*Explores dataset manipulation and graphical parameters necessary for barplots and pareto charts*

**EXECUTIVE SUMMARY**

Es

ALY6000 Introduction To Analytics

Module 3 Project 3

**PREPERATION:**

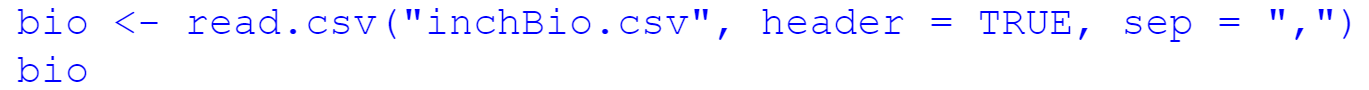
By: John DiSessa

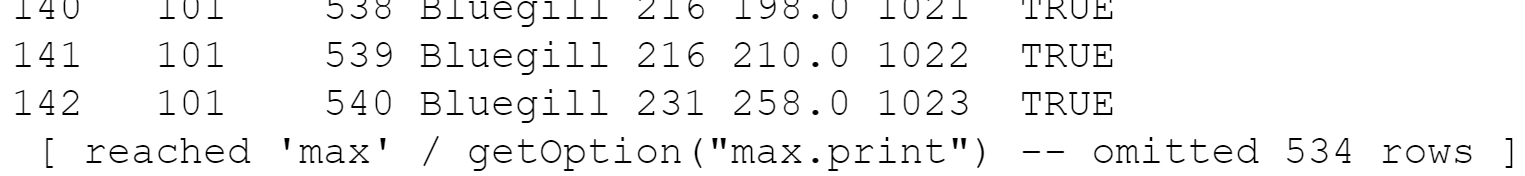
For: Professor Goulding

On: February 21st,2021

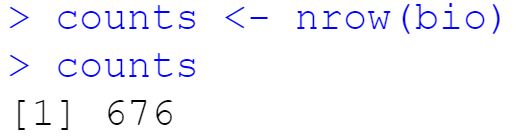
Introduction

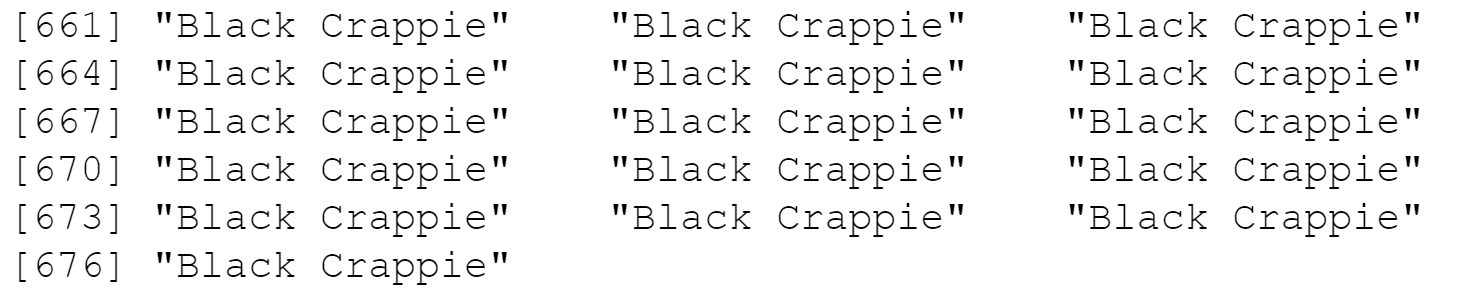
After opening the inchBio.csv file, I was surprised to find that I could not display the entire dataset because it reached the maximum allowed rows.



