ITA0448- STATISTICS WITH R PROGRAMMING

192011223

ASSESSMENT-4 part 2

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?

A.

To find the median of the given data, we need to arrange the age values 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

The median is the middle value in the dataset when it is arranged in order. If there are an odd number of values, the median is the middle value. If there are an even number of values, the median is the average of the two middle values.

In this case, there are 27 values in the dataset, which is an odd number. Therefore, the median is the middle value, which is the 14th value when the dataset is arranged in order.

So, the median age is 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?

A.

To find the quartiles of the given data, we need to arrange the age values 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

There are different methods to calculate quartiles, but one common method is to use the median. The first quartile (Q1) is the median of the lower half of the data, and the third quartile (Q3) is the median of the upper half of the data.

In this case, the median is 25, which divides the data into two halves:

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25 and 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

The lower half of the data has 14 values, so the median of the lower half is the 7th value when the lower half is arranged in order:

13, 15, 16, 16, 19, 20, 20

Therefore, the first quartile (Q1) is 20.

The upper half of the data also has 14 values, so the median of the upper half is the 7th value when the upper half is arranged in order:

30, 33, 33, 35, 35, 35, 35

Therefore, the third quartile (Q3) is 35.

So, the first quartile (Q1) is 20 and the third quartile (Q3) is 35, which means that the middle 50% of the data (i.e., the interquartile range) lies between 20 and 35

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

A.

SOURCE CODE:

To load the iris dataset in R, you can use the following code:

```
data(iris)
```

To explore the dataset in terms of dimension and summary statistics, you can use the following commands:

```
dim(iris)
summary(iris)
```

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

A.

In the iris dataset, the categorical column is "Species", which has three levels: setosa, versicolor, and virginica. To convert this column to factor form, you can use the factor() function in R:

```
iris$Species <- factor(iris$Species)
```

This will convert the "Species" column to a factor. To find the number of rows for each factor level, you can use the table() function:

```
table(iris$Species)
```

This will give you the count of rows for each factor level:

```
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")
```