Mathematical Analysis of the change of Environmental Phenomena and Climate Diversity in the United States from January 1951 to April 2025.

Thomas Jackson

2025-12-01

Midterm

Research Questions

Research Questions for the whole United States

1. What relation to the amount of casualties (Deaths/Injuries) per storm event do the daily averages by state of the variables precipitation (mm), maximum temperature ($°C$), and minimum temperature ($°C$) for the entire United States?
2. What relation to the monetary amount of property damage per storm event ($USD) do the daily averages by state of the variables precipitation (mm), maximum temperature ($°C$), and minimum temperature ($°C$) for the entire United States?
3. Do the Eigenvalues of the symmetric matrix $X'X_{cas}$, given by the values of precipitation, and maximum and minimum temperature for the days with recorded casualties from storm events show any meaningful trends in the casualty data?
4. What does the rank of our matrix $X'X_{cas}$ tell us about the independent variables?
5. Can we confirm that the trace of the $X'X_{cas}$ is equal to the sum of its Eigenvalues?
6. Is the determinant of $X'X_{cas}$ equal to the product of its Eigenvalues?
7. Do the Eigenvalues of the symmetric matrix $X'X_{prop}$, given by the values of precipitation, and maximum and minimum temperature for the days with recorded property damage from storms show any meaningful trends in the property damage data?
8. What does the rank of our matrix $X'X_{prop}$ tell us about the independent variables?
9. Does the trace of $X'X_{prop}$ equal the sum of the matrix's Eigenvalues?
10. Is the determinant of $X'X_{prop}$ equal to the product of its Eigenvalues?
11. What information does the singular value decomposition (SVD) of the matrix $X'X_{cas}$ give us?
12. What practical applications of the SVD of the matrix $X'X_{cas}$ are there?
13. What information does the SVD of the matrix $X'X_{prop}$ give us?
14. What practical applications of the SVD of the matrix $X'X_{prop}$ are there?
15. Does the Cholesky decomposition of the matrix $X'X_{cas}$ give us any extra information about the relation between the independent variables and the amount of casualties caused by the storm?

16. Does the Cholesky decomposition of the matrix $X'X_{prop}$ give us any extra information about the relation between the independent variables and the total cost from property damage caused by the storm?
17. What is the interpretation of the Spectral decomposition of the matrix $X'X_{cas}$? What are the possible reasons to use this decomposition?
18. What is the interpretation of the Spectral decomposition of the matrix $X'X_{prop}$? What use-cases does this decomposition have?
19. What conclusions can we draw about the relation of storm event casualties to daily averages over the entire U.S from our analysis?
20. What conclusions can we draw about the relation of property damage due to storm events to daily averages over the entire U.S. from our analysis?

Research Questions for New York State

21. What relation to the amount of casualties (Deaths/Injuries) per storm event do the daily averages in New York State of the variables precipitation (mm), maximum temperature ($^\circ C$), and minimum temperature ($^\circ C$) have when we preform Ordinary Least Squares on our data?
22. How does this relation differ from the relation of daily averages to casualties from storm events by state for the entire U.S. ?
23. What relation to the monetary amount of property damage per storm event ($USD) do the daily averages in New York State of the variables precipitation (mm), maximum temperature ($^\circ C$), and minimum temperature ($^\circ C$) have when we preform Ordinary Least Squares on our data?
24. How does this relation differ from the relation of daily averages to property damage from storm events by state for the entire U.S. ?
25. Do the Eigenvalues of the symmetric matrix $X'X_{cas}$, given by the values of precipitation, and maximum and minimum temperature for the days with recorded casualties from storm events in New York State show any meaningful trends in the casualty data from storm events?
26. What does the rank of our matrix $X'X_{cas}$ tell us about the independent variables?
27. Can we confirm that the trace of the $X'X_{cas}$ is equal to the sum of its Eigenvalues?
28. Is the determinant of $X'X_{cas}$ equal to the product of its Eigenvalues?
29. Do the Eigenvalues of the symmetric matrix $X'X_{prop}$, given by the values of precipitation, and maximum and minimum temperature for the days with recorded property damage from storms in New York State show any meaningful trends in the property damage data?
30. What does the rank of our matrix $X'X_{prop}$ tell us about the independent variables?
31. Does the trace of $X'X_{prop}$ equal the sum of the matrix's Eigenvalues?
32. Is the determinant of $X'X_{prop}$ equal to the product of its Eigenvalues?
33. What information does the singular value decomposition (SVD) of the matrix $X'X_{cas}$ give us?
34. What practical applications of the SVD of the matrix $X'X_{cas}$ are there?
35. What information does the SVD of the matrix $X'X_{prop}$ give us?
36. What practical applications of the SVD of the matrix $X'X_{prop}$ are there?
37. Does the Cholesky decomposition of the matrix $X'X_{cas}$ give us any extra information about the relation between the independent variables and the amount of casualties caused by the storm?

38. Does the Cholesky decomposition of the matrix $X'X_{prop}$ give us any extra information about the relation between the independent variables and the total cost from property damage caused by the storm?
39. What is the interpretation of the Spectral decomposition of the matrix $X'X_{cas}$? What are the possible reasons to use this decomposition?
40. What is the interpretation of the Spectral decomposition of the matrix $X'X_{prop}$? What use-cases does this decomposition have?
41. What conclusions can we draw about the relation of storm event casualties to daily averages for New York state from our analysis?
42. What conclusions can we draw about the relation of property damage due to storm events to daily averages for New York state from our analysis?
43. How do our conclusions about storm casualties for New York State differ from our conclusions for the entire United States?
44. How do our conclusions about storm property damage for New York State differ from our conclusions for the entire United States?

Mathematical Analysis of the Entire United States since 1951

Loading our data

First, we load in the Data set from our csv.

```r
weather_data = read.csv("Data/Weather.csv")
head(weather_data)%>%knitr::kable("markdown")
dat=weather_data[c("TMAX","TMIN","PRCP","DEATHS","INJURIES","PROPERTY_DAMAGE")]
dat=dat[complete.cases(weather_data),]
head(dat)%>%knitr::kable("markdown")
```

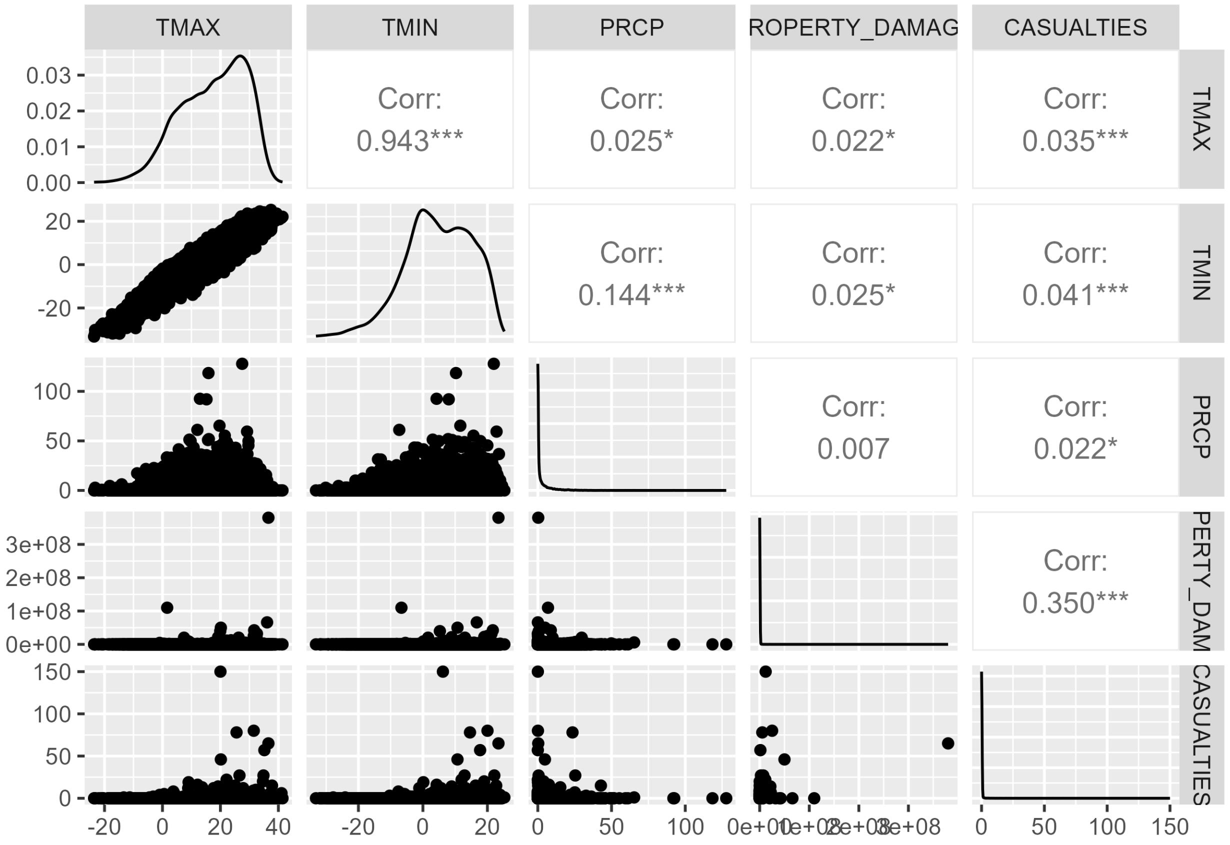
Then, we find only the cases with complete rows

```r
# we get the complete cases of the independent variables
dat=dat[complete.cases(weather_data[c("PRCP","TMAX","TMIN")]),]
head(dat)%>%knitr::kable("markdown")
summary(dat)
```

Now we'll perform Least Squares on the data.

```r
data_only=dat
data_only$CASUALTIES<-as.numeric(data_only$DEATHS)+as.numeric(data_only$INJURIES)

data_only=data_only[,c("TMAX","TMIN","PRCP","PROPERTY_DAMAGE","CASUALTIES")]


head(data_only)


# filter out only non-zero entries for the dependent variables
casualties=data_only%>%
  filter(CASUALTIES>0) %>%
  select(-"PROPERTY_DAMAGE")

y1=casualties$CASUALTIES
x_cas=as.matrix(casualties[,c("TMAX","TMIN","PRCP")])

prop_dmg=data_only%>%
  filter(PROPERTY_DAMAGE>0)%>%
  select(-"CASUALTIES")

# create y2 for dependent variable Property Damage (Dollars)
y2=prop_dmg$PROPERTY_DAMAGE
x_prop=as.matrix(prop_dmg[,c("TMAX","TMIN","PRCP")])


# sample then plot the variables
temp=data_only[sample(1:nrow(data_only),10000),]
ggpairs(temp)
```

```
ggsave("Images/pairs1.png", dpi=400,create.dir=T)
```



Linear Regression/ Ordinary Least Squares

Casualties

```
x=x_cas
y=y1
```

```r
x_1=as.matrix(bind_cols(1,x))
XT_X=t(x_1)%*%x_1
XT_X
```

```
##           ...1       TMAX      TMIN      PRCP
## ...1  22073.0   511569.2  244740.4  123921.5
## TMAX 511569.2 14497031.1 7984470.8 2753538.9
## TMIN 244740.4  7984470.8 4951583.9 1529082.9
## PRCP 123921.5  2753538.9 1529082.9 2710329.4
```

```r
b=solve(XT_X)%*%t(x_1)%*%y
m=lm(y~x)
b
```

