College of Professional Studies
Northeastern University San Jose

MPS Analytics

Course: ALY6000 - Introduction to Data Analytics

Assignment:

MODULE PROJECT - 3

EXECUTIVE SUMMARY REPORT – 3

Submitted on:

October 9, 2022

Submitted to:	**Submitted by:**
Professor: BEHZAD AHMADI	NIKSHITA RANGANATHAN

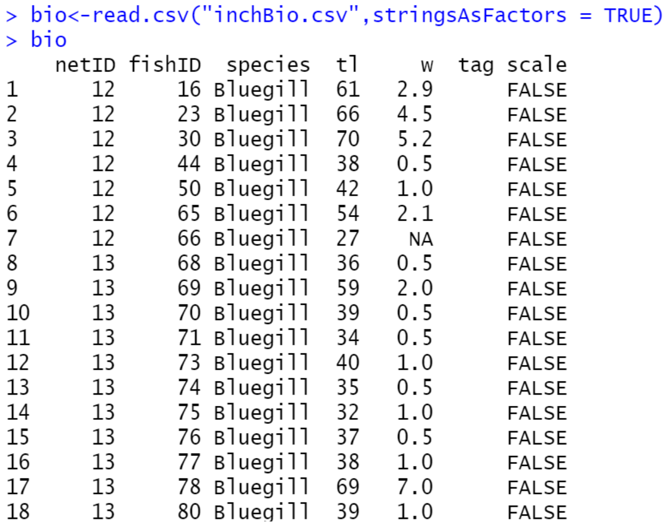
Introduction

This assignment aims to provide an explanatory analysis of descriptive characteristics of the data set that includes statistics including counts, cumulative counts, and frequency, percentages, boxplots, histograms, frequency and probability distributions, or bar plots (bar charts) & pareto plot. The goal of this is to provide not only the visual illustrations, but also to explain the significance of them.
The assignment focuses on the data analysis of dataset called InchBio that has data related to variety of fish species.

Key Findings

1. Importing dataset inchBio

- The dataset "inchbio" contains information of different fish species. I have imported the dataset as "bio" and stringsAsFactors () converts the data as factors.

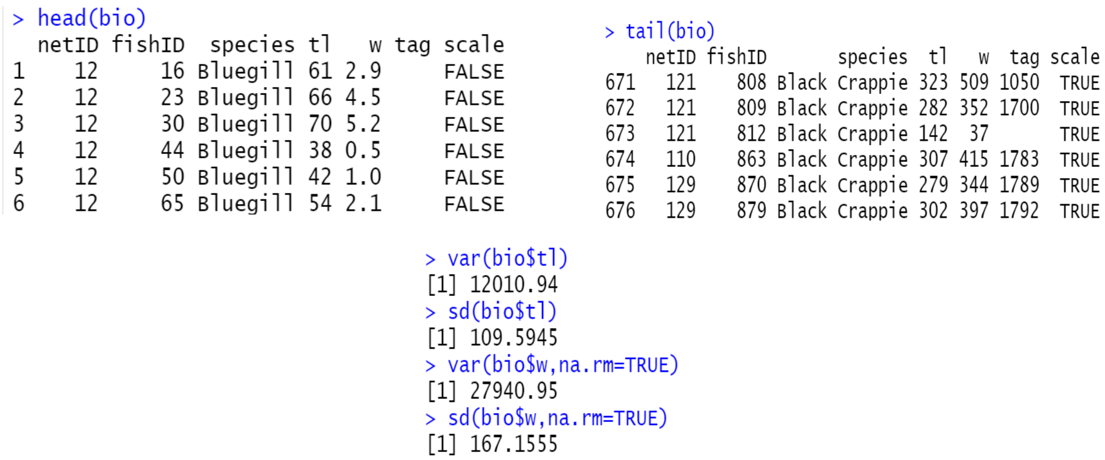


```
> bio<-read.csv("inchBio.csv",stringsAsFactors = TRUE)
> bio
    netID fishID  species  tl     w  tag scale
1      12     16 Bluegill  61   2.9      FALSE
2      12     23 Bluegill  66   4.5      FALSE
3      12     30 Bluegill  70   5.2      FALSE
4      12     44 Bluegill  38   0.5      FALSE
5      12     50 Bluegill  42   1.0      FALSE
6      12     65 Bluegill  54   2.1      FALSE
7      12     66 Bluegill  27    NA      FALSE
8      13     68 Bluegill  36   0.5      FALSE
9      13     69 Bluegill  59   2.0      FALSE
10     13     70 Bluegill  39   0.5      FALSE
11     13     71 Bluegill  34   0.5      FALSE
12     13     73 Bluegill  40   1.0      FALSE
13     13     74 Bluegill  35   0.5      FALSE
14     13     75 Bluegill  32   1.0      FALSE
15     13     76 Bluegill  37   0.5      FALSE
16     13     77 Bluegill  38   1.0      FALSE
17     13     78 Bluegill  69   7.0      FALSE
18     13     80 Bluegill  39   1.0      FALSE
```

Figure 1 – inchBio

2. Analysing bio dataset

- Head and tail functions returns the first and last records of "bio". In this case, we have the first 6 and last 6 records. I have calculated the variance and standard deviation of weight and total length columns of given dataset. na.rm=TRUE removes NA values in weight column.



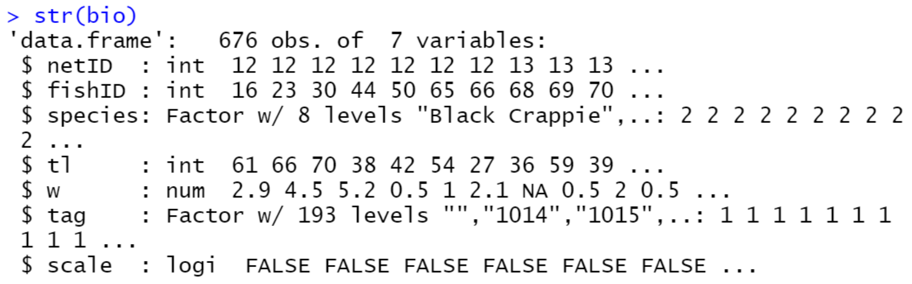
```
> head(bio)
  netID fishID  species tl   w tag scale
1    12     16 Bluegill 61 2.9     FALSE
2    12     23 Bluegill 66 4.5     FALSE
3    12     30 Bluegill 70 5.2     FALSE
4    12     44 Bluegill 38 0.5     FALSE
5    12     50 Bluegill 42 1.0     FALSE
6    12     65 Bluegill 54 2.1     FALSE
```

```
> tail(bio)
    netID fishID       species  tl   w  tag scale
671   121    808 Black Crappie 323 509 1050  TRUE
672   121    809 Black Crappie 282 352 1700  TRUE
673   121    812 Black Crappie 142  37        TRUE
674   110    863 Black Crappie 307 415 1783  TRUE
675   129    870 Black Crappie 279 344 1789  TRUE
676   129    879 Black Crappie 302 397 1792  TRUE
```

```
> var(bio$tl)
[1] 12010.94
> sd(bio$tl)
[1] 109.5945
> var(bio$w,na.rm=TRUE)
[1] 27940.95
> sd(bio$w,na.rm=TRUE)
[1] 167.1555
```

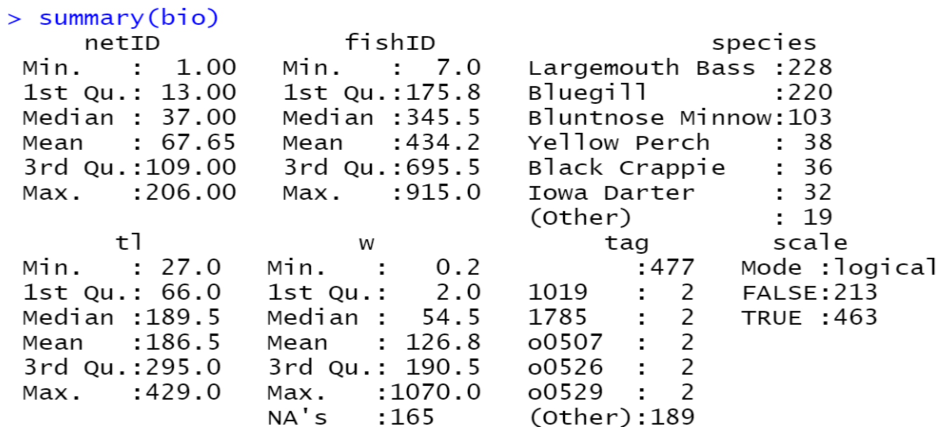
Figure 2 – head(),tail(),var() and sd()

