Data Cleaning Techniques

Chapter 1: Handling Missing Values

- **Task 1.1: Introduction to Missing Data**
  - Research and understand the different types of missing data: MCAR, MAR, NMAR.
  - Read relevant literature or documentation on missing data.
- **Task 1.2: Identifying Missing Values**
  - Load a dataset with missing values.
  - Use Python (Pandas) or R (dplyr) to identify and visualize missing values.
- **Task 1.3: Handling Missing Data**
  - Implement removal techniques (listwise and pairwise).
  - Apply imputation methods: mean, median, mode, KNN, MICE, and predictive modeling.
  - Compare the effects of different methods on data quality.
- **Task 1.4: Evaluation and Documentation**
  - Evaluate the impact of different handling methods on dataset completeness and accuracy.
  - Document your process and findings.

Chapter 2: Handling Outliers

- **Task 2.1: Introduction to Outliers**
  - Research and understand the types of outliers and their impact on data analysis.
  - Read relevant literature or documentation on outlier detection.
- **Task 2.2: Detecting Outliers**
  - Load a dataset with potential outliers.
  - Use Z-Score, IQR, and visual methods (boxplots, scatter plots) to detect outliers.
- **Task 2.3: Handling Outliers**
  - Implement methods for handling outliers, including removal or transformation.
  - Compare the results before and after handling outliers.
- **Task 2.4: Evaluation and Documentation**
  - Evaluate the impact of outlier handling on dataset distribution and analysis.

- Document your process and findings.

Chapter 3: Data Transformation

- **Task 3.1: Introduction to Data Transformation**
  - Research and understand normalization and standardization techniques.
  - Read relevant literature or documentation on data transformation.
- **Task 3.2: Normalization and Standardization**
  - Load a dataset and apply Min-Max Normalization and Z-Score Standardization.
  - Compare the effects of these transformations on data.
- **Task 3.3: Encoding Categorical Variables**
  - Apply One-Hot Encoding and other encoding techniques to categorical variables.
  - Evaluate the impact on dataset usability.
- **Task 3.4: Evaluation and Documentation**
  - Evaluate the impact of data transformation on dataset distribution and analysis.
  - Document your process and findings.

Chapter 4: Data Parsing and Text Data Cleaning

- **Task 4.1: Introduction to Text Data Cleaning**
  - Research and understand text parsing and cleaning techniques.
  - Read relevant literature or documentation on text data cleaning.
- **Task 4.2: Text Parsing**
  - Load a text dataset and apply tokenization, stopword removal, and other parsing techniques.
- **Task 4.3: Text Data Cleaning**
  - Implement stemming and lemmatization techniques to clean text data.
  - Evaluate the results of text cleaning.
- **Task 4.4: Evaluation and Documentation**
  - Evaluate the impact of text cleaning on data quality and analysis.
  - Document your process and findings.

Chapter 5: Dealing with Duplicate Data

- **Task 5.1: Introduction to Duplicate Data**

    - Research and understand the impact of duplicate data on analysis.
    - Read relevant literature or documentation on duplicate data handling.
- **Task 5.2: Detecting Duplicates**
    - Load a dataset with potential duplicates.
    - Use exact matching and fuzzy matching to identify duplicates.
- **Task 5.3: Handling Duplicates**
    - Implement methods for removing or consolidating duplicate records.
    - Compare the dataset before and after handling duplicates.
- **Task 5.4: Evaluation and Documentation**
    - Evaluate the impact of duplicate removal on dataset integrity.
    - Document your process and findings.

Chapter 6: Data Validation

- **Task 6.1: Introduction to Data Validation**
    - Research and understand data validation techniques and their importance.
    - Read relevant literature or documentation on data validation.
- **Task 6.2: Implementing Validation Checks**
    - Load a dataset and apply consistency checks and range checks.
    - Use Python (Pandas) or R (dplyr) to perform validation.
- **Task 6.3: Handling Validation Issues**
    - Address any issues identified during validation.
    - Implement strategies to ensure ongoing data quality.
- **Task 6.4: Evaluation and Documentation**
    - Evaluate the impact of validation checks on dataset accuracy.
    - Document your process and findings.

Chapter 7: Data Type Conversion

- **Task 7.1: Introduction to Data Types**
    - Research and understand different data types and their conversions.
    - Read relevant literature or documentation on data type conversion.
- **Task 7.2: Converting Data Types**

    - Load a dataset with mixed data types.
    - Implement conversion techniques to standardize data types.
- **Task 7.3: Standardizing Formats**
    - Apply standardization methods to ensure consistency in data formats.
    - Evaluate the results of data type conversion.
- **Task 7.4: Evaluation and Documentation**
    - Evaluate the impact of data type conversion on dataset usability.
    - Document your process and findings.

Chapter 8: Addressing Data Entry Errors

- **Task 8.1: Introduction to Data Entry Errors**
    - Research and understand common data entry errors and their impact.
    - Read relevant literature or documentation on data entry error correction.
- **Task 8.2: Identifying Data Entry Errors**
    - Load a dataset with potential data entry errors.
    - Use Python (Pandas) or R (dplyr) to identify and analyze errors.
- **Task 8.3: Correcting Data Entry Errors**
    - Implement methods for correcting typos and standardizing data formats.
    - Evaluate the impact of error correction on data quality.
- **Task 8.4: Evaluation and Documentation**
    - Evaluate the impact of error correction on dataset accuracy and analysis.
    - Document your process and findings.

Chapter 9: Handling Inconsistent Data

- **Task 9.1: Introduction to Inconsistent Data**
    - Research and understand the sources of data inconsistency and its impact.
    - Read relevant literature or documentation on handling inconsistent data.
- **Task 9.2: Identifying Inconsistencies**
    - Load a dataset with potential inconsistencies.
    - Use Python (Pandas) or R (dplyr) to identify and analyze inconsistencies.
- **Task 9.3: Harmonizing Data**

    - Implement methods for standardizing units and formats.
    - Use techniques to merge similar categories or harmonize data.
- **Task 9.4: Evaluation and Documentation**
    - Evaluate the impact of data harmonization on dataset consistency.
    - Document your process and findings.

Chapter 10: Data Enrichment and Feature Engineering

- **Task 10.1: Introduction to Data Enrichment and Feature Engineering**
    - Research and understand data enrichment and feature engineering techniques.
    - Read relevant literature or documentation on these advanced techniques.
- **Task 10.2: Enriching Data**
    - Load a dataset and merge it with external data sources.
    - Apply methods for enriching data with additional information.
- **Task 10.3: Feature Engineering**
    - Create new features from existing data (e.g., interaction terms, polynomial features).
    - Evaluate the impact of new features on data analysis.
- **Task 10.4: Evaluation and Documentation**
    - Evaluate the impact of data enrichment and feature engineering on dataset analysis.
    - Document your process and findings.