Executive Summary Report 2

Name - Shamim Sherafati

ALY 6000 – Introduction to Analytics

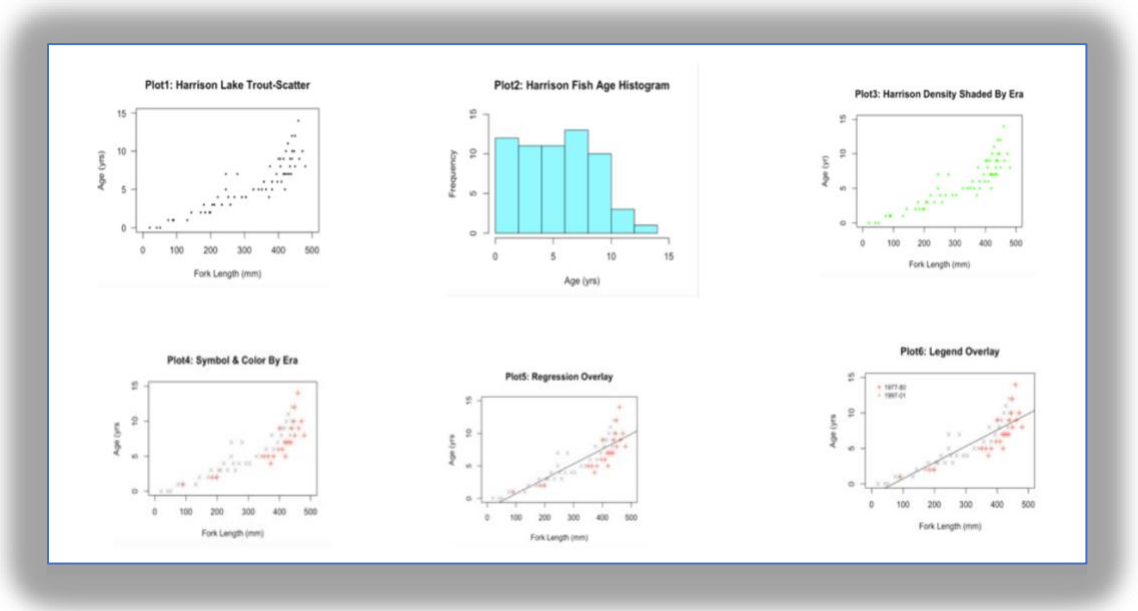
Date - 2022-10-05

Campus – Vancouver

NUID – 002742363

Abstract

The following is a Report based on the Module 2 of ALY 6000. This represents authentic work done by Ms. Shamim Sherafati for the subject and is duly submitting her work before the decided date and time



1. Print your name

```
name <- ("Shamim Sherafati")
r_name = paste("Plotting Basics : ",name)
print(r_name)
[1] "Plotting Basics :  Shamim Sherafati"
```

