Data Wrangling Report

Project Objective:

- Successfully wrangle the datasets using the 3 phases of gathering, assessing and cleaning all data effectively.
- Store the separate datasets into one master csv file and analyze the
- Create a visualization and gather 3 insights based on the end result

Phase 1: Gather

Three separate datasets that each share overlapping information that we need for this project were gathered and imported with the following methods:

- **I**mport Twitter dataset contained in 'twitter-archive-enhanced.csv' provided by Udacity through Udacity.
- The 'image file 'image-predictions.tsv' was directly imported into notebook programmatically using the requests function and the download url provided by Udacity
- 'Problem: could not access data through Tweepy due to Twitter's v2 api update; Solution: Used the .txt file provided in Udacity course and imported as a json file.

Phase 2: Assess

An assessment of the quality and tidiness was performed and resulted in the following assessments that were later fixed in the cleaning phase:

Quality

1. Certain columns don't have the appropriate data types.
2. Incorrectly extracted data in two datasets.
3. Completely empty columns in each dataset.
4. In the image dataset there are a lot of images where a dog was not detected at all, and in many cases detecting inanimate objects or other species of animals.
5. In the image predictions dataset there are multiple duplicate entries with the same jpg image url.
6. Source column in the tweets dataset displays the full html code for where the tweet was posted from, rather than just saying where it was coming from - Twitter for iPhone, Twitter Web Client, etc…
7. The dataset contains tweets that are retweets.
8. Column headers should be more descriptive and consistent across datasets

Tidiness

1. Dog stage (should be a single column but is incorrectly displayed in 4 separate columns.
2. Information about one type of observational unit (tweets) is unevenly spread across three different files/dataframes.

Phase 3: Clean

Using the previously made assessments on quality and tidiness, the following general changes were made to the datasets to improve accuracy, clarity, readability, and function:

Quality: Cleaning

1. Converted data types of several columns to optimize merging later on.
2. Revised incorrect and incomplete data across datasets into null values
3. Dropped all unnecessary columns
4. Identified images in the image dataset that are not dogs and removed them from the dataset.
5. Detected any duplicates in image dataset and removed them.
6. Replaced the html code in each of the 'source' column observations with readable text in each dataset
7. Filtered for any unoriginally posted tweet (retweet) - dropped retweets from the dataset
8. Analyzed and renamed several columns in all datasets

Tidiness: Cleaning

9. Consolidated 4 dog stage columns into one column and dropped the separated columns from database
10. Cross referenced tweets across the three different datasets and matched to keep only the ones that are contained in each dataset

The three databases were then merged together and double checked one last time.

End Result:

- Left with 609 separate entries that contain clean and organized data in one master file - 'twitter_archive_master.csv'
- Created clean visualizations representing the popularity of specific dog breeds
- Discovered 3 different insights based off of the final data and used the visualization to represent insights.