Answer Sketch for Homework 2

431 Staff and Professor Love

'Due 2018-09-14, version 2018-09-15

Contents

R Setup

Here's the complete R setup we used.

```
knitr::opts_chunk$set(comment=NA)
options(width = 60)

library(magrittr); library(tidyverse)
## make sure these packages are installed in R
```

An Introduction

This answer sketch borrows liberally from a case study entitled *Eruptions of the Old Faithful Geyser* from Chatterjee S Handcock MS Simonoff JS *A Casebook for a First Course in Statistics and Data Analysis* Wiley, 1995.

A geyser is a hot spring that occasionally becomes unstable and erupts hot water and steam into the air. The Old Faithful geyser at Yellowstone National Park in Wyoming is probably the most famous geyser in the world. Visitors to the park try to arrive at the geyser site to see it erupt without having to wait too long; the name of the geyser comes from the fact that eruptions follow a relatively stable pattern. The National Park Service web site which streams a live feed of the geyser includes a time frame during which the next eruption is predicted to occur. Thus, it is of interest to understand and predict the interval time until the next eruption. The main part of this assignment considers the `faithful` data frame, which describes eruption durations and waiting times for the Old Faithful geyser.

```
hw2 <- tbl_df(faithful)
hw2
```

```
# A tibble: 272 x 2
   eruptions waiting
 *     <dbl>   <dbl>
 1      3.6       79
 2      1.8       54
 3      3.33      74
 4      2.28      62
 5      4.53      85
 6      2.88      55
 7      4.7       88
 8      3.6       85
 9      1.95      51
10      4.35      85
# ... with 262 more rows
```

```
summary(hw2)
```

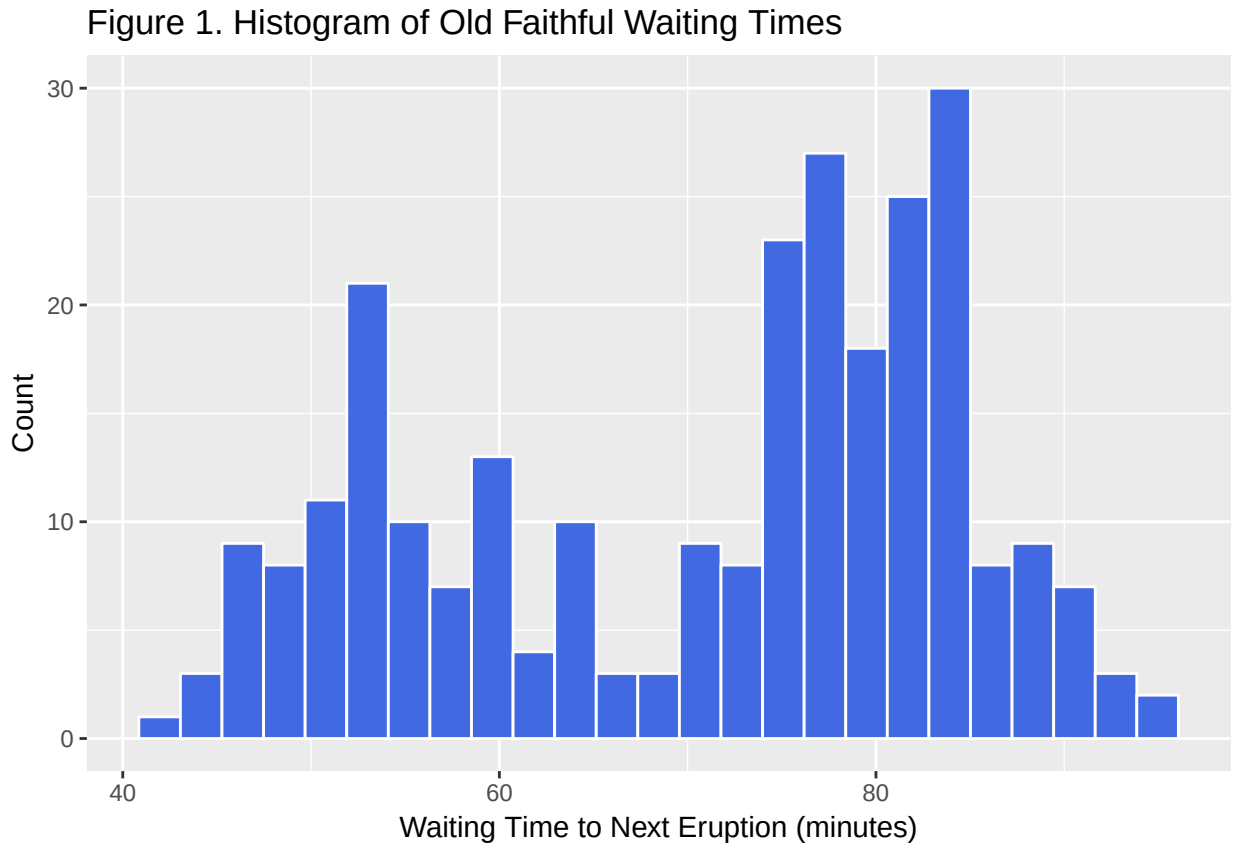
```
   eruptions        waiting
 Min.   :1.600   Min.   :43.0
 1st Qu.:2.163   1st Qu.:58.0
 Median :4.000   Median :76.0
 Mean   :3.488   Mean   :70.9
 3rd Qu.:4.454   3rd Qu.:82.0
 Max.   :5.100   Max.   :96.0
```

Question 1

> Plot a histogram or other summary plot which meaningfully describes the distribution of the waiting times. Be sure it is very clearly labeled.

