EXECUTIVE SUMMARY REPORT 2

Module 2 assignment

Abstract

In this assignment, I will create a summary of some data and learn how to use R

Mohammad Hossein Movahedi

Movahedi.m@northeastern.edu

Objective:

The objective of this assignment is to get used to following

- Calculate basic descriptive statistics to describe a set of data
- Create various types of graphs based on data provided
- Use R to visualize data

Introduction

In this summary, I will use the data from the BullTroutRML2 dataset and use the APA system for the bibliography. Also, the R code that I used to process the data is available on my GitHub account, which its address is mentioned in the bibliography part. I also put my code in the appendix part.

1. Printing my name at the top of the script
    - print("Plotting Basics:Movahedi")

2. Importing and loading required libraries
    - install.packages("FSA")
    - install.packages("FSAdata")
    - install.packages("magrittr")
    - install.packages("dplyr")
    - install.packages("plotrix")
    - install.packages("ggplot2")
    - install.packages("moments")
    - #loading installed libraries
    - library(FSA)
    - library(FSAdata)
    - library(magrittr)
    - library(dplyr)
    - library(plotrix)
    - library(ggplot2)
    - library(moments)
3. Loading the BullTroutRML2 dataset
    - data(BullTroutRML2)
4. Printing the first and last 3 records from the BullTroutRMS2 dataset
    - #showing first three lines

- print(head(BullTroutRML2, n = 3L))
- #showing last three lines
- print(tail(BullTroutRML2, n = 3L))

```
  age  fl     lake     era
1  14 459 Harrison 1977-80
2  12 449 Harrison 1977-80
3  10 471 Harrison 1977-80

   age  fl   lake     era
94   4 298 Osprey 1997-01
95   3 279 Osprey 1997-01
96   3 273 Osprey 1997-01
```

5. Removing all records except those from Harrison Lake
   - #make a copy of original dataset so my mistakes don't affect original data
   - Harrison_Lake <- BullTroutRML2
   - #cheak the structure of dataset to be able to set filter
   - str(Harrison_Lake)
   - #set the filter
   - Harrison_Lake <- dplyr::filter(Harrison_Lake, lake == "Harrison")
6. Displaying the first and last 5 records from the filtered BullTroutRML2 dataset
   - #display first and last 5 lines
   - head(Harrison_Lake, n = 5L)
   - tail(Harrison_Lake, n = 5L)

```
  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

  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
```

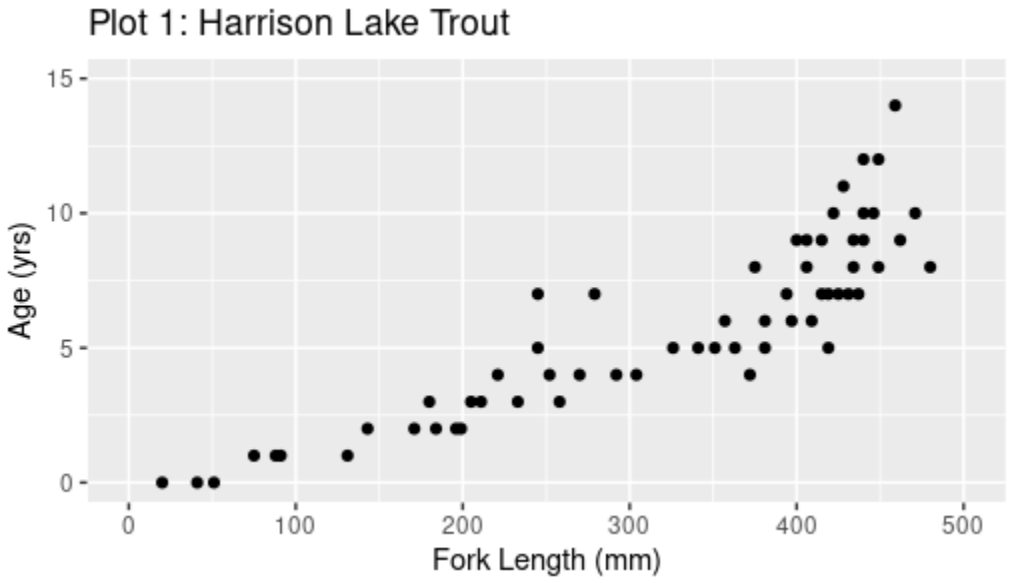
7. Displaying the structure of the filtered BullTroutRML2dataset
    - str(Harrison_Lake)

```
'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/ 2 levels "Harrison","Osprey": 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 ...
```

