Wrangle Report

1 Introduction

The dataset consist of three sources, first source from tweets archive of Twitter user @dog_rates (WeRateDogs), is a pre wrangled and cleaned data provided by UDACITY. The other is image prediction file with dogs breeds classification. The third source is a tweeter archive as json file, also provided by UDACITY for students that Tweeter API access wasn't not approved.

File list:

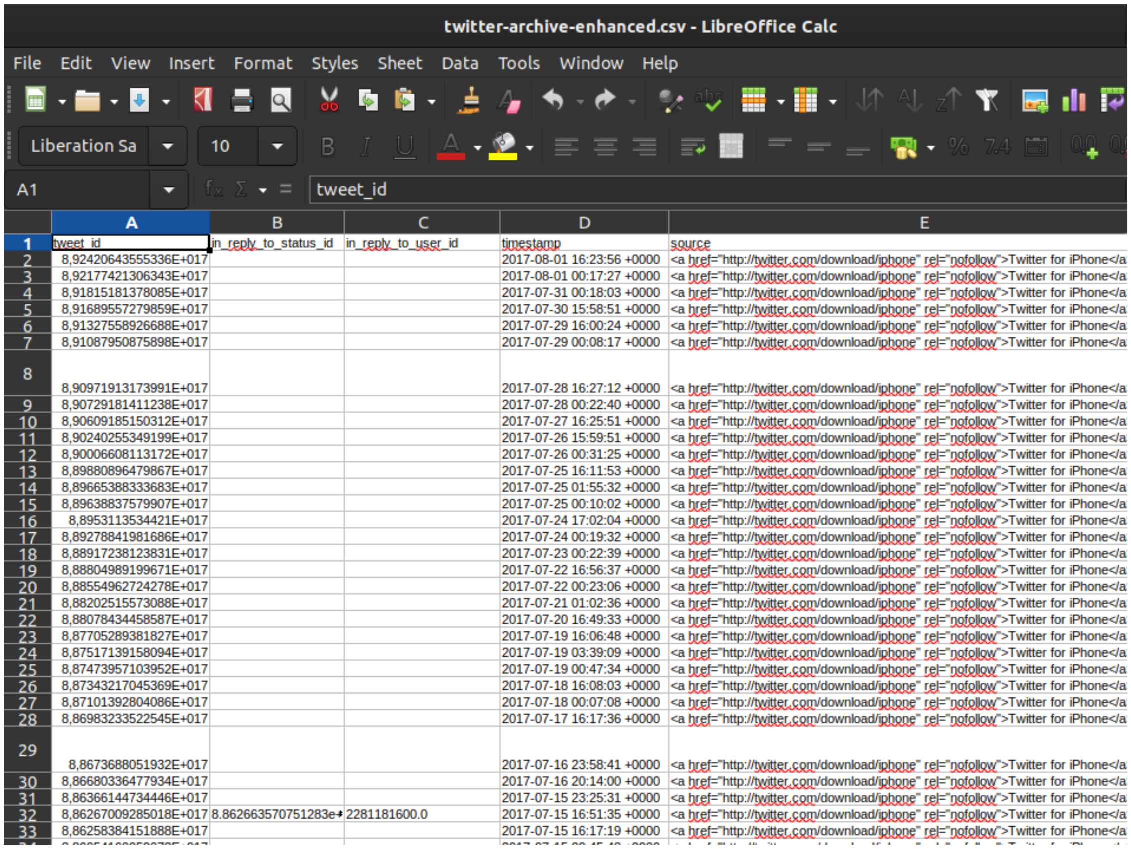
1. twitter-archive-enhanced.csv
2. image-predictions.tsv
3. tweet-jason.txt

# Wrangling Process

1.1 Preliminary data visualization and inspection.

Searching for gross errors, Datasets were first inspected using a spreadsheet software and text editors.

