Project1

Danny Brewer

February 24, 2019

Scraping the data from the HTML webpage. Html table is ran and the columns are set. Table is converted to a data frame

```
library(rvest)
```

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

```
library(tidyverse)
```

```
## -- Attaching packages ----------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts -------------- tidyverse_conflicts() --
## x dplyr::filter()         masks stats::filter()
## x readr::guess_encoding() masks rvest::guess_encoding()
## x dplyr::lag()            masks stats::lag()
## x purrr::pluck()          masks rvest::pluck()
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

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

```
library(dplyr)
library(tidyr)
library(readr)


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

solar_flare <- url %>%
  read_html() %>%
  html_node("table") %>%
  html_table() %>%
  set_colnames(c("rank","flare_classification","date","flare_region","start_time","max_time","en
d_time","movie")) %>%
  as_data_frame
```

```
## Warning: `as_data_frame()` is deprecated, use `as_tibble()` (but mind the new semantics).
## This warning is displayed once per session.
```

```
solar_flare <- as_tibble(solar_flare)

solar_flare
```

```
## # A tibble: 50 x 8
##     rank flare_classific~ date  flare_region start_time max_time end_time
##    <int> <chr>            <chr>        <int> <chr>      <chr>    <chr>
##  1     1 X28.0            2003~          486 19:29      19:53    20:06
##  2     2 X20.0            2001~         9393 21:32      21:51    22:03
##  3     3 X17.2            2003~          486 09:51      11:10    11:24
##  4     4 X17.0            2005~          808 17:17      17:40    18:03
##  5     5 X14.4            2001~         9415 13:19      13:50    13:55
##  6     6 X10.0            2003~          486 20:37      20:49    21:01
##  7     7 X9.4             1997~         8100 11:49      11:55    12:01
##  8     8 X9.3             2017~         2673 11:53      12:02    12:10
##  9     9 X9.0             2006~          930 10:18      10:35    10:45
## 10    10 X8.3             2003~          486 17:03      17:25    17:39
## # ... with 40 more rows, and 1 more variable: movie <chr>
```

Slight clean up and the datetime columns are created + converted to datetime objects

```
solar_flare <- solar_flare %>%
  select(-movie) %>%
  unite(start_datetime, date, start_time, sep=" ", remove = FALSE) %>%
  unite(max_datetime, date, max_time, sep=" ", remove = FALSE) %>%
  unite(end_datetime, date, end_time, sep=" ", remove = TRUE) %>%
  select(-start_time) %>%
  select(-max_time) %>%
  type_convert(col_types = cols(start_datetime = col_datetime(format = "%Y/%m/%d %H:%M"))) %>%
  type_convert(col_types = cols(max_datetime = col_datetime(format = "%Y/%m/%d %H:%M"))) %>%
  type_convert(col_types = cols(end_datetime = col_datetime(format = "%Y/%m/%d %H:%M")))

solar_flare
```

```
## # A tibble: 50 x 6
##     rank flare_classific~ 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 2 more variables: end_datetime <dttm>,
## #   flare_region <int>
```

Nasa TypeII bursts are scraped and placed into a dataframe with appropriate column names

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

nasa <- nasa_url %>% read_html() %>%
  html_node("pre") %>%
  html_text(trim =TRUE) %>%
  str_split("\n") %>%
  as_vector() %>%
  str_subset("^\\d{4}") %>%
  as_data_frame() %>%
  separate(value, c("start_date","start_time","end_date","end_time","start_frequency","end_frequ
ency","flare_location","flare_region","flare_classification","cme_date","cme_time","cme_angle",
"cme_width","cme_speed"), sep = "[\\s]+", extra ="drop")

nasa <- as_tibble(nasa)

nasa
```

```
## # A tibble: 511 x 14
##    start_date start_time end_date end_time start_frequency end_frequency
##    <chr>      <chr>      <chr>    <chr>    <chr>           <chr>
##  1 1997/04/01 14:00      04/01    14:15    8000            4000
##  2 1997/04/07 14:30      04/07    17:30    11000           1000
##  3 1997/05/12 05:15      05/14    16:00    12000           80
##  4 1997/05/21 20:20      05/21    22:00    5000            500
##  5 1997/09/23 21:53      09/23    22:16    6000            2000
##  6 1997/11/03 05:15      11/03    12:00    14000           250
##  7 1997/11/03 10:30      11/03    11:30    14000           5000
##  8 1997/11/04 06:00      11/05    04:30    14000           100
##  9 1997/11/06 12:20      11/07    08:30    14000           100
## 10 1997/11/27 13:30      11/27    14:00    14000           7000
## # ... with 501 more rows, and 8 more variables: flare_location <chr>,
## #   flare_region <chr>, flare_classification <chr>, cme_date <chr>,
## #   cme_time <chr>, cme_angle <chr>, cme_width <chr>, cme_speed <chr>
```

Missing information is filled with an NA as specified in the project document

```
nasa <- nasa %>%
  mutate(start_frequency = ifelse(start_frequency == "????", NA, start_frequency)) %>% mutate(en
d_frequency = ifelse(end_frequency == "????", NA, end_frequency)) %>%
  mutate(flare_location = ifelse(str_detect(flare_location, "-----"), NA, flare_location)) %>%
  mutate(flare_region = ifelse(str_detect(flare_region, "----"), NA, flare_region)) %>%
  mutate(flare_classification = ifelse(str_detect(flare_classification, "-"), NA, flare_classifi
cation)) %>%
  mutate(cme_date = ifelse(str_detect(cme_date, "--/--"), NA, cme_date)) %>%
  mutate(cme_time = ifelse(str_detect(cme_time, "--:--"), NA, cme_time)) %>%
  mutate(cme_angle = ifelse(str_detect(cme_angle, "-"), NA, cme_angle)) %>%
  mutate(cme_width = ifelse(str_detect(cme_width, "-"), NA, cme_width)) %>%
  mutate(cme_speed = ifelse(str_detect(cme_speed, "-"), NA, cme_speed)) %>% mutate(flare_classif
ication = ifelse(str_detect(flare_classification, "^[A-Z][0-9]{2}.$"), paste(flare_classificatio
n, "0", sep =""), flare_classification))

nasa
```

```
## # A tibble: 511 x 14
##    start_date start_time end_date end_time start_frequency end_frequency
##    <chr>      <chr>      <chr>    <chr>    <chr>           <chr>
##  1 1997/04/01 14:00      04/01    14:15    8000            4000
##  2 1997/04/07 14:30      04/07    17:30    11000           1000
##  3 1997/05/12 05:15      05/14    16:00    12000           80
##  4 1997/05/21 20:20      05/21    22:00    5000            500
##  5 1997/09/23 21:53      09/23    22:16    6000            2000
##  6 1997/11/03 05:15      11/03    12:00    14000           250
##  7 1997/11/03 10:30      11/03    11:30    14000           5000
##  8 1997/11/04 06:00      11/05    04:30    14000           100
##  9 1997/11/06 12:20      11/07    08:30    14000           100
## 10 1997/11/27 13:30      11/27    14:00    14000           7000
## # ... with 501 more rows, and 8 more variables: flare_location <chr>,
## #   flare_region <chr>, flare_classification <chr>, cme_date <chr>,
## #   cme_time <chr>, cme_angle <chr>, cme_width <chr>, cme_speed <chr>
```

