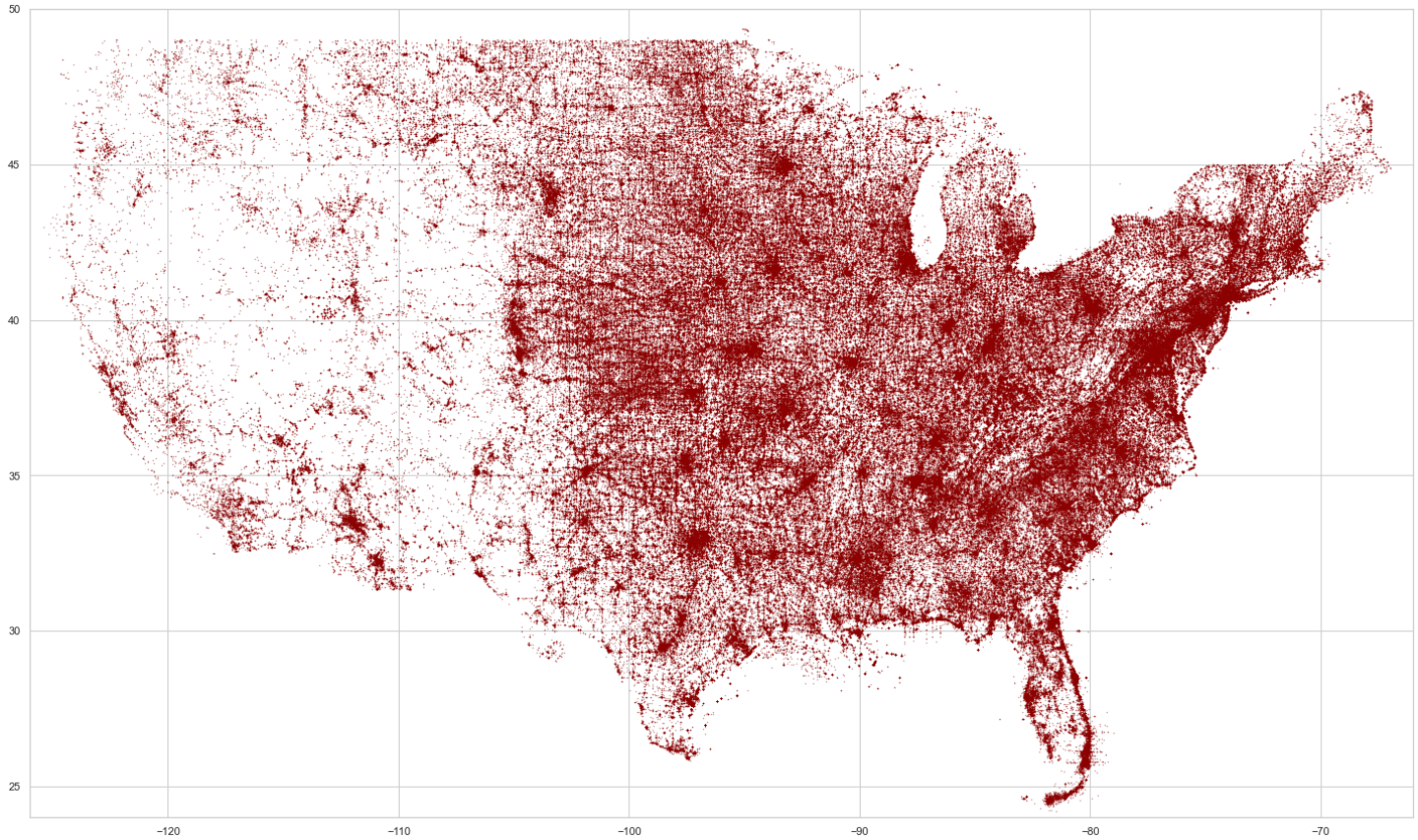
Predicting Property Damage From Storms - Summary Report



Predicting Property Damage From Storms - Summary Report

Github Repository

Tables of Contents

Introduction

Over 50,000 storm events occur in the United States every year. While the vast majority of storms do not cause any damage or fatalities, there are a number that cause significant damage. Over $20 billion in property damage occurs in the United States every year due to storms, with 2021 and 2022 being some of the most damaging years on record. According to the Office for Coast Management, the total cost of storm damage in the last five years is more than one-third of the disaster cost total of the last 42 years.

