ITMD/ITMS/STAT 514 Homework 1

Guruteja Kanderi

1/30/2024

Part I. Changing the author field and file name. (8 points)

(a) Change the `author:` field on the Rmd document

(b) Rename this file to "HW1_YourFirstInitialYourLastName.Rmd", (e.g. HW1_YDing.Rmd).

Part II. Cars data (80 points)

We'll look at data frame and plotting in much more detail in later classes. For a previous of what's to come, here's a very basic example.

For this example we'll use a very simple dataset. The `cars` data comes with the default installation of R. To see the first few rows of the data, just type `head(cars)`.

```
head(cars)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

```
# Create a dataframe with the given data
cars_data <- data.frame(speed = c(4, 4, 7, 7, 8, 9),
                        dist = c(2, 10, 4, 22, 16, 10))

# Calculate minimum, maximum, mean, and median values of speed
min_speed <- min(cars_data$speed)
max_speed <- max(cars_data$speed)
mean_speed <- mean(cars_data$speed)
median_speed <- median(cars_data$speed)

# Calculate minimum, maximum, mean, and median values of dist
min_dist <- min(cars_data$dist)
max_dist <- max(cars_data$dist)
mean_dist <- mean(cars_data$dist)
median_dist <- median(cars_data$dist)

# Print the results
cat("Speed: Min =", min_speed, ", Max =", max_speed, ", Mean =", mean_speed, ", Median =", median_speed
```

(a) Do a summary for cars. What the min,max, mean and median values of speed and dist? (10 points)

```
## Speed: Min = 4 , Max = 9 , Mean = 6.5 , Median = 7
```

```
cat("Dist: Min =", min_dist, ", Max =", max_dist, ", Mean =", mean_dist, ", Median =", median_dist, "\n
```

```
## Dist: Min = 2 , Max = 22 , Mean = 10.66667 , Median = 10
```

Speed: Min = 4 , Max = 9 , Mean = 6.5 , Median = 7 Dist: Min = 2 , Max = 22 , Mean = 10.66667 , Median = 10

```
# Create a vector with the speed values
speed <- c(4, 4, 7, 7, 8, 9)

# Calculate the variance of speed
variance_speed <- var(speed)

# Calculate the standard deviation of speed
std_dev_speed <- sd(speed)

# Print the results
cat("Variance of speed:", variance_speed, "\n")
```

(b) Calculate the variance and standard deviation of speed (5 points)

```
## Variance of speed: 4.3
```

```
cat("Standard deviation of speed:", std_dev_speed, "\n")
```

```
## Standard deviation of speed: 2.073644
```

Variance of speed: 4.3 Standard deviation of speed: 2.073644

```
# # Create a vector with the dist values
dist <- c(2, 10, 4, 22, 16, 10)

# Calculate the variance of dist
variance_dist <- var(dist)

# Calculate the standard deviation of dist
std_dev_dist <- sd(dist)

# Print the results
cat("Variance of dist:", variance_dist, "\n")
```

(c) Calculate the variance and standard deviation of dist (5 points)

```
## Variance of dist: 55.46667
```

```
cat("Standard deviation of dist:", std_dev_dist, "\n")
```

```
## Standard deviation of dist: 7.447595
```

Variance of dist: 55.46667 Standard deviation of dist: 7.447595

```r
# Create a dataframe with the given data
cars_data <- data.frame(speed = c(4, 4, 7, 7, 8, 9),
                        dist = c(2, 10, 4, 22, 16, 10))

# Filter the data where speed is greater than 15
filtered_data <- subset(cars_data, speed > 15)

# Calculate the average and standard deviation of dist
average_dist <- mean(filtered_data$dist)
std_dev_dist <- sd(filtered_data$dist)

# Print the results
cat("Average dist when speed > 15:", average_dist, "\n")
```

(d) Calulate the average and standard deviation of `dist` when `speed` is greater than 15. (5 points)

```
## Average dist when speed > 15: NaN
```

```r
cat("Standard deviation of dist when speed > 15:", std_dev_dist, "\n")
```

```
## Standard deviation of dist when speed > 15: NA
```

Average dist when speed > 15: NaN Standard deviation of dist when speed > 15: NA

```r
# # Filter the data where speed is less than or equal to 15
filtered_data <- subset(cars_data, speed <= 15)

