# Executive Summary Report 3

Tianyu Zhang

CPS: Project Management, Northeastern University

ALY 6000: Introduction to Analytics

Professor: Richard He

January 29, 2023

# Introduction

This project aims to analyze the inchBio dataset using statistical techniques to understand the patterns and relationships present in the data. The inchBio dataset is a collection of data collected from various sources, including Black Crappie, Bluegill, Bluntnose Minnow, Iowa Darter, Largemouth Bass, Pumpkinseed, Tadpole Madtom, and Yellow Perch. The analysis will begin with a thorough exploration of the data, including visualizations and summary statistics. The goal of this analysis is to create a species Pareto Plot that displays the count and relative frequency of fish species in a dataset. The plot will include the count of each species, the cumulative count, and the relative frequency of each species in the dataset.

## Key findings based on instruction

```
> str(tBio)
'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: chr  "Bluegill" "Bluegill" "Bluegill" "Bluegill" ...
 $ 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    : chr  "" "" "" "" ...
 $ scale  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
> summary(tBio)
     netID            fishID         species                tl              w         
 Min.   :  1.00   Min.   :  7.0   Length:676         Min.   : 27.0   Min.   :   0.2  
 1st Qu.: 13.00   1st Qu.:175.8   Class :character   1st Qu.: 66.0   1st Qu.:   2.0  
 Median : 37.00   Median :345.5   Mode  :character   Median :189.5   Median :  54.5  
 Mean   : 67.65   Mean   :434.2                      Mean   :186.5   Mean   : 126.8  
 3rd Qu.:109.00   3rd Qu.:695.5                      3rd Qu.:295.0   3rd Qu.: 190.5  
 Max.   :206.00   Max.   :915.0                      Max.   :429.0   Max.   :1070.0  
                                                                     NA's   :165     
     tag              scale        
 Length:676         Mode :logical  
 Class :character   FALSE:213      
 Mode  :character   TRUE :463      
```

From the output of the str() and summary() functions, it appears that the tBio object is a data frame with 676 observations and 7 variables. The variables include netID, fishID, species, tl (total length), w (weight), tag, and scale. The species variable is of class character, while the scale variable is of class logical. The tl and w variables contain numerical data, with the w variable containing some missing values (NA's). The scale variable has the most number of TRUE values, indicating that most of the observations have scale data.

```
cts <- table(tBio$species)
> cts

   Black Crappie         Bluegill Bluntnose Minnow      Iowa Darter  Largemouth Bass 
              36              220              103               32              228 
     Pumpkinseed   Tadpole Madtom     Yellow Perch 
              13                6               38 
```

The output of the table() function applied to tBio$species shows the frequency count of each unique species in the species column of the tBio data frame. From the output, it appears that there are 220 observations of Bluegill, 228 observations of Largemouth Bass, 103 observations of Bluntnose Minnow, 36 observations of Black Crappie, 32 observations of Iowa Darter, 38 observations of Yellow Perch, 13 observations of Pumpkinseed and 6 observations of Tadpole Madtom.

df

```
              Var1 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
```

This is the tabulation of the frequency count of each unique species in the species column of the tBio data frame. It is presented in a tabular format with two columns: "Var1" which represents the species and "Freq" representing the frequency count of the species in the data. From the table, it is clear that the most frequent species is "Largemouth Bass" with 228 observations, followed by "Bluegill" with 220 observations, "Bluntnose Minnow" with 103 observations, "Yellow Perch" with 38 observations, "Black Crappie" with 36 observations, "Iowa Darter" with 32 observations, "Pumpkinseed" with 13 observations and the least frequent species is "Tadpole Madtom" with 6 observations.

tSpecPct

```
   Black Crappie         Bluegill Bluntnose Minnow      Iowa Darter  Largemouth Bass
      0.05325444       0.32544379       0.15236686       0.04733728       0.33727811
     Pumpkinseed   Tadpole Madtom     Yellow Perch
      0.01923077       0.00887574       0.05621302
```

This is the relative frequency of each unique species in the species column of the tBio data frame. It is presented in a tabular format with each species and its corresponding relative frequency. The relative frequency is calculated by dividing the frequency count of each species by the total number of observations in the data. The resulting value represents the proportion of the observations that belong to each species. The sum of all relative frequencies will be 1.