2. Import libraries

```
install.packages ("plyr")
trying URL 'https://mirror.rcg.sfu.ca/mirror/CRAN/bin/macosx/contrib/4.2/plyr_1.8.7.tgz'
Content type 'application/octet-stream' length 1014519 bytes (990 KB)
==================================================
downloaded 990 KB


The downloaded binary packages are in
    /var/folders/cn/j3k0jn5d6qvctnpb78k90_rm0000gn/T//RtmptQ2Shf/downloaded_packages
install.packages ("FSA")
trying URL 'https://mirror.rcg.sfu.ca/mirror/CRAN/bin/macosx/contrib/4.2/FSA_0.9.3.tgz'
Content type 'application/octet-stream' length 1117449 bytes (1.1 MB)
==================================================
downloaded 1.1 MB


The downloaded binary packages are in
    /var/folders/cn/j3k0jn5d6qvctnpb78k90_rm0000gn/T//RtmptQ2Shf/downloaded_packages
install.packages ("FSAdata")
trying URL 'https://mirror.rcg.sfu.ca/mirror/CRAN/bin/macosx/contrib/4.2/FSAdata_0.3.9.tgz'
Content type 'application/octet-stream' length 919196 bytes (897 KB)
==================================================
downloaded 897 KB


The downloaded binary packages are in
    /var/folders/cn/j3k0jn5d6qvctnpb78k90_rm0000gn/T//RtmptQ2Shf/downloaded_packages
```

```
install.packages ("magrittr")

Error in install.packages : Updating loaded packages

install.packages ("dplyr")

Error in install.packages : Updating loaded packages

install.packages ("plotrix")

trying URL 'https://mirror.rcg.sfu.ca/mirror/CRAN/bin/macosx/contrib/4.2/plotrix_3.8-2
.tgz'

Content type 'application/octet-stream' length 1137125 bytes (1.1 MB)

==================================================

downloaded 1.1 MB


The downloaded binary packages are in

    /var/folders/cn/j3k0jn5d6qvctnpb78k90_rm0000gn/T//RtmptQ2Shf/downloaded_packages

install.packages ("ggplot2")

Error in install.packages : Updating loaded packages

install.packages ("moments")

trying URL 'https://mirror.rcg.sfu.ca/mirror/CRAN/bin/macosx/contrib/4.2/moments_0.14.
1.tgz'

Content type 'application/octet-stream' length 54374 bytes (53 KB)

==================================================

downloaded 53 KB


The downloaded binary packages are in

    /var/folders/cn/j3k0jn5d6qvctnpb78k90_rm0000gn/T//RtmptQ2Shf/downloaded_packages
```

3. Load the BullTroutRML2 dataset

```
data1 <- read.table (file="~/desktop/BullTroutRML2.txt", header= TRUE, sep=",", string
sAsFactors = FALSE)

data1
```

Describe number 3: Dataset of BullTroutRML2 which was automatically loaded while installing packages.

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

```
head (data1,  n=3 )

tail (data1, n=3 )

#headtail (data1, n=3)
```

Describe number 4: Used (head) for displaying the first records and used (tail) for displaying the last records with using “n” for the numbers of records which are shown.

5. Filter out all records except those from Harrison Lake

```
library(dplyr)


data_f = filter(data1, lake == 'Harrison')


data_f
```

Describe number 5: For filtering all datas except one of them, first I introduced an object “data_f” then write “filter” and the name of dataset and used == which means showing only that data which comes after ==.

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

```


head(data_f, n=3)
tail(data_f, n=3)
```

Describe number 6: Its like the previous question with this difference that we should show the head and tail for filtered data, so the object which I choose for filtered data, I write.

7. Display the structure of the filtered dataset

```
str(data_f)
'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" ...
```

Describe number 7: For displaying the structure, we should use (str) and then the object which choosed for filtered data which is data_f.

