MA 615 Midterm Project

Yuhan Pu

2023-11-06

MA 615 Midterm Project

Introduction: Based on these datasets from Fema, I could have a look on all disasters happened in USA, among them, storm is always the main reason for the cause of floods. In this case, I have download the datasets for Storm in USA of both 2020 and 2021 to figure out if there are any trend and law between the characatersitics of storm and the happens of flood.

library packages:

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(tidyr)
```

Import data:

```
StormEventsDetails2020<-read.csv("StormEvents_details-ftp_v1.0_d2020_c20230927.csv")
StormEventsDetails2021<-read.csv("StormEvents_details-ftp_v1.0_d2021_c20231017.csv")
```

First, since California and Massachusetts are two states that I have been living in. I want to focus on information of these two states.

Select data by states:

```
CA2020<-filter(StormEventsDetails2020,STATE=="CALIFORNIA")
CA2021<-filter(StormEventsDetails2021,STATE=="CALIFORNIA")
MA2020<-filter(StormEventsDetails2020,STATE=="MASSACHUSETTS")
MA2021<-filter(StormEventsDetails2021,STATE=="MASSACHUSETTS")
```

Then select data by floods:

```
CA2020flood<-filter(CA2020,EVENT_TYPE=="Flood" | EVENT_TYPE=="Flash Flood")
CA2021flood<-filter(CA2021,EVENT_TYPE=="Flood" | EVENT_TYPE=="Flash Flood")
MA2020flood<-filter(MA2020,EVENT_TYPE=="Flood" | EVENT_TYPE=="Flash Flood")
MA2021flood<-filter(MA2021,EVENT_TYPE=="Flood" | EVENT_TYPE=="Flash Flood")
```

Let's focus on California 2020 first: First thing I want to figure out is the time of duration of each floods or flash floods.

```
CA2020flood$DURATION<-(CA2020flood$END_TIME-CA2020flood$BEGIN_TIME)+(CA2020flood$END_DAY-CA2020flood$BE
```

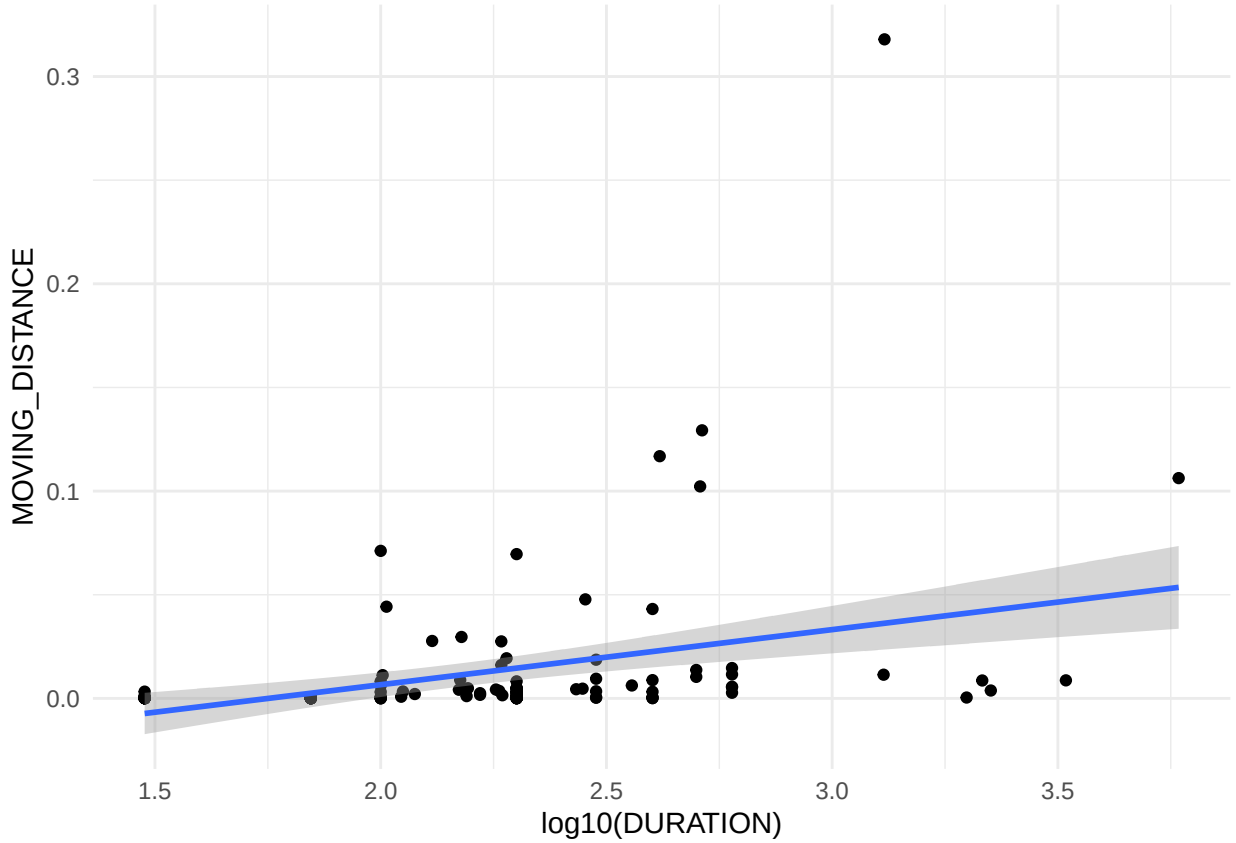
I want to know if there are any relationship between the moving distance of the storm and the duration. So I need to calculate a rate of moving distance by subtracting the latitude and longitude.

```
CA2020flood$MOVING_DISTANCE<-sqrt((CA2020flood$END_LAT-CA2020flood$BEGIN_LAT)^2+(CA2020flood$END_LON-CA
```

Then use ggplot to see if there any relationship between these two variables.

```
ggplot(data = CA2020flood, aes(x = log10(DURATION), y = MOVING_DISTANCE)) +
  geom_point() +
  geom_smooth(method='lm',se=TRUE) +
  labs(x = "log10(DURATION)", y = "MOVING_DISTANCE") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



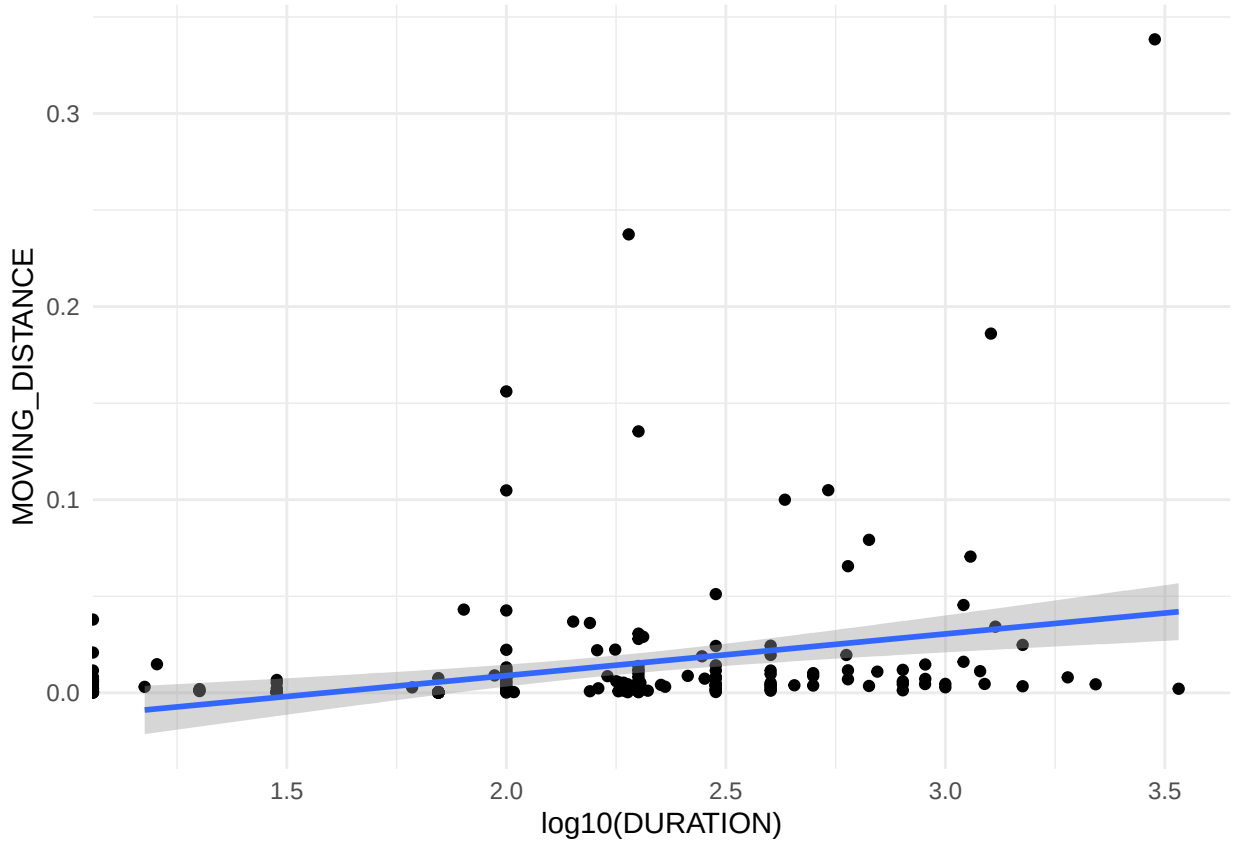
Then I want to do same things to California 2021:

```
CA2021flood$DURATION<-(CA2021flood$END_TIME-CA2021flood$BEGIN_TIME)+(CA2021flood$END_DAY-CA2021flood$BE
CA2021flood$MOVING_DISTANCE<-sqrt((CA2021flood$END_LAT-CA2021flood$BEGIN_LAT)^2+(CA2021flood$END_LON-CA
```

```
ggplot(data = CA2021flood, aes(x = log10(DURATION), y = MOVING_DISTANCE)) +
  geom_point() +
  geom_smooth(method='lm',se=TRUE) +
  labs(x = "log10(DURATION)", y = "MOVING_DISTANCE") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 19 rows containing non-finite values (`stat_smooth()`).
```



