Flood Project Final EDA

```
library("rfema")
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(readr)
library(sf)
```

```
Linking to GEOS 3.11.2, GDAL 3.7.2, PROJ 9.3.0; sf_use_s2() is TRUE
```

```
library(tigris)
```

```
To enable caching of data, set `options(tigris_use_cache = TRUE)`
in your R script or .Rprofile.
```

```
library(maps)
```

```
Warning: package 'maps' was built under R version 4.3.2
```

```
Attaching package: 'maps'

The following object is masked from 'package:purrr':

    map
```

```
library(tools)

storm_details_2020 <- read_csv("C:/Users/xianb/Downloads/MA 615/615-Flood-Project/StormEve
```

```
Rows: 61279 Columns: 51
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (26): STATE, MONTH_NAME, EVENT_TYPE, CZ_TYPE, CZ_NAME, WFO, BEGIN_DATE_T...
dbl (25): BEGIN_YEARMONTH, BEGIN_DAY, BEGIN_TIME, END_YEARMONTH, END_DAY, EN...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data2 <- read_csv("C:/Users/xianb/Downloads/MA 615/615-Flood-Project/FemaWebDisasterSummar
```

```
Rows: 3588 Columns: 14
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): hash, id
dbl  (9): disasterNumber, totalNumberIaApproved, totalAmountIhpApproved, tot...
dttm (3): paLoadDate, iaLoadDate, lastRefresh

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data_fatality_2020 <- read_csv("C:/Users/xianb/Downloads/MA 615/615-Flood-Project/StormEve
```

```
Rows: 904 Columns: 11
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (4): FATALITY_TYPE, FATALITY_DATE, FATALITY_SEX, FATALITY_LOCATION
```

```
dbl (7): FAT_YEARMONTH, FAT_DAY, FAT_TIME, FATALITY_ID, EVENT_ID, FATALITY_A...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
v2 <- read.csv("C:/Users/xianb/Downloads/DisasterDeclarationsSummaries.csv")
storm_details_2021 <- read.csv("C:/Users/xianb/Downloads/MA 615/615-Flood-Project/StormEve
data_fatality_2021 <- read.csv("C:/Users/xianb/Downloads/MA 615/615-Flood-Project/StormEve

data_fatality_2020 <- data_fatality_2020 %>%
  as_tibble()
data_fatality_2020 <- data_fatality_2020 %>%
  select(BEGIN_YEARMONTH = EVENT_YEARMONTH, c(names(data_fatality_2020)))

newdata <-storm_details_2020 %>%
  filter(EVENT_TYPE=="Flood"|EVENT_TYPE=="Coastal Flood"|
           EVENT_TYPE=="Flash Flood"|EVENT_TYPE=="Lakeshore Flood")

combine_data <- newdata %>%
  left_join(data_fatality_2020,
            by=c("EVENT_ID", "BEGIN_YEARMONTH"))
```

