**M2-PROJECT 2**

**Name: Sunil Raj Thota**

**NUID: 001099670**

**Date: 10/03/2020**

**Title: M2 Project 2 Executive Summary Report**

**ALY 6000 - 71618 (CRN Number)**

**Prof. Dee Chiluiza Reyes**

**Introduction:**

R is a programming language is used for statistical analysis and creating graphs. Data objects are used for calculations. The scope of R is vast and mainly used in the fields of data mining, Regression, Probability, ML etc., by using various packages.

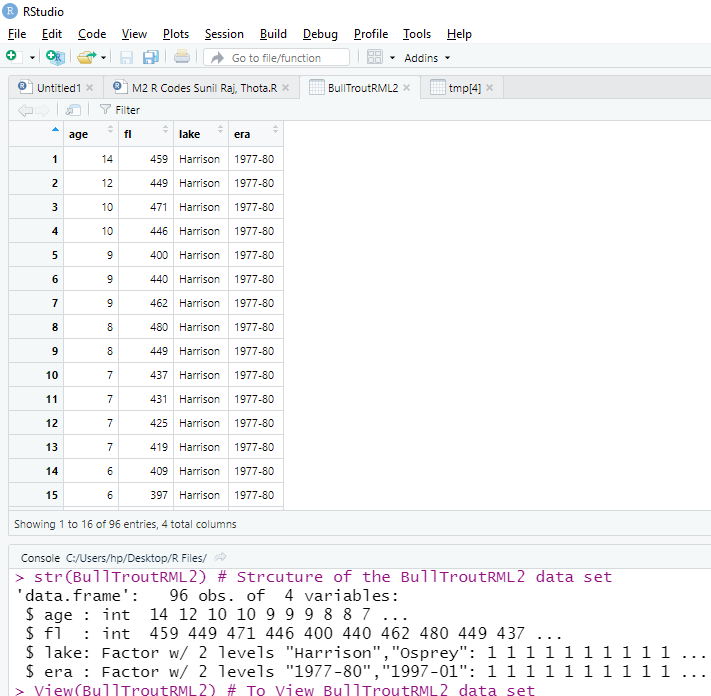
**Fisheries:**

* The effective management of fish populations requires knowledge of the growth rate of the fish. This requires determination of the age of fish to develop a relationship between the size and age of fish
* This is a prerequisite to generate the information on longevity, mortality and fluctuations in fisheries
* Such studies with proper statistical refinements are helpful in describing the present status and past history of a fish population along with the future course of the fishery

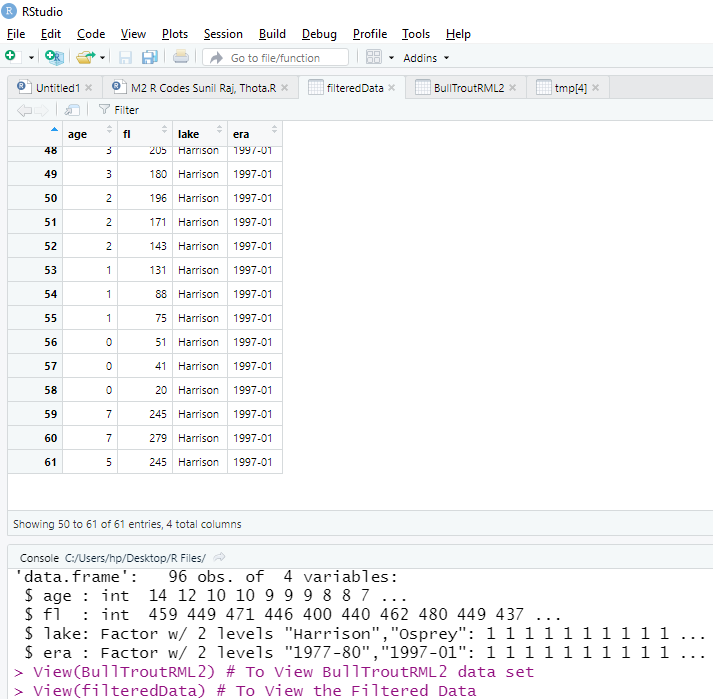
**Data Analysis:**

* The BullTroutRML2 data set showcases two different lakes called Harrison and Osprey in which the fisheries were maintained and cultivated.
* The data was framed in two timeframes; one in 1977 - 80 and the other in 1997 - 01
* The Age of the fishes is from 1 to 14 in years
* The fork length in mm is mostly ranged in between 250 to 420 for majority of the fishes
* There are 61 records of Harrison lake group and 35 records of Osprey lake group

