Variables	Description	Data type	Notes
craeted_at	date the tweet was created	datetime, continuous	converted from object to datetime
entities	any entity the tweet has (hashtags, symbols, mentions)	object, categorical	
favorite_count	number of likes the tweet has	float64, continuous	
id	id of the tweet	float64, continuous	
in_reply_to_screen_name	if the tweet has reply (show screen name)	object, categorical	
in_reply_to_status_id	if the tweet has reply (show tweet id)	float64, continuous	
in_reply_to_user_id	if the tweet has reply (show user id)	float64, continuous	
place	place the tweet was tweeted from	object, categorical	all the data was NaN so i excluded the column
possibly_sensitive	if there are any sensitive content in the tweet	object, categorical	all the data was False so i excluded the column
retweet_count	number of retweets	float64, continuous	
text	the text of the tweet	object, categorical	
name	the name of the neawspaper	object, categorical	i added this column
retweet_mean	average number of retweets	float64, continuous	i added thid column

- there are 23155 date, 22747 of them is unique, the top is 2017-02-22 20:05:04 with freq of 5.
    - count 23155
    - unique 22747
    - top 2017-02-22 20:05:04
    - freq 5
- there are 23155 value of entities, 22099 of them are unique, the top is {'hashtags': [], 'symbols': [], 'user_mentions': [], 'urls': []} with freq of 603.
    - count 23155
    - unique 22099
    - top {'hashtags': [], 'symbols': [], 'user_mentions...
    - freq 603
- there are 23155 value of favorite_count with mean of(2.128568), max(2015.000000), min(0.000000)
    - count 23155.000000
    - mean 2.128568
    - std 16.904183

    - min 0.000000
    - 25% 0.000000
    - 50% 1.000000
    - 75% 2.000000
    - max 2015.000000
- these numbers are meaningless since this is just an (id) except the count which represent the number of ids we have

    - count 2.315500e+04
    - mean 9.794991e+17
    - std 4.117512e+16
    - min 6.757187e+17
    - 25% 9.837335e+17
    - 50% 9.864680e+17
    - 75% 9.892307e+17
    - max 9.918028e+17
- there are 60 tweets that has reply, the tweets of 4 defferent newspapers, the newspapers that has the most replys on its tweets is AlseyassahNews with 41 tweets that has reply

    - count 60
    - unique 4
    - top AlseyassahNews
    - freq 41
- this column(in_reply_to_status_id) has 60 value

    - count 6.000000e+01
    - mean 9.810572e+17
    - std 1.101162e+16
    - min 9.138978e+17
    - 25% 9.768778e+17
    - 50% 9.836585e+17
    - 75% 9.878681e+17
    - max 9.917926e+17
- this column(in_reply_to_user_id) has 60 value

    - count 6.000000e+01
    - mean 2.484898e+08
    - std 2.650336e+08
    - min 7.000638e+07
    - 25% 1.509726e+08
    - 50% 2.660159e+08
    - 75% 2.660159e+08
    - max 2.175857e+09
- there are 23155 of the column(retweet_count), the mean of retweets is(5.116519), max(7635.000000), min(0.000000)

    - count 23155.000000
    - mean 5.116519
    - std 63.617464
    - min 0.000000

    - 25% 0.000000
    - 50% 1.000000
    - 75% 2.000000
    - max 7635.000000
- there are 23155 tweets, 23010 of them are unique, the top tweet is(#.. الصوت | فضيحة مالية ورقابية تهز\n#التطب...), it was repeated(5 times)

    - count 23155
    - unique 23010
    - top #تهز .. فضيحة مالية ورقابية | الصوت\n#التطب...
    - freq 5
- there are 23155 value of name with unique value of 8, the top value is alanba with freq of 3200

    - count 23155
    - unique 8
    - top alanba
    - freq 3200

- alseyaseyah has people reply more than other newspapers but retweets less than other people
- the cause may be that the people is replying with somthing they didn't like about the tweet
- alwatan have above average number of retweets and alseyassah have below average number of retweets
- even tho aldaar is below average it has the most retweets as outlayers
- alwatan have the highest average of tweets likes (is it because people like the tweets more than other newspapers tweets or because it have more follwers?)

