Data Visualization: Homework 2

Matt Isaac - A01515095

October 16, 2017

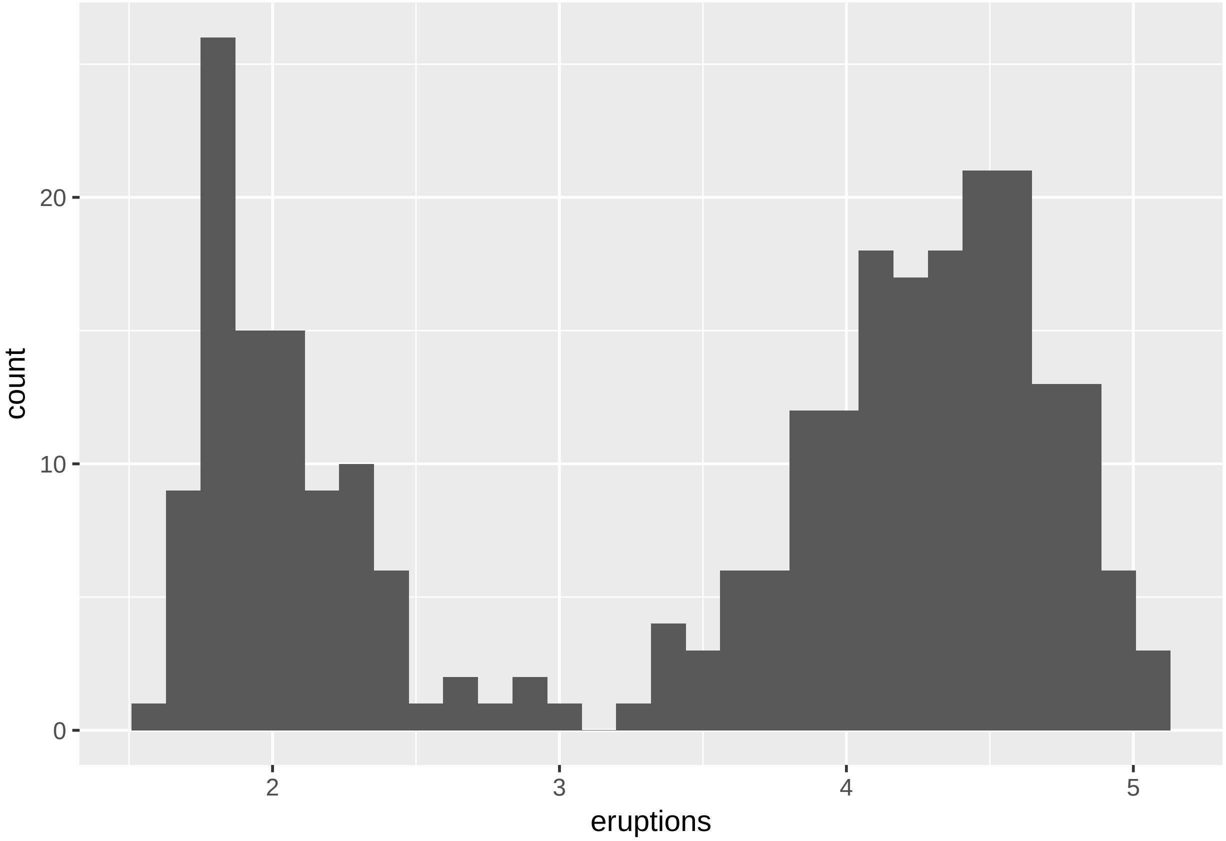
1.

a. Load all required R packages to answer this question.

```
library(ggplot2)
library(MASS)
```

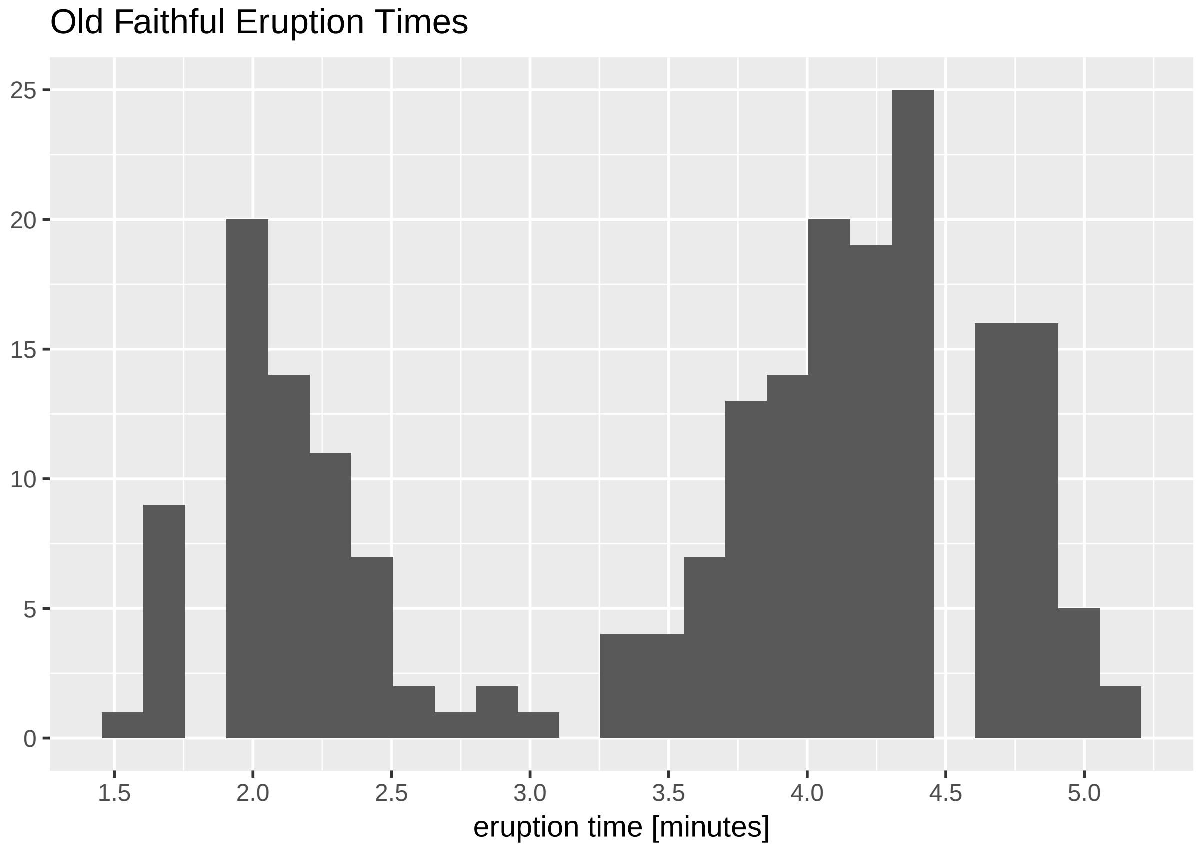
b. Draw a basic histogram for *eruptions* using ggplot2.

```
ggplot(faithful, aes(eruptions)) +
  geom_histogram()
```



c. Further improve your histogram from (b).

