Wrangle Report

Submitted by Bharat Garg

This report contains an overview of Data wrangling, which consists of:

1. Gathering data
2. Assessing data
3. Cleaning data
4. Storing, analyzing, and visualizing your wrangled data

Gathering Data

The following methods are used to gather data:

1. The WeRateDogs Twitter archive is downloaded manually (twitter_archive_enhanced.csv).
2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is downloaded programatically using requests library.
3. The tweet info for WeRateDogs Twitter archive is downloaded via api. The data is stored in JSON format and then read line by line to extract relevant information, such as retweet count etc.

Assessing Data

The assessment of data is done visually and programatically for quality and tidiness issues.

Visual Assessment

Tidiness Issues Display random observations from all 3 tables to visually inspect for tidines issues:

1. twitter_arc dataframe: None values for doggo, floofer, pupper, puppo columns, since there is only one value from four for every row filled, it means that data needs to be melted in a column dog_stage
2. tweet_image dataframe: dataframe can be joined onto twiter_arc
3. twitter_api dataframe: dataframe can be joined to twiter_arc

Quality Issues

1. twiter_arc dataframe: rating_numerator does not have correct numbers extracted from status tweets, for eg. tweet_id 883482846933004288 has a decimal 13.5.
2. twiter_arc dataframe: Remove the columns which are not needed - in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.
3. twiter_arc dataframe: Source column contains html links embeded in html as strings.
4. tweetimage dataframe:p1. p2 and p3 columns have a lot of variations in the values, for eg. - or between words, different capitalsiation - remove - and _ & capitalize.
5. tweet_image dataframe: 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.").

Programmatic Assessment

Quality Issues

1. twitter_arc dataframe: There are 6 columns that have non-null objects.
2. twitter_arc dataframe: Timestamp to be converted to time object.
3. twitter_arc dataframe: rating_denominator should be 10, so all other values to be converted to 10.
4. tweet_image dataframe: name column has strings such as "a", "an", "his", "my", "one".
5. twitter_api dataframe: created_at should be converted from object to timestamp.
6. In all tables, convert tweet_id to a string, instead of int.

Cleaning Data

The following steps were undertaken to clean the datasets:

1. Create copies of the original dataframes
2. In twiter_arc_copy dataframe, remove rows where retweeted_status_user_id is not null, i.e. they aren't original tweets
3. Convert tweet_id from int to string in all the 3 dataframes
4. merge twiter_arc_copy, tweet_image_copy and twitter_api_copy on twitter_id
5. Remove rows where jpg.url is NULL.
6. Remove uneccessary columns. Following issues identified in the assessment phase were defined, coded and tested:
7. rating_numerator was extracted from 'text' column using str.extract
8. Convert doggo, floofer, pupper, puppo None to NaN's. Combine to create a new column called dog_stage.
9. Name column has following non-name entried (a - 55, the - 7, an - 6, my - 1, one - 4, his - 1). Replace these values with NaN.
10. Convert time column (object) to timestamp.
11. rating_denominator should be 10, so all other values are converted to 10.
12. Clean html tags from source column.
13. for p1, p2 and p3 columns - remove rows which are not dogs, i.e. dog flag == False, then drop the p1_dog, p2_dog, p3_dog columns
14. p1, p2 and p3 columns: lower case and remove "_" and "-"
15. Convert created_at as datetime

After completing the above steps, the file is stored as twitter_archive_master.csv using pd.to_csv function.