Predicting severe storm events in Louisiana

Jenny Rhee

Problem Statement

The goal of this capstone project is to predict the number of storm events in Louisiana per month. Building this model can potentially assist cities and the general population in preparation for severe storm events (e.g., infrastructure planning, stocking up on supplies, insurance, familiarizing with evacuation routes, etc.). People will likely have experience with these events, but the prevalence is expected to increase with climate change. It is important to begin understanding the rising trends because two or more severe storm events within a small window of time can be exceptionally more devastating (e.g., a severe flooding with a hurricane in the following days with no time to recover from the flood).

Data

The National Weather Service provides storm data containing statistics on personal injuries and damage estimates from 1950 to present. There are 34 different storm events including hurricanes, thunderstorms, hail, etc. There are 51 columns including damage, injuries, deaths, etc. I used Python scripts to download all 213 .csv files, create a database, and import the data into the database. Additional supporting data includes historical meteorological data to analyze any potential correlations. NOAA's National Centers for Environmental Information has a Global Summary of the Month dataset, which includes 48 meteorological variables, from 1891 to present at numerous stations across the state.

Data Cleaning Summary

Storm Events Data

Parish

The parishes were in all caps, so they were reformatted to title case. There were two odd parish names, 'Laz038>040 - 056>070' and 'Laz067 - 070', for only two rows, so I dropped them. There was also a typo, 'Rapdies' (supposed to be 'Rapides) and was corrected.

There were several entries for parish that weren't exact parish names, mainly including a region within the parish (e.g., East Cameron vs Cameron). I used the FIPS DataFrame to check the FIPS code for each of the entries and compared them with the actual parish FIPS codes. Then

I checked for entries with coordinates and only one had them, Sabine and Natchitoches. I used the coordinates2politics API to find the parish based on the coordinates, and Sabine was returned. Some of the entries had invalid FIPS codes (from comparing with the Louisiana parish Wikipedia). I changed the names to remove the region and contain only the parish.

I then looked at the remaining null parishes (52 rows) and tried to find the parish using the coordinates2politics API again. 14 rows returned no parish data, and they were dropped. Finally, I mapped parishes to the correct FIPS code to create a FIPS column and dropped the latitude and longitude columns. There were a few rows with null FIPS, but they were counties outside of Louisiana and were dropped.

Damage

The property damage column consisted of strings with nulls, 0, or values appended by K (thousand), M (million), and B (billion). I converted them into the correct numerical values.

Monthly Aggregates by Parish

Finally, I needed to resample each parish by month. Event type was aggregated by count and renamed to event count while the remaining columns were aggregated by sum. The date column was converted to month-year periods.

Meteorological Data

I downloaded 18 variables that I thought may be of interest for this dataset including monthly average wind speed, different temperature variables, different precipitation variables, etc.

Parish/FIPS from Coordinates

I created a new DataFrame with name, latitude, and longitude from the original DataFrame and dropped duplicates. I used multi-processing (`pandarallel`) to use the coordinates2poltiics API to query for each parish based on the coordinates of each station. I then added a FIPS code column by mapping with the previous parish-FIPS dictionary. Finally, this DataFrame was merged with the original DataFrame, and the station, latitude, longitude, and elevation columns were dropped. The stations with null FIPS after this process were primarily out of state with the exception of two that were either an uninhabited parish or a bridge. These rows were dropped.

Missing Values

I looked at the percentage of missing data for each column and further inspected the variable if more than 20% was missing. I counted how many parishes had data for each variable

and if more than half of the parishes was missing, I dropped the variable. To further investigate, I checked which parishes had missing data for each variable and interestingly, St. James and Assumption were the only two with missing data for each of the variables.

Aggregating Station Data to Parish

I aggregated all the stations within a parish by calculating the means for each month and again used multi-processing.

Merged Storm Events and Meteorological DataFrames

I did an outer merge to include meteorological data where there was no storm data and filled NaN storm data values with 0 (assuming that no storm data for a particular month means there was no storm events in the parish for that month) and missing meteorological data were handled later. I dropped columns that were likely not useful and/or were missing a significant amount of data (cooling and heating degree days, extreme min and max temperature, number of days >= 1.0 inches of precipitation in the month, highest daily snow depth, highest daily snowfall).

Then, to handle the remaining missing data, I found the means for each variable for each parish and month. Missing values were replaced using the means. After this, there were still some rows with missing data, and as discovered earlier, were from Assumption and St. James parishes. The number of rows for these parishes and the number of rows with missing values were equal, so these parishes were dropped. I also renamed the meteorological columns into a more readable form ('highest_daily_total_precip', 'total_precip', 'avg_temp', ‘mean_max_temp’, ‘mean_min_temp’).

