1 收集

1.1 / import lib

In [524]:

```
# import libs

## official libs
import pprint as pp

## 3rd libs
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
from PIL import Image

## private libs
import wrangling1 as w

## paras
%matplotlib inline
```

executed in 9ms, finished 22:31:02 2019-08-09

In [525]:

```
# import libs
## official libs
import pprint  as pp

## third libs
import pandas as pd
import numpy as np

## private libs
import wrangling1 as w
# 包括了一些数据评估的简单功能
```

executed in 7ms, finished 22:31:02 2019-08-09

1.2 / load df

In [526]:

```
# load df
## read_jason 有很多参数,可以参考官方文档
## 此处要加 lines=True
df = pd.read_json('tweet_json.txt',lines=True)
# https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_
```

executed in 1.14s, finished 22:31:03 2019-08-09

In [527]:

```
# load df
## read_jason 有很多参数,可以参考官方文档
## 此处要加 lines=True
df = pd.read_json('tweet_json.txt',lines=True)
# https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_
```

executed in 817ms, finished 22:31:04 2019-08-09

1.3 / check df

In [528]:

```
w.check_sample(df)
```

executed in 18ms, finished 22:31:04 2019-08-09

```
----checking sample index: 187

- columns #1 : contributors
[nan]

- columns #2 : coordinates
[nan]

- columns #3 : created_at
['2017-04-22T18:55:51.000000000']

- columns #4 : display_text_range
[list([0, 89])]

- columns #5 : entities
[{'hashtags': [], 'symbols': [], 'user_mentions': [], 'urls': [{'url': 'https:
//t.co/sb73bV5Y7S', 'expanded_url': 'https://twitter.com/perfy/sta
tus/855857318168150016', 'display_url': 'twitter.com/perfy/status/
8…', 'indices': [90, 113]}]}]
```

In [529]:

```
1 df.columns
```

executed in 19ms, finished 22:31:04 2019-08-09

Out[529]:

```
Index(['contributors', 'coordinates', 'created_at', 'display_text_range',
       'entities', 'extended_entities', 'favorite_count', 'favorited',
       'full_text', 'geo', 'id', 'id_str', 'in_reply_to_screen_name',
       'in_reply_to_status_id', 'in_reply_to_status_id_str',
       'in_reply_to_user_id', 'in_reply_to_user_id_str', 'is_quote_status',
       'lang', 'place', 'possibly_sensitive', 'possibly_sensitive_appealable
',
       'quoted_status', 'quoted_status_id', 'quoted_status_id_str',
       'retweet_count', 'retweeted', 'retweeted_status', 'source', 'trunca
ted',
       'user'],
      dtype='object')
```

In [530]:

```
1 df.info()
```

executed in 21ms, finished 22:31:04 2019-08-09

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2352 entries, 0 to 2351
Data columns (total 31 columns):
contributors                    0 non-null float64
coordinates                     0 non-null float64
created_at                      2352 non-null datetime64[ns]
display_text_range              2352 non-null object
entities                        2352 non-null object
extended_entities               2073 non-null object
favorite_count                  2352 non-null int64
favorited                       2352 non-null bool
full_text                       2352 non-null object
geo                             0 non-null float64
id                              2352 non-null int64
id_str                          2352 non-null int64
in_reply_to_screen_name         78 non-null object
in_reply_to_status_id           78 non-null float64
in_reply_to_status_id_str       78 non-null float64
in_reply_to_user_id             78 non-null float64
in_reply_to_user_id_str         78 non-null float64
is_quote_status                 2352 non-null bool
lang                            2352 non-null object
place                           1 non-null object
possibly_sensitive              2211 non-null float64
possibly_sensitive_appealable   2211 non-null float64
quoted_status                   28 non-null object
quoted_status_id                29 non-null float64
quoted_status_id_str            29 non-null float64
retweet_count                   2352 non-null int64
retweeted                       2352 non-null bool
retweeted_status                177 non-null object
source                          2352 non-null object
truncated                       2352 non-null bool
user                            2352 non-null object
dtypes: bool(4), datetime64[ns](1), float64(11), int64(4), object(11)
memory usage: 505.4+ KB
```

In [531]:

```
1  df.describe()
```

executed in 79ms, finished 22:31:04 2019-08-09

Out[531]:

	contributors	coordinates	favorite_count	geo	id	
count	0.0	0.0	2352.000000	0.0	2.352000e+03	2.352
mean	NaN	NaN	8109.198980	NaN	7.425913e+17	7.425
std	NaN	NaN	11980.795669	NaN	6.846210e+16	6.846
min	NaN	NaN	0.000000	NaN	6.660209e+17	6.660
25%	NaN	NaN	1417.000000	NaN	6.783949e+17	6.783
50%	NaN	NaN	3596.500000	NaN	7.193536e+17	7.193
75%	NaN	NaN	10118.000000	NaN	7.991219e+17	7.991
max	NaN	NaN	132318.000000	NaN	8.924206e+17	8.924

2 评估

2.1 / quanlity

2.1.1 // drop1

In [532]:

```
1  # drop list1
2  ## 先删除重复和无意义的信息
3  droplist1 = ['contributors','coordinates','geo','place','id_str',
4               'in_reply_to_status_id_str','in_reply_to_user_id_str','quoted_status_id_s
```

executed in 6ms, finished 22:31:04 2019-08-09

In [533]:

```
## drop list1 excute
w.drop_column(df,droplist1)
```

executed in 12ms, finished 22:31:04 2019-08-09

---- proceding ----
– drop 8 columns: ['contributors', 'coordinates', 'geo', 'place', 'id_str',
'in_reply_to_status_id_str', 'in_reply_to_user_id_str', 'quoted_status_id
_str']
– remain 23 columns
– success : True

2.1.2 // check - inspect list

对一些怀疑是否有用的数据进行检视

In [534]:

```
# check1
## snip remained columns
w.check_sample(df)
```

executed in 23ms, finished 22:31:04 2019-08-09

----checking sample index: 1735

– columns #1 : created_at
['2015–12–23T03:58:25.000000000']

– columns #2 : display_text_range
[list([0, 133])]

– columns #3 : entities
[{'hashtags': [], 'symbols': [], 'user_mentions': [], 'urls': [], 'media': [{'
id': 679511347441328128, 'id_str': '679511347441328128', 'indices':
[110, 133], 'media_url': 'http://pbs.twimg.com/media/CW4b–GUWY
AAa8QO.jpg', 'media_url_https': 'https://pbs.twimg.com/media/CW
4b–GUWYAAa8QO.jpg', 'url': 'https://t.co/bwuV6FlRxr', 'display_url
': 'pic.twitter.com/bwuV6FlRxr', 'expanded_url': 'https://twitter.com
/dog_rates/status/679511351870550016/photo/1', 'type': 'photo',
'sizes': {'medium': {'w': 505, 'h': 639, 'resize': 'fit'}, 'large': {'w': 505,
'h': 639, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize': 'crop'},
'small': {'w': 505, 'h': 639, 'resize': 'fit'}}}]}]

– columns #4 : extended_entities
[{'media': [{'id': 679511347441328128, 'id_str': '67951134744132812
8', 'indices': [110, 133], 'media_url': 'http://pbs.twimg.com/media/C

W4b-GUWYAAa8QO.jpg', 'media_url_https': 'https://pbs.twimg.co
m/media/CW4b-GUWYAAa8QO.jpg', 'url': 'https://t.co/bwuV6FlRx
r', 'display_url': 'pic.twitter.com/bwuV6FlRxr', 'expanded_url': 'https:
//twitter.com/dog_rates/status/679511351870550016/photo/1', 'ty
pe': 'photo', 'sizes': {'medium': {'w': 505, 'h': 639, 'resize': 'fit'}, 'larg
e': {'w': 505, 'h': 639, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150,

'resize': 'crop'}, 'small': {'w': 505, 'h': 639, 'resize': 'fit'}}}]}]

- columns #5 : favorite_count
[3694]

- columns #6 : favorited
[False]

- columns #7 : full_text
["Say hello to William. He makes fun of others because he's terrifie
d of his own deep-seated insecurities. 7/10 https://t.co/bwuV6FlR
xr"] (https://t.co/bwuV6FlRxr"])

- columns #8 : id------
[679511351870550016]

- columns #9 : in_reply_to_screen_name
[None]

- columns #10 : in_reply_to_status_id
[nan]

- columns #11 : in_reply_to_user_id
[nan]

- columns #12 : is_quote_status
[False]

- columns #13 : lang----
['en']

- columns #14 : possibly_sensitive
[0.]

- columns #15 : possibly_sensitive_appealable
[0.]

- columns #16 : quoted_status
[nan]

- columns #17 : quoted_status_id
[nan]

– columns #18 : retweet_count
[1454]

– columns #19 : retweeted
[False]

– columns #20 : retweeted_status
[nan]

– columns #21 : source––
['Tw
itter for iPhone']

– columns #22 : truncated
[False]

– columns #23 : user––––
[{'id': 4196983835, 'id_str': '4196983835', 'name': 'SpookyWeRateD
ogs™', 'screen_name': 'dog_rates', 'location': 'MERCH⬎ DM DOGS
. WE WILL RATE', 'description': 'Only Legit Source for Professional
Dog Ratings STORE: @ShopWeRateDogs | IG, FB & SC: WeRateDo
gs | MOBILE APP: @GoodDogsGame Business: dogratingtwitter@g
mail.com', 'url': 'https://t.co/N7sNNHAEXS', 'entities': {'url': {'urls': [
{'url': 'https://t.co/N7sNNHAEXS', 'expanded_url': 'http://weratedo
gs.com', 'display_url': 'weratedogs.com', 'indices': [0, 23]}]}, 'descri
ption': {'urls': []}}, 'protected': False, 'followers_count': 3768911,
'friends_count': 107, 'listed_count': 3317, 'created_at': 'Sun Nov 15
21:41:29 +0000 2015', 'favourites_count': 120161, 'utc_offset': None,
'time_zone': None, 'geo_enabled': True, 'verified': True, 'statuses_c
ount': 5749, '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_ba
ckground_image_url_https': 'https://abs.twimg.com/images/theme
s/theme1/bg.png', 'profile_background_tile': False, 'profile_image_u
rl': 'http://pbs.twimg.com/profile_images/914581071265755136/2h
5uFpwU_normal.jpg', 'profile_image_url_https': 'https://pbs.twimg.c
om/profile_images/914581071265755136/2h5uFpwU_normal.jpg', '
profile_banner_url': 'https://pbs.twimg.com/profile_banners/41969
83835/1506888628', 'profile_link_color': 'F5ABB5', 'profile_sidebar
_border_color': '000000', 'profile_sidebar_fill_color': '000000', 'profi
le_text_color': '000000', 'profile_use_background_image': False, 'ha
s_extended_profile': True, 'default_profile': False, 'default_profile_i
mage': False, 'following': False, 'follow_request_sent': False, 'notific
ations': False, 'translator_type': 'none'}]
'––––checking complete––––'

In [535]:

```
## define inspect list1
inslist1 = ['favorited','in_reply_to_screen_name','in_reply_to_status_id',
            'in_reply_to_user_id','is_quote_status','lang','possibly_sensitive',
            'possibly_sensitive_appealable','quoted_status_id','retweeted','truncate
```

executed in 6ms, finished 22:31:04 2019-08-09

In [536]:

```
## inspect info
df[inslist1].info()
```

executed in 18ms, finished 22:31:04 2019-08-09

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2352 entries, 0 to 2351
Data columns (total 11 columns):
favorited                   2352 non-null bool
in_reply_to_screen_name     78 non-null object
in_reply_to_status_id       78 non-null float64
in_reply_to_user_id         78 non-null float64
is_quote_status             2352 non-null bool
lang                        2352 non-null object
possibly_sensitive          2211 non-null float64
possibly_sensitive_appealable  2211 non-null float64
quoted_status_id            29 non-null float64
retweeted                   2352 non-null bool
truncated                   2352 non-null bool
dtypes: bool(4), float64(5), object(2)
memory usage: 137.9+ KB
```

In [537]:

```
## inspect value
w.check_value(df,inslist1)
```

executed in 54ms, finished 22:31:04 2019-08-09

```
- columns #1 : favorited
False    2352
Name: favorited, dtype: int64

- columns #2 : in_reply_to_screen_name
dog_rates       47
markhoppus       2
jonnysun         1
xianmcguire      1
ComplicitOwl     1
```

Name: in_reply_to_screen_name, dtype: int64

– columns #3 : in_reply_to_status_id
6.671522e+17 2
8.562860e+17 1

8.131273e+17 1
6.754971e+17 1
6.827884e+17 1
Name: in_reply_to_status_id, dtype: int64

– columns #4 : in_reply_to_user_id
4.196984e+09 47
2.195506e+07 2
7.305050e+17 1
2.916630e+07 1
3.105441e+09 1
Name: in_reply_to_user_id, dtype: int64

– columns #5 : is_quote_status
False 2321
True 31
Name: is_quote_status, dtype: int64

– columns #6 : lang––––
en 2334
und 7
in 3
nl 3
ro 1
Name: lang, dtype: int64

– columns #7 : possibly_sensitive
0.0 2211
Name: possibly_sensitive, dtype: int64

– columns #8 : possibly_sensitive_appealable
0.0 2211
Name: possibly_sensitive_appealable, dtype: int64

– columns #9 : quoted_status_id
8.340867e+17 1
8.413114e+17 1
7.061659e+17 1
8.860534e+17 1
8.464848e+17 1
Name: quoted_status_id, dtype: int64

– columns #10 : retweeted

False 2352
Name: retweeted, dtype: int64

– columns #11 : truncated
False 2352
Name: truncated, dtype: int64

'----checking complete----'

2.1.3 // check - quoted_status

In [538]:

```
# special1 quoted_status
## quoted_status is a dict, move it to detlist
## check values (almost is null)
df.quoted_status.isnull().value_counts()
```

executed in 15ms, finished 22:31:04 2019-08-09

Out[538]:

True 2324
False 28
Name: quoted_status, dtype: int64

In [539]:

```
## check a sample
df[df.quoted_status.notnull()].sample(1).quoted_status.iloc[0]
### this is some extra info abou a forward
```

executed in 14ms, finished 22:31:04 2019-08-09

Out[539]:

{'created_at': 'Wed Apr 27 01:34:44 +0000 2016',
'id': 725136065078521856,
'id_str': '725136065078521856',
'full_text': 'Se nos metió otro jugador al partido de @dvotachira vs @
pumasmx en la #LibertadoresEnFD 🐶\nhttps://t.co/nPtdOeTxcW',
'truncated': False,
'display_text_range': [0, 113],
'entities': {'hashtags': [{'text': 'LibertadoresEnFD', 'indices': [70, 87]}],
'symbols': [],
'user_mentions': [{'screen_name': 'DvoTachira',
'name': 'Deportivo Táchira FC',
'id': 85361349,
'id_str': '85361349',
'indices': [40, 51]},

```
   {'screen_name': 'PumasMX',
    'name': 'PUMAS',
    'id': 78938710,
    'id_str': '78938710',
    'indices': [55, 63]}],
  'urls': [{'url': 'https://t.co/nPtdOeTxcW',
    'expanded_url': 'https://amp.twimg.com/v/47e2d017-ad5d-4716-
a8ce-5173b32e0a18',
    'display_url': 'amp.twimg.com/v/47e2d017-ad5…',
    'indices': [90, 113]}]},
 'source': '<a href="http://twitter.com" rel="nofollow">Twitter Web Cl
ient</a>',
 'in_reply_to_status_id': None,
 'in_reply_to_status_id_str': None,
 'in_reply_to_user_id': None,
 'in_reply_to_user_id_str': None,
 'in_reply_to_screen_name': None,
 'user': {'id': 195947234,
  'id_str': '195947234',
  'name': 'FOX Deportes',
  'screen_name': 'FOXDeportes',
  'location': '',
  'description': 'Somos la primera cadena de deportes en español en
USA. Síguenos para obtener las mejores noticias del mundo deportiv
o y recibe alertas de nuestra programación.',
  'url': 'http://t.co/4JtKgkdxJ0',
  'entities': {'url': {'urls': [{'url': 'http://t.co/4JtKgkdxJ0',
      'expanded_url': 'http://foxdeportes.com',
      'display_url': 'foxdeportes.com',
      'indices': [0, 22]}]},
   'description': {'urls': []}},
  'protected': False,
  'followers_count': 598720,
  'friends_count': 619,
  'listed_count': 2311,
  'created_at': 'Mon Sep 27 23:36:43 +0000 2010',
  'favourites_count': 1505,
  'utc_offset': -25200,
  'time_zone': 'Pacific Time (US & Canada)',
  'geo_enabled': True,
  'verified': True,
  'statuses_count': 147644,
  'lang': 'es',
  'contributors_enabled': False,
  'is_translator': False,
  'is_translation_enabled': False,
  'profile_background_color': '010206',
  'profile_background_image_url': 'http://pbs.twimg.com/profile_back
ground_images/378800000034230784/950b4d90f231bc9c11e564ee
```

