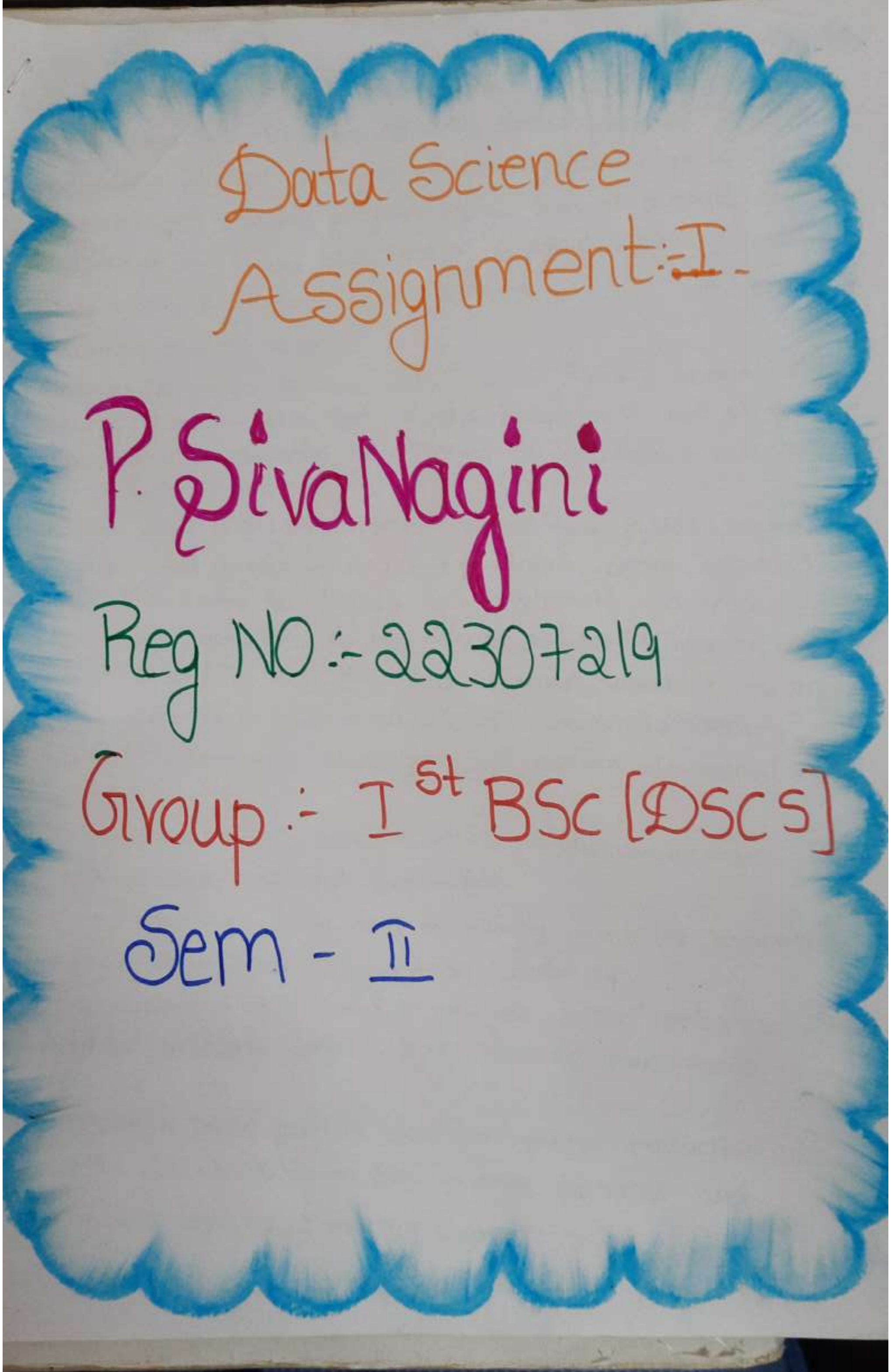


Data Science

Assignment:- I.

P. SivaNagini

Reg NO:- 22307219

Group:- Ist BSc [DSCS]

Sem - II

Define data science and applications of data science?

Ans- Data science:- It is the area of study which involves extracting knowledge from all the data. You can gather. Here we are transulating a business program into a research problem and then transulating the result back into a practical solution. People doing all this are called data scientists.

