Wrangling Report (Internal Document)

In this report, I describe the main steps followed during the wrangling process.

Gathering Data

First, I gathered the data from three different sources and in three different formats.

The first piece was a Twitter archive given. I downloaded twitter-archive-enhanced.csv from the link and uploaded it to the Jupyter Notebook workspace. Then I loaded it into a pandas dataframe (df1).

The second piece of data was a collection of image predictions. I downloaded image_predictions using the Requests library and URL given. Then I loaded it into a pandas dataframe (df2).

The third piece of data was Twitter additional data extracted through an API. 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. Then I read this .txt file into a pandas DataFrame (df3) keeping only useful data (tweet ids, number of retweets and number of likes).

Assessing Data

Then, I assessed the data: inspecting the data (visually or programmatically) and documenting issues I encountered.

First, I did a visual assessment. Each piece of gathered data was displayed in the Jupyter Notebook and then in an external application (Excel).

Secondly, I did a programmatic assessment using pandas' functions and/or methods.

These are the issues I documented, separated by quality (content issues) and tidiness (structural issues).

Quality issues

twitter-archive-enhanced.csv table

- missing data in in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
- timestamp is datetime not string
- name column: all records starting with a lower-case letter are not valid names
- some tweets have inconsistent rating_numerator and rating_denominator values (numerator is usually between 10 to 13 and denominator is almost always 10)
- doggo, floofer, pupper and puppo are categories not strings

image_predictions.tsv table

- underscores in p1, p2 and p3 and inconsistent upper-case letters
- p1, p2 and p3 are categories not strings

tweet_json.txt table

- tweet_id, favorite_count and retweet_count are integers not strings

Tidiness issues

twitter-archive-enhanced.csv table

- there are two columns (rating_numerator and rating_denominator) for one variable

There are three pieces of data but one type of observational unit.

Cleaning Data

Prior to cleaning, I made a copy of the original pieces of data.

The next step was to clean all issues identified in the assess phase using Python and pandas.

Finally, to complete the wrangling process I created a tidy and clean master dataset with all pieces of gathered data. This dataset will be used in the following analysis process.