In [78]:
```
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import numpy as np
%matplotlib inline
import matplotlib.dates as mdates
```

In [3]:
```
tweets_df = pd.read_csv("tweets_df.csv", encoding="utf-8")
```

In [3]:
```
tweets_df.head().T
```

Out[3]:

	0	1	2	
created_at	2018-05-02 21:40:52	2018-05-02 21:35:33	2018-05-02 21:30:48	2018-05-
entities	{'hashtags': [], 'symbols': [], 'user_mentions...	{'hashtags': [], 'symbols': [], 'user_mentions...	{'hashtags': [], 'symbols': [], 'user_mentions...	{'h ' 'user_
favorite_count	0	2	4	
id	991794679770484737	991793339270860809	991792142912802817	99179127
in_reply_to_screen_name	NaN	NaN	NaN	
in_reply_to_status_id	NaN	NaN	NaN	
in_reply_to_user_id	NaN	NaN	NaN	
place	NaN	NaN	NaN	
possibly_sensitive	False	False	False	
retweet_count	4	1	1	
text	محكمة أمريكية تأمر إيران بدفع 6 مليارات دولار ...	من سكان العالم يتنشقون 90 % هواء ملوثا\n\nhttp...	وصول الطائرة «فيلكا» بعد استكمال صيانتها في «ب...	..الأحد\n\nhttp
name	alwatan	alwatan	alwatan	

In [4]:
```
tweets_df.created_at.describe()
```

Out[4]:
```
count                   23155
unique                  22747
top       2017-02-22 20:05:04
freq                        5
Name: created_at, dtype: object
```

In [5]:
```
tweets_df.entities.describe()
```

Out[5]:
```
count                                                   23155
unique                                                  22099
top       {'hashtags': [], 'symbols': [], 'user_mentions...
freq                                                      603
Name: entities, dtype: object
```

In [6]:
```
tweets_df.favorite_count.describe()
```

Out[6]:
```
count    23155.000000
mean         2.128568
std         16.904183
min          0.000000
25%          0.000000
50%          1.000000
75%          2.000000
max       2015.000000
Name: favorite_count, dtype: float64
```

In [23]:
```
tweets_df.id.describe()
```

Out[23]:
```
count    2.315500e+04
mean     9.794991e+17
std      4.117512e+16
min      6.757187e+17
25%      9.837335e+17
50%      9.864680e+17
75%      9.892307e+17
max      9.918028e+17
Name: id, dtype: float64
```

In [7]:
```
tweets_df.in_reply_to_screen_name.describe()
```

Out[7]:
```
count                60
unique                4
top       AlseyassahNews
freq                 41
Name: in_reply_to_screen_name, dtype: object
```

In [8]:
```
tweets_df.in_reply_to_status_id.describe()
```

Out[8]:
```
count    6.000000e+01
mean     9.810572e+17
std      1.101162e+16
min      9.138978e+17
25%      9.768778e+17
50%      9.836585e+17
75%      9.878681e+17
max      9.917926e+17
Name: in_reply_to_status_id, dtype: float64
```

In [9]:
```
tweets_df.in_reply_to_user_id.describe()
```

Out[9]:
```
count    6.000000e+01
mean     2.484898e+08
std      2.650336e+08
min      7.000638e+07
25%      1.509726e+08
50%      2.660159e+08
75%      2.660159e+08
max      2.175857e+09
Name: in_reply_to_user_id, dtype: float64
```

In [10]:
```
tweets_df.place.describe()
```

Out[10]:
```
count    0.0
mean     NaN
std      NaN
min      NaN
25%      NaN
50%      NaN
75%      NaN
max      NaN
Name: place, dtype: float64
```

In [11]:
```
tweets_df.possibly_sensitive.describe()
```

Out[11]:
```
count     21380
unique        1
top       False
freq      21380
Name: possibly_sensitive, dtype: object
```

In [12]:
```
tweets_df.retweet_count.describe()
```

Out[12]:
```
count    23155.000000
mean         5.116519
std         63.617464
min          0.000000
25%          0.000000
50%          1.000000
75%          2.000000
max       7635.000000
Name: retweet_count, dtype: float64
```

In [13]:
```
tweets_df.text.describe()
```

Out[13]:
```
count                                                23155
unique                                               23010
top       #تهز .. فضيحة مالية ورقابية | الصوت\n#التطب...
freq                                                     5
Name: text, dtype: object
```

In [62]:
```
tweets_df.name.describe()
```

Out[62]:
```
count      23155
unique         8
top       alanba
freq        3200
Name: name, dtype: object
```

