Name: Yijun Wang

Class: Aly 6000

Date: 10/16/2021

Title: Module 3 Project — Executive Summary Report 3

Key finding

A. descriptive characteristics of the data set

This dataset contains 7 variables and 676 observations on eight different fish species: Bluegill, Largemouth Bass, Yellow Perch, Pumpkinseed, Tadpole Madtom, Iowa Darter, Bluntnose Minnow, and Black. The goal of this study is to examine the dataset statistically and provide relevant, in-depth findings about the dataset, including data levels, descriptive statistics, and data distribution.

```
> #18.Add new variables to <d> and call them cumfreq, counts, and cumcounts
> d<- mutate(d, cumfreq= cumsum(RelFreq),
+            counts= RelFreq*676,
+            cumcounts=cumsum(counts))
> d
           Species    RelFreq   cumfreq counts cumcounts
1  Largemouth Bass 0.33727811 0.3372781    228       228
2         Bluegill 0.32544379 0.6627219    220       448
3 Bluntnose Minnow 0.15236686 0.8150888    103       551
4     Yellow Perch 0.05621302 0.8713018     38       589
5    Black Crappie 0.05325444 0.9245562     36       625
6      Iowa Darter 0.04733728 0.9718935     32       657
7      Pumpkinseed 0.01923077 0.9911243     13       670
8   Tadpole Madtom 0.00887574 1.0000000      6       676
```

The overall kind and number of fish may be determined from the image above. We may deduce from the RelFreq column that Bluegill and Largemouth Bass account for the majority of the data. 0.33727+0.32544=0.66271. In other words, these two species accounts for around 67 percent of the research. (ZACH, 2021)

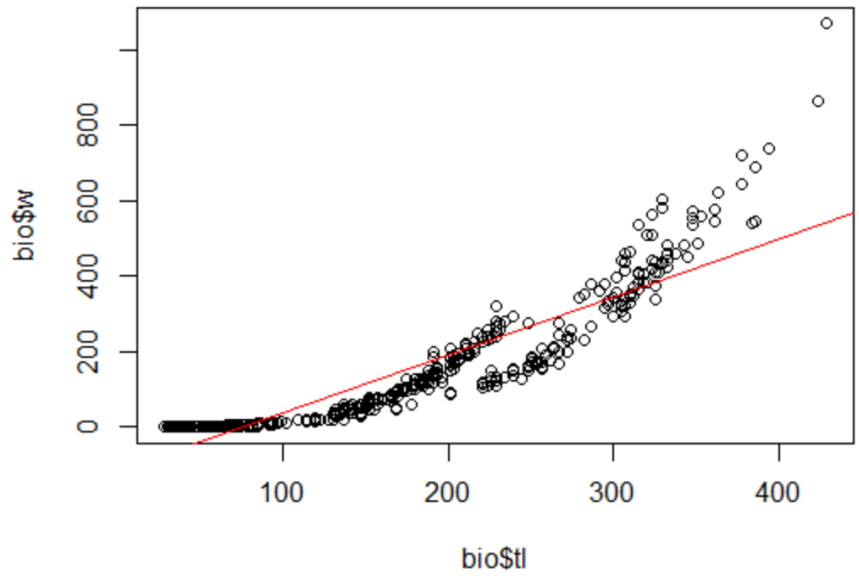
```
> summary(bio)
     netID            fishID                  species          tl       
 Min.   :  1.00   Min.   :  7.0   Largemouth Bass :228   Min.   : 27.0  
 1st Qu.: 13.00   1st Qu.:175.8   Bluegill        :220   1st Qu.: 66.0  
 Median : 37.00   Median :345.5   Bluntnose Minnow:103   Median :189.5  
 Mean   : 67.65   Mean   :434.2   Yellow Perch    : 38   Mean   :186.5  
 3rd Qu.:109.00   3rd Qu.:695.5   Black Crappie   : 36   3rd Qu.:295.0  
 Max.   :206.00   Max.   :915.0   Iowa Darter     : 32   Max.   :429.0  
                                  (Other)         : 19                  
       w                tag         scale       
 Min.   :   0.2          :477   Mode :logical  
 1st Qu.:   2.0   1019   :  2   FALSE:213      
 Median :  54.5   1785   :  2   TRUE :463      
 Mean   : 126.8   o0507  :  2                  
 3rd Qu.: 190.5   o0526  :  2                  
 Max.   :1070.0   o0529  :  2                  
 NA's   :165      (Other):189                  
>
```

