收集

In [1]:

```
import pandas as pd
```

In [2]:

```
# 将原始文件读取为数据框
df_archive = pd.read_csv('twitter-archive-enhanced.csv')
df_image = pd.read_csv('image-predictions.tsv', sep='\t')
```

In [3]:

```
# 原读法：
# df_tweet = pd.read_json('tweet_json.txt', lines=True, dtype=False)
# 审阅：直接把它读取进dataframe其实可以看出有些列中每个值本身是个字典，这样数据本身就存在结构性的问题，
# 这里我们可以按照如下步骤来读取
# 1，打开文件（如with...open
# 2， 申明一个dataframe/list
# 3， 遍历文件中所有数据，然后一条一条append到dataframe/存成字典append到list
# 4， 如果你是申明list，在最后把list中的数据传到dataframe

# 疑惑：原数据结构嵌套太多，感觉只能这样了，如果要让每列的值都纯净，好像只能把原数据大卸八块后读入列名
# 被重命名的数据框，因为像user里的id这种不重命名不行，不然和第一层级的id冲突
# 以下是新读法，感觉没啥意义
import json
with open('tweet_json.txt') as f:
    tweet = []
    for line in f:
        tweet.append(json.loads(line))

df_tweet = pd.DataFrame(tweet, columns=['id', 'retweet_count', 'favorite_count'])
```

评估

In [4]:

```
df_archive
```

Out[4]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	<a href r...
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a href r...
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a href r...
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a href r...
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<a href r...
5	891087950875897856	NaN	NaN	2017-07-29 00:08:17 +0000	<a href r...
6	890971913173991426	NaN	NaN	2017-07-28 16:27:12 +0000	<a href r...
7	890729181411237888	NaN	NaN	2017-07-28 00:22:40 +0000	<a href r...
8	890609185150312448	NaN	NaN	2017-07-27 16:25:51 +0000	<a href r...
9	890240255349198849	NaN	NaN	2017-07-26 15:59:51 +0000	<a href r...
10	890006608113172480	NaN	NaN	2017-07-26 00:31:25 +0000	<a href r...

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
11	889880896479866881	NaN	NaN	2017-07-25 16:11:53 +0000	<a href r...
12	889665388333682689	NaN	NaN	2017-07-25 01:55:32 +0000	<a href r...
13	889638837579907072	NaN	NaN	2017-07-25 00:10:02 +0000	<a href r...
14	889531135344209921	NaN	NaN	2017-07-24 17:02:04 +0000	<a href r...
15	889278841981685760	NaN	NaN	2017-07-24 00:19:32 +0000	<a href r...
16	888917238123831296	NaN	NaN	2017-07-23 00:22:39 +0000	<a href r...
17	888804989199671297	NaN	NaN	2017-07-22 16:56:37 +0000	<a href r...
18	888554962724278272	NaN	NaN	2017-07-22 00:23:06 +0000	<a href r...
19	888202515573088257	NaN	NaN	2017-07-21 01:02:36 +0000	<a href r...
20	887978434458587136	NaN	NaN	2017-07-20 16:49:33 +0000	<a href r...
21	887705289381826560	NaN	NaN	2017-07-19 16:06:48 +0000	<a href r...

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
22	887517139158093824	NaN	NaN	2017-07-19 03:39:09 +0000	<a href r...
23	887473957103951883	NaN	NaN	2017-07-19 00:47:34 +0000	<a href r...
24	887343217045368832	NaN	NaN	2017-07-18 16:08:03 +0000	<a href r...
25	887101392804085760	NaN	NaN	2017-07-18 00:07:08 +0000	<a href r...
26	886983233522544640	NaN	NaN	2017-07-17 16:17:36 +0000	<a href r...
27	886736880519319552	NaN	NaN	2017-07-16 23:58:41 +0000	<a href r...
28	886680336477933568	NaN	NaN	2017-07-16 20:14:00 +0000	<a href r...
29	886366144734445568	NaN	NaN	2017-07-15 23:25:31 +0000	<a href r...
...	...	...	...	...	...
2326	666411507551481857	NaN	NaN	2015-11-17 00:24:19 +0000	<a href r...
2327	666407126856765440	NaN	NaN	2015-11-17 00:06:54 +0000	<a href r...
2328	666396247373291520	NaN	NaN	2015-11-16 23:23:41	<a href r...

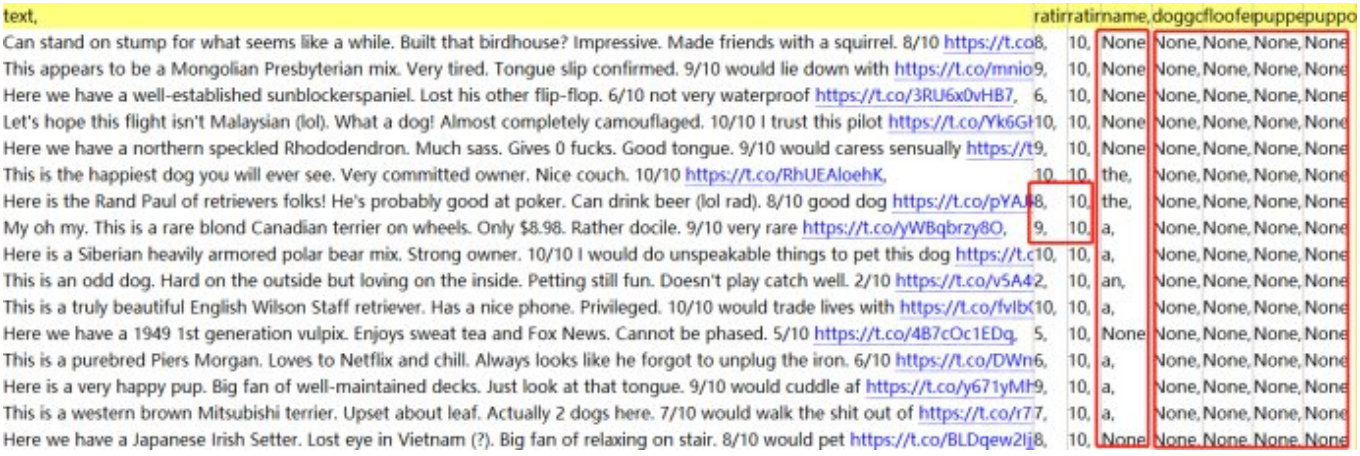
				+0000	...