Halo column is altered to where if there is a Halo, there will be a “Y”. Missing Halo information is replaced with NA

```
nasa <- nasa %>%
  mutate(halo = ifelse(str_detect(cme_angle, "Halo"), TRUE, FALSE)) %>%
  mutate(cme_angle = ifelse(str_detect(cme_angle, "Halo"), NA, cme_angle))

nasa
```

```
## # A tibble: 511 x 15
##    start_date start_time end_date end_time start_frequency end_frequency
##    <chr>      <chr>      <chr>    <chr>    <chr>           <chr>
##  1 1997/04/01 14:00      04/01    14:15    8000            4000
##  2 1997/04/07 14:30      04/07    17:30    11000           1000
##  3 1997/05/12 05:15      05/14    16:00    12000           80
##  4 1997/05/21 20:20      05/21    22:00    5000            500
##  5 1997/09/23 21:53      09/23    22:16    6000            2000
##  6 1997/11/03 05:15      11/03    12:00    14000           250
##  7 1997/11/03 10:30      11/03    11:30    14000           5000
##  8 1997/11/04 06:00      11/05    04:30    14000           100
##  9 1997/11/06 12:20      11/07    08:30    14000           100
## 10 1997/11/27 13:30      11/27    14:00    14000           7000
## # ... with 501 more rows, and 9 more variables: flare_location <chr>,
## #   flare_region <chr>, flare_classification <chr>, cme_date <chr>,
## #   cme_time <chr>, cme_angle <chr>, cme_width <chr>, cme_speed <chr>,
## #   halo <lgl>
```

Lower bound column in created and width column is altered to not include the “>” symbol

```
nasa <- nasa %>%
  mutate(cme_width_limit = ifelse(str_detect(cme_width, ">"), TRUE, FALSE)) %>%
  mutate(cme_width = ifelse(str_detect(cme_width, ">"), gsub(">","",cme_width), cme_width))


nasa
```

```
## # A tibble: 511 x 16
##    start_date start_time end_date end_time start_frequency end_frequency
##    <chr>      <chr>      <chr>    <chr>    <chr>           <chr>
##  1 1997/04/01 14:00      04/01    14:15    8000            4000
##  2 1997/04/07 14:30      04/07    17:30    11000           1000
##  3 1997/05/12 05:15      05/14    16:00    12000           80
##  4 1997/05/21 20:20      05/21    22:00    5000            500
##  5 1997/09/23 21:53      09/23    22:16    6000            2000
##  6 1997/11/03 05:15      11/03    12:00    14000           250
##  7 1997/11/03 10:30      11/03    11:30    14000           5000
##  8 1997/11/04 06:00      11/05    04:30    14000           100
##  9 1997/11/06 12:20      11/07    08:30    14000           100
## 10 1997/11/27 13:30      11/27    14:00    14000           7000
## # ... with 501 more rows, and 10 more variables: flare_location <chr>,
## #   flare_region <chr>, flare_classification <chr>, cme_date <chr>,
## #   cme_time <chr>, cme_angle <chr>, cme_width <chr>, cme_speed <chr>,
## #   halo <lgl>, cme_width_limit <lgl>
```

To prevent any potential issues, all 24:00 times were set back a minute and years were added to the end and cme date

```
nasa <- nasa %>% mutate(start_time = ifelse(str_detect(start_time, "24:00"), gsub("24:00","23:5
9",start_time), start_time)) %>%
  mutate(end_time = ifelse(str_detect(end_time, "24:00"), gsub("24:00","23:59",end_time), end_ti
me)) %>%
  mutate(cme_time = ifelse(str_detect(cme_time, "24:00"), gsub("24:00","23:59",cme_time), cme_ti
me)) %>%
  mutate(end_date = paste(substr(start_date,1,4), end_date, sep="/")) %>%
  mutate(cme_date = paste(substr(start_date,1,4), cme_date, sep="/"))

nasa
```

```
## # A tibble: 511 x 16
##    start_date start_time end_date end_time start_frequency end_frequency
##    <chr>      <chr>      <chr>    <chr>    <chr>           <chr>
##  1 1997/04/01 14:00      1997/04~ 14:15    8000            4000
##  2 1997/04/07 14:30      1997/04~ 17:30    11000           1000
##  3 1997/05/12 05:15      1997/05~ 16:00    12000           80
##  4 1997/05/21 20:20      1997/05~ 22:00    5000            500
##  5 1997/09/23 21:53      1997/09~ 22:16    6000            2000
##  6 1997/11/03 05:15      1997/11~ 12:00    14000           250
##  7 1997/11/03 10:30      1997/11~ 11:30    14000           5000
##  8 1997/11/04 06:00      1997/11~ 04:30    14000           100
##  9 1997/11/06 12:20      1997/11~ 08:30    14000           100
## 10 1997/11/27 13:30      1997/11~ 14:00    14000           7000
## # ... with 501 more rows, and 10 more variables: flare_location <chr>,
## #   flare_region <chr>, flare_classification <chr>, cme_date <chr>,
## #   cme_time <chr>, cme_angle <chr>, cme_width <chr>, cme_speed <chr>,
## #   halo <lgl>, cme_width_limit <lgl>
```

Date time columns are created

```
nasa <- nasa %>%
  unite(start_datetime, start_date, start_time, sep=" ", remove = TRUE) %>%
  unite(end_datetime, end_date, end_time, sep=" ", remove = TRUE) %>%
  unite(cme_datetime, cme_date, cme_time, sep=" ", remove = TRUE) %>%
  mutate(cme_datetime = ifelse(str_detect(cme_datetime, "NA"), NA, cme_datetime))

nasa
```

```
## # A tibble: 511 x 13
##    start_datetime end_datetime start_frequency end_frequency flare_location
##    <chr>          <chr>        <chr>           <chr>         <chr>
##  1 1997/04/01 14~ 1997/04/01 ~ 8000            4000          S25E16
##  2 1997/04/07 14~ 1997/04/07 ~ 11000           1000          S28E19
##  3 1997/05/12 05~ 1997/05/14 ~ 12000           80            N21W08
##  4 1997/05/21 20~ 1997/05/21 ~ 5000            500           N05W12
##  5 1997/09/23 21~ 1997/09/23 ~ 6000            2000          S29E25
##  6 1997/11/03 05~ 1997/11/03 ~ 14000           250           S20W13
##  7 1997/11/03 10~ 1997/11/03 ~ 14000           5000          S16W21
##  8 1997/11/04 06~ 1997/11/05 ~ 14000           100           S14W33
##  9 1997/11/06 12~ 1997/11/07 ~ 14000           100           S18W63
## 10 1997/11/27 13~ 1997/11/27 ~ 14000           7000          N17E63
## # ... with 501 more rows, and 8 more variables: flare_region <chr>,
## #   flare_classification <chr>, cme_datetime <chr>, cme_angle <chr>,
## #   cme_width <chr>, cme_speed <chr>, halo <lgl>, cme_width_limit <lgl>
```