From the table, it is clear that the most frequent species is "Largemouth Bass" with relative frequency of 0.33727811, followed by "Bluegill" with relative frequency of 0.32544379, "Bluntnose Minnow" with relative frequency of 0.15236686, "Yellow Perch" with relative frequency of 0.05621302, "Black Crappie" with relative frequency of 0.05325444, "Iowa Darter" with relative frequency of 0.04733728, "Pumpkinseed" with relative frequency of 0.01923077 and the least frequent species is "Tadpole Madtom" with relative frequency of 0.00887574

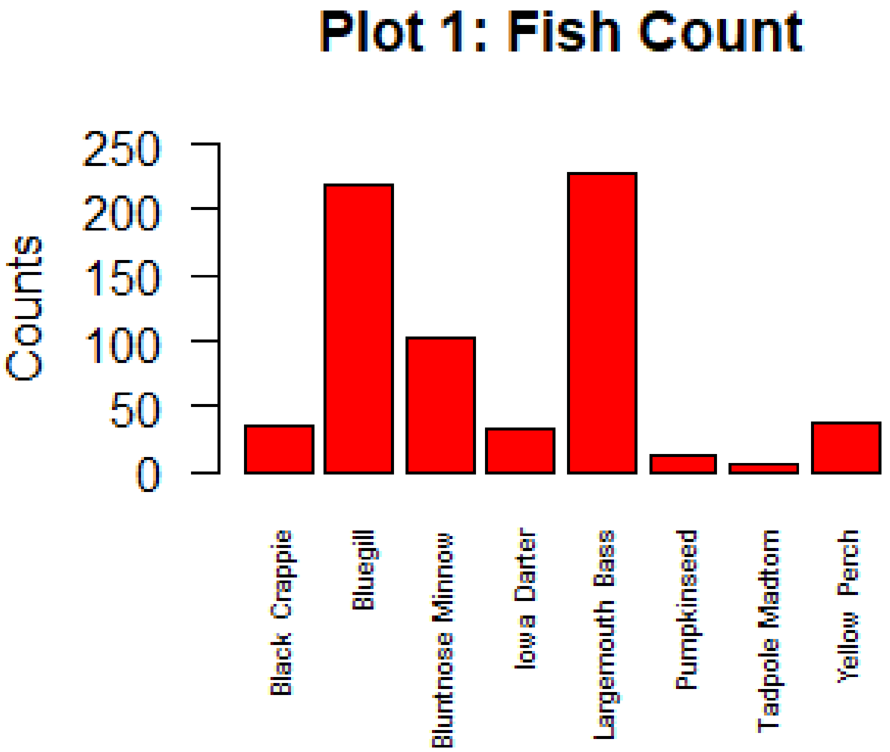
```
Species    RelFreq
1  Largemouth Bass 0.33727811
2         Bluegill 0.32544379
3 Bluntnose Minnow 0.15236686
4     Yellow Perch 0.05621302
5    Black Crappie 0.05325444
6      Iowa Darter 0.04733728
7      Pumpkinseed 0.01923077
8   Tadpole Madtom 0.00887574
```

This is the relative frequency of each unique species in the species column of the tBio data frame, but presented in a different format. It is presented in a tabular format with each species and its corresponding relative frequency, where the rows are ordered by decreasing relative frequency. The relative frequency is calculated by dividing the frequency count of each species by the total number of observations in the data. The resulting value represents the proportion of the observations that belong to each species. The sum of all relative frequencies will be 1.

From the table, it is clear that the most frequent species is "Largemouth Bass" with relative frequency of 0.33727811, followed by "Bluegill" with relative frequency of 0.32544379, "Bluntnose Minnow" with relative frequency of 0.15236686, "Yellow Perch" with relative frequency of 0.05621302, "Black Crappie" with relative frequency of 0.05325444, "Iowa Darter" with relative frequency of 0.04733728, "Pumpkinseed" with relative frequency of 0.01923077 and the least frequent species is "Tadpole Madtom" with relative frequency of 0.00887574