13eecda1.jpeg',
  'profile_background_image_url_https': 'https://pbs.twimg.com/profil
e_background_images/378800000034230784/950b4d90f231bc9c11
e564ee13eecda1.jpeg',
  'profile_background_tile': False,
  'profile_image_url': 'http://pbs.twimg.com/profile_images/80415535
6142153728/0mfaX5Zv_normal.jpg',
  'profile_image_url_https': 'https://pbs.twimg.com/profile_images/80
4155356142153728/0mfaX5Zv_normal.jpg',
  'profile_banner_url': 'https://pbs.twimg.com/profile_banners/195947
234/1507858560',
  'profile_link_color': 'D11313',
  'profile_sidebar_border_color': 'FFFFFF',
  'profile_sidebar_fill_color': 'E6E6E6',
  'profile_text_color': '333333',
  'profile_use_background_image': True,
  'has_extended_profile': False,
  'default_profile': False,
  'default_profile_image': False,
  'following': False,
  'follow_request_sent': False,
  'notifications': False,
  'translator_type': 'none'},
 'geo': None,
 'coordinates': None,
 'place': None,
 'contributors': None,
 'is_quote_status': False,
 'retweet_count': 4450,
 'favorite_count': 5040,
 'favorited': False,
 'retweeted': False,
 'possibly_sensitive': False,
 'possibly_sensitive_appealable': False,
 'lang': 'es'}

- 分析 quoted_status :
  - 是嵌套字典数据
  - 缺失很多(只有28个数据)
  - 内容无用信息比较多
- 结论:
  - 删除此列
  - user 列和此列类似,也删除

2.1.4 // check - in_reply_to_screen_name

In [541]:

```
# special2
## in_reply_to_screen_name have value dog_rates for 47 times
df.query('in_reply_to_screen_name == "dog_rates"')[:3]
```

executed in 60ms, finished 22:31:21 2019-08-09

Out[541]:

	created_at	display_text_range	entities	extended_entities
147	2017-05-12 17:12:53	[0, 139]	{'hashtags': [], 'symbols': [], 'user_mentions...	{'media': [{'id': 863079538779013120, 'id_str'...
181	2017-04-24 15:13:52	[0, 112]	{'hashtags': [], 'symbols': [], 'user_mentions...	{'media': [{'id': 856526604033556482, 'id_str'...
225	2017-04-01 16:41:12	[0, 135]	{'hashtags': [], 'symbols': [], 'user_mentions...	NaN

3 rows × 23 columns

In [542]:

```
df['in_reply_to_screen_name'].isnull().sum() / df.shape[0]
```

executed in 8ms, finished 22:31:24 2019-08-09

Out[542]:

0.9668367346938775

- 分析 in_reply_to_screen_name :
  - 可能 dog_rates 是默认回复名字
  - 数据缺失率为 97%
- 结论:
  - 删除数据

2.1.5 // drop2

根据上面 check 内容删除数据

In [543]:

```
# droplist2
droplist2 = inslist1.copy()
### use .copy to copy rather than llink
droplist2.append('quoted_status')
droplist2.append('retweeted_status')
droplist2.append('user')
#droplist2 = ['truncated','retweeted','possibly_sensitive_appealable','possibly_s
```

executed in 5ms, finished 22:31:26 2019-08-09

In [544]:

```
## drop2 excute
w.drop_column(df,droplist2)
```

executed in 23ms, finished 22:31:27 2019-08-09

---- proceding ----
– drop 14 columns: ['favorited', 'in_reply_to_screen_name', 'in_reply_t
o_status_id', 'in_reply_to_user_id', 'is_quote_status', 'lang', 'possibly_s
ensitive', 'possibly_sensitive_appealable', 'quoted_status_id', 'retweet
ed', 'truncated', 'quoted_status', 'retweeted_status', 'user']
– remain 9 columns
– success : True

2.1.6 // check - detail columns

对嵌套的数据进行检视

In [545]:

```
# check3
## recheck sample
w.check_sample(df)
```

executed in 20ms, finished 22:31:30 2019-08-09

----checking sample index: 316

– columns #1 : created_at
['2017-02-22T18:59:48.000000000']

– columns #2 : display_text_range
[list([0, 139])]

– columns #3 : entities
[{'hashtags': [], 'symbols': [], 'user_mentions': [{'screen_name':
'dog_rates', 'name': 'SpookyWeRateDogs™', 'id': 4196983835, 'id_str':
'4196983835', 'indices': [3, 13]}], 'urls': []}]

– columns #4 : extended_entities
[nan]

– columns #5 : favorite_count
[0]

– columns #6 : full_text
["RT @dog_rates: This is Leo. He was a skater pup. She said see ya l
ater pup. He wasn't good enough for her. 12/10 you're good enough f
or me..."]

– columns #7 : id------
[834477809192075265]

– columns #8 : retweet_count
[12146]

– columns #9 : source--
['Twitt
er for iPhone']
'----checking complete----'

In [546]:

```
# detail list1
## check dict long columns
detlist1 = ['entities','extended_entities']
```

executed in 4ms, finished 22:31:41 2019-08-09

In [547]:

```
# detail check
w.check_detail(df,detlist1)
### not new info  -> drop
```

executed in 22ms, finished 22:31:42 2019-08-09

```
– columns #1 : entities
{'hashtags': [],
 'media': [{'display_url': 'pic.twitter.com/MgUWQ76dJU',
            'expanded_url': 'https://twitter.com/dog_rates/status/89242
0643555336193/photo/1',
            'id': 8924206394868776 96,
            'id_str': '8924206394868776 96',
            'indices': [86, 109],
            'media_url': 'http://pbs.twimg.com/media/DGKD1–bXoAAIAU
K.jpg',
            'media_url_https': 'https://pbs.twimg.com/media/DGKD1–bX
oAAIAUK.jpg',
            'sizes': {'large': {'h': 528, 'resize': 'fit', 'w': 540},
                      'medium': {'h': 528, 'resize': 'fit', 'w': 540},
                      'small': {'h': 528, 'resize': 'fit', 'w': 540},
                      'thumb': {'h': 150, 'resize': 'crop', 'w': 150}},
            'type': 'photo',
            'url': 'https://t.co/MgUWQ76dJU'}],
 'symbols': [],
 'urls': [],
 'user_mentions': []}

– columns #2 : extended_entities
{'media': [{'display_url': 'pic.twitter.com/MgUWQ76dJU',
            'expanded_url': 'https://twitter.com/dog_rates/status/89242
0643555336193/photo/1',
            'id': 8924206394868776 96,
            'id_str': '8924206394868776 96',
            'indices': [86, 109],
            'media_url': 'http://pbs.twimg.com/media/DGKD1–bXoAAIAU
K.jpg',
            'media_url_https': 'https://pbs.twimg.com/media/DGKD1–bX
oAAIAUK.jpg',
            'sizes': {'large': {'h': 528, 'resize': 'fit', 'w': 540},
                      'medium': {'h': 528, 'resize': 'fit', 'w': 540},
                      'small': {'h': 528, 'resize': 'fit', 'w': 540},
                      'thumb': {'h': 150, 'resize': 'crop', 'w': 150}},
            'type': 'photo',
            'url': 'https://t.co/MgUWQ76dJU'}]}
```

- 分析:
  - 是嵌套字典数据
  - 缺失不多
  - 内容无用信息比较多(有些与其他列有重复)
- 结论:
  - 删除列

2.1.7 // drop3

In [548]:

```
# drop list3
droplist3 = detlist1.copy()

## drop3 excute
w.drop_column(df,droplist3)
```

executed in 21ms, finished 22:31:46 2019-08-09

---- proceding ----
– drop 2 columns: ['entities', 'extended_entities']
– remain 7 columns
– success : True

2.1.8 // check - display_text_range

使用函数check_value会在这一列报错,检查下是因为这列的列表嵌套数字的原因

In [549]:

```
# check specified
df.display_text_range.sample(5)
```

executed in 11ms, finished 22:31:47 2019-08-09

Out[549]:

281 [0, 112]
2236 [0, 139]
68 [0, 132]
1516 [0, 108]
717 [0, 107]
Name: display_text_range, dtype: object

In [550]:

```
dflist = ['created_at',
 'favorite_count',
 'full_text',
 'id',
 'retweet_count',
 'source']
```

executed in 4ms, finished 22:31:48 2019-08-09

In [551]:

```
w.check_value(df,dflist)
```

executed in 29ms, finished 22:31:50 2019-08-09

```
– columns #1 : created_at
2016–09–12 15:10:21    1
2016–06–03 01:07:16    1
2017–01–31 01:27:39    1
2016–10–13 23:23:56    1
2016–06–27 01:37:04    1
Name: created_at, dtype: int64

– columns #2 : favorite_count
0       177
1753      3
3548      3
689       3
1526      3
Name: favorite_count, dtype: int64

– columns #3 : full_text
Three generations of pupper. 11/10 for all https://t.co/tAmQYvzrau (h
ttps://t.co/tAmQYvzrau)
1
This is a rare Arctic Wubberfloof. Unamused by the happenings. No l
onger has the appetites. 12/10 would totally hug https://t.co/krvbacl
X0N (https://t.co/krvbacIX0N)    1
RT @rachaeleasler: these @dog_rates hats are 13/10 bean approved
https://t.co/nRCdq4g9gG (https://t.co/nRCdq4g9gG)
1
Here we see 33 dogs posing for a picture. All get 11/10 for superb co
operation https://t.co/TRAri5iHzd (https://t.co/TRAri5iHzd)
1
This is Beemo. He's a Chubberflop mix. 12/10 would cross the world
for https://t.co/kzMVMU8HBV (https://t.co/kzMVMU8HBV)
1
Name: full_text, dtype: int64
```

– columns #4 : id––––––
749075273010798592   1
741099773336379392   1
798644042770751489   1
825120256414846976   1
769212283578875904   1
Name: id, dtype: int64

– columns #5 : retweet_count
1280   5
312   5
745   5
1554   4
1103   4
Name: retweet_count, dtype: int64

– columns #6 : source––
Twitte
r for iPhone   2217
Vine – Make a Scene
91
Twitter Web Client
33
<a href="https://about.twitter.com/products/tweetdeck" rel="nofollo
w">TweetDeck   11
Name: source, dtype: int64
'––––checking complete––––'

2.1.9 // check - null data

In [552]:

```
1 df.info()
```

executed in 11ms, finished 22:31:51 2019-08-09

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2352 entries, 0 to 2351
Data columns (total 7 columns):
created_at            2352 non-null datetime64[ns]
display_text_range    2352 non-null object
favorite_count        2352 non-null int64
full_text             2352 non-null object
id                    2352 non-null int64
retweet_count         2352 non-null int64
source                2352 non-null object
dtypes: datetime64[ns](1), int64(3), object(3)
memory usage: 128.7+ KB
```

In [553]:

```
1 df.isnull().sum()
```

executed in 13ms, finished 22:31:52 2019-08-09

Out[553]:

```
created_at            0
display_text_range    0
favorite_count        0
full_text             0
id                    0
retweet_count         0
source                0
dtype: int64
```

2.1.10 // drop4

想来想去还是把 id 给drop了, 后续分析中用不到还有隐私隐患

In [554]:

```
droplist4 = ['id']
w.drop_column(df,droplist4)
```

executed in 8ms, finished 22:31:53 2019-08-09

---- proceding ----
– drop 1 columns: ['id']
– remain 6 columns
– success : True

2.1.11 // review (quanlity)

根据数据删除剩余7列 ['created_at', 'display_text_range', 'favorite_count', 'full_text', 'id', 'retweet_count', 'source']

- id 为标识列
- created_at 包括时间、日期,可以进行时序分析
- display_text_range 为文字长度
- favorite_count 为点赞数
- full_text 为文字内容
- retweet_count 为回复数
- source 为来源

2.1.12 // persistence

In [555]:

```
df.to_pickle('tweet.pickle.xz', compression='xz')
```

executed in 185ms, finished 22:31:55 2019-08-09

2.2 / tidyness

根据质量部分的输出,对于除id列之外的需要进行清洁度的整理

- created_at 包括时间、日期,可以进行时序分析
  - 转换为 datafame 的 datetime 格式
- display_text_range 为文字长度
  - 原格式为 [0-x] x实际为推文长度,需要提取 x, 有个别是 [x-y], 不知道为什么还有下限, 提取上限数据即可
  - 本列为非必须列,可以根据 full_text 得出回复长度
- favorite_count 为点赞数
  - 数字类型,无需转换
- full_text 为文字内容
  - 后续如果进行nlp的分析需要进行向量化
- retweet_count 为回复数
  - 数字类型,无需转换
- source 为来源
  - 来源为链接,中间为发布信息的设备
  - 需要使用 re 来完成提取
  - 最后输出为分类信息

2.2.1 / load clean df

In [556]:

```
1  dfclean = pd.read_pickle('tweet.pickle.xz', compression='xz')
2  dftest = dfclean.copy()
3  dfclean.sample()
```

executed in 34ms, finished 22:31:57 2019-08-09

Out[556]:

	created_at	display_text_range	favorite_count	full_text	retweet_cou
1636	2016-01-04 23:02:22	[0, 126]	3250	This is Sweets the English Bulldog. Waves back...	169

2.2.2 // created_at

define: 将数据转换为时间格式

- solution1 使用 dataframe 的 datatime 格式
  - 数据本身为 datetime 格式
  - 如果是时序的数据可以将时间转换为 index,非常方便筛选
    https://chrisalbon.com/python/data_wrangling/pandas_time_series_basics/
    (https://chrisalbon.com/python/data_wrangling/pandas_time_series_basics/)
- (solution2 使用 python datatime 格式、calendar格式)

In [557]:

```
1  # convert to datetime format
2  dftest.created_at = pd.to_datetime(dftest.created_at)
3  dftest.info()
```

executed in 14ms, finished 22:32:00 2019-08-09

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2352 entries, 0 to 2351
Data columns (total 6 columns):
created_at              2352 non-null datetime64[ns]
display_text_range      2352 non-null object
favorite_count          2352 non-null int64
full_text               2352 non-null object
retweet_count           2352 non-null int64
source                  2352 non-null object
dtypes: datetime64[ns](1), int64(2), object(3)
memory usage: 110.3+ KB
```

In [558]:

```
1 sample = dftest.sample()
2 sample
```

executed in 17ms, finished 22:32:00 2019-08-09

Out[558]:

	created_at	display_text_range	favorite_count	full_text	retweet_cou
2258	2015-11-20 03:35:20	[0, 140]	350	Here is George. George took a selfie of his ne...	13

In [559]:

```
1 sample.created_at.dt.month, sample.created_at.dt.day, sample.created_at.dt.ho
```

executed in 11ms, finished 22:32:03 2019-08-09

Out[559]:

```
(2258    11
 Name: created_at, dtype: int64, 2258    20
 Name: created_at, dtype: int64, 2258    3
 Name: created_at, dtype: int64)
```

In [560]:

```
1 dftest.tail(10)
```

executed in 20ms, finished 22:32:12 2019-08-09

Out[560]:

	created_at	display_text_range	favorite_count	full_text	retweet_co
2342	2015-11-16 01:01:59	[0, 135]	117	Here is the Rand Paul of retrievers folks! He'...	
2343	2015-11-16 00:55:59	[0, 124]	304	My oh my. This is a rare blond Canadian terrie...	

2344	2015-11-16 00:49:46	[0, 140]	449	Here is a Siberian heavily armored polar bear ...
2345	2015-11-16 00:35:11	[0, 138]	1250	This is an odd dog. Hard on the outside but lo...
2346	2015-11-16 00:30:50	[0, 140]	136	This is a truly beautiful English Wilson Staff...
2347	2015-11-16 00:24:50	[0, 120]	111	Here we have a 1949 1st generation vulpix. Enj...
2348	2015-11-16 00:04:52	[0, 137]	309	This is a purebred Piers Morgan. Loves to Netf...
2349	2015-11-15 23:21:54	[0, 130]	128	Here is a very happy pup. Big fan of well-main...
2350	2015-11-15 23:05:30	[0, 139]	132	This is a western brown Mitsubishi terrier. Up...
2351	2015-11-15 22:32:08	[0, 131]	2528	Here we have a Japanese Irish Setter. Lost eye...

