Proj1

Part 1: Data scraping and preparation

Step 1: Scrape your competitor’s data

This pipeline first reads the html from the provided url and finds the table based on the HTML node. Column names are added by using a vectpr of column names and the pipeline results in a dataframe

```
url <- "https://www.spaceweatherlive.com/en/solar-activity/top-50-solar-flares"

sf <- url %>%
  read_html() %>%
  html_node(".table-striped") %>%
  html_table() %>%
  set_colnames(c("rank", "flare_classification","date","flare_region", "start_time","maximum_time"
,"end_time","movie")) %>%
  as.data.frame()
head(sf, n = 10)
```

```
##    rank flare_classification       date flare_region start_time
## 1     1                X28.0 2003/11/04          486      19:29
## 2     2                X20.0 2001/04/02         9393      21:32
## 3     3                X17.2 2003/10/28          486      09:51
## 4     4                X17.0 2005/09/07          808      17:17
## 5     5                X14.4 2001/04/15         9415      13:19
## 6     6                X10.0 2003/10/29          486      20:37
## 7     7                 X9.4 1997/11/06         8100      11:49
## 8     8                 X9.3 2017/09/06         2673      11:53
## 9     9                 X9.0 2006/12/05          930      10:18
## 10   10                 X8.3 2003/11/02          486      17:03
##    maximum_time end_time             movie
## 1         19:53    20:06 MovieView archive
## 2         21:51    22:03 MovieView archive
## 3         11:10    11:24 MovieView archive
## 4         17:40    18:03 MovieView archive
## 5         13:50    13:55 MovieView archive
## 6         20:49    21:01 MovieView archive
## 7         11:55    12:01 MovieView archive
## 8         12:02    12:10      View archive
## 9         10:35    10:45 MovieView archive
## 10        17:25    17:39 MovieView archive
```

Step 2: Tidy the top 50 solar flare data

uses the data frame from Step 1 and first drops the movie column by using the select function that keeps everything but the movie col using the minus sign operator. Then the unite function is used to create a new column based on the column vectors passed in and finally the united column is converted to a POSIXct type

using type convert and using a column specification which uses a regex.

```
sf <- sf %>%
  select(-movie) %>%
  unite ("start_datetime", c("date", "start_time"), sep = " ", remove =    FALSE)%>%
  unite ("max_datetime",  c("date", "maximum_time"),sep = " ", remove = FALSE)%>%
  unite ("end_datetime", c("date", "end_time"), sep = " ", remove = TRUE)%>%
 type_convert(cols(start_datetime=col_datetime(format = "%Y/%m/%d %H:%M"),   max_datetime=col_date
time(format = "%Y/%m/%d %H:%M"), end_datetime = col_datetime(format = "%Y/%m/%d %H:%M")))
head(sf, n = 10)
```

```
##    rank flare_classification      start_datetime        max_datetime
## 1     1                X28.0 2003-11-04 19:29:00 2003-11-04 19:53:00
## 2     2                X20.0 2001-04-02 21:32:00 2001-04-02 21:51:00
## 3     3                X17.2 2003-10-28 09:51:00 2003-10-28 11:10:00
## 4     4                X17.0 2005-09-07 17:17:00 2005-09-07 17:40:00
## 5     5                X14.4 2001-04-15 13:19:00 2001-04-15 13:50:00
## 6     6                X10.0 2003-10-29 20:37:00 2003-10-29 20:49:00
## 7     7                 X9.4 1997-11-06 11:49:00 1997-11-06 11:55:00
## 8     8                 X9.3 2017-09-06 11:53:00 2017-09-06 12:02:00
## 9     9                 X9.0 2006-12-05 10:18:00 2006-12-05 10:35:00
## 10   10                 X8.3 2003-11-02 17:03:00 2003-11-02 17:25:00
##           end_datetime flare_region start_time maximum_time
## 1  2003-11-04 20:06:00          486   19:29:00     19:53:00
## 2  2001-04-02 22:03:00         9393   21:32:00     21:51:00
## 3  2003-10-28 11:24:00          486   09:51:00     11:10:00
## 4  2005-09-07 18:03:00          808   17:17:00     17:40:00
## 5  2001-04-15 13:55:00         9415   13:19:00     13:50:00
## 6  2003-10-29 21:01:00          486   20:37:00     20:49:00
## 7  1997-11-06 12:01:00         8100   11:49:00     11:55:00
## 8  2017-09-06 12:10:00         2673   11:53:00     12:02:00
## 9  2006-12-05 10:45:00          930   10:18:00     10:35:00
## 10 2003-11-02 17:39:00          486   17:03:00     17:25:00
```

