**M3 - PROJECT 3**

**Name: Sunil Raj Thota**

**NUID: 001099670**

**Date: 10/07/2020**

**Title: M3 Project 3 Executive Summary Report**

**ALY 6000 - 71618 (CRN Number)**

**Prof. Dee Chiluiza Reyes**

**Introduction:**

* R is a programming language is used for statistical analysis and creating graphs.
* Data objects are used for calculations.
* The scope of R is vast and mainly used in the fields of data mining, Regression, Probability, ML etc., by using various packages and libraries.

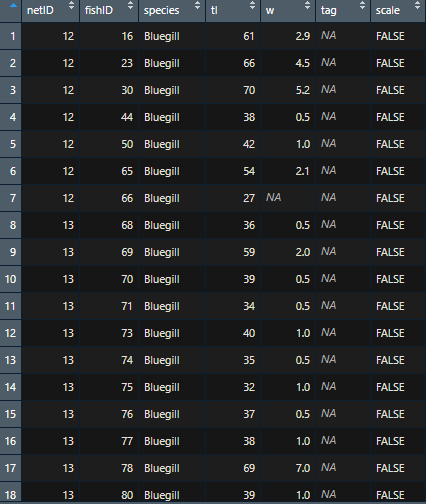
**Fisheries:**

* A structural analysis of the inchBio data set revealed that two species dominated this study. The Bluegill and Largemouth Bass accounted for 66% of the research data.
* As a result, Bluegill and Largemouth Bass subsets were created in order to facilitate an in-depth study of the fish species that dominate the dataset.

