Executive Summary Report 2

Tianyu Zhang

CPS: Project Management, Northeastern University

ALY 6000: Introduction to Analytics

Professor: Richard He

January 23, 2023

Introduction

This project aims to analyze the BullTroutRML2 dataset using statistical techniques to understand the patterns and relationships present in the data. The BullTroutRML2 dataset is a collection of data collected from various sources, including Harrison and Osprey. It includes information about the effects of different ages, locations, and eras on fork lengths. The goal of this analysis is to uncover any trends or patterns in the data that can provide insight into the behavior of the variable in question. The analysis will begin with a thorough exploration of the data, including visualizations and summary statistics. Next, we will apply various statistical techniques to identify patterns and relationships in the data. Finally, we will use the insights gained from the analysis to make predictions and draw conclusions about the behavior of the variable in the BullTroutRML2 dataset.

Key findings based on instruction

```
print(head(BullTroutRML2, 5))
  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
print(tail(BullTroutRML2, 5))
   age  fl   lake     era
92   5 289 Osprey 1997-01
93   4 313 Osprey 1997-01
94   4 298 Osprey 1997-01
95   3 279 Osprey 1997-01
96   3 273 Osprey 1997-01
```

From the first data set, we can see that it is a data set of fish from Harrison Lake during the time period 1977-1980. It includes information on the age and fork length (fl) of the fish as well as the era in which they were caught. The data shows that the fish range in age from 9 to 14 years old and in fork length from 400 to 471 mm.

From the second dataset, we can see that it is a data set of fish from Osprey Lake during the time period 1997-2001. It includes information on the age and fork length of the fish as well as the era in which they were caught. The data shows that the fish range in age from 3 to 5 years old and in fork length from 273 to 313 mm.

```
print(head(BullTroutRML2_filtered, 4))
  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
```

```
print(tail(BullTroutRML2_filtered, 4))
   age  fl     lake     era
58   0  20 Harrison 1997-01
59   7 245 Harrison 1997-01
60   7 279 Harrison 1997-01
61   5 245 Harrison 1997-01
```

These two datasets show the first and last 4 records of the filtered BullTroutRML2 dataset, which contains only the data from Harrison Lake. From the first 4 records, we can see that the fish have different ages and fork lengths, and the era is "1977-80". From the last 4 records, we can see that the fish have different ages and fork lengths, and the era is "1997-

01". We can also see that the minimum age of the fish is 0 and the maximum age is 14. The fork lengths of the fish range from 20mm to 471mm. These outputs give us an idea of the general range of ages and fork lengths of the fish in Harrison Lake and the time period in which the data was collected.

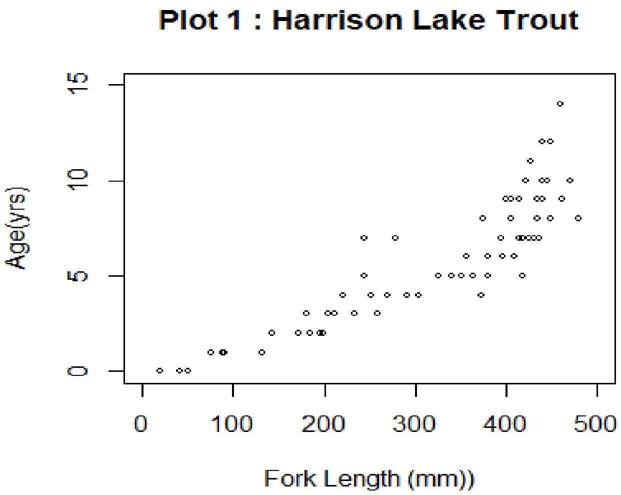
```
#Display the structure of the filtered BullTroutRML2 dataset
> str(BullTroutRML2_filtered)
'data.frame':	61 obs. of  4 variables:
 $ age : int  14 12 10 10 9 9 9 8 8 7 ...
 $ fl  : int  459 449 471 446 400 440 462 480 449 437 ...
 $ lake: Factor w/ 2 levels "Harrison","Osprey": 1 1 1 1 1 1 1 1 1 1 ...
 $ era : Factor w/ 2 levels "1977-80","1997-01": 1 1 1 1 1 1 1 1 1 1 ...
```