Changes in climate and weather are just one factor in this trend. Additionally, the increase in population and property costs over the last several decades is also a contributing factor. Much of this growth has taken place in vulnerable areas like coasts, further deepening the impact of increasingly dangerous storms.

Problem Statement & Objective

Question: Is it possible to predict the amount of property damage that a storm will cause, given the characteristics of that storm?

The objective of this project was to investigate the characteristics of storms that cause property damage. The end result is a predictive model where a storm can be input into the model, and the model will return how much property damage will be caused by the storm.

Potential users of the model

- Insurance companies - assess potential claims that could be caused by storms
- Emergency responders - identifying severe storm characteristics could help anticipate how many resources will be required

Data

| Data Collection and Refinement Notebook |

Data sources:

- Storm files were collected from the Iowa Environmental Mesonet
- Population density (census.gov)
- Home price index (FHFA.gov)

Data highlights:

- Collected information on 1,555,648 storms; from 2000 - 2022
- The dataset also contains a huge number of storms that did not have any property damage - 78% of observations do not have any damage.
- It also contains a high number of outliers - storms that caused incredible amounts of damage. There were 513 storms that were more than three standard deviations away from the mean.
- The high number of outliers and high number of zeros make this a fairly irregular dataset.

Data that I utilized for each storm:

- Storm details
    - Storm ID, date, state, storm type, storm begin time, storm end time, direct & indirect injuries, crop damage, storm magnitude, flood cause, category, tornado scale, tornado size, storm latitude and longitude, event narrative

- Population information - county population & population density

- Home price index - proxy for property values

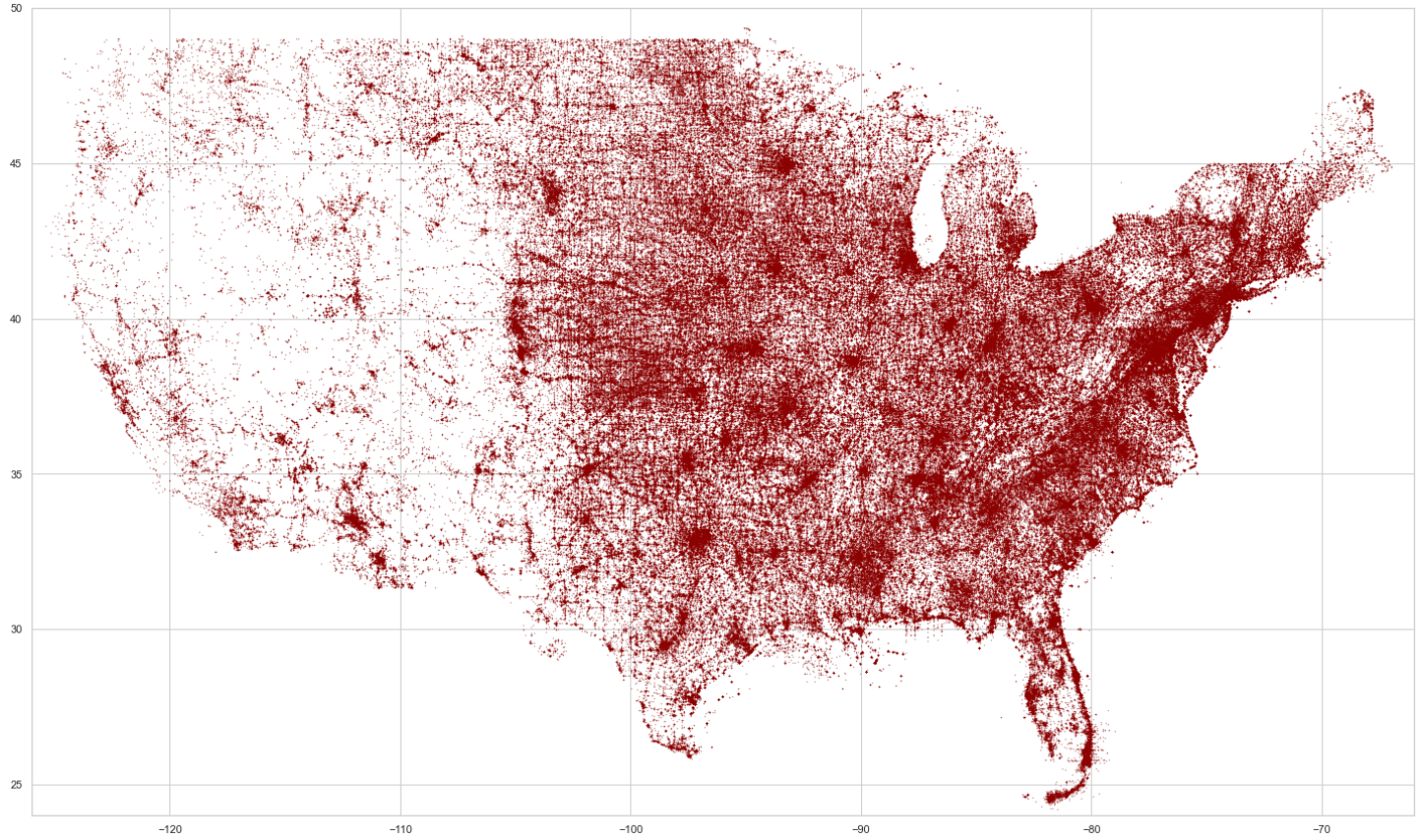
- Features that I created:
    - Concatenated state and county codes for identification
    - Storm duration
    - Storm area
    - County population density (merged from additional dataset)
    - Land values (merged from additional dataset)