8. Display the summary of the filtered BullTroutRML2dataset
    - summary(Harrison_Lake)

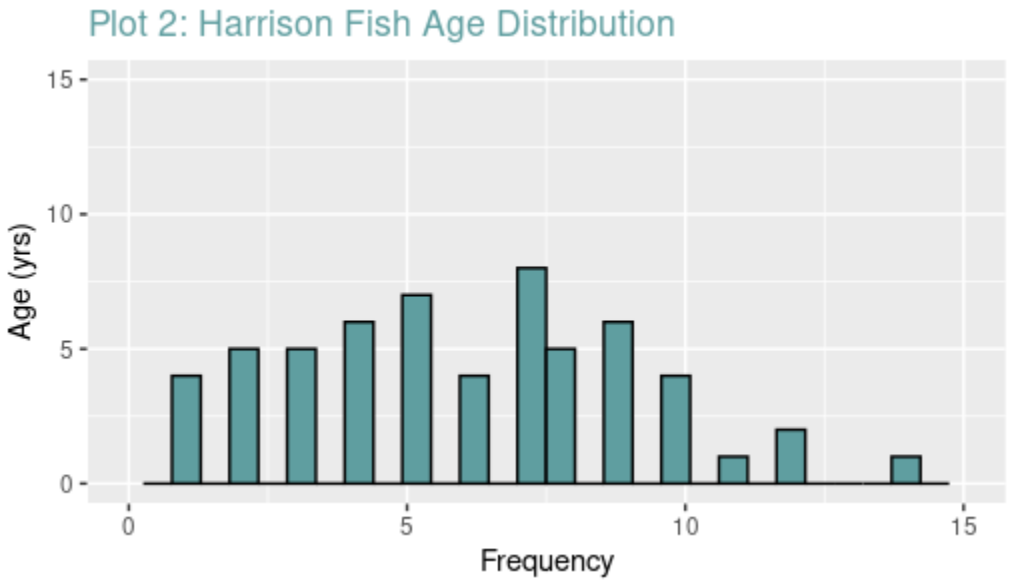
```
age              fl            lake         era
Min.   : 0.000   Min.   : 20   Harrison:61   1977-80:23
1st Qu.: 3.000   1st Qu.:221   Osprey  : 0   1997-01:38
Median : 6.000   Median :372
Mean   : 5.754   Mean   :319
3rd Qu.: 8.000   3rd Qu.:425
Max.   :14.000   Max.   :480
```

9. creating Plot 1: Harrison Lake Trout
    - ggplot(data = Harrison_Lake) +
    - geom_point(mapping = aes(x = fl, y = age))+
    - xlim(0, 500) + ylim(0,15)+
    - xlab("Fork Length (mm)")+ylab("Age (yrs)")+
    - ggtitle("Plot 1: Harrison Lake Trout")



