DATA SCIENCE

UNIT I

Introduction to Data Science and Life Cycle

Syllabus (12 Hrs):

Introduction to Data Science, definition and description, history and development, terminologies, basic framework and architecture, need for Data Science, benefits and uses, Data Science Life Cycle, overview of different Data Science techniques.

Contents:

Q.1 – Q.3 What is Data Science, history, terminologies ..... page 2

Q.4 Framework and architecture ..... page 5

Q.5 – Q.6 Need, benefits and uses ..... page 6

Q.7 Data Science Life Cycle ..... page 8

Q.8 Techniques of Data Science ..... page 10

Q.9 – Q.10 Comparisons, skills, roles and tools ..... page 12

Definitions, short and long questions ..... page 14

MCQs and last minute revision ..... page 18

Ink code: black = headings, blue = answers, red = important points.

SARTHI EDUCATION
Learn • Prepare • Achieve

Name: _____

Q.1 What is Data Science? Define and describe it. [5 Marks]

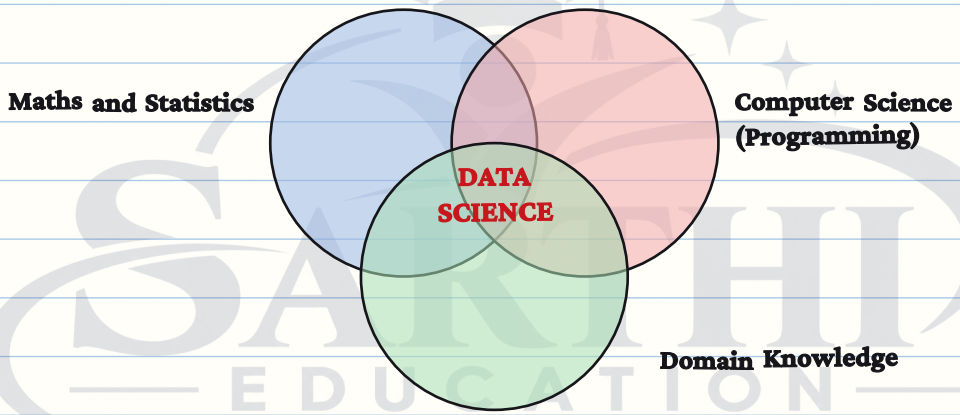
Definition:

Data Science is an interdisciplinary field that uses statistics, programming (algorithms) and domain knowledge to collect, clean, analyse and understand data, and to get useful insights and predictions from it.

Description:

Every day a huge amount of data is created by mobiles, websites, banks, hospitals and sensors. Data Science turns this raw data into useful knowledge that helps people take better decisions.

- *It works with structured data (tables) and unstructured data (text, images, video).*
- *It uses machine learning to find patterns and to predict the future.*
- *It shows results with charts and dashboards so that anyone can understand.*



Example:

Online shopping sites study what customers bought earlier and then suggest products to them. Banks study card use to find fraud. This is Data Science.

Data Science = Statistics + Programming + Domain knowledge

SARTHI EDUCATION Learn • Prepare • Achieve

Q.2 Explain the history and development of Data Science. [5 Marks]

Year	Event
1962	John Tukey writes "The Future of Data Analysis" (data analysis as a science)
1974	Peter Naur uses the term "data science" in a book on computer methods
1977	Tukey publishes "Exploratory Data Analysis"; IASC is founded
1990s	Data mining and KDD grow; data warehouses become common
1997	C. F. Jeff Wu suggests that statistics be renamed "data science"
2001	William Cleveland proposes data science as a separate field
2006	Hadoop is released for big data storage and processing
2008	"Data scientist" becomes a popular job title (Patil, Hammerbacher)
2012	Harvard Business Review calls it the sexiest job of the 21st century
2010s	Machine learning, deep learning, Spark, cloud and GPUs grow fast
2020s	Generative AI, AutoML and MLOps; used in every field

Stages of development:

(1) Statistics era: data was small and analysed by statisticians.

(2) Database and data mining era: databases, warehouses and mining tools.

(3) Big data era: internet, social media and sensors gave huge data, so Hadoop and Spark were made.

(4) AI era: machine learning and deep learning give accurate predictions.

Remember: Tukey (1962), Naur (1974), Wu (1997), Cleveland (2001).

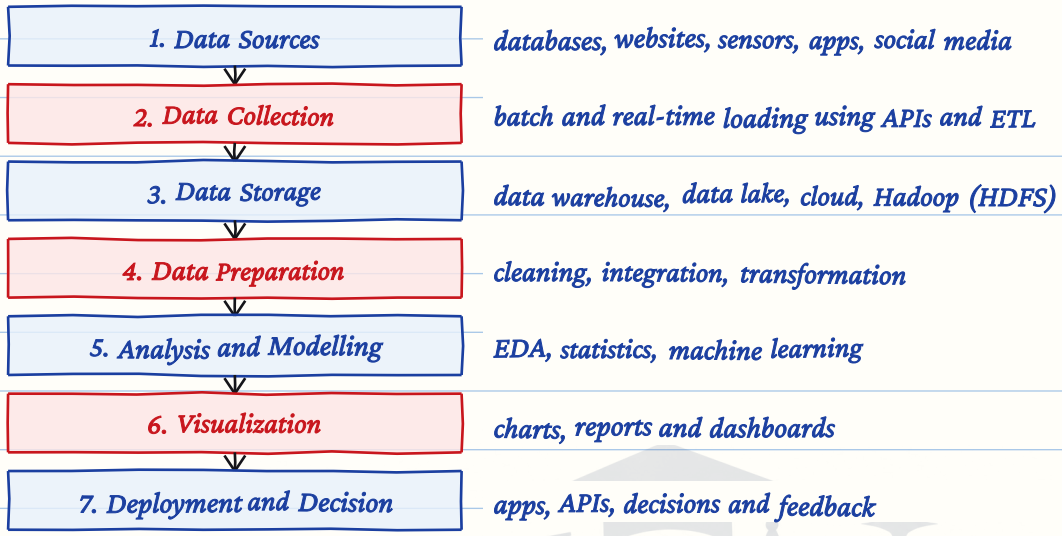
SARTHI EDUCATION Learn • Prepare • Achieve

Q.3 Terminologies related to Data Science. [5 Marks]

- *Data: raw facts and figures like numbers, text, images.*
- *Dataset: a collection of related data, usually in rows and columns.*
- *Big Data: very large and fast data that normal tools cannot handle. Described by Volume, Velocity and Variety.*
- *Structured data: data in fixed rows and columns, like a database table.*
- *Unstructured data: data without fixed format, like emails, photos, videos.*
- *Data mining: finding hidden patterns in large data.*
- *Machine Learning (ML): computers learn from data to make predictions without being clearly programmed.*
- *Artificial Intelligence (AI): machines that show human-like thinking and decisions.*
- *Deep Learning: ML that uses many-layer neural networks, used for images and speech.*
- *Algorithm: a step-by-step method to solve a problem.*
- *Model: a mathematical representation learned from data, used for prediction.*
- *Feature (variable): an input column used by the model, such as age or salary.*
- *Label / Target: the output the model must predict, such as house price.*
- *Training data: data used to teach (build) the model.*
- *Testing data: new data used to check how good the model is.*
- *Data wrangling: cleaning and changing raw data into a usable form.*
- *Data warehouse: a central store of cleaned, integrated and historical data.*
- *Data lake: a store of raw data in any format.*
- *Outlier: a value very different from other values.*
- *Data analytics: examining data to find conclusions and trends.*
- *Data visualization: showing data as charts, graphs and dashboards.*
- *Prediction: estimating a future or unknown value from data.*

Q.4 Explain the basic framework and architecture of Data Science. [10 Marks]

The framework shows how data flows from its source to the final decision. It has these layers:



Explanation of layers:

- **Sources and collection**: data comes from many places and is brought in by batch or real-time (streaming) loading.
- **Storage**: raw data goes to a data lake, cleaned data goes to a data warehouse.
- **Preparation**: errors, missing values and duplicates are removed and data is converted to a common format.
- **Analysis and modelling**: patterns are found and machine learning models are built and tested.
- **Visualization and deployment**: results are shown to users and the model is used in real applications. Feedback improves the model again.

Three parts of the framework: Data, Tools (Python, R, SQL, Hadoop, Spark) and People (data scientist, engineer, analyst).

Q.5 Explain the need for Data Science. [5 Marks]

Today data is called the "new oil" because it is very valuable if it is processed properly. Data Science is needed for the following reasons:

(1) Huge growth of data: mobiles, social media, sensors and internet create data every second. It cannot be handled by old manual methods.

(2) Better decisions: companies use facts and patterns instead of guesses.

(3) Prediction of future: sales, demand, weather and risk can be predicted.

(4) Competition: companies that use data give better service and win customers.

(5) Personalisation: suggesting movies, songs and products to each person.

(6) Automation: repeated tasks like sorting emails or checking documents are done by machines.

(7) Fraud and risk control: banks and insurance firms find suspicious activity quickly.

(8) Solving complex problems: disease detection, traffic control, climate study.

(9) Saving cost and time: finding waste and improving the process.

Example:

A hospital studies past patient records and predicts which patients may have heart disease. Doctors can then start treatment early.

In short: Data Science is needed to convert big raw data into useful decisions.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.6 Write the benefits and uses of Data Science. [5 Marks]

Benefits:

1. *Better and faster decision making.*
2. *Increases profit and reduces cost.*
3. *Improves customer experience (personal offers).*
4. *Predicts future trends, demand and risks.*
5. *Detects fraud, errors and unusual activity.*
6. *Creates new products and jobs.*

Uses (applications):

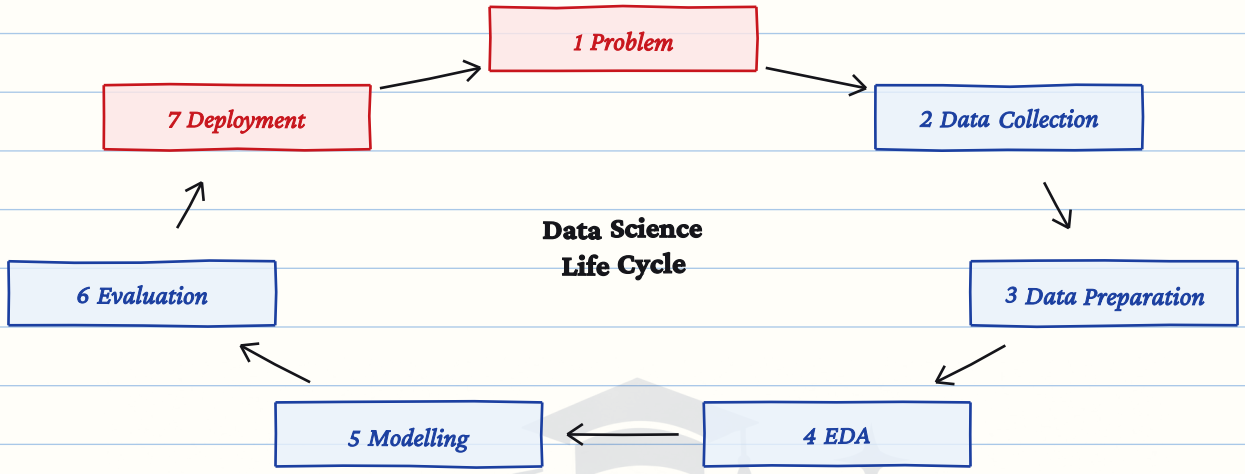
Field	Use of Data Science
Healthcare	*disease prediction, medical image study, drug research*
Banking and Finance	*fraud detection, credit scoring, risk analysis*
E-commerce and Retail	*product recommendation, demand forecasting*
Transport	*route planning, traffic and ride demand prediction*
Education	*student performance prediction, personal learning*
Social media	*sentiment analysis, friend and ad suggestions*
Agriculture	*crop yield and weather prediction*
Manufacturing	*predictive maintenance, quality control*
Entertainment	*movie and song recommendation*

Examples: Amazon and Netflix recommendations, Google Maps traffic, bank fraud alerts.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.7 What is the Data Science Life Cycle? Explain with a diagram. [10 Marks]

Data Science Life Cycle is the series of steps followed to complete a data science project, from understanding the problem to using the final model. The steps repeat, so it is a cycle.



Phases in short:

1. *Problem (business) understanding: define the goal.*
2. *Data collection: gather data from sources.*
3. *Data preparation: clean and transform data.*
4. *EDA: explore data using statistics and charts.*
5. *Modelling: build machine learning models.*
6. *Evaluation: check the accuracy of the model.*
7. *Deployment: put the model to use and monitor it.*

Details of each phase are on the next page.

Q.7 Data Science Life Cycle: details of each phase

1. Problem (Business) Understanding

Understand what the business wants. Define the goal, questions and success measure. Example: reduce customer loss by 10%.

2. Data Collection

Gather data from databases, files, websites, APIs, surveys and sensors. Check that data is enough and correct.

3. Data Preparation

Remove duplicates, fill missing values, correct errors, join tables, and change data to a suitable form. This takes most of the time (about 60-80%).

4. Exploratory Data Analysis (EDA)

Study the data using summary statistics, graphs and correlation to find patterns, outliers and useful features.

5. Modelling

Choose an algorithm (regression, classification, clustering). Train the model on training data.

6. Evaluation

Test the model on testing data. Measure accuracy and errors. If the result is poor, go back and improve.

7. Deployment and Monitoring

Release the model in a real application. Watch its performance, update it with new data, and share results with users.

The cycle repeats: results and feedback may send us back to earlier phases.

SARTHI EDUCATION

Q.8 Overview of Data Science techniques: types of learning. [5 Marks]

Most data science techniques are machine learning methods. They are divided into three types.

- Supervised learning: model learns from labelled data (input with known output). Used for classification and regression.
- Unsupervised learning: model finds hidden patterns in unlabelled data. Used for clustering and association.
- Reinforcement learning: an agent learns by trying actions and getting rewards or penalties (games, robots).

Supervised and Unsupervised learning:

Basis	Supervised	Unsupervised
Data	Labelled	Unlabelled
Goal	Predict known output	Find hidden patterns
Tasks	Classification, regression	Clustering, association
Algorithms	Linear regression, KNN, SVM	K-means, hierarchical, PCA
Example	Spam email detection	Customer grouping
Feedback	Yes (answers known)	No feedback

Semi-supervised learning uses a small labelled part and a large unlabelled part of data.

