Title: Gehlot_M3_Project3.pdf

Northeastern University – College of Professional Studies

ALY6000 – Introduction to Data Analytics

Instructor Name: Kayal Chandrasekaran

Due Date : January 29th,2023

Introduction

In this report, the inchBio dataset has been studied. The data contains different kinds of species of fish and will create understandable visualization and summarize this data for making major decisions. With the help of this data will calculate counts and probabilities based on categorical data and demonstrate a graph.

Key Findings

A. Analysis of descriptive characteristics

Upon loading the dataset, it has been found that the sample size of the dataset is 676 (using length()) and consists of 7 variables: netID, fishID, species, tl, w, tag and scale. Species data is a categorical data.There are eight types of different fishes: "Black Crappie", "Bluegill", "Bluntnose Minnow", "Iowa Darter", "Largemouth Bass", "Pumpkin seed", "Tadpole Madtom", "Yellow Perch". "Largemouth(228)" is found in major quantities in comparison to others while the "Tadpole Madtom(6)" count is the least.

```
> tmp

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

B. Visualizations

From the dataset, the following graphs have been analyzed: Fish count bar plot, Fish relative frequency and a specie Pareto chart. From Fish count bar plot its visible that Largemouth and Bluegill are a major portion of the data and Bluntnose Minnow (103) is also in good quantity. It shows that there is a high demand for these fish in the market. It is also visible in the Relative frequency graph that other fish's frequency is less than 0.2,

while Largemouth and Bluegill are 0.3 and 0.4 frequency. In the Specie Pareto graph new variable added to the dataset : cumfreq, count, cumcounts and further analysis has been done. The line shows cumulative frequency of the data.
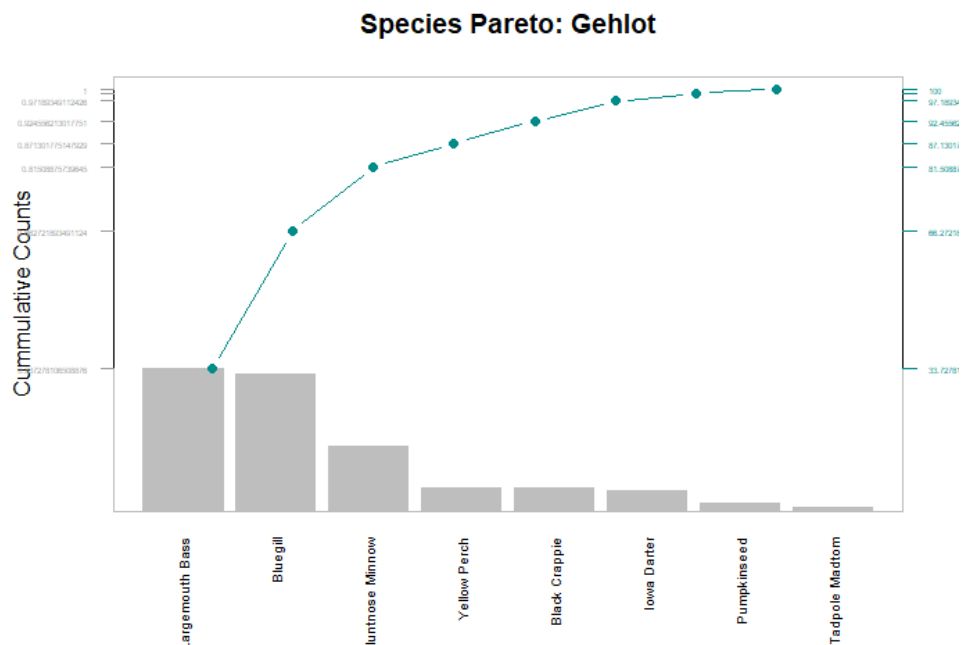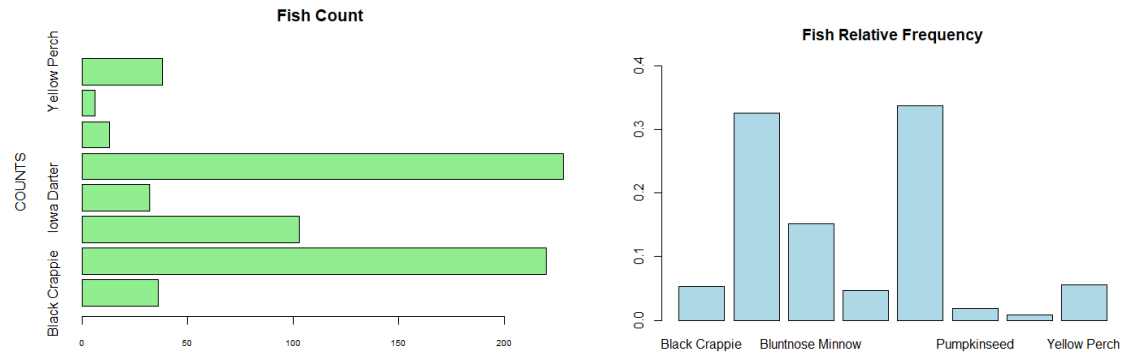
C. Summary of Key Points

Overall, from the inchBio dataset 676 fish has been analysed, where Largemouth and Bluegill are in major proportions following the Bluntnose Minnow and rest of them are in less proportion. This data can be analysed with more graphs and various inferences can be made.

References

- R, Programming. "How to Make Frequency Table in R - Programming R Tutorials." *ProgrammingR*, 3 Apr. 2021, https://www.programmingr.com/statistics/frequency-table/.
- Nnk. "Order Dataframe by One Descending and One Ascending Column in R." *Spark By {Examples}*, 25 Aug. 2022, https://sparkbyexamples.com/r-programming/order-dataframe-by-one-descending-and-one-ascending-column-in-r/.
- Suhani, et al. "Rename Data Frame Columns in R." *Datanovia*, 19 Oct. 2018, https://www.datanovia.com/en/lessons/rename-data-frame-columns-in-r/.
- SkarabSkarab 98744 gold badges1111 silver badges1414 bronze badges, and Joshua UlrichJoshua Ulrich 1. "How to Increase Size of Label Fonts in Barplot." *Cross Validated*, 1 Oct. 1957, https://stats.stackexchange.com/questions/3853/how-to-increase-size-of-label-fonts-in-barplot.
- Documentation, R. "Dra a Box around a Plot." *R: Draw A Box around a Plot*, Apr. 2023, https://stat.ethz.ch/R-manual/R-patched/library/graphics/html/box.html.