Homework 1

Xiaowei Zhao

9/21/2018

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

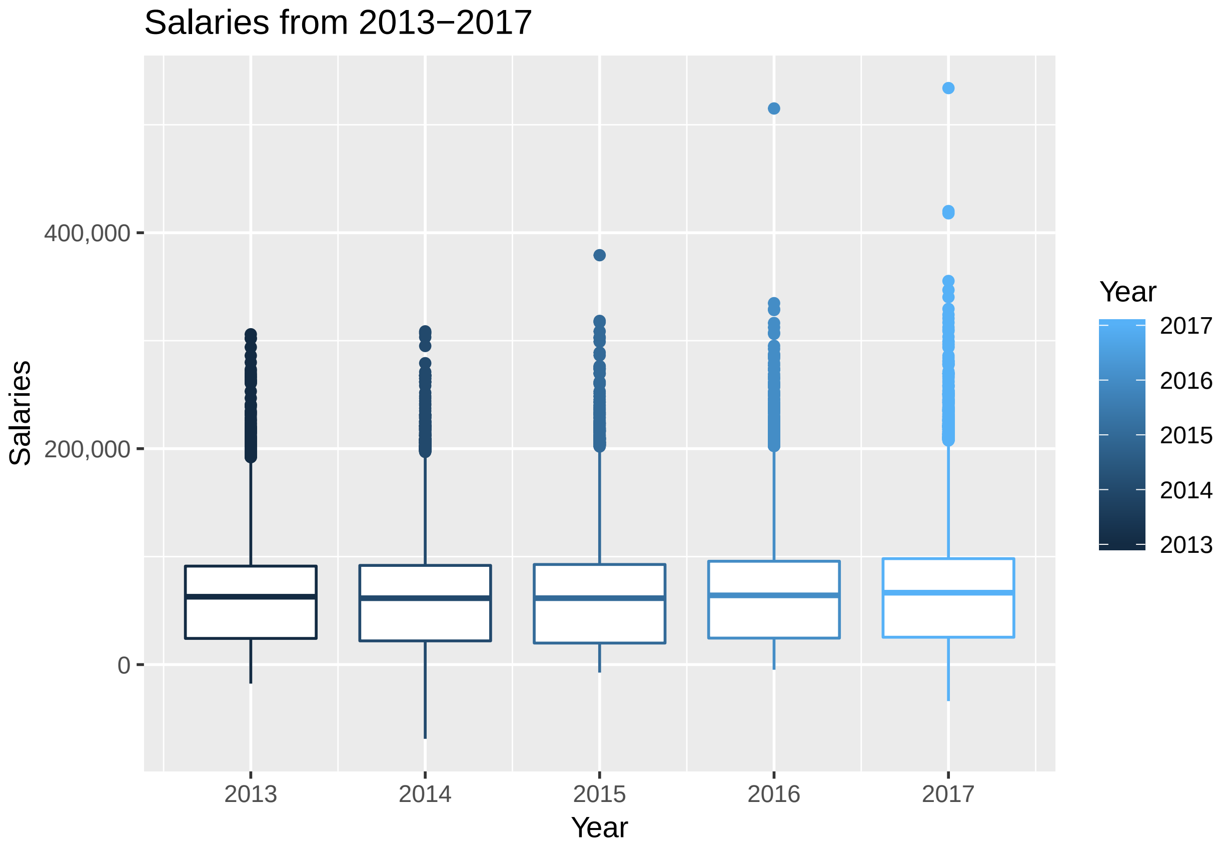
Read *Graphical Data Analysis with R*, Ch. 3

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). How do the distributions differ by year?

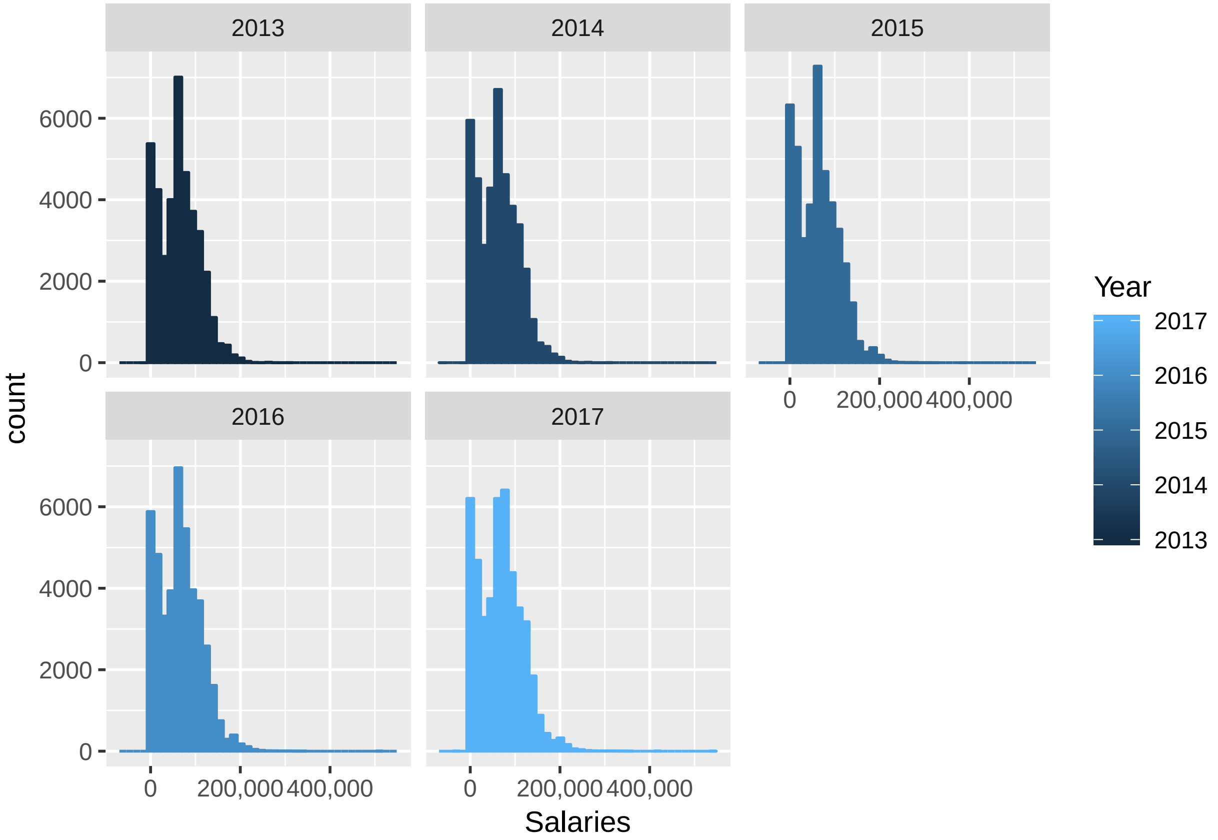
```
data=read.csv('employee.csv')
p1a=ggplot(data,aes(x=Year,y=Salaries,group=Year,col=Year)) +
  geom_boxplot() +
  scale_y_continuous(labels=comma) +
  ggtitle('Salaries from 2013-2017 ')
p1a
```



Actually, the median of salaries for each year has little difference, hence their distributions are similar.

