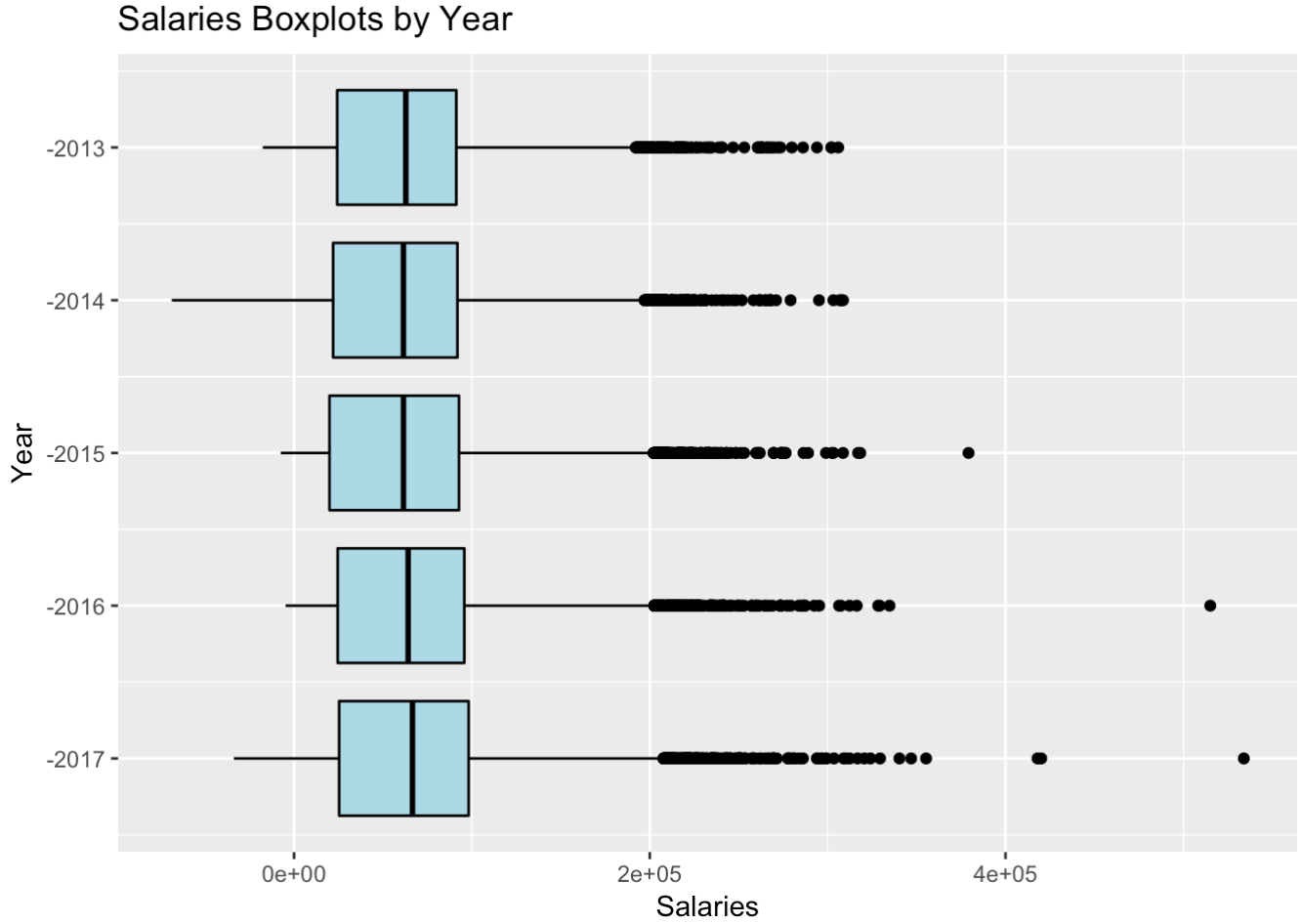
HW01

Yawen Han (yh3069)

9/24/2018

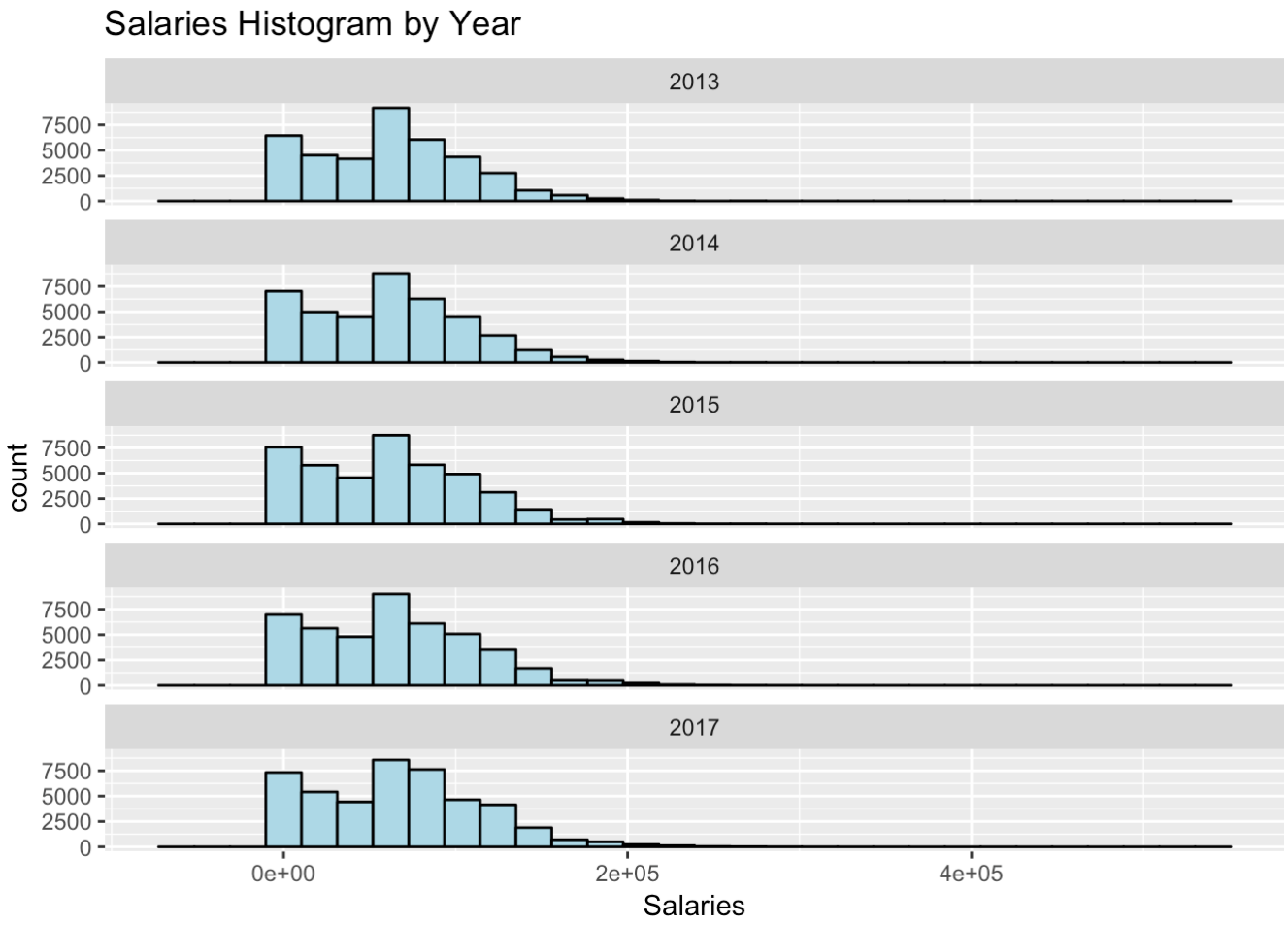
Problem 1

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?



