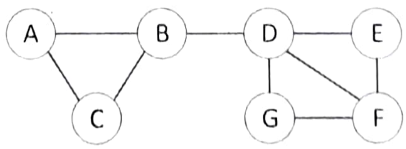
VIT Vidyalankar Institute of Technology Accredited A+ by NAAC (Autonomous College Affiliated to University of Mumbai)	End Semester Examination (CBCGS-C scheme) -(2022-23)	
Date: 17 Nov 2022	Branch: Computer Engineering	Time: 2 Hr.
Semester: VII	Subject: Big Data Analytics	Marks: 50

N B. :- All Questions are Compulsory

		CO	BL
Q.1)	Attempt any Five (2 Marks Each)		
a)	What is CAP theorem?	CO3	L1
b)	Explain any 2 characteristics of NoSQL database.	CO3	L2
c)	What is Sliding window sampling technique for big data.	CO4	L1
d)	List components of DSMS (Data Stream Management System) architecture.	CO4	L2
e)	Give any 4 examples of social network graph	CO5	L2
f)	Define social network and Social Network mining.	CO5	L1
g)	List any 2 basic features of R	CO6	L2
h)	Explain any 2 functions that allow users to handle data in R.	CO6	L1
Q.2)	Attempt any two. (5 Marks Each)		
a)	Compare NoSQL database with RDBMS.	CO3	L1
b)	Explain Key-Value store architecture pattern with example.	CO3	L2
c)	Explain Column family store architecture pattern.	CO3	L2
Q 3)	Attempt any One (10 Marks Each)		
a)	How bloom filter is useful for big data analytics? Explain with one example	CO4	L3
b)	Explain Flajolet-Martin (FM) algorithm. Using FM algorithm, compute number of distinct elements in the set S={1,3,2,1,2,3,4,3,1,2,3,1} with hash function h(x)=(6x+1) mod 5.	CO4	L2
Q 4)	Attempt any One (10 Marks Each)		
a)	Compare Content based filtering and Collaborative filtering in Recommendation System.	CO5	L4
b)	Explain the Detect communities in the below given social graph using Girvan-Newman (GN) algorithm 	CO5	L3
Q 5)	Attempt any One. (10 Marks Each)		
a)	What are the different data structures in R. Explain with example.	CO6	L1
b)	List and discuss different types of data visualizations techniques.	CO6	L1

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

START WRITING HERE

Space for Marks	Question No.	

1

a) In theoretical computer science, the CAP theorem states that any distributed database system can only provide 2 of the following 3:

i) Consistency
ii) Availability
iii) Partition tolerance.

In other words, if ~~a~~ network of distributed ~~system~~ data base system fails then it can either provide Consistency or Availability and not both.

b) 2 characteristics of NoSQL:

i) Volume :- NoSQL databases are highly scalable which makes them suitable for large amount of data.

ii) Variability :- NoSQL databases accomodate heterogeneous data. They ~~a~~ do not require data in any structure and hence are flexible in terms of their usage and reliability.

c) In sliding window sampling technique ~~is~~ ~~used~~, a window of specified length moves over the data, statistical analysis is performed over this window. ~~Ac~~ This window captures recent N entries of data and is used by

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

several algorithms like DGIM to detect number of '1's in the window.

d) Components of DSMS are.
1) Input Stream Data
2) Working Storage
3) Archieve Storage
4) Query processor
5) Repository of Standing queries.
6) output Buffer.
7) o/p stream. ~~of data~~.

e) Examples of social network graph are
i) Facebook friend graph
ii) Twitter follower graph.
iii) LinkedIn ~~conta~~ connect graph
iv) Publication co-author graph.
v) WhatsApp contact graph.
etc...

f) Social Network is a dedicated platform (website or application) which enables users to have social interactions with each other through posts, messages, comments, reactions, images, etc...
Social network mining is the process of obtaining information from user-generated big data on social media platforms.

