College of Professional Studies

Northeastern University San Jose

MPS Analytics

Course: ALY6000 - Introduction to Data Analytics

Assignment:

MODULE PROJECT - 2

EXECUTIVE SUMMARY REPORT – 2

Submitted on:

October 2, 2022

Submitted to:

Professor: BEHZAD AHMADI

Submitted by:

NIKSHITA RANGANATHAN

Introduction

This module of the assignment in Introduction to Analytics course has given an opportunity to brush up the R programming skills and gain perfection in the data analytics especially in statistical computing and graphical libraries. This assignment focussed on graphical plotting to convert the data into visually insightful elements like histograms, scatter plots, boxplots, frequency polygon about the shape of the data distribution.

The data visualization in this summary report is to initially focus on understanding the descriptive statistical differences in age and length between the Harrison Lake and Osprey BullTrout. The descriptive statistical analysis on the data of Harrison Lake bull trout derived from the BullTroutRML2 dataset and gives a key focus on the data analysis derived between three variables age, fork length and era of the Harrison Lake bull trout dataset.

Code and Outputs

1. Print your name at the top of the script. Include the prefix: "Plotting Basics:"

```
> print("Plotting Basics - Nikshita Ranganathan")
[1] "Plotting Basics - Nikshita Ranganathan"
```

2. Import libraries including: plyr, FSA, FSAdata, magrittr, dplyr, plotrix, ggplot2, and moments

```
> install.packages(c("plyr", "FSA", "FSAdata", "magrittr", "dply
r", "plotrix", "ggplot2", "moments"))
> library(plyr)
> library(FSA)
> library(FSAdata)
> library(magrittr)
> library(dplyr)
> library(plotrix)
> library(ggplot2)
> library(moments)
```

Install.packages() is utilized to download packages for CRAN. To load the packages, library() is applied.

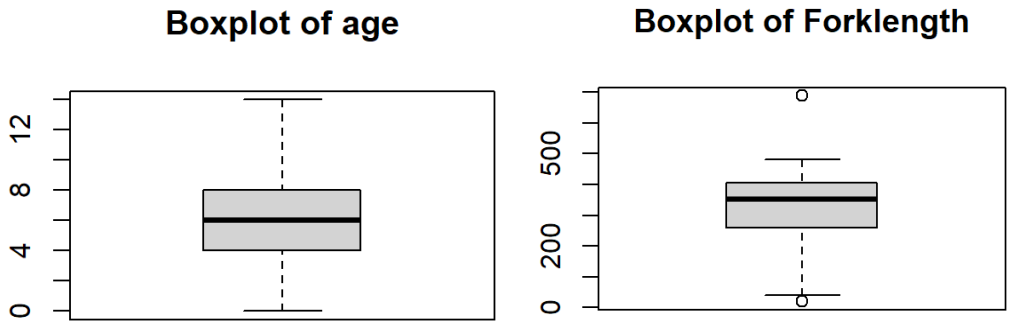
3. Load the BullTroutRML2 dataset

```
> 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
```

BulltroutRML2 is a dataframe made up of 4 variables and 96 observations. It is one of the dataset from FSAdata package and has columns age,forklength,lakes and eras. It gives information about ages and forklengths of Bulltrout found in Harrison and Osprey lakes from two eras (1977-80 amd 1997-01).

4. Print the first and last 3 records from the dataset

```
> bulltrout<-BullTroutRML2
> head(bulltrout,3)
  age  fl     lake     era
1  14 459 Harrison 1977-80
2  12 449 Harrison 1977-80
3  10 471 Harrison 1977-80
> tail(bulltrout,3)
   age  fl   lake     era
94   4 298 Osprey 1997-01
95   3 279 Osprey 1997-01
96   3 273 Osprey 1997-01
> summary(bulltrout)
      age              fl             lake         era
 Min.   : 0.000   Min.   : 20.0   Harrison:61   1977-80:38
 1st Qu.: 4.000   1st Qu.:258.0   Osprey  :35   1997-01:58
 Median : 6.000   Median :352.5
 Mean   : 5.771   Mean   :326.1
 3rd Qu.: 8.000   3rd Qu.:406.0
 Max.   :14.000   Max.   :688.0
> sd(bulltrout$age)
[1] 2.925313
> var(bulltrout$age)
[1] 8.557456
> sd(bulltrout$fl)
[1] 112.2022
> var(bulltrout$fl)
[1] 12589.34
> boxplot(bulltrout$age,main="Boxplot of age")
> boxplot(bulltrout$fl,main="Boxplot of Forklength")
```



