N Northeastern University

Toronto, Canada

A Report on

Executive Summary of Module 2

Introduction to Data Analytics

(ALY 6000)

Guided by:

Prof. Mohammad Shafiqul Islam

Submitted By:

Name of Student	NUID	Date of submission
Parth Shah	002956963	28 January, 2021

<u>INDEX</u>

1. <u>Introduction</u>

This report describes the statistics of BullTroutRML2 dataset in which there are two lakes Harrison Lake and Osprey and display the visualizations of the data.

2. Provide an analysis of descriptive characteristics of the data set provided by your instructor. This includes pertinent statistics including mean, median, quartiles, variance, standard deviation, skew, kurtosis, outliers etc. Include R console screen snipet to support your observations and conclusions. Below is a sample excerpt of an analysis of Harrison Lake fish from the BullTroutRML2 dataset.

After loading required packages through pacman that is package manager, I load BullTroutRML2, which have 96 results and there are different categories such as age, fl, lake and era.

By using head, I had display first 3 records of BullTroutRML2. Result shown below: -

```
#First 3 records
head(BullTroutRML2, n = 3)
##   age fl  lake      era
## 1 14 459 Harrison 1977-80
## 2 12 449 Harrison 1977-80
## 3 10 471 Harrison 1977-80
```

Similarly, by using tail, I had display last 3 records of given dataset. Result is as follow: -

```
#last 3 records
tail(BullTroutRML2, n = 3)
##   age fl   lake    era
## 94 4  298 Osprey 1997-01
## 95 3  279 Osprey 1997-01
## 96 3  273 Osprey 1997-01
```

After that I did remove all elements except Harrison Lake through following commands: -

```
HarrisonLake <- filter(BullTroutRML2, lake == "Harrison")
HarrisonLake
```

Here I had created new function HarrisonLake and I had display first 5 and last 5 records of HarrisonLake.

```
#Display first 5 records
head(HarrisonLake, n = 5)
```

```
#Display last 5 records
tail(HarrisonLake, n = 5)
```

By using Structure() function and Summary() function, I got all the data such as mean, median, minimum, maximum, quartiles.

```
#structure of Harrison lake
structure(HarrisonLake)
```

```
#summary of Harrison lake
summary(HarrisonLake)
##       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
```

3. Provide the executive with visualizations (at least 6) in that help them see the key characteristics you want to highlight. They can be boxplots, histograms, frequency and probability distributions, bar plots (bar charts) or pareto. Not only is the goal to present your visual results, but also to explain the significance of what the visuals are displaying.
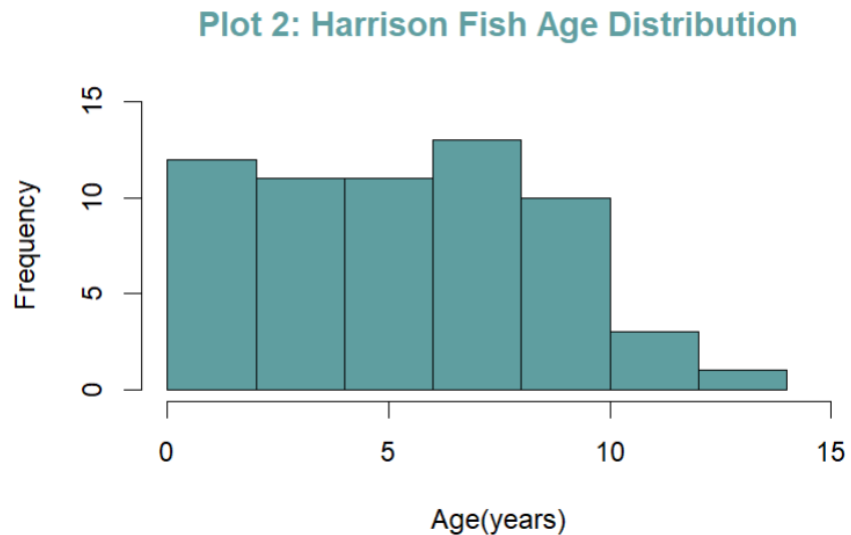
PLOT 1: - Harrison Lake Trout
The scatter plot shows that the graph of age (years) vs fork length (mm): -

