Northeastern University

Toronto, Canada

A Report on

Executive Summary of Module 1

Introduction to Data Analytics (ALY 6000)

Guided by:

Prof. Mohammad Shafiqul Islam

Submitted By:

Name of Student	NUID	Date of submission
Parth Shah	002956963	17th January, 2021

Index

1. Introduction

2. Following an introduction, provide an analysis of descriptive characteristics of the data set provided by your instructor. This includes pertinent statistics including counts, cumulative counts, and frequency, percentages, etc. Include R console screen snippets to support your observations and conclusions. Below is a sample excerpt.

3. Provide the executive with visualizations (at least 3) in that help them see the key characteristics you want to highlight. They can be boxplots, histograms, frequency and probability distributions, or bar plots (bar charts). A pareto plot as illustrated below must be included in this part of your report. Include screen snippets of your plots to support your findings and conclusions. The goal is not only to present your visual results, but also to explain the significance of them.

4. Summary
5. Reference
6. Appendix

1. Introduction

This report depicts the table of inchBio which contains 8 different species of fish and their specifications of the fish.

2. Following an introduction, provide an analysis of descriptive characteristics of the data set provided by your instructor. This includes pertinent statistics including counts, cumulative counts, and frequency, percentages, etc. Include R console screen snippets to support your observations and conclusions. Below is a sample excerpt.

I had created the counts object with reference the inchBio data.

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

Through unique function, I get know that there are 8 species

```
unique(Bio$species)
## [1] "Bluegill" "Bluntnose Minnow" "Iowa Darter" "Largemouth Bass"
## [5] "Pumpkinseed" "Tadpole Madtom" "Yellow Perch" "Black Crappie"
```

I had created a table with table name w which contains all the species names and check the class by class() function

As per question, I had converted w to t. Result is as follow: -

```
t <- as.data.frame(w)
```

```
t
##             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
```

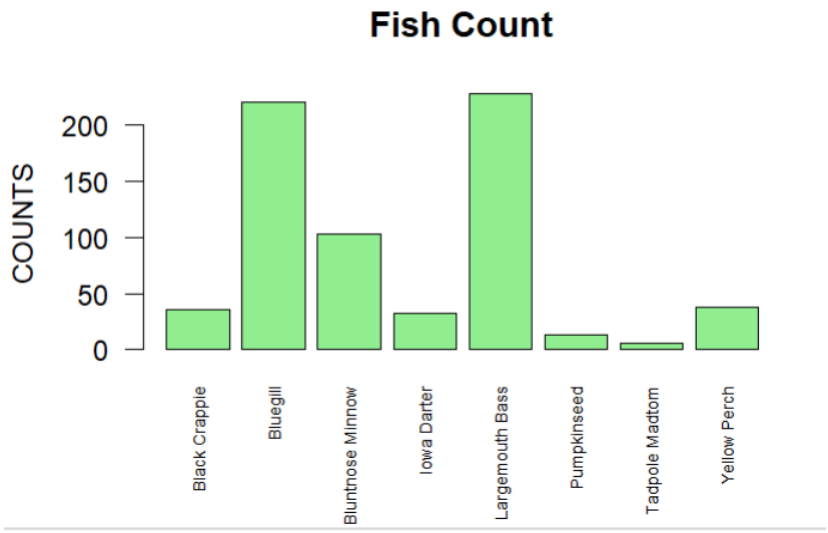
Displaying frequency through t$Freq. Result is as follow: -

```
## [1] 36 220 103 32 228 13 6 38
```

3. **Provide the executive with visualizations (at least 3) in that help them see the key characteristics you want to highlight. They can be boxplots, histograms, frequency and probability distributions, or bar plots (bar charts). A pareto plot as illustrated below must be included in this part of your report. Include screen snippets of your plots to support your findings and conclusions. The goal is not only to present your visual results, but also to explain the significance of them.**

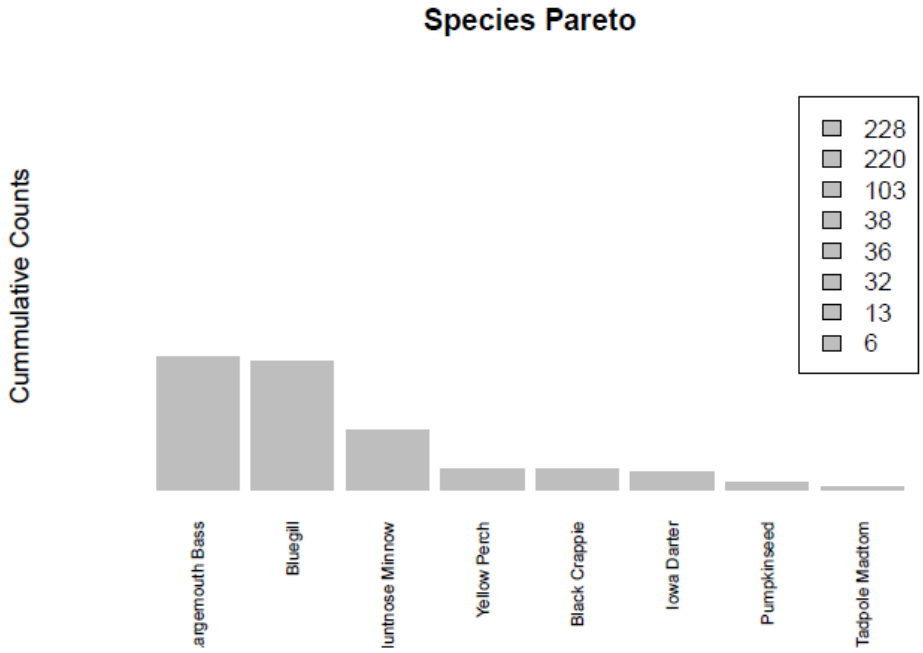
PLOT 1: - Fish Count

The bar plot illustrates that which species of fishes are highest.