The first step in any data analysis is simply to look at the data. A histogram gives a good deal of information about the distribution of eruption times, suggesting some interesting structure. Interval times are in the general range of 40 to 100 minutes, but there are apparently two subgroups in the data, centered at roughly 55 minutes, and 80 minutes, respectively, with a gap in the middle.

```
ggplot(hw2, aes(x = waiting)) +
  geom_histogram(bins = 25, fill = "royalblue", color = "white") +
  labs(title = "Figure 1. Histogram of Old Faithful Waiting Times",
       x = "Waiting Time to Next Eruption (minutes)", y = "Count")
```



Comments

This relatively simple histogram is just one of many possible plots we could use to describe the center, spread and shape of a distribution of data.

- We might consider a **stem-and-leaf display** to show the actual data values while retaining the shape of a histogram.

```
hw2 %$% stem(waiting)
```

```
  The decimal point is 1 digit(s) to the right of the |
```

```
  4 | 3
  4 | 55566666777788899999
  5 | 00000111111222223333333444444444
  5 | 555555666677788889999999
  6 | 00000022223334444
  6 | 555667899
  7 | 00001111123333333444444
  7 | 55555555666666666777777777777888888888888888999999999
  8 | 000000001111111111112222222222233333333333333344444444444
  8 | 55555566666677888888999
  9 | 00000012334
  9 | 6
```

- We might consider a **boxplot** or **box-and-whiskers plot** (as we'll see below in Question 4), or perhaps a variant of the boxplot called a **violin plot**.
- If we wanted to compare the distribution of the data to what we might expect from a Normal distribution, we might develop a histogram with an overplotted Normal density function (as I'll show in the discussion of Question 6), or, as we'll see, we might build a **Normal Q-Q plot** to facilitate such a comparison.

Question 2

> What appears to be a typical waiting time? Compare the mean, median and 80% trimmed mean (mean of the middle 80% of the observed waiting times.)

As noted previously, the waiting times appear to cluster into two groups: one centered around 55 minutes, and another, larger, group centered near 80 minutes.

The `summary` function in R provides the five-number summary (minimum, 25th, 50th [median] and 75th percentiles, maximum) and the mean, so that gets us two of our three needed summaries. To get the third, we can either use the `describe` function from the `psych` library, or we can calculate the trimmed mean using the `mean` function.

```
# summary provides mean and median, along with quartiles and min/max
hw2 %$% summary(waiting)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   43.0    58.0    76.0    70.9    82.0    96.0
```

```
# describe in the psych library also provides the trimmed mean we're looking for...
hw2 %$% psych::describe(waiting)
```

```
   vars   n mean    sd median trimmed   mad min max range
X1    1 272 70.9 13.59     76    71.5 11.86  43  96    53
    skew kurtosis   se
X1 -0.41    -1.16 0.82
```

```
# this trims 10% from the top and 10% of the bottom
# of the distribution, and then takes the mean of
# what remains, just as psych::describe does
hw2 %$% mean(waiting, trim = 0.1)
```

```
[1] 71.49541
```

Comments

In addition to writing full code chunks, we use Markdown to ask R to fill in the values as we go, rather than inserting them through copy and paste, or retyping. This substantially reduces the chance of errors, and lets us generate a revised document quickly if we find an error in the data.

Look at the Markdown file for this assignment to see, for instance, how we are using code to fill in the values in the next bullet.

- The distribution of 272 waiting times has mean 70.9 minutes, and median 76 minutes, with a trimmed mean of 71.5 minutes.
- Note that `signif` (which is used in the code to generate the previous sentence) is a function which rounds to the specified number of "significant figures" (digits). This has nothing to do with the notion of statistical *significance*.

Question 3

What is the inter-quartile range, and how does it compare to the standard deviation?

The 25th percentile is 58 and the 75th percentile is 82 so the inter-quartile range is 24 minutes, which is considerably larger than the standard deviation of 13.6 minutes. Specifically, the IQR is about 77% larger than the SD, since $\frac{IQR}{SD} = 24 \ / \ 13.6 = 1.77$.

Comments

- The **range** of the data is just the maximum minus the minimum, or 96 minus 43 or 53. Note that if you ask R for the `range` of the `hw2$waiting` variable with `range(hw2$waiting)`, this yields a vector with two values: the minimum and the maximum, for example 43, 96.
- If the data were Normally distributed, we would expect that about 68% of observations would fall within one standard deviation of the mean. For any distribution, the middle half of the distribution falls within the first and third quartiles. If the data followed a Normal distribution very closely, the IQR would be 25-50% larger than the standard deviation.
- The **median absolute deviation**, or MAD, is another candidate measure of dispersion or scale, which has a more direct relationship with the standard deviation (the population standard deviation is well estimated by the MAD for Normally distributed data).
  - The MAD is defined as the median of the absolute deviations of each observation from the data's median, multiplied by a constant (1.48 by default in R).
  - In this case, the MAD for the waiting times is 11.86 minutes, so the ratio of the standard deviation to the MAD for the Old Faithful waiting times is 13.59 / 11.86 = 1.15.

