Toronto

Soni_M2_Project2

Soni Manan[a]

[a] *College of Professional Studies, Master of Professional Studies in Analytics.* ***NUID****: 002982645*

Under the guidance of

Prof. Dr. M. Shafiqul Islam

Introduction

A short dataset of fishes found in North America. It includes fishes such as bluegill, largemouth bass, bluntnose minnow, yellow perch, black crappie, iowa darter, pumpkinseed, tadpole madtom. Dataset is largely cover by the samples of largemouth bass and bluegill.

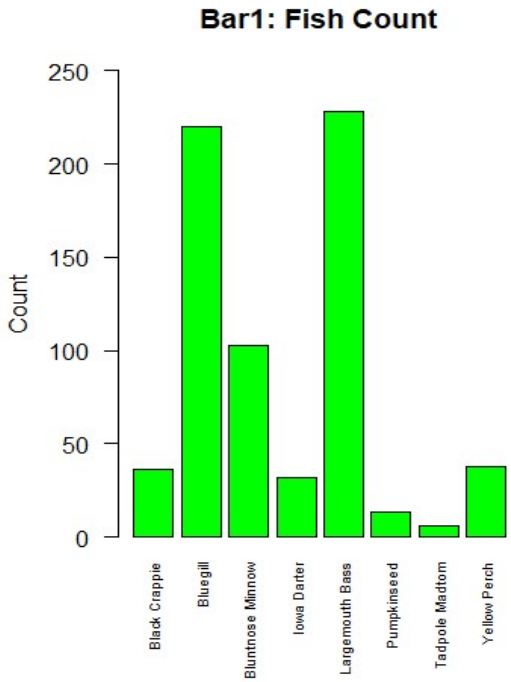
Objective

Primarily, a dataset of fishes has been provided. To get insights of it we need to separate out all of the fishes and finding their frequency distribution, cummulative frequency, relative frequency.

Methodology

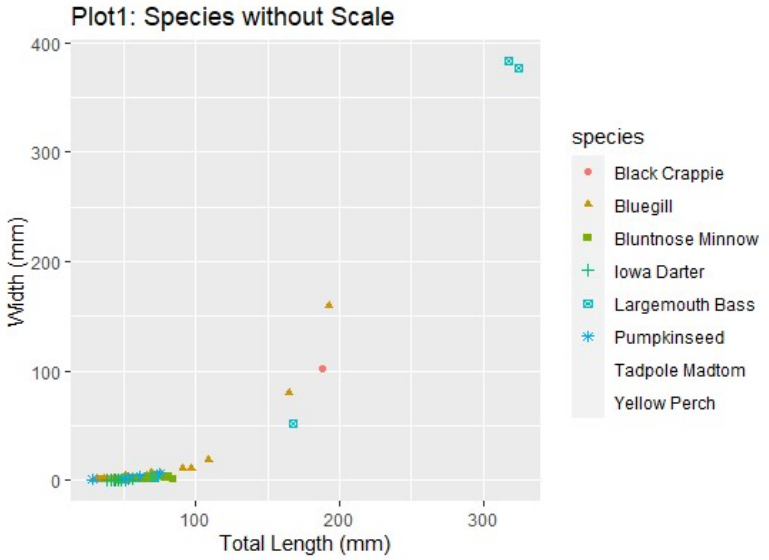
A. Observation of Dataset

Dataset contains feature variables such as Species, tl (Total Length) and w (width). It contains samples of 8 different species of fish. 165 values of 676 samples are missing in width. All of the characteristic are spread out due to different species so the standard deviation are meaning less until we separate out. As shown in Bar1, Large proportions have covered by bluegill and largemouth bass.