Step 3: Scrape the NASA data

```
url <-  "http://cdaw.gsfc.nasa.gov/CME_list/radio/waves_type2.html"

NASA <- url %>%
  read_html() %>%
  html_node("pre")%>%
  html_text() %>%
  strsplit("\n") %>%
  as.data.frame() %>%
  set_colnames(c("col")) %>%
  separate(col, into = c("start_date", "start_time", "end_date", "end_time", "start_freq", "end_fr
eq", "Location", "NOAA", "Importance",
                       "CME_date", "CME_time", "CME_CPA", "CME_width", "CME_speed", "Plots"), sep =
 "[ ]+")
```

```
## Warning: Too many values at 43 locations: 11, 28, 39, 55, 90, 98, 104, 109,
## 135, 136, 162, 163, 169, 183, 195, 196, 220, 242, 263, 264, ...
```

```
## Warning: Too few values at 13 locations: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12,
## 524, 525
```

```
tail(NASA, 513) %>%
  head(10)
```

```
##    start_date start_time end_date end_time start_freq end_freq Location
## 13 1997/04/01      14:00    04/01    14:15       8000     4000   S25E16
## 14 1997/04/07      14:30    04/07    17:30      11000     1000   S28E19
## 15 1997/05/12      05:15    05/14    16:00      12000       80   N21W08
## 16 1997/05/21      20:20    05/21    22:00       5000      500   N05W12
## 17 1997/09/23      21:53    09/23    22:16       6000     2000   S29E25
## 18 1997/11/03      05:15    11/03    12:00      14000      250   S20W13
## 19 1997/11/03      10:30    11/03    11:30      14000     5000   S16W21
## 20 1997/11/04      06:00    11/05    04:30      14000      100   S14W33
## 21 1997/11/06      12:20    11/07    08:30      14000      100   S18W63
## 22 1997/11/27      13:30    11/27    14:00      14000     7000   N17E63
##    NOAA Importance CME_date CME_time CME_CPA CME_width CME_speed Plots
## 13 8026       M1.3    04/01    15:18      74        79       312  PHTX
## 14 8027       C6.8    04/07    14:27    Halo       360       878  PHTX
## 15 8038       C1.3    05/12    05:30    Halo       360       464  PHTX
## 16 8040       M1.3    05/21    21:00     263       165       296  PHTX
## 17 8088       C1.4    09/23    22:02     133       155       712  PHTX
## 18 8100       C8.6    11/03    05:28     240       109       227  PHTX
## 19 8100       M4.2    11/03    11:11     233       122       352  PHTX
## 20 8100       X2.1    11/04    06:10    Halo       360       785  PHTX
## 21 8100       X9.4    11/06    12:10    Halo       360      1556  PHTX
## 22 8113       X2.6    11/27    13:56      98        91       441  PHTX
```

