Wrangle Report

Kexin Yao

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. From the dataset of the WeRateDogs we can get the interesting and trustworthy analyses and visualizations, but before the analysis, we need to wrangle the dataset to make it clean.

The data wrangle process contains the gathering data, assessing data and cleaning data.

Gathering data:

The dataset I'm wrangling is the tweet archive of Twitter user WeRateDogs which contained three dataset:

1.twitter_archive_enhanced.csv(Download the file manually)

2.image_predictions.tsv (Downloas programmatically using the Requests library and the URL)

3.tweet_json.txt. (Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data)

Assessing data

Using pandas functions like head(),info(),value_counts(),duplicated(),isnull(),etc…to check the three dataframe and recorded their quality issues and tidiness issues as following:

Twitter_archive_enhanced Table

1.Tweet_id is a int not a string.
2.There are too many null values in columns like 'in_reply_to_status_id' and 'in_reply_to_user_id'.
3.Erroneous datatypes(timestamp).
4.Retweeted twitter columns do not provide information for the analysis.
5.Rating_denominator not 10 sometimes.
6.Name has missing data and has names like 'a', 'an', 'the', etc. All names with problems initialed with lowercase letter.

Twitter_image_prediction Table

7.jpg_url has duplications
8.Tweet_id is a int not a string

Tweet_Count Table

no quality issues

Tidiness:

1.One variables in four columns in enhanced table(dog_stage).

2.All three tables could be merged on 'tweet_id' for analysis, as all variables belong to one observational unit - dog rating.

Cleaning data:

Since we only want original dog ratings, the first step of the cleaning data is to remove the retweets as a user can retweet their own tweet.

After remove the retweets, the 8 quality issues and 2 tidiness issues were cleaned by using pandas and pythons.

Finally, three cleaned dataset were merged to one dataframe and stored as a csv file”twitter_archive_master.csv”.