Paras Patel
Data Analytics – Udacity
Wrangle Report

Data wrangling effort's can be summarized in 3 stages: gathering, assessing, and cleaning data. After gathering each dataset in the project, they were assessed for at least (8) quality issues and (2) tidiness issues and later cleaned.

The first DataFrame consists of data from twitter_archive_enhanced.csv. This file was downloaded and the following quality and tidiness issues were assessed for cleaning. Cleaning method specified:

<u>Quality Issues</u>

1. Delete rows with a rating numerator greater than 14 and value of 0, and replace rating numerator with value specified in text.
    - Pandas 'query' for multiple conditions.
    - Regex python
2. Convert 'timestamp' column from object to date-time.
    - Pandas 'todatetime' method.
3. Delete rows with rating denominator value not equal to 10. We can investigate texts from these rows, but since there are only 23 rows out of 2300+ with a non-10 denominator, we'll delete these rows.
    - Pandas 'query' method.
4. Some dog names are not capitalized. Capitalize all dog names. Some rows do not specify dog names, search text to identify additional dog names.
    - Title function (python string method)
    - Regex python
5. Dog stages are missing for majority of tweet ID's. Scan text column for dog stage.
    - Write function to loop through 'text' column. Identify keywords (dog stages) in text. If keyword is present, that keywords is assigned as dog stage.
6. Rows in the expanded url column have duplicated url's.
    - Split string and index first value.
7. Convert 'tweet_id' column to object (string).
    - Pandas 'astype' converter
8. Filter only original tweets. Remove 'RT' rows.
    - String 'contains' method to filter. Store list of indexes and drop from df1

Tidiness Issues

1. Doggo, floofer, pupper, and puppo columns in twitter_archive_enhanced.csv should be combined into a single column, as this is the one variable that identifies the stage of the dog.

Other Issues

1. Remove the hyperlink and <a/a> tag for url's in the 'source' column.
    - Split string in hyperlink and index url.
2. Drop columns 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', and 'retweeted_status_timestamp'. No valuable information can be extracted from these columns.
    - Pandas 'drop' method.

The second DataFrame consists of data from image-predictions.tsv. This file was downloaded and the following issues were assessed for cleaning:

Quality Issues

1. Remove duplicated rows
    - Pandas 'duplicated' method
2. Convert 'tweet_id' column to object (string).
    - Pandas 'astype' converter
3. Create separate column for dog breed based on image prediction model results. If first prediction is identified as a dog, than dog breed will equal 'p1', if not, p2 will be evaluated next, and then p3 if necessary.
    - Write a for-loop to loop through each row to check 'p#_conf' columns and assign dog breed as appropriate.

Other Issues

1. Replace underscore with space for dog breeds.
    - String replace method.
2. Remove 'jpg_url' and 'img_num' columns - no valuable information.
    - Pandas 'drop' method

The third DataFrame consists of retweet and favorite counts for each Tweet ID identified in the first data frame. The following item was assessed for cleaning:

Quality Issues

1. Convert 'Created At' column to date-time.
    - Pandas 'todatetime' method.

2. Convert ‘tweet_id’ column to object (string).

Tidiness Issues

1. Merge critical metrics from all 3 dataframes by ‘tweet_id’ into one master dataframe.