WeRateDogs Twitter Archive - Wrangle Report

In this report, I will be outlining the various ways in which I gathered, assess then cleaned the data in preparation for the analysis and visualisation of the data to create insights into the WeRateDogs Twitter archive.

Data Gathering

In this project, data was gathered from three different sources:

1. The enhanced twitter archive was downloaded manually from the Udacity servers as a csv.
2. The image predictions file was downloaded programmatically from the Udacity servers using the requests library.
3. Additional tweet data was collected by querying the Twitter API using Python's Tweepy library. This was performed programatically and every available tweet's JSON data was then stored with each tweet as its own line so that relevant information such as tweet ID, retweet and favourite counts could then be retrieved.

Assess and Clean

These three datasets were then assessed and cleaned for a variety of quality and tidiness issues.

Issues that needed addressing include:

1. Dropping tweets that were replies, retweets or not about dogs.
2. Converting incorrect datatypes such as the timestamp to date time; source and breed to categorical; retweet counts and favourite counts to integers.
3. Simplifying some data such as the tweet source from html to just its names.
4. Dealing with erroneous data such as removing words that had incorrectly been recorded as dog names; amending incorrectly extracted ratings; and dropping tweets not about dogs.
5. Standardising case sensitivity for names.
6. Some column names such as the different dog stages were actually variables and should all be placed in a single column, dog_stage.
7. Merging relevant data into the Twitter archive DataFrame such as retweet and favourite counts.
8. Dropping redundant columns such as duplicated column (jpg_url) or rating denominators as they were all 10 after cleaning.
9. Rearranging the column order to improve readability.

The master DataFrame was then stored as a csv called 'twitter_archive_master.csv'. The image predictions and JSON DataFrames were not stored as they were no longer required for analysis.