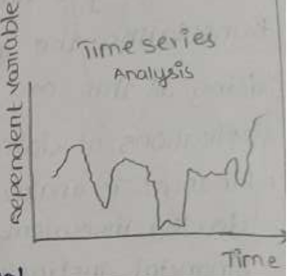
Applications of data science:-

1. Finance:- To analyze financial data, predict market trends, and develop invesment strategies. Banks, hedge funds, and other financial instructions rely heavily on data data science to make informed decisions.
2. Health care:- It is used to analyze patient data, predict disease outcomes and develop personalized treatment plants. also used in medical research to identify new treatments and cures.
3. Marketing:- Companies use data science to identify new markets, optimize pricing stategies and increase customer loyalty.
4. Education:- It is used in educational research to identify effective teaching strategies and improve educational policies.
5. Transportation:- It is used to improve safety reduce costs, and enhance customer experience.
6. Manufacturing:- It is used to identify areas for improvement, reduce downtime, and increase efficiently.
7. Government:- It is used to analyze public policy outcomes predict economic trends, and improve Government service.
8. Energy:- It is used to optimize energy production and predict distibution, predict energy demand and develop renewable energy sources.

9. Sports:- It is used to analyze player and team performance data, predict game outcomes, and develop game strategies.

10. E-commerce:- It is used to increase sales, improve customer loyalty, and optimize pricing ~~pr~~ strategies.

2. List the sources of data.

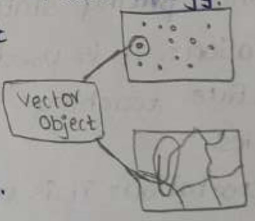


Time series data:- It refers to the data collected over a period of time, such as stock prices or weather conditions, allowing for analysis of patterns and trends.

Transition	Items
1	Biscuits
1	Cheese
1	Juice
2	Noodles

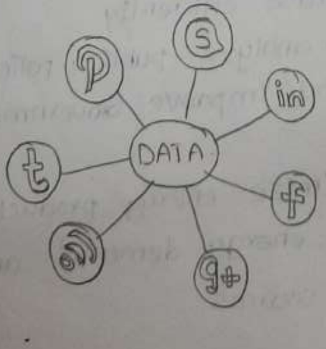
Transactional data:- This type of data records individual transactions, such as purchases or financial transactions, providing insights into customer behaviour and business operations.

Biological data:- It contains information related to living organisms, such as genetic sequences or physiological measurements. This data aids reasearch in areas like genetics medicine, and ecology.



Spatial data:- It contains information about physical locations and. This data helps in analysis and visualization of geographic patterns, such as maps or satellite imagery.

Social Network data:- It involves data about individuals and their relationships in a social network, offering insights into social interactions, influence, and community structures.



Write about Help function in R?

To access help function in R, we can use ? or help()

Example:- To get help about mean function in R

```
help (mean)
    or
  ?mean
```

Example-2:- To get help about rlm() in mass package

```
help (rlm, package = "MASS")
```

4. Write about Rstudio default display panes.?

A 1. source pane:- This is were you write your code.

2. console pane:- This is were output is displayed.

3. Environment R/history:- Here variables that are used in current program are displayed.

4. Files/ plots/ packages/ Help:- Here we can see the plots, files, Packages, etc.

5 How to use apply () on matrices in R. Explain with an example?

A

```
A = matrix(c(1,2,3,4))
apply (A, 2, mean)
                                        o/p:-
     [1]  [,2]
[1,]  1    2                            2    3
[2,]  3    4

apply (A, 1, mean)
        [,1]  [,2]                        o/p:-
  [1,]   1     2                          1.5    3.5
  [2,]   3     4
```

apply user define functions to matrices:-

```
f<- function (x)
{
  x/c(2,8)
}
A = matrix(c(1,2,3,4), nrow-2, ncol=2, by row=TRUE)
apply(A,1, F)
```

O/P:-

```
     [,1]  [,2]
[1,] 0.5   1.5
[2,] 0.25  0.5
```

5 a) Explain about the data science life cycle?

Step-1:- Define problem statement:- In meating with clients, data scientist must ask relevant mquestions to understand and define objectives of the problem that need to be tackled.

Step-2: Data collection:- If there is no data available, then you need to collect new data. This method is called Primary data collection.

Step-3:- Data preparation:- The most essential part of any data science project is data preparation. It consumes 60%. of the time spent on the project. steps in data preparation are.

