EDAV Fall 2019 Probem Set 1

Benjamin Livingston and Zhiyi Guo

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:

```
install.packages("devtools")
```

then,

```
devtools::install_github("coatless/ucidata")
```

```
devtools::install_github("coatless/ucidata")
library(ggplot2)
```

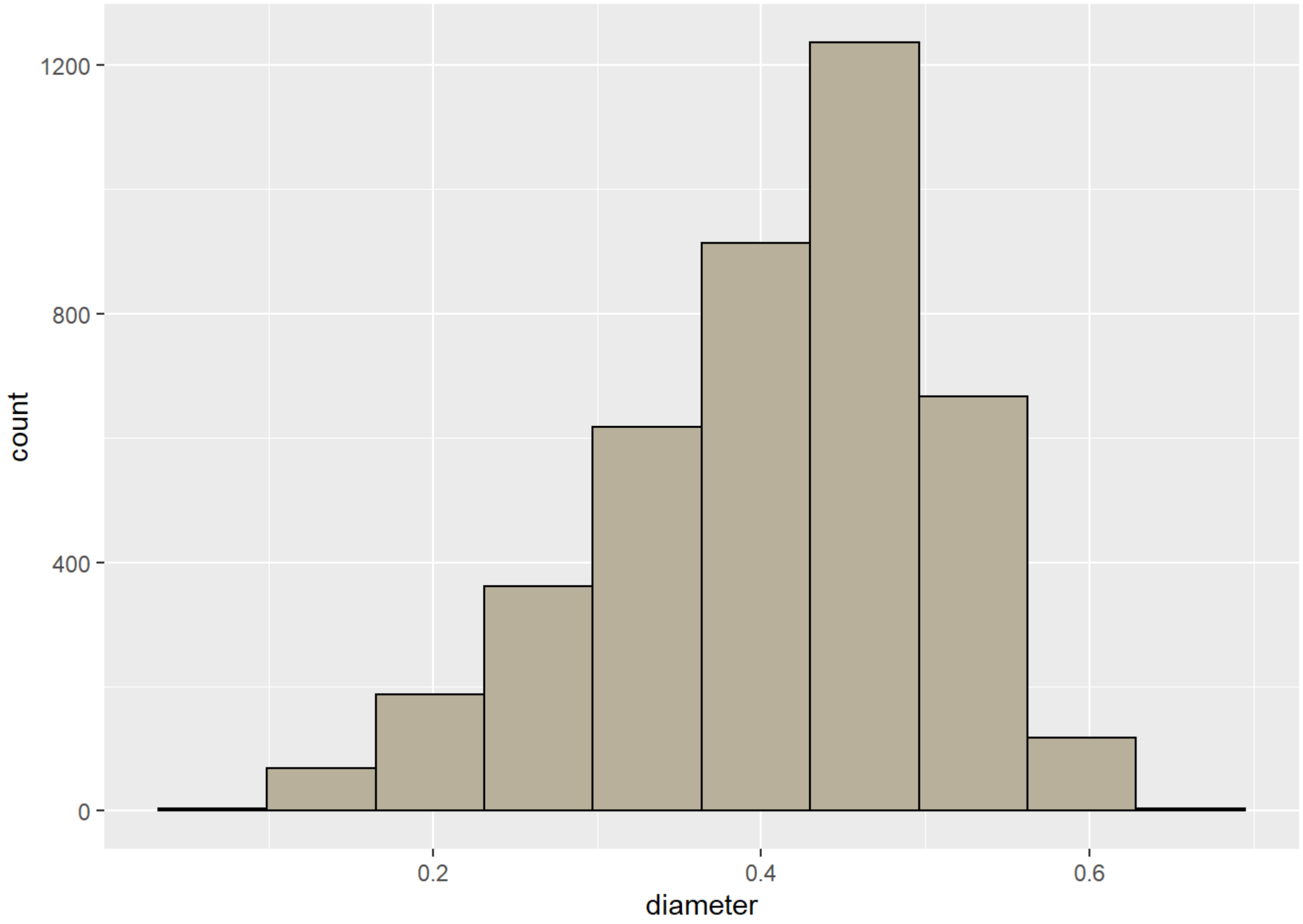
1. Abalone

[18 points]

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

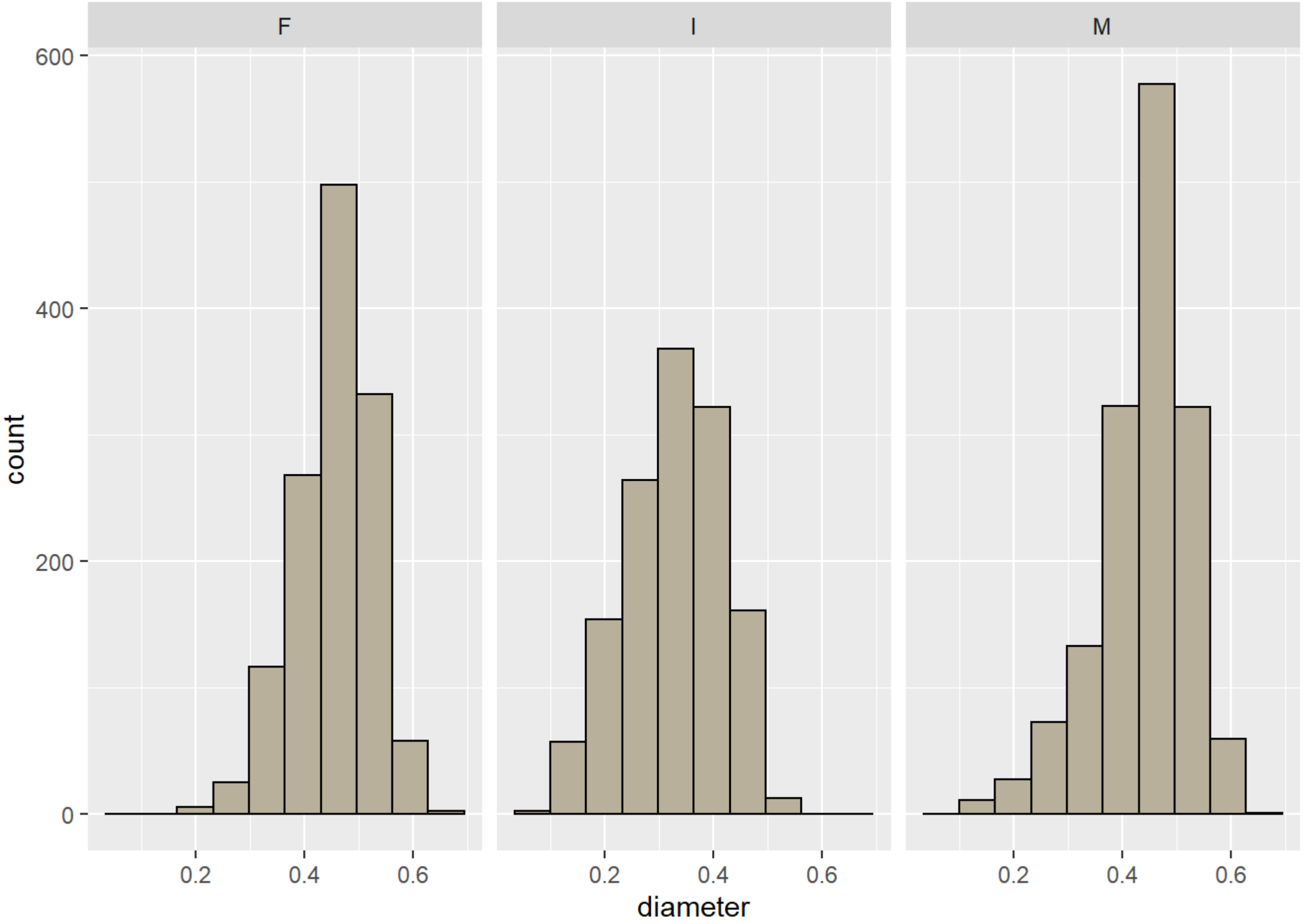
a. Plot a histogram of the variable.

```
abalone <- ucidata::abalone
ggplot(abalone, aes(x = diameter)) +
  geom_histogram(bins = 10, color = "black", fill = '#B8B09B')
```



