Exploring Data 01

Ian Vert

7/16/2021

This is an analysis of a Tornado Data set from Kaggle

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     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()
```

```
library(ggplot2)
```

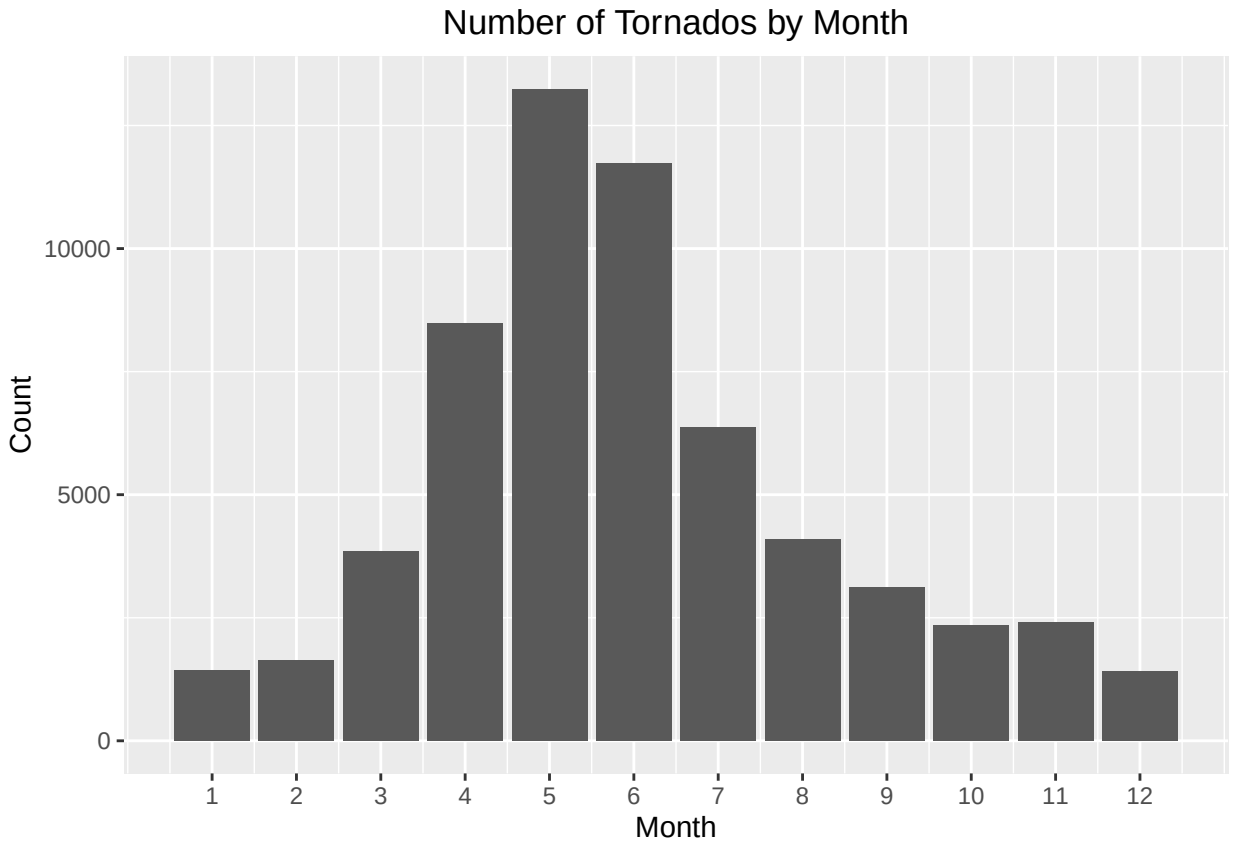
```
df <- read.csv('Tornadoes_SPC_1950to2015.csv')

