Data Analytics Project Report

AIT-580 2018 Fall Haowen Yu

1.Introduction

I found my interest in how modern people respond and prevent natural disasters, so I found the related data set on National Centers for Environmental Information website.

"NCEI is responsible for hosting and providing access to one of the most significant archives on Earth, with comprehensive oceanic, atmospheric, and geophysical data." [1]

The NCEI is merged by three different data centers in order to provide the public and researchers a more comprehensive database.

The full name of the dataset I select for my project is called NCDC Storm Events Database 2018 Jan-July. [2]

1.1 About Dataset

1.1.1 Privacy - The data set is open to public

1.1.2 Quality - There are some duplicate and blank records.

Some information collected from a third party may be unverified by the NWS.

1.1.3 Other Issues - Need to calculate and add new columns to answer some of my potential questions.

1.2 Metadata Information

There are totally 42,252 records and 51 columns in this dataset.

Since the data set include too many variables, I will only indicate metadata definitions of columns I use for data analysis.

BEGIN_YEARMONTH - Interval, Event start year and month.

BEGIN_DAY - Interval, Event start date.

BEGIN_TIME - Interval, Event start time.

END_YEARMONTH - Interval, Event end year and month.

END_DAY - Interval, Event end date.

END_TIME - Interval, Event end time.

TATE - Nominal, State name.

$YEAR - Interval, Year of event occurrence

MONTH_NAME - Nominal, Month of event occurrence

EVENT_TYPE - Nominal, Name of the different type of storms

INJURIES_DIRECT - Interval, Count of people directly injured in the event.

INJURIES_INDIRECT - Interval, Count of people indirectly injured in the event.

DEATHS_DIRECT - Interval, Count of people directly dead in the event.

DEATHS_INDIRECT - Interval, Count of people indirectly dead in the event.

DAMAGE_PROPERTY - Character string, Damaged property value

DAMAGE_CROPS - Character string, Damaged crops value

1.3Data exploration

Q1. How many unique Event Type is there in the dataset? Which Event happened the most?

```
dbGetQuery(con, "SELECT EVENT_TYPE, COUNT(*) AS n FROM stormevents GROUP BY EVENT_TYPE ORDER BY n DESC;")
```

##	EVENT_TYPE	n
## 1	Thunderstorm Wind	10822
## 2	Hail	6440
## 3	Winter Weather	3465
## 4	Flood	2821
## 5	Winter Storm	2778
## 6	High Wind	2173
## 7	Flash Flood	2068
## 8	Drought	1588
## 9	Heavy Snow	1550
## 10	Marine Thunderstorm Wind	1281
## 11	Heavy Rain	917
## 12	Heat	882
## 13	Tornado	706
## 14	Extreme Cold/Wind Chill	541
## 15	Cold/Wind Chill	503
## 16	Strong Wind	484
## 17	Blizzard	438
## 18	Frost/Freeze	431
## 19	Excessive Heat	347
## 20	Funnel Cloud	287
## 21	Dense Fog	266
## 22	High Surf	255
## 23	Wildfire	253
## 24	Lightning	235
## 25	Coastal Flood	133
## 26	Waterspout	105
## 27	Dust Storm	77
## 28	Debris Flow	70
## 29	Volcanic Ashfall	65
## 30	Rip Current	52
## 31	Lake-Effect Snow	42
## 32	Ice Storm	38
## 33	Marine Hail	22
## 34	Tropical Storm	22
## 35	Astronomical Low Tide	21
## 36	Avalanche	13
## 37	Tropical Depression	12
## 38	Marine High Wind	11
## 39	Sleet	8
## 40	Lakeshore Flood	8
## 41	Dense Smoke	6
## 42	Dust Devil	5
## 43	Marine Strong Wind	4
## 44	Storm Surge/Tide	2
## 45	Seiche	2
## 46	Marine Tropical Storm	2
## 47	Freezing Fog	1

As the result shows above, there are 47 event types in total. Thunderstorm Wind happened the most, and Freezing Fog only happened once.

