# Storms in the US and Their Fatalities

STAT 440 Data Project Report

Apurva Chakravorty, Tejaswi Rachapudi, Minji Song, Tonghui Tian

## Introduction

We will be analyzing data regarding storms and associated fatalities in the US in 2000. This data falls in line with patterns of climate change and geography. We want to explore if there are significant discrepancies between the fatalities of different storms or locations and try to figure out why that is. By doing so, we can better understand the climate patterns as they're changing based on location, and whether future storms will pose a greater threat in those areas. We can also investigate whether there exists a pattern in the times of year or times of day that the types of storms take place, which would give us insight into how to better prepare for future events.

This data about storms in the US is provided by the National Weather service, accessed through https://catalog.data.gov/dataset/ncdc-storm-events-database. The entire set of data files tracks storms and injuries from 1950's to the current year. It also includes what kinds of storms occurred, such as hail, blizzards, heavy storms, etc. On the following link, you can access the exact files we want to use: https://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/. We are using the 'Details' and 'Fatalities' files from 2000. Here are the exact files:

1. StormEvents_details-ftp_v1.0_d2000_c20190920.csv.gz (renamed to StormEvents_details_2000.csv)
2. StormEvents_fatalities-ftp_v1.0_d2000_c20190920.csv.gz (renamed StormEvents_fatalities_2000.csv)

https://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/Storm-Data-Export-Format.pdf gives us all the variables included as well as an in-depth description of each of them.

## Methods

StormEvents_details_2000.csv contains statistics on storm event data in 2000 including location, month, event type, event id, and many more. This dataset has 50 variables and 52,007 records. StormEvents_fatalities_2000.csv contains information on storm fatalities in 2000 including month, event id, fatality type, age, sex, and location. This dataset has 11 variables and 476 records.

In order to validate the data, we need to make sure there are no blanks or missing values in the datasets. The log does not show any errors or warnings about blanks or missing values, so we know the datasets are valid. We do not need all the variables in our datasets, since there are so many. In the details dataset, we only want to keep the following variables: Begin_yearmonth, begin_day, End_yearmonth, end_day, event_id, state, month_name, event_title, cz_timezone, injuries_direct, injuries_indirect, deaths_direct, deaths_indirect, source. The rest of the variables in this dataset are not useful to us because they are either redundant or not interesting to us. In the fatalities dataset, we wish to remove a few variables for the same reason. We are left with fat_yearmonth, fat_day, fatality_id, event_id, fatality_type, fatality_age, fatality_sex, and fatality_location. In order to clean the data, we import the raw data with all variables, then we create new datasets with only the variables we wish to keep by subsetting the datasets. We do this to both the raw details dataset and the raw fatalities dataset so we can work with the cleaned versions. We also set labels for the variables of interest. Next, we combine both cleaned datasets by merging into a single dataset named combined_storms_fatalities. We created new data based on the combined storm and fatality data, which is called FAT_STOR. Since they both have an event_ID column, we used inner join syntax to display both storm data and fatality data by event_ID. FAT_STOR data has 476 rows and 21 columns and this means that only 476 event IDs have information in both storm and fatality data. Below, we describe each table (which can be found in the Results section) we have created:

In Table A, we used sql to look specifically at the total number of injuries, total number of deaths, the proportion of deaths to the number of people affected, and the number of storm events, grouped by state.

In Table B and C, we used a similar approach using sql to group the data by month, and then by type of storm respectively. In Table B, we looked at total injuries, deaths and storm events by month, and in Table C, we specifically analysed the number of sources each type of storm was publicized by. In both data tables, we used the ‘count’ and ‘sum’ function to find these totals.

Looking at the data tables thus far, we noticed that Heat, Tornado, and Lightning storm types resulted in the highest injuries and fatalities. In Tables D, E and F, we used proc sql to specifically extract some relevant data about each of those storm types. In each table, we extracted event ID, location, month, injuries and deaths (direct and indirect) specific to Heat in Table D, Tornado in Table E and Lightning in Table F. This enables us to see every single occurrence of that storm type in 2000 and compare the event data to each other.

For Table H, we counted the number of fatal events that occurred in 2000 using sql. From table G, we already know that the number of people killed in 2000 was 476, and from table H, we can find that the number of fatalities was 338.

Table I is a frequency table created using SAS that shows how many women and men were killed due to a storm. This table shows a very interesting result: excluding 2 dead fetuses of unknown gender, 169 of those who were killed were women and 305 were men. Men are almost twice as likely as women. By combining the results of table C to analyze, because the highest fatality event is heat, the reason for the high proportion of male deaths may be due to the higher proportion of men engaged in high temperature or outdoor work.