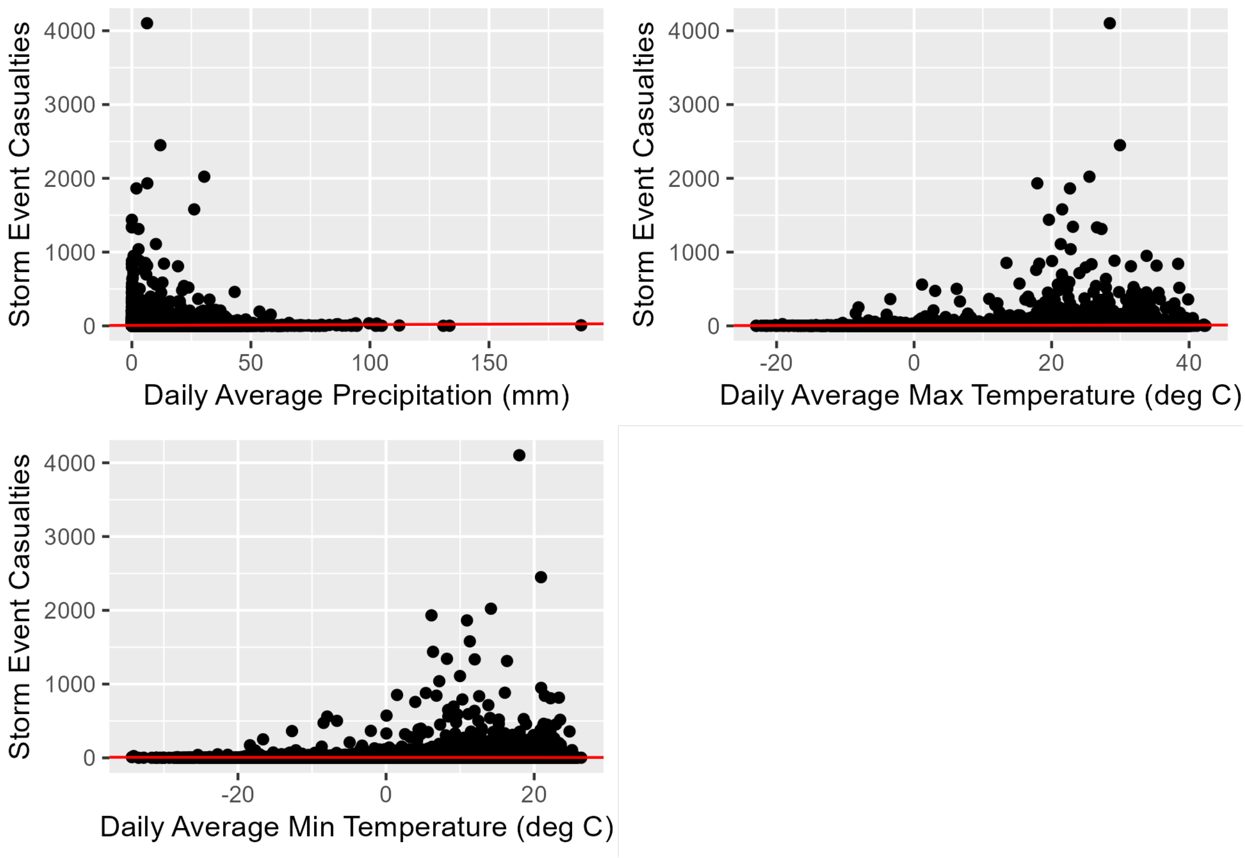
```
##             [,1]
## ...1  6.34167473
## TMAX  0.09293643
## TMIN -0.04454220
## PRCP  0.10675369
```

```r
m
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)        xTMAX        xTMIN        xPRCP
##     6.34167      0.09294     -0.04454      0.10675
```

```r
g1=ggplot(data=casualties,aes(PRCP,CASUALTIES))+
  geom_point()+
  geom_abline(slope=m$coefficients[["xPRCP"]],intercept=m$coefficients[["(Intercept)"]],col="red")+
  xlab("Daily Average Precipitation (mm)")+
  ylab("Storm Event Casualties")


g2=ggplot(data=casualties,aes(TMAX,CASUALTIES))+
  geom_point()+
  geom_abline(slope=m$coefficients[["xTMAX"]],intercept=m$coefficients[["(Intercept)"]],col="red")+
  xlab("Daily Average Max Temperature (deg C)")+
  ylab("Storm Event Casualties")

g3=ggplot(data=casualties,aes(TMIN,CASUALTIES))+
  geom_point()+
  geom_abline(slope=m$coefficients[["xTMIN"]],intercept=m$coefficients[["(Intercept)"]],col="red")+
  xlab("Daily Average Min Temperature (deg C)")+
  ylab("Storm Event Casualties")


plot_grid(g1,g2,g3)
```

```r
ggsave2("Images/grid11.png")
```



This gives and equation of the Form

$$\text{Casualties} = [0.1067537 \cdot \text{Precipation}] + [0.0929364 \cdot \text{Max Temperature`}] + [-0.0445422 \cdot \text{Min Temperature}] + [6.3416747]$$

This implies that deaths due to storms generally increase as the average maximum temperature of the state rises. This could be due to fact that storms generally take place in the more moderate temperatures ranges, or due to snow event deaths making up a large portion of deaths (further research and analysis needed). Also, the amount of deaths caused by storm related events decrease as the minimum temperature increases, this is likely due to the decrease in the loss of deaths from snow events and freezing temps. It is worth noting that tornadoes– which make up a large portion of the death count from storms in this database– can not occur until a dew point of roughly $55°F$ is reached. Then storm related deaths obviously occur more when precipitation is higher because stronger storms usually bring heavier rainfall.

However, we can notice here that the plots do not necessarily show a good relationship from the OLS that we preformed. This probably means there is some non-linear model that is a better fit for these variables.

Property Damage

```
x=x_prop
y=y2
x_1=as.matrix(bind_cols(1,x))
XT_X=t(x_1)%*%x_1
XT_X
```

```
##              ...1      TMAX        TMIN        PRCP
## ...1    78184.0   1930211    986564.8    455804.8
## TMAX 1930210.8  53957104 29786204.5 10324190.1
## TMIN  986564.8 29786205 17811836.5  5755203.7
## PRCP  455804.8 10324190  5755203.7  9092162.5
```

```
b=solve(XT_X)%*%t(x_1)%*%y
m=lm(y~x)
b
```

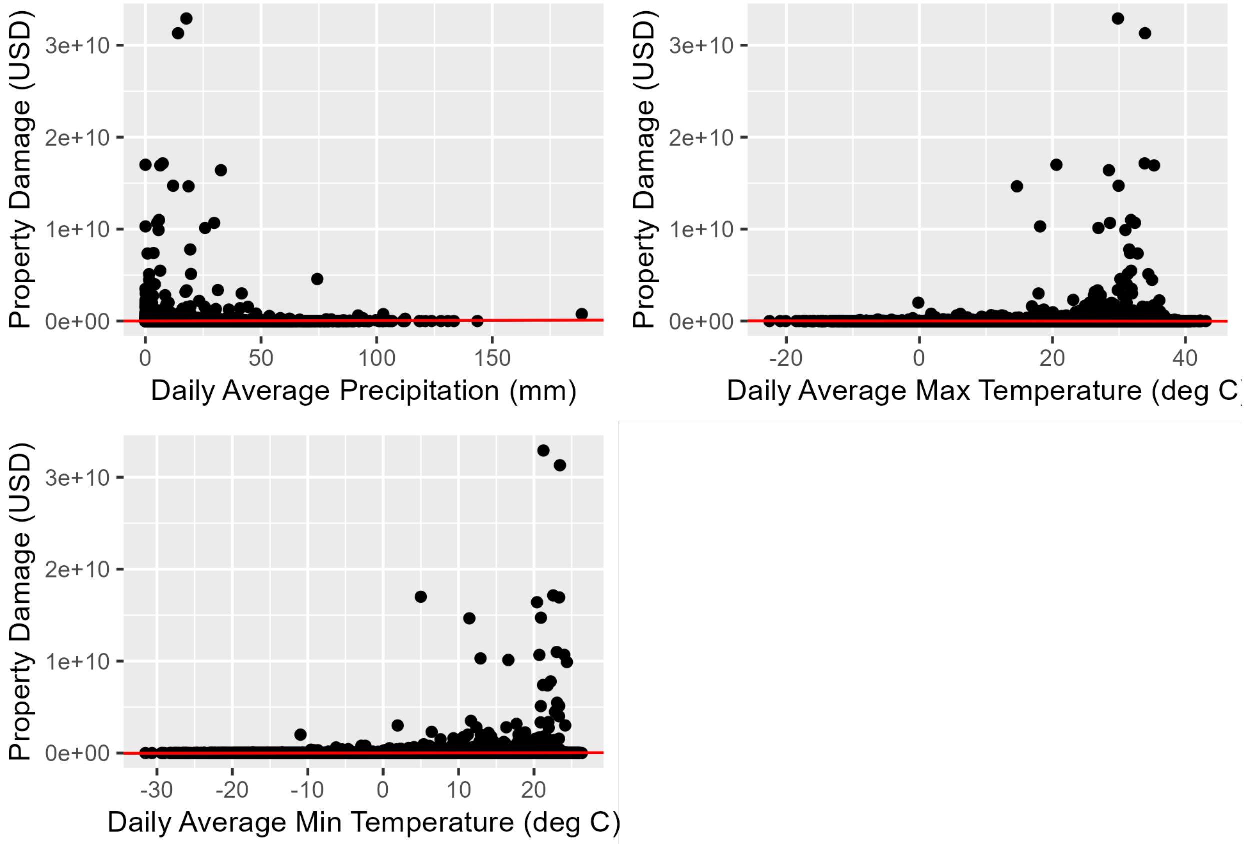
```
##           [,1]
## ...1 6418956.7
## TMAX -608538.7
## TMIN 1027723.9
## PRCP  457756.0
```

```
m
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)        xTMAX        xTMIN        xPRCP
##     6418957      -608539      1027724       457756
```

```
g1=ggplot(data=prop_dmg,aes(PRCP,PROPERTY_DAMAGE))+
  geom_point()+
  geom_abline(slope=m$coefficients[["xPRCP"]],intercept=m$coefficients[["(Intercept)"]],col="red")+
  xlab("Daily Average Precipitation (mm)")+
  ylab("Property Damage (USD)")


g2=ggplot(data=prop_dmg,aes(TMAX,PROPERTY_DAMAGE))+
  geom_point()+
  geom_abline(slope=m$coefficients[["xTMAX"]],intercept=m$coefficients[["(Intercept)"]],col="red")+
  xlab("Daily Average Max Temperature (deg C)")+
  ylab("Property Damage (USD)")

g3=ggplot(data=prop_dmg,aes(TMIN,PROPERTY_DAMAGE))+
  geom_point()+
  geom_abline(slope=m$coefficients[["xTMIN"]],intercept=m$coefficients[["(Intercept)"]],col="red")+
  xlab("Daily Average Min Temperature (deg C)")+
  ylab("Property Damage (USD)")

plot_grid(g1,g2,g3)
```

```
ggsave2("Images/grid12.png")
```



Here we see that property damage ($USD) increases as the daily minimum average for the state increase, and decreases as the max temperature increases. This is likely due to storms occurring at more moderate temperatures, especially tornadoes which often account for more severe damages than other storms. The damage also increases with more precipitation. While the precipitation variable has a similar trend to the precipitation variable in the casualty data, the signs of the TMAX and TMIN coefficients are both swapped. This means the amount of casualties is highest at extreme ranges, while property damage is more likely to be higher at moderate temperatures, with both rising due to increased precipitation.

Eigen Decomposition

Casualties

Now we find the Eigen-Decomposition of the matrix $X'X_{cas}$

```
x=x_cas
XT_X = t(x)%*%x
EI_XT_X = eigen(XT_X)
P = EI_XT_X$vectors
D = diag(EI_XT_X$values)
E = P%*%D%*%solve(P)
E
```

The Eigen-Decomposition of the Matrix $X'X_{cas}$ is

$$X'X_{cas} = \begin{bmatrix} 0.8552 & -0.1511 & 0.4958 \\ 0.4848 & -0.1048 & -0.8683 \\ 0.1832 & 0.9829 & -0.0163 \end{bmatrix} \begin{bmatrix} 19613328.6843 & 0 & 0 \\ 0 & 2123962.2379 & 0 \\ 0 & 0 & 421653.4149 \end{bmatrix} \begin{bmatrix} 0.8552 & 0.4848 & 0.1832 \\ -0.1511 & -0.1048 & 0.9829 \\ 0.4958 & -0.8683 & -0.0163 \end{bmatrix}$$

```
x=x_cas
det_XT_X=det(XT_X)
det_XT_X
prod(EI_XT_X$values)
```

The Determinant of the Matrix $X'X_{cas}$ is 1.7565225×10^{19}. A non-zero determinant means that our matrix is non-singular, i.e. it has an inverse.

Then our determinant is also equal to the product of our eigenvalues of $X'X_{cas} = \prod_{\lambda\in\Lambda}\lambda = 1.7565225 \times 10^{19}$

```
rank1 = rankMatrix(XT_X)
rank1[[1]]
```

The rank of the Matrix $X'X_{cas}$ is 3

```
trace_XT_X = sum(diag(XT_X))
trace_XT_X
sum(EI_XT_X$values)
```

The trace of the Matrix $X'X_{cas}$ is 2.2158944×10^{7}, this is equal to the sum of our eigenvalues, which is 2.2158944×10^{7}

```
E_10=P%*%(D^0.10)%*%solve(P)
E_10
```

The matrix $(X'X_{cas})^{0.10}$ is

$$(X'X_{cas})^{0.10} = \begin{bmatrix} 14497031.0564 & 7984470.8438 & 2753538.891 \\ 7984470.8438 & 4951583.914 & 1529082.9159 \\ 2753538.891 & 1529082.9159 & 2710329.3667 \end{bmatrix}^{0.10} =$$

$$\begin{bmatrix} 0.8552 & -0.1511 & 0.4958 \\ 0.4848 & -0.1048 & -0.8683 \\ 0.1832 & 0.9829 & -0.0163 \end{bmatrix} \begin{bmatrix} 19613328.6843 & 0 & 0 \\ 0 & 2123962.2379 & 0 \\ 0 & 0 & 421653.4149 \end{bmatrix}^{0.10} \begin{bmatrix} 0.8552 & 0.4848 & 0.1832 \\ -0.1511 & -0.1048 & 0.9829 \\ 0.4958 & -0.8683 & -0.0163 \end{bmatrix} =$$

$$\begin{bmatrix} 4.9166 & 0.7189 & 0.1726 \\ 0.7189 & 4.0606 & 0.0858 \\ 0.1726 & 0.0858 & 4.3282 \end{bmatrix}$$

Property Damage

Now we find the Eigen-Decomposition of the matrix $X'X_{prop}$

```
x=x_prop
XT_X = t(x)%*%x
XT_X
EI_XT_X = eigen(XT_X)
P = EI_XT_X$vectors
D = diag(EI_XT_X$values)
E = P%*%D%*%solve(P)
E
```

The Eigen-Decomposition of the Matrix $X'X_{prop}$ is

$$X'X_{prop} = \begin{bmatrix} 0.8569 & -0.1546 & 0.4917 \\ 0.4822 & -0.0965 & -0.8707 \\ 0.1821 & 0.9832 & -0.0082 \end{bmatrix} \begin{bmatrix} 72913705.0831 & 0 & 0 \\ 0 & 6903509.2294 & 0 \\ 0 & 0 & 1043888.9378 \end{bmatrix} \begin{bmatrix} 0.8569 & 0.4822 & 0.1821 \\ -0.1546 & -0.0965 & 0.9832 \\ 0.4917 & -0.8707 & -0.0082 \end{bmatrix}$$

```
det_XT_X=det(XT_X)
det_XT_X
```

The Determinant of the Matrix $X'X_{prop}$ is 5.2545239×10^{20}. A non-zero determinant means that our matrix is non-singular, i.e. it has an inverse.

