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 **Signature of Vice-Principal**

Name of Course Faculty

Assistant Professor,

Department of Computer Science and Engineering

Sri Ramachandra Faculty of Engineering and Technology,

SRIHER, Porur, Chennai-600 116.

Prof. M. Prema

Vice-Principal,

Department of Computer Science and Engineering

Sri Ramachandra Faculty of Engineering and Technology,

SRIHER, Porur, Chennai-600 116.

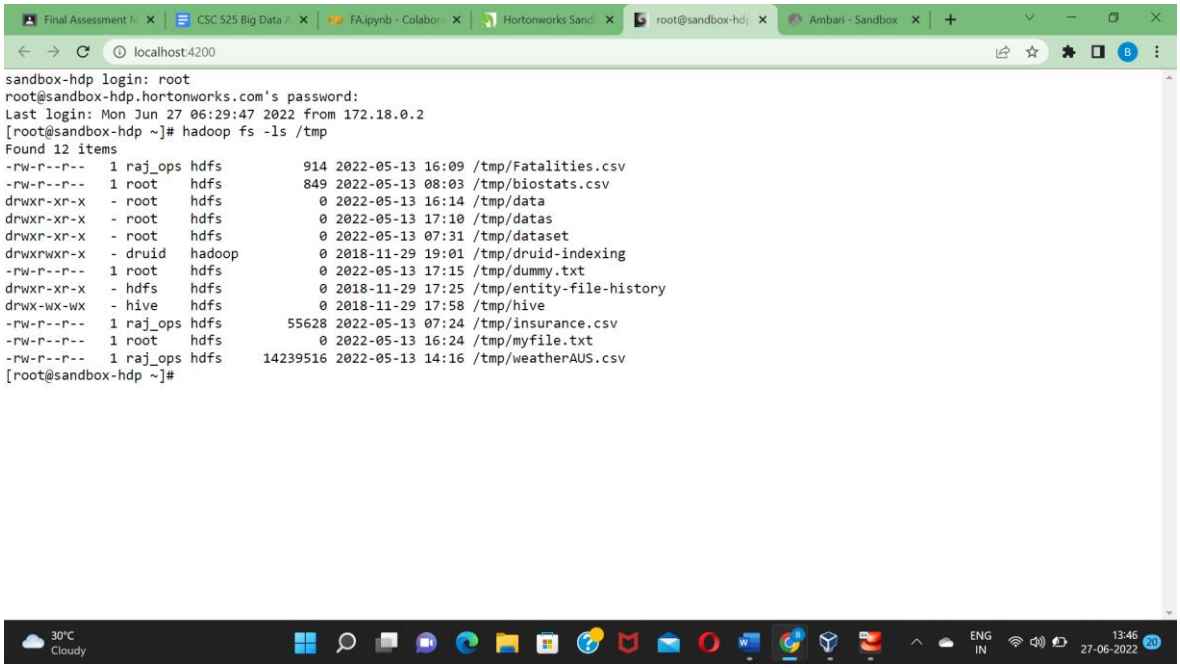
Evaluation Date:

TABLE OF CONTENTS

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

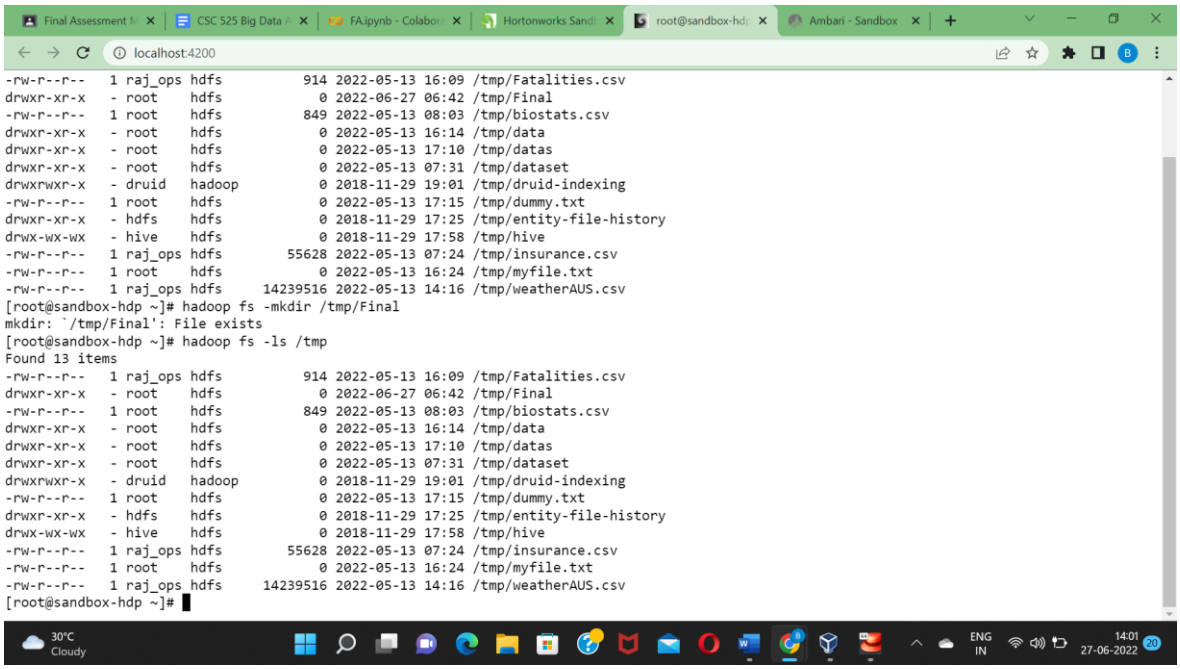
A. Perform the following Hadoop commands to check the feasibility of the Hadoop environment configured.

• List all the files in the data directory.



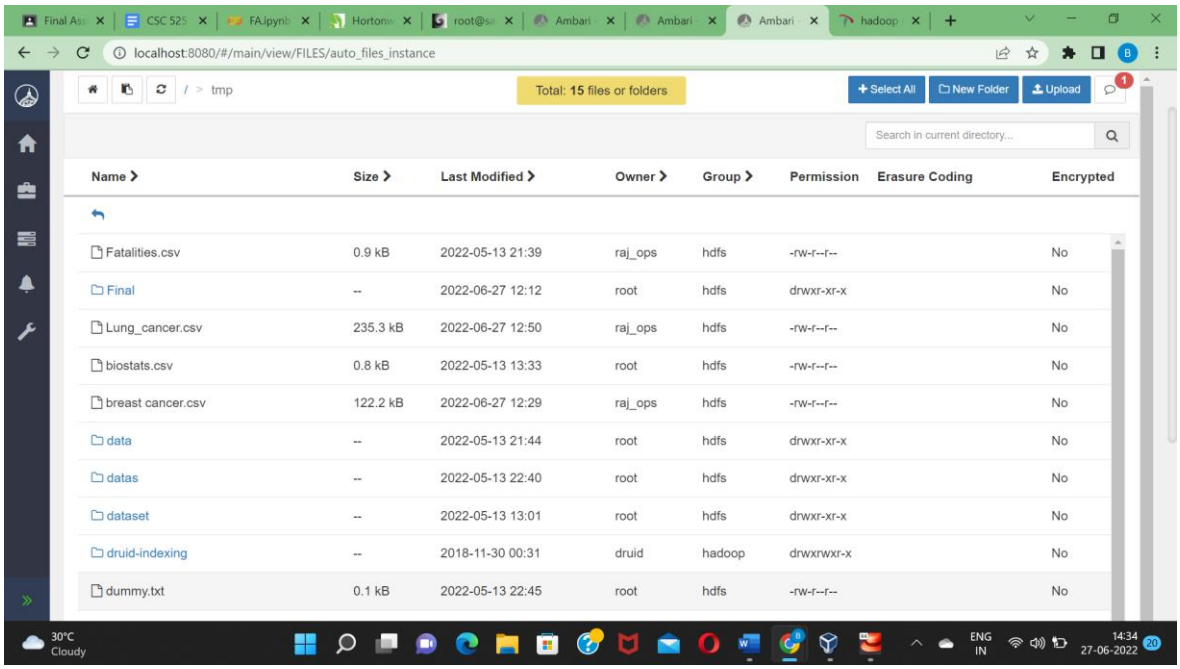
```
sandbox-hdp login: root
root@sandbox-hdp.hortonworks.com's password:
Last login: Mon Jun 27 06:29:47 2022 from 172.18.0.2
[root@sandbox-hdp ~]# hadoop fs -ls /tmp
Found 12 items
-rw-r--r--   1 raj_ops hdfs          914 2022-05-13 16:09 /tmp/Fatalities.csv
-rw-r--r--   1 root    hdfs          849 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
drwx-wx-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

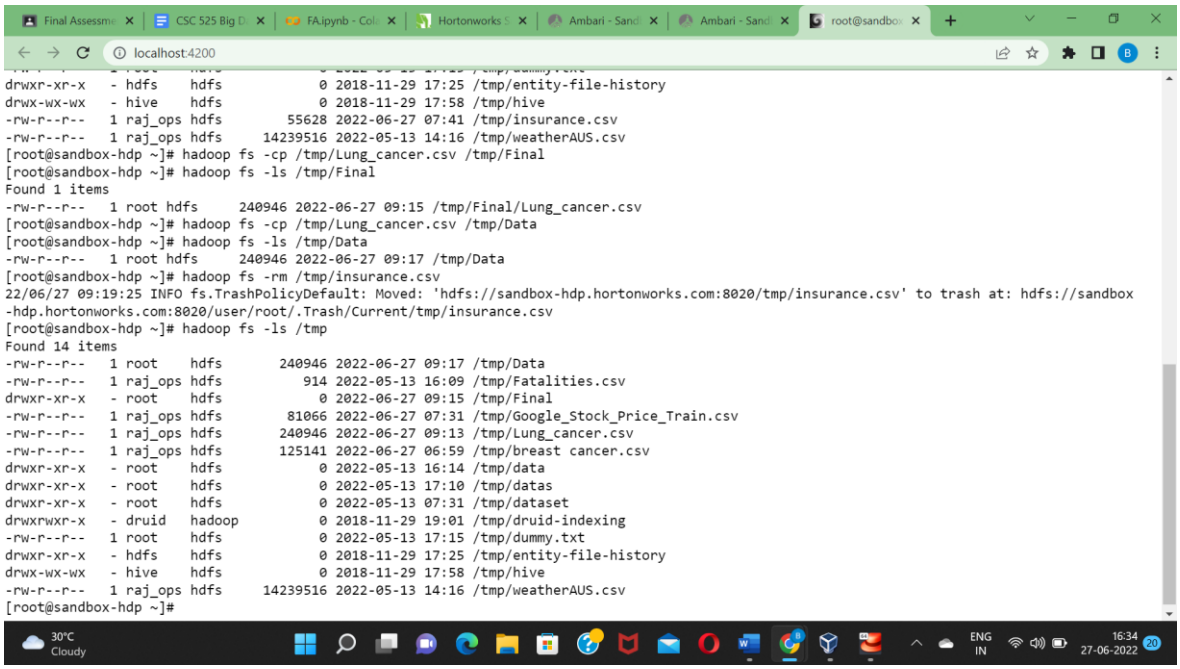


