Reproducible Research: Peer Assessment 2

Title: PA2 NOAA Storm Event Data Analysis

Course: JHU-DSCI Reproducible Research

Assignment: PA2 NOAA Storm Data

Name: Walter Yu

Date: July 2020

Part 1: Synopsis

This document is a reproducible analysis of the NOAA Storm Data set as follows:

1. Download and read the raw NOAA storm dataset
2. Review and process data before analysis
3. Analyze data to answer assignment questions
4. Plot data to communicate results
5. Document steps and findings to be reproducible
6. Assignment is organized into data processing, analysis and results
7. Each section is further organized by step or question being answered

Submission Notes

1. This markdown file is an analysis and visualization of the NOAA weather dataset. Source data available here (https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2).
2. This markdown project template is based on a fork of the course assignment Github repo available here (https://github.com/rdpeng/RepData_PeerAssessment1).
3. This assignment is completed for the JHU Coursera Data Science Program, which is a 10-course certification. More info about this program is available here (https://www.coursera.org/specializations/jhu-data-science).
4. All links and code sources referenced for this assignment are cited within each section so please refer to them accordingly.
5. Assignment was initially generated with R Studio Cloud (https://rstudio.cloud) due to package installation errors on my local machine; as a result, file was run and report rendered there instead.

Assignment Questions

1. Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?

2. Across the United States, which types of events have the greatest economic consequences?

Part 2: Data Processing

Overview:

1. Raw CSV data was read in from zip file
2. Data was reviewed and processed prior to analysis
3. Data was subset and aggregated during analysis

Notes:

1. Data processing is a large part of this assignment, so additional effort was made prior to analysis
2. The course discussion forum was reviewed, and this post (https://www.coursera.org/learn/reproducible-research/discussions/weeks/4/threads/38y35MMiEeiERhLphT2-QA) was referenced for approach
3. In general, duplicate event types/records and total damage calculations needed to be addressed

Part 2A: Data Import

Question 1 - Analyze Event Type by Impacts to Population Health

Steps:

1. Initially attempted to load data from url
2. However, attempts resulted in error so loaded from csv.bz2 zip file instead
3. As a result, zip file was downloaded and read into program
4. Data import step is saved as its own chunk and cached (https://bookdown.org/yihui/rmarkdown-cookbook/cache.html)

Analysis:

1. Initially used na.omit function to remove null values
2. However, na.omit removed most values do not used on dataset
3. As a result, data processing completed in following sections

```
# part 2a: data import - read from csv.bz2 format

# note: course discussion forum consulted for these steps:
# https://www.coursera.org/learn/reproducible-research/discussions/weeks/4/threads/38y35MMiEeiERhLphT2-QA

# note: read url unsuccessful after several attempts, so revert to reading from bz2 file
# source: https://rpubs.com/otienodominic/398952
# url <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
# temp <- tempfile()
# download.file(url, temp, mode="wb")
# data <- read.csv("repdata_data_StormData.csv.bz2", header=TRUE, sep=",")
# unlink(zipfile)

# source: assignment instructions
# https://www.coursera.org/learn/reproducible-research/discussions/weeks/4/threads/38y35MMiEeiERhLphT2-QA
data <- read.csv("data/repdata_data_StormData.csv.bz2")
```

```
# part 2a: data import - review data

# review dataset
# conclusion: dataset contains many columns, only some of which seem to be useful
# https://www.statmethods.net/stats/descriptives.html
# head(data, n=5L)
# names(data)
# summary(data)

# attempt remove null values
# source: https://www.statmethods.net/input/missingdata.html
# data_omit <- na.omit(data)

# review results
# conclusion: na.omit removes most records, so not used on dataset
# source: https://www.statmethods.net/stats/descriptives.html
# head(data_omit, n=5L)
# names(data_omit)
# summary(data_omit)

# identify event types
# source: https://www.statmethods.net/stats/descriptives.html
# conclusion: results show a large number of event types, some duplicated
# unique(data$EVTYPE)
```

Part 2: Data Processing

Part 2B: Data Analysis

Question 1 - Analyze Event Type by Impacts to Population Health

Steps:

1. Subset and analyze data for each question/relevant data
2. For question 1, data subset/analyzed for fatalities/injuries
3. For question 2, data subset/analyzed for property/crop damage

Analysis:

1. There are 37 variables and ~900k records, only some of which are needed for analysis
2. NOAA documentation quantifies health impacts with fatalities/injuries
3. As a result, aggregate fatalities/injuries by event type

Observations:

1. Aggregation showed that storm events with highest impacts make up large majority of impacts
2. In addition, top 5-10 storm events made up good portion of population impacts
3. As a result, those events were used to develop results

```
# part 2b: data analysis - analyze event types by fatalities

# aggregate by fatality or injury, then apply calculations
# source: https://www.statmethods.net/management/aggregate.html
event_fatal_total <- aggregate(FATALITIES ~ EVTYPE, data, sum)

# rank aggregate results:
# https://stackoverflow.com/questions/23659241/rank-in-the-aggregate-function
event_fatal_total <- event_fatal_total[order(event_fatal_total$FATALITIES, decreasing=TRUE),]

# subset for top 50 event types with highest count
# https://stackoverflow.com/questions/2667673/select-first-4-rows-of-a-data-frame-in-r/47400307
event_fatal_50 <- event_fatal_total[1:50,]
# event_fatal_50

# ratio of fatalities due to top 50 event types / all event types
# conclusion: top events make up large majority of impacts
event_fatal_50_sum <- sum(event_fatal_50$FATALITIES)
event_fatal_total_sum <- sum(event_fatal_total$FATALITIES)
event_fatal_ratio <- event_fatal_50_sum / event_fatal_total_sum
print("ratio of fatalities from top 50 / all event types:")
```

```
## [1] "ratio of fatalities from top 50 / all event types:"
```

```
print(event_fatal_ratio)
```

```
## [1] 0.9757016
```

```
# note: based on results above, storm-related and heat have highest totals
# conclusion: calculate ratios for these events to create plots
event_fatal_8 <- event_fatal_total[1:8,]
event_fatal_8
```

```
##             EVTYPE FATALITIES
## 834        TORNADO       5633
## 130 EXCESSIVE HEAT       1903
## 153    FLASH FLOOD        978
## 275           HEAT        937
## 464      LIGHTNING        816
## 856      TSTM WIND        504
## 170          FLOOD        470
## 585    RIP CURRENT        368
```

Part 2: Data Processing

Part 2C: Data Analysis

Question 1 - Analyze Event Type by Impacts to Population Health

Steps:

1. Subset and analyze data for each question/relevant data
2. For question 1, data subset/analyzed for fatalities/injuries
3. For question 2, data subset/analyzed for property/crop damage

Analysis:

1. There are 37 variables and ~900k records, only some of which are needed for analysis
2. NOAA documentation quantifies health impacts with fatalities/injuries
3. As a result, aggregate fatalities/injuries by event type

Observations:

1. Aggregation showed that storm events with highest impacts make up large majority of impacts
2. In addition, top 5-10 storm events made up good portion of population impacts
3. As a result, those events were used to develop results

```
# part 2c: data analysis - analyze event types by injuries

# aggregate by fatality or injury, then apply calculations
# source: https://www.statmethods.net/management/aggregate.html
event_injury_total <- aggregate(INJURIES ~ EVTYPE, data, sum)

# rank aggregate results:
# https://stackoverflow.com/questions/23659241/rank-in-the-aggregate-function
event_injury_total <- event_injury_total[order(event_injury_total$INJURIES, decreasing=TRUE),]

# subset for top 50 event types with highest count
# https://stackoverflow.com/questions/2667673/select-first-4-rows-of-a-data-frame-in-r/47400307
event_injury_50 <- event_injury_total[1:50,]
# event_injury_50

# ratio of injuryities due to top 50 event types / all event types
event_injury_50_sum <- sum(event_injury_50$INJURIES)
event_injury_total_sum <- sum(event_injury_total$INJURIES)
event_injury_ratio <- event_injury_50_sum / event_injury_total_sum
print("ratio of injuries from top 50 / all event types:")
```

```
## [1] "ratio of injuries from top 50 / all event types:"
```

```
print(event_injury_ratio)
```

```
## [1] 0.9931829
```

```
# note: based on results above, storm-related and heat have highest totals
# conclusion: calculate ratios for these events to create plots
event_injury_6 <- event_injury_total[1:6,]
event_injury_6
```

```
##             EVTYPE INJURIES
## 834        TORNADO    91346
## 856      TSTM WIND     6957
## 170          FLOOD     6789
## 130 EXCESSIVE HEAT     6525
## 464      LIGHTNING     5230
## 275           HEAT     2100
```

Part 3: Results

Part 3A: Plot Results

Question 1 - Plot Event Type by Impacts to Population Health

Overview:

1. Data processing produced subsets of aggregated, ranked data by population impact
2. Each subset was reviewed, then visualized with bar plots
3. Results show which events had the most impact on population health
4. Since fatalities have a larger impact, then it was the factor used for results
5. Injury plot created but not presented; assignment limited to 3 plots total

Steps:

1. Aggregated, subset data was ranked, then plot into bar plot
2. Plot is color coded by event type and axis labeled

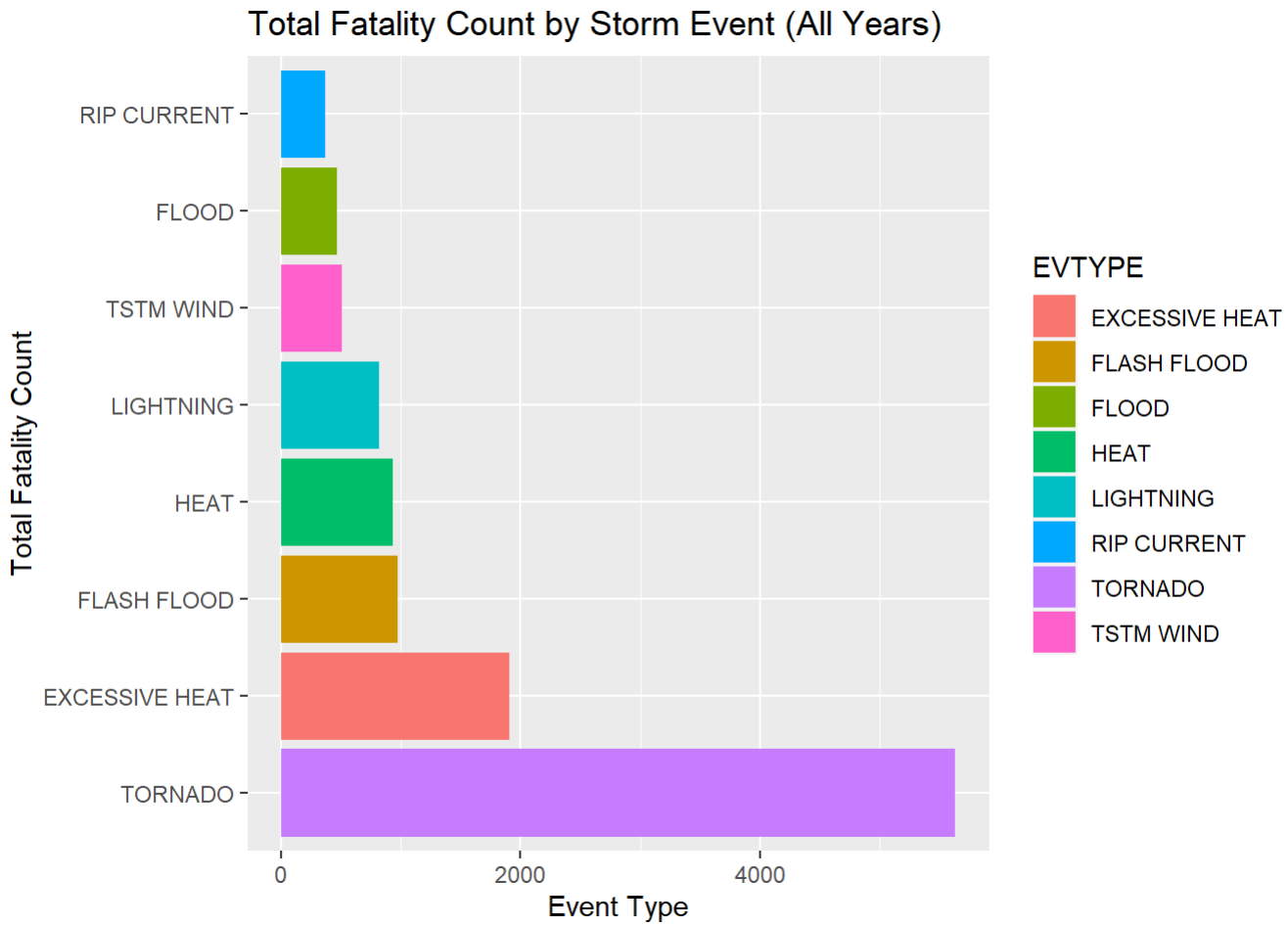
Observations:

1. Fatalities are assumed to have large impact than injuries alone
2. As a result, tornados had the largest impact followed by thunder storm winds, floods and high heat
3. Based on these results, storm-related events and heat had the largest impact to population health

```
# part 3a: data processing - plot event types by fatalities
# http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visua
lization
# https://stackoverflow.com/questions/16961921/plot-data-in-descending-order-as-appears-in-data-
frame
# install.packages("ggplot2")
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
p_fatal <- ggplot(data=event_fatal_8, aes(x=reorder(EVTYPE, -FATALITIES), y=FATALITIES, fill=EVT
YPE)) +
    geom_bar(stat="identity") +
    labs(title="Total Fatality Count by Storm Event (All Years)", x="Total Fatality Count", y="E
vent Type")
p_fatal + coord_flip()
```




```
# part 3a: data processing - plot event types by injuries
# http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization
# https://stackoverflow.com/questions/16961921/plot-data-in-descending-order-as-appears-in-data-frame
# install.packages("ggplot2")
# library(ggplot2)
# p_injury <- ggplot(data=event_injury_6, aes(x=reorder(EVTYPE, -INJURIES), y=INJURIES, fill=EVTYPE)) +
#     geom_bar(stat="identity") +
#     labs(title="Total Injury Count by Storm Event (All Years)", x="Total Injury Count", y="Event Type")
# p_injury + coord_flip()
```