From these two plots above, we can see that if we fit a line for both of them, the slope is a little bit flat so I want to check whether there are still relationship between these two variables in California.

Fit two models here:

```
model2020CA<-lm(CA2020flood$MOVING_DISTANCE~CA2020flood$DURATION)
model2021CA<-lm(CA2021flood$MOVING_DISTANCE~CA2021flood$DURATION)
```

Then summary them to get the p-value:

```
summary(model2020CA)
```

```
##
## Call:
## lm(formula = CA2020flood$MOVING_DISTANCE ~ CA2020flood$DURATION)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.054371 -0.007559 -0.006266 -0.004328  0.289490
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.692e-03  3.169e-03   1.796  0.07482 .
## CA2020flood$DURATION 1.742e-05  4.392e-06   3.966  0.00012 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.03298 on 130 degrees of freedom
## Multiple R-squared:  0.1079, Adjusted R-squared:  0.1011
## F-statistic: 15.73 on 1 and 130 DF,  p-value: 0.00012
```

Since p-value is smaller than 0.05, we reject the null hypothesis and conclude that DURATION is significant in affecting the MOVING_DISTANCE for California in 2020.

The same for California 2021:

```
summary(model2021CA)
```

```
##
## Call:
## lm(formula = CA2021flood$MOVING_DISTANCE ~ CA2021flood$DURATION)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.103638 -0.009116 -0.005751 -0.001103  0.244527
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.856e-03  2.725e-03   1.782   0.0761 .
## CA2021flood$DURATION 2.967e-05  5.173e-06   5.736 3.23e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03363 on 217 degrees of freedom
## Multiple R-squared:  0.1317, Adjusted R-squared:  0.1277
## F-statistic:  32.9 on 1 and 217 DF,  p-value: 3.226e-08
```

Since p-value is smaller than 0.05, we reject the null hypothesis and conclude that DURATION can have an affect on the MOVING_DISTANCE for California in 2020.

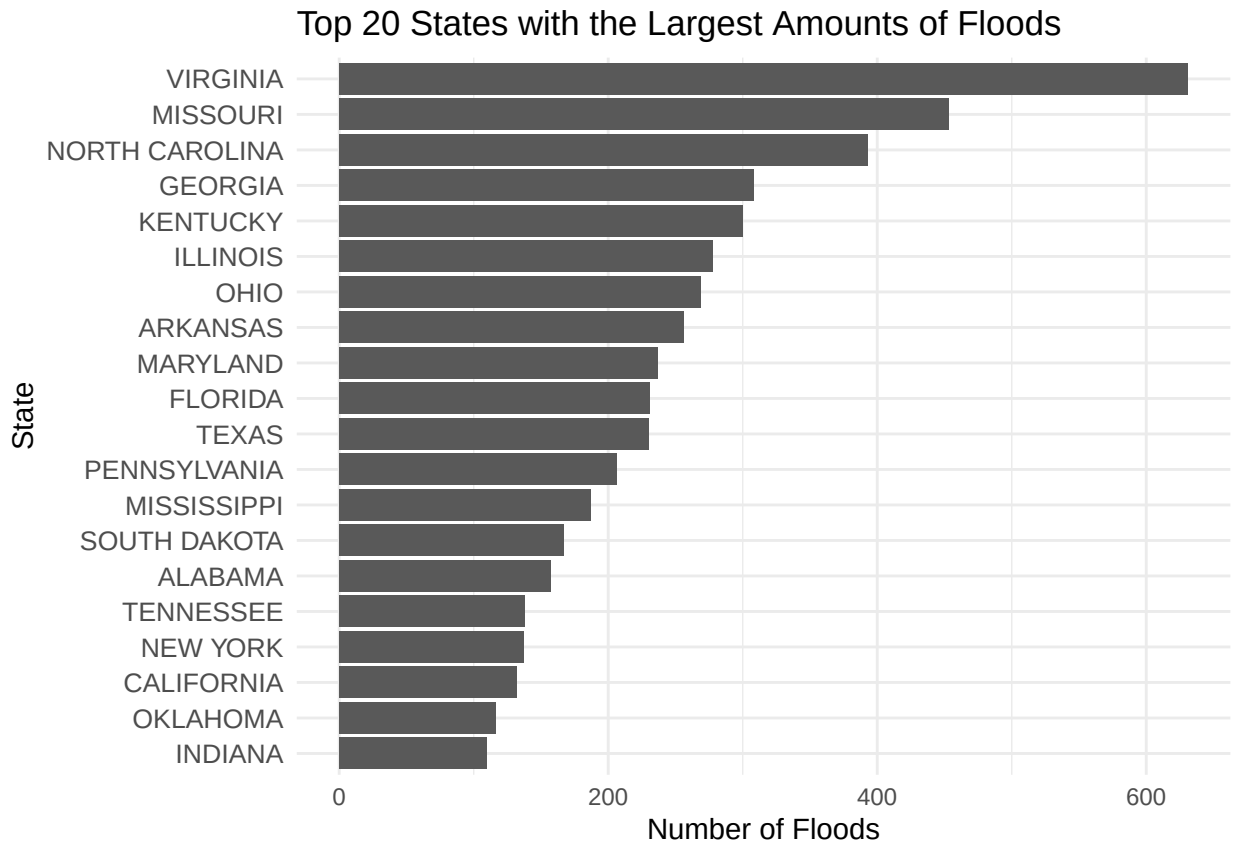
Both models can conclude that, for California, the moving distance of storms here is related with the duration of them.

After knowing this, I notice that the datas for only California and Massachussets seems to be so few that it's hard to find some law or frequence of data. In this case, I plan to go back to the overall datasets ("StormEventsDetails2020" and "StormEventsDetails2021") to search for some details.