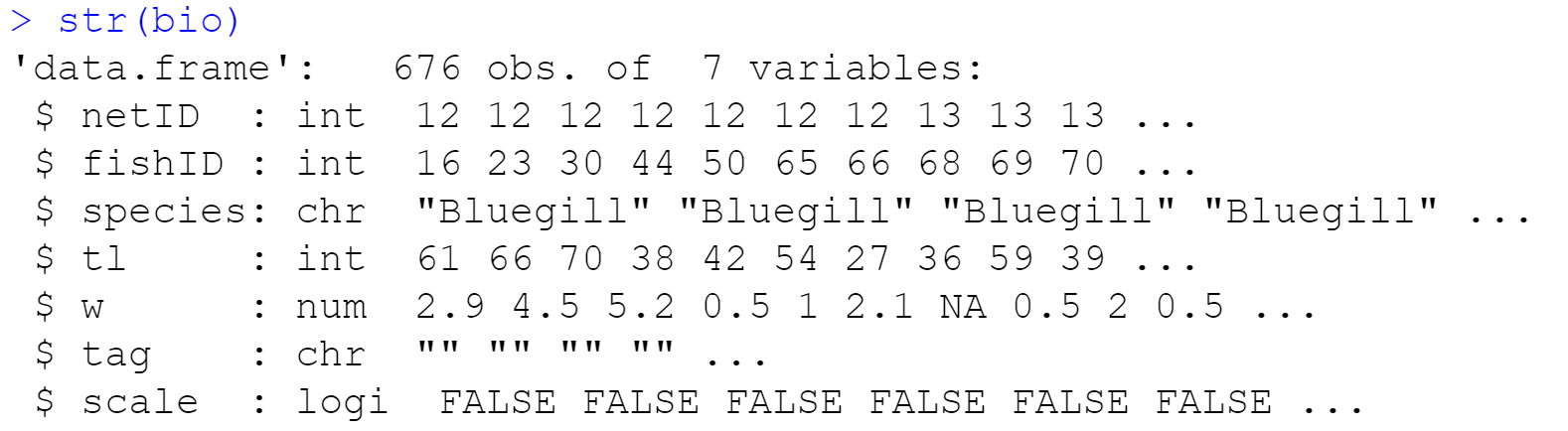


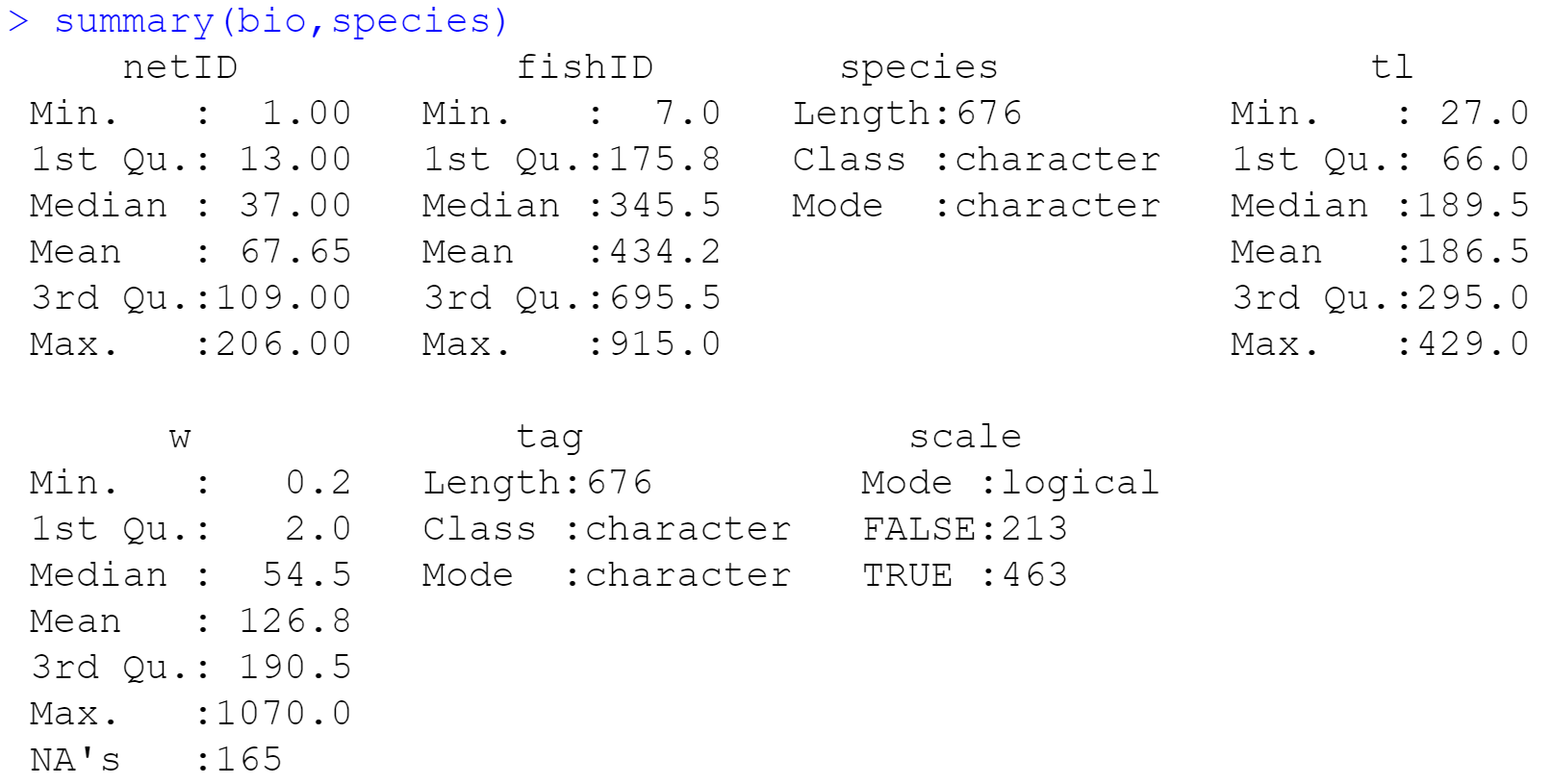
This was the first time this has ever happened to me, so I had to explore and find out the best way to summarize the dataset. I counted the number of rows (676) and listed all of the species in the data set to confirm there are 676.