```
-rw-r--r--   1 raj_ops hdfs          914 2022-05-13 16:09 /tmp/Fatalities.csv
drwxr-xr-x   - root    hdfs            0 2022-06-27 06:42 /tmp/Final
-rw-r--r--   1 root    hdfs          849 2022-05-13 08:03 /tmp/biostats.csv
drwxr-xr-x   - root    hdfs            0 2022-05-13 16:14 /tmp/data
drwxr-xr-x   - 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
drwx-wx-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 ~]# 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
drwx-wx-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 ~]#
```

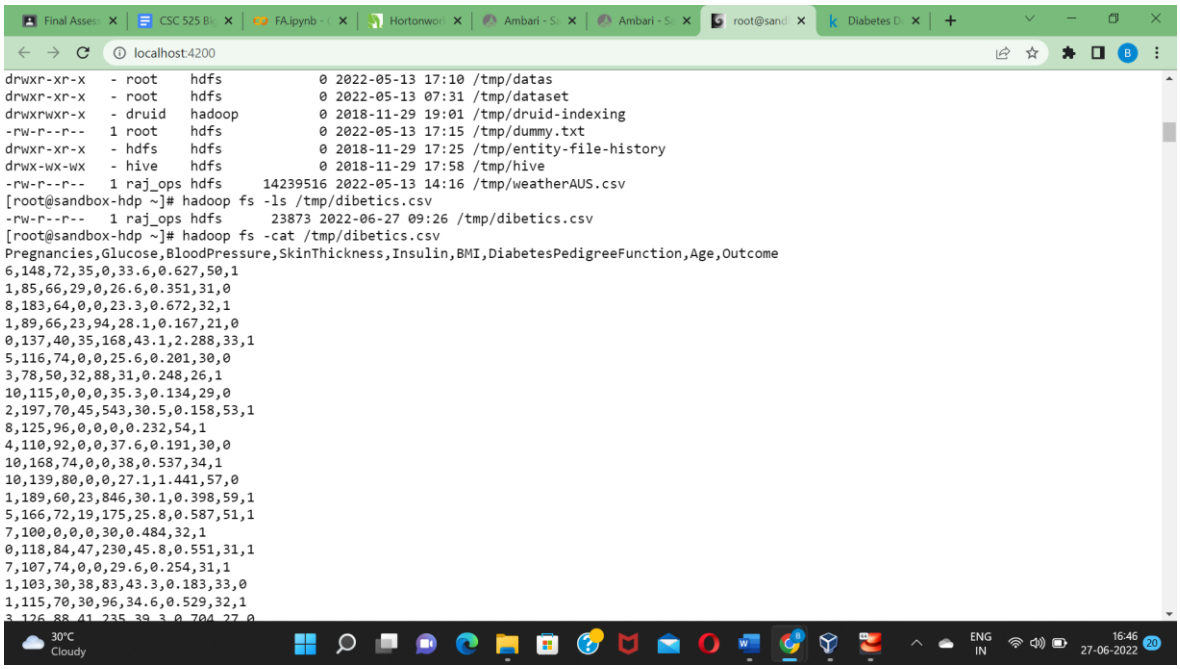
• Load csv file from local to data directory.

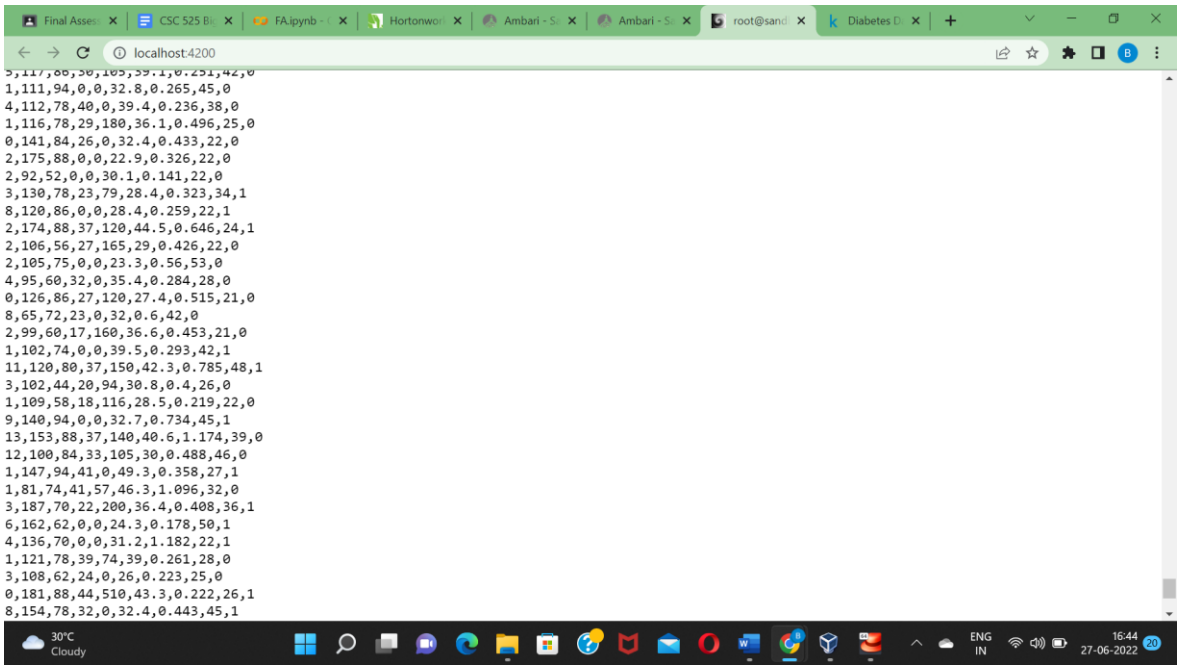


• Remove a particular file from tmp directory.

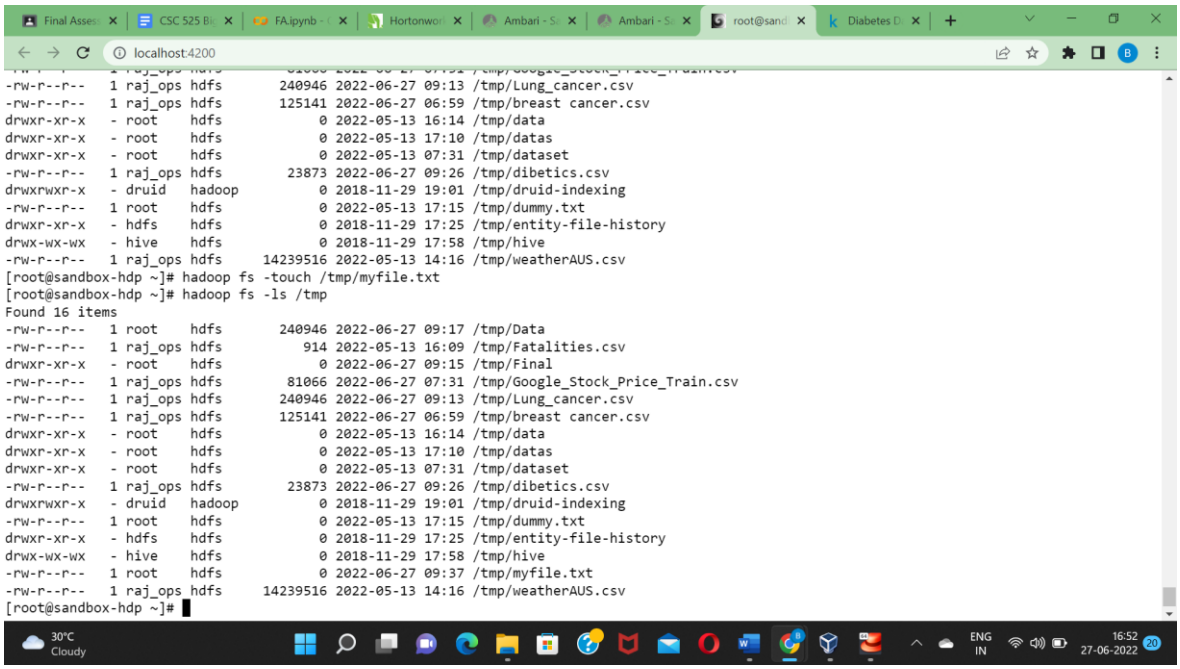


• Display the content present in the dibetics.csv file.

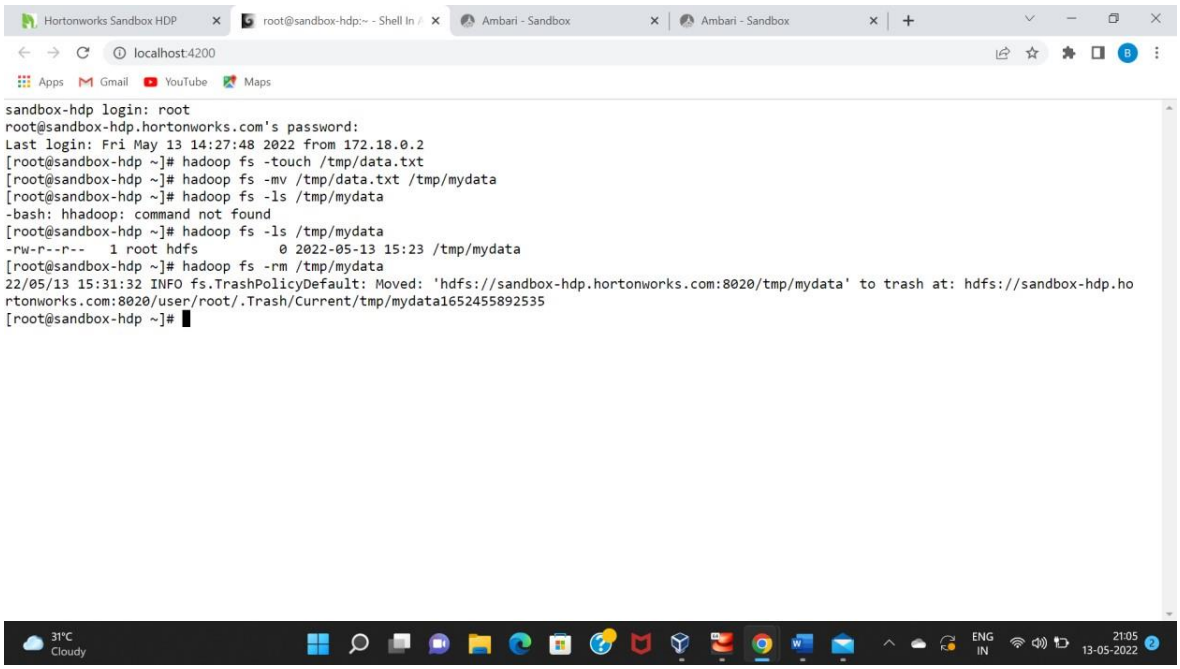




• Create an empty file called myfile.txt



• Load txt file from local to tmp directory.



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

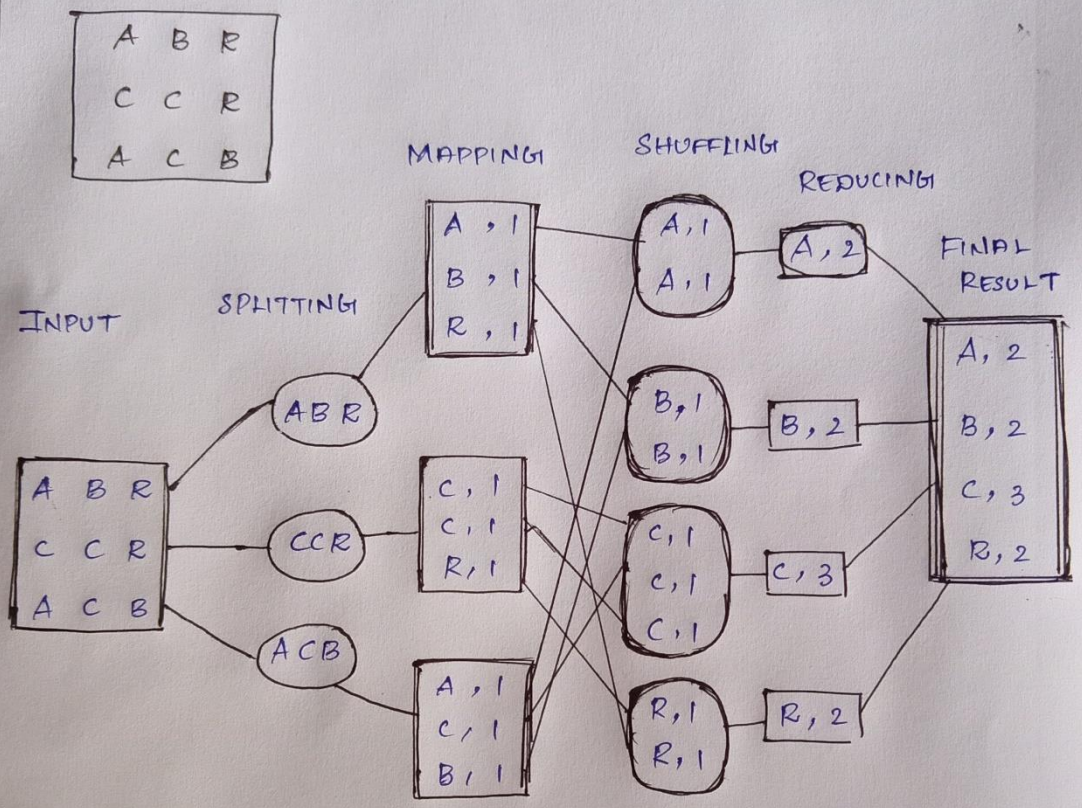
while visiting the Amazon website or through email. This is an example of a data-driven business decision.

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

CSC 525 - BIG DATA ANALYTICS

UNIQUE ID: ET3210004

1) c) Map Reduce

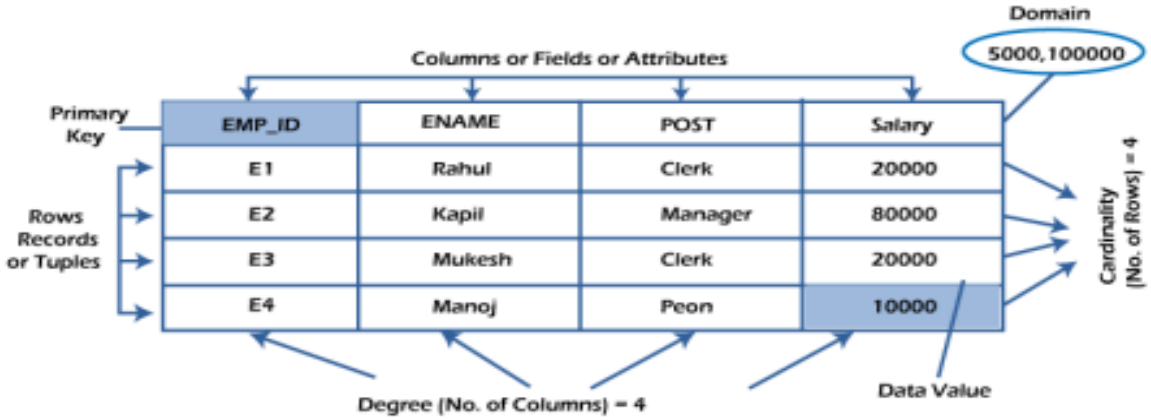


The Five common Steps Involved in Map Reduce are,

* 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.

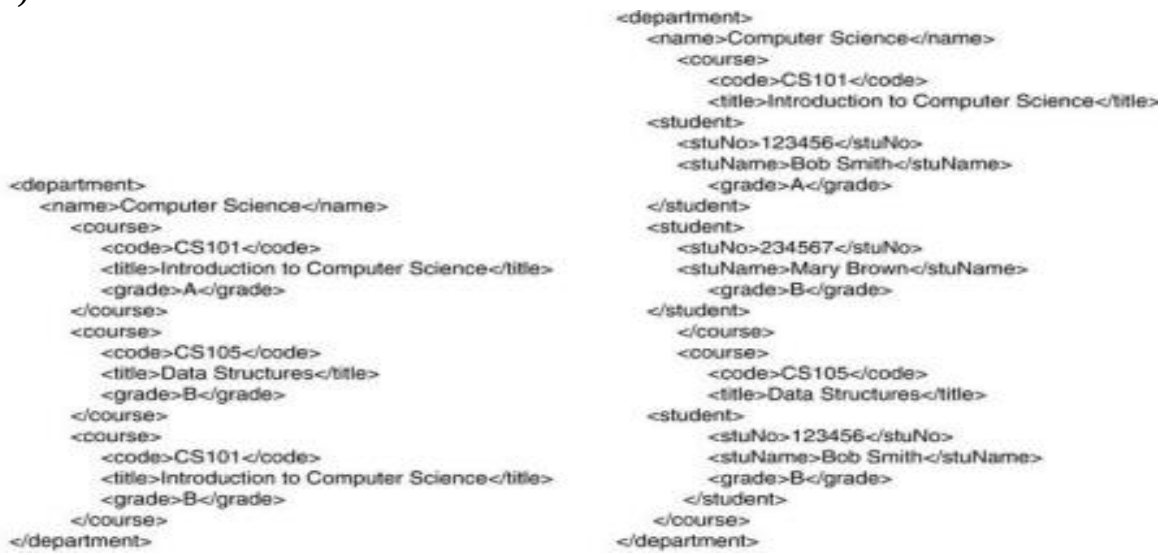
Examples:

*Text documents

*PDF document

*Images and Vedio

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>
```

```
<department>
  <name>Computer Science</name>
    <course>
      <code>CS101</code>
      <title>Introduction to Computer Science</title>
  <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>
    <course>
      <code>CS105</code>
      <title>Data Structures</title>
  <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 tat 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) $h(x) = 6x + 2 \mod 5$

$h(1) = 6(1) + 2 \mod 5$

$= 8 \mod 5$

$= \boxed{3}$

$h(3) = 6(3) + 2 \mod 5$

$= 20 \mod 5$

$= \boxed{0}$

$h(2) = 6(2) + 2 \mod 5$

$= 14 \mod 5$

$= \boxed{4}$

$h(4) = 6(4) + 2 \mod 5$

$26 \mod 5$

$= \boxed{1}$

Binary value of 3 is,

$\boxed{011}$ → 3 Bits

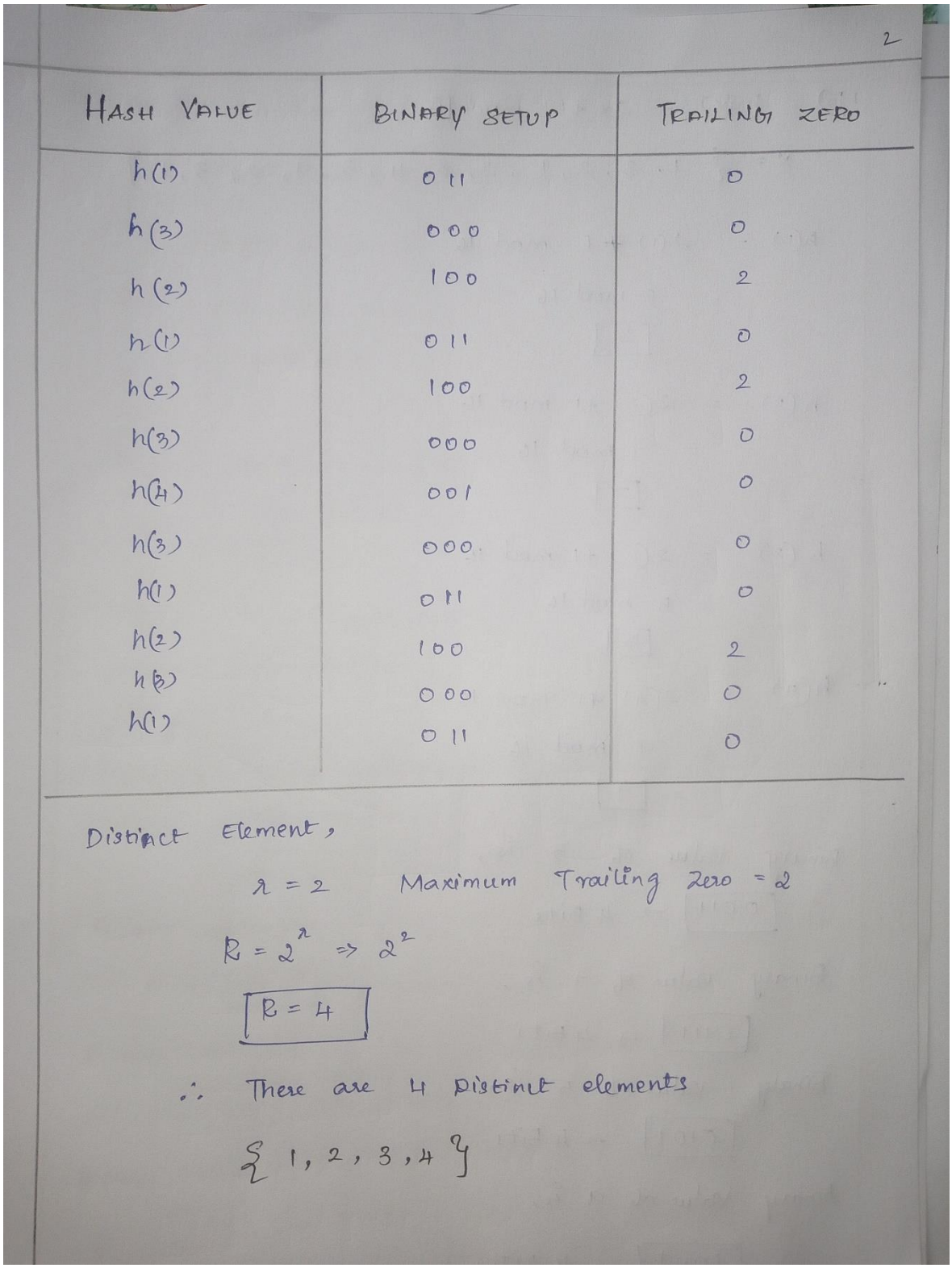
Binary value of 0 is

$\boxed{000}$ → 3 Bits

Binary value of 4 is

$\boxed{100}$ → 3 Bits

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



2

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,

$r = 2$ Maximum Trailing Zero = 2

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

$\boxed{R = 4}$

$\therefore$ There are 4 distinct elements

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

B. h(x) = 2x + 1 mod 16

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

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

$h(1) = 2(1) + 1 \mod 16$

$= 3 \mod 16$

$= \boxed{3}$

$h(3) = 2(3) + 1 \mod 16$

$= 7 \mod 16$

$\boxed{7}$

$h(2) = 2(2) + 1 \mod 16$

$= 5 \mod 16$

$\boxed{5}$

$h(4) = 2(4) + 1 \mod 16$

$= 9 \mod 16$

$= \boxed{9}$

Binary value of 3 is,

$\boxed{0011}$ → 4 Bits

Binary value of 7 is,

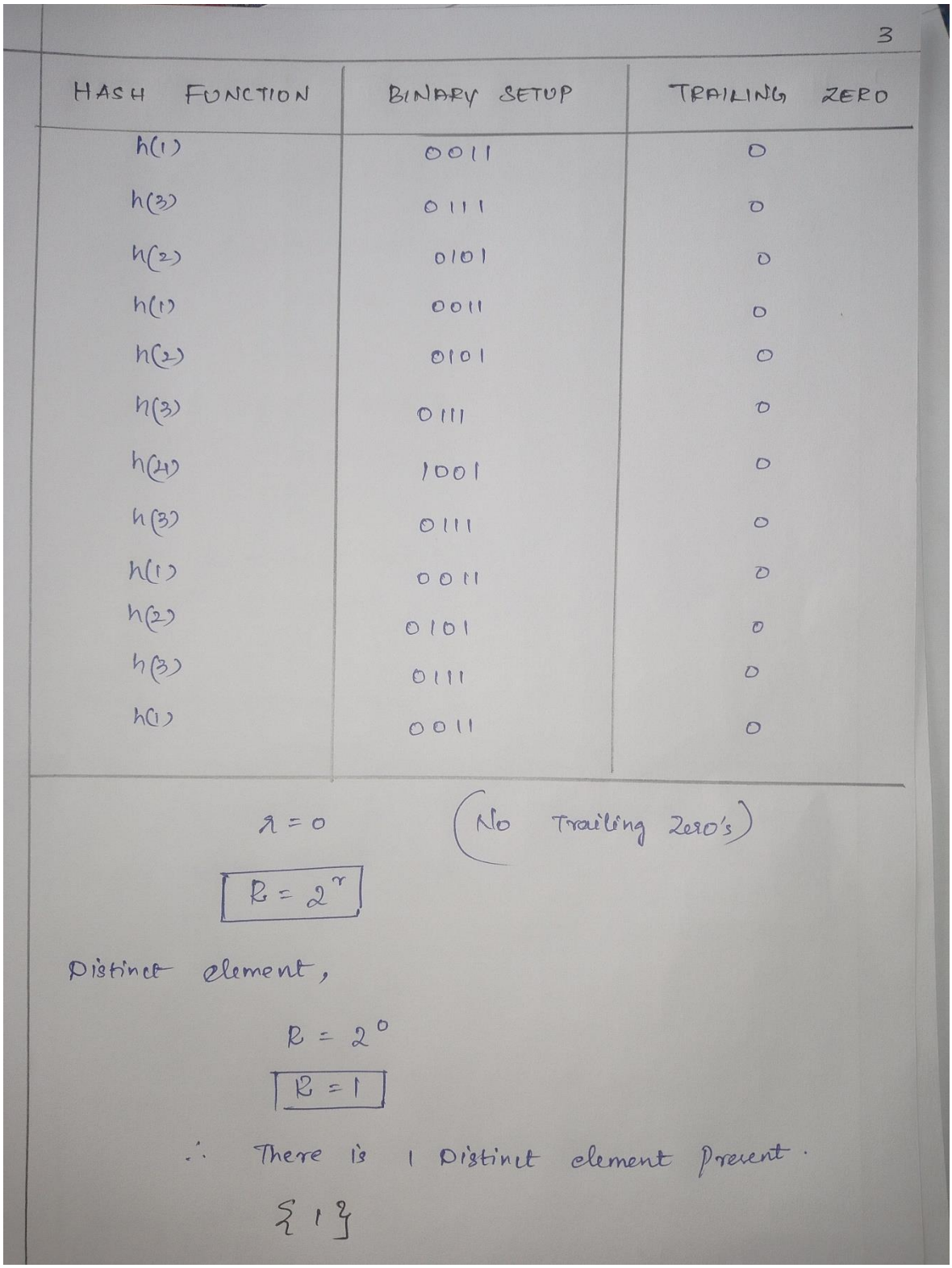
$\boxed{0111}$ → 4 Bits

Binary value of 5 is,

$\boxed{0101}$ → 4 Bits

Binary value of 9 is,

$\boxed{1001}$ → 4 Bits.



3

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

$r = 0$ (No Trailing Zero's)

$R = 2^r$

Distinct element,

$R = 2^0$

$R = 1$

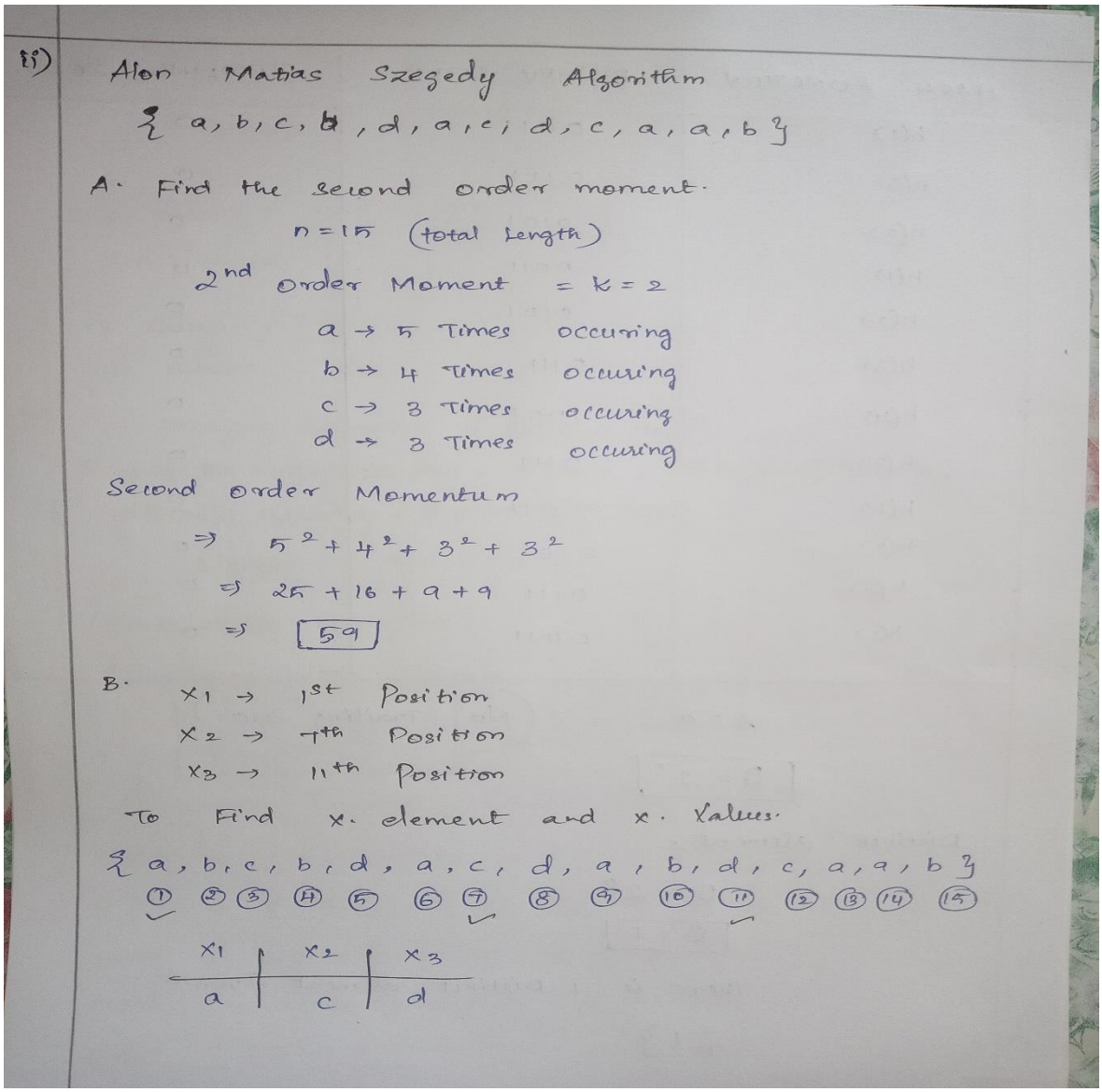
∴ There is 1 Distinct element present.

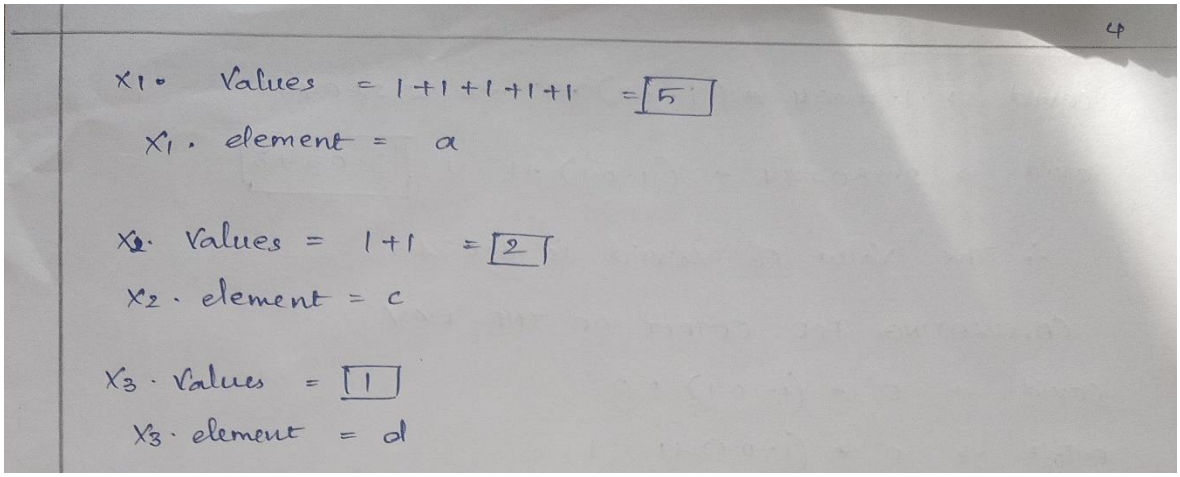
{1}

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.



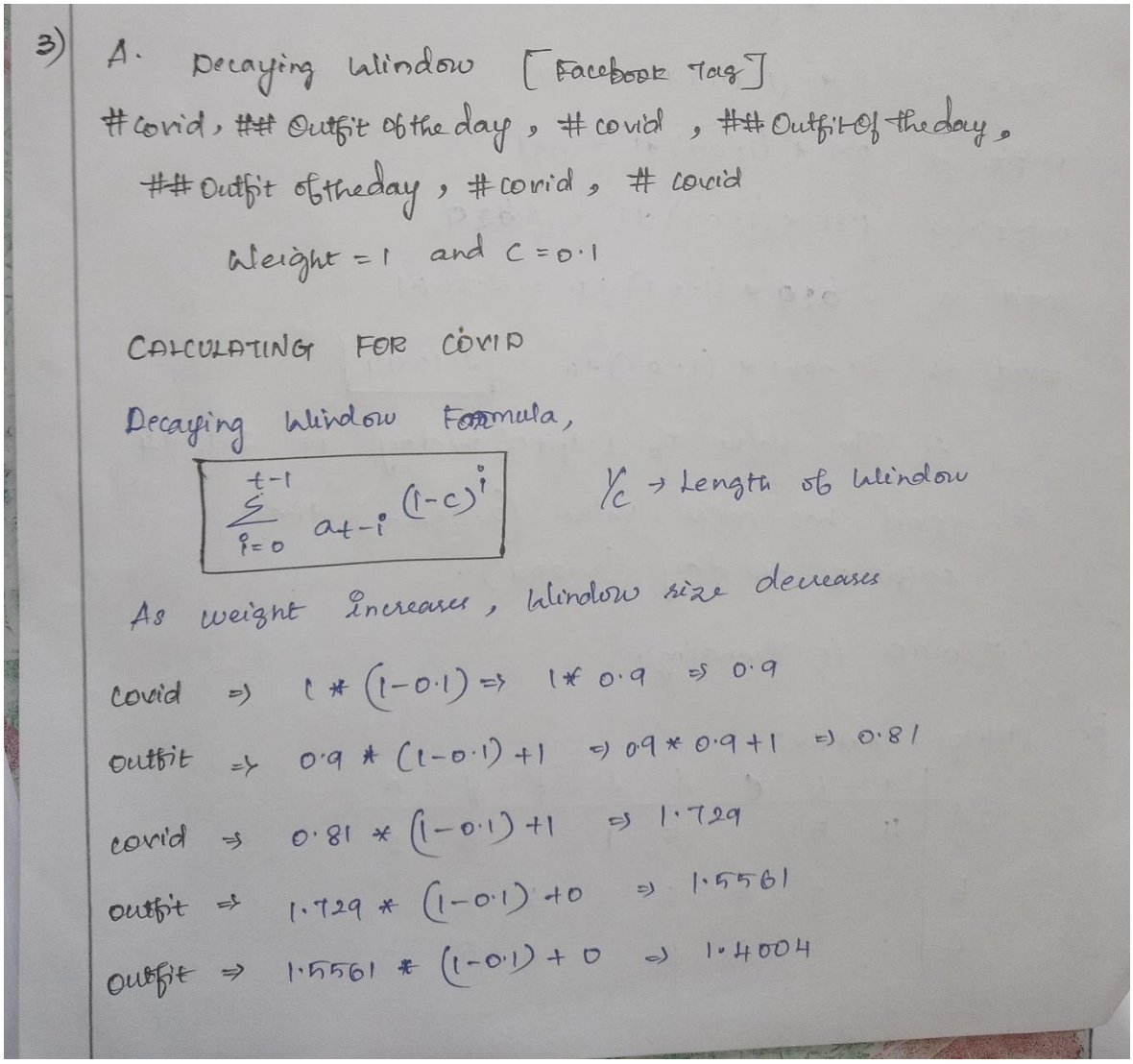


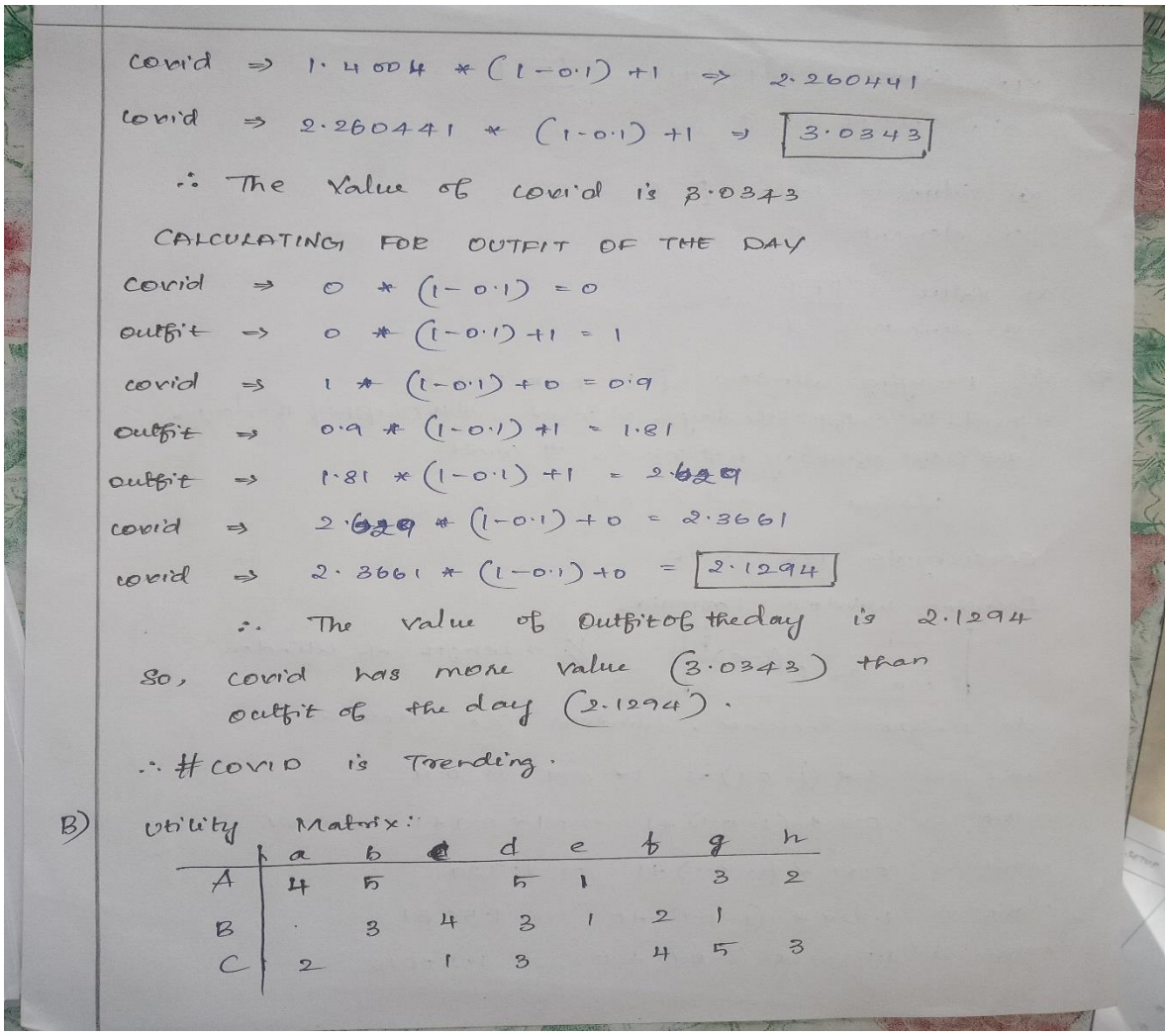
3. A. Apply decaying window algorithm to analyse the trends in facebook data for the following facebook tags

#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.





covid ⇒ 1.4004 * (1−0.1) +1 ⇒ 2.260441

covid ⇒ 2.260441 * (1−0.1) +1 ⇒ 3.0343

∴ The value of covid is 3.0343

CALCULATING FOR OUTFIT OF THE DAY

covid ⇒ 0 * (1−0.1) = 0

outfit ⇒ 0 * (1−0.1) +1 = 1

covid ⇒ 1 * (1−0.1) + 0 = 0.9

outfit ⇒ 0.9 * (1−0.1) +1 = 1.81

outfit ⇒ 1.81 * (1−0.1) +1 = 2.629

covid ⇒ 2.629 * (1−0.1) +0 = 2.3661

covid ⇒ 2.3661 * (1−0.1) +0 = 2.1294

∴ The value of Outfit of the day is 2.1294

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

∴ #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

i) Treat the utility matrix as Boolean and compute the Jaccard distance between each pair of users.

ii) Find the cosine distance between each pair of users

iii) 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

iv) 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

Jaccard distance $(A, B) = \mathfrak{R}_A \cap \mathfrak{R}_B \,/\, \mathfrak{R}_A \cup \mathfrak{R}_B$

$(A, B) = 4/8 = 1/2 = \boxed{0.5}$

Jaccard distance $(A, C) = \mathfrak{R}_A \cap \mathfrak{R}_C \,/\, \mathfrak{R}_A \cup \mathfrak{R}_C$

$(A, C) = 4/8 = 1/2 = \boxed{0.5}$

Jaccard distance $(B, C) = \mathfrak{R}_B \cap \mathfrak{R}_C \,/\, \mathfrak{R}_B \cup \mathfrak{R}_C$

$(B, C) = 4/8 = 1/2 = \boxed{0.5}$

∴ 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.

$$\text{Cosine}(A,B) = \cos(\angle A, \angle B)$$

$$= \frac{(5\times3)+(5\times3)+(1\times1)+(3\times1)}{\sqrt{4^2+5^2+5^2+1^2+3^2+2^2}\sqrt{3^2+4^2+3^2+1^2+2^2+1^2}}$$

$$\Rightarrow \frac{34}{8.944 \times 6.32455} \Rightarrow \frac{34}{56.5668}$$

$$\Rightarrow \boxed{0.6010}$$

Cosine Distance Between (A, C).

$$\text{Cosine}(A,C) = \cos(\angle A, \angle C)$$

$$= \frac{(4\times2)+(5\times3)+(3\times5)+(2\times3)}{\sqrt{4^2+5^2+5^2+1^2+3^2+1^2}\times\sqrt{2^2+1^2+3^2+4^2+5^2+3^2}}$$

$$\Rightarrow \frac{44}{8.944 \times 8} = \frac{44}{71.552}$$

$$\Rightarrow \boxed{0.61493}$$

Cosine Distance Between (B, C)

$$\text{cosine}(B,C) = \cos(\lambda B, \lambda C)$$

$$\Rightarrow \frac{(4\times1)+(3\times3)+(2\times4)+(1\times5)}{\sqrt{3^2+4^2+3^2+1^2+2^2+1^2}\times\sqrt{2^2+1^2+3^2+4^2+5^2+3^2}}$$

$$= 26/51.768 = \boxed{0.5022}$$

∴ The cosine Distance (A, C) is more similar 0.61493.

(ii) 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.

Treating 4 and 5 as 1
1, 2, 3 as 0

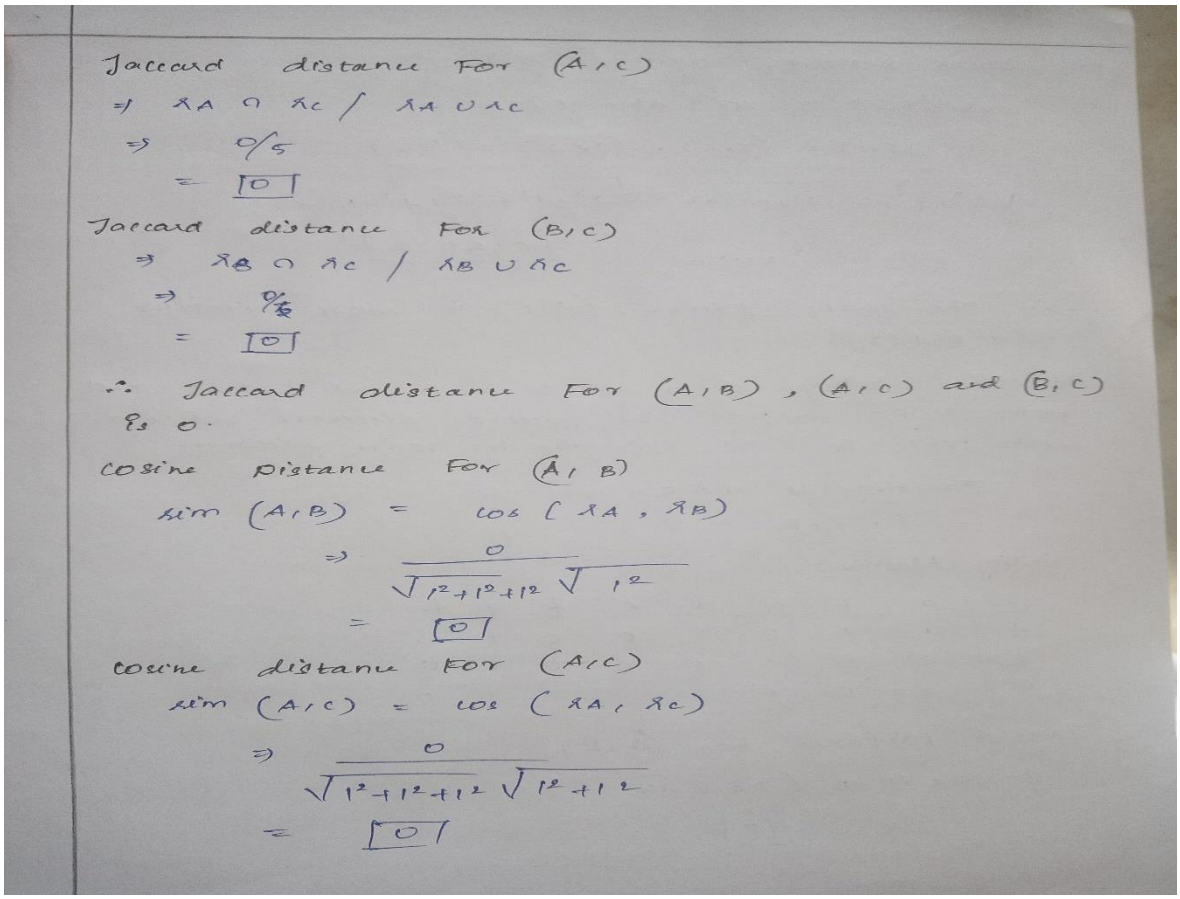
Utility Matrix:

	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)

$$\Rightarrow \lambda A \cap \lambda B / \lambda A \cup \lambda B$$

$$\Rightarrow 0/4 = \boxed{0}$$


Jaccard distance For (A, C)
⇒ A_A ∩ A_C / A_A ∪ A_C
⇒ 0/5
= 0
Jaccard distance For (B, C)
⇒ A_B ∩ A_C / A_B ∪ A_C
⇒ 0/5
= 0
∴ Jaccard distance For (A, B), (A, C) and (B, C) is 0.
Cosine Distance For (A, B)
sim (A, B) = cos (A_A, A_B)
⇒ 0 / √(1²+1²+1²) √(1²)
= 0
Cosine distance For (A, C)
sim (A, C) = cos (A_A, A_C)
⇒ 0 / √(1²+1²+1²) √(1²+1²)
= 0

Cosine Distance For (B,C)

$Sim(B,C) = \cos(\vec{B}, \vec{C})$

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

$$= \boxed{0}$$

∴ The cosine Distance For (A,B), (A,C) and (B,C) are 0.

iv) 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,

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

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

Subtracting 10/3 to calculate A,

$4 - 10/3 = \boxed{2/3}$ $\quad$ $1 - 10/3 = \boxed{-7/3}$

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

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

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

Average Rating For B,

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

$\Rightarrow 14/6$

$= \boxed{7/3}$

Subtracting 7/3 to calculate B values

$3 - 7/3 = \boxed{2/3}$ $\quad 2 - 7/3 = \boxed{-1/3}$

$4 - 7/3 = \boxed{5/3}$ $\quad 1 - 7/3 = \boxed{-4/3}$

Average Rating For C

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

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	2/3	5/3		5/3	-7/3		-1/3	-4/3
B		2/3	5/3	2/3	-4/3	-1/3	-4/3	
C	-1		-2	0		1	2	0

Cosine Distance For (A, B)

$$sim(A,B) = \text{Cosine}(\overrightarrow{AA}, \overrightarrow{AB})$$

$$\Rightarrow \frac{(5/3 \times 2/3) + (5/3 \times 2/3) + (-7/3 \times -4/3) + (-1/3 + -4/3)}{\sqrt{(2/3)^2 + (5/3)^2 + (5/3)^2 + (-7/3)^2 + (1/3)^2 + (-4/3)^2} \times \sqrt{(2/3)^2 + (5/3)^2 + (2/3)^2 + (-4/3)^2 + (-1/3)^2 + (-4/3)^2}}$$

$$\Rightarrow \frac{10/9 + 10/9 + 28/9 + 4/9}{\sqrt{4/9 + 25/9 + 25/9 + 49/9 + 1/9 + 16/9} \times \sqrt{4/9 + 25/9 + 4/9 + 16/9 + 1/9 + 16/9}}$$

$$\Rightarrow \frac{52/3}{\sqrt{120/9} \times \sqrt{66/9}}$$

$$\Rightarrow \frac{5.7777}{9.8869}$$

$$\Rightarrow \boxed{0.5843}$$

Cosine Distance For (A, C)

$$sim(A,C) = \text{Cosine}(\overrightarrow{AA}, \overrightarrow{AC})$$

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

$$\Rightarrow \frac{-2/3 + -2/3}{\sqrt{\frac{120}{9}} \times \sqrt{10}} \Rightarrow \frac{-4/3}{3.6515 \times 3.1622}$$

$$\Rightarrow \frac{-1.333}{11.5451} \Rightarrow \boxed{-0.115460}$$

Cosine Distance For (B, C)

$$sim(B,C) = \text{Cosine}(B,C)$$

$$\Rightarrow \frac{(5/3 \times -2) + (-1/3 \times 1) + (-4/3 \times 2)}{\sqrt{66/9} \times \sqrt{10}}$$

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

$$\Rightarrow \frac{-19/3}{8.5632}$$

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

$$\Rightarrow \boxed{-0.7395}$$

∴ The Cosine Distance (A, B) is more similar.

v) Interpret your results.

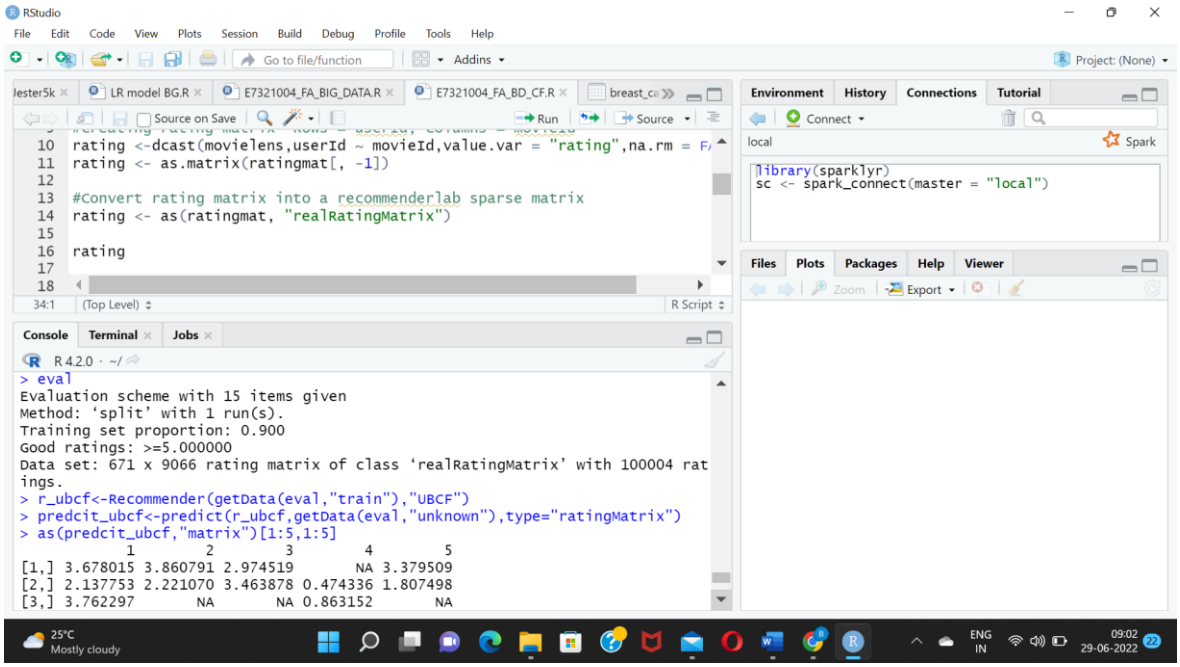
INTERPRETATION:

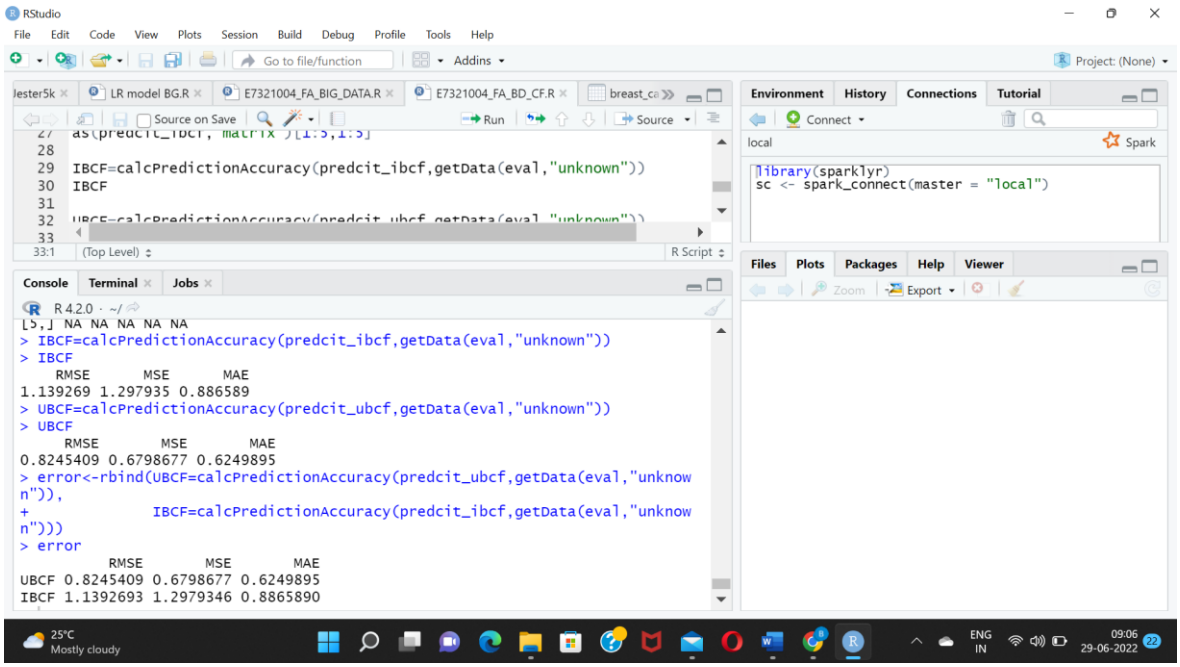
*For the above Boolean Matrix , when convert the ratings 4,5 as 1 and 1,2,3 and blank as 0 The values of jaccard distance and cosine distance became 0 There is no similarity between them.

*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.



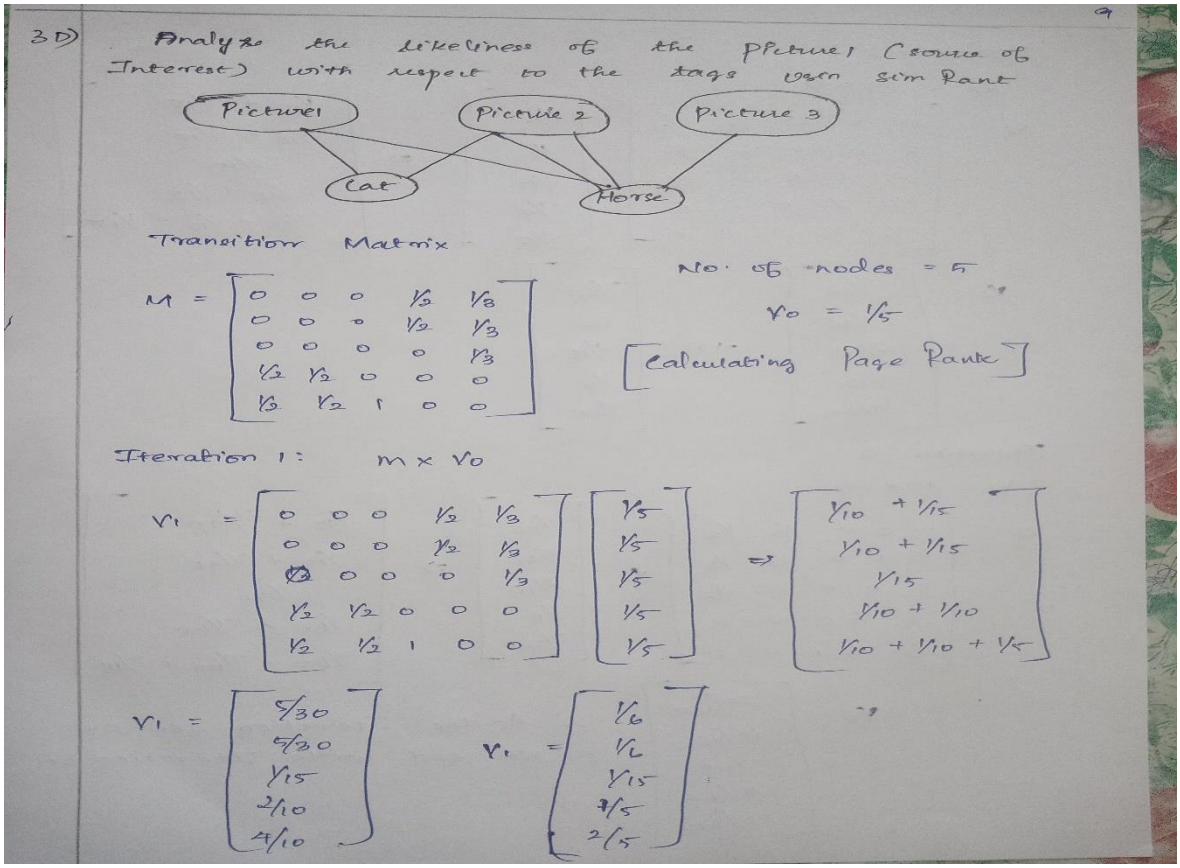


CODE:

```
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 :

$V_2 = m \times V_1$

$$V_2 = \begin{bmatrix} 0 & 0 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/2 & 1/3 \\ 0 & 0 & 0 & 0 & 1/3 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1/6 \\ 1/6 \\ 1/15 \\ 1/5 \\ 2/5 \end{bmatrix} \Rightarrow \begin{bmatrix} 1/10 + 2/15 \\ 1/10 + 2/15 \\ 2/15 \\ 1/12 + 1/12 \\ 1/12 + 1/12 + 1/15 \end{bmatrix}$$

$$V_2 = \begin{bmatrix} 7/30 \\ 7/30 \\ 2/15 \\ 2/12 \\ 14/60 \end{bmatrix} \Rightarrow V_2 = \begin{bmatrix} 7/30 \\ 7/30 \\ 2/15 \\ 1/6 \\ 7/30 \end{bmatrix}$$

Iteration 3 :

$V_3 = m \times V_2$

$$V_3 = \begin{bmatrix} 0 & 0 & 0 & 1/2 & 1/3 \\ 0 & 0 & 0 & 1/2 & 1/3 \\ 0 & 0 & 0 & 0 & 1/3 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 1 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 7/30 \\ 7/30 \\ 2/15 \\ 1/6 \\ 7/30 \end{bmatrix} \Rightarrow \begin{bmatrix} 1/12 + 7/90 \\ 1/12 + 7/90 \\ 7/90 \\ 7/60 + 7/60 \\ 7/60 + 7/60 + 2/15 \end{bmatrix}$$

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

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

Calculating Trust Rank for Sim Rank.

$$M = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ C \\ H \end{array} \begin{array}{c} \begin{array}{ccccc} P_1 & P_2 & P_3 & C & H \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 1/2 & 1/3 \\ 0 & 0 & 0 & 1/2 & 1/3 \\ 0 & 0 & 0 & 0 & 1/3 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 1 & 0 & 0 \end{bmatrix} \end{array}$$

n = no. of nodes = 5

$S = \{$ Picture 1 $\}$

$V = 1$

$$V' = \beta M v + (1-\beta) e_s$$

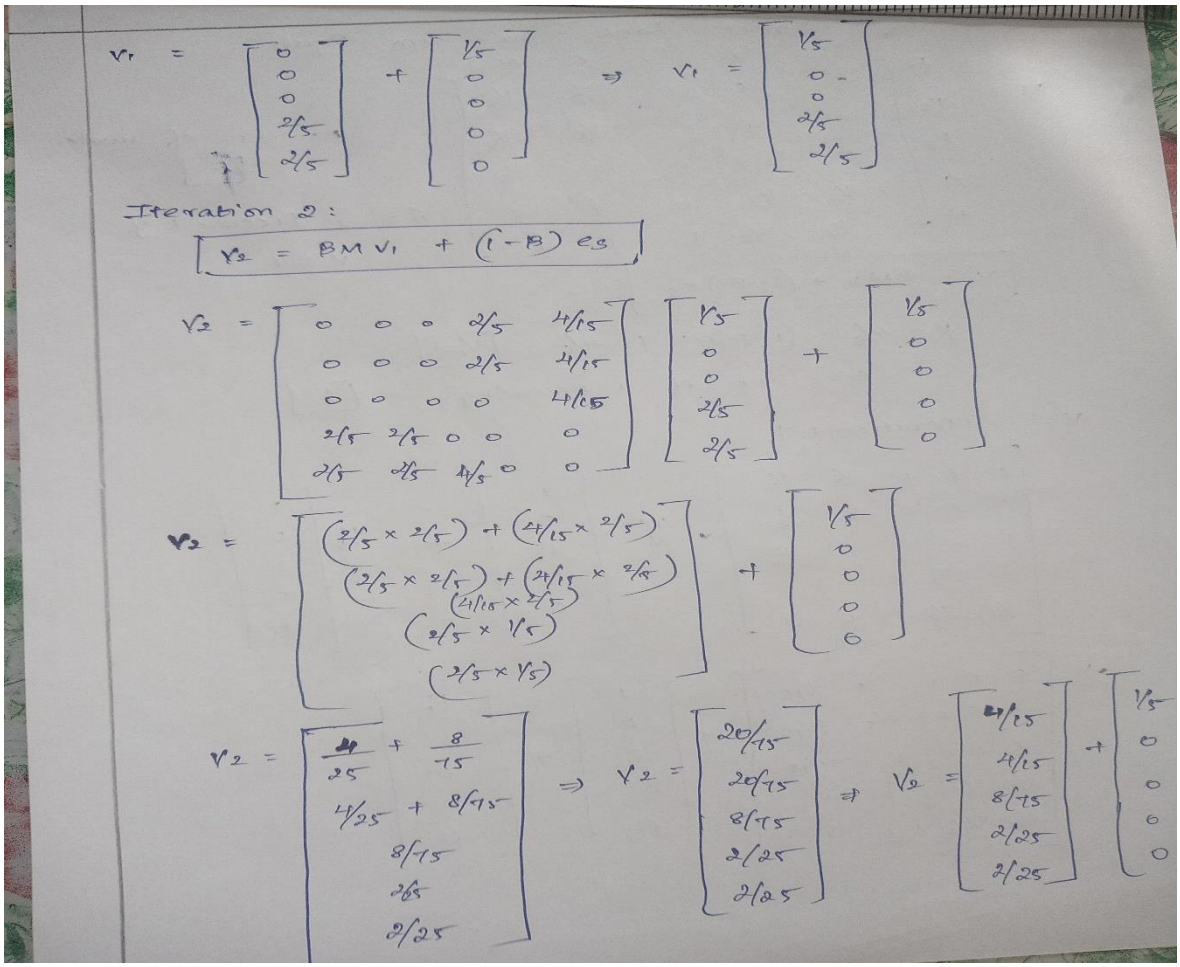
$\beta = 4/5 \quad (1-\beta) = 1 - 4/5 = 1/5$

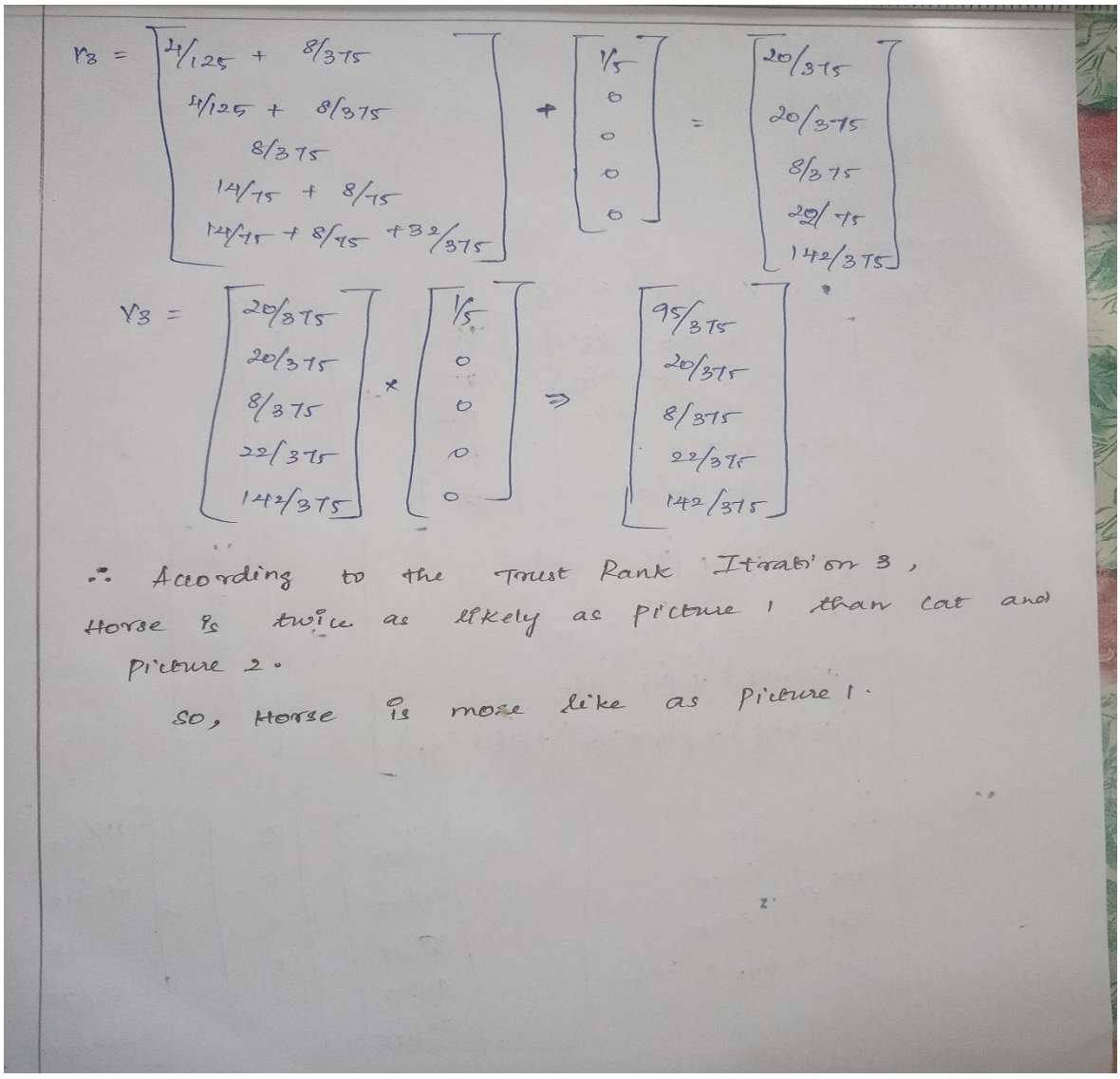
e_s = no. of 1 occuring = 1

V' = Iteration 1 :

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

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


Iteration 2:
$V_2 = BMV_1 + (1-B)e_3$



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

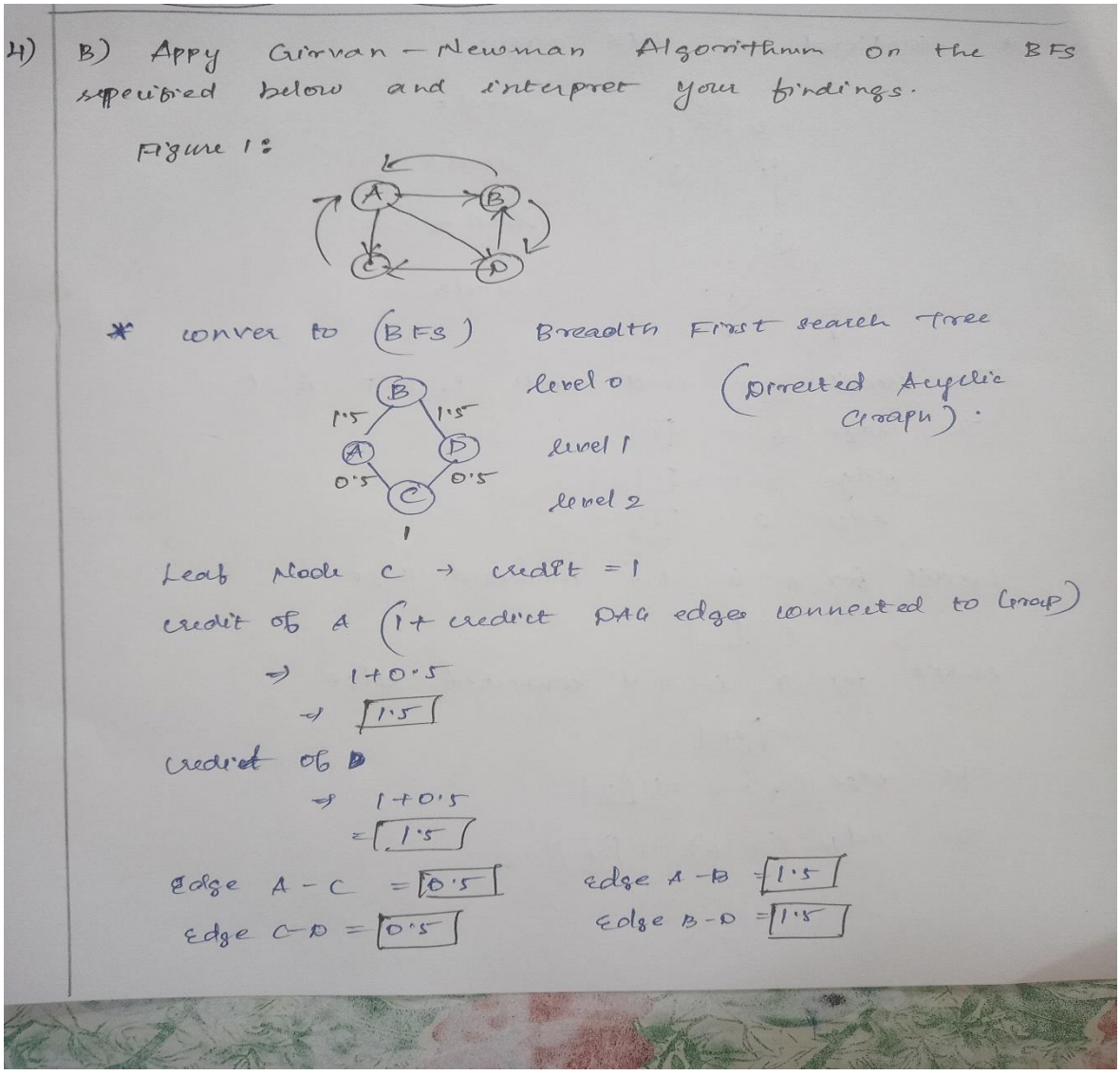
Edges – 2987624

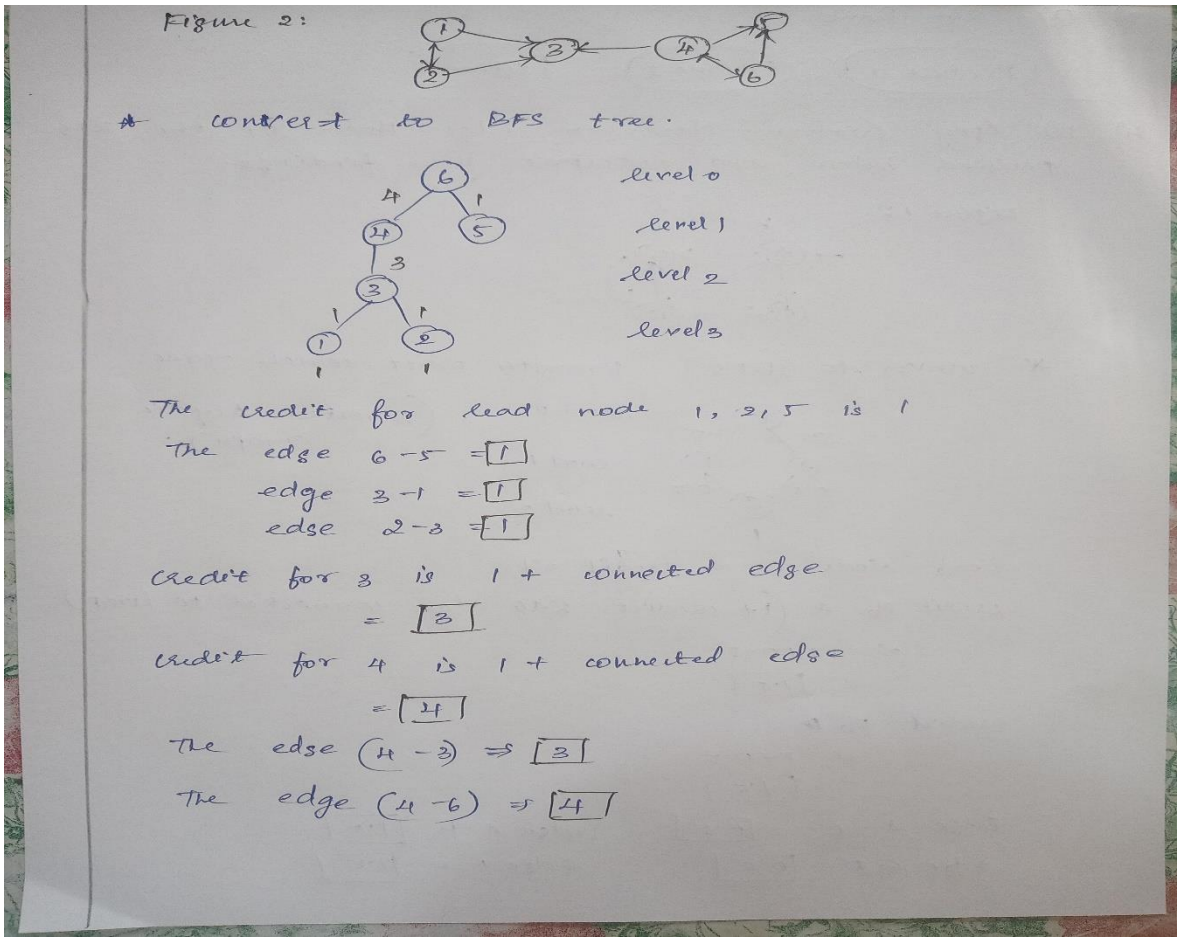
Number of Triangles- 3056386

Average clustering coefficient – 0.0808

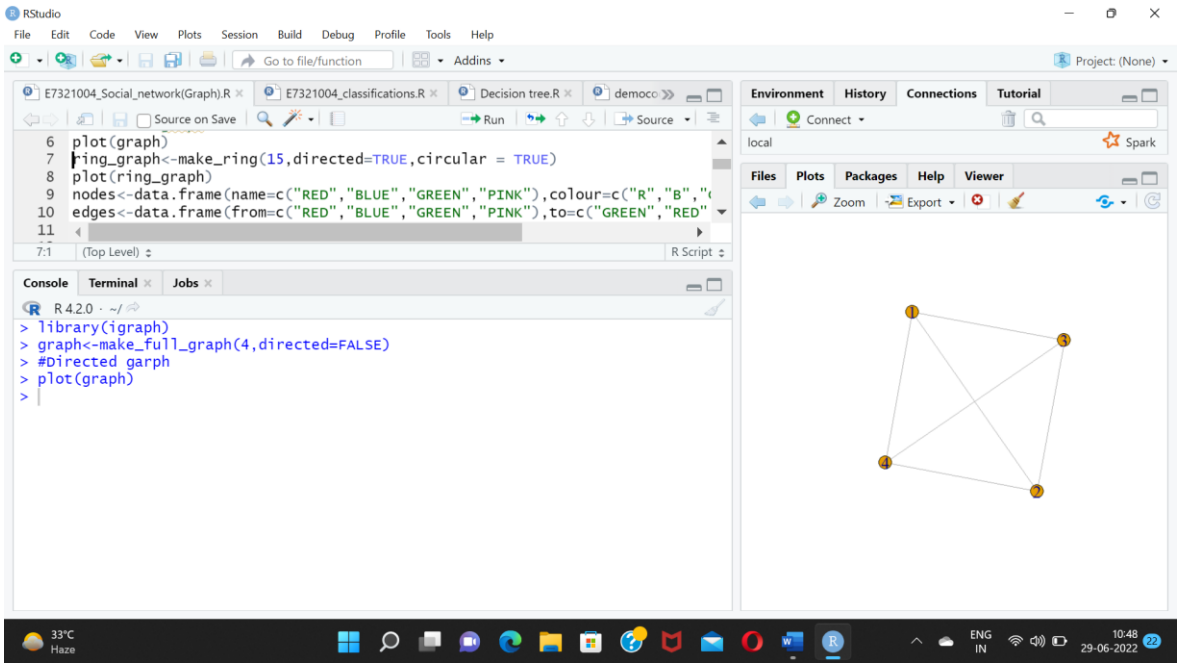
Number of communities – 8385

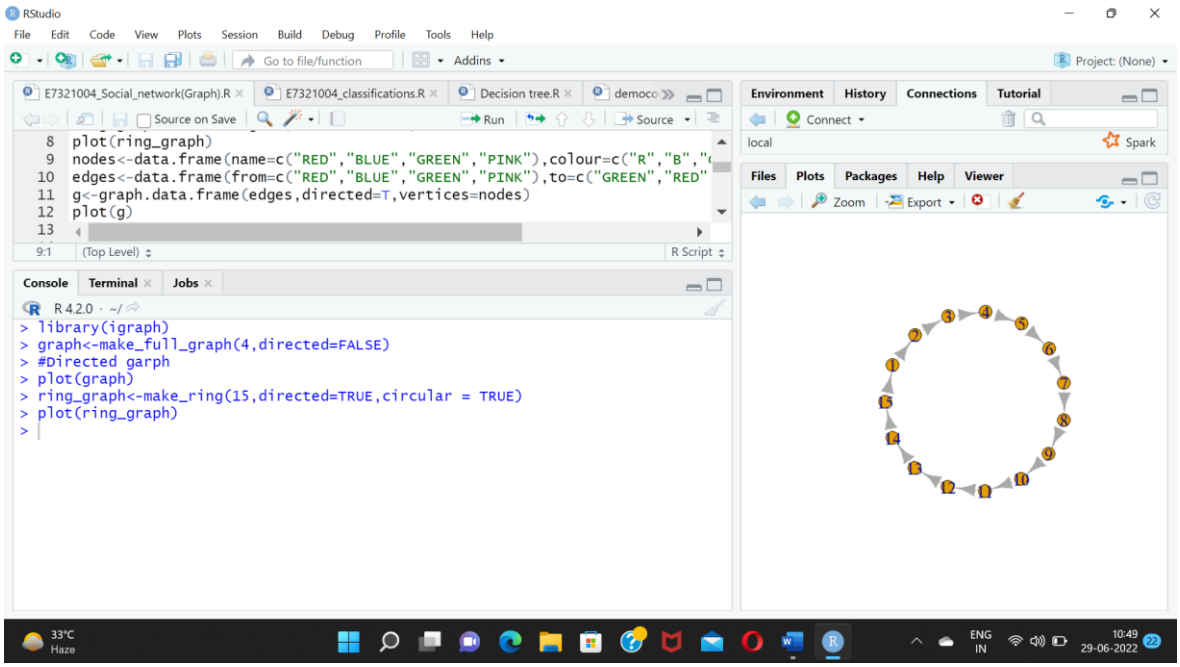
B. Apply Girvan-Newman Algorithm on the BFS specified below and interpret your findings for the figures 1 and 2 specified below.

