Homework 1

Youki Cao

Sep 25 2018

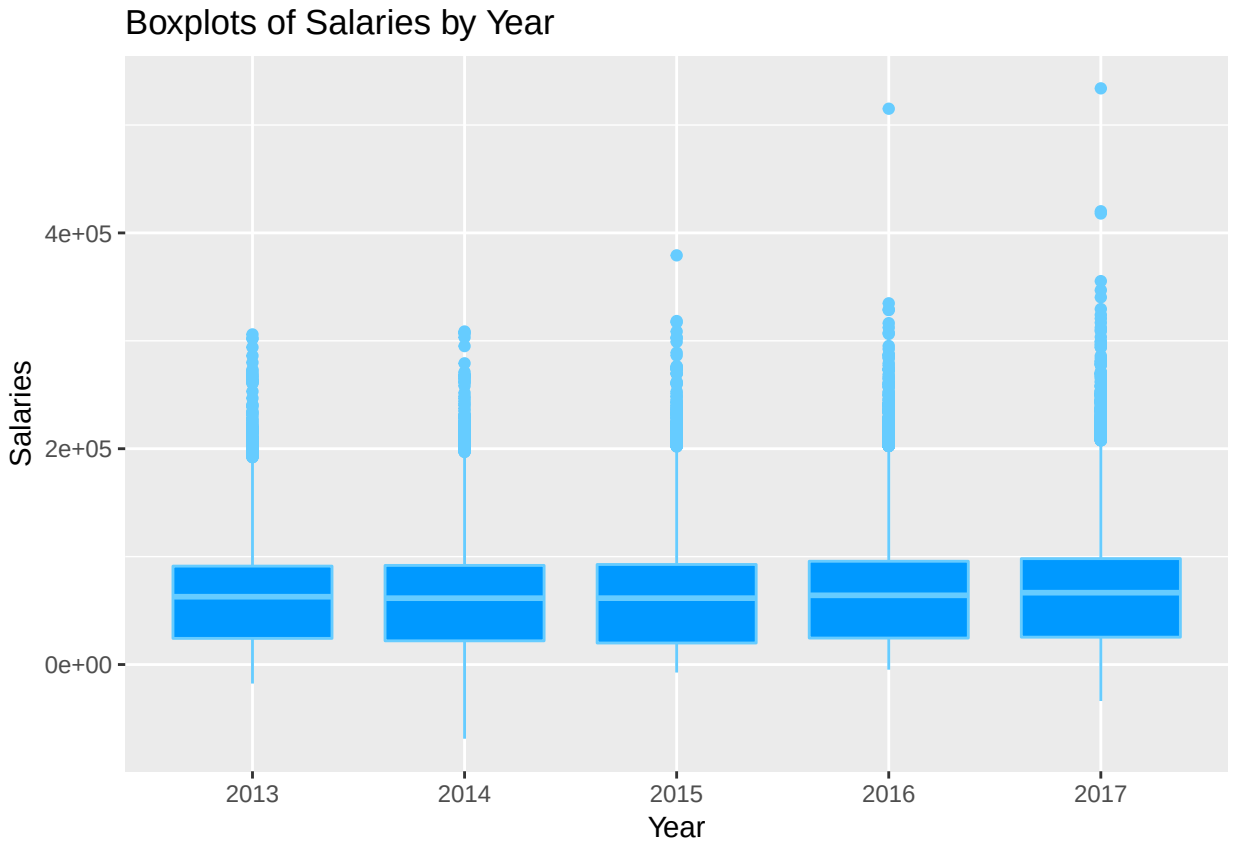
1. Salary

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?

```
library(ggplot2)  # plotting

employee <- read.csv("Employee.csv")  # read data from file "Employee.csv"
year <- factor(employee$Year)
salary <- employee$Salaries

# boxplot
p1a <- ggplot(employee, aes(x = year, y = salary)) +
  geom_boxplot(fill = "#0099FF", color = "#66CCFF") +
  ggtitle("Boxplots of Salaries by Year") +
  labs(x = "Year", y = "Salaries")
p1a
```



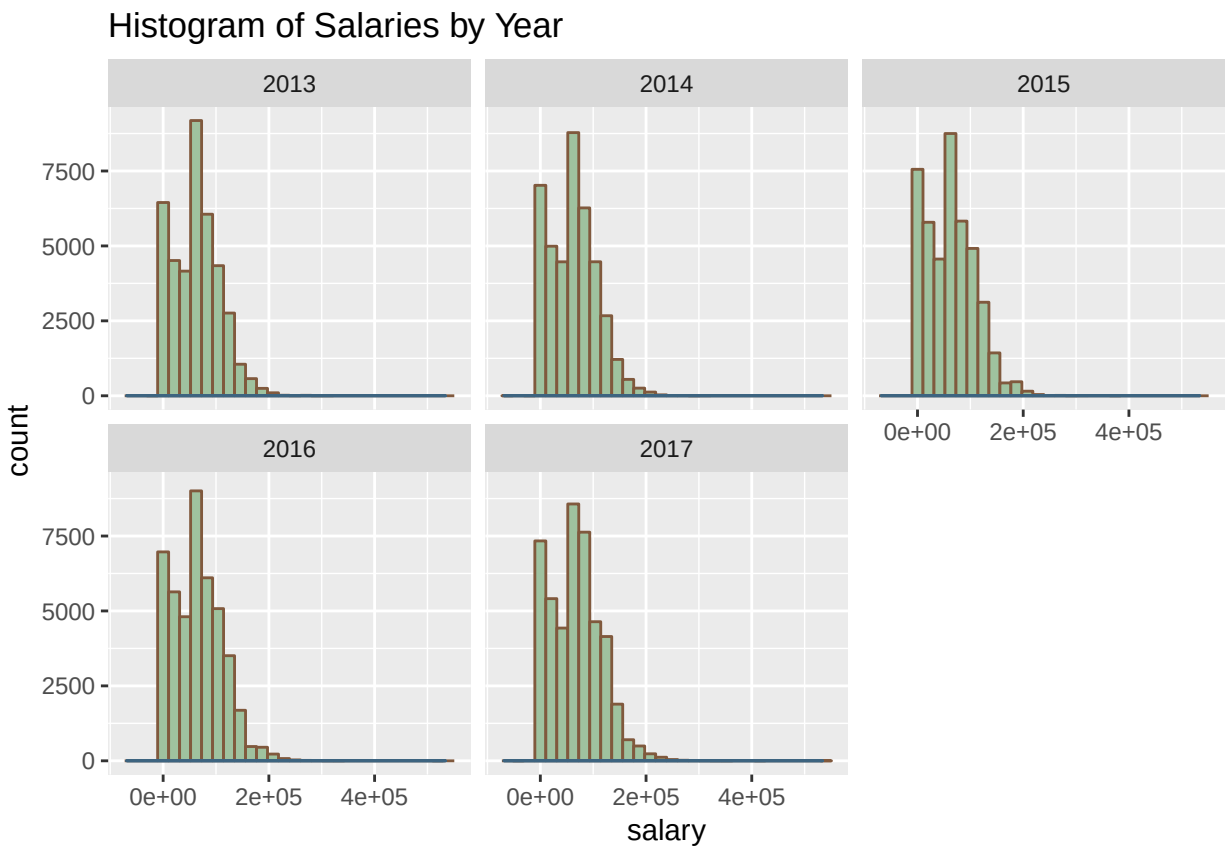
- From the boxplots, we figure that during the period from 2013 to 2017. There are more outliers, and more extreme high values. This shows that the salaries of the employees with extremely high incomes

are still increasing over years.

