NORTHEASTERN UNIVERSITY

Toronto, Canada

Executive Summary of Module 2

Introduction to Data Analysis

(ALY6000)

Submitted by:

Name: Dhruvang Patel

NUID: 002195090

Date: 25th January, 2022

Under the guidance of:

Prof. Mohammad Shafiqul Islam

Key Findings:

- In the following report, the analysis of data is of BullTroutRML2 in which assigned ages and fork lengths of Bull Trout from two Rocky Mountain (Harrison and Osprey) lakes in Alberta, CAN before and after a regulation change.

- For narrowing the analysis of Harrison Lake, we need to filter out other data except for Harrison Lake. So, the program to perform
  Input:
  Harrisonlake<-filter(BullTroutRML2, lake=="Harrison")
  Harrisonlake

- If there is a specific sort of data which you want to do analysis of other than the whole data, it is also done by creating a object in which we can load data.
  For example, I want to make different set of data for first and last 3 records of main data then an object is created and that records are inserted in the object.
  Input:
  tmp <- headtail(Harrisonlake,3)

```
> tmp
   age  fl     lake era
1   14 459 Harrison   1
2   12 449 Harrison   1
3   10 471 Harrison   1
59   7 245 Harrison   2
60   7 279 Harrison   2
61   5 245 Harrison   2
```

where tmp is the new object.

- To access one variable from a particular set of data, we can get it with the $ symbol.

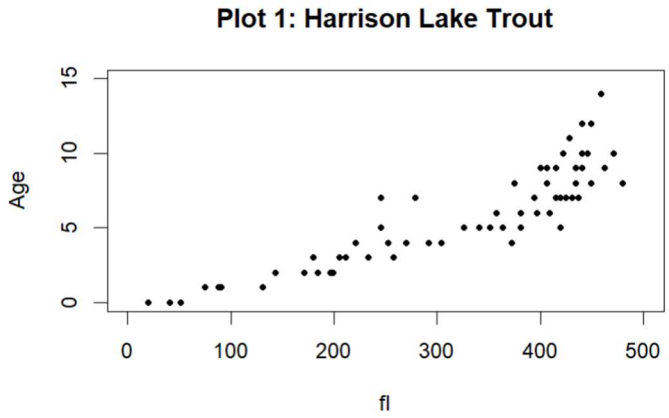
```
> tmp$era
[1] 1 1 1 2 2 2
>
```

- Different kind of vectors are also created, and those vectors can also be initiated into the values of the data.

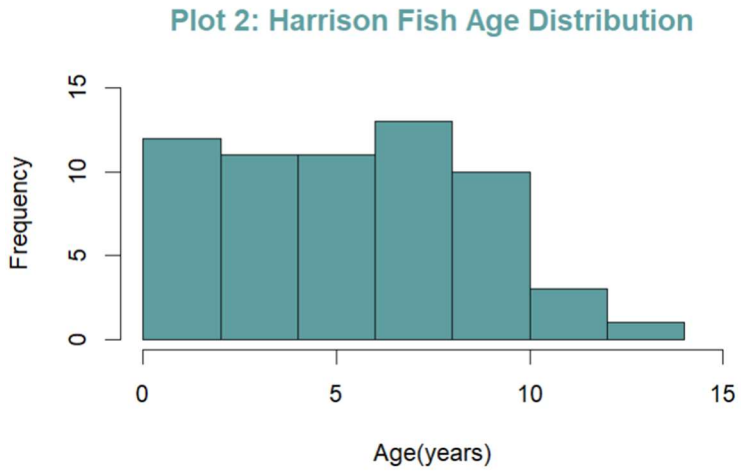
```
> #14. Create pchs vector
> pchs <- c("+","x")
> pchs
[1] "+" "x"
> #15. Create cols vector
> cols<-c("red", "gray60")
> cols
[1] "red"    "gray60"
> #17. Combine cols vector to tmp era values
> cols[tmp$era]
[1] "red"    "red"    "red"    "gray60" "gray60" "gray60"
>
```

- To understand the data properly, graphs are plotted to visualize the data clearly and have a better acknowledgement of the data.
  The different types of graphs are shown below:

  Plot 1:
  The following scatterplot depicts the graph between age and fork length of the distinct fishes over the years in Harrison Lake Trout.

