EDAV Fall 2019 Probem Set 1

Hima Bindu Bhardwaj(hb2635), Huayun Xu(hx2283)

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")
```

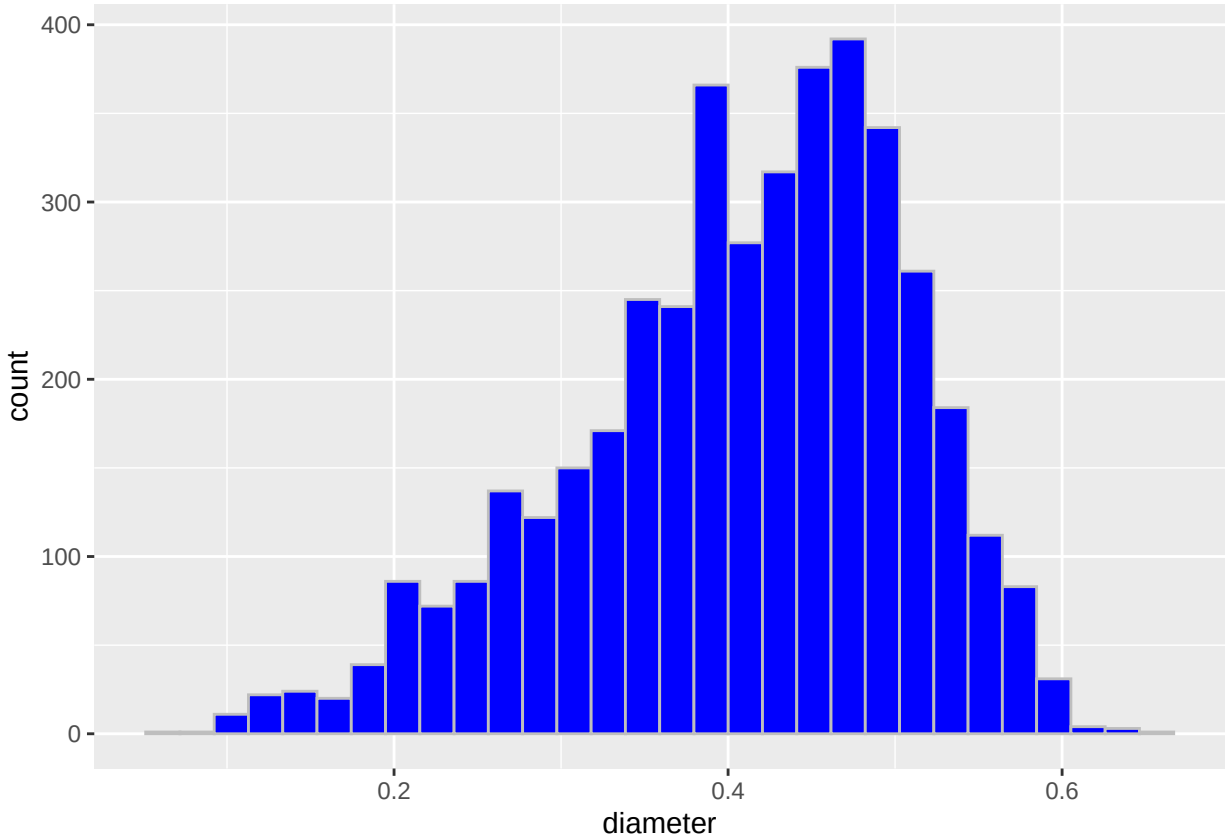
1. Abalone

[18 points]

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

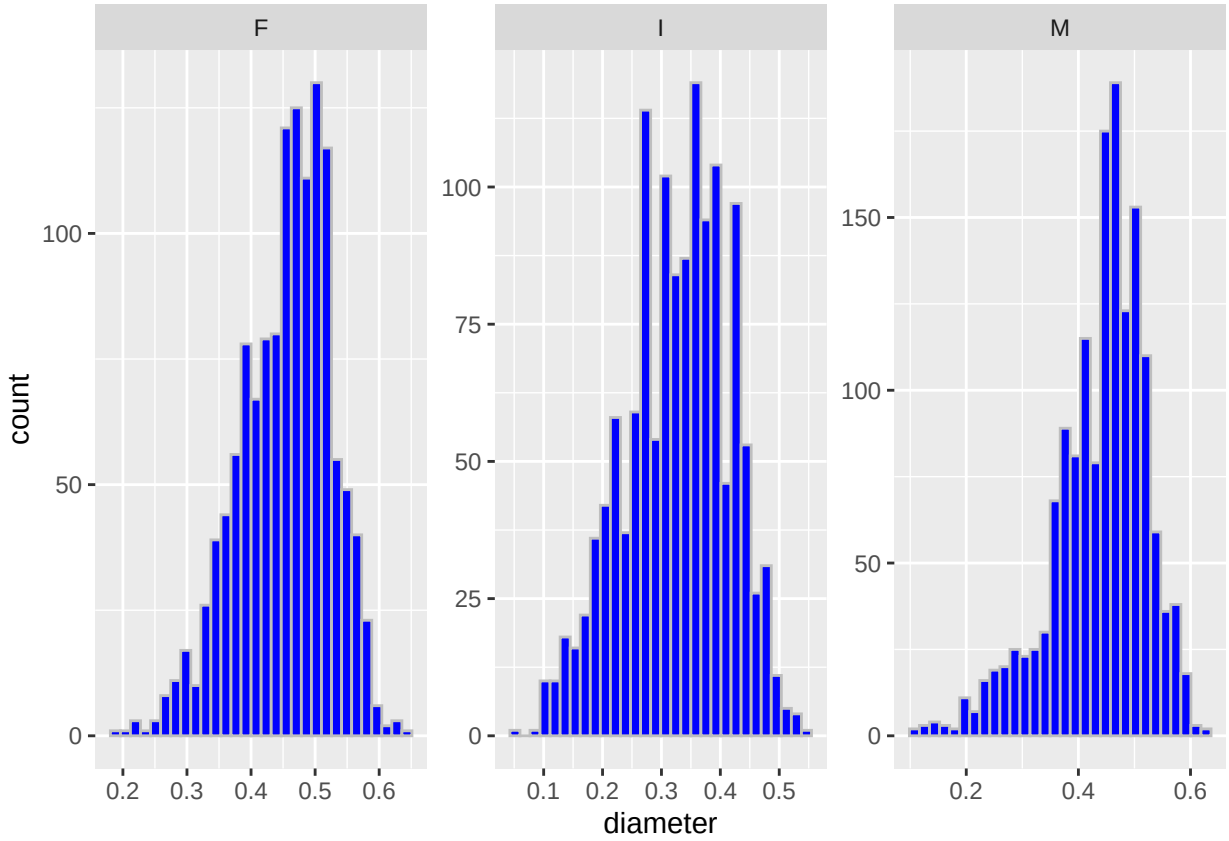
a) Plot a histogram of the variable.

