

Exploratory Data Analysis

Capstone Step 5

Choices Regarding Data Cleaning/Transformation

During prototyping I observed a few high level improvements possible via treating the raw data. Specific choices are annotated in screenshots on the subsequent slides.

1. Location file is entirely redundant and will be descoped.
2. A number of columns are redundant and will be removed after verifying the other locations offer equivalent data
3. A number of columns can be joined/consolidated
4. A number of columns can be converted from varchar to numerical type

Choices Regarding Pipeline Automation

The prototyping phase helped form the following approach to automation in order to achieve the optimal reliability:

1. To handle embedded commas within csv columns, use Linux csvkit's csvcut tool from Python (cut and awk both have tremendous difficulty).
- 2 . Use sqlalchemy create_engine, pandas read_csv and dataframe.to_sql instead of looping through each row to execute mysql_connector_python insert statement. The latter fails in cases of NULL values.
3. Do not declare any columns NOT NULL, and err on the side of caution with strings (TEXT instead of VARCHAR).
4. Delete files after processing to keep to manageable level.

	A	R	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V			
	BEGIN YEAR	MONTH	BEGIN DAY	END YEAR	MONTH	END DAY	PROG ID	EVENT ID	STATE	STATE ABB	YEAR	MONTH	NAME	EVENT TYPE	CZ TYPE	CZ RPS	CZ NAME	WFO	BEGIN DATE	TIME	END DATE	TIME	INJURIES	DIRECT	INDIRECT
2	202004	23	207	202004	23	228	47469	887009	MISSISSIPPI	28	2020	April	Tornado	C	5	AMITE	PHX	23-APR-20 02:07	00	24-APR-20 02:00	00	0	0	0	
3	202004	4	950	202008	4	951	48216	904094	PENNSYLVANIA	42	2020	August	Tornado	C	17	BUCKS	LIE	04-AUG-20 09:55	00	05-AUG-20 09:00	00	0	0	0	
4	202002	6	528	202002	6	529	45823	875896	GEORGIA	13	2020	February	Tornado	C	15	BARTOW	PHX	06-FEB-20 05:05	00	07-FEB-20 05:00	00	0	0	0	
5	202002	8	1055	202002	8	1100	45214	90338	NORTH CAROLINA	37	2020	February	Flood	C	45	CLEVELAND	PHX	08-FEB-20 05:05	00	09-FEB-20 05:00	00	0	0	0	
6	202007	25	1315	202007	25	1300	50353	913730	TEXAS	48	2020	July	Hurricane	Z	256	COASTAL WILACY	BPC	25-JUL-20 13:00	00	26-JUL-20 05:00	00	0	0	0	
7	202007	25	1400	202007	25	1800	51486	912708	TEXAS	48	2020	July	Hurricane	Z	442	KLEBERG COUNTY ISLANDS	CRP	25-JUL-20 14:00	00	26-JUL-20 05:00	00	0	0	0	
8	202005	26	1605	202005	27	600	50434	881096	TEXAS	48	2020	May	ORANGE	C	156	WEBSTER	PHX	26-MAY-20 16:05	00	27-MAY-20 05:00	00	0	0	0	
9	202005	25	1700	202005	25	1800	47310	885808	WEST VIRGINIA	54	2020	May	Flash Flood	C	101	WEBSTER	CRP	25-MAY-20 17:00	00	26-MAY-20 05:00	00	0	0	0	
10	202004	13	30	202004	13	330	46523	880416	VIRGINIA	51	2020	April	Flood	C	51	DICKENSON	ILN	13-APR-20 03:30	00	14-APR-20 03:00	00	0	0	0	
11	202003	28	1830	202003	28	1930	45618	874556	OHIO	39	2020	March	Flash Flood	C	37	DARKE	ILN	28-MAR-20 18:30	00	29-MAR-20 03:00	00	0	0	0	
12	202003	28	1430	202003	28	1530	45617	874465	OHIO	39	2020	March	Flood	C	103	LICKING	ILN	28-MAR-20 14:30	00	29-MAR-20 03:00	00	0	0	0	
13	202004	13	1900	202004	14	545	46730	883021	OHIO	39	2020	April	Flood	C	167	WASHINGTON	ILN	13-APR-20 19:00	00	14-APR-20 04:00	00	0	0	0	
14	202009	9	1401	202009	9	1615	52853	920637	PUERTO RICO	99	2020	September	Flash Flood	C	127	SAN JUAN	SLU	09-SEP-20 14:01	00	09-SEP-20 14:00	00	0	0	0	
15	202009	9	1405	202009	9	1615	52853	920638	PUERTO RICO	99	2020	September	Flash Flood	C	127	SAN JUAN	SLU	09-SEP-20 14:05	00	09-SEP-20 14:00	00	0	0	0	
16	202003	31	1405	202003	31	1405	46986	883678	SOUTH CAROLINA	45	2020	March	Thunderstorm Wind	C	17	JASPER	CHS	31-MAR-20 14:05	00	31-MAR-20 14:00	00	0	0	0	
17	202003	31	1451	202003	31	1452	46987	883679	SOUTH CAROLINA	45	2020	March	Thunderstorm Wind	C	5	ALLENDALE	CHS	31-MAR-20 14:51	00	31-MAR-20 14:00	00	0	0	0	
18	202002	6	7230	202002	6	1310	44988	870696	ALABAMA	12	2020	February	High Wind	Z	136	EASTERN ALABAMA	JAX	06-FEB-20 12:30	00	07-FEB-20 03:00	00	0	0	0	
19	202004	13	1751	202004	13	816</																			

