A Report on

Executive Summary of Module – 2

Introduction to Analytics

(ALY 6000)

Guided By:

Prof. Shahram Sattar

Submitted By:

Krutarth Jaiswal (002965528)

Date of Submission:

26/01/2022

A. Descriptive Analysis of the dataset (BullTroutRML2):

- The given dataset contains four columns (Age, Fork Length, Lake, and Era) and 96 rows.
- This data described two lakes (Harrison and Osprey) and two time periods (1977 to 80 and 1997 to 01). Also, it gives information about the ages of the fishes and fork length, which were living in these two lakes in a given time.

```
> headtail(data,n=3)
   age  fl     lake     era
1   14 459 Harrison 1977-80
2   12 449 Harrison 1977-80
3   10 471 Harrison 1977-80
94   4 298   Osprey 1997-01
95   3 279   Osprey 1997-01
96   3 273   Osprey 1997-01
```

1. **Mean:** It is derived by taking the sum of the values and dividing with the number of values in a data series.

- The mean value of age and fork length is around 5.7 and 326.1.

```
> mean(data$age)
[1] 5.770833
> mean(data$fl)
[1] 326.1146
```

2. **Median:** The middle value of the sorted data is called median.

- The median value of age and fork length is 6 and 352.5.

```
> median(data$age)
[1] 6
> median(data$fl)
[1] 352.5
```

3. **Mode:** The mode is the value that has highest number of occurrences in the dataset. Mode can have numeric as well as character data.
   - R does not have in-built function to find a mode. So, I have created a user function to calculate mode of the given data.
   - Mode values of Fork length, Age, Lake, and Era are 357, 7, Harrison, and 1997-01 respectively.

```
> cat("Mode value of Fork Length :",result)
Mode value of Fork Length : 357
> cat("Mode value of Age :",result1)
Mode value of Age : 7
> result2
[1] Harrison
Levels: Harrison Osprey
> result3
[1] 1997-01
Levels: 1977-80 1997-01
```

4. **Quartile:** Quartiles of an ordered dataset are divided mainly into 3 parts (first quartile-25%, second quartile-50% and third quartile-75%).
   - The first, second, and third quartiles of age are 4, 6, and 8 respectively.
   - The first, second, and third quartiles of fork length are 258, 352, and 406 respectively.

```
> quantile(data$age)
  0%  25%  50%  75% 100%
   0    4    6    8   14
> quantile(data$fl)
    0%    25%    50%    75%   100%
 20.0 258.0 352.5 406.0 688.0
```

5. **Variance:** It is the measure of how much values is away from the mean value.
   - The variance is always positive and greater values will indicate higher dispersion.
   - Variance values of age and fork length are 8.55 and 12589 respectively.

```
> var(data$age)
[1] 8.557456
> var(data$fl)
[1] 12589.34
```

6. **Standard Deviation:** The standard deviation is the positive square root of variance.
   - Standard deviation of age and fork length are 2.9 and 112.2 respectively.

```
> sd(data$age)
[1] 2.925313
> sd(data$fl)
[1] 112.2022
```

7. **Skewness:** Skewness is the parameter of measurement which tells about the shape of the data distribution.
   - It helps to check irregularity and asymmetry of the distribution.
   - To calculate skewness in R, **moments** package is required.
   - The skewness value of age is 0.21, which is greater than 0, then the graph is said to be positively skewed with the majority of data values less than the mean.
   - The skewness value of age is -0.51, which is less than 0, then the graph is said to be negatively skewed with the majority of data values greater than the mean.

```
> skewness(data$age)
[1] 0.212223
> skewness(data$fl)
[1] -0.5103268
```

8. **Kurtosis:** Kurtosis is a statistical method that measures the sharpness of the peak in the data distribution.
   - Kurtosis value of age is 2.73, which is less than 3 then it is said to be negative kurtosis.
   - Kurtosis value of fork length is 3.9, which is greater than 3 then it is called positive kurtosis.

```
> kurtosis(data$age)
[1] 2.735144
> kurtosis(data$fl)
[1] 3.914769
```

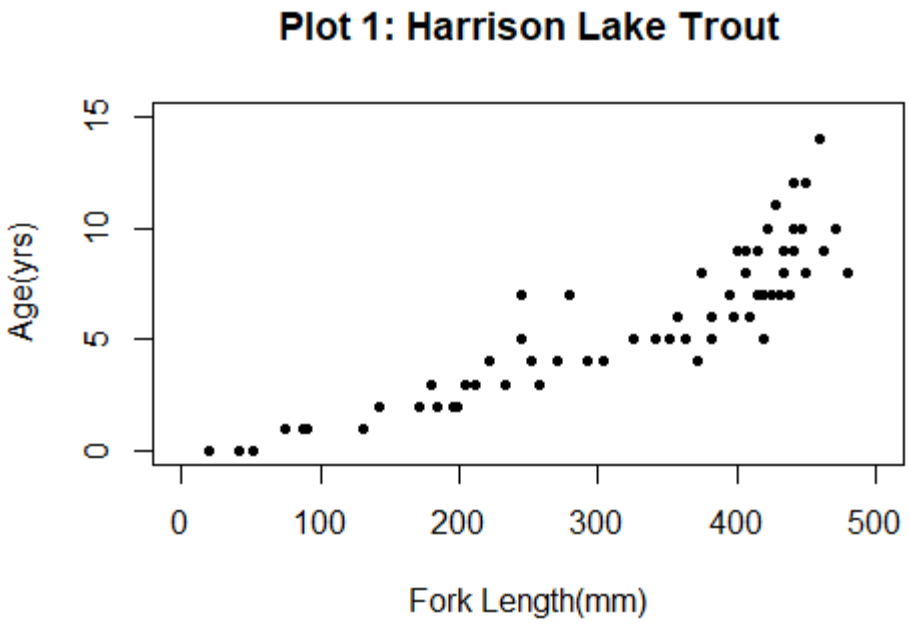
9. **Outliers:** An outlier is a value or an observation that is distant from other values.
   - There are lots of methods to find outliers. But I have used Grubbs's test to detect whether the highest or lowest value is an outlier in the dataset.
   - The highest value of 14 is an outlier in age and the highest value 688 is also an outlier in fork length.

```
> grubbs.test(data$age)

        Grubbs test for one outlier

data:  data$age
G = 2.81309, U = 0.91582, p-value = 0.1986
alternative hypothesis: highest value 14 is an outlier

> grubbs.test(data$fl)

        Grubbs test for one outlier

data:  data$fl
G = 3.22530, U = 0.88935, p-value = 0.04456
alternative hypothesis: highest value 688 is an outlier
```

A. Visualization of the dataset:

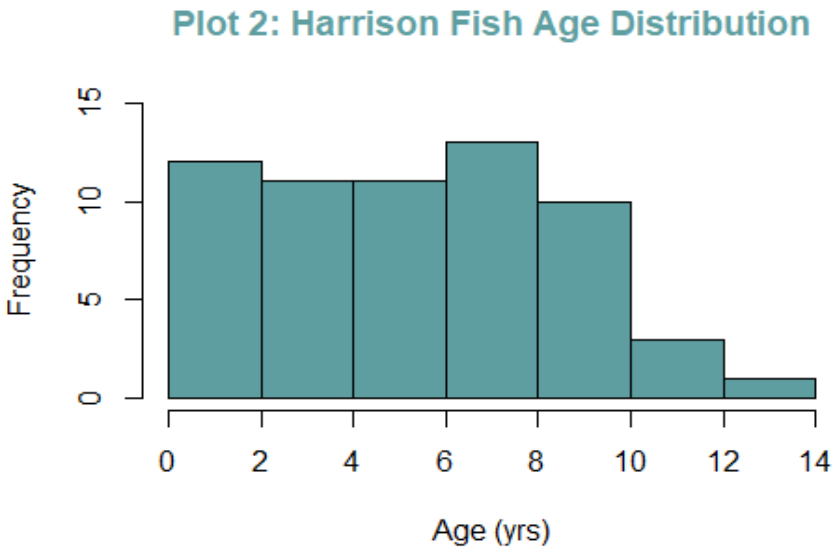
1. Scatter plot of Age and Fork length:

- Firstly, I have created a scatter plot of Age and Fork length with the help of BullTroutRML2 dataset.
- Scatter plot helps to show relationship between two variables and how much one variable can be affected by another.
- The relationship between age and fork length can be referred as having positive correlation because values of both the variables have been increasing.



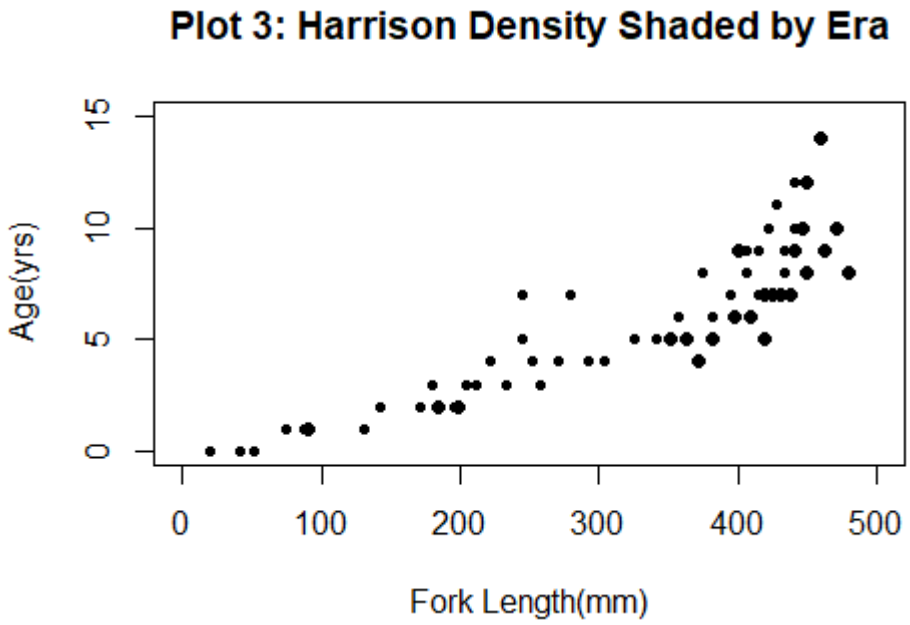
2. Histogram of Age:

- A histogram represents the number of occurrences of values of a variable. It can provide information of the distribution pattern of the data.
- It also helps to identify the maximum and minimum values of the variable. In this graph, the maximum age of fish is 14 and minimum age is 3.
- It shows that the age span of 6 to 8 years has the highest frequency.



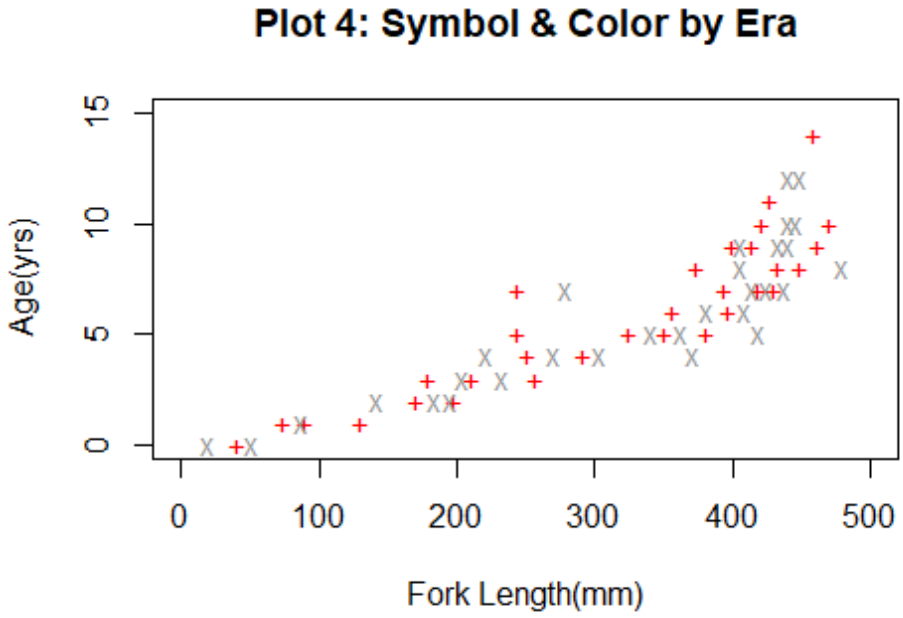
3. Scatter plot with density:

- It is a scatter plot between age and fork length with two levels of shading.
- If the points are darker, those points are denser than others.



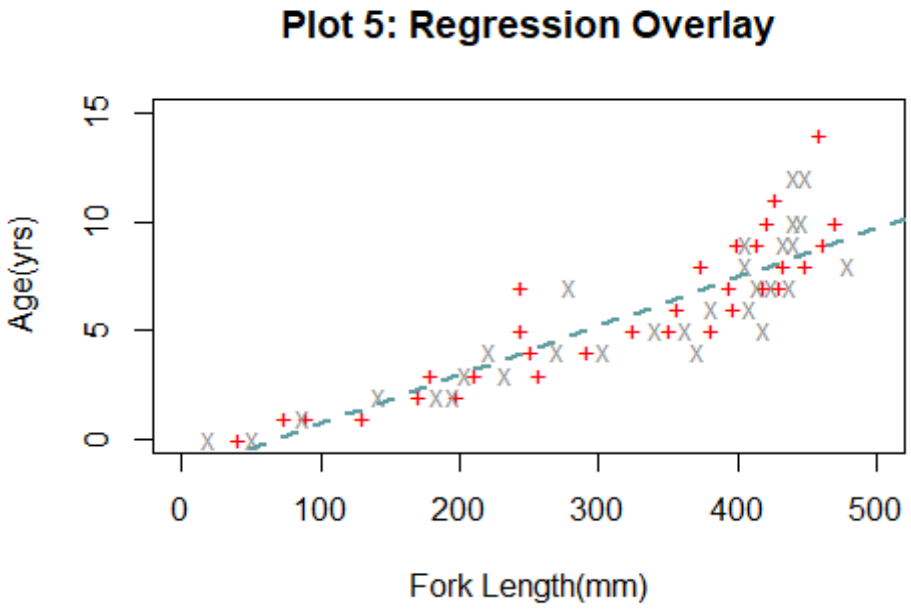
4. Scatterplot with symbols and colours:

- It is similar scatter plot of Age and Fork length of the fish, but it displays points with different point symbols such as '+' and 'x'.
- Also, value points are coloured red (+) and grey (x).

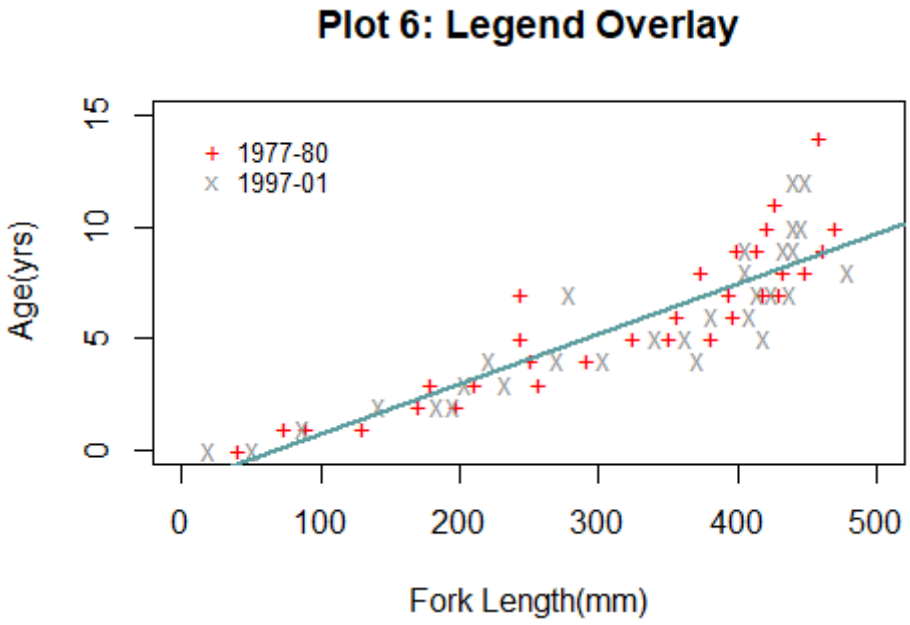


5. Scatter plot with regression line:

- The scatter plot is presented with a regression line which helps to understand the flow of the data.
- In this graph, I have used a dashed line of width 2. There are many types of lines such as 'dotted', 'solid', 'dotdash' and many more.



6. **Scatter plot with legend of levels:**
   - A scatter plot with legends and it helps to simplify and make a graph more understandable.
   - I've created a scatter plot with legend of levels by era.
   - It also has a solid regression line.



B. Summary:

In this report, I've represented descriptive analysis of BullTroutRML2 dataset by finding mean, median, mode, quartiles, variance, standard deviation, skewness, kurtosis and outliers. I've used scatter plot and histogram to understand the distribution of the data values of Age and Fork length. Scatter plots are created with different parameters such as regression line and legend.

Bibliography:

- *Mean median mode*. (n.d.). Tutorialspoint. https://www.tutorialspoint.com/r/r_mean_median_mode.htm
- *Quartile*. (n.d.). Quartiles. http://www.r-tutor.com/elementary-statistics/numerical-measures/quartile
- *Skewness and kutosis*. (n.d.). Skewness and Kurtosis. https://www.geeksforgeeks.org/skewness-and-kurtosis-in-r-programming/
- *Variance and standard deviation*. (n.d.). Variance and Standard Deviation. https://r-coder.com/standard-deviation-variance-r/
- *R in action*. (2011). Manning Publications Co. http://www.cs.uni.edu/~jacobson/4772/week11/R_in_Action.pdf

Appendix:

```
print ("Plotting Basics: Krutarth Jaiswal")

## Plotting Basics: Krutarth Jaiswal

#Install Packages

install.packages(c("plyr", "FSA", "FSAdata", "magrittr", "dplyr", "plotrix
", "ggplot2","moments"))

#Import Libraries
library(FSAdata)

## ## FSAdata v0.3.8. See ?FSAdata to find data for specific fisheries ana
lyses.

library(plyr)
library(FSA)

## ## FSA v0.9.1. See citation('FSA') if used in publication.
## ## Run fishR() for related website and fishR('IFAR') for related book.

##
## Attaching package: 'FSA'

## The following object is masked from 'package:plyr':
##
##     mapvalues

library(magrittr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(plotrix)
library(ggplot2)
library(moments)
```

```r
#Load the BullTroutRML2 dataset
data <- BullTroutRML2
data
```

```
##    age  fl     lake     era
## 1   14 459 Harrison 1977-80
## 2   12 449 Harrison 1977-80
## 3   10 471 Harrison 1977-80
## 4   10 446 Harrison 1977-80
## 5    9 400 Harrison 1977-80
## 6    9 440 Harrison 1977-80
## 7    9 462 Harrison 1977-80
## 8    8 480 Harrison 1977-80
## 9    8 449 Harrison 1977-80
## 10   7 437 Harrison 1977-80
## 11   7 431 Harrison 1977-80
## 12   7 425 Harrison 1977-80
## 13   7 419 Harrison 1977-80
## 14   6 409 Harrison 1977-80
## 15   6 397 Harrison 1977-80
## 16   5 419 Harrison 1977-80
## 17   5 381 Harrison 1977-80
## 18   5 363 Harrison 1977-80
## 19   5 351 Harrison 1977-80
## 20   4 372 Harrison 1977-80
## 21   2 199 Harrison 1977-80
## 22   2 184 Harrison 1977-80
## 23   1  91 Harrison 1977-80
## 24  12 440 Harrison 1997-01
## 25  11 428 Harrison 1997-01
## 26  10 440 Harrison 1997-01
## 27  10 422 Harrison 1997-01
## 28   9 434 Harrison 1997-01
## 29   9 415 Harrison 1997-01
## 30   9 406 Harrison 1997-01
## 31   8 434 Harrison 1997-01
## 32   8 406 Harrison 1997-01
## 33   8 375 Harrison 1997-01
## 34   7 415 Harrison 1997-01
## 35   7 394 Harrison 1997-01
## 36   6 381 Harrison 1997-01
## 37   6 357 Harrison 1997-01
## 38   5 341 Harrison 1997-01
## 39   5 326 Harrison 1997-01
## 40   4 304 Harrison 1997-01
## 41   4 292 Harrison 1997-01
## 42   4 270 Harrison 1997-01
## 43   4 252 Harrison 1997-01
## 44   4 221 Harrison 1997-01
## 45   3 258 Harrison 1997-01
## 46   3 233 Harrison 1997-01
## 47   3 211 Harrison 1997-01
## 48   3 205 Harrison 1997-01
```

```
## 49   3 180 Harrison 1997-01
## 50   2 196 Harrison 1997-01
## 51   2 171 Harrison 1997-01
## 52   2 143 Harrison 1997-01
## 53   1 131 Harrison 1997-01
## 54   1  88 Harrison 1997-01
## 55   1  75 Harrison 1997-01
## 56   0  51 Harrison 1997-01
## 57   0  41 Harrison 1997-01
## 58   0  20 Harrison 1997-01
## 59   7 245 Harrison 1997-01
## 60   7 279 Harrison 1997-01
## 61   5 245 Harrison 1997-01
## 62   8 360   Osprey 1977-80
## 63   8 357   Osprey 1977-80
## 64   7 357   Osprey 1977-80
## 65   7 329   Osprey 1977-80
## 66   6 385   Osprey 1977-80
## 67   6 323   Osprey 1977-80
## 68   5 369   Osprey 1977-80
## 69   5 326   Osprey 1977-80
## 70   4 357   Osprey 1977-80
## 71   4 326   Osprey 1977-80
## 72   4 258   Osprey 1977-80
## 73   4 239   Osprey 1977-80
## 74   3 221   Osprey 1977-80
## 75   3 258   Osprey 1977-80
## 76   3 276   Osprey 1977-80
## 77  11 688   Osprey 1997-01
## 78  10 369   Osprey 1997-01
## 79   9 400   Osprey 1997-01
## 80   8 381   Osprey 1997-01
## 81   8 332   Osprey 1997-01
## 82   7 394   Osprey 1997-01
## 83   7 388   Osprey 1997-01
## 84   7 354   Osprey 1997-01
## 85   7 320   Osprey 1997-01
## 86   6 320   Osprey 1997-01
## 87   6 347   Osprey 1997-01
## 88   6 360   Osprey 1997-01
## 89   5 354   Osprey 1997-01
## 90   5 335   Osprey 1997-01
## 91   5 313   Osprey 1997-01
## 92   5 289   Osprey 1997-01
## 93   4 313   Osprey 1997-01
## 94   4 298   Osprey 1997-01
## 95   3 279   Osprey 1997-01
## 96   3 273   Osprey 1997-01

#Print the first and last 3 records from the dataset
headtail(data,n=3)
```

```
##    age  fl     lake     era
## 1   14 459 Harrison 1977-80
## 2   12 449 Harrison 1977-80
## 3   10 471 Harrison 1977-80
## 94   4 298   Osprey 1997-01
## 95   3 279   Osprey 1997-01
## 96   3 273   Osprey 1997-01
```

```
#Filter out all records except those from Harrison Lake
filtered_data <- filter(data,data["lake"]=="Harrison")
filtered_data
```

```
##    age  fl     lake     era
## 1   14 459 Harrison 1977-80
## 2   12 449 Harrison 1977-80
## 3   10 471 Harrison 1977-80
## 4   10 446 Harrison 1977-80
## 5    9 400 Harrison 1977-80
## 6    9 440 Harrison 1977-80
## 7    9 462 Harrison 1977-80
## 8    8 480 Harrison 1977-80
## 9    8 449 Harrison 1977-80
## 10   7 437 Harrison 1977-80
## 11   7 431 Harrison 1977-80
## 12   7 425 Harrison 1977-80
## 13   7 419 Harrison 1977-80
## 14   6 409 Harrison 1977-80
## 15   6 397 Harrison 1977-80
## 16   5 419 Harrison 1977-80
## 17   5 381 Harrison 1977-80
## 18   5 363 Harrison 1977-80
## 19   5 351 Harrison 1977-80
## 20   4 372 Harrison 1977-80
## 21   2 199 Harrison 1977-80
## 22   2 184 Harrison 1977-80
## 23   1  91 Harrison 1977-80
## 24  12 440 Harrison 1997-01
## 25  11 428 Harrison 1997-01
## 26  10 440 Harrison 1997-01
## 27  10 422 Harrison 1997-01
## 28   9 434 Harrison 1997-01
## 29   9 415 Harrison 1997-01
## 30   9 406 Harrison 1997-01
## 31   8 434 Harrison 1997-01
## 32   8 406 Harrison 1997-01
## 33   8 375 Harrison 1997-01
## 34   7 415 Harrison 1997-01
## 35   7 394 Harrison 1997-01
## 36   6 381 Harrison 1997-01
## 37   6 357 Harrison 1997-01
## 38   5 341 Harrison 1997-01
## 39   5 326 Harrison 1997-01
## 40   4 304 Harrison 1997-01
```