```
library(ucidata)
library(ggplot2)
ggplot(abalone, aes(diameter)) + geom_histogram(colour = "grey", fill = 'blue')
```



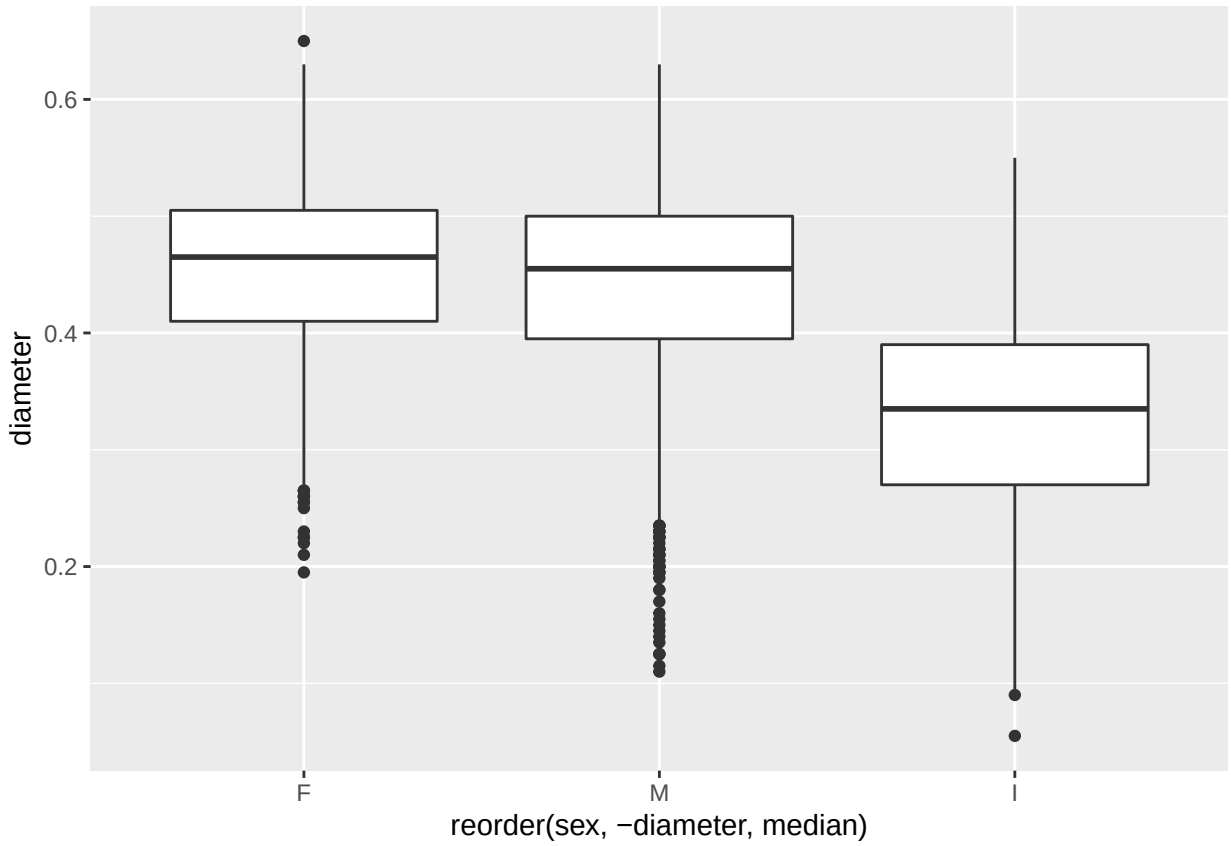
b) Plot histograms, faceted by `sex`, for the same variable.

