```
df = dataframe (class = c('fy', 'sy', 'Ty'),
                pass = c(80, 60, 90),
                fail = c(20, 40, 10))

summary (df)

# Type of

class (df)
class (df $ pass)
class (df $ class)

# To retrive values
```

1- df [3,]
→ [Ty, 90, 10]

2- df [,3]
→ [20, 40, 10]

3- df $ pass
→ [80, 60, 90]

Relational data

→ It refers to data that is organised in a tables or dataframe that are related to each other.
→ These relationships can be through of as connection between different sets of data often structured in a way that allows them to be easily combine filter or analysis in a context.
→ The concept of relational data typically involves multiple table or dataframes and how they are related through common variables or column.
→ These is similar to the concept of relational database such as SQL database where different tables can be related via keys (like primary key or foreign key).

key Concepts

1- Tables

2- primary keys

3- foreign keys

4- Joins :- Inner, left, right, full

26/11/24

* Date and time

```
x ← as.date ('1970-01-07')
sys.date()      # current system date and time
sys.time()      # only time current
sys.timezone()
 date()         # Current date and time
```

Q. Operations on Date

1- Subtract dates in R

→
```
x ← as.Date ('1970-01-07')
y ← as.Date ('1970-02-06')
 sys.Date (x-y)

x ← as.Date ('1970-01-07') - as.Date ('2000-01-06')
 print (x)
```

2- Add days to dates.

→
```
y ← as.Date ('2021-03-10') +3
 y ← x1 +3
```

3- find the internal between the dates.

→
```
z ← c ('1950-04-02', '1987-05-04')
  as.Date (diff (z))
```

4- Generate Sequence of dates

→
```
a ← (as Date ('2021-05-10'),
        length = 5,
         by = month)
```

lubricate

library (lubricate)

→ now()

→ a <- c('2020-10-01', '2021-10-02', 2022-10-04')

years (a) # print all year

month (a) # print all month

mday (a) # print all day.

Manipulate multiple date values in R

dates <- c('2020-01-02', '2020-02-04', '2020-04-20')

print (dates + years (1))

print (dates + month (1))

mday (date) <- c(22, 18, 15)

print (date)

Update multiple dates values in R.

30/11/24

Page No.
Date

* Time

Using POSIXCT (calendra)

```
timect <- as.POSIXct('2022-01-10 22:10:20 PST')
print (timect)
class (timect)
```

Using POSIXlt (local)

```
timelt <- as.POSIXlt('2024-10-20 10:50:45 PST')
print (timelt)
class (timelt)
```

Operations on time

1- Subtract time in R
   timect - timelt

2- Extract parts of time
   timect $ sec

Importing Data in R

1- Using the combine command

Eg.
```
v1 <- c(1, 2, 3)
v2 <- c(4, 5, 6)
v3 <- c(7, 8, 9)
combined_vector <- c(v1, v2, v3)
```

2- Entering Numerical Items

```
data1 <- c(5,8,7,3,2,9,6,1)
data2 <- c(data1, 4,5,7,6)
```

3- Entering Text time

```
day1 <- c('Mon', 'Tue', 'Wed', 'Thu')
day1 <- c(day1, 'fri', 'Sat')
```

4- Using Scan() Command

```
data <- data.frame(x1 = c(1,2,3), x2 = c(4,5,6), x3 = t(7,8,9))
write.table(data1, file = "data1.txt", row.names
                                = FALSE).
getwd()
scan_data <- scan("data1.txt", what = "character")
```

5- Using scan() to retrive

```
write.table(data1, file = "data2.csv", row.names = FALSE)
getwd()
scan_data2 <- scan("data2.csv", what = "character")
scan_data3 <- scan("data3.csv",
                   list = c(" ", " ", " ")).
scan_data4 <- scan("data2.csv", skip = 1).
```

Read csv file

```
data1 <- read.csv('filepath', show_col_type = FALSE)
data2 <- read.table('filepath', sep = ",", header = 1)
```

Import txt file

```
data3 <- read.delim('filepath', header = F)
```

Import Excel

```
library(readxl)
data1 <- read_excel('filepath', sheet = 1)
```

Import Json

```
library(rjson)
data5 <- fromJson('filepath')
```

Import SQL

```
library(RSqLite)
conn <- dbconnect(RSqLite::SqLite), 'filepath')

dblistTables(conn)
dbGetQuery(conn, "Select * from Table_name")
```

2/12/24

Page No.
Date

* Data Wrangling

→ It is a process reimaging the raw data to a most structured format which will help to get better inside and make better decisions from the data.
eg. tableau

* Tribbles

→ Tribbles are the core datastructure of the tidy verse and is used to faciliate the display, and analysis of information in a tidy format.
→ Tribbles is a new dataframe where dataframes are the most common data structure used to stored datasets in R.