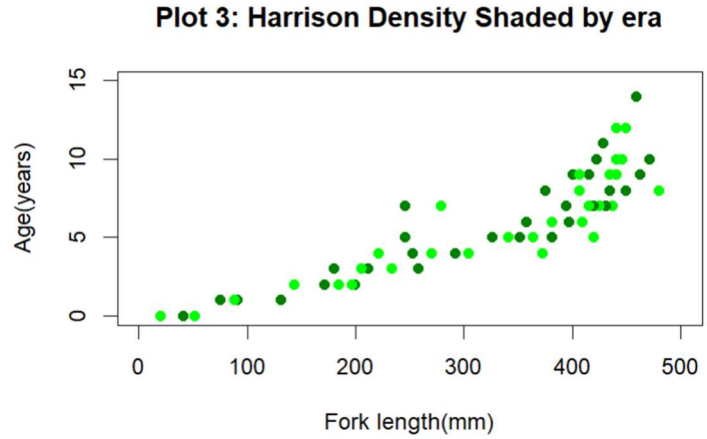


  Plot 2:
  This is a histogram of the age of the fish over the years with the frequency.



Plot 3:

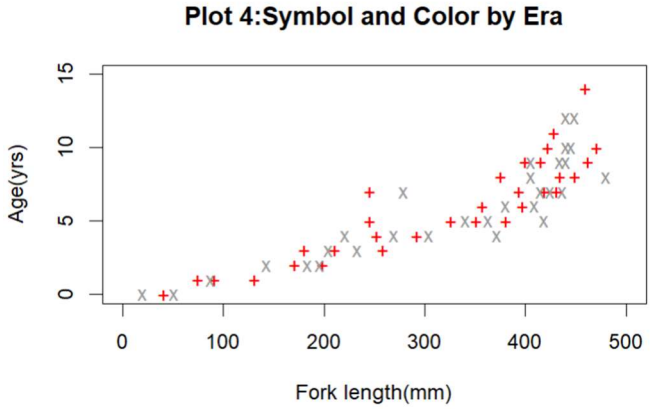
This graph is same as the first graph except for the colours by which the graph is shaded. There are two different shades to understand the two different eras.

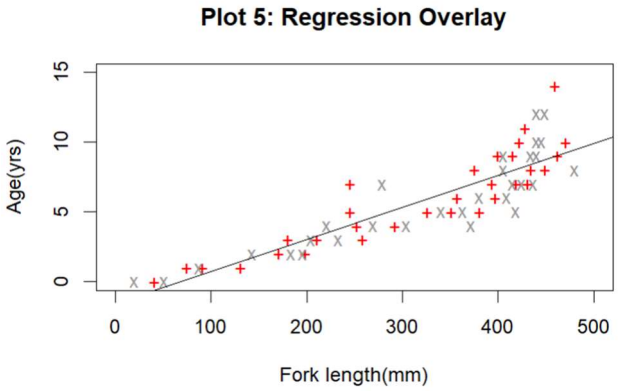


Plot 4 and 5:

In 4th graph different values are assigned to different eras for much better understanding and in 5th one, a regression line is added which gives comprehensive analysis of the graph.

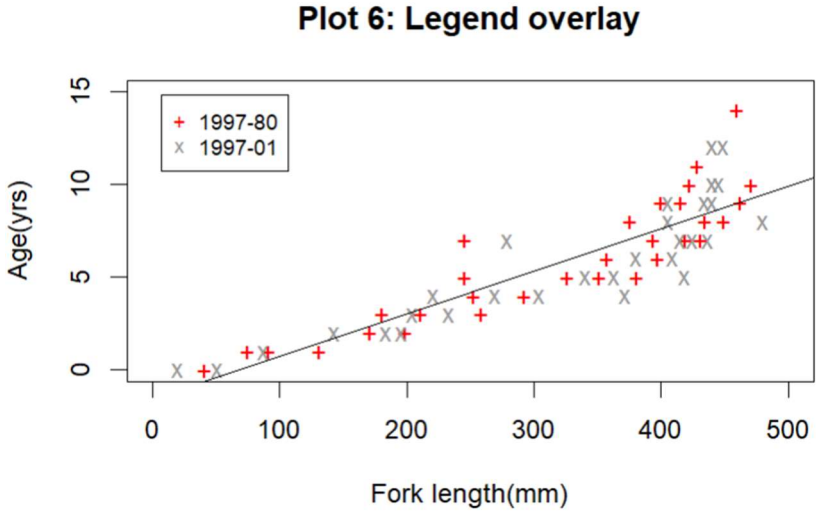
A regression line is used to predict the value of y for a given value of x.