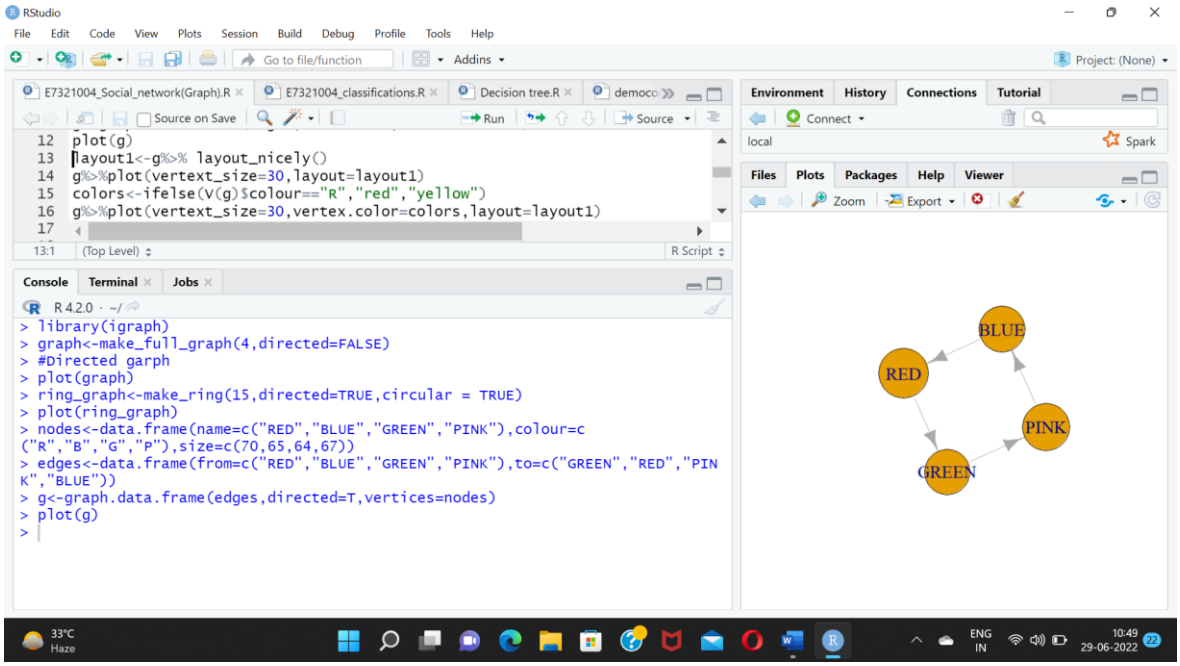


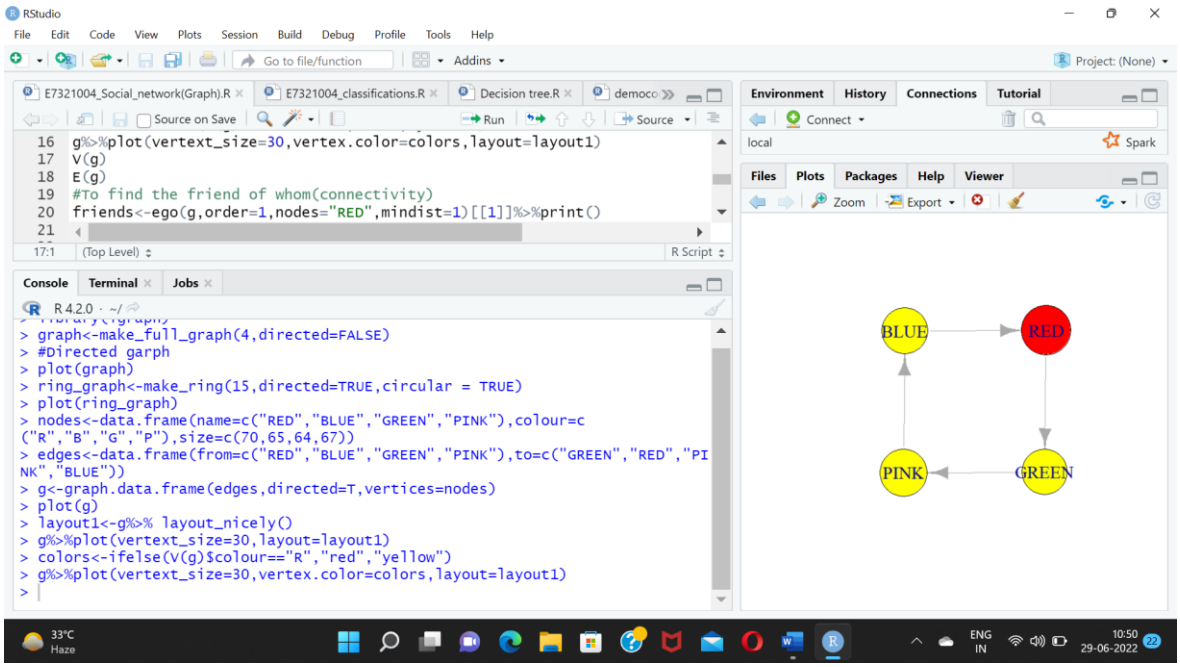


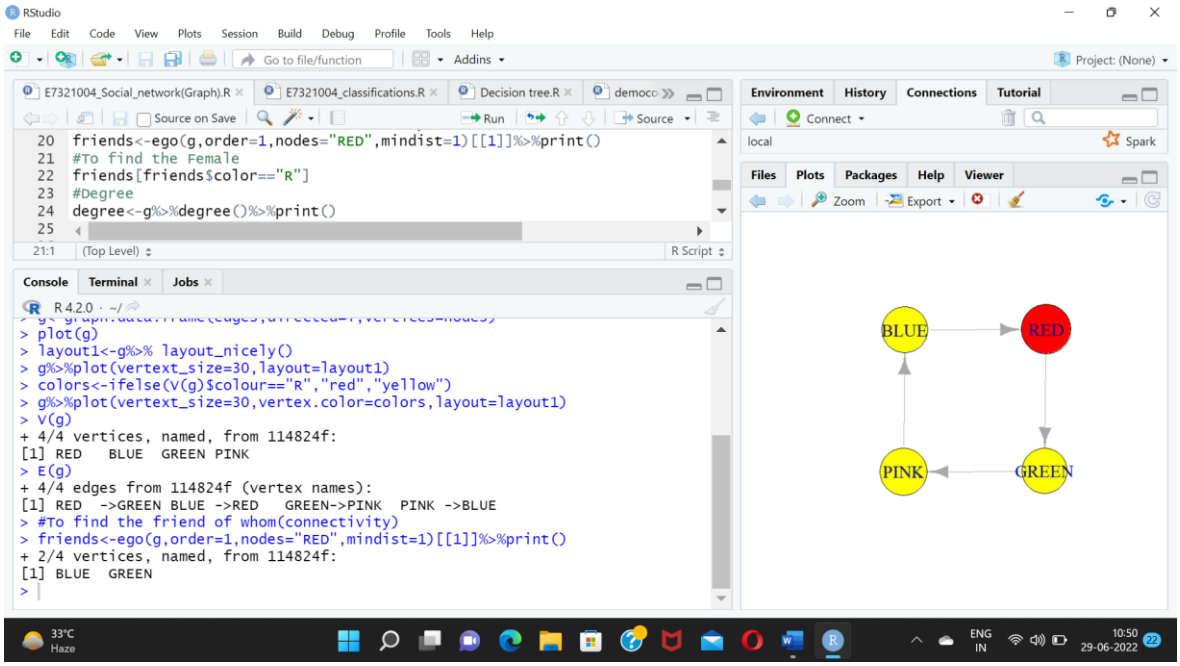
C. Implement the relationship between friends and family in Google+ through R programming.

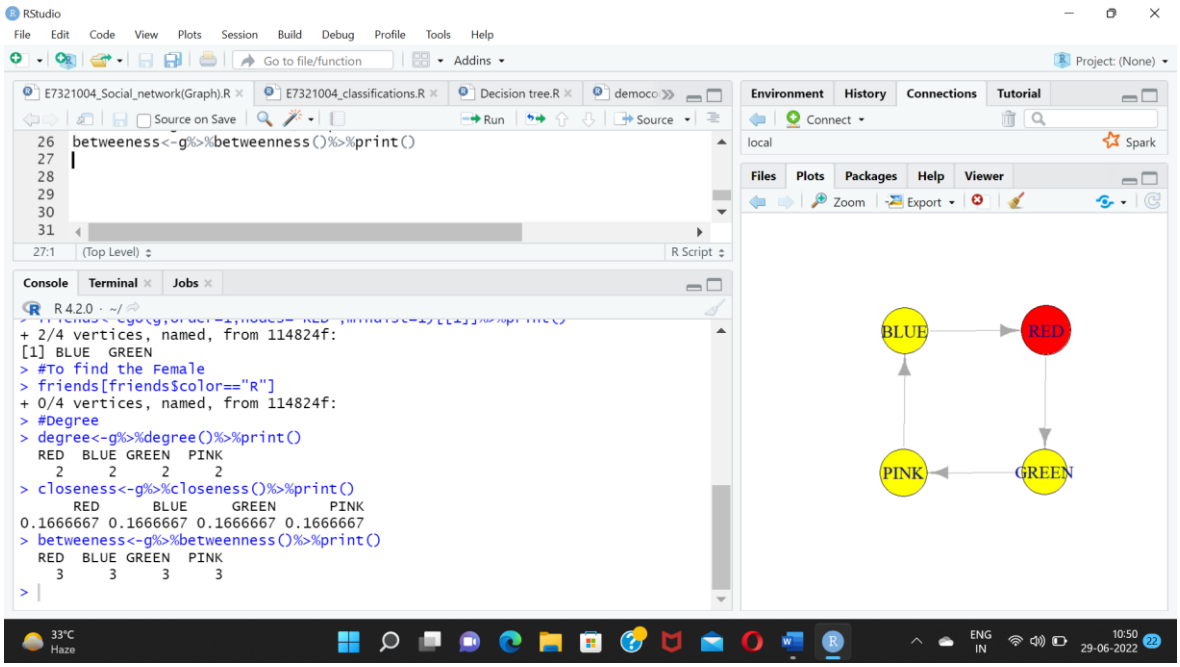



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
E7321004_Social_network(Graph).R
E7321004_classifications.R
Decision tree.R
plot(ring_graph)
nodes<-data.frame(name=c("RED","BLUE","GREEN","PINK"),colour=c("R","B",
edges<-data.frame(from=c("RED","BLUE","GREEN","PINK"),to=c("GREEN","RED"
g<-graph.data.frame(edges,directed=T,vertices=nodes)
plot(g)
Console Terminal Jobs
R 4.2.0 · ~/
> library(igraph)
> graph<-make_full_graph(4,directed=FALSE)
> #Directed garph
> plot(graph)
> ring_graph<-make_ring(15,directed=TRUE,circular = TRUE)
> plot(ring_graph)
Environment History Connections Tutorial
Files Plots Packages Help Viewer
33°C Haze
ENG IN
10:49 29-06-2022


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
E7321004_Social_network(Graph).R
E7321004_classifications.R
Decision tree.R
plot(g)
layout1<-g%>% layout_nicely()
g%>%plot(vertext_size=30,layout=layout1)
colors<-ifelse(V(g)$colour=="R","red","yellow")
g%>%plot(vertext_size=30,vertex.color=colors,layout=layout1)
Console Terminal Jobs
R 4.2.0 · ~/
> library(igraph)
> graph<-make_full_graph(4,directed=FALSE)
> #Directed garph
> 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","PIN
K","BLUE"))
> g<-graph.data.frame(edges,directed=T,vertices=nodes)
> plot(g)
Environment History Connections Tutorial
Files Plots Packages Help Viewer
BLUE
RED
PINK
GREEN
33°C Haze
ENG IN
10:49 29-06-2022


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
E7321004_Social_network(Graph).R
E7321004_classifications.R
Decision tree.R
Environment History Connections Tutorial
Files Plots Packages Help Viewer
16 g%>%plot(vertext_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()
> graph<-make_full_graph(4,directed=FALSE)
> #Directed garph
> 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","PI
NK","BLUE"))
> g<-graph.data.frame(edges,directed=T,vertices=nodes)
> plot(g)
> layout1<-g%>% layout_nicely()
> g%>%plot(vertext_size=30,layout=layout1)
> colors<-ifelse(V(g)$colour=="R","red","yellow")
> g%>%plot(vertext_size=30,vertex.color=colors,layout=layout1)
BLUE
RED
PINK
GREEN


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
E7321004_Social_network(Graph).R
E7321004_classifications.R
Decision tree.R
Environment History Connections Tutorial
Files Plots Packages Help Viewer
20 friends<-ego(g,order=1,nodes="RED",mindist=1)[[1]]%>%print()
21 #To find the Female
22 friends[friends$color=="R"]
23 #Degree
24 degree<-g%>%degree()%>%print()
> plot(g)
> layout1<-g%>% layout_nicely()
> g%>%plot(vertext_size=30,layout=layout1)
> colors<-ifelse(V(g)$colour=="R","red","yellow")
> g%>%plot(vertext_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
BLUE
RED
PINK
GREEN



CODE:

#Installing the package

install.packages("igraph")

library(igraph)

graph<-make_full_graph(4,directed=FALSE)

#Directed garph

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)

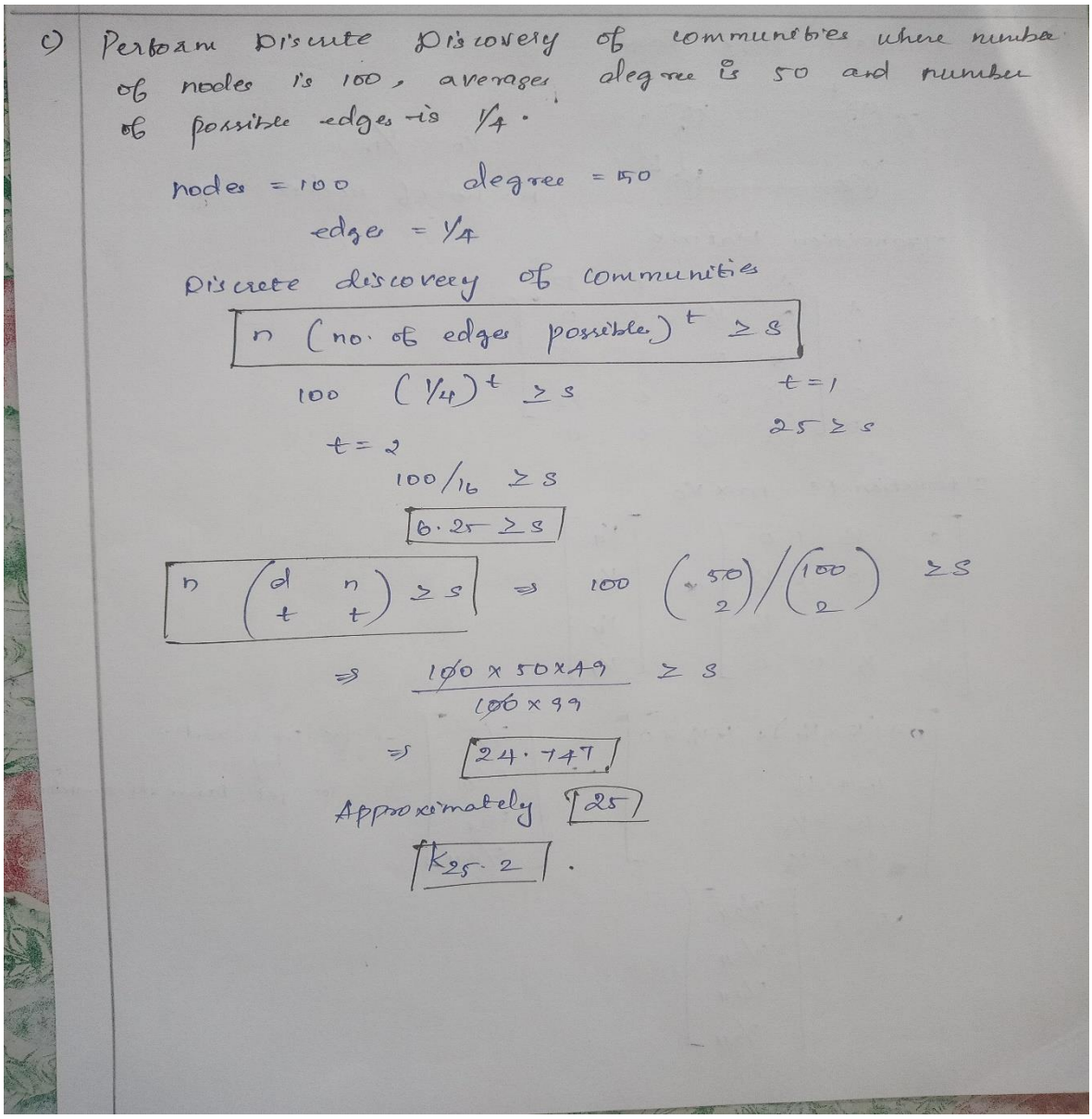
layout1<-g%>% layout_nicely()

