Practical No. 1
Aim:- Practical of Data collection, Data curation and management for Unstructured data (NoSQL)

Steps:

1. Download CouchDB: https://couchdb.apache.org/ (Select Option ***Windows (x64)***)

2. Install CouchDB. While Installing set Username as ***bnb*** and Password as ***bnb***.

3. To Open CouchDB enter following Link on address bar of browser: http://localhost:5984/_utils/#login

4. Enter Username and Password for login.

5. Download R and RStudio: https://posit.co/download/rstudio-desktop/

6. Install R.

7. Install Rstudio.

R Code with Output:

```
> install.packages('sofa')
> library('sofa')
> #create connection object
> x<-Cushion$new(user='bnb', pwd='bnb')
> x
> #Write Output Here.
>
> #to check whether object created
> x$ping()
> #Write Output Here.
>
> #create database tycs
> db_create(x,dbname = 'tycs')
> #Write Output Here.
>
> db_list(x)
> #Write Output Here.
>
> #create json doc
> doc1 = '{"rollno":"01","name":"ABC","GRADE":"A"}'
> doc_create(x,doc1,dbname = "tycs",docid = "a_1")
> #Write Output Here.
>
> doc2 = '{"rollno":"02","name":"PQR","GRADE":"A"}'
> doc_create(x,doc2,dbname = "tycs",docid = "a_2")
> #Write Output Here.
>
> doc3 = '{"rollno":"03","name":"XYZ","GRADE":"B","REMARK":"PASS"}'
> doc_create(x,doc3,dbname = "tycs",docid = "a_3")
> #Write Output Here.
>
> #CHANGES FEED
> db_changes(x,"tycs")
> #Write Output Here.
>
> #search for id > null so all docs will display
> db_query(x,dbname = "tycs", selector =
list('_id'=list('$gt'=NULL)))$docs
```

```
>
> #search for students with grade is A
> db_query(x,dbname = "tycs",selector = list(GRADE="A"))$docs
> #Write Output Here.
>

> #search for students with remark =pass
> db_query(x,dbname = "tycs",selector = list(REMARK="PASS"))$docs
> #Write Output Here.
>
> #return only certain fields where rollno>2
> db_query(x,dbname = "tycs",selector =
list(rollno=list('$gt'='02')),fields=c("name","GRADE"))$docs
> #Write Output Here.
>
> #convert the result of a query into a data frame using jsonlite
> install.packages(“jsonlite”)
> library("jsonlite")
> res = db_query(x,dbname = "tycs",selector =
list('_id'=list('$gt'=NULL)),fields=c("name","rollno","GRADE","REMARK"),as
="json")
> #display json doc
> fromJSON(res)$docs
> #Write Output Here.
>
> doc_delete(x,dbname = "tycs",docid = "a_2")
> #Write Output Here.
>
> doc_get(x,dbname = "tycs",docid = "a_2")
> #Write Output Here.
>
> doc2 = '{"name":"Sdrink","beer":"TEST","note":"yummy","note2":"yay"}'
> doc_update(x,dbname = "ty",doc=doc2,docid="a_3",rev = "3-
b1fb56db955b142c6efd3b3c52fe9e1b")
> #Write Output Here.
>
> doc3 = '{"rollno":"01", "name":"UZMA", "GRADE":"A"}'
> doc_update(x,dbname = "tycs",doc=doc3,docid = "a_1",rev = "1-
be7c98bddf8ea7c46f4f401ff387593d")
> #Write Output Here.
```

Practical No. 8

Aim:- Practical of Hypothesis testing

Small Sample Test:

a) t – test for single population mean:

Example 1:-

The random sample of 10 boys had following IQ 70, 120, 110, 101, 88, 83, 95, 89, 107, 125. Do this data support the assumption that population mean IQ is 100?

R – Code:-

H_0: Population mean IQ is 100.

