Report

On

Data Wrangling Steps: Gather, Assess, and clean

By:

Ramya Ramachandra

Project Details

Our tasks in this project are as follows:

- Data wrangling, consists of:
  - Gathering data.
  - Assessing data
  - Cleaning data
- Storing , analyzing and visualizing our wrangled data
- Reporting

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 by Udacity Platform. This archive contains tweet data(tweet ID, timestamp,text etc).
- The tweet image predictions, i.e, what breed of dog is present in each tweet according to a neural network. This file was provided by Udacity Platform.

Assessing Data:

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

Quality Issues

Archive:

1. 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 table)

2. Validity:
    - Dog names: some dogs have ‘None’ as a name, or ‘a’, or ‘an’.
    - This dataset 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)

3. Accuracy:
    - timestamp is an object
    - retweeted_status_timestamp is also an object(the other retweeted status’s are float)

4. Consistency:
    - rating_denominator should be a standard 10, but there are a multitude of other values
    - the source column still has the HTML tags

Image_predictions:

1. Validity:
    - p1, p2 and p3 columns have invalid data..why would the algorithm labeled a dog photo as a starfish, boathouse, or mailbox.

2. 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

1. Completeness:
    - Missing some data

Tidiness Issue

archive:

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

images:

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

twitter_counts:

- 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 steps:

Define, Code and Test

1. Merge the clean versions of archive, images and twitter_counts_df dataframes correct the dog types.
2. 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.

3. Delete retweets
4. Remove columns no longer needed
5. Change tweet_id from an integer to a string
6. Change the timestamp to correct datetime format
7. Correct naming issues
8. Standardize dog ratings
9. Creating a new dog_breed column using the image prediction data.