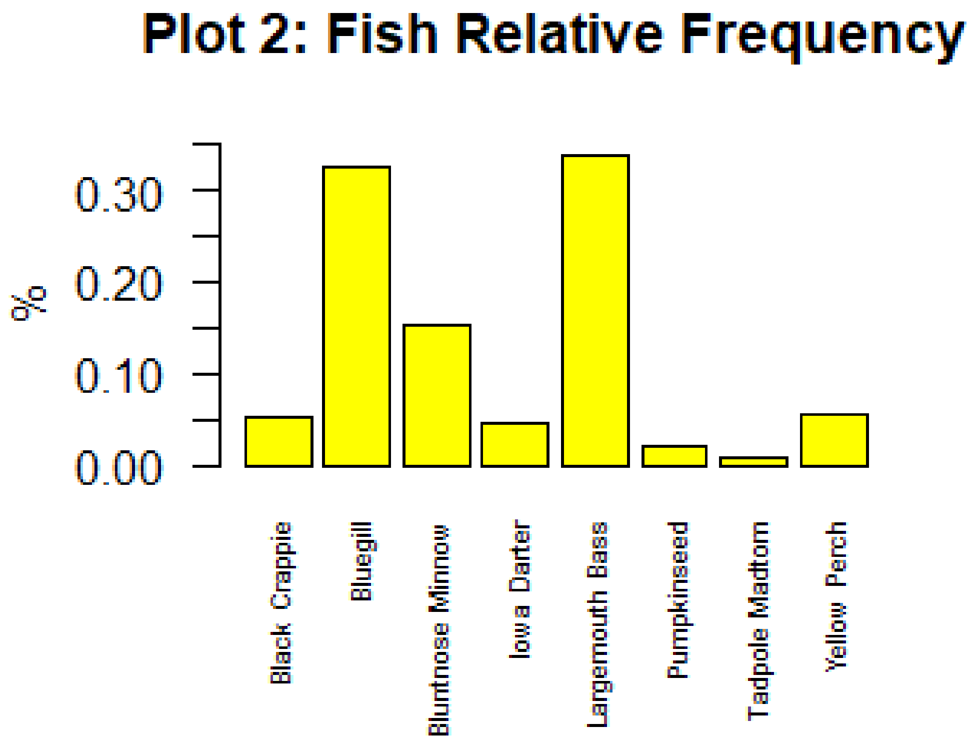
```
           Species    RelFreq   cumFreq cts cumCts
1  Largemouth Bass 0.33727811 0.3372781  36     36
2         Bluegill 0.32544379 0.6627219 220    256
3 Bluntnose Minnow 0.15236686 0.8150888 103    359
4     Yellow Perch 0.05621302 0.8713018  32    391
5    Black Crappie 0.05325444 0.9245562 228    619
6      Iowa Darter 0.04733728 0.9718935  13    632
7      Pumpkinseed 0.01923077 0.9911243   6    638
8   Tadpole Madtom 0.00887574 1.0000000  38    676
```

From this table, we can see the relative frequency and cumulative frequency of each species in the dataset, as well as the total count (cts) and cumulative count (cumCts) of each species. The "RelFreq" column shows the proportion of each species in the dataset, while the "cumFreq" column shows the proportion of each species and all the species that came before it in the table. The "cts" column shows the total number of individuals of each species in the dataset, while the "cumCts" column shows the total number of individuals of each species and all the species that came before it in the table. The last row shows that the total of all the species is 676.



This code creates a barplot using the data in the tSpec table, with the species on the x-axis and the counts on the y-axis. The title of the plot is "Plot 1: Fish Count", the y-axis is labeled "Counts" and the y-axis limits are set to 0 to 250. The bars are colored red, the x-axis labels are rotated to be vertical and the x-axis font magnification is set to 55% of the nominal size (Ryan Moore, 2019).