Exploratory Data Analysis

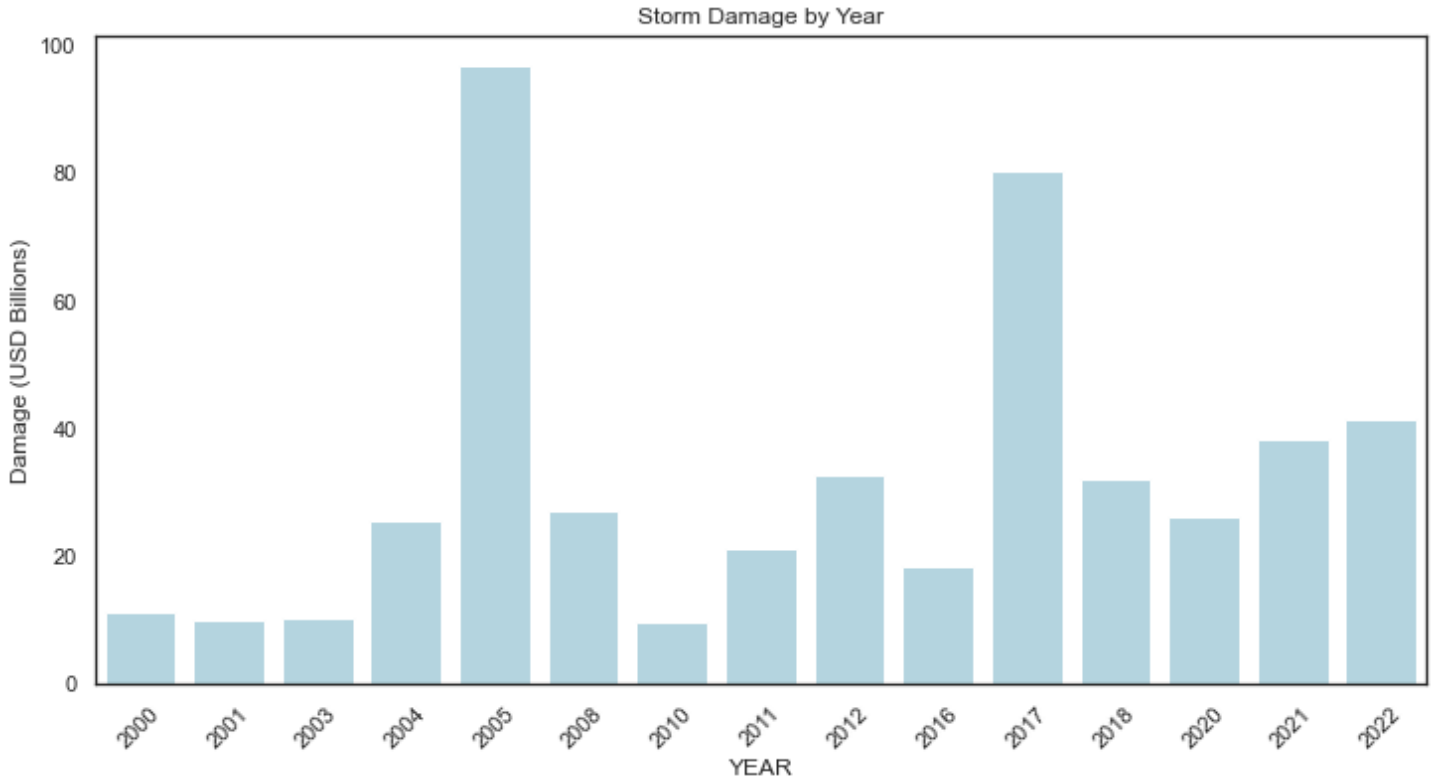
| EDA Notebook |

Below I will highlight a few of the many interesting trends that emerged from the data. A few general trends emerged: the prevalence of storms in late spring and summer months, high concentration of storms in the Midwest and Central states, and majority of damage concentrated in the Southeast.

