Homework 1

Hrishikesh Telang (hnt2107)

25th September,2018

Note: Grading is based both on your graphs and verbal explanations. Follow all best practices as discussed in class.

Read *Graphical Data Analysis with R*, Ch. 3

1. Salary

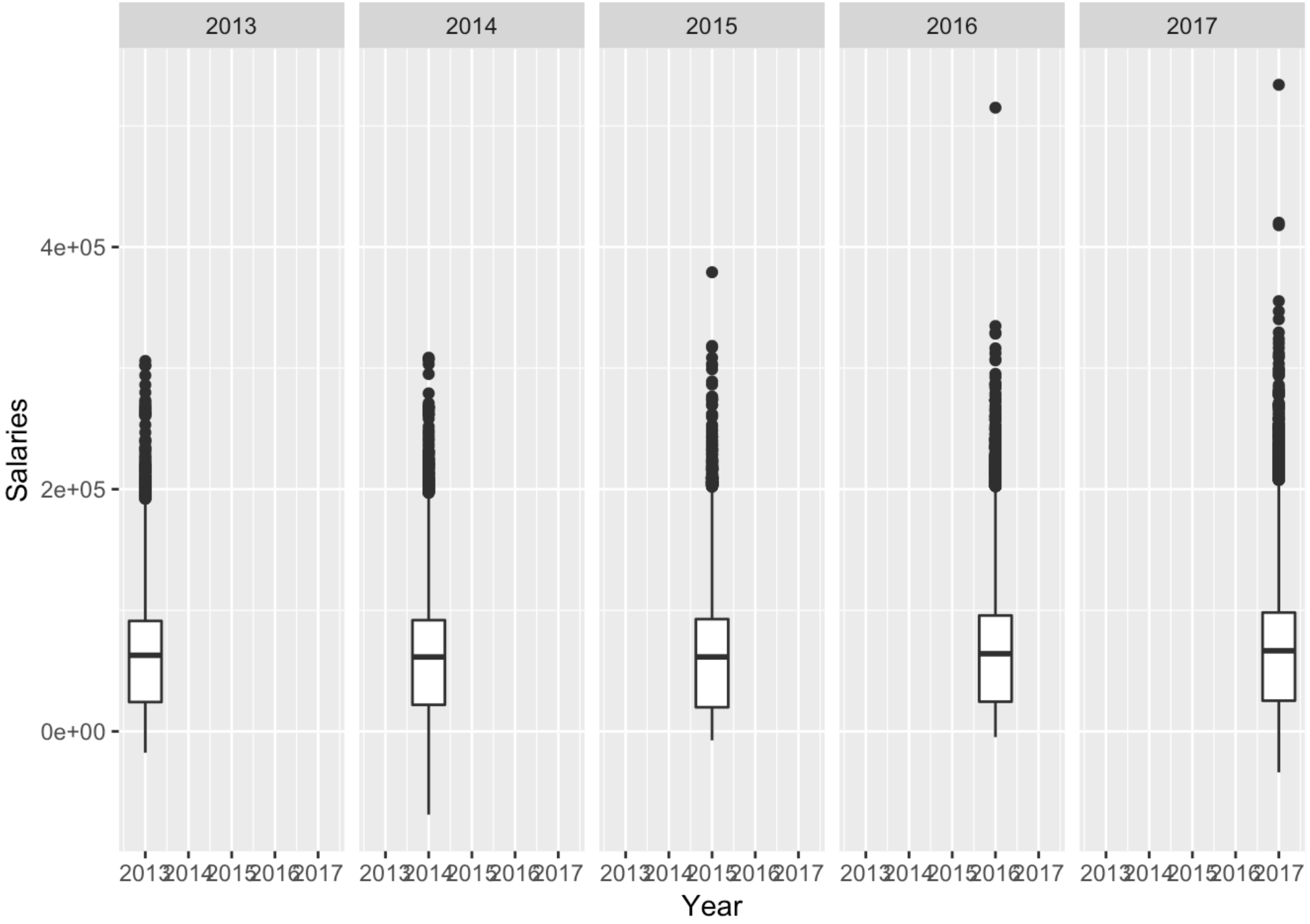
[15 points]

a) Draw multiple boxplots, by year, for the `Salaries` variable in *Employee.csv* (Available in the Data folder in the Files section of CourseWorks, original source: https://catalog.data.gov/dataset/employee-compensation-53987 (https://catalog.data.gov/dataset/employee-compensation-53987)). How do the distributions differ by year?

```
data=read.csv("/Users/hrishikeshtelang/Desktop/DSI acads/EDAV/Employee.csv")
library(ggplot2)
library(dplyr)
names(data)
```

```
## [1] "Year"               "Organization.Group" "Salaries"
## [4] "Overtime"
```

```
p<-ggplot(data,aes(Year,Salaries))+geom_boxplot()+facet_grid(~ Year)
p
```



```
data%>% group_by(Year)%>% summarise(Mean=mean(Salaries),Median=median(Salaries))
```

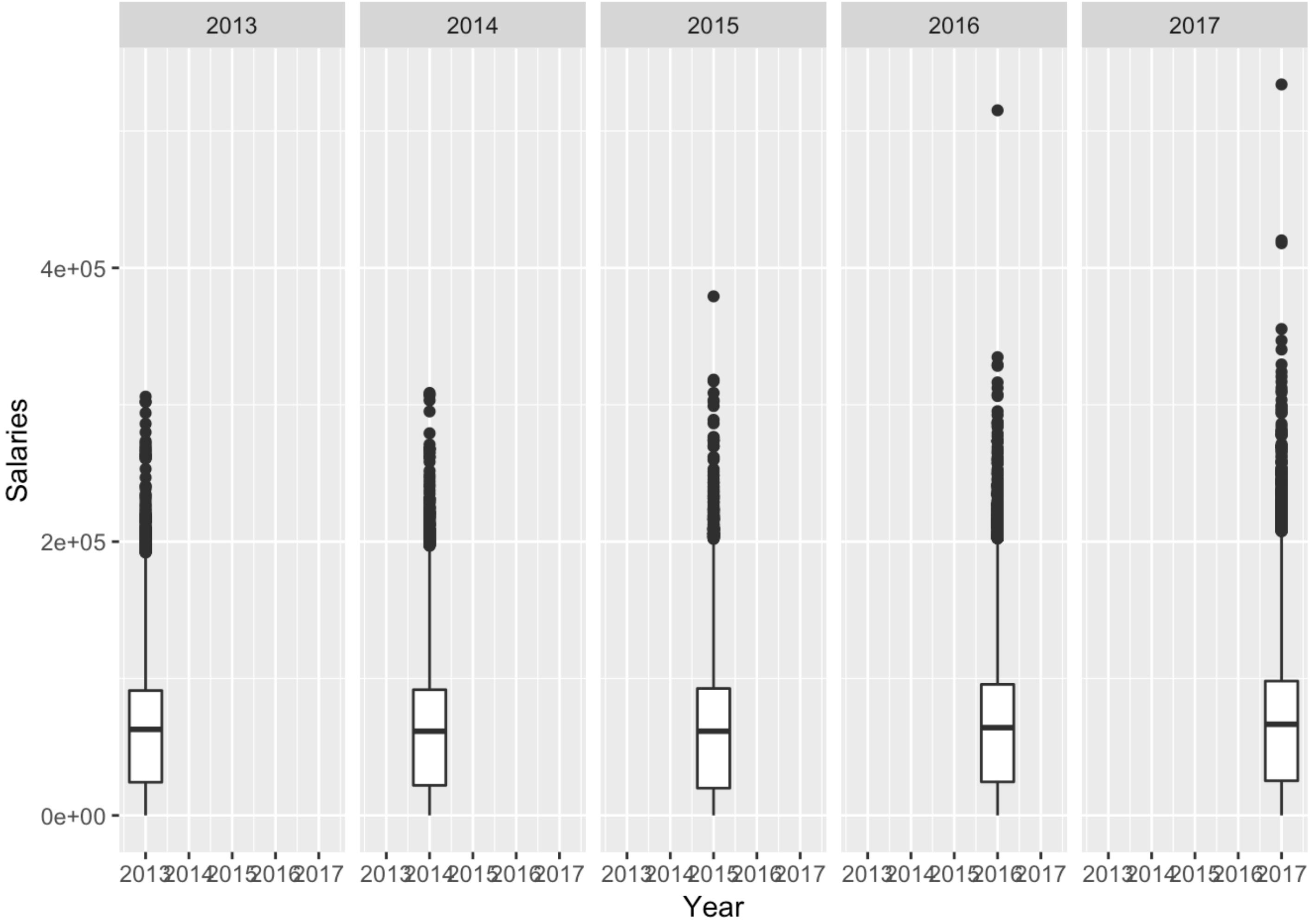
```
## # A tibble: 5 x 3
##    Year   Mean Median
##   <int>  <dbl>  <dbl>
## 1  2013 62876. 62848.
## 2  2014 61914. 61482.
## 3  2015 61969. 61509.
## 4  2016 64874. 64128.
## 5  2017 67062. 66585.
```

