The harm and impact of floods on humans

Wuge Li

2023-10-31

Background of the project

Floods are widely recognized as one of the most devastating natural disasters in the U.S. I aim to delve deeper into the dangers posed by floods, the economic implications of these events, and to discern any patterns regarding the communities most impacted by them. To address these queries, I will be utilizing a dataset spanning the years 2020 to 2021, sourced from NOAA.

The first step I need to do is to read the data.

```
d_2020 <- read.csv("StormEvents_details-ftp_v1.0_d2020_c20230927.csv", header = T)
d_2021 <- read.csv("StormEvents_details-ftp_v1.0_d2021_c20231017.csv", header = T)

f_2020 <- read.csv("StormEvents_fatalities-ftp_v1.0_d2020_c20230927.csv", header = T)
f_2021 <- read.csv("StormEvents_fatalities-ftp_v1.0_d2021_c20231017.csv", header = T)
l_2020 <- read.csv("StormEvents_locations-ftp_v1.0_d2020_c20230927.csv", header = T)
l_2021 <- read.csv("StormEvents_locations-ftp_v1.0_d2021_c20231017.csv", header = T)

diaster <- read.csv("DisasterDeclarationsSummaries.csv", header = T)
sum <- read.csv("FemaWebDisasterSummaries.csv", header = T)
```

In my view, column names are crucial to understanding a data. They provide insights into the content and context of the data.

```
colnames(d_2020)
```

```
##  [1] "BEGIN_YEARMONTH"    "BEGIN_DAY"          "BEGIN_TIME"
##  [4] "END_YEARMONTH"      "END_DAY"            "END_TIME"
##  [7] "EPISODE_ID"         "EVENT_ID"           "STATE"
## [10] "STATE_FIPS"         "YEAR"               "MONTH_NAME"
## [13] "EVENT_TYPE"         "CZ_TYPE"            "CZ_FIPS"
## [16] "CZ_NAME"            "WFO"                "BEGIN_DATE_TIME"
## [19] "CZ_TIMEZONE"        "END_DATE_TIME"      "INJURIES_DIRECT"
## [22] "INJURIES_INDIRECT"  "DEATHS_DIRECT"      "DEATHS_INDIRECT"
## [25] "DAMAGE_PROPERTY"    "DAMAGE_CROPS"       "SOURCE"
## [28] "MAGNITUDE"          "MAGNITUDE_TYPE"     "FLOOD_CAUSE"
## [31] "CATEGORY"           "TOR_F_SCALE"        "TOR_LENGTH"
## [34] "TOR_WIDTH"          "TOR_OTHER_WFO"      "TOR_OTHER_CZ_STATE"
## [37] "TOR_OTHER_CZ_FIPS"  "TOR_OTHER_CZ_NAME"  "BEGIN_RANGE"
## [40] "BEGIN_AZIMUTH"      "BEGIN_LOCATION"     "END_RANGE"
## [43] "END_AZIMUTH"        "END_LOCATION"       "BEGIN_LAT"
## [46] "BEGIN_LON"          "END_LAT"            "END_LON"
## [49] "EPISODE_NARRATIVE"  "EVENT_NARRATIVE"    "DATA_SOURCE"
```

In this data, it's talking about

1. the impact of natural disasters in 2020.
2. It shows when the disasters begin and end. (exact time)
3. Where the disasters happen. (State, begin and end location, begin and end longitude and latitude)
4. What kind of the disasters happened. (For example, Thunderstorm Wind, flood and so on)
5. Damages that the disasters lead. (injury direct and indirect, deaths direct and indirect, damage property and crops)

Since, column names for both d_2020 and d_2021 are the same, so I only show the column names of d_2020 here

Questions about the data

1. How dangerous are floods?
2. How expensive?
3. Is there any pattern to the kinds of communities that suffer losses from floods?

Clean, tidy and show the data

The next step we need to do is to tidy the data.

