D. Tenisha

20CSEG34

Data Analytics using R

Assignment – 1

1. Histogram analysis for dataset Insurance

The data given in data frame Insurance consist of the numbers of policyholders of an insurance company who were exposed to risk, and the numbers of car insurance claims made by those policyholders in the third quarter of 1973.

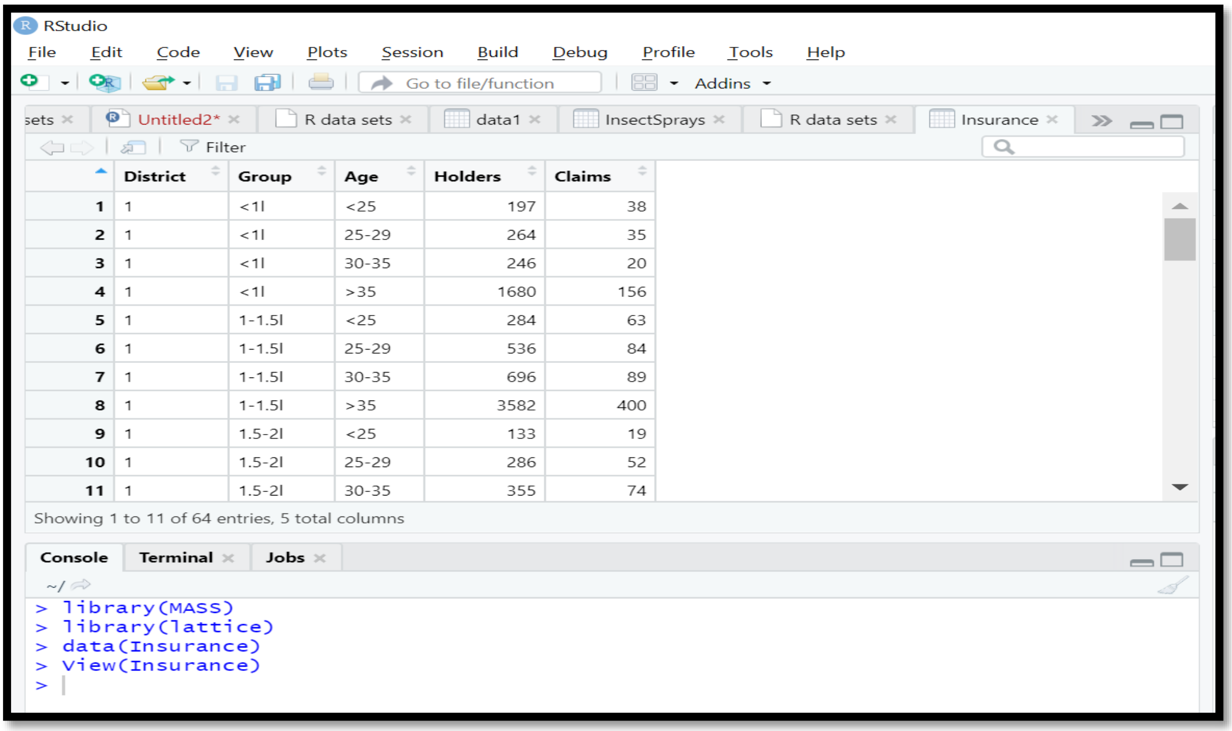
District - factor: district of residence of policyholder (1 to 4): 4 is major cities.

Group - an ordered factor: group of car with levels <1 litre, 1–1.5 litre, 1.5–2 litre, >2 litre.

Age - an ordered factor: the age of the insured in 4 groups labelled <25, 25–29, 30–35, >35.

Holders - numbers of policyholders.

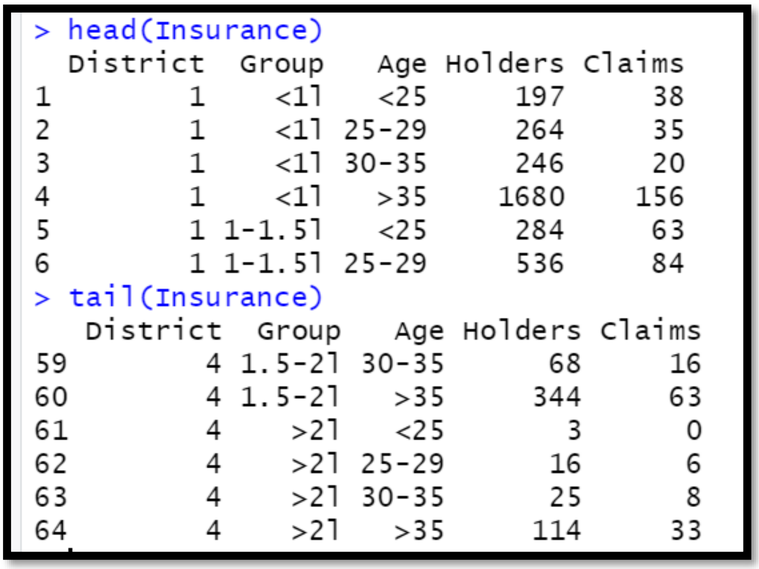
Claims - numbers of claims



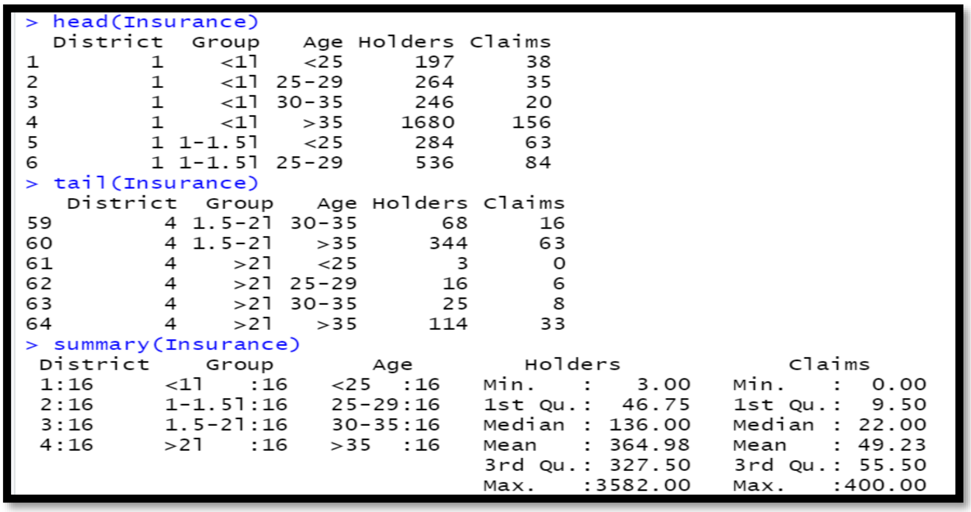
	District	Group	Age	Holders	Claims
1	1	<1l	<25	197	38
2	1	<1l	25-29	264	35
3	1	<1l	30-35	246	20
4	1	<1l	>35	1680	156
5	1	1-1.5l	<25	284	63
6	1	1-1.5l	25-29	536	84
7	1	1-1.5l	30-35	696	89
8	1	1-1.5l	>35	3582	400
9	1	1.5-2l	<25	133	19
10	1	1.5-2l	25-29	286	52
11	1	1.5-2l	30-35	355	74

