Mustafa Olomi

Wrangle Report

Twitter Archive Table

- Timestamp needs to be formatted correctly
    - Changed these to datetimes

- Data also includes (RT) Retweet data for 181 items that need to be removed and in their own table (tw_retweets)

- Missing 59 expanded_urls (2297 vs 2356 total)
    - These turned out to be replies to other people and I put them in their own dataframe (tw_replies)

- Rating Denominator has values other than 10 extracted from text wrong
    - Some incorrectly extracted terms like 24/7 as the rating, 7/11, and dates which I corrected

- Missing doggo, fluffer, pupper data.
    - Extracted 37 from text with regular expression

- Drop retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp (not used)

- Dog names are incorrect and have random words at times
    - ‘a’ was used as a name 55 times, ‘the’ was used 8 times etc
        - I used regular expressions to find and drop these
- Putt hashtags in their own column, used a lot for marketing and should be helpful

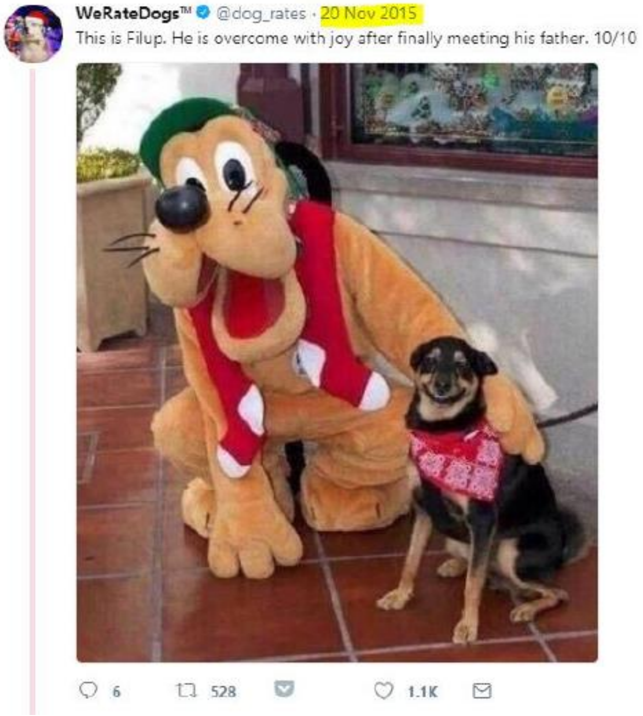
Twitter Archive Table

- Retweets of 181 items that need to be in their own table
- Contains 78 replies (in_reply_to_user_id) need to be on their own table
- Doggo, Fluffer, and Pupper are not fields so merged them in one column and removed the others

Cleaning and wrangling the data was a finetuning project that turned out to be easy but took many steps
Looking back now the things that helped most was creating save points in the data with .copy() of my dataframes so that if I had an error in merging, filtering, or deleting a dataset I did not have to start from the beginning.

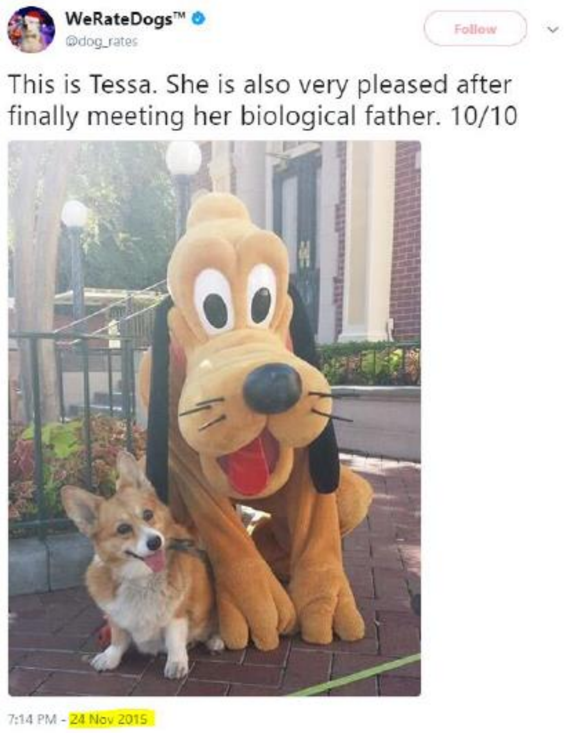
I also spent a lot of time googling the exact tweets to look at characteristics of the tweet itself on twitter so I could understand the information I was working with. That is how I found that if the reply status is not null and there is an expanded URL then it is essentially a regular post with the same characteristics like a rating and a picture, except it’s in reply to a previous post they made. It would not make sense to categorize these as replies or retweets because they are a new post it just has something in common with a previous post, ie. Same theme or update on a pup.

See below example.



WeRateDogs™ @dog_rates · 20 Nov 2015

This is Filup. He is overcome with joy after finally meeting his father. 10/10



WeRateDogs™ @dog_rates

This is Tessa. She is also very pleased after finally meeting her biological father. 10/10

7:14 PM - 24 Nov 2015