From the boxplot, we can infer that there are certain values for salary which are negative. Removing these values and creating a new dataframe:

```
data[,3][data[,3]<0]<-0
```

Plotting the new boxplot:

```
p1<-ggplot(data,aes(Year,Salaries))+geom_boxplot()+facet_grid(~ Year)
p1
```



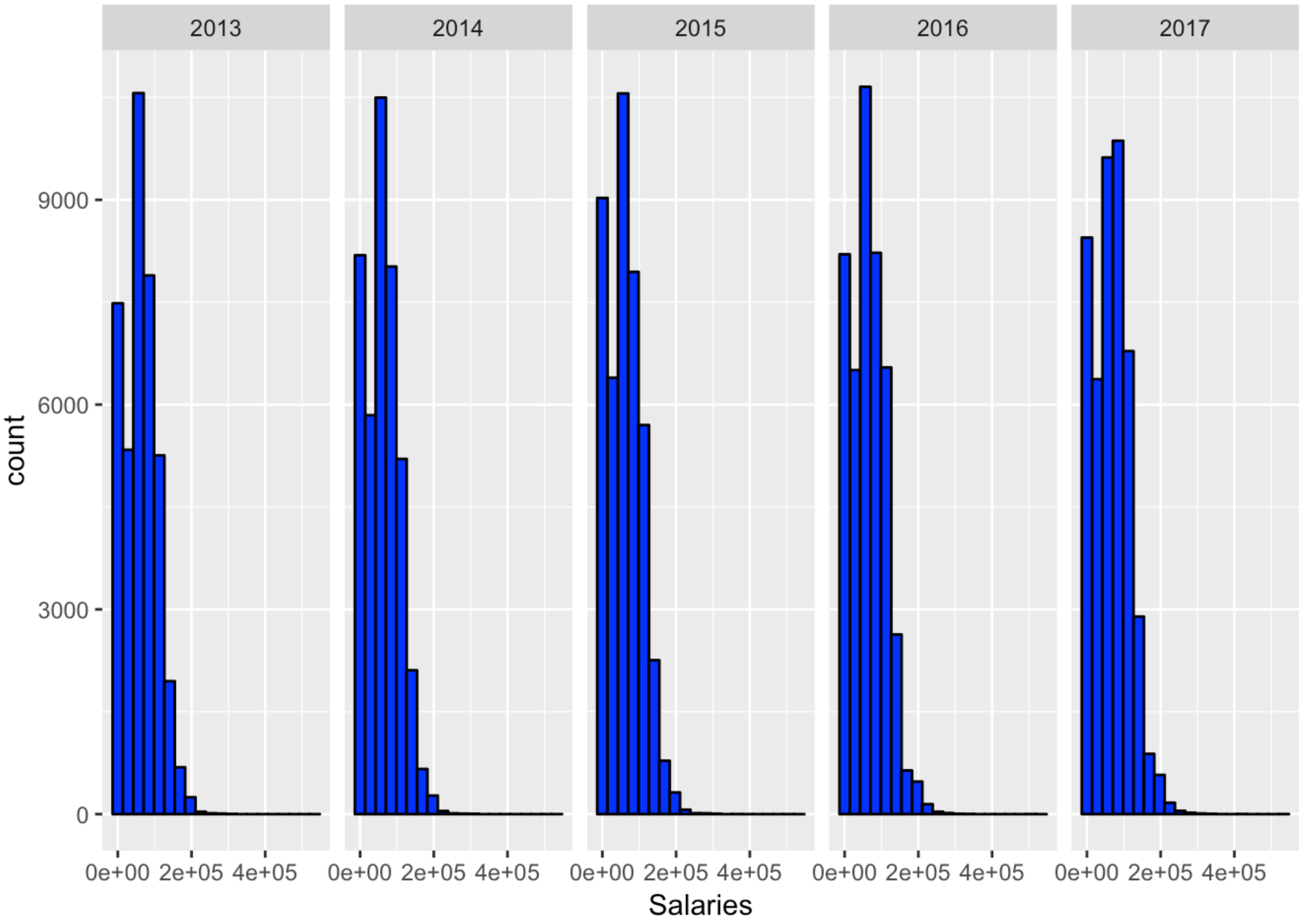
```
data%>% group_by(Year)%>% summarise(Mean=mean(Salaries),Median=median(Salaries),Stand
ard=sd(Salaries))
```

```
## # A tibble: 5 x 4
##    Year   Mean Median Standard
##   <int>  <dbl>  <dbl>    <dbl>
## 1  2013 62877. 62848.   43368.
## 2  2014 61917. 61482.   44042.
## 3  2015 61970. 61509.   45336.
## 4  2016 64874. 64128.   46114.
## 5  2017 67063. 66585.   47585.
```

We can also see that 2017 has extreme outliers which has lead to it having the maximum mean and standard deviation.

b) Draw histograms, faceted by year, for the same data. What additional information do the histograms provide?