Plot the state count map

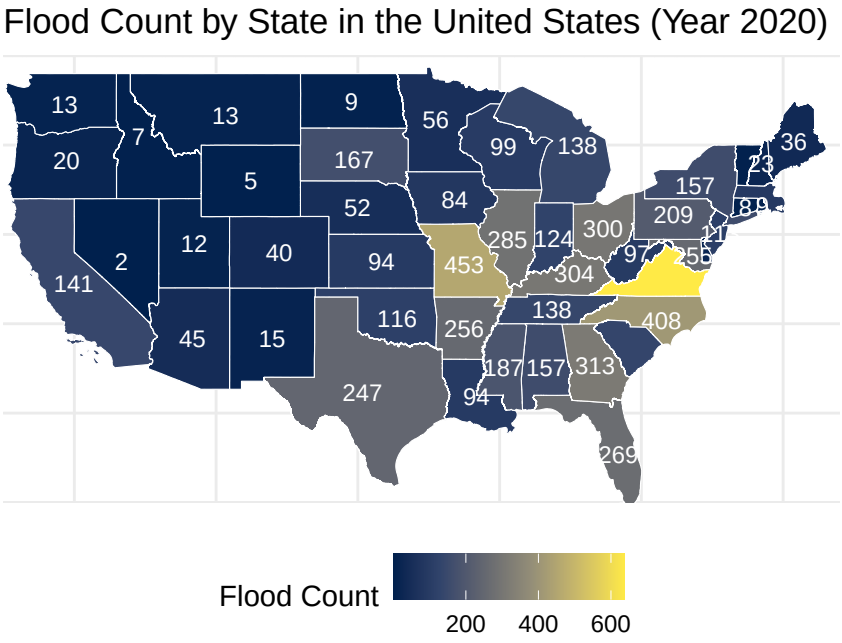
```
state_count <-combine_data %>%
  group_by(STATE)%>%
  summarise(count=n())%>%
  rename(state=STATE)
options(tigris_use_cache = TRUE)

state_count <- state_count %>%
  mutate(state=toTitleCase(tolower(state)))

states_sf <- st_as_sf(tigris::states())
```

```
Retrieving data for the year 2021
```

```r
states_sf$NAME <- as.character(states_sf$NAME)
map_data <- left_join(states_sf, state_count, by=c("NAME"="state"))

us_flood_map <- ggplot(map_data)+
  geom_sf(aes(fill=count),color="white", size=0.25)+
  scale_fill_viridis_c(name="Flood Count", option = "E")+
  theme_minimal()+
  geom_sf_text(aes(label=count),color="white",
               size=3,check_overlap = T)+
  labs(title="Flood Count by State in the United States (Year 2020)")+
  theme(legend.position = "bottom",
        axis.text.x = element_blank(),
        axis.text.y=element_blank(),
        axis.ticks = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank())+
  coord_sf(xlim = c(-125,-66),
           ylim = c(25,50),expand = F)
us_flood_map
```

```
Warning in st_point_on_surface.sfc(sf::st_zm(x)): st_point_on_surface may not
give correct results for longitude/latitude data
```

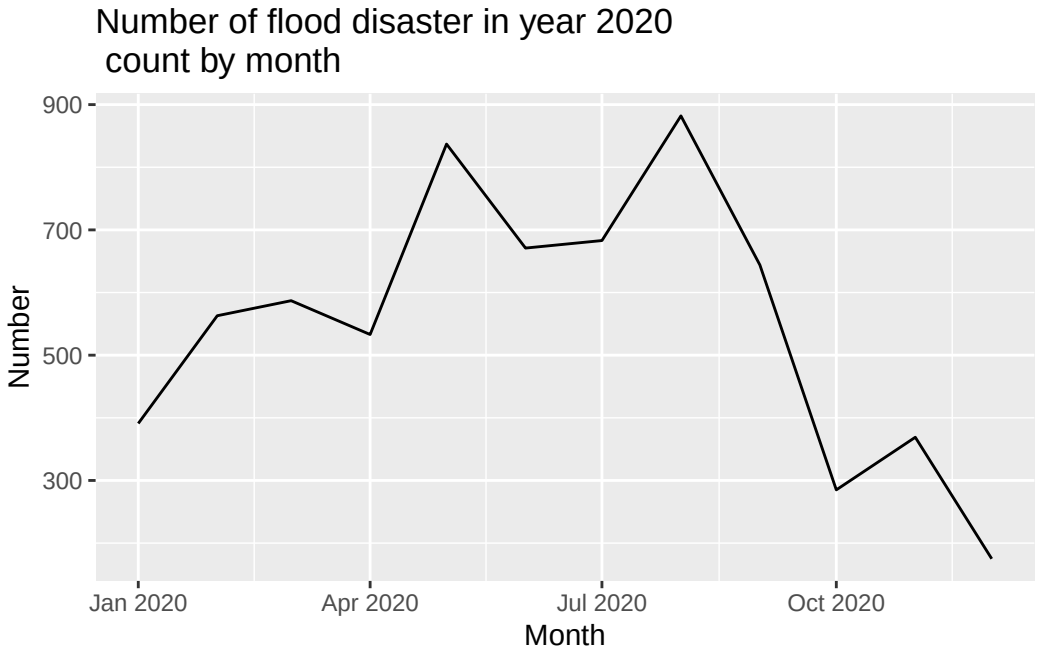
```
Warning: Removed 2 rows containing missing values (`geom_text()`).
```



As shown above, the eastern half of the United states suffer more floods than the western half. The largest amount of flood took place in Virginia in 2020.

