Vel Tech
Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology
(Deemed to be University Estd. u/s 3 of UGC Act, 1956)

DEPARTMENT OF MATHEMATICS
SCHOOL OF SCIENCE AND HUMANITIES

R-Language Laboratory Record

(Winter Semester)

ACADEMIC YEAR
2021-2022

Vel Tech
Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology
(Deemed to be University Estd. u/s 3 of UGC Act, 1956)

R-LANGUAGE LABORATORY RECORD

Name of the Student : ……………………………………….

VTU / Register Number : ……………………………………….

Subject : ……………………………………….

Subject Code : ……………………………………….

Degree : ……………………………………….

Department : ……………………………………….

Faculty In charge/Slot : ……………………………………….

Semester / Academic Year : ……………………………………….

R-LANGUAGE LABORATORY RECORD

Certified that this is the bonafide record of work done by

Mr. / Ms. …………………………………………………. Reg. No. / VTU No ……..…………………………..

of ……………. year ………………… semester, in the branch ………………………………………………

during the academic year 2021-2022.

Signature of faculty In-Charge with date

Signature of the Head of the Department with date

Submitted for the University Practical Examination held on ……………………………………….

Signature of the Internal Examiner with date

Signature of the External Examiner with date

Content

1. DIAGRAMMATICAL AND GRAPHICAL REPRESENTATION OF DATA

Aim: To Construct the Box plot, Stem and Leaf Diagram, Probability plot, Histogram and Time Series plot for the given data using R - Language.

Problem: The US petroleum imports as a percentage of totals, and Persian Gulf imports as percentage of all imports by year since 197 was given in the following table.

Sl. NO	Year(Y)	Petroleum Imports(A)	Total Petroleum Imports(B)	GP Import(C)
1	1975	6055	37.1	19.2
2	1976	7313	41.8	25.1
3	1977	8807	47.7	27.8
4	1978	8363	44.3	26.5
5	1979	8456	45.6	24.4
6	1980	6909	40.5	21.9
7	1981	5996	37.3	20.3
8	1982	5113	33.4	13.6
9	1983	5051	33.1	8.7
10	1984	5437	34.5	9.3
11	1985	5067	32.2	6.1
12	1986	6224	38.2	14.6
13	1987	6678	40	16.1
14	1988	7402	42.8	20.8
15	1989	8061	46.5	23
16	1990	8018	47.1	24.5
17	1991	7627	45.6	24.1
18	1992	7888	46.3	22.5
19	1993	8620	50	20.6
20	1994	8996	50.7	19.2
21	1995	8835	49.8	17.8
22	1996	9478	51.7	16.9
23	1997	10162	54.5	17.2
24	1998	10708	56.6	19.9
25	1999	10852	55.5	22.7
26	2000	11459	58.1	21.7
27	2001	11871	60.4	23.2
28	2002	11530	58.3	19.6
29	2003	12264	61.2	20.3
30	2004	13145	63.4	18.9

Using External Data:

R offers plenty of options for loading external data, including Excel, Minitab, SAS and SPSS files.

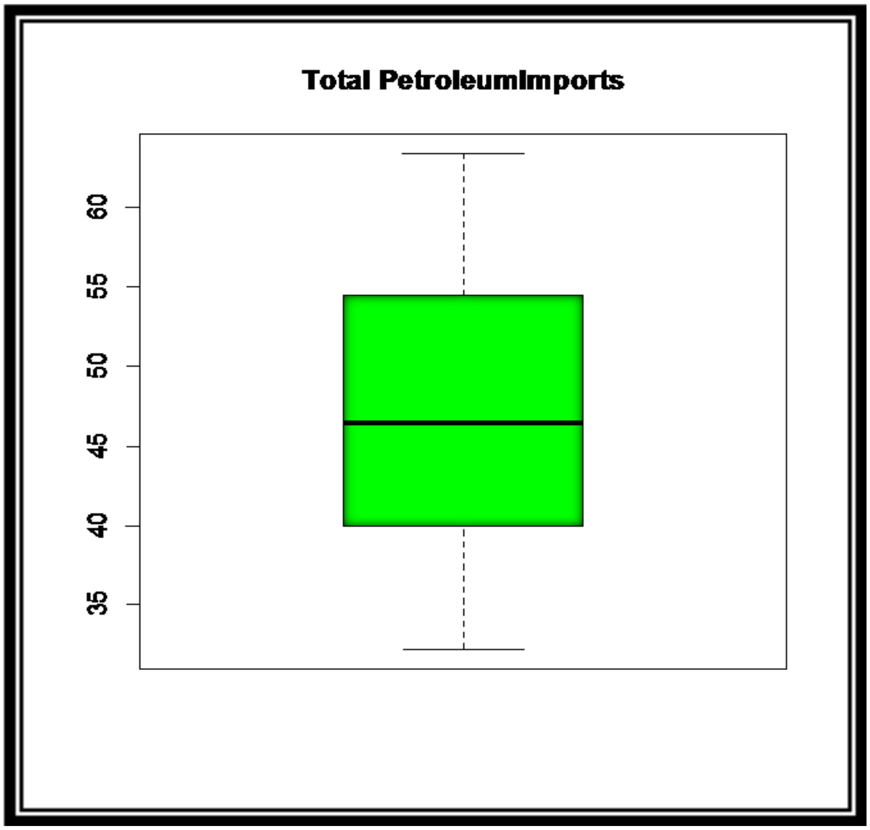
- **First enter the given data in Excel file.**
- **Save the excel file data in Command Separated Values file (CSV).**
- **In R Console window:**
  Data=read.csv("C:\\Users\\Student\\Desktop\\petroleum import data.csv")
  Data
  # this will read the file and display data in the window and we can store this data under a name Data