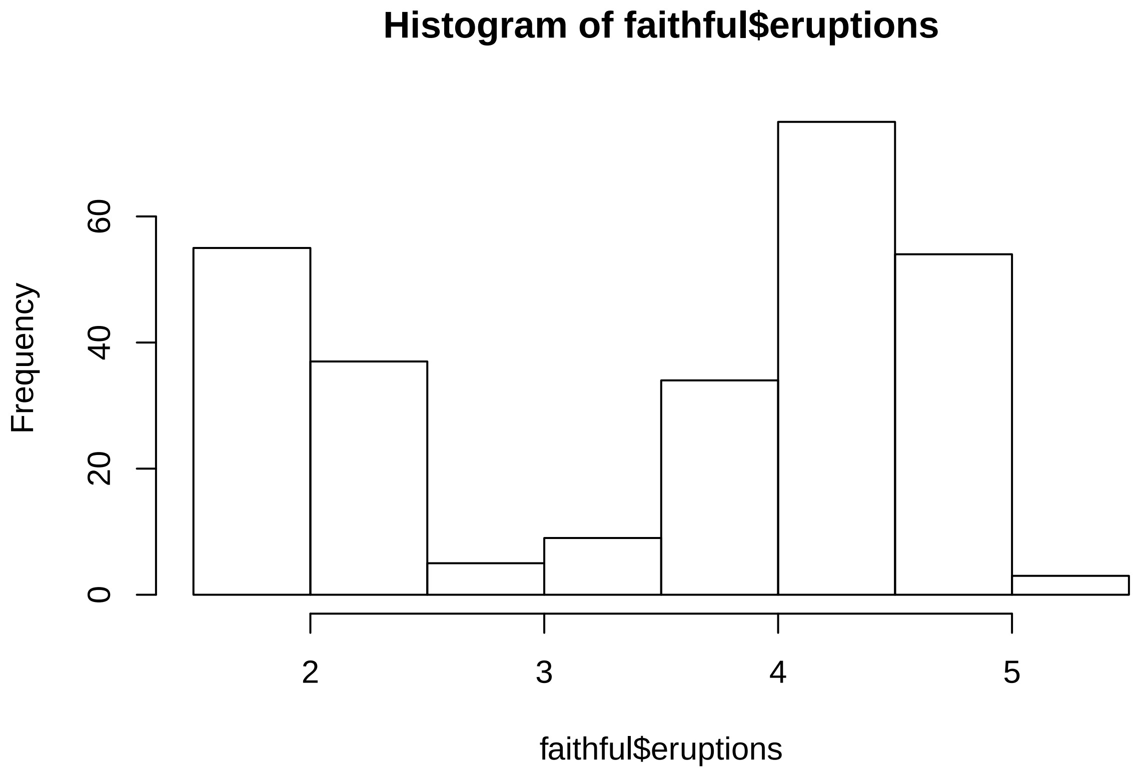
```
ggplot(faithful, aes(eruptions)) +
  xlab("eruption time [minutes]") +
  ylab("") +
  xlim(1, 5) +
  ylim(0, 25) +
  scale_x_continuous(breaks = seq(1, 6, .5))+
  geom_histogram(binwidth = 0.15, center = .03) +
  ggtitle("Old Faithful Eruption Times")
```



I first clarified the x-axis label, changing it from 'eruptions' to 'Eruption Time [minutes]'. This makes it clear that the histogram is displaying the frequency of geyser eruptions of various times. I also felt like it was important to include the time unit. Someone unfamiliar with the Old Faithful geyser (or with geysers in general) might mistakenly assume a different unit of time (seconds or hours). I also removed the y-axis label, as frequency is assumed with histograms unless otherwise indicated. I changed the x-axis limits to include 1 to 5, and the y-axis limits to include 0 to 25 in order to have a little room at each end of the graph. The scale on the x-axis took a lot of experimenting, but I decided to use 0.5 as steps on the scale. After more experimenting, I chose a binwidth of 0.15. It was this binwidth that made my x-axis scale need to be a bit unconventional. However, a binwidth of 0.10 was too small, and a binwidth of 0.20 was too big. I recognize that my bars in the histogram don't line up very well with the gridlines, but I felt that it was more important to have an intuitive scale than to have it line up perfectly. Lastly, I added a title to clarify what the eruption times data were from.

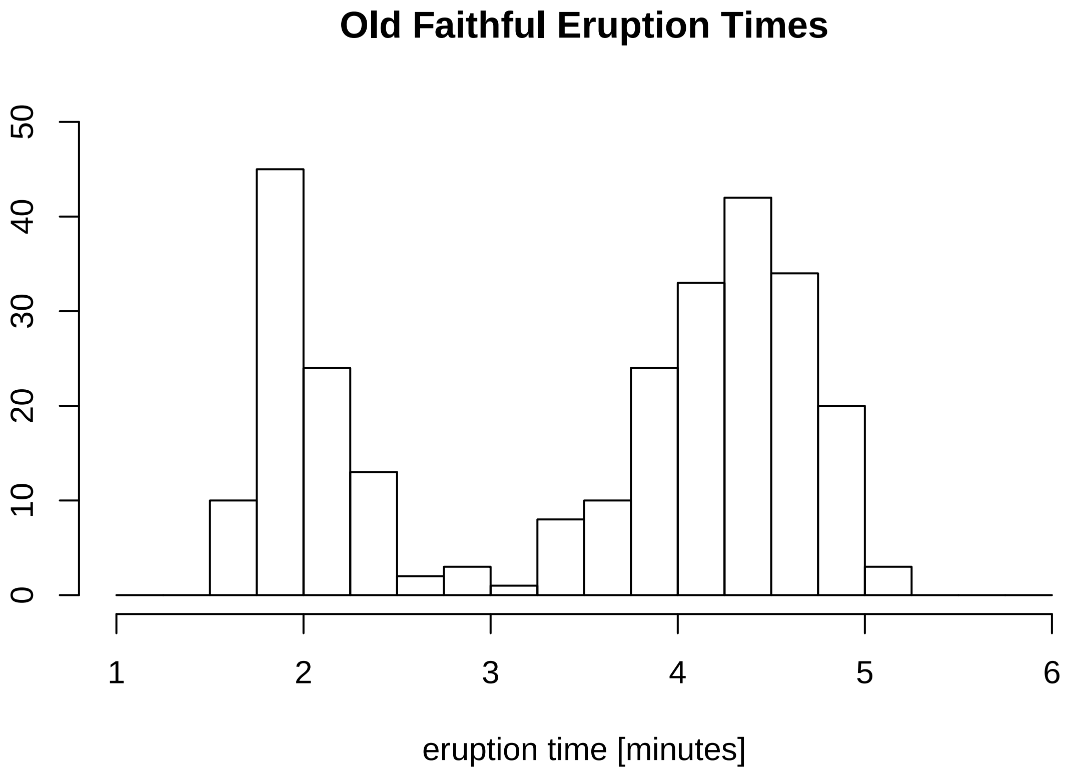
d. Repeat (b) from above, using the `hist()` function from baseR.

```r
hist(faithful$eruptions)
```



e. Repeat (c) from above, using the `hist()` function from baseR.

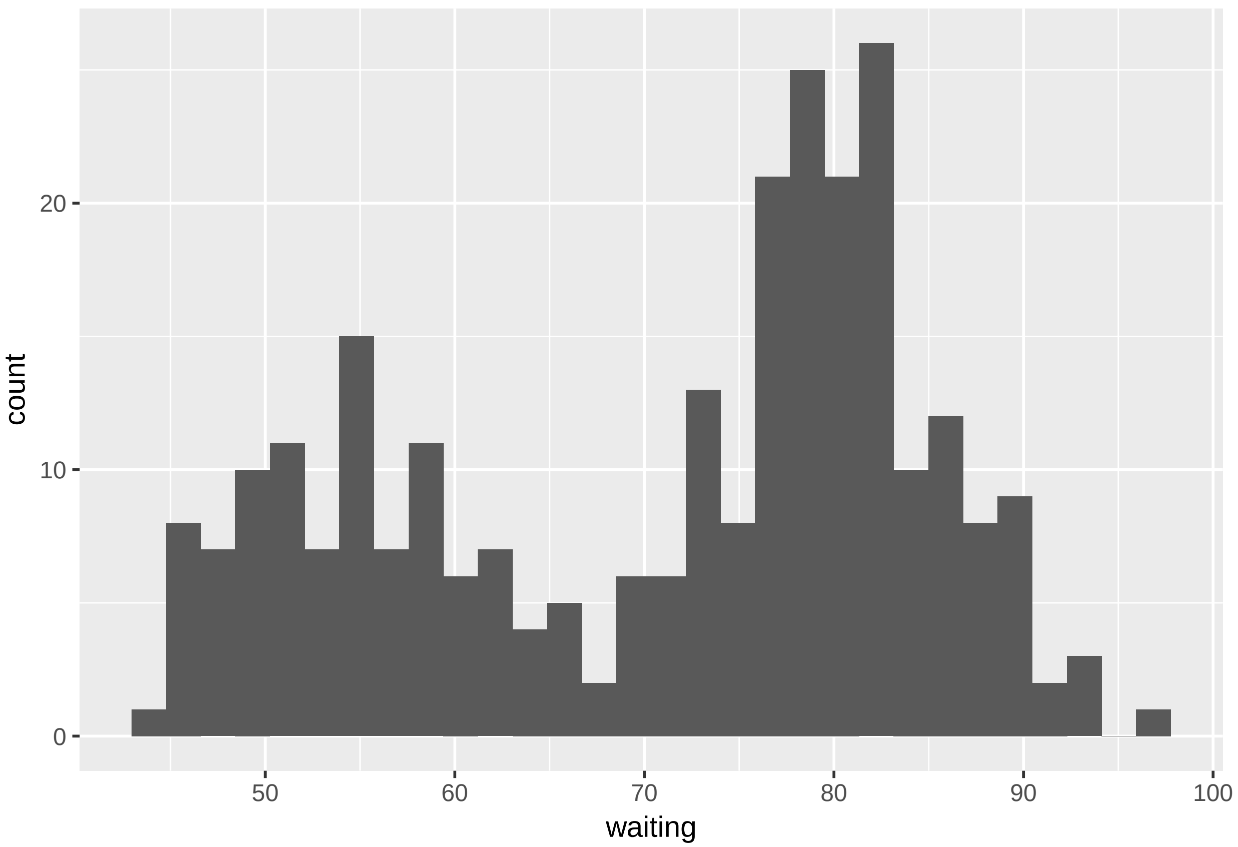
```r
hist(faithful$eruptions,
     main = "Old Faithful Eruption Times",
     xlab = "eruption time [minutes]",
     ylab = "", xlim = c(1,6),
     breaks=seq(1, 6, by=.25),
     ylim = range(0,50))
```



I was surprised by how the different package led me to create this graph a little differently. Even though the binwidths are different in the histogram in baseR than in the plot created with ggplot, I feel that in each case, a valid and informative histogram was produced, and that each will lead to the same conclusions. In this histogram, I made changes to the x and y axis labels and for the x and y axis limits for the same reasons as in part (c). For the binwidths, I felt that a width of 0.25 gave us a very good understanding of the data without having excessive definition and detail.