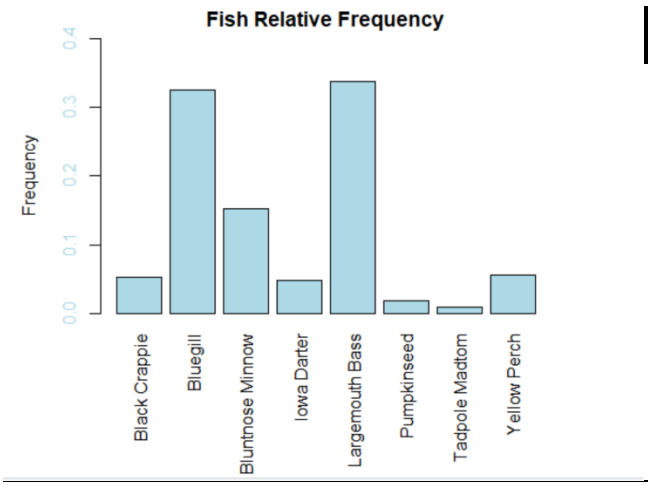
We can see from the second image above that there are 165 NAs in the weight data. This demonstrates that 165 of the weights are missing from a total of 676 pieces. When different sets of data are compared, the data clearly shows the difference distribution of each group, which is highly useful for generating conclusions. As a result, the weight data is insufficiently compelling and not accurate, so I will not consider this part. And, using species data, we can demonstrate that the number of Largemouth and Bluegill is, in fact, the bulk of the dataset.

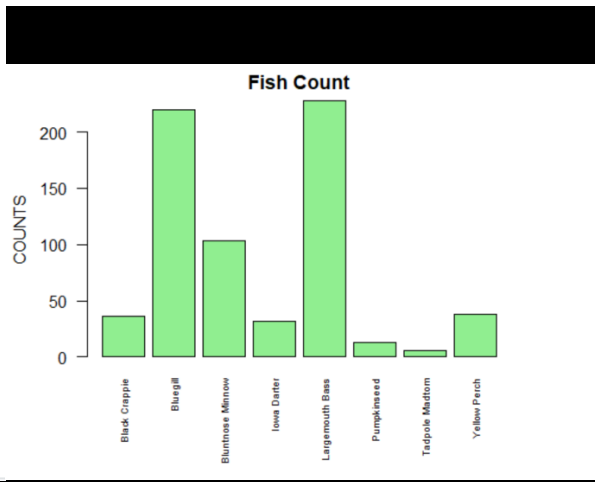
For the part of **tl**, the min number is 27, first quartile number is 66, median number is 189.5, mean is 186.5, third quartile number is 295 and max number is 429. Fl's IQR=Q3-Q1= 229, then upper limit = 295+1.5*229=638.5, lower limit= 66-1.5*229 = -277.5. So, the range is from -277.5 to 638.5. According to Lakefilter's summary, Q2 is present nearer to Q3, and mean < median, I can know that is negatively skewed distribution.

