Exploratory Data Analysis (EDA) for floods in 2020-2021

Jiaxin Li (U77902942)

Abstract

"Floods are one of the most devastating natural disasters". With the help of various data resources such as NOAA data, FEMA data, etc., I would like to explore these datasets by using EDA tools such as visualization, transformation, etc., to explore the characteristics of the flood events that occurred in various parts of the United States in 2020-2021, the casualties and property damage caused by the flood disasters, and the damage to the people living in the different counties and counties, etc. The report presents the main results of my EDA work.

Catalogue

1 Main question to be explored

How dangerous are floods?

How expensive?

Is there any pattern to the kinds of communities that suffer losses from floods? And so on.

2 Data origin and preparation

2.1 Data origin

(1) NOAA Dataset

Data frames I used:

a) StormEvents_details-ftp_v1.0_d2020_c20230927.csv.gz
b) StormEvents_details-ftp_v1.0_d2021_c20231017.csv.gz
c) StormEvents_locations-ftp_v1.0_d2020_c20230927.csv.gz
d) StormEvents_locations-ftp_v1.0_d2021_c20231017.csv.gz

(2) Open FEMA Dataset: Disaster Declarations Summaries

Disaster Declarations Summaries is a summarized dataset describing all federally declared disasters. This dataset lists all official FEMA Disaster Declarations, beginning with the first disaster declaration in 1953 and features all three disaster declaration types: major disaster, emergency, and fire management assistance. The dataset includes declared recovery programs and geographic areas.

(3) Open FEMA Dataset: FEMA Web Disaster Summaries

This data set contains financial assistance values, including the number of approved applications, as well as individual, public assistance, and hazard mitigation grant amounts.This is raw, unedited data from FEMA's National Emergency Management Information System (NEMIS) and as such is subject to a small percentage of human error. The financial information is derived from NEMIS and not FEMA's official financial systems.

(4) Census datasets:

Data frames I used:

a) File Census Download_2023-10-23T135225.zip
b) File Census Download_2023-10-23T140133.zip
c) File Census Download_2023-10-23T140147.zip

2.2 Data preparation for analysis

After obtaining the necessary dataset for analysis, I primarily conducted the following data cleaning tasks on the dataset:

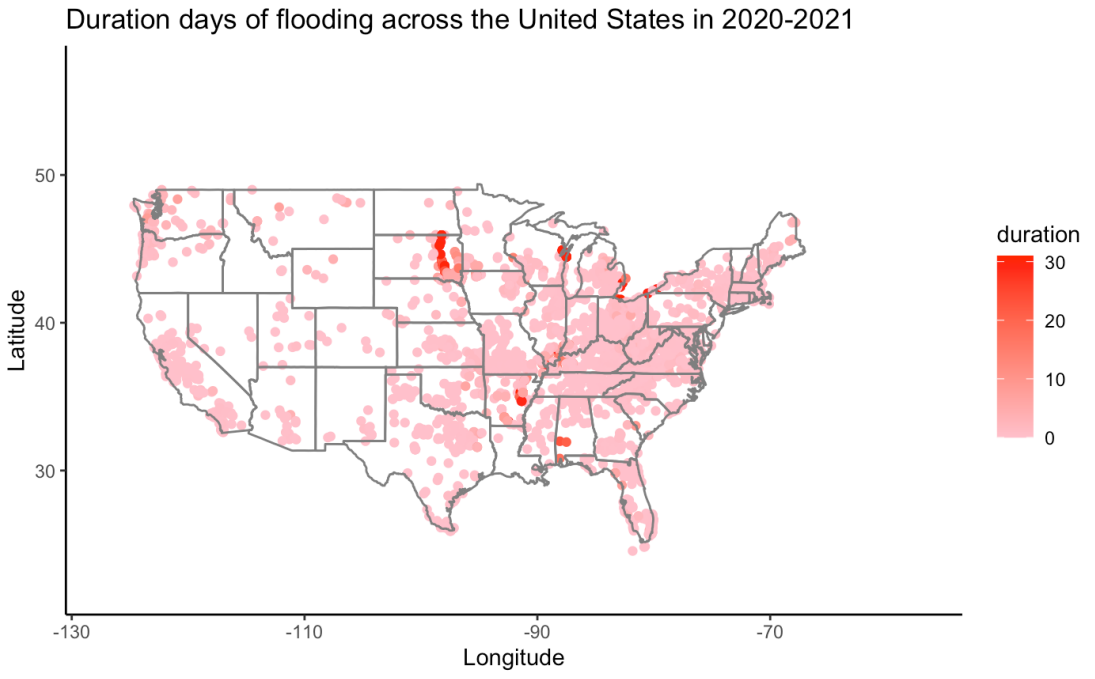
(1) I filtered out disaster data occurring between 2020 and 2021. Additionally, I created a subset of disaster records within this period that specifically pertained to the types of disasters categorized as "Flash Flood" and "Flood." I calculated the duration of each disaster event by comparing their start and end times.

(2) For the census dataset, I merged three distinct tables containing various demographic information of the affected population. This integration was achieved by utilizing the common "CZ_FIPS" column variable, which indicates the county where the disaster occurred. This process resulted in a comprehensive dataset encompassing information on the affected population's numbers, gender, age, and other relevant data.