[illegible]

AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ		
TOR_F_SCALE	TOR_LENGTH	TOR_WIDTH	TOR_OTHER_WFO	TOR_OTHER_CZ	STATE	TOR_OTHER_CZ	RPSTOR_OTHER_CZ	NAME	BEGIN_RANGE	BEGIN_AZMUTH	BEGIN_LOCATION	END_RANGE	END_AZ
EF2	18.3	1760	LIX	MS	113	PIKE	4	W	EAST FORK	11	E		
EF2	0.96	500	PH	PA	101	PHILADELPHIA	2	W	CORNWELLS HGTS	2	W		
EF0	0.33	200	FG	GA	129	GORDON	1	NNW	POLSON	2	NW		
EF2	5.14	150	GSP	NC	71	GASTON	1	ESE	ARNDTDALE	4	SSE		
							1	NNW	ERBACON	2	E		
							1	E	OSBORNS GAP	2	NE		
							1	E	GREENVILLE	2	S		
							2	ESE	PATASKALA ZEUNE ARPT	2	NW		
							1	N	MACKSBURG	1	NE		
							4	NW	SAN JUAN	3	NW		
							4	NW	SAN JUAN	4	NW		
							2	NNW	ALLENDIALE OSPWALD ARP	2	NNW		
							1	WSW	FOLLY BEACH	1	WSW		
							1	E	SULLIVANS ISLAND	1	E		
							0	N	CHESAPEAKE BAY BRIDGE TUNNEL 3RD ISLAMO		N		
							0	N	STANNARD ROCK LIGHT	0	N		

[illegible][illegible]

Details (cont)

	AQ	AR	AS	AT	AU	AV	AW	AX	AY
	END_AZIMUTH	END_LOCATION	BEGIN_LAT	BEGIN_LON	END_LAT	END_LON	EPISODE	N_EVENT	NAT_DATA_SOURCE
2	E	MARS HILL	31.2047	-90.7432	31.2705	-90.446	A strong co/A very large	25V	
3	W	CORNWELLS HGTS	40.0815	-74.9592	40.0822	-74.9599	Tropical Storm tornado	25V	
4	NW	FOL SOM	34.3951	-84.8631	34.3968	-84.8576	A line of thru a National	25V	
5	SSE	KINGS MTN	35.177	-81.413	35.196	-81.325	Unusually h/NWS storm	25V	
6							Hurricane HWHurricane H	25V	
7							Hurricane HWTCOON site	25V	
8							A tropical wt there was	25V	
9	E	ERBACON	38.5367	-80.5887	38.5186	-80.5378	Scattered ap County Road	25V	
10	NE	BARTLICK	37.1988	-82.5269	37.2685	-82.3034	An intense WAt the inter	25V	
11	S	WEIMERS MILL	40.1	-84.62	40.0942	-84.6748	Thunderstorm-high water	25V	
12	NW	KIRKERSVILLE	39.97	-82.62	39.969	-82.6201	Scattered thHigh water	25V	
13	NE	MACKSBURG	39.6391	-81.4691	39.6392	-81.4539	An intense WOn the West	25V	
14	NW	SAN JUAN	18.4148	-66.0901	18.4152	-66.0876	An upper-levFlooding rep	25V	
15	NW	SAN JUAN	18.414	-66.1016	18.4139	-66.101	An upper-levFlooding rep	25V	
16							Strong gradLow enforce	25V	
17	WNW	ALLENDALE OSWALD ARP	33.01	-81.4	33.01	-81.4	An area of h/Alendale C	25V	
18							Very strong Strong grad	25V	
19	WSW	FOLLY BEACH	32.65	-79.94	32.65	-79.94	A severe quThe Weather	25V	
20	E	SULLIVANS ISLAND	32.76	-79.82	32.76	-79.82	A severe quThe Weather	25V	
21	N	CHESAPEAKE BAY BRIDGE TUNNEL 3RD ISLAND	37.04	-76.08	37.04	-76.08	Scattered shWind gust	25V	
22	N	STANNARD ROCK LIGHT	47.18	-87.23	47.18	-87.23	A potent stoThe Stannard	25V	
23	V	V	F	F	F	F	V	V	Delete

Fatalities

	A	B	C	D	E	F	G	H	I	J	K
1	FAT_YEARMONTH	FAT_DAY	FAT_TIME	FATALITY_ID	EVENT_ID	FATALITY_TYPE	FATALITY_DATE	FATALITY_AGE	FATALITY_SEX	FATALITY_LOCATION	EVENT_YEARMONTH
2	202106	9	0	42960	953511	D	06/09/2021 00:00:00	70	M	Golfing	202106
3	202107	20	0	43206	961309	I	07/20/2021 00:00:00	32	M	Other	202107
4	202107	20	0	43207	961309	I	07/20/2021 00:00:00	31	M	Other	202107
5	202107	20	0	43208	961309	I	07/20/2021 00:00:00		M	Other	202107
6	202107	27	0	44279	964033	I	07/27/2021 00:00:00		F	Other	202107
7	202103	28	0	42898	951050	D	03/28/2021 00:00:00	61	F	In Water	202103
8	202104	24	0	42962	954336	D	04/24/2021 00:00:00	2		Permanent Home	202104
9	202106	10	0	42963	954408	D	06/10/2021 00:00:00		M	In Water	202106
10	202106	14	0	44478	970299	D	06/14/2021 00:00:00	23	M	Outside/Open Areas	202106
11	202106	13	0	44482	970319	D	06/13/2021 00:00:00	32	F	Outside/Open Areas	202106
12	202106	13	0	44483	970319	D	06/13/2021 00:00:00	42	M	Outside/Open Areas	202106
13	202106	14	0	44484	970319	D	06/13/2021 00:00:00	29	M	Outside/Open Areas	202106
14	202106	15	0	44485	970319	D	06/15/2021 00:00:00	34	F	Outside/Open Areas	202106
15	202106	17	0	44487	970319	D	06/17/2021 00:00:00	28	M	Outside/Open Areas	202106
16	202106	20	0	44488	970319	D	06/20/2021 00:00:00	19	M	Outside/Open Areas	202106
17	202106	20	0	44489	970319	D	06/20/2021 00:00:00	42	M	Outside/Open Areas	202106
18	202106	15	0	44490	970319	D	06/15/2021 00:00:00	35	M	Outside/Open Areas	202106
19	202106	17	0	44491	970319	D	06/17/2021 00:00:00	36	M	Outside/Open Areas	202106
20	202106	19	0	44492	970319	D	06/19/2021 00:00:00	54	M	Outside/Open Areas	202106
21	202106	13	0	44479	970320	D	06/13/2021 00:00:00	35	F	Outside/Open Areas	202106
22	202106	12	0	44480	970320	D	06/12/2021 00:00:00	22	M	Outside/Open Areas	202106
23	202106	15	0	44481	970320	D	06/15/2021 00:00:00	37	M	Outside/Open Areas	202106
24	202106	2	0	43761	970487	D	06/02/2021 00:00:00		M	Outside/Open Areas	202106
25	202106	2	0	43762	970487	D	06/02/2021 00:00:00	20	M	Outside/Open Areas	202106
26	202106	7	0	43770	970488	D	06/07/2021 00:00:00	40	M	Outside/Open Areas	202106
27	202106	6	0	43771	970490	D	06/06/2021 00:00:00	47	M	Outside/Open Areas	202106
28											
29											
30											
31											
32											
33											
34											
35											
36											
37											
38											
39											
40											
41											
42											
43											
44											
45											
46											
47											
48											
49											
50											
51											
52											
53											
54											