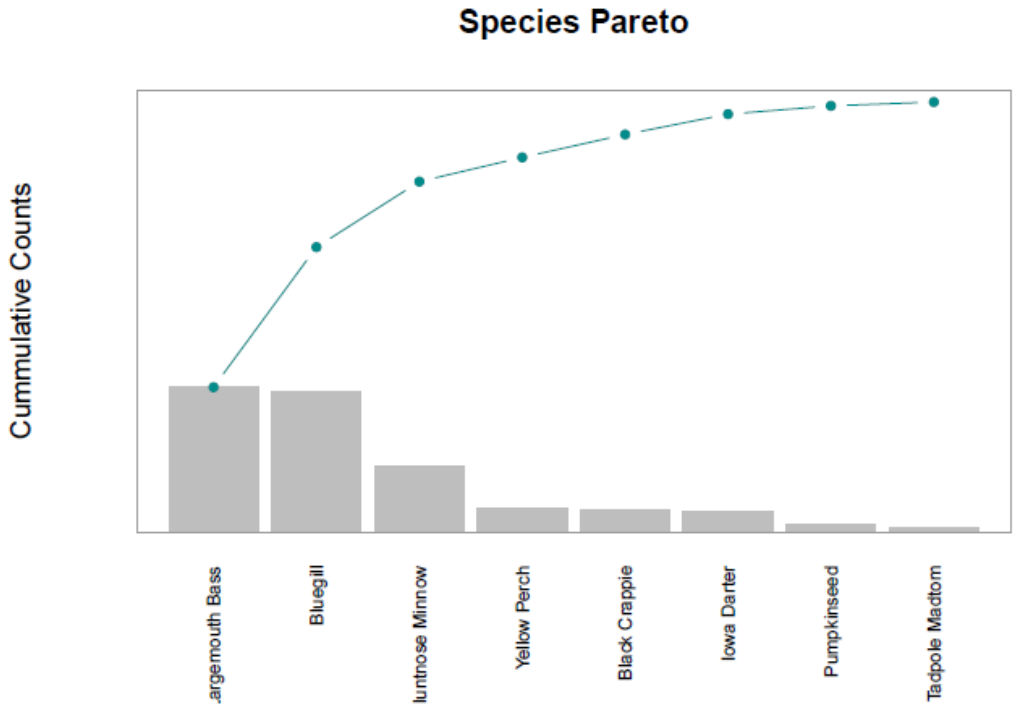
PLOT 2: - <u>Species Pareto</u>

The bar plot demonstrate which species is greater in cumulative counts.



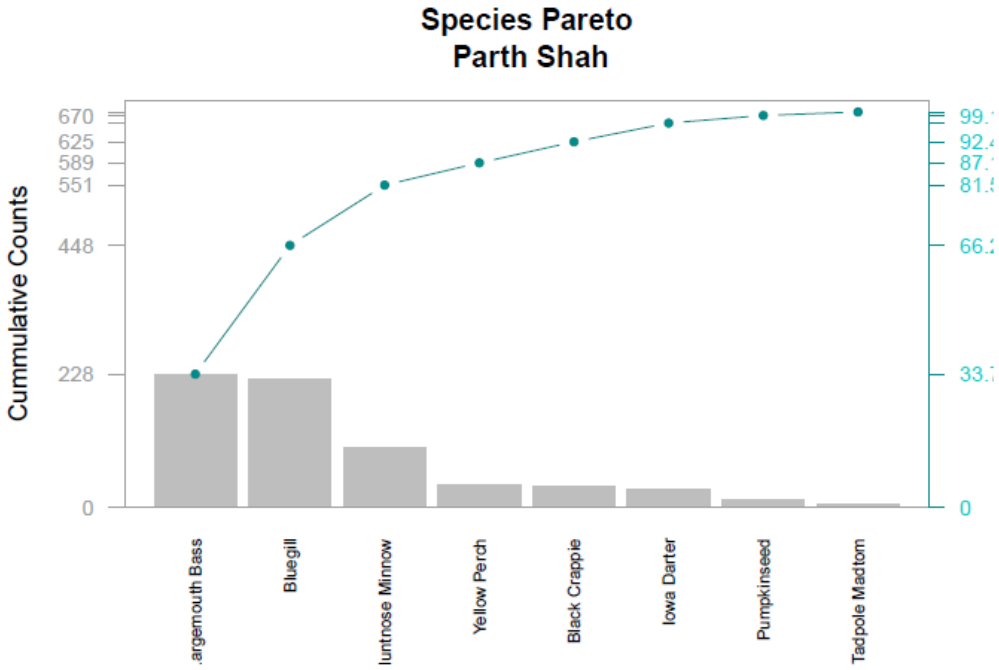
PLOT 3: - <u>Species Pareto with some extra Function</u>

This bar plot demonstrates how many various functions are used to create the plot's box, lines, and values, among other things.



