Tornadoes Analysis

Abigail Castro, Maggie Salomonsky, Aditi Jain, Sarah Saas, Ethan Thurmond

Note: Divide the task among yourself. Each team member should contribute to at least one part of the assignment. The person whose name starts last in the alphabetic order shares screen and compiles the Rmd file.

Tidy Tornadoes

The US Storm Prediction Center make severe weather data available from the website http://www.spc.noaa.gov/wcm/#data. This data is used by insurance companies to help with their claims evaluation and forecasting. A description of the data can be found http://www.spc.noaa.gov/wcm/data/SPC_severe_database_description.pdf.

1. Download the tornado event data and import it in R

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
tornado <- read.csv(file = 'tornado.csv')
```

2. Calculate the number of tornadoes by *year* and *Fujita score* (`f`)

```
d1<- tornado %>%  group_by(yr, f) %>%
     summarise(count = n())
```

```
## `summarise()` has grouped output by 'yr'. You can override using the `.groups` argument.
```

```
d1
```

```
## # A tibble: 49 x 3
## # Groups:   yr [9]
##       yr f     count
##    <int> <chr> <int>
##  1  2007 EF-0    681
##  2  2007 EF-1    306
##  3  2007 EF-2     97
##  4  2007 EF-3     27
##  5  2007 EF-4      4
##  6  2007 EF-5      1
##  7  2008 EF-0    997
##  8  2008 EF-1    515
##  9  2008 EF-2    158
## 10  2008 EF-3     56
## # ... with 39 more rows
```

3. Convert the results to a table (use pivot_wider). Note: Some years have 0 EF-5 tornadoes. The final result should look like this

year	EF-0	EF-1	EF-2	EF-3	EF-4	EF-5
2007	681	306	97	27	4	1
2008	997	515	158	56	11	1
2009	709	355	94	21	3	0
2010	776	351	129	42	17	0
2011	821	638	212	72	25	9
2012	577	242	100	32	5	0
2013	508	314	86	22	8	1
2014	478	325	76	20	7	0
2015	704	415	69	19	5	0

```r
d1 <- d1 %>%
    pivot_wider(names_from = f, values_from = count)
d1[is.na(d1)] <- 0
d1
```

```
## # A tibble: 9 x 7
## # Groups:   yr [9]
##      yr ‘EF-0‘ ‘EF-1‘ ‘EF-2‘ ‘EF-3‘ ‘EF-4‘ ‘EF-5‘
##   <int>  <int>  <int>  <int>  <int>  <int>  <int>
## 1  2007    681    306     97     27      4      1
## 2  2008    997    515    158     56     11      1
## 3  2009    709    355     94     21      3      0
## 4  2010    776    351    129     42     17      0
## 5  2011    821    638    212     72     25      9
## 6  2012    577    242    100     32      5      0
## 7  2013    508    314     86     22      8      1
## 8  2014    478    325     76     20      7      0
## 9  2015    704    415     69     19      5      0
```

4. What is the type of EF variables? How do you determine that?