Initially, I discovered that the data encompasses various disaster types. However, my interest lies solely in flood-related incidents, so I employed a filter to isolate all flood events.

```
tidy_d_2020 <- d_2020 |>
  filter(str_detect(string = EVENT_TYPE, pattern = "(?i)flood"))

tidy_d_2021 <- d_2021 |>
  filter(str_detect(string = EVENT_TYPE, pattern = "(?i)flood"))

unique(tidy_d_2020$EVENT_TYPE)
```

```
## [1] "Flash Flood"     "Flood"           "Coastal Flood"   "Lakeshore Flood"
```

After applying the filter, I used a unique function to examine the residual event types and determined that they all pertained to floods. Next, we aim to determine the number of flood events that occurred in the years 2020 and 2021.

```r
total_floods_2020 <- nrow(tidy_d_2020)
total_floods_2021 <- nrow(tidy_d_2021)
total_floods_2020
```

```
## [1] 6602
```

```r
total_floods_2021
```

```
## [1] 7059
```

Utilizing the nrow function, it was determined that there were 6,602 flood incidents in 2020, and 7,059 flood incidents in 2021. Not all flood events result in significant damage. Therefore, I aim to ascertain the number of floods that qualified as disasters. To do this, I will refer to the disaster form data that I have previously reviewed.

```r
tidy_disaster2020 <- diaster |>
  filter(fyDeclared == 2020) |>
  filter(incidentType == "Flood")
unique(tidy_disaster2020$disasterNumber)
```

```
## [1] 4553 4539 4519 4477 4475 4466
```

```r
disaster_floods_2020 <- length(unique(tidy_disaster2020$disasterNumber))
disaster_floods_2020
```

```
## [1] 6
```

```r
tidy_disaster2021 <- diaster |>
  filter(fyDeclared == 2021) |>
  filter(incidentType == "Flood")
unique(tidy_disaster2021$disasterNumber)
```

```
## [1] 4621 4620 4609 4606 4605 4604 4595 4571
```

```r
disaster_floods_2021 <- length(unique(tidy_disaster2021$disasterNumber))
disaster_floods_2021
```

```
## [1] 8
```

I organized the disaster form data and applied the unique function. It was revealed that there were only 6 flood events classified as disasters in 2020 and 8 such events in 2021.

```r
percentage_2020 <- (disaster_floods_2020 / total_floods_2020) * 100
percentage_2021 <- (disaster_floods_2021 / total_floods_2021) * 100


print(paste0(percentage_2020, "%"))
```

```
## [1] "0.0908815510451378%"
```

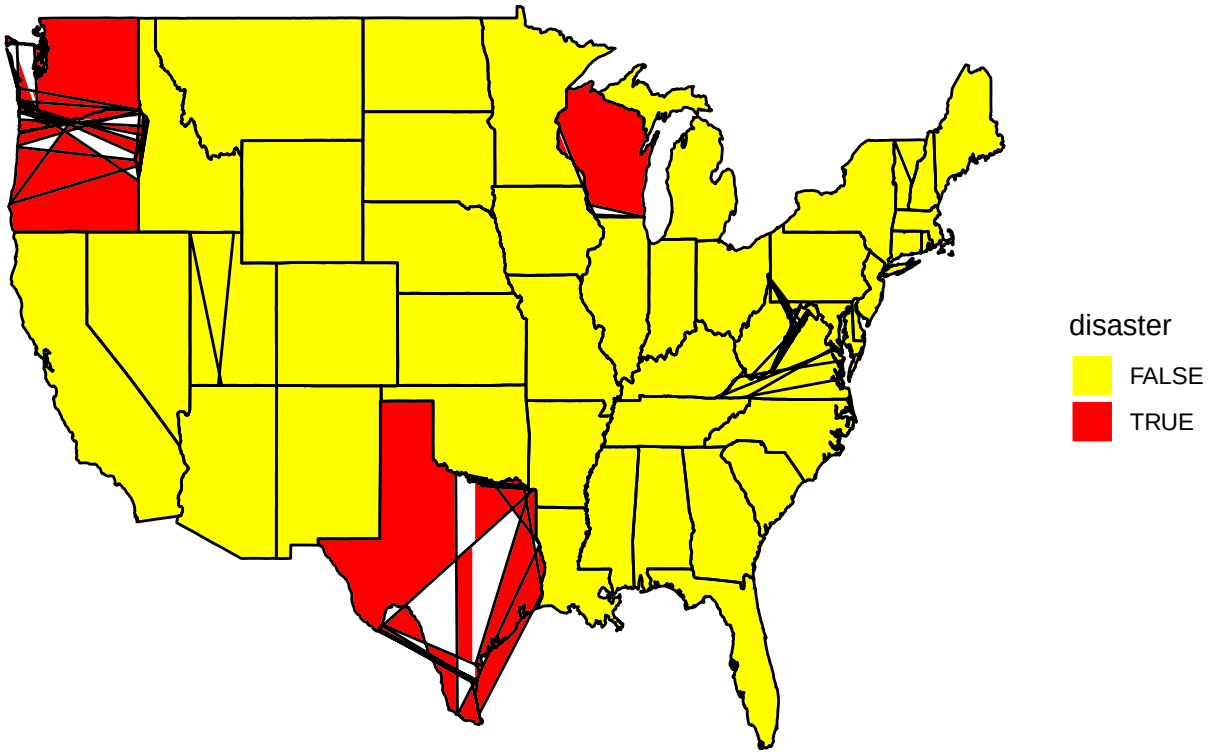
```r
print(paste0(percentage_2021, "%"))
```

```
## [1] "0.113330500070832%"
```

