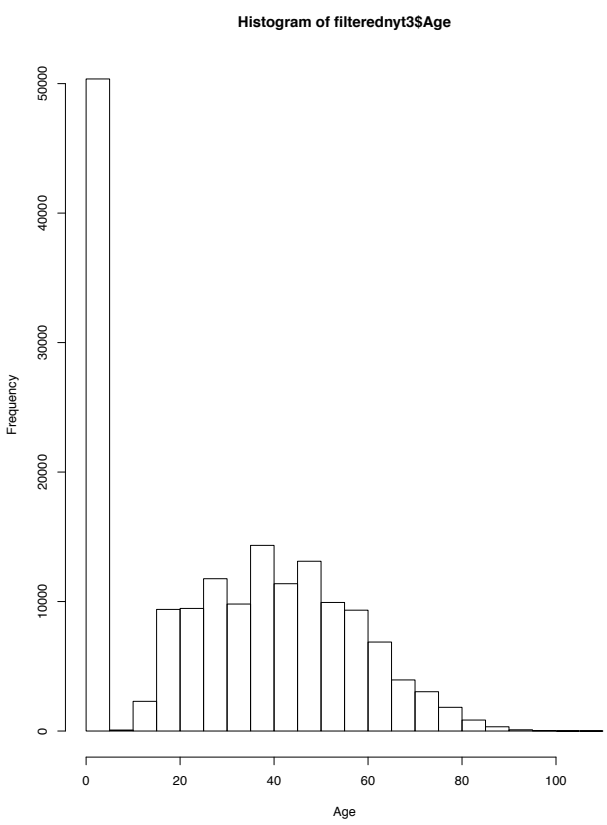


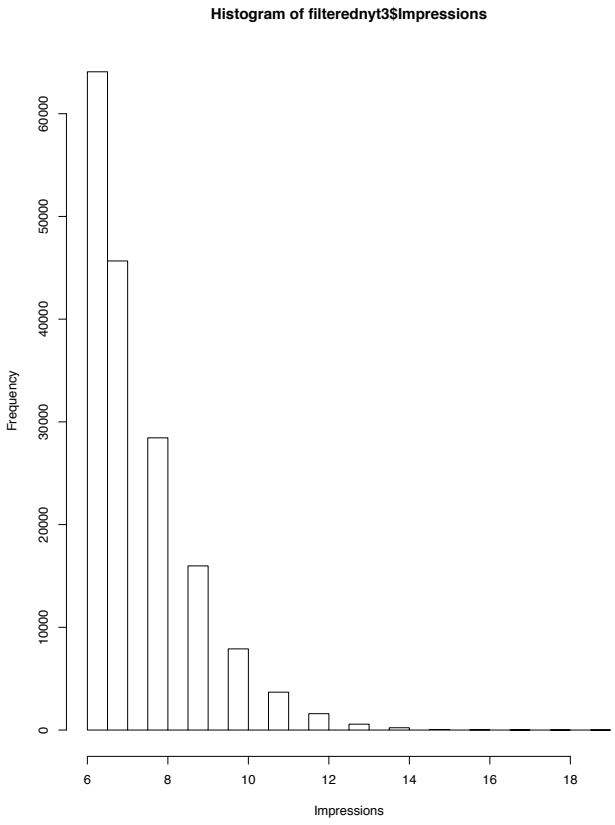
Distributions of the Age and Impressions variable after filtering out all the observations with Gender = 1(considering 1 as Male).
From the histogram of Age variable we can observe that the number of Males around the age of 40 years are the highest. The Age variable ranges from around 10 to 80 in the nyt2 dataset. The frequency of Males with age around the 40 year mark is around 210000.
From the histogram of Impressions variable we can observe that the number of Males with Impressions value of 5 are the highest. The Age variable ranges from around 0 to 18 in the nyt2 dataset. The frequency of Males with Impression variable around the 5 mark is around 30,000.

Dataset: nyt3
I have filtered nyt3 on the basis of Impressions variable such that Impressions >5
> filterednyt3 = subset(nyt3, Impressions >5)
> filterednyt3

	Age	Gender	Impressions	Clicks	Signed_In
2	75	0	9	0	1
6	50	0	8	0	1
7	23	1	6	0	1
10	20	0	6	0	1
16	71	0	8	0	1
20	51	1	6	0	1
21	0	0	8	0	0
22	0	0	9	0	0
30	0	0	6	0	0
48	21	0	8	0	1
50	45	0	9	0	1

> hist(filterednyt3$Age,xlab="Age")
> hist(filterednyt2$Impressions,xlab="Impressions")





Distributions of the Age and Impressions variable after filtering out all the observations with Impressions>5