Step 4: Tidy the NASA Table

```
NASA[NASA == "------"] <- NA
NASA[NASA == "-----"] <- NA
NASA[NASA == "----"] <- NA
NASA[NASA == "--/--"] <- NA
NASA[NASA == "--:--"] <- NA
NASA[NASA == "????"] <- NA
NASA[NASA == "BACK"] <- NA
NASA[NASA == "24:00"] <- "23:59"
NASA <- tail(NASA,513) %>% head(511) %>% mutate (Halo = (CME_CPA == "Halo"))
NASA$CME_CPA[NASA$CME_CPA == "Halo"] <- NA
NASA <- NASA %>% mutate (Lower_bound = substring(CME_width, 1, 1) == ">")
NASA$CME_width <- sub('>', '', NASA$CME_width)
NASA <- NASA %>% unite ("start", c("start_date", "start_time"), sep = " ", remove = TRUE)%>%
  mutate(year = substring(start,1,5)) %>%
  unite("CME_date", c("year","CME_date"), sep = "", remove = FALSE) %>%
  unite("end_date", c("year","end_date"), sep = "", remove = TRUE) %>%
  unite ("end",  c("end_date", "end_time"),sep = " ", remove = TRUE)%>%
  unite ("cme", c("CME_date", "CME_time"), sep = " ", remove = TRUE)
NASA$cme[substring(NASA$cme,6,7) == "NA"] <- "temp"

NASA <- NASA %>%  type_convert(cols(start=col_datetime(format = "%Y/%m/%d %H:%M"), end=col_datetim
e(format = "%Y/%m/%d %H:%M"), cme = col_datetime(format = "%Y/%m/%d %H:%M")), na = c("temp")) %>%
  set_colnames(c("start_datetime", "end_datetime","start_frequency","end_frequency", "flare_locati
on","flare_region","flare_classification","cme_datetime", "cme_angle", "cme_width", "cme_speed",
"plots", "halo", "cme_width_limit"))

head(NASA,10)
```

```
##         start_datetime        end_datetime start_frequency end_frequency
## 1  1997-04-01 14:00:00 1997-04-01 14:15:00            8000          4000
## 2  1997-04-07 14:30:00 1997-04-07 17:30:00           11000          1000
## 3  1997-05-12 05:15:00 1997-05-14 16:00:00           12000            80
## 4  1997-05-21 20:20:00 1997-05-21 22:00:00            5000           500
## 5  1997-09-23 21:53:00 1997-09-23 22:16:00            6000          2000
## 6  1997-11-03 05:15:00 1997-11-03 12:00:00           14000           250
## 7  1997-11-03 10:30:00 1997-11-03 11:30:00           14000          5000
## 8  1997-11-04 06:00:00 1997-11-05 04:30:00           14000           100
## 9  1997-11-06 12:20:00 1997-11-07 08:30:00           14000           100
## 10 1997-11-27 13:30:00 1997-11-27 14:00:00           14000          7000
##    flare_location flare_region flare_classification        cme_datetime
## 1          S25E16         8026                 M1.3 1997-04-01 15:18:00
## 2          S28E19         8027                 C6.8 1997-04-07 14:27:00
## 3          N21W08         8038                 C1.3 1997-05-12 05:30:00
## 4          N05W12         8040                 M1.3 1997-05-21 21:00:00
## 5          S29E25         8088                 C1.4 1997-09-23 22:02:00
## 6          S20W13         8100                 C8.6 1997-11-03 05:28:00
## 7          S16W21         8100                 M4.2 1997-11-03 11:11:00
## 8          S14W33         8100                 X2.1 1997-11-04 06:10:00
## 9          S18W63         8100                 X9.4 1997-11-06 12:10:00
## 10         N17E63         8113                 X2.6 1997-11-27 13:56:00
##    cme_angle cme_width cme_speed plots  halo cme_width_limit
## 1         74        79       312  PHTX FALSE           FALSE
## 2         NA       360       878  PHTX  TRUE           FALSE
## 3         NA       360       464  PHTX  TRUE           FALSE
## 4        263       165       296  PHTX FALSE           FALSE
## 5        133       155       712  PHTX FALSE           FALSE
## 6        240       109       227  PHTX FALSE           FALSE
## 7        233       122       352  PHTX FALSE           FALSE
## 8         NA       360       785  PHTX  TRUE           FALSE
## 9         NA       360      1556  PHTX  TRUE           FALSE
## 10        98        91       441  PHTX FALSE           FALSE
```