Parish Coordinates Data

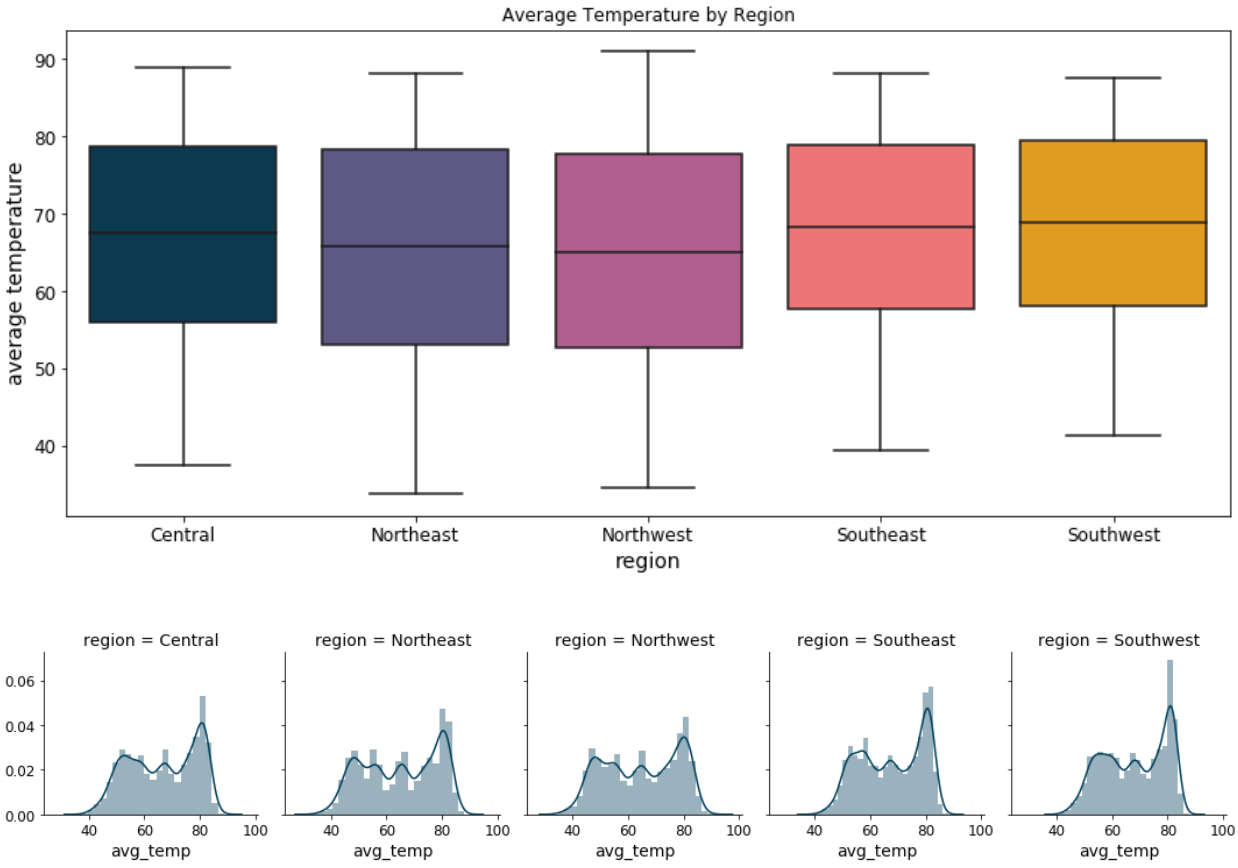
I added interpolated centroid coordinates for each parish (Source). The coordinates had to be formatted to remove the degree and plus symbols and then convert the string values into floats.

Region Data

I created a region column (Source) - Northeast, Northwest, Central, Southeast, Southwest. Finally, I converted the parish and region parishes to category data type.

