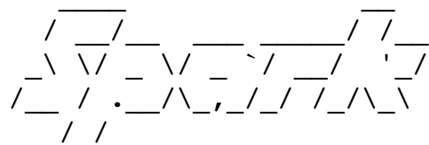


```
In [1]: import os
execfile(os.path.join(os.environ["SPARK_HOME"], 'python/pyspark/shell.py'))
```

Welcome to

 version 2.3.1

Using Python version 2.7.10 (default, Feb 7 2017 00:08:15)
SparkSession available as 'spark'.

```
In [2]: sc
```

Out[2]: **SparkContext**

[Spark UI \(http://10.43.3.52:4040\)](http://10.43.3.52:4040)

Version

v2.3.1

Master

local[*]

AppName

pyspark-shell

```
In [3]: from pyspark.sql import SparkSession
from pyspark.sql.functions import explode
from pyspark.sql import Window
import pyspark.sql.functions as f
```

```
In [26]: import numpy as np
import pandas as pd
import seaborn as sn
sn.set_style("darkgrid")

import matplotlib.pyplot as plt
import matplotlib.mlab as mlab
from sklearn.preprocessing import scale
```

```
In [4]: jsonData = spark.read.json("hdfs:///jason/beer")
```

```
In [ ]: jsonData.printSchema()
```

Top Beer Brands by Social Media Presence

```
In [7]: brands = jsonData.select('user.name', 'user.followers_count', 'user.verified')
brands.createOrReplaceTempView("brands")
brandsDF = spark.sql("SELECT brands.name, brands.followers_count, brands.description")
brandsDF = brandsDF.filter(brandsDF.description.contains('21+') | brandsDF.name.contains('21+'))
brandcounts = brandsDF.groupBy('name').count()
brands = brandsDF.dropDuplicates(['name']).drop('description')
Brands = brandcounts.join(brands, on='name')
Brands.show(50)
```

name	count	followers_count	favourites_count
BrewDog	3	125732	26156
Brooklyn Brew Shop	3	9815	6229
Michelob ULTRA	1	57886	292
Wild Rose Brewery...	1	17530	5544
Simon Brewer	1	22577	5829
MadTree Brewing	2	19292	13337
SweetWater Brewery	1	70929	7107
Ballast Point Beer	2	94783	10706
Blue Point Brewery	1	22078	12347
3 Floyds Brewpub	1	25476	116
Loddon Brewery	1	7053	3649
HonestBrew	1	13911	7967
Anchor Brewing	2	66906	14654
Night Shift Brewing	2	36506	14308
Brew Studs	12	100773	51421
Blue Moon Brewing Co	1	52712	9631
10 Barrel Brewing	3	18696	20050
Goose Island Beer Co	2	79502	35414
Lancaster Brewing Co	1	9485	9698
Summit Brewing	3	52433	9786
Budweiser	2	162713	2934
3 Floyds Brewing	1	88231	3498
Yards Brewing Co.	2	22508	24356
New Belgium Brewing	1	244809	8532
RITAS	1	27611	263
Pike Brewing Company	3	15229	5399
Lift Bridge Brewing	1	17822	10101
MillerCoors	1	26413	6476
Lone Star Beer	1	8578	2937
Harpoon Brewery	2	69830	47482
Deschutes Brewery	1	114047	30392
Short's Brewing	1	29069	9197
Tropicana AC	1	45044	12137
Saugatuck Brewing	5	7861	10610
SABreweries	1	25613	6096
Nickel Brook Brewing	2	13760	14717
Dogfish Head Brewery	2	298979	19915
Labatt Breweries	2	1189	433
Maestro Dobel	2	17827	4908
Two Roads Brewing	1	15756	26886
Allagash Brewing Co	1	32629	33973
Big Rock Brewery	2	21744	17378
Guinness US	1	16303	7414
Meantime Brewing Co.	1	33567	16697
8th Wonder Brewer...	1	22902	11619

Smuttynose Brewing	1	38061	13292
GreatLakesBrewingCo.	3	52224	8554
Brewers Association	3	76115	4357
+-----+-----+-----+-----+			

```
In [21]: brands = Brands.toPandas()
```

```
In [22]: brands.head()
```

Out[22]:

	name	count	followers_count	favourites_count
0	BrewDog	3	125732	26156
1	Brooklyn Brew Shop	3	9815	6229
2	Michelob ULTRA	1	57886	292
3	Wild Rose Brewery Ltd	1	17530	5544
4	Simon Brewer	1	22577	5829

```
In [ ]: brands = brands.set_index('name')
```

```
In [56]: brands['score'] = scale(brands['count']) + scale(brands['followers_count'])
brands.sort_values('score', ascending=False).head(10)
```

