Data Wrangling

this report described the data wrangling processes in this project.

Project objectives

The project main objectives were:

- ❖ Data Wrangling Efforts.
  - I. Gathering Data
  - II. Assessing Data
  - III. Cleaning Data

Gathering Data

In this phase, the three pieces of data were gathered and represented as pandas Dataframes:

• The “@WeRateDogs” Twitter archive (file on hand, manual download of 'twitter-archiveenhanced.csv')

• The tweet image predictions ('image-predictions.tsv'). This file was be downloaded programmatically using the Requests library from a provided URL.

• Each tweet's entire set of JSON data (with at minimum tweet ID, retweet count, and favorite count) in a file called 'tweet_json.txt' were stored using Twitter API and Python's Tweepy library. Each tweet's JSON data was written to its own line.

Assessing and Cleaning Data

While working with data, a number of observations were made. In the below table there are the observations along with actions taken in the Cleaning Step.

i. Quality

1) in_reply_to_status_id and retweeted_status_id variables are numeric
2) timestamp and retweeted_status_timestamp are not a datetime variable.
3) source value are formatted as <a>
4) rating_numerator are not always correctly accounting for decimals.
5) the dog names are not standardized.
6) Columns doggo, floofer, pupperand puppo has None for missing values
7) rating_denominator column has values less than 10 and values more than 10 for ratings more than one dog.
8) rating_numerator are not always correctly accounting for decimals.
9) text column has the link for the tweets and ratings at the end we can remove it
10) tweet_id variable are somtimes integers or floats (numeric).

ii. Tidiness

1) more than one stage is filled for a particular dog
2) source and expanded_urls have several information inside them.
3) columns doggo, floofer, pupper and puppo refer to the same measurement unit, i.e, dog stage
4) All datasets should be combined into 1 dataset only

Cleaning Data

1) Convert in_reply_to_status_id and retweeted_status_id to string.
2) Convert timestamp and retweeted_status_timestamp to datetime variable
3) remove <a> from source
4) Converting rating_numerator to integer by rounding the number to the nearest unit.
5) Replaced names None and unvalid names with np.nan.
6) Remove None values form doggo, floofer, pupperand puppo
7) Removed any rows with denominator more than 10.
8) Removed text rows from data.
9) Convert tweet_id to string by using astype() functions.

Tidiness

1) Created one column dog_stage and removed the 4 columns.
2) Delete unnecessary information related with source and expanded_urls
3) Combined all the 3 datasets into one pandas dataframe.