The str() function provides the structure of the filtered BullTroutRML2 dataset, which includes the data types of the variables (age and fl are integers, lake and era are factors)

```
#Display the summary of the filtered BullTroutRML2 dataset
> summary(BullTroutRML2_filtered)
      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                             
```

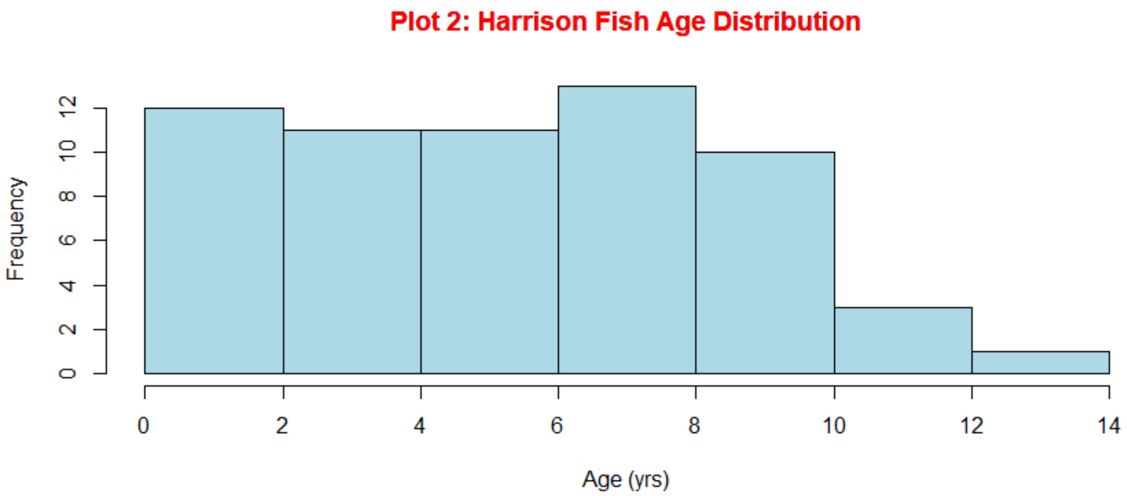
The summary() function provides a summary of the dataset, including the mean, median, min, and max. The summary can provide the general information of the dataset, for example, the average age of fish is around 5.754 years old and the fork length is around 319mm.

Create a scatterplot for “Age (yrs)” (y variable) and “Fork Length (mm)” (x variable) (Gangwar, 2022).



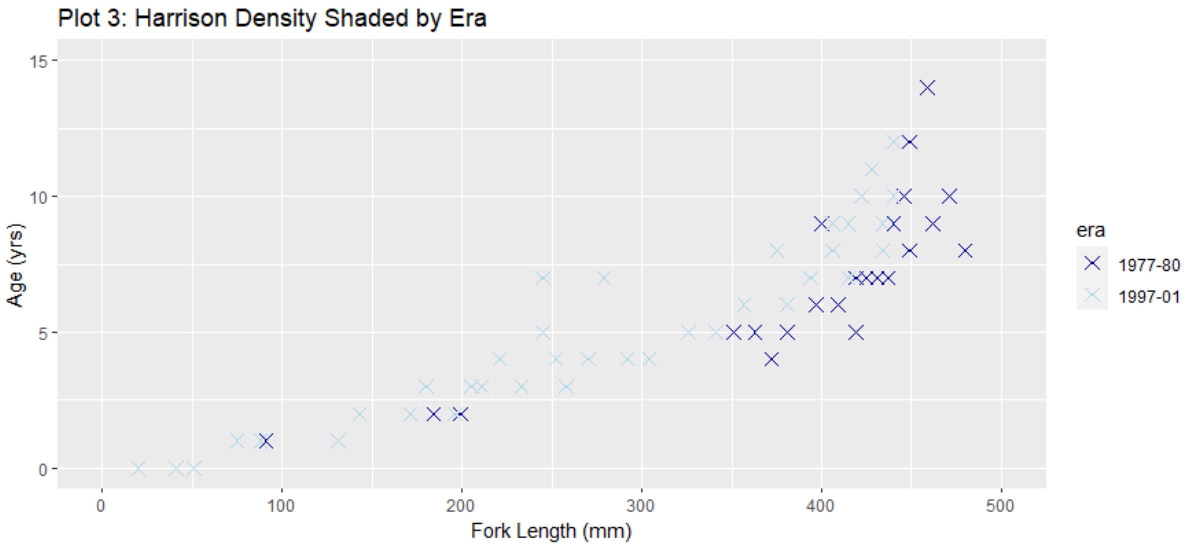
From this graph, we can see that there is a positive correlation between age and fork length for the Harrison Lake trout, as the points on the graph tend to slope upward from left to right. Additionally, we can see that the majority of the trout fall within the range of fork length between 0 and 500mm and the age range between 0 and 15 years. We can also see that there are some outliers, with a few trout having much larger fork lengths than the majority of the trout.

Plot an "Age" histogram with the following specifications (DataMentor, 2018).



This histogram shows the distribution of ages for fish caught in Harrison Lake. The x-axis represents the age of the fish in years and the y-axis represents the frequency of fish caught at that age. The histogram is colored in light blue and the title is in red. From this graph, we can see that the majority of the fish caught in Harrison Lake are between 0 and 10 years old. We can also see that the distribution appears to be roughly symmetric. Additionally, the title of the graph is "Plot 2: Harrison Fish Age Distribution" which tells us that the data being presented is specifically for fish caught in Harrison Lake.

Create a plot using the same specifications as the previous scatterplot (Tidyverse, 2023).



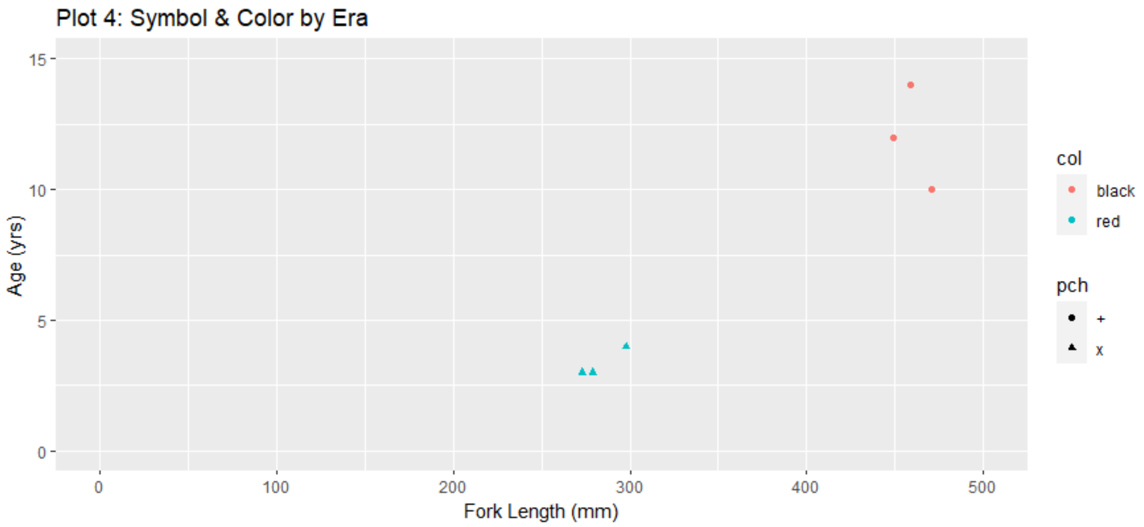
This graph is a scatter plot that shows the relationship between the age and fork length of Harrison Lake trout. The color of the data points indicates the era in which the fish were caught, with dark blue points representing 1977-80 and light blue points representing 1997-01. From this graph, we can see that there are more data points representing fish caught in the later era and that the distribution of ages and fork lengths for fish caught in both eras overlap. Additionally, we can see that there appears to be a positive correlation between age and fork length, with older fish generally having longer fork lengths.

Create a new object called “tmp” that includes the first 3 and last 3 records of the BullTroutRML2 dataset.