Upon calculating the percentages, it was found that only 0.09% of flood events in 2020 were classified as disasters, and for 2021, the figure was slightly higher at 0.11%. I want to create a graph to display the areas affected by flood disasters in 2020 and 2021.

```r
disaster_data <- tidy_disaster2020

disaster_data$disaster <- TRUE

disaster_data_unique <- aggregate(disaster ~ state, data = disaster_data, FUN = any)

state_names <- c(ND = "north nakota", OR = "oregon", TX = "texas", WA = "washington",
                 WI = "wisconsin")

disaster_data_unique <- disaster_data_unique |>
  mutate(state = state_names[state])

us_map <- map_data("state")

map_data <- merge(us_map, disaster_data_unique, by.x = "region", by.y = "state",
                  all.x = TRUE)

map_data$disaster[is.na(map_data$disaster)] <- FALSE

ggplot(data = map_data, aes(x = long, y = lat, group = group, fill = disaster)) +
  geom_polygon(color = "white") +
  geom_path() +
  scale_fill_manual(values = c("FALSE" = "yellow", "TRUE" = "red")) +
  labs(title = "Disaster States in the US in 2020") +
  theme_void()
```

Disaster States in the US in 2020



This graph illustrates that the majority of flood disasters which happened in 2020 occurred predominantly in:

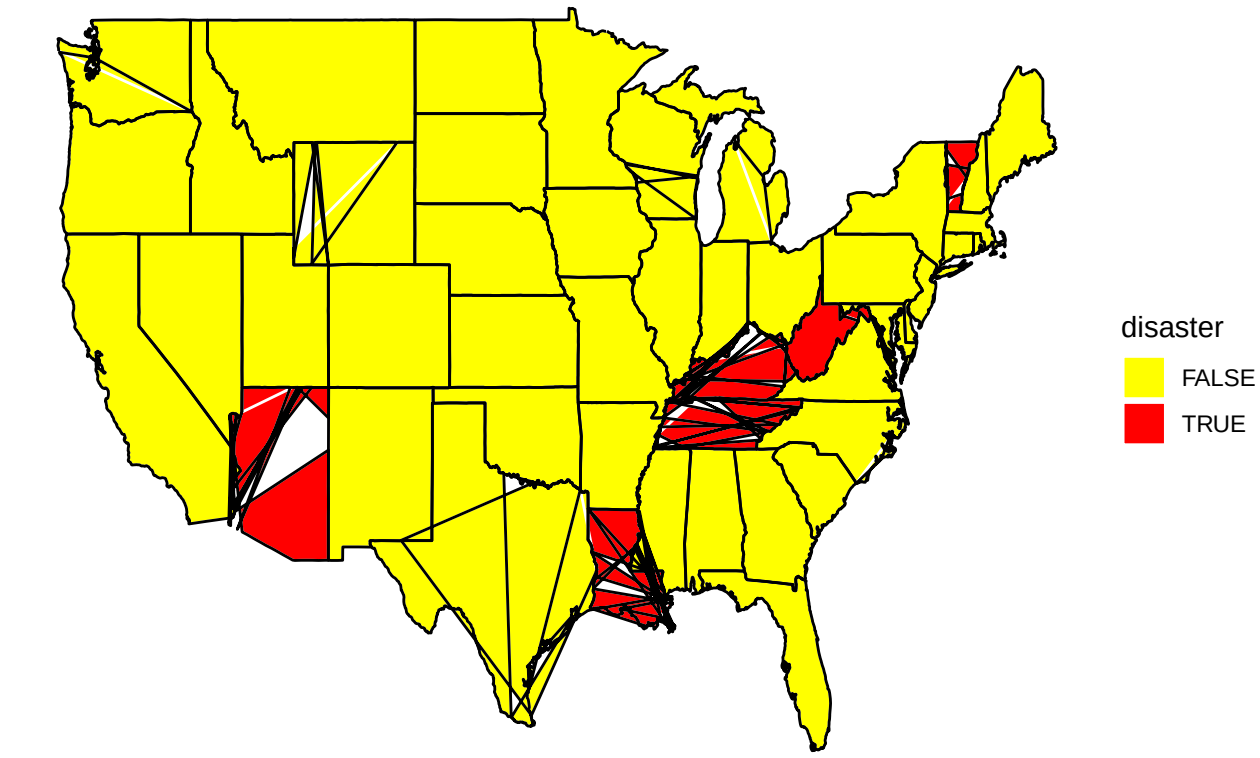
Washington, located in the upper left corner

