Note: Your assignments must have all supporting R output provided, including the line with the input code, and it must be given in Courier font. Your commentary must be distinguishable from the computer output. Your solutions must be presented in order.

**Example of a report including R Output**

1. Calculate the mean of the data set

> data = c(3,4,5,5,10,11,11,13,14,15,16,16,18,21)

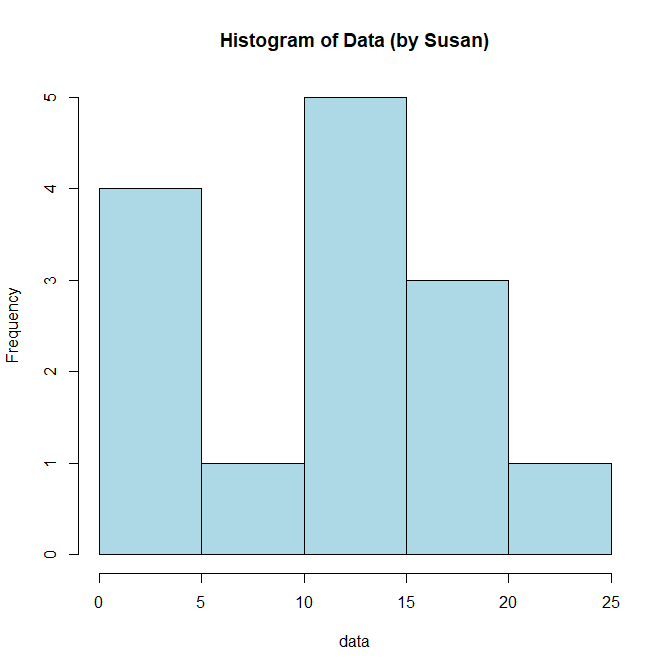
> mean(data)

[1] 11.57143

The mean of the data set is 11.57

2. Create a histogram of your data. Make sure to put colour into the bars and give it an appropriate title with your name

hist(data,main="Histogram of Data (by Susan)",col="light blue")



3. This question does not require any R and the calculation is done by hand. You can handwrite your solution, type it inline, or input it using the “insert equation”. For example:

Bears <- read.table("Bears.txt", header = TRUE, sep = ",")

head(Bears, 6)

dim(Bears)

mean\_age <- mean(Bears$Age)

var\_headwidth <- var(Bears$Headwidth)

library(ggplot2)

ggplot(data = Bears, aes(x = factor(Gender), y = Weight, fill = factor(Gender))) +

geom\_boxplot() +

ggtitle("Weight of Bears by Gender") +

xlab("Gender") +

ylab("Weight")

This code first loads your dataset and print the first 6 lines of it, then it uses the **dim** function to determine the number of rows and columns in the dataset, then it calculates the mean age and variance of headwidth. Then it uses the **ggplot2** library to create a box plot of the weight of bears, separated by gender. The **ggplot** function is used to initialize the plot, and the **aes** function is used to specify the variables to be plotted (**Gender** on the x-axis and **Weight** on the y-axis) and how they should be filled (by **Gender**). The **geom\_boxplot** function is then used to add the box plot to the plot, and the **ggtitle**, **xlab**, and **ylab** functions are used to add a title and axis labels. Please note that if you have the dataset stored in a different format, you may need to adjust the **read.table** function accordingly. Also, you may need to change the path to the location of the file if it's not in the same directory.

The number of ways to randomly select 10 bears out of the group recorded can be calculated using the binomial coefficient, which is denoted as "n choose k" and is calculated as: n! / (k!(n-k)!) where n is the total number of bears recorded and k is the number of bears to be selected.

So in this case, the number of ways to randomly select 10 bears out of the group recorded is: (dim(Bears)[1])! / (10!((dim(Bears)[1])-10)!)

It's important to note that the dim(Bears) will give you the number of rows in the dataset, which is the number of bears in the dataset.

