Data wrangling Report

Dataset description

In this project we work with 3 datasets, 2 coming from twitter and one coming from the result of a prediction algorithm. These are the Twitter archives of WeRateDogs. This data contains information from different dogs.

For a good analysis of the data, we made two evaluations which are:

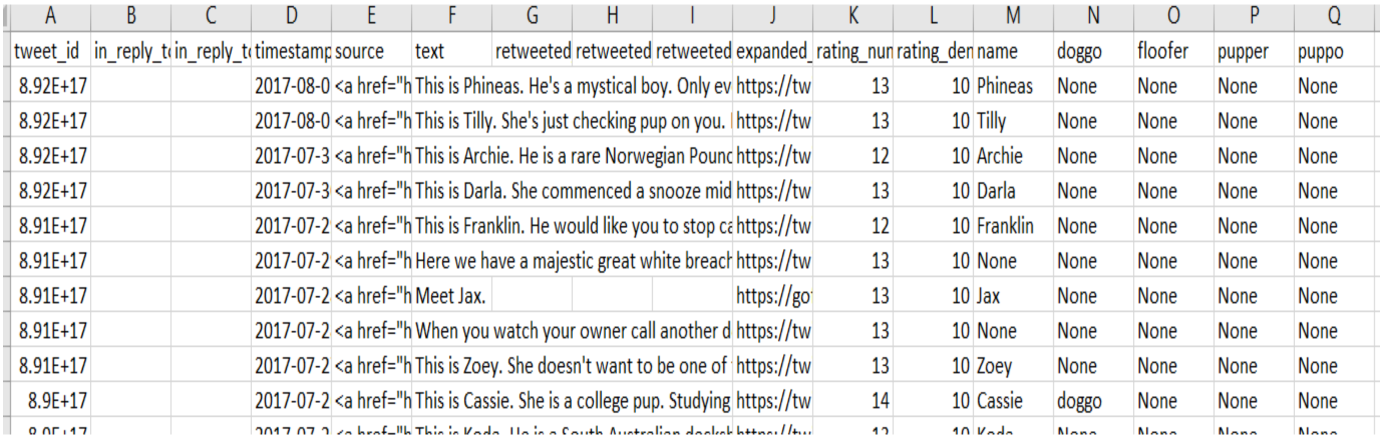
- The visual evaluation on Excel

- The programmatic evaluation

Visual evaluations

During the visual evaluation of our data, we inspected each column to detect anomalies

