HW1

Michael Demos

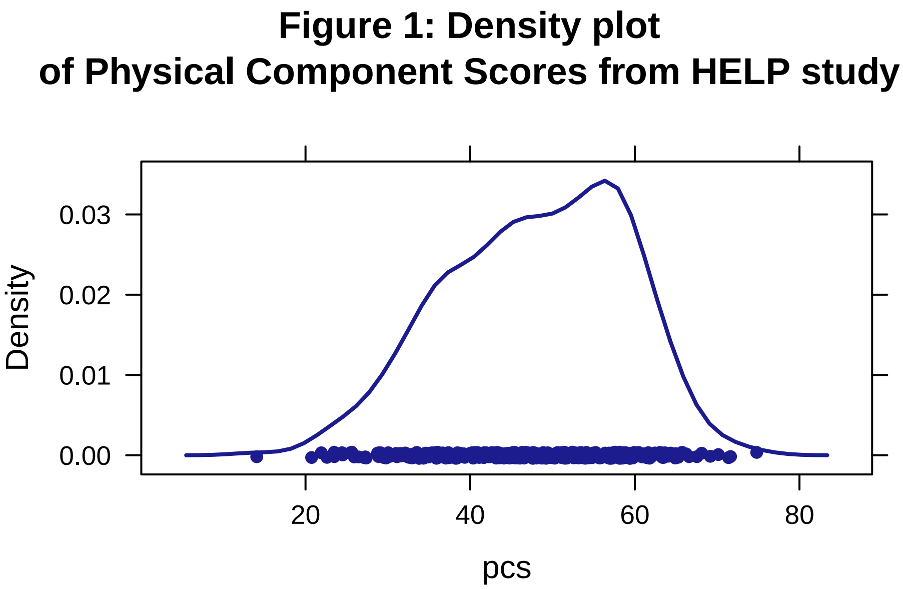
Problem 1

The **HELPrct** dataset in the mosaicData package includes data from the Health Evaluation and Linkage to Primary Care study, which was conducted in Boston 10 years ago. One of the study variables is a measure of physical function, with higher scores being better (possible scores can range from 0 to 100 points). Describe the sample size plus CENTER, SPREAD and SHAPE of this distribution, providing only a single measure of center and a single measure of spread. Be sure to provide an interpretation in the context of the problem.Could you provide any different graph to describe the distribution of this variable?

```
favstats(~ pcs, data=HELPrct)
```

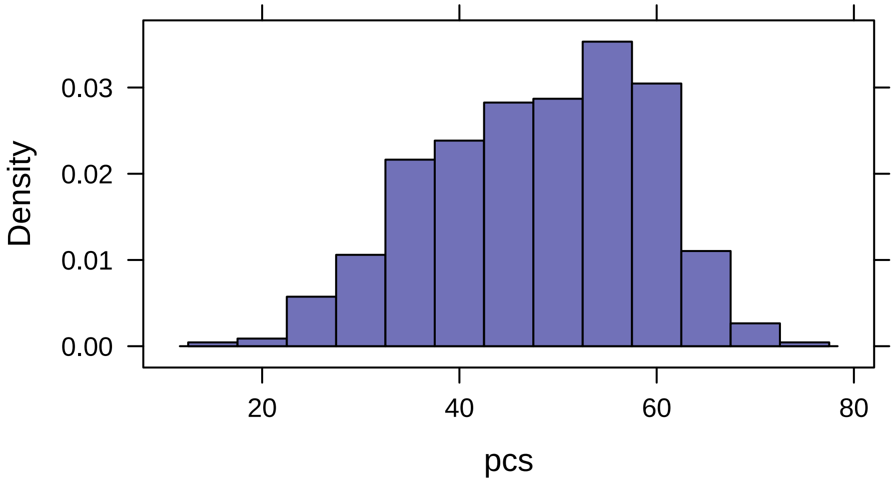
```
##       min       Q1   median       Q3      max     mean      sd   n missing
##  14.07429 40.38438 48.87681 56.95329 74.80633 48.04854 10.7846 453       0
```

```
densityplot(~ pcs,
  main="Figure 1: Density plot\nof Physical Component Scores from HELP study",
  data=HELPrct)
```



SOLUTION:

```
histogram(~ pcs, data=HELPrct, width = 5)
```



The sample size is 453 people. The center of the distribution (mean) is 48 points. The distribution has a mound shape. The spread (standard deviation) is about 11 points. Most people received a score of about 49, which is 1 point under the middle on the physical function scale, which means that the sample as a whole shows an overall physical function score almost perfectly in the middle, reflecting the fact that some people are less physically able than others, and that some are more physically able than others.

