ALY6000 Introduction to Analytics

Northeastern University

Module 2 Project – Executive Summary Report 2

Date: 04/24/2023

Submitted To: Richard He

Submitted By: Shree Vipulbhai Tejani

Analysis

This dataset 'BullTroutRML2' is imported from the FSA library. It contains 4 columns and 96 rows of data. The dataset is about fish information which includes the age of the fish, fish length, location of where it was found and year.

The key findings for the given set of instructions are as follows:

Importing the libraries required to support various functions required to import dataset, plot visualizations and more.

```
R 4.3.0 · ~/
> print("Plotting Basics: Shree Tejani")  # print author name
[1] "Plotting Basics: Shree Tejani"
> # use installed libraries
> library(FSA)
```

```
> library(FSAdata)
## FSAdata v0.4.0. See ?FSAdata to find data for specific fisheries analyses.
> library(magrittr)
> library(dplyr)
```

```
> library(plotrix)
> library(ggplot2)
> library(moments)
> library(Hmisc)
```

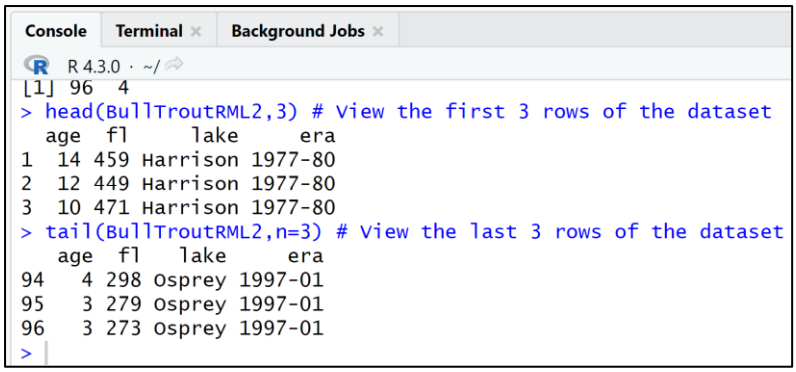
Imported the BullTroutRML2 dataset from FSA library

```
Console  Terminal  Background Jobs
R 4.3.0 · ~/
> #load the required dataset in df
> library(FSA)
> data(BullTroutRML2)  # import BullTroutRML2 dataset
> 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
```

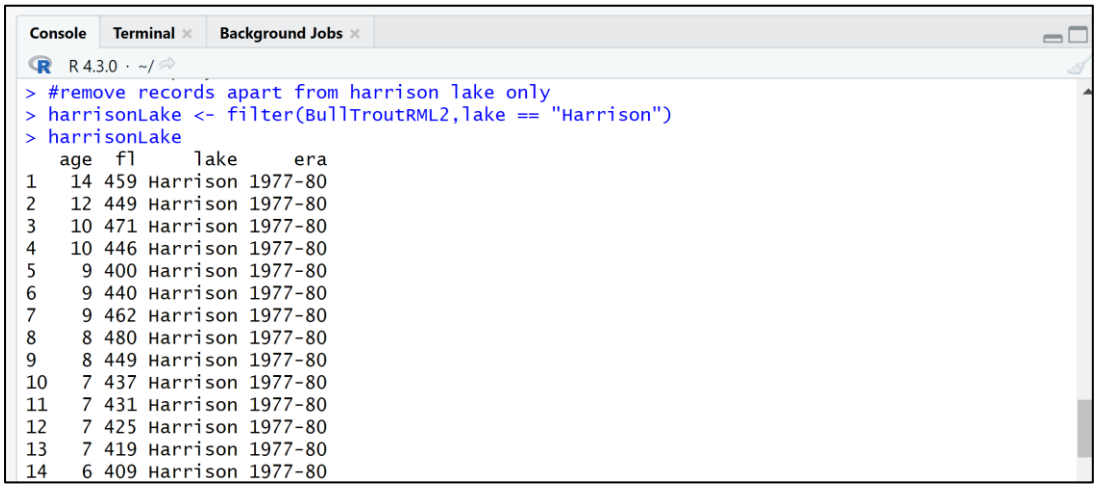
The structure for the BullTroutRML2 dataset displays the different variables and its datatype information.

```
Console  Terminal  Background Jobs
R 4.3.0 · ~/
96   3 273   Osprey 1997-01
> str(BullTroutRML2)  # view structure of the BullTroutRML2 dataset
'data.frame':   96 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 ...
> dim(BullTroutRML2)  # view number of rows and cols of BullTroutRML2 dataset
[1] 96  4
```