R – Code:-

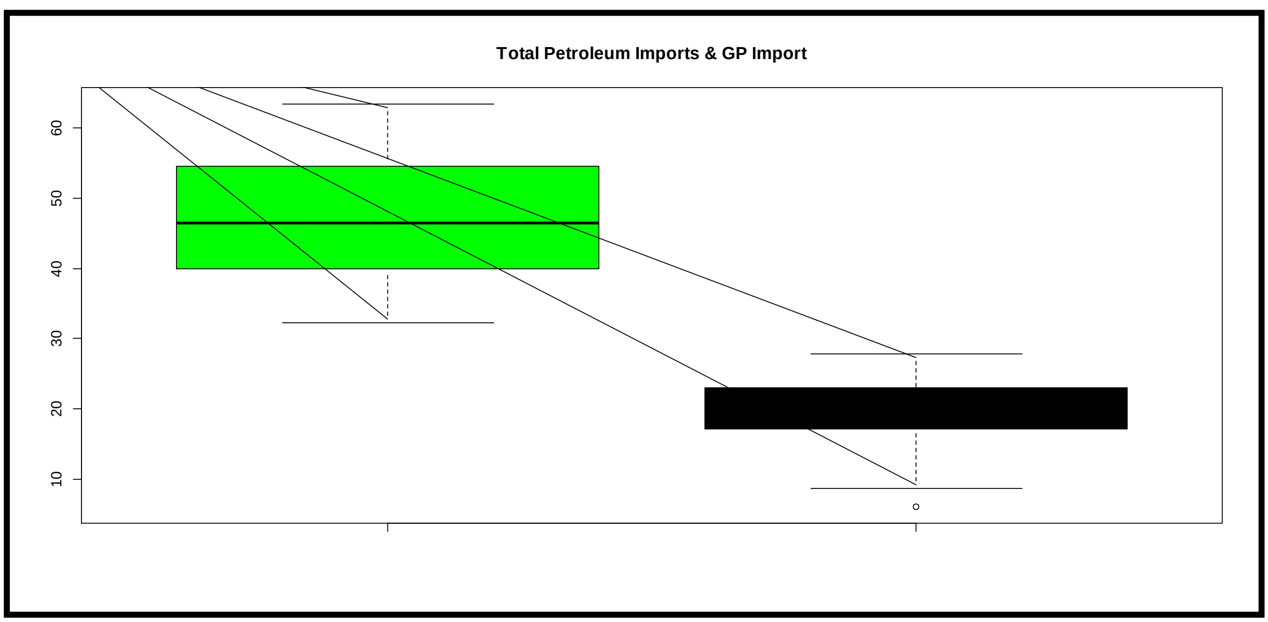
```
Data=read.csv("C:\\Users\\Student\\Desktop\\petroleum import data.csv")
 # Importing the  External data
>Y=Data$Year
>A=Data$PetroleumImports
>B= Data$TotalPetroleumImports
>C= Data$GPImport
> print(Y)
[1] 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991
1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004
> print(A)
 [1]  6055  7313  8807  8363  8456  6909  5996  5113  5051  5437  5067  6224  6678  7402  8061
8018  7627 7888  8620  8996  8835  9478 10162 10708 10852 11459 11871 11530 12264 13145
> print(B)
[1] 37.1 41.8 47.7 44.3 45.6 40.5 37.3 33.4 33.1 34.5 32.2 38.2 40.0 42.8 46.5 47.1 45.6 46.3 50.0
50.7 49.8 51.7 54.5 56.6 55.5 58.1 60.4 58.3 61.2 63.4
> print(C)
 [1] 19.2 25.1 27.8 26.5 24.4 21.9 20.3 13.6 8.7 9.3  6.1 14.6 16.1 20.8 23.0 24.5 24.1 22.5 20.6
19.2 17.8 16.9 17.2 19.9 22.7 21.7 23.2 19.6 20.3 18.9
```

Box Plot

```
> boxplot(B,main="Total PetroleumImports",col=c("green"))
```




```
> boxplot(B,C,main="Total Petroleum Imports & GP Import",col=c("green","black"))
```

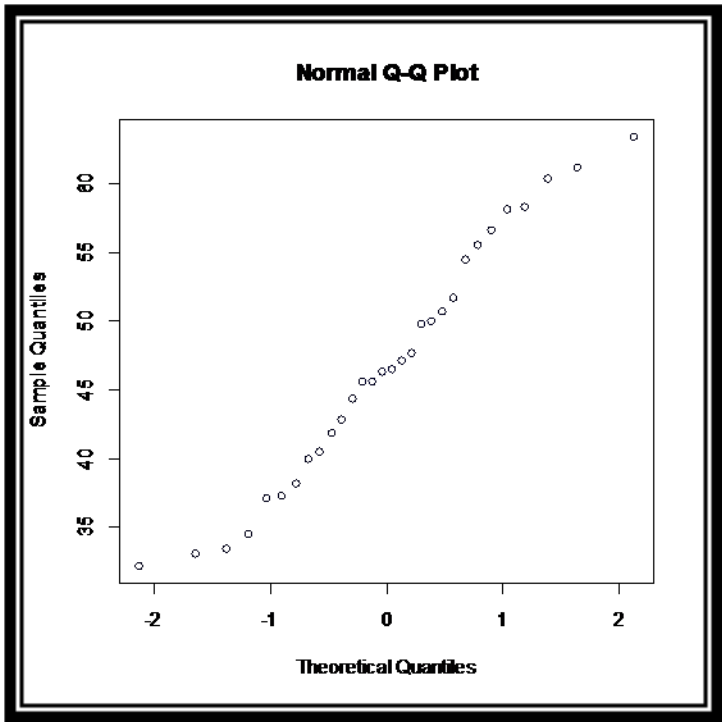


Stem and Leaf Diagram

```
> stem(C)

The decimal point is at the |

   6 | 1
   8 | 73
  10 |
  12 | 6
  14 | 6
  16 | 1928
  18 | 92269
  20 | 336879
  22 | 5702
  24 | 1451
  26 | 58

> stem(B,scale=2)

The decimal point is at the |

  32 | 214
  34 | 5
  36 | 13
  38 | 2
  40 | 058
  42 | 8
  44 | 366
  46 | 3517
  48 | 8
  50 | 077
  52 |
  54 | 55
  56 | 6
  58 | 13
  60 | 42
  62 | 4
```