1.1.1 twitter-archive-enhanced.csv



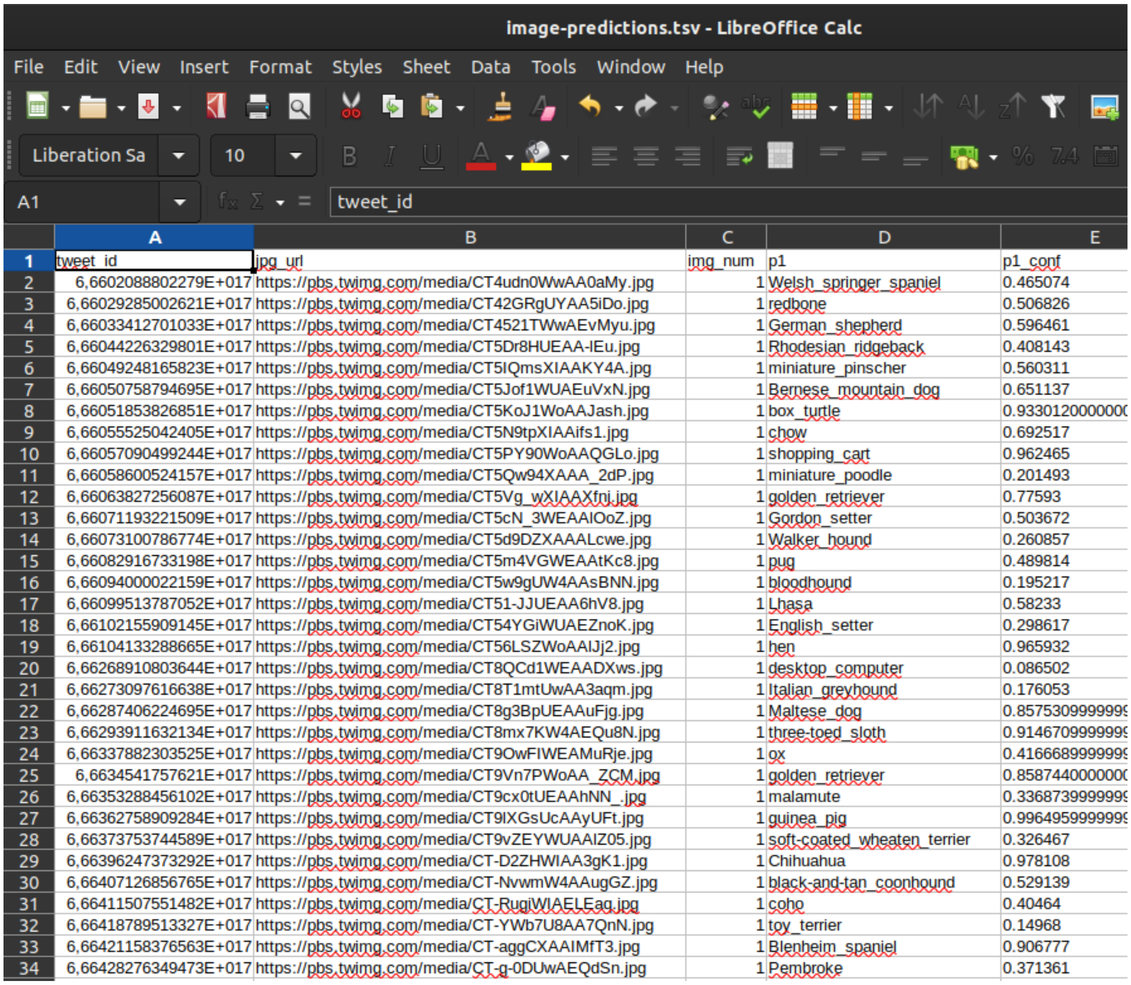
	A	B	C	D	E
1	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source
2	8,92420643555336E+017			2017-08-01 16:23:56 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
3	8,92177421306343E+017			2017-08-01 00:17:27 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
4	8,91815181378085E+017			2017-07-31 00:18:03 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
5	8,91689557279859E+017			2017-07-30 15:58:51 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
6	8,91327558926688E+017			2017-07-29 16:00:24 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
7	8,91087950875898E+017			2017-07-29 00:08:17 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
8	8,90971913173991E+017			2017-07-28 16:27:12 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
9	8,90729181411238E+017			2017-07-28 00:22:40 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
10	8,90609185150312E+017			2017-07-27 16:25:51 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
11	8,90240255349199E+017			2017-07-26 15:59:51 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
12	8,90006608113172E+017			2017-07-26 00:31:25 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
13	8,89880896479867E+017			2017-07-25 16:11:53 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
14	8,89665388333683E+017			2017-07-25 01:55:32 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
15	8,89638837579907E+017			2017-07-25 00:10:02 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
16	8,8953113534421E+017			2017-07-24 17:02:04 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
17	8,89278841981686E+017			2017-07-24 00:19:32 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
18	8,88917238123831E+017			2017-07-23 00:22:39 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
19	8,88804989199671E+017			2017-07-22 16:56:37 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
20	8,88554962724278E+017			2017-07-22 00:23:06 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
21	8,88202515573088E+017			2017-07-21 01:02:36 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
22	8,88078434458587E+017			2017-07-20 16:49:33 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
23	8,87705289381827E+017			2017-07-19 16:06:48 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
24	8,87517139158094E+017			2017-07-19 03:39:09 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
25	8,87473957103952E+017			2017-07-19 00:47:34 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
26	8,87343217045369E+017			2017-07-18 16:08:03 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
27	8,87101392804086E+017			2017-07-18 00:07:08 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
28	8,86983233522545E+017			2017-07-17 16:17:36 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
29	8,8673688051932E+017			2017-07-16 23:58:41 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
30	8,86680336477934E+017			2017-07-16 20:14:00 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
31	8,86366144734446E+017			2017-07-15 23:25:31 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
32	8,86267009285018E+017	8.862663570751283e+	2281181600.0	2017-07-15 16:51:35 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
33	8,86258384151888E+017			2017-07-15 16:17:19 +0000	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a