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

g) R has many features such as:

i) It is open source

ii) It has powerful graphics

iii) It has effective data handling & storage facilities.

iv) It can perform complex statistical calculations on ~~b~~ big data.

etc...

h) R has facility to read different types of files. Functions available include.

i) read.table () : To read text files

ii) read.csv () : To import ~~from~~ CSV file

iii) read.xls () : — " — xls file.

etc..

It also supports directory operations like.

(i) getwd () (ii) setwd()

And to ~~read~~ list files of directory it has list.files () function.

~~i)~~

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE
	2	
	(a)	

No SQL	RDBMS
1) Designed to handle lage volume of data	Designed to handle moderate volume of data
2) It handles all types of data (structured, semi structured and un-structured)	It is designed to handle only structured data (set of relations) ~~only~~
3) It is designed to support analytical operations & simple transactions	It is designed to support complex transactions & perform simple analysis.
4) It supports BASE property	It supports ACID property.
5) It is designed to ~~support~~ handle data coming from many sources and at high velocity	It is designed to handle data ~~from~~ coming from one or few sources and at low velocity.
6) Does not support changes in database	Supports frequent updates to database
7) Ex: Amazon Dynamo, MongoDB.	Ex: Oracle 8i, PostgreSQL.

Total Marks of Question no.		
		Examiner
		Moderator
		Re-Assessor

Space for Marks	Question No.	START WRITING HERE
2	b)	~~Key~~ Key-value store is the most simple ~~and a~~ NoSQL architechure pattern. In this model, a simple string or a collection of values or a complex entity are all treated as values. These values are identified by a single key such that each key uniquely identifies its value. Although same key can associate to a set of values.

Databases designed on key-value store pattern, use an efficient and compact indexing or hashing structure ~~to~~ for the key. This enables rapid search of values associated with the key.

Key → Value

such as

empid → 1

ename → xyz

city → Mumbai

Qualification → {SSC, HSC, BE, ME}

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

Advantages of Key-value store.

→ Easy to use due to simplicity

→ Efficient (quick response time)

→ Scalable (handles large volume)

→ Reliable

Drawbacks of Key-value store

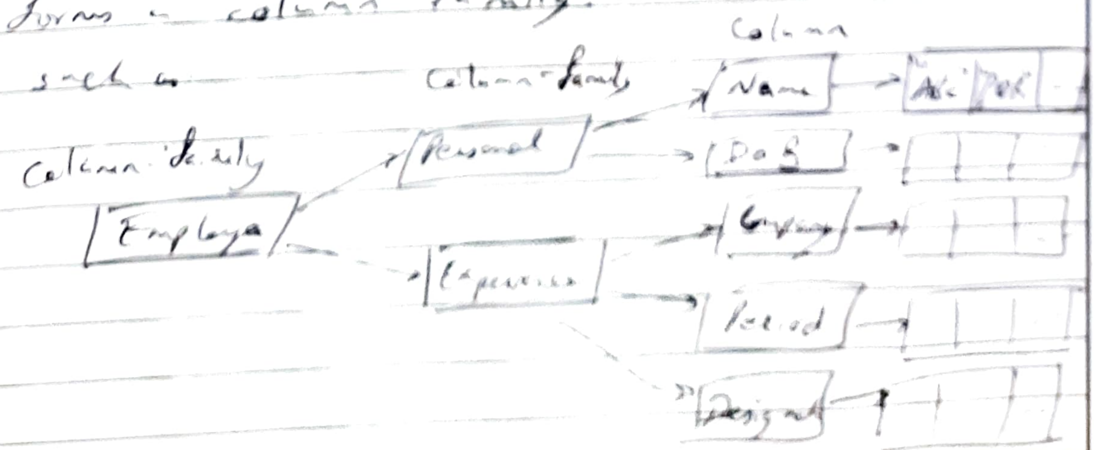
→ Cannot retrieve values without key.

→ No query language support.

Examples: Amazon DynamoDB, Berkeley DB, Couchbase, etc..

2 c) Column family store architecture pattern can be seen as a flipped version of relational database wherein each column of RDBMS is treated as a separate row of column store. A column need not always be associated with a set of values, it can also have a group of columns within it. Thereby it forms a column family.

such as



Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

Every column is treated seperately.
Every column format can vary from one row to another.

