14/09/2022

REPORT

On

Data Wrangling Steps: Gather, Assess, and Clean

By:

Calvin Baraka

Wrangle Report:

The dataset to wrangle in the project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

The goals of this project included:

- Wrangling the twitter data through the following processes:
  - Gathering Data
  - Assessing Data
  - Cleaning Data
- Storing, analyzing and visualizing the wrangled data
- Reporting on the data wrangling efforts and analysis and visualizations.

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. This data consisted of tweets that spanned from 2015-2017
- 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. This was supposed to be obtained programmatically.
- 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.

  In this particular step I encountered various issues and had to refer with my instructor for continuous guidance.

Assessing Data:

I had to then asses the overall quality of the data gathered:

Quality Issue:

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

a) Validity:

i. dog names: some dogs have 'None' as a name, or 'a', or 'an.'

ii. This data-set includes retweets, which means there is duplicated data. As a result, multiple columns had to be dropped. These included the:

- retweeted_status_id
- retweeted_status_user_id
- retweeted_status_timestamp).

b) Completeness:

i. 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.

ii. tweet_id is an int (applies to all tables)

c) Accuracy:

i. retweeted_status_timestamp is also an object (the otherretweeted statuses are floats)

ii. Time-stamp is an object

d) Consistency:

i. The Source column still has the HTML tags so I separated the source from the html

ii. rating_denominator should be a standard 10, but there were instances of values that were >10

'image_predictions.tsv':

- **Validity:**
  - p1, p2 and p3 columns have invalid data and wrongful naming of the dogs e.g., naming a dog as a starfish.
- **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. For this I replaced the spaces and dashes with an underscore
    similarly, I also converted the uppercase to lowercase
  - In p1, p2 and p3 columns there is an underscore for multi-word dog breeds.

'tweet_json':

- **Completeness:**
  - This dataset was missing some data.

Tidiness Issue:

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

- The last four columns all relate to the same variable (dogoo, floofer, pupper, puppo). These had to be combined into one column and consequently eliminate the "none".

'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 dataset was basically one table with a basic information about the dog ratings.

Cleaning Data:

The data was afterwards cleaned using the following steps

Define, Code and Test:

i. Merge the clean versions of archive, images, and twitter_counts_df data frames Correct the dog types.
ii. Create one column for the various dog types: doggo, floofer, pupper, puppo
iii. Remove columns no longer needed
iv. Delete retweets to avoid duplicated data
v. Change tweet_id from an integer to a string.
vi. Change the timestamp to correct datetime format.
vii. Correct naming issues and Standardize dog ratings.
viii. Creating a new dog_breed column using the image prediction data.
ix.