Executive Summary Report 2

6th October 2021

HARSHIT GAUR

NUID: 001093079

To: Prof. Richard He

KEY FINDINGS

1. Import of Libraries & Load of 'BullTroutRML2.csv' data set

```
#STEP 2: Installing the packages.
install.packages("FSA")
install.packages("FSAdata")
install.packages("magrittr")
install.packages("dplyr")
install.packages("plotrix")
install.packages("ggplot2")
install.packages("moments")

#STEP 2: Importing the packages.
library("FSA")
library("FSAdata")
library("magrittr")
library("dplyr")
library("plotrix")
library("ggplot2")
library("moments")
```

Figure 1.1

	age	fl	lake	era
1	14	459	Harrison	1977-80
2	12	449	Harrison	1977-80
3	10	471	Harrison	1977-80
4	10	446	Harrison	1977-80
5	9	400	Harrison	1977-80
6	9	440	Harrison	1977-80
7	9	462	Harrison	1977-80
8	8	480	Harrison	1977-80
9	8	449	Harrison	1977-80
10	7	437	Harrison	1977-80
11	7	431	Harrison	1977-80
12	7	425	Harrison	1977-80

Figure 1.3

```
#STEP 3: Import 'BullTroutRML2.csv' data set
data(BullTroutRML2)
View(BullTroutRML2)
```

Figure 1.2

Data	
BullTroutRML2	96 obs. of 4 variables

Figure 1.4

a. The libraries - FSA, FSAdata, magrittr, dplyr, plotrix, ggplot2, and moments were installed and imported successfully in the project.
b. BullTroutRML2 CSV data set was loaded into the memory of the project and used further. A screenshot has been attached of a snippet of the data present in it.

2. Operations on the data set

```
> #STEP 4: Print first 3 records of 'BullTroutRML2.csv' data set
> head(BullTroutRML2, 3)
  age  fl     lake     era
1  14 459 Harrison 1977-80
2  12 449 Harrison 1977-80
3  10 471 Harrison 1977-80
```

Figure 2.1

```
> #STEP 4: Print last 3 records of 'BullTroutRML2.csv' data set
> tail(BullTroutRML2, 3)
   age  fl   lake     era
94   4 298 Osprey 1997-01
95   3 279 Osprey 1997-01
96   3 273 Osprey 1997-01
```

Figure 2.2

a. Basic R operations on the data set were performed.
    i. *head(x, n)* function allows to select first *n* records from *x* data set.
    ii. *tail(x, n)* function allows to select last *n* records from *x* data set.

3. Filter on the data set

```
#STEP 5: Remove all records except those of 'Harrison' lake
#from 'BullTroutRML2.csv' data set
BullTroutRML2 <- filterD(BullTroutRML2, lake == 'Harrison')
```

Figure 3.1

Data	
BullTroutRML2	61 obs. of 4 variables

Figure 3.2

a. Filter functions can be used to filter out the data set and get the desired data set
    i. filterD(x, ..., except = NULL) function allows to filter the data set *x* with second argument as the condition of the filter.

4. Verification of the data set after filtrations

	age	fl	lake	era
1	14	459	Harrison	1977-80
2	12	449	Harrison	1977-80
3	10	471	Harrison	1977-80
4	10	446	Harrison	1977-80
5	9	400	Harrison	1977-80

Figure 4.1

	age	fl	lake	era
57	0	41	Harrison	1997-01
58	0	20	Harrison	1997-01
59	7	245	Harrison	1997-01
60	7	279	Harrison	1997-01
61	5	245	Harrison	1997-01

Figure 4.2

5. Structure of the BullTroutRML2 data set

```
> #STEP 7: Display the structure of the filtered 'BullTroutRML2.csv' data set
> str(BullTroutRML2)
'data.frame':   61 obs. of  4 variables:
 $ age : int  14 12 10 10 9 9 9 8 8 7 ...
 $ fl  : int  459 449 471 446 400 440 462 480 449 437 ...
 $ lake: Factor w/ 1 level "Harrison": 1 1 1 1 1 1 1 1 1 1 ...
 $ era : Factor w/ 2 levels "1977-80","1997-01": 1 1 1 1 1 1 1 1 1 1 ...
```

Figure 5.1

a. The BullTroutRML data set contains 96 observations which has been filtered out to contain *'Harrison'* lake records only. There are 61 observations belonging to it.
b. The variables *'age'* & *'fl'* contain observations in integer data type.
c. The variables *'lake'* & *'era'* contain observations in character data type which has been converted implicitly & internally into *'Factors'* by the R-script for efficiency.

6. Summary of the BullTroutRML2 data set

```
> #STEP 8: Display the summary of the filtered 'BullTroutRML2.csv' data set
> summary(BullTroutRML2)
      age               fl            lake         era
 Min.   : 0.000   Min.   : 20   Harrison:61   1977-80:23
 1st Qu.: 3.000   1st Qu.:221                 1997-01:38
 Median : 6.000   Median :372
 Mean   : 5.754   Mean   :319
 3rd Qu.: 8.000   3rd Qu.:425
 Max.   :14.000   Max.   :480
```

Figure 6.1

a. AGE :
   i. The minimum age is **0** years old and maximum age is **14** years old.
   ii. Median is **6.000** years old.
   iii. Mean is **5.754** years old.
   iv. The median and mean values are close to each other and around the middle of the data set. Therefore, outliers will not affect the data set adversely.

b. FL (Fork Length) :
   i. The minimum fork length is **20** mm and maximum fork length is **480** mm.
   ii. Median is **372** mm.
   iii. Mean is **319** mm.
   iv. Since, the median and mean values are not very close to each other and the graph is also **Left Skewed**, outliers can affect the statistics a little.

c. LAKE : We have already filtered out the data set to contain records of 'Harrison' lake only.

d. ERA :
   i. The whole filtered data set belongs to 2 eras only which have been factored.
   ii. **1977-80 (23 records)** and **1997-01 (38 records)**

7. Scatter Plot of Harrison Lake Trout (fl ~ age)

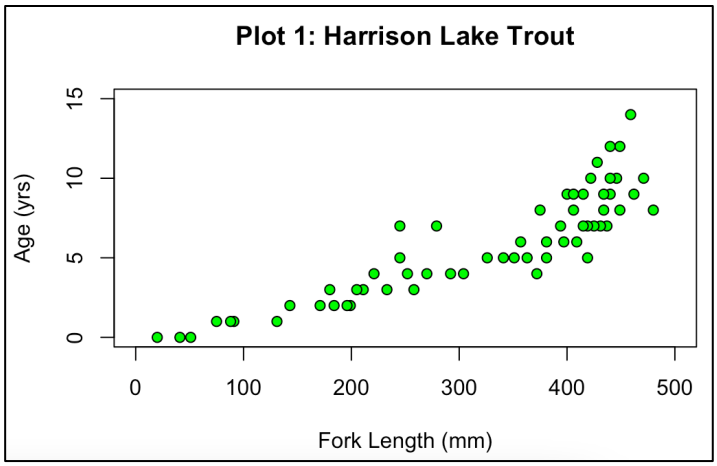

Figure 7.1 - Plot 1: Harrison Lake Trout

a. The graph depicts that the density of the scatter plot is more in the **range of 350 mm to 480 mm.**
b. From the graph, the fork length of Bull Trout living in Harrison lake increases sharply from **20 mm up to 380 mm** with **a slope around 0.5 up to the age of 7-8 years old**. After **7-8 years till 15 years old**, the **growth rate of fork length gets decreased**.