So I want to first know the top 20 states that floods are most likely to happen here.

```
flood2020<-filter(StormEventsDetails2020,EVENT_TYPE=="Flood" | EVENT_TYPE=="Flash Flood")
summary_data<-flood2020 %>%
  group_by(STATE) %>%
  summarize(counts=n()) %>%
  arrange(desc(counts)) %>%
  head(20)
ggplot(summary_data, aes(x = reorder(STATE, counts), y = counts)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(
    x = "State",
    y = "Number of Floods",
```

```
    title = "Top 20 States with the Largest Amounts of Floods"
  ) +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 10))
```



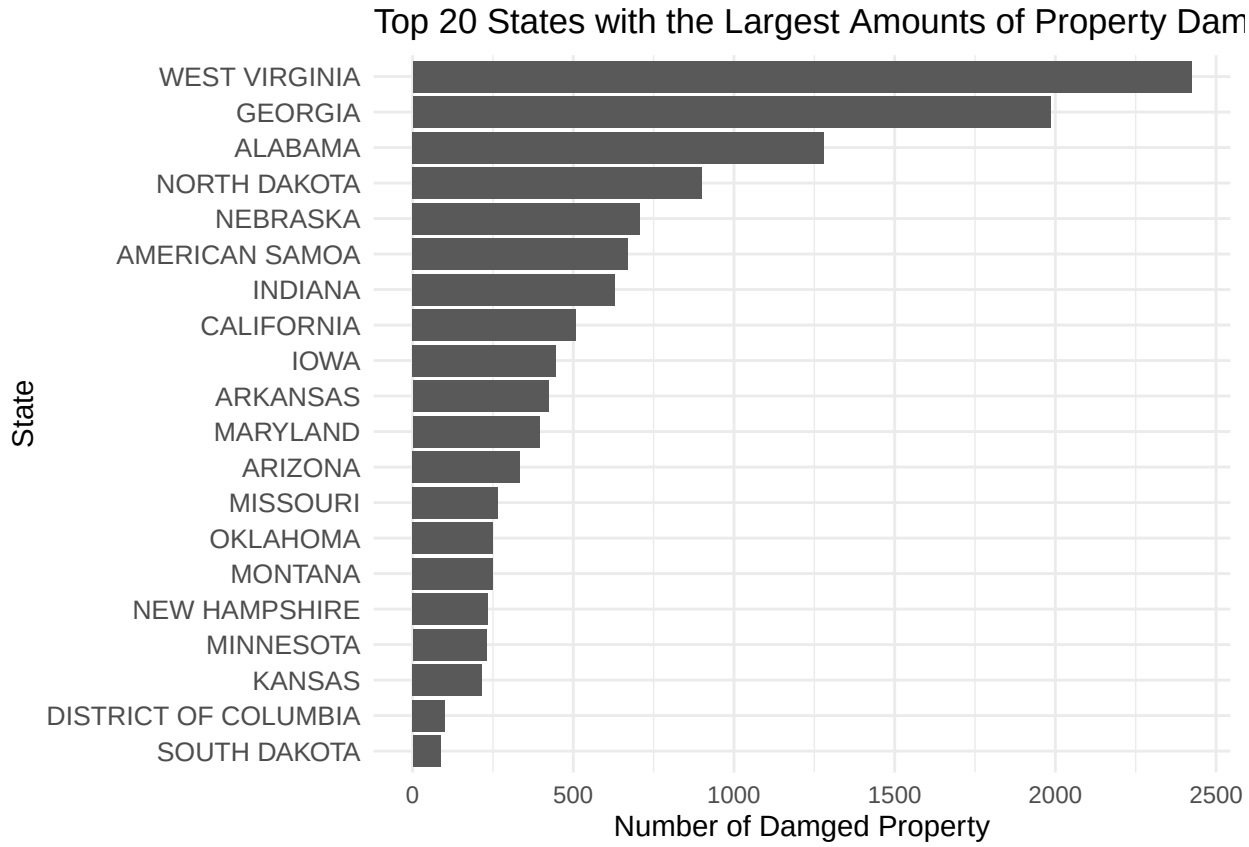
Also I want to know the top 20 states with the largest amount of property damaged.

```
flood2020$DAMAGE_PROPERTY <- gsub("K","",flood2020$DAMAGE_PROPERTY)
flood2020$DAMAGE_CROPS <- gsub("K","",flood2020$DAMAGE_CROPS)
```

```
summary_data<-flood2020 %>%
  group_by(STATE) %>%
  summarize(TotalPropertyDamaged=sum(as.numeric(DAMAGE_PROPERTY))) %>%
  arrange(desc(TotalPropertyDamaged),na.rm=TRUE) %>%
  head(20)
```

```
## Warning: There were 23 warnings in `summarize()`.
## The first warning was:
## i In argument: `TotalPropertyDamaged = sum(as.numeric(DAMAGE_PROPERTY))`.
## i In group 2: `STATE = "ALASKA"`.
## Caused by warning:
## ! NAs introduced by coercion
## i Run `dplyr::last_dplyr_warnings()` to see the 22 remaining warnings.
```

```
ggplot(summary_data, aes(x = reorder(STATE,TotalPropertyDamaged), y = TotalPropertyDamaged)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(
    x = "State",
    y = "Number of Damged Property",
    title = "Top 20 States with the Largest Amounts of Property Damaged"
  ) +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 10))
```



West Virgnia has the largest amount of property damaged in the flood.

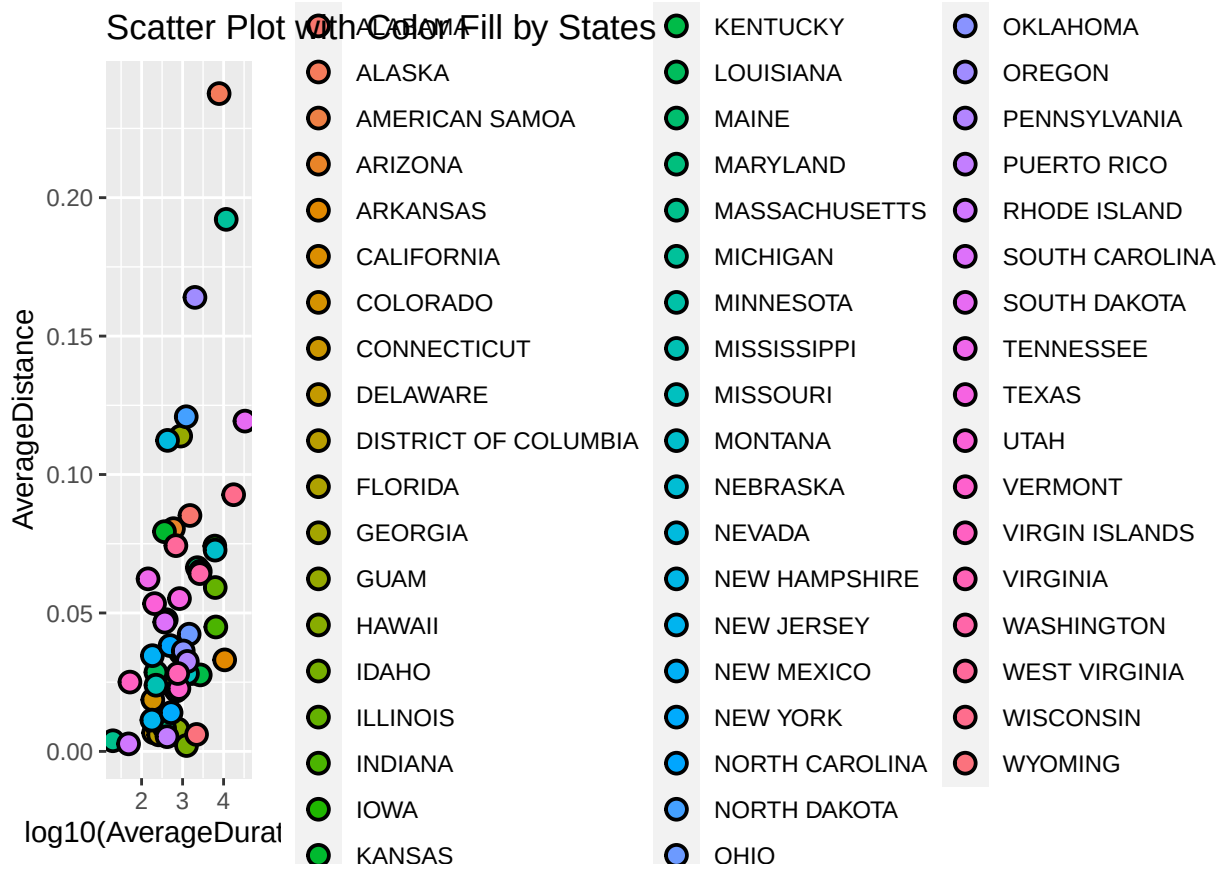
I want to create a new column to show the duration of each floods. Based on the findings in California that the duration of floods might have a relationship with the moving distance.

```
flood2020$DURATION<-(flood2020$END_TIME-flood2020$BEGIN_TIME)+(flood2020$END_DAY-flood2020$BEGIN_DAY)*24
flood2020$MOVING_DISTANCE<-sqrt((flood2020$END_LAT-flood2020$BEGIN_LAT)^2+(flood2020$END_LON-flood2020$
```