- It is hard to notice the distributions over the year. The inter-quantile range becomes a little bit wider over these years. The median of salaries increases a little bit. But these tiny changes are so hard to notice.

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

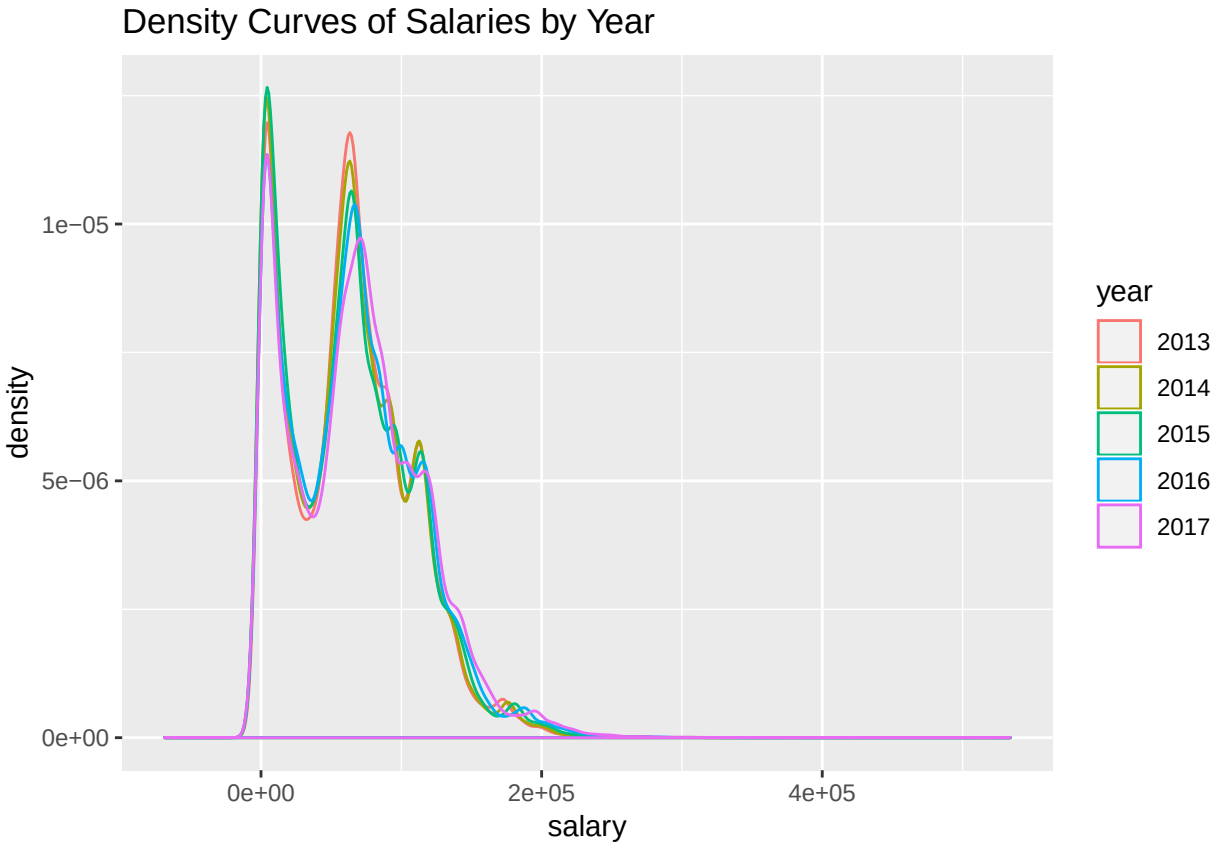
```
# histogram
p1b <- ggplot(employee, aes(x = salary)) +
  geom_histogram(colour = "#80593D", fill = "#9FC29F") +
  geom_density(color = "#3D6480") +
  facet_wrap(~Year) +
  # formatting
  ggtitle("Histogram of Salaries by Year ")
p1b
```



- The histogram provides the general distribution pattern of the salaries of each year. It shows that the distribution pattern is similar in 2013 - 2017, that is skewed right, and has double peaks.
- Besides, I can get the count of the data from histograms.

