Wrangle Report with WeRateDog Data

July 26, 2019

1 Wrangle Report with WeRateDog Data

1.1 Introduction of Data Set :

There are three data sets in total for this project : twitter-archive-enhanced.csv,image-predictions.tsv , tweet-json.txt

1.2 The Goal :

The goal of the data wrangle process is to generate a clean data set called twitter_archive_master.csv for data visualization and analysis later.

1.3 Data Issues: I found 12 quality issues and 4 tidiness issues

** Tidiness Issues (Content Issues) ** 1) numerator_rating and denominator should me merge in one coloumn instead of two column.

2) Split date and time in two seperate column instead of one column.

3) There is a incomplete data in some file so i perform merge operation to combine all the file for analysis.

4) There is inappropriate column name for json file.

** Quality Issues **

5) In name column there are multiple expression which are wrongly name convert these wrongly name to none.

6) Remove '+0000' from date column of df_clean file.

7) change 'tweet_id','in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id' from float to integer

8) Drop the rows when the values of retweet_ids are not none.

9) There are missing values in 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'. The columns can be dropped since they are not relevant for analysis purpose.

10) Create dog stage variable and remove individual dog stage columns.

11) Display full content of 'text' column.

12) Change missing values in 'name' from 'None' to NaN

13) Remove tweet without rating

14) Remove extra character after '&' in text column of text file.

15) Remove "-" in P1, P2, P3 Using replace function

16) Change all string in p1, p2, p3 in lower case.

Clean Data: Tidiness issues first , then quality issues

1) numerator_rating and denominator should me merge in one coloumn instead of two column.

** I simply add the numerator and denominator rating to one column called rating by using astype function and passing str as parameter.

2) Split date and time in two seperate column instead of one column.

** I use str.split function on timestamp column to split in two different column called date and time column.

3) There is a incomplete data in some file so i perform merge operation to combine all the file for analysis.

** I simply performe merge operation by calling merge function on the given three file and perform join on 'twitter_id'

4) There is inappropriate column name for json file.

** I use column renaming method and rename column twitter_id in place of id.

5) In name column there are multiple expression which are wrongly name, convert these wrongly name to none

** I use replace function to convert wrong name to none. First I make list of wrong name after that i perform replace opertaion on that.

6) Remove '+0000' from date column of df_clean file.

** I use rstrip function to remove '+0000' part from whole text.

7) change 'tweet_id','in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id' from float to integer

** I use astype function on the above column which have inappropriate data types.

8) Drop the rows when the values of retweet_ids are not none.

** By using drop function i drop retweet_ids whose value is equal to none.

9) There are missing values in 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'. The columns can be dropped since they are not relevant for analysis purpose.

** The above column are not useful for futher analysis because it have plenty of missing value. So i drop those column by using drop function.

10) Create dog stage variable and remove individual dog stage columns.

** I extract (puppo | pupper | floofer | doggo) in one column by using extract method. Because few variable is present in two many column.

11) Display full content of 'text' column.

** I call set_option function to display the full text by setting the colwidth to infinite. I did this because i was not able to read the whole tweet. So by doing this whole text is readable.

12) Change missing values in 'name' from 'None' to NaN

** Again using replace method replace None to NaN.

13) Remove tweet without rating

** I remove the tweet without rating by using not equal to operator this operator mtches the tweet without rating.

14) Remove extra character after '&' in text column of text file.

** By using replace I replace '&amp' with '&'

15) Remove "-" in P1, P2, P3 Using replace function

** By using replace function. I replace "-" with blank space.

16) Change all string in p1, p2, p3 in lower case.

** By using lower function i changes all the text in p1,p2,p3 column in to the lower text.

In [ ]: