1. **Data Quality Report Generation:**

You are an experienced data quality analyst tasked with evaluating the provided dataset for its readiness in machine learning applications. Focus your analysis on the dataset, with 'Y' as the target variable. Generate a detailed data quality report covering the following aspects:

1. Numerical Features

For each numerical feature, include:

* Percentage of Missing Values
* Cardinality (Number of unique values)
* Minimum Value
* 1st Quartile (25th percentile)
* Mean Value
* Median Value
* 3rd Quartile (75th percentile)
* Maximum Value
* Standard Deviation

2. Categorical Features

For each categorical feature, include:

* Percentage of Missing Values
* Cardinality (Number of unique categories)
* Mode (Most frequent category)
* Frequency of the Mode
* Percentage Frequency of the Mode
* Second Mode (Second most frequent category)
* Frequency of the Second Mode
* Percentage Frequency of the Second Mode

Ensure that the report is well-structured, and clearly labeled.

1. **Data Quality Issue Identification and Handling Strategy:**

You are a data quality expert tasked with analyzing a dataset based on a previously generated data quality report. Your objective is to identify data quality issues and propose appropriate handling strategies.

**Instructions:**

1. Use the data quality report to pinpoint issues, including:
   * Missing values
   * Outliers
   * Irregular cardinality (e.g., extremely high or low unique values)
   * Other potential data quality concerns
2. Create a table summarizing the identified issues and their respective handling strategies. Ensure the table includes the following columns:
   * **Feature Name**: Name of the feature with the identified issue
   * **Issue**: Description of the problem (e.g., "20% missing values," "Outliers detected beyond 3 standard deviations")
   * **Handling Strategy**: Proposed method for addressing the issue (e.g., "Impute missing values with median," "Remove outliers," "Reduce cardinality by grouping rare categories")

**Deliverables:**

Provide the output as a clear and concise table with actionable handling strategies. Ensure that the handling strategies are contextually appropriate for the dataset and align with best practices in data preprocessing for machine learning.

1. **Apply Data Quality Handling Strategies:**

You are a data preprocessing specialist tasked with improving the dataset quality by applying the identified handling strategies from the data quality issues report.

**Instructions:**

1. Refer to the table of identified data quality issues and their respective handling strategies.
2. Implement the strategies for each feature to address the following issues:
   * Missing values (e.g., imputation, removal)
   * Outliers (e.g., removal, transformation)
   * Irregular cardinality (e.g., grouping, one-hot encoding)
   * Other identified data quality concerns
3. Modify the dataset accordingly while maintaining its structure and integrity. Ensure no critical information is lost during the process.

**Deliverables:**

Provide the following:

* The cleaned and processed dataset after applying all handling strategies.
* A summary table highlighting each feature, the applied strategy, and the transformation details (e.g., "Missing values imputed with median," "Outliers capped at 3 standard deviations").
* Briefly explain any key decisions made during preprocessing that were not explicitly covered in the handling strategy table.

1. **Dataset Cleaning Instructions (Dataset 1)**

You are a data wrangling expert tasked with cleaning the attached CSV dataset while preserving its overall structure. The target feature is 'Y'. Perform the following operations precisely as instructed:

**Dataset 1 Cleaning Instructions:**

1. **Unify Units**:
   * Standardize the units for the following columns: Weight, Memory, and Inches. Ensure all values are consistent and converted to a common unit where applicable.
2. **Split Columns**:
   * **Memory**: Split the Memory column into two separate columns:
     + Memory Size (numeric values representing the size)
     + Memory Type (categorical values such as SSD or HDD)
   * **ScreenResolution**: Split the ScreenResolution column into:
     + Resolution (e.g., 1920x1080)
     + Display Type (e.g., IPS, OLED)
3. **Reduce Cardinality**:
   * For the GPU and CPU columns, reduce cardinality by grouping similar entries into broader categories (e.g., group GPUs by brand or series).

**Requirements:**

* Make only the requested changes; leave the rest of the dataset unchanged.
* Return the cleaned dataset as the output in CSV format.

