



SRI RAMACHANDRA
INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Category - I Deemed to be University) Porur, Chennai
SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

CSC 525
BIG DATA ANALYTICS

Submitted by

BRINDA G – E7321004

MASTER OF SCIENCE
In
BIG DATA ANALYTICS

Sri Ramachandra Faculty of Engineering and Technology
Sri Ramachandra Institute of Higher Education and Research, Porur,
Chennai -600116

JUNE, 2022

BONAFIDE CERTIFICATE

Certified that this project report is the bonafide record of work done by "**BRINDA G**
– E7321004”

Signature of the Course Faculty

Name of Course Faculty

Assistant Professor,

Department of Computer Science and Engineering

Sri Ramachandra Faculty of Engineering and Technology,

SRIHER, Porur, Chennai-600 116.

Signature of Vice-Principal

Prof. M. Prema

Vice-Principal,

Department of Computer Science and Engineering

Sri Ramachandra Faculty of Engineering and Technology,

SRIHER, Porur, Chennai-600 116.

Evaluation Date:



SRI RAMACHANDRA

INSTITUTE OF HIGHER EDUCATION AND RESEARCH

(Category - I Deemed to be University) Porur, Chennai

SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

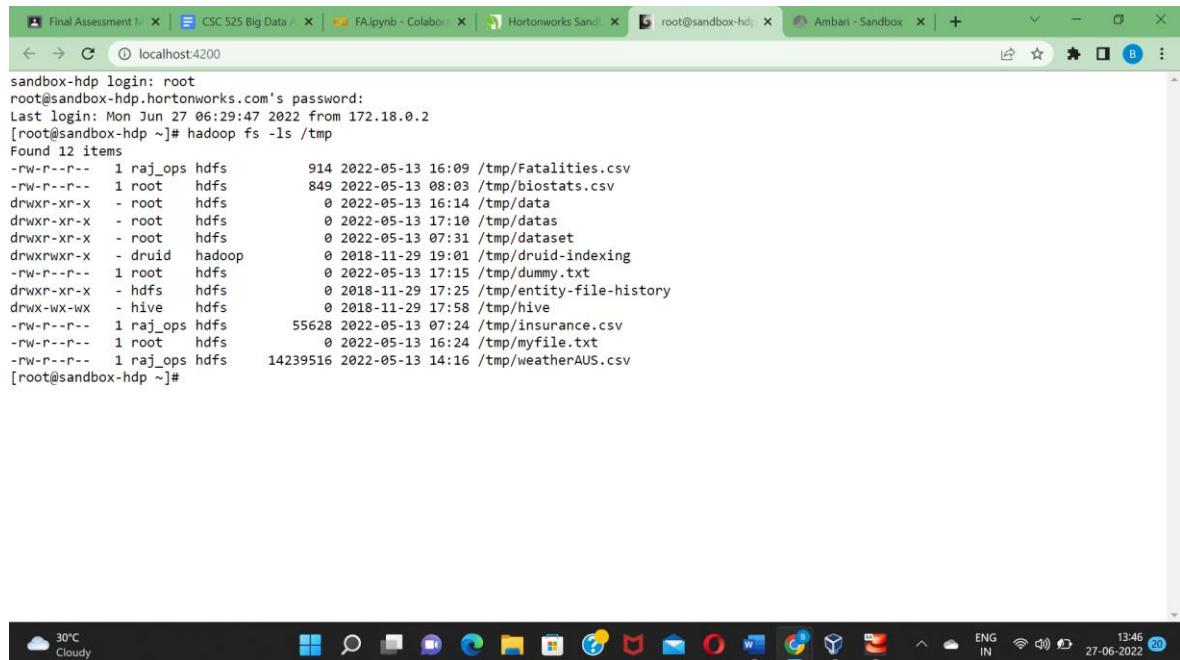
TABLE OF CONTENTS

QUESTION NO.	PAGE NO.	CO's
1	4	CO1
2	14	CO2
3	19	CO3
4	35	CO4
5	56	CO5

1. Implement the hadoop architecture in your local machine and perform the following tasks to analyse the streaming dataset.

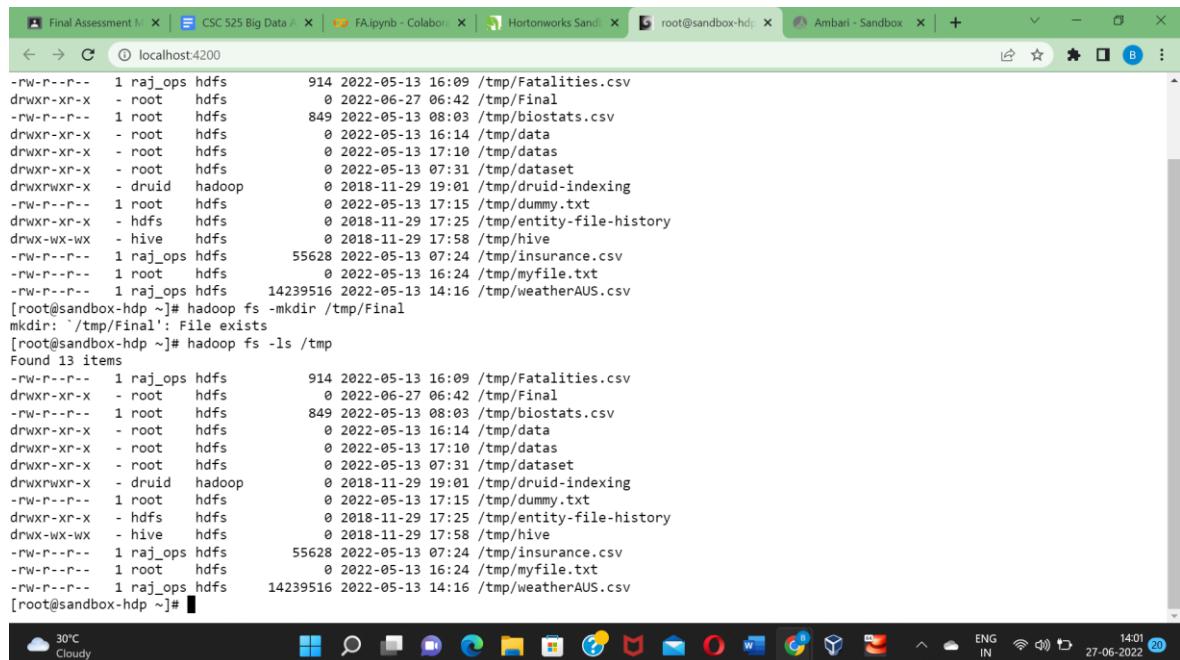
A. Perform the following Hadoop commands to check the feasibility of the Hadoop environment configured.

- List all the files in the data directory.



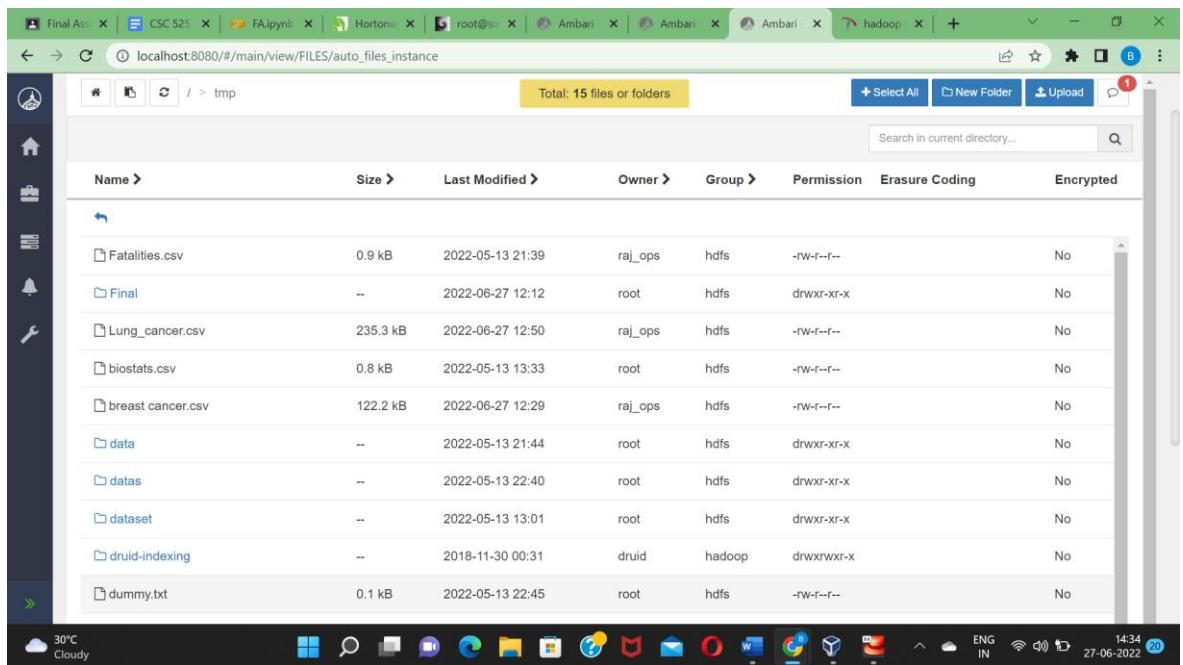
```
Final Assessment M X CSC 525 Big Data A X FA.ipynb - Colabor X Hortonworks Sand X root@sandbox-hdp: ~# Ambari - Sandbox X + localhost:4200
[sandbox-hdp login: root
root@sandbox-hdp.hortonworks.com's password:
Last login: Mon Jun 27 06:29:47 2022 from 172.18.0.2
[root@sandbox-hdp ~]# hadoop fs -ls /tmp
Found 12 items
-rw-r--r-- 1 raj_ops hdfs 914 2022-05-13 16:09 /tmp/Fatalities.csv
-rw-r--r-- 1 root hdfs 849 2022-05-13 08:03 /tmp/biostats.csv
drwxr-xr-x - root hdfs 0 2022-05-13 16:14 /tmp/data
drwxr-xr-x - root hdfs 0 2022-05-13 17:10 /tmp/datas
drwxr-xr-x - root hdfs 0 2022-05-13 07:31 /tmp/dataset
drwxrwxr-x - druid hadoop 0 2018-11-29 19:01 /tmp/druid-indexing
-rw-r--r-- 1 root hdfs 0 2022-05-13 17:15 /tmp/dummy.txt
drwxr-xr-x - hdfs hdfs 0 2018-11-29 17:25 /tmp/entity-file-history
drwxrwx-wx - hive hdfs 0 2018-11-29 17:58 /tmp/hive
-rw-r--r-- 1 raj_ops hdfs 55628 2022-05-13 07:24 /tmp/insurance.csv
-rw-r--r-- 1 root hdfs 0 2022-05-13 16:24 /tmp/myfile.txt
-rw-r--r-- 1 raj_ops hdfs 14239516 2022-05-13 14:16 /tmp/weatherAUS.csv
[root@sandbox-hdp ~]# ]
```

- Create a directory called final



```
Final Assessment M X CSC 525 Big Data A X FA.ipynb - Colabor X Hortonworks Sand X root@sandbox-hdp: ~# Ambari - Sandbox X + localhost:4200
[sandbox-hdp login: root
root@sandbox-hdp.hortonworks.com's password:
Last login: Mon Jun 27 06:29:47 2022 from 172.18.0.2
[root@sandbox-hdp ~]# hadoop fs -mkdir /tmp/Final
mkdir: '/tmp/Final': File exists
[root@sandbox-hdp ~]# hadoop fs -ls /tmp
Found 13 items
-rw-r--r-- 1 raj_ops hdfs 914 2022-05-13 16:09 /tmp/Fatalities.csv
drwxr-xr-x - root hdfs 0 2022-06-27 06:42 /tmp/Final
-rw-r--r-- 1 root hdfs 849 2022-05-13 08:03 /tmp/biostats.csv
drwxr-xr-x - root hdfs 0 2022-05-13 16:14 /tmp/data
drwxr-xr-x - root hdfs 0 2022-05-13 17:10 /tmp/datas
drwxr-xr-x - root hdfs 0 2022-05-13 07:31 /tmp/dataset
drwxrwxr-x - druid hadoop 0 2018-11-29 19:01 /tmp/druid-indexing
-rw-r--r-- 1 root hdfs 0 2022-05-13 17:15 /tmp/dummy.txt
drwxr-xr-x - hdfs hdfs 0 2018-11-29 17:25 /tmp/entity-file-history
drwxrwx-wx - hive hdfs 0 2018-11-29 17:58 /tmp/hive
-rw-r--r-- 1 raj_ops hdfs 55628 2022-05-13 07:24 /tmp/insurance.csv
-rw-r--r-- 1 root hdfs 0 2022-05-13 16:24 /tmp/myfile.txt
-rw-r--r-- 1 raj_ops hdfs 14239516 2022-05-13 14:16 /tmp/weatherAUS.csv
[root@sandbox-hdp ~]# ]
```

- Load csv file from local to data directory.



```

Hortonworks Sandbox | root@sandbox-hdp- | Ambari - Sandbox | root@sandbox-hdp- | + | - | X
localhost:4200 | Apps | Gmail | YouTube | Maps | ENG IN | 14:34 | 27-06-2022 | 20 | : |

25912 -- process information unavailable
30585 DataNode
12537 JobHistoryServer
12860 Kafka
6653 EmbeddedServer
[root@sandbox-hdp ~]# wget https://people.sc.fsu.edu/~jburkardt/data/csv/biostats.csv
--2022-05-13 08:03:14-- https://people.sc.fsu.edu/~jburkardt/data/csv/biostats.csv
Resolving people.sc.fsu.edu (people.sc.fsu.edu)... 144.174.16.102
Connecting to people.sc.fsu.edu (people.sc.fsu.edu)|144.174.16.102|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 849 [text/csv]
Saving to: 'biostats.csv'

100%[=====] 849 --.-K/s in 0s

2022-05-13 08:03:16 (62.9 MB/s) - 'biostats.csv' saved [849/849]

[root@sandbox-hdp ~]# hadoop fs -copyFromLocal biostats.csv /tmp
[root@sandbox-hdp ~]# hadoop fs -ls /tmp
Found 6 items
-rw-r--r-- 1 root    hdfs      849 2022-05-13 08:03 /tmp/biostats.csv
drwxr-xr-x  - root    hdfs      0 2022-05-13 07:31 /tmp/dataset
drwxrwxr-x  - druid   hadoop     0 2018-11-29 19:01 /tmp/druid-indexing
drwxr-xr-x  - hdfs    hdfs      0 2018-11-29 17:25 /tmp/entity-file-history
drwxrwx-wx  - hive    hdfs      0 2018-11-29 17:58 /tmp/hive
-rw-r--r--  1 raj_ops hdfs    55628 2022-05-13 07:24 /tmp/insurance.csv
[root@sandbox-hdp ~]#

```

- Remove a particular file from tmp directory.

```
Final Assesme... x CSC 525 Big D... x FAipynb - Col... x Hortonworks S... | Ambari - Sand... x | Ambari - Sand... x root@sandbox: ~ + _ -
```

localhost:4200

```
drwxr-xr-x - hdfs hdfs 0 2018-11-29 17:25 /tmp/entity-file-history  
drwxrwxr-x - hive hdfs 0 2018-11-29 17:58 /tmp/hive  
-rw-r--r-- 1 raj_ops hdfs 55628 2022-06-27 07:41 /tmp/insurance.csv  
-rw-r--r-- 1 raj_ops hdfs 14239516 2022-05-13 14:16 /tmp/weatherAUS.csv  
[root@sandbox-hdp ~]# hadoop fs -cp /tmp/Lung_cancer.csv /tmp/Final  
[root@sandbox-hdp ~]# hadoop fs -ls /tmp/Final  
Found 1 items  
-rw-r--r-- 1 root hdfs 240946 2022-06-27 09:15 /tmp/Final/Lung_cancer.csv  
[root@sandbox-hdp ~]# hadoop fs -cp /tmp/Lung_cancer.csv /tmp/Data  
[root@sandbox-hdp ~]# hadoop fs -ls /tmp/Data  
-rw-r--r-- 1 root hdfs 240946 2022-06-27 09:17 /tmp/Data  
[root@sandbox-hdp ~]# hadoop fs -rm /tmp/insurance.csv  
22/06/27 09:15:25 INFO fs.TrashPolicyDefault: Moved: 'hdfs://sandbox-hdp.hortonworks.com:8020/tmp/insurance.csv' to trash at: hdfs://sandbox-hdp.hortonworks.com:8020/user/root/.Trash/Current/tmp/insurance.csv  
[root@sandbox-hdp ~]# hadoop fs -ls /tmp  
Found 14 items  
-rw-r--r-- 1 root hdfs 240946 2022-06-27 09:17 /tmp/Data  
-rw-r--r-- 1 raj_ops hdfs 914 2022-05-13 16:09 /tmp/Fatalities.csv  
drwxr-xr-x - root hdfs 0 2022-06-27 09:15 /tmp/Final  
-rw-r--r-- 1 raj_ops hdfs 81066 2022-06-27 07:31 /tmp/Google_Stock_Price_Train.csv  
-rw-r--r-- 1 raj_ops hdfs 240946 2022-06-27 09:13 /tmp/Lung_Cancer.csv  
-rw-r--r-- 1 raj_ops hdfs 125141 2022-06-27 06:59 /tmp/breast_cancer.csv  
drwxr-xr-x - root hdfs 0 2022-05-13 16:14 /tmp/data  
drwxr-xr-x - root hdfs 0 2022-05-13 17:10 /tmp/datas  
drwxr-xr-x - root hdfs 0 2022-05-13 07:31 /tmp/dataset  
drwxrwxr-x - druid hadoop 0 2018-11-29 19:01 /tmp/druid-indexing  
-rw-r--r-- 1 root hdfs 0 2022-05-13 17:15 /tmp/dummy.txt  
drwxr-xr-x - hdfs hdfs 0 2018-11-29 17:25 /tmp/entity-file-history  
drwxrwxr-x - hive hdfs 0 2018-11-29 17:58 /tmp/hive  
-rw-r--r-- 1 raj_ops hdfs 14239516 2022-05-13 14:16 /tmp/weatherAUS.csv  
[root@sandbox-hdp ~]#
```

- Display the content present in the dibetics.csv file.

```
Final Ans x | CSC 525 B x | FAJpynb x | Hortonwo x | Ambari - S x | Ambari - S x | root@sand x | Diabetes D x | + | localhost:4200
drwxr-xr-x - root hdfs 0 2022-05-13 17:10 /tmp/datas
drwxr-xr-x - root hdfs 0 2022-05-13 07:31 /tmp/dataset
drwxrwxr-x - druid hadoop 0 2018-11-29 19:01 /tmp/druid-indexing
-rw-r--r-- 1 root hdfs 0 2022-05-13 17:15 /tmp/dummy.txt
drwxr-xr-x - hdfs hdfs 0 2018-11-29 17:25 /tmp/entity-file-history
drwxrwx-wx - hive hdfs 0 2018-11-29 17:58 /tmp/hive
-rw-r--r-- 1 raj_ops hdfs 14239516 2022-05-13 14:16 /tmp/weatherAUS.csv
[root@sandbox-hdp ~]# hadoop fs -ls /tmp/dibetics.csv
-rw-r--r-- 1 raj_ops hdfs 23873 2022-06-27 09:26 /tmp/dibetics.csv
[root@sandbox-hdp ~]# hadoop fs -cat /tmp/dibetics.csv
Pregnancies,Glucose,BloodPressure,SkinThickness,Insulin,BMI,DiabetesPedigreeFunction,Age,Outcome
6,148,72,35,0,33.6,0.627,58,1
1,85,66,29,0,26.6,0.351,31,0
8,183,64,0,0,23.3,0.672,32,1
1,89,66,23,94,28,1,0.167,21,0
0,137,40,35,168,43,1,2.288,33,1
5,116,74,0,0,25.6,0.201,30,0
3,78,50,32,88,31,0.248,26,1
10,115,0,0,0,35,3,0.134,29,0
2,197,70,45,543,30,5,0.158,53,1
8,125,96,0,0,0,0.232,54,1
4,110,92,0,0,0,37,6,0.191,30,0
10,168,74,0,0,38,0.537,34,1
10,139,80,0,0,27,1,1.441,57,0
1,189,60,23,846,30,1,0.398,59,1
5,166,72,19,175,25,8,0.587,51,1
7,100,0,0,0,30,0,0.484,32,1
0,118,84,47,230,45,8,0.551,31,1
7,107,74,0,0,29,6,0.254,31,1
1,103,38,38,83,43,3,0.183,33,0
1,115,70,30,96,34,6,0.529,32,1
3,126,88,41,235,39,3,0.704,27,0
```

```

5,117,66,50,105,59,1,0,251,42,0
1,111,94,0,0,32,8,0,265,45,0
4,112,78,40,0,39,4,0,236,38,0
1,116,78,29,180,36,1,0,496,25,0
0,141,84,26,0,32,4,0,433,22,0
2,175,88,0,0,22,9,0,326,22,0
2,92,52,0,0,30,1,0,141,22,0
3,130,78,23,79,28,4,0,323,34,1
8,120,86,0,0,28,4,0,259,22,1
2,174,88,37,120,44,5,0,646,24,1
2,106,56,27,165,29,0,426,22,0
2,185,75,0,0,23,3,0,56,53,0
4,95,60,32,0,35,4,0,284,28,0
0,126,86,27,120,27,4,0,515,21,0
8,65,72,23,0,32,0,6,42,0
2,99,60,17,160,36,6,0,453,21,0
1,102,74,0,0,39,5,0,293,42,1
11,120,80,37,150,42,3,0,785,48,1
3,182,44,20,94,30,8,0,4,26,0
1,189,58,18,116,28,5,0,219,22,0
9,140,94,0,0,32,7,0,734,45,1
13,153,88,37,140,40,6,1,174,39,0
12,100,84,33,105,30,0,488,46,0
1,147,94,41,0,49,3,0,358,27,1
1,81,74,41,57,46,3,1,096,32,0
3,187,70,22,200,36,4,0,408,36,1
6,162,62,0,0,24,3,0,178,50,1
4,136,70,0,0,31,2,1,182,22,1
1,121,78,39,74,39,0,261,28,0
3,108,62,24,0,26,0,223,25,0
0,181,88,44,510,43,3,0,222,26,1
8,154,78,32,0,32,4,0,443,45,1

