Project

Hsiu Yuan Yang

Fall 2022

US Storm / Disaster Analysis

Weather data is closely related to our lives. People would check on the current temperature and the raining probability when they go out, and they would monitor the latest storm / atmospheric event information in order to get prepared.

While I come from a country which has a lot of weather disasters, I was curious about what the natural disasters in US are like. I would like to know more about the disaster types, the occurring patterns (if any), etc. Hence, I decided to choose this topic as my R project.

The data sets used in this project are from the Storm Events Database owned by National Centers for Environmental Information - National Oceanic and Atmospheric Administration (NOAA). I will be using several data sets from NOAA, i.e. the storm / event details data sets for 2010 to 2022. The datasets are downloaded from https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/ (https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/), and detailed documentation about the fields / columns can be found on https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/Storm-Data-Bulk-csv-Format.pdf (https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/Storm-Data-Bulk-csv-Format.pdf).

The NOAA data sets are categorized by year, i.e. each csv file downloaded from its website is only for a specific year. While I planned to use the data set for a whole year cycle, it is noted that the data for the current 2022 cycle is not yet complete (with data only up to June), therefore I have manually created a data set to filter out information from July 2021 to June 2022. Besides, since I would like to compare if there are differences between the past (choosing 10 years before as a target) and the present, I also manually created a data set to filter out information from July 2011 to June 2012. In addition, to do future time series predictions, I also manually consolidated the storm / event data from January 2010 to the latest June 2022.

In short, 3 sets of data are in use for this project:

1. Storm / event details for the current cycle (i.e. Jul 2021 to Jun 2022)
2. Storm / event details for the past cycle (i.e. Jul 2011 to Jun 2012)
3. Storm / event details consolidated (i.e. Jan 2010 to Jun 2022)

Below outlines the details of the storm / event data sets (using the current cycle (i.e. Jul 2021 to Jun 2022) as an example):

```r
# data set for storm / event details of the current cycle (i.e. Jul 2021 to Jun 2022)
storm.current.raw <- read.csv("C:/Users/Yang Hsiu Yuan/Desktop/CMU/2022 Fall semester/94842 Progra
        mming R for Analytics/Project/StormEvents_details_current.csv", stringsAsFactors=TRUE)


# data set for storm / event details of the past cycle (i.e. Jul 2011 to Jun 2012)
storm.past.raw <- read.csv("C:/Users/Yang Hsiu Yuan/Desktop/CMU/2022 Fall semester/94842 Programmi
        ng R for Analytics/Project/StormEvents_details_previous.csv", stringsAsFactors=TRUE)


# data set for storm / event details consolidated (i.e. Jan 2010 to Jun 2022)
storm.consolidated.raw <- read.csv("C:/Users/Yang Hsiu Yuan/Desktop/CMU/2022 Fall semester/94842 P
        rogramming R for Analytics/Project/StormEvents_details_consolidated.csv", stringsAsFactors
        =TRUE)


# illustrate data set (using the current cycle one as an example)
colnames(storm.current.raw)
```

```
##  [1] "BEGIN_YEARMONTH"    "BEGIN_DAY"          "BEGIN_TIME"
##  [4] "END_YEARMONTH"      "END_DAY"            "END_TIME"
##  [7] "EPISODE_ID"         "EVENT_ID"           "STATE"
## [10] "STATE_FIPS"         "YEAR"               "MONTH_NAME"
## [13] "EVENT_TYPE"         "CZ_TYPE"            "CZ_FIPS"
## [16] "CZ_NAME"            "WFO"                "BEGIN_DATE_TIME"
## [19] "CZ_TIMEZONE"        "END_DATE_TIME"      "INJURIES_DIRECT"
## [22] "INJURIES_INDIRECT"  "DEATHS_DIRECT"      "DEATHS_INDIRECT"
## [25] "DAMAGE_PROPERTY"    "DAMAGE_CROPS"       "SOURCE"
## [28] "MAGNITUDE"          "MAGNITUDE_TYPE"     "FLOOD_CAUSE"
## [31] "CATEGORY"           "TOR_F_SCALE"        "TOR_LENGTH"
## [34] "TOR_WIDTH"          "TOR_OTHER_WFO"      "TOR_OTHER_CZ_STATE"
## [37] "TOR_OTHER_CZ_FIPS"  "TOR_OTHER_CZ_NAME"  "BEGIN_RANGE"
## [40] "BEGIN_AZIMUTH"      "BEGIN_LOCATION"     "END_RANGE"
## [43] "END_AZIMUTH"        "END_LOCATION"       "BEGIN_LAT"
## [46] "BEGIN_LON"          "END_LAT"            "END_LON"
## [49] "EPISODE_NARRATIVE"  "EVENT_NARRATIVE"    "DATA_SOURCE"
```

```r
str(storm.current.raw)
```

```
## 'data.frame':	67872 obs. of  51 variables:
##  $ BEGIN_YEARMONTH   : int  202107 202107 202107 202107 202107 202107 202107 202107 202107 2021
07 ...
##  $ BEGIN_DAY         : int  20 22 30 31 20 31 3 3 3 11 ...
##  $ BEGIN_TIME        : int  2230 1449 1910 1330 2025 1630 2000 100 2136 1602 ...
##  $ END_YEARMONTH     : int  202107 202107 202107 202107 202107 202107 202107 202107 202107 2021
07 ...
##  $ END_DAY           : int  20 22 30 31 20 31 7 4 4 11 ...
##  $ END_TIME          : int  2230 1449 1910 1330 2025 1630 1131 0 700 1602 ...
##  $ EPISODE_ID        : int  159008 159623 159709 159711 159008 159711 162430 162430 162430 1592
46 ...
##  $ EVENT_ID          : int  961536 965330 965533 965535 961538 965545 980717 980718 980716 9770
11 ...
##  $ STATE             : Factor w/ 66 levels "ALABAMA","ALASKA",..: 9 9 9 9 9 9 3 3 3 1 ...
##  $ STATE_FIPS        : int  8 8 8 8 8 8 97 97 97 1 ...
##  $ YEAR              : int  2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 ...
##  $ MONTH_NAME        : Factor w/ 12 levels "April","August",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ EVENT_TYPE        : Factor w/ 48 levels "Astronomical Low Tide",..: 6 6 6 6 6 6 21 39 21 40
...
##  $ CZ_TYPE           : Factor w/ 2 levels "C","Z": 1 1 1 1 1 1 2 2 2 1 ...
##  $ CZ_FIPS           : int  97 37 91 113 45 45 3 2 2 95 ...
##  $ CZ_NAME           : Factor w/ 3262 levels "5NM E OF FAIRPORT MI TO ROCK ISLAND PASSAGE",..:
2217 807 2139 2438 1052 1052 1658 2910 2910 1674 ...
##  $ WFO               : Factor w/ 123 levels "ABQ","ABR","AFC",..: 45 45 45 45 45 45 11 11 11 55
...
##  $ BEGIN_DATE_TIME   : Factor w/ 36326 levels "01-APR-22 00:00:00",..: 24651 26905 34926 35933
24628 35961 2575 2517 2577 12395 ...
##  $ CZ_TIMEZONE       : Factor w/ 12 levels "AKST-9","AST-4",..: 9 9 9 9 9 9 12 12 12 4 ...
##  $ END_DATE_TIME     : Factor w/ 35216 levels "01-APR-22 07:00:00",..: 23731 25862 33747 34802
23706 34826 6421 3008 3010 11802 ...
##  $ INJURIES_DIRECT   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ INJURIES_INDIRECT : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ DEATHS_DIRECT     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ DEATHS_INDIRECT   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ DAMAGE_PROPERTY   : Factor w/ 298 levels "","0.00K","0.01K",..: 216 219 209 46 129 127 2 2 2
13 ...
##  $ DAMAGE_CROPS      : Factor w/ 188 levels "","0.00K","0.01K",..: 2 2 2 2 2 2 12 28 70 2 ...
##  $ SOURCE            : Factor w/ 42 levels "911 Call Center",..: 14 31 14 19 14 14 7 25 31 16
...
##  $ MAGNITUDE         : num  NA NA NA NA NA NA NA 26 NA 43 ...
##  $ MAGNITUDE_TYPE    : Factor w/ 5 levels "","EG","ES","MG",..: 1 1 1 1 1 1 3 1 2 ...
##  $ FLOOD_CAUSE       : Factor w/ 8 levels "","Dam / Levee Break",..: 3 3 3 3 4 4 1 1 1 1 ...
##  $ CATEGORY          : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ TOR_F_SCALE       : Factor w/ 7 levels "","EF0","EF1",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ TOR_LENGTH        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ TOR_WIDTH         : num  NA NA NA NA NA NA NA NA NA NA ...
```

```
##  $ TOR_OTHER_WFO     : Factor w/ 56 levels "","AKQ","APX",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ TOR_OTHER_CZ_STATE: Factor w/ 31 levels "","AL","AR","DC",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ TOR_OTHER_CZ_FIPS : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ TOR_OTHER_CZ_NAME : Factor w/ 204 levels "","ADAMS","ALEXANDER",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ BEGIN_RANGE       : int  2 1 4 1 2 4 NA NA NA 1 ...
##  $ BEGIN_AZIMUTH     : Factor w/ 17 levels "","E","ENE","ESE",..: 5 5 2 6 8 9 1 1 1 16 ...
##  $ BEGIN_LOCATION    : Factor w/ 16054 levels "","(0E4)PAYSON ARPT",..: 12066 1065 2540 12793 6
22 13127 1 1 1 6040 ...
##  $ END_RANGE         : int  2 1 4 1 2 4 NA NA NA 1 ...
##  $ END_AZIMUTH       : Factor w/ 17 levels "","E","ENE","ESE",..: 5 5 2 6 8 9 1 1 1 16 ...
##  $ END_LOCATION      : Factor w/ 16070 levels "","(0E4)PAYSON ARPT",..: 12107 1057 2539 12838 6
20 13165 1 1 1 6083 ...
##  $ BEGIN_LAT         : num  39.2 39.6 38 38 39.6 ...
##  $ BEGIN_LON         : num  -107 -107 -108 -108 -107 ...
##  $ END_LAT           : num  39.2 39.6 38 38 39.6 ...
##  $ END_LON           : num  -107 -107 -108 -108 -107 ...
##  $ EPISODE_NARRATIVE : Factor w/ 9435 levels "A  fast moving cold front brought breezy winds an
d a period of light snow showers through the region Wednesday "| __truncated__,..: 6349 5362 8594
5635 6349 5635 3857 3857 3857 6752 ...
##  $ EVENT_NARRATIVE   : Factor w/ 48445 levels "","A  24-hour rainfall measurement of 3.25 inche
s was observed.",..: 22361 23814 3410 2198 12404 22404 35256 34891 14518 8972 ...
##  $ DATA_SOURCE       : Factor w/ 1 level "CSV": 1 1 1 1 1 1 1 1 1 1 ...
```

The data types covered in this set include some dates, event description such as event type, event location, and some event characteristics such as numbers of injuries and deaths.

