Name: Yijun Wang

Class: Aly 6000

Date: 10/04/2021

Title: Module 2 Project — Executive Summary Report 2

Key finding

A. Below is the summary of Lakefilter.

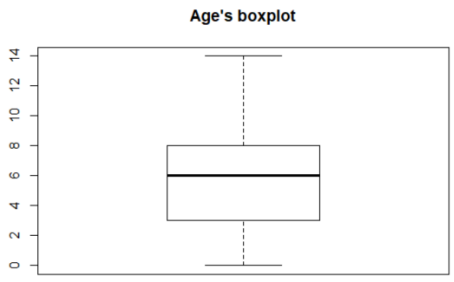
```
> summary(Lakefilter)
      age               fl            lake         era
 Min.   : 0.000   Min.   : 20   Harrison:61   1977-80:23
 1st Qu.: 3.000   1st Qu.:221                 1997-01:38
 Median : 6.000   Median :372
 Mean   : 5.754   Mean   :319
 3rd Qu.: 8.000   3rd Qu.:425
 Max.   :14.000   Max.   :480
```

The quartile deviation is described by splitting the data to eliminate the impact of extremely big and extremely tiny outliers, and determining the period where the data is largely concentrated by the density of the data. In other words, it is not influenced by extreme values. The lower the number, the more concentrated the data in the middle, the bigger the value, the more spread the data in the center.

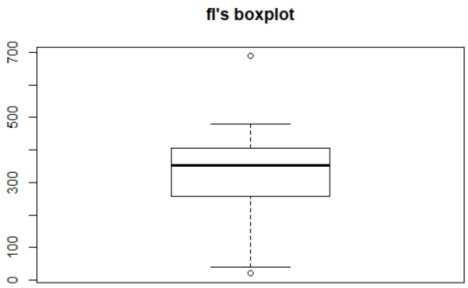
We can figure out that the **Age's** min number is 0, first quartile number is 3, median number is 6, mean is 5.754, third quartile number is 8 and max number is 14. Age's IQR=Q3-Q1= 5, then upper limit= 8+1.5*5= 15.5, lower limit= 3-1.5*5= -4.5. So, the range is from -4.5 to 15.5. According to Lakefilter's summary, Q2 is present nearer to Q3, and mean < median, I can know that is negatively skewed distribution. (Sharma, 2020)

For the part of **fl**, the min number is 20, first quartile number is 221, median number is 372, mean is 319, third quartile number is 425 and max number is 480. Fl's IQR=Q3-Q1= 204, then upper limit = 425+1.5*204=731, lower limit= 221-1.5*204 = -85. So, the range is from -85 to 731. According to Lakefilter's summary, Q2 is present nearer to Q3, and mean < median, I can know that is negatively skewed distribution as well. (Sharma, 2020)

```
> boxplot(BullTroutRML2$age,
+         main="age's boxplot")
>
```

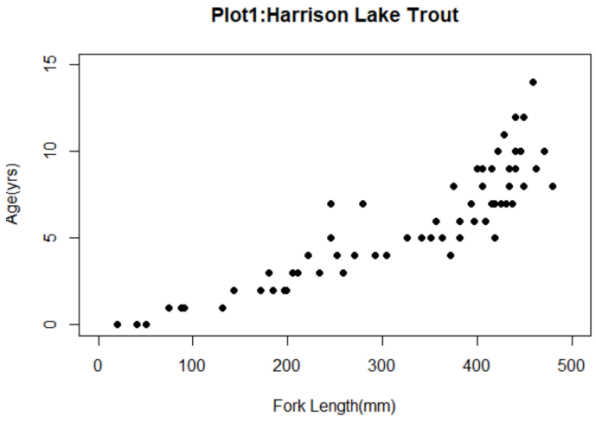


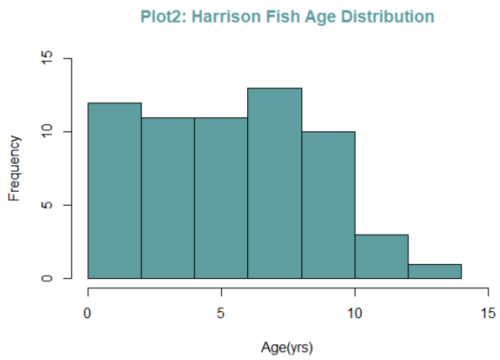

```
> boxplot(BullTroutRML2$fl,
+         main="fl's boxplot")
>
```

