Project1

Joon

March 7, 2018

start of code

```
##scrapes the top 50 data using html methods

  library(rvest)
```

```
## Loading required package: xml2
```

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

space_data <- url %>%
  read_html() %>%
  html_node(".table-striped") %>%
  html_table() %>%
  set_colnames(c("rank","flare_classification","date","flare_region","start_time","maximum_time","end_t
  as.data.frame()



space_data
```

```
##    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
## 11   11                 X8.2 2017/09/10         2673      15:35
## 12   12                 X7.1 2005/01/20          720      06:36
## 13   13                 X6.9 2011/08/09         1263      07:48
## 14   14                 X6.5 2006/12/06          930      18:29
## 15   15                 X6.2 2005/09/09          808      19:13
## 16   16                 X6.2 2001/12/13         9733      14:20
## 17   17                 X5.7 2000/07/14         9077      10:03
## 18   18                 X5.6 2001/04/06         9415      19:10
## 19   19                 X5.4 2012/03/07         1429      00:02
## 20   20                 X5.4 2003/10/23          486      08:19
## 21   21                 X5.4 2005/09/08          808      20:52
## 22   22                 X5.3 2001/08/25         9591      16:23
## 23   23                 X4.9 1998/08/18         8307      22:10
## 24   24                 X4.9 2014/02/25         1990      00:39
## 25   25                 X4.8 2002/07/23           39      00:18
## 26   26                 X4.0 2000/11/26         9236      16:34
## 27   27                 X3.9 1998/08/19         8307      21:35
## 28   28                 X3.9 2003/11/03          488      09:43
## 29   29                 X3.8 2005/01/17          720      06:59
## 30   30                 X3.7 1998/11/22         8384      06:30
## 31   31                 X3.6 2003/05/28          365      00:17
## 32   32                 X3.6 2004/07/16          649      13:49
## 33   33                 X3.6 2005/09/09          808      09:42
## 34   34                 X3.4 2006/12/13          930      02:14
## 35   35                 X3.4 2001/12/28         9767      20:02
## 36   36                 X3.3 1998/11/28         8395      04:54
## 37   37                 X3.3 2002/07/20           39      21:04
## 38   38                 X3.3 2013/11/05         1890      22:07
## 39   39                 X3.2 2013/05/14         1748      00:00
## 40   40                 X3.1 2014/10/24         2192      21:07
## 41   41                 X3.1 2002/08/24           69      00:49
## 42   42                 X3.0 2002/07/15           30      19:59
## 43   43                 X2.8 1998/08/18         8307      08:14
## 44   44                 X2.8 2001/12/11         9733      07:58
## 45   45                 X2.8 2013/05/13         1748      15:48
## 46   46                 X2.7 2015/05/05         2339      22:05
## 47   47                 X2.7 1998/05/06         8210      07:58
## 48   48                 X2.7 2003/11/03          488      01:09
## 49   49                 X2.6 2005/01/15          720      22:25
## 50   50                 X2.6 1997/11/27         8113      12:59
##    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
## 11        16:06    16:31      View archive
## 12        07:01    07:26 MovieView archive
## 13        08:05    08:08 MovieView archive
## 14        18:47    19:00 MovieView archive
## 15        20:04    20:36 MovieView archive
## 16        14:30    14:35 MovieView archive
## 17        10:24    10:43 MovieView archive
## 18        19:21    19:31 MovieView archive
## 19        00:24    00:40 MovieView archive
## 20        08:35    08:49 MovieView archive
## 21        21:06    21:17 MovieView archive
## 22        16:45    17:04 MovieView archive
## 23        22:19    22:28      View archive
## 24        00:49    01:03 MovieView archive
## 25        00:35    00:47 MovieView archive
## 26        16:48    16:56 MovieView archive
## 27        21:45    21:50      View archive
## 28        09:55    10:19 MovieView archive
## 29        09:52    10:07 MovieView archive
## 30        06:42    06:49 MovieView archive
## 31        00:27    00:39 MovieView archive
## 32        13:55    14:01 MovieView archive
## 33        09:59    10:08 MovieView archive
## 34        02:40    02:57 MovieView archive
## 35        20:45    21:32 MovieView archive
## 36        05:52    06:13 MovieView archive
## 37        21:30    21:54 MovieView archive
## 38        22:12    22:15 MovieView archive
## 39        01:11    01:20 MovieView archive
## 40        21:41    22:13 MovieView archive
## 41        01:12    01:31 MovieView archive
## 42        20:08    20:14 MovieView archive
## 43        08:24    08:32      View archive
## 44        08:08    08:14 MovieView archive
## 45        16:05    16:16 MovieView archive
## 46        22:11    22:15 MovieView archive
## 47        08:09    08:20 MovieView archive
## 48        01:30    01:45 MovieView archive
## 49        23:02    23:31 MovieView archive
## 50        13:17    13:20 MovieView archive
```