Showing 1 to 11 of 64 entries, 5 total columns

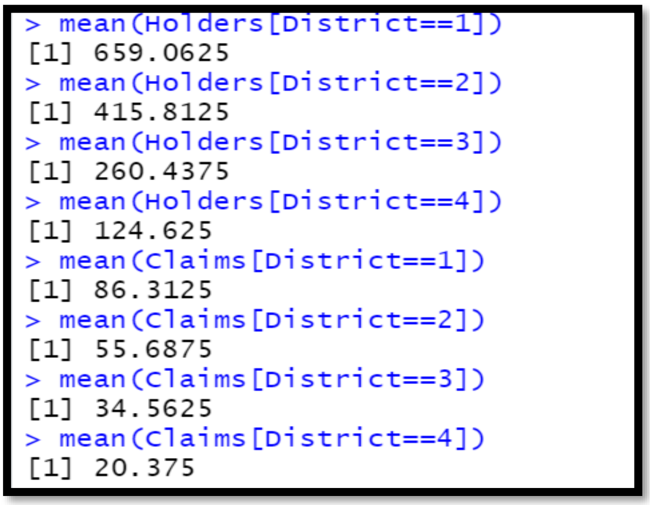
```
> library(MASS)
> library(lattice)
> data(Insurance)
> View(Insurance)
> 
```



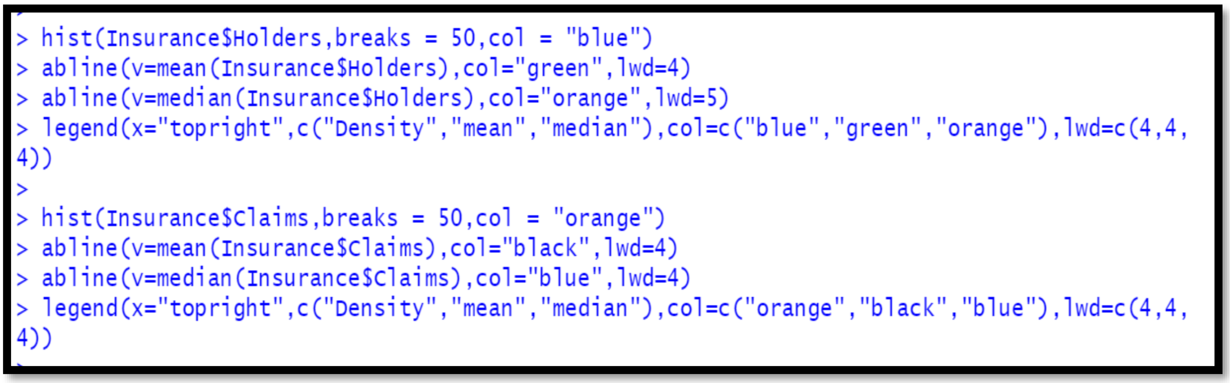
```
> head(Insurance)
  District  Group   Age Holders Claims
1        1    <1l   <25     197     38
2        1    <1l 25-29     264     35
3        1    <1l 30-35     246     20
4        1    <1l   >35    1680    156
5        1 1-1.5l   <25     284     63
6        1 1-1.5l 25-29     536     84
> tail(Insurance)
   District  Group   Age Holders Claims
59        4 1.5-2l 30-35      68     16
60        4 1.5-2l   >35     344     63
61        4    >2l   <25       3      0
62        4    >2l 25-29      16      6
63        4    >2l 30-35      25      8
64        4    >2l   >35     114     33
```



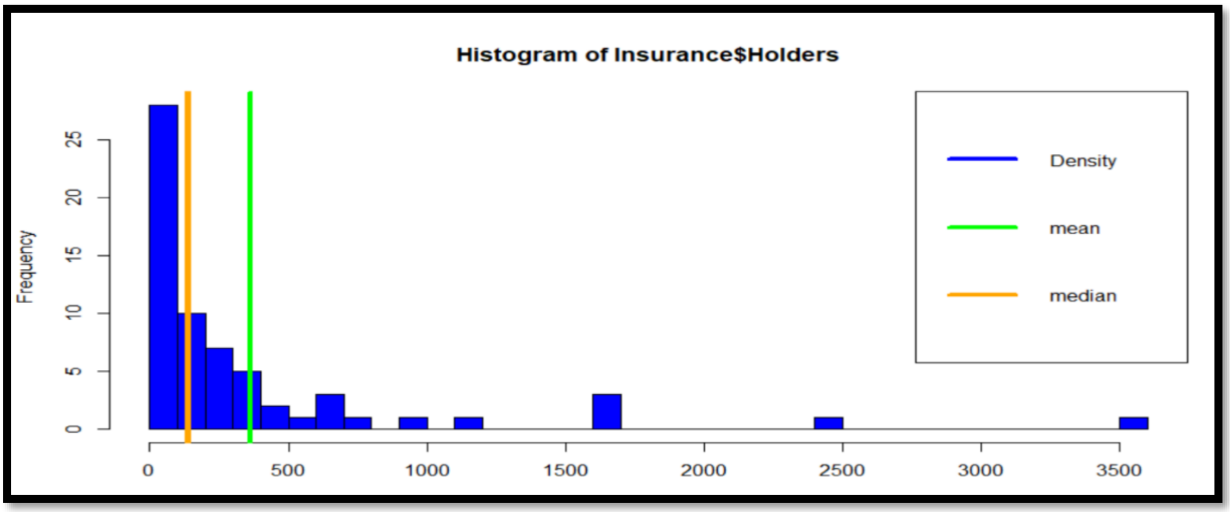
```
> head(Insurance)
  District  Group   Age Holders Claims
1        1    <1l   <25     197     38
2        1    <1l 25-29     264     35
3        1    <1l 30-35     246     20
4        1    <1l   >35    1680    156
5        1 1-1.5l   <25     284     63
6        1 1-1.5l 25-29     536     84
> tail(Insurance)
   District  Group   Age Holders Claims
59        4 1.5-2l 30-35      68     16
60        4 1.5-2l   >35     344     63
61        4    >2l   <25       3      0
62        4    >2l 25-29      16      6
63        4    >2l 30-35      25      8
64        4    >2l   >35     114     33
> summary(Insurance)
 District      Group         Age          Holders             Claims      
 1:16     <1l   :16   <25  :16   Min.   :   3.00   Min.   :  0.00  
 2:16     1-1.5l:16   25-29:16   1st Qu.:  46.75   1st Qu.:  9.50  
 3:16     1.5-2l:16   30-35:16   Median : 136.00   Median : 22.00  
 4:16     >2l   :16   >35  :16   Mean   : 364.98   Mean   : 49.23  
                                 3rd Qu.: 327.50   3rd Qu.: 55.50  
                                 Max.   :3582.00   Max.   :400.00  
```