Then our determinant is also equal to the product of our eigenvalues of $X'X_{cas} = \prod_{\lambda\in\Lambda}\lambda = 5.2545239 \times 10^{20}$

```
rank1 = rankMatrix(XT_X)
rank1[[1]]
```

The rank of the Matrix $X'X_{prop}$ is 3. This is the number of independent variables (pivot columns) in $X'X_{prop}$

```
trace_XT_X = sum(diag(XT_X))
trace_XT_X
```

The trace of the Matrix $X'X_{prop}$ is 8.0861103×10^{7}, this is equal to the sum of our eigenvalues, which is 8.0861103×10^{7}

```
E_10=P%*%(D^0.10)%*%solve(P)
E_10
```

The matrix $(X'X_{prop})^{0.10}$ is
$(X'X_{prop})^{0.10} =$

$$\begin{bmatrix} 53957104.1945 & 29786204.5125 & 10324190.1025 \\ 29786204.5125 & 17811836.5295 & 5755203.674 \\ 10324190.1025 & 5755203.674 & 9092162.5263 \end{bmatrix}^{0.10} =$$

$$\begin{bmatrix} 0.8569 & -0.1546 & 0.4917 \\ 0.4822 & -0.0965 & -0.8707 \\ 0.1821 & 0.9832 & -0.0082 \end{bmatrix} \begin{bmatrix} 72913705.0831 & 0 & 0 \\ 0 & 6903509.2294 & 0 \\ 0 & 0 & 1043888.9378 \end{bmatrix}^{0.10} \begin{bmatrix} 0.8569 & 0.4822 & 0.1821 \\ -0.1546 & -0.0965 & 0.9832 \\ 0.4917 & -0.8707 & -0.0082 \end{bmatrix} =$$

$$\begin{bmatrix} 5.5712 & 0.8865 & 0.2037 \\ 0.8865 & 4.4978 & 0.1068 \\ 0.2037 & 0.1068 & 4.8721 \end{bmatrix}$$

Eigenvectors are defined to be a vector v, with eigenvalues λ such that $Av = \lambda v$ for some nxn matrix A. So the Eigendecomposition involves applying the transformation of the set of eigenvectors (P) to the standard basis, then scaling the resulting matrix by the eigenvalues, then transforming back into the standard basis by applying P^{-1}, giving us A.

So, in our case this lets us do computations involving $X'X$ in the eigenspace, then transform it back into the standard basis. Namely, computing the powers of the matrix $X'X$ utilized the fact that $(PDP^{-1})^n = PD^nP^{-1}$, which allows us to compute the matrix $(X'X)^{.10}$.

This gives us a square matrix E such that $E^{10} = X'X$

Singluar Value Decomposition

Casualties

```
x=x_cas
svd_X=svd(t(x)%*%x)
svd_X
D = svd_X$d
U = svd_X$u
V = svd_X$v
```

```
U %*% diag(D) %*% t(V)

# singular values of X
sapply(svd_X$d,sqrt)
```

With $X'X_{cas} = \begin{bmatrix} 14497031.0564 & 7984470.8438 & 2753538.891 \\ 7984470.8438 & 4951583.914 & 1529082.9159 \\ 2753538.891 & 1529082.9159 & 2710329.3667 \end{bmatrix}$ The Singular Value Decomposition of the matrix $X'X_{cas}$ is

$$X'X_{cas} = \begin{bmatrix} -0.8552 & 0.1511 & 0.4958 \\ -0.4848 & 0.1048 & -0.8683 \\ -0.1832 & -0.9829 & -0.0163 \end{bmatrix} \begin{bmatrix} 19613328.6843 & 0 & 0 \\ 0 & 2123962.2379 & 0 \\ 0 & 0 & 421653.4149 \end{bmatrix} \begin{bmatrix} -0.8552 & -0.4848 & -0.1832 \\ 0.1511 & 0.1048 & -0.9829 \\ 0.4958 & -0.8683 & -0.0163 \end{bmatrix}$$

Singular value decomposition can be thought of as breaking down the matrix M into a rotation of the basis, followed by a scaling by the singular values, then rotating again to achieve the matrix M.

Above we have the case where the matrix is square, however if we wanted to preform this on the original X, we can also find the SVD. In fact, the SVD of the matrix X has singular values that are the square root of the singular values of the matrix $X'X$.

Property Damage

```
x=x_prop
svd_X=svd(t(x)%*%x)
svd_X
D = svd_X$d
U = svd_X$u
V = svd_X$v

U %*% diag(D) %*% t(V)
cat(D,sep=" & ")
```

With $X'X_{prop} = \begin{bmatrix} 53957104.1945 & 29786204.5125 & 10324190.1025 \\ 29786204.5125 & 17811836.5295 & 5755203.674 \\ 10324190.1025 & 5755203.674 & 9092162.5263 \end{bmatrix}$ The Singular Value Decomposition of the matrix $X'X_{prop}$ is

$$X'X_{prop} = \begin{bmatrix} -0.8569 & 0.1546 & 0.4917 \\ -0.4822 & 0.0965 & -0.8707 \\ -0.1821 & -0.9832 & -0.0082 \end{bmatrix} \begin{bmatrix} 72913705.0831 & 0 & 0 \\ 0 & 6903509.2294 & 0 \\ 0 & 0 & 1043888.9378 \end{bmatrix} \begin{bmatrix} -0.8569 & -0.4822 & -0.1821 \\ 0.1546 & 0.0965 & -0.9832 \\ 0.4917 & -0.8707 & -0.0082 \end{bmatrix}$$

Cholesky Decomposition

Casualties

The Cholesky Decomposition is the decomposition of the matrix $X'X$ into the form $X'X = LL^*$, where L is a lower triangular matrix

```
x=x_cas
c=chol(t(x)%*%x)
c
t(c)%*%c
```

$$X'X_{cas} = \begin{bmatrix} 53957104.1945 & 29786204.5125 & 10324190.1025 \\ 29786204.5125 & 17811836.5295 & 5755203.674 \\ 10324190.1025 & 5755203.674 & 9092162.5263 \end{bmatrix}$$

$$\begin{bmatrix} 3807.4967 & 0 & 0 \\ 2097.0395 & 744.3181 & 0 \\ 723.1888 & 16.8308 & 1478.8658 \end{bmatrix} \begin{bmatrix} 3807.4967 & 2097.0395 & 723.1888 \\ 0 & 744.3181 & 16.8308 \\ 0 & 0 & 1478.8658 \end{bmatrix} =$$

$$\begin{bmatrix} 14497031.0564 & 7984470.8438 & 2753538.891 \\ 7984470.8438 & 4951583.914 & 1529082.9159 \\ 2753538.891 & 1529082.9159 & 2710329.3667 \end{bmatrix}$$

Property Damage

```r
x=x_prop
c=chol(t(x)%*%x)
c
t(c)%*%c
```

The Cholesky Decomposition for $X'X_{prop}$ is

$$X'X_{prop} = \begin{bmatrix} 53957104.1945 & 29786204.5125 & 10324190.1025 \\ 29786204.5125 & 17811836.5295 & 5755203.674 \\ 10324190.1025 & 5755203.674 & 9092162.5263 \end{bmatrix}$$

$$\begin{bmatrix} 7345.55 & 0 & 0 \\ 4054.9999 & 1169.9625 & 0 \\ 1405.5027 & 47.7711 & 2667.2913 \end{bmatrix} \begin{bmatrix} 7345.55 & 4054.9999 & 1405.5027 \\ 0 & 1169.9625 & 47.7711 \\ 0 & 0 & 2667.2913 \end{bmatrix} =$$

$$\begin{bmatrix} 53957104.1945 & 29786204.5125 & 10324190.1025 \\ 29786204.5125 & 17811836.5295 & 5755203.674 \\ 10324190.1025 & 5755203.674 & 9092162.5263 \end{bmatrix}$$

Spectral Decomposition

Our spectral decomposition uses the same matrices P and D from our Eigen-Decomposition

Casualties

```r
x=x_cas
# solving for eigen components
XT_X = t(x)%*%x
EI_XT_X = eigen(XT_X)
P = EI_XT_X$vectors
D = diag(EI_XT_X$values)
D=round(D,5)
P=round(P,5)

# spectral calculations
P%*%D%*%t(P)
```

Here we see that we can use the transpose of the matrix P instead of the inverse to find a similar decomposition for the matrix $X'X_{cas}$ in the form of PDP'.

$$\begin{bmatrix} 0.8552 & -0.1511 & 0.4958 \\ 0.4848 & -0.1048 & -0.8683 \\ 0.1832 & 0.9829 & -0.0163 \end{bmatrix} \begin{bmatrix} 19613328.6843 & 0 & 0 \\ 0 & 2123962.2379 & 0 \\ 0 & 0 & 421653.4148 \end{bmatrix} \begin{bmatrix} 0.8552 & 0.4848 & 0.1832 \\ -0.1511 & -0.1048 & 0.9829 \\ 0.4958 & -0.8683 & -0.0163 \end{bmatrix} =$$

$$\begin{bmatrix} 14497029.042 & 7984425.4375 & 2753451.9795 \\ 7984425.4375 & 4951536.5373 & 1529035.0003 \\ 2753451.9795 & 1529035.0003 & 2710274.9263 \end{bmatrix}$$

Property Damages

```
x=x_prop

# Solving for eigen
XT_X = t(x)%*%x
EI_XT_X = eigen(XT_X)
P = EI_XT_X$vectors
D = diag(EI_XT_X$values)
E = P%*%D%*%solve(P)

P=EI_XT_X$vectors
D=diag(EI_XT_X$values)
P%*%D%*%t(P)
```

Here we see that we can use the transpose of the matrix P instead of the inverse to find a similar decomposition for the matrix $X'X_{prop}$ in the form of PDP'.

$$\begin{bmatrix} 0.8569 & -0.1546 & 0.4917 \\ 0.4822 & -0.0965 & -0.8707 \\ 0.1821 & 0.9832 & -0.0082 \end{bmatrix} \begin{bmatrix} 72913705.0831 & 0 & 0 \\ 0 & 6903509.2294 & 0 \\ 0 & 0 & 1043888.9378 \end{bmatrix} \begin{bmatrix} 0.8569 & 0.4822 & 0.1821 \\ -0.1546 & -0.0965 & 0.9832 \\ 0.4917 & -0.8707 & -0.0082 \end{bmatrix} =$$

$$\begin{bmatrix} 53957104.1945 & 29786204.5125 & 10324190.1025 \\ 29786204.5125 & 17811836.5295 & 5755203.674 \\ 10324190.1025 & 5755203.674 & 9092162.5263 \end{bmatrix}$$

Findings:

We found that the property damage is more likely to occur at moderate to warm temperatures, while casualties are more likely to occur at the temperature extremes. We see that both of these dependent variables increase with precipitation. We also found that the matrix $X'X$ has rank 3 for both the property damage and casualty matrices, meaning each of the variables are not linearly dependent on one another in either of these matrices.

Mathematical Analysis of New York State since 1951

```
weather_data = read.csv("Data/Weather.csv")
weather_data=weather_data %>%
  filter(if_any(STATE,~.x=="New York"))
head(weather_data)%>%knitr::kable("markdown")
dat=weather_data[c("TMAX","TMIN","PRCP","DEATHS","INJURIES","PROPERTY_DAMAGE")]
dat=dat[complete.cases(weather_data),]
head(dat)%>%knitr::kable("markdown")
```

Then, we find only the cases with complete rows

```
dat=dat[complete.cases(weather_data[c("PRCP","TMAX","TMIN")]),]
head(dat)%>%knitr::kable("markdown")
```

