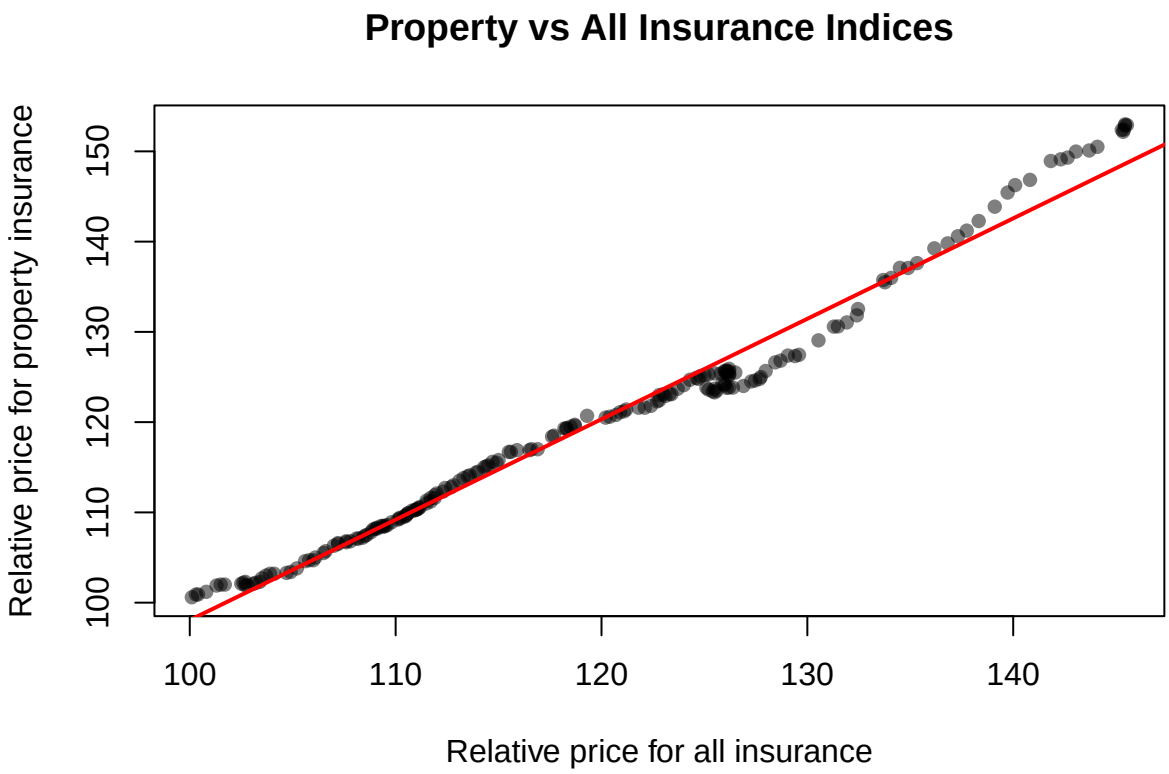
31_DataVisualization

```r
# Load your data files (you can comment out one if not needed)
rm(list = ls())
source("00_requirements.R")
load("22_processedData.RData")

# Convert key columns to correct types

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

all_insurance$All_Insurance_Index       <- as.numeric(all_insurance$All_Insurance_Index)
all_insurance$Property_Insurance_Index  <- as.numeric(all_insurance$Property_Insurance_Index)
```

Visualization

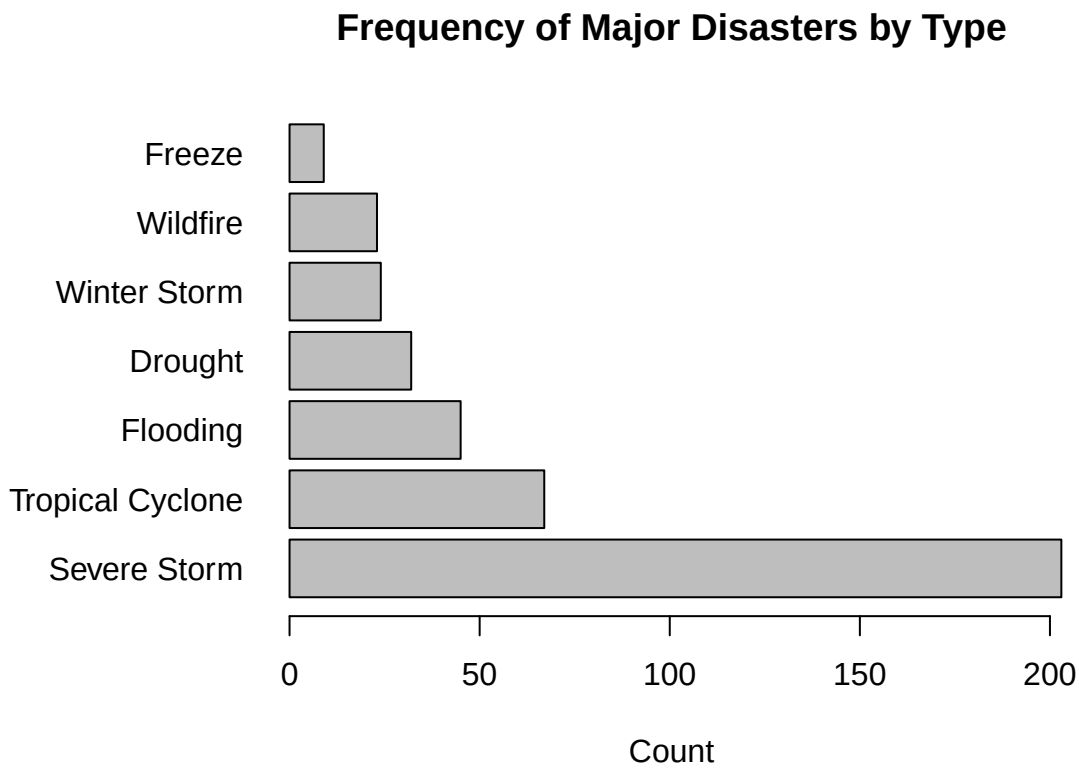
We can compare the property insurance prices to overall insurance prices to see if there's a significant difference between the two.

```r
plot(
  all_insurance$All_Insurance_Index,
  all_insurance$Property_Insurance_Index,
  xlab = "Relative price for all insurance",
  ylab = "Relative price for property insurance",
  main = "Property vs All Insurance Indices",
  pch = 16, col = rgb(0,0,0,0.5)
)
abline(lm(Property_Insurance_Index ~ All_Insurance_Index, data = all_insurance),
       col = "red", lwd = 2)
```



