Title: Gehlot_M2_Project2.pdf

Northeastern University – College of Professional Studies

ALY6000 – Introduction to Data Analytics

Instructor Name: Kayal Chandrasekaran

Due Date : January 22nd,2023

Analysis

A. The given BullTroutRML2 dataset consists of five variables such as ID,age, fl, lake and era.
Object ID,“age” and “fl” are numeric data types and “lake” and “era” are character data type.
Upon printing the first and last three records by using head() and tail(), there is a large difference in age and era.

```
> df1 <-rbind(head(BullTroutRML2,1),tail(BullTroutRML2,3))
> df1
# A tibble: 4 × 5
     ID   age    fl lake     era
  <dbl> <dbl> <dbl> <chr>    <chr>
1     1    14   459 Harrison 1977-80
2    94     4   298 Osprey   1997-01
3    95     3   279 Osprey   1997-01
4    96     3   273 Osprey   1997-01
> |
```

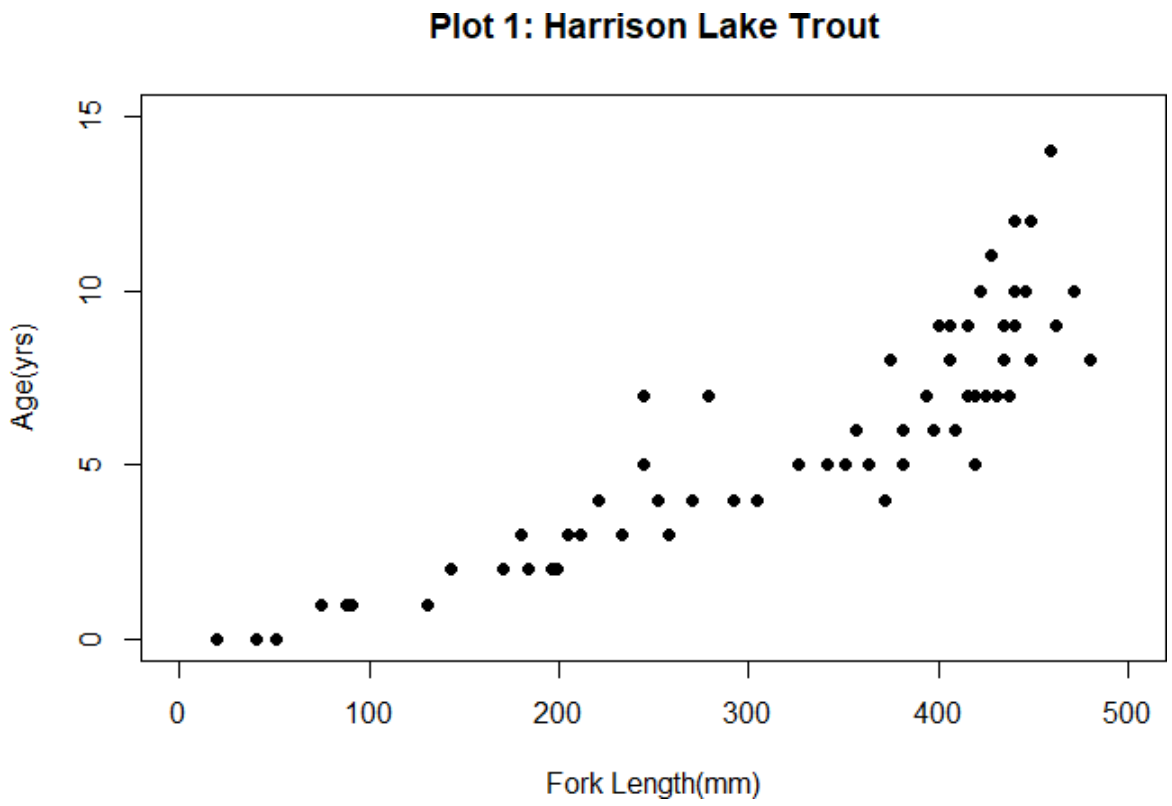
Later the dataset has been filtered for more visualization of only Harrison Lake. Dataset has reduced from 96 observations to 61. As it’s visible in the filtered data’s first and last records that there is a sudden drop in age from 14 yrs to 0yr and then fluctuation.

```
> #display first and last 5 records from filtered data
> df <- rbind(head(harrilake,1),tail(harrilake,5))
> df
# A tibble: 6 × 5
     ID   age    fl lake     era
  <dbl> <dbl> <dbl> <chr>    <chr>
1     1    14   459 Harrison 1977-80
2    57     0    41 Harrison 1997-01
3    58     0    20 Harrison 1997-01
4    59     7   245 Harrison 1997-01
5    60     7   279 Harrison 1997-01
6    61     5   245 Harrison 1997-01
> |
```