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(employee, aes(x = salary, color = year)) +
  geom_density() +
  ggtitle("Density Curves of Salaries by Year")
p1c
```



- The overlapping density curves indicate a more detailed view of distribution from 2013 to 2017, and make it easy to compare the distribution. It also shows how the distribution changes over time.
- I find that the distributions pattern are very similar these years. But density of relative low salaries decreases, and density of relative high salaries increases a little bit.

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?

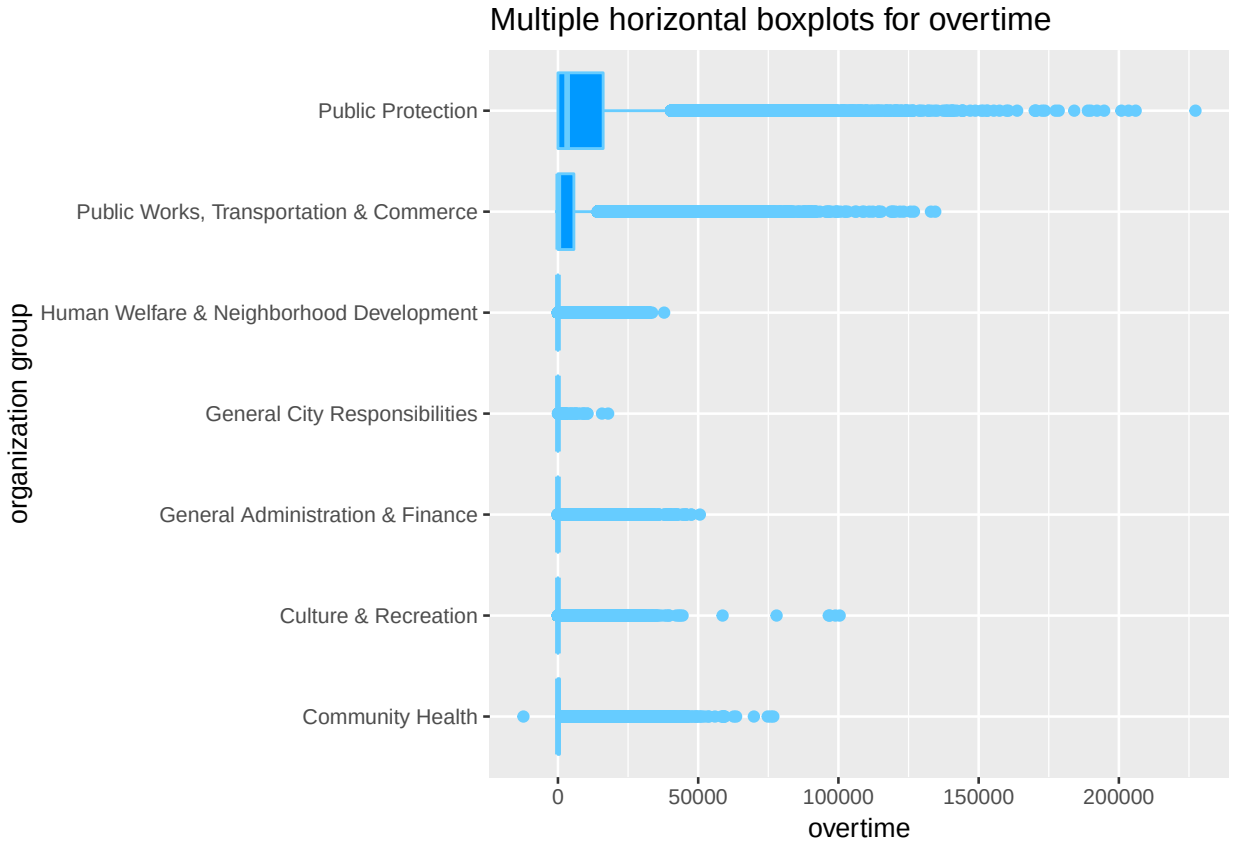
- The boxplot indicates the statistical data, like quantile, median directtly. It is also useful for identying outliers. In this data, we find the median of the data is not very high and almost unchanged over the years. And we can clearly observe the outliers and their values.
- The histogram shows the distribution pattern of emperical data, like skewness, multimodality. It is also a good way to indicate the exact count of particular range of salaries. We can find the general distribution modality, but not easy to observe outliers.
- The density curves shows the density of the data. As we plot overlapping curves based on different years, we can compare the similarity and difference of the distribution of different years, and also get the change of the distribution over the years.

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?

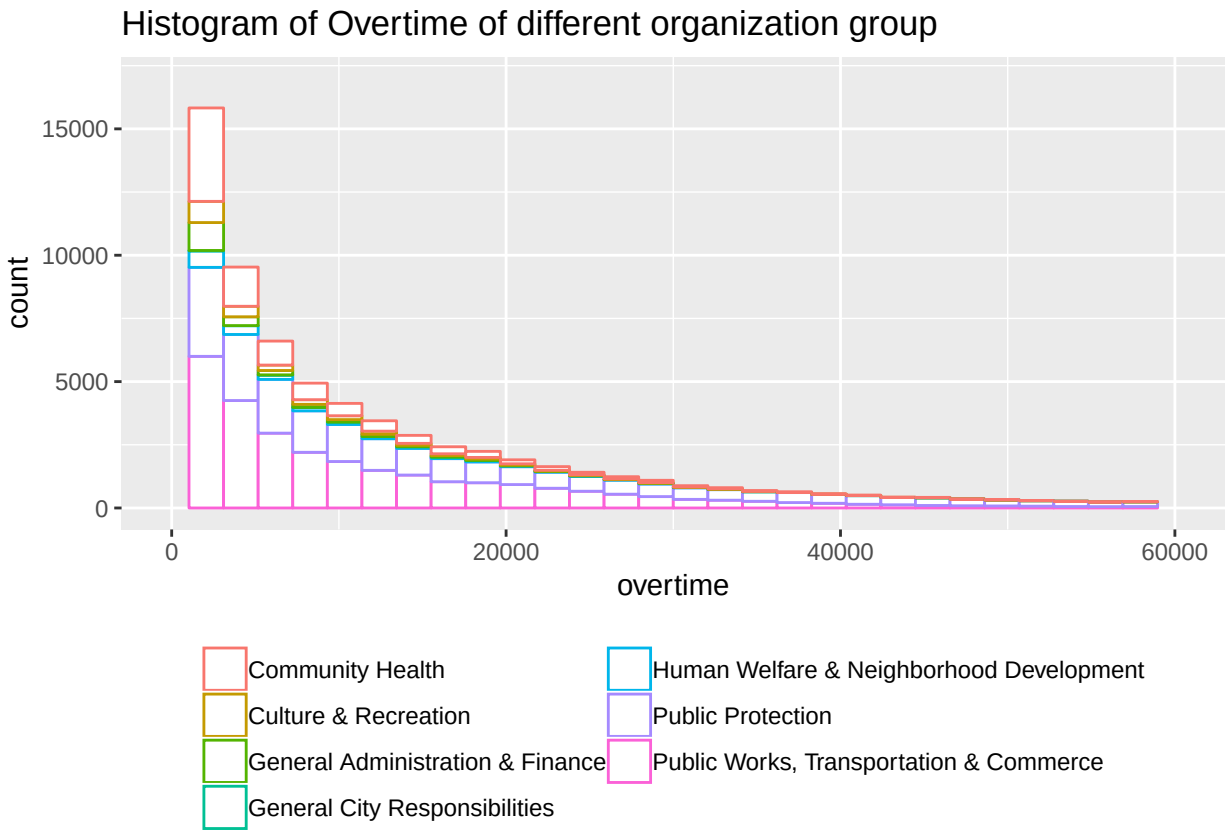
```
organization.group = employee$Organization.Group
overtime = employee$Overtime
p2a <- ggplot(employee, aes(x = reorder(organization.group, overtime, median),
                            y = overtime)) +
  geom_boxplot(fill = "#0099FF", color = "#66CCFF") +
  coord_flip() +
  theme_grey(10) +
  ggtitle("Multiple horizontal boxplots for overtime") +
  labs(x = "organization group")
p2a
```



- It's not that useful because there are so many 0 in the overtime data. So the data in the inter-quatile is too concentrated, and therefore inter-quatile range is quite small. That makes quite a lot data outliers.

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).