head() function returns the first n number of rows and tail() function returns the last n rows. In this question n = 3. Standard deviation and variance of ages in Bulltrout dataset are approximately 2.92 and 8.55. On the other hand, standard deviation, and variance of forklength columns are 112.2 and 12589.3. There are two outliers for forklength boxplot and no outliers for age boxplot.

5. Filter out all records except those from Harrison Lake

```
> library(dplyr)
> Harrison<-filter(bulltrout,lake=="Harrison")
> Harrison
   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
```

Filter() command is a part of dplyr package. It helps us select rows based on a specific condition. All the information related to Osprey lake have been removed for this question.

6. Display the first and last 3 records from the filtered dataset

```
> head(Harrison,3)
  age  fl     lake     era
1  14 459 Harrison 1977-80
2  12 449 Harrison 1977-80
3  10 471 Harrison 1977-80
> tail(Harrison,3)
   age  fl     lake     era
59   7 245 Harrison 1997-01
60   7 279 Harrison 1997-01
61   5 245 Harrison 1997-01
```

7. Display the structure of the filtered dataset

```
> str(Harrison)
'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: chr  "Harrison" "Harrison" "Harrison" "Harrison" ...
 $ era : chr  "1977-80" "1977-80" "1977-80" "1977-80" ...
```

The filtered dataset (Harrison) contains 61 observations and 4 variables which shows ages and forklengths of bulltrout in Harrison Lake during eras 1977-80s and 1977-01s.

8. Display the summary of the filtered dataset and save it as <t>

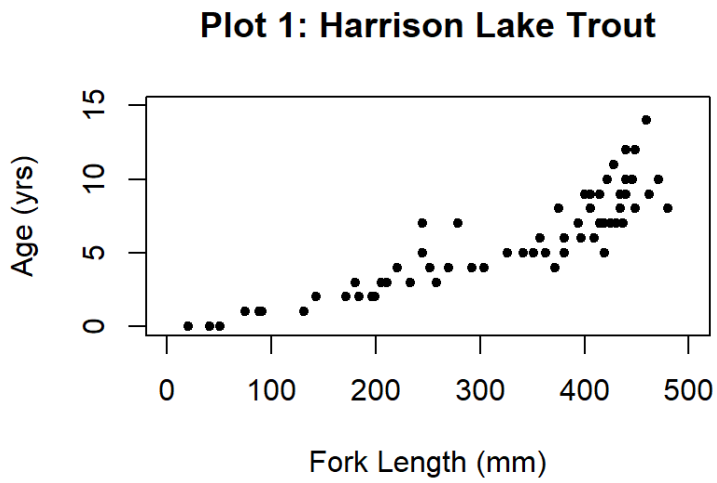
```
> summary(Harrison)
      age              fl          lake              era
 Min.   : 0.000   Min.   : 20   Length:61          Length:61
 1st Qu.: 3.000   1st Qu.:221   Class :character   Class :character
 Median : 6.000   Median :372   Mode  :character   Mode  :character
 Mean   : 5.754   Mean   :319
 3rd Qu.: 8.000   3rd Qu.:425
 Max.   :14.000   Max.   :480
> t<-summary(Harrison)
> t
      age              fl          lake              era
 Min.   : 0.000   Min.   : 20   Length:61          Length:61
 1st Qu.: 3.000   1st Qu.:221   Class :character   Class :character
 Median : 6.000   Median :372   Mode  :character   Mode  :character
 Mean   : 5.754   Mean   :319
 3rd Qu.: 8.000   3rd Qu.:425
 Max.   :14.000   Max.   :480
> qage<-quantile(Harrison$age,c(0,0.25,0.5,0.75,1))
> qage
  0%  25%  50%  75% 100%
   0    3    6    8   14
> qfl<-quantile(Harrison$fl,c(0,0.25,0.5,0.75,1))
> qfl
  0%  25%  50%  75% 100%
  20  221  372  425  480
```

Summary() gives us a basic understanding of the Harrison dataframe. Quartiles for age in Harrison dataset at 0%,25%,50%,75% and 100% are 0,3,6,8,14, respectively. On the other hand, for the fork length, it is observed to be 20,221,372,425 and 480.