In [54]:
```
tweets_text = tweets_df.text.value_counts().index
```

In [61]:
```
tweets_text[0]
```

Out[61]:
```
'#تهز .. فضيحة مالية ورقابية | الصوت\n#كافة يعطل المالي المراقب .. التطبيقي\nمن أكثر ومكتسبات حقوق
طالب الف50\nوطال… https://t.co/e4XWf3Eu3B' (https://t.co/e4XWf3Eu3B')
```

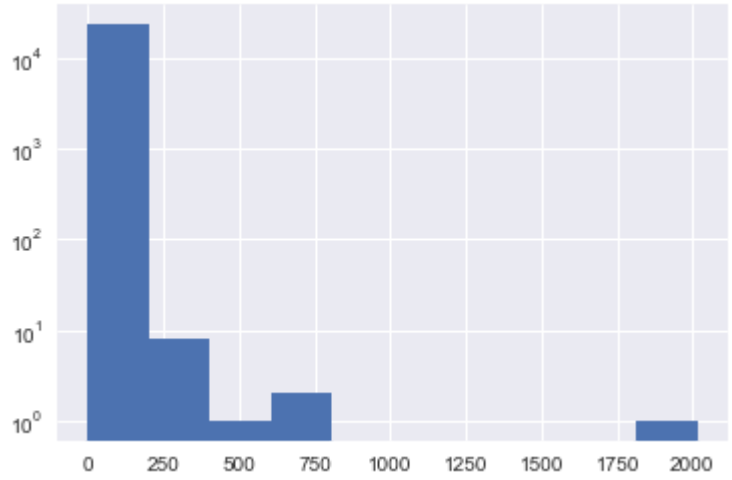
In [14]:
```
tweets_df.head().T
```

Out[14]:

	0	1	2	
created_at	2018-05-02 21:40:52	2018-05-02 21:35:33	2018-05-02 21:30:48	2018-05-
entities	{'hashtags': [], 'symbols': [], 'user_mentions...	{'hashtags': [], 'symbols': [], 'user_mentions...	{'hashtags': [], 'symbols': [], 'user_mentions...	{'h 'user
favorite_count	0	2	4	
id	991794679770484737	991793339270860809	991792142912802817	99179127
in_reply_to_screen_name	NaN	NaN	NaN	
in_reply_to_status_id	NaN	NaN	NaN	
in_reply_to_user_id	NaN	NaN	NaN	
place	NaN	NaN	NaN	
possibly_sensitive	False	False	False	
retweet_count	4	1	1	
text	محكمة أمريكية تأمر إيران بدفع 6 مليارات دولار ...	من سكان العالم يتنشقون 90 % هواء ملوثا\n\nhttp...	وصول الطائرة «فيلكا» بعد استكمال صيانتها في «ب...	..الأحد\n\nhttp
name	alwatan	alwatan	alwatan	

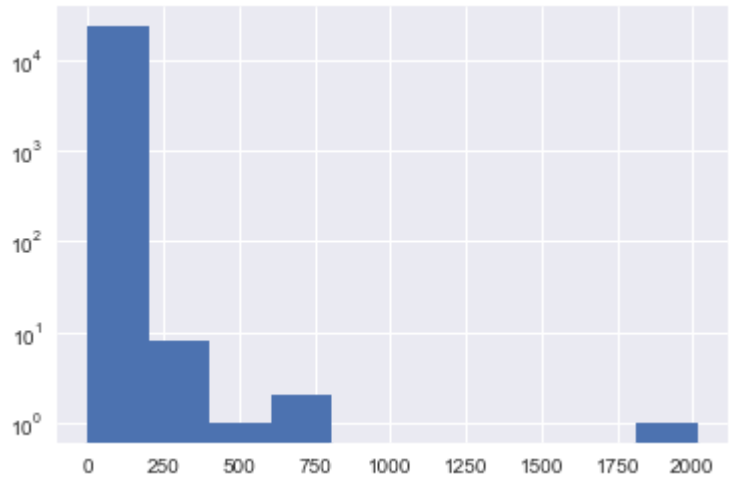
In [15]:
```
a = tweets_df.favorite_count.hist()
a.set(yscale="log")
```

Out[15]: [None]



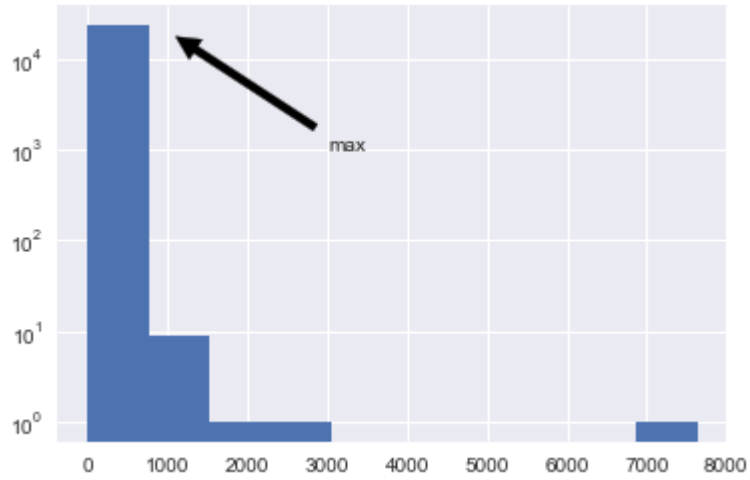
In [106]:
```
n = tweets_df.favorite_count.hist()
n.set(yscale="log")
```

Out[106]: [None]



In [73]:
```
b = tweets_df.retweet_count.hist()
b.set(yscale="log")
plt.annotate("max", xy=(1000, 20000), xytext= (3000, 1000), arrowprops=dict(facec
```

Out[73]: <matplotlib.text.Annotation at 0x300bb45fd0>



In [44]:
```
tweets_df[~(tweets_df.in_reply_to_screen_name.isnull())].count()
```

Out[44]:
```
created_at                60
entities                  60
favorite_count            60
id                        60
in_reply_to_screen_name   60
in_reply_to_status_id     60
in_reply_to_user_id       60
place                      0
possibly_sensitive        11
retweet_count             60
text                      60
name                      60
retweet_mean              60
dtype: int64
```

In [18]:
```
tweets_df["retweet_mean"] = tweets_df.retweet_count.mean()
```

In [ ]:
```
sns.factorplot(data=tweets_df, x="created_at", y="retweet_count", kind="box")
```

In [19]:
```
tweets_df.created_at
```

Out[19]:
```
0        2018-05-02 21:40:52
1        2018-05-02 21:35:33
2        2018-05-02 21:30:48
3        2018-05-02 21:27:22
4        2018-05-02 21:24:03
5        2018-05-02 21:20:20
6        2018-05-02 21:18:03
7        2018-05-02 21:16:07
8        2018-05-02 21:12:47
9        2018-05-02 21:10:52
10       2018-05-02 20:58:49
11       2018-05-02 20:55:07
12       2018-05-02 20:51:37
13       2018-05-02 20:40:20
14       2018-05-02 20:33:51
15       2018-05-02 19:49:21
16       2018-05-02 19:49:17
17       2018-05-02 19:32:27
18       2018-05-02 19:28:51
19       2018-05-02 19:09:48
20       2018-05-02 18:48:55
21       2018-05-02 18:38:47
22       2018-05-02 18:21:19
23       2018-05-02 18:18:14
24       2018-05-02 18:08:59
25       2018-05-02 18:06:38
26       2018-05-02 18:06:17
27       2018-05-02 18:05:52
28       2018-05-02 17:53:13
29       2018-05-02 17:49:39
                  ...
23125    2018-04-14 03:26:07
23126    2018-04-14 03:22:55
23127    2018-04-14 03:18:08
23128    2018-04-14 03:07:31
23129    2018-04-14 03:05:11
23130    2018-04-14 03:02:45
23131    2018-04-14 02:57:55
23132    2018-04-14 02:55:39
23133    2018-04-14 02:47:28
23134    2018-04-14 02:45:59
23135    2018-04-14 02:45:51
23136    2018-04-14 02:42:20
23137    2018-04-14 02:38:53
23138    2018-04-14 02:29:14
23139    2018-04-14 02:20:33
23140    2018-04-14 02:19:42
23141    2018-04-14 02:16:25
23142    2018-04-14 02:10:27
23143    2018-04-14 02:09:41
23144    2018-04-14 02:09:22
23145    2018-04-14 02:07:42
23146    2018-04-14 02:06:26
23147    2018-04-14 02:05:05
23148    2018-04-14 02:03:20
```

```
23149    2018-04-14 02:00:27
23150    2018-04-14 01:51:57
23151    2018-04-14 01:50:22
23152    2018-04-14 01:48:13
23153    2018-04-14 01:40:28
23154    2018-04-14 01:39:18
Name: created_at, Length: 23155, dtype: object
```

In [21]:
```
tweets_df["created_at"] = pd.to_datetime(tweets_df.created_at)
```

In [22]:
```
tweets_df["created_at"]
```

Out[22]:
```
0       2018-05-02 21:40:52
1       2018-05-02 21:35:33
2       2018-05-02 21:30:48
3       2018-05-02 21:27:22
4       2018-05-02 21:24:03
5       2018-05-02 21:20:20
6       2018-05-02 21:18:03
7       2018-05-02 21:16:07
8       2018-05-02 21:12:47
9       2018-05-02 21:10:52
10      2018-05-02 20:58:49
11      2018-05-02 20:55:07
12      2018-05-02 20:51:37
13      2018-05-02 20:40:20
14      2018-05-02 20:33:51
15      2018-05-02 19:49:21
16      2018-05-02 19:49:17
17      2018-05-02 19:32:27
18      2018-05-02 19:28:51
19      2018-05-02 19:09:48
20      2018-05-02 18:48:55
21      2018-05-02 18:38:47
22      2018-05-02 18:21:19
23      2018-05-02 18:18:14
24      2018-05-02 18:08:59
25      2018-05-02 18:06:38
26      2018-05-02 18:06:17
27      2018-05-02 18:05:52
28      2018-05-02 17:53:13
29      2018-05-02 17:49:39
              ...
23125   2018-04-14 03:26:07
23126   2018-04-14 03:22:55
23127   2018-04-14 03:18:08
23128   2018-04-14 03:07:31
23129   2018-04-14 03:05:11
23130   2018-04-14 03:02:45
23131   2018-04-14 02:57:55
23132   2018-04-14 02:55:39
23133   2018-04-14 02:47:28
23134   2018-04-14 02:45:59
23135   2018-04-14 02:45:51
23136   2018-04-14 02:42:20
23137   2018-04-14 02:38:53
23138   2018-04-14 02:29:14
23139   2018-04-14 02:20:33
23140   2018-04-14 02:19:42
23141   2018-04-14 02:16:25
23142   2018-04-14 02:10:27
23143   2018-04-14 02:09:41
23144   2018-04-14 02:09:22
23145   2018-04-14 02:07:42
23146   2018-04-14 02:06:26
23147   2018-04-14 02:05:05
23148   2018-04-14 02:03:20
```

```
23149   2018-04-14 02:00:27
23150   2018-04-14 01:51:57
23151   2018-04-14 01:50:22
23152   2018-04-14 01:48:13
23153   2018-04-14 01:40:28
23154   2018-04-14 01:39:18
Name: created_at, Length: 23155, dtype: datetime64[ns]
```

In [28]:
```
tweets_df[~(tweets_df.place.isnull())]
```

Out[28]:

created_at	entities	favorite_count	id	in_reply_to_screen_name	in_reply_to_status_id	in_reply_t

In [35]:
```
tweets_df[tweets_df.possibly_sensitive == True]
```

Out[35]:

created_at	entities	favorite_count	id	in_reply_to_screen_name	in_reply_to_status_id	in_reply_t

In [37]:
```
tweets_df.entities.value_counts()
```

Out[37]:
```
{'hashtags': [], 'symbols': [], 'user_mentions': [], 'urls': []}
                                                               603
{'hashtags': [], 'symbols': [], 'user_mentions': [{'screen_name': 'alqabas_
tv', 'name': 'تلفزيون القبس', 'id': 928669202386444288, 'id_str': '928669202386
444288', 'indices': [3, 14]}], 'urls': []}
```

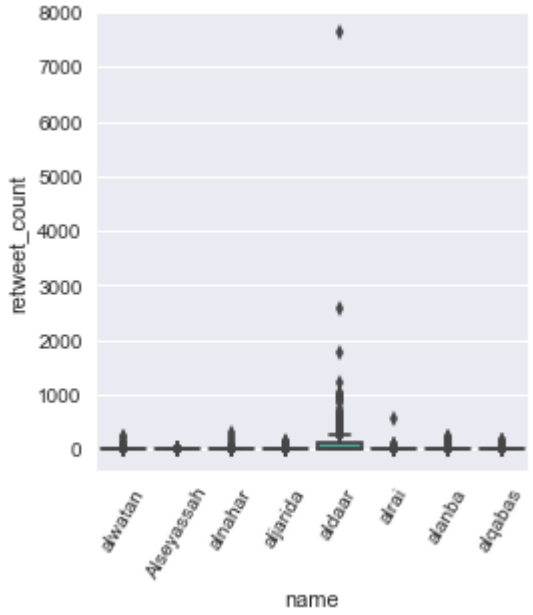
In [45]:
```
tweets_df.retweet_count
```

