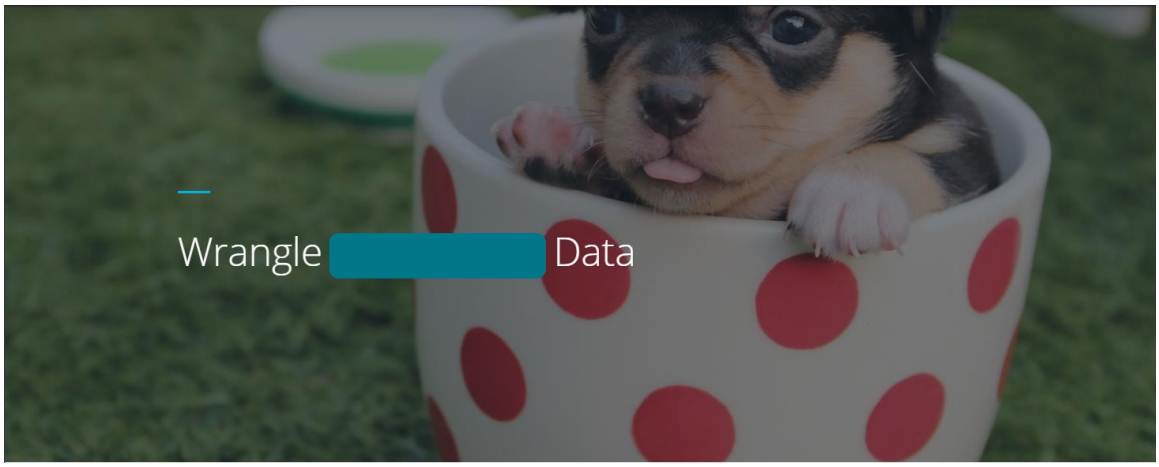

Wrangle report

GATHER, ASSESS, AND CLEAN

Kristijan Bakaric | Udacity – Wrangle Data | 13th of February 2018

GATHERING DATA

Gathering Data for this Project was done in three parts, from three data sources as described below:

- The WeRateDogs Twitter archive via **manual** download of csv file.
- **Programatically** for Tweet image predictions via https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv using **requests** library
- Query the **Twitter API** via tweet_ids from weratedogs twitter archive for each tweet's JSON data using Python's **Tweepy** library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's **JSON** data was written to its own line. Then read this .txt file line by line into a pandas DataFrame. Minimum was defined as retweet count and favorite ("like") count at minimum.

ASSESSING DATA

After gathering each of the above pieces of data, data was assesd visually and programmatically for quality and tidiness issues. I detect and document eight quality issues and two tidiness issues

Visual assessment

Tidiness

During the visual assessment I got familiar with the structure of the 3 separate tables and noted:

- twiter_arc_enc dataframe: there are None values for doggo, floofer, pupper, puppo columns and since it there is always one value from four filled it means that data needs to be condensed in a column dog_stage
- dataframe twiter_arc_enc can be joined onto twiter_arc_enc
- twitter_api dataframe dataframe can be joined to twiter_arc_enc
    - all three can be joined on tweet_id since they are unique per row

Quality

- twiter_arc_enc dataframe: Source column contains html links embeded in html as strings

- tweet_image_pred dataframe: p1. p2 and p3 columns have a lot of variations in the cells, like for example - or as a space between words, varied capitalization – lower capitalisation and remove dash and underscore
- remove p1_dog, p2_dog and p3_dog == False ("Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.")

Programatic assessment

Quality

- timestamp should be converted to time object (done)
- rating_denumerator and rating_numerator have max values of 1776 and 170 - investigate since 75th percentile is 12 and 10 respectively. (will not fix since it is anyway fictional)
- rating_denominator should be 10, so everything that is different than 10 should be converted to 10.
- tweet_image_pred dataframe: name column has non names as "a", "an", "his", "my", "one"
- twitter_api dataframe: created at should be converted from object to timestamp
- in all dataframes tweet_id should be converted to object from integer

CLEANING DATA

First steps were to:

- copies of the original dataframes
- in twiter_arc_enc_copy dataframe keep only rows where retweeted_status_user_id is NaN
- convert in all dataframes tweet_id from int to object
- join twiter_arc_enc_copy and tweet_image_pred_copy on twitter_id
- join merged and twitter_api_copy on twitter_id
- keep only rows where jpg_url is not NaN
- remove uneccessary columns

Following issues identified in the assessment phase were defined, coded and tested:

1. rating_numerator was extracted one more time from status tweets column to include also decimal numbers
2. replace "None" with NaN and convert doggo, floofer, pupper, puppo None to NaN's. Create a flag for a dog stage and use the flag to create a new column called dog_stage.

3. Name column is not always representing a name (a, an, my, one, his and None). Replace these values with NaN.
4. Parse time column (object) as timestamp.
5. rating_denominator should be 10, so everything that is different than 10 should be converted to 10.
6. Clean html tags from for_cleaning_copy["source"] strings.
7. for p1, p2 and p3 columns - remove non dogs, i.e. dog flag == False
8. p1, p2 and p3 columns: lower case and remove "_" and "-"
9. Parse created at as datetime

Cleaning phase was finalized by exporting cleaned dataframe into
twitter_archive_master.csv