UNIT I

INTRODUCTION TO DATASCIENCE – CLASS NOTES

Reg. No:

DATASCIENCE

Data science is a term given to the practice of analysing raw data to discover any hidden patterns.

Components of Datascience

- **Statistics:**

  Statistics is the method of collecting and analyzing numerical data in large quantities to get useful insights.

- **Visualization:**

  Visualization technique helps you to access huge amounts of data in easy to understand and digestible visuals.

- **Machine Learning:**

  Machine Learning explores the building and study of algorithms which learn to make predictions about unforeseen/future data.

Data Science Lifecycle

1. **Discovery:**
   a. Discovery step involves acquiring data from all the identified internal & external sources which helps you to answer the business question.
   b. The data can be:
      i. Logs from webservers
      ii. Data gathered from social media
      iii. Census datasets
      iv. Data streamed from online sources using APIs
2. **Data Preparation:**
   a. Data can have lots of inconsistencies like missing value, blank columns, incorrect data format which needs to be cleaned.
   b. Need to process, explore, and condition data before modeling.
   c. The cleaner the data, the better are the predictions.
3. **Model Planning:**
   a. This stage will help to determine the method and technique to draw the relation between input variables.

UNIT I

INTRODUCTION TO DATASCIENCE – CLASS NOTES

Reg. No:

b. Planning for a model is performed by using different statistical formulas and visualization tools.

4. **Model Building:**
    a. Here, Data scientist distributes datasets for training and testing. Techniques like association, classification, and clustering are applied to the training data set. The model once prepared is tested against the "testing" dataset.
5. **Operationalize:**
    a. In this stage, you deliver the final baselined model with reports, code, and technical documents. Model is deployed into a real-time production environment after thorough testing.
6. **Communicate Results:**
    a. In this stage, the key findings are communicated to all stakeholders. This helps you to decide if the results of the project are a success or a failure based on the inputs from the model.

Properties of Data

Structured Data

- Structured data exists in a predefined format.
- Example
    - Relational database consisting of tables with rows and columns
    - Tables like excel files and Google Docs spreadsheets.
- The programming language SQL (structured query language) is used for managing the structured data.