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
2329	666373753744588802	NaN	NaN	2015-11-16 21:54:18 +0000	<a href r...
2330	666362758909284353	NaN	NaN	2015-11-16 21:10:36 +0000	<a href r...
2331	666353288456101888	NaN	NaN	2015-11-16 20:32:58 +0000	<a href r...
2332	666345417576210432	NaN	NaN	2015-11-16 20:01:42 +0000	<a href r...
2333	666337882303524864	NaN	NaN	2015-11-16 19:31:45 +0000	<a href r...
2334	666293911632134144	NaN	NaN	2015-11-16 16:37:02 +0000	<a href r...
2335	666287406224695296	NaN	NaN	2015-11-16 16:11:11 +0000	<a href r...
2336	666273097616637952	NaN	NaN	2015-11-16 15:14:19 +0000	<a href r...
2337	666268910803644416	NaN	NaN	2015-11-16 14:57:41 +0000	<a href r...
2338	666104133288665088	NaN	NaN	2015-11-16 04:02:55 +0000	<a href r...
2339	666102155909144576	NaN	NaN	2015-11-16 03:55:04 +0000	<a href r...

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
2340	666099513787052032	NaN	NaN	2015-11-16 03:44:34 +0000	<a href r...
2341	666094000022159362	NaN	NaN	2015-11-16 03:22:39 +0000	<a href r...
2342	666082916733198337	NaN	NaN	2015-11-16 02:38:37 +0000	<a href r...
2343	666073100786774016	NaN	NaN	2015-11-16 01:59:36 +0000	<a href r...
2344	666071193221509120	NaN	NaN	2015-11-16 01:52:02 +0000	<a href r...
2345	666063827256086533	NaN	NaN	2015-11-16 01:22:45 +0000	<a href r...
2346	666058600524156928	NaN	NaN	2015-11-16 01:01:59 +0000	<a href r...
2347	666057090499244032	NaN	NaN	2015-11-16 00:55:59 +0000	<a href r...
2348	666055525042405380	NaN	NaN	2015-11-16 00:49:46 +0000	<a href r...
2349	666051853826850816	NaN	NaN	2015-11-16 00:35:11 +0000	<a href r...
2350	666050758794694657	NaN	NaN	2015-11-16 00:30:50 +0000	<a href r...

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
2351	666049248165822465	NaN	NaN	2015-11-16 00:24:50 +0000	<a href r...
2352	666044226329800704	NaN	NaN	2015-11-16 00:04:52 +0000	<a href r...
2353	666033412701032449	NaN	NaN	2015-11-15 23:21:54 +0000	<a href r...
2354	666029285002620928	NaN	NaN	2015-11-15 23:05:30 +0000	<a href r...
2355	666020888022790149	NaN	NaN	2015-11-15 22:32:08 +0000	<a href r...

2356 rows × 17 columns



text,	rati	rati	rname,	doggo	floofe	puppe	puppo
Can stand on stump for what seems like a while. Built that birdhouse? Impressive. Made friends with a squirrel. 8/10 https://t.co	8,	10,	None	None,	None,	None,	None
This appears to be a Mongolian Presbyterian mix. Very tired. Tongue slip confirmed. 9/10 would lie down with https://t.co/mnio	9,	10,	None	None,	None,	None,	None
Here we have a well-established sunblockerspaniel. Lost his other flip-flop. 6/10 not very waterproof https://t.co/3RU6x0vHB7,	6,	10,	None	None,	None,	None,	None
Let's hope this flight isn't Malaysian (lol). What a dog! Almost completely camouflaged. 10/10 I trust this pilot https://t.co/Yk6GI	10,	10,	None	None,	None,	None,	None
Here we have a northern speckled Rhododendron. Much sass. Gives 0 fucks. Good tongue. 9/10 would caress sensually https://t	9,	10,	None	None,	None,	None,	None
This is the happiest dog you will ever see. Very committed owner. Nice couch. 10/10 https://t.co/RhUEAloehK,	10	10,	the,	None,	None,	None,	None
Here is the Rand Paul of retrievers folks! He's probably good at poker. Can drink beer (lol rad). 8/10 good dog https://t.co/pYA	8,	10,	the,	None,	None,	None,	None
My oh my. This is a rare blond Canadian terrier on wheels. Only $8.98. Rather docile. 9/10 very rare https://t.co/yWBqbrzy8O,	9,	10,	a,	None,	None,	None,	None
Here is a Siberian heavily armored polar bear mix. Strong owner. 10/10 I would do unspeakable things to pet this dog https://t.c	10,	10,	a,	None,	None,	None,	None
This is an odd dog. Hard on the outside but loving on the inside. Petting still fun. Doesn't play catch well. 2/10 https://t.co/v5A4	2,	10,	an,	None,	None,	None,	None
This is a truly beautiful English Wilson Staff retriever. Has a nice phone. Privileged. 10/10 would trade lives with https://t.co/fvIb	10,	10,	a,	None,	None,	None,	None
Here we have a 1949 1st generation vulpix. Enjoys sweat tea and Fox News. Cannot be phased. 5/10 https://t.co/4B7cOc1EDq,	5,	10,	None	None,	None,	None,	None
This is a purebred Piers Morgan. Loves to Netflix and chill. Always looks like he forgot to unplug the iron. 6/10 https://t.co/DWn	6,	10,	a,	None,	None,	None,	None
Here is a very happy pup. Big fan of well-maintained decks. Just look at that tongue. 9/10 would cuddle af https://t.co/y671yMH	9,	10,	a,	None,	None,	None,	None
This is a western brown Mitsubishi terrier. Upset about leaf. Actually 2 dogs here. 7/10 would walk the shit out of https://t.co/r7	7,	10,	a,	None,	None,	None,	None
Here we have a Japanese Irish Setter. Lost eye in Vietnam (?). Big fan of relaxing on stair. 8/10 would pet https://t.co/BLDqew2Ij	8,	10,	None	None,	None,	None,	None

