Exploratory Data Analysis

Capstone Step 4

Document Purpose

This document serves as a high level overview of the process by which this Capstone project was carried out.

It tells the story of how I will arrive at the final state in a way that any engineer could follow.

Insights of Exploratory Data Analysis

The Storm Events Database contains the records used to create the official NOAA Storm Data publication, documenting:

- The occurrence of storms and other significant weather phenomena having sufficient intensity to cause loss of life, injuries, significant property damage, and/or disruption to commerce;
- Rare, unusual, weather phenomena that generate media attention, such as snow flurries in South Florida or the San Diego coastal area; and
- Other significant meteorological events, such as record maximum or minimum temperatures or precipitation that occur in connection with another event.

The database currently contains data from January 1950 to November 2021, as entered by NOAA's National Weather Service (NWS). Due to changes in the data collection and processing procedures over time, there are unique periods of record available depending on the event type. NCEI has performed data reformatting and standardization of event types but has not changed any data values for locations, fatalities, injuries, damage, narratives and any other event specific information.

The source data exists as CSVs which have been created since 1950 until present:

- storm_details_*.csv
- storm_locations_*.csv
- storm_fatalities_*.csv

Each has an EVENT_ID column available.
The details/locations data also have an EPISODE_ID
The fatalities data has FATALITY_ID

Insights of Exploratory Data Analysis

The Storm Events Database contains the records used to create the official NOAA Storm Data publication, documenting:

- The occurrence of storms and other significant weather phenomena having sufficient intensity to cause loss of life, injuries, significant property damage, and/or disruption to commerce;
- Rare, unusual, weather phenomena that generate media attention, such as snow flurries in South Florida or the San Diego coastal area; and
- Other significant meteorological events, such as record maximum or minimum temperatures or precipitation that occur in connection with another event.

The database currently contains data from January 1950 to November 2021, as entered by NOAA's National Weather Service (NWS). Due to changes in the data collection and processing procedures over time, there are unique periods of record available depending on the event type. NCEI has performed data reformatting and standardization of event types but has not changed any data values for locations, fatalities, injuries, damage, narratives and any other event specific information.

The source data exists as CSVs which have been created since 1950 until present:

- storm_details_*.csv
- storm_locations_*.csv
- storm_fatalities_*.csv

Each has an EVENT_ID column available.
The details/locations data also have an EPISODE_ID
The fatalities data has FATALITY_ID

event_id is unique for each row in details

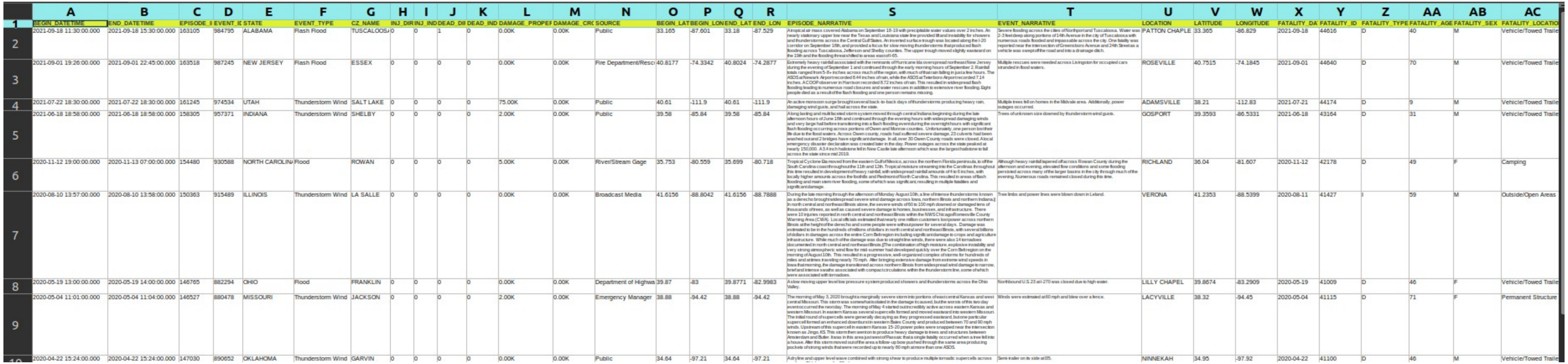
every fatality has an event_id match in detail
  those have repeated values, so the distinct subset is less
some fatalities have an event_id match in location
  those have repeated values, so the distinct subset is less

every location has an event_id match in detail
  those have repeated values, so the distinct subset is less
some locations have an event_id match in fatality
  those have repeated values, so the distinct subset is less

Visuals

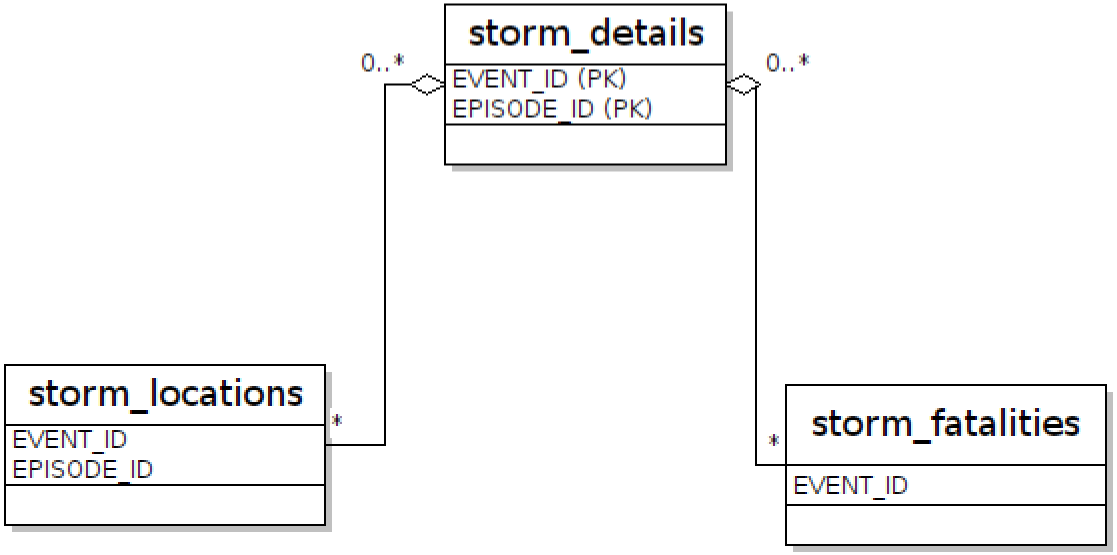
Total row count of details: 1680127
Total row count of locations: 1387001
Total row count of fatalities: 19133

I created a csv of a random sample of 10 records from the result of joining all 3 sets together, selecting a subset of critical columns, concatening columns and casting from strings to more suitable datatypes, and filtering to only non-null rows: randfull.csv