```
## 41   4 292 Harrison 1997-01
## 42   4 270 Harrison 1997-01
## 43   4 252 Harrison 1997-01
## 44   4 221 Harrison 1997-01
## 45   3 258 Harrison 1997-01
## 46   3 233 Harrison 1997-01
## 47   3 211 Harrison 1997-01
## 48   3 205 Harrison 1997-01
## 49   3 180 Harrison 1997-01
## 50   2 196 Harrison 1997-01
## 51   2 171 Harrison 1997-01
## 52   2 143 Harrison 1997-01
## 53   1 131 Harrison 1997-01
## 54   1  88 Harrison 1997-01
## 55   1  75 Harrison 1997-01
## 56   0  51 Harrison 1997-01
## 57   0  41 Harrison 1997-01
## 58   0  20 Harrison 1997-01
## 59   7 245 Harrison 1997-01
## 60   7 279 Harrison 1997-01
## 61   5 245 Harrison 1997-01

#Display the first and last 3 records from the filtered dataset
head(filtered_data,3)

##   age  fl     lake     era
## 1  14 459 Harrison 1977-80
## 2  12 449 Harrison 1977-80
## 3  10 471 Harrison 1977-80

tail(filtered_data,3)

##    age  fl     lake     era
## 59   7 245 Harrison 1997-01
## 60   7 279 Harrison 1997-01
## 61   5 245 Harrison 1997-01

#Display the structure of the filtered dataset
str(filtered_data)

## 'data.frame':    61 obs. of  4 variables:
##  $ age : int  14 12 10 10 9 9 9 8 8 7 ...
##  $ fl  : int  459 449 471 446 400 440 462 480 449 437 ...
##  $ lake: Factor w/ 2 levels "Harrison","Osprey": 1 1 1 1 1 1 1 1 1 1 ..
.
##  $ era : Factor w/ 2 levels "1977-80","1997-01": 1 1 1 1 1 1 1 1 1 1 ..
.

# Display the summary of the filtered dataset and save it as <t>
t <- summary(filtered_data)
t

##       age               fl            lake         era
##  Min.   : 0.000   Min.   : 20   Harrison:61   1977-80:23
```

```
##  1st Qu.: 3.000   1st Qu.:221   Osprey  : 0   1997-01:38
##  Median : 6.000   Median :372
##  Mean   : 5.754   Mean   :319
##  3rd Qu.: 8.000   3rd Qu.:425
##  Max.   :14.000   Max.   :480

#Create a new object called “tmp” that includes the first 3 and last 3 rec
ords of the whole data set
tmp <- headtail(data,n=3L)
tmp

##    age  fl     lake     era
## 1   14 459 Harrison 1977-80
## 2   12 449 Harrison 1977-80
## 3   10 471 Harrison 1977-80
## 94   4 298   Osprey 1997-01
## 95   3 279   Osprey 1997-01
## 96   3 273   Osprey 1997-01

#Display the “era” column in the new “tmp” object
tmp["era"]

##        era
## 1  1977-80
## 2  1977-80
## 3  1977-80
## 94 1997-01
## 95 1997-01
## 96 1997-01

#Create a pchs vector with the argument values for + and x. Then create a
cols vector with the two elements “red” and “gray60”
pchs <- c("+","x")
pchs

## [1] "+" "x"

cols <- c("red","gray60")

#Convert the tmp object values to numeric values. Then create a numeric nu
mEra object from the tmp$era object
numEra <- as.numeric(tmp$era)
numEra

## [1] 1 1 1 2 2 2

#Associate the cols vector with the tmp era values
cols[tmp$era]

## [1] "red"    "red"    "red"    "gray60" "gray60" "gray60"
```