9. Create a scatterplot for “age” (y variable) and “fl” (x variable) with the following specifications:
   - Limit of x axis is (0,500)
   - Limit of y axis is (0,15)
   - Title of graph is “Plot 1: Harrison Lake Trout
   - Y axis label is “Age (yrs)”

• X axis label is “Fork Length (mm)”
• Use a small, filled circle for the plotted data points

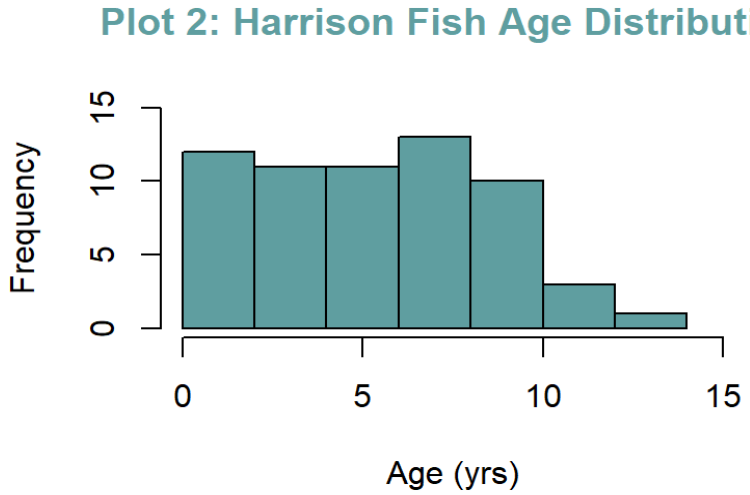
```
> plot(Harrison$fl,Harrison$age,ylim=c(0,15),xlim=c(0,500),main = "Plot 1:
 Harrison Lake Trout",xlab = "Fork Length (mm)",ylab = "Age (yrs)",pch=20 )
```



Scatterplot depicts an increasing relationship between age(years) and forklength(mm) of Harrison Lake trout.

10. Plot an “Age” histogram with the following specifications
    • Y axis label is “Frequency”
    • X axis label is “Age (yrs)”
    • Title of the histogram is “Plot 2: Harrison Fish Age Distribution”
    • The color of the frequency plots is “cadetblue”
    • The color of the Title is “cadetblue”

```
> attach(Harrison)
> hist(age,main="Plot 2: Harrison Fish Age Distribution",xlim=c(0,15),ylim=c
(0,15),xlab="Age (yrs)",ylab="Frequency",col.main="cadetblue",col="cadetblue")
```

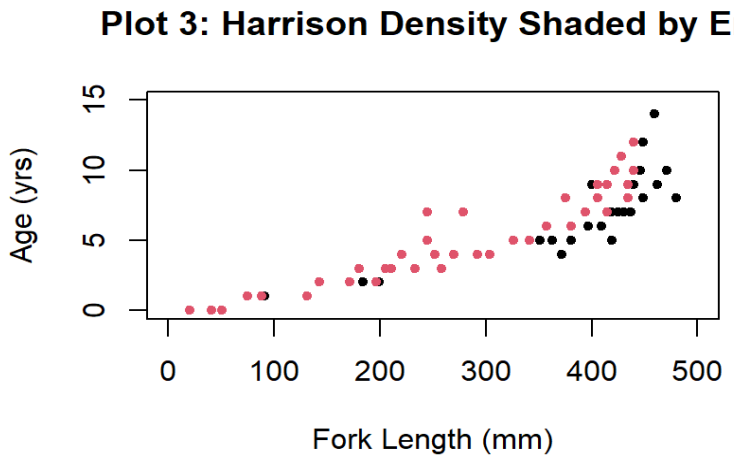



```
> mean(Harrison$age)
[1] 5.754098
> median(Harrison$age)
[1] 6
```

The above histogram shape is positively skewed. Mean of age elements in Harrison dataset is 5.75 approximately and median is 6.