Part 2: Data Processing

Overview:

1. Property/crop data differs from fatality/injury due to total damage ($)
2. Units are stored as exponents in the PROPDMGEXP and CROPDMGEXP variables
3. As a result, conversions were made to total dollars prior to analysis

Part 2C: Data Analysis

Question 2 - Analyze Event Type by Impacts to Economic Consequences

Steps:

1. Subset and analyze data for each question/relevant data
2. For question 1, data subset/analyzed for fatalities/injuries
3. For question 2, data subset/analyzed for property/crop damage
4. For question 2, total damage was converted to total dollars
5. Once total damaged was converted, then it was aggregated/totaled

Analysis:

1. There are 37 variables, only some of which are needed for analysis
2. NOAA documentation quantifies economic impacts with property/crop damage
3. As a result, aggregate property/crop damage by event type

Observations:

1. Aggregation showed that storm events with highest impacts make up large majority of impacts
2. In addition, top 5-10 storm events made up good portion of economic impacts
3. As a result, those events were used to develop results

```
# part 2c: data analysis - analyze event types by economic impact
# note: analyze property damage per course discussion post
# https://www.coursera.org/learn/reproducible-research/discussions/weeks/4/threads/38y35MMiEeiERhLphT2-QA
names(data)
```

```
##  [1] "STATE__"    "BGN_DATE"   "BGN_TIME"   "TIME_ZONE"  "COUNTY"
##  [6] "COUNTYNAME" "STATE"      "EVTYPE"     "BGN_RANGE"  "BGN_AZI"
## [11] "BGN_LOCATI" "END_DATE"   "END_TIME"   "COUNTY_END" "COUNTYENDN"
## [16] "END_RANGE"  "END_AZI"    "END_LOCATI" "LENGTH"     "WIDTH"
## [21] "F"          "MAG"        "FATALITIES" "INJURIES"   "PROPDMG"
## [26] "PROPDMGEXP" "CROPDMG"    "CROPDMGEXP" "WFO"        "STATEOFFIC"
## [31] "ZONENAMES"  "LATITUDE"   "LONGITUDE"  "LATITUDE_E" "LONGITUDE_"
## [36] "REMARKS"    "REFNUM"
```

```
summary(data$PROPDMG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   12.06    0.50 5000.00
```

```
summary(data$PROPDMGEXP)
```

```
##             -      ?      +      0      1      2      3      4      5
## 465934      1      8      5    216     25     13      4      4     28
##      6      7      8      B      h      H      K      m      M
##      4      5      1     40      1      6 424665      7  11330
```

```
summary(data$CROPDMG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.527   0.000 990.000
```

```
summary(data$CROPDMGEXP)
```

```
##             ?      0      2      B      k      K      m      M
## 618413      7     19      1      9     21 281832      1   1994
```

Part 2: Data Processing

Overview:

1. Property/crop data differs from fatality/injury due to total damage ($)
2. Units are stored as exponents in the PROPDMGEXP and CROPDMGEXP variables
3. As a result, conversions were made to total dollars prior to analysis
4. For question 2, total damage was converted to total dollars
5. Once total damaged was converted, then it was aggregated/totaled

Part 2D: Data Analysis

Question 2 - Analyze Event Type by Impacts to Economic Consequences

Steps:

1. Subset and analyze data for each question/relevant data
2. For question 1, data subset/analyzed for fatalities/injuries
3. For question 2, data subset/analyzed for property/crop damage
4. For question 2, total damage was converted to total dollars
5. Once total damaged was converted, then it was aggregated/totaled

Analysis:

1. There are 37 variables, only some of which are needed for analysis
2. NOAA documentation quantifies economic impacts with property/crop damage
3. As a result, aggregate property/crop damage by event type

Observations:

1. Aggregation showed that storm events with highest impacts make up large majority of impacts
2. In addition, top 5-10 storm events made up good portion of economic impacts
3. As a result, those events were used to develop results

```
# part 2d: data processing - analyze event types by property damage exp.

# *** note: sources for code below ***
# subset and verify results:
# https://stats.idre.ucla.edu/r/faq/frequently-asked-questions-about-rhow-can-i-subset-a-data-se
tthe-r-program-as-a-text-file-for-all-the-code-on-this-page-subsetting-is-a-very-important-compo
nent/
# https://stackoverflow.com/questions/15016723/how-to-add-column-into-a-dataframe-based-on-condi
tion
# convert powers of 10:
# https://en.wikipedia.org/wiki/Exponentiation#Powers_of_ten

# note: please refer to cited code sources above
prop_0e_subset <- subset(data, PROPDMGEXP == 1, select = c(EVTYPE, PROPDMG, PROPDMGEXP))
# prop_0e_subset
prop_0e_subset$PROPDMGTOT <- with(prop_0e_subset, ifelse(PROPDMGEXP == 1, PROPDMG*0, "NA"))
# prop_0e_subset

# note: please refer to cited code sources above
prop_1e_subset <- subset(data, PROPDMGEXP == 0, select = c(EVTYPE, PROPDMG, PROPDMGEXP))
# prop_1e_subset
prop_1e_subset$PROPDMGTOT <- with(prop_1e_subset, ifelse(PROPDMGEXP == 0, PROPDMG*10, "NA"))
# prop_1e_subset

# note: please refer to cited code sources above
prop_2e_subset <- subset(data, PROPDMGEXP == 2, select = c(EVTYPE, PROPDMG, PROPDMGEXP))
# prop_2e_subset
prop_2e_subset$PROPDMGTOT <- with(prop_2e_subset, ifelse(PROPDMGEXP == 2, PROPDMG*10^2, "NA"))
# prop_2e_subset

# note: please refer to cited code sources above
prop_3e_subset <- subset(data, PROPDMGEXP == 3 | PROPDMGEXP == "K", select = c(EVTYPE, PROPDMG,
 PROPDMGEXP))
# prop_3e_subset
prop_3e_subset$PROPDMGTOT <- with(prop_3e_subset, ifelse(PROPDMGEXP == 3 | PROPDMGEXP == "K", PR
OPDMG*10^3, "NA"))
# prop_3e_subset

# note: please refer to cited code sources above
prop_4e_subset <- subset(data, PROPDMGEXP == 4, select = c(EVTYPE, PROPDMG, PROPDMGEXP))
# prop_4e_subset
prop_4e_subset$PROPDMGTOT <- with(prop_4e_subset, ifelse(PROPDMGEXP == 4, PROPDMG*10^4, "NA"))
# prop_4e_subset

# note: please refer to cited code sources above
prop_5e_subset <- subset(data, PROPDMGEXP == 5, select = c(EVTYPE, PROPDMG, PROPDMGEXP))
# prop_5e_subset
prop_5e_subset$PROPDMGTOT <- with(prop_5e_subset, ifelse(PROPDMGEXP == 5, PROPDMG*10^5, "NA"))
# prop_5e_subset

# note: please refer to cited code sources above
prop_6e_subset <- subset(data, PROPDMGEXP == 6 | PROPDMGEXP == "M", select = c(EVTYPE, PROPDMG,
 PROPDMGEXP))
```

```
# prop_6e_subset
prop_6e_subset$PROPDMGTOT <- with(prop_6e_subset, ifelse(PROPDMGEXP == 6 | PROPDMGEXP == "M", PR
