192021002

SUDHARSANAM.J

1. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. What is the median?

code :

```
ages <- c(13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)

sorted_ages <- sort(ages)

median_age <- median(sorted_ages)
```

output:

```
[1] 25
```

2. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

code:

```
ages <- c(13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)

Q1 <- quantile(ages, 0.25)

Q3 <- quantile(ages, 0.75)
```

```
> Q1
 25%
  20


> Q3
 75%
  35
```

3.Load iris Dataset which is inbuilt in R .explore the dataset in terms of dimension and summary statistics (2M)

code:

```
data(iris)
dim(iris)
```

[1] 150 5

```
head(iris)
summary(iris)
```

```
 Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

4.Find the categorical column data and convert that to factor form, also find the number of rows for each factors in dataset. (2)

```
iris$Species <- as.factor(iris$Species)
table(iris$Species)
```

```
setosa versicolor  virginica
   50         50         50
```

5. Find mean of numeric data in dataset based on Species group. and plot Bar chart (use ggplot) to interpret same (8m)

```
library(dplyr)
library(ggplot2)

dataset <- read.csv("my_dataset.csv")
```

```
species_means <- dataset %>%
  group_by(Species) %>%
  summarize(mean = mean(NumericData))
ggplot(species_means, aes(x = Species, y = mean)) +
  geom_bar(stat = "identity") +
  labs(title = "Mean Numeric Data by Species",
       x = "Species",
       y = "Mean Numeric Data")

library(ggplot2)
data(iris)
```

6.Draw a suitable plot which summaries statistical parameter of Sepal.Width based on Species group(6m)

```
ggplot(iris, aes(x = Species, y = Sepal.Width, fill = Species)) +
  geom_boxplot() +
  labs(x = "Species", y = "Sepal Width", title = "Box plot of Sepal Width by Species")
```

7. Draw a suitable plot to find the skewness of the data for Sepal.Width and print the comment about skewness. (6m)

library(ggplot2)

```
data(iris)

ggplot(iris, aes(x = Sepal.Width)) +
  geom_histogram(aes(y = ..density..), bins = 20, color = "black
```

8.Draw ggplot2 scatterplot showing the variables Sepal.Length and Petal.Length grouped by

the three-level factor “Species”. (6m)

```
library(ggplot2)


data(iris)
ggplot(iris, aes(x = Sepal.Length, y = Petal.Length, color = Species)) +
  geom_point() +
  labs(x = "Sepal Length", y = "Petal Length", color = "Species")
```

DAY 4 ASSESSMENT

Sudharsanam.j

192021002

1.Children of th,1ree ages are asked to indicate their preference for three photographs of adults.

Do the data suggest that there is a significant relationship between age and photograph

preference? What is wrong with this study?

Photograph:

Age of child A B C

5-6 years: 18 22 20

7-8 years: 2 28 40

9-10ears: 20 10 40

(i) Use cov() to calculate the sample covariance between B and C.

program :

```
preferences <- matrix(c(18, 22, 20, 2, 28, 40, 20, 10, 40), nrow=3, byrow=TRUE)

cov(preferences[2:3,])
```

output :

```
>180
```

(ii) Use another call to cov() to calculate the sample covariance matrix for the preferences.

program:

```
cov(preferences)
```

output :

```
>      [,1]      [,2]      [,3]
[1,]  64.33333  -5.66667 -58.66667
[2,]  -5.66667 157.33333 150.33333
[3,] -58.66667 150.33333 197.33333
```

(iii)Use cor() to calculate the sample correlation between B and C.

program:

```
cor(preferences[2:3,])
```

output :

0.8586

2. Gopal travels daily from his house located at santhom to his office located at OMR road by

his car and he wants know how much time he spends on travel. He does record the time taken

to reach the off from his home for about a week and has the following value: 46.45, 34.34, 30,

56,12,44.67,43,36.45,48, 35.67,
37.23,32.7,39.20,40.01,45.02,34.12,33.19. Help Gopal to

analyse the time data using skewness and kurtosis and give your interpretation.

travel_time <- c(46.45, 34.34, 30, 56, 12, 44.67, 43, 36.45, 48, 35.67, 37.23, 32.7, 39.20, 40.01, 45.02, 34.12, 33.19)

program:

```
library(moments)
skewness(travel_time)
kurtosis(travel_time)
```

output :

