Data Quality Assessment & Cleaning

Dataset: Kansas City 311 Call Center Service Requests

1. Specific data problems observed

- Missing values: drove the high row_has_null = 'Y' count.
- Dates stored as strings: creation_date, closed_date often text with various formats.
- Invalid chronology: closed_date before creation_date ⇒ negative days_to_close.
- Closed_date and creation_date are in V_String format.
- days_to_close stored as text such as '879.0' → casting failures in SQL and visuals.
- Time fields: creation_time present but not always aligned with creation_date.
- Outliers in days_to_close: very large values likely due to data entry errors or very old cases reopened.
- Inconsistent categorical values: mixed case, trailing spaces, synonyms like Open, OPEN, open.
- Geography: NULL or (0,0) coordinates; a handful clearly outside KC.
- ZIP not always 5-digit; sometimes blank.
- Address fields with extra punctuation/special chars.

2. What I profiled (Alteryx)

- Row counts: full file (raw) loaded; created a row_has_null flag at the record level.
  - Y (has any NULL): 1,031,465 rows
  - N (no NULLs): 54,757 rows
- Uniqueness: used Summarize (Group By case_id) → Count; no duplicate case_id found.
- Schema scan: Auto Field + Browse to infer types and spot long text, dates stored as text, and outliers.
- Date consistency: compared creation_date vs closed_date; found records where closed < creation leading to negative durations.
- Address quality: street_address includes blanks, stray punctuation, inconsistent abbreviations.
- Created more columns: **File_Name** VARCHAR, **User_Name** VARCHAR, **Load_Date** DATE

3. SQL

- Uses TRY_CONVERT(date, creation_date) and a safe year-month string for month-level trend lines.
- **Top 10** cases with smallest days_to_close after converting the text field to numeric with TRY_CONVERT(float, days_to_close) and ensuring **non-negative** values. Categorizes output by category1 and type.
- **Geography: ZIP** top-10, **street address** top-10, and **exact lat/long** clusters grouped by identical coordinates to surface hotspots.
- Breaks down counts by department, work_group for stacked/treemap views.
- Per-department **Requests, AvgDays, MinDays, MaxDays**, with TRY_CONVERT(float, days_to_close) and **>= 0** filter to exclude invalid negatives.
- Averages of days_to_close by category1 (numeric conversion + non-negative filter) and returns Top-10 slowest categories.
- Counted number of rows present