```

- Create an empty file called myfile.txt

```

< -> C localhost:4200
Found 16 items
-rw-r--r-- 1 raj_ops hdfs 240946 2022-06-27 09:13 /tmp/Lung_cancer.csv
-rw-r--r-- 1 raj_ops hdfs 125141 2022-06-27 06:59 /tmp/breast_cancer.csv
drwxr-xr-x - root hdfs 0 2022-05-13 16:14 /tmp/data
drwxr-xr-x - root hdfs 0 2022-05-13 17:10 /tmp/datas
drwxr-xr-x - root hdfs 0 2022-05-13 07:31 /tmp/dataset
-rw-r--r-- 1 raj_ops hdfs 23873 2022-06-27 09:26 /tmp/dibetics.csv
drwxrwxr-x - druid hadoop 0 2018-11-29 19:01 /tmp/druid-indexing
-rw-r--r-- 1 root hdfs 0 2022-05-13 17:15 /tmp/dummy.txt
drwxr-xr-x - hdfs hdfs 0 2018-11-29 17:25 /tmp/entity-file-history
drwxrwx-wx - hive hdfs 0 2018-11-29 17:58 /tmp/hive
-rw-r--r-- 1 raj_ops hdfs 14239516 2022-05-13 14:16 /tmp/weatherAUS.csv
[root@sandbox-hdp ~]# hadoop fs -touch /tmp/myfile.txt
[root@sandbox-hdp ~]# hadoop fs -ls /tmp
Found 16 items
-rw-r--r-- 1 root hdfs 240946 2022-06-27 09:17 /tmp/Data
-rw-r--r-- 1 raj_ops hdfs 914 2022-05-13 16:09 /tmp/Fatalities.csv
drwxr-xr-x - root hdfs 0 2022-06-27 09:15 /tmp/Final
-rw-r--r-- 1 raj_ops hdfs 81066 2022-06-27 07:31 /tmp/Google_Stock_Price_Train.csv
-rw-r--r-- 1 raj_ops hdfs 240946 2022-06-27 09:13 /tmp/Lung_cancer.csv
-rw-r--r-- 1 raj_ops hdfs 125141 2022-06-27 06:59 /tmp/breast_cancer.csv
drwxr-xr-x - root hdfs 0 2022-05-13 16:14 /tmp/data
drwxr-xr-x - root hdfs 0 2022-05-13 17:10 /tmp/datas
drwxr-xr-x - root hdfs 0 2022-05-13 07:31 /tmp/dataset
-rw-r--r-- 1 raj_ops hdfs 23873 2022-06-27 09:26 /tmp/dibetics.csv
drwxrwxr-x - druid hadoop 0 2018-11-29 19:01 /tmp/druid-indexing
-rw-r--r-- 1 root hdfs 0 2022-05-13 17:15 /tmp/dummy.txt
drwxr-xr-x - hdfs hdfs 0 2018-11-29 17:25 /tmp/entity-file-history
drwxrwx-wx - hive hdfs 0 2018-11-29 17:58 /tmp/hive
-rw-r--r-- 1 root hdfs 0 2022-06-27 09:37 /tmp/myfile.txt
-rw-r--r-- 1 raj_ops hdfs 14239516 2022-05-13 14:16 /tmp/weatherAUS.csv

```

- Load txt file from local to tmp directory.



```

Hortonworks Sandbox HDP | root@sandbox-hdp:~ - Shell In | Ambari - Sandbox | Ambari - Sandbox
localhost:4200 | Apps | Gmail | YouTube | Maps

sandbox-hdp login: root
root@sandbox-hdp.hortonworks.com's password:
Last login: Fri May 13 14:27:48 2022 from 172.18.0.2
[root@sandbox-hdp ~]# hadoop fs -touch /tmp/data.txt
[root@sandbox-hdp ~]# hadoop fs -mv /tmp/data.txt /tmp/mydata
[root@sandbox-hdp ~]# hadoop fs -ls /tmp/mydata
-bash: hadoop: command not found
[root@sandbox-hdp ~]# hadoop fs -ls /tmp/mydata
-rw-r--r-- 1 root hdfs 0 2022-05-13 15:23 /tmp/mydata
[root@sandbox-hdp ~]# hadoop fs -rm /tmp/mydata
22/05/13 15:31:32 INFO fs.TrashPolicyDefault: Moved: 'hdfs://sandbox-hdp.hortonworks.com:8020/tmp/mydata' to trash at: hdfs://sandbox-hdp.hortonworks.com:8020/user/root/.Trash/Current/tmp/mydata1652455892535
[root@sandbox-hdp ~]#

```

B. Analyze the impact of data driven organizations using real time applications.

*The ability to work in real-time and respond to a customers needs or prevent issues before they arise ends up benefitting the bottom line by reducing risk enhancing accuracy and to monitor the incoming orders and other product and to detect for short term contract labour if production, packing or shipping is falling behind the target.

*Organizations are making investment in real-time data analytics , so that data is available to be analyzed , interpreted and visualized as it is created or changes in their source system.

Example:

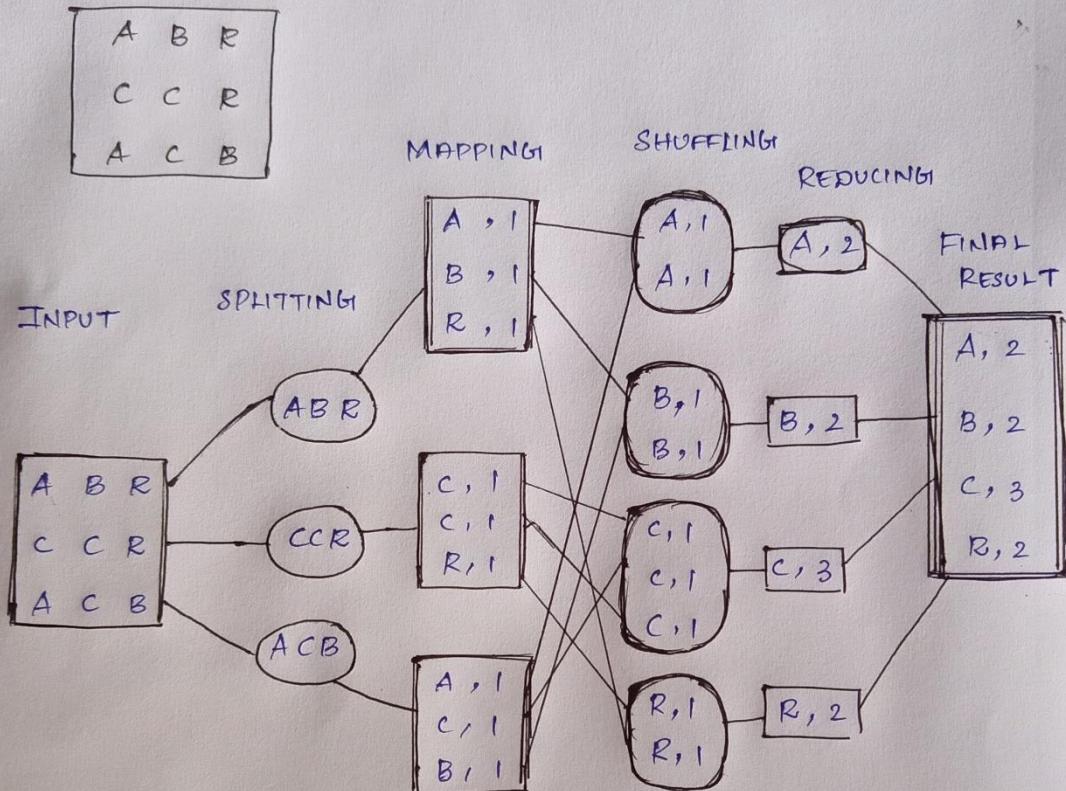
*When call centre agents look up customer records, they should be able to see information on the local outages, unusual high bill, cancelled flights or other issues that prompted the call, while talking the customer. This kind of insight driven by real time applications is standard practice now.

*Ecommerce sites typically use data to drive profits and sales. If you've ever shopped at Amazon you have probably received a product recommendation

while visiting the Amazon website or through email. This is an example of a data-driven business decision.

C. Diagrammatically represent the steps involved in MapReduce for the Data block given below for the wordcount analysis.

D) c) Map Reduce

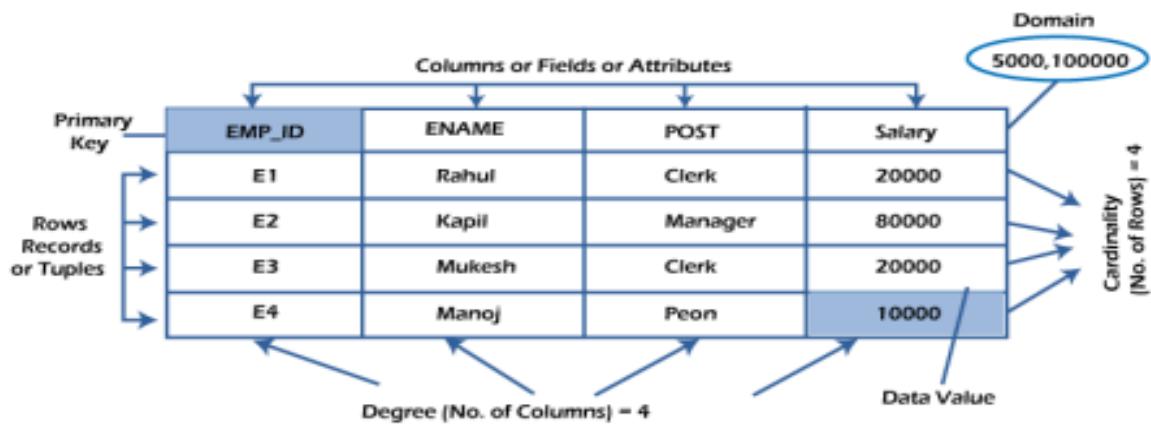


The Five common Steps Involved In Map Reduce are,

- * Preparing the Map() Input (splitting)
- * User-Provided Map() (Mapping)
- * Shuffling the Map output
- * Reducing the Map output
- * Produce the Final output.

D. Consider the data tables given below. Identify the data structure to which it belongs to and justify your answer.

i)



The table belongs to the Data structure of Structured data

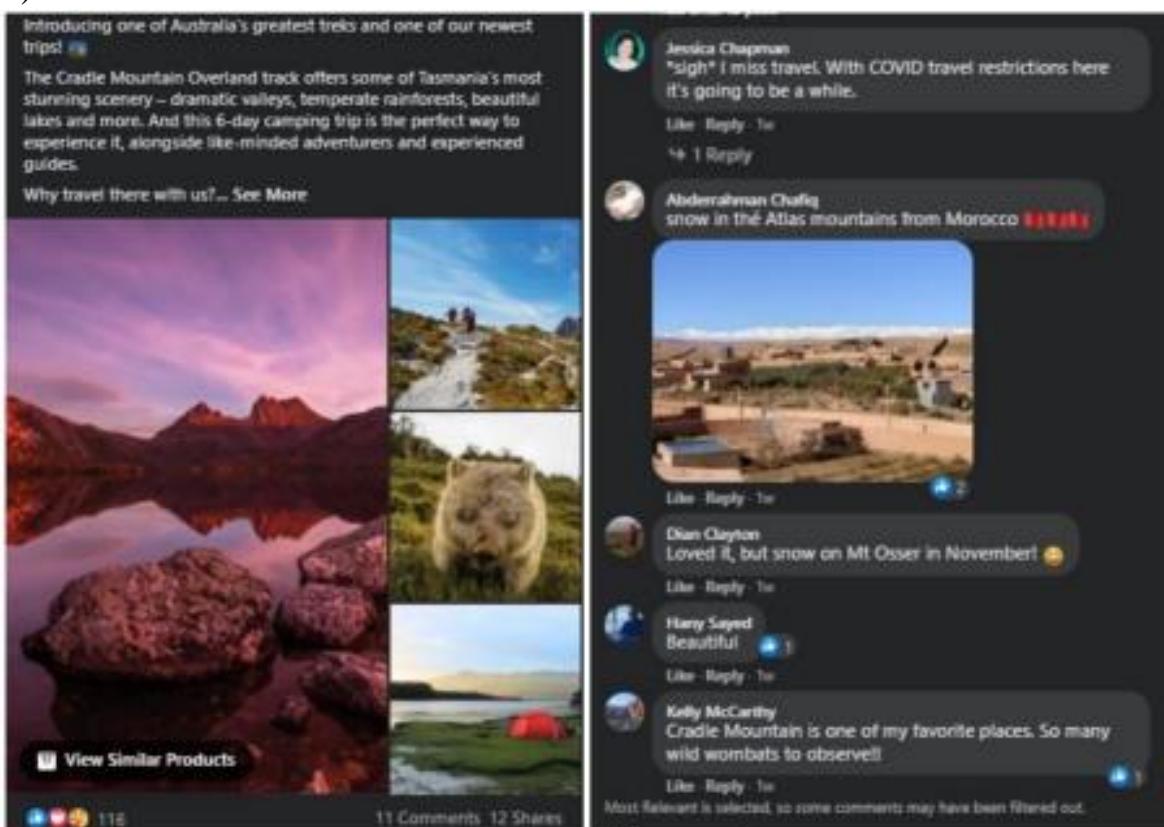
STRUCTURED DATA:

Structured data can be defined as the data that resides in a fixed field within a record. Data containing a defined data type, format and structure. In Structured data all the data has the same set of properties.

Examples:

- * Spreadsheet
- * Csv file
- * RDBMS file
- * Online Analytics processing(OLAP)

ii)



The type of the data structure is Unstructured data

UNSTRUCTURED DATA:

Unstructured data is the kind of the data that does not adhere to any definite schema or set of rules. Its arrangement is unplanned and the data has no inherent structure.

Examples:

- *Text documents
- *PDF document
- *Images and Video

iii)

```
<department>
  <name>Computer Science</name>
  <course>
    <code>CS101</code>
    <title>Introduction to Computer Science</title>
    <grade>A</grade>
  </course>
  <course>
    <code>CS105</code>
    <title>Data Structures</title>
    <grade>B</grade>
  </course>
  <course>
    <code>CS101</code>
    <title>Introduction to Computer Science</title>
    <grade>B</grade>
  </course>
</department>
<student>
  <stuNo>123456</stuNo>
  <stuName>Bob Smith</stuName>
  <grade>A</grade>
</student>
<student>
  <stuNo>234567</stuNo>
  <stuName>Mary Brown</stuName>
  <grade>B</grade>
</student>
<course>
  <code>CS105</code>
  <title>Data Structures</title>
</course>
<student>
  <stuNo>123456</stuNo>
  <stuName>Bob Smith</stuName>
  <grade>B</grade>
</student>
</course>
</department>
```

The type of the data Structure is Semi-Structured data

SEMI STRUCTURED DATA:

Semi-Structured data is not bound by any rigid scheme for data storage and handling . Semi-Structured data is a type that does not hold to the tabular structure of data models. The data is not in the relational format .It is the combination of Unstructured and Structured data.

Examples:

*XML

*JSON

*YAML

*HTML

E. An organization wanted to implement fault tolerance and promote portability across heterogeneous hardware and software platforms with data accessibility using master slave architecture. What architecture you would suggest for them and why?

*The suggested Architecture is Hadoop HDFS Architecture

*Hadoop Distributed File System (HDFS) is the world's most reliable storage system. It is best known for its **fault tolerance** and **high availability** Hadoop Distributed File System (HDFS) is the world's most reliable storage system. It is best known for its **fault tolerance** and **high availability**

*HDFS is designed with the portable property so that it should be portable from one platform to another. This enables the widespread adoption of HDFS. It is the best platform while dealing with a large set of data.

***HDFS** stores very large files running on a cluster of commodity hardware. It works on the principle of storage of less number of large files rather than the huge number of small files. HDFS stores data reliably even in the case of hardware failure. It provides high throughput by providing the data access in parallel.

2.i) One of the serious problems to be tackled in streaming data is the space requirement. Consider the input data streams represented as $X=1,3,2,1,2,3,4,3,1,2,3,1$. Perform the relevant algorithm to optimize the storage of the streaming data using the following Hash Function to find the distinct elements to be stored.

A. $h(x)=6x+2 \bmod 5$

2)

Hash Function

$$x = 1, 3, 2, 1, 2, 3, 4, 3, 1, 2, 3, 1$$

A) $h(x) = 6x + 2 \pmod{5}$

$$\begin{aligned} h(1) &= 6(1) + 2 \pmod{5} \\ &= 8 \pmod{5} \\ &= \boxed{3} \end{aligned}$$

$$\begin{aligned} h(3) &= 6(3) + 2 \pmod{5} \\ &= 20 \pmod{5} \\ &= \boxed{0} \end{aligned}$$

$$\begin{aligned} h(2) &= 6(2) + 2 \pmod{5} \\ &= 14 \pmod{5} \\ &= \boxed{4} \end{aligned}$$

$$\begin{aligned} h(4) &= 6(4) + 2 \pmod{5} \\ &= 26 \pmod{5} \\ &= \boxed{1} \end{aligned}$$

Binary value of 3 is,

$$\boxed{011} \rightarrow 3 \text{ Bits}$$

Binary value of 0 is

$$\boxed{000} \rightarrow 3 \text{ Bits}$$

Binary value of 4 is

$$\boxed{100} \rightarrow 3 \text{ Bits}$$

Binary value of 1 is $\boxed{001} \rightarrow 3 \text{ Bits}$.

HASH VALUE	BINARY SETUP	TRAILING ZERO
$h(1)$	011	0
$h(3)$	000	0
$h(2)$	100	2
$h(1)$	011	0
$h(2)$	100	2
$h(3)$	000	0
$h(4)$	001	0
$h(3)$	000	0
$h(1)$	011	0
$h(2)$	100	2
$h(3)$	000	0
$h(1)$	011	0

Distinct Element,

$$r = 2 \quad \text{Maximum Trailing Zero} = 2$$

$$R = 2^2 \Rightarrow 2^2$$

$$\boxed{R = 4}$$

\therefore There are 4 distinct elements

$$\{1, 2, 3, 4\}$$

B. $h(x) = 2x + 1 \bmod 16$

$$B) h(x) = 2x + 1 \pmod{16}$$

$$x = \{ 1, 3, 2, 1, 2, 3, 4, 3, 1, 2, 3, 1 \}$$

$$h(1) = 2(1) + 1 \pmod{16}$$

$$= 3 \pmod{16}$$

$$= \boxed{3}$$

$$h(3) = 2(3) + 1 \pmod{16}$$

$$= 7 \pmod{16}$$

$$\boxed{7}$$

$$h(2) = 2(2) + 1 \pmod{16}$$

$$= 5 \pmod{16}$$

$$\boxed{5}$$

$$h(4) = 2(4) + 1 \pmod{16}$$

$$= 9 \pmod{16}$$

$$= \boxed{9}$$

Binary value of 3 is,

$$\boxed{0011} \rightarrow 4 \text{ Bits}$$

Binary value of 7 is,

$$\boxed{0111} \rightarrow 4 \text{ Bits}$$

Binary value of 5 is,

$$\boxed{0101} \rightarrow 4 \text{ Bits}$$

Binary value of 9 is,

$$\boxed{1001} \rightarrow 4 \text{ Bits.}$$

HASH FUNCTION	BINARY SETUP	TRAILING ZERO
$h(1)$	0011	0
$h(3)$	0111	0
$h(2)$	0101	0
$h(1)$	0011	0
$h(2)$	0101	0
$h(3)$	0111	0
$h(4)$	1001	0
$h(3)$	0111	0
$h(1)$	0011	0
$h(2)$	0101	0
$h(3)$	0111	0
$h(1)$	0011	0

$\lambda = 0$ $(\text{No Trailing Zero's})$

$R = 2^0$

$\boxed{R = 1}$

Distinct element,

$R = 2^0$

$\boxed{R = 1}$

$\therefore \text{There is 1 distinct element present.}$

$\{1\}$

ii) Perform Alon Matias Szegedy algorithm for second order moment for data stream {a,b,c,b,d,a,c,d,a,b,d,c,a,a,b}

A. Find the second order moment.

B. Given the random 3 variables X1, X2 and X3 at 1st, 7th and 11th Position. Find the X.element and X.values.

i) Alien Matias Szegedy Algorithm
 $\{a, b, c, d, a, c, d, c, a, a, b\}$

A. Find the second order moment.

$$n = 15 \text{ (total length)}$$

$$2^{\text{nd}} \text{ Order Moment} = k = 2$$

$a \rightarrow 5$ Times occurring

$b \rightarrow 4$ Times occurring

$c \rightarrow 3$ Times occurring

$d \rightarrow 3$ Times occurring

Second Order Momentum

$$\Rightarrow 5^2 + 4^2 + 3^2 + 3^2$$

$$\Rightarrow 25 + 16 + 9 + 9$$

$$\Rightarrow \boxed{59}$$

B. $x_1 \rightarrow 1^{\text{st}}$ Position

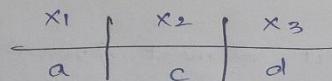
$x_2 \rightarrow 7^{\text{th}}$ Position

$x_3 \rightarrow 11^{\text{th}}$ Position

To Find x_i element and x_i values.

$\{a, b, c, b, d, a, c, d, a, b, d, c, a, a, b\}$

① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩ ⑪ ⑫ ⑬ ⑭ ⑮



$$x_1 \cdot \text{Values} = 1+1+1+1+1 = \boxed{5}$$

$x_1 \cdot \text{element} = a$

$$x_2 \cdot \text{Values} = 1+1 = \boxed{2}$$

$x_2 \cdot \text{element} = c$

$$x_3 \cdot \text{Values} = \boxed{1}$$

$x_3 \cdot \text{element} = d$

3. A. Apply decaying window algorithm to analyse the trends in facebook data for the following facebook tags

#Covid , # #Outfitoftheday, #Covid, # #Outfitoftheday, # #Outfitoftheday, #Covid, #Covid

Consider weight=1 and C=0.1

Find the most trending tag related to the Covid and Outfitoftheday and state the reason for the same mathematically.

3)

A. Decaying Window [Facebook Tag]

#covid, ## Outfit of the day, # covid, ## Outfit Of the day,
Outfit of the day, # covid, # covid

Weight = 1 and C = 0.1

CALCULATING FOR COVID

Decaying Window Formula,

$$\sum_{i=0}^{t-1} a^{t-i} (1-c)^i \quad |c \rightarrow \text{Length of window}$$

As weight increases, window size decreases

$$\text{covid} \Rightarrow 1 * (1-0.1) \Rightarrow 1 * 0.9 \Rightarrow 0.9$$

$$\text{outfit} \Rightarrow 0.9 * (1-0.1) + 1 \Rightarrow 0.9 * 0.9 + 1 \Rightarrow 0.81$$

$$\text{covid} \Rightarrow 0.81 * (1-0.1) + 1 \Rightarrow 1.729$$

$$\text{outfit} \Rightarrow 1.729 * (1-0.1) + 0 \Rightarrow 1.5561$$

$$\text{outfit} \Rightarrow 1.5561 * (1-0.1) + 0 \Rightarrow 1.4004$$

$$\text{covid} \Rightarrow 1.4004 * (1-0.1) + 1 \Rightarrow 2.260441$$

$$\text{covid} \Rightarrow 2.260441 * (1-0.1) + 1 \Rightarrow [3.0343]$$

\therefore The value of covid is 3.0343

CALCULATING FOR OUTFIT OF THE DAY

$$\text{covid} \Rightarrow 0 * (1-0.1) = 0$$

$$\text{outfit} \Rightarrow 0 * (1-0.1) + 1 = 1$$

$$\text{covid} \Rightarrow 1 * (1-0.1) + 0 = 0.9$$

$$\text{outfit} \Rightarrow 0.9 * (1-0.1) + 1 = 1.81$$

$$\text{outfit} \Rightarrow 1.81 * (1-0.1) + 1 = 2.629$$

$$\text{covid} \Rightarrow 2.629 * (1-0.1) + 0 = 2.3661$$

$$\text{covid} \Rightarrow 2.3661 * (1-0.1) + 0 = [2.1294]$$

\therefore The value of outfit of the day is 2.1294

So, covid has more value (3.0343) than outfit of the day (2.1294).