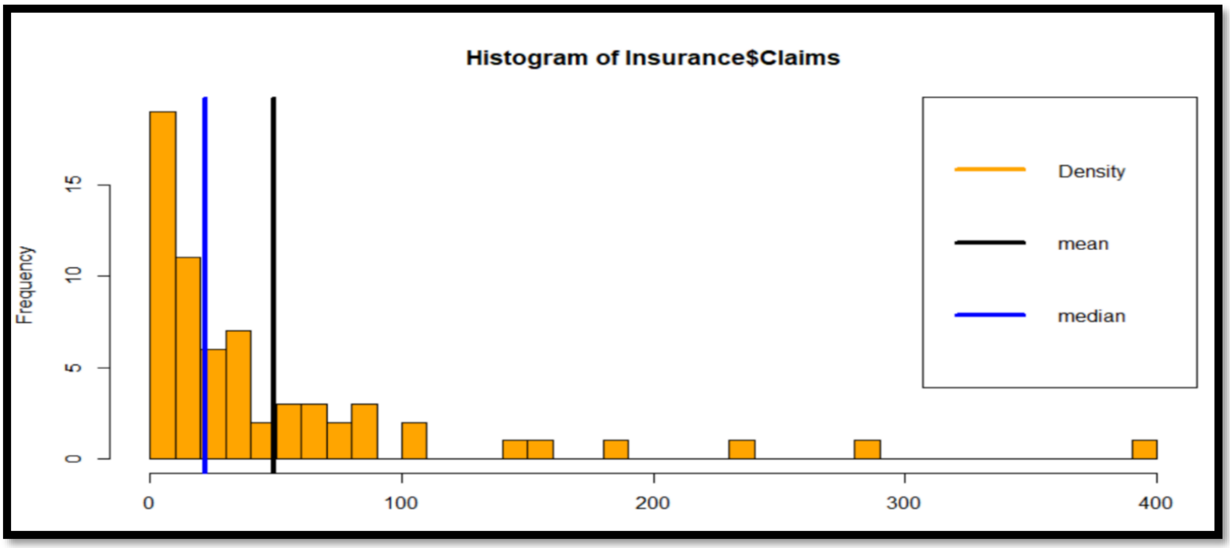
```
> mean(Holders[District==1])
[1] 659.0625
> mean(Holders[District==2])
[1] 415.8125
> mean(Holders[District==3])
[1] 260.4375
> mean(Holders[District==4])
[1] 124.625
> mean(Claims[District==1])
[1] 86.3125
> mean(Claims[District==2])
[1] 55.6875
> mean(Claims[District==3])
[1] 34.5625
> mean(Claims[District==4])
[1] 20.375
```



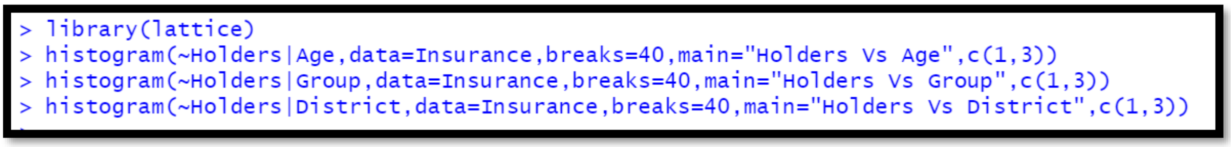
```
> hist(Insurance$Holders,breaks = 50,col = "blue")
> abline(v=mean(Insurance$Holders),col="green",lwd=4)
> abline(v=median(Insurance$Holders),col="orange",lwd=5)
> legend(x="topright",c("Density","mean","median"),col=c("blue","green","orange"),lwd=c(4,4,
4))
>
> hist(Insurance$Claims,breaks = 50,col = "orange")
> abline(v=mean(Insurance$Claims),col="black",lwd=4)
> abline(v=median(Insurance$Claims),col="blue",lwd=4)
> legend(x="topright",c("Density","mean","median"),col=c("orange","black","blue"),lwd=c(4,4,
4))
```



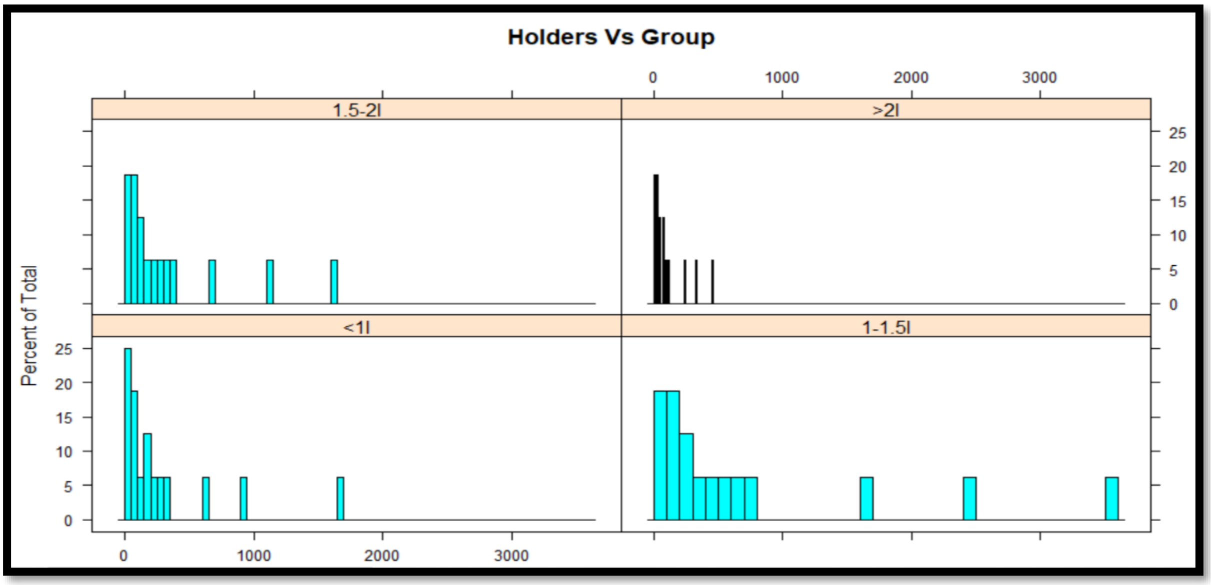
The above plot shows the Histogram of Holders in Insurance dataset. The green line represents mean and the orange line represents medians. The outlier here is above 3500. The plot is right skewed, it means positively skewed.



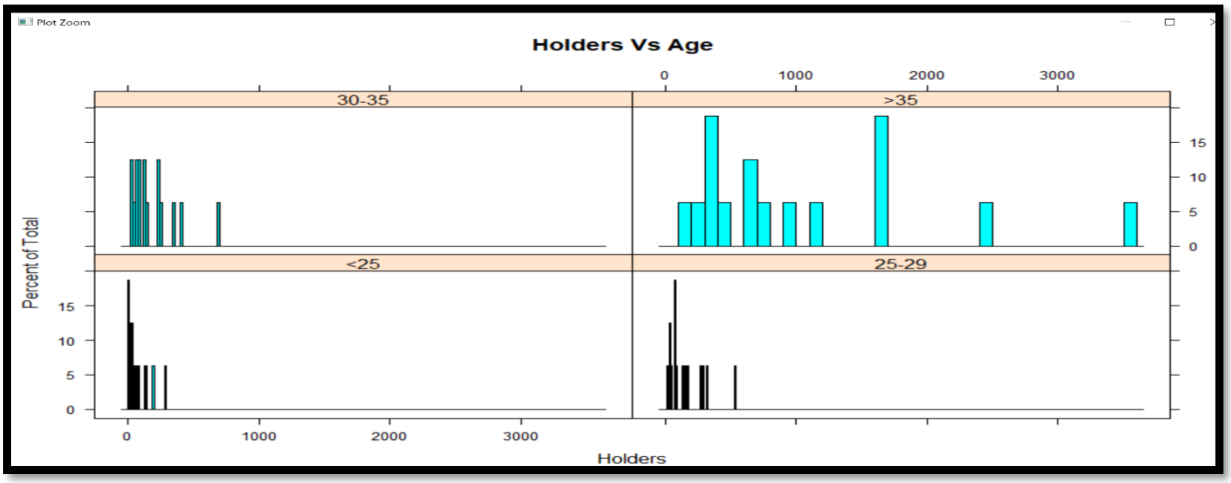
The above plot shows the Histogram of Holders in Insurance dataset. The black line represents mean and the blue line represents medians. The outlier is near to 400. The plot is right skewed, it means positively skewed.