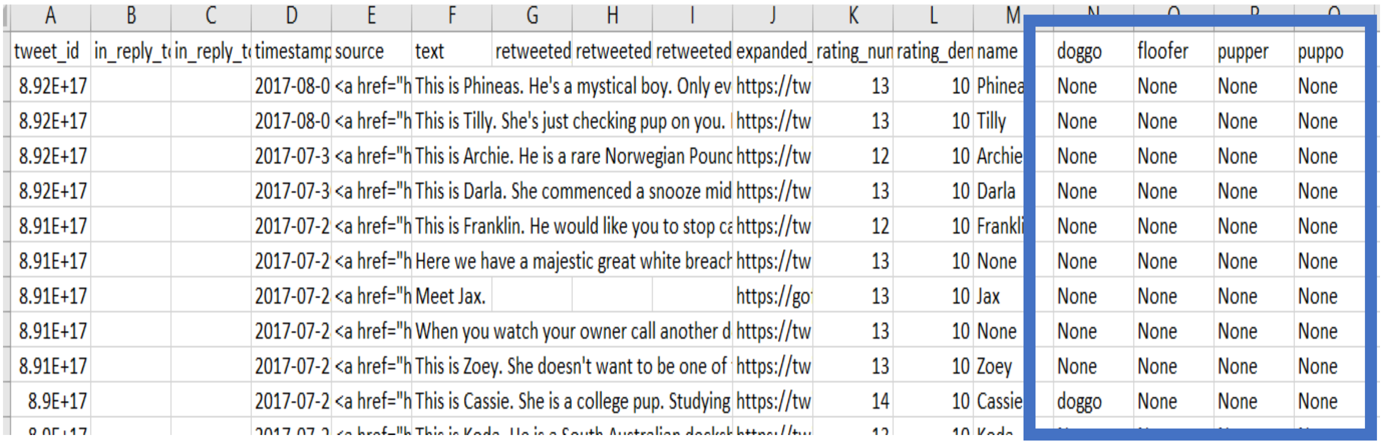
❖ Twitter_archive_enhanced.csv



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
tweet_id	in_reply_t	in_reply_t	timestamp	source	text	retweeted	retweeted	retweeted	expanded_	rating_num	rating_den	name	doggo	floofer	pupper	puppo
8.92E+17			2017-08-0	<a href="h	This is Phineas. He's a mystical boy. Only ev				https://tw	13	10	Phineas	None	None	None	None
8.92E+17			2017-08-0	<a href="h	This is Tilly. She's just checking pup on you.				https://tw	13	10	Tilly	None	None	None	None
8.92E+17			2017-07-3	<a href="h	This is Archie. He is a rare Norwegian Pounc				https://tw	12	10	Archie	None	None	None	None
8.92E+17			2017-07-3	<a href="h	This is Darla. She commenced a snooze mid				https://tw	13	10	Darla	None	None	None	None
8.91E+17			2017-07-2	<a href="h	This is Franklin. He would like you to stop ca				https://tw	12	10	Franklin	None	None	None	None
8.91E+17			2017-07-2	<a href="h	Here we have a majestic great white breach				https://tw	13	10	None	None	None	None	None
8.91E+17			2017-07-2	<a href="h	Meet Jax.				https://go	13	10	Jax	None	None	None	None
8.91E+17			2017-07-2	<a href="h	When you watch your owner call another d				https://tw	13	10	None	None	None	None	None
8.91E+17			2017-07-2	<a href="h	This is Zoey. She doesn't want to be one of				https://tw	13	10	Zoey	None	None	None	None
8.9E+17			2017-07-2	<a href="h	This is Cassie. She is a college pup. Studying				https://tw	14	10	Cassie	doggo	None	None	None
8.9E+17			2017-07-2	<a href="h	This is Koda. He is a South Australian deckh				https://tw	13	10	Koda	None	None	None	None

While exploring these data we have detected some quality and order problems

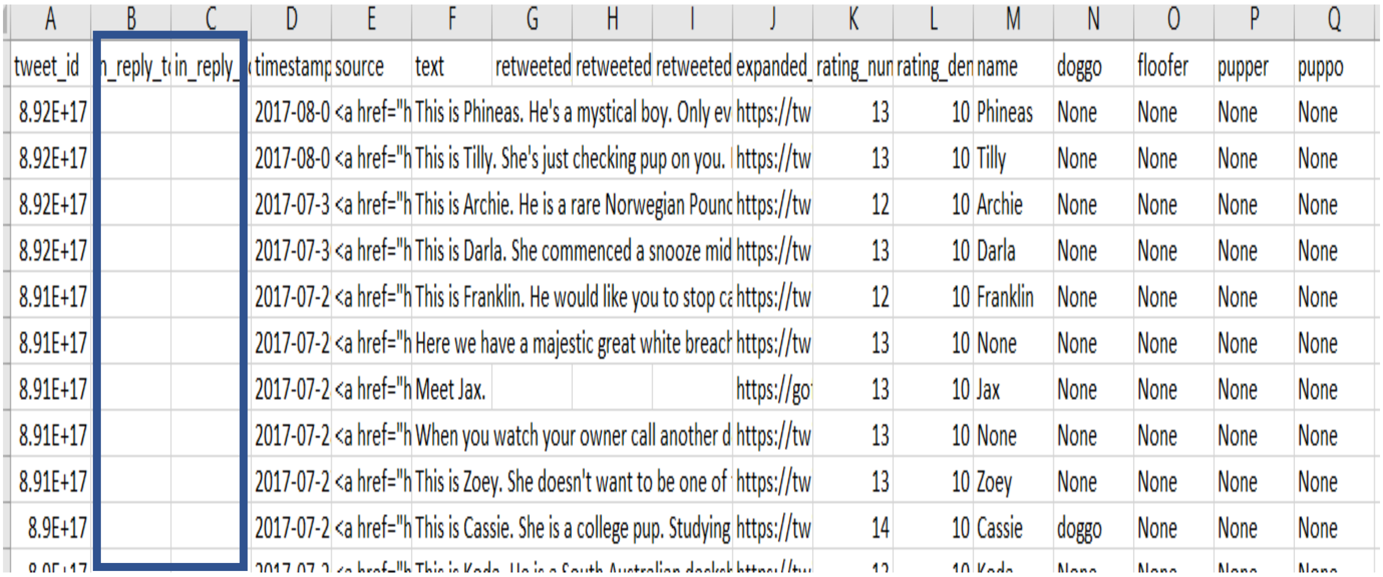
At first glance we can detect a problem of order with the columns doggo, flopper, pupper and puppo which must be transformed into a single column.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
tweet_id	in_reply_t	in_reply_t	timestamp	source	text	retweeted	retweeted	retweeted	expanded_	rating_num	rating_den	name	doggo	floofer	pupper	puppo
8.92E+17			2017-08-0	<a href="h	This is Phineas. He's a mystical boy. Only ev				https://tw	13	10	Phinea	None	None	None	None
8.92E+17			2017-08-0	<a href="h	This is Tilly. She's just checking pup on you.				https://tw	13	10	Tilly	None	None	None	None
8.92E+17			2017-07-3	<a href="h	This is Archie. He is a rare Norwegian Pounc				https://tw	12	10	Archie	None	None	None	None
8.92E+17			2017-07-3	<a href="h	This is Darla. She commenced a snooze mid				https://tw	13	10	Darla	None	None	None	None
8.91E+17			2017-07-2	<a href="h	This is Franklin. He would like you to stop ca				https://tw	12	10	Frankli	None	None	None	None
8.91E+17			2017-07-2	<a href="h	Here we have a majestic great white breach				https://tw	13	10	None	None	None	None	None
8.91E+17			2017-07-2	<a href="h	Meet Jax.				https://go	13	10	Jax	None	None	None	None
8.91E+17			2017-07-2	<a href="h	When you watch your owner call another d				https://tw	13	10	None	None	None	None	None
8.91E+17			2017-07-2	<a href="h	This is Zoey. She doesn't want to be one of				https://tw	13	10	Zoey	None	None	None	None
8.9E+17			2017-07-2	<a href="h	This is Cassie. She is a college pup. Studying				https://tw	14	10	Cassie	doggo	None	None	None
8.9E+17			2017-07-2	<a href="h	This is Koda. He is a South Australian deckh				https://tw	13	10	Koda				