figure 01 - Preliminary visualization twitter archive enhanced.

Findings:

- found dog names as “a”
- Missing Values Tagged as None and Nan
- consistent with comman separated file format

1.1.2 image-predictions.tsv



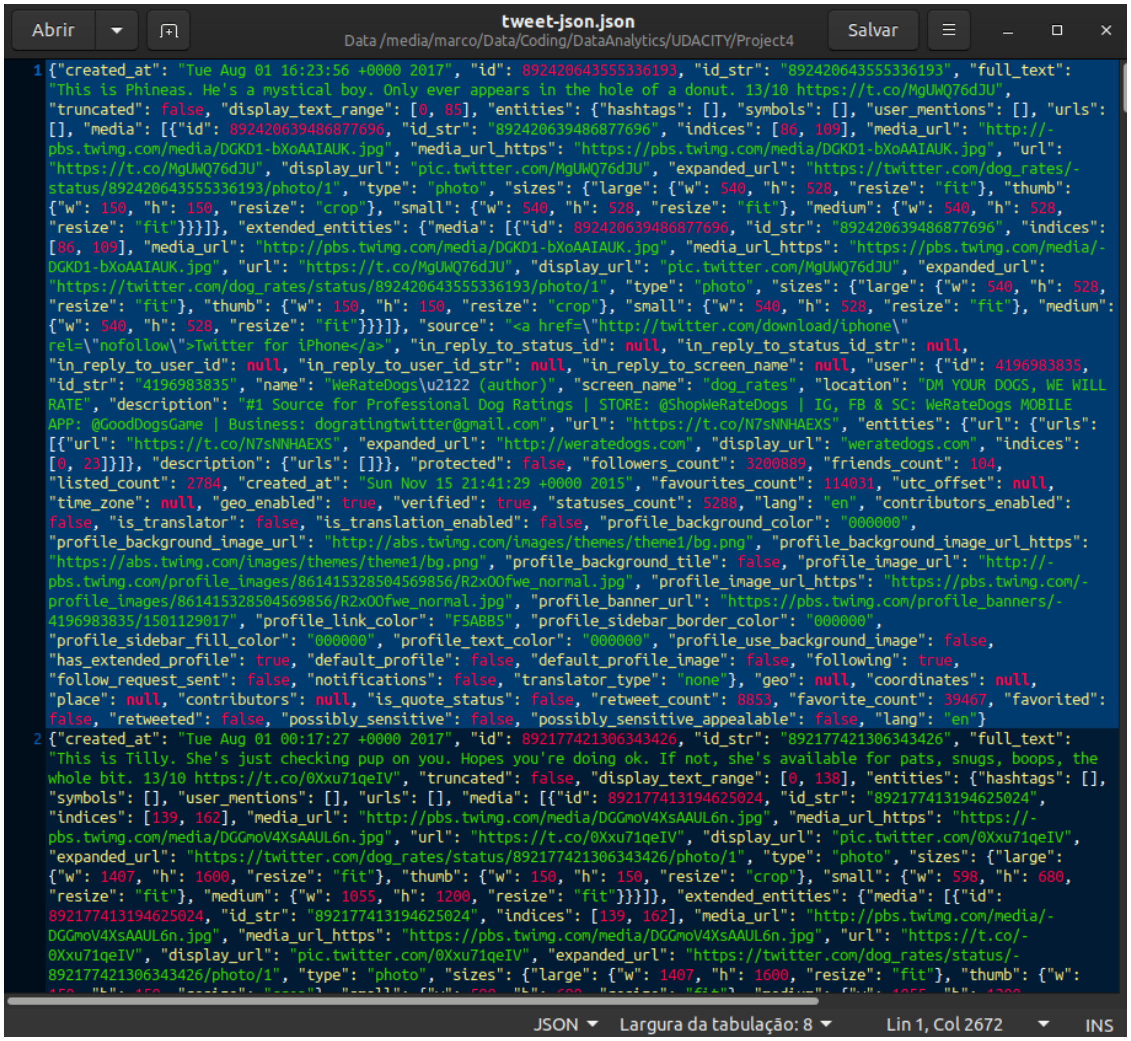
	A	B	C	D	E
1	tweet_id	jpg_url	img_num	p1	p1_conf
2	6,6602088802279E+017	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh_springer_spaniel	0.465074
3	6,66029285002621E+017	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	redbone	0.506826
4	6,66033412701033E+017	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	German_shepherd	0.596461
5	6,66044226329801E+017	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	Rhodesian_ridgeback	0.408143
6	6,66049248165823E+017	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	miniature_pinscher	0.560311
7	6,66050758794695E+017	https://pbs.twimg.com/media/CT5Jof1WUAEuVxN.jpg	1	Bernese_mountain_dog	0.651137
8	6,66051853826851E+017	https://pbs.twimg.com/media/CT5KoJ1WoAAJash.jpg	1	box_turtle	0.933012000000
9	6,66055525042405E+017	https://pbs.twimg.com/media/CT5N9tpXIAAifs1.jpg	1	chow	0.692517
10	6,66057090499244E+017	https://pbs.twimg.com/media/CT5PY90WoAAQGLo.jpg	1	shopping_cart	0.962465
11	6,66058600524157E+017	https://pbs.twimg.com/media/CT5Qw94XAAA_2dP.jpg	1	miniature_poodle	0.201493
12	6,66063827256087E+017	https://pbs.twimg.com/media/CT5Vg_wXIAAXfnj.jpg	1	golden_retriever	0.77593
13	6,66071193221509E+017	https://pbs.twimg.com/media/CT5cN_3WEAAlOoZ.jpg	1	Gordon_setter	0.503672
14	6,66073100786774E+017	https://pbs.twimg.com/media/CT5d9DZXAAALcwe.jpg	1	Walker_hound	0.260857
15	6,66082916733198E+017	https://pbs.twimg.com/media/CT5m4VGWEAAtKc8.jpg	1	pug	0.489814
16	6,66094000022159E+017	https://pbs.twimg.com/media/CT5w9gUW4AAsBNN.jpg	1	bloodhound	0.195217