Remember: Supervised = labelled, Unsupervised = unlabelled.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.8 Overview of different Data Science techniques (continued). [5 Marks]

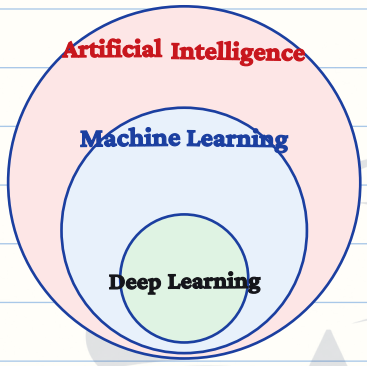
- *Classification: puts data into classes. Example: email is spam or not spam; loan is safe or risky.*
- *Regression: predicts a number. Example: house price, temperature, sales.*
- *Clustering: groups similar items without labels. Example: customer segments (K-means).*
- *Association rule mining: finds items that occur together. Example: people who buy bread also buy butter.*
- *Anomaly (outlier) detection: finds unusual data. Example: credit card fraud.*
- *Dimensionality reduction: reduces the number of variables but keeps information (PCA).*
- *Time series forecasting: predicts future values from past values over time. Example: stock, sales.*
- *Text mining / NLP: understands human language. Example: sentiment analysis, chatbots.*
- *Recommendation systems: suggest items to users. Example: Netflix, Amazon.*
- *Deep learning: neural networks for images, speech and text. Example: face recognition.*
- *Data visualization: shows data as charts so that patterns are easy to see.*

SARTHI EDUCATION Learn • Prepare • Achieve

Q.9 Differences between Data Science, Data Analytics, Machine Learning and AI. [5 Marks]

Basis	Data Science	Data Analytics
Focus	Predict future, new insights	Study past data
Scope	Broad (whole life cycle)	Narrow (analysis part)
Data	Structured and unstructured	Mostly structured
Skills	Programming, stats, ML	Statistics, SQL, Excel
Tools	Python, R, Spark	Excel, SQL, Power BI
Output	Models, predictions	Reports, dashboards

Relation of AI, ML and Deep Learning:



Deep Learning is a part of ML,
ML is a part of AI.

Data Science uses all of them.

- AI: machines that act intelligently. ML: machines learn from data. Data Science: whole process of getting insight from data.
- Big Data is the large data; Data Science is the method to analyse it.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.10 Roles, skills and tools of a Data Scientist. [5 Marks]

Skills needed:

1. Mathematics and statistics.
2. Programming (Python, R, SQL).
3. Machine learning knowledge.
4. Data visualization and storytelling.
5. Domain (business) knowledge.
6. Communication and teamwork.

Roles in a data science team:

Role	Work
Data Scientist	builds models and finds insights
Data Analyst	analyses data, makes reports
Data Engineer	builds data pipelines and storage
ML Engineer	deploys models into applications
Business Analyst	links business needs with data results

Common tools:

Purpose	Tools
Programming	Python, R
Database	SQL, MySQL, MongoDB
Big data	Hadoop, Spark
Data analysis	NumPy, Pandas, SciPy
Visualization	Matplotlib, Seaborn, Tableau, Power BI
Machine learning	scikit-learn, TensorFlow
Notebook	Jupyter Notebook

SARTHI EDUCATION Learn • Prepare • Achieve

Important Definitions (Unit I)

- **Data Science**: field that uses statistics, programming and domain knowledge to get insights from data.
- **Data**: raw facts and figures.
- **Big Data**: extremely large, fast and varied data (3 Vs: Volume, Velocity, Variety).
- **Structured data**: data in rows and columns, like a table.
- **Unstructured data**: data without a fixed format, like text, image, video.
- **Machine Learning**: computers learning from data without being clearly programmed.
- **Artificial Intelligence**: machines showing human-like intelligence.
- **Data mining**: finding hidden patterns in large data.
- **Data wrangling**: cleaning and converting raw data to a usable form.
- **Data warehouse**: central store of cleaned and historical data.
- **Data lake**: store of raw data in any format.
- **EDA**: Exploratory Data Analysis, studying data using statistics and charts.
- **Model**: mathematical form learned from data, used for prediction.
- **Supervised learning**: learning from labelled data.
- **Unsupervised learning**: finding patterns in unlabelled data.
- **Data Science Life Cycle**: the steps from problem definition to deployment of a data science project.

GARHI EDUCATION Learn • Prepare • Achieve

Short Answer Questions (2 / 3 Marks) – Part 1

Q1. Define Data Science.

Ans: It is a field that uses statistics, programming and domain knowledge to get insights from data.

Q2. Who first used the term "data science"?

Ans: Peter Naur used it in 1974.

Q3. What are the three parts of Data Science?

Ans: Statistics, computer science (programming) and domain knowledge.

Q4. What is Big Data?

Ans: Very large, fast and varied data (Volume, Velocity, Variety).

Q5. Differentiate structured and unstructured data.

Ans: Structured has rows and columns; unstructured has no fixed format (images, text).

Q6. What is machine learning?

Ans: Computers learn from data and make predictions.

Q7. What is a data warehouse?

Ans: A central store of cleaned, integrated, historical data.

Q8. What is a data lake?

Ans: A store of raw data in any format.

Q9. Name any four uses of Data Science.

Ans: Healthcare, banking fraud detection, e-commerce recommendation, transport.

Q10. What is a model?

Ans: A mathematical form learned from data used for prediction.

Short Answer Questions (2 / 3 Marks) – Part 2

Q11. Name the phases of the life cycle.

Ans: Problem, collection, preparation, EDA, modelling, evaluation, deployment.

Q12. What is data preparation?

Ans: Cleaning, integrating and changing data to a usable form.

Q13. What is EDA?

Ans: Studying data with statistics and graphs to find patterns.

Q14. Why is evaluation done?

Ans: To check the accuracy of the model on testing data.

Q15. What is supervised learning?

Ans: Learning from labelled data, used for classification and regression.

Q16. What is clustering?

Ans: Grouping similar items without labels.

Q17. Give an example of classification.

Ans: Spam or not spam email.

Q18. Give an example of regression.

Ans: Predicting house price.

Q19. Which tools are used in Data Science?

Ans: Python, R, SQL, Hadoop, Spark, Tableau.

Q20. Why is Data Science needed?

Ans: To turn huge raw data into useful decisions.

SARTHI EDUCATION Learn • Apply • Achieve

Important University Questions (Long Answer)

Find the full answer on the page given in brackets.

1. What is Data Science? Define and describe it. (5) [page 2]
2. Explain the history and development of Data Science. (5) [page 3]
3. Explain the terminologies related to Data Science. (5) [page 4]
4. Explain the basic framework and architecture of Data Science. (10) [page 5]
5. Explain the need for Data Science. (5) [page 6]
6. Write the benefits and uses of Data Science. (5) [page 7]
7. Explain the Data Science Life Cycle with a diagram. (10) [page 8]
8. Explain different Data Science techniques. (5/10) [page 10]
9. Differentiate Data Science and Data Analytics. (5) [page 12]
10. Write about roles, skills and tools of a data scientist. (5) [page 13]

Most repeated: Life cycle, need and uses, framework, techniques.

SARTHI EDUCATION Learn • Prepare • Achieve

Multiple Choice Questions (Part 1)

1. Data Science mainly extracts _ from data

(a) insights (b) hardware (c) paper (d) electricity

2. Which is NOT a part of Data Science?

(a) Statistics (b) Programming (c) Domain knowledge (d) Welding

3. Peter Naur used the term data science in

(a) 1974 (b) 1990 (c) 2008 (d) 2012

4. Hadoop was released around

(a) 1980 (b) 1995 (c) 2006 (d) 2020

5. Photos and videos are

(a) structured data (b) unstructured data (c) only numbers (d) none

6. A database table with rows and columns is

(a) structured data (b) unstructured data (c) audio data (d) none

7. Volume, Velocity and Variety are the

(a) 3 Vs of Big Data (b) types of errors (c) life cycle steps (d) tools

8. First step of the life cycle is

(a) problem understanding (b) modelling (c) deployment (d) visualization

9. Removing errors and missing values is part of

(a) modelling (b) data preparation (c) deployment (d) evaluation

10. EDA stands for

(a) Extra Data Analysis (b) Exploratory Data Analysis (c) Exact Data Algorithm (d) Estimated Data Analysis

11. A model is tested in the _ phase

(a) collection (b) evaluation (c) problem (d) none

12. Supervised learning uses

(a) labelled data (b) no data (c) only images (d) only text

Answers: 1-a 2-d 3-a 4-c 5-b 6-a 7-a 8-a 9-b 10-b 11-b 12-a

Multiple Choice Questions (Part 2)

13. Clustering is a type of

(a) supervised learning (b) unsupervised learning (c) reinforcement (d) none

14. Predicting house price is

(a) classification (b) clustering (c) regression (d) association

15. Spam detection is an example of

(a) classification (b) clustering (c) storage (d) sorting

16. K-means is used for

(a) clustering (b) regression (c) cleaning (d) storage

17. A data warehouse stores

(a) integrated historical data (b) only images (c) only code (d) nothing

18. A data lake stores

(a) only reports (b) raw data in any format (c) only numbers (d) only charts

19. Which is a data science language?

(a) Python (b) Photoshop (c) Word (d) Paint

20. Product suggestions in online shopping are

(a) a data science use (b) hardware fault (c) manual work (d) none

21. Data analytics mainly studies

(a) past data for insights (b) robots (c) chips (d) typing

22. Relation of AI, ML and DL is

(a) DL ⊂ ML ⊂ AI (b) AI ⊂ ML ⊂ DL (c) ML ⊂ DL (d) all same

23. Credit card fraud finding is

(a) anomaly detection (b) sorting (c) data storage (d) typing

24. Final phase of the life cycle is

(a) problem (b) deployment (c) EDA (d) collection

Answers: 13-b 14-c 15-a 16-a 17-a 18-b 19-a 20-a 21-a 22-a 23-a 24-b

Last Minute Revision – Unit I

Introduction

- *Data Science = Statistics + Programming + Domain knowledge.*
- *History: Tukey 1962, Naur 1974, Wu 1997, Cleveland 2001, Hadoop 2006, HBR 2012.*

Framework and need

- *Layers: Sources, Collection, Storage, Preparation, Modelling, Visualization, Deployment.*
- *Need: huge data, better decisions, prediction, competition, automation, fraud control.*

Life cycle (remember the order)

Problem → Collection → Preparation → EDA → Modelling → Evaluation → Deployment

Techniques

- *Supervised: classification, regression. Unsupervised: clustering, association.*
- *Others: anomaly detection, time series, NLP, recommendation, deep learning.*

Differences to remember

- *Data Science is broad (future, models). Data Analytics is narrow (past, reports).*
- *Data warehouse = cleaned data. Data lake = raw data.*
- *Structured = rows and columns. Unstructured = text, image, video.*

Draw the life cycle diagram in every 10 mark answer.

SARTHI EDUCATION Learn Prepare Achieve

DATA SCIENCE

UNIT II

Data Science Processes

Syllabus (11 Hrs):

Six steps of the data science process: define research goals, data retrieval, cleansing data and correcting errors, integrating data from different sources, transforming data, exploratory data analysis, data modelling, model and variable selection, model execution, model diagnostic and model comparison, presentation and automation.

Contents:

Q.1 Overview of the six steps ..... page 2

Q.2 – Q.3 Research goal and retrieving data ..... page 3

Q.4 – Q.6 Cleansing, integrating, transforming ..... page 5

Q.7 Exploratory data analysis ..... page 9

Q.8 Data modelling (selection, execution, diagnostics, comparison) ..... page 11

Q.9 – Q.10 Presentation, automation, summary ..... page 15

Definitions, short and long questions, practice ..... page 17

MCQs and last minute revision ..... page 22

Ink code: black = headings, blue = answers, red = important points.

Name: _____

Q.1 What is the Data Science Process? Explain the six steps with a diagram. [10 Marks]

The data science process is a step-by-step method to complete a data science project. It has six steps. The steps are repeated if the result is not good, so the process is iterative.



Step	Main work
1 Set research goal	understand the problem, write project charter
2 Retrieve data	find and collect data from sources
3 Data preparation	cleansing, integrating and transforming data
4 Data exploration	EDA with graphs and statistics
5 Data modelling	select model, execute it, diagnose and compare
6 Presentation, automation	show results and automate the work

Data preparation takes the most time (about 80% of the project).

SARTHI EDUCATION
Learn • Prepare • Achieve

Q.2 Explain Step 1: Defining research goals and the project charter. [5 Marks]

Meaning:

The first step is to understand the problem and say clearly what, why and how the project will be done. A project without a clear goal can fail even if the models are good.

Questions to ask the client (stakeholders):

- *What is the problem? Why is it important to the business?*
- *What data and resources are available? Who will use the result?*
- *In which form must the result be given (report, app, dashboard)? What is the deadline?*

Project charter:

After understanding the goal we write a project charter. It is a short document that is shown to the client and approved before work starts. It contains:

(1) A clear research goal.

(2) The project mission and context.

(3) How the analysis will be done (the approach).

(4) The resources needed (data, people, tools).

(5) Proof that the project is possible (early results or a prototype).

(6) Deliverables and the way success will be measured.

(7) A timeline of the work.

Example:

A bank wants fewer loan defaults. Research goal: predict which customers will fail to repay in the next 12 months. Deliverable: a model and a monthly report. Success: 85% correct predictions.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.3 Explain Step 2: Retrieving data. [5 Marks]

Meaning:

Step 2 is to find and collect the data needed for the goal. We check what data exists, where it is stored and whether we are allowed to use it.

1. Start with data inside the company:

Most companies already store data in different places. The data is usually cleaned and easy to use.

- *Databases: store current data in tables.*
- *Data marts: small part of a warehouse for one department.*
- *Data warehouses: central store of integrated historical data.*
- *Data lakes: store raw data in any format.*

2. Data from outside the company:

If inside data is not enough, shop around for external data. Many open sources are free:

Source type	Examples
Open government data	*data.gov.in, data.gov, Eurostat*
Public datasets	*Kaggle, UCI repository, World Bank*
APIs and web	*Twitter (X) API, web scraping*
Files and sensors	*CSV, Excel, logs, IoT sensors*

Data quality check:

Check at this stage whether the data is correct, complete and in the expected format. Errors found early are cheap to fix. Also take permission and access rights, as this may take time.

Remember: internal data first, then external data.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.4 Explain Step 3: Data cleansing and correcting errors. [10 Marks]

Data cleansing means removing errors from the data so that it is a true and consistent picture of the real world. It is the first part of data preparation. Errors are of two types:

- *Interpretation errors: impossible values, like age = 250 or a negative price.*
- *Inconsistencies between sources: "F" and "Female", rupees and dollars.*

Error	Example	Solution
Data entry error	"Godo" instead of "Good"	correct by code or manually
Redundant whitespace	"Delhi " and "Delhi"	remove spaces (strip)
Capital letter mismatch	"India" and "india"	convert to lower case
Impossible values	age = 250	use rules and sanity checks
Outliers	salary 50 lakh, others 30 thousand	plot, check, remove or keep
Missing values	blank age	omit, null or impute
Deviation from code book	gender code "X" not in book	match with the code book
Different units	km and miles	convert to one unit
Different aggregation level	weekly and monthly sales	aggregate to same level

Python commands:

```
df['city'].str.strip().str.lower()
df[df['age'] > 120] ← find impossible values
```

Remember: Garbage in, garbage out. Clean data gives a good model.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.4 Data cleansing (continued): missing values and outliers. [5 Marks]

Handling missing values:

Technique	Advantage	Disadvantage
Omit the rows	Easy	Information is lost
Set value to null	Easy	Not all methods accept null
Impute a static value (0, mean)	Keeps all rows	May create false data
Impute using a model	Keeps rows, less bias	Harder to do

Example (mean imputation): Ages are 4, 6, 8, ?, 12. Mean of known values = (4 + 6 + 8 + 12)/4 = 30/4 = 7.5. So the missing value is replaced by 7.5.

```
df['age'].fillna(df['age'].mean())
```

Finding outliers:

An outlier is a value much different from the others. We find it with a box plot or the IQR rule:

$$IQR = Q3 - Q1$$

$$\text{Lower fence} = Q1 - 1.5 \times IQR$$

$$\text{Upper fence} = Q3 + 1.5 \times IQR$$

Values outside the fences are outliers.