head(df)
```

```
##   om   yr mo dy      date     time tz st stf stn mag inj fat loss closs  slat
## 1  1 1950  1  3  1/3/1950 11:00:00  3 MO  29   1   3   3   0    6     0 38.77
## 2  2 1950  1  3  1/3/1950 11:55:00  3 IL  17   2   3   3   0    5     0 39.10
## 3  3 1950  1  3  1/3/1950 16:00:00  3 OH  39   1   1   1   0    4     0 40.88
## 4  4 1950  1 13 1/13/1950  5:25:00  3 AR   5   1   3   1   1    3     0 34.40
## 5  5 1950  1 25 1/25/1950 19:30:00  3 MO  29   2   2   5   0    5     0 37.60
## 6  6 1950  1 25 1/25/1950 21:00:00  3 IL  17   3   2   0   0    5     0 41.17
##     slon  elat   elon len wid fc
## 1 -90.22 38.83 -90.03 9.5 150  0
## 2 -89.30 39.12 -89.23 3.6 130  0
## 3 -84.58  0.00   0.00 0.1  10  0
## 4 -94.37  0.00   0.00 0.6  17  0
## 5 -90.68 37.63 -90.65 2.3 300  0
## 6 -87.33  0.00   0.00 0.1 100  0
```

Key Question: What months have the highest frequency of tornadoes?