```
## scrapestidys the top 50 solar flare data
  library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
  library(tidyr)
```

```
##
## Attaching package: 'tidyr'

## The following object is masked from 'package:magrittr':
##
##     extract
```

```
  library(readr)
```

```
##
## Attaching package: 'readr'

## The following object is masked from 'package:rvest':
##
##     guess_encoding
```

```
##uniting the times to make datetimes for tidying it pipes each result with false so we can reuse the da

tidy_space_data <- unite(space_data, start_datetime,c(date,start_time),sep = " ",remove = FALSE) %>%
              unite(maximum_datetime,c(date,maximum_time),sep = " ", remove = FALSE) %>%
              unite(end_datetime,c(date,end_time),sep = " ") %>%
              subset(select = -movie)

##converting to the right values using posixct
              tidy_space_data$start_datetime<- as.POSIXct(tidy_space_data$start_datetime)
              tidy_space_data$maximum_datetime <- as.POSIXct(tidy_space_data$maximum_datetime)
              tidy_space_data$end_datetime <- as.POSIXct(tidy_space_data$end_datetime)

tidy_space_data
```

```
##    rank flare_classification      start_datetime    maximum_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
## 11   11                 X8.2 2017-09-10 15:35:00 2017-09-10 16:06:00
## 12   12                 X7.1 2005-01-20 06:36:00 2005-01-20 07:01:00
## 13   13                 X6.9 2011-08-09 07:48:00 2011-08-09 08:05:00
## 14   14                 X6.5 2006-12-06 18:29:00 2006-12-06 18:47:00
## 15   15                 X6.2 2005-09-09 19:13:00 2005-09-09 20:04:00
## 16   16                 X6.2 2001-12-13 14:20:00 2001-12-13 14:30:00
## 17   17                 X5.7 2000-07-14 10:03:00 2000-07-14 10:24:00
## 18   18                 X5.6 2001-04-06 19:10:00 2001-04-06 19:21:00
## 19   19                 X5.4 2012-03-07 00:02:00 2012-03-07 00:24:00
## 20   20                 X5.4 2003-10-23 08:19:00 2003-10-23 08:35:00
## 21   21                 X5.4 2005-09-08 20:52:00 2005-09-08 21:06:00
## 22   22                 X5.3 2001-08-25 16:23:00 2001-08-25 16:45:00
## 23   23                 X4.9 1998-08-18 22:10:00 1998-08-18 22:19:00
## 24   24                 X4.9 2014-02-25 00:39:00 2014-02-25 00:49:00
## 25   25                 X4.8 2002-07-23 00:18:00 2002-07-23 00:35:00
## 26   26                 X4.0 2000-11-26 16:34:00 2000-11-26 16:48:00
## 27   27                 X3.9 1998-08-19 21:35:00 1998-08-19 21:45:00
## 28   28                 X3.9 2003-11-03 09:43:00 2003-11-03 09:55:00
## 29   29                 X3.8 2005-01-17 06:59:00 2005-01-17 09:52:00
## 30   30                 X3.7 1998-11-22 06:30:00 1998-11-22 06:42:00
## 31   31                 X3.6 2003-05-28 00:17:00 2003-05-28 00:27:00
## 32   32                 X3.6 2004-07-16 13:49:00 2004-07-16 13:55:00
## 33   33                 X3.6 2005-09-09 09:42:00 2005-09-09 09:59:00
## 34   34                 X3.4 2006-12-13 02:14:00 2006-12-13 02:40:00
## 35   35                 X3.4 2001-12-28 20:02:00 2001-12-28 20:45:00
## 36   36                 X3.3 1998-11-28 04:54:00 1998-11-28 05:52:00
## 37   37                 X3.3 2002-07-20 21:04:00 2002-07-20 21:30:00
## 38   38                 X3.3 2013-11-05 22:07:00 2013-11-05 22:12:00
## 39   39                 X3.2 2013-05-14 00:00:00 2013-05-14 01:11:00
## 40   40                 X3.1 2014-10-24 21:07:00 2014-10-24 21:41:00
## 41   41                 X3.1 2002-08-24 00:49:00 2002-08-24 01:12:00
## 42   42                 X3.0 2002-07-15 19:59:00 2002-07-15 20:08:00
## 43   43                 X2.8 1998-08-18 08:14:00 1998-08-18 08:24:00
## 44   44                 X2.8 