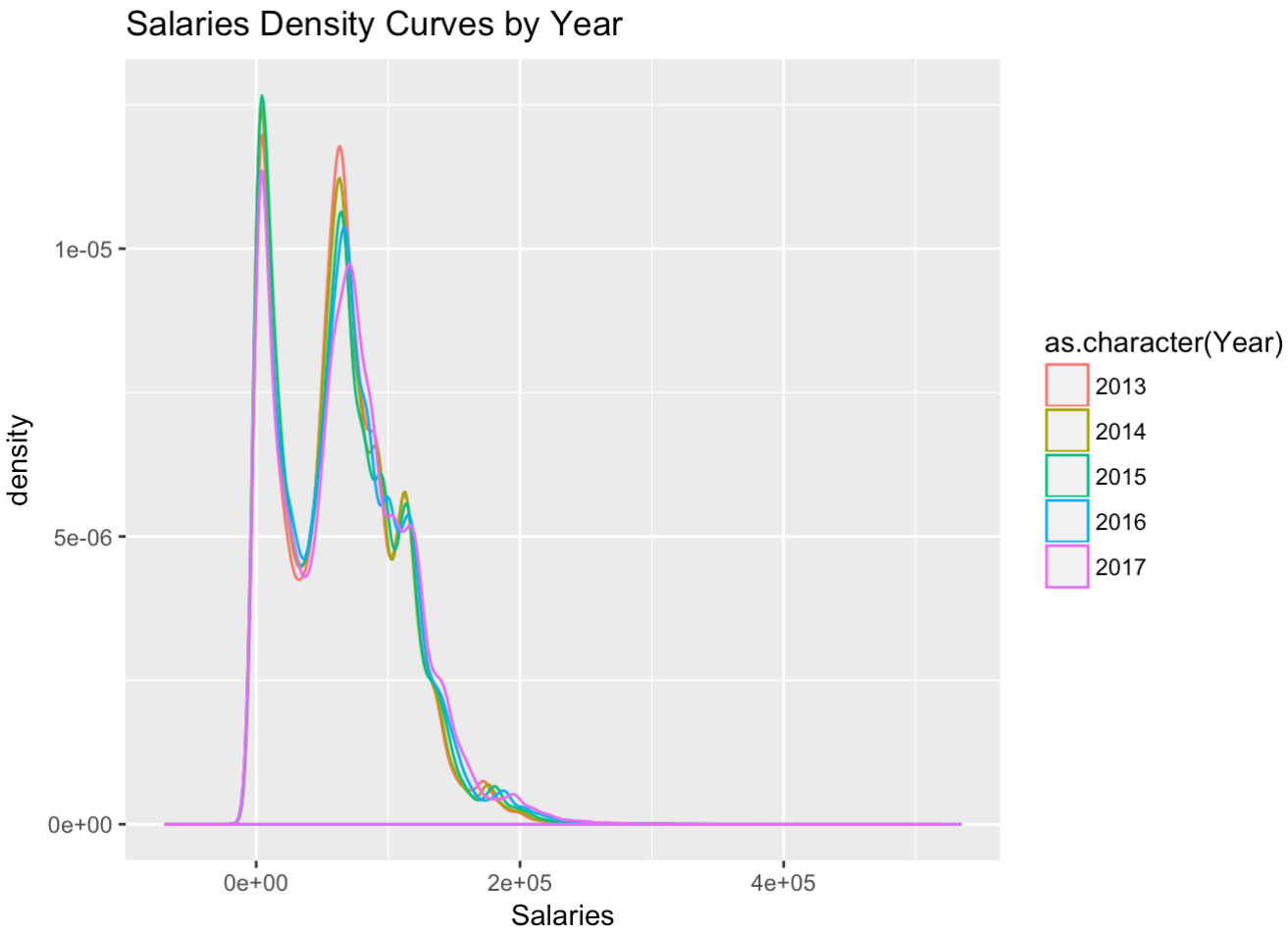
From the boxplot above, the distribution of employee Salaries for the different year are compared as follows:
1)Centers-The median of employee salaries is represented by the line in the box. As the center lines are all at the same level, it seems no obvious difference between the median employee salaries between years;
2)Dispersion:The interquartile ranges are similar, though the overall range of the data set is greater as outliers go further with the increasing of year;
3)Skewness-All distributions are skewed to the right-hand-side, but Year2015-2017 are more skewed than Year2013-2014;
4)Outliers-As outliers go further with the increasing of year, more overall variation might occur;
In conclusion, all five batches of Salaries data look as if they were generally distributed in a similar way, but their distribution become more right-skewed year by year.

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



From the histogram above, we can reconfirm that the five batches of data are distributed in a similar way, and are all right-skewed. However, outliers are not shown as clear as boxplots.

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?



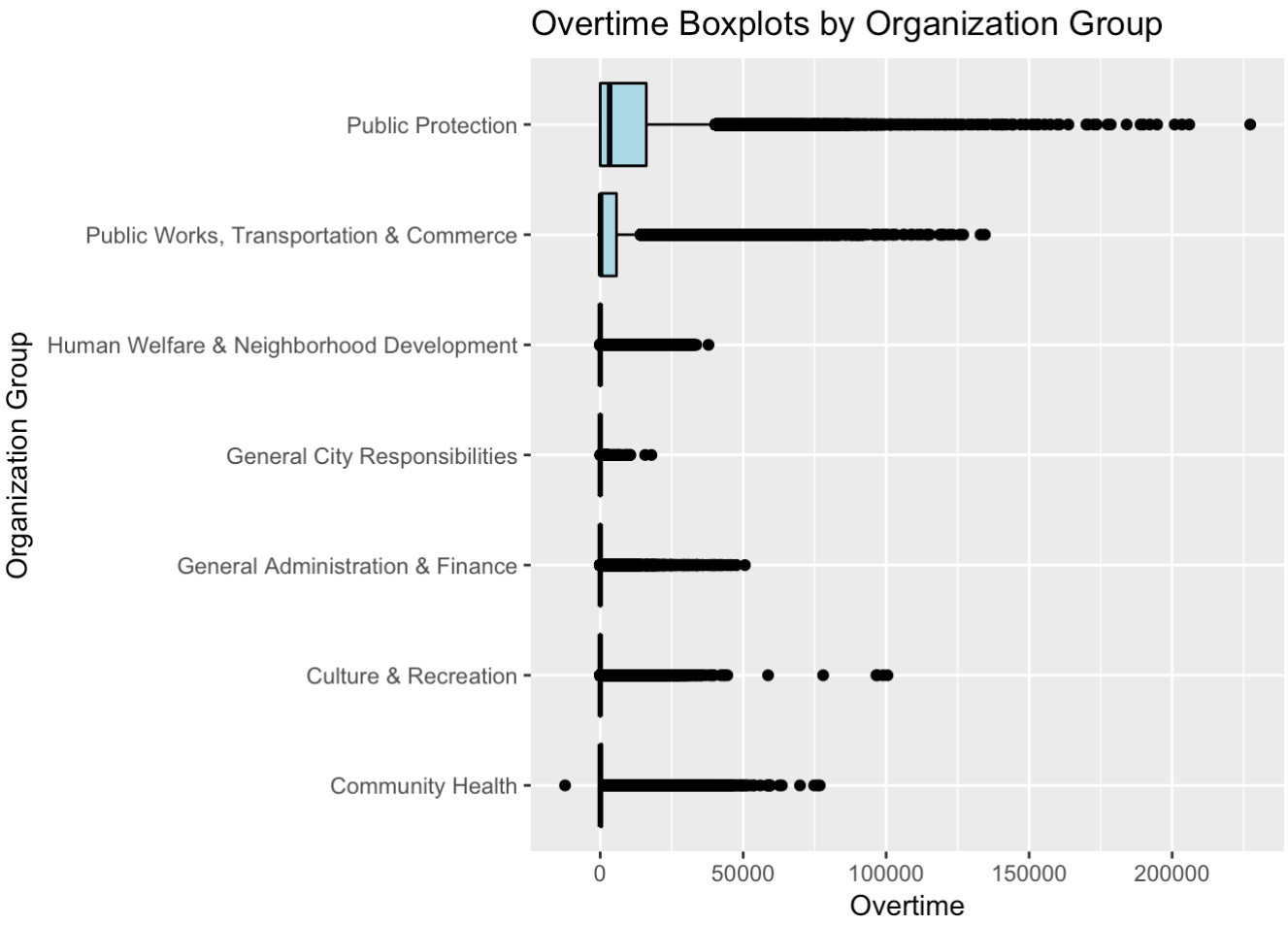
From the density curve, it is observed that all distribution are bimodal. With the lower peak value year by year, the curve tends to spread out to the right.

