Wrangle Report

WeRateDogs Data

This project is about data wrangling and it includes Gathering using a variety of gathering techniques including API, Assessing the data for quality and tidiness and Cleaning the data using a define, code, test. and then analyze and visualize the cleaned data.

Data gathering:

- Twitter archive: downloaded manually.
- Image predictions: downloaded Programmatically.
- tweets data: Gathered using twitter API and tweepy package in python

Data Assessing:

- Missing values for doggo, floofer, pupper and puppo
- NaNs is "None" (str) in (name, doggo, floofer, pupper, puppo)
- Some dog stage values extracted with errors
- Some extracted name not names
- Some extracted values for rating_numerator and rating_denominator has errors
- Erroneous datatypes
- The dataset contains Retweets
- The dataset contains Replies

Data Cleaning:

- **Fill the missing data** from the tweet text by extracting the word of the dog stages (doggo, floofer, pupper and puppo)
- Convert the string “None” with Nan
- Correct the mistaken dog stage by listing the values then address the wrong ones
- Correct the mistaken dog stage by listing the values then address the wrong ones
- Delete mistaken name by extracting the world that begin with capital letter
- Correct the data types (tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, doggo, floofer, pupper, puppo)
- Delete the retweets by delete the rows that has values
- Delete the replies by delete the rows that has values