- With the help of str(), different datatypes can be noticed. The dataset consists of 676 observations and 7 variables and the variables are netID, fishID, species, total length, weight, tag and scale.
- It can be observed in the output of summary(bio) that scale is logical and has 213 observations under false category and 463 observations under true category. Fish and net ID is for the identification of the fishes. There are some NA values in weight data.



```
> str(bio)
'data.frame':	676 obs. of  7 variables:
 $ netID  : int  12 12 12 12 12 12 12 13 13 13 ...
 $ fishID : int  16 23 30 44 50 65 66 68 69 70 ...
 $ species: Factor w/ 8 levels "Black Crappie",..: 2 2 2 2 2 2 2 2 2
 2 ...
 $ tl     : int  61 66 70 38 42 54 27 36 59 39 ...
 $ w      : num  2.9 4.5 5.2 0.5 1 2.1 NA 0.5 2 0.5 ...
 $ tag    : Factor w/ 193 levels "","1014","1015",..: 1 1 1 1 1 1 1
 1 1 1 ...
 $ scale  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

Figure 3 - str()

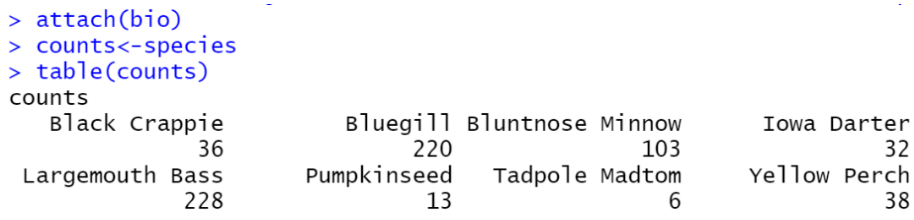


```
> summary(bio)
     netID            fishID                   species   
 Min.   :  1.00   Min.   :  7.0   Largemouth Bass :228  
 1st Qu.: 13.00   1st Qu.:175.8   Bluegill        :220  
 Median : 37.00   Median :345.5   Bluntnose Minnow:103  
 Mean   : 67.65   Mean   :434.2   Yellow Perch    : 38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

Figure 4 - summary()

3. Creating counts, tmp, tmp2 and t

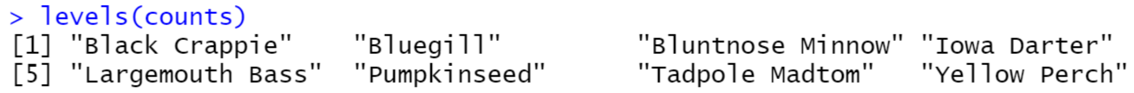
- <counts> comprises of all the species recorded. Fish species include Black Crappie, Bluegill, Largemouth Bass, Bluntnose Minnow, Tadpole Madtom, Iowa Darter, Pumpkinseed and Yellow Perch. Bio dataset gets stored in R's search directory with the help of attach().



```
> attach(bio)
> counts<-species
> table(counts)
counts
   Black Crappie         Bluegill Bluntnose Minnow      Iowa Darter 
              36              220              103               32 
 Largemouth Bass      Pumpkinseed   Tadpole Madtom     Yellow Perch 
             228               13                6               38 
```

Figure 5 - <counts>

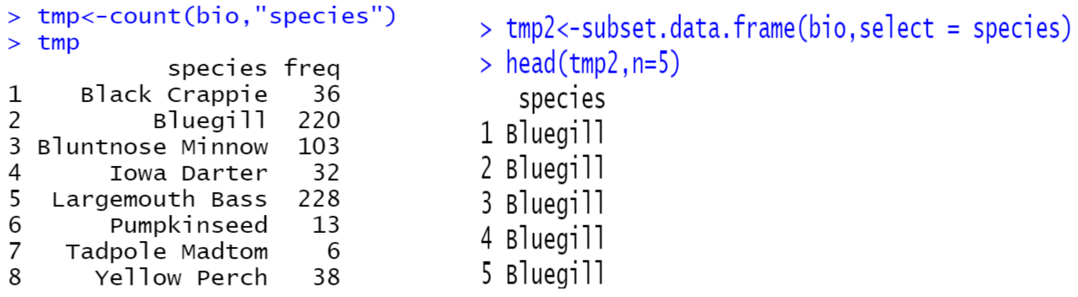
- levels() provides the details of the factor levels. 8 levels mean the 8 fish species of the dataset.



```
> levels(counts)
[1] "Black Crappie"    "Bluegill"         "Bluntnose Minnow" "Iowa Darter"     
[5] "Largemouth Bass"  "Pumpkinseed"      "Tadpole Madtom"   "Yellow Perch"    
```

Figure 6 - levels()

- count() provides the number of occurrences of the species. The highest frequency of records are for Largemouth Bass (228) and Bluegill(220). The fish species Tadpole Madtom has only 6 observations in “bio”. tmp2 is a subset of bio with species column. By using head(), first 5 rows of tmp2 are filtered.

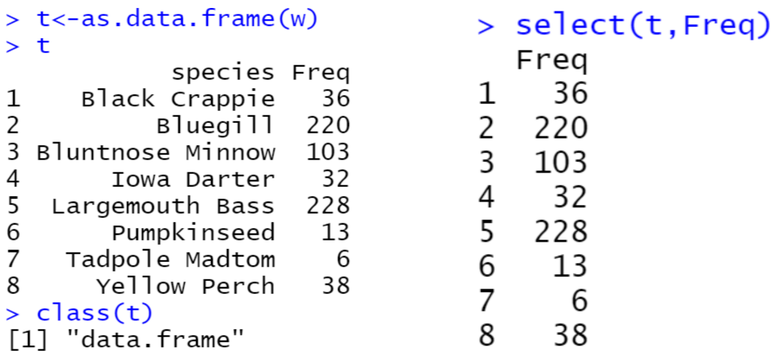


```
> tmp<-count(bio,"species")
> tmp
          species freq
1    Black Crappie   36
2         Bluegill  220
3 Bluntnose Minnow  103
4      Iowa Darter   32
5  Largemouth Bass  228
6      Pumpkinseed   13
7    Tadpole Madtom    6
8     Yellow Perch   38
```

```
> tmp2<-subset.data.frame(bio,select = species)
> head(tmp2,n=5)
   species
1 Bluegill
2 Bluegill
3 Bluegill
4 Bluegill
5 Bluegill
```

Figure 7 - tmp and tmp2

- In order to change w to a dataframe, I have used as.dataframe(w) and saved it as <t> object. class() checks the datatype of the data. select() helps in selecting only the required column Freq from dataframe t.



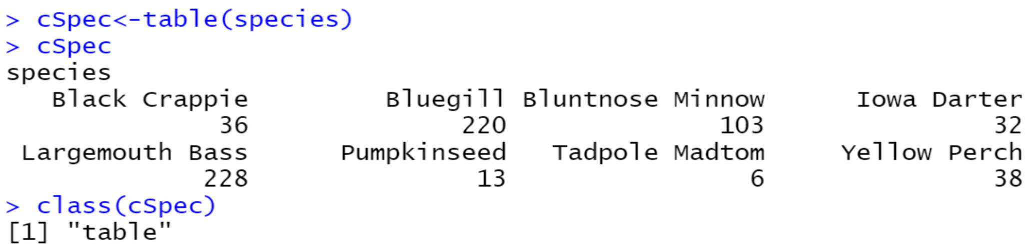
```
> t<-as.data.frame(w)
> t
           species Freq
1    Black Crappie   36
2         Bluegill  220
3 Bluntnose Minnow  103
4      Iowa Darter   32
5  Largemouth Bass  228
6      Pumpkinseed   13
7    Tadpole Madtom    6
8     Yellow Perch   38
> class(t)
[1] "data.frame"
```

```
> select(t,Freq)
  Freq
1   36
2  220
3  103
4   32
5  228
6   13
7    6
8   38
```