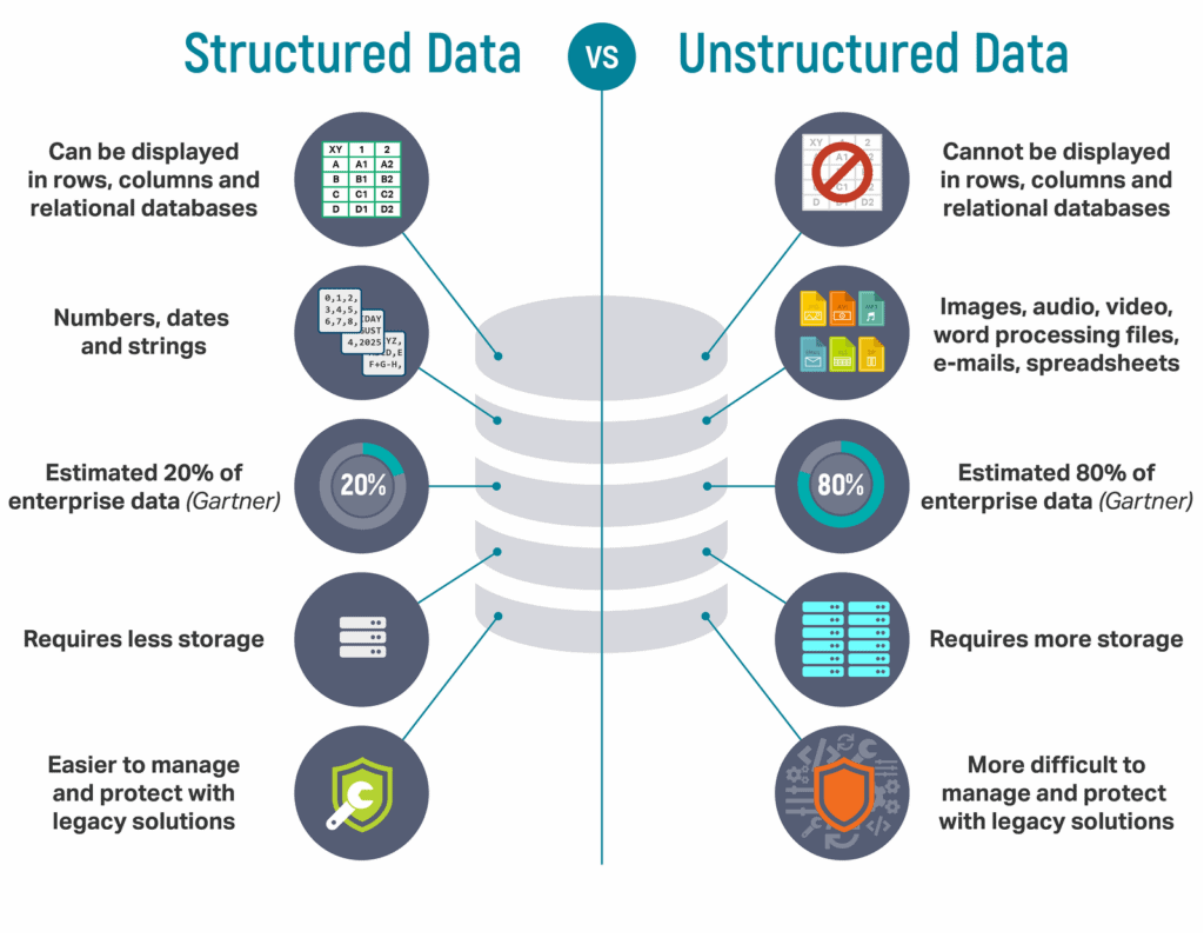
Unstructured Data

- Unstructured data has no predefined format or organization, making it much more difficult to collect, process, and analyze.
    - Text, images, audio and video files.
- It is qualitative in nature and sometimes stored in a non-relational database or NO-SQL.

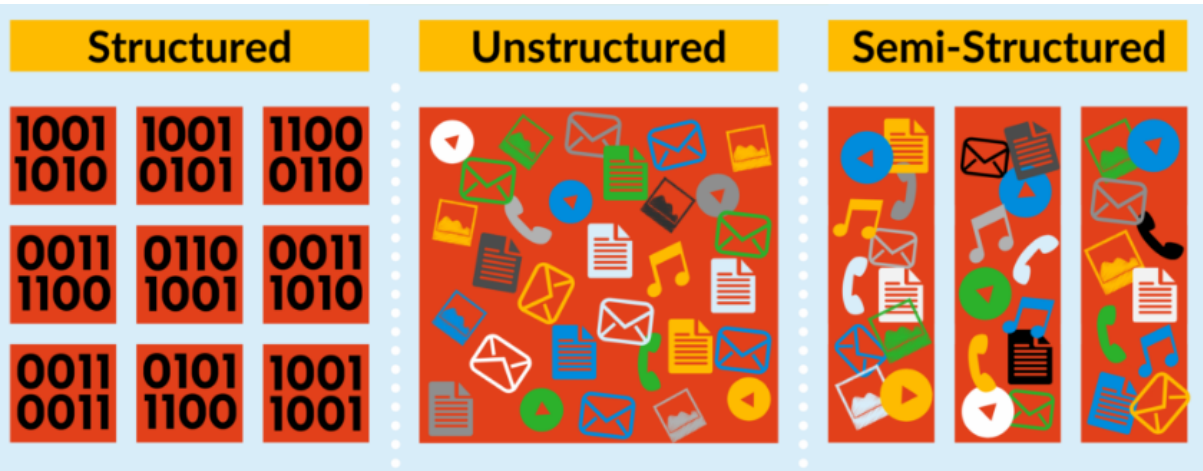
UNIT I

INTRODUCTION TO DATASCIENCE – CLASS NOTES

Reg. No:



Structured/Unstructured/Semi-Structured



Types of Data

- **Qualitative Data Type (Categorical data)**

UNIT I

INTRODUCTION TO DATASCIENCE – CLASS NOTES

Reg. No:

- Values or observations that can be sorted into groups or categories. (Qualitative)
- Bar charts and pie graphs are used to graph categorical data.
- Types of Qualitative Data
    - Nominal
    - Ordinal

- **Quantitative Data Type**
    - This data type tries to quantify things and it does by considering numerical values that make it countable in nature.
    - Types of Quantitative Data
        - Discrete
        - Continuous

Categorical Data	Quantitative Data
• Deals with descriptions.	• Deals with numbers.
• Data can be observed but not measured.	•Data which can be measured.
•Colors, textures, smells, tastes, appearance, beauty, etc.	• Length, height, area, volume, weight, speed, time, cost, age, etc.
• Categorical → Description	• Quantitative → Quantity

Name	Type	Appearance	Examples
Discrete	Numeric	Integers	pairs of shoes, books owned, children
Continuous	Numeric	Decimals	Time spent, speed, weight
Nominal	Categorical	Words no order	Race, shoe color, car type
Ordinal	Categorical	Words with order	Education, happiness

Python Libraries

UNIT I

INTRODUCTION TO DATASCIENCE – CLASS NOTES

Reg. No:

- **NumPy**
    - Provides support for multidimensional arrays with basic operations in them
    - consist of useful linear algebra function
- **SciPy**
    - Provides toolboxes for Signal processing, optimization and statistics.
    - Has plotting library Matplotlib.
    - Has tools for data visualization
- **SCIKIT-Learn**
    - Machine learning libraries built from NumPy, SciPy and MatPlotlib.
    - Efficient tool for data analysis such as classification, regression, clustering, dimensionality reduction, model selection and preprocessing.
- **PANDAS (Python Data Analysis Library)**
    - High performance data structure and analysis tools. Fast and efficient data frame format.
    - Datasets can be reshaped, add or remove columns and rows, aggregating, merging and joining datasets.
    - Easy import and export of data in various formats.