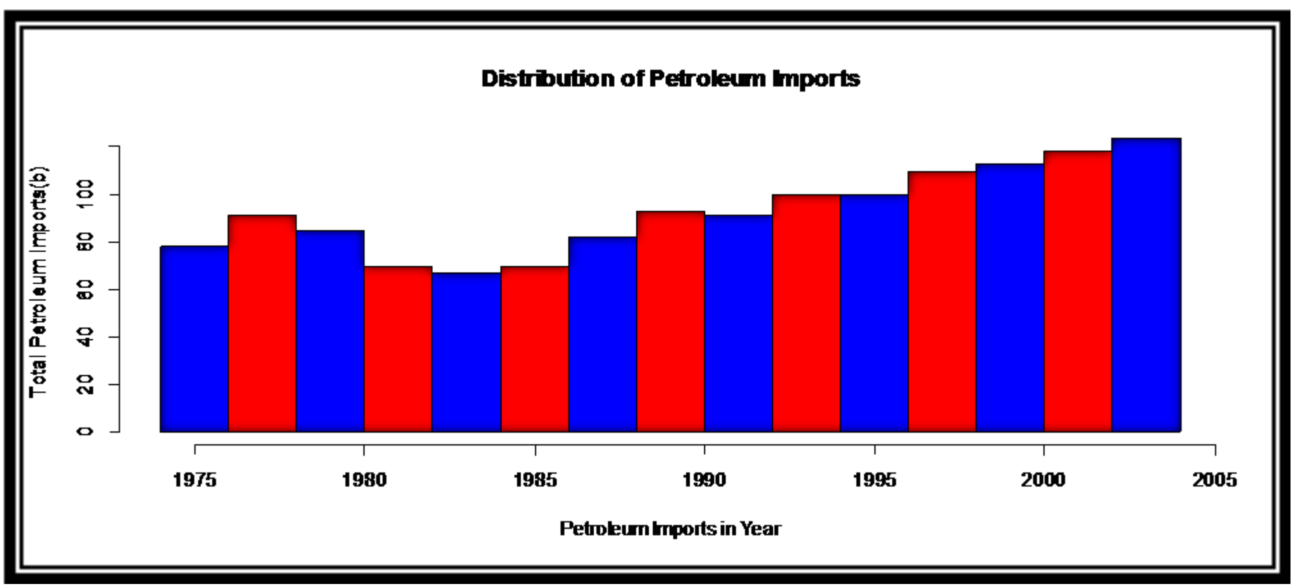
Probability Plot

>qqnorm(B,col="blue")



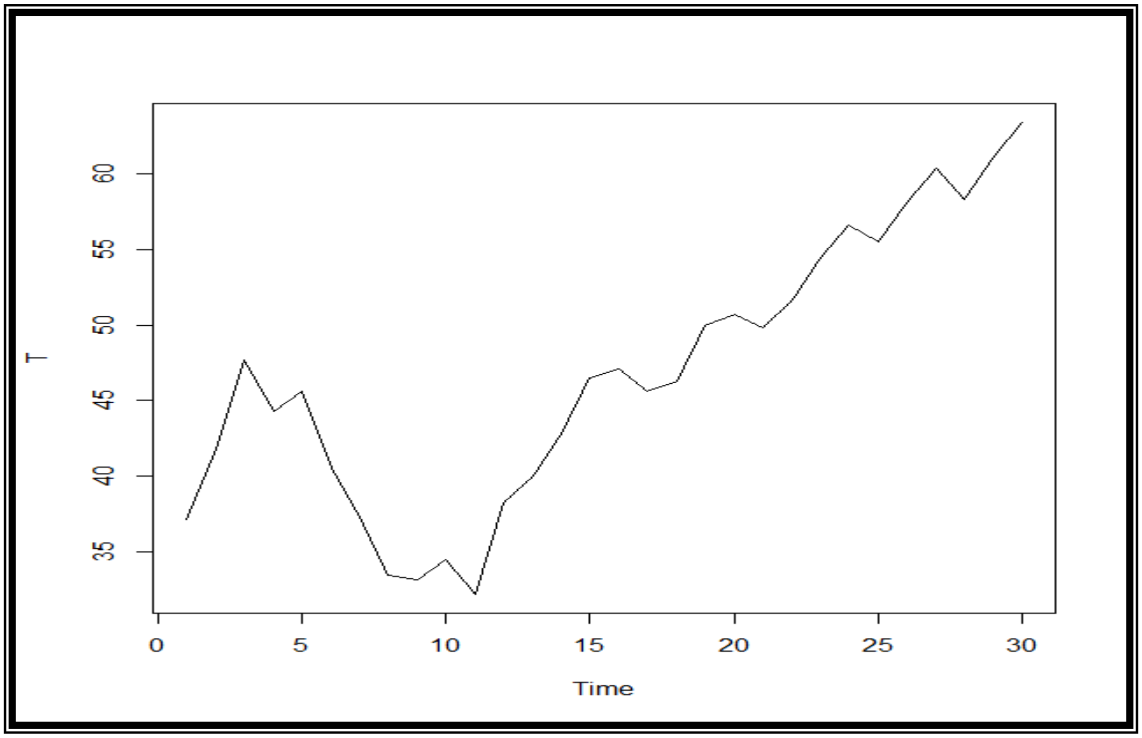
Histogram

> hist(rep(Y,B), main="Distribution of Petroleum Imports", xlab="Petroleum Imports in Year",ylab=" Total Petroleum Imports(b)",col=c("blue","red"))



Time Series Plot

```
> T=ts(B)
> T
Time Series:
Start = 1
End = 30
Frequency = 1
 [1] 37.1 41.8 47.7 44.3 45.6 40.5 37.3 33.4 33.1 34.5 32.2 38.2 40.0 42.8 46.5
[16] 47.1 45.6 46.3 50.0 50.7 49.8 51.7 54.5 56.6 55.5 58.1 60.4 58.3 61.2 63.4
> plot(T)
```



2. MEASURES OF CENTRAL TENDENCY, MEASURES OF DISPERSION FOR RAW DATA AND DISCRETE FREQUENCY DISTRIBUTION

Aim: To find the measures of central tendency , dispersion of the given raw data and discrete frequency distribution data respectively.