Part 2: Analysis

Question 1: Replication

From the created table based on NASA's data we can easily see that it is missing some values that are used in SpaceWeatherLive's data. This is because the data sets used by NASA and SpaceWeatherLive do not all have the same observations(some observations are missing from one table while they are present in another).

```
NASA_data <- NASA
NASA <- NASA %>% separate(flare_classification, into = c("Letter", "Number"), sep = 1, remove = FA
LSE) %>% filter(Letter == "X")
NASA$Number <- as.numeric(as.character(NASA$Number))
NASA_top50 <- NASA %>% arrange(desc(Number)) %>% select(-Letter,-Number) %>% head(50)
NASA_top50
```

```
##         start_datetime        end_datetime start_frequency end_frequency
## 1  2003-11-04 20:00:00 2003-11-04 23:59:00           10000           200
## 2  2001-04-02 22:05:00 2001-04-03 02:30:00           14000           250
## 3  2003-10-28 11:10:00 2003-10-29 23:59:00           14000            40
## 4  2001-04-15 14:05:00 2001-04-16 13:00:00           14000            40
## 5  2003-10-29 20:55:00 2003-10-29 23:59:00           11000           500
## 6  1997-11-06 12:20:00 1997-11-07 08:30:00           14000           100
## 7  2006-12-05 10:50:00 2006-12-05 20:00:00           14000           250
## 8  2003-11-02 17:30:00 2003-11-03 01:00:00           12000           250
## 9  2005-01-20 07:15:00 2005-01-20 16:30:00           14000            25
## 10 2011-08-09 08:20:00 2011-08-09 08:35:00           16000          4000
## 11 2006-12-06 19:00:00 2006-12-08 23:59:00           16000            30
## 12 2005-09-09 19:45:00 2005-09-09 22:00:00           10000            50
## 13 2000-07-14 10:30:00 2000-07-15 14:30:00           14000            80
## 14 2001-04-06 19:35:00 2001-04-07 01:50:00           14000           230
## 15 2012-03-07 01:00:00 2012-03-08 19:00:00           16000            30
## 16 2001-08-25 16:50:00 2001-08-25 23:00:00            8000           170
## 17 2014-02-25 00:56:00 2014-02-25 11:28:00           14000           100
## 18 2002-07-23 00:50:00 2002-07-23 04:00:00           11000           400
## 19 2000-11-26 17:00:00 2000-11-26 17:15:00           14000          7000
## 20 2003-11-03 10:00:00 2003-11-03 12:30:00            6000           400
## 21 2005-01-17 10:00:00 2005-01-17 10:35:00            6100          1500
## 22 2003-05-28 01:00:00 2003-05-29 00:30:00            1000           200
## 23 2001-12-28 20:35:00 2001-12-29 03:00:00           14000           350
## 24 2006-12-13 02:45:00 2006-12-13 10:40:00           12000           150
## 25 2002-07-20 21:30:00 2002-07-20 22:20:00           10000          2000
## 26 2013-05-14 01:16:00 2013-05-14 02:35:00           16000           700
## 27 2002-08-24 01:45:00 2002-08-24 03:25:00            5000           400
## 28 2013-05-13 16:15:00 2013-05-13 19:10:00           16000           300
## 29 1998-05-06 08:25:00 1998-05-06 08:35:00           14000          5000
## 30 2003-11-03 01:15:00 2003-11-03 01:25:00            3000          1500
## 31 2015-05-05 22:24:00 2015-05-05 23:14:00           14000           500
## 32 1997-11-27 13:30:00 1997-11-27 14:00:00           14000          7000
## 33 2001-09-24 10:45:00 2001-09-25 20:00:00            7000            30
## 34 2005-01-15 23:00:00 2005-01-15 00:00:00            3000            40
## 35 2004-11-10 02:25:00 2004-11-10 03:40:00           14000          1000
## 36 2000-06-06 15:20:00 2000-06-08 09:00:00           14000            40
## 37 2000-11-24 15:25:00 2000-11-24 22:00:00           14000           200
## 38 2001-04-10 05:24:00 2001-04-10 23:59:00           14000           100
## 39 2011-02-15 02:10:00 2011-02-15 07:00:00           16000           400
## 40 1997-11-04 06:00:00 1997-11-05 04:30:00           14000           100
## 41 2005-09-10 21:45:00 2005-09-10 01:00:00           14000           300
## 42 2011-09-06 22:30:00 2011-09-07 15:40:00           16000           150
## 43 2013-10-25 15:08:00 2013-10-25 22:32:00           16000           200
## 44 2000-11-24 05:10:00 2000-11-24 15:00:00           14000           100
## 45 2001-04-12 10:20:00 2001-04-12 10:40:00           14000          7000
## 46 2004-11-07 16:25:00 2004-11-08 20:00:00           14000            60
## 47 2005-01-17 09:25:00 2005-01-17 16:00:00           14000            30
## 48 2000-11-25 19:00:00 2000-11-25 19:35:00            6000          2000
## 49 1999-10-14 09:10:00 1999-10-14 10:00:00           14000          4000
## 50 2000-11-24 22:24:00 2000-11-24 22:36:00            4000          3000
##    flare_location flare_region flare_classification        cme_datetime
## 1          S19W83        10486                X28. 2003-11-04 19:54:00
```

```
## 2          N19W72         9393                X20. 2001-04-02 22:06:00
## 3          S16E08        10486                X17. 2003-10-28 11:30:00
## 4          S20W85         9415                X14. 2001-04-15 14:06:00
## 5          S15W02        10486                X10. 2003-10-29 20:54:00
## 6          S18W63         8100                X9.4 1997-11-06 12:10:00
## 7          S07E68        10930                X9.0                <NA>
## 8          S14W56        10486                X8.3 2003-11-02 17:30:00
## 9          N14W61        10720                X7.1 2005-01-20 06:54:00
## 10         N17W69        11263                X6.9 2011-08-09 08:12:00
## 11         S05E64        10930                X6.5                <NA>
## 12         S12E67        10808                X6.2 2005-09-09 19:48:00
## 13         N22W07         9077                X5.7 2000-07-14 10:54:00
## 14         S21E31         9415                X5.6 2001-04-06 19:30:00
## 15         N17E27        11429                X5.4 2012-03-07 00:24:00
## 16         S17E34         9591                X5.3 2001-08-25 16:50:00
## 17         S13E82        11990                X4.9 2014-02-25 01:25:00
## 18         S13E72        10039                X4.8 2002-07-23 00:42:00
## 19         N18W38         9236                X4.0 2000-11-26 17:06:00
## 20         N08W77        10488                X3.9 2003-11-03 10:06:00
## 21         N15W25        10720                X3.8 2005-01-17 09:54:00
## 22         S06W21        10365                X3.6 2003-05-28 00:50:00
## 23         S26E90         9756                X3.4 2001-12-28 20:30:00
## 24         S06W23        10930                X3.4 2006-12-13 02:54:00
## 25          SE90b        10039                X3.3 2002-07-20 22:06:00
## 26         N08E77        11748                X3.2 2013-05-14 01:25:00
## 27         S02W81        10069                X3.1 2002-08-24 01:27:00
## 28         N11E85        11748                X2.8 2013-05-13 16:07:00
## 29         S11W65         8210                X2.7 1998-05-06 08:29:00
## 30         N10W83        10488                X2.7 2003-11-03 01:59:00
## 31         N15E79        12339                X2.7 2015-05-05 22:24:00
## 32         N17E63         8113                X2.6 1997-11-27 13:56:00
## 33         S16E23         9632                X2.6 2001-09-24 10:30:00
## 34         N15W05        10720                X2.6 2005-01-15 23:06:00
## 35         N09W49        10696                X2.5 2004-11-10 02:26:00
## 36         N20E18         9026                X2.3 2000-06-06 15:54:00
## 37         N22W07         9236                X2.3 2000-11-24 15:30:00
## 38         S23W09         9415                X2.3 2001-04-10 05:30:00
## 39         S20W12        11158                X2.2 2011-02-15 02:24:00
## 40         S14W33         8100                X2.1 1997-11-04 06:10:00
## 41         S13E47        10808                X2.1 2005-09-10 21:52:00
## 42         N14W18        11283                