\therefore # covid is trending.

B) utility Matrix:

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

B. Provided the utility matrix specified in figure 1 below with the rating from 1-5 scale having 8 items a,b,c,d,e,f,g,h and 3 users A,B and C

- Treat the utility matrix as Boolean and compute the Jaccard distance between each pair of users.
- Find the cosine distance between each pair of users
- Treat ratings of 4 and 5 as 1 and 1,2,3 and blank as 0. Compute the Jaccard distance between each pair of users and also the cosine distance
- Normalize the matrix and then compute the cosine distance between each pair of users.

- i) Treat the Utility matrix as Boolean and compute the Jaccard distance of each pair.

UTILITY MATRIX AS BOOLEAN

	a	b	c	d	e	f	g	h
A	1	1	0	1	1	0	1	1
B	0	1	1	1	1	1	1	0
C	1	0	1	1	0	1	1	1

JACCARD DISTANCE

$$\text{Jaccard distance } (A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$$(A, B) = \frac{4}{8} = \frac{1}{2} = [0.5]$$

$$\text{Jaccard distance } (A, C) = \frac{|A \cap C|}{|A \cup C|}$$

$$(A, C) = \frac{4}{8} = \frac{1}{2} = [0.5]$$

$$\text{Jaccard distance } (B, C) = \frac{|B \cap C|}{|B \cup C|}$$

$$(B, C) = \frac{4}{8} = \frac{1}{2} = [0.5]$$

\therefore Jaccard distance between (A, B) , (A, C) and (B, C) are similar (0.5).

- ii) Find the cosine distance between each pair of users.

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

cosine distance between A and B.

$$\begin{aligned}
 \text{cosine } (A, B) &= \cos (\alpha_A, \alpha_B) \\
 &= \frac{(5 \times 3) + (5 \times 3) + (1 \times 1) + (3 \times 1)}{\sqrt{4^2 + 5^2 + 5^2 + 1^2 + 3^2 + 2^2} \sqrt{3^2 + 4^2 + 3^2 + 1^2 + 2^2 + 1^2}} \\
 &\Rightarrow \frac{34}{\sqrt{56.5668}} \\
 &\Rightarrow 0.6010
 \end{aligned}$$

cosine distance between (A, C).

$$\begin{aligned}
 \text{cosine } (A, C) &= \cos (\alpha_A, \alpha_C) \\
 &= \frac{(4 \times 2) + (5 \times 3) + (3 \times 5) + (2 \times 3)}{\sqrt{4^2 + 5^2 + 5^2 + 1^2 + 3^2 + 1^2} \sqrt{2^2 + 1^2 + 3^2 + 4^2 + 5^2 + 3^2}} \\
 &\Rightarrow \frac{44}{\sqrt{71.552}} \\
 &\Rightarrow 0.61493
 \end{aligned}$$

Cosine Distance Between (B, C)

$$\cos(\theta_{B,C}) = \cos(\alpha_B, \alpha_C)$$

$$\Rightarrow \frac{(4 \times 1) + (3 \times 3) + (2 \times 4) + (1 \times 5)}{\sqrt{3^2 + 4^2 + 3^2 + 1^2 + 2^2 + 1^2} \times \sqrt{2^2 + 1^2 + 3^2 + 1^2 + 5^2 + 3^2}} \\ = \frac{26}{\sqrt{51} \cdot \sqrt{76}} = [0.5022]$$

\therefore The cosine distance (A, C) is more similar
than B, C .

- (ii) Treat ratings of 4 and 5 as 1 and 2, 3 and blank as 0. compute the Jaccard distance between each pair of user and also the cosine distance.

Treating 4 and 5 as 1

1, 2, 3 as 0

Utility Matrix:

	a	b	c	d	e	f	g	h
A	1	1	0	1	0	0	0	0
B	0	0	1	0	0	0	0	0
C	0	0	0	0	0	1	1	0

Jaccard Distance For (A, B)

$$\Rightarrow \frac{|A \cap B|}{|A \cup B|}$$

$$\Rightarrow \frac{0}{4} = [0]$$

Jaccard distance for (A, C)

$$= |x_A \cap x_C| / |x_A \cup x_C|$$

$$\Rightarrow \frac{0}{5}$$

$$= \boxed{0}$$

Jaccard distance for (B, C)

$$= |x_B \cap x_C| / |x_B \cup x_C|$$

$$\Rightarrow \frac{0}{5}$$

$$= \boxed{0}$$

∴ Jaccard distance for (A, B), (A, C) and (B, C)

is 0.

Cosine distance for (A, B)

$$\text{sim}(A, B) = \cos(x_A, x_B)$$

$$\Rightarrow \frac{0}{\sqrt{1^2+1^2+1^2}} \sqrt{1^2}$$

$$= \boxed{0}$$

Cosine distance for (A, C)

$$\text{sim}(A, C) = \cos(x_A, x_C)$$

$$\Rightarrow \frac{0}{\sqrt{1^2+1^2+1^2}} \sqrt{1^2+1^2}$$

$$= \boxed{0}$$

Cosine Distance For (B, C)

$$\text{Sim}(B, C) = \cos(\alpha_B, \alpha_C)$$

$$\Rightarrow \frac{0}{\sqrt{1^2 + 1^2}}$$

$$= [0]$$

\therefore The cosine distance for (A, B), (A, C) and (B, C) are 0.

- v) Normalize the matrix and then compute the cosine distance between each pair.

Utility Matrix:

	a	b	c	d	e	f	g	h
A	4	5	5	1		3	2	
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

Average Rating For A,

$$\Rightarrow \frac{4+5+5+1+3+2}{6} \Rightarrow \frac{20}{6}$$

$$\Rightarrow \boxed{\frac{10}{3}}$$

Subtracting $\frac{10}{3}$ to calculate A,

$$4 - \frac{10}{3} = \boxed{\frac{2}{3}}$$

$$1 - \frac{10}{3} = \boxed{-\frac{7}{3}}$$

$$5 - \frac{10}{3} = \boxed{\frac{5}{3}}$$

$$3 - \frac{10}{3} = \boxed{-\frac{1}{3}}$$

$$2 - \frac{10}{3} = \boxed{-\frac{4}{3}}$$

Average Rating for B,

$$\Rightarrow \frac{3+4+3+1+2+1}{6}$$

$$\Rightarrow \frac{14}{6}$$

$$= \boxed{\frac{7}{3}}$$

Subtracting $\frac{7}{3}$ to calculate B values

$$3 - \frac{7}{3} = \boxed{\frac{2}{3}}$$

$$2 - \frac{7}{3} = \boxed{-\frac{1}{3}}$$

$$4 - \frac{7}{3} = \boxed{\frac{5}{3}}$$

$$1 - \frac{7}{3} = \boxed{-\frac{4}{3}}$$

Average Rating for C

$$\Rightarrow \frac{2+1+3+4+5+3}{6} \Rightarrow \frac{18}{6} = \boxed{3}$$

Subtracting 3 from original value to calculate C

$$2-3 \Rightarrow \boxed{-1} \quad 4-3 = \boxed{1}$$

$$1-3 \Rightarrow \boxed{-2} \quad 5-3 = \boxed{2}$$

$$3-3 \Rightarrow \boxed{0}$$

Normalized Utility Matrix Data.

	a	b	c	d	e	f	g	h
A	$\frac{2}{3}$	$\frac{5}{3}$		$\frac{5}{3}$	$-\frac{1}{3}$		$-\frac{1}{3}$	$-\frac{4}{3}$
B		$\frac{2}{3}$	$\frac{5}{3}$	$\frac{2}{3}$	$-\frac{4}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{4}{3}$
C	-1		-2	0		1	2	0

Cosine Distance FOR (A, B)

$$\text{sim}(A, B) = \text{cosine}(x_A, x_B)$$

$$\Rightarrow \frac{(5/3 \times -2/3) + (5/3 \times 2/3) + (-1/3 \times -4/3) + (-1/3 \times 4/3)}{\sqrt{(\frac{2}{3})^2 + (\frac{5}{3})^2 + (\frac{5}{3})^2 + (-\frac{1}{3})^2} \times \sqrt{(\frac{2}{3})^2 + (\frac{5}{3})^2 + (\frac{5}{3})^2 + (-\frac{4}{3})^2 + (-\frac{1}{3})^2 + (-\frac{4}{3})^2}}$$

$$\Rightarrow \frac{10/9 + 10/9 + 2/9 + 4/9}{\sqrt{4/9 + 25/9 + 25/9 + 1/9} \times \sqrt{4/9 + 25/9 + 4/9 + 16/9 + 1/9 + 16/9}}$$

$$\Rightarrow \frac{5/3}{\sqrt{120/9} \times \sqrt{60/9}}$$

$$\Rightarrow \frac{5 \cdot 7777}{9 \cdot 8869}$$

$$\Rightarrow [0.5843]$$

Cosine Distance FOR (A, C)

$$\text{sim}(A, C) = \text{cosine}(x_A, x_C)$$

$$\Rightarrow \frac{(2/3 \times -1) + (-1/3 \times 2)}{\sqrt{\frac{120}{9}} \times \sqrt{(-1)^2 + (-2)^2 + (1)^2 + (2)^2}}$$

$$\Rightarrow \frac{-2/3 + -2/3}{\sqrt{120/9} \times \sqrt{10}} \Rightarrow \frac{-4/3}{3.641 \times 3.1622}$$

$$\Rightarrow \frac{-10.333}{11.5451} \Rightarrow [-0.115460]$$

Cosine Distance FOR (B, C)

$$\text{sim}(B, C) = \text{cosine}(x_B, x_C)$$

$$\Rightarrow \frac{(5/3 \times -2) + (-1/3 \times 1) + (-1/3 \times 2)}{\sqrt{60/9} \times \sqrt{10}}$$

$$\Rightarrow \frac{(-10/3)(-1/3)(-8/3)}{2.7080 \times 3.1622}$$

$$\Rightarrow \frac{-10/3}{8.5632}$$

$$\Rightarrow \frac{-6.33}{8.5632}$$

$$\Rightarrow [-0.7395]$$

\therefore The cosine distance (A, B) is more similar.

v) Interpret your results.

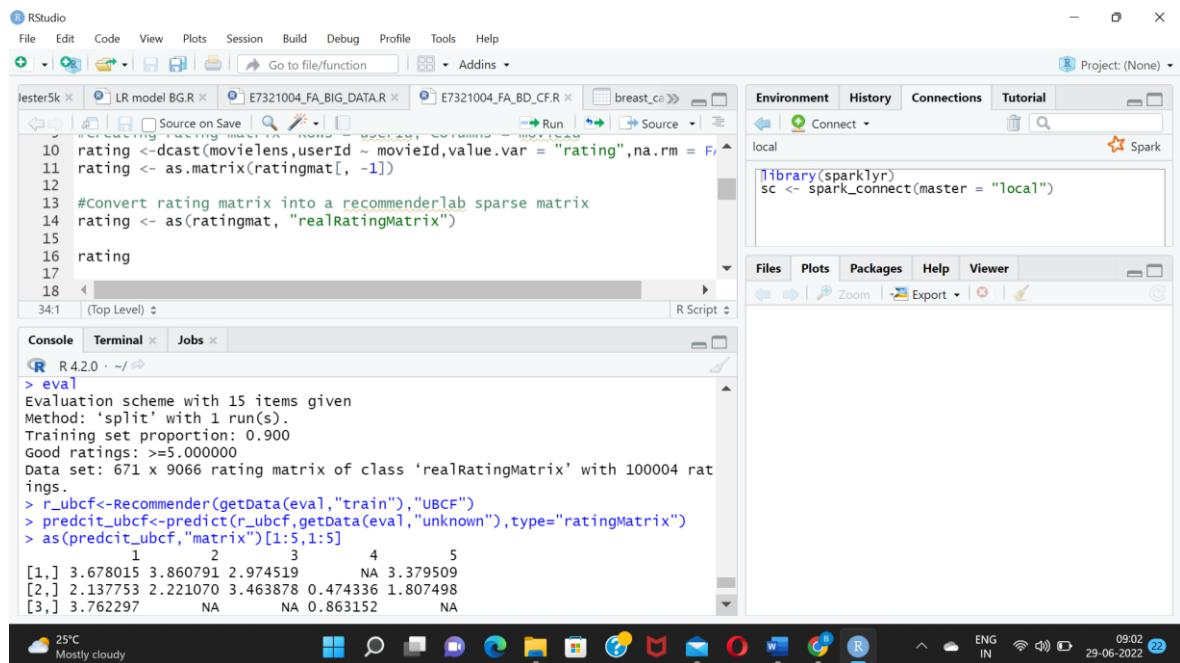
INTERPRETATION:

*For the above Boolean Matrix , when convert the ratings 4,5 as 1 and 1,2,3 and blank as 0 The values of jaccard distance and cosine distance became 0 There is no similarity between them.

*When we normalize the utility matrix we can analyse that there is a cosine similarity for user (A,B)

*So, normalizing the data is efficient.

C. Implement a simple collaborative filtering-based recommendation algorithm using the movielens datasets and interpret the effectiveness of your model.

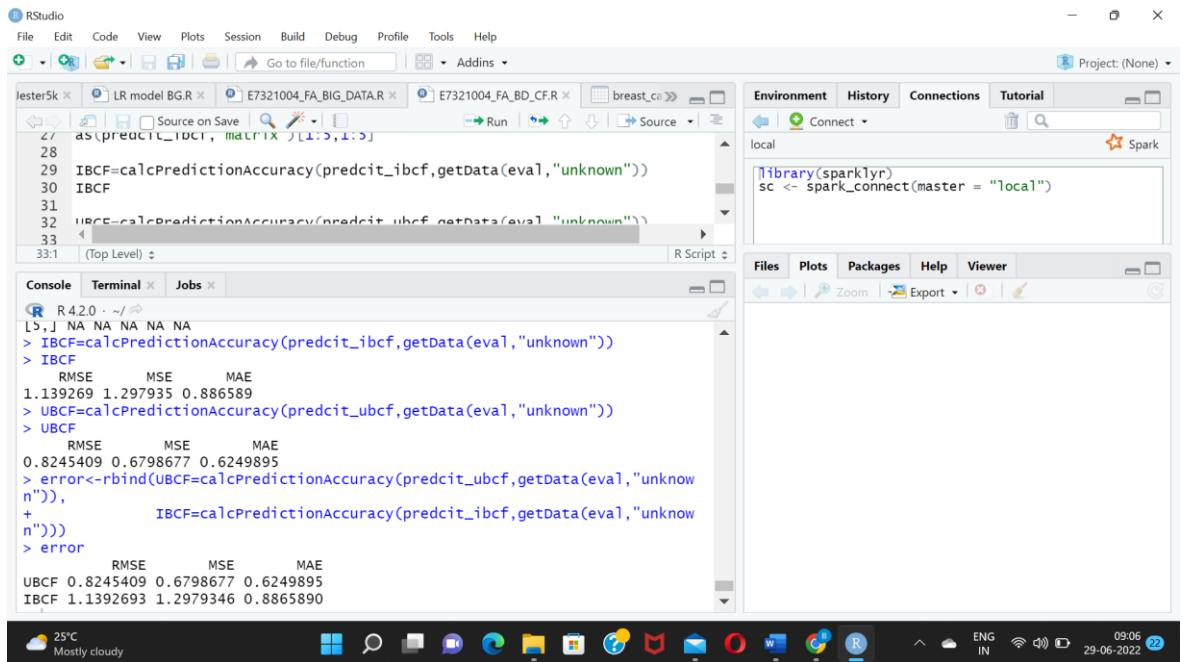


The screenshot shows the RStudio interface with the following components:

- Top Bar:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Project Bar:** Project: (None).
- Code Editor:** Shows R script code for a UBCF recommendation system, including data loading, matrix conversion, and prediction.
- Environment Tab:** Displays library(sparklyr) and sc <- spark_connect(master = "local")
- Console Tab:** Shows the R session output, including evaluation scheme details and a sample prediction matrix.
- System Tray:** Shows weather (25°C, Mostly cloudy), system icons, and a date/time stamp (29-06-2022 09:02).

```
rating <- dcast(movieLens,userId ~ movieId,value.var = "rating",na.rm = F)
rating <- as.matrix(rating[, -1])
#Convert rating matrix into a recommenderlab sparse matrix
rating <- as(ratingmat, "realRatingMatrix")
rating
<-
library(sparklyr)
sc <- spark_connect(master = "local")
```

```
> eval
Evaluation scheme with 15 items given
Method: 'split' with 1 run(s).
Training set proportion: 0.900
Good ratings: >=5.000000
Data set: 671 x 9066 rating matrix of class 'realRatingMatrix' with 100004 ratings.
> r_ubcf<-Recommender(getData(eval,"train"),"UBCF")
> predict_ubcf<-predict(r_ubcf,getData(eval,"unknown"),type="ratingMatrix")
> as(predict_ubcf,"matrix")[1:5,1:5]
      1     2     3     4     5
[1,] 3.678015 3.860791 2.974519    NA 3.379509
[2,] 2.137753 2.221070 3.463878 0.474336 1.807498
[3,] 3.762297    NA      NA 0.863152    NA
```



CODE:

```

install.packages("dslabs")
library(dslabs)
library(recommenderlab)
install.packages("reshape2")
library(reshape2)
data(movielens)
head(movielens)
View(movielens)

#Creating rating matrix Rows = userId, Columns = movieId
rating <- dcast(movielens,userId ~ movieId,value.var = "rating",na.rm = FALSE)
rating <- as.matrix(rating[, -1])

#Convert rating matrix into a recommenderlab sparse matrix
rating <- as(rating, "realRatingMatrix")
rating
eval<-evaluationScheme(rating,method="split",train=0.9,given=15,goodRating=5)
eval
r_ubcf<-Recommender(getData(eval,"train"),"UBCF")
predcit_ubcf<-predict(r_ubcf,getData(eval,"unknown"),type="ratingMatrix")

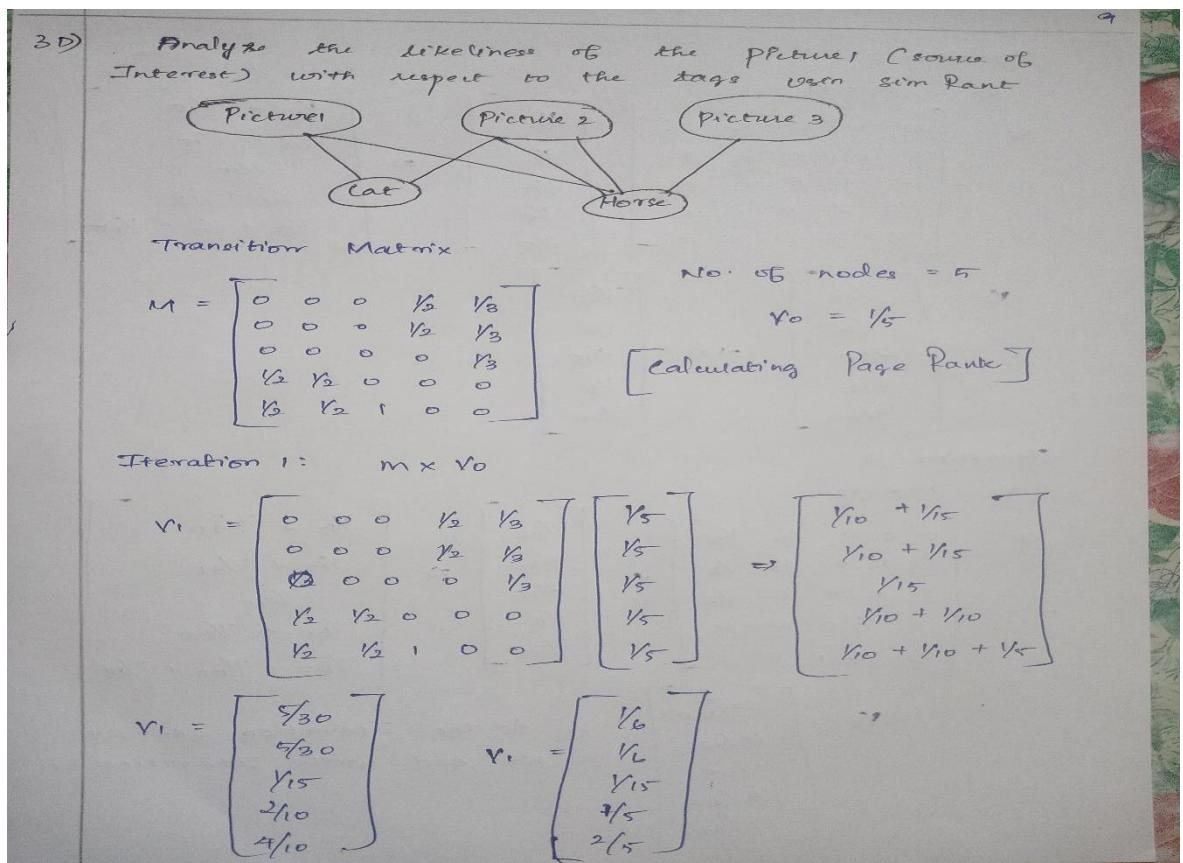
```

```

as(predcit_ubcf,"matrix")[1:5,1:5]
i_ubcf<-Recommender(getData(eval,"train"),"IBCF")
predcit_ibcf<-predict(i_ubcf,getData(eval,"unknown"),type="ratingMatrix")
as(predcit_ibcf,"matrix")[1:5,1:5]
IBCF=calcPredictionAccuracy(predcit_ibcf,getData(eval,"unknown"))
IBCF
UBCF=calcPredictionAccuracy(predcit_ubcf,getData(eval,"unknown"))
UBCF
error<-rbind(UBCF=calcPredictionAccuracy(predcit_ubcf,getData(eval,"unknown")),
IBCF=calcPredictionAccuracy(predcit_ibcf,getData(eval,"unknown")))
error

```

D. Analyse the likeliness of the Picture 1(Source of Interest) with respect to the tags represented in the graph below using SimRank.



Iteration 2 :

$$V_2 = m \times V_1$$

$$V_2 = \begin{bmatrix} 0 & 0 & 0 & V_2 & V_3 \\ 0 & 0 & 0 & V_2 & V_3 \\ 0 & 0 & 0 & 0 & V_3 \\ V_2 & V_2 & 0 & 0 & 0 \\ V_2 & V_2 & 1 & 0 & 0 \end{bmatrix} \xrightarrow{\dots} \begin{bmatrix} V_6 \\ V_6 \\ V_{15} \\ V_5 \\ V_5 \end{bmatrix} \Rightarrow \begin{bmatrix} V_{10} + V_{15} \\ V_{10} + V_{15} \\ V_{15} \\ V_{12} + V_{12} \\ V_{12} + V_{12} + V_{15} \end{bmatrix}$$

$$V_2 = \begin{bmatrix} V_{30} \\ V_{30} \\ 2/15 \\ 2/12 \\ 14/60 \end{bmatrix} \Rightarrow V_2 = \begin{bmatrix} V_{30} \\ V_{30} \\ 2/15 \\ V_6 \\ V_{30} \end{bmatrix}$$

Iteration 3 :

$$V_3 = m \times V_2$$

$$V_3 = \begin{bmatrix} 0 & 0 & 0 & V_2 & V_3 \\ 0 & 0 & 0 & V_2 - V_3 & V_3 \\ 0 & 0 & 0 & 0 & V_3 \\ V_2 & V_2 & 0 & 0 & 0 \\ V_2 & V_2 & 1 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} V_{30} \\ V_{30} \\ 2/15 \\ V_6 \\ V_{30} \end{bmatrix} \Rightarrow \begin{bmatrix} V_{12} + V_{90} \\ V_{12} + V_{90} \\ V_{90} \\ V_{60} + V_{60} \\ V_{60} + V_{60} + 2/15 \end{bmatrix}$$

$$V_3 = \begin{bmatrix} 29/180 \\ 29/180 \\ 7/90 \\ 14/60 \\ 22/60 \end{bmatrix}$$

∴ As the Iteration goes on
we end with Iteration 3.