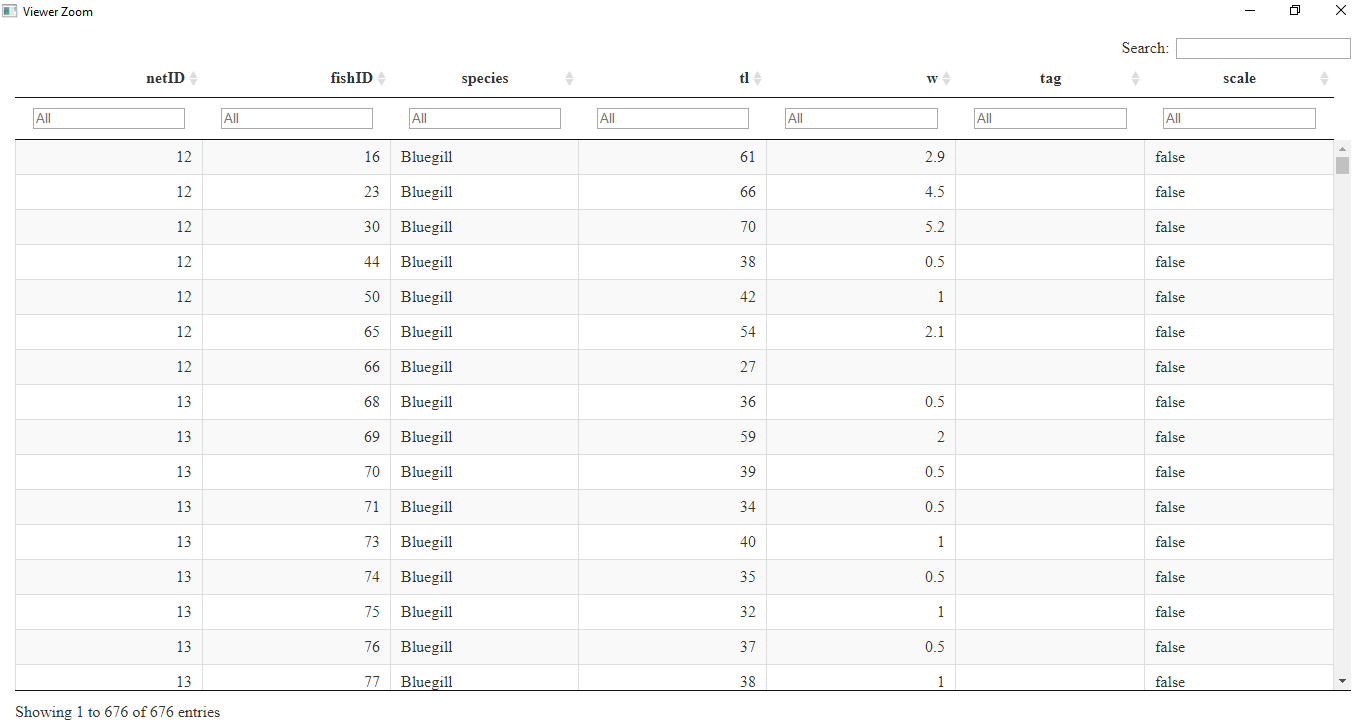
**Data Analysis:**

* The inchBio data set showcases 8 different species called Black Crappie, Bluegill Bluntnose Minnow, Iowa Darter, Largemouth Bass, Pumpkinseed, Tadpole Madtom, and Yellow Perch
* The data was categorized into 8 main frequencies as 36, 220, 103, 32, 228, 13, 6, 38 for Black Crappie, Bluegill Bluntnose Minnow, Iowa Darter, Largemouth Bass, Pumpkinseed, Tadpole Madtom, and Yellow Perch respectively
* The Total length of the fishes is from 27 to 429 in mm
* The overall length of Tag column is 676
* There are 213 False Scale fish species and 463 True Scale fish species

**Full InchBio data set:**

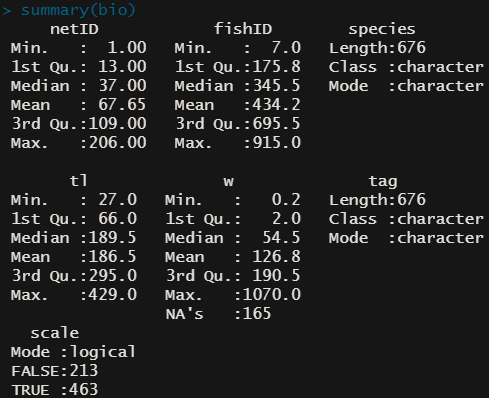


**Interactive DT Table:**



**Descriptive Statistics:**

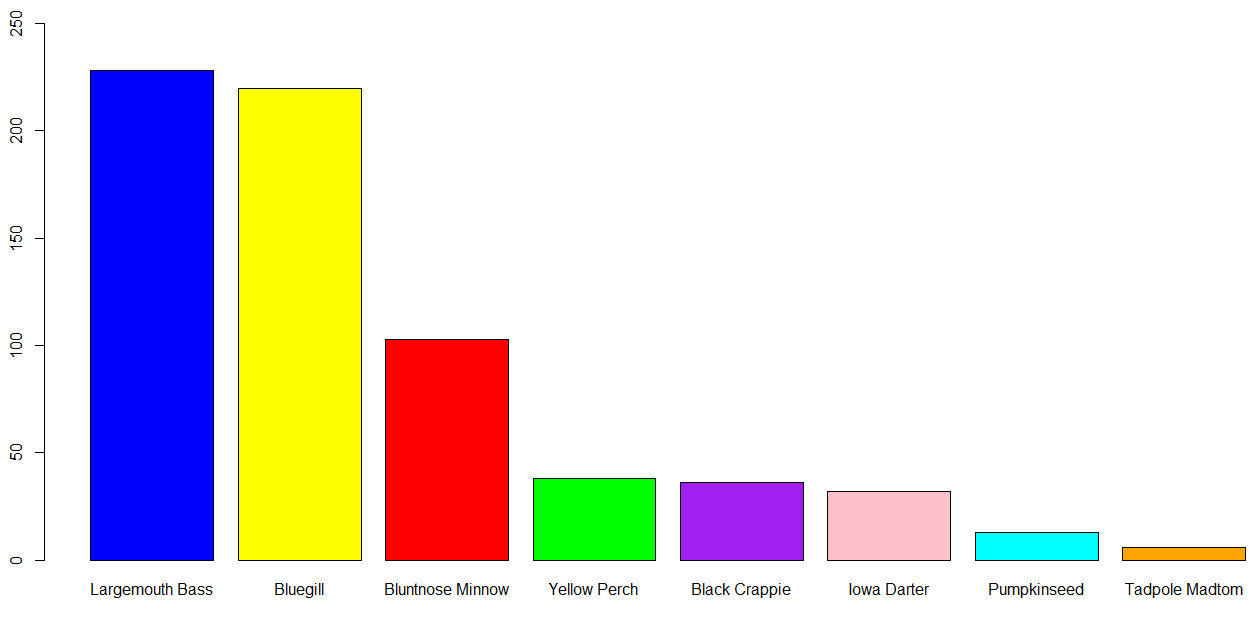
* It is an is important part in any statistical analysis
* Checks the quality of the data. Anyone can get a brief overview from this
* The below image show cases the summary stats of the incBio data set
* In this, the median and mean of the tl is almost similar
* There are almost 165 fields with NA’s in the w column



