Project 2

Disaster Counts And Loss: A Shiny Journey Through U.S. Disaster Data

Shoufei Meng, Tianyi Xia, Xiangjing Hu, Yang Yu

Columbia University

01 Motivation

02 Research Questions

03 Methodology

04 Findings

05 Future Extensions

Motivation and Data Introduction

Disasters strike the United States on a regular basis, and we're curious about exactly what kinds of disasters hit different states in the U.S. each year in a big way. Likewise, we're curious about what different disasters cost each state between years. That's why we chose the Disaster Information Dataset. In this dataset there are two types of datasets, V1 and V2, but V1 is not all the same. The Denied Declarations dataset lists all denied applications for major disaster declarations and emergency declarations. The Disaster Declarations dataset contains a list of FEMA declaration types and the types of assistance authorized. The Disaster Summaries dataset contains financial assistance values, including the number of approved applications, as well as individual, public assistance, and hazard mitigation grant amounts.

Motivation and Data Introduction(cont.)

Mission Assignment dataset is a work order issued by FEMA to another Federal agency directing completion of a specified task and citing funding, other managerial controls, and guidance. V2 is a summarized dataset describing all federally declared disasters. Because we started out wanting multiple variables and a long timeline to explore the problem we wanted to study. A lot of the information we needed to use was included in the disaster declaration dataset, and the declaration area dataset was also included. So we chose this as our V1.

Motivation and Data Introduction(cont.)

Because we started out wanting multiple variables as well as different variable types and a long timeline to explore the problem we wanted to study. A lot of the information we needed to use was included in the Disaster Declaration dataset, and the Declaration Area dataset was also included. So we chose that as our V1. In order to include losses encountered, we introduced a third dataset, the Public Assistance Applicants Program Deliveries dataset, to the current V1 and V2 datasets. But we only extracted the information we needed.

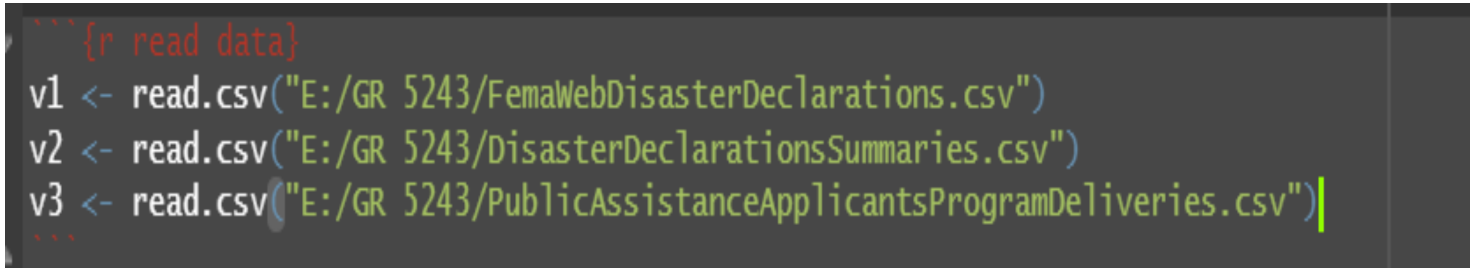
Research Questions

1. How do disaster frequencies vary among different states in the U.S. on an annual basis, and what patterns can be identified from these variations?

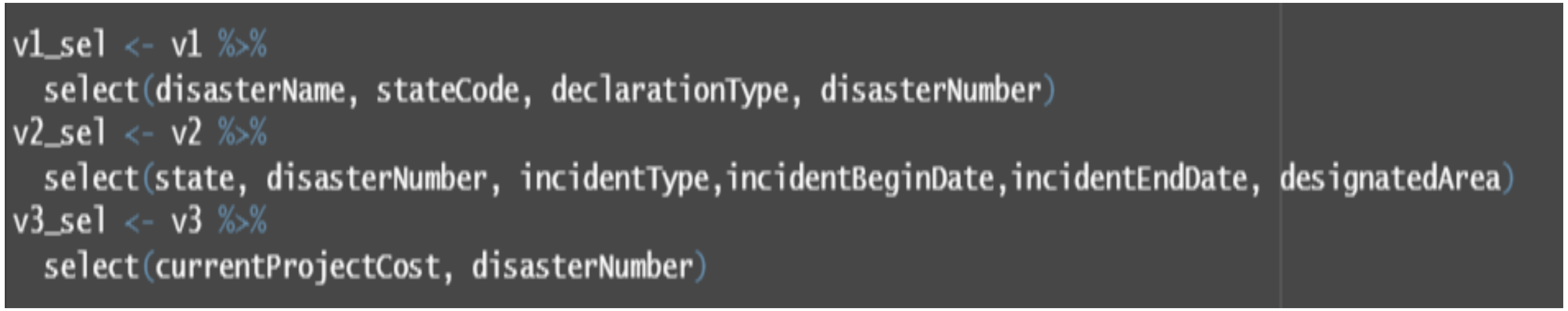
2. Which states have consistently experienced the highest number of different disasters per year, and how do these events correlate with losses in those states?