**Full BullTroutRML2 data set**

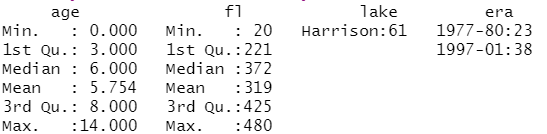


**Filtered Harrison - BullTroutRML2 data set**



**Descriptive Statistics:**

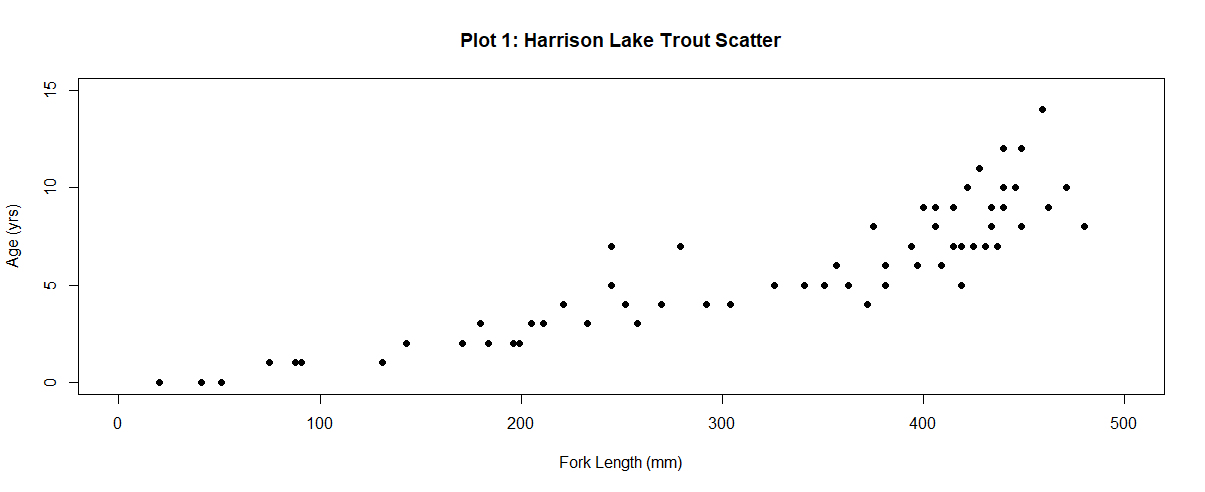
* Descriptive statistics is often the first step and an important part in any statistical analysis
* It allows to check the quality of the data and it helps to understand the data by having a clear overview of it
* It is already a good starting point for further analyses. There exist many measures to summarize a dataset
* They are divided into two types: location and dispersion measures
* Location measures give an understanding about the central tendency of the data
* Dispersion measures give an understanding about the spread of the data