Out[45]:
```
0        4
1        1
2        1
3        1
4        1
5        3
6        2
7        2
8        4
9        2
10       1
11       2
12       6
13       3
14       6
15       1
16       3
17       3
18       2
19       3
20       1
21       3
22       2
23       4
24       0
25       0
26       2
27       0
28       1
29       2
        ..
23125    1
23126    1
23127    1
23128    0
23129    1
23130    0
23131    1
23132    1
23133    1
23134    0
23135    6
23136    1
23137    2
23138    2
23139    1
23140    0
23141    2
23142    0
23143    0
23144    0
23145    1
23146    1
23147    0
23148    1
```

```
23149    3
23150    2
23151    0
23152    2
23153    0
23154    2
Name: retweet_count, Length: 23155, dtype: int64
```

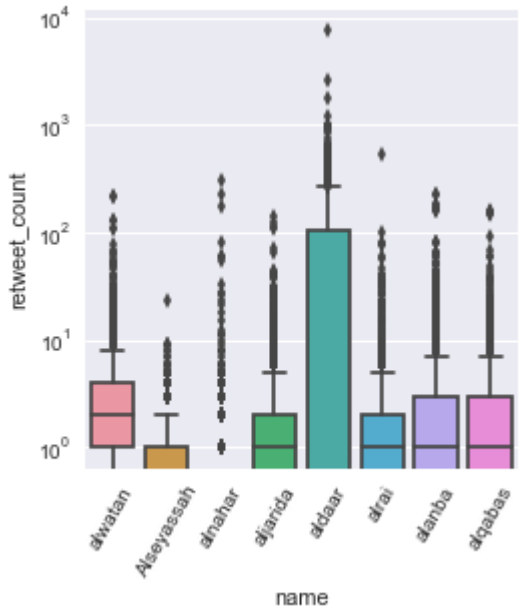
In [78]:
```
retweet_and_count = sns.factorplot(data=tweets_df, x="name", y="retweet_count", k
plt.xticks(rotation=60)
```

Out[78]: (array([0, 1, 2, 3, 4, 5, 6, 7]), <a list of 8 Text xticklabel objects>)



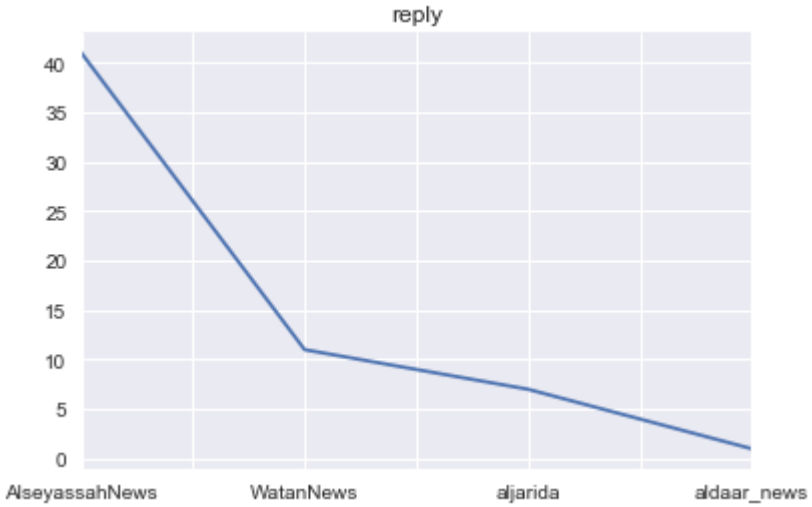
In [77]:
```
retweet_and_count = sns.factorplot(data=tweets_df, x="name", y="retweet_count", k
retweet_and_count.set(yscale="log")
plt.xticks(rotation=60)     # alwatan have above average number of retweets and a
                            # even tho aldaar is below average it has the most re
```