2001-12-11 07:58:00 2001-12-11 08:08:00
## 45   45                 X2.8 2013-05-13 15:48:00 2013-05-13 16:05:00
## 46   46                 X2.7 2015-05-05 22:05:00 2015-05-05 22:11:00
## 47   47                 X2.7 1998-05-06 07:58:00 1998-05-06 08:09:00
## 48   48                 X2.7 2003-11-03 01:09:00 2003-11-03 01:30:00
## 49   49                 X2.6 2005-01-15 22:25:00 2005-01-15 23:02:00
## 50   50                 X2.6 1997-11-27 12:59:00 1997-11-27 13:17:00
##           end_datetime flare_region start_time maximum_time
## 1  2003-11-04 20:06:00          486      19:29        19:53
## 2  2001-04-02 22:03:00         9393      21:32        21:51
## 3  2003-10-28 11:24:00          486      09:51        11:10
## 4  2005-09-07 18:03:00          808      17:17        17:40
## 5  2001-04-15 13:55:00         9415      13:19        13:50
## 6  2003-10-29 21:01:00          486      20:37        20:49
## 7  1997-11-06 12:01:00         8100      11:49        11:55
## 8  2017-09-06 12:10:00         2673      11:53        12:02
## 9  2006-12-05 10:45:00          930      10:18        10:35
## 10 2003-11-02 17:39:00          486      17:03        17:25
## 11 2017-09-10 16:31:00         2673      15:35        16:06
## 12 2005-01-20 07:26:00          720      06:36        07:01
## 13 2011-08-09 08:08:00         1263      07:48        08:05
## 14 2006-12-06 19:00:00          930      18:29        18:47
## 15 2005-09-09 20:36:00          808      19:13        20:04
## 16 2001-12-13 14:35:00         9733      14:20        14:30
## 17 2000-07-14 10:43:00         9077      10:03        10:24
## 18 2001-04-06 19:31:00         9415      19:10        19:21
## 19 2012-03-07 00:40:00         1429      00:02        00:24
## 20 2003-10-23 08:49:00          486      08:19        08:35
## 21 2005-09-08 21:17:00          808      20:52        21:06
## 22 2001-08-25 17:04:00         9591      16:23        16:45
## 23 1998-08-18 22:28:00         8307      22:10        22:19
## 24 2014-02-25 01:03:00         1990      00:39        00:49
## 25 2002-07-23 00:47:00           39      00:18        00:35
## 26 2000-11-26 16:56:00         9236      16:34        16:48
## 27 1998-08-19 21:50:00         8307      21:35        21:45
## 28 2003-11-03 10:19:00          488      09:43        09:55
## 29 2005-01-17 10:07:00          720      06:59        09:52
## 30 1998-11-22 06:49:00         8384      06:30        06:42
## 31 2003-05-28 00:39:00          365      00:17        00:27
## 32 2004-07-16 14:01:00          649      13:49        13:55
## 33 2005-09-09 10:08:00          808      09:42        09:59
## 34 2006-12-13 02:57:00          930      02:14        02:40
## 35 2001-12-28 21:32:00         9767      20:02        20:45
## 36 1998-11-28 06:13:00         8395      04:54        05:52
## 37 2002-07-20 21:54:00           39      21:04        21:30
## 38 2013-11-05 22:15:00         1890      22:07        22:12
## 39 2013-05-14 01:20:00         1748      00:00        01:11
## 40 2014-10-24 22:13:00         2192      21:07        21:41
## 41 2002-08-24 01:31:00           69      00:49        01:12
## 42 2002-07-15 20:14:00           30      19:59        20:08
## 43 1998-08-18 08:32:00         8307      08:14        08:24
## 44 2001-12-11 08:14:00         9733      07:58        08:08
## 45 2013-05-13 16:16:00         1748      15:48        16:05
## 46 2015-05-05 22:15:00         2339      22:05        22:11
## 47 1998-05-06 08:20:00         8210     07:58        08:09
## 48 2003-11-03 01:45:00          488     01:09        01:30
## 49 2005-01-15 23:31:00          720     22:25        23:02
## 50 1997-11-27 13:20:00         8113     12:59        13:17
```