```
library(ucidata)
library(ggplot2)
ggplot(abalone, aes(diameter)) +
  geom_histogram(colour = "grey", fill = 'blue') +
  facet_wrap(~sex,scales="free")
```



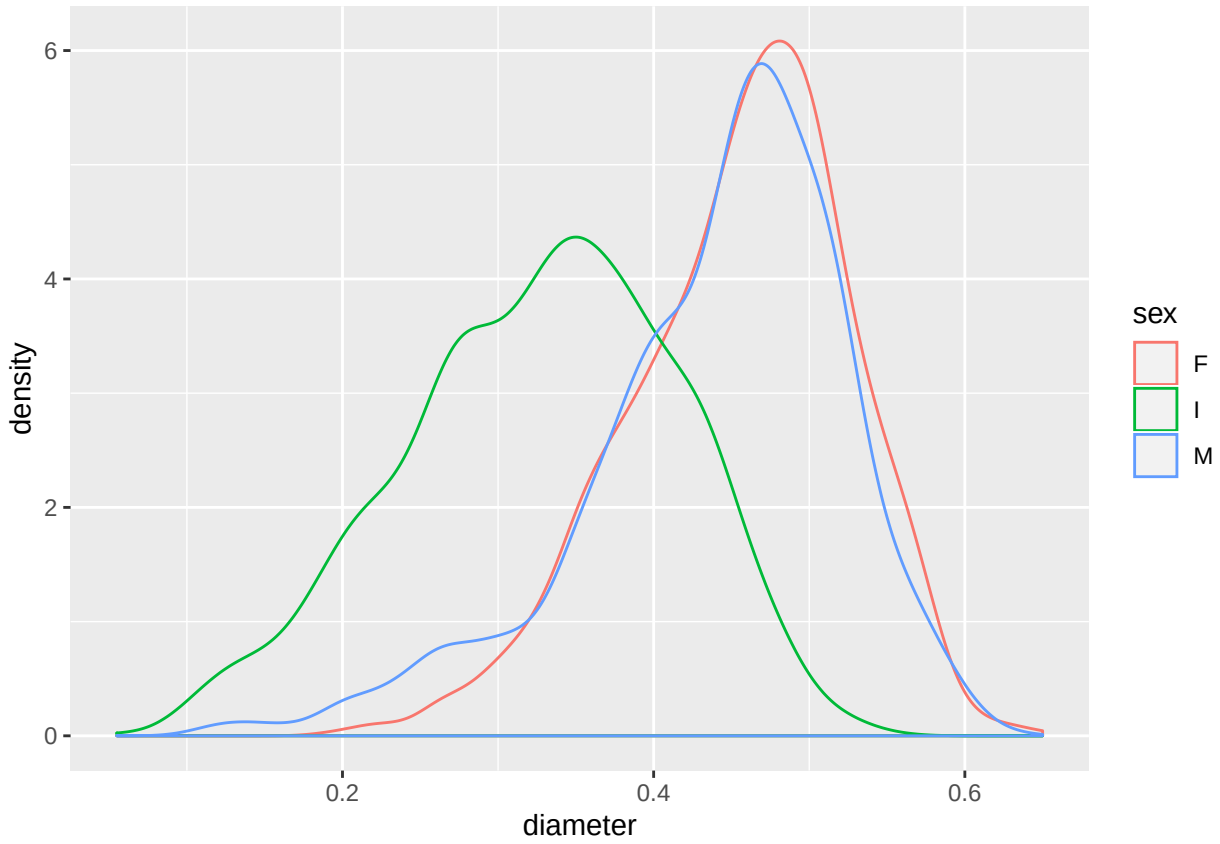
c) Plot multiple boxplots, grouped by `sex` for the same variable. The boxplots should be ordered by decreasing median from left to right.

```
library(ggplot2)
ggplot(abalone, aes(x = reorder(sex, -diameter, median), y = diameter)) +
  geom_boxplot()
```



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.

```
library(ucidata)
library(ggplot2)
ggplot(abalone, aes(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?

From the histograms faceted by sex, we can infer the following: i) Infants have a larger population of values at smaller values of diameters (eg: values$<$ 0.2) when compared to females and males. ii) Males have the largest population of abalones with a wide diameter, followed by females and infants have a small population of abalones with comparatively high values of diameter. i) and ii) can be attributed to the fact that infants are not yet grown and developed when compared to the adult female and male population of abalones.

From the multiple boxplots the following can be inferred: i) Females have a median value which is the highest among all three classes of sex. ii) Female and male boxplots are narrower than that of infant. This goes to show that the diameter values of grown females and males are concentrated in a narrow range of values whereas the infant boxplot seems to be longer since abalone's at different stages of infancy will have different diameter values that keeps growing as they grow into adulthood.

From the density curves, it follows that the mode of females is highest than compare to makes and infants have the smallest value of mode in comparison. The density plots of females and males is skewed left whereas that of infants seems to be more symmetric in comparison.

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

It doesn't seem right, since the value in the dataset is too small, the unit of measurement of diameter shouldn't be 'mm'. The actual size of an abalone ranges from 20mm to 200mm but the values present in the dataset are not of that order.

