**Consolidated-Disbursement-01-Apr-2024-to-31-Mar-2025-Sample-data**

**Data Cleaning and Transformation**

**A. Handle Missing Values**

* **Columns with Missing Values**:
  + Center Name: 164 missing.
  + Group Type: 225 missing.
  + Disbursement Mode: 739 missing.
  + Product Name, Product Code, Loan Purpose: 1,653 missing each.
  + Customer Caste, Customer Religious: 1,653 missing each.
  + Officer Emp Code Number: 164 missing.
* **Problem:** The dataset had missing values in several categorical columns. Missing values can cause issues in data analysis, modeling, and reporting, as they result in incomplete or skewed insights.
* **Affected Columns:**
  + Center Name
  + Group Type
  + Disbursement Mode
  + Product Name
  + Product Code
  + Loan Purpose
  + Customer Caste
  + Customer Religious
  + Officer Emp Code Number

**Cleaning Action Taken:**

* **Approach:**  
  For missing categorical values, we replaced them with the placeholder 'Unknown'. This ensures that the data remains consistent and analyzable without introducing bias or misrepresenting the information.
* **Why 'Unknown'?**
  + It clearly identifies the records where information was unavailable.
  + The placeholder value ensures transparency and flexibility for future updates.

**B. Variations in capitalization and extra whitespace**

* **Problem:**  
  Inconsistent formatting of text data across multiple columns was observed. Variations in capitalization and extra whitespace caused inefficiencies in data analysis and made it challenging to standardize insights.
* **Affected Columns:**
  + Center Name
  + Group Type
  + Customer Name
  + Disbursement Mode
  + Loan Purpose
  + Installment Frequency
  + Customer Caste
  + Customer Religious
  + Type
  + Officer Name
  + Branch
  + Region
  + State
* **Only whitespace removed, No lowercase is applied on the below colunms:**
* Product Name
* Product Code
* Center Code

**Cleaning Action Taken:**

* **Approach:**  
  For the specified columns, we applied two standardization techniques:
  1. **Stripping Whitespaces:**
     + Removed leading and trailing spaces from text data to eliminate formatting inconsistencies.
  2. **Converting to Lowercase:**
     + Transformed all text values to lowercase to ensure uniformity in capitalization.

**Why Was This Necessary?**

1. **Uniformity in Data:**  
   Standardizing the text values ensures that rows with the same logical value (e.g., New York and new york) are treated identically.
2. **Improved Analysis:**  
   Consistent data formatting avoids redundant categories during grouping, filtering, or reporting.

**C.** **Some rows had values such as 12 or 24 (numeric), while others had 12 Months or 24 Months (textual)**.

**Problem:**  
The column **'loan tenure'** contained inconsistent data formatting. Some rows had values such as 12 or 24 (numeric), while others had 12 Months or 24 Months (textual). This inconsistency can cause errors in analysis, aggregation, and visualization, especially in tools like Power BI.

**Cleaning Action Taken:**

1. **Standardizing Column Names:**
   * Stripped extra spaces from column names to ensure uniformity and prevent errors during referencing or processing.
   * Example: A column named 'loan tenure ' was cleaned to loan tenure.
2. **Standardizing the 'loan tenure' Column:**
   * Converted the **'loan tenure'** column to string format to handle mixed numeric and textual values consistently.
   * Removed any leading or trailing spaces in the column to eliminate formatting inconsistencies.
   * Replaced numeric values (12 and 24) with their standardized textual equivalents (12 Months and 24 Months) to ensure uniform representation.

**Why Was This Necessary?**

1. **Data Uniformity:**  
   By having all values in the same format (e.g., 12 Months), we ensure consistent analysis, reporting, and visualization.
2. **Avoid Confusion:**  
   Mixed data formats can cause errors in filters, groupings, and calculations, especially in tools that depend on clean, structured data.
3. **Enhanced Readability:**  
   Representing loan tenure as 12 Months or 24 Months is clearer and more intuitive for stakeholders.

**Key Benefits of the Cleaning Process:**

1. **Improved Data Quality:**  
   Ensures that all entries in the **'loan tenure'** column are standardized, free of ambiguity, and ready for further use.
2. **Ease of Reporting:**  
   Power BI and similar tools can now process this column without errors or manual adjustments.

**D. DateTime Format**

**Problem:**  
The dataset contains columns such as **'disbursement date'** and **'first repayment date'**, which store date values. However, these columns might not be in a proper datetime format and could include invalid or inconsistent entries. This can cause challenges during analysis or reporting, especially when performing date-based calculations, filtering, or visualizations.

**Cleaning Action Taken:**

1. **Identified Date Columns:**
   * Targeted the following columns for cleaning:
     + **'disbursement date'**
     + **'first repayment date'**
2. **Converted Columns to Datetime Format:**
   * Applied pd.to\_datetime() to these columns to convert them into a consistent datetime format.