Analysis Preparation and Potential Questions in Mind

To start with, considering the data sets used are rather large, and I would only like to focus on the storm / disaster details, I removed the columns that have a majority of NA values and only kept those that I would like to conduct my analysis on.

```
# create subset for the current cycle
storm.current <- subset(storm.current.raw, select = -c(EPISODE_ID, SOURCE, CATEGORY, TOR_OTHER_WF
        O:DATA_SOURCE))
storm.current <- storm.current[order(storm.current$BEGIN_YEARMONTH),]


# create subset for the past cycle
storm.past <- subset(storm.past.raw, select = -c(EPISODE_ID, SOURCE, CATEGORY, TOR_OTHER_WFO:DATA_
        SOURCE))
storm.past <- storm.past[order(storm.past$BEGIN_YEARMONTH),]


# create subset for the consolidated set
storm.consolidated <- subset(storm.consolidated.raw, select = -c(EPISODE_ID, SOURCE, CATEGORY, TOR
        _OTHER_WFO:DATA_SOURCE))
storm.consolidated <- storm.consolidated[order(storm.consolidated$BEGIN_YEARMONTH),]


# print out the updated column names for the current cycle as an example, the column names are the
        same for all 3 datasets
colnames(storm.current)
```

```
##  [1] "BEGIN_YEARMONTH"   "BEGIN_DAY"         "BEGIN_TIME"
##  [4] "END_YEARMONTH"     "END_DAY"           "END_TIME"
##  [7] "EVENT_ID"          "STATE"             "STATE_FIPS"
## [10] "YEAR"              "MONTH_NAME"        "EVENT_TYPE"
## [13] "CZ_TYPE"           "CZ_FIPS"           "CZ_NAME"
## [16] "WFO"               "BEGIN_DATE_TIME"   "CZ_TIMEZONE"
## [19] "END_DATE_TIME"     "INJURIES_DIRECT"   "INJURIES_INDIRECT"
## [22] "DEATHS_DIRECT"     "DEATHS_INDIRECT"   "DAMAGE_PROPERTY"
## [25] "DAMAGE_CROPS"      "MAGNITUDE"         "MAGNITUDE_TYPE"
## [28] "FLOOD_CAUSE"       "TOR_F_SCALE"       "TOR_LENGTH"
## [31] "TOR_WIDTH"
```

```
summary(storm.current)
```

```
##  BEGIN_YEARMONTH    BEGIN_DAY      BEGIN_TIME   END_YEARMONTH
##  Min.   :202107   Min.   : 1.00   Min.   :   0   Min.   :202107
##  1st Qu.:202109   1st Qu.: 7.00   1st Qu.: 700   1st Qu.:202109
##  Median :202201   Median :13.00   Median :1424   Median :202201
##  Mean   :202163   Mean   :14.07   Mean   :1233   Mean   :202163
##  3rd Qu.:202204   3rd Qu.:21.00   3rd Qu.:1800   3rd Qu.:202204
##  Max.   :202206   Max.   :31.00   Max.   :2359   Max.   :202206
##
##     END_DAY         END_TIME       EVENT_ID               STATE
##  Min.   : 1.00   Min.   :   0   Min.   : 957393   TEXAS       : 4425
##  1st Qu.: 9.00   1st Qu.:1117   1st Qu.: 983924   MINNESOTA   : 2990
##  Median :15.00   Median :1620   Median :1002261   SOUTH DAKOTA: 2819
##  Mean   :16.35   Mean   :1488   Mean   :1002237   CALIFORNIA  : 2663
##  3rd Qu.:24.00   3rd Qu.:1918   3rd Qu.:1020215   NEW YORK    : 2561
##  Max.   :31.00   Max.   :2359   Max.   :1044700   KANSAS      : 2403
##                                                   (Other)     :50011
##    STATE_FIPS         YEAR         MONTH_NAME               EVENT_TYPE
##  Min.   : 1.00   Min.   :2021   June    : 8583   Thunderstorm Wind:18957
##  1st Qu.:20.00   1st Qu.:2021   May     : 8304   Hail             : 7306
##  Median :33.00   Median :2022   July    : 7856   High Wind        : 5998
##  Mean   :33.88   Mean   :2022   August  : 7744   Drought          : 5113
##  3rd Qu.:46.00   3rd Qu.:2022   December: 6009   Flash Flood      : 3959
##  Max.   :99.00   Max.   :2022   April   : 5986   Winter Weather   : 3891
##                                 (Other) :23390   (Other)          :22648
##  CZ_TYPE      CZ_FIPS             CZ_NAME            WFO
##  C:36097   Min.   :  1.0   WASHINGTON:  512   LWX    : 2808
##  Z:31775   1st Qu.: 25.0   MONTGOMERY:  472   FSD    : 1776
##            Median : 63.0   JEFFERSON :  408   ALY    : 1361
##            Mean   :107.4   JACKSON   :  398   PHI    : 1332
##            3rd Qu.:123.0   LINCOLN   :  395   OUN    : 1320
##            Max.   :873.0   MADISON   :  379   FGF    : 1281
##                            (Other)   :65308   (Other):57994
##            BEGIN_DATE_TIME   CZ_TIMEZONE               END_DATE_TIME
##  01-JAN-22 00:00:00:  538   CST-6  :29759   31-DEC-21 23:59:00:  476
##  01-APR-22 00:00:00:  524   EST-5  :23821   30-APR-22 23:59:00:  469
##  01-MAR-22 00:00:00:  510   MST-7  : 8871   31-MAY-22 23:59:00:  426
##  01-MAY-22 00:00:00:  490   PST-8  : 4001   31-JAN-22 23:59:00:  412
##  01-AUG-21 00:00:00:  447   HST-10 :  713   31-JUL-21 23:59:00:  412
##  01-JUN-22 00:00:00:  442   AKST-9 :  341   28-FEB-22 23:59:00:  410
##  (Other)           :64921   (Other):  366   (Other)           :65267
##  INJURIES_DIRECT     INJURIES_INDIRECT   DEATHS_DIRECT      DEATHS_INDIRECT
##  Min.   :  0.00000   Min.   : 0.000000   Min.   : 0.00000   Min.   : 0.000000
##  1st Qu.:  0.00000   1st Qu.: 0.000000   1st Qu.: 0.00000   1st Qu.: 0.000000
##  Median :  0.00000   Median : 0.000000   Median : 0.00000   Median : 0.000000
##  Mean   :  0.02095   Mean   : 0.004774   Mean   : 0.00853   Mean   : 0.002328
##  3rd Qu.:  0.00000   3rd Qu.: 0.000000   3rd Qu.: 0.00000   3rd Qu.: 0.000000
```

```
## Max.   :210.00000   Max.   :30.000000   Max.   :53.00000   Max.   :12.000000
##
## DAMAGE_PROPERTY  DAMAGE_CROPS     MAGNITUDE      MAGNITUDE_TYPE
## 0.00K  :40075   0.00K  :51968   Min.   :  0.25     :38832
##        :15589          :14946   1st Qu.: 37.00   EG:17346
## 1.00K  : 2095   1.00K  :  203   Median : 50.00   ES:   22
## 5.00K  : 1325   2.00K  :  102   Mean   : 41.99   MG:11295
## 2.00K  : 1243   0.50K  :   61   3rd Qu.: 55.00   MS:  377
## 10.00K : 1161   3.00K  :   59   Max.   :138.00
## (Other): 6384   (Other):  533   NA's   :31494
##                       FLOOD_CAUSE    TOR_F_SCALE   TOR_LENGTH
##                             :61626      :66031   Min.   : 0.01
## Heavy Rain                 : 5836   EF0:  673   1st Qu.: 0.68
## Heavy Rain / Burn Area      :  169   EF1:  728   Median : 2.16
## Heavy Rain / Tropical System:  142   EF2:  189   Mean   : 3.63
## Heavy Rain / Snow Melt      :   63   EF3:   37   3rd Qu.: 4.97
## Ice Jam                     :   29   EF4:   10   Max.   :33.97
## (Other)                     :    7   EFU:  204   NA's   :66031
##   TOR_WIDTH
## Min.   :   1
## 1st Qu.:  50
## Median : 100
## Mean   : 183
## 3rd Qu.: 200
## Max.   :2600
## NA's   :66031
```

```
head(storm.current)
```

```
##   BEGIN_YEARMONTH BEGIN_DAY BEGIN_TIME END_YEARMONTH END_DAY END_TIME EVENT_ID
## 1          202107        20       2230        202107      20     2230   961536
## 2          202107        22       1449        202107      22     1449   965330
## 3          202107        30       1910        202107      30     1910   965533
## 4          202107        31       1330        202107      31     1330   965535
## 5          202107        20       2025        202107      20     2025   961538
## 6          202107        31       1630        202107      31     1630   965545
##      STATE STATE_FIPS YEAR MONTH_NAME  EVENT_TYPE CZ_TYPE CZ_FIPS    CZ_NAME
## 1 COLORADO          8 2021       July Debris Flow       C      97     PITKIN
## 2 COLORADO          8 2021       July Debris Flow       C      37      EAGLE
## 3 COLORADO          8 2021       July Debris Flow       C      91      OURAY
## 4 COLORADO          8 2021       July Debris Flow       C     113 SAN MIGUEL
## 5 COLORADO          8 2021       July Debris Flow       C      45   GARFIELD
## 6 COLORADO          8 2021       July Debris Flow       C      45   GARFIELD
##   WFO    BEGIN_DATE_TIME CZ_TIMEZONE      END_DATE_TIME INJURIES_DIRECT
## 1 GJT 20-JUL-21 22:30:00       MST-7 20-JUL-21 22:30:00               0
## 2 GJT 22-JUL-21 14:49:00       MST-7 22-JUL-21 14:49:00               0
## 3 GJT 30-JUL-21 19:10:00       MST-7 30-JUL-21 19:10:00               0
## 4 GJT 31-JUL-21 13:30:00       MST-7 31-JUL-21 13:30:00               0
## 5 GJT 20-JUL-21 20:25:00       MST-7 20-JUL-21 20:25:00               0
## 6 GJT 31-JUL-21 16:30:00       MST-7 31-JUL-21 16:30:00               0
##   INJURIES_INDIRECT DEATHS_DIRECT DEATHS_INDIRECT DAMAGE_PROPERTY DAMAGE_CROPS
## 1                 0             0               0          50.00K        0.00K
## 2                 0             0               0         500.00K        0.00K
## 3                 0             0               0           5.00K        0.00K
## 4                 0             0               0          10.00K        0.00K
## 5                 0             0               0         250.00K        0.00K
## 6                 0             0               0          25.00M        0.00K
##   MAGNITUDE MAGNITUDE_TYPE            FLOOD_CAUSE TOR_F_SCALE TOR_LENGTH
## 1        NA                            Heavy Rain                     NA
## 2        NA                            Heavy Rain                     NA
## 3        NA                            Heavy Rain                     NA
## 4        NA                            Heavy Rain                     NA
## 5        NA                Heavy Rain / Burn Area                     NA
## 6        NA                Heavy Rain / Burn Area                     NA
##   TOR_WIDTH
## 1        NA
## 2        NA
## 3        NA
## 4        NA
## 5        NA
## 6        NA
```

Before starting the analysis, install the libraries / packages that will be used.

```
# install libraries
install.packages("tidyverse")
install.packages("plyr")
install.packages("usmap")
install.packages("caret")
```

Below is a list of questions I would like to examine throughout this project:

I.Exploratory Data Analysis for the Current Cycle:

- Which type(s) of event occur the most often?
- Is there a specific month / season when more events happen? Throughout the year, how does the event occurrence fluctuate through time?
- Which state(s) have potentially more events compared to other states?
- For the states that occur more events than others, what are the most frequent events?
- Were there high injuries and deaths for the events?
- Is there a relationship between the length and the width of tornadoes?

II.Comparison of the Current Cycle and the Past Cycle:

- Are the top 5 events that occur the most in the current cycle same as these in the past?
- Are there any differences of the time fluctuation of event occurrence?
- Has the wind speed changed throughout the years?
- Has the hail size changed throughout the years?

III.Prediction using Long Term Time Series Data:

- Can we predict the tornado width with tornado length for the next 20 tornadoes using previous collected data?

I. Exploratory Data Analysis for the Current Cycle

Q1a: Which type(s) of event occur the most often?

Before starting the analysis, understand what the potential types / option for event type are.Relevant libraries should also be loaded.

```
unique(storm.current$EVENT_TYPE)
```

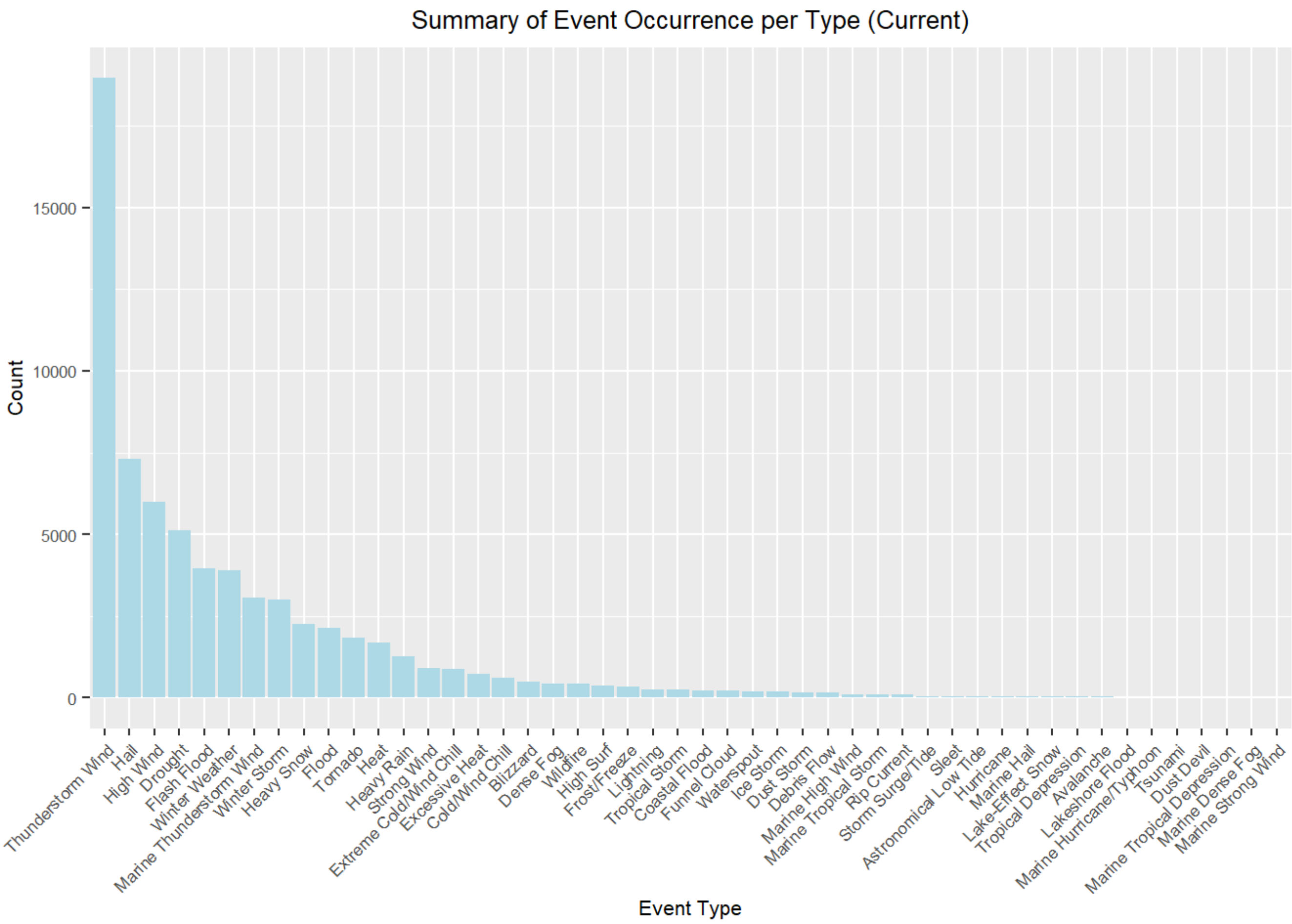
```
##  [1] Debris Flow                High Surf
##  [3] Strong Wind                Thunderstorm Wind
##  [5] Hail                       Wildfire
##  [7] Heat                       Drought
##  [9] Funnel Cloud               Flash Flood
## [11] Lightning                  Marine Thunderstorm Wind
## [13] Heavy Rain                 Flood
## [15] Dust Storm                 High Wind
## [17] Tropical Storm             Waterspout
## [19] Rip Current                Tornado
## [21] Marine Tropical Storm      Tropical Depression
## [23] Excessive Heat             Marine Hail
## [25] Coastal Flood              Storm Surge/Tide
## [27] Marine Tropical Depression Marine High Wind
## [29] Marine Strong Wind         Dense Fog
## [31] Astronomical Low Tide      Hurricane
## [33] Marine Hurricane/Typhoon   Marine Dense Fog
## [35] Frost/Freeze               Avalanche
## [37] Winter Weather             Lakeshore Flood
## [39] Heavy Snow                 Winter Storm
## [41] Blizzard                   Lake-Effect Snow
## [43] Cold/Wind Chill            Extreme Cold/Wind Chill
## [45] Ice Storm                  Tsunami
## [47] Sleet                      Dust Devil
## 48 Levels: Astronomical Low Tide Avalanche Blizzard ... Winter Weather
```

```
library(plyr)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.3.6      ✓ purrr   0.3.4
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.2      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::arrange()   masks plyr::arrange()
## ✕ purrr::compact()   masks plyr::compact()
## ✕ dplyr::count()     masks plyr::count()
## ✕ dplyr::failwith()  masks plyr::failwith()
## ✕ dplyr::filter()    masks stats::filter()
## ✕ dplyr::id()        masks plyr::id()
## ✕ dplyr::lag()       masks stats::lag()
## ✕ dplyr::mutate()    masks plyr::mutate()
## ✕ dplyr::rename()    masks plyr::rename()
## ✕ dplyr::summarise() masks plyr::summarise()
## ✕ dplyr::summarize() masks plyr::summarize()
```

A barplot is illustrated below to examine the occurrence of each event type.

```
# plot number of events occurred per type
event.plot <- ggplot(data = storm.current, aes(x = fct_infreq(EVENT_TYPE)))
event.plot + geom_bar(fill = 'lightblue') + xlab("Event Type") + ylab("Count") + ggtitle("Summary
         of Event Occurrence per Type (Current)") + theme(text = element_text(size=8),
        axis.text.x = element_text(angle=45, hjust=1), plot.title = element_text(hjust = 0.5))
```

