Coursea: Reproducible Research (Project 2)

01/05/2018 – Holger Speckter (Jerebai)

A. Brief Summary/ Synopsis

With this project we analyse the US National Oceanic & Atmospheric Administration's (NOAA) storm database to decide the effects of weather events on US population and economy. Impact on the fatalities, populous, and measured in injuries, was caused by similar weather event patterns, with Tornados inflicting the harshest toll. The economic impact, measured in crop and property damage, followed a very different pattern of weather events, with Floods causing the largest total damage.

B. Steps for Structure

Weather, e.g. storms and other severe weather events may cause both public health and economic problems and severe events may result in fatalities, injuries, and property damage. With aiming a prevention of such outcomes is a key concern and with this project explores the US National Oceanic & Atmospheric Administration's (NOAA) storm database. This database tracks characteristics of major storms and weather events in the United States, including when and where they occur, as well as estimates of any fatalities, injuries, and property damage.

We use R as for:
- a subset of the data was then manipulated to estimate e.g. the total economic value
- packages such as dplyr were used to select and rank the manipulated data
- ggplot2 was utilitized for all plots.

C. Data Processing

The data we use came in the form of a CSV compressed file. The Data file was provided via the coursera website as part of the course for reproducible research.

In the earlier years of the database fewer events recorded, most likely due to a lack of good records keeping. More recent years should be considered more complete.

D. Loading Data and Sub setting

The initial data set was loaded into R. The raw data set remained unchanged (Storm.Data.Raw), data manipulation was completed on the subset Storm.Data.

```r
library(dplyr)

library(tidyr)

library(ggplot2)

library(DT)
```

```
myDir <- "C:/Users/hspeckter/Documents/_R/Class5_ReproducibleResearch/Project2"
setwd(myDir)
download.file("https://www.coursera.org/learn/reproducible-research/
              peer/OMZ37/course-project-2/repdata-data-StormData.csv.bz2","data.csv.biz2")
dateDownloaded<- date()
## list.files()
## Read Data
Storm.Data.Raw<- read.csv("repdata-data-StormData.csv.bz2")

## Subsetting data and Raw Data remains untouched
Storm.Data<- select(Storm.Data.Raw,EVTYPE,FATALITIES,INJURIES,PROPDMG,PROPDMGEXP,   CROPDMG,CROPDMGEXP)
```

E. Calculating Damage Values (Property)

Once the levels of the exponent were identified, they were converted into numerical values. Total property damage in $ was simply calculated by multiplying DMG data with modfified EXP.

```
Event.Type.Labels<- sort(unique(Storm.Data$EVTYPE))
## Replacing Property Damage Exponent & calculating Property Damage.Value
Property.Damage.Exponent.Old<- as.character(sort(unique(Storm.Data$PROPDMGEXP)))

# Replacing Values
# Replacing Values
Storm.Data$PROPDMGEXP<- gsub(pattern = "?|-|+", replacement = "0" ,Storm.Data$PROPDMGEXP)
Storm.Data$PROPDMGEXP<- gsub("1"    , "10",Storm.Data$PROPDMGEXP)
Storm.Data$PROPDMGEXP<- gsub("H|h|2", "100" ,Storm.Data$PROPDMGEXP)
Storm.Data$PROPDMGEXP<- gsub("K|k|3", "1000" ,Storm.Data$PROPDMGEXP)
Storm.Data$PROPDMGEXP<- gsub("4"    , "10000" ,Storm.Data$PROPDMGEXP)
Storm.Data$PROPDMGEXP<- gsub("5"    , "100000" ,Storm.Data$PROPDMGEXP)
Storm.Data$PROPDMGEXP<- gsub("M|m|6", "1000000" ,Storm.Data$PROPDMGEXP)
Storm.Data$PROPDMGEXP<- gsub("7"    , "10000000" ,Storm.Data$PROPDMGEXP)
Storm.Data$PROPDMGEXP<- gsub("8"    , "100000000" ,Storm.Data$PROPDMGEXP)
Storm.Data$PROPDMGEXP<- gsub("B|b"  , "1000000000" ,Storm.Data$PROPDMGEXP)
Storm.Data$PROPDMGEXP.NUM<- as.numeric(as.character(Storm.Data$PROPDMGEXP))


# Check to make sure everything was replaced
Property.Damage.Exponent.New<- sort(unique(Storm.Data$PROPDMGEXP.NUM))
# Calculating the Property.Damage.Value
Storm.Data$PropValue<- Storm.Data$PROPDMG * Storm.Data$PROPDMGEXP.NUM
```

