EDAV Fall 2019 Probem Set 1

Swapnav Deka and Huazhang Liu

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

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

The datasets in this assignment are from the **ucidata** package which can be installed from GitHub. You will first need to install the `devtools` package if you don't have it:

1. Abalone

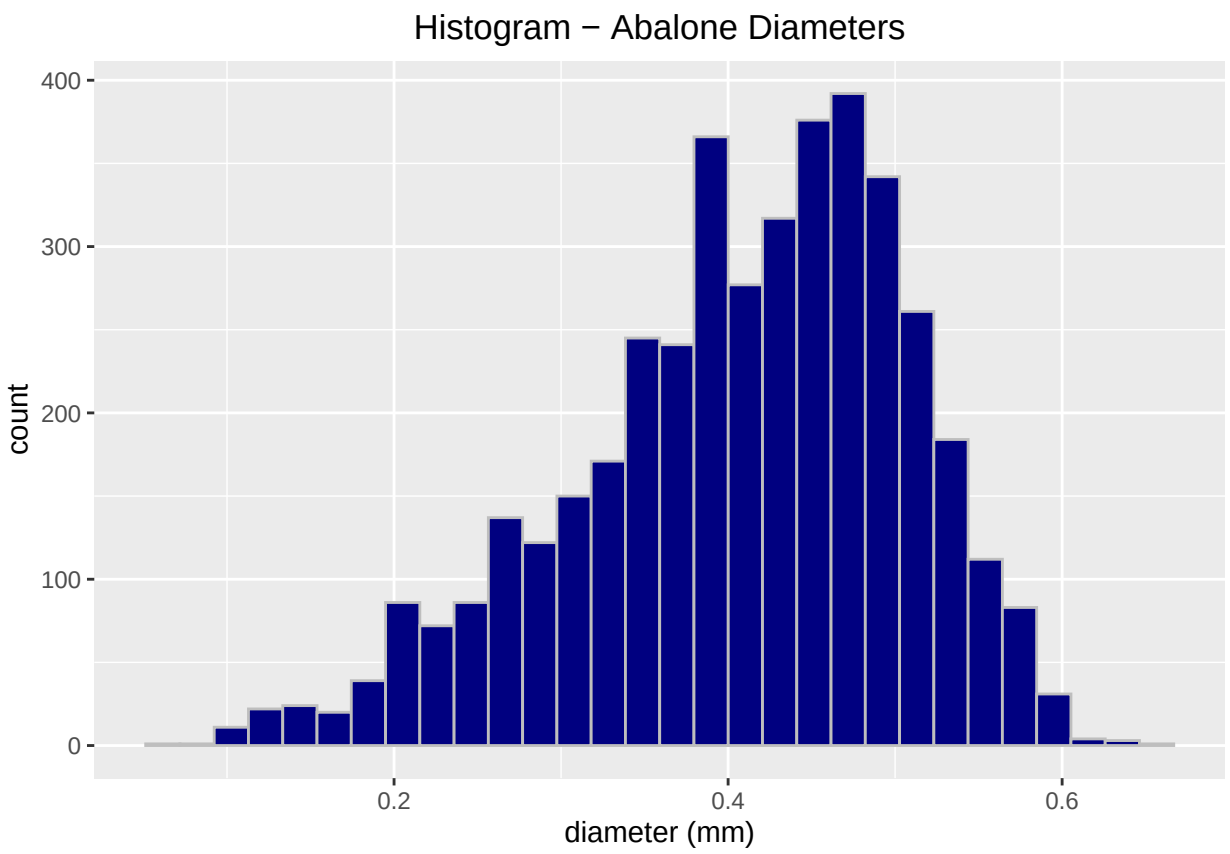
[18 points]

Choose one of the numeric variables in the `abalone` dataset.

```
library(ucidata)
library(ggplot2)
#head(abalone)
```

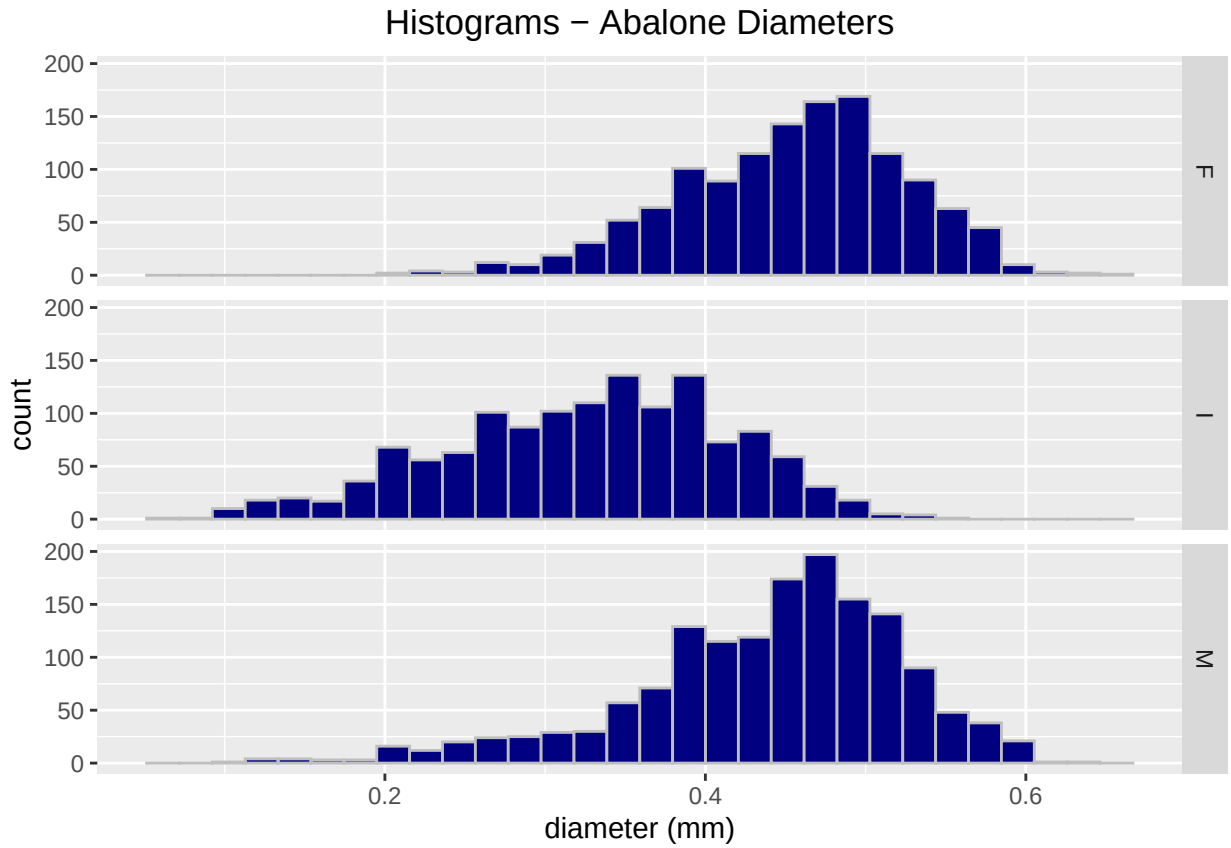
a) Plot a histogram of the variable.

```
ggplot(abalone, aes(x=diameter)) +
  geom_histogram(color = 'grey', fill = "navy") +
  labs(x = 'diameter (mm)', title = "Histogram - Abalone Diameters") +
  theme(plot.title = element_text(hjust = 0.5))
```



