WRANGLE REPORT

1.Introduction

The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The dataset that is wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10.

2.Gathering Data

The data was gathered from three sources:

- **Enhanced twitter archive** -This part is kind of on-hand data, stored in the twitter_archive_enhanced.csv
- **Image prediction -** We get the image prediction data from web scraping
- **Twitter API** -This dataset is archived from Twitter's API and parsed from JSON to csv

Assessing data

Quality Issues

- **p1**,**p2**,**p3**: dog breed names are not all in lowercase
- **tweet_id**, **timestamp**: wrong data types
- missing data probably due to retweets in twitter_archive
- jpg url duplicates
- Many entries are not dogs, e.g., jaguar, mailbox, peacock, cloak, etc
- timestamp is str, should be datetime, remove +0000 in timestamp
- Abnormal values in rating_denominator, e.g., 170, 150, 130, etc. The rating_denominator is almost always 10
- Abnormal values in rating_numerator, e.g., 1776, 960, 666, 204, 165, etc. make no sense.
- source info redundant, not easy to read

Cleaning

- Merge the clean versions of archive, images, and twitter_counts_df data frames Correct the dog types.
- Create one column for the various dog types: doggo, floofer, pupper, puppo Remove columns no longer needed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.
- Delete retweets.
- Remove columns no longer needed.
- Change tweet_id from an integer to a string.
- Change the timestamp to correct datetime format.
- Correct naming issues and Standardize dog ratings.
- Creating a new dog_breed column using the image prediction data.