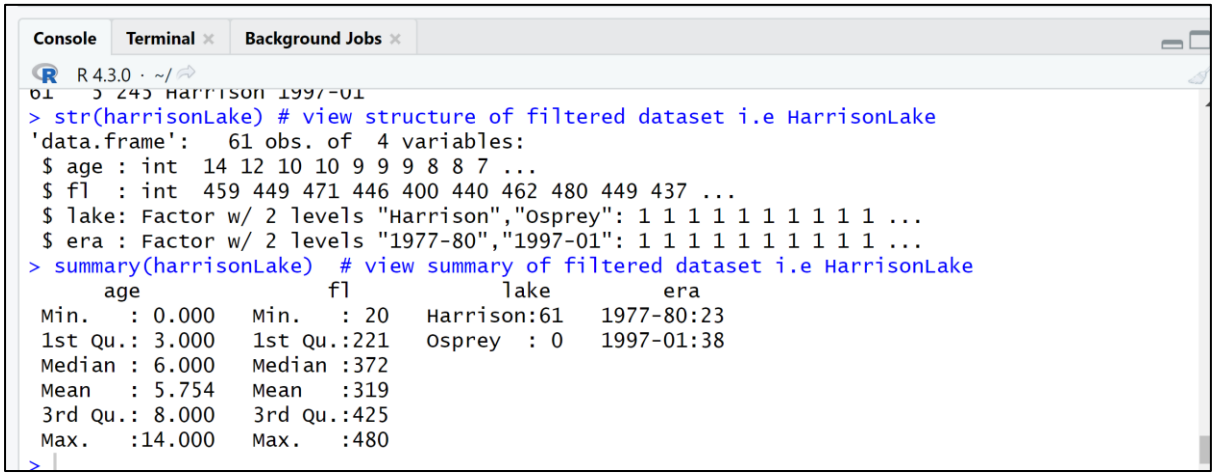
The below output displays the first 3 and last 3 records of the BullTroutRML2 dataset. I've used head() and tail() function for this.



The below output shows filtered data for just Harrison lake. I've used filter() function for it.