- Plot 6:

  This is the final graph from which we can have the full understanding from the graph that fork length of fish increases with their age and even in different time era. Here we have mentioned the value of the dots which are plotted in the graph which are different timeframes.



Summary

- All the descriptive statistics which includes mean, median, variance etc are key findings in recognizing the pattern of the dataset.
- Scatterplots, histograms, frequency and probability distributions, bar plots (bar charts) are different types of graphs taken into consideration for explaining the visuals thoroughly.
- Visualization of data through R makes it so easy for analysts to get the clear picture of the dataset.
- In this assignment, the graphical representation of dataset gives a clear and proper understanding of the data.

Bibliography

- https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/filter
- https://libraryguides.mcgill.ca/c.php?g=699776&p=4968546
- https://www.statmethods.net/graphs/scatterplot.html
- https://www.datacamp.com/community/tutorials/make-histogram-basic-r
- https://www.geeksforgeeks.org/how-to-create-a-scatterplot-with-a-regression-line-in-r/
- http://www.sthda.com/english/wiki/add-legends-to-plots-in-r-software-the-easiest-way

APPENDIX

#name
print("Plotting Basics:Dhruvang Patel")

```
r=getOption("repos")
r["CRAN"]="http://cran.us.r-project.org"
options(repos=r)
install.packages("vcd")
library(vcd)

#install plyr package
install.packages("plyr")
library(plyr)

#install FSA package
install.packages("FSA")
library(FSA)

#install FSAdata package
install.packages("FSAdata")
library(FSAdata)

#install magrittr package
install.packages("magrittr")
library(magrittr)

#install dplyr package
install.packages("dplyr")
library(dplyr)

#install plotrix package
install.packages("plotrix")
library(plotrix)

#install ggplot2 package
install.packages("ggplot2")
library(ggplot2)

#install moments package
install.packages("moments")
library(moments)
```

```
#Load the dataset
data(BullTroutRML2)
BullTroutRML2

#4. Print first and last three records
head(BullTroutRML2,3)
tail(BullTroutRML2,3)

#5. Remove all except Harrison Lake
Harrisonlake<-filter(BullTroutRML2, lake=="Harrison")
Harrisonlake

#6. Display first and last 5 records of new dataset
#first 5
head(Harrisonlake,5)
#last 5
tail(Harrisonlake,5)

#7. Structure of a dataset
structure(Harrisonlake)

#8. Summary of a dataset
summary(Harrisonlake)

#9. Create a scatterplot with specifications
#assign values
fl<-Harrisonlake$fl
age<-Harrisonlake$age
#plot the data
par("mar")
par(mar=c(5.1,4.1,4.1,2.1))
plot(age~fl)
#plot with specifications
plot(age~fl,
    data = Harrisonlake, xlim=c(0,500), ylim=c(0,15),
    main="Plot 1: Harrison Lake Trout", xlab="fl", ylab="Age",
    pch=20)

#10. Plot a Histogram
hist(Harrisonlake$age,
    xlab = "Age(years)", ylab = "Frequency", main = "Plot 2: Harrison Fish Age Distribution",
    xlim=c(0,15), ylim=c(0,15),
    col = "cadetblue", col.main="cadetblue")
```

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

#12. New object tmp for first and last 3 records
tmp <- headtail(Harrisonlake,3)
tmp

#13. Display era column from tmp
tmp$era

#14. Create pchs vector
pchs <- c("+","x")
pchs

#15. Create cols vector
cols<-c("red", "gray60")
cols

#16. Convert era to numeric
tmp$era <- as.numeric(tmp$era)
tmp$era
is.numeric(tmp$era)

#17. Combine cols vector to tmp era values
cols[tmp$era]

#18. Create plot with specifications
par("mar")
par(mar=c(5,4,4,2))
plot(age~fl,
    data = Harrisonlake,
    main="Plot 4:Symbol and Color by Era",
    xlim=c(0,500),
    ylim=c(0,15),
    ylab="Age(yrs)",
    xlab = "Fork length(mm)",
```

```
    pch=pchs,
    col=cols)
```

#19. Plot regression line

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

#20. Placing a legend

```
plot(age~fl,
    data = Harrisonlake, main="Plot 6: Legend overlay",
    xlim=c(0,500), ylim=c(0,15),
    ylab="Age(yrs)", xlab = "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)
```