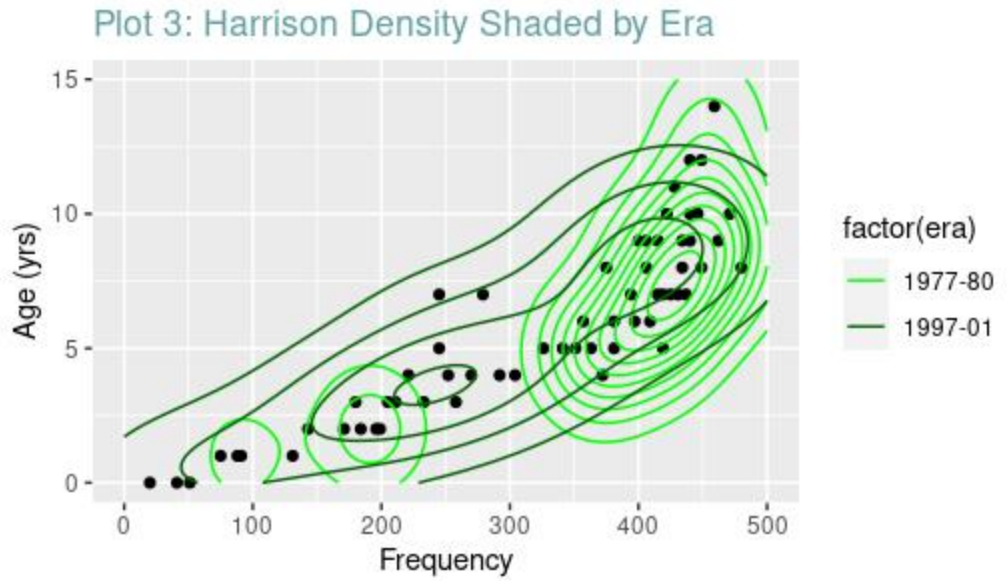
10. Plotting an “Age” histogram

- ggplot(data = Harrison_Lake, aes(x=age)) + geom_histogram(color="black", fill="cadetblue")+
- xlab("Frequency")+ylab("Age (yrs)")+
- labs(title="Plot 2: Harrison Fish Age Distribution")+
- xlim(0,15) + ylim(0,15)+
- theme(plot.title = element_text(color = "cadetblue"))



11. Creating an over dense plot using the same specifications as the previous scatterplot

- #Create 2d plot
-
- p <- ggplot(data = Harrison_Lake, aes(y=age,x=fl))+
- xlab("Frequency")+ylab("Age (yrs)")+
- labs(title="Plot 3: Harrison Density Shaded by Era")+
- theme(plot.title = element_text(color = "cadetblue"))+
- xlim(0,500) + ylim(0,15)+
- geom_point(colour = "black")
-
- # color 2d plot
- p +
- stat_density2d(aes(color = factor(era))) +
- scale_colour_manual(values = c("green", "darkgreen"))

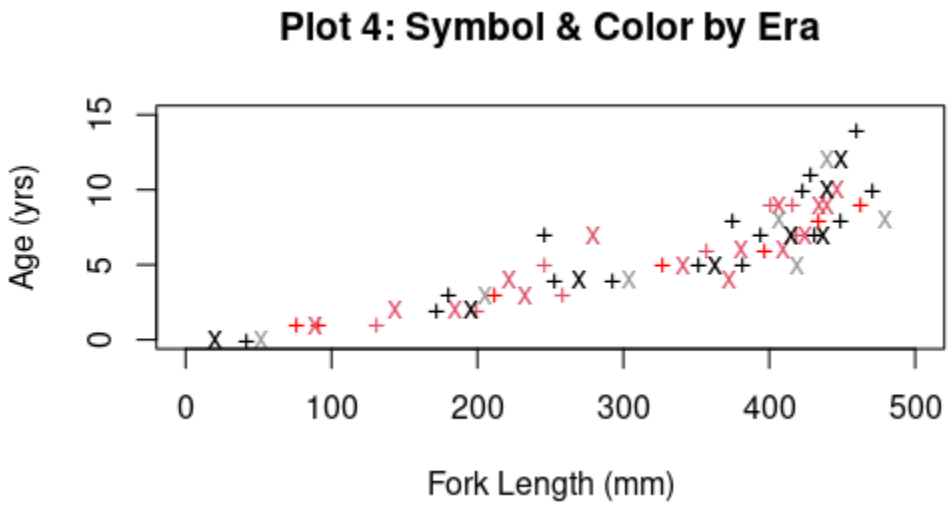


12. Creating a new object called “tmp” that includes the first 3 and last 3 records of the BullTroutRML2 data set
    - tmp <- head(Harrison_Lake, n = 3L)
    - tmp <- rbind(tmp,tail(Harrison_Lake, n = 3L))
13. Display the “era” column (variable) in the new “tmp” object
    - era <- subset(tmp,select=c("era"))
    - print(era)

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

14. Creating a pchs vector with the argument values for + and x.
    - pchs <- c("+","x")
15. Creating a cols vector with the two elements “red” and “gray60”
    - cols <- c("red","gray60")
16. Convert the tmp era values to numeric values.
    - tmp$era <- as.numeric(as.factor(tmp$era))
17. Initialize the cols vector with the tmp era values
    - cols <- append(tmp$era,cols)
18. creating Plot 4: Symbol & Color by Era
    - plot(y=Harrison_Lake$age,x=Harrison_Lake$fl,main="Plot 4: Symbol & Color by Era",