```
p2b <- ggplot(employee, aes(x = overtime, color = organization.group)) +
  geom_histogram(fill = "white") +
  xlim(1, 60000) +
  ylim(0, 17000) +
  theme(legend.position="bottom") +
  theme(legend.title = element_blank()) +
  guides(col = guide_legend(ncol=2)) +
  ggtitle("Histogram of Overtime of different organization group")
p2b
```



- This histogram shows clearly the distribution pattern of overtime data of each organization group by using different colors. We can compare data both by overtime and by different organization groups. Almost for all groups, the distribution is right skewed. And for different groups we can get which group's overtime is higher or lower than another according to the histogram.
- Also histogram provides the accurate count of overtime, whereas boxplots cannot.

3. Boundaries

[10 points]

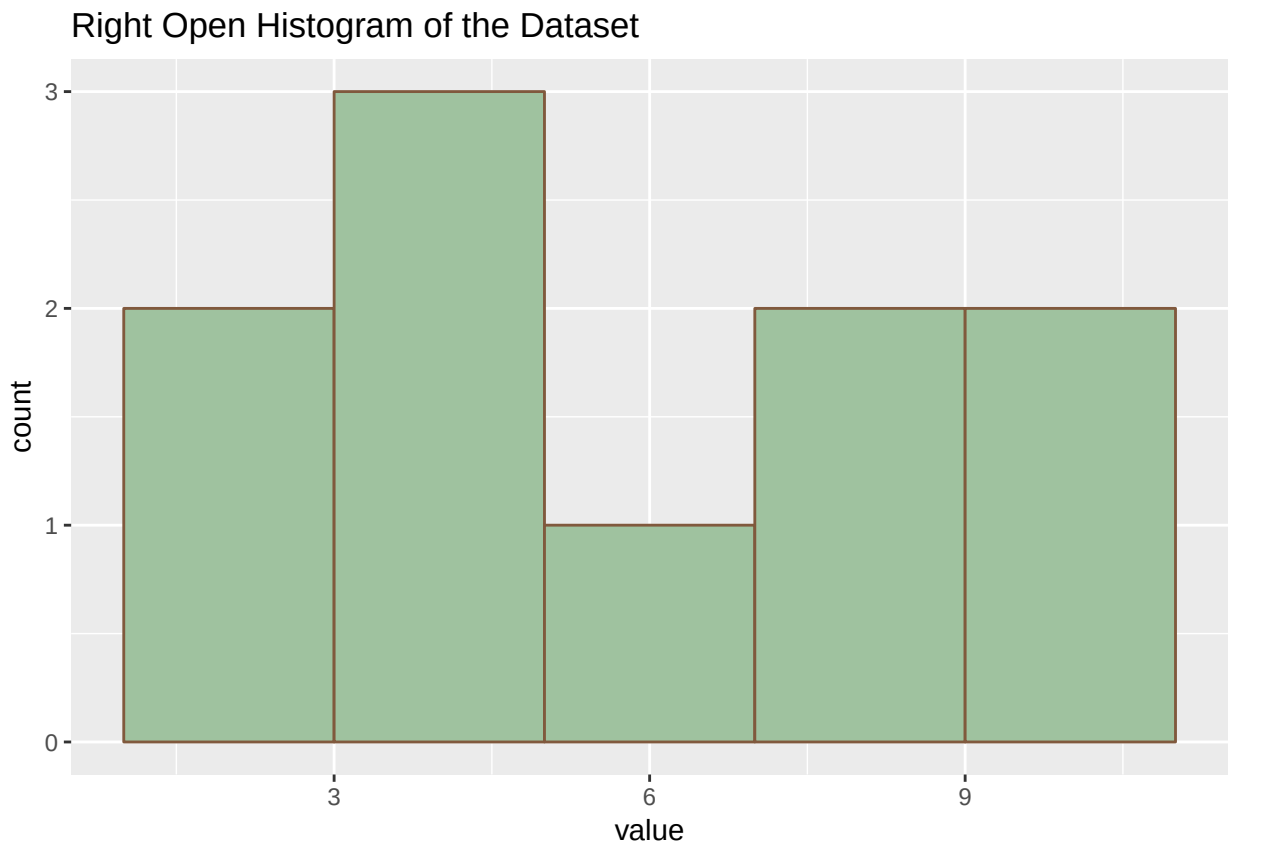
a) Find or create a small dataset (< 100 observations) for which right open and right closed histograms for the same parameters are not identical. Display the full dataset (that is, show the numbers) and the plots of the two forms.

```
# Create dataset
value <- c(1, 2, 3, 3, 4, 6, 7, 8, 9, 10)
dataset <- data.frame(value)
dataset
```

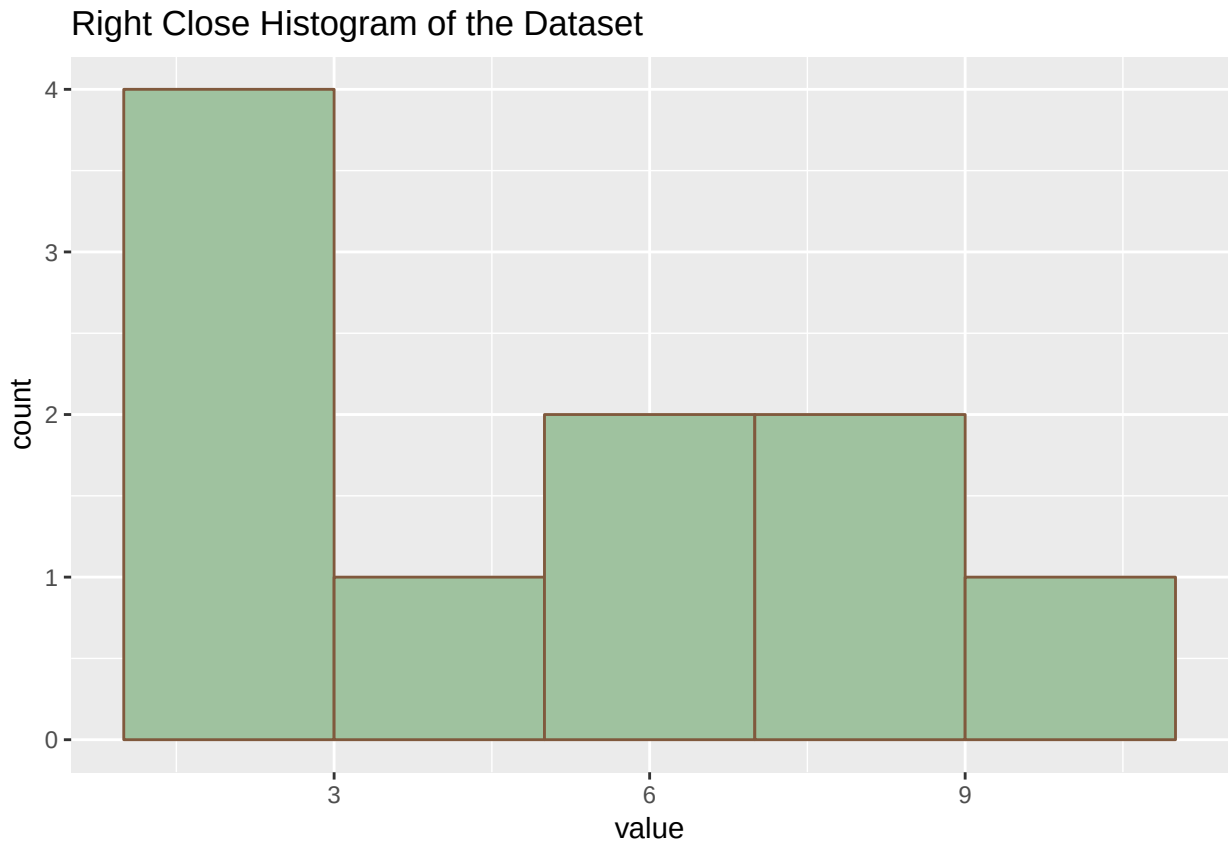
```
##    value
## 1      1
## 2      2
## 3      3
## 4      3
## 5      4
## 6      6
## 7      7
## 8      8
```