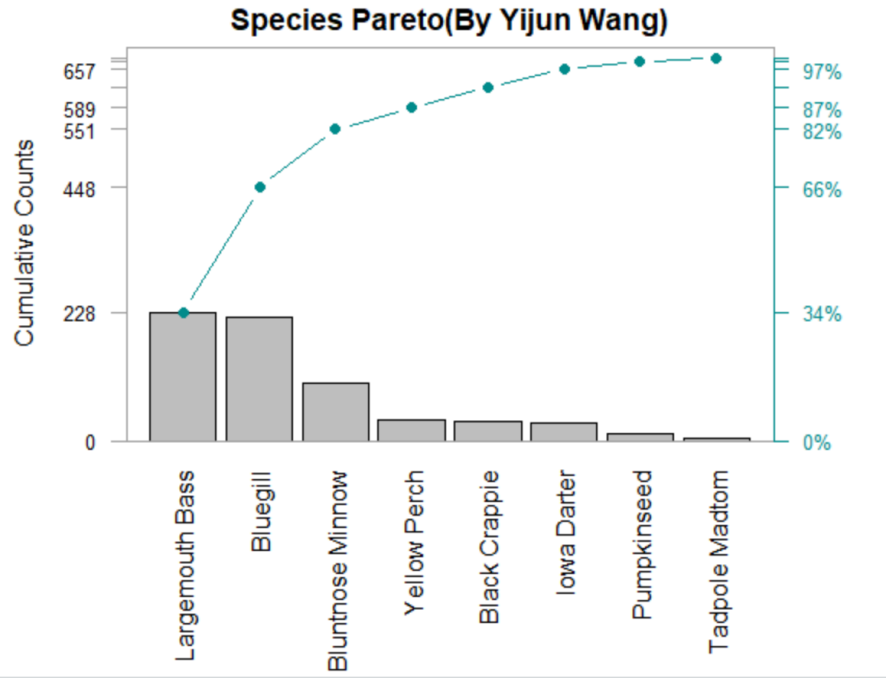
Largemouth Bass's average tl is 298.5657895, Bluegill's average tl is 140.8318182. Largemouth Bass's average is more than other species. In summary, I think Lagremouth Bass is the biggest species and has the most counts in this data set.

B. Provide 3 visualization









(Schork, 2020)

The data distribution in the regression line demonstrates that when there is a connection between tail length and weight, data is frequently concentrated towards a central value. And, when I use the R programming language to perform Linear Regression, the output is exactly proportional. As a result of the finding, as tail length increases, so does weight.

I used two historgram to show fish relative frequency and fish count, both historgram looks similar, specially for the fish of Bluegill and Largemouth Bass. It demonstrates how these two fish species have a huge population and a higher relative frequency.

The last one displays the cumulative frequencies of the dataset's eight species. Largemouth Bass and Bluegill's counts are more than other species, cumulative counts line was increasing, but not that high as Largemouth Bass and Bluegill.

C. summary

While bar charts and boxplots are good for comparing variables in a dataset, pareto charts are the most effective for determining which variables are dominant in the data. By creating this sort of graphic, statisticians will be able to quickly see which aspects should be investigated in order to make less time-consuming but still effective progress. As a result, rather than tracking all variables, researchers will save time and effort by concentrating on select ones to investigate. Largemouth Bass, Bluegill, and Bluntnose Minnow appear 551 times out of 676 times, indicating that these three variables should be closely monitored in this dataset. (am, 2016)

Furthermore, the average tail length of a Largemouth Bass is longer than that of other species, implying that this type of fish is heavier and larger than others. In other words, because they have a greater capacity to adapt to survive than other species, they can live for a long period and are enormous in size.

Bibliography

am, E. (2016, 01 24). *Data visualization in R: Axis Labels (color, size and rotation)*. Retrieved from coders-corner: https://coders-corner.net/2016/01/24/data-visualization-in-r-axis-labels-color-size-and-rotation/

Schork, J. (2020, 11 20). *Change Colors of Axis Labels & Values of Base R Plot (2 Examples)*. Retrieved from statisticsglobe: https://statisticsglobe.com/r-change-colors-axis-labels-values-of-plot

ZACH. (2021, 02 04). *Format Numbers as Percentages in R (With Examples)*. Retrieved from statology: https://www.statology.org/percentage-in-r/