Advantages:
→ Data is readily available.
→ Easier to perform aggregate operations on column such as SUM, AVG, ...

Drawbacks:
→ Designing effective working schema is difficult.
→ Although similar to relational model, this model is less ~~segu~~ secured and not suitable for transaction processing

Examples
HBase
Big table by Google
Cassandra
etc - -

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE
	3a	

Bloom filter is used to check whether an element is a member of a set or not. It ~~dos~~ uses hash seach technique. Instead of using regular hash search technique, Bloom filter uses a space-efficient model where in for each has position it stores a 0 or a 1.

0 → No element hashed so far

1 → An element has hashed to the location.

To reduce number of false positives (~~can~~ caused ~~by~~ due to collision) it uses multiple ~~hash of~~ hash functions. Number of hash functions to use depends on the probablistic analysis of Bloom filter

An element x is said to be present if for ~~ea~~ every hash function there is a 1 then the element x is considered to be a member of set S, else not a member

Example Consider Set $S = \{3, 5, 8\}$ and hash functions

$h_1 = (key * 7) \bmod 13$

$h_2 = (key * 3) \bmod 13$

$h_3 = (key * 5) \bmod 13$

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

Initial Bit Vector

0	1	2	3	4	5	6	7	8	9	10	11	12
0	0	0	0	0	0	0	0	0	0	0	0	0

Element 3

$h_1 = (3 * 7) \bmod 13 = 8$

$h_2 = (3 * 3) \bmod 13 = 9$

$h_3 = (3 * 5) \bmod 13 = 2$

Hence Bit vector become

0	1	2	3	4	5	6	7	8	9	10	11	12
0	0	1	0	0	0	0	0	1	1	0	0	0

Similarly, for element 5 & 8

$h_1 = (5 * 7) \bmod 13 = 9$ $\quad (8 * 7) \bmod 13 = 4$

$h_2 = (5 * 3) \bmod 13 = 2$ $\quad (8 * 3) \bmod 13 = 11$

$h_3 = (5 * 5) \bmod 13 = 2$ $\quad (8 * 5) \bmod 13 = 1$

Bit vector after all element ~~inserted~~

0	1	2	3	4	5	6	7	8	9	10	11	12
0	1	1	0	1	0	0	0	1	1	0	1	0

Now consider an element 4

h_1 gives $(4 * 7) \bmod 13 = 8$ (it is 1)

h_2 gives $(4 * 3) \bmod 13 = 12$ (it is 0)

Therefore 4 is not present in the set.

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE
	3b	FM algorithm is used to count number of distinct element in the set.

~~It~~ It hashes every element in the set using hash function and the hash value obtained is represented in binary.

For each Binary representation it ~~it~~ then determines number of trailing zeroes by counting number of '0's from LSB till it gets a 1. This count is represented ~~as~~ set r.

Consider given set $S = \{1,3,2,1,2,3,4,3,1,2,3,1\}$ and hash function $h(x) = (6x+1) \bmod 5$

Applying the above step we get.

x	$h(x)$	In Binary	no of trailing ~~No. of trailing~~ zeros (r)
1	$(6*1+1) \bmod 5 = 2$	10	1
3	$(6*3+1) \bmod 5 = 4$	100	2
2	$(6*2+1) \bmod 5 = 3$	11	0
1	$(6*1+1) \bmod 5 = 2$	10	1
2	$(6*2+1) \bmod 5 = 3$	11	0
3	$(6*3+1) \bmod 5 = 4$	100	2
4	$(6*4+1) \bmod 5 = 0$	0	-
3	$(6*3+1) \bmod 5 = 4$	100	2
1	$((6*1)+1) \bmod 5 = 2$	10	1
2	$((6*2)+1) \bmod 5 = 3$	11	0
3	$(6*3+1) \bmod 5 = 4$	100	2
1	$(6*1+1) \bmod 5 = 2$	10	1