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

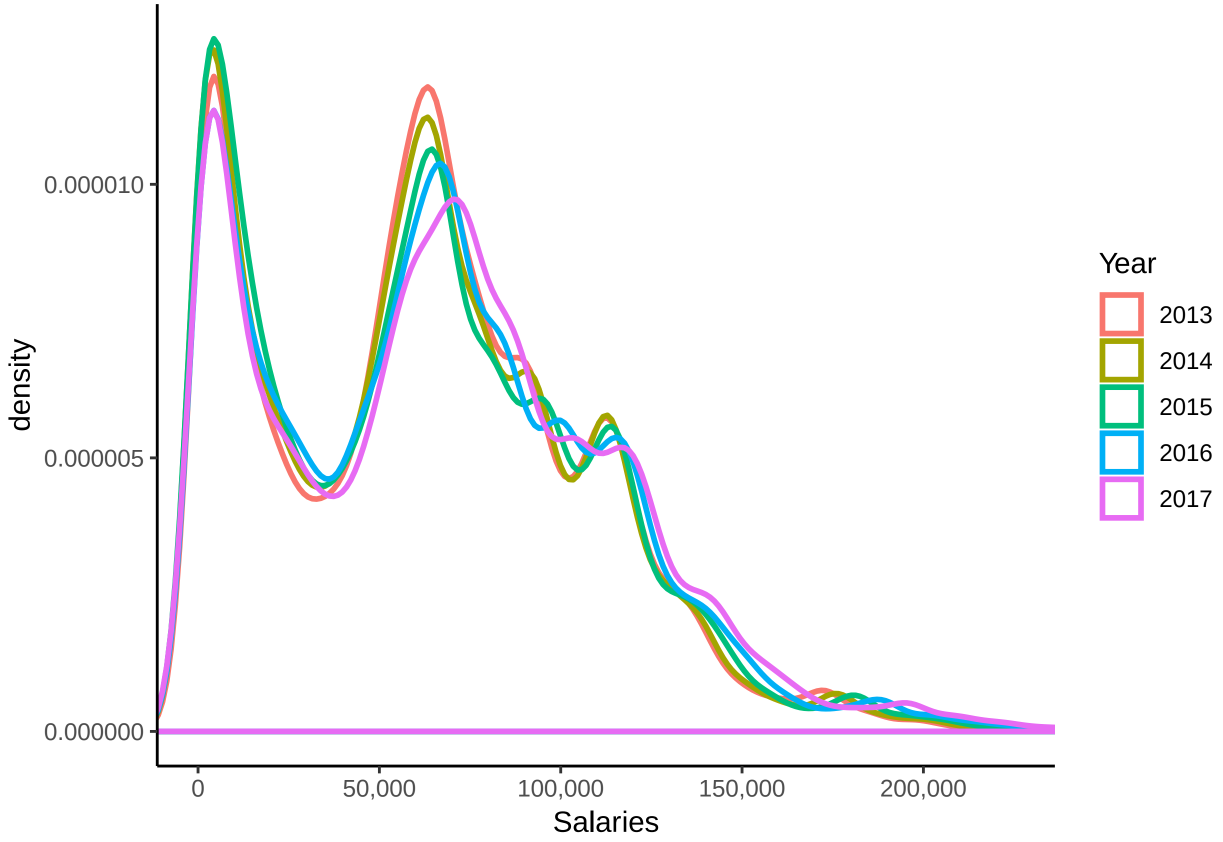
```
p1b=ggplot(data,aes(x=Salaries,col=Year,fill=Year)) +
  geom_histogram(bins=40) +
  scale_x_continuous(labels=comma) +
  facet_wrap(~Year)
p1b
```



The histogram provides us the detail of data spread. From the graph above, we can easily capture the exact and specific salary distribution from 2013 to 2017.

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?

```
p1c=ggplot(data,aes(x=Salaries,col=Year)) +
  geom_density(aes(col=factor(Year)),size=1) +
  scale_y_continuous(labels = comma) +
  scale_x_continuous(labels=comma) +
  coord_cartesian(xlim=c(0,225000)) +
  theme_classic()
p1c
```



The density curve for each year are kind of similar. The density curve can shows probability. As we can see in the graph, they are right-skewed. To make it more clear, I use coord_cartesian function to zoom in. There are two peaks in the graph in the intervals (0,50000) and (50000,100000). And also, the salary level are increasing year by year. The salary level of 2017 is rightmost.

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?