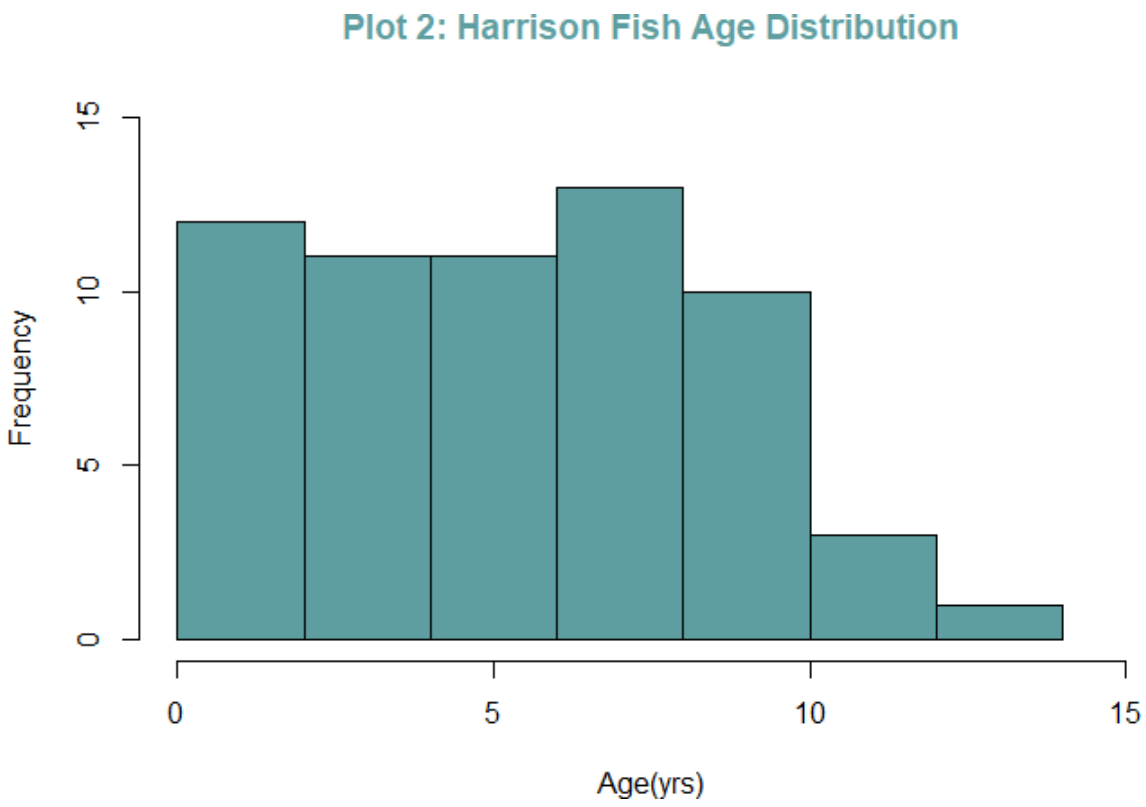
The filtered data frame renamed as “Harrilake” and the data summary is as follows:
The average is 5.75 and range is from 0 to 14. The median of Fork length is 372 and mean is 319.