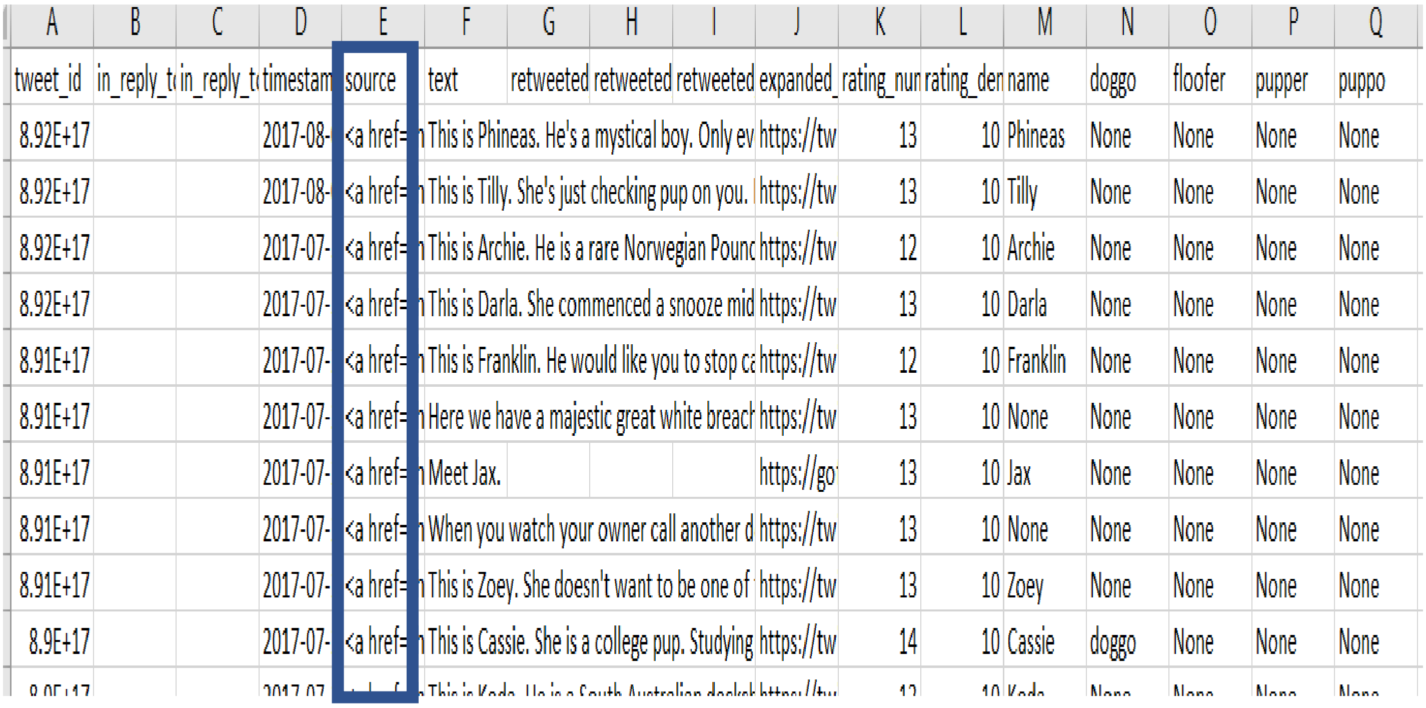
- Image_predictions.tsv

We can also see that there are missing values, another observation would be to notice that the extraction code for the value that it could not extract or that does not exist has filled the information with None. This information helped us a lot, but we did not decide to delete these values because they contain important information.



