Data Pipeline and Preprocessing Detailed Outline

1. Data Collection

1. Identify data sources (e.g., APIs, databases, files, sensors).
2. Fetch raw data from sources.
3. Store raw data for initial inspection and backup.

2. Data Ingestion

1. Load data into a processing workspace or system.
2. Automate recurring data pulls (scheduling, batch processing, streaming).
3. Verify data integrity after ingestion.

3. Data Exploration & Initial Assessment

1. Inspect data structure and types.
2. Summarize key statistics (mean, missing values, distributions).
3. Visualize sample data (histograms, boxplots, scatter plots).

4. Data Cleaning

1. Remove duplicates.
2. Handle missing values (imputation, removal, tagging).
3. Fix errors and inconsistencies (data type conversion, standardization).
4. Filter outliers and erroneous records if needed.

5. Data Transformation

1. Encode categorical variables.
2. Scale or normalize numerical features.
3. Feature engineering (create, combine, or extract features).
4. Aggregate or pivot data if required for analysis.

6. Data Splitting

1. Partition data into training, validation, and test sets.
2. Ensure representative distribution across splits.
3. Document the splitting methodology for reproducibility.

7. Pipeline Validation

1. Test each pipeline stage with sample data.

2. Check for data leakage or improper transformations.
3. Implement logging and error tracking.

8. Documentation & Versioning

1. Document pipeline steps, logic, and assumptions.
2. Maintain version control of pipeline scripts and configurations.
3. Keep change logs for improvements or bug fixes.