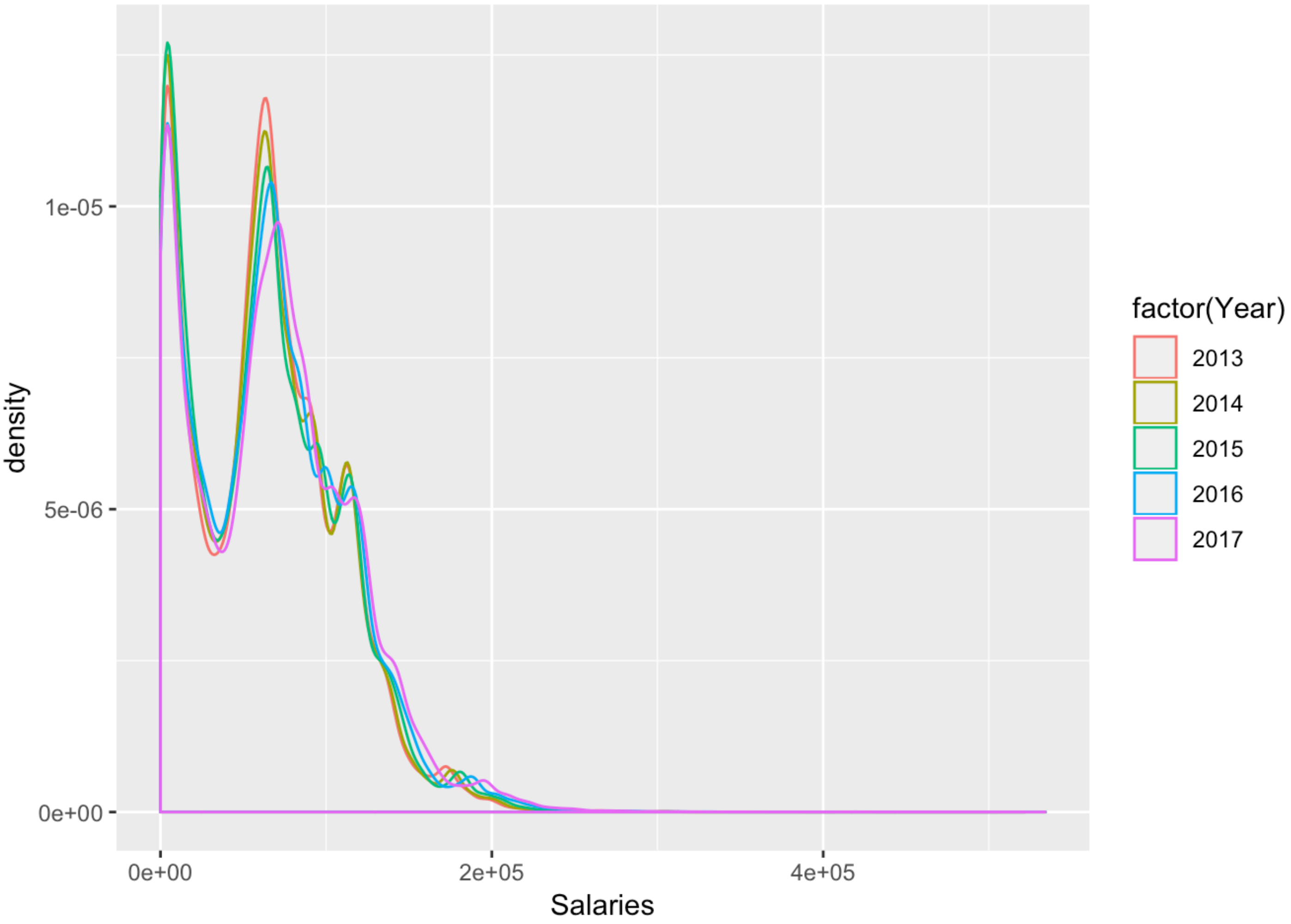
```
p2<-ggplot(data,aes(Salaries))+geom_histogram(bins=20,col="black",fill="blue")+facet_
grid(~ Year)
p2
```



The histograms provide us the most frequently occuring value of salary by year, i.e. helps us find the Mode of data by year. We can also see the general distribution of data

c) Plot overlapping density curves of the same data, one curve per year, on a single set of axes. Each curve should be a different color. What additional information do you learn?

```
p5 <- ggplot(data,aes(Salaries))+geom_density(aes(group=Year,color=factor(Year)))
p5
```



We learn from the density curves that more people earned higher values of salaries in 2017 compared to other years which leads it to have a density curve which is more spread out than the others. At the same time, we can see that the distribution of salaries in all years followed almost the same trajectory. We can observe that while a lot of people earned higher salaries in 2013 as evidenced by the second orange peak, the proportion dropped in the following years, with 2015(the first green peak) indicating that most people earned lower salaries on an average compared to 2013.

d) Sum up the results of a), b) and c): what kinds of questions, specific to this dataset, would be best answered about the data by each of the three graphical forms?

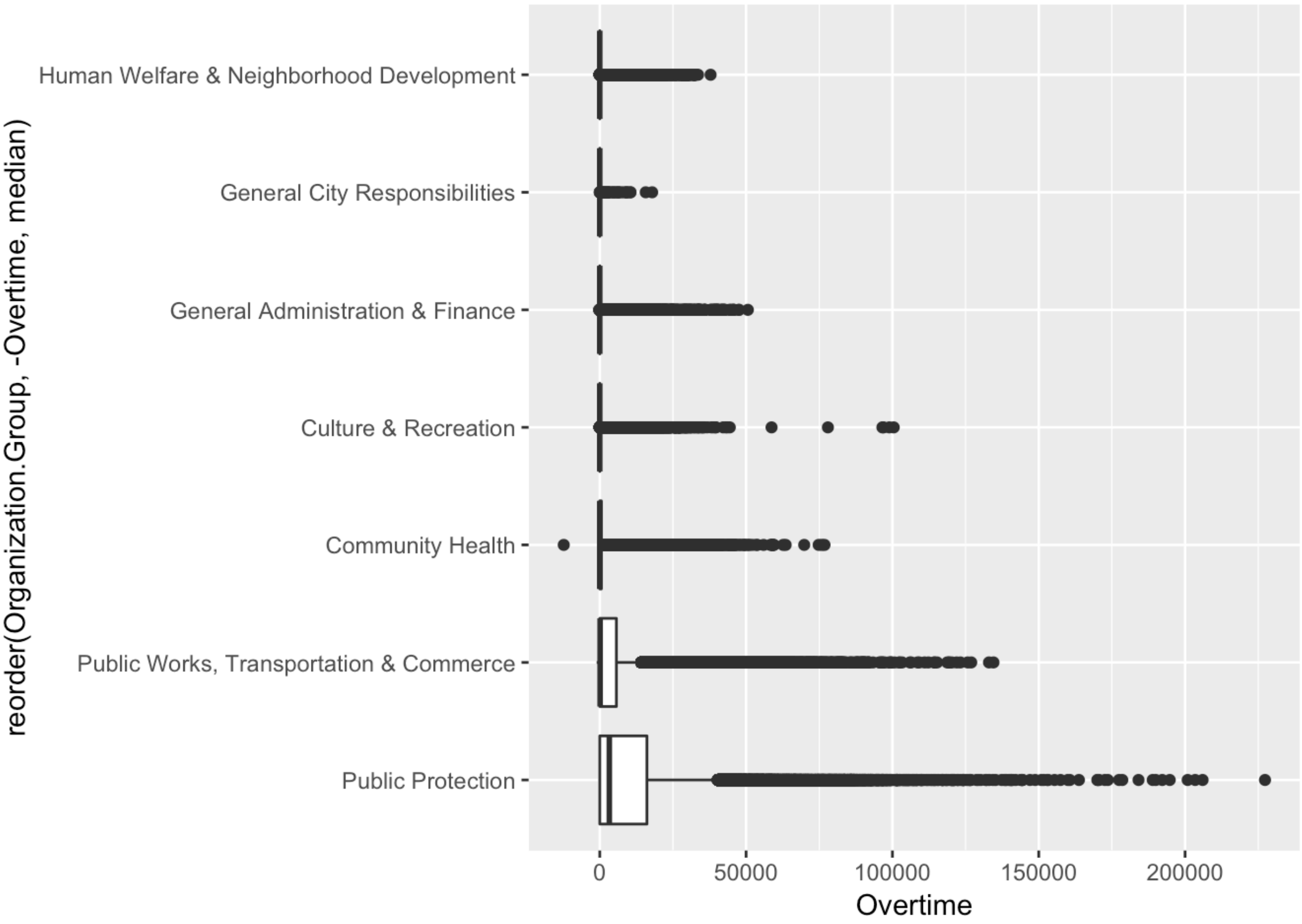
The graphs give us an overview of salary distribution through a period of 5 years. Questions that can be answered by the graphs include which year saw the greatest rise or drop in average salary, which salary bracket did most people fall into, is the data clean or not(it wasn't, as evidenced by presence of negative values), which years had the most outliers?

2. Overtime

[10 points]

a) Draw multiple horizontal boxplots, grouped by `Organization Group` for the `Overtime` variable in *Employee.csv* The boxplots should be sorted by group median. Why aren’t the boxplots particularly useful?