Texas at the bottom center

Wisconsin at the upper right of the map. Then we looked at the flood disaster at 2021.

```
disaster_data <- tidy_disaster2021

disaster_data$disaster <- TRUE

disaster_data_unique <- aggregate(disaster ~ state, data = disaster_data, FUN = any)

state_names <- c(AZ = "arizona", HI = "hawaii", KY = "kentucky", LA = "louisiana",
                 PR = "puerto rico", TN = "tennessee", VT = "vermont",
                 WV = "west virginia")

disaster_data_unique <- disaster_data_unique |>
  mutate(state = state_names[state])

us_map <- map_data("state")

map_data <- merge(us_map, disaster_data_unique, by.x = "region", by.y = "state",
                  all.x = TRUE)

map_data$disaster[is.na(map_data$disaster)] <- FALSE
```

```
ggplot(data = map_data, aes(x = long, y = lat, group = group, fill = disaster)) +
  geom_polygon(color = "white") +
  geom_path() +
  scale_fill_manual(values = c("FALSE" = "yellow", "TRUE" = "red")) +
  labs(title = "Disaster States in the US in 2021") +
  theme_void()
```



This graph illustrates that the majority of flood disasters which happened in 2021 occurred predominantly in:

Arizona, located in the bottom left

Louisiana at the bottom right

Kentucky, Tennessee and West Virginia at the middle right of the map

Vermont at the upper right corner of the map.

Then, I want to focus on the influence of these flood disasters. I will use the Disaster Summaries form

This is the description of the form

```
data_fields <- data.frame(
  Name = c(
    "disasterNumber",
    "totalNumberIaApproved",
```

```
    "totalAmountIhpApproved",
    "totalAmountHaApproved",
    "totalAmountOnaApproved",
    "totalObligatedAmountPa",
    "totalObligatedAmountCatAb",
    "totalObligatedAmountCatC2g",
    "paLoadDate",
    "iaLoadDate",
    "totalObligatedAmountHmgp",
    "hash",
    "lastRefresh",
    "id"
  ),
  Description = c(
    "A number which show the unique disaster",
    "The number of disaster approved applications for IA",
    "Dollars approved for IHP",
    "Dollars approved for Housing Assistance",
    "Dollars approved for Other Needs Assistance",
    "Dollars available for Public Assistance grants",
    "Dollars for Emergency Work Public Assistance (Categories A & B)",
    "Dollars for Permanent Work Public Assistance (Categories C to G)",
    "Date PA data was updated",
    "Date IA data was updated",
    "Dollars obligated for the Hazard Mitigation Grant Program",
    "MD5 hash of fileds and values of the record",
    "Date the record was last updated",
    "Unique ID assigned to the record"
  ),
  stringsAsFactors = FALSE
)

kable(data_fields, caption = "Data Fields Description") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"), full_width = F)
```