In [5]:

```
df_tweet
```

Out[5]:

	id	retweet_count	favorite_count
0	892420643555336193	8842	39492
1	892177421306343426	6480	33786
2	891815181378084864	4301	25445
3	891689557279858688	8925	42863
4	891327558926688256	9721	41016
5	891087950875897856	3240	20548
6	890971913173991426	2142	12053
7	890729181411237888	19548	66596
8	890609185150312448	4403	28187
9	890240255349198849	7684	32467
10	890006608113172480	7584	31127
11	889880896479866881	5116	28208
12	889665388333682689	8502	38745
13	889638837579907072	4705	27633
14	889531135344209921	2309	15329
15	889278841981685760	5635	25712
16	888917238123831296	4681	29555
17	888804989199671297	4535	26021
18	888554962724278272	3722	20267
19	888078434458587136	3637	22144
20	887705289381826560	5584	30690
21	887517139158093824	12053	46940
22	887473957103951883	18813	70007
23	887343217045368832	10713	34223
24	887101392804085760	6147	31045
25	886983233522544640	8045	35786
26	886736880519319552	3420	12286
27	886680336477933568	4597	22802
28	886366144734445568	3297	21488
29	886267009285017600	4	117
...	...	...	...
2322	666411507551481857	337	457
2323	666407126856765440	43	113

2324	666396247373291520	91	171

	id	retweet_count	favorite_count
2325	666373753744588802	99	194
2326	666362758909284353	590	801
2327	666353288456101888	76	228
2328	666345417576210432	146	308
2329	666337882303524864	96	203
2330	666293911632134144	365	519
2331	666287406224695296	71	152
2332	666273097616637952	81	183
2333	666268910803644416	37	108
2334	666104133288665088	6835	14703
2335	666102155909144576	15	81
2336	666099513787052032	73	160
2337	666094000022159362	78	168
2338	666082916733198337	47	121
2339	666073100786774016	173	334
2340	666071193221509120	67	154
2341	666063827256086533	230	494
2342	666058600524156928	61	117
2343	666057090499244032	146	304
2344	666055525042405380	261	449
2345	666051853826850816	877	1250
2346	666050758794694657	60	136
2347	666049248165822465	41	111
2348	666044226329800704	147	309
2349	666033412701032449	47	128
2350	666029285002620928	48	132
2351	666020888022790149	530	2528

2352 rows × 3 columns

In [6]:

```
df_image
```

Out[6]:

	tweet_id	jpg_url	img_
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1
5	666050758794694657	https://pbs.twimg.com/media/CT5Jof1WUAEuVxN.jpg	1
6	666051853826850816	https://pbs.twimg.com/media/CT5KoJ1WoAAJash.jpg	1
7	666055525042405380	https://pbs.twimg.com/media/CT5N9tpXIAAifs1.jpg	1
8	666057090499244032	https://pbs.twimg.com/media/CT5PY90WoAAQGLo.jpg	1
9	666058600524156928	https://pbs.twimg.com/media/CT5Qw94XAAA_2dP.jpg	1
10	666063827256086533	https://pbs.twimg.com/media/CT5Vg_wXIAAXfnj.jpg	1
11	666071193221509120	https://pbs.twimg.com/media/CT5cN_3WEAAlOoZ.jpg	1
12	666073100786774016	https://pbs.twimg.com/media/CT5d9DZXAAALcwe.jpg	1
13	666082916733198337	https://pbs.twimg.com/media/CT5m4VGWEAAtKc8.jpg	1
14	666094000022159362	https://pbs.twimg.com/media/CT5w9gUW4AAsBNN.jpg	1
15	666099513787052032	https://pbs.twimg.com/media/CT51-JJUEAA6hV8.jpg	1
16	666102155909144576	https://pbs.twimg.com/media/CT54YGiWUAEZnoK.jpg	1
17	666104133288665088	https://pbs.twimg.com/media/CT56LSZWoAAlJj2.jpg	1
18	666268910803644416	https://pbs.twimg.com/media/CT8QCd1WEAADXws.jpg	1
19	666273097616637952	https://pbs.twimg.com/media/CT8T1mtUwAA3aqm.jpg	1
20	666287406224695296	https://pbs.twimg.com/media/CT8g3BpUEAAuFjg.jpg	1
21	666293911632134144	https://pbs.twimg.com/media/CT8mx7KW4AEQu8N.jpg	1
22	666337882303524864	https://pbs.twimg.com/media/CT9OwFIWEAMuRje.jpg	1
23	666345417576210432	https://pbs.twimg.com/media/CT9Vn7PWoAA_ZCM.jpg	1
24	666353288456101888	https://pbs.twimg.com/media/CT9cx0tUEAAhNN_.jpg	1
25	666362758909284353	https://pbs.twimg.com/media/CT9lXGsUcAAyUFt.jpg	1
26	666373753744588802	https://pbs.twimg.com/media/CT9vZEYWUAAlZ05.jpg	1
27	666396247373291520	https://pbs.twimg.com/media/CT-D2ZHWIAA3gK1.jpg	1
28	666407126856765440	https://pbs.twimg.com/media/CT-NvwmW4AAugGZ.jpg	1
29	666411507551481857	https://pbs.twimg.com/media/CT-RugiWIAELEaq.jpg	1
...	...	...	...
2045	886366144734445568	https://pbs.twimg.com/media/DE0BTnQUwAApKEH.jpg	1
2046	886680336477933568	https://pbs.twimg.com/media/DE4fEDzWAAAyHMM.jpg	1

