Wrangle Report

#WeRateDogs

The goal for this project is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. I will be using Python and its libraries to gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it.

The dataset that I will be wrangling is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

Gathering Data

Data from 3 sources:

1. **Enhanced Twitter Archive:** contains basic tweet data for all 5000+ of their tweets.
2. **Additional Data via the Twitter API:** contains retweet count and favorite count which are two of the notable column omissions from the Twitter Archive.
3. **Image Predictions File:** contains a table full of image predictions alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

Assessing Data

This step allows us to identify quality and tidiness issues. Assess data for:

- Quality: issues with content. Low quality data is also known as dirty data such as missing, invalid, inaccurate and inconsistent data.
- Tidiness: issues with structure that prevent easy analysis. Untidy data is also known as messy data.

Two types of assessment are done to identify the issues:

- Visual assessment: scrolling through the data
- Programmatic assessment: using code to view specific portions and summaries of the data

The following is the list of issues that have been observed and will proceed to clean some of them in the next section:

Quality Issues

archive_df

- variables need to convert to the right datatype (Tweet_id, timestamp, source, doggo, floofer, pupper and puppo)
- source is in HTML format with a and \a tags surrounding the text
- There are missing name and some of the name fields are prepositions (e.g. 'a', 'actually', 'all', etc).
- column headers not descriptive
- Data contains retweets (ie. rows where retweeted_status_id and retweeted_status_user_id have a number instead of NaN)
- Numerous not required columns to be deleted

predictive_df

- variable need to convert to the right datatype (img_num)
- numerous column on image prediction result

api_df

- p1, p2, and p3 contain underscores instead of spaces in the labels

Tidiness Issues

- To combine 3 dataframes into one using tweet_id with a (inner) join condition.
- 1 variable (dog stage) in 4 different columns (doggo, floofer, pupper, and puppo)

Cleaning Data

The data have been cleaned using the programmatic methods. They are shown under the define, code and test format.

- Define: definition or issue to clean of fix.
- Code: the code use to clean or fix the issue.
- Test: to assure that the cleaning operations worked correctly.

Conclusion

Through the data wrangling process, we managed to fix a number of problems and cleaned the data for further analysis to be done.