Question 4

Is the distribution multi-modal or unimodal? How do you know?

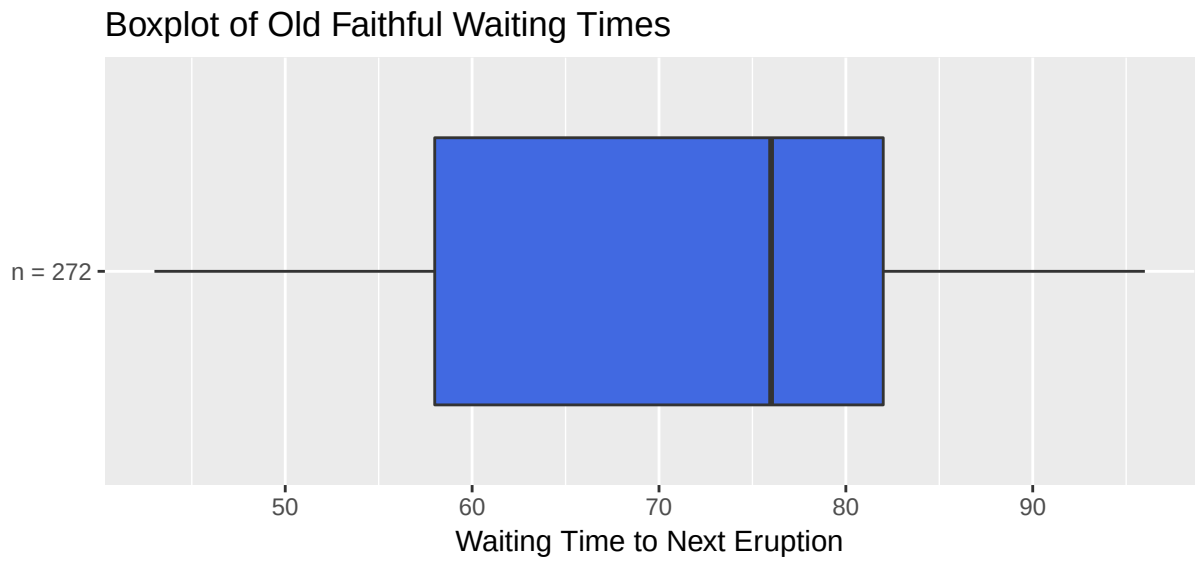
The distribution clearly has one cluster of waiting times centered at 50-55 minutes and another, larger, cluster centered at 80 minutes. The fact that the distribution has multiple local maxima would usually suggest that we interpret this as multi-modal (specifically, because there are two local maxima, we'd say bimodal) data, where a single summary of the center might not be as useful as it would be with unimodal data.

Comments

Not all exploratory techniques are equally effective for these data. A **boxplot** shows that the waiting times are in the general range of 40-100 minutes, but the bimodal distribution is hidden by the form of the plot. Boxplots are mostly used to make comparisons.

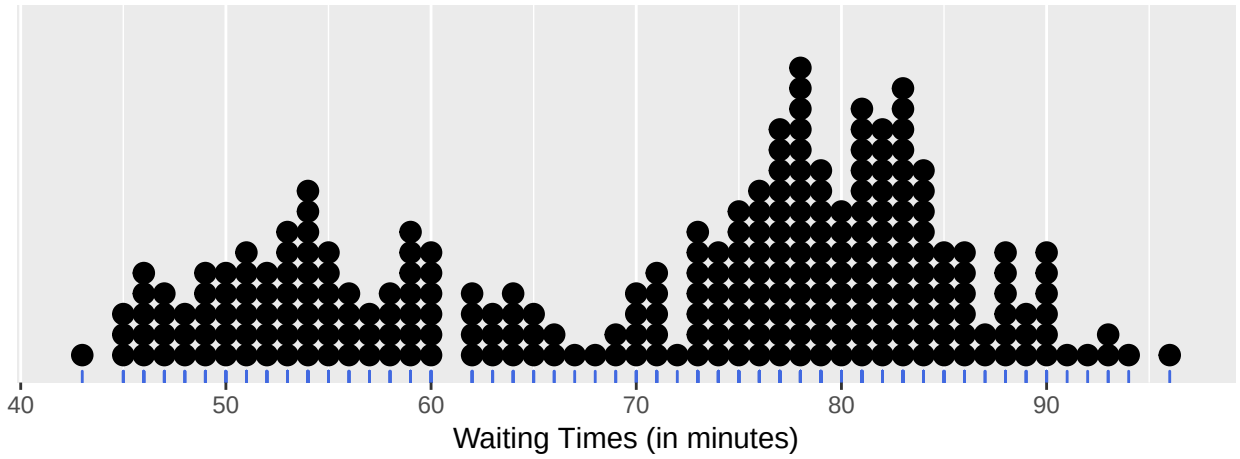
In the notes, we described one way to get a boxplot for a single distribution. That was as follows.

```
ggplot(hw2, aes(x = "n = 272", y = waiting)) +
  geom_boxplot(fill = "royalblue") +
  coord_flip() +
  labs(title = "Boxplot of Old Faithful Waiting Times",
       x = "", y = "Waiting Time to Next Eruption")
```



A **dotplot** might help here. Though there are other ways to generate these plots, we like the following approach, which creates a dot plot augmented by a rug plot.

```
ggplot(hw2, aes(x=waiting)) +
  geom_dotplot(binwidth=1) +      ## create dot plot
  geom_rug(col = "royalblue") +  ## add rug
  scale_y_continuous(breaks=NULL) + ## Remove ticks
  theme(axis.title.y=element_blank()) + ## Remove label
  labs(x = "Waiting Times (in minutes)")
```



Question 5

> Is the distribution skewed (and if so, in which direction) or is it essentially symmetric? How do you know?