Calculating Trust Rank for Sim Rank.

$$M = \begin{bmatrix} P_1 & & & & & \\ & P_1 & P_2 & P_3 & C & H \\ P_2 & & & & & \\ & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{3} \\ P_3 & & & & & \\ & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{3} \\ C & & & & & \\ & 0 & 0 & 0 & 0 & \frac{1}{3} \\ H & & & & & \\ & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 & 0 \\ & & & & 0 & 0 \end{bmatrix}$$

$$n = \text{no. of nodes} = 5$$

$$\delta = \frac{2}{5} \text{ Picture 1 } \frac{3}{5}$$

$$\nu = .1$$

$$v' = BMv + (1-B)e_s$$

$$B = \frac{4}{5} \quad (1-B) = 1 - \frac{4}{5} = \frac{1}{5}$$

$$e_s = \text{no. of 1 occurring} = 1$$

v' = Iteration 1 :

$$\Rightarrow \begin{bmatrix} 0 & 0 & 0 & \frac{1}{2} & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 & 0 \end{bmatrix} \times \frac{4}{5} \times \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \frac{1}{5} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 0 & 0 & 0 & \frac{2}{5} & \frac{4}{15} \\ 0 & 0 & 0 & \frac{2}{5} & \frac{4}{15} \\ 0 & 0 & 0 & 0 & \frac{4}{15} \\ \frac{2}{5} & \frac{2}{5} & 0 & 0 & 0 \\ \frac{2}{5} & \frac{2}{5} & 1 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{1}{5} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$v_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{2}{5} \\ \frac{2}{5} \end{bmatrix} + \begin{bmatrix} \frac{1}{5} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow v_1 = \begin{bmatrix} \frac{1}{5} \\ 0 \\ 0 \\ \frac{2}{5} \\ \frac{2}{5} \end{bmatrix}$$

Iteration 2:

$$v_2 = BMv_1 + (I - B)v_{es}$$

$$v_2 = \begin{bmatrix} 0 & 0 & 0 & \frac{2}{5} & \frac{4}{15} \\ 0 & 0 & 0 & \frac{2}{5} & \frac{4}{15} \\ 0 & 0 & 0 & 0 & \frac{4}{15} \\ \frac{2}{5} & \frac{2}{5} & 0 & 0 & 0 \\ \frac{2}{5} & \frac{2}{5} & \frac{4}{5} & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{5} \\ 0 \\ 0 \\ \frac{2}{5} \\ \frac{2}{5} \end{bmatrix} + \begin{bmatrix} \frac{1}{5} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$v_2 = \begin{bmatrix} \left(\frac{2}{5} \times \frac{2}{5}\right) + \left(\frac{4}{15} \times \frac{2}{5}\right) \\ \left(\frac{2}{5} \times \frac{2}{5}\right) + \left(\frac{4}{15} \times \frac{2}{5}\right) \\ \left(\frac{4}{15} \times \frac{2}{5}\right) \\ \left(\frac{2}{5} \times \frac{1}{5}\right) \\ \left(\frac{2}{5} \times \frac{1}{5}\right) \end{bmatrix} + \begin{bmatrix} \frac{1}{5} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$v_2 = \begin{bmatrix} \frac{4}{25} + \frac{8}{75} \\ \frac{4}{25} + \frac{8}{75} \\ \frac{8}{75} \\ \frac{2}{25} \\ \frac{2}{25} \end{bmatrix} \Rightarrow v_2 = \begin{bmatrix} \frac{20}{75} \\ \frac{20}{75} \\ \frac{8}{75} \\ \frac{2}{25} \\ \frac{2}{25} \end{bmatrix} \Rightarrow v_2 = \begin{bmatrix} \frac{4}{15} \\ \frac{4}{15} \\ \frac{8}{75} \\ \frac{2}{25} \\ \frac{2}{25} \end{bmatrix} + \begin{bmatrix} \frac{1}{5} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$r_3 = \begin{bmatrix} \frac{1}{125} + \frac{8}{375} \\ \frac{4}{125} + \frac{8}{375} \\ \frac{8}{375} \\ \frac{14}{125} + \frac{8}{125} \\ \frac{14}{125} + \frac{8}{125} + \frac{32}{375} \end{bmatrix} + \begin{bmatrix} \frac{1}{5} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{20}{375} \\ \frac{20}{375} \\ \frac{8}{375} \\ \frac{20}{125} \\ \frac{142}{375} \end{bmatrix}$$

$$v_3 = \begin{bmatrix} \frac{20}{375} \\ \frac{20}{375} \\ \frac{8}{375} \\ \frac{22}{375} \\ \frac{142}{375} \end{bmatrix} \times \begin{bmatrix} \frac{1}{5} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} \frac{95}{375} \\ \frac{20}{375} \\ \frac{8}{375} \\ \frac{22}{375} \\ \frac{142}{375} \end{bmatrix}$$

∴ According to the Trust Rank Iteration 3,
Horse is twice as likely as Picture 1 than Cat and
Picture 2.

so, Horse is more like as Picture 1.

4. A. Analyze ego-gplus and com-Youtube social network dataset provided by Stanford SNAP.

EGO-GPLUS

Nodes – 107614

Edges- 13673453

Average clustering coefficient -0.4901

COM-YOUTUBE

Nodes- 1134890

Edges – 2987624

Number of Triangles- 3056386

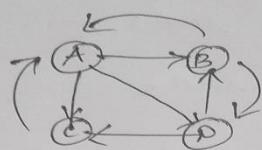
Average clustering coefficient – 0.0808

Number of communities – 8385

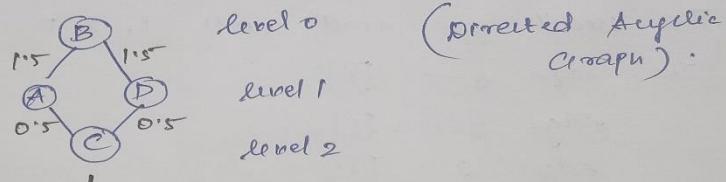
B. Apply Girvan-Newman Algorithm on the BFS specified below and interpret your findings for the figures 1 and 2 specified below.

4) B) Apply Girvan - Newman Algorithm on the BFS specified below and interpret your findings.

Figure 1:



* convert to (BFS) Breadth First search tree



Leaf Node C \rightarrow credit = 1

credit of A ($1 + \text{credit}$ of all edges connected to group)

$$\Rightarrow 1 + 0.5 \\ \Rightarrow [1.5]$$

credit of D

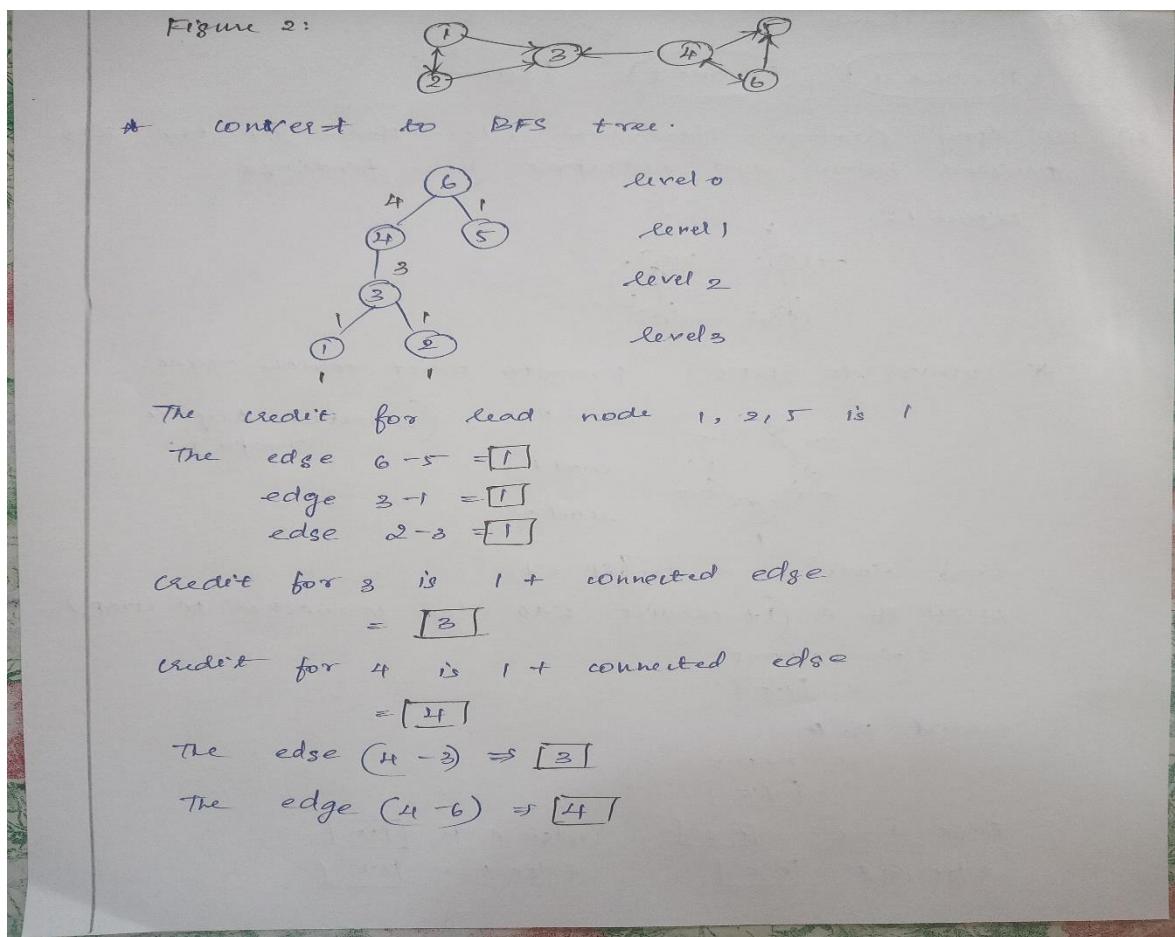
$$\Rightarrow 1 + 0.5 \\ \Rightarrow [1.5]$$

edge A - C = $[0.5]$

edge C - D = $[0.5]$

edge A - B = $[1.5]$

edge B - D = $[1.5]$



C. Implement the relationship between friends and family in Google+ through R programming.

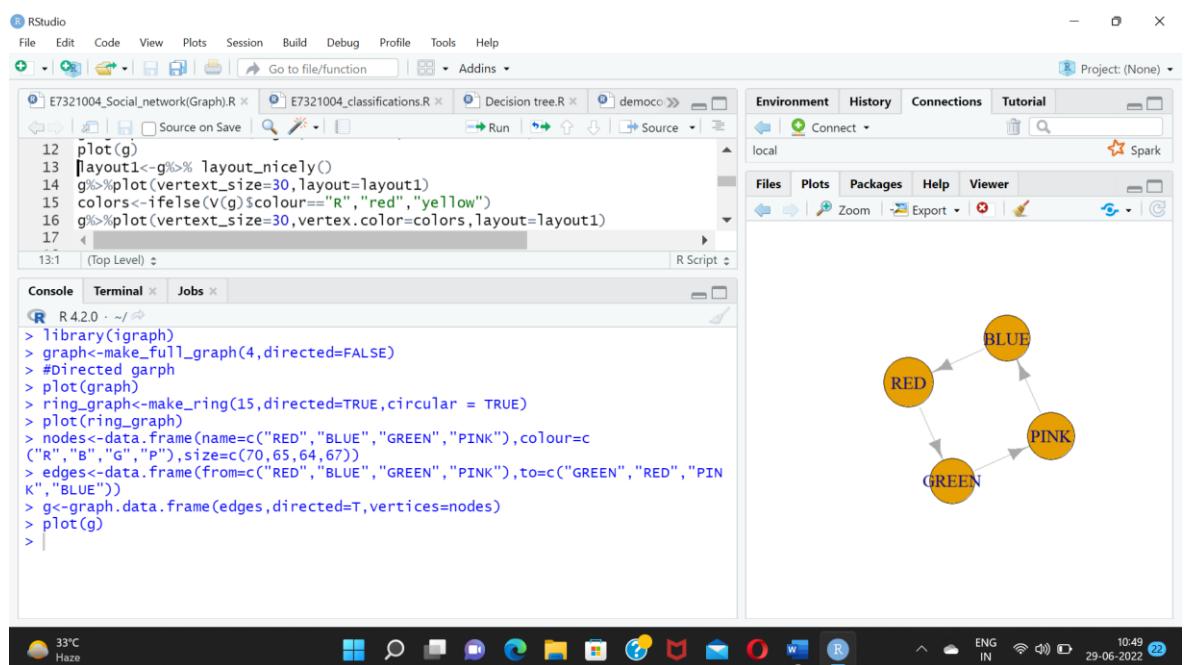
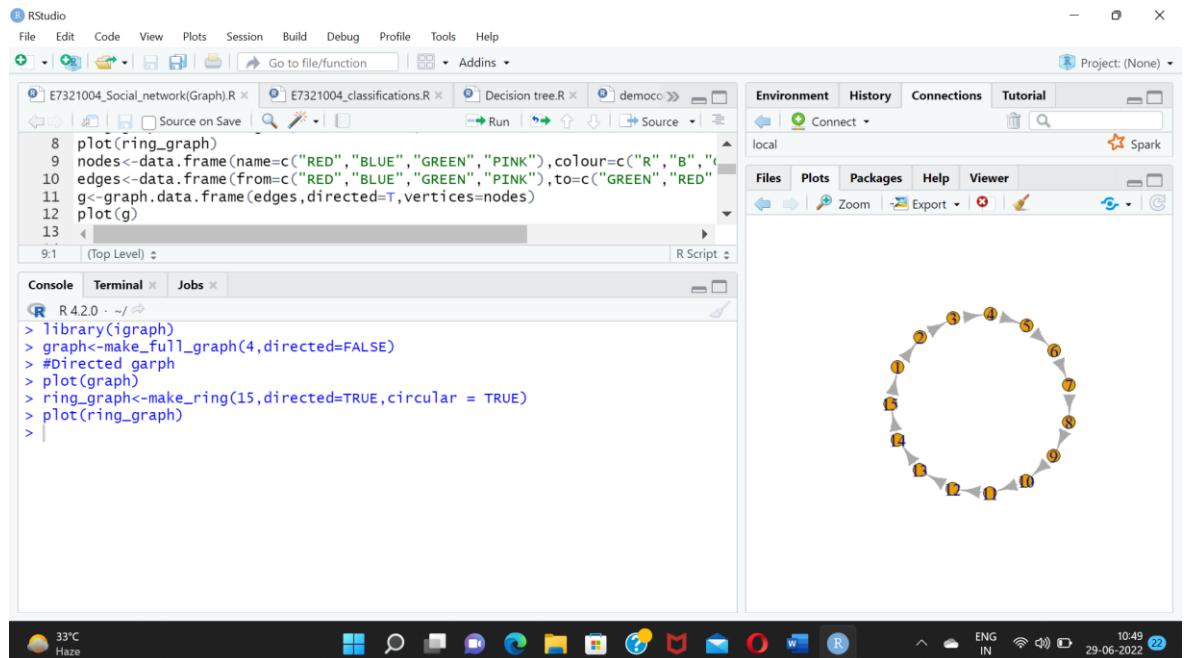
```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
+ E7321004_Social_network(Graph).R x E7321004_classifications.R x Decision tree.R x demo00 > Run Source Addins
Source on Save Go to file/function Environment History Connections Tutorial Project (None)
local Connect Spark
Files Plots Packages Help Viewer
Zoom Export R Script
Console Terminal Jobs
R 4.2.0 : ~/ ~
> library(igraph)
> graph<-make_full_graph(4,directed=FALSE)
> #Directed graph
> plot(graph)
>

```

33°C Haze

10:48 29-06-2022



R Studio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins Project: (None)

```

16 g %>% plot(vertex_size=30, vertex.color=colors, layout=layout1)
17 V(g)
18 E(g)
19 #To find the friend of whom(connectivity)
20 friends<-ego(g,order=1,nodes="RED",mindist=1)[[1]]%>%print()
21
21:1 | (Top Level) R Script

```

Console Terminal Jobs

```

R 4.2.0 - ~/ ~
> graph<-make_full_graph(4,directed=FALSE)
> #Directed graph
> plot(graph)
> ring_graph<-make_ring(15,directed=TRUE,circular = TRUE)
> plot(ring_graph)
> nodes<-data.frame(name=c("RED","BLUE","GREEN","PINK"),colour=c("R","B","G","P"),size=c(70,65,64,67))
> edges<-data.frame(from=c("RED","BLUE","GREEN","PINK"),to=c("GREEN","RED","PINK","BLUE"))
> g<-graph.data.frame(edges,directed=T,vertices=nodes)
> plot(g)
> layout1<-g%>% layout_nicely()
> g%>%plot(vertex_size=30,layout=layout1)
> colors<-ifelse(V(g)$colour=="R","red","yellow")
> g%>%plot(vertex_size=30,vertex.color=colors,layout=layout1)
>

```

33°C Haze 10:50 29-06-2022

R Studio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins Project: (None)

```

20 friends<-ego(g,order=1,nodes="RED",mindist=1)[[1]]%>%print()
21 #To find the Female
22 friends[friends$color=="R"]
23 #Degree
24 degree<-g%>%degree()%>%print()
25
25:1 | (Top Level) R Script

```

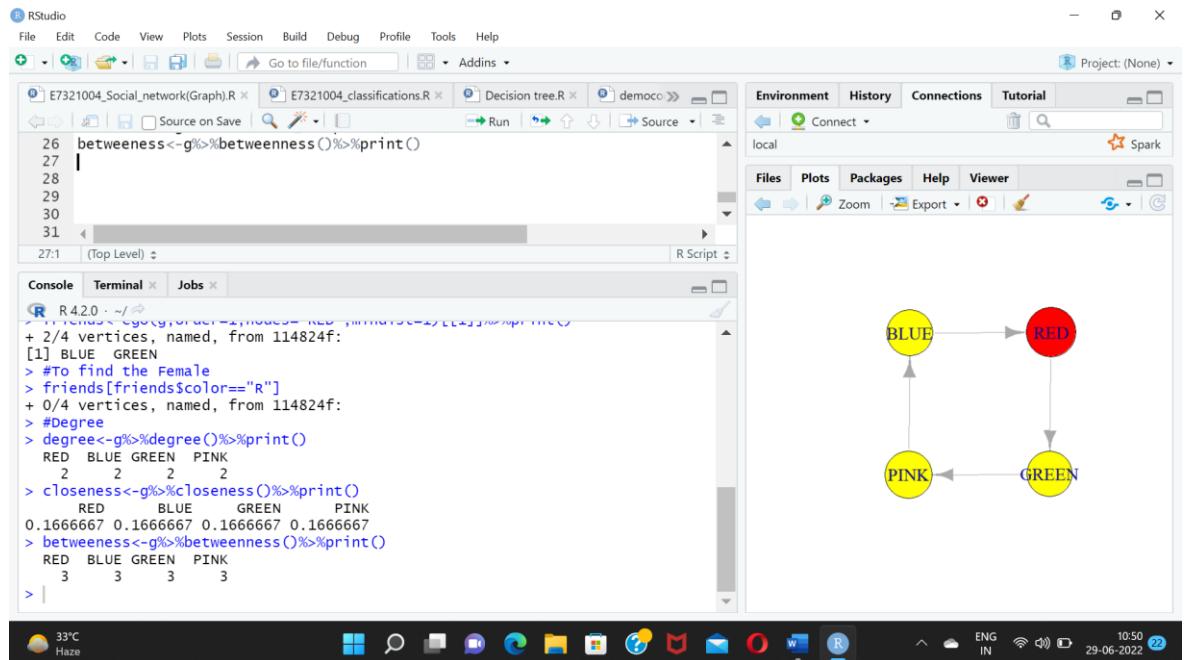
Console Terminal Jobs

```

R 4.2.0 - ~/ ~
> g<-graph.data.frame(edges,directed=TRUE,vertices=nodes)
> plot(g)
> layout1<-g%>% layout_nicely()
> g%>%plot(vertex_size=30,layout=layout1)
> colors<-ifelse(V(g)$colour=="R","red","yellow")
> g%>%plot(vertex_size=30,vertex.color=colors,layout=layout1)
> V(g)
+ 4/4 vertices, named, from 114824f:
[1] RED  BLUE  GREEN  PINK 
> E(g)
+ 4/4 edges from 114824f (vertex names):
[1] RED -->GREEN  BLUE -->RED  GREEN-->PINK  PINK -->BLUE
> #To find the friend of whom(connectivity)
> friends<-ego(g,order=1,nodes="RED",mindist=1)[[1]]%>%print()
+ 2/4 vertices, named, from 114824f:
[1] BLUE  GREEN 
>

```

33°C Haze 10:50 29-06-2022



CODE:

```

#Installing the package
install.packages("igraph")

library(igraph)

graph<-make_full_graph(4,directed=FALSE)

#Directed graph
plot(graph)

ring_graph<-make_ring(15,directed=TRUE,circular = TRUE)

plot(ring_graph)

nodes<-
data.frame(name=c("RED","BLUE","GREEN","PINK"),colour=c("R","B","G","P"),size=
c(70,65,64,67))

edges<-
data.frame(from=c("RED","BLUE","GREEN","PINK"),to=c("GREEN","RED","PINK","
BLUE"))

g<-graph.data.frame(edges,directed=T,vertices=nodes)

plot(g)

layout1<-g%>% layout_nicely()

g%>%plot(vertex_size=30,layout=layout1)

colors<-ifelse(V(g)$colour=="R","red","yellow")
    
```

```

g%>%plot(vertex_text_size=30,vertex_color=colors,layout=layout1)

V(g)

E(g)

#To find the friend of whom(connectivity)

friends<-ego(g,order=1,nodes="RED",mindist=1)[[1]]%>%print()

#To find the Female

friends[friends$color=="R"]

#Degree

degree<-g%>%degree()%>%print()

closeness<-g%>%closeness()%>%print()

betweenness<-g%>%betweenness()%>%print()

```

Perform Discrete discovery of communities where number of nodes is 100, averages degree is 50 and Number of possible edges is 1/4.

c) Perform Discrete Discovery of communities where number of nodes is 100, average degree is 50 and number of possible edges is $\frac{1}{4}$.

$$\text{nodes} = 100 \quad \text{degree} = 50$$

$$\text{edges} = \frac{1}{4}$$

Discrete discovery of communities

$$\left[n \left(\text{no. of edges possible} \right)^t \geq s \right]$$

$$100 \left(\frac{1}{4} \right)^t \geq s \quad t=1$$

$$t=2$$

$$100/16 \geq s$$

$$\left[6.25 \geq s \right]$$

$$\left[n \left(\frac{d}{n} \right)^t \geq s \right] \Rightarrow 100 \left(\frac{50}{100} \right)^t \geq s$$

$$\Rightarrow \frac{100 \times 50 \times 4^9}{100 \times 99} \geq s$$

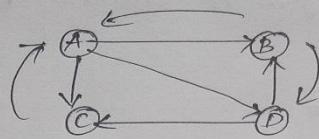
$$\Rightarrow \left[24.747 \right]$$

Approximately 25

$$\lceil k_{25} \rceil$$

D. Calculate the Page Rank for the graphs specified below

4) D) calculate Page Rank.



$$v_0 = 1/n$$

$n = \text{no. of nodes} = 4$

Transition Matrix

$$M = \begin{matrix} & \begin{matrix} A & B & C & D \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} & \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \end{matrix}$$

Iteration 1: $m \times v_0$

$$\begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$

$$\begin{bmatrix} 0 + (1/2 \times 1/4) + 1 \times 1/4 + 0 \\ 1/2 + 1/8 \\ 1/2 + 1/8 \\ 1/2 + 1/8 \end{bmatrix} \Rightarrow \begin{bmatrix} 3/8 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix} \rightarrow \begin{matrix} \times \text{ by } 2 \text{ and } \div \\ \text{by } 3 \\ \text{to get same denominator} \end{matrix}$$

$$v_1 = \begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}$$

Iteration 2:

$$[M \times V_0 = V_2]$$

$$V_2 = \begin{bmatrix} 0 & V_2 & 1 & 0 \\ V_2 & 0 & 0 & V_2 \\ V_2 & 0 & 0 & V_2 \\ V_2 & V_2 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 5/48 + 5/24 \\ 2/24 + 5/48 \\ 3/24 + 5/48 \\ 3/24 + 5/48 \end{bmatrix} \Rightarrow \begin{bmatrix} 15/48 \\ 11/48 \\ 11/48 \\ 11/48 \end{bmatrix} \Rightarrow \cancel{15/48}$$

$$V_2 = \begin{bmatrix} 15/48 \\ 11/48 \\ 11/48 \\ 11/48 \end{bmatrix}$$

Iteration 3:

$$V_3 = M \times V_2$$

$$\Rightarrow \begin{bmatrix} 0 & V_2 & 1 & 0 \\ V_2 & 0 & 0 & V_2 \\ V_2 & 0 & 0 & V_2 \\ V_2 & V_2 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 15/48 \\ 11/48 \\ 11/48 \\ 11/48 \end{bmatrix}$$

$$V_3 = \begin{bmatrix} 11/96 + 11/48 \\ 5/48 + 11/96 \\ 5/48 + 11/96 \\ 5/48 + 11/96 \end{bmatrix} \Rightarrow \begin{bmatrix} 11/32 \\ 7/32 \\ 7/32 \\ 7/32 \end{bmatrix}$$

$$V_B = \begin{bmatrix} 1/8 & 2 \\ 1/8 & 2 \\ 1/8 & 2 \\ 1/8 & 2 \end{bmatrix}$$

\therefore As the Iteration goes on, we stop at the Iteration 3.