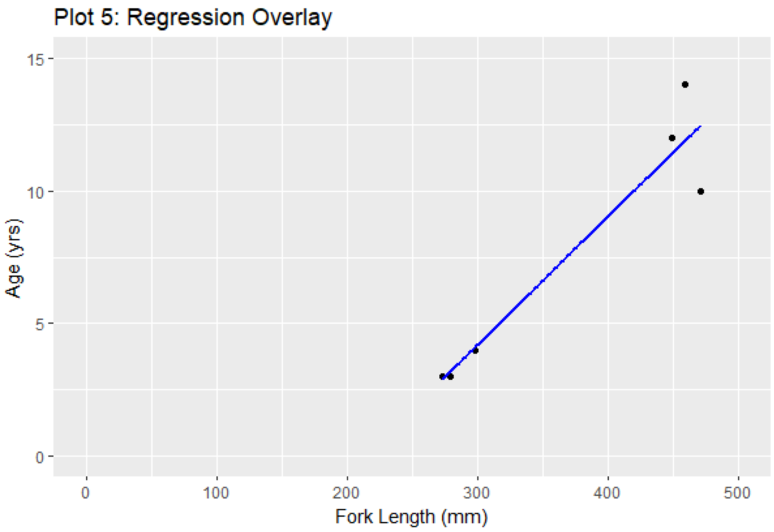
```
     age  fl     lake    era pch   col
1    14 459 Harrison   1   +   black
2    12 449 Harrison   1   +   black
3    10 471 Harrison   1   +   black
94   4 298   Osprey   2   x   red
95   3 279   Osprey   2   x   red
96   3 273   Osprey   2   x   red
```

```
tmp$era
[1] 1977-80 1977-80 1977-80 1997-01 1997-01 1997-01
Levels: 1977-80 1997-01
```

This indicates that the data in the "tmp" object includes three records from the "1977-80" era and three records from the "1997-01" era. It also shows that the "era" variable is a factor variable with two levels, "1977-80" and "1997-01".

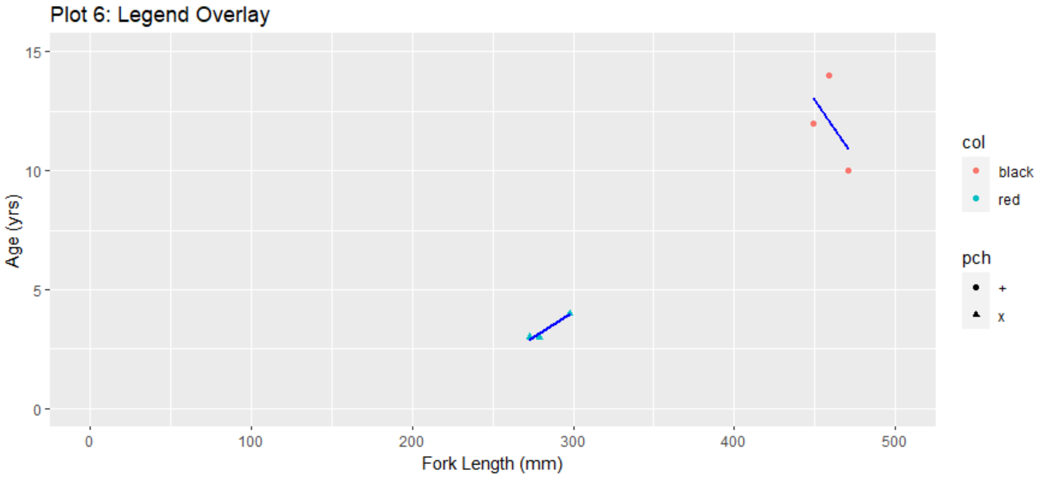


This graph is a scatter plot that shows the relationship between the fork length (x-axis) and age (y-axis) of fish in the tmp dataset. The points on the graph are represented by different symbols (pch) and colors (col), which correspond to the era of the fish. The x-axis has a limit of 0 to 500 and the y-axis has a limit of 0 to 15. The title of the graph is "Plot 4: Symbol & Color by Era" and the x-axis is labeled "Fork Length (mm)" and the y-axis is labeled "Age (yrs)". From this graph, we can see that different eras of fish may have different fork lengths and ages.