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

Next FM algorithm determines the max of r as R.
For above example $R = \max\{0, 1, 2\}$
$= 2$

Hence, number of distinct elements as per FM algorithm are 2^R

For the given example,
Number of distinct elements $= 2^2 = \underline{\underline{4}}$

4a Comparison between Content & Collaborative Filtering for recommendation systems.

i) Content based filtering has lots of information about item ~~...~~ features. Such as, for a movie, its genre, release year, director, actor, etc. Whereas, Collaborative based filtering does not rely on items features. It is based on similarity between two users.

ii) Content based filtering groups items based on their similar features & user history of user to create item profile for user. Whereas, collaborative filtering groups users having similar interests and recommends items based on ~~...~~ likings by other members of the group.

iii) Content based filtering can recommend

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

a new item to user because it matches the item profile interest of the user. Whereas, collaborative filtering has difficulty in recommending a new item to any user unless it is liked by others in group.

iv) Content based filtering can create profile for new user by explicitly asking questions relevant to create user's profile. Whereas, collaborative filtering cannot ~~create~~ recommend items to new user, ~~unless there is a history~~

v) Content based recommends similar items to user and hence gets ~~a~~ over specialized at times.
Whereas, Collaborative based recommends items ~~on~~ with varied features and can cater to multiple interests of a user group.

~~Eg~~ Example: Consider a case of movie recommendations.

Content based: will recommend movies to a user only based on the user's history & movie's feature. Thereby, it can cater to multiple users but can get over specialized,

collaborative based: will create groups of users who like similar movies and if a user X from group likes a movie, it will recommend that movie to all others in group.

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE
	4b	Girvan - Newman algorithm for Community detection.

Approach of GN algorithm is to determine edge betweenness for all edges of a community graph ~~and~~ (connected graph) and remove the edge having highest edge betweenness. This leads to creating 2 or ~~more~~ more communities ~~of~~ of the graph.

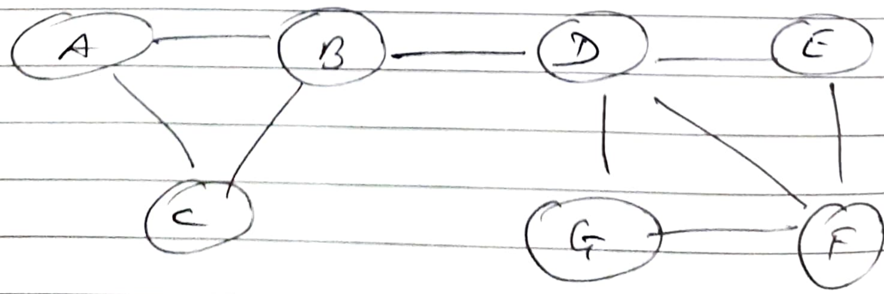
The above procedure is repeated for each community till the threshold is met.

It is based on hierarchical decomposition technique of clustering.

Steps of GN are as follows:-

1) Consider any one node of the given graph and apply modified BFS on it. (The modified BFS allows one child node to have multiple parents and hence o/p is a DAG instead of tree).