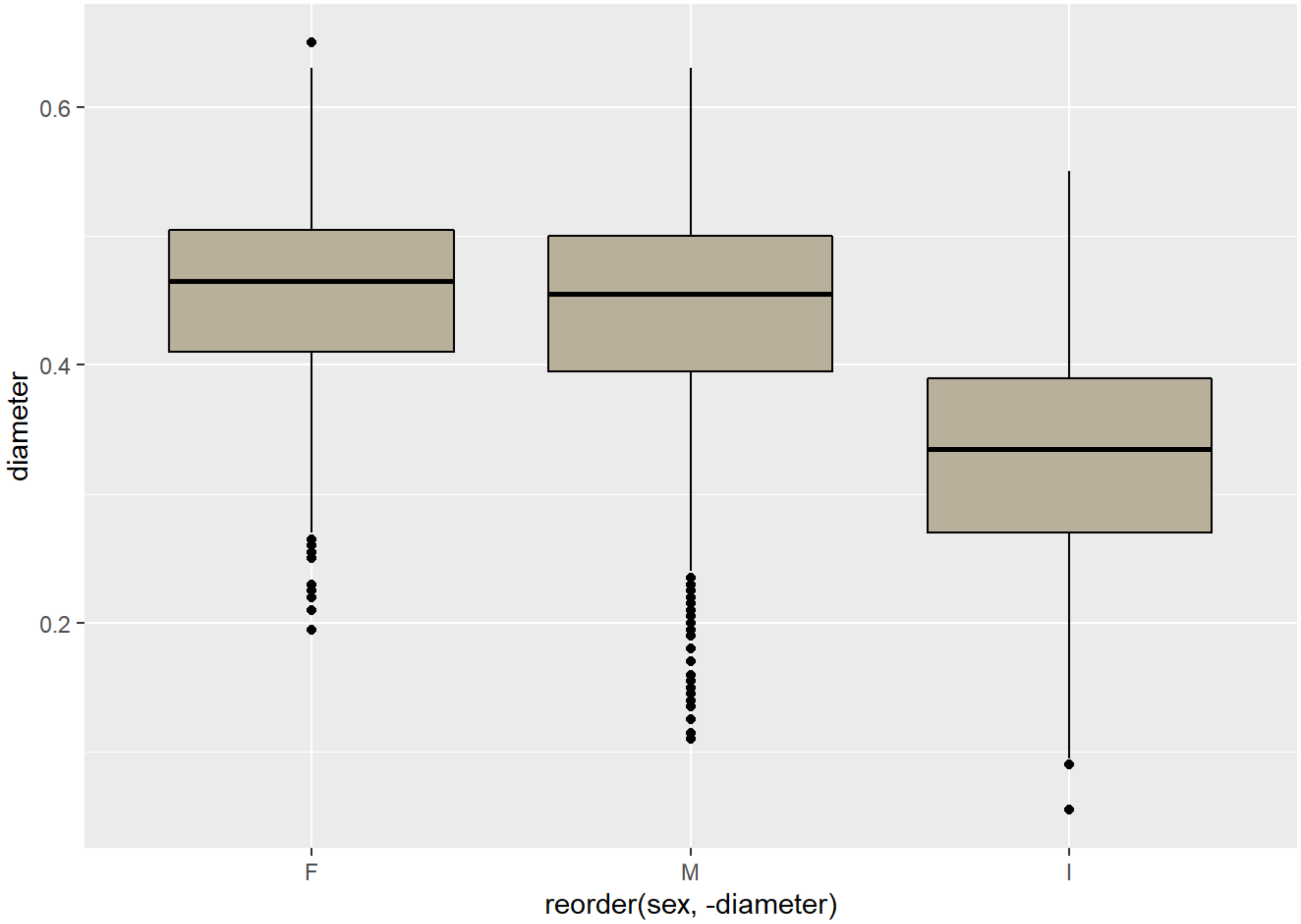
b. Plot histograms, faceted by `sex` , for the same variable.

```
ggplot(abalone, aes(x = diameter)) +
  geom_histogram(bins = 10, color = "black", fill = '#B8B09B') +
  facet_wrap(~sex)
```



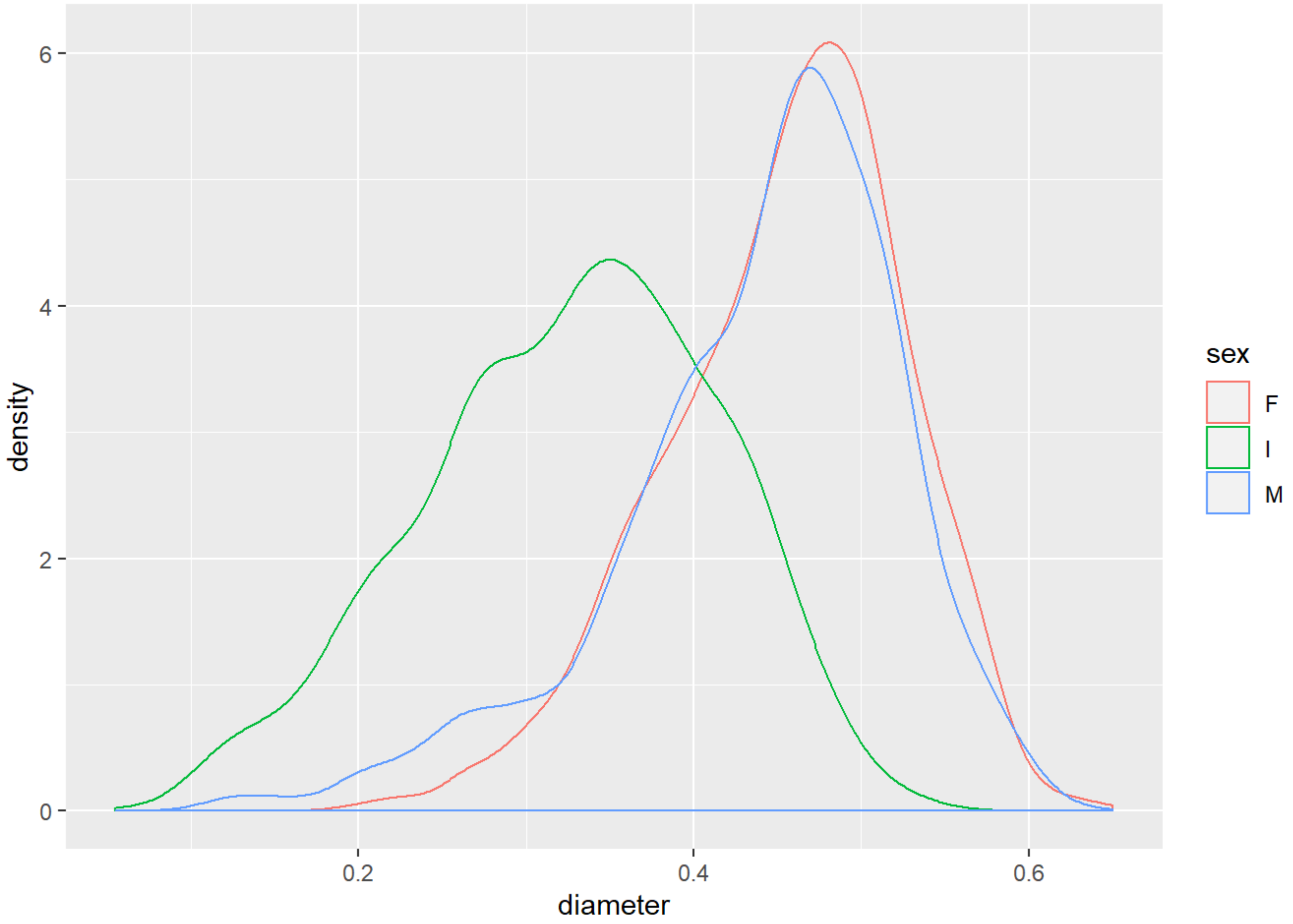
c. Plot multiple boxplots, grouped by `sex` for the same variable. The boxplots should be ordered by decreasing median from left to right.

```
ggplot(abalone, aes(x = reorder(sex,-diameter), y = diameter)) +
  geom_boxplot(fill = '#B8B09B', color = 'black')
```



