Individual Project Task 2 - Data Visualization

Laura Zhang

3/22/2021

This project involves the dataset `Severe Weather Database Files (1950-2019)` from the U.S. National Oceanic and Atmospheric Administration (NOAA). The original data can be downloaded from https://www.spc.noaa.gov/wcm/#ATP.

Print the original dataset which records all the tornadoes occurred in the U.S. from 1950 to 2019:

```
tornado_data <- read.csv(file = "D:/1950-2019_all_tornadoes.csv", header = T, sep=",")
head(tornado_data)
```

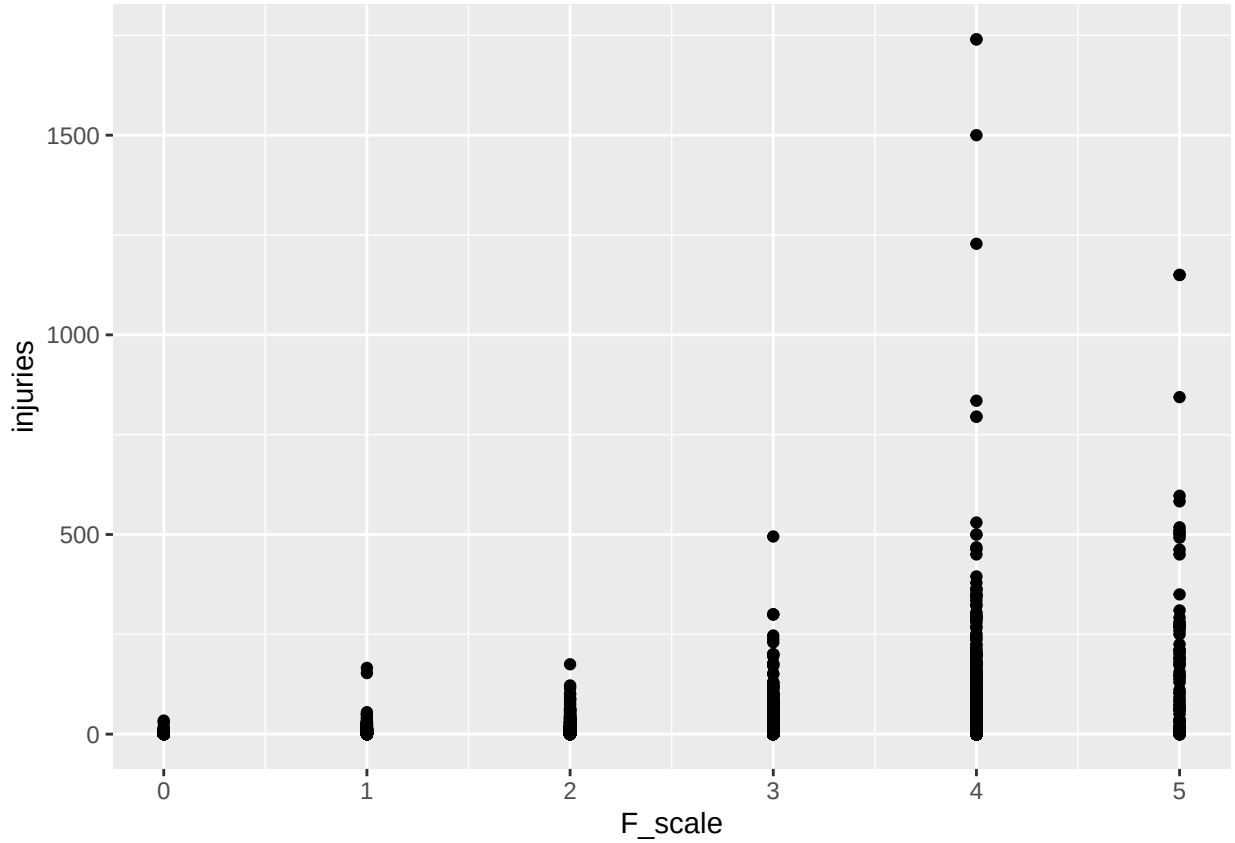
```
##   tornado_number year month day       date     time time_zone state
## 1              1 1950     1   3 1950-01-03 11:00:00         3    MO
## 2              1 1950     1   3 1950-01-03 11:00:00         3    MO
## 3              1 1950     1   3 1950-01-03 11:10:00         3    IL
## 4              2 1950     1   3 1950-01-03 11:55:00         3    IL
## 5              3 1950     1   3 1950-01-03 16:00:00         3    OH
## 6              4 1950     1  13 1950-01-13 05:25:00         3    AR
##   state_FIPS_number state_number F_scale injuries fatalities loss crop_loss
## 1                29            1       3        3          0    6         0
## 2                29            1       3        3          0    6         0
## 3                17            1       3        0          0    5         0
## 4                17            2       3        3          0    5         0
## 5                39            1       1        1          0    4         0
## 6                 5            1       3        1          1    3         0
##   starting_latitude starting_longitude ending_latitude ending_longitude
## 1             38.77              -90.22           38.83           -90.03
## 2             38.77              -90.22           38.82           -90.12
## 3             38.82              -90.12           38.83           -90.03
## 4             39.10              -89.30           39.12           -89.23
## 5             40.88              -84.58            0.00             0.00
## 6             34.40              -94.37            0.00             0.00
##   length_in_miles width_in_yards number_of_states_affected_by_this_tornado
## 1             9.5            150                                         2
## 2             6.2            150                                         2
## 3             3.3            100                                         2
## 4             3.6            130                                         1
## 5             0.1             10                                         1
## 6             0.6             17                                         1
##   state_number.1 tornado_segment_number X1st_county_FIPS_code
## 1              0                      1                    0
## 2              1                      2                  189
## 3              1                      2                  119
## 4              1                      1                  135
## 5              1                      1                  161
```

```
## 6              1                       1                     113
##   X2nd_county_FIPS_code X3rd_county_FIPS_code X4th_county_FIPS_code
## 1                     0                     0                     0
## 2                     0                     0                     0
## 3                     0                     0                     0
## 4                     0                     0                     0
## 5                     0                     0                     0
## 6                     0                     0                     0
##   F_scale_rating
## 1              0
## 2              0
## 3              0
## 4              0
## 5              0
## 6              0
```