Trust Rank for source $\{B, D\}$

$$\boxed{(1-\beta)es / |S|}$$

$$\beta = 4/5$$

$$(1-\beta) = 1 - 4/5 = 1/5$$

$es \rightarrow$ No. of Vertices in 1

$s \rightarrow$ size of S in 2

$$\text{(To get Trust rank)} = (Y_5) \times 1/2$$

$$= Y_{10}$$

$$\boxed{(1-\beta)es / |S| = Y_{10}}$$

$$\boxed{V' = BMV + (1-\beta)es / |S|} \quad (\text{Iteration 1:})$$

$$V_1 \Rightarrow \begin{bmatrix} 0 & Y_2 & 1 & 0 \\ Y_2 & 0 & 0 & Y_2 \\ Y_2 & 0 & 0 & Y_2 \\ Y_2 & Y_2 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 4/5 \\ 4/5 \\ 4/5 \\ 4/5 \end{bmatrix} \times \begin{bmatrix} 0 \\ Y_2 \\ 0 \\ Y_2 \end{bmatrix} + \begin{bmatrix} 0 \\ Y_{10} \\ 0 \\ Y_{10} \end{bmatrix}$$

$$v_1 \Rightarrow \begin{bmatrix} 0 & 4/10 & 4/5 & 0 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 4/10 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0 \\ Y_2 \\ 0 \\ Y_2 \end{bmatrix} + \begin{bmatrix} 0 \\ Y_{10} \\ 0 \\ Y_{10} \end{bmatrix}$$

$$v_1 = \begin{bmatrix} 4/20 \\ 4/20 \\ 4/20 \\ 4/20 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix} \Rightarrow \begin{bmatrix} 4/20 \\ 6/20 \\ 4/20 \\ 5/20 \end{bmatrix}$$

$$v_1 = \begin{bmatrix} 1/5 \\ -3/10 \\ 1/5 \\ 3/10 \end{bmatrix}$$

Iteration 2:

$$[PM v_1 + (I - P) es / |es|]$$

$$v_2 = \begin{bmatrix} 0 & 4/10 & 4/15 & 0 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 4/10 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 1/5 \\ 3/10 \\ 1/5 \\ 3/10 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ -0 \\ 1/10 \end{bmatrix}$$

$$v_2 = \begin{bmatrix} 12/100 & + 4/25 \\ 4/15 & + 12/100 \\ 4/15 & + 12/100 \\ 4/15 & + 12/100 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

$$v_2 = \begin{bmatrix} 28/100 \\ 52/300 \\ 52/300 \\ 52/300 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

$$v_2 \Rightarrow \begin{bmatrix} 7/25 \\ 41/150 \\ 13/75 \\ 14/150 \end{bmatrix}$$

$$\text{Iteration 3: } \boxed{PMV_2 + (1-P) e_3 / s_1}$$

$$V_3 = \begin{bmatrix} 0 & 4/10 & 4/10 & 0 \\ 4/10 & 0 & 0 & 4/10 \\ 4/10 & 0 & 0 & 4/10 \\ 4/10 & 4/10 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 1/25 \\ 4/150 \\ 12/45 \\ 4/150 \end{bmatrix} + \begin{bmatrix} 0 \\ V_{10} \\ 0 \\ V_{10} \end{bmatrix}$$

$$V_3 \Rightarrow \begin{bmatrix} \frac{164}{1500} + \frac{52}{375} \\ \frac{28}{375} + \frac{164}{1500} \\ \frac{28}{375} + \frac{164}{1500} \\ \frac{28}{375} + \frac{164}{1500} \end{bmatrix} \Rightarrow \begin{bmatrix} 93/375 \\ 23/125 \\ 23/125 \\ 23/125 \end{bmatrix} + \begin{bmatrix} 0 \\ V_{10} \\ 0 \\ V_{10} \end{bmatrix}$$

$$V_3 = \begin{bmatrix} 93/375 \\ 7/250 \\ 23/125 \\ 7/250 \end{bmatrix}$$

Spam Mass : $\frac{R-t}{R}$

$$T = \begin{bmatrix} 93/375 \\ 7/250 \\ 23/125 \\ 7/250 \end{bmatrix}$$

$$R = \begin{bmatrix} 1/32 \\ 1/32 \\ 1/32 \\ 1/32 \end{bmatrix}$$

$$\text{calculating For A} \Rightarrow \frac{0.3487 - 0.248}{0.3487} = \boxed{0.2757}$$

$$\text{calculating For B} \Rightarrow \frac{0.21875 - 0.284}{0.21875} = \boxed{-0.2982}$$

calculating For C,

$$\frac{0.21875 - 0.184}{0.21875} = \boxed{0.15885}$$

Node	Page Rank	Trust Rank	Spam Mass
A	1/32	9/375	0.2757
B	7/32	7/250	-0.2982
C	7/32	28/125	0.158857
D	7/32	7/250	-0.2982

∴ The conclusion is that the nodes D and B were determined not to be spam, as it has negative spam mass and therefore not spam.

The other nodes A & C have positive spam Mass are 0.27 and 0.2 which is closer to 0 than, so it is not spam.

Trust Rank for source {C, D}

$$\boxed{(1-\beta)e_s / |S|}$$

$$\beta = 4/5$$

$$(1-\beta) = 1 - 4/5 = 1/5$$

$e_s \rightarrow$ No. of Vectors in 1

$s \rightarrow$ size of s in 2

$$(1-\beta)e_s / |S| = 1/5 \times 1/2$$

$$\boxed{\underline{(1-\beta)e_s / |S| = 1/10}}$$

Iteration 1:

$$V_1 = BMV + (1-B)es / |s|$$

$$V_1 = \begin{bmatrix} 0 & 4/10 & 4/15 & 0 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 4/10 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0 \\ 0 \\ 1/2 \\ 1/2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1/10 \\ 1/10 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 2/5 \\ 2/10 \\ 2/10 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1/10 \\ 1/10 \end{bmatrix} \Rightarrow V_1 = \begin{bmatrix} 2/5 \\ 1/5 \\ 2/10 \\ 1/10 \end{bmatrix}$$

Iteration 2:

$$V_2 = BMV_1 + (1-B)es / |s|$$

$$\Rightarrow \begin{bmatrix} 0 & 4/10 & 4/15 & 0 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 4/10 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 2/5 \\ 1/5 \\ 2/10 \\ 1/10 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1/10 \\ 1/10 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 4/50 + 12/50 \\ 8/75 + 4/100 \\ 8/75 + 4/100 \\ 8/75 + 4/50 \end{bmatrix} \Rightarrow \begin{bmatrix} 0 \\ 0 \\ 1/10 \\ 1/10 \end{bmatrix} + \begin{bmatrix} 8/25 \\ 4/75 \\ 4/75 \\ 12/75 \end{bmatrix}$$

$$V_2 = \begin{bmatrix} 8/25 \\ 11/75 \\ 37/150 \\ 48/100 \end{bmatrix}$$

Iteration 3:

$$v_3 = B \cdot v_2 + (1-B) e^s / \|e\|$$

$$v_3 = \begin{bmatrix} 0 & 4/10 & 4/5 & 0 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 4/10 & 0 & 0 \end{bmatrix} \begin{bmatrix} 8/25 \\ 11/75 \\ 37/150 \\ 43/150 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1/10 \\ 1/10 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 44/150 + 148/150 \\ 32/375 + 172/1500 \\ 32/375 + 172/1500 \\ 32/375 + 44/150 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1/10 \\ 1/10 \end{bmatrix} \Rightarrow v_3 = \begin{bmatrix} 32/125 \\ 1/5 \\ 3/10 \\ 183/150 \end{bmatrix}$$

$$\text{Span Mass : } \frac{R - E}{R}$$

$$t = \begin{bmatrix} 32/125 \\ 1/5 \\ 3/10 \\ 183/150 \end{bmatrix} \quad (\text{Trust Rank})$$

$$R = \text{Page Rank} \begin{bmatrix} 1/32 \\ 1/32 \\ 1/32 \\ 1/32 \end{bmatrix}$$

calculating for A,

$$\frac{0.345 - 0.256}{0.343} \Rightarrow 0.2536$$

$$\text{calculating for B, } \frac{0.21875 - 0.2}{0.21875} = 0.08571$$

$$\text{calculating for C, } \frac{0.21875 - 0.3}{0.21875} = -0.3714$$

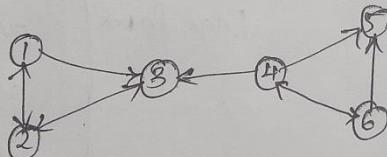
$$\text{calculating for D, } \frac{0.21875 - 0.244}{0.21875} = -0.1154$$

Node	Page Rank (R)	Trust Rank (t)	$R - \epsilon / R$
A	$\frac{1}{32}$	$\frac{32}{125}$	0.2531
B	$\frac{1}{32}$	$\frac{1}{5}$	0.08571
C	$\frac{1}{32}$	$\frac{3}{10}$	-0.3714
D	$\frac{1}{32}$	$\frac{183}{750}$	-0.1184

∴ The conclusion is that the Nodes C and D were determined to be not spam, as it has negative spam mass and therefore not spam.

The other two nodes A and B have positive spam mass 0.25 and 0.08 which is close to 0 than 1 so it is not spam.

Page Rank,



no. of nodes = 6

$n = 6$

$$v_0 = \frac{1}{6}$$

Transition Matrix

$$m = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 2 & \frac{1}{2} & 0 & 1 & 0 & 0 & 0 \\ 3 & \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{3} & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 5 & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{2} \\ 6 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 \end{bmatrix}$$

Iteration 1 : $m \times v_0$

$$v_1 = \begin{bmatrix} 0 & y_2 & 0 & 0 & 0 & 0 \\ y_2 & 0 & 1 & 0 & 0 & 0 \\ y_2 & y_2 & 0 & y_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & y_2 \\ 0 & 0 & 0 & y_3 & 0 & y_2 \\ 0 & 0 & 0 & y_3 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_6 \\ y_6 \\ y_6 \\ y_6 \\ y_6 \\ y_6 \end{bmatrix} \Rightarrow \begin{bmatrix} y_{12} \\ y_{12} + y_6 \\ y_{12} + y_{12} + y_{18} \\ y_{12} \\ y_{18} + y_{12} \\ y_{18} \end{bmatrix}$$

$$v_1 = \begin{bmatrix} y_{12} \\ 3y_{12} \\ 8/36 \\ y_{12} \\ 5/36 \\ y_{18} \end{bmatrix} \Rightarrow \begin{bmatrix} y_{12} \\ y_4 \\ 2/9 \\ y_{12} \\ 5/36 \\ y_{18} \end{bmatrix}$$

Iteration 2 : $m \times v_1$

$v_2 = m \times v_1$

$$v_2 = \begin{bmatrix} 0 & y_2 & 0 & 0 & 0 & 0 \\ y_2 & 0 & 1 & 0 & 0 & 0 \\ y_2 & y_2 & 0 & y_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & y_2 \\ 0 & 0 & 0 & y_3 & 0 & y_2 \\ 0 & 0 & 0 & y_3 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} y_{12} \\ 3/12 \\ 8/36 \\ y_{12} \\ 5/36 \\ y_{18} \end{bmatrix}$$

$$v_2 \Rightarrow \begin{bmatrix} 3/24 \\ y_{24} + 8/36 \\ y_{24} + 3/24 + y_{36} \\ y_{36} \\ y_{36} + y_{36} \end{bmatrix} \Rightarrow \begin{bmatrix} 3/24 \\ 19/72 \\ 14/72 \\ y_{36} \\ 2y_{36} \\ y_{36} \end{bmatrix}$$

Iteration 3:

$$V_3 = m \times V_2$$

$$V_3 = \begin{bmatrix} 0 & V_2 & 0 & 0 & 0 & 0 \\ V_2 & 0 & 1 & 0 & 0 & 0 \\ V_2 & V_2 & 0 & V_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & V_2 \\ 0 & 0 & 0 & V_2 & 0 & V_2 \\ 0 & 0 & 0 & V_2 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 3/50 \\ 19/72 \\ 14/72 \\ V_{30} \\ V_{18} \\ V_{36} \end{bmatrix}$$

$$V_3 = \begin{bmatrix} 19/144 \\ 3/48 + 14/72 \\ 3/48 + 19/144 + V_{108} \\ 1/72 \\ V_{108} + V_{12} \\ V_{108} \end{bmatrix}$$

$$V_3 = \begin{bmatrix} 19/144 \\ 3/48/144 \\ 1/54 \\ 1/72 \\ 5/216 \\ V_{108} \end{bmatrix}$$

\therefore Since the iteration does not converge we stop at Iteration 3.

Trust Rank

$$n = 6$$

$$\text{Source} = \{2, 6, 7\}$$

$$\beta = 4/5 \quad (1-\beta) = 1 - 4/5 = [1/5]$$

$$s = 2 \quad \left[(1-\beta)^{\infty}/\beta \right] = V_{10}$$

Iteration 1:

$$\boxed{V_1 = BMV_0 + (I - B)e^s / (\beta s)}$$

$$V_1 = \begin{bmatrix} 0 & 4/10 & 0 & 0 & 0 & 0 \\ 4/10 & 0 & 4/15 & 0 & 0 & 0 \\ 4/10 & 4/10 & 0 & 4/15 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4/10 \\ 0 & 0 & 0 & 4/15 & 0 & 4/10 \\ 0 & 0 & 0 & 4/15 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0 \\ V_2 \\ 0 \\ 0 \\ 0 \\ V_2 \end{bmatrix} + \begin{bmatrix} 0 \\ V_{10} \\ 0 \\ 0 \\ 0 \\ V_{10} \end{bmatrix}$$

$$V_1 = \begin{bmatrix} V_5 \\ 0 \\ V_{15} \\ V_5 \\ V_5 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ V_{10} \\ 0 \\ 0 \\ 0 \\ V_{10} \end{bmatrix} = \begin{bmatrix} V_5 \\ V_{10} \\ V_{15} \\ V_5 \\ V_5 \\ V_{10} \end{bmatrix}$$

Iteration 2:

$$\boxed{V_2 = BMV_1 + (I - B)e^s / (\beta s)}$$

$$V_2 = \begin{bmatrix} 0 & 4/10 & 0 & 0 & 0 & 0 \\ 4/10 & 0 & 4/15 & 0 & 0 & 0 \\ 4/10 & 4/10 & 0 & 4/15 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4/10 \\ 0 & 0 & 0 & 4/15 & 0 & 4/10 \\ 0 & 0 & 0 & 4/15 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} V_5 \\ V_{10} \\ V_{15} \\ V_5 \\ V_5 \\ V_{10} \end{bmatrix} \Rightarrow \begin{bmatrix} V_{25} \\ 6/25 \\ 13/75 \\ V_{25} \\ V_{75} \\ 4/75 \end{bmatrix} + \begin{bmatrix} 0 \\ V_{10} \\ 0 \\ 0 \\ 0 \\ V_{10} \end{bmatrix}$$

$$v_2 = \begin{bmatrix} 9/25 \\ 17/50 \\ 13/75 \\ 1/25 \\ 7/25 \\ 23/150 \end{bmatrix}$$

$$v_3 = [BMv_2 + (1-B)e^s / \|s\|] \quad \text{Iteration 3:}$$

$$v_3 = \begin{bmatrix} 0 & 4/10 & 0 & 0 & 0 & 0 \\ 4/10 & 0 & 4/15 & 0 & 0 & 0 \\ 4/10 & 4/10 & 0 & 4/15 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4/10 \\ 0 & 0 & 0 & 4/15 & 0 & 4/10 \\ 0 & 0 & 0 & 4/15 & 0 & 0 \end{bmatrix} \begin{bmatrix} 9/25 \\ 17/50 \\ 13/75 \\ 1/25 \\ 7/25 \\ 23/150 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 0 \\ 0 \\ 1/10 \end{bmatrix}$$

$$v_3 = \begin{bmatrix} 17/25 \\ -4/250 + 52/375 \\ 4/250 + \frac{34}{250} + 9/375 \\ 9/1500 \\ 4/375 + 9/1500 \\ 4/375 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 0 \\ 0 \\ 1/10 \end{bmatrix} \Rightarrow v_8 = \begin{bmatrix} 17/125 \\ 191/150 \\ 61/375 \\ 23/375 \\ 27/375 \\ 83/150 \end{bmatrix}$$

Spam Mass: $\frac{R-t}{R}$ $R = \text{page rank}$
 $t = \text{trust rank}$

$$R = \begin{matrix} 19/144 \\ 37/144 \\ 11/54 \\ 17/72 \\ 5/216 \\ 1/108 \end{matrix}$$

$$T = \begin{matrix} 17/125 \\ 191/750 \\ 61/375 \\ 23/375 \\ 27/375 \\ 83/750 \end{matrix}$$

$$\text{calculation for } ① = \frac{0.13944 - 0.136}{0.13944} = [-0.029]$$

$$\text{calculation for } ② = \frac{0.25674 - 0.25466}{0.25674} = [0.0088]$$

$$\text{calculation for } ③ = \frac{0.2037 - 0.1626}{0.2037} = [0.20176]$$

$$\text{calculation for } ④ = \frac{0.01388 - 0.0613}{0.01388} = [-3.416]$$

$$\text{calculation for } ⑤ = \frac{0.02314 - 0.72}{0.02314} = [-30.444]$$

$$\text{calculation for } ⑥ = \frac{0.00095 - 0.1106}{0.00095} = [-10.95]$$

Node	Page Rank (R)	Tauent Rank (T)	spam
1	19/144	17/125	-0.029
2	37/144	191/750	0.0088
3	11/54	61/375	0.20176
4	17/72	23/375	-3.416
5	5/216	27/375	-30.444
6	1/108	83/750	-10.95

∴ since most of the value are negative and nearly 0
It is not spam.

5. Perform the following analysis on the Breast Cancer dataset.

A. Perform reading of the dataset in the following formats.

i) Read the dataset in .txt file and print the structure of the dataset.

The screenshot shows the RStudio interface with the following details:

- Code Editor:** Contains R script code for reading a dataset and printing its statistics.
- Console:** Displays the output of the R script, showing the first few rows of the dataset and the result of the `str(data1)` command.
- Environment:** Shows the local environment with the `sc` variable defined.
- Plots:** No plots are present.
- Packages:** No packages are listed.
- Help:** No help pages are listed.
- Viewer:** No files are listed.

The screenshot shows the RStudio interface with the following details:

- Code Editor:** Continues the R script from the previous screenshot.
- Console:** Displays the output of the R script, showing the first few rows of the dataset and the result of the `str(data1)` command.
- Environment:** Shows the local environment with the `sc` variable defined.
- Plots:** No plots are present.
- Packages:** No packages are listed.
- Help:** No help pages are listed.
- Viewer:** No files are listed.

ii) Read the dataset in .csv file format and print the statistics of the dataset.

The screenshot shows the RStudio interface with the following components:

- File Explorer:** Shows files like "1).R", "Collaborative filtering based on recom...", "E7321004_FA_BIG_DATA.R", and "movielens".
- Code Editor:** Displays R code for reading datasets and connecting to a local Spark cluster.
- Data View:** Shows a data frame with columns: age, albumin, alk_phos, ascites, bilirubin, cholesterol, edema, edema_tmt, and edema_tm.
- Console:** Displays R session output, including the loading of R 4.2.0 and the execution of R code to read and print a dataset.
- Environment:** Shows the current environment variables.
- Connections:** Shows the connection to a local Spark cluster.
- Viewer:** Shows the data frame from the Data View.

The screenshot shows the RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Addins:** A dropdown menu showing various addins.
- Project:** Project: (None)
- Code Editor:** An R script named "1).R" containing code to read a CSV file, print its summary, and use sparklyr to connect to a local Spark master.
- Console:** Displays the output of running the R script, including a summary of the "data2" dataset and a table of descriptive statistics for various variables like age, albumin, alk_phos, ascites, etc.
- Environment:** Shows the local environment with a library(sparklyr) and sc <- spark_connect(master = "local") entry.
- Files:** A tab bar with Files, Plots, Packages, Help, and Viewer.

iii) Download a dataset in .xml format and convert it to a dataframe.