11. Create an overdense plot using the same specifications as the previous scatterplot. But include two levels of shading for the “black” data points. Title the plot “Plot 3: Harrison Density Shaded by Era”

```
> plot(fl,age,main="Plot 3: Harrison Density Shaded by Era",xlab = "Fork Length (mm)",ylab
 = "Age (yrs)",ylim=c(0,15),xlim=c(0,500),pch=20,col=as.factor(Harrison$era))
```



Red data points represent Era 1997-01s and black points denote Era 1977-80s. The scatterplot signifies that there is an increase in forklength with the increase of age.

12. Create a new object called “tmp” that includes the first 3 and last 3 records of the wholedata set.

```
> library(FSA)
## FSA v0.9.3. See citation('FSA') if used in publication.
## Run fishR() for related website and fishR('IFAR') for related book.
> tmp<-headtail(Harrison,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
```

headtail() is an inbuilt function of FSA package. This combines first and last rows of the data.

13. Display the “era” column in the new “tmp” object

```
> era<-select(tmp,era)
> era
       era
1  1977-80
2  1977-80
3  1977-80
94 1997-01
95 1997-01
96 1997-01
```

select() is one of the functions stored in dplyr package. It selects a part of the dataframe and displays it.

14. 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(3,4)
> pchs
[1] 3 4
> cols<-c("red","gray60")
> cols
[1] "red"    "gray60"
```

15. Convert the tmp object values to numeric values. Then create a numeric numEra object from the tmp$era object

```
> class(tmp)
[1] "data.frame"
> numtmp<-as.numeric(unlist(tmp))
> numtmp
 [1]  14  12  10   4   3   3 459 449 471 298 279 273   1   1   1   2   2
[18]   2   1   1   1   2   2   2
> class(tmp$era)
[1] "factor"
> numEra<-as.numeric(tmp$era)
> numEra
[1] 1 1 1 2 2 2
> class(numEra)
[1] "numeric"
```

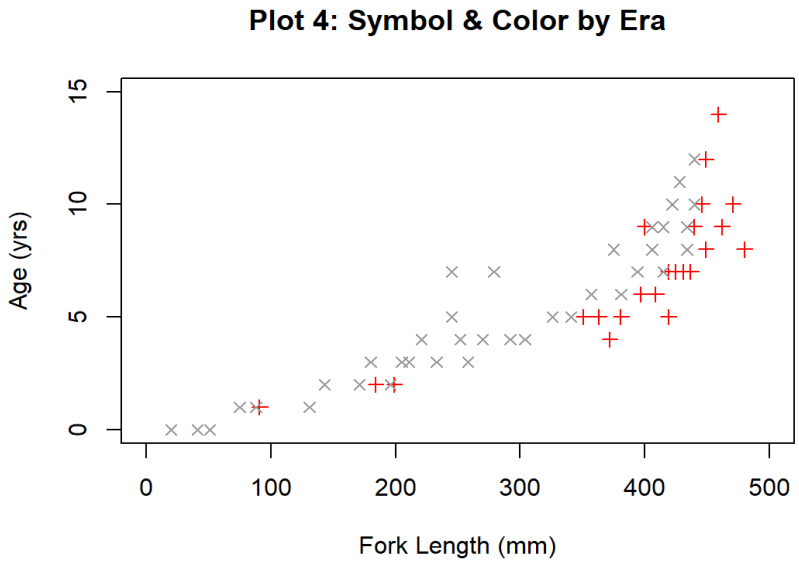
class() returns the datatype of the elements. In this case, tmp was a dataframe and it got changed into numeric. Similarly, era column in tmp dataframe was in factor form and it got converted into number format.

16. Associate the cols vector with the tmp era values