8. Histogram Graph for Harrison Fish Age Distribution

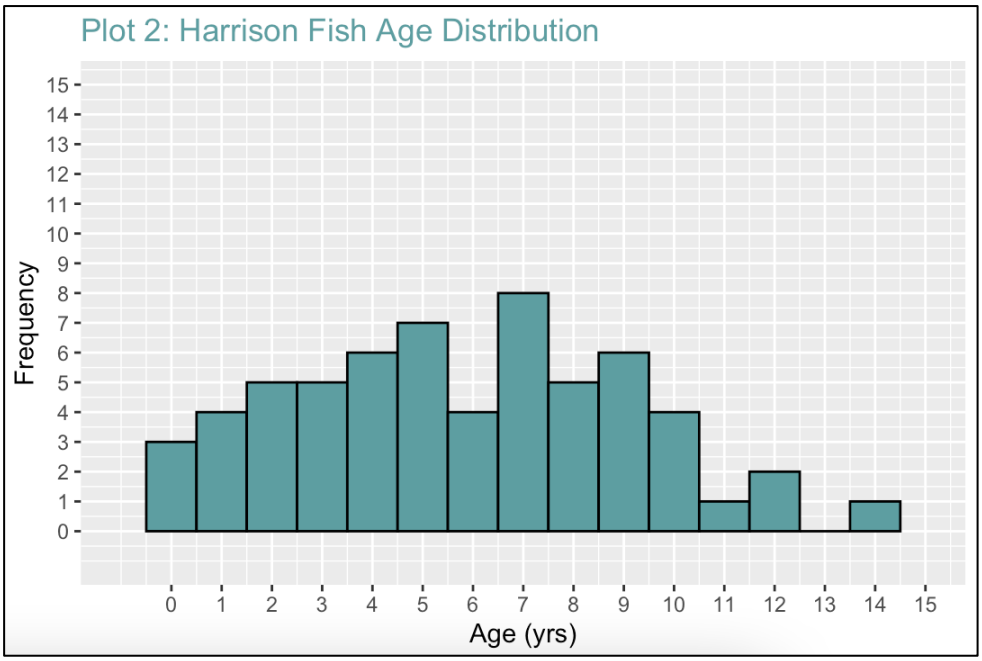

Figure 8.1 - Plot 2: Harrison Fish Age Distribution

a. The histogram depicts that most number of records (8) are present in the age group of 7.
b. It shows that the most of the age groups have close frequency ranges. The ranges of frequency vary from 4 to 7 records in these groups.
c. It also conveys that there are no Bull Trout belonging to the age group of 13.

9. Over-Dense Plot - Harrison Density Shaded by Era

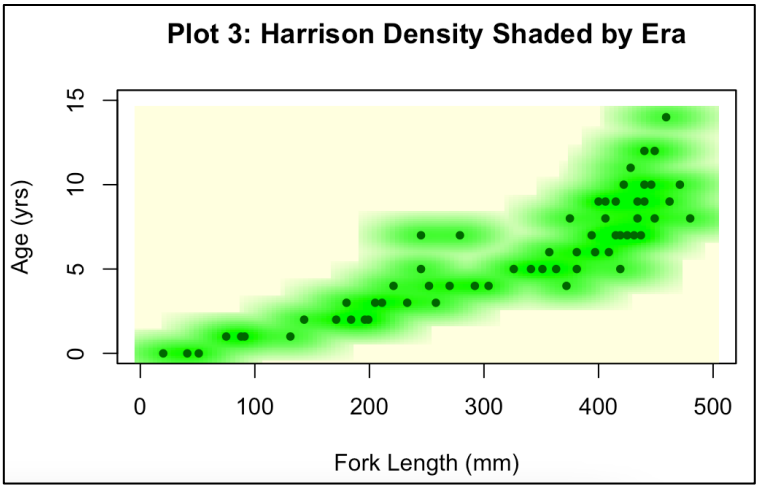

Figure 9.1 - Plot 3: Harrison Density Shaded by Era

a. This graph also depicts in a more clear way that the density of the scatter plot is more in the **range of 350 mm to 480 mm.**
b. We can infer the above points inferred in the section of 'Plot 1' more clearly from this graph.
c. The growth rate of fork length is sharp with a **slope around 0.5** in the age group of 0 to 7-8 years and then gets decreased in the age group of 7-8 years to 15 years.

10. Add First 5 and Last 5 Records in a new object

```
tmp <- headtail(BullTroutRML2, 3)
View(tmp)
```

Figure 10.1

	age	fl	lake	era
1	14	459	Harrison	1977-80
2	12	449	Harrison	1977-80
3	10	471	Harrison	1977-80
59	7	245	Harrison	1997-01
60	7	279	Harrison	1997-01
61	5	245	Harrison	1997-01

Figure 10.2

11. Display the 'era' variable in the new object

```
#STEP 13: Display 'era' column from the 'tmp' variable.
View(tmp['era'])
```