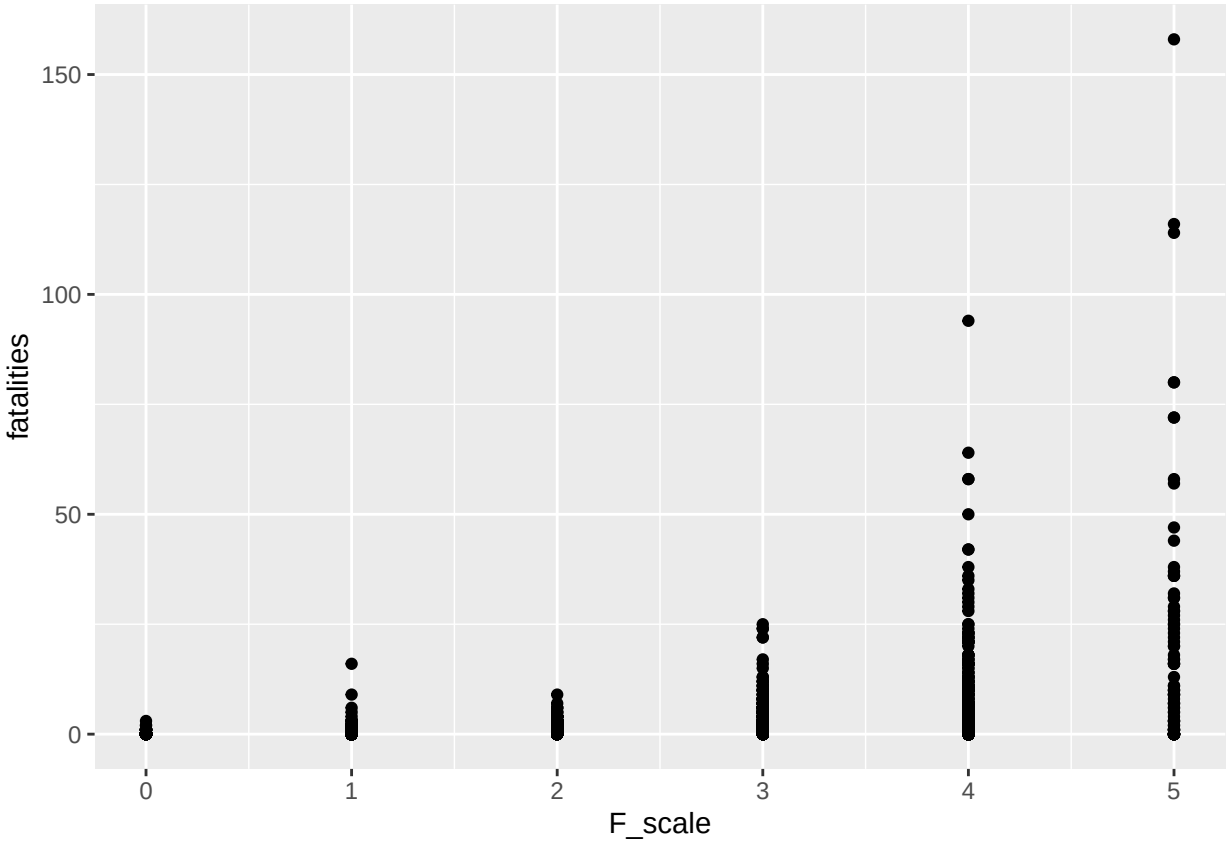
1. The original dataset is not a tidy data, for its column 2 to column 4, whose names are `year`, `month` and `day`, refer to the same as column 5 `date`.
2. `Injuries`, `fatalities` and `loss` are dependent on `F-scale`(or `EF-scale`), which has been used to assign tornado ratings since 1971. The larger `F-scale` value is, the more injuries, fatalities and loss the tornado will cause.