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)) +
  geom_density()
```



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 histogram in Part a shows us that the diameters are left-skewed. Part b reveals that this skewness is primarily a product of the female and male abalone, and the infants have diameters that are more normally distributed (although there is some skewness there, too). The boxplot in Part c reveals a number of male and female diameter outliers that are likely causing this skewness. The density curves further reveal that the male and female abalone diameters have very similar, left-skewed distributions, whereas the infant diameters are more normally distributed with a much lower mean value.

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

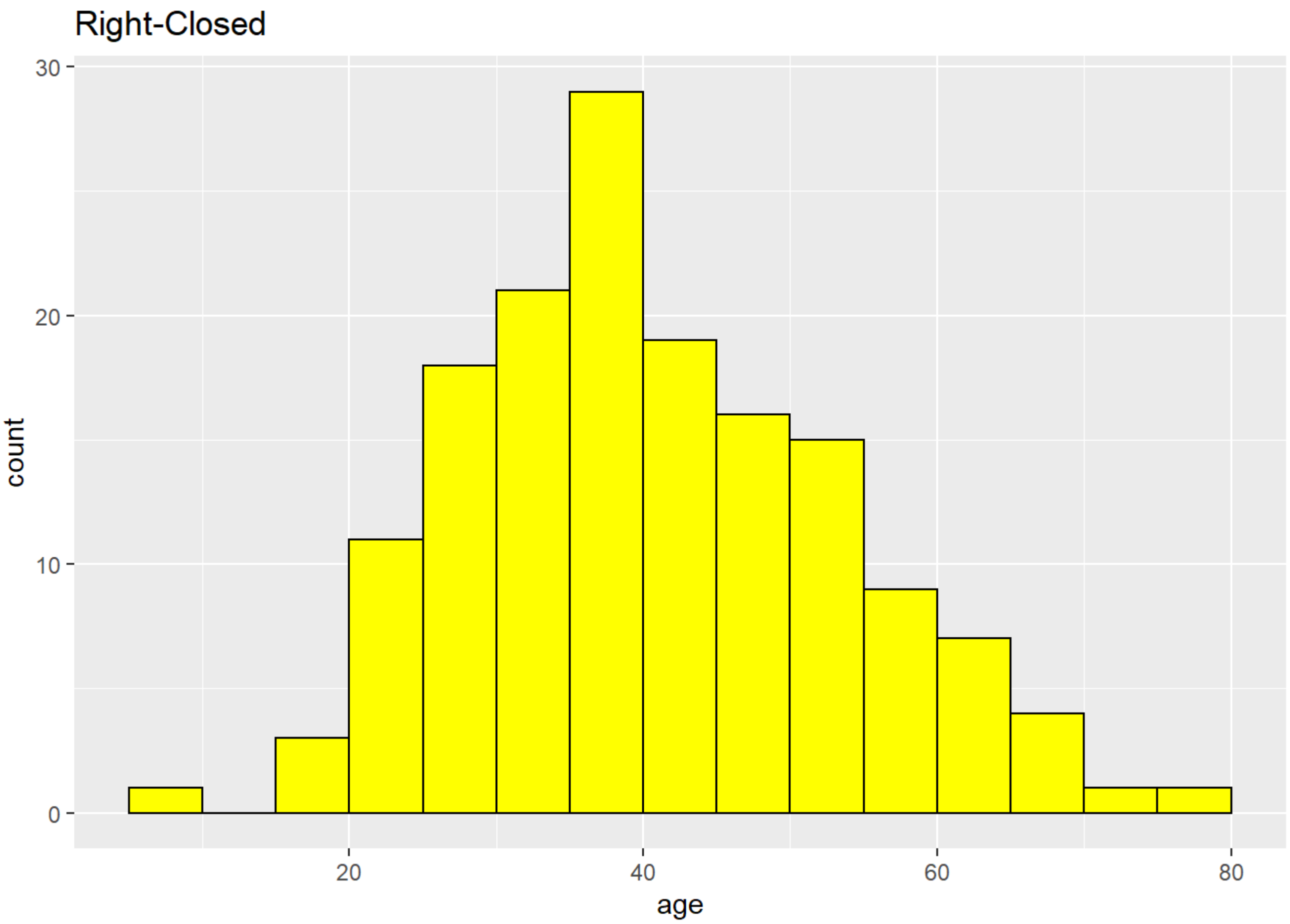
The measurements themselves seem right, but given that Abalone are amorphous and certainly not rectangular, it seems unreasonable to concretely distill their dimensions into length, diameter, and height.

2. Hepatitis

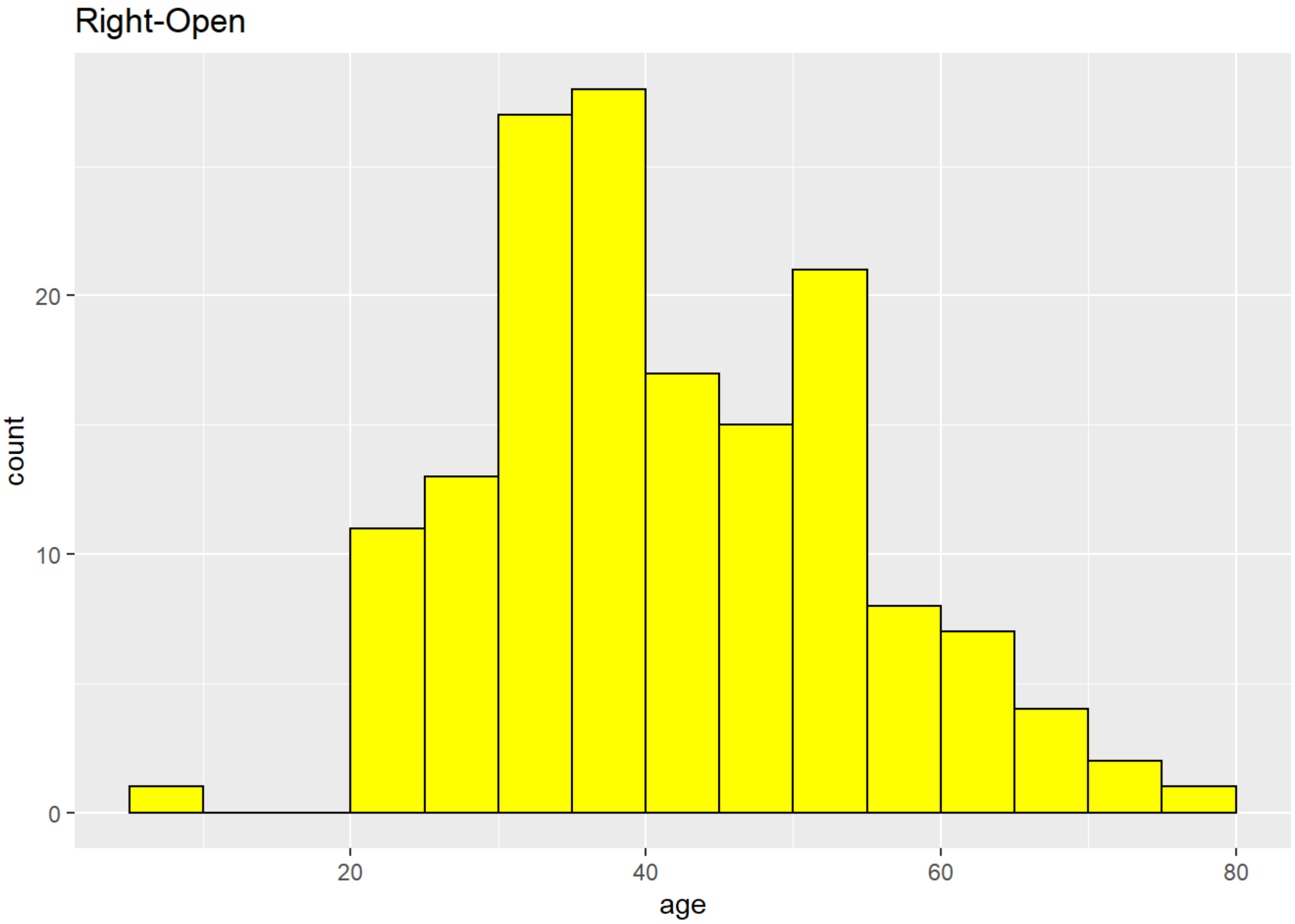
[6 points]

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?