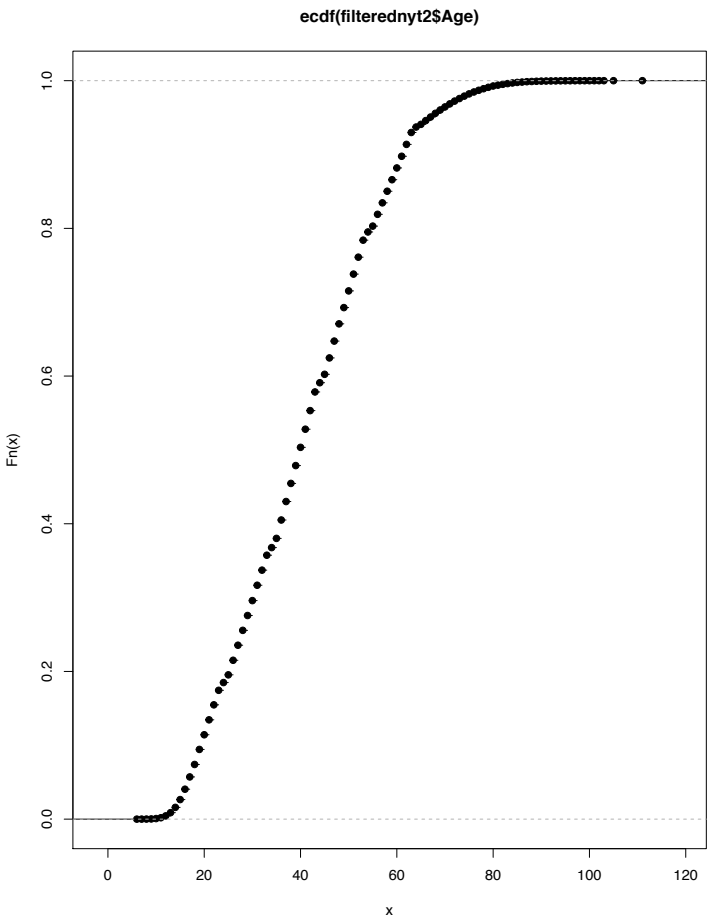
From the histogram of the Age variable we can see that the number of people with Age values 0 is still the highest. Apart form the value 0, the frequency is highest for people around 40 years in age. The frequency of people with Age around 80-90 years mark is the lowest. The Age variable first increases upto 40 and then decreases in frequency for the time it reaches 90 years in Age.

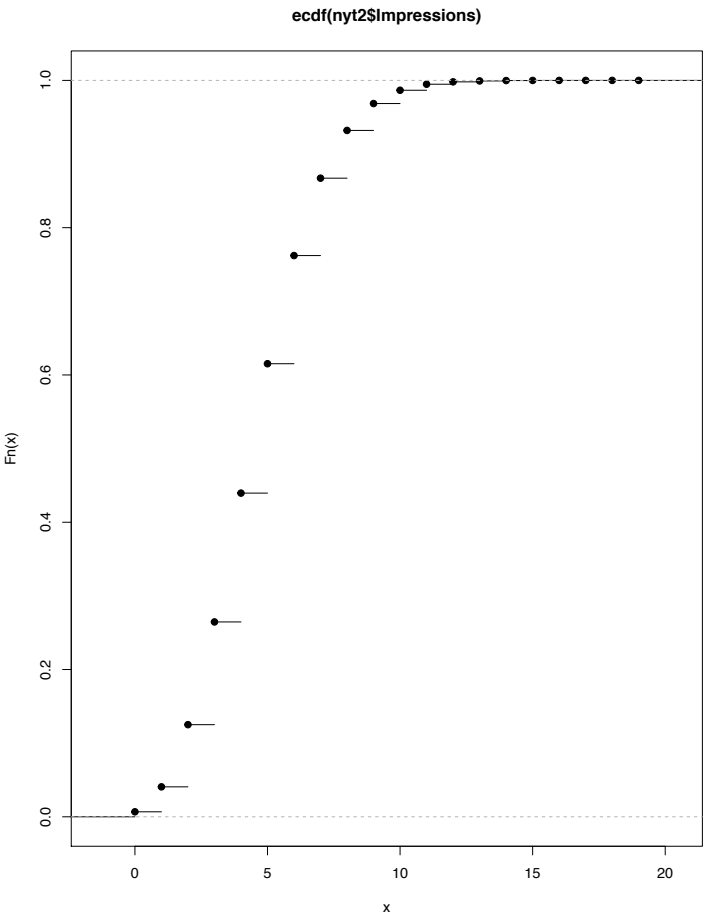
From the histogram of Impressions variable we can see that the number of people with Impression value of 6 are the highest. The frequency of people decreases as the value of Impression variable increases. The highest value for Impression variable is around 15.