Figure 8 – t dataframe

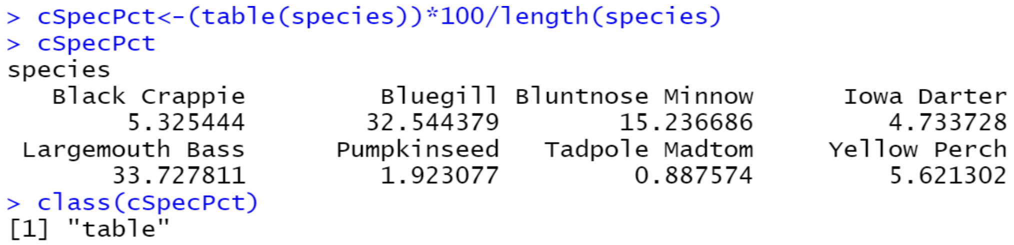
4. Working with cSpec , cSpecPct

- In Figure 9, cSpec is a table with frequencies of observations in each fish species. Whereas Figure 10 represents a table cSpecPct calculating the percentage of species as per the observation count.



```
> cSpec<-table(species)
> cSpec
species
   Black Crappie         Bluegill Bluntnose Minnow      Iowa Darter 
              36              220              103               32 
 Largemouth Bass      Pumpkinseed   Tadpole Madtom     Yellow Perch 
             228               13                6               38 
> class(cSpec)
[1] "table"
```

Figure 9 - cSpec Table

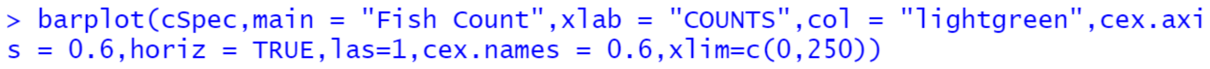


```
> cSpecPct<-(table(species))*100/length(species)
> cSpecPct
species
   Black Crappie        Bluegill Bluntnose Minnow      Iowa Darter
        5.325444       32.544379        15.236686         4.733728
 Largemouth Bass     Pumpkinseed   Tadpole Madtom     Yellow Perch
       33.727811        1.923077         0.887574         5.621302
> class(cSpecPct)
[1] "table"
```

Figure 10 – cSpecPct Table

5. Bar plot of Fish Count

- The bar plot shows the number of fishes in each species. The main data for this graph is cSpec. Largemouth bass and Bluegill have count values more than 200.On the other hand, Pumpkinseed and Tadpole Madtom have the least values.



```
> barplot(cSpec,main = "Fish Count",xlab = "COUNTS",col = "lightgreen",cex.axi
s = 0.6,horiz = TRUE,las=1,cex.names = 0.6,xlim=c(0,250))
```

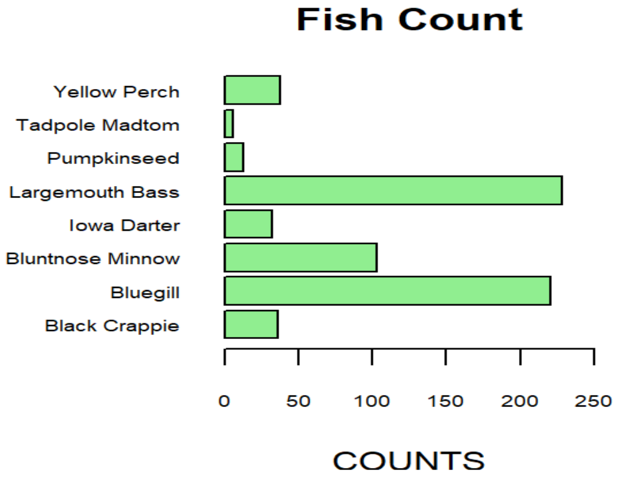
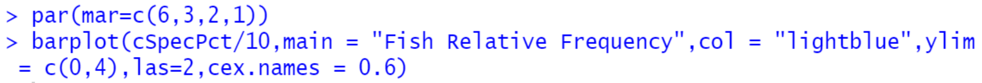


Figure 11 – Bar plot A

6. Bar plot of Fish Relative Frequency

- The graph in Figure 12 is a graph displaying the relative frequency of the fishes for each species. Similar trends as the previous graph can be seen here as well.
- mar() sets margins of the plot in R.



```
> par(mar=c(6,3,2,1))
> barplot(cSpecPct/10,main = "Fish Relative Frequency",col = "lightblue",ylim
 = c(0,4),las=2,cex.names = 0.6)
```

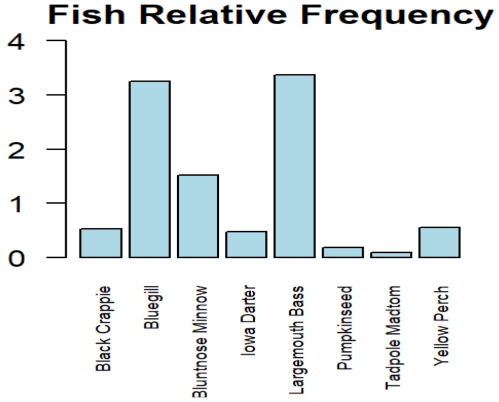
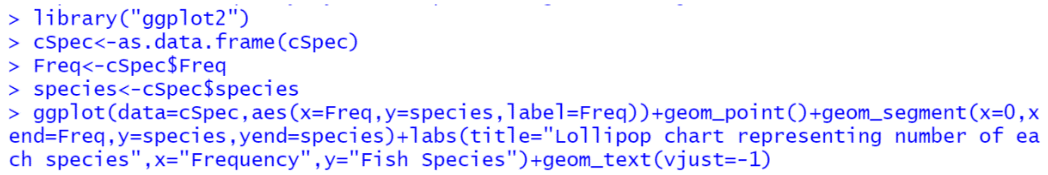


Figure 12 – Bar plot B

7. Other data visualizations

- Lollipop charts are like bar plots. This type of chart is easy to understand and is useful in situations where there are many categories. Frequency of fishes are shown above the lines.
- X axis depicts the frequency and y axis depicts the species of fish. vjust is to adjust the heading of the plot vertically.



```
> library("ggplot2")
> cSpec<-as.data.frame(cSpec)
> Freq<-cSpec$Freq
> species<-cSpec$species
> ggplot(data=cSpec,aes(x=Freq,y=species,label=Freq))+geom_point()+geom_segment(x=0,x
end=Freq,y=species,yend=species)+labs(title="Lollipop chart representing number of ea
ch species",x="Frequency",y="Fish Species")+geom_text(vjust=-1)
```

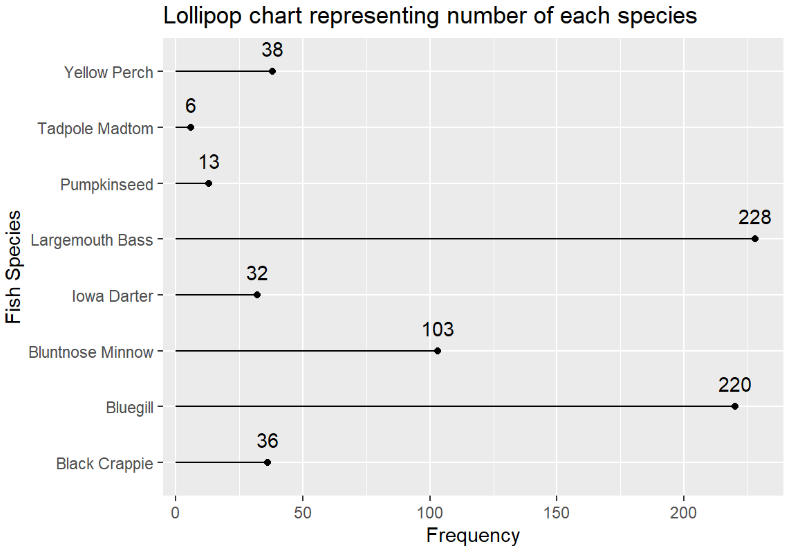
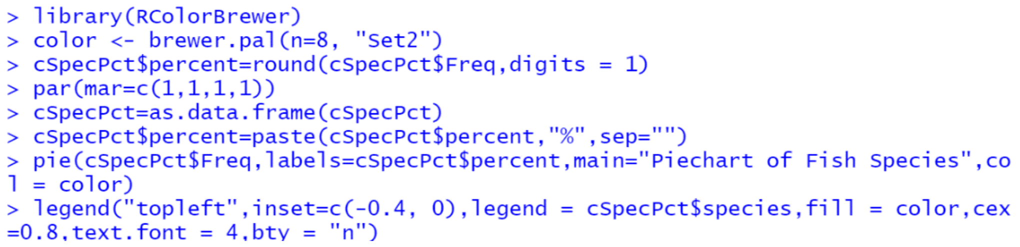