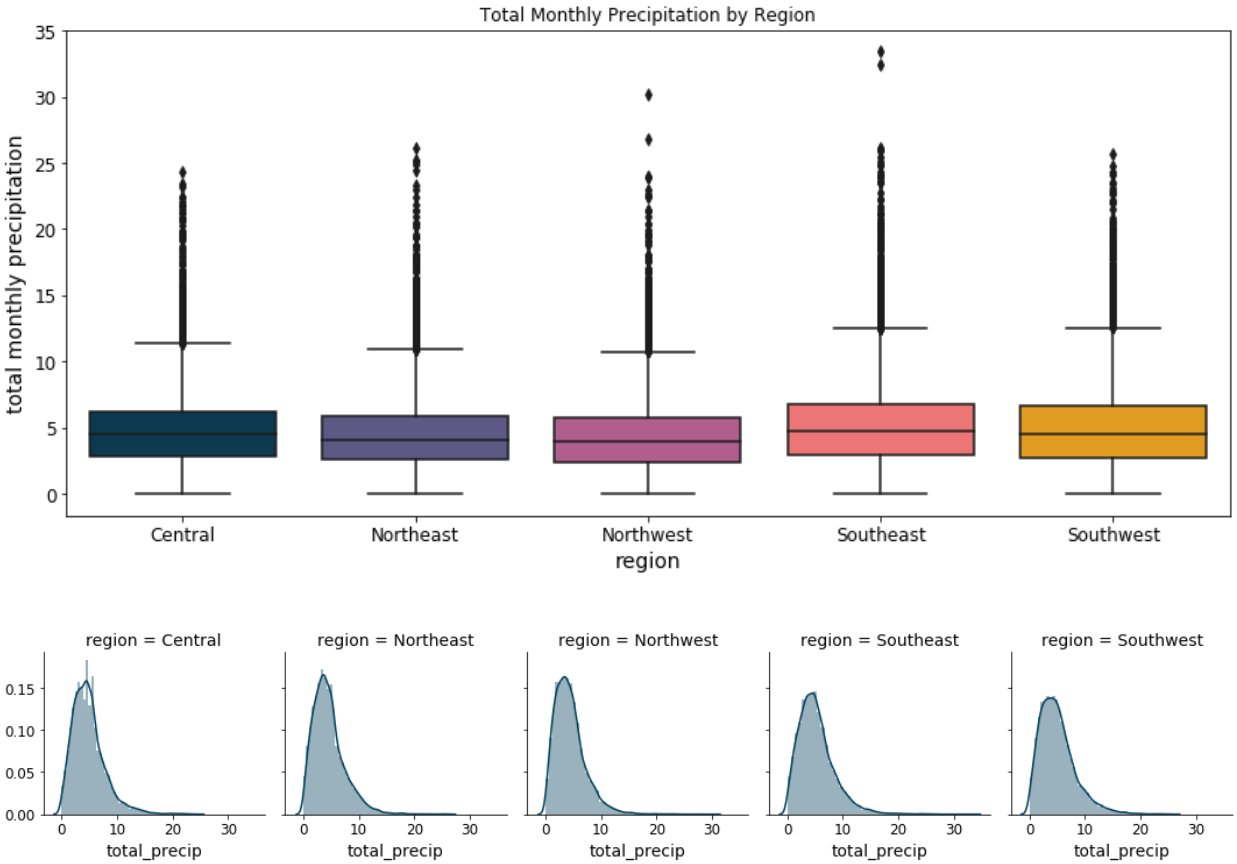
Exploratory Analysis Summary

Temperature



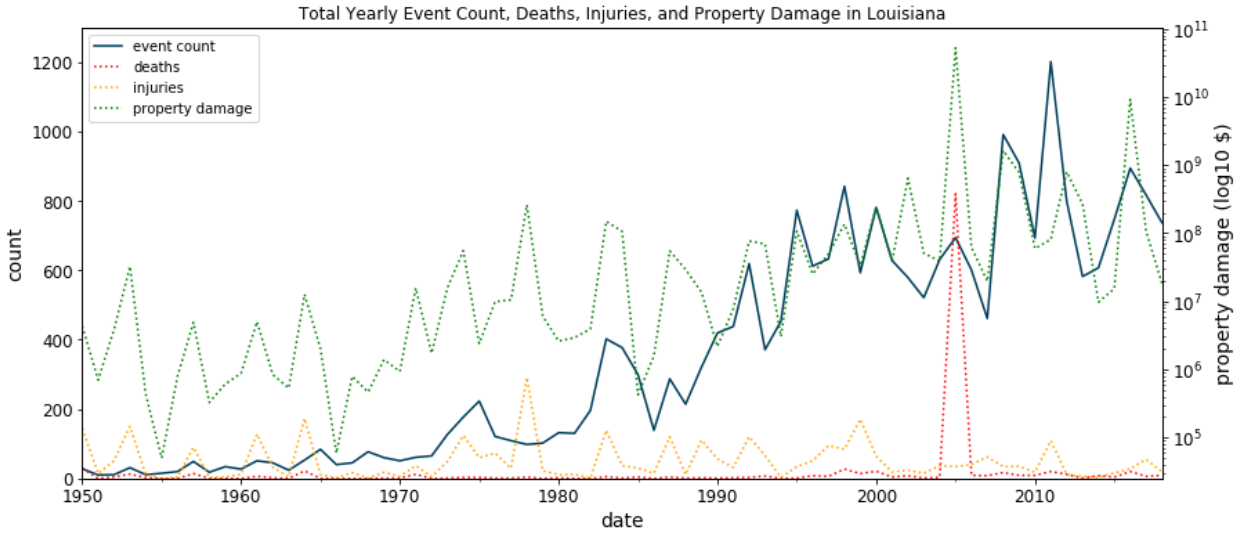
As expected, temperature distributions are similar throughout the state, but Southern regions seem to be slightly higher with less variability than other regions. To test for regional differences, I set my null hypothesis to be that Northern regions and Southern regions have the same average temperatures. My alternative hypothesis was that Northern regions have lower average temperatures than Southern regions with a significance level of 0.05. I ran a two sided t-test for two samples and obtained t-statistic = -25.29 and p-value = $5.46e^{-140}$, which allowed me to reject the null hypothesis.

Precipitation



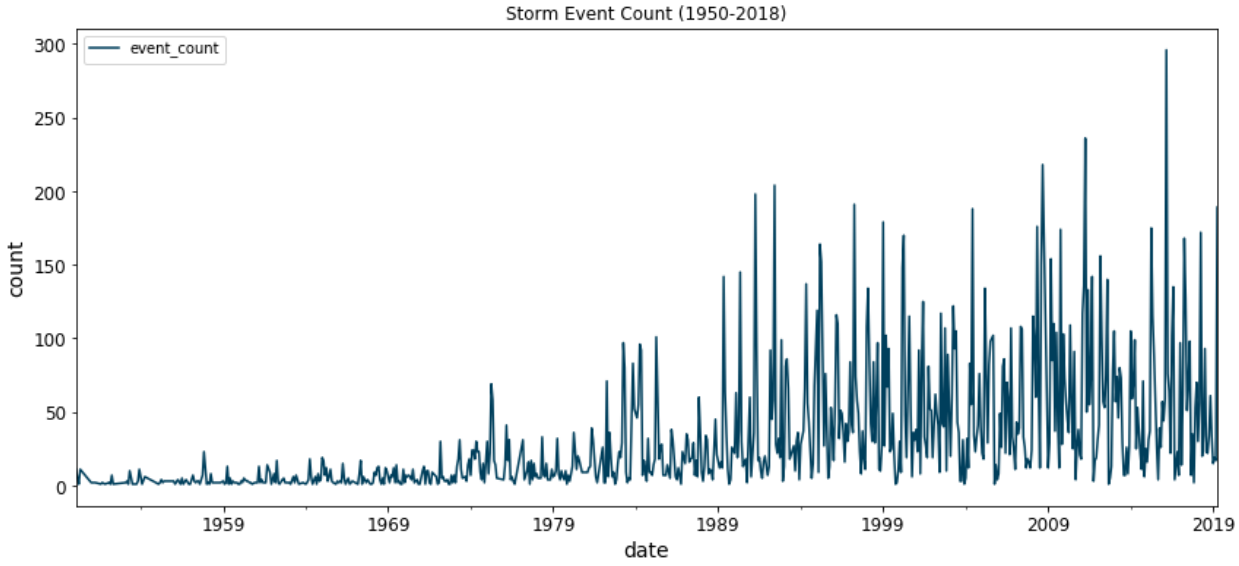
Again, total monthly precipitation is similar throughout the state, but Southern regions seem to have slightly more volumes. To test for regional differences, I set my null hypothesis to be that Northern regions and Southern regions have the same amounts of total monthly precipitation. My alternative hypothesis was that Northern regions have lower amounts of total monthly precipitation compared to Southern regions with a significance level of 0.05. Again, I ran a two-sided t-test for two samples and obtained t-statistic = -22.04 and p-value = $4.51e^{-107}$, which allowed me to reject the null hypothesis.