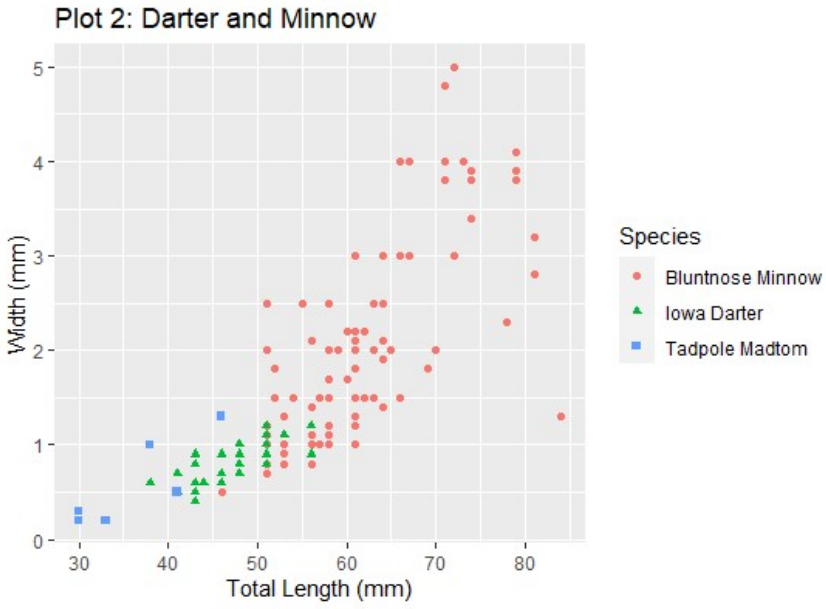
B. Findings

spread out due to different species so the standard deviation are meaning less until we separate out. As shown in Plot1, most of the fishes who aren't scaled have length less than 100,, and width is around 0 mm and some outliers are there as well.



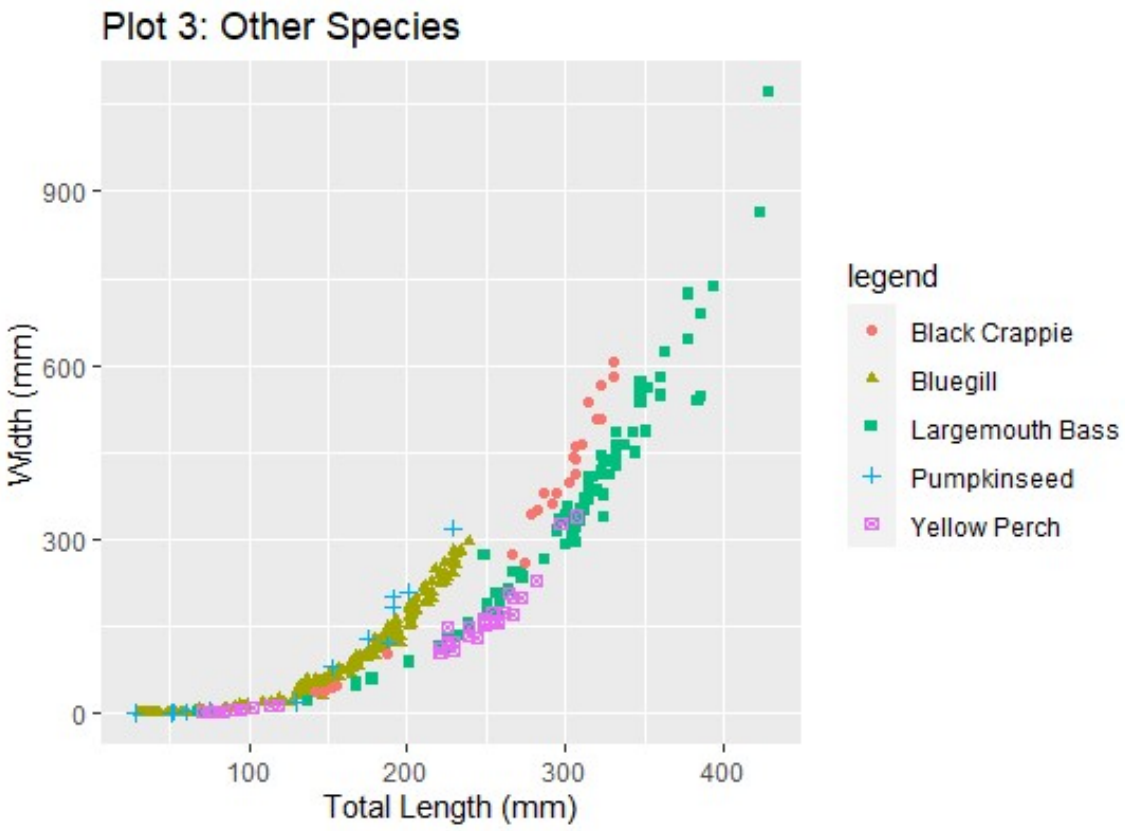
B.1 Tadpole Madtom, Iowa Darter & Bluntnose Minnow

According to Plot 2, there is linear incliment but data is spread out as standard deviation of length is 10.33 but width is pretty close to mean (1.68) because standard deviation is 1.05 of width. There are some outliers as Bluntnose Minnow with length of 80mm should have width of greater than 4 but its less than 2. However, for tadpole madtom data is ambigiuos as mean width of tadpole madtom is 0.58 mm which is very low and mean of egg diameter is 1.95 mm for Tadpole Madtom[1].