Problem: 1 Find the measures of central tendency and dispersion of the given raw data:
2, 2, 2, 3, 1, 1, 1, 4, 5, 6, 8, 8, 9.

R - Code:-.

```
>X=c(2,2,2,3,1,1,1,4,5,6,8,8,9)
> cat("given data =", X,"\n")
given data = 2 2 2 3 1 1 1 4 5 6 8 8 9
>Mean=mean(X)
> Median=median(X)
> mode=function(X)
+  {
+ uni=unique(X)
+ uni[which.max(tabulate(match(X,uni)))]
+ }
> Mode=mode(X)
> S.D=sd(X)
> Variance=var(X)
> Range=range(X)
> Quartile=quantile(X,c(0.25,0.50,0.75))
> Deciles=quantile(X,c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1))
> Percentile=quantile(X,c(0.26,0.76,0.85))
> IQR=IQR(X)
> QD=IQR/2
> n=length(X)
> CV=(S.D/Mean)*100
> MDMean= (sum(abs(X-Mean)))/n
> MD
> MDMedian= (sum(abs(X-Median)))/n
```

```
> GM=prod(X)^(1/n)
> HM=n/sum(1/X)
> cat("Mean of the given data =", Mean,"\n")
Mean of the given data = 4
> cat("Median of the given data =", Median,"\n")
Median of the given data = 3
> cat("Mode of the given data=",Mode,"\n")
Mode of the given data= 2
> cat("Standard Deviation of the given data=",S.D,"\n")
Standard Deviation of the given data= 2.915476
> cat("Variance of the given data=",Variance,"\n")
Variance of the given data= 8.5
> cat("Range of the given data=",Range,"\n")
Range of the given data= 1 9
> cat("Quartile of the given data:",Quartile,"\n")
Quartile of the given data: 2 3 6
> cat("Deciles of the given data:",Deciles,"\n")
Deciles of the given data: 1 1.4 2 2 3 4.2 5.4 7.2 8 9
> cat("Percentiles of the given data for 26%,76% and 85%:",Percentile,"\n")
Percentiles of the given data for 26%,76% and 85%: 2 6.24 8
> cat("Inter Quartile Range of the given data =",IQR,"\n")
Inter Quartile Range of the given data = 4
> cat("Quartile Deviation of the given data =",QD,"\n")
Quartile Deviation of the given data = 2
> cat("Coefficient of Variation of the given data =",CV,"\n")
Coefficient of Variation of the given data = 72.8869
> cat("Mean Deviation about Mean of the given data =",MDMean,"\n")
Mean Deviation about Mean of the given data = 2.461538
> cat("Mean Deviation about Median of the given data =",MDMedian,"\n")
Mean Deviation about Median of the given data = 2.384615
> cat("Geometric Mean of the given data =",GM,"\n")
Geometric Mean of the given data = 3.009174
> cat("Harmonic Mean of the given data =",HM,"\n")
Harmonic Mean of the given data = 2.237094
```

Problem: 2 Find the measures of central tendency and measures dispersion of the following discrete frequency distribution.

x	34	56	57	87	90
f	3	4	8	17	4

R-Code:-

```
> x=c(34,56,76,87,90)
> f=c(3,4,8,17,4)
> Mean= mean(rep(x,f))
> Median= median(rep(x,f))
> Mode=mode(rep(x,f))
> S.D= sd(rep(x,f))
> Variance= var(rep(x,f))
> Quartile=quantile(rep(x,f),c(0.25,0.50,0.75))
> Deciles=quantile(rep(x,f),c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1))
> Percentile=quantile(rep(x,f),c(0.26,0.76,0.85))
> IQR=IQR(rep(x,f))
> QD=IQR/2
> N=sum(f)
> CV=(S.D/Mean)*100
> MDMean=sum(abs(x-Mean))/N
> MDMedian=sum(abs(x-Median))/N
> GM=prod(x*f)^(1/N)
> HM=N/sum(x*f)
> cat("Mean of the given data =", Mean,"\n")
Mean of the given data = 77.02778
> cat("Median of the given data =", Median,"\n")
Median of the given data = 87
> cat("Mode of the given data=",Mode,"\n")
Mode of the given data= 87
> cat("Standard Deviation of the given data=",S.D,"\n")
Standard Deviation of the given data= 16.64329
> cat("Variance of the given data=",Variance,"\n")
Variance of the given data= 276.9992
> cat("Quartile of the given data:",Quartile,"\n")
Quartile of the given data: 76 87 87
> cat("Deciles of the given data:",Deciles,"\n")
Deciles of the given data: 56 76 76 76 87 87 87 87 88.5 90
> cat("Percentiles of the given data for 26%,76% and 85%:",Percentile,"\n")
```

Percentiles of the given data for 26%,76% and 85%: 76 87 87

```
> cat("Inter Quartile Range of the given data =",IQR,"\n")
```

Inter Quartile Range of the given data = 11

```
> cat("Quartile Deviation of the given data =",QD,"\n")
```

Quartile Deviation of the given data = 5.5

```
> cat("Coefficient of Variation of the given data =",CV,"\n")
```

Coefficient of Variation of the given data = 21.60687

```
> cat("Mean Deviation about Mean of the given data =",MDMean,"\n")
```

Mean Deviation about Mean of the given data = 2.445216

```
> cat("Mean Deviation about Median of the given data =",MDMedian,"\n")
```

Mean Deviation about Median of the given data = 2.722222

```
> cat("Geometric Mean of the given data =",GM,"\n")
```

Geometric Mean of the given data = 2.277576

```
> cat("Harmonic Mean of the given data =",HM,"\n")
```

Harmonic Mean of the given data = 0.01298233

3. CORRELATION AND REGRESSION ANALYSIS OF THE DATA

Aim: To find the correlation and regression of the given data.