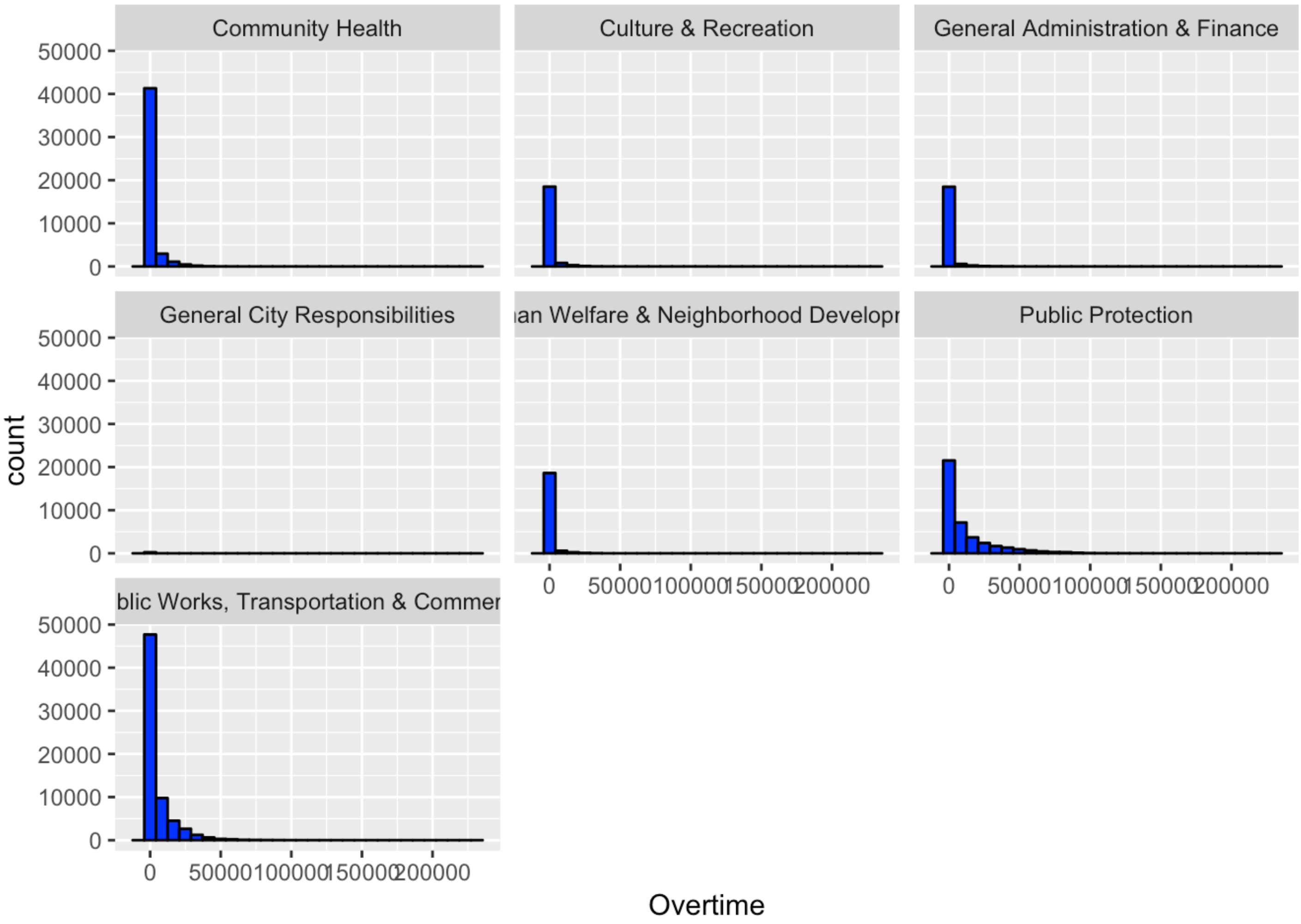
```
p3<-ggplot(data,aes(reorder(Organization.Group,-Overtime,median),Overtime))

p3+geom_boxplot()+coord_flip()
```



Because most of the values for Overtime are zero, The fields which have non zero values appear as outliers as they lie outside 1.5 times the Inter-Quartile Range, and most of the boxplots have a mean of zero.

b) Either subset the data or choose another graphical form (or both) to display the distributions of `Overtime` by `Organization Group` in a more meaningful way. Explain how this form improves on the plots in part a).

```
p4<-ggplot(data,aes(Overtime))+geom_histogram(col="black",fill="blue")
p4+facet_wrap(~Organization.Group)
```



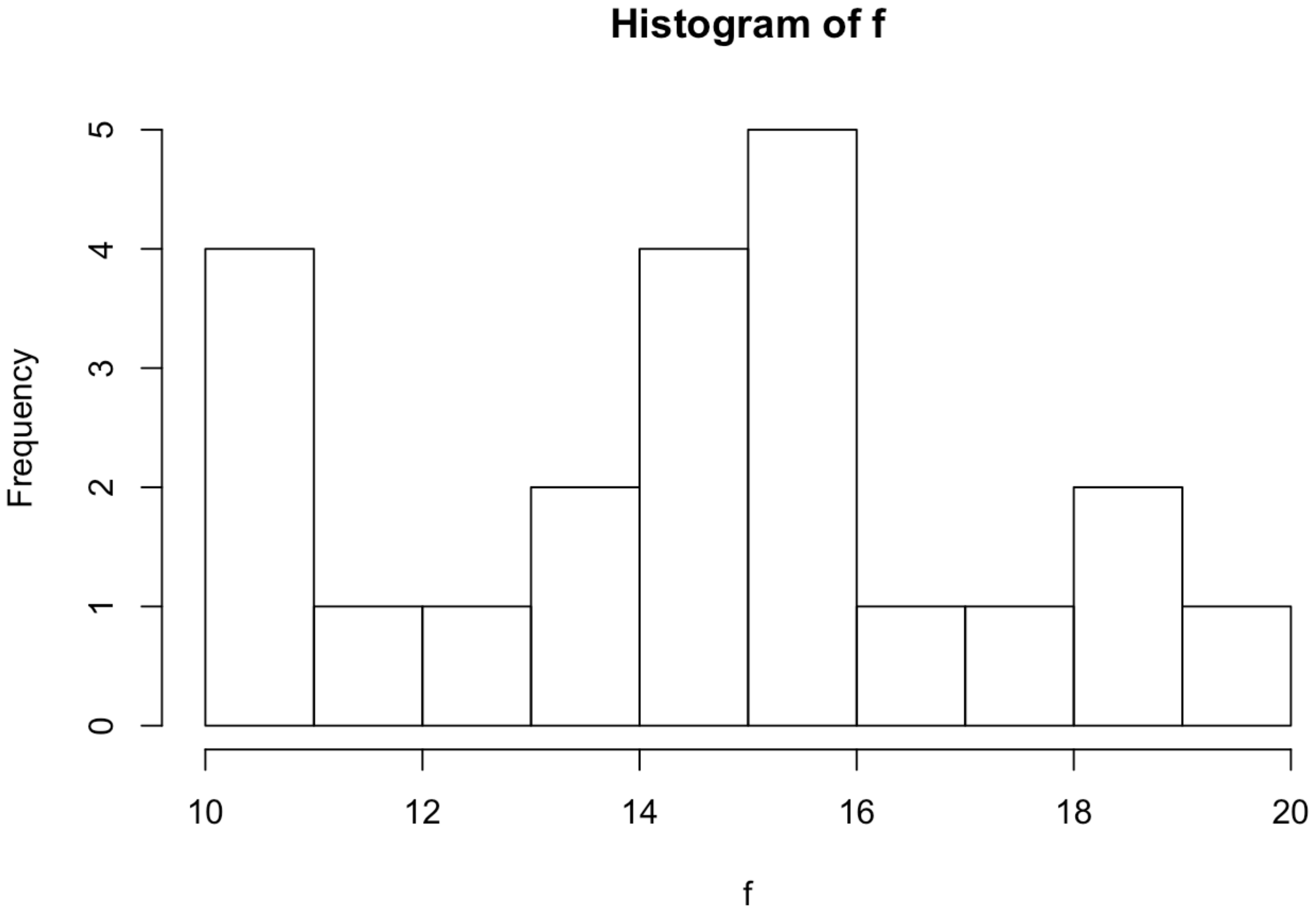
In the Histogram, we can clearly see that the mode value is zero and the overtime hours for the last two Groups are distinct, unlike in the Box Plot.

