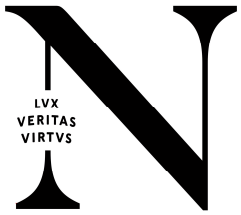


Northeastern University

Toronto, Canada

Executive Summary of module 2

Introduction to Analytics

(ALY 6000)

Submitted by

Name: Parva Patel

NUID: 002195186

Date: 28th January, 2022

Guided By

Prof. Mohammad Shafiqul Islam

Key Findings

- This paper is about a BullTroutRML2 dataset that was analysed. When we add the FSA and FSAdata libraries, the BullTroutRML2 dataset is imported. This dataset contains data on the ages and fork lengths of Bull Trout from two Rocky Mountain (Harrison and Osprey) lakes in Alberta, Canada before and after a regulatory change.

- The relevant libraries FSA, FSAdata, magrittr, dplyr, plotrix, ggplot2, and moments will provide many crucial functions required for data analysis.

- To perform analysis on data linked to Harrison Lake, we must first generate a subset of the dataset that only contains data about Harrison Lake. filter() is a function that will assist us in obtaining our desired data. As a result, the programme to execute

  ```
  Harrisonlake<-filter(BullTroutRML2, lake=="Harrison")
  Harrisonlake
  ```

- Performing analysis on the entire data set is not always essential, and we prefer to analyse specific data. This goal can be accomplished by building an object that can store certain information. For example, if I want to create a distinct collection of data for the first and last three rows of primary data, I'll construct an object and insert those records into it.

  Input:

  ```
  tmp <- headtail(Harrisonlake,3)
  ```

  Output :

  ```
  >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
  ```

- Different types of vectors are also formed, and those vectors can be injected into the data values.

Input :

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

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

Output :

```
>pchs
[1] "+" "x"

>cols
[1] "red"    "gray60"
```

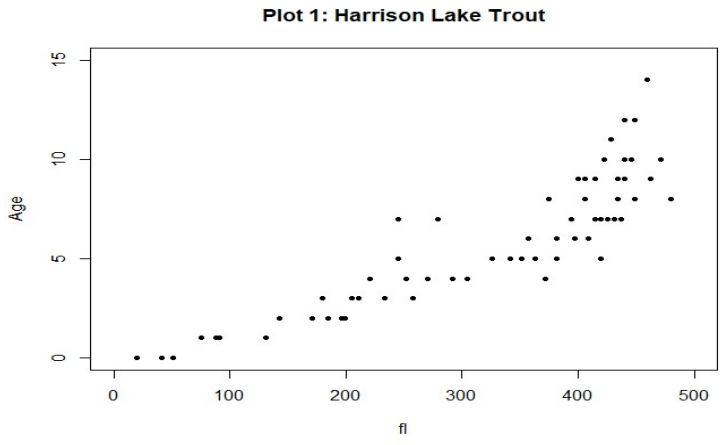
- The $ symbol can be used to get a single variable from a specific set of data. For example : tmp$era

- Graphs are plotted to show the data clearly and have a better understanding of the data in order to grasp it effectively.

  Several graph types are illustrated below:

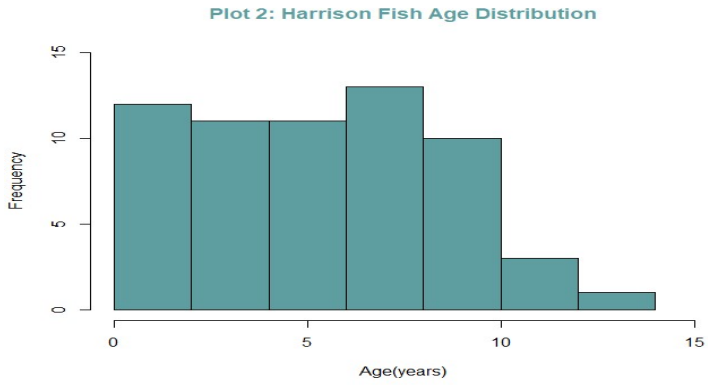
  Graph 1 :

  The link between two types of data is depicted using a scatter diagram. The age of fish and their fork length are plotted here, and we may deduce that folk length grows with age.