Problem: 1 Find the correlation of the following data and also draw the scatter diagram.

X	10	12	13	16	17	20	25
Y	19	22	26	27	29	33	37

R-Code:-

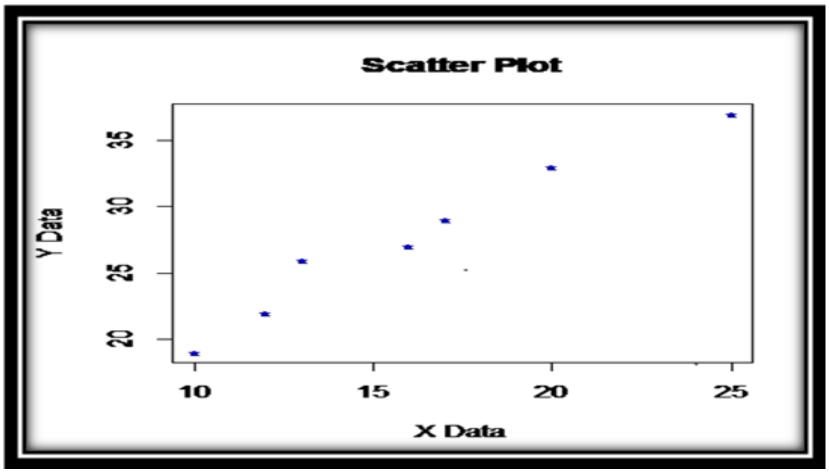
```
>  X=c(10,12,13,16,17,20,25)
>  Y=c(19,22,26,27,29,33,37)
>  r=cor(X,Y, method="pearson")
> cat("Correlation coefficient of the given data =",r,"\n")
```

Correlation coefficient of the given data = 0.9801983

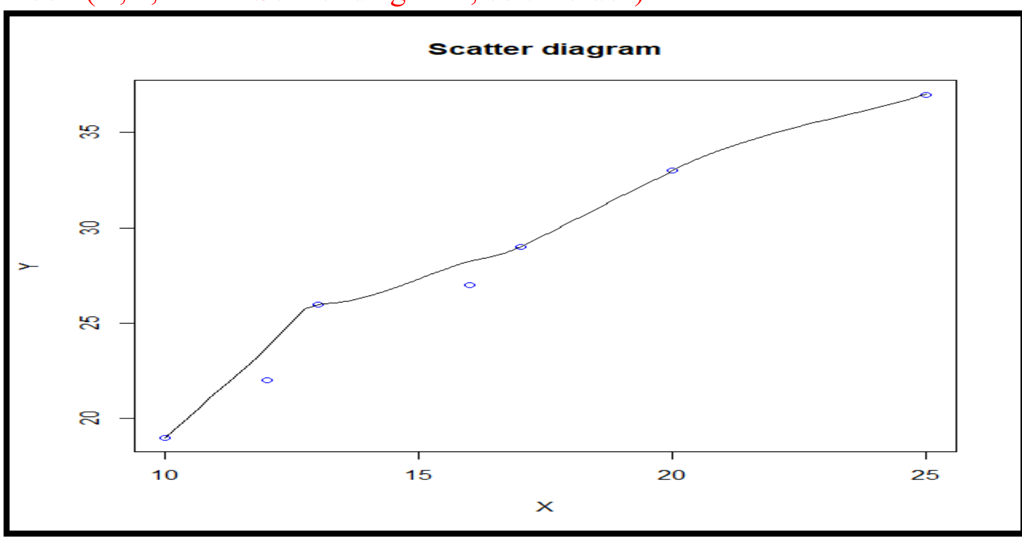
```
> Rho=cor(X,Y, method="spearman")
> cat("Rank Correlation coefficient of the given data =",Rho,"\n")
```

Rank Correlation coefficient of the given data = 1

```
> plot(X,Y, main="Scatter Plot",xlab="X Data", ylab="Y Data", col="blue",pch="*")
```




```
> scatter.smooth(X,Y,main="Scatter diagram", col.="blue")
```



Problem: 2 Find the regression of the following data

X	151	174	138	18	128	136	179	163	152	131
Y	63	81	56	91	47	57	76	72	62	48

R-Code:-

```
> X=c(151,174,138,186,128,136,179,163,152,131)
> Y=c(3,81,56,91,47,57,7,72,62,48)
> reg1=lm(X~Y)
> reg1
Call:
lm(formula = X ~ Y)

Coefficients:
(Intercept)          Y
  146.1282      0.1464

> summary(reg1)
Call:
lm(formula = X ~ Y)

Residuals:
    Min      1Q  Median      3Q     Max
-25.009 -17.937   0.614  13.592  31.847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 146.1282    14.9235   9.792 9.93e-06 ***
Y             0.1464     0.2529   0.579    0.579
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.71 on 8 degrees of freedom
Multiple R-squared:  0.04021,   Adjusted R-squared:  -0.07976
F-statistic: 0.3352 on 1 and 8 DF,  p-value: 0.5785

> reg2=lm(Y~X)
> reg2
Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)          X
   10.1555      0.2747

> summary(reg2)
```

```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-52.322   1.731   8.714  15.328  29.756

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.1555    73.5694   0.138    0.894
X             0.2747     0.4744   0.579    0.579

Residual standard error: 29.73 on 8 degrees of freedom
Multiple R-squared:  0.04021,   Adjusted R-squared:  -0.07976
F-statistic: 0.3352 on 1 and 8 DF,  p-value: 0.5785
```

4. CONFIDENCE INTERVAL FOR THE POPULATION MEAN

FOR LARGE SAMPLE

Aim: To find the confidence interval of the population mean

(i) when the variance is known

(ii) when the variance is Unknown

Problem: 1

A machine produces components, which have a standard deviation of 1.6cm in length. A random sample of 64 parts is selected from the output and this sample has a mean length of 90cm. The customer will reject the part if it is either less than 88cm or more than 92cm. Does the 95% confidence interval for the true mean length of all the components produced ensure acceptance by the customer?

