Data Wrangling Report

Project objectives

The project main objectives were:

1. Perform data wrangling (gathering, assessing and cleaning) on provided thee sources of data.
2. Store, analyze, and visualize the wrangled data.
3. Reporting on 1) data wrangling efforts and 2) data analyses and visualizations.

Step 1: Gathering Data

In this phase, the three pieces of data were gathered and represented as pandas dataframes:

1. The WeRateDogs Twitter archive (file on hand, manual download of 'twitter-archiveenhanced.csv')
2. The tweet image predictions ('image-predictions.tsv'). This file was be downloaded programmatically using the Requests library from a provided URL.
3. Each tweet's entire set of JSON data (with at minimum tweet ID, retweet count, and favorite count) in a file called 'tweet_json.txt' were stored using Twitter API and Python's Tweepy library. Each tweet's JSON data was written to its own line.

Step 2 and 3: Assessing and Cleaning Data

While working with data, a number of observations were made. In the below table there are the observations along with actions taken in the Cleaning Step.

Note:

df_new – name of dataset of the enhanced-twitter-archive after removing the retweets

Tweet_df – Is the name of the wrangled tweets using twitter API

Image_df – is the image prediction dataset

Quality

Dataset	Observation	Solution
df_new	1. rename this names of dog missing --'a', 'by', 'all', etc as None	1. replace all this names with none as this names were named after 'This is ....' In the text column

	2. rename wrongly written dogs e.g. GÃ²rdÃ³n, AmÃ©lie, FrÃ¶nq, O 3. varying rating denominators -- not consistent 4. timestamp datatype should be made datetime 5. rows more than the image rows 6. Some names are same but different spellings e.g. Ed and Edd, Fillup and Filup, 8. The name 'o' ought to be 'O'Malley' 9. The rating denominator had 0 -- this should be made 13/10 as stated in the text	2. Correct all these names taking a cue from the text column. Was only able to correct one cause it was correctly written in the text column 3. This was ignored 4. timestamp column was changed to datetime 5. as cleaning process went on, it became correct e.g. dealing with the tweets and not the retweets 6. Correct those names 7. This was done earlier which led to the name df_new 8. Rename it correctly 9. Correct it
Image_df	1. Not all the names of dogs are capitalized. 2. Inconsistent number of decimal place in the probability column: p1_conf, p2_conf, p3_conf	1. Capitalize the name column 2. Made them all 5 decimal place
Tweets_df	1. favorite_count is 0 2. column name 'id' and not 'tweet_id' as others	1. Favorite_count being 0 is an outlier. Hence I dealt with the dataset not having that 2.Change the 'id' column to 'tweet_id'

Tidiness

Dataset	Observation	Solution
df_new	doggo, floofer, pupper, puppo columnsDog stages should be in one column	The 4 columns were melted into one
all	Too many datasets	Reduced to 1 eventually

Eventually,

One tidy dataset is formed

```
twitter_archive_master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1994 entries, 0 to 1993
Data columns (total 27 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   tweet_id                    1994 non-null   int64
 1   jpg_url                     1994 non-null   object
 2   img_num                     1994 non-null   int64
 3   p1                          1994 non-null   object
 4   p1_conf                     1994 non-null   float64
 5   p1_dog                      1994 non-null   bool
 6   p2                          1994 non-null   object
 7   p2_conf                     1994 non-null   float64
 8   p2_dog                      1994 non-null   bool
 9   p3                          1994 non-null   object
 10  p3_conf                     1994 non-null   float64
 11  p3_dog                      1994 non-null   bool
 12  retweet_count               1994 non-null   int64
 13  favorite_count              1994 non-null   int64
 14  in_reply_to_status_id       23 non-null     float64
 15  in_reply_to_user_id         23 non-null     float64
 16  timestamp                   1994 non-null   datetime64[ns, UTC]
 17  source                      1994 non-null   object
 18  text                        1994 non-null   object
 19  retweeted_status_id         0 non-null      float64
 20  retweeted_status_user_id    0 non-null      float64
 21  retweeted_status_timestamp  0 non-null      object
 22  expanded_urls               1994 non-null   object
 23  rating_numerator            1994 non-null   int64
 24  rating_denominator          1994 non-null   int64
 25  name                        1994 non-null   object
 26  stage                       1994 non-null   object
dtypes: bool(3), datetime64[ns, UTC](1), float64(7), int64(6), object(10)
memory usage: 395.3+ KB
```