Wrangle Report

Data Wrangling

The dataset used for this wrangling exercise 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 funny thing is that the numerator is almost always greater than 10.

Gather

There are 3 sources of data required for this wrangling exercise. The first is the twitter_archive_enhanced.csv file. The second is the image_predictions.tsv file, and the last file needs to be created using the Tweepy library through the Twitter API.

Twitter archive CSV file

I started the project by downloading the WeRateDogs Twitter archive file manually using the following link to my hard drive : https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv and then imported the file into the following dataframe: twitter_archive = pd.read_csv ('C:\\Users\\Teresa\\twitter-archive-enhanced.csv')

Tweet image predictions

The tweet image predictions file, which is hosted on Udacity's servers was downloaded programmatically using the Requests library at the following URL:

https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

Twitter API

Now it was time to get each tweet's retweet count and favorite ("like") count. Using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file.

Assess

When accessing the data, I used both a visual and a programmatic assessment to detect Quality and Tidiness issues. The issues are as follows:

Quality Issues

1. Dataset contains retweets

2. Some of the tweets do not have images

3. Some of the rating denominators are < 10

4. Some of the rating denominators are < 10

5. Some of the dog names are incorrect

6. Column headers "p1", "p2", "p3", "p1_conf", "p2_conf", "p3_conf", "p1_dog", "p2_dog", and "p3_dog"" are not intuitive to what it contains and should be changed

7. Replace instances of "& amp" with just "&"

8. timestamp, retweeted_status_timestamp have incorrect datatype. They are showing as datatype string

Tidiness Issues

1. The column headers "doggo", "floofer", "pupper", and "puppo" should be merged into one column

2. The three separate datasets should be merged into one dataset

Clean

After assessing the data, I moved to cleaning the dataset. Before proceeding I made copies of the original pieces of data. I used both types of cleaning but mostly used the programmatic data cleaning process. I followed the format structure of Define-Code-Test

Store

To wrap up the project I stored the clean data frame to a CSV file named twitter_archive_master.csv.