Executive Summary Report 2

Data Description with R

ALY6000: Introduction to Analytics

Prepared by: Heejae Roh
Presented to: Professor Behzad Ahmadi

Date: Oct 3rd, 2022

Summary

I think visualizing data is a great help for understanding data at once and could have a big impact on explanation and persuasion. In data 'BullTroutRML2', Harrison Lake Trout grows in length as age adds up. Those from '1997-01' were shorter than those from '1977-80' in the same y (age) variable (Plot 3). When looking at the entire data, the length also gradually increases with age (Plot 4). The median Fork Length of whole data is 352.5, the mean is 326.1, Q1 is 258, and Q3 is 406 (Plot 5).

1-2. Name & Import libraries

```
#Plotting Basics by Heejae Roh#
install.packages("plyr")
install.packages("FSA")
install.packages("FSAdata")
install.packages("magrittr")
install.packages("moments")
install.packages("berryFunctions")
library(plyr)
library(FSA)
library(FSAdata)
library(magrittr)
library(dplyr)
library(plotrix)
library(ggplot2)
library(moments)
```

1-2

3-4. Load BullTroutRML2 & Print 1st and last 3 records

```
> setwd("C:\\Users\\14083\\Desktop\\exacutive summary\\Project 2")
> BullTroutRML2 <- read.csv("BullTroutRML2.csv", header=TRUE)
> head(BullTroutRML2,1)
  age  fl     lake     era
1  14 459 Harrison 1977-80
> tail(BullTroutRML2,3)
   age  fl   lake     era
94   4 298 Osprey 1997-01
95   3 279 Osprey 1997-01
96   3 273 Osprey 1997-01
```

3-4

5-6. Filter out all records except Harrison& Display first and last 3 records (left, below)

```
> filtered.Bull <- BullTroutRML2[BullTroutRML2$lake %in% "Harrison", ]
> head(filtered.Bull,1)
  age  fl     lake     era
1  14 459 Harrison 1977-80
> tail(filtered.Bull,3)
   age  fl     lake     era
59   7 245 Harrison 1997-01
60   7 279 Harrison 1997-01
61   5 245 Harrison 1997-01
```

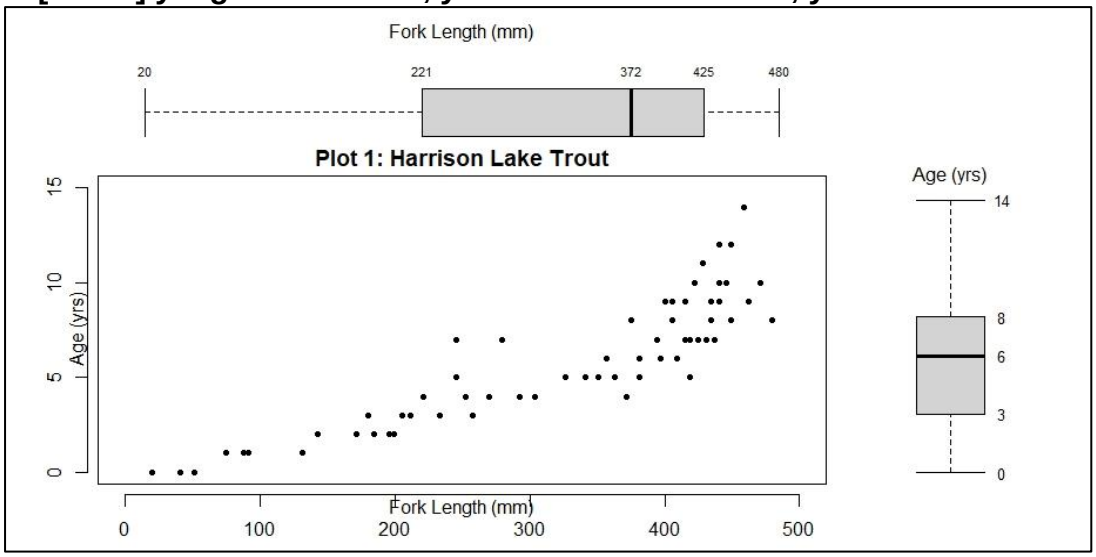
5-6

7-8. Display Structure of Filtered dataset & Summary it as <t> (right)

```
> str(filtered.Bull)
'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" ...
> t <- summary(filtered.Bull)
> 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
```