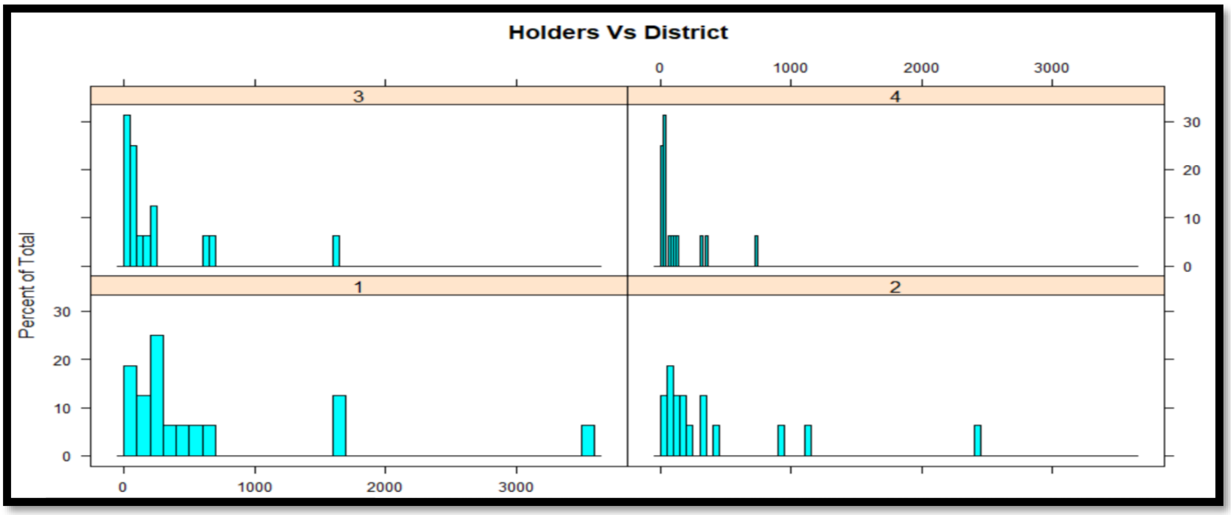
```
> library(lattice)
> histogram(~Holders|Age,data=Insurance,breaks=40,main="Holders Vs Age",c(1,3))
> histogram(~Holders|Group,data=Insurance,breaks=40,main="Holders Vs Group",c(1,3))
> histogram(~Holders|District,data=Insurance,breaks=40,main="Holders Vs District",c(1,3))
```



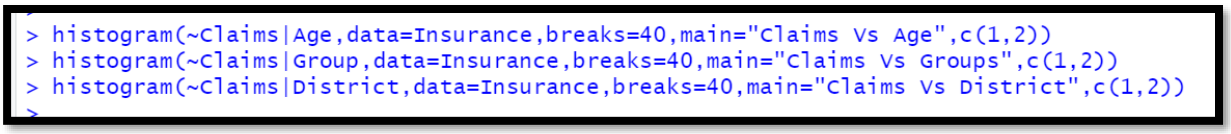
This image shows the histogram of Insurance Holders and Liter groups. <1 liter holds more number insurance than others.



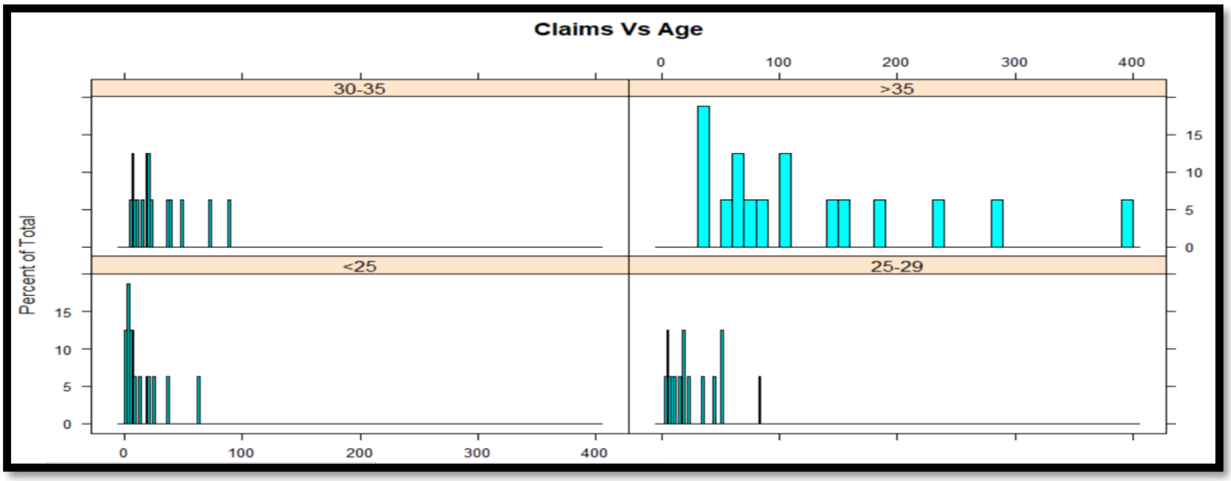
This image shows the comparison of Insurance holders and different age groups. >35 age groups hold more insurance. 30-35 age groups holds less number of insutrance.