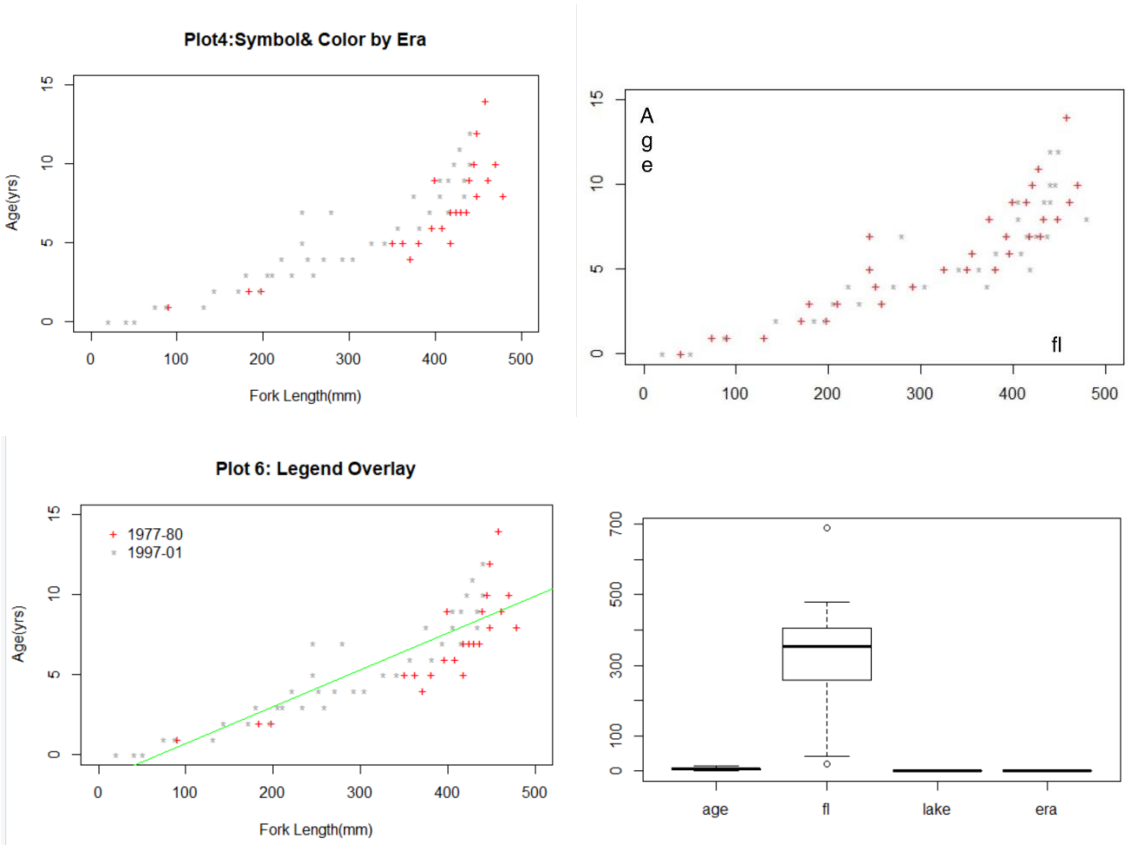


In summation, we can observe that there are no points outside the minimum and maximum on the age's boxplot, implying that there are no points outside the range of the dataset, implying that Age's outlier is zero. And, in terms of fork length's, we can observe that there is one obviously out of range point, indicating that an outlier exists.