This scatterplot compares overall insurance prices with property insurance prices. It helps us see whether property insurance behaves differently or just follows the same pattern as the broader market. The points form a clear upward trend, which means both usually rise and fall together. This is expected since damages recorded in our data result in more damage than just property. Because the two move so closely, we might later focus on the difference between them (the "spread") to find subtle property-specific changes.

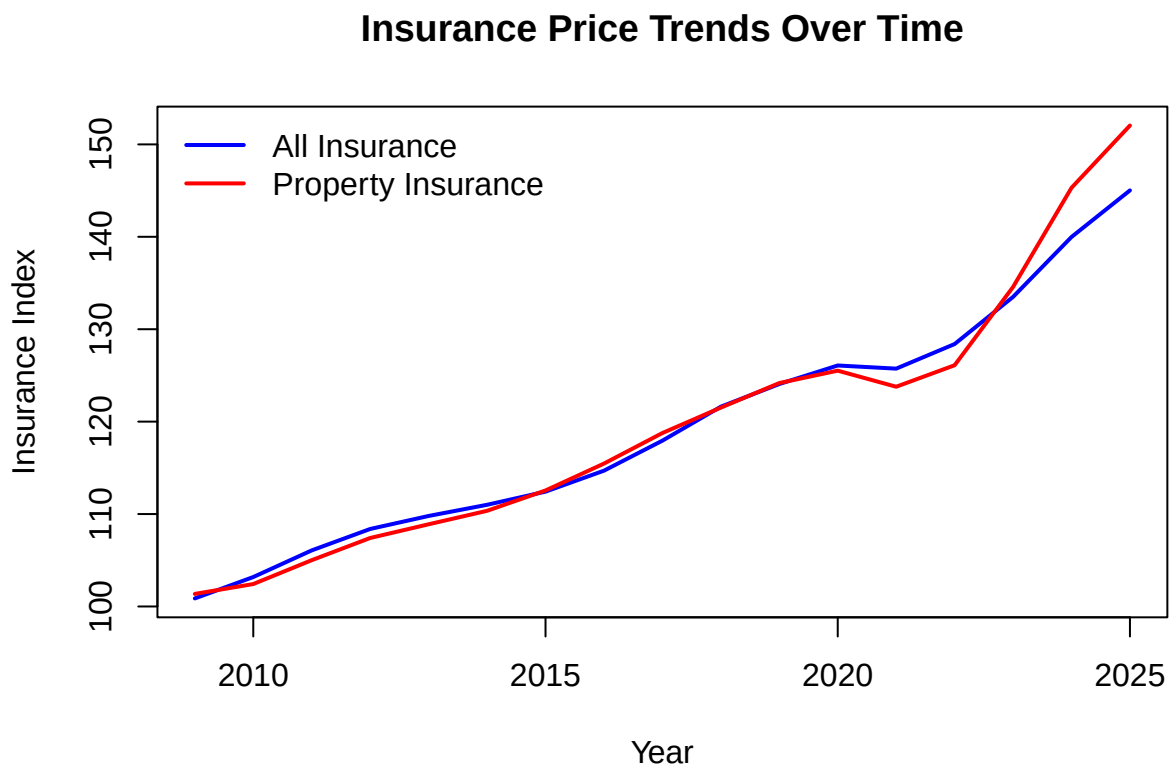
```
tbl <- sort(table(events_data$Disaster_Type), decreasing = TRUE)
par(mar = c(5,10,4,2) + 0.1)
barplot(tbl, horiz = TRUE, las = 1,
        main = "Frequency of Major Disasters by Type",
        xlab = "Count")
```



The bar chart shows which natural disaster happens the most within out data set. Seeing that Severe Storms have been recorded the most by a large margin we can no longer generalize that the most total money spent on Stors means that they are the most expensive on a one to one ratio. However severe storms have the most overall impact on insuarance.

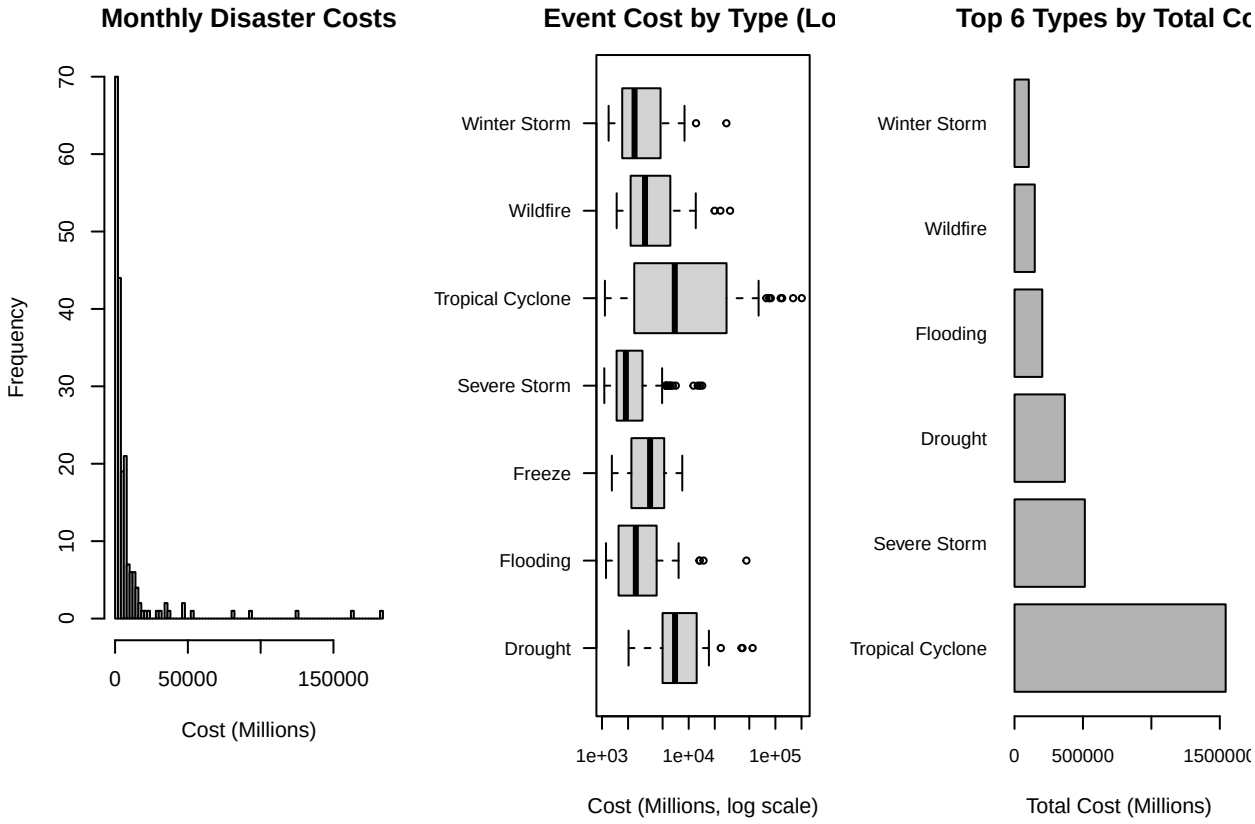
```
## Time Series Analysis

# 1. Extract the year from the 'Insurance_Month' Date column
# We create a new column 'Calc_Year' to aggregate by.
all_insurance$Calc_Year <- format(all_insurance$Insurance_Month, "%Y")

# 2. Annual mean indices from monthly insurance table
# We now aggregate using the newly created 'Calc_Year'
annual_all  <- aggregate(all_insurance$All_Insurance_Index ~ all_insurance$Calc_Year,
                         FUN = mean, na.rm = TRUE)
annual_prop <- aggregate(all_insurance$Property_Insurance_Index ~ all_insurance$Calc_Year,
                         FUN = mean, na.rm = TRUE)

# 3. Rename the columns (The grouping column is now the 'Year')
names(annual_all)  <- c("Year","AllIdx")
names(annual_prop) <- c("Year","PropIdx")

# 4. Align on common years
ii <- merge(annual_all, annual_prop, by="Year")
ii$Year <- as.integer(as.character(ii$Year))
# convert to numeric years
```

```r
# 5. Plot the time series
rng <- range(c(ii$AllIdx, ii$PropIdx), na.rm = TRUE)
plot(ii$Year, ii$AllIdx,
     type = "l",
     col = "blue",
     lwd = 2,
     xlab = "Year",
     ylab = "Insurance Index",
     main = "Insurance Price Trends Over Time",
     ylim = rng)
lines(ii$Year, ii$PropIdx,
      col = "red",
      lwd = 2)
legend("topleft",
       legend = c("All Insurance","Property Insurance"),
       col = c("blue","red"),
       lwd = 2,
       bty = "n")
```