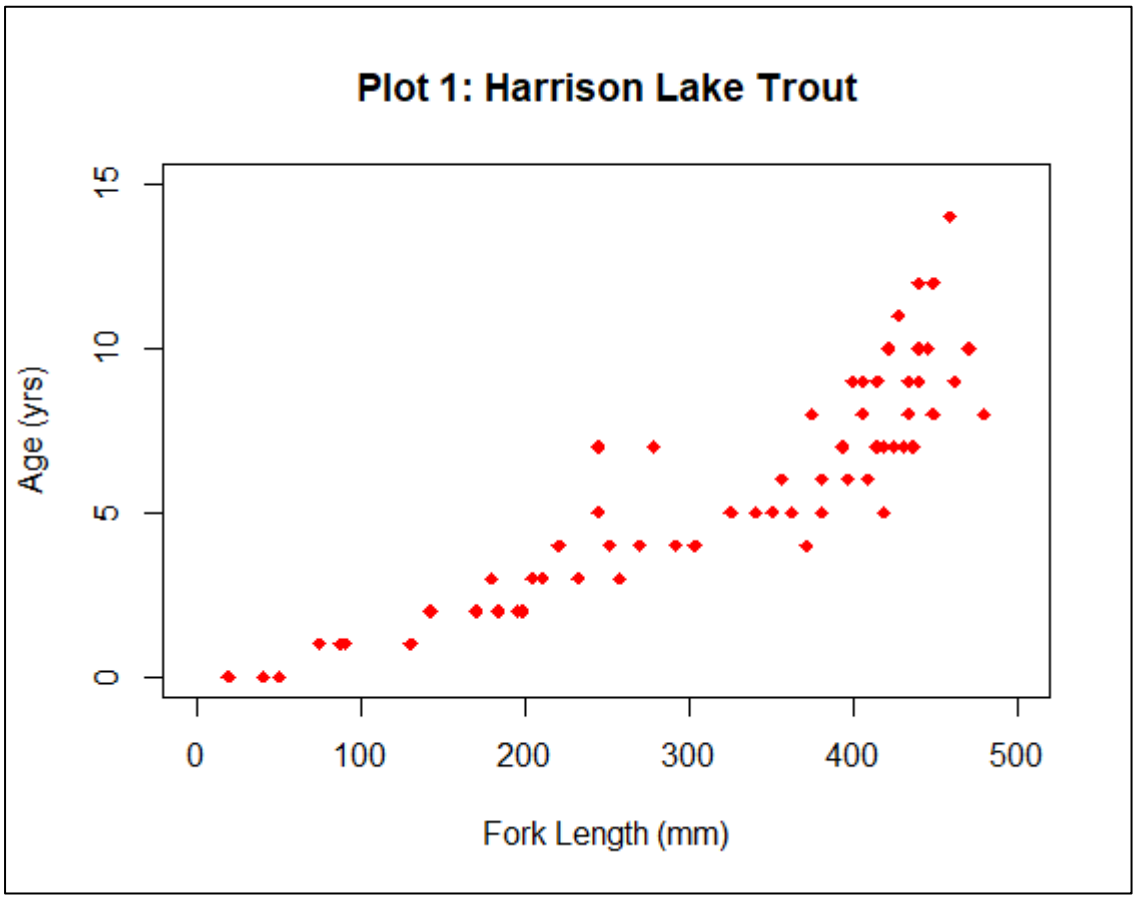
This output showcases Structure and Summary details of the Harrison lake filtered data. The structure of the data would remain the same as we have just filtered some rows and haven't changed any variables or inputted data. While the Summary function displays the various measures of central tendency like mean, median, mode, quartiles, etc.



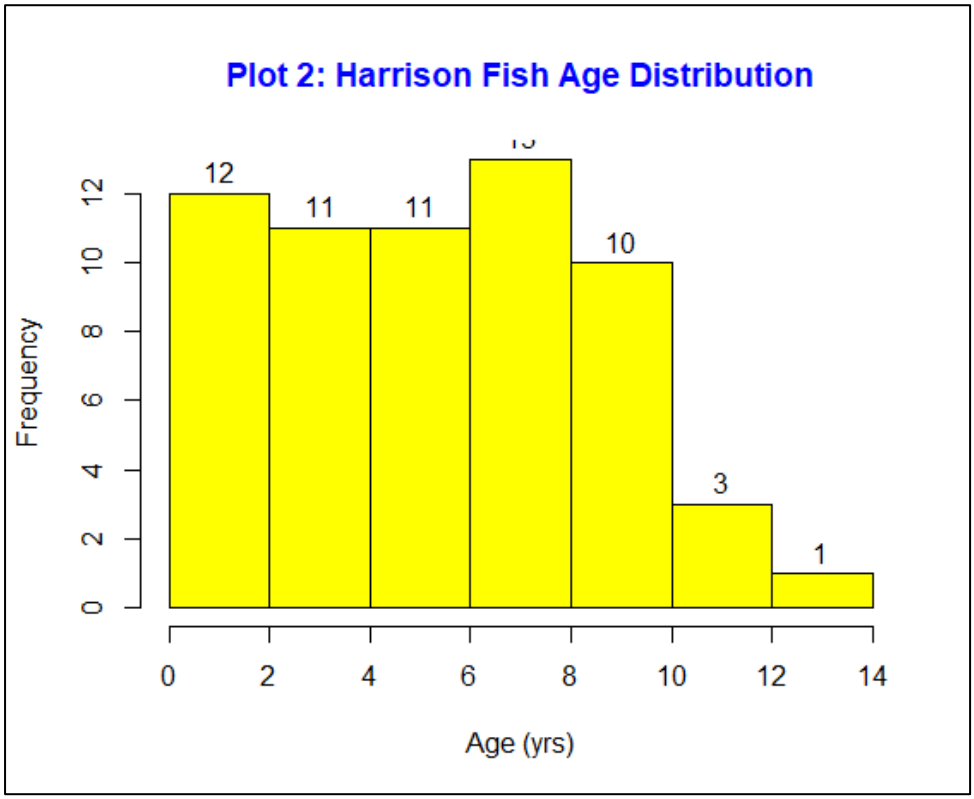
The below output displays the first 5 and last 5 records of the Harrison Lake dataset. I've used head() and tail() function for this.

```
Console   Terminal   Background Jobs
R 4.3.0 · ~/
 Max.   :14.000   Max.   :480
> head(harrisonLake,5) # View the first 5 rows of the filtered dataset i.e Harrison lake
  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
> tail(harrisonLake,n=5) # View the last 5 rows of the filtered dataset i.e Harrison lake
   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
>
```

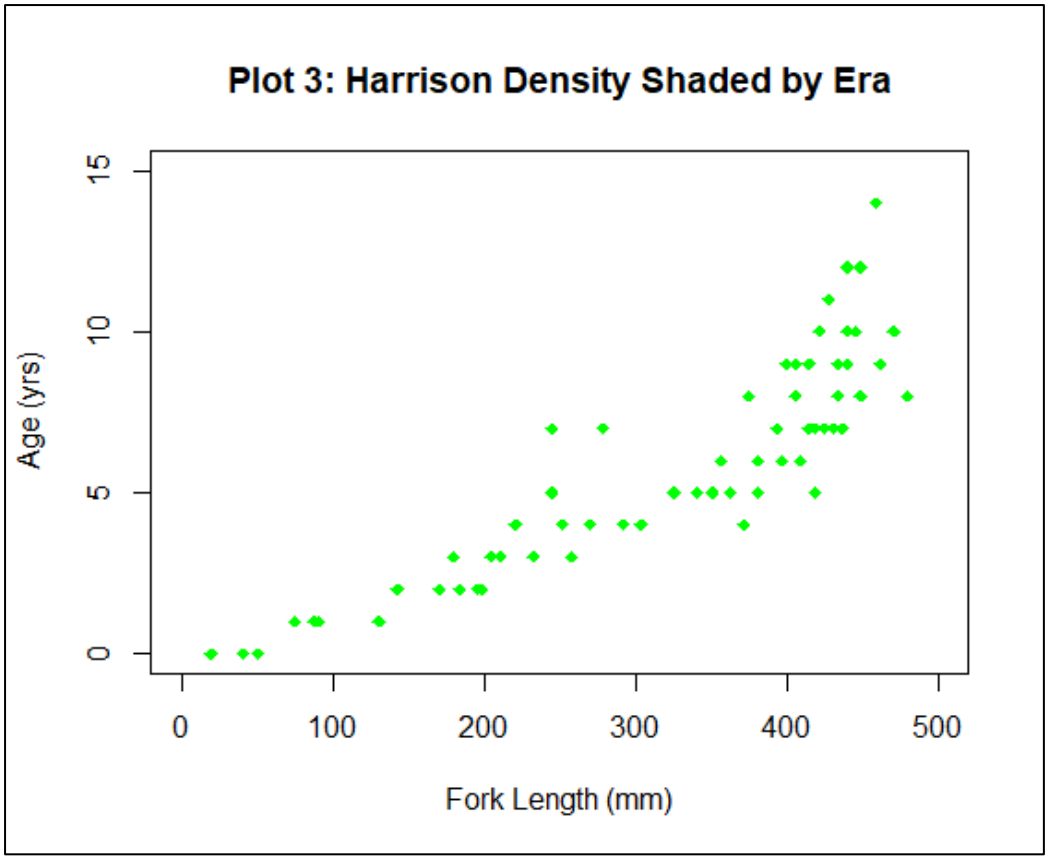
The graph below describes Fork length of Fish to its Age.



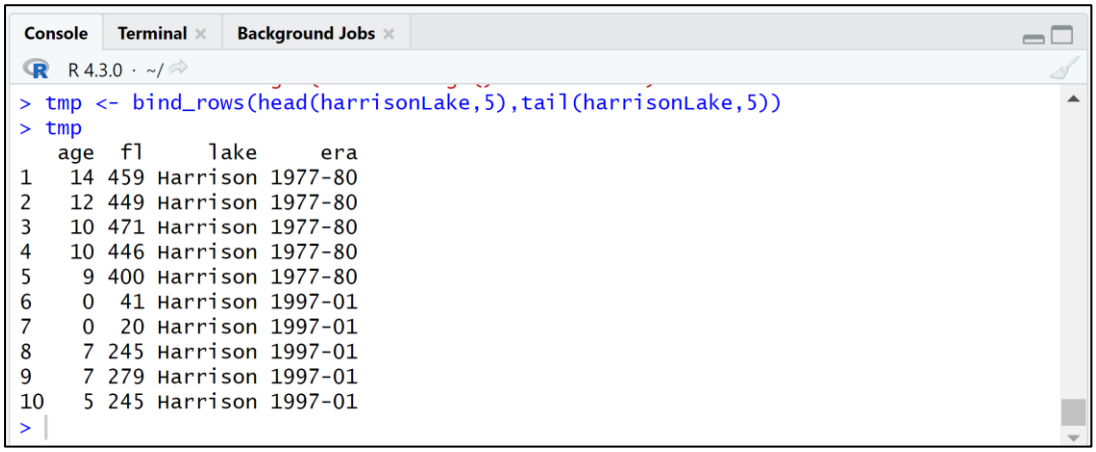
The graph below describes Fish distribution based on their age.



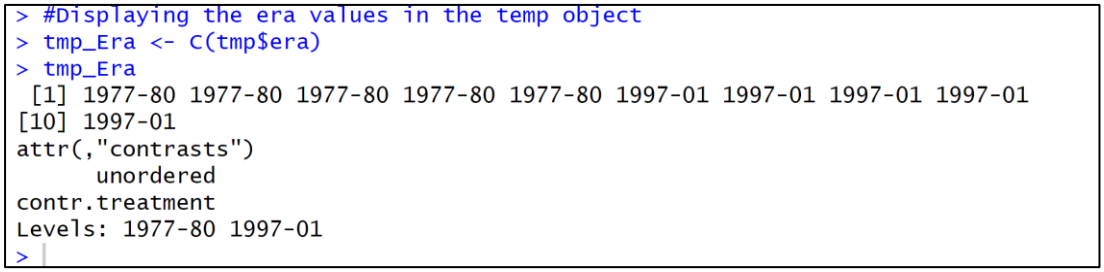
The graph below describes Fork length of Fish to its Age using density shaded by era.



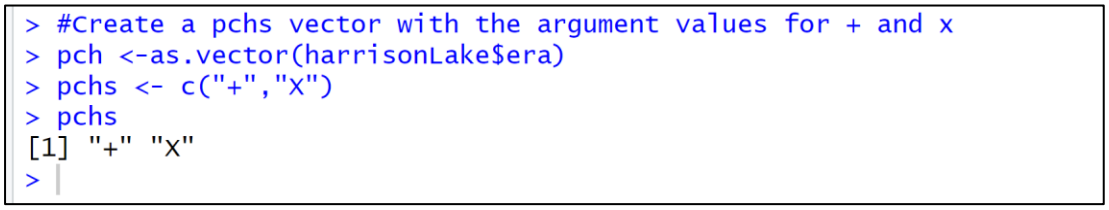
Created a new object ‘tmp’ to store first 5 and last 5 records of Harrison lake dataset.



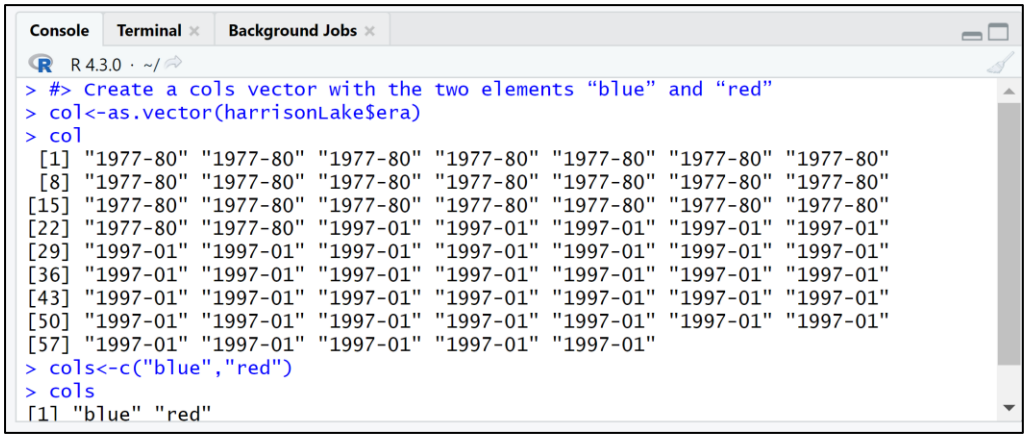
Displaying ‘era’ variable in the above ‘tmp’ object.



Created pchs vector for values ‘+’ and ‘X’.



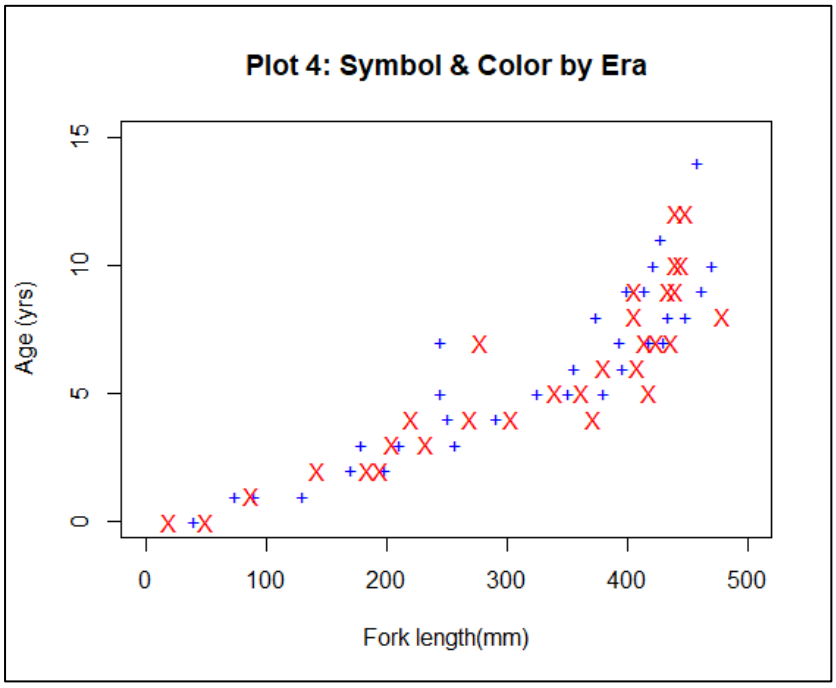
Created cols vector for elements ‘blue’ and ‘red’.



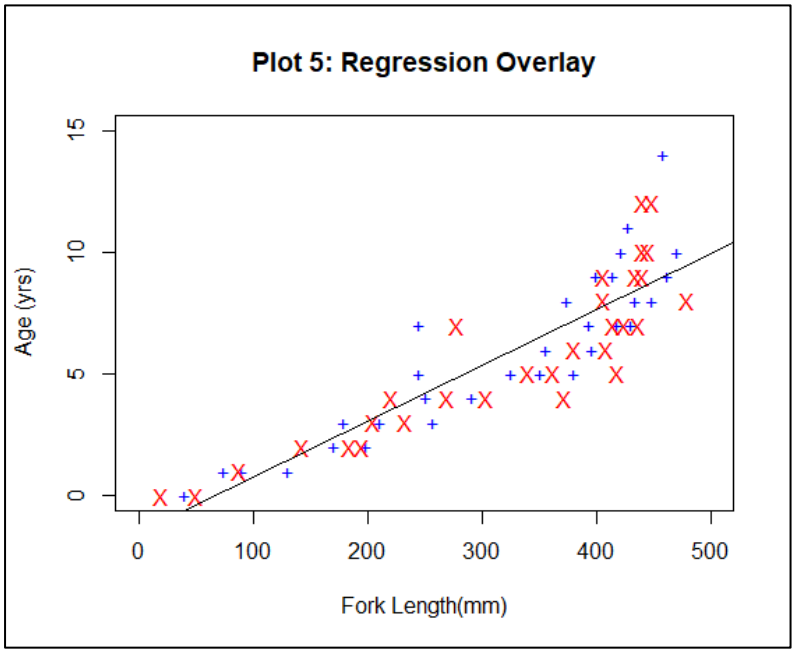
Converted tmp$era values to numeric values.

```
>
> #convert temp$era into numeric values
> tmp$era = as.numeric(tmp_Era)
> tmp$era
 [1] 1 1 1 1 1 2 2 2 2 2
>
```

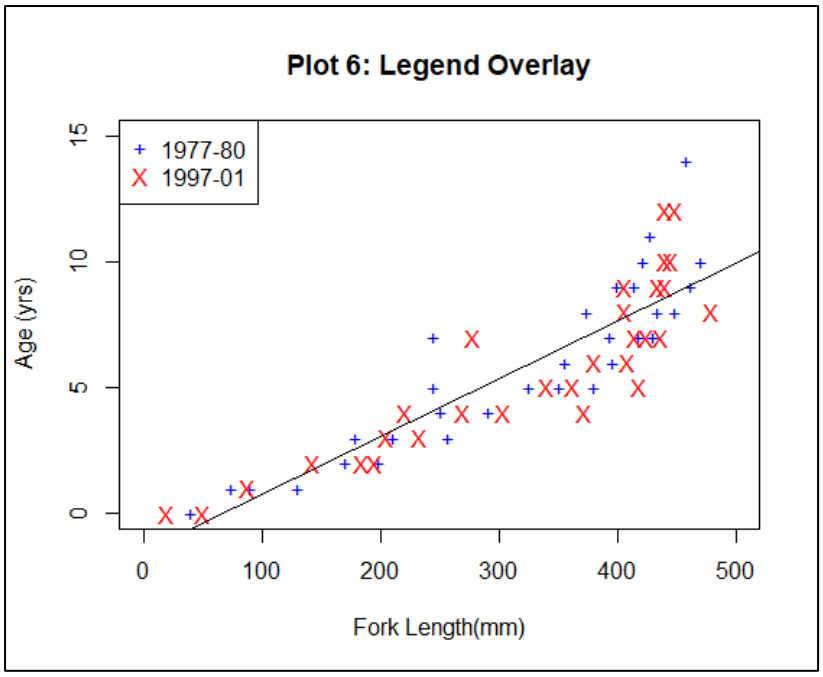
The graph below describes Fork length of a fish by age based on era data.



The graph below describes Fork length of a fish by age based on era data using a regression line.



The graph below describes Fork length of a fish by age based on era data using a regression line and placing a legend to help for better understanding.



Summary

The above dataset of the BullTroutRML2 file allows us to perform various functions to derive statistical information to help in further analysis. Due to the dataset being a discrete numerical data, generating statistical output was easy to interpret. I have supported this analysis by outputting the mean values, filtering the data and creating a scatter plot for the same.

I have added regression for the scatter plot which is derived after creating tmp object. The analysis displays results of various mathematical function like Min, Max, Mean, Median value by using the summary() function for the original BullTroutRML2 dataset. The above data was very limited and accurate to perform all kinds of analysis, which made it easier to create various kinds of graphical visualizations.

Bibliograpghy

Kabacoff, R. I. (2015). *R in action: Data analysis and graphics with R*. Manning.

Appendix

```
print("Plotting Basics: Shree Tejani")  # print author name

#importing the libraries

install.packages("FSA")

install.packages("FSAdata")

install.packages("magrittr")

install.packages("plotrix")

install.packages("ggplot2")

install.packages("moments")

# use installed libraries

library(FSA)

library(FSAdata)

library(magrittr)

library(dplyr)

library(plotrix)

library(ggplot2)

library(moments)

library(Hmisc)



#load the required dataset in df

library(FSA)

data(BullTroutRML2)  # import BullTroutRML2 dataset

BullTroutRML2

str(BullTroutRML2)  # view structure of the BullTroutRML2 dataset
```

```
dim(BullTroutRML2)  # view number of rows and cols of BullTroutRML2 dataset


head(BullTroutRML2,3) # View the first 3 rows of the dataset
tail(BullTroutRML2,n=3) # View the last 3 rows of the dataset


#remove records apart from harrison lake only
harrisonLake <- filter(BullTroutRML2,lake == "Harrison")
harrisonLake
str(harrisonLake) # view structure of filtered dataset i.e HarrisonLake
summary(harrisonLake)  # view summary of filtered dataset i.e HarrisonLake


head(harrisonLake,5) # View the first 5 rows of the filtered dataset i.e Harrison lake
tail(harrisonLake,n=5) # View the last 5 rows of the filtered dataset i.e Harrison lake


# create scatter plot for age and fl for harrison lake dataset
plot(harrisonLake$fl, harrisonLake$age,
     xlim = c(0,500), ylim = c(0,15),
     xlab = "Fork Length (mm)", ylab = "Age (yrs)",
     pch=18, col="red",
     main = "Plot 1: Harrison Lake Trout")


# create histogram for age and frequency for harrison lake dataset
hist(harrisonLake$age, main ="Plot 2: Harrison Fish Age Distribution",
    xlab="Age (yrs)", ylab="Frequency",col="yellow", col.main="blue",labels = TRUE)
```

```
# create scatter Plot 3: Harrison Density Shaded by Era

plot(data = harrisonLake, age~fl, main ="Plot 3: Harrison Density Shaded by Era", xlim
=c(0,500),ylim= c(0,15),

   xlab="Fork Length (mm)", ylab="Age (yrs)", pch=18,

   colramp = colorRampPalette(c('lightgreen','white')),col = 'green')


#Entering the first and the last five records of the BULLTORNT data in the new object "tmp"

tmp <- bind_rows(head(harrisonLake,5),tail(harrisonLake,5))

tmp


#Displaying the era values in the temp object

tmp_Era <- C(tmp$era)

tmp_Era


#Create a pchs vector with the argument values for + and x

pch <-as.vector(harrisonLake$era)

pchs <- c("+","X")

pchs


#> Create a cols vector with the two elements “blue” and “red”

col<-as.vector(harrisonLake$era)

col

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

cols


#convert temp$era into numeric values
```

```
tmp$era = as.numeric(tmp_Era)

tmp$era



#intialize cols vector with temp era values

#plot 4: Symbol & Color by Era

plot(data=harrisonLake,age~fl, xlab="Fork length(mm)", ylab ="Age (yrs)",xlim = c(0,500),
ylim = c(0,15),

    pch=pchs, col=cols, main="Plot 4: Symbol & Color by Era")



# regression overlay

plot(data=harrisonLake,age~fl, xlab="Fork Length(mm)", ylab = "Age (yrs)", xlim =
c(0,500), ylim = c(0,15),

    pch=pchs,col=cols, main="Plot 5: Regression Overlay")

abline(lm(age~fl,data = harrisonLake))



# legend

plot(data=harrisonLake,age~fl, xlab="Fork Length(mm)", ylab = "Age (yrs)", xlim =
c(0,500), ylim = c(0,15),

    pch=pchs,col=cols, main="Plot 6: Legend Overlay")

legend(x="topleft", legend = paste(levels(harrisonLake$era)),pch = pchs,col=cols)

abline(lm(age~fl,data = harrisonLake))
```