After analyzing the sex ratio of the victims, we are also interested in the age distribution of the victims. Through table J, by displaying important statistics, we found that the average and median age of the victims was 48. The maximum age of the victims is 98 years old, and the minimum is 1 year old children or fetuses, whose Q1 and Q3 are 28 years old and 69 years old, respectively. We can conclude that the age distribution of the deceased is close to the normal distribution.

For Table K, we wanted to analyze the location where people die. Before we used the whole combined_storms_fatalities data, we eliminated missing values first. We created a table to display the locations people die and how many people in which age group died there using proc freq statement. A total of 461 people died in a clear location and 16 values are unknown which are excluded from the table.

For Table L and Table M, we used FAT_STOR data for the rest of the data analyzing. Based on FAT_STOR data, we analyzed fatalities by states using proc sql. The variables are STATE and the number of fatalities in each state. Through the order by statement, we extracted the top 5 the most and least fatality states.

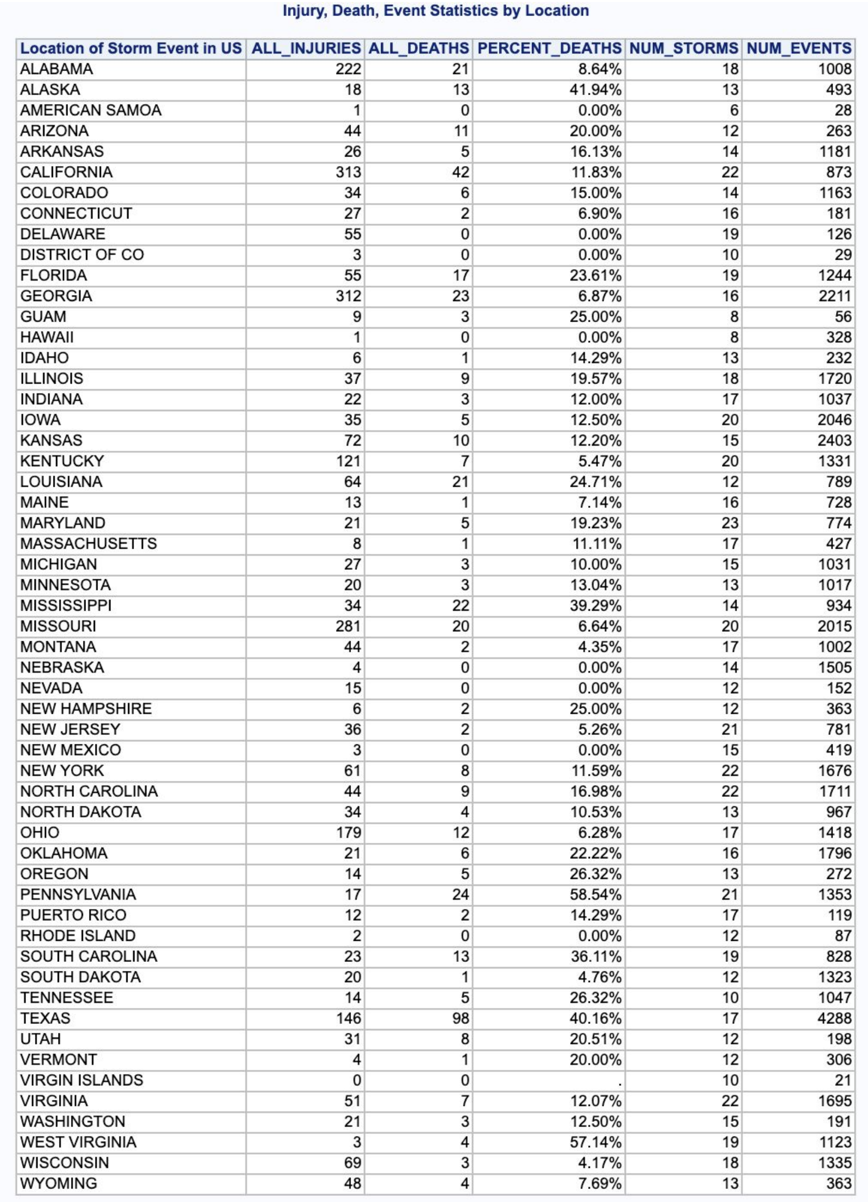
For Table N, we wanted to know which month has the highest fatality. Using sql, we were able to find the number of fatalities each month. The table has two columns: month name and the number of fatalities.

For Table O, after we figured out which state and when has the high fatality, we analyzed the main cause of death and its fatality numbers in the state which has the most fatalities. We added non correlated-subquery to evaluate how many people died in Texas by event_type first and found out the event_type that has the most fatality by using having statement. The table has three variables: state, event type and the number of fatalities in a state.

For Table P, we created the format for ages to classify the age group into 6 groups in order to analyze the relationship between age group and event type. We used proc freq statement to find the number of fatalities of each age group and event type. The table shows how many people in each age group died by which events and it also shows the fatality rate by age group.

## Results

Table A:

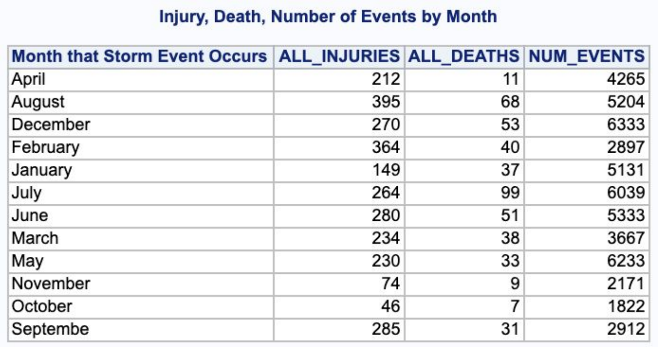


Injury, Death, Event Statistics by Location

| Location of Storm Event in US | ALL_INJURIES | ALL_DEATHS | PERCENT_DEATHS | NUM_STORMS | NUM_EVENTS |
|---|---|---|---|---|---|
| ALABAMA | 222 | 21 | 8.64% | 18 | 1008 |
| ALASKA | 18 | 13 | 41.94% | 13 | 493 |
| AMERICAN SAMOA | 1 | 0 | 0.00% | 6 | 28 |
| ARIZONA | 44 | 11 | 20.00% | 12 | 263 |
| ARKANSAS | 26 | 5 | 16.13% | 14 | 1181 |
| CALIFORNIA | 313 | 42 | 11.83% | 22 | 873 |
| COLORADO | 34 | 6 | 15.00% | 14 | 1163 |
| CONNECTICUT | 27 | 2 | 6.90% | 16 | 181 |
| DELAWARE | 55 | 0 | 0.00% | 19 | 126 |
| DISTRICT OF CO | 3 | 0 | 0.00% | 10 | 29 |
| FLORIDA | 55 | 17 | 23.61% | 19 | 1244 |
| GEORGIA | 312 | 23 | 6.87% | 16 | 2211 |
| GUAM | 9 | 3 | 25.00% | 8 | 56 |
| HAWAII | 1 | 0 | 0.00% | 8 | 328 |
| IDAHO | 6 | 1 | 14.29% | 13 | 232 |
| ILLINOIS | 37 | 9 | 19.57% | 18 | 1720 |
| INDIANA | 22 | 3 | 12.00% | 17 | 1037 |
| IOWA | 35 | 5 | 12.50% | 20 | 2046 |
| KANSAS | 72 | 10 | 12.20% | 15 | 2403 |
| KENTUCKY | 121 | 7 | 5.47% | 20 | 1331 |
| LOUISIANA | 64 | 21 | 24.71% | 12 | 789 |
| MAINE | 13 | 1 | 7.14% | 16 | 728 |
| MARYLAND | 21 | 5 | 19.23% | 23 | 774 |
| MASSACHUSETTS | 8 | 1 | 11.11% | 17 | 427 |
| MICHIGAN | 27 | 3 | 10.00% | 15 | 1031 |
| MINNESOTA | 20 | 3 | 13.04% | 13 | 1017 |
| MISSISSIPPI | 34 | 22 | 39.29% | 14 | 934 |
| MISSOURI | 281 | 20 | 6.64% | 20 | 2015 |
| MONTANA | 44 | 2 | 4.35% | 17 | 1002 |
| NEBRASKA | 4 | 0 | 0.00% | 14 | 1505 |
| NEVADA | 15 | 0 | 0.00% | 12 | 152 |
| NEW HAMPSHIRE | 6 | 2 | 25.00% | 12 | 363 |
| NEW JERSEY | 36 | 2 | 5.26% | 21 | 781 |
| NEW MEXICO | 3 | 0 | 0.00% | 15 | 419 |
| NEW YORK | 61 | 8 | 11.59% | 22 | 1676 |
| NORTH CAROLINA | 44 | 9 | 16.98% | 22 | 1711 |
| NORTH DAKOTA | 34 | 4 | 10.53% | 13 | 967 |
| OHIO | 179 | 12 | 6.28% | 17 | 1418 |
| OKLAHOMA | 21 | 6 | 22.22% | 16 | 1796 |
| OREGON | 14 | 5 | 26.32% | 13 | 272 |
| PENNSYLVANIA | 17 | 24 | 58.54% | 21 | 1353 |
| PUERTO RICO | 12 | 2 | 14.29% | 17 | 119 |
| RHODE ISLAND | 2 | 0 | 0.00% | 12 | 87 |
| SOUTH CAROLINA | 23 | 13 | 36.11% | 19 | 828 |
| SOUTH DAKOTA | 20 | 1 | 4.76% | 12 | 1323 |
| TENNESSEE | 14 | 5 | 26.32% | 10 | 1047 |
| TEXAS | 146 | 98 | 40.16% | 17 | 4288 |
| UTAH | 31 | 8 | 20.51% | 12 | 198 |
| VERMONT | 4 | 1 | 20.00% | 12 | 306 |
| VIRGIN ISLANDS | 0 | 0 | . | 10 | 21 |
| VIRGINIA | 51 | 7 | 12.07% | 22 | 1695 |
| WASHINGTON | 21 | 3 | 12.50% | 15 | 191 |
| WEST VIRGINIA | 3 | 4 | 57.14% | 19 | 1123 |
| WISCONSIN | 69 | 3 | 4.17% | 18 | 1335 |
| WYOMING | 48 | 4 | 7.69% | 13 | 363 |