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

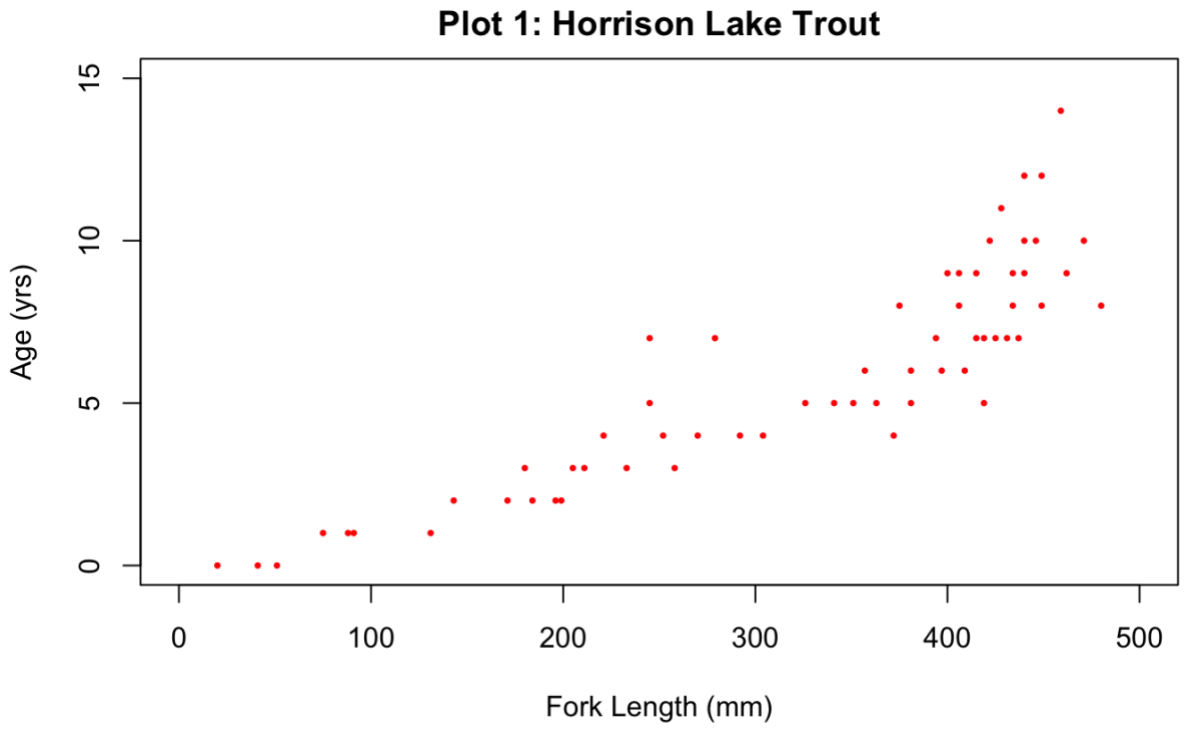
```
t <- summary(data_f)
t
      age               fl          lake               era
 Min.   : 0.000   Min.   : 20   Length:61          Length:61
```

```
 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
```

Describe number 8: First I choose a name for the summary (t) and then use summary + the name of filtered data.

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)

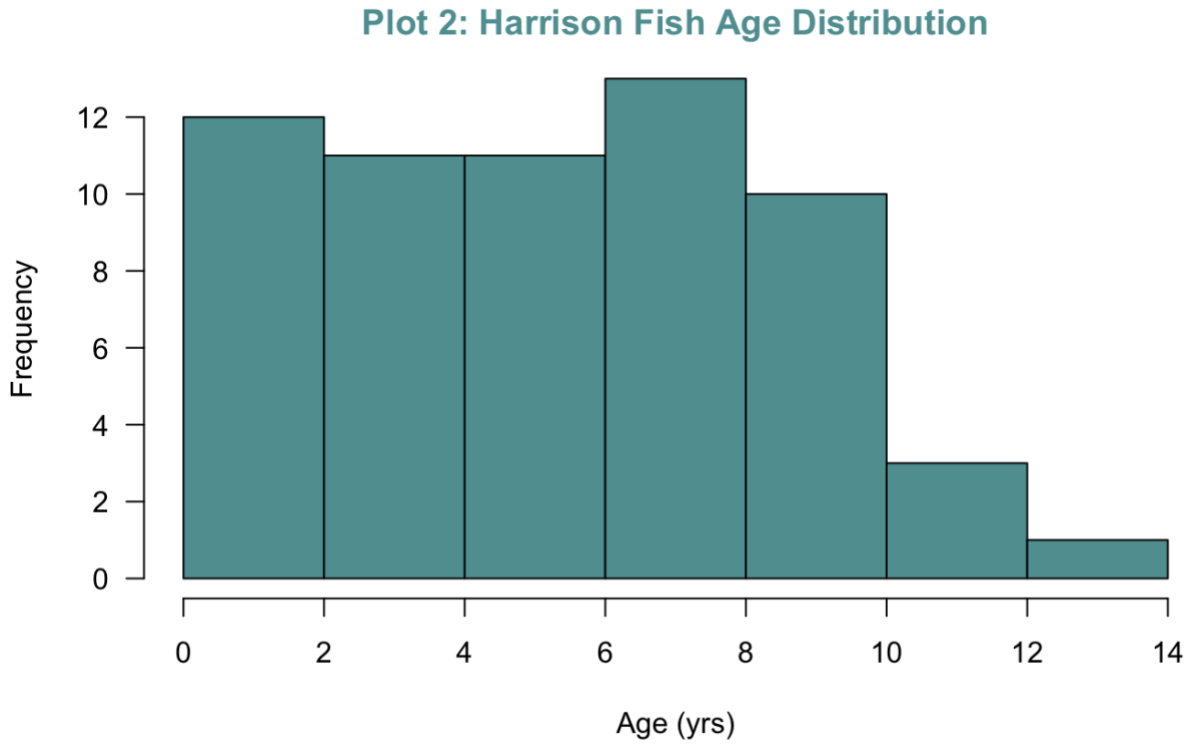
```
plot (data_f$fl, data_f$age, main = ("Plot 1: Horrison Lake Trout"), xlim = c(0, 500),
ylim = c(0, 15), xlab = ("Fork Length (mm)"), ylab = ("Age (yrs)"),col="red", cex=0.5,
pch=16)
```