```
## 9      9
## 10    10
```

```
# histogram with right open
p3a1 = ggplot(dataset, aes(x = value)) +
  geom_histogram(right = FALSE, binwidth = 2, colour = "#80593D", fill = "#9FC29F") +
  ggtitle("Right Open Histogram of the Dataset")
p3a1
```

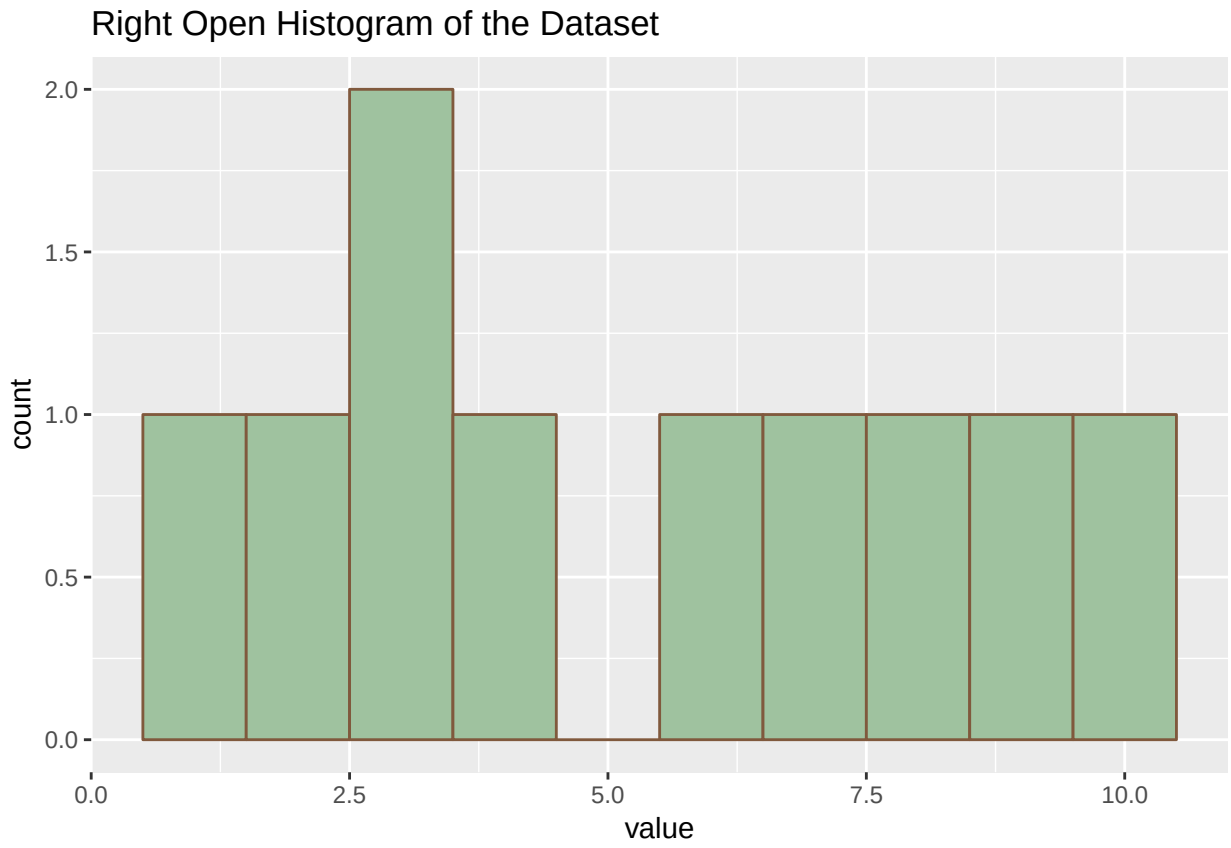



