Northeastern University

Toronto, Canada

A report on

Executive Summary Report 3

Subject

Introduction to Analytics – ALY 6000

Guided by

Prof. Mohammad Shafiqul Islam

Submitted by

Name	NUID	Date of Submission
Raj Tank	002988601	03/02/2022

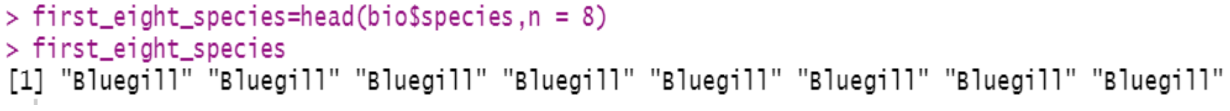
Introduction

The dataset of FSAdata(InchLake2) is primarily focuses on total weights and lengths of different fish species captured in InchLake2 from May, 2007 to May, 2008 by Derek H. Ogle, personal collection. In addition, this report consists of key findings, methodology, conclusion, bibliography, and appendix.

Key findings

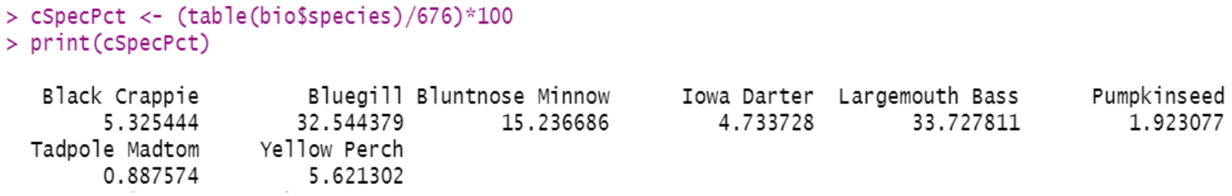
The dataset has only 676 observations, and 7variables (NetID, FishID, Species, Length, Weight, and Year of capture).

The first five species of the dataset by analyzing it via head function



```
> first_eight_species=head(bio$species,n = 8)
> first_eight_species
[1] "Bluegill" "Bluegill" "Bluegill" "Bluegill" "Bluegill" "Bluegill" "Bluegill" "Bluegill"
```

Group by count of all the species in the dataset by table function according to 676 observations

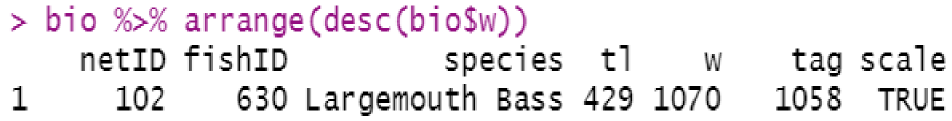


```
> cSpecPct <- (table(bio$species)/676)*100
> print(cSpecPct)

   Black Crappie        Bluegill Bluntnose Minnow      Iowa Darter  Largemouth Bass      Pumpkinseed 
        5.325444       32.544379       15.236686         4.733728        33.727811         1.923077 
  Tadpole Madtom    Yellow Perch 
        0.887574        5.621302 
```

As seen from the figure, Largemouth Bass accounted for 33 percent, which was the highest amongst all these species; while, Tadpole Madtom has the least amount of species 0.887574.

LargeMouth Bass species has the highest weight and length around 1070 and 429, respectively; whilst, Bluegill accounted for the lowest.



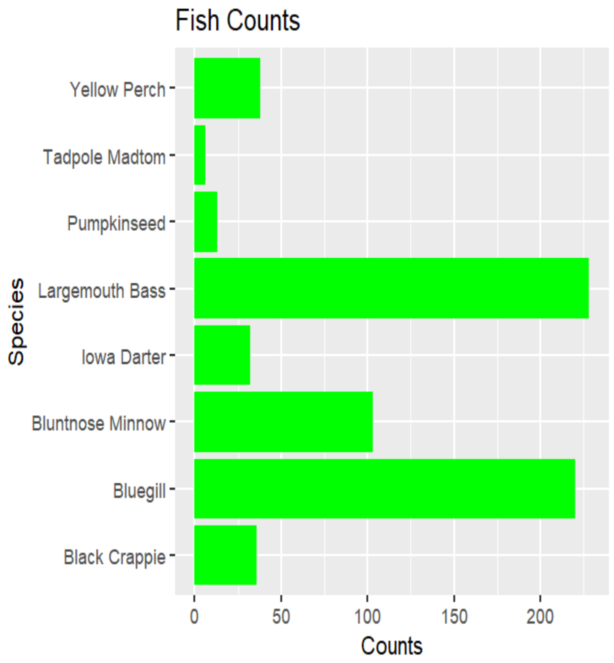
```
> bio %>% arrange(desc(bio$w))
    netID fishID         species  tl    w   tag scale
1     102    630 Largemouth Bass 429 1070  1058  TRUE
```

