Project: Investigating WeRateDogs Twitter Data

Data Wrangling Project Report

Data Gathering

Three datasets were collected :

- Enhanced Twitter Archive(Table 1),
- Image Predictions File (Table 2)
- Additional Data via the Twitter API (Table 3)

The first table was downloaded and read into a pandas Dataframe. The second was downloaded programmatically using the Requests library where it is hosted by Udacity server. The third dataset was retrieved as a json file having been queried from the tweeter API.**[N:B, due to delays from Tweeter, the alternative option provided was used, downloading the .txt file directly.]**

Assessing Data

The datasets were assessed but manually and programmatically. Some quality and tidiness issues were identified. These quality and tidiness issues are discussed thus:

Quality Issues:

Table_1

- Duplicated tweet_id
- Wrong datatypes (tweet_id,in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,retweeted_status_user_id)
- Some errors in numerator ratings and some denominators are not decimally rated
- Missing data; NaN in (in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp,expanded_urls)

Table_2

- tweet_id is not string
- Some lowercase prediction names (p1,p2,p3)
- Non descriptive column headers

Table_3

- tweet_id is not string

Tidiness Issues:

- Dog-type variable in 4 four different columns in Table 1
- Table 3 which contains the retweet count and favorite count should be part of Table

Cleaning Data

The cleaning process began by dropping the duplicated tweet_id in Table 1.

The Nulls in Table 1[in_reply_to_status_id , in_reply_to_user_id colums, retweeted_status_user_id, retweeted_status_user_id and retweeted_status_timestamp] were filled with "not available" while the null values in expanded_urls were filled with "missing".

The tidiness issue of 4 dog-type variable in four different columns was cleaned using the melt function. Which collapsed all four(*doggo, floofer, pupper ,puppo)* columns as 'growth_stage' and check for duplicated tweet_id as a result of multiple dog stage entry in rows and drop them. Then create dataframe for those duplicate and concatenate with parent Dataframe. And where all 4 growth_stage is 'None', a new Dataframe made, which is then joined together as one. The New Dataframe has just a column for the dog stage with some of them having double dog stage.

Table 1 and Table 3 were merged as 1 Dataframe for the twitter information while Table 2 as the prediction dataset.

Outrageous ratings were dropped from Table 1. Tweet IDs were converted to string in all tables. The inconsistency in Case were fixed as the lower cases were converted to Titlecase.

Non- descriptive column headers were renamed for descriptive purposes in Table 2.

Although the dataset was merged as “twitter_archive_master”, for analysis, the dataset are grouped into 2, the first is the tweeter information, while the second is the image prediction data.