Redundant
(Will remove after
verifying equivalent
to FATALITY_DATE)

I I V DT I V V I

Datatypes:
I - Int
V - Varchar
DT - Datetime

The event can
occur prior to
the fatality, so
this is not
redundant

After running transform.py, both tables are fully populated in mysql:

```
mysql> select count(*) from fatalities;
+-----+
| count(*) |
+-----+
| 19288 |
+-----+
1 row in set (0.01 sec)

mysql> select count(*) from details;
+-----+
| count(*) |
+-----+
| 1682037 |
+-----+
1 row in set (0.64 sec)

mysql> select BEGIN_YEARMONTH,BEGIN_DAY,BEGIN_TIME,END_YEARMONTH,END_DAY,END_TIME, EPISODE_ID,EVENT_ID,STATE,STATE_FIPS,EVENT_TYPE,CZ_TYPE from details limit 5;
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| BEGIN_YEARMONTH | BEGIN_DAY | BEGIN_TIME | END_YEARMONTH | END_DAY | END_TIME | EPISODE_ID | EVENT_ID | STATE | STATE_FIPS | EVENT_TYPE | CZ_TYPE |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 197009 | 4 | 1650 | 197009 | 4 | 1650 | NULL | 9881965 | ARKANSAS | 5 | Thunderstorm Wind | C |
| 197010 | 8 | 1655 | 197010 | 8 | 1655 | NULL | 10058129 | MISSOURI | 29 | Tornado | C |
| 197006 | 4 | 1615 | 197006 | 4 | 1615 | NULL | 10047491 | MISSISSIPPI | 28 | Thunderstorm Wind | C |
| 197006 | 11 | 2200 | 197006 | 11 | 2200 | NULL | 10056944 | MISSOURI | 29 | Thunderstorm Wind | C |
| 197007 | 1 | 2100 | 197007 | 1 | 2100 | NULL | 10052308 | MINNESOTA | 27 | Thunderstorm Wind | C |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
5 rows in set (0.00 sec)

mysql> select CZ_FIPS,CZ_NAME,WFO,CZ_TIMEZONE,INJURIES_DIRECT,INJURIES_INDIRECT,DEATHS_DIRECT,DEATHS_INDIRECT,DAMAGE_PROPERTY,DAMAGE_CROPS,SOURCE,MAGNITUDE,MAGNITUDE_TYPE,FLOOD_CAUSE,CATEGORY from details limit 5;
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| CZ_FIPS | CZ_NAME | WFO | CZ_TIMEZONE | INJURIES_DIRECT | INJURIES_INDIRECT | DEATHS_DIRECT | DEATHS_INDIRECT | DAMAGE_PROPERTY | DAMAGE_CROPS | SOURCE | MAGNITUDE | MAGNITUDE_TYPE | FLOOD_CAUSE | CATEGORY |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 93 | MISSISSIPPI | NULL | CST | 0 | 0 | 0 | 0 | 0 | 0 | 0 | NULL | 0 | NULL | NULL |
| 197 | SCHUYLER | NULL | CST | 0 | 0 | 0 | 0 | 25K | 0 | 0 | NULL | 0 | NULL | NULL |
| 159 | WINSTON | NULL | CST | 0 | 0 | 0 | 0 | 0 | 0 | 0 | NULL | 0 | NULL | NULL |
| 9 | BARRY | NULL | CST | 0 | 0 | 0 | 0 | 0 | 0 | 0 | NULL | 0 | NULL | NULL |
| 93 | MEEKER | NULL | CST | 0 | 0 | 0 | 0 | 0 | 0 | 0 | NULL | 0 | NULL | NULL |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
5 rows in set (0.00 sec)

mysql> select BEGIN_RANGE,BEGIN_AZIMUTH,BEGIN_LOCATION,END_RANGE,END_AZIMUTH,END_LOCATION,BEGIN_LAT,BEGIN_LON,END_LAT,END_LON from details limit 5;
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| BEGIN_RANGE | BEGIN_AZIMUTH | BEGIN_LOCATION | END_RANGE | END_AZIMUTH | END_LOCATION | BEGIN_LAT | BEGIN_LON | END_LAT | END_LON |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 0 | NULL | NULL | 0 | NULL | NULL | 36 | -90 | NULL | NULL |
| 0 | NULL | NULL | 0 | NULL | NULL | 40 | -93 | 40 | -93 |
| 0 | NULL | NULL | 0 | NULL | NULL | 33 | -89 | NULL | NULL |
| 0 | NULL | NULL | 0 | NULL | NULL | 37 | -94 | NULL | NULL |
| 0 | NULL | NULL | 0 | NULL | NULL | 45 | -94 | NULL | NULL |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
5 rows in set (0.00 sec)

mysql> select EVENT_NARRATIVE,EPISODE_NARRATIVE from details ORDER BY BEGIN_YEARMONTH desc limit 1;
+-----+-----+
| EVENT_NARRATIVE | EPISODE_NARRATIVE |
+-----+-----+
| In Danvers, water was up to the floor boards of a SUV (about 8 to 12 inches deep) on Village Post Road. | Strong southerly wind flow out ahead of an advancing cold front produced strong to damaging wind gusts and some heavy rain that caused some street and basement flooding. |
+-----+-----+
1 row in set (2.29 sec)

mysql>
```