1- Data cleaning: It handles missing values, NULL or unwanted values, duplicate values, misspelt attributes, inconsistent data types. Handling outliers: outliers are observations which are distant from the rest of the data. outliers can be good or bad for the data. outliers can be used for fraud detection. scatter plots and box plots help to identify the outliers in the data.

2. Data transformation :- turns raw data formats into desired outputs It also normalizes the data. Normalization is done in order to scale the data values in a specified range (-1.0 to 1.0) or (0.0to1.0)

For example:- consider the data below:

2001 pens 300   2002 pens 800

2001 pencils 400   2002 pencils 200

The data can be transformed as shown below. This table format helps us to summarise quickly

	2001	2002
Pens	300	800
pencils	400	200

3. Data integration:- when data from multiple sources are integrated the data after integration must be accurabte and reliable. Primary keys and foreign keys are handled while integrating data.

4. Data reduction:- Here we reduce the size of data by eliminating duplicate columns, unnecessary columns etc.

Step-4:- Data mining or Exploratory data analysis (EDA):- EDA helps us understand what we can actually do with the data. EDA helps understand the relationships between data and helps us in selecting the variables that will be used in model development. It also help us in identifying the right algorithm

Saftwares available:- tableau

Step-5:- Model building:- The model is built by selecting a machine learning algorithm that suits the data. Regression is used to predict continuous values.

Example:- Predicting house prices, temperature etc... and classification is used to predict discrete values.

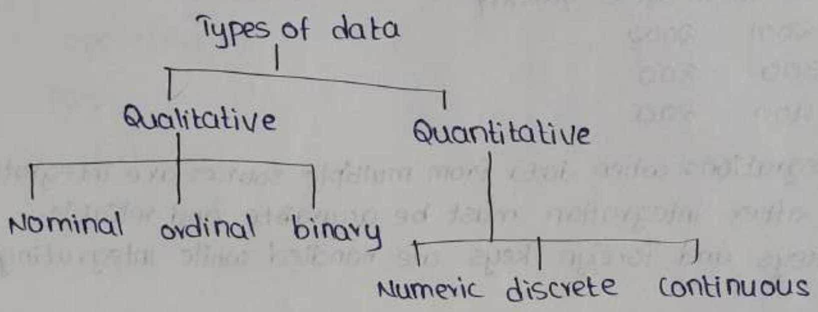
Example:- classifying whether the email is spam or not customer will buy a product or not.

For modeling the data, we can use the tools: R, python or SAS.

Step-6:- Visualization and communication :- the finding are communicated to business clients using simple and effective manner to convince the client. The visualization tools like tableau, power Bi, ~~Qu~~ ~~Qulf~~ Qlik view can be used to create powerful reports and dashboards.

* b) i) Explain about types of data?

A-



i) Qualitative data refers to categorical attributes (or) columns (or) features.

ii) Nominal attributes has no order [ranks, position]

Ex:- colours → Black, Brown, white.

iii) Attributes in binary data: It has only 2 values.

They are two types of binary data.

1. Symmetric :- Both values are equal import [Gender]
2. Asymmetric :- Both values are not equal important [result]

* ordinal attributes: It has meaning full order or ranking between them.

Ex:- Exam rank.

→ Quantitative data: It is related to number. It is measurable quantity.

1. Numeric data: It has divides into two types

a. Intervel b. Ratio

2. Discrete:- Discrete data have finite values it can be numerical and can also be in categorical form. These attributes has finite or countable infinite set of values

Ex:- Profession, zipcode

3. continuous:- continuous data have an infinite no. of states.

Ex:- Height, weight.

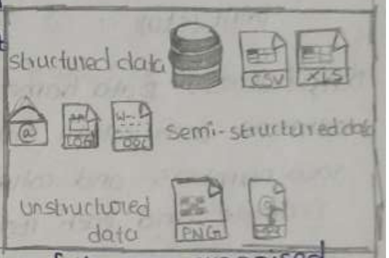
ii] Classification of digital data.

structured data:- It is created using a fixed Structure and is maintained in table format

Ex:- Relational data, SQL databases

Example for relational data:-

S.ID	S.Name	S.Address	S.EMail
1001	A	Delhi	A@gmail.com
1002	B	Mumbai	B@gmail.com



Unstructedured data:- The data in which does not follow any organised format.

Ex:- NO-SQL database, facebook posts, ~~twiets~~ tweets.

3 semistructured:- semi-structured data is information that doesn't have a strict and fixed format like a table. But still has some organization or tags to make it somewhat structured.

