Wrangle Report

by Haotong Li

May 2019

This short report describes the wrangling efforts involved in completing the “WeRateDogs” project as part of Udacity’s DAND.

The Data Wrangling process consists of 4 parts:

1. Gathering Data
2. Assessing Data
3. Cleaning Data
4. Storing, Analyzing and Visualizing Data

1. Gathering Data

There are 3 sources for data gathering:

1. ***twitter_archive_enhanced.csv***: Directly download CSV file

   Use *pd.read_csv* import into pandas data frame.

2. ***image_predictions.tsv***: Programmatic download from Udacity's server

   The tweet image predictions is present in each tweet according to a neural network. This file is hosted on Udacity's servers and downloaded programmatically using the *requests* library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3. ***tweets_df***: Query from Twitter API

   Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's *tweepy* library and store each tweet's entire set of JSON data in a file named *tweet_json.txt* file.

2. Assessing Data

The three saved data frames were first assessed programmatically in Jupyter Notebook with *pandas*, then visually in Excel/Google Sheets.

Several issues were detected and listed below:

Quality Issue (issues with content)

1. *twitter_archive_df*:
   1.1 Only want original ratings (Delete the 181 retweets and 78 replies)
   1.2 Don't need those columns: 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'img_num', 'expanded_urls' and 'jpg_url'
   1.3 All rating_denominator should be "10" and some rating_numerators are extreme values
   1.4 Since all the denominator is 10 after last step, we can get rid of rating_denominator column and change rating_numerators to 'rating'
   1.5 Many dog names are meesed up, such as "such" "a" "quite"
   1.6 timestamp have extra "+0000"
   1.7 timestamp's datatype should be converted to "datatime"
2. *img_predictions_df*:
   2.1 Remove "_" and capitalize the image predictions.(p1, p2, p3 column names)

Tidiness Issue (issues with structure)

0. Join 3 DataFrames.
1. *twitter_archive_df*:
   1.1 Dog stage's 4 variables: doggo, floofer, pupper, puppo should be in single column of categorical variable
   1.2 Dog stage have 'None' instead of np.nan
2. *img_predictions_df*:
   2.1 Image prediction should be summarized to one column 'dog_breed'
3. *tweets_df*:
   3.1 Renamed the column id to tweet_id for easy merging. (Already done when create tweets_df)

3. Cleaning Data

Tidiness Issues:

Issue 0: Inner join *twitter_archive_df_clean*, *img_predictions_df_clean*, and *tweets_df_clean* on tweet_id

Issue 1.1: Create *'dog_stage'* variable which is made by extracting the dog stage variables from the text column

Issue 1.2: Dog stage have 'None' and replace 'None' to np.nan

Issue 2.1: Use the ture prediction to fill in *dog_breed* column. If no ture prediction, fill in use np.nan

Quality Issues:

Issue 1.1: Select the rows from *twitter_archive_df* that *retweeted_status_id* and *in_reply_to_user_id* columns that is null

Issue 1.2: Remove columns: 1.*in_reply_to_status_id*, 2.*in_reply_to_user_id*, 3.*retweeted_status_id*, 4.*retweeted_status_user_id*, 5.*retweeted_status_timestamp*, 6.*img_num*

Issue 1.3: Drop rows where denominator of rating != 10 and where numerator rating >> 10

Issue 1.4: Drop *rating_denominator* column

Issue 1.5: We find all the incorrect names have lowercase first letters. We will change those names to None, then change all the None to np.nan

Issue 1.6 &1.7: Use *str.strip* to remove "+0000" and use *pd.to_datetime* convert timestamp's datatype

Issue 1.8: Use regular expression and *Series.str.extract* to find real source between tags > and <

Issue 2.1: Use *Series.str.replace* to remove '_' and use *Series.str.capitalize* to convert 'p1' 'p2' 'p3'

4. Storing Data

Store the clean df in CSV file with name using *.to_csv('twitter_archive_master.csv')*