Count the number of disaster by year

```
combine_data$BEGIN_YEARMONTH <- as.numeric(combine_data$BEGIN_YEARMONTH)
monthly_count <-combine_data %>%
  group_by(BEGIN_YEARMONTH)%>%
  summarise(count=n())

monthly_count$BEGIN_YEARMONTH <- ym(monthly_count$BEGIN_YEARMONTH)

monthly_count %>%
  ggplot(aes(x=BEGIN_YEARMONTH,y=count))+
  geom_line()+
  ggtitle("Number of flood disaster in year 2020\n count by month")+
  xlab("Month")+
  ylab("Number")
```



The flood in 2020 mainly took place during spring and summer.

####Plot the state death map

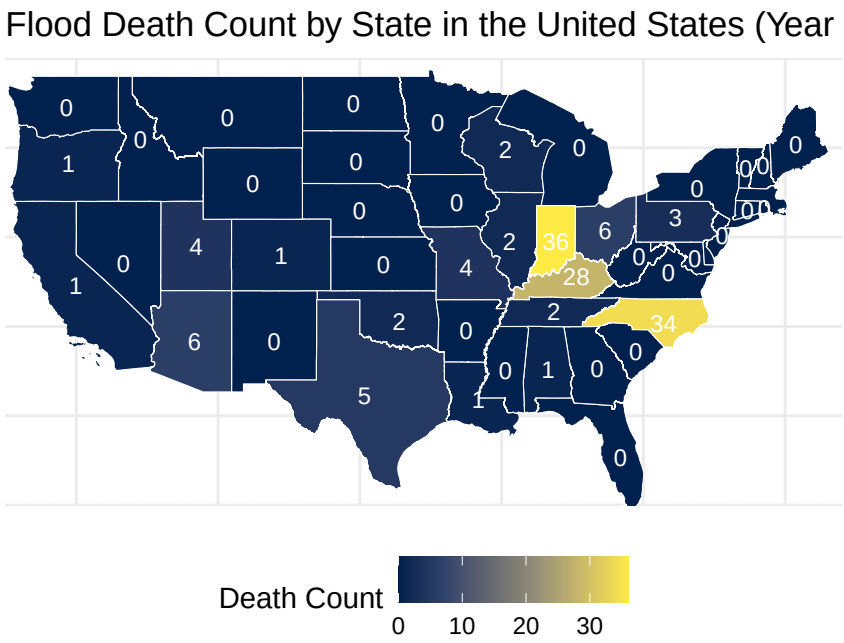
```
state_death <-
  combine_data %>%
  group_by(STATE)%>%
  summarise(Direct_death=sum(DEATHS_DIRECT),
            Indirect_death=sum(DEATHS_INDIRECT))

state_death <- state_death %>%
  mutate(state=toTitleCase(tolower(STATE)))

states_sf <- st_as_sf(tigris::states())
```

```
Retrieving data for the year 2021
```

```
states_sf$NAME <- as.character(states_sf$NAME)
map_death_data <- left_join(states_sf, state_death, by=c("NAME"="state"))
map_death_data <- map_death_data %>%
  mutate(Direct_death=replace(Direct_death,is.na(Direct_death),0),
```

```
          Indirect_death=replace(Indirect_death,is.na(Indirect_death),0))
map_death_data$Death <- map_death_data$Direct_death+map_death_data$Indirect_death

flood_death_map <- ggplot(map_death_data)+
  geom_sf(aes(fill=Death),color="white", size=0.25)+
  scale_fill_viridis_c(name="Death Count", option = "E")+
  theme_minimal()+
  geom_sf_text(aes(label=Death),color="white", size=3,check_overlap = T)+
  labs(title="Flood Death Count by State in the United States (Year 2020)")+
  theme(legend.position = "bottom",
        axis.text.x = element_blank(),
        axis.text.y=element_blank(),
        axis.ticks = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank())+
  coord_sf(xlim = c(-125,-66),
           ylim = c(25,50),expand = F)
flood_death_map
```