OPDMG*10^6, "NA"))
# prop_6e_subset

# note: please refer to cited code sources above
prop_7e_subset <- subset(data, PROPDMGEXP == 7, select = c(EVTYPE, PROPDMG, PROPDMGEXP))
# prop_7e_subset
prop_7e_subset$PROPDMGTOT <- with(prop_7e_subset, ifelse(PROPDMGEXP == 7, PROPDMG*10^7, "NA"))
# prop_7e_subset

# note: please refer to cited code sources above
prop_8e_subset <- subset(data, PROPDMGEXP == 8, select = c(EVTYPE, PROPDMG, PROPDMGEXP))
# prop_8e_subset
prop_8e_subset$PROPDMGTOT <- with(prop_8e_subset, ifelse(PROPDMGEXP == 8, PROPDMG*10^8, "NA"))
# prop_8e_subset

# note: please refer to cited code sources above
prop_9e_subset <- subset(data, PROPDMGEXP == "B", select = c(EVTYPE, PROPDMG, PROPDMGEXP))
# prop_9e_subset
prop_9e_subset$PROPDMGTOT <- with(prop_9e_subset, ifelse(PROPDMGEXP == "B", PROPDMG*10^9, "NA"))
# prop_9e_subset

propdmgtot_subset <- rbind(prop_0e_subset, prop_1e_subset, prop_2e_subset, prop_3e_subset, prop_
4e_subset, prop_5e_subset, prop_6e_subset, prop_7e_subset, prop_8e_subset, prop_9e_subset)
# propdmgtot_subset

# aggregate by cropdmg or propdmg, then apply calculations
# source: https://www.statmethods.net/management/aggregate.html
# event_propdmg_total <- aggregate(PROPDMGTOT ~ EVTYPE, propdmgtot_subset, FUN=length)
event_propdmg_total <- aggregate(PROPDMGTOT ~ EVTYPE, propdmgtot_subset, sum)

# rank aggregate results:
# https://stackoverflow.com/questions/23659241/rank-in-the-aggregate-function
event_propdmg_total <- event_propdmg_total[order(event_propdmg_total$PROPDMGTOT, decreasing=TRUE
),]

# subset for top 50 event types with highest count
# https://stackoverflow.com/questions/2667673/select-first-4-rows-of-a-data-frame-in-r/47400307
event_propdmg_10 <- event_propdmg_total[1:10,]
event_propdmg_10
```

```
##                EVTYPE   PROPDMGTOT
## 62              FLOOD 144657709800
## 179 HURRICANE/TYPHOON  69305840000
## 332           TORNADO  56935881815
## 281       STORM SURGE  43323536000
## 50        FLASH FLOOD  16822676125
## 103              HAIL  15730369077
## 171         HURRICANE  11868319010
## 340    TROPICAL STORM   7703890550
## 399      WINTER STORM   6688497260
## 156         HIGH WIND   5270046260
```

```
# ratio of propdmgities due to top 10 event types / all event types
event_propdmg_10_sum <- sum(event_propdmg_10$PROPDMGTOT, na.rm=TRUE)
event_propdmg_total_sum <- sum(event_propdmg_total$PROPDMGTOT, na.rm=TRUE)
event_propdmg_ratio <- event_propdmg_10_sum / event_propdmg_total_sum
print("ratio of property damage from top 10 / all event types:")
```

```
## [1] "ratio of property damage from top 10 / all event types:"
```

```
print(event_propdmg_ratio)
```

```
## [1] 0.8835103
```