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

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

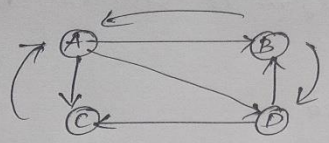
```
g%>%plot(vertext_size=30,vertex.color=colors,layout=layout1)

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()

betweeness<-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.



D. Calculate the Page Rank for the graphs specified below

4) D) Calculate Page Rank.



$V_0 = 1/n$

n = no. of nodes = 4

Transition Matrix

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

Iteration 1: $m \times V_0$

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

$$\begin{bmatrix} 0 + (1/2 \times 1/4) + 1 \times 1/4 + 0 \\ 1/2 + 1/8 \\ 1/12 + 1/8 \\ 1/12 + 1/8 \end{bmatrix} \Rightarrow \begin{bmatrix} 3/8 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix} \rightarrow$$ (× By 2 and ÷ by 3 to get same denominator)

$$V_1 = \begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}$$

Iteration 2:

$$\boxed{M \times V_0 = V_2}$$

$$V_2 = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}$$

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

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

Iteration 3:

$$V_3 = M \times V_2$$

$$\Rightarrow \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 15/48 \\ 11/48 \\ 11/48 \\ 11/48 \end{bmatrix}$$

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

$$V_3 = \begin{bmatrix} 11/32 \\ 7/32 \\ 7/32 \\ 7/32 \end{bmatrix}$$

∴ As the Iteration goes on, we stop at the Iteration 3.

Trust Rank for source {B, 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

(To Get Trust rank) $= (1/5) \times 1/2$

$= 1/10$

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

$\boxed{v' = \beta M r + (1-\beta)e_s/|s|}$ (Iteration 1:)

$$v_1 \Rightarrow \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 4/5 \\ 4/5 \\ 4/5 \\ 4/5 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1/2 \\ 0 \\ 1/2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

$$a_1 \Rightarrow \begin{bmatrix} 0 & 4/10 & 4/5 & 0 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 4/10 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1/2 \\ 0 \\ 1/2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

$$V_1 = \begin{bmatrix} 4/20 \\ 4/20 \\ 4/20 \\ 4/20 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix} = \begin{bmatrix} 4/20 \\ 6/20 \\ 4/20 \\ 6/20 \end{bmatrix}$$

$$V_1 = \begin{bmatrix} 1/5 \\ 3/10 \\ 1/5 \\ 3/10 \end{bmatrix}$$

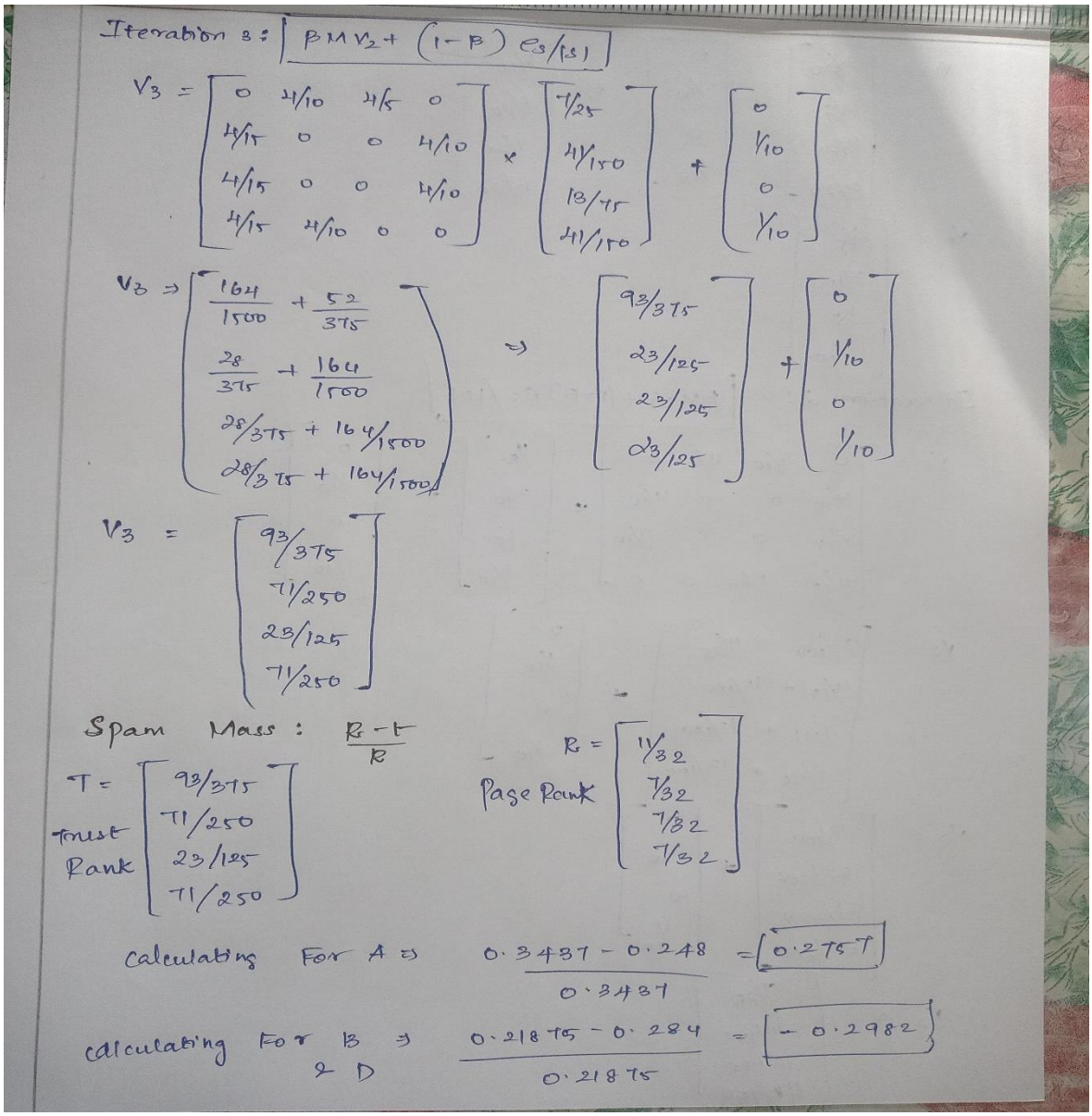
Iteration 2: $\boxed{BM\,V_1 + (1-B)\,e_S/|S|}$

$$V_2 = \begin{bmatrix} 0 & 4/10 & 4/5 & 0 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 4/10 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 1/5 \\ 3/10 \\ 1/5 \\ 3/10 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

$$V_2 = \begin{bmatrix} 12/100 + 4/25 \\ 4/15 + 12/100 \\ 4/15 + 12/100 \\ 4/15 + 12/100 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

$$V_2 = \begin{bmatrix} 28/100 \\ 52/300 \\ 52/300 \\ 52/300 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

$$V_2 \Rightarrow \begin{bmatrix} 7/25 \\ 41/150 \\ 13/75 \\ 41/150 \end{bmatrix}$$


Iteration 3 : βMV2 + (1−β) e3/|e|
V3 = [0 4/10 4/5 0; 4/15 0 0 4/10; 4/15 0 0 4/10; 4/15 4/10 0 0] × [7/25; 41/150; 13/75; 41/150] + [0; 1/10; 0; 1/10]
V3 ⇒ [164/1500 + 52/375; 28/375 + 164/1500; 28/375 + 164/1500; 28/375 + 164/1500]
⇒ [93/375; 23/125; 23/125; 23/125] + [0; 1/10; 0; 1/10]
V3 = [93/375; 71/250; 23/125; 71/250]
Spam Mass : R − T / R
T = Trust Rank [93/375; 71/250; 23/125; 71/250]
R = Page Rank [11/32; 7/32; 7/32; 7/32]
Calculating For A ⇒ 0.3437 − 0.248 / 0.3437 = 0.275
Calculating For B ⇒ 0.21875 − 0.284 / 0.21875 = −0.2982
& D

Calculating For C,

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

Node	Page Rank	Trust Rank	Spam Mass
A	11/32	93/375	0.2757
B	7/32	71/250	-0.2982
C	7/32	23/125	0.158857
D	7/32	71/250	-0.2982

∴ The conculusion 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 1, 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

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

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

Iteration 1:

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

$$V_1 = \begin{bmatrix} 0 & 4/10 & 4/5 & 0 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 4/10 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0 \\ 0 \\ 1/2 \\ 1/2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1/10 \\ 1/10 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 2/5 \\ 2/10 \\ 2/10 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1/10 \\ 1/10 \end{bmatrix} \Rightarrow V_1 = \begin{bmatrix} 2/5 \\ 1/5 \\ 3/10 \\ 1/10 \end{bmatrix}$$

Iteration 2:

$$V_2 = BMV_1 + 1-B\,es/|s|$$

$$\Rightarrow \begin{bmatrix} 0 & 4/10 & 4/5 & 0 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 4/10 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 2/5 \\ 1/5 \\ 3/20 \\ 1/10 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1/10 \\ 1/10 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 4/50 + 12/50 \\ 8/75 + 4/100 \\ 8/75 + 4/100 \\ 8/75 + 4/50 \end{bmatrix} \Rightarrow \begin{bmatrix} 0 \\ 0 \\ 1/10 \\ 1/10 \end{bmatrix} + \begin{bmatrix} 8/25 \\ 11/75 \\ 11/75 \\ 14/75 \end{bmatrix}$$

$$V_2 \Rightarrow \begin{bmatrix} 8/25 \\ 11/75 \\ 37/150 \\ 43/100 \end{bmatrix}$$

Iteration 3:

$$V_3 = \beta M V_2 + (1-\beta)\, e^s / |s|$$

$$V_3 = \begin{bmatrix} 0 & 4/10 & 4/5 & 0 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 0 & 0 & 4/10 \\ 4/15 & 4/10 & 0 & 0 \end{bmatrix} \begin{bmatrix} 8/25 \\ 11/75 \\ 37/150 \\ 43/150 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1/10 \\ 1/10 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 44/150 + 148/450 \\ 32/375 + 172/1500 \\ 32/375 + 172/1500 \\ 32/375 + 44/150 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1/10 \\ 1/10 \end{bmatrix} \Rightarrow V_3 = \begin{bmatrix} 32/125 \\ 1/5 \\ 3/10 \\ 183/450 \end{bmatrix}$$