Plot 1: Harrison Lake Trout
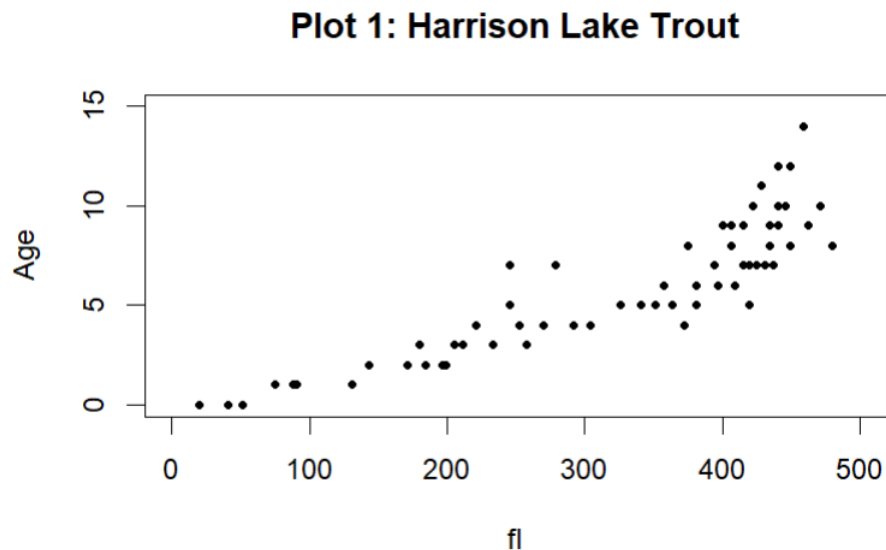
PLOT 2: - <u>Harrison Fish Age Distribution</u>

The histogram reveals that the graph of fish age distribution with cadet blue color and age (years) vs fork length (mm): -

Plot 2: Harrison Fish Age Distribution

PLOT 3: - <u>Harrison Density Shaded by Era</u>

This graph is similar with graph one, but the change is in color. There are two shades of green which represent different eras.

Plot 3: Harrison Density Shaded by era
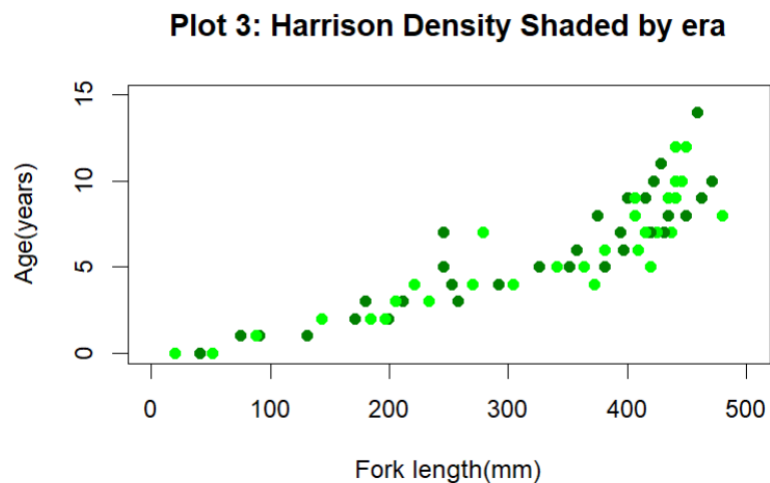
PLOT 4: - <u>Symbol & Color by Era</u>

    This plot displays the age vs fl with + and x signs with red and gray60 colors of different eras.