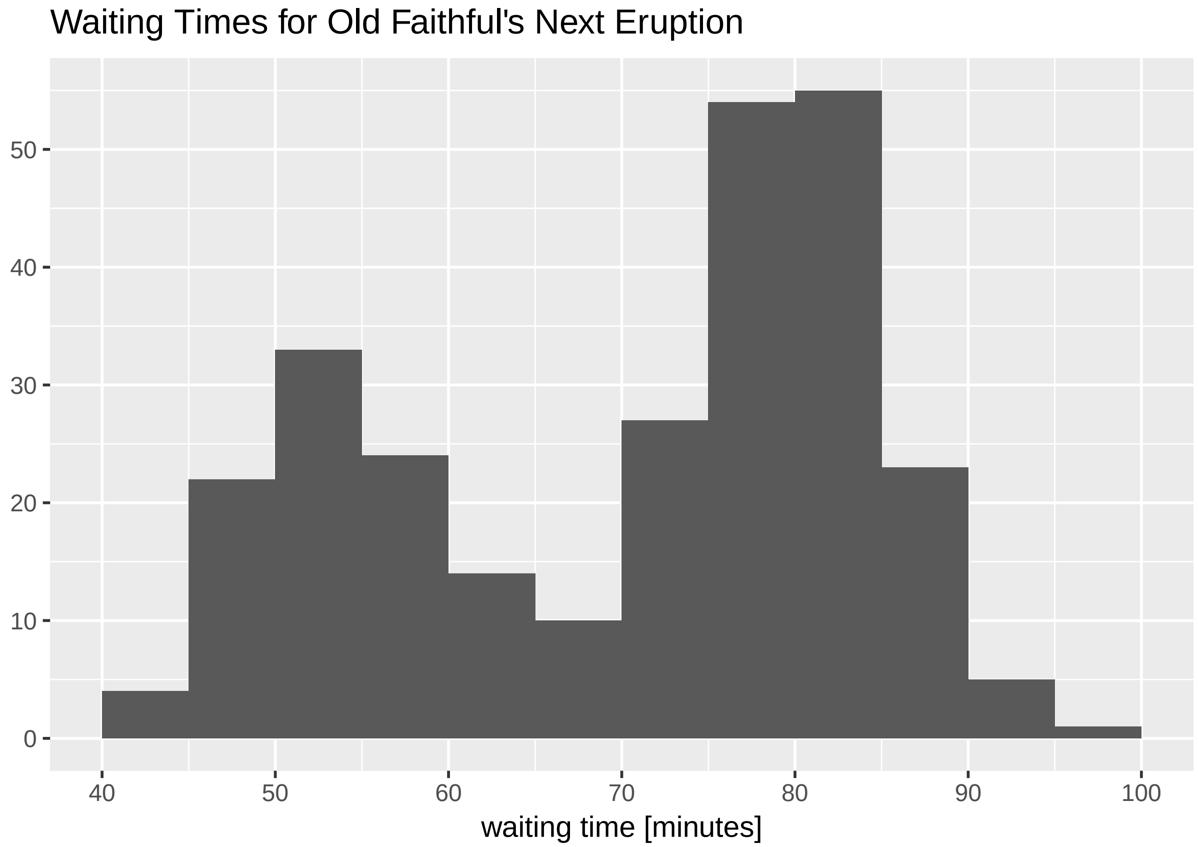
f. Repeat (b) from above, now for *waiting*.

```
ggplot(faithful, aes(waiting)) +
  geom_histogram()
```



g. Repeat (c) from above, now for *waiting*

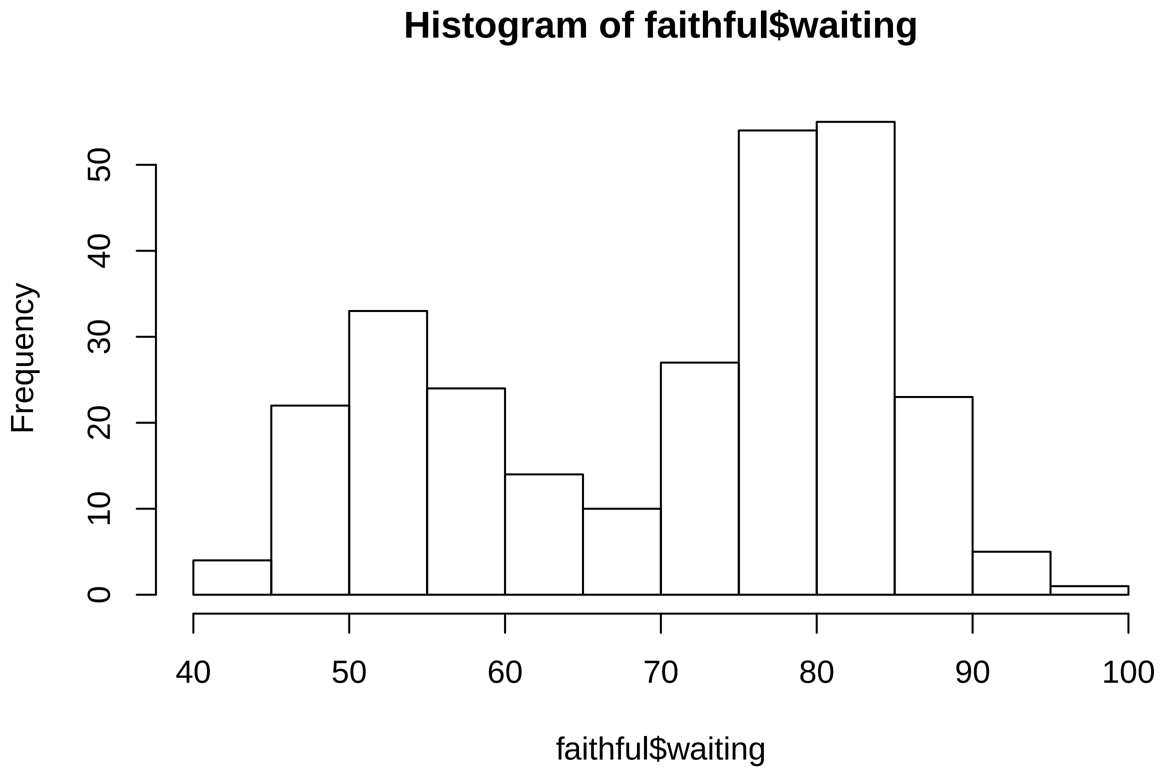
```
ggplot(faithful, aes(waiting)) +
  xlab("waiting time [minutes]") +
  ylab("") +
  xlim(35, 105) +
  ylim(0, 60)+
  scale_y_continuous(breaks = seq(0,65,by=10)) +
  scale_x_continuous(breaks = seq(40, 100, by=10))+
  ggtitle("Waiting Times for Old Faithful's Next Eruption") +
  geom_histogram(binwidth=5, center = 2.5)
```



As on the previous graphs, I clarified the x-axis label so there can be no misunderstanding about what this histogram is portraying. I also added a title for more precision of what this histogram shows. The y-axis label was removed because this histogram displays frequency on the y-axis, which is assumed. The scales on the x and y axes were extended to reach past the extreme data values. I selected a binwidth of 5 because I felt that it helped to display the data well without excessive detail. I also set the center of the bins to 2.5 so the bins would line up with the major grid lines on the x-axis.