The screenshot shows the RStudio interface. In the top-left pane, a script editor window displays R code for reading an XML file and converting it to a DataFrame. The code includes `install.packages("XML")`, `library(XML)`, and `results<-xmlToDataFrame("C:\\Users\\Brinda\\Downloads\\student.xml")`. The top-right pane shows the Environment tab with a local workspace containing a `spark` object. The bottom-left pane is the Console, showing the output of the R code, which prints the XML structure of the student data. The bottom-right pane is the Spark tab, showing the command `library(sparklyr)` and `sc <- spark_connect(master = "local")`.

```
14 install.packages("XML")
15 library(XML)
16 data3<-xmlParse("C:\\Users\\Brinda\\Downloads\\student.xml")
17 data3
18 results<-xmlToDataFrame("C:\\Users\\Brinda\\Downloads\\student.xml")
19 #converting the xml file to a data format
20 results
21 #iv)Download the dataset in .dbf format and display the data
22 
18:1 (Top Level) ▾
```

```
> data3
<?xml version="1.0" encoding="utf-8"?>
<STUDENTS>
  <STUDENT>
    <StudentID>1</StudentID>
    <Name>John Smith</Name>
    <Marks>200</Marks>
  </STUDENT>
  <STUDENT>
    <StudentID>2</StudentID>
    <Name>Mark Johnson</Name>
    <Marks>300</Marks>
  </STUDENT>
</STUDENTS>
```

This screenshot is similar to the one above, but it includes additional steps. After reading the XML file, the code uses `library(foreign)` and `write.dbf` to save the data to a dbf file named `sample.dbf`. The console output shows the XML structure and the resulting dbf file content, which is a data frame with columns `StudentID`, `Name`, and `Marks` containing two rows of data for students 1 and 2.

```
17 data3
18 #converting the xml file to a data format
19 results<-xmlToDataFrame("C:\\Users\\Brinda\\Downloads\\student.xml")
20 results
21 #iv)Download the dataset in .dbf format and display the data
22 library(foreign)
23 data4<-read.dbf("c:\\Users\\Brinda\\Downloads\\sample.dbf")
24 str(data4)
25 
21:1 (Top Level) ▾
```

```
R 4.2.0 - ~/ ...
<StudentID>2</StudentID>
<Name>Mark Johnson</Name>
<Marks>300</Marks>
</STUDENT>
</STUDENTS>

> #converting the xml file to a data format
> results<-xmlToDataFrame("C:\\Users\\Brinda\\Downloads\\student.xml")
> results
  StudentID      Name Marks
1       1 John Smith    200
2       2 Mark Johnson   300
> |
```

iv)Download the dataset in .dbf format and display the data.

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins Project: (None)
1.R x Collaborative filtering based on recom... E7321004_FA_BIG_DATA.R movielens ...
23 data4<-read.dbf("C:\\Users\\Brinda\\Downloads\\sample.dbf")
24 str(data4)
25 print(data4)
26
27 #v) Plot the graph using ggplotgui on various attributes and its
28 #relationship with class label.
29 library(devtools)
30 install.packages("devtools")
31 <-- (Top Level) R Script
Console Terminal x Jobs x
R 4.2.0 ~/ ~
'data.frame': 279 obs. of 4 variables:
$ LEISTNR: int 1 2 3 15 100 200 300 999999 999999 1 ...
$ LEISTUNG: Factor w/ 266 levels "Drehb.x8lhne pro sendung",...: 67 140 184 51
252 111 110 13 130 161 ...
$ PREIS : num 44 48 100 44 36.9 44 50 0 0 ...
$ KNAME : Factor w/ 98 levels "B\\x81ro", "B\\x81hgw",...: 45 61 69 45 98 36 32
87 65 64 ...
- attr(*, "data_types")= chr [1:4] "N" "C" "N" "C"
> print(data4)
  LEISTNR LEISTUNG      PREIS    KNAME
1       1   das ist doch keine leistung 44.00    hw
2       2           Meister        48.00    ms
3       3   Programmierer     100.00  progr
4      15 asdfasdfafas        44.00    hw