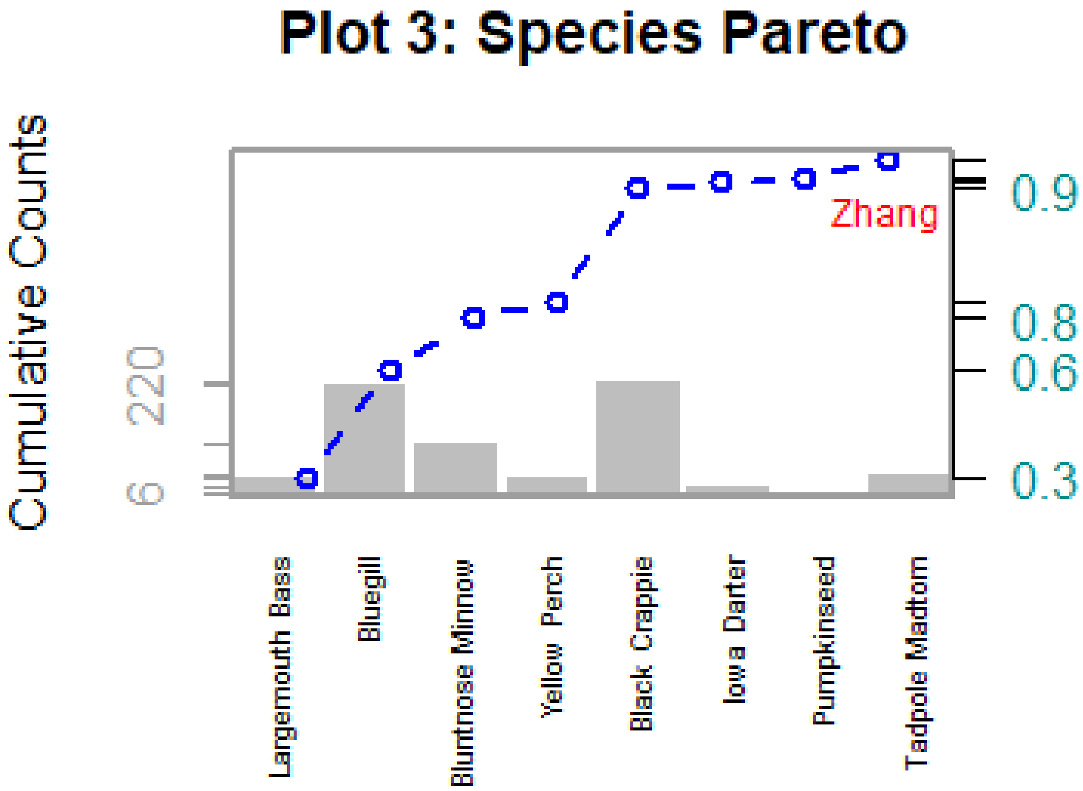
From this plot, one can learn about the relative abundance of different fish species in the dataset, and it can also be used to compare the relative abundance of different species. It can also be used to identify patterns, such as which species are the most common or least common, or which species are over or under-represented in the sample.



The barplot of tSpecPct shows the relative frequency of each fish species in the sample. The y-axis is labeled "%" and has a limit of 0 to 0.35, which allows for easy comparison of the relative frequency of each species. The bars are colored yellow, making it easy to distinguish each species. The title of the plot is "Plot 2: Fish Relative Frequency".

From the plot, we can see that Largemouth Bass is the most common species in the sample, with a relative frequency of about 0.34 or 34%. Bluegill is the second most common species with a relative frequency of about 0.32 or 32%. The remaining species have lower relative frequencies, with Bluntnose Minnow and Yellow Perch having the next highest relative frequencies of about 0.15 and 0.06 respectively (Kabacoff , 2017).

Overall, the plot provides a clear visual representation of the relative frequency of each fish species in the sample, which can be useful in understanding the distribution and abundance of different species in the sample.



This command is creating a barplot with the data from the variable "data$cts" and the X-axis labels are coming from the "data$Species" variable. The Y axis limit is set to 0 and 3.05 times the maximum value of the "data$cts" variable with the Y label as "Cumulative Counts". The X axis font magnification is set to 55% of the nominal value. The plot is titled as "Plot 3: Species Pareto" and the labels on the X-axis are rotated by 2(rebcarebca, 1959).

In addition, it is adding a cumulative counts line to the same plot with the line type as "b", the plotting text scaled at 75%, and the data values are marked with a solid triangle symbol and in blue color (Youtube, 2020).

# Bibliography

Kabacoff, R. (2017). *Graphical parameters*. Quick-R: Graphical Parameters. Retrieved January 30, 2023, from https://www.statmethods.net/advgraphs/parameters.html

rebcarebca. (1959, September 1). *How to change the font size and color of X-axis and y-axis label in a scatterplot with plot function in R?* Stack Overflow. Retrieved January 30, 2023, from https://stackoverflow.com/questions/12504549/how-to-change-the-font-size-and-color-of-x-axis-and-y-axis-label-in-a-scatterplo

Ryan Moore. (2019, April 25). *Rotating axis labels in R plots*. Tender Is The Byte. Retrieved January 30, 2023, from https://www.tenderisthebyte.com/blog/2019/04/25/rotating-axis-labels-in-r/

