Data Wrangling Report

1. Gathering Data

Gather Twitter archive CSV file

Using the link provided by Udacity, I downloaded the WeRateDogs Twitter archive manually as twitter_archive_enhanced.csv and imported this file into a dataframe (`twitter_archive`).

Gather tweet image predictions

Download the tweet image predictions file hosted on Udacity's servers programmatically using Python's Requests library and saved it locally to `image_predictions.tsv` file. Then, I imported this file into a Python Pandas dataframe (`image_prediction_original`).

Gather data from Twitter API

By using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

2. Assessing Data

In total, I have spotted 9 quality issues and 3 tidiness issues.

Visual Assessment

I opened the `twitter_archive_enhanced.csv` and `image_predictions.tsv` in Excel and scrolled through them, looking for quality and tidiness issues. I was able to spot the following **quality** and **tidiness** issues:

Quality:

1. unnecessary html tags in source column of twitter archive in place of utility name e.g.`<a href=""http://twitter.com/download/iphone"" rel=""nofollow"">Twitter for iPhone</a>`
2. text column of twitter archive contains untruncated text instead of displayable text
3. Twitter archive data without any duplicates (i.e. retweets) will have empty retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp columns, which can be dropped

Tidiness:

1. doggo, floofer, pupper and puppo columns in `twitter_archive` table should be merged into one column named "**stage**"
2. there is some twitter that contains multiple dog stages, we need to separate the two stages using a comma after we combine them into one column **stage**

Programmatic Assessment

I used pandas' `info,value_counts and other` methods on `twitter_archive` to spot the following **quality** and **tidiness** issues:

Quality:

1. contains retweets and therefore, duplicates
2. many tweet_id(s) of `twitter_archive` table are missing in `image_prediction` table
3. erroneous datatypes (in_reply_to_status_id, in_reply_to_user_id and timestamp columns, rating_numerator and rating_denominator)
4. rating_denominator column has values other than 10
5. there isn't **rating** variable(= rating_numerator/ rating_denominator) that is more representative of how people rate the dog.
6. erroneous dog names starting with lowercase characters (e.g. a, an, actually, by)

Tidiness:

1. "breed" column should be added in `twitter_archive` table; its values based on p1_conf and p1_dog columns of `image_prediction` (image predictions) table
2. The third rule of tidy data says: "each type of observational unit forms a table". The **retweet_count** and **favorite_count** from `status_df_original` (tweet status) table are part of the same observational unit as `twitter_archive_original` table so should be merged into the same table and stored in a file called `twitter_archive_master.csv`

3. Cleaning Data

As all the quality and tidiness issues were related to `twitter_archive` table, I created a copy of only this table and named it `archive_clean`. For each quality/tidiness issue, I performed the programmatic data cleaning process in 3 stages - Define, Code & Test. During the cleaning process, I converted the datatypes of source and newly created stage columns of `archive_clean` to `category` datatype.

Storing Data

After the completion of the cleaning process, I stored the `archive_clean` DataFrame in `twitter_archive_master.csv` file.

(540 words)