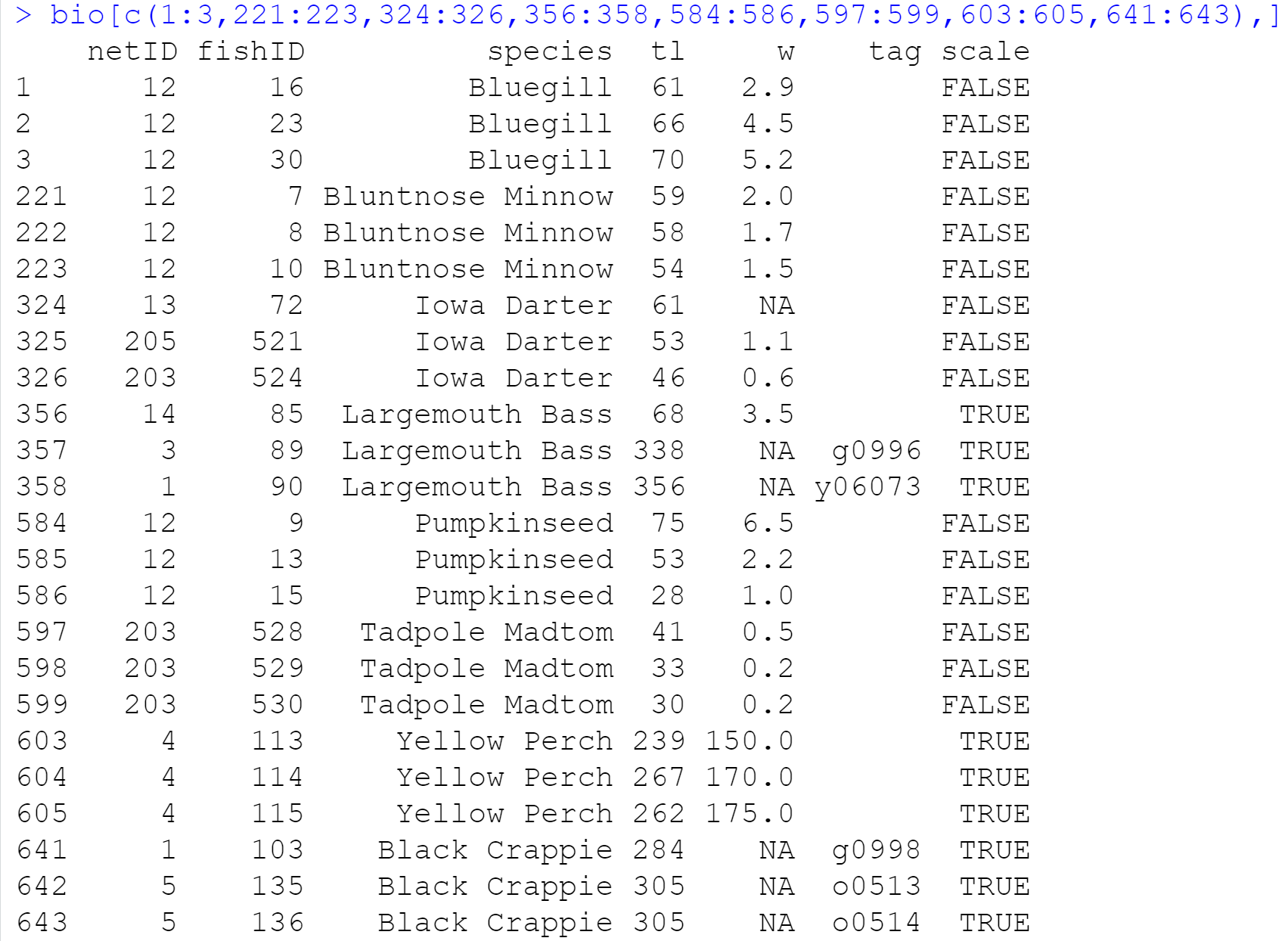
Even after displaying the list of species, I did not have a good understanding of the dataset as a whole. I looked at the structure, summary, variance, and standard deviation of inchBio.