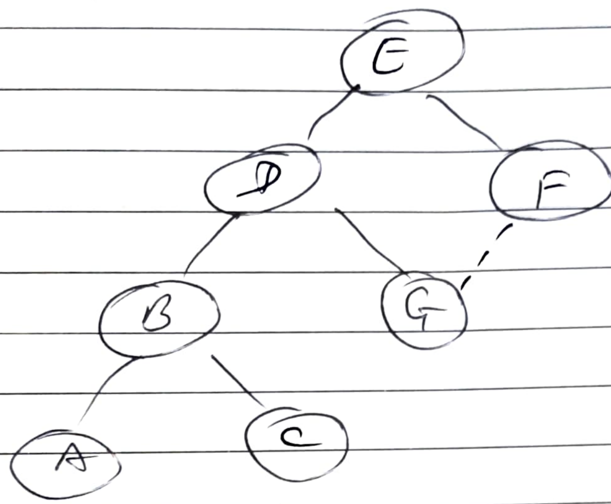
Given graph



Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

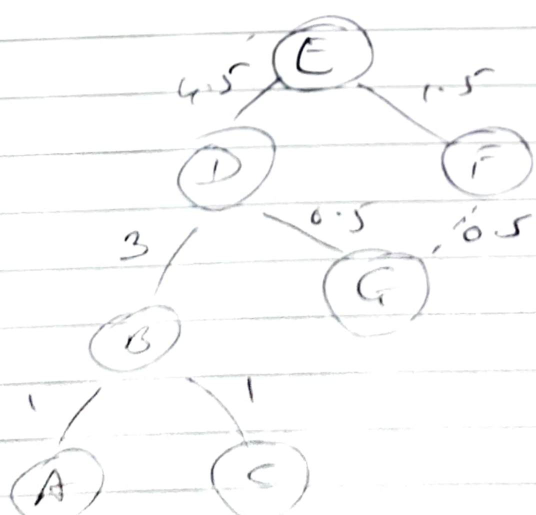
Space for Marks	Question No.	START WRITING HERE

Consider a node (say E).
Applying modified BFS on E, we get



2) Assign values for each edge from bottom to top. A child having multiple parents gives 0.5 $\left(\frac{1}{2}\right)$ to both parents or $\left(\frac{1}{3}\right)$ to all 3 parents, and so on...

Values for each edge in above case is:



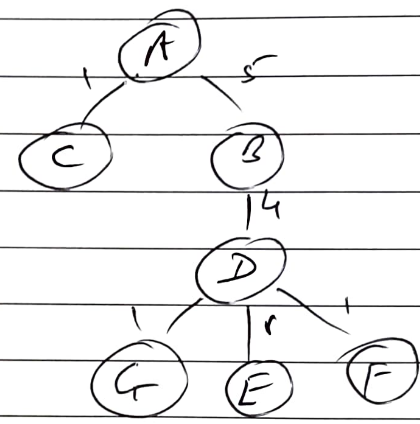
Hence, with E as root node (ie source), there are 3 shortest paths that travel from D to B.

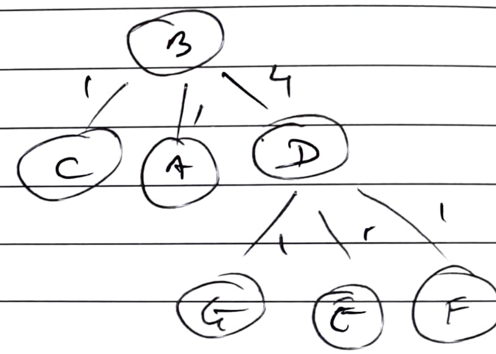
Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

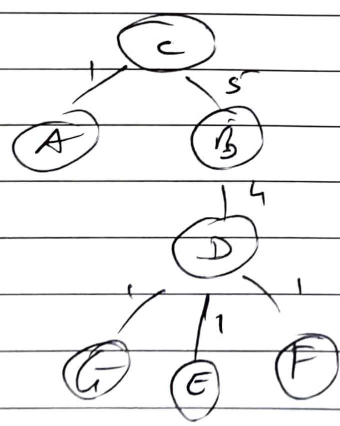
Space for Marks	Question No.	START WRITING HERE

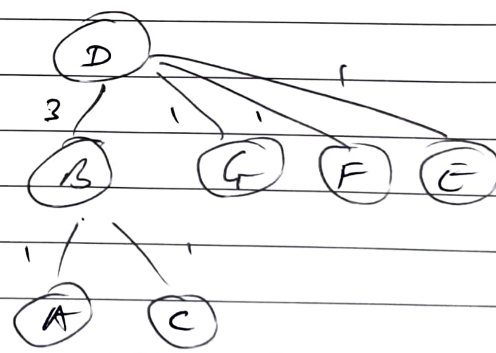
~~Applying similar step~~

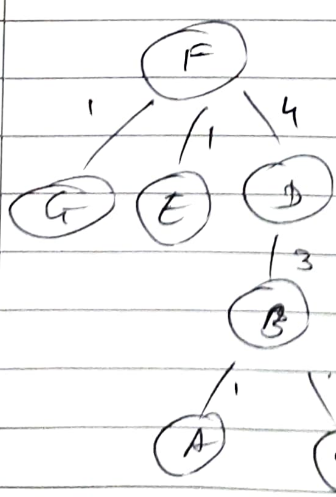
Similarly we apply previous 2 steps for all nodes and get