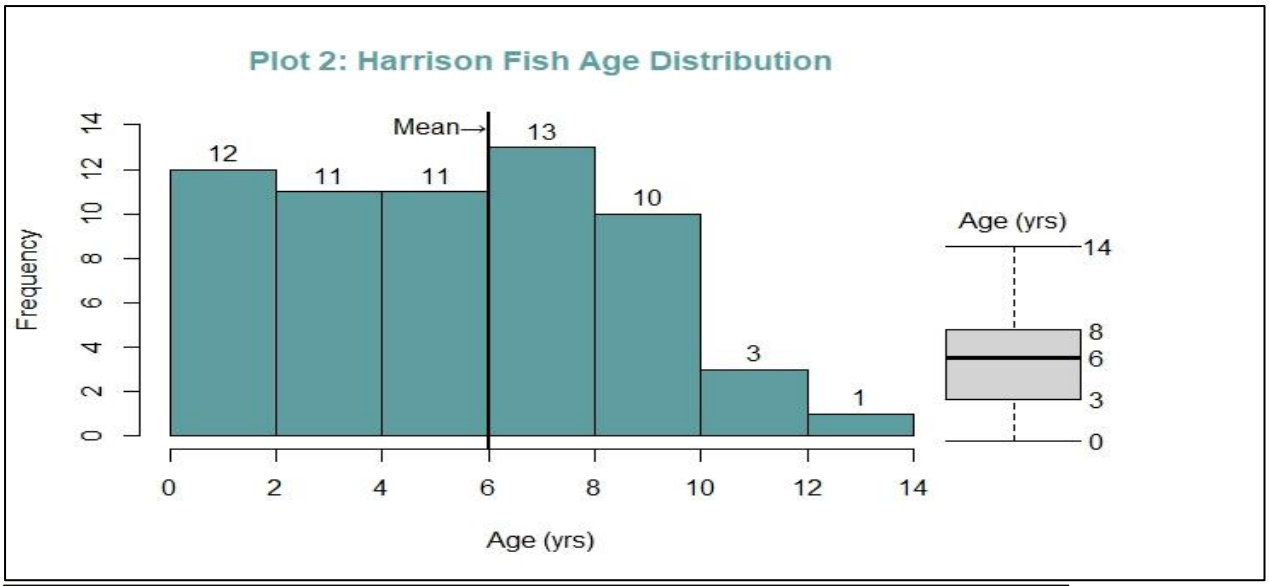
7-8

9. [Plot 1] y=age x=fl. limit x, y/ title/ axis labels for x, y/ small filled circle



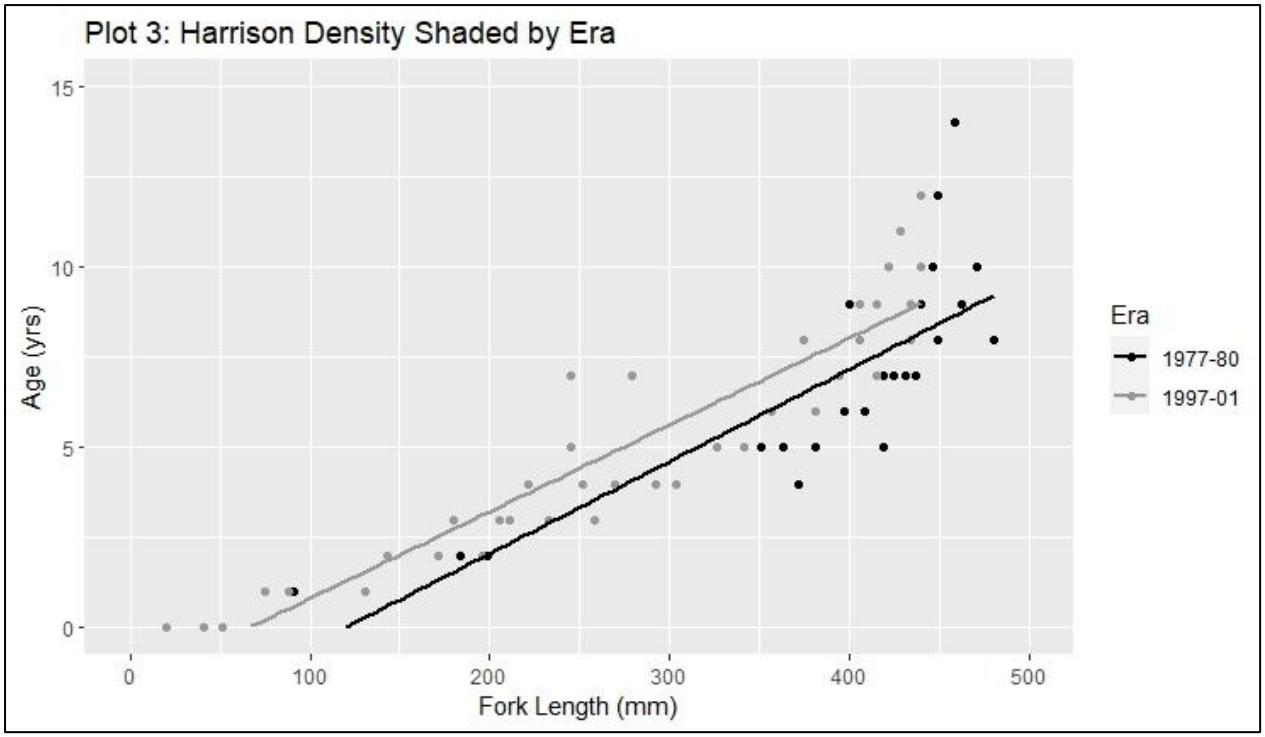
*Add **two Boxplots of Fork Length & Age** for enhancing understand of data*