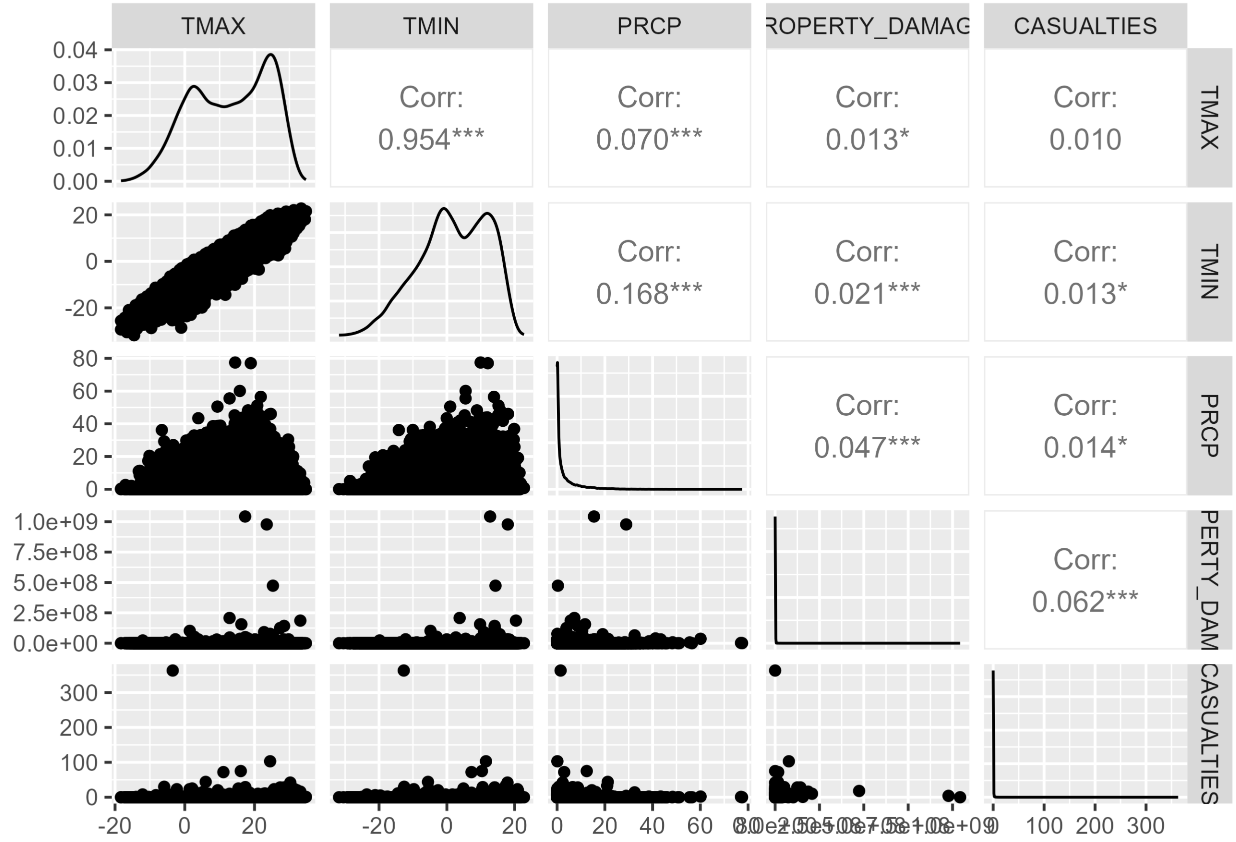
Now we'll perform Least Squares on the data.

```
data_only=dat
data_only$CASUALTIES<-as.numeric(data_only$DEATHS)+as.numeric(data_only$INJURIES)

data_only=data_only[,c("TMAX","TMIN","PRCP","PROPERTY_DAMAGE","CASUALTIES")]



head(data_only)
```

```
##    TMAX  TMIN  PRCP PROPERTY_DAMAGE CASUALTIES
## 1  8.58 -1.16  0.35               0          0
## 2 14.74  3.03  1.33               0          0
## 3 11.37  3.17  7.50               0          0
## 4  9.81  1.88  2.94               0          0
## 5 12.27  1.79 17.33               0          0
## 6 10.08  1.97  3.17               0          0
```

```
# filter out only non-zero entries for the dependent variables
casualties=data_only%>%
  filter(CASUALTIES>0) %>%
  select(-"PROPERTY_DAMAGE")

y1=casualties$CASUALTIES
x_cas=as.matrix(casualties[,c("TMAX","TMIN","PRCP")])

prop_dmg=data_only%>%
  filter(PROPERTY_DAMAGE>0)%>%
  select(-"CASUALTIES")

# create y2 for dependent variable Property Damage (Dollars)
y2=prop_dmg$PROPERTY_DAMAGE
x_prop=as.matrix(prop_dmg[,c("TMAX","TMIN","PRCP")])


# plot the pairs of variables
ggpairs(data_only)
```

```
ggsave("Images/pairs2.png", dpi=400)
```



Least Squares / Linear Regression

Casualties

```
x=x_cas
y=y1
x_1=as.matrix(bind_cols(1,x))
XT_X=t(x_1)%*%x_1
XT_X
```

```
##           ...1      TMAX      TMIN     PRCP
## ...1   523.00  10466.31   4668.12  2796.24
## TMAX 10466.31 269178.12 147135.02 52227.22
## TMIN  4668.12 147135.02  94175.02 25783.96
## PRCP  2796.24  52227.22  25783.96 46914.52
```

```
b=solve(XT_X)%*%t(x_1)%*%y
m=lm(y~x)
b
```

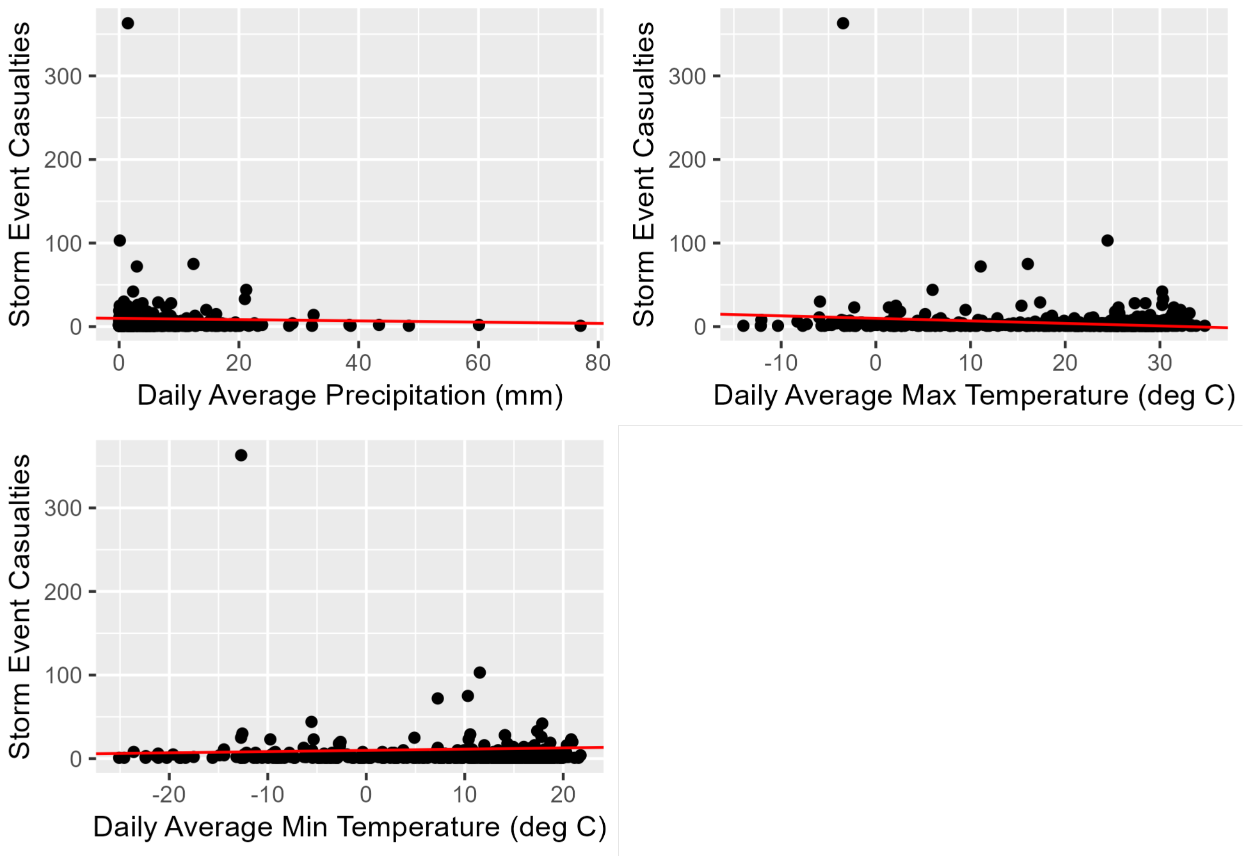
```
##              [,1]
## ...1   9.89323474
## TMAX -0.30103280
## TMIN  0.14683184
## PRCP -0.07506075
```

```
m
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)        xTMAX        xTMIN        xPRCP
##     9.89323     -0.30103      0.14683     -0.07506
```

```
g1=ggplot(data=casualties,aes(PRCP,CASUALTIES))+
  geom_point()+
  geom_abline(slope=m$coefficients[["xPRCP"]],intercept=m$coefficients[["(Intercept)"]],col="red")+
  xlab("Daily Average Precipitation (mm)")+
  ylab("Storm Event Casualties")


g2=ggplot(data=casualties,aes(TMAX,CASUALTIES))+
  geom_point()+
  geom_abline(slope=m$coefficients[["xTMAX"]],intercept=m$coefficients[["(Intercept)"]],col="red")+
  xlab("Daily Average Max Temperature (deg C)")+
  ylab("Storm Event Casualties")

g3=ggplot(data=casualties,aes(TMIN,CASUALTIES))+
  geom_point()+
  geom_abline(slope=m$coefficients[["xTMIN"]],intercept=m$coefficients[["(Intercept)"]],col="red")+
  xlab("Daily Average Min Temperature (deg C)")+
  ylab("Storm Event Casualties")


plot_grid(g1,g2,g3)
```

```
ggsave2("Images/grid21.png")
```



This gives and equation of the form

$$\text{Casualties} = [-0.0750608\cdot\text{Precipation}]+[-0.3010328\cdot\text{Max Temperature}]+[0.1468318\cdot\text{Min Temperature}]+[9.8932347]$$

from the R built-in lm function, and this seems to match our OLS calculations

This implies that casualties due to storms generally ***decrease*** as the average maximum temperature of the state rises, this is different from the national analysis. Since New York likely has less frequent extreme storm events than other parts of the country during the warmer months, a large portion of New York state's storm related casualties may be due to snowstorms, which occur at lower temperatures. Also, the amount of casualties due to storm events ***decreases*** as precipitation rises, this is counter-intuitive but, New York has a relatively moderate climate, so again this could be due to most casualties occurring at lower temperatures with less rain. However, I suspect that all of these variables have better fittings with non-linear models, as the extreme values of these variables likely bring more deaths, implying a non-linear relation.

Property Damage

```
x=x_prop
y=y2
x_1=as.matrix(bind_cols(1,x))
XT_X=t(x_1)%*%x_1
XT_X
```

```
##           ...1       TMAX      TMIN      PRCP
## ...1  2506.00   41474.84  14540.86  13206.66
## TMAX 41474.84 1040005.27 555246.63 214492.86
## TMIN 14540.86  555246.63 384669.36  88605.70
## PRCP 13206.66  214492.86  88605.70 203417.88
```

```
b=solve(XT_X)%*%t(x_1)%*%y
m=lm(y~x)
b
```

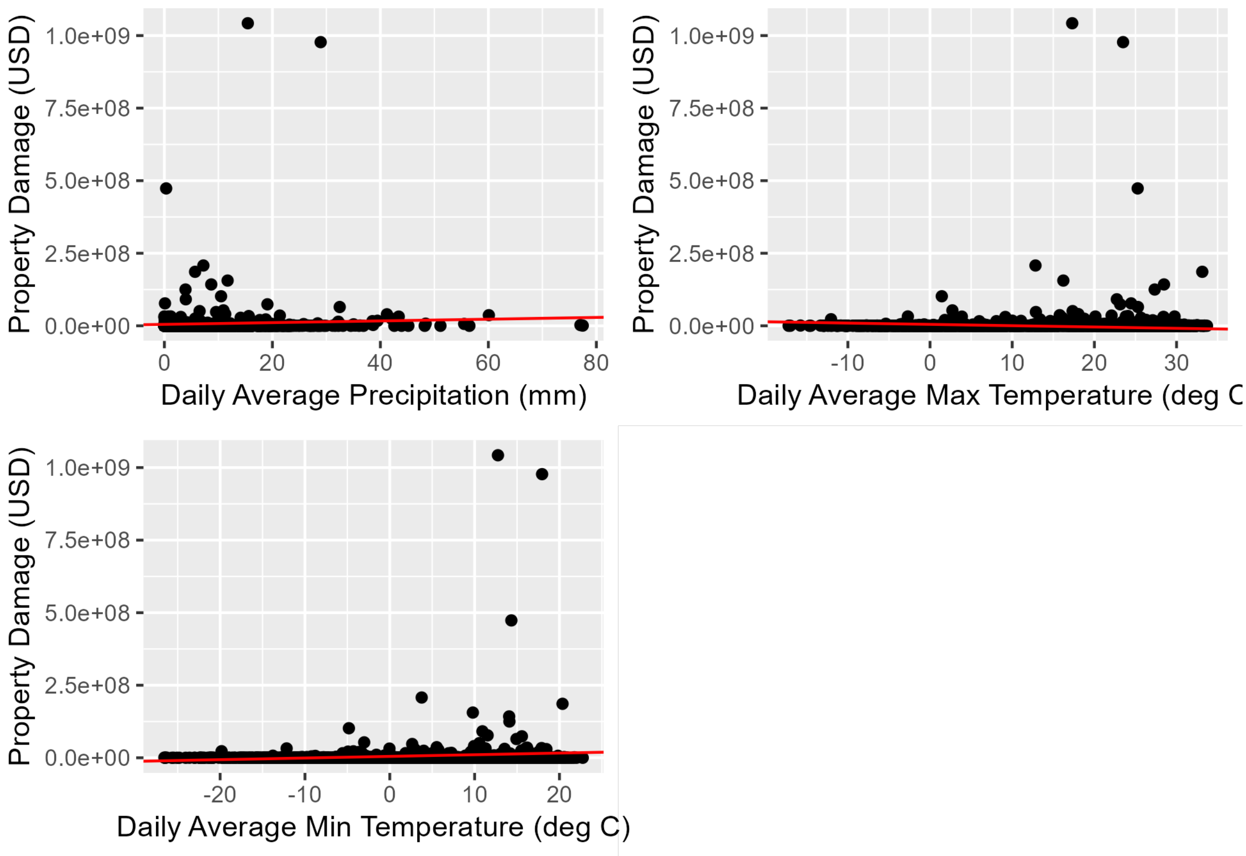
```
##            [,1]
## ...1 4882243.6
## TMAX -450475.9
## TMIN  568786.1
## PRCP  297986.7
```

```
m
```

```
## 
## Call:
## lm(formula = y ~ x)
## 
## Coefficients:
## (Intercept)        xTMAX        xTMIN        xPRCP  
##     4882244      -450476       568786       297987
```

```
g1=ggplot(data=prop_dmg,aes(PRCP,PROPERTY_DAMAGE))+
  geom_point()+
  geom_abline(slope=m$coefficients[["xPRCP"]],intercept=m$coefficients[["(Intercept)"]],col="red")+
  xlab("Daily Average Precipitation (mm)")+
  ylab("Property Damage (USD)")


g2=ggplot(data=prop_dmg,aes(TMAX,PROPERTY_DAMAGE))+
  geom_point()+
  geom_abline(slope=m$coefficients[["xTMAX"]],intercept=m$coefficients[["(Intercept)"]],col="red")+
  xlab("Daily Average Max Temperature (deg C)")+
  ylab("Property Damage (USD)")

g3=ggplot(data=prop_dmg,aes(TMIN,PROPERTY_DAMAGE))+
  geom_point()+
  geom_abline(slope=m$coefficients[["xTMIN"]],intercept=m$coefficients[["(Intercept)"]],col="red")+
  xlab("Daily Average Min Temperature (deg C)")+
  ylab("Property Damage (USD)")

plot_grid(g1,g2,g3)
```

```
ggsave2("Images/grid22.png")
```



This gives and equation of the form

Property Damage in Dollars = $[2.9798668 \times 10^5 \cdot \text{Precipation}] + [-4.5047591 \times 10^5 \cdot \text{Max Temperature`}] + [5.6878607 \times 10^5 \cdot \text{Min Te}$

from the R built-in lm function and this matches the matrix calculations for OLS.