```
hepatitis <- ucidata::hepatitis
ggplot(hepatitis,aes(x=age)) +
  geom_histogram(binwidth = 5, boundary = 0,color = 'black', fill = 'yellow') +
  ggtitle('Right-Closed')
```



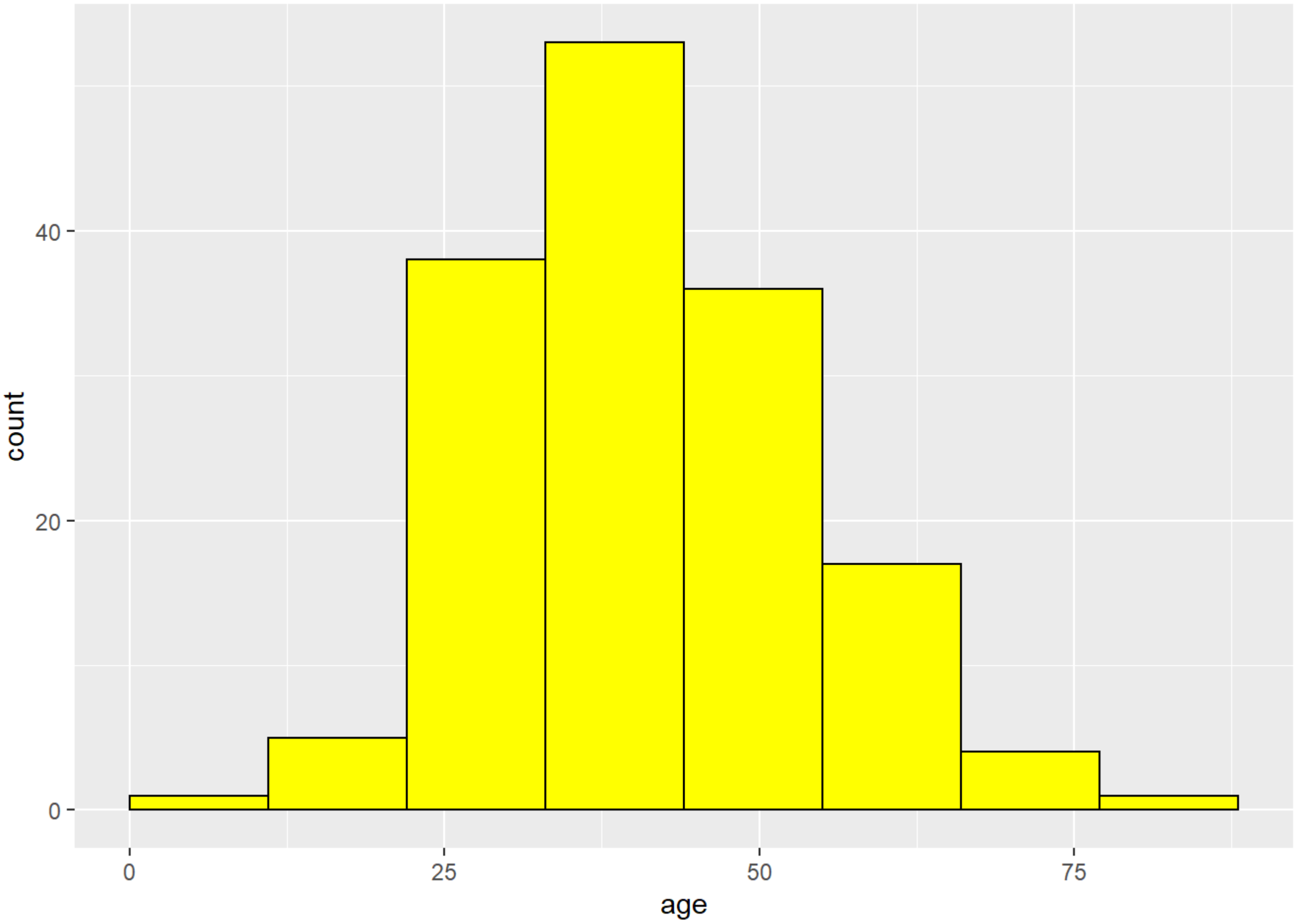
```
ggplot(hepatitis,aes(x=age)) +
  geom_histogram(binwidth = 5, boundary = 0,color = 'black', right = FALSE, fill = 'yellow') +
  ggtitle('Right-Open')
```



The right-open histogram looks almost bimodal, and at the very least we can say it doesn't look normally distributed. The right-closed histogram makes the distribution of age look right-skewed, but not bimodal. Clearly more needs to be done to more sensibly present the data.

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(binwidth = 11, boundary = 0,color = 'black', fill = 'yellow')
```



We chose to keep everything the same as it was originally (and leave the histogram right-closed) except the binwidth, which we increased to 11. By doing so, we found that the binwidth of 5 actually gave us a fairly accurate representation of the data. It was indeed right-skewed, but did not have the bimodality we suspected earlier. While it is possible that the data is bimodal, binwidths of 5 and 11 make it appear that this is random variation moreso than it is an overall trend in the data.

3. Glass

[18 points]

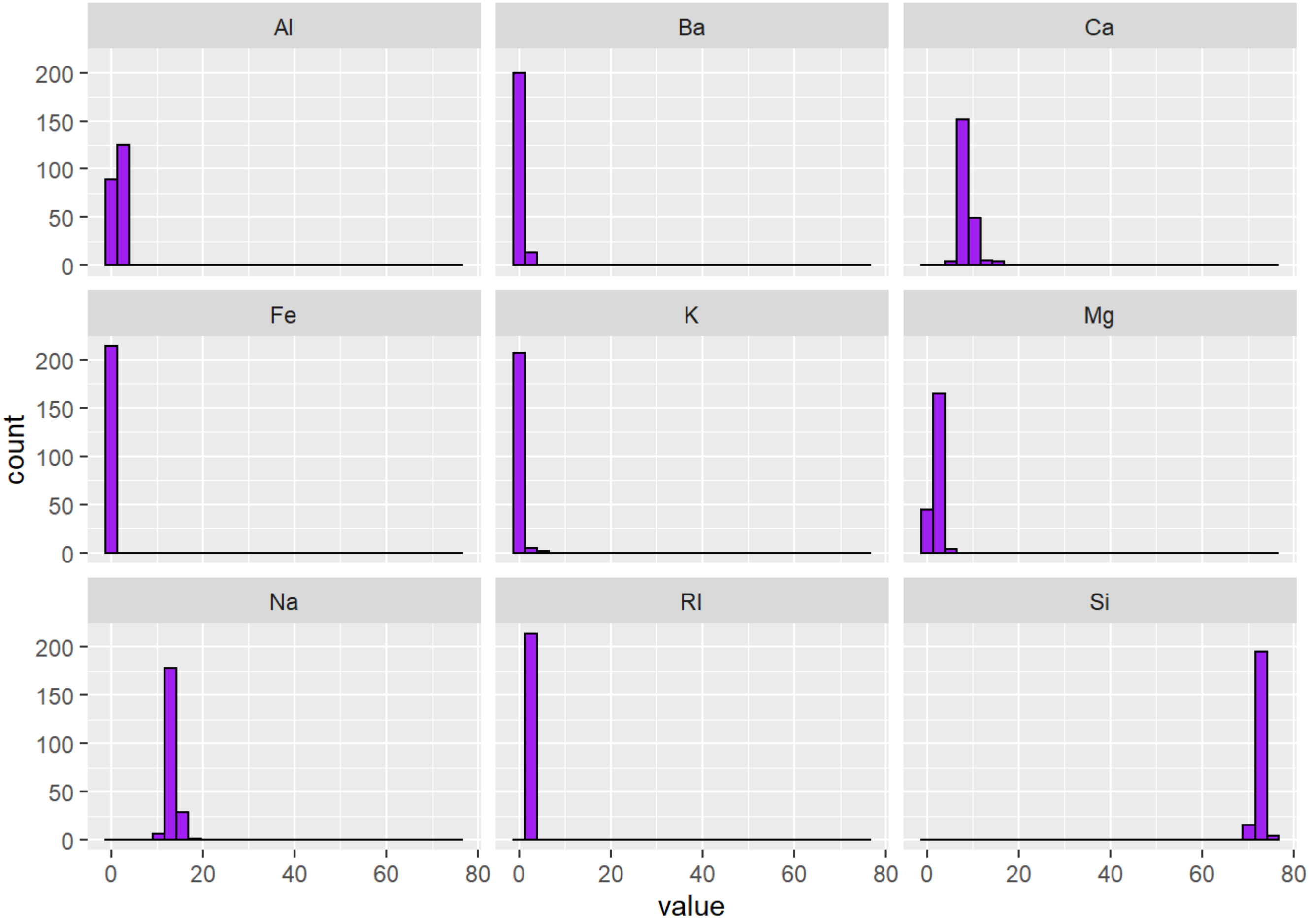
a. Use `tidyr::gather()` to convert the numeric columns in the `glass` dataset in the **ucidata** package to two columns: `variable` and `value` . The first few rows should be:

```
  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
```

```
library(tidyr)
glasstemp <- ucidata::glass
glasstemp2 <- glasstemp[,-11]
glass <- gather(glasstemp2,RI,Na,Mg,Al,Si,K,Ca,Ba,Fe,key = variable,value = value)
```