Datetime columns converted to datetime objects and other columns converted to proper column types

```
nasa <- nasa[c(1,2,8,3:7,9:13)] %>%
  type_convert(col_types = cols(start_datetime = col_datetime(format = "%Y/%m/%d %H:%M"))) %>%
  type_convert(col_types = cols(end_datetime = col_datetime(format = "%Y/%m/%d %H:%M"))) %>%
  type_convert(col_types = cols(cme_datetime = col_datetime(format = "%Y/%m/%d %H:%M"))) %>% typ
e_convert(col_types = cols(start_frequency = col_integer())) %>% type_convert(col_types = cols(c
me_angle = col_integer())) %>% type_convert(col_types = cols(cme_speed = col_integer())) %>% typ
e_convert(col_types = cols(cme_width = col_integer())) %>% type_convert(col_types = cols(end_fre
quency = col_integer()))

nasa
```

```
## # 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 1997-05-14 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 1997-11-05 04:30:00 1997-11-04 06:10:00
##  9 1997-11-06 12:20:00 1997-11-07 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 <dbl>,
## #   end_frequency <dbl>, flare_location <chr>, flare_region <chr>,
## #   flare_classification <chr>, cme_angle <dbl>, cme_width <int>,
## #   cme_speed <dbl>, halo <lgl>, cme_width_limit <lgl>
```

ANALYSIS

Part 1

The replication can not be exact because the top 50 solar table in SpaceWeatherLive.com does have some entities that are missing in the NASA data. Also, for double digit flare_classifications there was not any additional numbers after the decimal, which caused some issues as well. You are able to get most the classifications, their start times, their regions, and their end time. The regions are slightly off, but this can be fixed by removing any leading “1”

```
options(digits = 9)

nasa_top50 <- nasa %>% separate(flare_classification, c("flare_letter", "flare_num"), sep=1)

#Due to the nature of the project and the way I attempted to classify and sort, as.numeric was g
iving a warning due to the NA's that were added prior. In order to prevent this, I supporessed t
he warnings around the numeric conversion. This does not cause any problems with the code due to
NA's being ignored during classification anyways
nasa_top50 <- nasa_top50 %>% mutate(flare_region = ifelse (str_detect(flare_region, "^[0-9]{5}$"
), substring(flare_region,2), flare_region)) %>% mutate(flare_num = suppressWarnings(as.numeric
(flare_num))) %>% mutate(flare_region = ifelse (str_detect(flare_region, "^[0]"), substring(flar
e_region,2), flare_region))


#More cleanup plus combination of flare_classification
nasa_top50 <- nasa_top50[with(nasa_top50, order(flare_letter, flare_num, decreasing = TRUE)),] %
>% unite(flare_classification, c("flare_letter", "flare_num"), sep = "", remove = TRUE) %>% slic
e(1:50)  %>% mutate(flare_classification = ifelse(str_detect(flare_classification, "^[A-Z][0-9]
{1,2}$"), paste(flare_classification, ".0", sep=""), flare_classification))

nasa_top50 <- as_tibble(nasa_top50)

nasa_top50
```

```
## # A tibble: 50 x 13
##    start_datetime      end_datetime        cme_datetime
##    <dttm>              <dttm>              <dttm>
##  1 2003-11-04 20:00:00 2003-11-04 23:59:00 2003-11-04 19:54:00
##  2 2001-04-02 22:05:00 2001-04-03 02:30:00 2001-04-02 22:06:00
##  3 2003-10-28 11:10:00 2003-10-29 23:59:00 2003-10-28 11:30:00
##  4 2001-04-15 14:05:00 2001-04-16 13:00:00 2001-04-15 14:06:00
##  5 2003-10-29 20:55:00 2003-10-29 23:59:00 2003-10-29 20:54:00
##  6 1997-11-06 12:20:00 1997-11-07 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 2003-11-03 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 <dbl>,
## #   end_frequency <dbl>, flare_location <chr>, flare_region <chr>,
## #   flare_classification <chr>, cme_angle <dbl>, cme_width <int>,
## #   cme_speed <dbl>, halo <lgl>, cme_width_limit <lgl>
```

Part 2

For my similarity calculation, I took the four similar attributes between NASA and SpaceLive. These were the region,

classification, start time, and end time. My similarity function was broken into four functions. Three to compute the similarity between the four variables, and one to bring them together and compute a number between 0 and 10. For the end and start time, I simply computed the difference, took the negative exponent of it and multiplied it by 10. A score of 10 meant the dates were 100% similar, so I felt it was appropriate to use a multiplier of 10.

For the class_similarity function, I knew that there was some missing data after the decimal points in the NASA dataset. I filled these with a 0, which felt like the best option due to many of them already being 0. With this in mind, I computed the similarity between the alphabetic class and the number following it seperately. This would at least allow some similarity between those with wrong data from cleaning the NASA data set. EAch of these two classifications would bring a number between 0 and 10, and was then dividied by two to keep with the “10 rule”

Due to already tidying data, calculating the region similarity was also simple. If they

matched, then 10, if not then 0.