```
[1] -0.1299792
```

3(i). Generate a sample of 5000 random numbers and create a normal distribution

with a mean value of 70 and respectively fix the Standard deviation to

program:

```
set.seed(123) # set seed for reproducibility
random_numbers <- rnorm(5000, mean=70, sd=10)
```

(ii). Calculate the skewness of the normal distribution along with kurtosis and

interpret your results.

program:

```
library(moments)
skewness(random_numbers)
kurtosis(random_numbers)
```

output :

```
[1] -0.003522449
[1] 0.05940867
```

(iii)Write suitable R code to compute the median of the following values.

12,7,3,4.2,18,2,54, -21,8, -5

program:

```
values <- c(12, 7, 3, 4.2, 18, 2, 54, -21, 8, -5)
median(values)
```

(iv) A student recorded her scores on weekly math quizzes that were marked out of a

possible 10 points. Her scores were as follows:

[1] 2.823521

program:

```
scores <- c(8, 9, 7, 6, 5, 10, 8, 7, 8, 9, 6, 4, 5, 8, 7)
mean(scores)
median(scores)
```

output:

```
library(modeest)
mfv(scores)
```

4. The following table of grouped data represents the weight (in kg) of 100 students. Calculate

the mean weight for a student.

Weight (pounds)	Number of Student
21kg	8
30kg	25
56kg	45
73kg	18

110kg 4

program:

```
sum_weights <- 46.2972*8 + 66.1387*25 + 123.4587*45 +
160.9376*18 + 242.5082*4

mean_weight <- sum_weights / 100

mean_weight
```

code:

```
>76.138
```

5. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in

increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40,

45, 46, 52, 70. Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

program:

```
data <- c(13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33,
33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)

Q1 <- quantile(data, 0.25)

Q3 <- quantile(data, 0.75)
```

```
cat("First quartile (Q1) is approximately", Q1, "\n")
cat("Third quartile (Q3) is approximately", Q3, "\n")
```

output:

```
First quartile (Q1) is approximately 20
Third quartile (Q3) is approximately 35
```

6. Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the

following result

age 23 23 27 27 39 41 47 49 50

%fat 9.5 26.5 7.8 17.8 31.4 25.9 27.4 27.2 31.2

age 52 54 54 56 57 58 58 60 61

%fat 34.6 42.5 28.8 33.4 30.2 34.1 32.9 41.2 35.7

a. Calculate the standard deviation of age and %fat.

```
age <- c(23, 23, 27, 27, 39, 41, 47, 49, 50, 52, 54, 54, 56, 57, 58, 58, 60, 61)
fat_perc <- c(9.5, 26.5, 7.8, 17.8, 31.4, 25.9, 27.4, 27.2, 31.2, 34.6, 42.5, 28.8, 33.4, 30.2, 34.1, 32.9, 41.2, 35.7)

age_sd <- sd(age)
```

```
fat_perc_sd <- sd(fat_perc)

age_var <- var(age)
fat_perc_var <- var(fat_perc)
```

program:

```
cat("Standard deviation of age is", age_sd, "\n")
cat("Standard deviation of %fat is", fat_perc_sd, "\n")
cat("Variance of age is", age_var, "\n")
cat("Variance of %fat is", fat_perc_var, "\n")
```

output :

```
Standard deviation of age is 13.01894
Standard deviation of %fat is 10.73649
```

b. Calculate the Variance of age and %fat for the above dataset.

program:

```
cat("Variance of age is", age_var, "\n")
cat("Variance of %fat is", fat_perc_var, "\n")
```

output:

```
Variance of age is 169.7353
Variance of %fat is 115.6055
```

7. Find the H.M of the values 20.0, 35.5, 40.0 and 37.0 with their respective weights 7.0, 8.5,

3.0 and 6.0

program:

```
values <- c(20.0, 35.5, 40.0, 37.0)
weights <- c(7.0, 8.5, 3.0, 6.0)

hm <- sum(weights) / sum(weights / values)

print(hm)
```

output:

[1] 32.59398

8. The demand for a product on each of 20 days was as follows, (in units). 3, 12, 7, 17, 3, 14,

program:

```
demand <- c(3, 12, 7, 17, 3, 14, 9, 6, 11, 10, 1, 4, 19, 7, 15, 6, 9, 12, 12, 8)

mean_demand <- mean(demand)
```

```
cat("Arithmetic mean of demand is:", mean_demand)
```

output :

```
 mean is: 9.2
```