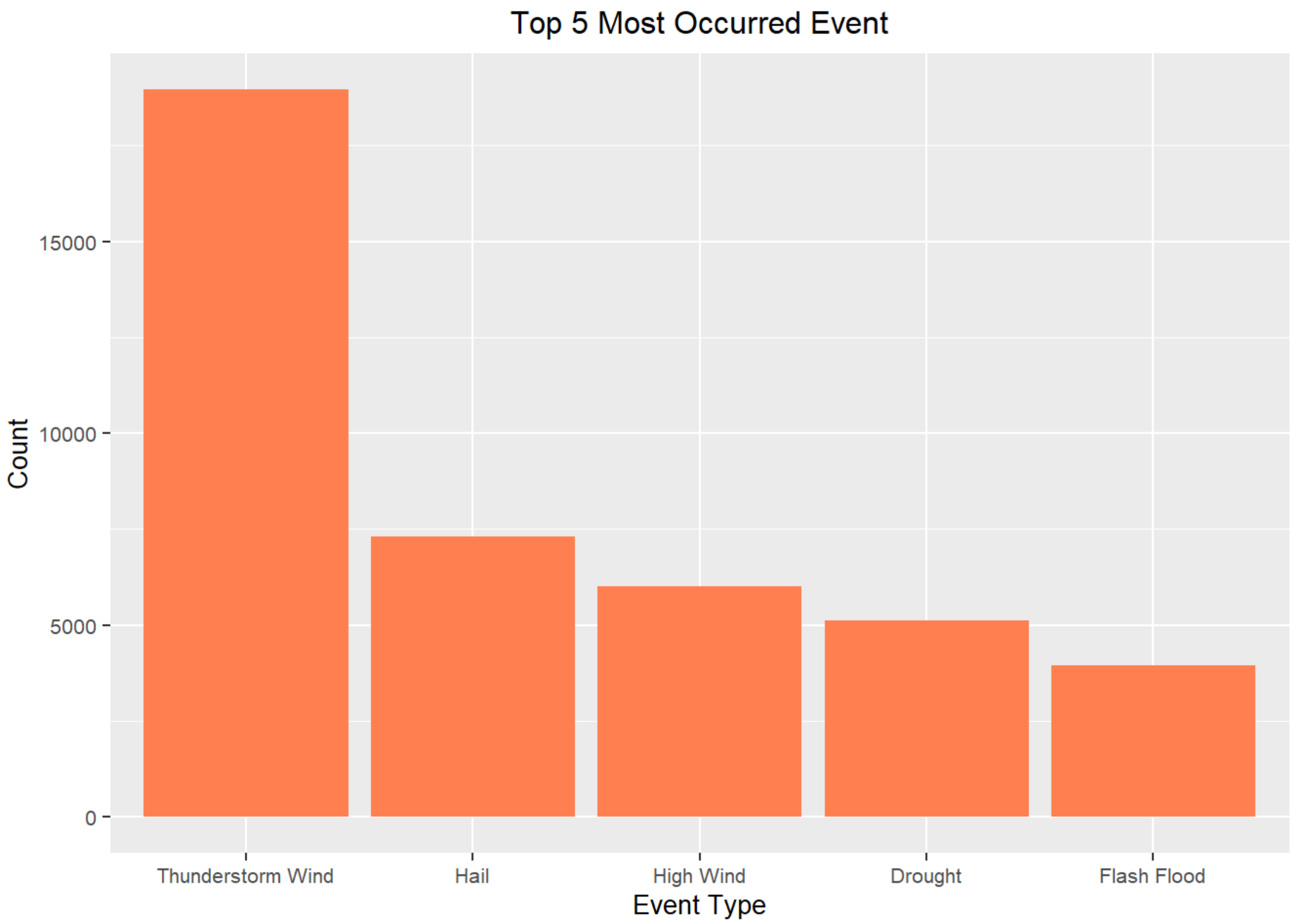



```
# a closer look on the top 5 event types
table(storm.current$EVENT_TYPE)
```

```
## 
##     Astronomical Low Tide                    Avalanche 
##                        38                           26 
##                  Blizzard                Coastal Flood 
##                       484                          216 
##           Cold/Wind Chill                  Debris Flow 
##                       612                          146 
##                 Dense Fog                      Drought 
##                       428                         5113 
##                Dust Devil                   Dust Storm 
##                         8                          168 
##            Excessive Heat      Extreme Cold/Wind Chill 
##                       714                          865 
##               Flash Flood                        Flood 
##                      3959                         2141 
##              Frost/Freeze                 Funnel Cloud 
##                       338                          213 
##                      Hail                         Heat 
##                      7306                         1688 
##                Heavy Rain                   Heavy Snow 
##                      1265                         2241 
##                 High Surf                    High Wind 
##                       370                         5998 
##                 Hurricane                    Ice Storm 
##                        38                          177 
##          Lake-Effect Snow               Lakeshore Flood 
##                        31                           12 
##                 Lightning              Marine Dense Fog 
##                       253                            3 
##               Marine Hail             Marine High Wind 
##                        32                          102 
##   Marine Hurricane/Typhoon           Marine Strong Wind 
##                        12                            2 
##   Marine Thunderstorm Wind Marine Tropical Depression 
##                      3061                            8 
##     Marine Tropical Storm                  Rip Current 
##                        97                           93 
##                     Sleet             Storm Surge/Tide 
##                        42                           50 
##               Strong Wind            Thunderstorm Wind 
##                       920                        18957 
##                   Tornado          Tropical Depression 
##                      1841                           28 
##            Tropical Storm                      Tsunami 
##                       253                            9 
##                Waterspout                     Wildfire 
```

```
##                 202                 427
##        Winter Storm      Winter Weather
##                2994                3891
```

```
# the top 5 most occurred events are Thunderstorm Wind, Hail, High Wind, Drought, Flash Flood
top5event <- c('Thunderstorm Wind', 'Hail', 'High Wind', 'Drought', 'Flash Flood')

top5event.plot <- ggplot(data = storm.current[storm.current$EVENT_TYPE %in% top5event,], aes(x = f
        ct_infreq(EVENT_TYPE)))
top5event.plot + geom_bar(fill = 'coral') + ylab("Count") + xlab("Event Type") + ggtitle("Top 5 Mo
        st Occurred Event") + theme(text = element_text(size=10),
        axis.text.x = element_text(angle=0), plot.title = element_text(hjust = 0.5))
```

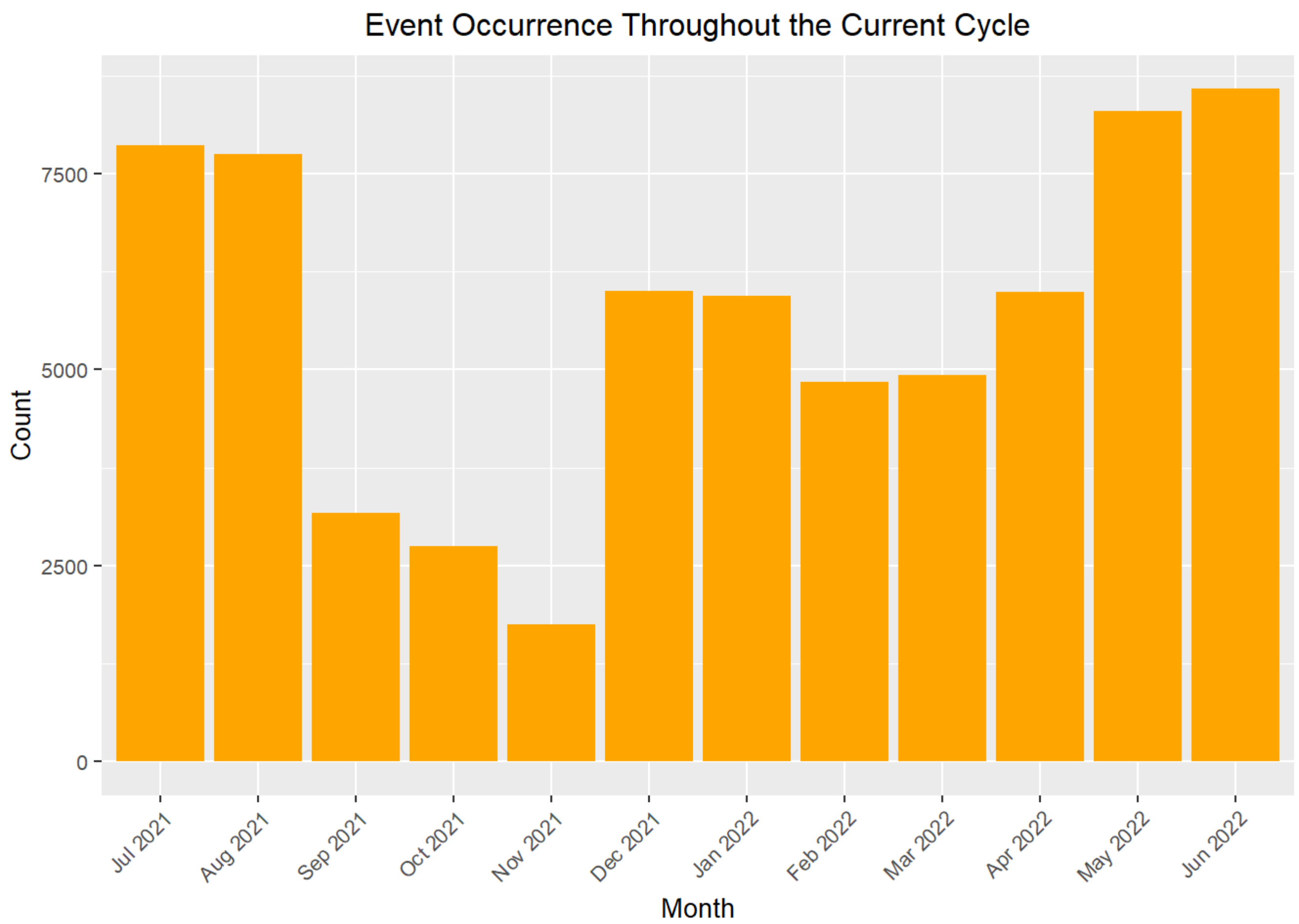


From the barplot above, it is noted that during the current cycle, thunderstorm wind occurs the most often, following by hail, high wind, drought, and flash flood. It is also noted that thunderstorm wind occurs more than twice of the number of hails.

Q1b: Is there a specific month / season when more events happen? Throughout the year, how does the event occurrence fluctuate through time?

A barplot is illustrated below to examine the occurrence of events throughout the year.

```
# plot of event occurrence throughout the current cycle
month.order <- c('July', 'August', 'September', 'October', 'November', 'December', 'January', 'February', 'March', 'April', 'May', 'June')
month.labels <- c('Jul 2021', 'Aug 2021', ' Sep 2021', 'Oct 2021', 'Nov 2021', 'Dec 2021', 'Jan 2022', 'Feb 2022', 'Mar 2022', 'Apr 2022', 'May 2022', 'Jun 2022')
month.plot <- ggplot(data = storm.current, aes(x = MONTH_NAME))
month.plot + geom_bar(fill = 'orange') + xlab("Month") + ylab("Count") + ggtitle("Event Occurrence Throughout the Current Cycle") + theme(text = element_text(size=10), axis.text.x = element_text(angle=45, hjust=1), plot.title = element_text(hjust = 0.5)) + scale_x_discrete(limits = month.order, labels = month.labels)
```



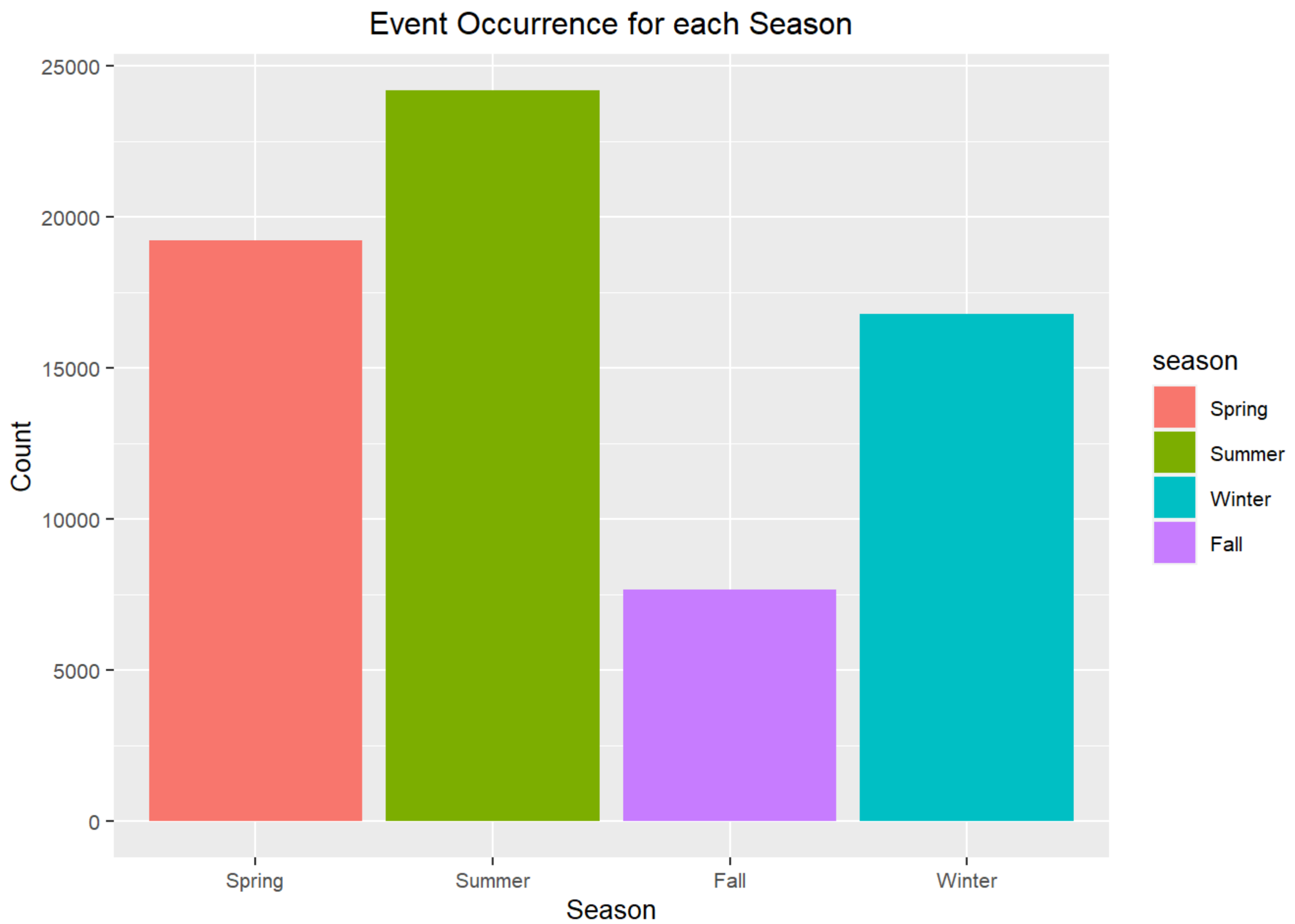
From the barplot above, it is noted that throughout the current cycle, it is more likely to have more events during summer. For the current cycle, June 2022 has occurred the most events, and November 2021 has occurred the least events.

Let's also view the differences between each season. Following meteorological seasons, divide the months into the seasons as the following:

- Spring: March, April, May
- Summer: June, July, August
- Fall: September, October, November
- Winter: December, January, February

A barplot is illustrated below to examine the occurrence of events for each season.

```
# assign months to each season category
storm.current.season <- mutate(storm.current,
                               season = as.factor(plyr::mapvalues(MONTH_NAME, c('July', 'Augus
t', 'September', 'October', 'November', 'December', 'January', 'February', 'March', 'Apri
l', 'May', 'June'), c('Summer', 'Summer', 'Fall', 'Fall', 'Fall', 'Winter', 'Winter', 'Win
ter', 'Spring', 'Spring', 'Spring', 'Summer'))))


# draw the barplot for each season
season.order <- c('Spring', 'Summer', 'Fall', 'Winter')
season.plot <- ggplot(data = storm.current.season, aes(x = season, fill = season))
season.plot + geom_bar() + xlab("Season") + ylab("Count") + ggtitle("Event Occurrence for each Sea
son") + theme(text = element_text(size=10), plot.title = element_text(hjust = 0.5)) + scal
e_x_discrete(limits = season.order, labels = season.order)
```



From the barplot above, it shows clearly that summer is the season which occurs the most events. Whereas the number of events in spring and winter are similar, and fall is the season which has the least events.

Q1c: Which state(s) have potentially more events compared to other states?

