Data Wrangling Report

By Haitham M. Adham
January 2021

As an assignment for Udacity Data Analysis Professional Nanodegree Program, This report illustrates the wrangling efforts involved in completing the “WeRateDogs” project.

Data Gathering

In this step data were collected from three main sources:

1. **Twitter Enhanced archive.CSV,** manually downloaded from the Udacity servers.

2. **Image_predictions.tsv** , The tweet image predictions are present in each tweet 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. **Twitter API** , downloaded by querying the Twitter API via Tweepy library and stored in a file named tweet_json.tx.

After gathering i loaded files into data frames.

Data Assessment

In this step started with Visual Assessment By checking through them and i noticed the the HTML tags in source column which its a quality issues and the unwanted columns of retweets and replies, also the 4 columns of dog stages.
Then began with Programmatic Assessment using panda’s functions and detected the following issues :

Quality Issue:

Twitter archive:

1. Remove Retweets and Replies as it's a key point.
2. Drop unnecessary columns filled with missing values “in_reply_to_status_id”, “in_reply_to_user_id”, "retweeted_status_id", “retweeted_status_user_id”, “retweeted_status_timestamp”.
3. Convert invalid lowercase entries in the "name" column to NaN.
4. String "none" in the 4 dog stage columns should be replaced with empty string("") to merge the 4 columns later.
5. Modify decimal rating_numerators to be valid.
6. Replace entries that it doesn't contain rating at all by missing values.
7. Modifying ratings that is multiplied by the number of dogs in pictures.
8. Assign all denominator ratings to be equal to 10.
9. Correct rating_numerator less than 6.

10. Remove "+0000" from the 'timestamp' column.
11. Convert 'tweet_id' datatype in archive_clean, image_prediction_clean, and api_clean to object and data type of 'timestamp' to datetime.
12. Extract HTML link from the content of the source column to make it more readable.
13. Remove duplicates in the 'expanded_urls' column.

Image Predictions:
14. Rename columns of image_predections dataset to be more descriptive.
15. Remove duplicate urls in the 'jpg_url' column.

Tidiness Issues
Twitter archive:
1. Merge the 4 dog stage columns into a single column in the archive table.

Image Predictions:
2. Reshape Image Predictions from wide to long.

Twitter API Data:
3. Merge cleaned Twitter API Data and cleaned the archive data table into a single data frame.

Data Cleaning
In this step quality and tidiness issues were cleaned in 3 steps: Define, Code and test.