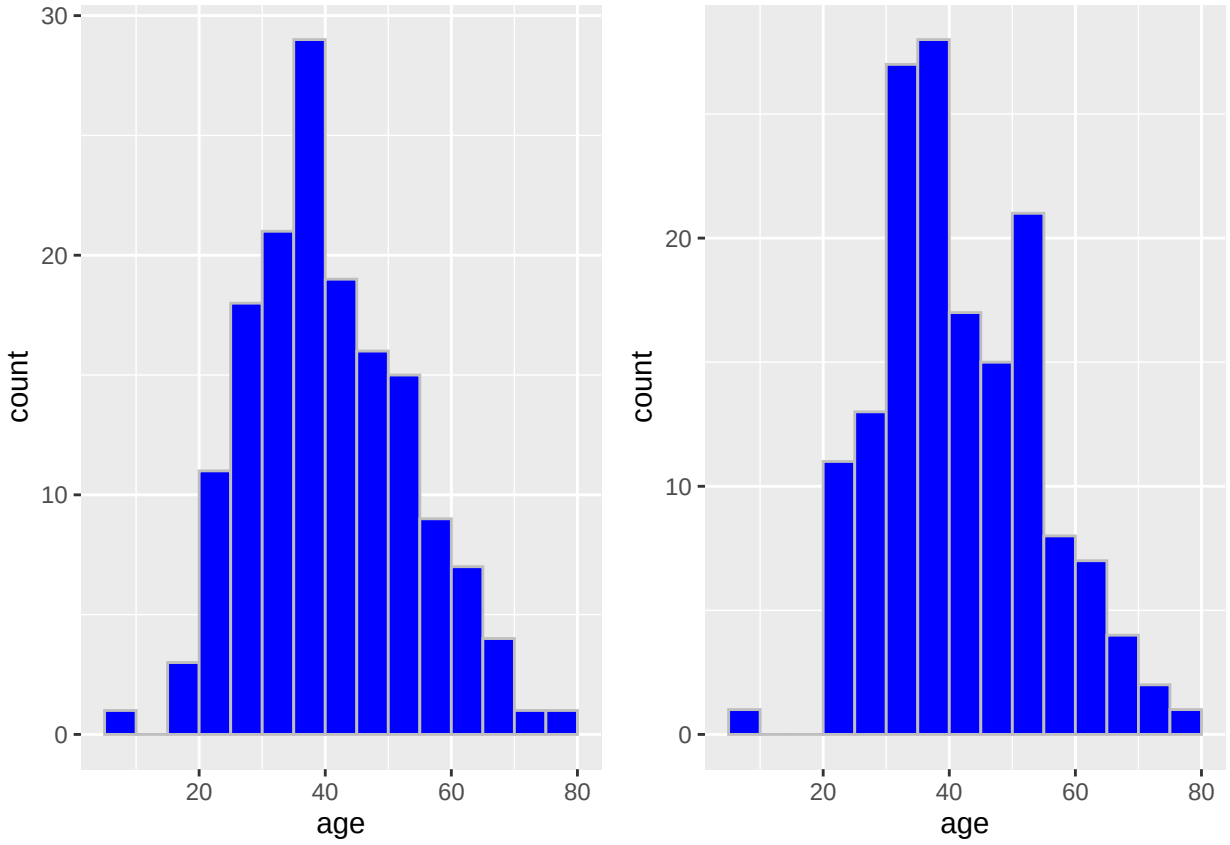
2. Hepatitis

[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?

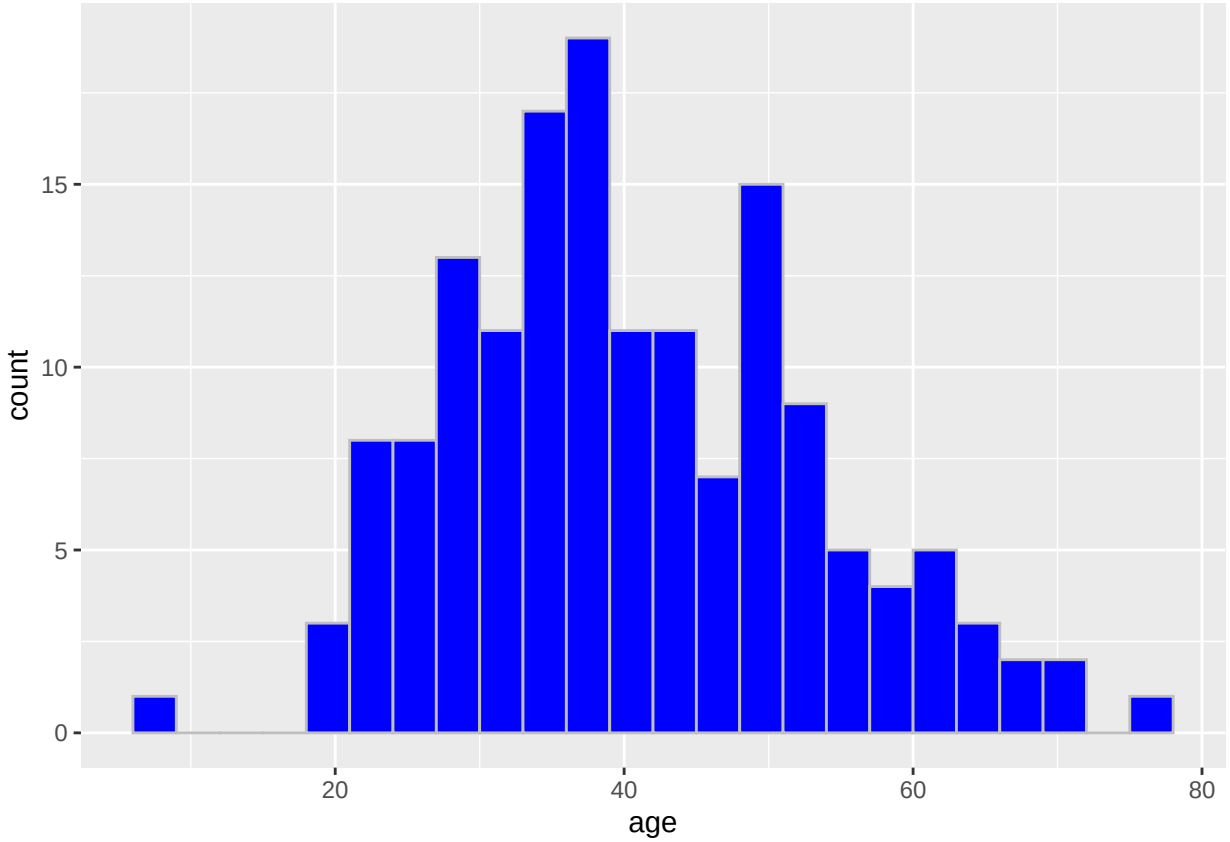
```
library(ucidata)
library(ggplot2)
library(gridExtra)
p1<-ggplot(hepatitis, aes(age)) +
  geom_histogram(binwidth = 5, boundary = 0, colour = "grey", fill = 'blue')
p2<-ggplot(hepatitis, aes(age)) +
  geom_histogram(binwidth = 5, boundary = 0, right = FALSE, colour = "grey", fill = 'blue')
grid.arrange(p1, p2, ncol = 2)
```