While there are 50 states and 3243 counties in US, I was also curious on which states were more prone to these natural disasters.The current dataset in use has in fact more than 50 values on its STATE column; this is because areas such as E Pacific, GULF OF MEXICO, LAKE HURON, etc. were also included.

```
# generate the count of events per state
state.event.summary <- storm.current %>%
  group_by(STATE) %>%
  dplyr::summarize(n = n()) %>%
  arrange(., desc(n))
colnames(state.event.summary) <- c("STATE", "Count")
knitr::kable(state.event.summary, caption = "Count of Events per State")
```

Count of Events per State

STATE	Count
TEXAS	4425
MINNESOTA	2990
SOUTH DAKOTA	2819
CALIFORNIA	2663
NEW YORK	2561
KANSAS	2403
VIRGINIA	2330
PENNSYLVANIA	2165
NEBRASKA	1946
COLORADO	1823
OKLAHOMA	1805
IOWA	1788
ILLINOIS	1779
KENTUCKY	1746
MISSOURI	1587
NORTH CAROLINA	1489
NORTH DAKOTA	1466
MONTANA	1446
WISCONSIN	1431
OHIO	1419
TENNESSEE	1335
NEW MEXICO	1308

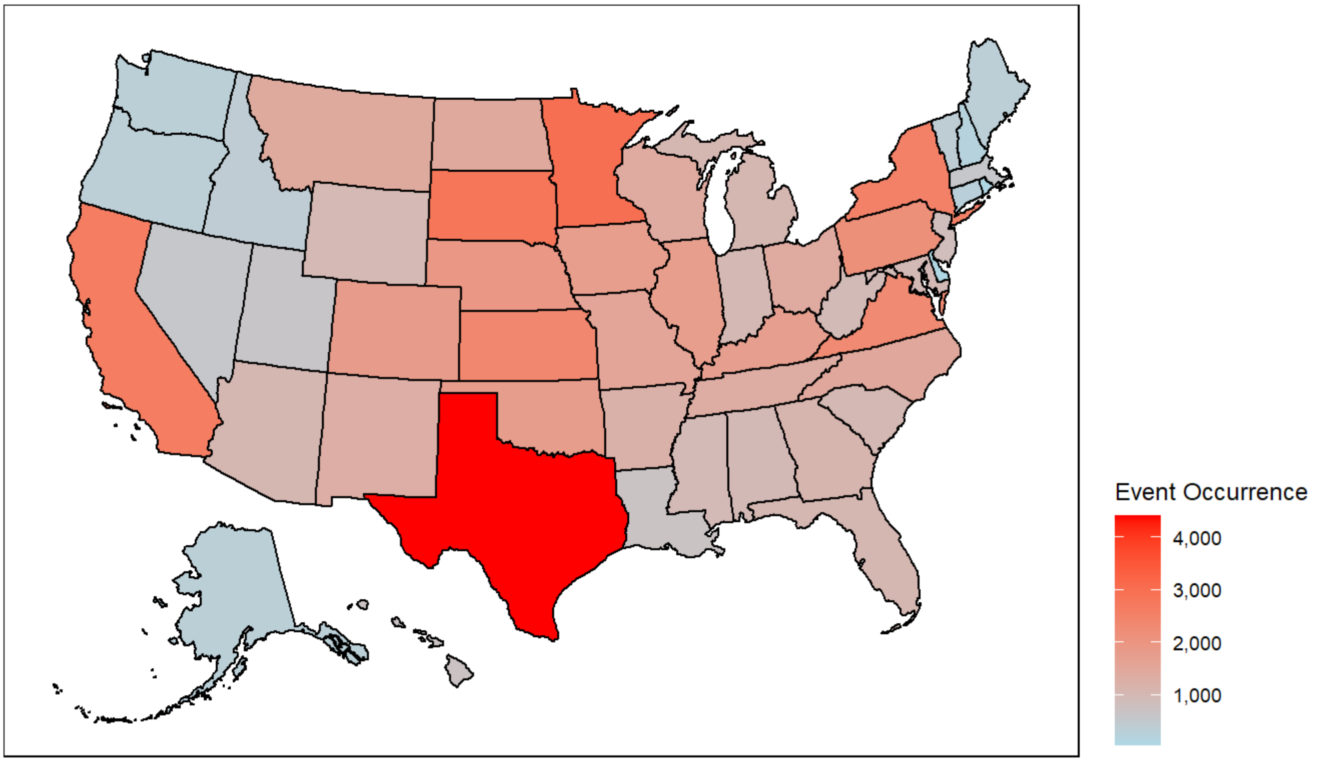
STATE	Count
GULF OF MEXICO	1237
ARKANSAS	1223
ATLANTIC NORTH	1157
GEORGIA	1103
MICHIGAN	1056
ARIZONA	1052
FLORIDA	1043
INDIANA	1026
SOUTH CAROLINA	1025
WYOMING	1005
ALABAMA	988
MARYLAND	981
MISSISSIPPI	975
WEST VIRGINIA	958
NEW JERSEY	855
HAWAII	713
LOUISIANA	705
UTAH	626
ATLANTIC SOUTH	579
MASSACHUSETTS	579
NEVADA	576
IDAHO	419
VERMONT	399
MAINE	346
OREGON	345
ALASKA	341
WASHINGTON	333
CONNECTICUT	280

STATE	Count
NEW HAMPSHIRE	233
LAKE MICHIGAN	189
PUERTO RICO	183
LAKE SUPERIOR	129
LAKE ERIE	108
DELAWARE	85
RHODE ISLAND	80
LAKE HURON	55
DISTRICT OF COLUMBIA	49
LAKE ST CLAIR	40
AMERICAN SAMOA	23
LAKE ONTARIO	17
GUAM	16
VIRGIN ISLANDS	8
ST LAWRENCE R	5
E PACIFIC	3

From the above table, it is noted that Texas, Minnesota, South Dakota, California and New York are the top 5 states with the most events during the current cycle. While there is a lot of information in this table, for the reader's easier view, a US map is also created as below:

```
library(usmap)

# add a state column for usmap
state.event.map <- mutate(state.event.summary,
                        state = STATE)
plot_usmap(data = state.event.map, value = 'Count', labels = FALSE) +
  scale_fill_continuous(low = 'lightblue', high = 'red',
                        name = 'Event Occurrence', label = scales::comma) +
  theme(legend.position = 'right') +
  theme(panel.background = element_rect(color = 'black')) +
  labs(title = 'Event Occurrence for each State')
```

Event Occurrence for each State

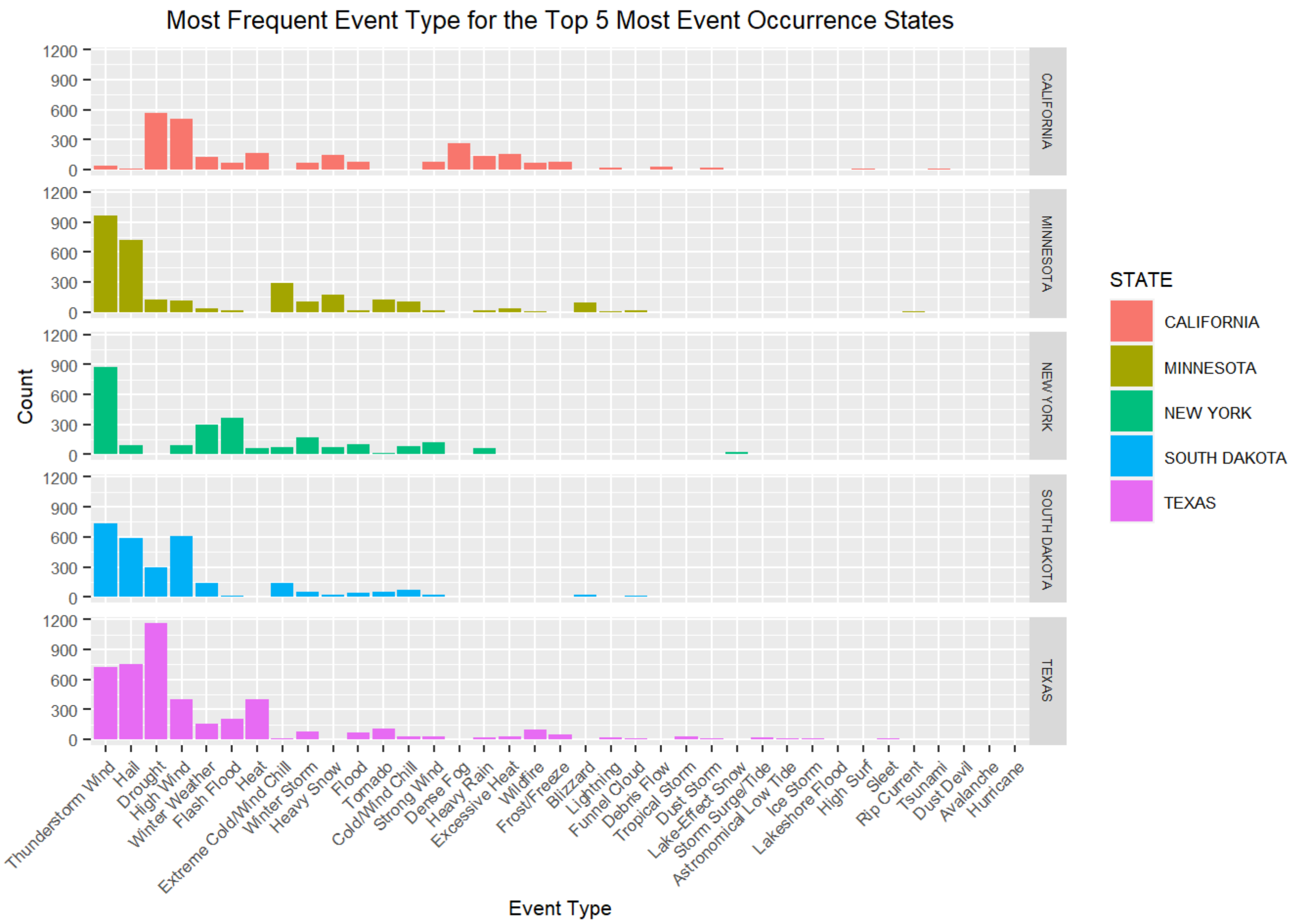


From the above US map, it is very clear that Texas, with the largest event occurrence frequency, stands out of the other states. The east and north part of US also has larger event occurrences compared to the states on the west side (excluding california).

Q1d: For the states that occur more events than others, what are the most frequent events?

Now that knowing which states have higher event occurrences, I was also curious about their corresponding top event types. The following identifies the top event types occurring in Texas, Minnesota, South Dakota, California, and New York.

```
top5states <- c('TEXAS', 'MINNESOTA', 'SOUTH DAKOTA', 'CALIFORNIA', 'NEW YORK')


# plot top events for the 5 states
top5states.plot <- ggplot(data = storm.current[storm.current$STATE %in% top5states,], aes(x = fct_
         infreq(EVENT_TYPE), fill = STATE))


top5states.plot + geom_bar() + xlab("Event Type") + ylab("Count") + facet_grid(STATE~.) + ggtitle
         ("Most Frequent Event Type for the Top 5 Most Event Occurrence States") + theme(strip.text
         = element_text(size = 5), text = element_text(size=8),
        axis.text.x = element_text(angle=45, hjust=1),  plot.title = element_text(hjust = 0.5))
```

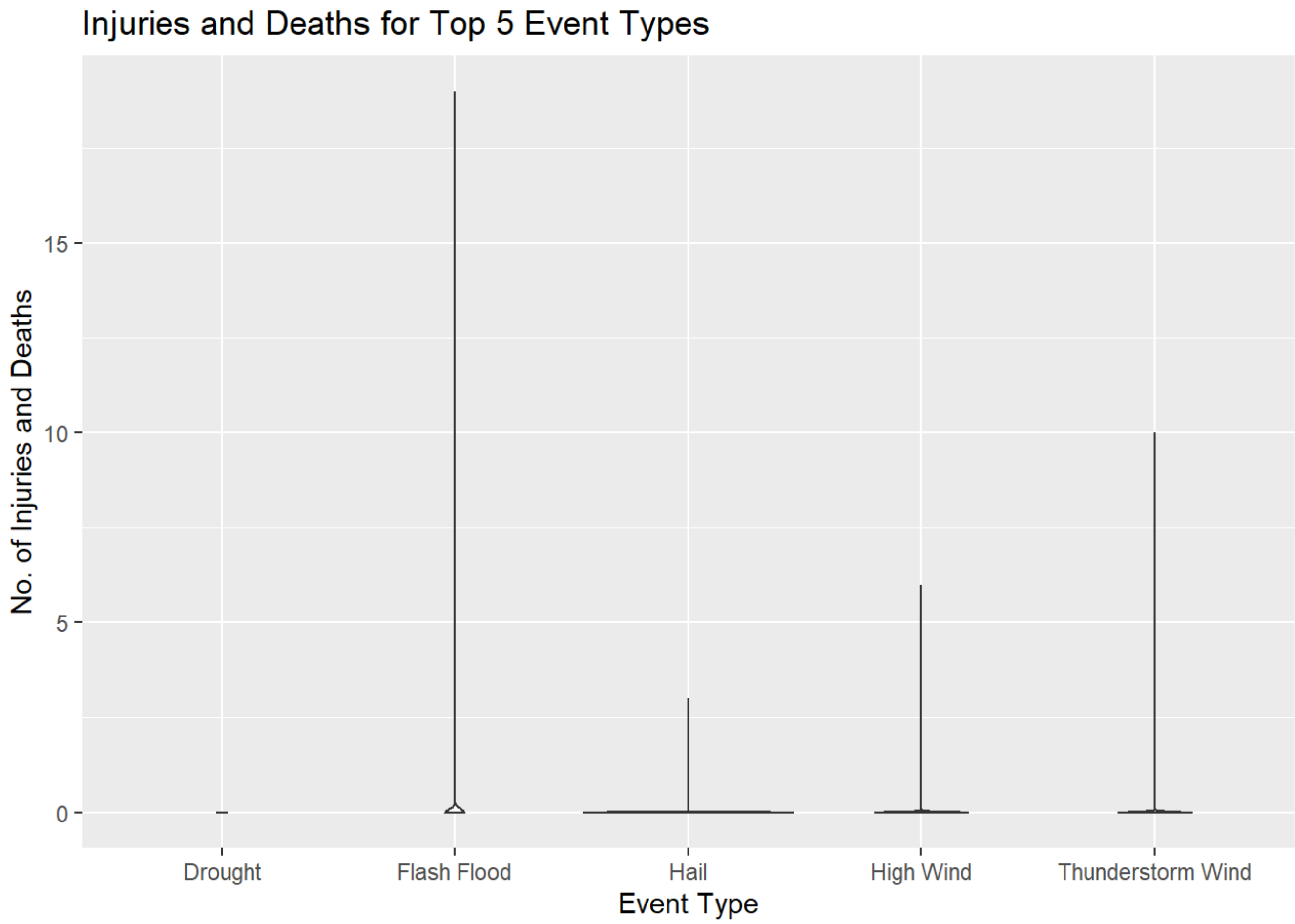


From the grouped graph above, it is noted that other than California, all other 4 has a higher occurrence of thunderstorm wind and hail, which could be related to their locations (since they are on the central / east parts of US). Moreover, Texas has more droughts compared to the others, considering its geographic location is also farther to seas / oceans. A special thing about California is that it has a significant occurrence of dense fog, perhaps due to its location near Pacific ocean and its topography.

Q1e: Were there high injuries and deaths for the events?

Assuming there is a positive relationship between injury / death amount and the frequency of the events, I would like to plot a violin plot for the top 5 event types identified in 1a (i.e. 'Thunderstorm Wind', 'Hail', 'High Wind', 'Drought', 'Flash Flood'.

```
top5event <- c('Thunderstorm Wind', 'Hail', 'High Wind', 'Drought', 'Flash Flood')


# aggregate injury and death data
storm.current.injuries <- mutate(storm.current,
                          INJURY_DEATH = INJURIES_DIRECT + INJURIES_INDIRECT + DEATHS_DIRECT + DEATH
        S_INDIRECT)
storm.current.injuries <- filter(storm.current.injuries, EVENT_TYPE %in% top5event)


storm.current.injuries.plot <- ggplot(storm.current.injuries, aes(x = EVENT_TYPE, y = INJURY_DEAT
        H))
storm.current.injuries.plot + geom_violin() + xlab("Event Type") + ylab("No. of Injuries and Death
        s") + ggtitle("Injuries and Deaths for Top 5 Event Types")
```



From the above violin plot, it is noted that actually the injuries and deaths rate for the top 5 events are mostly zero. For this current cycle, drought has not led to any injuries and deaths. However, it is also noticed that there are several injuries and deaths for flash flood, with the largest number of injuries and deaths per event around 18.

Q1f: Is there a relationship between the length and the width of tornadoes?

While tornadoes are very common in US, I was curious if a longer tornado always has a wider width. Understood that tornadoes may have different shapes and sizes, I decided to do an examination of the correlation between the length and the width of tornadoes.

Since the dataset uses miles for length and feet for width, I decided to transform the unit of width into miles by dividing it with 5280 to make easier interpretation and get clearer plots. Before directly drawing the linear model plot, I also decided to conduct a t-test to inspect the relationship between tornado length and width.