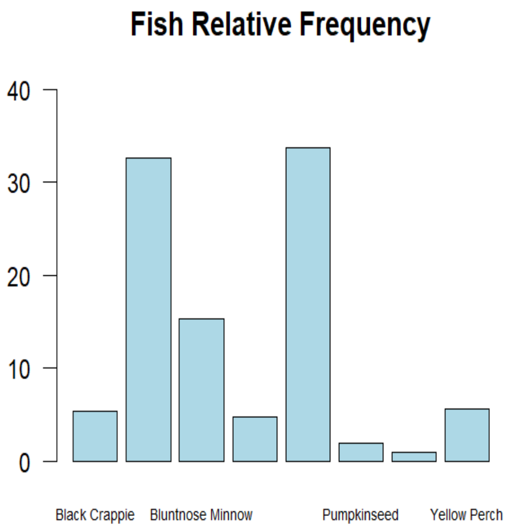
Many species of the fish do not have tag number in the dataset.

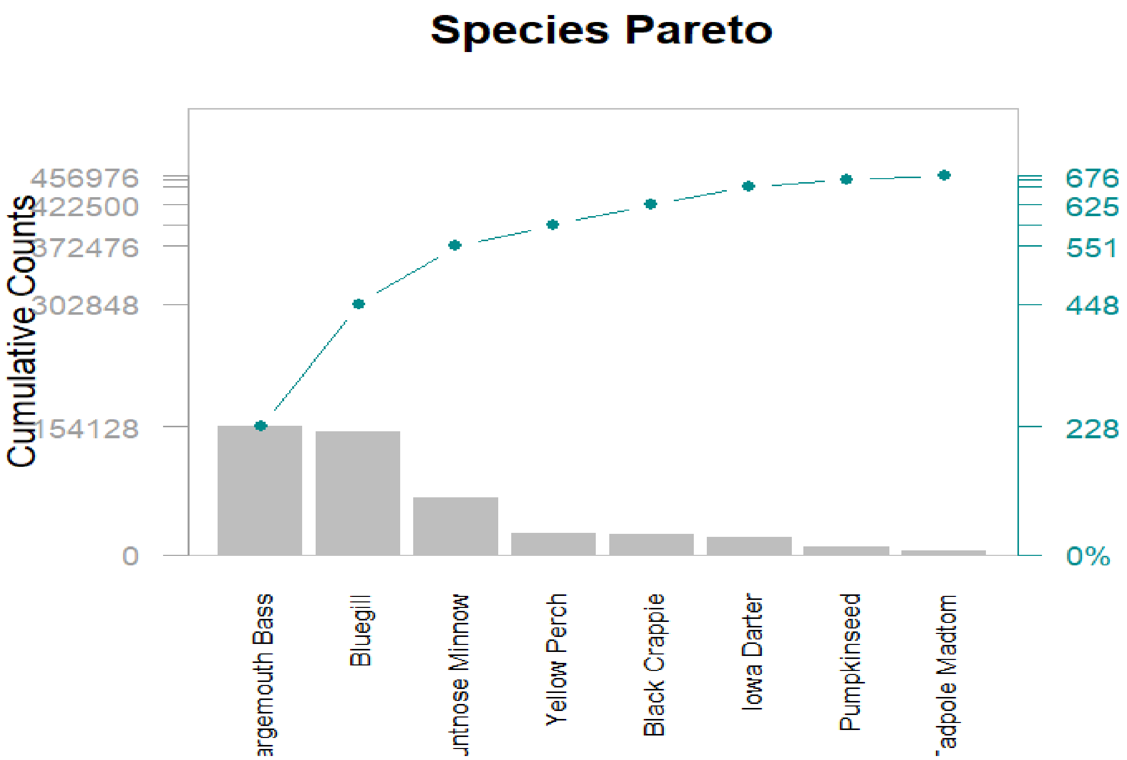
Also, the data provided and executed in R language has several Null values that need to be replaced by mean or 0 in order to get the exact analysis and for model preparation.