PLOT 4: - <u>Species Pareto with Name and Additional Function</u>

The box, lines, axis, and values, among other things, are all created using a variety of methods in this bar plot.



4. Summary

To summarize, I learned how to create a bar plot, set a box around it, define their structures and summaries, and display data using plots with a variety of functions and axes.

5. Reference

- Rui Barradas(2020). How to plot a line in R from Stack Overflow https://stackoverflow.com/questions/62793950/how-can-i-plot-a-line-on-bar-chart-in-r

- Harry. How to add column to data.frame in R from Analytics vidhya https://discuss.analyticsvidhya.com/t/how-to-add-a-column-to-a-data-frame-in-r/3278

- Data Mentor https://discuss.analyticsvidhya.com/t/how-to-add-a-column-to-a-data-frame-in-r/3278

- STHDA http://www.sthda.com/english/wiki/add-an-axis-to-a-plot-with-r-software

- GitHub Link: https://github.com/iparth0611/Module3.git

6. Appendix

Shah_M3_Project3.R

prbsh

2022-02-04

```
#Print name
print("Parth Shah")
```

```
## [1] "Parth Shah"
```

```
#Loading packages through pacman
pacman::p_load(FSA, FSAdata, magrittr, dplyr, tidyr, plyr, tidyverse)

#Import inchBio dataset
library(readr)
Bio <- read_csv("~/R/ALY6000/Module3/inchBio.csv")
```

```
## Rows: 676 Columns: 7
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): species, tag
## dbl (4): netID, fishID, tl, w
## lgl (1): scale
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(Bio)

#Head, Tail, Structure
headtail(Bio)
```

```
##     netID fishID       species  tl     w  tag scale
## 1      12     16      Bluegill  61   2.9 <NA> FALSE
## 2      12     23      Bluegill  66   4.5 <NA> FALSE
## 3      12     30      Bluegill  70   5.2 <NA> FALSE
## 674   110    863 Black Crappie 307 415.0 1783  TRUE
## 675   129    870 Black Crappie 279 344.0 1789  TRUE
## 676   129    879 Black Crappie 302 397.0 1792  TRUE
```

```
structure(Bio)
```

```
## # A tibble: 676 x 7
##    netID fishID species     tl     w tag   scale
##    <dbl>  <dbl> <chr>    <dbl> <dbl> <chr> <lgl>
##  1    12     16 Bluegill    61   2.9 <NA>  FALSE
##  2    12     23 Bluegill    66   4.5 <NA>  FALSE
##  3    12     30 Bluegill    70   5.2 <NA>  FALSE
##  4    12     44 Bluegill    38   0.5 <NA>  FALSE
##  5    12     50 Bluegill    42   1   <NA>  FALSE
##  6    12     65 Bluegill    54   2.1 <NA>  FALSE
##  7    12     66 Bluegill    27  NA   <NA>  FALSE
##  8    13     68 Bluegill    36   0.5 <NA>  FALSE
##  9    13     69 Bluegill    59   2   <NA>  FALSE
## 10    13     70 Bluegill    39   0.5 <NA>  FALSE
## # ... with 666 more rows
```

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

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

```
#Display just the 8 levels (names) of the species
unique(Bio$species)
```

```
## [1] "Bluegill"         "Bluntnose Minnow" "Iowa Darter"      "Largemouth Bass"
## [5] "Pumpkinseed"      "Tadpole Madtom"   "Yellow Perch"     "Black Crappie"
```

```
#Create a <tmp> object that displays the different species and the number of record of each species in
tmp <- table(Bio$species)
tmp
```

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

```
#Create a subset, <tmp2>, of just the species variable and display the first five records
tmp2 <- subset(Bio, select = species)
head(tmp2, 5)
```

```
## # A tibble: 5 x 1
##   species
##   <chr>
```

```
## 1 Bluegill
## 2 Bluegill
## 3 Bluegill
## 4 Bluegill
## 5 Bluegill
```

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

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

```
class(w)
```

```
## [1] "table"
```

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

```
##               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
```

```
#Extract and display the frequency values from the <t> data frame
t$Freq
```

```
## [1]  36 220 103  32 228  13   6  38
```

```
#Create a table named <cSpec> from the bio species attribute (variable) and confirm that you created a
cSpec <- table(Bio$species)
cSpec
```

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

```
#Create a table named <cSpecPct> that displays the species and percentage of records for each species.
CSpecPct <- prop.table(cSpec)*100
CSpecPct
```

```
##
##    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"
```