The property damage linear model looks similar to the model for the entire United States, with the magnitude of the coefficients of each of the independent variables being slightly less. This likely means when New York sees storm related property damages they are less extreme than what would be seen in the rest of United States, due to moderate temperatures in New York.

Eigen Decomposition

Casualties

Now we find the Eigen-Decomposition of the matrix $X'X_{cas}$

```
x=x_cas
XT_X = t(x)%*%x
EI_XT_X = eigen(XT_X)
P = EI_XT_X$vectors
D = diag(EI_XT_X$values)
E = P%*%D%*%solve(P)
E
```

The Eigen-Decomposition of the Matrix $X'X_{cas}$ is

$$X'X_{cas} = \begin{bmatrix} 0.8558 & -0.0931 & 0.5089 \\ 0.4847 & -0.1991 & -0.8517 \\ 0.1806 & 0.9755 & -0.1253 \end{bmatrix} \begin{bmatrix} 363543.1384 & 0 & 0 \\ 0 & 36666.0901 & 0 \\ 0 & 0 & 10058.4379 \end{bmatrix} \begin{bmatrix} 0.8558 & 0.4847 & 0.1806 \\ -0.0931 & -0.1991 & 0.9755 \\ 0.5089 & -0.8517 & -0.1253 \end{bmatrix}$$

```
x=x_cas
det_XT_X=det(XT_X)
det_XT_X
```

The Determinant of the Matrix $X'X_{cas}$ is 1.3407602×10^{14}. A non-zero determinant means that our matrix is non-singular, i.e. it has an inverse

```
rank1 = rankMatrix(XT_X)
rank1[[1]]
```

The rank of the Matrix $X'X_{cas}$ is 3, meaning we have 3 independent columns in $X'X_{cas}$

```
trace_XT_X = sum(diag(XT_X))
trace_XT_X
```

The trace of the Matrix $X'X_{cas}$ is 4.1026767×10^5

```
E_10=P%*%(D^0.10)%*%solve(P)
E_10
```

The matrix $(X'X_{cas})^{0.10}$ is
$(X'X_{cas})^{0.10} =$

$$\begin{bmatrix} 269178.1241 & 147135.0184 & 52227.2216 \\ 147135.0184 & 94175.025 & 25783.9624 \\ 52227.2216 & 25783.9624 & 46914.5174 \end{bmatrix}^{0.10} =$$

$$\begin{bmatrix} 0.8558 & -0.0931 & 0.5089 \\ 0.4847 & -0.1991 & -0.8517 \\ 0.1806 & 0.9755 & -0.1253 \end{bmatrix} \begin{bmatrix} 363543.1384 & 0 & 0 \\ 0 & 36666.0901 & 0 \\ 0 & 0 & 10058.4379 \end{bmatrix}^{0.10} \begin{bmatrix} 0.8558 & 0.4847 & 0.1806 \\ -0.0931 & -0.1991 & 0.9755 \\ 0.5089 & -0.8517 & -0.1253 \end{bmatrix} =$$

$$\begin{bmatrix} 3.3107 & 0.4564 & 0.1361 \\ 0.4564 & 2.782 & 0.0276 \\ 0.1361 & 0.0276 & 2.879 \end{bmatrix}$$

Property Damage

Now we find the Eigen-Decomposition of the matrix $X'X_{prop}$

```
x=x_prop
XT_X = t(x)%*%x
EI_XT_X = eigen(XT_X)
P = EI_XT_X$vectors
D = diag(EI_XT_X$values)
E = P%*%D%*%solve(P)
E
```

The Eigen-Decomposition of the Matrix $X'X_{prop}$ is

$$X'X_{prop} = \begin{bmatrix} 0.8548 & -0.0208 & 0.5185 \\ 0.4833 & -0.3316 & -0.8102 \\ 0.1888 & 0.9432 & -0.2734 \end{bmatrix} \begin{bmatrix} 1401320.3378 & 0 & 0 \\ 0 & 167528.3566 & 0 \\ 0 & 0 & 59243.8197 \end{bmatrix} \begin{bmatrix} 0.8548 & 0.4833 & 0.1888 \\ -0.0208 & -0.3316 & 0.9432 \\ 0.5185 & -0.8102 & -0.2734 \end{bmatrix}$$

```
det_XT_X=det(XT_X)
det_XT_X
```

The Determinant of the Matrix $X'X_{prop}$ is 1.3908132×10^{16}. A non-zero determinant means that our matrix is non-singular, i.e. it has an inverse. Then, our determinant is also equal to the product of our eigenvalues of $X'X_{prop} = \prod_{\lambda \in \Lambda} \lambda = 1.3908132 \times 10^{16}$

```
rank1 = rankMatrix(XT_X)
rank1[[1]]
```

The rank of the Matrix $X'X_{prop}$ is 3, this is the number of independent columns (pivots)

```
trace_XT_X = sum(diag(XT_X))
trace_XT_X
```

The trace of the Matrix $X'X_{prop}$ is 1.6280925×10^6, this is equal to the sum of the eigenvalues

```
E_10=P%*%(D^0.10)%*%solve(P)
E_10
```

The matrix $(X'X_{prop})^{0.10}$ is
$(X'X_{prop})^{0.10} =$

$$\begin{bmatrix} 1040005.2694 & 555246.6295 & 214492.8642 \\ 555246.6295 & 384669.3644 & 88605.6956 \\ 214492.8642 & 88605.6956 & 203417.8804 \end{bmatrix}^{0.10} =$$

$$\begin{bmatrix} 0.8548 & -0.0208 & 0.5185 \\ 0.4833 & -0.3316 & -0.8102 \\ 0.1888 & 0.9432 & -0.2734 \end{bmatrix} \begin{bmatrix} 1401320.3378 & 0 & 0 \\ 0 & 167528.3566 & 0 \\ 0 & 0 & 59243.8197 \end{bmatrix}^{0.10} \begin{bmatrix} 0.8548 & 0.4833 & 0.1888 \\ -0.0208 & -0.3316 & 0.9432 \\ 0.5185 & -0.8102 & -0.2734 \end{bmatrix} =$$

$$\begin{bmatrix} 3.8172 & 0.4637 & 0.1738 \\ 0.4637 & 3.298 & -9e-04 \\ 0.1738 & -9e-04 & 3.3332 \end{bmatrix}$$

Singluar Value Decomposition

Casualties

```
x=x_cas
svd_X=svd(t(x)%*%x)
D = svd_X$d
U = svd_X$u
V = svd_X$v

U %*% diag(D) %*% t(V)
```

With $X'X_{cas} = \begin{bmatrix} 269178.1241 & 147135.0184 & 52227.2216 \\ 147135.0184 & 94175.025 & 25783.9624 \\ 52227.2216 & 25783.9624 & 46914.5174 \end{bmatrix}$ The Singular Value Decomposition of the matrix $X'X_{cas}$ is

$$X'X_{cas} = \begin{bmatrix} -0.8558 & 0.0931 & 0.5089 \\ -0.4847 & 0.1991 & -0.8517 \\ -0.1806 & -0.9755 & -0.1253 \end{bmatrix} \begin{bmatrix} 363543.1384 & 0 & 0 \\ 0 & 36666.0901 & 0 \\ 0 & 0 & 10058.4379 \end{bmatrix} \begin{bmatrix} -0.8558 & -0.4847 & -0.1806 \\ 0.0931 & 0.1991 & -0.9755 \\ 0.5089 & -0.8517 & -0.1253 \end{bmatrix}$$

Property Damage

```
x=x_prop
svd_X=svd(t(x)%*%x)
D = svd_X$d
```

```
U = svd_X$u
V = svd_X$v

U %*% diag(D) %*% t(V)
```

With $X'X_{prop} = \begin{bmatrix} 1040005.2694 & 555246.6295 & 214492.8642 \\ 555246.6295 & 384669.3644 & 88605.6956 \\ 214492.8642 & 88605.6956 & 203417.8804 \end{bmatrix}$ The Singular Value Decomposition of the matrix $X'X_{prop}$ is

$$X'X_{prop} = \begin{bmatrix} -0.8548 & -0.0208 & -0.5185 \\ -0.4833 & -0.3316 & 0.8102 \\ -0.1888 & 0.9432 & 0.2734 \end{bmatrix} \begin{bmatrix} 1401320.3378 & 0 & 0 \\ 0 & 167528.3566 & 0 \\ 0 & 0 & 59243.8197 \end{bmatrix} \begin{bmatrix} -0.8548 & -0.4833 & -0.1888 \\ -0.0208 & -0.3316 & 0.9432 \\ -0.5185 & 0.8102 & 0.2734 \end{bmatrix}$$

Cholesky Decomposition

Casualties

The Cholesky Decomposition is the decomposition of the matrix $X'X$ into the form $X'X = LL^*$, where L is a lower triangular matrix

```
x=x_cas
c=chol(t(x)%*%x)
c
t(c)%*%c
```

$$X'X_{cas} = \begin{bmatrix} 1040005.2694 & 555246.6295 & 214492.8642 \\ 555246.6295 & 384669.3644 & 88605.6956 \\ 214492.8642 & 88605.6956 & 203417.8804 \end{bmatrix}$$

$$\begin{bmatrix} 518.8238 & 0 & 0 \\ 283.5934 & 117.2595 & 0 \\ 100.6647 & -23.5706 & 190.3302 \end{bmatrix} \begin{bmatrix} 518.8238 & 283.5934 & 100.6647 \\ 0 & 117.2595 & -23.5706 \\ 0 & 0 & 190.3302 \end{bmatrix} =$$

$$\begin{bmatrix} 269178.1241 & 147135.0184 & 52227.2216 \\ 147135.0184 & 94175.025 & 25783.9624 \\ 52227.2216 & 25783.9624 & 46914.5174 \end{bmatrix}$$

Property Damage

```
x=x_prop
c=chol(t(x)%*%x)
c
t(c)%*%c
```

The Cholesky Decomposition for $X'X_{prop}$ is

$$X'X_{prop} = \begin{bmatrix} 1040005.2694 & 555246.6295 & 214492.8642 \\ 555246.6295 & 384669.3644 & 88605.6956 \\ 214492.8642 & 88605.6956 & 203417.8804 \end{bmatrix}$$

$$\begin{bmatrix} 1019.8065 & 0 & 0 \\ 544.4627 & 297.0348 & 0 \\ 210.327 & -87.2272 & 389.3223 \end{bmatrix} \begin{bmatrix} 1019.8065 & 544.4627 & 210.327 \\ 0 & 297.0348 & -87.2272 \\ 0 & 0 & 389.3223 \end{bmatrix} =$$

$$\begin{bmatrix} 1040005.2694 & 555246.6295 & 214492.8642 \\ 555246.6295 & 384669.3644 & 88605.6956 \\ 214492.8642 & 88605.6956 & 203417.8804 \end{bmatrix}$$

Spectral Decomposition

Our spectral decomposition uses the same matrices P and D from our Eigen-Decomposition

Casualties

```r
x=x_cas
# solving for eigen components
XT_X = t(x)%*%x
EI_XT_X = eigen(XT_X)
P = EI_XT_X$vectors
D = diag(EI_XT_X$values)
D=round(D,5)
P=round(P,5)

# spectral calculations
P%*%D%*%t(P)
```

Here we see that we can use the transpose of the matrix P instead of the inverse to find a similar decomposition for the matrix $X'X_{cas}$ in the form of PDP'.

$$\begin{bmatrix} 0.8558 & -0.0931 & 0.5089 \\ 0.4848 & -0.1991 & -0.8517 \\ 0.1806 & 0.9755 & -0.1252 \end{bmatrix} \begin{bmatrix} 363543.1384 & 0 & 0 \\ 0 & 36666.0901 & 0 \\ 0 & 0 & 10058.4379 \end{bmatrix} \begin{bmatrix} 0.8558 & 0.4848 & 0.1806 \\ -0.0931 & -0.1991 & 0.9755 \\ 0.5089 & -0.8517 & -0.1252 \end{bmatrix} =$$

$$\begin{bmatrix} 269179.2115 & 147136.1929 & 52228.4954 \\ 147136.1929 & 94176.0658 & 25784.7048 \\ 52228.4954 & 25784.7048 & 46914.8213 \end{bmatrix}$$

Property Damages

```r
x=x_prop

# Solving for eigen
XT_X = t(x)%*%x
EI_XT_X = eigen(XT_X)
P = EI_XT_X$vectors
D = diag(EI_XT_X$values)
E = P%*%D%*%solve(P)

P=EI_XT_X$vectors
D=diag(EI_XT_X$values)
P%*%D%*%t(P)
```

Here we see that we can use the transpose of the matrix P instead of the inverse to find a similar decomposition for the matrix $X'X_{prop}$ in the form of PDP'.

$$\begin{bmatrix} 0.8548 & -0.0208 & 0.5185 \\ 0.4833 & -0.3316 & -0.8102 \\ 0.1888 & 0.9432 & -0.2734 \end{bmatrix} \begin{bmatrix} 1401320.3378 & 0 & 0 \\ 0 & 167528.3566 & 0 \\ 0 & 0 & 59243.8197 \end{bmatrix} \begin{bmatrix} 0.8548 & 0.4833 & 0.1888 \\ -0.0208 & -0.3316 & 0.9432 \\ 0.5185 & -0.8102 & -0.2734 \end{bmatrix} =$$

$$\begin{bmatrix} 1040005.2694 & 555246.6295 & 214492.8642 \\ 555246.6295 & 384669.3644 & 88605.6956 \\ 214492.8642 & 88605.6956 & 203417.8804 \end{bmatrix}$$

Midterm Conclusion

The data from the regression model for casualty for the entire United States had different trends for each of the independent variables, when compared to the regression model for New York State. This shows that weather related casualty predictions in New York State should use a separate model for prediction based on these variables. We also found that there is less monetary damage to properties received from storms in New York than in the rest of the United States, with the trends of the independent variables being the same. We see that property damage decreases as the maximum daily average for the state increases, and increases as the daily average minimum temperature and average amount of precipitation increase. Overall, I believe that from the results, the linear model for this data is not the best suited model, and further analysis could be done on this data set with non-linear methods.