Figure 13 – Lollipop Chart

- I created a piechart with cSpecPct data and could get a clear data visualization of percentages under each category.
- Bluegill is made up of 32.5% and the Largemouth Bass has 32.5% of the total count. Tadpole Madtom, Pumpkinseed, Iowa Darter have the percentages of 0.9% ,1.9% and 4.7% respectively.
- RcolorBrewer is a package, and it has many color palettes stored in it. bty fixes the type of box around the legend of the plot. bty=”n” means no box surrounding the legend.



```
> library(RColorBrewer)
> color <- brewer.pal(n=8, "Set2")
> cSpecPct$percent=round(cSpecPct$Freq,digits = 1)
> par(mar=c(1,1,1,1))
> cSpecPct=as.data.frame(cSpecPct)
> cSpecPct$percent=paste(cSpecPct$percent,"%",sep="")
> pie(cSpecPct$Freq,labels=cSpecPct$percent,main="Piechart of Fish Species",co
l = color)
> legend("topleft",inset=c(-0.4, 0),legend = cSpecPct$species,fill = color,cex
=0.8,text.font = 4,bty = "n")
```

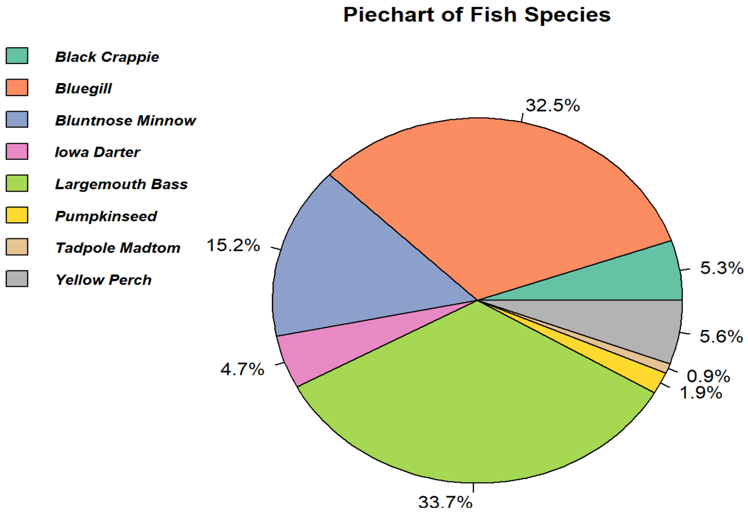
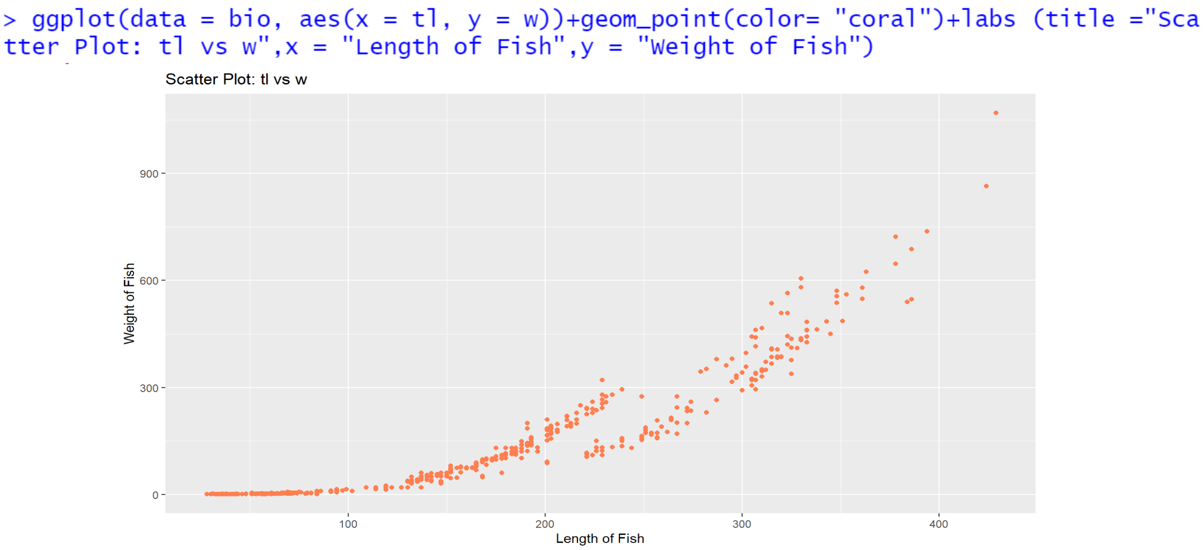


Figure 14 - Pie Chart

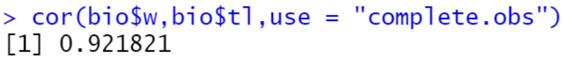
- Scatterplot signifies relationship between two variables and here we can see both variables (length and weight) are proportional to each other. There is an increase in length of the fish with the increase in weight.
- ggplot() is a part of ggplot2 package and is a tool for creating charts and graphs in R.



```
> ggplot(data = bio, aes(x = tl, y = w))+geom_point(color= "coral")+labs (title ="Sca
tter Plot: tl vs w",x = "Length of Fish",y = "Weight of Fish")
```

Figure 15 - Scatterplot between length and weight of the fishes

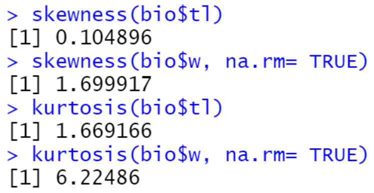
- I verified the correlation by using cor() command. This tells us about the strength of the relationship among the variables. use=”complete.obs” is for removing the NA values from weight column. Correlation coefficient is 0.92 which is very close to 1, this means very strong and positive type correlation.



```
> cor(bio$w,bio$tl,use = "complete.obs")
[1] 0.921821
```

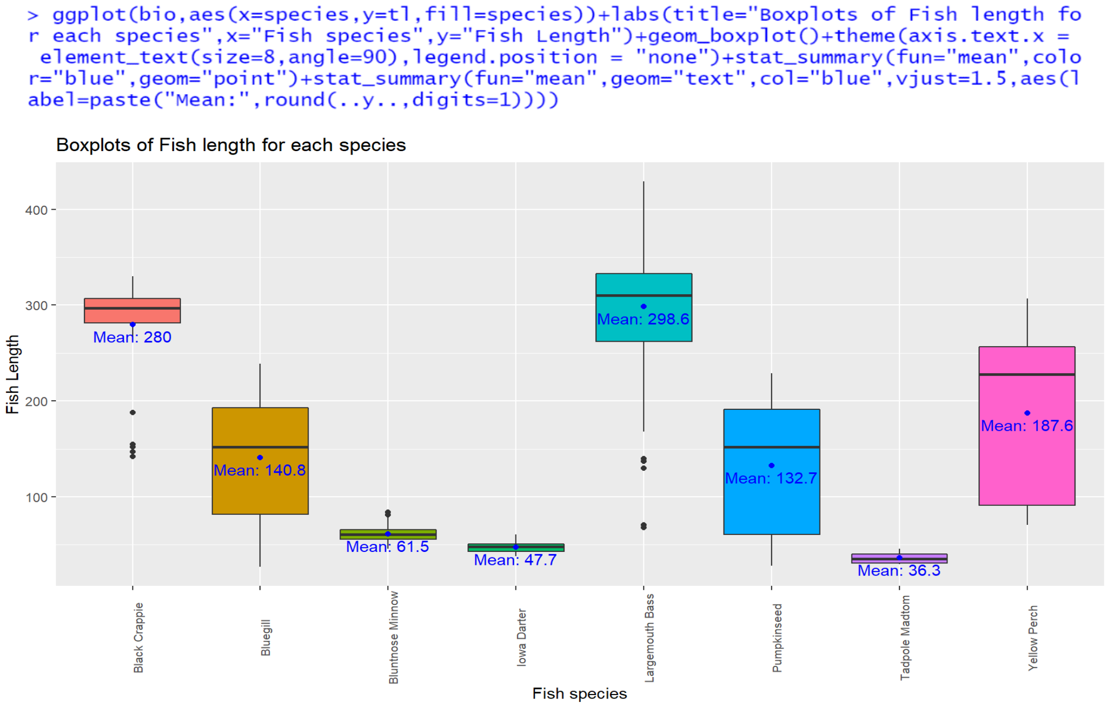
Figure 16 – Correlation Coefficient

- Both length and weight have positive skew. This is also known as right skewed.
- Kurtosis of length is approximately 1.7 which explains it will have fewer outliers. However, weight data has kurtosis of 6.22 denoting more outliers.



```
> skewness(bio$tl)
[1] 0.104896
> skewness(bio$w, na.rm= TRUE)
[1] 1.699917
> kurtosis(bio$tl)
[1] 1.669166
> kurtosis(bio$w, na.rm= TRUE)
[1] 6.22486
```