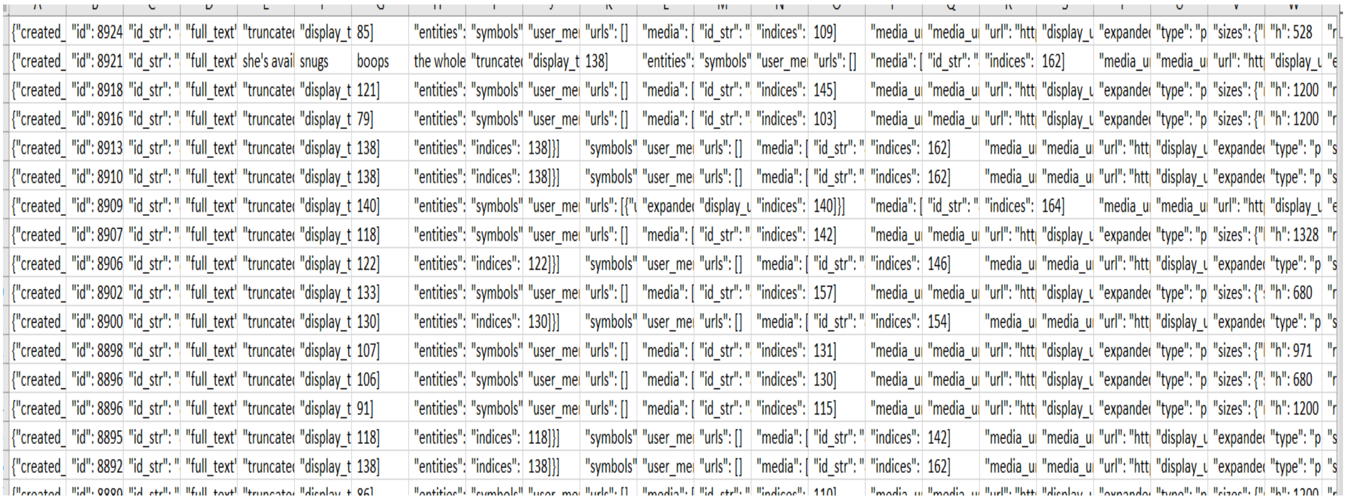
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
tweet_id	n_reply_t	in_reply_	timestamp	source	text	retweeted	retweeted	retweeted	expanded_	rating_nun	rating_den	name	doggo	floofer	pupper	puppo
8.92E+17			2017-08-0	<a href="h	This is Phineas. He's a mystical boy. Only ev				https://tw	13	10	Phineas	None	None	None	None
8.92E+17			2017-08-0	<a href="h	This is Tilly. She's just checking pup on you.				https://tw	13	10	Tilly	None	None	None	None
8.92E+17			2017-07-3	<a href="h	This is Archie. He is a rare Norwegian Pounc				https://tw	12	10	Archie	None	None	None	None
8.92E+17			2017-07-3	<a href="h	This is Darla. She commenced a snooze mid				https://tw	13	10	Darla	None	None	None	None
8.91E+17			2017-07-2	<a href="h	This is Franklin. He would like you to stop ca				https://tw	12	10	Franklin	None	None	None	None
8.91E+17			2017-07-2	<a href="h	Here we have a majestic great white breach				https://tw	13	10	None	None	None	None	None
8.91E+17			2017-07-2	<a href="h	Meet Jax.				https://go	13	10	Jax	None	None	None	None
8.91E+17			2017-07-2	<a href="h	When you watch your owner call another d				https://tw	13	10	None	None	None	None	None
8.91E+17			2017-07-2	<a href="h	This is Zoey. She doesn't want to be one of				https://tw	13	10	Zoey	None	None	None	None
8.9E+17			2017-07-2	<a href="h	This is Cassie. She is a college pup. Studying				https://tw	14	10	Cassie	doggo	None	None	None
8.9E+17			2017-07-2	<a href="h	This is Koda. He is a South Australian decksh				https://tw	13	10	Koda	None	None	None	None

You can see some html tags in the source column that we have removed with replace in the data cleaning section.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
tweet_id	in_reply_t	in_reply_t	timestam	source	text	retweeted	retweeted	retweeted	expanded_	rating_nun	rating_den	name	doggo	floofer	pupper	puppo
8.92E+17			2017-08-	<a href=	This is Phineas. He's a mystical boy. Only ev				https://tw	13	10	Phineas	None	None	None	None
8.92E+17			2017-08-	<a href=	This is Tilly. She's just checking pup on you.				https://tw	13	10	Tilly	None	None	None	None
8.92E+17			2017-07-	<a href=	This is Archie. He is a rare Norwegian Pounc				https://tw	12	10	Archie	None	None	None	None
8.92E+17			2017-07-	<a href=	This is Darla. She commenced a snooze mid				https://tw	13	10	Darla	None	None	None	None
8.91E+17			2017-07-	<a href=	This is Franklin. He would like you to stop ca				https://tw	12	10	Franklin	None	None	None	None
8.91E+17			2017-07-	<a href=	Here we have a majestic great white breach				https://tw	13	10	None	None	None	None	None
8.91E+17			2017-07-	<a href=	Meet Jax.				https://go	13	10	Jax	None	None	None	None
8.91E+17			2017-07-	<a href=	When you watch your owner call another d				https://tw	13	10	None	None	None	None	None
8.91E+17			2017-07-	<a href=	This is Zoey. She doesn't want to be one of				https://tw	13	10	Zoey	None	None	None	None
8.9E+17			2017-07-	<a href=	This is Cassie. She is a college pup. Studying				https://tw	14	10	Cassie	doggo	None	None	None
8.9E+17			2017-07	<a href=	This is Koda. He is a South Australian decksh				https://tw	13	10	Koda	None	None	None	None