Final Project

Research Questions

Research Questions for Entire United States

1. How can we interpret the Matrix $X'X_{prop}$ as a quadratic form?
2. How can we interpret the Matrix $X'X_{cas}$ as a quadratic form?
3. How do we perform Weighted Least Squares on the property damage data for the United States?
4. What weights should our weight matrix W_{prop} take?
5. What is the format of the matrix W_{prop}?
6. How does this compare to the OLS model we found earlier?
7. How do we perform weighted Least squares on the casualty data for the United States?
8. What weights should our weight matrix W_{cas} take?
9. What is the format of the matrix W_{cas}?
10. How does this compare to the OLS model we found earlier?
11. What is the Principle Component Analysis for the property damage data?
12. Does this Principle Component Analysis imply we can reduce the dimension of our data, without losing substantial information about the relation between weather readings and the amount of property damage that storm events cause?
13. What is the Principle Component Analysis for the Casualty data?
14. Does this Principle Component Analysis imply we can reduce the dimension of our data, without losing substantial information about the relation between weather readings and casualties caused by storm events?

Research Questions for New York State

15. How can we interpret the Matrix $X'X_{prop}$ as a quadratic form?
16. How can we interpret the Matrix $X'X_{cas}$ as a quadratic form?
17. How do we perform Weighted Least Squares on the property damage data for New York State?
18. What weights should our weight matrix W_{prop} take?
19. What is the format of the matrix W_{prop}?
20. How does this compare to the OLS model we found earlier?
21. How do we perform weighted Least squares on the casualty data for New York State?
22. What weights should our weight matrix W_{cas} take?
23. What is the format of the matrix W_{cas}?
24. How does this compare to the OLS model we found earlier?
25. What is the Principle Component Analysis for the property damage data?
26. Does this Principle Component Analysis imply we can reduce the dimension of our data, without losing substantial information about the relation between weather readings and the amount of property damage that storm events cause?
27. What is the Principle Component Analysis for the Casualty data?

28. Does this Principle Component Analysis imply we can reduce the dimension of our data, without losing substantial information about the relation between weather readings and casualties caused by storm events?

Analysis of the Entire United States

Quadratic Form

Property Damage

```
x=x_prop
XT_X=t(x)%*%x
XT_X
```

This matrix $X'X_{prop} = \begin{bmatrix} 53957104.1945 & 29786204.5125 & 10324190.1025 \\ 29786204.5125 & 17811836.5295 & 5755203.674 \\ 10324190.1025 & 5755203.674 & 9092162.5263 \end{bmatrix}$ gives us an equation with quadratic form $5.3957104 \times 10^7 x_1^2 + 1.7811837 \times 10^7 x_2^2 + 9.0921625 \times 10^6 x_3^2 + 5.9572409 \times 10^7 x_1 x_2 + 2.064838 \times 10^7 x_1 x_3 + 1.1510407 \times 10^7 x_2 x_3$

Casualties

```
x=x_cas
XT_X=t(x)%*%x
XT_X
```

The matrix $X'X_{cas} = \begin{bmatrix} 14497031.0564 & 7984470.8438 & 2753538.891 \\ 7984470.8438 & 4951583.914 & 1529082.9159 \\ 2753538.891 & 1529082.9159 & 2710329.3667 \end{bmatrix}$ gives us an equation with quadratic form $1.4497031 \times 10^7 x_1^2 + 4.9515839 \times 10^6 x_2^2 + 2.7103294 \times 10^6 x_3^2 + 1.5968942 \times 10^7 x_1 x_2 + 5.5070778 \times 10^6 x_1 x_3 + 3.0581658 \times 10^6 x_2 x_3$

Weighted Least Squares

Property Damage

```
x=x_prop
y=y2
x_1=as.matrix(bind_cols(1,x))
m=lm(y~x)
as.matrix(m$coefficients)


beta=solve(t(x_1)%*%x_1)%*%t(x_1)%*%y


W=bandSparse(nrow(x_1),nrow(x_1),0,list(1/(abs(y-(x_1%*%beta)))))
wls=solve(t(x_1)%*%W%*%x_1)%*%t(x_1)%*%W%*%y
wls
```

For weighted least squares we have the solution $\hat{\beta} = (X'WX)^{-1}t(X)WY$, where W is our weight matrix. To define W we make a square matrix with 78184 rows and columns, so that we have $dim(X'WX) = dim(X'X)$. This square matrix has the reciprocal of the residuals of each measurement on the diagonal. Since this matrix is very large with 6.1127379×10^9 entries, it would be very hard to fit in memory with the standard matrix object. Luckily since the majority of these entries are 0, we have a sparse matrix which let's us fix this by using the bandSparse function from the matrix package.

The weights in W are given according to $w_i = \dfrac{1}{|y_i - X_i\beta|}$, this formulation is called Least Absolute Residual Regression.

Comparing our OLS to the new model gives us:

$$\beta_{OLS} = \begin{bmatrix} 6418956.6576 \\ -608538.7418 \\ 1027723.8575 \\ 457756.015 \end{bmatrix} \text{ and } \beta_{WLS} = \begin{bmatrix} 1135115.6386 \\ -102337.2002 \\ 181201.9885 \\ 206537.9576 \end{bmatrix}$$

The Weighted Least Squares model drastically scales down all the components of β compared to the Ordinary Least Squares. The sign of each component stays the same, our weights are all positive.

Casualties

```r
x=x_cas
y=y1

x_1=as.matrix(bind_cols(1,x))

m=lm(y~x)
as.matrix(m$coefficients)

beta=solve(t(x_1)%*%x_1)%*%t(x_1)%*%y
W=bandSparse(nrow(x_1),nrow(x_1),0,list(1/(abs(y-(x_1%*%beta)))))
wls=solve(t(x_1)%*%W%*%x_1)%*%t(x_1)%*%W%*%y
wls
```

The weights in W are given according to $w_i = \dfrac{1}{|y_i - X_i\beta|}$ again. So we have W is a 22073 × 22073 matrix.

Comparing our OLS to the new model gives us:

$$\beta_{OLS} = \begin{bmatrix} 6.3417 \\ 0.0929 \\ -0.0445 \\ 0.1068 \end{bmatrix} \text{ and } \beta_{WLS} = \begin{bmatrix} 4.0731 \\ 0.1204 \\ -0.0964 \\ 0.0755 \end{bmatrix}$$

This shows that we have a smaller intercept, more weight given to TMAX, more negative weight given to TMIN, and less weight given to precipitation.

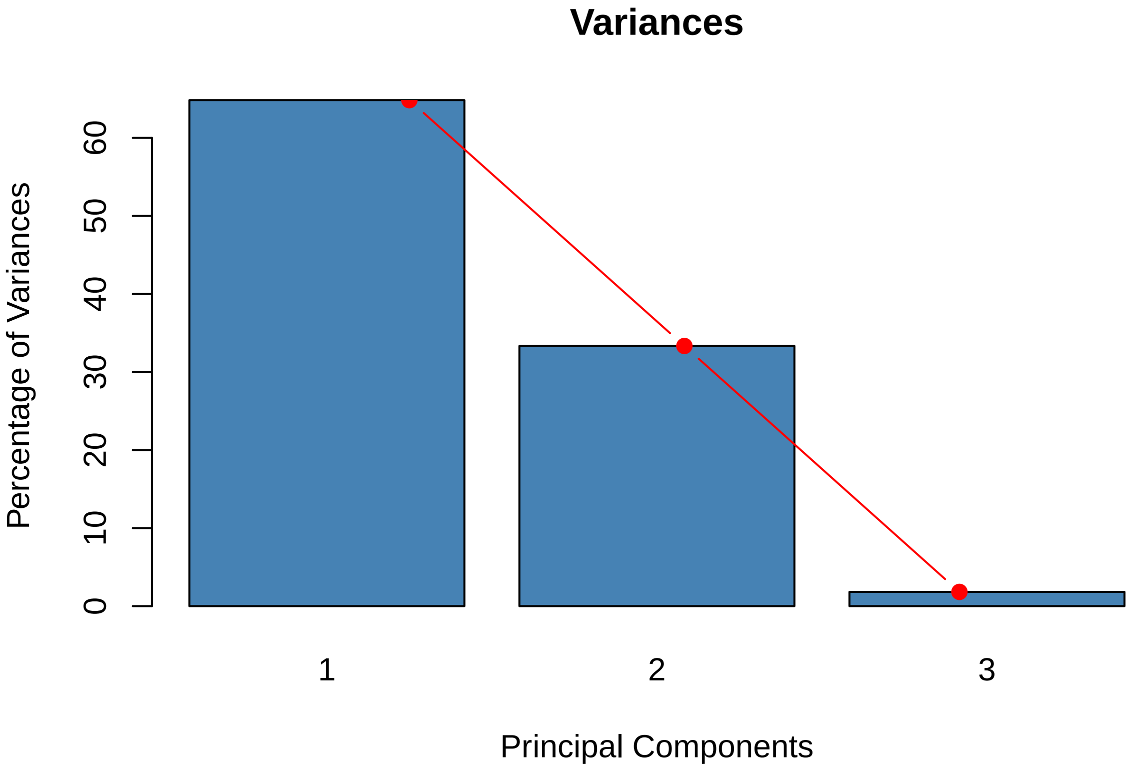
Principle Component Analysis

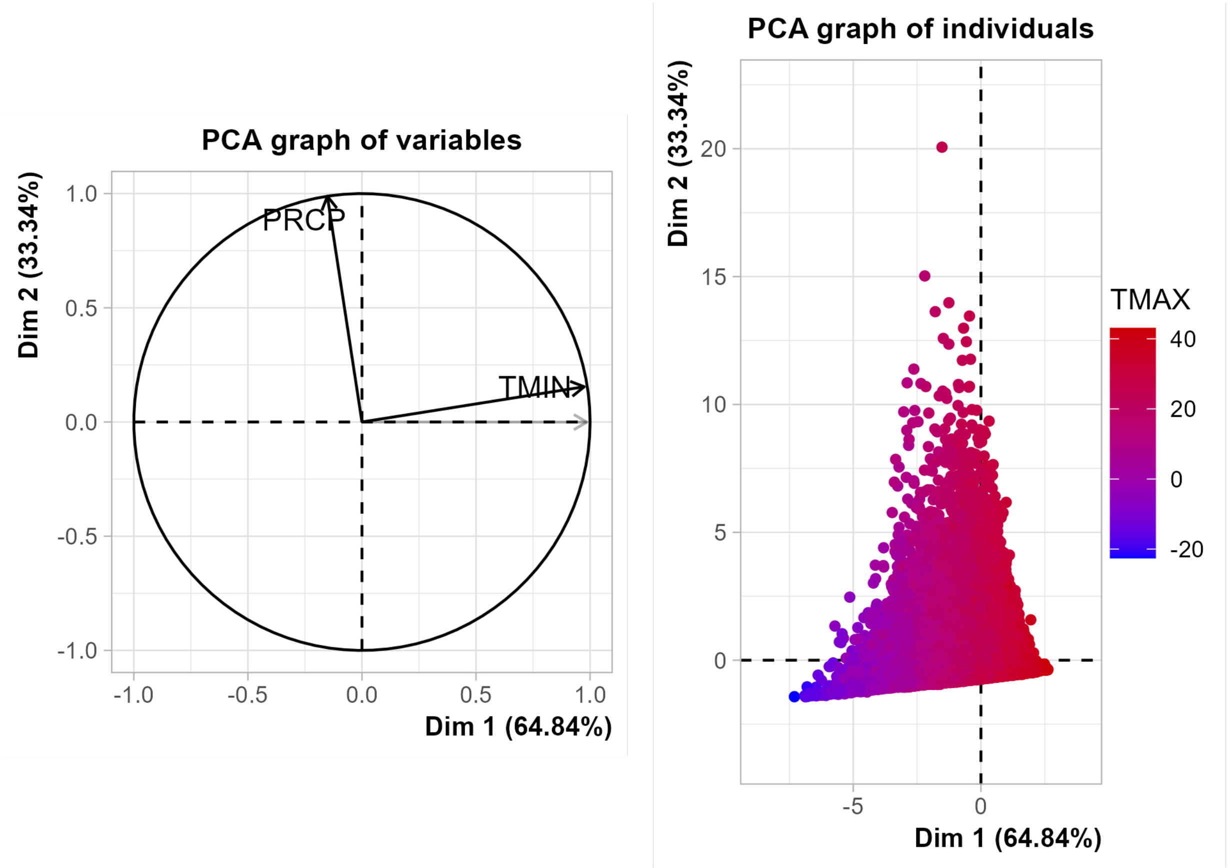
Property Damage

```r
x=x_prop

pca_x=PCA(x,graph=F)
g1=plot.PCA(pca_x, choix='var',select='contrib 2')
g2=plot.PCA(pca_x,axes=c(1,2), cex=1,choix='ind',label = 'none', habillage = 1)



plot_grid(g1,g2)
```

```r
ggsave2("Images/PCA_Prop_USA.png")
```

```r
eigenvalues=pca_x$eig
barplot(eigenvalues[, 2], names.arg=1:nrow(eigenvalues),
       main = "Variances",
       xlab = "Principal Components",
       ylab = "Percentage of Variances",
       col ="steelblue")
lines(x = 1:nrow(eigenvalues), eigenvalues[, 2],
      type="b", pch=19, col = "red")
```





This shows that Minimum Temperature and Precipitation account for ~98% of the variance in the property damage data. This indicates that we can apply dimension reduction to our data in order to make a more concise model, only depending on Minimum temperature and Precipitation.