This line graph tracks how both the overall and property insurance indices change year by year. It's useful for spotting trends and any time periods when property insurance rises faster than the general insurance market.The red (property) line pulls above the blue (overall) line in more recent years indicating that there may be more natural disaster and from what we learned in the last graph more severe storms more recently affecting property.

```r
# 1) Histogram of monthly disaster costs
# 2) Boxplot of event cost by type (log-scale)
# 3) Top 6 disaster types by total CPI-adjusted cost
# 4. Align on common years

op <- par(no.readonly = TRUE); on.exit(par(op))
par(mfrow = c(1, 3), mar = c(8, 4, 3, 1))  # set layout BEFORE plotting

## (1) Histogram (monthly disaster costs)
hist(all_insurance$Disaster_Cost, breaks = "FD",
     main = "Monthly Disaster Costs",
     xlab = "Cost (Millions)")

## (2) Boxplot by disaster type (horizontal, log-scale)
par(mar = c(5, 7, 3, 1), cex.axis = 0.9)  # extra left room for labels
boxplot(CPI_Adjusted_Cost_Millions ~ Disaster_Type, data = events_data,
        horizontal = TRUE, log = "x", las = 1,
        xlab = "Cost (Millions, log scale)",
        ylab =  "",
        main = "Event Cost by Type (Log)")

## (3) Top 6 types by total CPI-adjusted cost  (rotation fix: horizontal bars)
tot_by_type <- tapply(events_data$CPI_Adjusted_Cost_Millions,
                      events_data$Disaster_Type, sum, na.rm = TRUE)
tot_by_type <- sort(tot_by_type, decreasing = TRUE)
top6 <- head(tot_by_type, 6)
barplot(top6,
        horiz = TRUE, las = 1, xlab = "Total Cost (Millions)",
        main = "Top 6 Types by Total Cost", col = "gray70")
```



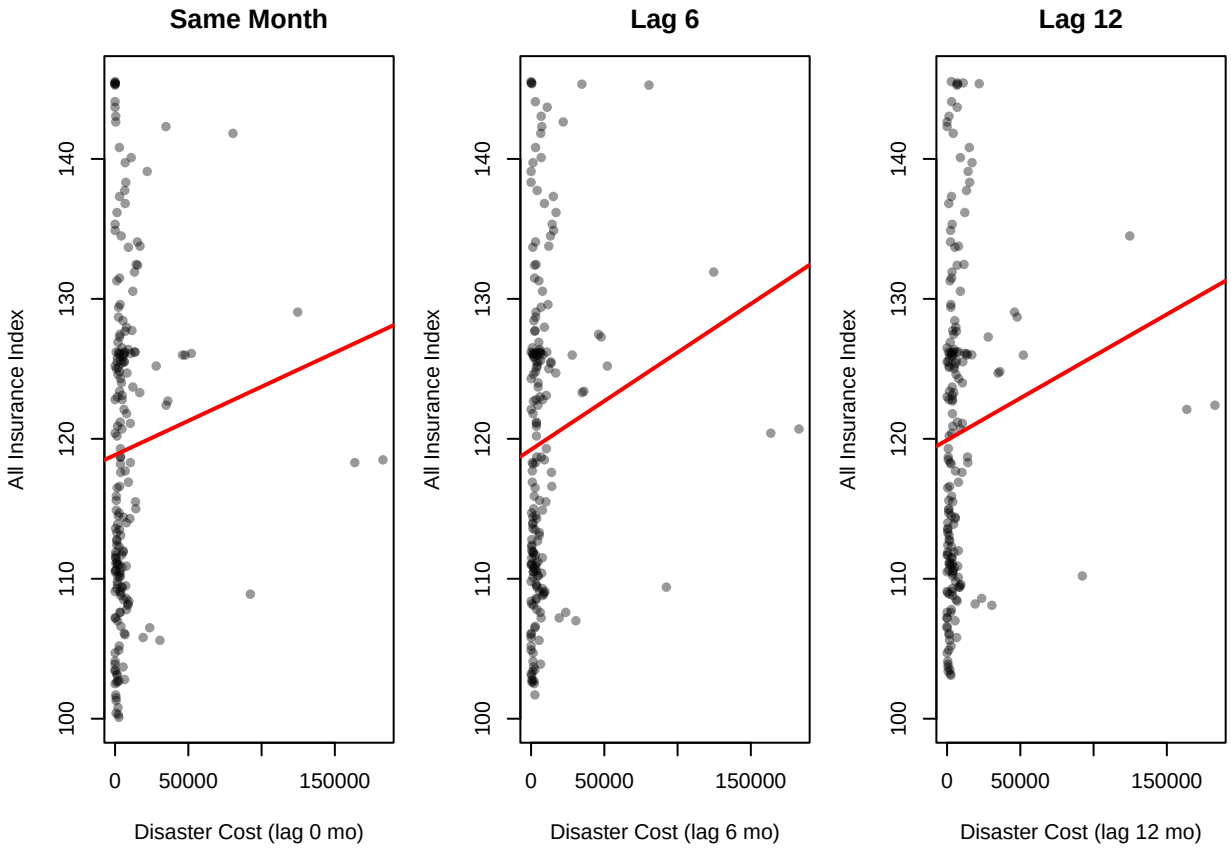
Monthly Disaster Costs (Histogram) #1 This histogram shows how disaster costs are distributed each month. Most months have relatively small costs, but a few extreme months stand out with huge spikes. This is predicted since disasters happen at random times. This pattern tells us the data are very right-skewed, meaning a few big events dominate the rest. That's why we use log scales or medians later instead of simple averages, since they can handle big outliers better.

Event Cost by Type (Boxplot) #2 The boxplot compares how expensive different disaster types are. Each box shows the spread and median cost, using a log scale to handle large differences. You can quickly see which types—like hurricanes or severe storms—tend to cause higher damages. This helps decide which disaster categories might have the strongest connection to insurance pricing later on. In a practictical sense it tells us what properties not to invest in and insurances the same.

Top 6 Types by Total CPI-Adjusted Cost #3 This bar chart focuses on the total combined cost of each disaster type. It highlights which ones have the biggest overall financial impact, not just which are common. A few types, such as tropical cyclones and severe storms, make up most of the total losses. These will probably play the biggest role when we study how disasters affect insurance rates since the total costs are so grand. In the grand scheme of disasters there are less overall cyclones so that is important to keep in mind.

```
# Check whether prior disaster costs relate to current insurance index
# Uses 0, 6, and 12 month lags

op <- par(no.readonly = TRUE); on.exit(par(op))
par(mfrow = c(1, 3), mar = c(4,4,3,1))

ord <- seq_len(nrow(all_insurance))
dc  <- all_insurance$Disaster_Cost[ord]
idx <- all_insurance$All_Insurance_Index[ord]
```

```r
lagK <- function(x, k) c(rep(NA, k), head(x, -k))

lags <- list(
  "Same Month" = 0L,
  "Lag 6"      = 6L,
  "Lag 12"     = 12L
)

i <- 1
for (nm in names(lags)) {
  k  <- lags[[nm]]
  lx <- if (k == 0L) dc else lagK(dc, k)
  plot(lx, idx,
       xlab = paste0("Disaster Cost (lag ", k, " mo)"),
       ylab = "All Insurance Index",
       main = nm,
       pch = 16, col = rgb(0,0,0,0.4))
  ok <- is.finite(lx) & is.finite(idx)
  if (sum(ok) > 2) abline(lm(idx[ok] ~ lx[ok]), col = "red", lwd = 2)
  i <- i + 1
}
```



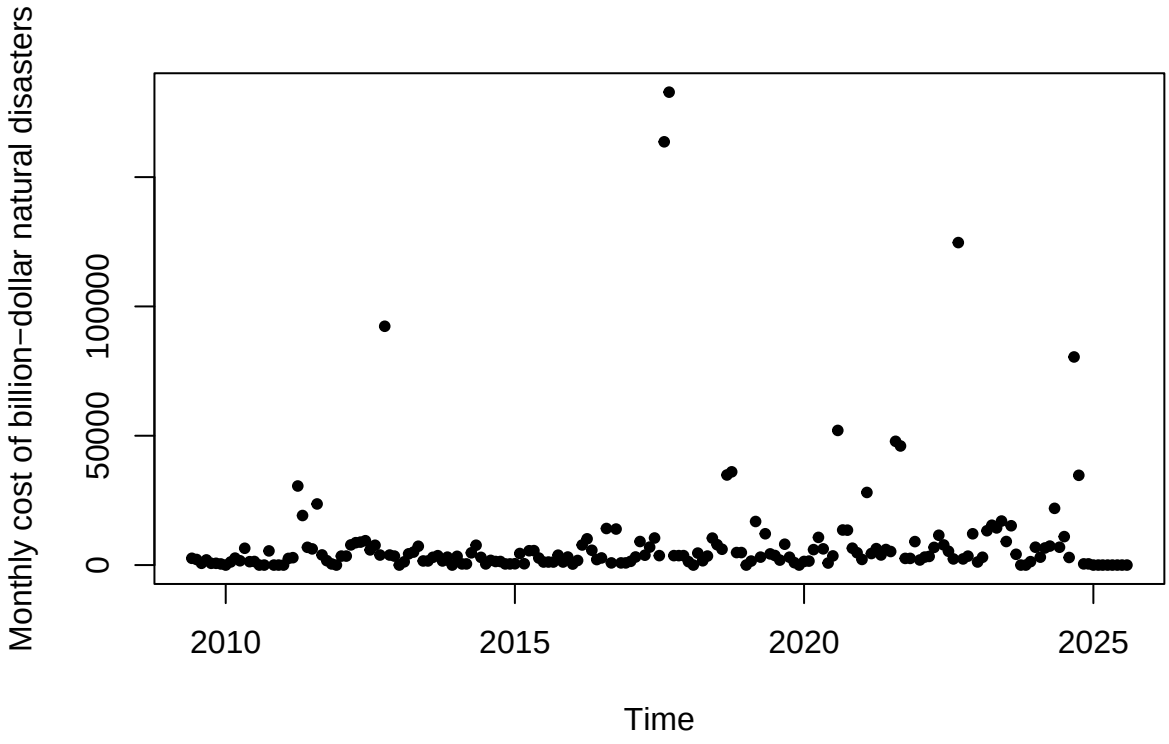
Same-Month Disaster Cost vs Insurance Index #1 This scatterplot looks at whether insurance prices go up immediately during months with expensive disasters. The points don't show much of a pattern, which suggests that prices don't react right away. It makes sense—insurers usually adjust rates after assessing the

damage, not in the same month. Also those affected need time to get to safety. It would be unfair to increase insurance rates directly after a severe tragety.