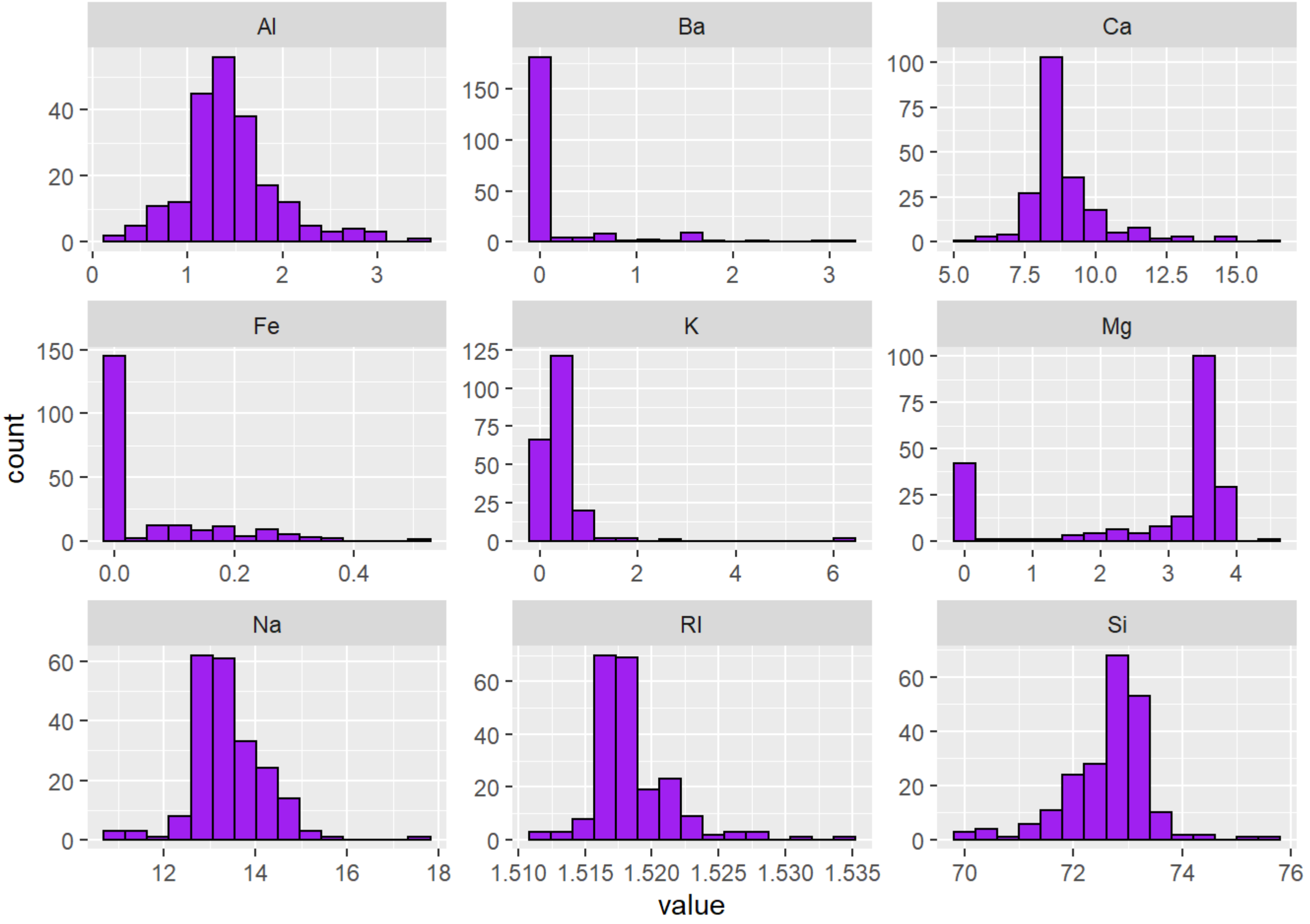
Use this form to plot histograms of all of the variables in one plot by faceting on `variable`. What patterns do you observe?

```
ggplot(glass, aes(x=value)) +
  geom_histogram(fill = 'purple',color='black') +
  facet_wrap(~variable)
```



From this, we can see that there are varying amounts of each element in each sample of glass. However, this chart is very difficult to read because of the clustering, so we are going to give each histogram its own scale.

```
ggplot(glass, aes(x=value)) +
  geom_histogram(fill = 'purple',color='black',bins=15) +
  facet_wrap(~variable,scales = 'free')
```