Example:

Data: 10, 12, 13, 14, 15, 16, 18, 60. Lower half 10, 12, 13, 14 gives Q1 = 12.5. Upper half 15, 16, 18, 60 gives Q3 = 17.

IQR = 17 − 12.5 = 4.5. Upper fence = 17 + 1.5 × 4.5 = 23.75. Lower fence = 12.5 − 6.75 = 5.75.

Since 60 > 23.75, 60 is an outlier.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.5 Explain data integration: combining data from different sources. [5 Marks]

Data integration means combining data from different tables or sources into one dataset. There are three ways:

1. Joining tables:

Columns of two tables are combined using a common key column (like ID). Example:

Client (ID, Name, Region code)	Region (Region code, Name)
1, Asha, 10	10, North
2, Ravi, 20	20, South

Join on Region code gives: 1, Asha, North and 2, Ravi, South.

```
pd.merge(clients, regions, on='code')
```

Join types: Inner (only matching rows), Left (all rows of left table), Right, Full outer (all rows).

2. Appending (stacking) tables:

Rows of one table are added below another table that has the same columns. Example: January sales + February sales = two-month sales.

```
pd.concat([jan, feb])
```

3. Using views:

A view is a virtual table made by a query. It does not copy data, so it saves disk space.

Enriching data:

New measures can be added, like total sales per region or profit = price − cost, after combining the data.

Join = add columns, Append = add rows.

PARTH EDUCATION

Q.6 Explain data transformation with examples. [10 Marks]

Data transformation changes data into a form that is suitable for modelling. It is the third part of data preparation.

1. Scaling (normalization):

$$\text{Min-max: } x' = \frac{x - min}{max - min}$$

$$\text{z-score: } z = \frac{x - mean}{S.D.}$$

Example (min-max): values 10, 20, 30, 50 (min = 10, max = 50) → (x − 10)/40 = 0, 0.25, 0.5, 1.

Example (z-score): values 2, 4, 6, 8. Mean = 5, S.D. = √5 = 2.24 → z = −1.34, −0.45, 0.45, 1.34.

2. Other transformations:

(1) *Log transform: reduces skewness of data. Example: log10 of 10, 100, 1000 = 1, 2, 3.*

(2) *Dummy variables: change a category into 0/1 columns. Colour = Red, Blue → Is Red (1/0), Is Blue (1/0).*

(3) *Reducing variables: remove extra variables or use PCA, which keeps most information in fewer variables.*

(4) *Aggregating: summarise data, like daily sales into monthly sales.*

(5) *Combining variables: new variable from old ones, like price per square foot = price/area.*

(6) *Binning: group ages into 0-18, 19-40, 41+.*

Why transform? Many models work better when variables are on the same scale.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.7 Explain Step 4: Exploratory Data Analysis (EDA). [5 Marks]

Meaning:

EDA is studying the data in detail, mostly with graphs, to understand its patterns, relations, errors and outliers before modelling. The idea was given by John Tukey.

Aim of EDA:

- *Understand the data and its shape, find relations between variables.*
- *Find errors, missing values and outliers that were missed in cleaning.*
- *Get ideas for choosing the model and variables.*

Techniques:

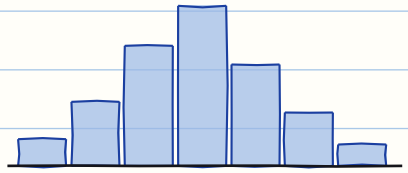
Technique	Use
Bar chart	*compare categories*
Line graph	*trend over time*
Histogram	*shape and spread of one variable*
Box plot	*median, quartiles, outliers*
Scatter plot	*relation of two variables*
Pareto diagram	*main causes (80/20 rule)*
Link and brush	*select points in one graph, see them in another*
Summary statistics	*mean, median, S.D., min, max, correlation*

Python: df.describe(), df.hist(), df.corr()

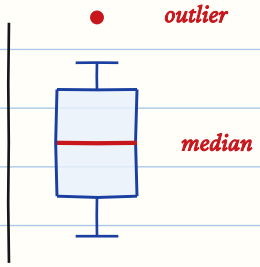
EDA is an eye test of data before building the model.

SARTHI EDUCATION Learn • Prepare • Achieve

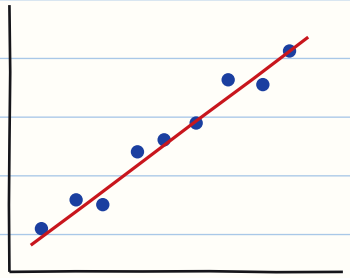
Q.7 EDA: common graphs (diagrams)



Histogram



Box plot



Scatter plot

What each graph tells us:

- *Histogram: shows the shape (normal, skewed) and spread of one variable.*
- *Box plot: shows median, Q1, Q3, whiskers and outliers as dots.*
- *Scatter plot: shows the relation of two variables. A rising line means positive correlation.*
- *Bar chart: compares categories, like sales by city.*
- *Line graph: shows change over time, like monthly sales.*
- *Pie chart: shows parts of a whole, like market share.*

Tools: Matplotlib, Seaborn, Tableau, Power BI.

SARTHI EDUCATION
Learn • Prepare • Achieve

Q.8 Explain Step 5: Data modelling. Model and variable selection. [5 Marks]

Data modelling means building a model from the prepared data to make predictions or to find patterns. It has four parts:

1. *Model and variable selection*
2. *Model execution*
3. *Model diagnostics*
4. *Model comparison*

Model selection:

We choose a technique that suits the problem. For example:

Problem	Possible model
Predict a number (price)	*Regression*
Predict a class (spam / not spam)	*Classification (decision tree, KNN)*
Group similar items	*Clustering (K-means)*
Items bought together	*Association rules*

Points to consider when choosing a model:

- *Does the model fit the type of data and problem?*
- *Is it easy to implement, maintain and explain?*
- *Does it have to run in production (real use) quickly?*
- *How much data and how many variables are available?*

Variable selection:

We choose the predictor variables (inputs) that really help the model.

- *Remove irrelevant variables and variables that are highly related to each other (they repeat the same information).*
- *Use domain knowledge and the correlation matrix. Keep the model simple.*

Fewer good variables are better than many weak variables.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.8 Model execution with a solved example (linear regression). [10 Marks]

Model execution means running (training) the chosen model on the data using tools like Python (scikit-learn, StatsModels) or R.

```
from sklearn.linear_model import LinearRegression
model = LinearRegression().fit(X, y)
```

Simple linear regression line: y = a + b x

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

$$a = \bar{y} - b\,\bar{x}$$

Example:

x (hours studied) = 1, 2, 3, 4, 5 and y (marks) = 2, 4, 5, 4, 5. Fit the line and predict y for x = 6.

x	y	$x - \bar{x}$	$y - \bar{y}$	product	$(x - \bar{x})^2$
1	2	−2	−2	4	4
2	4	−1	0	0	1
3	5	0	1	0	0
4	4	1	0	0	1
5	5	2	1	2	4
Sum 15	Sum 20			6	10

$\bar{x} = 15/5 = 3$, $\bar{y} = 20/5 = 4$. $b = 6/10 = 0.6$, $a = 4 - 0.6 \times 3 = 2.2$.

Model: $y = 2.2 + 0.6\,x$. For $x = 6$: $y = 2.2 + 3.6 = 5.8$.

Meaning: each extra study hour increases marks by 0.6 on average.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.8 Model diagnostics: how good is the model? [10 Marks]

Model diagnostics means checking the quality of the model. We split the data into *training data* (about 70-80%, to build the model) and *test data* (about 20-30%, to check it on unseen data).

`train_test_split(X, y, test_size=0.2)`

1. For numbers (regression): Mean Squared Error

$$MSE = \Sigma(actual - predicted)^2 / n$$

Example: actual 3, 5, 7; predicted 2, 5, 9. Errors –1, 0, 2 → squares 1, 0, 4. MSE = 5/3 = 1.67. Smaller MSE is better.

2. For classes: Confusion matrix

	Predicted Yes	Predicted No
Actual Yes	TP = 40	FN = 5
Actual No	FP = 10	TN = 45

$$Accuracy = (TP + TN)/Total$$
$$Precision = TP/(TP + FP),\ Recall = TP/(TP + FN)$$

Accuracy = (40 + 45)/100 = 0.85. Precision = 40/50 = 0.80. Recall = 40/45 = 0.89.

Overfitting and underfitting:

- *Overfitting*: model is very good on training data but poor on test data (it memorised the data).
- *Underfitting*: model is poor on both because it is too simple.

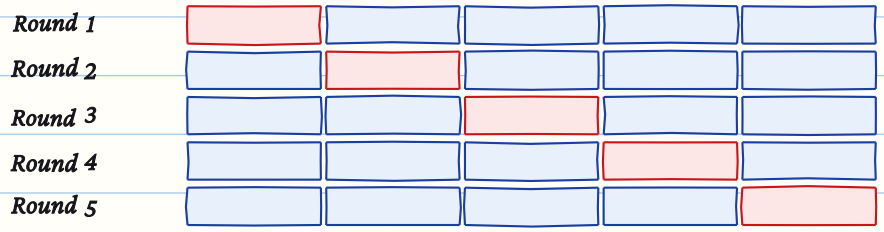
SARTHI EDUCATION Learn • Prepare • Achieve

Q.8 Model comparison: choosing the best model. [5 Marks]

Model comparison means trying more than one model and choosing the best one. All models are tested with the same test data and the same error measure.

Methods:

- *Hold-out method*: one split into training and test data.
- *K-fold cross-validation*: data is divided into k parts. Each part is used once as test data and the other parts for training. The average error is taken.



red = test part, blue = training parts (5-fold)

Example:

Model	Training error	Test error	Result
A (very complex)	2	14	Overfitting
B (moderate)	8	9	Good, choose B
C (very simple)	15	16	Underfitting

Model B has the lowest test error and the two errors are close, so Model B is selected.

Rule: compare on test error, not on training error. If errors are equal, prefer the simpler model.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.9 Explain Step 6: Presentation and automation. [5 Marks]

Presentation:

After modelling we present the results to the client or management in a form they can understand. A good result is useless if nobody understands it.

- Forms: reports, slides, dashboards (Power BI, Tableau), web apps, notebooks.
- Tips: know your audience, begin with the main finding, use simple charts, avoid technical words, tell a story, give a clear recommendation.

Automation:

Many projects must run again and again with new data. So the work is automated instead of repeating it by hand.

- Convert the code into scripts and pipelines (clean data → run model → create report).
- Use scheduled jobs so the work runs daily or weekly by itself.
- Deploy the model as an application or API so that other systems can use it.
- Monitor the model and re-train it when new data arrives or when accuracy falls.

Also in this step:

- Write documentation of data, code and results.
- Take feedback from users and improve the next version.

Remember: Step 6 = show the results (presentation) + make it run by itself (automation).

SARTHI EDUCATION Learn • Prepare • Achieve

Q.10 Summary of the six steps with activities, outputs and tools. [5 Marks]

Step	Main activities	Output
1 Research goal	understand problem, project charter	Project charter
2 Retrieve data	internal and external data, quality check	Raw data
3a Cleansing	remove errors, missing values, outliers	Clean data
3b Integrating	join, append, views	One dataset
3c Transforming	scaling, dummies, reduce variables	Model-ready data
4 Exploration	graphs, summary statistics	Insights
5 Modelling	select, execute, diagnose, compare	Best model
6 Presentation	report, dashboard, automate	Final result

Tools used:

Step	Tools
Retrieve and prepare	SQL, Python (Pandas), Excel
Exploration	Matplotlib, Seaborn, Tableau
Modelling	scikit-learn, StatsModels, R
Presentation, automation	Power BI, Jupyter, scheduled scripts

Data preparation = cleansing + integrating + transforming (Step 3).

SARTHI EDUCATION Learn • Prepare • Achieve

Important Definitions (Unit II)

- *Data science process: six steps followed to finish a data science project.*
- *Project charter: short document with goal, context, approach, resources, deliverables and timeline.*
- *Data retrieval: finding and collecting the required data from internal or external sources.*
- *Data cleansing: removing errors and inconsistencies from data.*
- *Missing value: a blank where data should be present.*
- *Imputation: replacing missing values by a mean or another estimated value.*
- *Outlier: a value far from the other values.*
- *Data integration: combining data from different sources.*
- *Join: combining tables column-wise using a key. Append: stacking rows.*
- *View: a virtual table created by a query that stores no data copy.*
- *Data transformation: changing data into a suitable form for modelling.*
- *Dummy variable: 0/1 column made from a category.*
- *EDA: studying data using graphs and statistics to understand it.*
- *Model diagnostics: checking how good a model is using test data.*
- *Overfitting: model fits training data too closely and fails on new data.*
- *Cross-validation: testing a model on k different parts of the data.*
- *Automation: making the analysis run again and again without manual work.*

SARTHI EDUCATION Learn • Prepare • Achieve

Short Answer Questions (2 / 3 Marks) – Part 1

Q1. Name the six steps of the data science process.

Ans: Research goal, retrieving data, data preparation, exploration, modelling, presentation and automation.

Q2. What is a project charter?

Ans: A document stating goal, context, approach, resources, deliverables and timeline.

Q3. Name sources of internal data.

Ans: Databases, data marts, data warehouses and data lakes.

Q4. Name two sources of external data.

Ans: Open government data (data.gov.in) and Kaggle.

Q5. What is data cleansing?

Ans: Removing errors and inconsistencies from data.

Q6. Give two kinds of data errors.

Ans: Interpretation errors and inconsistencies between sources.

Q7. How can missing values be handled?

Ans: Omit, set to null, impute a value or impute using a model.

Q8. What is an outlier?

Ans: A value very different from the other values.

Q9. Differentiate join and append.

Ans: Join adds columns using a key; append adds rows.

Q10. What is a view?

Ans: A virtual table that saves space because data is not copied.

Short Answer Questions (2 / 3 Marks) – Part 2

Q11. What is data transformation?

Ans: Converting data into a suitable form for modelling.

Q12. What is a dummy variable?

Ans: A 0/1 variable made from a category.

Q13. Write the min-max formula.

Ans: $x' = (x - min)/(max - min)$.

Q14. What is EDA?

Ans: Studying data with graphs and statistics before modelling.

Q15. What does a box plot show?

Ans: Median, quartiles, whiskers and outliers.

Q16. What are the parts of the modelling step?

Ans: Selection, execution, diagnostics and comparison.

Q17. Why do we split data into training and test sets?

Ans: To check the model on unseen data.

Q18. Define overfitting.

Ans: Good on training data but poor on new data.

Q19. What is cross-validation?

Ans: Using each of k parts once as test data and averaging the error.

Q20. What is done in the last step?

Ans: Results are presented and the work is automated.

SARTHI EDUCATION Learn & Achieve

Important University Questions (Long Answer)

Find the full answer on the page given in brackets.