Plot 4:Symbol and Color by Era
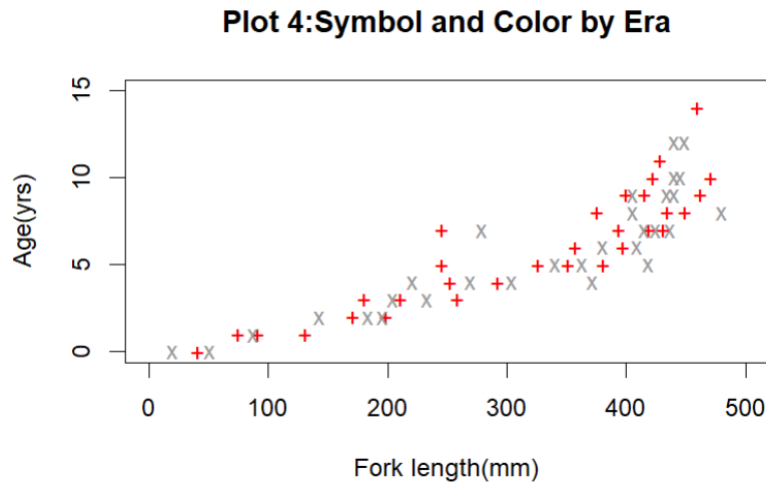
PLOT 5: - <u>Regression Overlay</u>

    This plot shows that the relation of age vs fl which is same as PLOT 4 with a regression line which shows the trend of the data.

Plot 5: Regression Overlay

PLOT 6: - <u>Legend Overlay</u>

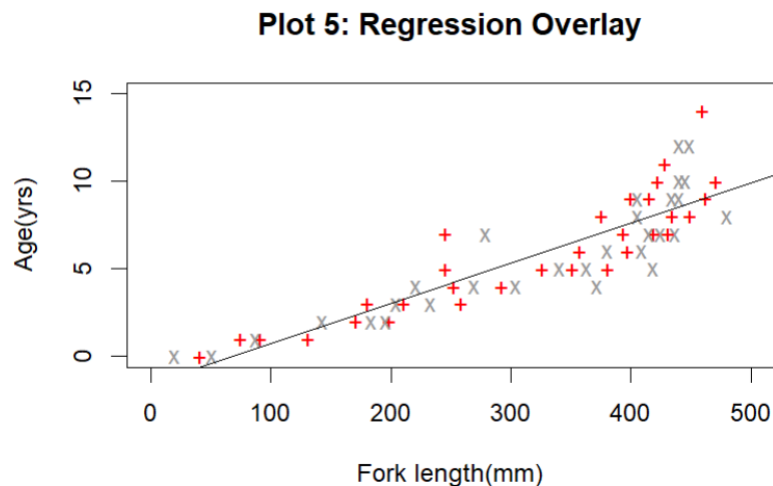   This plot displays same as PLOT 5 with left hand side upper corner data of eras, revealing that 1997-80 with red + sign and 1997-01 with gray60 and x sign.

