Bharati Vidyapeeth
Deemed to be University
Department of Engineering and Technology
Plot no. KC-1, Sector 3, Kharghar, Navi Mumbai-410210

Subject: Computing Lab - III | Experiment No - 03 (3rd YEAR CSE-AIML 2023-2024)

Roll No: 11	Name: Kamran Khan
Class: CSE-AIML	Batch: B1
PRN: 2143110133	Date of Experiment: 02 / 02 / 2024
Marks (Out of 25):	Date of Submission: __ / __ / 2024

Aim: To study Descriptive statistics/Exploratory Data Analysis in R Programming.

Theory:

A. Univariate Analysis

1. Reading the inbuilt dataset iris:

In R, datasets are often provided for experimentation. We read the iris dataset, a popular dataset in data science, which contains measurements of iris flowers.

```
d <- iris
```

2.Using `head()`:

The `head()` function displays the first few rows of a dataset, providing a glimpse of its structure and contents.

```
head(d)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

3. Accessing Columns with their names:

Columns in a dataset can be accessed using the `$` operator followed by the column name.

```
d$Species
```

```
##   [1] setosa     setosa     setosa     setosa     setosa     setosa
##   [7] setosa     setosa     setosa     setosa     setosa     setosa
##  [13] setosa     setosa     setosa     setosa     setosa     setosa
##  [19] setosa     setosa     setosa     setosa     setosa     setosa
##  [25] setosa     setosa     setosa     setosa     setosa     setosa
##  [31] setosa     setosa     setosa     setosa     setosa     setosa
##  [37] setosa     setosa     setosa     setosa     setosa     setosa
##  [43] setosa     setosa     setosa     setosa     setosa     setosa
##  [49] setosa     setosa     versicolor versicolor versicolor versicolor
##  [55] versicolor versicolor versicolor versicolor versicolor versicolor
##  [61] versicolor versicolor versicolor versicolor versicolor versicolor
##  [67] versicolor versicolor versicolor versicolor versicolor versicolor
##  [73] versicolor versicolor versicolor versicolor versicolor versicolor
##  [79] versicolor versicolor versicolor versicolor versicolor versicolor
##  [85] versicolor versicolor versicolor versicolor versicolor versicolor
##  [91] versicolor versicolor versicolor versicolor versicolor versicolor
##  [97] versicolor versicolor versicolor versicolor virginica  virginica
## [103] virginica  virginica  virginica  virginica  virginica  virginica
## [109] virginica  virginica  virginica  virginica  virginica  virginica
## [115] virginica  virginica  virginica  virginica  virginica  virginica
## [121] virginica  virginica  virginica  virginica  virginica  virginica
## [127] virginica  virginica  virginica  virginica  virginica  virginica
## [133] virginica  virginica  virginica  virginica  virginica  virginica
## [139] virginica  virginica  virginica  virginica  virginica  virginica
## [145] virginica  virginica  virginica  virginica  virginica  virginica
## Levels: setosa versicolor virginica
```

4. Structure of iris dataset:

The `str()` function provides information about the structure of a dataset, including the data types of its columns.

```
str(d)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

5. Minimum and Maximum of a column:

The `min()` and `max()` functions calculate the minimum and maximum values of a column, respectively.

```
min(d$Petal.Length)
```

```
## [1] 1
```

```
max(d$Petal.Length)
```

```
## [1] 6.9
```

6. Quantile of Column:

Quantiles divide a dataset into equally sized portions. The `quantile()` function computes the quantile values of a column.

```
quantile(d$Sepal.Length, 0.25)
```

```
## 25%
## 5.1
```

7. Range of Column:

The `range()` function calculates the range (difference between the maximum and minimum values) of a column.

```
range(d$Petal.Width)
```

```
## [1] 0.1 2.5
```

8. Standard Deviation of a column:

Standard deviation measures the dispersion of values around the mean. The `sd()` function computes the standard deviation of a column.

```
sd(d$Petal.Length)
```

```
## [1] 1.765298
```

9. Variance of a Column:

Variance measures the average squared deviation from the mean. The `var()` function calculates the variance of a column.

```
var(d$Petal.Length)
```

```
## [1] 3.116278
```

10. Summary of a dataset:

The `summary()` function provides a summary of statistics for each column in the dataset.

```
summary(d)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  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
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

11. Using `by()` function on a column:

The `by()` function allows splitting a dataset by a factor and applying a function to each subset. Here, we use it to calculate summary statistics for `Petal.Length` grouped by `Species` .

```
by(d$Petal.Length, d$Species, summary)
```

```
## d$Species: setosa
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.400   1.500   1.462   1.575   1.900
## ------------------------------------------------------------
## d$Species: versicolor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    4.00    4.35    4.26    4.60    5.10
## ------------------------------------------------------------
## d$Species: virginica
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.500   5.100   5.550   5.552   5.875   6.900
```

12. Using Stat. Desc Function:

The `stat.desc()` function from the `pastecs` package computes a comprehensive set of descriptive statistics for a dataset.

```
library(pastecs)
stat.desc(d)
```

```
##              Sepal.Length  Sepal.Width Petal.Length  Petal.Width Species
## nbr.val      150.00000000 150.00000000  150.0000000 150.00000000      NA
## nbr.null       0.00000000   0.00000000    0.0000000   0.00000000      NA
## nbr.na         0.00000000   0.00000000    0.0000000   0.00000000      NA
## min            4.30000000   2.00000000    1.0000000   0.10000000      NA
## max            7.90000000   4.40000000    6.9000000   2.50000000      NA
## range          3.60000000   2.40000000    5.9000000   2.40000000      NA
## sum          876.50000000 458.60000000  563.7000000 179.90000000      NA
## median         5.80000000   3.00000000    4.3500000   1.30000000      NA
## mean           5.84333333   3.05733333    3.7580000   1.19933333      NA
## SE.mean        0.06761132   0.03558833    0.1441360   0.06223645      NA
## CI.mean.0.95   0.13360085   0.07032302    0.2848146   0.12298004      NA
## var            0.68569351   0.18997942    3.1162779   0.58100626      NA
## std.dev        0.82806613   0.43586628    1.7652982   0.76223767      NA
## coef.var       0.14171126   0.14256420    0.4697441   0.63555114      NA
```

13. IQR of a Column:

The Interquartile Range (IQR) represents the range between the first and third quartiles, indicating the spread of the middle 50% of the data.

```
IQR(d$Petal.Length)
```

```
## [1] 3.5
```

B. Multivariate Analysis:

1. Using Corelation function:

Correlation measures the strength and direction of the linear relationship between two variables. The `cor()` function calculates the correlation coefficient between two columns.

```
cor(d$Sepal.Length, d$Sepal.Width)
```

```
## [1] -0.1175698
```

2. Using Covariance function:

Covariance measures how two variables change together. The `cov()` function computes the covariance between two columns.

```
cov(d$Sepal.Length, d$Petal.Length)
```

```
## [1] 1.274315
```

Conclusion:

In this experiment through the analysis of the iris dataset, we learned various univariate and multivariate techniques to summarize and interpret data. These techniques are fundamental for gaining insights into datasets and forming hypotheses for further analysis in data science projects.

Signature of Lab Incharge

(Prof. Supriya Khaitan)