1. *Explain the six steps of the data science process with a diagram. (10) [page 2]*
2. *What is a project charter? Explain defining research goals. (5) [page 3]*
3. *Explain retrieving data and its sources. (5) [page 4]*
4. *Explain data cleansing and correcting errors. (10) [page 5]*
5. *How are missing values and outliers handled? (5) [page 6]*
6. *Explain data integration (join, append, view). (5) [page 7]*
7. *Explain data transformation with examples. (10) [page 8]*
8. *What is EDA? Explain its techniques. (5) [page 9]*
9. *Explain model and variable selection. (5) [page 11]*
10. *Explain model execution with an example. (10) [page 12]*
11. *Explain model diagnostics. (5/10) [page 13]*
12. *Explain model comparison. (5) [page 14]*
13. *Explain presentation and automation. (5) [page 15]*

Most repeated: six steps with diagram, data cleansing, transformation, model diagnostics.

SARTHI EDUCATION Learn • Prepare • Achieve

Practice Numericals with Answers

1. Min-max scale the values 5, 15, 25, 35, 45.

   Ans: $(x - 5)/40 = 0, 0.25, 0.5, 0.75, 1$

2. Impute the missing value in 10, 20, ?, 40 by the mean.

   Ans: $(10 + 20 + 40)/3 = 23.33$

3. Q1 = 20 and Q3 = 30. Find the fences. Is 52 an outlier?

   Ans: IQR = 10; lower = 5, upper = 45; yes, 52 > 45

4. TP = 30, TN = 50, FP = 10, FN = 10. Find accuracy, precision, recall.

   Ans: Accuracy = 80/100 = 0.8; precision = 30/40 = 0.75; recall = 30/40 = 0.75

5. Actual 2, 4, 6 and predicted 3, 4, 8. Find MSE.

   Ans: $(1 + 0 + 4)/3 = 1.67$

6. $x = 1, 2, 3$ and $y = 2, 4, 6$. Fit $y = a + bx$ and predict for $x = 5$.

   Ans: $b = 2$, $a = 0$, $y = 2x$, so $y = 10$

7. Z-score of 8 if mean = 5 and S.D. = 2.

   Ans: $z = (8 - 5)/2 = 1.5$

SARTHI EDUCATION
Learn • Prepare • Achieve

Multiple Choice Questions (Part 1)

1. How many steps are in the data science process?

(a) 4 (b) 5 (c) 6 (d) 8

2. The first step is

(a) modelling (b) setting the research goal (c) presentation (d) cleansing

3. A project charter is made in

(a) step 1 (b) step 3 (c) step 5 (d) step 6

4. A data warehouse is an example of

(a) internal data source (b) graph (c) model (d) error

5. data.gov.in is a source of

(a) open data (b) private data (c) model (d) code

6. Removing errors from data is called

(a) integration (b) cleansing (c) modelling (d) automation

7. Age = 250 is a(n)

(a) impossible value (b) dummy variable (c) view (d) model

8. Replacing missing values by the mean is

(a) omission (b) imputation (c) joining (d) scaling

9. Outliers can be found by

(a) box plot (b) join (c) view (d) dummy

10. Adding columns of two tables using a key is

(a) append (b) join (c) scaling (d) binning

11. Stacking rows of tables with same columns is

(a) join (b) append (c) view (d) PCA

12. A virtual table made by a query is a

(a) view (b) model (c) graph (d) key

Answers: 1-c 2-b 3-a 4-a 5-a 6-b 7-a 8-b 9-a 10-b 11-b 12-a

EARTHI EDUCATION Learn • Prepare • Achieve

Multiple Choice Questions (Part 2)

13. Min-max scaling gives values between

(a) –1 and 1 (b) 0 and 1 (c) 0 and 100 (d) 1 and 10

14. A 0/1 column made from a category is a

(a) dummy variable (b) outlier (c) key (d) view

15. PCA is used for

(a) reducing variables (b) adding rows (c) cleaning spaces (d) joining

16. EDA is mainly done using

(a) graphs and statistics (b) only SQL (c) only code (d) hardware

17. A scatter plot shows

(a) relation of two variables (b) one category (c) time only (d) errors only

18. Which is NOT a part of modelling?

(a) Selection (b) Execution (c) Diagnostics (d) Data entry

19. Training data is used to

(a) build the model (b) test the model (c) clean data (d) present data

20. A model good on training but poor on test data shows

(a) overfitting (b) underfitting (c) good fit (d) no error

21. Accuracy in a confusion matrix is

(a) (TP + TN)/Total (b) TP/(TP + FP) (c) TP/(TP + FN) (d) FP/Total

22. Smaller MSE means

(a) better model (b) worse model (c) no change (d) more error

23. k-fold cross-validation uses each part as test data

(a) once (b) never (c) twice only (d) k + 1 times

24. The last step of the process is

(a) presentation and automation (b) retrieval (c) EDA (d) goal setting

Answers: 13-b 14-a 15-a 16-a 17-a 18-d 19-a 20-a 21-a 22-a 23-a 24-a

Last Minute Revision – Unit II

Six steps: Goal → Retrieve → Prepare → Explore → Model → Present and automate

Step 1 and 2

- *Project charter: goal, context, approach, resources, deliverables, timeline.*
- *Internal data first (database, mart, warehouse, lake), then external (open data).*

Step 3: Data preparation

- *Cleansing: spaces, capital letters, impossible values, outliers, missing values.*
- *Integrating: join (columns), append (rows), views.*
- *Transforming: scaling, log, dummy variables, PCA, aggregation.*

Step 4 and 5

- *EDA: histogram, box plot, scatter plot, summary statistics.*
- *Modelling: selection → execution → diagnostics → comparison.*
- *Formulas: IQR fences, min-max, z-score, regression* $b = \Sigma(x - \bar{x})(y - \bar{y})/\Sigma(x - \bar{x})^2$, *MSE, accuracy.*

Step 6

- *Present results simply; automate with scripts, scheduling, deployment and monitoring.*

Always draw the six-step diagram in a 10 mark answer.

SARTHI EDUCATION Learn • Prepare • Achieve

DATA SCIENCE

UNIT III

Facets of Data, Python, NumPy and Pandas

Syllabus (11 Hrs):

Facets of Data: structured data, unstructured data, natural language, machine-generated data, network data, audio, images and video, streaming data. Data science process: overview, defining the goal, retrieving data, data preparation, data exploration, build the models. Python in Data Science: NumPy, Pandas, SciPy, Matplotlib. NumPy basics: arrays and vectorized computation, getting started with Pandas, data loading, storage and file formats.

Contents:

Q.1 – Q.5 Facets of data ..... page 2

Q.6 – Q.8 Data science process overview and case study ..... page 7

Q.9 – Q.11 Python, NumPy arrays and vectorized computation ..... page 11

Q.12 – Q.13 SciPy and Matplotlib ..... page 15

Q.14 – Q.16 Pandas, file loading and differences ..... page 17

Definitions, short and long questions, practice ..... page 21

MCQs and last minute revision ..... page 26

Ink code: black = headings, blue = answers, red = important points.

SARTHI EDUCATION
Learn • Prepare • Achieve

Name: _____

Q.1 What are the facets of data? Explain each facet in short. [10 Marks]

Data comes in many forms. In data science these forms are called the facets of data. The main facets are:

Facet	Meaning	Example
Structured data	*fixed rows and columns*	*bank table, Excel sheet*
Unstructured data	*no fixed format*	*emails, PDFs, posts*
Natural language	*text in human language*	*reviews, news, chats*
Machine-generated	*made by machines automatically*	*server logs, sensor data*
Graph / network	*nodes and links*	*friends on social media*
Audio, image, video	*sound and pictures*	*photos, songs, CCTV*
Streaming data	*continuous real-time flow*	*stock ticks, live tweets*

Points to remember:

- *Each facet needs a different way of storing and analysing.*
- *Structured data is easy to analyse. The others need special techniques like text mining, graph analysis, computer vision.*
- *A project may use more than one facet together.*

Remember the seven facets: Structured, Unstructured, Natural language, Machine-generated, Graph, Audio-image-video, Streaming.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.2 Explain structured and unstructured data. Differentiate them. [5 Marks]

Structured data:

Data that is organised in a fixed format of rows and columns. It follows a predefined model (schema) and is stored in relational databases and queried by SQL.

EmpID	Name	Dept	Salary
101	Asha	HR	30000
102	Ravi	IT	45000

- *Examples: bank records, student records, Excel sheets, sales tables.*
- *Easy to store, search and analyse.*

Unstructured data:

Data that has no fixed format or model. It cannot be easily put in rows and columns. About 80% of the world's data is unstructured.

- *Examples: emails, Word and PDF files, social media posts, photos, audio, video.*
- *Needs special tools (text mining, image processing) and more storage.*

Difference:

Basis	Structured	Unstructured
Format	Fixed rows and columns	No fixed format
Storage	Relational database	Data lake, files, NoSQL
Query	SQL	Special tools
Analysis	Easy	Difficult
Example	Student table	Email, video

Semi-structured data (JSON, XML) has tags but no strict table.

Q.3 Explain natural language as a facet of data. [5 Marks]

Meaning:

Natural language is the language humans speak and write (English, Hindi, Marathi). It is a special kind of unstructured data. Computers find it difficult because language has meaning, context and many styles.

Why is it difficult?

(1) Ambiguity: one word can have many meanings (bank of a river and a bank for money).

(2) Context: the same sentence can mean different things in different places.

(3) Different languages and spellings, slang and short forms.

(4) Sarcasm: "Great, it rained again!" has a negative meaning.

Natural Language Processing (NLP) tasks:

- *Sentiment analysis: finding if a review is positive or negative.*
- *Entity recognition: finding names of persons, places, dates.*
- *Topic recognition and text summarisation.*
- *Machine translation and chatbots.*

Simple steps of text analysis:

Collect text → remove punctuation and stop words (is, the) → split into words (tokens) → count words or convert to numbers → apply a model.

Tools: NLTK, spaCy. Applications: spam filter, Google Translate, voice assistants.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.4 Explain machine-generated data and graph (network) data. [5 Marks]

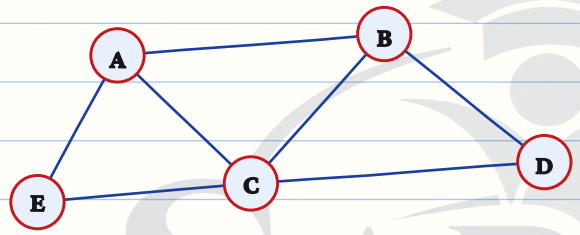
Machine-generated data:

Data that is created automatically by machines without human help.

- *Examples: web server logs, sensor readings, call detail records, application error logs, smart meter data.*
- *It is produced in huge volume and high speed, so it must be analysed with automated tools.*
- *Uses: finding system failures, security threats, predictive maintenance.*

Graph (network) data:

Data that shows relationships between things. A graph has nodes (objects) and edges (links between them).



node = person, edge = friendship
(a small social network)

- *Examples: Facebook and LinkedIn friends, road maps, web page links, flight routes.*
- *Stored in graph databases like Neo4j and queried by Cypher or SPARQL.*
- *Measures: degree (number of links of a node), centrality (importance of a node), clusters (groups).*

Uses: friend suggestion, finding influencers, fraud rings, shortest route.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.5 Explain audio, image, video and streaming data. [5 Marks]

Audio, image and video data:

These are multimedia data. They are unstructured and very large, and hard for computers to understand. Special fields are used:

- *Computer vision: understanding images and video (face recognition, medical X-ray, self-driving cars).*
- *Speech recognition: converting audio to text (Siri, Google Assistant).*
- *Deep learning (neural networks) gives the best results for such data.*
- *Examples: photos, songs, CCTV video, MRI scans, YouTube videos.*
- *Challenge: very large storage and heavy computing power (GPU) is needed.*

Streaming data:

Data that arrives continuously in real time, is processed immediately and is not stored first.

- *Examples: stock market prices, live tweets, sensor data, online payments, live video.*
- *Needs real-time processing to give instant results, like fraud alerts.*
- *Tools: Apache Kafka and Spark Streaming.*

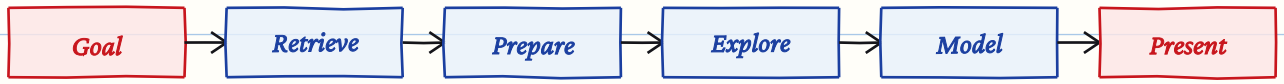
Batch and streaming:

Batch data	Streaming data
Collected and processed later	Processed as it arrives
Result after some time	Instant result
Example: monthly report	Example: fraud alert

Remember: Vision for images, speech recognition for audio, Kafka for streams.

Q.6 Overview of the data science process: defining the goal and retrieving data. [5 Marks]

The data science process is a set of steps to finish a project in an organised way. The steps are: defining the goal, retrieving data, data preparation, data exploration, building the models, presentation and automation.



Step 1: Defining the goal

Understand the problem, the client's needs, the expected result and the deadline. Write a project charter. A clear goal saves time and cost.

Step 2: Retrieving data

1. Look for data inside the company (databases, warehouses, data lakes).
2. If it is not enough, get data from outside (open data, APIs, purchased data).
3. Check quality of data and get permission to use it.

Example: To predict house prices, the goal is a price prediction model, and data is taken from property records and open city data.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.7 Overview of data preparation and data exploration. [5 Marks]

Data preparation:

Raw data is full of errors, so it is made ready for analysis. It has three parts:

- *Cleansing: remove errors, spaces, impossible values, outliers; handle missing values (omit or fill with mean).*
- *Integrating: combine data from many sources by joining tables (columns) or appending tables (rows).*
- *Transforming: scale values, make dummy variables (0/1), reduce the number of variables.*

This step takes about 80% of the total project time.

Data exploration (EDA):

Study the cleaned data to understand it before building the model.

- *Use graphs: histogram, box plot, scatter plot, bar chart.*
- *Use statistics: mean, median, standard deviation, correlation.*
- *Find patterns, relations between variables, and hidden errors or outliers.*

Python commands:

df.dropna() df.fillna(0) df.describe() df.corr()

Example:

For house price data: remove houses with price 0, fill missing area by mean, plot price against area. The scatter plot shows that a bigger area gives a higher price, so area is an important variable.