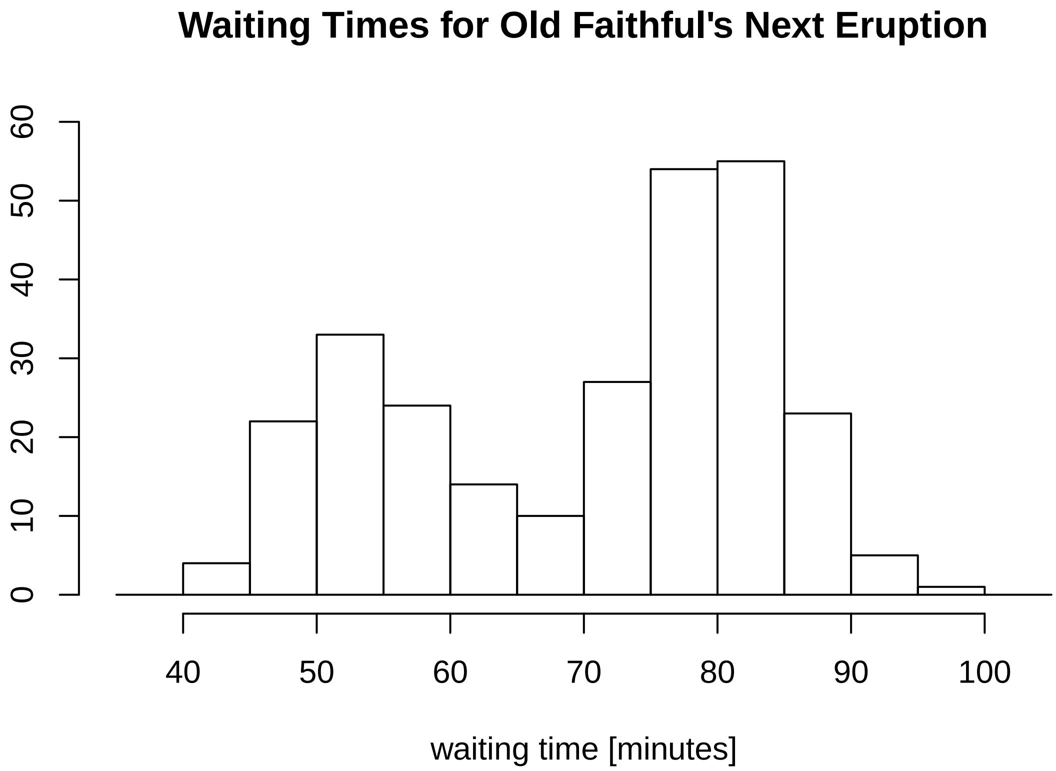
h. Repeat (d) from above, now for *waiting*.

```
hist(faithful$waiting)
```



i. Repeat (e) from above, not for *waiting*.

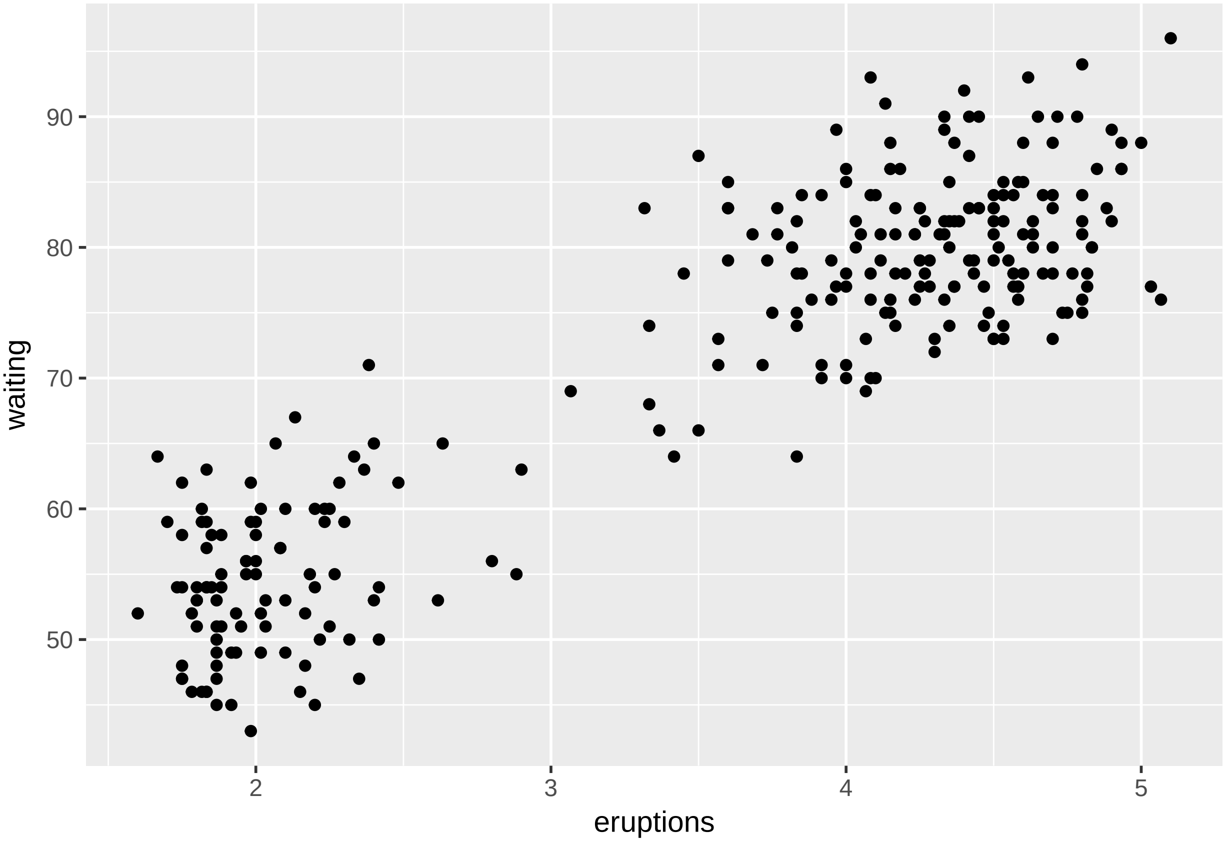
```r
hist(faithful$waiting,
     xlab = "waiting time [minutes]",
     ylab = "",
     main = "Waiting Times for Old Faithful's Next Eruption",
     breaks = seq(35,105,by=5),
     xlim = c(35,105),
     ylim = c(0, 60))
```



I first changes the x and y axis labels as well as the main title because I knew (from the previous graph) what I thought made for clear labels. I then experimented with differnt binwidths, and ended up choosing a binwidth of 5 again, which happened to be the default choice by the `hist()` function. Other than these changes, there wasn't anything else that I felt needed to be modified. I felt like the scales on both the x and y axes worked well for these data, and didn't need to be extended any farther.

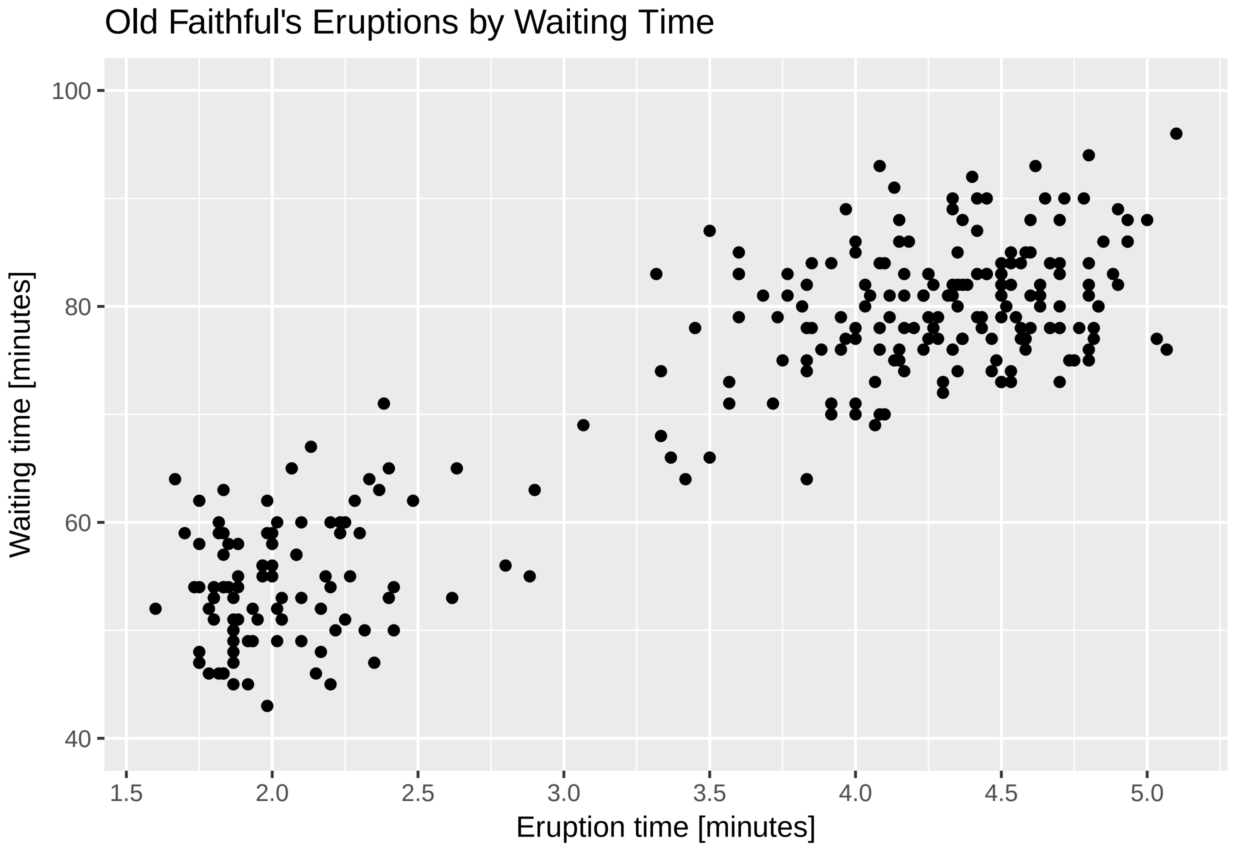
j. Draw a basic scatterplot of the two variables using ggplot2. Assume that eruptions is the explanatory variable.

```r
ggplot(faithful, aes(x=eruptions, y=waiting)) +
  geom_point()
```



k. Further improve your scatterplot from (j). Clearly indicate which changes you made and why you made those changes.

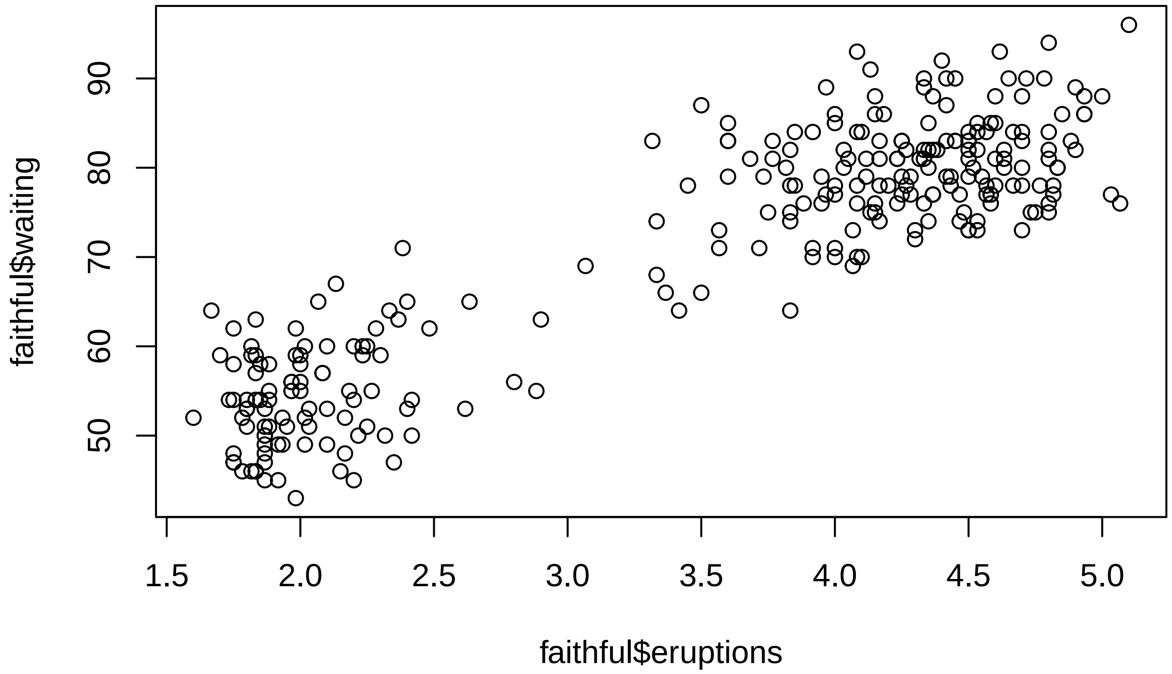
```
ggplot(faithful, aes(x=eruptions, y=waiting)) +
  xlab("Eruption time [minutes]") +
  ylab("Waiting time [minutes]") +
  ggtitle("Old Faithful's Eruptions by Waiting Time") +
  xlim(c(1,5)) +
  ylim(c(40, 100)) +
  scale_x_continuous(breaks = seq(1,5,by=0.5)) +
  geom_point()
```



I first changed the x and y axis labels to 'eruptions' and 'waiting time [minutes]'. I didn't feel like the variable names themselves were enough to give a clear understanding of what the variables mean. I added a little more resolution on the x-axis scale, and extended the endpoints of the scale out to encompass all of our data. I did the same thing on the y axis. Unfortunately, I couldn't figure out how to increase the resolution on the y-axis while still maintaining the `40` and `100` points on the scale. I opted to go with a lower resolution and on the y-axis while keeping those endpoints there.

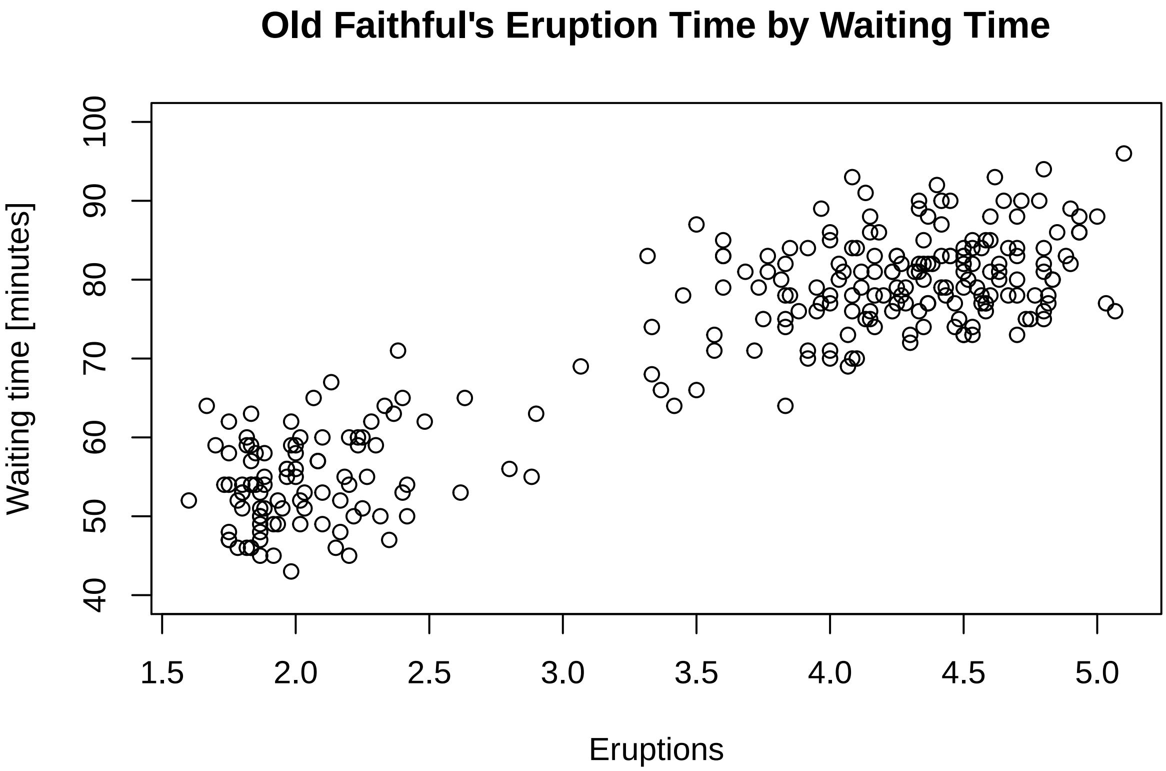
l. Repeat (j) from above, using the the `plot()` function from baseR.

```r
plot(faithful$eruptions, faithful$waiting)
```



m. Repeat (k) from above, using the `plot()` function from baseR.

```
plot(faithful$eruptions,
     faithful$waiting,
     main = "Old Faithful's Eruption Time by Waiting Time",
     xlab = "Eruptions",
     ylab = "Waiting time [minutes]",
     ylim = c(40,100))
```