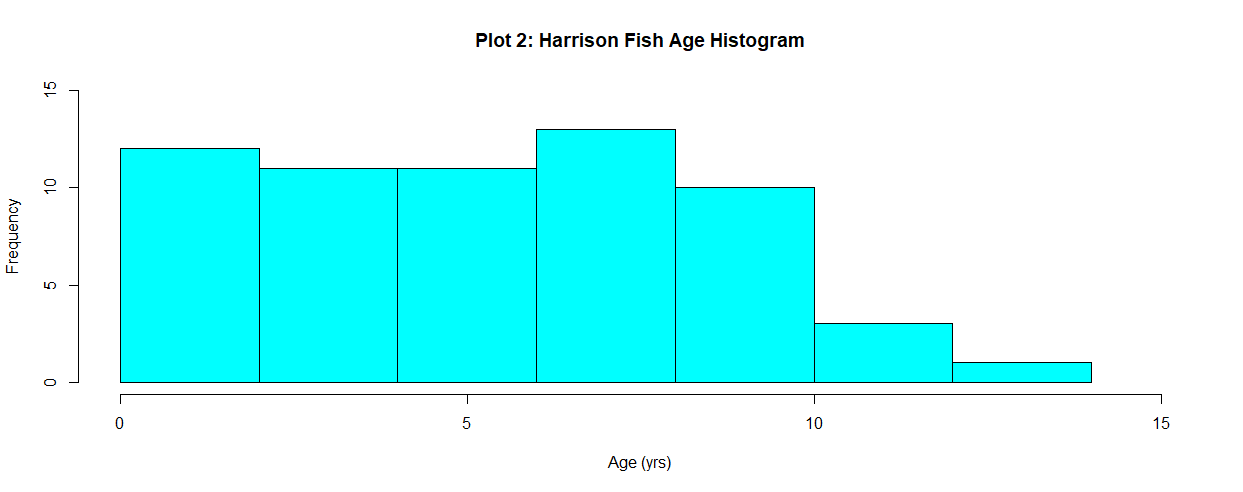
**R Applications:**

* Many data analysts and research programmers use R because R is the most prevalent language
* Hence, R is used as a fundamental tool for finance
* Many quantitative analysts use R as their programming tool
* Hence, R helps in data importing and cleaning, depending on what manner of strategy you are using on
* R is best for data Science because it gives a broad variety of statistics
* In addition, R provides the environment for statistical computing and design

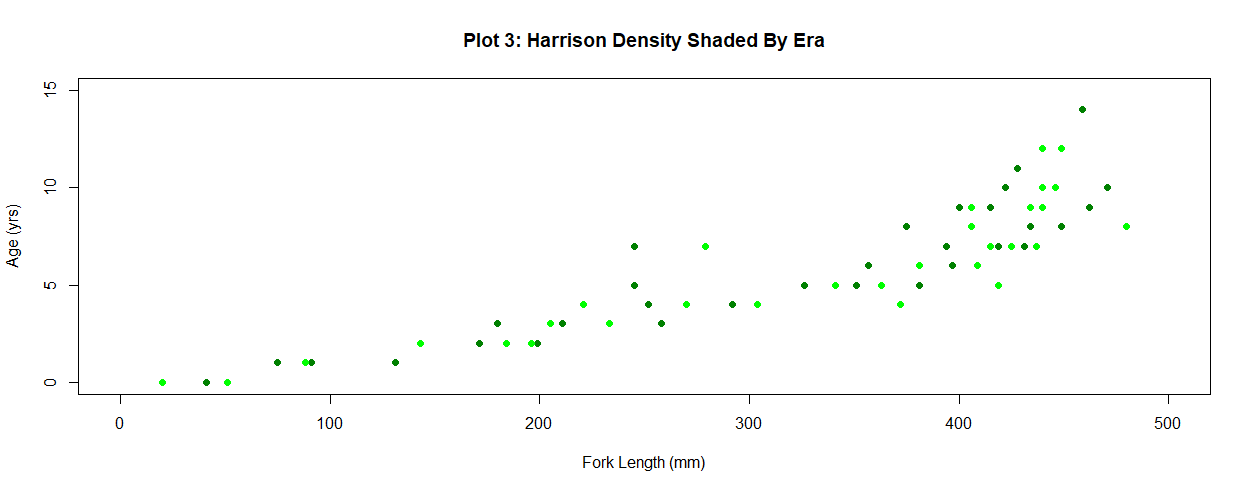
**Plot 1: Harrison Lake Trout Scatter Plot**