Data is returning meaningful queries for analysis:

phpMyAdmin

Recent Favorites

New

Capstone

New

details

Columns

New

BEGIN_AZIMUTH

BEGIN_DAY

BEGIN_LAT

BEGIN_LOCATION

BEGIN_LON

BEGIN_RANGE

BEGIN_TIME

BEGIN_YEARMON

CATEGORY

CZ_FIPS

CZ_NAME

CZ_TIMEZONE

CZ_TYPE

DAMAGE_CROPS

DAMAGE_PROPER

DEATHS_DIRECT

DEATHS_INDIREC

END_AZIMUTH

END_DAY

END_LAT

END_LOCATION

END_LON

END_RANGE

END_TIME

END_YEARMONTH

EPISODE_ID

Server: localhost:3306 » Database: Capstone » Table: details

Browse Structure SQL Search Insert Export Import Privileges Operations Triggers

Show query box

⚠ Current selection does not contain a unique column. Grid edit, checkbox, Edit, Copy and Delete features are not available.

✔ Showing rows 0 - 24 (70 total, Query took 8.8693 seconds.)

```
SELECT EVENT_TYPE,SUM(DEATHS_DIRECT),SUM(DEATHS_INDIRECT),SUM(INJURIES_DIRECT),SUM(INJURIES_INDIRECT) FROM details GROUP BY 1 ORDER BY 2 DESC,3 DESC, 4 DESC, 5 DESC
```

1 > >> ☐ Show all Number of rows: 25 Filter rows:

+ Options

EVENT_TYPE	SUM(DEATHS_DIRECT)	SUM(DEATHS_INDIRECT)	SUM(INJURIES_DIRECT)	SUM(INJURIES_INDIRECT)
Tornado	3648	37	64662	268
Heat	2533	246	10293	210
Flash Flood	1695	60	6469	58
Rip Current	1140	8	813	2
Hurricane (Typhoon)	1097	25	1382	2409
Thunderstorm Wind	1018	87	11244	375
Excessive Heat	1005	152	5815	68
Lightning	887	47	5222	322
Flood	598	57	2387	51
Cold/Wind Chill	590	52	298	31
Avalanche	383	6	247	5
High Wind	350	93	1532	438
Wildfire	324	47	2171	454
High Surf	262	17	329	6
Strong Wind	243	64	551	74
Winter Storm	237	302	1444	1690
Extreme Cold/Wind Chill	206	34	26	1
Winter Weather	163	777	2368	3620
Heavy Snow	124	127	712	582
Debris Flow	121	3	242	2
Heavy Rain	119	109	279	355
Tropical Storm	109	87	391	20
Ice Storm	98	84	334	1487
Blizzard	97	90	411	141
Dense Fog	69	148	852	696