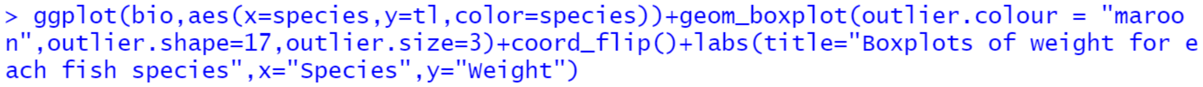
Figure 17 – Kurtosis and Skewness



```
> ggplot(bio,aes(x=species,y=tl,fill=species))+labs(title="Boxplots of Fish length fo
r each species",x="Fish species",y="Fish Length")+geom_boxplot()+theme(axis.text.x =
 element_text(size=8,angle=90),legend.position = "none")+stat_summary(fun="mean",colo
r="blue",geom="point")+stat_summary(fun="mean",geom="text",col="blue",vjust=1.5,aes(l
abel=paste("Mean:",round(..y..,digits=1))))
```

Figure 18 – Boxplot A

- It can be interpreted from the boxplot between species and fish length that Largemouth Bass are the longest of all species. It is followed by Black Crappie and Yellow Parch. Bluegill and Pumpkinseed fall in the midrange. Bluntnose Minnow, Iowa Darter and Tadpole Madtom are shorter with respect to others in the group.
- I have inserted mean value of lengths in the graph. This is an appropriate measure for comparison between different varieties. Largemouth Bass species have a mean length of 298.6. Black dots are the outliers.



```
> ggplot(bio,aes(x=species,y=tl,color=species))+geom_boxplot(outlier.colour = "maroo
n",outlier.shape=17,outlier.size=3)+coord_flip()+labs(title="Boxplots of weight for e
ach fish species",x="Species",y="Weight")
```

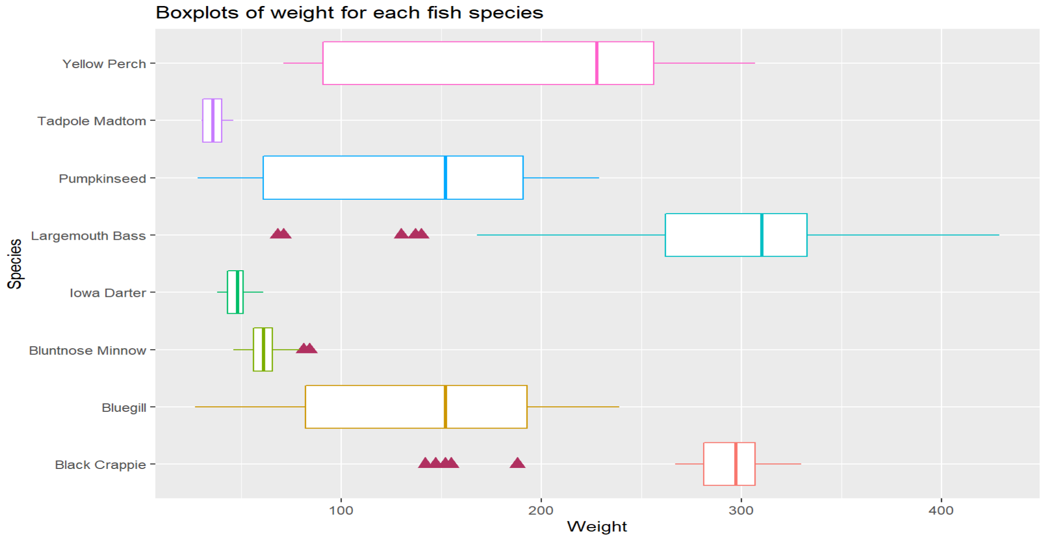
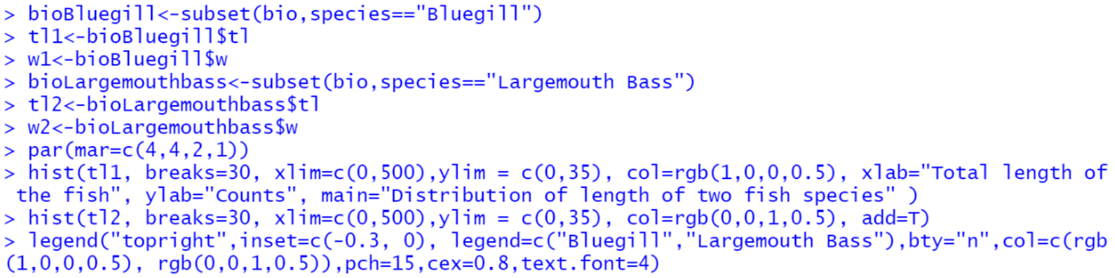


Figure 19 - Boxplot B

- Kurtosis of weight (6.22) can be verified using Boxplot B. The maroon triangles portray the outliers in weight column.
- Both boxplots A and B have similar sequence of values. This proves the linear relationship between weight and length of the fishes which corresponds to the scatterplot (Figure 15).



```
> bioBluegill<-subset(bio,species=="Bluegill")
> tl1<-bioBluegill$tl
> w1<-bioBluegill$w
> bioLargemouthbass<-subset(bio,species=="Largemouth Bass")
> tl2<-bioLargemouthbass$tl
> w2<-bioLargemouthbass$w
> par(mar=c(4,4,2,1))
> hist(tl1, breaks=30, xlim=c(0,500),ylim = c(0,35), col=rgb(1,0,0,0.5), xlab="Total length of
 the fish", ylab="Counts", main="Distribution of length of two fish species" )
> hist(tl2, breaks=30, xlim=c(0,500),ylim = c(0,35), col=rgb(0,0,1,0.5), add=T)
> legend("topright",inset=c(-0.3, 0), legend=c("Bluegill","Largemouth Bass"),bty="n",col=c(rgb
(1,0,0,0.5), rgb(0,0,1,0.5)),pch=15,cex=0.8,text.font=4)
```

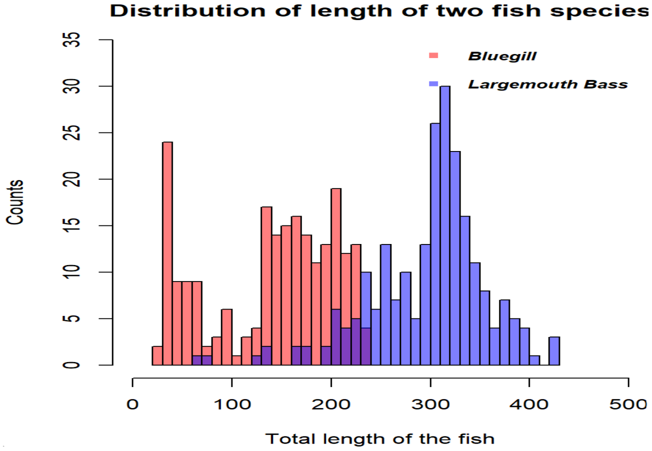


Figure 20 – Multiple Histogram A

- In Multiple Histogram A, I have compared total length of two species – Largemouth Bass and Bluegill. Length of Bluegill ranges from 0 to approx 240. Largemouth Bass has values distributed.
- Multiple histograms compare two distributions in the same plot.

```
> hist(w1, breaks=20, xlim=c(0,750),ylim = c(0,70), col=rgb(1,0,0,0.5), xlab="Weight
 of the fish", ylab="Counts", main="Distribution of weight of two fish species" )
> hist(w2, breaks=60, xlim=c(0,750),ylim = c(0,70), col=rgb(0,0,1,0.5), add=T)
> legend("topright",inset=c(-0.2, 0), legend=c("Bluegill","Largemouth Bass"),bty="n",
col=c(rgb(1,0,0,0.5), rgb(0,0,1,0.5)),pch=15,cex=0.8,text.font=4)
```