I want to get the average duration of each states and the average moving distance of each states:

```
summary_data<-flood2020 %>%
  group_by(STATE) %>%
  summarize(AverageDuration=mean(as.numeric(DURATION)),
            AverageDistance=mean(as.numeric(MOVING_DISTANCE)))
ggplot(summary_data, aes(x = log10(AverageDuration), y = AverageDistance, fill = STATE)) +
```

```
  geom_point(shape = 21, size = 3, stroke = 1) +
  labs(
    x = "log10(AverageDuration)",
    y = "AverageDistance",
    title = "Scatter Plot with Color Fill by States")
```



I notice that the duration and distance can have a more obvious relationship through the picture above. So I want to fit a model.

```
model2020<-lm(AverageDistance~log10(AverageDuration), data=summary_data)
summary(model2020)
```

```
##
## Call:
## lm(formula = AverageDistance ~ log10(AverageDuration), data = summary_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06138 -0.02600 -0.01034  0.01358  0.14852
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -0.064703   0.025764  -2.511   0.0151 *
## log10(AverageDuration)  0.039519   0.008589   4.601 2.66e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04119 on 53 degrees of freedom
## Multiple R-squared:  0.2854, Adjusted R-squared:  0.272
## F-statistic: 21.17 on 1 and 53 DF,  p-value: 2.655e-05
```

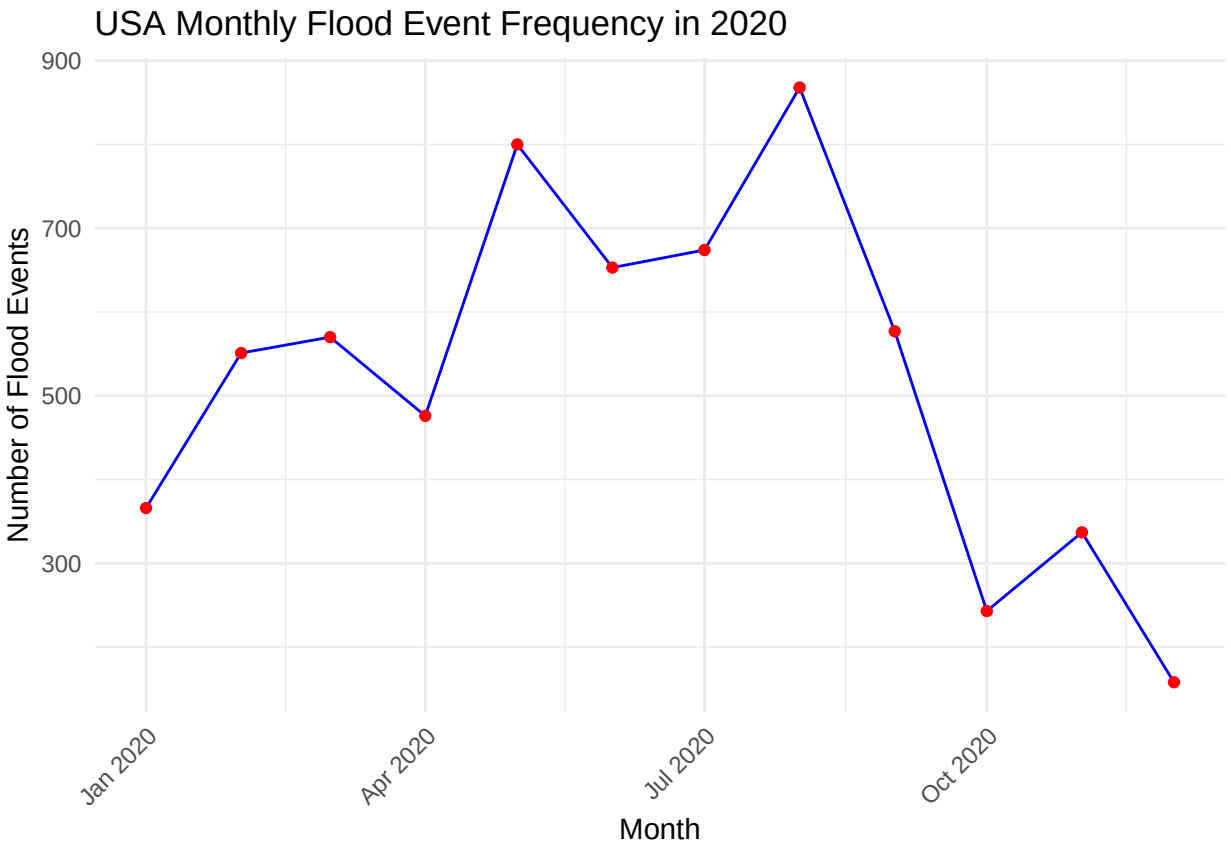
We see the p-value here is pretty smaller than 0.05 which indicates a strong relationship between Average Duration and Average Distance.

After this, I want to know the frequency of flood at a certain period of time.

```
flood2020hap<-flood2020%>%
  mutate(
    YEAR=floor(BEGIN_YEARMONTH/100),
    MONTH=BEGIN_YEARMONTH%%100
  )
flood2020freq<-flood2020hap%>%
  count(YEAR,MONTH)%>%
  rename(FloodCount=n)
```

After get the frequency in each month, let's make a plot:

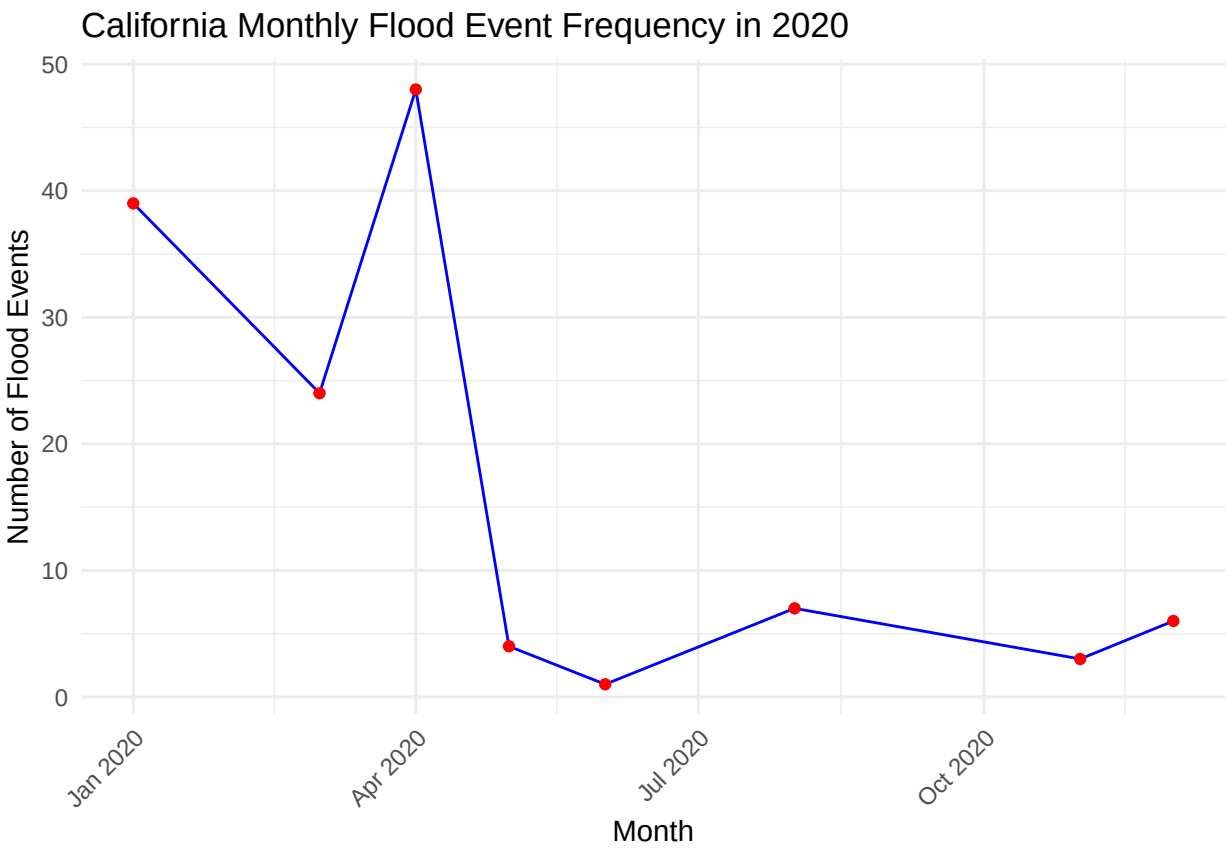
```
flood2020freq$DATE<-make_date(flood2020freq$YEAR,flood2020freq$MONTH)
ggplot(flood2020freq, aes(x = DATE, y = FloodCount)) +
  geom_line(group = 1, colour = "blue") +
  geom_point(colour = "red") +
  labs(title = "USA Monthly Flood Event Frequency in 2020",
       x = "Month",
       y = "Number of Flood Events") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



We saw that the highest number of flood events happen at around May and August.

As I just mentioned, California and Massachusetts are two states that I am most curious about. So I want to know the frequency of flood on these two states: For California:

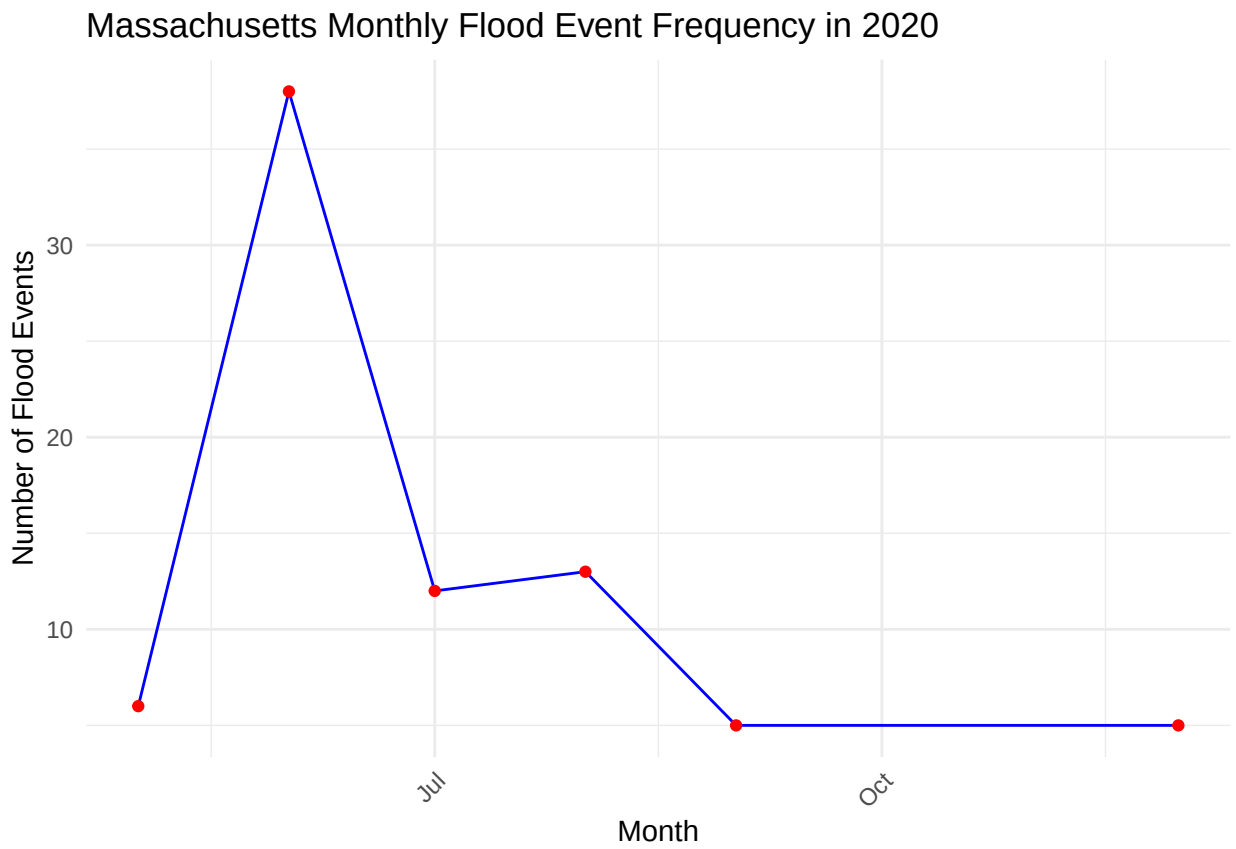
```
CAflood2020hap<-CA2020flood%>%
  mutate(
    YEAR=floor(BEGIN_YEARMONTH/100),
    MONTH=BEGIN_YEARMONTH%%100
  )
CAflood2020freq<-CAflood2020hap%>%
  count(YEAR,MONTH)%>%
  rename(FloodCount=n)
CAflood2020freq$DATE<-make_date(CAflood2020freq$YEAR,CAflood2020freq$MONTH)
ggplot(CAflood2020freq, aes(x = DATE, y = FloodCount)) +
  geom_line(group = 1, colour = "blue") +
  geom_point(colour = "red") +
  labs(title = "California Monthly Flood Event Frequency in 2020",
       x = "Month",
       y = "Number of Flood Events") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Different from the overall trend in USA, California have the largest amount of flood around April and Janurary.

Then for Massachusetts:

```
MAflood2020hap<-MA2020flood%>%
  mutate(
    YEAR=floor(BEGIN_YEARMONTH/100),
    MONTH=BEGIN_YEARMONTH%%100
  )
MAflood2020freq<-MAflood2020hap%>%
  count(YEAR,MONTH)%>%
  rename(FloodCount=n)
MAflood2020freq$DATE<-make_date(MAflood2020freq$YEAR,MAflood2020freq$MONTH)
ggplot(MAflood2020freq, aes(x = DATE, y = FloodCount)) +
  geom_line(group = 1, colour = "blue") +
  geom_point(colour = "red") +
  labs(title = "Massachusetts Monthly Flood Event Frequency in 2020",
       x = "Month",
       y = "Number of Flood Events") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

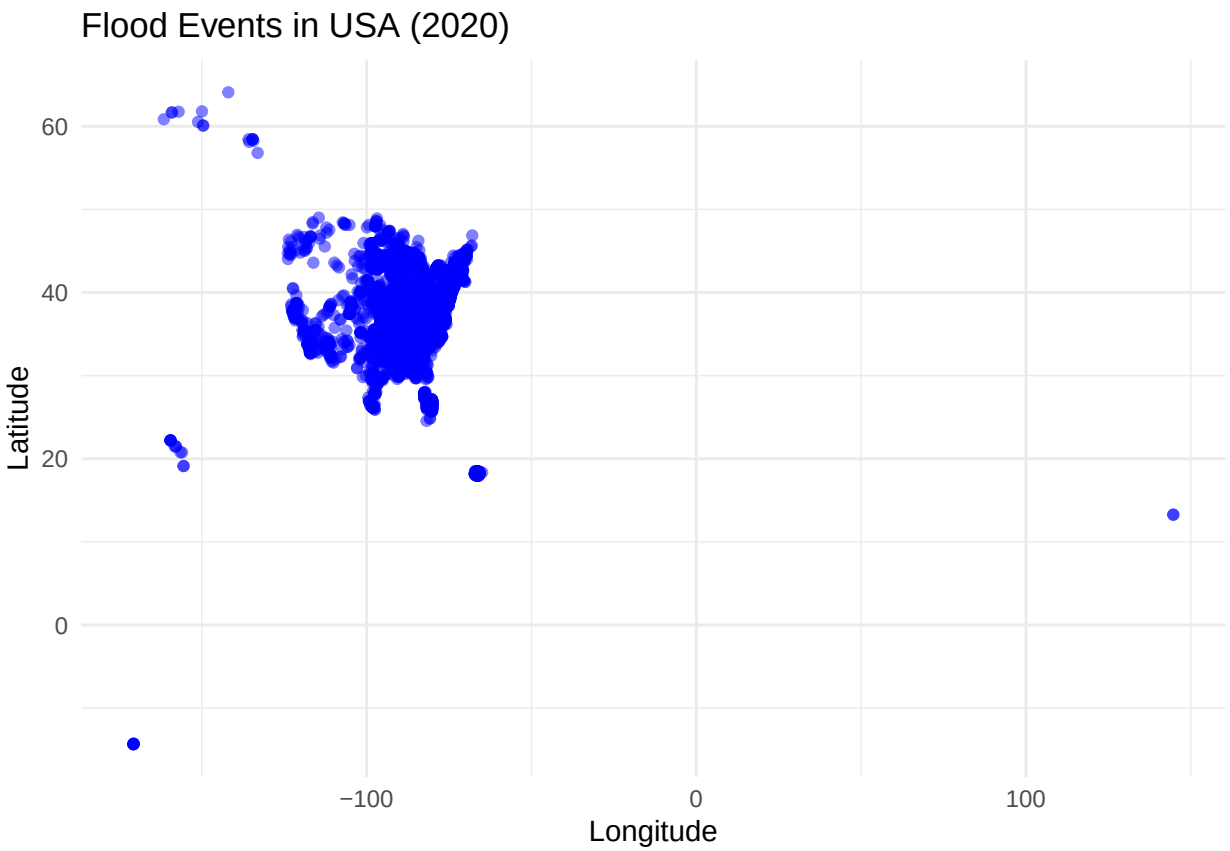


Similar to the overall trend, Massachusetts have the largest amount of floods around May.

I want to know the occurance of floods at different latitude and longitude in USA, California and Massachusetts:

The first is for the whole USA:

```
ggplot(flood2020, aes(x=BEGIN_LON,y=BEGIN_LAT))+
  geom_point(alpha=0.5,color='blue')+
  labs(title='Flood Events in USA (2020)',
       x='Longitude',
       y='Latitude') +
  theme_minimal()
```

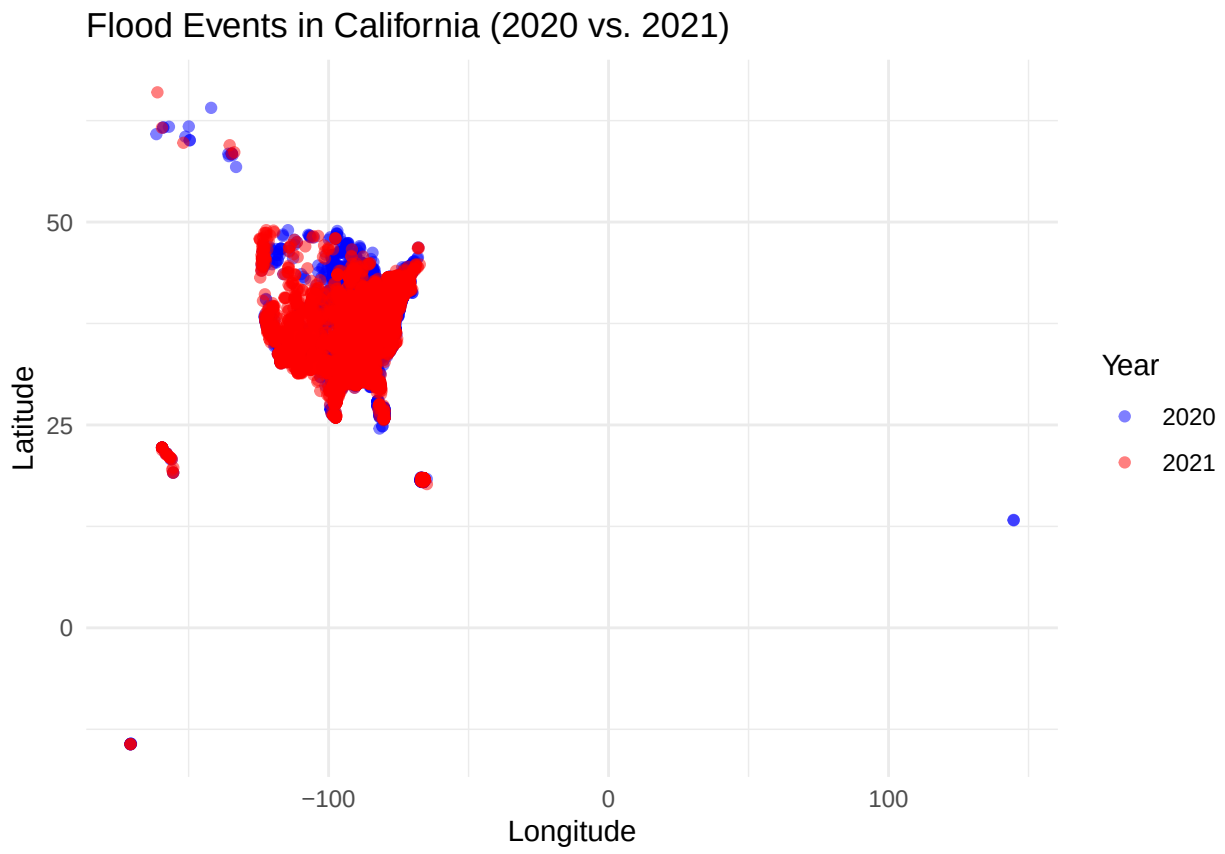


We see that the occurance of floods is mostly often in around 40 latitude and -100 longitude.

I want to see if it is the same for California and Massachusetts.

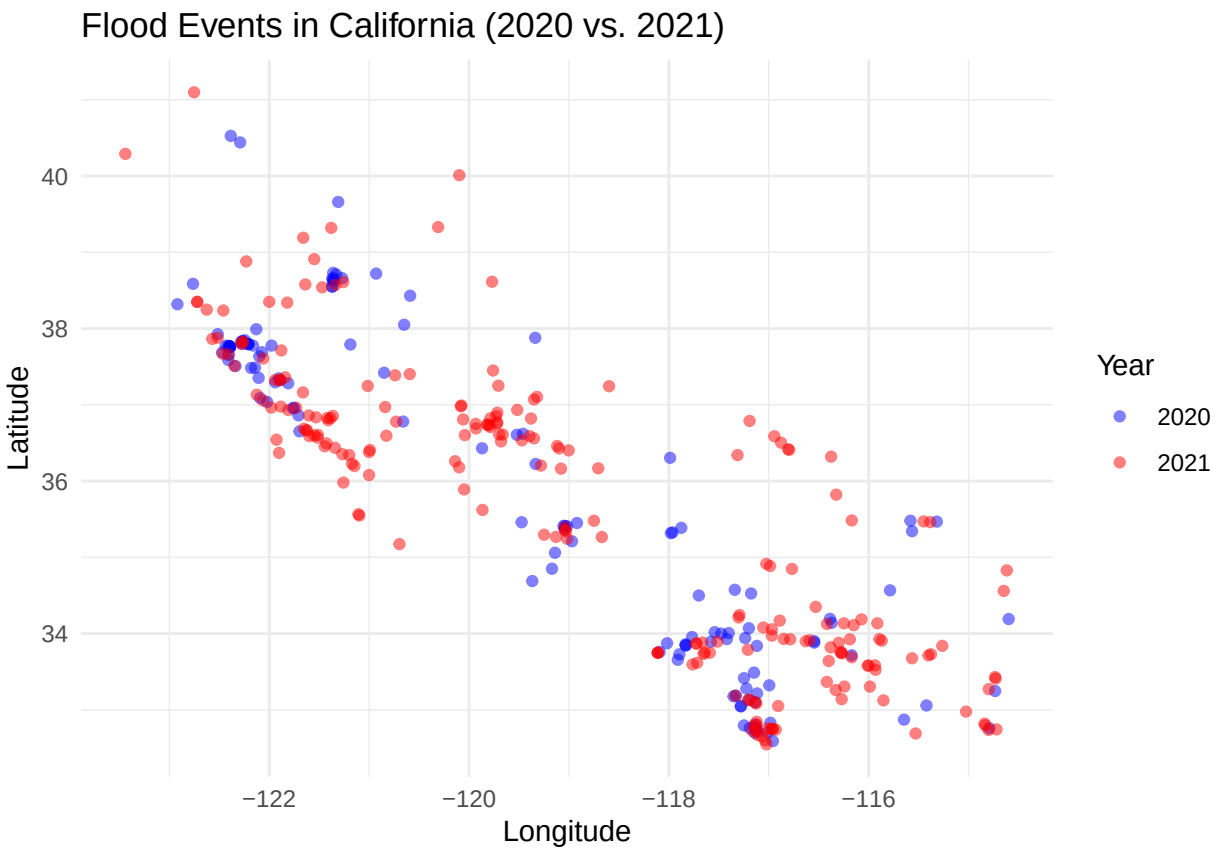
I also collect the data from 2021, I want to plot them with 2020 to see if there are significant difference: For the whole USA:

```
flood2021<-filter(StormEventsDetails2021,EVENT_TYPE=="Flood" | EVENT_TYPE=="Flash Flood")
USAfloodcombined<-bind_rows(flood2020,flood2021)
ggplot(USAfloodcombined, aes(x = BEGIN_LON, y = BEGIN_LAT, color = as.factor(YEAR))) +
  geom_point(alpha = 0.5) +
  scale_color_manual(values = c('blue', 'red')) +
  labs(title = 'Flood Events in California (2020 vs. 2021)',
       x = 'Longitude',
       y = 'Latitude',
       color = 'Year') +
  theme_minimal()
```



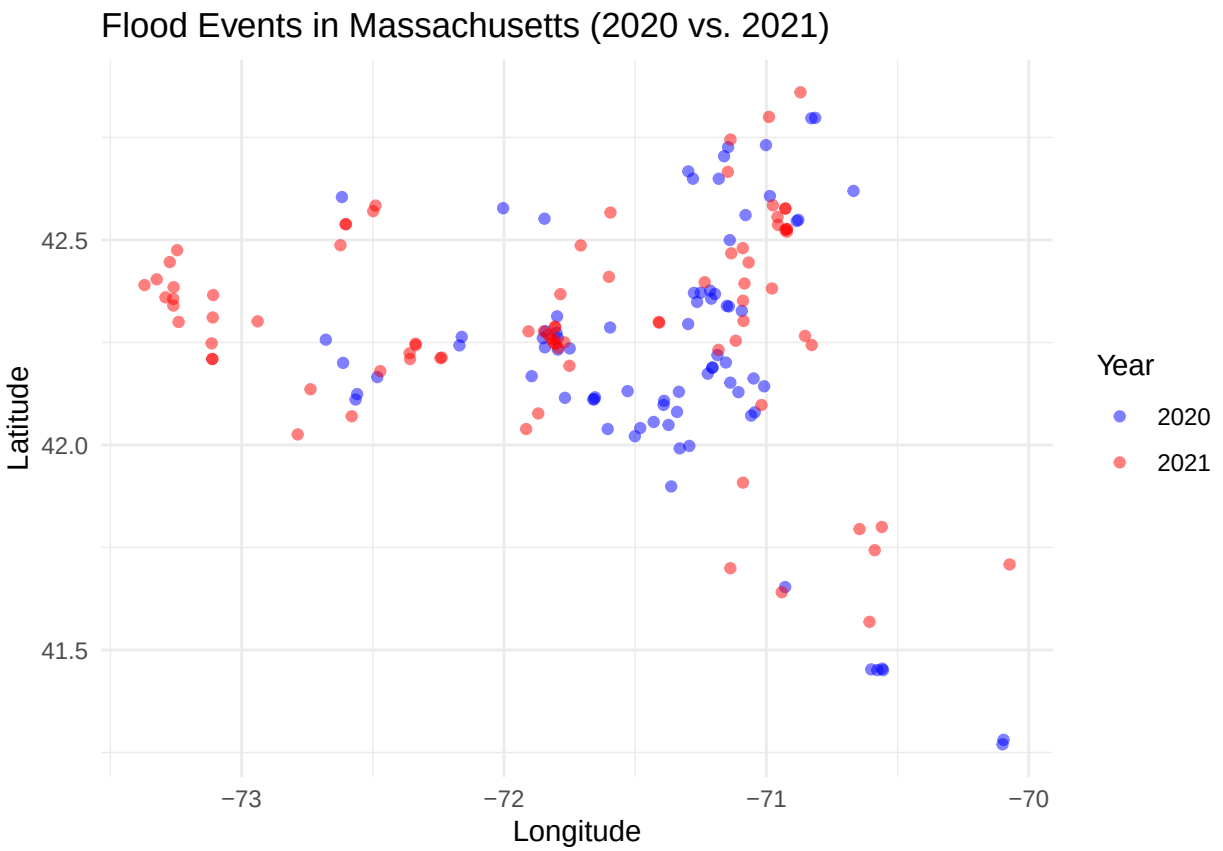
For California:

```
CAfloodcombined<-bind_rows(CA2020flood, CA2021flood)
ggplot(CAfloodcombined, aes(x = BEGIN_LON, y = BEGIN_LAT, color = as.factor(YEAR))) +
  geom_point(alpha = 0.5) +
  scale_color_manual(values = c('blue', 'red')) +
  labs(title = 'Flood Events in California (2020 vs. 2021)',
       x = 'Longitude',
       y = 'Latitude',
       color = 'Year') +
  theme_minimal()
```



For Massachusetts

```
MAfloodcombined<-bind_rows(MA2020flood, MA2021flood)
ggplot(MAfloodcombined, aes(x = BEGIN_LON, y = BEGIN_LAT, color = as.factor(YEAR))) +
  geom_point(alpha = 0.5) +
  scale_color_manual(values = c('blue', 'red')) +
  labs(title = 'Flood Events in Massachusetts (2020 vs. 2021)',
       x = 'Longitude',
       y = 'Latitude',
       color = 'Year') +
  theme_minimal()
```



We found that for the whole USA, California and Massachusetts, there are no obvious difference in the position of occurance of floods in 2020 or 2021.

Besides this, the damage and hazard of floods are also important, so I want to know the death and injuries brought by floods.

```
total_deaths <- sum(flood2020$DEATHS_DIRECT, flood2020$DEATHS_INDIRECT, na.rm = TRUE)
total_injuries <- sum(flood2020$INJURIES_DIRECT, flood2020$INJURIES_INDIRECT, na.rm = TRUE)
```

define a function to remove “K” ,“M”, and “B” (I got help from Chatgpt on this part)

```
convert_damage <- function(damage) {
  if (is.na(damage)) {
    return(0)
  }
  factor <- 1
  if (grepl("K", damage)) {
    factor <- 1e3
    damage <- str_remove(damage, "K")
  } else if (grepl("M", damage)) {
    factor <- 1e6
    damage <- str_remove(damage, "M")
  } else if (grepl("B", damage)) {
    factor <- 1e9
    damage <- str_remove(damage, "B")
  }
```

```
  return(as.numeric(damage) * factor)
}
```

Then we count the number of death, injuries, and damages:

```
flood2020$DAMAGE_PROPERTY_NUM <- sapply(flood2020$DAMAGE_PROPERTY, convert_damage)
flood2020$DAMAGE_CROPS_NUM <- sapply(flood2020$DAMAGE_CROPS, convert_damage)
total_damage_property <- sum(flood2020$DAMAGE_PROPERTY_NUM, na.rm = TRUE)
total_damage_crops <- sum(flood2020$DAMAGE_CROPS_NUM, na.rm = TRUE)
```

```
list(
  Total_Deaths = total_deaths,
  Total_Injuries = total_injuries,
  Total_Property_Damage = total_damage_property,
  Total_Crop_Damage = total_damage_crops
)
```

```
## $Total_Deaths
## [1] 61
##
## $Total_Injuries
## [1] 10
##
## $Total_Property_Damage
## [1] 854039355
##
## $Total_Crop_Damage
## [1] 93306005
```

I want to do the similar analysis to each states:

```
severity_by_state<-flood2020%>%
  group_by(STATE)%>%
  summarise(
    Total_Deaths=sum(DEATHS_DIRECT, DEATHS_INDIRECT,na.rm=TRUE),
    Total_Injuries=sum(INJURIES_DIRECT, INJURIES_INDIRECT,na.rm=TRUE)
  )

damage_by_state<-flood2020%>%
  group_by(STATE)%>%
  summarise(
    Total_Property_Damage=sum(DAMAGE_PROPERTY_NUM,na.rm=TRUE),
    Total_Crop_Damage=sum(DAMAGE_CROPS_NUM,na.rm=TRUE)
  )
output<-left_join(severity_by_state,damage_by_state,by='STATE')
output
```

