Wrangle Report

Introduction:

Real-world data rarely comes clean. The dataset that i wrangled in the project 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. WeRateDogs has over 4 million followers and has received international media coverage.

Data:

- Twitter Archive File
- Image Predictions File
- Additional Data via the Twitter API

Project Details:

- Introduction
- Data
- Gathering Data
- Assessing Data (Quality & Tidiness)
- Cleaning Data (Define, Code & Test)
- Analyzing & Visualizing Data

Gathering Data:

Data was gathered from three different sources:

1. **Twitter Archive File:** the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually. This archive contains basic tweet data (tweet_id, timestamp, name, etc.) for all 5000+ of their tweets as they stood on Augest 1, 2017.
2. **Image Prediction File:** The tweet image predictions file was downloaded programmatically using the Requests library from Udacity's servers.
3. **Additional Data via the Twitter API:** The third data was extracted from Twitter API using python's tweepy library.

Assessing Data:

After the data was gathered, assessment was performed using the following methods:

- .info()
- .head()
- .describe()
- .value_counts()

Quality Issues that were cleaned:

1. Remove "in_reply_to_status_id" and "in_reply_to_user_id" since they are all NaN
2. Change the timestamp to datetime format
3. Standardized dog ratings (ratings = rating_numerator / rating_denominator), then drop rating_numerator and rating_denominator
4. Remove retweets columns (i.e 'retweeted_status_id', 'retweeted_status_user_id' and 'retweeted_status_timestamp' columns will be removed.)
5. In archive_clean, removing HTML tags from source column
6. Convert the tweet_id in archive_clean into object(str) for merging
7. Get rid of image prediction columns in image_clean
8. Drop duplicated in image_clean
9. Convert the tweet_id in image_clean into object(str) for merging

Tidiness Issues that were cleaned:

1. Create one column for the various dog types: doggo, floofer, pupper, puppo
2. Combining 3 datasets into a big data since they are information about the same tweet

Cleaning Data:

This part of data wrangling was divided in three parts: Define, Code and Test the code. These three steps were on each of the issues described in the assess section.

Before cleaning any data from the dataset, it is benefits to create a copy of the three original dataframes. And then, I wrote the codes to manipulate the copies. If there was an error, I could create a new copy from the original. It's also help to debug quickly.

The issues found during the assessment process were cleaned and tested using the following methods and techniques:

- .replace()
- .astype()
- .drop()
- .sample()
- .to_datatime()
- .isnan()
- print()
- .split()
- .append()
- .apply()
- list()
- .duplicated()
- .merge()
- .to_csv()
- .loc[]
- Loops