

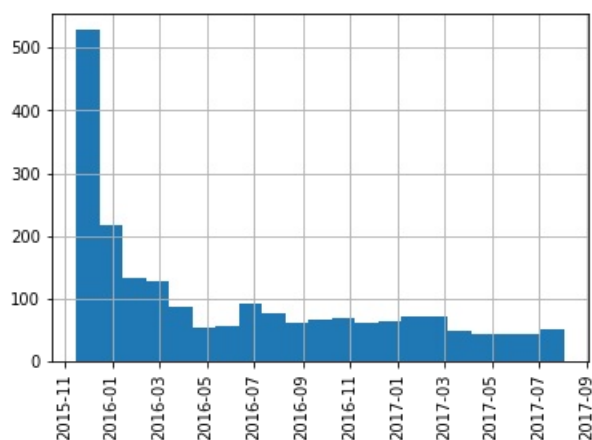
# Acting on Wrangled Data

How many days can we see the activity of this account?

```
print('We have data from:', master.timestamp.min(), 'to:', master.timestamp.max())
```

We have data from: 2015-11-15 22:32:08 to: 2017-08-01 16:23:56

```
master.timestamp.hist(bins=21, xrot=90);
```



WeRateDogs had most tweets at December 2015, but at other intervals, we can see less 100 tweets in a month.

## Top 10 of scores in the rating system.

```
rate = lambda row: str(row['rating_numerator']) + '/' + \
            + str(row['rating_denominator'])
master.apply(rate, axis=1).value_counts()[:10]
```

```
12.0/10.0 474
10.0/10.0 429
11.0/10.0 413
13.0/10.0 281
9.0/10.0 150
8.0/10.0 95
7.0/10.0 51
14.0/10.0 40
```

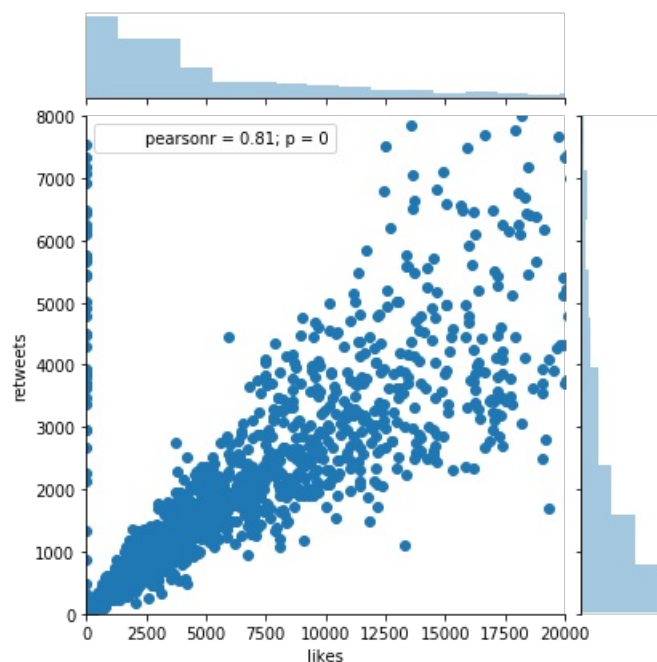
5.0/10.0 34  
6.0/10.0 32  
dtype: int64

```
print(round(sum(master.apply(rate, axis=1).value_counts()[:4].values)
        / len(master) * 100), '% tweets have rate from 10/10 to 13/10')
```

77.0 % tweets have rate from 10/10 to 13/10

## About correlation between likes and retweets

```
sns.jointplot(x=master["favorite_count"], y=master["retweet_count"],
              kind='scatter', marginal_kws=dict(bins=100), xlim=(0, 20000),
              ylim=(0, 8000)).set_axis_labels("likes", "retweets");
```



Likes correlate to retweets about 0.76, and we can see that every two likes bring one retweet.

```
print(len(master[(master.favorite_count == 0) & (master.retweet_count > 0)]),
      'tweets have retweets without any like.')
```

78 tweets have retweets without any like.

## Top 10 names of dogs

```
print(master[master.name.notnull() == True].name.value_counts()[:10])

names = ' '.join(list(master[master.name.notnull() == True].name))
wordcloud = WordCloud().generate(names)

plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off");
```

```
Charlie 11
Tucker 10
Lucy 10
Oliver 10
Penny 10
Cooper 10
Bo 8
Lola 8
Sadie 8
Name: name, dtype: int64
```



# About recognize picture of dogs

```
# how many dogs recognize
breeds = master[(master.p1_dog == True) |
                 (master.p2_dog == True) |
                 (master.p3_dog == True)].copy()
print(len(breeds), 'pictures from', len(master),
      'tweets had recognized successful as a picture of the dog.')
```

1748 pictures from 2072 tweets had recognized successful as a picture of the dog.

```
def get_breed(row):
    breed, conf = '', 0.0
    if (row['p1_dog'] == True) and (row['p1_conf'] > conf):
        breed, conf = row['p1'], row['p1_conf']

    if (row['p2_dog'] == True) and (row['p2_conf'] > conf):
        breed, conf = row['p2'], row['p2_conf']

    if (row['p3_dog'] == True) and (row['p3_conf'] > conf):
        breed, conf = row['p3'], row['p3_conf']

    row['breed'] = breed
    row['conf'] = conf

    return row

breeds['breed'] = ''
breeds['conf'] = 0
breeds = breeds.apply(get_breed, axis=1)
```

```
print('We got', len(breeds.breed.unique()), 'names of a breed.')
```

We got 113 names of a breed.

```
print('Top 10 breeds')
breeds.breed.value_counts()[:10]
```

Top 10 breeds

```
golden_retriever 173
Labrador_retriever 113
Pembroke 95
Chihuahua 94
pug 65
toy_poodle 52
chow 51
Samoyed 46
Pomeranian 42
malamute 34
Name: breed, dtype: int64
```

```
print(round(len(breeds[breeds.conf >= 0.9]) / len(breeds) * 100, 2),
      '% breeds recognized with 90% confidence.')
```

16.93 % breeds recognized with 90% confidence.

```
print('Top 10 breeds recognized with 90% confidence.')
breeds[breeds.conf >= 0.9].breed.value_counts()[:10]
```

Top 10 breeds recognized with 90% confidence.

golden\_retriever 40  
pug 26  
Pembroke 26  
Samoyed 22  
Labrador\_retriever 19  
Pomeranian 16  
Chihuahua 13  
French\_bulldog 12  
chow 9  
Blenheim\_spaniel 6  
Name: breed, dtype: int64

```
# how many breeds not recognize
pure_rec = set(breeds.breed.unique()) - set(breeds[breeds.conf >= 0.9].breed.unique())
print(len(pure_rec), '=', round(len(pure_rec) / len(breeds.breed.unique()) * 100, 2),
      '% breeds NOT recognized with 90% confidence.')
```

55 = 48.67 % breeds NOT recognized with 90% confidence.