DATA WRANGLING REPORT

By Thu Phung Ngoc Minh

May 2020

I. Introduction

The report describes the wrangling process to complete the project WeRateDogs from Udacity

The Data Wrangling process included:

1. Gathering data
2. Accessing data
3. Cleaning data

II. Implementation

1. Gathering data:

Sources of data included:

- A CSV file "twitter-archieve-enhanced" : downloaded directly from Udacity server
- A website link to download file "prediction" directly and read file directly by using Pandas
- A Twitter API: Created account on Twitter and register application development account. Get Keys and Token passwords, by using package Tweepy to query the Twitter API.

2. Accessing data

- Explore each dataframe by using head(), shape(), info(), duplicated() and value_counts()
- Identify quality issues and tidiness issues:
    - Quality Issues:
        - Retweets and replies columns should be deleted.
        - Expand URL has 281 entries without URLs, these entries also should not take into account.
        - Timestamp columns should not be in object(str) type.
        - Ratings denominators is over than 10. They may be from wrong typing
        - Numerators have some incorrect values.
        - Some dog's names are incorrect.
        - Dog' names should be in same format with first letter capital.
        - P1, P2, and P3 in prediction dataframe need to be fixed in same forma
    - Tidiness Issues:
        - 3 dataframs can be merged together

- Dog's stages are split into 4 columnns. They should be merge in 1 colum

3. <u>Cleaning data</u>

❖ With Tidiness Issues:

1. Part 1: Merge 3 dataframes into 1 dataframe “Twitter_archieve_master”
- By using merge() function to join firstly 2 dataframes archieved.csv and tweet.csv by tweet_id
- Then join continue with prediction.csv, also by tweet_id
2. Part 2: merge 4 types of dog into 1 column “dog_type”
- Using function melt() to identify column headers in dataframe and add 1 column dog_type
- Drop 1 column is called variable

❖ With Quality Issues

1. Drop expanded URLs with Null values
   - Using function dropna() to drop all Null values in this column
   - Check again if Null values are dropped
2. Change sources name into easier name for reading and visualization
   - Using function value_counts() to see how many categories of source type
   - Using function replace() to change name
3. Change data type of timestamp
   - Using function pd.to_datetime() to change type of column “timestamp” from string to datetime64[ns]
4. Drop columns relevant to retweets and replies from 1st dataframe “twitter-archive-enhance”
   - Create the list of these columns want to drop ('retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp','in_reply_to_status_id', 'in_reply_to_user_id')
   - Drop all list out of dataframe
5. Check incorrect values in rating_denominator and rating_numerator
   - As denominator rating score normally not over 10, so we will using function value_counts() to see how many values have rating_denominator over 10
   - Then we locate these tweet_ids with rating_denominator over 10
   - Explore the text column, to see what they noticed or gave information.
   - Then, I recognized 8 rows have information right about the raiting of denominator
   - Besides, the score of raiting numerator also were mentioned in their text, so it is adjusted also.
6. Check dog name values and change format
   - Using function unique() to list all names of dogs
   - Then, I merged these weird names such as ‘a','all','an'… into 1 variables called None.

7. Change format of P1, P2, P3

- Using function str.capitalize() to convert these 3 columns with values always start with capital letters.

- Export the completed final file into CSV file under name “master_archieve.csv” for insights analysis and visualization