Wrangle Report

Our goal is to use Twitter's WeRateDogs data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. It takes extra gathering, then assessing and cleaned-worthy analyses and visualizations.

Gathering data

We gathered data from 3 sources:

Twitter Archive, which contains basic tweet data for all 5000+ of their

tweets .

Additional Data via the Twitter API, which contains retweet count and favorite

count are two of the notable column omissions. Fortunately, this additional data can

be gathered by anyone from Twitter's API .

Image Predictions File, which is a table full of image predictions (the top three only)

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).

Assess data

The data is assessment in two stages:

Visual assessment: scrolling through the data in your preferred software application (Google Sheets, Excel, a text editor, etc.).

Programmatic assessment: using code to view specific portions and summaries of the data (pandas' head, tail, and info methods, for example).

Assess data:

•**Quality**: issues with content. Low quality data is also known as dirty data.

•**Tidiness**: issues with structure that prevent easy analysis. Untidy data is also known as messy data.

Here’s a list of the issues that have been highlighted:

Quality Issues:

- tweet_id should be object(string)
- timestamp should be datatime
- Sources to be clearly defined such as Twitter for iPhone..etc
- Same rating_denominator are higher or less than 10
- You only want original ratings no retweets
- Dog name is not always accurate: a ,an
- Change the name of the 'p' like "p1,p2,p3,p1_conf,p1_dog, etc..) to (prediction1, prediction2, etc..)
- Dog Stages replaced by 'None'

Tidiness Issues:

- Combine the datasets together.
- Joining the dog stages into a single column instead of four

Clean data

The data was cleaned up with programmatic process:

We defined convert our assessments into specific cleaning tasks. Then convert our definitions into code and run that code. Test our dataset, visually or code, to make sure cleanups are successful

Conclusion

Through the data wrangling and analysis, we used many libraries such as pandas, NumPy,requests, tweepy, and json, which allow us to gather, assess, and clean the data.