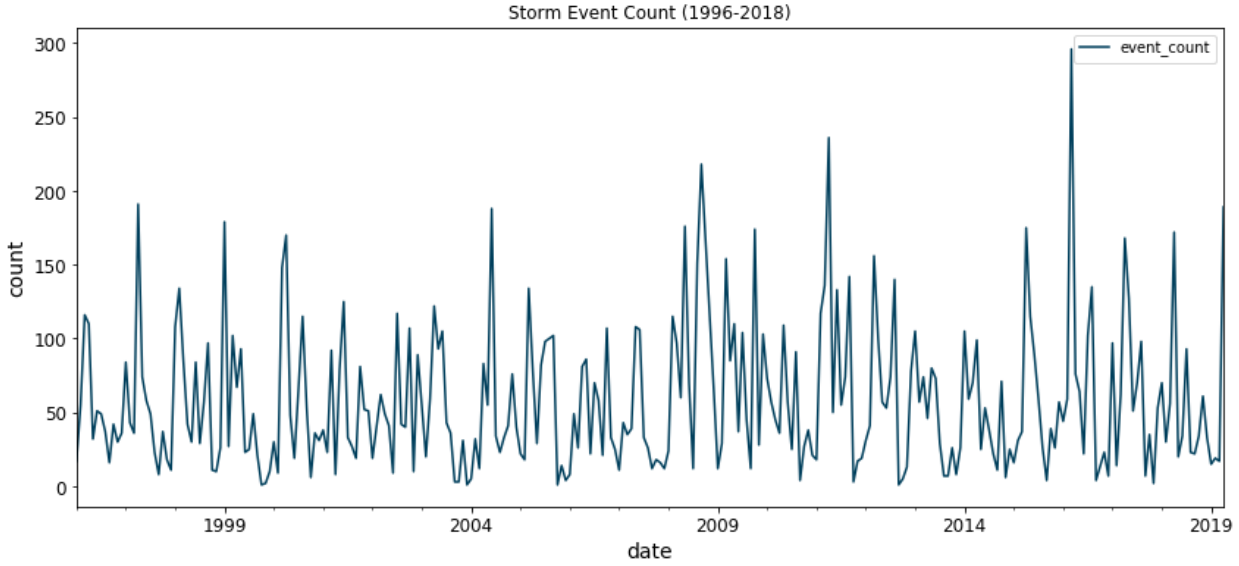
Total Yearly Event Count, Deaths, Injuries, and Property Damage



There is an increasing trend in number of storm events with potentially interesting peaks during 1973-1975, 1983-1985, 2008, 2011, and 2015-2018. The spikes in property damage and deaths were in 2005 when Hurricane Katrina devastated South Louisiana. Otherwise, death count was generally low. The second highest peak in property damage was in 2016 when there was a 1000-year flood in Central Louisiana. I further explored the following questions:

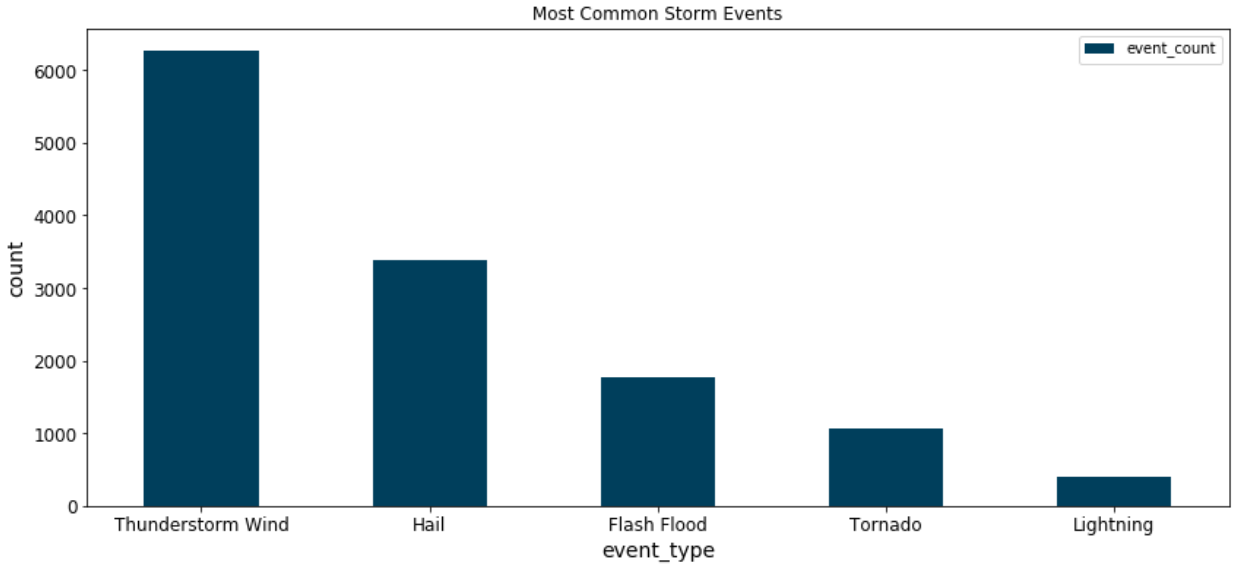
1. What could be the reason(s) for such a drastic change in the number of storm events from 1950 to now? (Source)
    a. Tornado: From 1950 through 1954, only tornado events were recorded.
    b. Tornado, Thunderstorm Wind and Hail: From 1955 through 1992, only tornado, thunderstorm wind and hail events were keyed from the paper publications into digital data. From 1993 to 1995, only tornado, thunderstorm wind and hail events have been extracted from the Unformatted Text Files.
    c. **All Event Types (48 from Directive 10-1605): From 1996 to present, 48 event types are recorded as defined in NWS Directive 10-1605.**





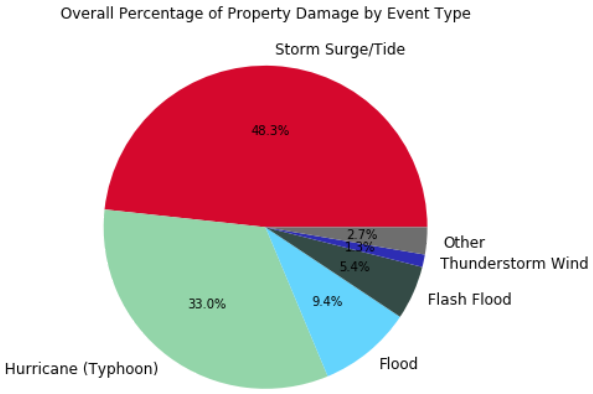
For modeling and the remaining analyses, I will only use data from 1996 to present.

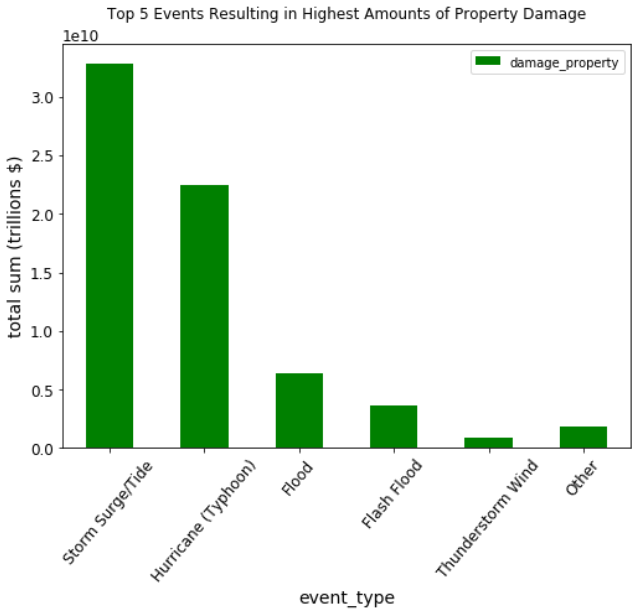
2. What are the most common storm event types?