F. Calculating Damage Values (Crop)

The following process outlined above was applied to the conversion of the property exponent values as well.

```
Crop.Damage.Exponent.Old<- as.character(sort(unique(Storm.Data$CROPDMGEXP))
)


Storm.Data$CROPDMGEXP<- gsub("?", "0" ,Storm.Data$CROPDMGEXP)
Storm.Data$CROPDMGEXP<- gsub("H|h|2", "100" ,Storm.Data$CROPDMGEXP)
Storm.Data$CROPDMGEXP<- gsub("K|k|3", "1000" ,Storm.Data$CROPDMGEXP)
Storm.Data$CROPDMGEXP<- gsub("M|m"  , "1000000" ,Storm.Data$CROPDMGEXP)
Storm.Data$CROPDMGEXP<- gsub("B|b"  , "1000000000" ,Storm.Data$CROPDMGEXP)


Storm.Data$CropExpNumeric<- as.numeric(as.character(Storm.Data$CROPDMGEXP))


# Check to make sure everything was replaced
Crop.Damage.Exponent.New<- sort(unique(Storm.Data$CropExpNumeric))
# Calculating the Property.Damage.Value
Storm.Data$CropDamaValue<- Storm.Data$CROPDMG * Storm.Data$CropExpNumeric
```

G. Summarizing: Fatality Data by Weather Event Type

Coursera Question: Across the United States, which types of events are more harmful with respect to population healthy?

In order to answer question 1 of the assignment, dplyr was used to group, summarize, and rank the data in the df Sum.Fatalities. The data had to be ordered in descending order using the order comand, so that the data can be displayed in descending order in ggplot.

```
## Summarizing Fatalities
Sum.Fatalities<- select(Storm.Data,EVTYPE,FATALITIES)
Sum.Fatalities<- Sum.Fatalities %>%
                        group_by(EVTYPE) %>%
                        summarise(FATALITIES = sum(FATALITIES)) %>%
                        top_n(n=10) %>%
                        arrange(desc(FATALITIES))
## Sorting factor EVTYPE by Fatalities for ggplot
Sum.Fatalities$EVTYPE <- factor(Sum.Fatalities$EVTYPE, levels = Sum.Fatalities$EVTYPE[order(-Sum.Fatalities$FATALITIES)])
datatable(Sum.Fatalities, caption = 'Table 1: Top 10: Weather Events causing fatalities',options = list(pageLength = 5))
```

Table 1: Top 10: Weather Events causing fatalities

	EVTYPE	
1	TORNADO	
2	EXCESSIVE HEAT	
3	FLASH FLOOD	
4	HEAT	
5	LIGHTNING	

H. Summarizing Injury Data by Weather Event Type

Dplyr was also used to group, summarize, and rank the data in the df Sum.Injuries. Again, The data had to be ordered in descending order using the order comand, so that the data can be displayed in descending order in ggplot.

```
## Summarizing Injuries
Sum.Injuries<- select(Storm.Data,EVTYPE,INJURIES)
Sum.Injuries<- Sum.Injuries %>%
      group_by(EVTYPE) %>%
      summarise(INJURIES = sum(INJURIES)) %>%
      top_n(n=10) %>%
      arrange(desc(INJURIES))

## Sorting factor EVTYPE by Injuries for ggplot
```

```
Sum.Injuries$EVTYPE <- factor(Sum.Injuries$EVTYPE, levels = Sum.Injuries$EV
TYPE[order(-Sum.Injuries$INJURIES)])
datatable(Sum.Injuries, caption = 'Table 2: Top 10: Weather Events causing
Injuries',options = list(pageLength = 5))
```

Table 2: Top 10: Weather Events causing Injuries

	EVTYPE	
1	TORNADO	
2	TSTM WIND	
3	FLOOD	
4	EXCESSIVE HEAT	
5	LIGHTNING	

I. Modifying data for Plot 3

Second question: Across the United States, which types of events have the greatest economic consequences?