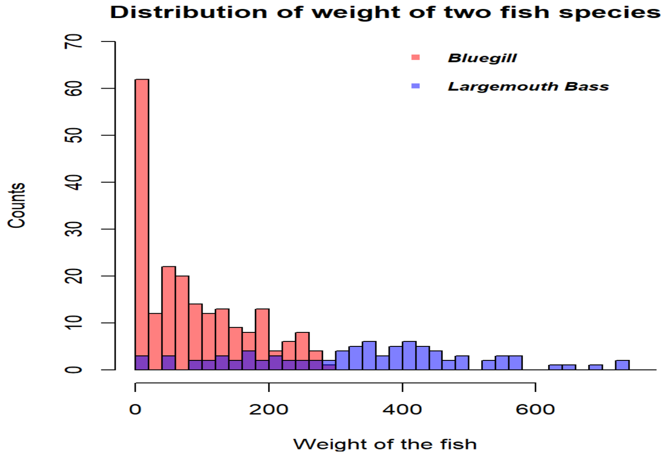


Figure 21 – Histogram B

- In Histogram B, I have compared weight of two species – Largemouth Bass and Bluegill. Bluegill species do not weigh more than 300. Largemouth Bass species are spread all over the plot.
- The red one is skewed right and the blue one seems multimodal.

```
> ggplot(data=bio,aes(x=tl,group=species,fill=species))+geom_density(adjust=1.5,alpha
=0.4)+labs(title="Multiple density plot",x="Fish length",y="Density")
```

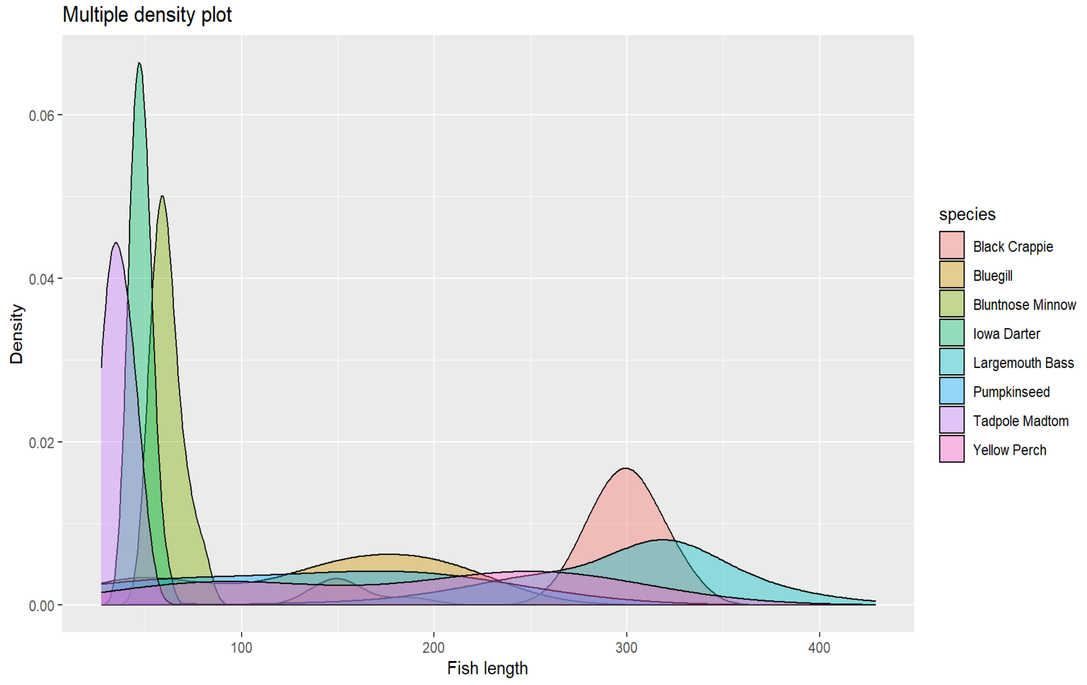
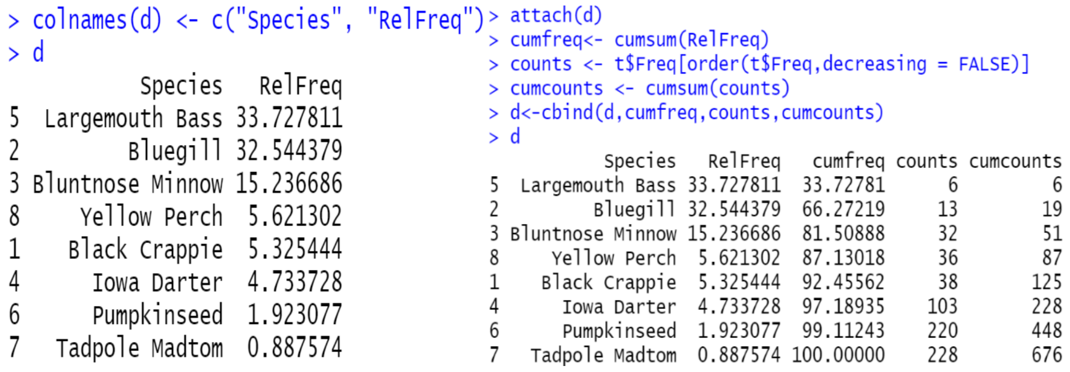


Figure 22 – Multiple density plot

- I have plotted multiple density plot against the fish length. We can see the variations of density in each fish species.