# Calculate the average and standard deviation of dist
average_dist <- mean(filtered_data$dist)
std_dev_dist <- sd(filtered_data$dist)

# Print the results
cat("Average dist when speed <= 15:", average_dist, "\n")
```

(e) Calulate the average and standard deviation of `dist` when `speed` is less than and equal to 15. (5 points)

```
## Average dist when speed <= 15: 10.66667
```

```r
cat("Standard deviation of dist when speed <= 15:", std_dev_dist, "\n")
```

```
## Standard deviation of dist when speed <= 15: 7.447595
```

Average dist when speed <= 15: 10.66667 Standard deviation of dist when speed <= 15: 7.447595

```r
# Filter the data where speed is greater than 12 and less than 19
filtered_data <- subset(cars_data, speed > 12 & speed < 19)

# Calculate the average and standard deviation of dist
average_dist <- mean(filtered_data$dist)
std_dev_dist <- sd(filtered_data$dist)
```

```r
# Print the results
cat("Average dist when speed > 12 and speed < 19:", average_dist, "\n")
```

(f) Calulate the average and standard deviation of `dist` when `speed` is greater than 12 and less than 19. (10 points)

```
## Average dist when speed > 12 and speed < 19: NaN
```

```r
cat("Standard deviation of dist when speed > 12 and speed < 19:", std_dev_dist, "\n")
```

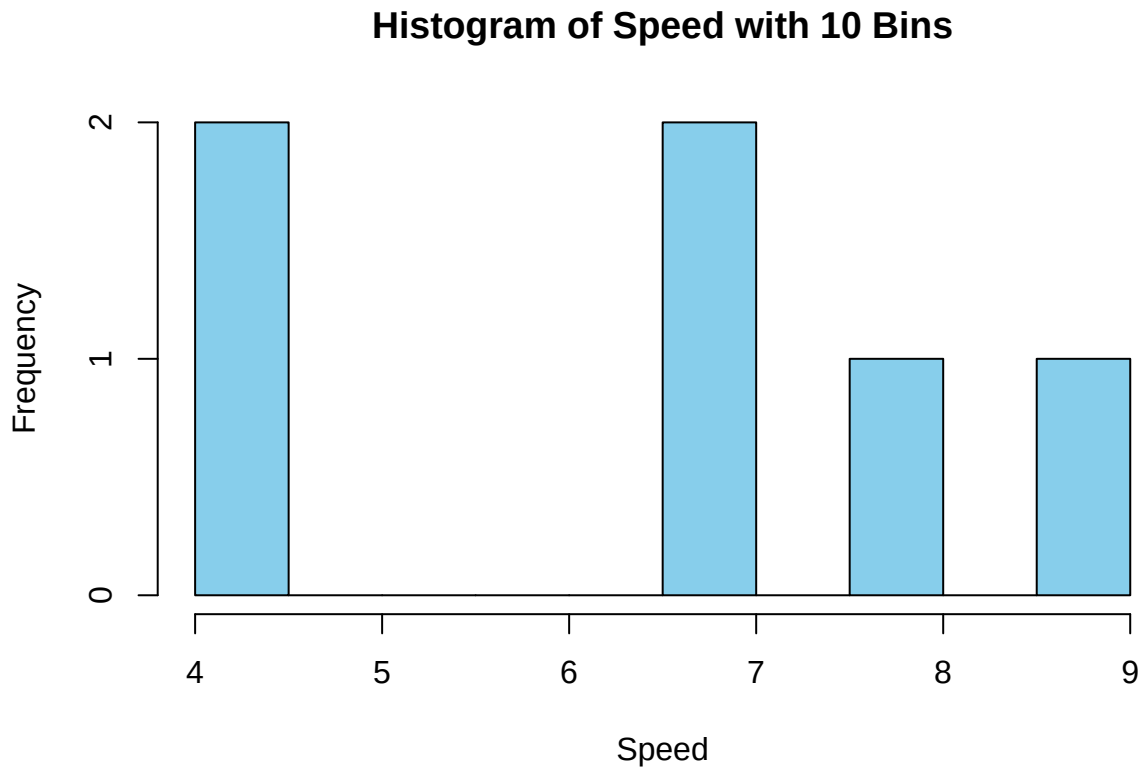
```
## Standard deviation of dist when speed > 12 and speed < 19: NA
```

Average dist when speed > 12 and speed < 19: NaN Standard deviation of dist when speed > 12 and speed < 19: NA