Part 2: Data Processing

Overview:

1. Property/crop data differs from fatality/injury due to total damage ($)
2. Units are stored as exponents in the PROPDMGEXP and CROPDMGEXP variables
3. As a result, conversions were made to total dollars prior to analysis
4. For question 2, total damage was converted to total dollars
5. Once total damaged was converted, then it was aggregated/totaled

Part 2E: Data Analysis

Question 2 - Analyze Event Type by Impacts to Economic Consequences

Steps:

1. Subset and analyze data for each question/relevant data
2. For question 1, data subset/analyzed for fatalities/injuries
3. For question 2, data subset/analyzed for property/crop damage
4. For question 2, total damage was converted to total dollars
5. Once total damaged was converted, then it was aggregated/totaled

Analysis:

1. There are 37 variables, only some of which are needed for analysis
2. NOAA documentation quantifies economic impacts with property/crop damage
3. As a result, aggregate property/crop damage by event type

Observations:

1. Aggregation showed that storm events with highest impacts make up large majority of impacts
2. In addition, top 5-10 storm events made up good portion of economic impacts
3. As a result, those events were used to develop results

```
# part 2e: data processing - analyze event types by crop damage

# *** note: sources for code below ***
# subset and verify results:
# https://stats.idre.ucla.edu/r/faq/frequently-asked-questions-about-rhow-can-i-subset-a-data-se
tthe-r-program-as-a-text-file-for-all-the-code-on-this-page-subsetting-is-a-very-important-compo
nent/
# https://stackoverflow.com/questions/15016723/how-to-add-column-into-a-dataframe-based-on-condi
tion
# convert powers of 10:
# https://en.wikipedia.org/wiki/Exponentiation#Powers_of_ten

# note: please refer to cited code sources above
crop_1e_subset <- subset(data, CROPDMGEXP == 0, select = c(EVTYPE, CROPDMG, CROPDMGEXP))
# crop_1e_subset
crop_1e_subset$CROPDMGTOT <- with(crop_1e_subset, ifelse(CROPDMGEXP == 0, CROPDMG*10, "NA"))
# crop_1e_subset

# note: please refer to cited code sources above
crop_2e_subset <- subset(data, CROPDMGEXP == 2, select = c(EVTYPE, CROPDMG, CROPDMGEXP))
# crop_2e_subset
crop_2e_subset$CROPDMGTOT <- with(crop_2e_subset, ifelse(CROPDMGEXP == 2, CROPDMG*10^2, "NA"))
# crop_2e_subset

# note: please refer to cited code sources above
crop_3e_subset <- subset(data, CROPDMGEXP == "k" | CROPDMGEXP == "K", select = c(EVTYPE, CROPDM
G, CROPDMGEXP))
# crop_3e_subset
crop_3e_subset$CROPDMGTOT <- with(crop_3e_subset, ifelse(CROPDMGEXP == "k" | CROPDMGEXP == "K",
 CROPDMG*10^3, "NA"))
# crop_3e_subset

# note: please refer to cited code sources above
crop_6e_subset <- subset(data, CROPDMGEXP == "m" | CROPDMGEXP == "M", select = c(EVTYPE, CROPDM
G, CROPDMGEXP))
# crop_6e_subset
crop_6e_subset$CROPDMGTOT <- with(crop_6e_subset, ifelse(CROPDMGEXP == "m" | CROPDMGEXP == "M",
 CROPDMG*10^6, "NA"))
# crop_6e_subset

# note: please refer to cited code sources above
crop_9e_subset <- subset(data, CROPDMGEXP == "B", select = c(EVTYPE, CROPDMG, CROPDMGEXP))
# crop_9e_subset
crop_9e_subset$CROPDMGTOT <- with(crop_9e_subset, ifelse(CROPDMGEXP == "B", CROPDMG*10^9, "NA"))
# crop_9e_subset

cropdmgtot_subset <- rbind(crop_1e_subset, crop_2e_subset, crop_3e_subset, crop_6e_subset, crop_
9e_subset)
# cropdmgtot_subset

# aggregate by cropdmg or cropdmg, then apply calculations
# source: https://www.statmethods.net/management/aggregate.html
# event_cropdmg_total <- aggregate(CROPDMGTOT ~ EVTYPE, cropdmgtot_subset, FUN=length)
```

```
event_cropdmg_total <- aggregate(CROPDMGTOT ~ EVTYPE, cropdmgtot_subset, sum)

# rank aggregate results:
# https://stackoverflow.com/questions/23659241/rank-in-the-aggregate-function
event_cropdmg_total <- event_cropdmg_total[order(event_cropdmg_total$CROPDMGTOT, decreasing=TRUE
),]

# subset for top 50 event types with highest count
# https://stackoverflow.com/questions/2667673/select-first-4-rows-of-a-data-frame-in-r/47400307
event_cropdmg_10 <- event_cropdmg_total[1:10,]
event_cropdmg_10
```

```
##               EVTYPE  CROPDMGTOT
## 16           DROUGHT 13972566000
## 34             FLOOD  5661968450
## 98       RIVER FLOOD  5029459000
## 85         ICE STORM  5022113500
## 52              HAIL  3025954650
## 77         HURRICANE  2741910000
## 82 HURRICANE/TYPHOON  2607872800
## 30       FLASH FLOOD  1421317100
## 26      EXTREME COLD  1292973000
## 46      FROST/FREEZE  1094086000
```

```
# ratio of cropdmg due to top 10 event types / all event types
event_cropdmg_10_sum <- sum(event_cropdmg_10$CROPDMGTOT, na.rm=TRUE)
event_cropdmg_total_sum <- sum(event_cropdmg_total$CROPDMGTOT, na.rm=TRUE)
event_cropdmg_ratio <- event_cropdmg_10_sum / event_cropdmg_total_sum
print("ratio of croperty damage from top 10 / all event types:")
```

```
## [1] "ratio of croperty damage from top 10 / all event types:"
```

```
print(event_cropdmg_ratio)
```

```
## [1] 0.8526811
```

Part 3: Results

Part 3B: Plot Results

Question 2 - Analyze Event Type by Impacts to Economic Consequences

Overview:

1. Data processing produced subsets of aggregated, ranked data by economic impact
2. Each subset was reviewed, then visualized with bar plots
3. Results show which events had the most impact on economic consequences

4. Both property and crop damage are visualized for comparison
5. Based on results, flood and drought have largest economic impact

