Wrangle Report

The data was available from 3 sources:
- A downloadable CSV file (Twitter archive) which contains the tweet text, various tweet metadata, as well as output from a text analysis of the tweet (dog name, rating and stage)
- A TSV file hosted on Cloudflare which contains computer vision predictions of what is on the tweet image, and whether the image is of a dog
- Twitter itself through its API

I started out importing the data and creating a dataframe for each:
- The Twitter archive was imported directly from the locally stored CSV
- The image predictions from Cloudflare was downloaded using the requests library
- The tweet data directly from the Twitter API using the Tweepy library, and stored as a JSON file as well as imported into a dataframe

Afterwards, I inspected the data for quality and tidiness issues.

I found the following issues, and cleaned some of them.

Twitter archive issues:

Quality issue 1:
'name' is incorrect in several cases, eg names like 'a', 'an' and 'the' appear frequently.
Cleaned by removing the names that I manually identified as being incorrect.

Quality issue 2:
'source' is not actually the URL of the tweet. Instead, it's simply a link to download the Twitter app.
Cleaned by replacing the incorrect URL with the correct one, which was inside the 'text' variable.

Quality issue 3:
Some dogs are allocated to multiple stages.
Cleaned by allocating dogs to the first stage they were allocated to.

Quality issue 4:
181 tweets are retweets and should be removed.
Cleaned by removing those tweets.

Quality issue 5:
23 tweets have a rating_denominator different from 10.
Cleaned by removing those tweets.

Tidiness issue 1:
'text' is actually two variables: the text of the tweet plus the URL of the tweet, which should instead be in 'source'. This means that the dataset has the issue of multiple variables stored in one column.
Cleaned by removing the URL from the 'text' variable and putting it in the 'source' variable instead.

Tidiness issue 2:
Dog stage is a set of columns (doggo, pupper etc) although it should be a single column ('stage') with the stage of the dog as the variable. This means that the dataset has the issue of Column headers are values, not variable names.
Cleaned by creating a new 'stage' variable and putting the dog stage there, and removing the other stage variables.

Tidiness issue 3:
The table contains both fundamental data about the tweets as well as output of a text analysis that has been made on the text (columns rating_numerator, rating_denominator, name, stage). To avoid having multiple data types in one table, this text output should be put in a separate table.
Cleaned by creating a new dataframe with the text extraction data.

Image predictions issues:

Quality issue 6:
Some tweets in the Twitter archive don't have image predictions associated with them.
Not cleaned.

Twitter API data issues:

Quality issue 7:
Some tweets in the Twitter archive don't exist anymore, and therefore it's not possible to get retweet data through the Twitter API for them.
Not cleaned.

After the data wrangling, I have the following clean dataframes:
- twitter_archive_master: Contains the actual tweet data from the Twitter archive
- text_extractions: Contains the data extracted from the tweet text
- image_predictions: Contains the predictions for the image of the tweet
- tweet_popularity: Contains the popularity data (retweets, likes etc) of the tweet

Finally, I stored the 4 tables as CSVs as well as in a SQLite database.