**R Applications:**

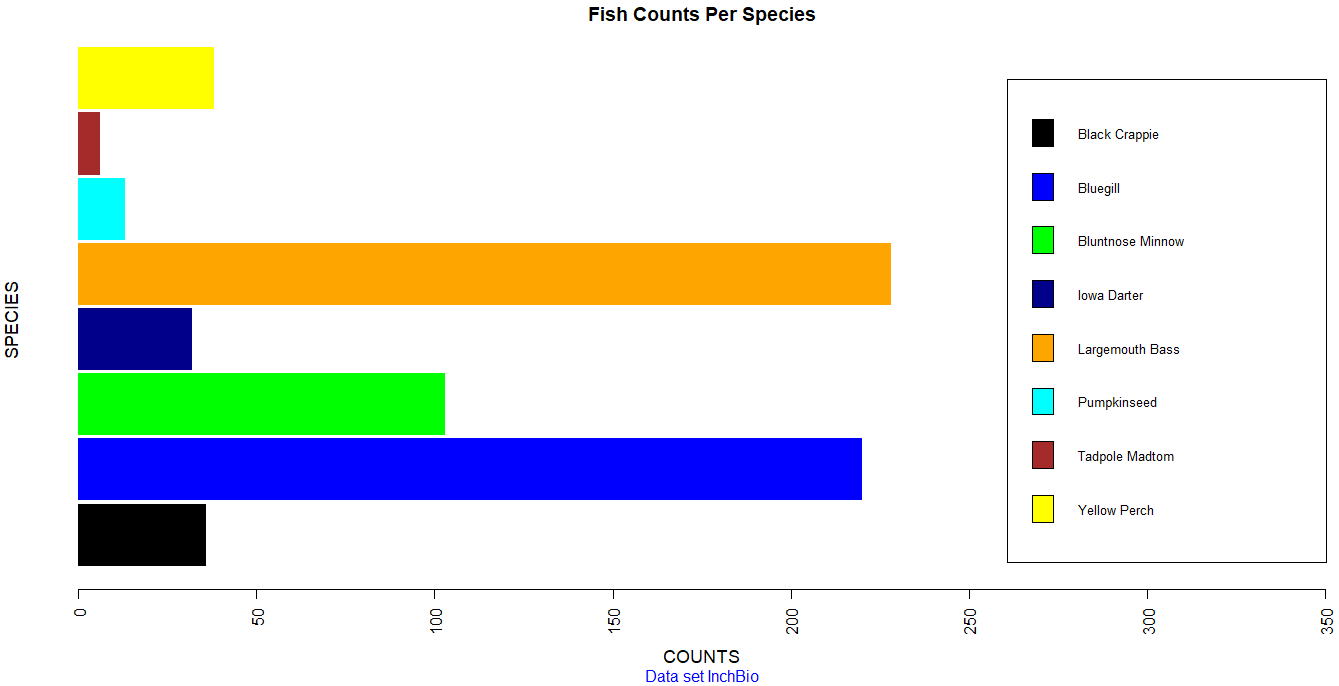
* Data analysts and researchers uses R
* Mostly used in finance domain
* R can import, clean, alter, and modify the data and generate graphs accordingly
* Provides glimpses of Descriptive Stats easily

**Basic Bar Plot of the InchBio Species:**



In this we can observe that the number of Largemouth Bass and Bluegill alone allocate around 66% of total species where as the least number is Tadpole Madtom species. Yellow Perch and Black Crappie are also similar in count.