This is a left-skewed distribution, with the mean substantially less than the median.

On Assessing Skew

A reasonable measure of skewness or asymmetry in a distribution, sometimes called $\mathbf{skew_1}$ or *non-parametric skew*, compares the mean to the median, while using the standard deviation as the unit of measurement.

Skewness is a far more meaningful concept with unimodal data than with multi-modal data like this. We can declare the skew to be positive or negative regardless of whether the data are in fact multi-modal or follow any other particular pattern.

The formula is $\text{skew}_1 = \frac{mean - median}{SD}$ where:

- A positive skew_1 value indicates right (sometimes called positive) skew where the mean exceeds the median, and
- a negative skew_1 value indicates left skew, where the mean is less than the median.
- $\text{skew}_1 = 0$ when the mean is equal to the median, an indication of potential symmetry.
- skew_1 values exceeding 0.2 in absolute value are sometimes taken to indicate fairly substantial skew (far enough from a Normal distribution to call into question whether the mean and standard deviation alone are sufficient to approximate the data well).
- If skew_1 exceeds 0.5 in absolute value, that indicates very strong skew.

In our data, we can calculate skew_1 with, for instance:

```
hw2 %>%
  summarize(skew1 = (mean(waiting) - median(waiting))/sd(waiting))
```

```
# A tibble: 1 x 1
   skew1
   <dbl>
1 -0.375
```

Interpreting a Graph with help from $skew_1$

Generally, if the mean is more than 20% of a standard deviation away from the median, I would expect a graph of the data to show substantial skew. If the mean is within 20% of a standard deviation of the median, I wouldn't necessarily expect the data to look meaningfully asymmetric.

For the waiting times, $skew_1$ is -0.38. So the mean is about 38% of a standard deviation below the median, indicating fairly substantial left skew.

Question 6

> Are there any unusual (outlier) values in the distribution, and if so, what are they?

No. For instance, a boxplot of the waiting times (see Question 4) shows no outliers in the distribution.

The boxplot identifies as an outlier any point that is more than 1.5 IQR outside of the middle half of the data. We define the **inner fences** as falling at Q1 - 1.5 IQR and Q3 + 1.5 IQR, and any points outside those fences are identified by the boxplot and considered to be, at the least, outlier candidates. Sometimes we'll define more serious outliers using a tougher standard. We define the **outer fences** as falling at Q1 - 3 IQR and Q3 + 3 IQR, so that any points outside the outer fences are then described as serious outliers.

Z Scores for Most Outlying Values

Another approach to assessing how outlier-prone the data appear to be, in comparison to what we might expect from a Normal distribution, is to calculate the maximum (and minimum) Z scores for the data set.

The Z score for any particular observation X is (X - mean) / SD , so that our $skew_1$ measure, for instance, may be interpreted as the negative of the Z score for the median. If the data were really drawn from a Normal distribution, then we'd expect:

- roughly 10% of observations to have a Z score greater than 1.645 in absolute value.
- roughly 5% of observations to have a Z score greater than 1.96 in absolute value.
- roughly 1% of observations to have a Z score greater than 2.57 in absolute value.
- less than 3 in 1,000 observations to have a Z score greater than 3 in absolute value.
- less than 1 in 10,000 observations to have a Z score greater than 4 in absolute value.

In this case, the maximum observed waiting time was 96, which has a Z score of 1.85.

How do I know this? Well. . .

```
hw2 %>% summarize(max(waiting), z = (max(waiting) - mean(waiting)) / sd(waiting))
```

```
# A tibble: 1 x 2
  `max(waiting)`     z
           <dbl> <dbl>
1             96  1.85
```