The two histograms are not identical. The right closed graph looks approximately normal whereas in the left closed graph, it seems like some elements have been pushed to the right side bins (since it is right open, the values which would've joined a bin will now join the one on it's right)

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

```
library(ucidata)
library(ggplot2)
library(gridExtra)
ggplot(hepatitis, aes(age)) +
  geom_histogram(binwidth = 3, boundary = 0, right = TRUE, colour = "grey", fill = 'blue')
```



When the binwidth is set to 3, the detail or insights visualized don't seem too detailed or too coarse/broad.

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:

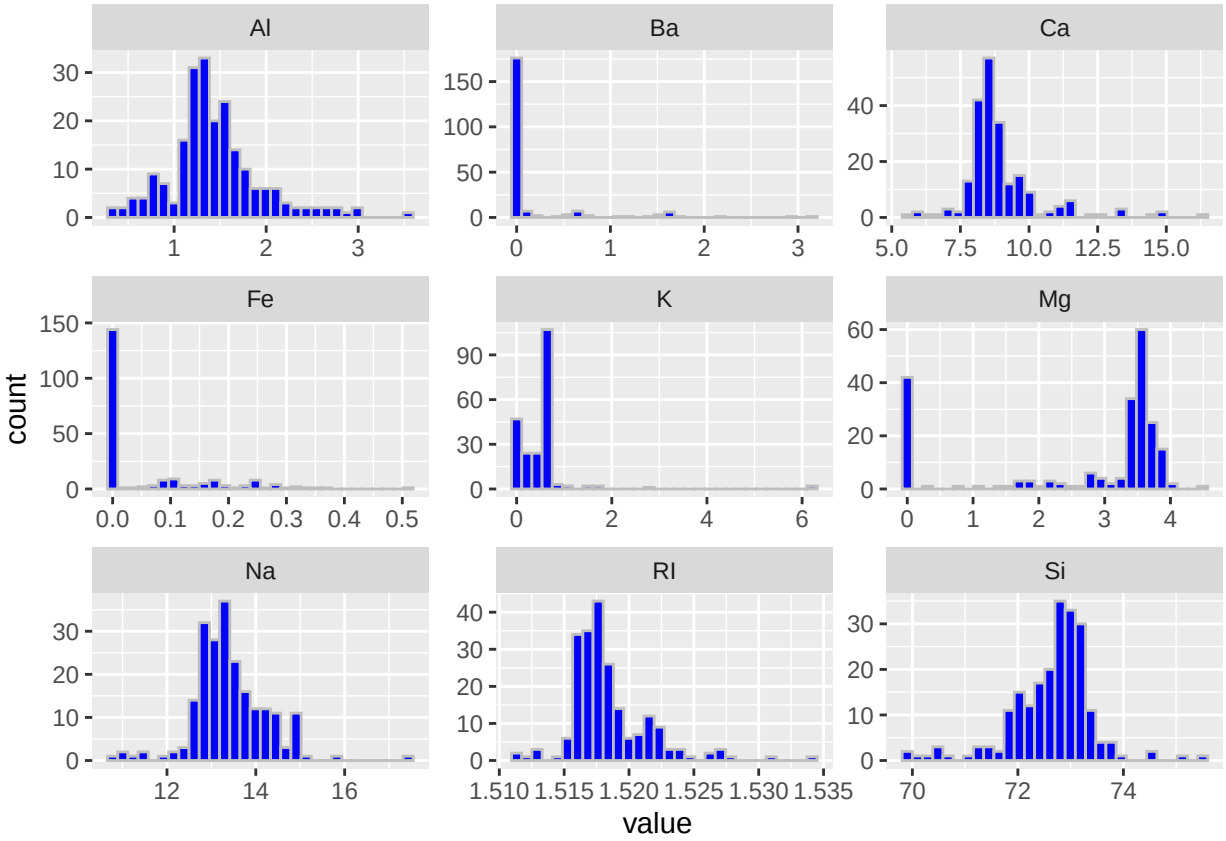
```
  variable   value
1       RI 1.52101
2       RI 1.51761
3       RI 1.51618
4       RI 1.51766
5       RI 1.51742
6       RI 1.51596
```

```
library(ucidata)
library(ggplot2)
a<-tidyr::gather(ucidata::glass[,2:10], key = "variable", value = "value")
head(a)
```

```
##   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?