```
# histogram with right close
p3a2 = ggplot(dataset, aes(x = value)) +
  geom_histogram(right = TRUE, binwidth = 2, colour = "#80593D", fill = "#9FC29F") +
  ggtitle("Right Close Histogram of the Dataset")
p3a2
```

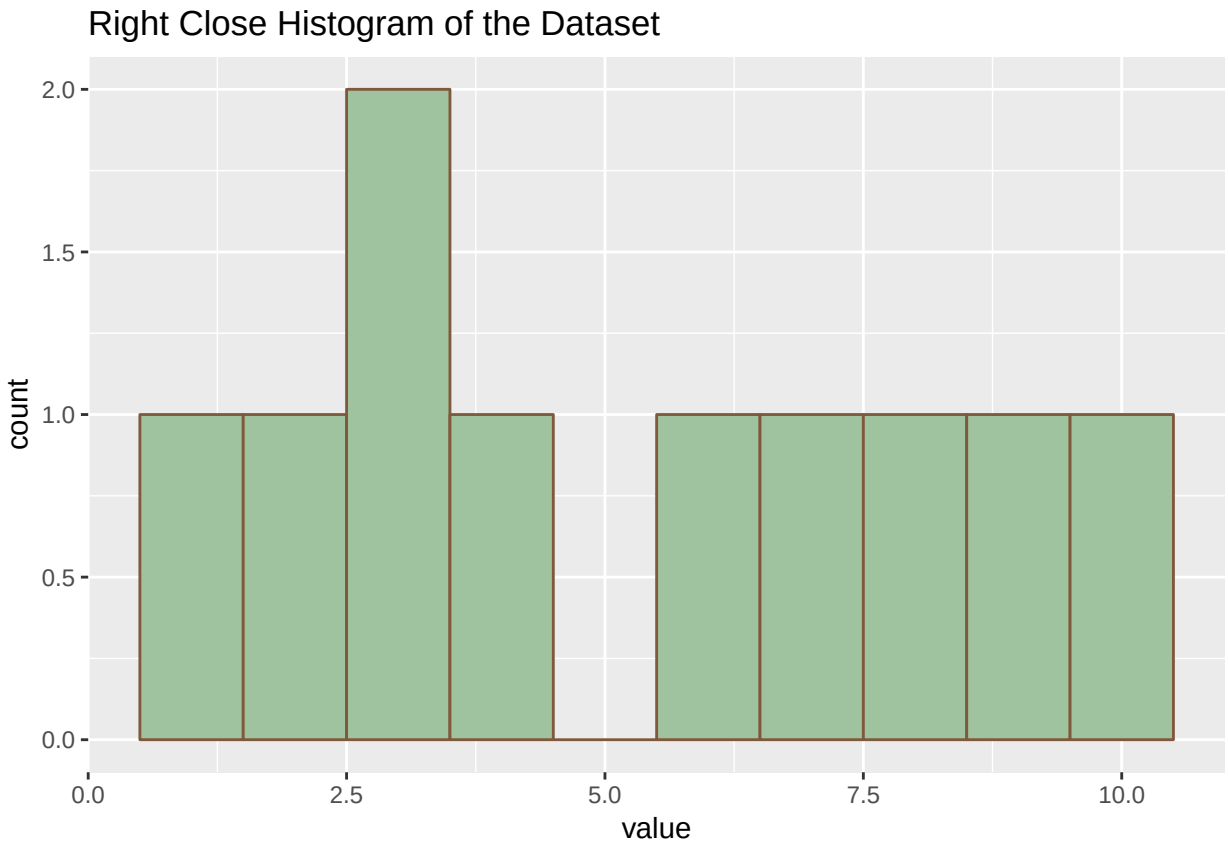


b) Adjust parameters–the same for both–so that the right open and right closed versions become identical. Explain your strategy.

```
# histogram with right open
p3b1 = ggplot(dataset, aes(x = value)) +
  geom_histogram(right = FALSE, binwidth = 1, colour = "#80593D", fill = "#9FC29F") +
  ggtitle("Right Open Histogram of the Dataset")
