Wrangle Report

Project: Wrangle and Analyze Data
Jungun Goo
September 2018

1. Gathering Data

- Data is gathered from 3 different resources and saved into 3 dataframes;
  1.1) df <- pd.read_csv('twitter-archive-enhanced.csv')
  1.2) img_prdt <- pd.read_csv('image-predictions.tsv', sep='\t')
  1.3) df_retwt <- extracted from Twitter API with tweet_id
    - I collected data, the number of favourites and retweets on tweet_id. And I saved data as text file, tweet_json.txt, then read the file into dataframe and handled it for wrangling.

2. Assessing Data

2.1) Quality Issues

- Capitalizing all dog's name, and fix 'None' to Null_value.
- The denominator columns cannot have zero value.
- The following columns have wrong data types: tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp
- The following columns don't seem to be necessary as we won't consider the data retweeted: retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.
- the favorites and retweets columns' data type are wrong. Should be fixed to an integer type.
- The entries don't have pictures should be removed from the dataset.
- In df, nulls represented as 'None' in columns 'name', 'doggo', 'floofer', 'pupper','puppo'. After condensing these column into one column, fix 'None' value to Null_value so it can be counted as empty data.
- Dog's breed column is needed to be capitalized in the dog prediction dataframe to compare with other columns.(p1, p2, p3)

- Wrong data types of 'tweet_id' in the both dataframes, df_clean and breed_prediction.

2.2) Tideness Issues

- The two dataframes(df, df_twt) can be unified into one on tweet_id.
- The doggo, pupper, puppo, and floofer columns can be condensed.
- Rating_numerator and denominator should be one variable rating.

3. Cleaning Data

- I copied 3 dataframes into; df_clean, img_prdt_clean and df_twt_clean

3.1) *Define:* The columns 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp' are not needed.
- Dropped 3 columns and saved it.

3.2) *Define:* Condensing dog type(doggo, floofer, pupper, puppo) into a column.
- 394 entries are classified dogs into 4 types.
- However, it turned out that 14 entries have 2 types.
- So, total 380 entries have dog types value after cleaning.

3.3) *Define:* We only consider the ratings with images. Getting rid of the entries without a value of the column(expanded_urls).
- 59 entries didn't have expanded_urls value, and it's removed from the dataframe.

3.4) *Define:* Retweet/favorites data can be joined into the main dataframe on tweet_id.
- Two dataframes, df_clean and df_twt_clean, are joined into one on tweet_id.
- df_clean has two more columns, favorites and retweets.

3.5) *Define:* Some favorites and retweets value have Null-value. Find these value by connecting API one more.
- To collect as many as data, I reconnected to twitter API one more with tweet_id.
- And 3 more entries are gathered and added to the dataframe.
- There are 2283 entries in the main dataframe.

3.6) *Define:* Data type of favorites and retweets columns seem to be wrong. Should be fixed to an integer not a float.
- I used the function, astype(int), to convert data type.

3.7) *Define:* Wrong data types (in_reply_to_status_id, in_reply_to_user_id, timestamp)
- I used the function, astype(str), to convert data type.
- To convert time type, pd.to_datetime is used.

3.8) *Define:* Wrong data types of 'tweet_id' in the both dataframes, df_clean and breed_prediction.
- Data type of tweed_id is converted to an object type in both dataframes, df_clean and img_prdt_clean.

3.9) *Define:* The denominator value cannot be zero-value. Create a new column 'rating'(=rating_numerator/rating_denominator). Drop the both of the columns(rating_numerator and rating_denominator)
- Denominator cannot have zero value, and it's removed.
- Added a new column, rating, with a value of (numerator/denominator).
- Dropped two columns of rating_numerator, rating_denominator.

3.10) *Define:* Capitalizing all dog's name, and fix 'None' to Null_value.
- To have a consistency for data.

3.11) *Define:* Dog's breed column is needed to be capitalized in the dog prediction dataframe to compare with other columns.(p1, p2, p3).
- To have a consistency for data.

3.12) *Define:* Combining two dataframes into one for the convenience of analyzing and managing data.
- Joined df_clean and img_prdt_clean on tweet_id.

3.13) Final Test
- df_clean has 2283 valid data entries with rating, favorites and retweets value.
- However, the columns related to type and prediction have 2067 valid entries.

4. Storing Data
- Store the final and clean dataframe into a CSV file, twitter_archive_master.csv