Steps:
- Total Damage, the summation of property and crop damage, was calculated
- The data is ranked bythe top 15 data points selected using dplyr
- Again the data had to be ordered using the order comand.

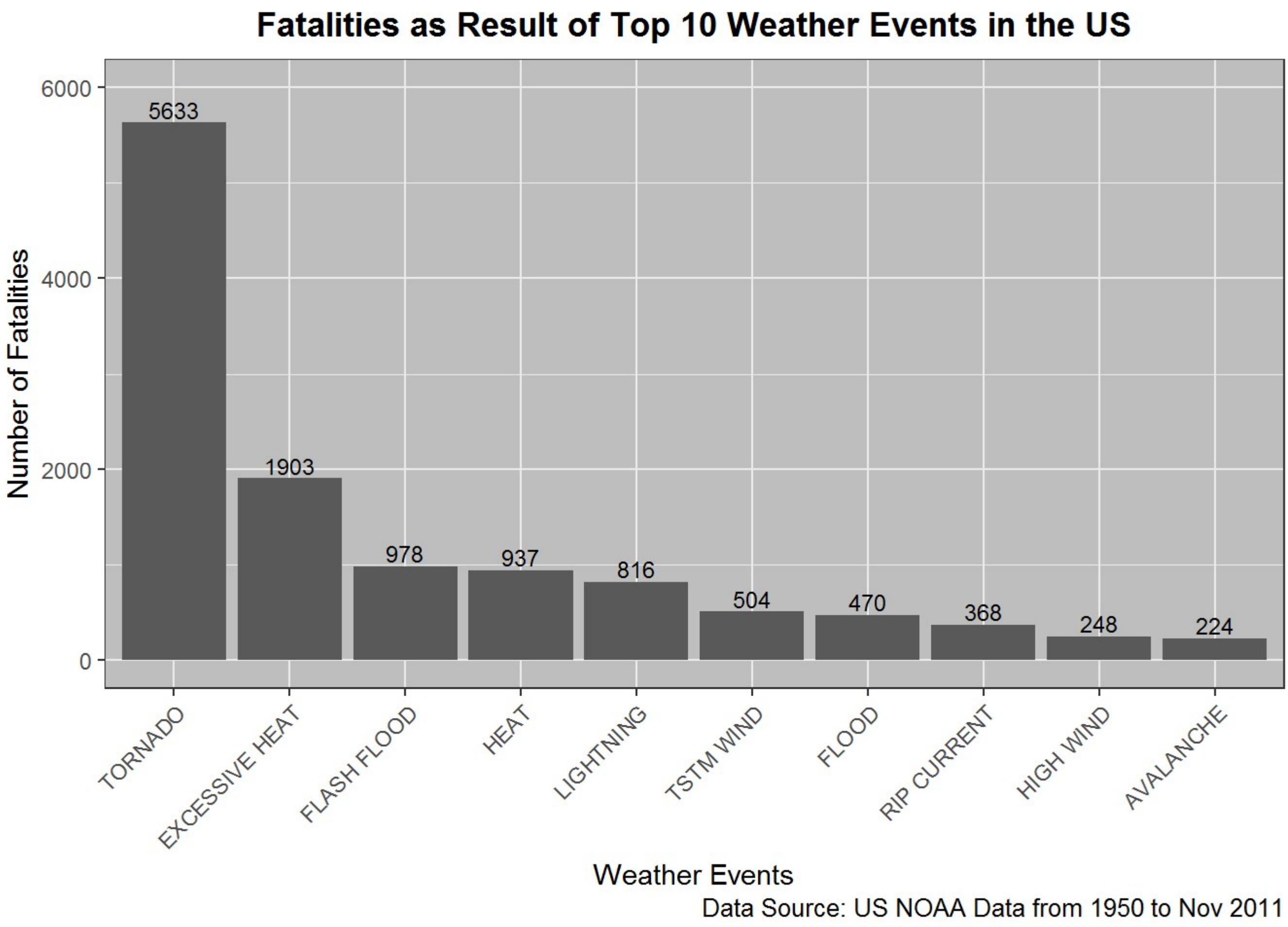
```
Sum.Damage<- select(Storm.Data,EVTYPE,PropValue, CropDamaValue)
Sum.Damage<- mutate(Sum.Damage, TotalDamage = PropValue + CropDamaValue)

Sum.Damage<- Sum.Damage %>%
      group_by(EVTYPE) %>%
      summarise_each(funs(sum)) %>%
      top_n(n=15) %>%
      arrange(desc(TotalDamage))

Sum.Damage$EVTYPE <- factor(Sum.Damage$EVTYPE,levels = Sum.Damage$EVTYPE[or
der(-Sum.Damage$TotalDamage)])
## converting data (into long data format) for stacked bar plot in ggplot
Q2<- select(Sum.Damage,-TotalDamage) %>%
            gather(variable, value, -EVTYPE)


datatable(Sum.Damage, caption = 'Table 3: Top 10: Weather Events causing la
rgest economic impact',options = list(pageLength = 5))
```

