# *Data Wrangling Report*

## BY AHMED SHETA

February 2021

This report is to identify the wrangling activities performed to collect the data of twitter account WeRateDogs which is an assignment in Data analysis professional Nano degree from UDACITY.com

## 1. Gathering Data

Data for this project was collected by three means:

1. Twitter_archive_matser.csv file was downloaded manually and was uploaded to the working directory and extracted to a Data frame using pd.df.read_csv()
2. Images-prediction.tsv file is hosted on a webpage, and it was downloaded by using its URL, by library request and imported to data frame
3. The last source of data was to scrap the account 'WeRateDog' using API tweepy to extract the required data, I used the hosted file because my Developer account application for twitter still under review, I downloaded the file and uploaded it to the working directory then import data from it into a data frame using json library

## 2. Assessment

1. In this step I managed to examine the imported data into data frames with both approaches; visual assessment by opening the data frame as a sheet and review it visually, while programmatically in notebook to get the required statistics and information necessary to assess data
2. Several quality and tidiness issues were found, findings like data accuracy, missing data, inconsistency,
3. Other data quality issues were addressed by project scope like keeping inly tweets with images entries.
4. The detailed findings and the taken actions are mentioned in the table below

## 3. Cleaning Data:

1. In this step all the quality and tidiness findings were addressed to clean the data and make it ready to be analyzed
2. The detailed cleaning steps are mentioned in the table below

## 4. Detailed Findings and action taken:

| DATA SET | FINDINGS | ACTION TAKEN |
|---|---|---|
| | **Quality Issues** | |
| DF_TWITTER | • Some tweets are replies and retweets | • Drop rows of replies and retweets |
| | • Some tweets are not with expanded_urls (no Image) | • Drop rows of tweets with no expanded_urls (no Image) |
| | • timestamp column dtype not correct | • Modify timestamp column dtype |
| | • tweet_id column dtype not correct | • Modify tweet_id column dtype |
| | • columns not required in master sheet (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls) | • columns not required in master sheet (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls) |
| | • dog stages columns contain 'None' as values | • convert 'None' in dog stages columns to Nan |
| | • rating_numerator contains wrong inputs | • correct rating_numerator values by reviewing text |
| | • rating_denominator contains wrong inputs | • correct rating_denominator values by reviewing text. |
| | • source of tweeting included in the source URL | • extract the tweets source from column 'source' |
| *DF_IMAGES* | • There are entries in df_twitter_raw without images data | • To be dropped but after merging with df_twitter to merge by tweet id |
| | • columns label not expressive | • Change the columns labels to be expressive |
| | • tweet_id column dtype not correct | • Change tweet_id column dtype to str |
| | • images breed prediction and if it's dog (distributed over 9 columns) | • identify the Non-dog images for tweet ids<br>• identify the most probable breed in a separate column |
| | • tweets contain Non-dog images | • identify the if image is dog in a separate column |
| *DF_TWEETS* | • column name from id not matching with df_twitter 'tweet_id' | • change the column name from id to tweet_id |

| | | |
|---|---|---|
| | • there are not required columns | • keep only the required columns ('id','retweet_count','favorite_count') |
| | • tweet_id column dtype not correct | • Change tweet_id column dtype to str |
| | **Tidiness Issues** | |
| | • Dog stage values are distributed over 4 columns | • combine the 4 dog stages columns into a 'dog_stages' column |
| | • data are in 3 separate data frames | • Set column (tweet_id) as index before merging<br>• combine a master data frame by concatenating the 3 clean Data frames with index (tweet_id) |

## 5. Outcomes

A clean master data frame is ready to be analyzed.