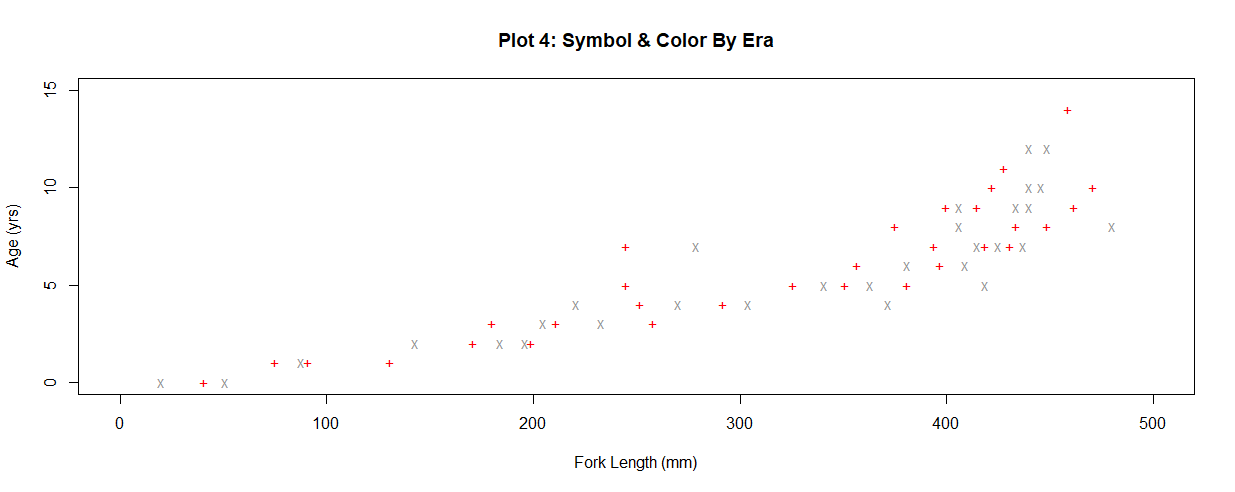


**Plot 2: Harrison Fish Age Histogram Plot**

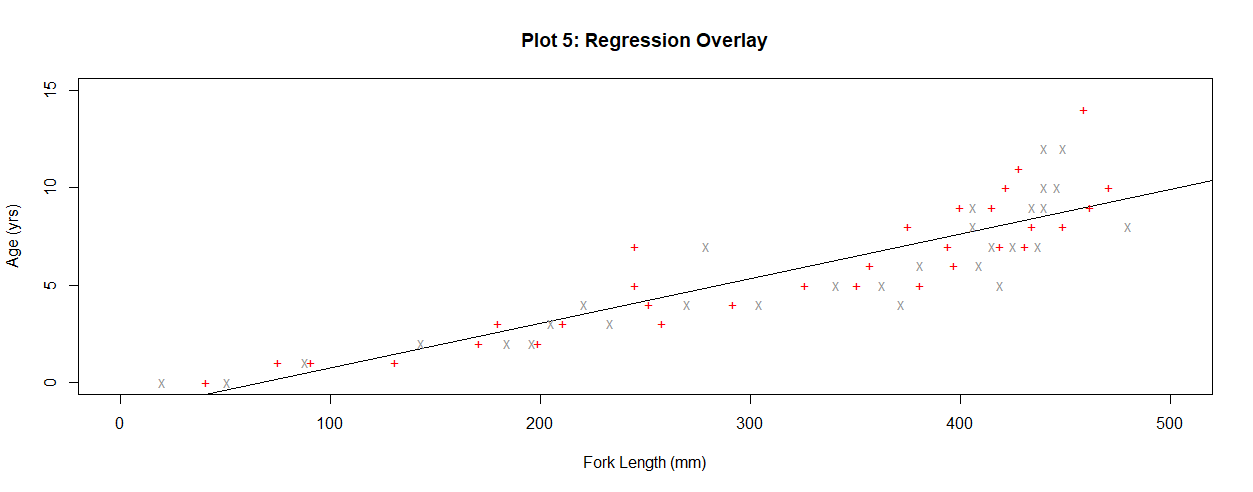


**Plot 3: Harrison Density Shaded by Era**

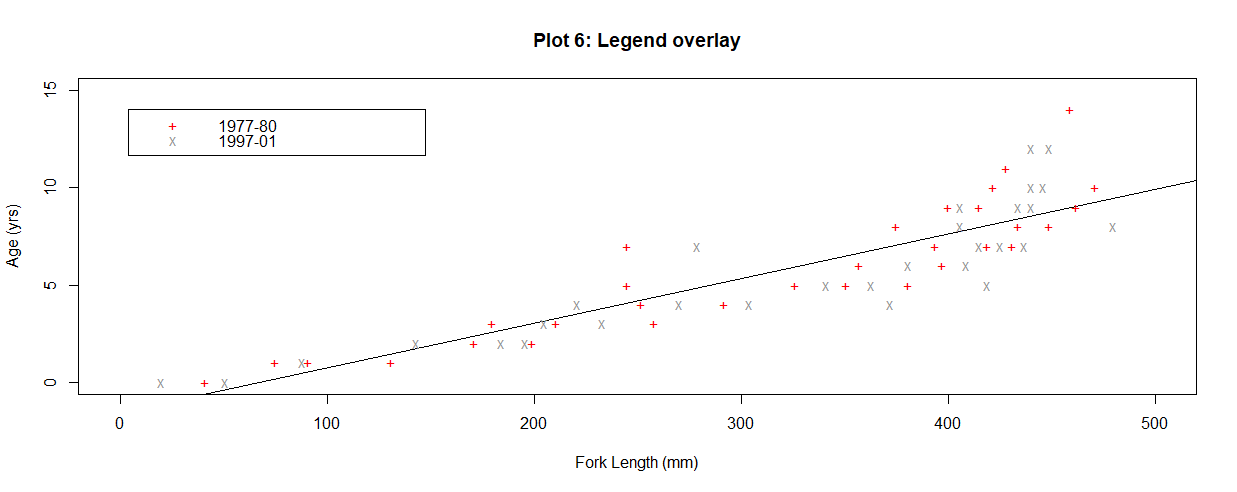


**Plot 4: Harrison Lake Trout Scatter Plot**

**Plot 5: Regression Overlay**



**Plot 6: Legend Overlay**



**Bibliography:**

1. <https://stackoverflow.com/questions/27688754/bar-chart-legend-position-avoiding-operlap-in-r>
2. <https://www.r-project.org/doc/bib/R-books.html>
3. <https://richardlent.github.io/post/rstudio-as-a-research-and-writing-platform/>
4. <https://www.rdocumentation.org/packages/FSAdata/versions/0.3.8/topics/BullTroutRML2>
5. <https://www.r-graph-gallery.com/>
6. <https://stackoverflow.com/questions/3932038/plot-a-legend-outside-of-the-plotting-area-in-base-graphics>
7. <https://www.dataanalytics.org.uk/legends-on-graphs-and-charts/>
8. <https://www.statmethods.net/graphs/bar.html>

**Appendix**

# ALY 6000 - 71618 (CRN Number)

# Module 2 Project

# 10/03/2020

# Sunil Raj Thota

# NUID: 001099670

#---------------------------------------------#

# Q1: Print "Plotting Basics: Your Last Name"

# "Plotting Basics: Thota"

#---------------------------------------------#

# Q2: Import Libraries: FSA, FSAdata, magrittr, dplyr, plotrix, ggplot2, moments

install.packages("FSA")

install.packages("FSAdata")

install.packages("magrittr")