R-Code:-

```
>n=64
>sigma=1.6
>xbar=90
>sem=sigma / sqrt(n);
>sem
[1] 0.2
>E=qnorm(0.975)*sem;
>E
[1] 0.3919928
>xbar+c(-E,E)
[1] 89.081    90.39199
```

Inference:-

This implies that the true value of the population mean length of the components will fail in this interval hence we infer that the 95% confidence interval ensure confidence by the customer.

Problem:2

A Stock market analyst wants to estimate the average return on a certain stock. A random sample of 15 days yields an average return of mean is 10.37% and a standard deviation of s = 3.5 %. Assuming a normal population of returns, give a 95% confidence interval for the average return on this stock.

R-Code:-

```
> n=15

> xbar=10.37

> s=3.5

> sem=s/sqrt(n);

>sem

[1] 0.9036961

> E=qnorm(0.975)*sem;

>E

[1] 1.771212

> xbar+c(-E,E)

[1]  8.598788 12.141212
```

Inference:-

This implies that the analyst may be 95% sure that the average annualized return on the stock is anywhere from 8.60 % to 12.14%.

5. Z-TEST FOR SINGLE MEAN AND DIFFERENCE OF TWO SAMPLE MEANS

Aim: To test whether there is significant difference between

(i) sample mean and population mean,

(ii) two sample means

Problem:1

Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. In a sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the population standard deviation is 120 hours. At 0.05 significance level, can we reject the claim by the manufacturer?

Null hypothesis (H_0) : $\mu \geq 10000$.

Alternative hypothesis (H_1) : $\mu < 10000$

We begin with computing the test statistic

R-Code:-

```
>xbar=9900              #sample mean
> mu0=10000             #hypothesized value
>sigma=120              #population standard deviation
> n=30                  #sample size
> z=abs((xbar-mu0)/(sigma/sqrt(n)))
>z                      # test statistic
[1] 4.564355
```

Compute the critical value at .05 significance level :

```
>alpha=0.05
>z.alpha=qnorm(1-alpha)
>z.alpha
[1]1.644854
```

Inference:-

The test statistic **4.5644** is less than the critical value of **1.6449**. Hence, at .05 significance level, we reject the claims that mean lifetime of a light bulb is above 10,000 hours.

Problem:2

A college conduct both day and night classes intended to be identical. A sample of 100 day students yields examination results as $\overline{X} = 72.4,\ \sigma = 14.8$ and a sample of 200 night students as $\overline{X} = 73.9,\ \sigma = 17.9$. Are the two means statistically equal at 5% level?

Null hypothesis (H_0) : $\mu_1 = \mu_2$

Alternative hypothesis (H_1) : $\mu_1 \neq \mu_2$

We begin with computing the test statistic.

R code:-

```
> x1bar=72.4          # Sample1 mean
> x2bar=73.9          # Sample2 mean
> s1=14.8             # Sample1 SD
> s2=17.9             # Sample2 SD
> n1=100              # Sample1 size
> n2=200              # Sample2 size
> Test_Statistics=abs(((x1bar-x2bar)/sqrt((s1^2/n1)+(s2^2/n2))))
> Test_Statistics
```

[1] 0.7702493

Compute the critical value at .05 significance level:

```
>alpha=0.05
>z.alpha=qnorm(1-alpha/2)
> z.alpha
```

[1] 1.959964

Inference:-

The test statistic **0.7702493** is less than the critical value of **1.959964** at .05 significance level. Hence, we accept H_0. This implies that the two means are statistically equal.

6. Z-TEST FOR SINGLE PROPORTION AND DIFFERENCE OF TWO SAMPLE PROPORTIONS

Aim: To test whether there is significant difference between

(i) sample proportion and population proportion

(ii) two sample proportion.

Problem:1

Suppose that 12% of apples harvested in an orchard last year was rotten. 30 out of 214 apples in a harvest sample this year turns out to be rotten. At .05 significance level, can we reject the null hypothesis that the proportion of rotten apples in harvest stays below 12% this year?

Null hypothesis (H_0) : $p \leq 0.12$
Alternative hypothesis (H_1) : $p > 0.12$
We begin with computing the test statistic,

R-Code:-

```
>pbar=30/214                    # sample proportion
> p0=0.12                       # hypothesized value
> n=214                         # sample size
> z=abs((pbar-p0)/sqrt(p0*(1-p0)/n))
>z                              # test statistic
[1] 0.908751
```

Compute the critical value at .05 significance level :

```
>alpha=0.05
>z.alpha=qnorm(1-alpha)
>z.alpha                  # critical value

[1] 1.644854
```

Inference:-

The test statistic 0.90875 is not greater than the critical value of 1.6449. Hence, at .05 significance level, we do not reject the null hypothesis that the proportion of rotten apples in harvest stays below 12% this year.

Problem:2

Random samples of 400 men and 600 women asked whether they would like to have a fly over near their residence. 200 men and 325 women were in favour of it. Test the equality of men and women in the proposal.

R – Code:-

Null hypothesis (H_0) : p1=p2

Alternative hypothesis (H_1) : p1 ≠ p2

We begin with computing the test statistic,

```
> n1=400                    # Sample1 size
> p1=200/n1                 # Sample1 proportion
> n2=600                    # Sample2 size
> p2=325/n2                 # Sample2 proportion
> P=(n1*p1+n2*p2)/(n1+n2)
> Q=1-P
> z=abs((p1-p2)/sqrt(P*Q*(1/n1+1/n2)))
> z                         # test statistic
[1] 1.292611
```

Compute the critical value at .05 significance level:

```
>alpha=0.05
>z.alpha=qnorm(1-alpha/2)
>z.alpha                    # critical value

[1] 1.959964
```

Inference:-

The test statistic **1.292611** is lesser than the critical value of **1.959964** at .05 significance level. Hence, we accept the null hypothesis. Thus, the men and women are equally favour in the proposal.