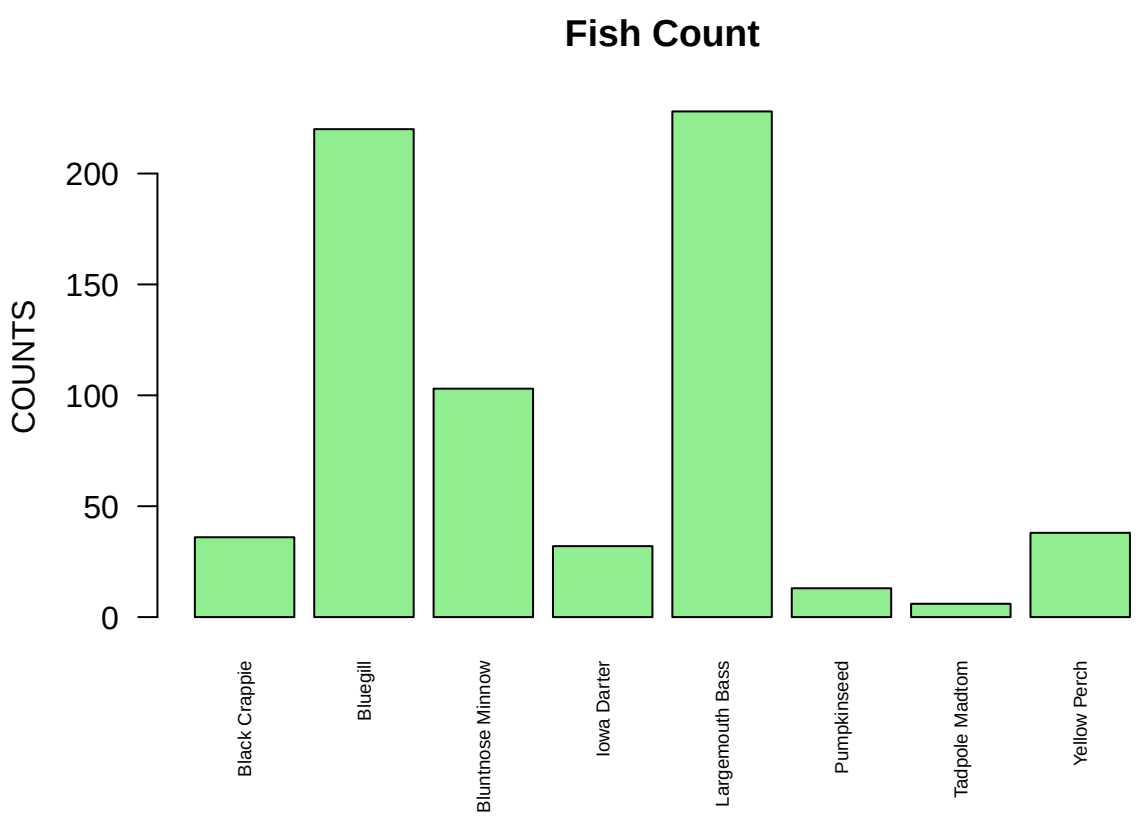
```
#Convert the table, <cSpecPct>, to a data frame named <u> and confirm that <u> is a data frame
u <- as.data.frame(CSpecPct)
u
```

```
##               Var1      Freq
## 1    Black Crappie  5.325444
## 2         Bluegill 32.544379
## 3 Bluntnose Minnow 15.236686
## 4      Iowa Darter  4.733728
## 5  Largemouth Bass 33.727811
## 6      Pumpkinseed  1.923077
## 7   Tadpole Madtom  0.887574
## 8     Yellow Perch  5.621302
```

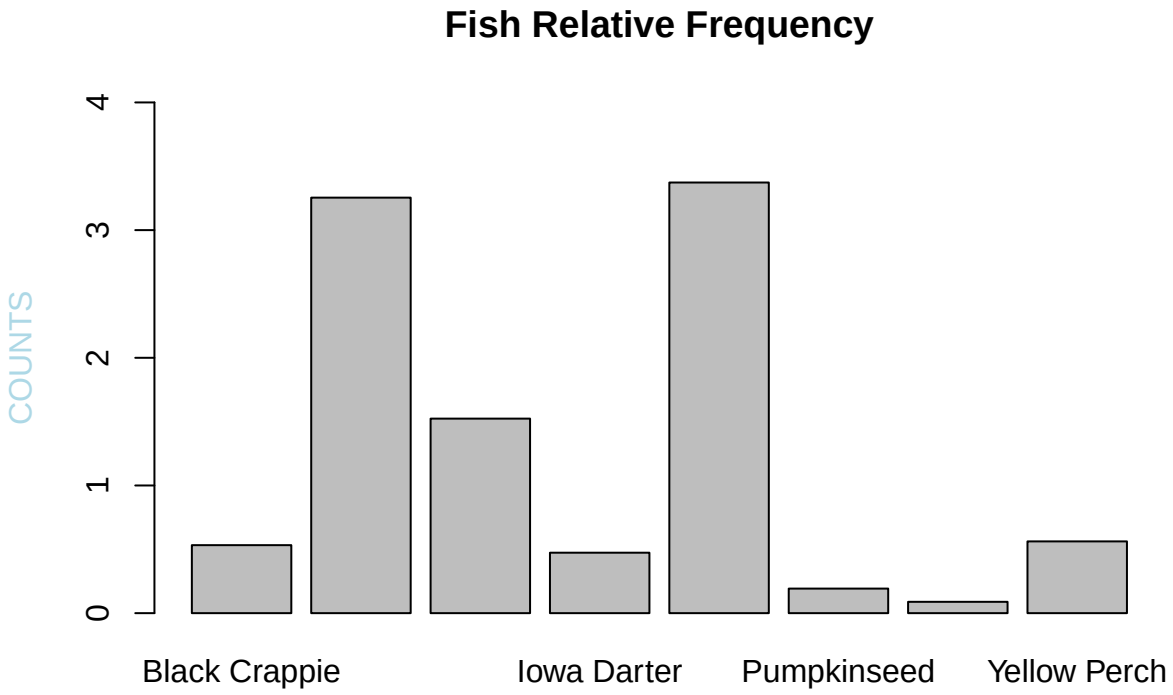
```
class(u)
```

```
## [1] "data.frame"
```

```
#Barplot of <cSpec> with the following: titled Fish Count
barplot(cSpec,
     main = "Fish Count",
     ylab = "COUNTS",
     col = "lightgreen",
     cex.names = 0.60,
     las = 2
     )
```




```
#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/10,
        ylim = c(0,4),
        main = "Fish Relative Frequency",
        ylab = "COUNTS",
        col.lab = "lightblue")
```




```
#Rearrange the <u> cSpec Pct data frame in descending order of relative frequency. Save the rearranged
d <- u[order(-u$Freq),]
d
```

```
##               Var1      Freq
## 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
```

```
#Rename the <d> columns Var 1 to Species, and Freq to RelFreq
d
```

```
##               Var1      Freq
## 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
```

```
names(d)[names(d)=="Var1"] <- "Species"
d
```

```
##            Species      Freq
## 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
```

```
names(d)[names(d)=="Freq"] <- "RelFreq"
print(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
```

```
#Add new variables to <d> and call them cumfreq, counts, and cumcounts
counts
```