10. [Plot 2] "Age" histogram. Axis Labels x, y/ Title/ Color frequency, title



*Add **Mean line** on the histogram graph and add **Boxplot of Age***

11. [Overdense Plot 3] Two black shading levels/ Title



*Add **ablines to compare 1977-80 with 1997-01** About Fork Length in the same age (y variable)*

12-13. tmp including 1st and last 3 records of whole/ Display era column (left, below)

```
> library(berryFunctions)
> tmp <- headtail(BullTroutRML2,3)
> tmp
   age  fl     lake     era
1   14 459 Harrison 1977-80
2   12 449 Harrison 1977-80
3   10 471 Harrison 1977-80
94   4 298   Osprey 1997-01
95   3 279   Osprey 1997-01
96   3 273   Osprey 1997-01
> data.frame(tmp$era)
  tmp.era
1 1977-80
2 1977-80
3 1977-80
4 1997-01
5 1997-01
6 1997-01
```

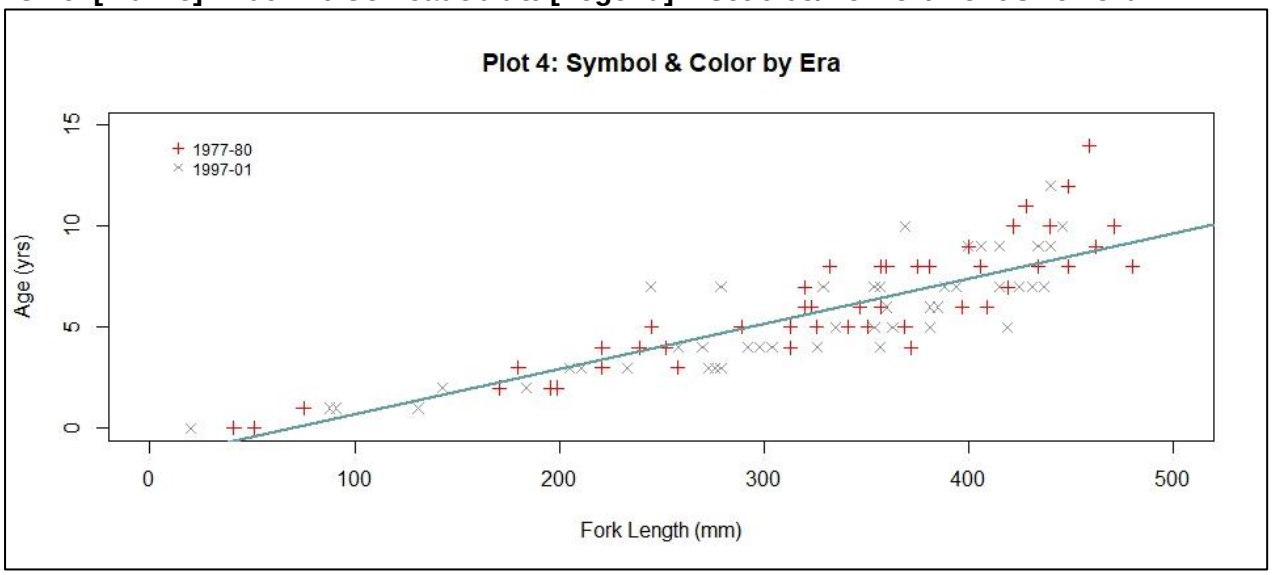
12-13

```
> pchs <- c(3,4)
> cols <- c("red","gray60")
> class(tmp$era)
[1] "character"
> levels(tmp$era)
NULL
> era.tmp <- factor(c("1977-80", "1977-80", "1977-80", "1997-01",
 "1997-01", "1997-01"), levels=c("1977-80", "1997-01"))
> numEra <- as.numeric(era.tmp)
> numEra
[1] 1 1 1 2 2 2
> cols[numEra]
[1] "red"   "red"   "red"   "gray60" "gray60" "gray60"
```