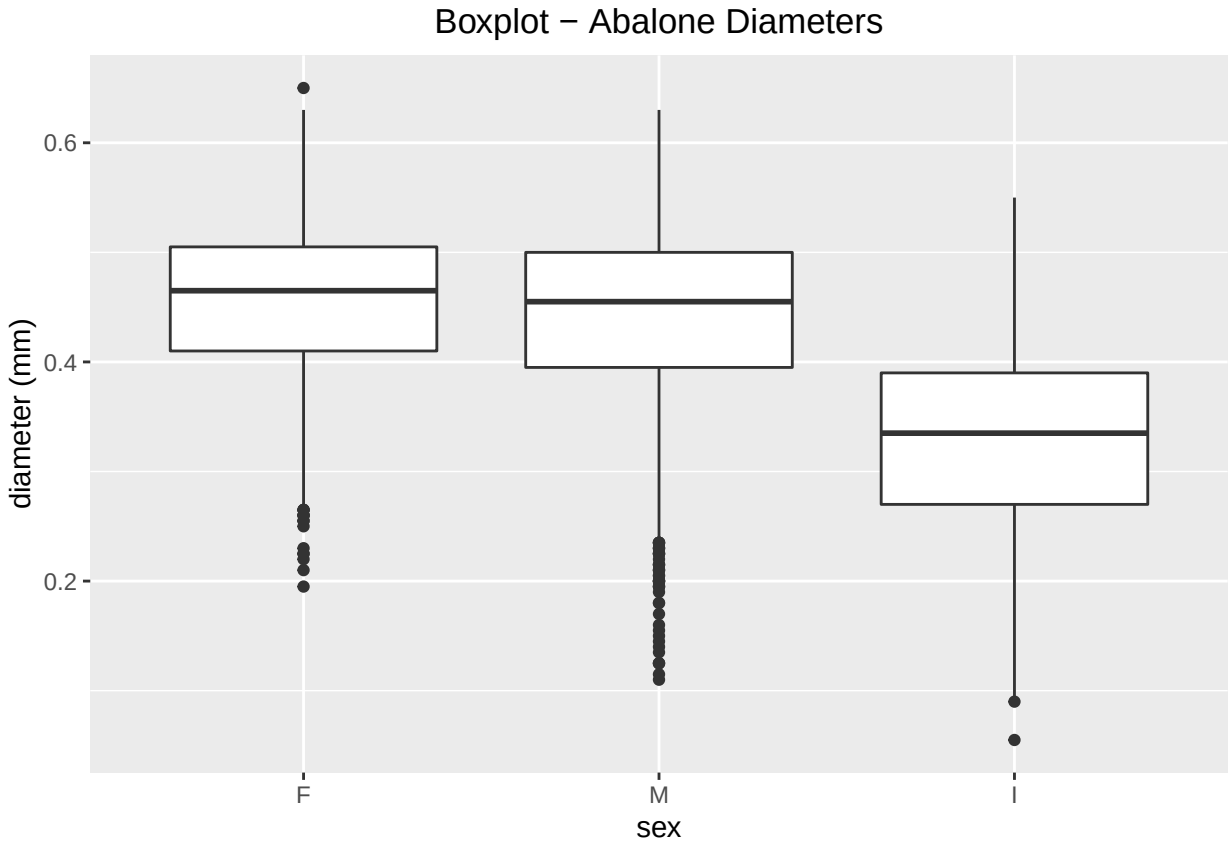
b) Plot histograms, faceted by **sex**, for the same variable.

```
ggplot(abalone, aes(x=diameter)) +
  geom_histogram(color = 'grey', fill = 'navy') +
  facet_grid(sex~.) +
  labs(x = 'diameter (mm)', title = "Histograms - Abalone Diameters") +
  theme(plot.title = element_text(hjust = 0.5))
```



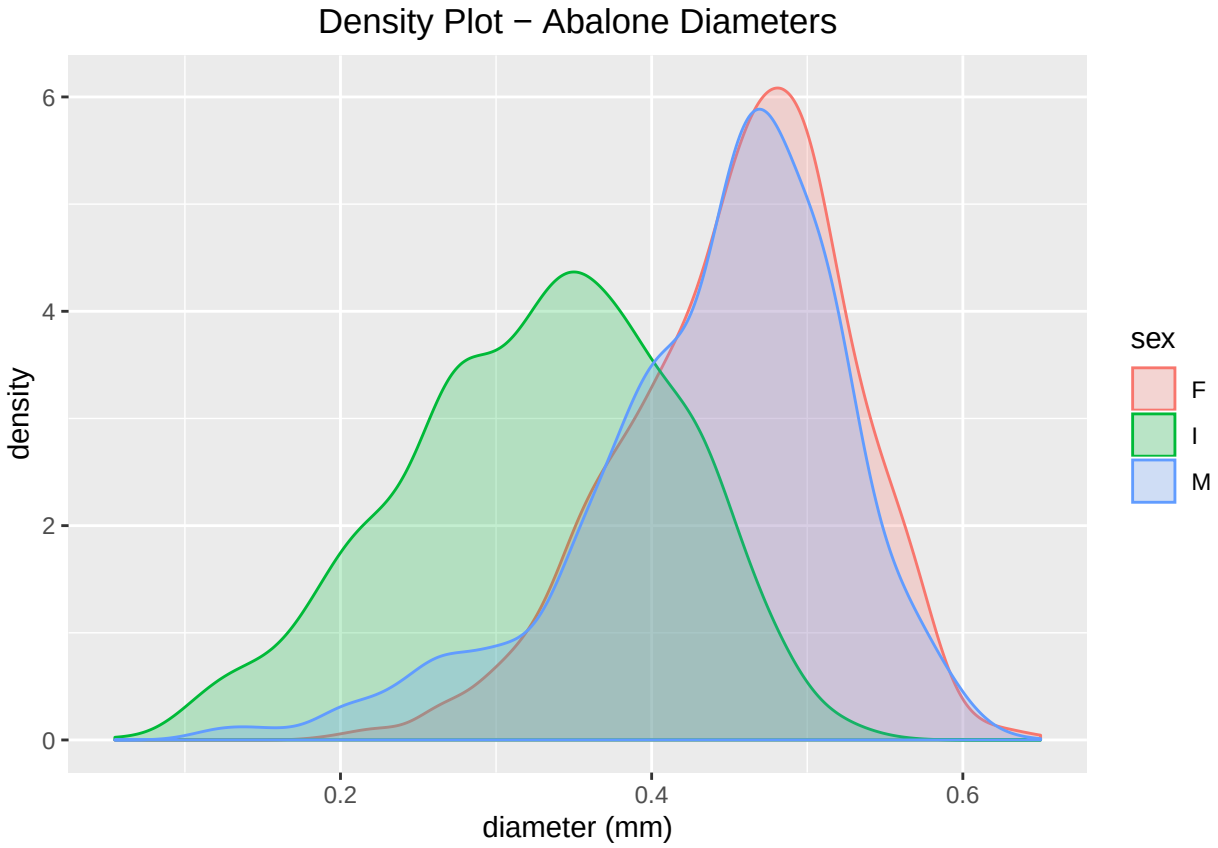
c) Plot multiple boxplots, grouped by **sex** for the same variable. The boxplots should be ordered by decreasing median from left to right.

```
ggplot(abalone, aes(x=reorder(sex, -diameter, median), y=diameter)) +
  geom_boxplot() +
  labs(x = 'sex', y = 'diameter (mm)', title = "Boxplot - Abalone Diameters") +
  theme(plot.title = element_text(hjust = 0.5))
```



d) Plot overlapping density curves of the same variable, one curve per factor level of `sex`, on a single set of axes. Each curve should be a different color.

```
ggplot(abalone, aes(x=diameter, color = sex, fill = sex)) +
  geom_density(alpha = 0.24) +
  labs(x = 'diameter (mm)', title = "Density Plot - Abalone Diameters") +
  theme(plot.title = element_text(hjust = 0.5))
```



e) Summarize the results of b), c) and d): what unique information, *specific to this variable*, is provided by each of the three graphical forms?