According to the `SPC_severe_database_description`, the column name of `F-scale` has changed to `EF-scale` after January 2007, which values -9, 0, 1, 2, 3, 4, 5 (-9 = unknown). To analyze the `F-scale` (or `EF-scale`) data, rows whose `F-scale` are -9 must be removed.

```
library(ggplot2)
knownFscale<-subset(tornado_data, F_scale>=0)
ggplot(knownFscale, aes(x=F_scale, y=injuries))+geom_point()
```



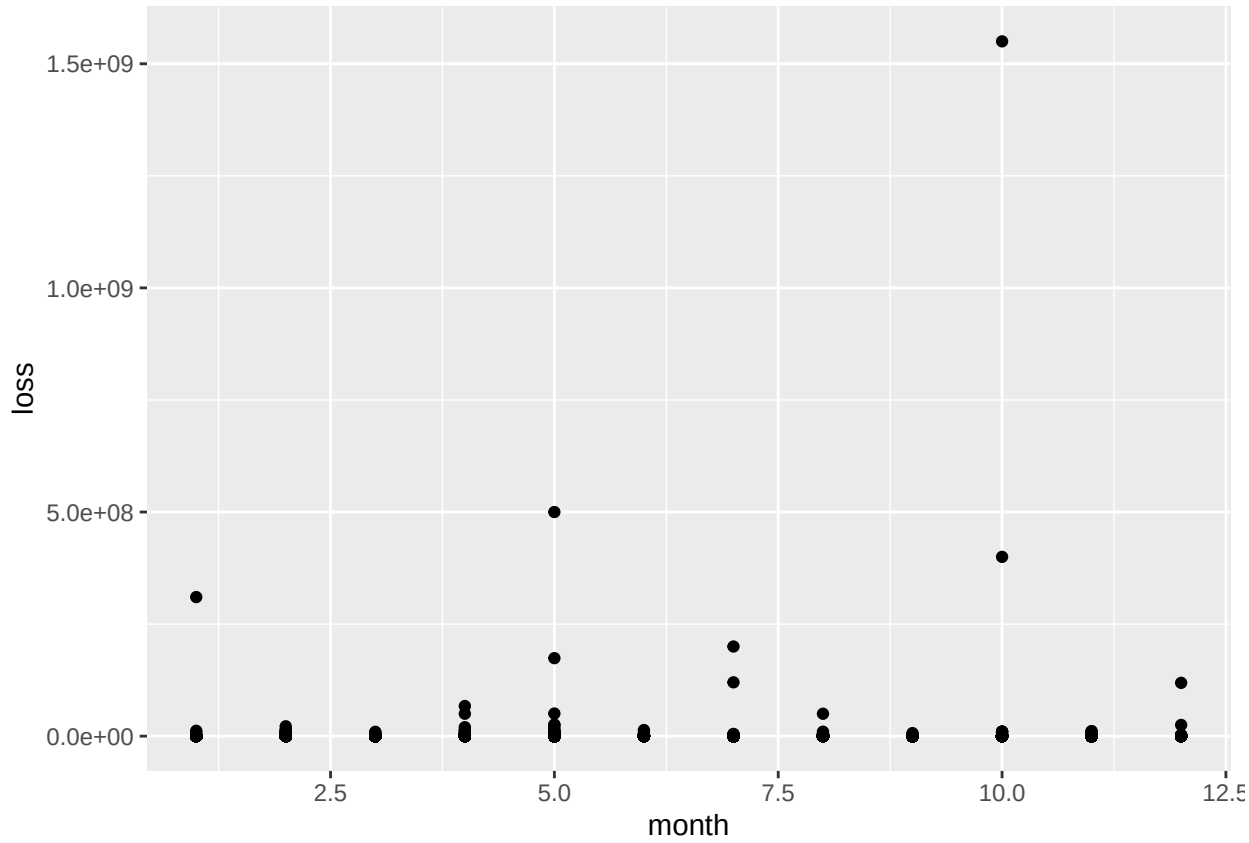

```
ggplot(knownFscale, aes(x=F_scale, y=fatalities))+geom_point()
```



From the plots above we can see a positive correlation between `F-scale` and `fatalities`, and things are roughly the same between `F-scale` and `injuries`. However, for cases whose `F-scale` is 4, reported injuries in three of them (1979-04-10, 2011-04-27, 1953-06-09) are more than all reported injuries in cases whose `F-scale` is 5.

3. To find any seasonal difference in the loss of each tornado, plot all the `loss` in each month.

```
library(ggplot2)
ggplot(tornado_data, aes(x=month, y=loss))+geom_point()
```



It is worth noting that there are outliers, which imply that the corresponding tornadoes have caused significant property losses, especially the case on 2019-10-20 in Texas, which caused the most property loss in the record ($1,550,000,000) but no injuries or fatalities were reported.

According to Wikipedia:
> The tornado outbreak of October 20–22, 2019 was a significant severe weather event across the South Central United States. On the evening of October 20, discrete supercell thunderstorms developed across the Dallas–Fort Worth metroplex, contributing to several tornadoes. One of those tornadoes caused EF3 damage in the Dallas suburbs, becoming the costliest tornado event in Texas history, at $1.55 billion.

```
sc_201910<-subset(tornado_data, year==2019&month==10&day>=20&day<=22)
sum(sc_201910$injuries)
```

```
## [1] 4
```

```
sum(sc_201910$fatalities)
```