Plot 5 is a scatter plot with a regression line overlay (Porras, 2018). The data points are represented by small solid circles. The x-axis represents the fork length of the fish in millimeters, and the y-axis represents the age of the fish in years. The x-axis is limited between 0 and 500 millimeters and the y-axis is limited between 0 and 15 years. From this graph, we can see a positive correlation between age and fork length. The regression line shows the overall trend of the data, indicating that as the fork length increases, the age of the fish also increases.

Place a legend on Plot 5 and call the new graph "Plot 6: Legend Overlay (GreeksforGreeks, 2021).



The Legend Overlay graph shows the relationship between the fork length and the age of the fish. The addition of the legend on Plot 6 helps to indicate which color and shape correspond to which era of fish.

Bibliography

Gangwar, M. (2022, August 3). *Understanding plot() Function in R - Basics of Graph Plotting*. DigitalOcean. Retrieved January 23, 2023, from https://www.digitalocean.com/community/tutorials/plot-function-in-r

GeeksforGeeks. (2021, May 16). *Add legend to plot in R*. GeeksforGeeks. Retrieved January 23, 2023, from https://www.geeksforgeeks.org/add-legend-to-plot-in-r/

Porras, E. M. (2018, July 18). *R linear regression tutorial: LM function in R with code examples*. DataCamp. Retrieved January 23, 2023, from https://www.datacamp.com/tutorial/linear-regression-R

R hist() to create histograms (with numerous examples). DataMentor. (2018, April 11). Retrieved January 23, 2023, from https://www.datamentor.io/r-programming/histogram/

Tidyverse. (2023, January 12). *GGPLOT2/plot.r at 2c5a78cbf1cbba8dac090dff4d7de467df6cb22b · Tidyverse/GGPLOT2*. GitHub. Retrieved January 23, 2023, from https://github.com/tidyverse/ggplot2/blob/HEAD/R/plot.r