3. How has the frequency of different disasters evolved over time across the U.S., and can any short-term or long-term trends be discerned from the data?

Methodology (Data Preprocessing)



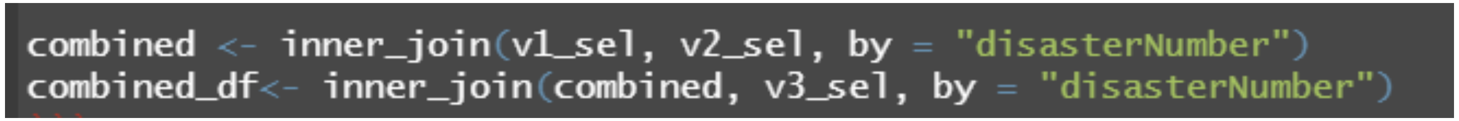
~~~
```{r read data}
v1 <- read.csv("E:/GR 5243/FemaWebDisasterDeclarations.csv")
v2 <- read.csv("E:/GR 5243/DisasterDeclarationsSummaries.csv")
v3 <- read.csv("E:/GR 5243/PublicAssistanceApplicantsProgramDeliveries.csv")
```
~~~



```
v1_sel <- v1 %>%
  select(disasterName, stateCode, declarationType, disasterNumber)
v2_sel <- v2 %>%
  select(state, disasterNumber, incidentType,incidentBeginDate,incidentEndDate, designatedArea)
v3_sel <- v3 %>%
  select(currentProjectCost, disasterNumber)
```

We first downloaded and loaded the packages we needed to use. Then we imported our dataset. Then we selected the variables we needed to use in each dataset.

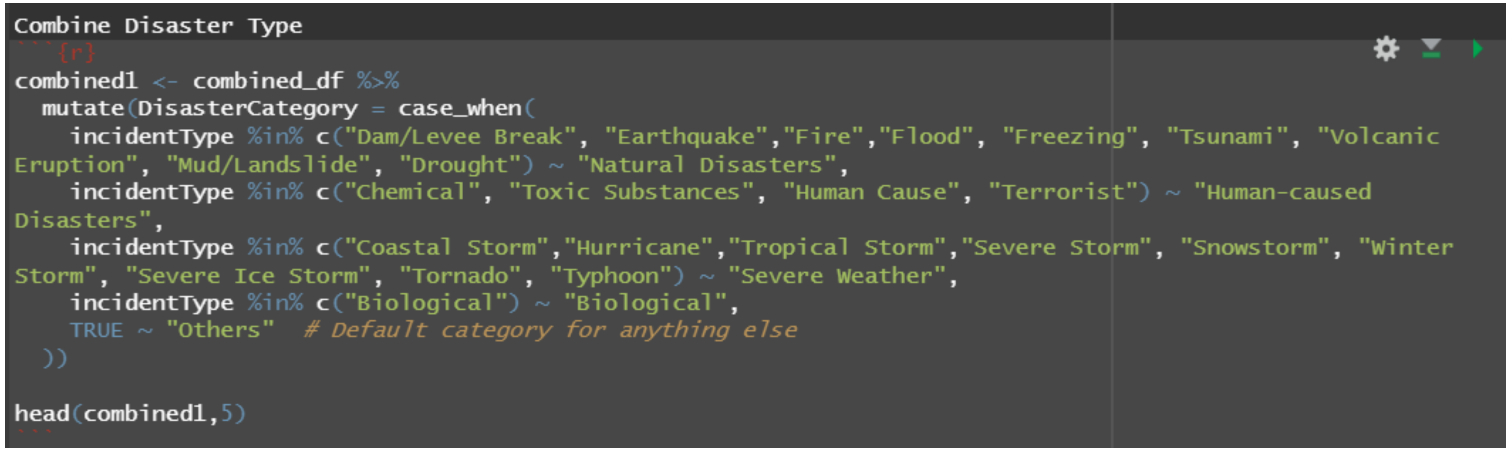
Methodology (Data Preprocessing)(cont.)



```
combined <- inner_join(v1_sel, v2_sel, by = "disasterNumber")
combined_df<- inner_join(combined, v3_sel, by = "disasterNumber")
```

Here we need to find the right foreign keys to merge our selected datasets into new data frames for subsequent use.

Methodology (Data Preprocessing)(cont.)

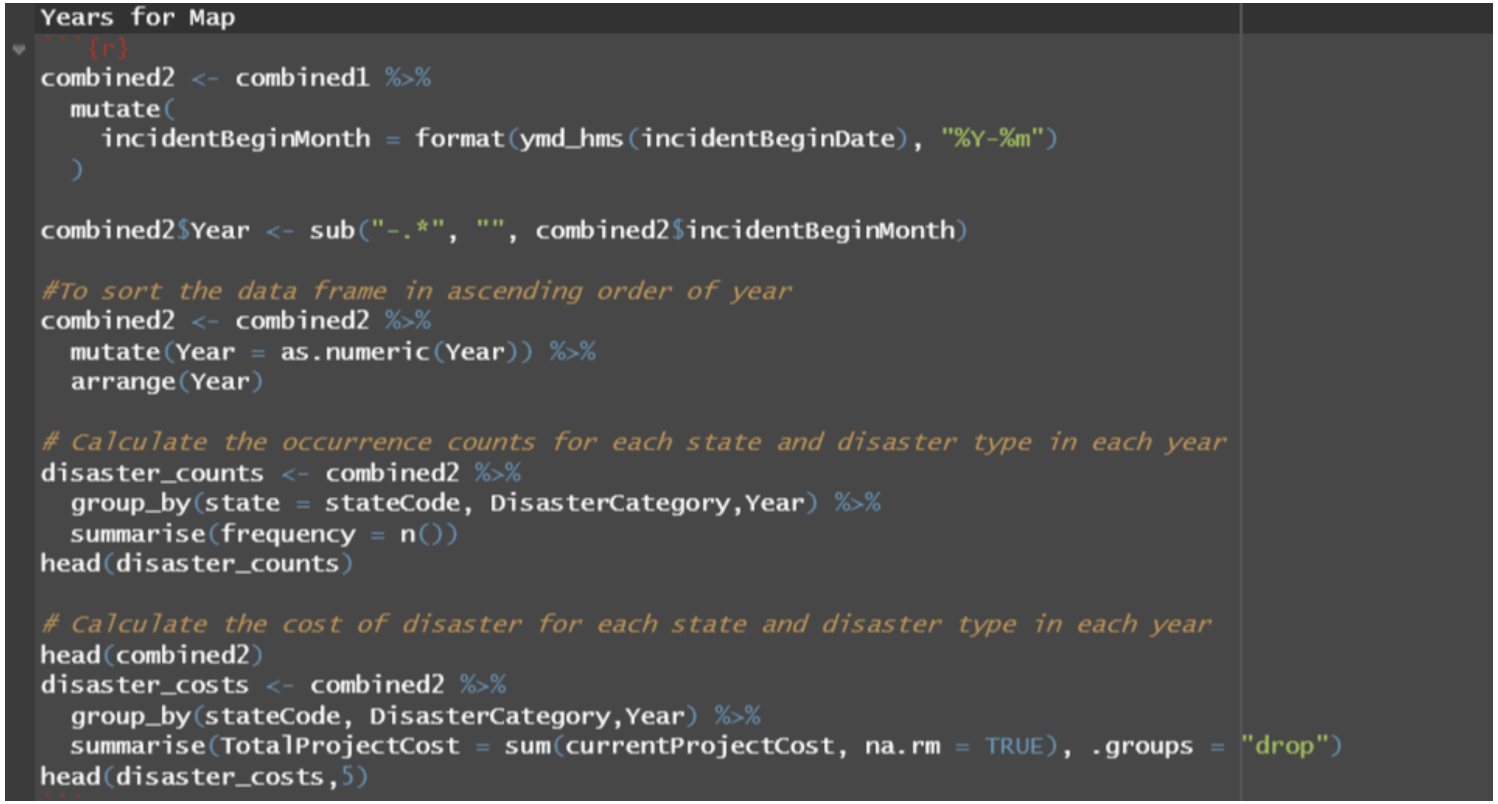


Combine Disaster Type

~~~
```{r}
combined1 <- combined_df %>%
  mutate(DisasterCategory = case_when(
    incidentType %in% c("Dam/Levee Break", "Earthquake","Fire","Flood", "Freezing", "Tsunami", "Volcanic
Eruption", "Mud/Landslide", "Drought") ~ "Natural Disasters",
    incidentType %in% c("Chemical", "Toxic Substances", "Human Cause", "Terrorist") ~ "Human-caused
Disasters",
    incidentType %in% c("Coastal Storm","Hurricane","Tropical Storm","Severe Storm", "Snowstorm", "Winter
Storm", "Severe Ice Storm", "Tornado", "Typhoon") ~ "Severe Weather",
    incidentType %in% c("Biological") ~ "Biological",
    TRUE ~ "Others"   # Default category for anything else
  ))

head(combined1,5)
```
~~~

Because there are many types of disasters, here we have divided all the disaster types into five major categories so that the disaster types are not too redundant in the shiny app.

Methodology (Data Pre For Map)(cont.)

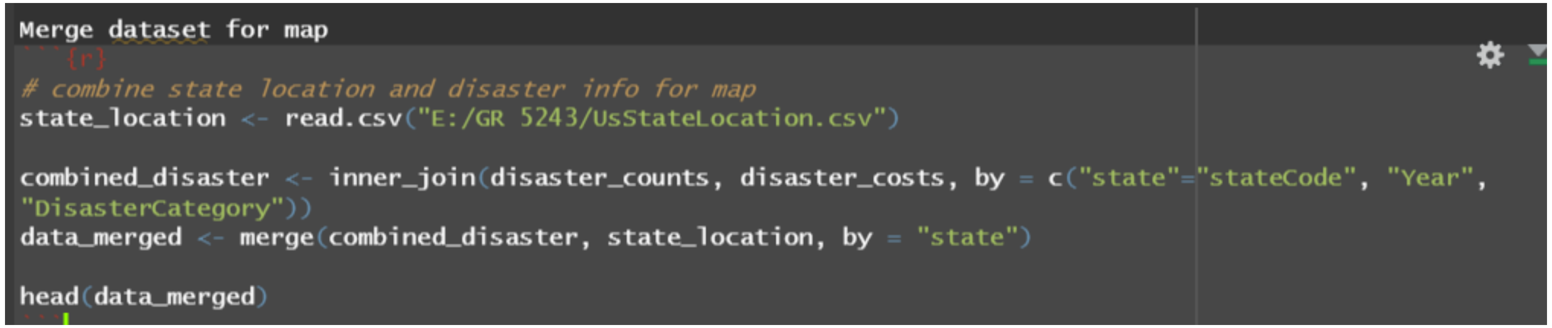


~~~
Years for Map
```{r}
combined2 <- combined1 %>%
  mutate(
    incidentBeginMonth = format(ymd_hms(incidentBeginDate), "%Y-%m")
  )

combined2$Year <- sub("-.*", "", combined2$incidentBeginMonth)

#To sort the data frame in ascending order of year
combined2 <- combined2 %>%
  mutate(Year = as.numeric(Year)) %>%
  arrange(Year)

# Calculate the occurrence counts for each state and disaster type in each year
disaster_counts <- combined2 %>%
  group_by(state = stateCode, DisasterCategory,Year) %>%
  summarise(frequency = n())
head(disaster_counts)

# Calculate the cost of disaster for each state and disaster type in each year
head(combined2)
disaster_costs <- combined2 %>%
  group_by(stateCode, DisasterCategory,Year) %>%
  summarise(TotalProjectCost = sum(currentProjectCost, na.rm = TRUE), .groups = "drop")
head(disaster_costs,5)
```
~~~

We changed the date format in the data frame to year and month, for subsequent model construction. We then calculated the number and cost of disaster events for each state.