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

```
t$Freq
```

```
## [1]  36 220 103  32 228  13   6  38
```

```
tdesc <- t[order(-t$Freq),]
tdesc$Freq
```

```
## [1] 228 220 103  38  36  32  13   6
```

```
d <- d %>% mutate(cumfreq=cumsum(d$RelFreq), counts=tdesc$Freq, cumcounts=cumsum(tdesc$Freq))
d
```

```
##            Species   RelFreq   cumfreq counts cumcounts
## 5  Largemouth Bass 33.727811  33.72781    228       228
## 2         Bluegill 32.544379  66.27219    220       448
```

```
## 3 Bluntnose Minnow 15.236686  81.50888    103       551
## 8     Yellow Perch  5.621302  87.13018     38       589
## 1    Black Crappie  5.325444  92.45562     36       625
## 4      Iowa Darter  4.733728  97.18935     32       657
## 6      Pumpkinseed  1.923077  99.11243     13       670
## 7   Tadpole Madtom  0.887574 100.00000      6       676
```

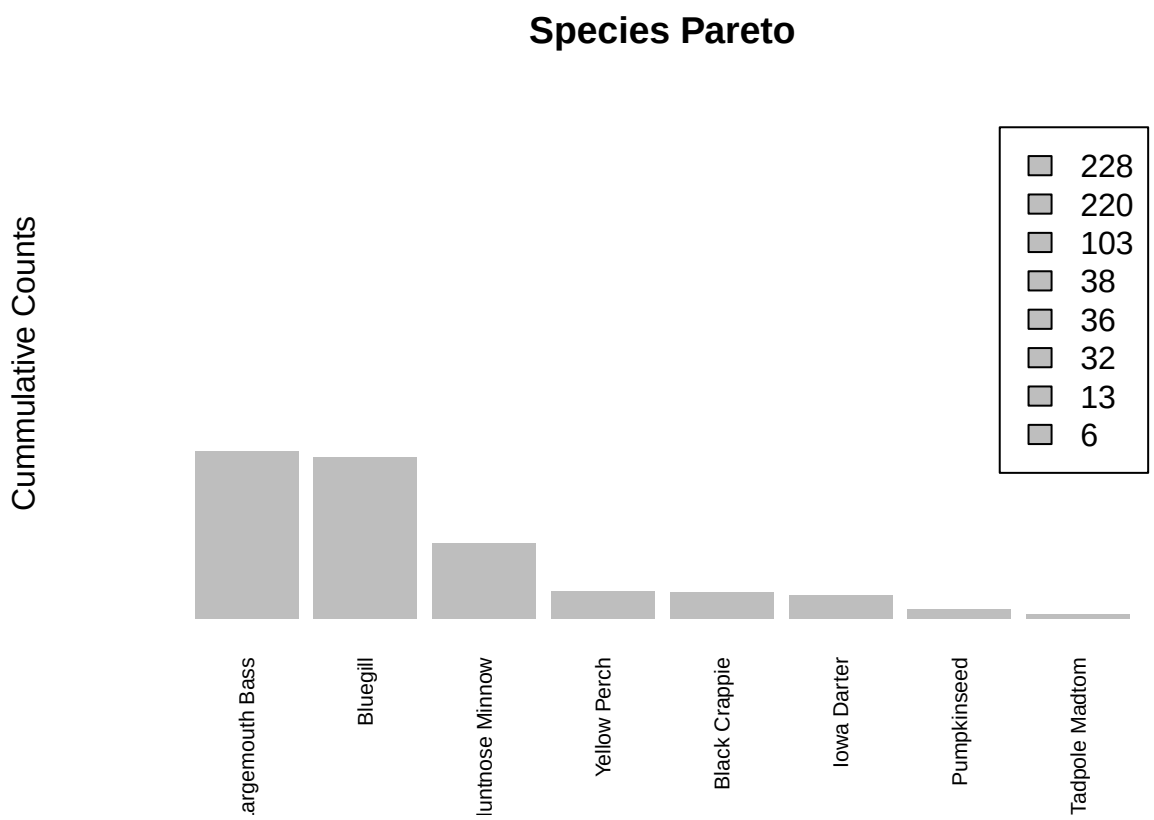
```
#Create a parameter variable <def_par> to store parameter variables
def_par <- par(no.readonly = TRUE)

# barplot <pc>
pc <- barplot(d$counts, width = 1, space = 0.15,border = NA, axes = F, ylim = c(0,3.05*228), ylab = "Cu
```