Q2. Count Thunderstorm Wind for each state and arrange in descending order.

```
dbGetQuery(
  con,
  "SELECT STATE, COUNT(*) AS n
  FROM StormEvents
  WHERE (EVENT_TYPE IN ('Thunderstorm Wind'))
  GROUP BY STATE
  ORDER BY n DESC"
)
```

	STATE	n
1	KANSAS	631
2	GEORGIA	514
3	TEXAS	503
4	KENTUCKY	475
5	TENNESSEE	458
6	NORTH CAROLINA	448
7	OKLAHOMA	445
8	ALABAMA	412
9	SOUTH CAROLINA	406
10	VIRGINIA	399
11	NEW YORK	393
12	MISSISSIPPI	388
13	MISSOURI	384
14	OHIO	361
15	NEBRASKA	357
16	PENNSYLVANIA	348
17	IOWA	338
18	ARKANSAS	337
19	SOUTH DAKOTA	333
20	ILLINOIS	313
21	FLORIDA	293
22	MINNESOTA	257
23	INDIANA	251
24	NORTH DAKOTA	243
25	LOUISIANA	186
26	WISCONSIN	151
27	MONTANA	143
28	WEST VIRGINIA	135
29	MICHIGAN	121
30	COLORADO	106
31	MARYLAND	99
32	ARIZONA	91
33	NEW HAMPSHIRE	86
34	VERMONT	58
35	MASSACHUSETTS	55
36	WYOMING	54
37	NEW MEXICO	46
38	MAINE	43
39	CONNECTICUT	31
40	NEW JERSEY	26
41	IDAHO	26
42	NEVADA	21
43	UTAH	17
44	WASHINGTON	16
45	CALIFORNIA	6
46	DELAWARE	5
47	OREGON	4
48	DISTRICT OF COLUMBIA	3
49	HAWAII	2
50	PUERTO RICO	2
51	RHODE ISLAND	2

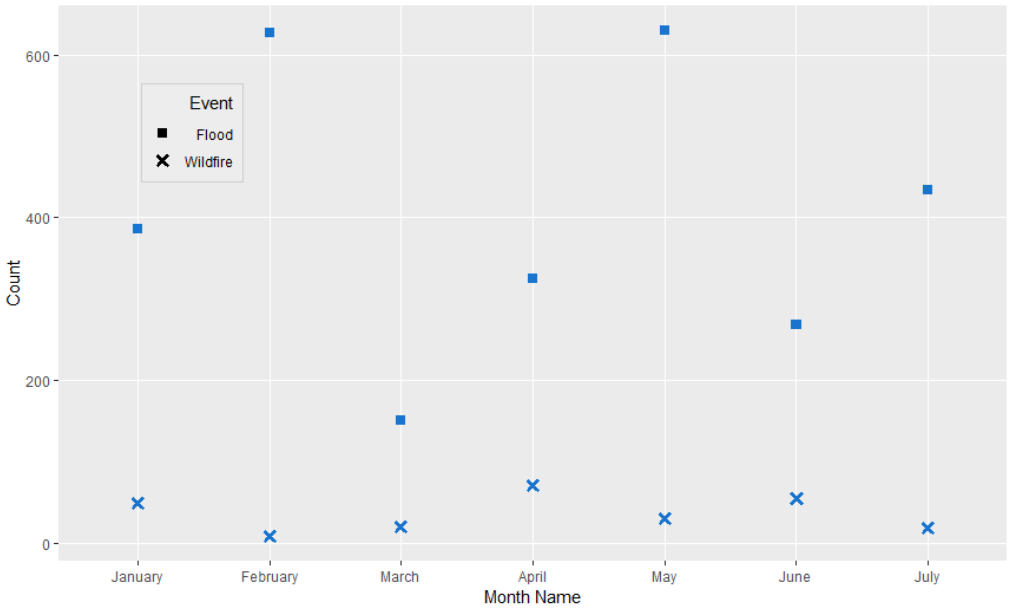
As the result shows above Kansas has the has the highest Thunderstorm Wind count, Rhode Island has the lowest Thunderstorm Wind count.

Please see attachment for SQL schema and query.

2.Data Analytics, visualizations and Interpretation

2.1 Scatterplot

Q3. Compare the incident count of Flood and Wildfire from January to July in 2018. Which month has the highest and lowest flood or wildfire incident count?



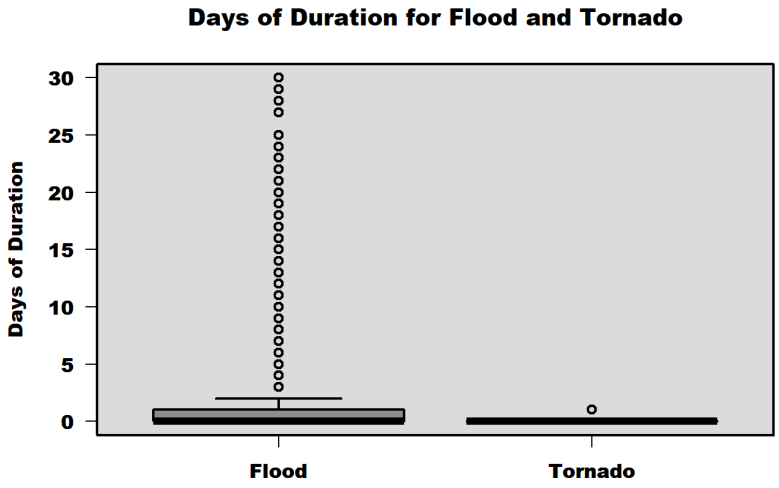
Square represents Flood, x represents Wildfire.