```
ggplot(a, aes(value), color = variable) +
  geom_histogram(colour = "grey", fill = 'blue') +
  facet_wrap(~variable,scales="free")
```

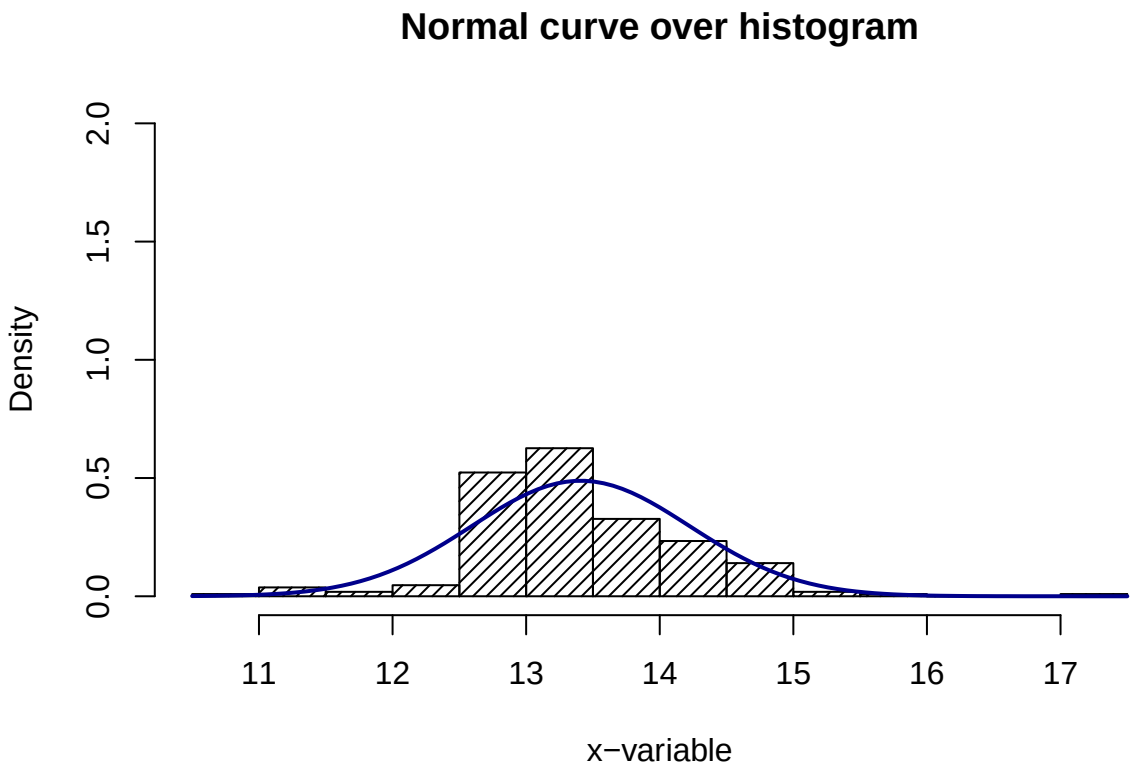


Some elements have values ranging from very high to very low, such as Ba, Fe, K and Mg. Al, Ca, Na and RI seem to have graphs which are skewed right whereas Si has a left skewed graph. And Al, Ca, Na and RI seem to share some characteristic of the normal curve, such as unimodal.

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?

```
g = glass$Na
m<-mean(g)
std<-sqrt(var(g))
hist(g, density=20, breaks=20, prob=TRUE,
     xlab="x-variable", ylim=c(0, 2),
     main="Normal curve over histogram")
curve(dnorm(x, mean=m, sd=std),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```



The normal curve with the mean and standard deviation of the data seems to be tracing an approximately normal curve over the histogram bars.