Describe number 9: I create a scatter plot with “age” (y variable) and “fl” (x variable). This Scatter Plot shows strong Positive and Linear Relationship between Fork Length and Age of the Dataset.

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”

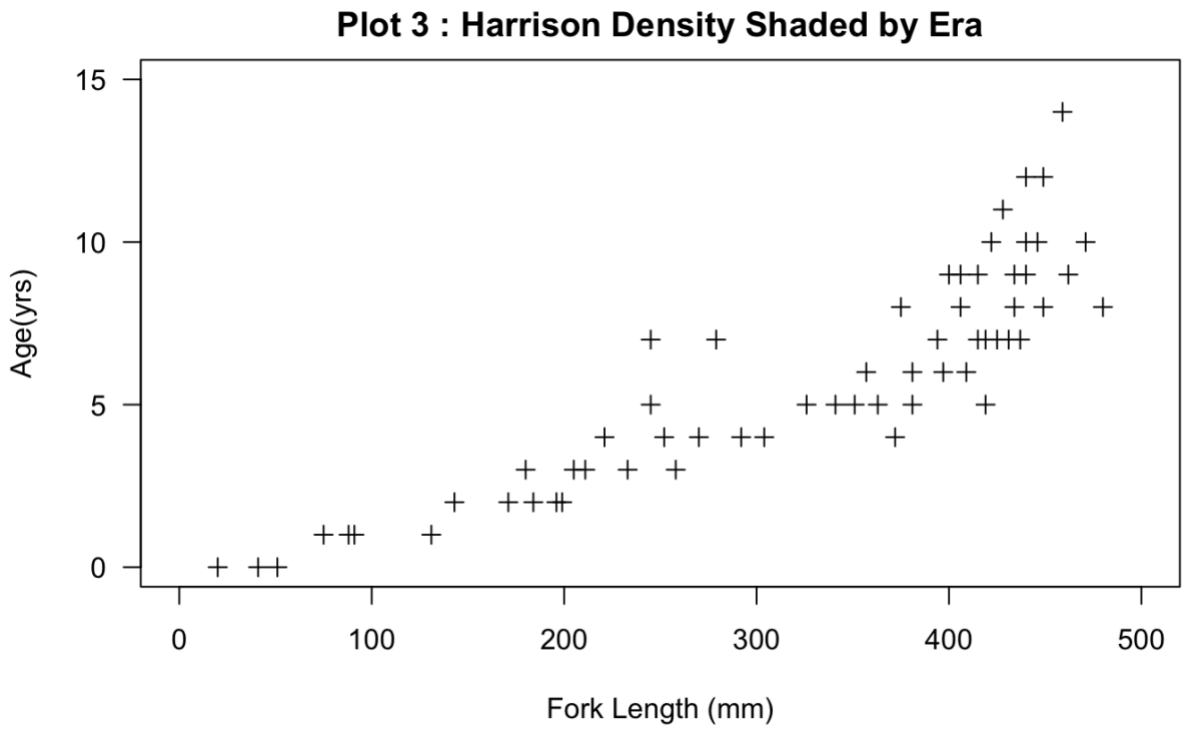
```
hist(data_f$age, main= "Plot 2: Harrison Fish Age Distribution" , xlab= "Age (yrs)" ,
ylab= "Frequency", col="cadetblue", col.main="cadetblue", las=1)
```



Describe number 10: Now, with the same dataset and x and y variable, We can customize this Histogram by adding colours to the bars as well as to the Title with choosing “cadetblue” for both of them.

