Data Scraping and Cleaning

Zackary Frazier

3/9/2020

Part 1: Data scraping and preparation

Step 1: Scrape your competitor's data (10 pts)

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

# retrieve the table
weather_tab <- weather_url %>%
  read_html() %>%
  html_node('table') %>%
  html_table()

# rename columns
colnames(weather_tab) <- c('rank', 'flare_classification', 'date', 'flare_region',
                           'start_time', 'maximum_time', 'end_time', 'movie')

head(weather_tab)
```

```
##   rank flare_classification       date flare_region start_time maximum_time
## 1    1                 X28+ 2003/11/04          486      19:29        19:53
## 2    2                 X20+ 2001/04/02         9393      21:32        21:51
## 3    3               X17.2+ 2003/10/28          486      09:51        11:10
## 4    4                 X17+ 2005/09/07          808      17:17        17:40
## 5    5                X14.4 2001/04/15         9415      13:19        13:50
## 6    6                  X10 2003/10/29          486      20:37        20:49
##   end_time           movie
## 1    20:06 MovieView archive
## 2    22:03 MovieView archive
## 3    11:24 MovieView archive
## 4    18:03 MovieView archive
## 5    13:55 MovieView archive
## 6    21:01 MovieView archive
```

Step 2: Tidy the top 50 solar flare data (10 pts)

```r
# drop last column
weather_tab <- select(weather_tab, -'movie')

# combine the date and times
weather_tab <- weather_tab %>%
```

```r
  mutate(start_time = paste(date, start_time, sep = ' ')) %>%
  mutate(maximum_time = paste(date, maximum_time, sep = ' ')) %>%
  mutate(end_time = paste(date, end_time, sep = ' '))

# convert the combined columns into datetime objects
weather_tab <- weather_tab %>%
  mutate(start_time = make_datetime(
    year = strtoi(str_sub(start_time, start = 1L, end = 4L), base = 10L),
    month = strtoi(str_sub(start_time, 6L, 7L), base = 10L),
    day = strtoi(str_sub(start_time, 9L, 10L), base = 10L),
    hour = strtoi(str_sub(start_time, 12L, 13L), base = 10L),
    min = strtoi(str_sub(start_time, 15L, 16L), base = 10L),
    sec = 0)) %>%
  mutate(end_time = make_datetime(
    year = strtoi(str_sub(end_time, start = 1L, end = 4L), base = 10L),
    month = strtoi(str_sub(end_time, 6L, 7L), base = 10L),
    day = strtoi(str_sub(end_time, 9L, 10L), base = 10L),
    hour = strtoi(str_sub(end_time, 12L, 13L), base = 10L),
    min = strtoi(str_sub(end_time, 15L, 16L), base = 10L),
    sec = 0)) %>%
  mutate(maximum_time = make_datetime(
    year = strtoi(str_sub(maximum_time, start = 1L, end = 4L), base = 10L),
    month = strtoi(str_sub(maximum_time, 6L, 7L), base = 10L),
    day = strtoi(str_sub(maximum_time, 9L, 10L), base = 10L),
    hour = strtoi(str_sub(maximum_time, 12L, 13L), base = 10L),
    min = strtoi(str_sub(maximum_time, 15L, 16L), base = 10L),
    sec = 0)) %>%
  select(rank, flare_classification, flare_region, start_datetime = start_time,
         maximum_datetime = maximum_time, end_datetime = end_time, -date)