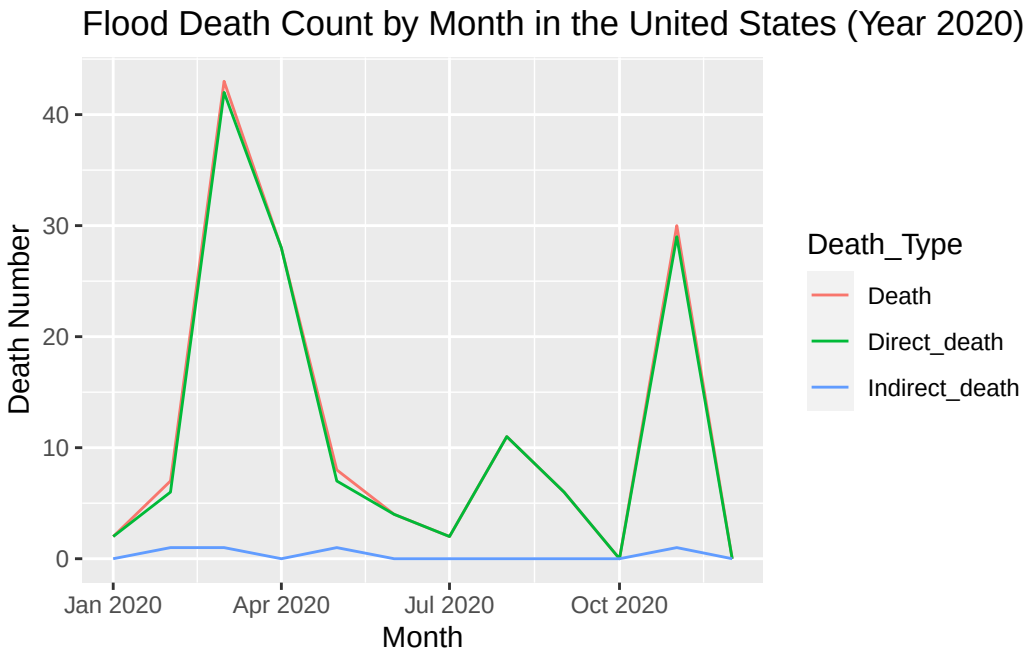
```
Warning in st_point_on_surface.sfc(sf::st_zm(x)): st_point_on_surface may not
give correct results for longitude/latitude data
```



The same as the amount of flood took place, the virginia had the most severe death account in US.

####plot death people by month

```
month_death <-
  combine_data %>%
  group_by(BEGIN_YEARMONTH)%>%
  summarise(Direct_death=sum(DEATHS_DIRECT),
            Indirect_death=sum(DEATHS_INDIRECT),
            Death=sum(DEATHS_DIRECT)+ sum(DEATHS_INDIRECT)) %>%
  mutate(Month=ym(BEGIN_YEARMONTH))

month_death_long <-month_death %>%
  pivot_longer(cols = c(-Month,-BEGIN_YEARMONTH),
               names_to="Death_Type",
               values_to = "Death_number")
month_death_long %>%
  ggplot(aes(x=Month,y=Death_number, colour=Death_Type))+
  geom_line()+ggtitle("Flood Death Count by Month in the United States (Year 2020)")+
  ylab("Death Number")
```



Most death occured during spring and winter. And it seems that most of the death are direct