	tweet_id	jpg_url	img_
2047	886736880519319552	https://pbs.twimg.com/media/DE5Se8FXcAAJFx4.jpg	1
2048	886983233522544640	https://pbs.twimg.com/media/DE8yicJW0AAAvBJ.jpg	2
2049	887101392804085760	https://pbs.twimg.com/media/DE-eAq6UwAA-jaE.jpg	1
2050	887343217045368832	https://pbs.twimg.com/ext_tw_video_thumb/88734...	1
2051	887473957103951883	https://pbs.twimg.com/media/DFDw2tyUQAAAFke.jpg	2
2052	887517139158093824	https://pbs.twimg.com/ext_tw_video_thumb/88751...	1
2053	887705289381826560	https://pbs.twimg.com/media/DFHDQBbXgAEqY7t.jpg	1
2054	888078434458587136	https://pbs.twimg.com/media/DFMWn56WsAAkA7B.jpg	1
2055	888202515573088257	https://pbs.twimg.com/media/DFDw2tyUQAAAFke.jpg	2
2056	888554962724278272	https://pbs.twimg.com/media/DFTH_O-UQAACu20.jpg	3
2057	888804989199671297	https://pbs.twimg.com/media/DFWra-3VYAA2piG.jpg	1
2058	888917238123831296	https://pbs.twimg.com/media/DFYRgsOUQAARGhO.jpg	1
2059	889278841981685760	https://pbs.twimg.com/ext_tw_video_thumb/88927...	1
2060	889531135344209921	https://pbs.twimg.com/media/DFg_2PVW0AEHN3p.jpg	1
2061	889638837579907072	https://pbs.twimg.com/media/DFihzFfXsAYGDPR.jpg	1
2062	889665388333682689	https://pbs.twimg.com/media/DFi579UWsAAatzw.jpg	1
2063	889880896479866881	https://pbs.twimg.com/media/DFl99B1WsAITKsg.jpg	1
2064	890006608113172480	https://pbs.twimg.com/media/DFnwSY4WAAAMliS.jpg	1
2065	890240255349198849	https://pbs.twimg.com/media/DFrEyVuW0AAO3t9.jpg	1
2066	890609185150312448	https://pbs.twimg.com/media/DFwUU__XcAEpyXI.jpg	1
2067	890729181411237888	https://pbs.twimg.com/media/DFyBahAVwAAhUTd.jpg	2
2068	890971913173991426	https://pbs.twimg.com/media/DF1eOmZXUAALUcq.jpg	1
2069	891087950875897856	https://pbs.twimg.com/media/DF3HwyEWsAABqE6.jpg	1
2070	891327558926688256	https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg	2
2071	891689557279858688	https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg	1
2072	891815181378084864	https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg	1
2073	892177421306343426	https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg	1
2074	892420643555336193	https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg	1

2075 rows × 12 columns

In [7]:

```
df_archive.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                      2356 non-null int64
in_reply_to_status_id         78 non-null float64
in_reply_to_user_id           78 non-null float64
timestamp                     2356 non-null object
source                        2356 non-null object
text                          2356 non-null object
retweeted_status_id           181 non-null float64
retweeted_status_user_id      181 non-null float64
retweeted_status_timestamp    181 non-null object
expanded_urls                 2297 non-null object
rating_numerator              2356 non-null int64
rating_denominator            2356 non-null int64
name                          2356 non-null object
doggo                         2356 non-null object
floofer                       2356 non-null object
pupper                        2356 non-null object
puppo                         2356 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

In [8]:

```
df_archive.name.value_counts()
```

Out[8]:

```
None        745
a            55
Charlie      12
Cooper       11
Lucy         11
Oliver       11
Lola         10
Penny        10
Tucker       10
Winston       9
Bo            9
the           8
Sadie         8
Toby          7
Buddy         7
Daisy         7
Bailey        7
an            7
Milo          6
Rusty         6
Oscar         6
Jack          6
Koda          6
Stanley       6
Leo           6
Bella         6
Jax           6
Dave          6
Scout         6
George        5
             ...
Maisey        1
Torque        1
Skye          1
Chevy         1
Simba         1
Vince         1
Hero          1
Linus         1
Raphael       1
Cannon        1
Aja           1
Binky         1
Pippin        1
Jeb           1
Herb          1
Ruffles       1
Napolean      1
Sephie        1
Jeremy        1
Darby         1
Mookie        1
Trevith       1
Jiminus       1
Obi           1
Billl         1
Scott         1
Meatball      1
Kota          1
Ron           1
```

Koh 1

Tino 1
Name: name, Length: 957, dtype: int64

In [9]:

```
pd.set_option('max_colwidth',200)
df_archive[df_archive.name=='None'].text
```

Out[9]:

5 Here we have a majestic great white breaching off South Africa's coast. Absolutely h*ckin breathtaking. 13/10 (IG: tucker_marlo) #BarkWeek https://t.co/kQ04fDDRmh

7 When you watch your owner call another dog a good boy but then they turn back to you and say you're a great boy. 13/10 https://t.co/vOnONBcwxq

12 Here's a puppo that seems to be on the fence about something haha no but seriously someone help her. 13/10 https://t.co/BxvuXkOUCm

24 You may not have known you needed to see this today. 13/10 please enjoy (IG: emmylouroo) https://t.co/WZqNqygEyV

25 This... is a Jubilant Antarctic House Bear. We only rate dogs. Please only send dogs. Thank you... 12/10 would suffocate in floof https://t.co/4Ad1jzJSdp

30 @NonWhiteHat @MayhewMayhem omg hello tanner you are a scary good boy 12/10 would pet with extreme caution

32 RT @Athletics: 12/10 #BATP https://t.co/WxwJmvjfxo

35 I have a new hero and his name is Howard. 14/10 https://t.co/gzLHboL7Sk

37 Here we have a corgi undercover as a malamute. Pawbably doing important investigative work. Zero control over tongue happenings. 13/10 https://t.co/44ItaMubBf

41 I present to you, Pup in Hat. Pup in Hat is great for all occasions. Extremely versatile. Compact as h*ck. 14/10 (IG: itselizabethgales) https://t.co/vvBOcC2VdC

42 OMG HE DIDN'T MEAN TO HE WAS JUST TRYING A LITTLE BARKOUR HE'S SUPER SORRY 13/10 WOULD FORGIVE IMMEDIATE https://t.co/uF3pQ8Wubj

47 Please only send dogs. We don't rate mechanics, no matter how h*ckin good. Thank you... 13/10 would sneak a pat https://t.co/Se5fZ9wp5E

55 @roushfenway These are good dogs but 17/10 is an emotional impulse rating. More like 13/10s

59 Ugh not again. We only rate dogs. Please don't send in well-dressed floppy-tongued street penguins. Dogs only please. Thank you... 12/10 https://t.co/WiAMbTkDPf

62 Please don't send in photos without dogs in them. We're not @porch_rates. Insubordinate and churlish. Pretty good porch tho 11/10 https://t.co/HauE8M3Bu4

64 @RealKentMurphy 14/10 confirmed

72 Martha is stunning how h*ckin dare you. 13/10 https://t.co/9uABQXgjwa

78 RT @rachel2195: @dog_rates the boyfriend and his soaking wet pupper h*cking love his new hat 14/10 https://t.co/dJx4Gzc50G

83 I can say with the pupmost confidence that the dogos who assisted with this search are heroic as h*ck. 14/10 for all https://t.co/8yoc1CNTsu

88 You'll get your package when that precious man is done appreciating the pups. 13/10 for everyone https://t.co/PEp4MghzBW

fp4mgnzDw
89 Guys please stop sending pictures without any do
gs in th- oh never mind hello excuse me sir. 12/10 stealthy as h*ck https://t.co/b
rCQoqc8AW
93 I can't believe this keeps happening. This, is a birb taking a
bath. We only rate dogs. Please only send dogs. Thank you... 12/10 https://t.co/pw
Y9PQhtP2
96 We usually don't rate Deck-bound Saskatoon Black Bears, but this one
is h*ckin flawless. Sneaky tongue slip too. 13/10 would hug firmly https://t.co/mN
uMH9400n
99 Here's a very large dog. He has a date later. Politely asked this
water person to check if his breath is bad. 12/10 good to go doggo https://t.co/EM
YIdob1MR
100 Here are my favorite #dogsatpollingstations \nMost voted for a more consis
tent walking schedule and to increase daily pats tenfold. All 13/10 https://t.co/1
7FVM14VZ5
101 RT @loganamnosis: Penelope here is doing me quite
a divertir. Well done, @dog_rates! Loving the pupdate. 14/10, je jouerais de nouve
au. htt…
103 We. Only. Rate. Dogs. Do not send in other things like this fluffy
floor shark clearly ready to attack. Get it together guys... 12/10 https://t.co/BZ
HiKx3FpQ
110
Never doubt a doggo 14/10 https://t.co/A
bBLh2FZCH
112 Real funny guys. Sending in a pic without a dog in it. Hilarious. W
e'll rate the rug tho because it's giving off a very good vibe. 11/10 https://t.c
o/GCD1JccCyi
113
@ComplicitOwl @ShopWeRateDogs >10/10 is reserved
for dogs

...

2294 Extremely intelligent dog here. Has lear
ned to walk like human. Even has his own dog. Very impressive 10/10 https://t.co/0
DvHAMdA4V
2298
After much debate this dog is being upgraded to 10/10. I rep
eat 10/10
2299 Here we have a Hufflepuff. Loves vest. Eyes
wide af. Flaccid tail. Matches carpet. Always a little blurry. 8/10 https://t.co/7
JdgVqDnvR
2301
12/10 gimme now https://t.co/Q
ZAnwgnOMB
2305 My goodness. Very rare dog here. Large. Tail da
ngerous. Kinda fat. Only eats leaves. Doesn't come when called 3/10 https://t.co/x
YGdBrMS9h
2306 These are Peruvian Feldspars. Their names are Cupit a
nd Prencer. Both resemble Rand Paul. Sick outfits 10/10 & 10/10 https://t.co/Z
nEMHBsAs1
2307
12/10 simply brilliant pup https://t.co/V
6ZzG45zzG
2310 Unfamiliar with this breed. Ears pointy af. Wo
n't let go of seashell. Won't eat kibble. Not very fast. Bad dog 2/10 https://t.c
o/EIn5kE1Y1S
2316 Cool dog. Enjoys couch. Low monotone bark. Very
nice kicks. Pisses milk (must be rare). Can't go down stairs. 4/10 https://t.co/vX
MKrJC81s
2320 Here we see a lone northeastern Cumberbatch. Ha

lf ladybug. Only builds with bricks. Very confident with body. 7/10 https://t.co/7LtjBSOGPK
2321 "Can you behave? You're ruining my wedding day"\nDOG: idgaf this flashlight tastes good as hell\n\n10/10 https://t.co/GlFZPzqcEU
2322 Oh boy what a pup! Sunglasses take this one to the next level. Weirdly folds front legs. Pretty big. 6/10 https://t.co/yECbFrSArM
2323 Here we have an Austrian Pulitzer. Collectors edition. Levitates (?). 7/10 would garden with https://t.co/NMQq6HIglK
2324 *internally screaming* 12/10 https://t.co/YMcrXC2Y6R
2328 Oh goodness. A super rare northeast Qdoba kangaroo mix. Massive feet. No pouch (disappointing). Seems alert. 9/10 https://t.co/Dc7bOE8qFE
2329 Those are sunglasses and a jean jacket. 11/10 dog cool af https://t.co/uHXrPkUEyl
2330 Unique dog here. Very small. Lives in container of Frosted Flakes (?). Short legs. Must be rare 6/10 would still pet https://t.co/XMD9CwjEnM
2331 Here we have a mixed Asiago from the Galápagos Islands. Only one ear working. Big fan of marijuana carpet. 8/10 https://t.co/t1tQ5w9aUO
2332 Look at this jokester thinking seat belt laws don't apply to him. Great tongue tho 10/10 https://t.co/VFKG1vxGjB
2336 Can take selfies 11/10 https://t.co/ws2AMaNwPW
2337 Very concerned about fellow dog trapped in computer. 10/10 https://t.co/0yxApIikpk
2338 Not familiar with this breed. No tail (weird). Only 2 legs. Doesn't bark. Surprisingly quick. Shits eggs. 1/10 https://t.co/Asgdc6kuLX
2339 Oh my. Here you are seeing an Adobe Setter giving birth to twins!!! The world is an amazing place. 11/10 https://t.co/11LvqN4WLq
2340 Can stand on stump for what seems like a while. Built that birdhouse? Impressive. Made friends with a squirrel. 8/10 https://t.co/Ri4nMTLq5C
2341 This appears to be a Mongolian Presbyterian mix. Very tired. Tongue slip confirmed. 9/10 would lie down with https://t.co/mnioXo3IfP
2342 Here we have a well-established sunblockerspaniel. Lost his other flip-flop. 6/10 not very waterproof https://t.co/3RU6xOvHB7
2343 Let's hope this flight isn't Malaysian (lol). What a dog! Almost completely camouflaged. 10/10 I trust this pilot https://t.co/Yk6GHE9tOY
2344 Here we have a northern speckled Rhododendron. Much sass. Gives 0 fucks. Good tongue. 9/10 would caress sensually https://t.co/ZoL8kq2XFx
2351 Here we have a 1949 1st generation vulpix. Enjoys sweat tea and Fox News. Cannot be phased. 5/10 https://t.co/4B7cOc1EDq
2355 Here we have a Japanese Irish Setter. Lost eye in Vietnam (?). Big fan of relaxing on stair. 8/10 would pet https://t.co/B

LDqew2Ijj
Name: text, Length: 745, dtype: object

In [10]:

```
df_archive[df_archive.name=='a'].text
```

Out[10]:

```
56      Here is a pupper approaching maximum borkdrive. Zooming at never before se
en speeds. 14/10 paw-inspiring af \n(IG: puffie_the_chow) https://t.co/ghXBIIeQZF
649                                 Here is a perfect example of someone who has th
eir priorities in order. 13/10 for both owner and Forrest https://t.co/LRyMrU7Wfq
801                     Guys this is getting so out of hand. We only rate dogs. Thi
s is a Galapagos Speed Panda. Pls only send dogs... 10/10 https://t.co/8lpAGaZRFn
1002                    This is a mighty rare blue-tailed hammer sherk. Human almos
t lost a limb trying to take these. Be careful guys. 8/10 https://t.co/TGenMeXreW
1004                     Viewer discretion is advised. This is a terrible attack i
n progress. Not even in water (tragic af). 4/10 bad sherk https://t.co/L3UOj14N5R
1017                          This is a carrot. We only rate dogs. Please only send
in dogs. You all really should know this by now ...11/10 https://t.co/9e48aPrBm2
1049                     This is a very rare Great Alaskan Bush Pupper. Hard to st
umble upon without spooking. 12/10 would pet passionately https://t.co/xOBKCdpzaa
1193                    People please. This is a Deadly Mediterranean Plop T-Rex.
We only rate dogs. Only send in dogs. Thanks you... 11/10 https://t.co/2ATDsgHD4n
1207                      This is a taco. We only rate dogs. Please only send in d
ogs. Dogs are what we rate. Not tacos. Thank you... 10/10 https://t.co/cx16xGY8B9
1340                                                Here is a heartbreaking scene of
an incredible pupper being laid to rest. 10/10 RIP pupper https://t.co/81mvJOrGRu
1351                                                                              H
ere is a whole flock of puppers.  60/50 I'll take the lot https://t.co/9dpcw6MdWa
1361                    This is a Butternut Cumberfloof. It's not windy they just
look like that. 11/10 back at it again with the red socks https://t.co/hMjzhdUHaW
1368                         This is a Wild Tuscan Poofwiggle. Careful not to startl
e. Rare tongue slip. One eye magical. 12/10 would def pet https://t.co/4EnShAQjv6
1382                                                    "Pupper is a present to
world. Here is a bow for pupper." 12/10 precious as hell https://t.co/ItSsE92gCW
1499                    This is a rare Arctic Wubberfloof. Unamused by the happeni
ngs. No longer has the appetites. 12/10 would totally hug https://t.co/krvbacIXON
1737                    Guys this really needs to stop. We've been over this way to
o many times. This is a giraffe. We only rate dogs.. 7/10 https://t.co/yavgkHYPOC
1785                                                     This is a dog swingin
g. I really enjoyed it so I hope you all do as well. 11/10 https://t.co/Ozo9KHTRND
1853                    This is a Sizzlin Menorah spaniel from Brooklyn named Wyli
e. Lovable eyes. Chiller as hell. 10/10 and I'm out.. poof https://t.co/7EOAiJXPmI
1854                                                Seriously guys?! Only send in d
ogs. I only rate dogs. This is a baby black bear... 11/10 https://t.co/H7kpabTfLj
1877                      C'mon guys. We've been over this. We only rate dogs. Thi
s is a cow. Please only submit dogs. Thank you..... 9/10 https://t.co/WjcELNEqN2
1878                                           This is a fluffy albino Bacardi Colu
mbia mix. Excellent at the tweets. 11/10 would hug gently https://t.co/diboDRUuEI
1923                                                                  This is a Sag
itariot Baklava mix. Loves her new hat. 11/10 radiant pup https://t.co/Bko5kFJYUU
1941                          This is a heavily opinionated dog. Loves walls. Nobod
y knows how the hair works. Always ready for a kiss. 4/10 https://t.co/dFiaKZ9cDl
1955                    This is a Lofted Aphrodisiac Terrier named Kip. Big fan of
bed n breakfasts. Fits perfectly. 10/10 would pet firmly https://t.co/gKlLpNzI13
1994                                                          This is a baby Rand
Paul. Curls for days. 11/10 would cuddle the hell out of https://t.co/xHXNaPAYRe
2034                      This is a Tuscaloosa Alcatraz named Jacob (Yacōb). Love
s to sit in swing. Stellar tongue. 11/10 look at his feet https://t.co/2IslQ8ZSc7
2066                    This is a Helvetica Listerine named Rufus. This time Rufus
will be ready for the UPS guy. He'll never expect it 9/10 https://t.co/34OhVhMkVr
2116                    This is a Deciduous Trimester mix named Spork. Only 1 ear
works. No seat belt. Incredibly reckless. 9/10 still cute https://t.co/CtuJoLHiDo
2125                      This is a Rich Mahogany Seltzer named Cherokee. Just got
destroyed by a snowball. Isn't very happy about it. 9/10 https://t.co/98ZBi6o4dj
2128                     This is a Speckled Cauliflower Yosemite named Hemry. He's
terrified of intruder dog. Not one bit comfortable. 9/10 https://t.co/vV3Oqjb8jN
```

terrified of intruder dog. Not one bit comfortable. 9/10 https://t.co/yv3Qgjn6IN
2146 This is a spotted Lipitor Rumpelstiltskin named Alphred. He can't wait for the Turkey. 10/10 would pet really well https://t.co/6GUGO7azNX
2153 This is a brave dog. Excellent free climber. Trying to get closer to God. Not very loyal though. Doesn't bark. 5/10 https://t.co/ODnILTr4QM
2161 This is a Coriander Baton Rouge named Alfredo. Loves to cuddle with smaller well-dressed dog. 10/10 would hug lots https://t.co/eCRdwouKCl
2191 This is a Slovakian Helter Skelter Feta named Leroi. Likes to skip on roofs. Good traction. Much balance. 10/10 wow! https://t.co/Dmy2mY2Qj5
2198 This is a wild Toblerone from Papua New Guinea. Mouth always open. Addicted to hay. Acts blind. 7/10 handsome dog https://t.co/IGmVbzO7tZ
2211 Here is a horned dog. Much grace. Can jump over moons (dam!). Paws not soft. Bad at barking. 7/10 can still pet tho https://t.co/2Su7gmsnZm
2218 This is a Birmingham Quagmire named Chuk. Loves to relax and watch the game while sippin on that iced mocha. 10/10 https://t.co/HvNg9JWxFt
2222 Here is a mother dog caring for her pups. Snazzy red mohawk. Doesn't wag tail. Pups look confused. Overall 4/10 https://t.co/YOHe61fO9m
2235 This is a Trans Siberian Kellogg named Alfonso. Huge ass eyeballs. Actually Dobby from Harry Potter. 7/10 https://t.co/XpseHBlAAb
2249 This is a Shotokon Macadamia mix named Cheryl. Sophisticated af. Looks like a disappointed librarian. Shh (lol) 9/10 https://t.co/J4GnJ5Swba
2255 This is a rare Hungarian Pinot named Jessiga. She is either mid-stroke or got stuck in the washing machine. 8/10 https://t.co/ZUOiOKJyqD
2264 This is a southwest Coriander named Klint. Hat looks expensive. Still on house arrest :(\n9/10 https://t.co/IQTOMqDUIe
2273 This is a northern Wahoo named Kohl. He runs this town. Chases tumbleweeds. Draws gun wicked fast. 11/10 legendary https://t.co/J4vn2rOYFk
2287 This is a Dasani Kingfisher from Maine. His name is Daryl. Daryl doesn't like being swallowed by a panda. 8/10 https://t.co/jpaeu6LNmW
2304 This is a curly Ticonderoga named Pepe. No feet. Loves to jet ski. 11/10 would hug until forever https://t.co/cyDfaK8NBc
2311 This is a purebred Bacardi named Octaviath. Can shoot spaghetti out of mouth. 10/10 https://t.co/uEvsGLOFHa
2314 This is a golden Buckminsterfullerene named Johm. Drives trucks. Lumberjack (?). Enjoys wall. 8/10 would hug softly https://t.co/uQbZJM2DQB
2327 This is a southern Vesuvius bumblegruff. Can drive a truck (wow). Made friends with 5 other nifty dogs (neat). 7/10 https://t.co/LopTBkKa8h
2334 This is a funny dog. Weird toes. Won't come down. Loves branch. Refuses to eat his food. Hard to cuddle with. 3/10 https://t.co/IIXisOzta0
2347 My oh my. This is a rare blond Canadian terrier on wheels. Only $8.98. Rather docile. 9/10 very rare https://t.co/yWBqbrzy8O
2348 Here is a Siberian heavily armored polar bear mix. Strong owner. 10/10 I would do unspeakable things to pet this dog https://t.co/rdivxLiqEt
2350 This is a truly beautiful English Wilson Staff retriever. Has a nice phone. Privileged. 10/10 would trade lives with https://t.co/fvIbQfHjIe
2352 This is a purebred Piers Morgan. Loves to Netflix and chill. Always looks like he forgot to unplug the iron. 6/10 https://t.co/DWnyCjf2mx
2353 Here is a very happy pup. Big fan of well-maintained decks. Just look at that tongue. 9/10 would cuddle af https://t.co/y671yMhoiR
2354 This is a western brown Mitsubishi terrier. Upset about leaf. Actually 2 dogs here. 7/10 would walk the shit out of https://t.co/r7mOb2m0UI
Name: text, dtype: object

In [11]:

```
df_archive.rating_numerator.value_counts()
```

Out[11]:

```
12      558
11      464
10      461
13      351
9       158
8       102
7        55
14       54
5        37
6        32
3        19
4        17
1         9
2         9
420       2
0         2
15        2
75        2
80        1
20        1
24        1
26        1
44        1
50        1
60        1
165       1
84        1
88        1
144       1
182       1
143       1
666       1
960       1
1776      1
17        1
27        1
45        1
99        1
121       1
204       1
Name: rating_numerator, dtype: int64
```

In [12]:

```
df_archive[df_archive.rating_numerator==1776].text
```

Out[12]:

```
979    This is Atticus. He's quite simply America af. 1776/10 https://t.co/GRXwMxL
Bkh
Name: text, dtype: object
```

In [13]:

```
df_tweet.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2352 entries, 0 to 2351
Data columns (total 3 columns):
id                2352 non-null int64
retweet_count     2352 non-null int64
favorite_count    2352 non-null int64
dtypes: int64(3)
memory usage: 55.2 KB
```

In [14]:

```
df_image.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2075 non-null int64
jpg_url     2075 non-null object
img_num     2075 non-null int64
p1          2075 non-null object
p1_conf     2075 non-null float64
p1_dog      2075 non-null bool
p2          2075 non-null object
p2_conf     2075 non-null float64
p2_dog      2075 non-null bool
p3          2075 non-null object
p3_conf     2075 non-null float64
p3_dog      2075 non-null bool
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

质量

df_archive ***表格***

- 含有转发的数据
- twitter id应为字符串
- 分子评分有低于10分的
- 分子分母有不合规的评分，如1776的分子评分，及420/10，165/150，143/130等
- 狗名有a，an，the等明显错误，还有大量“None”值
- 狗的地位有空值

df_tweet ***表格***

- id 应为字符串类型

df_image ***表格***

- id应为字符串

清洁度

- df_archive 表格中的狗地位应为一列
- 数据分布在三个表中

清理

In [15]:

```
archive_clean = df_archive.copy()
tweet_clean = df_tweet.copy()
image_clean = df_image.copy()
```

清洁度

狗地位不在一列

方案

狗地位应置于一列，为避免原数据可能的错误，故在质量方案中重新提取

质量

含有转发数据

方案

清理转发数据,并删除数据缺失严重的列及无用列

代码

In [16]:

```
archive_clean = archive_clean[archive_clean.retweeted_status_id.isnull()]
archive_clean = archive_clean[['tweet_id', 'text', 'rating_numerator', 'rating_denominator', 'na
me']]
```