```
p <- df %>% group_by(mo) %>% ggplot(mapping=aes(x = mo )) +
  geom_bar() + scale_x_continuous(breaks = seq(1,12,by = 1)) +
  labs(x = 'Month',
       y = 'Count',
       title = 'Number of Tornados by Month') +
  theme(plot.title = element_text(hjust = 0.5))
```

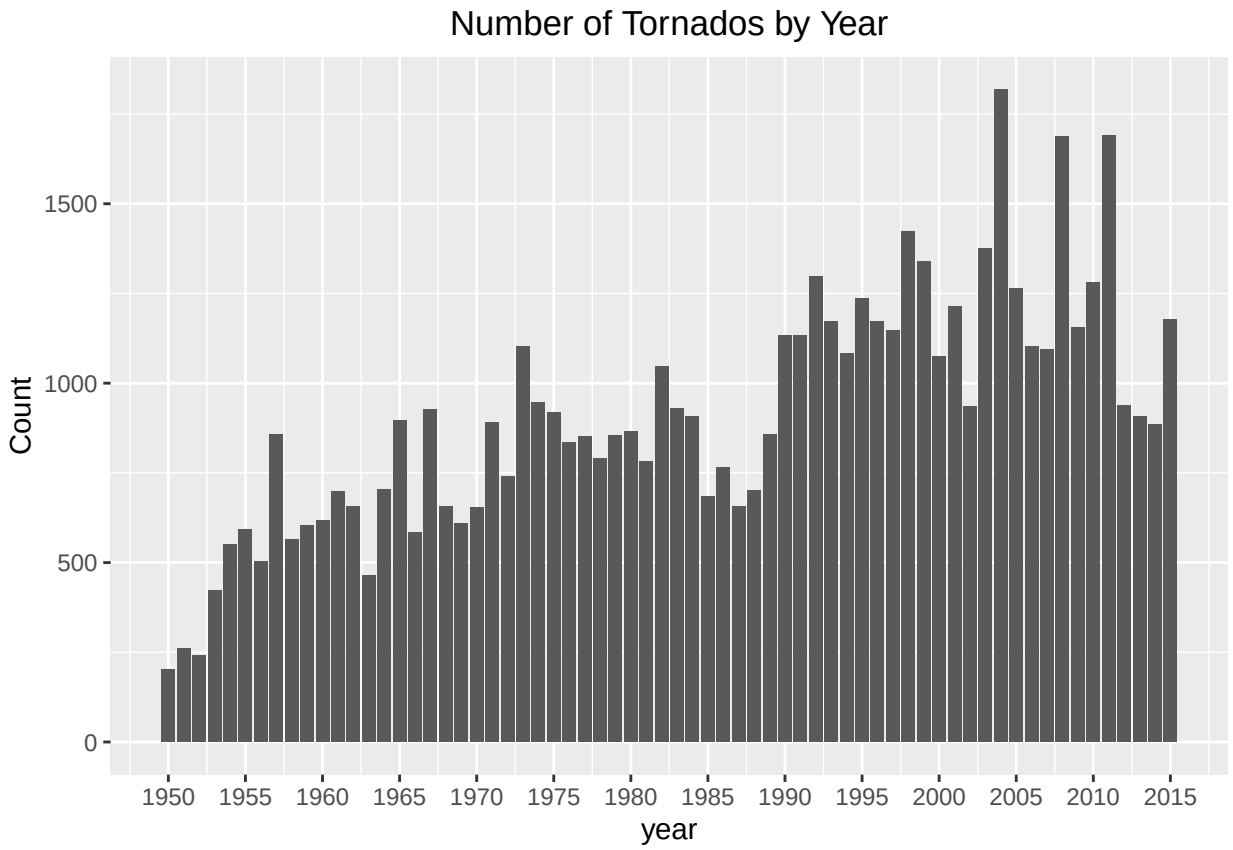
```
plot(p)
```



Looking at the graph above the spring time months of April, May, June are the most common months for Tornado activity.

Key Question : How have the number of Tornados changed over time?

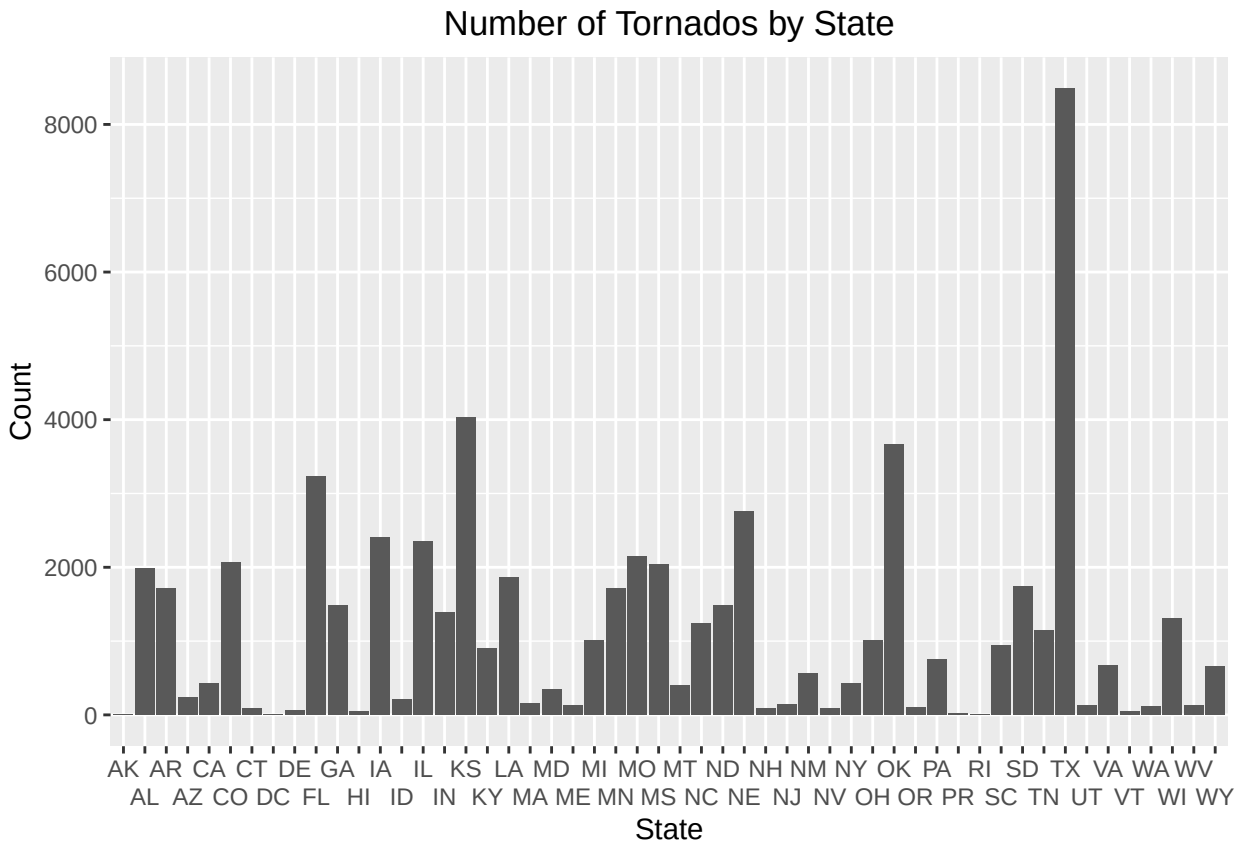
```
p <- df %>% group_by(yr) %>% ggplot(mapping=aes(x = yr)) +
  geom_bar() + scale_x_continuous(breaks = seq(1950,2015,by = 5)) +
  labs(x = 'year',
       y = 'Count',
       title = 'Number of Tornados by Year') +
  theme(plot.title = element_text(hjust = 0.5))