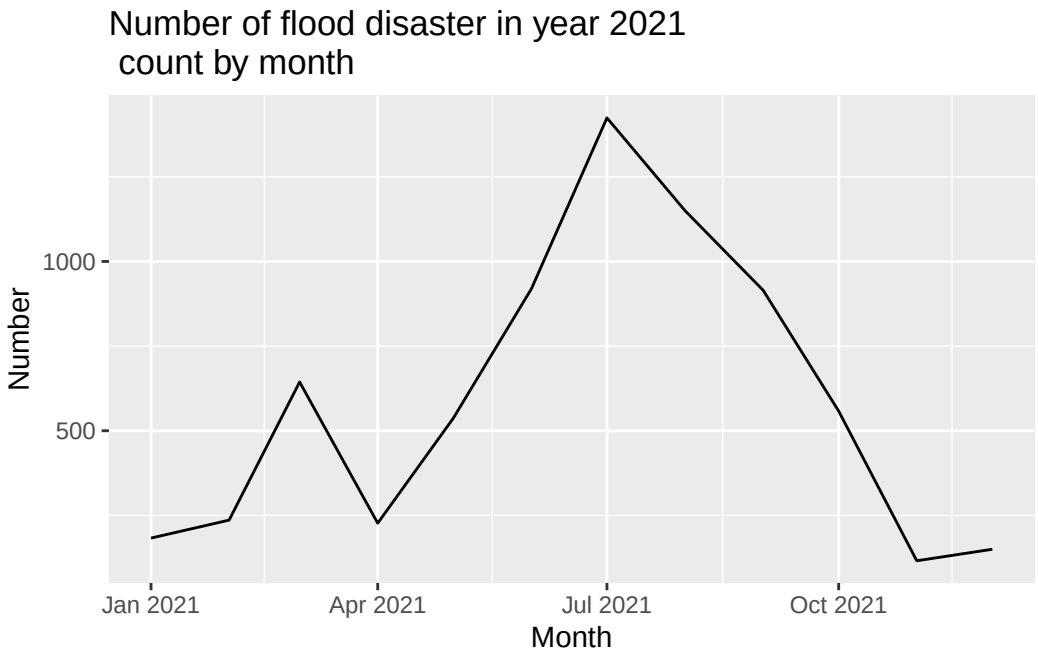
death since the line of direct death (green) is almost coverd with the total death line (red).

Case of 2021

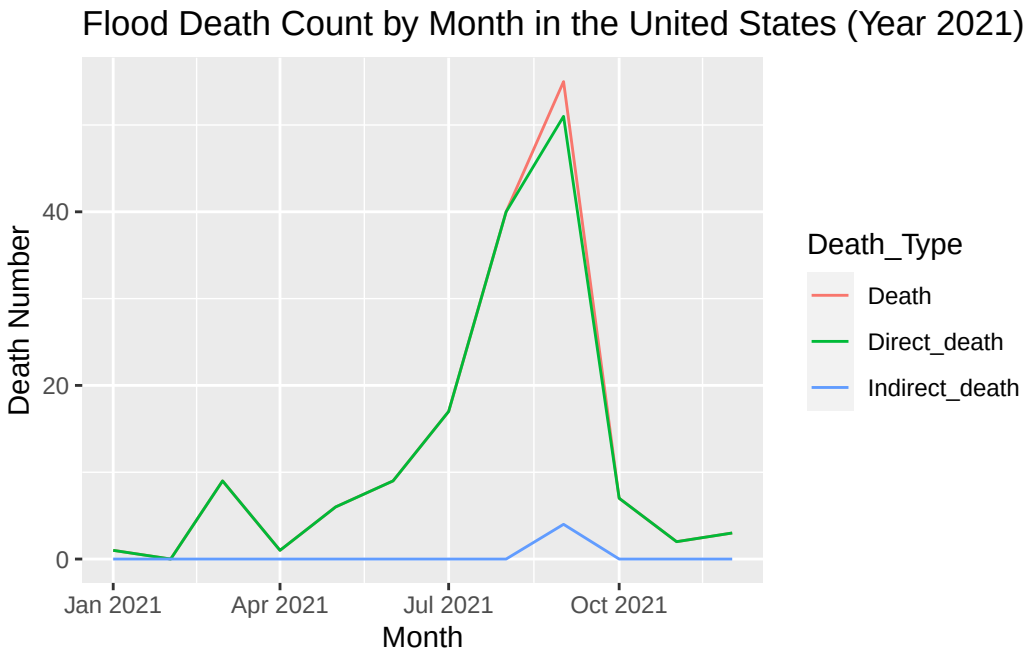
```
data_fatality_2021 <- data_fatality_2021 %>%
  as_tibble()

new_2021 <-storm_details_2021 %>%
  filter(EVENT_TYPE=="Flood"|EVENT_TYPE=="Coastal Flood"|
           EVENT_TYPE=="Flash Flood"|EVENT_TYPE=="Lakeshore Flood")


###consider damage property

combine_data_2021 <- new_2021 %>%
  mutate(DAMAGE_PROPERTY= as.numeric(as.numeric(str_replace(DAMAGE_PROPERTY, "K", ""))*100
  mutate(DAMAGE_PROPERTY=replace(DAMAGE_PROPERTY,is.na(DAMAGE_PROPERTY),0))
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `DAMAGE_PROPERTY = as.numeric(...)`.
Caused by warning:
! NAs introduced by coercion
```

```
combine_data_2021$BEGIN_YEARMONTH <- as.numeric(combine_data_2021$BEGIN_YEARMONTH)
monthly_count_2021 <-combine_data_2021 %>%
  group_by(BEGIN_YEARMONTH)%>%
  summarise(count=n())

monthly_count_2021$BEGIN_YEARMONTH <- ym(monthly_count_2021$BEGIN_YEARMONTH)

monthly_count_2021 %>%
  ggplot(aes(x=BEGIN_YEARMONTH,y=count))+
  geom_line()+
  ggtitle("Number of flood disaster in year 2021\n count by month")+
  xlab("Month")+
  ylab("Number")
```