Here, we grouped by state and counted the total injuries and deaths per state in 2000. We can easily see that California and Georgia had the highest number of injuries in 2000, but Pennsylvania, West Virginia and Alaska all have significantly high death rates in proportion to the number of people affected there. We then wanted to see if there might exist a correlation between the number of storms or type of storm, and the number of deaths in the area. At a quick glance, there is no obvious trend.

Table B:

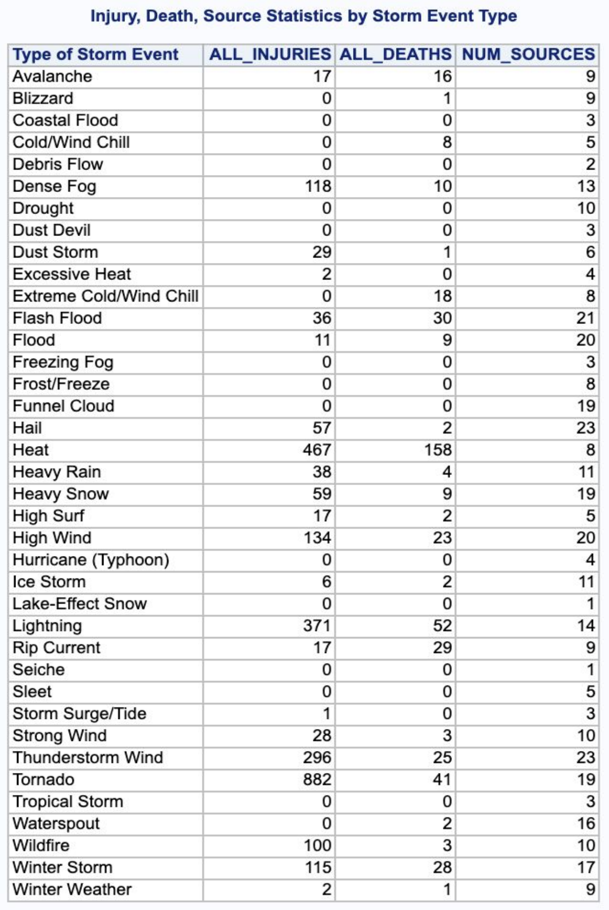


Injury, Death, Number of Events by Month

| Month that Storm Event Occurs | ALL_INJURIES | ALL_DEATHS | NUM_EVENTS |
|---|---|---|---|
| April | 212 | 11 | 4265 |
| August | 395 | 68 | 5204 |
| December | 270 | 53 | 6333 |
| February | 364 | 40 | 2897 |
| January | 149 | 37 | 5131 |
| July | 264 | 99 | 6039 |
| June | 280 | 51 | 5333 |
| March | 234 | 38 | 3667 |
| May | 230 | 33 | 6233 |
| November | 74 | 9 | 2171 |
| October | 46 | 7 | 1822 |
| Septembe | 285 | 31 | 2912 |

In this table, we organized the total injuries, deaths and storm events by Month, to see if there is any trend between the number of casualties and time of year, and whether the number of storm events increase or decrease with season. There is again, no obvious correlation, however, in October and November, there is a drop in number of storm events as well as injuries and death.