Then, I want to tidy this form; I only need the flood disaster data for 2020 and 2021.

```
tidy_sum <- sum |>
  filter(disasterNumber %in% c(4553, 4539, 4519, 4477, 4475, 4466, 4621,
                               4620, 4609, 4606, 4605, 4604, 4595, 4571)) |>
  mutate(year = ifelse(row_number() <= 6, "2020", "2021"))
```

```
ggplot(data = tidy_sum, aes(x = as.factor(disasterNumber),
                            y = totalNumberIaApproved, fill = year)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = totalNumberIaApproved), position = position_dodge(width = 0.9),
            vjust = -0.25) +
  labs(x = "Disaster Number", y = "The number of approved applications",
       title = "The number of approved applications for each disaster") +
  scale_fill_manual(values = c("2020" = "blue", "2021" = "red")) +
```

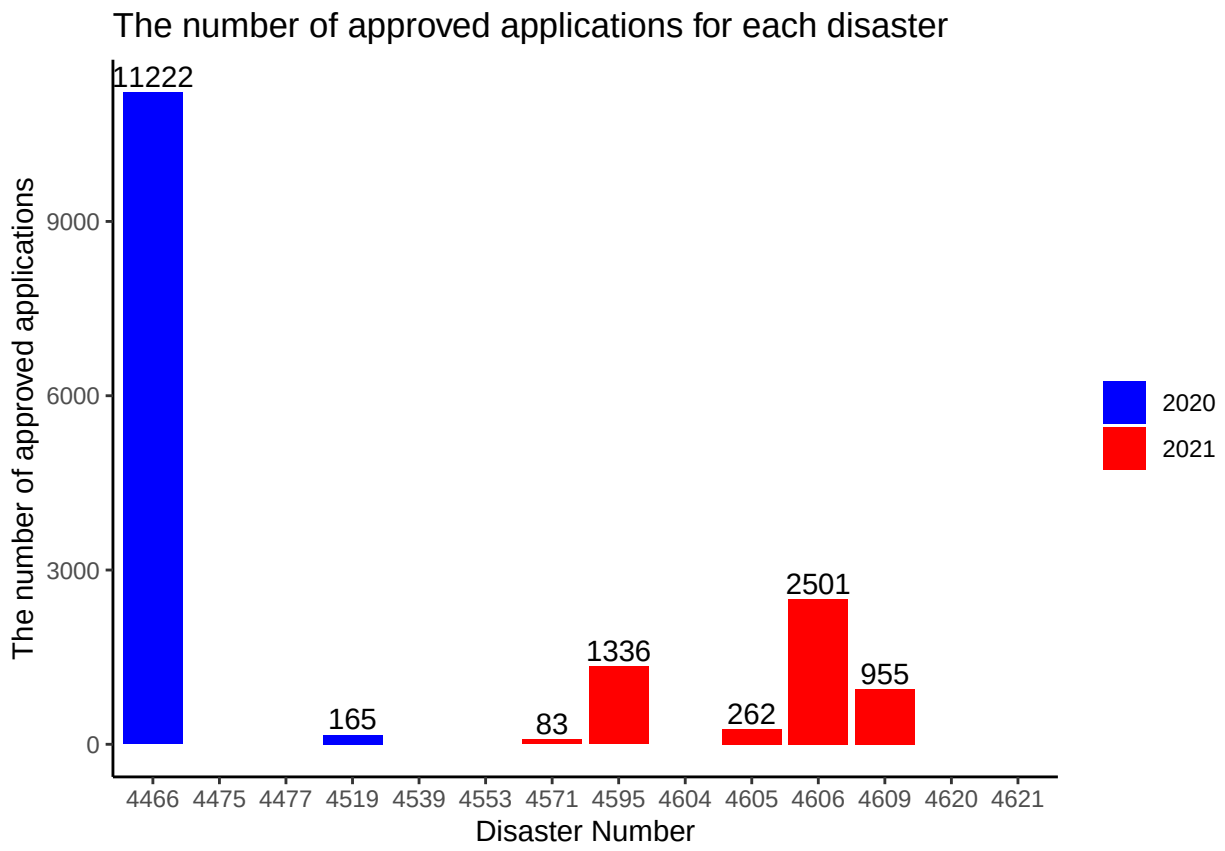
Table 1: Data Fields Description

Name	Description
disasterNumber	A number which show the unique disaster
totalNumberIaApproved	The number of disaster approved applications for IA
