**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  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 | 1. replace all this names with none as this names were named after ‘This is ….’ In the text column  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