B. Below are different forms of the data visualization. They are plot, boxplot, histogram, and regression line.







The first scatter plot depicts the overall pattern of age changing with fork length variable, indicating that as fork length grows, so does age. The histogram focuses on the individual's description and can properly determine the size of each item. As a result, we may conclude that the fork length is greatest when the age of seven. Figures 3 and 4 show a scatter plot comparison. To tell Rstudio which point to use, what color to use, and what icon to use, we must supply a vector in the parameter that includes the icon and color of each point. Figure 3 shows the clean scatter plot after the filter, while Figure 4 shows the scatter plot without transferring the vector entirely to the parameter. In the meantime, Figure 3's data distribution clearly shows that there is a direct relationship between Fork Length and Ages. Data is frequently concentrated at a single value. And, when I use R programming language to perform Linear Regression, the output is exactly proportional, which is the following regression line.

Fork length shows one outlier in the last boxplot when compared to age, lake, and era. In general, depending on the circumstance, data analysis and processing

procedures differ. An outlier might be a data value that was incorrectly recorded. If this is the case, it should be addressed before continuing with the analysis. An outlier might also be a value that was mistakenly included in the data collection. If this is the case, it is possible to delete it. An aberrant value can also be an anomalous data value that was successfully collected and is now part of the data set. It should remain in this state.

C. Plot 1 demonstrates a substantial positive relationship between BullTout's Fork Length and age. The Fork Length expands with age, but it stops between 400 and 500mm, indicating that the mature BullTrout's Fork Length is in this range. The age distribution of BullTrout is clearly shown in Plot 2. The average life span is between 0 and 10 years, while the majority of the population is between the ages of 5 and 10.

Plots 4 and Plot 6 depict the current growth rate and longevity disparity. BullTrout seemed to have developed quicker in 1977 than in 1997 at the same age. As a result, we found more 1977 Era examples between 350mm and 500mm Fork Length ranges for different ages larger than 5 years.

Bibliography

Bibliography

Ogle, D. H. (2020). *DEPRECATED - Subsets/filters a data frame and drops the unused levels.* Retrieved from derekogle: https://derekogle.com/FSA/reference/filterD-deprecated.html

Schork, J. (2019, 04 13). *Convert Data Frame Column to Numeric in R (2 Examples) | Change Factor, Character & Integer*. Retrieved from statisticsglobe: https://statisticsglobe.com/convert-data-frame-column-to-numeric-in-r

Sharma, A. (2020, July 6). *Statistics for Data Science: What is Skewness and Why is it Important?* Retrieved from analyticsvidhya: https://www.analyticsvidhya.com/blog/2020/07/what-is-skewness-statistics/

Appendix

```r
#Print the name at the top of the script
paste("Plotting Basics: Wang")

#Import multiple libraries
install.packages("FSA")
library(FSA)
install.packages("FSAdata")
library(FSAdata)
install.packages("magrittr")
library(magrittr)
install.packages("dplyr")
library(dplyr)
install.packages("plotrix")
library(plotrix)
install.packages("ggplot2")
library(ggplot2)
install.packages("moments")
library(moments)

#load dataset
dataset <- BullTroutRML2
dataset

#Print first records and last 3 records
head(dataset, 1)
tail(dataset,3)

```

```r

#remove all records except Harrison Lake
BullTroutRML2
Lakefilter<- c(filterD(BullTroutRML2,lake == "Harrison"))
Lakefilter<- (data.frame(Lakefilter))
Lakefilter

#the first and last 5 records
head(Lakefilter, 1)
tail(Lakefilter,5)

#structure of the filtered
str(Lakefilter)

#summary of the filtered
summary(Lakefilter)


#Plot 1: scatterplot for "age" and "fl"
attach(Lakefilter)
plot(fl,age,main="Plot1:Harrison Lake Trout",
     xlab="Fork Length(mm)",ylab="Age(yrs)",pch=16,xlim=c(0,500),ylim=c(0,15))

