ETL Project Report

By Adam Newlyn and Jay Schwan

Synopsis

Our team chose to simulate the merging of two animal shelters by taking two sets of data, munging the data for consistency, and then loading the data into a Postgres database.

1. Data Sources (Extraction)
   * We pulled two CSVs from Kaggle.com that contained animal shelter data for the cities of Austin, TX and Los Angeles, CA. The LA contained extraneous csv files for animal outcomes that we chose not to use, focusing instead on animal intake data.
2. Type of Transformation (Data Transformation)
   * The Austin dataset had less to massage, so we chose to make more changes to the LA data to get them to match.
     1. Removed extraneous columns: “age\_upon\_intake”, “color”, “datetime2”, “name”, “sex\_upon\_intake”
     2. Renamed the column headers to match the database schema.
     3. Reset index to animal\_id.
   * The LA dataset required more attention in order to get the data read to load. We mostly transformed the LA data to match the Austin data.
     1. First, we changed title cases from all caps to title casing
     2. Changed values that had similar phrasing to be consistent across both data sets. For example, we changed “Owner Surr” to “Owner Surrender” for Intake Type.
     3. Merged the “Group”, “Breed 1”, and “Breed 2” into one column to better match the Austin heading for breed.
     4. Removed extraneous columns: “Breed 1”, “Breed 2”, and “Group”.
     5. Renamed the column headers to match the database schema.
     6. Reset index to animal\_id.
3. Database (Loading the Data)
   * For the final database, we chose to load the data into a Postgres database. The data was already tabular and Postgres seemed to be the best choice as it is made to handle tabular data.
   * We chose to load the data into two separate tables that are easily mergeable with queries. We felt that keeping the two data sets distinct would provide historical context for future users.