Report on Data Wrangling: Gather, Assess and Clean

By:

Michael A. Olojo

Wrangle Report:

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

The WeRateDogs Twitter project includes wrangling of data through the following processes:

- Gathering Data
- Assessing Data
- Cleaning Data

In addition, I have to store, analyze and visualize the wrangled data.

Also, I have to report the data wrangling efforts, data analysis and visualization.

Gathering Data:

My wrangling effort for the WeRateDogs project involves gathering data from the following sources:

- The WeRateDogs Twitter Archive. The twitter_archive_enhanced.csv file was provided to students by Udacity, which contains tweet data e.g tweet ID, timestamp, text etc for all 5000+ of their tweet.

- The tweet image prediction. The image prediction file contains the breed of dogs that is present in each tweet according to a neural network. This was also provided by Udacity to the students.

- Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite count at minimum. Alternatively, the data was provided on the online content by Udacity, if students are not satisfied or comfortable with the first option.

Assessing Data:

Once I have gathered the data, then I begin to assess the data based on quality and tidiness issues.

Quality Issue:

'Twitter_archive_enhanced.csv'

1. id column name should be "tweet_id" instead of "id"
2. tweet_id should be "str" and not "int"

3. Some values in rating denominator column is not "10"

4. Some values in rating numerator column is less than "10"

5. Some values in rating numerator is equal to zero.

6. retweeted_status_id should be removed since we are interested in the tweet.

7. retweeted_status_user_id should be removed since we are interested in the tweet.

8. retweeted_status_timestamp should be removed since we are interested in the tweet.

9. Nulls represented as "none" in the name column.

10. In the columns "conf" should be confident

Tidiness:

1. doggo, floofer, pupper, and puppo should be in 1 column not 4 columns
2. combining the three data frames in one data frame

Cleaning Data:

After the data have been assessed, I cleaned the data using three steps which are: Define, Code and Test.

1. id column name should be "tweet_id" instead of "id"

2. tweet_id should be "str" and not "int"

3. Some values in rating denominator column is not "10"

4. Some values in rating numerator column is less than "10"

5. Some values in rating numerator is equal to zero.

6. retweeted_status_id should be removed since we are interested in the tweet.

7. retweeted_status_user_id should be removed since we are interested in the tweet.

8. retweeted_status_timestamp should be removed since we are interested in the tweet.

9. Nulls represented as "none" in the name column.

10. In the columns "conf" should be confident