**Data Types**

* Numerical -Quantitative values, continuous or discrete (Height, weight, count of items)
* Categorical -Qualitative data, nominal or ordinal (Gender, rating, color)
* Text -String / natural language data (Reviews, emails)
* Time-Series / Sequential -Data points indexed over time (Stock prices, sensor readings)
* Images / Audio / Video -Multidimensional arrays or signals (Photos, audio files, videos)

**Data Collection Sources**

* Data Collection Sources - Where data comes from (Databases, CSV, APIs, web scraping, sensors)

**Data Cleaning / Preprocessing**

* Handling Missing Values -Strategies for missing data (Drop, mean/median imputation, interpolation)
* Outlier Detection -Identify and handle abnormal values (Z-score, IQR, winsorization)
* Scaling / Normalization -Standardize features (Min-Max scaling, Z-score standardization)
* Encoding Categorical Data -Convert categories to numbers (One-hot, label, target encoding)
* Datetime Handling -Extract features from timestamps (Year, month, lag features)
* Text Preprocessing -Clean text data (Tokenization, stopwords removal, stemming)

**Exploratory Data Analysis (EDA)**

* Descriptive Stats -Summary metrics of data (Mean, median, variance, skewness)
* Visualizations -Plot data to understand patterns (Histogram, boxplot, scatter, heatmap)
* Correlation Analysis -Relationships between variables (Pearson, Spearman, Kendall)
* Distribution Checks -Test if data follows a distribution (Shapiro-Wilk, Kolmogorov-Smirnov)

**Feature Engineering**

* Feature Creation -Create new features from data (Ratios, interactions, polynomial features)
* Feature Transformation -Transform features for better modelling (Log, sqrt, Box-Cox)
* Dimensionality Reduction -Reduce feature count while preserving info (PCA, t-SNE, UMAP)
* Feature Selection -Select most important features (SelectKBest, RFE, variance threshold)

**Sampling / Dataset Split**

* Train/Test/Validation Split -Split data for training and evaluation (70/30, 80/20, 60/20/20)
* Cross-Validation -Evaluate model robustness (K-Fold, Stratified K-Fold, LOOCV)
* Handling Imbalanced Data -Deal with skewed class distributions (SMOTE, oversampling, undersampling)

**Data Quality / Integrity Checks**

* Data Quality / Integrity Checks -Ensure clean and consistent data (Missing %, duplicates, outliers, consistency checks)

**Basic Statistics / Probability**

* Descriptive Stats -Measures of central tendency & spread (Mean, median, mode, variance, std dev)
* Probability -Likelihood of events (Conditional, joint, Bayes’ theorem)
* Distributions -Common probability distributions (Normal, Binomial, Poisson, Exponential, Uniform)

**Data Visualization**

* Tools -Libraries and platforms for plotting (Matplotlib, Seaborn, Plotly, Tableau)
* Charts -Visual methods for different data types (Bar, pie, histogram, boxplot, scatter, heatmap)
* Dashboards / Storytelling -Combine plots to communicate insights (Multi-plot dashboards, interactive charts)

**Data Ethics / Bias**

* Bias Detection -Detect sampling, label, or feature bias (Statistical checks, fairness metrics)
* Privacy -Protect sensitive information (Anonymization, encryption, GDPR compliance)
* Fairness -Avoid discriminatory decisions (Avoid sensitive features, monitor outputs