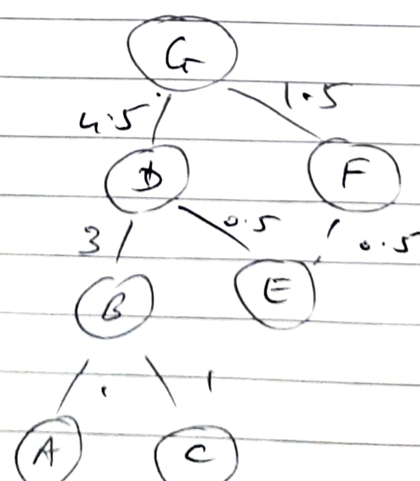








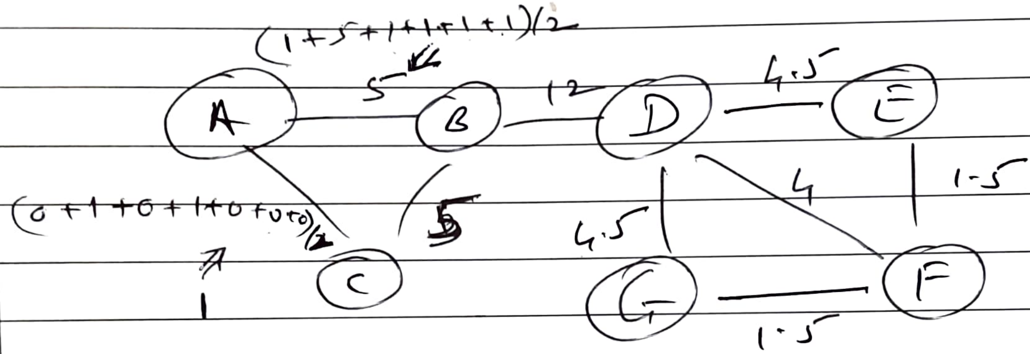




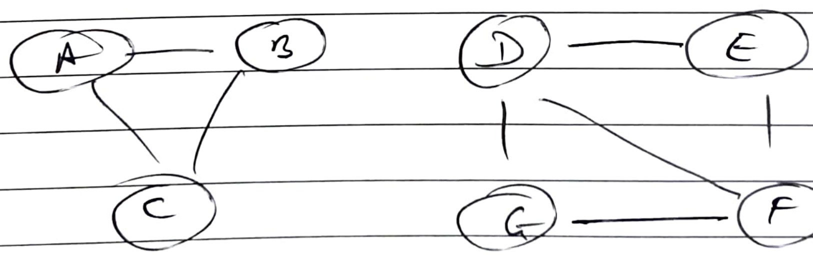
Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

Summing up the paths from all above graphs we get total. This total when divided by 2 gives edge betweenness As follows :-



Removing edge BD that has highest edge betweenness (12), we get 2 communities as shown



The above procedure is to be repeated seperately for both the communities {A, B, C} & {D E, F, G} to detect further communities.

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

5a The most commonly used ~~da~~ data structures in R include:

- Vectors
- Matrix
- Data Frame
- List

Vector:

A vector is a collection of elements that are commonly of type character, logical, integer or numeric.

To create a vector ~~you~~, we call function vector().

Example:

vector("character", length = 5)

creates a vector of type character with 5 elements ~~such~~ as:

" " " " " " " " " "

The above can be directly constructed with character(5)

numeric(5) creates 0 0 0 0 0

By default, the type of vector is logical and each element of logical vector is FALSE

Vectors can be created directly as

x <- c(1, 2, 3)

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

will create a numeric vector of length 3 having ~~val~~ elements 1, 2 and 3.