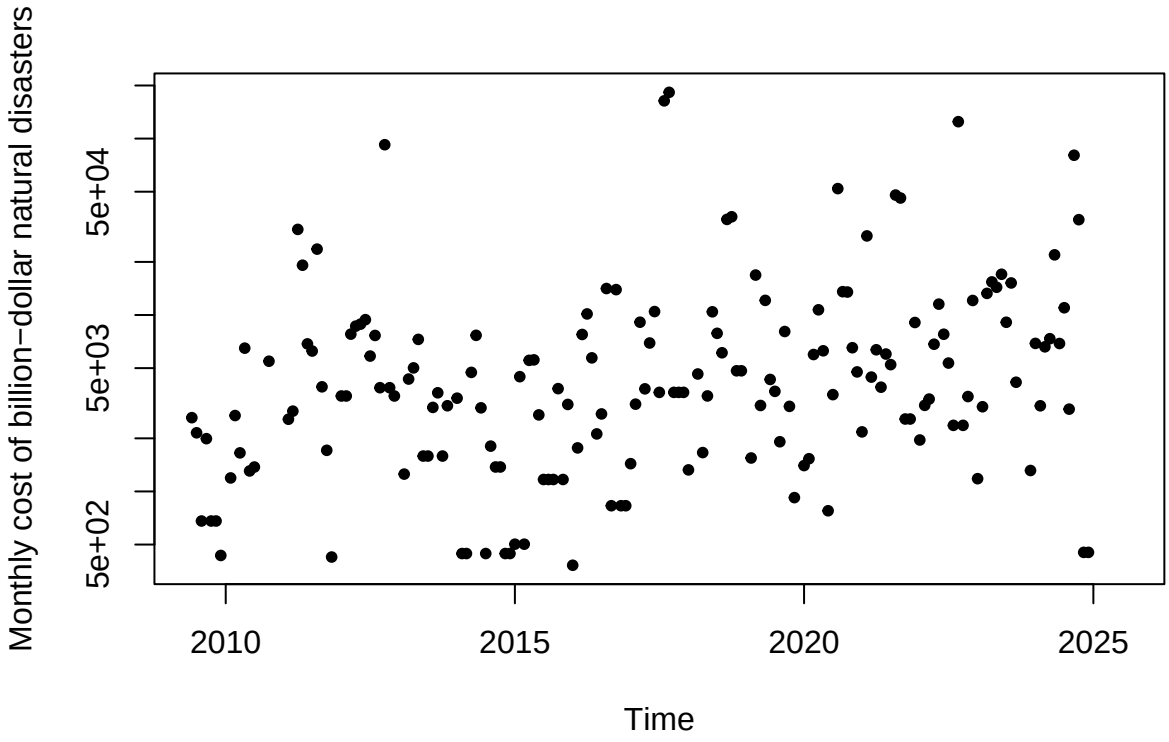
Lag 6 Months (Disaster Cost vs Insurance Index) #2 Here we shift the disaster costs back by six months to see if there's a delayed effect. The pattern starts to look a bit stronger in correlation, with higher disaster costs linked to slightly higher insurance prices half a year later. This hints that the market might take several months to reflect the impact of big losses. It will likely take even longer to recover from them.

Lag 12 Months (Disaster Cost vs Insurance Index) #3 Finally, we check a one-year lag. The relationship either weakens or stays similar to the six-month one. This helps us understand how long it takes for disasters to influence pricing trends. Whether the response fades after a year or continues. We'll use this to choose the right lag period in our later models. Looking at this graph disasters seem to have a more short term impact on insurance prices overall.

```
#these are our graphs of disaster costs over time, first is linear scale second is log scale - not real

plot(Disaster_Cost ~ Insurance_Month, data = all_insurance, pch = 20, ylab = "Monthly cost of billion-d
```




```
plot(Disaster_Cost ~ Insurance_Month, data = all_insurance, pch = 20, log = "y", ylab = "Monthly cost o
```



running linear regression on these two graphs:

```
disaster_time <- lm(Disaster_Cost ~ Insurance_Month, data = all_insurance)
summary(disaster_time)
```

```
##
## Call:
## lm(formula = Disaster_Cost ~ Insurance_Month, data = all_insurance)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13038  -7045  -4936  -1712 173785
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.491e+04  1.605e+04  -0.929    0.354
## Insurance_Month  1.377e+00  9.208e-01   1.495    0.137
##
## Residual standard error: 22030 on 193 degrees of freedom
## Multiple R-squared:  0.01145,    Adjusted R-squared:  0.006324
## F-statistic: 2.235 on 1 and 193 DF,  p-value: 0.1366
```

```
all_insurance$ModifiedDisaster <- all_insurance$Disaster_Cost + 1
log_disaster_time <- lm(log(ModifiedDisaster) ~ Insurance_Month, data = all_insurance)
summary(log_disaster_time)
```

```
##
## Call:
## lm(formula = log(ModifiedDisaster) ~ Insurance_Month, data = all_insurance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4894 -0.1936  0.7646  1.5059  4.7357
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.726e+00  2.103e+00   3.199  0.00161 **
## Insurance_Month 3.761e-05  1.206e-04   0.312  0.75554
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.886 on 193 degrees of freedom
## Multiple R-squared:  0.0005034,  Adjusted R-squared:  -0.004675
## F-statistic: 0.09721 on 1 and 193 DF,  p-value: 0.7555
```