Table 3: Top 10: Weather Events causing largest economic impact

	EVTYPE	PropValue	CropDamaValue
1	FLOOD	1446577098000	56619684500
2	HURRICANE/TYPHOON	693058400000	26078728000
3	STORM SURGE	433235360000	50000
4	DROUGHT	10461060000	139725660000
5	HURRICANE	118683190100	27419100000

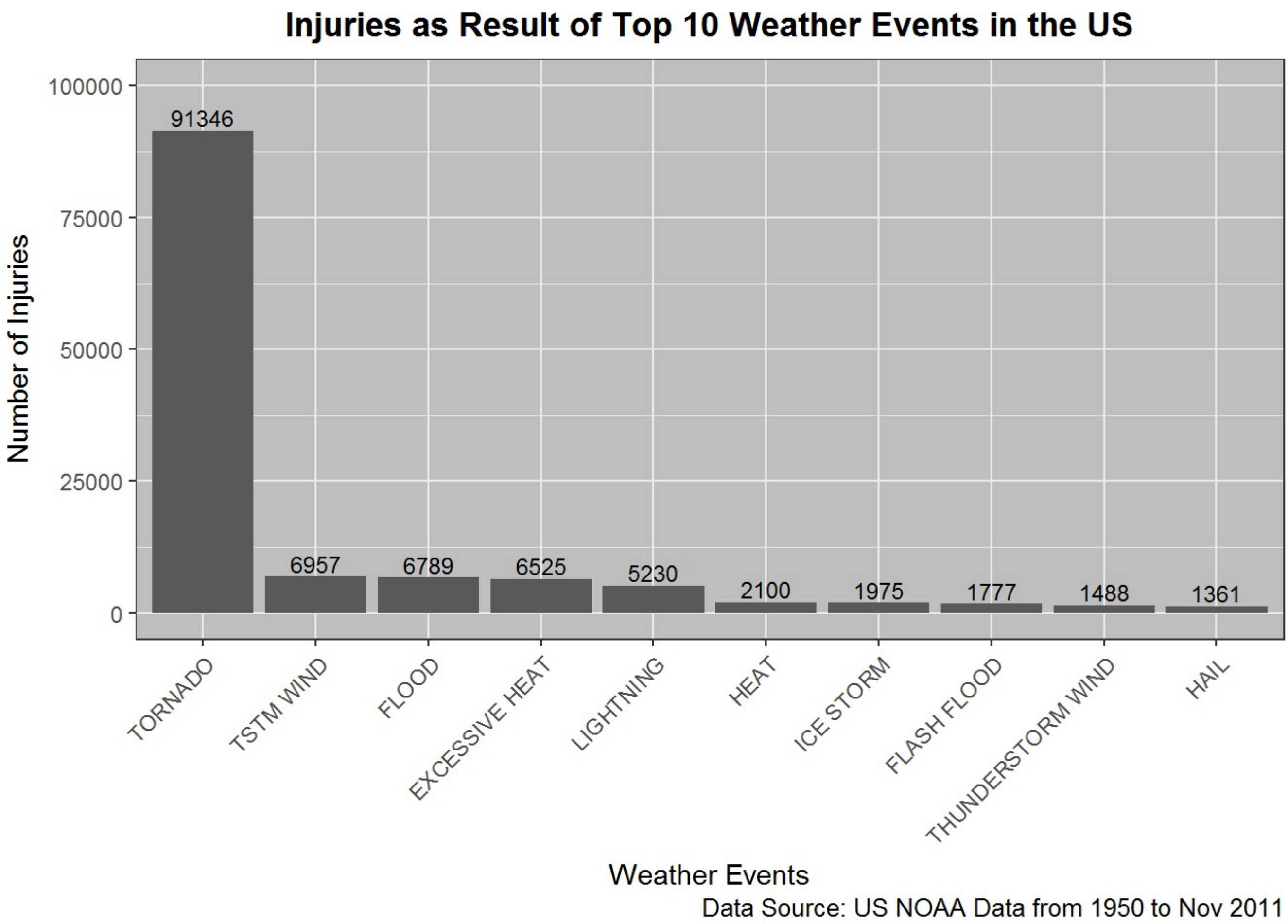
J. Fatalities by Weather Event Type

By plotting Fatalities by Weather Event in descending order clearly highlights the top 10 weather events causing the most significant human toll over the past 60+ years.