Advantage of tribbles over dataframe

→ All tidy verse package support tribbles.
→ It print in a much clear format then dataframe.
→ A dataframe often convert characters string to factor and analysis obtain have to override the setting while tribbles doesn't try to make these conversion automatically.

Page No.
Date

Q Compare tibbles and dataframe

→

	Tribble	Dataframe
①	Modern dataframe variant from tidyverse	① Traditional R data structure
②	prints first few rows intelligently	② prints entire datasets
③	Disables partial matching	③ Allows partial column name matching
④	Discourages row names	④ Support row names
⑤	Requires Conversion	⑤ Native R structure
⑥	Optimized for performance	⑥ Slower for large datasets
⑦	Tidyverse ecosystem	⑦ Base R
⑧	preserves original data types	⑧ Converts strings to factors
⑨	More memory efficient	⑨ less memory efficient
⑩	Consistent tibble subsetting	⑩ Traditional R subsetting
⑪	To column Access :- df$column	⑪ To column Access :- df$column
⑫	tibble() or as_tibble()	⑫ data.frame()

* Tidyverse

→ A tidyverse is a set of packages that work in harmony because they share common data representation and api design the tidyverse packages is designed to make it easy to install and read core packages from the tidy verse in single command

install.packages ("tidyverse")

library (tidyverse)

1- ggplot 2 - for data Visualization

2- dplyr, for data manipulation

3- tidyr, for data tidying

4- readr, for data import

5- purrr, for functional programming

6- tibble, A modern re-imaging of dataframe

7- stringr, for strings

8- forcats, for factors

9- lubricate, for data and time

* Data Analytics

→ Data information is a raw format the increase in size of data has to lead to arise in need for coming out inspection, data cleaning, transformation as well as data modelling to gain insight from the data in order to derive conclusions for better decision making process. This process is known as data Analytics.
→ Data mining is popular type of data Analysis technique carry out data modelling as well as knowledge, discovery that is geared towards predictive process.
→ Business Intelligence operation provide various data analysis capabilities that ready on data aggregation as well as focuses on domain expertise of the business.
→ In statical application business analytics can be divided into EDA, CDA.
→ EDA focuses on discovery new feature in the data.
→ CDA focuses on confirming or specifying existing hypothesis.

EDA
→ It is only approach with the help of descriptive, statics and visual method it is not a formal process that contain a strict a set of rules.

Q. Why do we used EDA graph in data analysis.
→ EDA graphs are used in data analysis to visually explore data, identify patterns, detect outliers, understand relationships between variables, check for missing values, and asses data quality, all of which guide further analysis or model building.

7/12/24

Page No.
Date

* Terminologies in EDA

1- Variable :- It is quality, quantity or property that you can measure.

Types of Variable

1- quality :- Variables takes on value that are names or labels.

e.g Color of a ball it can green, blue, yellow.

Types of quality Variable

1- Nominal :- Basically it displays graphical data or all ordering are equally meaningful.

e.g Student Religion

2- Ordinal :- A categorical variable whose categories can be meaningful ordered is called ordinal

e.g Students grade

2- quantity :- Variables that can measure on a numeric or quantitive scale.

eg age, count of anything

Types of quantity

1- Discrete :- A discrete variable is one that cannot take on all variables within the limits of the variable

e.g no. of students

2- Continue :- The variable can take on any value between two specify values

e.g age

2- Value :- It is the state of a variable when you measure it the value of a variable may change from measurement to measurement.

3- Observation :- It is a set of measurement made under similar conditions and observation will contain several values each associated with a different variable. It will sometime refer to an observation as a data point.

4- tabular data :- Basically it is set of values each associated with a variable an observation tabular data is tidy if each value is it own "cell". Each variable in its own column and Each observation in its own row.

5- DataSet :-

Page No.
Date

* Data Transformation

→ It is a technique used to convert the raw data into a suitable format that efficiently eases data mining and retrives strategic informatic.
→ It changes the format, structured or values of the data and convert them into clean usable data.

Q. Which are the process to perform data transformation.

11/12/24

Page No.
Date

* Model building

It is critical phase in the Data Science process where data is transform into action able insight and predictions There are multiple steps to build a model

1- Define a problem and set objective
① Clearly define a problem and aim to solve and establish objectives.
② Understand the scope of desired outcomes of model
③ This steps ensure that the model align the problem hand and provide meaningful

2- Gather proper data
① Collect the revelant data require for model building.
② Clean and preprocess the data handling missing values and outliers and Unseen data.
③ perform factor engineering and extract meaningful predictor or ensure data quality

3- Split the data into training and testing set
The training set is used to train the model while the testing set serve as Unseen data set for evaluating the model performance.
e.g. Cross Validation.