- ylab="Age (yrs)",xlab="Fork Length (mm)",
- ylim=c(0,15),xlim=c(0,500),
- pch=pchs,col=cols)



19. Plot 5: Regression Overlay

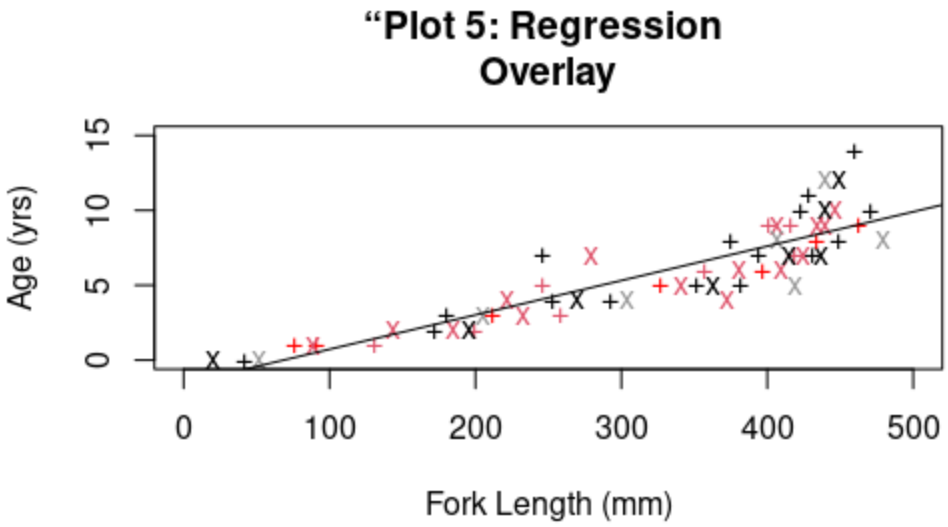
- plot(y=Harrison_Lake$age,x=Harrison_Lake$fl,main=""“Plot 5: Regression

,"Overlay

,"ylab="Age (yrs)",xlab="Fork Length (mm)

,(･’ᐞ･･)ylim=c(0,15),xlim=c

pch=pchs,col=cols)+abline(lm(Harrison_Lake$age ~ Harrison_Lake$fl))



20. Place a legend of on Plot 5 and call the new graph “Plot 6: :Legend Overlay”

,"- plot(y=Harrison_Lake$age,x=Harrison_Lake$fl,main="“Plot 6: :Legend Overlay

,"ylab="Age (yrs)",xlab="Fork Length (mm)

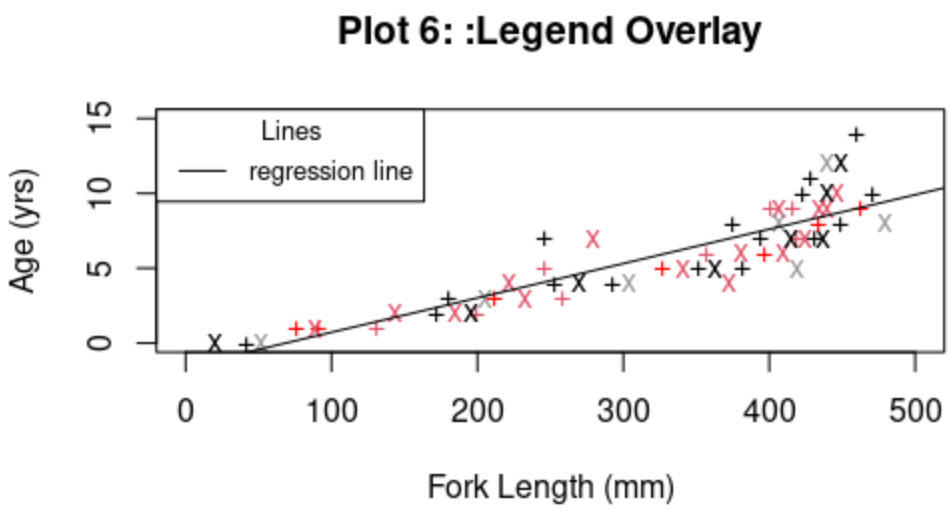
,(･’ᴖ･･)ylim=c(0,15),xlim=c

+((pch=pchs,col=cols)+abline(lm(Harrison_Lake$age ~ Harrison_Lake$fl

,("legend("topleft", legend=c("regression line

,col=c("black"), lty=1:2, cex=0.8

title="Lines", text.font=1, bg="white")