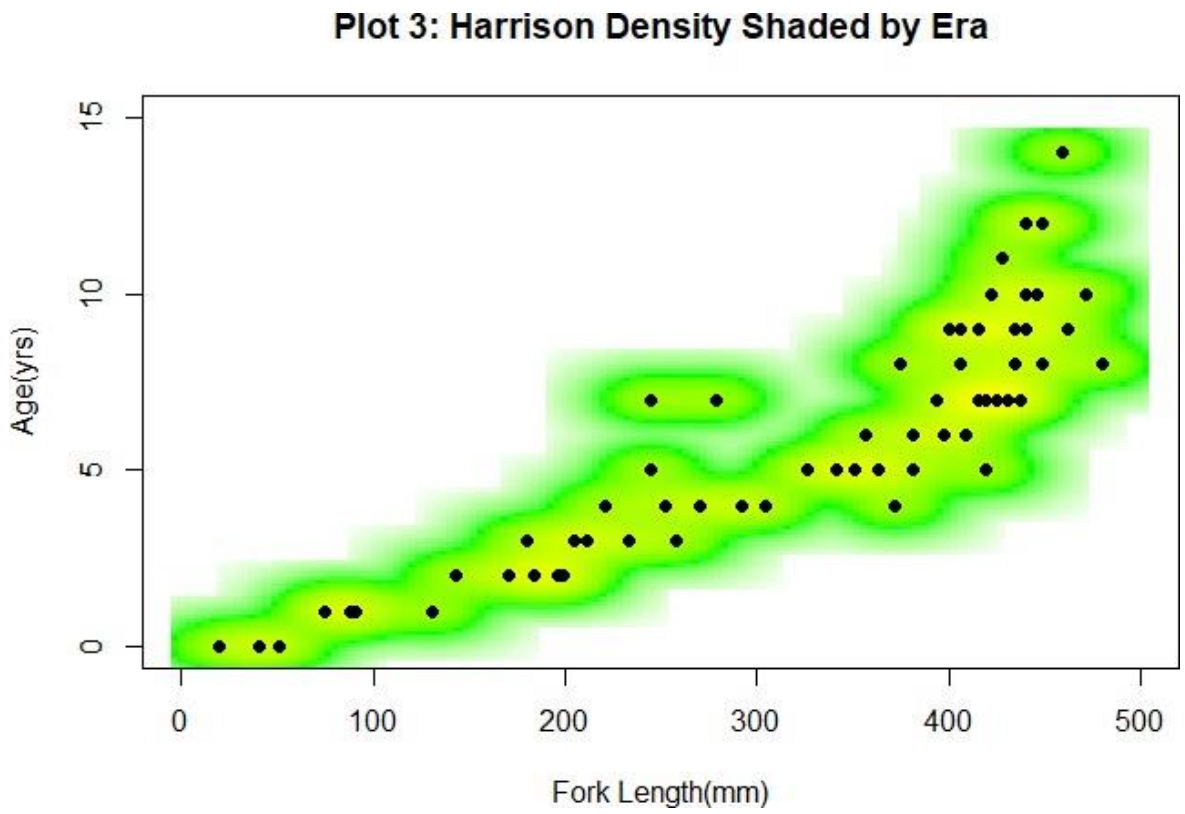
```
> summary(harrilake)
       ID             age              fl          lake               era
 Min.   : 1   Min.   : 0.000   Min.   : 20   Length:61          Length:61
 1st Qu.:16   1st Qu.: 3.000   1st Qu.:221   Class :character   Class :character
 Median :31   Median : 6.000   Median :372   Mode  :character   Mode  :character
 Mean   :31   Mean   : 5.754   Mean   :319
 3rd Qu.:46   3rd Qu.: 8.000   3rd Qu.:425
 Max.   :61   Max.   :14.000   Max.   :480
> str(harrilake)
tibble [61 × 5] (S3: tbl_df/tbl/data.frame)
 $ ID  : num [1:61] 1 2 3 4 5 6 7 8 9 10 ...
 $ age : num [1:61] 14 12 10 10 9 9 9 8 8 7 ...
 $ fl  : num [1:61] 459 449 471 446 400 440 462 480 449 437 ...
 $ lake: chr [1:61] "Harrison" "Harrison" "Harrison" "Harrison" ...
 $ era : chr [1:61] "1977-80" "1977-80" "1977-80" "1977-80" ...
>
```