Columbia University

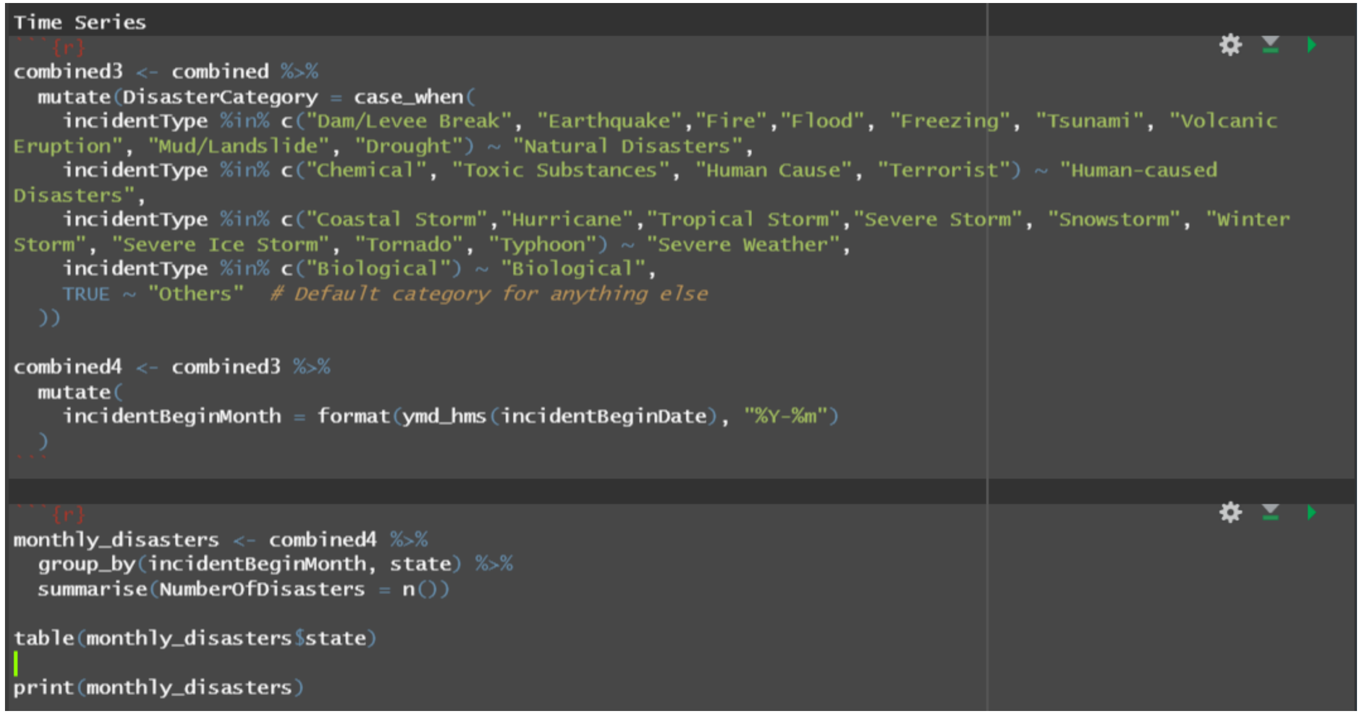
Methodology (Data Pre For Map)(cont.)



~~~
Merge dataset for map
```{r}
# combine state location and disaster info for map
state_location <- read.csv("E:/GR 5243/UsStateLocation.csv")

combined_disaster <- inner_join(disaster_counts, disaster_costs, by = c("state"="stateCode", "Year", 
"DisasterCategory"))
data_merged <- merge(combined_disaster, state_location, by = "state")

head(data_merged)
```
~~~

We map the number of disasters and the cost of disasters obtained against each state in the map.

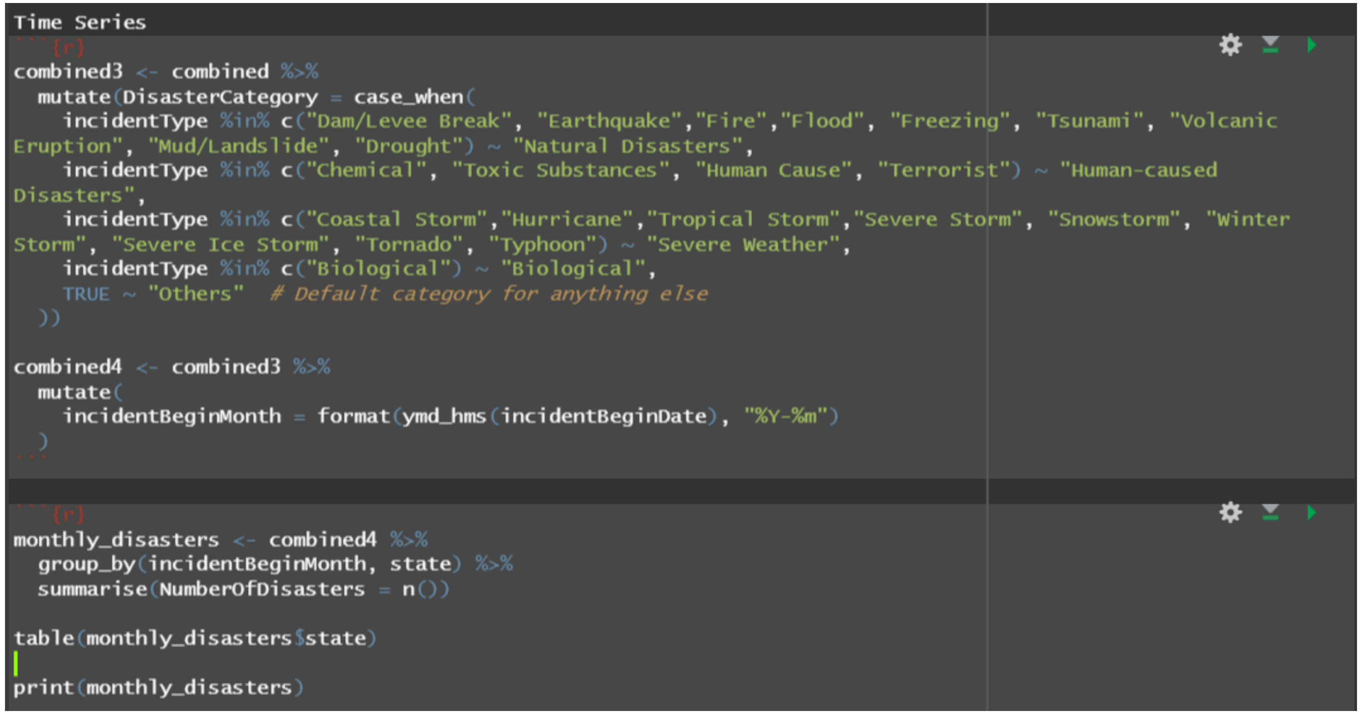
Methodology (Data Pre For Time Series)(cont.)



~~~
Time Series
```{r}
combined3 <- combined %>%
  mutate(DisasterCategory = case_when(
    incidentType %in% c("Dam/Levee Break", "Earthquake","Fire","Flood", "Freezing", "Tsunami", "Volcanic
Eruption", "Mud/Landslide", "Drought") ~ "Natural Disasters",
    incidentType %in% c("Chemical", "Toxic Substances", "Human Cause", "Terrorist") ~ "Human-caused
Disasters",
    incidentType %in% c("Coastal Storm","Hurricane","Tropical Storm","Severe Storm", "Snowstorm", "Winter
Storm", "Severe Ice Storm", "Tornado", "Typhoon") ~ "Severe Weather",
    incidentType %in% c("Biological") ~ "Biological",
    TRUE ~ "Others"  # Default category for anything else
  ))