**Deliverables:**

1. The cleaned dataset in its updated format.
2. A summary of the transformations applied to each column, specifying how units were unified, how columns were split, and how cardinality was reduced.
3. **Dataset Cleaning Instructions (Dataset 2)**

You are a data wrangling specialist responsible for cleaning the attached CSV dataset while preserving its integrity. The target feature is 'Y'. Perform the following operations as specified:

**Dataset 2 Cleaning Instructions:**

[Insert your specific instructions here, such as:

* Splitting or merging columns
* Standardizing units
* Normalization
* Reducing cardinality for categorical variables
* Renaming columns for clarity or consistency
* ………

**Requirements:**

* Execute only the specified changes; leave the rest of the dataset unchanged.
* Ensure the dataset structure remains intact and all transformations align with the provided instructions.

**Deliverables:**

1. The cleaned dataset in its updated format (CSV).
2. A summary of all performed transformations, detailing how each instruction was implemented.
3. **Dataset Splitting: Training and Testing Sets [Both Datasets]**

You are a machine learning engineer tasked with preparing data for model development. Using the attached CSV dataset with 'Y' as the target feature, split the dataset into training and testing sets.

**Instructions:**

1. **Split Ratio**: Divide the dataset into:
   * **Training Set**: 80% of the data
   * **Testing Set**: 20% of the data
2. **Stratify**: Ensure the split is stratified based on the target feature 'Y' to maintain the same class distribution in both sets.
3. **Random State**: Use a fixed random state to ensure reproducibility of the split.
4. **Output Format**: Provide the resulting training and testing sets as separate CSV files.

**Deliverables:**

1. training\_set.csv: Contains 80% of the data with stratified distribution of 'Y'.
2. testing\_set.csv: Contains 20% of the data with stratified distribution of 'Y'.
3. A brief summary of the class distribution in both training and testing sets to confirm stratification.
4. **Train a Machine Learning Model (Dataset 1, 2) [change the Metrics information]**

You are a machine learning practitioner tasked with building and evaluating a predictive model using the provided training dataset. The target feature is 'Price\_euros'.

**Instructions:**

1. **Algorithm**: Train a machine learning model using the specified algorithm [Insert Algorithm Here].
2. **Cross-Validation**: Perform 5-fold cross-validation to ensure robust evaluation and reduce overfitting.
3. **Evaluation Metrics**: Calculate and report the following metrics for the model's performance on the training set:
   * **R-Squared**
   * **Adjusted R-Squared** (if applicable)
   * **Mean Squared Error (MSE)**
   * Additional relevant metrics based on the algorithm or problem type (e.g., RMSE, MAE, classification accuracy).

**Deliverables:**

1. **Model Training Accuracy**: Summary of the model's training performance across all folds.
2. **Evaluation Metrics**: Provide the metrics' values and explain their significance in assessing the model's performance.
3. **Validation Results**: Include the mean and standard deviation of the metrics across the folds.
4. **Test the Machine Learning Model (Dataset 1,2) [change the Metrics information]**

You are a machine learning practitioner tasked with evaluating the performance of the trained model on unseen data. Use the provided testing dataset with 'Price\_euros' as the target feature.

**Instructions:**

1. **Model Testing**: Use the trained model to make predictions on the testing dataset.
2. **Performance Metrics**: Evaluate the model's performance on the testing dataset by calculating the following metrics:
   * **R-Squared**
   * **Adjusted R-Squared** (if applicable)
   * **Mean Squared Error (MSE)**
   * Additional relevant metrics (e.g., RMSE, MAE) depending on the problem type.
3. **Insights and Generalization**: Analyze the results to provide insights into the model's ability to generalize to unseen data. Discuss:
   * The alignment or discrepancies between training and testing performance.
   * Possible reasons for any gaps (e.g., overfitting, insufficient data, noise).
   * Recommendations for further improvement, if needed.

**Deliverables:**

1. **Testing Metrics**: Report the calculated metrics for the testing dataset.
2. **Generalization Analysis**: Provide a brief analysis of the model's performance and generalization capability.
3. **Recommendations**: Highlight potential next steps to enhance the model's performance, if applicable.