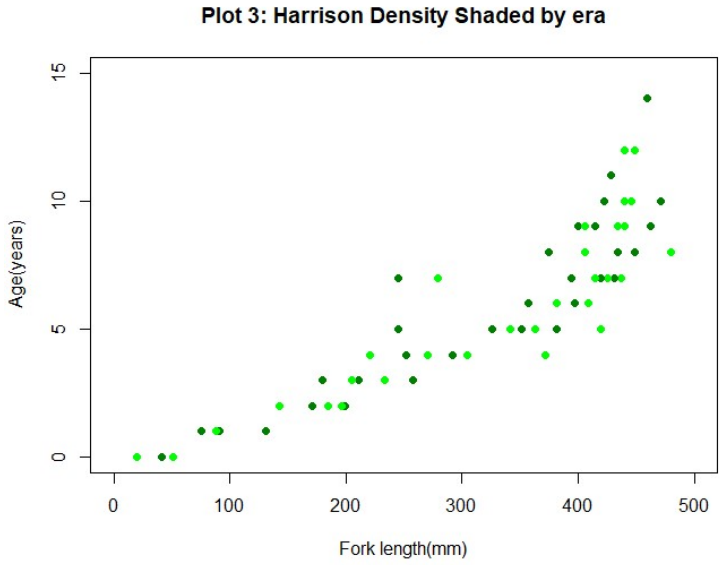
Graph 2 : Histogram

The frequency of the fish's age utilised in this data is summarised by plotting age in a histogram.



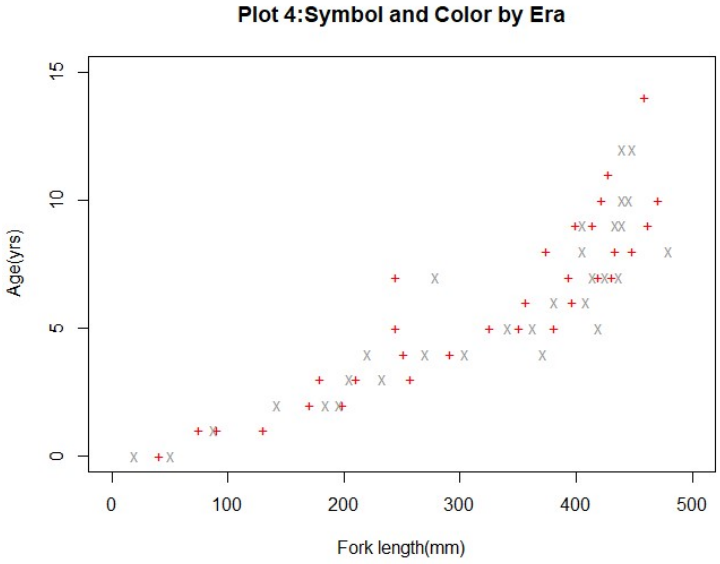
Graph 3 : Scatter plot

The first scatter plot and this plot have the same content. The difference between them is the use of various colours in plotting the graph, which aids in understanding the two eras.



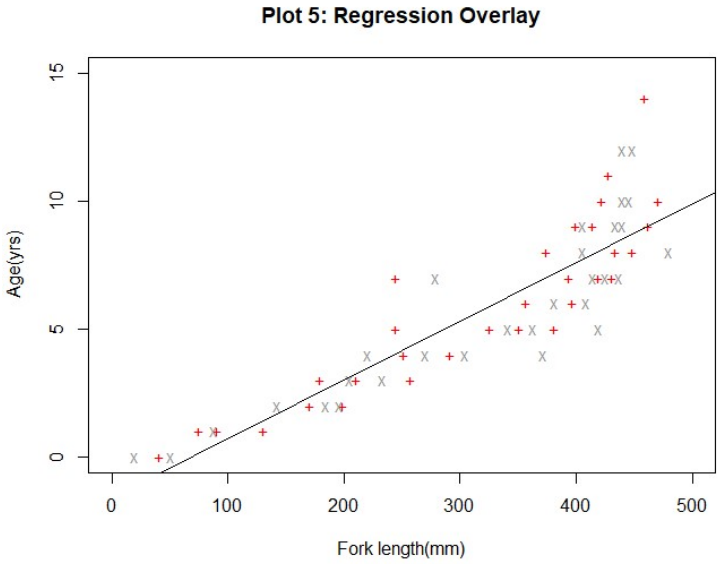
Graph 4 :

To become more comfortable with the data in this graph, various values have been allocated to different eras, which will aid in better data analysis.