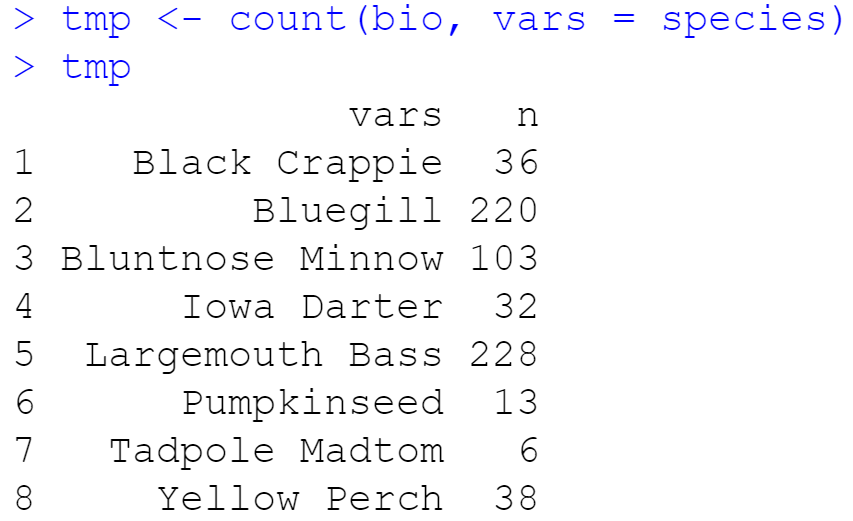


|  |  |  |
| --- | --- | --- |
|  | Variance | Standard Deviation |
| tl (tail length) | 12,010.94 | 109.59 |
| w (weight) | 27,940.96 | 167.16 |

These results were incredibly concerning to me and indicated there could be sample size issues between different species in the dataset or that each species was vastly different. I displayed the first three rows of each species from the dataset to get a quick look at any stark discrepancies.



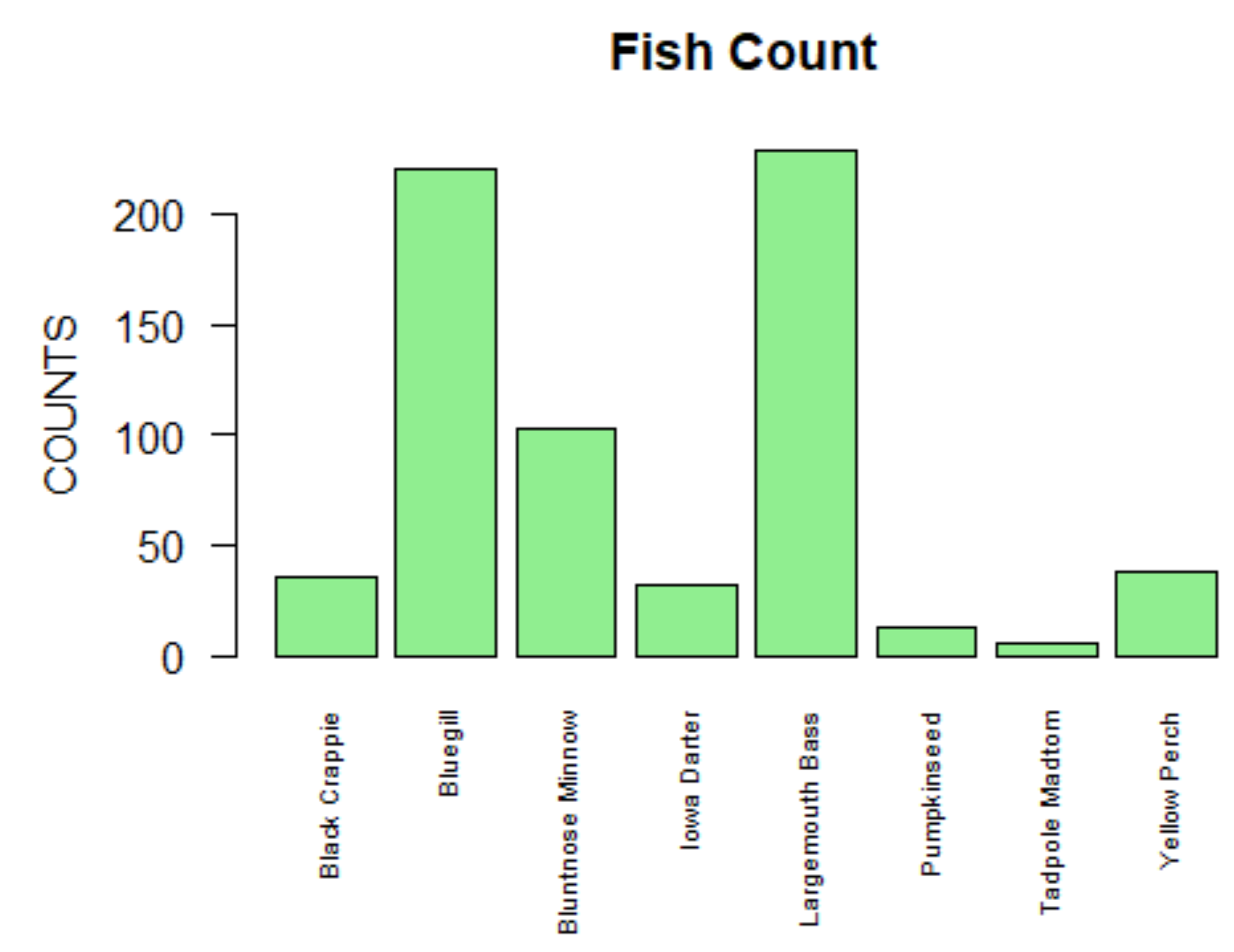
I then displayed the counts of each species in the dataset to confirm my sampling issue suspicions.