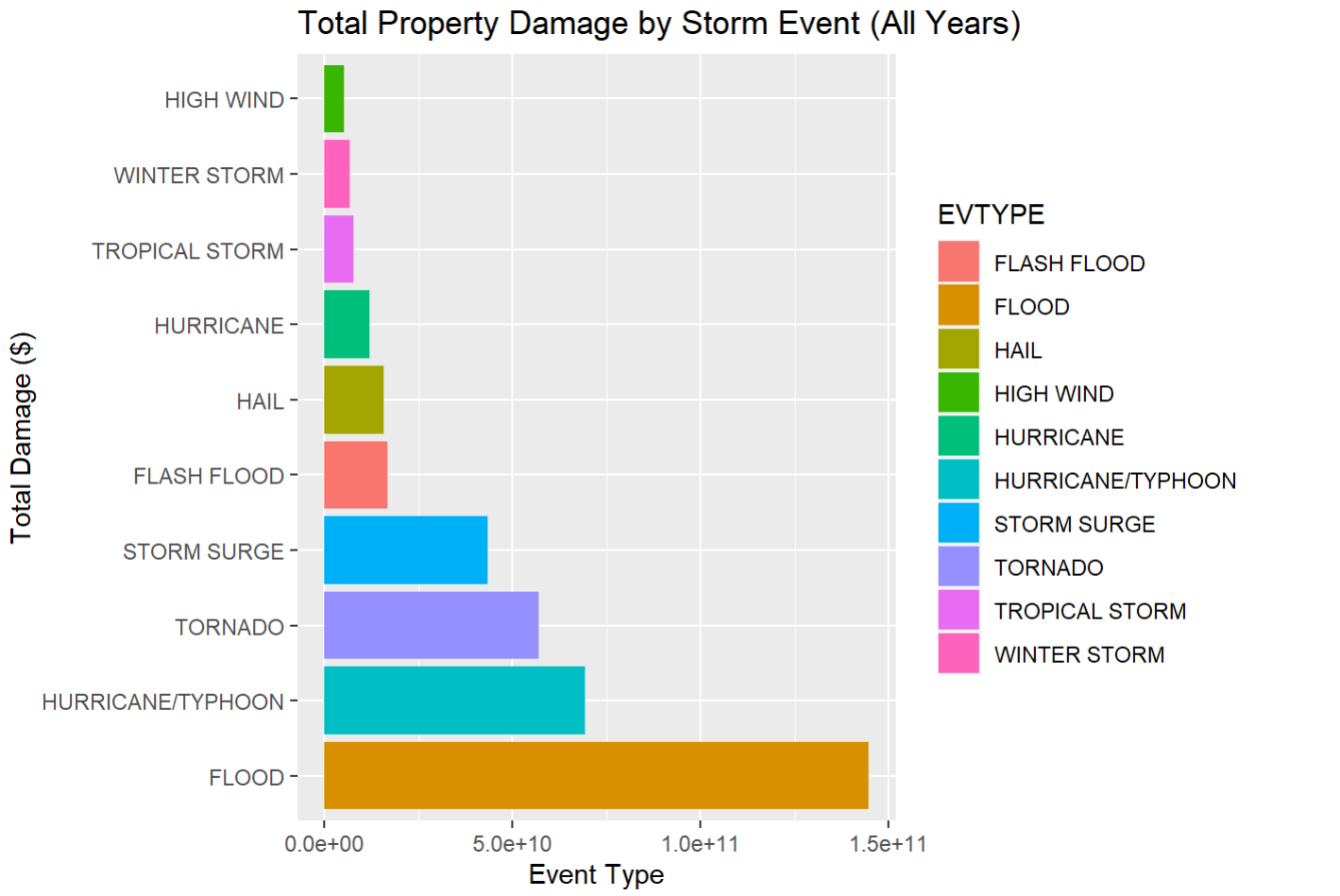
Steps:

1. Aggregated, subset data was ranked, then plot into bar plot
2. Plot is color coded by event type and axis labeled

Observations:

1. Property and crop damage compared for total economic impact
2. As a result, flood/drought had the largest impact
3. Based on results, flood and drought have largest economic impact

```
# part 3b: data analysis - plot event types by property damage
# http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization
# https://stackoverflow.com/questions/16961921/plot-data-in-descending-order-as-appears-in-data-frame
# install.packages("ggplot2")
# library(ggplot2)
p_propdmg <- ggplot(data=event_propdmg_10, aes(x=reorder(EVTYPE, -PROPDMGTOT), y=PROPDMGTOT, fill=EVTYPE)) +
    geom_bar(stat="identity") +
    labs(title="Total Property Damage by Storm Event (All Years)", x="Total Damage ($)", y="Event Type")
p_propdmg + coord_flip()
```



```
# http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization
# https://stackoverflow.com/questions/16961921/plot-data-in-descending-order-as-appears-in-data-frame
# install.packages("ggplot2")
# library(ggplot2)
p_cropdmg <- ggplot(data=event_cropdmg_10, aes(x=reorder(EVTYPE, -CROPDMGTOT), y=CROPDMGTOT, fill=EVTYPE)) +
    geom_bar(stat="identity") +
    labs(title="Total Crop Damage by Storm Event (All Years)", x="Total Damage ($)", y="Event Type")
p_cropdmg + coord_flip()
```