```r
# Given data
speed <- c(4, 4, 7, 7, 8, 9)

# Plot histogram
hist(speed, breaks = 10, main = "Histogram of Speed with 10 Bins", xlab = "Speed", ylab = "Frequency",
```

(g) Produce a histogram of `speed` with 10 bins. Please illustrate your observation from the out-



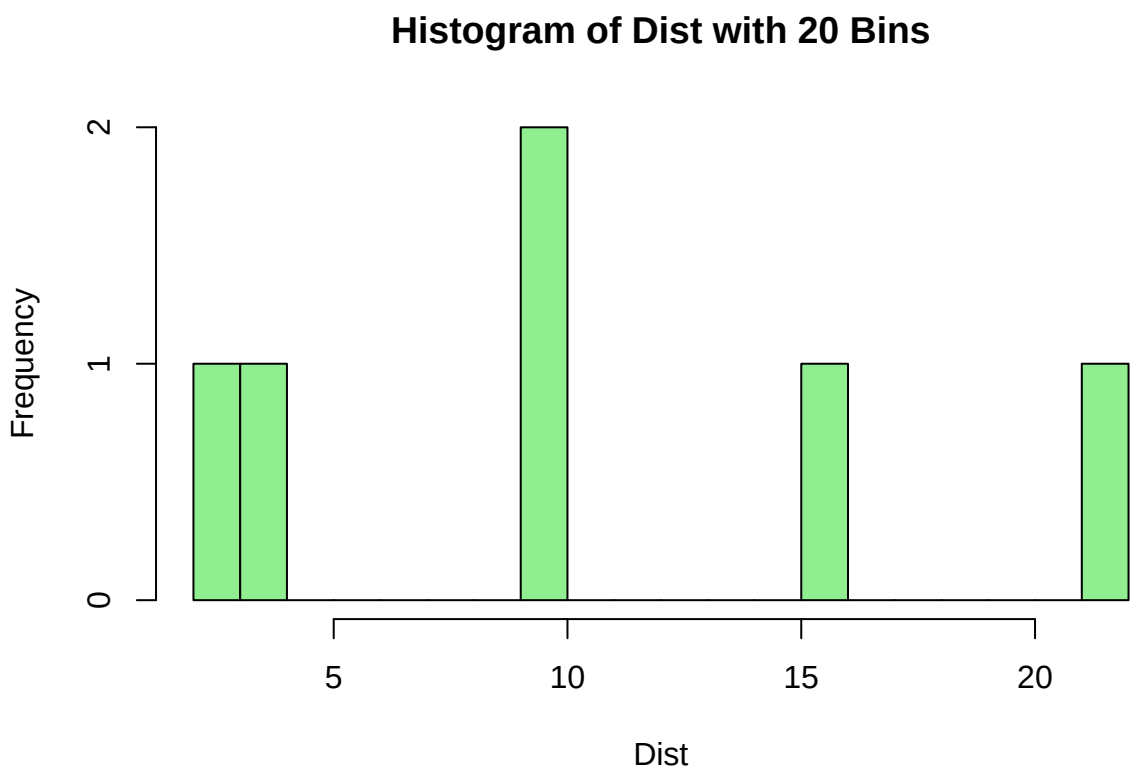
put. (10 points)

Observation: Based on the histogram, you can observe the distribution of speeds and identify any patterns or outliers present in the data. For example, you can see the frequency of speeds within each bin and identify if there are any peaks or clusters indicating common speed ranges. Additionally, you can observe the overall shape of the distribution to assess its symmetry or skewness.