The first plot, from the result of part b, shows the distribution of diameters among abalone divided into three separate histograms: one for males, one for females, and one for infants. We can see that the distribution of males and females is nearly identical, while the distribution for infants is shifted left significantly. This makes sense because infants, given their age, are supposed to be much smaller than adults.

The second plot, from the result of part c, provides some insight on the differences betweent the male and female data. The infant data is still easily differentiable, as we can see that the lower 75% of all infant data is located below the 25th percentile (first quartile) of all male and female data. When comparing male and female data, we see that the lower quartile (1st/3rd) and median values. The male data also has a larger number of outliers than the female data.

The third plot, from the result of part d, shows denisty curves. The area under each curve represents 100% of all probabilities, so the relative area under a specific slice shows the likelihood of that diameter. You can see that infants, obviously, are much more likely to be smaller. The male and female curves are similar, but there are sections, such as the slice from 0.2 to 0.3 that shows are higher probability among males, that showcase specific fluctuations in probability between the two groups. Females seem to have a (slightly) larger area under the curve among higher diameters, indicating a slightly larger probability to be larger in size.

f) Look at photos of an abalone. Do the measurements in the dataset seem right? What's the issue?

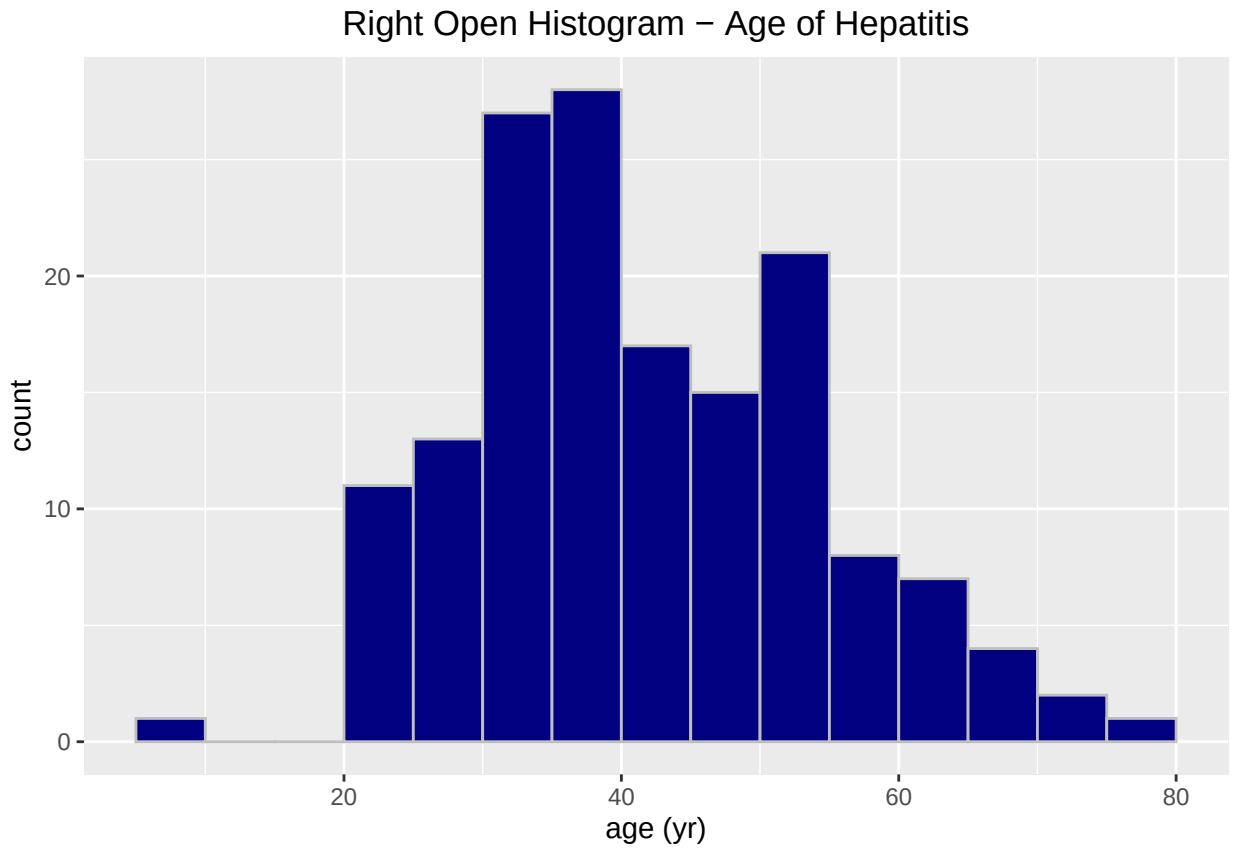
The measurements in the dataset seem to be off by some large factor, perhaps a magnitude of 100. Most abalone range in size from 20 to 200mm, but the dataset (that is recorded in mm) has all measurements as decimal values under 1.

2. Hepatitis

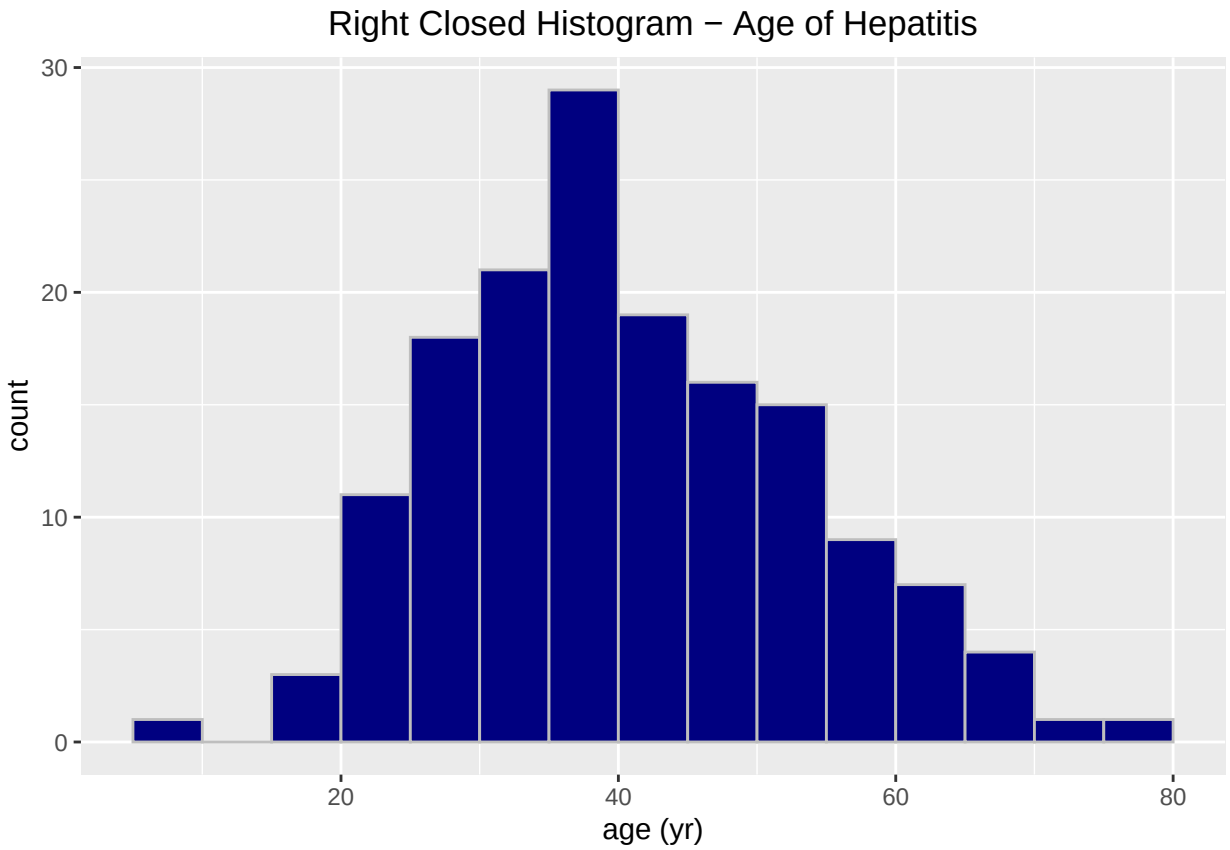
[6 points]

