Data Wrangling Report

By Aya Ghorabah

December 2020

This report is an illustration for data wrangling's of Twitter account "WeRateDogs".

A) Data Gathering:

This is data collection from different sources. In this project, I used three data files.

1- The first file was given with the project assignment's files. Its format was Comma Separated Value (CSV), I used pandas read_csv function to open and read it as a data frame "twitter_archive".

2- The second file is downloaded from the given link using the requests library and the get method. Then saved to a CSV file. Finally, I use the same method used with the first file to open it as a dataframe "img_predict".

3- The third file is a downloaded text file using the Twitter API library "tweepy". Then used the JSON library to convert it to a JSON file. Finally, convert it to dataframe using pandas library and DataFrame method. This dataframe is "twitter_api_archive".

B) Data Assessing:

This Assessing the data for quality and tidiness issues. It contains two main steps:

1- Visual assessment using Microsoft Excel.

2- Programmatic assessment using Jupyter notebook and different python libraries.

The results are:

1- twitter_archive dataframe:

a) Quality issues

- Some columns have inappropriate names (time_stamp, source,…etc.).
- Timestamp column is an object.
- Name column some values <= 2 letters.
- "rating_denominator" some values! = 10.
- "rating_numerator" some values > 15 or <=5.

b) Tidiness issues

- It contains retweet and replies columns.
- Four columns for the dog stage.

2- img_predict dataframe:

a) Quality issues

- Some columns have inappropriate names (p1_conf, p1, ..etc.).

- False values in "p1_dog".

b) Tidiness issues

- “img_predict” and "twitter_archive" dataframe as both contain data for the same object.
- There are three columns for prediction algorithms and the same for results and is it a dog.

3- twitter_api_archive dataframe:

b) Tidiness Assessment

- “twitter_api_archive” and "twitter_archive" dataframe as both contain data for the same object.

C) Data Cleaning:

This is a three stages process:

a) Define: specify the problem.
b) Code: fix the problem.
c) Test: is the problem fixed?

First, create a copy for the dataframes to apply the changes to. They are twitter_archive_new, img_predict_new, and twitter_api_new.

Quality / Tidiness	Problem	Solution
Quality	“time_stamp” type is an object.	Converted to date-time format.
	“dog_stage” column contains values with combined stages	Find the rows with the problem. Then split the two stages using”- “
	Some column names not appropriate ('timestamp', 'source', 'p1', 'p1_conf', 'p1_dog', …. etc.)	Replace with appropriate ones ('tweet_timestamp', 'tweet_text', 'predicted_type', …. etc.)
	"rating_denominator”! = 10	Investigate tweet text and fix or drop
	"rating_numerator" <= 5 or > 15	Investigate tweet text and fix or drop
	"dog_name” <= 2	Investigate tweet text and fix or drop
	Dog_name is No_name or None	Investigate tweet text and fix or drop

Tidiness	Four columns contain dog stage data.	Remove the “None” values. Then combine the four columns in one column “dog_stage” and then drop them.
	Three columns contain predicted type and the same for accuracy percentage and if it is a dog	As the prediction "p1" is the most likely one. Drop the other predictions columns and depend only on "p1" results
	"img_predict_new" and "twitter_archive_new" have data for the same object	Combine both dataframes to “tweets_df”
	“tweets_df” contains retweet and tweet's replies columns	drop rows with reply and retweet is not NaN. Then drop columns of retweets and replies
	Some pics are not dogs	Drop rows with "is_dog" is False. Then drop the column as it has no indication.
	Some rows do not have pics	Drop rows with "jpg_url" is NaN
	"tweets_df" and "twitter_api_archive" are about the same object	Combine both dataframes to twitter_df