(3) Leveraging the county where each disaster took place as an intermediary factor, I merged the census information dataset with the storm events information dataset. This fusion enabled the creation of a dataset for analyzing the impact of different storm events, particularly "Flash Flood" and "Flood," on the population in various counties across the United States during the years 2020 and 2021.

3 The story of floods in 2020-2021 (Main findings of the EDA)

3.1 The relationship between flood duration and hazard (injuries, deaths, damage)



Firstly, through visualization I found that the duration days of flooding varied in different parts of the US, and I wondered if it was safe to say that those places that flooded for a longer period were more severely affected. So, I intend to investigate whether there is a positive correlation between the duration days of floods and the extent of damage they cause.

It seems safe to first assume that most floods occurring between 2020 and 2021 that last for a longer period would cause more casualties and damage. Then I analyzed whether the duration days of a flood can be considered to have a positive correlation with the number of casualties and damages caused by the flood, i.e., the shorter the duration days of the flood, the smaller the number of casualties and damages caused by the flood.

The output from R:

Pearson Correlation between duration and injuries_total: -0.01055809

Pearson Correlation between duration and deaths_total: -0.01004481

Pearson Correlation between duration and damage_total: 0.01144618

Based on the Pearson correlation coefficient results, I found that:

(1) The Pearson correlation between "duration" and "injuries_total" is approximately -0.01055809. This value is close to zero, and the negative correlation suggests that there may be a weak negative association between the duration of the flood and the total injuries;

(2) The Pearson correlation between "duration" and "deaths_total" is approximately -0.01004481. This value is also very close to zero, and the negative correlation implies that there may be a very weak or almost no linear relationship between the duration of the flood and the number of deaths.

(3) The Pearson correlation between "duration" and "damage_total" is approximately 0.01144618. This value is likewise close to zero, and the positive correlation suggests there is very little or almost no linear relationship between the duration of the flood and the total damage to property and crops.

In summary, the Pearson correlation coefficient results suggest that there is no strong linear relationship between these variables. This may imply that the linear relationships are very weak or potentially nonlinear. Therefore, we cannot simply and arbitrarily say that the shorter the duration of a flood, the fewer casualties, damages, etc. it causes.

3.2 How dangerous are floods? And how expensive?

3.2.1 Analysis of casualties caused by floods in 2020-2021

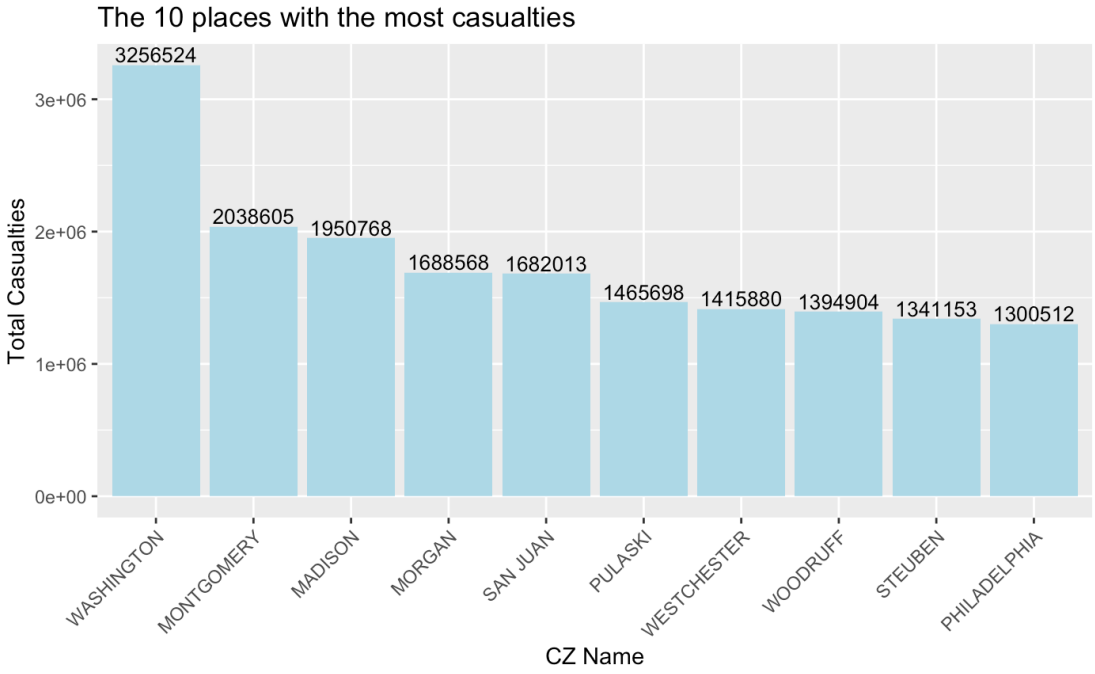
The distribution of casualties caused by different flooding events:

The output from R:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1311	3933	13110	88236	62600	3256524

(Column variables calculated: "INJURIES_DIRECT", "INJURIES_INDIRECT", "DEATHS_DIRECT", "DEATHS_INDIRECT")

By analyzing the data on casualties caused by floods that occurred in the United States between 2020 and 2021, the distribution of total casualties caused by different floods shows that there were many casualties due to floods during these two years. On average, a disaster-level flood caused close to 90,000 casualties, with the worst flood disasters causing more than three million casualties. Thus, floods during the two years posed a huge threat to people's lives and needed to be taken seriously.



3.2.2 Analysis of economic loss caused by floods in 2020-2021

To calculate the economic loss caused by floods, I used the column variable: totalObligatedAmountHmgp (Total Obligated Amount HMGP), which represents the total amount obligated under the Hazard Mitigation Grant Program in dollars. The total amount obligated represents the amount obligated for Regular Project Costs, Planning Costs, Initiative Project Costs, Recipient Management Costs, Recipient Admin Costs and Subrecipient Admin Costs.

The distribution of casualties caused by different flooding events:

The output from R:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	3747827	670979	1671835350

(Column variables calculated: "totalObligatedAmountHmgp")

Analysis of the data on the amount of economic assistance expenditures due to floods occurring in the United States between 2020 and 2021 shows that the distribution of economic losses due to different floods (i.e., the amount of economic assistance needed for disaster recovery for a single flood disaster) indicates that the amount of economic assistance expenditures needed due to floods during these two years was significant. On average, a disaster-level flood required more than $3.7 million in economic assistance, and the worst flood disaster required more than $1.6 billion in economic assistance. Thus, the floods during the two-year period caused enormous economic losses and high economic costs to the Government and the people.

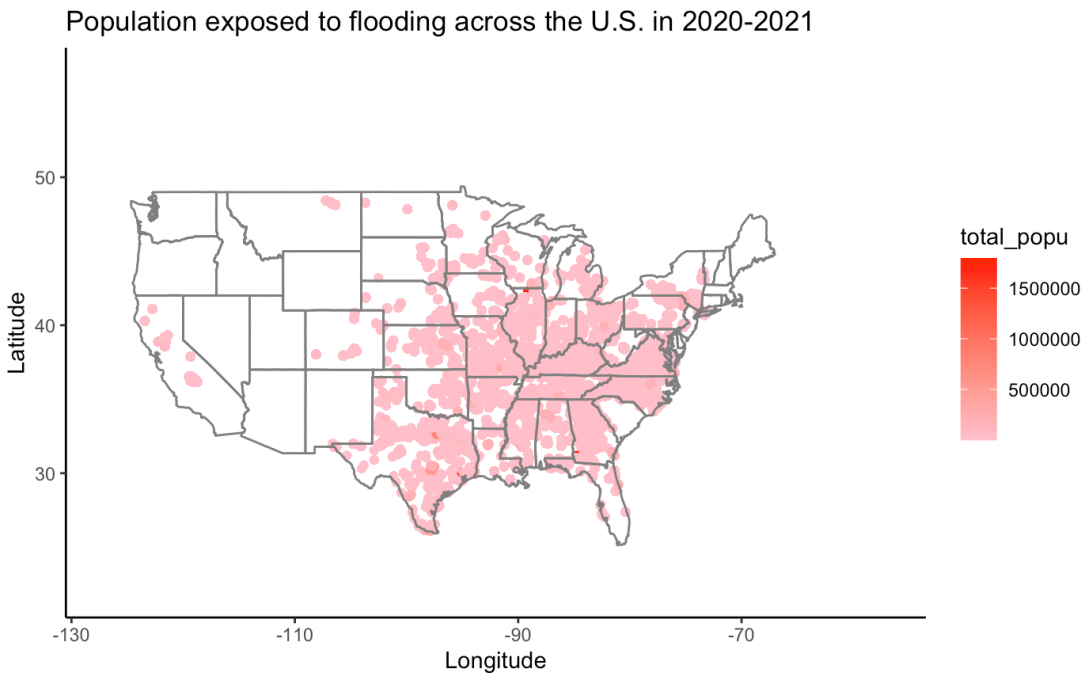
3.3 Analysis of the impacts of flood events on people

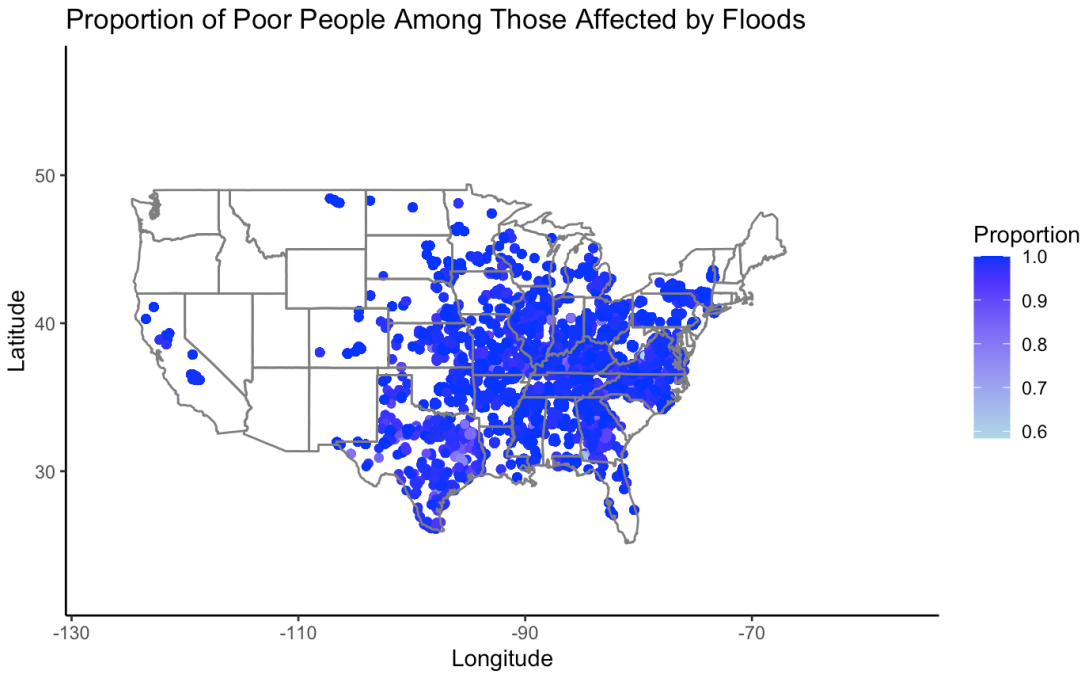
Based on the analyses in 3.1 and 3.2, I have a general impression of the flood events between 2020 and 2021. However, I am even more interested in knowing how the lives and livelihoods of people living in different parts of the United States have been affected.

