Project 1 Assignment

February 24, 2022

Jay Desmarais CMSC 320 Project 1

Part 1: Data scraping and preparation

Step 1: Scrape your competitor's data

[81]:

```
# 1. Necessary utilities are imported.
import requests
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import re
from datetime import datetime
from bs4 import BeautifulSoup
```

[82]:

```
# 2. HTTP GET request from a mirror of a SpaceWeatherLive.com archive.
r = requests.get("https://cmsc320.github.io/files/top-50-solar-flares.html")
# 3. The text is extracted from the HTTP GET request response.
extract = r.content
# 4. Raw content from SpaceWatherLive.com is parsed.
root = BeautifulSoup(extract, 'lxml')
# 5. The HTML content is passed into Prettify so we can locate the desired␣
 ↪content.
pretty = root.prettify()
# 6. The HTML Table is found and isolated as a str type.
html = str(root.find("table"))
# 7. The HTML is converted into a pandas DataFrame.
swl = pd.read_html(html)[0]
# 8. Names are assigned to all columns of the DataFrame.
swl.columns = [
    'rank',
    'x_classification',
    'date',
    'region',
    'start_time',
    'max_time',
    'end_time',
    'movie'
]
```

```
# Display the final DataFrame at the end of this step.
swl
```

[82]:

```
    rank x_classification        date  region start_time max_time end_time  \
0      1             X28+  2003/11/04     486      19:29    19:53    20:06
1      2             X20+  2001/04/02    9393      21:32    21:51    22:03
2      3           X17.2+  2003/10/28     486      09:51    11:10    11:24
3      4             X17+  2005/09/07     808      17:17    17:40    18:03
4      5            X14.4  2001/04/15    9415      13:19    13:50    13:55
5      6              X10  2003/10/29     486      20:37    20:49    21:01
6      7             X9.4  1997/11/06    8100      11:49    11:55    12:01
7      8             X9.3  2017/09/06    2673      11:53    12:02    12:10
8      9               X9  2006/12/05     930      10:18    10:35    10:45
9     10             X8.3  2003/11/02     486      17:03    17:25    17:39
10    11             X8.2  2017/09/10    2673      15:35    16:06    16:31
11    12             X7.1  2005/01/20     720      06:36    07:01    07:26
12    13             X6.9  2011/08/09    1263      07:48    08:05    08:08
13    14             X6.5  2006/12/06     930      18:29    18:47    19:00
14    15             X6.2  2005/09/09     808      19:13    20:04    20:36
15    16             X6.2  2001/12/13    9733      14:20    14:30    14:35
16    17             X5.7  2000/07/14    9077      10:03    10:24    10:43
17    18             X5.6  2001/04/06    9415      19:10    19:21    19:31
18    19             X5.4  2012/03/07    1429      00:02    00:24    00:40
19    20             X5.4  2005/09/08     808      20:52    21:06    21:17
20    21             X5.4  2003/10/23     486      08:19    08:35    08:49
21    22             X5.3  2001/08/25    9591      16:23    16:45    17:04
22    23             X4.9  2014/02/25    1990      00:39    00:49    01:03
23    24             X4.9  1998/08/18    8307      22:10    22:19    22:28
24    25             X4.8  2002/07/23      39      00:18    00:35    00:47
25    26               X4  2000/11/26    9236      16:34    16:48    16:56
26    27             X3.9  2003/11/03     488      09:43    09:55    10:19
27    28             X3.9  1998/08/19    8307      21:35    21:45    21:50
28    29             X3.8  2005/01/17     720      06:59    09:52    10:07
29    30             X3.7  1998/11/22    8384      06:30    06:42    06:49
30    31             X3.6  2005/09/09     808      09:42    09:59    10:08
31    32             X3.6  2004/07/16     649      13:49    13:55    14:01
32    33             X3.6  2003/05/28     365      00:17    00:27    00:39
33    34             X3.4  2006/12/13     930      02:14    02:40    02:57
34    35             X3.4  2001/12/28    9767      20:02    20:45    21:32
35    36             X3.3  2013/11/05    1890      22:07    22:12    22:15
36    37             X3.3  2002/07/20      39      21:04    21:30    21:54
37    38             X3.3  1998/11/28    8395      04:54    05:52    06:13
38    39             X3.2  2013/05/14    1748      00:00    01:11    01:20
39    40             X3.1  2014/10/24    2192      21:07    21:41    22:13
40    41             X3.1  2002/08/24      69      00:49    01:12    01:31
41    42               X3  2002/07/15      30      19:59    20:08    20:14
42    43             X2.8  2013/05/13    1748      15:48    16:05    16:16
```

```
43    44             X2.8  2001/12/11    9733      07:58    08:08    08:14
44    45             X2.8  1998/08/18    8307      08:14    08:24    08:32
45    46             X2.7  2015/05/05    2339      22:05    22:11    22:15
46    47             X2.7  2003/11/03     488      01:09    01:30    01:45
47    48             X2.7  1998/05/06    8210      07:58    08:09    08:20
48    49             X2.6  2005/01/15     720      22:25    23:02    23:31
49    50             X2.6  2001/09/24    9632      09:32    10:38    11:09

                movie
0   MovieView archive
1   MovieView archive
2   MovieView archive
3   MovieView archive
4   MovieView archive
5   MovieView archive
6   MovieView archive
7   MovieView archive
8   MovieView archive
9   MovieView archive
10  MovieView archive
11  MovieView archive
12  MovieView archive
13  MovieView archive
14  MovieView archive
15  MovieView archive
16  MovieView archive
17  MovieView archive
18  MovieView archive
19  MovieView archive
20  MovieView archive
21  MovieView archive
22  MovieView archive
23       View archive
24  MovieView archive
25  MovieView archive
26  MovieView archive
27       View archive
28  MovieView archive
29  MovieView archive
30  MovieView archive
31  MovieView archive
32  MovieView archive
33  MovieView archive
34  MovieView archive
35  MovieView archive
36  MovieView archive
37  MovieView archive
```

```
38  MovieView archive
39  MovieView archive
40  MovieView archive
41  MovieView archive
42  MovieView archive
43  MovieView archive
44       View archive
45  MovieView archive
46  MovieView archive
47  MovieView archive
48  MovieView archive
49  MovieView archive
```