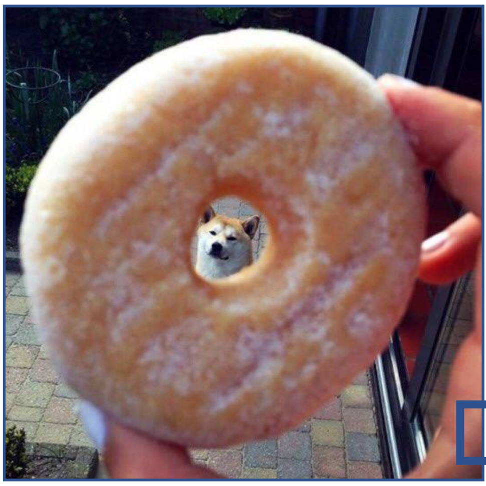
- Tweet_json.txt

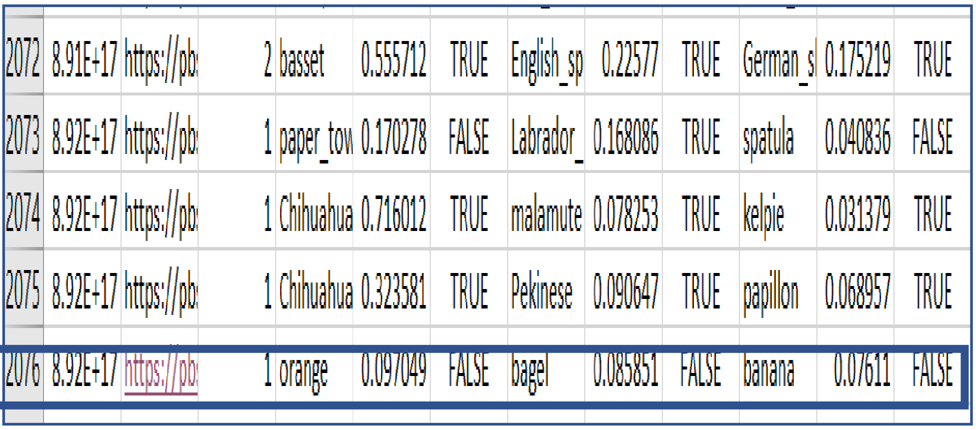


{"created_	"id": 8924	"id_str": "	"full_text"	"truncate	"display_t	85]	"entities":	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	109]	"media_u	"media_u	"url": "htt	"display_u	"expande	"type": "p	"sizes": {"	"h": 528	"r
{"created_	"id": 8921	"id_str": "	"full_text"	she's avail	snugs	boops	the whole	"truncate	"display_t	138]	"entities":	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	162]	"media_u	"media_u	"url": "htt	"display_u	"e
{"created_	"id": 8918	"id_str": "	"full_text"	"truncate	"display_t	121]	"entities":	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	145]	"media_u	"media_u	"url": "htt	"display_u	"expande	"type": "p	"sizes": {"	"h": 1200	"r
{"created_	"id": 8916	"id_str": "	"full_text"	"truncate	"display_t	79]	"entities":	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	103]	"media_u	"media_u	"url": "htt	"display_u	"expande	"type": "p	"sizes": {"	"h": 1200	"r
{"created_	"id": 8913	"id_str": "	"full_text"	"truncate	"display_t	138]	"entities":	"indices":	138]}]	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	162]	"media_u	"media_u	"url": "htt	"display_u	"expande	"type": "p	"s
{"created_	"id": 8910	"id_str": "	"full_text"	"truncate	"display_t	138]	"entities":	"indices":	138]}]	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	162]	"media_u	"media_u	"url": "htt	"display_u	"expande	"type": "p	"s
{"created_	"id": 8909	"id_str": "	"full_text"	"truncate	"display_t	140]	"entities":	"symbols"	"user_me	"urls": [{"u	"expande	"display_u	"indices":	140]}]	"media": [	"id_str": "	"indices":	164]	"media_u	"media_u	"url": "htt	"display_u	"e
{"created_	"id": 8907	"id_str": "	"full_text"	"truncate	"display_t	118]	"entities":	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	142]	"media_u	"media_u	"url": "htt	"display_u	"expande	"type": "p	"sizes": {"	"h": 1328	"r
{"created_	"id": 8906	"id_str": "	"full_text"	"truncate	"display_t	122]	"entities":	"indices":	122]}]	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	146]	"media_u	"media_u	"url": "htt	"display_u	"expande	"type": "p	"s
{"created_	"id": 8902	"id_str": "	"full_text"	"truncate	"display_t	133]	"entities":	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	157]	"media_u	"media_u	"url": "htt	"display_u	"expande	"type": "p	"sizes": {"	"h": 680	"r
{"created_	"id": 8900	"id_str": "	"full_text"	"truncate	"display_t	130]	"entities":	"indices":	130]}]	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	154]	"media_u	"media_u	"url": "htt	"display_u	"expande	"type": "p	"s
{"created_	"id": 8898	"id_str": "	"full_text"	"truncate	"display_t	107]	"entities":	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	131]	"media_u	"media_u	"url": "htt	"display_u	"expande	"type": "p	"sizes": {"	"h": 971	"r
{"created_	"id": 8896	"id_str": "	"full_text"	"truncate	"display_t	106]	"entities":	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	130]	"media_u	"media_u	"url": "htt	"display_u	"expande	"type": "p	"sizes": {"	"h": 680	"r
{"created_	"id": 8896	"id_str": "	"full_text"	"truncate	"display_t	91]	"entities":	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	115]	"media_u	"media_u	"url": "htt	"display_u	"expande	"type": "p	"sizes": {"	"h": 1200	"r
{"created_	"id": 8895	"id_str": "	"full_text"	"truncate	"display_t	118]	"entities":	"indices":	118]}]	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	142]	"media_u	"media_u	"url": "htt	"display_u	"expande	"type": "p	"s
{"created_	"id": 8892	"id_str": "	"full_text"	"truncate	"display_t	138]	"entities":	"indices":	138]}]	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	162]	"media_u	"media_u	"url": "htt	"display_u	"expande	"type": "p	"s
{"created_	"id": 8889	"id_str": "	"full_text"	"truncate	"display_t	86]	"entities":	"symbols"	"user_me	"urls": []	"media": [	"id_str": "	"indices":	110]	"media_u	"media_u	"url": "htt	"display_u	"expande	"type": "p	"sizes": {"	"h": 1200	"r