Analyses of the flood dataset showed that it was difficult to gain insight into the impacts of flooding on people in different areas between 2020 and 2021 without combining detailed data on the affected populations in different counties. Specifically, I wanted to know how many people were affected in different areas, the gender and age distribution of the affected population, which counties were more severely affected, and so on. Therefore, in 3.3, I will use a dataset that combines information on floods and affected populations to explore the flood-affected residents in different counties between 2020 and 2021.

3.3.1 Number of people affected by floods and proportion of poverty in 2020-2021

(1) Number of people affected by floods and proportion



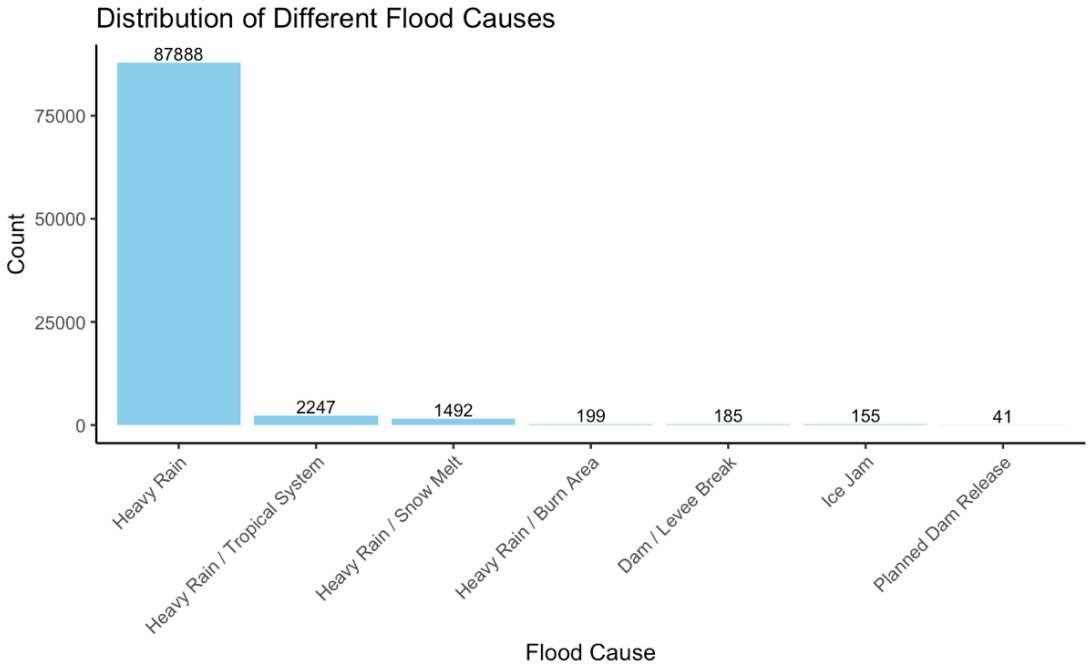


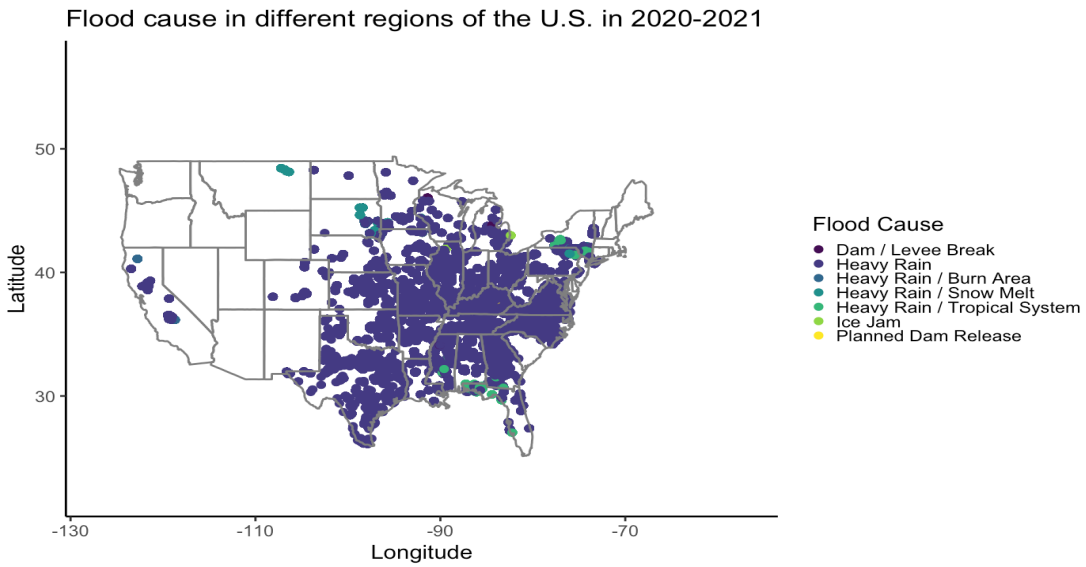
The population affected by flooding between 2020 and 2021 is concentrated in the central, eastern, and southern regions. The fact that most of the affected population will be classified as poor, most of the affected population will be poor, supports the findings of the previous section

3.2 that the floods are indeed dangerous and costly in terms of the security risks and economic losses that they pose to the affected population.