Table C:



Injury, Death, Source Statistics by Storm Event Type

| Type of Storm Event | ALL_INJURIES | ALL_DEATHS | NUM_SOURCES |
|---|---|---|---|
| Avalanche | 17 | 16 | 9 |
| Blizzard | 0 | 1 | 9 |
| Coastal Flood | 0 | 0 | 3 |
| Cold/Wind Chill | 0 | 8 | 5 |
| Debris Flow | 0 | 0 | 2 |
| Dense Fog | 118 | 10 | 13 |
| Drought | 0 | 0 | 10 |
| Dust Devil | 0 | 0 | 3 |
| Dust Storm | 29 | 1 | 6 |
| Excessive Heat | 2 | 0 | 4 |
| Extreme Cold/Wind Chill | 0 | 18 | 8 |
| Flash Flood | 36 | 30 | 21 |
| Flood | 11 | 9 | 20 |
| Freezing Fog | 0 | 0 | 3 |
| Frost/Freeze | 0 | 0 | 8 |
| Funnel Cloud | 0 | 0 | 19 |
| Hail | 57 | 2 | 23 |
| Heat | 467 | 158 | 8 |
| Heavy Rain | 38 | 4 | 11 |
| Heavy Snow | 59 | 9 | 19 |
| High Surf | 17 | 2 | 5 |
| High Wind | 134 | 23 | 20 |
| Hurricane (Typhoon) | 0 | 0 | 4 |
| Ice Storm | 6 | 2 | 11 |
| Lake-Effect Snow | 0 | 0 | 1 |
| Lightning | 371 | 52 | 14 |
| Rip Current | 17 | 29 | 9 |
| Seiche | 0 | 0 | 1 |
| Sleet | 0 | 0 | 5 |
| Storm Surge/Tide | 1 | 0 | 3 |
| Strong Wind | 28 | 3 | 10 |
| Thunderstorm Wind | 296 | 25 | 23 |
| Tornado | 882 | 41 | 19 |
| Tropical Storm | 0 | 0 | 3 |
| Waterspout | 0 | 2 | 16 |
| Wildfire | 100 | 3 | 10 |
| Winter Storm | 115 | 28 | 17 |
| Winter Weather | 2 | 1 | 9 |

This data table gives us the most information about the storm types and the number of injuries and deaths each one results in. Heat is the most effective storm with 467 injuries and the highest 158 fatalities. Tornado caused the most injuries with 882 injuries, and Lightning comes in third with 371 injuries.

Table G:

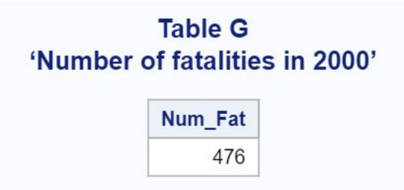


Table G
'Number of fatalities in 2000'

| Num_Fat |
|---|
| 476 |

Table H:

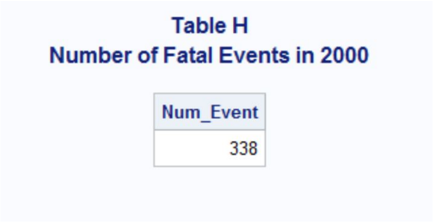


Table H
Number of Fatal Events in 2000

| Num_Event |
|---|
| 338 |

This table shows that the number of fatal events that occurred in 2000 was 338.

Table I:

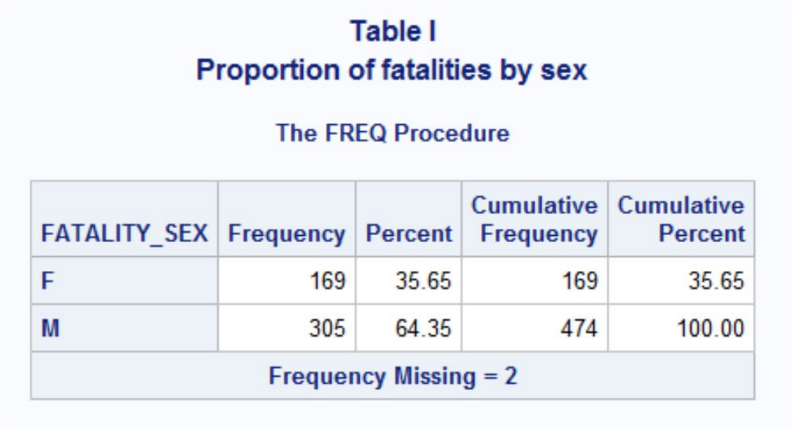


Table I
Proportion of fatalities by sex

The FREQ Procedure

| FATALITY_SEX | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| F | 169 | 35.65 | 169 | 35.65 |
| M | 305 | 64.35 | 474 | 100.00 |
| Frequency Missing = 2 | | | | |

Table J:

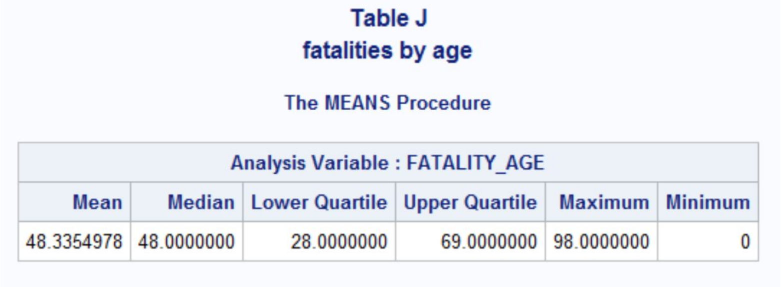
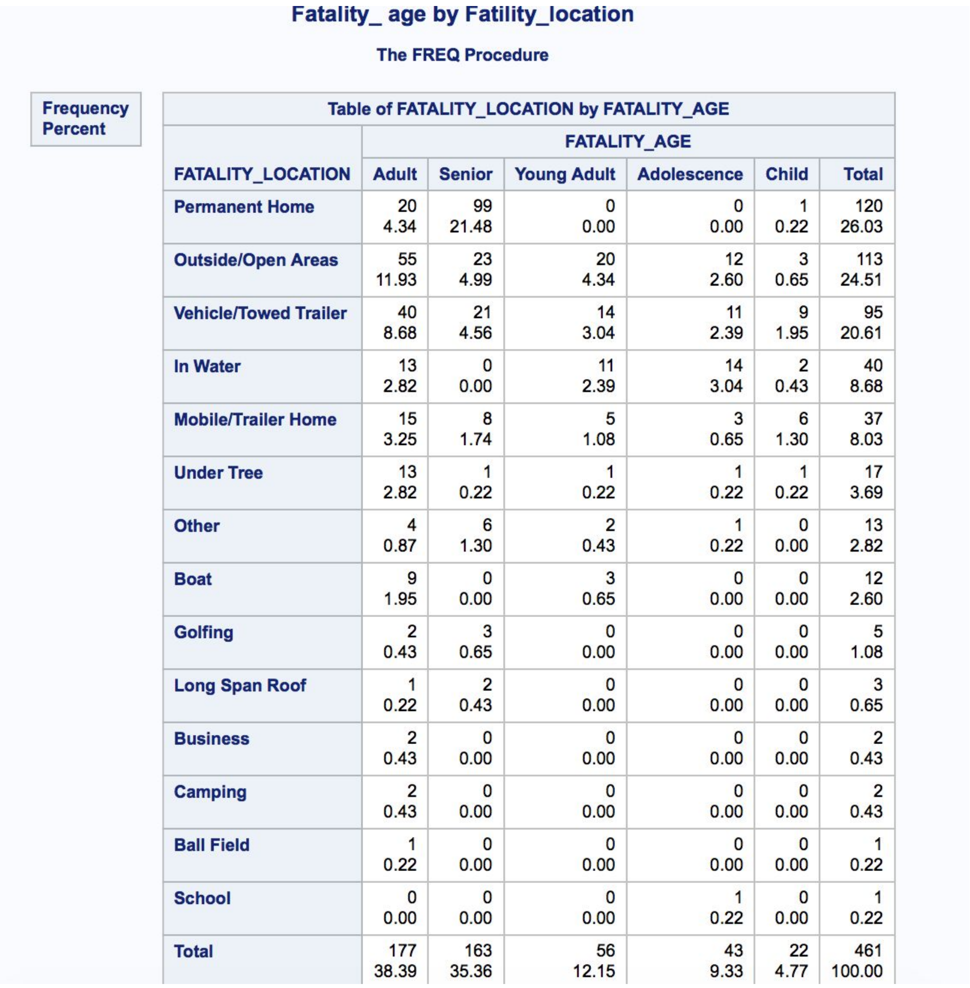