1. How do the distribution of salaries change year by year?

2. In each year, the probability of different salary level.

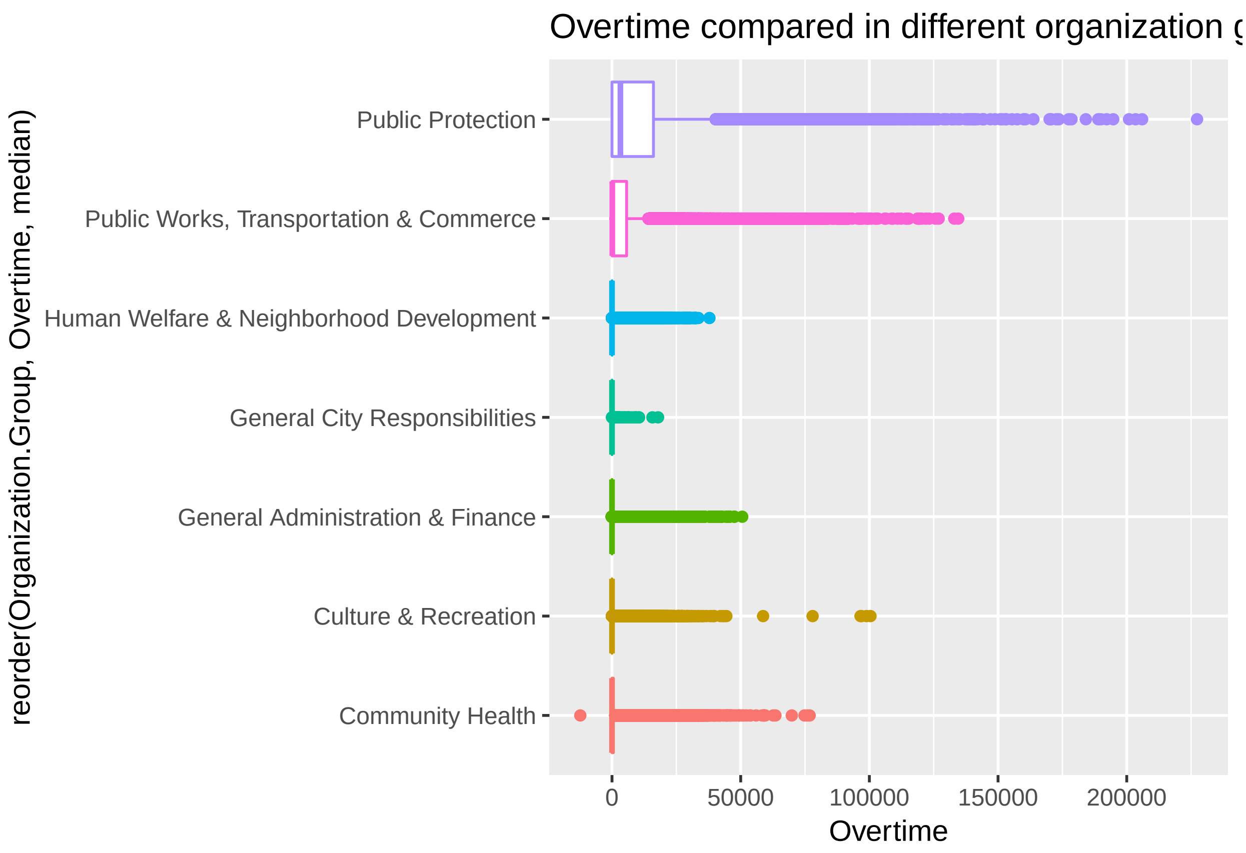
3. Some statistical data about the salary.

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?

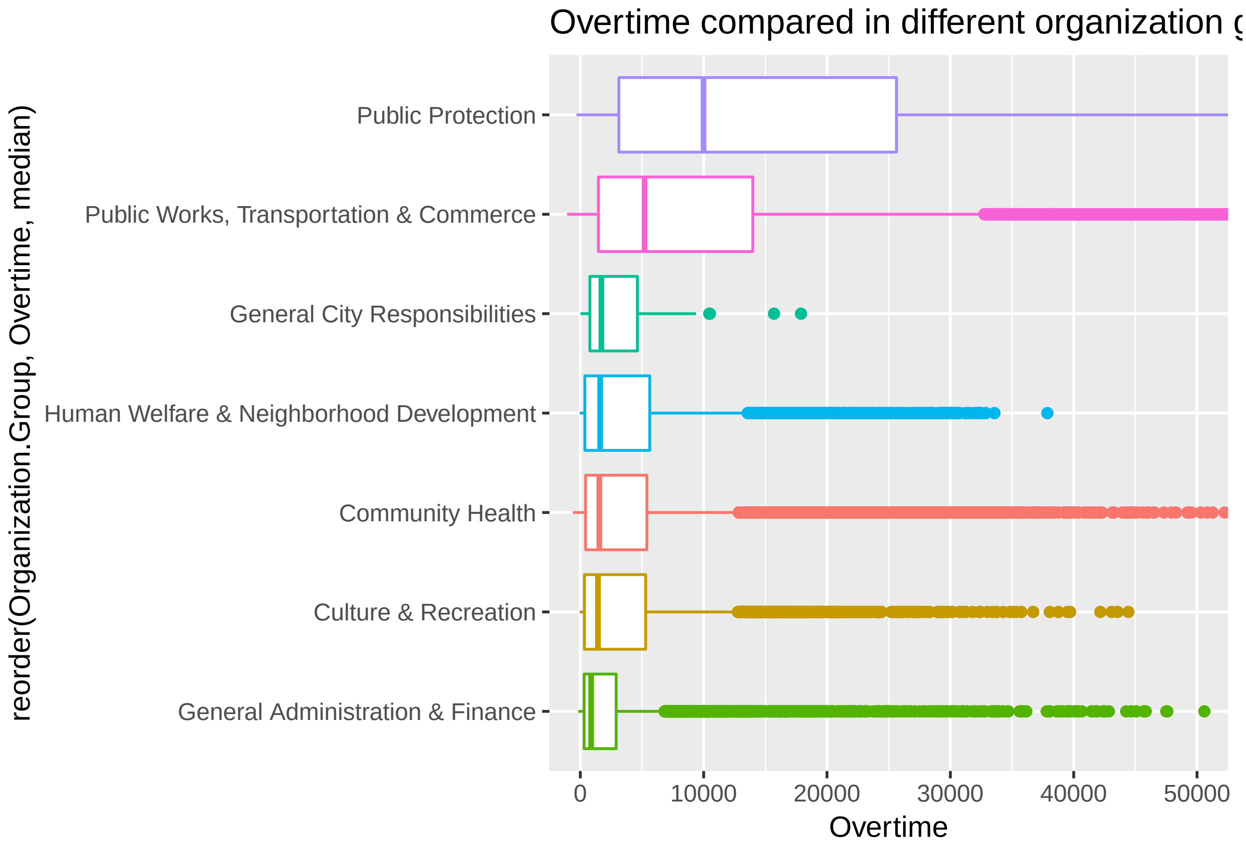
```
p2a=ggplot(data,aes(x= reorder(Organization.Group, Overtime, median),y=Overtime,group=Organization.Grou
  geom_boxplot(show.legend = FALSE) +
  ggtitle('Overtime compared in different organization group') +
  coord_flip()
p2a
```



The boxplots are not usefull here since there are too much zero in our data. Hence, the hinges of boxplot are mostly 0 here.

