Wrangling Report

In this project, I have analyzed We Rate Dogs tweets data. You can find more information about We Rate Dogs following this link. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. It was started in 2015 by college student Matt Nelson, and has received international media coverage both for its popularity and for the attention drawn to social media copyright law when it was suspended by Twitter.

Gathering data

Data were gathered from three different sources.

1. Twitter_archive_enhanced.csv was provided directly by Udacity, the CSV file was loaded using pandas and imported into a pandas DataFrame. This file contains general information about the tweets.
2. The prediction file image_predictions.tsv was downloaded from the Internet using the request library, this file contains the predictions made by neural networks on the images of the tweets. After the download, the file was imported into a pandas DataFrame.
3. The tweet_json.txt file was created by using Tweepy and the Twitter API. The data gathered from the Twitter API was saved to a text file called tweet_json.txt. This text file was later parsed line by line and imported into a pandas DataFrame. This file contains information about the favorite and retweet count.

Assessing the Data

The three DataFrames were assessed visually and programmatically for quality and tidiness issues. During this phase where each DataFrames was assessed individually, I found 10 quality issues and 2 tidiness issues

Quality

1. Retweeted tweets in the tweet json DataFrame should be removed in order to only have original posts
2. Keep only original ratings, avoiding retweets, that have images in the twitter_archive DataFrame
3. ID columns should be objects type, they are not intended to be numers and perform math calculations.

4. Correct numerator values extracting the values from the tweet text
5. Many dog names should be properly cleaned because aren't dog names, such as "a, an, the" etc...
6. The rating numerator and rating denominator columns in the twitter_archive DataFrame should be float type in order to handle decimals
7. Duplicated jpg url in the image predictions DataFrame should be removed
8. Retweet and favorite count should be int and not object in the tweet json DataFrame
9. Timestamp in the twitter_archive DataFrame should be DateTime datatype
10. Remove empty dog names

Tidiness

1. The 4 different columns doggo, floofer, pupper and puppo, refer to the dog stages of one dog, so they should be combined into one column
2. The three datasets should be merged into one master DataFrame

Clean

The issues identified in the previously were cleaned during this phase, using a define code and test pattern. At the end everything was merged into a one master pandas DataFrame

Store

The master pandas DataFrame was saved into a CSV file `twitter_archive_master.csv`