WeRateDogs Data Wrangling Project

23 May 2019

Introduction

This project is part of the data wrangling section of the Udacity Data Analyst Nanodegree program and is primarily focused on wrangling data from the WeRateDogs, which is a Twitter account that rates people's dogs with a humorous comment about the dog. The rating denominator is usually 10, however, the numerators are usually greater than 10. This aspect was not cleaned as it is part of the humor and popularity of WeRateDogs.

Project Goal

Fully assessing and cleaning the entire dataset would require exceptional effort so only a subset of its issues (eight quality issues and two tidiness issues at minimum) needed to be assessed and cleaned.

The tasks for this project were:

- Data wrangling, which consisted of:
  - Gathering data
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing the wrangled data

- Reporting on my data analyses and visualizations (act_report.pdf)

Data Wrangling Process

Gather:

I gathered data from three sources:

- The Twitter archive #WeRateDogs, a csv file that contains the tweet, rating, dog name, and dog 'stage' in life (such as puppy). The file is provides by the Udacity Course and I use pd.read_csv() to import them into dataframe.
- An 'image prediction' file, or what breed of dog is in each tweet, according to a neural network. I downloaded the image prediction file programmatically from Udacity's servers using the Requests library and the provided url
- tweet_json.txt fie: Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called 'tweet_json.txt' file. Each tweet's JSON data is written to its own line.

Assess:

I assessed the data based on quality i.e. content issues like missing, duplicate, or incorrect data and tidiness i.e. structural issues

Tidiness Issues

- Dog Stages (i.e doggo, floofer, pupper & puppo) in “twitter-archive-enhanced.csv” should be one column
- The 3 provided files need to merge into one file. the twitter_archive_master.csv

Quality Issues

- Erroneous Datatype: tweet_id (in 3 files), timestamp (in twitter-archive-enhanced.csv)
- source column contains <> tag in “twitter-archive-enhanced.csv”
- Denominator have differnt values, not only 10 in “twitter-archive-enhanced.csv”
- One record have so big nominator's value (1776) and also some other big values in “twitter-archive-enhanced.csv”
- Some expanded_urls contain more than one URL and some have missing value in “twitter-archive-enhanced.csv”
- Invalid names in “twitter-archive-enhanced.csv”
- rate checks in “twitter-archive-enhanced.csv”
- There are retweets data in “twitter-archive-enhanced.csv”
- Not needed columns: in_reply_to_status_id , in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp in “twitter-archive-enhanced.csv”

- Some predictions have no dog image in “image_predictions.tsv”
- The dataset should only contain p, p_conf (take the highest prediction with dog spieces for each observation) in “image_predictions.tsv”

Clean:

Quality Issues

- Dropped unnecessary columns
- Converted erroneous data types
- Solved the NaN values in expanded_urls column from twitter_archive dataset
- Solved the Invalid dogs' names in twitter_archive dataset, and capitalized the first letter of dog name for consistence
- Changed the incorrect rating_numerator and rating_denominator values then created a new column rating = rating_numerator / rating_denominator dropped oberservations with extreme ratings (removing outliers.)
- Chose only a dog breed with the highest confidence each row. Tweets which are not predicted as a dog are set as missing value.
- Removed observations which are predicted as non-dog.
- Optimized the source content by ‘Twitter for iphone’, ‘Vine - Make a Scene’, ‘Twitter Web Client’, and ‘TweetDeck’.
- Created a ‘prediction’ column

Tidiness Issues

- Combined Dog stages columns (i.e doggo, floofer, pupper & puppo)
- Merged the 3 data sets into one data set

Finally, I stored the “twitter_archive_clean” to the file ‘twitter_archive_master.csv’.