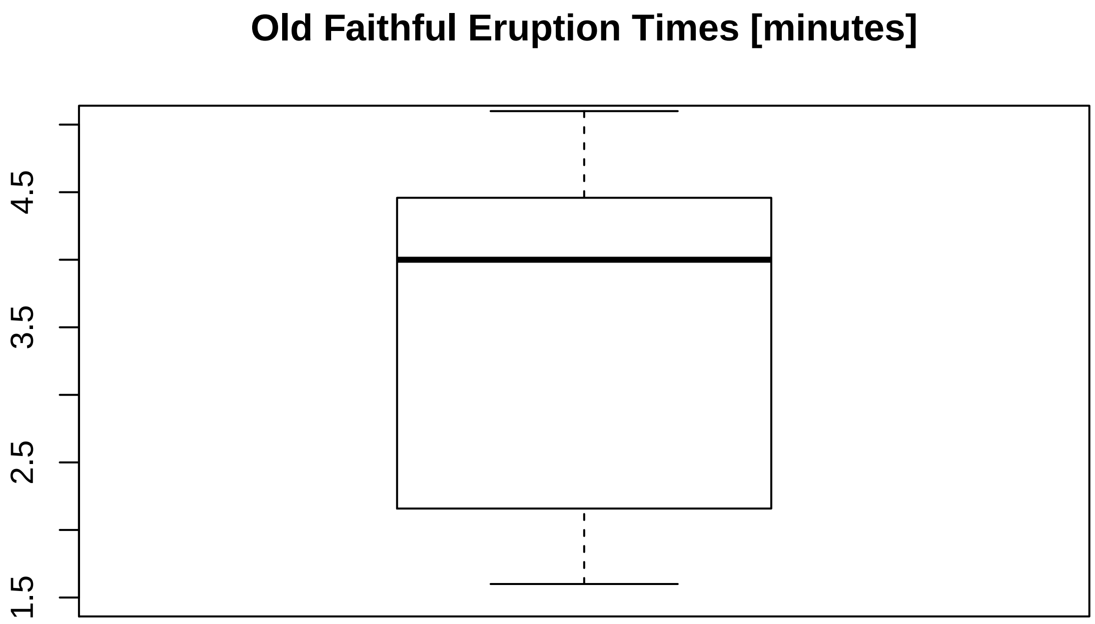
I essentially made the same changes on the baseR plot as I did on the ggplot version. I first added a title, and changes the x and y axis labels from the defualt names to more clear descriptions of the variables. I then manipulated the y-axis scale to include 40 and 100 on the endpoints. In the baseR version, I was able to add those endpoints and keep the scale stepping up by 10. I like this aspect of the baseR scatterplot better than the ggplot version. The y-axis scale needed no changes.

n. Provide a careful discussion of the three final graphs.

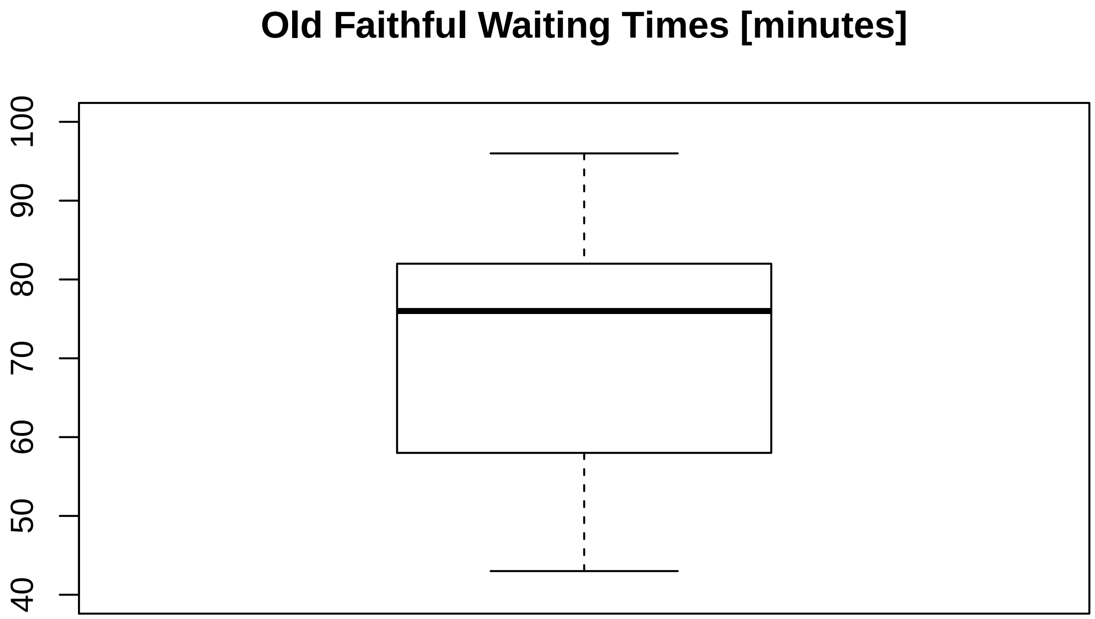
The most distinguishing feature of the first histogram showing the the distribution of eruption times is that it is clearly bimodal. One grouping of the data is centered around 2 minutes, and the other is centered at about 4.5 minutes. This is rather interesting in the context of these data, because it indicates that there are two types of eruptions that one has the possibility of witnessing: a shorter eruption lasting around 2 minutes, or a longer eruption lasting just under 5 minutes. There could also be some interesting implications regarding the "recharging" stages of the geyser, and how different circumstances lead to two groupings of eruption lengths. This all becomes even more interesting when we look at the next histogram, displaying the waiting time for the next eruption of Old Faithful. Again, we see that these data are clearly bimodal. The center of one mode is at about 55 minutes, and the other is at about 80 minutes. These data range from a wait time of 40 minutes to 100 minutes. This is interesting because there may be some connection between the two groupings of eruption times and the two groupings of waiting times. It would make sense to me that a a longer wait time would mean that the geyser is recharging for a longer time, and therefore increase the chances of a longer eruption.
Finally, the scatterplot of eruption time by wait time confirms the presence of positive correlation between eruption time and wait time. Our correlation coefficient is 0.901. This correlation coefficient and the scatterplot tell us that a longer eruption time is associated with a longer waiting time, and vice versa.

o. Would boxplots be good replacements for the two histograms? Creat two basic boxplots with a package of your choice.

```
boxplot(faithful$eruptions,
        main = "Old Faithful Eruption Times [minutes]",
        ylim = c(1.5,5))
```



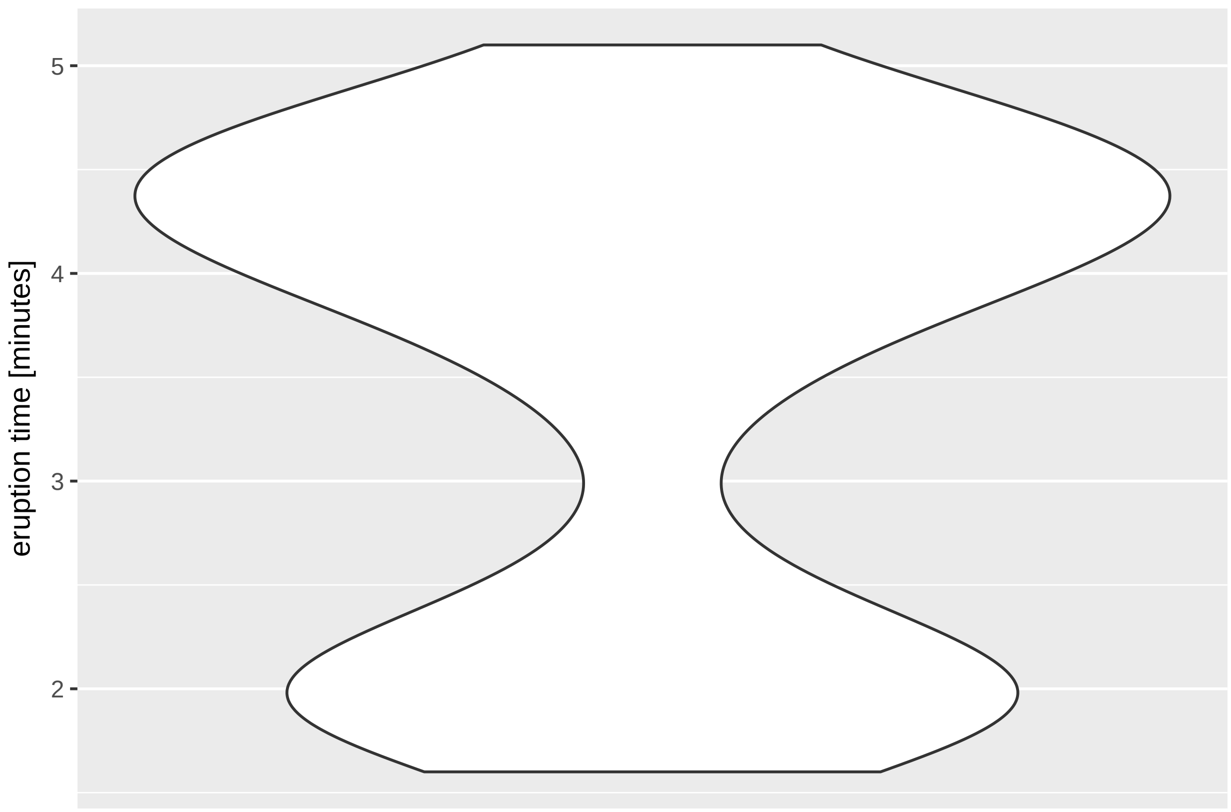
```
boxplot(faithful$waiting,
        main = "Old Faithful Waiting Times [minutes]",
        ylim = c(40, 100))
```



