Wrangling report

In this report, I discuss the sources and the process of gathering my data. I then proceed to discuss the process of cleaning my data.

Gathering

I gathered data from three sources:

1. A csv file which was provided to me, which contained the twitter archive of WeRateDogs twitter account @dog_rates.
2. A url link to a tsv file, which I downloaded from the internet using the requests library.
3. I use the Twitter API – Tweepy to get the json of the tweets using their tweet ids. I save all the json to a text file, from which I extracted the retweet count as well as the favourites count for each tweet.

Cleaning

In this part, I document the quality and tidiness issues I discovered in my gathered data, and how I wrangled them.

1. The dataframe which contained the retweet and favourite counts, as well as the image prediction dataframe were merged into the cleaned twitter archive dataframe.

2. dog_stage (doggo, floofer, pupper and puppo columns) were not correctly extracted from text column.

- I combined the doggo, floofer, pupper and puppo columns to form one column called ‘dog_stage’. Then I dropped the doggo, floofer, pupper and puppo columns.

3. name column contains None as the NaN values, 'very' and 'a' are also a common names.

- I replaced None, very and a with nan values.

4. source column does not contain url in right format

- I replaced the source to either Twitter for iPhone, Vine, Twitter Web or TweetDeck to reflect the true source and changed the data type to category

5. rating_denominator is not 10 for all dogs

- changed rating_denominator to be 10 for all observations because this is the standard of the twitter account

6. changed dog stage column to category data type

7. extra characters in text column '&', '>

- Replaced & with '&' and replace > with '>'

8. maximum of rating_numerator seems unusually high

- changed the rating_numerators from the tweet texts to reflect the actual ratings from the tweet’s text.

9. decimal ratings do not show as decimal in rating_numerator

- I changed the rating numerators containing decimals to reflect their actual ratings from the tweet’s text.

10. timestamp is of object type

11. retweeted_status_timestamp is of object type

- change timestamp and retweeted_status_timestamp columns to datetime data type