3. Boundaries

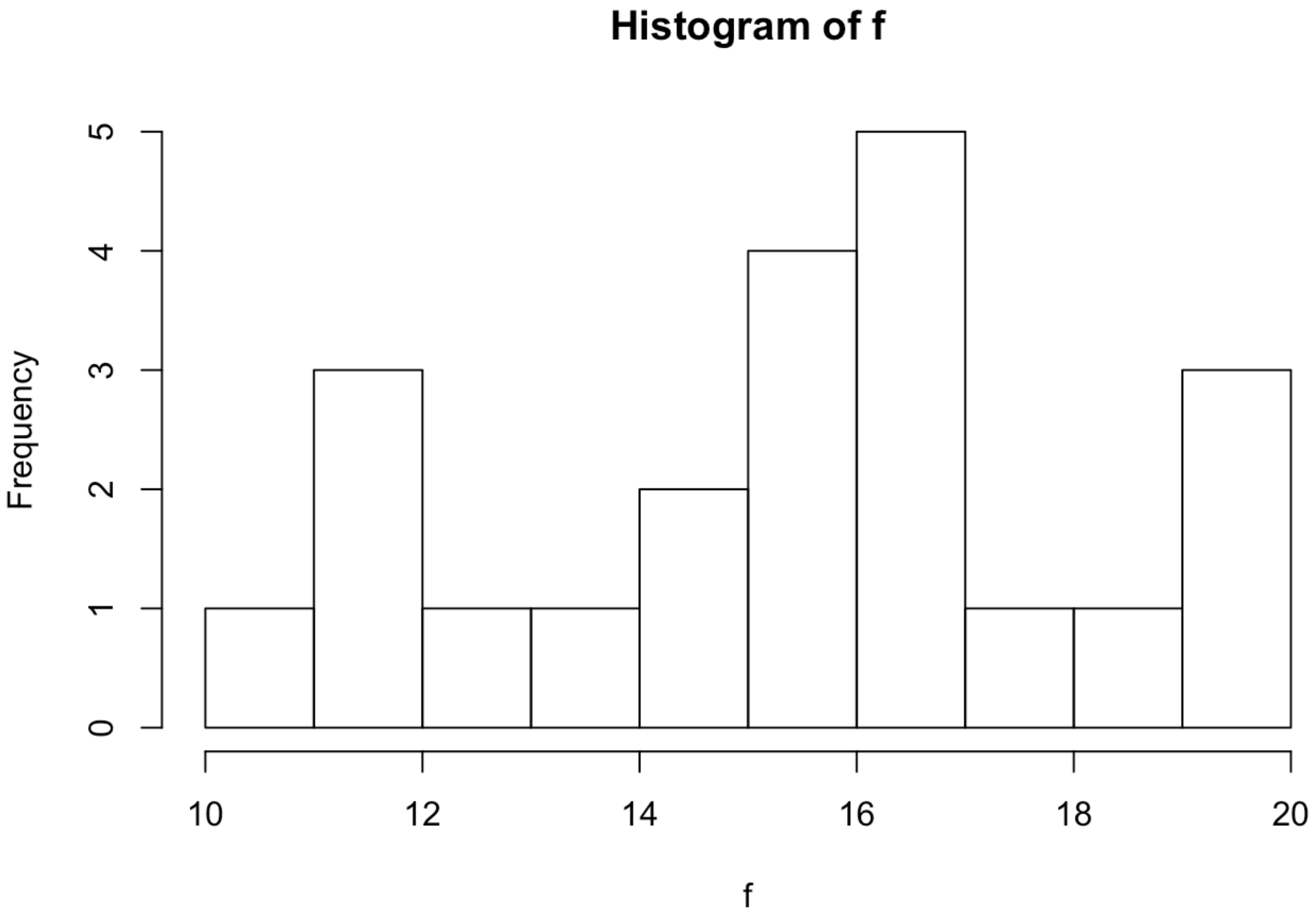
[10 points]

a) Find or create a small dataset (< 100 observations) for which right open and right closed histograms for the same parameters are not identical. Display the full dataset (that is, show the numbers) and the plots of the two forms.

```
f<-c(10,11,11,11,12,13,14,14,15,15,15,15,16,16,16,16,16,17,18,19,19,20)
hist(f,breaks=seq(10,20,1),right=TRUE)
```

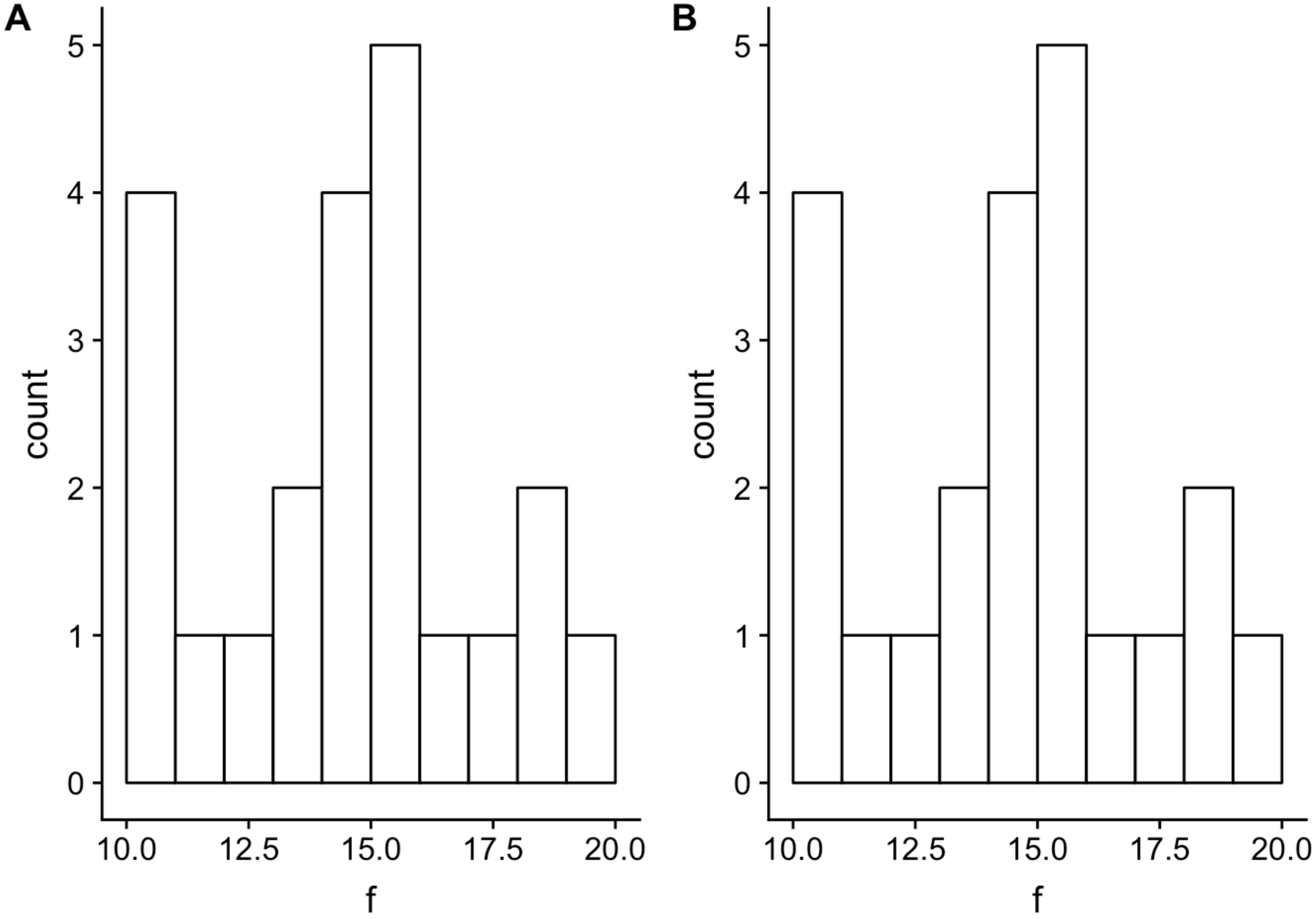


```
hist(f,breaks=seq(10,20,1),right=FALSE)
```



b) Adjust parameters–the same for both–so that the right open and right closed versions become identical. Explain your strategy.

```
x<-data.frame(f)
g1<-ggplot(data=x,aes(f))+geom_histogram(binwidth=1,boundary=1,color="black",fill="wh
ite")
g2<-ggplot(data=x,aes(f))+geom_histogram(binwidth=1,boundary=0,color="black",fill="wh
ite")
library(cowplot)
plot_grid(g1, g2, labels = "AUTO")
```



Thus instead of using base-R we can use ggplot() function to plot data such that we can set the binwidth to an appropriate value leading to similar boxplots.