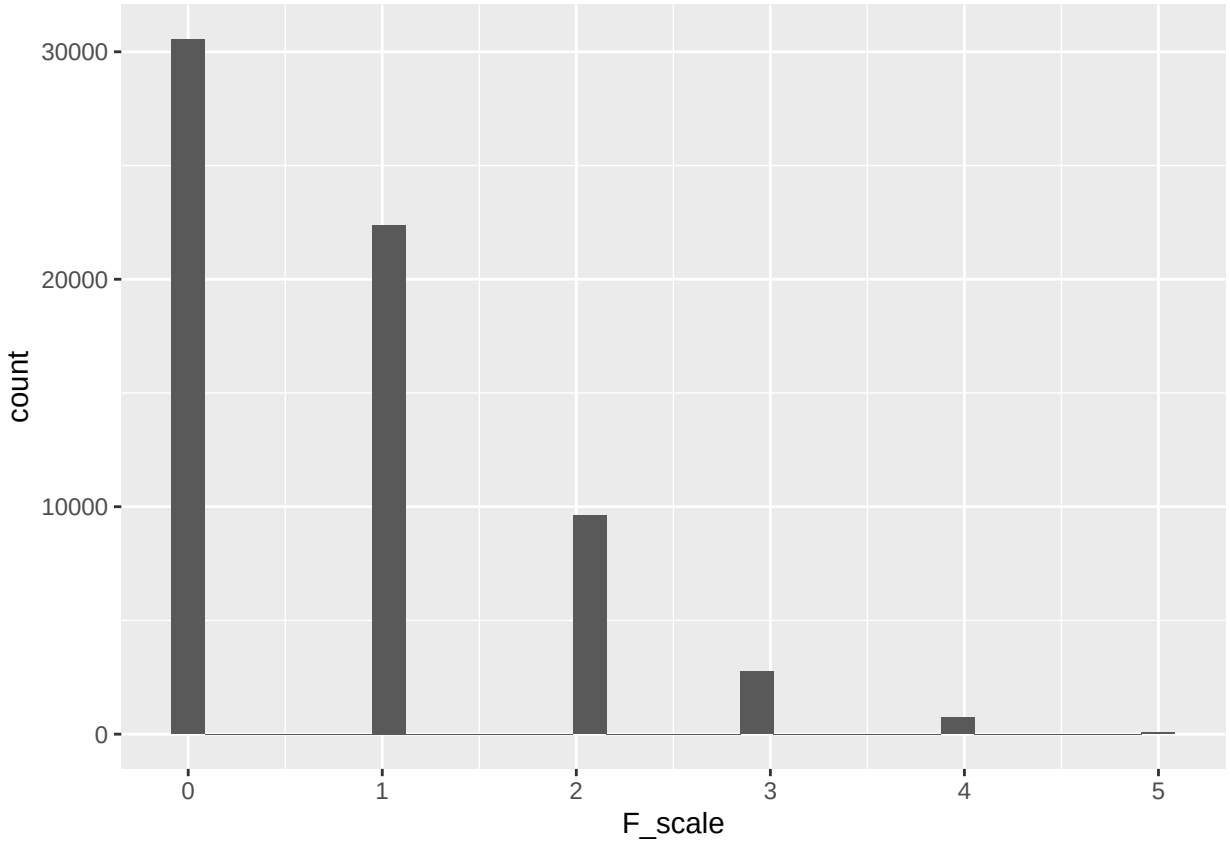
```
## [1] 0
```

By selecting all rows and create a subset of the cases occurred between 2019-10-20 to 2019-10-22, we can see that the tornado outbreak of October 20–22, 2019 only caused 4 injuries and 0 fatality, even if it was the costliest tornado event in Texas history.