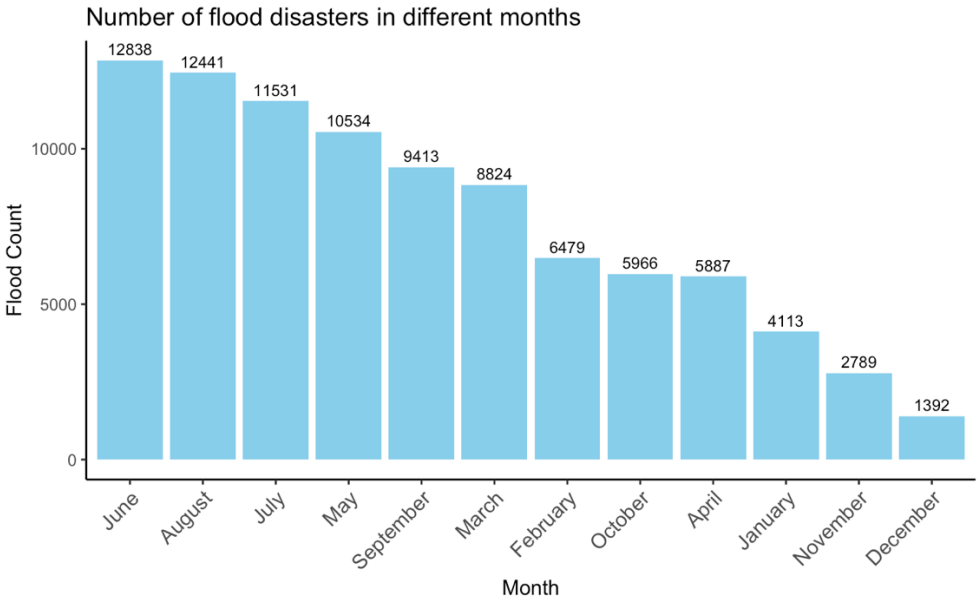
(2) Flood cause in 2020-2021

Having learnt about the distribution of the affected population, I would like to explore why this is the case, that is, why the center, the east and the south were more severely affected, and what factors are influencing the vulnerability of some areas to flooding. I therefore visualized data on the causes of flooding over the two years.



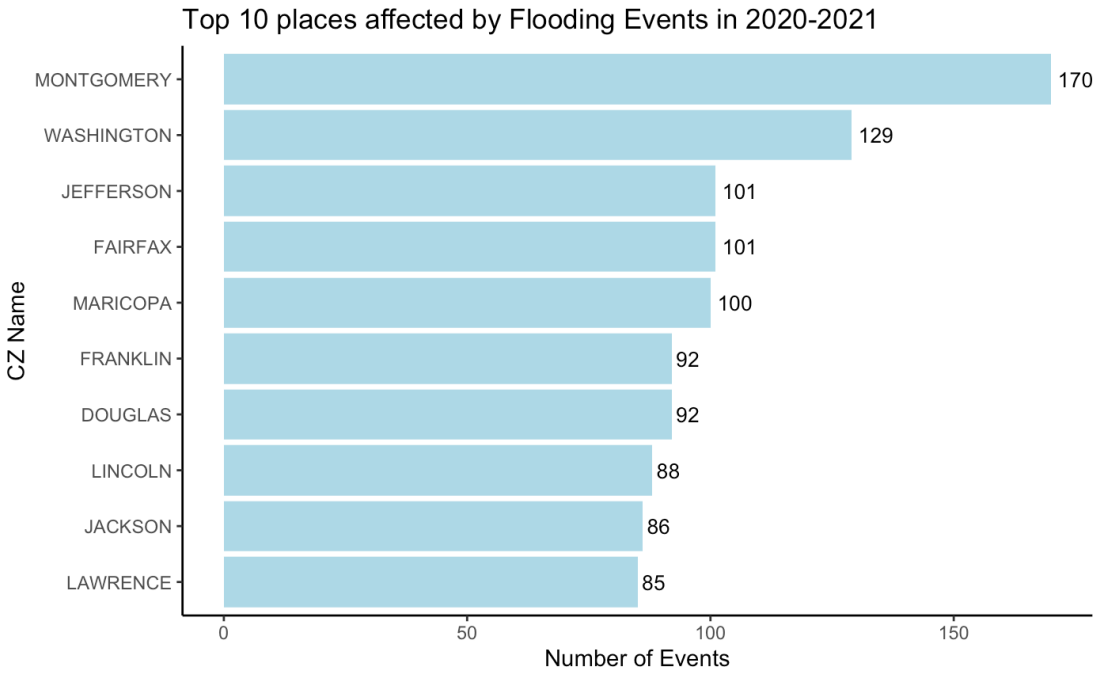


The first histogram of the different causes of floods shows that the most likely cause of floods in these two years is “Heavy Rain”, areas prone to heavy rainfall will be more prone to flooding. The histograms of the causes of floods in different regions between 2020 and 2021 show that most of the floods in the central, eastern, and southern regions are caused by "Heavy Rain". As a result, the Central, Eastern and Southern regions are more prone to heavy rainfall due to climatic conditions and are therefore more likely to be affected by flooding.



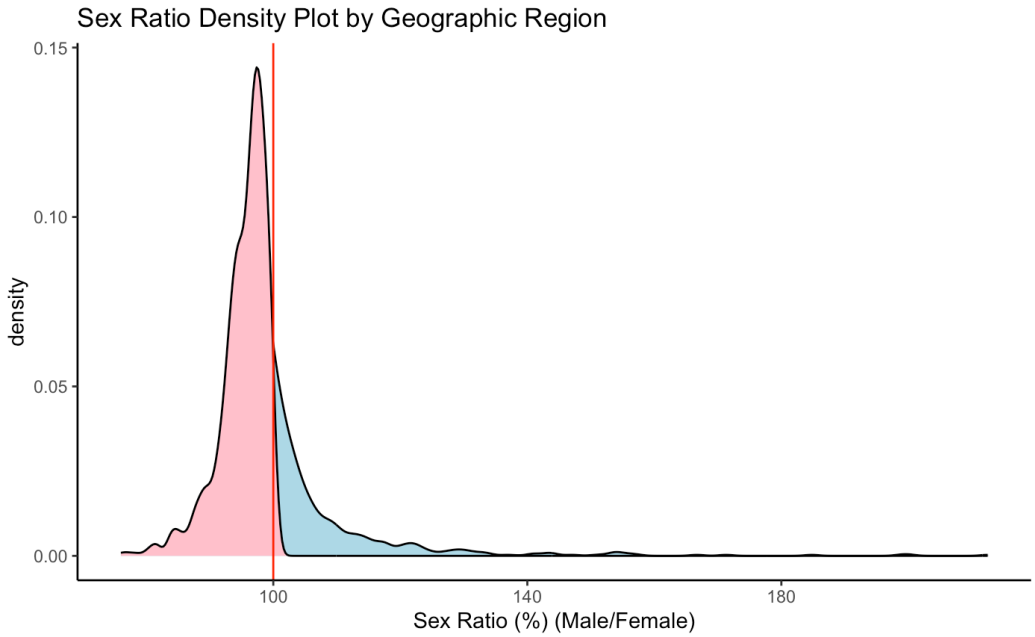
In addition, visualization of the number of flood events in the different months between 2020 and 2021 shows that many flood events occur in the summer months, with fewer occurring in the spring and autumn months compared to the summer months, and relatively fewer

occurring in the winter months. This is in conjunction with the previous exploration that heavy rainfall is most likely to lead to flooding, and that months with more precipitation are more prone to heavy rainfall leading to more flooding events.



3.3.2 Gender and age of the flood-affected population in 2020-2021

(2) Sex Ratio (%) (Male / Female)



5 counties where Male were significantly more severely affected than Female (Sex Ratio Top5):

Geography	Geographic Area Name	Sex Ratio (%)
0500000US13259	Stewart County, Georgia	212.5
0500000US22125	West Feliciana Parish, Louisiana	200.2
0500000US12125	Union County, Florida	198.9
0500000US13309	Wheeler County, Georgia	184.8
0500000US48169	Garza County, Texas	171.3

5 counties where Female were significantly more severely affected than Male (Sex Ratio Tail5):

Geography	Geographic Area Name	Sex Ratio (%)
0500000US13235	Pulaski County, Georgia	76.0
0500000US21189	Owsley County, Kentucky	77.1
0500000US51595	Emporia city, Virginia	78.4
0500000US21201	Robertson County, Kentucky	80.1
0500000US29117	Livingston County, Missouri	80.5

Six counties where the severity was the same for males and females (Sex Ratio = 100%)

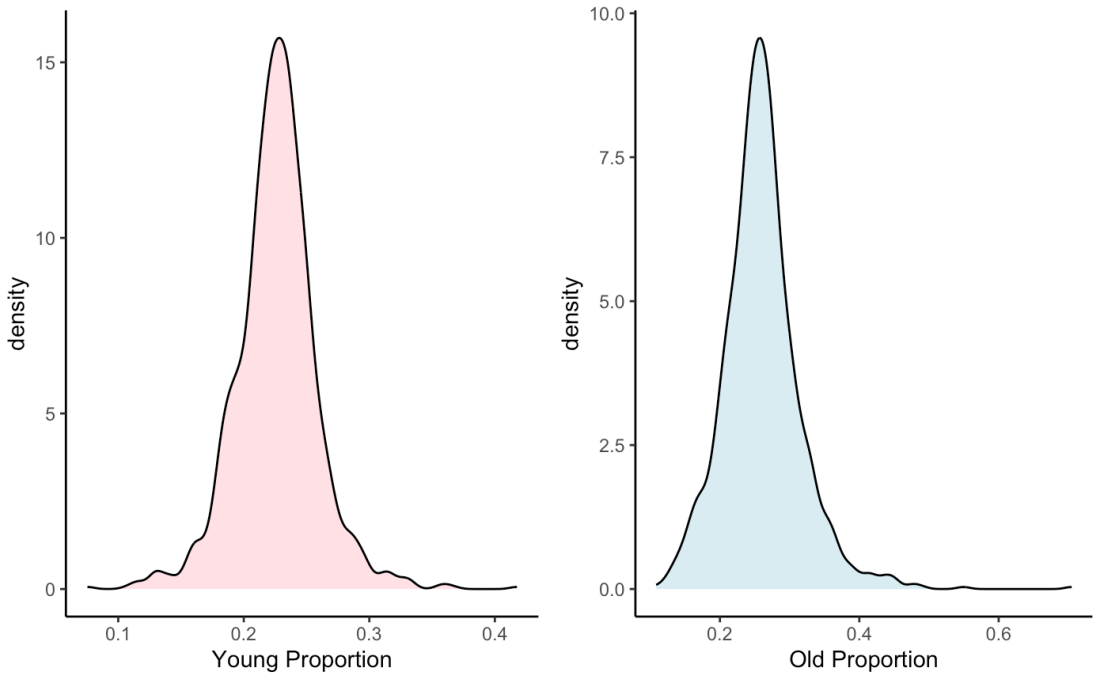
Geography	Geographic Area Name	Sex Ratio (%)
0500000US06107	Tulare County, California	100.0
0500000US17195	Whiteside County, Illinois	100.0
0500000US21177	Muhlenberg County, Kentucky	100.0
0500000US47139	Polk County, Tennessee	100.0
0500000US48493	Wilson County, Texasri	100.0
0500000US72107	Orocovis Municipio, Puerto Rico	100.0

Most of the affected population is not uniformly male or female, with some counties having more females than males as flood victims and others having more males as flood victims. Of all the affected counties, only 6 counties had a sex ratio of 1 male to 1 female.

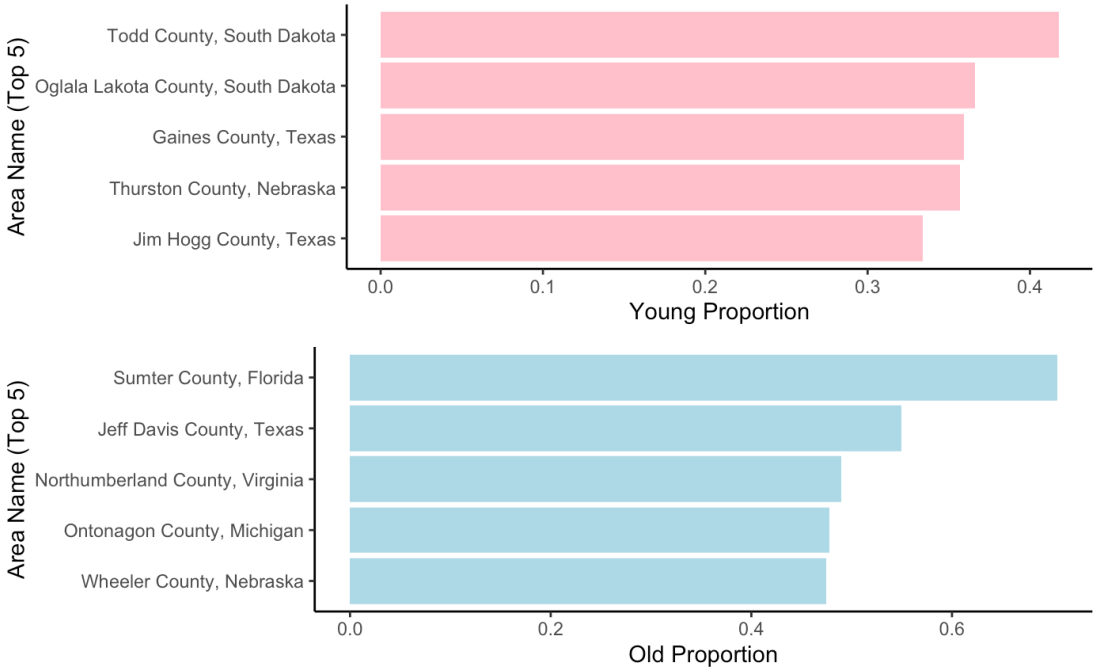
(2) Proportion of young (under 18) and old (over 60) people in the affected population

I focus on two age groups of the affected population (young people under the age of 18 and people over the age of 60) as a proportion of the total affected population in the local area, to explore whether these two groups, which are more vulnerable than young adults, are more susceptible to flooding.