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 (data_f$fl, data_f$age, xlim=c(0, 500), ylim=c(0,15), xlab= "Fork Length (mm)",
      ylab= "Age(yrs)", main="Plot 3 : Harrison Density Shaded by Era",  col=c("black"
, "black")  , cex=1,pch=3, las=1 )
```




```
NA
NA
```

Describe number 11: This ScatterPlot shows the strong,positive and Linear relationship between Fork length and Age of the Dataset: data1_Harrison, with including two levels of shading for the "black" data points.

12. Create a new object called "tmp" that includes the first 3 and last 3 records of the whole data set

```
tmp <- data.frame(rbind(head(data1,3),tail(data1, 3)))
tmp
```

Describe number 12: Here, Create a new Dataframe(tmp) with two dataframes (First 3 Records which should used head and Last 3 records (tail) of BullTroutRML2 dataset which is called (data1), having each values in the column.

13. Display the "era" column in the new "tmp" object

```
eratmp <- tmp$era
eratmp
[1] "1977-80" "1977-80" "1977-80" "1997-01" "1997-01" "1997-01"
```

Describe number 13: Create new object called(eratmp) to show the “era” column in the new “tmp” object.

14.Create a pchs vector with the argument values for + and x. Then create a cols vector with the two elements “red” and “gray60”

```
pch <-c("+", "x")
pch
[1] "+" "x"
cols <-c ("red", "gray60")
cols
[1] "red"    "gray60"
```

Describe number 14: Now in this section, I have included the Shape of the Points and Colour of the Points in vectors - pchs and cols respectively.

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

```
tmp$era <- as.numeric(tmp$era)
Warning: NAs introduced by coercion
tmp$era
[1] NA NA NA NA NA NA

numEra <- ("tmp$era")
numEra
[1] "tmp$era"
```

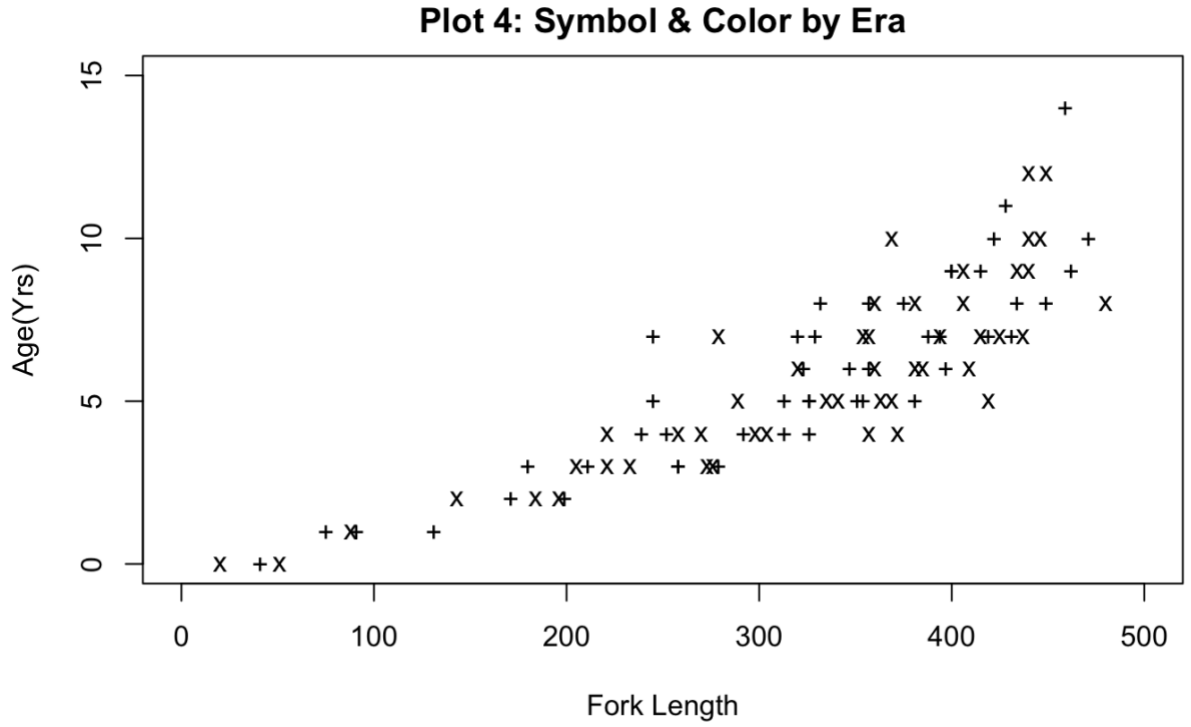
Describe number 15: Here, first we need to convert tmp object to numeric, so I created tmp$era and then use its function. Then as the question mentioned, I created a new numeric object called numEra from the tmp$era.

16. Associate the cols vector with the tmp era values

