Quality issues

twitter_archive_enhanced_df

1- wrong data types:

any column describe an ID should as a string

These ID column should be strings instead of numerical data types because I don't need to deal with them as number "no need for statistics, visualization"

timestamp should be at date format

Timestamp is represented as string but it should be at date format so I can use pandas date functions on it .

timestamp 2017-08-01 16:23:56 +0000 2017-08-01 00:17:27 +0000 2017-07-31 00:18:03 +0000

2- 'source' column has the tag appear in it .

source
Twitter for iPhone

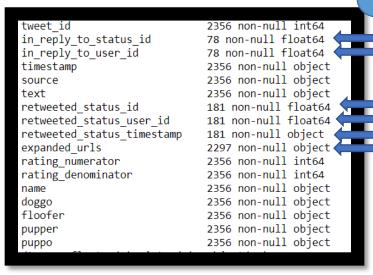
The HTML tag should be removed so the data can be easily to read and manipulate. In the previous example it easy to deal with source when it represented like this

→ http://twitter.com/download/iphone

I use split function then I take the second index

source Twitter for iPhone





As you see there is many null values especially in 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'expanded_urls'.

So it's so hard to use this column I will remove them

4- invalid dog names.

there is some wired dog name which clearly not dog name . e.g. : ['a', 'quite', 'General']

5- invalid rating_numerator and rating_denominator

there is some rating numerator and rating denominator that doesn't make sense as you see

numerator all value counts

denominator weird value count

10	2333
11	3
50	3
80	2
20	2
2	1
16	1
40	1
70	1
15	1
90	1
110	1
120	1
130	1
150	1
170	1
7	1
0	1

There is 23 numerator value which isn't valid because it supposed to be 10

I compare this value with the original tweet text and replace them with the write values manually .

420	2
75	2
60	1
27	1
99	1
165	1
80	1
144	1
204	1
45	1
88	1
143	1
1776	1
44	1
50	1
26	1
84	1
182	1
24	1
121	1
666	1
960	1

There is 24 denominator value which doesn't make sense because it supposed to be less than 20

For example :-

763

Here the numerator is 960 and the dementor is 0 but in the real text there are 13-10 so I replaced them manually . for example

188 855862651834028034 -> no problem 189 855860136149123072 -> no problem 313 835246439529840640 -> 13 340 832215909146226688 -> 9.75 433 820690176645140481 ->group 695 786709082849828864 -> 9.75

778027034220126208 -> 11.27

342 832088576586297345 -> wrong data 433 820690176645140481 -> group of dogs 516 810984652412424192 -> 7 784 775096608509886464 -> 10 1068 740373189193256964 -> 10 1662 682962037429899265 -> 10 902 758467244762497024 -> group of dogs 2335 666287406224695296 -> 10

313 835246439529840640 -> 10

The red mean wrong data so I will drop it

In the case of group I will divide them by there number $88/80 \rightarrow 11/10$

Green data will be modified

the rest is true but there are outliers so I will delete them to make my analysis clear

6-there is wrong tweet text

There are some text that isn't clear :-

@docmisterio account started on 11/15/15

There is many text cell which star with "RT @dog_rates:"

RT @dog_rates: This is Moreton. He's the Good Boy Who Lived. 13/10 magical as h*ck https://t.co/rLHGx3VAF3

The text include the image url → tidness "I will not solve it "

7-Wrong representation of NONE values in name column

image_prediction_df → wrong ID data type, it should be string

1- in image_prediction_df (p1, p2 and p3) aren't descriptive names

tweets information df \rightarrow wrong ID data type, it should be string

tidiness issues

2- in twitter_archive_enhancment there is 1 variable in 4 column ['doggo', 'floofer', 'pupper', 'puppo'] but it better to be represented in one column called dog stage.

loggo	None	None	None	doggo
Vone	None	None	None	None
Vone	None	None	None	None
Vone	None	None	puppo	puppo

3- the text column has the tweet text and the picture url

RT @dog rates: This is Loki. He smiles like Elvis. Ain't nothin but a hound doggo. 12/10 https://t.co/QV5nx6otZR

4- data in 3 separate tables.

it's better to concatenate all related data in the same data frame

as example: - retweet count, and favorite count should be in twitter archive enhancment.

Α	В	C	D	E	F	G	H	1	J	K	L	M	N	0	Р	Q	R	S
tweet_id	timestamp	source	text	rating_numera	rating_denom	name	dog_stage	favorite_count	retweet_count	predictio_1	p1_conf	p1_dog	prediction_2	p2_conf	p2_dog	prediction	p3_conf	p3_dog
8.92421E+17	8/1/2017 16:23	http://twit	This is Phir	13	10	Phineas		39467	8853	orange	0.097049	FALSE	bagel	0.085851	FALSE	banana	0.07611	FALSE
8.92177E+17	8/1/2017 0:17	http://twi	This is Tilly	13	10	Tilly		33819	6514	Chihuahua	0.323581	TRUE	Pekinese	0.090647	TRUE	papillon	0.068957	TRUE
8.91815E+17	7/31/2017 0:18	http://twi	This is Arch	12	10	Archie		25461	4328	Chihuahua	0.716012	TRUE	malamute	0.078253	TRUE	kelpie	0.031379	TRUE