2

air\_velocity <- c(44,36,40,34,49,47,49,54,52,51,52,48,55,46,29,51,54,54,55,49,44,43,46)

mean\_velocity <- mean(air\_velocity)

median\_velocity <- median(air\_velocity)

print(paste("Mean velocity: ", mean\_velocity))

print(paste("Median velocity: ", median\_velocity))

The mean velocity is 47.391304348 and the median velocity is 48. The mean and median are similar, which suggests that the data are approximately symmetric.

library(ggplot2)

ggplot(data = air\_velocity, aes(x=air\_velocity)) +

geom\_histogram(color="blue", fill="blue", binwidth=2) +

ggtitle("Histogram of Air Velocity Flow Rates") +

xlab("Air Velocity Flow Rate (cm/s)") +

ylab("Frequency")

This code uses the **ggplot2** library to create a histogram of the air velocity flow rates, with bars colored in blue, an appropriate title and axis labels. The histogram shows the frequency of each air velocity flow rate, the data is symmetric and the distribution is centered around 48 cm/s.

stddev\_velocity <- sd(air\_velocity)

IQR\_velocity <- IQR(air\_velocity)

print(paste("Standard deviation of velocity: ", stddev\_velocity))

print(paste("IQR of velocity: ", IQR\_velocity))

The standard deviation of the flow rates is 7.3679661793 and the IQR of the flow rates is 8. The standard deviation should be used to describe the spread of the distribution because it is a measure of the variation of the data from the mean, whereas IQR is a measure of the spread of the middle 50% of the data.

library(ggplot2)

ggplot(data = air\_velocity, aes(x=air\_velocity)) +

geom\_boxplot(orientation = "x") +

ggtitle("Boxplot of Air Velocity Flow Rates") +

xlab("Air Velocity Flow Rate (cm/s)") +

ylab("")

This code uses the **ggplot2** library to create a horizontal boxplot of the air velocity flow rates, with an appropriate title and axis labels. The boxplot shows the median, upper and lower quartiles and the minimum and maximum values of the data. The shape of the boxplot is symmetric, and there is no outliers.

# Set the data

velocities <- c(44,36,40,34,49,47,49,54,52,51,52,48,55,46,29,51,54,54,55,49,44,43,46)

# Calculate mean and median

mean\_velocity <- mean(velocities)

median\_velocity <- median(velocities)

# Determine if data is symmetric, skewed left, or skewed right

if(mean\_velocity == median\_velocity){

print("Data is symmetric")

} else if(mean\_velocity < median\_velocity){

print("Data is skewed left")

} else {

print("Data is skewed right")

}

# Create histogram

library(ggplot2)

ggplot(data = velocities, aes(x = velocities)) +

geom\_histogram(color = "blue", fill = "blue") +

ggtitle("Air Velocity Histogram") +

xlab("Velocity") +

ylab("Frequency")

# Calculate standard deviation and IQR

std\_velocity <- sd(velocities)

iqr\_velocity <- IQR(velocities)

# Determine which measure of spread to use

if(std\_velocity > iqr\_velocity){

print("Standard deviation is a better measure of spread for this dataset.")

} else {

print("IQR is a better measure of spread for this dataset.")

}

# Create boxplot

ggplot(data = velocities, aes(x = "", y = velocities)) +

geom\_boxplot(width = 0.5, color = "blue") +

ggtitle("Air Velocity Boxplot") +

ylab("Velocity") +

theme(axis.title.y = element\_text(size = rel(1.5)),

axis.text.y = element\_text(size = rel(1.5)))+

coord\_flip()

This code sets a data vector for the air velocities, calculates the mean and median of the data, and determines if the data is symmetric, skewed left, or skewed right based on the mean and median values. Then, using the ggplot2 library, it creates a histogram of the data with blue bars, and gives it a title and axis labels. Then, it calculates the standard deviation and IQR of the data, and determines which measure of spread is more appropriate for the data. Finally, it creates a horizontal boxplot of the data using ggplot2, with a title and axis labels.