Ex:- Email, the mails are in structured format under in box, sent etc. But the text inside Email is not structured.

2 · Explain about data structures in the R language.

Data structure:- It is the way of organising data.

The most used data structure in R are.

1. vectors 2. lists 3. Dataframes 4. Matrices 5 Arrays 6 Factors

1 vector:- A vector is a collection of elements of mode or automic modes. character, integer, logical or Numeric.

vector is ordered collection.

Index starts from one.

Ex:- How to create a vector!

```
v=c(1,2,3)            O/P:-
  print (v)           [1] 1 2 3
```

lists:- In R the list is a container. A list is a special type of vector in which element can be different type.

```
Ex:-                                   O/P:-
    id = c(1,2,3)                   [[1]]
    name=c("siva","sri", chinni")     [1] 1 2 3
    stu = list(id,name)             [[2]]
    print (stu)                       [1] "siva", "sri", "chinni"
```

Dataframes:- Data frames is a 2D structure containing rows and columns. Each colum must have same number of items. It has row numbers and column names.

Example and each item must be of same type.

```
Ex:- data.frame (id,name)              O/P:-
       print (df)               rowno.   id   name
                                  1      1    Siva
                                  2      2    Sri
                                  3      3    chinni
```

Matrices:- Matrix is a 2-D structure containing rows & columns.

All the data must have be of same type.

Matrix will not give you id & name

Data frame will " " " " "

```
Ex:-  A=matrix (id,nrow=3, ncol=1)
      print(A)
                                        OUTPUT
                                          [,1]
                                        [1,] 1
                                        [2,] 2
                                        [3,] 3
```

Arrays:- Array can store more than 2 dimensions

Example an array of dimensions.

(2,4,3) creats 3 matrices. Each matrix has 2 rows & 4 columns.

```
Ex:-  A= array(c(1,2,3,4,5,6,7,8) dim =c(2,2,2))
      print (A)
```

o/p:-

```
,,1                          ,,2
   [,1] [,2]                    [,1] [,2]
[1,]  1    3                 [1,]  5    7
[2,]  2    4                 [2,]  6    8
```

Factors:- Factors categorize the data and store it as levels. They categorize unique values.

Ex:-

```
Fac = Factor(c(1,1,2,2,2))                    o/p:-
print(Fac)                                    [1] 1 1 2 2 2
                                              levels: 1 2
```

b. Explain about common vector operations ?.

1. vector addition:-

Ex:-

```
var1 = c(4,5)                                 o/p:-
var2 = c(2,4)                                 addition of two vectors
print("addition of two vectors")                    6 9
print (var1+var2)
```

vector subtraction:-

```
Ex:- print(var1÷var2)                         o/p
                                              2 1
```

vector multiplication

```
Ex! print(var1 * var2)                        o/p
                                              8 20
```

vector divison

```
Ex:- print(var1 /var2)                        o/p:-
                                              2.00  1.25
```

Modulous of two vectors:-

```
print (var1 %%.var2)                          o/p
                                              0 1  [remainders]
```

Floor division of two vectors

```
print (var1 %/% var2)                         o/p:
                                              2 1
```

Exponent of two vectors.

```
print(var1 ^ var2)                            o/p:-
                                              6  625
```

who develop trageted advertising compaigns. companies use data science to identify new markets. or optimize pricing strategies, and increase coustomer loyalty.

9. a) Explain about various functions applied on matrix rows & columns?

A Addition of matrices:-

```
A = matrix (c(3,5,4,6,7,8,9,3,5), nrow=3, ncol=3, byrow = TRUE)
print (A)
B= matrix (c(5,6,7,2,4,3,6,9,7), nrow=3, ncol=3, by row = TRUE)
print(B)
print(A+B)
```