above shows that disaster costs are not correlated with time, what about disaster costs leading to increase in property insurance?

```
disaster_insurance <- lm(Property_Insurance_Index  ~ Disaster_Cost, data = all_insurance)
summary(disaster_insurance)
```

```
##
## Call:
## lm(formula = Property_Insurance_Index ~ Disaster_Cost, data = all_insurance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.607 -10.405  -0.867   5.820  33.918
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.191e+02  1.026e+00 116.003   <2e-16 ***
## Disaster_Cost 5.028e-05  4.313e-05   1.166    0.245
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.28 on 193 degrees of freedom
## Multiple R-squared:  0.006994,   Adjusted R-squared:  0.001849
## F-statistic: 1.359 on 1 and 193 DF,  p-value: 0.2451
```

```
log_disaster_insurance <- lm(Property_Insurance_Index ~ log(ModifiedDisaster) , data = all_insurance)
summary(log_disaster_insurance)
```

```
##
## Call:
## lm(formula = Property_Insurance_Index ~ log(ModifiedDisaster),
##     data = all_insurance)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.9703 -10.4236  -0.0703   6.3369  31.2880
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           123.0703     2.6161  47.043   <2e-16 ***
## log(ModifiedDisaster)  -0.4803     0.3304  -1.454    0.148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.25 on 193 degrees of freedom
## Multiple R-squared:  0.01083,    Adjusted R-squared:  0.005704
## F-statistic: 2.113 on 1 and 193 DF,  p-value: 0.1477
```

We checked both linear disaster cost as well a log(disaster cost), but neither had any correlation. We also check overall insurance prices, although I personally doubt it will be different

```
disaster_all <- lm(All_Insurance_Index ~ Disaster_Cost, data = all_insurance)
summary(disaster_all)
```

```
##
## Call:
## lm(formula = All_Insurance_Index ~ Disaster_Cost, data = all_insurance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.875  -9.292  -1.433   7.008  26.679
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.188e+02  9.134e-01  130.12   <2e-16 ***
## Disaster_Cost 4.875e-05  3.838e-05    1.27    0.206
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.81 on 193 degrees of freedom
## Multiple R-squared:  0.008291,   Adjusted R-squared:  0.003153
## F-statistic: 1.614 on 1 and 193 DF,  p-value: 0.2055
```

```
log_disaster_all <- lm(All_Insurance_Index ~ log(ModifiedDisaster), data = all_insurance)
summary(log_disaster_all)
```

```
##
## Call:
## lm(formula = All_Insurance_Index ~ log(ModifiedDisaster), data = all_insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.0849  -9.4599  -0.6147   7.4439  24.5113
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           121.0147     2.3383  51.754   <2e-16 ***
## log(ModifiedDisaster)  -0.2345     0.2953  -0.794    0.428
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.84 on 193 degrees of freedom
## Multiple R-squared:  0.003256,   Adjusted R-squared:  -0.001908
## F-statistic: 0.6305 on 1 and 193 DF,  p-value: 0.4281
```

We have determined there is no linear correlation beteween insurance prices and the scale of natural disasters, but what if insruance prices just keep going up with time?

```
insurance_time <- lm(Property_Insurance_Index ~ Insurance_Month, data = all_insurance)
summary(insurance_time)
```

```
##
## Call:
## lm(formula = Property_Insurance_Index ~ Insurance_Month, data = all_insurance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6342 -1.4355 -0.5002  1.5542 12.2332
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -8.2913724  2.9592976  -2.802   0.0056 **
## Insurance_Month  0.0073676  0.0001698  43.402   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.061 on 193 degrees of freedom
## Multiple R-squared:  0.9071, Adjusted R-squared:  0.9066
## F-statistic:  1884 on 1 and 193 DF,  p-value: < 2.2e-16
```

this does seem to be the case. We can also check to see if there is a delay in insurance prices increasing as a cuase of large natural disasters - we can check 1 month 6 month and 12 month delays

```
delayed_damage <- data.frame(
  Property_Insurance = all_insurance$Property_Insurance_Index,
  Monthly_Damage = all_insurance$Disaster_Cost
)

diff_one_month <- c()
for (num in 1:(nrow(delayed_damage))) {
  diff_one_month <- append(diff_one_month, (delayed_damage$Property_Insurance[num+1]/ delayed_damage$Pr
}

delayed_damage$diff_1_month <- diff_one_month

diff_6_month <- c()
for (num in 1:(nrow(delayed_damage))) {
  diff_6_month <- append(diff_6_month, (delayed_damage$Property_Insurance[num+6]/ delayed_damage$Proper
```

```
}
delayed_damage$diff_6_month <- diff_6_month

diff_12_month <- c()

for (num in 1:(nrow(delayed_damage))) {
  diff_12_month <- append(diff_12_month, (delayed_damage$Property_Insurance[num+12]/ delayed_damage$Pro
}

delayed_damage$diff_1_year <- diff_12_month
delayed_damage$ModifiedDamage <- all_insurance$ModifiedDisaster
```