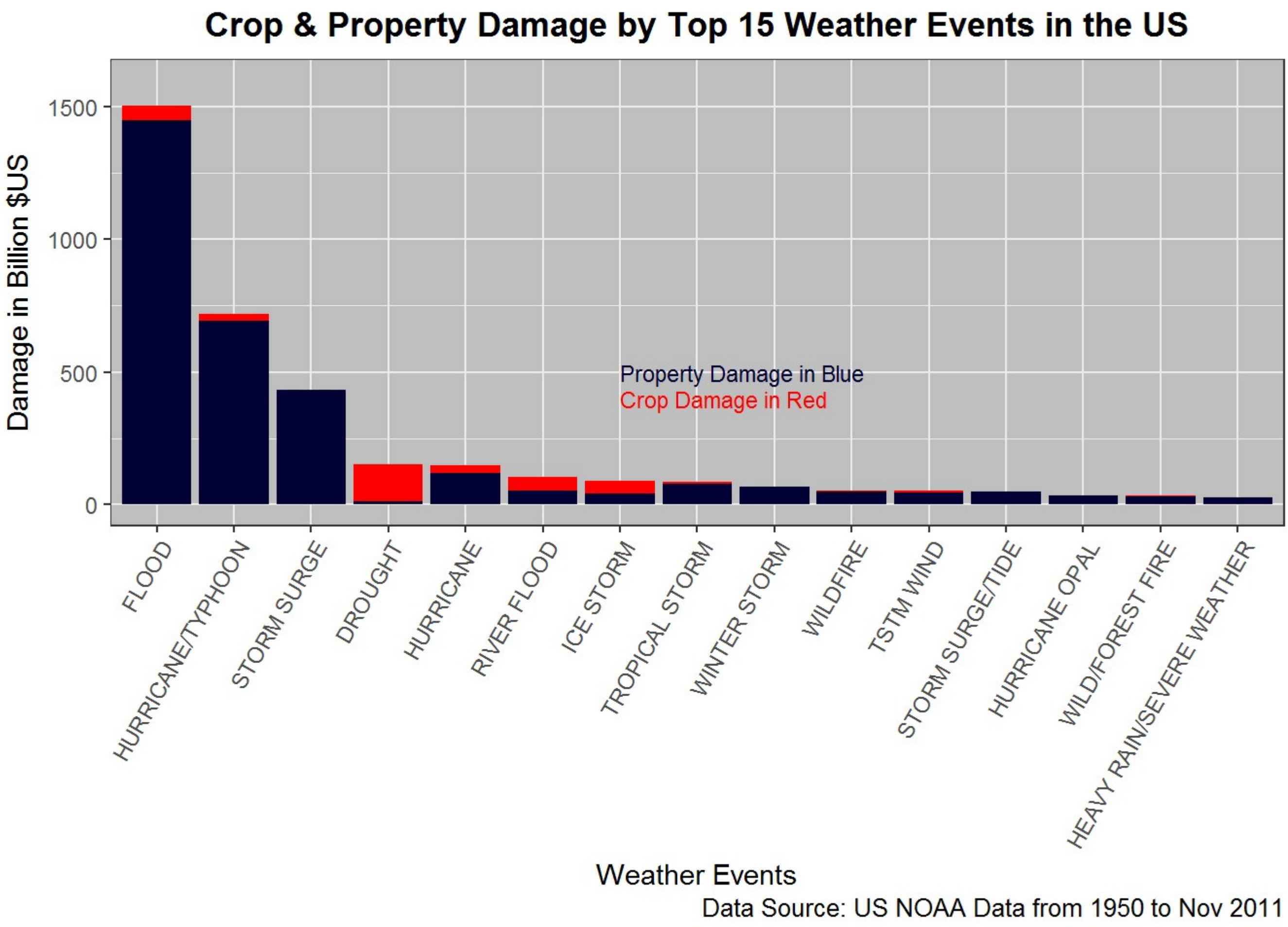
K. The Injuries by Weather Event Type

We will plotting now Injuries by Weather Event in descending order clearly highlights the top 10 weather events causing the most significant number of injuries inflicted on the US population over the past 60+ years.



L. Crop & Property Damage by Weather Event Type

We now, plot the data in ggplot, damage values were displayed in billion US dollars. As one can see, total economic damage followed a very different weather event pattern as in the earlier analysis. The Tornados may be the deadliest weather event, yet Floods are responsible for the largest economic damage.



X. Appendix

This Appendix contains the sample code used to generate the chars.

Code for first Plot

```
## Plotting Fatalities

p1<- ggplot(Sum.Fatalities, aes(x = EVTYPE, y = FATALITIES)) + geom_bar(sta
t = "identity") +

      labs(x="Weather Events", y = "Number of Fatalities",

           title="Fatalities as Result of Top 10 Weather Events in the US",

           caption = "Data Source: US NOAA Data from 1950 to Nov 2011") +

      ylim(0, 6000)

p1<- p1 + geom_text(aes(label=FATALITIES), position=position_dodge(width=0.
9), vjust=-0.25,size = 3)


p1<- p1 + theme_bw() + theme(panel.background = element_rect(fill = 'grey')
)

p1<- p1 + theme(legend.position="none")
```

```
p1<- p1 + theme(plot.title = element_text(lineheight=.8, face="bold", hjust
= 0.5)) +
      theme(plot.subtitle = element_text(lineheight=.8,hjust = 0.5)) +
      theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Code for second Plot

```
## Plotting Injuries
p2<- ggplot(Sum.Injuries, aes(x = EVTYPE, y = INJURIES)) + geom_bar(stat =
"identity") +
