Data Wrangling Report

By Mai Ezzat Fraghaly Abdellatif

December 2020

This report illustrates the main steps for the wrangling process that has been applied for the "WeRateDogs" twitter account

Data Gathering

In this step data were gathered from 3 main resources to work on later

1. The twitter Archive file that was provided by Udacity and downloaded manually. The file was in csv format and then was uploaded the Jupyter notebook to be read with the read_csv method

2. The image predictions file which was downloaded programmatically using the requests library and the given url. The file was in tsv format and it was uploaded to the Jupyter notebook in the same directory of the twitter archive file. Then it was red by the read_csv method with assigning the separator parameter to tab.

3. The third dataset was gathered from twitter using the API. To perform this process, a developer twitter account has been created to query the additional data. Then query the additional information by using the API object and the tweepy library. This information was about the favorit count of tweet and the retweet for it.

Data Assesment

In this step date were assessed visually and programmatically

1. The visual assessment was done by spreadsheets using Excel

2. The programmatic assessment was done on Jupyter notebook using some useful methods and functions of pandas such as .info(), .describe(), .value_counts(), .sample(), .shape, .head()

3. Quality issues and tidiness issues has been recorded to clean later

Data Cleaning

In this step the cleaning strategy was applied by defining, coding then testing for following issues

1. Quality

Archive table

- Timestamp column is string not datetime
- Missing values in 5 columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)
- lowercase names in the name column which are not actually names
- In the name column there is "O" and it should be "O'Malley"
- The column header "expanded_urls" is not descriptive
- The url and the rating exists with the text column
- In the timestamp there is "+0000"
- Missing values in the expanded_urls column
- The column header "name"
- Timestamp column should be 2 columns one for date and the other for timing
- Extra column should be done for the rating (numerator / denominator)
- Columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) should be removed
- The 3 tables should be merged on the "tweet_id"
- The source column
- The rating in the text
- the data type in rating_numerator and rating_denominator is integer

Images table

- The columns header are not descriptive (p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog)
- The names given in p1, p2, p3 is a mixture of lowercase and uppercase
- Some names is separated by "-" or " _" or space
- The img_num column

Favorit count & retweet count table

- The header "id" should be "tweet_id"

2. Tidiness

Archive table

- columns (doggo, floofer, pupper, puppo) should be one column
- Merge the 3 tables