combined4 <- combined3 %>%
  mutate(
    incidentBeginMonth = format(ymd_hms(incidentBeginDate), "%Y-%m")
  )
```

```{r}
monthly_disasters <- combined4 %>%
  group_by(incidentBeginMonth, state) %>%
  summarise(NumberOfDisasters = n())

table(monthly_disasters$state)

print(monthly_disasters)
~~~

We still represent the data in the form of years and months. In addition we based the problem we wanted to study the change in disaster frequency over time for each state. So we changed the data again to create a new data frame.

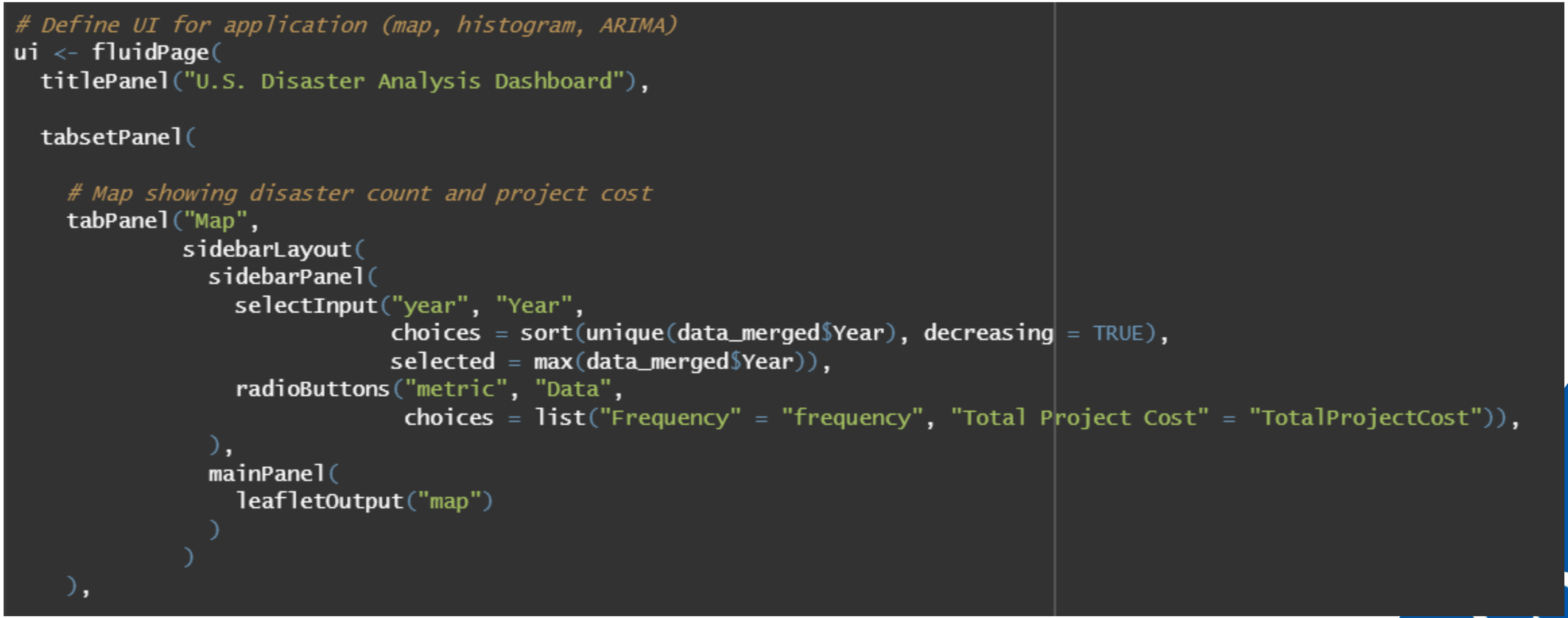
Methodology (Data Pre For Time Series)(cont.)