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	BEGIN_DATETIME	END_DATETIME	EPISODE_I	EVENT_I	STATE	EVENT_TYPE	CZ_NAME	INJ_DIR	INJ_IND	DEAD_DII	DEAD_IND	DAMAGE_PROPEF	DAMAGE_CRC	SOURCE	BEGIN_LAT	BEGIN_LON	END_LAT	END_LON	EPISODE_NARRATIVE	EVENT_NARRATIVE	LOCATION	LATITUDE	LONGITUDE	FATALITY_DA	FATALITY_ID	FATALITY_TYPE	FATALITY_AGE	FATALITY_SEX	FATALITY_LOCATIO
2	2021-09-18 11:30:00.000	2021-09-18 15:30:00.000	163105	984795	ALABAMA	Flash Flood	TUSCALOOSA	0	0	1	0	0.00K	0.00K	Public	33.165	-87.601	33.18	-87.529	A tropical air mass covered Alabama on September 18-19 with precipitable water values over 2 inches. An nearly stationary upper low near the Texas and Louisiana state line provided lift and instability for showers and thunderstorms across the Central Gulf States. An inverted surface trough was located along the I-20 corridor on September 18th, and provided a focus for slow moving thunderstorms that produced flash flooding across Tuscaloosa, Jefferson and Shelby counties. The upper trough moved slightly eastward on the 19th and the flooding threat shifted to areas east of I-65.	Severe flooding across the cities of Northport and Tuscaloosa. Water was 2-3 feet deep along portions of 14th Avenue in the city of Tuscaloosa with numerous roads flooded and impassable across the city. One fatality was reported near the intersection of Greensboro Avenue and 24th Street as a vehicle was swept off the road and into a drainage ditch.	PATTON CHAPLE	33.365	-86.829	2021-09-18	44616	D	40	M	Vehicle/Towed Traile
3	2021-09-01 19:26:00.000	2021-09-01 22:45:00.000	163518	987245	NEW JERSEY	Flash Flood	ESSEX	0	0	0	0	0.00K	0.00K	Fire Department/Resc	40.8177	-74.3342	40.8024	-74.2877	Extremely heavy rainfall associated with the remnants of Hurricane Ida overspread northeast New Jersey during the evening of September 1 and continued through the early morning hours of September 2. Rainfall totals ranged from 5-8+ inches across much of the region, with much of that rain falling in just a few hours. The ASOS at Newark Airport recorded 8.44 inches of rain, while the ASOS at Teterboro Airport recorded 7.14 inches. A COOP observer in Harrison recorded 8.72 inches of rain. This resulted in widespread flash flooding leading to numerous road closures and water rescues in addition to extensive river flooding. Eight people died as a result of the flash flooding and one person remains missing.	Multiple rescues were needed across Livingston for occupied cars stranded in flood waters.	ROSEVILLE	40.7515	-74.1845	2021-09-01	44640	D	70	M	Vehicle/Towed Traile
4	2021-07-22 18:30:00.000	2021-07-22 18:30:00.000	161245	974534	UTAH	Thunderstorm Wind	SALT LAKE	0	0	0	0	75.00K	0.00K	Public	40.61	-111.9	40.61	-111.9	An active monsoon surge brought several back-to-back days of thunderstorms producing heavy rain, damaging wind gusts, and hail across the state.	Multiple trees fell on homes in the Midvale area. Additionally, power outages occurred.	ADAMSVILLE	38.21	-112.83	2021-07-21	44174	D	9	M	Vehicle/Towed Traile
5	2021-06-18 18:58:00.000	2021-06-18 18:58:00.000	158305	957371	INDIANA	Thunderstorm Wind	SHELBY	0	0	0	0	2.00K	0.00K	Public	39.58	-85.84	39.58	-85.84	A long lasting and multi faceted storm system moved through central Indiana beginning during the late afternoon hours of June 18th and continued through the evening hours with widespread damaging winds and very large hail before transitioning into a flash flooding event during the overnight hours with significant flash flooding occurring across portions of Owen and Monroe counties. Unfortunately, one person lost their life due to the flood waters. Across Owen county, roads had suffered severe damage, 23 culverts had been washed out and 2 bridges have significant damage. In all, over 30 Owen County roads were closed. A local emergency disaster declaration was created later in the day. Power outages across the state peaked at nearly 150,000. A 3.4 inch hailstone fell in New Castle late afternoon which was the largest hailstone to fall across the state since mid 2019.	Trees of unknown size downed by thunderstorm wind gusts.	GOSPORT	39.3593	-86.5331	2021-06-18	43164	D	31	M	Vehicle/Towed Traile
6	2020-11-12 19:00:00.000	2020-11-13 07:00:00.000	154480	930588	NORTH CAROLINA	Flood	ROWAN	0	0	0	0	5.00K	0.00K	River/Stream Gage	35.753	-80.559	35.699	-80.718	Tropical Cyclone Eta moved from the eastern Gulf of Mexico, across the northern Florida peninsula, to off the South Carolina coast throughout the 11th and 12th. Tropical moisture streaming into the Carolinas throughout this time resulted in development of heavy rainfall, with widespread rainfall amounts of 4 to 6 inches, with locally higher amounts across the foothills and Piedmont of North Carolina. This resulted in areas of flash flooding and main stem river flooding, some of which was significant, resulting in multiple fatalities and significant damage.	Although heavy rainfall tapered off across Rowan County during the afternoon and evening, elevated flow conditions and some flooding persisted across many of the larger basins in the city through much of the evening. Numerous roads remained closed during this time.	RICHLAND	36.04	-81.607	2020-11-12	42178	D	49	F	Camping
