Capstone 1: Data Wrangling

Project Proposal: A Home for All

LEARNING OBJECTIVES

1. Collect Data
2. Clean the dataset and address issues (missing values, outliers)
3. Apply data wrangling techniques

DATASET : https://www.kaggle.com/c/petfinder-adoption-prediction

APPROACH

1. Review the available files:
    - For Kaggle Dataset: Will be using the files “test.csv” and “train.csv” as well as all the tables for breed labels, color labels, and state labels.
    - Overall the Kaggle dataset is very clean, which is not unexpected.
2. Examine the “test’ dataset.
    - Found null values in the “names” and “description” columns, both of which make sense. Cursory examination of the data showed no outliers or unexpected values.
    - Created an extra column “adoption speed” and filled will nulls to allow for combination with “train” dataset.
3. Examine the “train” data
    - Similar null values, in name and description. Leave these as nulls.
4. Combine the datasets, drop the cat specific data and only keep dog specific data.
    - Notice that the null values for Name are now missing.
    - Go back and correct them by identifying that they’ve been converted to “No Description”.
    - Further examination of the name column reveals several “please name me”, “unnamed”, etc. Word clouds during EDA would yield more information and allow for a potentially more effective replacement string. However, further investigation shows a large number of names that are not actually names. Pulled data into excel and manually cleaned entirety of name column.
    - Description column is in a similar state. Anything that is just a string of characters and not an actual description manually cleaned out via excel.
5. Examine the breed columns. Notice that there are several blank values.
    - Create a “no breed” value and assign it to zero.
    - Fill null breeds with zero value.
    - Create a column for “breed count” to distinguish between mixed/pure breed dogs.
6. Retrieve dataset for big bones canine rescue from Petfinder KPI.
    - Missing values are dog specific data
        - Color, hair type, breeds, “declawed”.
        - “Declawed” is specific to cats and can be removed.
    - Also notice that petfinder dataset only shows ACTIVE pets, not adopted pets.
    - Contacted the shelter to see if they have a more complete dataset than that which is listed on Petfinder.
    - Eliminated columns with missing data from the data frame.

See Work via Github Link: Capstone_1_Data_Wrangling