B. The correlation between Fork length(mm) vs Age(yrs) graph is 0.8848, which signifies the strong positive correlation between two variables. There is a growth in age as fork length increases. This statistical analysis helps in understanding of the fisheries.

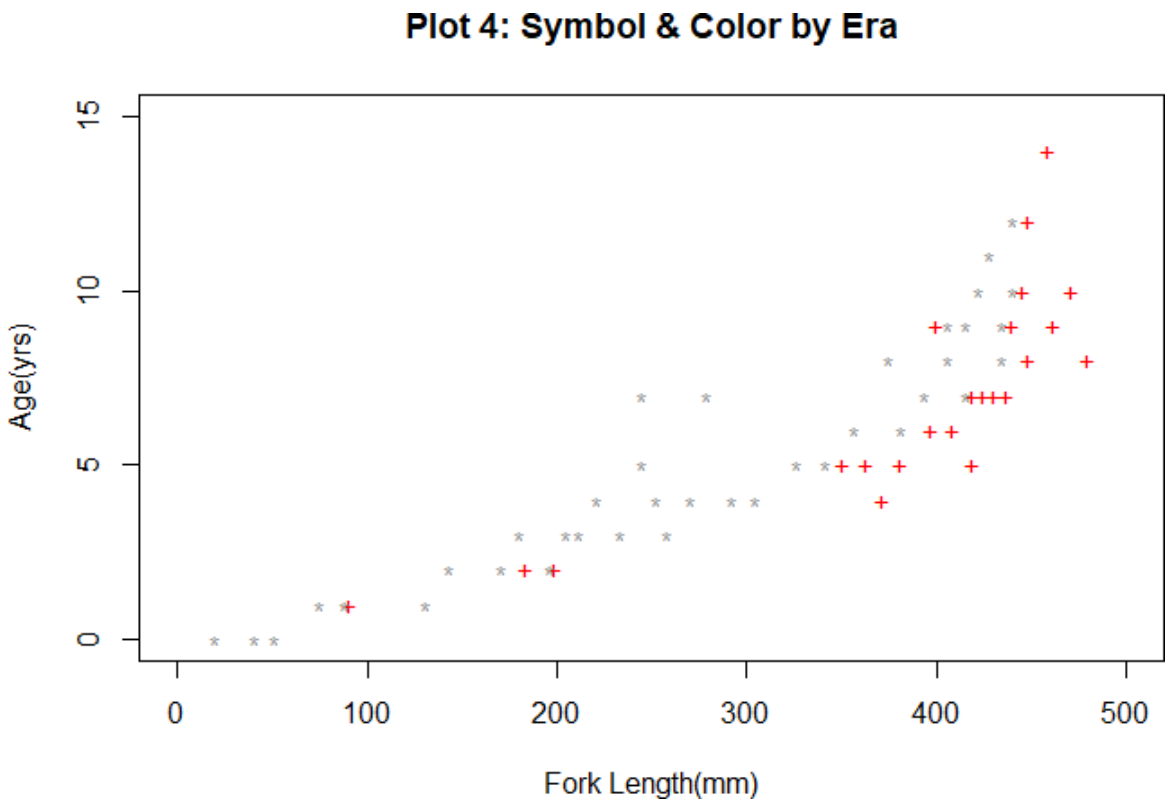


“Harrison Fish Age Distribution” histogram is skewed right where mean is greater than median. It shows that the Fish between age from 0yrs to 10 yrs exists more in quantity while above 15 yrs their population keeps declining.

