Project: Data Wrangling Report

Objective:

Wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations.

Gathering Data

Data was gathered from multiple sources using different approaches.

1. Twitter Archive data was provided as a direct download from the website
2. Image predictions were downloaded programmatically using the ‘request’ module in python
3. Additional Twitter data was downloading using Twitter API and the Tweepy Python module. The data was downloaded in JSON format and written to text

Assessing Data

All three data sources were assessed visually and programmatically. Microsoft Excel and Notepad++ was used to assess the data visually

Various python functions were used to assess the data programmatically, such describe and info of the various datasets. Data was checked for missing values and incorrect datatypes. Duplicates were checked as well

Several issues were identified:

- Quality issues
  - API Tweets (df_tweets)
    - The data is incomplete because some tweets failed to download, at least 14 tweets were unsuccessful
    - Columns doggo, floofer, pupper & puppo represent a single variable and should have been in one column
  - Image Predictions (df_images)
    - Several tweets had more than one image, but the dataset does not indicate which image was used for prediction
    - There are 66 duplicate image URLs, which could indicate that there are some retweets, since all tweet ids are unique
    - 324 tweets do not contain any dog images or AI failed to identify the dog
  - Twitter Archive (df_achtweets)
    - Incorrect data types. Date/time stored as objects
    - 84% of the tweets do not have a personality for the dog
    - There are two tweets without a 0 numerator and one tweet with a 0 denominator

Clean the datasets

All the datasets were cleaned programmatically using the Define-Code-Test methodology.

The following steps were taken:

- All retweets were removed. Leaving them is akin to working with duplicates.
- Removed tweets with no images or images that are not dogs.
- The datasets were reshaped to remove unnecessary columns
- Test edeach field where dog is True, and created new column where dog is identified in the p1,2 or 3 -- If no dog is identified. This was to ensure that only dogs were kept
- Cleaned up dog names. They were not written in proper sentence case
- Corrected ‘datestamp’ field data type that was stored as object. This was changed to date/time format.
- Dropped invalid ratings, where the denominator was not equal to 10
- Created rating column (numerator/denominator) as float.
- Merge all three datasets on tweetids/id

Storing data

Saved the Tweets Master dataset to a csv file named "twitter_archive_master.csv"