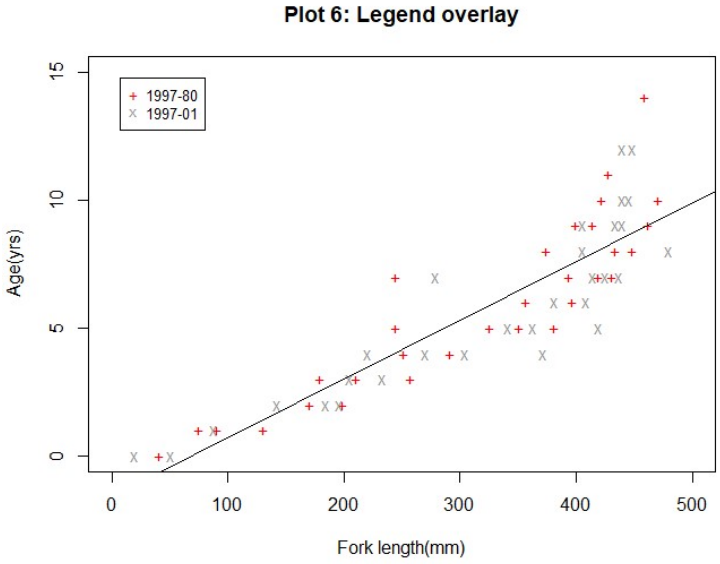
Graph 5 :

The regression line can be added to the scatter plot to highlight the overall trend of a group of data. When the value of x is known, it is simple to calculate the value of y using a regression line.



Graph 6 :

A complete scatter plot is depicted in this graph. It's the same graph as the last one, but this time we've included the values for which the dots in the graph are plotted. As a result of this graph, it is clear that the length of a fish's fork increases with age and even across time periods.


Plot 6: Legend overlay
1997-80
1997-01
Age(yrs)
Fork length(mm)
0
5
10
15
0
100
200
300
400
500

Summary

- This project focuses on the various forms of graphs that can be used to obtain relevant information and a thorough comprehension of the data.

- To depict data or extract information from a data collection, graphs like scatterplots, histograms, frequency and probability distributions, and bar plots (bar charts) are used.

- How to work with enormous data sets by creating subsets or extracting data that can predict for the entire data set, making data analysis easier

- All descriptive statistics, such as mean, median, and variance, are useful in deciphering the pattern of the dataset.

- The usage of R makes it very simple for analysts to extract relevant information from a dataset by visualising it.

Bibliography

- **https://www.rstudio.com/products/rpackages/**
- **https://dplyr.tidyverse.org/reference/filter.html**
- **https://r-lang.com/as-numeric-r/**
- **https://r-coder.com/scatter-plot-r/**
- **https://matplotlib.org/stable/gallery/lines_bars_and_markers/scatter_with_legend.html**

My GitHub repository : https://github.com/ParvaPatel10/Module2

Appendix

 # 1. Name

```
print("Plotting Basics:Parva Patel")

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

 # 2. Install plyr package

```
install.packages("plyr")

library(plyr)
```

 # Install dplyr package

```
install.packages("dplyr")

library(dplyr)
```

 # 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 plotrix package

```
install.packages("plotrix")

library(plotrix)
```

 # install ggplot2 package

```
install.packages("ggplot2")

library(ggplot2)
```

 # install moments package

```
install.packages("moments")

library(moments)
```

 # 3. Load the dataset

```
data(BullTroutRML2)

BullTroutRML2
```

#4. Print first and last three records

```
# First 5
head(BullTroutRML2,3)

# Last 5
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),
    xlab="Fork length(mm)",
    xlim=c(0,500),
    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. pchs vector

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

#15. 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)",

    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)
```