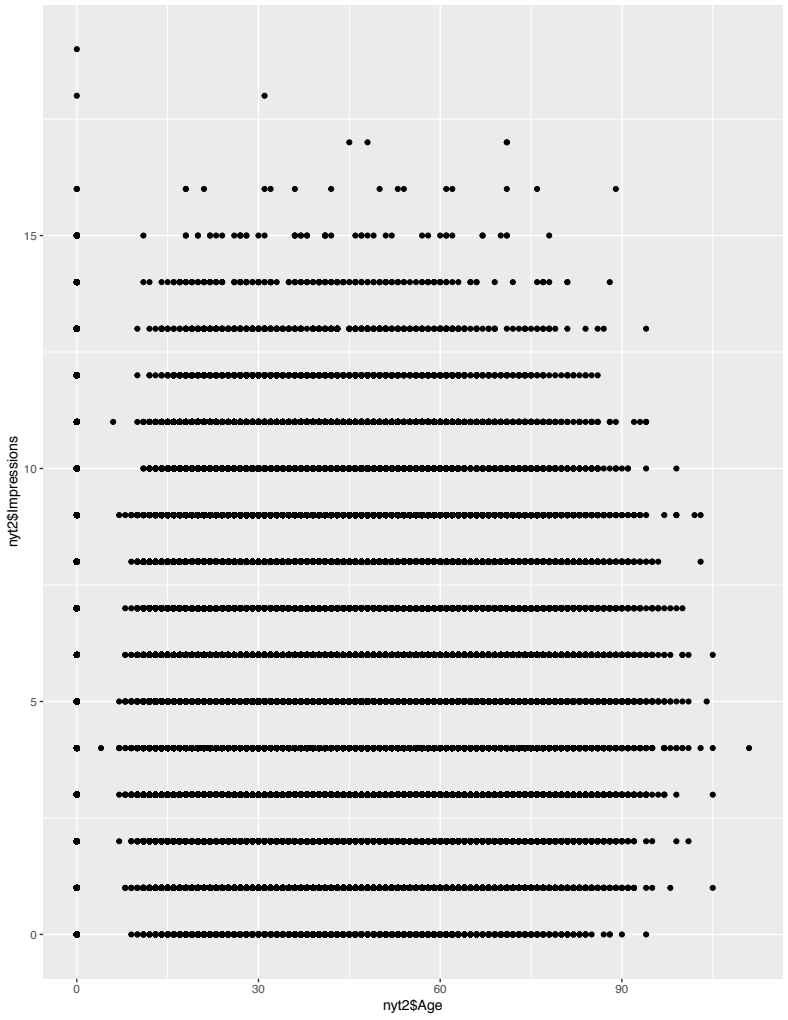
Q1c) Plotting the ecdf and qqplot for the two filtered variables

> age=ecdf(filterednyt2$Age)
> plot(age, ylab="Fn(x)", verticals = FALSE, col.01line = "gray70", pch = 19)
> imp=ecdf(nyt2$Impressions)
> plot(imp, ylab="Fn(x)", verticals = FALSE, col.01line = "gray70", pch = 19)

> plot(imp, ylab="Fn(x)", verticals = FALSE, col.01line = "gray70", pch = 19)


ecdf(filterednyt2$Age)
Fn(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120


ecdf(nyt2$Impressions)
Fn(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20


nyt2$Impressions
nyt2$Age
0
5
10
15
0
30
60
90

The e.c.d.f. (empirical cumulative distribution function) Fn is a step function with jumps i/n at observation values, where i is the number of tied observations at that value. Missing values are ignored. For observations x=(x1,x2, … xn), Fn is the fraction of observations less or equal to t,

i.e., Fn(t)=#{xi≤t} /n=1nn∑i=11[xi≤t].

Reference: https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/ecdf

So, we end up plotting the function fn , plot a feature of your data in order from least to greatest and see the whole feature as if is distributed across the data set. This suggests that there are lesser number of tied observations in the variable.

So, we can see that in case of Age variable, it is mostly continuous with small steps. The number of values of the Age variable are largely concentrated as it reaches 1.

For the impressions variable ecdf comes out to be somewhat non-continuous function with large steps. This suggests us that there are large number of tied observations in the Impressions variable. The number of values increases as the function tends to zero.

Q1d) Performing the significance tests of the filtered variables

```
> cor.test(filterednyt2$Age, filterednyt2$Impressions)

data:  filterednyt2$Age and filterednyt2$Impressions
t = 0.055769, df = 168454, p-value = 0.9555
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.004639475  0.004911229
sample estimates:
        cor
0.0001358799

> cor.test(filterednyt3$Age, filterednyt3$Impressions)

data:  filterednyt3$Age and filterednyt3$Impressions
t = -1.9461, df = 168229, p-value = 0.05164
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -9.523061e-03  3.381325e-05
sample estimates:
        cor
-0.004744732
```

For the two filtered datasets taken(nyt2 and nyt3 after filtering), the true correlation between the variables Age and Impressions is not equal to zero. The null hypothesis is not valid in any in any of the tow filtered datasets. The p-value is same for the nyt2 dataset even after filtering. The p-value for the filtered nyt3 dataset increases by a large margin from the unfiltered one. The lower the p-value, the lesser the chance that this much correlation happened as a matter of chance. The lower the p-value (< 0.01 or 0.05 typically), stronger is the significance of the relationship. The p-value of the filtered dataset nyt2 is on the lower side which is good. Higher positive correlation means higher relation between the variables of the data. Negative correlation means that an increase in one variable reliably predicts a decrease in the other variable. Here for the filtered datasets, the

variables in the nyt3 dataset are negatively correlated. The correlation for the nyt2 dataset is close to zero.