4. Beavers

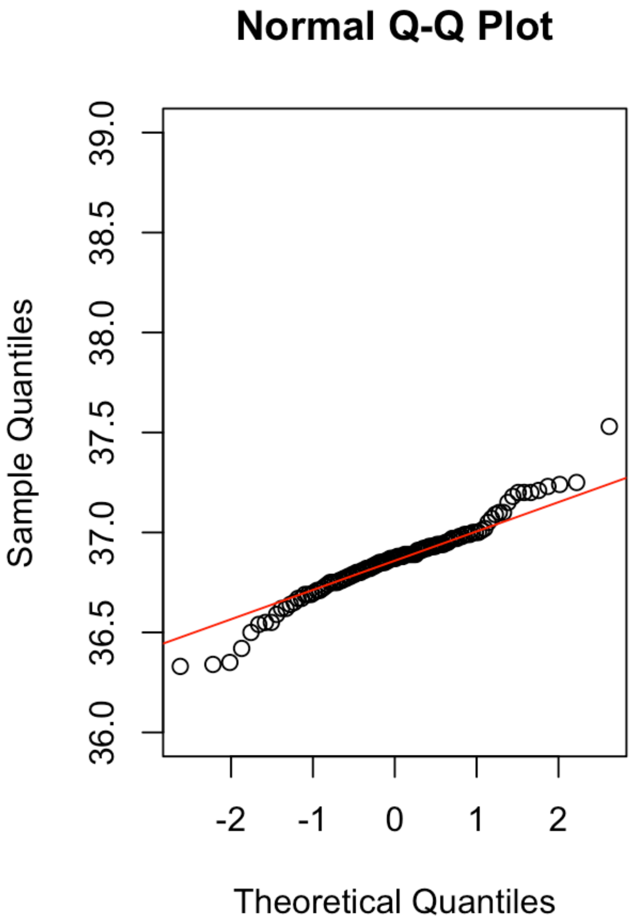
[10 points]

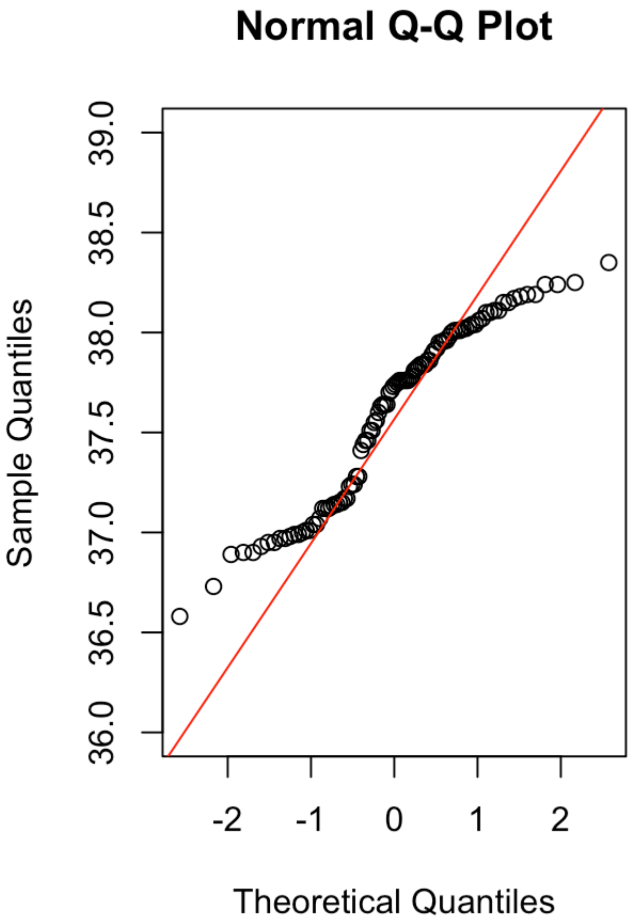
a) Use QQ (quantile-quantile) plots with theoretical normal lines to compare `temp` for the built-in *beaver1* and *beaver2* datasets. Which appears to be more normally distributed?

```
library(datasets)
data(beavers)
head(beaver1)
```

```
##   day time  temp activ
## 1 346  840 36.33     0
## 2 346  850 36.34     0
## 3 346  900 36.35     0
## 4 346  910 36.42     0
## 5 346  920 36.55     0
## 6 346  930 36.69     0
```

```
x<-beaver1$temp
y<-beaver2$temp
par(mfrow=c(1,2))
qqnorm(x,ylim=c(36,39))
qqline(x,col="red")
qqnorm(y,ylim=c(36,39))
qqline(y,col="red")
```