These four similarity variables were added up, divided by 4.0, and then multiplied by 2 in order to get a number between 0 and 10. I came to the conclusion through observation that any entity with a similarity score less than 9 was not sufficiently similar, and therefore that was the threshold for determining whether or not there was a matching entitity.

```
#Computes similarity between dates
date_similarity <- function(d1, d2){

  d <- (as.integer(d1 - d2))^2

  exp(-d) *10.0

}

#Computes the class similarity
class_similarity <- function(c1, c2){

  c1class <- substring(c1, 1, 1)
  c1num <- as.numeric(substring(c1, 2, nchar(c1)))
  c2class <- substring(c2, 1, 1)
  c2num <- as.numeric(substring(c2, 2, nchar(c2)))

  class_rank <- ifelse(c1class == c2class, 10, 0)

  num_rank <- exp(-((c1num-c2num)^2))*10.0

  class_similarity <- (class_rank + num_rank)/2


}


#Computes the similarity between the regions
region_similarity <- function(r1, r2){

  r1 <- as.numeric(r1)
  r2 <- as.numeric(r2)

  ifelse(r1 == r2, 10 ,0)

}


#Computes a similarity between four similarity variables
flare_similarity <- function(sim1, sim2, sim3, sim4){

  ((sim1+sim2+sim3+sim4)/ 4.0)*2.0

}

flare_match <- function(E1, E2){

#Create df with pairwise combinations of row indices from both df's
index_df <- E1 %>% rowid_to_column(var = "rowid") %>%
  select(df1_id="rowid") %>%
  mutate(df2_id=NA) %>%
  bind_rows(E2 %>% rowid_to_column(var = "rowid") %>%
              select(df2_id="rowid") %>%
```

```
              mutate(df1_id = NA)) %>%
  tidyr::expand(df1_id, df2_id) %>%
  tidyr::drop_na()
index_df

#Join dataframe to populate it with attributes from both
similarity_df <- index_df %>%
  inner_join(E1 %>% rowid_to_column(var = "rowid") %>%
               select(rowid, flare_classification.E1=flare_classification,
                      start_datetime.E1=start_datetime, end_datetime.E1=end_datetime, flare_regi
on.E1=flare_region),
                      by=c(df1_id = "rowid")) %>%
  inner_join(E2 %>% rowid_to_column(var = "rowid") %>%
               select(rowid, flare_classification.E2=flare_classification,
                      start_datetime.E2=start_datetime, end_datetime.E2=end_datetime, flare_regi
on.E2=flare_region),
                      by=c(df2_id = "rowid"),
                      suffix=c(".ind", ".df2"))


#Compute Similarities of all relevant attributes
similarity_df <- similarity_df %>%
  mutate(start_date_sim = date_similarity(start_datetime.E1, start_datetime.E2)) %>% mutate(end_
date_sim = date_similarity(end_datetime.E1, end_datetime.E2)) %>%
  mutate(class_sim = class_similarity(flare_classification.E1, flare_classification.E2)) %>% mut
ate(region_sim = region_similarity(flare_region.E1, flare_region.E2)) %>%
  mutate(similarity = flare_similarity(start_date_sim, end_date_sim, class_sim, region_sim)) %>%
  select(df1_id, df2_id, similarity)

#Group and find the maximum similarities
similarity_df <- similarity_df %>%
  group_by(df1_id) %>%
  summarize(max_sim = max(similarity), df2_match_id=df2_id[which.max(similarity)])  %>% mutate(d
f2_match_id=ifelse(max_sim < 9, NA, df2_match_id)) %>% select(df1_id, df2_match_id)

colnames(similarity_df)[which(names(similarity_df) == "df1_id")] <- "rank"


similarity_df

}

#Merge the similartity df and solar_flare dataframe to get the matching entities

sim <- flare_match(solar_flare, nasa_top50)

solar_flare <- merge(solar_flare, sim)


solar_flare
```

```
##    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
## 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 df2_match_id
## 1   2003-11-04 20:06:00          486            1
```

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

PART 3 Q1

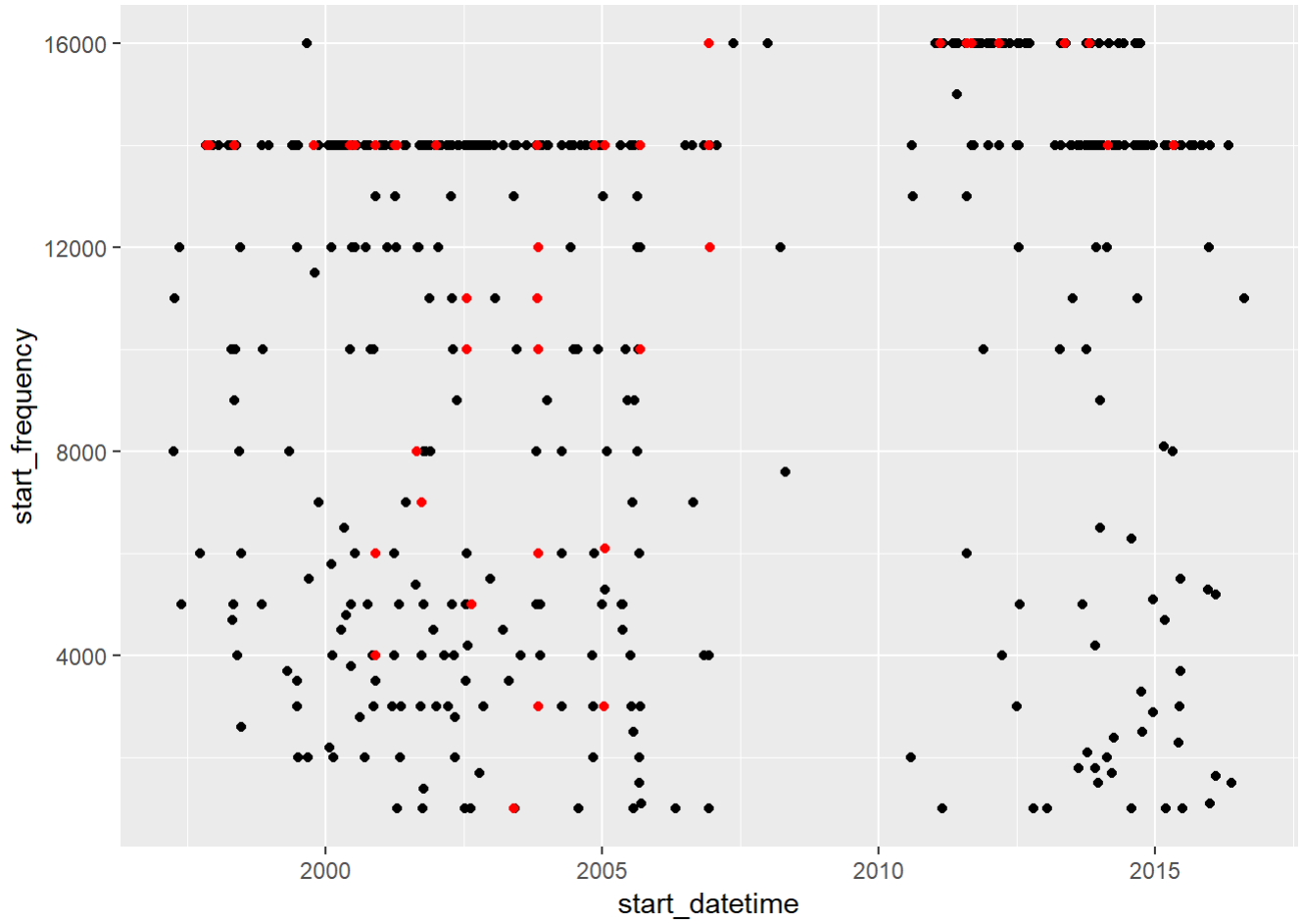
Intention: The intention of these plots is to show any variance of variables over time. Potentially to see how, as time progressed, solar flares have changed

Description: These four plots each show particular data points in relation to the time that these solar flares took place. The top 50 solar flare points are highlighted in red

