RMD_21_DataProcessing

2025-11-04

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
```

```
load("12_cleanedData.RData")
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
```

```
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
```

```
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
```

Our overall insurance data does not contain the months of March 2009 to May 2009, so we will remove them from the property insurance data so they can match 1 to 1

```
prop_insurance <- prop_insurance[-c(1:3),]
```

We can now merge the 2 datasets, since they match 1 to 1 in terms of date. We can also get rid of useless columns

```
all_insurance <- merge(prop_insurance, overall_insurance)
head(all_insurance)
```

```
##   Year Month_Code Time_Period   Series_ID Property_Insurance_Index
## 1 2009        M06    2009 Jun WPU41110401                    100.6
## 2 2009        M07    2009 Jul WPU41110401                    100.9
## 3 2009        M08    2009 Aug WPU41110401                    100.9
## 4 2009        M09    2009 Sep WPU41110401                    101.2
## 5 2009        M10    2009 Oct WPU41110401                    101.9
## 6 2009        M11    2009 Nov WPU41110401                    102.0
##   PPI_Series_ID All_Insurance_Index
## 1        WPS411               100.1
## 2        WPS411               100.3
## 3        WPS411               100.4
## 4        WPS411               100.8
## 5        WPS411               101.3
## 6        WPS411               101.5
```

```
#we can now get rid of all and prop insurance
rm(overall_insurance)
rm(prop_insurance)
```

```
all_insurance <- all_insurance[ -c(3,4,6)]
head(all_insurance)
```

```
##   Year Month_Code Property_Insurance_Index All_Insurance_Index
## 1 2009        M06                    100.6               100.1
## 2 2009        M07                    100.9               100.3
## 3 2009        M08                    100.9               100.4
## 4 2009        M09                    101.2               100.8
## 5 2009        M10                    101.9               101.3
## 6 2009        M11                    102.0               101.5
```

Finding the cost of each month

```
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
```

All the events have a begin date and an end date. For simplicity, we will be assuming the cost of one of the disasters is split evenly over the months in which it occured. To do so, since we don’t care about the specific days, just the months and years, first we will floor all the dates to make them easier to work with

```
events_data$Begin_Date <- floor_date(events_data$Begin_Date, "month")
events_data$End_Date <- floor_date(events_data$End_Date, "month")
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 1980-04-01 1980-04-01                     2749.4
## 4 Tropical Cyclone 1980-08-01 1980-08-01                     2236.2
## 5          Drought 1980-06-01 1980-11-01                    40681.2
## 6           Freeze 1981-01-01 1981-01-01                     2076.4
## 7     Severe Storm 1981-05-01 1981-05-01                     1409.1
## 8     Winter Storm 1982-01-01 1982-01-01                     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
```

Next, we will get the number of months between the begin and end dates. Add one (to account for the starting month), and that will be the duration of our event (in months). We can then find the cost per month

(I got the at period code off stackexchange)

```
events_data$Duration_Interval <- interval(events_data$Begin_Date, events_data$End_Date)

events_data$Duration_Months <- as.period(events_data$Duration_Interval
) %/% months(1) + 1
#events_data$Duration_Months

events_data$Adjusted_CPM_Millions <- events_data$CPI_Adjusted_Cost_Millions / events_data$Duration_Months

events_data$Unadjusted_CPM_Millions <- events_data$Unadjusted_Cost_Millions / events_data$Duration_Months

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 1980-04-01 1980-04-01                     2749.4
## 4 Tropical Cyclone 1980-08-01 1980-08-01                     2236.2
## 5          Drought 1980-06-01 1980-11-01                    40681.2
## 6           Freeze 1981-01-01 1981-01-01                     2076.4
## 7     Severe Storm 1981-05-01 1981-05-01                     1409.1
## 8     Winter Storm 1982-01-01 1982-01-01                     2217.8
##   Unadjusted_Cost_Millions Deaths              Duration_Interval
## 3                    706.8      7 1980-04-01 UTC--1980-04-01 UTC
## 4                    590.0     13 1980-08-01 UTC--1980-08-01 UTC
## 5                  10020.0   1260 1980-06-01 UTC--1980-11-01 UTC
## 6                    572.0      0 1981-01-01 UTC--1981-01-01 UTC
## 7                    401.4     20 1981-05-01 UTC--1981-05-01 UTC
## 8                    662.0     85 1982-01-01 UTC--1982-01-01 UTC
##   Duration_Months Adjusted_CPM_Millions Unadjusted_CPM_Millions
## 3               1                2749.4                   706.8
## 4               1                2236.2                   590.0
## 5               6                6780.2                  1670.0
## 6               1                2076.4                   572.0
## 7               1                1409.1                   401.4
## 8               1                2217.8                   662.0
```

We can use this data to view the relative cost of disasters - perhaps the longer lasting ones are less destructive per month or equally as so. We will do so in the preliminary visualization. However, now we have to get the cost of each disaster into the insurance data. First, we will remove the M from the month code so we have a year and month, then we can turn that into a date

```
all_insurance$Month <- substr(all_insurance$Month_Code, 2, 3)
all_insurance$Calc_Date <- as.Date(
  paste(all_insurance$Year, all_insurance$Month, "1", sep = "-")
)
head(all_insurance)
```

```
##   Year Month_Code Property_Insurance_Index All_Insurance_Index Month  Calc_Date
## 1 2009        M06                    100.6               100.1    06 2009-06-01
## 2 2009        M07                    100.9               100.3    07 2009-07-01
## 3 2009        M08                    100.9               100.4    08 2009-08-01
## 4 2009        M09                    101.2               100.8    09 2009-09-01
## 5 2009        M10                    101.9               101.3    10 2009-10-01
## 6 2009        M11                    102.0               101.5    11 2009-11-01
```

Now, we can compare the calculated date to the interval in the events_data in order to see if that month incurred the cost of the event.

```r
all_insurance$Disaster_Cost <- 0

A <- all_insurance[,c(6:7)] #get date
B <- events_data[,c(3,4,10)] #get date interval and disaster cost
#idk if this is neccessary but its the example on the rdocumentation website so ill go with what i know works https://www.rdocumentation.org/packages/data.table/versions/1.17.8/topics/foverlaps
A$start <- A$Calc_Date
A$end <- A$Calc_Date
B$start <- B$Begin_Date
B$end <- B$End_Date
B <- data.table(B)
A <- data.table(A)
setkey(B, start, end)
overlaps <- foverlaps(A, B, type = "any", which = TRUE)
overlaps <- data.frame(overlaps)
for (i in 1:nrow(overlaps)) {
  monthindex <- overlaps[i,1]
  costindex <- overlaps[i,2]
  all_insurance[monthindex,]$Disaster_Cost <- all_insurance[monthindex,]$Disaster_Cost + B[costindex,]$Adjusted_CPM_Millions
}
rm(A, B, i, costindex, monthindex)
head(all_insurance)
```

```
##   Year Month_Code Property_Insurance_Index All_Insurance_Index Month  Calc_Date
## 1 2009        M06                    100.6               100.1    06 2009-06-01
## 2 2009        M07                    100.9               100.3    07 2009-07-01
## 3 2009        M08                    100.9               100.4    08 2009-08-01
## 4 2009        M09                    101.2               100.8    09 2009-09-01
## 5 2009        M10                    101.9               101.3    10 2009-10-01
## 6 2009        M11                    102.0               101.5    11 2009-11-01
##   Disaster_Cost
## 1     2618.8333
## 2     2149.8333
## 3      679.8333
## 4     1991.6333
## 5      679.8333
## 6      679.8333
```

```r
#i testesd this with an inefficeint brute force alg and it matches up, although leaves N/A where the cost is 0, thats fine
```

Now we can clean up (we’re keeping overlaps in case its useful later)

```
#we can use the month date instead with month(date) and year(date)
all_insurance$Year <- NULL
all_insurance$Month <- NULL
all_insurance$Month_Code <- NULL
all_insurance <- all_insurance[, c("Calc_Date", "Disaster_Cost", "Property_Insurance_Index", "Al
l_Insurance_Index")]
colnames(all_insurance)[1] <- "Insurance_Month"

#if the sum was 0 the cost is n/a, turn that back into 0
#taken from tidyverse replace na reference
all_insurance$Disaster_Cost <- replace_na(all_insurance$Disaster_Cost, 0)

head(all_insurance)
```

```
##   Insurance_Month Disaster_Cost Property_Insurance_Index All_Insurance_Index
## 1      2009-06-01     2618.8333                    100.6               100.1
## 2      2009-07-01     2149.8333                    100.9               100.3
## 3      2009-08-01      679.8333                    100.9               100.4
## 4      2009-09-01     1991.6333                    101.2               100.8
## 5      2009-10-01      679.8333                    101.9               101.3
## 6      2009-11-01      679.8333                    102.0               101.5
```

```
events_data$Duration_Interval <- NULL
colnames(events_data)[3] <- "Begin_Month"
colnames(events_data)[4] <- "End_Month"
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_Month  End_Month CPI_Adjusted_Cost_Millions
## 3         Flooding  1980-04-01 1980-04-01                     2749.4
## 4 Tropical Cyclone  1980-08-01 1980-08-01                     2236.2
## 5          Drought  1980-06-01 1980-11-01                    40681.2
## 6           Freeze  1981-01-01 1981-01-01                     2076.4
## 7     Severe Storm  1981-05-01 1981-05-01                     1409.1
## 8     Winter Storm  1982-01-01 1982-01-01                     2217.8
##   Unadjusted_Cost_Millions Deaths Duration_Months Adjusted_CPM_Millions
## 3                    706.8      7               1                2749.4
## 4                    590.0     13               1                2236.2
## 5                  10020.0   1260               6                6780.2
## 6                    572.0      0               1                2076.4
## 7                    401.4     20               1                1409.1
## 8                    662.0     85               1                2217.8
##   Unadjusted_CPM_Millions
## 3                   706.8
## 4                   590.0
## 5                  1670.0
## 6                   572.0
## 7                   401.4
## 8                   662.0
```

```
save.image("13_processedData.RData")
```