~~~
Time Series
```{r}
combined3 <- combined %>%
  mutate(DisasterCategory = case_when(
    incidentType %in% c("Dam/Levee Break", "Earthquake","Fire","Flood", "Freezing", "Tsunami", "Volcanic
Eruption", "Mud/Landslide", "Drought") ~ "Natural Disasters",
    incidentType %in% c("Chemical", "Toxic Substances", "Human Cause", "Terrorist") ~ "Human-caused
Disasters",
    incidentType %in% c("Coastal Storm","Hurricane","Tropical Storm","Severe Storm", "Snowstorm", "Winter
Storm", "Severe Ice Storm", "Tornado", "Typhoon") ~ "Severe Weather",
    incidentType %in% c("Biological") ~ "Biological",
    TRUE ~ "Others"  # Default category for anything else
  ))

combined4 <- combined3 %>%
  mutate(
    incidentBeginMonth = format(ymd_hms(incidentBeginDate), "%Y-%m")
  )
```

```{r}
monthly_disasters <- combined4 %>%
  group_by(incidentBeginMonth, state) %>%
  summarise(NumberOfDisasters = n())

table(monthly_disasters$state)

print(monthly_disasters)
~~~

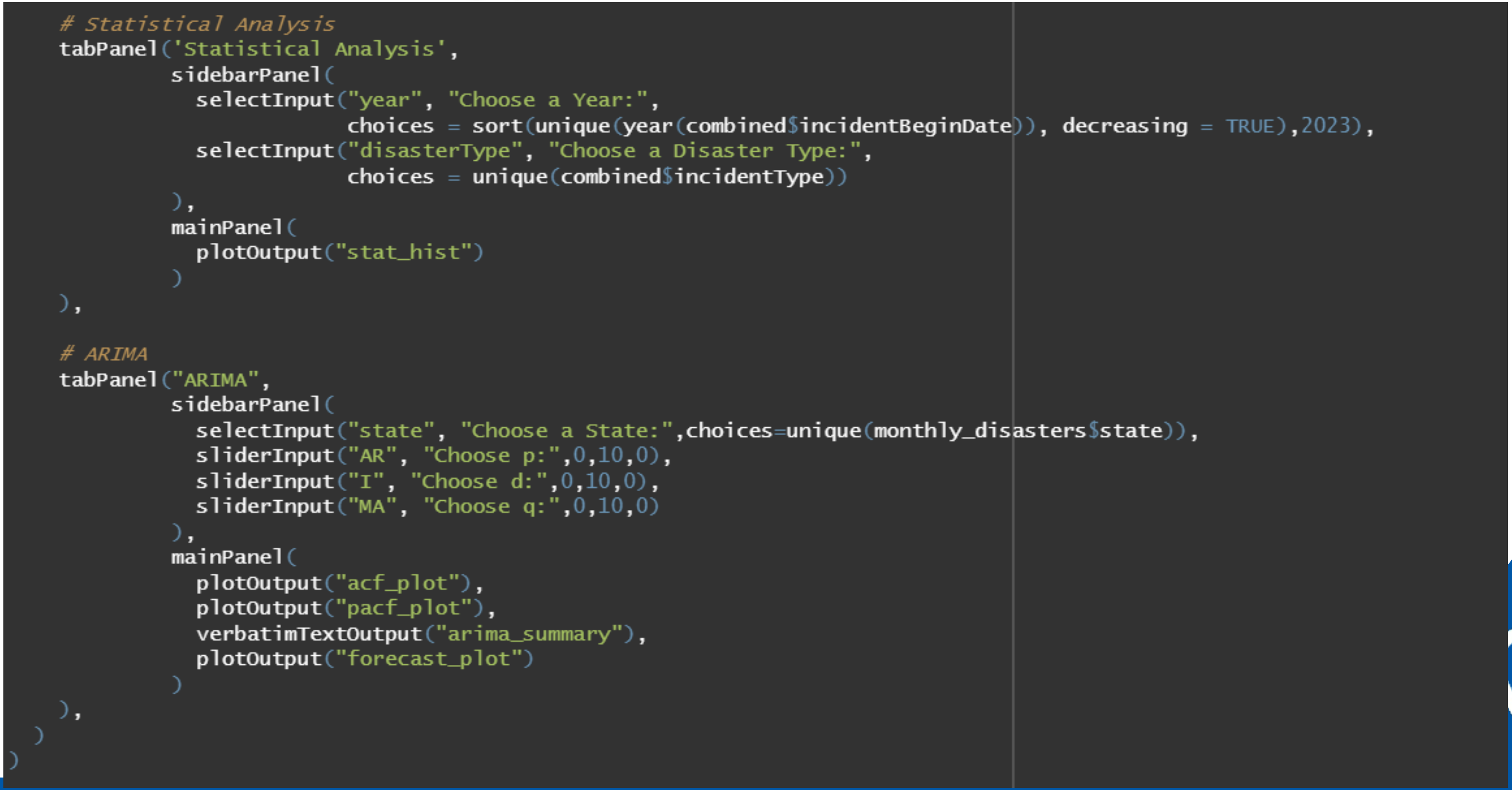
We still represent the data in the form of years and months. In addition we based the problem we wanted to study the change in disaster frequency over time for each state. So we changed the data again to create a new data frame.