Problem 2 (Old Faithful)

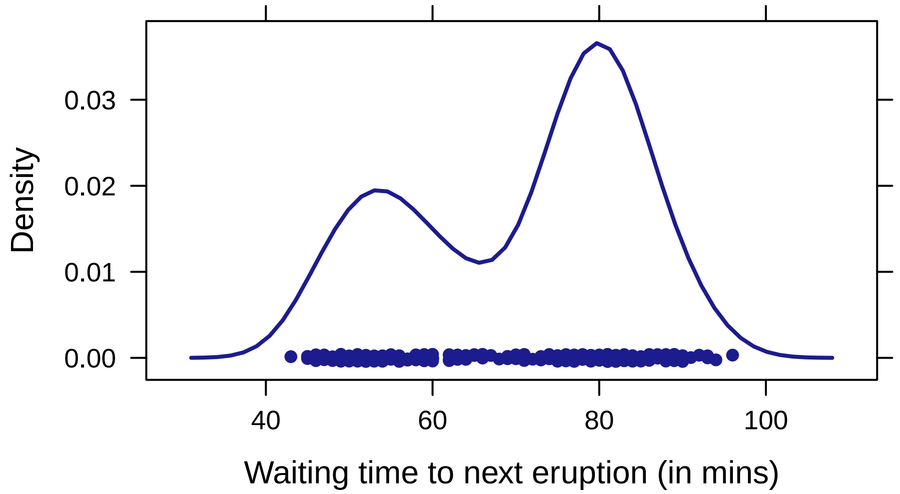
The **faithful** dataset contains the waiting time (in minutes) to the next eruption of the Old Faithful geyser in Yellowstone National Park in Wyoming. Describe the sample size plus CENTER, SPREAD and SHAPE of this distribution, providing only a single measure of center and a single measure of spread. Be sure to provide an interpretation in the context of the problem (and don't forget to specify units).Could you provide any different graph to describe the distribution of this variable?

```
favstats(~ waiting, data=faithful)
```

```
##  min Q1 median Q3 max     mean       sd   n missing
##   43 58     76 82  96 70.89706 13.59497 272       0
```

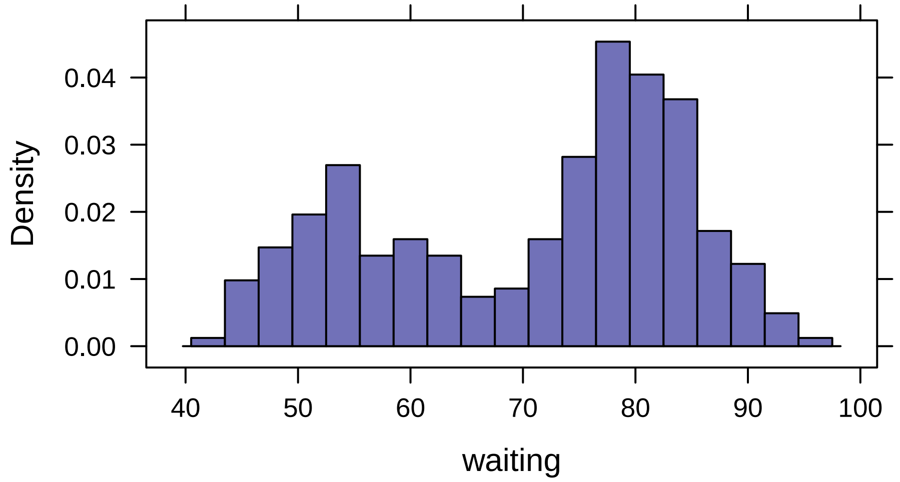
```
densityplot(~ waiting,
  xlab="Waiting time to next eruption (in mins)",
  main="Figure 2: Density plot of Old Faithful geyser dataset", data=faithful)
```

Figure 2: Density plot of Old Faithful geyser dataset



SOLUTION:

```
histogram(~ waiting, data = faithful, width = 3)
```



The sample size is 272 eruptions recorded. The data is bimodal has the shape of two mounds. The center of distribution (median) is 76min. The spread (IQR) of the data is about 24min. This means that the two most likely wait times for an eruption is about 52min (less likely) and about 76min (more likely). In other words, the geyser erupts either every 52 or 76 minutes, on average.