Data Wrangling Report

This report is on the data wrangling process carried out on @dog_rates (WeRateDogs) twitter account. WeRateDogs rates people's dogs with a humorous comment about the dog.

Data Wrangling Process

- Data Gathering
- Data Assessing
- Data Cleaning

1. Data Gathering

Dataset was gathered from three different data sources:

- The WeRateDogs Twitter archive (a CSV file that was provided by WeRateDogs via udacity). The archive contained basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.
- The tweet Image prediction file (tsv file). This file contained the confident levels predictions of the dogs using a neural network.
- The additional data gotten from Twitter API (containing retweet count and favorite count of each tweet ID). Unfortunately, my request for twitter API was rejected, but Udacity provided a tweet_json.txt file which contained the retweet count and favorite count for each tweet ID.

The gathered data was loaded into three different DataFrame:

- archive_data: Loaded data from twitter_archive_enhanced.csv
- prediction: Loaded data from image-predictions.tsv
- tweet: Loaded data from tweet-json.txt

2. Data Assessing

The gathered dataset was assessed using two methods namely,

- Visual Assessment: Each piece of gathered data was displayed in the jupyter Notebook. Once displayed, data was additionally assessed using an external application (Ms.Excel).
- Programmatic Assessment: Pandas' functions were used to assess the data.

Legend:

- Methods: Visual (V) / Programmatic (P)
- Issues: Quality (Q) / Tidiness (T)

Table 1: Issues Identified

DataFrame	Methods	Issues	Column/Features	Description
archive_data	P	Q	timestamp, retweet, tweet_id	Erroneous datatype
	V	Q	In_reply_to_status_id, in_reply_to_user_id, retweeted_status_timestamp, retweeted_status_id, retweeted_status_user_id	original tweets (no retweets/reply) must be considered as per project requirement
	V	Q	rating_denominator, rating_numerator	Extremely large values
	P	T	source	Contains HTML tags, URL and content in a single column
	V	Q	name	Invalid names
	P	Q	In_reply_to_status_id, in_reply_to_user_id, retweeted_status_timestamp, retweeted_status_id, retweeted_status_user_id	Uninterested columns
	V	Q	rating_numerator, rating_denominator	Inconsistent rating
	P	T	doggo, floofer, pupper and puppo	Categorical variables represented in a seperate column
prediction	P	Q	p1, p2 and p3	Same observation in multiple columns
	P	Q	p1, p2 and p3	Dog breed has no standard (upper and lower case names)
prediction, tweet	P	Q	all	Missing records of observation i.e prediction (2074 values), tweet (2354)

3. Data Cleaning

After the data gathered was assessed, it was further cleaned to solve the previously identified issues. The three DataFrames were copied to protect the original data

- archieve_data_clean = archive_data.copy()
- prediction_clean = prediction.copy()
- tweet_clean = tweet.copy()

cleaning process:

- Merged all four columns (doggo, pupper, puppo, and floffer) into one column named stage. Also merged the favorite_count, retweet_count and the prediction table to the archive_data table using tweet_id.
- Dropped the uninterested observations for reply (78 values) and retweet (181 values) by rows. Also the features were dropped.
- Erroneous datatype were converted i.e to string using .astype() for tweet_id and pd.to_datatime() for timestamp.
- Source name in the source column was extracted using regex.
- The inconsistency in the rating was solved by performing feature engineering (where: rating = rating_numerator / rating_denominator).
- The invalid names starting with lower case were replaced with 'None'.
- Finally, the cleaned dataset was stored in the master DataFrame.

4. Conclusion

In the first iteration, eleven issues have been documented about the dataset. However, the master dataset is not free of issues, as Data Wrangling is an iterative process.

The wrangled data was stored in the twitter_archive_master.csv file with minor issues, and ready for Data Analysis. The file has 1971 observations and 15 features.