X2.1 2011-09-06 23:05:00
## 43         S06E69        11882                X2.1 2013-10-25 15:12:00
## 44         N20W05         9236                X2.0 2000-11-24 05:30:00
## 45         S19W43         9415                X2.0 2001-04-12 10:31:00
## 46         N09W17        10696                X2.0 2004-11-07 16:54:00
## 47         N15W25        10720                X2.0 2005-01-17 09:30:00
## 48         N20W23         9236                X1.9 2000-11-25 19:31:00
## 49         N11E32         8731                X1.8 1999-10-14 09:26:00
## 50         N21W14         9236                X1.8 2000-11-24 22:06:00
##    cme_angle cme_width cme_speed plots  halo cme_width_limit
## 1         NA       360      2657  PHTX  TRUE           FALSE
## 2        261       244      2505  PHTX FALSE           FALSE
## 3         NA       360      2459  PHTX  TRUE           FALSE
## 4        245       167      1199  PHTX FALSE           FALSE
```

```
## 5         NA       360      2029  PHTX  TRUE          FALSE
## 6         NA       360      1556  PHTX  TRUE          FALSE
## 7         NA      <NA>        NA  PHTX    NA             NA
## 8         NA       360      2598  PHTX  TRUE          FALSE
## 9         NA       360       882  PHTX  TRUE          FALSE
## 10        NA       360      1610  PHTX  TRUE          FALSE
## 11        NA      <NA>        NA  PHTX    NA             NA
## 12        NA       360      2257  PHTX  TRUE          FALSE
## 13        NA       360      1674  PHTX  TRUE          FALSE
## 14        NA       360      1270  PHTX  TRUE          FALSE
## 15        NA       360      2684  PHTX  TRUE          FALSE
## 16        NA       360      1433  PHTX  TRUE          FALSE
## 17        NA       360      2147  PHTX  TRUE          FALSE
## 18        NA       360      2285  PHTX  TRUE          FALSE
## 19        NA       360       980  PHTX  TRUE          FALSE
## 20       293       103      1420  PHTX FALSE          FALSE
## 21        NA       360      2547  PHTX  TRUE          FALSE
## 22        NA       360      1366  PHTX  TRUE          FALSE
## 23        NA       360      2216  PHTX  TRUE          FALSE
## 24        NA       360      1774  PHTX  TRUE          FALSE
## 25        NA       360      1941  PHTX  TRUE          FALSE
## 26        NA       360      2625  PHTX  TRUE          FALSE
## 27        NA       360      1913  PHTX  TRUE          FALSE
## 28        NA       360      1850  PHTX  TRUE          FALSE
## 29       309       190      1099  PHTX FALSE          FALSE
## 30       304        65       827  PHTX FALSE          FALSE
## 31        NA       360       715  PHTX  TRUE          FALSE
## 32        98        91       441  PHTX FALSE          FALSE
## 33        NA       360      2402  PHTX  TRUE          FALSE
## 34        NA       360      2861  PHTX  TRUE          FALSE
## 35        NA       360      3387  PHTX  TRUE          FALSE
## 36        NA       360      1119  PHTX  TRUE          FALSE
## 37        NA       360      1245  PHTX  TRUE          FALSE
## 38        NA       360      2411  PHTX  TRUE          FALSE
## 39        NA       360       669  PHTX  TRUE          FALSE
## 40        NA       360       785  PHTX  TRUE          FALSE
## 41        NA       360      1893  PHTX  TRUE          FALSE
## 42        NA       360       575  PHTX  TRUE          FALSE
## 43        NA       360      1081  PHTX  TRUE          FALSE
## 44        NA       360      1289  PHTX  TRUE          FALSE
## 45        NA       360      1184  PHTX  TRUE          FALSE
## 46        NA       360      1759  PHTX  TRUE          FALSE
## 47        NA       360      2094  PHTX  TRUE          FALSE
## 48        NA       360       671  PHTX  TRUE          FALSE
## 49        NA       360      1250  PHTX  TRUE          FALSE
## 50        NA       360      1005  PHTX  TRUE          FALSE
```