```
cols <- tmp$era
cols
[1] "1977-80" "1977-80" "1977-80" "1997-01" "1997-01" "1997-01"
```

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(data1$fl,data1$age, main = "Plot 4: Symbol & Color by Era", xlim=c(0, 500), ylim=
c(0,15), xlab="Fork Length", ylab = "Age(Yrs)", pch = pchs, cols=col)
```

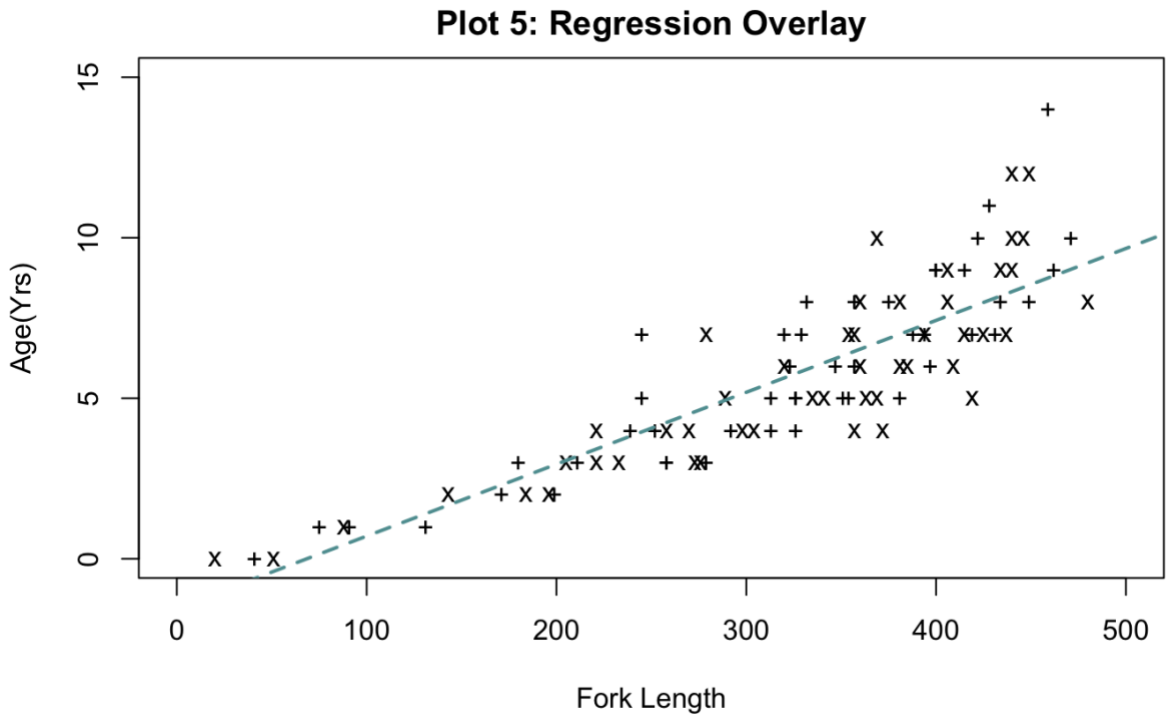


Describe number 17: As per the above graph, This Scatter plot shows weak, positive and Linear relationship with no potential outliers between the Fork length and Age of the tmp Dataset.