H_1:Population mean IQ is not 100.

```
> x=c(70, 120, 110, 101, 88, 83, 95, 89, 107, 125)
> Result=t.test(x)
> print(Result)
          One Sample t-test
data:  x
t = 18.244, df = 9, p-value = 2.039e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  86.54903 111.05097
sample estimates:
mean of x 
     98.8 
```

Conclusion:- Here p-value is less than LOS(5%). So we Reject Null Hypothesis.

b) Difference between two sample means:

Example 1:-

Two groups of 10 subjects each were given the digit span subtest from the Wechsler Adult Intelligence Scale. One group consisted of regular smokers of marijuana, while the other group consisted of nonsmokers. The scores are given below:

Nonsmokers		22	21	17	20	17	23	20	22	21
Smokers	16	20	14	21	20	18	13	15	17	21

Test the hypothesis that both there is no significant effect on score due to smoking.

R – Code:

H_0: There is no significant effect on score due to smoking

H_1: There is no significant effect on score due to smoking

```
> nonsmokers=c(18,22,21,17,20,17,23,20,22,21)
> smokers=c(16,20,14,21,20,18,13,15,17,21)
> Result=t.test(nonsmokers,smokers)
> print(Result)
          Welch Two Sample t-test
data:  nonsmokers and smokers
t = 2.2573, df = 16.376, p-value = 0.03798
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1628205 5.0371795
sample estimates:
mean of x mean of y 
     20.1      17.5 
```

Conclusion:-

c) Paired t – test:

Example 1:-

An IQ test was administrated to 5 persons before and after they were trained. The results are given below:

Before	110	120	123	132	125
After	120	118	125	136	121

R – Code:

H_0:

H_1:

```
> x=c(110,120,123,132,125)
> y=c(120,118,125,136,121)
> Result=t.test(x,y,paired = TRUE,alternative = "less")
> print(Result)
          Paired t-test
data:  x and y
t = -0.8165, df = 4, p-value = 0.23
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 3.221937
sample estimates:
mean of the differences 
                     -2 
```

Conclusion:-

Example 2:-

School athletics has taken a new instructor, and want to test the effectiveness of the new type of training proposed by comparing the average times of 10 runners in the 100 meters. The time in seconds before and after training for each athlete are given below:

Before Training	12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3
After Training	12.0, 12.2, 11.2, 13.0, 15.0, 15.8, 12.2, 13.4, 12.9, 11.0

R – Code:-

H_0:

H_1:

```
> x=c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3)
> y=c(12.0, 12.2, 11.2, 13.0, 15.0, 15.8, 12.2, 13.4, 12.9, 11.0)
> Result=t.test(x,y,paired = TRUE,alternative = "greater")
> print(Result)
          Paired t-test
data:  x and y
t = 5.2671, df = 9, p-value = 0.0002579
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.049675      Inf
sample estimates:
mean of the differences 
                   1.61 
```

Conclusion:-

Practical No. 9

Aim:- Practical of Analysis of Variance

Program:

Analysis of Variance (ANOVA) is a commonly used statistical technique for investigating data by comparing the means of subsets of the data. In **One-Way ANOVA** the data is subdivided into groups based on a single classification factor and the standard terminology used to describe the set of factor levels is treatment even though this might not always having meaning for the particular application.

R provides two commands **Oneway.test ()** and **aov ()** for One-way ANOVA.

Example 1:

The following data gives effect of three treatments.

A	2, 3, 7, 2, 6
B	10, 8, 7, 5, 10
C	10, 13, 14, 13, 15

Test the hypothesis that all treatments are equally effective.

R Code:-

H_0:

H_1:

```
> Group1 = c(2,3,7,2,6)
> Group2 = c(10,8,7,5,10)
> Group3 = c(10,13,14,13,15)
> Combined_Group = data.frame(Group1,Group2,Group3)
> Combined_Group
  Group1 Group2 Group3
1      2     10     10
2      3      8     13
3      7      7     14
4      2      5     13
5      6     10     15
> Stacked_Group = stack(Combined_Group)
> Stacked_Group
   values    ind
1       2 Group1
2       3 Group1
3       7 Group1
4       2 Group1
5       6 Group1
6      10 Group2
7       8 Group2
8       7 Group2
9       5 Group2
10     10 Group2
11     10 Group3
12     13 Group3
13     14 Group3
14     13 Group3
15     15 Group3
> Result = aov(values~ind,data=Stacked_Group)
> summary(Result)
            Df Sum Sq Mean Sq F value   Pr(>F)    
ind          2  203.3   101.7   22.59 8.54e-05 ***
Residuals   12   54.0     4.5                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion:-

Example 2:-

The following data gives life of tires of four brands.

A	20, 23, 18, 17, 18, 22, 24
B	19, 15,17, 20, 16, 17
C	21, 19, 22, 17, 20
D	15, 14, 16, 18, 14, 16

Test the hypothesis that average life for each brand is same.

R Code:-

H_0:

H_1:

```
> x1=c(20,23,18,17,18,22,24)
> x2=c(19,15,17,20,16,17)
> x3=c(21,19,22,17,20)
> x4=c(15,14,16,18,14,16)
> Combined_Group=list(b1=x1,b2=x2,b3=x3,b4=x4)
> Stacked_Group=stack(Combined_Group)
> Stacked_Group
   values ind
1      20  b1
2      23  b1
3      18  b1
4      17  b1
5      18  b1
6      22  b1
7      24  b1
8      19  b2
9      15  b2
10     17  b2
11     20  b2
12     16  b2
13     17  b2
14     21  b3
15     19  b3
16     22  b3
17     17  b3
18     20  b3
19     15  b4
20     14  b4
21     16  b4
22     18  b4
23     14  b4
24     16  b4
> Result=aov(values~ind,data = Stacked_Group)
> summary(Result)
            Df Sum Sq Mean Sq F value  Pr(>F)
ind          3  91.44  30.479   6.845 0.00235 **
Residuals   20  89.06   4.453
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion:-

Two-Way ANOVA:- Two-Way ANOVA is used to compare the means of populations that are classified in two different ways, or mean responses in an experiment with two factors without interaction. We fit two-way ANOVA models in R using the function aov ().
aov(Response ~ FactorA + FactorB)

Example3:-
A tea company appoints four salesmen A, B, C and D and observes their sales in three seasons – summer, winter and monsoon. The figures (in lakhs) of sales are given in the following table:

Season	**Salesmen**			
	A	**B**	**C**	**D**
Summer	36	32	21	30
Winter	24	25	20	22
Monsoon	20	18	19	15

(i) Do the salesmen significantly differ in performance?
(ii) Is there significant difference between the season?

R - Code:-

H_{01}:

H_{11}:

H_{02}:

H_{12}:

```
> sales=c(36,32,21,30,24,25,20,22,20,18,19,15)
> f1=c(rep(1:3,rep(4,3)))
> f2=rep(c("A","B","C","D"),3)
> season=factor(f1)
> salesmen=factor(f2)
> Result=aov(sales~season+salesmen)
> summary(Result)
            Df Sum Sq Mean Sq F value  Pr(>F)   
season       2 279.50  139.75  11.673 0.00855 **
salesmen     3  77.67   25.89   2.162 0.19358   
Residuals    6  71.83   11.97                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion:-

Example2:-

Five different fertilizers are used and three types of seeds are sown. The yield obtained (in kgs) is tabulated below:

	Fert I	**Fert II**	**Fert III**	**Fert IV**	**Fert V**
Seed A	110	100	107	104	102
Seed B	112	99	101	112	107
Seed C	97	87	99	101	98

Carry out ANOVA to test the significance between types of seeds and fertilizers used.

R – Code:-

H_{01}:

H_{11}:

H_{02}:

H_{12}:

```
> yield=c(110,100,107,104,102,112,99,101,112,107,97,87,99,101,98)
> f1=c(rep(1:3,rep(5,3)))
> f2=rep(c(1:5),3)
> sed=factor(f1)
> seed=factor(f1)
> fertilizer=factor(f2)
> Result=aov(yield~seed+fertilizer)
> summary(Result)
            Df Sum Sq Mean Sq F value  Pr(>F)
seed         2  276.4  138.20  10.954 0.00512 **
fertilizer   4  228.3   57.07   4.523 0.03335 *
Residuals    8  100.9   12.62
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion:-