O/P:-

```
      [,1] [,2] [,3]        [,1] [,2] [,3]           [,1]  [,2]  [,3]
[1,]  3    5    4     [1,]  5    6    7        [1,]  8     11    11
[2,]  6    7    8     [2,]  2    4    3        [2,]  8     11    11
[3,]  9    3    5     [3,]  6    9    7        [3,]  15    12    12
```

subtraction:-

```
print (A-B)
```

O/P:-

```
      [,1] [,2] [,3]
[1,]  -2   -1   -3
[2,]  4    3    5
[3,]  3    -6   -2
```

Multiplication.

```
print(A*B)
```

O/P:-

```
      [,1] [,2] [,3]
[1,]  15   30   28
[2,]  12   28   24
[3,]  54   27   35
```

Division:-

```
print(A/B)
```

O/P:-

```
      [,1]  [,2]  [,3]
[1,]  0.6   0.83  0.57
[2,]  3.0   1.75  2.66
[3,]  1.5   0.33  0.71
```

Applying functions to matrix rows & columns:-

```
apply (A,2,mean)

    [,1]  [,2]                o/p:-
[1,]  1     2
[2,]  3   4                   2    3
```

```
apply (A,1,mean)
      [,1] [,2]               o/p:-
 [1,]   1   2
 [2,]   3   4                1.5   3.5
```

apply user define functions to matrices:-

```
f← function (x)
{
  x/c (2,8)
}
A=matrix (c(1,2,3,4),nrow=2, ncol=2, byrow=TRUE)
 apply (A,1,F)
 o/p:-
      [,1]  [,2]
   [1,] 0.5  1.5
   [2,]0.25  0.5
```

b] What is Data frame? Explain procedure to create Data frame with an example?

A. Data frame :- A data frame is a special type of list where every moment of list has the same length. Data frame is a tabular data type.

Row concatenation :-

```
print(rbind(A,B))
```

o/p:-

```
     [,1] [,2] [,3]
[1,]  3    5    4
[2,]  6    7    8
[3,]  9    3    5
[4,]  5    6    7
[5,]  2    4    3
[6,]  6    9    7
```

Deleting second column:-

```
A=matrix(c(c(3,5,4,6,7,8,9,3,5),nrow=3,
          ncol=3, byrow=TRUE)
print("Before deleting second column")
print(A)
A=A[,-2]
print("After deleting second column")
print(A)
```

o/p:-

~~o/p:~~

```
[1] "Befor deleting 2nd column
     [,1] [,2] [,3]
[1,]  3    5    4
[2,]  6    7    8
[3,]  9    3    5
```

```
[1] after deleting 2nd column
     [,1] [,2]
[1,]  3    4
[2,]  6    8
[3,]  9    5
```

updating second row:-

```
A-matrix(c(c(3,5,4,6,7,8,9,3,5),nrow=3, ncol=3, byrow=TRUE)
print ("Before updating Second sec row")
print(A)
A[2,] = c(11,12,15)
print ("After updating second row")
print(A)
```

o/p:-

```
[1]"Before updating second row
     [,1] [,2] [,3]
[1,]  3    5    4
[2,]  6    7    8
[3,]  9    3    5
```

```
[1]"After updating second row
     [,1] [,2] [,3]
[1,]  3    5    4
[2,]  11   12   15
[3,]  9    3    5
```

Creating a data frame:-

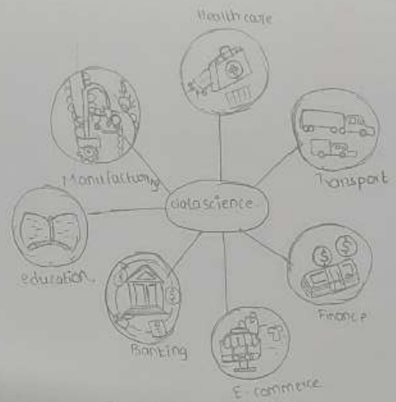
using data.frame () function

```
name = c("siva", "chinni")
age  = c(18, 20)
df = data.frame(Student name =name, Student age =age)
print(df)
```

O|P:-

```
   Studentname    Studentage
1    Siva            18
2    chinni          20
```

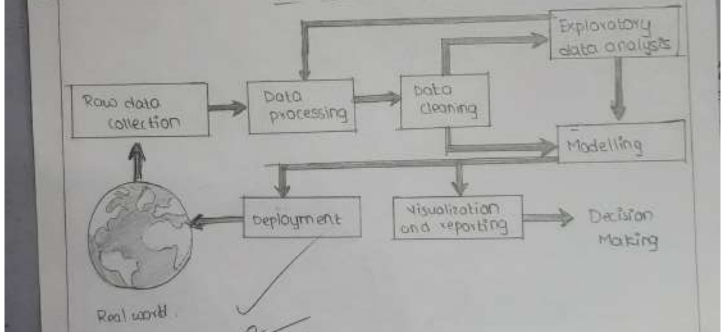
data frame will give row numbers automatically.



Data science applications.

6A.

Data science life cycle.



14/07/2023