4- Choose the right algorithm
Select appropriate annale algorithm based on problem type
e.g. Classification and Regression.

5- Train the model
fit the selected algorithm to the training data adjust the models parameters and hypoparameter to optimize its performance

6- Evaluate the model
Acceses the model performance and appropriate evalution matrices.
e.g Confusion matrix

7- find tune and Optimize
Refine model to enhance its performance.

8- The Result
① Understand the models output to gain insight into underline pattern and relationship in the data
② Analyse ~~fraction~~ feature importance co-efficient or decision boundary to explain model behaviour.

9- deploy and monitor

17/12/24

Page No.
Date

* R Markdown

→ R Markdown is a free and open source R package that provide a workplace for creating data science project.
→ The main advantage of R Markdown is that it allow you to combine add text and data visualization in a single polished, shareable fully reproducible document that can be rendered in a wide variable of output formats both static and interactive.
→ Some of its popular output formats are HTML, PDF, Msword, presentations, application website dashboard, report, templates, articles, books, etc.
→ R Markdown allow easy version control tracking and support may programming languages besides R, including python and SQL.

* NoSQL

→ It is a type of dbms that is designed to handle and store large volumes of Unstructured and semi-structured data.

→ NoSQL databases use flexible data models that can adapt to changes in data structures and capable of scaling horizontally to handle growing amounts of data.

Types of NoSQL databases

1- Document databases :- Store data in documents similar to JSON objects. Each document contains pairs of fields and values. The value can typically be a variety of types including things like strings, numbers, booleans, arrays or objects.

2- key-value databases :- They are a simple type of database where each item contain keys and values.

3- Wide-Column stores :- They store data in tables, rows and dynamic columns.

4- Graph databases :- Store data in nodes and edges. Nodes typically store information about people, places and things, while edges store information about the relationships between the nodes.

Brief history

1- 1998 - Carlo Strozzi
Use the term for lightweight, open source relational database.

2- 2000 - Graph DB Neo4j is launched.

3- 2004 - Google Big table is launched

4- 2005 - CouchDB is launched

5- 2007 - The research paper Amazon Dynamo is released.

6- 2008 - Facebook open source the cassandra project.

7- 2009 - The term NoSQL was reintroduced.

24/12/25

Page No.
Date

* Big data

Ex :- 1- Snapchat - 2.1 million
2- Search engine - 3.8 million
3- Facebook - 1.0 million
4- youtube - 4.5 million
5- Email - 188 million

Q. How do you classified data as big data or characteristic of big data

→ 5v's

Eg Healthcare Industries

1- Volume :- (High amount of data) Hospital and clinics across the world generated massive volume of data 2314 Exabyte of data are collected annually.

2- Velocity :- (High speed of data) in the form of patient records and test results all these data is generated at a very high speed which attributes to the velocity of big data.

3- Variety :- (different format of data from various sources) refer the various datatype such as structured, Unstructured, semi- structured data for Eg. Excel records logs files and X- Ray images.

4- Veracity :- (In consitency of Uncertain in data) Accuracy trust worthyness of generated the data is terms Veracity.

Date

5- Value :- (Useful data) Analysis all these data with benefit the medical sector by mainly faster disease detection, better treatment and reduced cost is known as value of big data.

Q. How do we stored and process big data.

7/01/25

Page No.
Date

* High dimensional data

It refers to data set in which the Number of features 'P' is greater than number of observation 'N' for example :- A dataset has p = 6 features and N = 3 observation would be considered as high dimensional data.

Q. Why is high dimensional data problem ?

Examples of high dimensional
Healthcare Industries
financial data
Bionomics

Q. How to handle high dimensional data.
→ 1- Choose to include fewer features
1- The most obvious way to avoid dealing with high dimensional data is to simply include fewer feature is the dataset.
2- Drop feature with many missing value.
3- Drop feature with low variance.
4- Drop feature with low correlation with the response variable.
2- Use a regularization method.
Another way to handle high dimensional data without dropping feature from the dataset is to use a regularization technique such as :-
i- PCA
ii- Ridge
iii- casso.

Q. Why
→ Whe
of

→ In
th
va
e

C

1- (
2- [

Page No.
Date

Q. Why is high dimensional data problem?

→ When the no. of features in a dataset exceeds the no. of observation we will never have a deterministic answer.

→ In other word it becomes impossible to find a model that can describe the relationship between the predictor variable and the response variable because we don't enough observation to ~~the~~ train the model on.

Case Study

1- Google big data Service
2- Big data on aws.

Case Study : Spotify and Google Cloud Big Data Services

→ Challenges :- Managing and analyzing vast amount of user data for personalized recommendations and real-time insights.

→ Solutions :-