```

(Ogle, 2020)

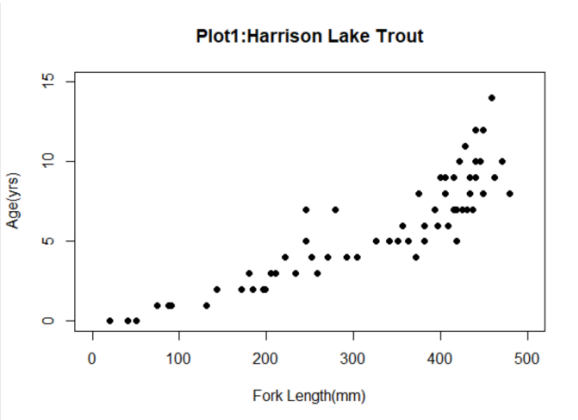
```
49
50 #Plot 2: "Age" historgram
51 hist(age,main="Plot2: Harrison Fish Age Distribution",
52      xlab="Age(yrs)",ylab="Frequency",
53      xlim=c(0,15),ylim=c(0,15),
54      col="cadetblue",col.main="cadetblue")
55
56 #Plot 3: Overdense plot
57 attach(Lakefilter)
58 smoothScatter(fl,age,main="Plot3:Harrison Density Shaded by Era",
59               xlab="Fork Length(mm)",
60               ylab="Age(yrs)",
61               xlim=c(0,500),ylim=c(0,15),pch=16,col="green")
62
63 #create new "tmp" object
64 tmp<- headtail(dataset, n=3)
65 tmp
66
67 #Display the "era"column
68 era<- tmp[,c("era")]
69 era<- (data.frame(era))
70 era
71
72 #create a pchs
73 pchs <- c("+","*")
74
```

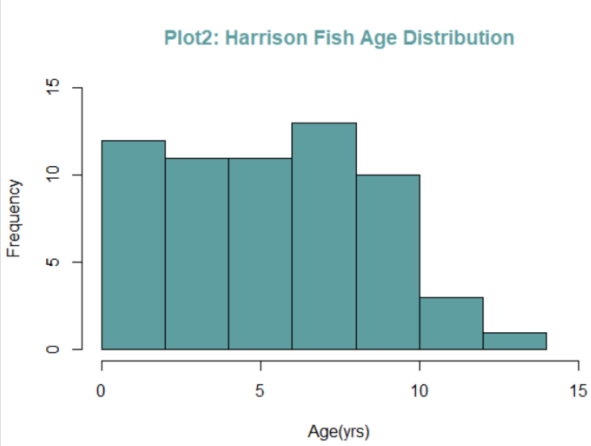
```
74
75 #create a cols
76 cols<- c("red","gray60")
77 cols
78
79 #convert the tmp era values to numberic values
80 num<- as.numeric(era)
81 num
82
83 #initialize the cols vector with tmp era values
84 colors_num <-cols[num]
85 colors_num
86
87 #plot 4
88 all_cols<- cols(Lakefilter$era)
89 all_cols
90 all_pchs<- pchs(Lakefilter$era)
91 all_pchs
92 attach(Lakefilter)
93 plot(fl, age,main="Plot4:Symbol& Color by Era",
94      xlab="Fork Length(mm)",ylab="Age(yrs)",xlim=c(0,500),ylim=c(0,15),
95      col=all_cols,pch=all_pchs)
96
```

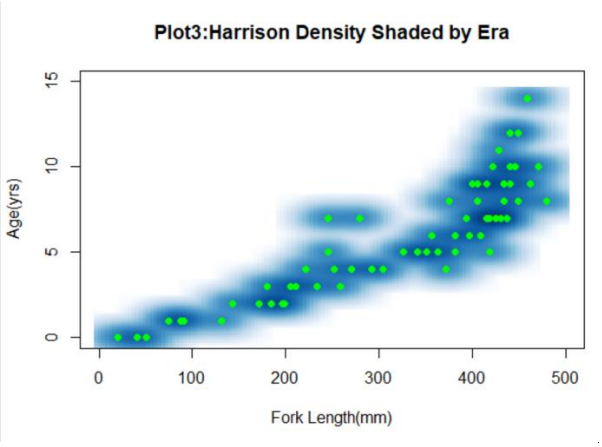
(Schork, 2019)

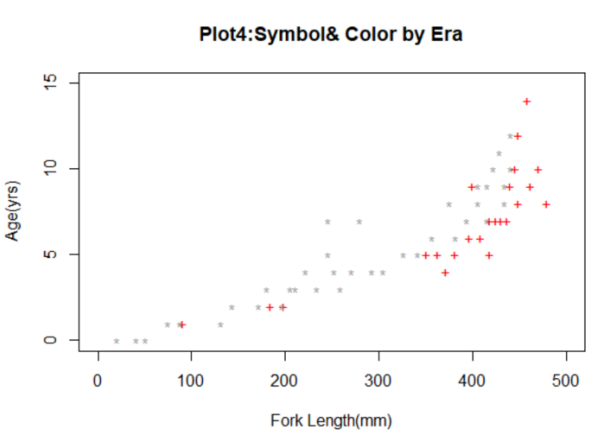
```
96
97  #Plot 5: regression line
98  all_cols<- cols(Lakefilter$era)
99  all_cols
100 all_pchs<- pchs(Lakefilter$era)
101 all_pchs
102 attach(Lakefilter)
103 plot(fl, age,main="Plot 5: Regression Overlay",
104      xlab="Fork Length(mm)",ylab="Age(yrs)",xlim=c(0,500),ylim=c(0,15),
105      col=all_cols,pch=all_pchs)
106 reg_model<- lm(age~fl,data=Lakefilter)
107 abline(reg_model,col = "green")
108
109 #Plot 6: legend of plot 6
110 all_cols<- cols(Lakefilter$era)
111 all_cols
112 all_pchs<- pchs(Lakefilter$era)
113 all_pchs
114 attach(Lakefilter)
115 plot(fl, age,main="Plot 6: Legend Overlay",
116      xlab="Fork Length(mm)",ylab="Age(yrs)",xlim=c(0,500),ylim=c(0,15),
117      col=all_cols,pch=all_pchs)
118 reg_model<- lm(age~fl,data=Lakefilter)
119 abline(reg_model,col = "green")
120 legend(2,15,c("1977-80","1997-01"),pch=pchs,col=cols,box.lty=0)
121
```

Output:

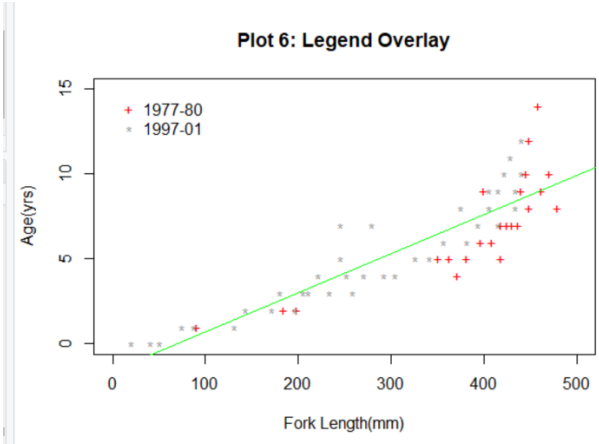





Plot3:Harrison Density Shaded by Era
Age(yrs)
Fork Length(mm)
0
5
10
15
0
100
200
300
400
500


Plot4:Symbol& Color by Era
Age(yrs)
Fork Length(mm)
0
5
10
15
0
100
200
300
400
500


Plot 5: Regression Overlay
Age(yrs)
Fork Length(mm)
0
5
10
15
0
100
200
300
400
500


Plot 6: Legend Overlay
1977-80
1997-01
Age(yrs)
Fork Length(mm)
0
5
10
15
0
100
200
300
400
500