Data Quality Issues and Cleaning Plan for Kansas City 311 Call Center Service Requests Dataset

1. Missing Values

- **Problem**: Certain fields in the dataset, such as Department, Category1, Response Time, and Request Source, have missing values.
- **Observation**: Missing values may lead to incomplete analysis and visualization, affecting insights on department efficiency, request categorization, and trends.
- **Plan to Clean**:
    - Identify fields with a significant number of missing values.
    - For categorical fields (Department, Category1), consider using the mode (most common value) to impute missing data, or create a separate category such as "Unknown".
    - For numerical fields like Response Time, impute missing values using the mean or median, or exclude them from analysis if imputation isn't feasible.

2. Duplicate Records

- **Problem**: Duplicate service requests exist, potentially inflating the total number of requests and distorting the analysis.
- **Observation**: Duplicates can mislead analysis by exaggerating service workload and response times.
- **Plan to Clean**:
    - Use a unique identifier such as Case ID (or Request ID) to detect and remove duplicate records.
    - If Case ID is missing, check for identical combinations of Address, Request Type, Date, and Status fields to flag duplicates.

3. Inconsistent Date Formats

- **Problem**: Dates are inconsistently formatted, with some records using MM/DD/YYYY while others use YYYY-MM-DD or even text representations.
- **Observation**: Date inconsistencies make it difficult to analyze service requests over time (e.g., trends by year, month).
- **Plan to Clean**:
    - Standardize all date fields to a consistent format (YYYY-MM-DD), ensuring uniformity for temporal analysis.
    - Convert any text-based date fields into proper date formats.

4. Special Characters and Formatting Issues

- **Problem**: Special characters (e.g., quotes, commas, ampersands) in text fields like Request Description and Address may cause parsing errors and affect data integrity.
- **Observation**: This can lead to incomplete data import into SQL or visualization tools and may also affect readability.
- **Plan to Clean**:
    - Remove or replace special characters with appropriate alternatives (e.g., converting & to "and", removing excessive commas).
    - Utilize string cleaning functions in Alteryx or SQL to clean the affected fields.

5. Inconsistent Categorical Values

- **Problem**: Some fields, such as Category1 and Department, contain inconsistent labels (e.g., "Parks & Rec" vs. "Parks and Recreation").
- **Observation**: Inconsistent categorization will cause incorrect grouping in analysis, leading to inaccurate insights.
- **Plan to Clean**:
    - Normalize values by identifying and consolidating similar categories (e.g., use a consistent naming convention like "Parks and Recreation").
    - Use case-insensitive comparisons to group similar categories.

6. Incorrect or Outdated Geographic Data

- **Problem**: Geographic fields such as Zip Code, Latitude, and Longitude contain inaccuracies or missing values.
- **Observation**: This affects geographical visualizations, making it difficult to analyze service requests by location.
- **Plan to Clean**:
    - Validate geographic data using external reference sources (e.g., correct zip codes, validate coordinates).
    - For missing or inaccurate data, use third-party APIs (e.g., Google Maps API) to fetch correct geographic details based on addresses.

7. Invalid or Unrealistic Values in Response Time

- **Problem**: The Response Time field contains negative or unrealistically large values (e.g., -5 days or 1000 days to close a request).
- **Observation**: Invalid values distort response time analysis and affect departmental performance metrics.

- **Plan to Clean**:
    - Remove or flag records with negative or extreme response times as outliers.
    - If possible, investigate and correct any errors, otherwise exclude these records from the analysis.

8. Inconsistent Status Labels

- **Problem**: The Status field (e.g., Open, Closed, In Progress) contains inconsistent labels (e.g., "InProgress", "In Progress", "Open", "Closed-Resolved").
- **Observation**: This inconsistency can skew the analysis of service request statuses over time.
- **Plan to Clean**:
    - Standardize status labels by consolidating similar terms (e.g., "In Progress" instead of "InProgress").
    - Group similar statuses under a consistent format.

9. Improper Data Types

- **Problem**: Some numeric fields (e.g., Response Time, Days to Close) are stored as strings, while date fields might be stored as text.
- **Observation**: Incorrect data types prevent accurate calculations and aggregations.
- **Plan to Clean**:
    - Convert fields to their appropriate data types (e.g., convert Response Time to integer/decimal, and date fields to date type).
    - Validate data after conversion to ensure proper functioning.

10. No Proper Tracking of File Metadata

- **Problem**: The dataset lacks columns to track the file's metadata, such as filename, user, and load date.
- **Observation**: Lack of metadata tracking can cause issues with auditing and managing data loads.
- **Plan to Clean**:
    - Add three new columns:
        - File_Name: To capture the name of the file loaded.
        - User_Name: To capture the user who loaded the data (using GetEnvironmentVariable("USERNAME")).

- Load_Date: To capture the date of the data load (using DateTimeNow()).