Name : Manreen Kaur A/P Jagjit Singh

Total Marks

/ 50

TP Number : 071290

Intake : APU2F2411CS(AI)

**Duration: 1 hour 30 minutes**

**Answer ALL questions. Please snap the picture of your solutions and paste it below.**

Question 1:

# Install and load necessary packages

install.packages("ggplot2")

install.packages("moments")

library(ggplot2)

library(moments)

# Define the dataset

Time\_Spent <- c(5, 6, 7, 8, 9, 10, 11, 12, 13)

Employees <- c(18, 30, 42, 36, 24, 15, 9, 4, 2)

(a)

# Create a histogram for time spent in training

barplot(Employees, names.arg = Time\_Spent, col = "lightblue",

main = "Histogram of Time Spent on Training",

xlab = "Time Spent (Hours)", ylab = "Number of Employees",

border = "black")

A graph of blue rectangular bars

AI-generated content may be incorrect.

(b)

# Create a boxplot

boxplot(rep(Time\_Spent, Employees), col = "lightgreen",

main = "Boxplot of Time Spent on Training",

ylab = "Time Spent (Hours)")

A screenshot of a computer

AI-generated content may be incorrect.

(c)

**Histogram:**

* The histogram shows a unimodal distribution of time spent on training.
* Most employees spent around 7–8 hours in training, as indicated by the peak in this range.
* The distribution appears slightly left-skewed, as there are fewer employees who spent less than 6 hours compared to those spending more time (6-9 hours).

**Boxplot:**

* The boxplot shows the central tendency and variability of the time spent on training.
* The interquartile range (IQR) is between approximately 6 and 9 hours.
* There are no outliers, as no points lie outside the whiskers of the boxplot.
* The median is around 7.5 hours, showing that half of the employees spent at least this much time on training

(d)  
(i)

# Compute percentiles

data\_expanded <- rep(Time\_Spent, Employees)

quantile(data\_expanded, probs = c(0.10, 0.25, 0.75))

A close-up of a computer code

AI-generated content may be incorrect.

The output is 10% = 5.9, 25% = 6.0, 75% = 9.0

(ii)

A close-up of a computer code

AI-generated content may be incorrect.

Based on the output;

* **10% (5.9):** 10% of employees spent 5.9 hours or less on training.
* **25% (6.0):** 25% of employees spent 6.0 hours or less on training (first quartile).
* **75% (9.0):** 75% of employees spent 9.0 hours or less on training (third quartile).

Interpretation of the Data:

IQR = Q3 - Q1 = 9.0 - 6.0

= 3.0

This indicates that the middle 50% of employees spent between 6.0 and 9.0 hours on training.

1. The 10th percentile (5.9) suggests that only a small proportion (10%) of employees spent significantly less time on training, just under the first quartile.
2. Most employees spent time closer to the 25th percentile (6.0) to the 75th percentile (9.0), indicating a concentration of data around this range.

(iii)  
# Interquartile range  
IQR\_value <- IQR(data\_expanded)

IQR\_value

A close-up of a white background

AI-generated content may be incorrect.

# IQR is 3

(e)

mean\_value <- mean(data\_expanded)

median\_value <- median(data\_expanded)

sd\_value <- sd(data\_expanded)

round(c(mean\_value, median\_value, sd\_value), 1)

A close-up of a computer code

AI-generated content may be incorrect.

(f)

skewness(data\_expanded)



This value represents the Pearson correlation coefficient between the variables (e.g., training time and another variable like performance or productivity).

**Strength of the Relationship:**  
The correlation coefficient r=0.5717366r = 0.5717366r=0.5717366 indicates a **moderate positive correlation** between the two variables.

**Direction of the Relationship:**  
Since r>0r > 0r>0, the relationship is positive, meaning that as one variable increases (e.g., training time), the other variable (e.g., performance) tends to increase as well.

**Practical Implication:**  
There is evidence to suggest that employees who spend more time on training generally show improved outcomes in the associated variable, though the relationship is not perfectly strong.

(g)

Since the histogram and boxplot indicate a slightly left-skewed distribution, the median is the best measure of central tendency. The median (7.5 hours) better represents the typical time spent in training as it is not influenced by extreme values, unlike the mean.

Question 2:

# Install and Load

install.packages("MASS")

library(MASS)

data(cats)

(a)  
(i)

# Build the linear regression model

anatomy <- lm(Bwt ~ Hwt, data = cats)

# Display the summary of the regression model

summary(anatomy)

A screenshot of a computer program

AI-generated content may be incorrect.

(ii)

y=a+bx

Where:

* y = Body weight (Bwt)
* x = Heart weight (Hwt)
* a = Intercept
* b = Slope

From the summary output:

Bwt=a+b×Hwt

Replace a and b with the actual values from the regression output.

The regression equation in the form y=a+bx is:

Bwt=1.0196+0.1603×Hwt = 1.0196 + 0.1603 times HwtBwt=1.0196+0.1603×Hwt

Where:

* Bwt = Body weight (kg)
* Hwt = Heart weight (g)
* a=1.0196 (Intercept)
* =0.1603 (Slope)

This means that for each 1g increase in heart weight, the body weight increases by 0.1603 kg.

(iii) Interpretation of a and b

* a (Intercept): The estimated body weight when the heart weight is 0 grams. While not realistic, it is a statistical reference.
* b (Slope): Indicates how much the body weight changes for each additional gram of heart weight. A positive slope means that an increase in heart weight corresponds to an increase in body weight.

(iv)

# (iv)

cor(cats$Hwt, cats$Bwt)

A close-up of numbers

AI-generated content may be incorrect.

This computes the **product-moment correlation coefficient (r)**, which measures the strength and direction of the linear relationship.

* **r is close to 1** → Strong positive correlation.

(v)

A close-up of a math equation

AI-generated content may be incorrect.

(vi)

plot(cats$Hwt, cats$Bwt, main = "Scatter Plot of Heart Weight vs Body Weight",

xlab = "Heart Weight (g)", ylab = "Body Weight (kg)", pch = 16, col = "blue")

A screen shot of a graph

AI-generated content may be incorrect.  
(vii)

abline(anatomy, col = "red", lwd = 2)

A diagram of a graph

AI-generated content may be incorrect.

(b)

(i)

A close-up of a computer code

AI-generated content may be incorrect.

Spearman’s Rank Correlation Coefficient

This will output ρ, the non-parametric measure of correlation.

(ii)

The value 0.5196 indicates a moderate positive correlation between body weight (Bwt) and heart weight (Hwt).

This means that as heart weight increases, body weight tends to increase as well, but the relationship is not perfectly strong.  
Since the correlation is positive, there is a direct relationship between the two variables: higher heart weight is generally associated with higher body weight.