      labs(x="Weather Events", y = "Number of Injuries",
           title="Injuries as Result of Top 10 Weather Events in the US",
           caption = "Data Source: US NOAA Data from 1950 to Nov 2011") +
      ylim(0, 100000)
p2<- p2 + geom_text(aes(label=INJURIES), position=position_dodge(width=0.9)
, vjust=-0.25,size = 3)


p2<- p2 + theme_bw() + theme(panel.background = element_rect(fill = 'grey')
)
p2<- p2 + theme(legend.position="none")


p2<- p2 + theme(plot.title = element_text(lineheight=.8, face="bold", hjust
= 0.5)) +
      theme(plot.subtitle = element_text(lineheight=.8,hjust = 0.5)) +
      theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Code for third Plot

```
p3<- ggplot(Q2, aes(x = EVTYPE, y = value/1000000000, fill = variable)) +
      geom_bar(stat = "identity") +
      labs(x="Weather Events", y = "Damage in Billion $US",
           title="Crop & Property Damage by Top 15 Weather Events in the US
",
           caption = "Data Source: US NOAA Data from 1950 to Nov 2011") +
           ylim(0, 160)

p3<- p3 + theme_bw() + theme(panel.background = element_rect(fill = 'grey')
)
p3<- p3 + theme(legend.position="none") +
      scale_fill_manual(values=c("red", "#000033"))
```

```r
p3<- p3 + theme(plot.title = element_text(lineheight=.8, face="bold", hjust = 0.5)) +
      theme(plot.subtitle = element_text(lineheight=.8,hjust = 0.5)) +
      theme(axis.text.x = element_text(angle = 60, hjust = 1))
p3<- p3 + annotate("text", x = 7, y = 50, size = 3,
                   label = "Property Damage in Blue", color= "#000033",hjust = 0 )
p3<- p3 + annotate("text", x = 7, y = 40, size = 3,
                   label = "Crop Damage in Red", color= "red",hjust = 0 )
```