Question 2: Entity Resolution

Similarity will be based on flare_classification, start_datetime, and flare_region because i view these as important attributes that can uniquely describe each flare and they are also attributes that both data sets contain. Flare_classification is weighted higher than the other 2 attributes but flares are not considered similar if only their

classifications are the same; they need to match in either datetime or flare region or both along with flare_classification. If two flares are similar, then they are a match. If multiple matches, it is arbitarily decided but this should be a very rare case as it is very unlikely.

```
flare_similarity <- function(e1,e2)
{
  score = 0
  if (e1$start_datetime == e2$start_datetime)
    score = score + .3
  if (e1$flare_classification == e2$flare_classification)
    score = score + .7
  if (e1$flare_region == e2$flare_region)
    score = score + .3
  return(score)
}

flare_match <- function(e1,e2)
{
  if (flare_similarity(e1,e2) > 1){
   index = which((NASA_data == e1$flare_classification) && (NASA_data == e1$datetime) || (NASA_dat
a = e1$flare_region), arr.ind = TRUE)
  return(index)
  }else
    return(NA)
}
sim_vec = numeric(50)
begin <- function() {
  for (x in 1:50) {
    for(y in 1:nrow(NASA)) {
      value = flare_match(sf[x,], NASA[y,])
      sim_vec[x] = value
    }
  }
return(sim_vec)
}

matches <- begin()
with_index <- sf %>% mutate(index = matches)
as_tibble(with_index)
```

```
## # A tibble: 50 x 9
##     rank flare_classification start_datetime      max_datetime
##    <int> <chr>                <dttm>              <dttm>
##  1     1 X28.0                2003-11-04 19:29:00 2003-11-04 19:53:00
##  2     2 X20.0                2001-04-02 21:32:00 2001-04-02 21:51:00
##  3     3 X17.2                2003-10-28 09:51:00 2003-10-28 11:10:00
##  4     4 X17.0                2005-09-07 17:17:00 2005-09-07 17:40:00
##  5     5 X14.4                2001-04-15 13:19:00 2001-04-15 13:50:00
##  6     6 X10.0                2003-10-29 20:37:00 2003-10-29 20:49:00
##  7     7 X9.4                 1997-11-06 11:49:00 1997-11-06 11:55:00
##  8     8 X9.3                 2017-09-06 11:53:00 2017-09-06 12:02:00
##  9     9 X9.0                 2006-12-05 10:18:00 2006-12-05 10:35:00
## 10    10 X8.3                 2003-11-02 17:03:00 2003-11-02 17:25:00
## # ... with 40 more rows, and 5 more variables: end_datetime <dttm>,
## #   flare_region <int>, start_time <time>, maximum_time <time>,
## #   index <dbl>
```