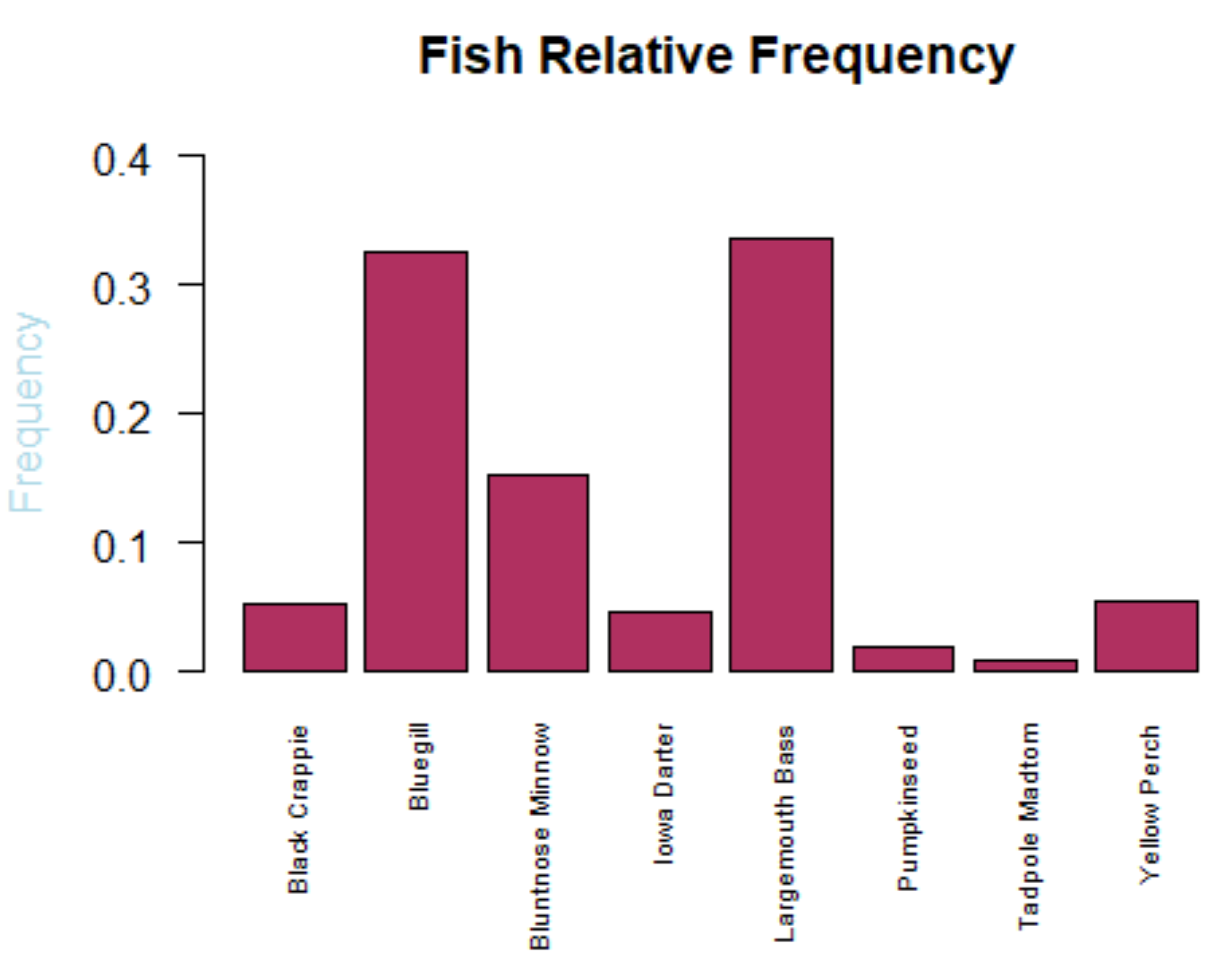
Now that I understood the issues with my initial data, I cautiously proceeded with my analysis.