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

$t = \begin{bmatrix} 32/125 \\ 1/5 \\ 3/10 \\ 183/450 \end{bmatrix}$ (Trust Rank) $R =$ Page Rank $\begin{bmatrix} 11/32 \\ 7/32 \\ 7/32 \\ 7/32 \end{bmatrix}$

calculating for A, $\frac{0.345 - 0.256}{0.343} \Rightarrow 0.2536$

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

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

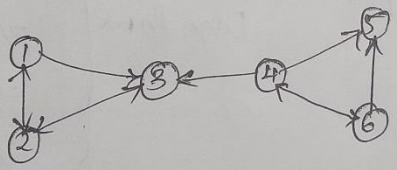
calculating for D, $\frac{0.21875 - 0.244}{0.21875} = -0.1154$

Node	Page Rank (R)	Trust Rank (t)	R-t/R
A	11/32	32/125	0.2536
B	7/32	1/5	0.08571
C	7/32	3/10	-0.3714
D	7/32	183/750	-0.1154

∴ 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 closer to 0 than 1 so it is not spam.

Page Rank,



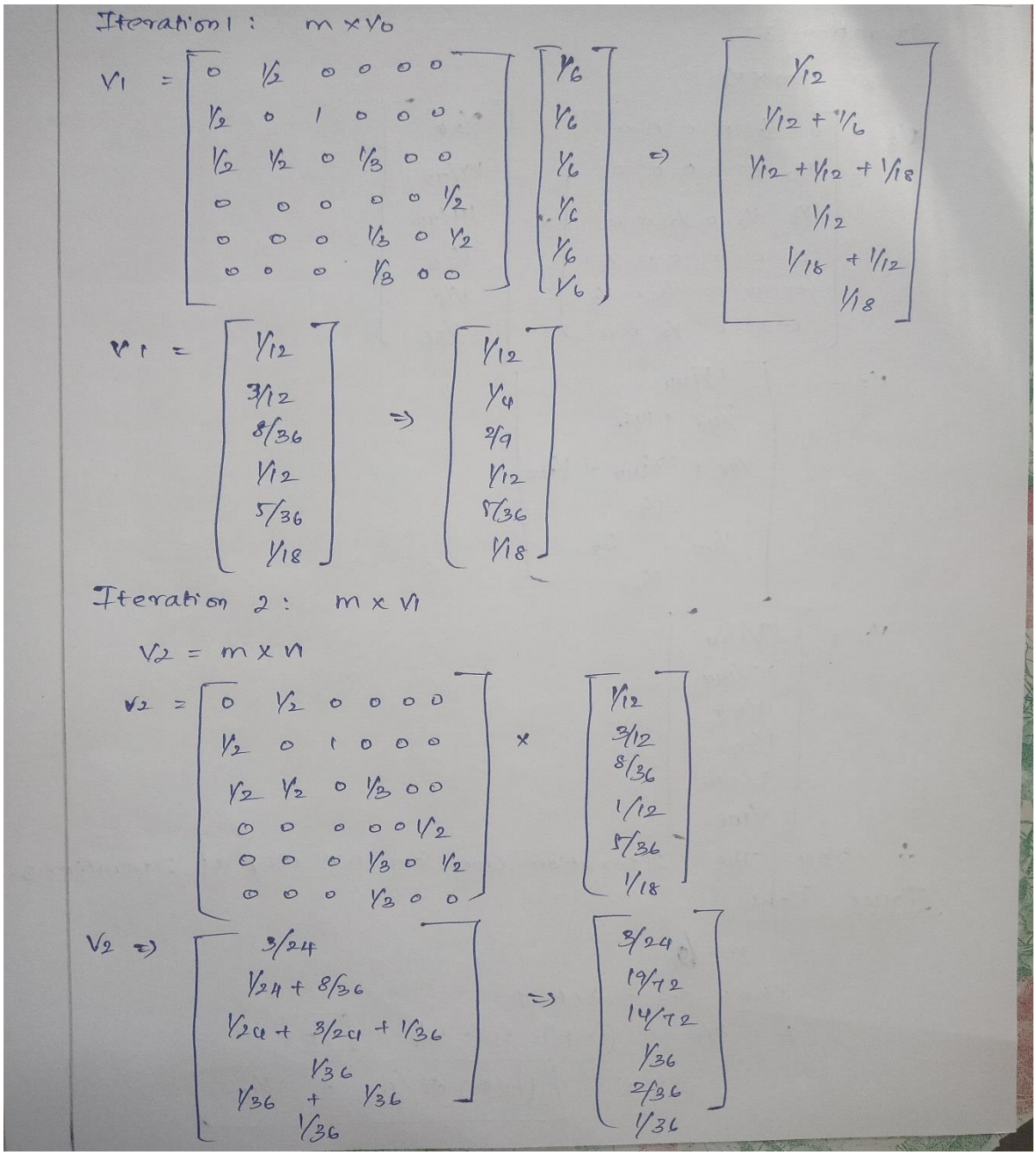
no. of nodes = 6

$n = 6$

$V_0 = 1/6$

Transition Matrix

$$m = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} 0 & 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1/3 & 0 & 1/2 \\ 0 & 0 & 0 & 1/3 & 0 & 0 \end{bmatrix} \end{array}$$


Iteration 1 : m x V0
V1 =
⇒
V1 =
⇒
Iteration 2 : m x V1
V2 = m x V1
V2 =
×
V2 ⇒
⇒

Iteration 3:

$V_3 = m \times V_2$

$$V_3 = \begin{bmatrix} 0 & 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1/3 & 0 & 1/2 \\ 0 & 0 & 0 & 1/3 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 3/24 \\ 19/72 \\ 14/72 \\ 1/36 \\ 1/18 \\ 1/36 \end{bmatrix}$$

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

$$V_3 = \begin{bmatrix} 19/144 \\ 37/144 \\ 11/54 \\ 1/72 \\ 5/216 \\ 1/108 \end{bmatrix}$$

∴ Since the Iteration goes on we stop at Iteration 3.

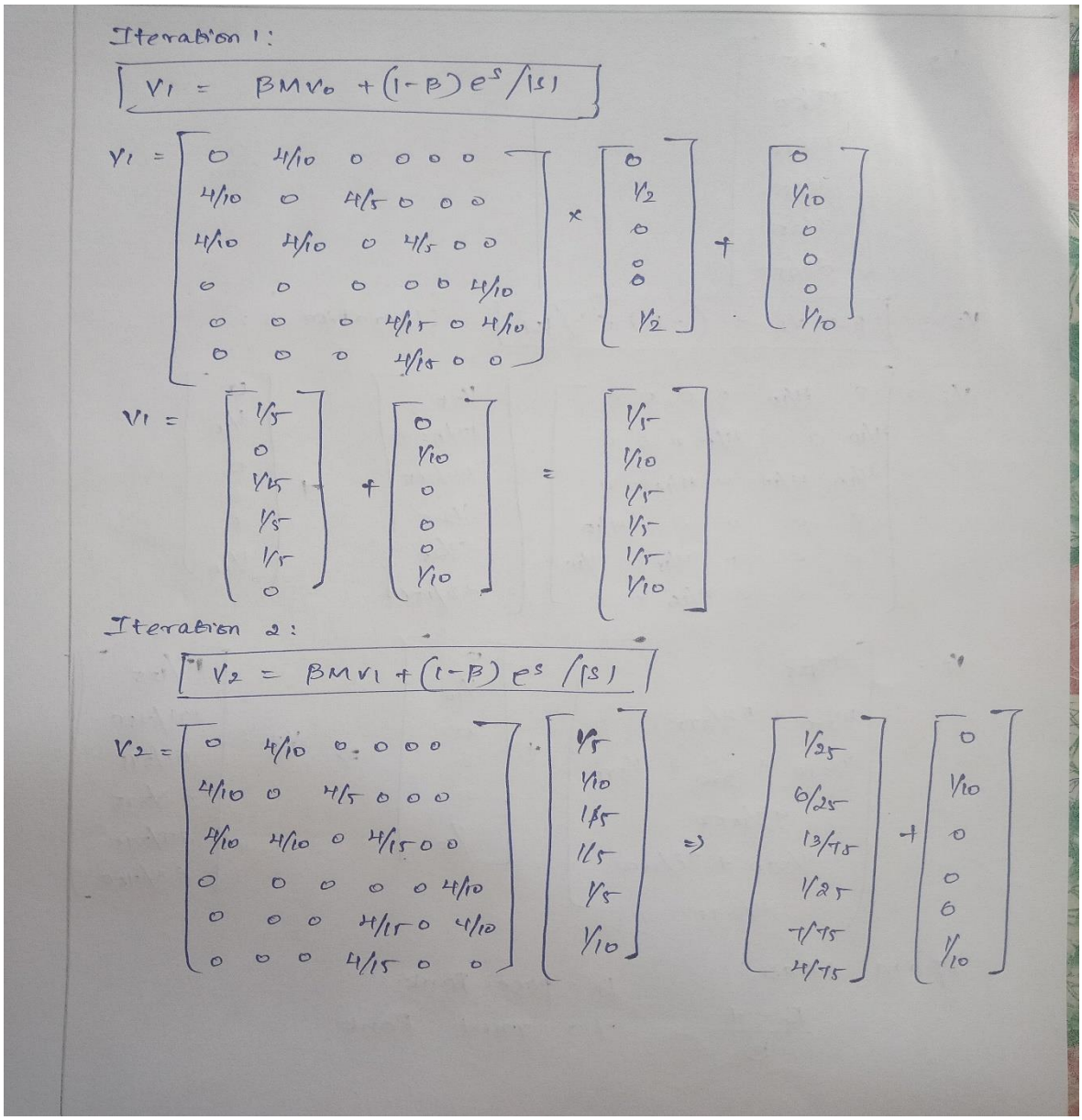
Trust Rank

$n = 6$

Source = {2, 6}

$\beta = 4/5$ $\quad (1-\beta) = 1 - 4/5 = \boxed{1/5}$

$s = 2$ $\quad \boxed{(1-\beta)\,es/|S| = 1/10}$


Iteration 1:
$V_1 = BMV_0 + (1-B)e^s/|s|$
$V_1 = \begin{bmatrix} 0 & 4/10 & 0 & 0 & 0 & 0 \\ 4/10 & 0 & 4/5 & 0 & 0 & 0 \\ 4/10 & 4/10 & 0 & 4/5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4/10 \\ 0 & 0 & 0 & 4/15 & 0 & 4/10 \\ 0 & 0 & 0 & 4/15 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1/2 \\ 0 \\ 0 \\ 0 \\ 1/2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 0 \\ 0 \\ 1/10 \end{bmatrix}$
$V_1 = \begin{bmatrix} 1/5 \\ 0 \\ 1/15 \\ 1/5 \\ 1/15 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 0 \\ 0 \\ 1/10 \end{bmatrix} = \begin{bmatrix} 1/5 \\ 1/10 \\ 1/15 \\ 1/5 \\ 1/15 \\ 1/10 \end{bmatrix}$
Iteration 2:
$V_2 = BMV_1 + (1-B)e^s/|s|$
$V_2 = \begin{bmatrix} 0 & 4/10 & 0 & 0 & 0 & 0 \\ 4/10 & 0 & 4/5 & 0 & 0 & 0 \\ 4/10 & 4/10 & 0 & 4/15 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4/10 \\ 0 & 0 & 0 & 4/15 & 0 & 4/10 \\ 0 & 0 & 0 & 4/15 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1/5 \\ 1/10 \\ 1/5 \\ 1/5 \\ 1/5 \\ 1/10 \end{bmatrix} \Rightarrow \begin{bmatrix} 1/25 \\ 6/25 \\ 13/75 \\ 1/25 \\ 7/75 \\ 4/75 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 0 \\ 0 \\ 1/10 \end{bmatrix}$

$$V_2 = \begin{bmatrix} 1/25 \\ 17/50 \\ 13/75 \\ 1/25 \\ 7/25 \\ 23/150 \end{bmatrix}$$

$$V_3 = \boxed{\beta M V_2 + (1-\beta) e^3/|3|} \qquad \text{Iteration 3 :}$$

$$V_3 = \begin{bmatrix} 0 & 4/10 & 0 & 0 & 0 & 0 \\ 4/10 & 0 & 4/5 & 0 & 0 & 0 \\ 4/10 & 4/10 & 0 & 4/15 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4/10 \\ 0 & 0 & 0 & 4/15 & 0 & 4/10 \\ 0 & 0 & 0 & 4/15 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1/25 \\ 17/50 \\ 13/75 \\ 1/25 \\ 7/25 \\ 23/150 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 0 \\ 0 \\ 1/10 \end{bmatrix}$$

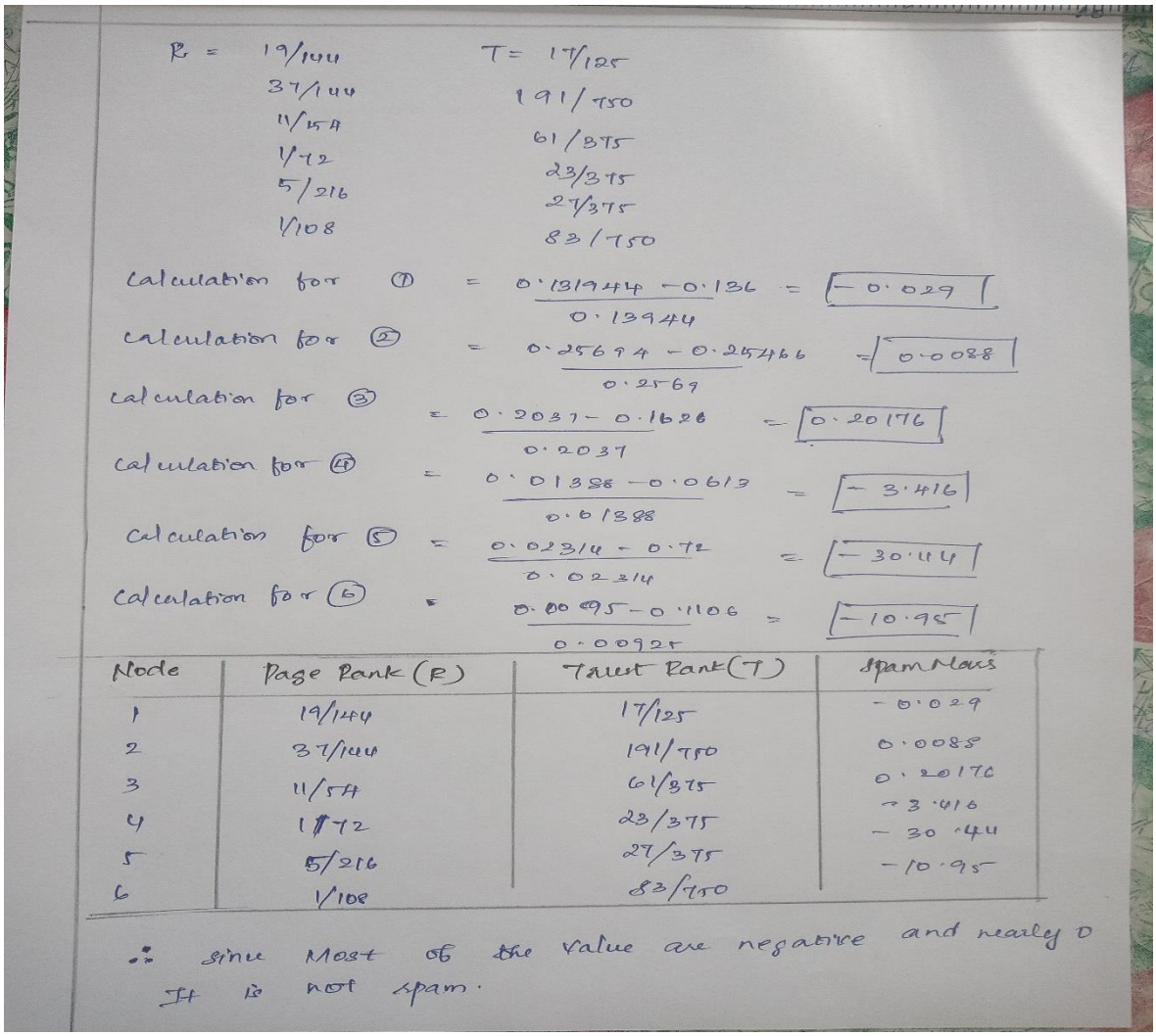
$$V_3 = \begin{bmatrix} 17/25 \\ 4/250 + 52/375 \\ 4/250 + \frac{34}{250} + 9/375 \\ 92/1500 \\ 4/375 + 92/1500 \\ 4/375 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 0 \\ 0 \\ 1/10 \end{bmatrix} \Rightarrow V_3 = \begin{bmatrix} 17/125 \\ 191/750 \\ 61/375 \\ 23/375 \\ 27/375 \\ 83/750 \end{bmatrix}$$

Spam Mass :

$\frac{R-t}{R}$

R = page Rank

t = trust Rank



R = 19/144, 37/144, 11/54, 1/72, 5/216, 1/108

T = 17/125, 191/750, 61/375, 23/375, 27/375, 82/750

Calculation for ① = (0.131944 − 0.136) / 0.131944 = −0.029

Calculation for ② = (0.25694 − 0.25466) / 0.2569 = 0.0088

Calculation for ③ = (0.2037 − 0.1626) / 0.2037 = 0.20176

Calculation for ④ = (0.01388 − 0.0613) / 0.01388 = −3.416

Calculation for ⑤ = (0.02314 − 0.72) / 0.02314 = −30.144

Calculation for ⑥ = (0.00695 − 0.1106) / 0.00925 = −10.95

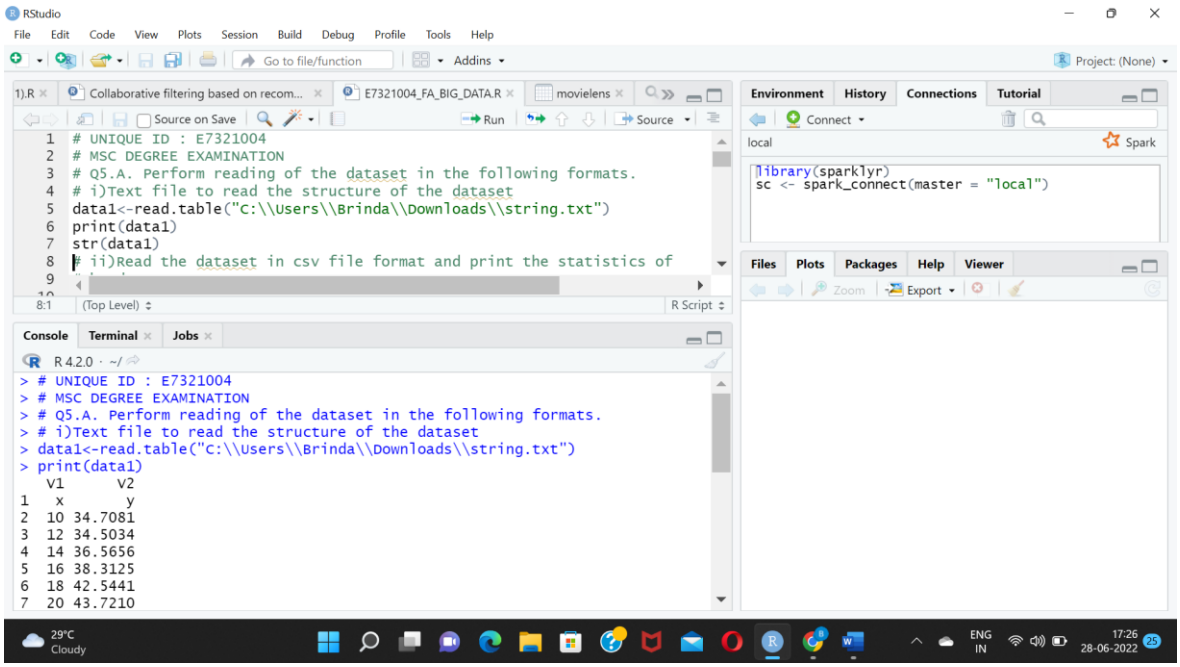
Node	Page Rank (R)	Tweet Rank (T)	Spam Mass
1	19/144	17/125	−0.029
2	37/144	191/750	0.0085
3	11/54	61/375	0.20176
4	1/72	23/375	−3.416
5	5/216	27/375	−30.44
6	1/108	82/750	−10.95

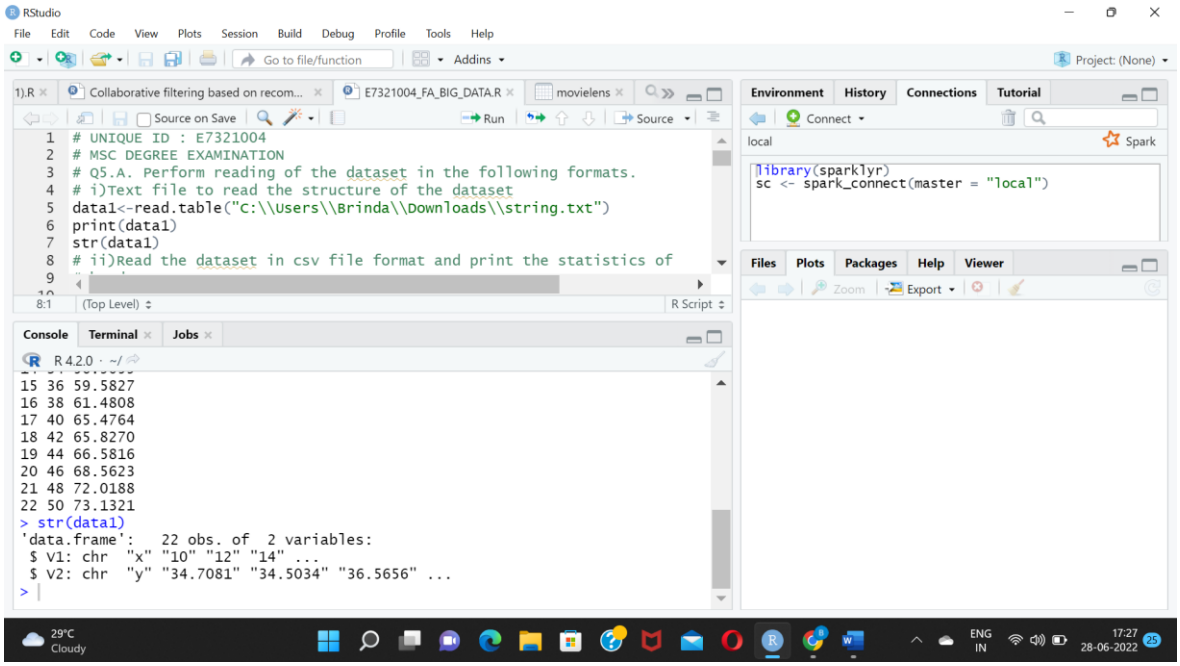
∴ Since Most of the value are negative and nearly 0 It is not spam.

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

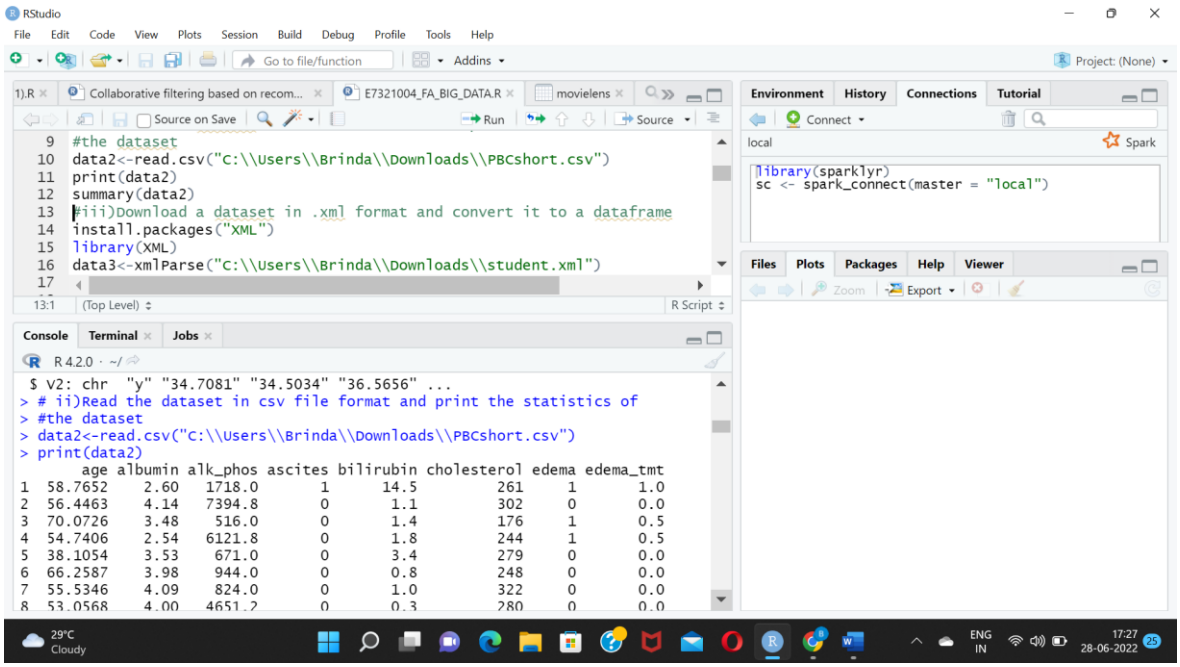
A. Perform reading of the dataset in the following formats.

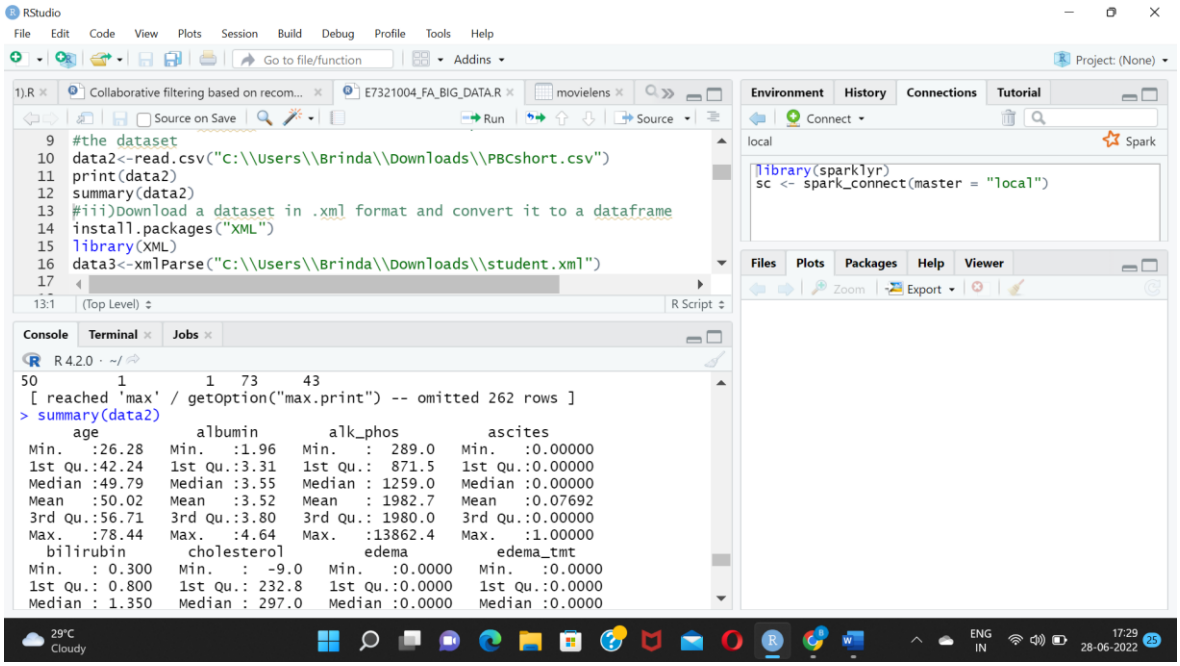
i) Read the dataset in .txt file and print the structure of the dataset.



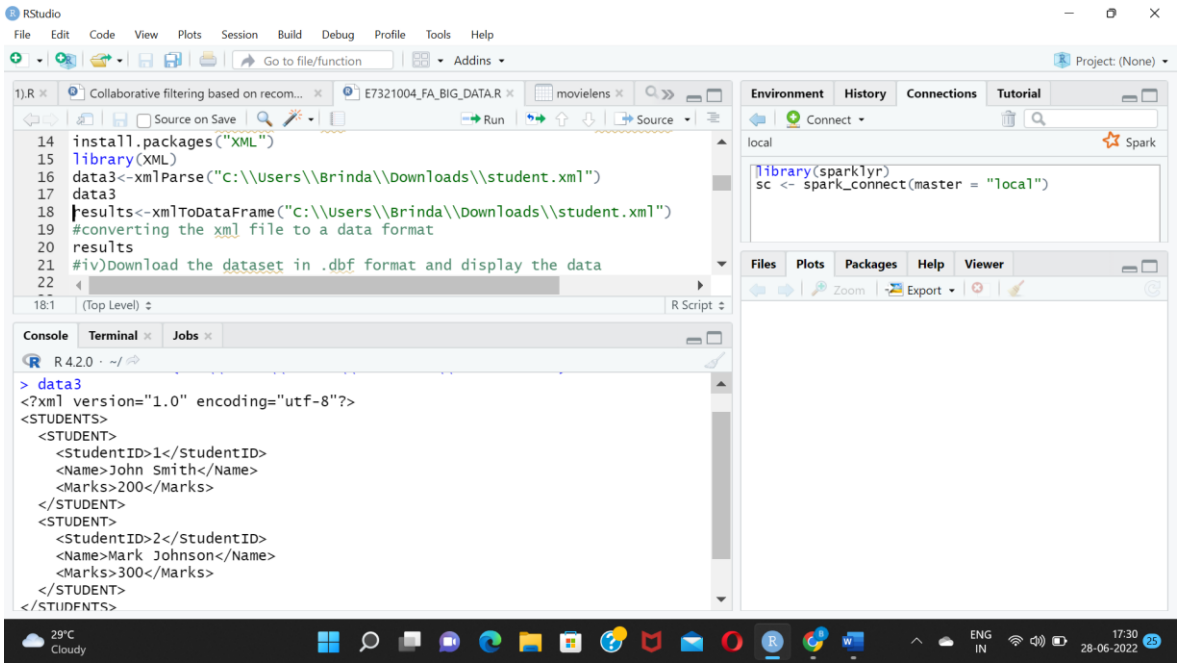


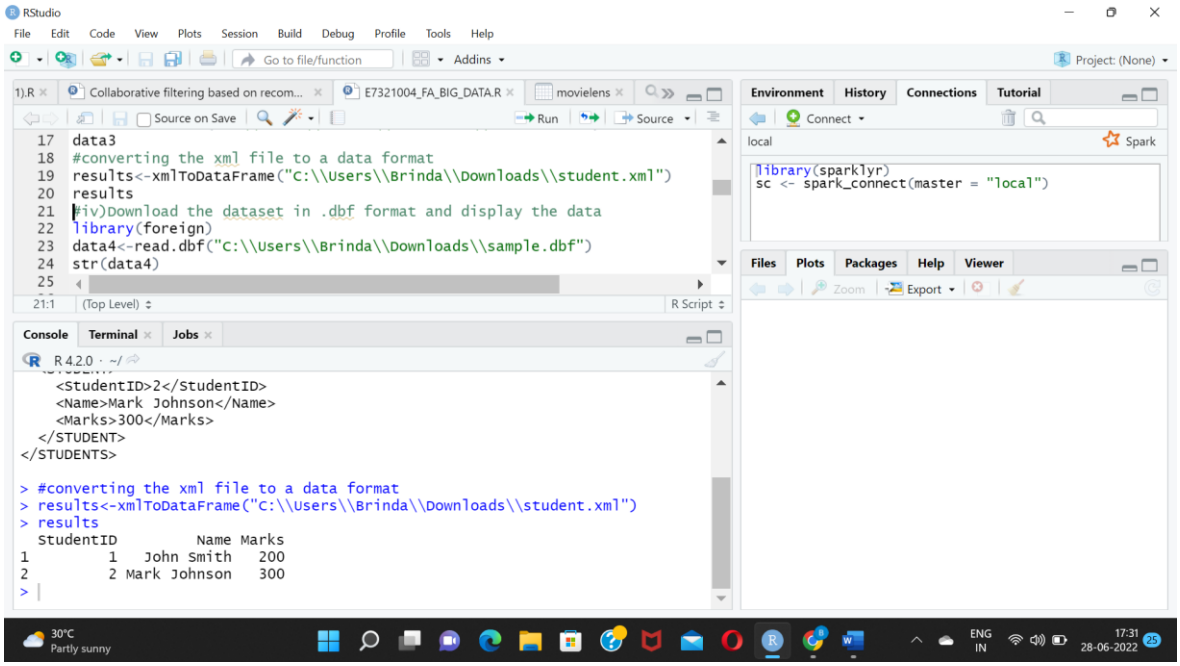
ii) Read the dataset in .csv file format and print the statistics of the dataset.



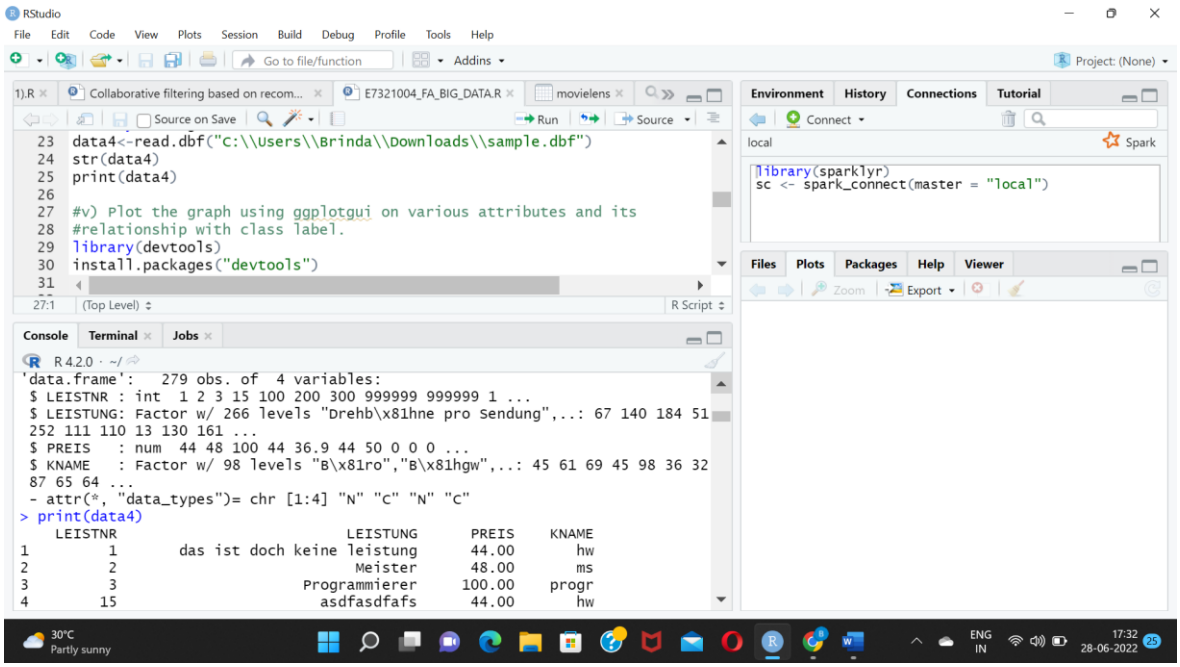


iii)Download a dataset in .xml format and convert it to a dataframe.



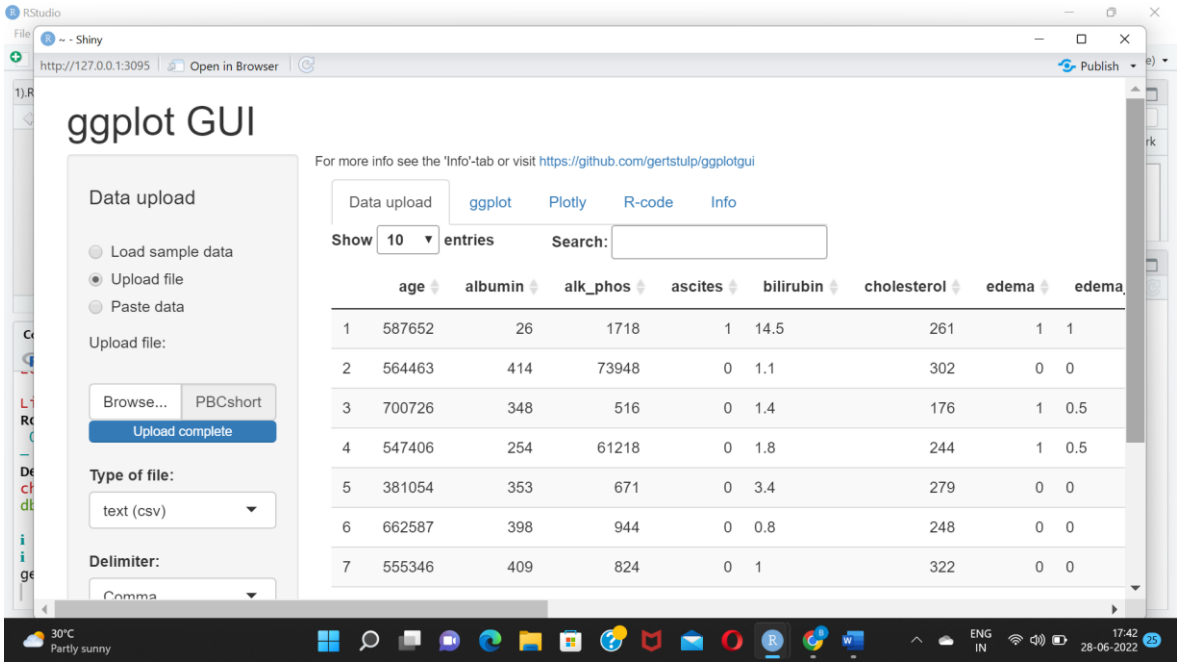


iv)Download the dataset in .dbf format and display the data.

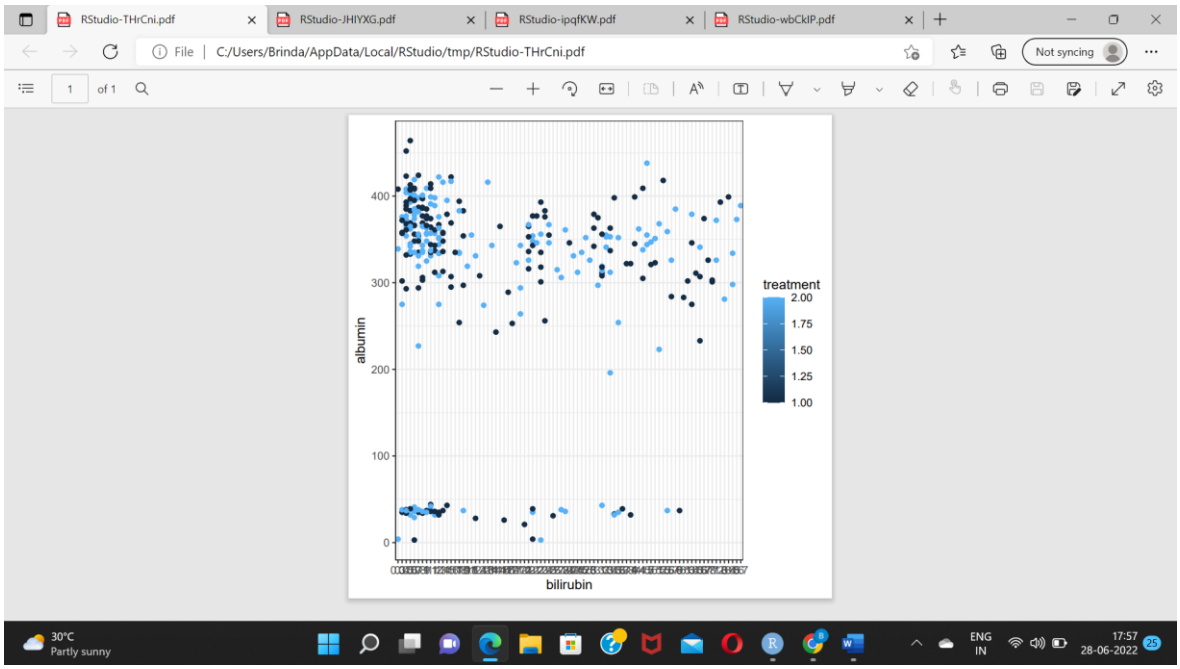


