UDACITY
Data Analysis Nanodegree

PROJECT 5: Data Wrangling Report on 'WeRateDogs' Twitter Page

Deepak Nandipati
Date Created: May 7, 2019

Introduction

The purpose of this project is to utilize data wrangling and cleaning skills learnt from Udacity classroom to sort tweet archive of Twitter page "WeRateDogs", going by username @dog_rates. This twitter page currently consists of 8.11 Million followers with more than 10.3 thousand tweets that mainly constitute of cute dog pictures.

Gathering Data

Three pieces of data was needed for the completion of this project. Two of which were extracted from Udacity servers and one was derived from Twitter API and JSON using '*tweepy'* function.

- **Twitter Archive File:** File was provided from Udacity as "twitter_archive_enhanced.csv" which contains the twitter archive data of "WeRateDogs"

- **Tweet Image Predictions:** File was hosted on Udacity's servers and pulled from server using Request function. Document consists of neural network of image predictions (object, animal, breed of dog, etc.) present in each tweet. URL is as follows: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

- **Twitter API and JSON:** Acquired consumer key/secret along with access token and secret from Twitter Dev. Using tweet IDs in Twitter archive, queried the Twitter API for each tweets JSON data extracted using 'Tweepy' library. The tweet's entire set of JSON data was written to its own line and saved in a file called 'tweet_json.txt'. This file was uploaded to notebook and converted to pandas DataFrame for wrangling efforts.

Data Assessment

Data was assessed via visual and programming tools for quality and issues of tidiness. Below are the initial issues that were found for each data source:

Twitter Image Prediction Quality Issues:

- P1_dog pertaining to false needs to fixed, as it pertains to image predictions of objects or other animals.
- 'jpg_url' and 'img_num' column are not needed for analysis
- P2_dog and p3_dog column with low probability of being dog should also be removed from dataset for consistency
- Need to remove existing duplicates
- Keep only data pertaining to dogs and drop other animals/objects

Twitter API Data:

- Need to merge data frames and remove duplicates if present
- Have to rename id to same as other data frames and convert to string variable

Twitter Enhanced Archive Quality Issues:

- Denominator is inconsistent, there are 23 cases in which denominator are greater or less than 10; we can drop those datasets.
- Unnecessary columns such as retweets or responses information
- 181 retweets which need to be removed
- Replace string 'NaN' with datatype error NaN – and making them blank
- Some Dog names are incorrect; all names without a capital letter are not pertaining to names.
- Numerators need to be rectified for outliers and float values
- Need to delete duplicate values

Tidiness Issues:

- Need to remove p1_dog = False dataset as it does not pertain to dogs
- A new column needs to be added for dog category and categorize these columns ‘doggo’, ‘floofer’, ‘pupper’ and ‘puppo’ to a single column
- All tables need to be merged to a single dataframe

Data Cleaning

Before any of the data was cleaned, I made a copy of the data frames and worked with the copies to make changes. This process avoids any actual changes to original data, and can save the cleaner version as new file. Once the data was cleaned for individual data frames, the datasets have been merged into a master dataset and excess columns have then been removed.

Brief summary of cleaning I did for this dataset:

- I have successfully removed irrelevant columns needed for analysis, deleted columns that do not pertain to dogs from ‘df_written_clean’, and removed columns with low image predictability for accurate data content.
- Removed duplicates, and null values but also replaced empty values with blanks
- Extracted dog types and compiled to single column: ‘stage’
- Sorted dog names so all are actual names
- Standardized denominator to 10 and removed any other denominator for consistency
- Numerators that did not amount the decimal values were fixed

All datasets were tested for verification of changes and once confirmed datasets were combined and analyzed.