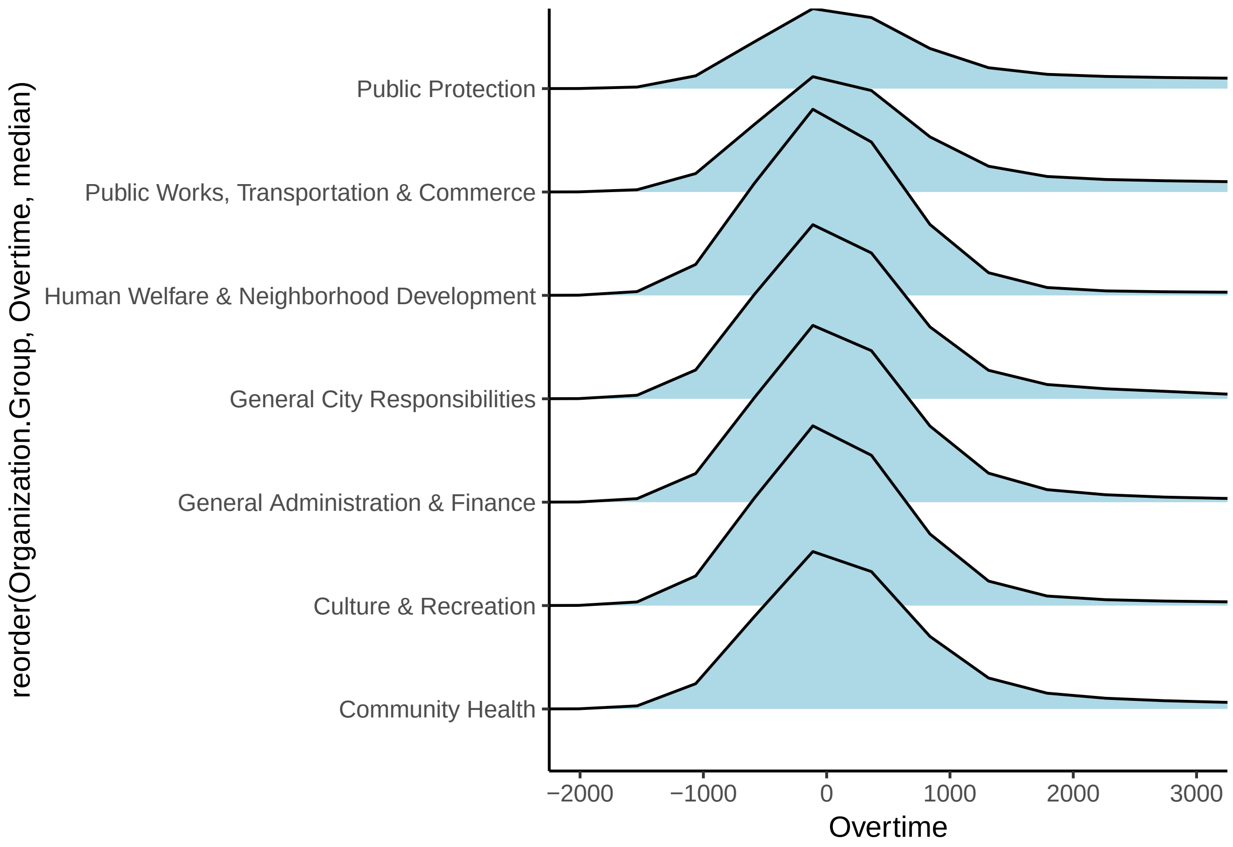
b) Either subset the data or choose another graphical form (or both) to display the distributions of `Overtime` by `Organization Group` in a more meaningful way. Explain how this form improves on the plots in part a).

```
tempdata=tbl_df(data)
tempdata = filter(tempdata,Overtime!=0)
p2b1=ggplot(tempdata,aes(x= reorder(Organization.Group, Overtime, median),y=Overtime,group=Organization
  geom_boxplot(show.legend = FALSE) +
  ggtitle('Overtime compared in different organization group') +
  coord_flip(ylim = c(0,50000))
p2b1
```



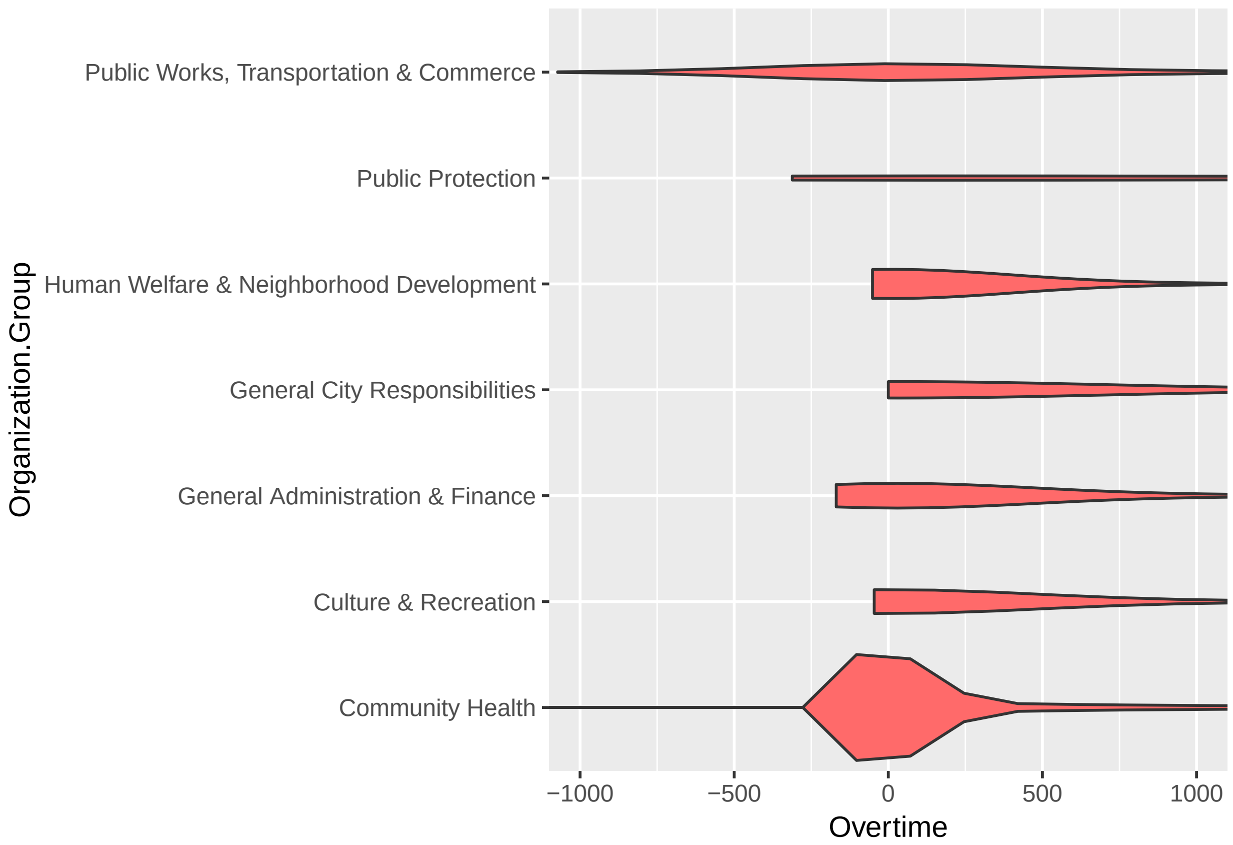
For this plot, I filter the row without 0 to make a new dataframe. As we can see here, the three hinges in the boxplot are clear. I zoom in the data to observe and compare it more comfortably. We can easily tell how their quantiles differ here.

```
p2b2=ggplot(data, aes(x = Overtime, y = reorder(Organization.Group, Overtime, median))) +
  geom_density_ridges(fill='lightblue') +
  coord_cartesian(xlim=c(-2000,3000)) +
  theme_classic()
p2b2
```



For this plot, I use ggridges package to draw a geom__density__ridges graph from original data. It first estimates data densities and then draws those using ridgelines. We can clearly observe the distribution here.

```
p2b22=ggplot(data,aes(x=Organization.Group,y=Overtime)) +
  geom_violin(width=1,fill='indianred1') +
  coord_flip(ylim=c(-1000,1000))
p2b22
```



For this plot, I draw violin plot from original data. It allows us to visualize the distribution of overtime.