Q.8 Explain how to build the models and present the results. [5 Marks]

In this step we use the prepared data to create a model that can predict or explain. It has four parts:

1. *Model and variable selection: choose a suitable technique (regression, classification, clustering) and the useful input variables.*
2. *Model execution: train the model on training data using Python libraries such as scikit-learn or StatsModels.*
3. *Model diagnostics: test the model on unseen test data. Use error measures like MSE and accuracy. Check overfitting.*
4. *Model comparison: try more than one model and choose the one with the best test result.*

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
model = LinearRegression().fit(X_train, y_train)
```

Presentation and automation:

- *Show the results to the client using reports, charts and dashboards in simple words.*
- *Automate the work with scripts and scheduled jobs so it repeats with new data.*

The process is a loop. If results are poor, go back to earlier steps.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.8 Case study: applying the data science process (example). [5 Marks]

Problem: A college wants to find which students may fail, so that extra help can be given early.

Step	What is done in this project
1 Goal	predict pass or fail of a student; target 85% accuracy
2 Retrieve	attendance, internal marks, previous result from college database
3 Prepare	remove duplicates, fill missing attendance by mean, Yes/No to 1/0
4 Explore	scatter plot: more attendance gives higher marks (correlation 0.8)
5 Model	decision tree and logistic regression, 80% train and 20% test data
6 Present	dashboard of at-risk students for teachers
Automate	script runs every week with new marks

Results:

Decision tree test accuracy = 82%, logistic regression test accuracy = 87%. The college selects logistic regression because it has the higher test accuracy and is easy to explain.

Lesson: every step is needed. Bad data in step 3 gives a bad model in step 5.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.9 Why is Python used in Data Science? Introduce its main libraries. [5 Marks]

Reasons to use Python:

1. Simple and easy to learn, with readable code.
2. Free and open source, with a very large community.
3. Thousands of ready-made libraries for data work.
4. Works well with big data tools and databases.
5. Good for machine learning (scikit-learn, TensorFlow).
6. Works on every operating system.

Main libraries:

Library	Use
NumPy	fast arrays and mathematical operations
Pandas	tables (DataFrame), data loading and cleaning
SciPy	scientific computing: statistics, optimisation, integration
Matplotlib	plots and charts
Seaborn	attractive statistical graphs
scikit-learn	machine learning models

Installing and importing:

```
pip install numpy pandas scipy matplotlib
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats
```

Tool: Jupyter Notebook is used to write code and see the output together.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.10 Introduction to NumPy and creating arrays. [5 Marks]

Meaning:

NumPy (Numerical Python) is a library for fast numerical computing. Its main object is the ndarray, an N-dimensional array that holds items of the same data type.

Features:

- *Much faster and uses less memory than Python lists.*
- *Vectorized operations without loops, and broadcasting.*
- *Many maths functions: sum, mean, sqrt, dot product, linear algebra.*

Creating arrays:

a = np.array([10, 20, 30, 40]) → [10 20 30 40]

np.zeros(3) → [0. 0. 0.]

np.ones((2, 2)) → [[1. 1.] [1. 1.]]

np.arange(0, 10, 2) → [0 2 4 6 8]

np.linspace(0, 1, 5) → [0. 0.25 0.5 0.75 1.]

np.eye(2) → [[1. 0.] [0. 1.]]

np.random.rand(3) → 3 random numbers 0 to 1

Attributes of an array:

Code	Meaning	Output for a
a.shape	*size of each dimension*	*(4,)*
a.ndim	*number of dimensions*	*1*
a.size	*total elements*	*4*
a.dtype	*data type*	*int64*

Remember: array items must be of one type.

Q.11 NumPy: arrays and vectorized computation. [10 Marks]

Vectorization means doing an operation on the whole array at once without writing a loop. It is shorter and much faster.

a = np.array([1, 2, 3, 4]) b = np.array([10, 20, 30, 40])

Operation	Result
a + b	[11 22 33 44]
a * b	[10 40 90 160]
a ** 2	[1 4 9 16]
b / a	[10. 10. 10. 10.]
a > 2	[False False True True]
np.sqrt(a)	[1. 1.414 1.732 2.]

Loop and vectorized code:

*Loop: [x * 2 for x in a_list]*

*Vectorized: a * 2 → [2 4 6 8]*

Aggregate (summary) functions:

Code	Output
a.sum()	10
a.mean()	2.5
a.max(), a.min()	4, 1
a.std()	1.118
np.argmax(a)	3 (index of largest)

Universal functions (ufuncs) like np.sqrt, np.exp, np.log work element by element.

Matrix product: A @ B or np.dot(A, B).

SARTHI EDUCATION Learn • Prepare • Achieve

Q.11 NumPy: indexing, slicing, reshaping and broadcasting.

m = np.array([[1, 2, 3], [4, 5, 6]]) shape (2, 3)

Code	Output	Meaning
m[1, 2]	6	row 1, column 2
m[0]	[1 2 3]	first row
m[:, 1]	[2 5]	second column
m[:, :2]	[[1 2] [4 5]]	first two columns
m.sum(axis=0)	[5 7 9]	column sums
m.sum(axis=1)	[6 15]	row sums
m.T	shape (3, 2)	transpose
m.reshape(3, 2)	[[1 2] [3 4] [5 6]]	new shape
m.flatten()	[1 2 3 4 5 6]	make 1-D

1-D slicing and boolean indexing:

a = np.array([1, 2, 3, 4])

a[1:3] → [2 3] a[::-1] → [4 3 2 1]

a[a > 2] → [3 4] (boolean indexing)

Broadcasting:

NumPy stretches a smaller array to match a bigger one so that operations work on different shapes.

m + np.array([10, 20, 30]) → [[11 22 33] [14 25 36]]

Row index first, column index next: m[row, column].

SARTHI EDUCATION Learn • Prepare • Achieve

Q.12 Explain SciPy with its modules and examples. [5 Marks]

SciPy (Scientific Python) is built on NumPy. It gives many functions for scientific and technical computing.

Important modules:

Module	Use
scipy.stats	statistics, distributions, tests
scipy.optimize	finding minimum, curve fitting
scipy.integrate	integration
scipy.linalg	linear algebra (solve equations)
scipy.interpolate	interpolation
scipy.signal	signal processing

Examples:

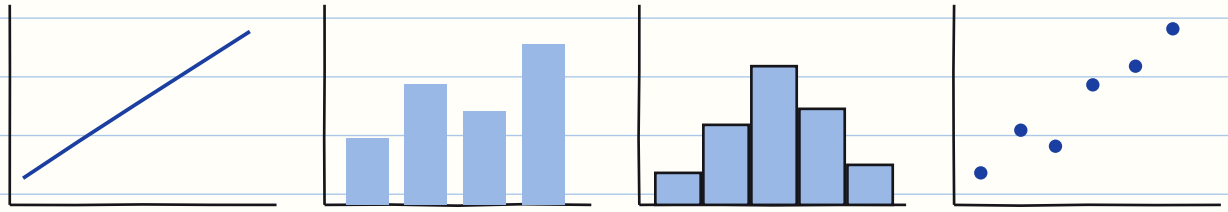
```
from scipy import stats, integrate, linalg
stats.norm.cdf(1.96) → 0.975
stats.norm.ppf(0.975) → 1.96
integrate.quad(lambda x: x**2, 0, 3) → (9.0, ...) area under x²
A = np.array([[2, 1], [1, -1]]); b = np.array([5, 1])
linalg.solve(A, b) → [2. 1.] (x = 2, y = 1)
stats.ttest_1samp(data, 50) → t value and p value
```

- *NumPy gives arrays; SciPy gives advanced scientific functions on those arrays.*

SARTHI EDUCATION Learn • Prepare • Achieve

Q.13 Explain Matplotlib and common plots with code. [5 Marks]

Matplotlib is the basic Python library to draw charts. We use its pyplot module.




```
import matplotlib.pyplot as plt
x = [1, 2, 3, 4]; y = [2, 4, 6, 8]
plt.plot(x, y) # line plot
plt.xlabel('x'); plt.ylabel('y'); plt.title('Line plot')
plt.show()
```

Plot	Code	Use
Line	plt.plot(x, y)	trend
Bar	plt.bar(names, values)	compare categories
Histogram	plt.hist(data, bins=5)	distribution
Scatter	plt.scatter(x, y)	relation
Pie	plt.pie(sizes, labels=labels)	parts of whole

Other: plt.legend(), plt.subplots(1, 2) for many plots, plt.savefig('plot.png') to save.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.14 Getting started with Pandas: Series and DataFrame. [10 Marks]

Pandas is the library for working with tables of data. It has two main structures:

- *Series: one-dimensional labelled array, like one column.*
- *DataFrame: two-dimensional table with rows and columns, like an Excel sheet.*

```
import pandas as pd
s = pd.Series([10, 20, 30], index=['a', 'b', 'c'])
df = pd.DataFrame({'Name': ['Asha', 'Ravi', 'Meena'], 'Age': [21, 22, 20], 'City': ['Pune', 'Nagpur', 'Mumbai']})
```

	Name	Age	City
0	Asha	21	Pune
1	Ravi	22	Nagpur
2	Meena	20	Mumbai

Inspecting a DataFrame:

Code	Output / use
df.shape	(3, 3) rows and columns
df.head()	first 5 rows
df.info()	column names, types, null counts
df.describe()	count, mean, std, min, max (Age mean = 21.0)
df.columns	names of columns

Series = one column, DataFrame = many columns. Every row has an index label.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.14 Pandas: selecting, filtering, sorting, grouping and missing data.

df has columns Name, Age, City (Asha 21, Ravi 22, Meena 20)

Task	Code	Result
One column	df['Age']	Series 21, 22, 20
Many columns	df[['Name', 'Age']]	two columns
By label	df.loc[0]	row of Asha
By position	df.iloc[0, 1]	21
Filter rows	df[df['Age'] > 20]	Asha, Ravi
New column	df['Marks'] = [80, 75, 90]	column added
Sort	df.sort_values('Age')	Meena, Asha, Ravi
Mean	df['Age'].mean()	21.0
Drop column	df.drop('City', axis=1)	City removed

Group by:

df.groupby('City')['Marks'].mean()→ average marks for each city

Missing data:

Code	Meaning
df.isnull().sum()	count of missing values in each column
df.dropna()	remove rows with missing values
df.fillna(0)	fill missing values with 0
df['Age'].fillna(df['Age'].mean())	fill with the mean

loc uses labels, iloc uses numbers (positions).

Q.15 Data loading, storage and file formats. [10 Marks]

Data science starts by loading data from files into Python and ends by saving the results. Pandas has a reader (read) and a writer (to) for each format.

Format	Ext.	Read in Pandas	Save in Pandas
CSV (comma separated)	.csv	pd.read_csv()	df.to_csv()
Excel	.xlsx	pd.read_excel()	df.to_excel()
JSON	.json	pd.read_json()	df.to_json()
SQL database		pd.read_sql()	df.to_sql()
HTML table	.html	pd.read_html()	df.to_html()
Pickle (Python object)	.pkl	pd.read_pickle()	df.to_pickle()

Reading a CSV file:

```
df = pd.read_csv('data.csv')
pd.read_csv('data.csv', sep=';', header=None, names=['a', 'b'],
index_col=0, skiprows=1, nrows=100, na_values=['NA'])
df.to_csv('out.csv', index=False)
```

NumPy files:

```
np.loadtxt('d.txt', delimiter=',') np.savetxt('a.csv', a, delimiter=',')
np.genfromtxt('d.csv', delimiter=',', skip_header=1)
np.save('arr.npy', a) np.load('arr.npy')
```

Notes on formats:

- CSV: simple text, opens in Excel. JSON: key-value data used by web APIs.
- Pickle keeps Python objects. HDF5/Parquet are fast and compressed for big data.
- If the file is in another folder give the full path, e.g. 'C:/data/data.csv'.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.16 Differences: list vs NumPy array, NumPy vs Pandas, Series vs DataFrame. [5 Marks]

Basis	Python list	NumPy array
Data type	mixed types allowed	one type only
Speed	slower	much faster
Memory	more	less
x * 2 on [1, 2]	[1, 2, 1, 2]	[2 4]
Maths operations	need loops	vectorized

Basis	NumPy	Pandas
Main object	ndarray	Series, DataFrame
Labels	no row or column names	row and column labels
Data	same type	different types per column
Missing data	limited support	good support (NaN)
Best for	numerical computing	data analysis, tables

Basis	Series	DataFrame
Dimension	1-D	2-D
Meaning	one column	table of columns
Create	pd.Series([1, 2])	pd.DataFrame(dict)

Pandas is built on NumPy, and both work with Matplotlib and SciPy.

SARTHI EDUCATION Learn • Prepare • Achieve

Important Definitions (Unit III)

- *Structured data: data in fixed rows and columns.*
- *Unstructured data: data with no fixed format (text, images, video).*
- *Natural language: human language, a type of unstructured data.*
- *NLP: Natural Language Processing, making computers understand human language.*
- *Machine-generated data: data created automatically by machines (logs, sensors).*
- *Graph data: data of nodes and edges showing relationships.*
- *Streaming data: continuous data processed in real time.*
- *Computer vision: understanding images and video by computers.*
- *NumPy: Python library for fast numerical arrays.*
- *ndarray: N-dimensional array of one data type in NumPy.*
- *Vectorization: operating on a whole array without loops.*
- *Broadcasting: stretching a smaller array to match a bigger array in operations.*
- *Pandas: Python library for tables (Series and DataFrame).*
- *DataFrame: 2-D labelled table of rows and columns.*
- *SciPy: library for scientific computing built on NumPy.*
- *Matplotlib: library for drawing plots and charts.*
- *CSV: text file of values separated by commas.*

SARTHI EDUCATION Learn • Prepare • Achieve

Short Answer Questions (2 / 3 Marks) – Part 1

Q1. Name the facets of data.

Ans: Structured, unstructured, natural language, machine-generated, graph, audio-image-video, streaming.

Q2. What is structured data?

Ans: Data in fixed rows and columns, stored in relational databases.

Q3. Give two examples of unstructured data.

Ans: Emails and photos.

Q4. What is NLP?

Ans: Making computers understand and use human language.

Q5. Give two NLP tasks.

Ans: Sentiment analysis and machine translation.

Q6. What is machine-generated data?

Ans: Data created automatically by machines, like server logs.

Q7. What are nodes and edges?

Ans: Nodes are objects and edges are links between them in graph data.

Q8. Name a graph database.

Ans: Neo4j.

Q9. What is streaming data?

Ans: Continuous real-time data like stock prices.

Q10. Name a tool for streaming data.

Ans: Apache Kafka or Spark Streaming.

SARTHI EDUCATION Learn • Prepare • Achieve

Short Answer Questions (2 / 3 Marks) – Part 2

Q11. Why is Python used in Data Science?

Ans: It is simple, free and has many libraries.

Q12. What is NumPy?

Ans: A library for fast numerical computing with arrays.

Q13. What is an ndarray?

Ans: An N-dimensional array holding items of one type.

Q14. What is vectorization?

Ans: Operating on a whole array without loops.

Q15. What is broadcasting?

Ans: Stretching a smaller array to fit a larger one in operations.

Q16. What is SciPy used for?

Ans: Scientific computing like statistics, optimisation, integration.

Q17. Which library is used for plots?

Ans: Matplotlib (pyplot).

Q18. What are Series and DataFrame?

Ans: Series is a 1-D column; DataFrame is a 2-D table.

Q19. Which function reads a CSV file?

Ans: pd.read_csv().

Q20. Which function saves a NumPy array?

Ans: np.save() saves it as a .npy file.

SARTHI EDUCATION Learn • Prepare • Achieve

Important University Questions (Long Answer)

Find the full answer on the page given in brackets.

1. What are the facets of data? Explain. (10) [page 2]
2. Differentiate structured and unstructured data. (5) [page 3]
3. Explain natural language data and NLP. (5) [page 4]
4. Explain machine-generated data and graph data. (5) [page 5]
5. Explain audio, image, video and streaming data. (5) [page 6]
6. Explain the data science process overview. (5/10) [page 7]
7. Explain data preparation and data exploration. (5) [page 8]
8. Explain building the models. (5) [page 9]
9. Why is Python used? Explain its libraries. (5) [page 11]
10. Explain NumPy and create arrays. (5) [page 12]
11. Explain vectorized computation in NumPy. (10) [page 13]
12. Explain SciPy and Matplotlib. (5) [page 15]
13. Explain Pandas Series and DataFrame. (10) [page 17]
14. Explain data loading, storage and file formats. (10) [page 19]

Most repeated: facets of data, NumPy operations, Pandas DataFrame, file loading.

SARTHI EDUCATION
Learn • Prepare • Achieve

Practice: Find the Output (with answers)

1. *np.array([1, 2, 3, 4]) ** 2*

   Ans: [1 4 9 16]

2. *m = np.array([[1, 2], [3, 4]]); m.sum(axis=0) and m.sum(axis=1)*

   Ans: [4 6] and [3 7]

3. *np.arange(1, 10, 3)*

   Ans: [1 4 7]

4. *a = np.array([5, 8, 2, 9]); a[a > 4]*

   Ans: [5 8 9]

5. *np.linspace(0, 10, 5)*

   Ans: [0. 2.5 5. 7.5 10.]

6. *integrate.quad(lambda x: x, 0, 4) (first value)*

   Ans: 8.0, because the area is $x^2/2$ from 0 to 4

7. *Ages 21, 22, 20 in df['Age']. Find df['Age'].mean()*

   Ans: 21.0

8. *np.zeros((2, 3)).shape*

   Ans: (2, 3)

SARTHI EDUCATION Learn • Prepare • Achieve

Multiple Choice Questions (Part 1)

1. Data in rows and columns is

(a) structured (b) unstructured (c) streaming (d) audio

2. Emails and social media posts are

(a) structured (b) unstructured (c) relational (d) numeric only

3. NLP deals with

(a) images (b) human language (c) sensor logs (d) networks

4. Server logs and sensor readings are

(a) natural language (b) machine-generated (c) graph data (d) audio

5. Social network data is stored as

(a) tables only (b) nodes and edges (c) audio (d) plain text only

6. Neo4j is a

(a) graph database (b) spreadsheet (c) language (d) library

7. Continuous real-time flow of data is

(a) batch data (b) streaming data (c) structured data (d) static data

8. A streaming data example is

(a) stock market ticks (b) printed book (c) old report (d) photo album

9. Computer vision works on

(a) images and video (b) only numbers (c) only text (d) none

10. The first step of the data science process is

(a) build model (b) defining the goal (c) presentation (d) exploration

11. EDA belongs to

(a) data retrieval (b) data exploration (c) modelling (d) deployment

12. Training and test split is part of

(a) defining the goal (b) building the models (c) retrieving data (d) none

Answers: 1-a 2-b 3-b 4-b 5-b 6-a 7-b 8-a 9-a 10-b 11-b 12-b

Multiple Choice Questions (Part 2)

13. NumPy stands for

(a) Numerical Python (b) New Python (c) Number Pandas (d) None

14. The main object of NumPy is

(a) DataFrame (b) ndarray (c) Series (d) tuple

15. The 2-D table structure of Pandas is

(a) Series (b) DataFrame (c) ndarray (d) list

16. Which library is used for plotting?

(a) NumPy (b) Pandas (c) Matplotlib (d) SciPy

17. SciPy is used for

(a) plotting only (b) scientific computing (c) file reading only (d) web design

18. np.array([1, 2, 3]) * 2 gives

(a) [1 2 3 1 2 3] (b) [2 4 6] (c) [3 4 5] (d) error

19. Shape of np.zeros((2, 3)) is

(a) (3, 2) (b) (2, 3) (c) (6,) (d) (2,)

20. Function to read a CSV file in Pandas

(a) load_csv() (b) read_csv() (c) open_csv() (d) get_csv()

21. df.head() shows

(a) last 5 rows (b) first 5 rows (c) column names (d) shape

22. Which function fills missing values?

(a) fillna() (b) plot() (c) head() (d) sort()

23. A NumPy array is saved in .npy format by

(a) np.save() (b) np.write() (c) np.dump() (d) np.keep()

24. Operating on a whole array without a loop is called

(a) iteration (b) vectorization (c) recursion (d) compilation

Answers: 13-a 14-b 15-b 16-c 17-b 18-b 19-b 20-b 21-b 22-a 23-a 24-b

SARTHI EDUCATION Learn • Prepare • Achieve

Last Minute Revision – Unit III

Facets of data (7)

Structured, Unstructured, Natural language, Machine-generated, Graph, Audio-image-video, Streaming

- *Structured = SQL tables. Graph = nodes and edges (Neo4j). Streaming = real time (Kafka).*
- *NLP tasks: sentiment analysis, entity recognition, translation, summarisation.*

Process overview

Goal → Retrieve → Prepare (clean, integrate, transform) → Explore (EDA) → Build models → Present and automate.

Python libraries

- *NumPy: ndarray, vectorization, broadcasting. m[row, column], axis=0 columns, axis=1 rows.*
- *Pandas: Series (1-D), DataFrame (2-D). head, info, describe, loc, iloc, groupby, fillna.*
- *SciPy: stats, optimize, integrate, linalg. Matplotlib: plot, bar, hist, scatter, pie.*

File loading

- *pd.read_csv / read_excel / read_json / read_sql. np.loadtxt, np.genfromtxt, np.save, np.load.*

Remember: write the import lines (import numpy as np, import pandas as pd) in every code answer.

SARTHI EDUCATION Learn • Prepare • Achieve

DATA SCIENCE

UNIT IV

Data Visualization, Correlation and Regression

Syllabus (11 Hrs):

Data Visualization: definition, importance in data science, exploratory data analysis, visualization tools. Chart types: tabular data, dot and line plots, scatter plots, bar plots, histograms and pie charts, visualization with Seaborn. Correlation: correlation coefficient for quantitative data, computational formula, regression line, least squares regression line, standard error of estimate, interpretation of R^2, multiple regression equations, regression towards the mean.

Contents:

Q.1 – Q.4 Visualization, EDA, tools, tabular data ..... page 2

Q.5 – Q.10 Dot, line, scatter, bar, histogram, pie, Seaborn ..... page 6

Q.11 – Q.14 Correlation and numericals ..... page 12

Q.15 – Q.16 Regression line and numericals ..... page 16

Q.17 – Q.20 Standard error, R^2, multiple regression, regression to the mean ..... page 18

Q.21 – Q.22 Differences and Python ..... page 22

Definitions, short and long questions, practice ..... page 24

MCQs, formulas and last minute revision ..... page 29

Ink code: black = headings, blue = answers, red = important points.

Name: _____

Q.1 What is data visualization? Explain its importance in Data Science. [5 Marks]

Definition:

Data visualization is the graphical presentation of data using charts, graphs, maps and dashboards so that patterns, trends and relations can be seen easily.

Importance in Data Science:

(1) Understand big data quickly: a chart is easier than thousands of rows.

(2) Find patterns, trends and outliers that tables hide.

(3) Helps in EDA to choose variables and models.

(4) Communicates results to non-technical people and managers.

(5) Supports faster and better decisions.

(6) Shows the performance of a model (accuracy, errors) visually.

(7) Helps to find mistakes and missing data in the data.

Rules of a good visualization:

- *Choose the right chart for the data and the purpose.*
- *Give a title, axis labels, units and a legend.*
- *Keep it simple: avoid too many colours and decoration.*
- *Start the bar chart axis at zero, so that the picture is honest.*

Steps:

Data → Clean → Choose chart → Draw → Interpret and share

Example:

A table of monthly sales for a year is hard to read, but a line plot shows at once that sales rise in October and November (festival season).

SARTHI EDUCATION Learn • Prepare • Achieve

Q.2 Explain Exploratory Data Analysis (EDA) and its techniques. [10 Marks]

EDA is the study of data using statistics and graphs to understand its structure, find patterns, outliers and relations, and test assumptions before modelling. It was promoted by John Tukey.

Aims of EDA:

- Know the data: size, types of variables, missing values.
- Find relations between variables and detect outliers and errors.
- Decide which variables and models to use.

Steps of EDA:

1. Check the structure: rows, columns, data types.
2. Summary statistics: mean, median, S.D., min, max, quartiles.
3. Find and treat missing values and outliers.
4. Univariate analysis (one variable): histogram, box plot.
5. Bivariate analysis (two variables): scatter plot, correlation.
6. Multivariate analysis: heatmap, pair plot.

Types of EDA:

Non-graphical	Graphical
summary statistics, tables	charts and plots
mean, median, correlation	histogram, box plot, scatter

EARTH EDUCATION Learn • Prepare • Achieve

`df.shape` `df.info()` `df.describe()` `df.isnull().sum()`

`df.hist()` `df.corr()` `sns.pairplot(df)`

EDA = look at the data before building the model.

Q.3 Explain data visualization tools and how to choose a chart. [5 Marks]

Tool	Type	Main use
Matplotlib	Python library	basic static plots
Seaborn	Python library	attractive statistical plots
Plotly, Bokeh	Python libraries	interactive web charts
Tableau	BI software	drag-and-drop dashboards
Power BI	Microsoft BI tool	reports and dashboards
Excel	spreadsheet	quick charts
Looker Studio	Google web tool	free dashboards
ggplot2	R library	layered graphics
D3.js	JavaScript library	custom web visuals

Choosing the right chart:

Purpose	Chart to use
Compare categories	bar chart
Trend over time	line plot
Distribution of one variable	histogram, box plot
Relation of two variables	scatter plot
Parts of a whole	pie chart
Many variables together	heatmap, pair plot

Tableau and Power BI need little coding. Matplotlib and Seaborn give full control with Python code.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.4 Explain tabular data as a way of presenting data. [5 Marks]

Tabular data is data arranged in rows and columns. Each row is a record (observation) and each column is a variable. It is the simplest form of presentation and the base of all charts.

Student	Marks	Result
Asha	78	Pass
Ravi	35	Fail
Meena	62	Pass
Kiran	88	Pass

Frequency table:

It shows how many times each value occurs.

Result	Frequency	Percent
Pass	3	75%
Fail	1	25%
Total	4	100%

Uses and limits:

- *Good for exact values and small data. Easy to make and read.*
- *Difficult to see patterns when data is large, so charts are used.*

```
df['Result'].value_counts()
pd.crosstab(df['Class'], df['Result']) (cross table)
df.pivot_table(values='Marks', index='Class', aggfunc='mean')
```

Tip: use a table when exact numbers are needed, a chart when the pattern is needed.

SARTHI EDUCATION Learn • Prepare • Achieve

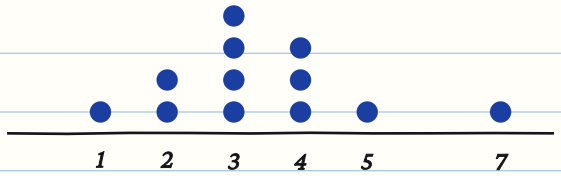
Q.5 Explain dot plots and line plots with examples. [5 Marks]

Dot plot:

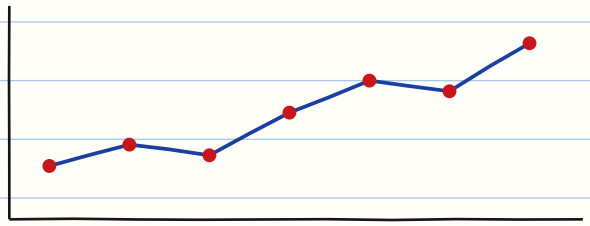
A dot plot shows each data value as a dot above a number line. Equal values are stacked. It is good for small data to see the shape, clusters and outliers.

Line plot:

A line plot joins data points by lines, usually with time on the x-axis. It shows the trend (rise, fall, repeat).



Dot plot



Line plot (trend over time)

Python:

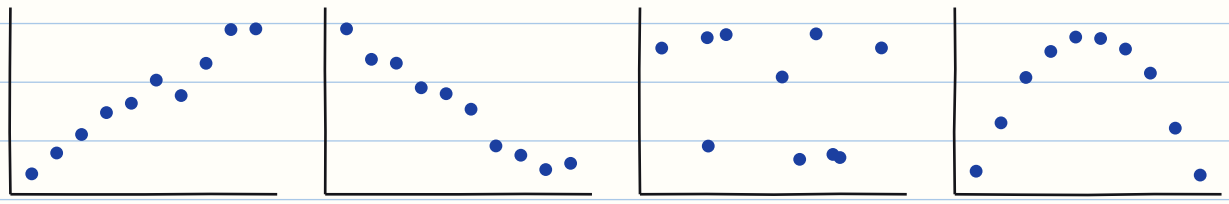
```
plt.plot(months, sales, marker='o')  # line plot
plt.plot(x, y, 'o'); plt.yticks([]) # simple dots
sns.stripplot(x=data) or sns.swarmplot(x=data) # dot style plot
```

Use line plots for time series, dot plots for small data.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.6 Explain scatter plots and the patterns they show. [5 Marks]

A scatter plot (scatter diagram) shows the relation between two numerical variables. Each point is one observation with x and y coordinates. It is the first step in correlation and regression.



What we read from it:

- Positive: both variables rise together (hours studied and marks).
- Negative: one rises and the other falls (speed and travel time).
- No correlation: points are spread randomly.
- Non-linear: points follow a curve. Outliers are points far from the others.

```
plt.scatter(x, y); plt.xlabel('Hours');plt.ylabel('Marks')
sns.scatterplot(x='Hours', y='Marks', data=df)
```

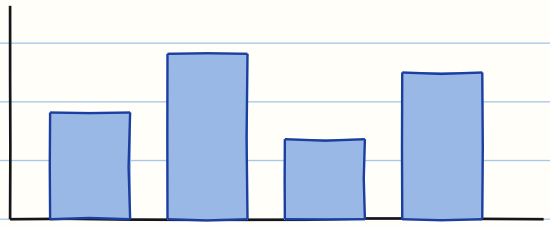
Remember: scatter plot = relation of two variables.

SARTHI EDUCATION Learn • Prepare • Achieve

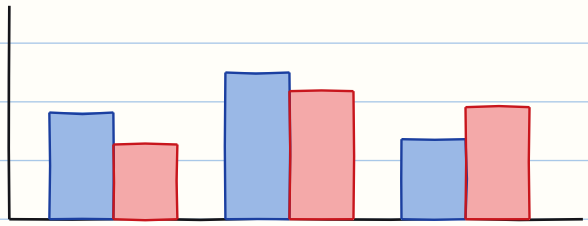
Q.7 Explain bar plots (types and uses). [5 Marks]

A bar plot uses bars of equal width whose length shows the value of each category.

There are gaps between the bars because categories are separate.



Simple bar



Grouped bar (two groups)

Types:

- Vertical and horizontal bar plots.
- Grouped (clustered): bars of groups side by side for comparison.
- Stacked: bars of groups put on top of each other to show the total.

Uses:

Comparing sales of different cities, marks of different subjects, counts of categories.