In [561]:

```
# 根据上述观察, 发现时间是按照发生顺序倒序排列的
## 时序分析入门 https://ourcodingclub.github.io/2019/01/07/pandas-time-ser
## 需要转换为 datetime index (方便筛选)
## 将在正式数据上实现
### df.where 可以直接替换,有空测试
dfclean.index = pd.to_datetime(dfclean.created_at)
dfclean.index.name = 'time_index'
```

executed in 7ms, finished 22:32:14 2019-08-09

In [562]:

```
droplist = ['created_at']
w.drop_column(dfclean, droplist)
```

executed in 10ms, finished 22:32:17 2019-08-09

---- proceding ----
– drop 1 columns: ['created_at']
– remain 5 columns
– success : True

In [563]:

```
dfclean['20170101']
```

executed in 23ms, finished 22:32:20 2019-08-09

Out[563]:

time_index	display_text_range	favorite_count	full_text	retweet_count	
2017-01-01 19:22:38	[0, 100]	9130	This is Titan. His nose is quite chilly. Reque...	1901	hre
2017-01-01 02:53:20	[0, 44]	11423	Happy New Year from the squad! 13/10 for all h...	4388	hre

2.2.3 // display_text_range

define: 抽取出 text 的长度,存为整数

- solution1 使用 python standard re lib
  - 抽出字符
  - 转换为 int
- https://pandas.pydata.org/pandas-docs/stable/user_guide/text.html (https://pandas.pydata.org/pandas-docs/stable/user_guide/text.html) 非常全面的介绍

In [564]:

```
# code
## extract str
dfclean.display_text_range[:10]
```

executed in 11ms, finished 22:32:22 2019-08-09

Out[564]:

```
time_index
2017-08-01 16:23:56     [0, 85]
2017-08-01 00:17:27    [0, 138]
2017-07-31 00:18:03    [0, 121]
2017-07-30 15:58:51     [0, 79]
2017-07-29 16:00:24    [0, 138]
2017-07-29 00:08:17    [0, 138]
2017-07-28 16:27:12    [0, 140]
2017-07-28 00:22:40    [0, 118]
2017-07-27 16:25:51    [0, 122]
2017-07-26 15:59:51    [0, 133]
Name: display_text_range, dtype: object
```

In [565]:

```
## 使用.str[slice] 直接解析相应位置的数字
dfclean.display_text_range = dfclean.display_text_range.str[1]
```

executed in 11ms, finished 22:32:25 2019-08-09

In [566]:

```
1  dfclean.display_text_range = dfclean.display_text_range.astype(int)
2  dfclean.info()
```

executed in 17ms, finished 22:32:26 2019-08-09

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 2352 entries, 2017-08-01 16:23:56 to 2015-11-15 22:
32:08
Data columns (total 5 columns):
display_text_range    2352 non-null int64
favorite_count        2352 non-null int64
full_text             2352 non-null object
retweet_count         2352 non-null int64
source                2352 non-null object
dtypes: int64(3), object(2)
memory usage: 110.2+ KB
```

In [567]:

```
1  dfclean.describe()
```

executed in 30ms, finished 22:32:30 2019-08-09

Out[567]:

	display_text_range	favorite_count	retweet_count
count	2352.000000	2352.000000	2352.000000
mean	111.179847	8109.198980	3134.932398
std	27.364336	11980.795669	5237.846296
min	11.000000	0.000000	0.000000
25%	93.000000	1417.000000	618.000000
50%	116.000000	3596.500000	1456.500000
75%	137.000000	10118.000000	3628.750000
max	165.000000	132318.000000	79116.000000

2.2.4 // full_text

define:

- 每个评价后面都有一个分值和链接 11/10 https://t.co/8W5iSOgXfx (https://t.co/8W5iSOgXfx)
- 评分为 10/10 或 11/10,没找到说明, 链接科学上网也不能访问
- 需要删除后保存
- 此处不做处理,词云的制作最后再做
- try solution
  - str.replace
  - str[i]
  - str.extract(r'ab (%5Cd))
  - pat = / str.match
  - str.contains
  - get.dummies(sep=',')

In [568]:

```
# code
detlist = ['full_text']
dfclean.full_text[:10]
```

executed in 12ms, finished 22:32:33 2019-08-09

Out[568]:

```
time_index
2017-08-01 16:23:56    This is Phineas. He's a mystical boy. Only eve
...
2017-08-01 00:17:27    This is Tilly. She's just checking pup on you...
.
2017-07-31 00:18:03    This is Archie. He is a rare Norwegian Pounci
n...
2017-07-30 15:58:51    This is Darla. She commenced a snooze mid
meal...
2017-07-29 16:00:24    This is Franklin. He would like you to stop ca
...
2017-07-29 00:08:17    Here we have a majestic great white breachi
ng ...
2017-07-28 16:27:12    Meet Jax. He enjoys ice cream so much he g
ets ...
2017-07-28 00:22:40    When you watch your owner call another do
g a g...
2017-07-27 16:25:51    This is Zoey. She doesn't want to be one of t
h...
2017-07-26 15:59:51    This is Cassie. She is a college pup. Studying
...
Name: full_text, dtype: object
```

In [569]:

```
# extrac
dfclean.full_text[1]
```

executed in 8ms, finished 22:32:35 2019-08-09

Out[569]:

```
"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 htt
ps://t.co/0Xxu71qeIV" (https://t.co/0Xxu71qeIV")
```

In [570]:

```
## extract testing
s = pd.Series(['a1', 'b2', 'c3'])
s.str.extract(r'([ab])(\d)')
```

executed in 14ms, finished 22:32:36 2019-08-09

Out[570]:

	0	1
0	a	1
1	b	2
2	NaN	NaN

In [571]:

```
test = dfclean.full_text.str.lower()
```

executed in 6ms, finished 22:32:41 2019-08-09

In [572]:

```
test[117]
```

executed in 8ms, finished 22:32:41 2019-08-09

Out[572]:

'this is dewey (pronounced "covfefe"). he\'s having a good walk. argu
ably the best walk. 13/10 would snug softly https://t.co/hcieajkc4d'
(https://t.co/hcieajkc4d')

In [573]:

```
1  test.str.extract('(\d\d\/\d\d)')[:5]
```

executed in 30ms, finished 22:32:42 2019-08-09

Out[573]:

time_index	0
2017-08-01 16:23:56	13/10
2017-08-01 00:17:27	13/10
2017-07-31 00:18:03	12/10
2017-07-30 15:58:51	13/10
2017-07-29 16:00:24	12/10

In [574]:

```
1  test.str.extract('(.*)(\d{2}\/\d{2})')[:5]
```

executed in 171ms, finished 22:32:44 2019-08-09

Out[574]:

time_index	0	1
2017-08-01 16:23:56	this is phineas. he's a mystical boy. only eve...	13/10
2017-08-01 00:17:27	this is tilly. she's just checking pup on you....	13/10
2017-07-31 00:18:03	this is archie. he is a rare norwegian pouncin...	12/10
2017-07-30 15:58:51	this is darla. she commenced a snooze mid meal.	13/10
2017-07-29 16:00:24	this is franklin. he would like you to stop ca...	12/10

In [575]:

```
1  test.str.extract('(.*)(\d{2}\/\d{2})')[0].str.strip()[:5]
```

executed in 172ms, finished 22:32:47 2019-08-09

Out[575]:

```
time_index
2017-08-01 16:23:56    this is phineas. he's a mystical boy. only eve..
.
2017-08-01 00:17:27    this is tilly. she's just checking pup on you....
2017-07-31 00:18:03    this is archie. he is a rare norwegian pouncin.
..
2017-07-30 15:58:51     this is darla. she commenced a snooze mid
meal.
2017-07-29 16:00:24    this is franklin. he would like you to stop ca...
Name: 0, dtype: object
```

In [576]:

```
1  dfclean['clean_text'] = test.str.extract('(.*)(\d{2}\/\d{2})')[0]
```

executed in 168ms, finished 22:32:48 2019-08-09

In [577]:

```
1  droplist = ['full_text']
2  w.drop_column(dfclean,droplist)
```

executed in 9ms, finished 22:32:48 2019-08-09

```
---- proceding ----
- drop 1 columns: ['full_text']
- remain 5 columns
- success : True
```

2.2.5 // source

define: 抽取出发 tweet 使用的设备

- 信息是这样的 Twitter for iPhone
- 需要抽取出 `Twitter for iPhone`, 并定义分类为 iphone
- 将本列做成分类数据
- 更新
  - 根据value_counts的输出, 95%的来源为iphone, 失去分析价值(Android的去哪里了)
  - 不过起码说明移动的登陆要比网页多很多

In [578]:

```
# code
## 观察数据
dfclean.source.value_counts()
```

executed in 13ms, finished 22:32:50 2019-08-09

Out[578]:

```
<a href="http://twitter.com/download/iphone" rel="nofollow">Twitte
r for iPhone</a>      2217
<a href="http://vine.co" rel="nofollow">Vine – Make a Scene</a>
91
<a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>
33
<a href="https://about.twitter.com/products/tweetdeck" rel="nofollo
w">TweetDeck</a>       11
Name: source, dtype: int64
```

In [579]:

```
## drop
droplist = ['source']
w.drop_column(dfclean, droplist)
```

executed in 10ms, finished 22:32:51 2019-08-09

---- proceding ----
– drop 1 columns: ['source']
– remain 4 columns
– success : True

2.2.6 // persistence

In [580]:

```
# code
dfclean.to_pickle('tweetclean.pickle.xz', compression='xz')
```

executed in 114ms, finished 22:32:56 2019-08-09

3 探索

3.1 / load df

In [581]:

```
# code
df = pd.read_pickle('tweetclean.pickle.xz', compression='xz')
df.sample()
```

executed in 25ms, finished 22:32:58 2019-08-09

Out[581]:

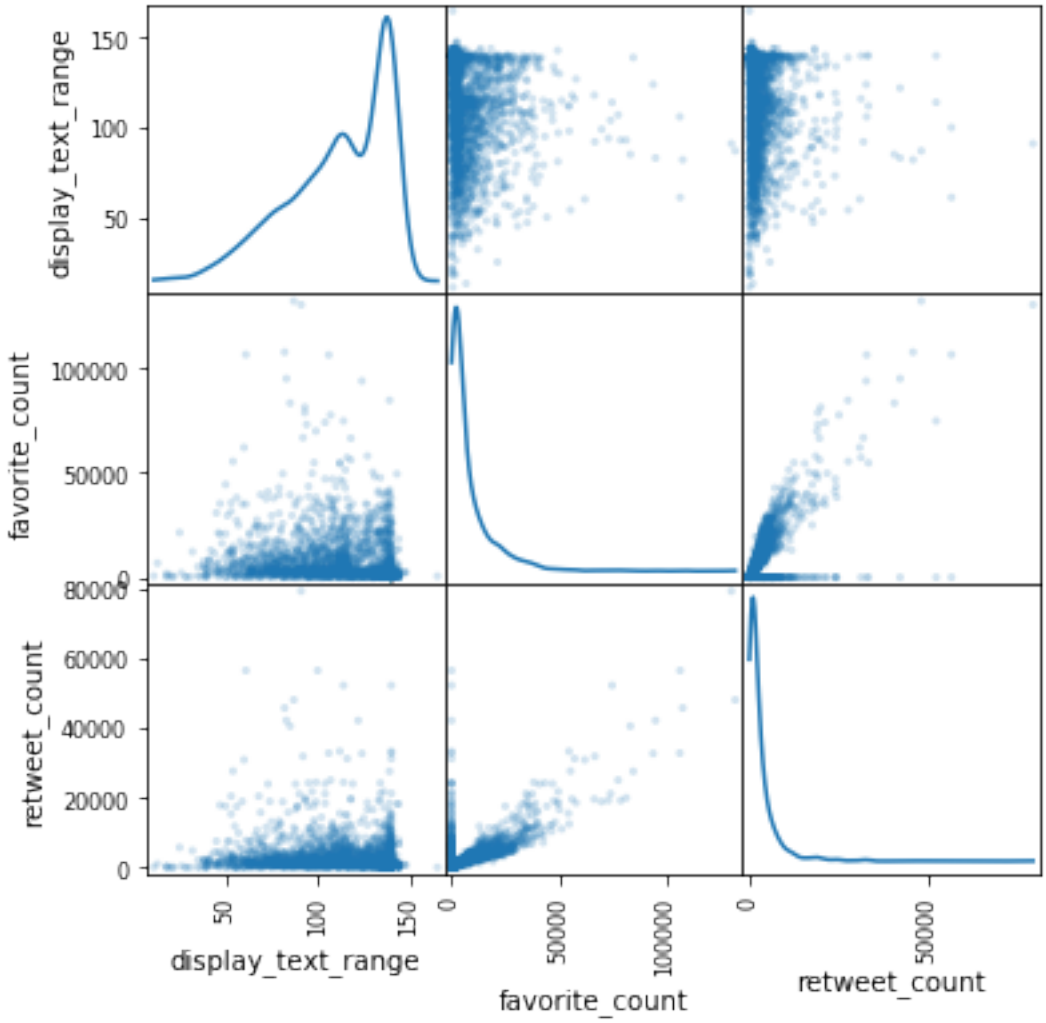
time_index	display_text_range	favorite_count	retweet_count	clean_text
2015-11-30 01:10:04	140	795	212	NaN

3.2 / data visulization

In [582]:

```
# code
# Scatter Matrix Plot
from pandas.plotting import scatter_matrix
scatter_matrix(df, alpha=0.2, figsize=(6, 6), diagonal='kde');
```

executed in 1.08s, finished 22:33:02 2019-08-09



In [583]:

```
df.columns
x = df.columns[0]
y = df.columns[1]
z = df.columns[2]
x, y, z
```

executed in 8ms, finished 22:33:03 2019-08-09

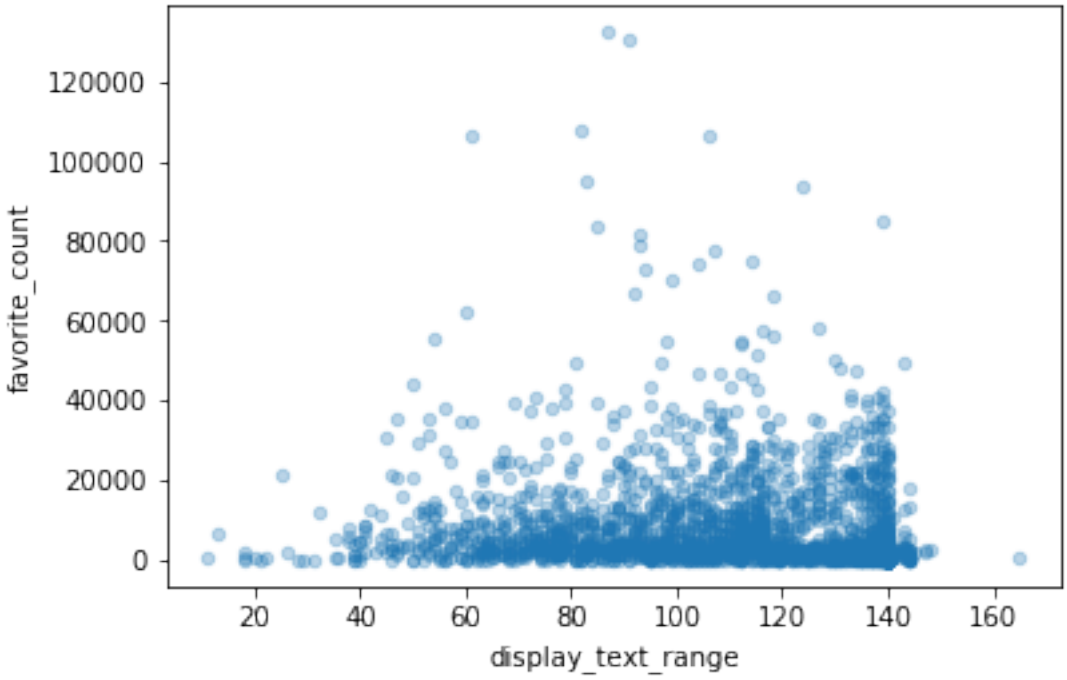
Out[583]:

('display_text_range', 'favorite_count', 'retweet_count')

In [584]:

```
1 df.plot.scatter(x,y,alpha=0.3);
```

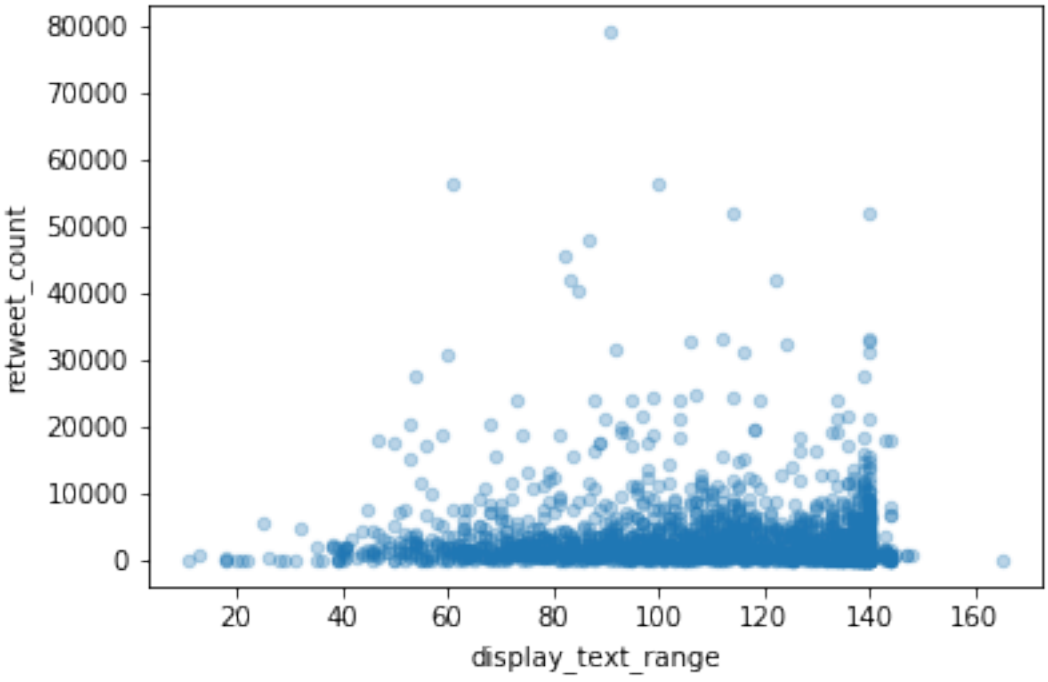
executed in 244ms, finished 22:33:04 2019-08-09



In [585]:

```
1 df.plot.scatter(x,z,alpha=0.3);
```

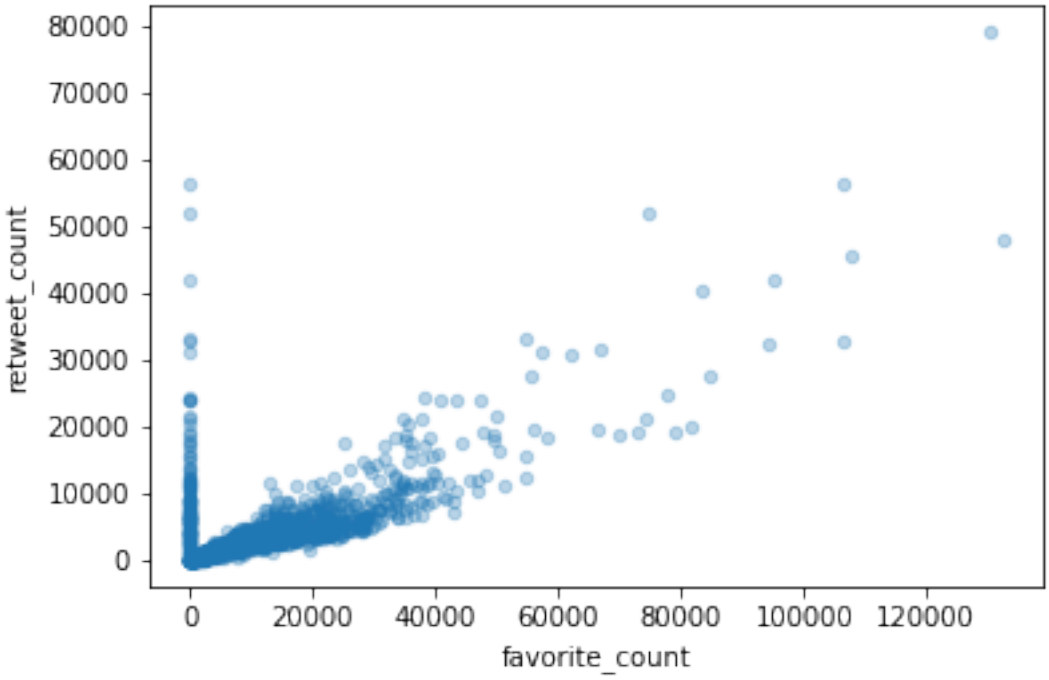
executed in 262ms, finished 22:33:04 2019-08-09



In [586]:

```
1 df.plot.scatter(y,z,alpha=0.3);
```

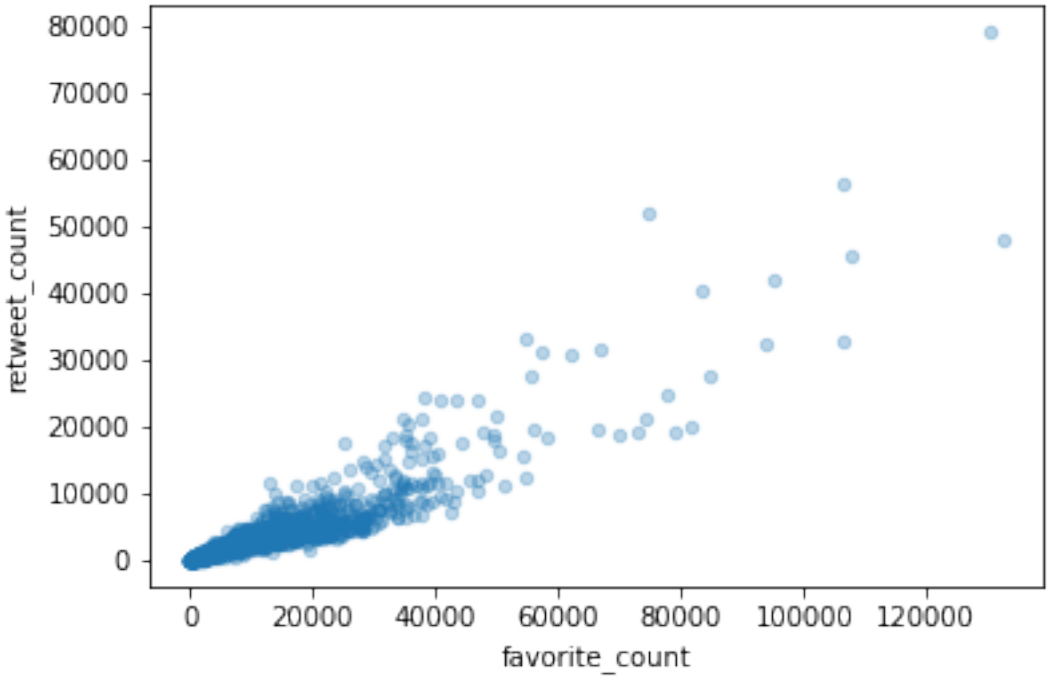
executed in 253ms, finished 22:33:05 2019-08-09



In [587]:

```
1 df.query('favorite_count > 0').plot.scatter(y,z,alpha=0.3);
```

executed in 236ms, finished 22:33:06 2019-08-09



3.3 / word cloud

3.3.1 // word cloud library

In [588]:

```
# code
#pip install wordcloud
```

executed in 4ms, finished 22:33:08 2019-08-09

3.3.2 // word cloud official example

In [589]:

```
from os import path
from PIL import Image
import numpy as np
import matplotlib.pyplot as plt
import os

from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator

# get data directory (using getcwd() is needed to support running example in g
d = path.dirname(__file__) if "__file__" in locals() else os.getcwd()

# Read the whole text.
text = open(path.join(d, 'alice.txt')).read()

# read the mask / color image taken from
# http://jirkavinse.deviantart.com/art/quot-Real-Life-quot-Alice-282261010
alice_coloring = np.array(Image.open(path.join(d, "alice_color.png")))
stopwords = set(STOPWORDS)
stopwords.add("said")

wc = WordCloud(background_color="white", max_words=2000, mask=alice_co
               stopwords=stopwords, max_font_size=40, random_state=42)
# generate word cloud
wc.generate(text)

# create coloring from image
image_colors = ImageColorGenerator(alice_coloring)

# show
fig, axes = plt.subplots(1, 3)
axes[0].imshow(wc, interpolation="bilinear")
```

```
# recolor wordcloud and show
# we could also give color_func=image_colors directly in the constructor
axes[1].imshow(wc.recolor(color_func=image_colors), interpolation="bilinear")
axes[2].imshow(alice_coloring, cmap=plt.cm.gray, interpolation="bilinear")
for ax in axes:
    ax.set_axis_off()
plt.show()
```

executed in 3.01s, finished 22:33:14 2019-08-09

3.3.3 // prepare word

In [590]:

```
df.clean_text[:5]
```

executed in 10ms, finished 22:33:17 2019-08-09

Out[590]:

```
time_index
2017-08-01 16:23:56    this is phineas. he's a mystical boy. only eve..
.
2017-08-01 00:17:27    this is tilly. she's just checking pup on you....
2017-07-31 00:18:03    this is archie. he is a rare norwegian pouncin.
..
2017-07-30 15:58:51     this is darla. she commenced a snooze mid
meal.
2017-07-29 16:00:24    this is franklin. he would like you to stop ca...
Name: clean_text, dtype: object
```

In [591]:

```
# 使用 sum 前要删除 null 值, 否则会报错
str_input = df.clean_text.dropna()
str_input.isnull().sum()
# 聚合方式可以参考
# https://stackoverflow.com/questions/47465542/how-to-concatenate-all-s
```

executed in 11ms, finished 22:33:19 2019-08-09

Out[591]:

0

In [592]:

```
text_twitter = str_input.sum()
text_twitter[:1000]
```

executed in 95ms, finished 22:33:20 2019-08-09

Out[592]:

'this is phineas. he\'s a mystical boy. only ever appears in the hole of
 a donut. 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
. this is archie. he is a rare norwegian pouncing corgo. lives in the tal
l grass. you never know when one may strike. this is darla. she comm
enced a snooze mid meal. this is franklin. he would like you to stop c
alling him "cute." he is a very fierce shark and should be respected a
s such. here we have a majestic great white breaching off south afric
a\'s coast. absolutely h*ckin breathtaking. meet jax. he enjoys ice cr
eam so much he gets nervous around it. when you watch your owner
call another dog a good boy but then they turn back to you and say y
ou\'re a great boy. this is zoey. she doesn\'t want to be one of the s
cary sharks. just wants to be a snuggly pettable boatpet. this is cassi
e. she is a college pup. studying international doggo communication
and stick theory. this is koda'

3.3.4 // word cloud

In [593]:

```
# 将图像转为 np 二维数据 (所以是png还是jpeg应该没有关系)
# read the mask / color image taken from
color1 = np.array(Image.open("tweet1.jpeg"))
color2 = np.array(Image.open("tweet2.jpeg"))
color3 = np.array(Image.open("t1.png"))
color4 = np.array(Image.open("t2.png"))

# 设置停用词
stopwords = set(STOPWORDS)
stopwords.add("said")
```

executed in 56ms, finished 22:33:23 2019-08-09

In [594]:

```
# wordcloud 参数
wc = WordCloud(background_color="white", max_words=2000,
               stopwords=stopwords, max_font_size=40, random_state=42)

## https://github.com/amueller/word_cloud
## git 中提供例子和cli(可以根据 text 和 pic 直接输出词云, 非常方便
```

executed in 7ms, finished 22:33:23 2019-08-09

In [595]:

```
# wc.generate(text);
```

executed in 10ms, finished 22:33:24 2019-08-09

In [596]:

```
wc.generate(text_twitter);
```

executed in 471ms, finished 22:33:25 2019-08-09

In [597]:

```
# create coloring from image
image_colors = ImageColorGenerator(color4)

# 可以直接在构造函数中直接给颜色
# 通过这种方式词云将会按照给定的图片颜色布局生成字体颜色策略

# show
fig, axes = plt.subplots(1, 3, figsize=(24,4))
axes[0].imshow(wc, interpolation="bilinear")
# recolor wordcloud and show
# we could also give color_func=image_colors directly in the constructor
axes[1].imshow(wc.recolor(color_func=image_colors), interpolation="bilinear")
axes[2].imshow(color4, cmap=plt.cm.gray, interpolation="bilinear")
for ax in axes:
    ax.set_axis_off();
plt.show();
```

executed in 648ms, finished 22:33:28 2019-08-09

In [598]:

```
# 增加 mask 蒙版系列
wc = WordCloud(background_color="white", max_words=2000, mask=color4,
               stopwords=stopwords, max_font_size=40, random_state=42)
wc.generate(text_twitter);
```

executed in 7.68s, finished 22:33:36 2019-08-09

In [599]:

```
# create coloring from image
image_colors = ImageColorGenerator(color4)

# 可以直接在构造函数中直接给颜色
# 通过这种方式词云将会按照给定的图片颜色布局生成字体颜色策略

# show
fig, axes = plt.subplots(1, 3, figsize=(24,4))
axes[0].imshow(wc, interpolation="bilinear")
# recolor wordcloud and show
# we could also give color_func=image_colors directly in the constructor
axes[1].imshow(wc.recolor(color_func=image_colors), interpolation="bilinear")
axes[2].imshow(color4, cmap=plt.cm.gray, interpolation="bilinear")
for ax in axes:
    ax.set_axis_off();
plt.show();
```

executed in 1.33s, finished 22:33:38 2019-08-09



In [600]:

```
# 输出两个图像做对比
# show
fig, axes = plt.subplots(2, 1, figsize=(20,4))
axes[0].imshow(wc.recolor(color_func=image_colors), interpolation="bilinear")
axes[1].imshow(color4, cmap=plt.cm.gray, interpolation="bilinear")
for ax in axes:
    ax.set_axis_off();
plt.show();
## 不是特别美观, 看来wordcloud如果使用mask和图像的样子关系很大
```

executed in 991ms, finished 22:33:39 2019-08-09

In [601]:

```
# 增加 mask 蒙版系列2
wc = WordCloud(background_color="white", max_words=200, mask=color2,
               stopwords=stopwords, max_font_size=40, random_state=42)
wc.generate(text_twitter);

# create coloring from image
image_colors = ImageColorGenerator(color2)

# 可以直接在构造函数中直接给颜色
# 通过这种方式词云将会按照给定的图片颜色布局生成字体颜色策略

# 输出两个图像做对比
# show
fig, axes = plt.subplots(2, 1)
axes[0].imshow(wc.recolor(color_func=image_colors), interpolation="bilinear")
axes[1].imshow(color2, cmap=plt.cm.gray, interpolation="bilinear")
for ax in axes:
    ax.set_axis_off();
plt.show();
## 不是特别美观, 看来wordcloud如果使用mask和图像的样子关系很大
### 大小和图片分辨率相同
### 遇到有的图片会报错
## 感觉对分词如果用 nltk 处理下可能会更好
### https://sqlshep.com/?p=971
# 更新! relative_scaling 参数特别重要(见结论图)
```

executed in 303ms, finished 22:33:39 2019-08-09

3.4 / time series analysis

https://ourcodingclub.github.io/2019/01/07/pandas-time-series.html
(https://ourcodingclub.github.io/2019/01/07/pandas-time-series.html)

In [602]:

```
1 df.head(10)
```

executed in 17ms, finished 22:33:40 2019-08-09

Out[602]:

time_index	display_text_range	favorite_count	retweet_count	clean_text
2017-08-01 16:23:56	85	39492	8842	this is phineas. he's a mystical boy. only eve...
2017-08-01 00:17:27	138	33786	6480	this is tilly. she's just checking pup on you....
2017-07-31 00:18:03	121	25445	4301	this is archie. he is a rare norwegian pouncin...
2017-07-30 15:58:51	79	42863	8925	this is darla. she commenced a snooze mid meal.
2017-07-29 16:00:24	138	41016	9721	this is franklin. he would like you to stop ca...
2017-07-29 00:08:17	138	20548	3240	here we have a majestic great white breaching ...
2017-07-				meet jax. he enjoys ice

28 16:27:12	140	12053	2142	cream so much he gets ...
2017-07-28 00:22:40	118	66596	19548	when you watch your owner call another dog a g...
2017-07-27 16:25:51	122	28187	4403	this is zoey. she doesn't want to be one of th...
2017-07-26 15:59:51	133	32467	7684	this is cassie. she is a college pup. studying...

In [603]:

```
# check intervals
print("Dataframe shape: ", df.shape)
dt = (df.index[0] - df.index[-1])
print("Number of hours between start and end dates: ", dt.total_seconds()/360
dt
```

executed in 15ms, finished 22:33:42 2019-08-09

```
Dataframe shape:  (2352, 4)
Number of hours between start and end dates:  14994.863333333333
```

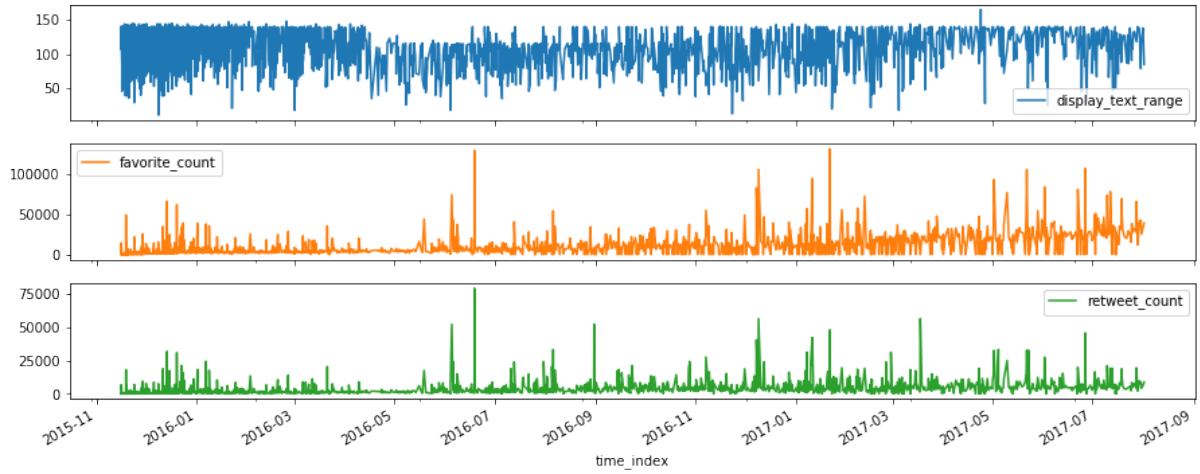
Out[603]:

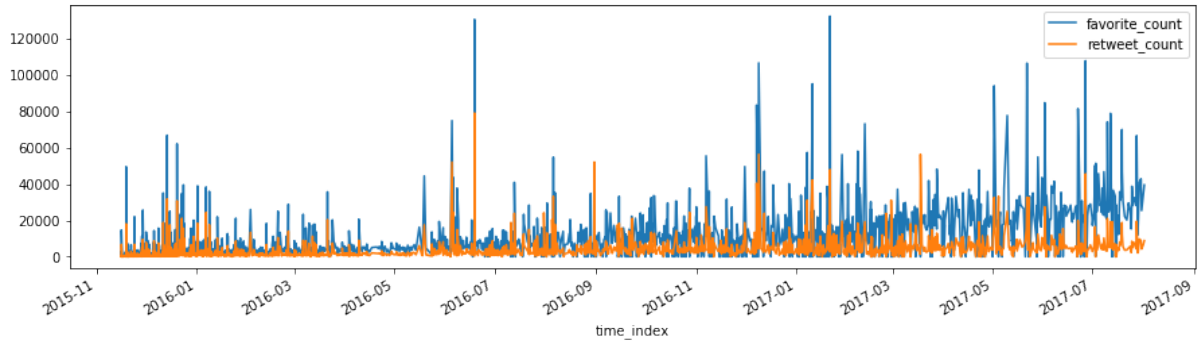
Timedelta('624 days 17:51:48')

In [604]:

```
#df.plot(figsize=(15,4))
df.plot(subplots=True, figsize=(15,6))
df.plot(y=["favorite_count", "retweet_count"], figsize=(15,4));
```

executed in 1.04s, finished 22:33:45 2019-08-09

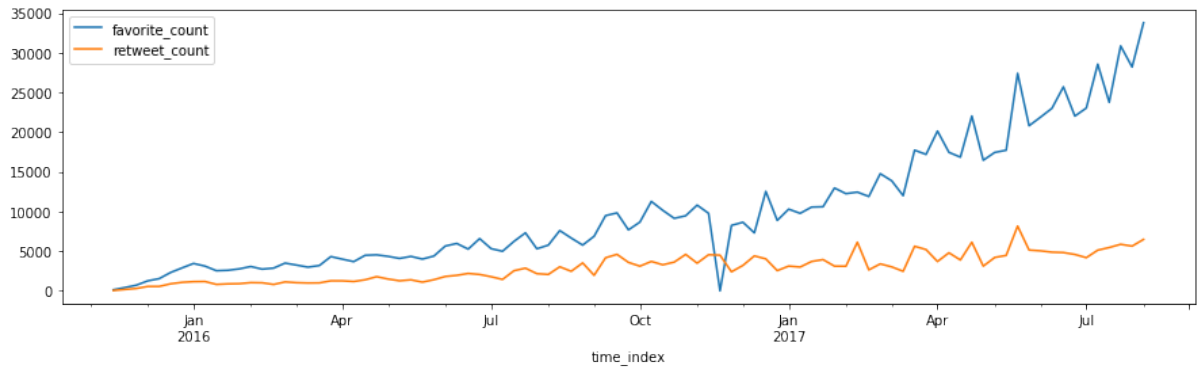




In [605]:

```
df[["favorite_count", "retweet_count"]].resample("1w").median().plot(figsize=(15
```

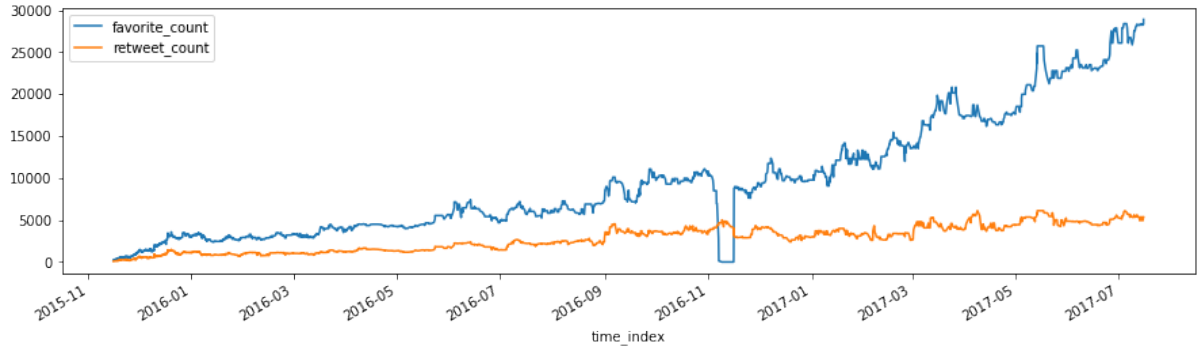
executed in 338ms, finished 22:33:45 2019-08-09



In [606]:

```
1  df[["favorite_count", "retweet_count"]].rolling(30).median().plot(figsize=(15,4));
```

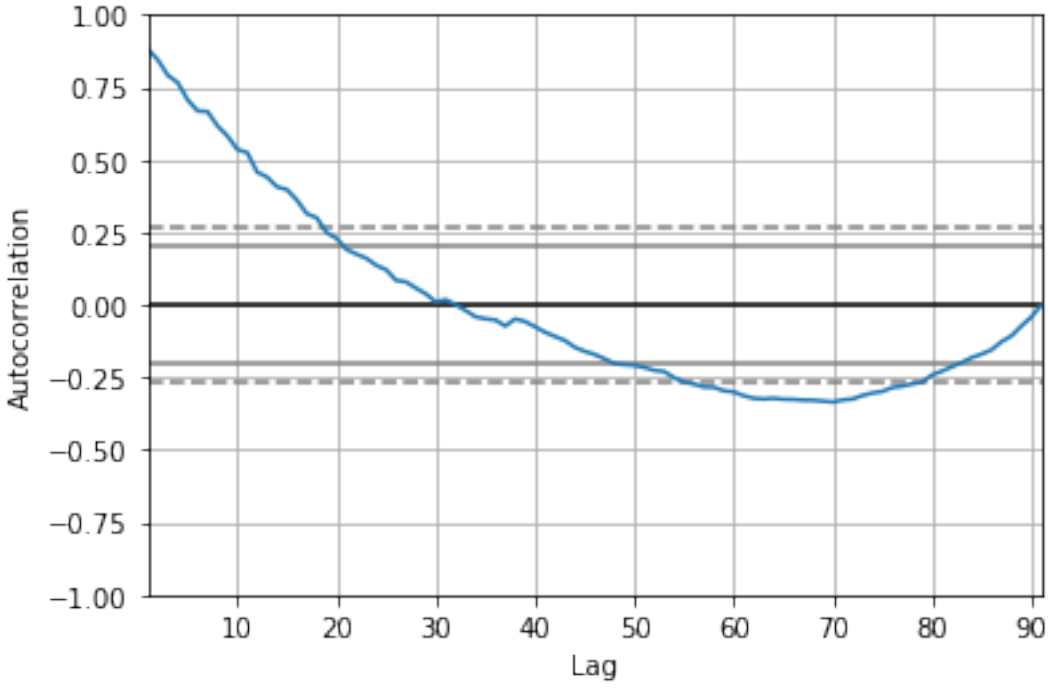
executed in 395ms, finished 22:33:47 2019-08-09



In [607]:

```
1  # 如果是周期的可以用这个(后续研究)
2  pd.plotting.autocorrelation_plot(df["favorite_count"].resample("1w").median());
```

executed in 242ms, finished 22:33:47 2019-08-09



3.5 / sentiment analysis

- 使用sklearn https://towardsdatascience.com/sentiment-analysis-with-python-part-1-5ce197074184 (https://towardsdatascience.com/sentiment-analysis-with-python-part-1-5ce197074184)
- 另外比较常见的是使用 nltk 库
- 此处先pass, 深度学习时候有空再深入

4 结论

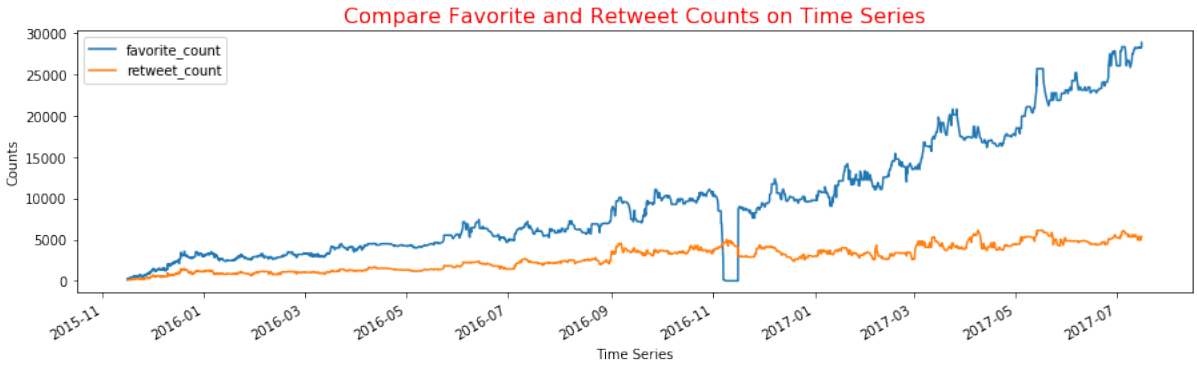
4.1 / favorite 和 retweet 时序分析

- 2016年上半年之前, favorite 数量大概是 retweet 的两倍
- 但再这之后, favorite 数量大量上涨, retweet 数量上涨十分缓慢(两者之比达到6倍)
- 推测相关因素如下:
  - 可以看出 twitter 增长非常迅速(可惜缺少用户量相关的数据)
  - 但是人们愿意付出更多一点时间 retweet 的时间在减少, 可能原因是当人接触到更多的 twitter 信息后, 能够 retweet 的注意力已经没有什么增长空间了(注意力处于饱和状态)

In [608]:

```
# 使用30天滚动平均值完成作图
df[["favorite_count", "retweet_count"]].rolling(30).median().plot(figsize=(15,4));
plt.xlabel('Time Series')
plt.ylabel('Counts')
plt.title('Compare Favorite and Retweet Counts on Time Series', color='r', fonts
```

executed in 317ms, finished 22:33:50 2019-08-09



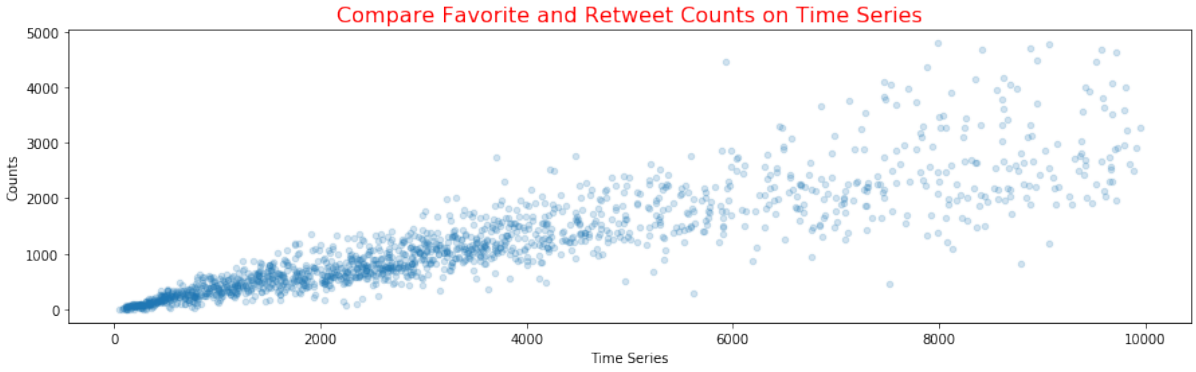
4.2 / favorate 和 retweet 相关性分析

- 分析中过滤掉了 retweet 为0的数据和大于1000的数据
- 此处考虑的是两个参数的对应关系, 和问题1的趋势并不冲突(因为数据做了过滤)
- 可以看出在 favorate 和 retweet 两个数据中间具有相关性
- 回归线要用到 sm 库或 sklearn 库, 后续研究
  https://nbviewer.jupyter.org/github/weecology/progbio/blob/master/ipynbs/statistics.i
  (https://nbviewer.jupyter.org/github/weecology/progbio/blob/master/ipynbs/statistics.

In [609]:

```
1  df.query('0 < favorite_count < 10000').plot.scatter(y,z,alpha=0.2,figsize=(15,4));
2  plt.xlabel('Time Series')
3  plt.ylabel('Counts')
4  plt.title('Compare Favorite and Retweet Counts on Time Series', color='r', fonts
```

executed in 241ms, finished 22:33:52 2019-08-09



4.3 / word cloud 分析

- 对评论使用 word cloud 进行分析
- 去掉了 stop words
- 图像为 twitter 英文字符(小鸟图不太美观)

In [610]:

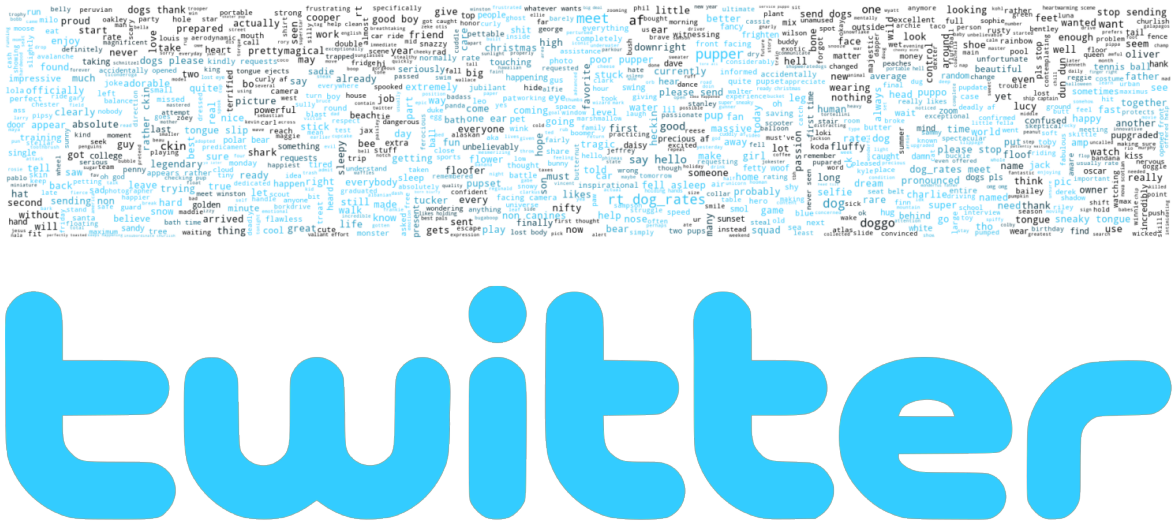
```
# set wc paras
wc = WordCloud(background_color="white", max_words=1000, mask=color4,
               stopwords=stopwords, max_font_size=24, relative_scaling=0.3, widt

# gen wc
wc.generate(text_twitter);

# create coloring from image
image_colors = ImageColorGenerator(color4)

# gen pic
fig, axes = plt.subplots(2, 1,figsize=(36,12))
axes[0].imshow(wc.recolor(color_func=image_colors), interpolation="bilinear")
axes[1].imshow(color4, cmap=plt.cm.gray, interpolation="bilinear")
for ax in axes:
    ax.set_axis_off();
plt.show();
```

executed in 9.91s, finished 22:34:03 2019-08-09



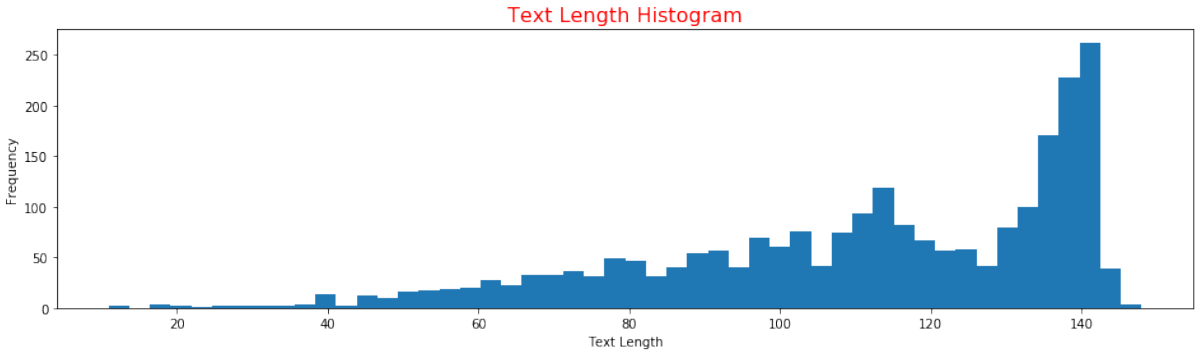
4.4 / text range 分析

- text range 改名为 text range 更为明确
- 数据做了过滤(过滤掉了个别 160 字符的)
- 数据有左偏斜趋势 (不能断定) 因为在140字的限制上有大量出现, 所以明显存在人为调整
- 有些数据超出了140
- 后续可以做异常值分析(按说不应该有超出, 也可能是正则化过滤时留下的问题)

In [611]:

```
1  df.query('display_text_range < 150').display_text_range.plot.hist(bins=50,figsize
2  plt.xlabel('Text Length')
3  plt.title('Text Length Histogram', color='r', fontsize=16);
```

executed in 377ms, finished 22:34:03 2019-08-09



4.5 / 后续完善

- 增加数据feature: 虽然原始数据 featrue 比较多, 但经过梳理发现所剩数据不多. 像用户日活, 注册量等信息缺失.
- 完善情感分析: 情感分析可以画出 积极/消极/主观/客观 两个维度的信息. 便于增加数据用以更多分析 (比如 140字的回复中, 是积极信息多还是消极信息多)
- 完善 source 分类数据: 本来很关注的feature, 因为数据收集的问题(可能是数据收集时ios比较好记录), 这点非常重要, 因为起码从尝试来讲 android 的不应该这么少. 这种情况会造成数据偏见, 可能带来错误的结论

In [ ]:

```
1
```

收集

/ import lib

In [524]:

In [525]:

/ load df

In [526]:

In [527]:

/ check df

In [528]:

```
----checking sample index: 187

- columns #1 : contributors
[nan]

- columns #2 : coordinates
[nan]

- columns #3 : created_at
['2017-04-22T18:55:51.000000000']

- columns #4 : display_text_range
[list([0, 89])]

- columns #5 : entities
[{'hashtags': [], 'symbols': [], 'user_mentions': [], 'urls': [{'url': 'https:
//t.co/sb73bV5Y7S', 'expanded_url': 'https://twitter.com/perfy/sta
tus/855857318168150016', 'display_url': 'twitter.com/perfy/status/
8…', 'indices': [90, 113]}]}]
```

In [529]:

Out[529]:

```
Index(['contributors', 'coordinates', 'created_at', 'display_text_range',
       'entities', 'extended_entities', 'favorite_count', 'favorited',
       'full_text', 'geo', 'id', 'id_str', 'in_reply_to_screen_name',
       'in_reply_to_status_id', 'in_reply_to_status_id_str',
       'in_reply_to_user_id', 'in_reply_to_user_id_str', 'is_quote_status',
       'lang', 'place', 'possibly_sensitive', 'possibly_sensitive_appealable
',
       'quoted_status', 'quoted_status_id', 'quoted_status_id_str',
       'retweet_count', 'retweeted', 'retweeted_status', 'source', 'trunca
ted',
       'user'],
      dtype='object')
```

In [530]:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2352 entries, 0 to 2351
Data columns (total 31 columns):
contributors                  0 non-null float64
coordinates                   0 non-null float64
created_at                    2352 non-null datetime64[ns]
display_text_range            2352 non-null object
entities                      2352 non-null object
extended_entities             2073 non-null object
favorite_count                2352 non-null int64
favorited                     2352 non-null bool
full_text                     2352 non-null object
geo                           0 non-null float64
id                            2352 non-null int64
id_str                        2352 non-null int64
in_reply_to_screen_name       78 non-null object
in_reply_to_status_id         78 non-null float64
in_reply_to_status_id_str     78 non-null float64
in_reply_to_user_id           78 non-null float64
in_reply_to_user_id_str       78 non-null float64
```

In [531]:

Out[531]:

	contributors	coordinates	favorite_count	geo	id	
count	0.0	0.0	2352.000000	0.0	2.352000e+03	2.352
mean	NaN	NaN	8109.198980	NaN	7.425913e+17	7.425
std	NaN	NaN	11980.795669	NaN	6.846210e+16	6.846
min	NaN	NaN	0.000000	NaN	6.660209e+17	6.660
25%	NaN	NaN	1417.000000	NaN	6.783949e+17	6.783
50%	NaN	NaN	3596.500000	NaN	7.193536e+17	7.193
75%	NaN	NaN	10118.000000	NaN	7.991219e+17	7.991
max	NaN	NaN	132318.000000	NaN	8.924206e+17	8.924

评估

/ quanlity

// drop1

In [532]:

In [533]:

```
---- proceding ----
- drop 8 columns: ['contributors', 'coordinates', 'geo', 'place', 'id_str',
'in_reply_to_status_id_str', 'in_reply_to_user_id_str', 'quoted_status_id
_str']
- remain 23 columns
- success : True
```

// check - inspect list

对一些怀疑是否有用的数据进行检视

In [534]:

```
----checking sample index: 1735

- columns #1 : created_at
['2015-12-23T03:58:25.000000000']

- columns #2 : display_text_range
[list([0, 133])]

- columns #3 : entities
[{'hashtags': [], 'symbols': [], 'user_mentions': [], 'urls': [], 'media': [{'
id': 679511347441328128, 'id_str': '679511347441328128', 'indices':
[110, 133], 'media_url': 'http://pbs.twimg.com/media/CW4b-GUWY
AAa8QO.jpg', 'media_url_https': 'https://pbs.twimg.com/media/CW
4b-GUWYAAa8QO.jpg', 'url': 'https://t.co/bwuV6FlRxr', 'display_url
': 'pic.twitter.com/bwuV6FlRxr', 'expanded_url': 'https://twitter.com
/dog_rates/status/679511351870550016/photo/1', 'type': 'photo',
'sizes': {'medium': {'w': 505, 'h': 639, 'resize': 'fit'}, 'large': {'w': 505,
'h': 639, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize': 'crop'},
'small': {'w': 505, 'h': 639, 'resize': 'fit'}}}]}]
```

In [535]:

In [536]:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2352 entries, 0 to 2351
Data columns (total 11 columns):
favorited                   2352 non-null bool
in_reply_to_screen_name      78 non-null object
in_reply_to_status_id        78 non-null float64
in_reply_to_user_id          78 non-null float64
is_quote_status             2352 non-null bool
lang                        2352 non-null object
possibly_sensitive          2211 non-null float64
possibly_sensitive_appealable  2211 non-null float64
quoted_status_id              29 non-null float64
retweeted                   2352 non-null bool
truncated                   2352 non-null bool
dtypes: bool(4), float64(5), object(2)
memory usage: 137.9+ KB
```

In [537]:

```
- columns #1 : favorited
False    2352
Name: favorited, dtype: int64

- columns #2 : in_reply_to_screen_name
dog_rates        47
markhoppus        2
jonnysun          1
xianmcguire       1
ComplicitOwl      1
Name: in_reply_to_screen_name, dtype: int64

- columns #3 : in_reply_to_status_id
6.671522e+17    2
8.562860e+17    1
8.131273e+17    1
6.754971e+17    1
6.827884e+17    1
Name: in_reply_to_status_id, dtype: int64
```

// check - quoted_status

In [538]:

Out[538]:

```
True     2324
False      28
Name: quoted_status, dtype: int64
```

In [539]:

Out[539]:


```
{'created_at': 'Wed Apr 27 01:34:44 +0000 2016',
 'id': 725136065078521856,
 'id_str': '725136065078521856',
 'full_text': 'Se nos metió otro jugador al partido de @dvotachira vs
 @pumasmx en la #LibertadoresEnFD 🐶\nhttps://t.co/nPtdOeTxc
W',
 'truncated': False,
 'display_text_range': [0, 113],
 'entities': {'hashtags': [{'text': 'LibertadoresEnFD', 'indices': [70, 87]
}],
  'symbols': [],
  'user_mentions': [{'screen_name': 'DvoTachira',
    'name': 'Deportivo Táchira FC',
    'id': 85361349,
    'id_str': '85361349',
    'indices': [40, 51]},
   {'screen_name': 'PumasMX',
    'name': 'PUMAS',
```


- 分析 quoted_status :
  - 是嵌套字典数据
  - 缺失很多(只有28个数据)
  - 内容无用信息比较多
- 结论:
  - 删除此列
  - user 列和此列类似,也删除

// check - in_reply_to_screen_name

In [541]:

Out[541]:

	created_at	display_text_range	entities	extended_entities
147	2017-05-12 17:12:53	[0, 139]	{'hashtags': [], 'symbols': [], 'user_mentions...	{'media': [{'id': 863079538779013120, 'id_str'...
181	2017-04-24 15:13:52	[0, 112]	{'hashtags': [], 'symbols': [], 'user_mentions...	{'media': [{'id': 856526604033556482, 'id_str'...
225	2017-04-01 16:41:12	[0, 135]	{'hashtags': [], 'symbols': [], 'user_mentions...	NaN

3 rows × 23 columns

In [542]:

Out[542]:

0.9668367346938775

- 分析 in_reply_to_screen_name :
  - 可能 dog_rates 是默认回复名字
  - 数据缺失率为 97%
- 结论:
  - 删除数据

// drop2

根据上面 check 内容删除数据

In [543]:

In [544]:

```
---- proceding ----
- drop 14 columns: ['favorited', 'in_reply_to_screen_name', 'in_reply_t
o_status_id', 'in_reply_to_user_id', 'is_quote_status', 'lang', 'possibly_s
ensitive', 'possibly_sensitive_appealable', 'quoted_status_id', 'retweet
ed', 'truncated', 'quoted_status', 'retweeted_status', 'user']
- remain 9 columns
- success : True
```

// check - detail columns

对嵌套的数据进行检视

In [545]:

```
----checking sample index: 316

- columns #1 : created_at
['2017-02-22T18:59:48.000000000']

- columns #2 : display_text_range
[list([0, 139])]

- columns #3 : entities
[{'hashtags': [], 'symbols': [], 'user_mentions': [{'screen_name': 'dog
_rates', 'name': 'SpookyWeRateDogs™', 'id': 4196983835, 'id_str': '
4196983835', 'indices': [3, 13]}], 'urls': []}]

- columns #4 : extended_entities
[nan]

- columns #5 : favorite_count
[0]
```

In [546]:

In [547]:

```
– columns #1 : entities
{'hashtags': [],
 'media': [{'display_url': 'pic.twitter.com/MgUWQ76dJU',
            'expanded_url': 'https://twitter.com/dog_rates/status/892
420643555336193/photo/1',
            'id': 8924206394868776 96,
            'id_str': '8924206394868776 96',
            'indices': [86, 109],
            'media_url': 'http://pbs.twimg.com/media/DGKD1–bXoAAI
AUK.jpg',
            'media_url_https': 'https://pbs.twimg.com/media/DGKD1–
bXoAAIAUK.jpg',
            'sizes': {'large': {'h': 528, 'resize': 'fit', 'w': 540},
                      'medium': {'h': 528, 'resize': 'fit', 'w': 540},
                      'small': {'h': 528, 'resize': 'fit', 'w': 540},
                      'thumb': {'h': 150, 'resize': 'crop', 'w': 150}},
            'type': 'photo',
            'url': 'https://t.co/MgUWQ76dJU'}],
```

- 分析:
  - 是嵌套字典数据
  - 缺失不多
  - 内容无用信息比较多(有些与其他列有重复)
- 结论:
  - 删除列

// drop3

In [548]:

```
```

---- proceding ----
– drop 2 columns: ['entities', 'extended_entities']
– remain 7 columns
– success : True

// check - display_text_range

使用函数check_value会在这一列报错,检查下是因为这列的列表嵌套数字的原因

In [549]:

```
```

Out[549]:

```
281     [0, 112]
2236    [0, 139]
68      [0, 132]
1516    [0, 108]
717     [0, 107]
Name: display_text_range, dtype: object
```

In [550]:

```
```

In [551]:

```
– columns #1 : created_at
2016–09–12 15:10:21    1
2016–06–03 01:07:16    1
2017–01–31 01:27:39    1
2016–10–13 23:23:56    1
2016–06–27 01:37:04    1
Name: created_at, dtype: int64

– columns #2 : favorite_count
0       177
1753      3
3548      3
689       3
1526      3
Name: favorite_count, dtype: int64

– columns #3 : full_text
Three generations of pupper. 11/10 for all https://t.co/tAmQYvzrau
```

// check - null data

In [552]:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2352 entries, 0 to 2351
Data columns (total 7 columns):
created_at              2352 non-null datetime64[ns]
display_text_range      2352 non-null object
favorite_count          2352 non-null int64
full_text               2352 non-null object
id                      2352 non-null int64
retweet_count           2352 non-null int64
source                  2352 non-null object
dtypes: datetime64[ns](1), int64(3), object(3)
memory usage: 128.7+ KB
```

In [553]:

Out[553]:

```
created_at              0
display_text_range      0
favorite_count          0
full_text               0
id                      0
retweet_count           0
source                  0
dtype: int64
```

// drop4

想来想去还是把 id 给drop了, 后续分析中用不到还有隐私隐患

In [554]:

```
```

---- proceding ----
– drop 1 columns: ['id']
– remain 6 columns
– success : True

// review (quanlity)

根据数据删除剩余7列 ['created_at', 'display_text_range', 'favorite_count', 'full_text', 'id', 'retweet_count', 'source']

- id 为标识列
- created_at 包括时间、日期,可以进行时序分析
- display_text_range 为文字长度
- favorite_count 为点赞数
- full_text 为文字内容
- retweet_count 为回复数
- source 为来源

// persistence

In [555]:

```
```

/ tidyness

根据质量部分的输出,对于除id列之外的需要进行清洁度的整理

- created_at 包括时间、日期,可以进行时序分析
  - 转换为 datafame 的 datetime 格式
- display_text_range 为文字长度
  - 原格式为 [0-x] x实际为推文长度,需要提取 x, 有个别是 [x-y], 不知道为什么还有下限, 提取上限数据即可
  - 本列为非必须列,可以根据 full_text 得出回复长度
- favorite_count 为点赞数
  - 数字类型,无需转换
- full_text 为文字内容
  - 后续如果进行nlp的分析需要进行向量化
- retweet_count 为回复数
  - 数字类型,无需转换
- source 为来源
  - 来源为链接,中间为发布信息的设备
  - 需要使用 re 来完成提取
  - 最后输出为分类信息

/ load clean df

In [556]:

Out[556]:

	created_at	display_text_range	favorite_count	full_text	retweet_cou
1636	2016-01-04 23:02:22	[0, 126]	3250	This is Sweets the English Bulldog. Waves back...	169

// created_at

define: 将数据转换为时间格式

- solution1 使用 dataframe 的 datatime 格式
  - 数据本身为 datetime 格式
  - 如果是时序的数据可以将时间转换为 index,非常方便筛选 https://chrisalbon.com/python/data_wrangling/pandas_time_series_basics/ (https://chrisalbon.com/python/data_wrangling/pandas_time_series_basics/)
- (solution2 使用 python datatime 格式、calendar格式)

In [557]:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2352 entries, 0 to 2351
Data columns (total 6 columns):
created_at              2352 non-null datetime64[ns]
display_text_range      2352 non-null object
favorite_count          2352 non-null int64
full_text               2352 non-null object
retweet_count           2352 non-null int64
source                  2352 non-null object
dtypes: datetime64[ns](1), int64(2), object(3)
memory usage: 110.3+ KB
```

In [558]:

Out[558]:

	created_at	display_text_range	favorite_count	full_text	retweet_cou
2258	2015-11-20 03:35:20	[0, 140]	350	Here is George. George took a selfie of his ne...	13

In [559]:

Out[559]:

(2258 11
Name: created_at, dtype: int64, 2258 20
Name: created_at, dtype: int64, 2258 3
Name: created_at, dtype: int64)

In [560]:

Out[560]:

	created_at	display_text_range	favorite_count	full_text	retweet_cc
	2015-11-			Here is the Rand	

2342	16 01:01:59	[0, 135]	117	Paul of retrievers folks! He'...
2343	2015-11-16 00:55:59	[0, 124]	304	My oh my. This is a rare blond Canadian terrie...
2344	2015-11-16 00:49:46	[0, 140]	449	Here is a Siberian heavily armored polar bear ...
2345	2015-11-16 00:35:11	[0, 138]	1250	This is an odd dog. Hard on the outside but lo...
2346	2015-11-16 00:30:50	[0, 140]	136	This is a truly beautiful English Wilson Staff...
2347	2015-11-16 00:24:50	[0, 120]	111	Here we have a 1949 1st generation vulpix. Enj...
2348	2015-11-16 00:04:52	[0, 137]	309	This is a purebred Piers Morgan. Loves to Netf...
2349	2015-11-15 23:21:54	[0, 130]	128	Here is a very happy pup. Big fan of well-main...
2350	2015-11-15 23:05:30	[0, 139]	132	This is a western brown Mitsubishi terrier. Up...

2351	2015-11-15 22:32:08	[0, 131]	2528	Here we have a Japanese Irish Setter. Lost eye...

In [561]:

In [562]:

---- proceding ----
– drop 1 columns: ['created_at']
– remain 5 columns
– success : True

In [563]:

Out[563]:

time_index	display_text_range	favorite_count	full_text	retweet_count	
2017-01-01 19:22:38	[0, 100]	9130	This is Titan. His nose is quite chilly. Reque...	1901	hre
2017-01-01 02:53:20	[0, 44]	11423	Happy New Year from the squad! 13/10 for all h...	4388	hre

// display_text_range

define: 抽取出 text 的长度,存为整数

- solution1 使用 python standard re lib
  - 抽出字符
  - 转换为 int
- https://pandas.pydata.org/pandas-docs/stable/user_guide/text.html (https://pandas.pydata.org/pandas-docs/stable/user_guide/text.html) 非常全面的介绍

In [564]:

```
```

Out[564]:

```
time_index
2017-08-01 16:23:56     [0, 85]
2017-08-01 00:17:27    [0, 138]
2017-07-31 00:18:03    [0, 121]
2017-07-30 15:58:51     [0, 79]
2017-07-29 16:00:24    [0, 138]
2017-07-29 00:08:17    [0, 138]
2017-07-28 16:27:12    [0, 140]
2017-07-28 00:22:40    [0, 118]
2017-07-27 16:25:51    [0, 122]
2017-07-26 15:59:51    [0, 133]
Name: display_text_range, dtype: object
```

In [565]:

```
```

In [566]:

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 2352 entries, 2017-08-01 16:23:56 to 2015-11-15 22:
32:08
Data columns (total 5 columns):
display_text_range    2352 non-null int64
favorite_count        2352 non-null int64
full_text             2352 non-null object
retweet_count         2352 non-null int64
source                2352 non-null object
dtypes: int64(3), object(2)
memory usage: 110.2+ KB
```

In [567]:

Out[567]:

	display_text_range	favorite_count	retweet_count
count	2352.000000	2352.000000	2352.000000
mean	111.179847	8109.198980	3134.932398
std	27.364336	11980.795669	5237.846296
min	11.000000	0.000000	0.000000
25%	93.000000	1417.000000	618.000000
50%	116.000000	3596.500000	1456.500000
75%	137.000000	10118.000000	3628.750000
max	165.000000	132318.000000	79116.000000

// full_text

define:

- 每个评价后面都有一个分值和链接 11/10 https://t.co/8W5iSOgXfx (https://t.co/8W5iSOgXfx)
- 评分为 10/10 或 11/10,没找到说明, 链接科学上网也不能访问
- 需要删除后保存
- 此处不做处理,词云的制作最后再做
- try solution
  - str.replace
  - str[i]
  - str.extract(r'ab (%5Cd))
  - pat = / str.match
  - str.contains
  - get.dummies(sep=',')

In [568]:

