Wrangle Report

This report aims to summarise the processes used during the data wrangling project for Udacity.

Gathering

a) Flat files were programmatically downloaded and stored locally in csv or tsv formats.
b) A twitter API was used to obtain json data that was stored as a txt file.
c) To assess the data, all three data sources were converted into separate data frames

Assessment and Cleaning

Prediction Data – TSV Flat File

Cleanliness

- The data appeared to be relatively clean. There were no missing values. Values appear to be consistent and are appropriately assigned to string, Boolean, float, or integer.

Tidiness

- This dataset can be tidied up by using prediction number as a separate column header. The columns could then be listed as:
  - tweet_id
  - img_num
  - prediction (which will take values 1 to 3)
  - type
  - conf
  - dog

This allows each row to be a prediction and reduces the number of columns from 12 to 6.

To Sort out this issue I pulled out data for each prediction separately and then appended them together. This resulted in 3 times as many rows in the dataset. However, the number of columns was reduced from 12 to 6.

Archived Data – CSV Flat File

Cleanliness

- The following columns had null values and suggest incomplete data:
  - in_reply_to_status_id
  - in_reply_to_user_id
  - retweeted_status_id
  - retweeted_status_user_id
  - retweeted_status_timestamp
- As this information is not very useful, I decided to remove these columns from the dataset.
- The column expanded_urls also has missing data but cannot fixed due to truncated tweets.
- It was also noted that a rating denominator had a value of 0. I converted this value to 10.
- It was noted that some denominators were not 10. I standardised this so that numerators correspond to a denominator of 10.
- I also set the numerator value to 10.

- There are some extreme values that were biased. For example, the largest value 1776 corresponds to Independence Day in US where the dog was wearing a Mr America Suit. I set a hard cap for the max value of the numerator to be 15.
- The name column simply tries to capture the word the follows “This is” in the tweet. To make this more robust we can remove cases where the name does not begin with a capital letter.
- The source column is cleaned so that we only use descriptor name.
- The short URL is separated from the rest of the text.
- The short_url has multiple sometimes has multiple URLs. Only the last (latest) one is kept.
- The timestamp is changed to a datetime value.
- The extended_urls column has multiple values that repeat the same URL in most cases. Only the last URL is kept.

Tidiness

We can create a category for dog type so that we eliminate columns:

- doggo
- pupper
- floofer
- puppo

API Data – JSON converted to TXT File

Cleanliness

Some tweet_ids are missing from the API data. This is probably because the tweets have been later deleted. This issue will be ignored for now. The alternative is to delete these tweets.

Tidiness

There are no tidiness issues with this data.

Merging Datasets

- The archived data was merged with the API data first using a full outer join on tweet_id.
- Then this data merged with the prediction data using a full outer join on tweet_id.
- The final dataset reported information at a prediction level not tweet_id level.
- A full outer join was used to preserve information about missing rows from datasets.

Storing Datasets

The dataset was stored using SQLite. It was also saved as a csv flat file.