Methodology

Data visualization in R language with InchLake2(InchBio.csv) dataset provided in this report for analysis.







By considering graphs, Largemouth bass and Bluegill has the most counts as compared to other species.

Fish relative frequency shows the proportion out of 676 observations.

This report includes descriptive as well as inferential statistics. Furthermore, Null values need to be fixed first before analyzing the data. After, the data was analyzed and processed in R such as (Exploratory Data Analytics, plotting via simple plot, and ggplot2, filtration by dplyr(chaining and piping).

Conclusion

Considering all the points, graphs, details generated above, it can be recapitulated that:

The pareto chart has all the species with two axis chart and cumulative frequency. The axis shown in the graph at left side depicts cumulative counts, and the right axis illustrates cumulative frequency.

Many observations do not have scale; thus, the data has to be strongly scaled before making analysis report.

Surprisingly, Pumkinseed and Tadpole Madtom has the least count of the species throughout the observations captured by the author.

Bibliography

Stack Overflow (2010, December 24), *cex equivalent in ggplot2*. Retrieved from*:* https://stackoverflow.com/questions/4528161/cex-equivalent-in-ggplot2, Last accessed: February 3, 2022.

Galili, T. (2021, August 9), *How to Create Pareto Chart in R*, R-Bloggers. Retrieved from: https://www.r-bloggers.com/2021/08/how-to-create-pareto-chart-in-r/, Last accessed: February 3, 2022.

Ogle, D. H. (2007, May 5), *Fisheries Data by Package*, FishR. Retrieved from: https://derekogle.com/fishR/data/byPackage, Last accessed: February 3, 2022.

Pedersen, T. L. (2020, March 24), *ggplot2 workshop part 1*, YouTube. Retrieved from: https://www.youtube.com/watch?v=h29g21z0a68&feature=youtu.be, Last accessed: February 3, 2022.

Zach, Z. (2021, January 27), *How to Add a Column to a Data Frame in R (With Examples)*, Statology. Retrieved from: https://www.statology.org/r-add-a-column-to-dataframe/, Last accessed: February 3, 2022.

Tank, R. (2022, February 3), *GitHub - rajtank/week_3_R_Project*, GitHub. Retrieved from: https://github.com/rajtank/week_3_R_Project, Last accessed: February 4, 2022.