YouTube. (2020, June 19). *Change font size of GGPLOT2 facet grid labels in R (example) | increase or decrease text sizes*. YouTube. Retrieved January 30, 2023, from https://www.youtube.com/watch?v=aMqcGM6VcjU

## Appendix

```
#Print your name at the top of the script and load these libraries: FSA, FSAdata, magrittr,
#dplyr, tidyr plyr and tidyverse
print("Tianyu Zhang")
install.packages("FSA")
install.packages("FSAdata")
install.packages("magrittr")
install.packages("dplyr")
install.packages("tidyr")
install.packages("plyr")
install.packages("tidyverse")
install.packages("ggplot2")
library("FSA")
library("FSAdata")
library("magrittr")
library("dplyr")
library("tidyr")
library("plyr")
library("tidyverse")
library("ggplot2")
#. Import the inchBio.csv and name the table <tBio>
tBio <- read.csv("C:/Users/张天羽/Desktop/inchBio.csv")

#Display the first 3 and last 3 records, summary and structure of <tBio>
head(tBio,3)
tail(tBio,3)

str(tBio)
summary(tBio)
```

```
#Create an object <cts>, that counts and lists all the species records
cts <- table(tBio$species)
cts

#Display just the 8 levels (names) of the species



head(unique(tBio$species),8)

#Create a <temp1> object that displays the different species and the number of records
#of each species in the dataset. Include this information in your report.

temp1 <- table(tBio$species)
temp1

#Create a subset, <temp2>, of just the species variable and display the first 3 records

temp2 <- subset(tBio, select = c("species"))
head(temp2, 3)



#Create a table, <t>, of the species variable. Display the class of <t>
t <- table(tBio$species)
class(t)

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

#Extract and display the frequency values from the <df> data frame as variable freq
```

```
freq <- df$Freq
print(freq)

#Create a table named <tSpec> from the <tBio> species attribute (variable) and confirm
#that you created a table which displays the number of species in the dataset <tBio>
tSpec <- table(tBio$species)
class(tSpec)
print(tSpec)

#. Create a table named <tSpecPct> that displays the species and percentage of records
#for each species. Confirm you created a table class.
tSpecPct <- prop.table(tSpec)
tSpecPct
class(tSpecPct)



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

#Create a barplot of <tSpec>
barplot(tSpec, col = "red", ylab = "Counts",
        main = "Plot 1: Fish Count", ylim = c(0, 250),las = 2, cex.names = 0.55 )

#Create a barplot of <tSpecPct>


barplot(tSpecPct, col = "yellow", ylab = "%",
        main = "Plot 2: Fish Relative Frequency",ylim = c(0,0.35), las = 2, cex.names = 0.55)
```

```
#Rearrange the <dfSP> data frame in descending order of relative frequency. Save the
#rearranged data frame as the object <data>
data <- dfSP %>% arrange(desc(dfSP$Freq))
data

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

#Add new variables to <data> and call them cumFreq, cts, and cumCts
data <- data %>%
  mutate(cumFreq = cumsum(RelFreq), cts = cts , cumCts = cumsum(cts))
data

#Create a parameter variable <varPar> to store graphical parameters
varPar <- list()
varPar
#Create a barplot for data$cts, <pc>,
barplot(data$cts, width = 1, space = .1, border = NA, axes = F,
        ylim = c(0, 3.05*max(data$cts, na.rm=T)), ylab = "Cumulative Counts",
        xaxs = "i", cex.names = 0.55, names.arg = data$Species,
        main = "Plot 3: Species Pareto", las = 2)



#Add a cumulative counts line to the <pc> plot
lines(data$cumCts, type = "b", col = "blue", lty = 2, lwd = 2)

#Place a grey(grey62) box around the pareto plot
box(col="grey62", lty=1, lwd=2)
```

```
#Add a left side axis with the following specifications

axis(side = 2, at = data$cts, col = "grey62", col.axis = "grey62", cex.names = 0.75)


#Add axis details on right side of box

axis(side = 4, at = data$cumCts, labels = data$cumFreq, las = 2, col.axis = "cyan4", col.lab = "cyan4", cex.name = 0.75, tick = TRUE)


#Display the finished Species Pareto Plot

text(x = nrow(data), y = max(data$cts)*2.5, col = "red", labels = "Zhang", cex = 0.75)
```