Figure 11.1

	era
1	1977-80
2	1977-80
3	1977-80
59	1997-01
60	1997-01
61	1997-01

Figure 11.2

12. Convert new object's 'era' variable to numeric values

```
#STEP 16: Convert 'temp$era' values to numeric values
original_tmp_era <- tmp$era
tmp$era <- as.numeric(tmp$era)

#STEP 17: Initialize 'cols' vector with 'temp$era' values
cols <- unique(tmp$era)
```

Figure 11.1

13.Plot of Age vs Fork Length - Symbol & Colour by Era

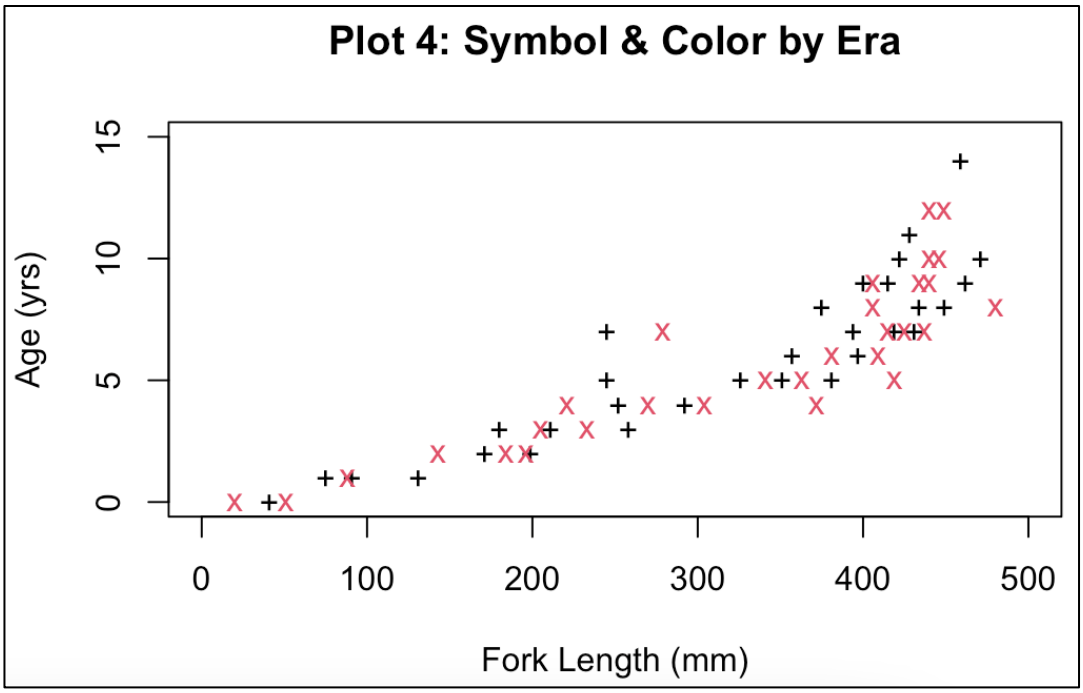

Figure 11.1

a. In this graph, we have separated both the eras with different symbols and colours.
b. The inferences remain the same as the plot 1 and plot 3.