Interpretation: Overall, I do not see much correlation. The points do not follow a trend in any way, including the top 50 points. I do not see any variation between top50 trends and regular plot trends as well. I did notice however that the cme_width and angle is fairly high in Top 50 solar flares

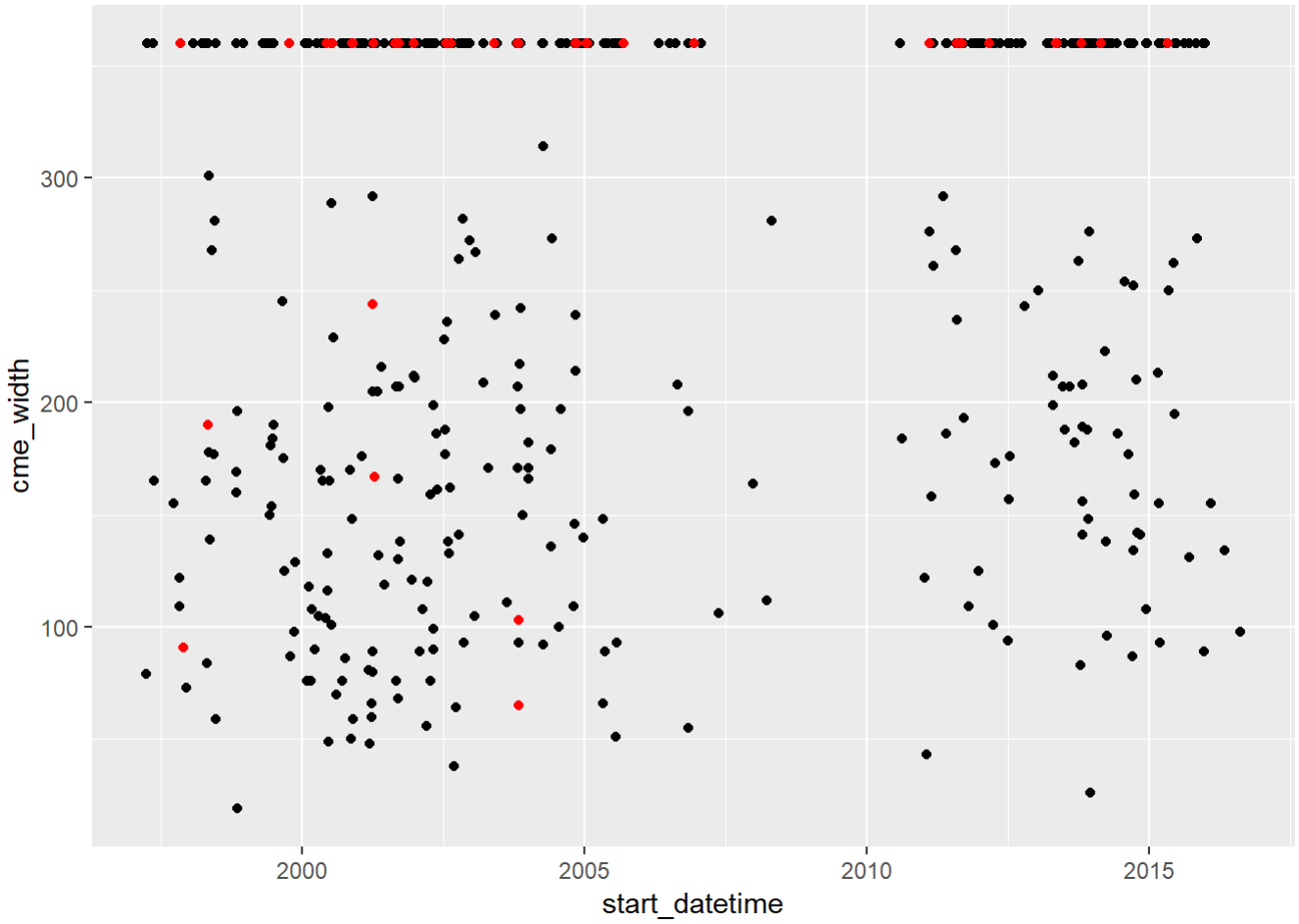
```
#Start frequency over time plot. Top 50 are highlighted in Red
start_freq_plot <- nasa %>% ggplot(mapping=aes(y=start_frequency, x= start_datetime), na.rm=TRUE
) + geom_point() + geom_point(data = nasa_top50, aes(y = start_frequency, x= start_datetime), co
lor = "red")

suppressWarnings(print(start_freq_plot))
```



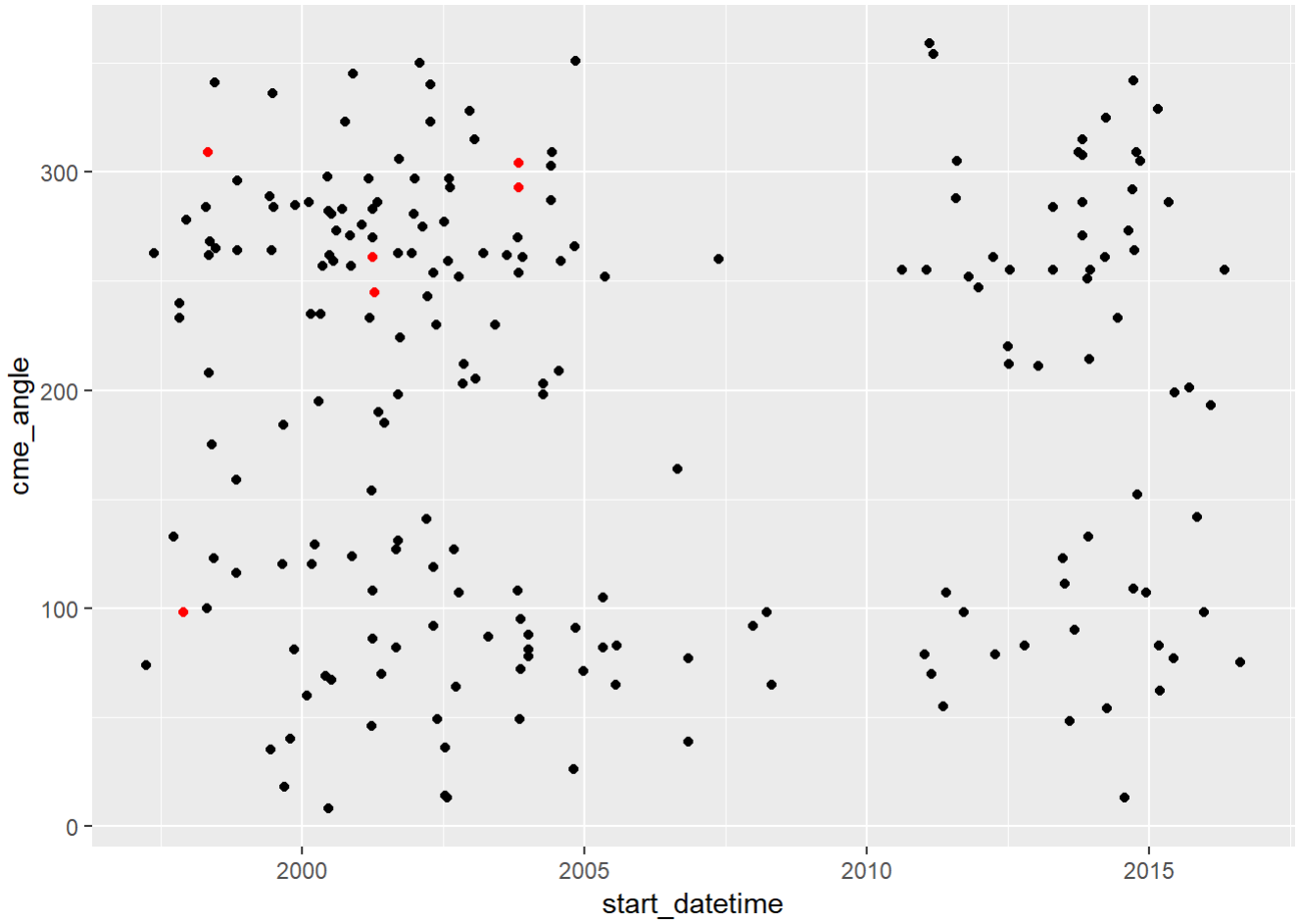

```
#Flare width over time plot. Top 50 are highlighted in Red
flare_width_plot <- nasa %>% ggplot(mapping=aes(y=cme_width, x= start_datetime), na.rm=TRUE) + g
eom_point() + geom_point(data = nasa_top50, aes(y = cme_width, x= start_datetime), color = "red"
)

suppressWarnings(print(flare_width_plot))
```



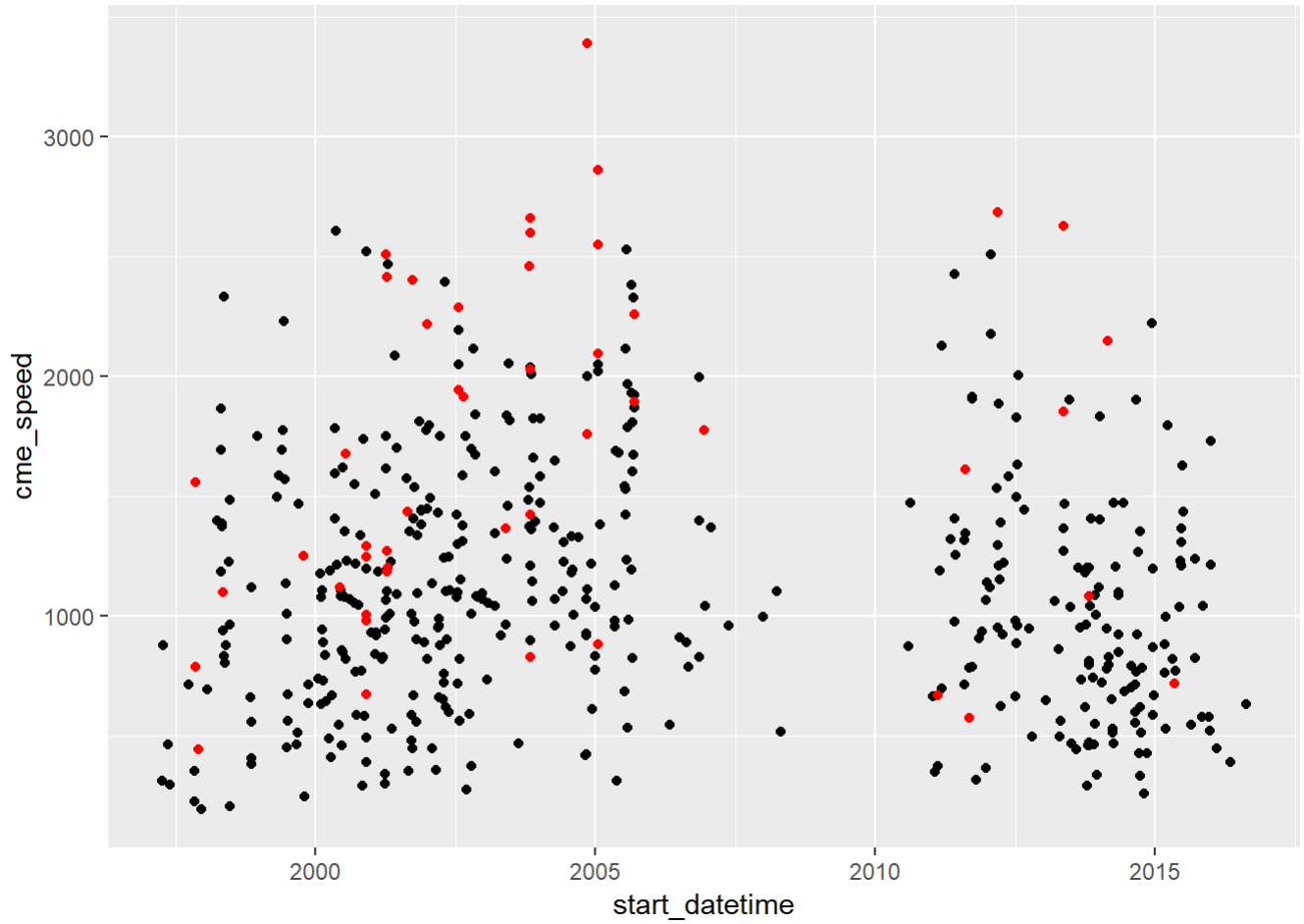

```
#Flare angle over time plot. Top 50 are highlighted in Red
flare_angle_plot <- nasa %>% ggplot(mapping=aes(y=cme_angle, x= start_datetime), na.rm=TRUE) + g
eom_point() + geom_point(data = nasa_top50, aes(y = cme_angle, x= start_datetime), color = "red"
)

suppressWarnings(print(flare_angle_plot))
```




```
#Flare speed over time plot. Top 50 are highlighted in Red
flare_speed_plot <- nasa %>% ggplot(mapping=aes(y=cme_speed, x= start_datetime), na.rm=TRUE) + g
eom_point() + geom_point(data = nasa_top50, aes(y = cme_speed, x= start_datetime), color = "red"
)

suppressWarnings(print(flare_speed_plot))
```



PART 3 Q2

The intention of this plot is to show whether or not strong solar flares tend to have halo's or not. It will show the variation between the proportion of strong solar flares with halos and weaker solar flares with halos

Description: This plot shows the number of halos that are or are not present in the top 50 solar flares versus those not in the top 50. The portion highlighted in blue are the strong top