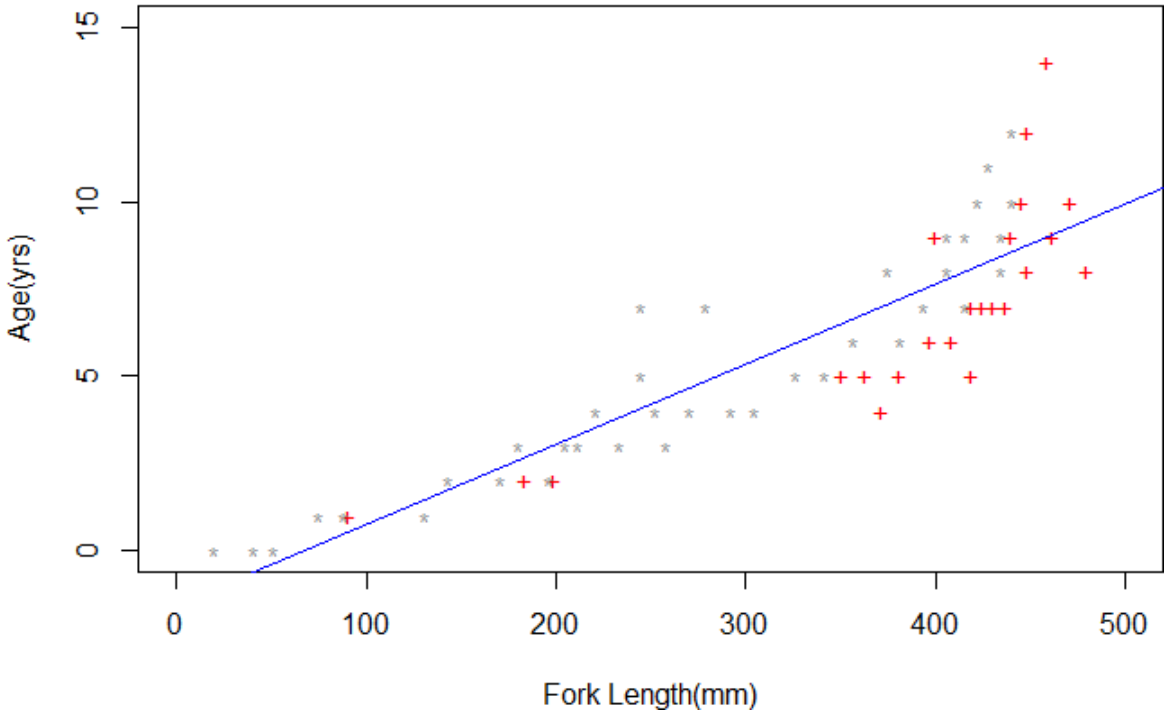


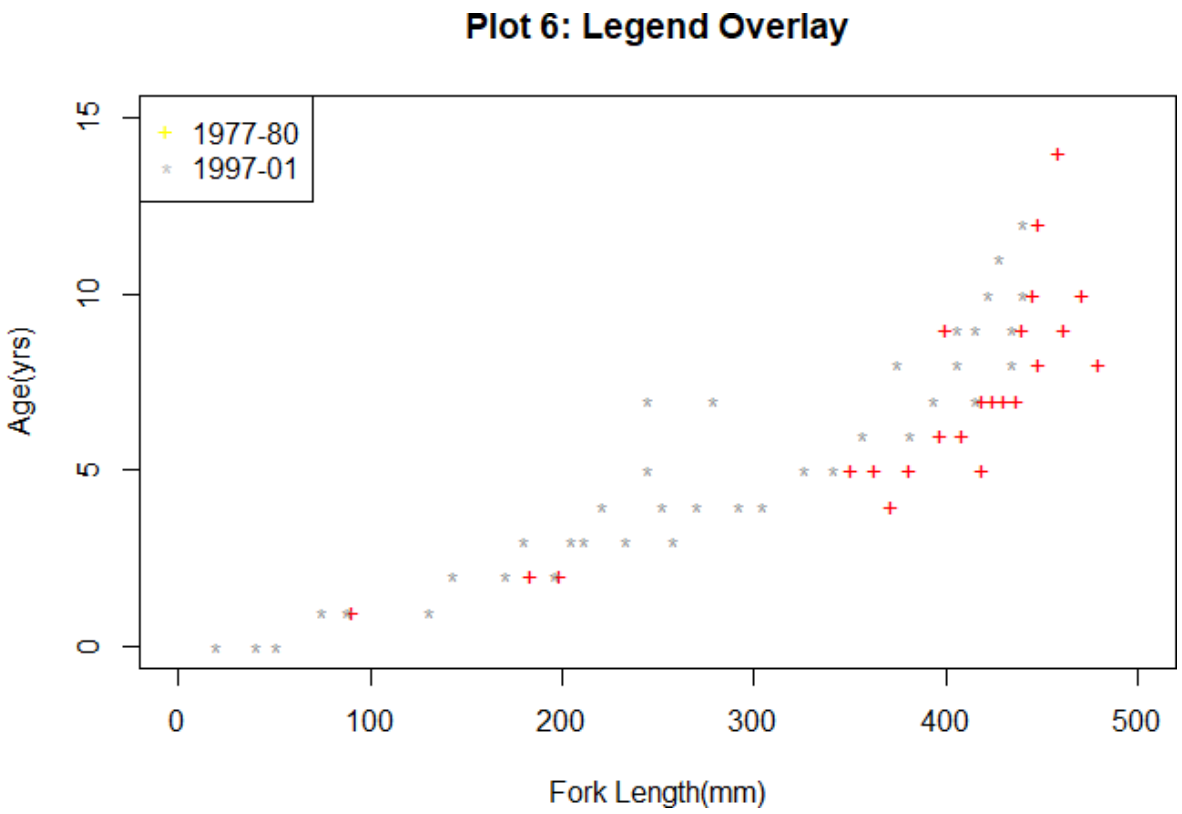

Plot 3: Harrison Density Shaded by Era
Age(yrs)
0
5
10
15
0
100
200
300
400
500
Fork Length(mm)

Ty


Plot 4: Symbol & Color by Era
Age(yrs)
0
5
10
15
0
100
200
300
400
500
Fork Length(mm)

Plot 5: Regression





C. From the above graphs, the dataset is visualized in various approaches. The density plot helps in analyzing the distributed data give an idea about outliers from graph age=15 is an outlier. From plot, it is identified that data points are not overlapping.

Bibliography:

- MarinStatsLectures-R Programming & Statistics. (2013, August 9). *Marinstatslectures-R Programming & Statistics*. YouTube. Retrieved January 22, 2023, from https://www.youtube.com/user/marinstatlectures
- R_UserR_User 10.4k2424 gold badges7777 silver badges120120 bronze badges *et al.* (1960) *Colorize parts of the title in a plot, Stack Overflow*. Available at: https://stackoverflow.com/questions/17083362/colorize-parts-of-the-title-in-a-plot (Accessed: January 22, 2023).
- *Convert data frame column to numeric in R (2 example codes)* (2022) *Statistics Globe*. Available at: https://statisticsglobe.com/convert-data-frame-column-to-numeric-in-r (Accessed: January 22, 2023).
- Jim (2022) *Remove rows from the data frame in R: R-bloggers, R*. Available at: https://www.r-bloggers.com/2022/06/remove-rows-from-the-data-frame-in-r/ (Accessed: January 22, 2023.