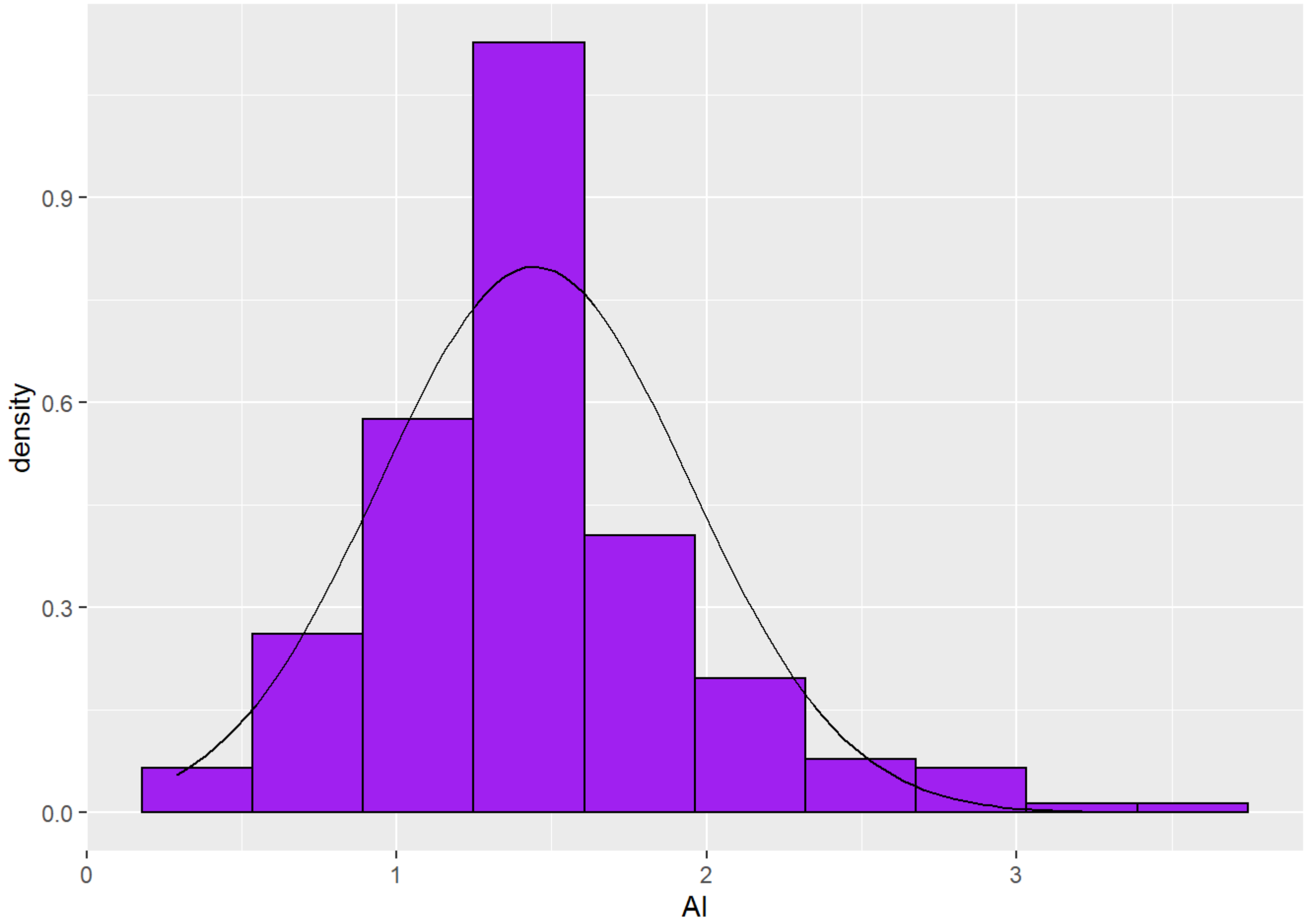


Now, we can see that the distributions vary significantly by element. Al, Ca, Na, and RI are approximately normal, but somewhat right-skewed. Same for Si, except left-skewed. Ba and Fe have either a few extreme outliers, or a number of zero values - we can't tell which from these faceted histograms. K could be left-skewed with a number of extreme outliers above the mean, but it's similarly difficult to tell. Mg appears to be left-skewed with a high number of zero values.

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(glasstemp) +
  geom_histogram(aes(x=Al,y=..density..),fill = 'purple',color='black',bins=10) +
  stat_function(fun = dnorm, args=list(mean=mean(glasstemp$Al),sd=sd(glasstemp$Al)))
```



The data is right-skewed - the bins left of the mean follow the normal distribution, while the bins right of the mean show right-skewedness. There's also a surprisingly high modal value.

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

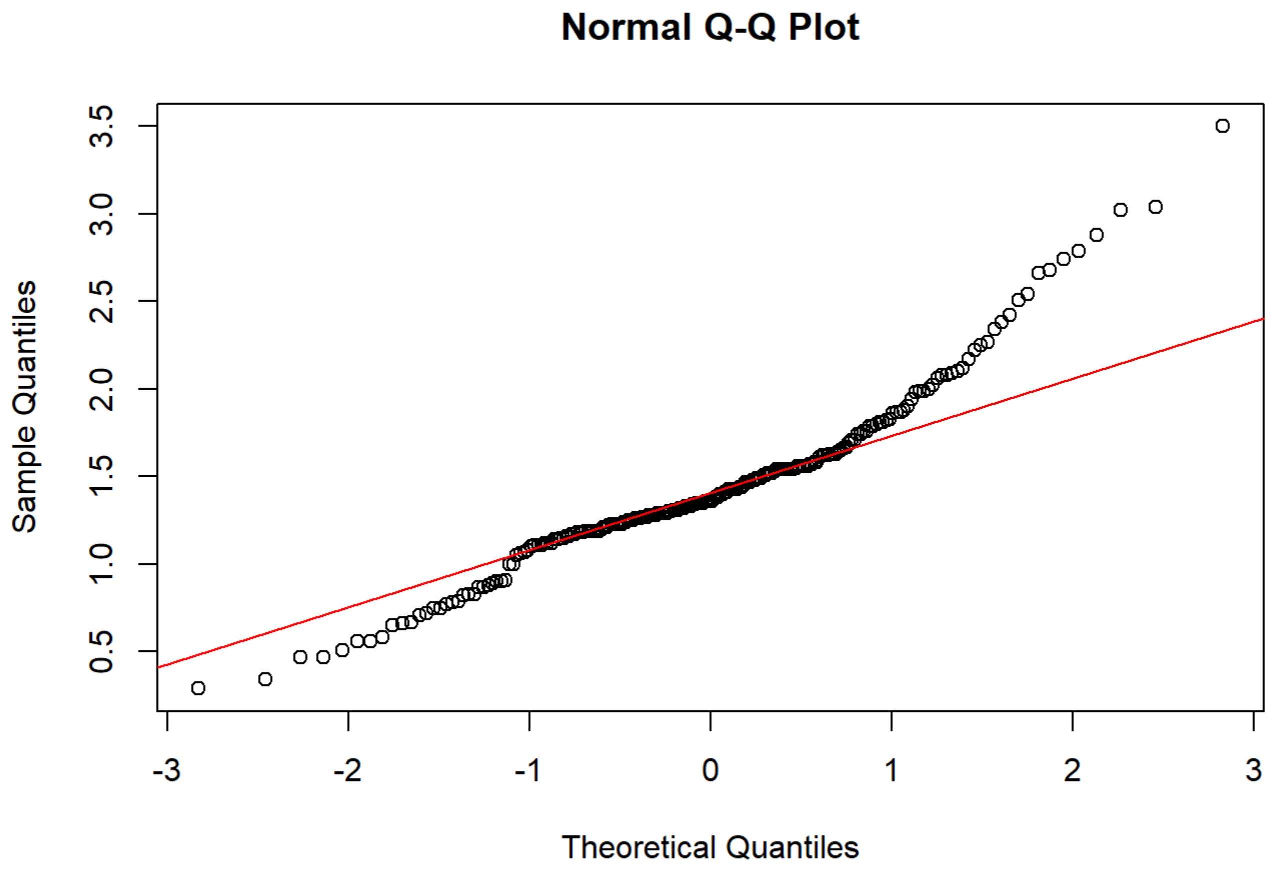
```
shapiro.test(glasstemp$Al)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  glasstemp$Al
## W = 0.94341, p-value = 2.083e-07
```

According to this p-value, we reject the null hypothesis of normality and say this distribution is not normal.

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

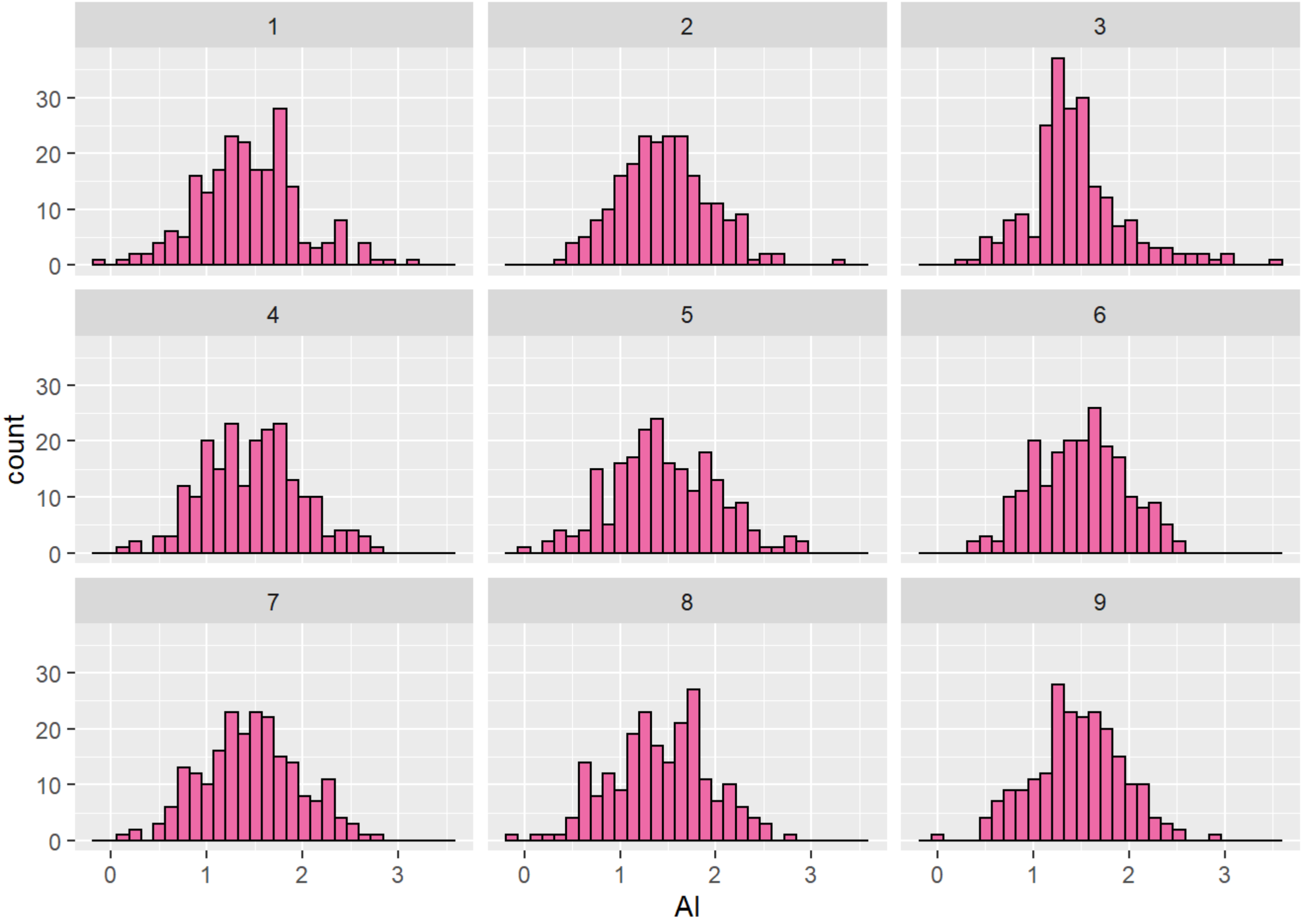
```
qqnorm(glasstemp$Al)
qqline(glasstemp$Al,col='red')
```



The QQ plot makes the distribution appear close to normal around the mean, but not normal around the outliers. It would be appropriate to call this distribution heavy-tailed if we were to exclusively judge by this plot.

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)
lp = lineup(null_dist('Al','norm'),glasstemp,n=9)
ggplot(lp, aes(Al)) +
  geom_histogram(color='black',fill='hotpink2') +
  facet_wrap(~ .sample)
```