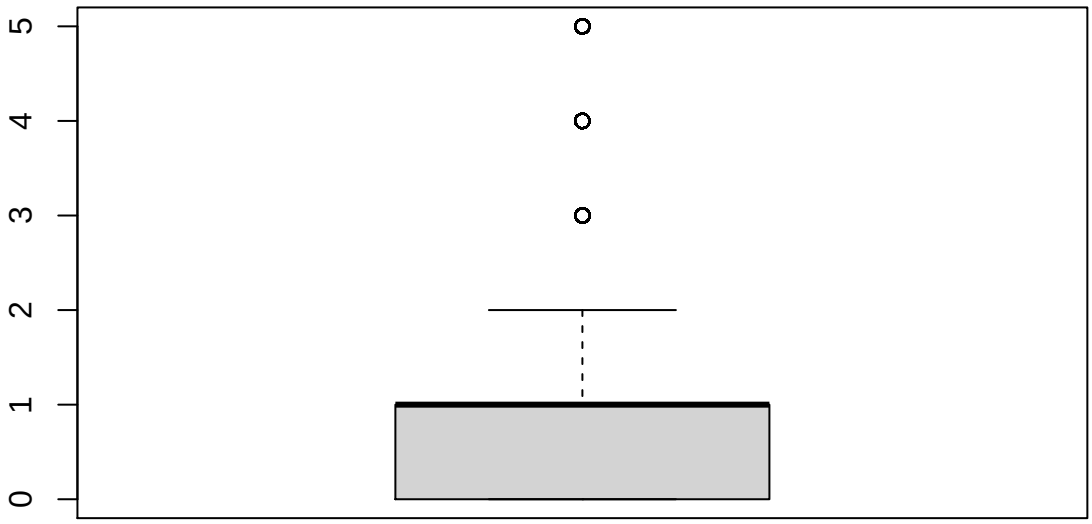
4. As an example of visualization by histogram, plot the count of cases in different `F-scale`. `F-scale` is not suitable for boxplots, because the plot has no meanings.

```
library(ggplot2)
ggplot(knownFscale, aes(x=F_scale))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

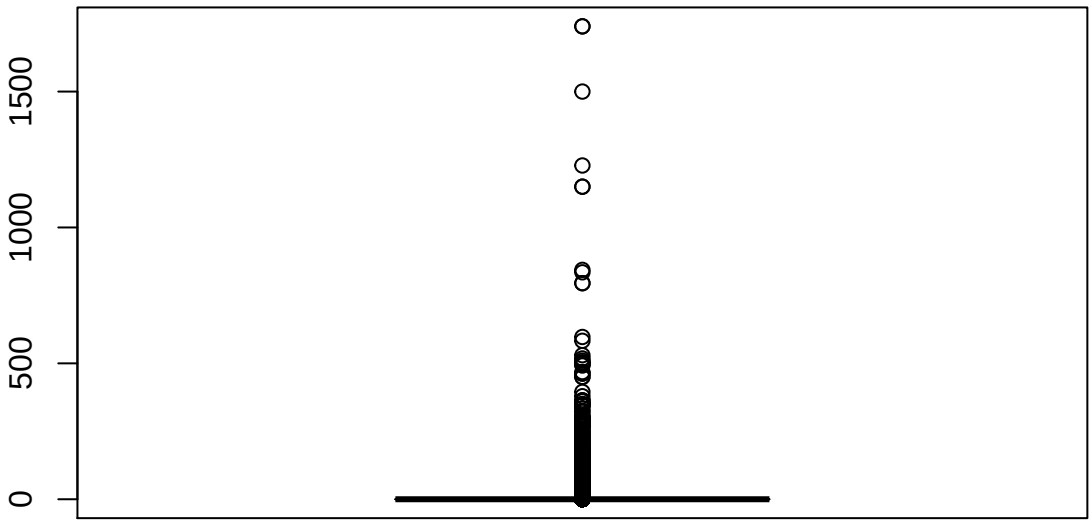



```
boxplot(knownFscale$F_scale)
```



Instead, plot the `injuries` as a boxplot.

```
library(ggplot2)
boxplot(tornado_data$injuries)
```



This plot shows injuries caused by tornadoes in most cases.