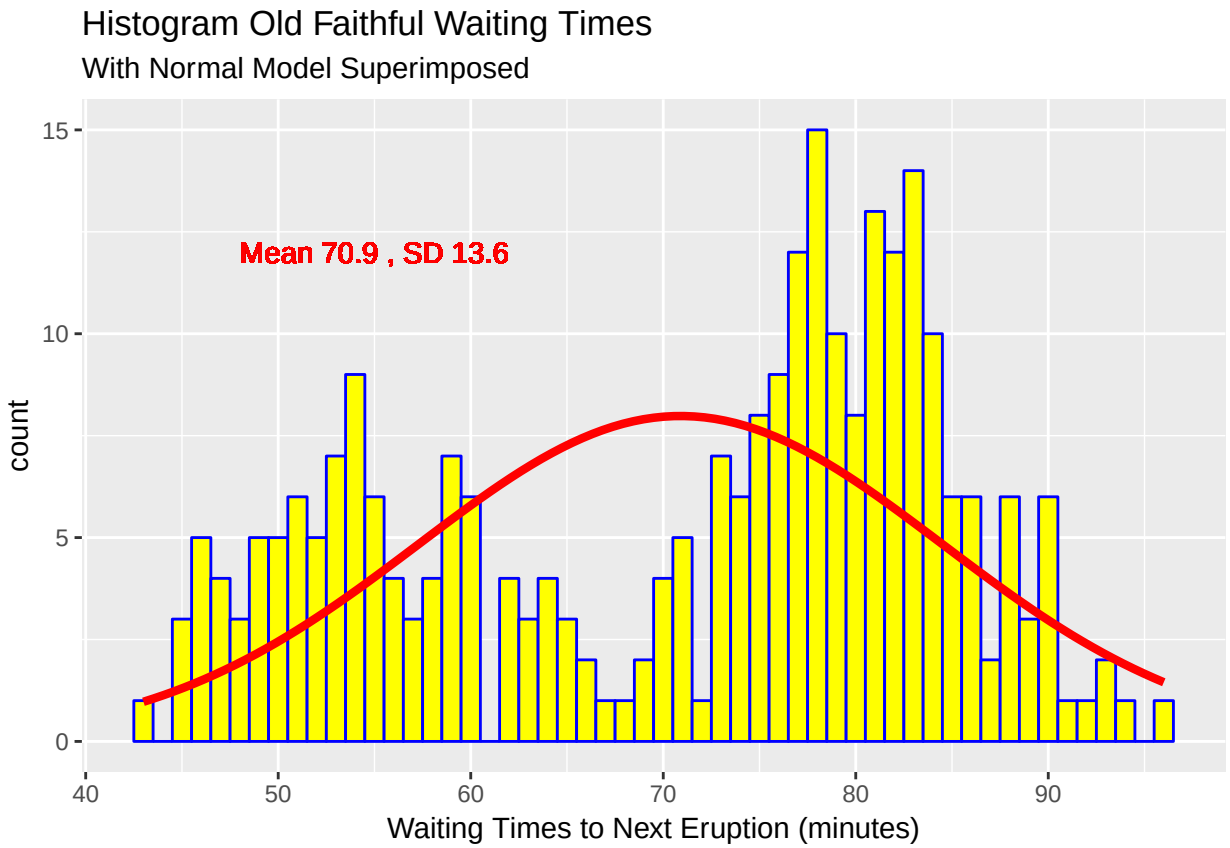
The minimum observed time was 43, which has a Z score of -2.05. With a sample of size 272 these particular values seem to suggest that the data is somewhat **less** outlier-prone than we might expect from a Normal distribution. This is also referred to as the distribution having **lighter tails** than the Normal distribution. But, in essence, we already know that the data don't follow a Normal distribution from our graphs.

Histogram vs. Expectation under a Normal Distribution

Sometimes, it's easier to see *light-tailed* (fewer outlying values than we'd expect from a Normal distribution) vs. *heavy-tailed* (more outliers than a Normal) distributions by directly comparing the histogram to the Normal distribution with the data's mean and standard deviation

```
## ggplot including histogram of waiting times
## with Normal model superimposed

ggplot(hw2, aes(x = waiting)) +
  geom_histogram(binwidth = 1, fill = "yellow",
                 col = "blue") +
  stat_function(fun = function(x, mean, sd, n)
    n * dnorm(x = x, mean = mean, sd = sd),
    args = with(hw2,
                c(mean = mean(waiting),
                  sd = sd(waiting),
                  n = length(waiting))),
    col = "red", lwd = 1.5) +
    geom_text(aes(label = paste("Mean", round(mean(hw2$waiting),1),
                                ", SD", round(sd(hw2$waiting),1))),
              x = 55, y = 12, color="red") +
  labs(title = "Histogram Old Faithful Waiting Times",
       subtitle = "With Normal Model Superimposed",
       x = "Waiting Times to Next Eruption (minutes)")
```



Question 7

> Would a model using the Normal distribution be an appropriate way to summarize the waiting time data? Why or why not?

No, a Normal distribution would not be an appropriate way to summarize this distribution, as the data are multi-modal, and substantially left skewed. Based on the histogram's appearance, the distribution might be well described as a mixture of two different (and perhaps close to Normal) distributions, one centered at 50-55 minutes, and another (which would be a more frequently observed component of the mixture), centered at about 80 minutes.

The mean waiting time of about 71 minutes, for example, seems informative, but it doesn't actually describe a typical result in either subgroup.

The Empirical Rule

A useful idea is that roughly 95% of the observations will lie within two standard deviations of the mean when the data follow a Normal distribution.

Here, that means within the range of 70.9 - 2(13.6) = 43.7 to 70.9 + 2(13.6) = 98.1 minutes. In this case, all but one (the minimum value of 43) of the 272 waiting times fall in this range, which is more than we would expect if the waiting times were Normally distributed.