v) Plot the graph using ggplotgui on various attributes and its relationship with class label.

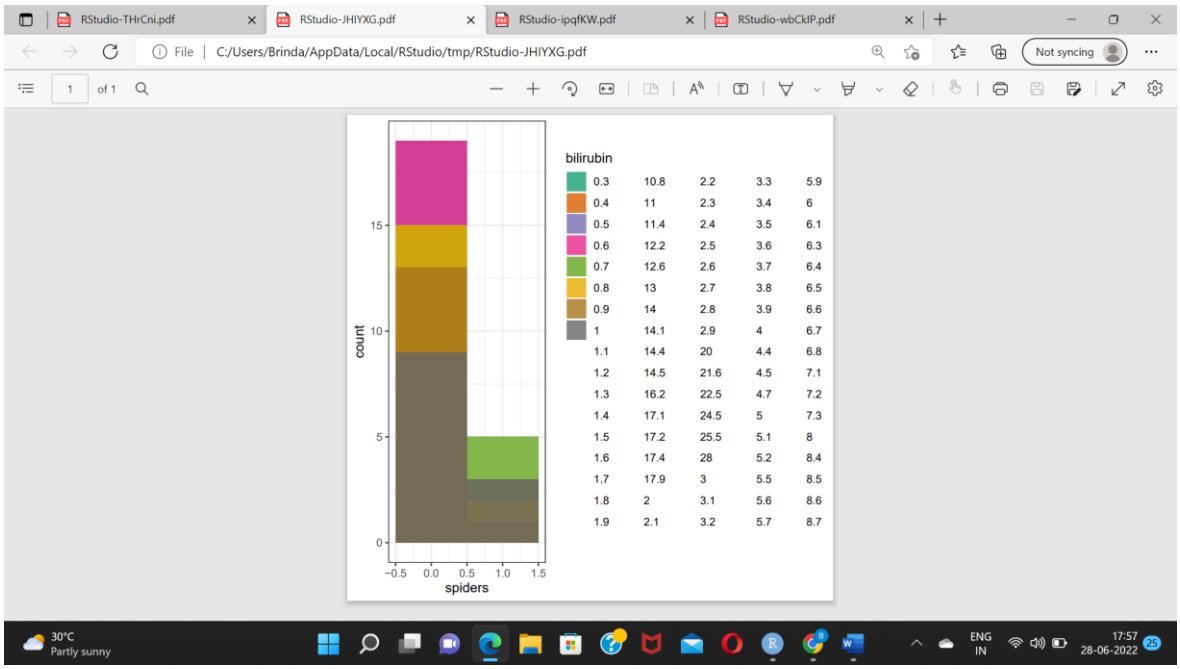
DATASET : PBCshort csv file



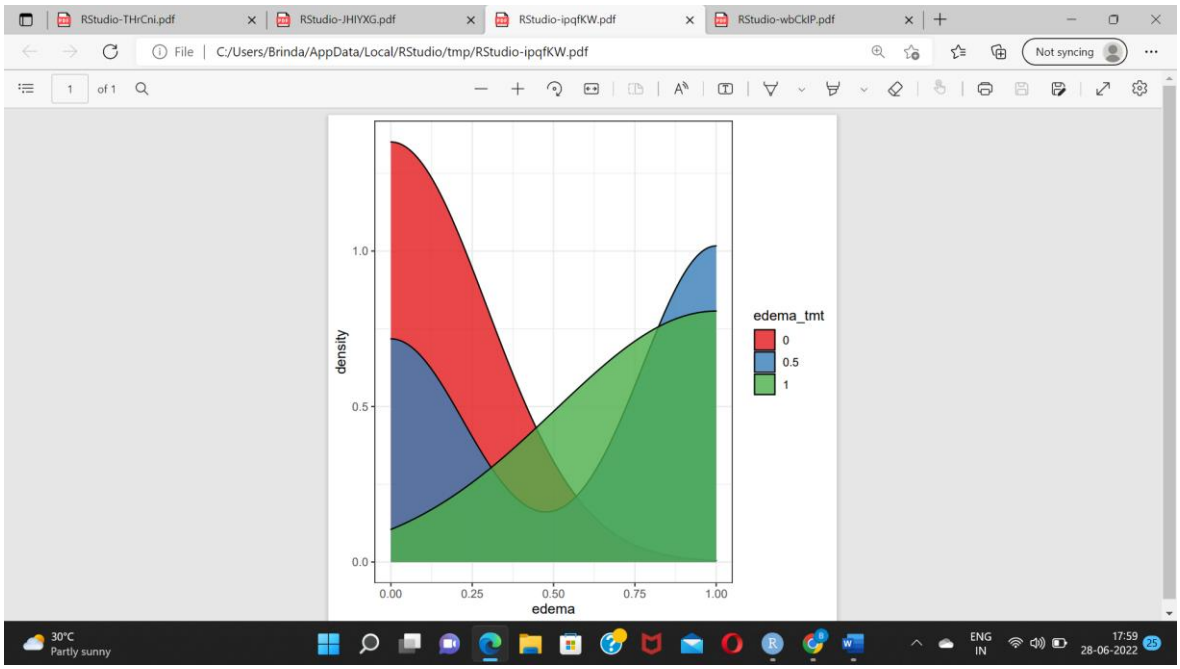
SCATTER PLOT



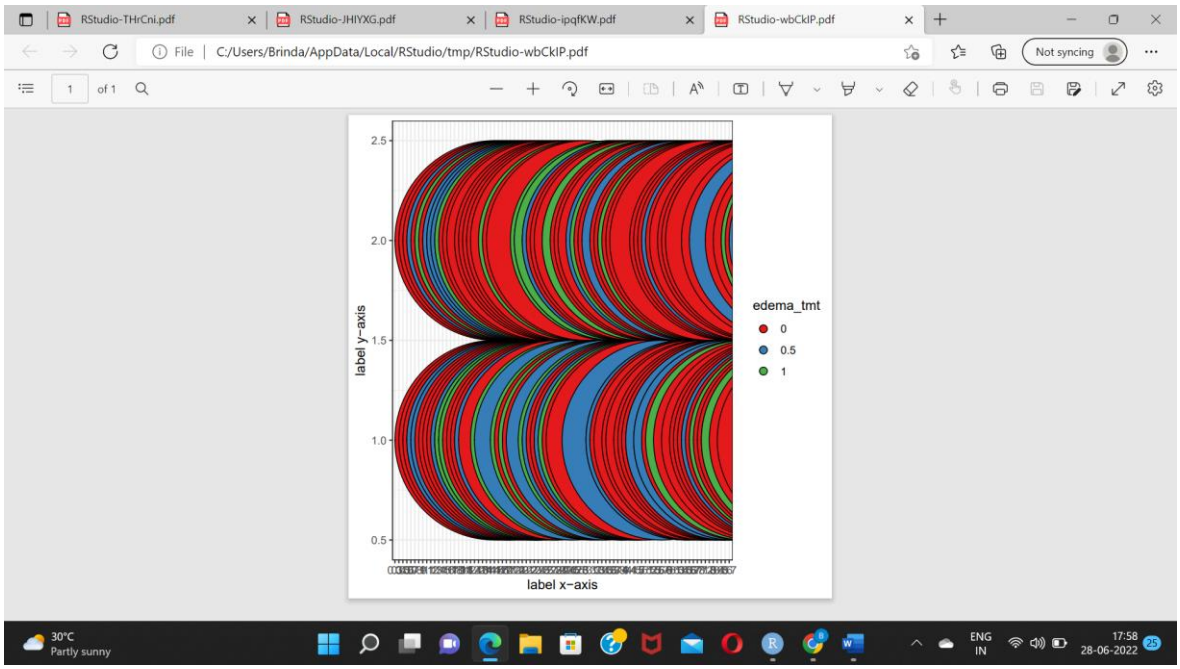
HISTOGRAM



DENSITY PLOT



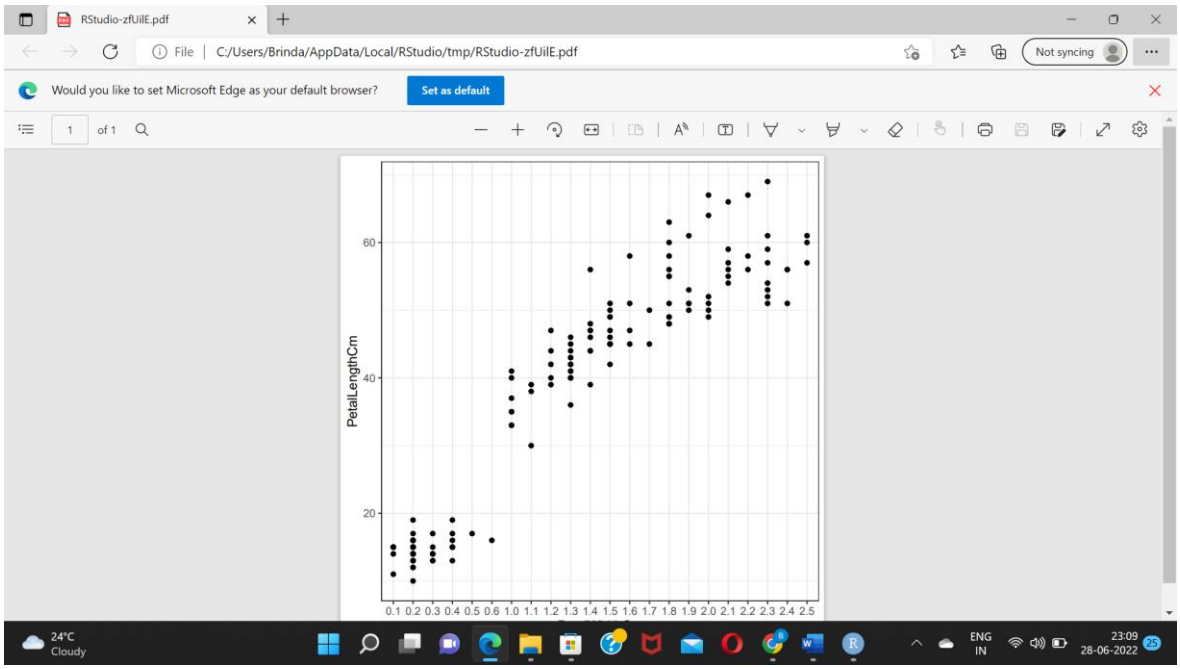
DOT PLOT



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 Statiscally significant linear relationship with petal Length.

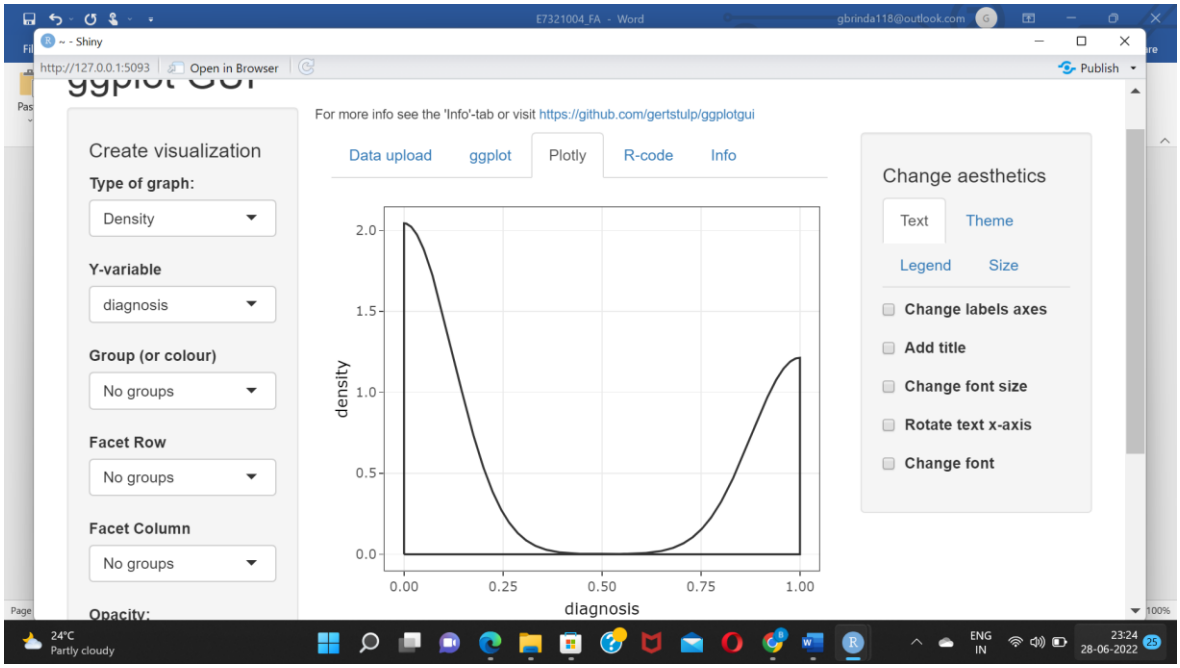


INTERPRETATION:

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:

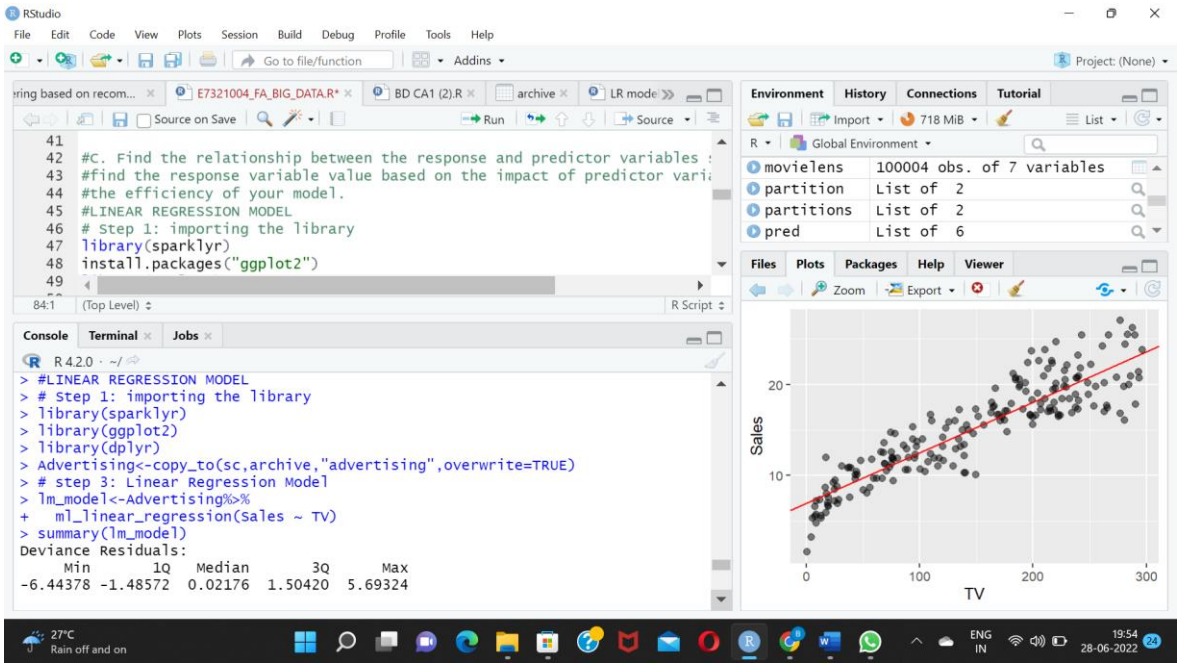
To Build a Binary Classification model on breast cancer dataset to analyse the data on which the patient has cancer or not.

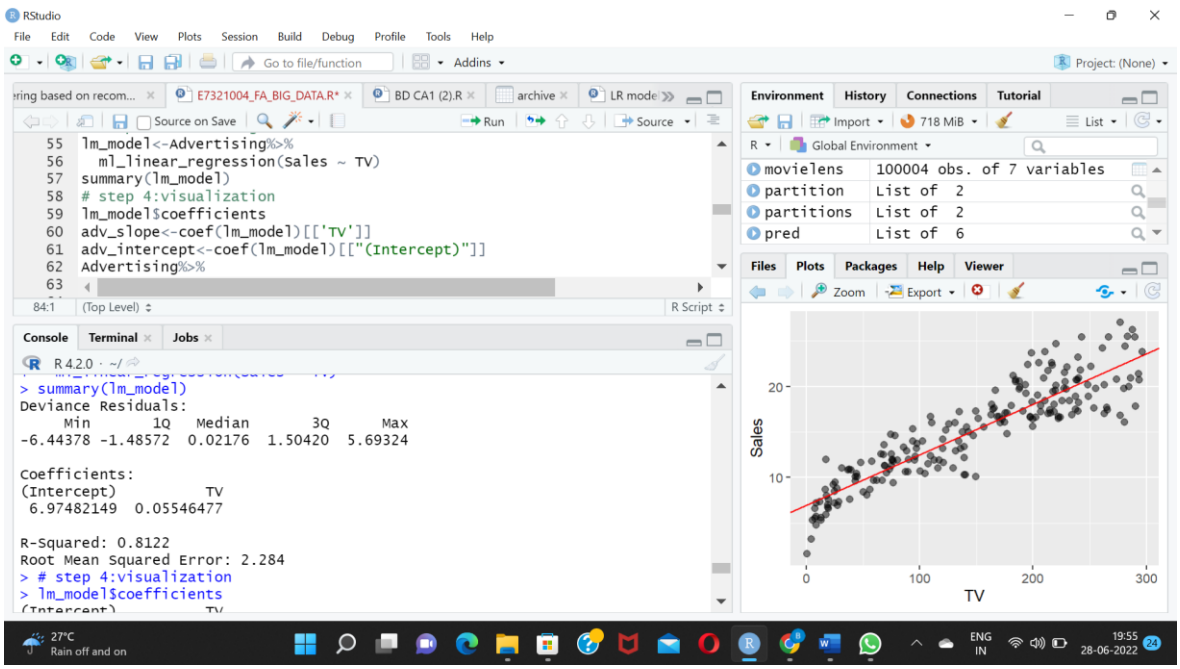


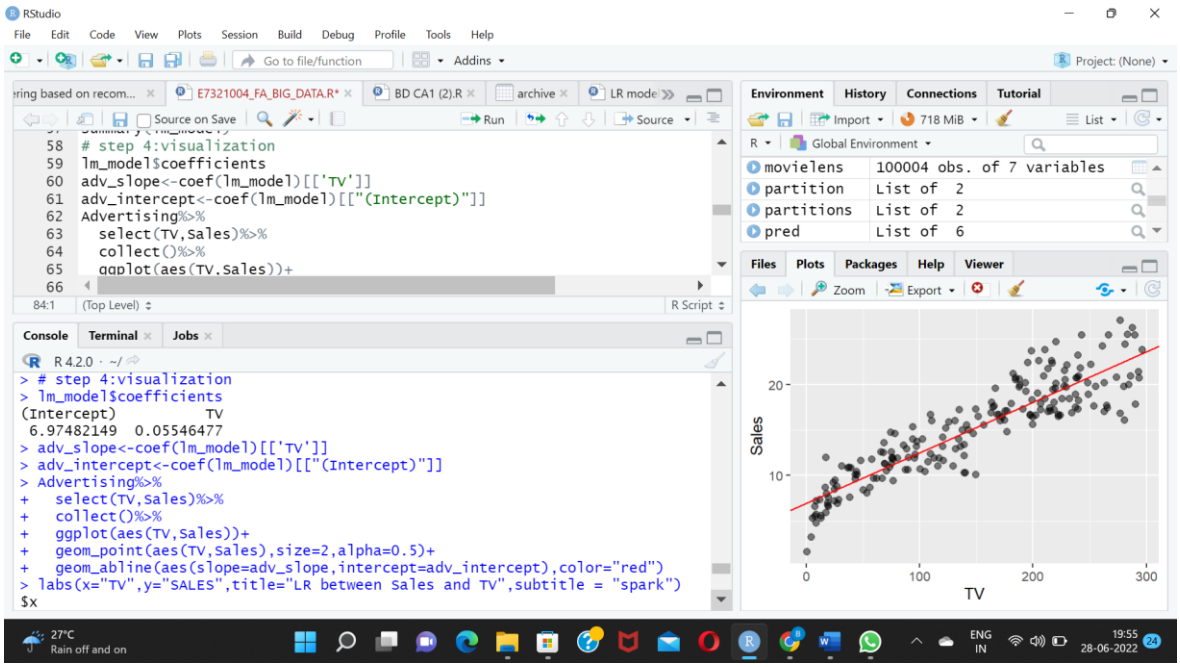
INTERPRETATION:

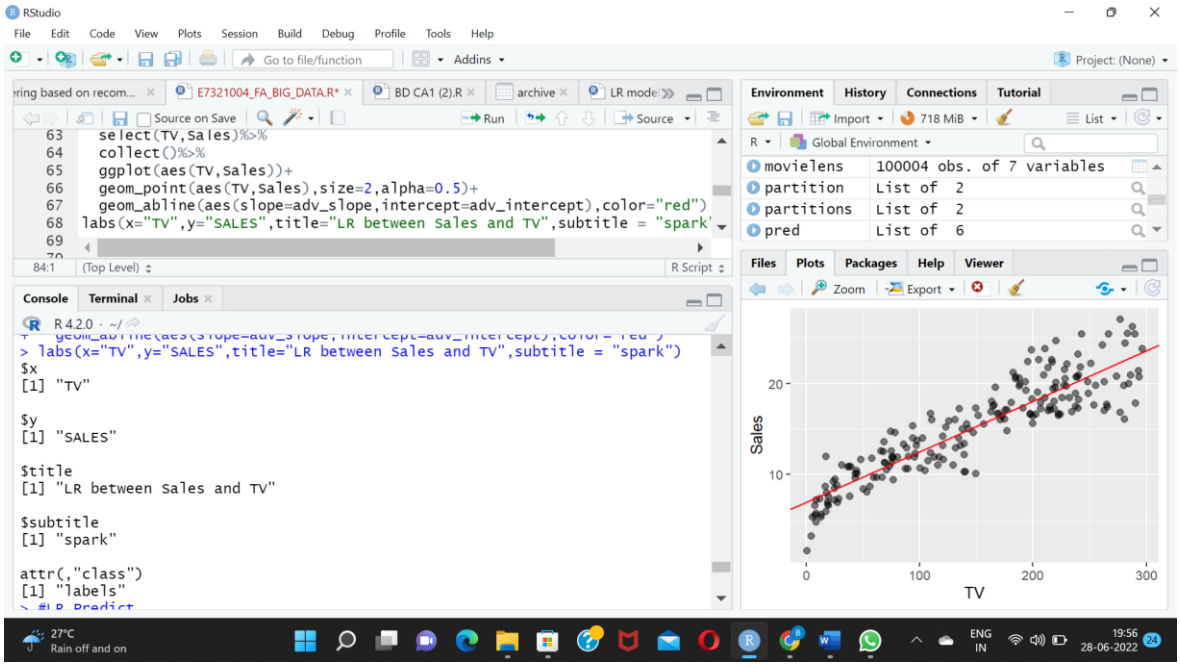
Using the Visualization we can interpret that Most of the patients are not affected by cancer (The value 2 indicates not having cancer).But still half of the patients are Affected by cancer.

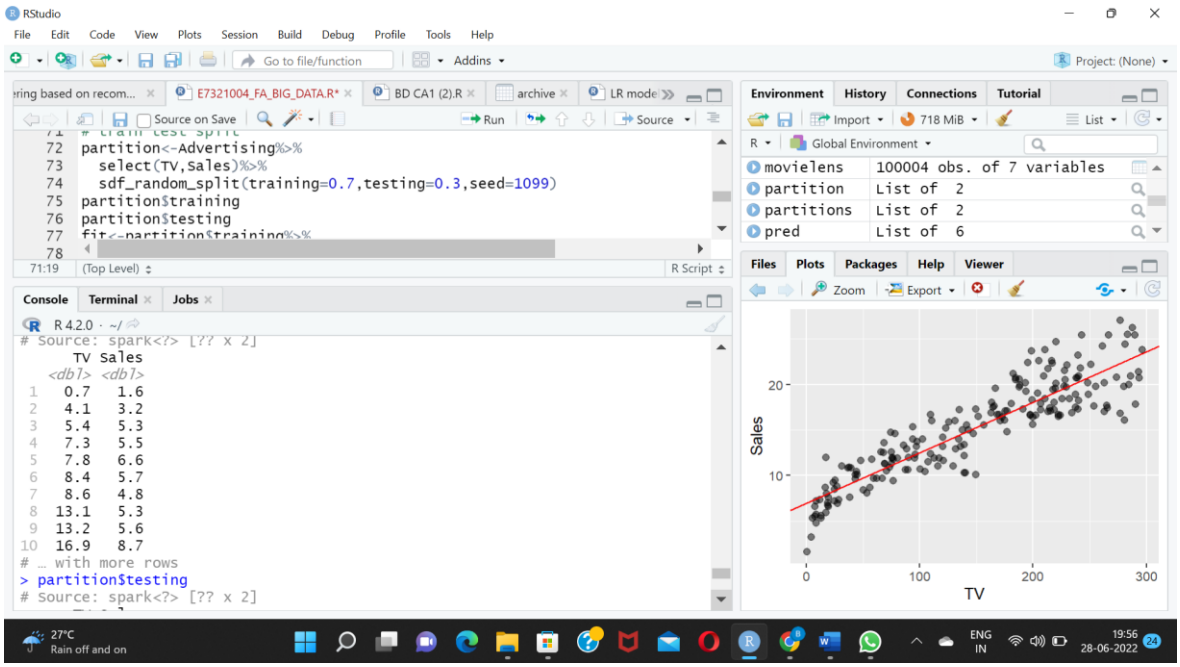
C. Find the relationship between the response and predictor variables statistically and find the response variable value based on the impact of predictor variables. Analyse the efficiency of your model.

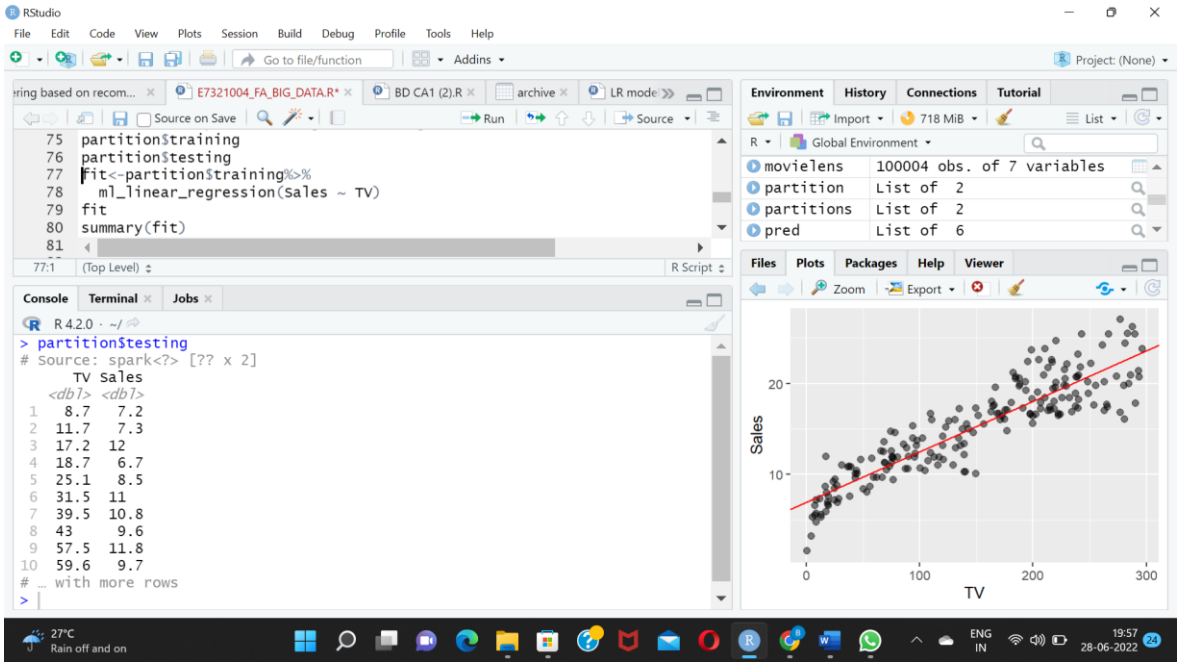


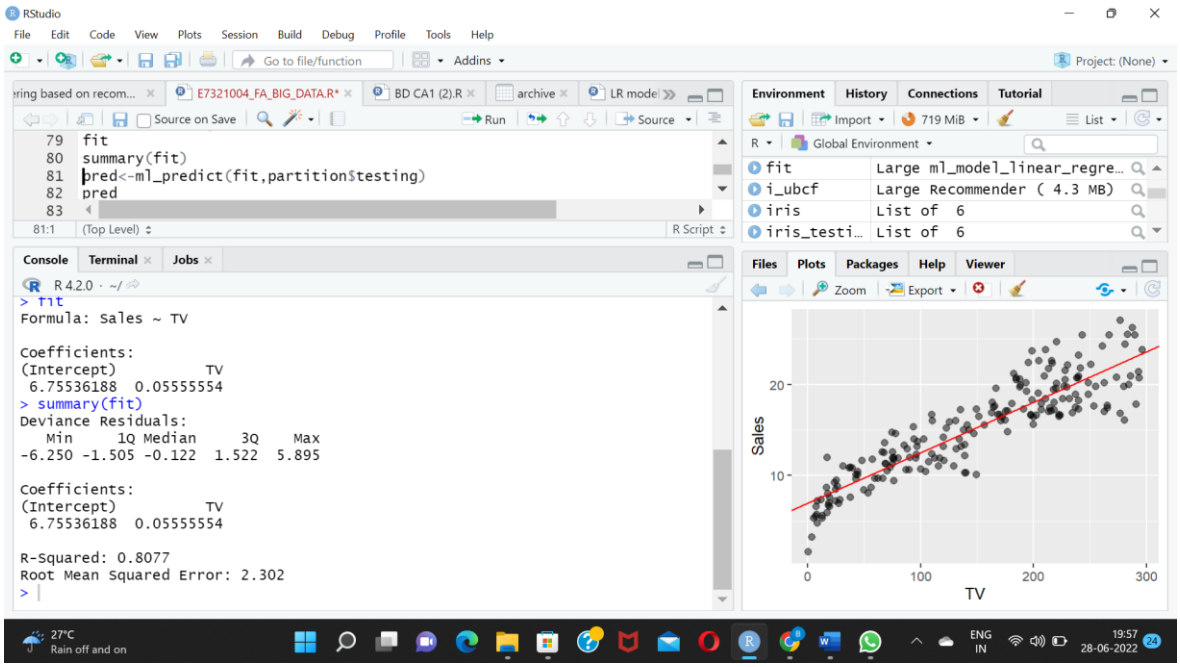


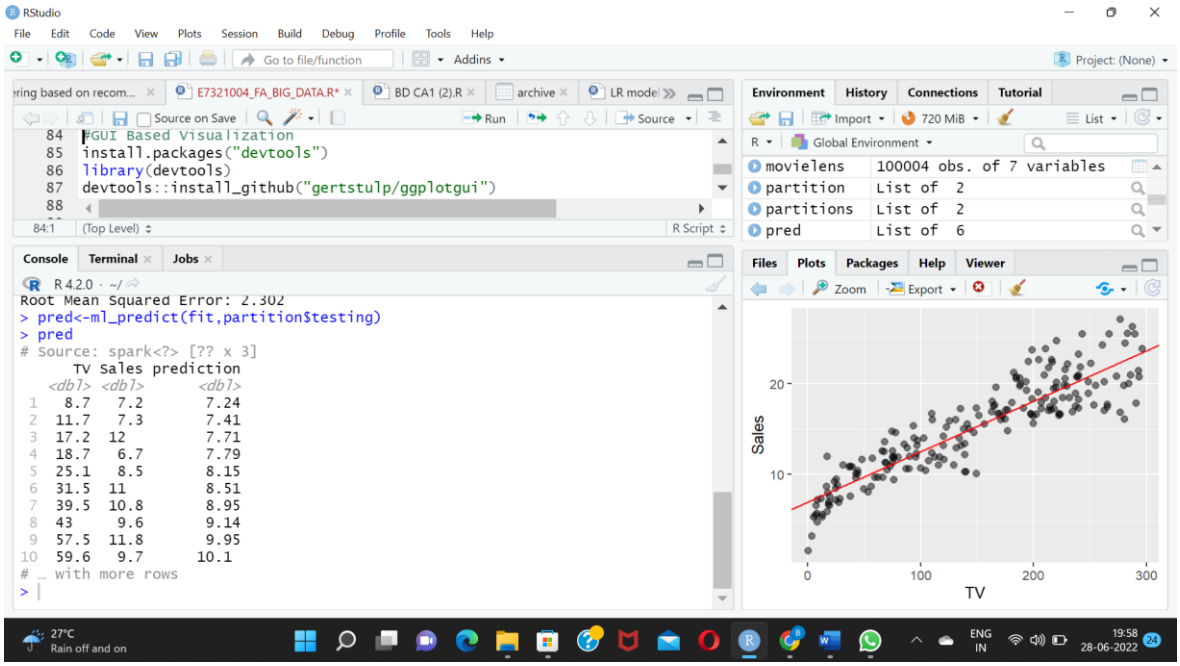

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
E7321004_FA_BIG_DATA.R*
BD CA1 (2).R
archive
LR mode
Source on Save
Run
Source
58 # step 4:visualization
59 lm_model$coefficients
60 adv_slope<-coef(lm_model)[['TV']]
61 adv_intercept<-coef(lm_model)[["(Intercept)"]]
62 Advertising%>%
63 select(TV,Sales)%>%
64 collect()%>%
65 ggplot(aes(TV,Sales))+
66
84:1 (Top Level) R Script
Console Terminal Jobs
R 4.2.0 · ~/
> # step 4:visualization
> lm_model$coefficients
(Intercept) TV
6.97482149 0.05546477
> adv_slope<-coef(lm_model)[['TV']]
> adv_intercept<-coef(lm_model)[["(Intercept)"]]
> Advertising%>%
+ select(TV,Sales)%>%
+ collect()%>%
+ ggplot(aes(TV,Sales))+
+ geom_point(aes(TV,Sales),size=2,alpha=0.5)+
+ geom_abline(aes(slope=adv_slope,intercept=adv_intercept),color="red")
> labs(x="TV",y="SALES",title="LR between Sales and TV",subtitle = "spark")
$x
Environment History Connections Tutorial
Import 718 MiB List
R Global Environment
movielens 100004 obs. of 7 variables
partition List of 2
partitions List of 2
pred List of 6
Files Plots Packages Help Viewer
Zoom Export
Sales
TV
0 100 200 300
10 20
27°C Rain off and on
ENG IN
19:55 28-06-2022


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
E7321004_FA_BIG_DATA.R*
BD CA1 (2).R
archive
LR mode
Source on Save
Run
Source
63 select(TV,Sales)%>%
64 collect()%>%
65 ggplot(aes(TV,Sales))+
66 geom_point(aes(TV,Sales),size=2,alpha=0.5)+
67 geom_abline(aes(slope=adv_slope,intercept=adv_intercept),color="red")
68 labs(x="TV",y="SALES",title="LR between Sales and TV",subtitle = "spark"
69
84:1 (Top Level) R Script
Console Terminal Jobs
R 4.2.0 · ~/
> labs(x="TV",y="SALES",title="LR between Sales and TV",subtitle = "spark")
$x
[1] "TV"
$y
[1] "SALES"
$title
[1] "LR between Sales and TV"
$subtitle
[1] "spark"
attr(,"class")
[1] "labels"
Environment History Connections Tutorial
Import 718 MiB List
R Global Environment
movielens 100004 obs. of 7 variables
partition List of 2
partitions List of 2
pred List of 6
Files Plots Packages Help Viewer
Zoom Export
Sales
TV
0 100 200 300
10 20
27°C Rain off and on
ENG IN
19:56 28-06-2022


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
E7321004_FA_BIG_DATA.R*
BD CA1 (2).R
archive
LR mode
Source on Save
Run
Source
# train test split
partition<-Advertising%>%
select(TV,Sales)%>%
sdf_random_split(training=0.7,testing=0.3,seed=1099)
partition$training
partition$testing
fit<-partition$training%>%
71:19
(Top Level)
R Script
Console
Terminal
Jobs
R 4.2.0 · ~/
Source: spark<?> [?? x 2]
TV Sales
<dbl> <dbl>
1 0.7 1.6
2 4.1 3.2
3 5.4 5.3
4 7.3 5.5
5 7.8 6.6
6 8.4 5.7
7 8.6 4.8
8 13.1 5.3
9 13.2 5.6
10 16.9 8.7
… with more rows
> partition$testing
Source: spark<?> [?? x 2]
Environment
History
Connections
Tutorial
Import
718 MiB
List
R
Global Environment
movielens 100004 obs. of 7 variables
partition List of 2
partitions List of 2
pred List of 6
Files
Plots
Packages
Help
Viewer
Zoom
Export
Sales
TV
0
100
200
300
10
20
27°C
Rain off and on
ENG
IN
19:56
28-06-2022


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
E7321004_FA_BIG_DATA.R*
BD CA1 (2).R
archive
LR mode
Source on Save
Run
Source
partition$training
partition$testing
fit<-partition$training%>%
ml_linear_regression(Sales ~ TV)
fit
summary(fit)
77:1
(Top Level)
R Script
Console
Terminal
Jobs
R 4.2.0 · ~/
> partition$testing
Source: spark<?> [?? x 2]
TV Sales
<dbl> <dbl>
1 8.7 7.2
2 11.7 7.3
3 17.2 12
4 18.7 6.7
5 25.1 8.5
6 31.5 11
7 39.5 10.8
8 43 9.6
9 57.5 11.8
10 59.6 9.7
… with more rows
Environment
History
Connections
Tutorial
Import
718 MiB
List
R
Global Environment
movielens 100004 obs. of 7 variables
partition List of 2
partitions List of 2
pred List of 6
Files
Plots
Packages
Help
Viewer
Zoom
Export
Sales
TV
0
100
200
300
10
20
27°C
Rain off and on
ENG
IN
19:57
28-06-2022

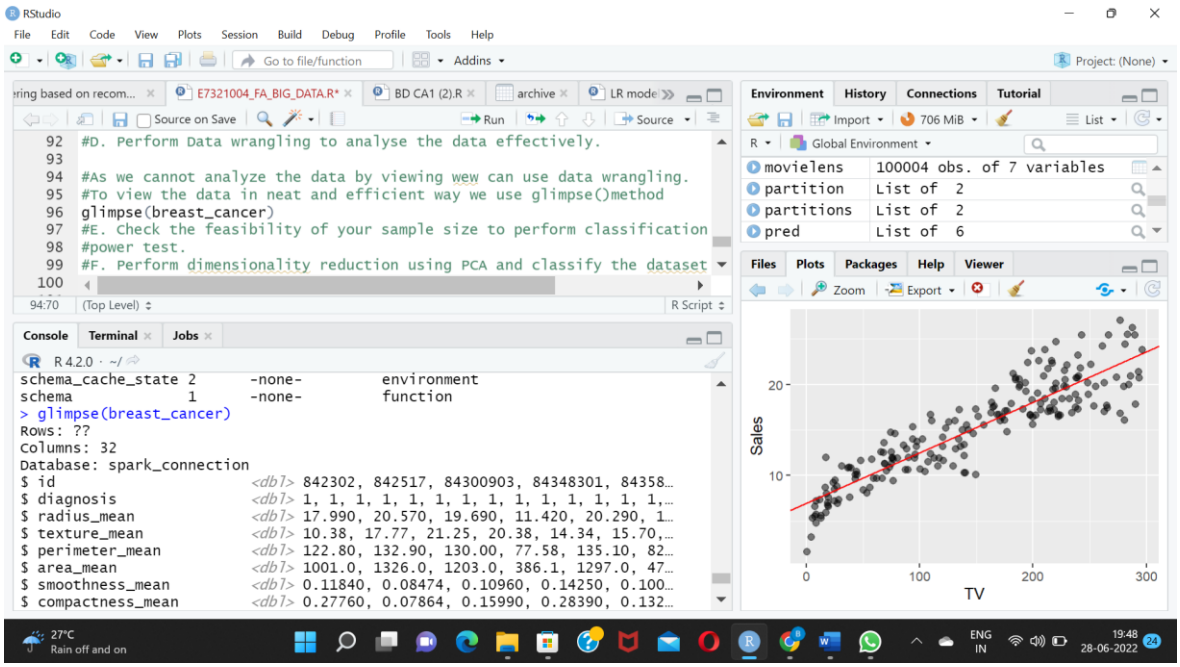




D. Perform Data wrangling to analyse the data effectively.

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

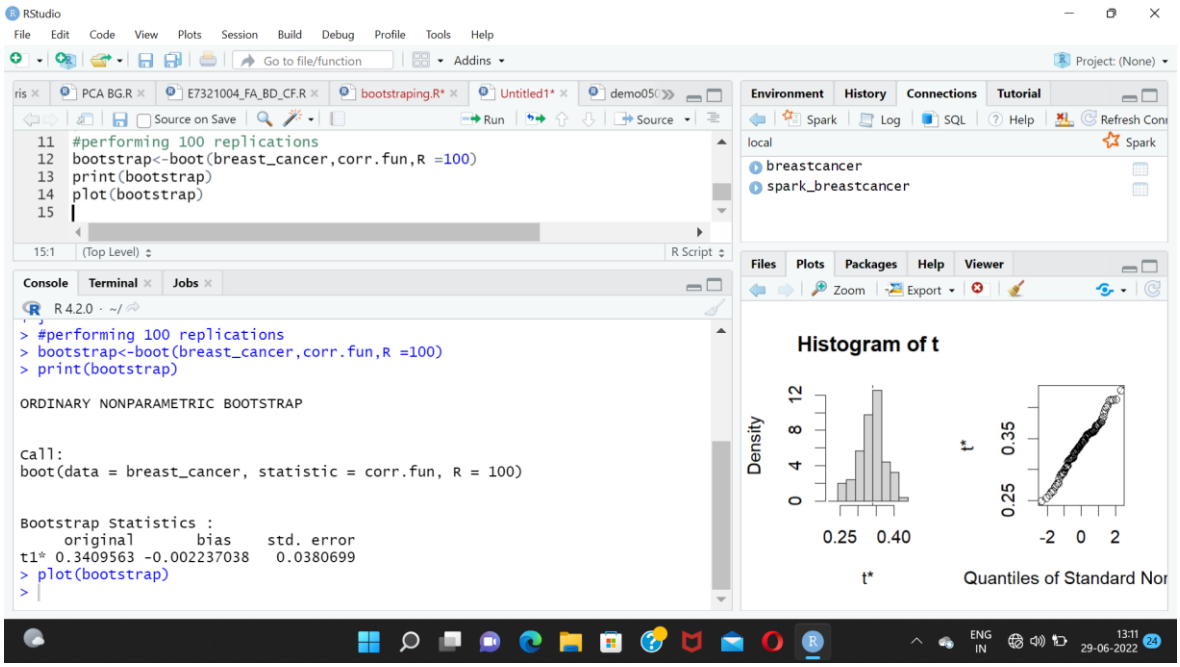
Using glimpse () we can read the data in a neat format.

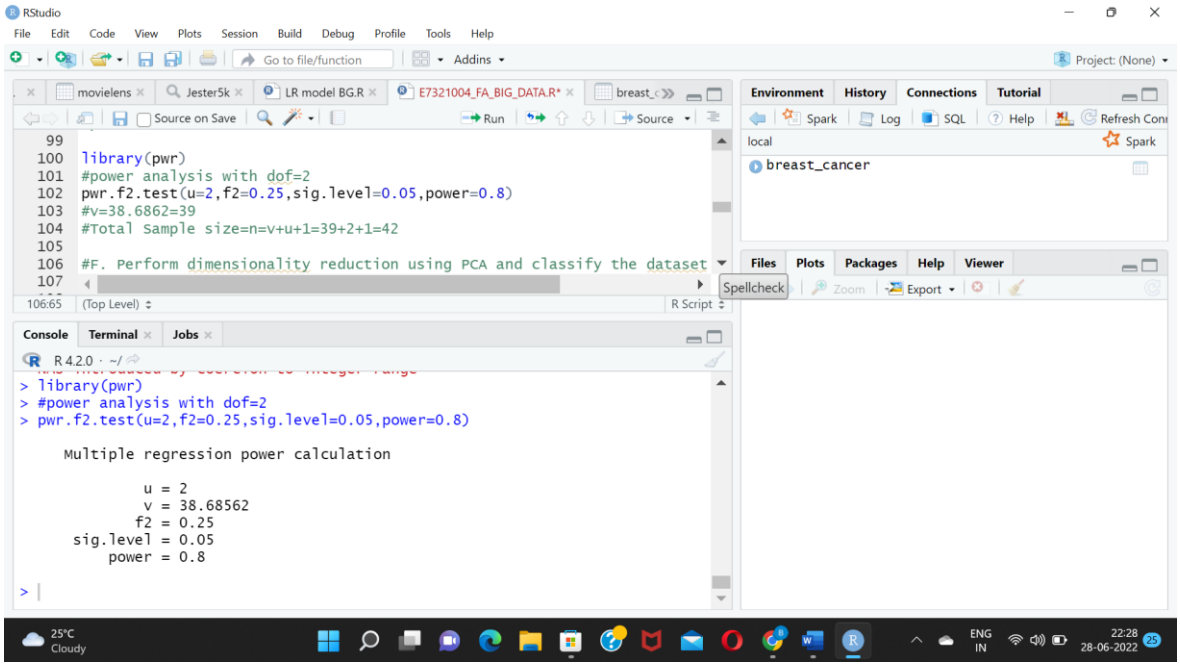


E. Check the feasibility of your sample size to perform classification using Bootstrap and power test.

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)
```