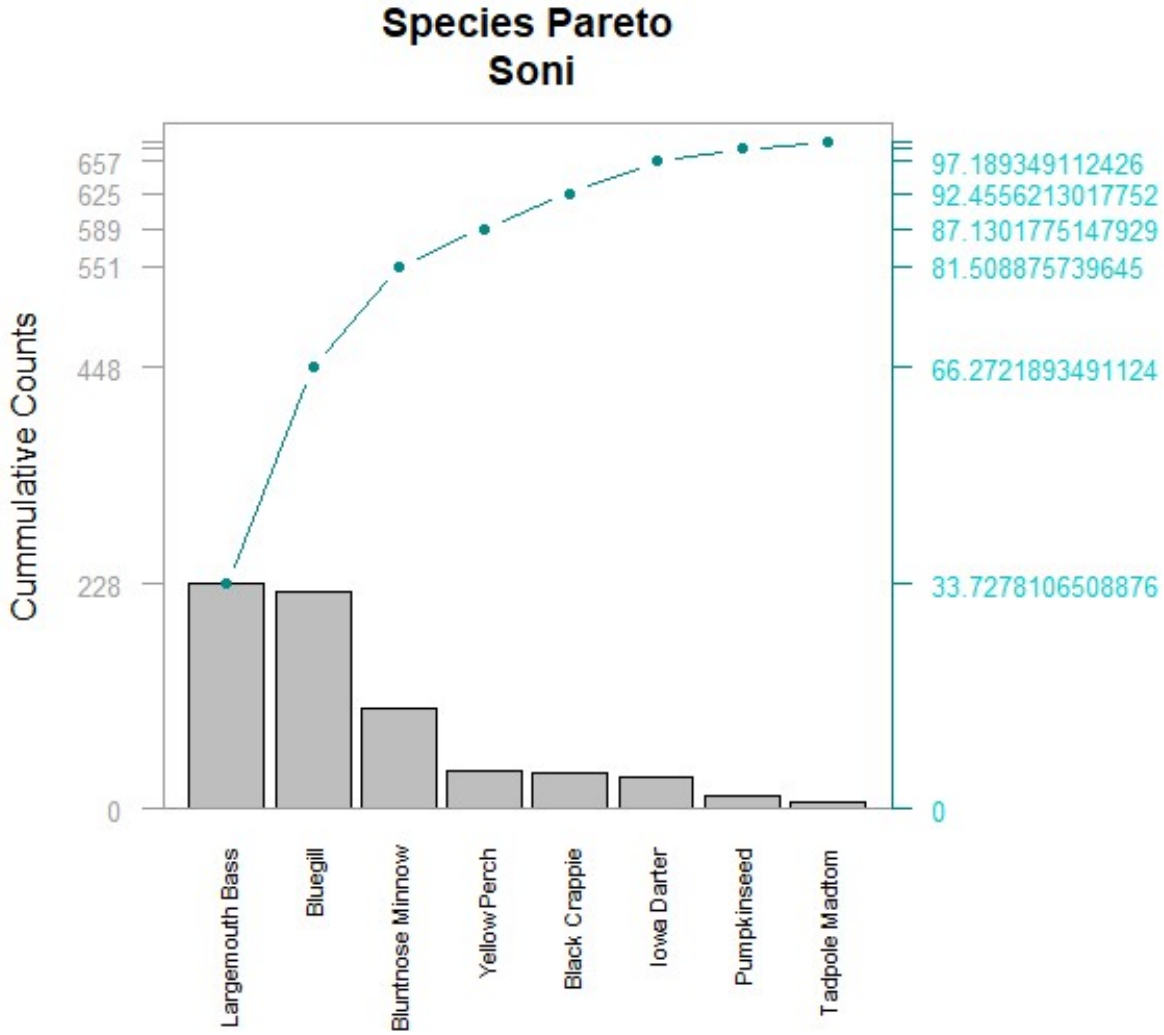
B.2 Remaining species

As shown in Plot 3, It's clear that all the fish is grows linearly in their lifespan. Biggest among them are Largemouth Bass as common length for it is 400 mm[2]. There are less number of samples collected of Pumpkinseed. To note, not a single Bluegill has crossed width of 310 mm.



Conclusion

As seen in pareto plot, 65% of samples are of Largemouth Bass and Bluegill. Bluntnose Minnow adds another 15% in samples and rest of them are around 20%. The place from where samples have been collected, has species as distributed in Pareto plot.