```
d1
```

```
## # A tibble: 9 x 7
## # Groups:   yr [9]
##        yr `EF-0` `EF-1` `EF-2` `EF-3` `EF-4` `EF-5`
##     <int>  <int>  <int>  <int>  <int>  <int>  <int>
## 1  2007    681    306     97     27      4      1
## 2  2008    997    515    158     56     11      1
## 3  2009    709    355     94     21      3      0
## 4  2010    776    351    129     42     17      0
## 5  2011    821    638    212     72     25      9
## 6  2012    577    242    100     32      5      0
## 7  2013    508    314     86     22      8      1
## 8  2014    478    325     76     20      7      0
## 9  2015    704    415     69     19      5      0
```

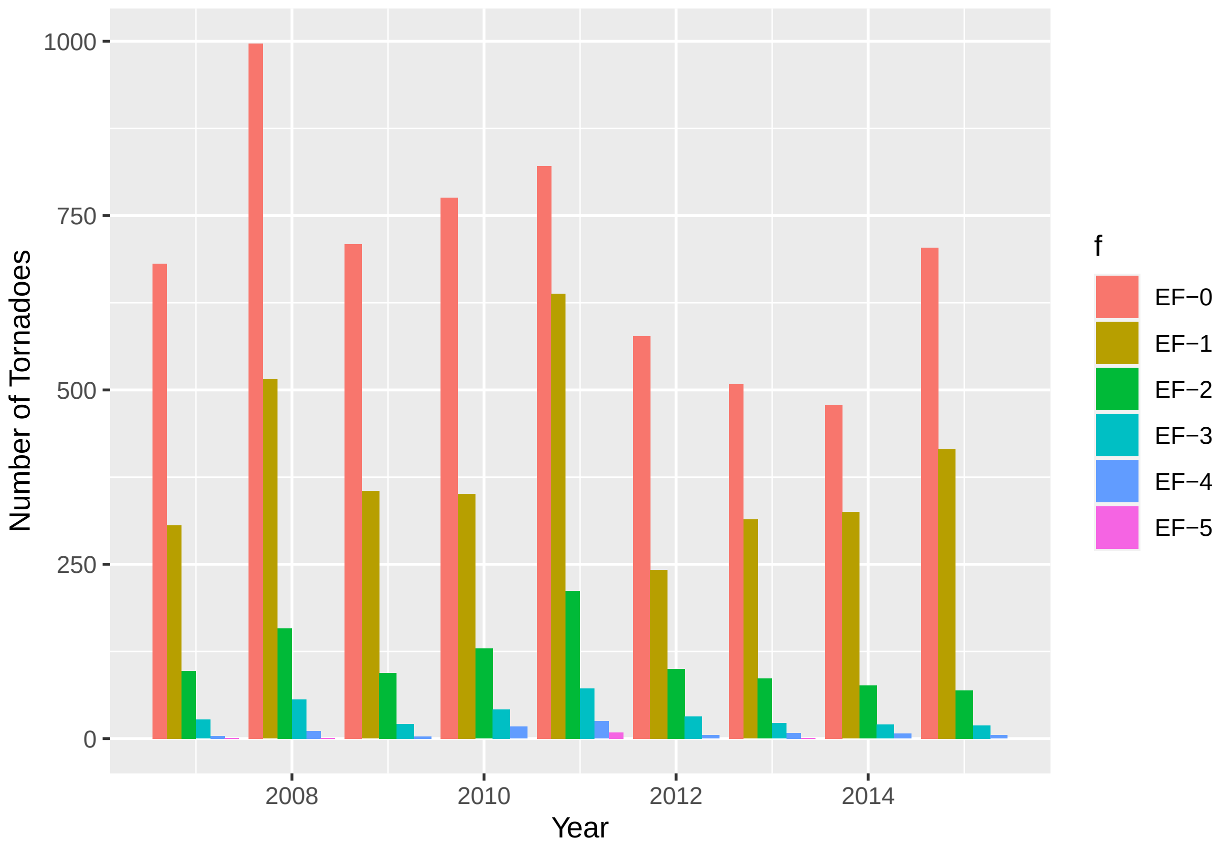
By using the above command, you can see that the EF variables are of type int.

5. Compare different types of tornadoes in terms of 1) frequency of occurrence and 2) trends over the years. What type of plots do you need to use? How do you determine that? Look at the data types to see if you need to convert the variable types. Since the year is a categorical variable and count is an int, we should use bar charts to visualize the data. There is no need to convert the variable types.

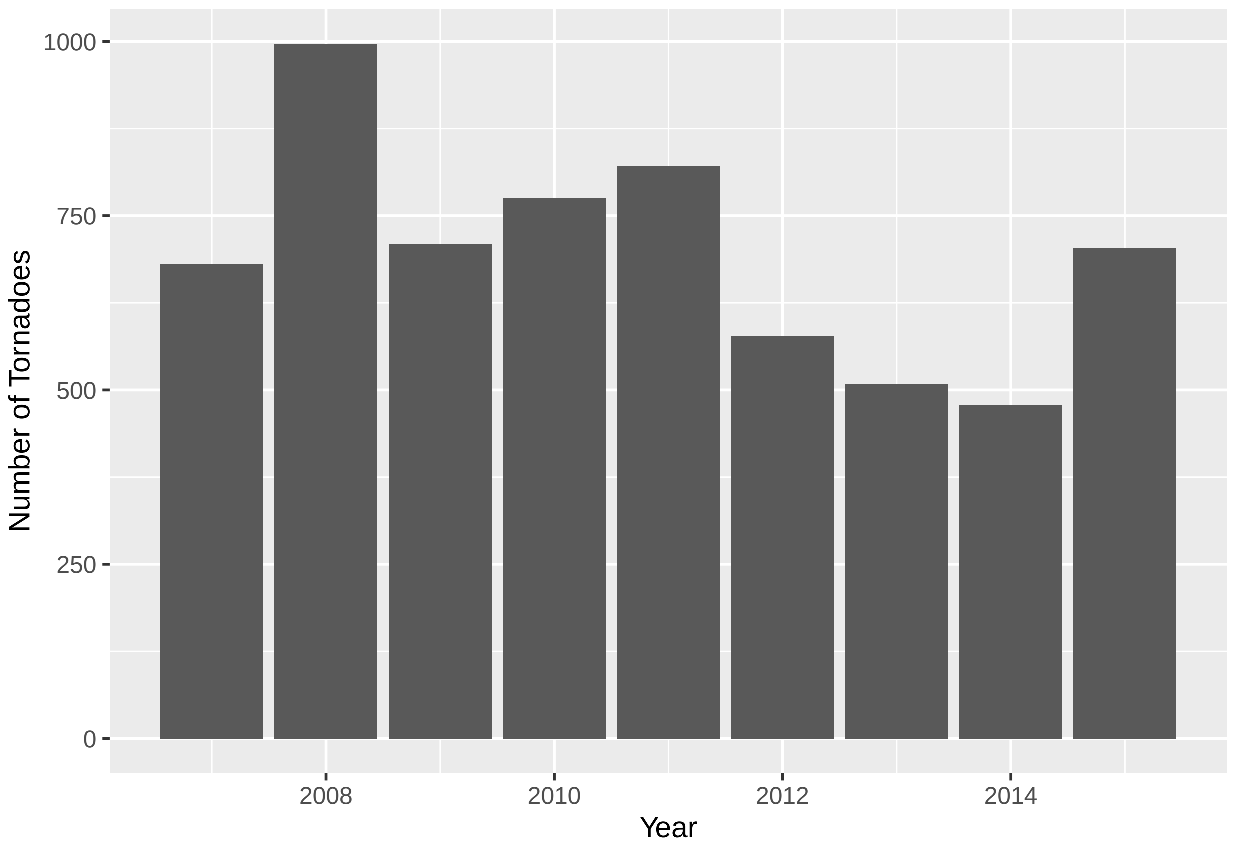
```
d2 <- tornado %>%  group_by(yr, f) %>%
     summarise(count = n())
```