totalAmountIhpApproved	Dollars approved for IHP
totalAmountHaApproved	Dollars approved for Housing Assistance
totalAmountOnaApproved	Dollars approved for Other Needs Assistance
totalObligatedAmountPa	Dollars available for Public Assistance grants
totalObligatedAmountCatAb	Dollars for Emergency Work Public Assistance (Categories A & B)
totalObligatedAmountCatC2g	Dollars for Permanent Work Public Assistance (Categories C to G)
paLoadDate	Date PA data was updated
iaLoadDate	Date IA data was updated
totalObligatedAmountHmgp	Dollars obligated for the Hazard Mitigation Grant Program
hash	MD5 hash of fileds and values of the record
lastRefresh	Date the record was last updated
id	Unique ID assigned to the record

```
  theme_classic() +
  theme(legend.title = element_blank())
```

```
## Warning: Removed 7 rows containing missing values (`position_stack()`).
```

```
## Warning: Removed 7 rows containing missing values (`geom_text()`).
```

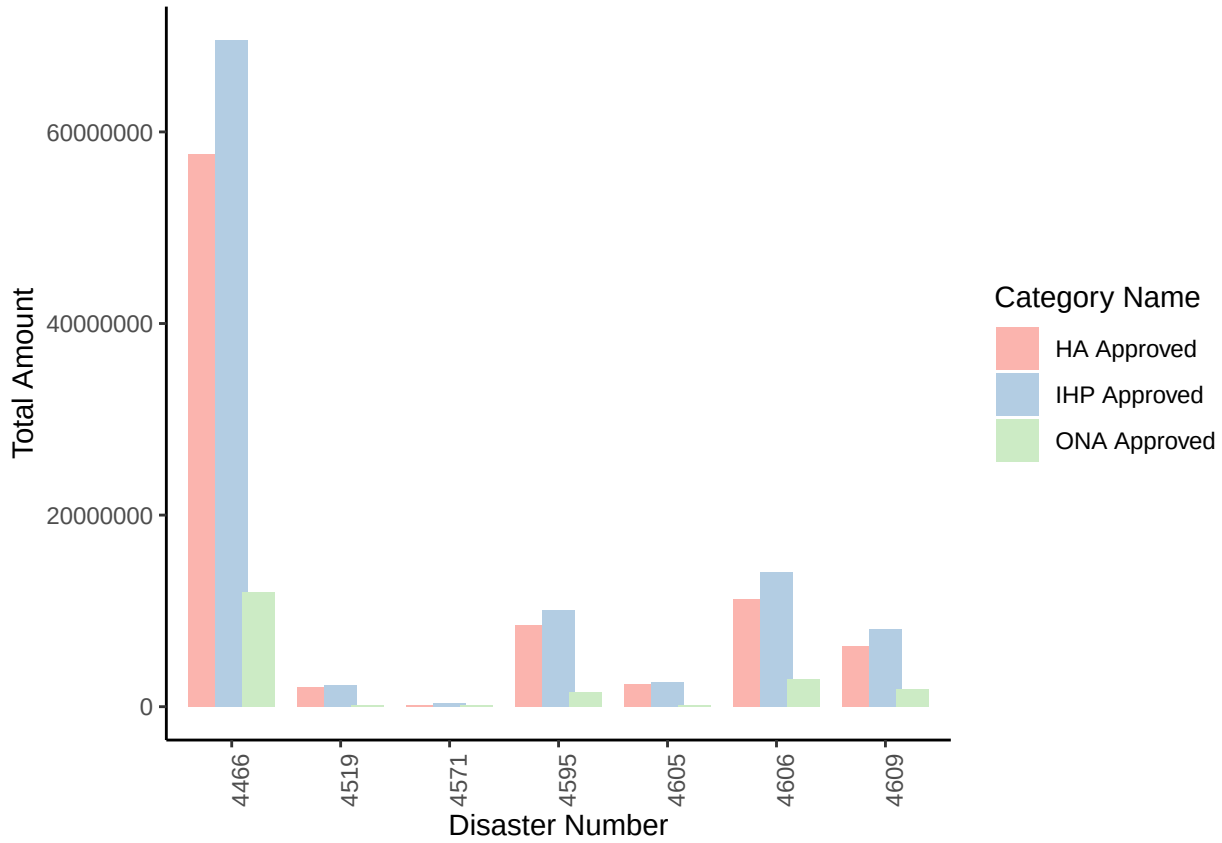


From the graph, we can observe the following:

1. Disaster 4466, which occurred in 2020, had a significantly higher number of approved applications compared to the other disasters, with a total of 11,222 applications.
2. For the year 2020, there are only two disasters represented (4466 and 4519), and the number of approved applications for disaster 4519 is relatively low (165 applications) compared to disaster 4466.
3. For the year 2021, there are more disasters represented (4571, 4595, 4604, 4605, 4606, 4609, 4620, 4621), with the number of approved applications ranging from a low of 83 (disaster 4571) to a high of 2501 (disaster 4605).
4. Among the 2021 disasters, 4605, 4606, and 4609 have the most approved applications, with 2501, 1336, and 955 applications, respectively.
5. It appears that there may be a general trend of fewer approved applications in 2021 compared to 2020, although this graph only shows one disaster for 2020.
6. The range of disaster numbers shows that the data set likely spans a sequence of events that are numerically close, possibly indicating they occurred within a similar time frame.
7. No applications are represented for disaster numbers 4475, 4477, 4519 (for 2021), and 4620 for the year 2020, which could indicate there were no applications or the data is not available/recorded for these particular disasters in those years.

From the description, we can ascertain that the fields totalAmountIhpApproved, totalAmountHaApproved, and totalAmountOnaApproved represent the same category of data, namely the amounts approved by different assistance programs.

```
totalA <- tidy_sum |>
  select(c(disasterNumber, totalAmountHaApproved, totalAmountIhpApproved,
           totalAmountOnaApproved, year)) |>
  na.omit()

totalA_long <- totalA |>
  select(disasterNumber, totalAmountHaApproved, totalAmountIhpApproved,
         totalAmountOnaApproved) |>
  pivot_longer(
    cols = -disasterNumber,
    names_to = "Category",
    values_to = "Amount"
  ) |>
  mutate(Category = factor(Category, levels = c("totalAmountHaApproved",
                                                "totalAmountIhpApproved",
                                                "totalAmountOnaApproved")))

ggplot(totalA_long, aes(x = factor(disasterNumber), y = Amount, fill = Category)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.75)) +
  scale_fill_brewer(palette = "Pastel1", labels = c("HA Approved", "IHP Approved", "ONA Approved")) +
