

## Data Wrangling Report - WeRateDogs Twitter Archive

The overall aim of this data wrangling process was to combine data from 3 different sources, all in relation to certain sets of tweets taken from the [WeRateDogs Twitter page](#). After gathering and assessing the data, my goal was to find at least 8 quality issues and 2 tidiness issues. However, to my benefit, I identified 12 quality issues and 3 tidiness issues.

To account for word tokenization and word cloud visualization later on, I installed **wordcloud** and **spacy** through the pip Python package installer. I also installed **scipy** and **statsmodels** to account for regression analysis and visualization.

### Gathering Data

Firstly, Udacity provided an exclusive Twitter archive sent by WeRateDogs for Udacity students to use in this project, named *twitter\_archive\_enhanced.csv*. This archive contains 5,000+ of their tweets as of August 1, 2017. I stored this CSV file in a Pandas DataFrame named `arc`.

Udacity provided yet a second dataset called *image\_predictions.tsv*. This file was hosted on Udacity's servers and I programmatically downloaded it using the Requests library. This file contains the breed predictions of each image associated with each tweet, all ran through a neural network. I stored this TSV file, with tab delimiters, in a Pandas DataFrame named `image_predictions`.

Lastly, I needed to gather more data about the tweets that were not given in *twitter\_archive\_enhanced.csv*. To do this, I used `tweepy` to create a Twitter API object. I then scraped the JSON data of each tweet id listed under that CSV file, then stored each tweet's JSON data to its own line in a file named *tweet\_json.txt*. The fields I picked to be included in the DataFrame were *retweet\_count*, *favorite\_count*, and *created\_at*. After successfully parsing the data, I read the TXT file line-by-line into a Pandas DataFrame named `tweets`.

### Assessing Data

#### Visual Assessment

I looked through the *twitter\_archive\_enhanced.csv* and *image\_predictions.tsv* files in Excel. I spotted the following 4 quality issues and 1 tidiness issue:

- Quality
  - Inconsistent and inaccurate readings mapped from text to numerator & denominator fields (e.g. 24/7 being mistaken as a rating).
  - There is one entry with no rating.
  - Nonsensical dog names such as *a*, *an*, *quite*, *by*, *actually*, *such*, *not*, *one*.

- Unnecessary anchor tags under the *source* field.
- Tidiness
  - *Doggo, floofer, pupper, and puppo* all are not under 1 category variable.

### ***Programmatic Assessment***

Between *arc* and *image\_predictions*, I used a helper function to identify the tweet ids they had in common and which ids were missing. I stored these ids in a separate DataFrame. Afterward, I identified the tweet ids in common between *arc* and *tweets*. Again, I stored these ids in separate DataFrames, then I checked which missing tweet ids that *image\_predictions* and *tweets* had in common. Afterwards, I discovered these 2 quality issues.

- Quality
  - *arc* contains 281 tweets that are missing from *image\_predictions*.
  - *arc* contains 22 tweets that are missing from *tweets*. After dropping previous entries, there should be 13 tweets to be dropped from *arc*.

After dealing with missing data, I invoked the `.info()` function on each DataFrame to check for inconsistent or wrong data types. I found the following 1 quality issue:

- Erroneous data types under *tweet\_id*, *in\_reply\_to\_status\_id*, *in\_reply\_to\_user\_id*, *retweeted\_status\_id*, *retweeted\_status\_user\_id*, *timestamp*, and *retweeted\_status\_timestamp* in *arc*.

I invoked `.value_counts()` on some of the fields in all of the DataFrames. Additionally, I looked at the text content of the entries which are retweets, which indicated that there may be duplicate tweets. I found the following 4 quality issues and 2 tidiness issues:

- Quality:
  - Retweeted data indicates duplicated tweets.
  - *rating\_denominator* has values besides 10.
  - Inconsistent timestamp format between *arc* and *tweets*.
  - Under *image\_predictions*, dog breeds under the *pl* column have untidy formatting.
  - Dog breeds under the *pl* column have untidy formatting.
- Tidiness:
  - *favorite\_count* and *retweet\_count* fields under *tweets* need to be merged with *arc*.
  - *pl*, *pl\_conf*, and *pl\_dog* under *image\_predictions* need to be merged with *arc*.

I also looked at each unique *rating\_numerator* individually under *arc* and spotted the following 1 quality issue:

- Some numerator values contain a decimal point (e.g. 9.75/10).

### **Cleaning Data**

Before I dealt with the aforementioned issues, I created a clean copy of *arc* into a new DataFrame named *arc\_clean*. I cleaned the data in 3 stages for each issue: Define, Code, and Test. For each of these stages, I defined an action I needed to take, implemented that action in code, and tested if it was done correctly.

### **Storing Data**

After I was satisfied with the cleaned data, I stored *arc\_clean* into a file named *twitter\_archive\_master.csv*.