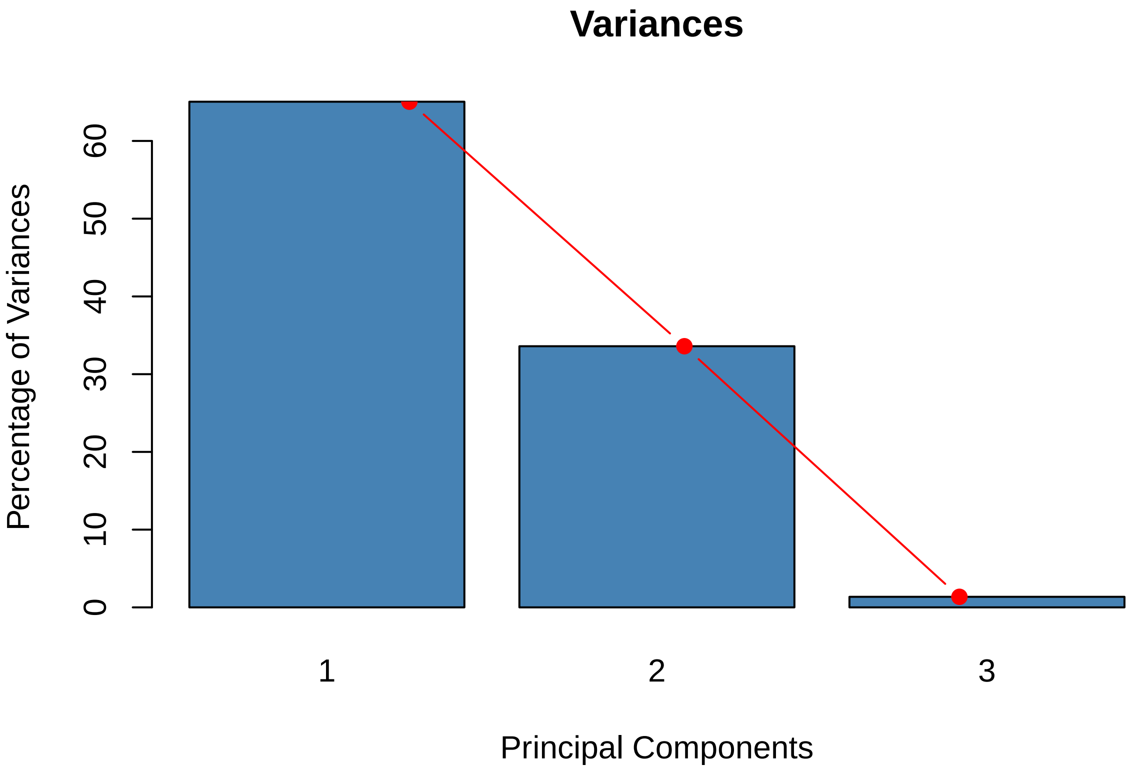
The results of the PCA for property damage are surprising to me, since I would assume that most property damage would occur in the southern U.S. during hurricane/storm season, or in tornado Alley during the peak months, leading to more weight being applied to the Maximum temperature. However, our PCA shows that more weight is applied to the Minimum temperature, which I do not have any explanation for.

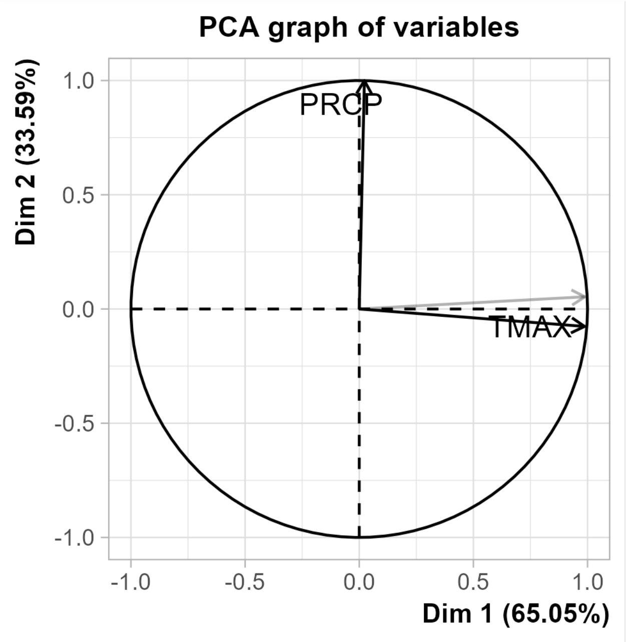
Casualties

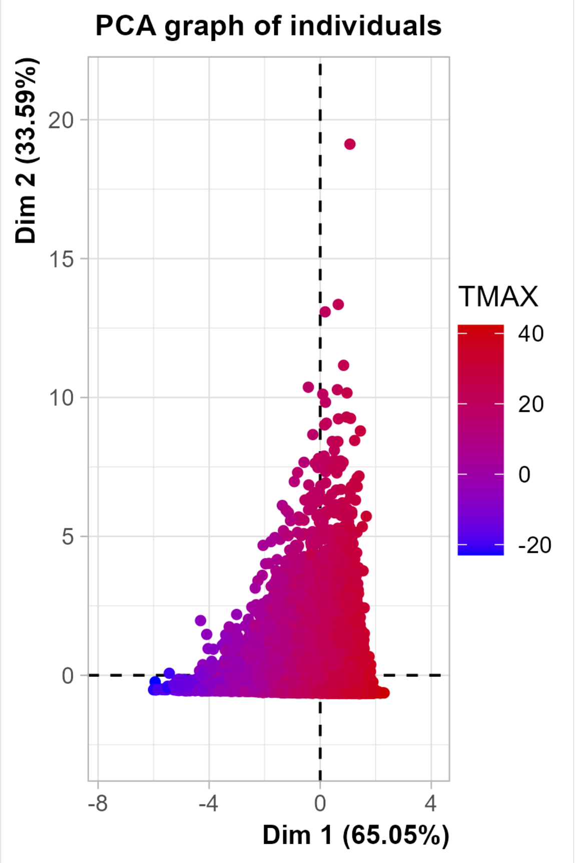
```
x=x_cas

pca_x=PCA(x,graph=F)
g1=plot.PCA(pca_x, choix='var',select='contrib 2')
g2=plot.PCA(pca_x,axes=c(1,2), cex=1,choix='ind',label = 'none', habillage = 1)


plot_grid(g1,g2)
```

```
ggsave2("Images/PCA_Cas_USA.png")
```

```
eigenvalues=pca_x$eig
barplot(eigenvalues[, 2], names.arg=1:nrow(eigenvalues),
       main = "Variances",
       xlab = "Principal Components",
       ylab = "Percentage of Variances",
        col ="steelblue")
lines(x = 1:nrow(eigenvalues), eigenvalues[, 2],
      type="b", pch=19, col = "red")
```







The results of the PCA on casualty data for the United States shows that the Precipitation and Max Temperature have the most weight to them.

They account for ~99% of the variance, and it is reasonable to assume that a majority of the US endures events that cause massive amounts of casualties during the warmer months, when extreme storms rip through the Southeast, and tornadoes tear up the Midwest.

Analysis of New York State

Quadratic Form

Property Damage

```
x=x_prop
XT_X=t(x)%*%x
XT_X
```

The matrix $X'X_{prop} = \begin{bmatrix} 1040005.2694 & 555246.6295 & 214492.8642 \\ 555246.6295 & 384669.3644 & 88605.6956 \\ 214492.8642 & 88605.6956 & 203417.8804 \end{bmatrix}$ gives us an equation with quadratic form $1.0400053 \times 10^6 x_1^2 + 3.8466936 \times 10^5 x_2^2 + 2.0341788 \times 10^5 x_3^2 + 1.1104933 \times 10^6 x_1 x_2 + 4.2898573 \times 10^5 x_1 x_3 + 1.7721139 \times 10^5 x_2 x_3$

Casualties

```
x=x_cas
XT_X=t(x)%*%x
XT_X
```

The matrix $X'X_{cas} = \begin{bmatrix} 269178.1241 & 147135.0184 & 52227.2216 \\ 147135.0184 & 94175.025 & 25783.9624 \\ 52227.2216 & 25783.9624 & 46914.5174 \end{bmatrix}$ gives us an equation with quadratic form $2.6917812 \times 10^5 x_1^2 + 9.4175025 \times 10^4 x_2^2 + 4.6914517 \times 10^4 x_3^2 + 2.9427004 \times 10^5 x_1 x_2 + 1.0445444 \times 10^5 x_1 x_3 + 5.1567925 \times 10^4 x_2 x_3$

Weighted Least Squares

Property Damage

```r
x=x_prop
y=y2

x_1=as.matrix(bind_cols(1,x))

m=lm(y~x)
as.matrix(m$coefficients)

beta=solve(t(x_1)%*%x_1)%*%t(x_1)%*%y
W=bandSparse(nrow(x_1),nrow(x_1),0,list(1/(abs(y-(x_1%*%beta)))))
wls=solve(t(x_1)%*%W%*%x_1)%*%t(x_1)%*%W%*%y
wls
```

The weights in W are given according to $w_i = \dfrac{1}{|y_i - X_i\beta|}$ again. So we have W is a 2506 × 2506 matrix.

Comparing our OLS to the new model gives us:

$$\beta_{OLS} = \begin{bmatrix} 4882243.6045 \\ -450475.9112 \\ 568786.0662 \\ 297986.6847 \end{bmatrix} \text{ and } \beta_{WLS} = \begin{bmatrix} 1521027.7869 \\ -133205.1846 \\ 167217.4394 \\ 92013.1014 \end{bmatrix}$$

The weights of the Weighted Least Squares model are significantly scaled down, indicating that the residuals are rather significant in the Ordinary Least Squares model. This means our WLS model is hopefully more accurate in predicting property damage.

Casualties

```r
x=x_cas
y=y1

x_1=as.matrix(bind_cols(1,x))

m=lm(y~x)
as.matrix(m$coefficients)

beta=solve(t(x_1)%*%x_1)%*%t(x_1)%*%y
W=bandSparse(nrow(x_1),nrow(x_1),0,list(1/(abs(y-(x_1%*%beta)))))
wls=solve(t(x_1)%*%W%*%x_1)%*%t(x_1)%*%W%*%y
wls
```

The weights in W are given according to $w_i = \dfrac{1}{|y_i - X_i\beta|}$ again. So we have W is a 523 × 523 matrix.

Comparing our OLS to the new model gives us:

$$\beta_{OLS} = \begin{bmatrix} 9.8932 \\ -0.301 \\ 0.1468 \\ -0.0751 \end{bmatrix} \text{ and } \beta_{WLS} = \begin{bmatrix} 6.6819 \\ -0.1408 \\ 0.0317 \\ -0.0476 \end{bmatrix}$$

Again, all of the weights are scaled down a good amount, none of the signs of the weights change– as expected.

