WeRateDogs

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The account was started in 2015 by college student Matt Nelson, and has received international media attention. Now the account has about 9 million followers.

I have wrangled three datasets from this account as a project in my Udacity's Data Analyst nanodegree. The wrangling process consisted of gathering, assessing and cleaning of three different datasets.

1. Data Gathering

The first step I took was data gathering. I have gathered three different datasets via different gathering techniques;

- I have directly downloaded the first dataset – *twitter-archive-enhanced.csv*
- For the second dataset, I have used request library in order to retrieve data from a given link
- The last dataset was collected by using twitter's APIs. But, in order to get this dataset, I set up twitter developer account.

2. Data Assessing

After gathering, the three datasets may contain a lot of content and structure issues that needs to be assessed first. So, I have both visually and programmatically assessed the three datasets. I found nine content issues and two structure issues from the first dataset (twitter-archive-enhanced.csv). The two other datasets were almost clean, so I haven't found any issues with the other two datasets. After detecting the issues, I have documented one by one for the purpose of cleaning easily. These are the issues and found all of these issues from the first dataset.

1. Missing values in these columns.
   - in_reply_to_status_id
   - in_reply_to_user_id
   - retweeted_status_user_id
   - retweeted_status_timestamp
2. Timestamp has unnecessary +0000
3. The timestamp has a datatype of object instead of datetime
4. Source columns has html tags
5. Duplicates and Missing values in expanded_urls
6. Rating_numerator is wrong in some of the tweets
7. Wrong denominator, most denominators are supposed to be 10, but there are some denominators greater than 10.
8. Dogs having an invalid name such as a, an, and the

9. Names starting with lower case letters

Before cleaning the nine above issues, I have extracted the original tweets, and removed the retweets from the first dataset.

3. Data Cleaning

After assessment, I have started to clean all these documented issues. The cleaning process starts defining the issue, cleaning it and testing if it cleaned or not. So, I have cleaned all the nine content issues and the two structure issues of the dataset.

I have dropped all the columns that have large content of null values such as in_reply_to_status_id, in_reply_to_user_id, retweeted_status_user_id, retweeted_status_timestamp. Also, I have cleaned the timestamp column where I have removed the unnecessary +0000 and its data type was changed into datetime. The rest of the issues was also cleaned.