head(weather_tab)
```

```
##   rank flare_classification flare_region      start_datetime
## 1    1                 X28+          486 2003-11-04 19:29:00
## 2    2                 X20+         9393 2001-04-02 21:32:00
## 3    3               X17.2+          486 2003-10-28 09:51:00
## 4    4                 X17+          808 2005-09-07 17:17:00
## 5    5                X14.4         9415 2001-04-15 13:19:00
## 6    6                  X10          486 2003-10-29 20:37:00
##      maximum_datetime        end_datetime
## 1 2003-11-04 19:53:00 2003-11-04 20:06:00
## 2 2001-04-02 21:51:00 2001-04-02 22:03:00
## 3 2003-10-28 11:10:00 2003-10-28 11:24:00
## 4 2005-09-07 17:40:00 2005-09-07 18:03:00
## 5 2001-04-15 13:50:00 2001-04-15 13:55:00
## 6 2003-10-29 20:49:00 2003-10-29 21:01:00
```

Step 3. Scrape the NASA data (15 pts)

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

# convert the HTML into a table-like structure
```

```
NASA_tab <- NASA_url %>%
  read_html() %>%
  html_node('pre') %>%
  html_text() %>%
  str_sub(802, -99) %>%
  strsplit(split = '\n')

# insert the vector of strings into a dataframe
NASA_tab <- data.frame(NASA_tab)
colnames(NASA_tab) <- c('raw_data')

# parse the raw data into separate columns
NASA_tab[, 'start_date'] <- NASA_tab$raw_data %>%
  substr(1, 10) %>%
  str_trim()
NASA_tab[, 'start_time'] <- NASA_tab$raw_data %>%
  substr(12, 17) %>%
  str_trim()
NASA_tab[, 'end_date'] <- NASA_tab$raw_data %>%
  substr(18, 22)
NASA_tab[, 'end_time'] <- NASA_tab$raw_data %>%
  substr(24, 29)
NASA_tab[, 'start_freq'] <- NASA_tab$raw_data %>%
  substr(30, 35) %>%
  str_trim()
NASA_tab[, 'end_freq'] <- NASA_tab$raw_data %>%
  substr(36, 41) %>%
  str_trim()
NASA_tab[, 'location'] <- NASA_tab$raw_data %>%
  substr(42, 49) %>%
  str_trim()
NASA_tab[, 'NOAA'] <- NASA_tab$raw_data %>%
  substr(50, 55) %>%
  str_trim()
NASA_tab[, 'imp'] <- NASA_tab$raw_data %>%
  substr(56, 62) %>%
  str_trim()
NASA_tab[, 'CME_date'] <- NASA_tab$raw_data %>%
  substr(63, 68) %>%
  str_trim()
NASA_tab[, 'CME_time'] <- NASA_tab$raw_data %>%
  substr(69, 74)
NASA_tab[, 'CME_angle'] <- NASA_tab$raw_data %>%
  substr(76, 80) %>%
  str_trim()
NASA_tab[, 'CME_width'] <- NASA_tab$raw_data %>%
  substr(81, 85) %>%
  str_trim()
NASA_tab[, 'CME_speed'] <- NASA_tab$raw_data %>%
  substr(86, 90) %>%
  str_trim()
head(NASA_tab)
```

```
##                                                                                                    raw_data
```

```
## 1 1997/04/01 14:00 04/01 14:15  8000  4000   S25E16  8026 M1.3   04/01 15:18   74   79  312   PHTX
## 2 1997/04/07 14:30 04/07 17:30 11000  1000   S28E19  8027 C6.8   04/07 14:27 Halo  360  878   PHTX
## 3 1997/05/12 05:15 05/14 16:00 12000    80   N21W08  8038 C1.3   05/12 05:30 Halo  360  464   PHTX
## 4 1997/05/21 20:20 05/21 22:00  5000   500   N05W12  8040 M1.3   05/21 21:00  263  165  296   PHTX
## 5 1997/09/23 21:53 09/23 22:16  6000  2000   S29E25  8088 C1.4   09/23 22:02  133  155  712   PHTX
## 6 1997/11/03 05:15 11/03 12:00 14000   250   S20W13  8100 C8.6   11/03 05:28  240  109  227   PHTX
##   start_date start_time end_date end_time start_freq end_freq location NOAA
## 1 1997/04/01      14:00    04/01    14:15       8000     4000   S25E16 8026
## 2 1997/04/07      14:30    04/07    17:30      11000     1000   S28E19 8027
## 3 1997/05/12      05:15    05/14    16:00      12000       80   N21W08 8038
## 4 1997/05/21      20:20    05/21    22:00       5000      500   N05W12 8040
## 5 1997/09/23      21:53    09/23    22:16       6000     2000   S29E25 8088
## 6 1997/11/03      05:15    11/03    12:00      14000      250   S20W13 8100
##    imp CME_date CME_time CME_angle CME_width CME_speed
## 1 M1.3    04/01    15:18        74        79       312
## 2 C6.8    04/07    14:27      Halo       360       878
## 3 C1.3    05/12    05:30      Halo       360       464
## 4 M1.3    05/21    21:00       263       165       296
## 5 C1.4    09/23    22:02       133       155       712
## 6 C8.6    11/03    05:28       240       109       227
```