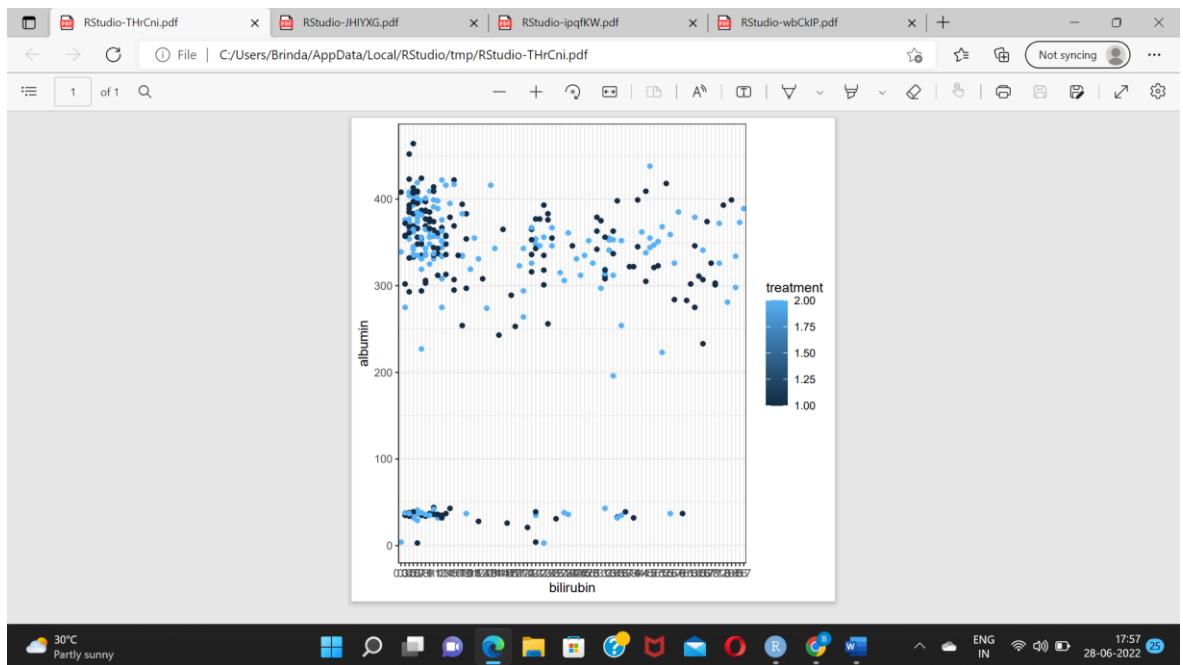
```

v) Plot the graph using ggplotgui on various attributes and its relationship with class label.

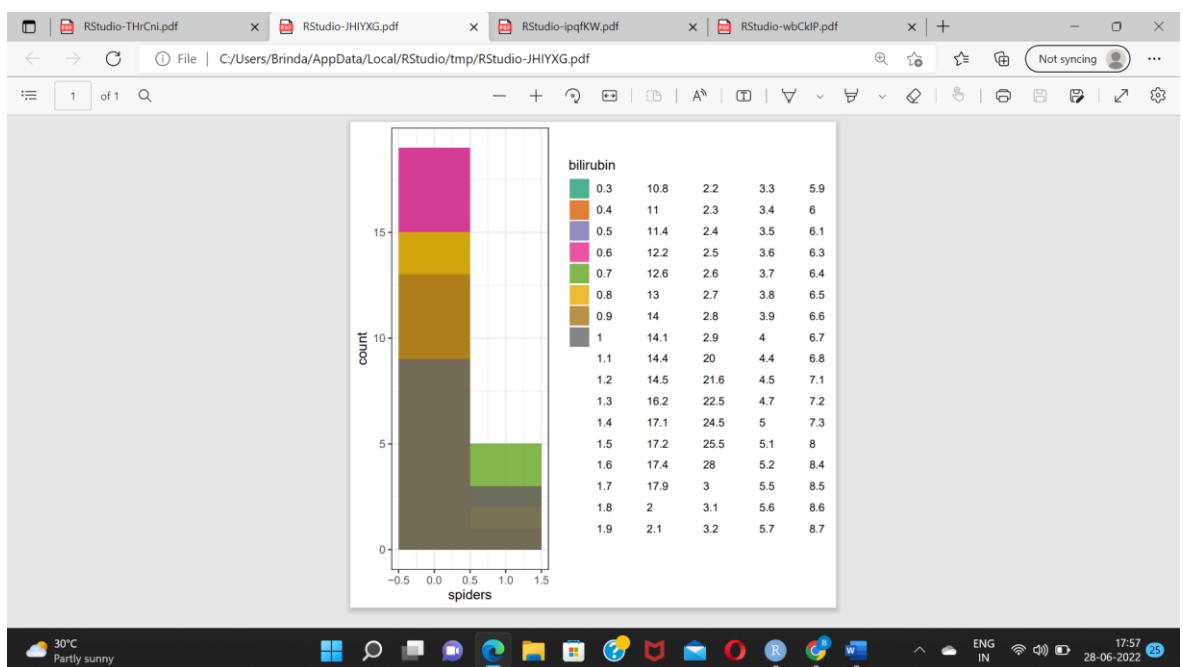
DATASET : PBCshort csv file

	age	albumin	alk_phos	ascites	bilirubin	cholesterol	edema	edema
1	587652	26	1718	1	14.5	261	1	1
2	564463	414	73948	0	1.1	302	0	0
3	700726	348	516	0	1.4	176	1	0.5
4	547406	254	61218	0	1.8	244	1	0.5
5	381054	353	671	0	3.4	279	0	0
6	662587	398	944	0	0.8	248	0	0
7	555346	409	824	0	1	322	0	0

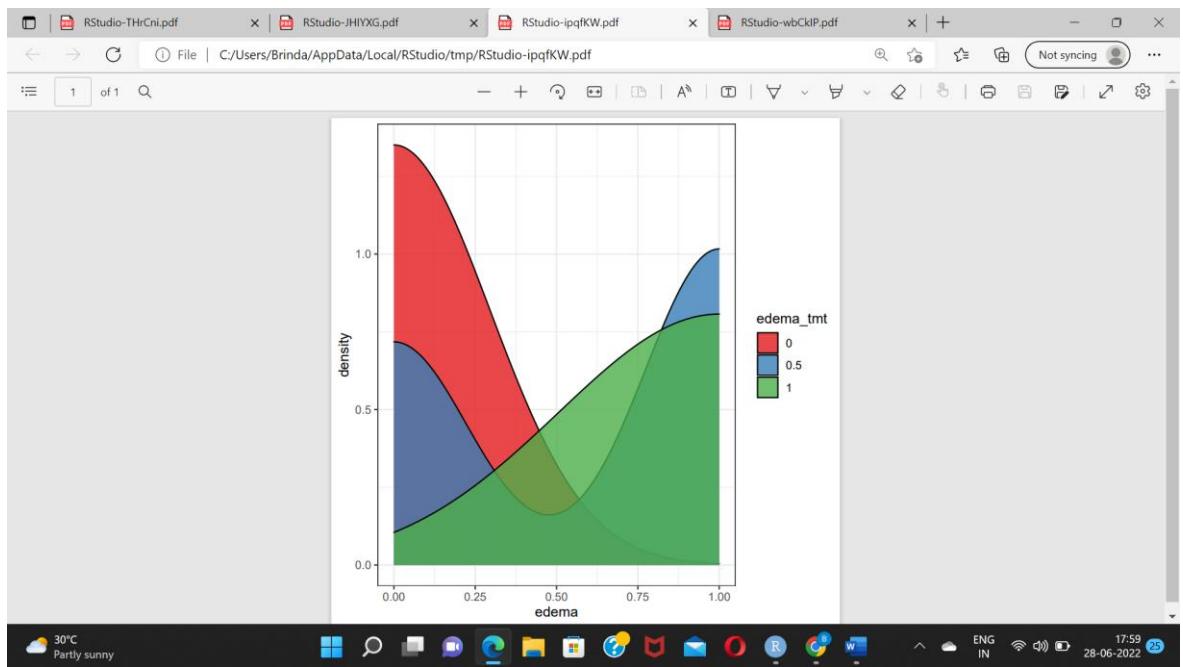
SCATTER PLOT



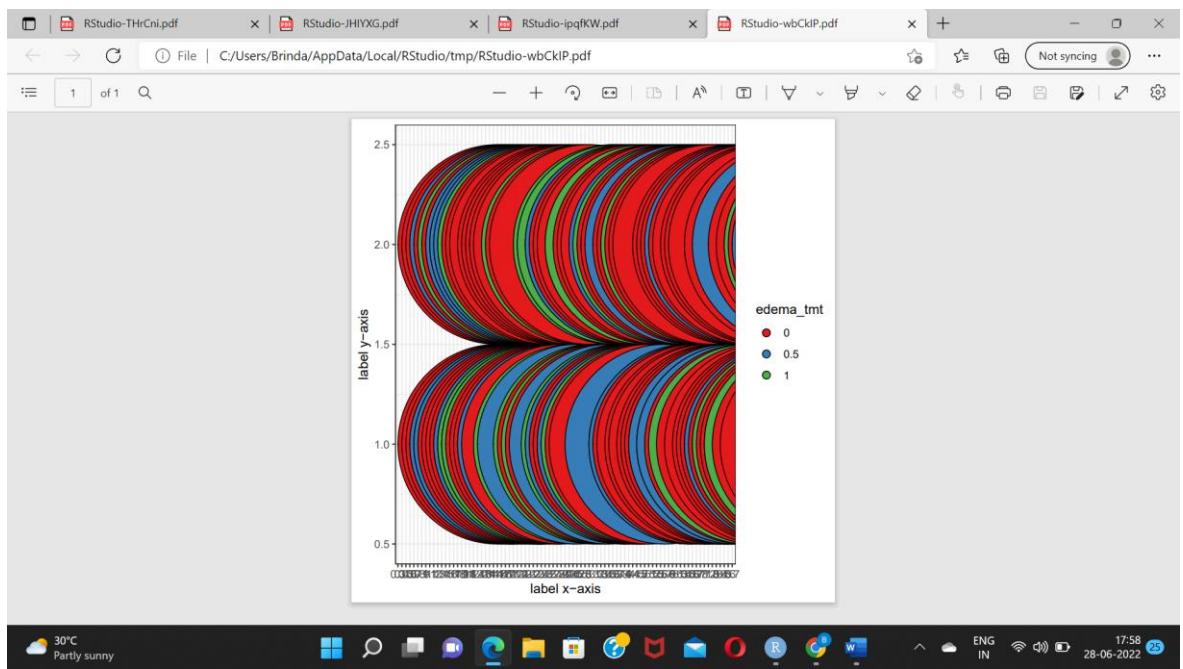
HISTOGRAM



DENSITY PLOT



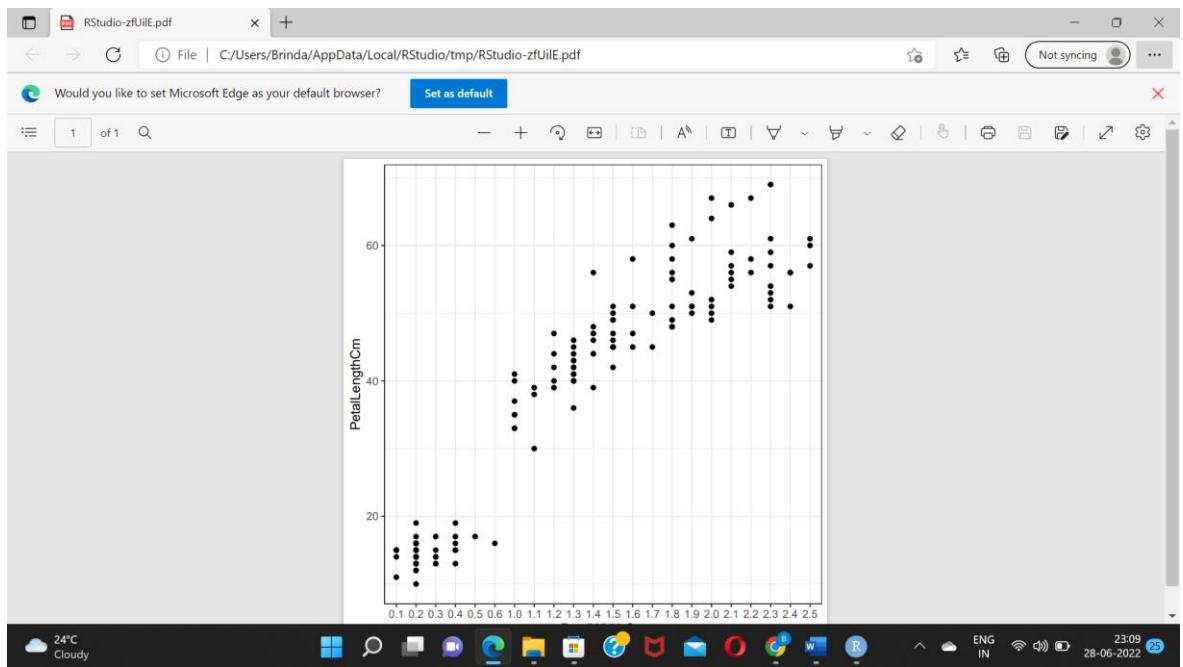
DOT PLOT



B. Formulate any two-research question on the data set and perform the relevant visualization technique to find the solution for the same and also draw statistical inferences on the same.

RESEARCH QUESTION 1:

To Build a simple Regression model that we can use to predict the petal width by establishing a Statistically significant linear relationship with petal Length.

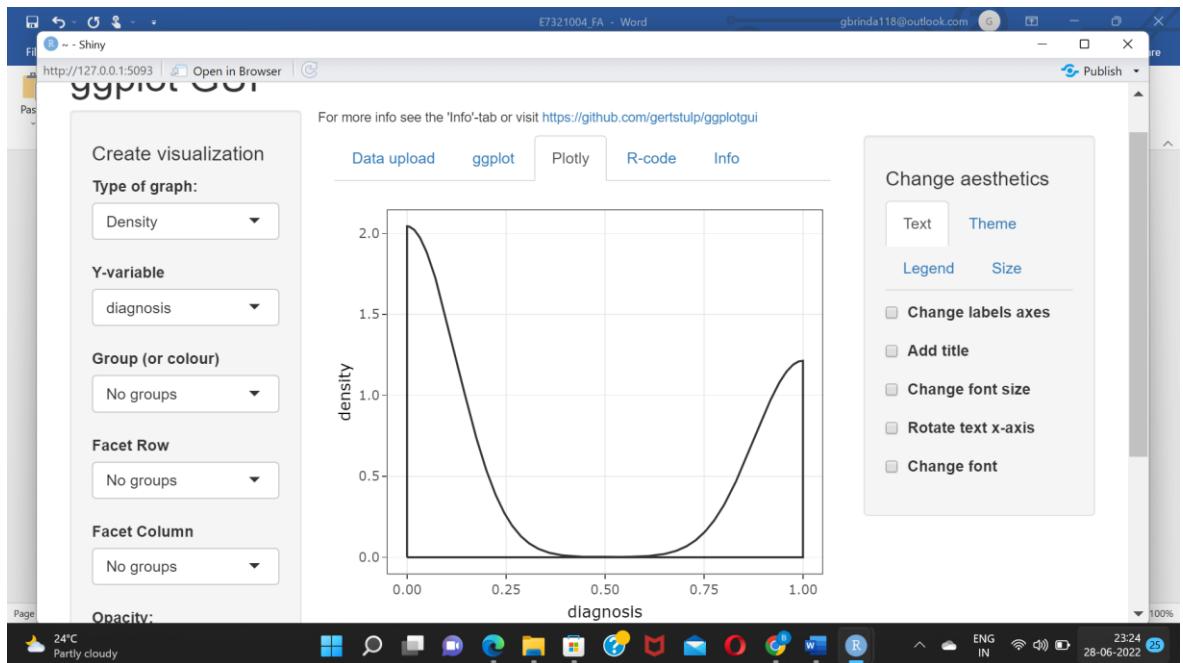


INTERPRETATION:

The slope is 0.4158, This states that on average for every 1cm increases in length petal Length, The petal width increases by 0.4158 cm

RESEARCH QUESTION 1:

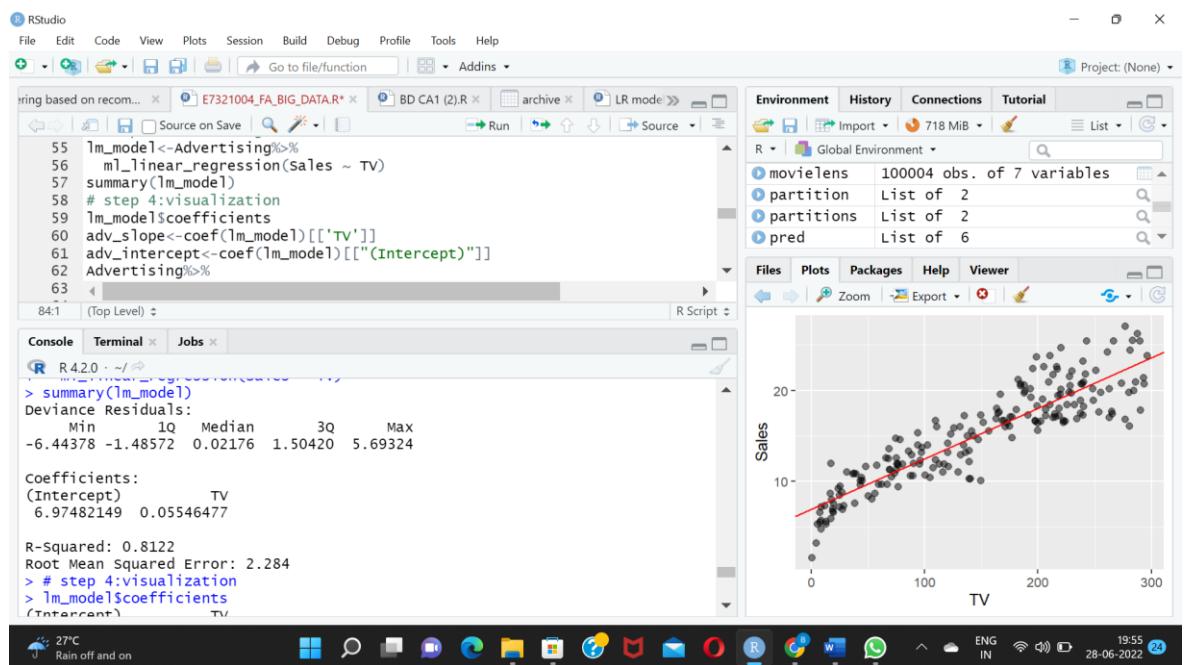
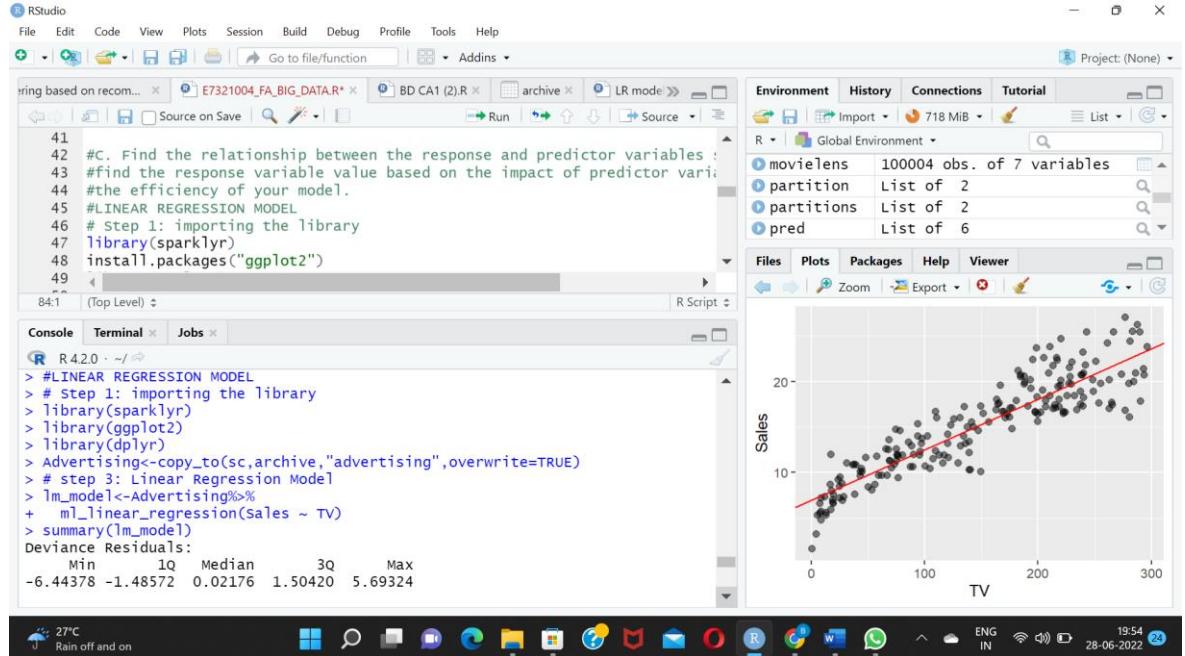
To Build a Binary Classification model on breast cancer dataset to analyse the data on which the patient has cancer or not.

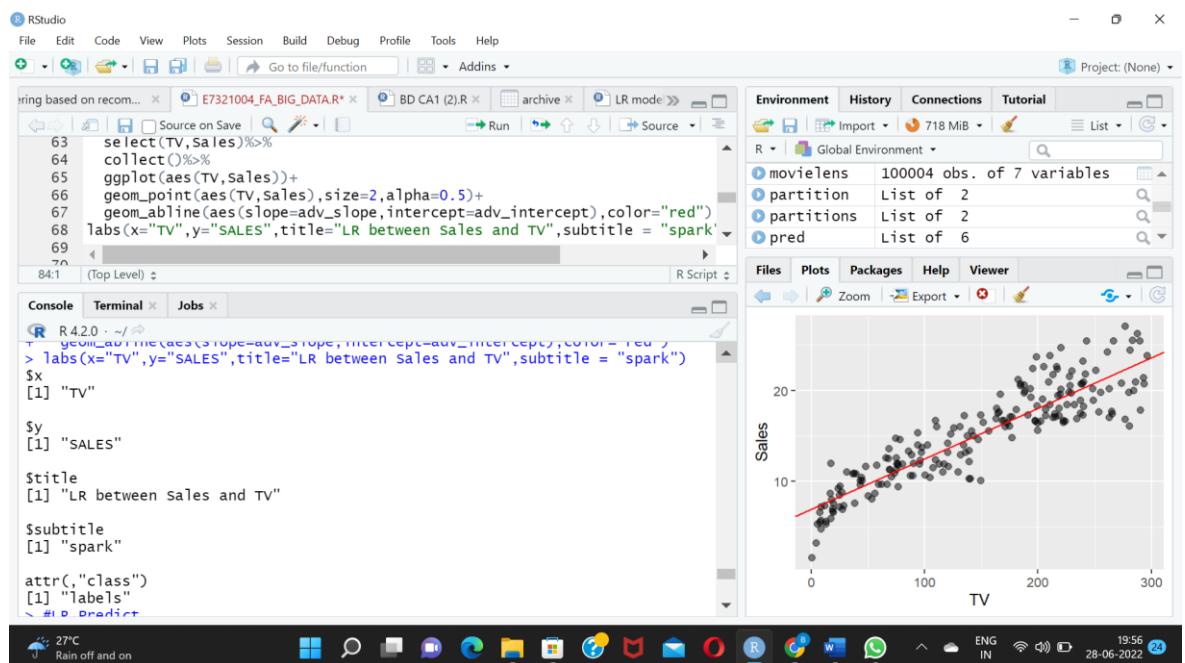
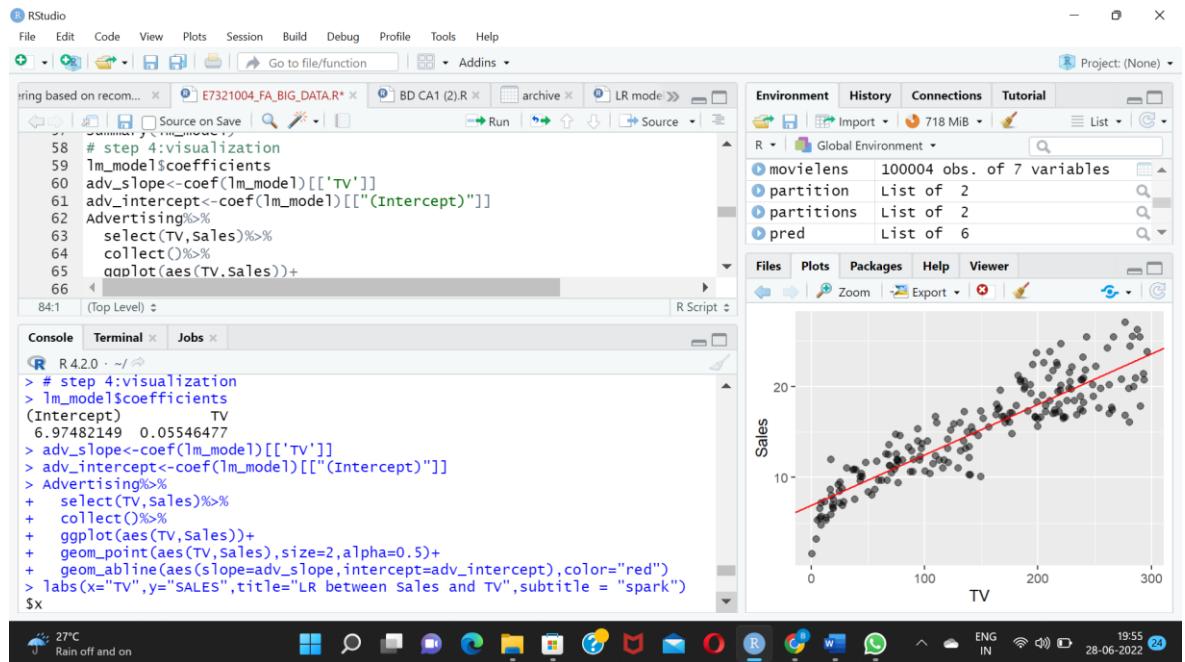


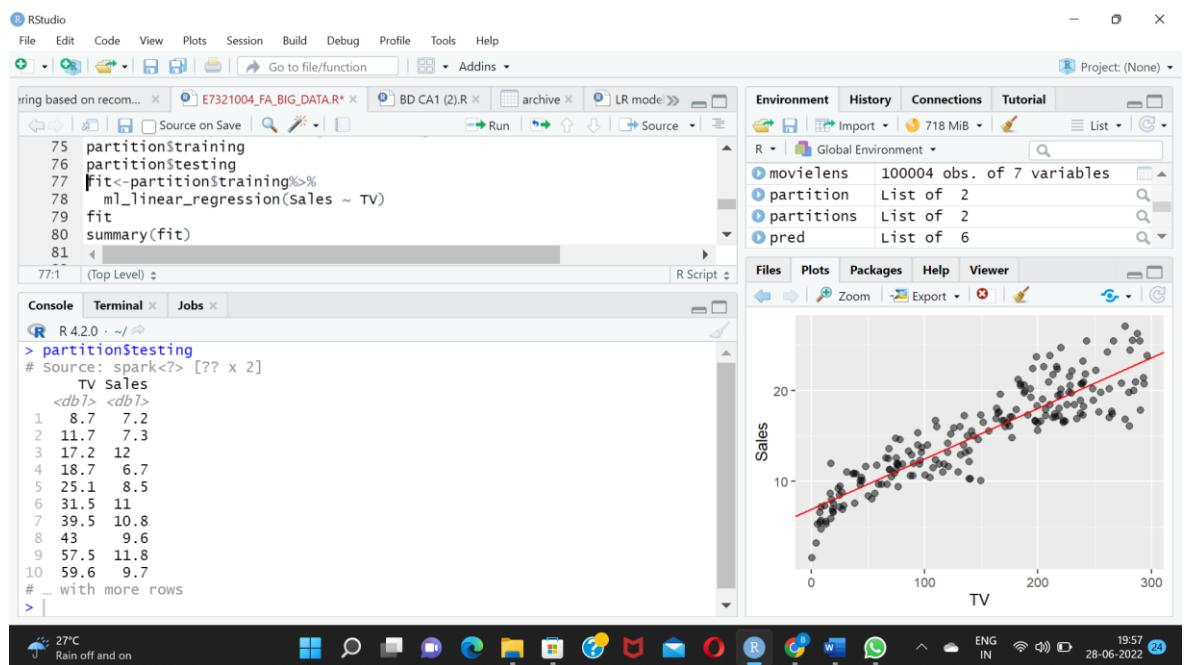
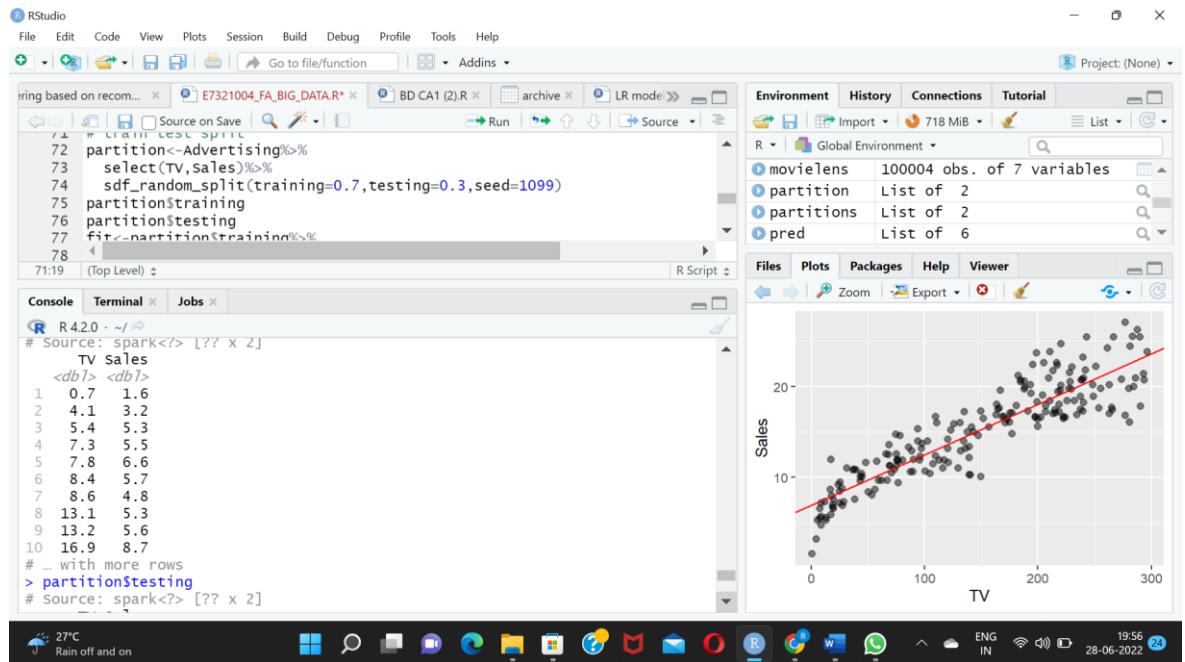
INTERPRETATION:

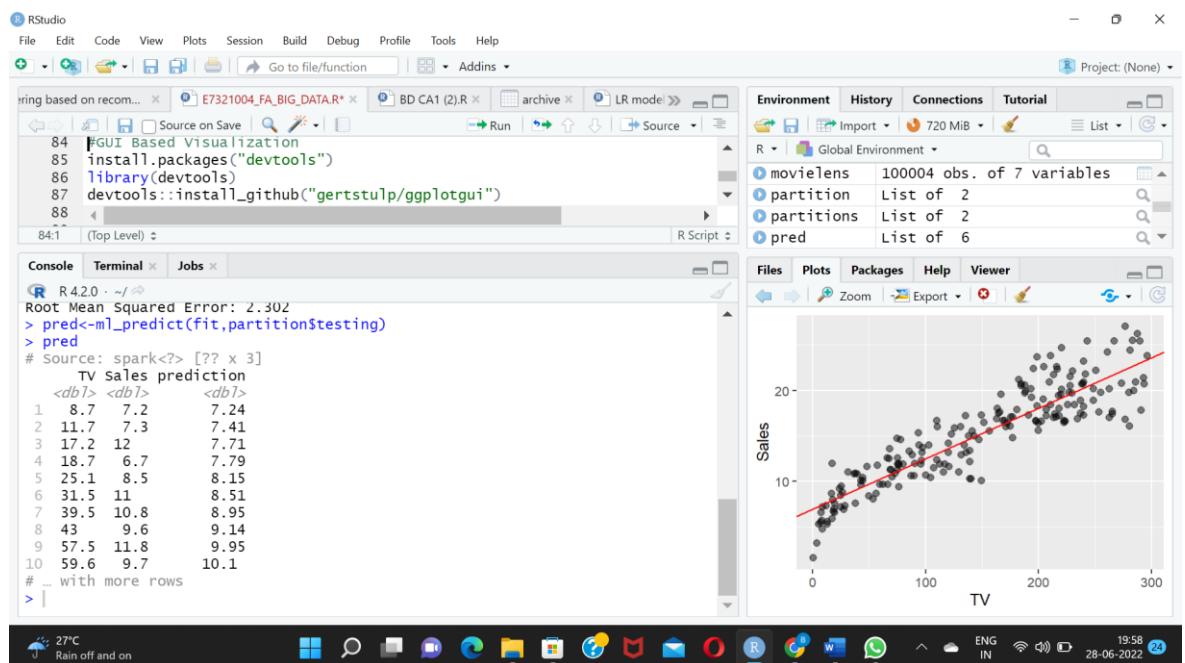
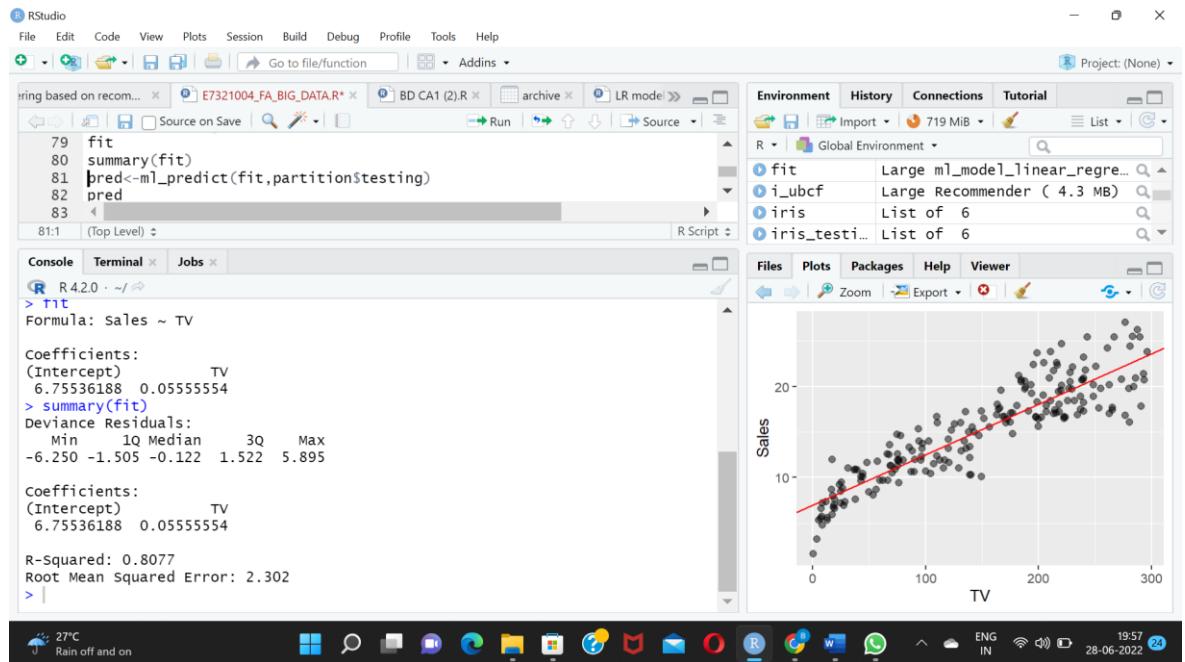
Using the Visualization we can interpret that Most of the patients are not affected by cancer (The value 2 indicates not having cancer).But still half of the patients are Affected by cancer.

C. Find the relationship between the response and predictor variables statistically and find the response variable value based on the impact of predictor variables. Analyse the efficiency of your model.





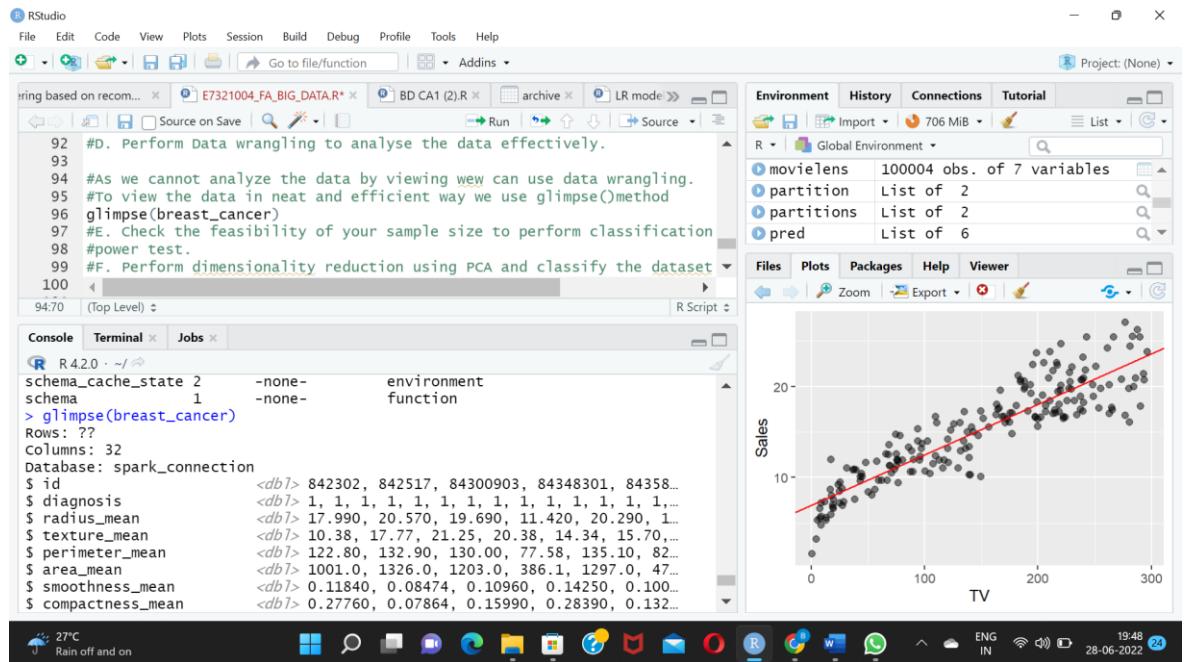




D. Perform Data wrangling to analyse the data effectively.

Data wrangling called data cleaning, data remediation, refers to **a variety of processes designed to transform raw data into more readily used formats**

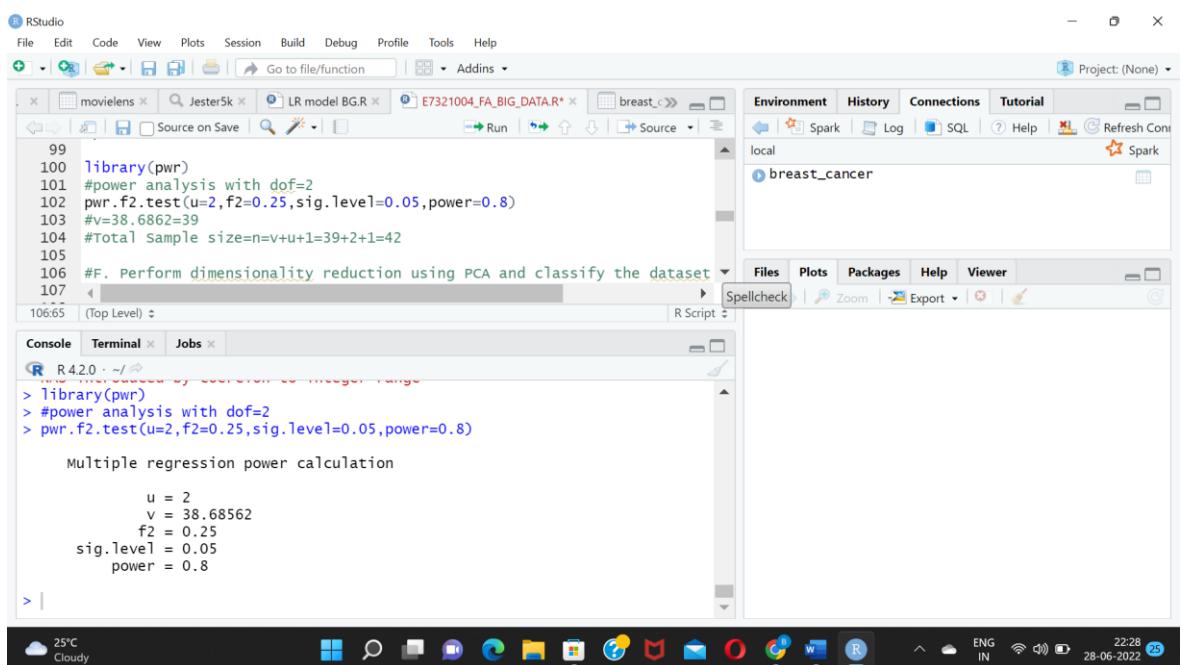
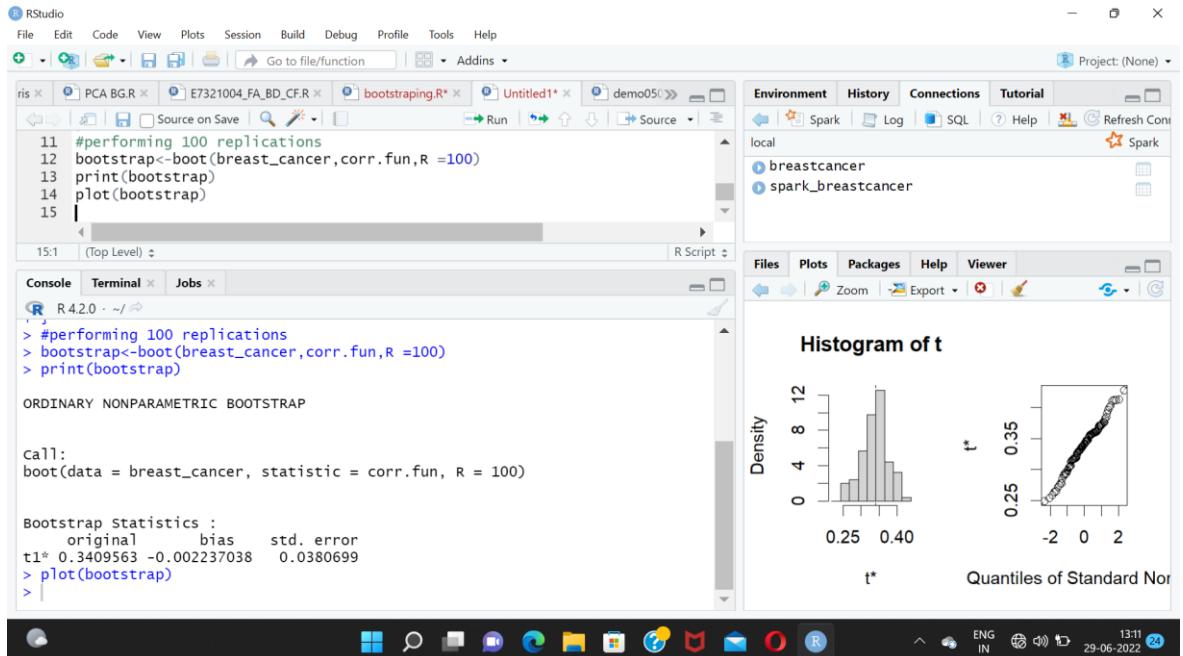
Using glimpse () we can read the data in a neat format.



E. Check the feasibility of your sample size to perform classification using Bootstrap and power test.

CODE:

```
#Bootstrap
library(sparklyr)
sc<-spark_connect(master="local")
data<-spark_read_csv(sc, "spark_breastcancer","C:\\Users\\Brinda\\Downloads\\breast
cancer.csv")
BC_tbl <- copy_to(sc,breast_cancer, "BreastCancer", overwrite = TRUE)
library(boot)
corr.fun<-function(data,idx){
  df<-data[idx,]
  c(cor(df[,3],df[,4],method='spearman'))
}
#performing 100 replications
bootstrap<-boot(breast_cancer,corr.fun,R =100)
print(bootstrap)
plot(bootstrap)
boot.ci(bootstrap,index = 1)
```



F. Perform dimensionality reduction using PCA and classify the dataset

PCA using logistic Regression CODE: (Breast Cancer Datset)

```
library(sparklyr)
```

```
sc<-spark_connect(master="local")
```

PCA Using logistic Regression Model

```
breast_cancer <- copy_to(sc, breast_cancer, "breast_cancer", overwrite = TRUE)
```

```
partition<- breast_cancer%>%
```

```
sdf_random_split(training=0.7,testing=0.3,seed=111)
```

```

breast_cancer_training<-partition$training
breast_cancer_testing<-partition$testing
lr_model<-breast_cancer_training%>%
  ml_logistic_regression(diagnosis ~.)
pred<-ml_predict(lr_model,breast_cancer_testing)
ml_binary_classification_evaluator(pred)
predict<-ml_predict(lr_model,breast_cancer_testing)%>%
  collect()
table(pred$diagnosis,pred$prediction)

glm_model <- breast_cancer %>%
  select_if(is.numeric) %>%
  ml_logistic_regression(diagnosis ~.)
summary(glm_model)
pred<-ml_predict(lr_model,breast_cancer_testing)
ml_binary_classification_evaluator(pred)
ml_predict(glm_model, breast_cancer_testing) %>%
  count(diagnosis, prediction)
pca_model <- tbl(sc, "breast_cancer") %>%
  select(-diagnosis) %>%
  ml_pca()

pca_model

```

```

132 count(diagnosis, prediction)
133 pca_model <-tbl(sc, "breast_cancer") %>%
134 select(-diagnosis) %>%
135 ml_pca()
136
137 pca_model
138
139 #G. Without dimensionality reduction, perform any three-classification
140
141
139.1 (Top Level) R Script

```

Console R 4.2.0 ~/

```

1.110223e-16 1.110223e-16 1.110223e-16 1.110223e-16
PC21 PC22 PC23 PC24 PC25
1.110223e-16 1.110223e-16 1.110223e-16 1.110223e-16 1.110223e-16
PC26 PC27 PC28 PC29 PC30
1.110223e-16 1.110223e-16 1.110223e-16 1.110223e-16 8.804741e-17
PC31
5.832333e-17

Rotation:
PC1 PC2 PC3
id -1.000000e+00 5.620736e-07 2.021130e-08
radius_mean -2.103552e-09 -5.101431e-03 9.258356e-03
texture_mean -3.432346e-09 -2.152559e-03 -2.790511e-03
perimeter_mean -1.421925e-08 -3.518688e-02 6.253861e-02

```

G. Without dimensionality reduction, perform any three-classification algorithm of your choice using sparklyr and analyse the performance of the algorithms

LOGISTIC REGRESSION :

```

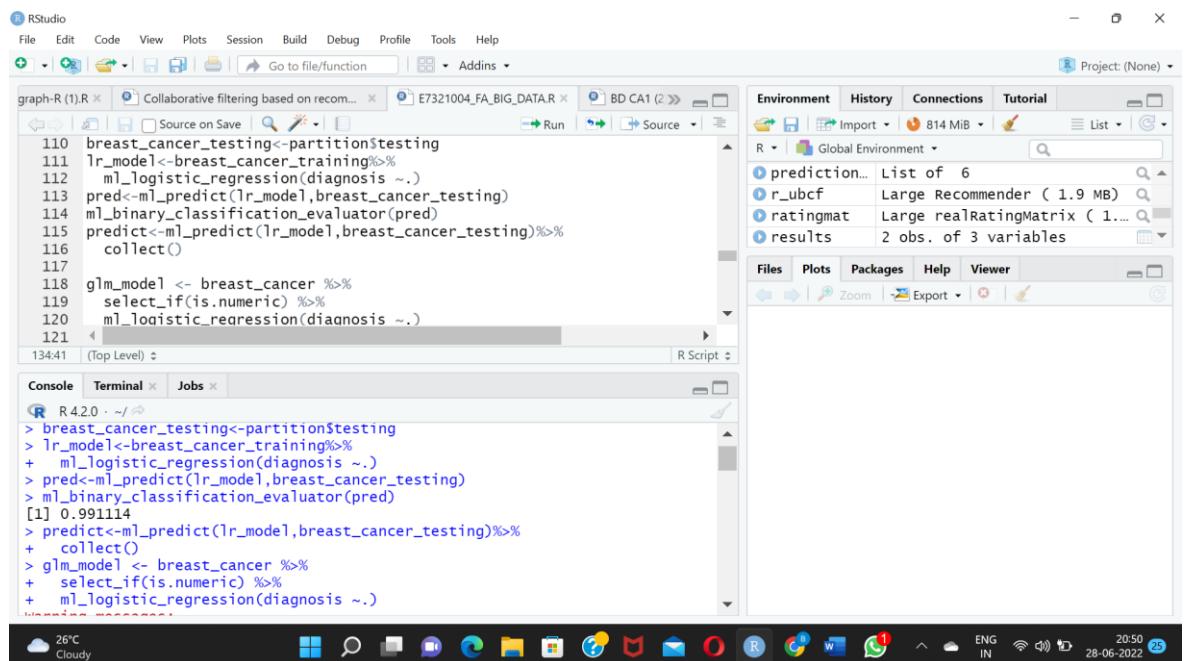
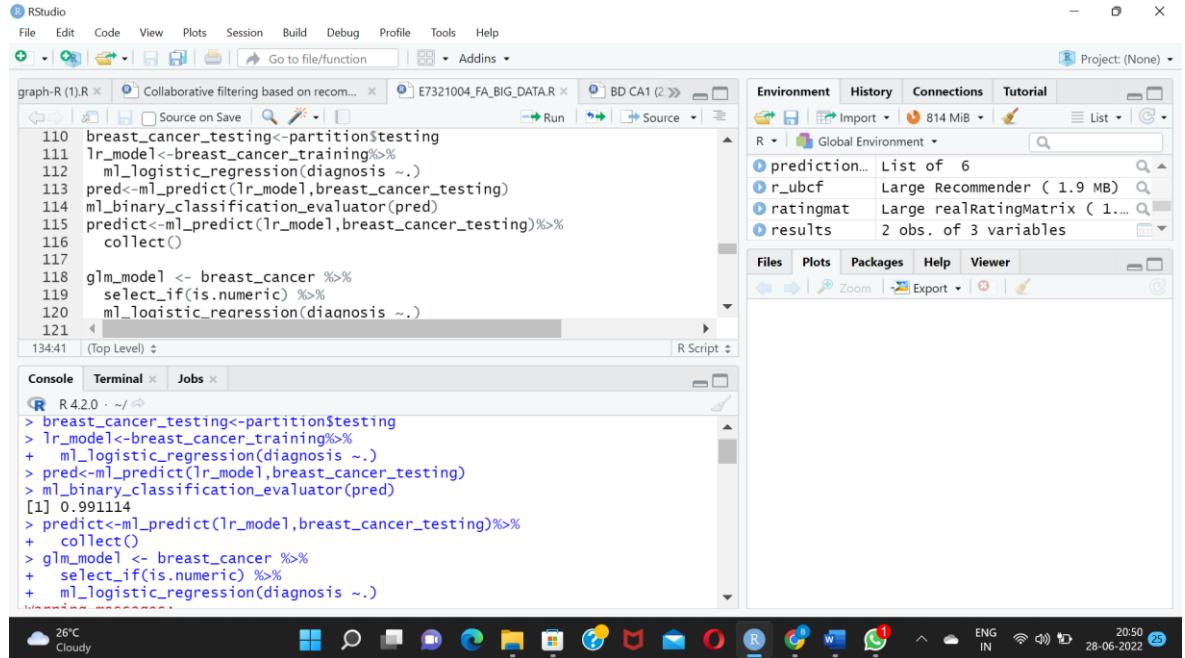
library(sparklyr)
sc<-spark_connect(master="local")
#Logistic Regression Classification
breast_cancer <- copy_to(sc, breast_cancer, "breast_cancer", overwrite = TRUE)
partition<- breast_cancer%>%
  sdf_random_split(training=0.7,testing=0.3,seed=111)
breast_cancer_training<-partition$training
breast_cancer_testing<-partition$testing
lr_model<-breast_cancer_training%>%
  ml_logistic_regression(diagnosis ~.)
pred<-ml_predict(lr_model,breast_cancer_testing)
ml_binary_classification_evaluator(pred)
predict<-ml_predict(lr_model,breast_cancer_testing)%>%
  collect()
glm_model <- breast_cancer %>%

```

```

select_if(is.numeric) %>%
ml_logistic_regression(diagnosis ~.)
summary(glm_model)
pred<-ml_predict(lr_model,breast_cancer_testing)
ml_binary_classification_evaluator(pred)
ml_predict(glm_model, breast_cancer_testing) %>%
count(diagnosis, prediction)

```



RStudio interface showing a script named graph-R (1).R containing R code for logistic regression. The console output shows the summary of the glm_model, displaying coefficients for various features like radius_mean, texture_mean, etc. The operating system taskbar at the bottom indicates it's running on a Windows machine with a 26°C temperature.

```

graph-R (1).R x [Collaborative filtering based on recom... x E7321004_FA_BIG_DATA.R x BD CA1 (2) x
105 #Logistic Regression Classification
106 breast_cancer <- copy_to(sc, breast_cancer, "breast_cancer", overwrite =
107 partition<- breast_cancer%>%
108   sdf_random_split(training=0.7,testing=0.3,seed=111)
109 breast_cancer_training<-partition$training
110 breast_cancer_testing<-partition$testing
111 lr_model<-breast_cancer_training%>%
112   ml_logistic_regression(diagnosis ~.)
113
134:41 (Top Level) R Script
```

Console output:

```

R 4.2.0 : ~/~
  length(x) = 32 > 1  in coercion to logical(1)
> summary(glm_model)
Coefficients:
            (Intercept)          id          radius_mean
-1.389233e+01 6.679944e-09 -4.879600e+00
  texture_mean    perimeter_mean      area_mean
  3.055822e-01 -5.612817e-01  3.364076e-02
smoothness_mean compactness_mean concavity_mean
  1.804683e+02 -8.565527e+01  8.829201e+01
concave_points_mean symmetry_mean fractal_dimension_mean
  2.641213e+02 -9.977615e+01 -3.656005e+02
  radius_se        texture_se      perimeter_se
  6.052642e+00 -3.415688e+00 -3.058667e+00
  area_se         smoothness_se compactness_se
```

RStudio interface showing a script named graph-R (1).R containing R code for a random forest model. The console output shows predictions being made on the testing data, followed by a confusion matrix for the classification results. The operating system taskbar at the bottom indicates it's running on a Windows machine with a 26°C temperature.

```

graph-R (1).R x [Collaborative filtering based on recom... x E7321004_FA_BIG_DATA.R x BD CA1 (2) x
105 #Logistic Regression Classification
106 breast_cancer <- copy_to(sc, breast_cancer, "breast_cancer", overwrite =
107 partition<- breast_cancer%>%
108   sdf_random_split(training=0.7,testing=0.3,seed=111)
109 breast_cancer_training<-partition$training
110 breast_cancer_testing<-partition$testing
111 lr_model<-breast_cancer_training%>%
112   ml_logistic_regression(diagnosis ~.)
113
134:41 (Top Level) R Script
```

Console output:

```

R 4.2.0 : ~/~
> pred<-ml_predict(lr_model,breast_cancer_testing)
> ml_binary_classification_evaluator(pred)
[1] 0.991114
> ml_predict(glm_model, breast_cancer_testing) %>%
+   count(diagnosis, prediction)
# Source: sparklyr [?? x 3]
# Groups: diagnosis
  diagnosis prediction n
  <dbl> <dbl> <dbl>
1     1         0     2
2     1         1     57
3     0         0    102
4     0         1     1
```

RANDOM FOREST CLASSIFICATION

#Random Forest Classifier using sparklyr on iris dataset

```
breast_cancer<-sdf_copy_to(sc,breast_cancer,name="breast_cancer",overwirte=TRUE)
```

```
partition<- breast_cancer%>%
```

```
sdf_random_split(training=0.7,testing=0.3,seed=111)
```

```
breast_cancer_training<-partition$training
```

```
breast_cancer_testing<-partition$testing
```

```
rf_model<-breast_cancer_training%>%
```

```

ml_random_forest(diagnosis~,type="classification")
prediction_score<-ml_predict(rf_model,breast_cancer_testing)
ml_binary_classification_evaluator(prediction_score)
pred<-ml_predict(rf_model,breast_cancer_testing)%>%
  collect()
table(pred$diagnosis,pred$prediction)

```

The screenshot shows the RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Project:** Project: (None)
- Environment Tab:** Shows objects: pred (162 obs. of 39 variables), predict_ib... (Large realRatingMatrix), predict_ub... (Large realRatingMatrix), and predict (162 obs. of 39 variables).
- Console Tab:** Displays the R script and its output. The output shows the execution of the provided R code, including the creation of a random forest model, prediction scores, and a confusion matrix.
- Output:**

```

R 4.2.0 : ~/ ~/
> rf_model<-breast_cancer_training%>%
+ ml_random_forest(diagnosis~,type="classification")
> prediction_score<-ml_predict(rf_model,breast_cancer_testing)
> ml_binary_classification_evaluator(prediction_score)
[1] 0.9947342
> pred<-ml_predict(rf_model,breast_cancer_testing)%>%
+ collect()
> table(pred$diagnosis,pred$prediction)

      0   1
 0 101  2
 1   5 54
>

```
- System Status Bar:** Shows the date (28-06-2022), time (21:22), and battery level (25%).

DECISION TREE CODE:

```
#Decision Tree Classification
```

```
breast_cancer<-sdf_copy_to(sc,breast_cancer,name="breast_cancer",overwrite = TRUE)
partitions<-breast_cancer%>%
```

```
sdf_random_split(training=0.7,test=0.3,seed=30)
```

```
breast_cancer_training<-partitions$training
```

```
breast_cancer_testing<-partitions$test
```

```
dt_model<-breast_cancer_training%>%
```

```
  ml_decision_tree(diagnosis ~., type="classification")
```

```
pred<-ml_predict(dt_model,breast_cancer_testing)
```

```
ml_binary_classification_evaluator(pred)
```

```
pred<-ml_predict(dt_model,breast_cancer_testing)%>%
```

```
  collect()
```

```

library("dplyr")
glimpse(pred)
pred$diagnosis
pred$prediction
table(pred$diagnosis,pred$prediction)

```

The screenshot shows the RStudio interface with the following details:

- File Bar:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Project Bar:** Project: (None).
- Code Editor:** Shows the R script with code lines 133 to 142. Line 142 shows the result of the `table` command.
- Environment View:** Shows objects `pred` (162 obs. of 39 variables), `predcit_ib...` (Large realRatingMatrix), `predcit_ub...` (Large realRatingMatrix), and `predict` (162 obs. of 39 variables).
- Console View:** Shows the R session output from line 142, which prints a 2x2 matrix:

0	1
0	101 2
1	5 54
- System Tray:** Shows the date and time (28-06-2022 21:22), battery level (25%), and system icons.

INTERPRETATION:

Among the three classification (Logistic Regression, Random forest and Decision Tree) . The Prediction Evaluation is Higher in **Decision Tree** classification(0.997) than the other classifications.

H. Remove the data label column in your dataset and apply clustering technique and interpret your findings

CODE:

```

kmeans_model<-breast_cancer %>%
  ml_kmeans(k=3,features=c("radius_worst","radius_se"))
kmeans_model
kmeans_model$feature_names
attributes(kmeans_model)
kmeans_model$centers
predicted<-ml_predict(kmeans_model,breast_cancer)%>%
  collect()

```

```

library(dplyr)
glimpse(predicted)
table(predicted$diagnosis,predicted$prediction)

```

The screenshot shows the RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Project:** Project: (None)
- Environment Tab:** Shows the `breast_cancer` dataset.
- Console Tab:**

```

R 4.2.0 -/-
+ ml_kmeans(k=3,features=c("radius_worst","radius_se"))
> kmeans_model
K-means clustering with 3 clusters

Cluster centers:
  radius_worst radius_se
1    25.13429 0.8209560
2    12.88396 0.2828911
3    17.80164 0.4078242

within set sum of squared Errors =  2163.937
> kmeans_model$feature_names
[1] "radius_worst" "radius_se"
> attributes(kmeans_model)

```
- System Status Bar:** 25°C Cloudy, 22:05, IN, 28-06-2022.

The screenshot shows the RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Project:** Project: (None)
- Environment Tab:** Shows the `breast_cancer` dataset.
- Console Tab:**

```

R 4.2.0 -/-
3 17.80164 0.4078242

within set sum of squared Errors =  2163.937
> kmeans_model$feature_names
[1] "radius_worst" "radius_se"
> attributes(kmeans_model)
$names
[1] "pipeline_model" "formula"      "dataset"       "pipeline"
[5] "model"          "features_col"   "feature_names" "summary"
[9] "centers"        "cost"

$class
[1] "ml_model_kmeans"     "ml_model_clustering" "ml_model"

```
- System Status Bar:** 25°C Cloudy, 22:05, IN, 28-06-2022.

The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The top toolbar has icons for file operations like Open, Save, and Run, along with a search bar labeled "Go to file/function". The left sidebar lists projects: "movieLens", "Jester5k", "LR model.BGR.R", and "E7321004_FA_BIG_DATA.R". The main workspace contains an R script:

```
168 attributes(kmeans_model)
169 kmeans_model$centers
170 predicted<-ml_predict(kmeans_model,breast_cancer)%>%
171   collect()
172 library(dplyr)
173 glimpse(predicted)
174 table(predicted$diagnosis,predicted$prediction)
175
```

The status bar at the bottom shows the current session ID as "R42202" and the date/time as "28-06-2022 22:06".

The screenshot shows an RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The toolbar contains icons for file operations like Open, Save, and Print, along with Go to file/function and Addins. The left pane displays a script editor with the following R code:

```
168 attributes(kmeans_model)
169 kmeans_model$centers
170 predicted<-mlr_predict(kmeans_model,breast_cancer)%>%
171   collect()
172 library(dplyr)
173 glimpse(predicted)
174 table(predicted$diagnosis,predicted$prediction)
175
```

The status bar at the bottom indicates Rows: 569 and Columns: 34. The right pane shows the Environment tab with a local variable named `breast_cancer`. The bottom right corner shows the system tray with a weather icon (25°C Cloudy), date (28-06-2022), and time (22:07).

The screenshot shows the RStudio interface with several tabs open. In the top-left, there are tabs for 'movielens', 'Jester5k', 'LR model BG.R', 'E7321004_FA_BIG_DATA.R', and 'breast_cancer'. The 'breast_cancer' tab is active. The top menu bar includes 'File', 'Edit', 'Code', 'View', 'Plots', 'Session', 'Build', 'Debug', 'Profile', 'Tools', and 'Help'. Below the menu is a toolbar with icons for file operations like Open, Save, and Run. The main workspace shows R script code and its output. The code includes:

```

168 attributes(kmeans_model)
169 kmeans_model$centers
170 predicted<-ml_predict(kmeans_model,breast_cancer)%>%
171   collect()
172 library(dplyr)
173 glimpse(predicted)
174 table(predicted$diagnosis,predicted$prediction)
175

```

The output in the console shows the centers of the clusters and a confusion matrix:

```

$ compactness_worst <dbl> 0.6656, 0.1866, 0.4245, 0.8663, 0.2050, 0...
$ concavity_worst <dbl> 0.71190, 0.24160, 0.45040, 0.68690, 0.400...
$ concave_points_worst <dbl> 0.26540, 0.18600, 0.24300, 0.25750, 0.162...
$ symmetry_worst <dbl> 0.4601, 0.2750, 0.3613, 0.6638, 0.2364, 0...
$ fractal_dimension_worst <dbl> 0.11890, 0.08902, 0.08758, 0.17300, 0.076...
$ features <list> <25.380, 1.095>, <24.9900, 0.5435>, <23.5400, 0.2050>, ...
$ prediction <int> 0, 0, 0, 1, 0, 2, 0, 2, 2, 1, 2, 2, 2, 2, ...
> table(predicted$diagnosis,predicted$prediction)

      0   1   2
0    0 300  57
1   93  11 108

```

The bottom status bar shows system information: 25°C Cloudy, 22:07, 28-06-2022.

INTERPRETATION:

Using clustering Technique, we have formulated 3 clusters for the features radius_worst and radius_se ,using clustering model we have formulated the centers of each clusters. And using that we have predicted the other features efficiently.

Perform comparative analysis of the classification model and clustering technique employed on the dataset after removing label column. Interpret the same and give your view about the outcome.

INTERPRETATION:

*Using classification technique we can classify the dataset based on the class label i.e. wheather the person is affected by cancer or not ,we can also predict the unkown values using the data.

*In Clustering technique ,after removing the class label we can only form the clusters of the features, we cannot predict anything using cluterling like(The person is affected by cancer or not)