7. STUDENT'S t-TEST FOR SINGLE MEAN AND DIFFERENCE OF TWO SAMPLE MEANS

Aim:- To test the significant difference between

(i) sample mean and population mean for one tailed and two tailed tests.
(ii) two sample means

Problem 1:- A random sample of 10 boys has the following IQ's: 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data supports the assumption of population mean IQ of 100.

Null hypothesis (H_0) : $\mu = 100$

Alternative hypothesis (H_1) : $\mu \neq 100$

We begin with computing the test statistic,

R – Code:-

```
> x=c(70,120,110,101,88,83,95,98,107,100)

>t.test(x,mu=100)

        One Sample t-test

data:  x

t = -0.62034, df = 9, p-value = 0.5504

Alternative hypothesis: true mean is not equal to 100

95 percent confidence interval:

  86.98934                    107.41066

sample estimates:

mean of x

    97.2
```

Compute the critical value at .05 significance level:

```
>qt(0.975,9)

[1] 2.262157
```

Inference:-

Since calculated value of t (0.62034) is lesser than the table value of t (2.262157) at 5% LOS with 9 dof. Accept H_0, ie. The mean IQ of the population can be assumed as 100.

Problem 2 :- Comparing two independent sample means, taken from two populations with unknown variance. The following data shows the heights of individuals of two different countries with unknown population variances. Is there any significant difference between the average heights of two groups?

A:	175	168	168	190	156	181	182	175	174	179
B:	120	180	125	188	130	190	110	185	112	188

Null hypothesis (H_0) : $\mu_1 = \mu_2$

Alternative hypothesis (H_1) : $\mu_1 \neq \mu_2$

We begin with computing the test statistic,

R - Code:-

```
> A=c(175,168,168,190,156,181,182,175,174,179)
> A
[1] 175 168 168 190 156 181 182 175 174 179
> B=c(120,180,125,188,130,190,110,185,112,188)
> B
 [1] 120 180 125 188 130 190 110 185 112 188
>t.test(A,B)

Welch Two Sample t-test

data:  A and B
t = 1.8827,      df = 10.224,    p-value = 0.08848
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.95955                    47.95955
sample estimates:
mean of x       mean of y
   174.8          152.8
> qt(0.975,10+10-2)
[1] 2.100922
```

Inference:-

Since calculated value of t (**1.8827**) is lesser than the table value of t (**2.100922**) at 5% LOS with 18 dof. Accept H_0, ie. There is no significant difference between the average heights of two groups.

8. STUDENT t-TEST FOR PAIRED OBSERVATION

Aim: To test whether there is any significant difference paired observations based on one tailed and two tailed test.

Problem: 1 An IQ test was administrated to five persons before and after they were trained. The results are given below:

Candidates :	I	II	III	IV	V
IQ before training	110	120	123	132	125
IQ after training	120	118	125	136	121

Test whether there is any change in IQ after the training programme.

R – Code:-

```
> Before_Training=c(110,120,123,132,125)
> Before_Training
[1] 110 120 123 132 125
> After_Training=c(120,118,125,136,121)
> After_Training
[1] 120 118 125 136 121
> t.test(Before_Training,After_Training,paired=TRUE)

Paired t-test

data:  Before_Training and After_Training
t = -0.8165,     df = 4,           p-value = 0.4601
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.800874       4.800874
sample estimates:
mean of the differences                    -2

> qt(0.975,4)

[1] 2.776445
```

Inference:-

Since calculated value of t (**0.8165**) is lesser than the table value of t (**2.776445**) at 5% LOS with 4 dof. Accept H_0, ie. There is no significant change in IQ due to the training programme.

Problem: 2 Consider the paired data below that represents cholesterol levels on 10men before and after a certain medication

Before(x)	237	289	257	228	303	275	262	304	244	233
After(y)	194	240	230	186	265	222	242	281	240	212

Test the claim that, on average, the drug lowers cholesterol in all men. i.e., test the claim that $\mu_d > 0$. Test this at the 0.05 significance level.

R - Code:-

```
> before=c(237,289,257,228,303,275,262,304,244,233)
> before
 [1] 237 289 257 228 303 275 262 304 244 233
> after=c(194,240,230,186,265,222,252,281,240,212)
> before
[1] 237 289 257 228 303 275 262 304 244 233
> t.test(before,after, paired=TRUE,alternative="greater",mu=0)

Paired t-test

data:  before and after
t = 5.9151,      df = 9,          p-value = 0.0001124
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 21.3929     Inf
sample estimates:
mean of the differences
                    31
> qt(0.975,9)
[1] 2.262157
```

Inference:-

Since calculated value of t (5.9151) is greater than the table value of t (2.262157) at 5% LOS with 9 dof.We can reject the null hypothesis and support the claim.

9. F- TEST OF EQUALITY OF VARIANCES

Aim: To test whether there is significant difference between population variances.

Problem: Five Measurements of the output of two units have given the following results (in kilograms of material per one hour of operation) .Assume that both samples have been obtained from normal populations, test at 5% significance level if two populations have the same variance.

Unit A	14.1	10.1	14.7	13.7	14.0
Unit B	14.0	14.5	13.7	12.7	14.1

$H_0: S_1^2 = S_2^2$

$H1: S_1^2 \neq S_2^2$

Level of Significance: 0.05

R-Code:-

```
> Unit_A=c(14.1,10.1,14.7,13.7,14.0)
> Unit_A
[1] 14.1 10.1 14.7 13.7 14.0
> Unit_B=c(14.0,14.5,13.7,12.7,14.1)
> Unit_B
[1] 14.0 14.5 13.7 12.7 14.1
> var.test(Unit_A,Unit_B)
F test to compare two variances
data:  Unit_A and Unit_B
F = 7.3304, num df = 4, denom df = 4, p-value = 0.07954
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.7632268 70.4053799
sample estimates:
ratio of variances
        7.330435
> qt(0.975,4,4)
[1] 12.3232
```

Inference:

Since calculated value of F (**7.3304**) is lessser than the table value of F(**12.3232**) at 5% LOS with 9 dof. Hence, we accept the null hypothesis.

