Wrangle report

Introduction

The dataset that we wrangled, analysed and visualised is 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 dogs. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 13/10, etc. Why? Because "they're good dogs Brent." source

Gathering data

Some of the data were sourced from the WeRateDogs Twitter archive made available exclusively to Udacity, where we gathered it by downloading it directly. Udacity ran it through a neural network to detect the presence of a dog in every image in the tweets and predicted the breed of the dog where applicable. A result of this was provided to us and we programmatically downloaded it into our workspace. We further queried Twitter's API using the Tweepy module, to enrich our data further.

Assessing data.

The Twitter archive data consisted of 2356 tweets and 17 columns which included tweet_id, timestamp, source, text, retweeted_status_id, expanded_urls, rating_numerator, rating_denominator, name, doggo, floofer, pupper, and puppo, among others. The image prediction data provided to us consisted of 2075 tweets and 12 columns which included tweet_id, jpg_url, img_num, p1, p1_conf, and p1_dog, among others. We successfully queried 2327 tweets with the columns tweet_id, retweet_count, and favorite_count.

Quality Issues

Various quality issues were observed, such as:

Twitter archive dataset:

- The Twitter archive dataset had 181 retweets and 78 reply tweets.
- Most tweets had no dog stage.
- The ‘timestamp’ column was an object and not a DateTime format.
- Many records didn't have dog names while others had incorrect names such as 'such', 'a', 'quite', 'not', 'one',etc.
- Some tweets appeared to have ratings that were not about dogs while others had incorrect values in the rating numerator and denominator.
- Some tweets were not about dogs.
- The tweet text also ended in a URL instead of just text.
- Some rows in the expanded URLs were empty.
- The source was an HTML tag.

Image predictions dataset:

- Unuseful columns such as image number.
- Some tweets (about 324) didn't have a dog image.

API dataset:

- Had no issues

Tidiness Issues

We also identified the following tidiness issues:

- Dogs stages column headers were variables.
- Three levels of image predictions.
- The three datasets should have been combined into one dataset.

Cleaning data

All the issues above were cleaned successfully, save for the two issues of some tweets that had ratings that were not about dogs while others had incorrect values in the rating numerator and denominator, and some tweets that were not about dogs.

In the cleaning stage, I first made a copy of the original data before cleaning. Cleaning included merging individual pieces of data, dropping irrelevant columns and invalid rows, creating relevant columns such as dog stage and dog breed and populating them with the correct data, changing the format of some data, and removing irrelevant data in some rows, among others.

Storing data

After the cleaning, we saved the gathered, assessed, and cleaned master dataset to a CSV file named "twitter_archive_master.csv".

Analyzing and visualizing data

In our analysis, we found that the tweet with the tweet_id 744234799360020481, tweeted on 2016-06-18 at 18:26:18, was the most favourited (144310 times) tweet and retweeted, (70366 times) in our dataset. The Golden Retriever was the most common dog breed in our dataset with 156 entries. Finally, we noted that there is a weak positive correlation between the favourite count and the rating numerator at 0.07 and plotted a scatter plot to visualise it.