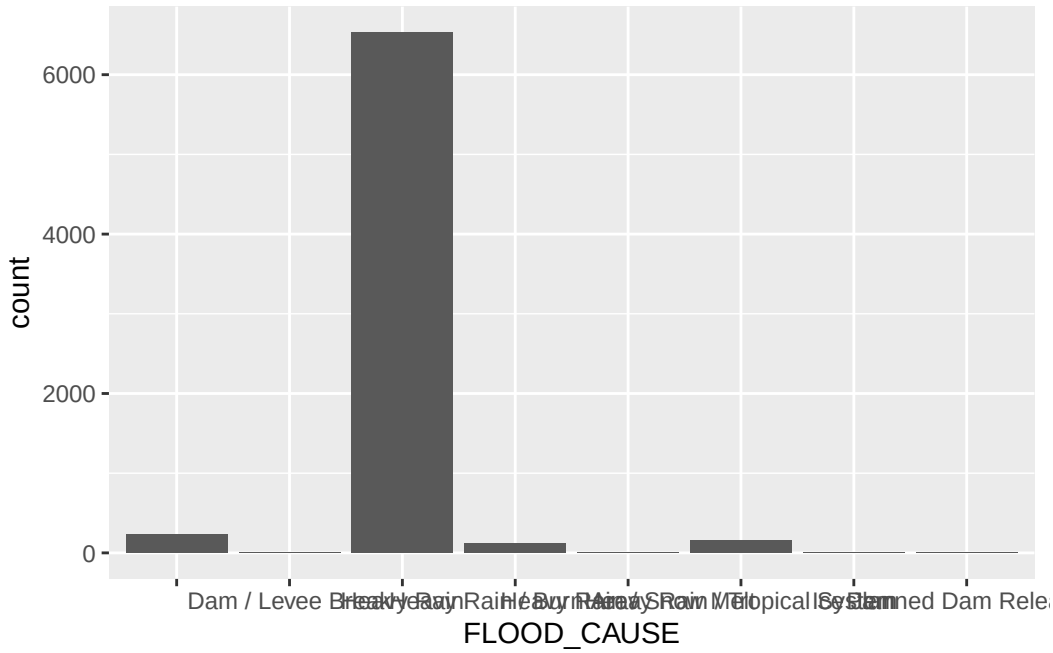

```
month_death2021 <-
  combine_data_2021 %>%
  group_by(BEGIN_YEARMONTH)%>%
  summarise(Direct_death=sum(DEATHS_DIRECT),
            Indirect_death=sum(DEATHS_INDIRECT),
            Death=sum(DEATHS_DIRECT)+ sum(DEATHS_INDIRECT)) %>%
  mutate(Month=ym(BEGIN_YEARMONTH))

month_death_long2021 <-month_death2021 %>%
  pivot_longer(cols = c(-Month,-BEGIN_YEARMONTH),
               names_to="Death_Type",
               values_to = "Death_number")
month_death_long2021 %>%
  ggplot(aes(x=Month,y=Death_number, colour=Death_Type))+
  geom_line()+ggtitle("Flood Death Count by Month in the United States (Year 2021)")+
  ylab("Death Number")
```