3. Boundaries

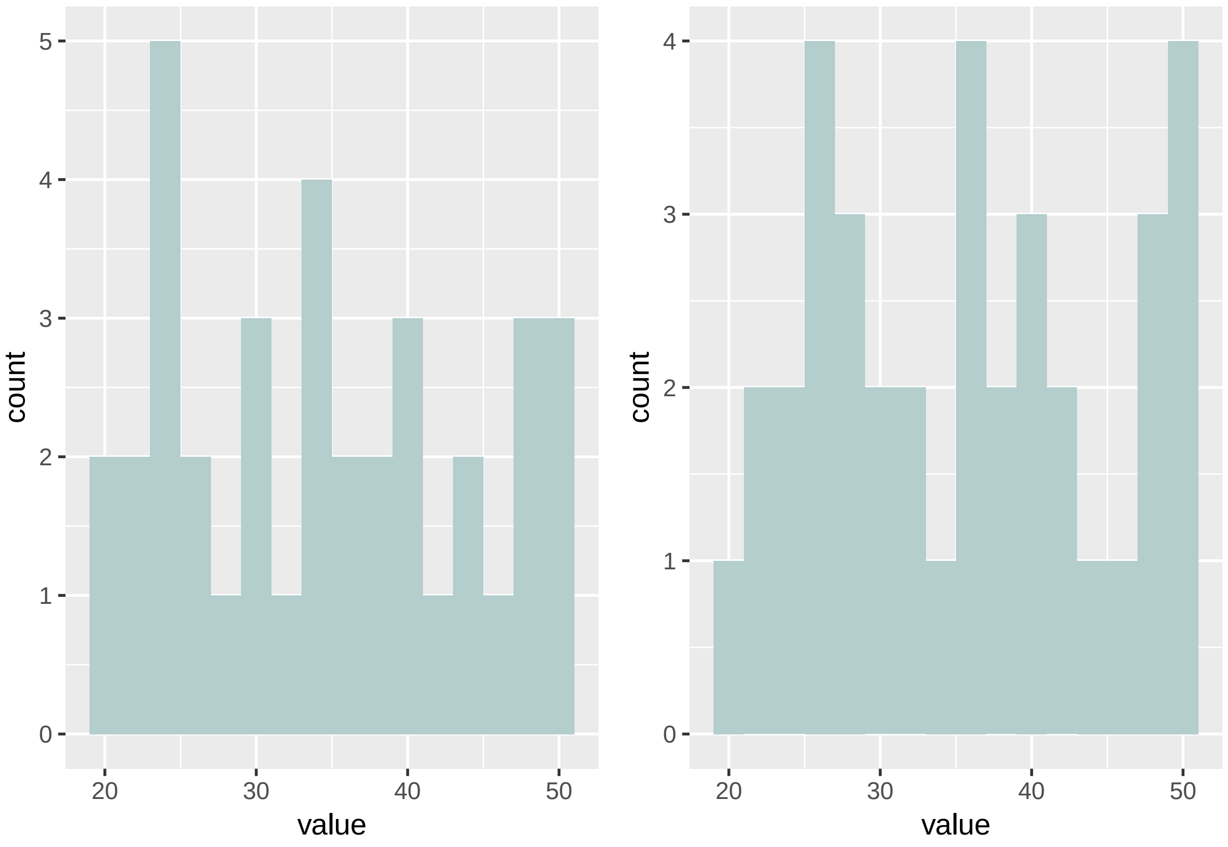
[10 points]

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.

```
mydata=tbl_df(c(20,21,22,23,24,25,25,25,25,27,27,28,30,30,31,32,34,35,35,35,36,37,38,39,40,40,41,42,44,
mybin=2
mydata
```

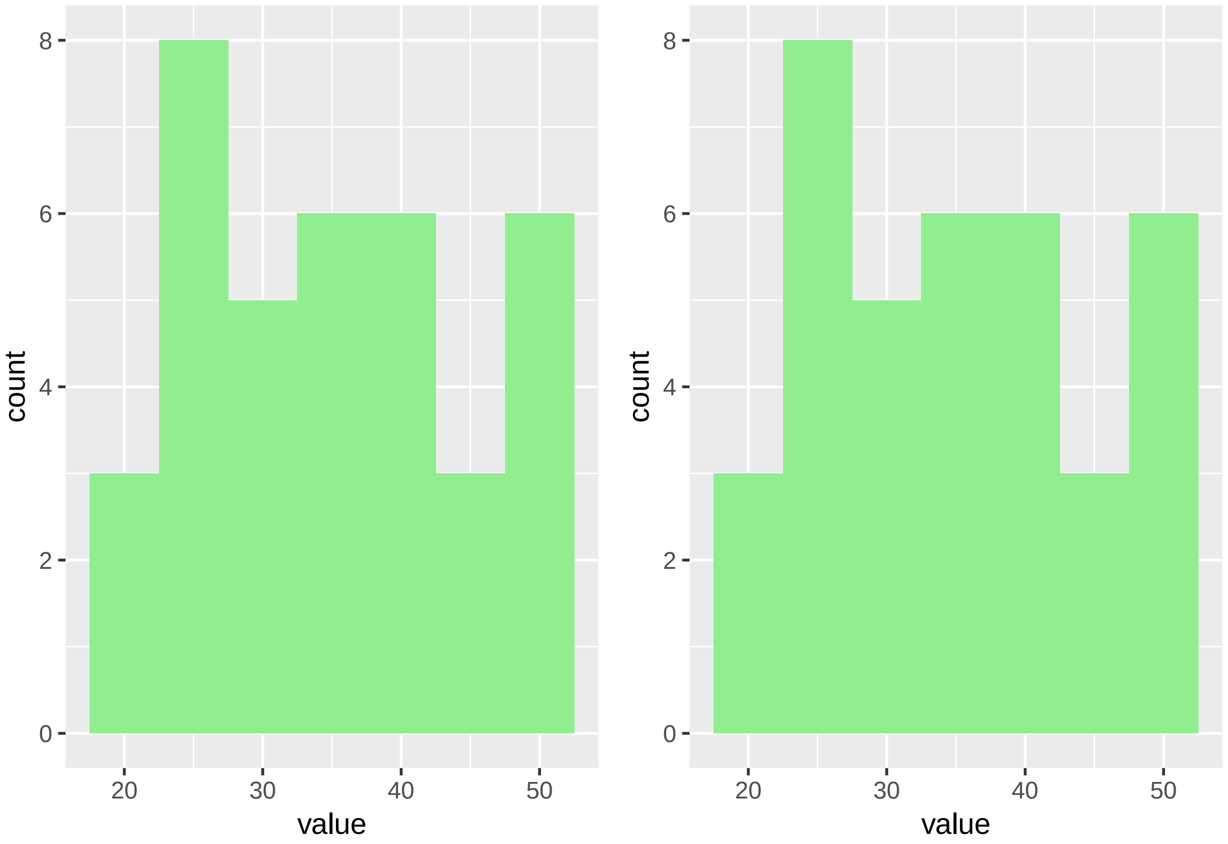
```
## # A tibble: 37 x 1
##    value
##    <dbl>
##  1    20
##  2    21
##  3    22
##  4    23
##  5    24
##  6    25
##  7    25
##  8    25
##  9    25
## 10    27
## # ... with 27 more rows
```

```
p3a1=ggplot(mydata,aes(x=value)) +
  geom_histogram(binwidth = mybin,fill='lightcyan3',right=TRUE)

p3a2=ggplot(mydata,aes(x=value)) +
  geom_histogram(binwidth = mybin,fill='lightcyan3',right=FALSE)

grid.arrange(p3a1, p3a2, nrow = 1)
```