8. Adding ‘cumfreq’, ‘counts’, ‘cumcounts’ to dataframe



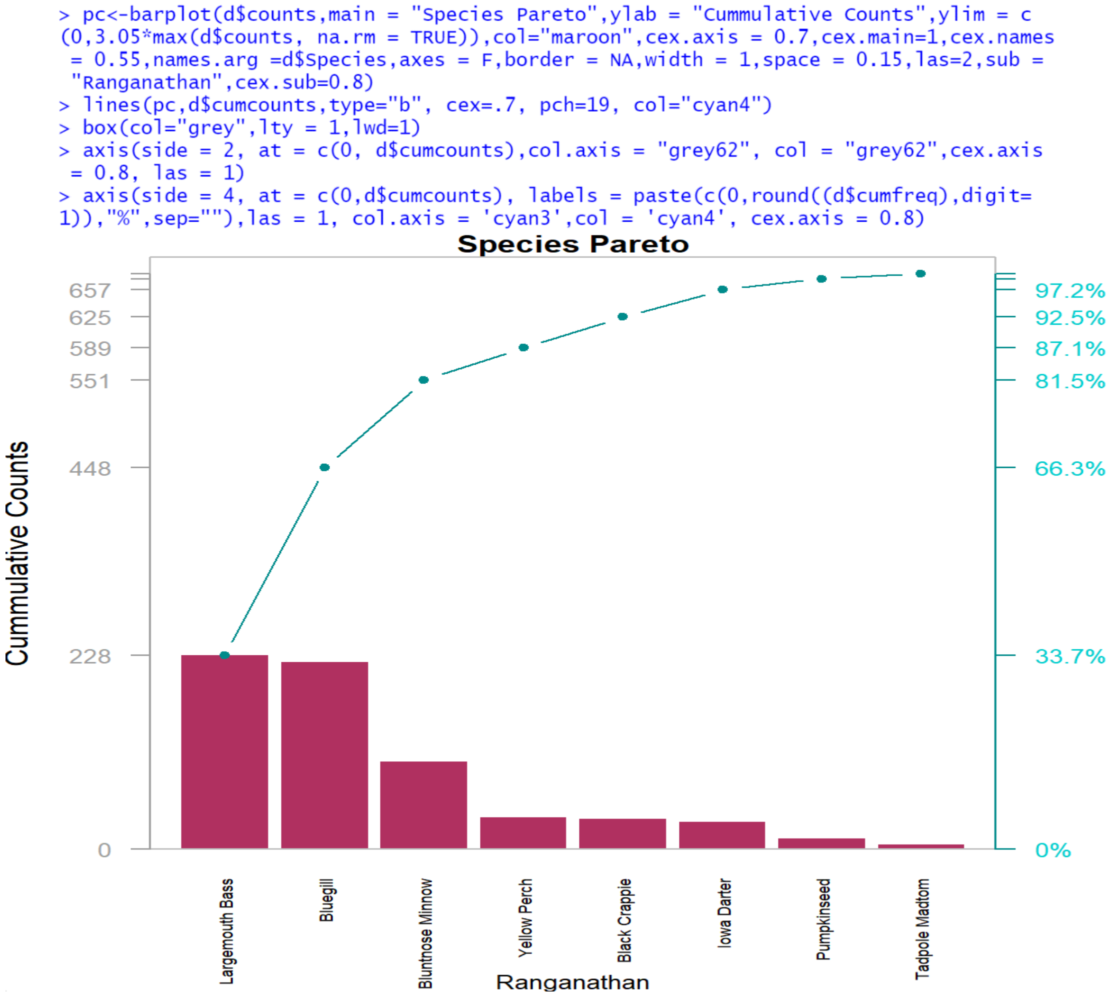
```
> colnames(d) <- c("Species", "RelFreq")
> d
          Species   RelFreq
5 Largemouth Bass 33.727811
2        Bluegill 32.544379
3 Bluntnose Minnow 15.236686
8     Yellow Perch  5.621302
1    Black Crappie  5.325444
4      Iowa Darter  4.733728
6      Pumpkinseed  1.923077
7   Tadpole Madtom  0.887574
```

```
> attach(d)
> cumfreq<- cumsum(RelFreq)
> counts <- t$Freq[order(t$Freq,decreasing = FALSE)]
> cumcounts <- cumsum(counts)
> d<-cbind(d,cumfreq,counts,cumcounts)
> d
           Species   RelFreq   cumfreq counts cumcounts
5  Largemouth Bass 33.727811  33.72781      6         6
2         Bluegill 32.544379  66.27219     13        19
3 Bluntnose Minnow 15.236686  81.50888     32        51
8     Yellow Perch  5.621302  87.13018     36        87
1    Black Crappie  5.325444  92.45562     38       125
4      Iowa Darter  4.733728  97.18935    103       228
6      Pumpkinseed  1.923077  99.11243    220       448
7   Tadpole Madtom  0.887574 100.00000    228       676
```

Figure 23 d dataframe

- d has the relative frequency of each species in descending order. I have applied cbind() in order to combine three columns (cumfreq, counts and cumcounts) to d.
- cumfreq is derived from running totals of relative frequency. Counts column is arranged in ascending order. cumcounts is worked out by adding the counts to the previous one.

9. Pareto Plot



```
> pc<-barplot(d$counts,main = "Species Pareto",ylab = "Cummulative Counts",ylim = c
(0,3.05*max(d$counts, na.rm = TRUE)),col="maroon",cex.axis = 0.7,cex.main=1,cex.names
 = 0.55,names.arg =d$Species,axes = F,border = NA,width = 1,space = 0.15,las=2,sub =
 "Ranganathan",cex.sub=0.8)
> lines(pc,d$cumcounts,type="b", cex=.7, pch=19, col="cyan4")
> box(col="grey",lty = 1,lwd=1)
> axis(side = 2, at = c(0, d$cumcounts),col.axis = "grey62", col = "grey62",cex.axis
 = 0.8, las = 1)
> axis(side = 4, at = c(0,d$cumcounts), labels = paste(c(0,round((d$cumfreq),digit=
1)),"%",sep=""),las = 1, col.axis = 'cyan3',col = 'cyan4', cex.axis = 0.8)
```

Figure 24 – Pareto Chart

- Pareto charts contain 2 Y axis and 1 X axis. It is made up of bar graph and line graph and the values are in descending order. Largemouth Bass and Bluegill constitute of around 66% of the fish population. Other species belong to remaining 34%.
- box() adds box around the plot. I have used subtitle() to include my name to the chart.

Summary

This assignment is an explanatory analysis that allows us to recognize the descriptive features of data, present them data visually and compare the observations of various fish species.

To summarize, we can see that two fish species Bluegill and Largemouth Bass are more in number as compared to other species. Also, according to the analysis, there is a rising curve seen which confirms positive correlation between weight and total length of the fish species. Tadpole Madtom and Pumpkinseed are at risk of disappearing in the future. Efforts and steps should be taken to protect and preserve them in order to restore balance in world's ecosystem.

The decline of fish population can be because of various factors like overfishing, global warming, adjustment of habitat and dominant species eating up the non-dominant ones.

Bibliography

Kabacoff, Robert.I. (2011). R in Action Data analysis and graphics with R. Manning

Bluman, Allan G. (2017). Elementary statistics A Step-by-step approach. McGraw Hill

The Data Digest. (2022 Jan 3). Scatterplots in R with geom_point() and geom_text/label() [Video File]. Retrieved from https://www.youtube.com/watch?v=sk59wjdmrd8

MarinStatsLectures-R Programming & Statistics. (2013 Aug 8). Subsetting (Sort/Select) Data in R with Square Brackets | R Tutorial 1.9| MarinStatsLectures [Video File]. Retrieved from https://www.youtube.com/watch?v=jGf7WNh-LX8

Statistics Globe. (2022 Feb 1). R How to Fix: Error in plot.new() : figure margins too large (Examples) | Change Plot Area | par mar [Video File]. Retrieved from https://www.youtube.com/watch?v=QrfRa9OG0dY

References (Websites):

https://www.geeksforgeeks.org/introduction-to-color-palettes-in-r-with-rcolorbrewer/