**Why Was This Necessary?**

1. **Data Consistency:**  
   Standardizing date formats ensures uniformity, making it easier to perform date-based operations like sorting, filtering, or aggregations.
2. **Facilitating Analysis:**  
   Date columns in proper datetime format are essential for calculations such as the time difference between two dates, identifying trends over time, or generating visualizations in tools like Power BI.

**Key Benefits of the Cleaning Process:**

1. **Improved Usability:**  
   Standardized date columns are now ready for direct use in analytics, dashboards, or reports.
2. **Enhanced Functionality:**  
   Ensures compatibility with date-related operations and functions in Python, Power BI, or other analytics tools.

**E**. **Disbursement Year** **and** **Effective Interest**

* **Objective:**  
  To enrich the dataset by adding two derived columns, **'disbursement year'** and **'effective interest'**, which provide additional insights and facilitate better analysis.

**Enhancement Actions Taken:**

1. **Derived the Year of Disbursement:**
   * **Column Added:**
     + **'disbursement year'**
2. **Calculated Effective Interest:**
   * **Column Added:**
     + **'effective interest'**
   * **Purpose:**
     + Provides a direct monetary value of the interest amount for each loan. This helps in understanding the financial impact of loans and evaluating revenue from interest.

**Why Was This Necessary?**

1. **Actionable Insights:**
   * The **'disbursement year'** allows for time-series analysis, enabling the identification of patterns or shifts in loan distribution over years.
   * The **'effective interest'** provides clarity on the financial value of interest generated for each loan.
2. **Power BI Integration:**
   * Including these columns ensures easier visualization and deeper insights in the dashboard, especially for revenue analysis or trend visualization.

**Consolidated Pos as on 30 Nov 2024 (Sample Data)**

**A. Handling Missing Value**

**Problem:** The Disb Type column contained missing values, which could lead to inconsistencies and inaccuracies in data analysis. Having incomplete data made it difficult to properly classify and analyze the distribution of disbursement types, potentially skewing results and insights.

**Affected Column:**

* **Disb Type**

**Cleaning Action Taken:**

* **Approach:** For the Disb Type column, the following action was taken:
  1. **Filling Missing Values:**
     + The missing values (NaN) in the Disb Type column were replaced with the most frequent value, "Normal", to ensure completeness of the data.

**Why Was This Necessary?**

1. **Data Completeness:**
   * Missing values in the dataset can cause gaps in analysis, making it difficult to draw accurate insights. By filling in the missing values, we ensured that each row in the dataset had a valid Disb Type, allowing for more reliable analysis.
2. **Consistency in Reporting:**
   * The "Normal" category is the most frequent disbursement type, representing the majority of the data. By assigning "Normal" to missing values, we maintained consistency with the distribution of disbursement types.
3. **Accurate Insights:**
   * With missing values filled, any analysis, aggregation, or grouping by the Disb Type column can now be performed without the risk of missing data affecting results, leading to more accurate insights.

**B. Handling Missing Value**

**Problem:** The Cycle Number column had missing values (NaN), which could lead to incomplete data analysis and potentially misclassify entries. The missing values needed to be addressed to maintain consistency and accuracy across the dataset.

**Affected Column:**

* **Cycle Number**

**Cleaning Action Taken:**

* **Approach:** For the Cycle Number column, the following action was taken:
  1. **Filling Missing Values:**
     + The missing values (NaN) in the Cycle Number column were replaced with the most common cycle number, 1.0, to ensure completeness of the data and maintain consistency.

**Why Was This Necessary?**

1. **Data Completeness:**
   * Missing values in the Cycle Number column could lead to gaps in the analysis, making it difficult to classify and evaluate the data accurately. By filling the missing values with 1.0, we ensured that all rows had a valid cycle number, allowing for a more complete dataset.
2. **Consistency in Analysis:**
   * The cycle number 1.0 was chosen because it is the most frequent value in the dataset, appearing in the majority of the records. Using this value for missing entries ensures that the dataset reflects the most common cycle number, minimizing the impact of missing data.
3. **Improved Accuracy in Reporting:**
   * With missing values addressed, any analysis or calculations based on the Cycle Number column can now be carried out without worrying about the influence of missing data, resulting in more reliable and accurate insights.

**C. Handling Missing Value**

**Problem:** The Officer Employee Number column contained missing values (NaN), which could lead to gaps in data analysis. It was crucial to handle these missing values to maintain the integrity and consistency of the dataset.

**Affected Column:**

* **Officer Employee Number**

**Cleaning Action Taken:**

* **Approach:** For the Officer Employee Number column, the following action was taken:
  1. **Filling Missing Values:**
     + The missing (NaN) values in the Officer Employee Number column were replaced with the string 'Unknown' to fill the gaps and ensure that the column does not have any missing data.

**Why Was This Necessary?**

1. **Data Completeness:**
   * Missing values in the Officer Employee Number column could cause inconsistencies in analysis or make it difficult to identify specific officers. By replacing NaN values with 'Unknown', we ensured that every entry has a valid value, avoiding incomplete data.
2. **Handling Null Values Consistently:**
   * In cases where an officer's employee number is not available, the value 'Unknown' serves as a placeholder, signaling the absence of a specific officer. This placeholder ensures that missing data does not disrupt the analysis process and provides a consistent approach to handling null values.
3. **Enabling Accurate Reporting:**
   * With no missing values, it becomes easier to generate reports, analyze officer-related metrics, and perform any further operations on the Officer Employee Number column without concerns about missing data affecting the outcomes.