Out[77]: (array([0, 1, 2, 3, 4, 5, 6, 7]), <a list of 8 Text xticklabel objects>)



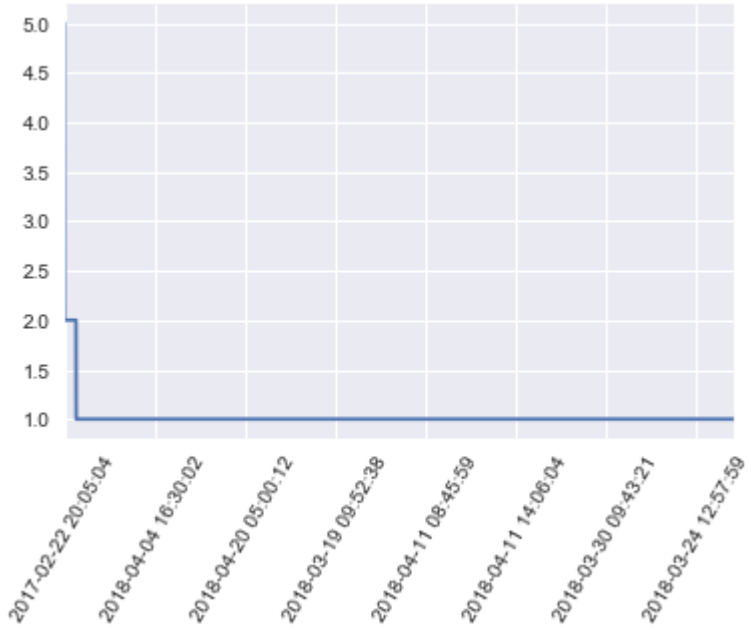
In [131]:
```
tweets_df.in_reply_to_screen_name.value_counts().plot()  # alseyaseyah has people
plt.title("reply")                                                    # the ca
```

Out[131]: <matplotlib.text.Text at 0x3014890da0>



In [121]:
```
date_retweets.created_at.value_counts().plot()
plt.xticks(rotation=60)
```

Out[121]:
```
(array([     0.,   2000.,   4000.,   6000.,   8000.,  10000.,  12000.,
         14000.,  16000.]), <a list of 9 Text xticklabel objects>)
```

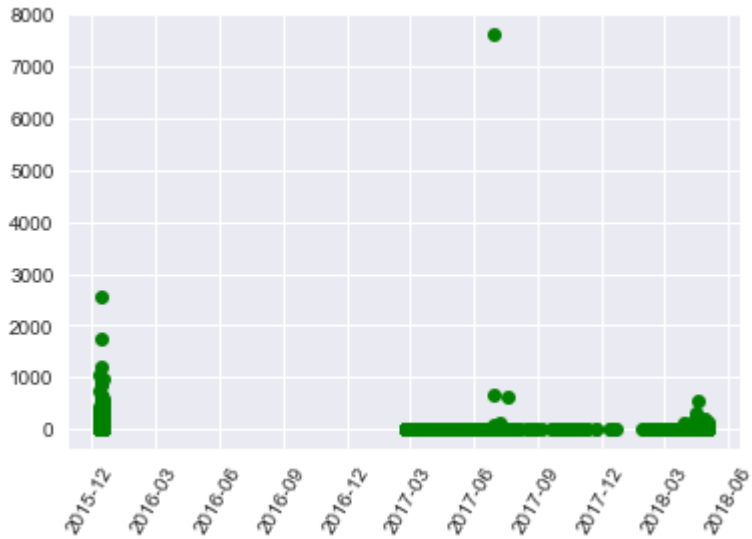


In [91]:
```
date_retweets = tweets_df[(tweets_df.created_at < "2018-04-23") & (~tweets_df.re
```

