Assignment No: 12

**Title:** Iris (flower data) data analytics.

**Theory:**

How to Find the Mean, Median, Mode, Range, and Standard Deviation

Simplify comparisons of sets of number, especially large sets of number, by calculating

the center values using mean, mode and median. Use the ranges and standard deviations

of the sets to examine the variability of data.

**Calculating Mean**

The mean identifies the average value of the set of numbers. For example, consider the

data set containing the values 20, 24, 25, 36, 25, 22, 23.

**Formula**

To find the mean, use the formula: Mean equals the sum of the numbers in the data set

divided by the number of values in the data set. In mathematical terms: Mean=(sum of all

terms)÷(how many terms or values in the set).

**Adding Data Set**

Add the numbers in the example data set: 20+24+25+36+25+22+23=175.

**Finding Divisor**

Divide by the number of data points in the set. This set has seven values so divide by 7.

**Finding Mean**

Insert the values into the formula to calculate the mean. The mean equals the sum of the

values (175) divided by the number of data points (7). Since 175÷7=25, the mean of this

data set equals 25. Not all mean values will equal a whole number.

**Calculating Range**

Range shows the mathematical distance between the lowest and highest values in the data

set. Range measures the variability of the data set. A wide range indicates greater

variability in the data, or perhaps a single outlier far from the rest of the data. Outliers

may skew, or shift, the mean value enough to impact data analysis.

**Identifying Low and High Values**

In the sample group, the lowest value is 20 and the highest value is 36.

**Calculating Range**

To calculate range, subtract the lowest value from the highest value. Since 36-20=16, the

range equals 16.

**Calculating Standard Deviation**

Standard deviation measures the variability of the data set. Like range, a smaller standard

deviation indicates less variability.

**Formula**

Finding standard deviation requires summing the squared difference between each data

point and the mean [∑(x-μ)2], adding all the squares, dividing that sum by one less

than the number of values (N-1), and finally calculating the square root of the

dividend.

Mathematically, start with calculating the mean.

**Calculating the Mean**

Calculate the mean by adding all the data point values, then dividing by the number of

data points. In the sample data set, 20+24+25+36+25+22+23=175. Divide the sum, 175,

by the number of data points, 7, or 175÷7=25. The mean equals 25.

**Squaring the Difference**

Next, subtract the mean from each data point, then square each difference. The formula

looks like this: ∑(x-μ)2, where ∑ means sum, x represents each data set value and μ

represents the mean value. Continuing with the example set, the values become: 20-25=-5

and -5^2=25; 24-25=-1 and -12=1; 25-25=0 and 02=0; 36-25=11 and 112=121; 25-25=0 and

0 2=0; 22-25=-3 and -3 2=9; and 23-25=-2 and -2 2=4. 74

**Adding the Squared Differences**

Adding the squared differences yields: 25+1+0+121+0+9+4=160.

Division by N-1

Divide the sum of the squared differences by one less than the number of data points. The

example data set has 7 values, so N-1 equals 7-1=6. The sum of the squared differences,

160, divided by 6 equals approximately 26.6667.

**Standard Deviation**

Calculate the standard deviation by finding the square root of the division by N-1. In the

example, the square root of 26.6667 equals approximately 5.164. Therefore, the standard

deviation equals approximately 5.164.

**Evaluating Standard Deviation**

Standard deviation helps evaluate data. Numbers in the data set that fall within one

standard deviation of the mean are part of the data set. Numbers that fall outside of two

standard deviations are extreme values or outliers. In the example set, the value 36 lies

more than two standard deviations from the mean, so 36 is an outlier. Outliers may

represent erroneous data or may suggest unforeseen circumstances and should be

carefully considered when interpreting data.

**Facilities :** Windows/Linux Operating Systems, RStudio, jdk.

**Applications:**

1. The histogram is suitable for visualizing distribution of numerical data over a

continuous interval, or a certain time period. The histogram organizes large amounts

of data, and produces a visualization quickly, using a single dimension.

2. The box plot allows quick graphical examination of one or more data sets. Box plots

may seem more primitive than a histogram but they do have some advantages. They take

up less space and are therefore particularly useful for comparing distributions between

several groups or sets of data. Choice of number and width of bins techniques can heavily

influence the appearance of a histogram, and choice of bandwidth can heavily influence

the appearance of a kernel density estimate.

3. Data Visualization Application lets you quickly create insightful data visualizations, in 75

minutes.

Data visualization tools allow anyone to organize and present information intuitively.

They enables users to share data visualizations with others.

**Program: -**

#print("HelloWorld")

#create a variable

#i <- "Sample"

#class(i) #returns datatype

#typeof(i) #returns datatype

#dim(iris) # dimention of dataset

#plot(pressure,pch=15,col="red",bg="lightblue") # Sample plot

#reads the dataset into df variable

df<-read.table("iris.data",sep=",")

print(df)

head(iris) #first Six instences of dataset

head(df)

#features/columns and datatypes

print("Structure of Dataset")

str(df) #provides the structure of dataframe

str(iris)

#datatypes

print("Datatypes of attributes")

class(df$V1)

class(df$V2)

class(df$V3)

class(df$V4)

class(df$V5) #typeof() can be used as alternative

typeof(df$V5)

is.numeric(df$V1)

is.integer(df$V1)

#Structure of Dataset

str(iris) #provides the structure of the data frame

max(df$V1)

min(df$V1)

mean(df$V1)

median(df$V1)

sd(df$V1) #Standard deviation

range(df$V1) #shows min and max value in given range

var(df$V1) #Variance

quantile(df$V1) #

summary(df) # show the summary of dataframe

#ploting the histogram of dataset column

hist(df$V1 ,xlab = "V1 scale", ylab = "Frequency Distribution", col="lightblue1", border="dodgerblue3")

hist(df$V2 ,xlab = "V2 scale", ylab = "Frequency Distribution", col="lightblue1", border="dodgerblue3")

hist(df$V3 ,xlab = "V3 scale", ylab = "Frequency Distribution", col="lightblue1", border="dodgerblue3")

hist(df$V4 ,xlab = "V4 scale", ylab = "Frequency Distribution", col="lightblue1", border="dodgerblue3")

#ploting the boxplot of dataset column

boxplot(df$V1)

boxplot(df$V2)

boxplot(df$V3)

boxplot(df$V4)

#All boxplots Combine into Single plot

boxplot(df$V1,df$V2,df$V3,df$V4,

main = "Multiple boxplots for comparision",

at = c(1,2,3,4),

names = c("V1", "V2", "V3", "V4"),

horizontal = TRUE,

notch = TRUE,

col = "lightBlue1",

border="dodgerblue3")

#Skewness

#if less than -1 or greater than 1 - Highly Skewed

#if -1 and -0.5 or 0.5 and 1 then -moderately skewed

#if -0.5 and 0.5 - approx symmetric

library(moments)

skewness(df$V1)

skewness(df$V2)

skewness(df$V3)

skewness(df$V4)

**Output:**