Bibliography

[1] Rohde, F.C. 1980. *Noturus gyrinus* (Mitchill), Tadpole madtom *Source:* http://txstate.fishesoftexas.org/noturus%20gyrinus.htm

[2] *Google Search*. (2022, Feb 4). Retrieved from Google: https://www.google.com/search?q=average+size+of+largemouth+bass

[3] Hadley, et al. "Subset rows using column values." the Grammar of Data Manipulation • dplyr, RStudio, https://dplyr.tidyverse.org/index.html.

[4] Kabacoff, Robert. R In Action: Data Analysis and Graphics with R. Manning, 2015.

[5] Tutorialspoint. R tutorial, https://www.tutorialspoint.com/r/index.htm

[6] Wickham, Hadley, et al. "Create Elegant Data Visualizations Using the Grammar of Graphics." Create Elegant Data Visualizations Using the Grammar of Graphics • ggplot2, RStudio, https://ggplot2.tidyverse.org/index.html.

Appendix

```
# SONI_M3_Project3

# install.packages(c('FSAdata','plyr','FSA','dplyr','plotrix','moments','ggplot2',,'magrittr',
'tidyr','tidyverse'))
library('FSAdata')
library('magrittr')
library('FSA')
library('plotrix')
library('moments')
library('dplyr')
### For setting working directory dynamically
#install.packages('rstudioapi')
setwd(dirname(rstudioapi::getActiveDocumentContext()$path  ))
print(getwd())

#2. Importing
bio = read.csv('./inchBio.csv')

# 3. Displaying last and first 3 rows of <bio>
headtail(bio, n=3)

# 4. List of species and their count
counts = bio %>% dplyr::count(species)
print(counts)

# 5. names of species
uniques = unique(bio$species)
print(uniques)

# 6.  displaying the different species and the number of record of each species
tmp = bio %>% group_by(species)
print(tmp)

#7. subset of species
tmp2 =tmp$`bio$species`
print(tmp2)

# 8. creating table <w>
w = table(bio$w,bio$species)
print(w)

# 9.
t = data.frame(w)

# 10.
t$Freq

# 11. Table
cSpec = table(bio$species)
cSpec
is.table(cSpec)
```

```
# 12. creating <cSpecPct>
library(ggplot2)
cSpecPct <- table(bio$species, bio$netID)
cSpecPct
class(cSpecPct)
# getting percentage
cSpecPct <- rowSums(prop.table(cSpecPct)) * 100
cSpecPct

# 13. Datafram of cSpecPct to <u>
u = data.frame(cSpecPct)
is.data.frame(u)

par(las=2)
par(mar=c(7,5,4,2))
# 14. Barplot of cSpec
barplot(cSpec, main= "Bar1: Fish Count",  ylim=c(0,250),ylab="Count", col="green", las=2, cex.
names = 0.60,
        names.arg = c("Black Crappie", "Bluegill", "Bluntnose Minnow", "Iowa Darter", " Largem
outh Bass",
                      "Pumpkinseed", "Tadpole Madtom", "Yellow Perch"))

# 15. Barplot of cSpecPct
barplot(u$cSpecPct, main="Fish Relative Frequency", ylab="Percentage", ylim = c(0,40),col="lig
htblue", cex.names = 0.60, las= 2,
        names.arg = c("Black Crappie", "Bluegill", "Bluntnose Minnow", "Iowa Darter", " Largem
outh Bass",
                      "Pumpkinseed", "Tadpole Madtom", "Yellow Perch"))

# 16.
d = bio %>% group_by(species) %>% dplyr::summarise(Freq = 100 * n()/nrow(bio))
d = d %>% arrange(desc(Freq))

# 17.
colnames(d) = c('Species','RelFreq')
d

# 18.
d['cumfreq'] = cumsum(d$RelFreq)
d['counts'] = counts$n
d['cumcounts'] = cumsum(counts$n)
d

# 19.
def_par = par(no.readonly = TRUE)
def_par

# 20
pc <- barplot(d$counts, width = 1, space = 0.15, axes = F, ylim = c(0,3.05*228), ylab = "Cummu
lative Counts",
              names.arg = d$Species, las=2,col='red', cex.names = 0.70, main = "Species Pareto
", d$counts, na.rm=TRUE)
```

```
# 21
pc <- barplot(d$counts, width = 1, space = 0.15, border = NA, axes = F, main = "Species Pareto
", ylim = c(0,3.05*228),col='cyan2', ylab = "Cummulative Counts", names.arg = d$Species, las=2