On the other hand, for the two datasets Tweet_json and images_predictions.tsv the visual evaluation did not allow me to detect any major anomalies. However, I could detect a quality problem in images_predictions some images had no classification, that is to say that the algorithm could not predict correctly the breed of the dog.

Example

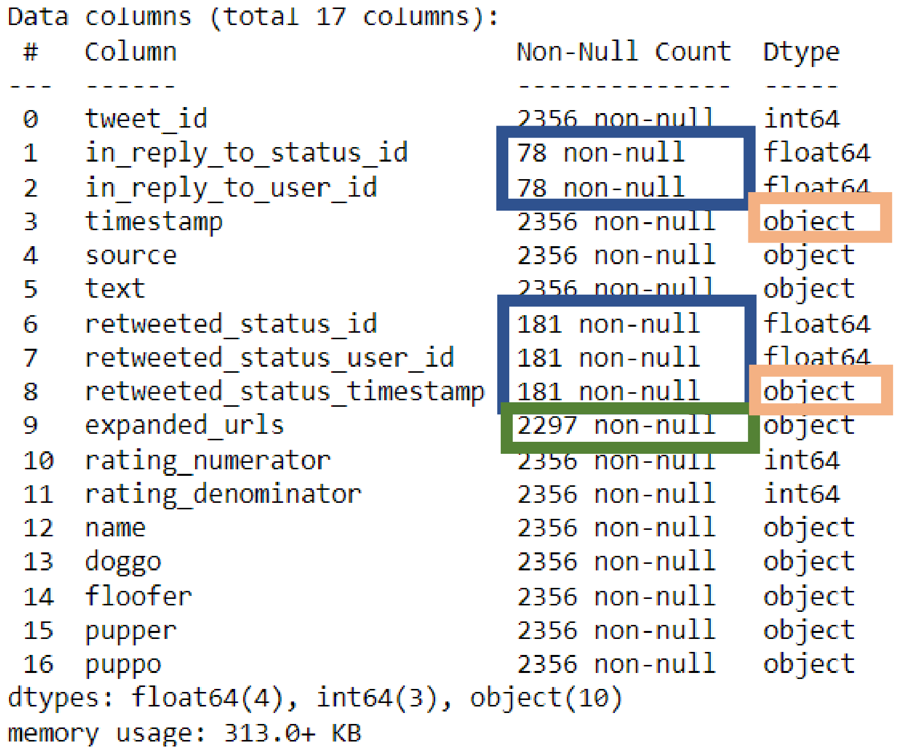




2072	8.91E+17	https://pb	2	basset	0.555712	TRUE	English_sp	0.22577	TRUE	German_s	0.175219	TRUE
2073	8.92E+17	https://pb	1	paper_tow	0.170278	FALSE	Labrador_	0.168086	TRUE	spatula	0.040836	FALSE
2074	8.92E+17	https://pb	1	Chihuahua	0.716012	TRUE	malamute	0.078253	TRUE	kelpie	0.031379	TRUE
2075	8.92E+17	https://pb	1	Chihuahua	0.323581	TRUE	Pekinese	0.090647	TRUE	papillon	0.068957	TRUE
2076	8.92E+17	https://pb	1	orange	0.097049	FALSE	bagel	0.085851	FALSE	banana	0.07611	FALSE

