­­­

**Procedures and Decisions**

**1. Data Loading**

**Objective**: Load the dataset into a suitable data structure for processing.

* **Tool/Library**: Pandas (Python), read.csv() function
* **Description**: The dataset was loaded from a CSV file into a DataFrame for analysis.

**2. Handling Missing Values**

**Objective**: Identify and handle missing values in the dataset.

* **Method**:
  + **Identification**: Missing values were identified using functions like isnull() and sum().
  + **Imputation**:
    - Numerical missing values were imputed with the mean of the column using SimpleImputer with strategy 'mean'.
    - Categorical missing values were imputed with the most frequent value using SimpleImputer with strategy 'most\_frequent'.
* **Reasoning**: Imputation was chosen to maintain dataset size and consistency.

**3. Handling Outliers**

**Objective**: Detect and handle outliers to ensure data quality.

* **Method**:
  + **Z-Score Method**: Calculated the Z-scores for numerical columns and removed rows with Z-scores greater than 3.
  + **IQR Method**: Calculated the Interquartile Range (IQR) and removed rows outside the range defined by 1.5 times the IQR.
* **Reasoning**: Outliers were removed to prevent distortion of statistical analyses and model performance.

**4. Addressing Inconsistencies**

**Objective**: Correct data inconsistencies.

* **Method**:
  + **Standardization**: Standardized categorical values to lowercase and stripped whitespace.
  + **Data Type Conversion**: Converted date columns to a datetime format using pd.to\_datetime().
* **Reasoning**: Consistent formatting ensures accurate analysis and reduces errors.

**5. Removing Duplicates**

**Objective**: Identify and remove duplicate records.

* **Method**: Used the drop\_duplicates() function to remove duplicate rows from the dataset.
* **Reasoning**: Removing duplicates avoids redundant data and ensures the uniqueness of each record.

**6. Feature Engineering**

**Objective**: Create new features to enhance predictive power.

* **Method**:
  + **New Features**: Created new variables such as interaction terms and date-related features (e.g., year, month, day).
  + **Feature Creation**: Used arithmetic operations and date functions to derive new variables.
* **Reasoning**: New features were created to capture additional information and improve model performance.

**7. Data Transformation (Normalization and Scaling)**

**Objective**: Apply normalization and scaling to standardize feature ranges.

* **Method**:
  + **Normalization**: Applied Min-Max Scaling to bring features into a range [0, 1].
  + **Standardization**: Applied Z-score standardization to center features around zero with unit variance.
* **Reasoning**: Transformation techniques improve model performance by ensuring features are on a comparable scale.

**8. Exploratory Data Analysis (EDA)**

**Objective**: Explore data distributions and relationships.

* **Method**:
  + **Visualization**: Used libraries such as Matplotlib and Seaborn to create histograms, pair plots, and heatmaps.
  + **Statistical Analysis**: Calculated correlation matrices and visualized data distributions.
* **Reasoning**: EDA provides insights into the data and helps identify patterns, trends, and anomalies.

**9. Ensuring Data Quality and Integrity**

**Objective**: Confirm the data quality and integrity throughout the cleaning and processing pipeline.

* **Method**:
  + **Integrity Check**: Verified that no missing values or duplicates remained after processing.
  + **Documentation**: Recorded the steps and decisions made during the data processing.
* **Reasoning**: Ensuring data quality and documenting the process are crucial for reproducibility and reliable analysis.
  + **New Features**: Created new variables such as interaction terms and date-related features (e.g., year, month, day).
  + **Feature Creation**: Used arithmetic operations and date functions to derive new variables.
* **Reasoning**: New features were created to capture additional information and improve model performance.