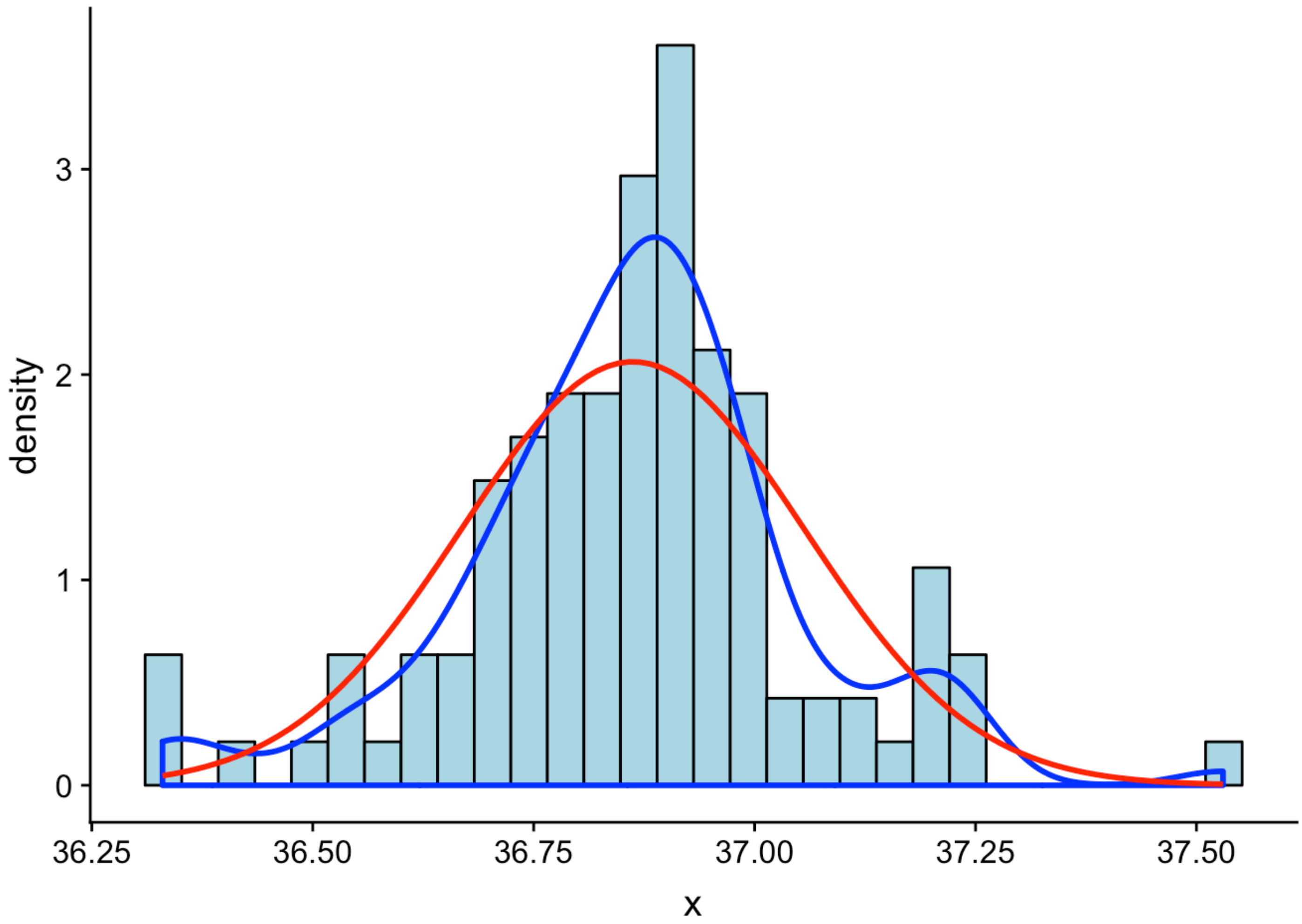




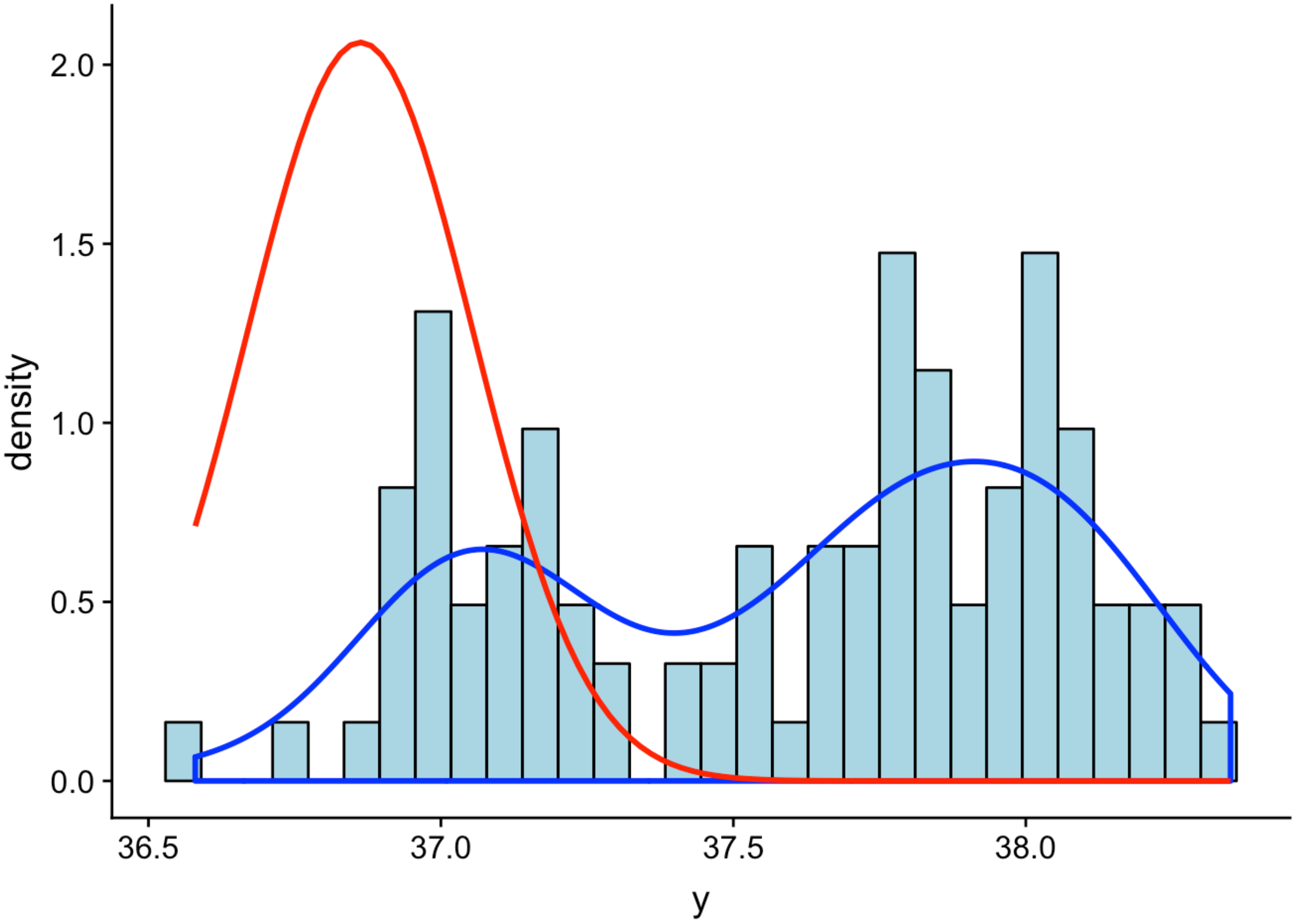
From the Q-Q plots, the temp variable in beaver1 appears to be more normally distributed

b) Draw density histograms with density curves and theoretical normal curves overlaid. Do you get the same results as in part a)?

```
p5 <- ggplot(beaver1, aes(x)) + geom_histogram(aes(y = ..density..),fill = "lightblue
",color ="black")
p5+geom_density(lwd=1,color="blue")+stat_function(fun=dnorm,args=list(mean=mean(x),sd
=sd(x)),color="red",lwd=1)
```



```
p6 <- ggplot(beaver2, aes(y)) + geom_histogram(aes(y = ..density..),fill = "lightblue
",color ="black")
p6+geom_density(lwd=1,color="blue")+stat_function(fun=dnorm,args=list(mean=mean(x),sd
=sd(x)),color="red",lwd=1)
```



After drawing the histograms with density curves and theoretical normal curves overlaid, we still get the same hypothesis as a). That is, temp of beaver1 appears to be more normally distributed than temp of beaver2

c) Perform the Shapiro-Wilk test for normality using the `shapiro.test()` function. How do the results compare to parts a) and b)?

```
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.97031, p-value = 0.01226
```

```
shapiro.test(y)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  y
## W = 0.93336, p-value = 7.764e-05
```

According to the tests, the Null hypothesis(i.e. data comes from a normal distribution) is rejected in both cases. From the Q-Q plot for beaver1, it is obvious that data isn’t completely normal for the lower and higher values. Thus neither data is normally distributed, but temp of beaver1 is more normal than temp of beaver2

5. Doctors

[5 points]

Draw two histograms of the number of deaths attributed to coronary artery disease among doctors in the *breslow* dataset (**boot package**), one for smokers and one for non-smokers. *Hint: read the help file ?breslow to understand the data.*

```
library(boot)
data(breslow)
breslow
```

```
##    age smoke     n   y    ns
## 1   40     0 18790   2     0
## 2   50     0 10673  12     0
## 3   60     0  5710  28     0
## 4   70     0  2585  28     0
## 5   80     0  1462  31     0
## 6   40     1 52407  32 52407
## 7   50     1 43248 104 43248
## 8   60     1 28612 206 28612
## 9   70     1 12663 186 12663
## 10  80     1  5317 102  5317
```

```
??breslow
sm<-breslow[,4][ breslow[,2] <1]
sm
```