It's fairly clear which histogram is built from the real data (we won't specify which one it is because it changes every time the code is run), although if we looked at that graph outside of a lineup it would probably look approximately normal. This lineup makes it look much less normally distributed than it would otherwise. It's very slightly more right-skewed, but the more noticeable thing is a couple strangely short bins - which, again, wouldn't be very noteworthy normally, but in this context is conspicuous compared to the eight other histograms.

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?

We showed it to a statistician, and even he couldn't pick it out on his first three tries - so maybe the distribution is more normal than we thought in Part e!

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

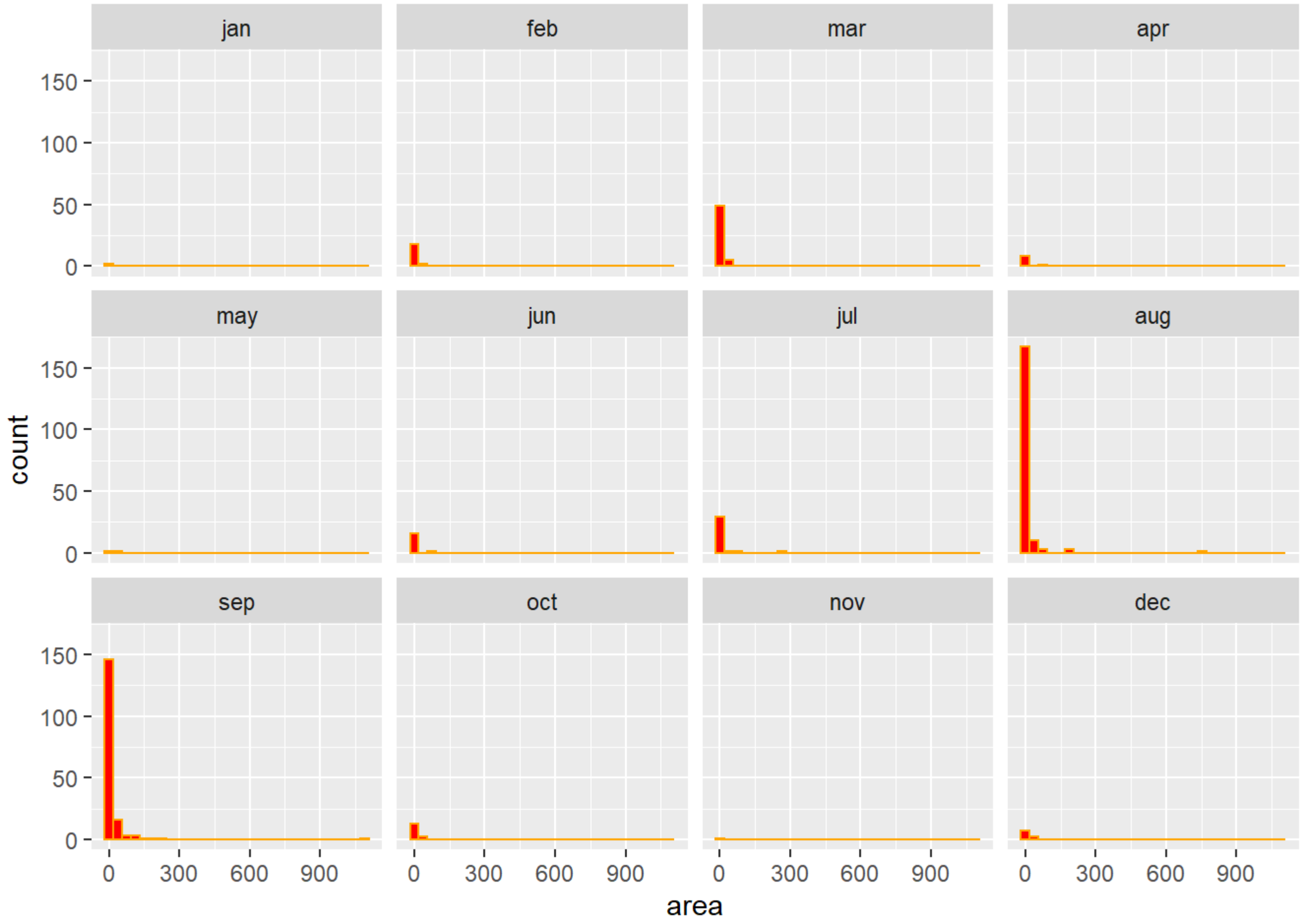
These methods all provided somewhat different answers, ranging from the distribution seeming approximately normal, to it being slightly skewed in one direction, to it having different types of skewness in both directions. It would probably be fair to say that the distribution is approximately normal, with some slight right-skewness.