a) Draw two histograms of the age variable in the `hepatitis` dataset in the **ucidata** package, with binwidths of 5 years and `boundary = 0`, one right open and one right closed. How do they compare?

```
ggplot(hepatitis, aes(x=age)) +
  geom_histogram(boundary = 0, binwidth = 5, closed = 'left', color = 'grey', fill = 'navy') +
  labs(x = 'age (yr)', title = "Right Open Histogram - Age of Hepatitis") +
  theme(plot.title = element_text(hjust = 0.5))
```



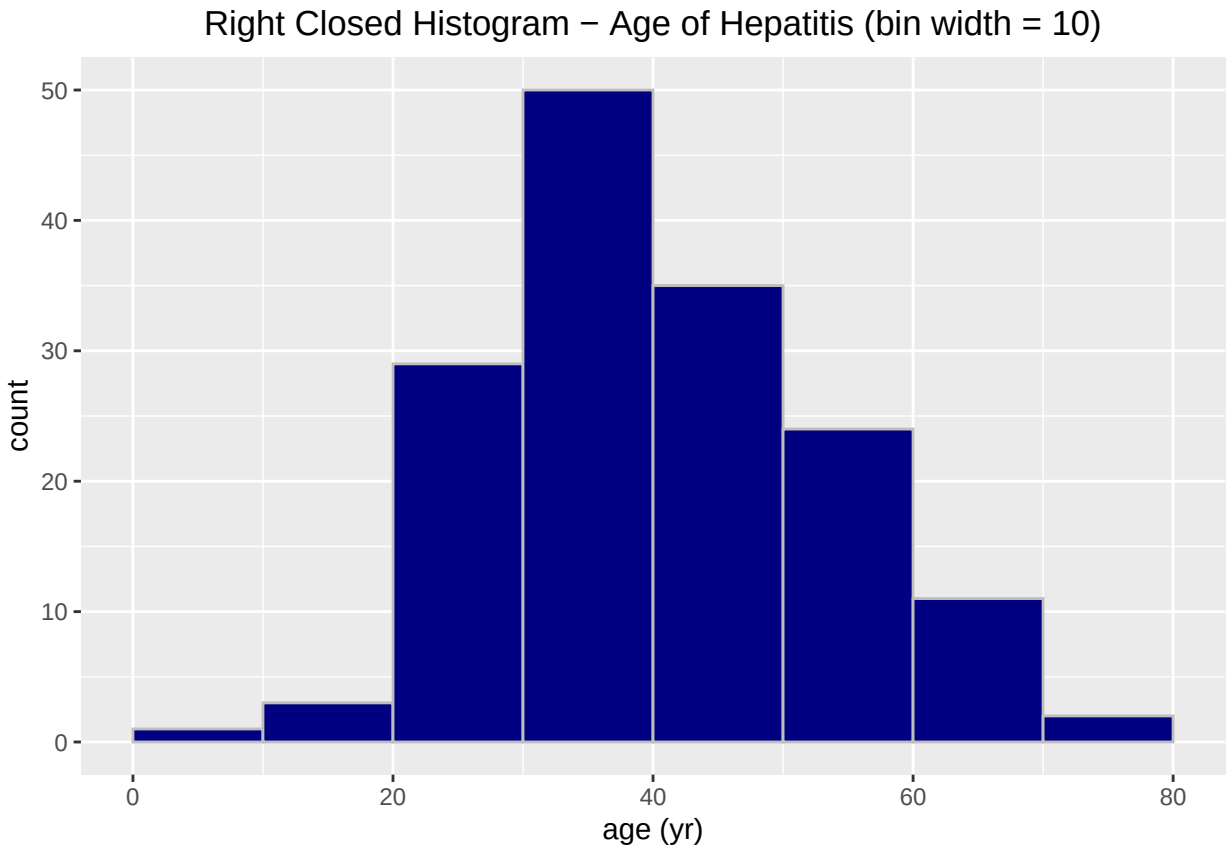

```
ggplot(hepatitis, aes(x=age)) +
  geom_histogram(boundary = 0, binwidth = 5, closed = 'right', color = 'grey', fill = 'navy') +
  labs(x = 'age (yr)', title = "Right Closed Histogram - Age of Hepatitis") +
  theme(plot.title = element_text(hjust = 0.5))
```



The right-closed histogram seems to appear symmetric, potentially indicating a normal distribution. The left-closed histogram is more sporadic, and does not have even tails or a symmetric distribution. This indicates that there are a good number of boundary values in the dataset, because if there were not, then the change in histograms would not be very noticable.

b) Redraw the histogram using the parameters that you consider most appropriate for the data. Explain why you chose the parameters that you chose.

```
ggplot(hepatitis, aes(x=age)) +
  geom_histogram(boundary = 0, binwidth = 10, closed = 'right', color = 'gray', fill = 'navy') +
  labs(x = 'age (yr)', title = "Right Closed Histogram - Age of Hepatitis (bin width = 10)") +
  theme(plot.title = element_text(hjust = 0.5))
```



I chose to redraw the histogram with bin widths of 10. The reason I chose to do this is because decades are an intuitive measure of age: teens, twenties, thirties, etc. Furthermore, there is (albeit just one) a bin not at the tail that is completely empty, so the more intuitive bin size also allows for smoother representation of the data. I chose to use a right closed histogram because when people are a certain age, they are never (except only for a moment) exactly that age. They immediately become second, minutes, days, or months older than the age, so it makes sense to include boundary values in the next bin instead of the preceeding group.

3. Glass

[18 points]

