Importing Data

Overview of Analytical Process

- Problem definition
- Data collection
- Data preprocessing
- Exploratory analysis
- Modeling and interpretation

Problem Definition

- Objective: What are you trying to achieve?
- Business Question: What question are you trying to answer?
- Scope and Constraints: Timeline, resources, and limitations

Data Collection

- **Sources**: Databases, APIs, web scraping, CSVs, surveys, etc
- **Types**: Structured (tables), semi-structured (JSON), unstructured (text, images)

Data Preprocessing

- Cleaning: Handle missing values, duplicates, errors
- Transformation: Normalize, encode categorical variables

Exploratory Data Analysis (EDA)

- Visualization: Histograms, boxplots, heatmaps, etc
- Statistics: Mean, median, correlations, distributions
- Outlier detection

Modeling and interpretation

- Modeling: Choose algorithms (e.g., logistic regression, random forest, etc.)
- Evaluation: Use metrics like accuracy, precision, recall, F1-Score
- Interpretation: Understand feature importance
- Validation: Cross-validation, test/train split

Structured vs Unstructured Data

- Structured: Organized in rows and columns (e.g.databases)
- Unstructured: No predefined format (e.g. images, videos)

Quantitative Data

- Numerical data used for calculations

  Examples: Age, salary, temperature

Qualitative Data

Descriptive data

Examples: Gender, occupation, feedback

Categorical Data

Nominal: No order (e.g. color)

Ordinal: order (e.g. rating scale)

Numerical Data

Discrete: Countable (e.g. number of items)

Continuous: Any value (e.g. height)

What Are Data Sources?

- Def:- Data Sources refer to the origins or repositories where data is collected from
- Importance

1) Informed Decision-Making
   2) Improved Efficiency
   3) Innovation and Research
   4) Market Understanding

open and paid data sources

- Free Data Sources

Ex. Kaggle.com, NASA Open Data

- Paid Data Sources(Premium)

Ex. IBM MarketScan for Health Industry data

Describe the metadata

- Def:-Data About data is called as metadata
- Data Identification

:- what the dataset contains

- Data Quality

:-Missing values, data format. Data Processing & Automation

- Data Processing & Automation

:- ETL pipelines for understand data