```r
# Given data
dist <- c(2, 10, 4, 22, 16, 10)

# Plot histogram
hist(dist, breaks = 20, main = "Histogram of Dist with 20 Bins", xlab = "Dist", ylab = "Frequency", col
```

(h) Produce a histogram of `dist` with 20 bins. Please illustrate your observation from the output.



(10 points)

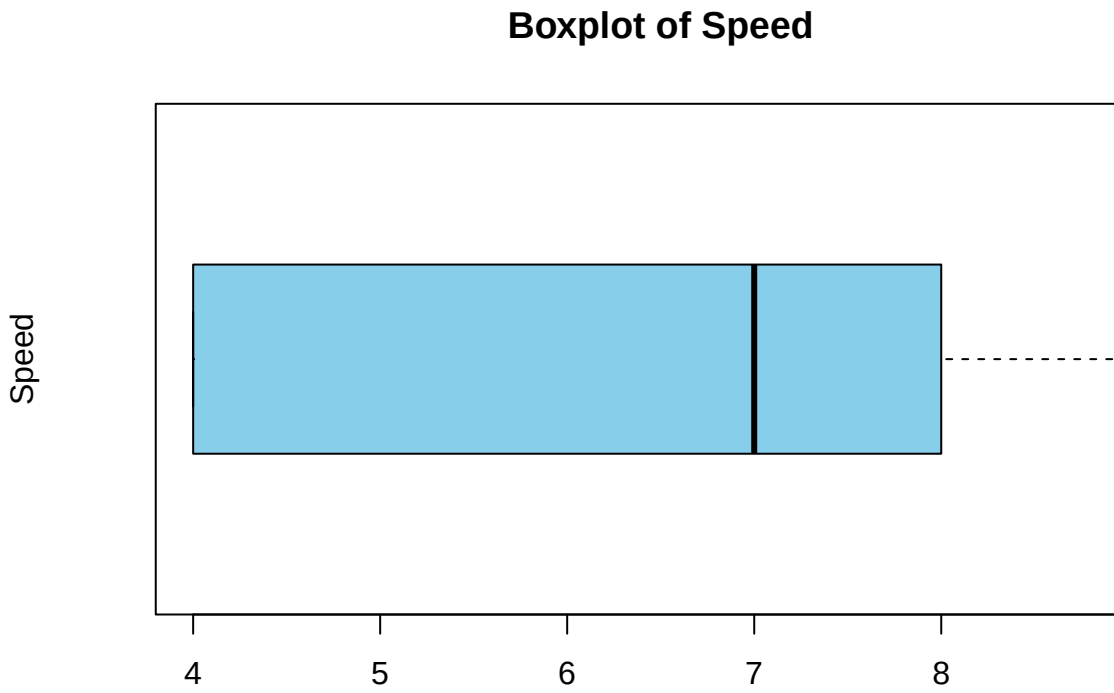
Observation: Based on the histogram, you can observe the distribution of distances and identify any patterns or outliers present in the data. For example, you can see the frequency of distances within each bin and identify if there are any peaks or clusters indicating common distance ranges. Additionally, you can observe the overall shape of the distribution to assess its symmetry or skewness.

```r
# Given data
speed <- c(4, 4, 7, 7, 8, 9)

# Create boxplot with notch set to FALSE
boxplot(speed,
        main = "Boxplot of Speed",
        ylab = "Speed",
        col = "skyblue",
        border = "black",
        horizontal = TRUE,  # Display boxplot horizontally
        notch = FALSE       # Disable notch for confidence interval
```

