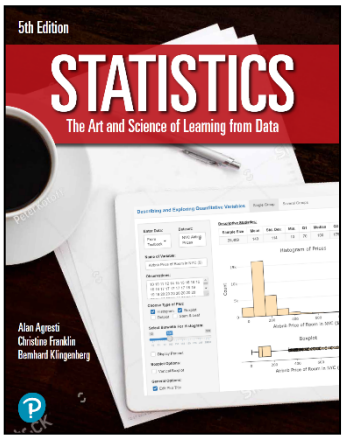

R Code for Examples in the book
"Statistics: The Art and Science of Learning from Data"
by Agresti, Franklin and Klingenberg, 5th edition

Chapter 2

Example 2: Shark Attacks – Distribution of a Variable

Creating dataset

```
region <- c('Florida', 'Hawaii', 'South Carolina', 'California',
            'North Carolina', 'Australia', 'South Africa',
            'Reunion Island', 'Brazil', 'Bahamas', 'Other')
frequency <- c(203, 51, 34, 33, 23, 125, 43, 17, 16, 6, 138)
attacks <- data.frame(region, frequency)
```

Display the entire dataset:

```
attacks
```

```
##            region frequency
## 1         Florida       203
## 2          Hawaii        51
## 3  South Carolina        34
## 4      California        33
## 5  North Carolina        23
## 6       Australia       125
## 7    South Africa        43
## 8  Reunion Island        17
## 9          Brazil        16
## 10        Bahamas         6
## 11          Other       138
```

Display only the first 6 lines:

```
head(attacks, 6)
```

```
##           region frequency
## 1        Florida       203
## 2         Hawaii        51
## 3 South Carolina        34
## 4     California        33
## 5 North Carolina        23
## 6      Australia       125
```

Or, you can read in the dataset via:

```
# > path <-
'https://raw.githubusercontent.com/artofstat/data/master/Chapter2/sharks.csv'
# > attacks <- read.csv(path)
```

Create column for the proportion in the dataframe:

```
attacks$proportion <- attacks$frequency / sum(attacks$frequency)
head(attacks,6)
```

```
##           region frequency proportion
## 1        Florida       203 0.29462990
## 2         Hawaii        51 0.07402032
## 3 South Carolina        34 0.04934688
## 4     California        33 0.04789550
## 5 North Carolina        23 0.03338171
## 6      Australia       125 0.18142235
```

Create column for the percentage:

```
attacks$percentage <- 100 * (attacks$frequency / sum(attacks$frequency))
head(attacks,6)
```

```
##           region frequency proportion percentage
## 1        Florida       203 0.29462990  29.462990
## 2         Hawaii        51 0.07402032   7.402032
## 3 South Carolina        34 0.04934688   4.934688
## 4     California        33 0.04789550   4.789550
## 5 North Carolina        23 0.03338171   3.338171
## 6      Australia       125 0.18142235  18.142235
```

For nicer printing in R, use dplyr package and declare data frame as a table, using function as_tibble(). To install dplyr package, use install.packages('dplyr'). Then, load package into R using library('dplyr'):

```
library(dplyr)
attacks <- as_tibble(attacks)
attacks
```

```
## # A tibble: 11 × 4
##    region         frequency proportion percentage
##    <chr>              <dbl>      <dbl>      <dbl>
##  1 Florida              203    0.295       29.5
##  2 Hawaii                51    0.0740       7.40
##  3 South Carolina        34    0.0493       4.93
##  4 California            33    0.0479       4.79
##  5 North Carolina        23    0.0334       3.34
##  6 Australia            125    0.181       18.1
##  7 South Africa          43    0.0624       6.24
##  8 Reunion Island        17    0.0247       2.47
##  9 Brazil                16    0.0232       2.32
## 10 Bahamas                6    0.00871      0.871
## 11 Other                138    0.200       20.0
```