```
## `summarise()` has grouped output by 'yr'. You can override using the `.groups` argument.
```

```
ggplot(d2, aes(yr, count , fill= f)) +
   geom_bar(stat="identity", position = "dodge") +
   labs(x= 'Year', y= 'Number of Tornadoes')
```



```
ggplot(d2, aes(yr, count)) +
   geom_bar(stat="identity", position = "dodge") +
   labs(x= 'Year', y= 'Number of Tornadoes')
```



6. Describe your observations. Generally, tornado frequency has decreased since 2008. There is no apparent correlation between severity of tornados and year, however, within each year there is a negative correlation between severity of tornados and frequency of tornados.

Time of occurance

The `time` column in the `tornado` data gives the time-of-day (24 hour clock, central time zone) when the tornado occurred. Ignoring the time zone issue, create a density plot of the fractional hour when tornadoes occur.

1. Use the `separate()` function to create three new columns (*hour*, *min*, *sec*) from the `time` column.

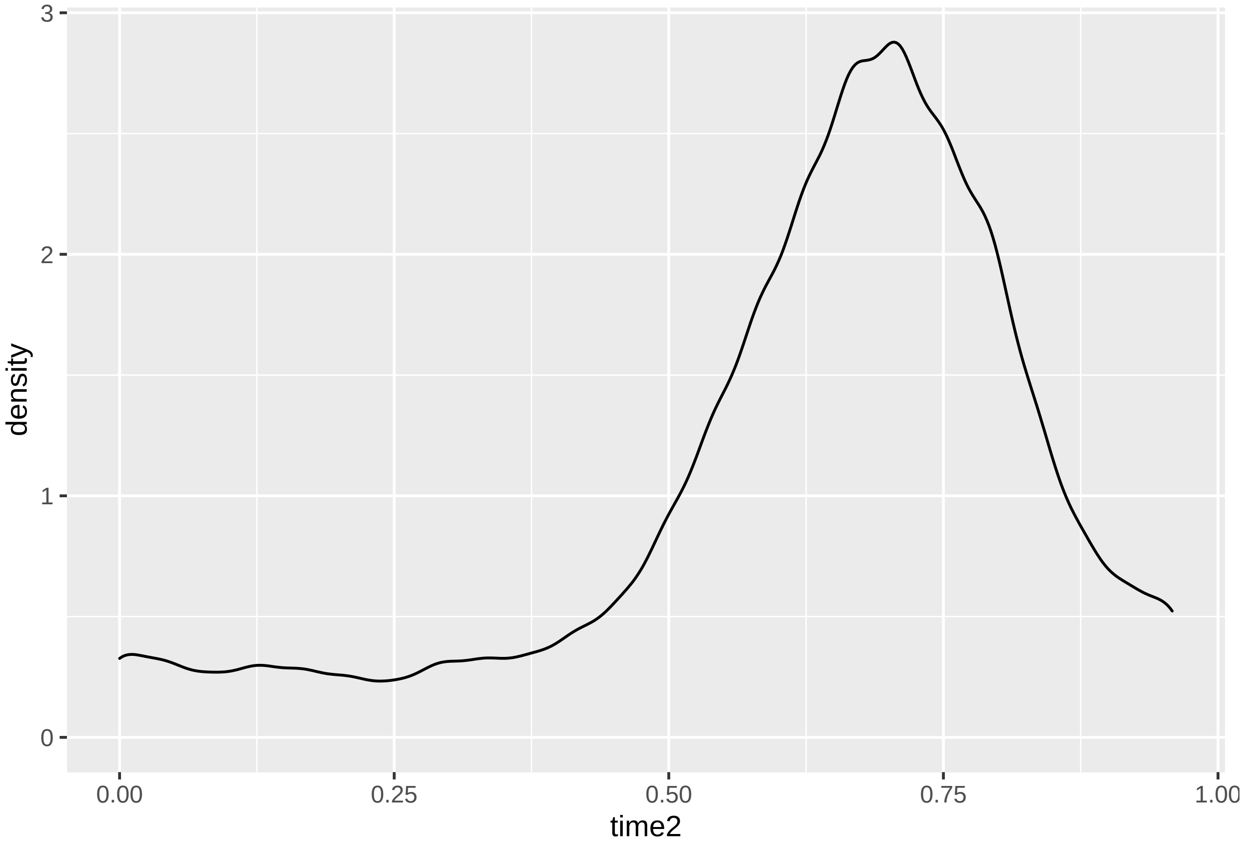
```
tornado_data_sep <- tornado %>%
  separate(time, into = c("hour", "min", "sec"), convert = TRUE)
```

2. Add another column, named `time2`, that gives the fractional number of hours that a tornado occurred.

```
tornado_data_sep <- tornado_data_sep %>% mutate(time2 = hour/24)
```

3. Generate a density plot of `time2`. Are there any differences by severity?

```
ggplot(tornado_data_sep, aes(x=time2)) + geom_density()
```



4. Describe your observations. As shown in the density plot there are certain hours of the day where tornados are more likely to occur.