Evaluations programmatiques

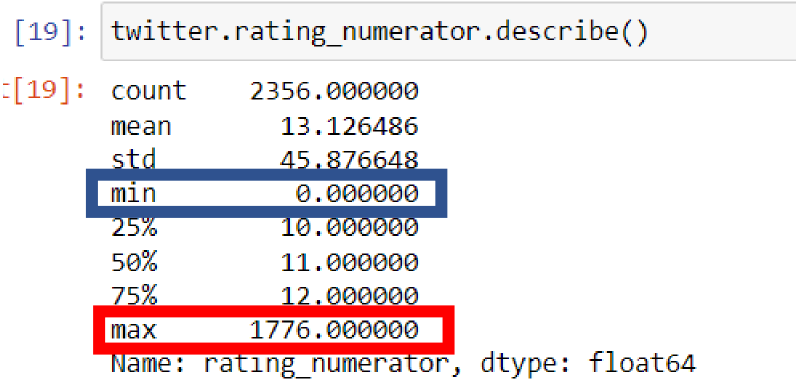
The programmatic evaluation consisted of detecting problems using the code:



```
Data columns (total 17 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   tweet_id                    2356 non-null   int64  
 1   in_reply_to_status_id       78 non-null     float64
 2   in_reply_to_user_id         78 non-null     float64
 3   timestamp                   2356 non-null   object 
 4   source                      2356 non-null   object 
 5   text                        2356 non-null   object 
 6   retweeted_status_id         181 non-null    float64
 7   retweeted_status_user_id    181 non-null    float64
 8   retweeted_status_timestamp  181 non-null    object 
 9   expanded_urls               2297 non-null   object 
 10  rating_numerator            2356 non-null   int64  
 11  rating_denominator          2356 non-null   int64  
 12  name                        2356 non-null   object 
 13  doggo                       2356 non-null   object 
 14  floofer                     2356 non-null   object 
 15  pupper                      2356 non-null   object 
 16  puppo                       2356 non-null   object 
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

With the info functions we confirmed the presence of missing values.

And the described commands on the numerator and denominator of the rating allowed me to detect a problem among the rating as numerator rating and or denominator rating equal to zero, which I solved by deleting the values equal to zero and creating a new column rate_dog that represents the ratio of these two values.

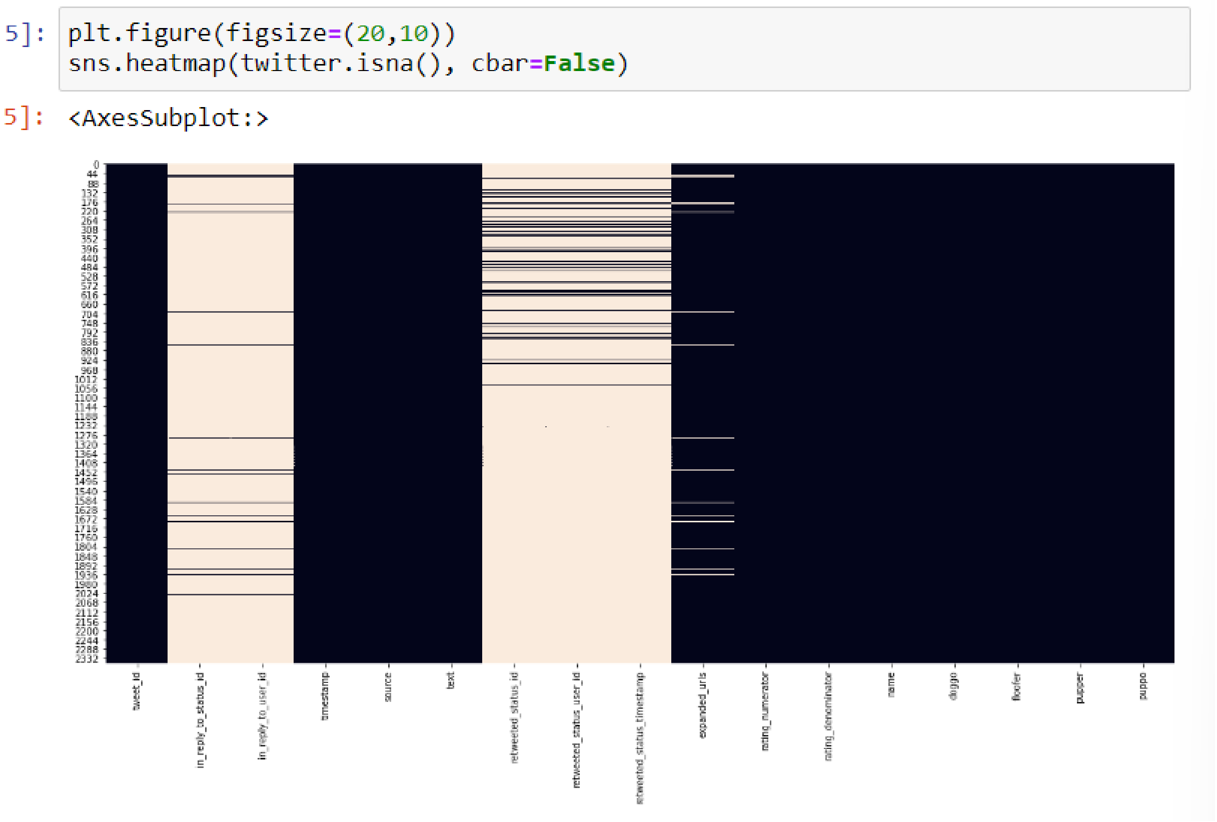


```
[19]: twitter.rating_numerator.describe()