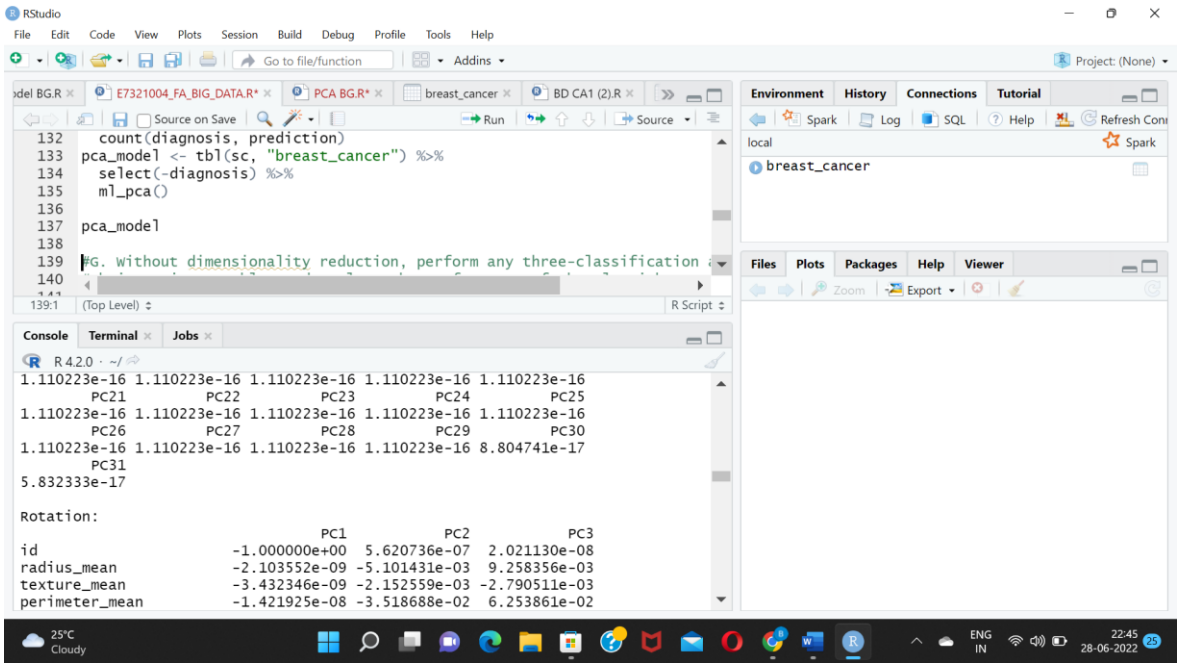


F. Perform dimensionality reduction using PCA and classify the dataset

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
```

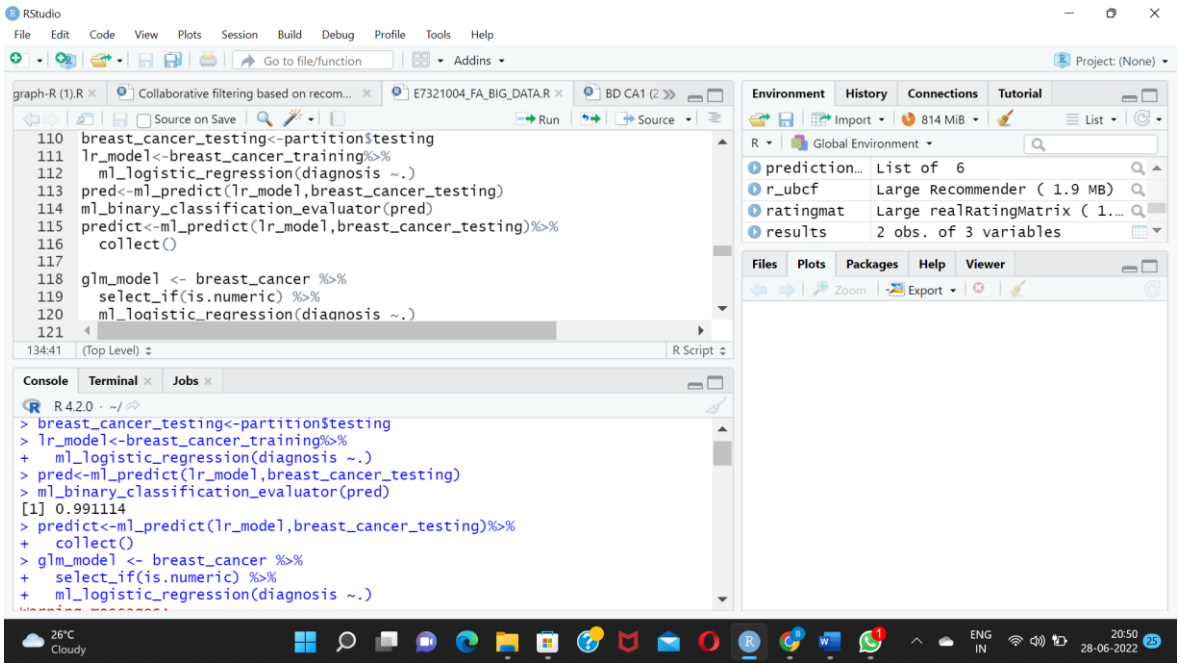


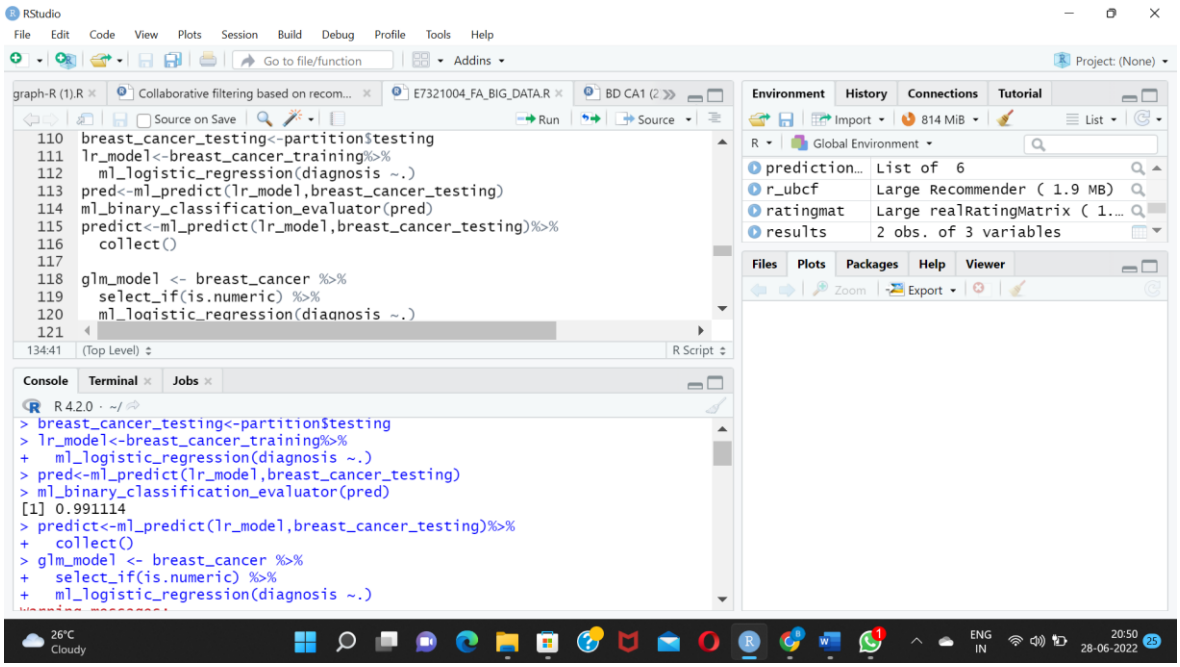
G. Without dimensionality reduction, perform any three-classification algorithm of your choice using sparklyr and analyse the performance of the algorithms

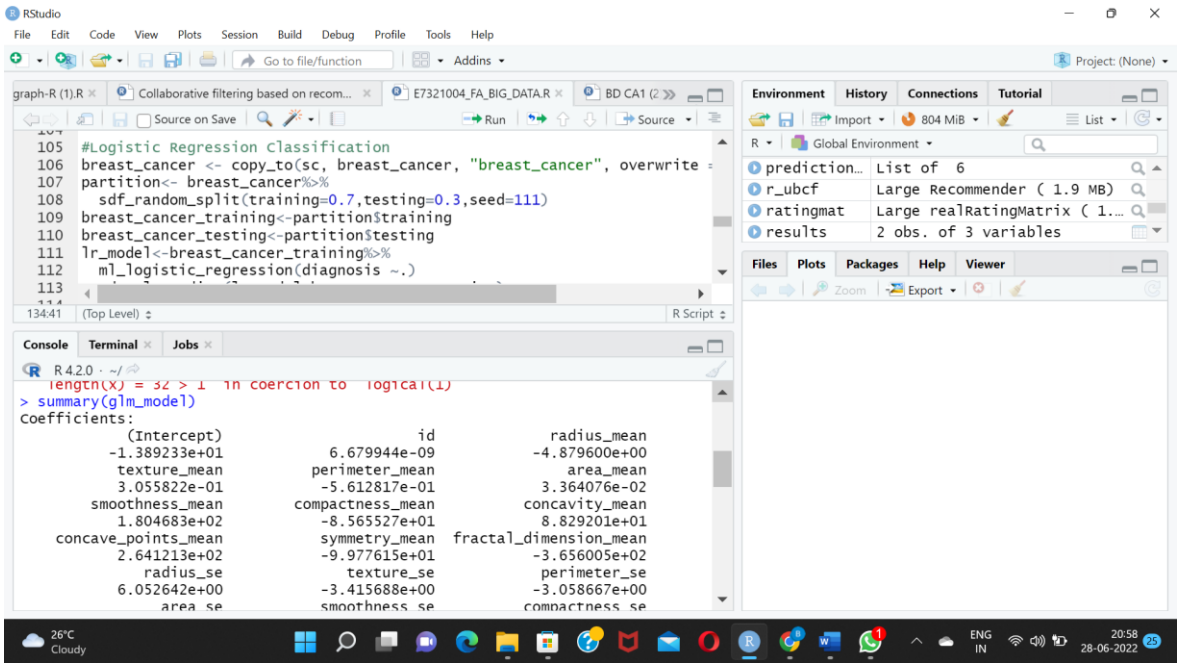
LOGISTIC REGRESSION :

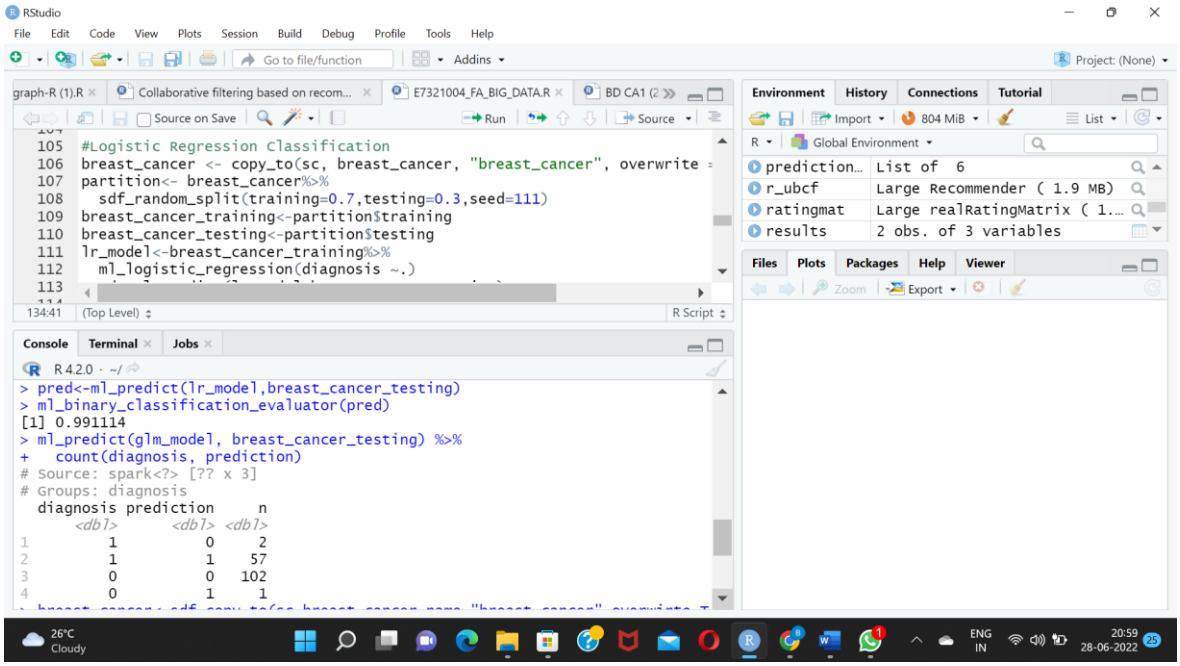
```
library(sparklyr)

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)
```









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%>%

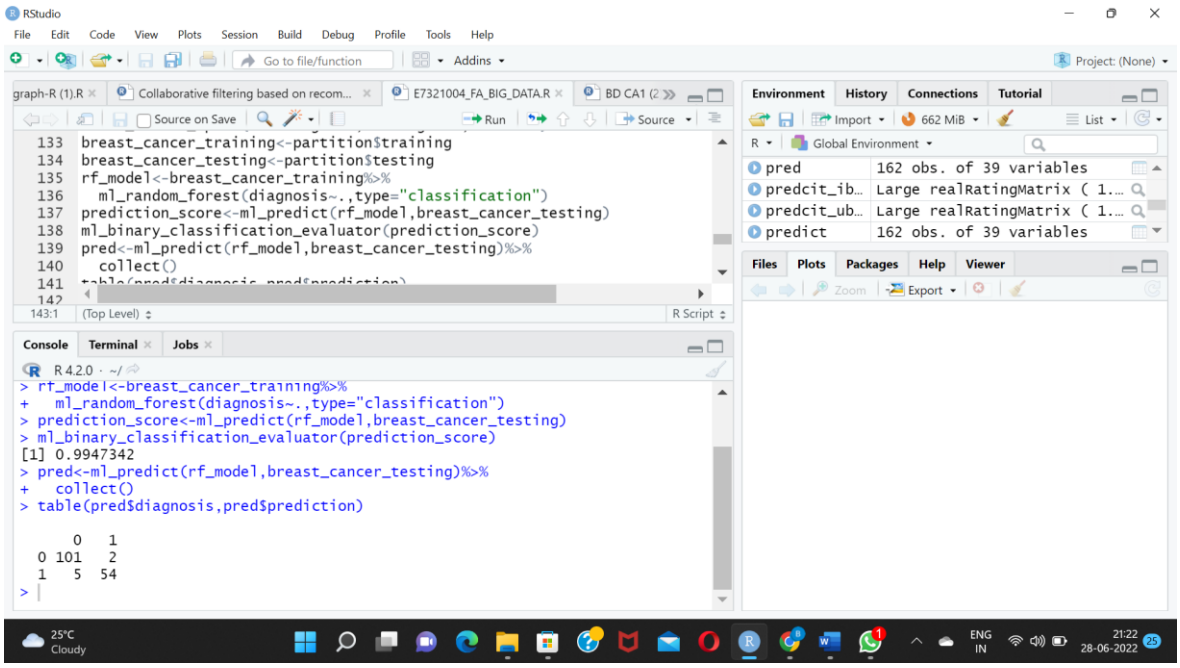
sdf_random_split(training=0.7,testing=0.3,seed=111)

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)
```

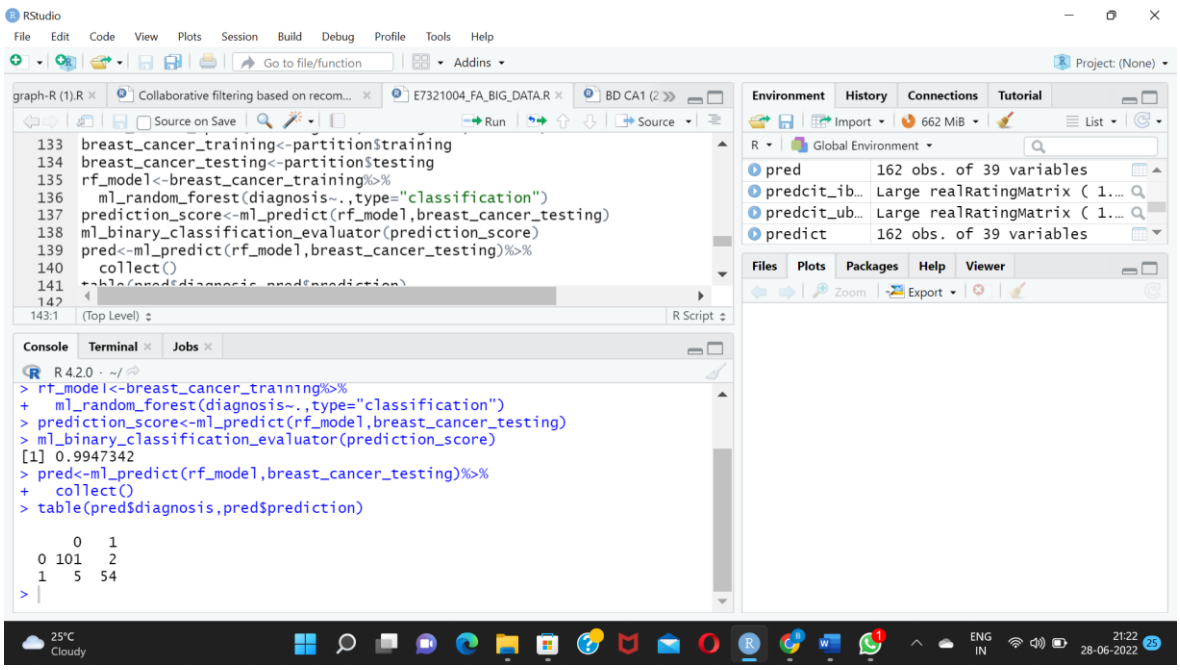


DECISION TREE CODE:

#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)
```



INTERPRETATION:

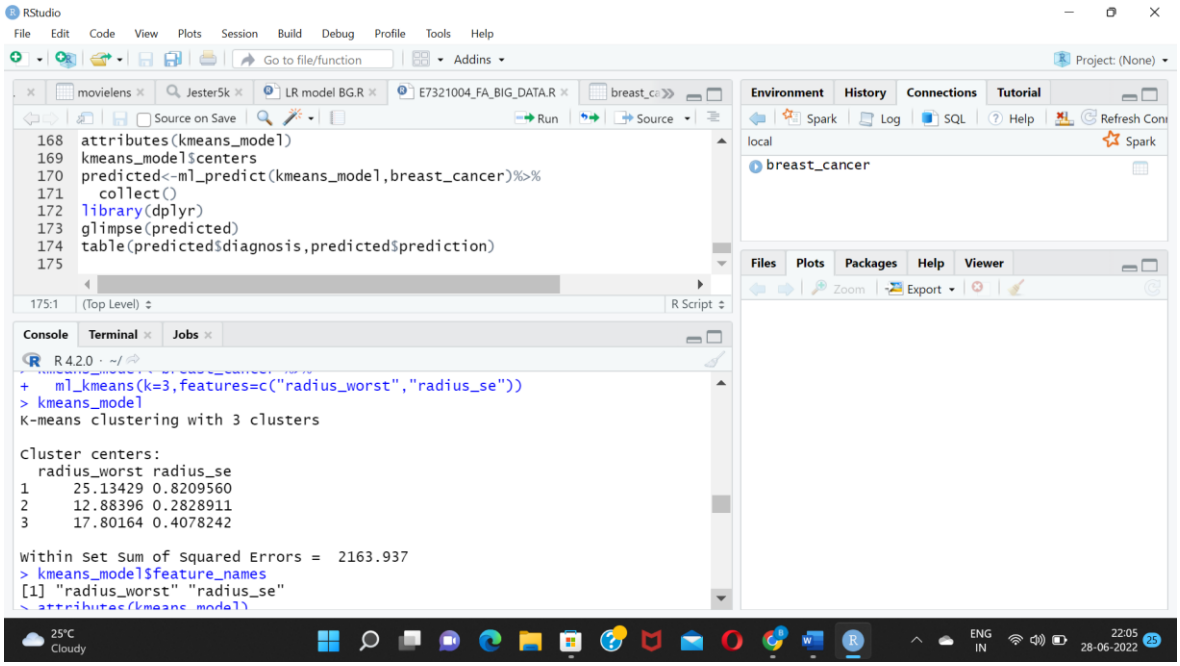
Among the three classification (Logistic Regression, Random forest and Decision Tree) . The Prediction Evaluation is Higher in **Decision Tree** classification(0.997) than the other classifications.

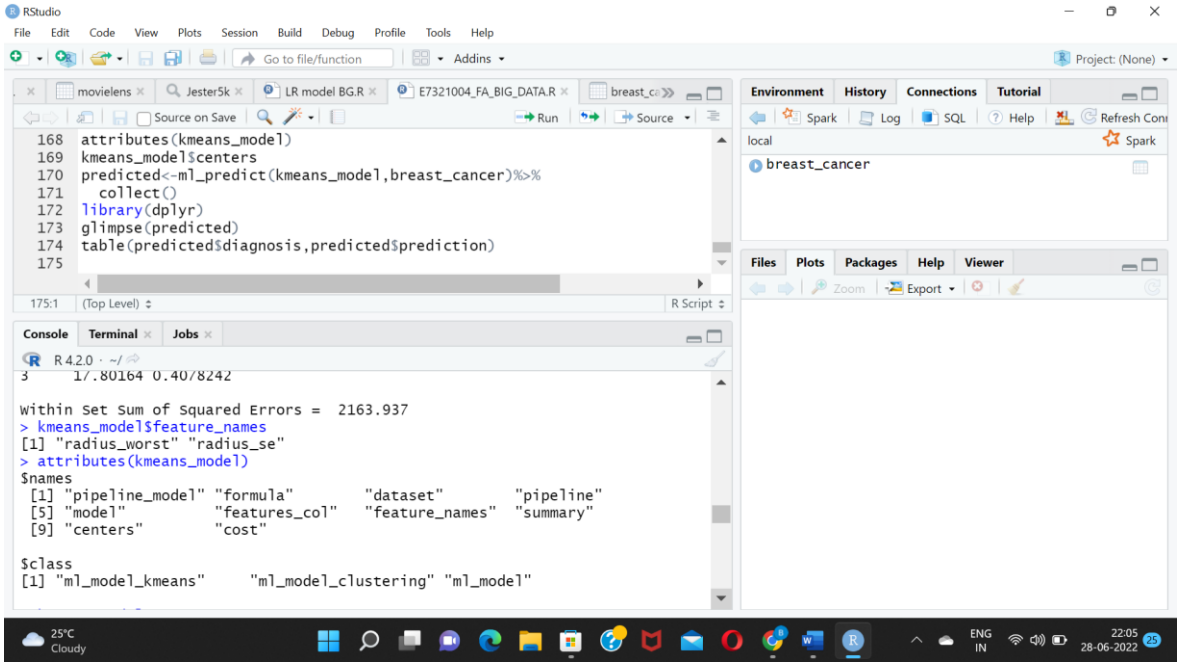
H. Remove the data label column in your dataset and apply clustering technique and interpret your findings

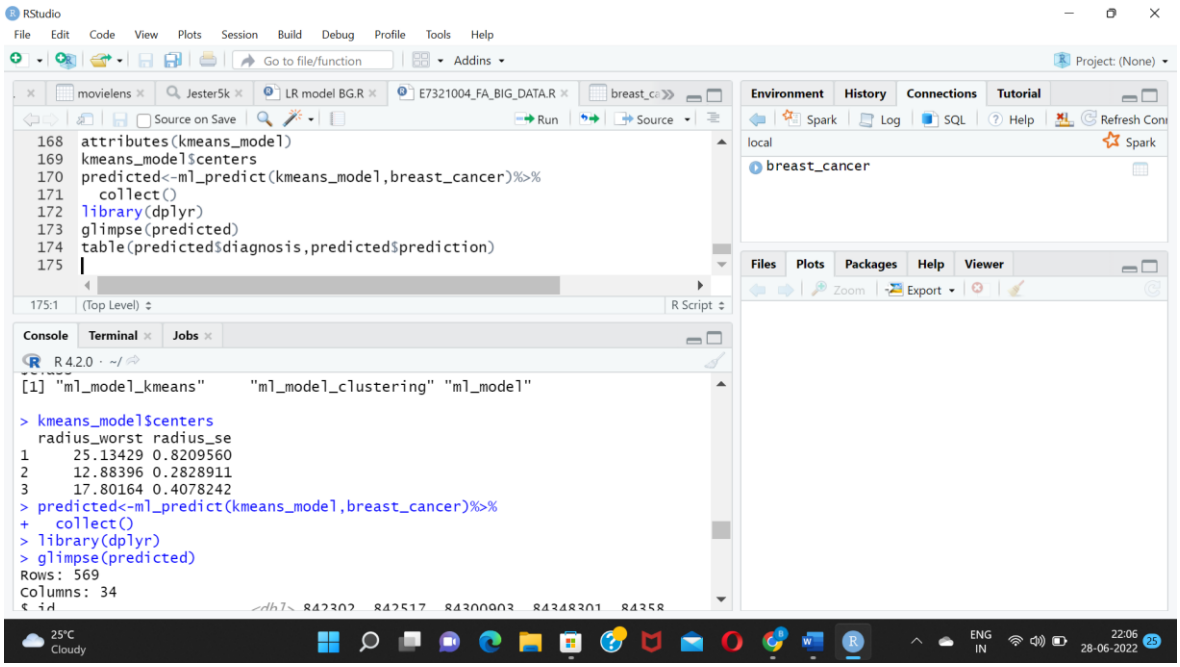
CODE:

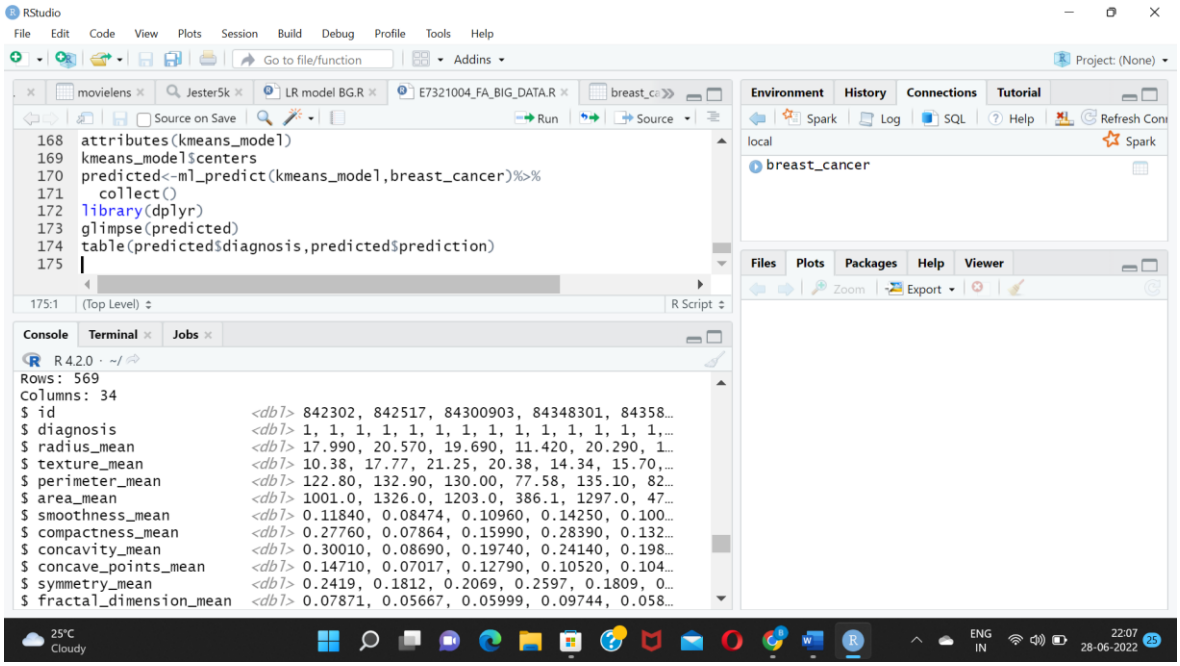
```
kmeans_model<-breast_cancer %>%
  ml_kmeans(k=3,features=c("radius_worst","radius_se"))

kmeans_model

kmeans_model$feature_names

attributes(kmeans_model)

kmeans_model$centers

predicted<-ml_predict(kmeans_model,breast_cancer)%>%
  collect()
```

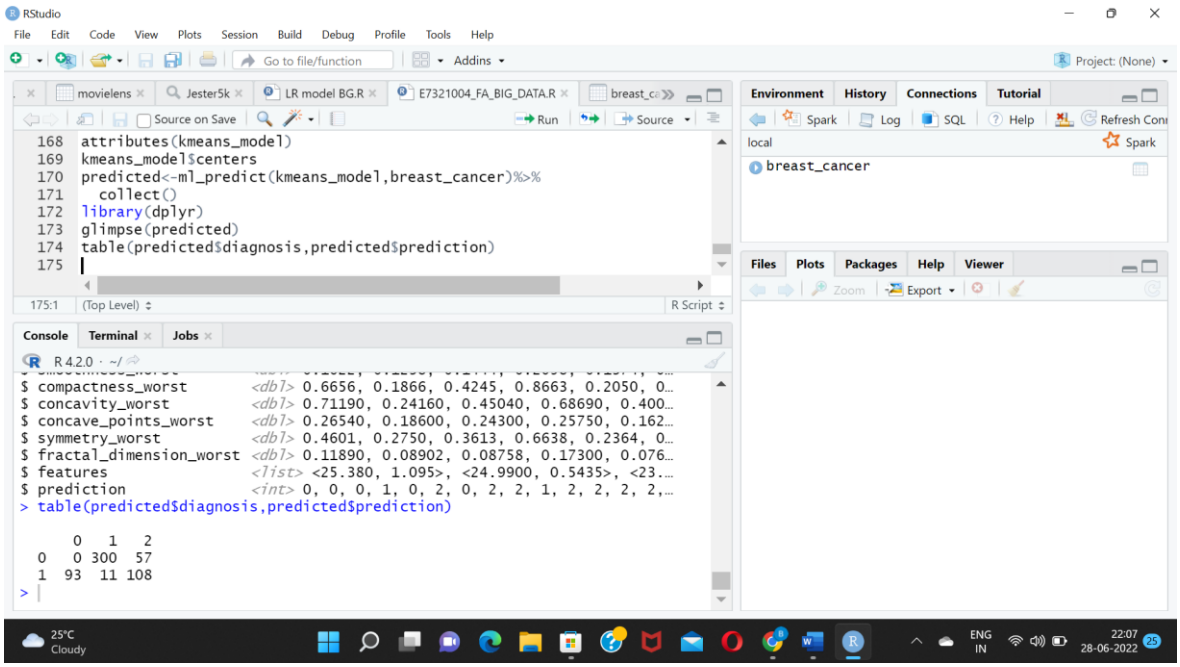
```
library(dplyr)
glimpse(predicted)
table(predicted$diagnosis,predicted$prediction)
```






RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
movielens
Jester5k
LR model BG.R
E7321004_FA_BIG_DATA.R
breast_ca
Source on Save
Run
Source
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$prediction)
175
175:1 (Top Level) R Script
Environment History Connections Tutorial
Spark Log SQL Help Refresh Con
local Spark
breast_cancer
Files Plots Packages Help Viewer
Zoom Export
Console Terminal Jobs
R 4.2.0 · ~/
[1] "ml_model_kmeans" "ml_model_clustering" "ml_model"
> kmeans_model$centers
radius_worst radius_se
1 25.13429 0.8209560
2 12.88396 0.2828911
3 17.80164 0.4078242
> predicted<-ml_predict(kmeans_model,breast_cancer)%>%
+ collect()
> library(dplyr)
> glimpse(predicted)
Rows: 569
Columns: 34
25°C Cloudy
ENG IN
22:06
28-06-2022


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
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
Environment History Connections Tutorial
breast_cancer
Files Plots Packages Help Viewer
Console Terminal Jobs
R 4.2.0 · ~/
Rows: 569
Columns: 34
$ id <dbl> 842302, 842517, 84300903, 84348301, 84358…
$ diagnosis <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
$ radius_mean <dbl> 17.990, 20.570, 19.690, 11.420, 20.290, 1…
$ texture_mean <dbl> 10.38, 17.77, 21.25, 20.38, 14.34, 15.70,…
$ perimeter_mean <dbl> 122.80, 132.90, 130.00, 77.58, 135.10, 82…
$ area_mean <dbl> 1001.0, 1326.0, 1203.0, 386.1, 1297.0, 47…
$ smoothness_mean <dbl> 0.11840, 0.08474, 0.10960, 0.14250, 0.100…
$ compactness_mean <dbl> 0.27760, 0.07864, 0.15990, 0.28390, 0.132…
$ concavity_mean <dbl> 0.30010, 0.08690, 0.19740, 0.24140, 0.198…
$ concave_points_mean <dbl> 0.14710, 0.07017, 0.12790, 0.10520, 0.104…
$ symmetry_mean <dbl> 0.2419, 0.1812, 0.2069, 0.2597, 0.1809, 0…
$ fractal_dimension_mean <dbl> 0.07871, 0.05667, 0.05999, 0.09744, 0.058…
25°C Cloudy
ENG IN
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 clutering like(The person is affected by cancer or not)