https://www.statology.org/legend-outside-plot-r/

https://r-coder.com/boxplot-r/

https://www.simplilearn.com/tutorials/statistics-tutorial/skewness-and-kurtosis

Appendix

```
---
title: "Module Project-3"
author: "Nikshita"
output:  word_document: default
date: "2022-10-09"
---
# 1. Print your name at the top of the script and load these libraries: FSA, FSAdata, magrittr,
dplyr, tidyr plyr and tidyverse
print("Nikshita Ranganathan")
```

```
install.packages(c("FSA","FSAdata","magrittr","dplyr","tidyr","plyr","tidyverse"))
library(FSA)
library(FSAdata)
library(magrittr)
library(dplyr)
library(plyr)
library(tidyr)
library(tidyverse)

# 2. Import the inchBio.csv and name the table <bio>
bio<-read.csv("inchBio.csv",stringsAsFactors = TRUE)
bio

# 3. Display the head, tail and structure of <bio>
head(bio)
tail(bio)
str(bio)
summary(bio)

# 4. Create an object, <counts>, that counts and lists all the species records
attach(bio)
counts<-species
table(counts)

# 5. Display just the 8 levels (names) of the species
levels(counts)

# 6. Create a <tmp> object that displays the different species and the number of record of
each species in the dataset. Include this information in your report.
tmp<-count(bio,"species")
tmp

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

# 8. Create a table, <w>, of the species variable. Display the class of w
w<-table(species)
w
class(w)

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

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

# 11. Create a table named <cSpec> from the bio species attribute (variable) and confirm that
you created a table which displays the number of species in the dataset <bio>
cSpec<-table(species)
cSpec
```

```
class(cSpec)

# 12. Create a table named <cSpecPct> that displays the species and percentage of records for
each species. Confirm you created a table class.
cSpecPct<-(table(species))*100/length(species)
cSpecPct
class(cSpecPct)

# 13. Convert the table, <cSpecPct>, to a data frame named <u> and confirm that <u> is a
data frame
u<-as.data.frame(cSpecPct)
u
class(u)

# 14. Create a barplot of <cSpec> with the following: titled Fish Count with the following
specifications:
# • Title: Fish Count
# • Y axis is labeled "COUNTS"
# • Color the bars Light Green
# • Rotate Y axis to be horizontal
# • Set the X axis font magnification to 60% of nominal
par(mar=c(4,5,3,1))
barplot(cSpec,main = "Fish Count",xlab = "COUNTS",col = "lightgreen",cex.axis = 0.6,horiz
= TRUE,las=1,cex.names = 0.5,xlim=c(0,250))

# 15. Create a barplot of <cSpecPct>, with the following specifications:
# • Y axis limits of 0 to 4
# • Y axis label color of Light Blue
# • Title of "Fish Relative Frequency"
par(mar=c(6,3,2,1))
barplot(cSpecPct/10,main = "Fish Relative Frequency",col = "lightblue",ylim =
c(0,4),las=2,cex.names = 0.6)

# 16. Rearrange the <u> cSpec Pct data frame in descending order of relative frequency. Save
the rearranged data frame as the object <d>
d<-u[order(u$Freq,decreasing = TRUE),]
d
class(d)

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

# 18. Add new variables to <d> and call them cumfreq, counts, and cumcounts
attach(d)
cumfreq<- cumsum(RelFreq)
counts <- t$Freq[order(t$Freq,decreasing = TRUE)]
cumcounts <- cumsum(counts)
d<-cbind(d,cumfreq,counts,cumcounts)
d

# 19. Create a parameter variable <def_par> to store parameter variables
def_par<-names(d)
```

```
def_par

# 20. Create a barplot, <pc>, with the following specifications:
# • d$counts of width 1, spacing of .15
# • no boarder
# • Axes: F
# • Yaxis limit 0,3.05*max
# • d$counts na.rm is true
# • y label is Cummulative Counts
# • scale x axis to 70%
# • names.arg: d$Species
# • Title of the barplot is “Species Pareto”
par(mar=c(5,3.8,1,3))
pc<-barplot(d$counts,main = "Species Pareto",ylab = "Cummulative Counts",ylim =
c(0,3.05*max(d$counts, na.rm = TRUE)),col="maroon",cex.axis =
0.7,cex.main=1,cex.names = 0.55,names.arg =d$Species,axes = F,border = NA,width =
1,space = 0.15,las=2)

# 21. Add a cumulative counts line to the <pc> plot with the following:
# • Spec line type is b
# • Scale plotting text at 70%
# • Data values are solid circles with color cyan4
lines(pc,d$cumcounts,type="b", cex=.7, pch=19, col="cyan4")

# 22. Place a grey box around the pareto plot
box(col="grey",lty = 1,lwd=1)

# 23. Add a left side axis with the following specifications
# • Horizontal values at tick marks at cumcounts on side 2
# • Tickmark color of grey62
# • Color of axis is grey62
# • Axis scaled to 80% of normal
axis(side = 2, at = c(0, d$cumcounts),col.axis = "grey62", col = "grey62",cex.axis = 0.8, las =
1)

# 24. Add axis details on right side of box with the specifications:
# • Spec: Side 4
# • Tickmarks at cumcounts with labels from 0 to cumfreq with %,
# • Axis color of cyan5 and label color of cyan4
# • Axis font scaled to 80% of nominal
axis(side = 4, at = c(0,d$cumcounts), labels =
paste(c(0,round((d$cumfreq),digit=1)),"%",sep=""),las = 1, col.axis = 'cyan3',col = 'cyan4',
cex.axis = 0.8)

# 25. Display the finished Species Pareto Plot (without the star watermarks). Have your last
name on the plot
pc<-barplot(d$counts,main = "Species Pareto",ylab = "Cummulative Counts",ylim =
c(0,3.05*max(d$counts, na.rm = TRUE)),col="maroon",cex.axis =
0.7,cex.main=1,cex.names = 0.55,names.arg =d$Species,axes = F,border = NA,width =
1,space = 0.15,las=2,sub = "Ranganathan",cex.sub=0.8)
lines(pc,d$cumcounts,type="b", cex=.7, pch=19, col="cyan4")
box(col="grey",lty = 1,lwd=1)
```

```
axis(side = 2, at = c(0, d$cumcounts),col.axis = "grey62", col = "grey62",cex.axis = 0.8, las = 1)
axis(side = 4, at = c(0,d$cumcounts), labels = paste(c(0,round((d$cumfreq),digit=1)),"%",sep=""),las = 1, col.axis = 'cyan3',col = 'cyan4', cex.axis = 0.8)

# Analyzing dataframe bio
var(bio$tl)
sd(bio$tl)
var(bio$w,na.rm=TRUE)
sd(bio$w,na.rm=TRUE)

# Lollipop chart
library("ggplot2")
cSpec<-as.data.frame(cSpec)
Freq<-cSpec$Freq
species<-cSpec$species
ggplot(data=cSpec,aes(x=Freq,y=species,label=Freq))+geom_point()+geom_segment(x=0,xend=Freq,y=species,yend=species)+labs(title="Lollipop chart representing number of each species",x="Frequency",y="Fish Species")+geom_text(vjust=-1)

# Pie Chart
library(RColorBrewer)
color <- brewer.pal(n=8, "Set2")
cSpecPct$percent=round(cSpecPct$Freq,digits = 1)
par(mar=c(1,1,1,1))
cSpecPct=as.data.frame(cSpecPct)
cSpecPct$percent=paste(cSpecPct$percent,"%",sep="")
pie(cSpecPct$Freq,labels=cSpecPct$percent,main="Piechart of Fish Species",col = color)
legend("topleft",inset=c(-0.4, 0),legend = cSpecPct$species,fill = color,cex=0.8,text.font = 4,bty = "n")

# Scatterplot
ggplot(data = bio, aes(x = tl, y = w))+geom_point(color= "coral")+labs (title ="Scatter Plot: tl vs w",x = "Length of Fish",y = "Weight of Fish")

cor(bio$w,bio$tl,use = "complete.obs")

library(moments)
skewness(bio$tl)
skewness(bio$w, na.rm= TRUE)
kurtosis(bio$tl)
kurtosis(bio$w, na.rm= TRUE)

# Boxplots
ggplot(bio,aes(x=species,y=tl,fill=species))+labs(title="Boxplots of Fish length for each species",x="Fish species",y="Fish Length")+geom_boxplot()+theme(axis.text.x = element_text(size=8,angle=90),legend.position = "none")+stat_summary(fun="mean",color="blue",geom="point")+stat_summary(fun="mean",geom="text",col="blue",vjust=1.5,aes(label=paste("Mean:",round(..y..,digits=1))))
ggplot(bio,aes(x=species,y=tl,color=species))+geom_boxplot(outlier.colour = "maroon",outlier.shape=17,outlier.size=3)+coord_flip()+labs(title="Boxplots of weight for each fish species",x="Species",y="Weight")
```

```
#Histogram
bioBluegill<-subset(bio,species=="Bluegill")
tl1<-bioBluegill$tl
w1<-bioBluegill$w
bioLargemouthbass<-subset(bio,species=="Largemouth Bass")
tl2<-bioLargemouthbass$tl
w2<-bioLargemouthbass$w
par(mar=c(4,4,2,1))
hist(tl1, breaks=30, xlim=c(0,500),ylim = c(0,35), col=rgb(1,0,0,0.5), xlab="Total length of
the fish", ylab="Counts", main="Distribution of length of two fish species" )
hist(tl2, breaks=30, xlim=c(0,500),ylim = c(0,35), col=rgb(0,0,1,0.5), add=T)
legend("topright",inset=c(-0.3, 0), legend=c("Bluegill","Largemouth
Bass"),bty="n",col=c(rgb(1,0,0,0.5), rgb(0,0,1,0.5)),pch=15,cex=0.8,text.font=4)
hist(w1, breaks=20, xlim=c(0,750),ylim = c(0,70), col=rgb(1,0,0,0.5), xlab="Weight of the
fish", ylab="Counts", main="Distribution of weight of two fish species" )
hist(w2, breaks=60, xlim=c(0,750),ylim = c(0,70), col=rgb(0,0,1,0.5), add=T)
legend("topright",inset=c(-0.2, 0), legend=c("Bluegill","Largemouth
Bass"),bty="n",col=c(rgb(1,0,0,0.5), rgb(0,0,1,0.5)),pch=15,cex=0.8,text.font=4)

# Multiple Density plot
ggplot(data=bio,aes(x=tl,group=species,fill=species))+geom_density(adjust=1.5,alpha=0.4)+l
abs(title="Multiple density plot",x="Fish length",y="Density")
```