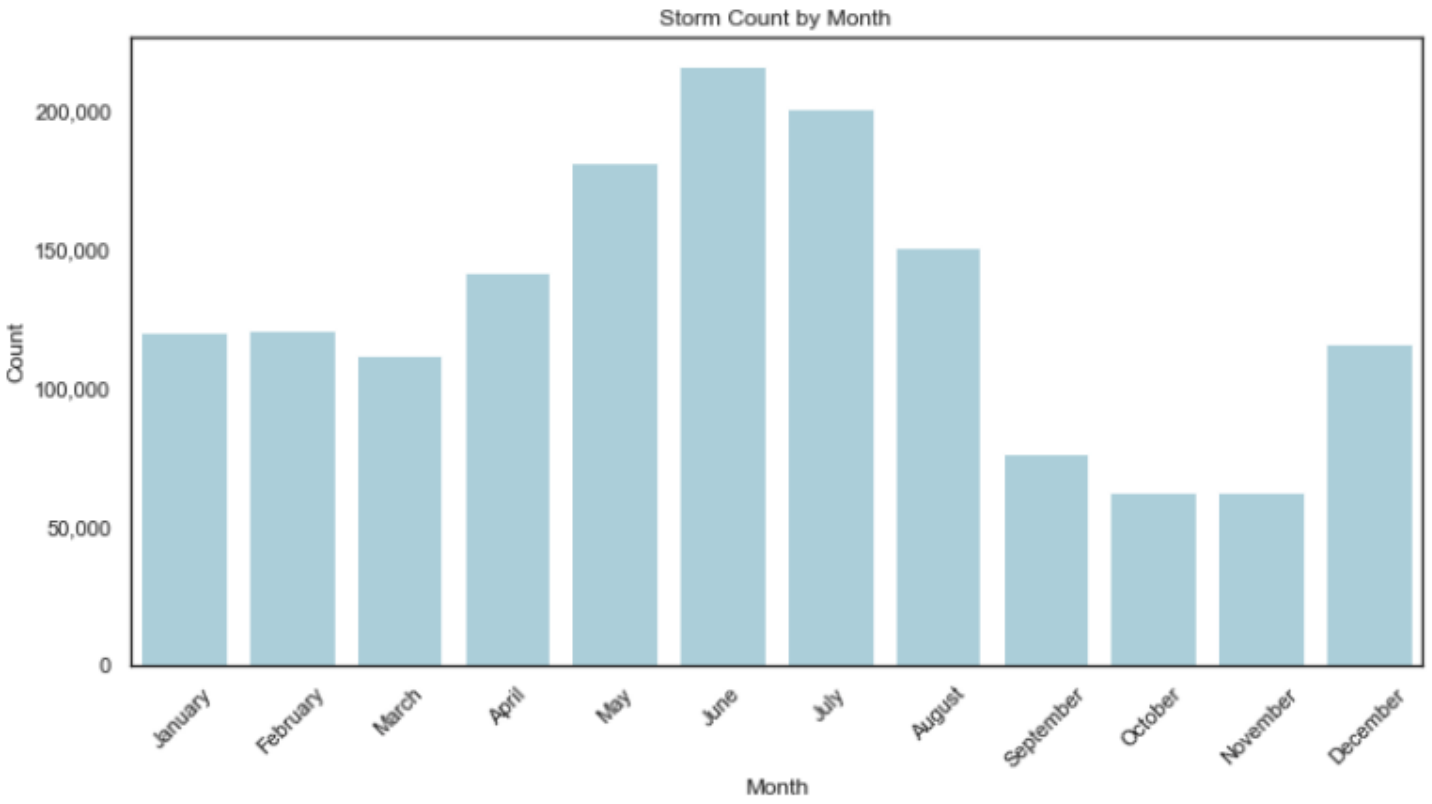
Geographic map of storms



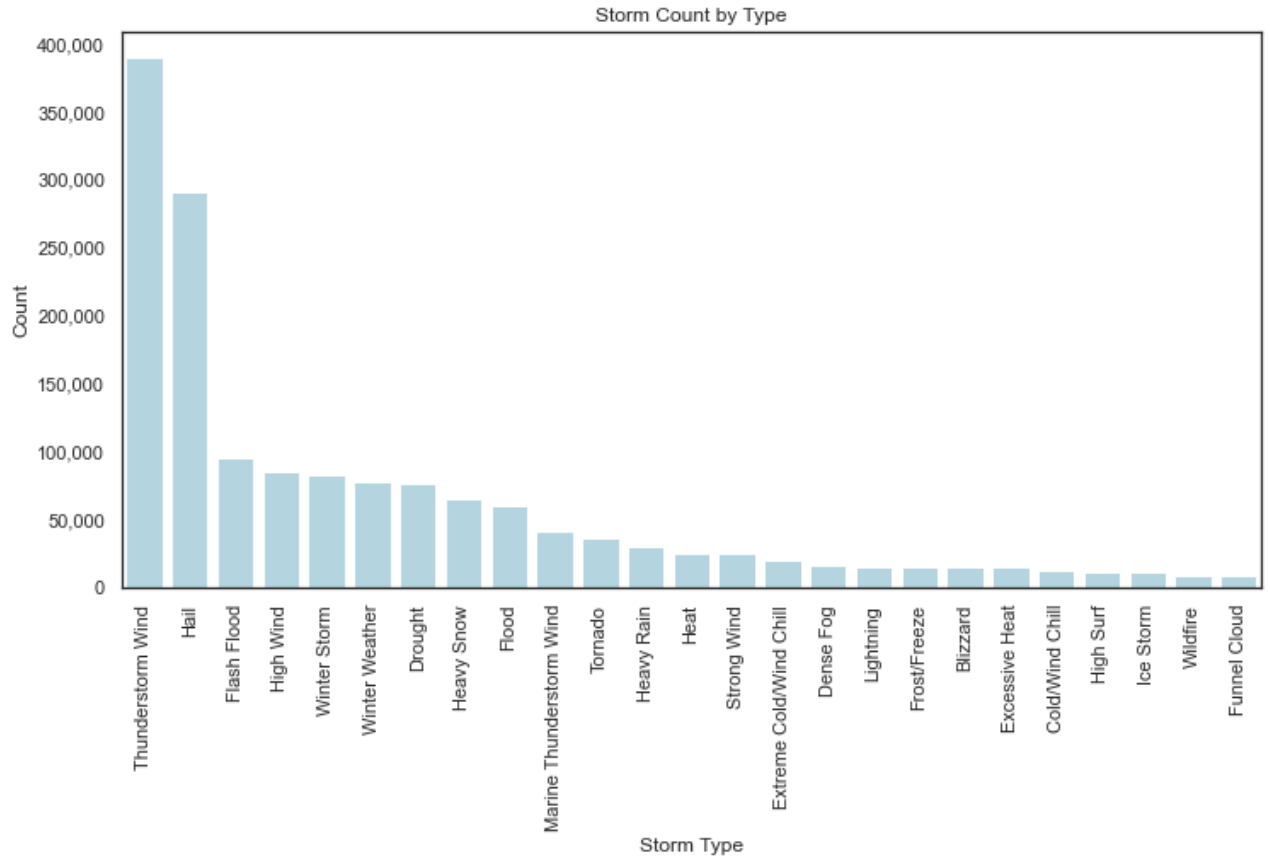