As the above graph shows February has the lowest Wildfire incident count, January has the lowest Flood incident count. April has the highest Wildfire incident count, June has the highest Flood incident count and February Flood incident count is almost as high as in June. There is no clear pattern between or correlation between Wildfire and Flood.

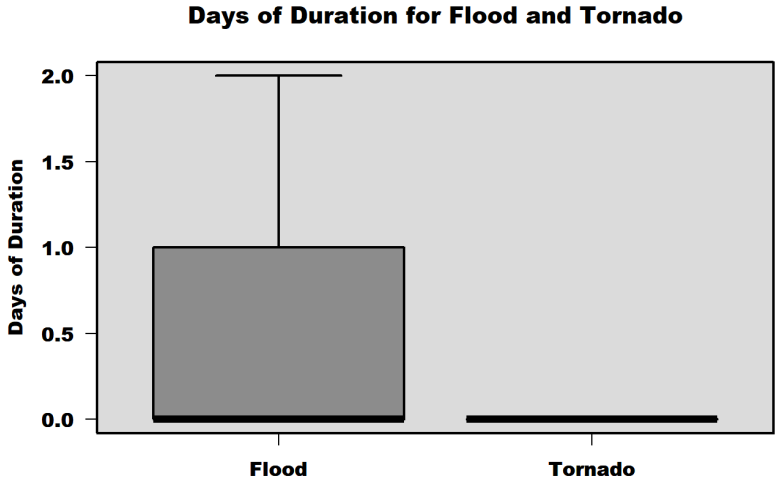
2.2 Boxplot

Q4. Compare the duration of Flood and Tornado. What can you find?

This question came out directly when I review this dataset. I believe it's a common knowledge that Flood has more duration time than Tornado, so I want to use boxplot to prove this. In order to find this answer, I will need to calculate the duration time first since it's not in the original data set.



The above graph came out first, but I think is a little bit hard to see the difference since the y-axis range is too big. So, I modified my R code, remove the outline and get the below graph.

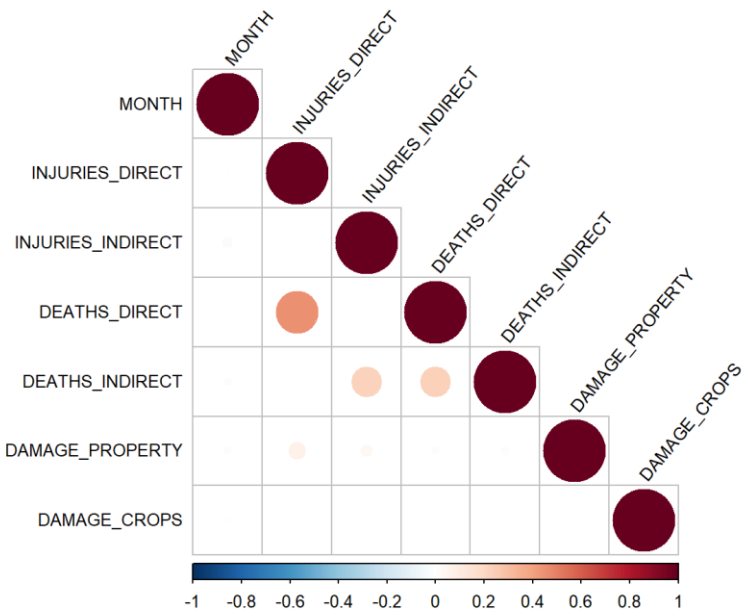


Now we have a very clear graph. Same as I expected, Flood has a longer duration than Tornado. Most Flood lasts almost 1 day, however most Tornado only lasts for a very short time.

2.3 Correlation Analysis

Q5. Is there any correlation you can find in this data set?