```
plt.bar(names, values) plt.barh(names, values)
sns.barplot(x='City', y='Sales', data=df) sns.countplot(x='Result', data=df)
```

Bars should start from zero. Use a bar plot for categories, not for continuous data.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.8 Explain histograms and pie charts with examples. [10 Marks]

Histogram:

A histogram shows the distribution of one continuous variable. Data is grouped into bins (class intervals) and the height of each bar is the frequency. The bars touch each other because the data is continuous.

- Shows shape: symmetric, skewed, or two peaks; centre and spread; outliers.

plt.hist(data, bins=5) sns.histplot(data, bins=5, kde=True)

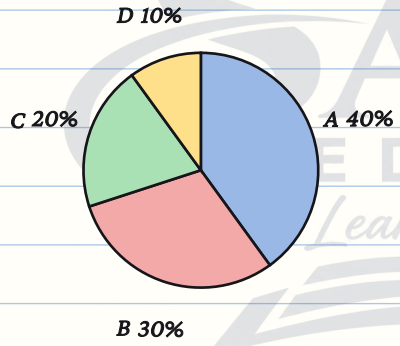
Pie chart:

A pie chart is a circle divided into slices. Each slice shows the share (part) of a whole.

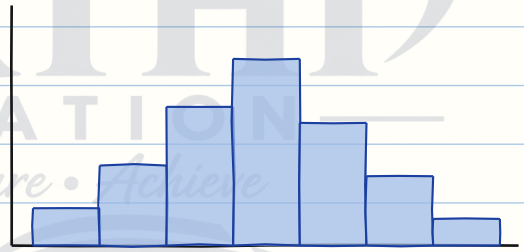
Angle of a slice:

$$\text{Angle} = \frac{\text{Value}}{\text{Total}} \times 360°$$

Example: marks share A = 40, B = 30, C = 20, D = 10 (total 100). Angles = 144°, 108°, 72°, 36°.



Pie chart



Histogram (bars touch)

plt.pie(sizes, labels=labels, autopct='%1.1f%%')

Q.9 Explain visualization with Seaborn. [10 Marks]

Seaborn is a Python library for statistical graphics. It is built on Matplotlib and works directly with Pandas DataFrames. It gives attractive default styles and colours with less code.

Advantages:

- *Beautiful plots with short code; works with DataFrame column names.*
- *Special plots for statistics: box plot, violin plot, heatmap, pair plot, regression line.*

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.set_theme()
df = sns.load_dataset('tips')  (sample dataset)
```

Plot	Function	Use
Scatter	sns.scatterplot(x=,y=, data=)	relation
Line	sns.lineplot()	trend
Bar	sns.barplot()	mean per category
Count	sns.countplot()	count of categories
Histogram	sns.histplot(kde=True)	distribution
Box	sns.boxplot()	median, outliers
Violin	sns.violinplot()	shape and spread
Heatmap	sns.heatmap(df.corr(), annot=True)	correlation matrix
Pair plot	sns.pairplot(df)	all pairs of variables
Regression	sns.regplot(x=, y=, data=)	scatter with line

```
sns.scatterplot(x='total_bill', y='tip', data=df); plt.show()
```

Matplotlib gives control, Seaborn gives beauty and statistics.

Q.10 Differences between common charts and libraries. [5 Marks]

Basis	Bar chart	Histogram
Data	categorical	continuous numerical
Bars	gaps between bars	bars touch
Order	can be rearranged	fixed order of bins
Shows	comparison	distribution

Basis	Line plot	Scatter plot
Points	joined by lines	separate dots
X-axis	usually time	numerical variable
Shows	trend	relation, correlation

Basis	Matplotlib	Seaborn
Level	low level, full control	high level, easy
Code	more lines	fewer lines
Looks	plain default	attractive default
Based on	NumPy	Matplotlib and Pandas
Best for	custom charts	statistical plots

Use the chart that fits the question: category → bar, time → line, relation → scatter, distribution → histogram.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.11 What is correlation? Explain its types and the correlation coefficient. [5 Marks]

Correlation is the statistical measure of the strength and direction of the linear relationship between two quantitative variables, such as hours studied and marks.

Types:

- *Positive: both increase together (height and weight).*
- *Negative: one increases while the other decreases (price and demand).*
- *Zero: no linear relation.*

Correlation coefficient (r):

It is a number between −1 and +1 that measures correlation. Pearson's coefficient is used for quantitative data.

Value of r	Meaning
r = +1	perfect positive correlation
0.7 to 0.99	strong positive
0.3 to 0.7	moderate positive
near 0	no or very weak linear relation
−0.3 to −0.7	moderate negative
−0.7 to −0.99	strong negative
r = −1	perfect negative correlation

Correlation is not causation: two variables moving together does not mean one causes the other (ice cream sales and drowning both rise in summer).

Scatter plot is used to see the correlation first.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.12 Write the formulas of the correlation coefficient (definition and computational). [5 Marks]

Definition formula (Karl Pearson):

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{[\,\Sigma(x - \bar{x})^2 \times \Sigma(y - \bar{y})^2\,]}}$$

$$r = Cov(x, y) \,/\, (\sigma_x\, \sigma_y)$$

Computational formula (used for numericals):

$$r = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{\sqrt{\{\,[n\,\Sigma x^2 - (\Sigma x)^2]\,[n\,\Sigma y^2 - (\Sigma y)^2]\,\}}}$$

Meaning of symbols:

Symbol	Meaning
n	number of pairs of observations
Σxy	sum of products of x and y
Σx, Σy	sums of x and of y
Σx^2, Σy^2	sums of squares of x and y
$(\Sigma x)^2$	square of the sum of x
σ_x, σ_y	standard deviations of x and y

Steps:

1. Make columns x, y, xy, x^2, y^2 and find the totals.
2. Put the totals in the formula.
3. Calculate and interpret r.

Properties:

- $-1 \le r \le +1$. It has no unit. It does not change if every x is increased by the same amount or multiplied by a positive number.
- r is symmetric: $r(x, y) = r(y, x)$.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.13 Numerical on correlation coefficient: Example 1. [10 Marks]

Find the correlation coefficient between hours studied (x) and marks scored in a test (y):

x: 1, 2, 3, 4, 5 y: 2, 4, 5, 4, 5

x	y	xy	x^2	y^2
1	2	2	1	4
2	4	8	4	16
3	5	15	9	25
4	4	16	16	16
5	5	25	25	25
$\Sigma x = 15$	$\Sigma y = 20$	$\Sigma xy = 66$	$\Sigma x^2 = 55$	$\Sigma y^2 = 86$

Here $n = 5$. Put the values in the formula:

$$r = (5 \times 66 - 15 \times 20) / \sqrt{[\,(5 \times 55 - 15^2)(5 \times 86 - 20^2)\,]}$$

$$r = (330 - 300) / \sqrt{[\,(275 - 225)(430 - 400)\,]} = 30 / \sqrt{(50 \times 30)}$$

$$r = 30 / \sqrt{1500} = 30 / 38.73 = 0.775.$$

Interpretation: $r = 0.775$ shows a strong positive correlation. More hours of study go with higher marks.

Note: $R^2 = r^2 = 0.775^2 = 0.60$, so 60% of the variation in marks is explained by study hours.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.14 Numericals on correlation coefficient: Examples 2 and 3. [10 Marks]

Example 2 (negative correlation):

x (price) = 2, 4, 6, 8, 10 and y (demand) = 9, 7, 5, 3, 1.

$n = 5$, $\Sigma x = 30$, $\Sigma y = 25$, $\Sigma xy = 18 + 28 + 30 + 24 + 10 = 110$, $\Sigma x^2 = 220$, $\Sigma y^2 = 81 + 49 + 25 + 9 + 1 = 165$.

Numerator $= 5 \times 110 - 30 \times 25 = 550 - 750 = -200$.

Denominator $= \sqrt{[(5 \times 220 - 900)(5 \times 165 - 625)]} = \sqrt{(200 \times 200)} = 200$.

$r = -200/200 = -1$. Perfect negative correlation: when price rises, demand falls exactly.

Example 3 (strong positive):

x (hours) = 2, 3, 5, 7, 8 and y (marks) = 40, 50, 65, 80, 85.

x	y	xy	x^2	y^2
2	40	80	4	1600
3	50	150	9	2500
5	65	325	25	4225
7	80	560	49	6400
8	85	680	64	7225
25	320	1795	151	21950

Numerator $= 5 \times 1795 - 25 \times 320 = 8975 - 8000 = 975$.

Denominator $= \sqrt{[(5 \times 151 - 625)(5 \times 21950 - 320^2)]} = \sqrt{(130 \times 7350)} = \sqrt{955500} = 977.5$.

$r = 975/977.5 = 0.997$. A very strong positive correlation.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.15 What is regression? Explain the least squares regression line. [10 Marks]

Regression is a method to find the relation between variables and to predict the value of the dependent variable (y) from the independent variable (x).

Regression line of y on x:

$$y = a + b\,x$$

y = dependent variable, x = independent variable, a = intercept (value of y when x = 0), b = slope (change in y for one unit change in x).

Method of least squares:

Many lines can be drawn through the points. The least squares line is the line for which the sum of squares of errors (vertical distances between actual y and predicted ŷ) is the smallest: minimise $\Sigma(y - \hat{y})^2$.

Formulas for a and b:

$$b = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2}$$

$$a = \bar{y} - b\,\bar{x}$$

Another form: $b = r \times \sigma_y / \sigma_x$.

Properties:

- *The line passes through the point* $(\bar{x}, \bar{y})$.
- *The sum of errors* $\Sigma(y - \hat{y}) = 0$.
- *There are two lines: y on x (predict y) and x on y (predict x).*

Interpretation:

If b = 0.6, then for each extra unit of x, y increases by 0.6 on average. A negative b means y falls as x rises.

Q.16 Numericals on the regression line. [10 Marks]

Example 1:

Using the data x: 1, 2, 3, 4, 5 and y: 2, 4, 5, 4, 5 ($\Sigma x = 15$, $\Sigma y = 20$, $\Sigma xy = 66$, $\Sigma x^2 = 55$, $n = 5$). Fit $y = a + bx$ and predict y for $x = 6$.

$b = (5 \times 66 - 15 \times 20)/(5 \times 55 - 15^2) = 30/50 = 0.6$

$\bar{x} = 3$, $\bar{y} = 4$, so $a = 4 - 0.6 \times 3 = 2.2$. Line: $y = 2.2 + 0.6\,x$.

For $x = 6$: $y = 2.2 + 0.6 \times 6 = 5.8$.

Example 2:

Hours x: 2, 3, 5, 7, 8 and marks y: 40, 50, 65, 80, 85 ($n = 5$, $\Sigma x = 25$, $\Sigma y = 320$, $\Sigma xy = 1795$, $\Sigma x^2 = 151$).

$b = (5 \times 1795 - 25 \times 320)/(5 \times 151 - 25^2) = 975/130 = 7.5$

$\bar{x} = 5$, $\bar{y} = 64$, so $a = 64 - 7.5 \times 5 = 26.5$. Line: $y = 26.5 + 7.5\,x$.

A student studies 6 hours: $y = 26.5 + 7.5 \times 6 = 71.5$ marks.

Meaning: each extra hour of study adds about 7.5 marks.

Example 3 (regression of x on y for Example 1 data):

$b_{xy} = (n\Sigma xy - \Sigma x\Sigma y)/(n\Sigma y^2 - (\Sigma y)^2) = 30/(430 - 400) = 30/30 = 1.$

$a' = \bar{x} - b_{xy}\,\bar{y} = 3 - 1 \times 4 = -1$. Line: $x = -1 + 1 \times y$.

Check: $r^2 = b_{yx} \times b_{xy} = 0.6 \times 1 = 0.6$, and $r = \sqrt{0.6} = 0.775$ (same as before).

Do not use the x on y line to predict y.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.17 Explain the standard error of estimate with an example. [5 Marks]

The standard error of estimate (S_e) measures how far the actual values of y are from the values predicted by the regression line. It is the standard deviation of the errors (residuals). Small S_e means good predictions.

$$S_e = \frac{\sqrt{\Sigma(y - \hat{y})^2}}{\sqrt{(n - 2)}}$$

Short form: $\Sigma(y - \hat{y})^2 = \Sigma y^2 - a\Sigma y - b\Sigma xy$. We divide by $n - 2$ because two values (a and b) are estimated from the data.

Example (data of Example 1, line $\hat{y} = 2.2 + 0.6x$):

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	2	2.8	−0.8	0.64
2	4	3.4	0.6	0.36
3	5	4.0	1.0	1.00
4	4	4.6	−0.6	0.36
5	5	5.2	−0.2	0.04
			$\Sigma = 0$	2.40

$S_e = \sqrt{(2.40/(5 - 2))} = \sqrt{0.8} = 0.894$.

Check with short form: $86 - 2.2 \times 20 - 0.6 \times 66 = 86 - 44 - 39.6 = 2.4$ ☐

Meaning: the predicted marks are wrong by about 0.9 marks on average.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.18 Explain the interpretation of R-squared (R^2). [5 Marks]

R^2 (coefficient of determination) is the fraction of the total variation in y that is explained by the regression model. Its value is between 0 and 1.

$$R^2 = \frac{SSR}{SST}$$

$$R^2 = 1 - SSE/SST$$

In simple regression: $R^2 = r^2$

Term	Meaning	Formula
SST	total variation in y	$\Sigma(y - \bar{y})^2$
SSR	explained by the line	$\Sigma(\hat{y} - \bar{y})^2$
SSE	unexplained (error)	$\Sigma(y - \hat{y})^2$

SST = SSR + SSE.

Example (Example 1 data):

$SST = \Sigma y^2 - n\bar{y}^2 = 86 - 5 \times 16 = 6$. SSE = 2.4. SSR = 6 − 2.4 = 3.6.

$R^2 = 3.6/6 = 0.60$ (also $r^2 = 0.775^2 = 0.60$).

Meaning: 60% of the variation in marks is explained by study hours. The other 40% is due to other factors.

Interpretation guide:

- *R^2 close to 1 → model fits very well. R^2 close to 0 → model explains very little.*
- *In multiple regression use adjusted R^2, because R^2 always rises when a variable is added.*
- *A high R^2 does not prove causation, and a low R^2 can still be useful in some fields.*

SARTHI EDUCATION Learn • Prepare • Achieve

Q.19 Explain multiple regression and its equation. [5 Marks]

Multiple regression predicts a dependent variable y using two or more independent variables.

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + e$$

b_0 = intercept, b_1, b_2, ... = regression coefficients, e = error. Each b_i shows the change in y for one unit change in x_i when the other variables are kept constant.

Example:

Salary (in thousand rupees) = 10 + 2.5 × (years of experience) + 1.5 × (years of education).

For a person with 4 years of experience and 16 years of education: Salary = 10 + 2.5 × 4 + 1.5 × 16 = 10 + 10 + 24 = 44 thousand.

Points:

- Coefficients are found by least squares using software.
- Check R^2, adjusted R^2, p-values of the coefficients ($p < 0.05$ means the variable is significant).
- Avoid multicollinearity (independent variables highly correlated with each other).
- Category variables are used as dummy variables (0/1).

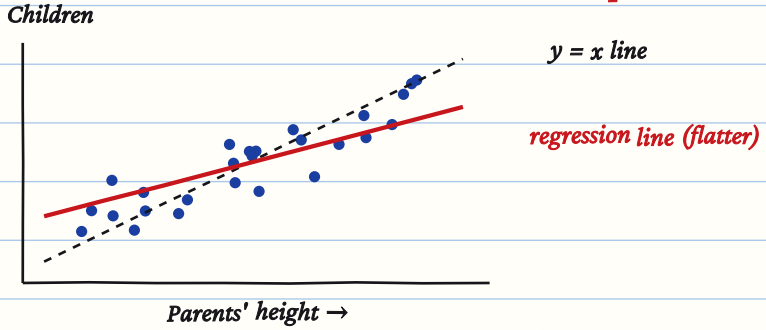
Python:

```
from sklearn.linear_model import LinearRegression
model = LinearRegression().fit(X,y) # X has many columns
model.coef_ model.intercept_ model.score(X, y) (R squared)
```

Simple regression: one x. Multiple regression: many x.