Table J
fatalities by age

The MEANS Procedure

| Analysis Variable : FATALITY_AGE | | | | | |
|---|---|---|---|---|---|
| Mean | Median | Lower Quartile | Upper Quartile | Maximum | Minimum |
| 48.3354978 | 48.0000000 | 28.0000000 | 69.0000000 | 98.0000000 | 0 |

Table K:



Fatality_ age by Fatility_location

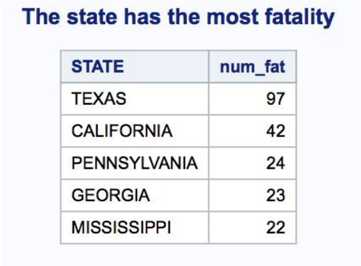
The FREQ Procedure

Frequency
Percent

Table of FATALITY_LOCATION by FATALITY_AGE

| | FATALITY_AGE | | | | | |
|---|---|---|---|---|---|---|
| FATALITY_LOCATION | Adult | Senior | Young Adult | Adolescence | Child | Total |
| Permanent Home | 20<br>4.34 | 99<br>21.48 | 0<br>0.00 | 0<br>0.00 | 1<br>0.22 | 120<br>26.03 |
| Outside/Open Areas | 55<br>11.93 | 23<br>4.99 | 20<br>4.34 | 12<br>2.60 | 3<br>0.65 | 113<br>24.51 |
| Vehicle/Towed Trailer | 40<br>8.68 | 21<br>4.56 | 14<br>3.04 | 11<br>2.39 | 9<br>1.95 | 95<br>20.61 |
| In Water | 13<br>2.82 | 0<br>0.00 | 11<br>2.39 | 14<br>3.04 | 2<br>0.43 | 40<br>8.68 |
| Mobile/Trailer Home | 15<br>3.25 | 8<br>1.74 | 5<br>1.08 | 3<br>0.65 | 6<br>1.30 | 37<br>8.03 |
| Under Tree | 13<br>2.82 | 1<br>0.22 | 1<br>0.22 | 1<br>0.22 | 1<br>0.22 | 17<br>3.69 |
| Other | 4<br>0.87 | 6<br>1.30 | 2<br>0.43 | 1<br>0.22 | 0<br>0.00 | 13<br>2.82 |
| Boat | 9<br>1.95 | 0<br>0.00 | 3<br>0.65 | 0<br>0.00 | 0<br>0.00 | 12<br>2.60 |
| Golfing | 2<br>0.43 | 3<br>0.65 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 5<br>1.08 |
| Long Span Roof | 1<br>0.22 | 2<br>0.43 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 3<br>0.65 |
| Business | 2<br>0.43 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 2<br>0.43 |
| Camping | 2<br>0.43 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 2<br>0.43 |
| Ball Field | 1<br>0.22 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 1<br>0.22 |
| School | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 1<br>0.22 | 0<br>0.00 | 1<br>0.22 |
| Total | 177<br>38.39 | 163<br>35.36 | 56<br>12.15 | 43<br>9.33 | 22<br>4.77 | 461<br>100.00 |

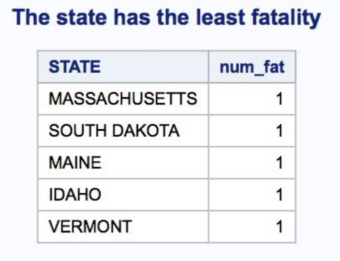
Table L:



The state has the most fatality

| STATE | num_fat |
|---|---:|
| TEXAS | 97 |
| CALIFORNIA | 42 |
| PENNSYLVANIA | 24 |
| GEORGIA | 23 |
| MISSISSIPPI | 22 |

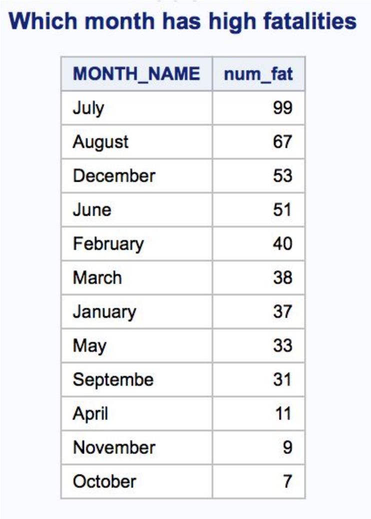
Table M:



The state has the least fatality

| STATE | num_fat |
|---|---:|
| MASSACHUSETTS | 1 |
| SOUTH DAKOTA | 1 |
| MAINE | 1 |
| IDAHO | 1 |
| VERMONT | 1 |

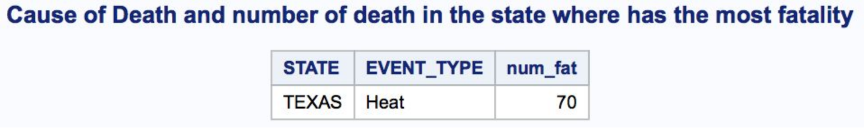
Table N:



Which month has high fatalities

| MONTH_NAME | num_fat |
|---|---:|
| July | 99 |
| August | 67 |
| December | 53 |
| June | 51 |
| February | 40 |
| March | 38 |
| January | 37 |
| May | 33 |
| Septembe | 31 |
| April | 11 |
| November | 9 |
| October | 7 |

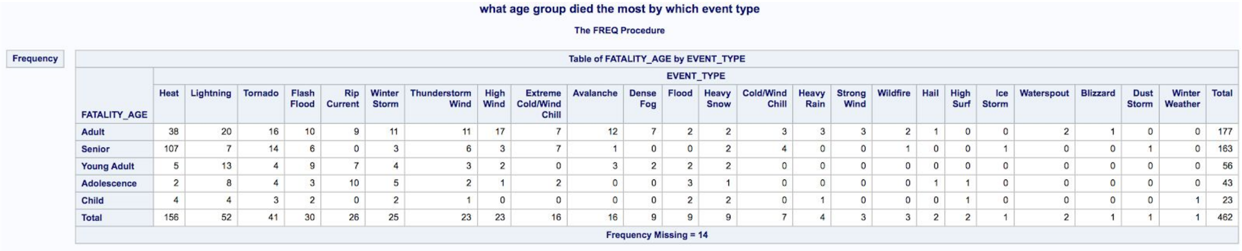
Table O:



Cause of Death and number of death in the state where has the most fatality

| STATE | EVENT_TYPE | num_fat |
|---|---|---|
| TEXAS | Heat | 70 |

Table P:



what age group died the most by which event type

The FREQ Procedure

Frequency

Table of FATALITY_AGE by EVENT_TYPE

| | EVENT_TYPE | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FATALITY_AGE | Heat | Lightning | Tornado | Flash Flood | Rip Current | Winter Storm | Thunderstorm Wind | High Wind | Extreme Cold/Wind Chill | Avalanche | Dense Fog | Flood | Heavy Snow | Cold/Wind Chill | Heavy Rain | Strong Wind | Wildfire | Hail | High Surf | Ice Storm | Waterspout | Blizzard | Dust Storm | Winter Weather | Total |
| Adult | 38 | 20 | 16 | 10 | 9 | 11 | 11 | 17 | 7 | 12 | 7 | 2 | 2 | 3 | 3 | 3 | 2 | 1 | 0 | 0 | 2 | 1 | 0 | 0 | 177 |
| Senior | 107 | 7 | 14 | 6 | 0 | 3 | 6 | 3 | 7 | 1 | 0 | 0 | 2 | 4 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 163 |
| Young Adult | 5 | 13 | 4 | 9 | 7 | 4 | 3 | 2 | 0 | 3 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 56 |
| Adolescence | 2 | 8 | 4 | 3 | 10 | 5 | 2 | 1 | 2 | 0 | 0 | 3 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 43 |
| Child | 4 | 4 | 3 | 2 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 23 |
| Total | 156 | 52 | 41 | 30 | 26 | 25 | 23 | 23 | 16 | 16 | 9 | 9 | 9 | 7 | 4 | 3 | 3 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 462 |

Frequency Missing = 14

Tables K-P show the following: The locations where people who died in 2000 the most are Permanent Home, Outside/Open Areas and Vehicle/Towed Trailer. Adult group died the most in age groups. Texas is the state that has the highest fatality in the U.S. as 97 people died in 2000 and only one person died in Massachusetts, South Dakota, Maine, Idaho, and Vermont which are the lowest fatality states. A total of 166 people died in July or August. We focused on Texas to analyze any factors that are related to death since it has the highest fatality. The main cause of death in Texas is Heat, 70 people died because of the heat. Seniors are the most vulnerable age group to heat and the next group is Adult.