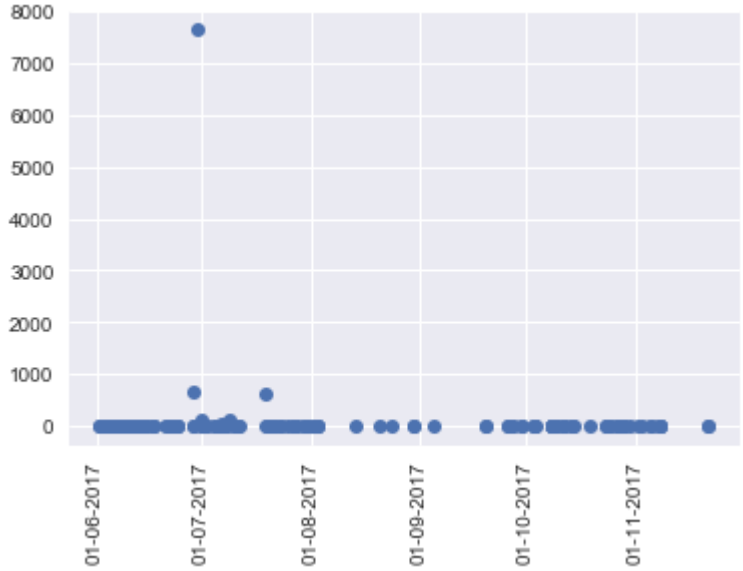
In [7]:
```
date_retweets_sample=date_retweets.sample(5000)
```

In [77]:
```
date= plt.plot_date(data=tweets_df, x="created_at", y="retweet_count", fmt="go")
plt.xticks(rotation=60)
```

Out[77]:
```
(array([ 735933.,  736024.,  736116.,  736208.,  736299.,  736389.,
         736481.,  736573.,  736664.,  736754.,  736846.]),
 <a list of 11 Text xticklabel objects>)
```

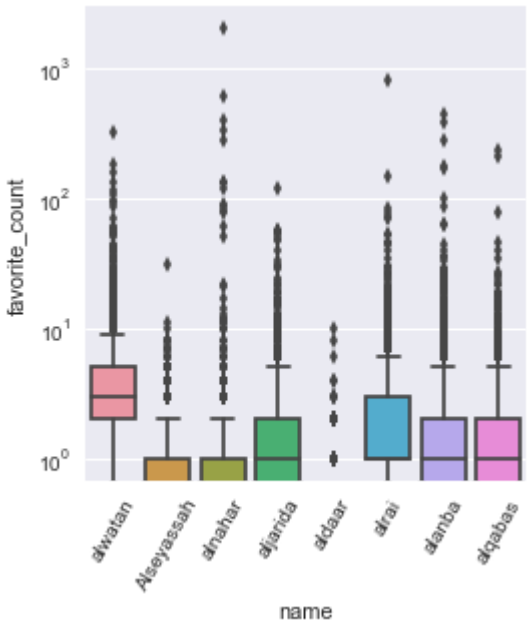


In [100]:
```
date2= plt.plot_date(data=date_retweets2, x="created_at", y="retweet_count")
plt.xticks(rotation=90)
ax = plt.subplot()
ax.xaxis.set_major_formatter(mdates.DateFormatter("%d-%m-%Y"))
```



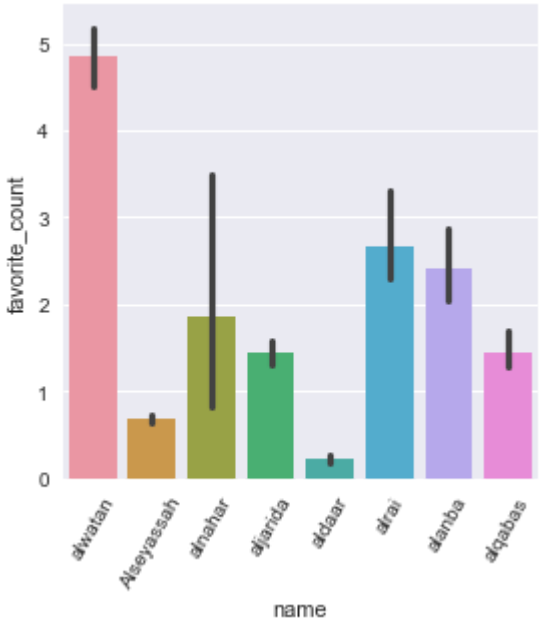
In [117]:
```
v=sns.factorplot(data=tweets_df, x="name", y="favorite_count", kind="box")
plt.xticks(rotation=60)
v.set(yscale="log")
```

Out[117]: <seaborn.axisgrid.FacetGrid at 0x300c78c978>



In [119]:
```
v=sns.factorplot(data=tweets_df, x="name", y="favorite_count", kind="bar")
plt.xticks(rotation=60)
```

Out[119]: (array([0, 1, 2, 3, 4, 5, 6, 7]), <a list of 8 Text xticklabel objects>)



In [ ]: