Report on Data Wrangling
By Aastha Arora

Data Wrangling

The tweet archive of Twitter user @dog_rates, also known as WeRateDogs was wrangled to make the data suitable for analysis and visualization. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Data wrangling process was divided into three parts:

1. Gathering Data

To analyze the twitter data for WeRateDogs, data was collected from three different sources.

a. Twitter Archive Data
The WeRateDogs Twitter archive data was provided as a csv file. The data was manually downloaded from the link
https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv
The file contains the basic data for 5000+ tweets which have been filtered for tweets with contain a rating (total 2356 tweets)

b. Tweet Image Predictions
Every image in the WeRateDogs Twitter archive was passed through a neural network to classify the breed of the dogs. The tsv file containing the tweet image predictions was programmatically downloaded from the URL-
https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

c. Additional Data via the Twitter API
Each tweet's retweet count, favorite ("like") count and URLs were saved by querying the Twitter API using the tweet IDs in the WeRateDogs Twitter archive. Python's Tweepy library was used to query the API and each tweet's JSON data was stored in tweet_json.txt file.

2. Accessing Data

Data was accessed visually and programmatically for quality and tidiness issues. To complete the requirements of the project at least 8 quality issues and 2 tidiness issues in the dataset were identified. The data issues listed below do not form a comprehensive list.

Data Quality Issues

twitter_archive table

- The columns- name, doggo, floofer, pupper and puppo have multiple rows with the value as None
- Some tweets are retweets (in_reply_to_status_id column not null), need to remove them. We only want original ratings (no retweets) that have images
- Values in rating_numerator column are incorrect if the given rating is a decimal number. For E.g. 9.75/10 has been parsed as rating with numerator 75
- Text containing date 9/11 is interpreted as rating for 2 tweets
- Text mentioning store 7/11 is interpreted as rating for a tweet
- Tweet index 2335 has incorrect rating
- Source column has html tags
- Timestamp column is a string. It needs to be converted to datetime column for anlysis
- Table has extra columns which are not needed for analysis

image_predictions table

- p1, p2, p3 columns have '_' instead of space between words. The needs to be corrected to make it more readable.
- Column names are not descriptive

Tidiness Issues

- Information about one type of observational unit (tweets) is spread across three different datasets.
- Variable dog_stage is spread across four columns (doggo, floofer, pupper, puppo)

3. Cleaning Data

After accessing the datasets for quality and tidiness issues, data was cleaned using a define, code, test methodology. A copy of the data was made to preserve the original data. Each data was issue handled one by one and the code was tested after making each change in the dataset.

4. Storing the Data

The cleaned dataset after the data wrangling process was saved in the ‘tweet_archive_master’ csv file