linear regression run below

```
#see if the price change from 1 year ago to now is correlated with the damage from a month
#plot(diff_1_year ~ Monthly_Damage, data = delayed_damage, pch = 20)
month_delay <- lm(diff_1_month ~ Monthly_Damage, data =delayed_damage)
summary(month_delay)
```

```
##
## Call:
## lm(formula = diff_1_month ~ Monthly_Damage, data = delayed_damage)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0157455 -0.0017979 -0.0006245  0.0007300  0.0202151
##
## Coefficients:
##                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     1.002e+00  2.746e-04 3649.551   <2e-16 ***
## Monthly_Damage -7.656e-09  1.151e-08   -0.665    0.507
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003541 on 192 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.0023, Adjusted R-squared:  -0.002897
## F-statistic: 0.4425 on 1 and 192 DF,  p-value: 0.5067
```

```
six_month_delay <- lm(diff_6_month ~ Monthly_Damage, data =delayed_damage)
summary(six_month_delay)
```

```
##
## Call:
## lm(formula = diff_6_month ~ Monthly_Damage, data = delayed_damage)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.030723 -0.006750 -0.002390  0.003016  0.041456
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

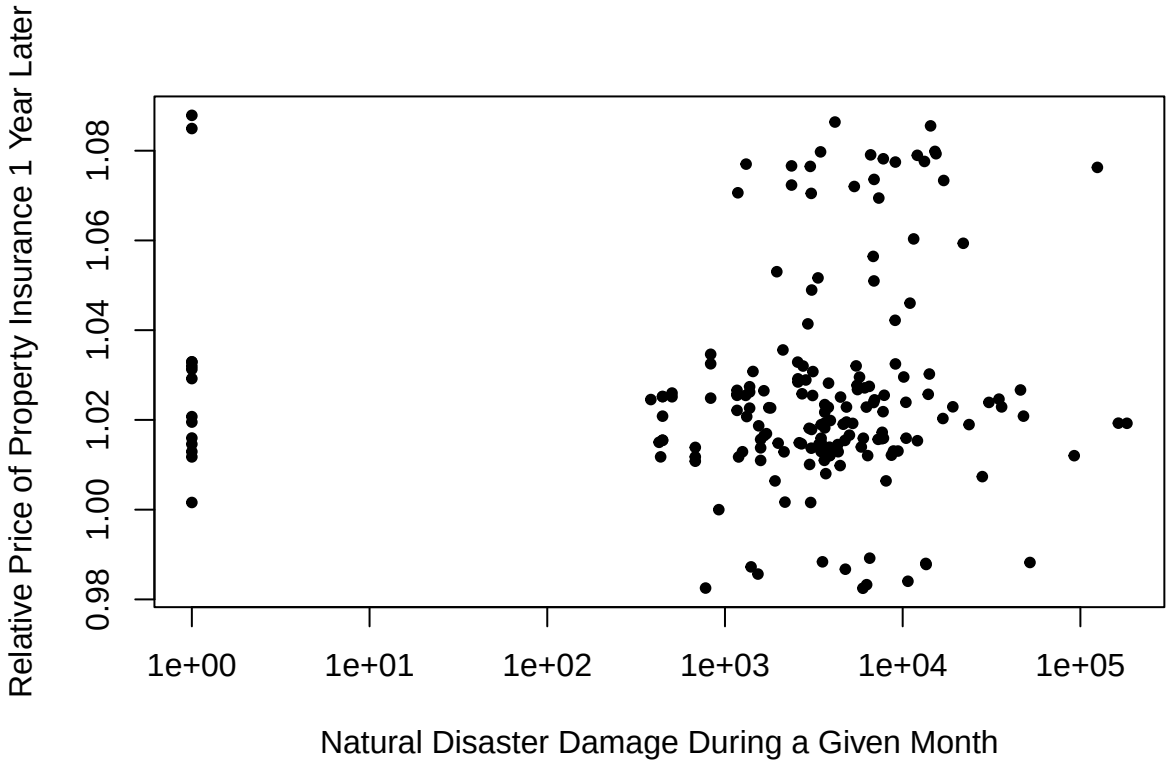
```
## (Intercept)    1.013e+00  9.803e-04 1033.49   <2e-16 ***
## Monthly_Damage 4.478e-09  4.055e-08    0.11    0.912
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01245 on 187 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  6.521e-05,  Adjusted R-squared:  -0.005282
## F-statistic: 0.01219 on 1 and 187 DF,  p-value: 0.9122
```

```
year_delay <- lm(diff_1_year~ Monthly_Damage, data = delayed_damage)
summary(year_delay)
```

```
##
## Call:
## lm(formula = diff_1_year ~ Monthly_Damage, data = delayed_damage)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.044023 -0.012529 -0.004635  0.003900  0.061576
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.026e+00  1.879e-03 546.098   <2e-16 ***
## Monthly_Damage 3.401e-08  7.930e-08   0.429    0.669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02355 on 181 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.001015,   Adjusted R-squared:  -0.004504
## F-statistic: 0.1839 on 1 and 181 DF,  p-value: 0.6685
```

we also run 1 year delayed with the log of the damage done, but to no results. we could run this for the other time groups but that feels excessive-it seems pretty clear there is no correlation.

```
plot(diff_1_year ~ ModifiedDamage, data = delayed_damage, pch = 20, log = "x", xlab = "Natural Disaster
```




```
year_delay <- lm(diff_1_year~ log(ModifiedDamage), data = delayed_damage)
summary(year_delay)
```

```
##
## Call:
## lm(formula = diff_1_year ~ log(ModifiedDamage), data = delayed_damage)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.044280 -0.012354 -0.004389  0.003945  0.062454
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.0254253  0.0056586  181.22   <2e-16 ***
## log(ModifiedDamage) 0.0001540  0.0007014    0.22    0.826
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02356 on 181 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.0002663,  Adjusted R-squared:  -0.005257
## F-statistic: 0.04822 on 1 and 181 DF,  p-value: 0.8264
```

```
print("end of 31")
```

```
## [1] "end of 31"
```