

This file is a documentation of my wrangling effort on this project:

## First of all Gathering Data:

- I found this part was the easiest and fastest part. I tried to import the files as required.
- The instructions to import these files was very straight to the point and all in the course content

## Second part Assessing Data:

- This part was very challenging for me as it was very large data and I took too much time to digest the data and understand what really needed to be done
- I've Assessed the data programmatically using jupyter notebook and Visually using excel
- I tried to address each Issue I found to the method detected with in the ipynb files
- I've used many pandas methods like `.info()` / `.head()` / `.describe()` / `.value_counts()` / `.nunique()`, and also tried to be very explanatory in the jupyter notebook files to see how is my workflow is going
- For the visual part I also transformed TSV file to a CSV file so that I can read it easily as shown, Atom was very hard to see the patterns in it

File Home Insert Page Layout Formulas Data Review View Help Tell me what you want to do

Calibri 11 A A

B I U A- A

Font

Wrap Text

Alignment

General

Number

Conditional Formatting

Normal Bad Good Neutral Calculation

Check Cell Explanatory... Followed By... Hyperlink Input

Styles

Cell

B1 =COUNTA(B3:B2358)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1		2356	78	78	2356	2356	2356	181	181	181	2297	2356	2356	2356	2356	2356	2356	2356
2	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source		retweeter	retweeter	retweeter	expanded	rating	rating	de	name	doggo	floof	pupper	puppo
3	8.92421E+17			2017-08-01	ca href="http://twitter						https://tw	13	10	Phineas	None	None	None	None
4	8.92177E+17			2017-08-01	ca href="http://twitter						https://tw	13	10	Tilly	None	None	None	None
5	8.91815E+17			2017-07-31	ca href="http://twitter						https://tw	12	10	Archie	None	None	None	None
6	8.91696E+17			2017-07-31	ca href="http://twitter						https://tw	13	10	Darla	None	None	None	None
7	8.91328E+17			2017-07-21	ca href="http://twitter						https://tw	12	10	Franklin	None	None	None	None
8	8.91088E+17			2017-07-21	ca href="http://twitter						https://tw	13	10	None	None	None	None	None
9	8.90972E+17			2017-07-21	ca href="http://twitter						https://gc	13	10	Jax	None	None	None	None
10	8.90729E+17			2017-07-21	ca href="http://twitter						https://tw	13	10	None	None	None	None	None
11	8.90609E+17			2017-07-21	ca href="http://twitter						https://tw	13	10	Zoe	None	None	None	None
12	8.9024E+17			2017-07-21	ca href="http://twitter						https://tw	14	10	Cassie	doggo	None	None	None
13	8.90007E+17			2017-07-21	ca href="http://twitter						https://tw	13	10	Koda	None	None	None	None
14	8.89881E+17			2017-07-21	ca href="http://twitter						https://tw	13	10	Bruno	None	None	None	None
15	8.89665E+17			2017-07-21	ca href="http://twitter						https://tw	13	10	None	None	None	None	puppo
16	8.89639E+17			2017-07-21	ca href="http://twitter						https://tw	12	10	Ted	None	None	None	None
17	8.89531E+17			2017-07-21	ca href="http://twitter						https://tw	13	10	Stuart	None	None	None	puppo
18	8.89279E+17			2017-07-21	ca href="http://twitter						https://tw	13	10	Oliver	None	None	None	None
19	8.88917E+17			2017-07-21	ca href="http://twitter						https://tw	12	10	Jim	None	None	None	None
20	8.88805E+17			2017-07-21	ca href="http://twitter						https://tw	13	10	Zeke	None	None	None	None
21	8.88555E+17			2017-07-21	ca href="http://twitter						https://tw	13	10	Ralphus	None	None	None	None
22	8.88203E+17			2017-07-21	ca href="http://twitter						https://tw	13	10	Canela	None	None	None	None
23	8.88078E+17			2017-07-21	ca href="http://twitter						https://tw	12	10	Gerald	None	None	None	None
24	8.87705E+17			2017-07-11	ca href="http://twitter						https://tw	13	10	Jeffrey	None	None	None	None
25	8.87517E+17			2017-07-11	ca href="http://twitter						https://tw	14	10	such	None	None	None	None
26	8.87474E+17			2017-07-11	ca href="http://twitter						https://tw	13	10	Canela	None	None	None	None

[illegible]

My wrangling effort ended with these findings:

## Tidiness Issues:

archive\_df

- related rows to retweets and replies (259 row)
- 4 columns of dogs classification
- 5 columns of un needed data related to retweets and replies

api\_df

- favorites and retweet count columns needed to be merged with the first table to create a unit of tweets' observation

## Quality Issues:

archive\_df

- tweet\_id: is int [#programmatically](#)
- timestamp: is object [#programmatically](#)
- None are represented as string not NaN object [#programmatically](#)
- tweets with empty images (extended\_url) [#programmatically](#)
- tweets has two classifications together [#visually](#)
- some dog names are not extracted correctly [#visually](#)
- there are (rating\_denominator) > 10 which manipulate the scale of its numerator likr rows 903 and 1121 [#programmatically](#)
- source column is not represented in clear naming [#visually](#)
- tweets with no dog classification [#visually](#)

image\_predictions\_df

- tweet\_id: is int [#programmatically](#)
- there images in this table is not in archive table -visually- [#visually](#)

api\_df

- tweet\_id: is int -programmatically-

To be honest I didn't manage to solve 2 points of them thus I was short in time.  
And yes I'm really looking forward for your help and criticize

-----

## Second part Assessing Data:

- Cleaning part was the most time consuming part I took a lot of time to search and learn how to fix the issues I found
- I ensured that my cleaning was under the specifications (**Define, Code, Test**)
- I've enjoyed so much about Pattern Matching with Regular Expressions and really was thrilled when I generated the dog names it was very exciting
- Also the part which I found it not done in a proper way was the melting the variables into one column, I tried my way but I believe there is a better way to achieve that

All my cleaning process is detailed in the jupyter files I hope you can understand it and help me improving my skills