```
## # A tibble: 55 x 5
##    STATE     Total_Deaths Total_Injuries Total_Property_Damage Total_Crop_Damage
##    <chr>            <int>          <int>                 <dbl>             <dbl>
##  1 ALABAMA              1              1                  1280                 0
##  2 ALASKA               0              0               8101474                 0
```

```
##  3 AMERICAN~            0             0               671               51
##  4 ARIZONA              4             3               334                0
##  5 ARKANSAS             0             0               423                0
##  6 CALIFORN~            1             0               508                1
##  7 COLORADO             1             0                 5.71             0.24
##  8 CONNECTI~            0             0                60                0
##  9 DELAWARE             0             0                 0                0
## 10 DISTRICT~            0             0               100                0
## # i 45 more rows
```

I want to sort this based on four columns:

```
outputsorted<-output%>%
  arrange(desc(Total_Deaths))%>%
  head(20)
outputsorted$STATE
```

```
##  [1] "NORTH CAROLINA" "KENTUCKY"       "INDIANA"        "ARIZONA"
##  [5] "MISSOURI"       "OHIO"           "PENNSYLVANIA"   "TEXAS"
##  [9] "ILLINOIS"       "OKLAHOMA"       "PUERTO RICO"    "TENNESSEE"
## [13] "UTAH"           "WISCONSIN"      "ALABAMA"        "CALIFORNIA"
## [17] "COLORADO"       "LOUISIANA"      "OREGON"         "ALASKA"
```

North Carolina has the largest total death.

```
outputsorted<-output%>%
  arrange(desc(Total_Injuries))%>%
  head(20)
outputsorted$STATE
```

```
##  [1] "ARIZONA"              "SOUTH CAROLINA"       "ALABAMA"
##  [4] "GEORGIA"              "KENTUCKY"             "NEBRASKA"
##  [7] "PUERTO RICO"          "ALASKA"               "AMERICAN SAMOA"
## [10] "ARKANSAS"             "CALIFORNIA"           "COLORADO"
## [13] "CONNECTICUT"          "DELAWARE"             "DISTRICT OF COLUMBIA"
## [16] "FLORIDA"              "GUAM"                 "HAWAII"
## [19] "IDAHO"                "ILLINOIS"
```

Arizona has the largest total injuries.

```
outputsorted<-output%>%
  arrange(desc(Total_Property_Damage))%>%
  head(20)
outputsorted$STATE
```

```
##  [1] "WASHINGTON"     "MICHIGAN"       "TEXAS"          "FLORIDA"
##  [5] "OREGON"         "NEW YORK"       "HAWAII"         "MASSACHUSETTS"
##  [9] "KENTUCKY"       "OHIO"           "PUERTO RICO"    "ILLINOIS"
## [13] "NORTH CAROLINA" "ALASKA"         "WISCONSIN"      "UTAH"
## [17] "VIRGINIA"       "PENNSYLVANIA"   "SOUTH CAROLINA" "NEW MEXICO"
```

Washington has the largest total property damage.

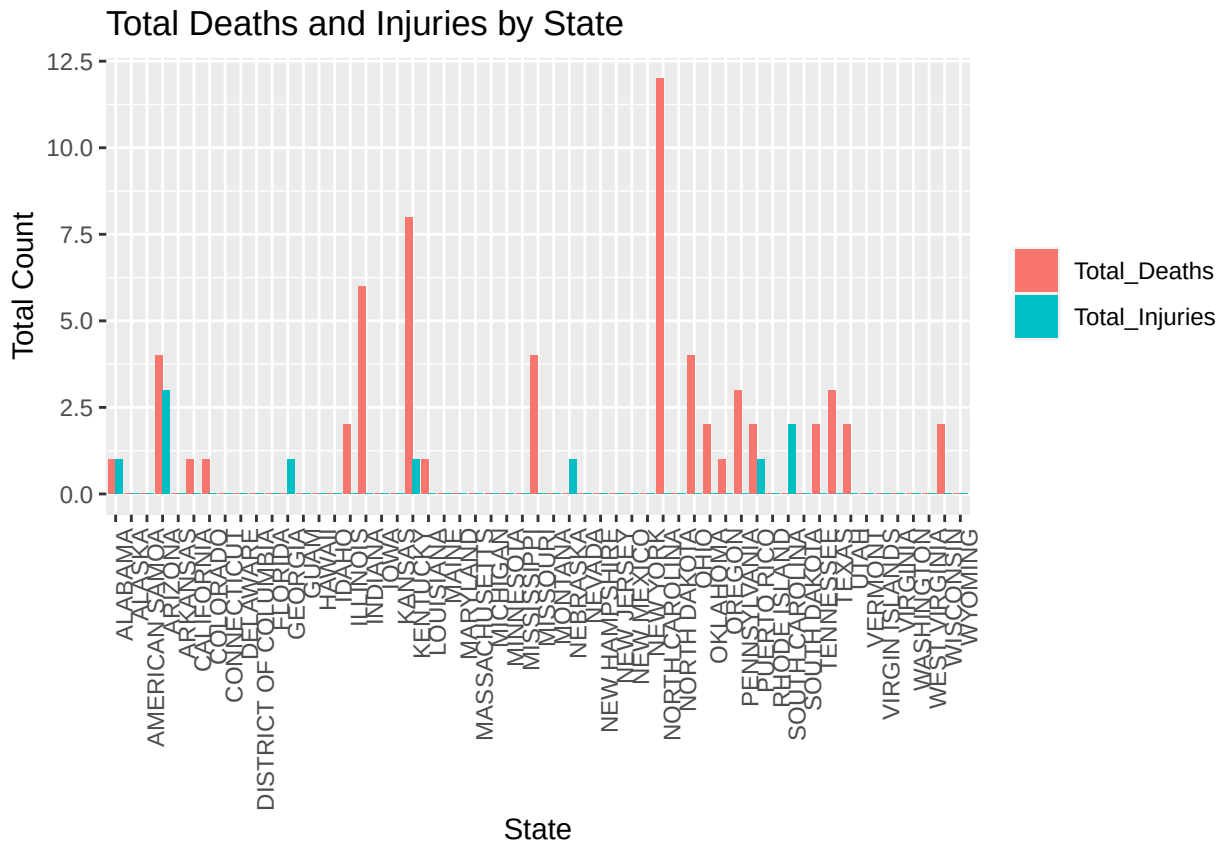
```r
outputsorted<-output%>%
  arrange(desc(Total_Crop_Damage))%>%
  head(20)
outputsorted$STATE
```

```
##  [1] "TEXAS"          "NEBRASKA"       "NORTH DAKOTA"   "MINNESOTA"
##  [5] "IOWA"           "ILLINOIS"       "WISCONSIN"      "INDIANA"
##  [9] "SOUTH DAKOTA"   "MISSOURI"       "AMERICAN SAMOA" "KENTUCKY"
## [13] "GUAM"           "IDAHO"          "PENNSYLVANIA"   "GEORGIA"
## [17] "SOUTH CAROLINA" "OHIO"           "CALIFORNIA"     "NEW JERSEY"
```

Texas has the largest total crop damage.

let's make a plot of them for total death and total injuries in each states:

```r
severity_by_state<-flood2020%>%
  group_by(STATE)%>%
  summarise(
    Total_Deaths=sum(DEATHS_DIRECT,DEATHS_INDIRECT,na.rm=TRUE),
    Total_Injuries=sum(INJURIES_DIRECT,INJURIES_INDIRECT,na.rm=TRUE)
  )%>%
  ungroup()
severity_long<-severity_by_state%>%
  pivot_longer(
    cols=c("Total_Deaths", "Total_Injuries"),
    names_to="Metric",
    values_to="Count"
  )
ggplot(severity_long,aes(x=STATE,y=Count,fill=Metric))+
  geom_bar(stat="identity",position="dodge")+
  labs(x="State",y="Total Count",title="Total Deaths and Injuries by State") +
  theme(axis.text.x=element_text(angle=90,hjust=1),
        legend.title=element_blank())
```



Conclusion: This project I estimated the data of the occurance of floods in the whole USA and especially focus on the situation in California and Massachusetts. I noticed that the occurance of floods is mainly caused by storms and it happens with the same trend in each year. It brought some deaths and injuries and damages to crops and properties.