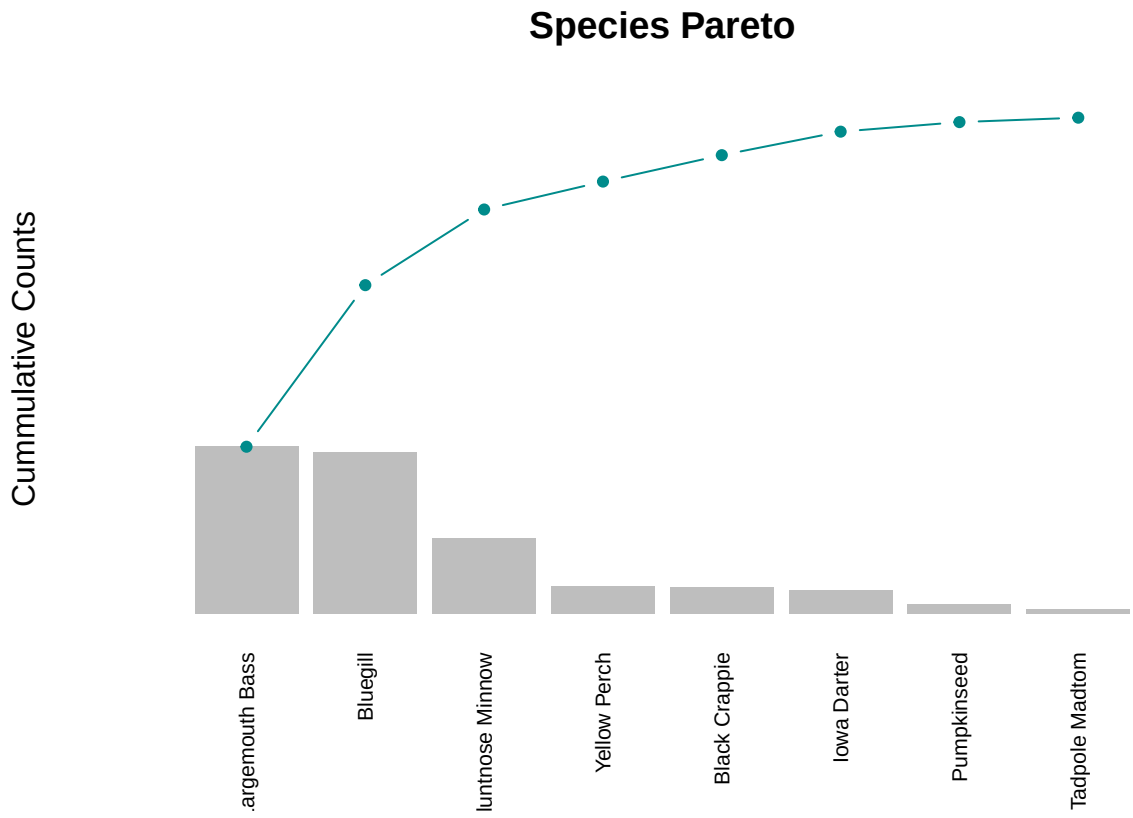
```
## Warning in plot.window(xlim, ylim, log = log, ...): "na.rm" is not a graphical
## parameter
```

```
## Warning in axis(if (horiz) 2 else 1, at = at.l, labels = names.arg, lty =
## axis.lty, : "na.rm" is not a graphical parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): "na.rm"
## is not a graphical parameter
```