Out[568]:

time_index
2017-08-01 16:23:56 This is Phineas. He's a mystical boy. Only eve...
2017-08-01 00:17:27 This is Tilly. She's just checking pup on you....
2017-07-31 00:18:03 This is Archie. He is a rare Norwegian Pouncin...
2017-07-30 15:58:51 This is Darla. She commenced a snooze mid meal...
2017-07-29 16:00:24 This is Franklin. He would like you to stop ca...
2017-07-29 00:08:17 Here we have a majestic great white breaching ...
2017-07-28 16:27:12 Meet Jax. He enjoys ice cream so much he gets ...
2017-07-28 00:22:40 When you watch your owner call another dog a g...
2017-07-27 16:25:51 This is Zoey. She doesn't want to be one o

In [569]:

Out[569]:

"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 htt
ps://t.co/0Xxu71qelV" (https://t.co/0Xxu71qelV")

In [570]:

Out[570]:

	0	1
0	a	1
1	b	2
2	NaN	NaN

In [571]:

In [572]:

Out[572]:

'this is dewey (pronounced "covfefe"). he\'s having a good walk. argu
ably the best walk. 13/10 would snug softly https://t.co/hcieajkc4d'
(https://t.co/hcieajkc4d')

In [573]:

Out[573]:

	0
time_index	
2017-08-01 16:23:56	13/10
2017-08-01 00:17:27	13/10
2017-07-31 00:18:03	12/10
2017-07-30 15:58:51	13/10
2017-07-29 16:00:24	12/10

In [574]:

Out[574]:

	0	1
time_index		
2017-08-01 16:23:56	this is phineas. he's a mystical boy. only eve...	13/10
2017-08-01 00:17:27	this is tilly. she's just checking pup on you....	13/10
2017-07-31 00:18:03	this is archie. he is a rare norwegian pouncin...	12/10
2017-07-30 15:58:51	this is darla. she commenced a snooze mid meal.	13/10
2017-07-29 16:00:24	this is franklin. he would like you to stop ca...	12/10

In [575]:

Out[575]:

```
time_index
2017-08-01 16:23:56    this is phineas. he's a mystical boy. only eve..
.
2017-08-01 00:17:27    this is tilly. she's just checking pup on you....
2017-07-31 00:18:03    this is archie. he is a rare norwegian pouncin.
..
2017-07-30 15:58:51      this is darla. she commenced a snooze mid
meal.
2017-07-29 16:00:24    this is franklin. he would like you to stop ca...
Name: 0, dtype: object
```

In [576]:

In [577]:

```
---- proceding ----
- drop 1 columns: ['full_text']
- remain 5 columns
- success : True
```

// source

define: 抽取出发 tweet 使用的设备

- 信息是这样的 Twitter for iPhone
- 需要抽取出 `Twitter for iPhone`, 并定义分类为 iphone
- 将本列做成分类数据
- 更新
  - 根据value_counts的输出, 95%的来源为iphone, 失去分析价值(Android的去哪里了)
  - 不过起码说明移动的登陆要比网页多很多

In [578]:

Out[578]:

```
<a href="http://twitter.com/download/iphone" rel="nofollow">Twitte
r for iPhone</a>      2217
<a href="http://vine.co" rel="nofollow">Vine – Make a Scene</a>
91
<a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>
33
<a href="https://about.twitter.com/products/tweetdeck" rel="nofollo
w">TweetDeck</a>       11
Name: source, dtype: int64
```

In [579]:

```
---- proceding ----
- drop 1 columns: ['source']
- remain 4 columns
- success : True
```

// persistence

In [580]:

探索

/ load df

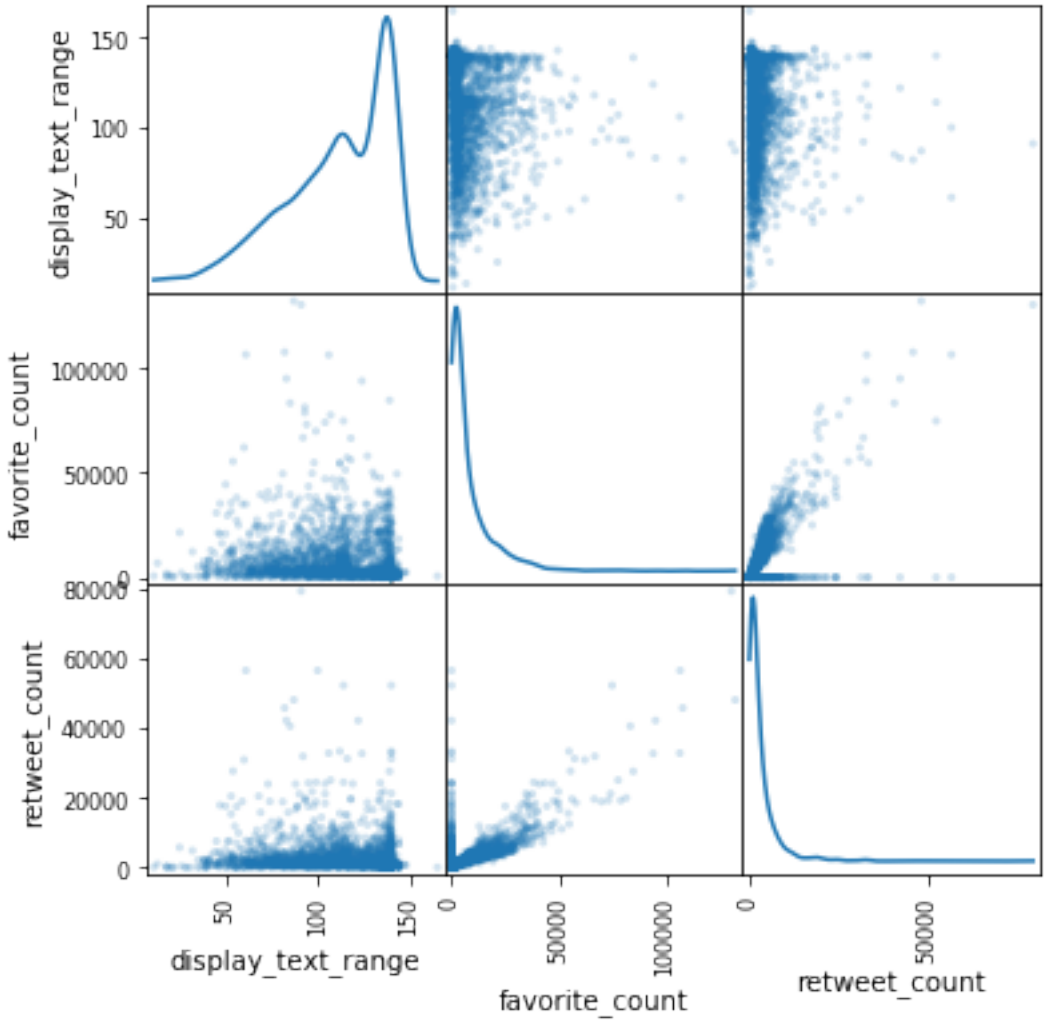
In [581]:

Out[581]:

time_index	display_text_range	favorite_count	retweet_count	clean_text
2015-11-30 01:10:04	140	795	212	NaN

/ data visulization

In [582]:

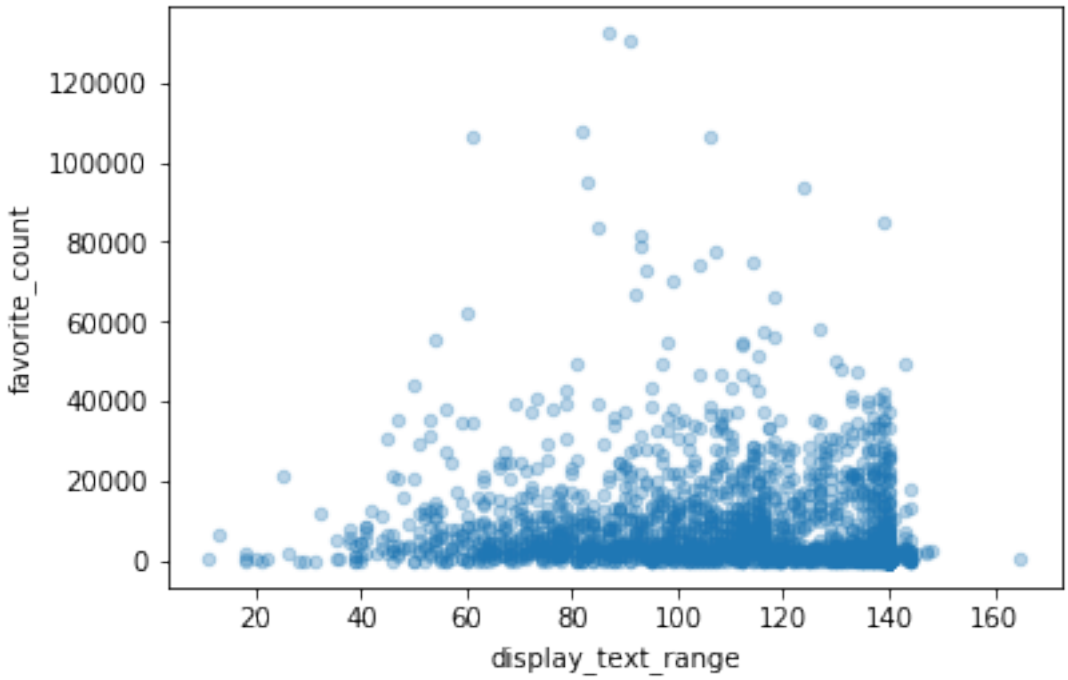


In [583]:

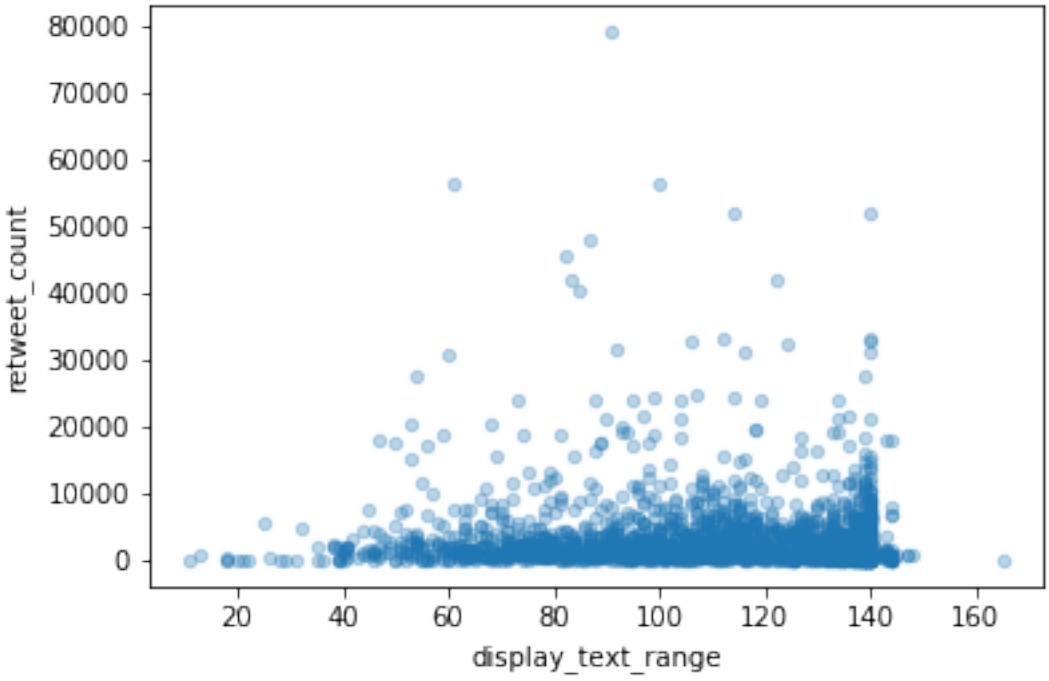
Out[583]:

('display_text_range', 'favorite_count', 'retweet_count')

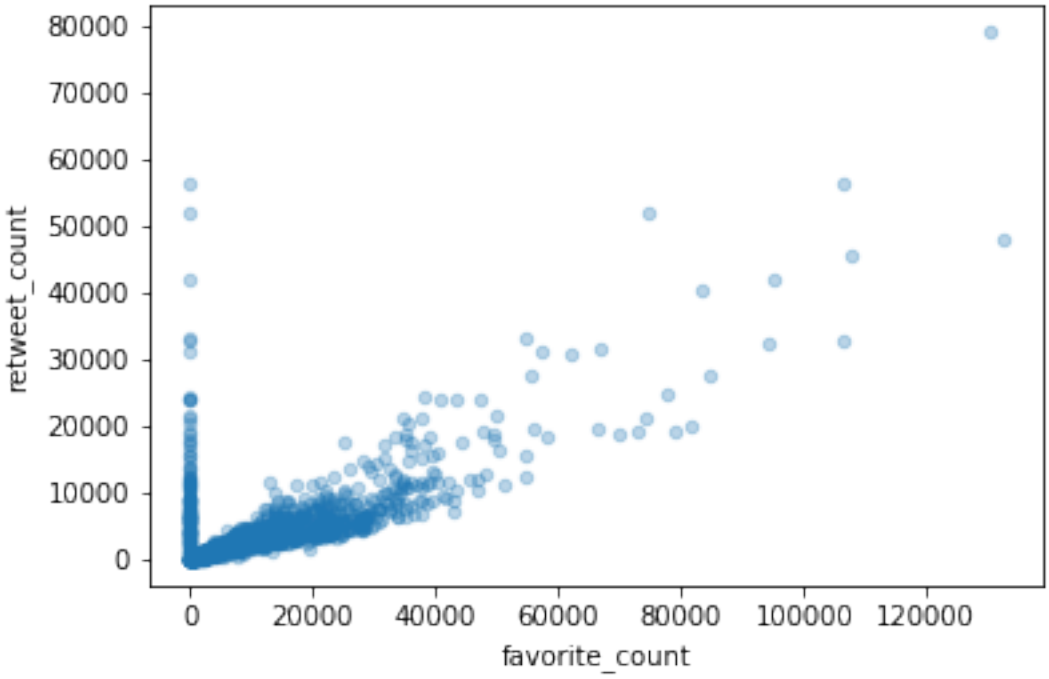
In [584]:



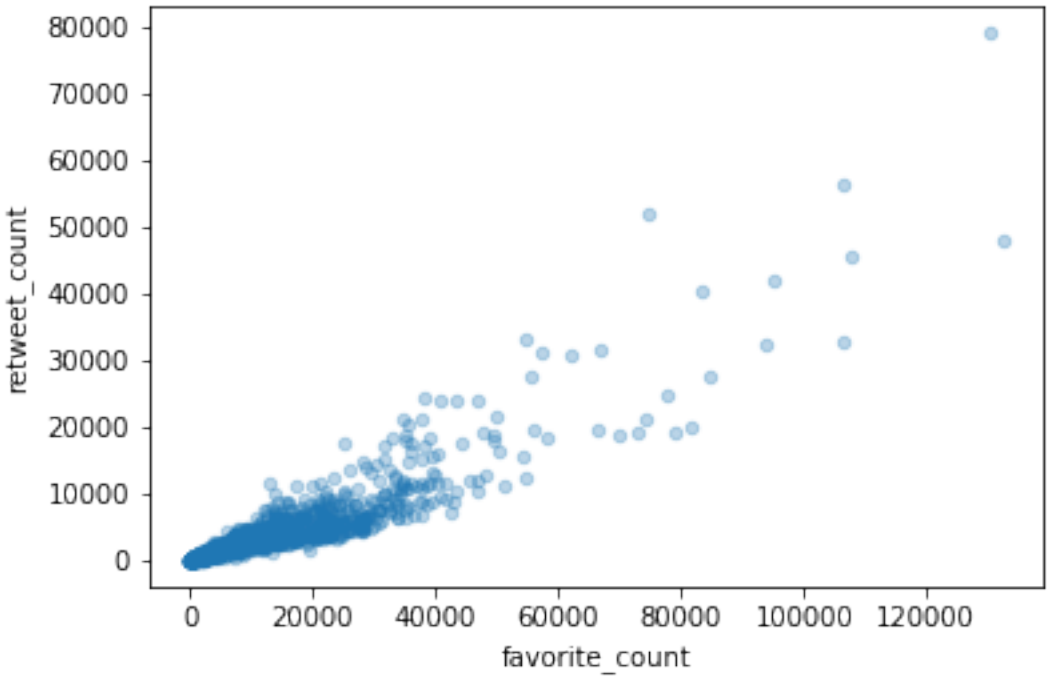
In [585]:



In [586]:



In [587]:



/ word cloud

// word cloud library

In [588]:

// word cloud official example

In [589]:

// prepare word

In [590]:

Out[590]:

```
time_index
2017-08-01 16:23:56     this is phineas. he's a mystical boy. only eve..
.
2017-08-01 00:17:27     this is tilly. she's just checking pup on you....
2017-07-31 00:18:03     this is archie. he is a rare norwegian pouncin.
..
2017-07-30 15:58:51      this is darla. she commenced a snooze mid
meal.
2017-07-29 16:00:24     this is franklin. he would like you to stop ca...
Name: clean_text, dtype: object
```

In [591]:

Out[591]:

0

In [592]:

Out[592]:

'this is phineas. he\'s a mystical boy. only ever appears in the hole of
 a donut. 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
. this is archie. he is a rare norwegian pouncing corgo. lives in the tal
l grass. you never know when one may strike. this is darla. she comm
enced a snooze mid meal. this is franklin. he would like you to stop c
alling him "cute." he is a very fierce shark and should be respected a
s such. here we have a majestic great white breaching off south afric
a\'s coast. absolutely h*ckin breathtaking. meet jax. he enjoys ice cr
eam so much he gets nervous around it. when you watch your owner
call another dog a good boy but then they turn back to you and say y
ou\'re a great boy. this is zoey. she doesn\'t want to be one of the s
cary sharks. just wants to be a snuggly pettable boatpet. this is cassi
e. she is a college pup. studying international doggo communication
and stick theory. this is koda'