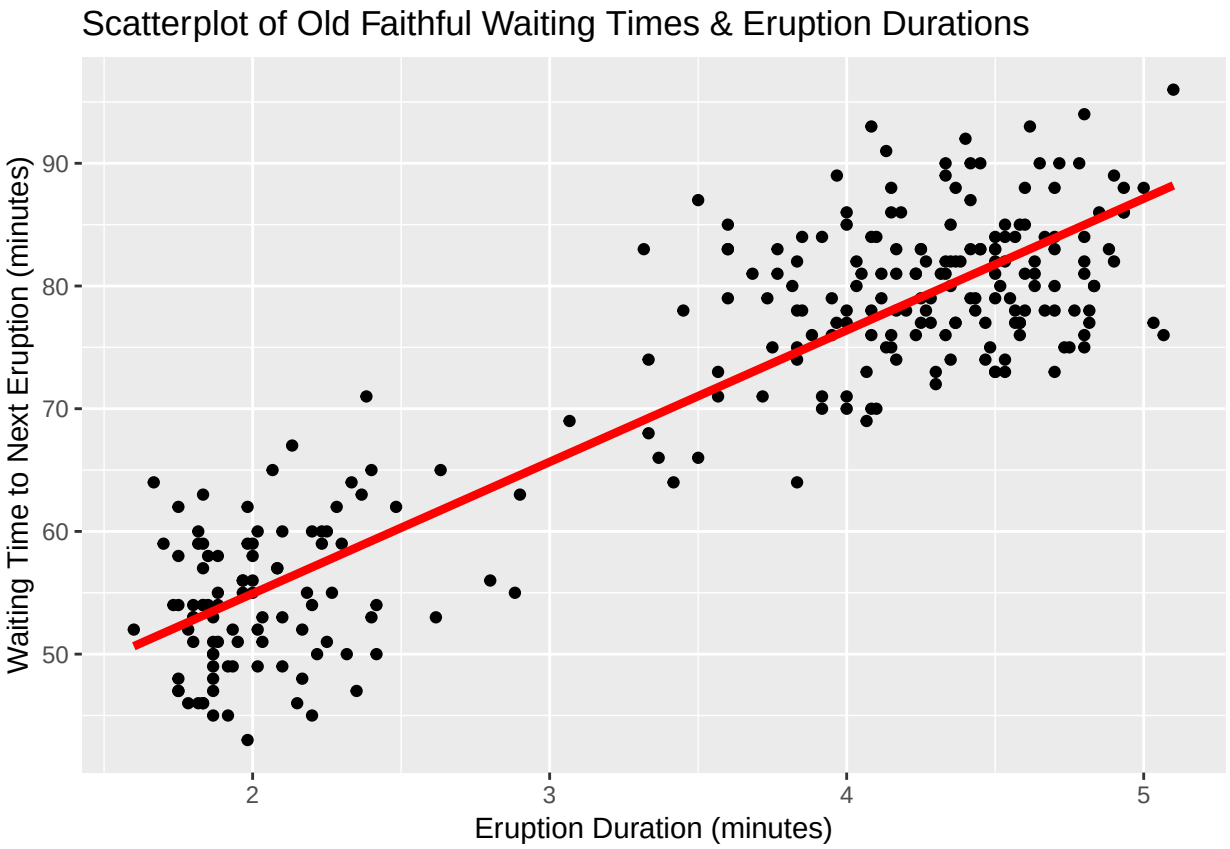
Question 8

> Plot a scatterplot of the waiting times (y-axis) vs. the eruption durations (x-axis), and be sure your plot is very clearly labeled. Describe your general impression of the plot. What sort of relationship do you see?

How, then, can we help the tourists? We need more information. One readily available characteristic of the geyser is the duration of the previous eruption. We can think of the `faithful` data as pairs of the form (eruption duration, time to next eruption) and then build a scatterplot of those pairs.

The plot reveals two clusters: it appears that eruption durations of 1.5 to 2.5 minutes are followed by shorter waiting times of 50-65 minutes, while longer eruption durations (of roughly 4 to 5 minutes) are followed by longer waiting times of 75-95 minutes.

```
ggplot(hw2, aes(x = eruptions, y = waiting)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red", lwd = 1.5) +
  labs(title = "Scatterplot of Old Faithful Waiting Times & Eruption Durations",
       y = "Waiting Time to Next Eruption (minutes)",
       x = "Eruption Duration (minutes)")
```



The existence of two subgroups in this type of data is rare, but not unheard of. J. S. Rinehart, in a 1969 paper in the *Journal of Geophysical Research*, provides a mechanism for this pattern based on the temperature level of the water at the bottom of a geyser tube at the time the water at the top reaches the boiling temperature. That a shorter eruption would be followed by a shorter waiting time (and a longer eruption would be followed by a longer waiting time) is also consistent with Rinehart's model, since a short eruption is characterized by having more water at the bottom of the geyser heated short of boiling temperature, and left in the tube. This water has been heated somewhat, however, so that it takes less time for the next eruption to occur. A long eruption results in the tube being emptied, so the water must be heated from a colder temperature, which takes longer.

Question 9

> What is the correlation of waiting time with eruption duration? How would you interpret this result?

The Pearson correlation coefficient is 0.901.

```
hw2 %$%
  cor(waiting, eruptions)
```

```
[1] 0.9008112
```

This indicates a strong positive (or direct) and nearly linear relationship between eruption duration and waiting time.