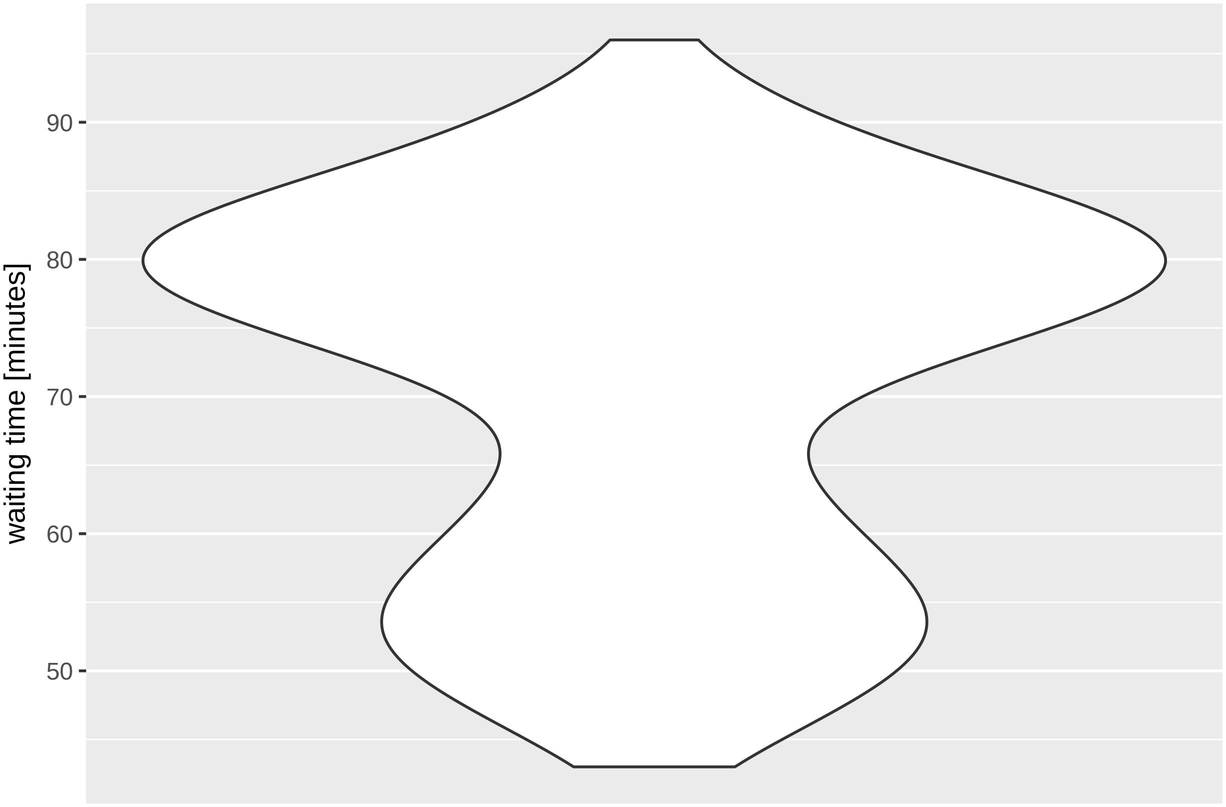
No, I don't think that boxplots would be a good replacement for the histograms in this case. The boxplots don't show us the bimodality that is so obvious in the histograms. Without that information clearly dislayed, I think that a lot of interesting and potentially valuable information is left undiscovered.

p. Create two basic violin plots with a package of your choice. Would violin plots be good replacements for the two histograms?

```r
ggplot(faithful, aes(y = eruptions, x = "var")) +
  geom_violin() +
  scale_x_discrete(breaks = NULL) +
  xlab("") +
  ylab("eruption time [minutes]")
```



```
ggplot(faithful, aes(y = waiting, "var")) +
  xlab("") +
  scale_x_discrete(breaks = NULL) +
  ylab("waiting time [minutes]") +
  geom_violin()
```



Yes, I do think that these violin plots would be a viable replacement for the histograms. They display the bimodality of the distributions of these variables, which is a crucial part of these data. I think that they display just as much valuable information as the histograms do.

2.

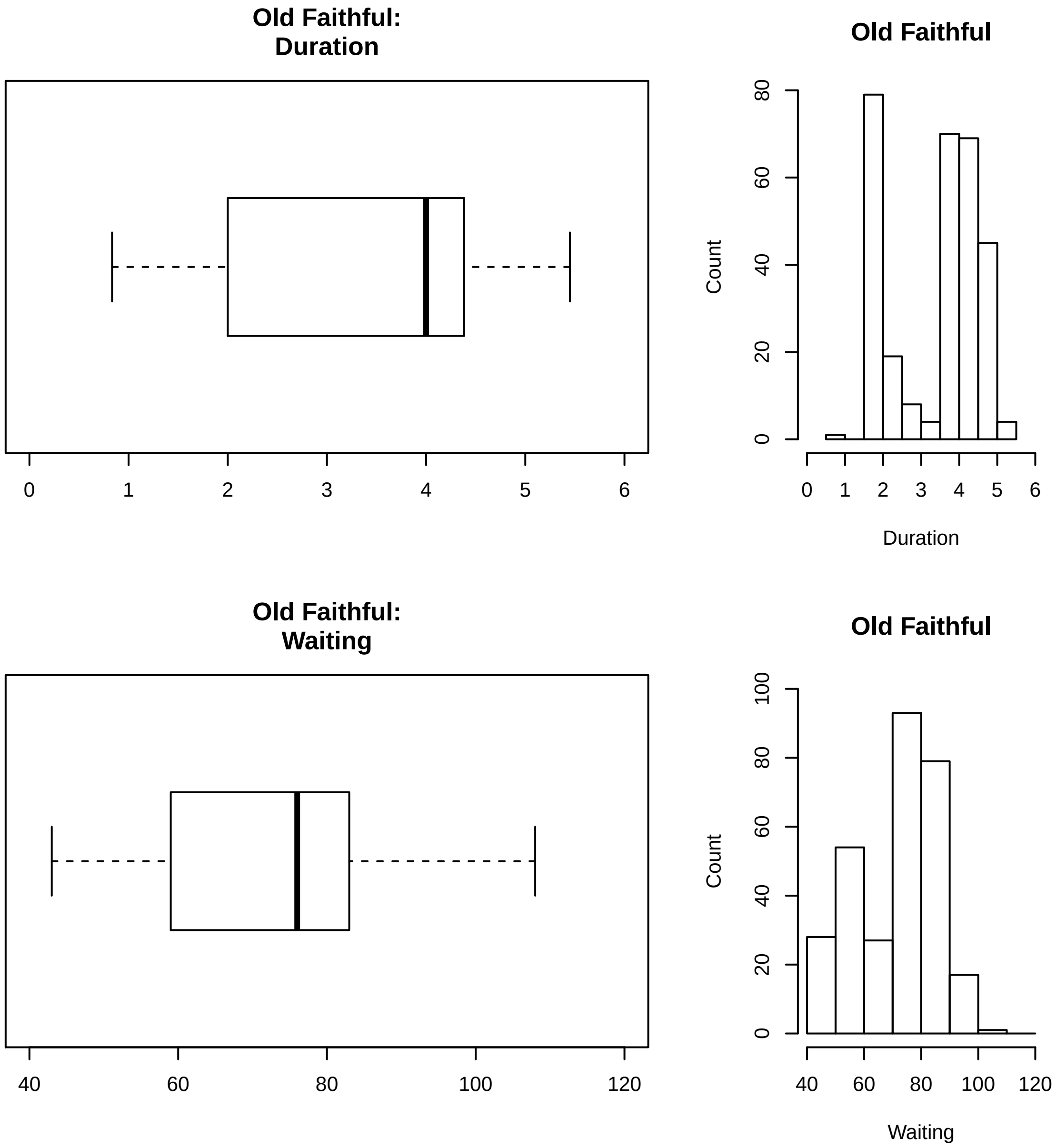
a. Recreate the graphs (and layout) below using baseR.

```
graphics::layout(matrix(c(1, 1, 2,
                          3, 3, 4),
                        2, 3, byrow = TRUE))
par(omar = c(0, 9, 0, 2))
boxplot(geyser$duration,
        main = "Old Faithful:\nDuration",
        horizontal = TRUE,
        breaks = c(1,7,by=1),
        ylim = c(0,6))

hist(geyser$duration,
     main = "Old Faithful",
     xlab = "Duration",
     ylab = "Count",
     xlim = c(0,6))

boxplot(geyser$waiting,
        main = "Old Faithful:\nWaiting",
        horizontal = TRUE,
        ylim = c(40, 120))

hist(geyser$waiting,
     main = "Old Faithful",
     xlab = "Waiting",
     ylab = "Count",
     breaks = seq(40, 120, by=10),
     ylim = c(0,100))
```