```
##scrapes and tidys the nasa table

    library(rvest)
    library(stringr)
    library(readr)
    library(tidyr)
    library(dplyr)

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

whitespace <-"\\s+"
solar_flare <- url %>%
        read_html() %>%
        html_node("pre") %>%
        html_text() %>%

# splits with  a newline as that is what separates the rows

        str_split("\n",simplify = TRUE) %>%

  ##finding all incomplete entries and setting to NA (based on the  website description)

        str_replace_all("\\?\\?\\?\\?","NA") %>%
        str_replace_all("--/--","NA") %>%
        str_replace_all("--:--","NA") %>%
        str_replace_all("-----","NA") %>%
        str_replace_all("----","NA") %>%
        str_replace_all("SW90b","NA") %>%
        str_replace_all("Back","NA") %>%
        str_replace_all("BACK", "NA") %>%
        str_replace_all("back\\?","NA") %>%
        str_subset(".*PHTX") %>%
        as_data_frame() %>%

  ##separating into new cols for tidy data separates using whitespace which == \\s+

        separate(value,
                c("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"),sep= whitespace ) %>%


  ## creating new cols halo and width_limit that take logical values true or false

        mutate(Halo = ifelse(cme_angle == "Halo",TRUE,FALSE)) %>%
        mutate(cme_width_limit = ifelse(grepl(">",cme_width),TRUE,FALSE)) %>%
  ##uniting the times and dates

        unite(start_datetime,c(start_date,start_time),sep = " ", remove = FALSE) %>%
        unite(end_datetime,c(start_date,end_time),sep = " ",remove = FALSE) %>%

  ## I united startdate and cme_time becuase cme_time didn't have the right format for posixct conversi
        unite(cme_datetime,c(start_date,cme_time),sep = " ", remove = FALSE) %>%
        subset(select = -c(start_date, start_time,end_date,end_time,cme_date,cme_time))
```

```
## Warning: Too many values at 511 locations: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
## 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...
```

```
## getting rid of non numerics in width and setting any Halo values in cme_angle to NA
##grepped to find where there was an NA united with start date as I noticed in the data
##that there was no cases where cme_date != NA while cme_time == NA
##can safely assume that cme_datetime can be NA if it has NA anywhere in the column
## I also convert every chr NA value to the actual NA value

        solar_flare$cme_datetime[grepl("NA",solar_flare$cme_datetime) == TRUE] <- NA
        solar_flare[solar_flare == "NA"] <- NA
        solar_flare$cme_width <- ifelse(grepl(">",solar_flare$cme_width),
                                 substring(solar_flare$cme_width,2),solar_flare$cme_width)
        solar_flare$cme_angle[solar_flare$cme_angle == "Halo"] <- NA

## converting types

        solar_flare$cme_datetime <- as.POSIXct(solar_flare$cme_datetime)
        solar_flare$start_datetime <- as.POSIXct(solar_flare$start_datetime)
        solar_flare$end_datetime <- as.POSIXct(solar_flare$end_datetime)
        solar_flare$cme_datetime <- as.POSIXct(solar_flare$cme_datetime)
        solar_flare$start_frequency <- as.integer(solar_flare$start_frequency)
        solar_flare$end_frequency <- as.integer(solar_flare$end_frequency)
        solar_flare$cme_angle <- as.integer(solar_flare$cme_angle)
        solar_flare$cme_speed <- as.integer(solar_flare$cme_speed)
        solar_flare$cme_width <- as.integer(solar_flare$cme_width)
```