Storm damage by year



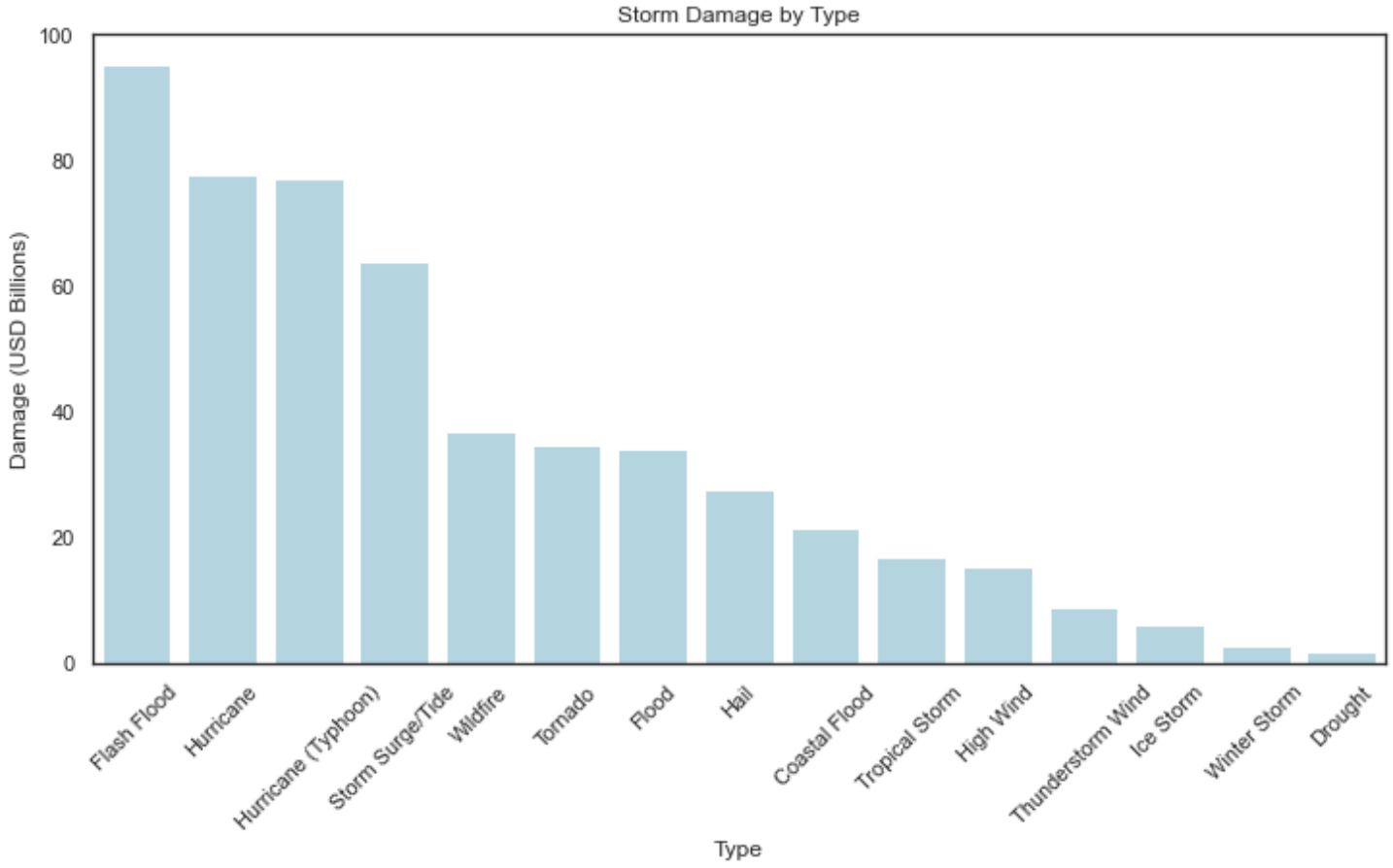
Storm count by month



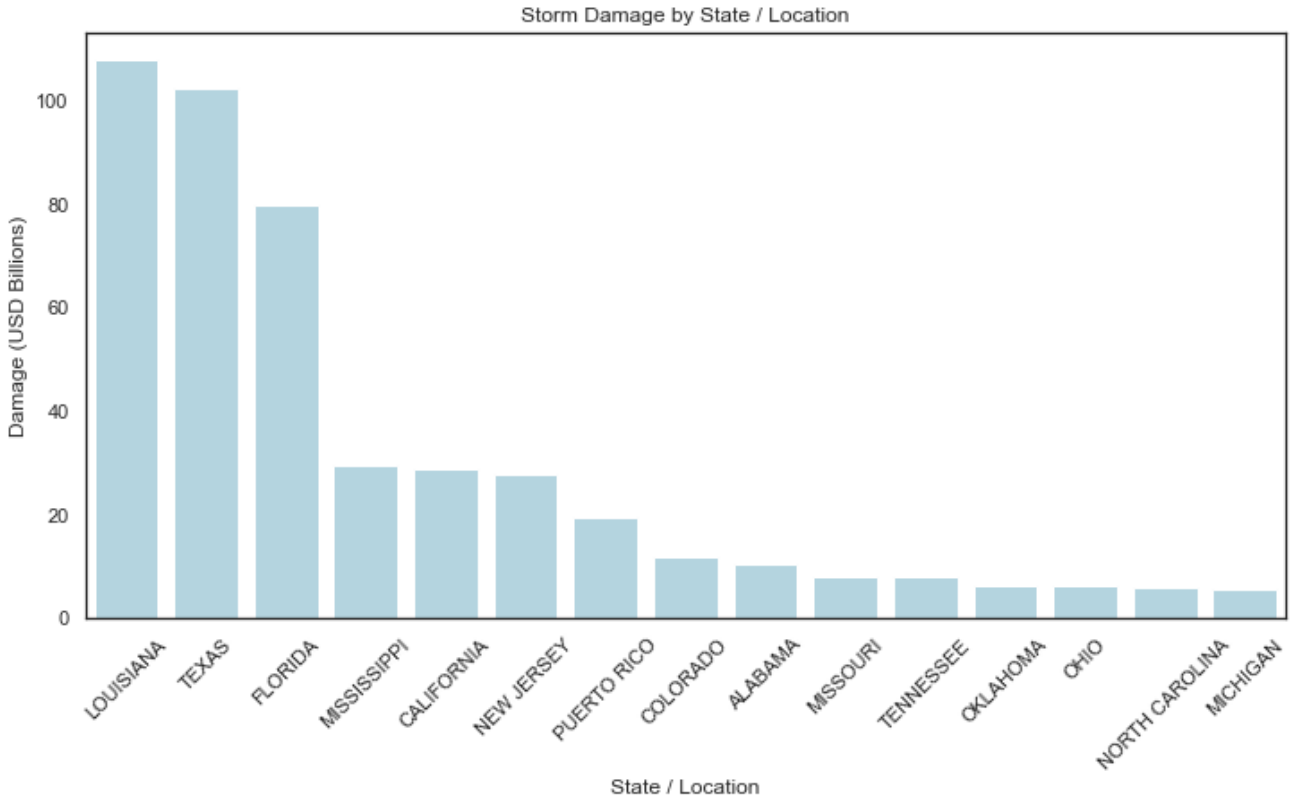