7	2020-08-10 13:57:00.000	2020-08-10 13:58:00.000	150363	915489	ILLINOIS	Thunderstorm Wind	LA SALLE	0	0	0	0	0.00K	0.00K	Broadcast Media	41.6156	-88.8042	41.6156	-88.7888	During the late morning through the afternoon of Monday August 10th, a line of intense thunderstorms known as a derecho brought widespread severe wind damage across Iowa, northern Illinois and northern Indiana.|| In north central and northeast Illinois alone, the severe winds of 60 to 100 mph downed or damaged tens of thousands of trees, as well as caused severe damage to homes, businesses, and infrastructure. There were 10 injuries reported in north central and northeast Illinois within the NWS Chicago/Romeoville County Warning Area (CWA). Local officials estimated that nearly one million customers lost power across northern Illinois at the height of the derecho and some people were without power for several days. Damage was estimated to be in the hundreds of millions of dollars in north central and northeast Illinois, with several billions of dollars in damages across the entire Corn Belt region including significant damage to crops and agriculture infrastructure. While much of the damage was due to straight line winds, there were also 14 tornadoes documented in north central and northeast Illinois.||The combination of high moisture, explosive instability and very strong atmospheric wind flow for mid-summer had developed quickly over the Corn Belt region on the morning of August 10th. This resulted in a progressive, well-organized complex of storms for hundreds of miles and at times traveling nearly 70 mph. After bringing extensive damage from extreme wind speeds in Iowa that morning, the damage transitioned across northern Illinois from widespread wind damage to narrow, brief and intense swaths associated with compact circulations within the thunderstorm line, some of which were associated with tornadoes.	Tree limbs and power lines were blown down in Leland.	VERONA	41.2353	-88.5399	2020-08-11	41427	I	59	M	Outside/Open Areas
8	2020-05-19 13:00:00.000	2020-05-19 14:00:00.000	146765	882294	OHIO	Flood	FRANKLIN	0	0	0	0	0.00K	0.00K	Department of Highwa	39.87	-83	39.8771	-82.9983	A slow moving upper level low pressure system produced showers and thunderstorms across the Ohio Valley.	Northbound U.S. 23 at I-270 was closed due to high water.	LILLY CHAPEL	39.8674	-83.2909	2020-05-19	41009	D	46	F	Vehicle/Towed Traile
9	2020-05-04 11:01:00.000	2020-05-04 11:04:00.000	146527	880478	MISSOURI	Thunderstorm Wind	JACKSON	0	0	0	0	2.00K	0.00K	Emergency Manager	38.88	-94.42	38.88	-94.42	The morning of May 3, 2020 brought a marginally severe storm into portions of east central Kansas and west central Missouri. This storm was somewhat isolated in the damage it caused, but the worst of this two day event occurred the next day. The morning of May 4 started out incredibly active across eastern Kansas and western Missouri. In eastern Kansas several supercells formed and moved eastward into western Missouri. The initial round of supercells were generally decaying as they progressed eastward, but one particular supercell formed an enhanced downburst in western Bates County and produced between 70 and 90 mph winds. Upstream of this supercell in eastern Kansas 15-20 power poles were snapped near the intersection known as Jingo, KS. This storm then went on to produce heavy damage to trees and structures between Amsterdam and Butler. It was in this area just west of Passaic that a single fatality occurred when a tree fell into a house. After this storm moved out of the area a follow-up bow pushed through the same area producing pockets of strong winds that were recorded up to nearly 80 mph at more than one ASOS.	Winds were estimated at 60 mph and blew over a fence.	LACYVILLE	38.32	-94.45	2020-05-04	41115	D	71	F	Permanent Structure
10	2020-04-22 15:24:00.000	2020-04-22 15:24:00.000	147030	890652	OKLAHOMA	Thunderstorm Wind	GARVIN	0	0	0	0	0.00K	0.00K	Public	34.64	-97.21	34.64	-97.21	A dryline and upper level wave combined with strong shear to produce multiple tornadic supercells across	Semi-trailer on its side at I35.	NINNEKAH	34.95	-97.92	2020-04-22	41100	D	46	M	Vehicle/Towed Traile

ERD Diagram


storm_details
EVENT_ID (PK)
EPISODE_ID (PK)
0..*
0..*
storm_locations
EVENT_ID
EPISODE_ID
*
storm_fatalities
EVENT_ID
*