, cex.names = 0.70)
lines(pc, d$cumcounts, type = "b", cex = 0.7, pch = 19, col="cyan4")

# 22
#Placeing a grey box around the pareto plot
lines(pc, d$cumcounts, type = "b", cex = 0.7, pch = 19, col="cyan4")
box(col = "grey62")

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

# 24
axis(side = 4, at = c(0, d$cumcounts), labels = c(0, d$cumfreq),las = 1,
     col.axis = "cyan3", col = "cyan4", cex.axis = 0.8)
# 25
par(mar=c(6,5,4,8))
pc <- barplot(d$counts, width = 1, space = 0.15,axes = F, main = "Species Pareto\nSoni",
                ylim = c(0,3.05*228), ylab = "Cummulative Counts", names.arg = d$Species,
                las=2, cex.names = 0.70)
  lines(pc, d$cumcounts, type = "b", cex = 0.7, pch = 19, col="cyan4")
  box(col = "grey62")
  axis(side = 2, at = c(0, d$cumcounts), las = 1, col.axis = "grey62", col = "grey62", cex.axi
s = 0.8)
  axis(side = 4, at = c(0, d$cumcounts), labels = c(0, d$cumfreq),las = 1,
       col.axis = "cyan3", col = "cyan4", cex.axis = 0.8)

# Basic barplot
p<-ggplot(data=data.frame(cSpec),aes(x=Var1,y=Freq)) +
  geom_bar(stat="identity",fill="green") + labs(x="Species",y="Frequency", )
p
# Horizontal bar plot
p + coord_flip()

ggplot(data=percentage, aes(x=species,y=rFreq)) +
  geom_bar(aes(y = ..prop.., group = 1)) + lims(y = c(0,4))

names(bio) # column heads
summary(bio) # Mean, Median, Mode

bluegill = bio %>% filter(bio$species == "Bluegill")
bluntnoseMinnow = bio %>% filter(bio$species == "Bluntnose Minnow")
darter = bio  %>% filter(bio$species == "Iowa Darter")
largemouthBass = bio  %>% filter(bio$species == "Largemouth Bass")
pumpkinSeed = bio  %>% filter(bio$species == "Pumpkinseed")
tadpoleMadtom = bio  %>% filter(bio$species == "Tadpole Madtom")
yellowPerch = bio  %>% filter(bio$species == "Yellow Perch")
blackCrappie = bio  %>% filter(bio$species == "Black Crappie")
```

```
temp = bio

# Remove Na rows
na.omit(largemouthBass)
na.omit(bluegill)
na.omit(darter)
na.omit(pumpkinSeed)
na.omit(tadpoleMadtom)
na.omit(yellowPerch)
na.omit(blackCrappie)
na.omit(bluntnoseMinnow)

big = bio %>% filter(bio$species != "Iowa Darter")
big = big %>% filter(big$species !=  'Bluntnose Minnow')
big = big %>% filter(big$species !=  'Tadpole Madtom')
big$legend = as.factor(big$species)
big

small = bio %>% filter(species %in% c("Iowa Darter",'Bluntnose Minnow','Tadpole Madtom'))
small$Species = as.factor(small$species)
small

plot(darter$tl,darter$w,xlab="Total Length (mm)", ylab="Width (mm)", main="Bluegill Species of
inchBio", col="orange", pch=2)

plot(bluntnoseMinnow$tl,bluntnoseMinnow$w,xlab="Total Length (mm)", ylab="Width (mm)", main="B
luegill Species of inchBio", col="orange", pch=2)

ggplot(big, aes(x=tl, y=w, shape=legend, color=legend)) +
  geom_point() +
  labs(x="Total Length (mm)",y="Width (mm)",title = "Plot 3: Other Species", # adds title)

ggplot(small, aes(x=tl, y=w, shape=Species, color=Species)) +
  geom_point() +
  labs(x="Total Length (mm)", y="Width (mm)", title = "Plot 1: Darter and Minnow")

nonscaled = bio %>% filter(scale == "FALSE")
nonscaled = na.omit(nonscaled)

ggplot(nonscaled, aes(x=tl, y=w, shape=species, color=species)) +
  geom_point() +
  labs(x="Total Length (mm)", y="Width (mm)", title = "Plot1: Species without Scale")

scaled = bio %>% filter(scale == "TRUE")
scaled = na.omit(scaled)

ggplot(scaled, aes(x=tl, y=w, shape=species, color=species)) +
  geom_point() + labs(x="Total Length (mm)",y="Width (mm)",title = "Plot:1 Darter and Minnow")
```