测试

In [17]:

```
archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2355
Data columns (total 5 columns):
tweet_id              2175 non-null int64
text                  2175 non-null object
rating_numerator      2175 non-null int64
rating_denominator    2175 non-null int64
name                  2175 non-null object
dtypes: int64(3), object(2)
memory usage: 102.0+ KB
```

分子评分有低于10分的，分子分母有不合规的评分，如1776的分子评分及420/10，165/150，143/130

狗名有a，an，the等明显错误，还有大量空值

方案

重新从文本中提取狗名，评分，且只提取分母为10，分子大于分母又没有大太多的评分

代码

In [18]:

```
# 添加新列：狗地位
archive_clean['status'] = None
```

In [19]:

```
archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2355
Data columns (total 6 columns):
tweet_id              2175 non-null int64
text                  2175 non-null object
rating_numerator      2175 non-null int64
rating_denominator    2175 non-null int64
name                  2175 non-null object
status                0 non-null object
dtypes: int64(3), object(3)
memory usage: 118.9+ KB
```

In [20]:

```
# 从text里提取狗名
archive_clean.name = archive_clean.text.str.extract('(?:This is|Meet|name is|Say hello to|named)
 ([A-Z][a-z]{2,12})', expand=False).values

# 提取地位
import re
for i in archive_clean.index:
    status_set = set(re.findall('(doggo|floofer|pupper|puppo)', archive_clean.loc[i,'text']))
    if len(status_set) > 0:
        status_value = ', '.join(status_set)
        archive_clean.loc[i, 'status'] = status_value

# 提取评分，之所以rating_numerator的范围限定为11~16，是因为视觉评估时发现正常评分基本都在11~15之间，提取时就
# 放弃420/10，165/150，143/130这种异常值
archive_clean.rating_numerator = archive_clean.text.str.extract('(1[1-6]\.?\d*)/10', expand=Fals
e).values
archive_clean.rating_denominator = archive_clean.text.str.extract('1[1-6]\.?\d*/(10)', expand=Fa
lse).values
```

测试

In [21]:

```
archive_clean.rating_numerator.value_counts()
```

Out[21]:

```
12       501
11       428
13       311
14        44
11.26      1
11.27      1
13.5       1
15         1
Name: rating_numerator, dtype: int64
```

In [22]:

```
archive_clean.status.value_counts()
```

Out[22]:

```
pupper           242
doggo             78
puppo             30
pupper, doggo      8
floofer            4
puppo, doggo       2
Name: status, dtype: int64
```

In [23]:

```
archive_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2175 entries, 0 to 2355
Data columns (total 6 columns):
tweet_id              2175 non-null int64
text                  2175 non-null object
rating_numerator      1288 non-null object
rating_denominator    1288 non-null object
name                  1399 non-null object
status                364 non-null object
dtypes: int64(1), object(5)
memory usage: 198.9+ KB
```

id应为str型

方案

合并数据框，将id统一改为str

代码

In [24]:

```
df_clean_transit = pd.merge(archive_clean, image_clean, on='tweet_id')
df_clean = pd.merge(tweet_clean, df_clean_transit, left_on='id', right_on='tweet_id')
df_clean.drop(columns='id', inplace=True)
```

测试

In [25]:

```
df_clean_transit.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1994 entries, 0 to 1993
Data columns (total 17 columns):
tweet_id              1994 non-null int64
text                  1994 non-null object
rating_numerator      1152 non-null object
rating_denominator    1152 non-null object
name                  1356 non-null object
status                326 non-null object
jpg_url               1994 non-null object
img_num               1994 non-null int64
p1                    1994 non-null object
p1_conf               1994 non-null float64
p1_dog                1994 non-null bool
p2                    1994 non-null object
p2_conf               1994 non-null float64
p2_dog                1994 non-null bool
p3                    1994 non-null object
p3_conf               1994 non-null float64
p3_dog                1994 non-null bool
dtypes: bool(3), float64(3), int64(2), object(9)
memory usage: 239.5+ KB
```

In [26]:

```
df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1994 entries, 0 to 1993
Data columns (total 19 columns):
retweet_count         1994 non-null int64
favorite_count        1994 non-null int64
tweet_id              1994 non-null int64
text                  1994 non-null object
rating_numerator      1152 non-null object
rating_denominator    1152 non-null object
name                  1356 non-null object
status                326 non-null object
jpg_url               1994 non-null object
img_num               1994 non-null int64
p1                    1994 non-null object
p1_conf               1994 non-null float64
p1_dog                1994 non-null bool
p2                    1994 non-null object
p2_conf               1994 non-null float64
p2_dog                1994 non-null bool
p3                    1994 non-null object
p3_conf               1994 non-null float64
p3_dog                1994 non-null bool
dtypes: bool(3), float64(3), int64(4), object(9)
memory usage: 270.7+ KB
```

可视化

In [27]:

```
import matplotlib.pyplot as plt
% matplotlib inline
```

In [28]:

```
df_clean['rating_numerator'] = df_clean['rating_numerator'].astype(float)
df_clean['rating_denominator'] = df_clean['rating_denominator'].astype(float)
```

In [29]:

```
df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1994 entries, 0 to 1993
Data columns (total 19 columns):
retweet_count         1994 non-null int64
favorite_count        1994 non-null int64
tweet_id              1994 non-null int64
text                  1994 non-null object
rating_numerator      1152 non-null float64
rating_denominator    1152 non-null float64
name                  1356 non-null object
status                326 non-null object
jpg_url               1994 non-null object
img_num               1994 non-null int64
p1                    1994 non-null object
p1_conf               1994 non-null float64
p1_dog                1994 non-null bool
p2                    1994 non-null object
p2_conf               1994 non-null float64
p2_dog                1994 non-null bool
p3                    1994 non-null object
p3_conf               1994 non-null float64
p3_dog                1994 non-null bool
dtypes: bool(3), float64(5), int64(4), object(7)
memory usage: 270.7+ KB
```

In [30]:

```
df_clean.rating_numerator.hist();# 此处能否忽略那几个含小数的值？或者设置分组让条柱都处在11，12，13，14的中间
```