Data Cleaning Plan

Client Name: [Insert Client Name]

Project Overview

The project involves cleaning a customer call list dataset to ensure its accuracy, consistency, and readiness for use by call agents. The key tasks include handling duplicate records, addressing missing values, standardizing inconsistent formats, and removing irrelevant fields, as per the client's requirements.

Dataset Description

- **Raw Data Shape:** 21 rows, 8 columns
- **Key Columns:**
    - CustomerID
    - First_Name
    - Last_Name
    - Phone_Number
    - Address
    - Paying Customer
    - Do_Not_Contact
    - Not_Useful_Column

Key Challenges

1. Duplicate rows in the dataset.
2. Missing values in columns: `Last_Name`, `Phone_Number`, and `Do_Not_Contact`.
3. Inconsistent phone number formats (e.g., 123-545-5421, 123/643/9775, 876|678|3469) and invalid placeholders like "N/a".
4. Erroneous characters in `Last_Name` (e.g., '/', '..').
5. Variations in categorical values for `Paying Customer` and `Do_Not_Contact` columns (e.g., "Yes", "Y", "No", "N").
6. An irrelevant column (`Not_Useful_Column`).

Cleaning Strategy

1. **Data Acquisition and Profiling**
    - Load the dataset for inspection and profiling.
    - Identify duplicate records, null values, and inconsistent formats.
2. **Documentation of Findings and Strategy**
    - Record identified issues and design the cleaning workflow based on client needs.
3. **Data Cleaning**

- o Remove duplicate rows.
  - o Handle missing values in `Last_Name` and `Phone_Number` by filling or removing based on business rules.
  - o Standardize `Phone_Number` formats to a consistent format (e.g., +123-456-7890).
  - o Clean `Last_Name` by removing invalid characters (e.g., '/', '..').
  - o Split `Address` into `Street_Address`, `State`, and `Zip_Code` columns.
  - o Standardize categorical values in `Paying Customer` and `Do_Not_Contact` to "Yes" and "No".
  - o Remove irrelevant columns like `Not_Useful_Column`.
  - o Identify and flag duplicate phone numbers.
4. **Cleaned Data Export**
  - o Export the final cleaned dataset as "Cleaned_Customer_Call_List.xlsx".
5. **Verification and Finalization**
  - o Validate the cleaned dataset for consistency and correctness.
  - o Cross-check against client requirements before submission.

Expected Output

- Cleaned dataset in Excel format with 10 rows and 10 columns.
- Data Cleaning Report summarizing the process and findings.

Timeline

- **Day 1:** Data profiling and cleaning strategy design.
- **Day 2:** Data cleaning and interim validation.
- **Day 3:** Verification, finalization, and delivery.