PROJECT REPORT ON WRANGLING AND ANALYZING DATA

By Onyedikachi B. Ogbonna

Introduction

In the world of data science, data analysis (or any field that has to do with data), getting the data is one part, but Real-world data rarely comes clean (maybe due to mising values, duplicate rows, etc), and hence the hard part is mostly wrangling, cleaning or preparing the data before analyzing or feeding it to a model for training. In this project, I will be wrangling the data (Tweets) gotten from twitter (WeRateDogs) and then do some analysis.

I used the following python packed for my project

- pandas
- NumPy
- matplotlib
- json

Gathering data

Three datasets were used for this project namely

- twitter-archive-enhanced.csv: contains tweets of dog ratings
- image-predictions.tsv: contains predictions for dog breeds
- tweet-json.txt: contains retweets and favorites

Assessing data

Theses datasets were imported and assessed, but I found some tidiness and quality problems:

Quality issues includes:

1. twitter_archive dataset has some columns with missing values
    - in_reply_to_status_id,
    - in_reply_to_user_id,
    - retweeted_status_id
    - retweeted_status_user_id
    - retweeted_status_timestamp
    - expanded_urls

2. tweet_id should be string, not int
3. timestamp should be in date-time format
4. There were many missing dog names as they weren't given real or correct names (eg None, a, etc).
5. They rating numerator had some issues, some values were more than 10, which was wrong (eg, 11, 15).
6. Some retweets were just same as main tweets.
7. There were some columns that weren't needed all to together
8. The source column contains some html structure that are not needed, it should be just plain text.

Tidiness issues includes:

1. The datasets have so many redundant information and hence should be merged.
2. The dog stage should have been values, not columns and also some dogs had multiple dog stages.

Cleaning data

This stage I made use of the define-code-test framework. But first I made a copy of the datasets before working on them as follows:

- Join three datasets using the unique feature/identifier, in this case, it's the tweet_id.
- Create one column for the various dog stages: doggo, floofer, pupper and puppo.
- Correct/change naming issue
- Standardize dog ratings (change dog rating)
- Change tweet_id from an integer to a string
- Change the timestamp to datetime format
- Remove HTML tags from source
- Drop null or missing values from expanded_urls
- Remove retweets records from the dataset
- Remove columns no longer needed

Storing data

After joining all three dataset and doing the necessary cleanup, I saved the new dataset as twitter_archive_master.csv for any future use

Analysis

And then I did some analysis on some features to get some insights, eg the most popular dog stage, the most given dog name, etc.

Conclusion

The data wrangling process is much needed in any data science or data analysis project as it gives one a clean and reliable data to work with so as to get an accurate result in the end.