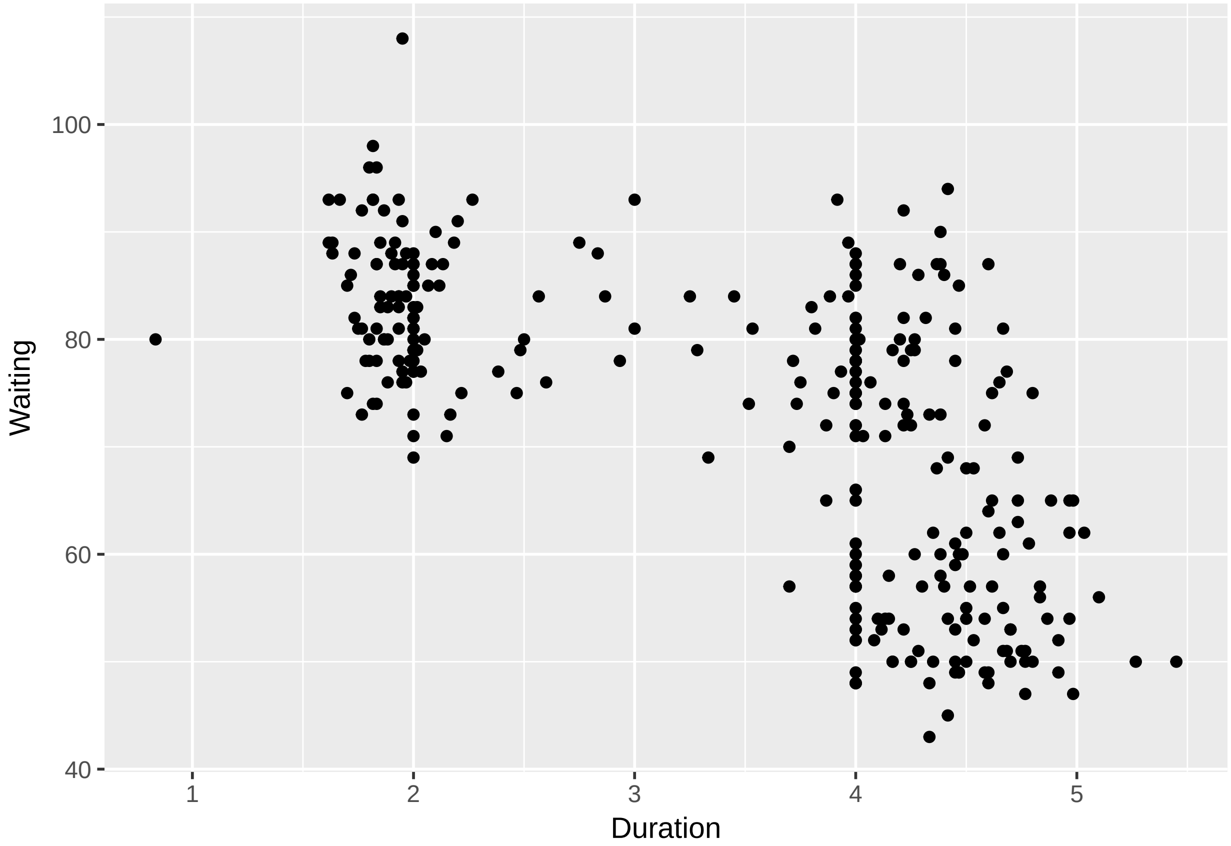


```r
dev.off()
```

```
## null device
##           1
```

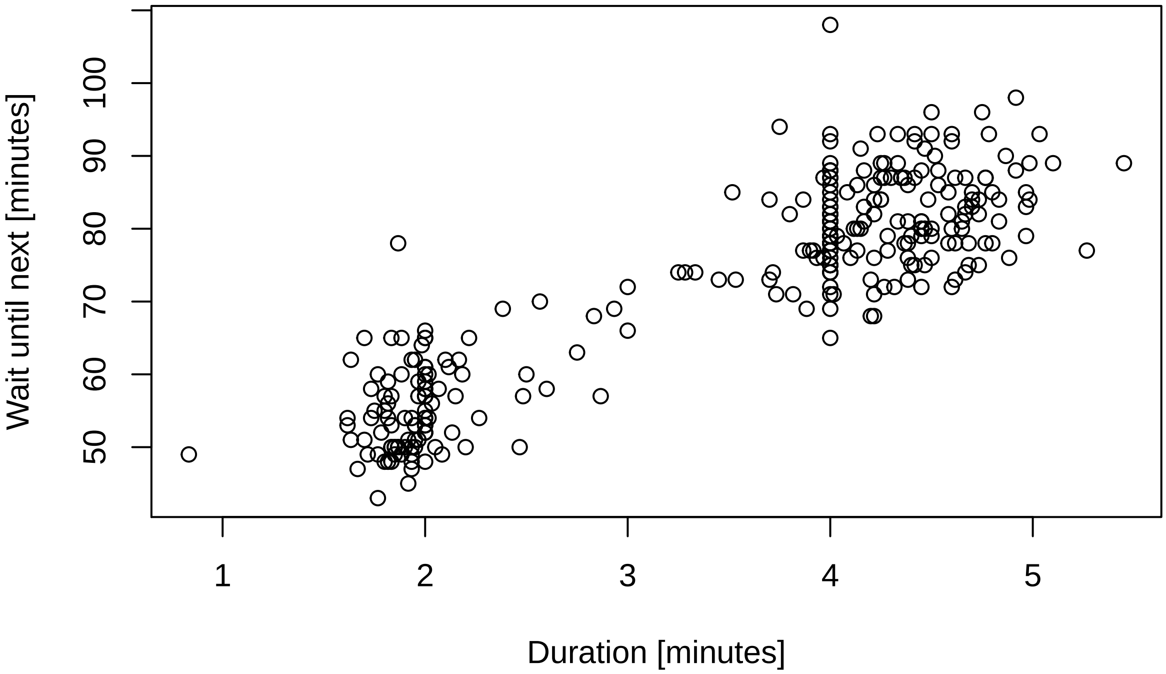
b. Recreate the scatterplot using *ggplot2*.

```r
ggplot(geyser, aes(x = duration, y = waiting)) +
  ylab("Waiting") +
  xlab("Duration") +
  geom_point()
```



c. Create a basic scatterplot for the *geyser* data that matches the overall appearance in question 1 (j).

```
waiting.next <- geyser$waiting[2:299]
duration.next <- geyser$duration[1:298]
plot(duration.next, waiting.next, xlab = "Duration [minutes]", ylab = "Wait until next [minutes]")
```


Wait until next [minutes]
50
60
70
80
90
100
1
2
3
4
5
Duration [minutes]