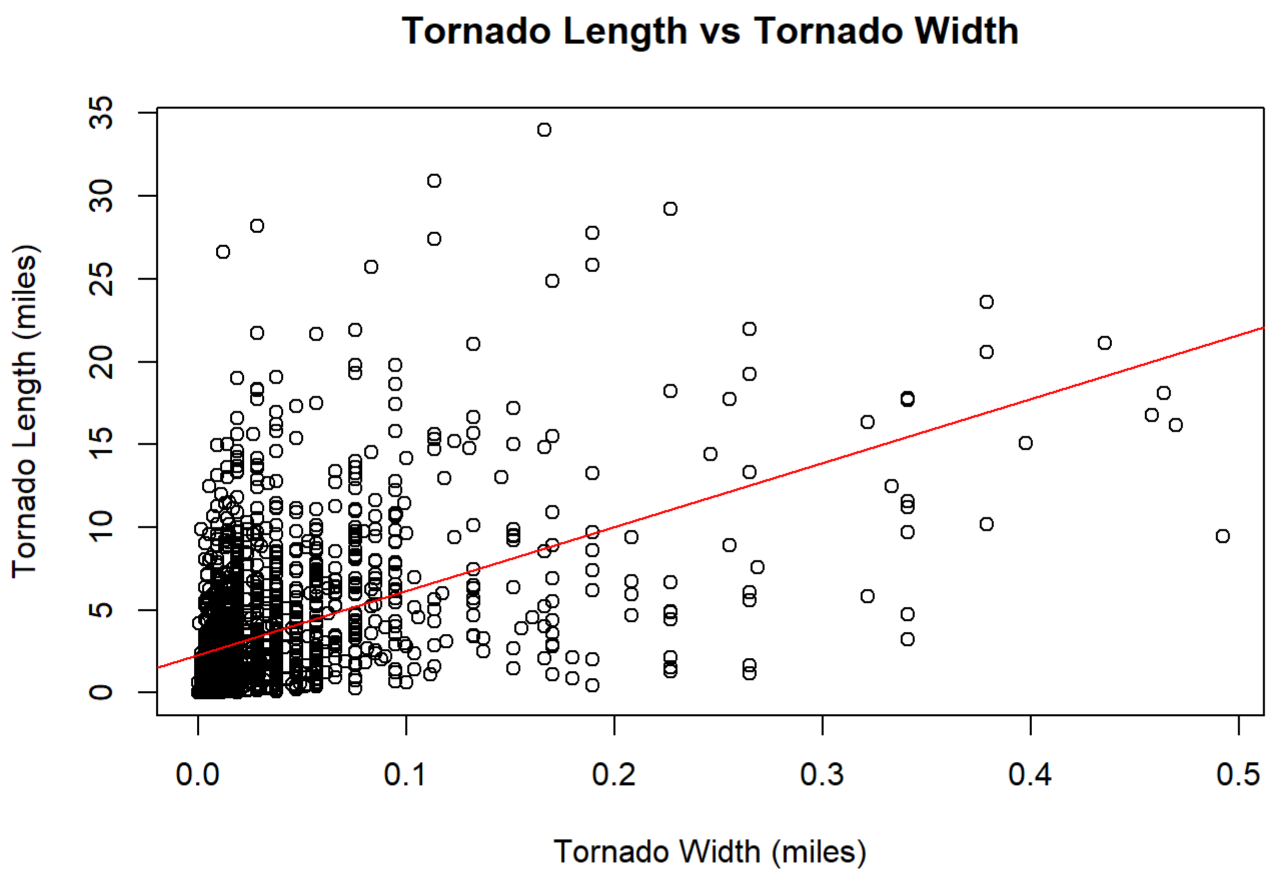
```
# add TOR_WIDTH2 to transform the unit of width into miles
storm.current <- mutate(storm.current,
                        TOR_WIDTH2 = TOR_WIDTH / 5280)

# conduct a t-test on TOR_LENGTH and TOR_WIDTH2 first
t.test(x = storm.current$TOR_WIDTH2, y = storm.current$TOR_LENGTH)
```

```
##
##  Welch Two Sample t-test
##
## data:  storm.current$TOR_WIDTH2 and storm.current$TOR_LENGTH
## t = -35.688, df = 1840.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.792876 -3.397716
## sample estimates:
##  mean of x  mean of y
## 0.03466595 3.62996198
```

```
# construct a linear model using TOR_LENGTH and TOR_WIDTH of the dataset
torn.lmfit <-lm(TOR_LENGTH ~ TOR_WIDTH2, data = storm.current)
summary(torn.lmfit)
```

```
##
## Call:
## lm(formula = TOR_LENGTH ~ TOR_WIDTH2, data = storm.current)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.241  -2.248  -1.195   1.228  25.232
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2885     0.1046   21.89   <2e-16 ***
## TOR_WIDTH2   38.6971     1.6252   23.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.78 on 1839 degrees of freedom
##   (66031 observations deleted due to missingness)
## Multiple R-squared:  0.2357, Adjusted R-squared:  0.2352
## F-statistic:   567 on 1 and 1839 DF,  p-value: < 2.2e-16
```

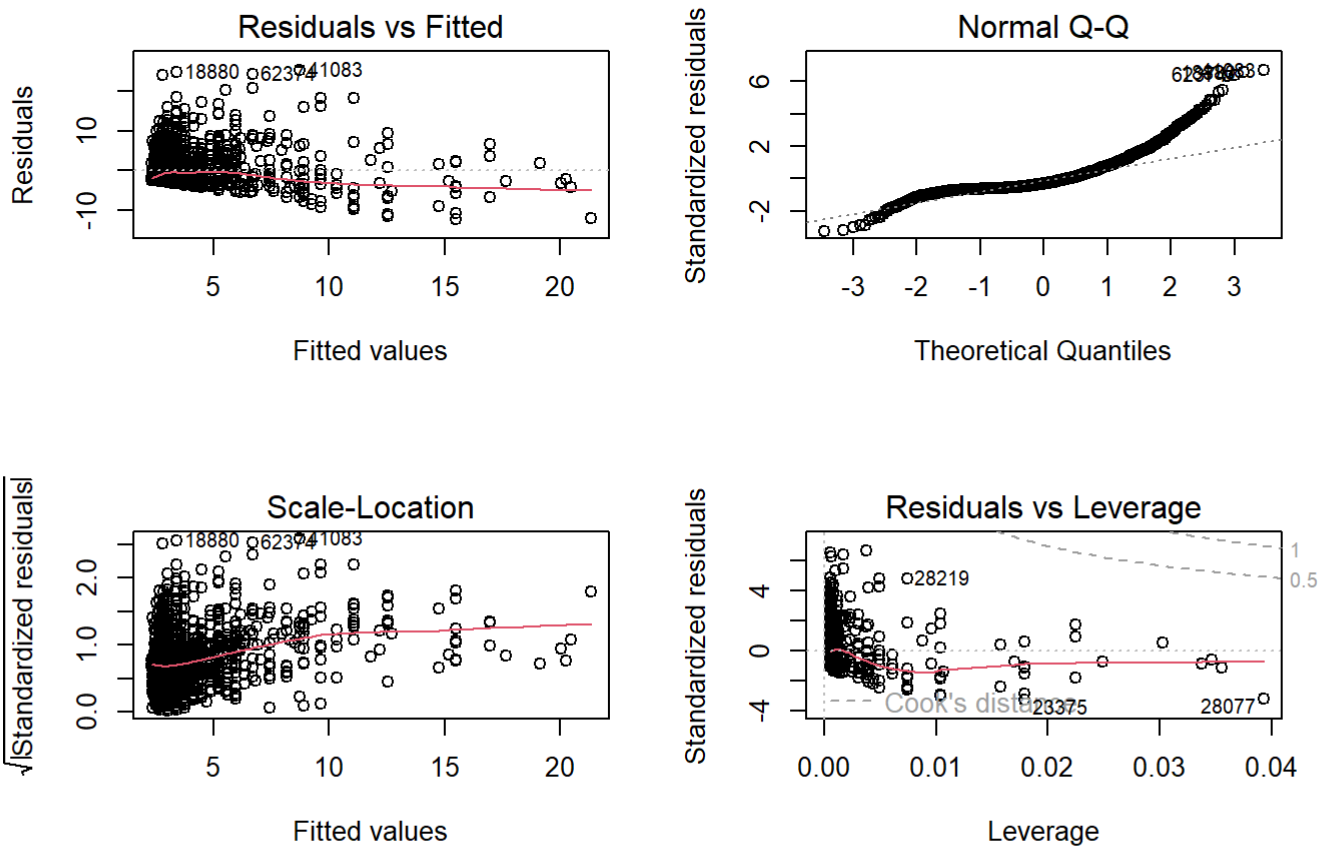
```
plot(storm.current$TOR_LENGTH ~ storm.current$TOR_WIDTH2,
     ylab = "Tornado Length (miles)",
     xlab = "Tornado Width (miles)",
     main = "Tornado Length vs Tornado Width")
abline(torn.lmfit, col = 'red')
```



From the regression statistics, it is likely that tornado width does have some impact on the tornado length as the p-value of the x variable (i.e. tornado width) is very small. However, it is also noted from the low R-squared value that the model does not fit well. Hence, by viewing at the plot and considering the very low R-squared value, I would not be able to state that there is a linear relationship between the tornado width and length.

More graphs regarding the linear model of tornado width and length are generated below:

```
par(mfrow = c(2, 2))
plot(torn.lmfit)
```



It is clear that there is a large difference between the residuals and the fitted line. This further supports a poor to no linear correlation.

Seeing this poor statistics, I was curious if taking some transformation to the model could help. Below outlines the linear model of tornado length and square of tornado width.

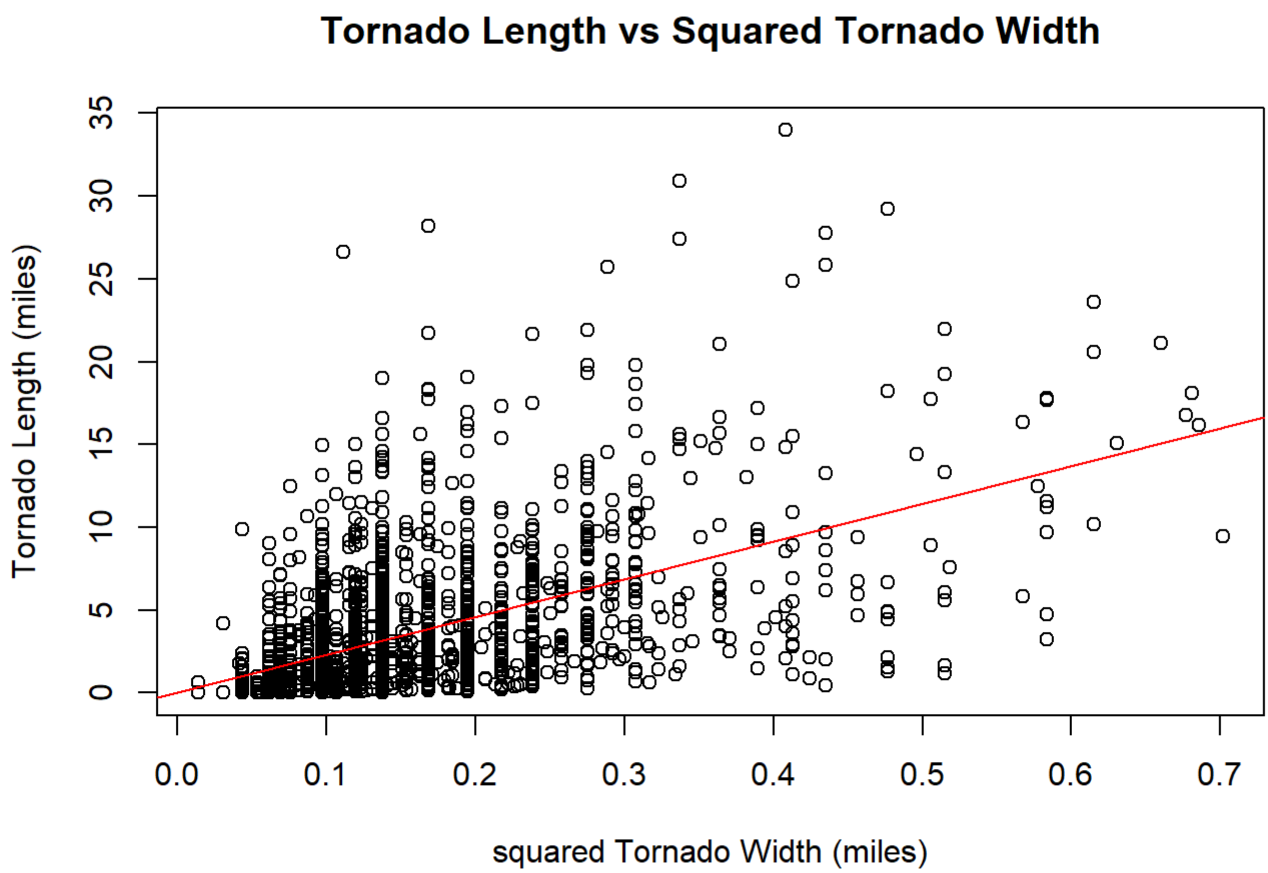
```
# conduct a t-test on TOR_LENGTH and TOR_WIDTH2 first
t.test(x = sqrt(storm.current$TOR_WIDTH2), y = storm.current$TOR_LENGTH)
```

```
##
##  Welch Two Sample t-test
##
## data:  sqrt(storm.current$TOR_WIDTH2) and storm.current$TOR_LENGTH
## t = -34.469, df = 1842, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.670718 -3.275483
## sample estimates:
## mean of x mean of y
## 0.1568612 3.6299620
```

```r
# construct a linear model using TOR_LENGTH and square(TOR_WIDTH) of the dataset
ttorn.lmfit <-lm(TOR_LENGTH ~ sqrt(TOR_WIDTH2), data = storm.current)
summary(ttorn.lmfit)
```

```
##
## Call:
## lm(formula = TOR_LENGTH ~ sqrt(TOR_WIDTH2), data = storm.current)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5961  -1.9764  -0.9582   1.1411  24.6139
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.05696    0.15876   0.359     0.72
## sqrt(TOR_WIDTH2) 22.77813    0.85267  26.714   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.67 on 1839 degrees of freedom
##   (66031 observations deleted due to missingness)
## Multiple R-squared:  0.2796, Adjusted R-squared:  0.2792
## F-statistic: 713.6 on 1 and 1839 DF,  p-value: < 2.2e-16
```