Step 2: Tidy the top 50 solar flare data

```
[83]: # 1. The movie column is dropped from the DataFrame.
swl.drop('movie', axis=1, inplace=True)
# 2. Each of the time columns in the DataFrame are combined with the date␣
 ↪column to create new datetime columns.
for column in ['end', 'max', 'start']:
    swl.insert(
        loc = 2,
        column = column + '_datetime',
        value = pd.to_datetime(swl['date'] + swl[column + '_time'], format='%Y/
 ↪%m/%d%H:%M')
    )
    # The time columns are dropped from the DataFrame.
    swl.drop(column + '_time', axis = 1, inplace=True)
# The date column is dropped from the DataFrame.
swl.drop('date', axis=1, inplace=True)
# 3. Remove or replace any "+" in the x_classification column if the preceding␣
 ↪number contains a decimal place or not, resepctively.
swl = swl\
    .replace(regex={r'(^X{1}[0-9].\.[0-9]?)\+$' : r'\1', r'(^X{1}[0-9]+)\+$' :␣
 ↪r'\1.0', r'(^X{1}[0-9\.]+)\+$' : r'\1'})
# 4. Any instances of "-" in the DataFrame are replaced with NaN.
swl = swl.replace('-', np.nan)
# Display the final DataFrame at the end of this step.
swl
```

```
[83]:    rank x_classification      start_datetime        max_datetime  \
0      1            X28.0 2003-11-04 19:29:00 2003-11-04 19:53:00
1      2            X20.0 2001-04-02 21:32:00 2001-04-02 21:51:00
2      3            X17.2 2003-10-28 09:51:00 2003-10-28 11:10:00
3      4            X17.0 2005-09-07 17:17:00 2005-09-07 17:40:00
4      5            X14.4 2001-04-15 13:19:00 2001-04-15 13:50:00
5      6              X10 2003-10-29 20:37:00 2003-10-29 20:49:00
```

```
6      7             X9.4 1997-11-06 11:49:00 1997-11-06 11:55:00
7      8             X9.3 2017-09-06 11:53:00 2017-09-06 12:02:00
8      9               X9 2006-12-05 10:18:00 2006-12-05 10:35:00
9     10             X8.3 2003-11-02 17:03:00 2003-11-02 17:25:00
10    11             X8.2 2017-09-10 15:35:00 2017-09-10 16:06:00
11    12             X7.1 2005-01-20 06:36:00 2005-01-20 07:01:00
12    13             X6.9 2011-08-09 07:48:00 2011-08-09 08:05:00
13    14             X6.5 2006-12-06 18:29:00 2006-12-06 18:47:00
14    15             X6.2 2005-09-09 19:13:00 2005-09-09 20:04:00
15    16             X6.2 2001-12-13 14:20:00 2001-12-13 14:30:00
16    17             X5.7 2000-07-14 10:03:00 2000-07-14 10:24:00
17    18             X5.6 2001-04-06 19:10:00 2001-04-06 19:21:00
18    19             X5.4 2012-03-07 00:02:00 2012-03-07 00:24:00
19    20             X5.4 2005-09-08 20:52:00 2005-09-08 21:06:00
20    21             X5.4 2003-10-23 08:19:00 2003-10-23 08:35:00
21    22             X5.3 2001-08-25 16:23:00 2001-08-25 16:45:00
22    23             X4.9 2014-02-25 00:39:00 2014-02-25 00:49:00
23    24             X4.9 1998-08-18 22:10:00 1998-08-18 22:19:00
24    25             X4.8 2002-07-23 00:18:00 2002-07-23 00:35:00
25    26               X4 2000-11-26 16:34:00 2000-11-26 16:48:00
26    27             X3.9 2003-11-03 09:43:00 2003-11-03 09:55:00
27    28             X3.9 1998-08-19 21:35:00 1998-08-19 21:45:00
28    29             X3.8 2005-01-17 06:59:00 2005-01-17 09:52:00
29    30             X3.7 1998-11-22 06:30:00 1998-11-22 06:42:00
30    31             X3.6 2005-09-09 09:42:00 2005-09-09 09:59:00
31    32             X3.6 2004-07-16 13:49:00 2004-07-16 13:55:00
32    33             X3.6 2003-05-28 00:17:00 2003-05-28 00:27:00
33    34             X3.4 2006-12-13 02:14:00 2006-12-13 02:40:00
34    35             X3.4 2001-12-28 20:02:00 2001-12-28 20:45:00
35    36             X3.3 2013-11-05 22:07:00 2013-11-05 22:12:00
36    37             X3.3 2002-07-20 21:04:00 2002-07-20 21:30:00
37    38             X3.3 1998-11-28 04:54:00 1998-11-28 05:52:00
38    39             X3.2 2013-05-14 00:00:00 2013-05-14 01:11:00
39    40             X3.1 2014-10-24 21:07:00 2014-10-24 21:41:00
40    41             X3.1 2002-08-24 00:49:00 2002-08-24 01:12:00
41    42               X3 2002-07-15 19:59:00 2002-07-15 20:08:00
42    43             X2.8 2013-05-13 15:48:00 2013-05-13 16:05:00
43    44             X2.8 2001-12-11 07:58:00 2001-12-11 08:08:00
44    45             X2.8 1998-08-18 08:14:00 1998-08-18 08:24:00
45    46             X2.7 2015-05-05 22:05:00 2015-05-05 22:11:00
46    47             X2.7 2003-11-03 01:09:00 2003-11-03 01:30:00
47    48             X2.7 1998-05-06 07:58:00 1998-05-06 08:09:00
48    49             X2.6 2005-01-15 22:25:00 2005-01-15 23:02:00
49    50             X2.6 2001-09-24 09:32:00 2001-09-24 10:38:00

          end_datetime  region
0  2003-11-04 20:06:00     486
```

```
1  2001-04-02 22:03:00    9393
2  2003-10-28 11:24:00     486
3  2005-09-07 18:03:00     808
4  2001-04-15 13:55:00    9415
5  2003-10-29 21:01:00     486
6  1997-11-06 12:01:00    8100
7  2017-09-06 12:10:00    2673
8  2006-12-05 10:45:00     930
9  2003-11-02 17:39:00     486
10 2017-09-10 16:31:00    2673
11 2005-01-20 07:26:00     720
12 2011-08-09 08:08:00    1263
13 2006-12-06 19:00:00     930
14 2005-09-09 20:36:00     808
15 2001-12-13 14:35:00    9733
16 2000-07-14 10:43:00    9077
17 2001-04-06 19:31:00    9415
18 2012-03-07 00:40:00    1429
19 2005-09-08 21:17:00     808
20 2003-10-23 08:49:00     486
21 2001-08-25 17:04:00    9591
22 2014-02-25 01:03:00    1990
23 1998-08-18 22:28:00    8307
24 2002-07-23 00:47:00      39
25 2000-11-26 16:56:00    9236
26 2003-11-03 10:19:00     488
27 1998-08-19 21:50:00    8307
28 2005-01-17 10:07:00     720
29 1998-11-22 06:49:00    8384
30 2005-09-09 10:08:00     808
31 2004-07-16 14:01:00     649
32 2003-05-28 00:39:00     365
33 2006-12-13 02:57:00     930
34 2001-12-28 21:32:00    9767
35 2013-11-05 22:15:00    1890
36 2002-07-20 21:54:00      39
37 1998-11-28 06:13:00    8395
38 2013-05-14 01:20:00    1748
39 2014-10-24 22:13:00    2192
40 2002-08-24 01:31:00      69
41 2002-07-15 20:14:00      30
42 2013-05-13 16:16:00    1748
43 2001-12-11 08:14:00    9733
44 1998-08-18 08:32:00    8307
45 2015-05-05 22:15:00    2339
46 2003-11-03 01:45:00     488
47 1998-05-06 08:20:00    8210
```

```
48 2005-01-15 23:31:00     720
49 2001-09-24 11:09:00    9632
```

