In [1]:

```
import findspark
findspark.init('/opt/anaconda3/lib/python3.8/site-packages/pyspark')
import praw
import pandas as pd
import pyspark
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
import datetime
from datetime import datetime, timezone

import matplotlib.pyplot as plt


spark=SparkSession.builder.appName('Reddit').getOrCreate()
```

In [3]:

```

start_time=[1514786400,1522558800,1530421200,1538370000,1546322400,155
stop_time=[1522558800,1530421200,1538370000,1546322400,1554094800,1561

title_list=[]
author_list=[]
created_utc_list=[]
is_org_content=[]
score_list=[]
id_list=[]
sub_red_name=[]
flair_text=[]
over_18=[]
text_only=[]
url=[]

import requests
import json
import time

df=pd.DataFrame()
def getPushshiftData(start,stop):
    url = 'https://api.pushshift.io/reddit/search/submission/?title=Da
    print(url)
    r = requests.get(url)
    data = json.loads(r.text)
    return data['data']
```

In [30]:

```

#for i in range(14):
data=getPushshiftData(start_time[13],stop_time[13])
for submission in data:
    #if submission.created_utc>=1625115600 and submission.created_utc<
    title_list.append(submission['title'])
    author_list.append(submission['author'])
    created_utc_list.append(submission['created_utc'])
    if 'is_original_content' in submission:
        is_org_content.append(submission['is_original_content'])
    else:
        is_org_content.append("Null")
    score_list.append(submission['score'])
    id_list.append(submission['id'])
    sub_red_name.append(submission['subreddit'])
    over_18.append(submission['over_18'])
    text_only.append(submission['is_self'])
    url.append(submission['url'])
    if 'link_flair_text' in submission:
        flair_text.append(submission['link_flair_text'])
    else:
        flair_text.append("Null")
#time.sleep(60)


print(len(data))
df1=pd.DataFrame({'author':author_list,'created_utc':created_utc_list,
df=df.append(df1)
```

```
https://api.pushshift.io/reddit/search/submission/?title=Daily
(https://api.pushshift.io/reddit/search/submission/?title=Daily)
Discussion Thread&limit=1000&after=1617253200&before=1625115600&subre
ddit=wallstreetbets
75
```

In [106]:

```
#df=pd.DataFrame({'author':author_list,'created_utc':created_utc_list,
#df.head()
df=pd.read_csv('/Users/vyshnavigovindankutty/Downloads/reddit_wallstre
df=df.drop_duplicates()
df['author']=df['author'].astype(str)
df['sub_red']=df['sub_red'].astype(str)
df['is_original_content']=df['is_original_content'].astype(bool)
df['over_18']=df['over_18'].astype(bool)
df['text_only']=df['text_only'].astype(bool)

df_py=spark.createDataFrame(df)
df_py=df_py.withColumn('dte',to_timestamp('created_utc'))
df_py=df_py.withColumn('tme', date_format('dte', 'HH:mm:ss'))
df_py=df_py.withColumn('dte',to_date('dte'))

df_py.select(min('dte'),max('dte')).show()

#df.to_csv('/Users/vyshnavigovindankutty/Downloads/reddit_wallstreetda
```

```
+----------+----------+
|  min(dte)|  max(dte)|
+----------+----------+
|2018-01-01|2021-06-30|
+----------+----------+
```

In [50]:

```
import psaw
from psaw import PushshiftAPI
import pandas as pd
import praw

reddit=praw.Reddit(client_id='Hkfh-1zV5w-6J5Z5iAkkeg',client_secret='V
api=PushshiftAPI()
df=pd.DataFrame()
sub_red=['funny','AskReddit','gaming','worldnews','todayilearned','new
for s in sub_red:
    submiss=api.search_submissions(subreddit=s,filter=['url','author',
    df=df.append([sub.d_ for sub in submiss])

df.tail()
```

```
/opt/anaconda3/lib/python3.8/site-packages/psaw/PushshiftAPI.py:192:
UserWarning: Got non 200 code 429
  warnings.warn("Got non 200 code %s" % response.status_code)
/opt/anaconda3/lib/python3.8/site-packages/psaw/PushshiftAPI.py:180:
UserWarning: Unable to connect to pushshift.io. Retrying after backof
f.
  warnings.warn("Unable to connect to pushshift.io. Retrying after ba
ckoff.")
```

```
/opt/anaconda3/lib/python3.8/site-packages/psaw/PushshiftAPI.py:192:
UserWarning: Got non 200 code 525
  warnings.warn("Got non 200 code %s" % response.status_code)
/opt/anaconda3/lib/python3.8/site-packages/psaw/PushshiftAPI.py:192:
UserWarning: Got non 200 code 522
  warnings.warn("Got non 200 code %s" % response.status_code)
/opt/anaconda3/lib/python3.8/site-packages/psaw/PushshiftAPI.py:192:
UserWarning: Got non 200 code 502
  warnings.warn("Got non 200 code %s" % response.status_code)
```

Out[50]:

	author	created_utc	id	score	subreddit	title	
1456	ocdude	1201987680	67h11	1	recipes	Going back to my (non-existant) eastern Europe...	http://www.bluewavedigital.n
1457	morgan420	1202118072	67l3m	1	recipes	Kesari Kulfi - Pakistani Indian Ice Cream Dessert	http://freemancooks.blogspot.c
1458	hamsterboy	1201271622	66lcg	7	recipes	Recipe: No-Knead Bread - New York Times	http://www.nytimes.com/2006/1
1459	ocdude	1201987680	67h11	1	recipes	Going back to my (non-existant) eastern Europe...	http://www.bluewavedigital.n
1460	hamsterboy	1201271622	66lcg	7	recipes	Recipe: No-Knead Bread - New York Times	http://www.nytimes.com/2006/1

In [62]:

```
df.size
```

Out[62]:

```
2347029
```

In [59]:

```
df['author']=df['author'].astype(str)
df['subreddit']=df['subreddit'].astype(str)
df['is_original_content']=df['is_original_content'].astype(bool)

df_py=spark.createDataFrame(df)
df_py=df_py.withColumn('dte',to_timestamp('created_utc'))
df_py=df_py.withColumn('tme', date_format('dte', 'HH:mm:ss'))
df_py=df_py.withColumn('dte',to_date('dte'))
```

In [60]:

```

df_py.select(min('dte'),max('dte')).show()
```

```
+----------+----------+
|  min(dte)|  max(dte)|
+----------+----------+
|2006-02-28|2021-06-18|
+----------+----------+
```

In [127]:

```
_py=df_py.filter((df_py['title'].startswith('D'))|df_py['title'].start
_py=df_py.filter((df_py['title']!='Daily XRP discussion thread')&(df_p
_py=df_py.filter((df_py['id']!='hujg5i')&(df_py['id']!='hujh1v')&(df_p
_py.filter(df_py['title'].contains('Daily Discussion Thread')==False).
1=df_py.groupby('dte').agg(max('score').alias('score'))
_py=df_py.join(df1,['dte','score'],'inner')
_py=df_py.filter((df_py['id']!='fotcp9')&(df_py['id']!='f1u0ub')&(df_p

_py.groupby('dte').count().filter('count>1').show(20)
```

```
+----------------------------------------+
|substring(title, 0, 40)                 |
+----------------------------------------+
|Daily discussion thread for february 9|
+----------------------------------------+
```

In [124]:

```
#df_py.filter(df_py['title'].contains('Daily Penny Stock Discussion'))
#df_py.select('title').show(50,truncate=False)
#df_py.select('title').distinct().count()
#df_py.toPandas().to_csv('/Users/vyshnavigovindankutty/Downloads/Wall_
#df_py=df_py.groupby('author','created_utc','is_original_content','tit

#df1=df_py.groupby('dte').agg(max('score').alias('score'))
#df1.show()

#df_py=df_py.join(df1,['dte','score'],'inner')
#df_py.count()
#df_py.toPandas().to_csv('/Users/vyshnavigovindankutty/Downloads/Wall_
df_py.groupby('dte').count().filter('count>1').show(20)
```

```
+----------+-----+
|       dte|count|
+----------+-----+
|2020-03-25|    2|
|2021-01-30|    3|
|2020-02-10|    4|
|2021-01-04|    2|
|2020-08-03|    2|
|2020-09-15|    3|
|2021-01-08|    2|
|2021-01-12|    2|
+----------+-----+
```

In [141]:

```
## Comments

id1=df_py.select('id').collect()
ids=[row.id for row in df_py.select('id').collect()]

reddit=praw.Reddit(client_id='Hkfh-1zV5w-6J5Z5iAkkeg',client_secret='V
comm_all=[]
comm=[]
created_utc=[]
n=0;
```

In [142]:

```
for i in ids:
    n=n+1
    subm=reddit.submission(id=i)
    subm.comments.replace_more(limit=0)
    comm=list([(comment.body) for comment in subm.comments])
    for comment in subm.comments:
        #if isinstance(comment, MoreComments):
        #    continue
        comm.append(comment.body)
        created_utc.append(comment.created_utc)
```

In [66]:

```
## Converting downloaded data to Pyspark
df_comm=pd.DataFrame({'comm':comm,'created_utc':created_utc})
df_comm.head()

py_comm=spark.createDataFrame(df_comm)
py_comm=py_comm.withColumn('dte',to_timestamp('created_utc'))
py_comm=py_comm.withColumn('tme', date_format('dte', 'HH:mm:ss'))
py_comm=py_comm.withColumn('dte',to_date('dte'))
#py_comm.show()

py_comm.printSchema()
#py_comm.select('comm').filter(py_comm['comm'].contains('$')).count()
#py_comm.count()
```

```
root
 |-- comm: string (nullable = true)
 |-- created_utc: string (nullable = true)
 |-- dte: date (nullable = true)
 |-- tme: string (nullable = true)
```

In [73]:

```
## Reading from local

df_comm=pd.read_csv('/Users/vyshnavigovindankutty/Desktop/AWS/Reddit_A
df_comm['created_utc']=df_comm['created_utc'].astype(int)
df_comm['comm']=df_comm['comm'].astype(str)
df_comm=df_comm[['comm','created_utc']]

py_comm=spark.createDataFrame(df_comm)
py_comm=py_comm.withColumn('dte',to_timestamp('created_utc'))
py_comm=py_comm.withColumn('tme', date_format('dte', 'HH:mm:ss'))
py_comm=py_comm.withColumn('dte',to_date('dte'))
```

In [3]:

```
## Sentiment Analysis Function

from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

def senti_score():
    obj=SentimentIntensityAnalyzer()
    senti_dict=[]
    for comment in comm:
        s=0

        for c in comment:
            s=s+obj.polarity_scores(c)['compound']

        senti_dict.append(s)
    return senti_dict
```

In [82]:

```
## Getting the sentiments score for comments

df1=py_comm.toPandas()
df1=df1.groupby('dte')['comm'].apply(list)
df1=df1.reset_index(name='comm')
comm=[]
comm=df1['comm'].tolist()
senti_dict=senti_score()
df1['score']=senti_dict
#df1.head()
```

In [227]:

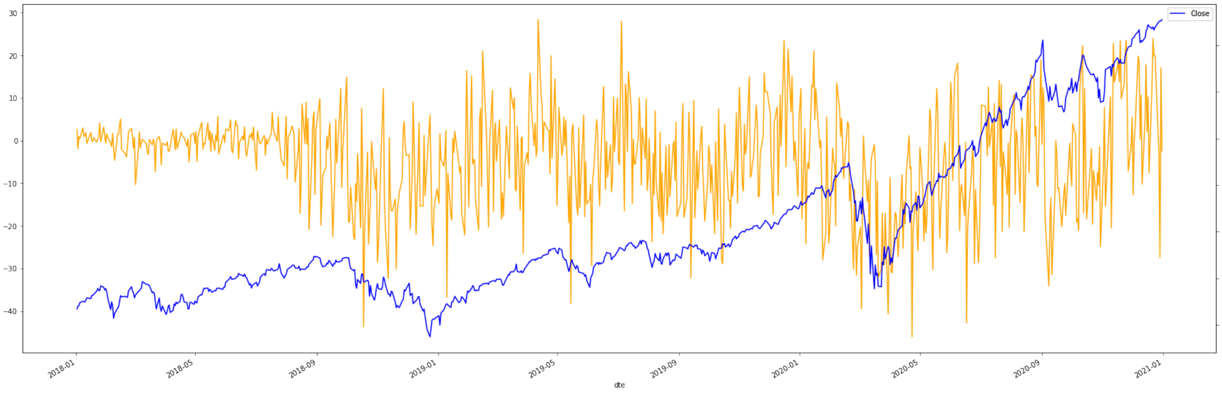
```
## Finance data for NASDAQ-100 -QQQ index

import pandas_datareader.data as web
import pandas as pd

df_finance = web.DataReader('QQQ', 'yahoo', start='2018-01-01', end='2
df_finance.head()

df_finance['dte']=df_finance.index
df1['dte']=pd.to_datetime(df1['dte'])
df1=pd.merge(df1,df_finance,on=['dte'])
```

Out[227]:

	High	Low	Open	Close	Volume	Adj Close
Date						
2018-01-02	158.529999	156.169998	156.559998	158.490005	32573300.0	154.620209
2018-01-03	160.169998	158.610001	158.639999	160.029999	29383600.0	156.122574
2018-01-04	160.789993	160.080002	160.580002	160.309998	24776100.0	156.395767
2018-01-05	162.029999	160.770004	161.070007	161.919998	26992300.0	157.966446
2018-01-08	162.630005	161.860001	161.919998	162.550003	23159100.0	158.581085

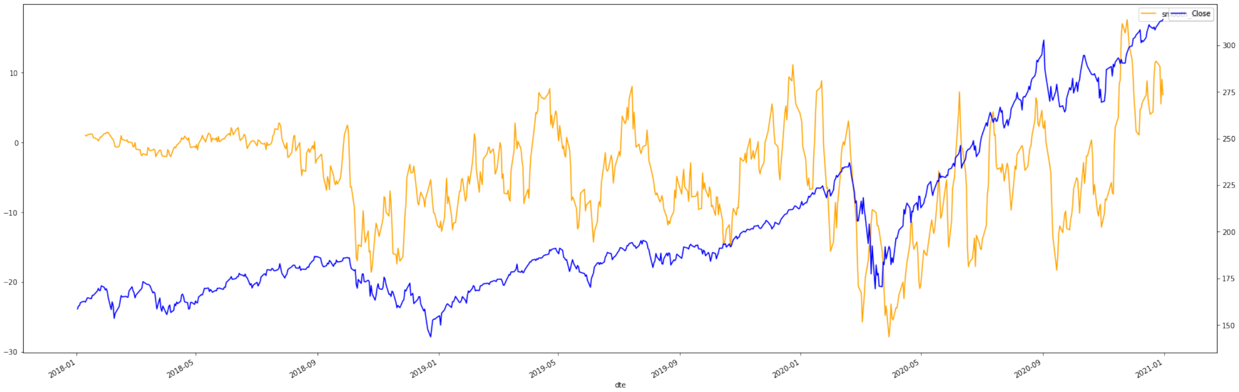
In [288]:

```
## Original plot

fig = plt.figure()
ax = fig.add_subplot(111)
ax2 = ax.twinx()

df1.plot(y='score',x='dte',color='orange',figsize=(30, 10),ax=ax)
df1.plot(y='Close',x='dte',color='blue',figsize=(30, 10), ax=ax2)
plt.show()
```



In [289]:

```
## Plotting score vs price- smoothed version

import numpy as np
import matplotlib.pyplot as plt

fig = plt.figure()
ax = fig.add_subplot(111)
ax2 = ax.twinx()

#df1.plot(y='score',x='dte',color='blue', figsize=(30, 10),ax=ax)
y_y=list(df1['score'])
smooth_data = pd.Series(y_y).rolling(window=7).mean()
df2=pd.DataFrame()
df2['smooth_score']=smooth_data
df2['dte']=df1['dte']

df2.plot(y='smooth_score',x='dte',color='orange',figsize=(30, 10),ax=a
df1.plot(y='Close',x='dte',color='blue',figsize=(30, 10), ax=ax2)
plt.show()
```

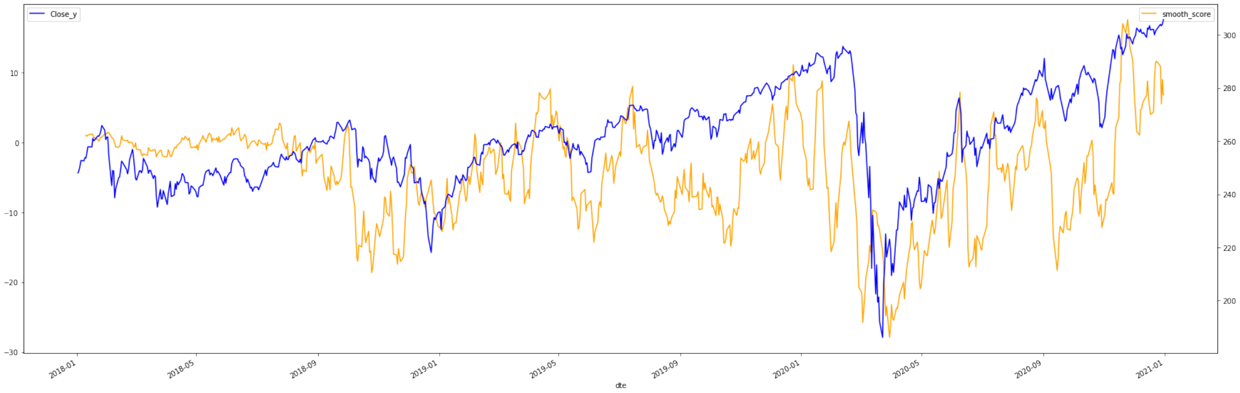


In [296]:

```
## Finance data DowJones- DIA index

import pandas_datareader.data as web
import pandas as pd

df_finance1 = web.DataReader('DIA', 'yahoo', start='2018-01-01', end='
df_finance1.head()

df_finance1['dte']=df_finance1.index
df2=df1
df2['dte']=pd.to_datetime(df2['dte'])
df2=pd.merge(df2,df_finance1,on=['dte'])
```

In [297]:

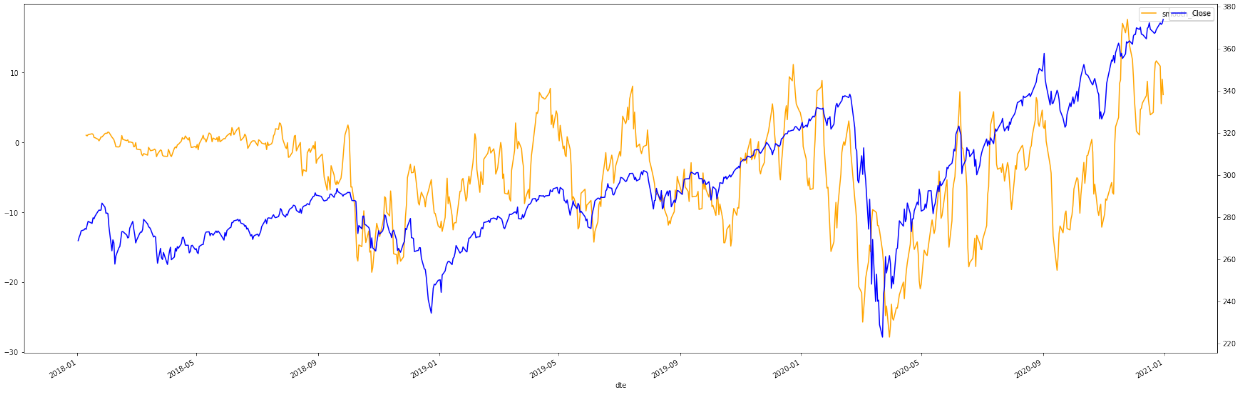
```
fig = plt.figure()
ax = fig.add_subplot(111)
ax2 = ax.twinx()

y_y=list(df2['score'])
smooth_data = pd.Series(y_y).rolling(window=7).mean()
df3=pd.DataFrame()
df3['smooth_score']=smooth_data
df3['dte']=df2['dte']
#x_x=list(df1['dte'])
df3.plot(y='smooth_score',x='dte',color='orange',figsize=(30, 10),ax=a
df2.plot(y='Close_y',x='dte',color='blue',figsize=(30, 10), ax=ax2)
plt.show()
```



In [84]:

```
## Finance data S&P500- SPY index

import pandas_datareader.data as web
import pandas as pd

df_finance1 = web.DataReader('SPY', 'yahoo', start='2018-01-01', end='
df_finance1.head()

df_finance1['dte']=df_finance1.index
df2=df1
df2['dte']=pd.to_datetime(df2['dte'])
df2=pd.merge(df2,df_finance1,on=['dte'])
```

In [87]:
```
fig = plt.figure()
ax = fig.add_subplot(111)
ax2 = ax.twinx()

y_y=list(df2['score'])
smooth_data = pd.Series(y_y).rolling(window=7).mean()
df3=pd.DataFrame()
df3['smooth_score']=smooth_data
df3['dte']=df2['dte']
#x_x=list(df1['dte'])
df3.plot(y='smooth_score',x='dte',color='orange',figsize=(30, 10),ax=a
df2.plot(y='Close',x='dte',color='blue',figsize=(30, 10), ax=ax2)
plt.show()
```



In [ ]:

In [ ]:

In [56]:
```
df
```

Out[56]:
```
2347029
```

In [65]:
```
df_py.filter((df_py['subreddit']=='pics')|(df_py['subreddit']=='wallst
```

```
+----------+----------+
|min(score)|max(score)|
+----------+----------+
|         0|    234711|
+----------+----------+
```

In [66]:
```
df_py.filter(df_py['subreddit']=='pics').sort(col('score').desc()).sho
```

```
+-------------------+-----------+------+------+---------+------------
--------+--------------------+-------------+-------------------+-----
-----+--------+
|             author|created_utc|    id| score|subreddit|            
title|                url|      created|is_original_content|       d
te|     tme|
+-------------------+-----------+------+------+---------+------------
--------+--------------------+-------------+-------------------+-----
-----+--------+
|          pdmcmahon| 1490560408|61ns2w|234711|     pics|Private Inte
rnet ...|http://i.imgur.co...|1.490578408E9|               true|2017-
03-26|15:33:28|
|          pdmcmahon| 1490560408|61ns2w|234711|     pics|Private Inte
rnet ...|http://i.imgur.co...|1.490578408E9|               true|2017-
03-26|15:33:28|
|    Theon_Graystark| 1591709669|gzn8tj|217126|     pics|At a protest
 in A...|https://i.redd.it...|1.591727669E9|              false|2020-0
6-09|08:34:29|
|          cursetenj| 1503357584|6v6y1k|210005|     pics|Hello Reddit
. I a...|https://i.redd.it...|1.503375584E9|               true|2017-
08-21|18:19:44|
|            Tibujon| 1505015888|6z699p|203110|     pics|My dad waits
 ever...|https://www.flick...|1.505033888E9|               true|2017-0
9-09|22:58:08|
|       effectivepep| 1550366784|arfjs5|202696|     pics|This person 
sold ...|https://i.imgur.c...|1.550384784E9|              false|2019-
02-16|19:26:24|
|        jacques4801| 1527043084|8lfwjv|201853|     pics|Found an old
 lett...|https://i.redd.it...|1.527061084E9|              false|2018-0
5-22|21:38:04|
|              Goal1| 1535988082|9cmazx|198234|     pics|They noticed
 ther...|https://i.redd.it...|1.536006082E9|              false|2018-0
9-03|10:21:22|
|  Itsjorgehernandez| 1478651245|5bx4bx|196587|     pics|      Thanks
, Obama.|https://i.redditu...|1.478669245E9|               true|2016-
11-08|18:27:25|
|  Itsjorgehernandez| 1478651245|5bx4bx|196587|     pics|      Thanks
, Obama.|https://i.redditu...|1.478669245E9|               true|2016-
11-08|18:27:25|
|  Itsjorgehernandez| 1478651245|5bx4bx|196587|     pics|      Thanks
, Obama.|https://i.redditu...|1.478669245E9|               true|2016-
11-08|18:27:25|
|  Itsjorgehernandez| 1478651245|5bx4bx|196587|     pics|      Thanks
, Obama.|https://i.redditu...|1.478669245E9|               true|2016-
11-08|18:27:25|
|             iKojan| 1554926298|bbql1i|194651|     pics|This is Dr K
atie ...|https://i.redd.it...|1.554944298E9|              false|2019-
04-10|14:58:18|
|          datbanter| 1486400800|5sfexx|193491|     pics|This is Shel
ia Fr...|https://i.redditu...|  1.4864188E9|               true|2017-
02-06|11:06:40|
|          datbanter| 1486400800|5sfexx|193491|     pics|This is Shel
ia Fr...|https://i.redditu...|  1.4864188E9|               true|2017-
```

```
02-06|11:06:40|
|AWildSketchAppeared| 1496153840|6e7luk|192394|     pics|And here it
is: L...|http://imgur.com/...| 1.49617184E9|              true|2017-
05-30|09:17:20|
|AWildSketchAppeared| 1496153840|6e7luk|192394|     pics|And here it
is: L...|http://imgur.com/...| 1.49617184E9|              true|2017-
05-30|09:17:20|
|       JaMollyAdams| 1534265171|979ywc|188329|     pics|A bird flew
in my...|https://i.imgur.c...|1.534283171E9|             false|2018-
08-14|11:46:11|
|          zombi3123| 1530017983|8tzsyf|188250|     pics|I work in a
kitch...|https://i.redd.it...|1.530035983E9|             false|2018-
06-26|07:59:43|
|        El_Torito23| 1534975156|99hl0j|183785|     pics|Teachers hom
ework...|https://i.redd.it...|1.534993156E9|             false|2018-
08-22|16:59:16|
+-------------------+-----------+------+------+---------+------------
--------+--------------------+-------------+------------------+-----
-----+--------+
only showing top 20 rows
```

In [100]:

```
reddit=praw.Reddit(client_id='Hkfh-1zV5w-6J5Z5iAkkeg',client_secret='V
api=PushshiftAPI()
df3=pd.DataFrame()
submiss=api.search_submissions(after=1626393600,subreddit="pics",sort_
    #'author', 'title', 'subreddit','score','is_original_content','id'
df3=df3.append([sub.d_ for sub in submiss])

#df1=pd.DataFrame({'author':author_list,'created_utc':created_utc_list
```

```
/opt/anaconda3/lib/python3.8/site-packages/psaw/PushshiftAPI.py:192:
UserWarning: Got non 200 code 525
  warnings.warn("Got non 200 code %s" % response.status_code)
/opt/anaconda3/lib/python3.8/site-packages/psaw/PushshiftAPI.py:180:
UserWarning: Unable to connect to pushshift.io. Retrying after backof
f.
  warnings.warn("Unable to connect to pushshift.io. Retrying after ba
ckoff.")
```

In [105]:

```
df4.show()
```

```
+---------------+-----------+------+------+---------+----------------
----+--------------------+-------------+------------------+---------
-+--------+
|         author|created_utc|    id| score|subreddit|               t
itle|                 url|      created|is_original_content|       dt
e|     tme|
+---------------+-----------+------+------+---------+----------------
----+--------------------+-------------+-------------------+---------
-+--------+
|       the_Diva| 1514430055|7mjw12|255208|    funny|My cab driver to
n...|https://i.redd.it...|1.514448055E9|               true|2017-12-2
7|21:00:55|
|        iH8myPP| 1480959674|5gn8ru|222208|    funny|Guardians of the
...|http://i.imgur.co...|1.480977674E9|               true|2016-12-05
|11:41:14|
|        namraka| 1507061450|7431qq|218358|    funny|Gas station work
e...|https://gfycat.co...| 1.50707945E9|               true|2017-10-0
3|15:10:50|
|        Romobyl| 1513711444|7kvjuz|204953|    funny|The conversation
...|https://i.imgur.c...|1.513729444E9|               true|2017-12-19
|13:24:04|
|deadleaf_shrimp| 1535746569|9bx0o9|184761|    funny|I get an email e
v...|https://i.redd.it...|1.535764569E9|              false|2018-08-3
1|15:16:09|
|      [deleted]| 1506465777|72o2zv|182767|    funny|Gonna request a 
r...|https://i.redd.it...|1.506483777E9|               true|2017-09-2
6|17:42:57|
|        iH8myPP| 1501332999|6qatmn|180654|    funny| Reddit's Immigr
ants|http://i.imgur.co...|1.501350999E9|               true|2017-07-2
9|07:56:39|
|  MalletsDarker| 1516282537|7r9ptc|180520|    funny|I took a few sho
t...|http://imgur.com/...|1.516300537E9|               true|2018-01-1
8|07:35:37|
|        guyi567| 1516116337|7qt032|178459|    funny|These damn ads a
r...|https://gfycat.co...|1.516134337E9|               true|2018-01-1
6|09:25:37|
|  foreverwasted| 1536242219|9ditu6|173629|    funny|Bill Burr on Goo
d...|https://imgur.com...|1.536260219E9|              false|2018-09-0
6|08:56:59|
|  TheDarkLord66| 1510671126|7cw2m9|171326|    funny|UPDATE. EA annou
n...|https://i.redd.it...|1.510689126E9|               true|2017-11-1
4|08:52:06|
|        alebrew| 1570996151|dhfigp|169588|    funny|Irish man leaves
...|https://v.redd.it...|1.571014151E9|              false|2019-10-13
|14:49:11|
| System32Comics| 1568139947|d2bwot|169177|    funny|            Prin
ters|https://i.redd.it...|1.568157947E9|              false|2019-09-1
0|13:25:47|
|     ronlechler| 1528553932|8pt2oc|168878|    funny|Shoutout to the 
1...|https://i.redd.it...|1.528571932E9|              false|2018-06-0
9|09:18:52|
|     bsurfn2day| 1513315334|7jxoon|166861|    funny|Bollywood at it 
f...|https://i.imgur.c...|1.513333334E9|               true|2017-12-1
4|23:22:14|
|      Timber371| 1572633946|dq8gm7|166507|    funny|My son happened 
a...|https://v.redd.it...|1.572651946E9|              false|2019-11-0
```

```
1|13:45:46|
|EverythingFerns| 1515508242|7p7ffi|164758|    funny|I found a bunch
o...|https://gfycat.co...|1.515526242E9|               true|2018-01-0
9|08:30:42|
|    Ginger_King| 1537390686|9h905h|161346|    funny|Today was "Meme
D...|https://i.redd.it...|1.537408686E9|              false|2018-09-1
9|15:58:06|
|therightcoaster| 1528158123|8omduo|160117|    funny|Girl takes cardb
o...|https://i.imgur.c...|1.528176123E9|              false|2018-06-0
4|19:22:03|
|         PNWndn| 1516069158|7qp34o|159718|    funny|   I'm that sib
ling|https://v.redd.it...|1.516087158E9|               true|2018-01-1
5|20:19:18|
+---------------+-----------+------+------+---------+----------------
----+--------------------+-------------+-------------------+---------
-+--------+
only showing top 20 rows
```

In [103]:

```
df3['author']=df3['author'].astype(str)
df3['subreddit']=df3['subreddit'].astype(str)
df3['is_original_content']=df3['is_original_content'].astype(bool)

df4=spark.createDataFrame(df)
df4=df4.withColumn('dte',to_timestamp('created_utc'))
df4=df4.withColumn('tme', date_format('dte', 'HH:mm:ss'))
df4=df4.withColumn('dte',to_date('dte'))
```

In [80]:

```python
## Write Data to S3
from io import StringIO
import boto3

buffer=StringIO()
df.to_csv(buffer)
s3_resource=boto3.resource('s3')
s3_resource.Object('employeee-bucket','reddit_big_data.csv').put(Body=
```

Out[80]:

```
{'ResponseMetadata': {'RequestId': 'JAWW5HKDMS59DCSZ',
  'HostId': 'RGpYMcF8bdXLJqwUwvgop0wgD/8/xFRAq2CgeswWCwI4hsAOhUcXPDGy
PpXMhBa/tPedYGwYkoU=',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'x-amz-id-2': 'RGpYMcF8bdXLJqwUwvgop0wgD/8/xFRAq2Cg
eswWCwI4hsAOhUcXPDGyPpXMhBa/tPedYGwYkoU=',
   'x-amz-request-id': 'JAWW5HKDMS59DCSZ',
   'date': 'Tue, 20 Jul 2021 15:30:30 GMT',
   'etag': '"35601a25abf3084385f4a20a02b526dd"',
   'server': 'AmazonS3',
   'content-length': '0'},
  'RetryAttempts': 0},
 'ETag': '"35601a25abf3084385f4a20a02b526dd"'}
```

In [86]:

```
## Import from S3

AWS_S3_BUCKET = "employeee-bucket"
AWS_ACCESS_KEY_ID = "AKIATLAVBA6KZGDRBTXL"
AWS_SECRET_ACCESS_KEY = "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"

s3_client = boto3.client(
    "s3",
    aws_access_key_id=AWS_ACCESS_KEY_ID,
    aws_secret_access_key=AWS_SECRET_ACCESS_KEY,
)

response = s3_client.get_object(Bucket=AWS_S3_BUCKET, Key="reddit_big_

df2=pd.read_csv(response.get("Body"))
```

Out[86]:

	Unnamed: 0	author	created_utc	id	score	subreddit	title	
0	0	the_Diva	1514430055	7mjw12	255208	funny	My cab driver tonight was so excited to share ...	
1	1	iH8myPP	1480959674	5gn8ru	222208	funny	Guardians of the Front Page	
2	2	namraka	1507061450	7431qq	218358	funny	Gas station worker takes precautionary measure...	https://gfy
3	3	Romobyl	1513711444	7kvjuz	204953	funny	The conversation my son and I will have on Chr...	
4	4	deadleaf_shrimp	1535746569	9bx0o9	184761	funny	I get an email every time I get a package deli...	

In [ ]:

In [ ]:

In [ ]:

```
In [ ]:
```