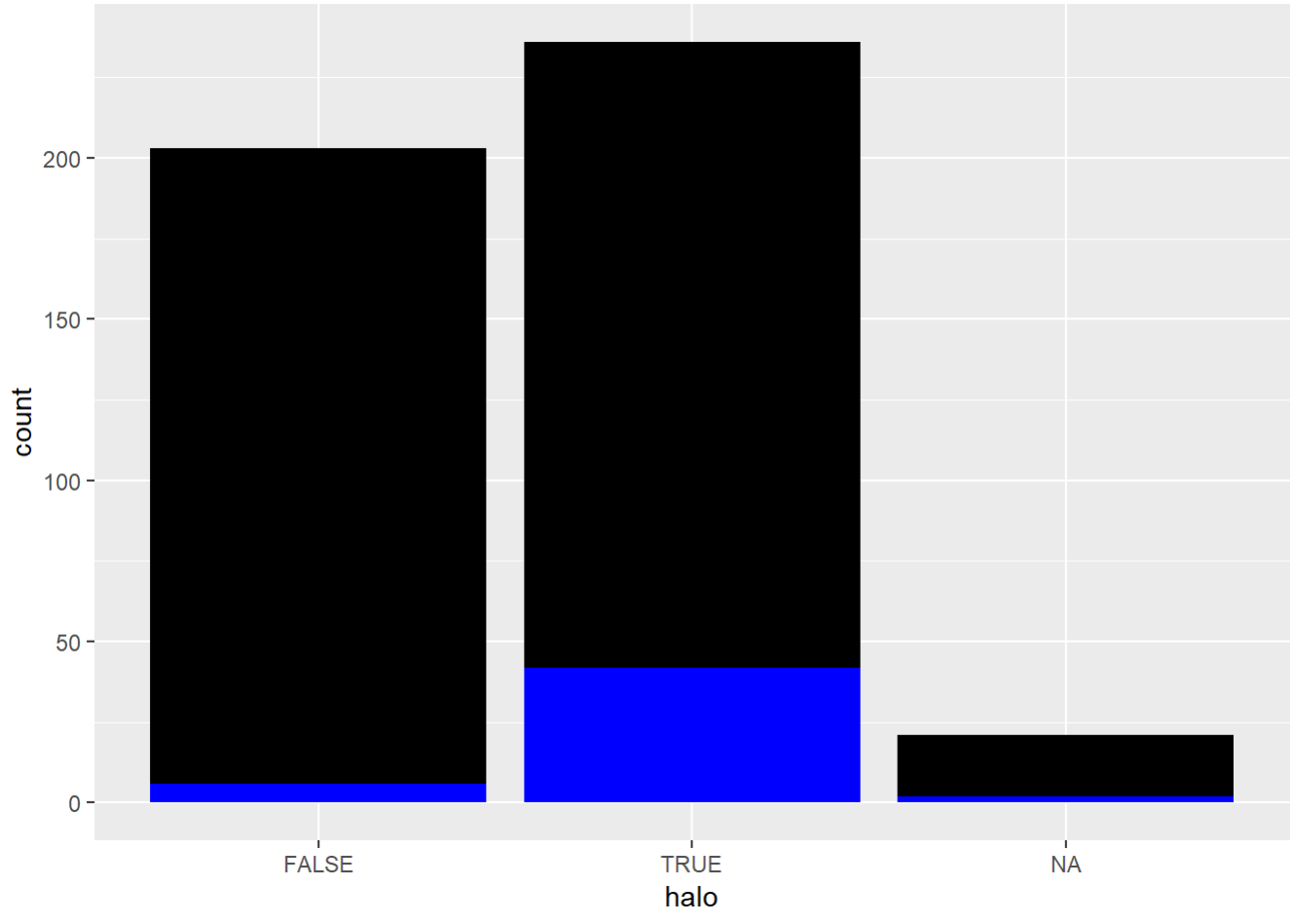
50 solar flares while the portion in black is everything else

Interpretation: According to the plot, one can see that the number of halos and strength of solar flare is correlated. The number of true halo's in comparison to false in the top 50 has a much higher variation than difference between the weaker solar flare data

```
#Create a dataframe without the top50 data
anti_frame <- anti_join(nasa, nasa_top50, by = "flare_classification")

q2_plot <- anti_frame %>% ggplot(mapping=aes(halo), na.rm=TRUE)+geom_bar(fill = "black") + geom_
bar(nasa_top50, mapping= aes(halo), fill = "blue")

suppressWarnings(print(q2_plot))
```



Part 3 Q3

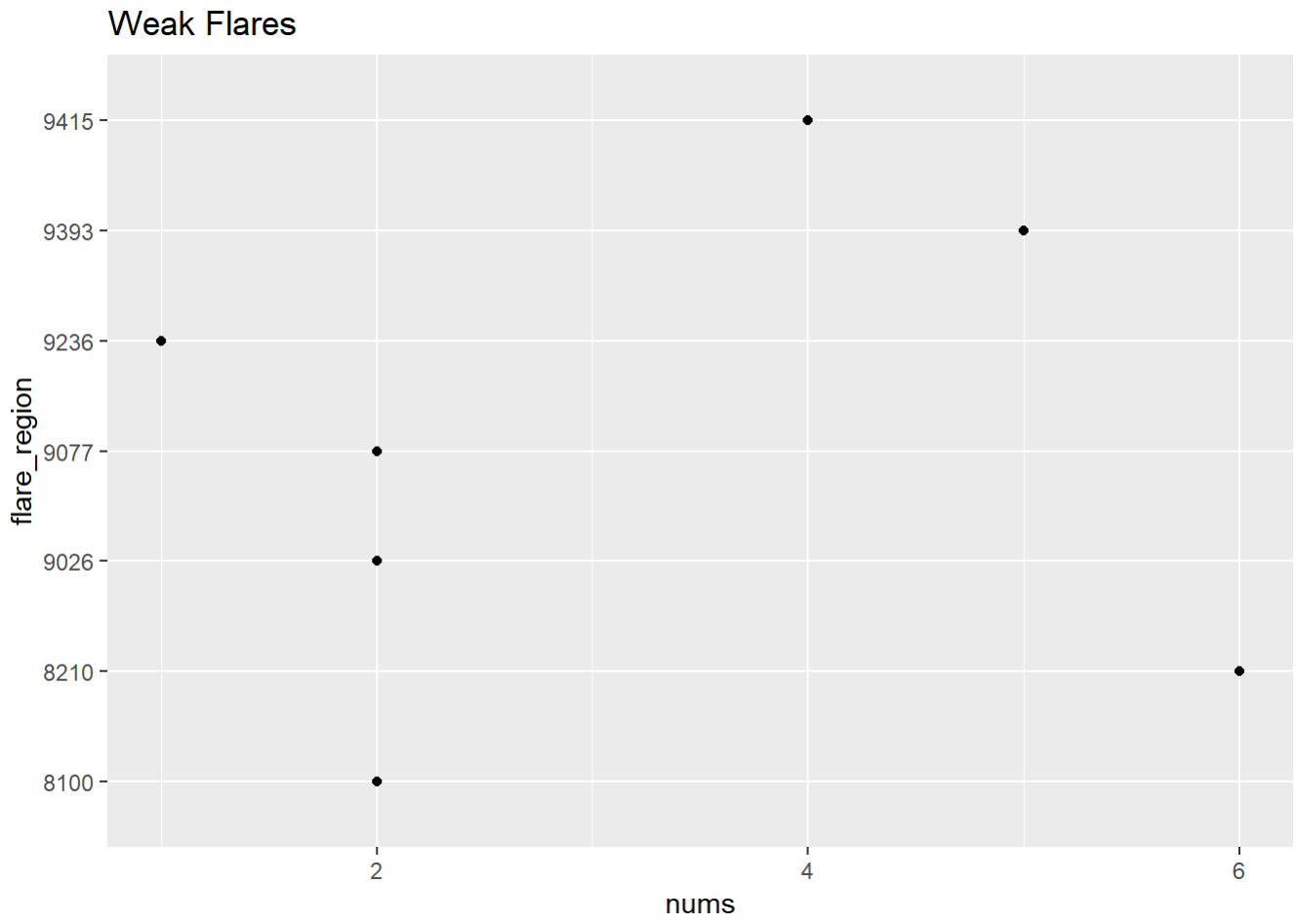
Intention: The intention of this plot is to show where strong solar flares cluster most

