# STATS_107_HalfwayProj 2025-11-02

Jessica Xia

2025-10-24

# Data Processing

```
rm(list = ls())
source("00_requirements.R")
```

```
## Loading required package: tidyverse
```

```
## Warning: package 'tidyverse' was built under R version 4.4.3
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
## Warning: package 'tibble' was built under R version 4.4.3
```

```
## Warning: package 'tidyr' was built under R version 4.4.3
```

```
## Warning: package 'purrr' was built under R version 4.4.3
```

```
## Warning: package 'dplyr' was built under R version 4.4.3
```

```
## Warning: package 'stringr' was built under R version 4.4.3
```

```
## Warning: package 'forcats' was built under R version 4.4.3
```

```
## Warning: package 'lubridate' was built under R version 4.4.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.1     ✔ stringr   1.5.2
## ✔ ggplot2   4.0.0     ✔ tibble    3.3.0
## ✔ lubridate 1.9.4     ✔ tidyr     1.3.1
## ✔ purrr     1.1.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
## Warning: package 'tidyverse' is in use and will not be installed
```

```
## Loading required package: data.table
```

```
## Warning: package 'data.table' was built under R version 4.4.3
```

```
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
## Warning: package 'data.table' is in use and will not be installed
```

First we will load the data:

```
#events in US from 1980 - 2024
events_data <- read.csv("data/events-US-1980-2024-Q4.csv")
events_data <- na.omit(events_data)
events_data <- events_data[-c(1, 2), ]

#insurance data
overall_insurance <- read.csv("data/all_insurance_costs.csv")
overall_insurance <- na.omit(overall_insurance)

#property insurance cost data
prop_insurance <- read.csv("data/property_insurance_costs.csv")
prop_insurance <- na.omit(prop_insurance)
```

Next we will fix the data types of our data

```
colnames(events_data) <- c("Name_Date", #chr
                           "Disaster_Type", #chr
                           "Begin_Date", #num -> chr (date format)
                           "End_Date", #num -> chr (date format)
                          "CPI_Adjusted_Cost_Millions", #num
                          "Unadjusted_Cost_Millions", #num
                          "Deaths") #int

events_data$CPI_Adjusted_Cost_Millions <- as.numeric(events_data$CPI_Adjusted_Cost_Millions)
events_data$Unadjusted_Cost_Millions <- as.numeric(events_data$Unadjusted_Cost_Millions)
events_data$Deaths <- as.integer(events_data$Deaths)


events_data$Begin_Date <- as.Date(as.character(events_data$Begin_Date),
                          format = "%Y%m%d")

events_data$End_Date <- as.Date(as.character(events_data$End_Date),
                        format = "%Y%m%d")

events_data$CPI_Adjusted_Cost_Millions <- as.numeric(events_data$CPI_Adjusted_Cost_Millions)

events_data$Unadjusted_Cost_Millions <- as.numeric(events_data$Unadjusted_Cost_Millions)

events_data$Deaths <- as.numeric(events_data$Deaths)

head(events_data)
```

```
##                                                             Name_Date
## 3                   Southern Severe Storms and Flooding (April 1980)
## 4                                      Hurricane Allen (August 1980)
## 5               Central/Eastern Drought/Heat Wave (Summer-Fall 1980)
## 6                                      Florida Freeze (January 1981)
## 7            Severe Storms, Flash Floods, Hail, Tornadoes (May 1981)
## 8 Midwest/Southeast/Northeast Winter Storm, Cold Wave (January 1982)
##      Disaster_Type Begin_Date   End_Date CPI_Adjusted_Cost_Millions
## 3         Flooding 1980-04-10 1980-04-17                     2749.4
## 4 Tropical Cyclone 1980-08-07 1980-08-11                     2236.2
## 5          Drought 1980-06-01 1980-11-30                    40681.2
## 6           Freeze 1981-01-12 1981-01-14                     2076.4
## 7     Severe Storm 1981-05-05 1981-05-10                     1409.1
## 8     Winter Storm 1982-01-08 1982-01-16                     2217.8
##   Unadjusted_Cost_Millions Deaths
## 3                    706.8      7
## 4                    590.0     13
## 5                  10020.0   1260
## 6                    572.0      0
## 7                    401.4     20
## 8                    662.0     85
```

```
dim(events_data)
```

```
## [1] 403   7
```

```
colnames(overall_insurance) <- c("PPI_Series_ID",
                              "Year",
                              "Month_Code",
                              "Time_Period",
                              "All_Insurance_Index")

head(overall_insurance)
```

```
##   PPI_Series_ID Year Month_Code Time_Period All_Insurance_Index
## 1        WPS411 2009        M06    2009 Jun               100.1
## 2        WPS411 2009        M07    2009 Jul               100.3
## 3        WPS411 2009        M08    2009 Aug               100.4
## 4        WPS411 2009        M09    2009 Sep               100.8
## 5        WPS411 2009        M10    2009 Oct               101.3
## 6        WPS411 2009        M11    2009 Nov               101.5
```

```
colnames(prop_insurance) <- c("Series_ID",
                  "Year",
                  "Month_Code",
                  "Time_Period",
                  "Property_Insurance_Index")

head(prop_insurance)
```

```
##     Series_ID Year Month_Code Time_Period Property_Insurance_Index
## 1 WPU41110401 2009        M03    2009 Mar                    100.0
## 2 WPU41110401 2009        M04    2009 Apr                    100.4
## 3 WPU41110401 2009        M05    2009 May                    100.3
## 4 WPU41110401 2009        M06    2009 Jun                    100.6
## 5 WPU41110401 2009        M07    2009 Jul                    100.9
## 6 WPU41110401 2009        M08    2009 Aug                    100.9
```

```
save.image("12_cleanedData.RData")
```