Appendix

```
#Print your name at the top of the script. Include the prefix: “Plotting Basics:” such that it
#appears “Plotting Basics: Firstname Lastname ”
print("Plotting Basics : Tianyu Zhang")
#Import libraries including: FSA, FSAdata, magrittr, dplyr, plotrix, ggplot2, and moments
#NOTE: You must use R version 3.6.3 to gain access to the FSA data set. If you installed a
#later version of R, you must uninstall Rstudio and R. Then reinstall R version 3.6.3; then
#reinstall Rstudio
install.packages("FSA")
install.packages("FSAdata")
install.packages("magrittr")
install.packages("dplyr")
install.packages("plotrix")
install.packages("ggplot2")
install.packages("moments")
library(FSA)
library(FSAdata)
library(magrittr)
library(dplyr)
library(plotrix)
library(ggplot2)
library(moments)
#Load the BullTroutRML2 dataset (BullTroutRML2.csv)
data(BullTroutRML2)
# Print the first 5 records
print(head(BullTroutRML2, 5))

# Print the last 5 records
print(tail(BullTroutRML2, 5))
```

```
#Remove all records except those from Harrison Lake (hint: use dplyr::filter() function)
#NOTE: From this point forward any reference to BullTroutRML2 always refers to the
#filtered dataset (Harrison Lake only data is used). You may choose to rename the
#dataset at this point.
BullTroutRML2_filtered <- dplyr::filter(BullTroutRML2, lake == "Harrison")

#Display the first 4 and last 4 records from the filtered BullTroutRML2 dataset
# Display the first 4 records
print(head(BullTroutRML2_filtered, 4))

# Display the last 4 records
print(tail(BullTroutRML2_filtered, 4))

#Display the structure of the filtered BullTroutRML2 dataset
str(BullTroutRML2_filtered)
#Display the summary of the filtered BullTroutRML2 dataset
summary(BullTroutRML2_filtered)

#Create a scatterplot for “Age (yrs)” (y variable) and “Fork Length (mm)” (x variable)
#with the following specifications:
#• Limit of x axis is (0,500)
#• Limit of y axis is (0,15)
#• Title of graph is “Plot 1: Harrison Lake Trout”
#• Y axis label is “Age (yrs)”
#• X axis label is “Fork Length (mm)”
#• Use small solid circles for the plotted data points
plot(BullTroutRML2_filtered$fl, BullTroutRML2_filtered$age,
    xlim = c(0,500), ylim = c(0,15),
    main = "Plot 1 : Harrison Lake Trout",
    xlab = "Fork Length (mm))", ylab = "Age(yrs)",
    pch = 21, cex = 0.5)
#Plot an “Age” histogram with the following specifications
#• Y axis label is “Frequency”
#• X axis label is “Age (yrs)”
#• Title of the histogram is “Plot 2: Harrison Fish Age Distribution”
#• The color of the frequency plots is “lightblue”
#• The color of the Title is “red”



hist(BullTroutRML2_filtered$age, xlab = "Age (yrs)", ylab = "Frequency",
    main = "Plot 2: Harrison Fish Age Distribution",
    col = "lightblue")
title(main = "Plot 2: Harrison Fish Age Distribution", col.main = "red")

#Create a plot using the same specifications as the previous scatterplot. But,
#• Title the plot “Plot 3: Harrison Density Shaded by Era”
#• Y axis label is “Age (yrs)”
#• Y axis limits are 0 to 15
#• X axis label is “Fork Length (mm)”
#• X axis limits are 0 to 500
#• include two levels of shading of blue for the data points based on era values.
#• Plot solid diamonds as data points
ggplot(BullTroutRML2_filtered, aes(x = fl, y = age, color = era)) +
  geom_point(shape = 4, size = 3) +
  ggtitle("Plot 3: Harrison Density Shaded by Era")+
  xlab("Fork Length (mm)") +
  ylab("Age (yrs)") +
  scale_color_manual(values = c("darkblue","lightblue")) +
  xlim(0,500) + ylim(0,15)
```

```
#Create a new object called “tmp” that includes the first 3 and last 3 records of the
#BullTroutRML2 dataset.

head(BullTroutRML2, 3)
tail(BullTroutRML2, 3)
tmp <- rbind(head(BullTroutRML2, 3), tail(BullTroutRML2, 3))

#Display the “era” column (variable) in the new “tmp” object

tmp$era
#Create a pchs vector with the argument values for + and x.
#Create a cols vector with the two elements “black” and “red”
pchs <- c("+", "x")
cols <- c("black", "red")

# Convert the tmp era values to numeric values.
# Initialize the pchs and cols vector conditional on the tmp era values

tmp$era <- as.numeric(tmp$era)
tmp$pch <- pchs[tmp$era]
tmp$col <- cols[tmp$era]

#Create a plot of “Age (yrs)” (y variable) versus “Fork Length (mm)” (x variable) with the
#following specifications:
 # • Title of graph is “Plot 4: Symbol & Color by Era”
#• Limit of x axis is (0,500)
#• Limit of y axis is (0,15)
#• Y axis label is “Age (yrs)”
#• X axis label is “Fork Length (mm)”
#• Set pch equal to pchs era values
#• Set col equal to cols era values
ggplot(tmp, aes(x = fl, y = age, pch = pch, col = col)) +
  geom_point() +
  xlim(0, 500) +
  ylim(0, 15) +
  xlab("Fork Length (mm)") +
  ylab("Age (yrs)") +
  ggtitle("Plot 4: Symbol & Color by Era")

#Plot a regression line (blue color) overlay on Plot 4 and title the new graph “Plot 5:
#Regression Overlay”.
ggplot(tmp, aes(x = fl, y = age)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  xlim(0, 500) +
  ylim(0, 15) +
  xlab("Fork Length (mm)") +
  ylab("Age (yrs)") +
  ggtitle("Plot 5: Regression Overlay")

#Place a legend of on Plot 5 and call the new graph “Plot 6: Legend Overlay”
ggplot(tmp, aes(x = fl, y = age, pch = pch, col = col)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  xlim(0, 500) +
  ylim(0, 15) +
  xlab("Fork Length (mm)") +
  ylab("Age (yrs)") +
ggtitle("Plot 6: Legend Overlay")
```