14.Plot of Age vs Fork Length - Regression Overlay

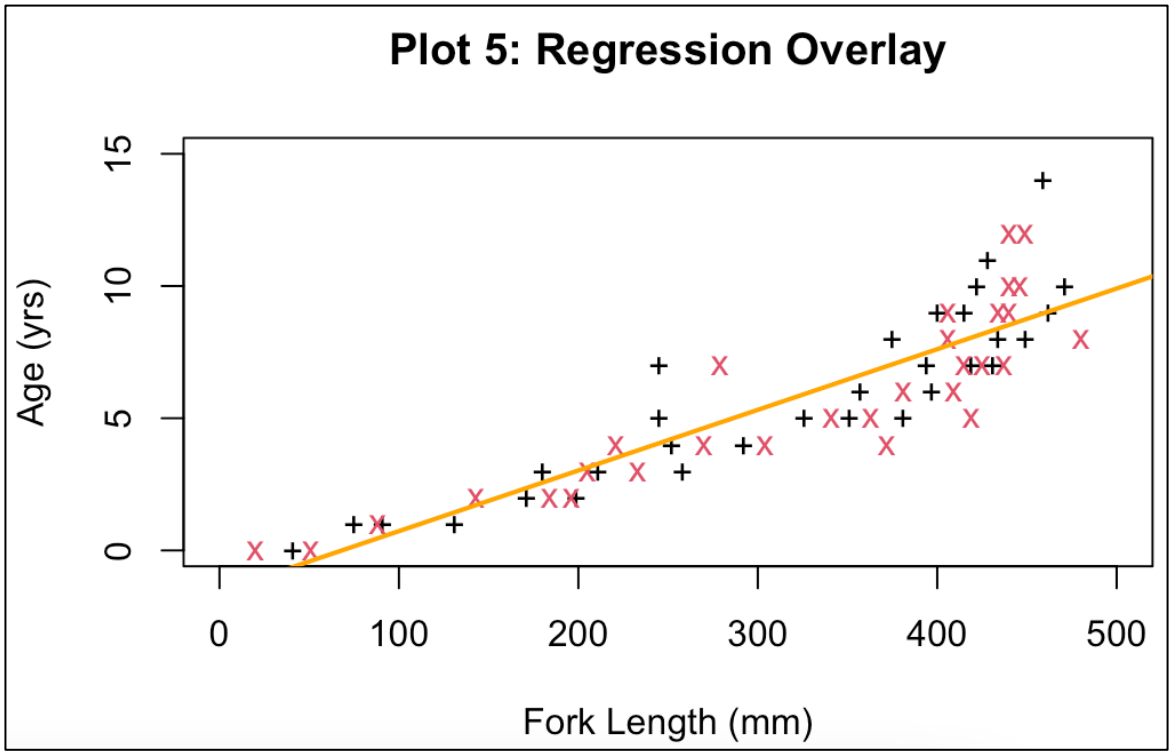

Figure 11.1

a. We have added a regression linear model to depict the relationship between the two variables.
b. The **slope of the graph is around 0.5**