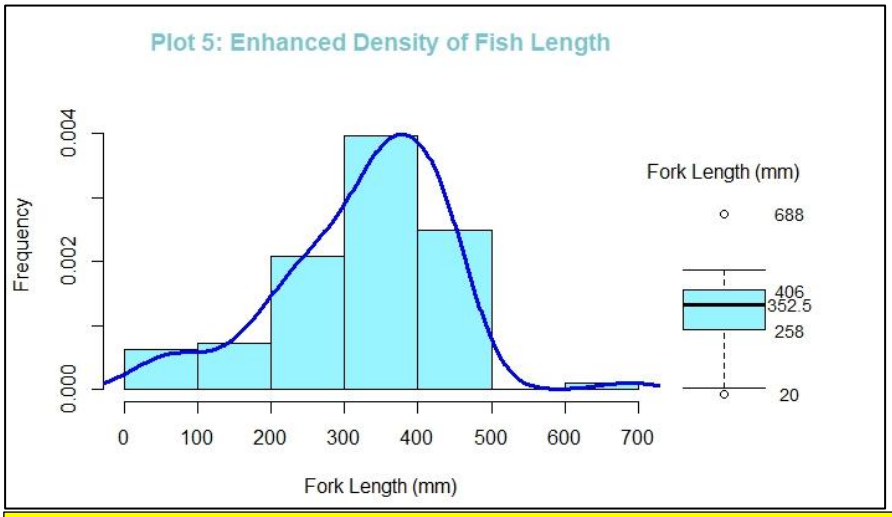
14-16

14-16. Create a pchs values for + & x and create cols vector/ Convert tmp to numeric/ Create numeric numEra from tmp$era/ Associate col vec with tmp$era (right, above)

17. [Plot 4] y=age x=fl/ Limit x, y/ Title/ Axis Label x, y// pch&col = pchs & cols value
18-19. [Abline] Width=2/ Col=cadeblue/ [Legend] Inset 0.05/No Box/ Font size 75%



+ [Plot 5] Add more. Enhanced Density of Fish Length



Shows Density of Fish Length, with lines with Boxplot.
Fork Length ***Median:352.5, Q1: 258, Q3: 406, one outlier at each side***

Bibliography

Kabacoff. Robert. (2015). R In Action: Data Analysis and Graphics with R. Manning.

Bluman, A. (2017). *Elementary Statistics: A Step By Step Approach* (10th ed.). McGraw-Hill Higher Education (US). https://reader2.yuzu.com/books/9781260042054

Alpha. (2013, Jun 13). Colorize parts of the title in a plot. Stackoverflow. Retrieved from https://stackoverflow.com/questions/17083362/colorize-parts-of-the-title-in-a-plot

joran. (2012, Oct 23). Avoiding NAs in as.numeric(). Stackoverflow. Retrieved from https://stackoverflow.com/questions/13022234/avoiding-nas-in-as-numeric

DATAMENTOR. (n.d). R Factors. DATAMENTOR. Retrieved from https://www.datamentor.io/r-programming/factor/

Peter Dalgaard. (2018, Jun 18). [R] Line width in graphs. n.d. Retrieved from https://stat.ethz.ch/pipermail/r-help/2002-June/022513.html

Zach. (2021, April 21). How to Create Horizontal Boxplots in R. STATOLOGY. Retrieved from https://www.statology.org/horizontal-boxplot-in-r/

eipi10. (2015, Sep 10). Explain ggplot2 warning: "Removed k rows containing missing values". Stackoverflow. Retrieved from https://stackoverflow.com/questions/32505298/explain-ggplot2-warning-removed-k-rows-containing-missing-values

mpalanco. (2016, Oct 19). How to put values on a boxplot for median. 1st quartile and last quartile?. Stackoverflow. Retrieved from https://stackoverflow.com/questions/13945434/how-to-put-values-on-a-boxplot-for-median-1st-quartile-and-last-quartile

STHDA. (n.d). abline R function : An easy way to add straight lines to a plot using R software. STHDA. Retrieved from http://www.sthda.com/english/wiki/abline-r-function-an-easy-way-to-add-straight-lines-to-a-plot-using-r-software

R-bloggers. (2012, Sep 27). Histogram + Density Plot Combo in R. R-bloggers. Retrieved from https://www.r-bloggers.com/2012/09/histogram-density-plot-combo-in-r/