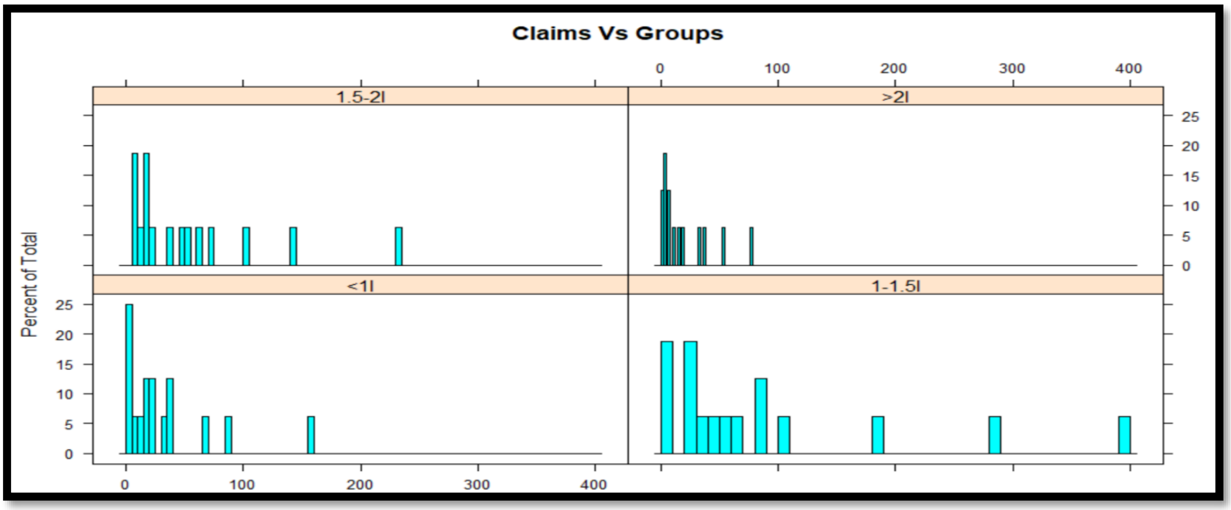


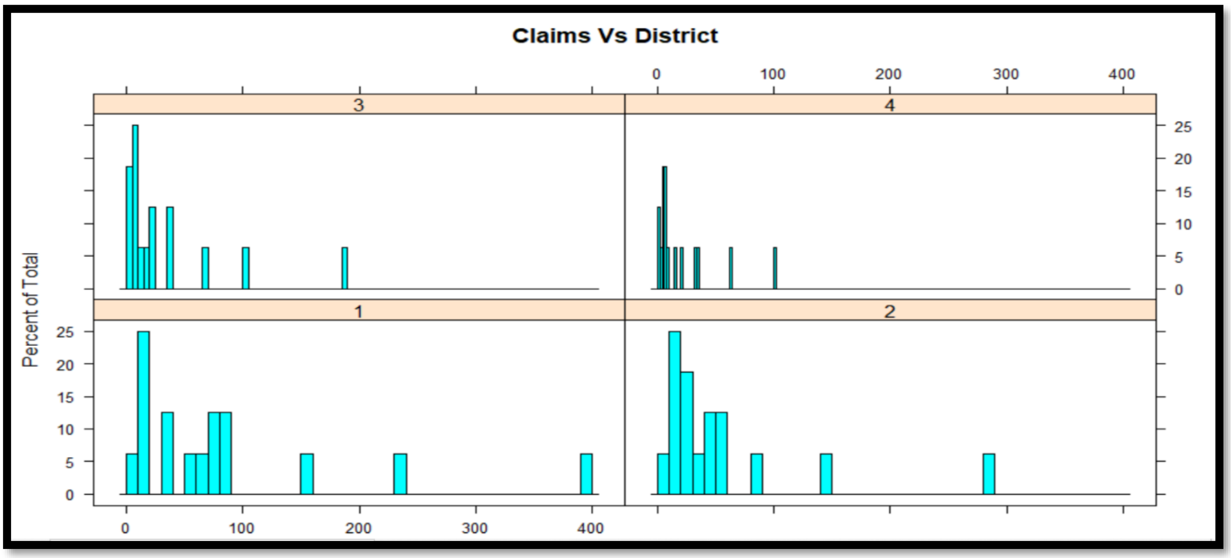
This image shows the comparison of Insurance holders vs different District (1,2,3,4).



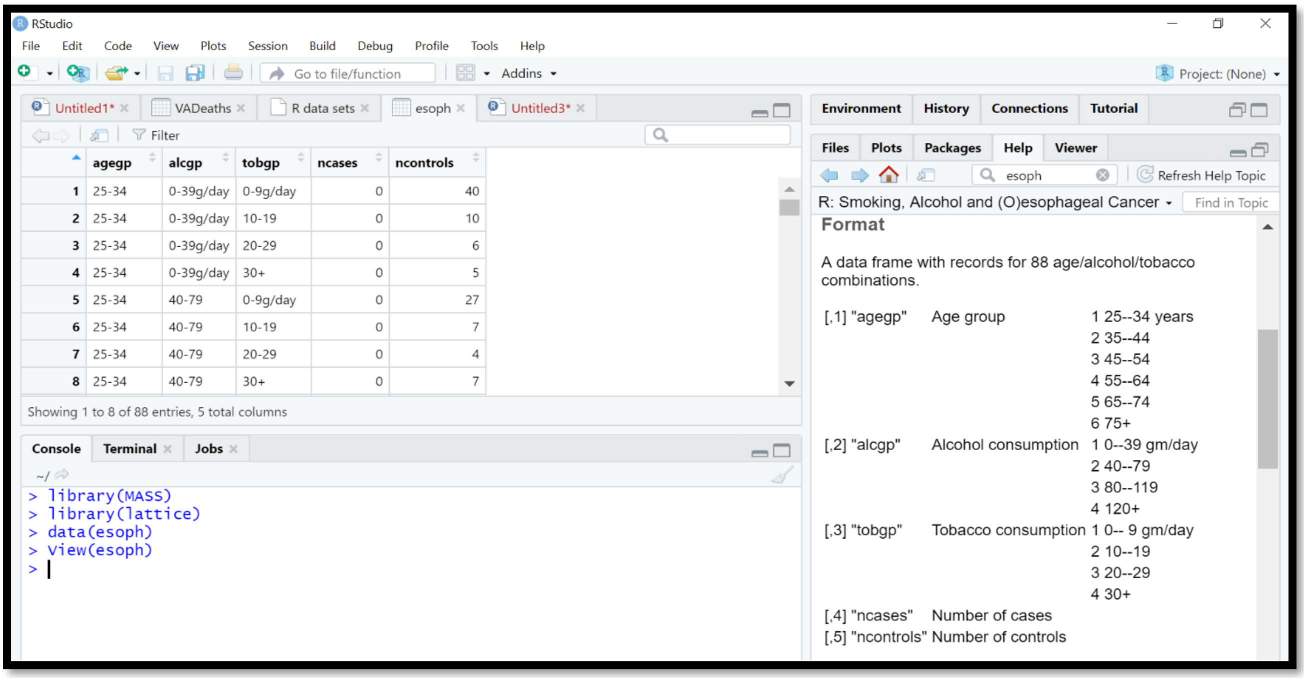
```
> histogram(~Claims|Age,data=Insurance,breaks=40,main="Claims Vs Age",c(1,2))
> histogram(~Claims|Group,data=Insurance,breaks=40,main="Claims Vs Groups",c(1,2))
> histogram(~Claims|District,data=Insurance,breaks=40,main="Claims Vs District",c(1,2))
>
```




Claims Vs Groups
1.5-2l
>2l
<1l
1-1.5l
Percent of Total
0
5
10
15
20
25
0
100
200
300
400


Claims Vs District
3
4
1
2
Percent of Total
0
5
10
15
20
25
0
100
200
300
400

Histogram Analysis using Esoph data set



	agegp	alcgp	tobgp	ncases	ncontrols
1	25-34	0-39g/day	0-9g/day	0	40
2	25-34	0-39g/day	10-19	0	10
3	25-34	0-39g/day	20-29	0	6
4	25-34	0-39g/day	30+	0	5
5	25-34	40-79	0-9g/day	0	27
6	25-34	40-79	10-19	0	7
7	25-34	40-79	20-29	0	4
8	25-34	40-79	30+	0	7

Showing 1 to 8 of 88 entries, 5 total columns

```
> library(MASS)
> library(lattice)
> data(esoph)
> View(esoph)
> 
```

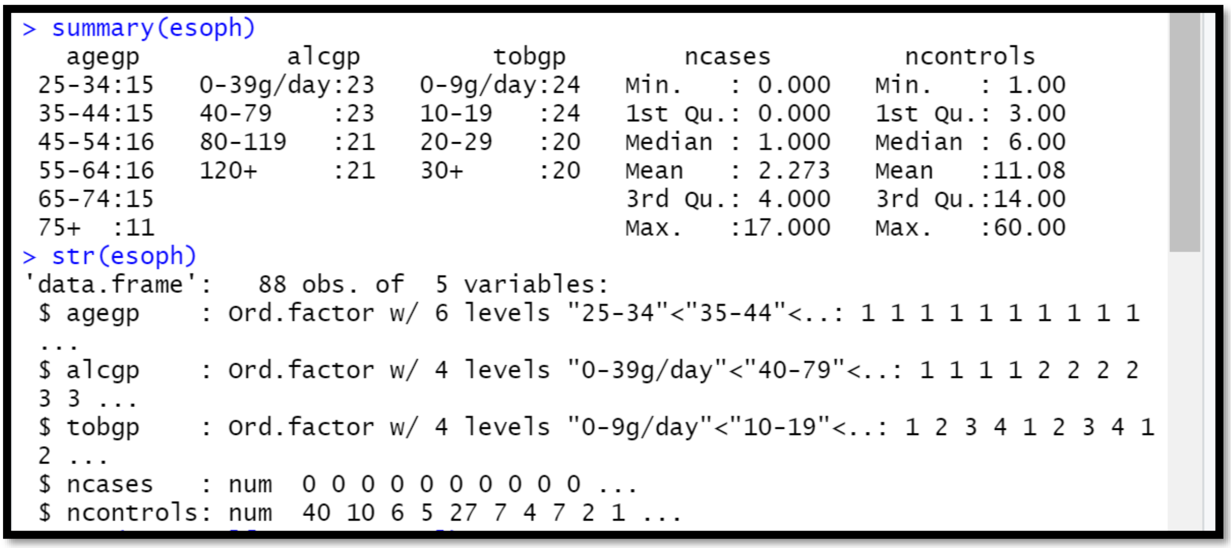
R: Smoking, Alcohol and (O)esophageal Cancer

Format

A data frame with records for 88 age/alcohol/tobacco combinations.