install.packages("dplyr")

install.packages("plotrix")

install.packages("ggplot2")

install.packages("moments")

library(FSA)

library(FSAdata)

library(magrittr)

library(ggplot2)

library(dplyr)

library(plotrix)

library(moments)

# Imported the necessary libraries into the workspace

#---------------------------------------------#

# Q3: Load the data set "BullTroutRML2.csv"

data(BullTroutRML2) # Loaded the BullTroutRML2 data set

str(BullTroutRML2) # Strcuture of the BullTroutRML2 data set

View(BullTroutRML2) # To View BullTroutRML2 data set

#---------------------------------------------#

# Q4: Print first and last 3 records dataset BullTroutRML2

firstLastThreeRecs <- headtail(BullTroutRML2, 3)

firstLastThreeRecs # Printed first and last 3 records of BullTroutRML2 data set

#---------------------------------------------#

# Q5: Remove all the records from BullTroutRML2. Except those from Harrison Lake

filteredData <- filterD(BullTroutRML2, lake == "Harrison")

filteredData # Removed all the records form BullTroutRML2 data set except those from harrison Lake

View(filteredData) # To View the Filtered Data

#---------------------------------------------#

# Q6: Display again the first and last 5 records from data set BullTroutRML2

firstLastFiveRecs <- headtail(filteredData, 5)

firstLastFiveRecs # Printed first and last 5 records of filtered BullTroutRML2 data set