Plot 6: Legend overlay
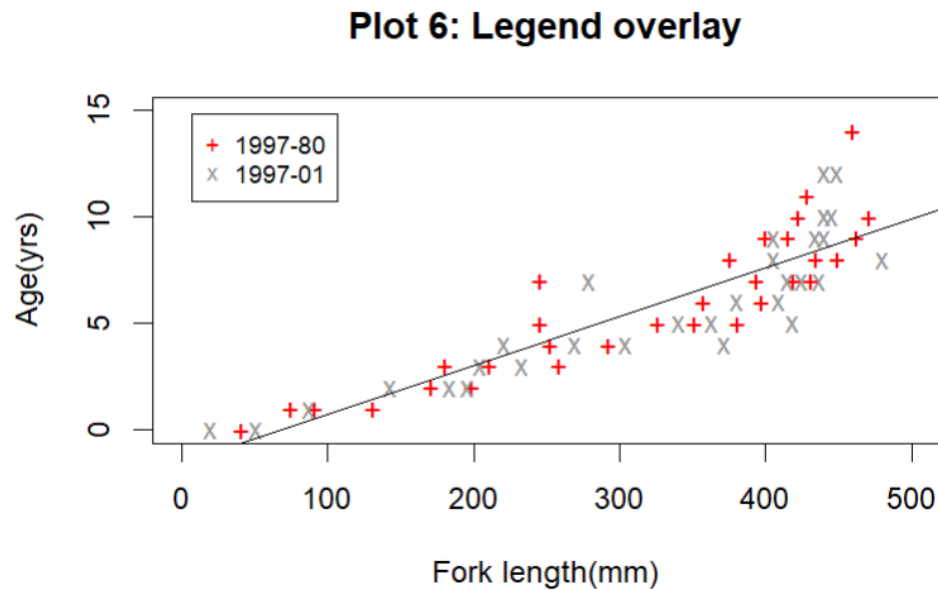
4. Summary

To summarize, I learned how to load the datasets, finding their structures and summary, finding head and tails of given data, remove an element and how to display the data via plot through plot, scatter plot and histogram by adding regression line and legend overlay to clarify the data with values (eras).

5. <u>References</u>

- Michael Toth(2019). How to filter data in R from R bloggers website named https://www.r-bloggers.com/2019/04/how-to-filter-in-r-a-detailed-introduction-to-the-dplyr-filter-function/

- Tutorials Point(2021). Scatterplots in R from Tutorial Point website named https://www.tutorialspoint.com/r/r_scatterplots.htm

- Venables, W. N. and Ripley. B. D. (2002) Modern Applied Statistics with S. Springer. Histograms in R from RDocumentation website named https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/hist

- Knitr and Jekyll. Changing types of vector in R from cookbook website named http://www.cookbook-r.com/Manipulating_data/Converting_between_vector_types/

- Codersgram9(2021). How to add legend in R from geeksforgeeks website named https://www.geeksforgeeks.org/add-legend-to-plot-in-r/

6. Appendix

   Github Link: - https://github.com/iparth0611/Module2.git

Assignment-2.R

prbsh

2022-01-27

```r
#Plotting last name
print("Plotting Basics: Shah")
```

```
## [1] "Plotting Basics: Shah"
```

```r
#install packages
pacman::p_load(FSA, FSAdata, magrittr,  dplyr, plotrix, ggplot2, moments)

#load data
data("BullTroutRML2")
BullTroutRML2
```

```
##    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
```

```r
#First 3 records
head(BullTroutRML2, n = 3)
```