Output

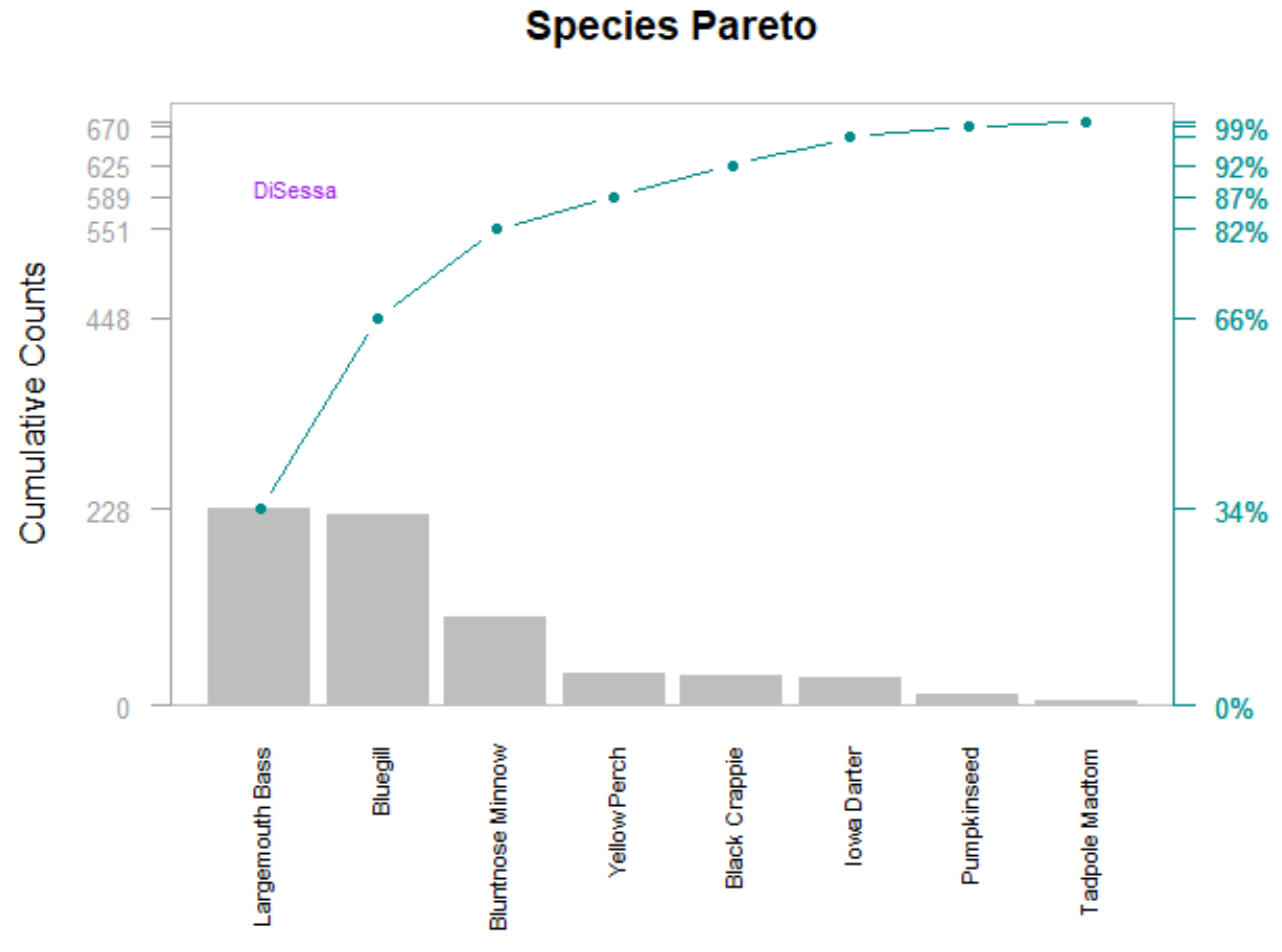
After initially summarizing the dataset, it was still difficult to grasp how much the Bluegill and Largemouth Bass samples skewed the results. I first created a barplot to visually show this sample disparity.



Next, I looked at the relative frequency of each species to further illustrate how the Bluegill and Largemouth Bass samples skew the dataset.