// word cloud

In [593]:

In [594]:

In [595]:

In [596]:

In [597]:

In [598]:

In [599]:

In [600]:

In [601]:

/ time series analysis

https://ourcodingclub.github.io/2019/01/07/pandas-time-series.html
(https://ourcodingclub.github.io/2019/01/07/pandas-time-series.html)

In [602]:

Out[602]:

time_index	display_text_range	favorite_count	retweet_count	clean_te
2017-08-01 16:23:56	85	39492	8842	this phinea he's mystic boy. on eve
2017-08-01 00:17:27	138	33786	6480	this is till she's ju checkir pup c you.
2017-07-31	121	25445	4301	this archie. he a ra

00:18:03				norwegia pouncin this is darl
2017-07-30 15:58:51	79	42863	8925	sh commence a snooz mid mea
2017-07-29 16:00:24	138	41016	9721	this franklin. h would lik you to sto ca
2017-07-29 00:08:17	138	20548	3240	here w have majest great whi breaching
2017-07-28 16:27:12	140	12053	2142	meet jax. h enjoys ic cream s much h gets
2017-07-28 00:22:40	118	66596	19548	when yc watch yo owner ca another dc a g
2017-07-27 16:25:51	122	28187	4403	this is zoe she doesr want to b one of th
2017-07-26 15:59:51	133	32467	7684	this cassie. sh is a colleg pu studying

In [603]:

```
```

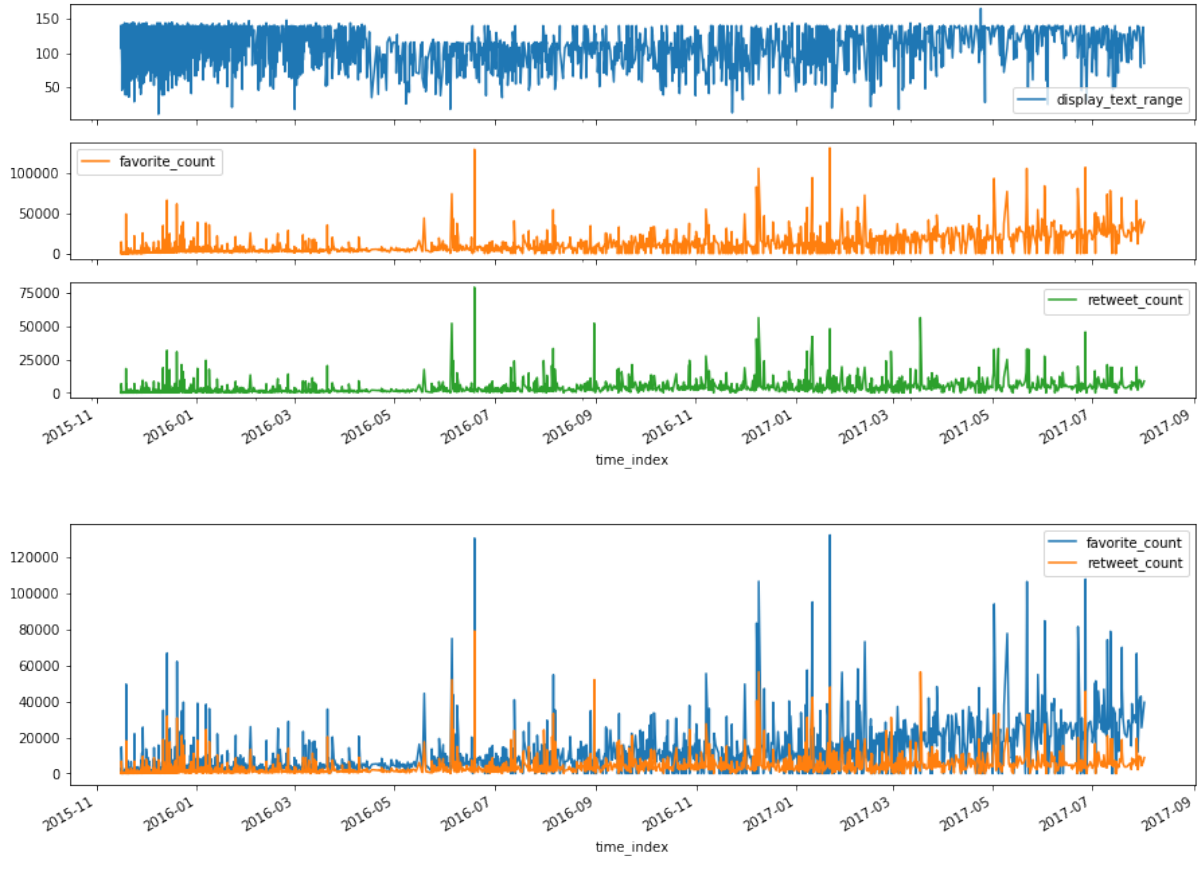
Dataframe shape: (2352, 4)
Number of hours between start and end dates: 14994.863333333333

Out[603]:

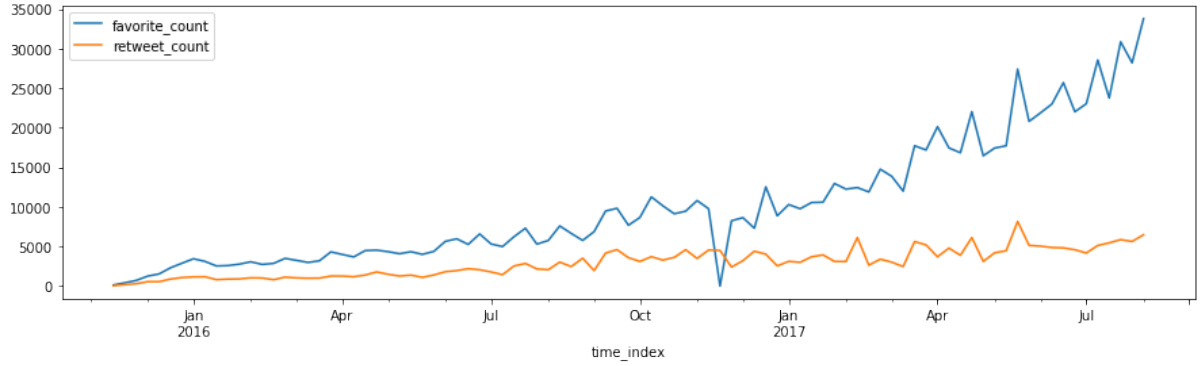
Timedelta('624 days 17:51:48')

In [604]:

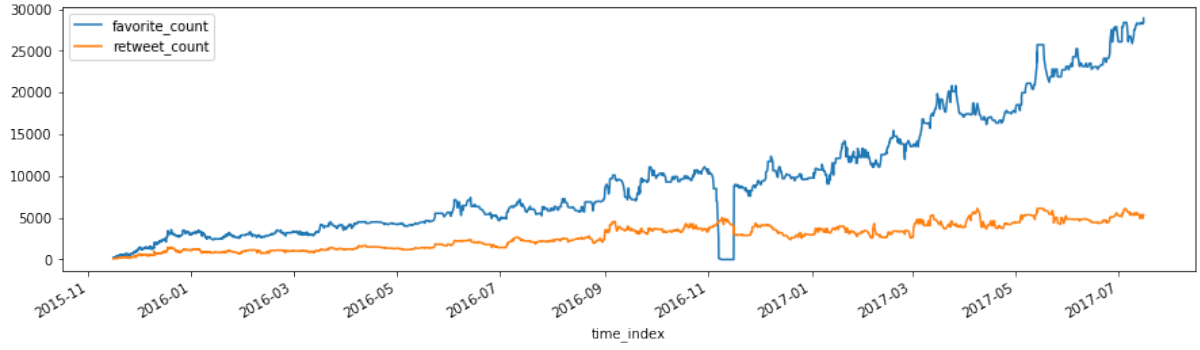
```
```



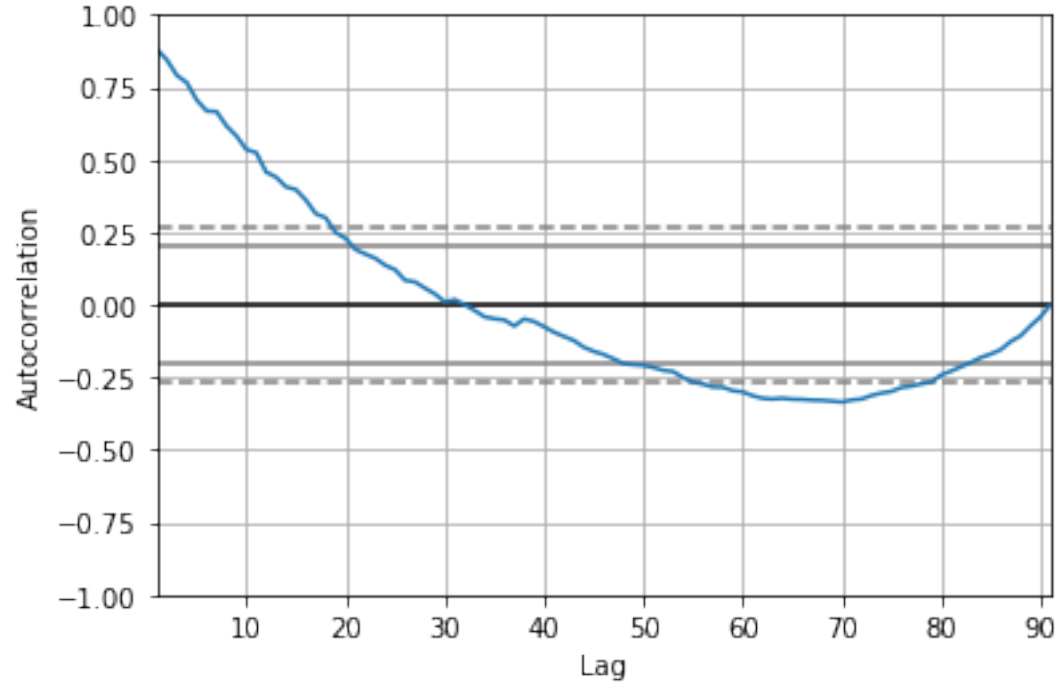
In [605]:



In [606]:



In [607]:



/ sentiment analysis

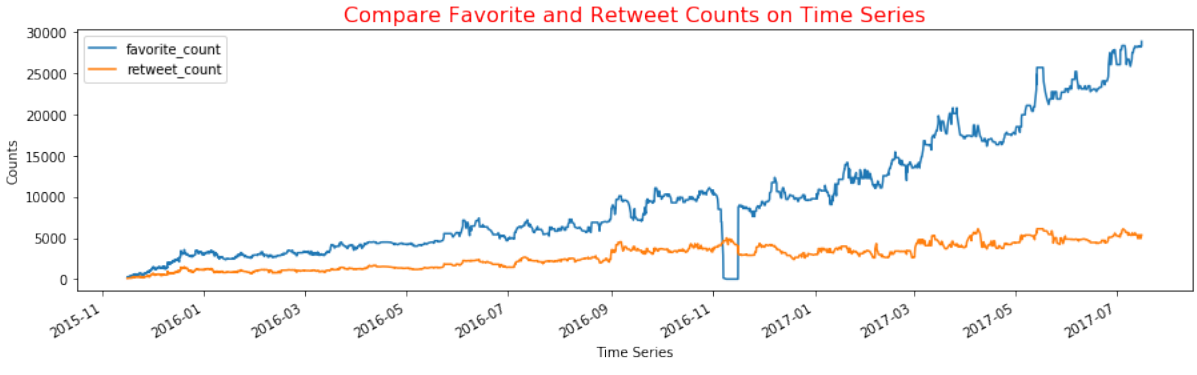
- 使用sklearn https://towardsdatascience.com/sentiment-analysis-with-python-part-1-5ce197074184 (https://towardsdatascience.com/sentiment-analysis-with-python-part-1-5ce197074184)
- 另外比较常见的是使用 nltk 库
- 此处先pass, 深度学习时候有空再深入

结论

/ favorite 和 retweet 时序分析

- 2016年上半年之前, favorite 数量大概是 retweet 的两倍
- 但再这之后, favorite 数量大量上涨, retweet 数量上涨十分缓慢(两者之比达到6倍)
- 推测相关因素如下:
  - 可以看出 twitter 增长非常迅速(可惜缺少用户量相关的数据)
  - 但是人们愿意付出更多一点时间 retweet 的时间在减少, 可能原因是当人接触到更多的 twitter 信息后, 能够 retweet 的注意力已经没有什么增长空间了(注意力处于饱和状态)

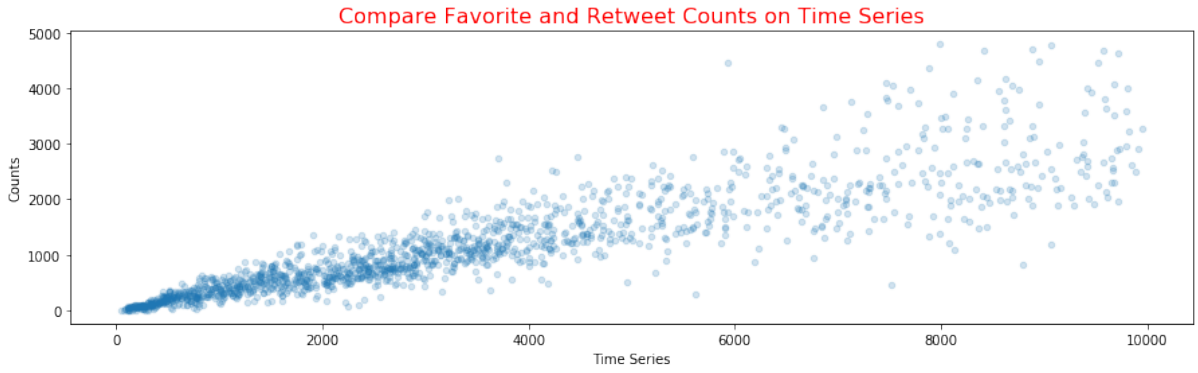
In [608]:



/ favorate 和 retweet 相关性分析

- 分析中过滤掉了 retweet 为0的数据和大于1000的数据
- 此处考虑的是两个参数的对应关系, 和问题1的趋势并不冲突(因为数据做了过滤)
- 可以看出在 favorate 和 retweet 两个数据中间具有相关性
- 回归线要用到 sm 库或 sklearn 库, 后续研究
  https://nbviewer.jupyter.org/github/weecology/progbio/blob/master/ipynbs/statistics.i
  (https://nbviewer.jupyter.org/github/weecology/progbio/blob/master/ipynbs/statistics.

In [609]:



/ word cloud 分析

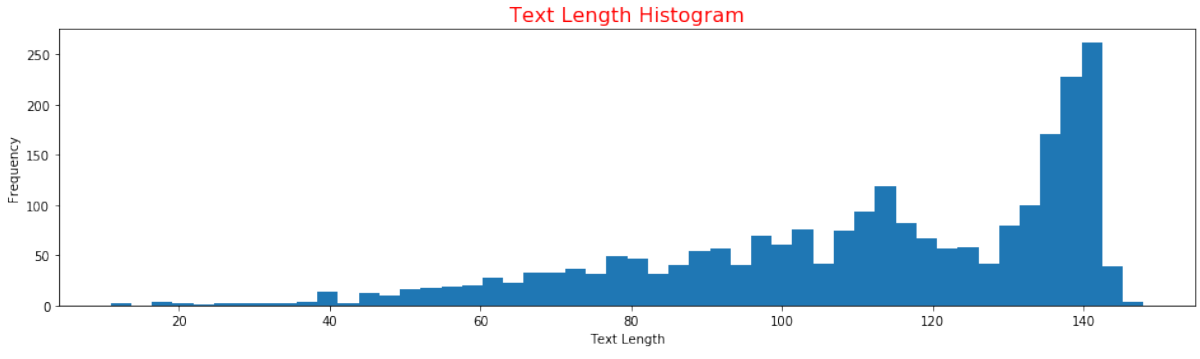
- 对评论使用 word cloud 进行分析
- 去掉了 stop words
- 图像为 twitter 英文字符(小鸟图不太美观)

In [610]:

/ text range 分析

- text range 改名为 text range 更为明确
- 数据做了过滤(过滤掉了个别 160 字符的)
- 数据有左偏斜趋势 (不能断定) 因为在140字的限制上有大量出现, 所以明显存在人为调整
- 有些数据超出了140
- 后续可以做异常值分析(按说不应该有超出, 也可能是正则化过滤时留下的问题)

In [611]:



/ 后续完善

- 增加数据feature: 虽然原始数据 featrue 比较多, 但经过梳理发现所剩数据不多. 像用户日活, 注册量等信息缺失.
- 完善情感分析: 情感分析可以画出 积极/消极/主观/客观 两个维度的信息. 便于增加数据用以更多分析 (比如 140字的回复中, 是积极信息多还是消极信息多)
- 完善 source 分类数据: 本来很关注的feature, 因为数据收集的问题(可能是数据收集时ios比较好记录), 这点非常重要, 因为起码从尝试来讲 android 的不应该这么少. 这种情况会造成数据偏见, 可能带来错误的结论

In [ ]:

```
```