Appendix: The R Script

```
#Plotting Basics by Heejae Roh#
install.packages("plyr")
install.packages("FSA")
install.packages("FSAdata")
install.packages("magrittr")
install.packages("moments")
install.packages("berryFunctions")
library(plyr)
library(FSA)
library(FSAdata)
library(magrittr)
library(dplyr)
library(plotrix)
library(ggplot2)
library(moments)

setwd("C:\\Users\\14083\\Desktop\\exacutive summary\\Project 2")
BullTroutRML2 <- read.csv("BullTroutRML2.csv", header=TRUE)
BullTroutRML2
head(BullTroutRML2,1)
tail(BullTroutRML2,3)
filtered.Bull <- BullTroutRML2[BullTroutRML2$lake %in% "Harrison", ]
filtered.Bull
str(filtered.Bull)
t <- summary(filtered.Bull)
t

attach(filtered.Bull)
opar <- par(no.readonly = TRUE)
par(fig=c(0, 0.8, 0, 0.8))
plot(fl, age, main = "Plot 1: Harrison Lake Trout", line=0.5, ylab="Age (yrs)", xlab="Fork
Length (mm)", ylim=c(0,15), xlim=c(0,500), pch=20)
par(fig=c(0.02, 0.78, 0.48, 1), new=TRUE)
boxplot(fl, horizontal = TRUE, axes=FALSE, staplewex=1)
mtext("Fork Length(mm)", side=3, line=0.3)
text(x=fivenum(fl), labels=fivenum(fl), y=1.35, cex=0.7)
par(fig=c(0.7, 1, 0, 0.78), new=TRUE)
boxplot(age, axes=FALSE, staplewex=1)
mtext("Age (yrs)", side=3, line=0.2)
text(y=fivenum(age), labels=fivenum(age), x=1.3, cex=0.8)
par(opar)

opar <- par(no.readonly = TRUE)
par(fig=c(0, 0.8, 0, 1))
hist(age, main="Plot 2: Harrison Fish Age Distribution", col.main="cadetblue",
ylab="Frequency", xlab="Age (yrs)", col="cadetblue", labels=TRUE, ylim=c(0,14))
```

```
abline(v=6, col="black", lwd=2)
text(5.1,14, "Mean→")
par(fig=c(0.55, 1, 0, 0.78), new=TRUE)
boxplot(age, axes=FALSE, staplewex=1)
mtext("Age (yrs)", side=3, line=0.2)
text(y=fivenum(age), labels=fivenum(age), x=1.25, cex=1)
par(opar)

filtered.Bull%>%ggplot(aes(fl,age, color=era))+geom_point(pch=19)+ggtitle("Plot 3: Harrison Density Shaded by Era")+ylab("Age (yrs)")+xlab("Fork Length (mm)")+xlim(0,500)+ylim(0,15)+scale_color_manual(values=c("#000000","#999999"))+labs(color="Era")+geom_smooth(method=lm, se=FALSE)
detach(filtered.Bull)

library(berryFunctions)
tmp <- headtail(BullTroutRML2,3)
tmp

data.frame(tmp$era)
pchs <- c(3,4)
cols <- c("red","gray60")
class(tmp$era)
levels(tmp$era)
era.tmp <- factor(c("1977-80", "1977-80", "1977-80", "1997-01", "1997-01", "1997-01"), levels=c("1977-80", "1997-01"))
numEra <- as.numeric(era.tmp)
numEra
cols[numEra]

attach(BullTroutRML2)
plot(fl,age, main = "Plot 4: Symbol & Color by Era", ylab="Age (yrs)", xlab="Fork Length(mm)", ylim=c(0,15), xlim=c(0,500), pch=pchs[numEra], col=cols[numEra])
abline(lm(age ~ fl), lwd=2, col="cadetblue")
legend("topleft", inset = 0.05, c("1977-80", "1997-01"), col=cols, pch=pchs, box.col="white", cex=0.75)

opar <- par(no.readonly = TRUE)
par(fig=c(0, 0.8, 0, 1))
hist(fl, freq=F, main="Plot 5: Enhanced Density of Fish Length", col.main="cadetblue3", xlab="Fork Length (mm)", ylab="Frequency", col="cadetblue1", ylim=c(0,0.0045))
lines(density(fl), col="blue", lwd=3)
par(fig=c(0.58, 1, 0, 0.78), new=TRUE)
boxplot(fl, axes=FALSE, staplewex=1, col="cadetblue1")
mtext("Fork Length (mm)", side=3, line=1)
text(y=fivenum(fl), labels=fivenum(fl), x=1.32, cex=0.85)
par(opar)
detach(BullTroutRML2)
```