plot(p)
```



The number of Tornados per year is increasing over time.

Key Question: What states have the most tornados?

```
p <- df %>% group_by(st) %>% ggplot(mapping=aes(x = st)) +
  geom_bar() +
  labs(x = 'State',
       y = 'Count',
       title = 'Number of Tornados by State') +
  scale_x_discrete(guide =  guide_axis(n.dodge = 2)) +
  theme(plot.title = element_text(hjust = 0.5))



plot(p)
```



The states that have portion of land in what commonly called 'tornado valley', such as Iowa, Nebraska, Kansas, Oklahoma, and Texas are some of the states with the most Tornados. While not in tornado valley, Florida is not surprising as it is a common spot for hurricanes.

Here I create a a dataset of only the 'Tornado Valley' states

```
df_valley <- df %>% filter(st == 'TX' |
                           st == 'OK' |
                           st =='KS'  |
                           st == 'IA' |
                           st == 'NE')
```

Key Question: Are tornados increasing just in Tornado valley or everywhere?

To answer this key question I will take two different decades as timepoints, the 1970s and the 2000s, and take the mean number of tornados of each state in the two decades. Then I will use percent change to determine what states were increasing or decreasing between the 1970s and the 2000s.

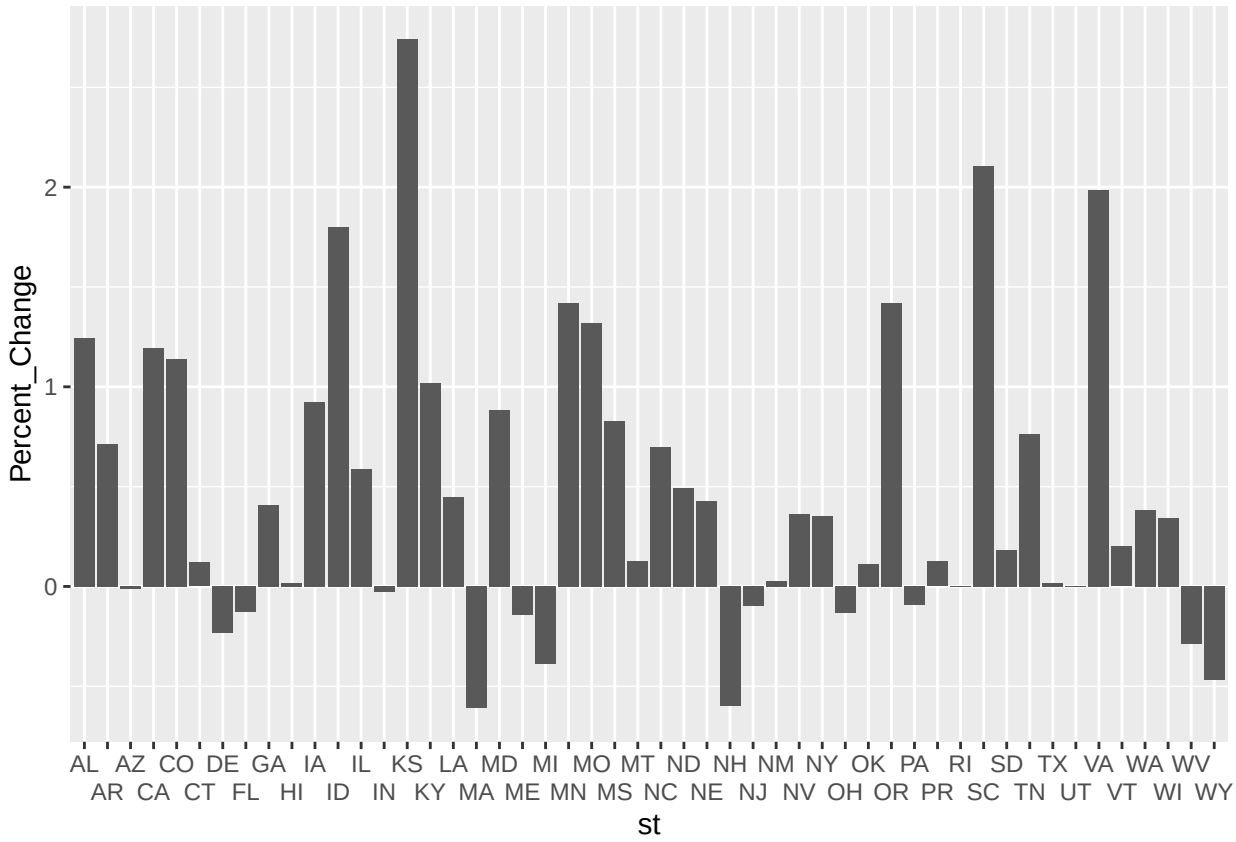
```
df_70s <- df %>% filter(yr %in% 1970:1979) %>% group_by(st) %>% count(yr) %>% mutate(mean(n))