Step 3: Scrape the NASA data

```
[110]: # 1. HTTP GET request from a mirror of a NASA's catalog.
r = requests.get("https://cmsc320.github.io/files/waves_type2.html")
# Raw content from SpaceWatherLive.com is parsed.
root = BeautifulSoup(r.content, 'html')
# The HTML is found and isolated by line as a list strings.
content = str.split(str(root), '\n')
# 2. Regex is defined to find and isolate different columns.
regex = re\
    .compile(r'(\d{4}\/\d{2}\/\d{2}) (\d{2}:\d{2}) (\d{2}\/\d{2}) (\d{2}:\d{2})␣
 ↪*(?:<[^>]*>)?(\w*|\?*)(?:<[^>]*>)? *(?:<[^>]*>)?(\w+|\?{4})(?:<[^>]*>)?␣
 ↪*(\w+\??|-{6}) *(\w+\??|-{5}) *(\w+\.?\d?|-{4}) *(?:<[^>]*>)?(\d+\/\d+|-{2}\/
 ↪-{2}|\w)(?:<[^>]*>)? *(\d{2}:\d{2}|-{2}:-{2}|\w+) *(\w+|-+) *((?:&gt\;)?
 ↪\d+|-+) *(?:<[^>]*>)?(\w+|-{2}\/-{2}|-{4})(?:<[^>]*>)?')
# If a line matches the above regex, the columns are seperated and added to a␣
 ↪new list.
html = []
for line in content:
    #print(line)
    if regex.match(line):
        result = regex.search(line)
        line = result.groups()
        html.append(line)
# The column names to be used to create the DataFrame.
columns= [
    'start_date',
    'start_time',
    'end_date',
    'end_time',
    'start_frequency',
    'end_frequency',
    'flare_location',
    'flare_region',
    'flare_classification',
    'cme_date',
    'cme_time',
    'cme_angle',
    'cme_width',
    'cme_speed'
]
# The HTML is converted into a pandas DataFrame using the above DataFrame.
nasa = pd.DataFrame(html, columns = columns)
# Display the final DataFrame at the end of this step.
```

```
nasa
```

[110]:

```
     start_date start_time end_date end_time start_frequency end_frequency  \
0    1997/04/01      14:00    04/01    14:15            8000          4000
1    1997/04/07      14:30    04/07    17:30           11000          1000
2    1997/05/12      05:15    05/14    16:00           12000            80
3    1997/05/21      20:20    05/21    22:00            5000           500
4    1997/09/23      21:53    09/23    22:16            6000          2000
..          …          …       …       …               …             …
513  2017/09/04      20:27    09/05    04:54           14000           210
514  2017/09/06      12:05    09/07    08:00           16000            70
515  2017/09/10      16:02    09/11    06:50           16000           150
516  2017/09/12      07:38    09/12    07:43           16000         13000
517  2017/09/17      11:45    09/17    12:35           16000           900

    flare_location flare_region flare_classification cme_date cme_time  \
0           S25E16         8026                 M1.3    04/01    15:18
1           S28E19         8027                 C6.8    04/07    14:27
2           N21W08         8038                 C1.3    05/12    05:30
3           N05W12         8040                 M1.3    05/21    21:00
4           S29E25         8088                 C1.4    09/23    22:02
..             …            …                  …       …        …
513         S10W12        12673                 M5.5    09/04    20:12
514         S08W33        12673                 X9.3    09/06    12:24
515         S09W92        -----                 X8.3    09/10    16:00
516         N08E48        12680                 C3.0    09/12    08:03
517        S08E170        -----                 ----    09/17    12:00

    cme_angle cme_width cme_speed
0          74        79       312
1        Halo       360       878
2        Halo       360       464
3         263       165       296
4         133       155       712
..        …         …         …
513      Halo       360      1418
514      Halo       360      1571
515      Halo       360      3163
516       124        96       252
517      Halo       360      1385

[518 rows x 14 columns]
```