```
## Warning: NAs introduced by coercion
```

```
solar_flare
```

```
## # A tibble: 511 x 13
##    start_datetime      end_datetime        cme_datetime
##    <dttm>              <dttm>              <dttm>
##  1 1997-04-01 14:00:00 1997-04-01 14:15:00 1997-04-01 15:18:00
##  2 1997-04-07 14:30:00 1997-04-07 17:30:00 1997-04-07 14:27:00
##  3 1997-05-12 05:15:00 1997-05-12 16:00:00 1997-05-12 05:30:00
##  4 1997-05-21 20:20:00 1997-05-21 22:00:00 1997-05-21 21:00:00
##  5 1997-09-23 21:53:00 1997-09-23 22:16:00 1997-09-23 22:02:00
##  6 1997-11-03 05:15:00 1997-11-03 12:00:00 1997-11-03 05:28:00
##  7 1997-11-03 10:30:00 1997-11-03 11:30:00 1997-11-03 11:11:00
##  8 1997-11-04 06:00:00 1997-11-04 04:30:00 1997-11-04 06:10:00
##  9 1997-11-06 12:20:00 1997-11-06 08:30:00 1997-11-06 12:10:00
## 10 1997-11-27 13:30:00 1997-11-27 14:00:00 1997-11-27 13:56:00
## # ... with 501 more rows, and 10 more variables: start_frequency <int>,
## #   end_frequency <int>, flare_location <chr>, flare_region <chr>,
## #   flare_classification <chr>, cme_angle <int>, cme_width <int>,
## #   cme_speed <int>, Halo <lgl>, cme_width_limit <lgl>
```

```
library(gtools)

replication <- solar_flare[mixedorder(solar_flare$flare_classification, decreasing = TRUE),] %>%
  filter(!is.na(flare_classification)) %>%
  slice(1:50)

## I was able to replicate most of the data however there seems to be discrepancies
##between the data given in the https://cdaw.gsfc.nasa.gov/CME_list/radio/waves_type2.html site vs the
##https://www.spaceweatherlive.com/en/solar-activity/top-50-solar-flares
##mainly due to the fact that the data just isn't recorded in the untidy data we had
##to tidy in my solar_flare method
## I checked to see if the data was there and it was not so I conclude that the
##discrepancy is mainly just from it not being recorded.

replication
```

```
## # A tibble: 50 x 13
##    start_datetime      end_datetime        cme_datetime
##    <dttm>              <dttm>              <dttm>
##  1 2003-11-04 20:00:00 2003-11-05 00:00:00 2003-11-04 19:54:00
##  2 2001-04-02 22:05:00 2001-04-02 02:30:00 2001-04-02 22:06:00
##  3 2003-10-28 11:10:00 2003-10-29 00:00:00 2003-10-28 11:30:00
##  4 2001-04-15 14:05:00 2001-04-15 13:00:00 2001-04-15 14:06:00
##  5 2003-10-29 20:55:00 2003-10-30 00:00:00 2003-10-29 20:54:00
##  6 1997-11-06 12:20:00 1997-11-06 08:30:00 1997-11-06 12:10:00
##  7 2006-12-05 10:50:00 2006-12-05 20:00:00 NA
##  8 2003-11-02 17:30:00 2003-11-02 01:00:00 2003-11-02 17:30:00
##  9 2005-01-20 07:15:00 2005-01-20 16:30:00 2005-01-20 06:54:00
## 10 2011-08-09 08:20:00 2011-08-09 08:35:00 2011-08-09 08:12:00
## # ... with 40 more rows, and 10 more variables: start_frequency <int>,
## #   end_frequency <int>, flare_location <chr>, flare_region <chr>,
## #   flare_classification <chr>, cme_angle <int>, cme_width <int>,
## #   cme_speed <int>, Halo <lgl>, cme_width_limit <lgl>
```

```
    library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```
    library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------------

## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v forcats 0.2.0

## -- Conflicts ---------------------------------------------------------------------------------------
## x lubridate::as.difftime() masks base::as.difftime()
## x lubridate::date()        masks base::date()
## x tidyr::extract()         masks magrittr::extract()
## x dplyr::filter()          masks stats::filter()
## x readr::guess_encoding()  masks rvest::guess_encoding()
## x lubridate::intersect()   masks base::intersect()
## x dplyr::lag()             masks stats::lag()
## x purrr::pluck()           masks rvest::pluck()
## x purrr::set_names()       masks magrittr::set_names()
## x lubridate::setdiff()     masks base::setdiff()
## x lubridate::union()       masks base::union()
```

```
    library(sqldf)
```

```
## Warning: package 'sqldf' was built under R version 3.4.4
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Warning: package 'proto' was built under R version 3.4.4
```

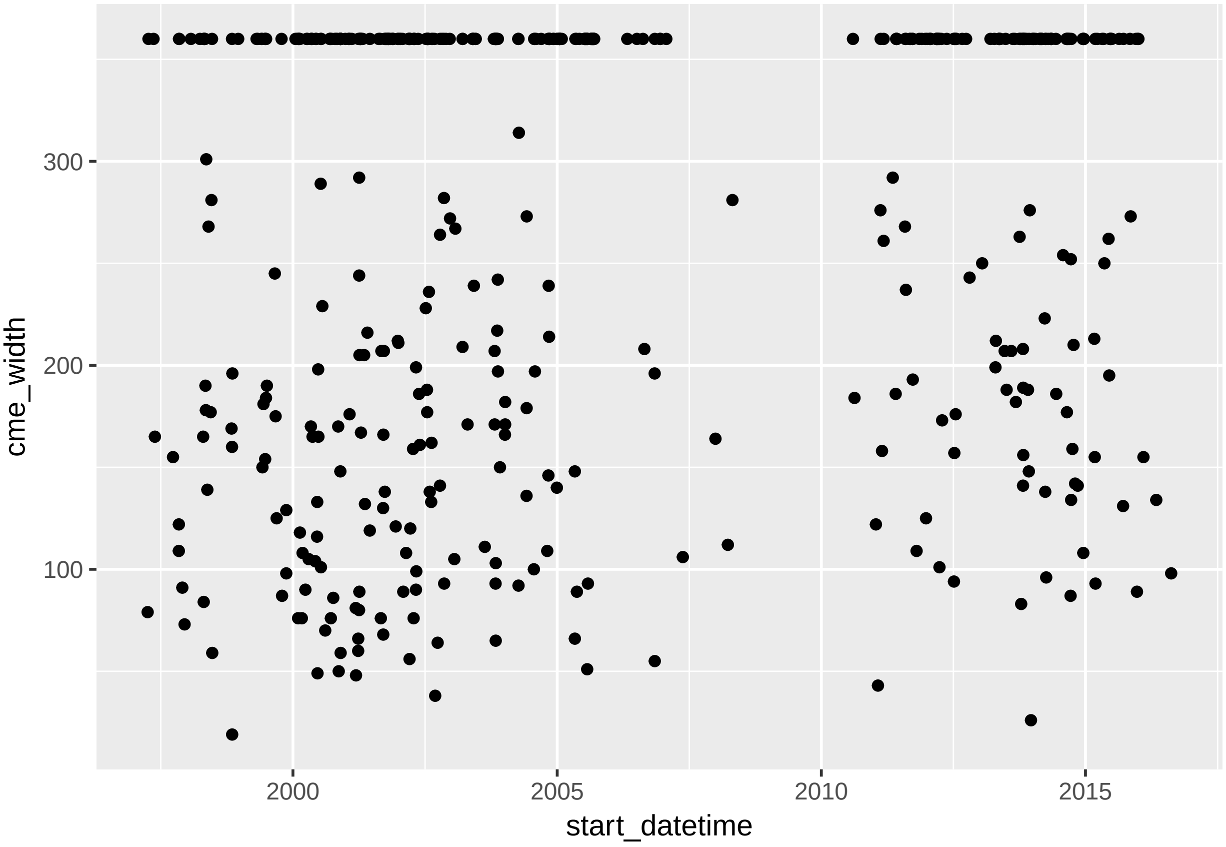
```
## Loading required package: RSQLite
```