```
)
```

(i) Use the `boxplot` function to create a boxplot of `speed`. Please illustrate your observation from the output. (10 points)

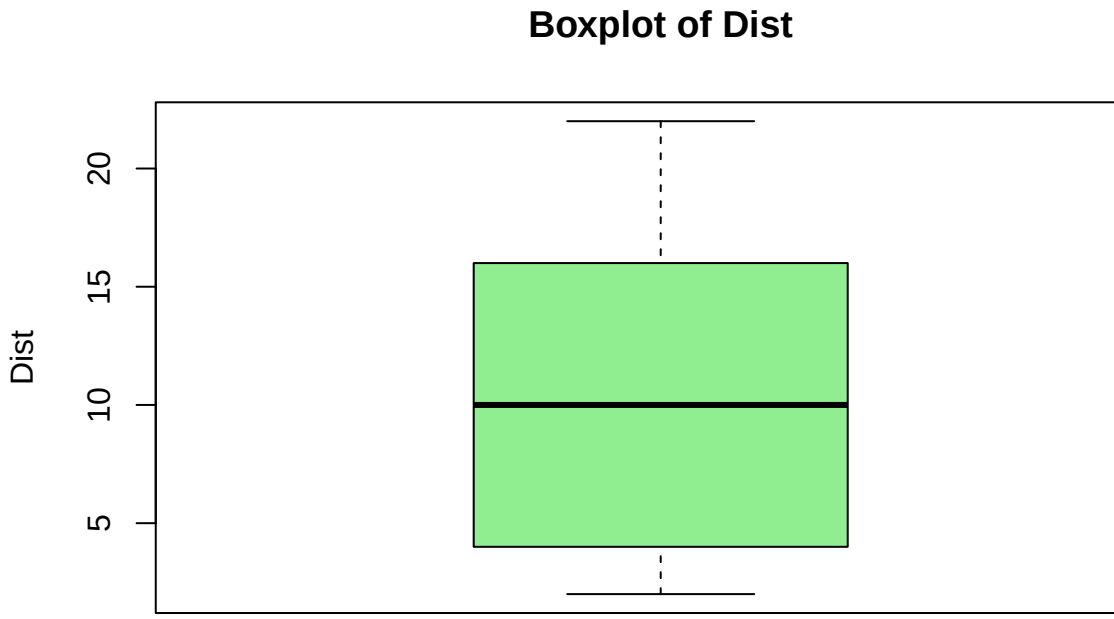


The speed data's distribution unfolds visually through the boxplot. The horizontal line within the blue box pinpoints the median speed, representing the data's center. The box itself encompasses the middle 50% of speeds, capturing the interquartile range (IQR). Extending from the box are the black whiskers, reaching the minimum and maximum values within 1.5 times the IQR from the lower and upper quartiles, respectively. These whiskers effectively illustrate the data's spread. Notably, the absence of any data points beyond the whiskers highlights the lack of outliers, indicating a consistent distribution without extreme values. This horizontal layout, with the blue box and black whiskers, offers a concise yet informative picture of the speed data's central tendency and variability, making comparisons and interpretation straightforward.

```r
# Given data
dist <- c(2, 10, 4, 22, 16, 10)

# Create boxplot of dist
boxplot(dist,
        main = "Boxplot of Dist",
        ylab = "Dist",
        col = "lightgreen",
        border = "black",
        horizontal = FALSE  # Display boxplot vertically
)
```

(j) Use the `boxplot` function to create a boxplot of `dist`. Please illustrate your observation from



the output. (10 points)

The boxplot condenses the key features of the "dist" data's distribution in a readily digestible format. The line within the box highlights the median value, representing the data's center. The box itself captures the middle 50% of the "dist" values, spanning the interquartile range (IQR). Extending from the box are the whiskers, reaching the minimum and maximum values within 1.5 times the IQR from the lower and upper quartiles, respectively. This visualizes the data's spread. Notably, the absence of outliers indicates a consistent distribution with no extreme values. Rendered vertically, the light green box and black whiskers offer a clear and intuitive picture of the "dist" data's central tendency and variability, simplifying interpretation and comparison.

Part III. Random Variables (12 points)

Classify the following random variables as discrete or continuous:

X: the number of automobile accidents per year in Illinois.

The annual count of automobile accidents in Illinois qualifies as a discrete random variable. This is because it's restricted to whole number values (like 0, 1, 2, and so on) and represents separate, quantifiable occurrences.

Y: the length of time to play 18 holes of golf.

The time it takes to complete 18 holes of golf exhibits the characteristics of a continuous random variable. Unlike discrete values, it can adopt any point within a seamless range (think fractions of hours or even minutes) and reflects a continuously changing measurement.

M: the amount of milk produced yearly by a particular cow.

The amount of milk produced yearly by a particular cow is a continuous random variable because

it can take on any value within a continuous range (e.g., liters or gallons), and it represents a continuous measurement.

N: the number of eggs laid each month by a hen.

The number of eggs laid each month by a hen is a discrete random variable because it can only take on integer values (e.g., 0, 1, 2, ...), and it represents distinct, countable events.

P: the number of building permits issued each month in a certain city.

The number of building permits issued each month in a certain city is a discrete random variable because it can only take on integer values (e.g., 0, 1, 2, ...), and it represents distinct, countable events.

Q: the weight of grain produced per acre.

The weight of grain produced per acre is a continuous random variable because it can take on any value within a continuous range (e.g., kilograms or pounds), and it represents a continuous measurement.