DATAWRANGLE OF THE WERATEDOGS'S TWEETS

Udacity - Nanodegree Data Analyst

Leonardo Simões

07/15/2020

Summary

1. Data Wrangling

The data Wrangling process is generally divided into the steps of gathering, assessing and cleaning data, this work also occurs, but the assessing and clean stages will be divided into two each, one for organization problems (tidiness) and one for quality of the data. At the end of the cleaning, the data was saved in a twitter_archive_master.csv file.

2. Gather

The data is collected from three different files, each with a different extension (.csv, .tsv, .txt (JSON)). The file 'twitter-archive-enhanced.csv' is normally read with the Pandas read_csv method on a dataframe called df. The content of the 'image-predictions.tsv' file was downloaded programmatically using the requests library, saved to a file and then opened also with the Pandas read_csv method on a dataframe called df2. The 'tweet-json.txt' file was opened as a text file, processing each line to separate the multiples of the tweets and then its content was read with the Pandas read_csv method on a dataframe called df3. Thus, each file was opened on a different dataframe.

3. Assessing of Data Organization

This step consisted of checking the structures of the data frames: number of rows, number of columns, columns in common, and whether the 'stage' feature was a column. The ‘tweet_id’ column was present in two of the datraframes, but in the other it was just named ‘id’. There was no column for the ‘stage’ feature, but its possible values were columns. df3 had many columns in addition to the three specified as necessary.

4. Clean of Data Organization

Copies of the dataframes were made for cleaning. The correction of each storage problem is divided into define, code and test. The df3_copy columns are discarded, leaving only 3 of them. The 'id' column of df3_copy is renamed to 'tweet_id'. A new df_clean daframe is generated by joining df1_copy, df2_copy and df3_copy using the 'tweet_id' column, and with that all operations are done in df_clean, and df1, df2 and df3 and copies are discarded. A new 'stage' column is created with the corresponding 'none', 'doggo', 'puppo', 'pupper', 'floofer' values. Columns 'doggo', 'puppo', 'pupper', 'floofer' are excluded.

5. Assessing of Data Quality

The info method is used to superficially view general dataframe information, such as name, data type and filled column values.

```
df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2356 entries, 0 to 2355
Data columns (total 27 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   tweet_id                    2356 non-null   int64  
 1   in_reply_to_status_id       78 non-null     float64
 2   in_reply_to_user_id         78 non-null     float64
 3   timestamp                   2356 non-null   object 
 4   source                      2356 non-null   object 
 5   text                        2356 non-null   object 
 6   retweeted_status_id         181 non-null    float64
 7   retweeted_status_user_id    181 non-null    float64
 8   retweeted_status_timestamp  181 non-null    object 
 9   expanded_urls               2297 non-null   object 
 10  rating_numerator            2356 non-null   int64  
 11  rating_denominator          2356 non-null   int64  
 12  name                        2356 non-null   object 
 13  jpg_url                     2075 non-null   object 
 14  img_num                     2075 non-null   float64
 15  p1                          2075 non-null   object 
 16  p1_conf                     2075 non-null   float64
 17  p1_dog                      2075 non-null   object 
 18  p2                          2075 non-null   object 
 19  p2_conf                     2075 non-null   float64
 20  p2_dog                      2075 non-null   object 
 21  p3                          2075 non-null   object 
 22  p3_conf                     2075 non-null   float64
 23  p3_dog                      2075 non-null   object 
 24  retweet_count               2354 non-null   float64
 25  favorite_count              2354 non-null   float64
 26  stage                       2356 non-null   object 
dtypes: float64(10), int64(3), object(14)
memory usage: 515.4+ KB
```

Figure 1 – df_clean.info()

Then, it was verified which columns have missing values, and to facilitate the visualization of this information, a horizontal bar graph was plotted.

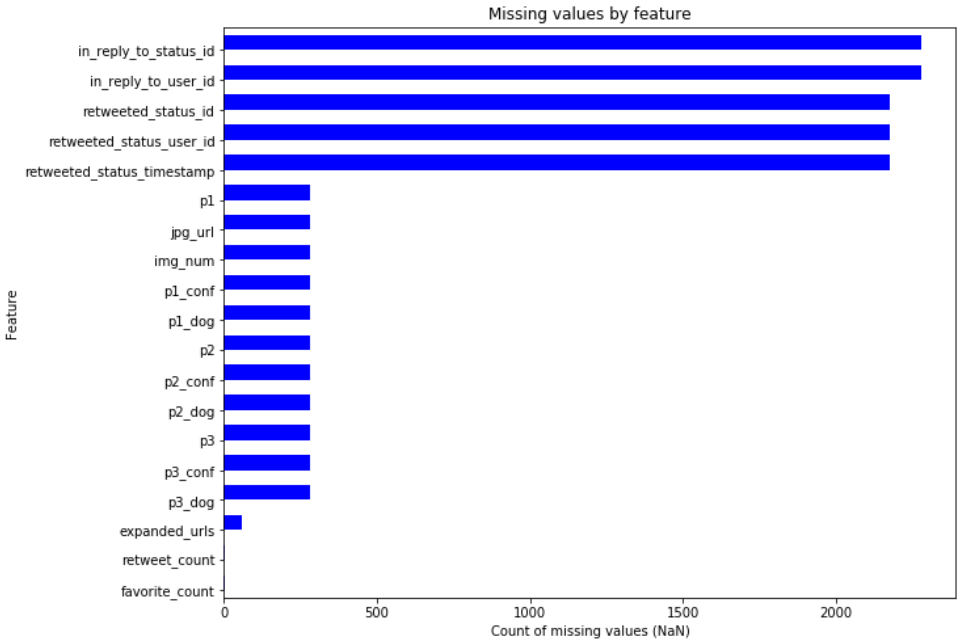

Figure 2 – Missing values by features

With this information and individual analysis of some columns, some data quality problems were identified. No duplicate lines were found.

The 'rating_denominator' column has a value of 0, which can be a problem in calculating a possible 'rating' feature. The 'source' column has only 4 values, but these should only be the values between the anchor tags.

It was observed that the columns 'p1', 'p1_conf', 'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog' had NaN values simultaneously in 281 lines. The 'retweet_count', 'favorite_count' columns had NaN values in just 2 rows.

The data type of the columns 'retweet_count', 'favorite_count' and 'img_num' were of type float64, but, as suggested by the prefixes 'count' and 'num' should be int64.

Only lines that do not represent a retweet should be kept. If confirmed that the tweets are not retweets, the columns for retweets are unnecessary.

6. Clean of Data Quality

The value 0 in 'rating_denominator' is replaced by 1. The values in 'source' become just the values between the anchors tags.

Missing values in 'retweet_count' and 'favorite_count' are filled in with 0. Lines that have missing values for 'p1', 'p1_conf', 'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3' , 'p3_conf', 'p3_dog' are dropped.

The data type of the 'img_num', 'favorite_count', 'retweet_count' columns becomes int64.

The lines where 'Retweeted_status_id' is not NaN, represent retweets, have been dropped. After confirming that tweets are not retweets, 'retweeted_status_id', 'retweeted_status_user_id', and 'retweeted_status_timestamp' columns are dropped.

REFERENCES

UDACITY - Data Analyst Nanodegree Program: https://www.udacity.com/course/data-analyst-nanodegree--nd002

WeRateDogs, Twitter profile (@dog_rates): https://twitter.com/dog_rates/status/749981277374128128