17	6,66099513787052E+017	https://pbs.twimg.com/media/CT51-JJUEAA6hV8.jpg	1	Lhasa	0.58233
18	6,66102155909145E+017	https://pbs.twimg.com/media/CT54YGiWUAEZnoK.jpg	1	English_setter	0.298617
19	6,66104133288665E+017	https://pbs.twimg.com/media/CT56LSZWoAAlJj2.jpg	1	hen	0.965932
20	6,66268910803644E+017	https://pbs.twimg.com/media/CT8QCd1WEAADXws.jpg	1	desktop_computer	0.086502
21	6,66273097616638E+017	https://pbs.twimg.com/media/CT8T1mtUwAA3aqm.jpg	1	Italian_greyhound	0.176053
22	6,66287406224695E+017	https://pbs.twimg.com/media/CT8g3BpUEAAuFjg.jpg	1	Maltese_dog	0.857530999999
23	6,66293911632134E+017	https://pbs.twimg.com/media/CT8mx7KW4AEQu8N.jpg	1	three-toed_sloth	0.914670999999
24	6,66337882303525E+017	https://pbs.twimg.com/media/CT9OwFIWEAMuRje.jpg	1	ox	0.416668999999
25	6,6634541757621E+017	https://pbs.twimg.com/media/CT9Vn7PWoAA_ZCM.jpg	1	golden_retriever	0.858744000000
26	6,66353288456102E+017	https://pbs.twimg.com/media/CT9cx0tUEAAhNN_.jpg	1	malamute	0.336873999999
27	6,66362758909284E+017	https://pbs.twimg.com/media/CT9lXGsUcAAyUFt.jpg	1	guinea_pig	0.996495999999
28	6,66373753744589E+017	https://pbs.twimg.com/media/CT9vZEYWUAAlZ05.jpg	1	soft-coated_wheaten_terrier	0.326467
29	6,66396247373292E+017	https://pbs.twimg.com/media/CT-D2ZHWIAA3gK1.jpg	1	Chihuahua	0.978108
30	6,66407126856765E+017	https://pbs.twimg.com/media/CT-NvwmW4AAugGZ.jpg	1	black-and-tan_coonhound	0.529139
31	6,66411507551482E+017	https://pbs.twimg.com/media/CT-RugiWIAELEaq.jpg	1	coho	0.40464
32	6,66418789513327E+017	https://pbs.twimg.com/media/CT-YWb7U8AA7QnN.jpg	1	toy_terrier	0.14968
33	6,66421158376563E+017	https://pbs.twimg.com/media/CT-aggCXAAIMfT3.jpg	1	Blenheim_spaniel	0.906777
34	6,66428276349473E+017	https://pbs.twimg.com/media/CT-g-0DUwAEQdSn.jpg	1	Pembroke	0.371361