Step 4: Tidy the NASA the table (15 pts)

```
# Remove empty values
NASA_tab$start_freq[startsWith(NASA_tab$start_freq, '?')] <- NA
NASA_tab$end_freq[startsWith(NASA_tab$end_freq, '?')] <- NA
NASA_tab$location[grepl('.*[bB][aA][cC][kK].*', NASA_tab$location)] <- NA
NASA_tab$NOAA[startsWith(NASA_tab$NOAA, '-')] <- NA
NASA_tab$imp[startsWith(NASA_tab$imp, '-')
             | grepl('.*[fF][iI][lL][aA].*', NASA_tab$imp)] <- NA
NASA_tab$CME_date[startsWith(NASA_tab$CME_date, '-')] <- NA
NASA_tab$CME_time[startsWith(NASA_tab$CME_time, '-')] <- NA
NASA_tab$CME_angle[startsWith(NASA_tab$CME_angle, '-')] <- NA
NASA_tab$CME_width[startsWith(NASA_tab$CME_width, '-')] <- NA
NASA_tab$CME_speed[startsWith(NASA_tab$CME_speed, '-')] <- NA

# Tidying the data
NASA_tab <- NASA_tab %>%
  mutate(start_time = paste(start_date, start_time, sep = ' ')) %>%
  mutate(end_time = paste(paste(substr(start_date, 1, 4), end_date, sep='/'),
                          end_time, sep = ' ')) %>%
  mutate(CME_time = paste(paste(substr(start_date, 1, 4), CME_date, sep='/'),
                          CME_time, sep = ' ')) %>%
  select(-raw_data, -start_date, -end_date, -CME_date)

# convert the combined columns into datetime objects
NASA_tab <- NASA_tab %>%
  mutate(start_time = make_datetime(
    year = strtoi(substr(start_time, 1, 4), base = 10),
    month = strtoi(substr(start_time, 6, 7), base=10),
    day = strtoi(substr(start_time, 9, 10), base=10),
    hour = strtoi(substr(start_time, 12, 13), base=10),
    min = strtoi(substr(start_time, 15, 16), base=10),
```

```r
    sec = 0)) %>%
  mutate(end_time = make_datetime(
    year = strtoi(substr(end_time, 1, 4), base=10),
    month = strtoi(substr(end_time, 6, 7), base=10),
    day = strtoi(substr(end_time, 9, 10), base=10),
    hour = strtoi(substr(end_time, 12, 13), base=10),
    min = strtoi(substr(end_time, 15, 16), base=10),
    sec = 0)) %>%
  mutate(CME_time = make_datetime(
    year = strtoi(substr(CME_time, 1, 4), base=10),
    month = strtoi(substr(CME_time, 6, 7), base=10),
    day = strtoi(substr(CME_time, 9, 10), base=10),
    hour = strtoi(substr(CME_time, 13, 14), base=10),
    min = strtoi(substr(CME_time, 16, 17), base=10),
    sec = 0)) %>%
  select(start_datetime = start_time, end_datetime = end_time, start_freq,
         end_freq, location, NOAA, imp, CME_datetime = CME_time,
         CME_angle, CME_width, CME_speed)

# removing Halo from CME_angle and > symbols from CME_width
# creating lower_bound and Halo columns
NASA_tab <- NASA_tab %>%
  mutate(Halo = (CME_angle == 'Halo'),
         lower_bound = startsWith(CME_width, '>')) %>%
  mutate(CME_width = ifelse(startsWith(CME_width, ">"),
                            str_sub(CME_width, 2, -1), CME_width)) %>%
  mutate(CME_angle = ifelse(CME_angle == 'Halo', NA, CME_angle))

# converting columns to appropriate datatypes
NASA_tab <- NASA_tab %>%
  mutate(start_freq = ifelse(is.na(start_freq), NA,
                             strtoi(start_freq, base=10))) %>%
  mutate(end_freq = ifelse(is.na(end_freq), NA, strtoi(end_freq, base=10))) %>%
  mutate(NOAA = ifelse(is.na(NOAA), NA, strtoi(NOAA, base=10))) %>%
  mutate(CME_angle = ifelse(is.na(CME_angle), NA,
                            strtoi(CME_angle, base=10))) %>%
  mutate(CME_width = ifelse(is.na(CME_width), NA, strtoi(CME_width))) %>%
  mutate(CME_speed = ifelse(is.na(CME_speed), NA, strtoi(CME_speed)))

head(NASA_tab)
```

```
##        start_datetime        end_datetime start_freq end_freq location NOAA
## 1 1997-04-01 14:00:00 1997-04-01 14:15:00       8000     4000   S25E16 8026
## 2 1997-04-07 14:30:00 1997-04-07 17:30:00      11000     1000   S28E19 8027
## 3 1997-05-12 05:15:00 1997-05-14 16:00:00      12000       80   N21W08 8038
## 4 1997-05-21 20:20:00 1997-05-21 22:00:00       5000      500   N05W12 8040
## 5 1997-09-23 21:53:00 1997-09-23 22:16:00       6000     2000   S29E25 8088
## 6 1997-11-03 05:15:00 1997-11-03 12:00:00      14000      250   S20W13 8100
##    imp        CME_datetime CME_angle CME_width CME_speed  Halo lower_bound
## 1 M1.3 1997-04-01 15:18:00        74        79       312 FALSE       FALSE
## 2 C6.8 1997-04-07 14:27:00        NA       360       878  TRUE       FALSE
## 3 C1.3 1997-05-12 05:30:00        NA       360       464  TRUE       FALSE
## 4 M1.3 1997-05-21 21:00:00       263       165       296 FALSE       FALSE
## 5 C1.4 1997-09-23 22:02:00       133       155       712 FALSE       FALSE
```

```
## 6 C8.6 1997-11-03 05:28:00       240       109       227 FALSE       FALSE
```

Part 2: Analysis

Question 1: Replication (10 pts)

Can you replicate the top 50 solar flare table in SpaceWeatherLive.com exactly using the data obtained from NASA? That is, if you get the top 50 solar flares from the NASA table based on their classification (e.g., X28 is the highest), do you get data for the same 50 solar flare events in the SpaceWeatherLive page? If not, why not?

```
# Replication

options(digits = 5)
NASA_top50 <- NASA_tab %>%
  filter(startsWith(imp, 'X')) %>%
  mutate(rank = as.double(str_sub(imp, start = 2, end = nchar(imp)))) %>%
  arrange(desc(rank)) %>%
  head(50) %>%
  mutate(rank = row_number()) %>%
  select(rank, flare_classification=imp, flare_region = NOAA, start_datetime,
         maximum_datetime=CME_datetime, end_datetime)

head(NASA_top50)
```

```
##   rank flare_classification flare_region      start_datetime
## 1    1                 X28.        10486 2003-11-04 20:00:00
## 2    2                 X20.         9393 2001-04-02 22:05:00
## 3    3                 X17.        10486 2003-10-28 11:10:00
## 4    4                 X14.         9415 2001-04-15 14:05:00
## 5    5                 X10.        10486 2003-10-29 20:55:00
## 6    6                 X9.4         8100 1997-11-06 12:20:00
##      maximum_datetime        end_datetime
## 1 2003-11-04 19:54:00 2003-11-05 00:00:00
## 2 2001-04-02 22:06:00 2001-04-03 02:30:00
## 3 2003-10-28 11:30:00 2003-10-30 00:00:00
## 4 2001-04-15 14:06:00 2001-04-16 13:00:00
## 5 2003-10-29 20:54:00 2003-10-30 00:00:00
## 6 1997-11-06 12:10:00 1997-11-07 08:30:00
```

You cannot get the same table with the NASA data, although you can get a similar one. It appears that the weather table got it's information from a different source than the NASA table. Also the NASA table does not record the maximum time, only the CME time. It does appear however that some of the solar flares in the NASA table are the same ones as the ones in the weather table.

Question 2: Entity Resolution (15 pts)

There are three similarity functions defined in flare_similarity.

- s_classification compares the classifications of flares based on how far away they are from each other numerically. If they have different classification IDs a zero is returned. I consider it highly unlikely that two flares with different classifications would be the same flare.

- s_region compares the categorical regions of each flare. Either these flares were in the same region, or they were not. If so it returns 1, if not it returns 0. Does not account for the fact that some of the regions have very similar region codes, such as the NOAA being 10486 an the region in the spaceweather table being 486 as there is no way to know whether that was a typo or if these tables are using slightly different regional categories.
- s_datetime measures the differences between datetimes and returns an exponential value such that the smaller the difference is, the larger the value that will be returned.

The threshold is set to 1.01 so that if something is in the same region it won't automatically be resolved to an entity in the NASA data.

```
flare_similarity <- function(E1, E2) {
  # Similarity Functions

  s_classification <- function(e1, e2) {
    if(substr(e1, 1, 1) == substr(e2, 1, 1)) {
      s1 <- ifelse(endsWith(e1, "+"), as.double(str_sub(e1, 2, -2)),
                   as.double(str_sub(e1, 2, -1)))
      s2 <- ifelse(endsWith(e2, "+"), as.double(str_sub(e2, 2, -2)),
                   as.double(str_sub(e2, 2, -1)))
      exp( -((s1 - s2)^2))
    } else 0
  }

  s_region <- function(e1, e2) {
    ifelse(e1 == e2, 1, 0)
  }

  s_datetime <- function(e1, e2) {
    exp( -(as.double(e1 - e2)^2))
  }

  # building the matrix
  m <- matrix(nrow = nrow(E1), ncol = nrow(E2))

  for(i in 1:nrow(E1)) {
    for(j in 1:nrow(E2)) {
      if(!(is.na(E1$flare_classification[i]) | is.na(E2$imp[j])))
        m[i, j] <- s_classification(E1$flare_classification[i], E2$imp[j])
      if(!(is.na(E1$flare_region[i]) | is.na(E2$NOAA[j])))
        m[i, j] <- m[i, j] + s_region(E1$flare_region[i], E2$NOAA[j])
      if(!(is.na(E1$start_datetime[i]) | is.na(E2$start_datetime[j])))
        m[i, j] <- m[i, j] + s_datetime(E1$start_datetime[i],
                                        E2$start_datetime[j])
      if(!(is.na(E1$maximum_datetime[i]) | is.na(E2$CME_datetime[j])))
        m[i, j] <- m[i, j] + s_datetime(E1$maximum_datetime[i],
                                        E2$CME_datetime[j])
      if(!(is.na(E1$end_datetime[i]) | is.na(E2$end_datetime[j])))
        m[i, j] <- m[i, j] + s_datetime(E1$end_datetime[i], E2$end_datetime[j])
    }
  }
  m
}
```

```r
# returns a vector of the index in E2 that e1 in E1 corresponds to
flare_match <- function(E1, E2) {
  sim_matrix <- flare_similarity(E1, E2)
  matches <- vector()
  index = vector(length=nrow(sim_matrix))

  for(i in 1:nrow(sim_matrix)) {
    matches <- append(matches, max(sim_matrix[i,], na.rm=TRUE))
    index[i] <- ifelse(matches[i] < 1.01, NA,
                       which(sim_matrix[i,] == max(sim_matrix[i,], na.rm=TRUE)))
  }
  index
}

weather_tab[, 'index'] <- flare_match(weather_tab, NASA_tab)

head(weather_tab)
```

```
##   rank flare_classification flare_region      start_datetime
## 1    1                 X28+          486 2003-11-04 19:29:00
## 2    2                 X20+         9393 2001-04-02 21:32:00
## 3    3               X17.2+          486 2003-10-28 09:51:00
## 4    4                 X17+          808 2005-09-07 17:17:00
## 5    5                X14.4         9415 2001-04-15 13:19:00
## 6    6                  X10          486 2003-10-29 20:37:00
##      maximum_datetime        end_datetime index
## 1 2003-11-04 19:53:00 2003-11-04 20:06:00   241
## 2 2001-04-02 21:51:00 2001-04-02 22:03:00   118
## 3 2003-10-28 11:10:00 2003-10-28 11:24:00   234
## 4 2005-09-07 17:40:00 2005-09-07 18:03:00    NA
## 5 2001-04-15 13:50:00 2001-04-15 13:55:00   127
## 6 2003-10-29 20:49:00 2003-10-29 21:01:00    NA
```

Question 3: Analysis (10 pts)

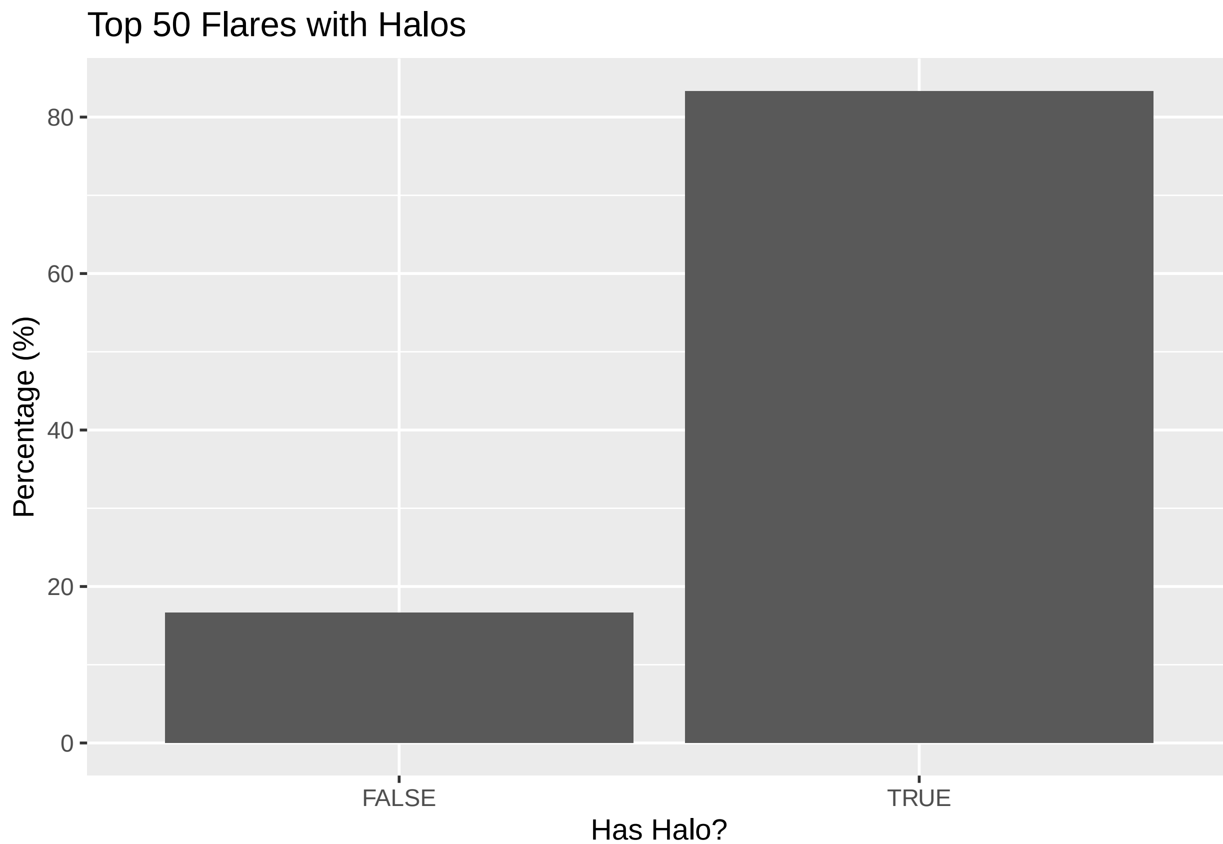
This plot analyzes the proportion of halo CMEs to non-halo CMEs based on the resolvable flares from the top 50 table. It seems that there is some sort of correlation between being a top 50 flare and having a halo CME.

```r
graph_tab <- filter(weather_tab, !is.na(index))

for( i in 1:nrow(graph_tab)) {
  graph_tab[i, 'Halo'] <- NASA_tab$Halo[graph_tab$index[i]]
}

valid_data <- length(which(!is.na(graph_tab$Halo)))

graph_tab %>%
  group_by(Halo) %>%
  summarize(true = (sum(Halo) / valid_data) * 100,
            false = (sum(!Halo) / valid_data) * 100) %>%
  ggplot(aes(x = Halo, y = true + false)) +
  geom_bar(stat = "identity") + ggtitle("Top 50 Flares with Halos") +
  labs(x = "Has Halo?", y = "Percentage (%)")
```


Top 50 Flares with Halos
80
60
40
20
0
Percentage (%)
FALSE
TRUE
Has Halo?