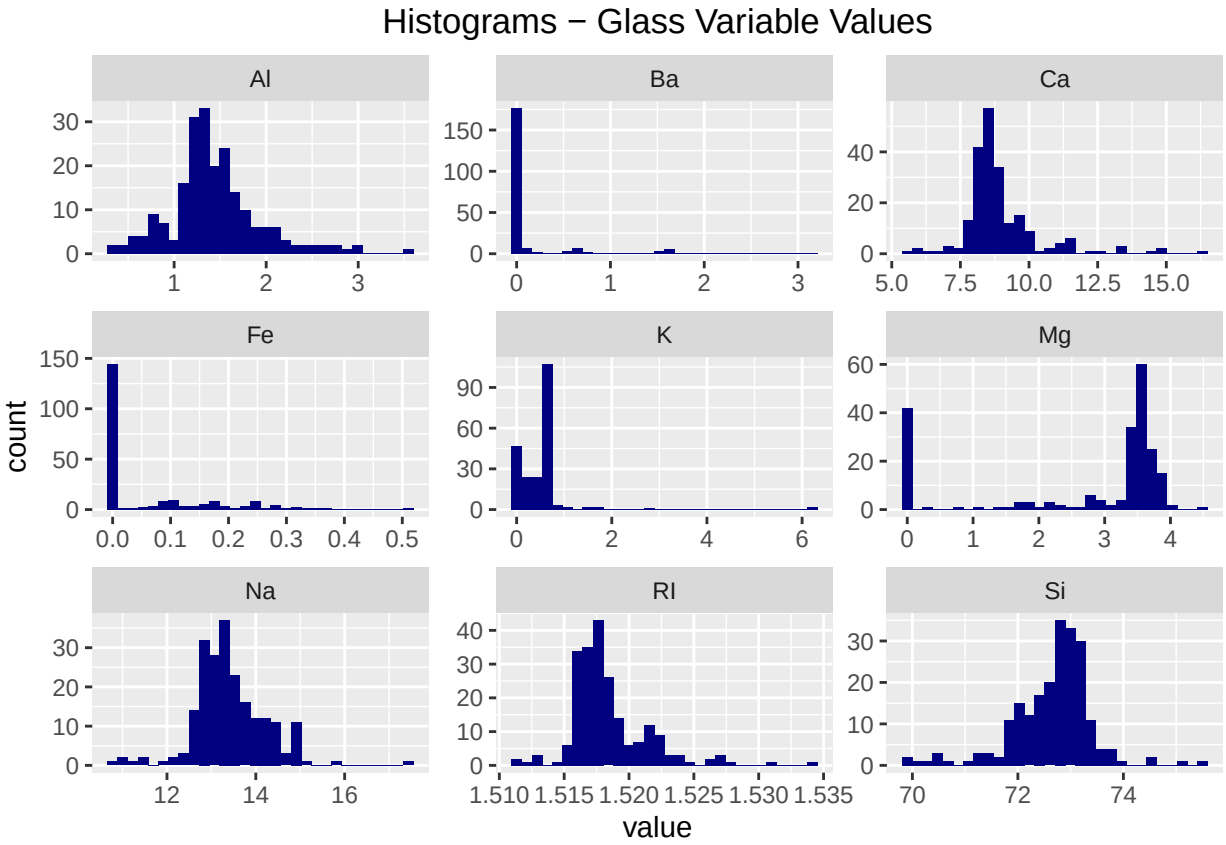
a) Use `tidyr::gather()` to convert the numeric columns in the `glass` dataset in the **ucidata** package to two columns: `variable` and `value`. The first few rows should be:

```
  variable   value
1       RI 1.52101
2       RI 1.51761
3       RI 1.51618
4       RI 1.51766
5       RI 1.51742
6       RI 1.51596
```

Use this form to plot histograms of all of the variables in one plot by faceting on `variable`. What patterns do you observe?

```
library(tidyverse)
newglass <- gather(glass, key = 'variable', value = 'value', 2:10)
```

```
ggplot(newglass, aes(x=value)) +
  geom_histogram(fill = 'navy') +
  facet_wrap(~variable, scales = 'free') +
  labs (title = "Histograms - Glass Variable Values") +
  theme(plot.title = element_text(hjust = 0.5))
```

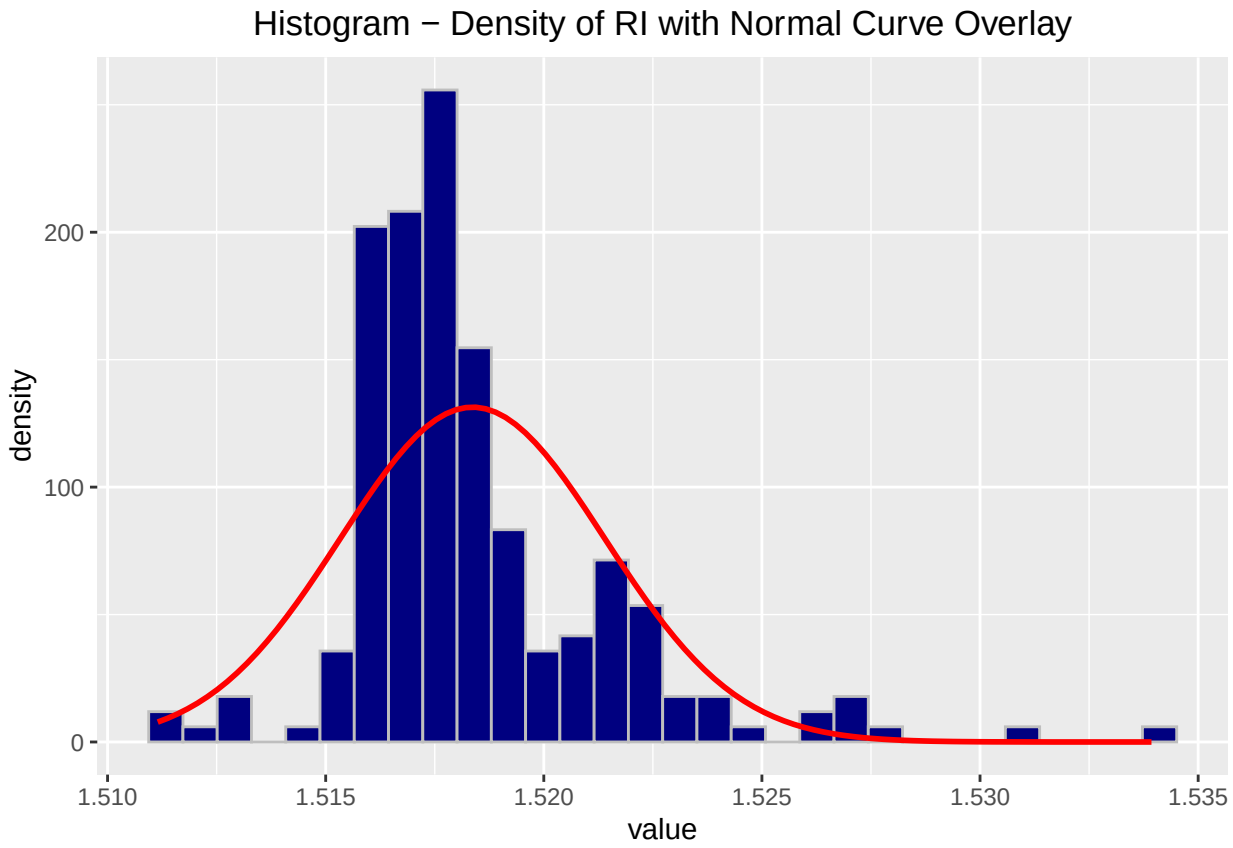


In these 9 histograms, most of them are unimodal (Al, Ba, Ca, Fe, Na, Ri and Si). Mg has a high amount of counts at 0 level and another peak between 3 and 4. K also has a relatively high amount of counts at 0 level, but also has a another peak between 0 to 2. Both Ka and Mg has a large gap between 2 peaks. In addition, Ba, Fe, K have most of the data points close to 0 level. But they also have a few outliers on the right, far from the right side. Ca has the most asymmetry graphs but most of the points are on the left set but not on the center. Al, Na, RI and Si are unlikely to be asymmetric. Most of data points of Al, Na and RI are on the left but the majority of Si is on the right. Al, Ca, Na, RI and Si have outliers on the right. Since the scale is quite different in each plot, the graphs are unlikely to compare with each other.