figure 02 - Preliminary visualization image prediction file.

Findings:

- file in a good shape
- consistent with tab separated file format

1.1.3 tweer-json.txt



```
{"created_at": "Tue Aug 01 16:23:56 +0000 2017", "id": 892420643555336193, "id_str": "892420643555336193", "full_text":
"This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10 https://t.co/MgUWQ76dJU",
"truncated": false, "display_text_range": [0, 85], "entities": {"hashtags": [], "symbols": [], "user_mentions": [], "urls":
[], "media": [{"id": 892420639486877696, "id_str": "892420639486877696", "indices": [86, 109], "media_url": "http://-
pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg", "media_url_https": "https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg", "url":
"https://t.co/MgUWQ76dJU", "display_url": "pic.twitter.com/MgUWQ76dJU", "expanded_url": "https://twitter.com/dog_rates/-
status/892420643555336193/photo/1", "type": "photo", "sizes": {"large": {"w": 540, "h": 528, "resize": "fit"}, "thumb":
{"w": 150, "h": 150, "resize": "crop"}, "small": {"w": 540, "h": 528, "resize": "fit"}, "medium": {"w": 540, "h": 528,
"resize": "fit"}}}]}, "extended_entities": {"media": [{"id": 892420639486877696, "id_str": "892420639486877696", "indices":
[86, 109], "media_url": "http://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg", "media_url_https": "https://pbs.twimg.com/media/-
DGKD1-bXoAAIAUK.jpg", "url": "https://t.co/MgUWQ76dJU", "display_url": "pic.twitter.com/MgUWQ76dJU", "expanded_url":
"https://twitter.com/dog_rates/status/892420643555336193/photo/1", "type": "photo", "sizes": {"large": {"w": 540, "h": 528,
"resize": "fit"}, "thumb": {"w": 150, "h": 150, "resize": "crop"}, "small": {"w": 540, "h": 528, "resize": "fit"}, "medium":
{"w": 540, "h": 528, "resize": "fit"}}}]}, "source": "<a href=\"http://twitter.com/download/iphone\"
rel=\"nofollow\">Twitter for iPhone</a>", "in_reply_to_status_id": null, "in_reply_to_status_id_str": null,
"in_reply_to_user_id": null, "in_reply_to_user_id_str": null, "in_reply_to_screen_name": null, "user": {"id": 4196983835,
"id_str": "4196983835", "name": "WeRateDogs\u2122 (author)", "screen_name": "dog_rates", "location": "DM YOUR DOGS, WE WILL
RATE", "description": "#1 Source for Professional Dog Ratings | STORE: @ShopWeRateDogs | IG, FB & SC: WeRateDogs MOBILE
APP: @GoodDogsGame | Business: dogratingtwitter@gmail.com", "url": "https://t.co/N7sNNHAEXS", "entities": {"url": {"urls":
[{"url": "https://t.co/N7sNNHAEXS", "expanded_url": "http://weratedogs.com", "display_url": "weratedogs.com", "indices":
[0, 23]}]}, "description": {"urls": []}}, "protected": false, "followers_count": 3200889, "friends_count": 104,
"listed_count": 2784, "created_at": "Sun Nov 15 21:41:29 +0000 2015", "favourites_count": 114031, "utc_offset": null,
"time_zone": null, "geo_enabled": true, "verified": true, "statuses_count": 5288, "lang": "en", "contributors_enabled":