[19]: count    2356.000000
      mean       13.126486
      std        45.876648
      min         0.000000
      25%        10.000000
      50%        11.000000
      75%        12.000000
      max      1776.000000
      Name: rating_numerator, dtype: float64
```

Thanks to a visualization with heatmap I could notice a disparity between the data in relation to the number of missing values which pushed me to remove from my final dataset the following columns:

[['retweeted_status_id', 'retweeted_status_user_id','retweeted_status_timestamp', 'in_reply_to_status_id', 'in_reply_to_user_id',]].

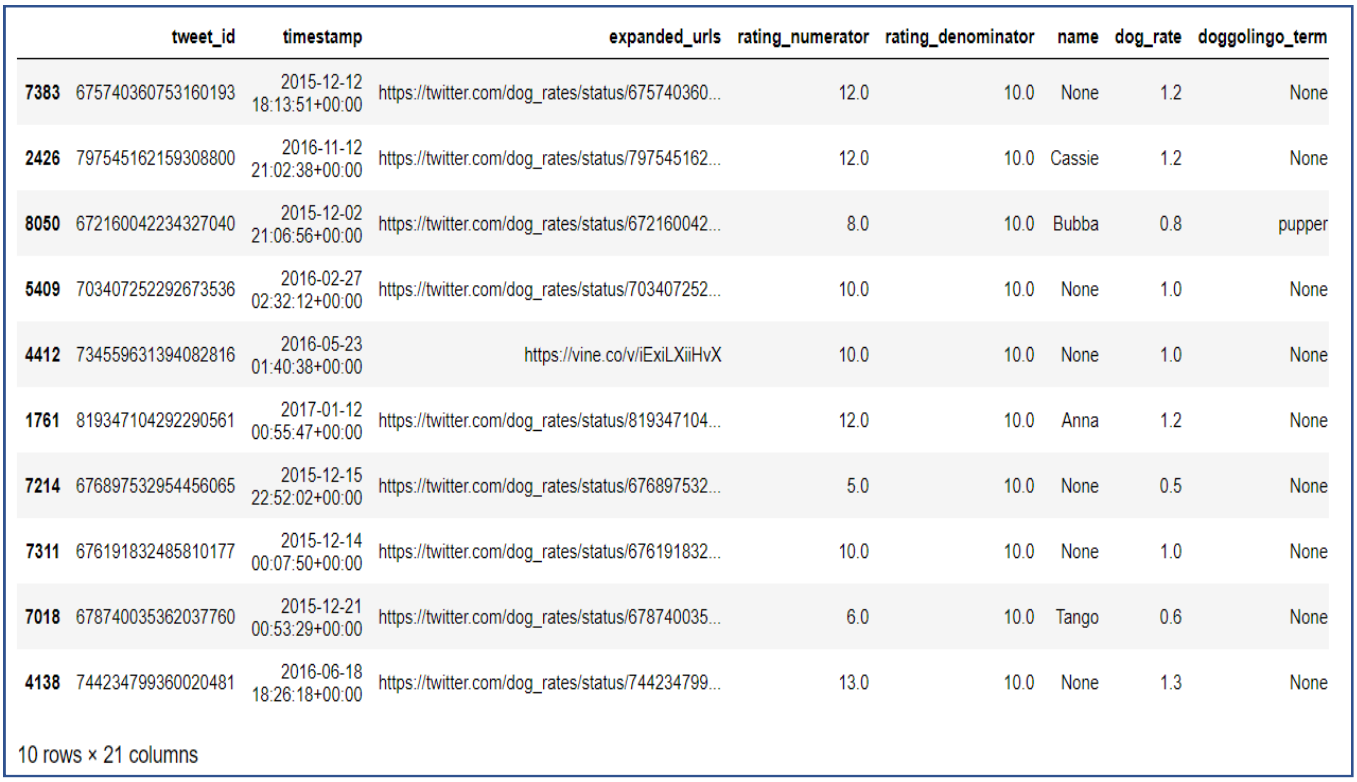


```
5]: plt.figure(figsize=(20,10))
    sns.heatmap(twitter.isna(), cbar=False)

5]: <AxesSubplot:>
```

Result

Final dataset



	tweet_id	timestamp	expanded_urls	rating_numerator	rating_denominator	name	dog_rate	doggolingo_term
7383	675740360753160193	2015-12-12 18:13:51+00:00	https://twitter.com/dog_rates/status/675740360...	12.0	10.0	None	1.2	None
2426	797545162159308800	2016-11-12 21:02:38+00:00	https://twitter.com/dog_rates/status/797545162...	12.0	10.0	Cassie	1.2	None
8050	672160042234327040	2015-12-02 21:06:56+00:00	https://twitter.com/dog_rates/status/672160042...	8.0	10.0	Bubba	0.8	pupper
5409	703407252292673536	2016-02-27 02:32:12+00:00	https://twitter.com/dog_rates/status/703407252...	10.0	10.0	None	1.0	None
4412	734559631394082816	2016-05-23 01:40:38+00:00	https://vine.co/v/iExiLXiiHvX	10.0	10.0	None	1.0	None
1761	819347104292290561	2017-01-12 00:55:47+00:00	https://twitter.com/dog_rates/status/819347104...	12.0	10.0	Anna	1.2	None
7214	676897532954456065	2015-12-15 22:52:02+00:00	https://twitter.com/dog_rates/status/676897532...	5.0	10.0	None	0.5	None
7311	676191832485810177	2015-12-14 00:07:50+00:00	https://twitter.com/dog_rates/status/676191832...	10.0	10.0	None	1.0	None
7018	678740035362037760	2015-12-21 00:53:29+00:00	https://twitter.com/dog_rates/status/678740035...	6.0	10.0	Tango	0.6	None
4138	744234799360020481	2016-06-18 18:26:18+00:00	https://twitter.com/dog_rates/status/744234799...	13.0	10.0	None	1.3	None

10 rows × 21 columns