p3b1
```




```
# histogram with right close
p3b2 = ggplot(dataset, aes(x = value)) +
  geom_histogram(right = TRUE, binwidth = 1, colour = "#80593D", fill = "#9FC29F") +
  ggtitle("Right Close Histogram of the Dataset")
p3b2
```



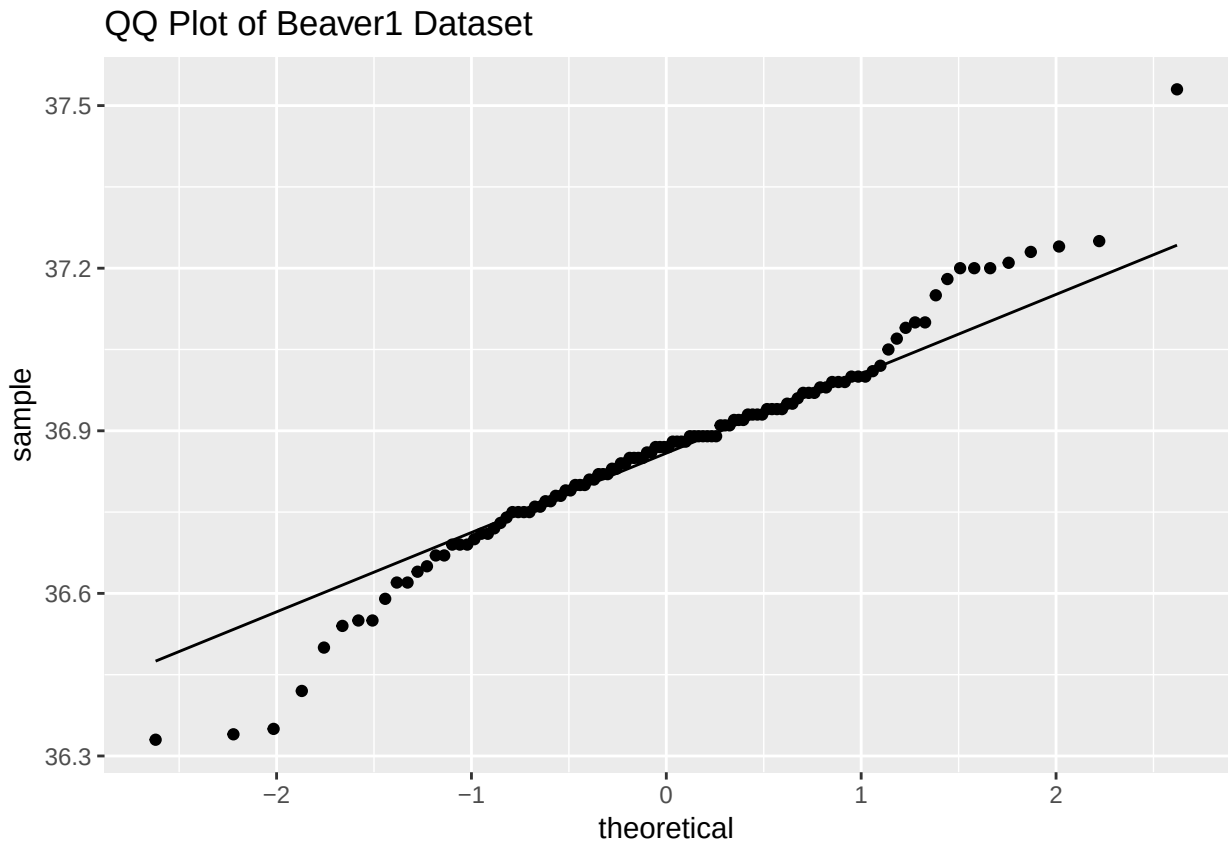
- My strategy is to decrease the bid width to 1, so that there is no number lying on the boundary. So the right open histogram and right close histogram is the same under such circumstance.

4. Beavers

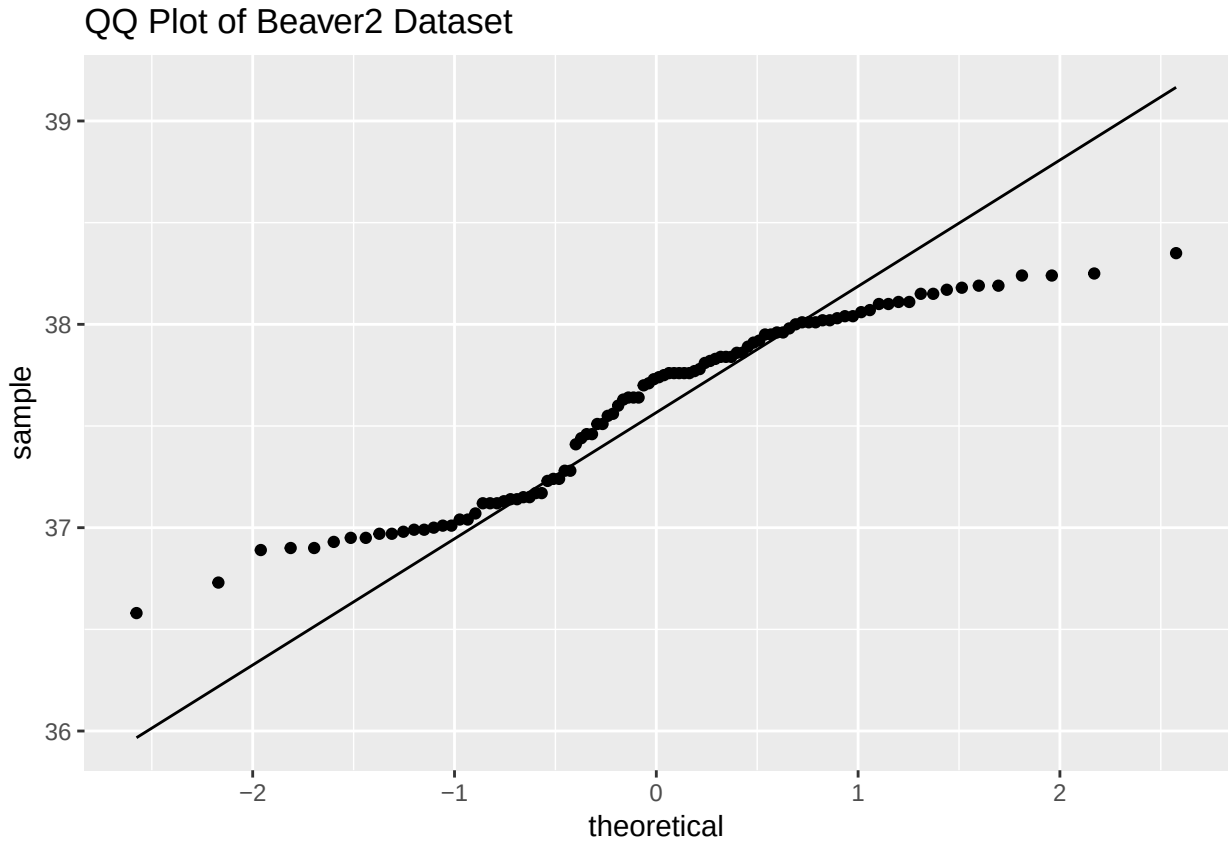
[10 points]

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?

```
# QQ plots for beaver1
p4a1 <- ggplot(beaver1, aes(sample = temp)) +
  stat_qq() +
  stat_qq_line() +
  ggtitle('QQ Plot of Beaver1 Dataset')
p4a1
```




```
# QQ plots for beaver2
p4a2 <- ggplot(beaver2, aes(sample = temp)) +
  stat_qq() +
  stat_qq_line() +
  ggtitle('QQ Plot of Beaver2 Dataset')
p4a2
```

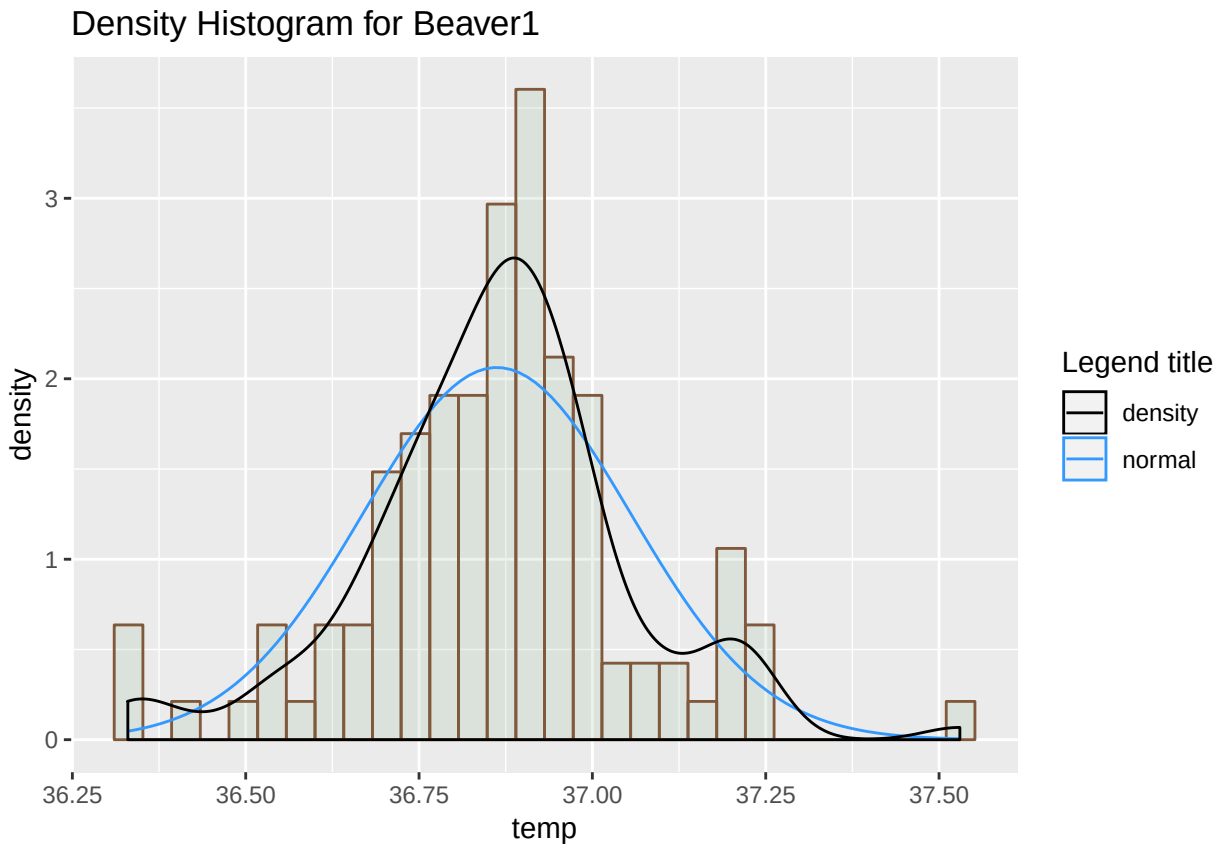


- Dataset beaver1 appears to be more normally distributed.
- The reason is that the QQ plots of beaver1 is closer to theoretical normal lines than beaver2. Actually the QQ plots of dataset beaver1 are nearly going along with the added lines which go through the 25% and 75% quantiles. But the QQ plots of dataset beaver2 are more different from the added line.

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

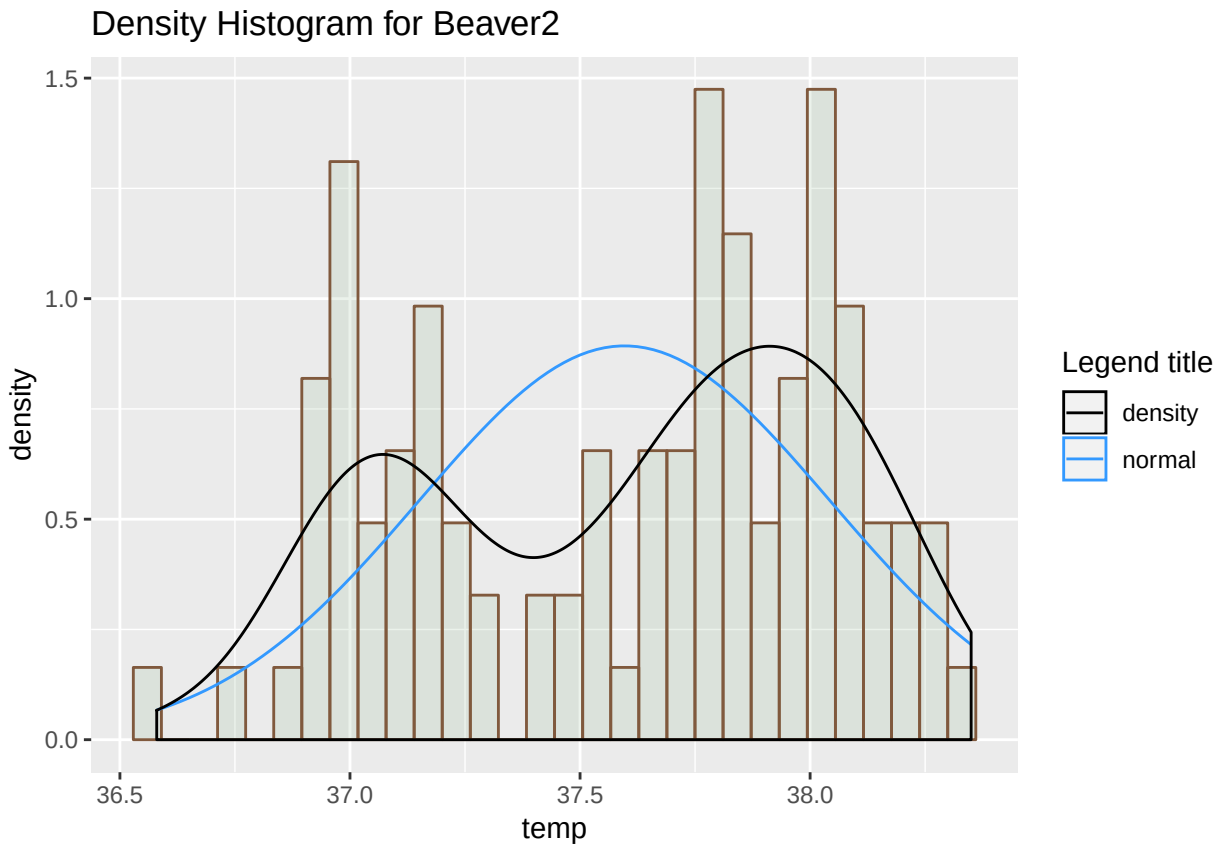
```
# density histogram for beaver1
p4b1 <- ggplot(beaver1, aes(x = temp)) +
  geom_histogram(aes(y = ..density.. ),  colour = "#80593D", fill = "#9FC29F", alpha = 0.2) +
  stat_function(fun = dnorm,
                args = list(mean = mean(beaver1$temp), sd = sd(beaver1$temp)),
                aes(color = "red")) +
  geom_density(aes(color = "Density")) +
  scale_colour_manual("Legend title", values = c("black", "#3399FF"), labels=c("density", "normal")) +
  ggtitle("Density Histogram for Beaver1")

p4b1
```




```
# density histogram for beaver2
p4b2 <- ggplot(beaver2, aes(x = temp)) +
  geom_histogram(aes(y = ..density.. ),  colour = "#80593D", fill = "#9FC29F", alpha = 0.2) +
  stat_function(fun = dnorm,
                args = list(mean = mean(beaver2$temp), sd = sd(beaver2$temp)),
                aes(color = "red")) +
  geom_density(aes(color = "Density")) +
  scale_colour_manual("Legend title", values = c("black", "#3399FF"), labels=c("density", "normal")) +
  ggtitle("Density Histogram for Beaver2")

p4b2
```



- Yes. Based on the histogram and density curve, I find beaver1 is more normally distributed.
- The density curve of dataset beaver1 has very similar shape and tendency compared to the standard normal distribution curve, although they are not totally overlapping. But the density curve of dataset beaver2 has double peaks, and are quite different from the 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)?

```
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
```

- According to the results of Shapiro-Wilk normality test, we find p-value of beaver1 is much larger than that of beaver2.
- Actually the null hypothesis is the sample comes from a normal distribution. If p-value$\leq \alpha$, (α is the significance level), then I will reject the null hypothesis.

- As the p-value of beaver2 is so small (much small than beaver1's), there is high probability that I reject the null hypothesis (and more likely to reject the hull hypothesis than beaver1). That is, data of beaver2 is not normally distributed (and less normally distributed than beaver1).
- So from this result, I conclude that beaver1 is more likely normally distributed. This result matches the conclusion of (a) and (b).

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.*

```
# prepare data
library(boot)

p5a <- ggplot(breslow, aes(x = age, y = y)) +
  geom_histogram(stat = "identity", colour = "#80593D", fill = "#9FC29F") +
  ylab("number of deaths") +
  facet_wrap(~smoke) +
  ggtitle("The Number of Deaths in Different Ages",
          subtitle = "Left for Non-Smokers and Right for Smokers") +
  theme(plot.title = element_text(face = "bold")) +
  theme(plot.subtitle = element_text(face = "bold", color = "grey35"))
p5a
```