Reporting: wrangle_report

Introduction

This is a real world data wrangling project using three different data set obtain from WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dogs. The three data set was provided by Udacity for analysis. They include the tweet_json, image-prediction and twitter-archive-enhanced data set. The data wrangling process was spitted into series of steps which are :

Step 1: Gathering data
Step 2: Assessing data
Step 3: Cleaning data
Step 4: Storing data
Step 5: Analyzing, and visualizing data
Step 6: Reporting

Gathering data

The data was gathered from three sources. Udacity gave us the twitter-archive-enhanced, tweet-json and image-predictions data set. The twitter-archive-enhanced file was loaded into dataframe and the image-prediction data set was downloaded programatically using the request and get method. Tweepy was used to download the tweet_json file from the Tweeter WeRateDogs API

Assessing data

Each data set was loaded into data frame in other to identify some quality and tidiness issues in the data set.

Quality issues

1. The rating_denorminator colunms have value greater and less than 10
2. The rating_numerator columns have value considerable values greater than 10
3. Their is a lot of NAN values in the in_reply_to_status_id, reply_to_user_id
4. The timestamp in df_arch as the wrong datatypes
5. The source columns as unnecessary html anchor and href tag
6. Errors of name in the name columns e.g (infuriating, a, by, the, space, etc)
7. Considerable number of tweet without image
8. timestamp and retweet_status_timestamp are not a datetime variable
9. The names columns are not in the standard form
10. Consederable numbers of retweet

Tidiness issues

1. Removing the three empty columns from df_arch_clean table (i.e retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp)
2. The source and expanded_url column have several information in them
3. doggo, floofer, pupper, puppo columns in arch_df should be merged into one column named dog_stage
4. The retweet_count and favorite_count columns in the df_api table should be joined to the df_arch table

Cleaning Data

All imperfection was removed from each of the data set making it ready for the analysis.

Storing data

The clean data set was stored into a single dataframe called merge_df using the inner join method.