```
#used hcorrada github similarity functions as a starting off point

#function to determine how similar start date is
# i give points based on year,month, and day

startyear_similarity <- function(d1,d2) {
  ifelse(year(d1) == year(d2),2.5,0)
}

startmonth_similarity <- function(d1,d2) {
  ifelse(month(d1) == month(d2),2.5,0)
}

startday_similarity <- function(d1,d2) {
  ifelse(day(d1)==day(d2),2.5,0)
}

#function to determine region similiarity
region_similarity <- function(v1,v2) {
  ifelse((v1+ 10000) == v2,2.5,0)

}


#function to determine if flare_classification is the same
class_similarity <-function(v1,v2) {
  v1str = substr(v1,1,1)
  v2str = substr(v2,1,1)
  ifelse(v1str == v2str,2.5,0)
}

#function that puts all the functions together and finds the similarity percentage
similarity_between <-function(v1,v2) {


  sum <-
    startyear_similarity(tidy_space_data$start_datetime[v1],solar_flare$start_datetime[v2])
  sum <- sum +
      startmonth_similarity(tidy_space_data$start_datetime[v1],solar_flare$start_datetime[v2])
  sum <- sum +
      startday_similarity(tidy_space_data$start_datetime[v1],solar_flare$start_datetime[v2])
  sum <- sum +
      region_similarity(tidy_space_data$flare_region[v1],solar_flare$flare_region[v2])
  sum <- sum +
    class_similarity(tidy_space_data$flare_classification[v1],solar_flare$flare_classification[v2])

  sim <- (sum/15) * 100
  return(sim)
}

#flare match function
flare_match <-function(df1,df2){
sim_matrix <- matrix(NA,nrow(df1),nrow(df2))

#finding the similarities between every combination

  for(i in seq(1,nrow(df1))) {
    for(j in seq(1,nrow(df2))){
      s <- similarity_between(i,j)
      ifelse(s == 0,sim_matrix[i,j] <- NA,sim_matrix[i,j] <- s)
    }
  }

# creating the sim_matrix as a data frame
sim_df <- sim_matrix %>%
  magrittr::set_colnames(seq(1,ncol(.))) %>%
  as_data_frame() %>%
  rowid_to_column("rank") %>%
  tidyr::gather(solar_flares, similarity, -rank) %>%
  mutate(solar_flares = as.integer(solar_flares)) %>%

#matching the row which has the highest similarity from top 50 to solar_flare %>%
    group_by(rank) %>%
    summarize(max_sim = max(similarity),index = solar_flares[which.max(similarity)])


#adding the index to the tidy top 50 data
# i use sql because I feel the natural join is the easiest way to join the tables
#I can then just subset the desired cols the matched col is named index as per the proj description

matched_tidy_space_data <- sqldf("select * from tidy_space_data
                                  natural join
                                 sim_df") %>%
  subset(select= -max_sim)
}

#exmaple for my similarity function
sim <- similarity_between(1,243)

#example for my flare_match function
```

```
matched_df_final <- flare_match(tidy_space_data,solar_flare)

##sim function DEFINITION
#given indexes you can compute similarities between the entities at those positions
#ex row 1 of the top 50 table is compared to row 243 in the nasa table. It uses start date, region, and
#classification
#to determine the percentage of similarity between the two entities.
#from there the flare match iterates over both tables to calculate every single similarity
#it then groups by rank (top 50) and finds the most similar (largest) match and gets that index
#SO the top 50 table now has the index of its best match on the NASA table.
# again the nasa data table really isnt accurate for example if u look at the first
#entry in the top 5 table it gives the starttime of 19:29 that same entry is located
# at row 243 in the nasa data table but the start time is rounded to 20:00
#this data is not accurate so i guess just getting a high percentage is enough
```

```
## plotting the solar flare data from NASA
#end me. Pls.
solar_flare %>%
    ggplot(mapping =aes(y=cme_width,x=start_datetime)) + geom_point()
```

```
## Warning: Removed 23 rows containing missing values (geom_point).
```


cme_width
start_datetime