**Plot 1: Fish Counts Per Species**

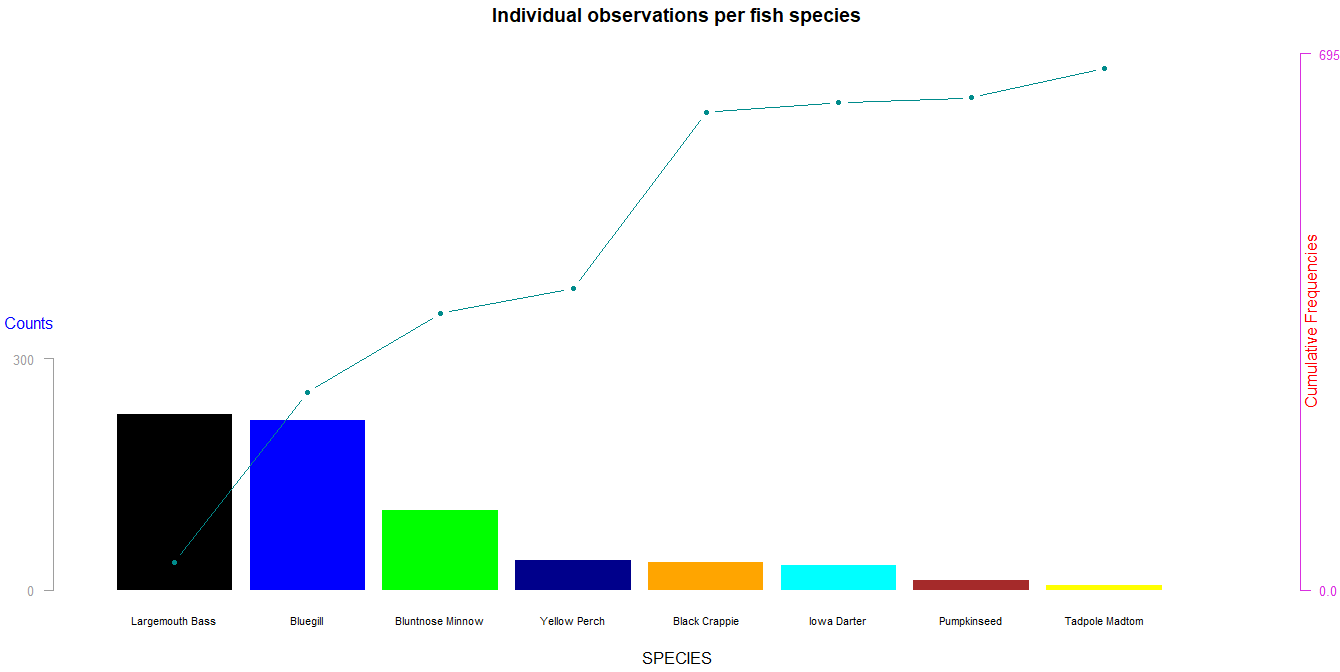


Plotted a Bar plot of Frequencies data and titled as "Fish Counts per species" with a Subtitle "Data set InchBio”. In this I have analysed the basic Fish Counts per species data and noticed that there is significance level of similarity between Largemouth Bass and Bluegill. We can also notice that the Black Crappie and Yellow Perch are somewhat similar

**Plot 2: Individual Observations per fish species – Pareto Chart**

* It is a blend of a bar chart and a line chart used for visualization.
* In this, the right axis is used for cum. frequency while the left axis is used for freq. The famous Pareto value tells that 80% work can be done from 20% of efforts.
* We have a bar chart that showcases the freq. of occurrence in various levels in descending order, and a line chart tells the cum. Percentages accordingly.