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

```
mybin=5
p3a1=ggplot(mydata,aes(x=value)) +
  geom_histogram(binwidth = mybin,fill='lightgreen',right=TRUE)

p3a2=ggplot(mydata,aes(x=value)) +
  geom_histogram(binwidth = mybin,fill='lightgreen',right=FALSE)

grid.arrange(p3a1, p3a2, nrow = 1)
```



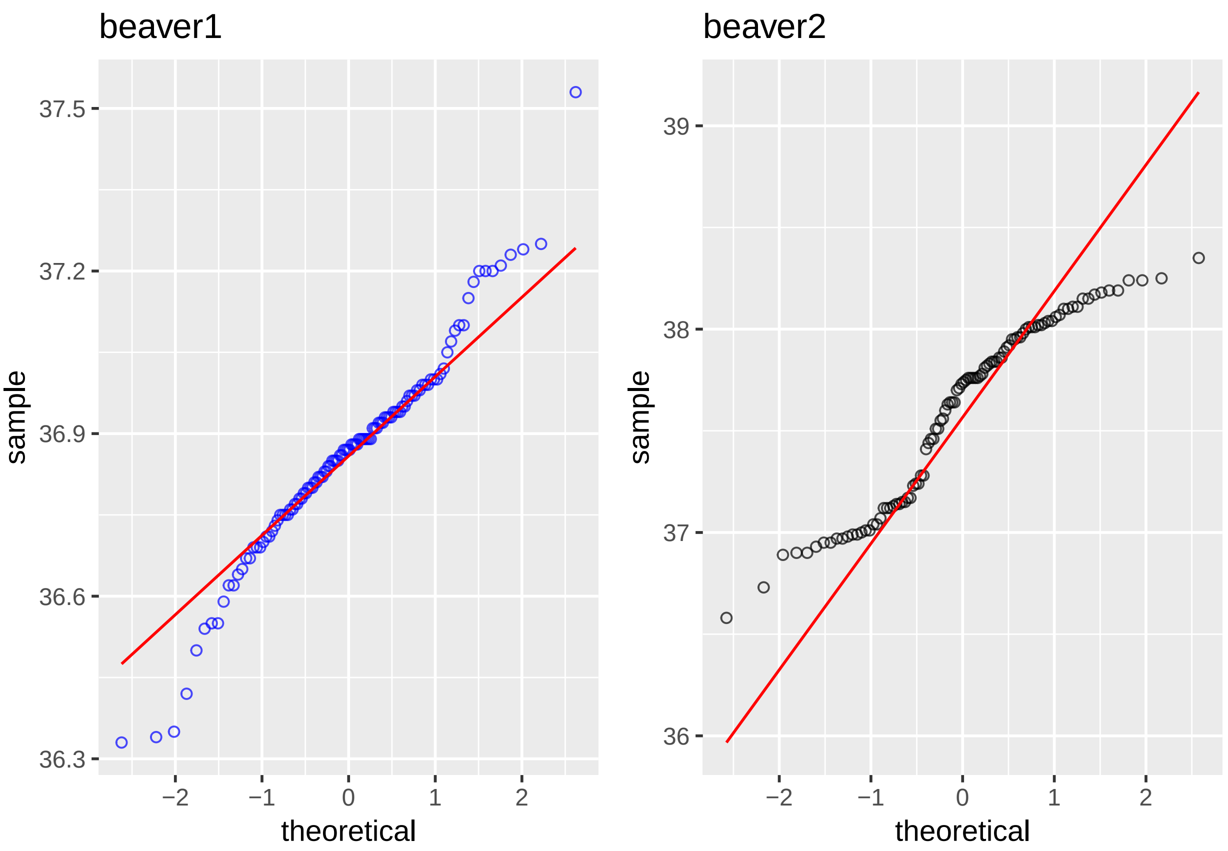
The trick here is the binwidth. In my data, there are some data just on the right boundary with closed interval and binwidth equals 2. When I change the binwidth to 5, it eliminates differences between two histograms.

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?