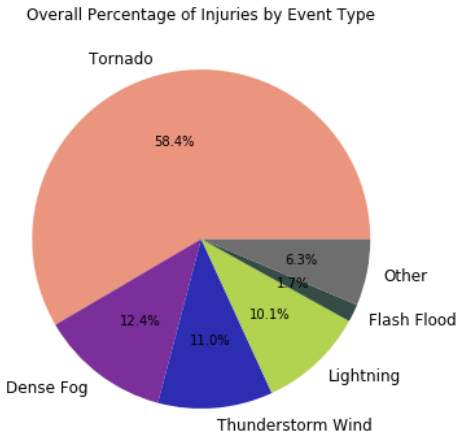
3. Which event types have resulted in the most amount of property damage, injuries, and deaths?

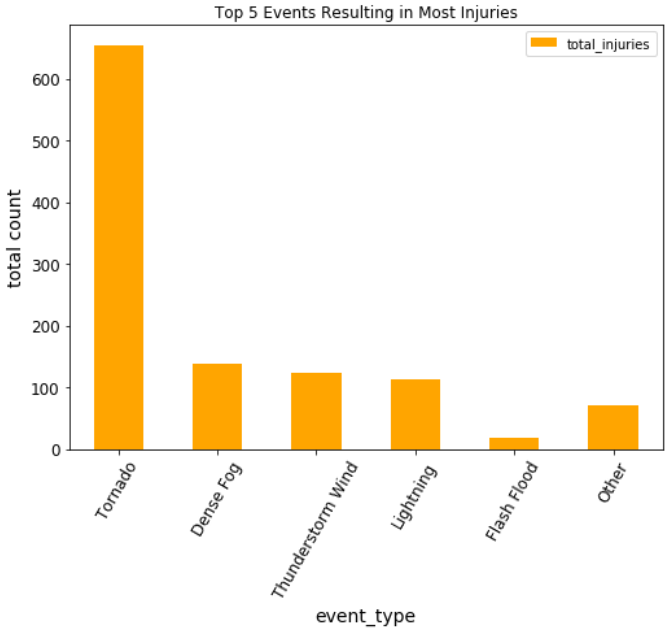
Property Damage



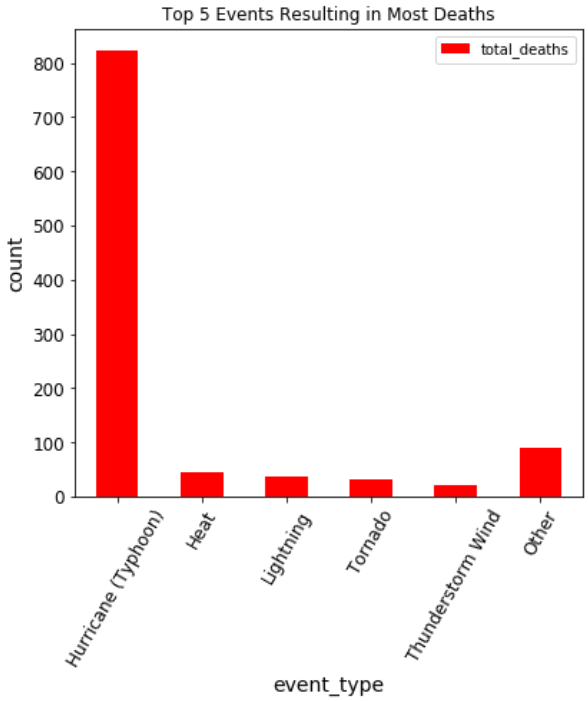


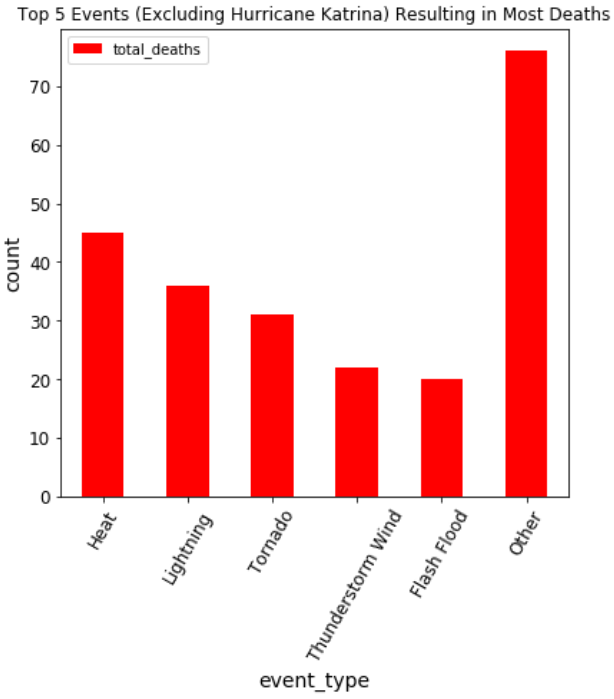
Injuries

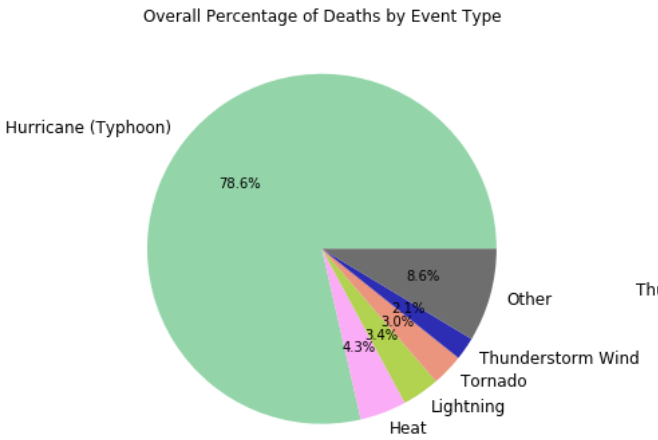


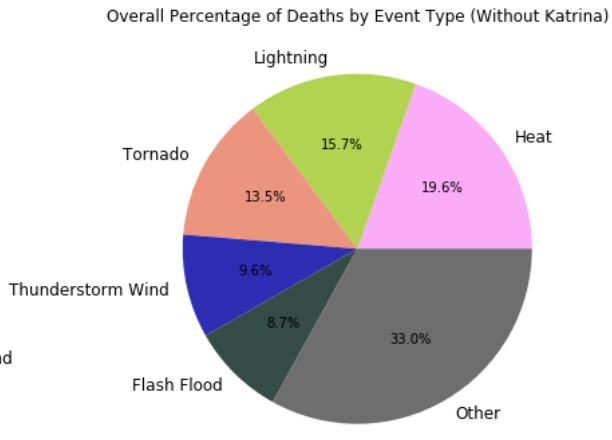


Deaths

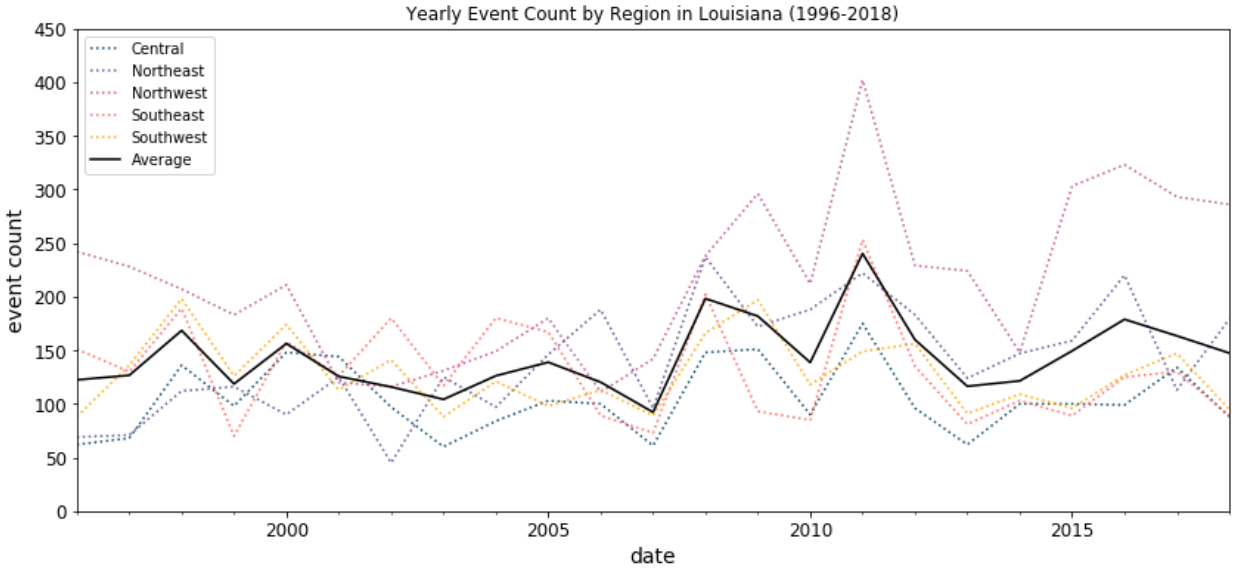






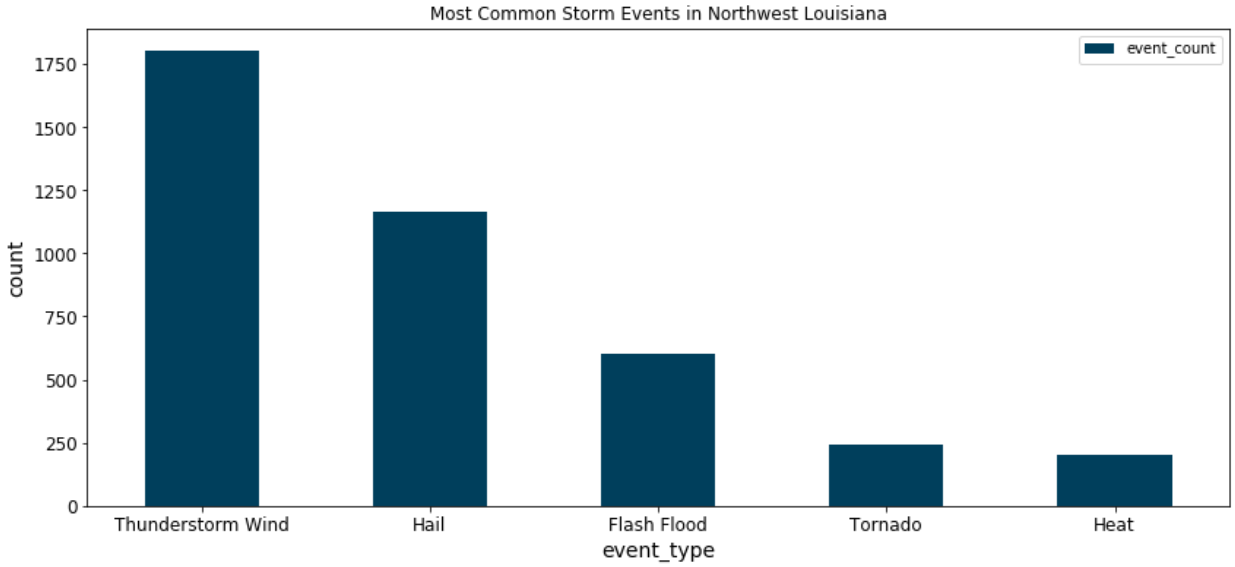


Yearly Event Count by Region

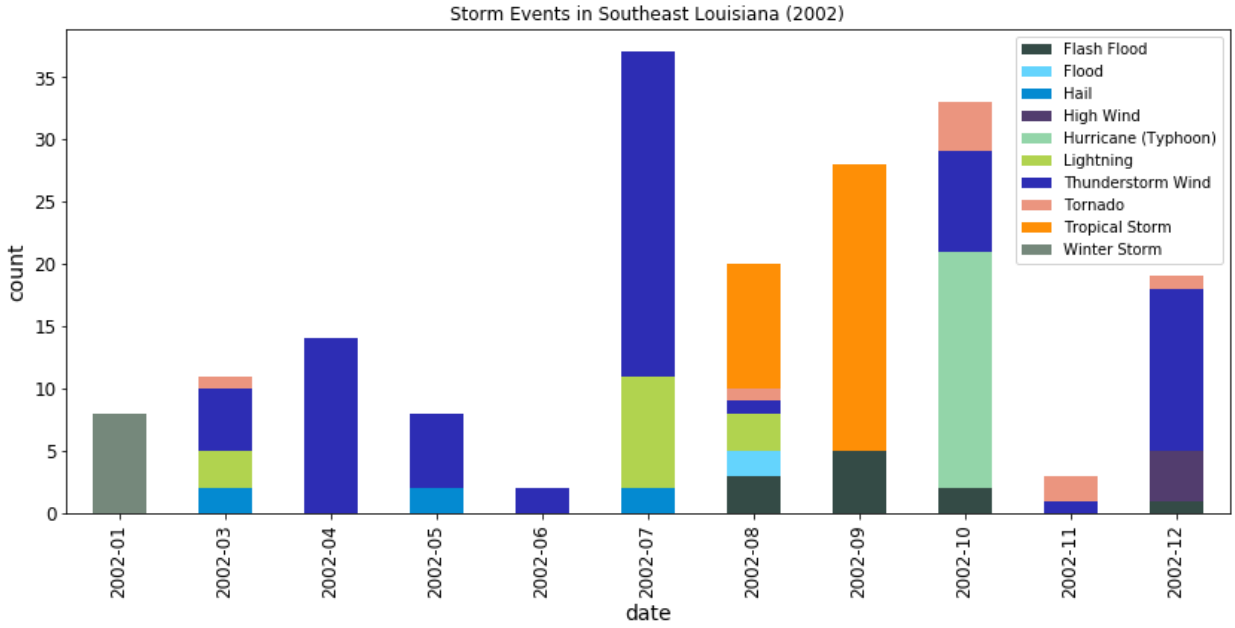


Northwest Louisiana has consistently had the most number of storm events (with some exceptions) while Central Louisiana generally has had the least. The trends are relatively similar among all regions with Northwest Louisiana being the exception. I explored the following questions:

1. It is surprising that Northwest Louisiana consistently has the most number of storm events. What type of storm events are common for this region?

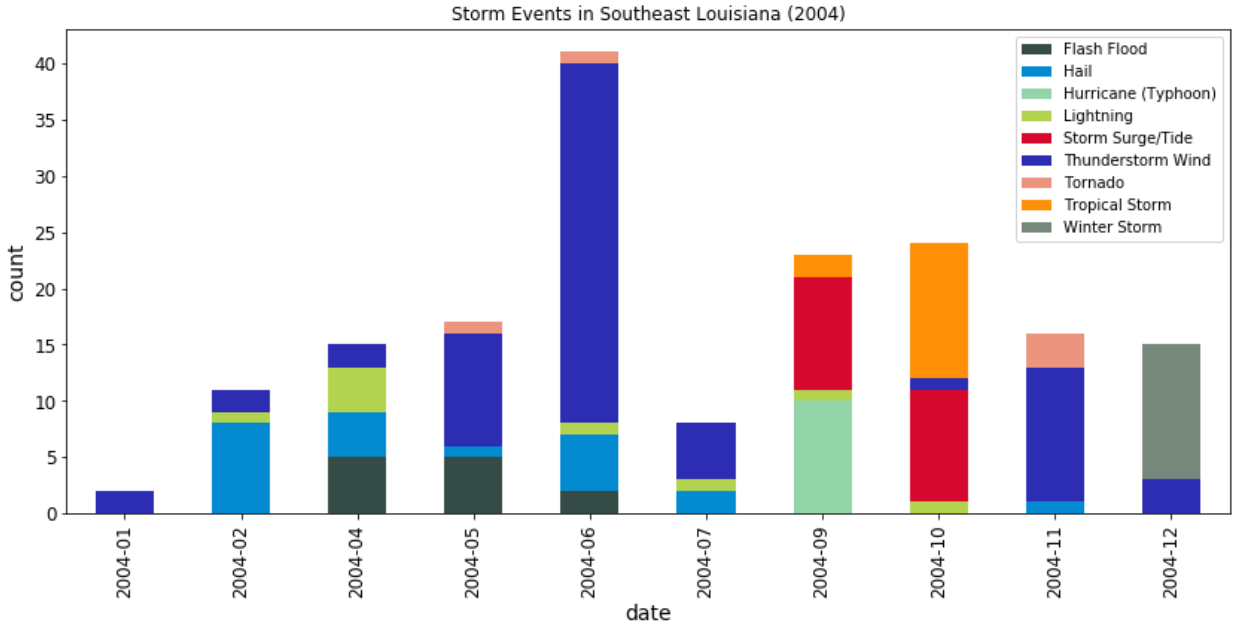


2. What happened in Southeast Louisiana in 2002 and 2004?



- **Winter storms** in January
    - [Snow hits an unpracticed South and shuts it down]
- Relatively high number of **lightning** events in July
- High number of **thunderstorm winds** especially in July, but also notably in April and December