Plot 6: :Legend Overlay
Lines
regression line
Age (yrs)
0
5
10
15
0
100
200
300
400
500
Fork Length (mm)

Bibliography

Add legends to plots in R software : the easiest way! - Easy Guides - Wiki - STHDA. (2020). Sthda.com. http://www.sthda.com/english/wiki/add-legends-to-plots-in-r-software-the-easiest-way

guest contributer. (2015, April 10). Linear Models in R: Plotting Regression Lines - The Analysis Factor. The Analysis Factor. https://www.theanalysisfactor.com/linear-models-r-plotting-/regression-lines

Krunal Lathiya. (2021, December 23). pch in R: How to Use Plot Character in R. R-Lang; Krunal /Lathiya. https://r-lang.com/pch-in-r

Plot ranges of data in R - Burns Statistics. (2013, February 21). Burns Statistics. /https://www.burns-stat.com/plot-ranges-of-data-in-r

Data Mentor. (2017, November 23). DataMentor. https://www.datamentor.io/r-/programming/plot-function

DataScience. (2015). How to create Vectors, Factors, Lists, Matrices and Datasets with R Programming | DataScience+. Datascienceplus.com. https://datascienceplus.com/how-to-/create-data-frames-in-r

Convert Data Frame Column to Numeric in R (2 Example Codes). (2022, January 10). Statistics Globe. https://statisticsglobe.com/convert-data-frame-column-to-numeric-in-r

ggplot2 - Title and Subtitle with Different Size and Color in R - GeeksforGeeks. (2021, May 13). GeeksforGeeks. https://www.geeksforgeeks.org/ggplot2-title-and-subtitle-with-different-size-/and-color-in-r

How to install packages on R + screenshots | R-bloggers. (2013, January 31). R-Bloggers. /https://www.r-bloggers.com/2013/01/how-to-install-packages-on-r-screenshots

r print specific columns Code Example. (2020). Codegrepper.com. https://www.codegrepper.com/code-examples/r/r+print+specific+columns

Krunal Lathiya. (2021, September 19). How to Append Data Frames in R. R-Lang; Krunal Lathiya. /https://r-lang.com/how-to-append-data-frames-in-r

ggplot2 histogram plot : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA. (2020). Sthda.com. http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization

ggplot2 title : main, axis and legend titles - Easy Guides - Wiki - STHDA. (2020). Sthda.com. http://www.sthda.com/english/wiki/ggplot2-title-main-axis-and-legend-titles

Set scale limits — lims. (2022). Tidyverse.org. https://ggplot2.tidyverse.org/reference/lims.html

Grolemund, G. (2022). 3 Data visualisation | R for Data Science. Had.co.nz. https://r4ds.had.co.nz/data-visualisation.html#statistical-transformations

Data Carpentry contributors. (2022). Data visualization with ggplot2. Datacarpentry.org. https://datacarpentry.org/R-ecology-lesson/04-visualization-ggplot2.html

Select first or last rows of a data frame | R-bloggers. (2020, June 12). R-Bloggers. https://www.r-/bloggers.com/2020/06/select-first-or-last-rows-of-a-data-frame

head function - RDocumentation. (2019). Rdocumentation.org. https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/head

Create your own discrete scale — scale_manual. (2022). Tidyverse.org. https://ggplot2.tidyverse.org/reference/scale_manual.html

Chang, W. (2022, January 28). 6.12 Making a Density Plot of Two-Dimensional Data | R Graphics Cookbook, 2nd edition. R-Graphics.org. https://r-graphics.org/recipe-distribution-density2d

Awesome List Of 657 R Color Names You Need to Know - Datanovia. (2018, November 17). /Datanovia. https://www.datanovia.com/en/blog/awesome-list-of-657-r-color-names

momova97. (2022, January 29). momova97/ALY6000_movahedi. GitHub. https://github.com/momova97/ALY6000_movahedi