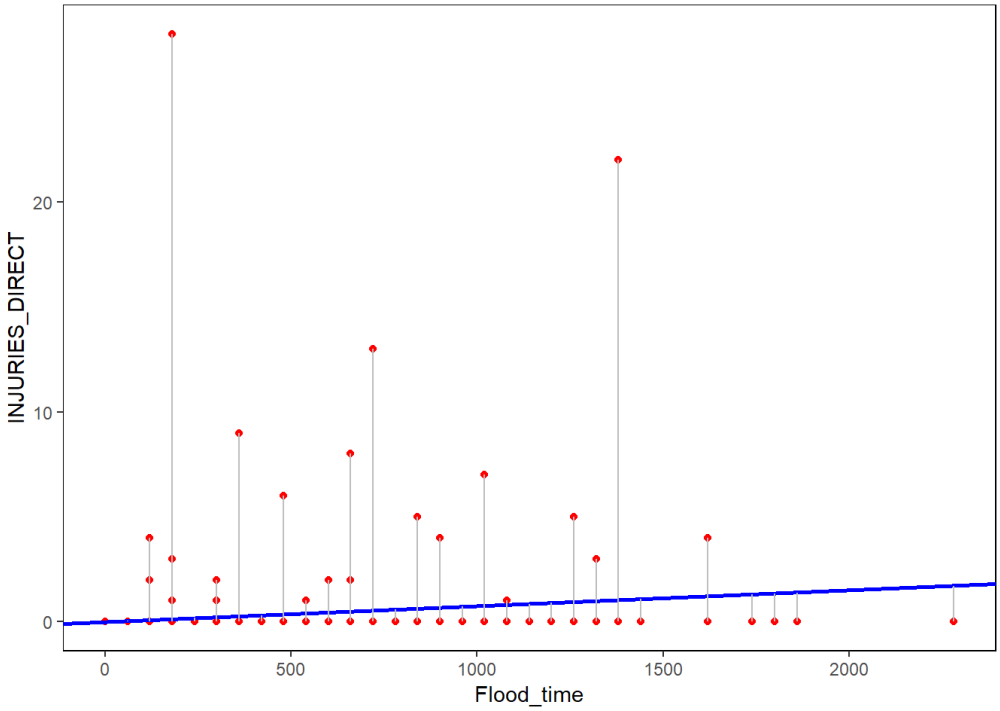
I create a corrplot trying to find correlation between injuries, deaths, and damage.



The darkest color shows between Injuries direct/indirect and death direct, so there is a significant relationship between these three variables.

2.4 Regression Analysis

Q6. Is there any direct relation between flood duration and injuries?



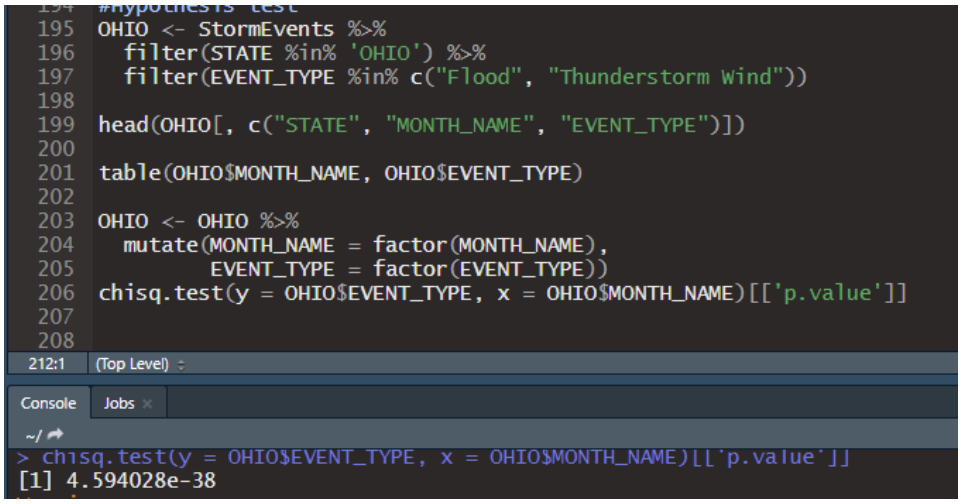
I did a linear regression to analyze this question. Red dots represent injuries counts and blue line represent the relation slope. However, as the output graph above shows, there is no direct relation between flood duration time and injuries.

2.5 Hypothesis Test

Q7. Does month/time have an impact on the events type?

Finally, I found I can use a hypothesis test to answer this question.

I assume month has an impact on the different event type, H0: month has no impact on event type; H1: month has significant impact on event type.



As the result shows p-value is 4.594028e-38 < α, null hypothesis is rejected.

2.6 Library in use

ggplot2

dplyr

RMySQL

DBI

Please see attachment for R code.

3.Definition

All technical terms are clear in this report, nothing needs to be included here.

4.References

[1] NOAA Official website, *About NOAA*. Retrieved from

https://www.nodc.noaa.gov/about/

[2] Data set reference

National Centers for Environmental Information, *NCDC Storm Events Database*, *Storm Events Data, (July 2018)*, Published byNOAA Customer Engagement Branch

Retrieved from https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ncdc:C00510

5.Attachements

sql scheme and query - Data Analystics Project sql Haowen Yu.R

R code - Data Analystics Project R Haowen Yu.R

data set - StormEvents_details-ftp_v1.0_d2018_c20181017.csv