**No Kill Colorado (NKC) Project: Data Wrangling**

This data wrangling project was carried out by me individually. Due to the extensive scope of work, I was not able to complete everything within the initial deadline. However, I am submitting the significant progress I’ve made so far, which I believe adds valuable insights to the overall project.

The key steps completed include:

* **Standardizing and merging multi-year data (2016–2024):** Cleaned and consolidated shelter and rescue data across these years to ensure consistency.
* **Facility location standardization:** Extracted and standardized available facility location details, including additional information from the Active PACFA Facilities master list.
* **Data cleaning and consistency checks:** Ensured data accuracy by handling discrepancies in facility names, event types, and location information.
* **Final consolidated dataset:** Combined all-year shelter data with the maximum available facility location information into a single standardized file.

**Limitations:**

* Data wrangling focused on the years 2016–2024 data. Earlier data (2007–2015) were not processed due to time constraints.
* Not all facility location records could be fully completed. Approximately **314 unique facilities** still lack location information (city, county, state, zipcode, latitude, and longitude) in the final cleaned file. While many records were updated using available datasets and PACFA information.
* Further enrichment (e.g., through web scraping) would require additional time due to the large number of facilities.
* Maximum efforts were made to standardize facility names, species, event types, and flow categories. However, minor inconsistencies may still remain and could be addressed through future validation and review.

**In summary, the current submission covers cleaned and consolidated animal shelter & rescue data from 2016–2024, with standardized location information integrated wherever possible.**

**Note:** All the detailed step-by-step transformations and full code explanations are included in the [Python script / Jupyter notebook document](https://drive.google.com/drive/folders/1--A36QZa1yCNwCc8RngMpI77JqZJJKg4?usp=drive_link). This document focuses on the **workflow summary and data structure**, while the notebook provides the cell-by-cell implementation, intermediate outputs, and clarification of each processing step.

**Workflow Documentation:**

**All-Year Animal and Shelter Data:**

* **Tools Used:** Excel and Python (Jupyter Notebook).
* **Initial Cleaning:** 
  + In Excel, removed extra header/footer rows, text, and obvious formatting issues that would break automated parsing.
  + Saved cleaned Excel versions to a working folder before loading to Python.
* **Data Transformation:**
  + The data was originally in wide format, making analysis difficult.
  + Converted wide format to long format using melt (unpivoting) in Python.
  + This significantly reduced the number of columns, placing all numeric values (animal counts) into one column and their related metrics into another.
  + Extracted Species, Age Group, and Event Types from the metrics column into separate columns.
  + Standardized names across Species, Age Groups, and Event Types for consistency.
  + Created a new column ‘Flow Type’ to categorize events and their values into broader groups (Start of Year, Intake, Outcomes, End of Year).
  + Added a ‘Reporting Year’ column for easy identification of data by year.
* **Consolidation:**
  + Repeated the above process for all years (2016–2024) with minor adjustments as needed.
  + Combined all cleaned data into a single consolidated file.

**Facility and Location Data:**

* Facility location details (Zip Code, Latitude, Longitude, City, County) were incomplete and inconsistent across years.
* Extracted all unique facility names and their available location details into a separate facility location table.
* Applied manual cleaning and fuzzy matching to standardize facility names and align location details.
* Supplemented missing facility details with data from the Active PACFA Facilities master list.
* Merged the cleaned facility location table with the consolidated all-year dataset to enrich it with standardized location details.

**Final Cleaned Data Structure:**

The standardized dataset is saved as: [**Cleaned\_2016\_to\_2024\_Shelter\_And\_Rescue\_Statistics\_final**](https://docs.google.com/spreadsheets/d/1K-knqXKoRbGbZLTWMKx68aSZwXfdiNv8/edit?usp=drive_link&ouid=108416728146964584886&rtpof=true&sd=true)

Columns:

1. Facility Name
2. City
3. County
4. State
5. Zip Code
6. Latitude
7. Longitude
8. Reporting Year
9. Species
10. Age Group
11. Flow Type
12. Event Type
13. Animal Count

**Distinct Categories:**

* **Reporting Year:** 2016–2024
* **Species:** Dogs, Cats, Birds, Small Mammals, Reptiles, Rabbits, Other, Reptiles & Amphibians
* **Age Group:** Adult, Juvenile, Unknown
* **Flow Type:**
  + Start of Year Count
  + Intake
  + Outcome (Positive)
  + Outcome (Negative)
  + End of Year Count
* **Event Type:**
  + Beginning Shelter Count
  + Beginning Foster Count
  + Stray
  + Owner Relinquished
  + Owner Requested Euthanasia upon intake
  + Transfer in from another Colorado organization
  + Transfer in from Out of State organization
  + Other: TNR / Protective Custody / Returns / Disaster Relief
  + Adoption
  + Returned To Owner (RTO)
  + Transferred out to another Colorado organization
  + Transferred to an Out of State organization
  + Other Live Outcomes (ie: TNR / SNR)
  + Died
  + Missing / Stolen
  + Shelter Euthanasia
  + Owner Requested Euthanasia
  + Ending Shelter Count
  + Ending Foster Count