Data Wrangling

I. Data Gathering

The data to be analyzed in this project is retrieved from several sources:

1. CSV file to be read using Pandas and stored in a dataframe
2. TSV file to be retrieved using Requests API then stored in a dataframe
3. Twitter using tweepy API, stored later in a json file to be read by Pandas and saved in a dataframe

II. Assessing

After removing retweets, each dataset was assessed separately to identify quality and order issues. For the first dataset, twitter archives, I noticed several issues:

- More Than 2/3 of the dogs in the dataset are not classified
- There are some dogs with 2 different classification
- 4 classsification colmuns that can be replaced to one column as the dog shall belong to only one category
- There are some missing values in the dataset
- Timestamp's data type shall be altered from object to datetime
- Some denominator values are 2-7 to be cleared
- I expected that the rating of WeDogsRate is always positive. However, a high rate is negative

For the second dataset, I focused on:

- Abstract columns' names which shall be altered to be more meaningful
- Duplicated images

For the third dataset, retrieved using Twitter API, I checked its shape. Its size is different than the second dataset

I will need to merge the three datasets together.

III. Cleaning

1. First, I merged the three datasets in one dataframe. To do this, I renamed the id column in the third dataset to tweet_id.
2. I created a column called classification to solve the multiple classification issues found in the dataset. Then, I dropped the 4 different classifications columns.
3. As most of the dogs are not classified, we assigned the most common classification value to null values
4. I set timestamp columns (timestamp, retweeted_status_timestamp) type to datetime
5. I removed rows with denominator lower than 10
6. For negative ratings, after diagnosing visually the dataset, I decided to add 10 to the value of rating_numerator with negative rating.
7. The columns' names of the second dataset are confusing. An end-user would not understand the meaning of p1, p2, p3. Hence, I renamed these columns.
8. I removed the rows containing duplicated images.

IV. Storing Data

After gathering, assessing and cleaning data, I stored the dataframe to a csv file called: "twitter_archive_master.csv"