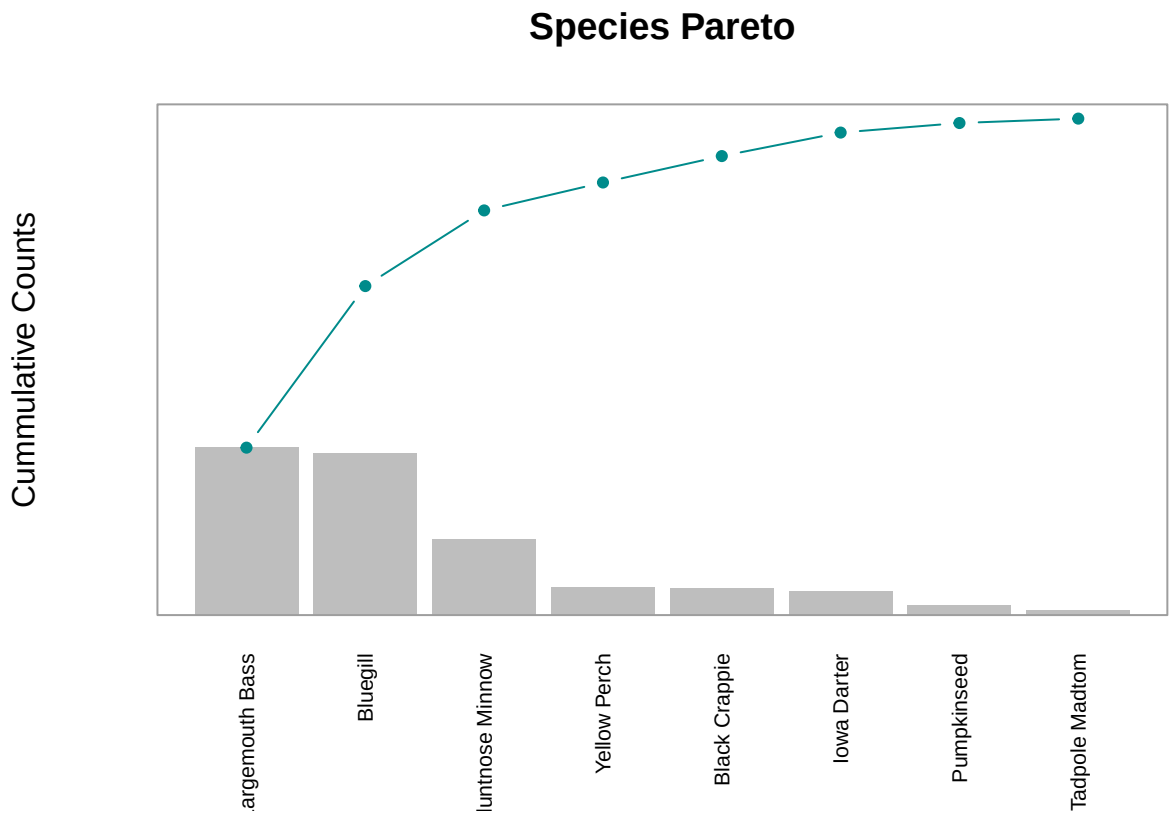


```
#Add a cumulative counts line to the <pc> plot with the following:
#Spec line type is b
#Scale plotting text at 70%
#Data values are solid circles with color cyan4
pc <- barplot(d$counts,
              width = 1,
              space = 0.15,
              border = NA,
              axes = F,
              main = "Species Pareto",
              ylim = c(0,3.05*228),
              ylab = "Cummulative Counts",
              names.arg = d$Species,
              las=2,
              cex.names = 0.70
              )
lines(pc, d$cumcounts, type = "b", cex = 0.7, pch = 19, col="cyan4")
```



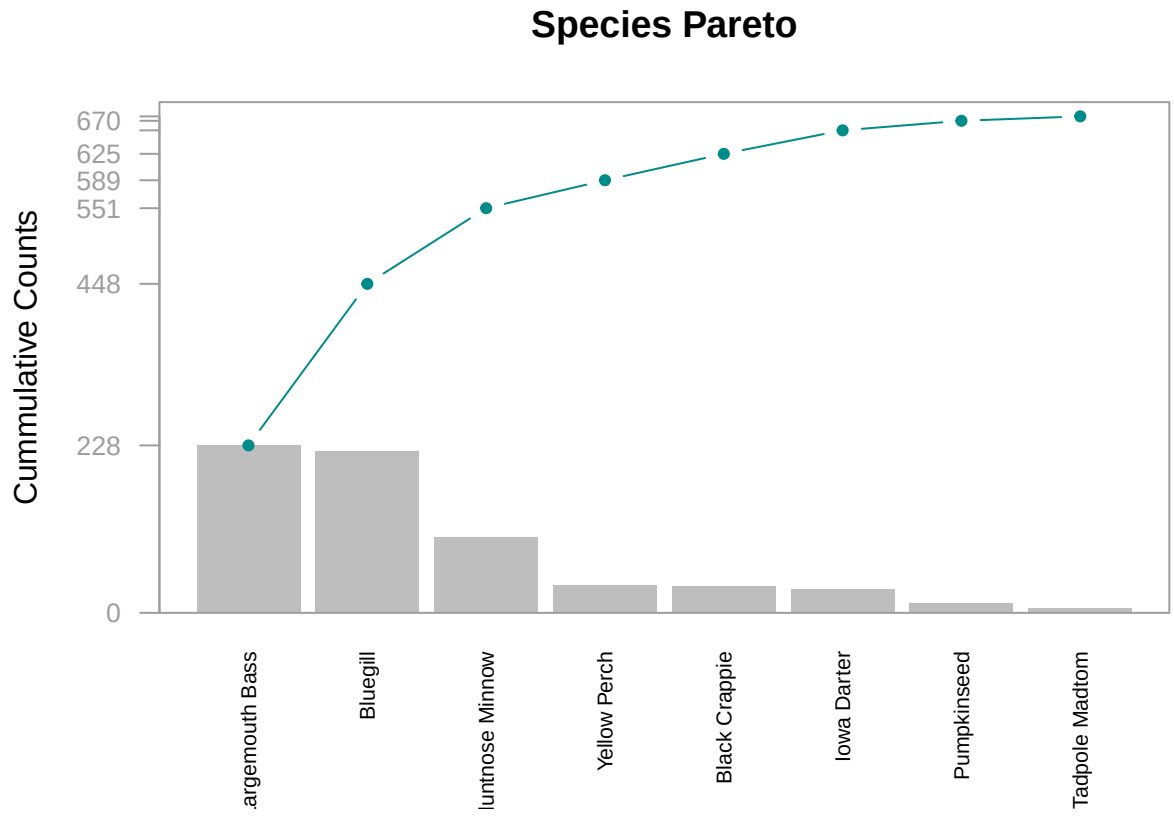

```
#Place a grey box around the pareto plot
pc <- barplot(d$counts,
              width = 1,
              space = 0.15,
              border = NA,
              axes = F,
              main = "Species Pareto",
```

```
              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")
```



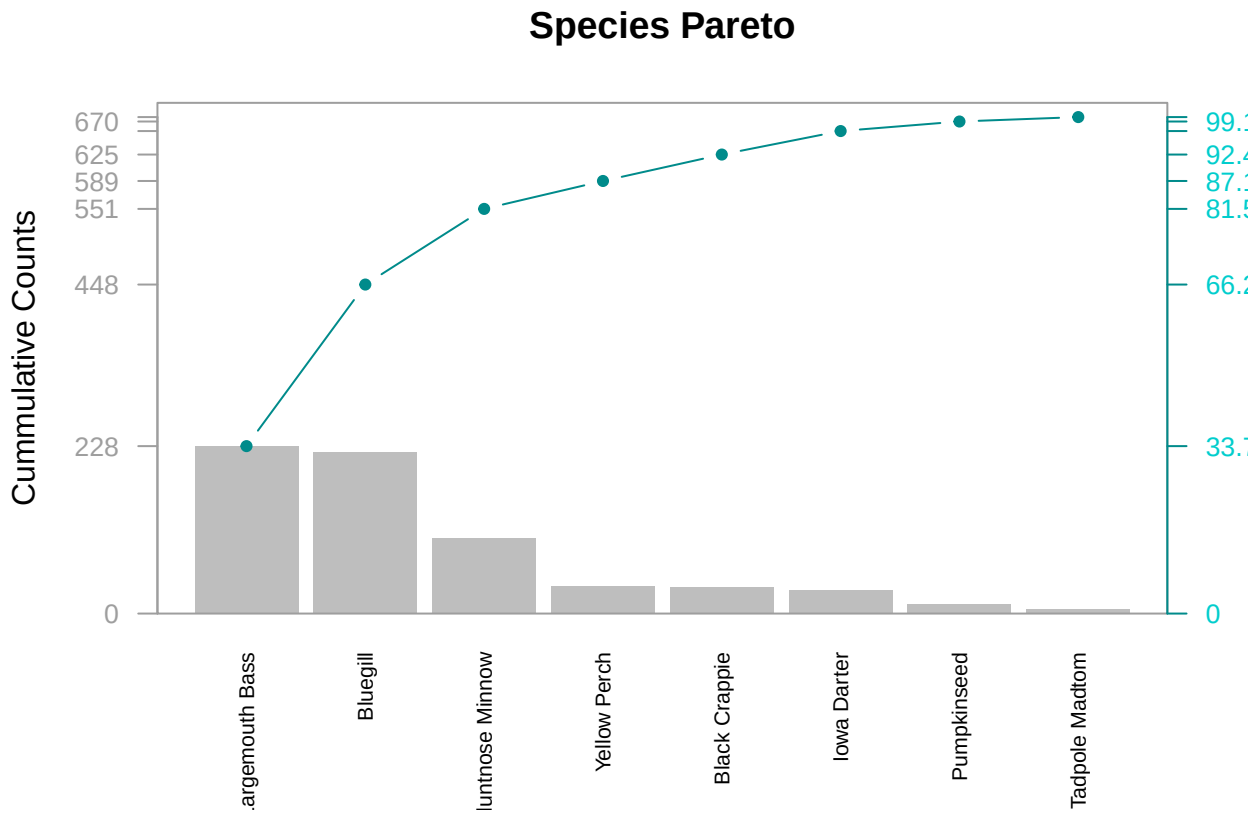

```
#Add a left side axis with the following specifications
#Horizontal values at tick marks at cumcounts on side 2
#Tickmark color of grey62
#Color of axis is grey62
#Axis scaled to 80% of normal
pc <- barplot(d$counts,
              width = 1,
              space = 0.15,
              border = NA,
              axes = F,
              main = "Species Pareto",
              ylim = c(0,3.05*228),
              ylab = "Cummulative Counts",
              names.arg = d$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.axis = 0.8)
```




```
#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
pc <- barplot(d$counts,
              width = 1,
              space = 0.15,
              border = NA,
              axes = F,
              main = "Species Pareto",
              ylim = c(0,3.05*228),
              ylab = "Cummulative Counts",
              names.arg = d$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.axis = 0.8)
```

```
axis(side = 4, at = c(0, d$cumcounts),
     labels = c(0, d$cumfreq),
     las = 1,
     col.axis = "cyan3", #Error throughs that invalid cyan5 color
     col = "cyan4",
     cex.axis = 0.8)
```




```
#Display the finished Species Pareto Plot (without the star watermarks). Have your last name on the plo
pc <- barplot(d$counts,
              width = 1,
              space = 0.15,
              border = NA,
              axes = F,
              main = "Species Pareto\n Parth Shah",
              ylim = c(0,3.05*228),
              ylab = "Cummulative Counts",
              names.arg = d$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.axis = 0.8)
axis(side = 4, at = c(0, d$cumcounts),
     labels = c(0, d$cumfreq),
```

```
    las = 1,
    col.axis = "cyan3", #Error throughs that invalid cyan5 color
    col = "cyan4",
    cex.axis = 0.8)
```