15.Plot of Age vs Fork Length - Legend Overlay

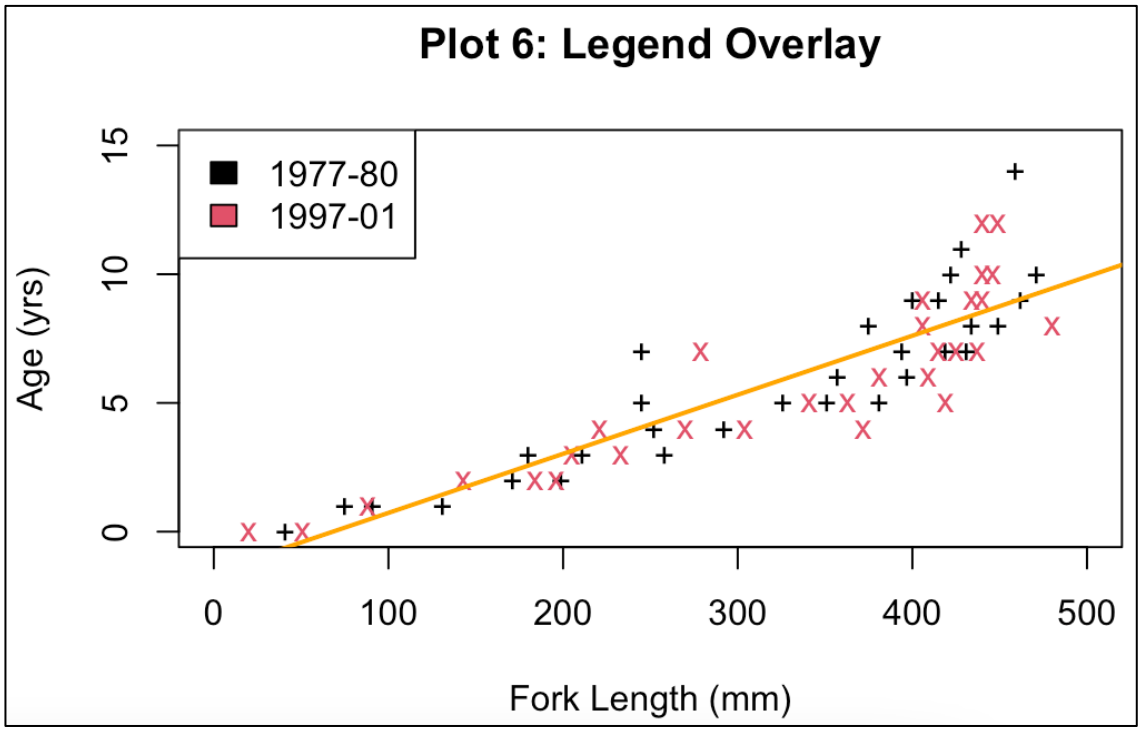

Figure 11.1

a. In this graph, legend has been added for both the factors of *'era'*
b. The presentation value of this graph has been increased with the addition of regression linear model line and legend for the eras.

SUMMARY

1. We were able to find out the statistics of the BullTroutRML2 data set and analysed the following :
   a. The variable ***'age'*** ranged from **0 to 14 years old** with a **mean as 5.754 years old** and **median as 6 years old**. The median and mean values are close to each other and around the middle of the data set. Therefore, outliers will not have adverse effect on the statistics of the 'age' data set.
   b. The variable ***'fl'*** (Fork Length) ranged from **20 mm to 480 mm** with **mean of 319 mm** and **median of 372 mm**. The graph is also **Left Skewed**.

2. The **density** of Bull Trout in the age group **8 years - 15 years** is **more** for the range of **fork lengths 350 mm - 480 mm**.

3. The fork length of Bull Trout living in Harrison lake increases sharply from **20 mm up to 380 mm** with **a slope around 0.5 up to the age of 7-8 years old** and afterwards **till 15 years old**, the **growth rate of fork length gets decreased**.

4. The relationship between age and fork length of Bull Trout is linear

BIBLIOGRAPHY

1. *R plot pch symbols : The different point shapes available in R - Easy Guides - Wiki - STHDA*. (2021). STHDA. http://www.sthda.com/english/wiki/r-plot-pch-symbols-the-different-point-shapes-available-in-r

2. *ggplot2 histogram plot : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA*. (2021). STHDA GGPLOT2. http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization

3. D. (2021a, May 4). *5 tips to make better histograms with ggplot2 in R*. Data Viz with Python and R. https://datavizpyr.com/histograms-with-ggplot2-in-r/

4. *How to set limits for axes in ggplot2 R plots?* (2010, August 31). Stack Overflow. https://stackoverflow.com/questions/3606697/how-to-set-limits-for-axes-in-ggplot2-r-plots

5. *ggplot2 point shapes - Easy Guides - Wiki - STHDA*. (2021). STHDA Ggplot Point Shapes. http://www.sthda.com/english/wiki/ggplot2-point-shapes

