Cherie Goodenough
2/9/2019

Wrangle Report for Udacity We Rate Dogs Project

Discussion regarding steps taken in wrangling process 300 wd minimum. No code.

For this project, Udacity asks students to practice the data wrangling process using data from the WeRateDogs Twitter account. This account accepts and post pictures of dogs sent by their owners. Each dog is rated out of 10, with the numerator usually being above 10 because, as the WRD owner famously stated, “They’re all good dogs, Brent.”

Students are to gather data (some provided, others to be obtained), assess, clean and provide some brief analysis an visualization. This document discusses the wrangling activities undertaken.

For this project, I chose the following analysis questions:

- Do the favorite and retweet counts track the WRD ratings (i.e., are the more highly rated dogs also favorited/retweeted more?
- What breed of dog is the most popular with WRD readers?
- Have ratings gone up over time?

1. Gather

Three separate data sources from the WeRateDogs Twitter (WRD) account were gathered for this project:

1.1. A tweet archive of the WRD feed provided to Udacity by WRD and provided by Udacity to students. This archive is in the form of a csv file and contains information for which WRD is the source, including the rating numerator and denominators, urls, timestamp and tweet text. Since many of the tweet texts contain the name of the dog, Udacity extracted names programmatically into a new column. WRD also refers to dogs by unique types: doggo, floofer, puppo, and pupper. Udacity also pulled these descriptors out of tweet texts into columns. The data file is twitter_archived_enhanced.csv and it is included in this submission.

1.2. A tsv data file created by Udacity that used an AI application (not provided) to analyze the tweet photos and predict the breed of the dog in the photo. There are first, second, and third best guess columns along with their p values and a Boolean for whether the guess predicted that the photo was a dog. Data is included in this submission

1.3. by using the Twitter API and python package tweepy. Using the tweet ids from the archive dataset, we used the api to get information about the tweet, specifically the favorite count, retweet count and url of each tweet. The results were stored in a dictionary using the tweet id as key and setting the value as a list containing the favorite count, retweet count and url. This dictionary was converted to a dataframe then saved as a csv file called likes_retweets.csv and is included in this submission.

2. Assess

Each of the datasets was assessed to identify quality and tidiness issues. Assessments were performed using visual and programmatic assessments. Visual assessments included using methods like df.head() or slicing the dataframe based on specific column values. Programmatic methods included using .info() and .describe() to locate missing values, outliers, incorrect data types, incorrect values, etc.

The schema was defined as follows:

- Only observations about dogs were included in the analysis
- An observation was considered a dog if at least one of the three dog/not-dog predictions was True
- Only original tweets are valid (no reply-to or retweets)
- Missing data in the dog type (doggo, floofer, etc) is acceptable and that information is retained.

3. Clean

Based on the assessment of the data sets, the following activities were undertaken.

3.1. Incorrect data like, for example, incorrectly extracted dog names or mis-handled decimal ratings were corrected
3.2. Best guess dog breed were extracted from the AI data
3.3. Retweets and replies for WRD account were dropped.
3.4. Data type issues (tweet_id as string not integer, timestamp as datetime not string).

4. Analyze

Analysis addressing the questions listed above were undertaken using python for statistical analysis and Tableau for analysis and visualization.