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)For boxplots: comparing the center, skewness and outliers for different batches of data;
2)For histogram: comparing the general distribution of different batches of data;
3)For density curve: comparing the general distribution from a smoothed version, and provide the spread of different batches of data.

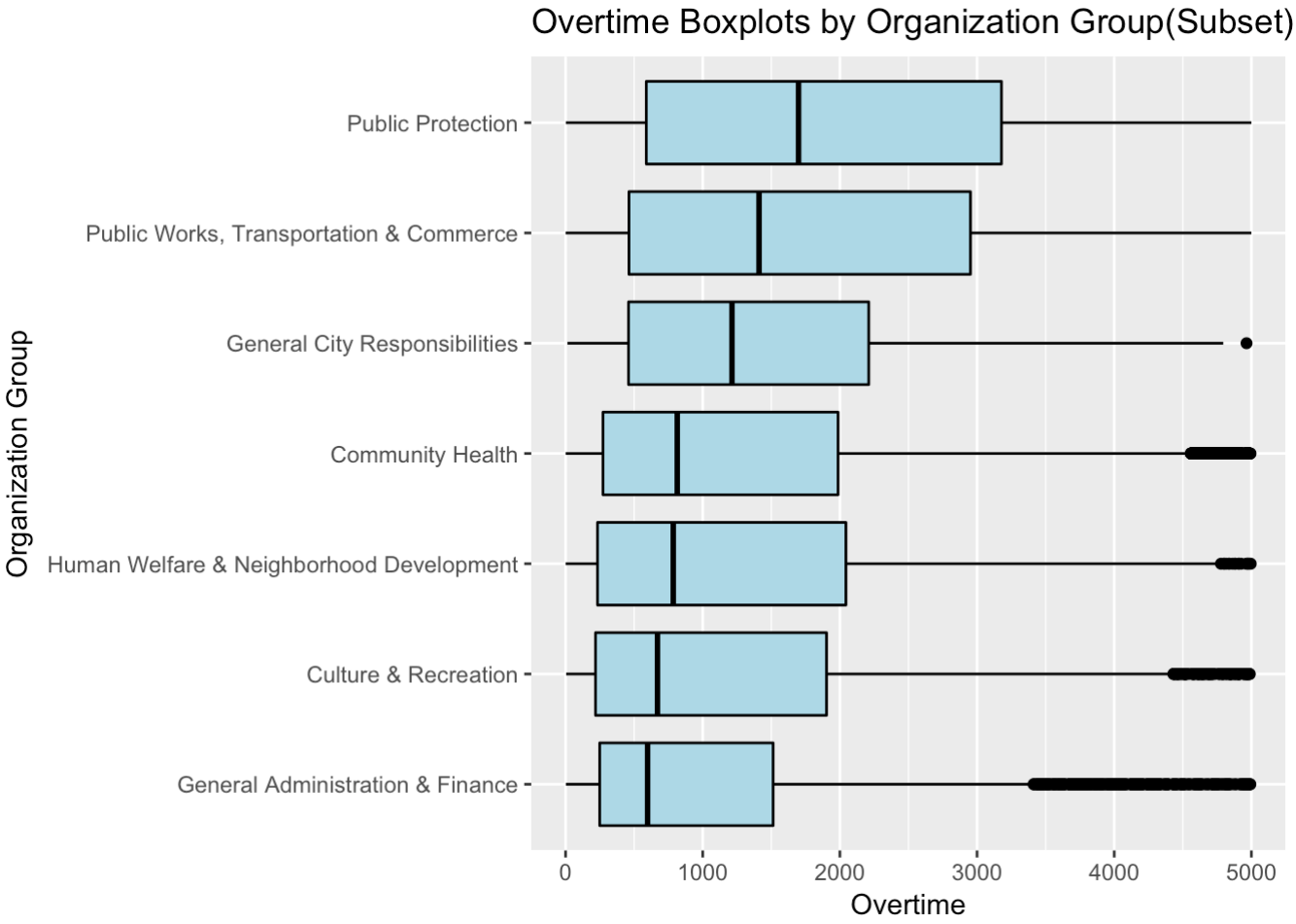
Problem 2

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?



As the spreads of the first two Organization Groups are much larger than others, the boxplots of other groups are squeezed together with only a “line” and outliers remaining. With the graph above, it’s hard to compare the distributions of different groups with the limited information shown, thus not particularly useful to interpret the dataset.

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).



Subset the data: use the boxplots to show the dataset with 0<Overtime<5000. In this case, it's much more clear to investigate the center and spread of the data. As we only focus on the distribution for the 0< Overtime< 5000 range, the boxes are never been squeezed as a "line", the comparison in case of center, skewness, and spread can be proceeded now.

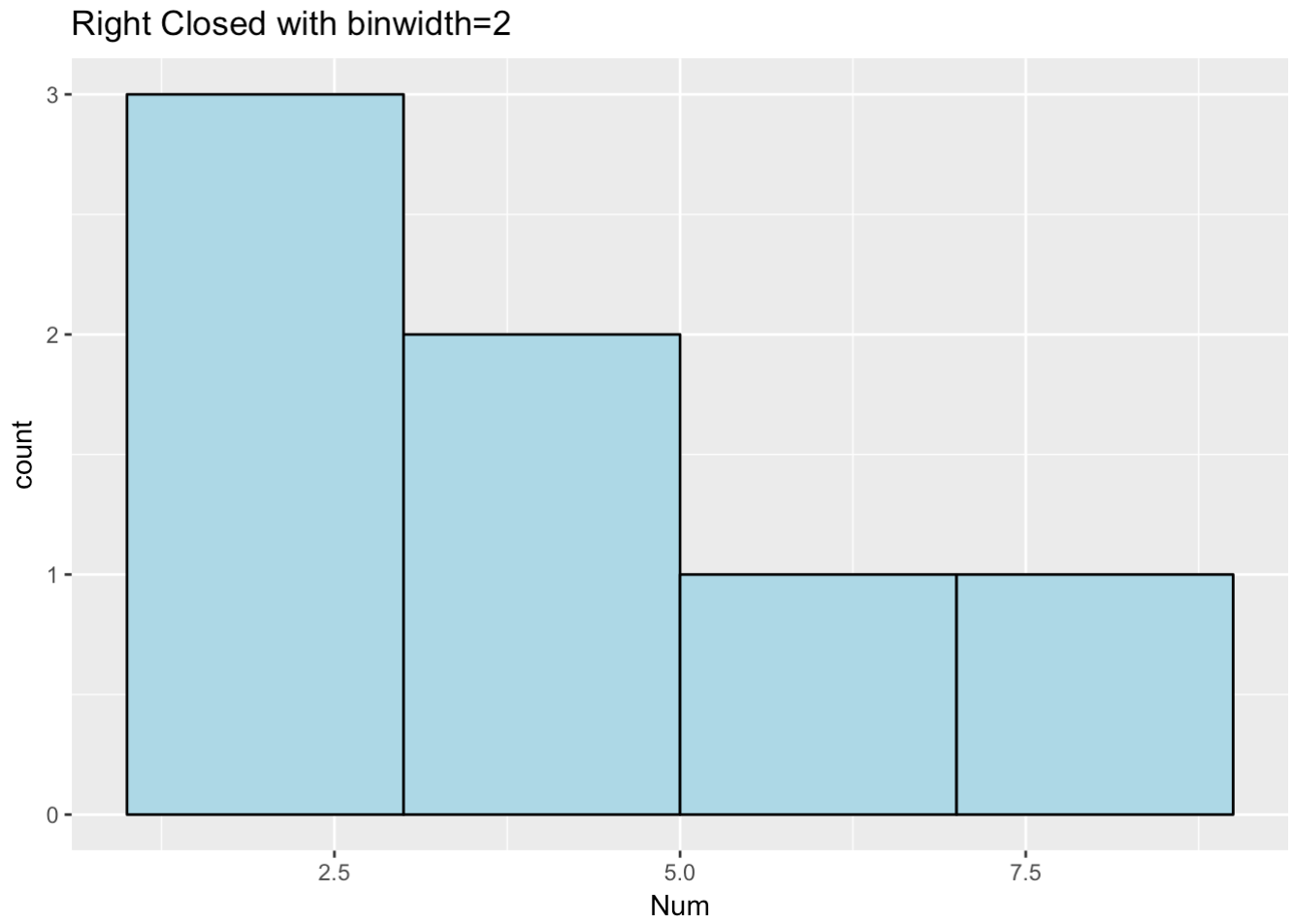
Problem 3

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.

```
##   Num
## 1   1
## 2   1
## 3   2
## 4   5
## 5   5
## 6   7
## 7   9
```

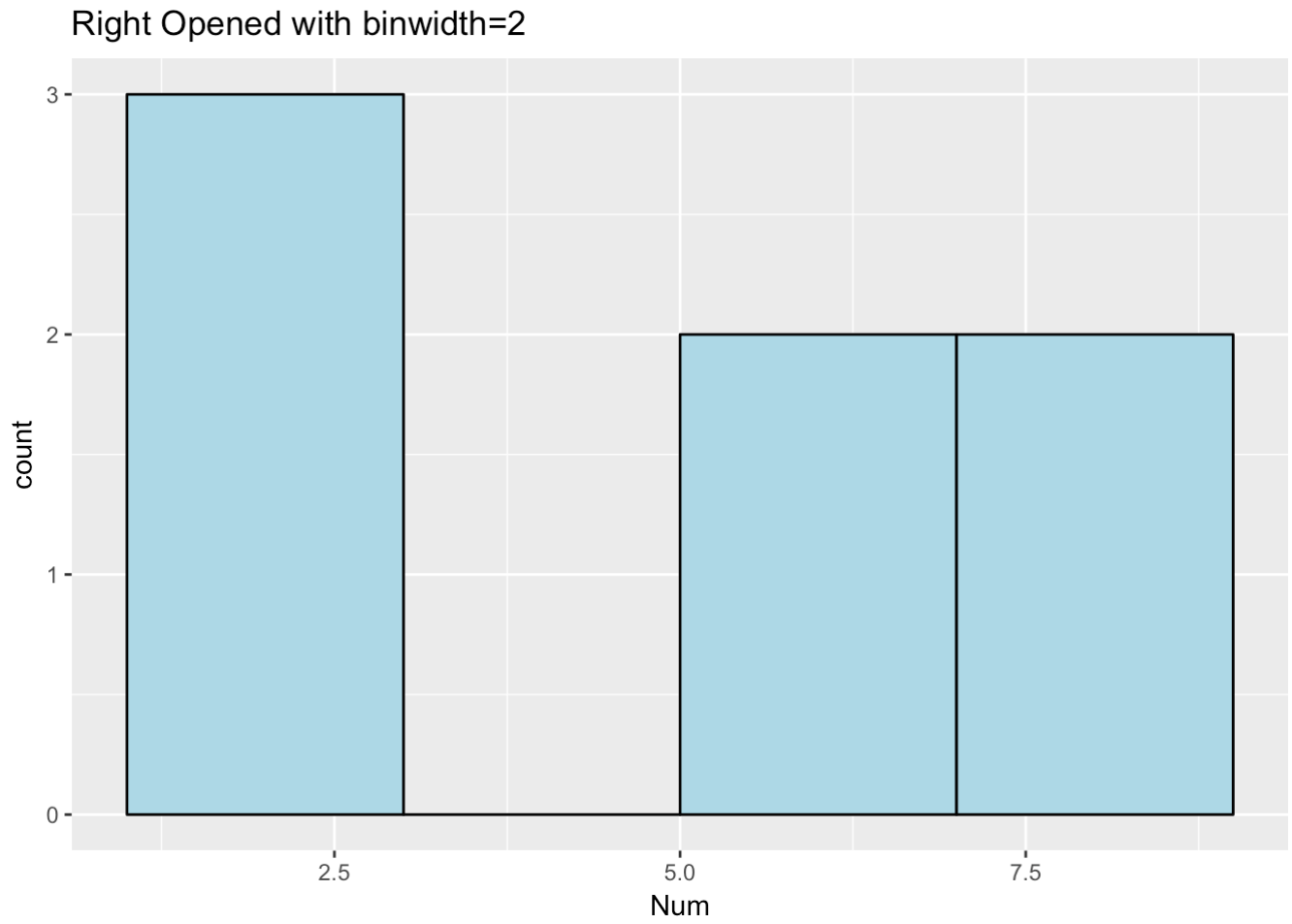
The dataset is show above as [1,1,2,5,5,7,9]

```
Ba1 <- ggplot(x,aes(Num))+geom_histogram(color="black",fill="lightblue",binwidth=2,right
=TRUE)+ggtitle("Right Closed with binwidth=2")
Ba1
```



The dataset is plotted as a right closed histograms with binwidth=2.

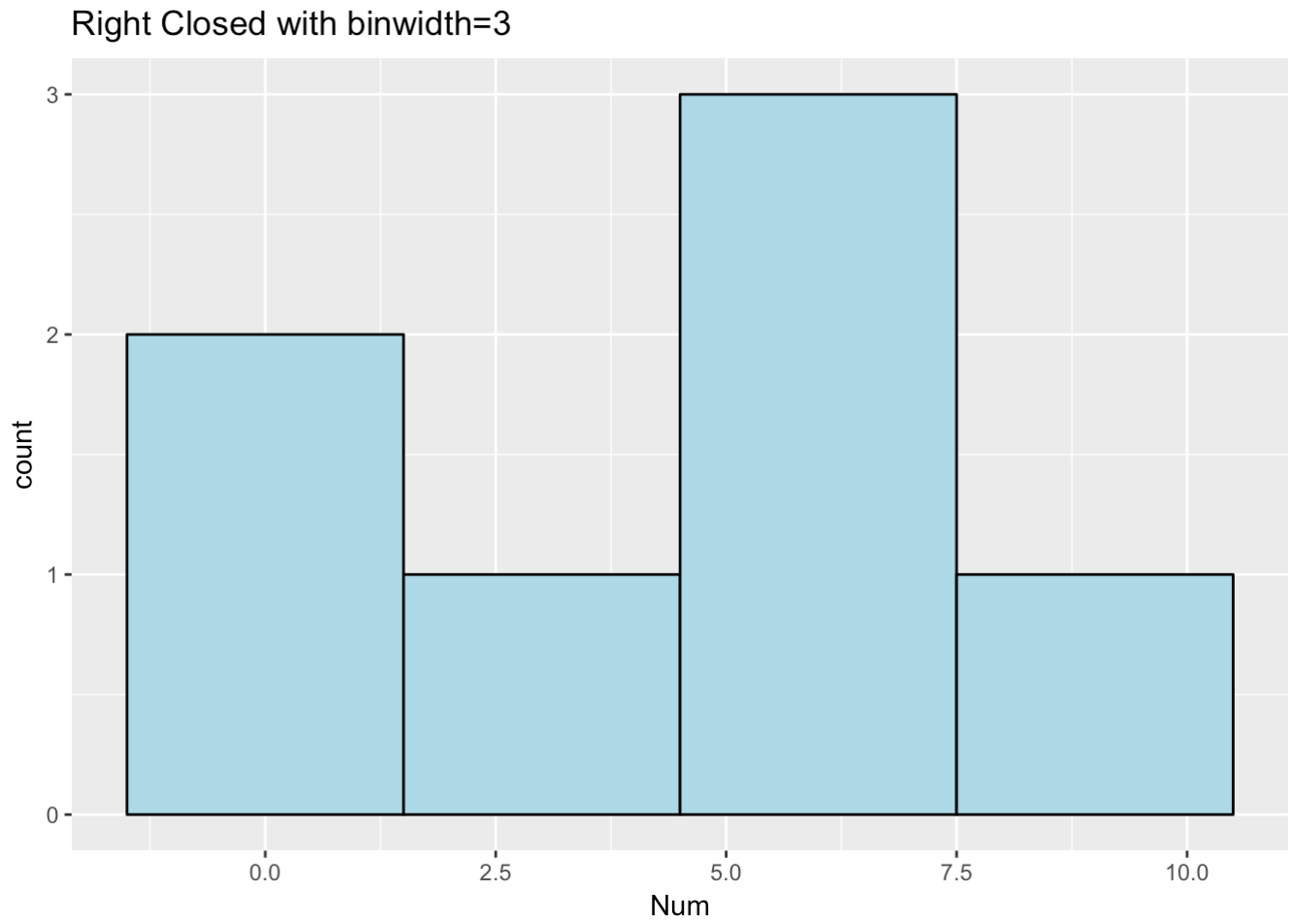
```
Ba2 <- ggplot(x,aes(Num))+geom_histogram(color="black",fill="lightblue",binwidth=2,right
=FALSE)+ggtitle("Right Opened with binwidth=2")
Ba2
```



The dataset is plotted as a right opened histogram with the same binwidth=2.It's easy to observe that the two histograms are not identical as there is a big "gap" occurs in the latter histogram.

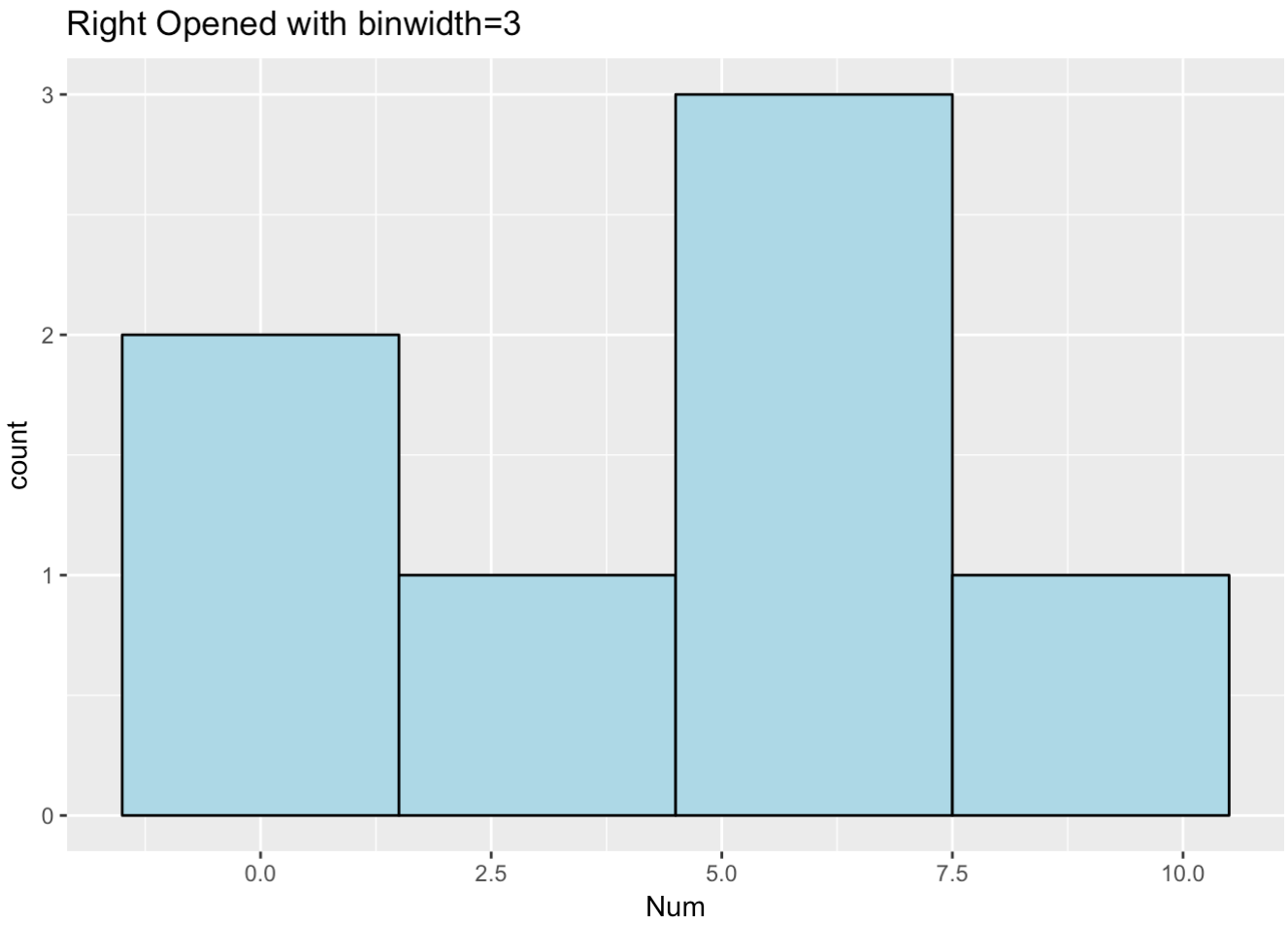
b.Adjust parameters–the same for both–so that the right open and right closed versions become identical. Explain your strategy.

```
Bb1 <- ggplot(x,aes(Num))+geom_histogram(color="black",fill="lightblue",binwidth=3,right
=TRUE)+ggtitle("Right Closed with binwidth=3")
Bb1
```



After changing the binwidth from 2 to 3, the dataset is plotted as a right closed histograms with binwidth=3.

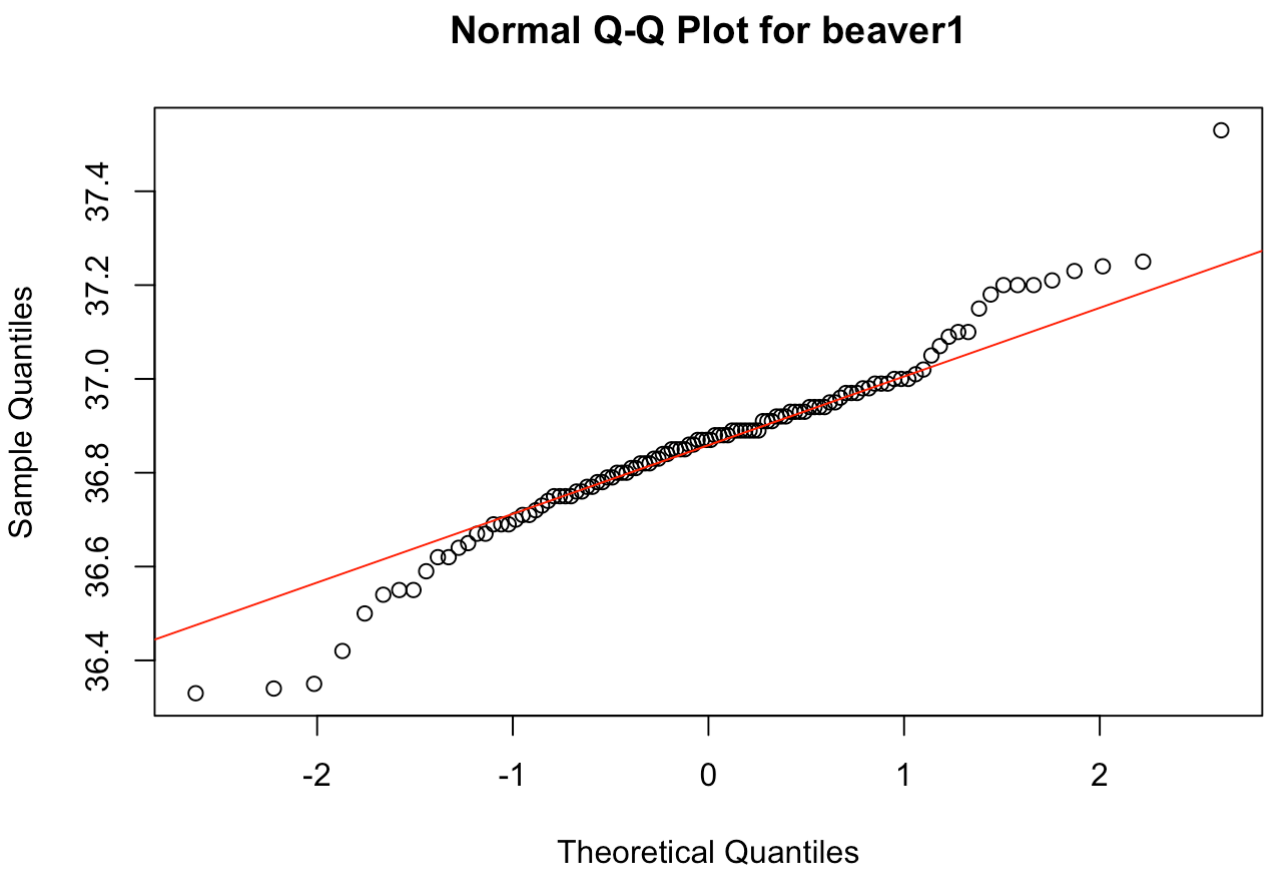
```
Bb2 <- ggplot(x,aes(Num))+geom_histogram(color="black",fill="lightblue",binwidth=3,right
=FALSE)+ggtitle("Right Opened with binwidth=3")
Bb2
```



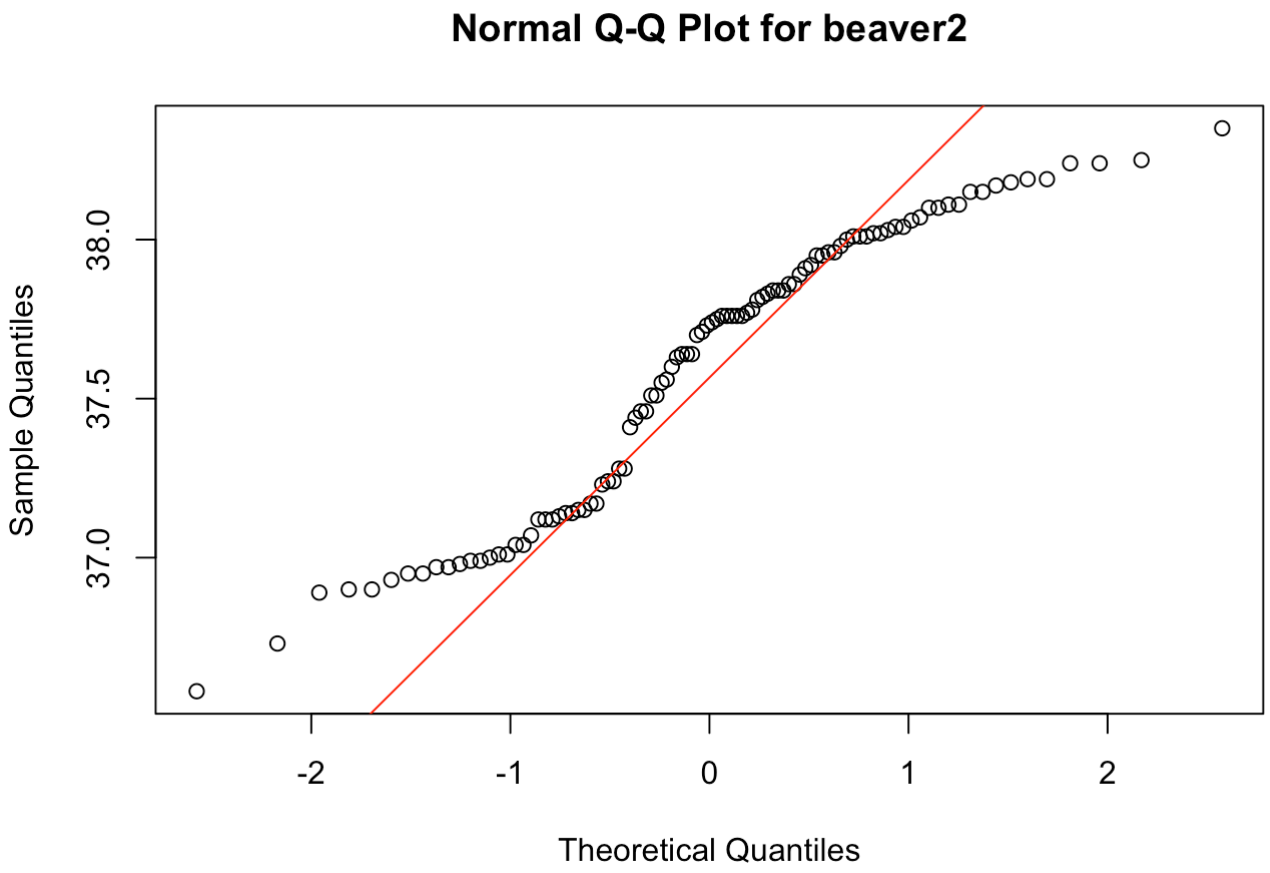
Then the dataset is plotted as a right opened histogram with the same binwidth=3, and the two histograms are identical this time. My strategy to avoid the discrepancy of right-closed and right-opened histogram is: to choose rational bins so that no data falls on the boundaries of bins. Thus, the boundaries would not affect the total counts in each bin.

Problem 4

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?


Normal Q-Q Plot for beaver1
Sample Quantiles
36.4
36.6
36.8
37.0
37.2
37.4
-2
-1
0
1
2
Theoretical Quantiles

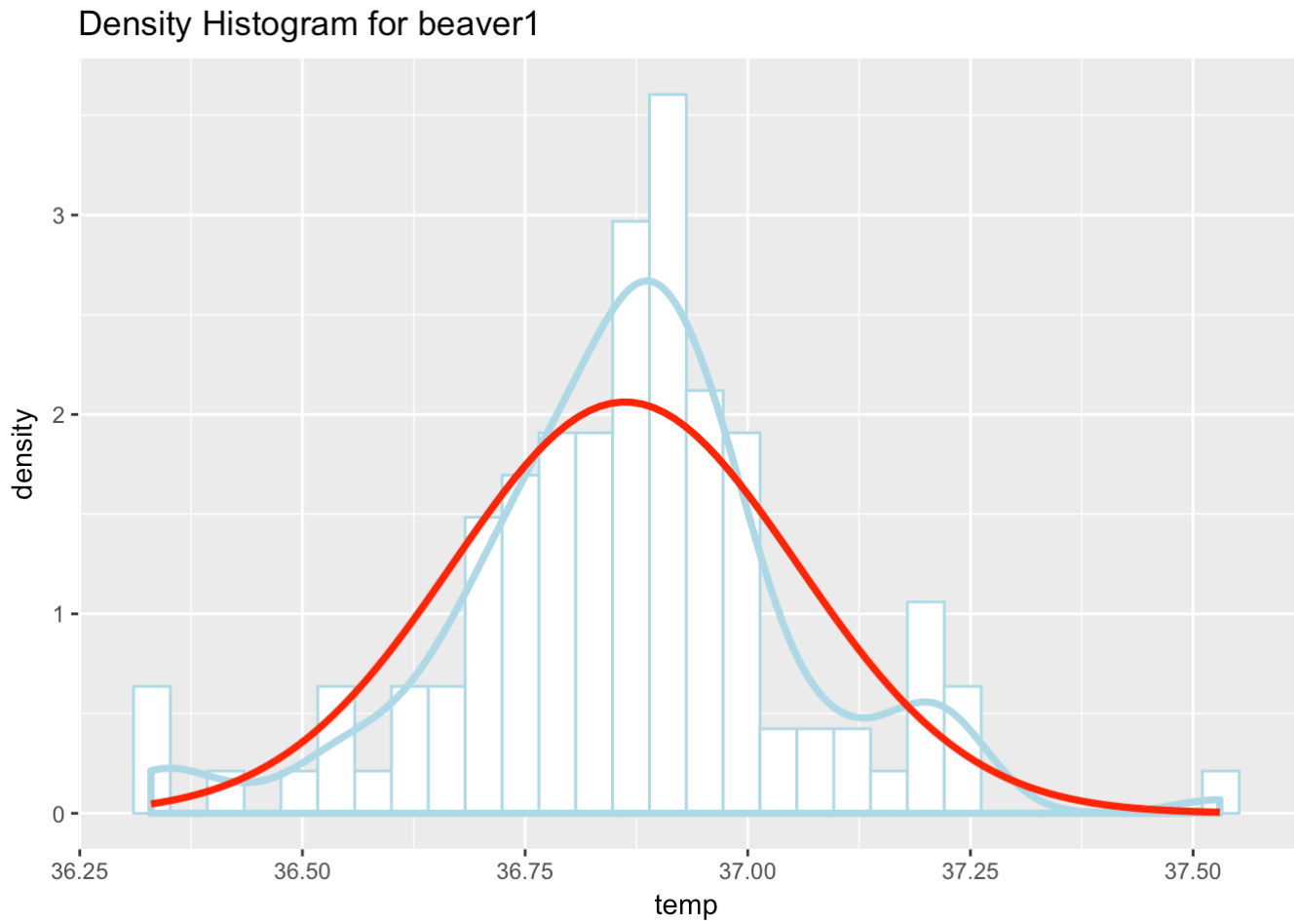
The Normal Q-Q Plot for beaver1 with the theoretical normal line is plotted above. The middle parts of the Q-Q plot overlap with the theoretical normal line.



The Normal Q-Q Plot for beaver2 with the theoretical normal line is also plotted above, and there exits obvious difference between its Q-Q plot and the theoretical normal line. Compared the two Q-Q plots, beaver1 appears to be more normally distributed than beaver2 as there are more overlapping parts between its Q-Q plot and theoretical normal line.

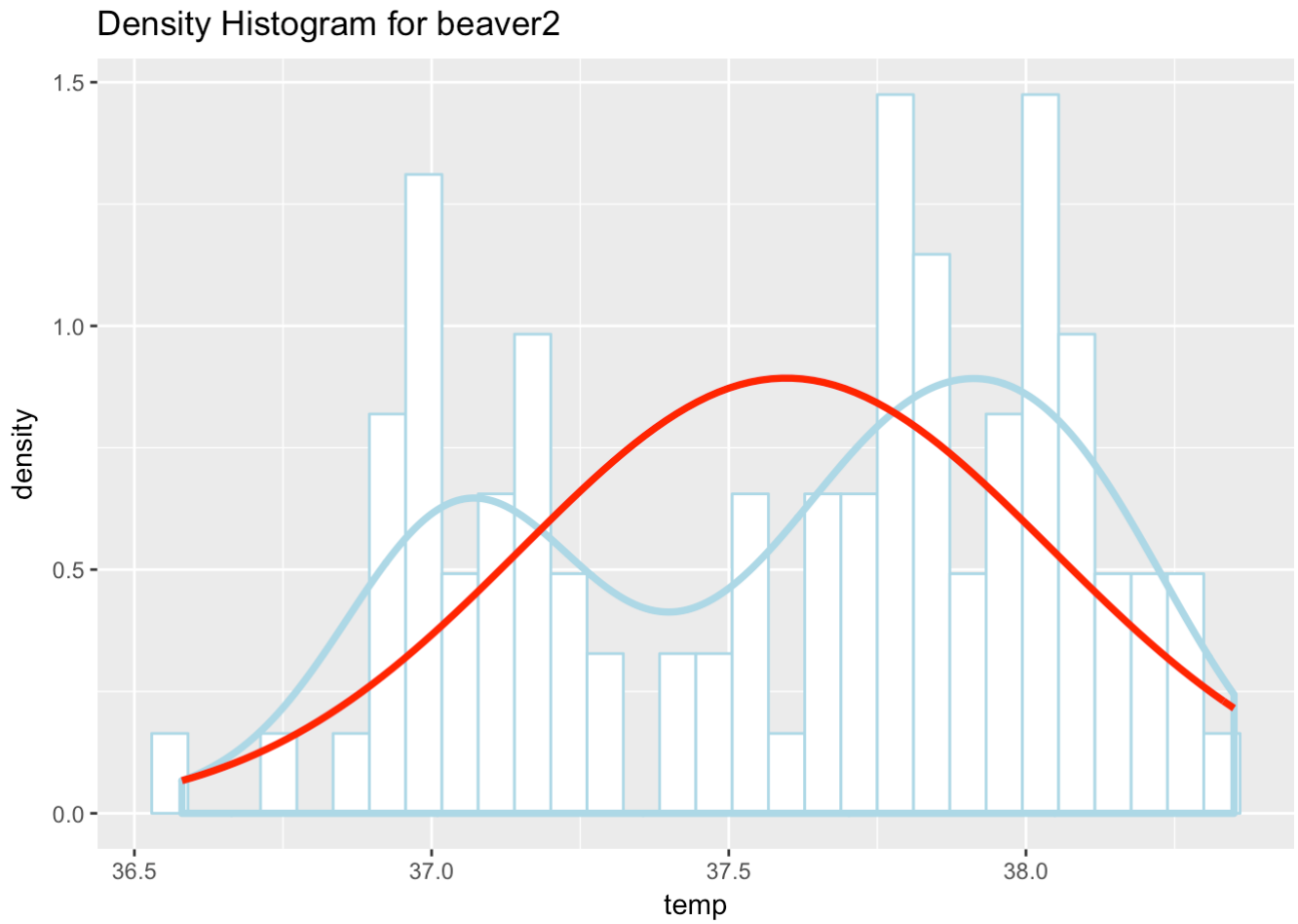
b.Draw density histograms with density curves and theoretical normal curves overlaid. Do you get the same results as in part a)?

```
ggplot(beaver1,aes(x=temp))+geom_histogram(aes(y=..density..),fill="white",color="lightb
lue")+geom_density(color="lightblue",lwd=1.3)+stat_function(fun=dnorm,args=list(mean=mea
n(x),sd=sd(x)),color="red",lwd=1.3)+ggtitle("Density Histogram for beaver1")
```