df_00s <- df %>% filter(yr %in% 2000:2009) %>% group_by(st) %>% count(yr) %>% mutate(mean(n))
```

```
df_merge <- merge(df_70s, df_00s, by= 'st')

df_merge <- df_merge %>% mutate(`mean(n).y` - `mean(n).x`)

colnames(df_merge)[8] <- 'Mean_Difference'
```

```
df_merge$Percent_Change <- ((df_merge$Mean_Difference / df_merge$`mean(n).x`) * 100)


df_change <- unique(subset(df_merge, select = c( st , Percent_Change)))

head(df_change)
```

```
##     st Percent_Change
## 1   AL     124.242424
## 101 AR      71.264368
## 201 AZ      -1.098901
## 271 CA     119.576720
## 361 CO     113.793103
## 461 CT      12.000000
```

```
p <- df_change %>% ggplot(aes(x = st, y = Percent_Change)) +
  geom_col() +
  scale_x_discrete(guide =  guide_axis(n.dodge = 2)) +
  scale_y_continuous(labels=function(x)x/100) +
  theme(plot.title = element_text(hjust = 0.5))

plot(p)
```



The above graph is showing the percent change in the mean per state from 1970s and the 2000s. The graph shows that tornados or not just increasing in the 'Tornado Valley' states.

Kansas, South Carolina, Virginia, Missouri, and Minnesota showed the highest increase.Only the state of Kansas is part of Tornado valley, showing that Tornados are increasing outside of

'Tornado Valley'. The States that showed a large percent decrease are Maryland, Michigan, Wyoming, and West Virginia. The decrease the same as increase is not subject to any specific geographical area.

Preparing for Future Work

The way the data was recorded changed in 2007, so I will do analyze 1950-2006 and 2007-2015 separately in the next part of the analysis.

I have three datasets for future analysis, 1950-2006, 2007-2015, only states part of tornado valley.

```
df_prior <- df %>% filter(yr %in% 1950:2006)

df_post <- df %>% filter(yr %in% 2007:2015)
```

```
invisible(write.csv(df_prior, 'df_prior.csv'))
invisible(write.csv(df_prior, 'df_post.csv'))
invisible(write.csv(df_prior, 'df_valley.csv'))
```