Wrangling Report

Data Wrangling:

- Read the `twitter-archive-enhanced`, `tweet-json.txt`

  Read `image-predictions.tsv` using requests module

Data Assessment:

Quality Issues:

- Missing Names
- Some posts are retweets
- Blatantly incorrect data in the `lang` column
- NaN Values
- Inaccurate data in `names` column as 55 dogs are named ‘s’
- `source_y` is identical to `source_x`
- Making the `p1_conf`, `p2_conf`, `p3_conf` in percentage rather than decimal form will make data easier to read
- Wrong Data Type in column `timestamp`

Tidiness Issues:

- Unnecessary columns that won’t affect Data Analysis
- `text` column in `twitter-archive` violated ‘Column-Variable Principle’
- Dog Stage variable is spread between many columns: `dogo` `floofer` `puppo` `pupper`

Cleaning Data

Unnecessary Columns

- Went from 58 columns to 31 columns

Missing Names

- Despite the name being empty it was present in the full tweet text in some cases. So using Spacy, a Python package, I created an algorithm to pass over each row and detect the name and replace the NaN value with it

Some Posts are Retweets

- All the columns that hinted that the post was a retweet were used to delete rows containing non-NaN values for Retweet Related Columns

Blatantly Incorrect Data in the lang Column

- Using the **`value_counts()`** function the column contained a multitude of different languages but upon looking carefully at the texts, it turns out that they were all in English just using English abnormally by spamming or removing spaces. Here are examples:
    - **Ohboyohboyohboyohboyohboy** *[was identified as Indonesian]*
    - **Omg omg oMG OMG OMG** *[was identified as Estonian]*

text Column Violated 'Column-Variable Principle'

- They contained the three variables **`text`**, **`twitter-link`**, **`rating`** which were split up using **`str.replace`** and Regular Expressions

Dog Age Variable is Spread Between Different Columns

- Using a well-detailed algorithm, the values in the four columns: **`doggo`**, **`floofer`**, **`pupper`**, **`puppo`** were all merged into one column named **`dog_stage`**

NaN Values

- All NaN values that were present after the previous cleaning processes were in **`dog_stage`** and **`name`**
    - NaN values in `dog_stage` were replaced by: None
    - NaN values in `name` were replaced by: Not Provided

Inaccurate Data in names Column

- About 55 dogs were named 'a' even though that wasn't the name given to them in the **`full_text`** column.
    - Using Spacy NLP all of them were given their correct value.

`source_y` is Identical to `source_x`

- As a result of the merge between the two datasets **`tweet.json`** and **`twitter_archive`**, which both contained the column **`source`**, duplicate columns were created. **`source_y`** was removed and **`source_x`** was renamed **`source`**

Making the `p1_conf`, `p2_conf`, `p3_conf` in Percentage Form Rather Than Decimal Form

- This was created for Data Readability, one of the pillars of Data Quality issues. It was done very simply by multiplying all values by 100 and rounding to 1 decimal place

Wrong Datatype in Column `timestamp`

- Data type should be ‘datetime64[ns, UTC]’ for the variable to be usable in Data Analysis.

Entire Log

- Dropped Columns:
  - `'in_reply_to_status_id_y'`, `'in_reply_to_status_id_str'`, `'in_reply_to_user_id_y'`, `'in_reply_to_user_id_str'`, `'in_reply_to_screen_name'`, `'in_reply_to_status_id_x'`, `'in_reply_to_user_id_x'`, `'truncated'`, `'geo'`, `'coordinates'`, `'contributors'`, `'is_quote_status'`, `'retweeted'`, `'place'`, `'possibly_sensitive'`, `'possibly_sensitive_appealable'`, `'quoted_status_id'`, `'quoted_status_id_str'`, `'quoted_status'`, `'retweeted_status_id'`, `'retweeted_status_user_id'`, `'retweeted_status_timestamp'`, `'retweeted_status'`, `'doggo'`, `'floofer'`, `'pupper'`, `'puppo'`, `'source_y'`

➔ `name` Column Changed:

- Read from `'names_filtered.csv'` and filtered further using NLP
- NaN values were changed to `'Not Provided'`

➔ Rows Removed:

- 79 rows were removed because they were retweets

➔ `lang` Column Changes:

- All values were turned to `'en'`

➔ `text` Column Changes:

- All ratings and Twitter links were removed

➔ `dog_stage` Column Creation:

- Storing the values of `'doggo'`, `'floofer'`, `'puppo'`, `'fluffer'`
- NaN values were changed to `'None'`

- `source_x` Renaming:
    - Renamed to `'source'`

- `p1_conf`, `p2_conf`, `p3_conf` Changes:
    - All were turned into percentage format (to the nearest decimal point) rather than decimals

- `timestamp` Data Type Change:
    - Changed to `datetime64[ns, UTC]`