Appendix

```r
#1. Print the name at the tip of the script
paste("Yijun Wang")

#Import the libraries
install.packages("FSA")
library(FSA)
install.packages("FSAdata")
library(FSAdata)
install.packages("magrittr")
library(magrittr)
install.packages("dplyr")
library(dplyr)
install.packages("tidyr")
library(tidyr)
install.packages("plyr")
library(plyr)
install.packages("tidyverse")
library(tidyverse)

#2. Import the inchBio.csv and name the table<bio>
bio <- read.csv("C:\\Users\\junni\\Downloads\\inchBio.csv",header=TRUE,sep=",")
bio

#3. display the head,tail and structure
head(bio,n=3)
tail(bio,n=3)
str(bio)

#4. create an object<counts>
count(bio[1:677],vars="species")
counts <- count(bio$species)
counts

#5. display 8 level names
counts$freq<- NULL
counts
```

```r
#6. create<tmp> display the different species and the number of record of each
tmp <- table(bio$species)
tmp <- (data.frame(tmp))
tmp

#7. Create a subset<tmp2>, species variable and display the first five records
tmp2 <- subset(bio,select = species)
head(tmp2,5)

#8. Create a table <w>, species variable.class w
w <- table(bio$species)
w
class(w)

#9. Convert <w> to a data frame named <t> and display the results
t <- (data.frame(w))
t

#10. Extract and display the frequency values from the <t> data frame
print(t$Freq)

#11. create a table <cSpec>
cSpec <- table(bio$species)
cSpec

#12. create a table <cSpecPct> and class
#&13. <cSpecPct> to data frame named <u>.
install.packages(scales)
library(scales)
cSpecPct <- table(bio$species)/676
u <- (data.frame(cSpecPct))
u
class(cSpecPct)
#13. confirm <u> is data frame
class(u)
```

```r

#14.create barplot <cSpec>
barplot(cSpec, ylab="COUNTS", main = "Fish Count",
        col = "Light Green", las = 2, cex.names = 0.6)


#15. create barplot <cSpecPct>
barplot(cSpecPct, ylim=c(0,.4),col="light blue",las=2, yaxt="none",
        main= "Fish Relative Frequency",ylab="Frequency")
axis(2, col.axis="Light Blue")

#16.Rearrange the <u> cSpecPct data frame in descending order of relative frequer
#Save the rearranged data frame as the object <d>
d = arrange(u,desc(Freq))
d

#17. Rename the <d> columns Var 1 to Species, and Freq to RelFreq
names(d) <- c("Species", "RelFreq")
d

```

```r
#18.Add new variables to <d> and call them cumfreq, counts, and cumcounts
d<- mutate(d, cumfreq= cumsum(RelFreq),
           counts= RelFreq*676,
           cumcounts=cumsum(counts))
d

#19. Create a parameter variable <def_par> to store parameter variables
def_par=par()
par(mar=c(8,5,2,5))

#20.Create a barplot <pc>
pc <- barplot(d$counts, width=1, space=0.15, boarder=NA, axes=F,
              ylim = c(0, 3.05 * max(d$counts, na.rm = TRUE)),
              ylab = "Cumulative Counts",cex.axis = 1.7,
              names.arg = d$Species,
              main= "Species Pareto(By Yijun Wang)", las=2)
pc

```

```r
#21. Add a cumulative counts line to the <pc> plot
lines(pc, d$cumcounts, type= 'b', pch= 19, col= 'cyan4')

#22. Place a grey box around the pareto plot
box(col = 'grey62')

#23.Add a left side axis
axis(side= 2, at= c(0, d$cumcounts),
     tick = TRUE ,line = NA,
     col.ticks = "grey62",
     col= "grey62", cex.axis=0.8, las=2)

#24. Add axis details on right side of box
axis(side= 4, at = c(0, d$cumcounts),
     col= "cyan4", cex.axis= 0.8, las= 2, tick= TRUE,
     line= NA, col.axis= "cyan4",
     labels = paste0(round( c(0,d$cumfreq) * 100,digits = 0),'%'))

#25.display the name
```