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

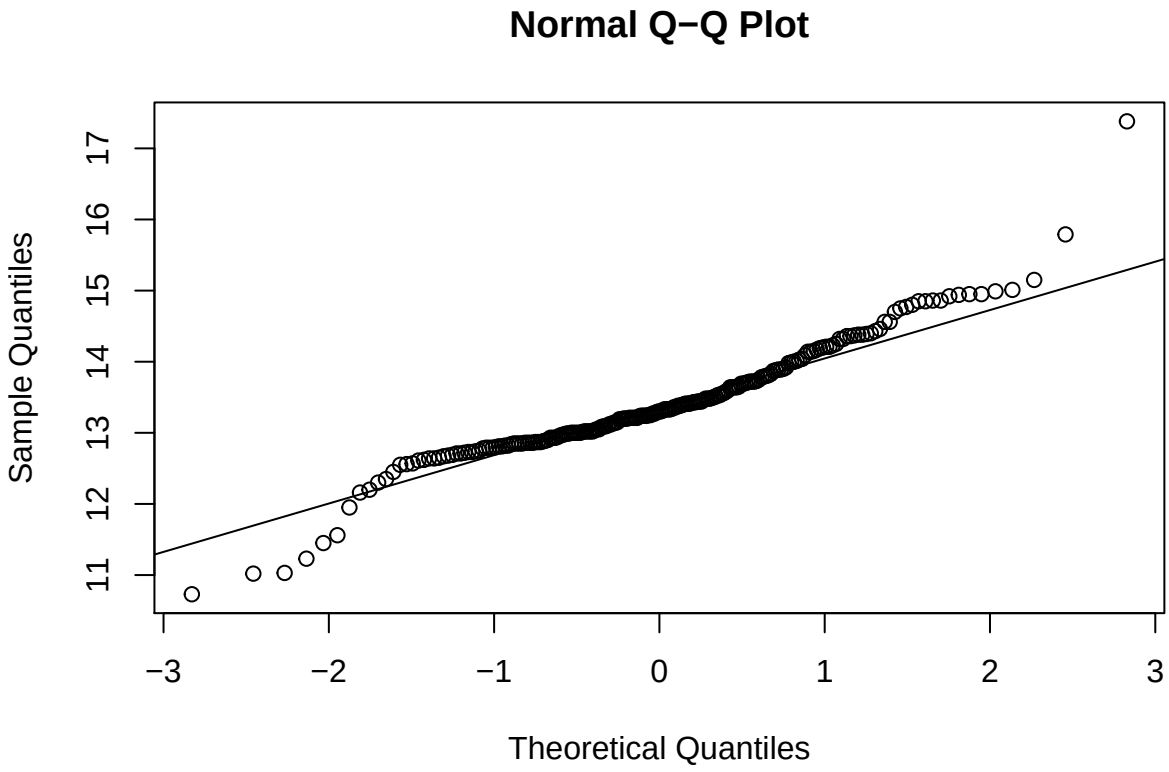
```
shapiro.test(glass$Na)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  glass$Na
## W = 0.94576, p-value = 3.466e-07
```

The p-value is 3.466e-07 which is less than 0.05. Since the value is less than 0.5, the test states that the distribution is not normal.

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

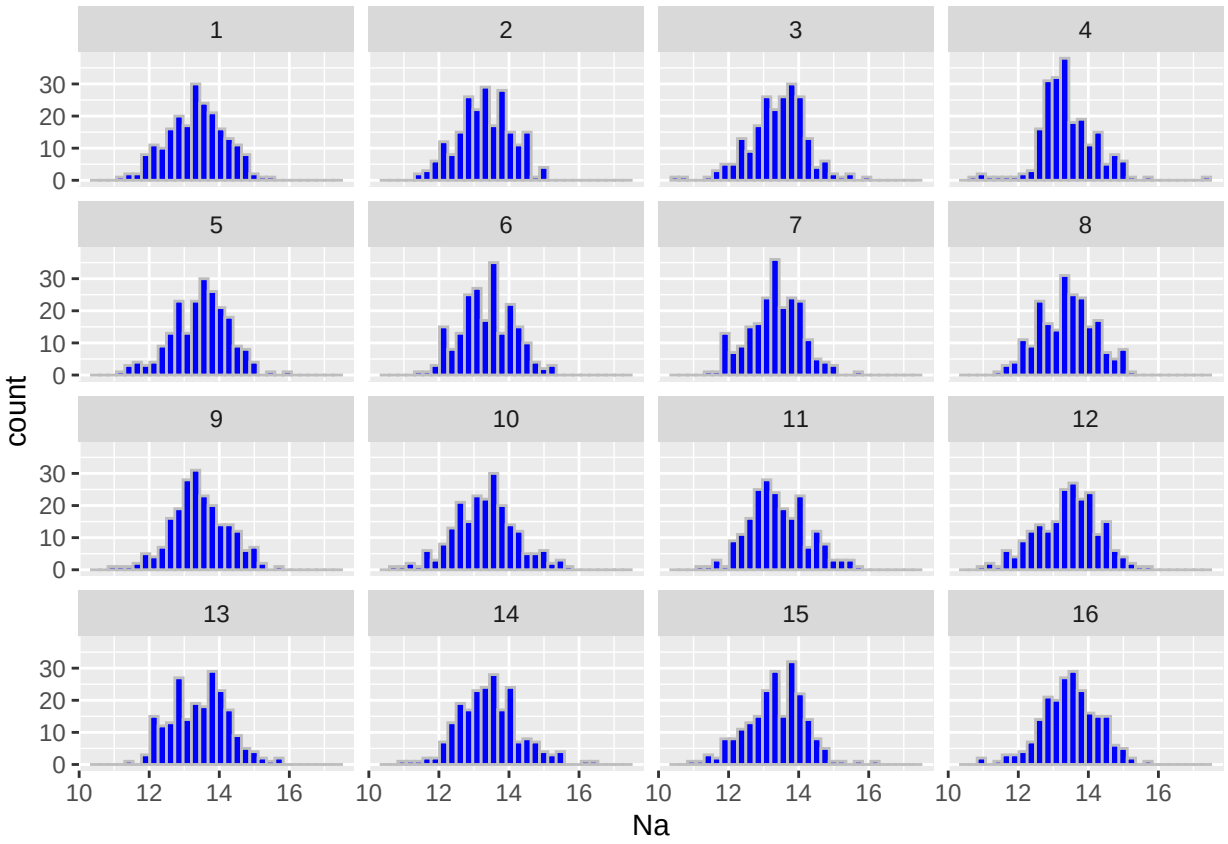
```
qqnorm(glass$Na)
qqline(glass$Na)
```



Yes, it appears to be approximately normal since the datapoints are close to the diagonal line.