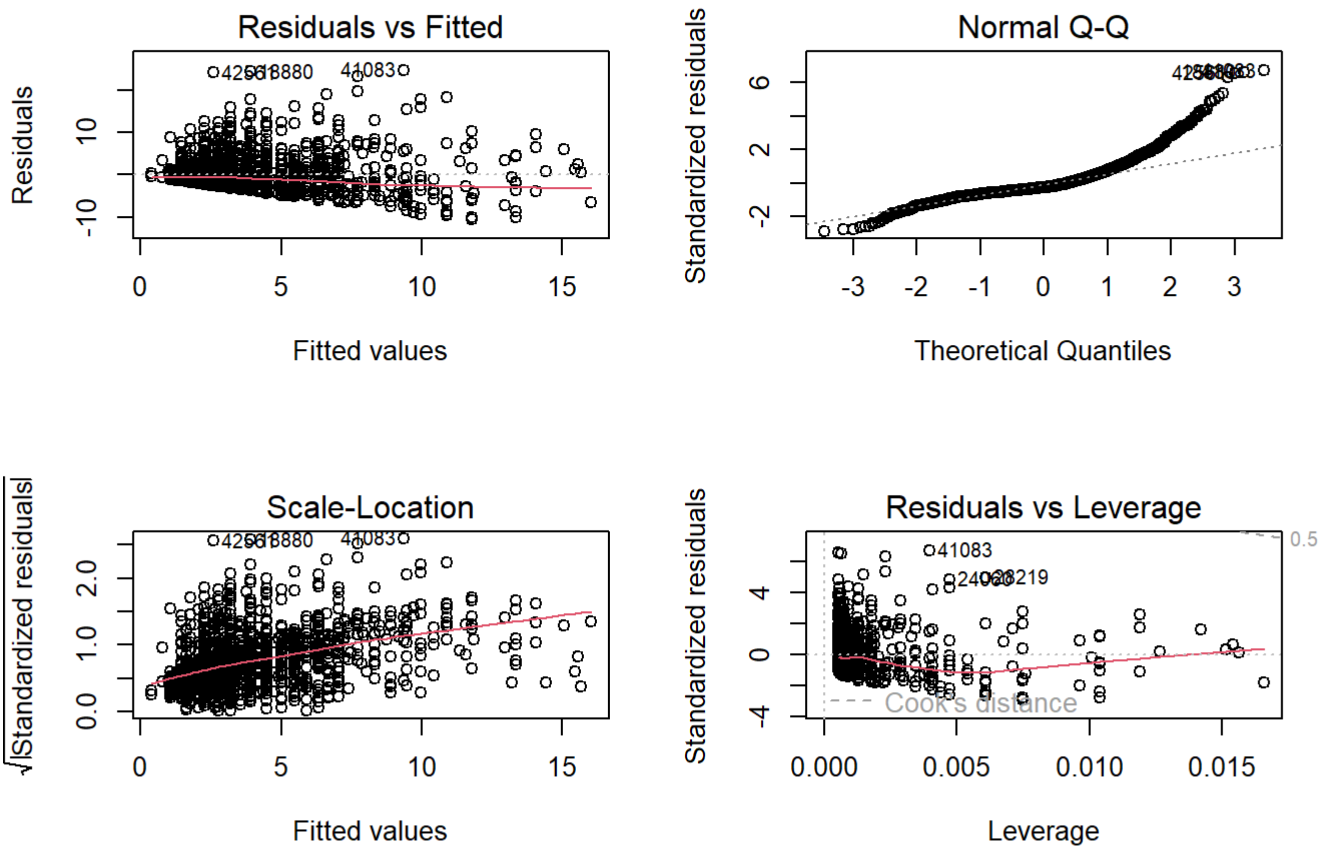
```r
plot(storm.current$TOR_LENGTH ~ sqrt(storm.current$TOR_WIDTH2),
     ylab = "Tornado Length (miles)",
     xlab = "squared Tornado Width (miles)",
     main = "Tornado Length vs Squared Tornado Width")
abline(ttorn.lmfit, col = 'red')
```



The above results show that by squaring the tornado width, the p-value for squared tornado width is also low, though the R-squared values are low as well and such result supports that there is poor linear correlation between tornado length and squared tornado width. Although there is a slight improvement of the model, the conclusion that they are linearly correlated still cannot be made.

More graphs regarding the linear model of squared tornado width and length are generated below:

```
par(mfrow = c(2, 2))
plot(ttorn.lmfit)
```

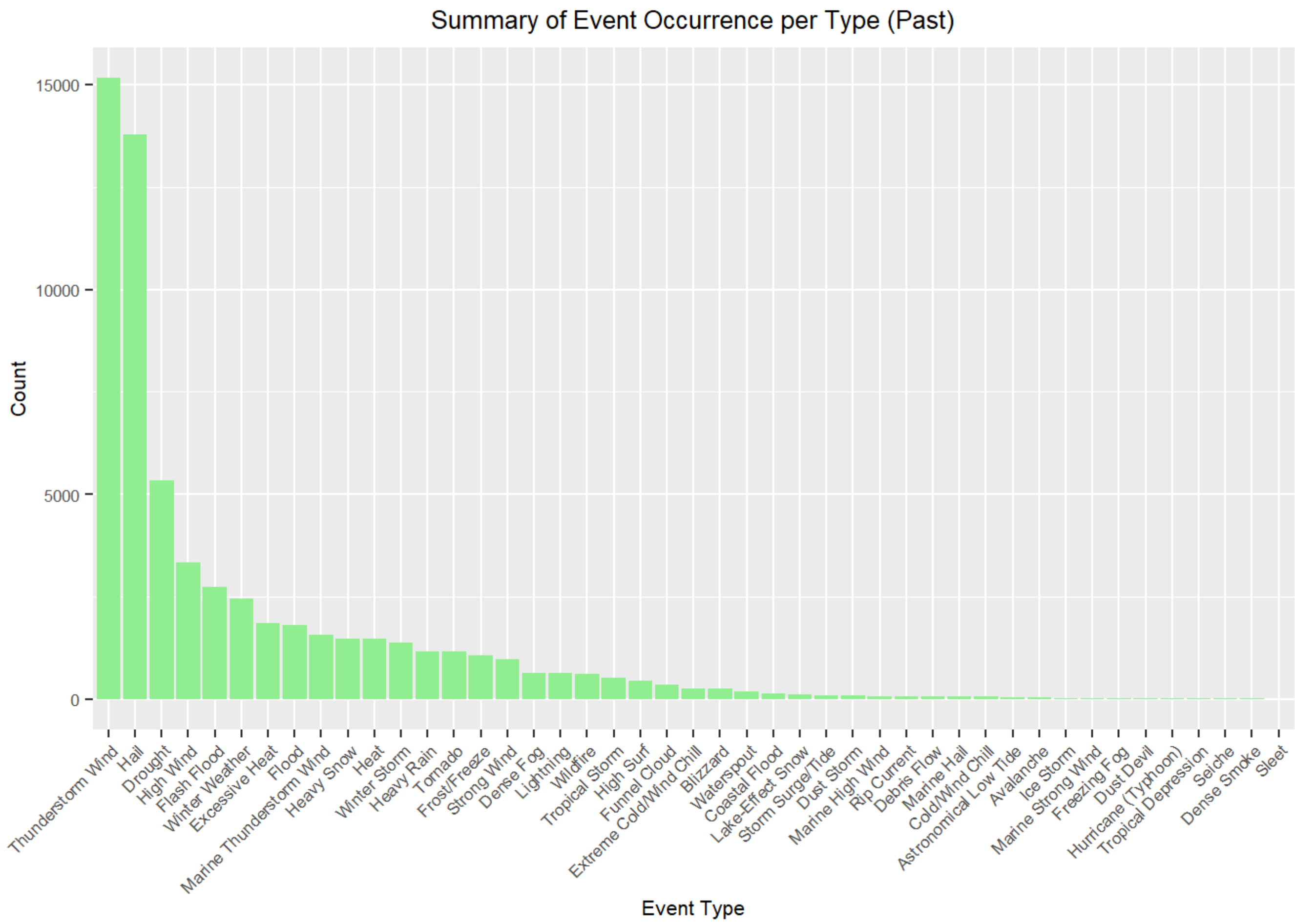


The 4 residual plots also look very similar like the previous result. It shows that taking square on tornado width does not really improve the model.

II. Comparison of the current cycle and the past cycle

Q2a: Are the top 5 events that occur the most in the current cycle same as these in the past?

```
past.event.plot <- ggplot(data = storm.past, aes(x = fct_infreq(EVENT_TYPE)))
past.event.plot + geom_bar(fill = 'lightgreen') + xlab("Event Type") + ylab("Count") + ggtitle("Su
        mmary of Event Occurrence per Type (Past)") + theme(text = element_text(size=8),
        axis.text.x = element_text(angle=45, hjust=1), plot.title = element_text(hjust = 0.5))
```



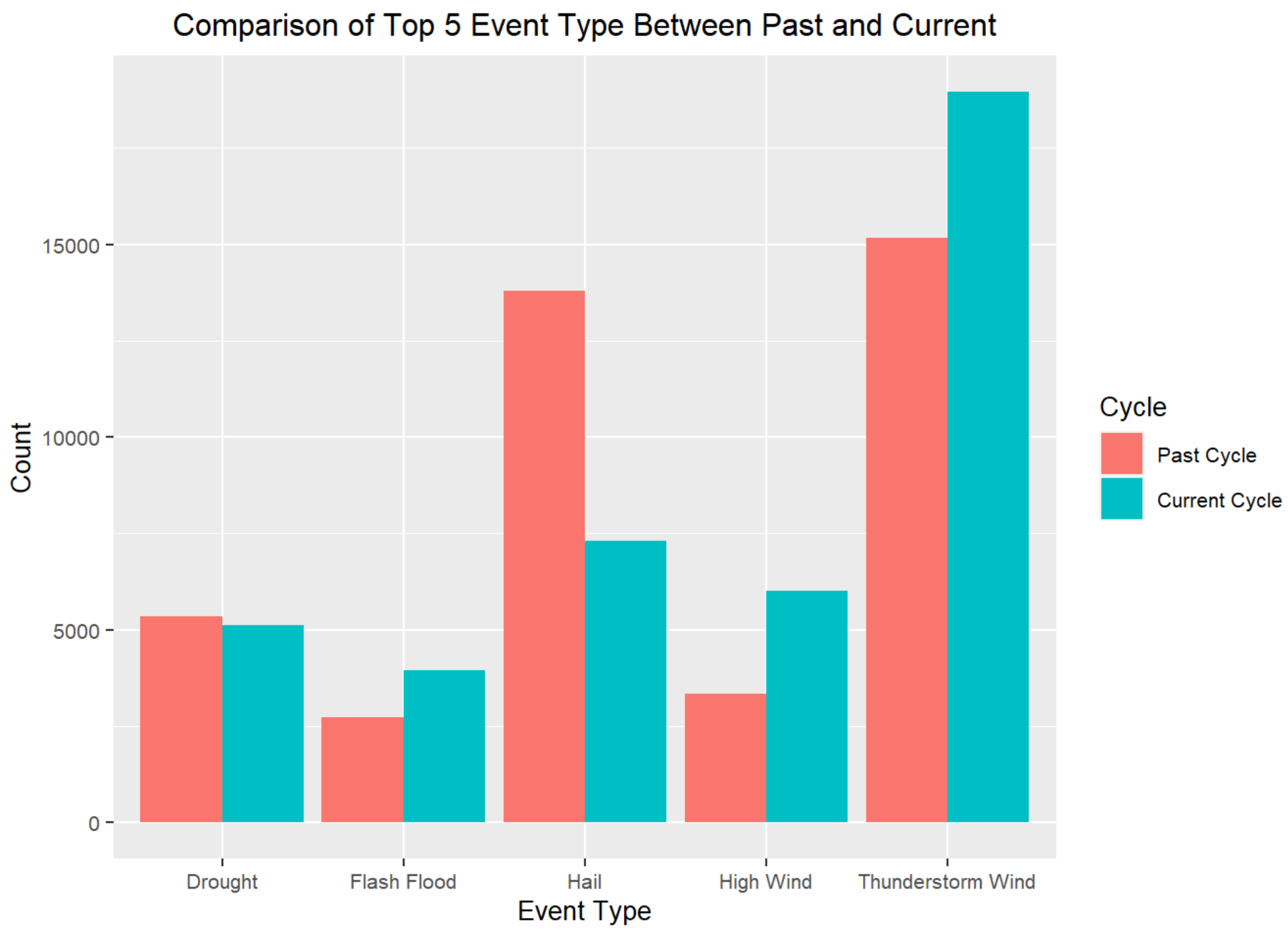

```
# a closer look on the top 5 event types
table(storm.past$EVENT_TYPE)
```

```
## 
##    Astronomical Low Tide                 Avalanche                  Blizzard 
##                       33                        31                       259 
##            Coastal Flood          Cold/Wind Chill               Debris Flow 
##                      135                        54                        61 
##                Dense Fog               Dense Smoke                   Drought 
##                      630                         4                      5344 
##               Dust Devil                Dust Storm            Excessive Heat 
##                       13                        82                      1861 
##  Extreme Cold/Wind Chill               Flash Flood                     Flood 
##                      260                      2729                      1807 
##             Freezing Fog              Frost/Freeze              Funnel Cloud 
##                       18                      1055                       341 
##                     Hail                      Heat                Heavy Rain 
##                    13793                      1463                      1168 
##               Heavy Snow                 High Surf                 High Wind 
##                     1482                       436                      3336 
##      Hurricane (Typhoon)                 Ice Storm          Lake-Effect Snow 
##                       13                        24                       108 
##                Lightning               Marine Hail          Marine High Wind 
##                      629                        57                        70 
##       Marine Strong Wind Marine Thunderstorm Wind               Rip Current 
##                       20                      1559                        64 
##                   Seiche                     Sleet          Storm Surge/Tide 
##                        9                         1                        98 
##              Strong Wind         Thunderstorm Wind                   Tornado 
##                      972                     15155                      1161 
##      Tropical Depression            Tropical Storm                Waterspout 
##                       10                       528                       177 
##                 Wildfire              Winter Storm            Winter Weather 
##                      603                      1386                      2446
```

Compared to the current cycle, whose top 5 event types are Thunderstorm Wind, Hail, High Wind, Drought, Flash Flood, the past cycle actually has the same top 5 events as it. The only difference is just the order of these 5 event types. For the past cycle, the order was Thunderstorm Wind, Hail, Drought, High Wind, Flash Flood, where the order of Drought and High Wind has swapped in the current cycle.

Let’s also do a comparison of the frequency of these 5 top events between the past cycle and the current cycle.

```r
# bind the current dataset and past dataset into a large data frame
total.raw <- rbind(storm.current.raw, storm.past.raw)
total <- subset(total.raw, select = -c(EPISODE_ID, SOURCE, CATEGORY, TOR_OTHER_WFO:DATA_SOURCE))


# for past cycle, cycle is 0; for current cycle, cycle is 1
total <- mutate(total.raw, cycle = as.factor(plyr::mapvalues(YEAR, c(2011, 2012, 2021, 2022), c('P
        ast Cycle', 'Past Cycle', 'Current Cycle', 'Current Cycle'))))


# top 5 event comparison
top5event <- c('Thunderstorm Wind', 'Hail', 'High Wind', 'Drought', 'Flash Flood')


cp.top5event.plot <- ggplot(data = total[total$EVENT_TYPE %in% top5event,], aes(x = EVENT_TYPE, fi
        ll = forcats::fct_rev(cycle)))
cp.top5event.plot + geom_bar(position="dodge") + xlab("Event Type") + ylab("Count") + ggtitle("Com
        parison of Top 5 Event Type Between Past and Current") + theme(text = element_text(size=1
        0),
       axis.text.x = element_text(angle=0), plot.title = element_text(hjust = 0.5)) + guides(fill
        = guide_legend(title = "Cycle"))
```

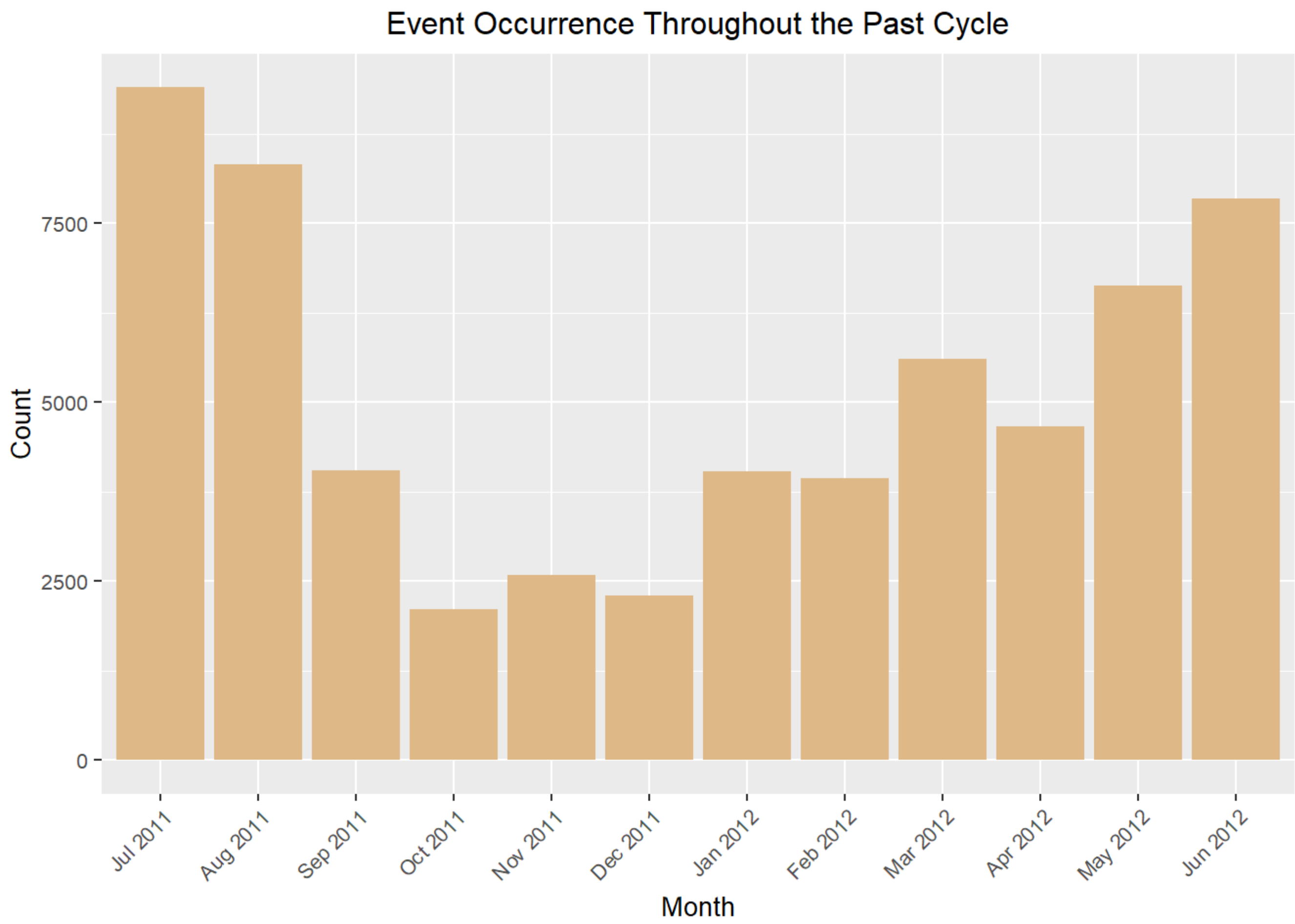


From the above comparison plot, it is noted that although the top 5 events remain the same for the current cycle compared to the past cycle, the frequency distribution of the events has changed. In short, there were more thunderstorm winds, high winds, and flash floods in this cycle, while the past cycle occurred more droughts and hails. Hail has largely decreased in the current cycle, which could be related to climate change.