Appendix

```
#1 printing my name and load the necessary libraries.
print("Raj Tank") #printing my name
#installing libraries
install.packages("FSA")
install.packages("FSAdata")
install.packages("magrittr")
install.packages("dplyr")
install.packages("tidyr")
install.packages("plyr")
install.packages("tidyverse")
#importing those libaries
library(FSA)
library(FSAdata)
library(magrittr)
library(dplyr)
library(tidyr)
library(plyr)
library(tidyverse)

#2 importing the dataset "inchBio.csv".
#setting up the working directory
setwd("C:/Users/baps/Downloads")
bio=read.csv("inchBio.csv") #naming the table bio
bio
```

```
#3 Displaying the headtail structure of the dataset.
head_tail=headtail(bio)
print(head_tail)

#4 creating "<counts>", which counts and lists all the species records.
counts=bio$species
counts
#displaying the species
spec_records=bio$species
spec_records

#5 Depicting 8 species' names.
first_eight_species=head(bio$species,n = 8)
first_eight_species

#6 creating "<tmp>" that contains different species and records of it.
tmp=count(bio$species)
tmp

#7 creating a subset "<tmp2>" of just species variable, and display 5 records.
tmp2=bio$species
head(tmp2,n=5) # to get only 5 recors of the dataframe

#8 creating a table "<w>", and display the class of it.
```

```
w<-table(bio$species)
w
#checking the class of the variable
class(w)

#9 conver"<w>" to dataframe named "<t>",and print results.
# converting table into dataframe
t=as.data.frame(w)
t
class(t)

#10 Extracting and displaying the frequency values from "<t>" dataframe
extracted_freq=t$Freq
#displaying the frequency
extracted_freq

#11 creating a table named "<cSpec>" from the bio and that table displays the
#number of species
cSpec<-table(bio$species)
cSpec
class(cSpec)

#12 creating a table name "<cSpecPct>" that displays the species and percentage
#of records. Also confirm class of it
cSpecPct <- (table(bio$species)/676)*100
```

```
print(cSpecPct)
class(cSpecPct)

#13 converting the table "<SpecPct>"to dataframe named"<u>" and confirm
#the class is dataframe
u=as.data.frame(cSpecPct)
class(u)
class(cSpec)
cSpec=as.data.frame(cSpec)
class(cSpec)
cSpec
class(cSpecPct)
cSpecPct=as.data.frame(cSpecPct)
cSpecPct
class(cSpecPct)

#14 creating a barplot of "<cSec>" with the following details mentioned below
ggplot(cSpec, aes(y=Freq, x=Var1)) +
  geom_bar(stat = "identity",fill="green") +
  coord_flip()+labs(y="Counts",x="Species",title = "Fish Counts")
#using barplot method
barplot(height = cSpec$Freq,names=cSpec$Var1,col = "green",
      horiz = TRUE,ylab ="Species",xlab = "COUNTS",main = "Fish
Counts",cex.lab =0.60)

#15 creating a braplot of "<cSpecPct>" with following details mentioned below
```

```
barplot(height = cSpecPct$Freq,names=cSpecPct$Var1,las=1,
        col = "lightblue",ylim=c(0,40),cex.names = 0.60,
        main ="Fish Relative Frequency")

#16 rearranging "<u>" cSpecPct dataframe in descneding order of relative
#frequency and save that object as "<d>"
d<-cSpec %>% arrange(desc(cSpec$Freq))
d

#17 rename <d> column and var 1 to species, and Freq to RelFreq
#renaming columns
colnames(d)[colnames(d)=="Var1"] <- "Species"
colnames(d)[colnames(d)=="Freq"] <- "RelFreq"
class(d)
d

#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))
print(d)
view(d)

#19 adding new variable ",def_par>" to store parameter variable
def_par=par(no.readonly = TRUE)
def_par
```

```
#20 creating "<pc>" with following details mentioned below
pc<-barplot(d$counts,width = 1,ylab="Cumulative Counts",
        main ="Species Pareto",,cex.names = 0.7,names.arg=d$Species,las=2,
        border = NA,axes =F, space= 0.15,
        ylim=c(0,3.5)*max(d$counts,na.rm = TRUE))

#21 adding cumulative count line to "<pc>" with following details
lines(pc,d$cumcounts,type="b",cex=0.7,pch=16,col="cyan4")

#22 placing a grey box around pareto chart
box(col="grey")
#reference:https://www.statmethods.net/advgraphs/parameters.html

#23 adding left side axiswith following details
axis(at = c(0,d$cumcounts),side = 2,las=1, col.axis="grey62",cex.axis=0.8,
    col = "grey62")

#24 adding axis on right side of the box with following details
axis(side = 4,cex.axis=0.80,col = "cyan4",col.axis="cyan4",at = c(0,d$cumcounts),
    las=1,labels = paste(c(0,round(d$cumfreq*100)),"%",sep=""))

#25 displaying the finished species pareto chart, having my last name on the plot
pc<-barplot(d$counts,width = 1,ylab="Cumulative Counts",
        main ="Species Pareto(Raj Tank)",,cex.names =
0.7,names.arg=d$Species,las=2,
          border = NA,axes =F, space= 0.15,
          ylim=c(0,3.5)*max(d$counts,na.rm = TRUE))
lines(pc,d$cumcounts,type="b",cex=0.7,pch=16,col="cyan4")
box(col="grey")
axis(at = c(0,d$cumcounts),side = 2,las=1, col.axis="grey62",cex.axis=0.8,
    col = "grey62")
axis(side = 4,cex.axis=0.80,col = "cyan4",col.axis="cyan4",at = c(0,d$cumcounts),
    las=1,labels = paste(c(0,round(d$cumfreq*100)),"%",sep=""))
```

#26 commit my repo in Github