10. ANALYSIS OF VARIANCE - ONE WAY CLASSIFICATION

Aim:- To test whether there is significant difference between the row means (column means).

Problem: As head of a department of a consumer's research organization we have the responsibility of testing and comparing life times of four brands of electric bulbs. Suppose we test the lifetime of three electric bulbs each of 4 brands, the data is given below, each entry representing the lifetime of an eclectic bulb, measured in hundreds of hours.

A	B	C	D
38	40	41	39
45	42	49	36
40	38	42	42

Can we infer that the mean lifetime of the four brands of electric bulbs are equal?

Null Hypothesis (H_0): $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (I.e. Average lifetimes of four different bulbs are same)

Alternate Hypothesis (H_1): $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$ (I.e. Average lifetimes of four different bulbs are different)

R – Code:-

```
>data
[1] 38 45 40 40 42 38 41 49 42 39 36 42
> brands=c("A","A","A","B","B","B","C","C","C","D","D","D")
>brands
[1] "A" "A" "A" "B" "B" "B" "C" "C" "C" "D" "D" "D"
>One_Way_Anova=aov(data~brands)
>summary(One_Way_Anova)
            Df    Sum Sq        Mean Sq     F value           Pr(>F)
brands      3     42       14.00            1.244       0.356
Residuals   8     90       11.25
>qf(.95, df1=3, df2=8)
[1] 4.066181
```

Inference:-

Since calculated value of F (**1.244**) is lesser than the table value of F (**4.066181**) at 5% LOS with dof (3,8). We accept H_0. The mean life times of the four electric bulbs are equal.

11. ANALYSIS OF VARIANCE - TWO WAY CLASSIFICATION

Aim:- To test whether there is significant difference between the row means as well as column means.

Problem: The following data represent the number of units of production per day turned out by different workers using 4 different types of machines.

	A	B	C	D
1	44	38	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	49	39

Test whether the five men differ with respect to mean productivity and test whether the mean productivity is the same for the four different machine types.

Null Hypothesis (H_0) : There is no significant difference between men and machines.
Alternate Hypothesis (H_1) : There is significant difference between men and machines.

R – Code:-

```
> data=c(44,38,47,36,46,40,52,43,34,36,44,32,43,38,46,33,38,42,49,39)
>data
[1] 44 38 47 36 46 40 52 43 34 36 44 32 43 38 46 33 38 42 49 39
>workers=gl(5,4)
>workers
 [1] 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5
Levels: 1 2 3 4 5
>machine=gl(4,1,20)
>machine
[1] 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4
Levels: 1 2 3 4
>Twoway_Anova=aov(data~workers+machine)
>summary(Twoway_Anova)
            Df   Sum Sq   Mean Sq   F value   Pr(>F)
workers     4    161.5    40.37     6.574     0.00485 **
machine     3    338.8    112.93    18.388    8.78e-05 ***
Residuals   12   73.7     6.14
>qf(.95, df1=4, df2=12)
[1] 3.259167

>qf(.95, df1=3, df2=12)
[1] 3.490295
```

Inference:-

Since calculated values of F (6.574 & 18.388) are greater than the table value of F (3.259167&3.490295) at 5% LOS with dof(**4,12**) & dof(**3,12**). We reject H_0. There is significance difference between men and machines.

12. ANALYSIS OF VARIANCE - LATIN SQUARE DESIGN

Aim:- To test whether there is significant difference between the row means, column means and treatments.

Problem: Four varieties A, B, C, D of a crop are tested in a Latin Square design with four replications. The plot yields are given in kgs as below:

A 6	C 5	D 6	B 9
C 8	A 4	B 6	D 4
B 7	D 6	C 10	A 6
D 7	B 4	A 8	C 9

Analyze the experimental yield and state your conclusion.

Null Hypothesis (H_0) : All the varieties are equal.

Alternate Hypothesis (H_1) : All the varieties are not equal.

R – Code:-

```
> crops=c(rep("crop1",1), rep("crop2",1), rep("crop3",1), rep("crop4",1))
>crops
[1] "crop1" "crop2" "crop3" "crop4"
> treat <- c(rep("treatA",4), rep("treatB",4), rep("treatC",4), rep("treatD",4))
> treat
 [1] "treatA" "treatA" "treatA" "treatA" "treatB" "treatB" "treatB" "treatB" "treatC" "treatC" "treatC" "treatC"
"treatD" "treatD" "treatD" "treatD"
> yields=c("A","C","D","B","C","A","B","D","B","D","C","A","D","B","A","C")
>yields
 [1] "A" "C" "D" "B" "C" "A" "B" "D" "B" "D" "C" "A" "D" "B" "A" "C"
> freq=c(6,5,6,9,8,4,6,4,7,6,10,6,7,4,8,9)
>freq
 [1]  6  5  6  9  8  4  6  4  7  6 10  6  7  4  8  9
>mydata=data.frame(treat, crops, yields, freq)
>myfitdata=lm(freq ~ crops+treat+yields, mydata)
> LSD=anova(myfitdata)
> LSD
Analysis of Variance Table
Response: freq
                  Df        Sum Sq      Mean Sq     F value     Pr(>F)
crops    3        18.1875   6.0625      2.5304      0.1536
treat    3        7.1875    2.3958      1.0000      0.4547
yields   3        12.1875   4.0625      1.6957      0.2663
Residuals         6      14.3750     2.3958
>qf(.95, df1=6, df2=3)
[1] 8.940645
```

Inference:-

Since calculated values of F (**2.5304, 1.0000 & 1.6957**) are lesser than the table value of F (**8.940645**) at 5% LOS with dof (**6,3**). We accept H_0. There is no significant difference between the row means, column means and treatments.

---*****---