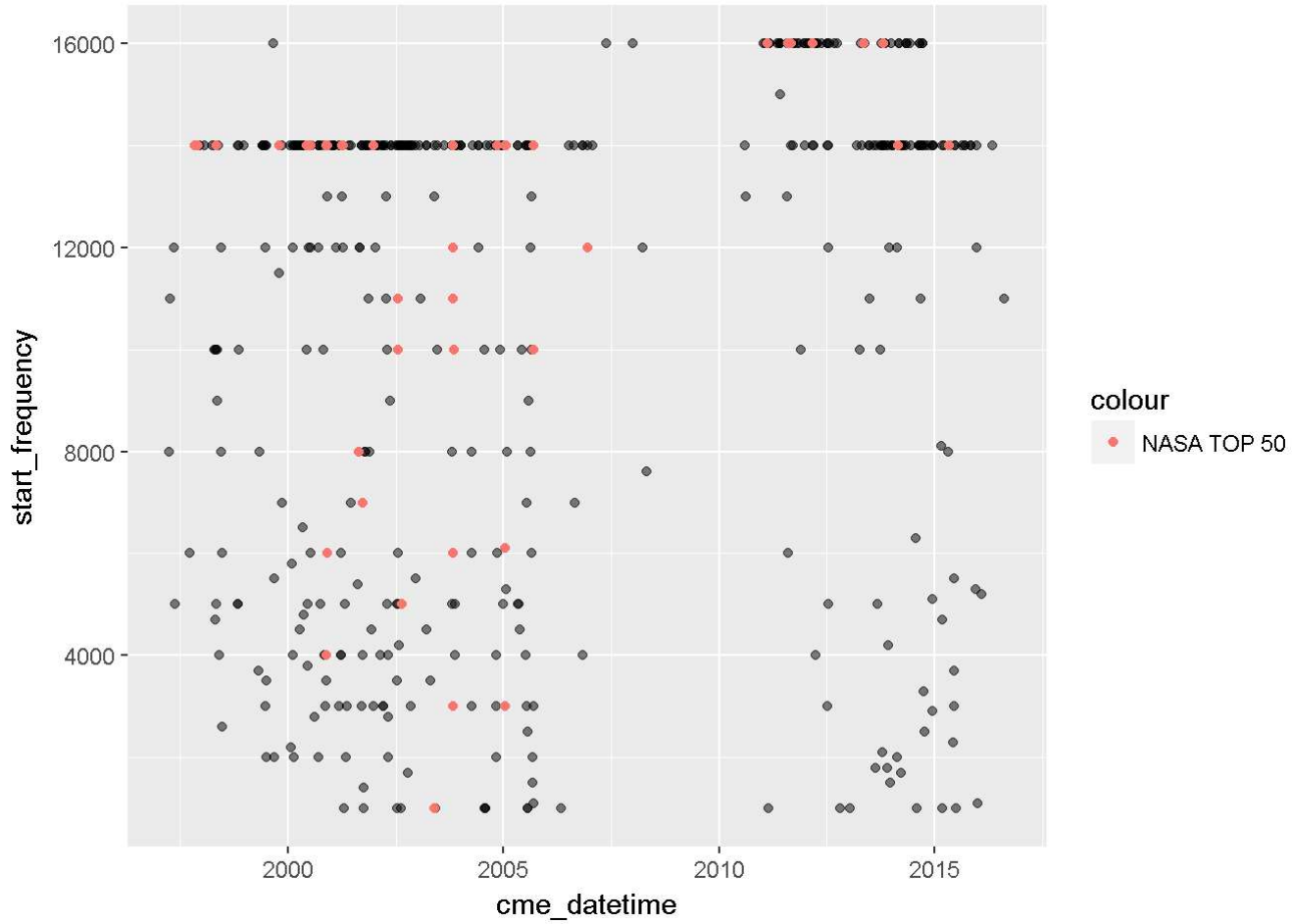
question 3: Analysis

I wanted to see if the starting frequency changed over time. The scatter plot is also useful to see how spread apart my data is or variance.

```
ggplot(NASA_data, mapping = aes(y = start_frequency, x = cme_datetime)) +
  geom_point(alpha = .5)+
  geom_point(data = NASA_top50, mapping = aes(y = start_frequency , x = cme_datetime, color = "NAS
A TOP 50"))
```

```
## Warning: Removed 28 rows containing missing values (geom_point).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



On the y-Axis i plotted the start_frequency because that is the dependent variable i will be analyzing. On the x-axis i plotted the datetime because that is the independent varaibl and I wanted tos ee how the frequency would change over time. It seems that most starting frequencys start at around 14000kHz and that the top 50 data set has starting frequencies all over the place. The variance would be a larger number because the data seems spread out ranging from <1000khz to >16000kHz.