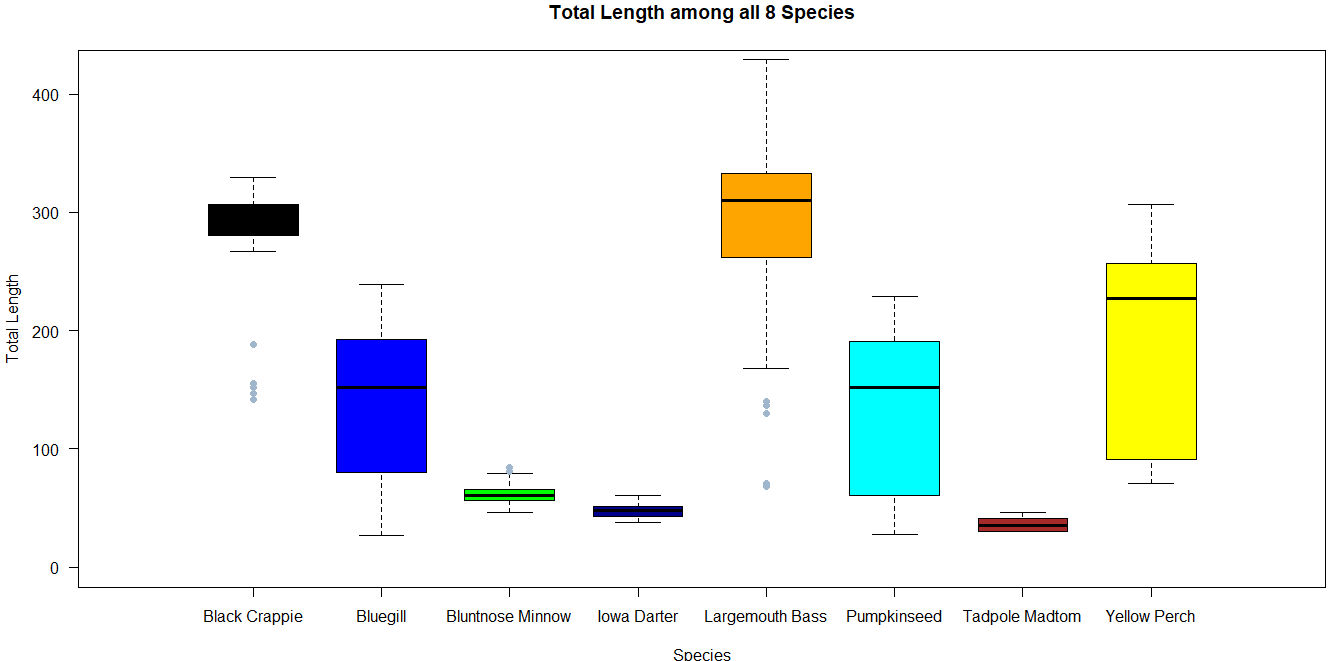
Instead of showing cumulative percentages, it shows cumulative counts. In this, the y-axis limit has been increased from its original height to ylim = c (0, 3.05 \* max (FinalTable$Freq, na.rm = TRUE)), where na.rm is to remove NA values from the data frame



In this we have used axis () function to create the axes on left and right of the graph which specifies the Counts and Cum. Frequencies accordingly. The axes follow this: 1 means below, 2 is left, 3 means above and 4 is right

Also, added custom text labels using mtext (). Las is a numeric value gives the orientation of the labels. 0: Always parallel to the axis, 1: Always horizontal, 2: Always perpendicular, and 3: Always vertical

**Plot 3: Total Length among all 8 species**



* Created a boxplot to compare the Total length among all the 8 species of the bio data set.
* It is used for graphically showing groups of numerical data
* They can be drawn horizontally or vertically depending on ones interests

**Conclusions:**

Analysed the InchBio Data set and drawn few conclusions and reports from the data. 66 % of the species were under Largemouth bass and Bluegill. Also, obtained a Pareto chart to showcase the 80% of effects are produced by the 20% of the causes.

**Bibliography:**

1. Axes and Text, <https://www.statmethods.net/advgraphs/axes.html>
2. Bar Chart Legend Position, <https://stackoverflow.com/questions/27688754/bar-chart-legend-position-avoiding-operlap-in-r>
3. Data camp’s R Tutorial, https://www.youtube.com/watch?v=9f2g7RN5N0I&list=PLjgj6kdf\_snYBkIsWQYcYtUZiDpam7ygg&index=12&ab\_channel=DataCamp
4. Books related to R, <https://www.r-project.org/doc/bib/R-books.html>
5. R Graph Gallery, <https://www.r-graph-gallery.com/>
6. Bar plot, <https://www.statmethods.net/graphs/bar.html>

**Appendix**

An additional file containing the R codes has been attached to this report. The name of this file is **M3\_Projet 3\_Thota, Sunil Raj**