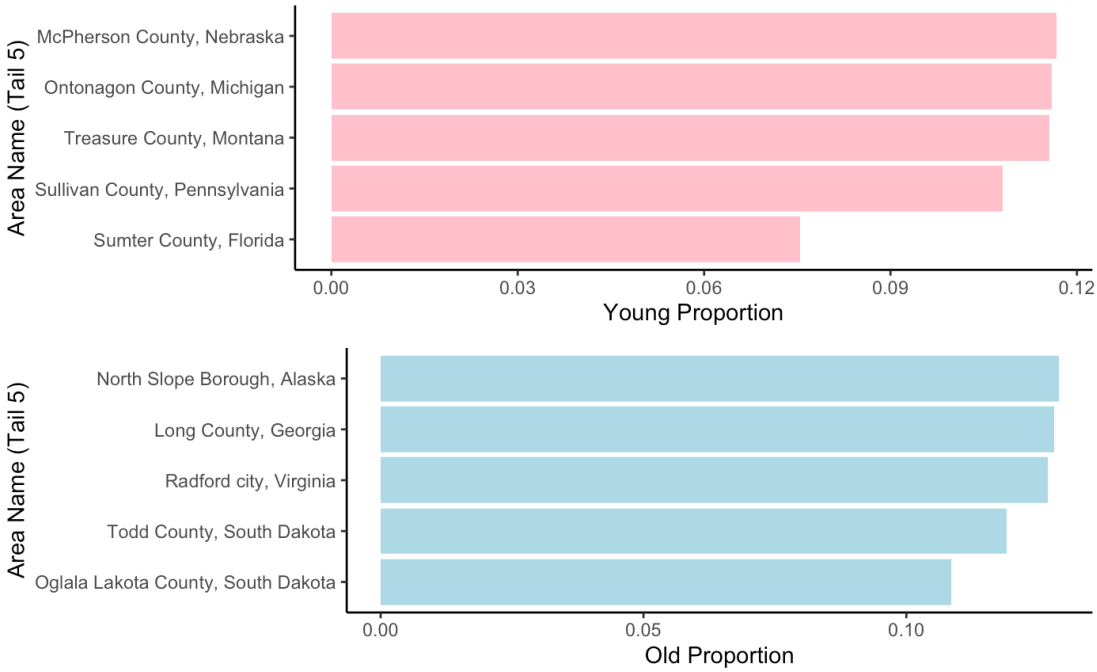
Distribution of the proportion of young people (under 18) or older people (over 60) in the affected population:



10 counties with the highest proportion of young people (under 18) or the highest proportion of older people (over 60):



10 counties with the lowest proportion of young people (under 18) or the highest proportion of older people (over 60):



The results of the data analysis suggest that there is no single trend or pattern here, and that the situation varies from county to county. The distribution of the proportional data shows that:

(1) Most counties have between fifteen and thirty percent of the affected population under the age of 18. Todd County has the highest percentage, with over forty percent of the total affected population under the age of 18, and Sumter County has the lowest percentage, with less than ten percent of the total affected population under the age of 18.

(2) Most counties had between ten and forty percent of the affected population over the age of 60. Sumter County had the highest percentage with over seventy percent of the total affected population over the age of 60, and Sumter County had the lowest percentage with over ten percent of the total affected population under the age of 18.

Overall, older adults over the age of 60 are more likely to be affected than younger adults under the age of 18, perhaps because older adults are more likely to be affected due to limited mobility.

4 Summary

In summary, I seem to be able to mentally map out the story line of the flood events in 2020-2021. As areas such as the central, eastern, and southern parts of the country are prone to heavy rainfall, which is the most likely cause of flooding, floods are more likely to occur in these heavy rainfall-prone areas and have a huge impact on the lives and properties of the people there. And many flood events occur in the summer months, with fewer occurring in the spring and autumn months compared to the summer months, and relatively fewer occurring in the winter months. This is in conjunction with the previous exploration that heavy rainfall is most likely to lead to flooding, and that months with more precipitation are more prone to heavy rainfall leading to more flooding events.

The hazards of these floods include, but are not limited to, deaths and injuries to residents, damage to local buildings, utilities, crops, and so on. Due to the cost of reconstruction and the high rate of poverty amongst the affected population, these floods result in high financial aid payments to governments and relief agencies. These are therefore very dangerous and costly disaster events.

Most of the affected population in each locality would be recognized as being economically deprived (with a high proportion of poor affected people in the total affected population). In addition, most of the affected population is not uniformly male or female, with some counties having more females than males as flood victims and others having more males as flood victims. Of all the affected counties, only 6 counties had a sex ratio of 1 male to 1 female. This may be due to factors such as the gender ratio of the population in different areas and the structure of local industries.

There is no single trend or pattern in age proportions, which vary from county to county. In most counties, between 15 and 30 per cent of the affected population is under 18 years of

age. The highest percentage is over 40 per cent and the lowest is less than 10 per cent. Between 10 and 40 per cent of the affected population is over 60 years of age. The highest percentage is over 70 per cent and the lowest is over 10 per cent. Overall, people over the age of 60 are more likely to be affected than young people under the age of 18, which may be since older people are more likely to be affected due to mobility constraints.

5 Limitations and future plan

The above exploratory data analysis is based on existing flood event data and affected population data for the period 2020-2021, and based on the granularity of the data set to the county level, there are some phenomena that are difficult to explain with the current granularity of the study and the scope of the data set, for example, why there is a large disparity between men and women in some of the counties, and why there are high proportions of youth or elderly amongst the affected population in some of the counties. why some counties have a high proportion of adolescents or elderly people in the affected population.

Therefore, in the following study, I plan to collect other external data, such as the demographic data of each county in the United States from 2020 to 2021, the gender and age distribution of the residents, the industrial structure of different counties, and the existence of aging. In the future, I can go into specific counties and take a deeper look at what is happening in those specific counties.

6 Reference

[1] NOAA Dataset (*https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/*)

[2] Open FEMA Dataset: Disaster Declarations Summaries (*https://www.fema.gov/openfema-data-page/disaster-declarations-summaries-v2*)

[3] Open FEMA Dataset: FEMA Web Disaster Summaries (*https://www.fema.gov/openfema-data-page/fema-web-disaster-summaries-v1*)

[4] Census dataset (*https://learn.bu.edu/ultra/courses/_101818_1/cl/outline*)

[5] R for Data Science