Hany Elshafey

Udacity

Data Analysis Professional

Nanodegree Program

16/3/2021

Data Wrangling Report

First of all, how proud I am to join you on this inspiring program. I thank everyone at Udacity, and I thank the Egyptian government as well. Based on the project submission system, i will show the action steps involved in the data wrangling of Twitter account " WeRateDogs ".

Prepare for the project

In this step, I downloaded the files available to work on, and also prepared the working environment using Anaconda, and then installed important auxiliary libraries (pandas ,numpy,mat -plotlip).

Data Gathering

And here begins collecting data from the three main different sources:

1- Twitter_archive_enhanced.csv file. I got it by downloading manually then I imported it to my work space using pandas.read_csv

2- Image_predictions.tsv file. To work on this file, I downloaded it programmatically using its link as it was hosted in Udacity's servers ,I downloaded it using requeast library (get Function), This file was represent the best three predictions for the tweets image and the first one prediction is the most accurate one so I kept it and remove the anothers.

3- tweet-json.txt is the final file. there was two ways to get this final dataset and I choose to get the file directly by downloading it as I could not get atwitter developer account. And to can work with it I used json library (json.loaad function) to extract the main important data .

Data Assessment

In this step I started the important datasets investigation in two ways visually and programmatically to can find quality and tidiness issues :

1-Visually: for visual assessment I used jupyter notebook ,To be able to this I changed the pandas options to display.max_rows and display.max_colwidth.

2- I divided the data Assessment method into three main steps:

A-twitter-archive-enhanced assess.

B-Image-predictions assess.

c- json file visual assess.

*Every step contains two subset steps (visually and programmatically).

*At the end of every main step I wort the quality and tidiness issues to help me remember all issues .

Data Cleaning

Table 1

Cleaning Table

Table	Issue	Solution
Twitter_archive_enhanced quality issues	There is rows for retweets without image	Remove rows without image by keeping only tweets_with_image(find by image_predictions **table)**
	No Need For retweets rows	Remove retweets rows by keeping only tweets with retweeted_status_id.notnull
	No need for replayes rows	Remove replayes rows by keeping only tweets with in_reply_to_status_id.notnull
	The tweet_id column is appear in scientific notation	change tweet_id column from numeric type to str(object)
	Timestamp Dtype shoud be Datetime type	Change timestamp Dtype to Datetime type
	'rating_numerator' with values bigger than 20	replac 'rating_numerator' with values bigger than 20 with nan
	There is Wrong values in 'rating_numerator' with index =1202 as its rating in tweet is 11/10	Replace Wrong values in 'rating_numerator' with index =1202 as rating in tweet is 11/10
	There is Wrong values in 'rating_denominator' with index =1202 as rating in tweet is 11/10	Replace Wrong values in 'rating_denominator' with index =1202 as rating in tweet is 11/10
	Source column represent the tweet source with wrong way	Replace Source column to represent the tweet source (iphone , Twitter Web Client ,TweetDeck) only
	replace expanded_urls dublicated values with only one value in same cell	replace expanded_urls dublicated values with only one value in same cell
	wrong values in 'name'	Replace wrong values in 'name' with np.NaN
	There is “None” string in the last four columns	Replace the “None” string in the last four columns with ' '
image_predictions quality issues	the tweet_id are appear in int64 (numeric) shoud be object dtype.	Change tweet_id dtype from int64 to object dtype
	the last 9 columns are named by confused way as not explain to what it contains.	Change the name of the columns
	Some Values in p1 , p2 and p3 start with small letter	Capitalize the letters
	Some values in p1 , p2 and p3 have _ or - .	Replace this value
	the 'p1' , 'p1_conf ’ ,'p1_dog' columns are named by confused way as not explain to what it contains	Replace _ or - in p1 , p2 and p3 which I named(prediction_1,prediction_2,prediction_3) to '
	2075 row for image_predictions while shoud be 1971 row	Dropping the retweets and replies ids from the image prediction dataframe

Table	Issue	Solution
tweet-json Quality issues	names column not meaningful	rename(columns) retweet_count':'retweets','favorite_count':'likes'
	retweets and replies shoud be remove from the image prediction dataframe	Dropping the retweets and replies ids from the image prediction dataframe
All files Tidiness issues	No need for in_reply_to_status_id in_reply_to_user_id , retweeted_status_id , retweeted_status_user_id , retweeted_status_timestamp as are irrelevant information or attributes to analysis and observational under study¶	Remove all of this columns by create a drop list for them so can remove by one step
	One variable is expressed in four columns(doggo , floof , pupper , puppo) should be one column with header(dog_kind)	combine four columns(doggo , floof , pupper , puppo) in one column with header(dog_class) and drop four columns
	The 'p2' , 'p2_conf ' ,'p2_dog' , 'p3' , 'p3_conf ' and' p3_dog' columns are not usefull because the P1 is the highest ratio and p1 cancelled another columns so it irrelevant columns or attributes to the observational under study	Drop the last 6 columns
	A single observational unit is stored in multiple tables(thetwitter-archive , image-predictions and tweet-json shoud be in same table)	Combine all of them in one file named twitter_archive_master

OUTPUT :-

One file named twitter_archive_master.