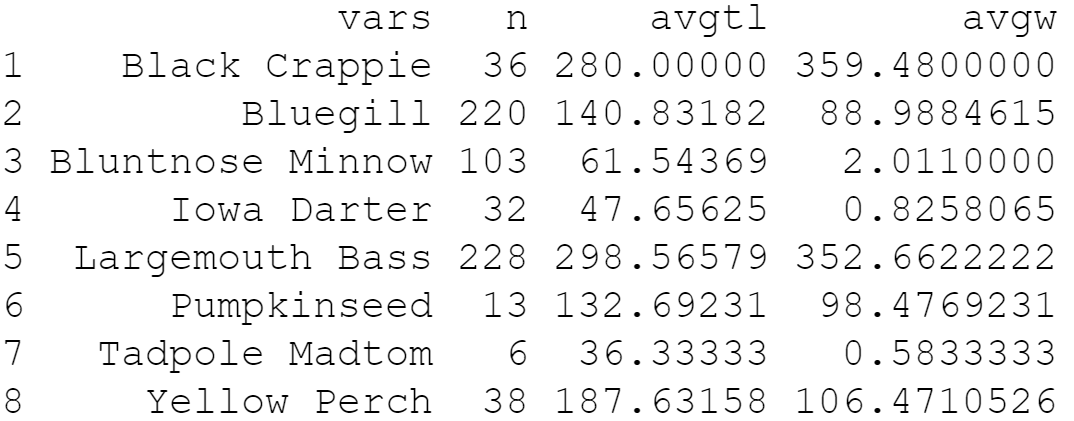


As you can see above, the two largest samples make up 66.2% of the entire inchBio dataset. I then created a Pareto chart as this is a simple graph that can illustrate both of the previous charts in one image.



Each species is sorted from most frequently appearing in the dataset to least frequent. Each y axis label on the left shows the cumulative count at each of the eight species. Each y axis label on the right shows the cumulative percentage of how much each species comprised the dataset. The line in the middle shows how fast the first two species make up most of the samples and how slow the remaining six species make up the rest.

Despite my concerns about the sampling, I looked at the average tail length and weight of each species. I was curious to see if the larger fish are preying on the smaller fish as a potential explanation as to the sample size discrepancy. As you can see below, neither variable seems to relate to the sample size (Black Crappies are the heaviest fish but only made up 5.3% of the data.



Summary

Even though I have proven how skewed our dataset is, there could be several factors contributing to this. First, there could be a sampling issue. Perhaps samples were taken from only certain parts or certain depths of the lake where Largemouth Bass and Bluegill fish primarily reside. It is also possible that the bait used by researchers was strongly preferred by these two fish and not the others. It is also possible that their nets were too big to easily capture small fish. I would be curious as to their collection methods and if they used different methods for different fish. If any of the above are true, the sampling issue must be corrected before any further analysis is conducted. My analysis above also does not indicate any evidence of larger fish preying on smaller fish so based on my current date, I have ruled out the theory that population differences are due to survival instincts.

Bibliography

Kabacoff, Robert. *R In Action: Data Analysis and Graphics with R*. Manning, 2015.

Bluman, Allan G. *Elementary Statistics: a Step by Step Approach*. McGraw-Hill Education, 2018.

“Calculate Mean across Rows with NA Values in R.” Stack Overflow, 1 Feb. 2014, stackoverflow.com/questions/22458644/calculate-mean-across-rows-with-na-values-in-r.

“Basic Statistical Analysis Using the R Statistical Package.” 1.9 Subgroup Analyses: Finding Means and Standard Deviations for Subgroups, sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R-Manual/R-Manual7.html.

“Text.” Function | R Documentation, www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/text.

robk@statmethods.net, Robert Kabacoff -. “Graphical Parameters.” Quick-R: Graphical Parameters, www.statmethods.net/advgraphs/parameters.html.

Appendix

##### John DiSessa #####

# Q-1 Load these libraries: FSA, FSAdata, magrittr,dplyr, tidyr plyr and tidyverse

library(FSA)

library(FSAdata)

library(magrittr)

library(plyr)

library(tidyr)

library(dplyr)

library(tidyverse)

# Q-2 Import the inchBio.csv and name the table <bio>

bio <- read.csv("inchBio.csv", header = TRUE, sep = ",")

bio

# Q-3 Display the head, tail and structure of <bio>

headtail(bio, n=5)

str(bio)

summary(bio,species)

var(bio$tl)

sd(bio$tl)

var(bio$w, na.rm = T)

sd(bio$w, na.rm = T )

lmb <- subset(bio, species == "Largemouth Bass", na.rm = T)

bg <- subset(bio, species == "Bluegill", na.rm = T)

bn <- subset(bio, species == "Bluntnose Minnow",na.rm = T)

yp <- subset(bio, species == "Yellow Perch",na.rm = T)

bc <- subset(bio, species == "Black Crappie",na.rm = T)

id <- subset(bio, species == "Iowa Darter",na.rm = T)

ps <- subset(bio, species == "Pumpkinseed",na.rm = T)

tm <- subset(bio, species == "Tadpole Madtom",na.rm = T)

avgtl <- as.matrix(c(mean(bc$tl),mean(bg$tl),mean(bn$tl),mean(id$tl),

mean(lmb$tl),mean(ps$tl),mean(tm$tl),mean(yp$tl)))

avgtl

avgw <- as.matrix(c(mean(bc$w, na.rm = T),mean(bg$w, na.rm = T),

mean(bn$w, na.rm = T),mean(id$w, na.rm = T),

mean(lmb$w, na.rm = T),mean(ps$w, na.rm = T),

mean(tm$w, na.rm = T),mean(yp$w, na.rm = T)))

avgw

bio[c(1:3,221:223,324:326,356:358,584:586,597:599,603:605,641:643),]

# Q-4 Create an object, <counts>, that counts and lists all the species records

counts <- nrow(bio)

counts

list(bio$species)

# Q-5 Display just the 8 levels (names) of the species

specuniq <- unique(bio$species)

class(specuniq)

specuniq

# Q-6 Create a <tmp> object that displays the different species and the number of record of each species in the dataset

tmp <- count(bio, vars = species)

tmp

tmp3 <- mutate(tmp,avgtl,avgw)

tmp3

# Q-7 Create a subset, <tmp2>, of just the species variable and display the first five records

tmp2 <- head(subset(bio,select = species),5)

tmp2

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

w <- table(bio$species)

class(w)

w

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

t <- as.data.frame(w)

t

# Q-10 Extract and display the frequency values from the <t> data frame

t$Freq

# Q-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(bio$species)

cSpec

class(cSpec)

# Q-12 Create a table named <cSpecPct> that displays the species and percentage of records for each species. Confirm you created a table class.

cSpecPct <- (proportions(cSpec))

cSpecPct

class(cSpecPct)

# Q-13 Convert the table, <cSpecPct>, to a data frame named <u> and confirm that <u> is a data frame

u <- as.data.frame(cSpecPct)

class(u)

u

# Q-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(pin = c(4,2))

barplot(cSpec, main = "Fish Count",

ylab = "COUNTS", las = 2, col = "LightGreen",

cex.names = .6)

# Q-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”

barplot(cSpecPct, main = "Fish Relative Frequency",

ylim = c(0,.4), las = 2, ylab = "Frequency",

col.lab = "LightBlue", col = "Maroon",

cex.names = .6)

# Q-16 Rearrange the <u> cSpec Pct data frame in descending order of relative frequency. Save the rearranged data frame as the object <d>

d <- arrange(u,desc(Freq))

d

# Q-17 Rename the <d> columns Var 1 to Species, and Freq to RelFreq

colnames(d)

d <- rename(d,c("Species" = "Var1","RelFreq" = "Freq"))

d

# Q-18 Add new variables to <d> and call them cumfreq, counts, and cumcounts

cumfreq <- cumsum(d$RelFreq)

cumfreq

count <- arrange(t,desc(Freq))

count

cumcounts <- cumsum(count$Freq)

cumcounts

d <- mutate(d,cumfreq,count$Freq,cumcounts)

d <- rename(d,c("counts" = "count$Freq"))

class(d)

d

# Q-19 Create a parameter variable <def\_par> to store parameter variables

def\_par <- par()

par(mar=c(6,5,3,5))

def\_par

# Q-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”

# las: 2

par(pin = c(5,3))

pc <- barplot(d$counts, width = 1, space = .15, border = NA, axes = F,

ylim = c(0,3.05\*max(d$counts, na.rm = TRUE)),

ylab = "Cumulative Counts", cex.names = .7,

names.arg = d$Species,main = "Species Pareto", las = 2)

pc

# Q-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")

# Q-22 Place a grey box around the pareto plot (hint: https://www.statmethods.net/advgraphs/parameters.html)

box(col="grey")

# Q-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 (hint: https://www.statmethods.net/advgraphs/axes.html)

axis(side = 2,at = c(0,d$cumcounts),col.axis = "grey62",

col = "grey62", cex.axis = .8, las = 1)

# Q-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 \* 100)),"%",sep = ""),las = 1,

col.axis = "cyan4", col = "cyan4", cex.axis = .8)

# Q-25 Display the finished Species Pareto Plot (without the star watermarks). Have your last name on the plot

axis(side = 4, at = c(0,d$cumcounts),

labels = paste(c(0,round(d$cumfreq \* 100)),"%",sep = ""),las = 1,

col.axis = "cyan4", col = "cyan4", cex.axis = .8,

text(1,600, "DiSessa", cex = .7, col = "purple"))