  labs(x = "Disaster Number", y = "Total Amount", fill = "Category Name") +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

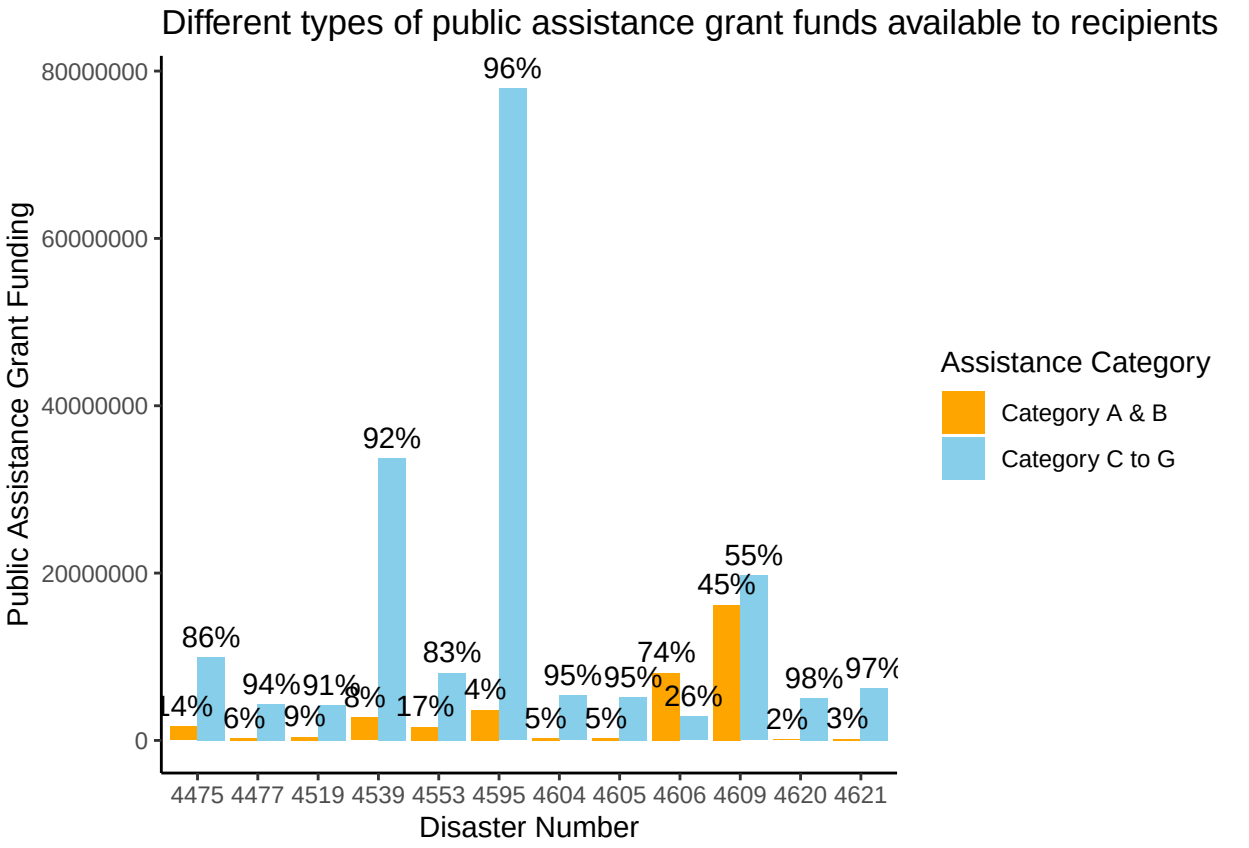


Based on this graph and the number of approved applications for each disaster graph, we know: Disaster number 4466 stands out in both graphs, indicating it had both a high total amount approved (particularly in the Ha category) and a high number of approved applications in 2020. This means that disaster 4466 is very dangerous and brings great losses. There is a variance in the distribution of amounts and the number of applications approved between different disasters and between the years 2020 and 2021.

Then, I want to look at totalObligatedAmountCatAb and totalObligatedAmountCatC2g.

```
totalO <- tidy_sum |>
  select(c(disasterNumber, totalObligatedAmountCatAb, totalObligatedAmountCatC2g)) |>
  na.omit()



# Calculate totals and percentages
totalO$total <- totalO$totalObligatedAmountCatAb + totalO$totalObligatedAmountCatC2g
totalO$percentAb <- (totalO$totalObligatedAmountCatAb / totalO$total) * 100
totalO$percentC2g <- (totalO$totalObligatedAmountCatC2g / totalO$total) * 100


# Reshape for plotting
totalO_long <- gather(totalO, category, amount, c(totalObligatedAmountCatAb,
                                                  totalObligatedAmountCatC2g))

ggplot(totalO_long, aes(x = factor(disasterNumber), y = amount, fill = category)) +
  geom_bar(stat = "identity", position = position_dodge()) +
```

```r
geom_text(aes(label = sprintf("%.0f%%", amount/total0$total*100)),
          vjust = -0.5, position = position_dodge(width = 0.9)) +
scale_fill_manual(values = c("orange", "skyblue"),
                  labels = c("Category A & B", "Category C to G")) +
labs(x = "Disaster Number",
     y = "Public Assistance Grant Funding",
     title = "Different types of public assistance grant funds available to recipients",
     fill = "Assistance Category") +  # This 'fill' attribute sets the legend title
theme_classic()
```



In summary, while there are exceptions, Category C to G generally receives a higher percentage of public assistance grant funds compared to Category A & B across various disasters. The reasons behind this distribution would likely be due to the nature of the costs and damages involved in those categories, which could include more extensive or expensive recovery and rebuilding efforts.