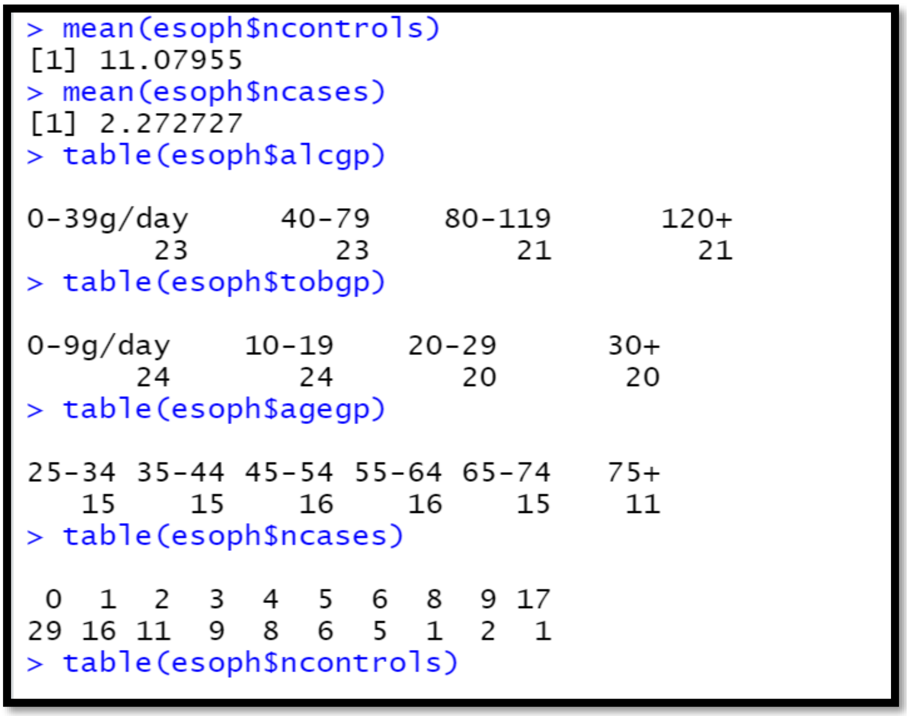
[,1] "agegp"	Age group	1 25--34 years 2 35--44 3 45--54 4 55--64 5 65--74 6 75+
[,2] "alcgp"	Alcohol consumption	1 0--39 gm/day 2 40--79 3 80--119 4 120+
[,3] "tobgp"	Tobacco consumption	1 0-- 9 gm/day 2 10--19 3 20--29 4 30+
[,4] "ncases"	Number of cases	
[,5] "ncontrols"	Number of controls	

The above picture gives the information about the data set which is “esoph” and shows the coding for loading and viewing the package.

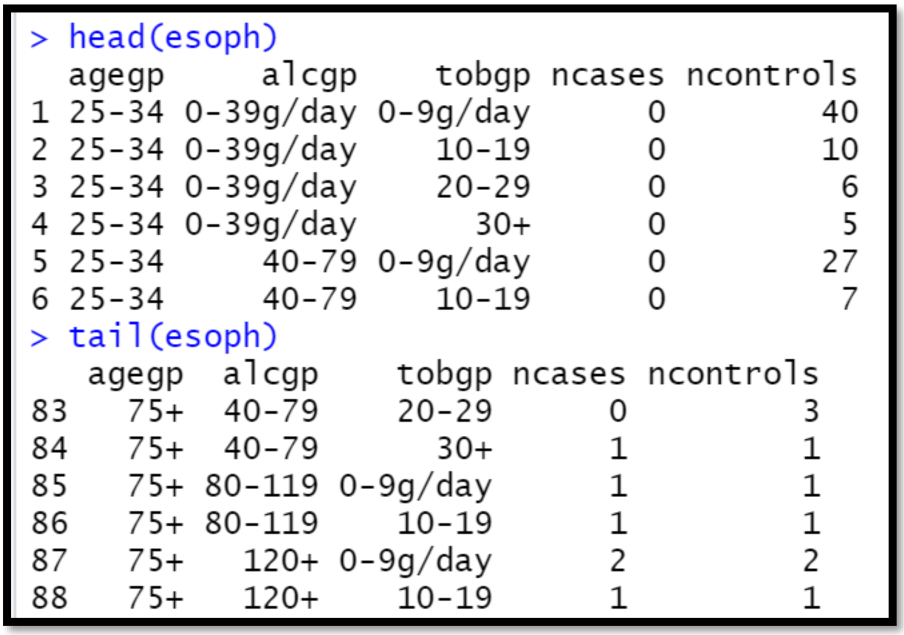


```
> summary(esoph)
   agegp          alcgp          tobgp          ncases         ncontrols    
 25-34:15   0-39g/day:23   0-9g/day:24   Min.   : 0.000   Min.   : 1.00  
 35-44:15   40-79    :23   10-19    :24   1st Qu.: 0.000   1st Qu.: 3.00  
 45-54:16   80-119   :21   20-29    :20   Median : 1.000   Median : 6.00  
 55-64:16   120+     :21   30+      :20   Mean   : 2.273   Mean   :11.08  
 65-74:15                                 3rd Qu.: 4.000   3rd Qu.:14.00  
 75+  :11                                 Max.   :17.000   Max.   :60.00  
> str(esoph)
'data.frame':	88 obs. of  5 variables:
 $ agegp    : Ord.factor w/ 6 levels "25-34"<"35-44"<..: 1 1 1 1 1 1 1 1 1 1
 ...
 $ alcgp    : Ord.factor w/ 4 levels "0-39g/day"<"40-79"<..: 1 1 1 1 2 2 2 2
 3 3 ...
 $ tobgp    : Ord.factor w/ 4 levels "0-9g/day"<"10-19"<..: 1 2 3 4 1 2 3 4 1
 2 ...
 $ ncases   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ncontrols: num  40 10 6 5 27 7 4 7 2 1 ...
```

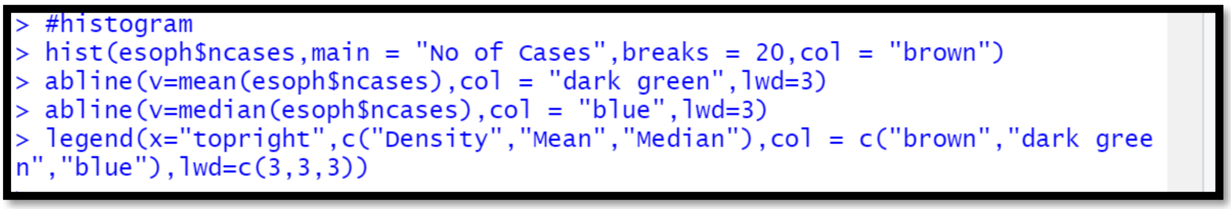
This image displays the coding and output of summary() and str() functions



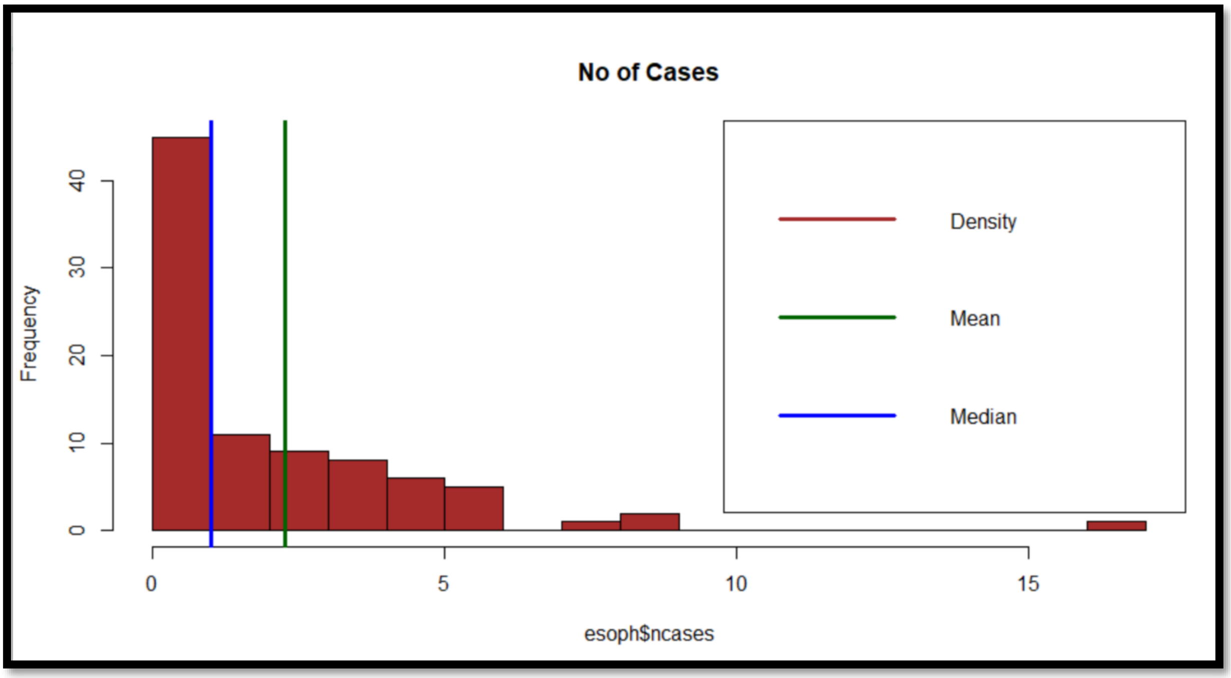
```
> mean(esoph$ncontrols)
[1] 11.07955
> mean(esoph$ncases)
[1] 2.272727
> table(esoph$alcgp)

0-39g/day     40-79    80-119      120+ 
       23        23        21        21 
> table(esoph$tobgp)

0-9g/day    10-19    20-29      30+ 
      24       24       20       20 
> table(esoph$agegp)

25-34 35-44 45-54 55-64 65-74   75+ 
   15    15    16    16    15    11 
> table(esoph$ncases)

 0  1  2  3  4  5  6  8  9 17 
29 16 11  9  8  6  5  1  2  1 
> table(esoph$ncontrols)
```

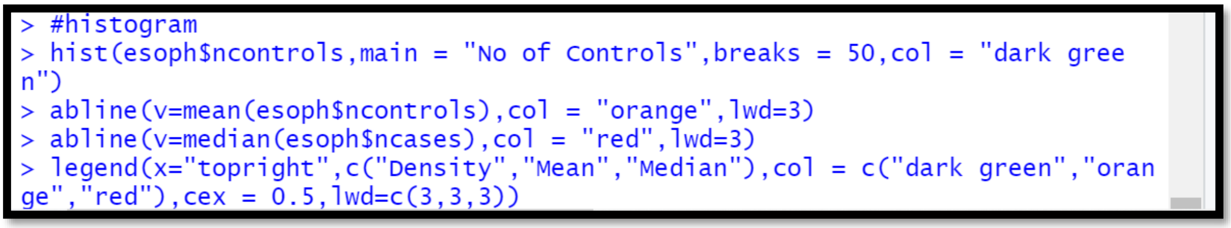


```
> head(esoph)
  agegp     alcgp    tobgp ncases ncontrols
1 25-34 0-39g/day 0-9g/day      0        40
2 25-34 0-39g/day    10-19      0        10
3 25-34 0-39g/day    20-29      0         6
4 25-34 0-39g/day      30+      0         5
5 25-34     40-79 0-9g/day      0        27
6 25-34     40-79    10-19      0         7
> tail(esoph)
   agegp  alcgp    tobgp ncases ncontrols
83   75+  40-79    20-29      0         3
84   75+  40-79      30+      1         1
85   75+ 80-119 0-9g/day      1         1
86   75+ 80-119    10-19      1         1
87   75+   120+ 0-9g/day      2         2
88   75+   120+    10-19      1         1
```