For the remaining parts we will consider different methods to test for normality.

b) Choose one of the variables with a unimodal shape histogram and draw a true normal curve on top on the histogram. How do the two compare?

```
ggplot(glass, aes(x = RI)) +
  geom_histogram(aes(y = ..density..), color = "gray", fill = "navy") +
  stat_function(fun = dnorm, args = list(mean = mean(glass$RI), sd = sd(glass$RI)),
                color = "red", size = 1) +
  labs(x = "value", title = "Histogram - Density of RI with Normal Curve Overlay") +
  theme(plot.title = element_text(hjust = 0.5))
```



The histogram has the similar peak with the normal curve. However, the peak is on the left a little bit and the right side of the peak is much lower than the peak. The majority of the data is on the left side but not distributed equally around the peak. Also, the shape of the histogram looks much different from the red normal curve. It also has the outliers on the right side. The data from the histogram seem to be right skewed.

c) Perform the Shapiro-Wilk test for normality of the variable using the `shapiro.test()` function. What do you conclude?

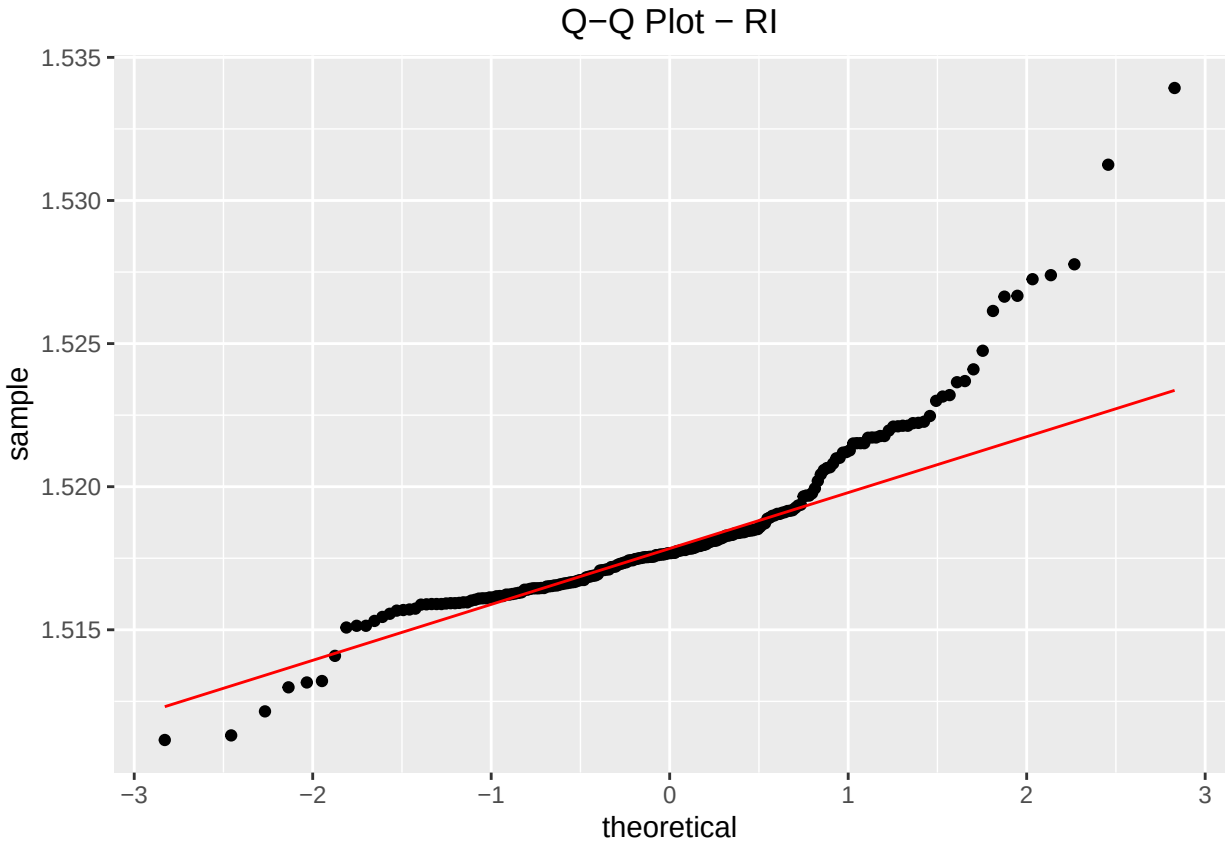
```
shapiro.test(glass$RI)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  glass$RI
## W = 0.86757, p-value = 1.077e-12
```

According to the Shapiro-Wilk normality test, the p-value is 1.077e-12 and much less than 0.05. So we will reject the null hypothesis, which is that data are normally distributed. Therefore, we can conclude that based on the test, the distrubution of RI in glass dataset is not normal.

d) Draw a quantile-quantile (QQ) plot of the variable. Does it appear to be normally distributed?

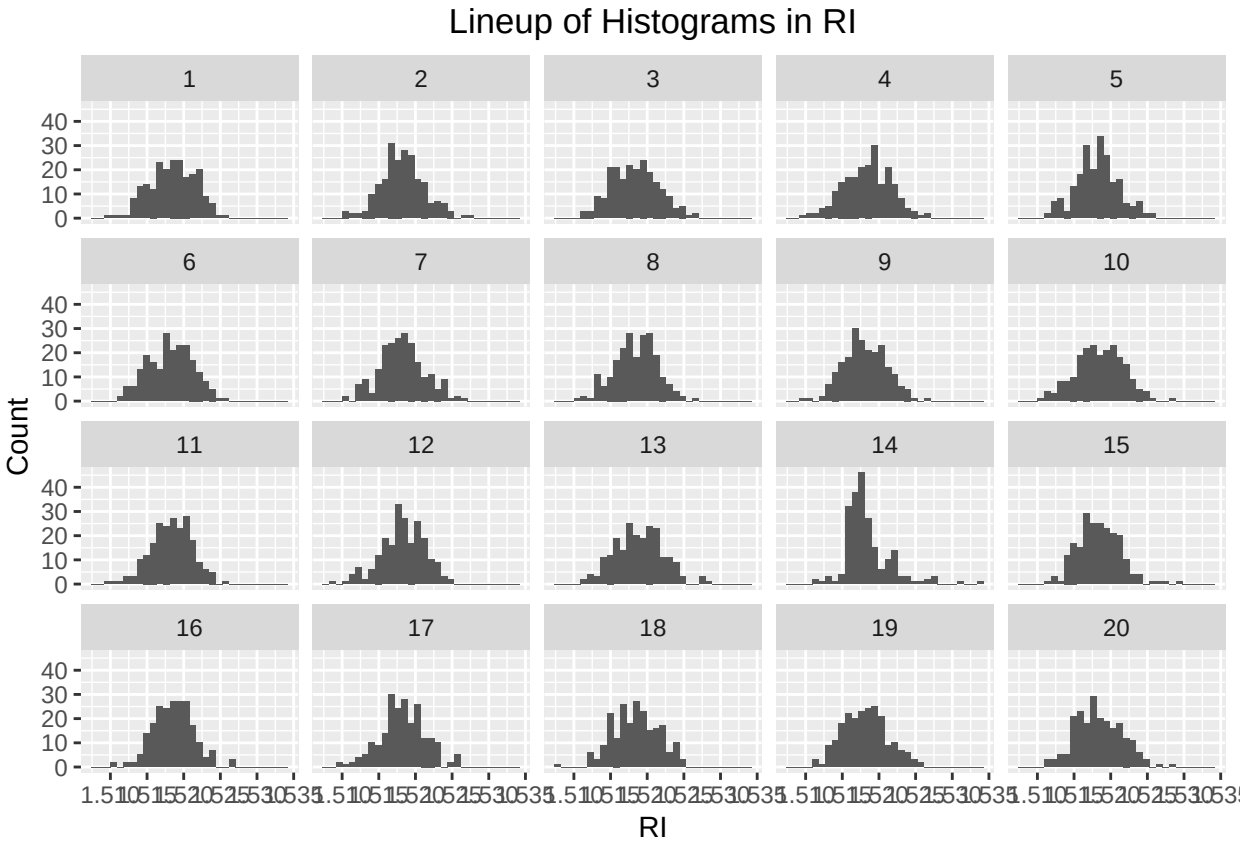
```
ggplot(glass, aes(sample = RI)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot - RI") +
  theme(plot.title = element_text(hjust = 0.5))
```



In this QQ-plot, although most of the dots between -1.5 to 0.5 of theoretical quantitles looks like a straight line, after 1 level at theoretical quantiles, the dots become divergent and do not appear to be a straight line. Similarly, before the -2 level at theoretical quantiles, the dots are also divergent. So it cannot be concluded that the variable appears to be normally distrubuted. Instead, the qqplot supports the previously stated observation that the data appeared to be right skewed. This can be seen by the exponential curve-like nature of the points after x = 1. The slight hump could also imply slight bimodality that we did not notice originally.

e) Use the **nullabor** package to create a lineup of histograms in which one panel is the real data and the others are fake data generated from a null hypothesis of normality. Can you pick out the real data? If so, how does the shape of its histogram differ from the others?

```
library(nullabor)
set.seed(20)
RI_lineup <- lineup(null_dist("RI", dist = "normal"), glass)

ggplot(data = RI_lineup, aes(x = RI)) +
  geom_histogram(bins = 30, aes(y = ..count..)) +
  facet_wrap(~.sample) +
  labs(x = "RI", y = "Count", title = "Lineup of Histograms in RI") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
#attr(RI_lineup, "pos")
```

We would predict the real data to be number 14. It appears to right skewed like our observations showed and Q-Q plot confirmed. The peak is further left than all the other plots. Further, we commented that the Q-Q plot showed a slight sign of possible bimodality, and this is also visible slightly on plot 14. In the remaining plots, the data peaks are centered in the middle. The peak of 14 is also significantly taller than the other plots.

f) Show the lineup to someone else, not in our class (anyone, no background knowledge required). Ask them which plot looks the most different from the others. Did they choose the real data?

The friend I asked also picked plot 14. He has some statistical background, so he also stated he thought it looked like a mixture distribution and recommended I evaluate it with "kolgomorov smirnoff." However, I don't know what that is and asked him what he thought visually. He stated that the same things we saw, citing the shifted, taller peak and the obvious visual differences from other plots.

g) Briefly summarize your investigations. Did all of the methods produce the same result?

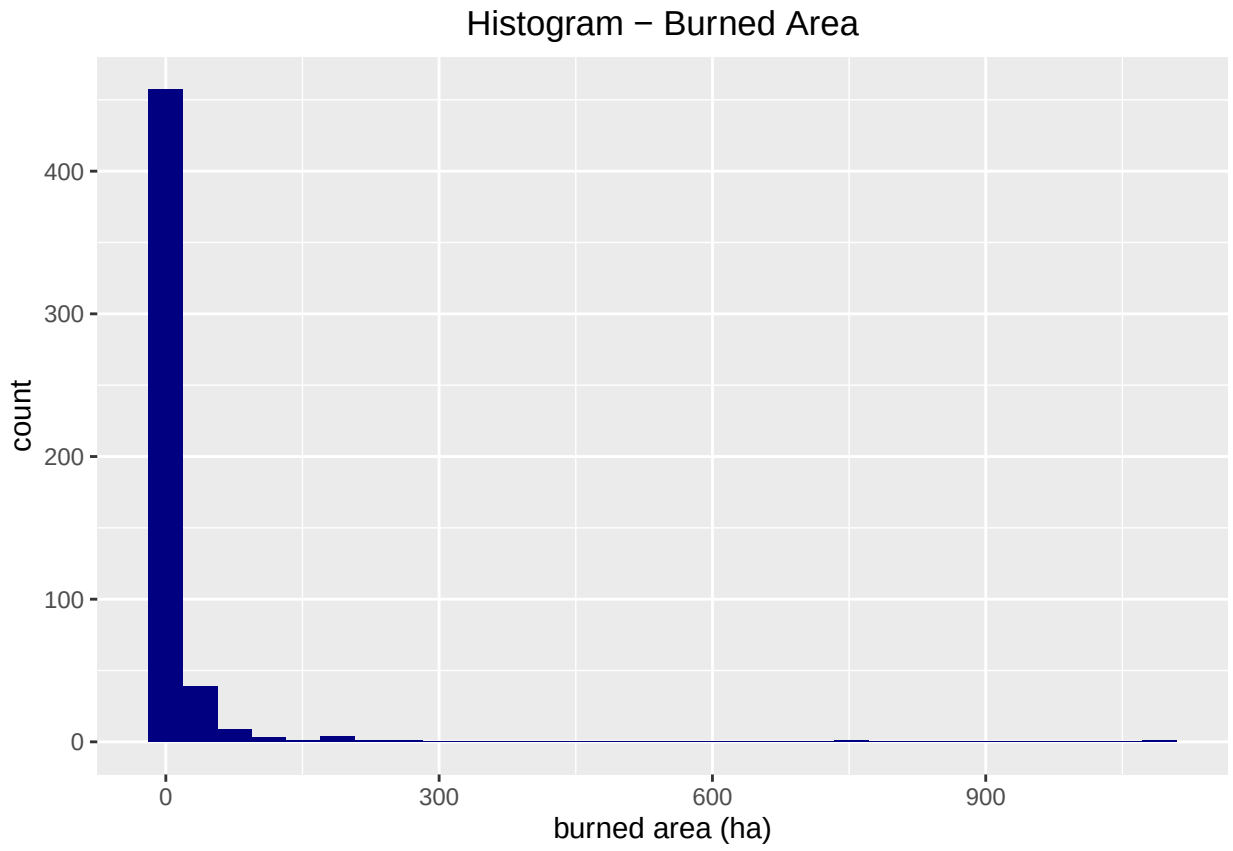
Based on all the investigations, the RI data are not likely to be normal distributed. Most of the data are concentrated on the left, and the peak is also on the left, not in the center, of the histogram. There is a large gap between the peak and other parts of the histograms, epscially on the left. The Q-Q plot supported our hypothesis of right skew and also showed some appearance of bimodality. It also overall indicated, along with the Shapiro test, that the data was not normal. Using the RI lineup plots, we were able to easily pick out the real data and saw that it differed very clearly from normal distributions. All the steps we took during this investigation lead to the same conclusion, that the data are not normal.