Appendix

```
Print my name at the top of the script#

print("Plotting Basics:Movahedi")

Importing required libraries#


install.packages("FSA")

install.packages("FSAdata")

install.packages("magrittr")

install.packages("dplyr")

install.packages("plotrix")

install.packages("ggplot2")

install.packages("moments")

loading installed libraries#

library(FSA)

library(FSAdata)

library(magrittr)

library(dplyr)

library(plotrix)

library(ggplot2)

library(moments)

loading data set#

data(BullTroutRML2)

showing first three lines#
```

```
print(head(BullTroutRML2, n = 3L))

showing last three lines#

print(tail(BullTroutRML2, n = 3L))

make a copy of original dataset so my mistakes don't affect original data#

Harrison_Lake <- BullTroutRML2

cheak the structure of dataset to be able to set filter#

str(Harrison_Lake)

set the filter#

Harrison_Lake <- dplyr::filter(Harrison_Lake, lake == "Harrison")

display first and last 5 lines#

head(Harrison_Lake, n = 5L)

tail(Harrison_Lake, n = 5L)

displaying structure of the filtered BullTroutRML2 data set#

str(Harrison_Lake)

Display the summary of the filtered BullTroutRML2 data set #

summary(Harrison_Lake)



creating scatter plot#

determining limits#

determining labels#

+ ggplot(data = Harrison_Lake)

+geom_point(mapping = aes(x = fl, y = age))

+(·’১৫)xlim(0, 500) + ylim
```

```
+xlab("Fork Length (mm)")+ylab("Age (yrs)")

ggtitle("Plot 1: Harrison Lake Trout")

ploting histogram for age#

+ggplot(data = Harrison_Lake, aes(x=age)) + geom_histogram(color="black", fill="cadetblue")

+xlab("Frequency")+ylab("Age (yrs)")

+labs(title="Plot 2: Harrison Fish Age Distribution")

+(٠’١٥)xlim(0,15) + ylim

theme(plot.title = element_text( color = "cadetblue"))

Create 2d plot#


+p <- ggplot(data = Harrison_Lake, aes(y=age,x=fl))

+xlab("Frequency")+ylab("Age (yrs)")

+labs(title="Plot 3: Harrison Density Shaded by Era")

+theme(plot.title = element_text( color = "cadetblue"))

+(٠’١٥)xlim(0,500) + ylim

geom_point(colour = "black")


color 2d plot #

+ p

+ stat_density2d(aes(color = factor(era)))

scale_colour_manual(values = c("green", "darkgreen"))


creating tmp object containing first and last 3 lines of BullTroutRML2#

tmp <- head(Harrison_Lake, n = 3L)
```

```
tmp <- rbind(tmp,tail(Harrison_Lake, n = 3L))

print(tmp)

display era only#

era <- subset(tmp,select=c("era"))

print(era)

creating pchs vector #

pchs <- c("+","x")

creating cols vector #

cols <- c("red","gray60")

turn era to numeric factor#

tmp$era <- as.numeric(as.factor(tmp$era))

Initialize the cols vector with the tmp era values #

cols <- append(tmp$era,cols)

create the + and x plot#

,"plot(y=Harrison_Lake$age,x=Harrison_Lake$fl,main="Plot 4: Symbol & Color by Era

,"ylab="Age (yrs)",xlab="Fork Length (mm)

,(・’۵・・)ylim=c(0,15),xlim=c

(pch=pchs,col=cols

create regression line plot 5#

plot(y=Harrison_Lake$age,x=Harrison_Lake$fl,main="“Plot 5: Regression

,"Overlay

,"ylab="Age (yrs)",xlab="Fork Length (mm)

,(・’۵・・)ylim=c(0,15),xlim=c

((pch=pchs,col=cols)+abline(lm(Harrison_Lake$age ~ Harrison_Lake$fl
```

create plot 6#

,"plot(y=Harrison_Lake$age,x=Harrison_Lake$fl,main="Plot 6: :Legend Overlay

,"ylab="Age (yrs)",xlab="Fork Length (mm)

,(·’ᔆ··)ylim=c(0,15),xlim=c

+((pch=pchs,col=cols)+abline(lm(Harrison_Lake$age ~ Harrison_Lake$fl

,("legend("topleft", legend=c("regression line

,col=c("black"), lty=1:2, cex=0.8

("title="Lines", text.font=1, bg="white