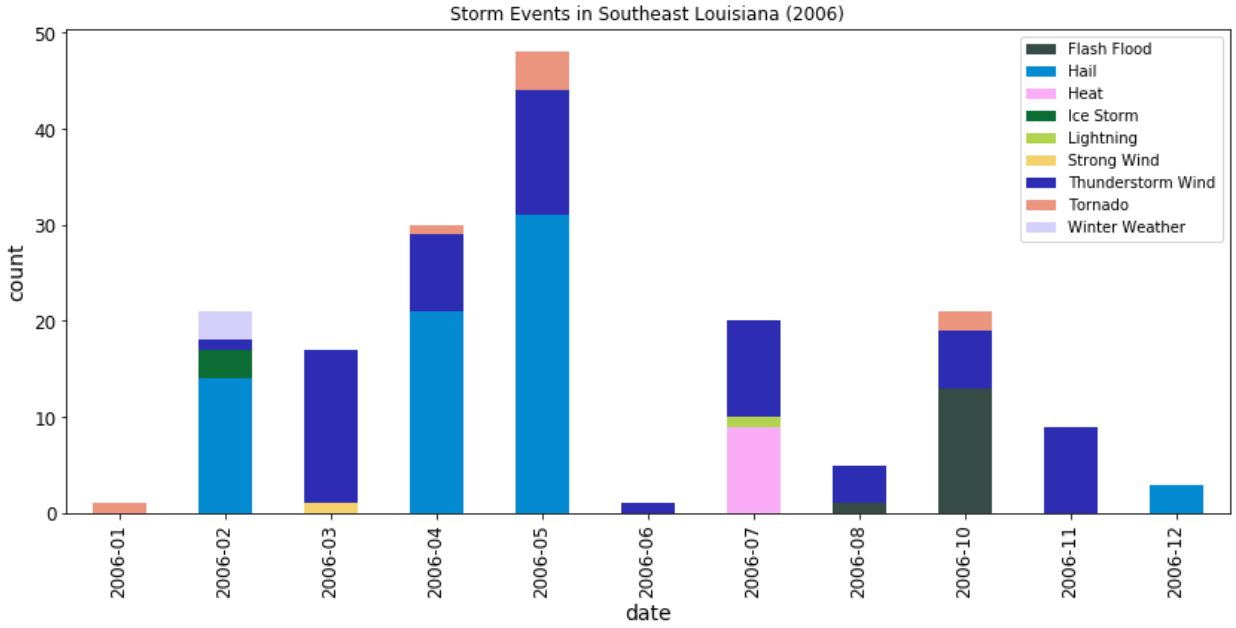
- **Tropical storms** in August and September
    - August 5, Tropical Storm Bertha - Damage in Louisiana totaled to $150,000 (2002 USD).
    - September 5, Tropical Storm Fay
    - September 14, Tropical Storm Hanna
    - September 26, Tropical Storm/Hurricane Isidore
- **Hurricanes** in October
    - October 3, Hurricane Lili caused over $790 million (2002 USD) in damage to Louisiana
- Tropical Storm/Hurricane Source



- Relatively high number of **hail** events in February and also notably in April and June
- **Flash floods** mainly in April and May
    - Stormy weather caused flooding in May over Southeast Louisiana and South Mississippi
- High number of **thunderstorm winds** especially in June, but also in May and November
- **Tropical storms**, **hurricanes**, and **storm surges/tides** in September and October
    - September, Hurricane Ivan - Four people died during evacuation and damage in the state reached roughly $7.9 million. Upon the storm's second landfall in Holly Beach resulted in minor coastal flooding, with damage totaling only about $15,000.
    - October, Tropical Storm Matthew caused $255,000 in damage (2004 USD).
- **Winter storms** in December
    - The Great Christmas Eve Storm of 2004

- [Tropical Storm/Hurricane Source]

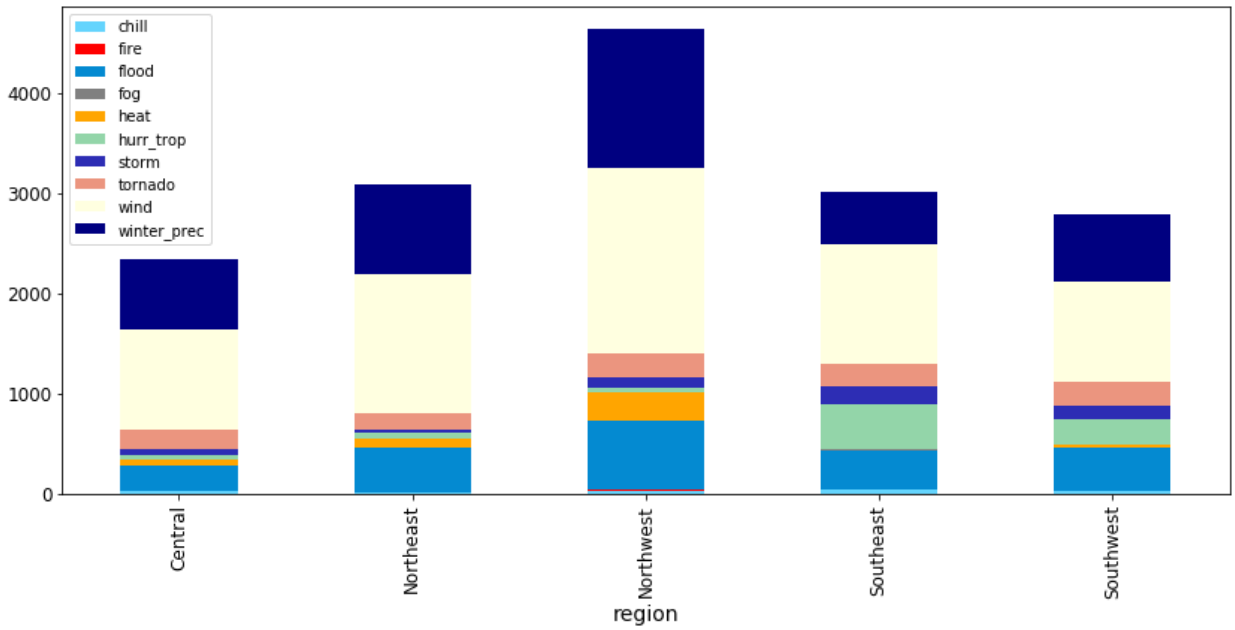
3. What happened in Northeast Louisiana in 2006?



- High number of **hail** events in February, April, and May
- **Thunderstorm events** throughout the year
- **Heat** in July

Categorizing Event Types

For the model, it makes sense to only look at events that have caused significant amounts of property damage, injuries, or deaths in the past. This excluded the following events: dust devils, funnel clouds, droughts, freezing fogs, seiches, waterspouts, rip currents, and astronomical low tides. I added another column to further generalize some events (flash floods, floods, coastal floods → flood), 26 remaining events into 10 categories. The following visualization is the number of events in each category from 1996 to present for each region.



Next Steps

I will be building a model to predict future number of severe storm events in Louisiana, which will likely involve some feature engineering.