4. Forest Fires

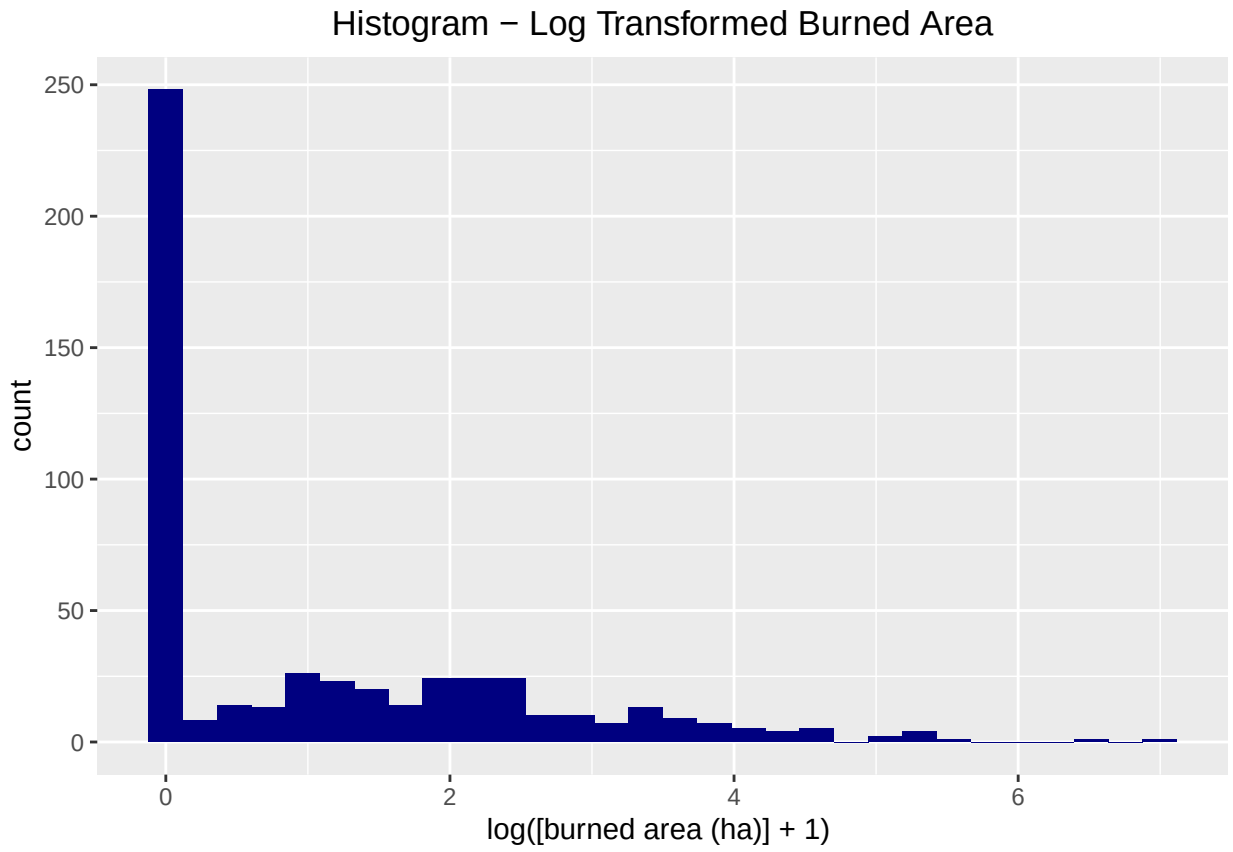
[8 points]

Using the `forest_fires` dataset in the **ucidata** package, analyze the burned area of the forest by month. Use whatever graphical forms you deem most appropriate. Describe important trends.

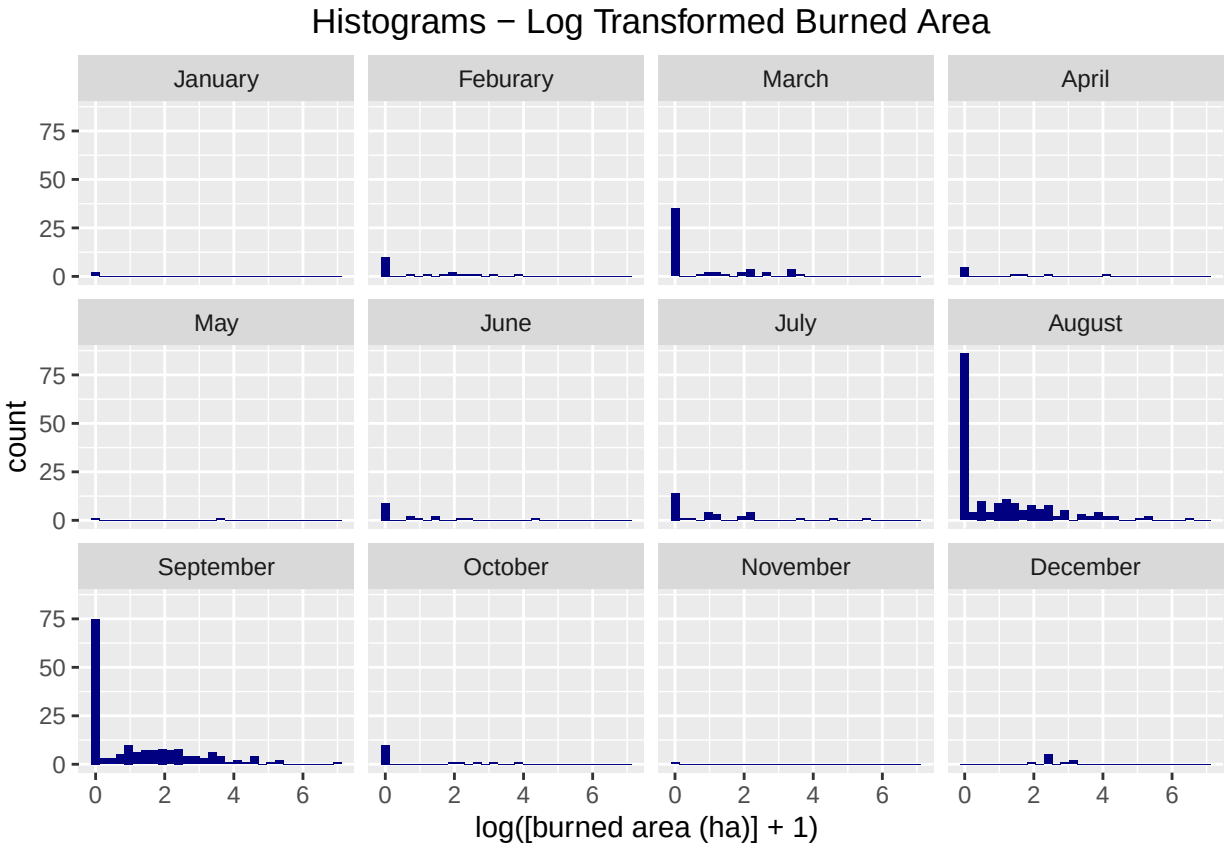
```
levels(forest_fires$month) <- c("April", "August", "December", "Feburary",
                                "January", "July", "June", "March", "May",
                                "November","October", "September")

forest_fires$month_timeorder <- factor(forest_fires$month,
                                       levels = c("January", "Feburary", "March",
                                                  "April", "May", "June", "July",
                                                  "August", "September", "October",
                                                  "November", "December"))
ggplot(forest_fires, aes(x=area)) +
  geom_histogram(fill = 'navy') +
  labs(x = "burned area (ha)", title = "Histogram - Burned Area") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(forest_fires, aes(x=log(area +1))) +
  geom_histogram(fill = 'navy') +
  labs(x = "log([burned area (ha)] + 1) ", title = "Histogram - Log Transformed Burned Area") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(forest_fires, aes(x=log(area +1))) +
  geom_histogram(fill = 'navy') +
  facet_wrap(~month_timeorder) +
  labs(x = "log([burned area (ha)] + 1) ", title = "Histograms - Log Transformed Burned Area") +
  theme(plot.title = element_text(hjust = 0.5))
```



I chose to use a histogram to visualize the burned area by month. Before splitting the data by month, I wanted to see what it looked like overall. The data is extremely right-skewed, so I applied the logarithmic function to help with skewness and symmetry. This definitely helped visualize the data. I moved forward with this choice when faceting the histogram by month. The histograms show that overwhelmingly the largest number of fires occur in the months of August and September. March seems to have a moderate number of small fires that occur, as well. The number of small fires seems to disproportionately outnumber larger fires in almost every month. Some other important trends include spikes in fires during months with traditional long breaks/vacations, including summer and spring breaks. The colder months seem to have much fewer fire incidents than the hot months. It could be interesting to further facet the data by day of week to see how working hours affect the trends as well.