APPENDIX

```
#---------------------- GAUR_M2_PROJECT2 ----------------------#

#STEP 1: Printing my last-name with a prefix.
print("Plotting Basics: GAUR")

#STEP 2: Installing the packages.
install.packages("FSA")
install.packages("FSAdata")
install.packages("magrittr")
install.packages("dplyr")
install.packages("plotrix")
install.packages("ggplot2")
install.packages("moments")

#STEP 2: Importing the packages.
library("FSA")
library("FSAdata")
library("magrittr")
library("dplyr")
library("plotrix")
library("ggplot2")
library("moments")

#STEP 3: Import 'BullTroutRML2.csv' data set
data(BullTroutRML2)
View(BullTroutRML2)

#STEP 4: Print first 3 records of 'BullTroutRML2.csv' data set
head(BullTroutRML2, 3)
#STEP 4: Print last 3 records of 'BullTroutRML2.csv' data set
tail(BullTroutRML2, 3)

#STEP 5: Remove all records except those of 'Harrison' lake
#from 'BullTroutRML2.csv' data set
BullTroutRML2 <- filterD(BullTroutRML2, lake == 'Harrison')
View(BullTroutRML2)

#STEP 6: Print first 5 and last 5 records of the filtered 'BullTroutRML2.csv' data set
View(head(BullTroutRML2, 5))
View(tail(BullTroutRML2, 5))

#STEP 7: Display the structure of the filtered 'BullTroutRML2.csv' data set
str(BullTroutRML2)

#STEP 8: Display the summary of the filtered 'BullTroutRML2.csv' data set
summary(BullTroutRML2)

#STEP 9: Plot a Scatter plot for fl ~ age
# ----------- Plot 1: Harrison Lake Trout ----------- #

plot(BullTroutRML2$fl, BullTroutRML2$age, xlim = c(0,500), ylim = c(0,15), main = "Plot 1: Harrison Lake Trout", xlab = "Fork
Length (mm)", ylab = "Age (yrs)", pch = 21, bg = "GREEN")


#STEP 10: Plot an histogram graph for age using 'ggplot'
# ----------- Plot 2: Harrison Fish Age Distribution ----------- #

ggplot(BullTroutRML2) +
  geom_histogram(
    mapping = aes(age),
    color="BLACK",
    fill="CADETBLUE",
    binwidth = 1
  ) +
  labs(
    title = "Plot 2: Harrison Fish Age Distribution",
    x = "Age (yrs)",
    y = "Frequency"
  ) +
  scale_x_continuous(limits = c(-1,15), breaks = c(0:15)) +
  scale_y_continuous(limits = c(-1,15), breaks = c(0:15)) +
  theme(plot.title = element_text(colour = "CADETBLUE"))
```

```
#STEP 11: Plot an Overlay Dense Plot (Smooth Scatter) for age using 'ggplot'
# ----------- Plot 3: Harrison Density Shaded by Era ----------- #

smoothScatter(x = BullTroutRML2$fl, y = BullTroutRML2$age, main = "Plot 3: Harrison Density Shaded by Era",
        xlab = "Fork Length (mm)", ylab = "Age (yrs)", xlim = c(0,500), ylim = c(0,15),
        pch = 20, colramp = colorRampPalette( c('LIGHTYELLOW','GREEN') ), col = 'DARKGREEN')

#STEP 12: Create a new variable 'tmp' the first 3 and last 3 records of the data set.
tmp <- headtail(BullTroutRML2, 3)
View(tmp)

#STEP 13: Display 'era' column from the 'tmp' variable.
View(tmp['era'])

#STEP 14: Create 'pchs' vector with arguments '+' and '*'
pchs <- c('+', 'x')

#STEP 15: Create 'cols' vector with arguments 'red' and 'gray60'
cols <- c('red', 'gray60')

#STEP 16: Convert 'temp$era' values to numeric values
original_tmp_era <- tmp$era
tmp$era <- as.numeric(tmp$era)

#STEP 17: Initialize 'cols' vector with 'temp$era' values
cols <- unique(tmp$era)

#STEP 18: Plot a graph for 'Age (yrs) versus Fork Length (mm)'
# ----------- Plot 4: Symbol & Color by Era ----------- #

plot(x = BullTroutRML2$fl, y = BullTroutRML2$age, main = "Plot 4: Symbol & Color by Era",
    xlab = "Fork Length (mm)", ylab = "Age (yrs)", xlim = c(0,500), ylim = c(0,15),
    pch = c(pchs), col = c(cols))


#STEP 18: Plot a graph for 'Age (yrs) versus Fork Length (mm)'
# ---------- Plot 5: Regression Overlay ---------- #

plot(x = BullTroutRML2$fl, y = BullTroutRML2$age, main = "Plot 5: Regression Overlay",
    xlab = "Fork Length (mm)", ylab = "Age (yrs)", xlim = c(0,500), ylim = c(0,15),
    pch = c(pchs), col = c(cols))
#Regression Line
abline(lm(age~fl, data = BullTroutRML2), col="ORANGE", lwd=2)


#STEP 19: Plot a graph for 'Age (yrs) versus Fork Length (mm)'
# ---------- Plot 6: Legend Overlay ---------- #

plot(x = BullTroutRML2$fl, y = BullTroutRML2$age, main = "Plot 6: Legend Overlay",
    xlab = "Fork Length (mm)", ylab = "Age (yrs)", xlim = c(0,500), ylim = c(0,15),
    pch = c(pchs), col = c(cols))
#Plot Regression Line
abline(lm(age~fl, data = BullTroutRML2), col="ORANGE", lwd=2)
#Add Legend
legend(x = "topleft", legend = unique(original_tmp_era), fill = unique(original_tmp_era))
```