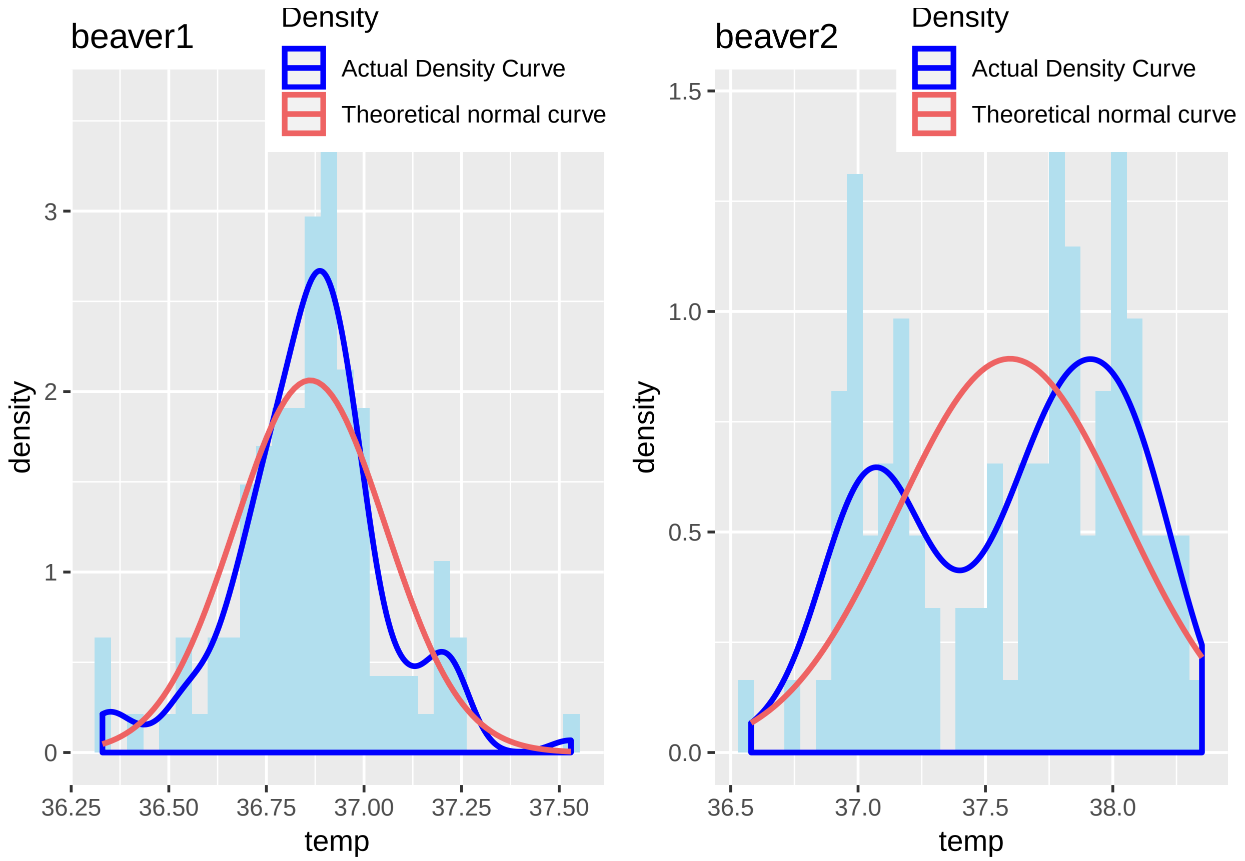
```r
p4a1=ggplot(beaver1,aes(sample=temp)) +
  geom_qq(shape=1,alpha=0.7,col='blue') +
  geom_qq_line(col='red') +
  ggtitle('beaver1')

p4a2=ggplot(beaver2,aes(sample=temp)) +
  geom_qq(shape=1,alpha=0.7) +
  geom_qq_line(col='red') +
  ggtitle('beaver2')

grid.arrange(p4a1, p4a2, nrow = 1)
```



‘temp’ for beaver1 appears to be more normally distributed than one for beaver2

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

```
p4b1=ggplot(beaver1,aes(x=temp)) +
  geom_histogram(aes(y=..density..),fill='lightblue2') +
  geom_density(aes(color='Actual Density Curve'),size=1) +
  stat_function(aes(color='Theoretical normal curve'),fun=dnorm,geom='line',args=list(mean=mean(beaver1
  ggtitle('beaver1') +
  scale_colour_manual(name="Density", breaks=c('Actual Density Curve','Theoretical normal curve'),value
  theme(legend.position=c(0.7,1))

p4b2=ggplot(beaver2,aes(x=temp)) +
  geom_histogram(aes(y=..density..),fill='lightblue2') +
  geom_density(aes(color='Actual Density Curve'),size=1) +
  stat_function(aes(color='Theoretical normal curve'),fun=dnorm,geom='line',args=list(mean=mean(beaver2
  ggtitle('beaver2') +
  scale_colour_manual(name="Density", breaks=c('Actual Density Curve','Theoretical normal curve'),value
  theme(legend.position=c(0.7,1))

grid.arrange(p4b1, p4b2, nrow = 1)
```



Yes. As we can see from the two graphs above, 'temp' for beaver1 appears to be more normally distributed than one for 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(beaver1$temp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  beaver1$temp
## W = 0.97031, p-value = 0.01226
```

```
shapiro.test(beaver2$temp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  beaver2$temp
## W = 0.93336, p-value = 7.764e-05
```

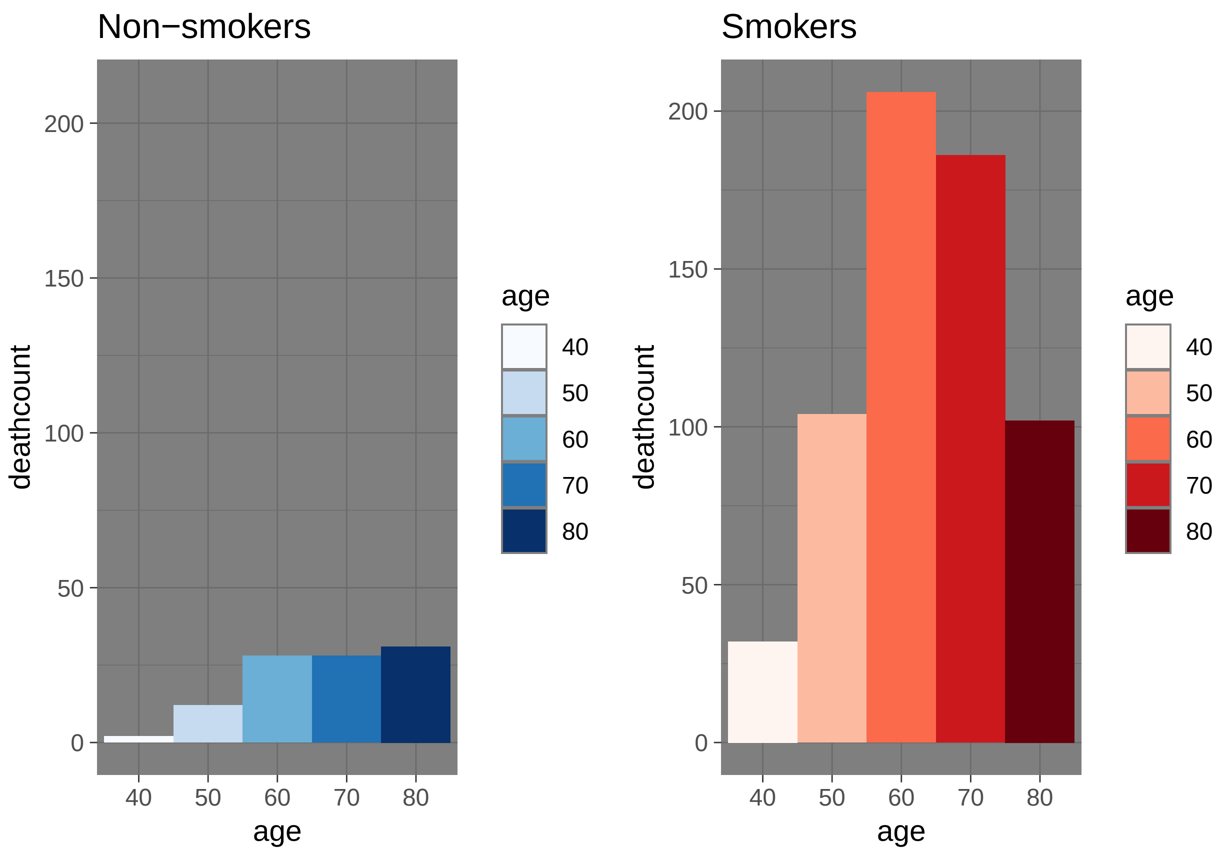
The results are consistent with the parts a and b since the P-value of 'temp' for beaver1 is greater than 'temp' for beaver2, hence, the 'temp' for beaver1 is more normally distributed. However, from the output, both the p-value < 0.05 implying that the distribution of the 'temp' for beaver1 and beaver2 are significantly different from normal distribution. In other words, they are not normally distributed.

5. Doctors

[5 points]

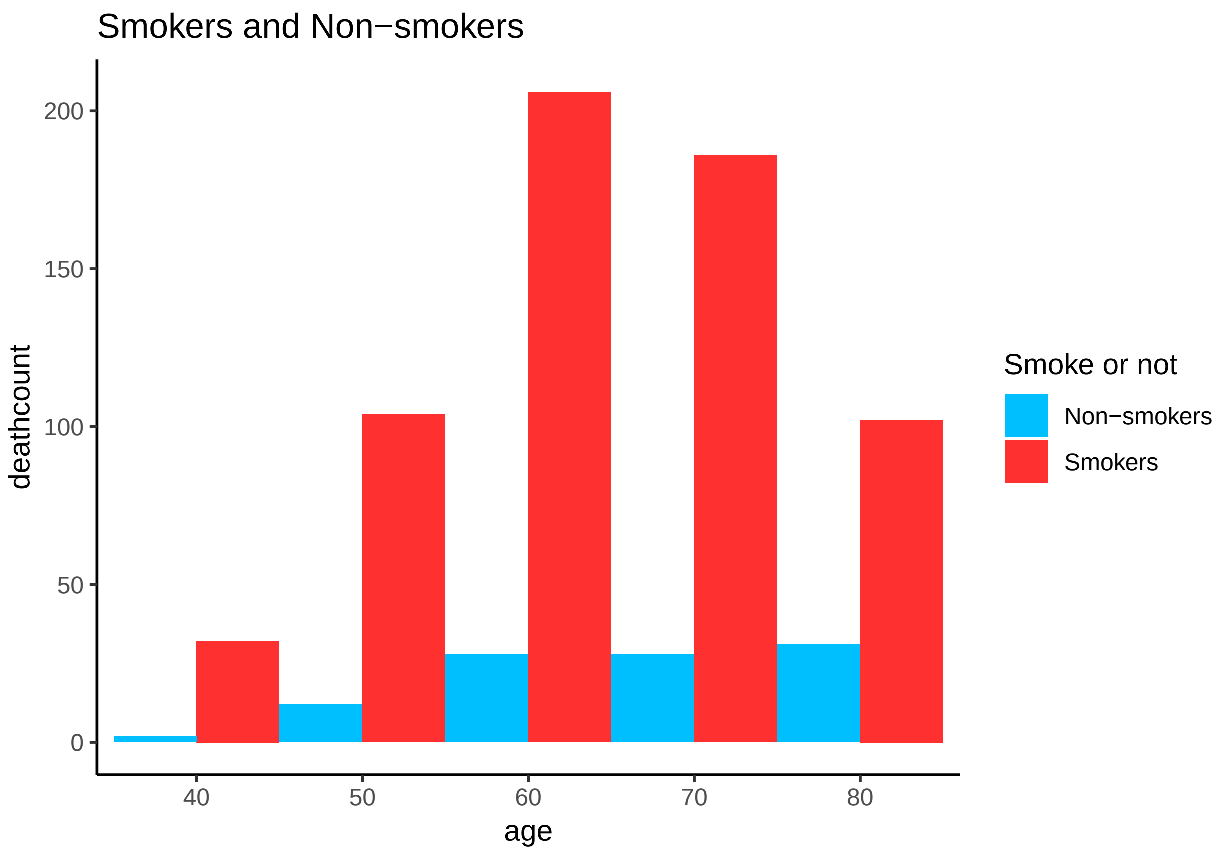
Draw two histograms of the number of deaths attributed to coronary artery disease among doctors in the *breslow* dataset (**boot** package), one for smokers and one for non-smokers. *Hint: read the help file ?breslow to understand the data.*

```r
non_smokers=tbl_df(breslow[1:5,])
non_smokers=mutate(non_smokers,deathcount=y)
smokers=tbl_df(breslow[6:10,])
smokers=mutate(smokers,deathcount=y)

blues <- brewer.pal(9, "Blues")
blue_range <- colorRampPalette(blues)
p5a=ggplot(non_smokers,aes(x=age,y=deathcount,fill=age)) +
  geom_histogram(stat="identity",width=1) +
  scale_y_continuous(limits = c(0,210))  +
  scale_fill_manual(values = blue_range(5)) +
  theme_dark() +
  ggtitle('Non-smokers')

reds <- brewer.pal(9, "Reds")
red_range <- colorRampPalette(reds)
p5b=ggplot(smokers,aes(x=age,y=deathcount,fill=age)) +
  geom_histogram(stat="identity",width=1) +
  scale_fill_manual(values = red_range(5)) +
  theme_dark() +
  ggtitle('Smokers')

grid.arrange(p5a, p5b, nrow = 1)
```



In a more clear way, we can plot like this to compare them.

```
all_smokers=tbl_df(breslow)
all_smokers=mutate(all_smokers,deathcount=y)

p5b=ggplot(all_smokers,aes(x=age,y=deathcount,fill=factor(smoke))) +
  geom_bar(stat='identity',position="dodge",width=1) +
  scale_fill_manual(values=c('deepskyblue','firebrick1'),name="Smoke or not", labels=c('Non-smokers','S
  theme_classic() +
  ggtitle('Smokers and Non-smokers')
p5b
```


Smokers and Non−smokers
deathcount
200
150
100
50
0
40
50
60
70
80
age
Smoke or not
Non−smokers
Smokers