false, "is_translator": false, "is_translation_enabled": false, "profile_background_color": "000000",
"profile_background_image_url": "http://abs.twimg.com/images/themes/theme1/bg.png", "profile_background_image_url_https":
"https://abs.twimg.com/images/themes/theme1/bg.png", "profile_background_tile": false, "profile_image_url": "http://-
pbs.twimg.com/profile_images/861415328504569856/R2xOOfwe_normal.jpg", "profile_image_url_https": "https://pbs.twimg.com/-
profile_images/861415328504569856/R2xOOfwe_normal.jpg", "profile_banner_url": "https://pbs.twimg.com/profile_banners/-
4196983835/1501129017", "profile_link_color": "F5ABB5", "profile_sidebar_border_color": "000000",
"profile_sidebar_fill_color": "000000", "profile_text_color": "000000", "profile_use_background_image": false,
"has_extended_profile": true, "default_profile": false, "default_profile_image": false, "following": true,
"follow_request_sent": false, "notifications": false, "translator_type": "none"}, "geo": null, "coordinates": null,
"place": null, "contributors": null, "is_quote_status": false, "retweet_count": 8853, "favorite_count": 39467, "favorited":
false, "retweeted": false, "possibly_sensitive": false, "possibly_sensitive_appealable": false, "lang": "en"}
{"created_at": "Tue Aug 01 00:17:27 +0000 2017", "id": 892177421306343426, "id_str": "892177421306343426", "full_text":
"This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not, she's available for pats, snugs, boops, the
whole bit. 13/10 https://t.co/0Xxu71qeIV", "truncated": false, "display_text_range": [0, 138], "entities": {"hashtags": [],
"symbols": [], "user_mentions": [], "urls": [], "media": [{"id": 892177413194625024, "id_str": "892177413194625024",
"indices": [139, 162], "media_url": "http://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg", "media_url_https": "https://-
pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg", "url": "https://t.co/0Xxu71qeIV", "display_url": "pic.twitter.com/0Xxu71qeIV",
"expanded_url": "https://twitter.com/dog_rates/status/892177421306343426/photo/1", "type": "photo", "sizes": {"large":
{"w": 1407, "h": 1600, "resize": "fit"}, "thumb": {"w": 150, "h": 150, "resize": "crop"}, "small": {"w": 598, "h": 680,
"resize": "fit"}, "medium": {"w": 1055, "h": 1200, "resize": "fit"}}}]}, "extended_entities": {"media": [{"id":
892177413194625024, "id_str": "892177413194625024", "indices": [139, 162], "media_url": "http://pbs.twimg.com/media/-
DGGmoV4XsAAUL6n.jpg", "media_url_https": "https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg", "url": "https://t.co/-
0Xxu71qeIV", "display_url": "pic.twitter.com/0Xxu71qeIV", "expanded_url": "https://twitter.com/dog_rates/status/-
892177421306343426/photo/1", "type": "photo", "sizes": {"large": {"w": 1407, "h": 1600, "resize": "fit"}, "thumb": {"w":
```