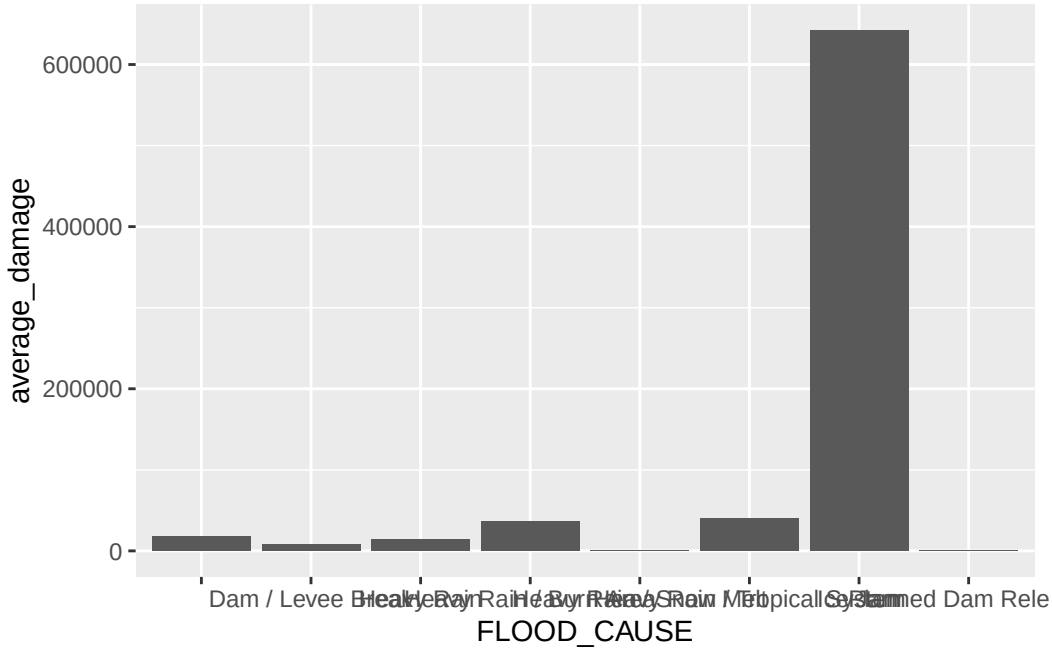
Similarly, in year 2021 most flood still occured during summer. And the amount of flood seems to be more than the amount of 2020. And most of the death are still direct instead of indirect. This indicates how fatal the floob could be.

Flood Statistics in Year 2021

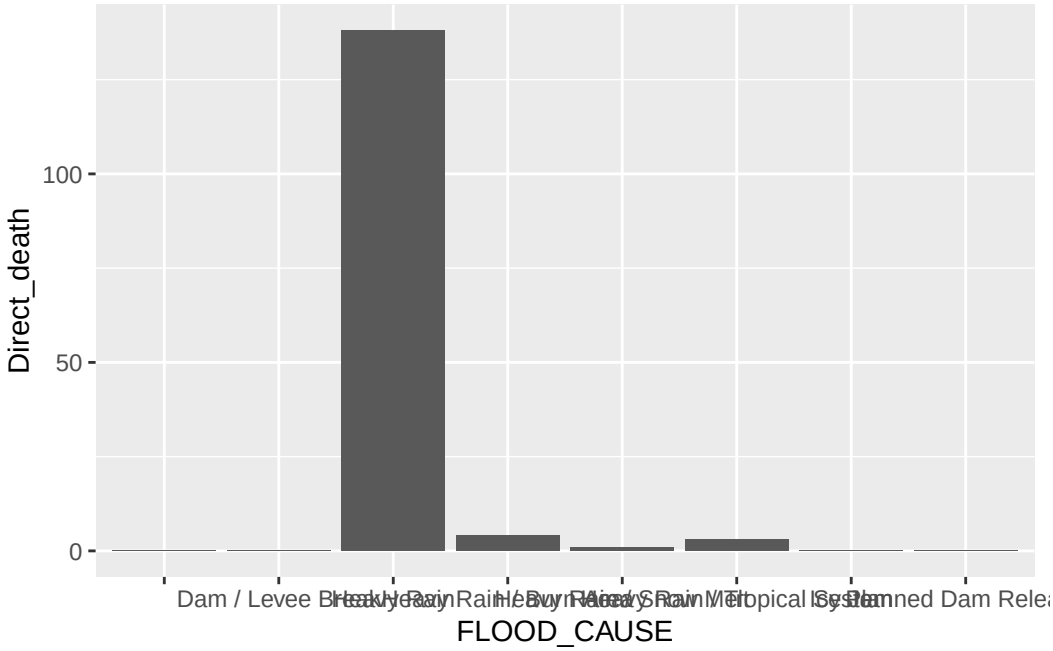
```
flood_cause_summary <- combine_data_2021 %>%
  group_by(FLOOD_CAUSE) %>%
  summarise(count=n(),
            average_damage=mean(DAMAGE_PROPERTY),
            maximum_damage=max(DAMAGE_PROPERTY),
            Direct_death=sum(DEATHS_DIRECT),
            Indirect_death=sum(DEATHS_INDIRECT),
            Death=sum(DEATHS_DIRECT)+ sum(DEATHS_INDIRECT))

flood_cause_summary %>%
  ggplot(aes(x=FLOOD_CAUSE,y=count))+
  geom_bar(stat = "identity")
```