On Correlation

Any two variables can be correlated. Any correlation that is not zero indicates some degree of correlation. Also, correlation is unitless: it's not a percentage of anything. The correlation is 0.9 here: undoubtedly a strong positive correlation. A perfect correlation would be +1 or -1 (depending on the direction of the relationship) and whether a correlation is strong depends powerfully on the context. For now, it's probably best to suggest that any correlation above about 0.5 in absolute value is usually fairly strong, and any correlation below 0.3 in absolute value is usually fairly weak.

On Rounding

The waiting time data in the `faithful` data frame are rounded to the nearest integer number of minutes. It is therefore silly to claim substantially more precision than 0 decimal places in evaluating summary statistics based on those data. Adding a single additional significant figure in summarizing data is usually somewhat justifiable, but any more than that is not.

Specifying a standard deviation, or a correlation coefficient to more than one decimal place in this case is likely to be inappropriate. The standard deviation of waiting times was about 14 minutes, or maybe 13.6 minutes, but not really 13.5949738 minutes. The correlation of waiting time and eruption duration is about 0.9, but not really 0.9008112.

Borrowing from a great line by John Tukey in a slightly different context:

> Be approximately right, rather than exactly wrong.

Question 10

> Would a linear model be an appropriate thing to use in attempting to predict the waiting time given the most recent eruption duration, based on these data? Why or why not? If you like, you can add a simple least squares regression line to the plot.

Yes, a linear model might well be a useful summary here, as the waiting time for the next eruption shows a nearly linear relationship with eruption duration. The scatter of points tracks with the regression line fairly closely across the range of eruption durations.

Question 11

> Investigate questions 8-10 again using the* `geyser` *data in the* `MASS` *package, and compare your results appropriately.

```
hw2extra <- tbl_df(MASS::geyser)
hw2extra
```

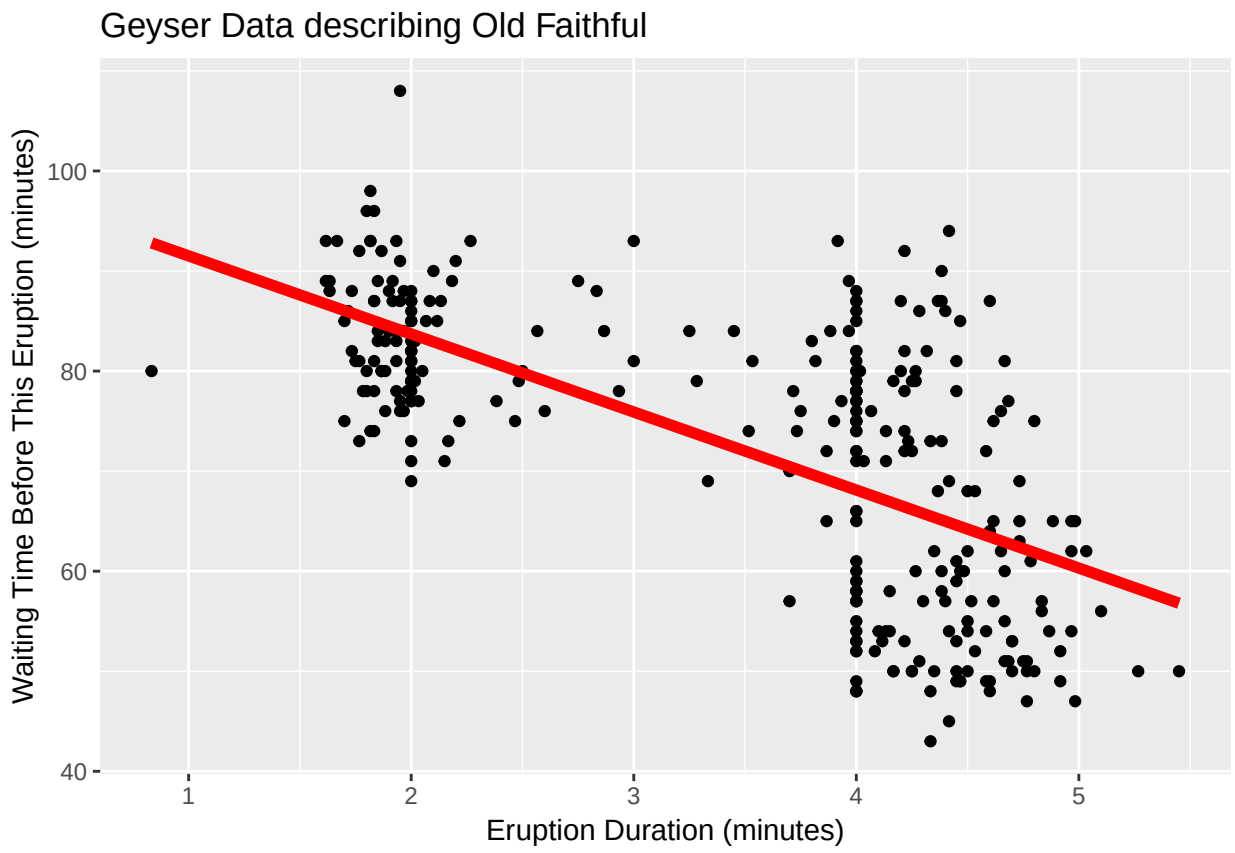
```
# A tibble: 299 x 2
   waiting duration
 *   <dbl>    <dbl>
 1      80     4.02
 2      71     2.15
 3      57     4
 4      80     4
 5      75     4
 6      77     2
 7      60     4.38
```

```
 8      86     4.28
 9      77     2.03
10      56     4.83
# ... with 289 more rows
```

There are several differences between the data frames.