The Density Histogram for beaver1 with theoretical normal curve is plotted above. The density curve of beaver1 has the similar “bell” shape as the theoretical normal curve, but the data seems more condensed with a higher peak value.

```
ggplot(beaver2,aes(x=temp))+geom_histogram(aes(y=..density..),fill="white",color="lightb
lue")+geom_density(color="lightblue",lwd=1.3)+stat_function(fun=dnorm,args=list(mean=mea
n(y),sd=sd(y)),color="red",lwd=1.3)+ggtitle("Density Histogram for beaver2")
```



The Density Histogram for beaver2 with theoretical normal curve is also plotted above. The density curve of beaver2 is bimodal, which not looks like a “bell” shape as the theoretical normal curve. Same as the conclusion in part(a), beaver1 appears to be more normally distributed than beaver2 with a higher similarity density curve with theoretical normal curve.

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

```
library("dplyr")
shapiro.test(x)
```

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

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

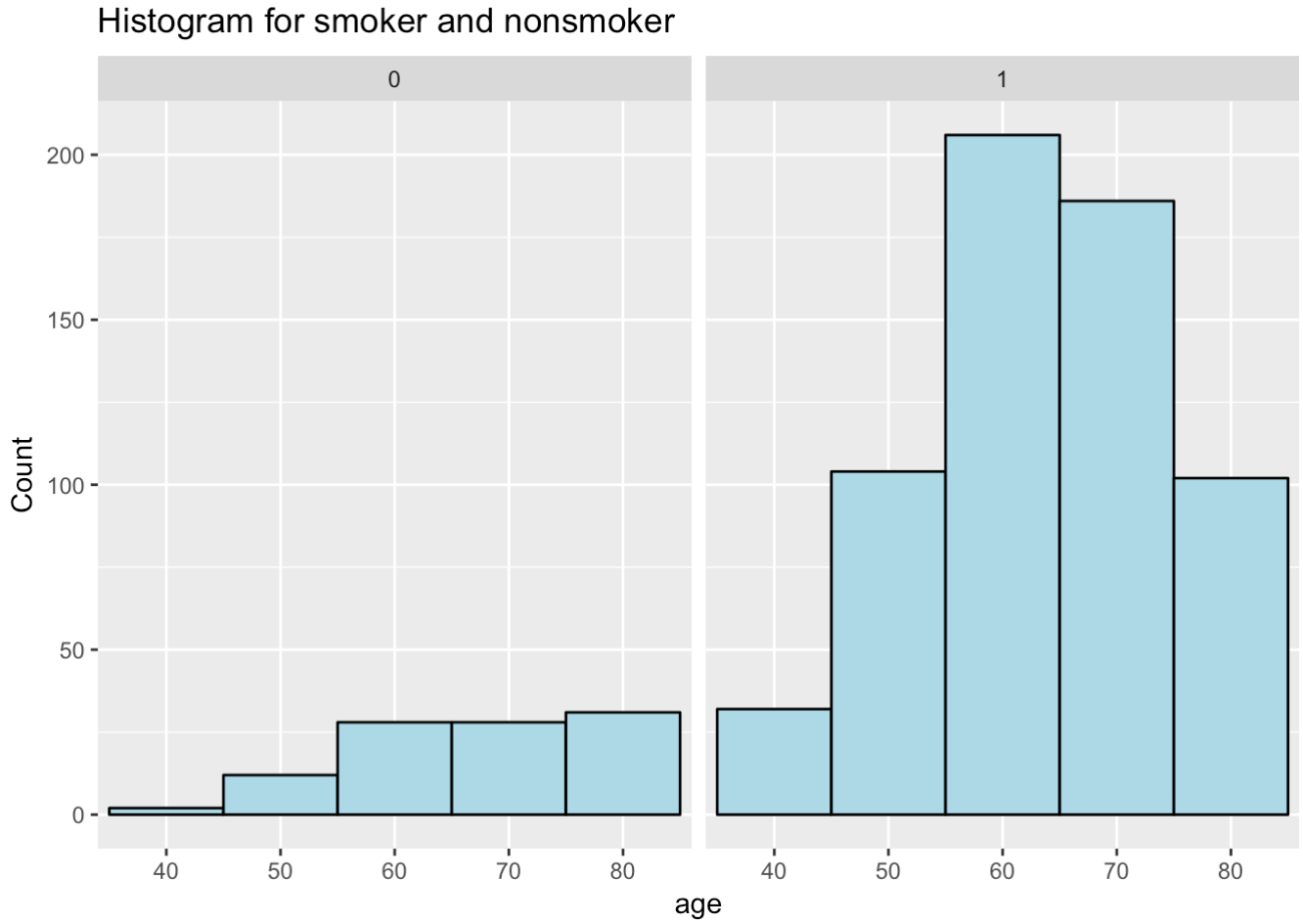
```
##
##  Shapiro-Wilk normality test
##
## data:  y
## W = 0.93336, p-value = 7.764e-05
```

From the output above, if we set alpha=0.05, both p-value < 0.05 implying that the distribution of the data is significantly different from the normal distribution. In other words, both datasets can be assumed not normally distributed. However, if we set alpha=0.01, the p-value of beaver1 is 0.01226>0.01, and the p-value of beaver2 is 7.7e-05<0.01. This implies that the distribution of beaver1 appears to be more normally distributed than beaver2, which agrees with the conclusion in parts(a) and (b).

Problem 5

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)
ggplot(breslow, aes(x=age,y=y))+geom_bar(color="black",fill="lightblue",stat="identity",
width=1.0)+facet_wrap("smoke")+ggtitle("Histogram for smoker and nonsmoker")+ylab("Coun
t")
```



This dataset is very special as it already provide the bins and count for each age range. Therefore, in this case, it's no more useful to apply the "geom_history()" to get histograms. Instead, barchart is a good choice to show the histogram of the number of deaths attributed to coronary artery disease among doctors in the breslow dataset. As the plots show above, the left plot represents the histogram for "nonsmoker", and the right plot represents the histogram for "smoker". The x-axis is the range of their age, and the y-axis is the count for each age interval.