Q2b: Are there any differences of the time fluctuation of event occurrence?

After comparing the event type difference, I would also like to look at whether the past cycle and the current cycle has similar time patterns. From 1b, the result shows that in the current cycle, June 2022 has occurred the most events, and November 2021 has occurred the least events. To do a comparison, a barplot is illustrated below to examine the occurrence of events through out the past cycle.

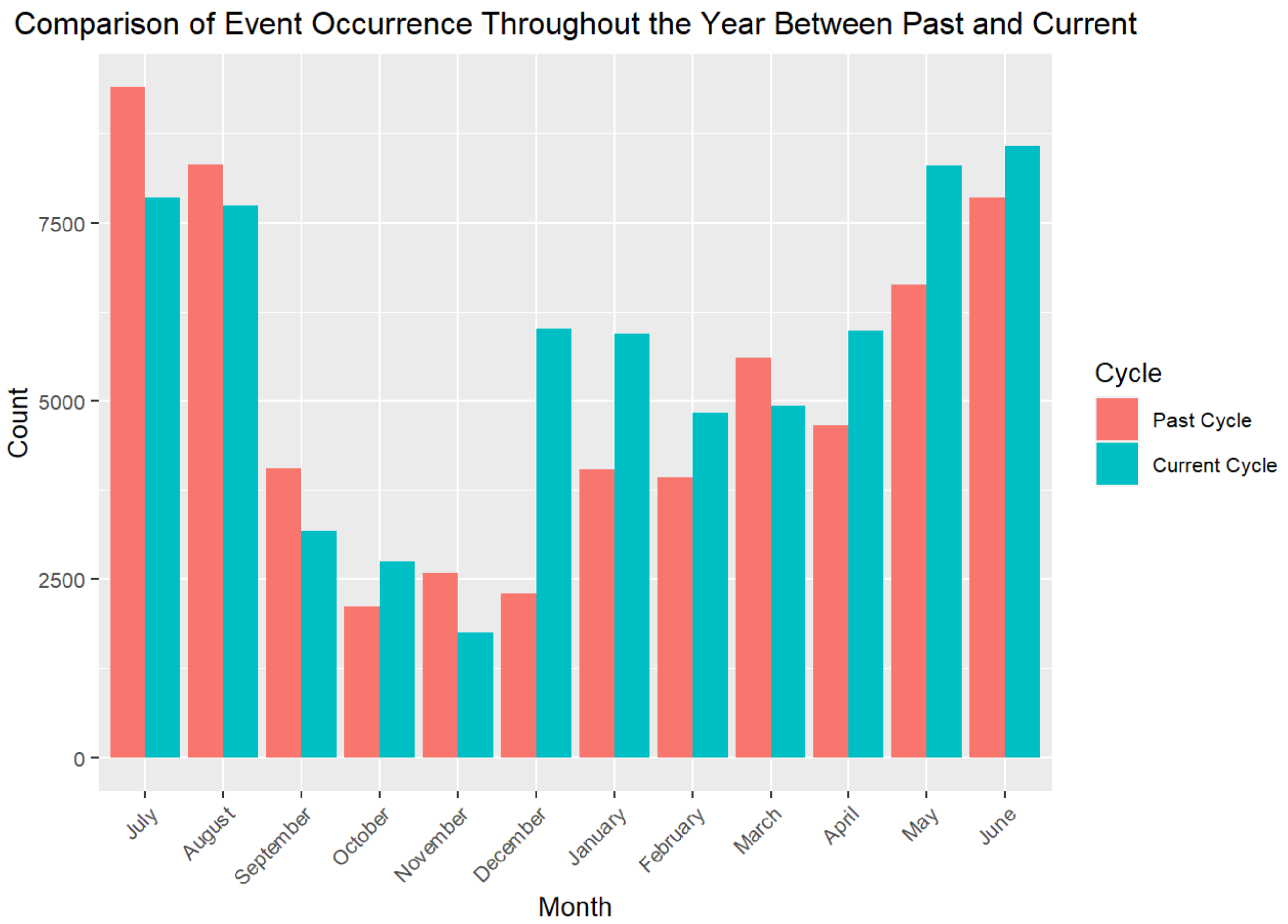
```
# plot of event occurrence throughout the past cycle
month.order <- c('July', 'August', 'September', 'October', 'November', 'December', 'January', 'Feb
        ruary', 'March', 'April', 'May', 'June')
month.labels <- c('Jul 2011', 'Aug 2011', ' Sep 2011', 'Oct 2011', 'Nov 2011', 'Dec 2011', 'Jan 20
        12', 'Feb 2012', 'Mar 2012', 'Apr 2012', 'May 2012', 'Jun 2012')
month.plot <- ggplot(data = storm.past, aes(x = MONTH_NAME))
month.plot + geom_bar(fill = 'burlywood') + xlab("Month") + ylab("Count") + ggtitle("Event Occurre
        nce Throughout the Past Cycle") + theme(text = element_text(size=10), axis.text.x = elemen
        t_text(angle=45, hjust=1), plot.title = element_text(hjust = 0.5)) + scale_x_discrete(limi
        ts = month.order, labels = month.labels)
```



Compared to 1b, it is noted that the barplot for the past cycle actually has a similar pattern like the current cycle. It is also more likely to have events during summer. For the past cycle, July 2011 has occurred the most events, and October 2011 has occurred the least events.

Similar as above, a comparison of event occurrence throughout the year for the past and current cycle is provided below.

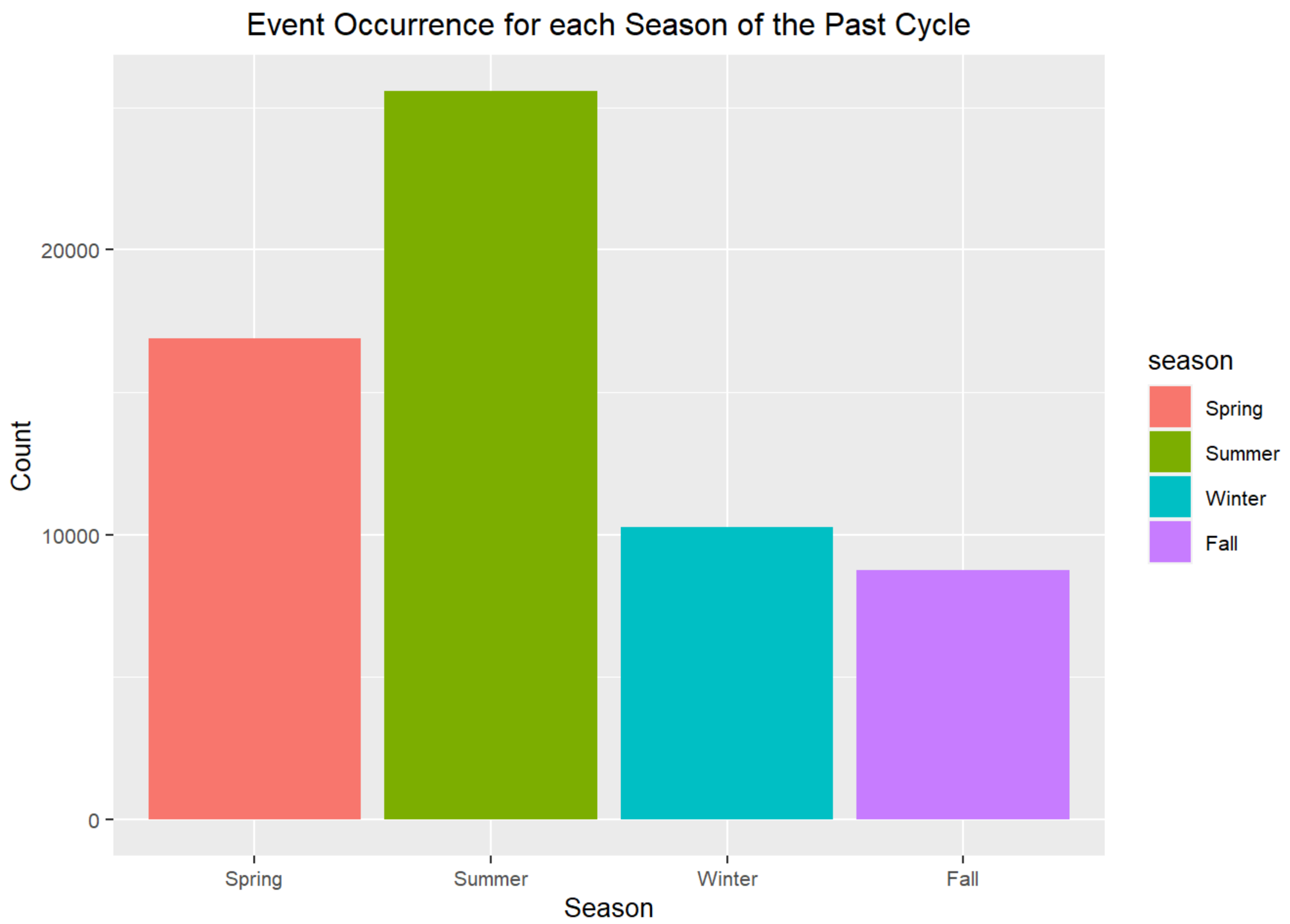
```
# use the bound total dataset from above
# for past cycle, cycle is 0; for current cycle, cycle is 1


# event occurrence comparison
month.order <- c('July', 'August', 'September', 'October', 'November', 'December', 'January', 'Feb
        ruary', 'March', 'April', 'May', 'June')


cp.month.plot <- ggplot(data = total, aes(x = MONTH_NAME, fill = forcats::fct_rev(cycle)))
cp.month.plot + geom_bar(position="dodge") + xlab("Month") + ylab("Count") + ggtitle("Comparison o
        f Event Occurrence Throughout the Year Between Past and Current") + theme(text = element_t
        ext(size=10), axis.text.x = element_text(angle=45, hjust=1), plot.title = element_text(hju
        st = 0.5)) + scale_x_discrete(limits = month.order, labels = month.order) + guides(fill =
        guide_legend(title = "Cycle"))
```



From the above figure, it is noted that there are no significant differences of event occurrence between the current cycle and the past cycle. However, it is noted that the past cycle has a higher maximum occurrence (i.e. in July) than that of the current cycle. In contrast, the current cycle has a smaller minimum occurrence (i.e. in November) compared to that of the past cycle. The largest difference of the months occurred in December, where the current cycle has around twice occurrences than the past cycle. In addition, the number of months having a higher occurrence than the other cycle is actually quite equal, with the current cycle having 7 months more occurrence than the past one (i.e. October, December, January, February, April, May, June), and the past cycle having 5 months more occurrence than the current one (i.e. July, August, September, November, March).

To take a closer look, below is a barplot for comparison of the occurrence of events for each season in the past cycle.

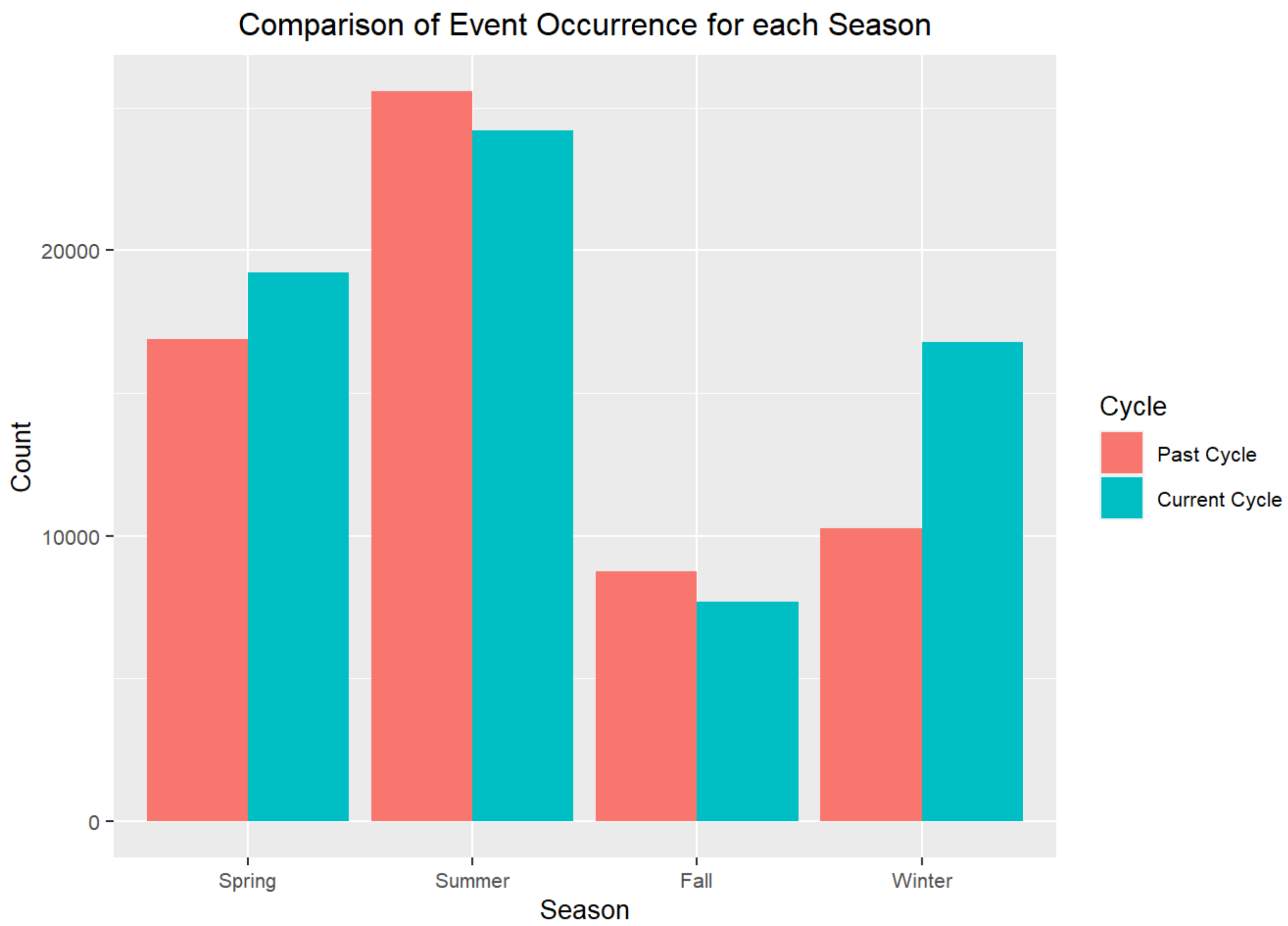
```
# assign months to each season category
storm.past.season <- mutate(storm.past,
                                    season = as.factor(plyr::mapvalues(MONTH_NAME, c('July', 'Augus
        t', 'September', 'October', 'November', 'December', 'January', 'February', 'March', 'Apri
        l', 'May', 'June'), c('Summer', 'Summer', 'Fall', 'Fall', 'Fall', 'Winter', 'Winter', 'Win
        ter', 'Spring', 'Spring', 'Spring', 'Summer'))))


# draw the barplot for each season
season.order <- c('Spring', 'Summer', 'Fall', 'Winter')
season.plot <- ggplot(data = storm.past.season, aes(x = season, fill = season))
season.plot + geom_bar() + xlab("Season") + ylab("Count") + ggtitle("Event Occurrence for each Sea
        son of the Past Cycle") + theme(text = element_text(size=10), plot.title = element_text(hj
        ust = 0.5)) + guides(fill = guide_legend(title = "Season") + scale_x_discrete(limits = sea
        son.order, labels = season.order))
```



Same as the current cycle, it is noted that summer is the season which occurs the most events and fall is the season with the least events in the past cycle as well. A slight difference would be the past cycle still has a higher occurrence in spring than fall and winter.