**D. Handling Officer Name and Id**

**Problem:** The Officer Name column contained both valid officer names and potential officer IDs, which could cause confusion and affect data analysis. Officer names and IDs follow different patterns, so it was important to identify and filter out rows where the Officer Name was not a valid name.

**Affected Column:**

* **Officer Name**

**Why Was This Necessary?**

1. **Data Accuracy:**
   * Officer IDs are different from officer names, and mixing the two could result in misleading data analysis or incorrect reporting. Filtering out rows with IDs ensures that the data only contains valid officer names, making it more accurate and reliable for analysis.
2. **Consistency in Data:**
   * By ensuring that the Officer Name column only contains names and not IDs, we standardized the data. This consistency improves the quality of reporting and reduces the likelihood of errors in downstream analysis.
3. **Improved Data Usability:**
   * Keeping only rows with officer names enhances the usability of the dataset, especially when generating reports or conducting detailed analysis involving officer-related data. The dataset becomes more focused and relevant to specific use cases, like identifying officer performance or creating officer-specific reports.

**Result:** This data cleaning step ensures that the Officer Name column now contains only valid officer names, making the dataset cleaner and ready for accurate and meaningful analysis.

**E. Date & Time**

**Problem:** The dataset contained several columns with date values stored as strings, which can create inconsistencies and errors when performing date-related analysis or calculations. These string values could contain incorrect formats or invalid entries, leading to potential issues when working with time-based analysis in downstream processes.

**Affected Columns:**

* **Disbursement Date**
* **First Installment Date**
* **Last Installment Date**

**Cleaning Action Taken:**

* **Approach:** To standardize and ensure proper handling of date values, the following steps were implemented:
  1. **Converting String Dates to Datetime Format**
  2. **Handling Invalid Date Entries**
  3. **Ensuring Data Consistency**

**Why Was This Necessary?**

1. **Consistency in Date Format:**
2. **Efficient Date Handling:**
3. **Handling Invalid Data:**

**Result:** This data cleaning step standardizes the date columns, ensuring that all date values are in the correct datetime format, and invalid date entries are safely handled. This improves the dataset’s consistency, facilitates accurate date-based analysis, and prevents errors related to improperly formatted date values.

**F. Stripping and Lowercasing**

**Problem:** The dataset contains multiple columns with text-based data, and inconsistencies in formatting were observed. These inconsistencies include leading or trailing whitespaces and variations in capitalization. Such formatting discrepancies can hinder analysis, cause inconsistencies in grouping or filtering, and result in inaccurate insights.

**Affected Columns:**

* **Officer Name**
* **Center Name**
* **Type**
* **OG Branch Name**
* **State**
* **Region**
* **Urban/Rural**
* **District**
* **Own/ Managed**
* **Agri / Non Agri**
* **Haryali / Samrudhi**
* **Disb Type**
* **Pre-Post**
* **Billing / Matured**

**Cleaning Action Taken:**

* **Approach:** To standardize the text values in the specified columns, the following steps were applied:
  1. **Stripping Leading and Trailing Whitespaces:**
     + The str.strip() method was used to remove any leading or trailing spaces from the text values in the specified columns. This ensures that any extra spaces that may have been inadvertently added during data entry or data extraction are eliminated.
  2. **Converting to Lowercase:**
     + The str.lower() method was applied to transform all text values to lowercase. This standardizes the capitalization, ensuring that variations like "New York" and "new york" are treated as the same value.
  3. **Ensuring Data Consistency:**
     + Both the stripping of whitespaces and the conversion to lowercase were applied to each of the affected columns, ensuring uniformity across the dataset and removing inconsistencies that could affect further analysis.

**Why Was This Necessary?**

1. **Consistency in Formatting:**
   * Text data with extra spaces or inconsistent capitalization can result in multiple entries for what should be the same value. For example, "New York" and " new york " would be treated as different values unless these issues are addressed. By stripping the whitespaces and converting text to lowercase, all values are standardized, improving data consistency.
2. **Improved Data Grouping and Analysis:**
   * Inconsistent formatting (such as variations in capitalization or extra spaces) could lead to issues when grouping, filtering, or aggregating data. For instance, grouping by state name might create multiple groups for the same state due to differences in formatting. Standardizing the text ensures accurate grouping and analysis.
3. **Data Quality Improvement:**
   * Cleaning the data by removing extra spaces and ensuring consistent capitalization improves the overall quality of the dataset. It reduces the chance of errors and ensures that insights derived from the dataset are based on accurate, uniform data.

**Result:** This cleaning step standardizes the text columns by removing extraneous spaces and ensuring all text values are in lowercase. The result is a cleaner, more consistent dataset, which enables accurate grouping, filtering, and analysis of categorical variables.