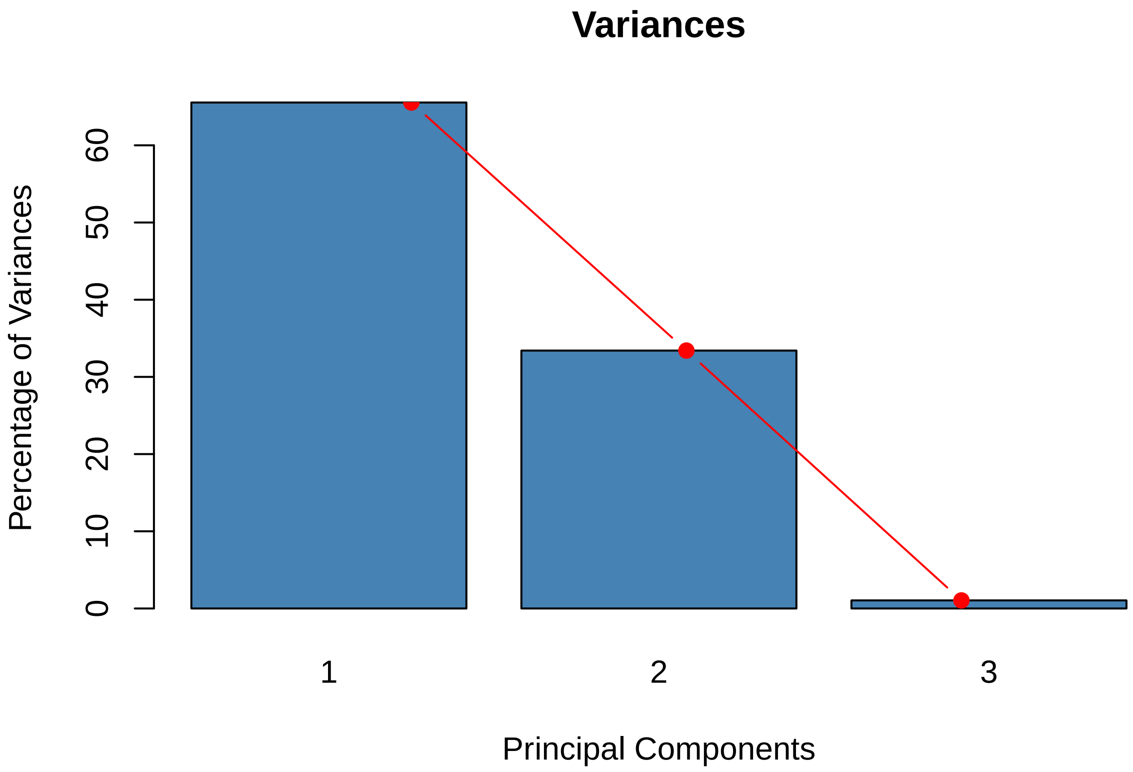
Principle Component Analysis

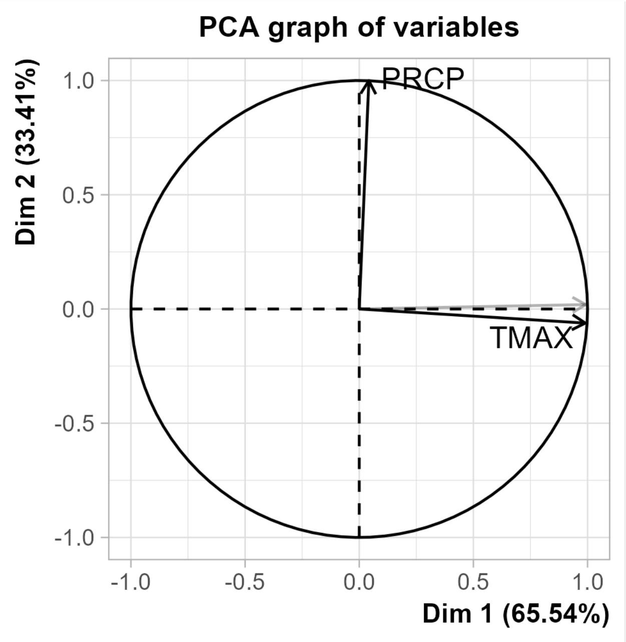
Property Damage

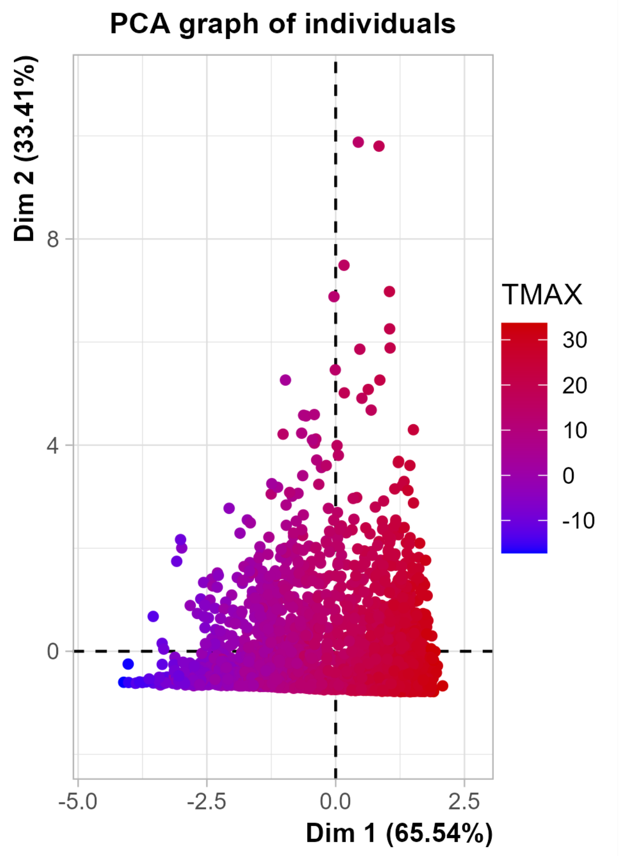
```
x=x_prop

pca_x=PCA(x,graph=F)
g1=plot.PCA(pca_x, choix='var',select='contrib 2')
g2=plot.PCA(pca_x,axes=c(1,2), cex=1,choix='ind',label = 'none', habillage = 1)


plot_grid(g1,g2)
```

```
ggsave2("Images/PCA_Prop_NYS.png")
```

```
eigenvalues=pca_x$eig
barplot(eigenvalues[, 2], names.arg=1:nrow(eigenvalues),
       main = "Variances",
       xlab = "Principal Components",
       ylab = "Percentage of Variances",
       col ="steelblue")
lines(x = 1:nrow(eigenvalues), eigenvalues[, 2],
      type="b", pch=19, col = "red")
```







Here we can see that the two most influential principle components on the number of property damage are the amount of precipitation and the maximum temperature for the day. Since these two combined account for ~99% of the variance, we can find a model that predicts the amount of property damage from storm events with similar accuracy to our current model without using Minimum Temperature as an independent variable.

These two variables probably account for most of the variance since damaging storm events like tornadoes and hurricanes occur at warmer temperatures, and bring more precipitation.

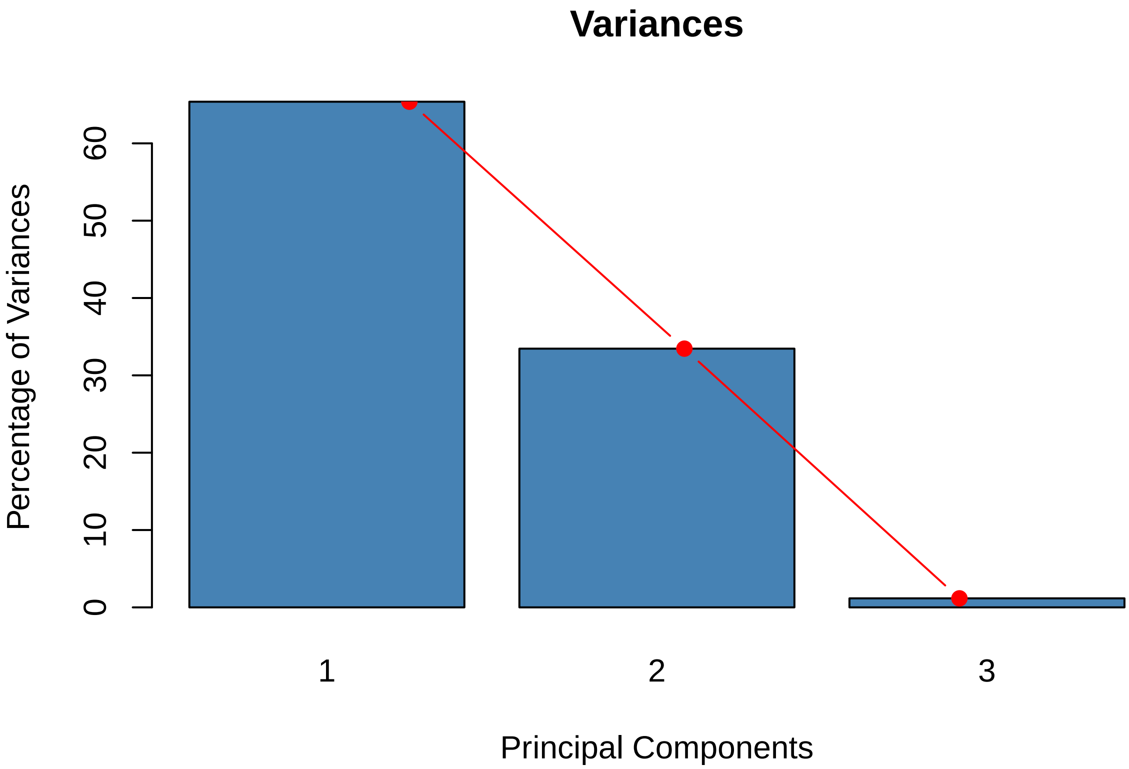
This result is different from the property damage result for the entire United States, further analysis of the data is needed to understand why this is the case.

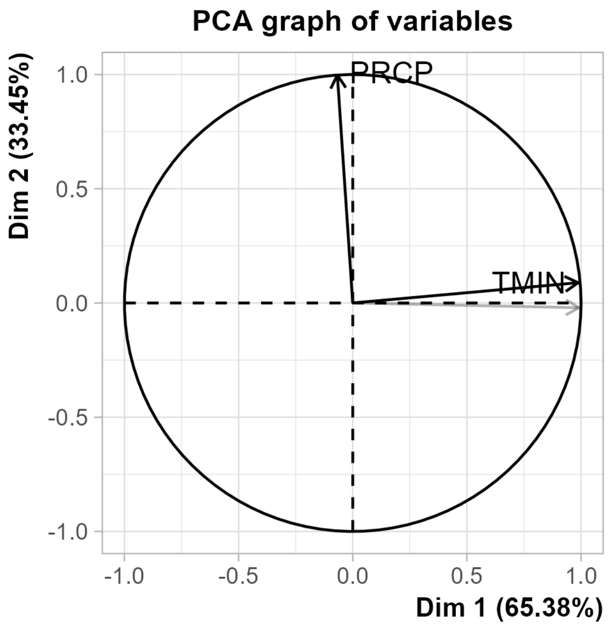
Casualties

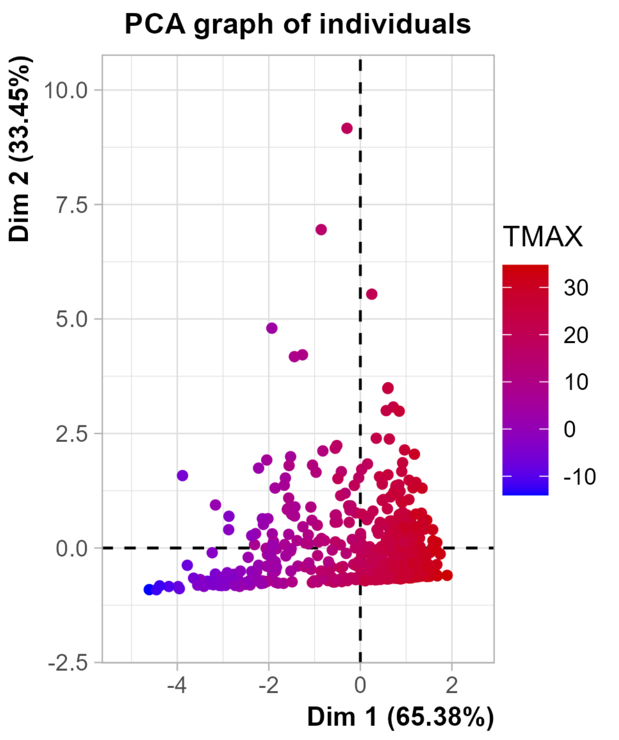
```
x=x_cas

pca_x=PCA(x,graph=F)

g1=plot.PCA(pca_x, choix='var',select='contrib 2')
g2=plot.PCA(pca_x,axes=c(1,2), cex=1,choix='ind',label = 'none', habillage = 1)
plot_grid(g1,g2)
```

```
ggsave2("Images/PCA_Cas_NYS.png")
```

```
eigenvalues=pca_x$eig
barplot(eigenvalues[, 2], names.arg=1:nrow(eigenvalues),
       main = "Variances",
       xlab = "Principal Components",
       ylab = "Percentage of Variances",
        col ="steelblue")
lines(x = 1:nrow(eigenvalues), eigenvalues[, 2],
      type="b", pch=19, col = "red")
```







For this we have that the amount of casualties is most dependent on precipitation and minimum temperature for New York State. This is likely because of poor driving conditions caused by frequent winter storms.

This also differs from the rest of the United States where Maximum temperature has more weight on the variance than minimum temperature.

Final Conclusion

The Weighted Least squares analysis provided us with a new predictive model for both the United States and New York state, for property damage and casualties. While the effectiveness of this model needs to be tested, these new models provide alternative results from the Ordinary Least squares models of the same data. The Principle Component Analysis gave us some insight on the differences between New York State and the United States that reinforce some of our findings from the midterm project that showed clear differences in these two geographic scopes for the trends of property damage and casualties from storm events. Overall, we have not found a great predicative model for our data, but we have made found enough evidence to conclude that our scope for our model should be state-wide, or at least more local than Country-wide.

Project Summary

We have shown that the trends for the casualties and property damage from weather events for the entire United States differ from trends for the State of New York. By employing various mathematical techniques, we can conclude that analysis of casualties and property damage from storm events should be done on a state-wide level, instead of a national level. The difference in biomes across the United States varies too much to have a good predictor that is solely based on Minimum Temperature, Maximum Temperature and Precipitation. Our results should tells us that we need to look for a model that relies on local weather data, and that we should consider using more predicative variables when we try to make a new model for predicting both casualties and property damage from storm damage. While we did not have any amazing findings from our analysis, we still have more information on how to build a better model for these data sets.

SOURCES:

NOAA:

National Climate Grid Daily v1-0-0: https://www.ncei.noaa.gov/pub/data/daily-grids/v1-0-0/

Severe Weather Data Inventory, Storm Events Database: https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/

TOOLS:

xan, the CSV Magician

Guillaume Plique, Béatrice Mazoyer, Laura Miguel, César Pichon, Anna Charles, & Julien Pontoire. (2025). xan, the CSV magician. (0.50.0). Zenodo. https://doi.org/10.5281/zenodo.15310200