Description: The first plot shows the number of flares of the whole ataset that are in the flare regions present in the top 50. The second plot groups the number of occurrences a solar flare occurs within a solar region from the top 50 dataset

Interpretation: Judging by variation in number of flare_regions present in the top 50 plot, it seems as though the top 50 flares are mostly in different regions. However, it seems as though some specific regions are more prone to strong solar flares. Region 9236 has only one flare but in the top 50 it contains the most with five flares. In contrast, it seems as though regions with low amounts of strong flares end up having more weak flares to replace. Flare region 9393 and 8210 had low top 50 flare presence but high weak flare presence

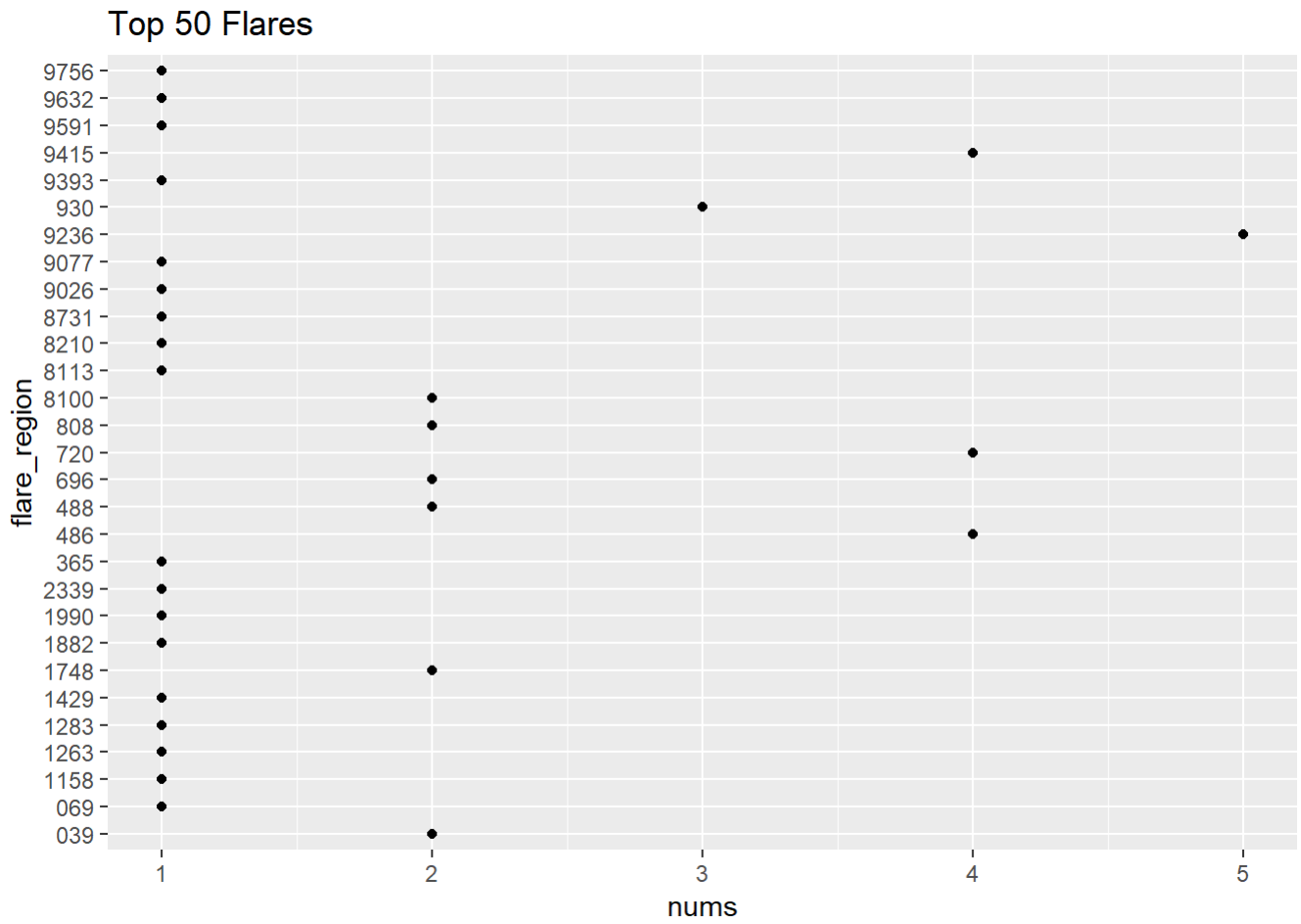
```
region_reg_plot <- anti_frame %>% filter(flare_region %in% nasa_top50$flare_region) %>% group_by
(flare_region) %>% summarize(nums = n()) %>% ggplot(mapping=aes(y=flare_region, x=nums)) + geom_
point() + ggtitle("Weak Flares")

suppressWarnings(print(region_reg_plot))
```




```
region_top50_plot <- nasa_top50 %>% group_by(flare_region) %>% summarize(nums = n()) %>% ggplot
(mapping=aes(y=flare_region, x=nums)) + geom_point() + ggtitle("Top 50 Flares")

suppressWarnings(print(region_top50_plot))
```


Top 50 Flares
flare_region
9756
9632
9591
9415
9393
930
9236
9077
9026
8731
8210
8113
8100
808
720
696
488
486
365
2339
1990
1882
1748
1429
1283
1263
1158
069
039
1
2
3
4
5
nums