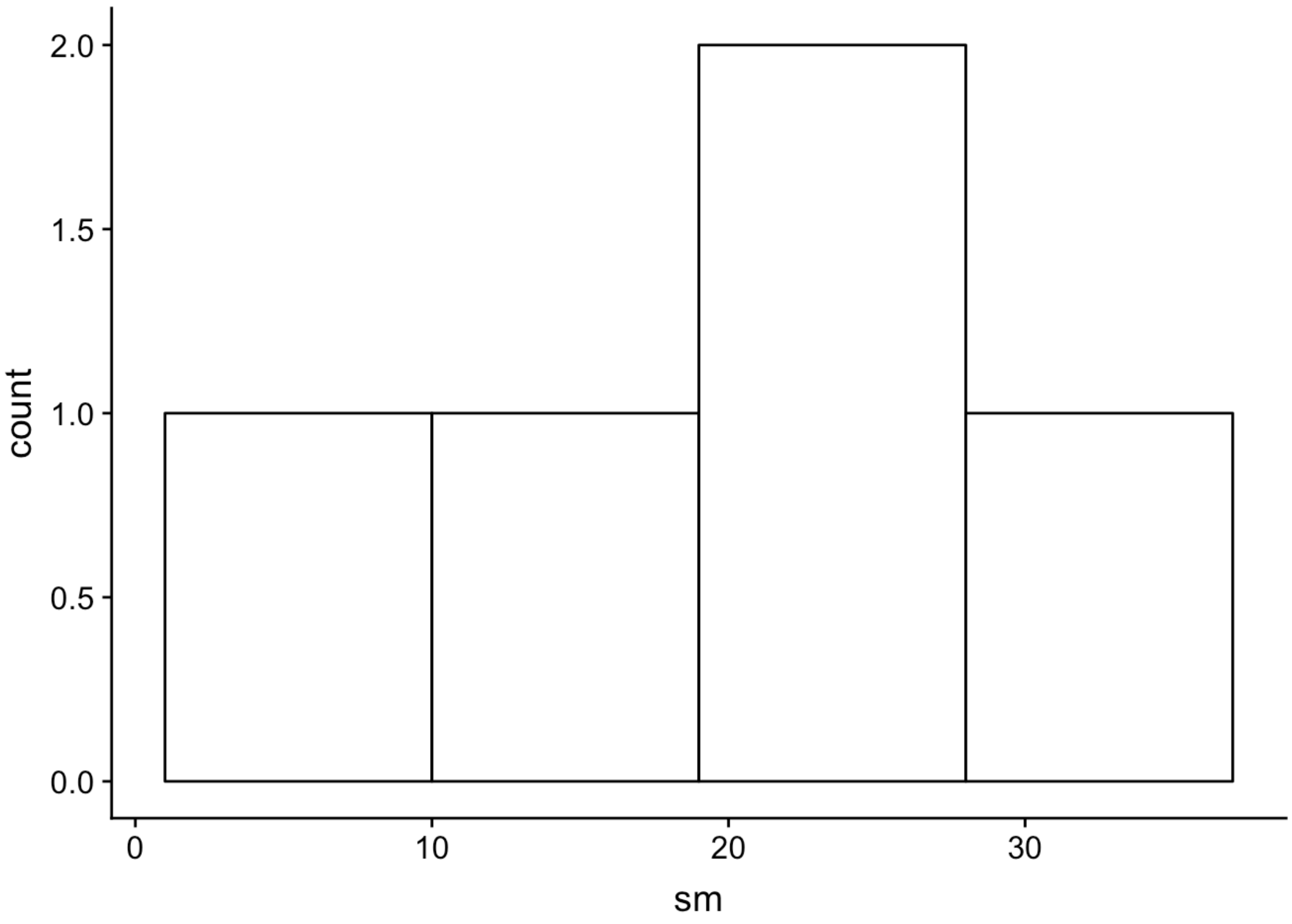
```
## [1]  2 12 28 28 31
```

```
nonsm<-breslow[,4][ breslow[,2] >0]
nonsm
```

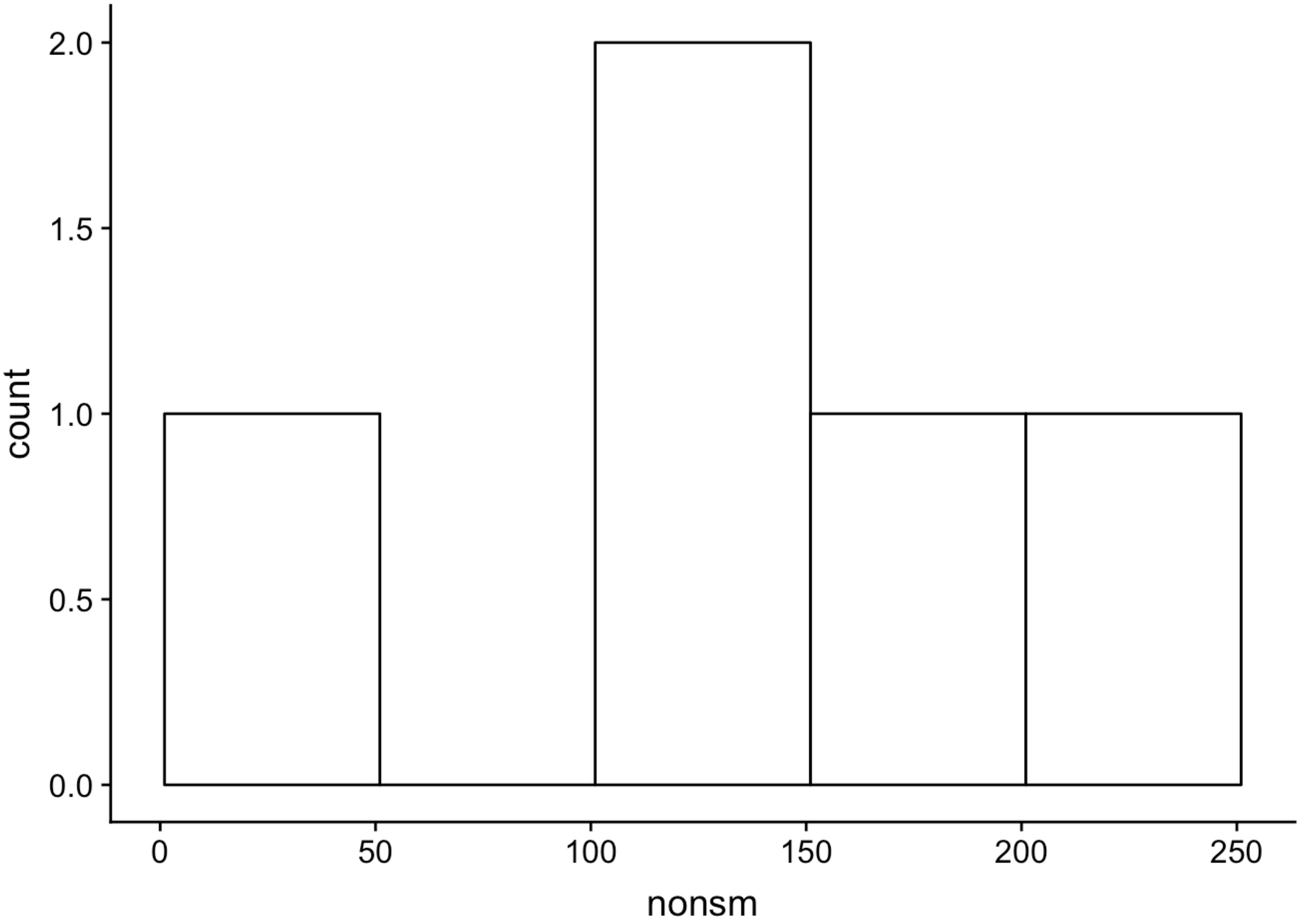
```
## [1]  32 104 206 186 102
```

We have thus filtered out the smokers and nonsmokers

```
s<-data.frame(sm)
non<-data.frame(nonsm)

p7<-ggplot(data=s,aes(sm))+geom_histogram(binwidth=9,boundary=1,color="black",fill="w
hite")
p7
```



```
p8<-ggplot(data=non,aes(nonsm))+geom_histogram(binwidth=50,boundary=1,color="black",f
ill="white")
p8
```


2.0
1.5
1.0
0.5
0.0
count
0
50
100
150
200
250
nonsm