4. Forest Fires

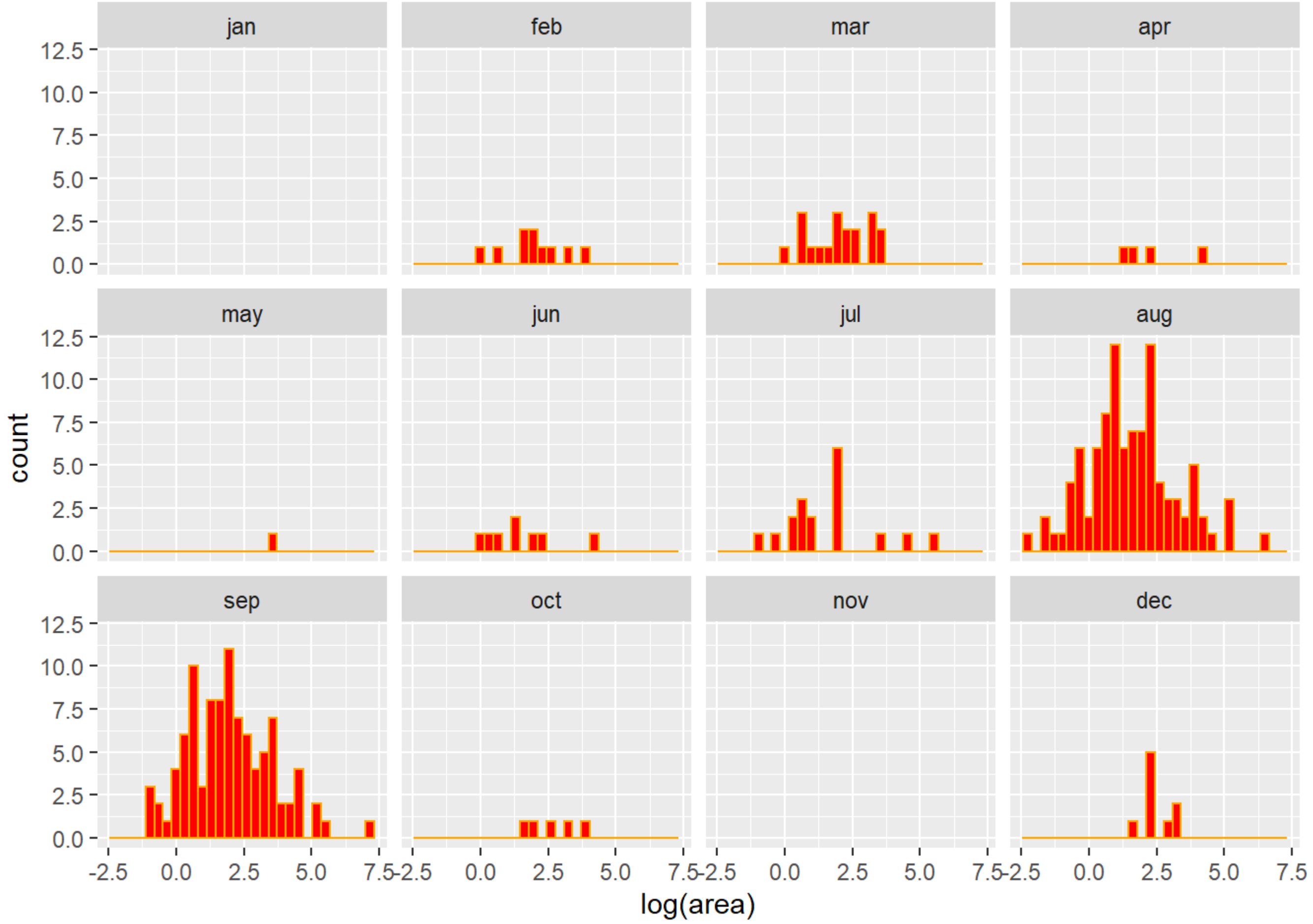
[8 points]

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.

```
fires = ucidata::forest_fires
ggplot(fires) +
  geom_histogram(aes(x=area), fill = 'red', color = 'orange') +
  facet_wrap(~factor(month,levels=c('jan','feb','mar','apr','may','jun','jul','aug','sep','oct',
'nov','dec')))
```



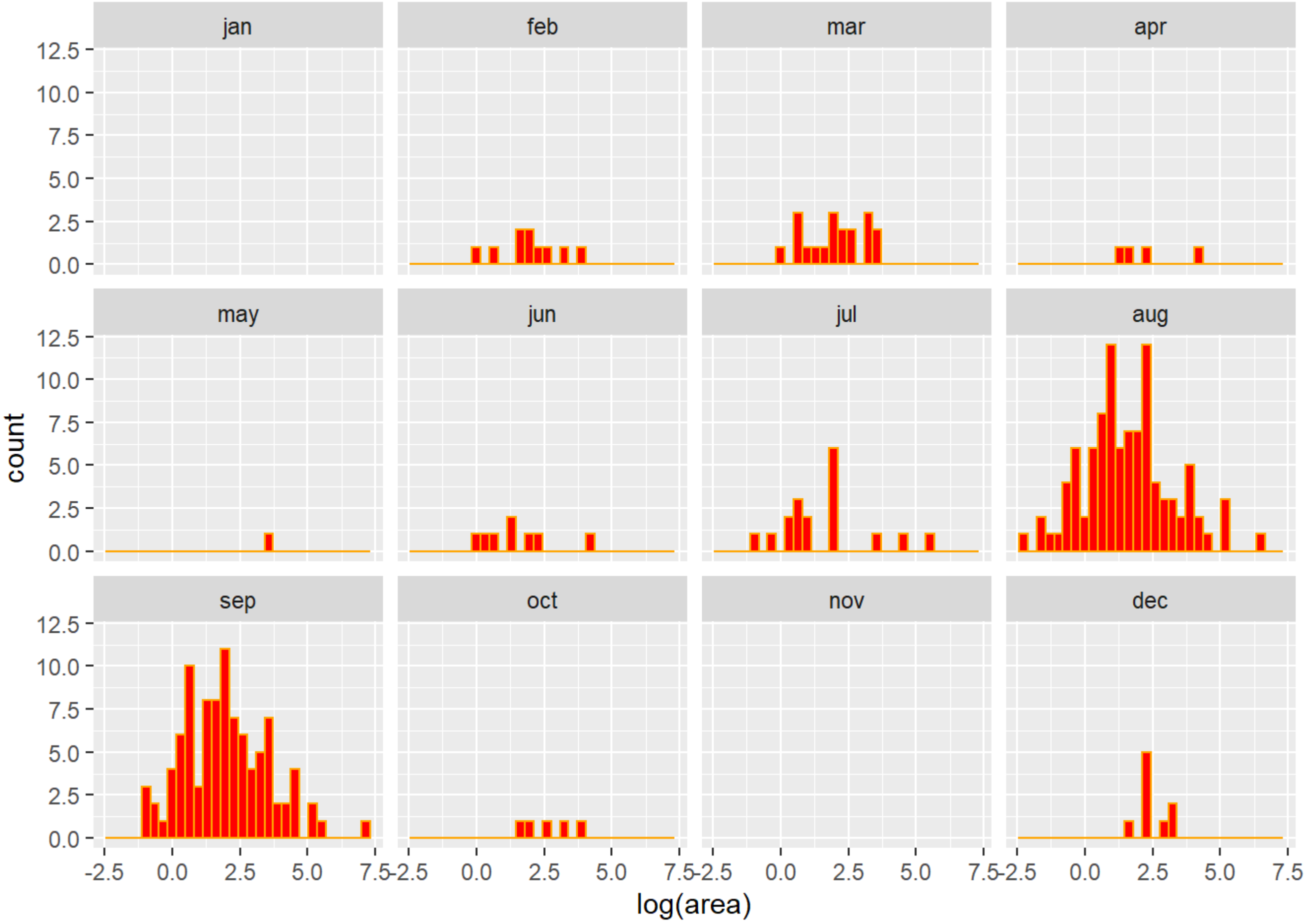
```
fires = ucidata::forest_fires
ggplot(fires) +
  geom_histogram(aes(x=log(area)), fill = 'red', color = 'orange') +
  facet_wrap(~factor(month,levels=c('jan','feb','mar','apr','may','jun','jul','aug','sep','oct',
'nov','dec')))
```



We can tell from these graphs that the fires are most concentrated in August and September. We had to apply a log transformation to the areas of the fires, because the sheer quantity of fires in August and September and the wide range of outliers within those months made the other charts unreadable. It's clear that the vast majority of large fires occur in those months.

We will now adjust each histogram to make its own distribution more easily visible by giving each histogram its own coordinate scales.

```
fires = ucidata::forest_fires
ggplot(fires) +
  geom_histogram(aes(x=log(area)), fill = 'red', color = 'orange') +
  facet_wrap(~factor(month,levels=c('jan','feb','mar','apr','may','jun','jul','aug','sep','oct',
'nov','dec')))
```



From this graph, it becomes clear that no trend can be discerned for the other ten months outside of August and September. There aren't enough fires to get a real sense of the distribution, and the fires that do occur are all relatively small compared to the large fires in August and September, giving us even less of a sense of the distributions of the areas in those months.

As for August and September, the distribution of fire areas is extremely right-skewed without a log transformation, and slightly right-skewed even with our log(10) transformation.

All this makes it clear that small fires are quite common for most of the year, but there are occasionally huge fires many, many times larger than these small fires, and those fires tend to come in August and September.