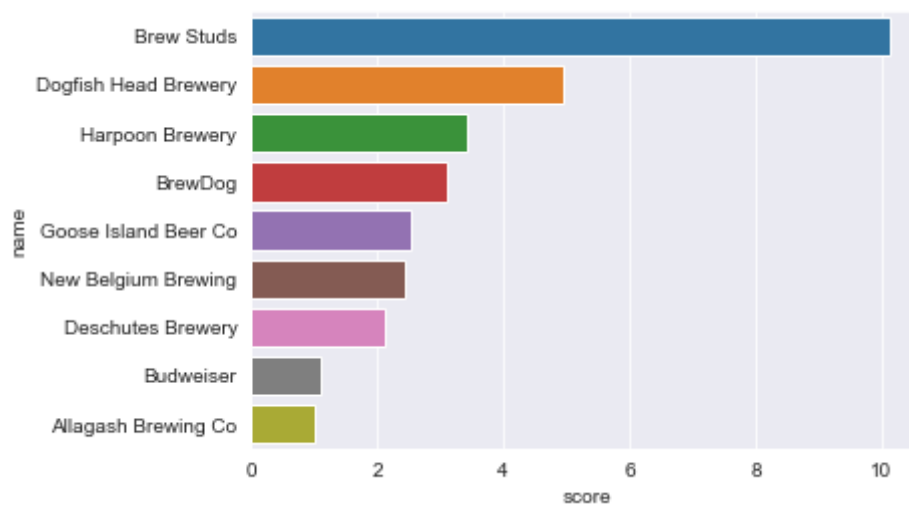
Out[56]:

	count	followers_count	favourites_count	score
name				
Brew Studs	12	100773	51421	10.137127
Dogfish Head Brewery	2	298979	19915	4.962605
Harpoon Brewery	2	69830	47482	3.426856
BrewDog	3	125732	26156	3.101009
Goose Island Beer Co	2	79502	35414	2.533159
New Belgium Brewing	1	244809	8532	2.443199
Deschutes Brewery	1	114047	30392	2.105596
Budweiser	2	162713	2934	1.115722
Allagash Brewing Co	1	32629	33973	1.013739
Saugatuck Brewing	5	7861	10610	0.863957

```
In [63]: TopBrands = brands.query('score > 1').reset_index().sort_values('score', as
```

```
In [64]: sn.barplot(x='score', y='name', data=TopBrands)
```

```
Out[64]: <matplotlib.axes._subplots.AxesSubplot at 0x110d33150>
```



Brand Mentions

```
In [11]: mentions = jsonData.select(explode("entities.user_mentions").alias("mention")
      select("mentions.name")
      mentions = mentions.dropna()
      MentionCounts = mentions.groupBy('name').count()
      BrandsTrim = Brands.select('name')
      BrMentions = BrandsTrim.join(MentionCounts, on='name')
      BrMentions.show(50)
```

name	count
BrewDog	144
Brooklyn Brew Shop	2
Michelob ULTRA	65
Wild Rose Brewery...	9
MadTree Brewing	23
SweetWater Brewery	31
Ballast Point Beer	46
3 Floyds Brewpub	1
Blue Point Brewery	12
Loddon Brewery	4
HonestBrew	2
Anchor Brewing	20
Night Shift Brewing	17
Brew Studs	55
Blue Moon Brewing Co	20
10 Barrel Brewing	24
Goose Island Beer Co	161
Lancaster Brewing Co	4
Summit Brewing	23
Budweiser	65
3 Floyds Brewing	48
Yards Brewing Co.	9
New Belgium Brewing	101
RITAS	10
Pike Brewing Company	8
Lift Bridge Brewing	7
MillerCoors	74
Lone Star Beer	15
Harpoon Brewery	50
Deschutes Brewery	57
Short's Brewing	33
Tropicana AC	1
Saugatuck Brewing	10
SABreweries	14
Nickel Brook Brewing	7
Dogfish Head Brewery	138
Maestro Dobel	3
Two Roads Brewing	30
Allagash Brewing Co	24
Big Rock Brewery	5
Guinness US	3
Meantime Brewing Co.	5
8th Wonder Brewer...	24
Smuttynose Brewing	9
GreatLakesBrewingCo.	49
Brewers Association	10

+-----+-----+

```
In [57]: Mentions = BrMentions.toPandas()
```

```
In [59]: Mentions = Mentions.sort_values('count', ascending=False)
Mentions.head(10)
```

Out[59]:

	name	count
16	Goose Island Beer Co	161
0	BrewDog	144
35	Dogfish Head Brewery	138
22	New Belgium Brewing	101
26	MillerCoors	74
2	Michelob ULTRA	65
19	Budweiser	65
29	Deschutes Brewery	57
13	Brew Studs	55
28	Harpoon Brewery	50

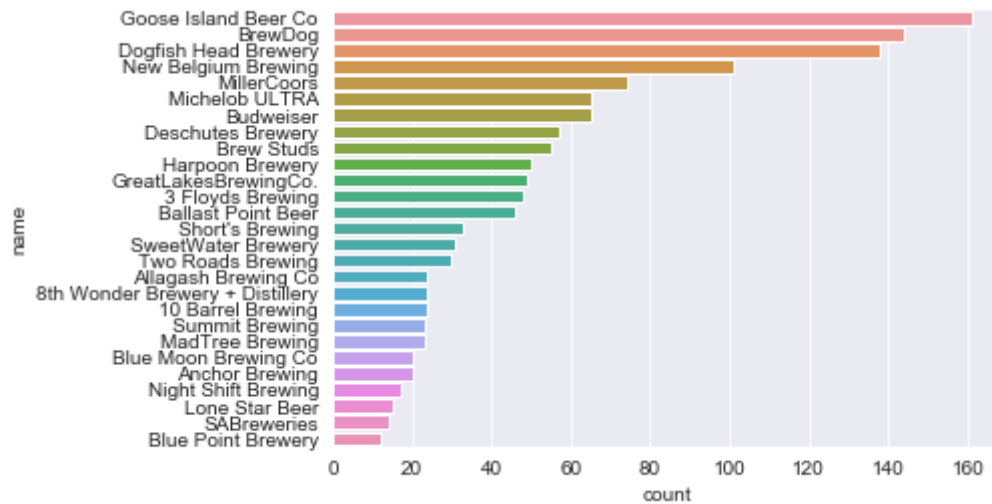
```
In [60]: Mentions.tail(10)
```

Out[60]:

	name	count
39	Big Rock Brewery	5
41	Meantime Brewing Co.	5
17	Lancaster Brewing Co	4
9	Loddon Brewery	4
36	Maestro Dobel	3
40	Guinness US	3
1	Brooklyn Brew Shop	2
10	HonestBrew	2
7	3 Floyds Brewpub	1
31	Tropicana AC	1

```
In [62]: Mentions = Mentions.query('count > 10')
sn.barplot(x='count', y='name', data=Mentions)
```

Out[62]: <matplotlib.axes._subplots.AxesSubplot at 0x1109b8190>



Top Beer Brands by Locality

```
In [14]: places = jsonData.select("place", explode("entities.user_mentions").alias("
select("mentions.name", "place.country")
places = places.dropna().groupBy('name', 'country').count()
BrandsTrim2 = Brands.drop('count')
Locality = places.join(BrandsTrim2, on='name')
Locality.sort(f.desc('country')).show(50)
```

```
+-----+-----+-----+-----+-----+
+-----+
|          name|          country|count|followers_count|favourites
_count|
+-----+-----+-----+-----+-----+
+-----+
| SweetWater Brewery| United States| 10| 70929|
7107|
| Ballast Point Beer| United States| 10| 94783|
10706|
| BrewDog| United States| 3| 125732|
26156|
| Michelob ULTRA| United States| 2| 57886|
292|
| MadTree Brewing| United States| 7| 19292|
13337|
| Lancaster Brewing Co| United States| 1| 9485|
9698|
| Summit Brewing| United States| 1| 52433|
9786|
| 3 Floyds Brewing| United States| 11| 88231|
3498|
| Yards Brewing Co.| United States| 3| 22508|
24356|
| Blue Point Brewery| United States| 3| 22078|
12347|
| Goose Island Beer Co| United States| 46| 79502|
35414|
| Pike Brewing Company| United States| 2| 15229|
5399|
| New Belgium Brewing| United States| 21| 244809|
8532|
| Lift Bridge Brewing| United States| 1| 17822|
10101|
| Harpoon Brewery| United States| 12| 69830|
47482|
| Night Shift Brewing| United States| 7| 36506|
14308|
| MillerCoors| United States| 16| 26413|
6476|
| Deschutes Brewery| United States| 14| 114047|
30392|
| Short's Brewing| United States| 8| 29069|
9197|
| Blue Moon Brewing Co| United States| 9| 52712|
9631|
| Dogfish Head Brewery| United States| 34| 298979|
19915|
| Two Roads Brewing| United States| 9| 15756|
26886|
```


	Drink_Local		
Anchor Brewing	United States	4	66906
14654			
Allagash Brewing Co	United States	3	32629
33973			
10 Barrel Brewing	United States	12	18696
20050			
GreatLakesBrewingCo.	United States	15	52224
8554			
8th Wonder Brewer...	United States	7	22902
11619			
Meantime Brewing Co.	United Kingdom	2	33567
16697			
BrewDog	United Kingdom	15	125732
26156			
Loddon Brewery	United Kingdom	1	7053
3649			
Goose Island Beer Co	United Kingdom	5	79502
35414			
Deschutes Brewery	Thailand	1	114047
30392			
BrewDog	Norway	1	125732
26156			
BrewDog	Nederland	2	125732
26156			
Ballast Point Beer	Mexico	1	94783
10706			
BrewDog	Italia	2	125732
26156			
BrewDog	Iceland	1	125732
26156			
MillerCoors	Dominican Republic	1	26413
6476			
BrewDog	Denmark	1	125732
26156			
Anchor Brewing	Canada	1	66906
14654			
Blue Moon Brewing Co	Canada	1	52712
9631			
Big Rock Brewery	Canada	1	21744
17378			
BrewDog	Brazil	1	125732
26156			
Goose Island Beer Co	Brasil	5	79502
35414			
BrewDog	Belgium	1	125732
26156			
Dogfish Head Brewery		2	298979
19915			
+-----+-----+-----+-----+			
-----+			

```
In [100]: locality = Locality.toPandas()
```

```
In [101]: locality = locality.sort_values(['country', 'count'], ascending=False)
locality.head(10)
```

Out[101]:

	name	country	count	followers_count	favourites_count
22	Goose Island Beer Co	United States	46	79502	35414
38	Dogfish Head Brewery	United States	34	298979	19915
29	New Belgium Brewing	United States	21	244809	8532
32	MillerCoors	United States	16	26413	6476
45	GreatLakesBrewingCo.	United States	15	52224	8554
36	Deschutes Brewery	United States	14	114047	30392
21	10 Barrel Brewing	United States	12	18696	20050
34	Harpoon Brewery	United States	12	69830	47482
27	3 Floyds Brewing	United States	11	88231	3498
11	SweetWater Brewery	United States	10	70929	7107

In [102]: locality

Out[102]:

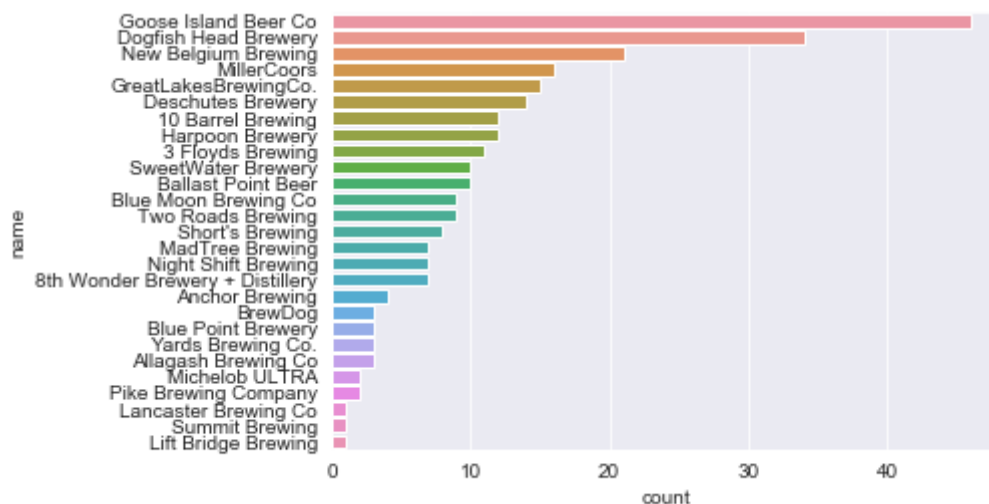
	name	country	count	followers_count	favourites_count
22	Goose Island Beer Co	United States	46	79502	35414
38	Dogfish Head Brewery	United States	34	298979	19915
29	New Belgium Brewing	United States	21	244809	8532
32	MillerCoors	United States	16	26413	6476
45	GreatLakesBrewingCo.	United States	15	52224	8554
36	Deschutes Brewery	United States	14	114047	30392
21	10 Barrel Brewing	United States	12	18696	20050
34	Harpoon Brewery	United States	12	69830	47482
27	3 Floyds Brewing	United States	11	88231	3498
11	SweetWater Brewery	United States	10	70929	7107
13	Ballast Point Beer	United States	10	94783	10706
20	Blue Moon Brewing Co	United States	9	52712	9631
40	Two Roads Brewing	United States	9	15756	26886
37	Short's Brewing	United States	8	29069	9197
10	MadTree Brewing	United States	7	19292	13337
18	Night Shift Brewing	United States	7	36506	14308
44	8th Wonder Brewery + Distillery	United States	7	22902	11619
17	Anchor Brewing	United States	4	66906	14654
6	BrewDog	United States	3	125732	26156
14	Blue Point Brewery	United States	3	22078	12347
28	Yards Brewing Co.	United States	3	22508	24356
41	Allagash Brewing Co	United States	3	32629	33973
9	Michelob ULTRA	United States	2	57886	292
30	Pike Brewing Company	United States	2	15229	5399
25	Lancaster Brewing Co	United States	1	9485	9698
26	Summit Brewing	United States	1	52433	9786
31	Lift Bridge Brewing	United States	1	17822	10101
3	BrewDog	United Kingdom	15	125732	26156
24	Goose Island Beer Co	United Kingdom	5	79502	35414
43	Meantime Brewing Co.	United Kingdom	2	33567	16697
15	Loddon Brewery	United Kingdom	1	7053	3649
35	Deschutes Brewery	Thailand	1	114047	30392
8	BrewDog	Norway	1	125732	26156

	name	country	count	followers_count	favourites_count
7	BrewDog	Nederland	2	125732	26156
12	Ballast Point Beer	Mexico	1	94783	10706
2	BrewDog	Italia	2	125732	26156
0	BrewDog	Iceland	1	125732	26156
33	MillerCoors	Dominican Republic	1	26413	6476
4	BrewDog	Denmark	1	125732	26156
16	Anchor Brewing	Canada	1	66906	14654
19	Blue Moon Brewing Co	Canada	1	52712	9631
42	Big Rock Brewery	Canada	1	21744	17378
1	BrewDog	Brazil	1	125732	26156
23	Goose Island Beer Co	Brasil	5	79502	35414
5	BrewDog	Belgium	1	125732	26156
39	Dogfish Head Brewery		2	298979	19915

```
In [103]: US = locality.query("country == 'United States'")
US = US.sort_values('count', ascending=False)
```

```
In [104]: sn.barplot(x='count', y='name', data=US)
```

```
Out[104]: <matplotlib.axes._subplots.AxesSubplot at 0x114220890>
```



```
In [106]: EA = locality.query("country != 'United States' & country != ''")
EA = EA.sort_values('count', ascending=False)
EA
```

Out[106]:

	name	country	count	followers_count	favourites_count
3	BrewDog	United Kingdom	15	125732	26156
23	Goose Island Beer Co	Brasil	5	79502	35414
24	Goose Island Beer Co	United Kingdom	5	79502	35414
43	Meantime Brewing Co.	United Kingdom	2	33567	16697
7	BrewDog	Nederland	2	125732	26156
2	BrewDog	Italia	2	125732	26156
4	BrewDog	Denmark	1	125732	26156
1	BrewDog	Brazil	1	125732	26156
42	Big Rock Brewery	Canada	1	21744	17378
19	Blue Moon Brewing Co	Canada	1	52712	9631
16	Anchor Brewing	Canada	1	66906	14654
0	BrewDog	Iceland	1	125732	26156
33	MillerCoors	Dominican Republic	1	26413	6476
12	Ballast Point Beer	Mexico	1	94783	10706
8	BrewDog	Norway	1	125732	26156
35	Deschutes Brewery	Thailand	1	114047	30392
15	Loddon Brewery	United Kingdom	1	7053	3649
5	BrewDog	Belgium	1	125732	26156

```
In [107]: sn.barplot(x='count', y='name', data=EA)
```

Out[107]: <matplotlib.axes._subplots.AxesSubplot at 0x114f99f50>