Step 4: Tidy the NASA table

[111]:

```
# 1. All missing entries are replaced with NaN.
```

```
nasa = nasa.replace(regex=r'--:--|--/--|-----?-?|\?\?\?\?
 ↪|BACK|altr|DSF|FILA|DIM|EP\??|^h$', value=np.nan)
# Clean up the classification entries missing a 0 after the decimal place
nasa = nasa.replace(regex={r'(^X[0-9]{2})\.$' : r'\1.0'})
# 2. The entries with Halo flares are replaced with NaN.
nasa['is_halo'] = nasa['cme_angle'].apply(lambda x: True if x == 'Halo' else␣
 ↪False)
# The same entries as above are specified as Halo flares in a new is_halo␣
 ↪column.
nasa['cme_angle'] = nasa['cme_angle'].apply(lambda x: np.nan if x == 'Halo'␣
 ↪else x)
# 3. The entries with lower bounds in the width column get the ">" removed.
nasa['width_lower_bound'] = nasa['cme_width'].apply(lambda x: True if '&gt;' in␣
 ↪str(x) else False)
# The same entries as above are specified as lower bounds in a new␣
 ↪width_lower_bound column.
nasa['cme_width'] = nasa['cme_width'].str.extract('(\d+)', expand=False)
# 4. Years are added to the 'end_date' and 'cme_date' columns.
for date in ['end', 'cme']:
    nasa[date + '_date'] = nasa['start_date'].str.extract('(\d{4})',␣
 ↪expand=False) + "/" + nasa[date + '_date']
# Each of the time columns in the DataFrame are combined with the date column␣
 ↪to create new datetime columns.
for column in ['end', 'start', 'cme']:
    # Columns that specify 24:00 for time are changes to 00:00.
    nasa[column + '_time'] = nasa[column + '_time'].apply(lambda x: '00:00' if␣
 ↪'24:00' in str(x) else x)
    nasa.insert(
        loc = 0 if not column == 'cme' else 7,
        column = column + '_datetime',
        value = pd.to_datetime(nasa[column + '_date'] + nasa[column + '_time'],␣
 ↪format='%Y/%m/%d%H:%M')
    )
    # The time columns are dropped from the DataFrame.
    nasa.drop(column + '_time', axis=1, inplace=True)
    # The date columns are dropped from the DataFrame.
    nasa.drop(column + '_date', axis=1, inplace=True)
# Display the final DataFrame at the end of this step.
nasa
```

```
[111]:        start_datetime        end_datetime start_frequency end_frequency  \
       0   1997-04-01 14:00:00 1997-04-01 14:15:00            8000          4000
       1   1997-04-07 14:30:00 1997-04-07 17:30:00           11000          1000
       2   1997-05-12 05:15:00 1997-05-14 16:00:00           12000            80
       3   1997-05-21 20:20:00 1997-05-21 22:00:00            5000           500
       4   1997-09-23 21:53:00 1997-09-23 22:16:00            6000          2000
```

```
..                  …                 …             …           …
513 2017-09-04 20:27:00 2017-09-05 04:54:00           14000          210
514 2017-09-06 12:05:00 2017-09-07 08:00:00           16000           70
515 2017-09-10 16:02:00 2017-09-11 06:50:00           16000          150
516 2017-09-12 07:38:00 2017-09-12 07:43:00           16000        13000
517 2017-09-17 11:45:00 2017-09-17 12:35:00           16000          900

    flare_location flare_region flare_classification        cme_datetime  \
0           S25E16         8026                 M1.3 1997-04-01 15:18:00
1           S28E19         8027                 C6.8 1997-04-07 14:27:00
2           N21W08         8038                 C1.3 1997-05-12 05:30:00
3           N05W12         8040                 M1.3 1997-05-21 21:00:00
4           S29E25         8088                 C1.4 1997-09-23 22:02:00
..             …          …                  …                 …
513         S10W12        12673                 M5.5 2017-09-04 20:12:00
514         S08W33        12673                 X9.3 2017-09-06 12:24:00
515         S09W92          NaN                 X8.3 2017-09-10 16:00:00
516         N08E48        12680                 C3.0 2017-09-12 08:03:00
517        S08E170          NaN                  NaN 2017-09-17 12:00:00

    cme_angle cme_width cme_speed  is_halo  width_lower_bound
0          74        79       312    False              False
1         NaN       360       878     True              False
2         NaN       360       464     True              False
3         263       165       296    False              False
4         133       155       712    False              False
..        …       …       …      …                …
513       NaN       360      1418     True              False
514       NaN       360      1571     True              False
515       NaN       360      3163     True              False
516       124        96       252    False              False
517       NaN       360      1385     True              False

[518 rows x 13 columns]
```

Part 2: Analysis

Question 1: Replication

[112]:

```
# All classifications that aren't rated X are downgraded to a 0.0 for sake of comparing the X classifactions flares.
x_class = nasa.replace(regex={r'^X([0-9]+\.[0-9])$' : r'\1', r'^[A-W][0-9]+\.[0-9]$' : 0.0, 'FILA' : 0.0})
# All entries are set to floats so they can be compares.
x_class = x_class.astype({'flare_classification': float}, errors='raise')
# The values are sorted based on their classification.
x_class = x_class.sort_values(by = ['flare_classification'], ascending = False)
# The first 50 results are returned.
```

```
x_class.head(50)

# The data here is very similar to that of SWL, but is lacking in some ways.␣
 ↪Just taking a look at the flare_classification,
# a few of the entries are slightly off or missing. The other good way to␣
 ↪double check which flares correlate with each other
# is by looking at the region. If you look at both the classification and␣
 ↪region, most match up with the exception of a few.
# For example, there are two flares rated at 17.0+ on the swl page, but only␣
 ↪one on NASA. This being said, NASA seems to have
# that same flare recorded by confirming dates and regions, but there is a lot␣
 ↪of data missing hence a note from NASA that
# reads "LASCO_DATA_GAP" for a lack of observations. This lack in some data␣
 ↪entries seems to make up for most discrepencies.
```

```
[112]:          start_datetime        end_datetime start_frequency end_frequency  \
       240 2003-11-04 20:00:00 2003-11-04 00:00:00           10000           200
       117 2001-04-02 22:05:00 2001-04-03 02:30:00           14000           250
       233 2003-10-28 11:10:00 2003-10-29 00:00:00           14000            40
       126 2001-04-15 14:05:00 2001-04-16 13:00:00           14000            40
       234 2003-10-29 20:55:00 2003-10-29 00:00:00           11000           500
       8   1997-11-06 12:20:00 1997-11-07 08:30:00           14000           100
       514 2017-09-06 12:05:00 2017-09-07 08:00:00           16000            70
       328 2006-12-05 10:50:00 2006-12-05 20:00:00           14000           250
       515 2017-09-10 16:02:00 2017-09-11 06:50:00           16000           150
       237 2003-11-02 17:30:00 2003-11-03 01:00:00           12000           250
       288 2005-01-20 07:15:00 2005-01-20 16:30:00           14000            25
       359 2011-08-09 08:20:00 2011-08-09 08:35:00           16000          4000
       331 2006-12-06 19:00:00 2006-12-08 00:00:00           16000            30
       317 2005-09-09 19:45:00 2005-09-09 22:00:00           10000            50
       82  2000-07-14 10:30:00 2000-07-15 14:30:00           14000            80
       121 2001-04-06 19:35:00 2001-04-07 01:50:00           14000           230
       375 2012-03-07 01:00:00 2012-03-08 19:00:00           16000            30
       135 2001-08-25 16:50:00 2001-08-25 23:00:00            8000           170
       443 2014-02-25 00:56:00 2014-02-25 11:28:00           14000           100
       193 2002-07-23 00:50:00 2002-07-23 04:00:00           11000           400
       104 2000-11-26 17:00:00 2000-11-26 17:15:00           14000          7000
       239 2003-11-03 10:00:00 2003-11-03 12:30:00            6000           400
       286 2005-01-17 10:00:00 2005-01-17 10:35:00            6100          1500
       222 2003-05-28 01:00:00 2003-05-29 00:30:00            1000           200
       160 2001-12-28 20:35:00 2001-12-29 03:00:00           14000           350
       332 2006-12-13 02:45:00 2006-12-13 10:40:00           12000           150
       192 2002-07-20 21:30:00 2002-07-20 22:20:00           10000          2000
       404 2013-05-14 01:16:00 2013-05-14 08:20:00           16000           240
       201 2002-08-24 01:45:00 2002-08-24 03:25:00            5000           400
       403 2013-05-13 16:15:00 2013-05-13 19:10:00           16000           300
       238 2003-11-03 01:15:00 2003-11-03 01:25:00            3000          1500
```

```
487 2015-05-05 22:24:00 2015-05-05 23:14:00           14000           500
19  1998-05-06 08:25:00 1998-05-06 08:35:00           14000          5000
142 2001-09-24 10:45:00 2001-09-25 20:00:00            7000            30
9   1997-11-27 13:30:00 1997-11-27 14:00:00           14000          7000
284 2005-01-15 23:00:00 2005-01-17 00:00:00            3000            40
276 2004-11-10 02:25:00 2004-11-10 03:40:00           14000          1000
73  2000-06-06 15:20:00 2000-06-08 09:00:00           14000            40
123 2001-04-10 05:24:00 2001-04-10 00:00:00           14000           100
99  2000-11-24 15:25:00 2000-11-24 22:00:00           14000           200
345 2011-02-15 02:10:00 2011-02-15 07:00:00           16000           400
318 2005-09-10 21:45:00 2005-09-11 01:00:00           14000           200
420 2013-10-25 15:08:00 2013-10-25 22:32:00           16000           200
7   1997-11-04 06:00:00 1997-11-05 04:30:00           14000           100
361 2011-09-06 22:30:00 2011-09-07 15:40:00           16000           150
125 2001-04-12 10:20:00 2001-04-12 10:40:00           14000          7000
98  2000-11-24 05:10:00 2000-11-24 15:00:00           14000           100
274 2004-11-07 16:25:00 2004-11-08 20:00:00           14000            60
285 2005-01-17 09:25:00 2005-01-17 16:00:00           14000            30
102 2000-11-25 19:00:00 2000-11-25 19:35:00            6000          2000

    flare_location flare_region  flare_classification         cme_datetime  \
240         S19W83        10486                  28.0 2003-11-04 19:54:00
117         N19W72         9393                  20.0 2001-04-02 22:06:00
233         S16E08        10486                  17.0 2003-10-28 11:30:00
126         S20W85         9415                  14.0 2001-04-15 14:06:00
234         S15W02        10486                  10.0 2003-10-29 20:54:00
8           S18W63         8100                   9.4 1997-11-06 12:10:00
514         S08W33        12673                   9.3 2017-09-06 12:24:00
328         S07E68        10930                   9.0                 NaT
515         S09W92          NaN                   8.3 2017-09-10 16:00:00
237         S14W56        10486                   8.3 2003-11-02 17:30:00
288         N14W61        10720                   7.1 2005-01-20 06:54:00
359         N17W69        11263                   6.9 2011-08-09 08:12:00
331         S05E64        10930                   6.5                 NaT
317         S12E67        10808                   6.2 2005-09-09 19:48:00
82          N22W07         9077                   5.7 2000-07-14 10:54:00
121         S21E31         9415                   5.6 2001-04-06 19:30:00
375         N17E27        11429                   5.4 2012-03-07 00:24:00
135         S17E34         9591                   5.3 2001-08-25 16:50:00
443         S12E82        11990                   4.9 2014-02-25 01:25:00
193         S13E72        10039                   4.8 2002-07-23 00:42:00
104         N18W38         9236                   4.0 2000-11-26 17:06:00
239         N08W77        10488                   3.9 2003-11-03 10:06:00
286         N15W25        10720                   3.8 2005-01-17 09:54:00
222         S07W20        10365                   3.6 2003-05-28 00:50:00
160         S26E90         9756                   3.4 2001-12-28 20:30:00
332         S06W23        10930                   3.4 2006-12-13 02:54:00
```

```
192         S13E90        10039                   3.3 2002-07-20 22:06:00
404         N08E77        11748                   3.2 2013-05-14 01:25:00
201         S02W81        10069                   3.1 2002-08-24 01:27:00
403         N11E85        11748                   2.8 2013-05-13 16:07:00
238         N10W83        10488                   2.7 2003-11-03 01:59:00
487         N15E79        12339                   2.7 2015-05-05 22:24:00
19          S11W65         8210                   2.7 1998-05-06 08:29:00
142         S16E23         9632                   2.6 2001-09-24 10:30:00
9           N17E63         8113                   2.6 1997-11-27 13:56:00
284         N15W05        10720                   2.6 2005-01-15 23:06:00
276         N09W49        10696                   2.5 2004-11-10 02:26:00
73          N20E18         9026                   2.3 2000-06-06 15:54:00
123         S23W09         9415                   2.3 2001-04-10 05:30:00
99          N22W07         9236                   2.3 2000-11-24 15:30:00
345         S20W12        11158                   2.2 2011-02-15 02:24:00
318         S13E47        10808                   2.1 2005-09-10 21:52:00
420         S06E69        11882                   2.1 2013-10-25 15:12:00
7           S14W33         8100                   2.1 1997-11-04 06:10:00
361         N14W18        11283                   2.1 2011-09-06 23:05:00
125         S19W43         9415                   2.0 2001-04-12 10:31:00
98          N20W05         9236                   2.0 2000-11-24 05:30:00
274         N09W17        10696                   2.0 2004-11-07 16:54:00
285         N15W25        10720                   2.0 2005-01-17 09:30:00
102         N20W23         9236                   1.9 2000-11-25 19:31:00

    cme_angle cme_width cme_speed  is_halo  width_lower_bound
240       NaN       360      2657     True              False
117       261       244      2505    False              False
233       NaN       360      2459     True              False
126       245       167      1199    False              False
234       NaN       360      2029     True              False
8         NaN       360      1556     True              False
514       NaN       360      1571     True              False
328       NaN       NaN       NaN    False              False
515       NaN       360      3163     True              False
237       NaN       360      2598     True              False
288       NaN       360       882     True              False
359       NaN       360      1610     True              False
331       NaN       NaN       NaN    False              False
317       NaN       360      2257     True              False
82        NaN       360      1674     True              False
121       NaN       360      1270     True              False
375       NaN       360      2684     True              False
135       NaN       360      1433     True              False
443       NaN       360      2147     True              False
193       NaN       360      2285     True              False
104       NaN       360       980     True              False
```

```
239       293       103     1420    False              False
286       NaN       360     2547     True              False
222       NaN       360     1366     True              False
160       NaN       360     2216     True              False
332       NaN       360     1774     True              False
192       NaN       360     1941     True              False
404       NaN       360     2625     True              False
201       NaN       360     1913     True              False
403       NaN       360     1850     True              False
238       304        65      827    False              False
487       NaN       360      715     True              False
19        309       190     1099    False              False
142       NaN       360     2402     True              False
9          98        91      441    False              False
284       NaN       360     2861     True              False
276       NaN       360     3387     True              False
73        NaN       360     1119     True              False
123       NaN       360     2411     True              False
99        NaN       360     1245     True              False
345       NaN       360      669     True              False
318       NaN       360     1893     True              False
420       NaN       360     1081     True              False
7         NaN       360      785     True              False
361       NaN       360      575     True              False
125       NaN       360     1184     True              False
98        NaN       360     1289     True              False
274       NaN       360     1759     True              False
285       NaN       360     2094     True              False
102       NaN       360      671     True              False
```

Question 2: Integration

[113]:
```
arr = []
# Add the strongest connections to the list. These occured on the same day, are␣
 ↪in the same region, and have an exact classification.
for i, nasa_row in nasa.iterrows():
    value = np.nan
    for j, swl_row in swl.iterrows():
        # This is where we check the strength of a
        if nasa_row['start_datetime']\
            .date() == swl_row['start_datetime'].date() and \
            str(nasa_row['flare_region'])[-3:] == str(swl_row['region'])[-3:]␣
 ↪and \
            str(nasa_row['flare_classification']) ==␣
 ↪(swl_row['x_classification']):
                value = swl_row['rank']
    arr.append(value)
```

```
# Add the next strongest connections to the list. These occured on the same day␣
 ↪and either were in the same region or have a very close classification score.
for i, nasa_row in nasa.iterrows():
    # This if statement prevents the stronger connections from being␣
 ↪overwritten.
    if np.isnan(arr[i]):
        value = np.nan
        for j, swl_row in swl.iterrows():
            if (str(nasa_row['flare_classification'])[:2] ==␣
 ↪str(swl_row['x_classification'])[:2] or \
                str(nasa_row['flare_region'])[-3:] == str(swl_row['region'])[-3:
 ↪]) \
                and nasa_row['start_datetime'].date() ==␣
 ↪swl_row['start_datetime'].date():
                    # This if statement prevents multiple nasa points from␣
 ↪mapping to a single swl entity.
                    if swl_row['rank'] in arr:
                        arr[arr.index(swl_row['rank'])] = np.nan
                        value = swl_row['rank']
                    else:
                        value = swl_row['rank']
            arr[i] = value
# Add the next strongest connections to the list. These supposedly occured on␣
 ↪the days surrounding what was listed on WSL.
for i, nasa_row in nasa.iterrows():
    # This if statement prevents the stronger connections from being␣
 ↪overwritten.
    if np.isnan(arr[i]):
        value = np.nan
        for j, swl_row in swl.iterrows():
            # Checking the for any data within a 2 week margin of error for␣
 ↪recording the flare.
            duration = nasa_row['start_datetime'].date() -␣
 ↪swl_row['start_datetime'].date()
            if duration.days < 13 and duration.days > -13:
                # This if statement prevents multiple nasa points from mapping␣
 ↪to a single swl entity.
                if not swl_row['rank'] in arr:
                    value = swl_row['rank']
            arr[i] = value
# Adding the mappings to the DataFrame and sorting it.
nasa["swt_map"] = arr
nasa = nasa.sort_values(by = ['swt_map'], ascending = True)
nasa.head(50)
# Add the last strongest connections to the list. These entries were much␣
 ↪further apart with much less in similar between the 1998 entries.
```

```
for i, nasa_row in nasa.iterrows():
    # This if statement prevents the stronger connections from being␣
 ↪overwritten.
    if np.isnan(arr[i]):
        value = np.nan
        for j, swl_row in swl.iterrows():
            # Checking the for any data within a little over 2 month margin of␣
 ↪error for recording the flare.
            duration = nasa_row['start_datetime'].date() -␣
 ↪swl_row['start_datetime'].date()
            if duration.days < 70 and duration.days > -70:
                # This if statement prevents multiple nasa points from mapping␣
 ↪to a single swl entity.
                if not swl_row['rank'] in arr:
                    value = swl_row['rank']
            arr[i] = value
# Adding the mappings to the DataFrame and sorting it.
nasa["swt_map"] = arr
nasa = nasa.sort_values(by = ['swt_map'], ascending = True)
nasa.head(50)

# To decide how to organize the matches between the two data sets, I wanted to␣
 ↪get what seemed to be most acurate first,
# so I started with the data with as much in common as possible, which meant␣
 ↪looking for entries with multiple of the same
# pieces of information, namely the day it occured, the rating, and the region.␣
 ↪After grabbing these matches, I broadened
# the criteria to get partial matches and things that were very simlar like␣
 ↪classifications that were a decimal point off.
# From there, the data had very little in common so I switched to finding␣
 ↪events that occured in the same general time frame,
# starting with events recorded 13 days apart and eventually to those in 1998␣
 ↪whose entries were a few months apart.
```

[113]:

```
         start_datetime        end_datetime start_frequency end_frequency  \
211 2002-11-11 16:15:00 2002-11-11 17:50:00           14000           600
74  2000-06-10 17:15:00 2000-06-10 18:45:00           10000          1000
204 2002-09-27 13:35:00 2002-09-27 14:30:00           14000          8000
299 2005-07-13 14:15:00 2005-07-13 15:05:00           14000          1000
84  2000-08-11 11:35:00 2000-08-11 11:59:00            2800          2000
205 2002-10-13 18:10:00 2002-10-13 18:40:00           14000          4000
328 2006-12-05 10:50:00 2006-12-05 20:00:00           14000           250
512 2017-07-23 05:27:00 2017-07-23 06:12:00            4400           900
313 2005-08-31 22:10:00 2005-08-31 23:00:00           14000          6000
208 2002-10-27 23:06:00 2002-10-28 01:20:00           14000           300
513 2017-09-04 20:27:00 2017-09-05 04:54:00           14000           210
```

```
268 2004-10-24 03:12:00 2004-10-24 03:21:00           14000          8000
349 2011-05-09 21:00:00 2011-05-10 04:00:00           16000           900
318 2005-09-10 21:45:00 2005-09-11 01:00:00           14000           200
300 2005-07-14 11:00:00 2005-07-14 12:54:00            3000           800
118 2001-04-03 03:40:00 2001-04-03 07:25:00           14000           400
39  1999-06-11 11:45:00 1999-06-11 17:00:00           14000           400
78  2000-06-25 08:10:00 2000-06-25 09:00:00           12000          2500
366 2011-11-09 13:30:00 2011-11-09 17:00:00           16000           400
295 2005-06-03 12:50:00 2005-06-03 15:00:00           10000           270
200 2002-08-22 02:50:00 2002-08-22 04:13:00           14000          3500
93  2000-11-08 23:20:00 2000-11-09 12:00:00            4000           200
438 2014-01-06 07:57:00 2014-01-06 22:30:00           14000            80
286 2005-01-17 10:00:00 2005-01-17 10:35:00            6100          1500
159 2001-12-26 05:20:00 2001-12-27 05:00:00           14000           150
61  2000-03-02 13:50:00 2000-03-02 14:03:00           14000          6000
210 2002-11-10 03:20:00 2002-11-10 06:00:00            3000           300
239 2003-11-03 10:00:00 2003-11-03 12:30:00            6000           400
266 2004-09-12 00:45:00 2004-09-13 21:00:00           14000            40
222 2003-05-28 01:00:00 2003-05-29 00:30:00            1000           200
294 2005-05-17 03:20:00 2005-05-17 03:35:00            4500          1500
236 2003-11-02 09:23:00 2003-11-02 11:22:00           14000           550
190 2002-07-18 20:55:00 2002-07-18 21:40:00            6000          3000
319 2005-09-11 13:10:00 2005-09-11 15:15:00            3000           350
120 2001-04-05 09:14:00 2001-04-05 09:34:00           14000          7500
414 2013-09-29 21:53:00 2013-09-30 21:00:00           14000            60
156 2001-11-22 22:40:00 2001-11-24 02:30:00           14000            40
261 2004-07-23 19:00:00 2004-07-23 19:35:00           10000          2500
396 2012-10-22 01:50:00 2012-10-22 11:15:00            1000           200
470 2014-09-23 23:41:00 2014-09-23 23:47:00           14000         12000
168 2002-03-12 00:00:00 2002-03-12 02:20:00           14000          2200
149 2001-10-19 01:15:00 2001-10-19 02:25:00           14000          1300
395 2012-09-27 23:55:00 2012-09-28 10:15:00           16000           250
116 2001-04-02 11:30:00 2001-04-02 12:00:00           14000          5000
104 2000-11-26 17:00:00 2000-11-26 17:15:00           14000          7000
484 2015-03-11 10:30:00 2015-03-11 14:50:00            1000           250
209 2002-11-09 13:20:00 2002-11-10 03:00:00           14000           100
312 2005-08-31 11:40:00 2005-08-31 12:10:00            6000           800
264 2004-07-31 07:10:00 2004-07-31 11:30:00            1000           200
100 2000-11-24 22:24:00 2000-11-24 22:36:00            4000          3000

    flare_location flare_region flare_classification        cme_datetime  \
211         S13W60        10180                 M1.8 2002-11-11 15:54:00
74          N22W38         9026                 M5.2 2000-06-10 17:08:00
204         N13E45        10134                 M1.8 2002-09-27 13:56:00
299         N11W90        10786                 M5.0 2005-07-13 14:30:00
84           NW90b          NaN                  NaN 2000-08-11 07:31:00
205         S07W54        10150                 C4.7 2002-10-13 19:35:00
```

```
328         S07E68        10930                X9.0                 NaT
512            NaN          NaN                 NaN 2017-07-23 04:48:00
313            NaN          NaN                 NaN 2005-08-31 22:30:00
208          SE90b          NaN                 NaN 2002-10-27 23:18:00
513         S10W12        12673                M5.5 2017-09-04 20:12:00
268         N11E19        10687                C1.7 2004-10-24 03:54:00
349         N16E88          NaN                C5.4 2011-05-09 20:57:00
318         S13E47        10808                X2.1 2005-09-10 21:52:00
300         N11W90        10786                X1.2 2005-07-14 10:54:00
118         S21E83         9415                X1.2 2001-04-03 03:26:00
39          N38E90          NaN                C8.8 1999-06-11 11:26:00
78          N16W55         9046                M1.9 2000-06-25 07:54:00
366         N24E35        11343                M1.1 2011-11-09 13:36:00
295         N15E90        10775                M1.0 2005-06-03 12:32:00
200         S07W62          NaN                M5.4 2002-08-22 02:06:00
93          N10W77         9213                M7.4 2000-11-08 23:06:00
438        S15W112        11936                 NaN 2014-01-06 08:00:00
286         N15W25        10720                X3.8 2005-01-17 09:54:00
159         N08W54         9742                M7.1 2001-12-26 05:30:00
61          S20W58         8882                M6.5 2000-03-02 13:54:00
210         S12W37        10180                M2.4 2002-11-10 03:30:00
239         N08W77        10488                X3.9 2003-11-03 10:06:00
266         N03E49        10672                M4.8 2004-09-12 00:36:00
222         S07W20        10365                X3.6 2003-05-28 00:50:00
294            NaN          NaN                 NaN 2005-05-17 03:06:00
236          SW90b          NaN                 NaN 2003-11-02 09:30:00
190           E90b        10036                C3.3 2002-07-18 19:31:00
319         S16E39        10808                M3.0 2005-09-11 13:00:00
120         N14W85         9393                M8.4 2001-04-05 09:06:00
414         N17W29          NaN                C1.3 2013-09-29 22:12:00
156         S17W36         9704                M9.9 2001-11-22 23:30:00
261         N04W05        10652                C4.1 2004-07-23 19:31:00
396         S10E76        11598                M1.3 2012-10-21 20:57:00
470         S13E33        12172                M2.3 2014-09-23 23:36:00
168         S15E45          NaN                M5.0 2002-03-11 23:30:00
149         N16W18         9661                X1.6 2001-10-19 01:27:00
395         N06W34        11577                C3.7 2012-09-28 00:12:00
116         N20W70         9393                X1.1 2001-04-02 11:26:00
104         N18W38         9236                X4.0 2000-11-26 17:06:00
484         S15E23        12297                M2.6 2015-03-11 08:24:00
209         S12W29        10180                M4.6 2002-11-09 13:31:00
312         N13W13        10803                C2.0 2005-08-31 11:30:00
264         N05W89        10652                C8.4 2004-07-31 05:54:00
100         N21W14         9236                X1.8 2000-11-24 22:06:00

    cme_angle cme_width cme_speed  is_halo  width_lower_bound  swt_map
211       212        93      1083    False              False      1.0
```

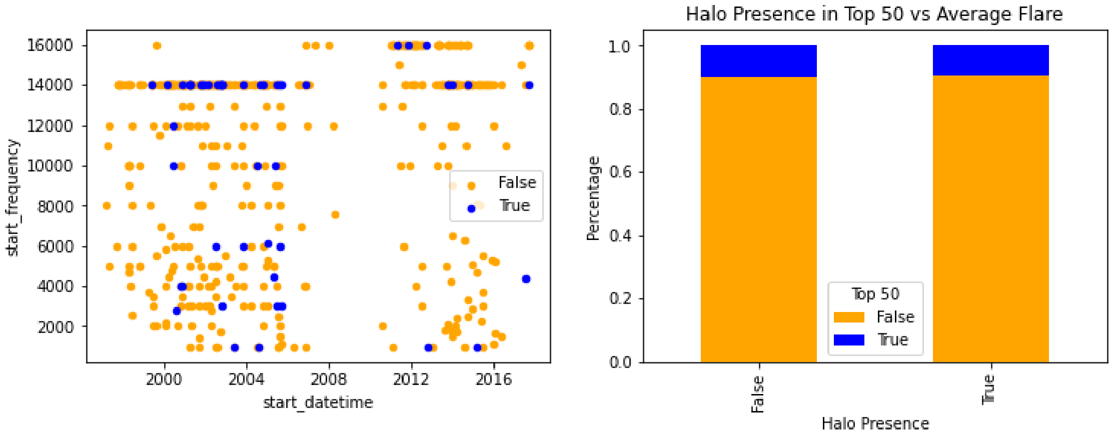
```
74        NaN       360      1108     True              False      2.0
204        64        64       591    False              False      3.0
299       NaN       360      1423     True              False      4.0
84        273        70      1071    False              False      5.0
205       252       141       373    False              False      6.0
328       NaN       NaN       NaN    False              False      7.0
512       NaN       360      1848     True              False      8.0
313       NaN       360      1808     True              False      9.0
208       NaN       360      2115     True              False     10.0
513       NaN       360      1418     True              False     11.0
268        26       109       417    False               True     12.0
349        55       292      1318    False              False     13.0
318       NaN       360      1893     True              False     14.0
300       NaN       360      2115     True              False     15.0
118       108       292      1613    False              False     16.0
39         35       181      1569    False               True     17.0
78        262       165      1617    False              False     18.0
366       NaN       360       907     True              False     19.0
295       NaN       360      1679     True              False     20.0
200       NaN       360       998     True              False     21.0
93        271       170      1738    False               True     22.0
438       NaN       360      1402     True              False     23.0
286       NaN       360      2547     True              False     24.0
159       281       212      1446    False               True     25.0
61        235        76       835    False              False     26.0
210       203       282      1670    False              False     27.0
239       293       103      1420    False              False     28.0
266       NaN       360      1328     True              False     29.0
222       NaN       360      1366     True              False     30.0
294       252        89       311    False              False     31.0
236       NaN       360      2036     True              False     32.0
190       NaN       360      2191     True              False     33.0
319       NaN       360      1922     True              False     34.0
120       283       205      1750    False              False     35.0
414       NaN       360      1179     True              False     36.0
156       NaN       360      1437     True              False     37.0
261       209       100       874    False              False     38.0
396        83       243       496    False              False     39.0
470       109       134       331    False              False     40.0
168       NaN       360       950     True              False     41.0
149       NaN       360       558     True              False     42.0
395       NaN       360       947     True              False     43.0
116       270        80       992    False              False     44.0
104       NaN       360       980     True              False     45.0
484        62        93       530    False              False     46.0
209       NaN       360      1838     True              False     47.0
312       NaN       360       825     True              False     48.0
```

```
264        259        197       1192     False                   True      49.0
100        NaN        360       1005      True                  False      50.0
```

Question 3: Analysis

[116]:

```
nasa = nasa.astype({'start_frequency': float}, errors='raise')
nasa["top_50"] = nasa["swt_map"]\
    .apply(lambda x: False if np.isnan(x) else True)

fig, ax = plt.subplots(1, 2)

# Graph for the frequency of halos in the top 50 solar flares as a percentage␣
 ↪of the whole.
#
# The intent of this plot is to depict how likely the classsification/size/
 ↪strength of a flare is to
# create a halo. This can be used to predict whether or not a flare will have a␣
 ↪halo based on it's size.
#
# My interpretation of this plot is that the size of a flare has little to no␣
 ↪effect on the presence of a
# halo. As seen on the chart, the percent of flares in the top 50 is almost␣
 ↪exactly the same given that
# whether a halo occured or not.
nasa.groupby('is_halo')['top_50']\
    .value_counts(normalize=True)\
    .unstack(level=1)\
    .plot(ax=ax[1], kind='bar', stacked=True, color=['orange', 'blue'])\
    .legend(title='Top 50')

# Graph for frequency of flare at start by the start datetime.
#
# The intent of this plot is to depict how much of an effect the start␣
 ↪frequency of a flare has on its
# classification as a top flare. This can be used to predict how a flare will␣
 ↪be classified based on its
# starting frequency.
#
# My interpretation of this plot is that the start frequency has little to no␣
 ↪effect on the classification
# of a flare. As seen on the chart, the percent of flares in the top 50 is␣
 ↪almost just as sparatic in
# frequency as those not in the top 50.
colors = {False:'orange', True:'blue'}

grouped = nasa.groupby('top_50')
for key, group in grouped:
    group.plot(ax=ax[0], kind='scatter', x='start_datetime',␣
↪y='start_frequency', label=key, color=colors[key])

plt.xlabel("Halo Presence")
plt.ylabel("Percentage")
plt.title("Halo Presence in Top 50 vs Average Flare")

fig.set_figwidth(12)
plt.show()
```


start_frequency
start_datetime
False
True
Halo Presence in Top 50 vs Average Flare
Percentage
Top 50
False
True
False
True
Halo Presence