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

```r
#last 3 records
tail(BullTroutRML2, n = 3)
```

```
##    age  fl   lake     era
## 94   4 298 Osprey 1997-01
## 95   3 279 Osprey 1997-01
## 96   3 273 Osprey 1997-01
```

```r
#remove all records exccept Harrison lake
HarrisonLake <- filter(BullTroutRML2, lake == "Harrison")
HarrisonLake
```

```
##    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
```

```r
#Display first 5 records
head(HarrisonLake, n = 5)
```

```
##   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
```

```
#Display last 5 records
tail(HarrisonLake, n = 5)
```

```
##    age  fl    lake     era
## 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
```

```
#structure of Harrison lake
structure(HarrisonLake)
```

```
##    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
```

```r
#summary of Harrison lake
summary(HarrisonLake)
```

```
##       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
```

```r
#Scatterplot of Harrison lake
fl <- HarrisonLake$fl
age <- HarrisonLake$age
#plot(age ~ fl)
plot(age ~ fl,
     data = HarrisonLake,
     xlim = c(0,500),
     ylim = c(0,15),
     main = "Plot 1: Harrison Lake Trout",
     xlab = "Fork Length (mm)",
     ylab = "Age (yrs)",
     pch = 20)
```
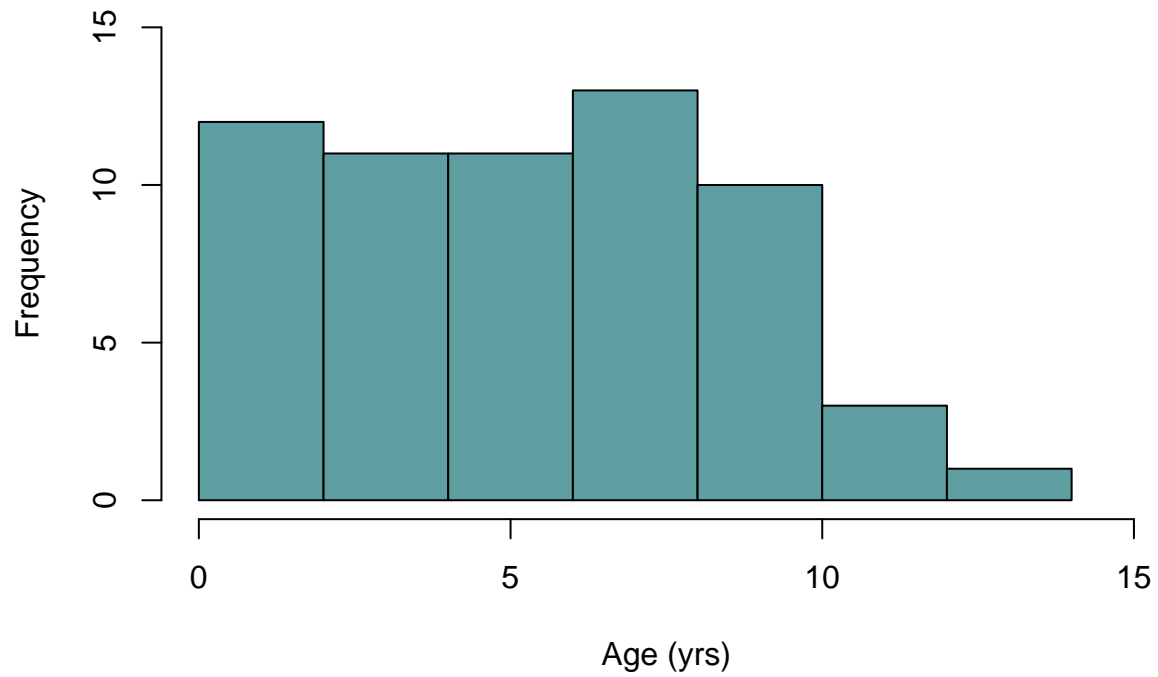
## Plot 1: Harrison Lake Trout
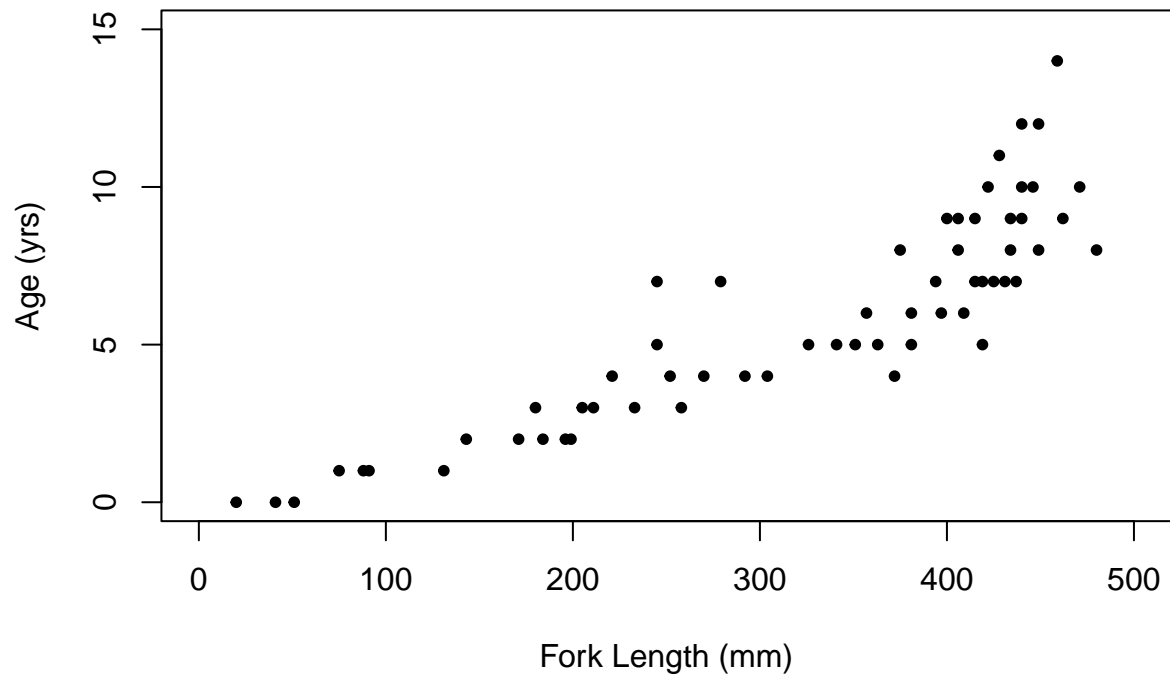

```
#Histogram of Harrison lake
hist(HarrisonLake$age,
  xlab = "Age (yrs)",
  ylab = "Frequency",
  main = ("Plot 2: Harrison Fish Age Distribution"),
  xlim = c(0,15),
  ylim = c(0,15),
  col = "cadetblue",
  col.main = "cadetblue")
```

Plot 2: Harrison Fish Age Distribution


```
#Harrison Density Shaded by Era
plot(age ~ fl,
    main = "Plot 3: Harrison Density Shaded by Era",
    xlab = "Fork length (mm)",
    ylab = "Age (yrs)",
    xlim = c(0,500),
    ylim = c(0,15),
    pch = 16,
    col = rgb(0,(1:2)/2,0))
```

Plot 3: Harrison Density Shaded by Era
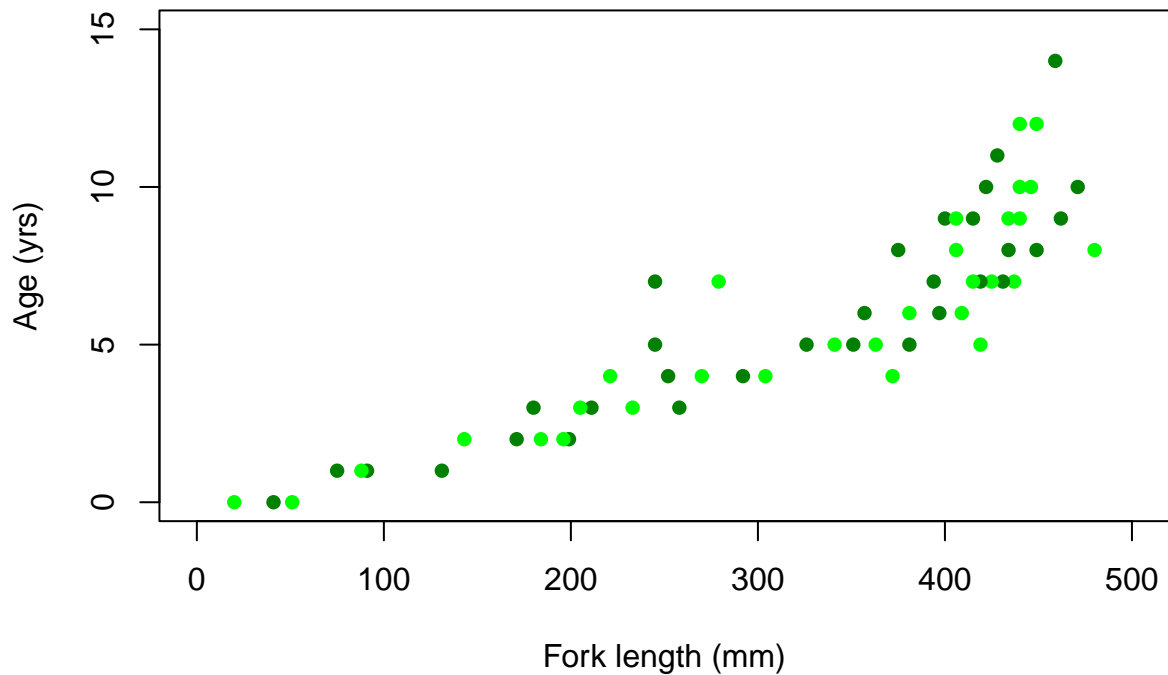

```r
#Create object "tmp" and display first 3 and last 3 records
tmp <- headtail(HarrisonLake, n = 3)
tmp
```

```
##    age  fl     lake      era
## 1   14 459 Harrison 1977-80
## 2   12 449 Harrison 1977-80
## 3   10 471 Harrison 1977-80
## 59   7 245 Harrison 1997-01
## 60   7 279 Harrison 1997-01
## 61   5 245 Harrison 1997-01
```

```r
#Display era in tmp
tmp$era
```

```
## [1] 1977-80 1977-80 1977-80 1997-01 1997-01 1997-01
## Levels: 1977-80 1997-01
```

```r
#pchs vector for + and X
pchs <- c("+", "x")
pchs
```

```
## [1] "+" "x"
```

```
#create cols vector with red and grey60
cols <- c("red", "gray60")
cols
```

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

```
#convert tmp into numeric values
tmp$era <- as.numeric(tmp$era)
tmp$era
```

```
## [1] 1 1 1 2 2 2
```

```
is.numeric(tmp$era)
```
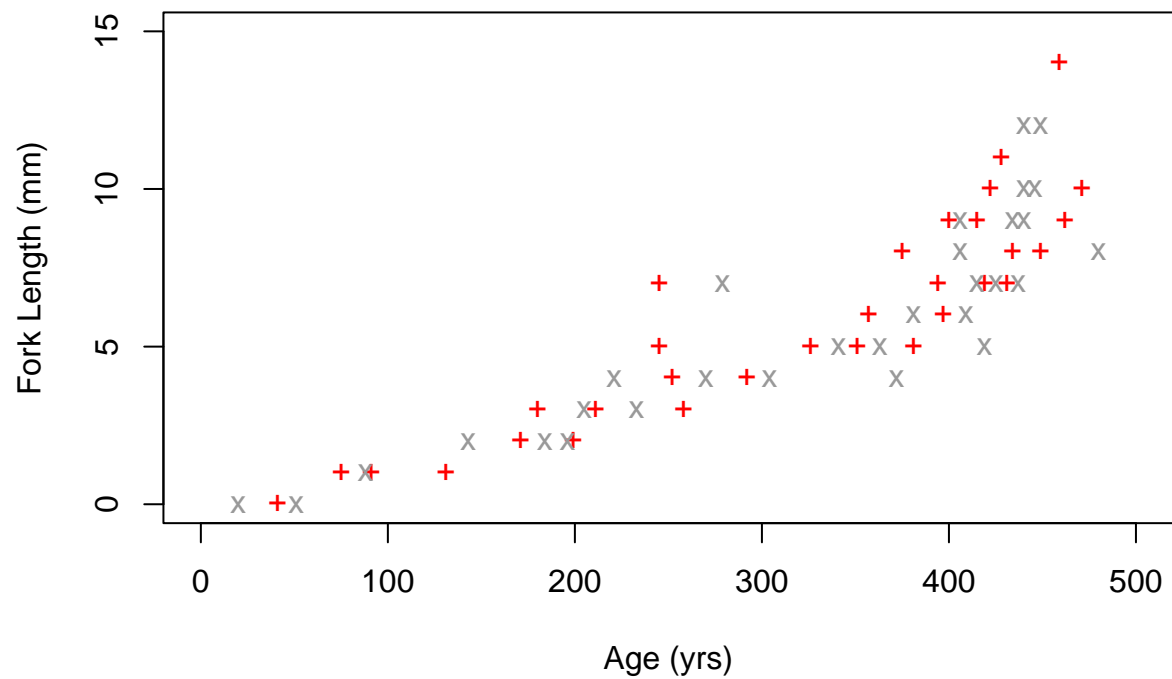
```
## [1] TRUE
```

```
#initialize cols with tmp era values
cols[tmp$era]
```

```
## [1] "red"    "red"    "red"    "gray60" "gray60" "gray60"
```
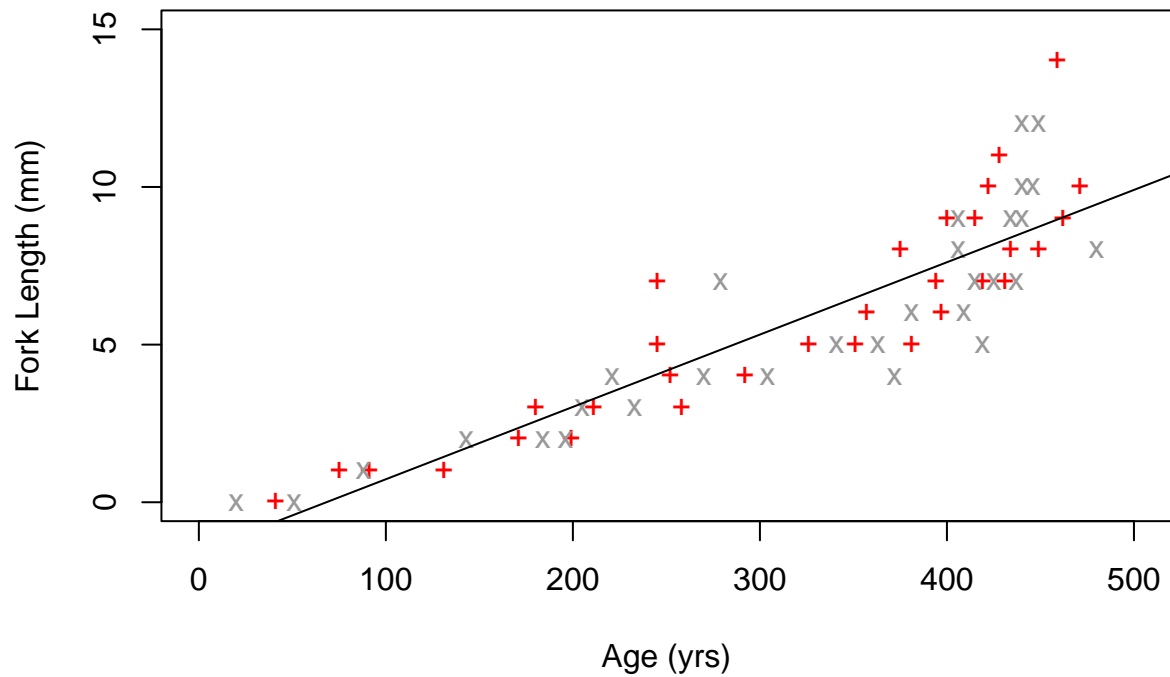
```
#Symbol & Color by Era
plot(age ~ fl,
     data = HarrisonLake,
     main = "Plot 4: Symbol & Color by Era",
     xlim = c(0,500),
     ylim = c(0,15),
     xlab = "Age (yrs)",
     ylab = "Fork Length (mm)",
     pch = pchs,
     col = cols)
```

Plot 4: Symbol & Color by Era


```
#Regression Overlay
plot(age ~ fl,
     data = HarrisonLake,
     main = "Plot 5: Regression Overlay",
     xlim = c(0,500),
     ylim = c(0,15),
     xlab = "Age (yrs)",
     ylab = "Fork Length (mm)",
     pch = pchs,
     col = cols)
abline(lm(age ~ fl, data = HarrisonLake))
```

Plot 5: Regression Overlay


```
#Legend Overlay
plot(age ~ fl,
     data = HarrisonLake,
     main = "Plot 6: :Legend Overlay",
     xlim = c(0,500),
     ylim = c(0,15),
     xlab = "Age (yrs)",
     ylab = "Fork Length (mm)",
     pch = pchs,
     col = cols)
abline(lm(age ~ fl, data = HarrisonLake))
legend("topleft", inset = 0.05,
       legend = c("1997-80","1997-01"),
       bty = "1",
       cex = 0.8,
       pch = pchs,
       col = cols)
```

Plot 6: :Legend Overlay