Methodology (Shiny App)



```
# Define UI for application (map, histogram, ARIMA)
ui <- fluidPage(
  titlePanel("U.S. Disaster Analysis Dashboard"),

  tabsetPanel(

    # Map showing disaster count and project cost
    tabPanel("Map",
             sidebarLayout(
               sidebarPanel(
                 selectInput("year", "Year",
                             choices = sort(unique(data_merged$Year), decreasing = TRUE),
                             selected = max(data_merged$Year)),
                 radioButtons("metric", "Data",
                              choices = list("Frequency" = "frequency", "Total Project Cost" = "TotalProjectCost")),
               ),
               mainPanel(
                 leafletOutput("map")
               )
             )
    ),
```

Methodology (Shiny App)



```
    # Statistical Analysis
    tabPanel('Statistical Analysis',
             sidebarPanel(
               selectInput("year", "Choose a Year:",
                           choices = sort(unique(year(combined$incidentBeginDate)), decreasing = TRUE),2023),
               selectInput("disasterType", "Choose a Disaster Type:",
                           choices = unique(combined$incidentType))
             ),
             mainPanel(
               plotOutput("stat_hist")
             )
    ),

    # ARIMA
    tabPanel("ARIMA",
             sidebarPanel(
               selectInput("state", "Choose a State:",choices=unique(monthly_disasters$state)),
               sliderInput("AR", "Choose p:",0,10,0),
               sliderInput("I", "Choose d:",0,10,0),
               sliderInput("MA", "Choose q:",0,10,0)
             ),
             mainPanel(
               plotOutput("acf_plot"),
               plotOutput("pacf_plot"),
               verbatimTextOutput("arima_summary"),
               plotOutput("forecast_plot")
             )
    ),
  )
)
```

Columbia University

Findings (Question1)

Severe weather and natural disasters are the two disasters that will have the greatest impact on the United States through 2020. The Southeast is highly impacted by severe weather. The central as well as the upper central part of the country is highly affected by natural disasters. However, in 2020, due to the emergence of a new coronavirus, every state in the U.S. is significantly impacted by it.The Lower Middle is significantly impacted by severe weather in 2021. Notably, California is significantly affected by severe weather and natural disasters in 2022 and 2023.

Findings (Question2)

According to our data framework, the disasters faced by different states are largely positively correlated with the costs they face. The size of the circle indicates the magnitude of the cost. However, it also appears that some disasters have a high frequency but not a high cost. Situations also emerged where the frequency of disasters was high but the costs were very large.

Findings (Question2)

We also ranked the frequency of different disasters in descending order on a long time series line, selecting the states where different disasters occur more frequently each year.

Findings (Question3)

For each state we constructed an ARIMA model and plotted its ACF and PACF to help us determine the order of the model. At the bottom we plotted the predictions of disaster frequency in the short term.

Future Extensions

In our map table there is only data for 2016-2023, indicating that there is still too little data available to us to be very powerful and convincing. In the histogram, some data for disasters are missing, which is why the x-axis is shown at positive infinity. If we want to show more convincing maps as well as time series predictions, we may need more data to show our ideas.