WERateDogs

Introduction

This project demonstrates the data wrangling process for the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. In this report I will provide a brief description of the data wrangling techniques that were used to gather, assess and clean the dog twitter archive.

Gather data

The following files were gathered for the analysis: **The WeRateDogs Twitter archive -** This file (archive.csv) was downloaded manually and consists of basic tweet data for 2300+ tweets from WeRateDogs. **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) was downloaded programmatically from Udacity.

Each tweet's retweet count and favorite ("like") count -This file (tweet_json) contains JSON data for each tweet indicating the retweet and like counts.

Assess data

The three files obtained in the gathering phase were loaded into individual Pandas data frames for assessment.
Each of the data frames were evaluated visually and programmatically. The following quality and tidiness issues were observed during the assessment.
Quality issues:

twitter archive table
  •Some name column entries are not names
  •Some denominators are incorrect
  •Remove unused dog stage columns
  •Remove retweets

image prediction table
  •Remove entries that have p1_dog, p2_dog, & p3_dog values

set to false. These are not dogs of any kind.
- •Remove duplicate jpg_url entries
- •Remove unused img_num column

api table

Just 3 columns needed `id`, `retweet_count`, `favorite_count`

Tidiness issues:

Clean data

The quality and tidiness issues were cleaned using programmatic techniques such as:

- • Dropping unnecessary columns from the tables
- • Removing rows that consisted of retweets
- • Removal of rows with duplicate information
- • Deleted rows that did not have any dog predictions at all
- • Combining all three data frames into a single data frame