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

```
plot(data1$fl,data1$age, main = "Plot 5: Regression Overlay", xlim=c(0, 500), ylim=c(0
,15), xlab="Fork Length", ylab = "Age(Yrs)", pch = pchs, cols=col)


abline(lm(age~fl, data = data1), lty=2, lwd=2, col = 'cadetblue')
```



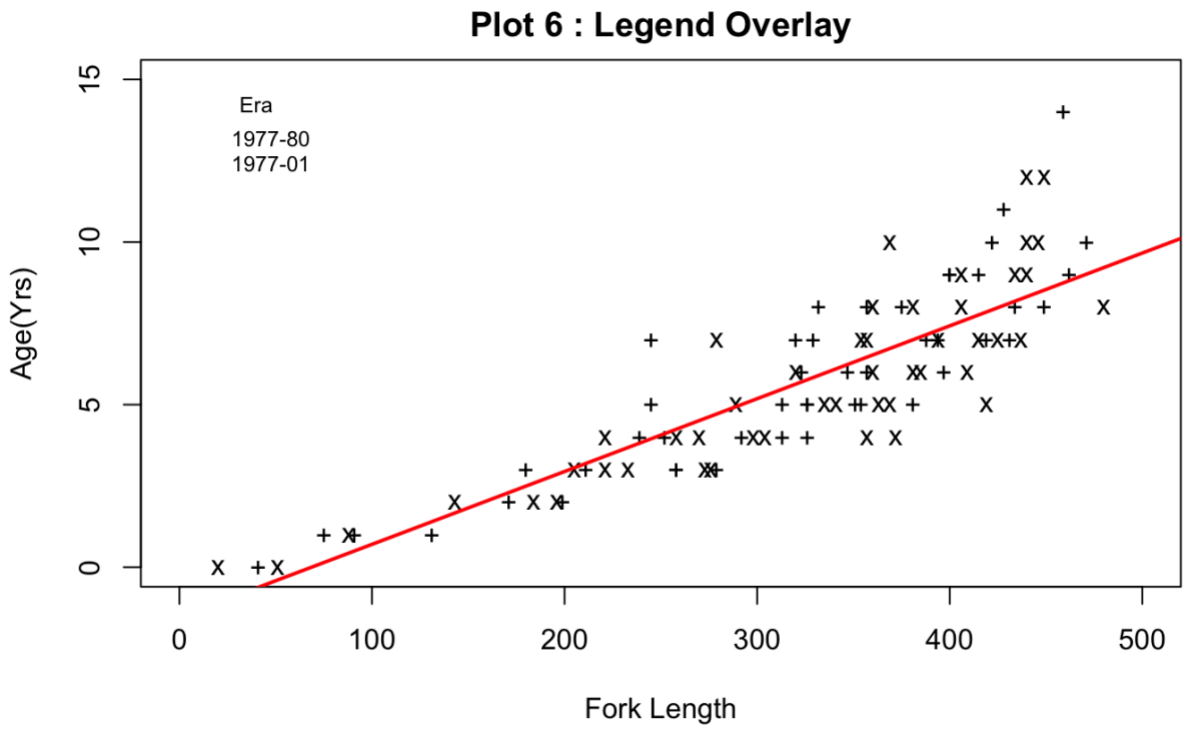
Describe number 18: Here, Only 1 Regression line has been overlayed. The regression line which can be seen, makes it easy to understand the data, by showing how the data points approximately fit in the line perfectly. This shows that Scatter plot here is showing Weak, Positive and Linear Relationship with no Outliers between Fork length and Age of the tmp Dataset.