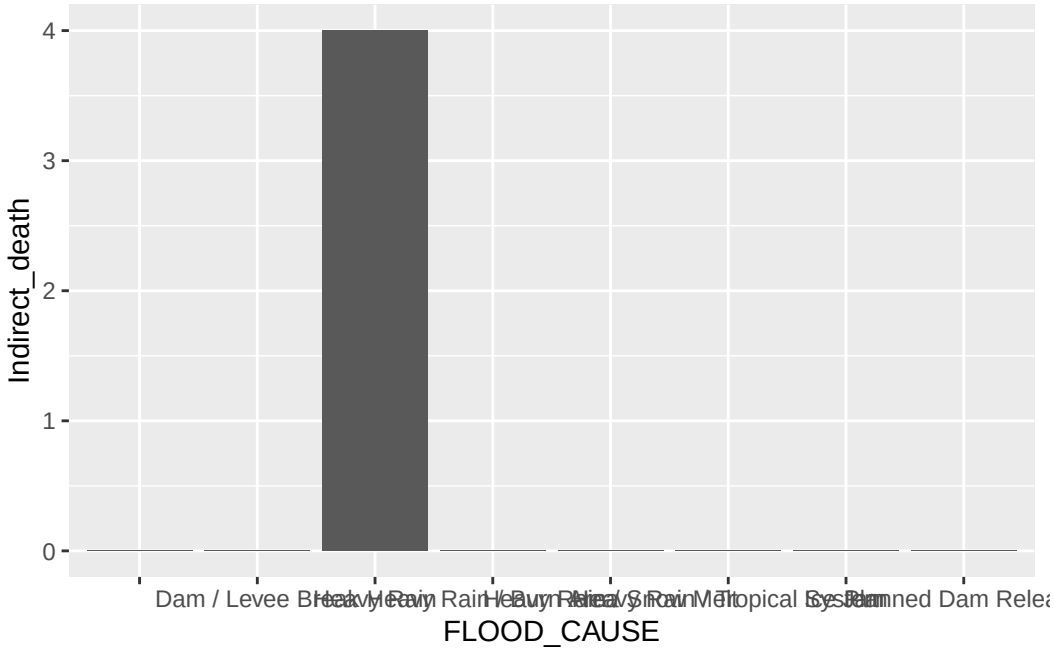

```
flood_cause_summary %>%
  ggplot(aes(x=FLOOD_CAUSE,y=average_damage))+
  geom_bar(stat = "identity")
```




```
flood_cause_summary %>%
  ggplot(aes(x=FLOOD_CAUSE,y=Direct_death))+
  geom_bar(stat = "identity")
```




```
flood_cause_summary %>%
  ggplot(aes(x=FLOOD_CAUSE,y=Indirect_death))+
  geom_bar(stat = "identity")
```



Most of the flood were caused by heavy rain; The most severe damage were caused by ice jam; And most death no matter what kinds were caused by heavy rain.