~~Matrix~~ Other functions on vectors include Typeof(), length(), class() and str().

Matrix :

Matrix in R is an extension of numeric of character vectors. It is created by specifying number of rows & columns as :-

m1 <- matrix (nrow = 2, ncol = 2)

This creates a matrix m1 having 2 rows & 2 columns as below

	[, 1]	[, 2]
[1,]	NA	NA
[2,]	NA	NA

Dimensionality of a vector can ~~be found~~ be determined using dim() function.

Ex : dim (m1) will return 2 2.

Functions on matrix

class(m1) shows that ~~typ~~ m1 is a matrix and typeof(m1) shows that matrix is of what type

A particular element of matrix ~~can~~ be referred by its row and column numbers in bracket as m1[2, 1]

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

~~Data Frame~~ list :

* Unlike vectors lists can have elements of different types.

Example :

x <- list(1, "a", TRUE, 3+4y)

list is itself treated to be a type of vector element & can be constructed as
x <- vector("list", length = 4)

Hence, the content of the list can be refered using double ~~square~~ square brackets such as
x[[2]]
will return "a" for the list declared on top.

DataFrame :-

A data frame is a special type of list where in each element of list have same length. Hence called frame.
It is very useful to represent data in a tabular format.
A data frame can be created as
df <- data.frame(id = letters[1:10], x = 11:20)

The content of df will be as follows

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

id	x
a	11
b	12
c	13
d	14
e	15
f	16
g	17
h	18
i	19
j	20

It is usually used to obtain data which is read from external files such as csv files or xls files.
Operations to read them ~~are~~ are read.csv() and read.xls().
This imports data from files into a data frame in R.

Other operations on data frame are.

head(), tail(), dim(), nrow(), ncol(), str(), etc...

Contents of data frames are also refered using double square brackets (like lists) or using $.

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE
	5b	Data Vizualization is a technique of representing data as a graph or in a pictorial format.

R support basic 4 types of plots :-
1) Barplots
2) Histogram
3) Box plots
& 4) scatter plots.

1) There are 2 types of Bar plots - horizontal and vertical which represent data points as horizontal or vertical bars of certain lengths proportional to the value of the data items. They are generally used for categorical variable plotting.

2) Ex : barplot (movierating $Movie, main = "Movie Ratings', xlab = 'user feedback', col = 'blue', horiz = ~~FAL~~ FALSE)

3) Bar plots are useful while comparing 2 or more categories in the data set, or analyze change of a variable over time in days or months.

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

2) Histogram:

A histogram is like a bar chart as it uses bars of varying height to represent data distribution. However, in a histogram, values are grouped into consecutive intervals called bins. Continous values are grouped and displayed ~~as~~ in these bins whose size can be varied.

Ex: hist (airquality $Temp, main " maximum Temperature (Daily)", xlab = " Temperature (Farhenheit)" xlim = c(50,125), col = "yellow", freq = TRUE)

Histogram are useful to verify an equal and symmetric distribution of data. and to identify deviations from expected values.

3) Box Plots:

Statistical summary of the given data is presented graphically using a boxplot.

A boxplot depicts information like the minimum and maximum data points, the median value, first and third quartiles, and ~~inter-qua~~ interquartile range.

Ex: boxplot (airquality $ Wind, main = "Wind Speed",

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

xlab = "miles per hour", col = "orange",
border = "brown", horizontal = "TRUE",
notch = "TRUE")

Box plot is useful to give a comprehensive statistical description of the data and identify the outlier points.

4) Scatter Plot:
A scatter plot is composed of many plots on cartesian plane. Each point denotes the value taken by two parameters and helps us easily identify the relationship between them.

Ex plot(airquality $Ozone, airquality $Month,
main = "Scatterplot", xlab = "Ozone concentration in parts per billion"
ylab = "Month of observation",
pch = (9))

Scatter plot is useful to show whether an correlation exist between bivariate data and to measure the strength and direction of such a relationship.

*Other types include Heat Map(), 3D graphs using persp() and showing in maps using library (maps).