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(data1$fl,data1$age, main = "Plot 6 : Legend Overlay", xlim=c(0, 500), ylim=c(0,15
), xlab="Fork Length", ylab = "Age(Yrs)", pch = pchs, cols=col)

abline(lm(age~fl, data = data1), lty=1, lwd=2, col = 'red')

legend('topleft', inset = 0.05, title="Era",c("1977-80","1977-01"), bty='n', cex=0.75)
```



Describe number 19: Here, as we write top left, the legend is appeared on that side; Legend Title - Era is displayed with:

Color black : 1977-80 (Shape Circle)

Color red : 1997- 01 (Shape circle)

BIBLIOGRAPHY:

https://r-coder.com/add-legend-r/

geeksforgeeks.com

Tutorialspoint.com

Stackoverflow.com

R in Action by R. Kabacoff

SUMMARY:

In the following assignment, we learnt the following

- Calculate basic descriptive statistics to describe a set of data
- Create various types of graph based on data provided
- Use R to visualize data
- Explain the significance of calculate statistics and graphs

Appendix:

```
name <- ("Shamim Sherafati")
r_name = paste("Plotting Basics :",name)
print(r_name)
install.packages ("plyr")
```

```
install.packages (“FSA”)

install.packages (“FSAdata”)

install.packages (“magrittr”)

install.packages (“dplyr”)

install.packages (“plotrix”)

install.packages (“ggplot2”)

install.packages (“moments”)

data1 <- read.table (file=“~/desktop/BullTroutRML2.txt”, header= TRUE, sep=“,”,
>stringsAsFactors = FALSE)

data1 head (data1, n=3 )

tail (data1, n=3 )

#headtail (data1, n=3)

library(dplyr)

data_f = filter(data1, lake == ‘Harrison’)

data_f

head(data_f, n=3)

tail(data_f, n=3)

str(data_f)

t <- summary(data_f)

t
```

```
plot (data_f\(fl, data_f\)age, main = (“Plot 1: Horrison Lake Trout”), xlim = c(0, 500), ylim = c(0, >15), xlab = (“Fork Length (mm)”), ylab = (“Age (yrs)”),col=“red”, cex=0.5, pch=16)

hist(data_f$age, main= “Plot 2: Harrison Fish Age Distribution” , xlab= “Age (yrs)” , ylab= >“Frequency”, col=“cadetblue”, col.main=“cadetblue”, las=1)

plot (data_f\(fl, data_f\)age, xlim=c(0, 500), ylim=c(0,15), xlab= “Fork Length (mm)”, ylab= “Age(yrs)”, main=“Plot 3 : Harrison Density Shaded by Era”, col=c(“black”, “gray”) , cex=0.5,pch=15, las=1 )

tmp <- data.frame(rbind(head(data1,3),tail(data1, 3)))

tmp

eratmp <- tmp$era

eratmp

pch <-c(“+”, “x”)

pch

cols <-c (“red”, “gray60”)

cols

tmp\(era <- as.numeric(tmp\)era)

tmp$era

numEra <- (“tmp$era”)

numEra

cols <- tmp$era

cols
```

```
plot(data1\(fl,data1\)age, main = “Plot 4: Symbol & Color by Era”, xlim=c(0, 500),
ylim=c(0,15), xlab=“Fork Length”, ylab = “Age(Yrs)”, pch = pchs, cols=col)
```

```
plot(data1\(fl,data1\)age, main = “Plot 5: Regression Overlay”, xlim=c(0, 500),
ylim=c(0,15), xlab=“Fork Length”, ylab = “Age(Yrs)”, pch = pchs, cols=col)
abline(lm(age~fl, data = data1), lty=2, lwd=2, col = ‘cadetblue’)
```

```
plot(data1\(fl,data1\)age, main = “Plot 6 : Legend Overlay”, xlim=c(0, 500),
ylim=c(0,15), xlab=“Fork Length”, ylab = “Age(Yrs)”, pch = pchs, cols=col)
abline(lm(age~fl, data = data1), lty=1, lwd=2, col = ‘red’) legend(‘topleft’, inset = 0.05,
title=“Era”,c(“1977-80”,“1977-01”), bty=‘n’, cex=0.75)
```