A comparison plot of event occurrence for each season is illustrated below.

```
# bind the current season dataset and past season dataset into a large data frame
total.season.raw <- rbind(storm.current.season, storm.past.season)


# for past cycle, cycle is 0; for current cycle, cycle is 1
total.season <- mutate(total.season.raw,
                        season = as.factor(plyr::mapvalues(MONTH_NAME, c('July', 'August', 'Septemb
        er', 'October', 'November', 'December', 'January', 'February', 'March', 'April', 'May', 'J
        une'), c('Summer', 'Summer', 'Fall', 'Fall', 'Fall', 'Winter', 'Winter', 'Winter', 'Sprin
        g', 'Spring', 'Spring', 'Summer'))),
                         cycle = as.factor(plyr::mapvalues(YEAR, c(2011, 2012, 2021, 2022), c('Past
        Cycle', 'Past Cycle', 'Current Cycle', 'Current Cycle'))))


# seasonal event occurrence comparison
season.order <- c('Spring', 'Summer', 'Fall', 'Winter')
cp.season.plot <- ggplot(data = total.season, aes(x = season, fill = forcats::fct_rev(cycle)))
cp.season.plot + geom_bar(position="dodge") + xlab("Season") + ylab("Count") + ggtitle("Comparison
        of Event Occurrence for each Season") + theme(text = element_text(size=10), plot.title = e
        lement_text(hjust = 0.5)) + scale_x_discrete(limits = season.order, labels = season.order)
        + guides(fill = guide_legend(title = "Cycle"))
```

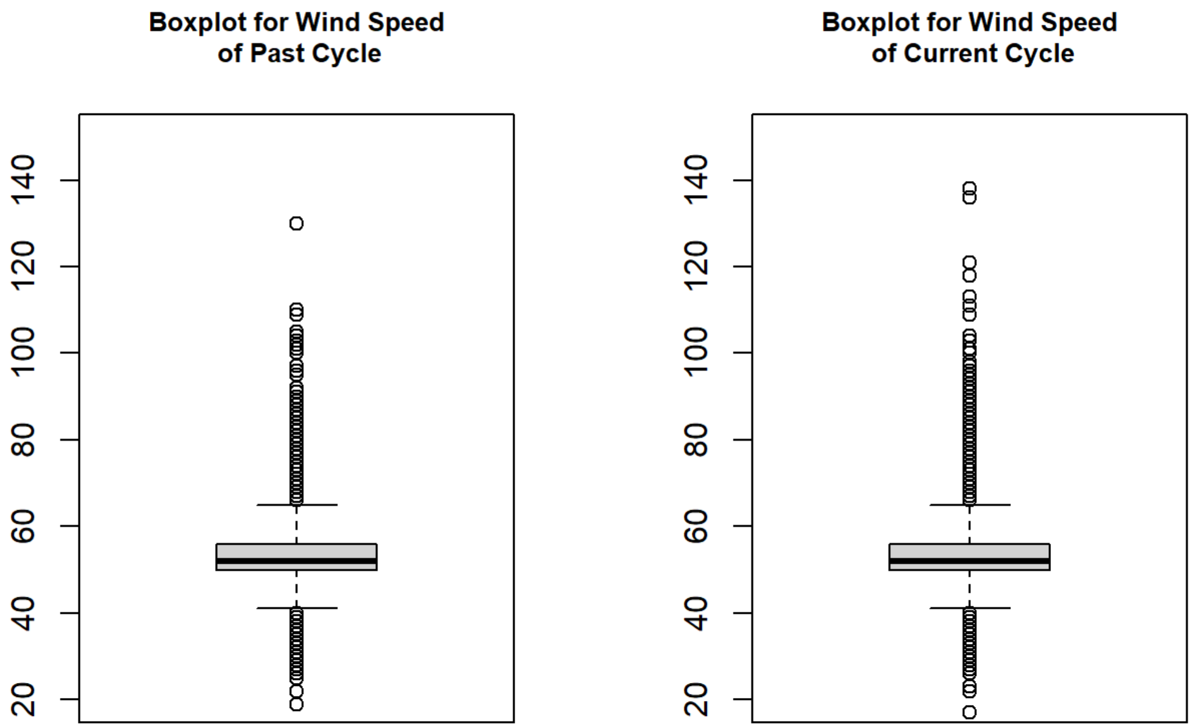


From the above figure, it is noted that both cycles have similar pattern, with summer as their peak for events. Compared to the past, the current cycle has less events in Summer and Fall, but more events in Spring and Winter.

Q2c: Has the wind speed changed throughout the years?

For high winds, marine high winds, marine strong winds, marine thunderstorm winds, strong winds, and thunderstorm winds, the magnitude field measures the wind speeds (in knots). To understand if wind speed has increased or decreased throughout the years, below outlines 2 box plots of wind speed for comparison.

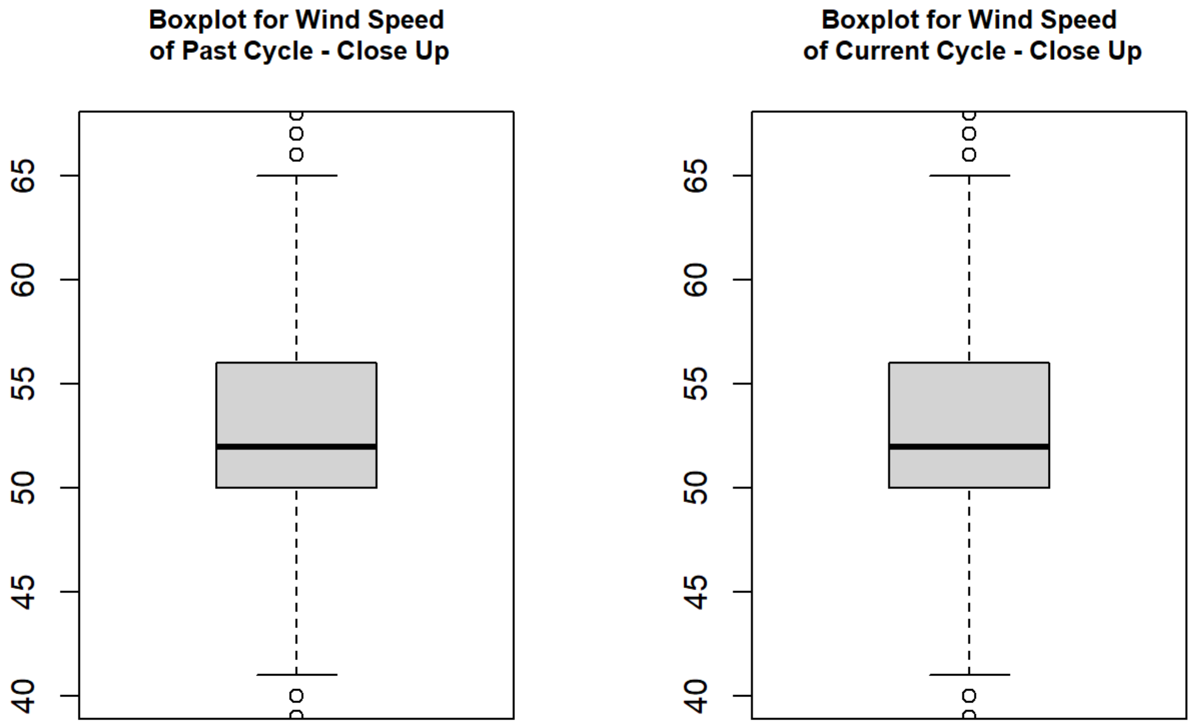
```
wind.types <- c("High Wind", "Marine High Wind", "Marine Strong Wind", "Marine Thunderstorm Wind",
         "Strong Wind", "Thunderstorm Wind")
storm.current.wind <- filter(storm.current, EVENT_TYPE %in% wind.types)
storm.past.wind <- filter(storm.past, EVENT_TYPE %in% wind.types)


par(cex.main = 0.8, mfrow = c(1, 2))
# boxplot for wind speed of past and current cycle
boxplot(storm.past.wind$MAGNITUDE, ylim = c(20, 150), main = "Boxplot for Wind Speed\n of Past Cyc
         le")
boxplot(storm.current.wind$MAGNITUDE, ylim = c(20, 150), main = "Boxplot for Wind Speed\n of Curre
         nt Cycle")
```



From the box plots above, it is noted that both box plots generate similar minimum value, 1st quartile, median, 3rd quartile, and maximum value, the current cycle has more outliers. The current cycle has more records with wind speed larger than 100, and the box plot for past cycle shows that in the past cycle, wind speed varied less and was more concentrated.

While the outliers vary a lot over range, below outlines a closer look for the box plot data points.

```
par(cex.main = 0.8, mfrow = c(1, 2))
# boxplot for wind speed of past and current cycle - close up
boxplot(storm.past.wind$MAGNITUDE, ylim = c(40, 67), main = "Boxplot for Wind Speed\n of Past Cycl
        e - Close Up")
boxplot(storm.current.wind$MAGNITUDE, ylim = c(40, 67), main = "Boxplot for Wind Speed\n of Curren
        t Cycle - Close Up")
```



By taking a closer look, it is noted that the minimum value, 1st quartile, median, 3rd quartile, and maximum value data points are quite close to each other in both cycles. Both of them have a median around 52. This implicates that although the wind speed for the current cycle is more scattered out, the distribution of wind speed is still similar to the past cycle.

Q2d: Has the hail size changed throughout the years?

For hails and marine hails, the magnitude field measures the hail size (in inches to the hundredth). To understand if the hail size has increased or decreased throughout the years, below outlines a comparison of the hail size distributions.

```
# define a function for calculation of basic statistics
basic.statistics <- function(x) {
  c(minimum = min(x),
    median = median(x),
    mean = mean(x),
    maximum = max(x),
    stddev = sd(x))
}

hail.types <- c("Hail", "Marine Hail")
storm.past.hail <- filter(storm.past, EVENT_TYPE %in% hail.types)
storm.current.hail <- filter(storm.current, EVENT_TYPE %in% hail.types)

# get hail statistics
hail.past.stats <- basic.statistics(storm.past.hail$MAGNITUDE) %>% round(3)
hail.current.stats <- basic.statistics(storm.current.hail$MAGNITUDE) %>% round(3)

# create a data frame for easier examination
data.frame(hail.past.stats, hail.current.stats)
```

```
##         hail.past.stats hail.current.stats
## minimum           0.250              0.250
## median            1.000              1.000
## mean              1.173              1.256
## maximum           5.000              6.000
## stddev            0.483              0.520
```

From the above data frame, it is noted that there is a slight increase in hail size in the current cycle. Although both cycles have the same minimum and median size, since the maximum size increased in the current cycle, the mean also increased by around 7 percent (i.e. (1.256 - 1.173) / 1.173 = 0.07). There is also a slight increase in the standard deviation.

III. Prediction using long term time series data

Q3a: Can we predict the tornado width with tornado length for the next 20 tornadoes using previous data?

Since I consolidated datasets from January 2010 till June 2022, I would like to predict the tornado width using tornado length for the future 20 tornadoes. Before I start to do modelling, I would like to check how many data points do I have to make sure there is sufficient data for training.

```
# check number of tornado length / width values in the consolidated data set
sum(!is.na(storm.consolidated$TOR_LENGTH))
```

```
## [1] 17513
```

```
sum(!is.na(storm.consolidated$TOR_WIDTH))
```

```
## [1] 17513
```

From the above result, it seems that the data is quite sufficient, and a training model can be built. By inspection, it is noted that the TOR_LENGTH and TOR_WIDTH values are recorded together, therefore, if one of them has value, the other one would also have value.

```
# load caret
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

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

```
# drop the na values first, only need to do this one either TOR_LENGTH or TOR_WIDTH
torn.train <- storm.consolidated %>% drop_na(TOR_LENGTH)

# define train control parameters
ctrl <- trainControl(method = 'repeatedcv',
                     repeats = 5)

# define parameters to be used for linear model
torn.lmtrain <- train(TOR_LENGTH ~ TOR_WIDTH,
                           data = torn.train,
                           method = "lm",
                           trControl = ctrl)
torn.lmtrain
```

```
## Linear Regression
##
## 17513 samples
##     1 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 15762, 15763, 15762, 15762, 15762, 15762, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   3.583192  0.2921491  2.336781
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
torn.preds <- predict(torn.lmtrain)
head(torn.preds, 20) %>% round(2)
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12    13
##  2.25  2.25  2.06  2.83  1.90  2.19  1.75  2.45  4.00 10.09 10.09  2.45  3.73
##    14    15    16    17    18    19    20
##  2.45  2.45  2.06  2.06  3.22  1.87  2.45
```

By leveraging caret library, 20 tornado widths has been predicted based on the relationship with tornado lengths from the consolidated dataset. However, it is noted that the R-squared value is rather small and RMSE / MAE results do not look good.

Another training model constructed with k-Nearest Neighbors is illustrated below.

```
# define train control parameters
ctrl <- trainControl(method = 'repeatedcv',
                     repeats = 5)

# define parameters to be used for linear model
torn.knnfit <- train(TOR_WIDTH ~ TOR_LENGTH ,
                           data = torn.train,
                           method = "knn",
                           trControl = ctrl)
torn.knnfit
```

```
## k-Nearest Neighbors
##
## 17513 samples
##     1 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 15762, 15762, 15761, 15764, 15761, 15761, ...
## Resampling results across tuning parameters:
##
##   k  RMSE      Rsquared   MAE
##   5  264.7876  0.2254318  144.9430
##   7  260.3439  0.2438890  143.5472
##   9  257.5687  0.2556710  142.6767
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 9.
```

```
torn.preds <- predict(torn.knnfit)
head(torn.preds, 20) %>% round(2)
```

```
##  [1] 143.64  90.29  96.88  82.79  50.09 179.08  46.55  46.55 317.38 569.44
## [11] 449.70  51.82 279.18 581.82 609.64 241.67 133.79 109.70  64.58 179.61
```

From the above result, it is noted that KNN also did not do a good job, though it could be because of the poor explainability of tornado width. The model also had poor R-squared, RMSE, and MAE values.