```
> cols[tmp$era]
[1] "red"    "red"    "red"    "gray60" "gray60" "gray60"
```

17. Create a plot of “Age (yrs)” (y variable) versus “Fork Length (mm)” (x variable) with the following specifications:
    - Limit of x axis is (0,500)
    - Limit of y axis is (0,15)
    - Title of graph is “Plot 4: Symbol & Color by Era”
    - X axis label is “Age (yrs)”
    - Y axis label is “Fork Length (mm)”
    - Set pch equal to pchs era values
    - Set col equal to cols era values

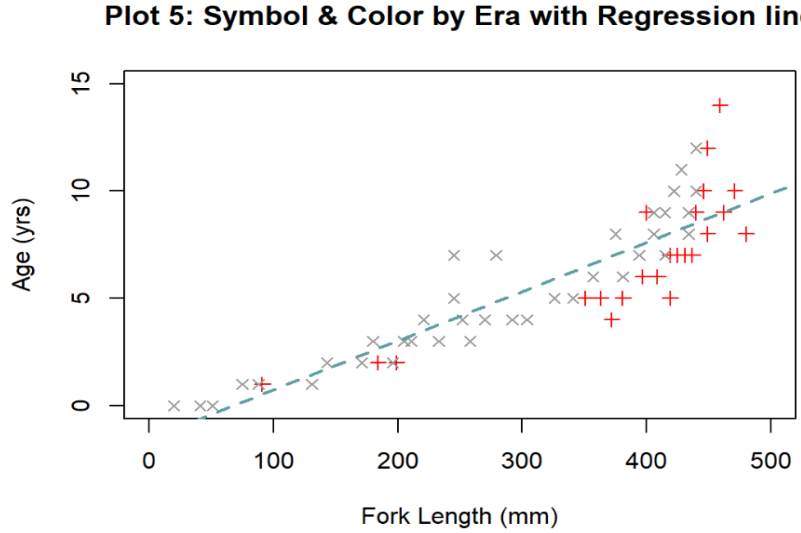
```
> plot(fl,age,main = "Plot 4: Symbol & Color by Era",xlab = "Fork Length (mm)",ylab = "Age
 (yrs)",xlim = c(0,500),ylim=c(0,15),pch=pchs[as.factor(Harrison$era)],col=cols[as.factor
(Harrison$era)])
```



Symbol "x" represents 1997-01s era and Symbol "+" represents 1977-80s era.

18. Plot a regression line of the previous plot with a dashed line with width 2 and color "cadetblue"

```
> plot(fl,age,main = "Plot 5: Symbol & Color by Era with Regression line",xlab = "Fork Len
gth (mm)",ylab = "Age (yrs)",xlim = c(0,500),ylim=c(0,15),pch=pchs[as.factor(Harrison$er
a)],col=cols[as.factor(Harrison$era)])
> abline(lm(formula=age~fl),lty=2,lwd=2,col="cadetblue")
```

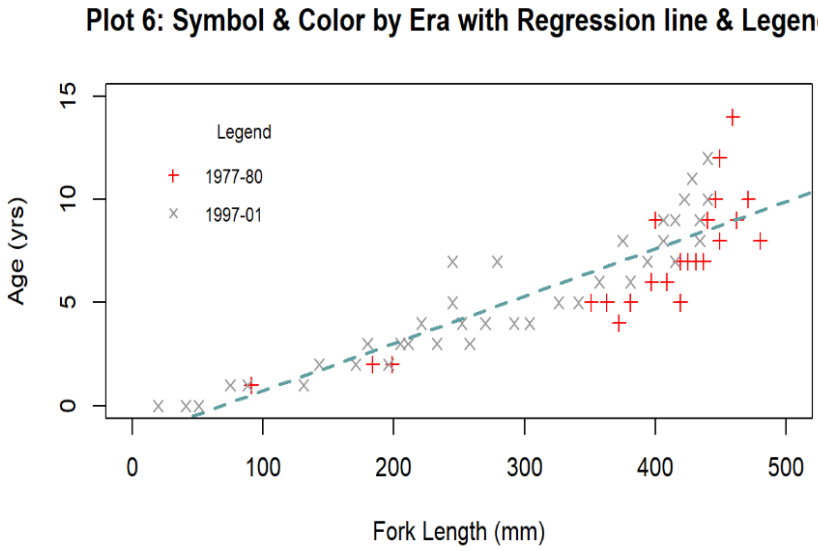


Regression line in the above graph displays linear relationship between age and forklength.

19. Place a legend of levels by era with pchs symbols in the top left of the plot with the following specifications:

- Inset of 0.05
- No box around the legend
- Font size: 75% of nominal

```
> plot(fl,age,main = "Plot 6: Symbol & Color by Era with Regression line & Legend",xlab = "Fork Length
 (mm)",ylab = "Age (yrs)",xlim = c(0,500),ylim=c(0,15),pch=pchs[as.factor(Harrison$era)],col=cols[as.f
actor(Harrison$era)])
> abline(lm(formula=age~fl),lty=2,lwd=2,col="cadetblue")
> legend("topleft",inset=0.05,cex=0.75,c("1977-80","1997-01"),box.col = "white",pch = c(3,4),col=cols,
title = "Legend")
```




```
> kurtosis(Harrison$fl)
[1] 2.345694
> kurtosis(Harrison$age)
[1] 2.357344
> skewness(Harrison$fl)
[1] -0.7353215
> skewness(Harrison$age)
[1] 0.1677228
```

Age in Harrison dataset has positive skew. Legend shows the symbols that characterizes the two eras. Contrarily, forklength has a skew value less than 0 and hence it is negatively skewed. Both forklength and age elements have coefficient of kurtosis which is less than 3 and it has a flat peak (Platykurtic).

Summary

This assignment is an explanatory analysis that allows us to demonstrate our skills to process data, present the data visually, and compare the dataset of observations of 1997-01 & 1977-08. While doing the comparison of two dataset, the analysis observed shows that Harrison Lake bull trout tends to grow broader and wider in 1977-08 compared to 1977-01.

Bibliography

Kabacoff, Robert.I. (2011). R in Action Data analysis and graphics with R. Manning

Bluman, Allan G. (2017). Elementary statistics A Step-by-step approach. McGraw Hill

Statistics Globe. (2019 Oct 26). plot() Function in R (8 Examples) | How Plot Data in RStudio | density() & lines() [Video File]. Retrieved from https://www.youtube.com/watch?v=utisKvP0HOM

DataCamp. (2015 Dec 4). R Tutorial - Customizing Your Plots In R [Video File]. Retrieved from https://www.youtube.com/watch?v=0MrYVzPxBIc

DataDaft. (2019 Sep 17). dplyr: filter [Video File]. Retrieved from https://www.youtube.com/watch?v=BkmYBBM2SdQ&list=PLiC1doDIe9rC8RgWPAWqDETE-VbKOWfWl&index=3

References (Websites):

https://www.tutorialspoint.com/r/r_scatterplots.htm

https://www.rdocumentation.org/

Appendix

```
---
title: "Module Project-2"
author: “Nikshita”
output:  word_document: default
date: "2022-10-02"
---
# 1. Print your name at the top of the script. Include the prefix: “Plotting Basics:”
print("Plotting Basics - Nikshita Ranganathan")

# 2. Import libraries including: plyr, FSA, FSAdata, magrittr, dplyr, plotrix, ggplot2, and moments
install.packages(c("plyr","FSA","FSAdata","magrittr","dplyr","plotrix","ggplot2","moments"))
library(plyr)
library(FSA)
library(FSAdata)
library(magrittr)
library(dplyr)
library(plotrix)
library(ggplot2)
library(moments)

# 3. Load the BullTroutRML2 dataset
data("BullTroutRML2")
BullTroutRML2

# 4. Print the first and last 3 records from the dataset
bulltrout<-BullTroutRML2
head(bulltrout,3)
tail(bulltrout,3)

# 5. Filter out all records except those from Harrison Lake
library(dplyr)
Harrison<-filter(bulltrout,lake=="Harrison")
Harrison

# 6. Display the first and last 3 records from the filtered dataset
```

```
head(Harrison,3)
tail(Harrison,3)

# 7. Display the structure of the filtered dataset
str(Harrison)

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

# 9. Create a scatterplot for “age” (y variable) and “fl” (x variable) with the following
specifications:
# • Limit of x axis is (0,500)
# • Limit of y axis is (0,15)
# • Title of graph is “Plot 1: Harrison Lake Trout
# • Y axis label is “Age (yrs)”
# • X axis label is “Fork Length (mm)”
# • Use a small filled circle for the plotted data points
plot(Harrison$fl, Harrison$age, ylim=c(0,15), xlim=c(0,500),main = "Plot 1: Harrison Lake
Trout", xlab = "Fork Length (mm)", ylab = "Age (yrs)",pch=20 )