figure 03 - Preliminary visualization tweeter jason file.

Findings:

- file in a good shape
- consistent with jason file format

1.2 Data Gathering with Pandas

The three files were loaded into Pandas dataframe without any occurrence.

1. twitter-archive-enhanced.csv -> df1
2. image-predictions.tsv -> df2
3. tweet-jason.txt -> df3

1.3 Data Assessing using Pandas

Dataframe where programmatic assessed and inspected using Pandas functions and results with additional findings for data quality and tidiness issues.

1.3.1 Findings

General issues

1. Check is all df3(json file) is in df1

Quality issues

1. Wrong Data types: (df1)

Column	type
in_reply_to_status_id	float64
in_reply_to_user_id	float64
timestamp	object
retweeted_status_id	float64
retweeted_status_user_id	float64
retweeted_status_timestamp	object

2.Lines with "Stage of" Dogs with more than one classification

3. Dog names column as 'None' (string) istead of Nan (null object) for missing Data
4. Retweets rows
5. In reply rows
6. Error getting the rate numbers like 5 instead 13.5, 75 instead 9.75, etc
7. Tweets with "This is a —" geting Dog name as "a"

8.Not necessary Columns (df1) - source - in_reply_to_status_id - in_reply_to_user_id - retweeted_status_id - retweeted_status_user_id - retweeted_status_timestamp

9. Remove unnecessary columns 'doggo', 'floofer', 'pupper', 'puppo', 'stg_count'
10. Missing values handling

1.3.2 Tidiness issues

1. Text Column with text and pictures URL (df1)
2. Stage of dogs in Columns (df1)

1.4 Cleaning Data

1.4.1 General Issue 01

Dataframe df3 where found entire at the df1. So df2 where left out of merging.

1.4.2 Quality issue - 01

Types where fixed for appropriated.

1.4.3 Quality issue - 02

Rows with more than one dog stage were removed

1.4.4 Quality issue - 03

Normalized all missing value representation to np.Nan for later missing value metrics and handling.

1.4.5 Quality issue - 04

Retweets rows dropped.

1.4.6 Quality issue - 05

In reply rows dropped

1.4.7 Quality issue - 06

Fixed number errors with proper rounding.

1.4.8 Quality issue - 07

Name column entry equal ‘a’ and changed for NaN.

1.4.9 Quality issue - 08

Dropped unnecessary columns.

1.4.10 Quality issue - 09

Dropped unnecessary columns after fixing tidiness issues.

1.4.11 Quality issue - 10

Filled all missing values with ‘Unkown’

