

# "Wrangle and Analyze Data" Project

## Wrangle report

- Gather
- Assess
  - Quality issues
  - Tidiness issues
- Clean

### Insights Revealing

## Gather

First, I gathered 3 different format files, `twitter_archive_enhanced.csv`, `image_predictions.tsv` and at last `twitter-json.txt` that holds tweets json data sets.

The first file was downloaded directly, the second was downloaded programmatically, the third was extracted from a zip file programmatically after failing of code of twitter API to get the data.

After that, I loaded their data in corresponding pandas dataframes as copies.

## Assess

Then I assessed the files dataframes tables:

1. Visually: using spreadsheets, text editor software and pandas.

2. Programmatically: using pandas functions and methods

like: `head()`, `info`, `value_counts()`, `uplicated()`, `sum()`, `sort_values()`, `describe()`

Below is what I had found out:

### Quality issues

#### *twitter\_archive table*

- Data type of `tweet_id` column is `int64` instead of `category`
- Data type of `timestamp` column is `object` instead of `datetime`
- Source column contains distracting HTML tags
- Duplicated url strings in `expanded_urls` column
- Inaccurate username in `expanded_urls` column like (4bonds2carbon, kaijohnson\_19,bbcworld) in `urls` column instead of (dog\_rates)
- Embedded Urls like (<https://www.gofundme.com/mingusneedsus>), and (<https://www.gofundme.com/3yd6y1c>), (<https://www.gofundme.com/help-my-baby-sierra-get-better>), strings in `expanded_urls` column
- Wrong urls of tweets in `expanded_urls` column
- Missing urls of tweets in `expanded_urls` column
- Data type of `rating_numerator` and `rating_denominator` columns is `int64` instead of `float`
- 'None's in (doggo ,floofer,pupper,puppo) instead of null
- 'None's instead of null in `name` column
- Ratings with decimal values incorrectly extracted
- Wrong assigned rating for tweets with ids 810984652412424192 , 675153376133427200, 670783437142401025, 667549055577362432, and 666104133288665088
- Rows that contain 'retweets' data that are not the original tweets meant for analysis
- Rows that contain 'replies' data that are not the original tweets meant for analysis
- Columns (`in_reply_to_status_id`, `in_reply_to_user_id` column, `retweeted_status_id`, `retweeted_status_user_id`, `retweeted_status_timestamp`) in `twitter_archive` table not needed in our analysis

### *image\_predictions table*

- img\_num data type is int64 instead of category

### *tweet\_json\_df table*

#### Tidiness issues

- (doggo ,floofer,pupper,puppo) in `twitter_archive` table represent one variable "stage" in four columns
- (retweet\_counts, favorite\_counts) columns in `tweet_json_df` table should be part of the `twitter_archive` table, also (jpg\_url,img\_num) columns in `image_predictions` table should be part of the `twitter_archive` table. In general the three datasets should be merged as they are part of the same observational unit

## Clean

### Tidiness

I made stage column of concatenating (doggo,floofer,pupper,puppo) columns in `twitter_archive` table, then dropped the four separated columns at the end of cleaning up.

Then, I merged `tweet_json_df_clean` dataframe to `twitter_archive_clean` dataframe to have the columns `retweet_counts` and `favorite_counts` within one `twitter_archive_clean` dataframe

After that I also merged `image_predictions_clean` dataframe to `twitter_archive_clean`.

### Quality

#### In a brief:

- I omitted rows retweets and replys:

by removing rows that have non-empty `retweeted_status_id`, `retweeted_status_user_id`, or `retweeted_status_timestamp` and `in_reply_to_status_id`.

- Then replacing distracting HTML tags in `source` column with the original sources like:

```
"Twitter for iPhone" replaced <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
```

- URLs in `expanded_urls` were rehandled with the string ("[https://twitter.com/dog\\_rates/status/](https://twitter.com/dog_rates/status/)" + tweet ID) where tweet ID was taken from the `tweet_id` column to set the tweet url correctly in its place on the table and to overcome issues like:

1. URL duplications within the same cell in `expanded_urls` column
2. Usernames like (4bonds2carbon, kaijohnson\_19,bbcworld) in urls on `expanded_urls` column cells instead of the accurate username (dog\_rates)
3. trange embedded urls like (<https://www.gofundme.com/mingusneedsus>), (<https://www.gofundme.com/3yd6y1c>), and (<https://www.gofundme.com/help-my-baby-sierra-get-better>), strings in `expanded_urls` column
4. Wrong urls of tweets in `expanded_urls` column
5. Missing urls in `expanded_urls` column

- Extracting correct decimal values ratings came later, then removing wrong assigned rating for a number of tweets.
- The 'None's in `name` column, and empty cells in `stage` column were changed to nulls.
- Columns `in_reply_to_user_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`, `retweeted_status_timestamp` were dropped off.

- Eventually, here were the conversions of columns' data types:

- `tweet_id` and `img_num` columns to category
- `favorite_count` and `retweet_count` to int
- `timestamp` column to datetime
- `rating_numerator` and `rating_denominator` to float

## Arranging `twitter_archive_clean` columns and exporting to a csv file

- I tended to sort the master tweets table by `timestamp` column ignoring maintaining original indexing of table.
- Then, I sorted the table columns in a list to rearrange columns order in the table as they fit.
- At last step I exported the master table to a csv file named '`twitter_archive_master.csv`'

## Insights revealed

- Finally, I tried to extract a number of insights and here are some:

**1st Insight:** The vast majority of the tweets were pushed up from "Twitter" app on an iPhone device with 1964 times and about 94% of total tweets

**2nd Insight:** 'Charlie', 'Lucy', 'Oliver' and Cooper are sharing almost the same times 11 or 10 to be a given name lying within the most given names, while 'Laika', 'Jeffri', 'Mollie', 'Leela', and 'Rhino' are lying on the tail with one time to be a given name.

**3rd Insight:** With 132810 likes, a tweet sent on 2017-01-21 6:26:02 PM from Twitter for iPhone hit the top score of favorite (like) count with the text "Here's a super supportive puppo participating in the Toronto #WomensMarch today. 13/10 <https://t.co/nTz3FtorBc>" amongst all other tweets till August,1 2017

**4th Insight:** Though not reaching the highest favorite count, and being sent from the least source of tweets, an 'Atticus' got the top rating numerator with 1776 from @dog\_rates amongst all other tweets since 2015-11-15 10:32:08 PM, till August,1 2017 4:23:56 PM, on a tweet sent on 2016-07-04 3:00:45 PM from TweetDeck with the text "This is Atticus. He's quite simply America af. 1776/10 <https://t.co/GRXwMxLBkh>"

**5th Insight:** The most retweet count was for the favor of a "doggo" captured in a video by Tina Conrad, on a tweet sent on 2016-06-18 6:26:18 PM with the text "Here's a doggo realizing you can stand in a pool. 13/10 enlightened af (vid by Tina Conrad) <https://t.co/7wE9LTEXC4>"