- One difference is that we have 299 observations in the `geyser` data, 27 more than we had in the `faithful` data.
- The second, and more important distinction is that the waiting times now refer to the *current* eruption, so that when we plot the results as they are given, they show the waiting time preceding *this* eruption, rather than the waiting time preceding *the next* eruption.
- Third, we see lots of eruption durations specified as exactly 2 or exactly 4, in the `geyser` data, creating vertical lines in the scatterplot.

```
ggplot(hw2extra, aes(x = duration, y = waiting)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red", lwd = 2) +
  labs(title = "Geyser Data describing Old Faithful",
       x = "Eruption Duration (minutes)",
       y = "Waiting Time Before This Eruption (minutes)")
```



The slope of the regression line is *negative* here, indicating that eruptions of shorter duration (say, 1.5 to 2.5 minutes) were preceded by longer waiting times while eruptions of longer duration (say, 4-5 minutes) were preceded by shorter waiting times. The correlation is -0.645, which indicates a strong negative association between the waiting time for this eruption and its duration.

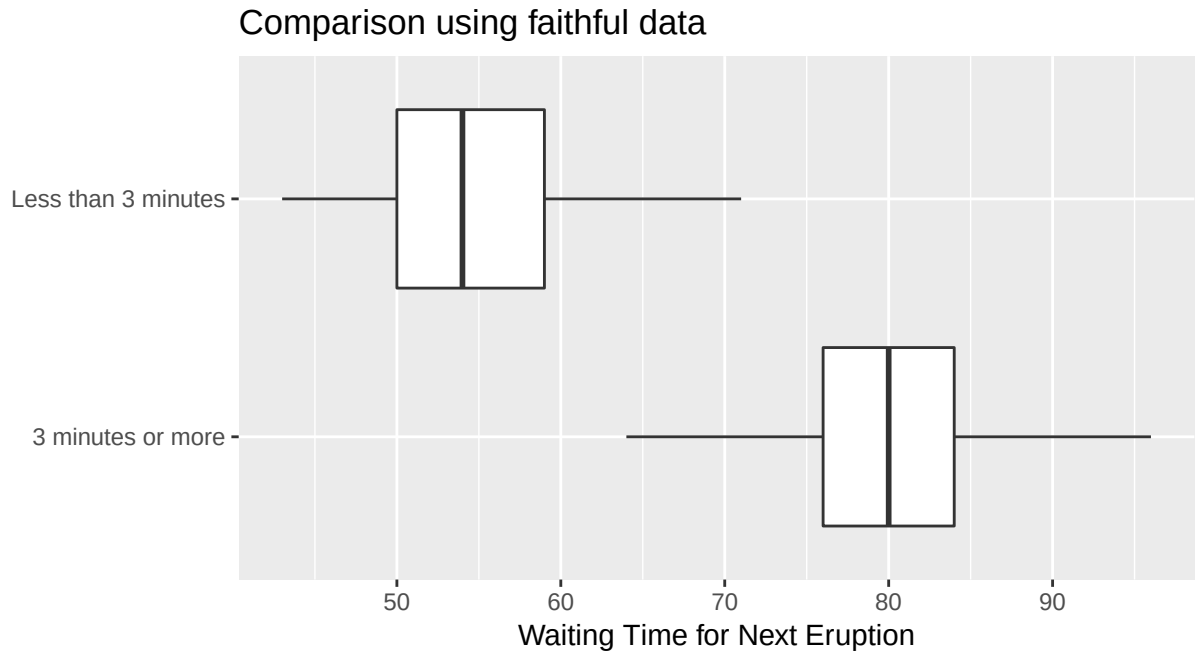
The conclusions we would draw here, are thus similar to those we developed for the original `faithful` data, but the available information is arranged a bit differently.

An Alternative Model for the `faithful` Data

We noticed two dominant effects in the `faithful` data: there are two different subgroups, and a longer eruption tends to be followed by a longer time interval until the next eruption. Suppose we separate the eruptions by whether the duration is less than three minutes.

```
hw2 <- hw2 %>%
  mutate(timegroup = ifelse(eruptions < 3,
                            "Less than 3 minutes",
                            "3 minutes or more"))

ggplot(hw2, aes(x = timegroup, y = waiting)) +
  geom_boxplot() +
  coord_flip() +
  labs(title = "Comparison using faithful data",
       x = "", y = "Waiting Time for Next Eruption")
```




```
mosaic::favstats(waiting ~ timegroup, data = hw2)
```

```
            timegroup min Q1 median Q3 max     mean
1   3 minutes or more  64 76     80 84  96 79.98857
2 Less than 3 minutes  43 50     54 59  71 54.49485
        sd   n missing
1 5.994239 175       0
2 5.840098  97       0
```

Based on these summaries, a simple prediction rule would be that an eruption of less than 3 minutes will be followed by a waiting time of about 55 minutes, while an eruption of duration 3 minutes or more will be followed by a waiting time of about 80 minutes. Further, the latter (longer) waiting time would be expected to occur about 2/3 of the time.