1.4.12 Tidiness issue - 01

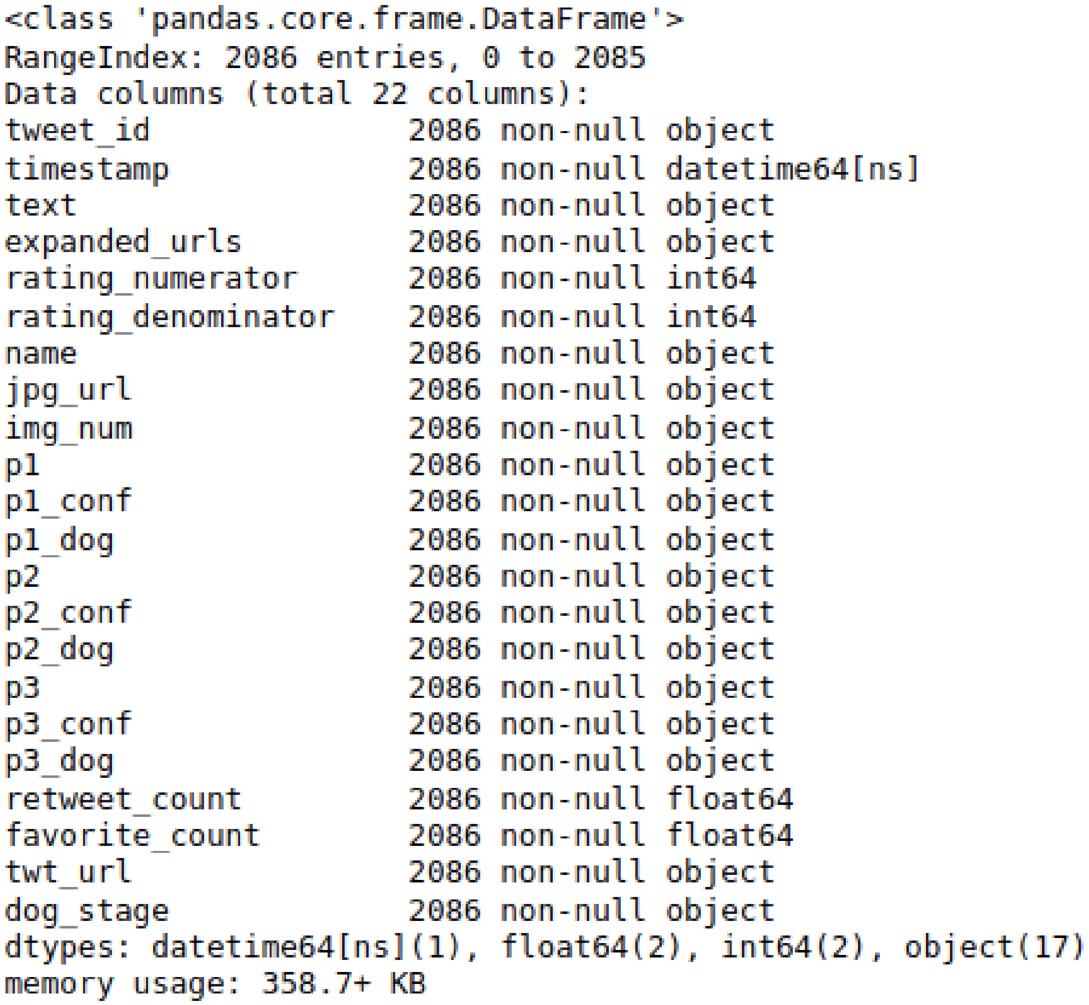
1.4.13 Tidiness issue - 02

Created new Dog Stage column with correspondent stage.

1.5 Storing final Result.

Final Data Set is stored in csv format as “twitter_archive_master.csv”.

With file md5sum: $ddd4689a510728b931a0ad4640b6206e$



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2086 entries, 0 to 2085
Data columns (total 22 columns):
tweet_id              2086 non-null object
timestamp             2086 non-null datetime64[ns]
text                  2086 non-null object
expanded_urls         2086 non-null object
rating_numerator      2086 non-null int64
rating_denominator    2086 non-null int64
name                  2086 non-null object
jpg_url               2086 non-null object
img_num               2086 non-null object
p1                    2086 non-null object
p1_conf               2086 non-null object
p1_dog                2086 non-null object
p2                    2086 non-null object
p2_conf               2086 non-null object
p2_dog                2086 non-null object
p3                    2086 non-null object
p3_conf               2086 non-null object
p3_dog                2086 non-null object
retweet_count         2086 non-null float64
favorite_count        2086 non-null float64
twt_url               2086 non-null object
dog_stage             2086 non-null object
dtypes: datetime64[ns](1), float64(2), int64(2), object(17)
memory usage: 358.7+ KB
```

figure 04 - Final data frame info.