# 10. Plot an “Age” histogram with the following specifications
# • Y axis label is “Frequency”
# • X axis label is “Age (yrs)”
# • Title of the histogram is “Plot 2: Harrison Fish Age Distribution”
# • The color of the frequency plots is “cadetblue”
# • The color of the Title is “cadetblue”
attach(Harrison)
hist(age, main="Plot 2: Harrison Fish Age Distribution", xlim=c(0,15), ylim=c(0,15),
xlab="Age (yrs)", ylab="Frequency",col.main="cadetblue",col="cadetblue")

# 11. Create an overdense plot using the same specifications as the previous scatterplot. But,
include two levels of shading for the “black” data points. Title the plot “Plot 3: Harrison
Density Shaded by Era”
plot(fl, age, main="Plot 3: Harrison Density Shaded by Era",xlab = "Fork Length (mm)",ylab
= "Age (yrs)",ylim=c(0,15),xlim=c(0,500),pch=20,col=as.factor(Harrison$era))

# 12. Create a new object called “tmp” that includes the first 3 and last 3 records of the whole
data set.
library(FSA)
tmp<-headtail(Harrison,n=3)
tmp

# 13. Display the “era” column in the new “tmp” object
era<-select(tmp,era)
```

era

# 14. 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(3,4)
pchs
cols<-c("red","gray60")
cols

# 15. Convert the tmp object values to numeric values. Then create a numeric numEra object from the tmp$era object
class(tmp)
numtmp<-as.numeric(unlist(tmp))
numtmp
class(tmp$era)
numEra<-as.numeric(tmp$era)
numEra
class(numEra)

16. Associate the cols vector with the tmp era values
cols[tmp$era]

# 17. Create a plot of “Age (yrs)” (y variable) versus “Fork Length (mm)” (x variable) with the following specifications:
# • Limit of x axis is (0,500)
# • Limit of y axis is (0,15)
# • Title of graph is “Plot 4: Symbol & Color by Era”
# • X axis label is “Age (yrs)”
# • Y axis label is “Fork Length (mm)”
# • Set pch equal to pchs era values
# • Set col equal to cols era values
plot(fl,age,main = "Plot 4: Symbol & Color by Era",xlab = "Fork Length (mm)",ylab = "Age (yrs)",xlim = c(0,500), ylim=c(0,15), pch=pchs[as.factor(Harrison$era)], col=cols[as.factor(Harrison$era)])

# 18. Plot a regression line of the previous plot with a dashed line with width 2 and color “cadetblue”
plot(fl,age,main = "Plot 5: Symbol & Color by Era with Regression line",xlab = "Fork Length (mm)",ylab = "Age (yrs)",xlim = c(0,500), ylim=c(0,15), pch= pchs[as.factor(Harrison$era)], col=cols[as.factor(Harrison$era)])
abline(lm(formula=age~fl),lty=2,lwd=2,col="cadetblue")

19. Place a legend of levels by era with pchs symbols in the top left of the plot with the following specifications:
• Inset of 0.05
• No box around the legend

```
# • Font size: 75% of nominal
plot(fl,age,main = "Plot 6: Symbol & Color by Era with Regression line & Legend",xlab =
"Fork Length (mm)",ylab = "Age (yrs)",xlim = c(0,500), ylim=c(0,15),
pch=pchs[as.factor(Harrison$era)], col=cols[as.factor(Harrison$era)])
abline(lm(formula=age~fl),lty=2,lwd=2,col="cadetblue")
legend("topleft",inset=0.05,cex=0.75,c("1977-80","1997-01"),box.col = "white",pch = c(3,4),
col=cols,title = "Legend")

# Additional Practice
summary(bulltrout)
str(bulltrout)
sd(bulltrout$age)
var(bulltrout$age)
sd(bulltrout$fl)
var(bulltrout$fl)
boxplot(bulltrout$age,main="Boxplot of age")
boxplot(bulltrout$fl,main="Boxplot of Forklength")
qage<-quantile(Harrison$age,c(0,0.25,0.5,0.75,1))
qage
qfl<-quantile(Harrison$fl,c(0,0.25,0.5,0.75,1))
qfl
IQR(Harrison$age)
IQR(Harrison$fl)
kurtosis(Harrison$fl)
kurtosis(Harrison$age)
skewness(Harrison$fl)
skewness(Harrison$age)
mean(Harrison$age)
median(Harrison$age)
```