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(ucidata)
library(ggplot2)
library(nullabor)
ggplot(lineup(method = null_dist("Na","norm"), true = glass, n = 16), aes(Na))+
  geom_histogram(colour = "grey", fill = 'blue') +
  facet_wrap(~.sample)
```



We couldn't pick out the real data without counter checking with the actual histogram of the real data.

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?

No, our friend chose the fake data.

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

All of the methods did not produce the same result. The Q-Q plot seemed to show that the data was normal since the datapoints were very close to the diagonal. Even the normal curve drawn over the histogram seemed to show that the data was normal. But the Shapiro wilk test failed and showed that the data was not normal.

4. Forest Fires

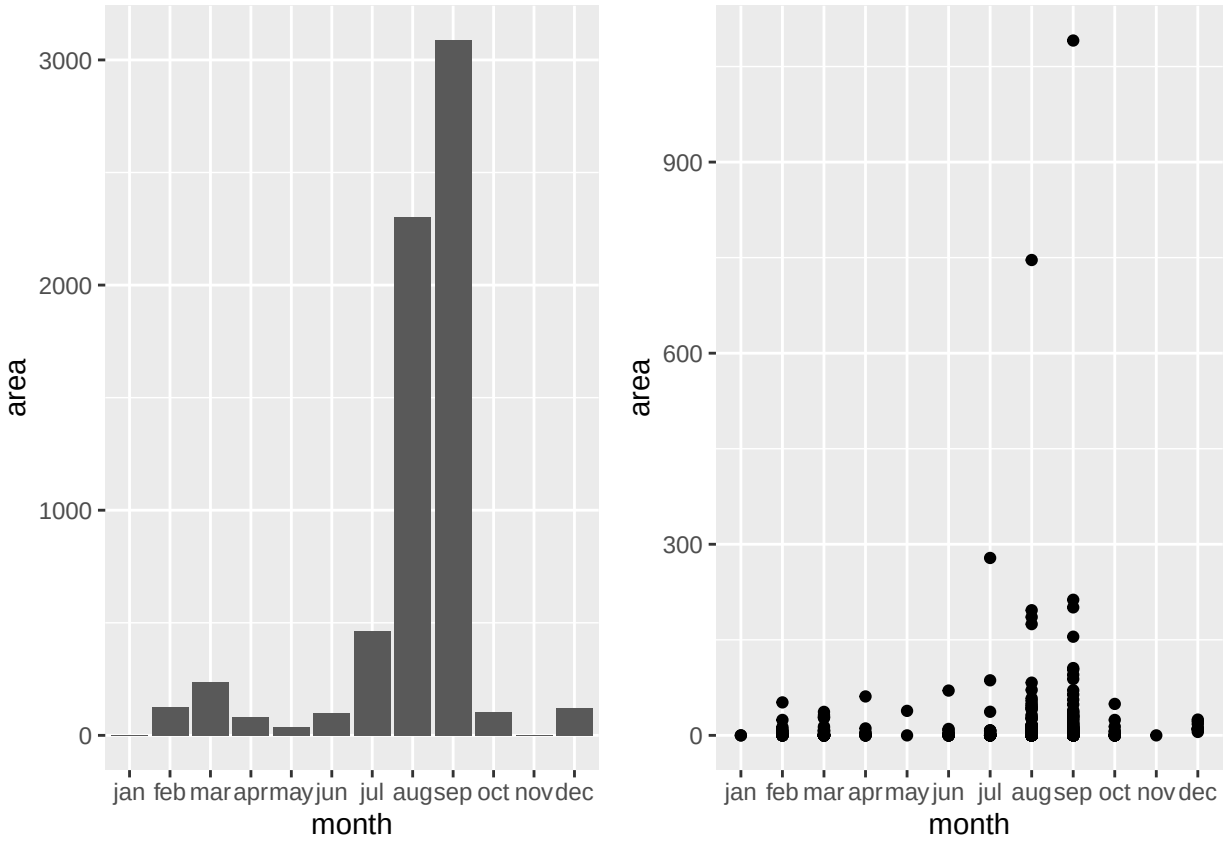
[8 points]

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.

```
library(ucidata)
library(ggplot2)
data = forest_fires
forest_fires$month <-factor(forest_fires$month,
                        levels = c("jan", "feb", "mar", "apr", "may", "jun",
                                   "jul", "aug", "sep", "oct", "nov", "dec"))

p1 <- ggplot(forest_fires, aes( x = month, y = area)) + geom_col()
p2 <- ggplot(forest_fires, aes(x = month, y = area)) +
geom_point()
```

```
grid.arrange(p1, p2, ncol = 2)
```



For the visualization of the burned area of forest fire incidents, we chose two visualizations both of which convey different insights. The column plot show the total area burnt by month and hence helped us to compare the areas burnt in various months. The scatterplot on the right shows the extent of damage or the severity of the forest fire by using the points. Scatterplot shows individual events and hence we can see that many incidents were similar to each other (clustered near the x axis between 0 and 300) in terms of magnitude of fires and there were some incidents that were severe (3 individual dots not near to x axis)

We can also infer from the graph that the number of forest fires were much more in summer months (june-AugSep) and very few in winter months (Nov - Feb).