#---------------------------------------------#

# Q7: Display the structure of the filtered BullTroutRML2 dataset

str(filteredData) # Structure of filtered BullTroutRML2 data set

#---------------------------------------------#

# Q8: Display the summary of the filtered BullTroutRML2 dataset

summary(filteredData) # Summary of filtered BullTroutRML2 data set

#---------------------------------------------#

# Q9: Plot 1: Harrison Lake Trout Scatter Plot

plot\_1 <- plot(

filteredData$fl,

filteredData$age,

xlim = c(0, 500),

ylim = c(0, 15),

main = "Plot 1: Harrison Lake Trout Scatter",

xlab = "Fork Length (mm)",

ylab = "Age (yrs)",

pch = 16

) # Plot 1: Scatter Plot of Harrison lake Trout

#---------------------------------------------#

# Q10: Plot 2: Harrison Fish Age Histogram

plot\_2 <- hist(

filteredData$age,

xlab = "Age (yrs)",

ylab = "Frequency",

main = "Plot 2: Harrison Fish Age Histogram",

xlim = c(0, 15),

ylim = c(0, 15),

col = "cyan"

) # Plot 2: Histogram Plot of Harrison Fish Age

#---------------------------------------------#

# Q11: Plot 3: Harrison Density Shaded By #Era

Plot\_3 <- plot(

filteredData$fl,

filteredData$age,

xlim = c(0, 500),

ylim = c(0, 15),

main = "Plot 3: Harrison Density Shaded By Era",

xlab = "Fork Length (mm)",

ylab = "Age (yrs)",

pch = 16,

col = rgb(0, (1:2) / 2, 0)

) # Plot 3: Scatter Plot of Harrison Density shaded by Era

#---------------------------------------------#

# Q12: Create tmp object with the first 3 and last 3 records of BullTroutRML2

tmp <- headtail(filteredData, 3)

tmp # Created tmp object with the first 3 and last 3 records of filtered BullTroutRML2 data set

#---------------------------------------------#

# Q13: Display the "era" column (variable) of the tmp object

displayTmp <- tmp[, c("era"), drop = FALSE]

displayTmp # Displayed the era column of the tmp object

#---------------------------------------------#

# Q14: Create a pchs vector with numerical arguments for + and x

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

pchs # Created a pchs vector with numerical arguments for + and x

#---------------------------------------------#

# Q15: Create a cols vector with the two elements: "red" and "gray60"

cols <- c("red", "gray60")

cols # Created a cols vector with the two elements: "red" and "gray60"

#---------------------------------------------#

# Q16: Convert the tmp era values to numeric

tmpAsNumeric <- as.numeric(sub("-", "", tmp$era))

tmpAsNumeric # Converted the tmp era values to numeric

#---------------------------------------------#

# Q17: Initialize the cols vector with tmp era values

initCols <- c(tmp$era)

initCols # Initialized the cols vector with the tmp era values

#---------------------------------------------#

# Q18: Plot 4: Symbol & #Color By Era

plot\_4 <- plot(

filteredData$fl,

filteredData$age,

xlim = c(0, 500),

ylim = c(0, 15),

main = "Plot 4: Symbol & Color By Era",

xlab = "Fork Length (mm)",

ylab = "Age (yrs)",

pch = pchs,

col = cols

) # Plot 4: Scatter Plot of Symbol & Color Era

#---------------------------------------------#

# Q19: PLOT 5: Regression Overlay

plot\_5 <- plot(

filteredData$fl,

filteredData$age,

xlim = c(0, 500),

ylim = c(0, 15),

main = "Plot 5: Regression Overlay",

xlab = "Fork Length (mm)",

ylab = "Age (yrs)",

pch = pchs,

col = cols,

# Regression Line

abline(lm(filteredData$age ~ filteredData$fl), col = "black")

) # Plot 5: Scatter Plot with Regression Overlay

#---------------------------------------------#

# Q20: Plot 6: Legend overlay

plot\_6 <- plot(

filteredData$fl,

filteredData$age,

xlim = c(0, 500),

ylim = c(0, 15),

main = "Plot 6: Legend overlay",

xlab = "Fork Length (mm)",

ylab = "Age (yrs)",

pch = pchs,

col = cols,

# Regression Line

abline(lm(filteredData$age ~ filteredData$fl), col = "black"),

) # Plot 6: Scatter Plot with legend Overlay

# Legend to indicate the Age in allocated colors

legend(

4,

14,

legend = c("1977-80", "1997-01"),

col = cols,

pch = pchs,

cex = 1

)

#---------------------------------------------#