Wrangle Report

Overview
For this project data was gathered from 3 different sources:

- twitter_archive_enhanced.csv: The WeRateDogs Twitter archive
- image_predictions.tsv: Udacity Server
- tweet_json.txt: Data queried from Twitter API using Python's Tweepy library

After gathering each of the above pieces of data, data was assessed for quality and tidiness issues. Clean data that was used for analysis and visualization was saved under the name : twitter_archive_master

Wrangling Process
After gathering the data, it was assessed both manually and programmatically for quality and tidiness issues. Assesment has been carried out with a few key points in mind:

- Only original ratings (no retweets) have images
- Not all tweets are dog ratings and some are retweets
- The fact that the rating numerators are greater than the denominators does not need to be cleaned. This unique rating system is a big part of the popularity of WeRateDogs.

The following assessment observation were made and addressed accordingly:

Quality:

- **Twitter Archive Dataset**
  - tweet_id : Data type is int. should be string
  - timestamp : Data type is string. should be datetime
  - doggo , floofer , pupper , puppo : Although in info report, there's no null value, there are many entries that read 'None'.
  - name : Many 'None' entries. Inacurate entries such as 'a', 'by',etc
  - rating_denominator : All entries should be 10
  - rating_numerator : Many invalid entries: 1776, 960, ...Values higher than 20 are not acceptable
  - Identify tweets that are not about dogs. Should be identified and dropped
  - source : Extract source from URL
  - Drop retweets

- **Image Prediction Dataset**
  - tweet_id : Data type is int. should be string
  - Columns p1 , p1_conf , p1_dog ,...: Column names are not informative
  - Entries for dog breed: Inconsistent coptalization. "_" should be replaced by space
  - p1_dog , p2_dog , p3_dog : In a number of cases, where highly probable prediction(1st) suggests it is not a dog, there are inconsistencies in 2nd and 3rd predictions

    - I dentify predictions that suggest the image is not of a dog. And label accordingly
- **Twitter API Dataset**
    - Column id_str : name should change to tweet_id for consisance needed in later stages

Tidiness:

- **Twitter Archive Dataset:**
    - 4 columns(doggo , floofer , pupper , puppo) represent one value (dog stage). Record the appropriate value under new columns dog_stage
- **Image Prediction Dataset**
    - p1_dog , p2_dog , p3_dog represent one value (is it a picture of a dog or not. Record the identifies (True,False) under a new column is_dog
    - p1, p2, p3 represent one value. It is stored under new columns dog_breed
- **Overall:**
    - All relevant data on original tweets about dogs and their attributes in one dataset twitter archive master

Data for name, dog stage and ratings were extracted from tweet text again to make sure, entries are correct. They were crossed checked what was available in the original dataset to ensure no entry is overlooked. In a number of cases manual cleaning had to be done through reading tweet text.

Entries for dog breed that was generated by image prediction algorithm were meticulously checked to ensure tweets that are not about dogs were taken out and that the most probable dog breed was recorded for each tweet.

In the end 3 datasets were merged on tweet_id to create twitter archive master dataset and irrelevant columns were dropped.