16-06-2020

Report

On

Data Wrangling Steps: Gather, Assess, and Clean

By:

Abhishek Tiwari

Wrangle Report:

The dataset wrangle in the project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a twitter account that rates people's dogs with humorous comment about the dog.

The WeRateDogs Twitter project goals included:

- Wrangling the twitter data through the following processes:
  - Gathering Data
  - Assessing Data
  - Cleaning Data
- Storing, analyzing and visualizing your wrangled data
- Reporting on the data wrangling efforts and data analyse and visualization

Gathering Data:

My wrangling efforts for the WeRateDogs Twitter project included gathering data from the following sources:

- The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students ("like me"). This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.
- The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity students ("Like me").
- Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite ("like") count at minimum, and any additional data I find interesting.

Assessing Data:

Once the data was gathered, I began to assess the data on both quality and tidiness issues.

Quality Issue:

'twitter-archive-enhanced-2.csv':

- Completeness:
  - missing data in the following columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls.
  - tweet_id is an int (applies to all tables)
- Validity:
  - dog names: some dogs have 'None' as a name, or 'a', or 'an.'

- This data-set includes retweets, which means there is duplicated data (as a result, these columns will be empty: retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp).

- Accuracy:
  - retweeted_status_timestamp is also an object (the other retweeted statuses are floats)
  - Time-stamp is an object
- Consistency:
  - The Source column still has the HTML tags
  - rating_denominator should be a standard 10, but there are a multitude of other values

'image_predictions.tsv':

- Validity:
  - p1, p2 and p3 columns have invalid data...why would the algorithm labelled a dog photo as a starfish, boathouse, or mailbox.
- Consistency:
  - p1, p2 and p3 columns aren't consistent when it comes to capitalization: sometimes the dog breed listed is all lowercase, sometimes it is written in Sentence Case.
  - In p1, p2 and p3 columns there is an underscore for multi-word dog breeds.

'tweet_json':

- Completeness:
  - Missing Some Data

Tidiness Issue:

'twitter-archive-enhanced-2.csv':

- The last four columns all relate to the same variable (dogoo, floofer, pupper, puppo).

'image_predictions.tsv':

- This data set is part of the same observational unit as the data in the 'twitter-archive-enhanced-2.csv' - one table with all basic information about the dog ratings.

'tweet_json':

- This data set is also part of the same observational unit - one table with all basic information about the dog ratings.

Cleaning Data:

After the assessment, I cleaned the data through the following means:

Define, Code and Test:

- Merge the clean versions of archive, images, and twitter_counts_df data frames Correct the dog types.
- Create one column for the various dog types: doggo, floofer, pupper, puppo Remove columns no longer needed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.
- Delete retweets.
- Remove columns no longer needed.
- Change tweet_id from an integer to a string.
- Change the timestamp to correct datetime format.
- Correct naming issues and Standardize dog ratings.
- Creating a new dog_breed column using the image prediction data.