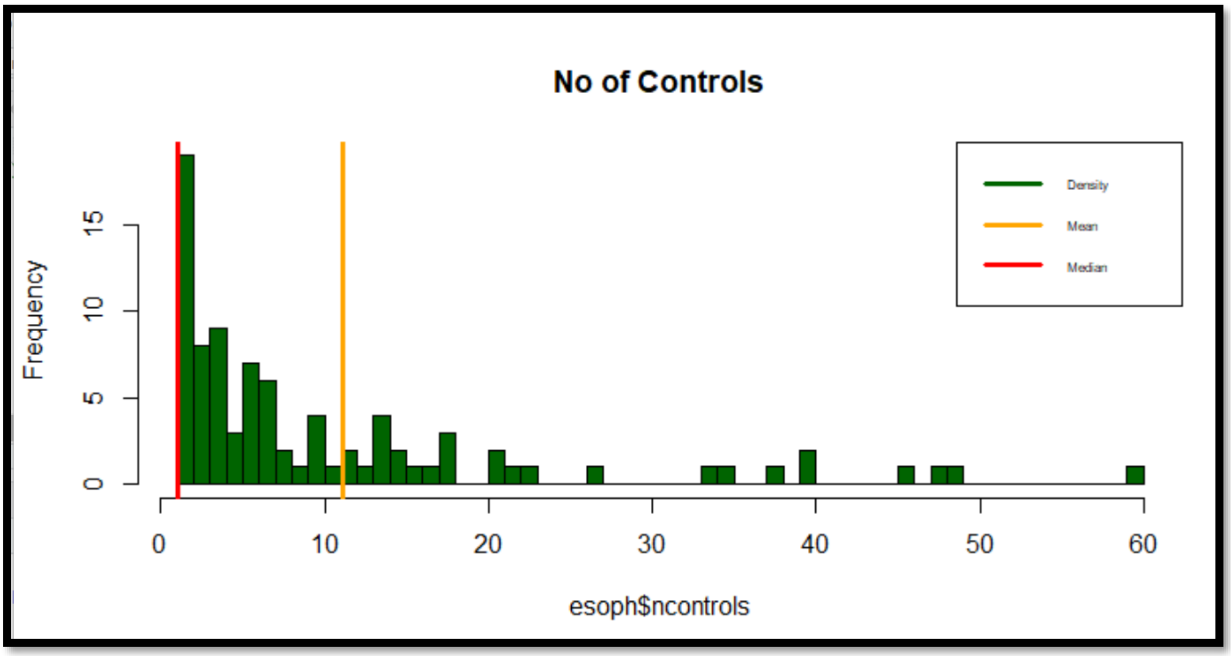


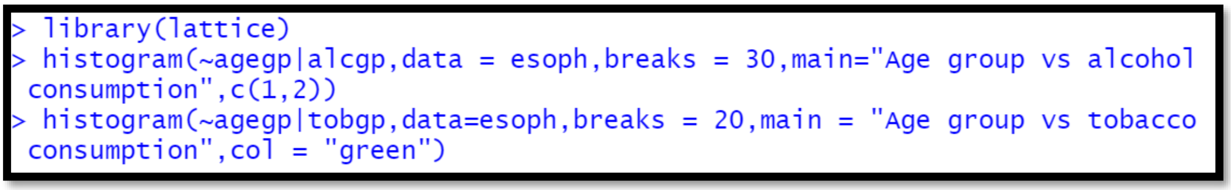
```
> #histogram
> hist(esoph$ncases,main = "No of Cases",breaks = 20,col = "brown")
> abline(v=mean(esoph$ncases),col = "dark green",lwd=3)
> abline(v=median(esoph$ncases),col = "blue",lwd=3)
> legend(x="topright",c("Density","Mean","Median"),col = c("brown","dark gree
n","blue"),lwd=c(3,3,3))
```



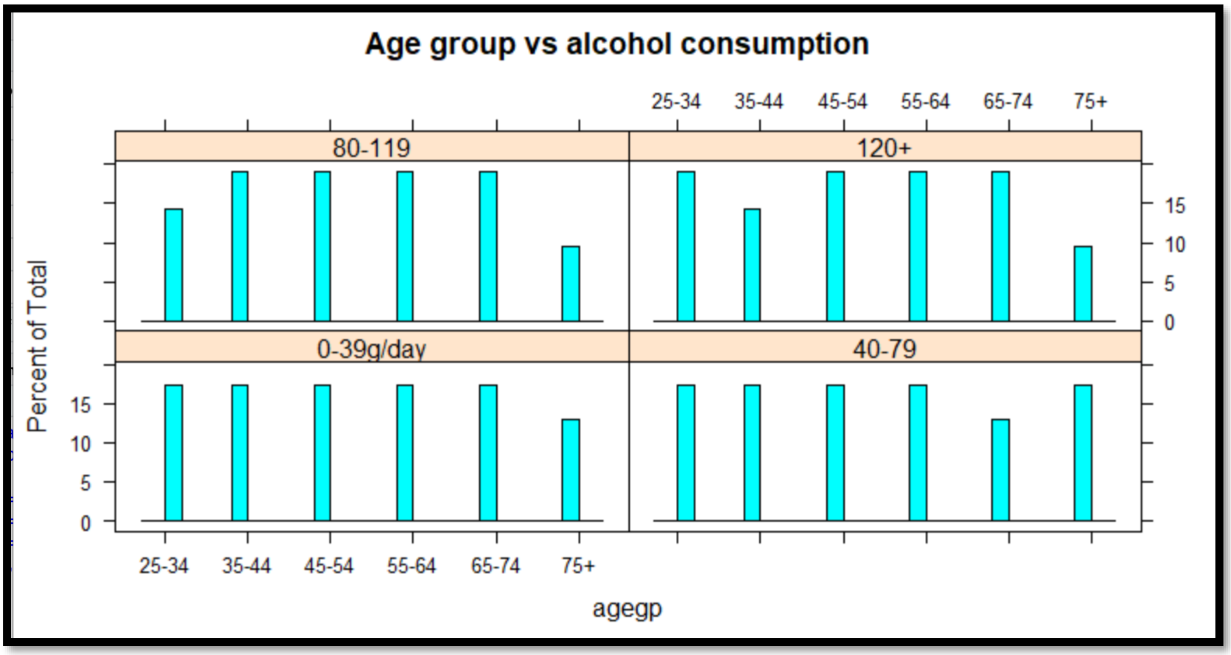


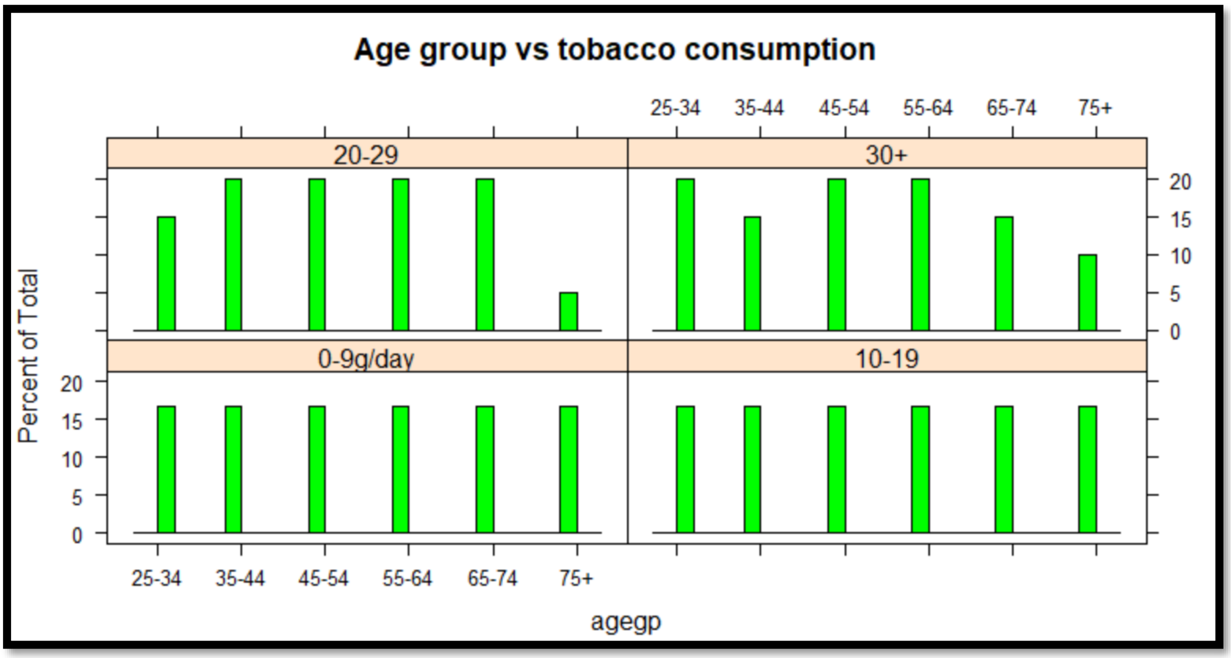
```
> #histogram
> hist(esoph$ncontrols,main = "No of Controls",breaks = 50,col = "dark gree
n")
> abline(v=mean(esoph$ncontrols),col = "orange",lwd=3)
> abline(v=median(esoph$ncases),col = "red",lwd=3)
> legend(x="topright",c("Density","Mean","Median"),col = c("dark green","oran
ge","red"),cex = 0.5,lwd=c(3,3,3))
```





```
> library(lattice)
> histogram(~agegp|alcgp,data = esoph,breaks = 30,main="Age group vs alcohol
 consumption",c(1,2))
> histogram(~agegp|tobgp,data=esoph,breaks = 20,main = "Age group vs tobacco
 consumption",col = "green")
```


Age group vs alcohol consumption
0-39g/day
Percent of Total
agegp


Age group vs tobacco consumption
0-9g/day
Percent of Total
agegp