Data wrangling is composed of three steps:

1-Data Gathering

2- Data Assessing(Quality and Tidiness)

3-Cleaning

Data Gathering:

we have three different resources:

1) enhanced twitter archive (.csv):

It is downloaded manually and imported using pd.read_csv().

The output is archive_df

-2) Additional data via the twitter API (tweepy):

Query Twitter's API for JSON data for each tweet ID in the Twitter archive and write each tweet's json to a text file and after that we extract information from text file into a panda's data frame.

The output is api_df

3) Image predictions file

it is downloaded programmatically using requests library from this link

the output is image_predictions_df

Assessing:

Data types (consistency issues)

Archive_df:

>> Created_at column is object

>> tweet_id is integer not string

>>Representation of null values with 'None'

Completeness issue

>>tweets with no images there is a discrepancy in the number of tweets between the **archive_df** dataset and the **image_prediction_df.**

>>names column has none values though the name may be in the tweet

>> Some tweets are actually retweets and replies
not original tweets that have to be deleted

accuracy issue

Archive_df:

Inaccuracy:

>>For the denominator; any value below or above 10

is suspected

>> the name column has some inaccurate values such as a and quiet

Tidiness:

>>The last four columns'headers in archive_df are variables we have columns for dog_stage which violates this rule

>>One observational unit(tweet_id) is found in more than one table so I stored api_df and archive_df in one table and image_predictions in another table

>>undescriptive column names in image_predictions_df

Cleaning:

I followed the following steps

** I converted the created_at column to datetime object

**replaced the None values with empty strings then NaN

I found the tweets that have images only from image_predictions_df and filtered the **archive_df by these ids then found the tweets that are retweets /replies removed them from archive_df and image_predictions_df

**removed the columns that are not important

**removed the tweets that have unreasonable numerator

Storing

I saved the data after cleaning in 2 csv files

1) twitter_archive_master.csv: that contains both archive_df and api_df
2) image_predictions.csv