Q.20 What is regression towards the mean? Explain with an example. [5 Marks]

Regression towards the mean means that an extreme value on one measurement is usually followed by a value closer to the average on the next measurement. It was observed by Francis Galton in the heights of parents and children: very tall parents had children who were tall but shorter than the parents.



Why does it happen?

Because the correlation r is less than 1, the regression line is flatter than the 45° line.

Predicted z-score of $y = r \times$ z-score of x.

Example:

Mean height = 170 cm, S.D. = 10 cm, $r = 0.5$. A father is 190 cm ($z = 2$). Expected son $z = 0.5 \times 2 = 1$, so the son is expected to be $170 + 1 \times 10 = 180$ cm, shorter than 190 but still above average.

Another example: students with the highest marks in test 1 usually score a little lower in test 2.

Do not think that the decrease is due to some cause. It is a statistical effect (regression fallacy).

SARTHI EDUCATION Learn • Prepare • Achieve

Q.21 Differences between correlation and regression; properties of regression coefficients. [5 Marks]

Basis	Correlation	Regression
Meaning	strength and direction of relation	prediction of one variable from other
Variables	both are treated equal	dependent and independent
Result	one number r (−1 to +1)	an equation $y = a + bx$
Purpose	measure the relation	estimate and predict
Symmetry	$r(x, y) = r(y, x)$	line y on x ≠ line x on y
Unit	no unit	has units of y per x

Properties of regression coefficients:

(1) Both b_{yx} and b_{xy} have the same sign as r.

(2) $r^2 = b_{yx} \times b_{xy}$ so $r = \pm\sqrt{(b_{yx} \times b_{xy})}$.

(3) Both regression lines pass through the point $(\bar{x}, \bar{y})$.

(4) If $r = \pm 1$ both lines are the same; if $r = 0$ the lines are at right angles.

(5) $b_{yx} = r\,\sigma_y/\sigma_x$ and $b_{xy} = r\,\sigma_x/\sigma_y$.

Example:

$b_{yx} = 0.9$ and $b_{xy} = 0.4$. Then $r = \sqrt{(0.9 \times 0.4)} = \sqrt{0.36} = 0.6$ (positive, as both b are positive).

Correlation tells us how strong, regression tells us how much.

SARTHI EDUCATION Learn • Prepare • Achieve

Q.22 Correlation and regression in Python. [5 Marks]

```
import numpy as np
from scipy import stats
x = np.array([1, 2, 3, 4, 5])
y = np.array([2, 4, 5, 4, 5])
```

Code	Output
np.corrcoef(x, y)[0, 1]	0.775 (correlation r)
stats.pearsonr(x, y)	r = 0.775 and a p-value
res = stats.linregress(x,y)	fits y = a + bx
res.slope	0.6
res.intercept	2.2
res.rvalue ** 2	0.6 (R squared)

With Pandas and scikit-learn:

```
df.corr() # correlation matrix
from sklearn.linear_model import LinearRegression
model = LinearRegression().fit(x.reshape(-1,1), y)
model.coef_ → [0.6] model.intercept_ → 2.2
model.predict([[6]]) → [5.8]
```

With Seaborn:

```
sns.heatmap(df.corr(), annot=True) # shows all correlations
sns.regplot(x='x',y='y', data=df) # scatter plot with regression line
```

The Python outputs match the hand calculation: slope 0.6, intercept 2.2, r = 0.775, R^2 = 0.6.

SARTHI EDUCATION Learn • Prepare • Achieve

Important Definitions (Unit IV)

- **Data visualization**: presenting data in graphs and charts.
- **EDA**: studying data with statistics and graphs before modelling.
- **Dot plot**: shows each value as a dot above a number line.
- **Line plot**: points joined by lines, to show a trend over time.
- **Scatter plot**: dots showing the relation of two numerical variables.
- **Bar plot**: bars showing values of categories.
- **Histogram**: bars showing frequency distribution of continuous data.
- **Pie chart**: circle divided into slices showing parts of a whole.
- **Seaborn**: Python library for statistical graphics built on Matplotlib.
- **Correlation**: strength and direction of the linear relation of two variables.
- **Correlation coefficient (r)**: value from −1 to +1 measuring correlation.
- **Regression**: method to predict one variable from other variables.
- **Least squares line**: line with the smallest sum of squared errors.
- **Standard error of estimate**: S.D. of the errors around the regression line.
- **R^2**: fraction of variation in y explained by the model.
- **Multiple regression**: regression with two or more independent variables.
- **Regression towards the mean**: extreme values tend to be followed by values nearer the average.

SARTHI EDUCATION Learn • Prepare • Achieve

Short Answer Questions (2 / 3 Marks) – Part 1

Q1. Define data visualization.

Ans: Presenting data using graphs and charts to see patterns.

Q2. Why is visualization important?

Ans: It shows patterns and outliers and explains results easily.

Q3. What is EDA?

Ans: Studying data using statistics and graphs before modelling.

Q4. Name four visualization tools.

Ans: Matplotlib, Seaborn, Tableau and Power BI.

Q5. Which chart shows a trend over time?

Ans: Line plot.

Q6. Which chart shows relation of two variables?

Ans: Scatter plot.

Q7. Difference between bar chart and histogram?

Ans: Bar chart is for categories (gaps); histogram is for continuous data (bars touch).

Q8. Write the angle formula for a pie slice.

Ans: Angle = (value/total) × 360°.

Q9. What is Seaborn?

Ans: A Python library for statistical plots, built on Matplotlib.

Q10. Name the Seaborn function for a heatmap.

Ans: sns.heatmap().

PARTH EDUCATION Learn • Prepare • Achieve

Short Answer Questions (2 / 3 Marks) – Part 2

Q11. What is correlation?

Ans: Strength and direction of the linear relation between two variables.

Q12. State the range of r.

Ans: –1 to +1.

Q13. What does r = 0 mean?

Ans: No linear correlation.

Q14. Write the computational formula of r.

Ans: $r = (n\Sigma xy - \Sigma x\Sigma y)/\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}$.

Q15. Write the regression line equation.

Ans: $y = a + bx$.

Q16. Write the formulas of b and a.

Ans: $b = (n\Sigma xy - \Sigma x\Sigma y)/(n\Sigma x^2 - (\Sigma x)^2)$ *and* $a = \bar{y} - b\,\bar{x}$.

Q17. What does the least squares method do?

Ans: Minimises the sum of squared errors.

Q18. What is the standard error of estimate?

Ans: S.D. of errors around the regression line, $\sqrt{[\Sigma(y - \hat{y})^2/(n - 2)]}$.

Q19. What does $R^2 = 0.8$ mean?

Ans: 80% of the variation in y is explained by the model.

Q20. What is regression towards the mean?

Ans: Extreme values tend to be followed by values closer to the average.

Important University Questions (Long Answer)

Find the full answer on the page given in brackets.

1. What is data visualization? Explain its importance. (5) [page 2]
2. Explain EDA and its techniques. (10) [page 3]
3. Explain visualization tools and how to choose a chart. (5) [page 4]
4. Explain dot plots and line plots. (5) [page 6]
5. Explain scatter plots and their patterns. (5) [page 7]
6. Explain bar plots, histograms and pie charts. (5/10) [page 9]
7. Explain visualization with Seaborn. (10) [page 10]
8. What is correlation? Write the formulas of r. (5) [page 13]
9. Numerical on correlation coefficient. (10) [page 14]
10. Explain the least squares regression line. (10) [page 16]
11. Numerical on regression line. (10) [page 17]
12. Explain standard error of estimate. (5) [page 18]
13. Explain the interpretation of R^2. (5) [page 19]
14. Explain multiple regression. (5) [page 20]
15. Explain regression towards the mean. (5) [page 21]
16. Differences between correlation and regression. (5) [page 22]

Most repeated: numericals on r and regression line, R^2, chart types, Seaborn.

SARTHI EDUCATION Learn • Prepare • Achieve

Practice Numericals with Answers

1. $x = 1, 2, 3$ and $y = 2, 4, 6$. Find r.

   Ans: points lie on a line, $r = +1$

2. $x = 1, 2, 3, 4$ and $y = 4, 3, 2, 1$. Find r.

   Ans: $r = -1$ (perfect negative)

3. $n = 5, \Sigma x = 20, \Sigma y = 30, \Sigma xy = 130, \Sigma x^2 = 90, \Sigma y^2 = 200$. Find r.

   Ans: num $= 650 - 600 = 50$; den $= \sqrt{(50 \times 100)} = 70.71$; $r = 0.707$

4. $x = 1, 2, 3, 4, 5$ and $y = 3, 5, 7, 9, 11$. Fit $y = a + bx$. Predict y at $x = 6$.

   Ans: $b = 2$, $a = 1$, $y = 1 + 2x$; $y(6) = 13$

5. $r = 0.8$. Find R^2 and interpret.

   Ans: $R^2 = 0.64$; 64% of variation in y is explained by x

6. $\Sigma(y - \hat{y})^2 = 20$ and $n = 7$. Find the standard error of estimate.

   Ans: $S_e = \sqrt{(20/5)} = 2$

7. $b_yx = 0.9$ and $b_xy = 0.4$. Find r.

   Ans: $r = \sqrt{0.36} = 0.6$

8. $r = 0.6$. A father has $z = 1.5$. Find the expected z of the son.

   Ans: $z = 0.6 \times 1.5 = 0.9$

9. A pie chart has $A = 25$, $B = 50$, $C = 25$. Find the angles.

   Ans: 90°, 180°, 90°

SARTHI EDUCATION Learn • Prepare • Achieve

Multiple Choice Questions (Part 1)

1. Data visualization means

(a) graphical presentation of data (b) storing data (c) deleting data (d) typing data

2. Best chart for the relation of two numeric variables

(a) pie chart (b) scatter plot (c) table (d) none

3. A pie chart shows

(a) trend over time (b) relation (c) parts of a whole (d) outliers

4. A histogram shows

(a) distribution of a numeric variable (b) parts of a whole (c) a map (d) a table

5. A line plot is best for

(a) categories (b) trend over time (c) parts (d) none

6. Seaborn is built on

(a) NumPy only (b) Matplotlib (c) SciPy only (d) Excel

7. Seaborn function for a heatmap

(a) sns.heatmap (b) sns.pie (c) sns.table (d) sns.cake

8. EDA was promoted by

(a) Karl Pearson (b) John Tukey (c) Galton (d) Fisher

9. Tableau and Power BI are

(a) visualization tools (b) databases (c) languages (d) viruses

10. In a bar chart the bars are

(a) separated by gaps (b) always touching (c) circles (d) lines

11. Bars of a histogram touch because

(a) data is continuous (b) data is categorical (c) it is a pie (d) no reason

12. Pie slice angle is

(a) (value/total) × 360 (b) value × total (c) total/value (d) value + 360

Answers: 1-a 2-b 3-c 4-a 5-b 6-b 7-a 8-b 9-a 10-a 11-a 12-a

Multiple Choice Questions (Part 2)

13. Range of the correlation coefficient

(a) 0 to 1 (b) −1 to +1 (c) −∞ to ∞ (d) 0 to 100

14. r = 0 means

(a) no linear relation (b) perfect positive (c) perfect negative (d) error

15. The correlation coefficient was given by

(a) Galton (b) Karl Pearson (c) Tukey (d) Gauss

16. Regression line equation

(a) $y = a + bx$ (b) $y = ab$ (c) $y = a/b$ (d) $y = a - x$

17. Least squares method minimises

(a) sum of errors (b) sum of squared errors (c) number of points (d) none

18. $R^2 = 0.64$ means

(a) 36% is explained (b) 64% of variation in y is explained (c) $r = 0.64$ (d) no relation

19. In simple regression R^2 equals

(a) r (b) r^2 (c) $1 - r$ (d) $\sqrt{r}$

20. Multiple regression has

(a) one x (b) more than one independent variable (c) no x (d) only y

21. Regression towards the mean was observed by

(a) Galton (b) Tukey (c) Pearson (d) Fisher

22. Standard error of estimate measures

(a) spread around the regression line (b) mean (c) median (d) mode

23. If r = −0.9 the correlation is

(a) weak positive (b) strong negative (c) zero (d) perfect positive

24. Correlation shows

(a) causation (b) association, not causation (c) proof (d) none

Answers: 13-b 14-a 15-b 16-a 17-b 18-b 19-b 20-b 21-a 22-a 23-b 24-b

Important Formulas – Unit IV

Topic	Formula
Pearson correlation	$r = \Sigma(x - \bar{x})(y - \bar{y}) / \sqrt{[\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2]}$
Computational r	$r = (n\Sigma xy - \Sigma x\Sigma y) / \sqrt{\{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]\}}$
Regression line	$y = a + bx$
Slope b (y on x)	$b = (n\Sigma xy - \Sigma x\Sigma y) / (n\Sigma x^2 - (\Sigma x)^2)$
Intercept a	$a = \bar{y} - b\,\bar{x}$
b in terms of r	$b = r\,\sigma_y / \sigma_x$
Regression of x on y	$b_{xy} = (n\Sigma xy - \Sigma x\Sigma y) / (n\Sigma y^2 - (\Sigma y)^2)$
Relation	$r^2 = b_{yx} \times b_{xy}$
Standard error of estimate	$S_e = \sqrt{[\Sigma(y - \hat{y})^2 / (n - 2)]}$
Short form of SSE	$\Sigma(y - \hat{y})^2 = \Sigma y^2 - a\Sigma y - b\Sigma xy$
R squared	$R^2 = SSR/SST = 1 - SSE/SST = r^2$
Multiple regression	$y = b_0 + b_1x_1 + \ldots + b_kx_k$
Regression to the mean	$z_y = r \times z_x$
Pie slice angle	$(value/total) \times 360°$

Interpretation: |r| near 1 = strong, near 0 = weak. $R^2 \times 100$ = percent of variation explained.

SARTHI EDUCATION
Learn • Prepare • Achieve

Last Minute Revision – Unit IV

Data visualization

- Chart choice: categories → bar, time → line, distribution → histogram, relation → scatter, parts → pie.
- Bar chart: gaps between bars. Histogram: bars touch.
- Seaborn: built on Matplotlib, works with Pandas. heatmap, pairplot, boxplot, regplot.
- EDA = statistics + graphs before the model (John Tukey).

Correlation

- r from -1 to $+1$. Use the computational formula. Correlation is not causation.
- Steps: make columns x, y, xy, x^2, y^2, total, put in the formula, interpret.

Regression

- $y = a + bx$, $b = (n\Sigma xy - \Sigma x\Sigma y)/(n\Sigma x^2 - (\Sigma x)^2)$, $a = \bar{y} - b\,\bar{x}$.
- Least squares: smallest sum of squared errors. Line passes through $(\bar{x}, \bar{y})$.
- $S_e = \sqrt{[\Sigma(y - \hat{y})^2/(n - 2)]}$. $R^2 = r^2$ (simple regression).
- Multiple regression: many x. Use adjusted R^2 and watch multicollinearity.
- Regression to the mean: $z_y = r\, z_x$, so extreme values move toward the average.

Write the formula, put the values, calculate, then interpret in words, for full marks.

SARTHI EDUCATION
Learn • Prepare • Achieve