Storm count by type



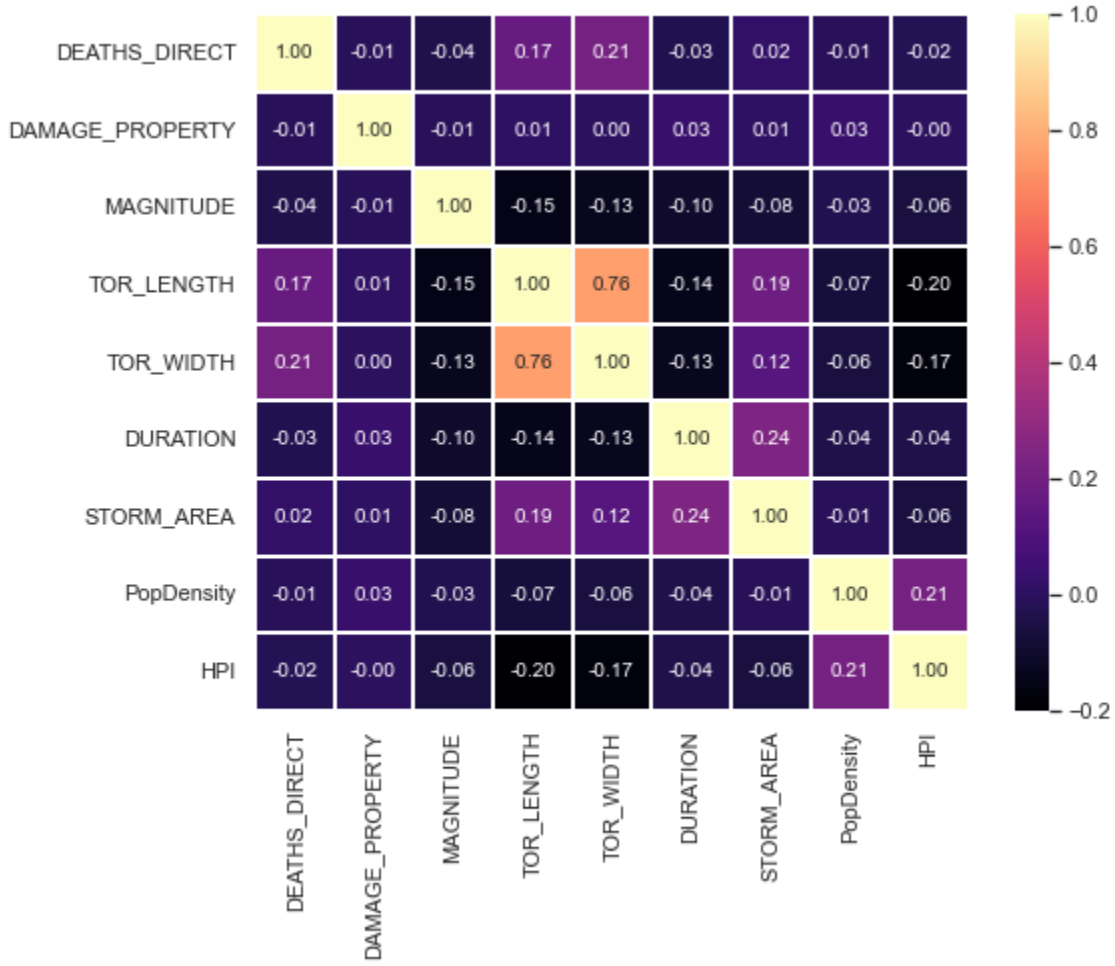
Storm damage by type



Storm damage by state / location



Correlation heatmap for numeric features (after removing 5,000 outliers)



Algorithms and Modeling

| Algorithms and Modeling Notebook |
| Pre-processing and Training Data Development |
| Metrics Table |

I initially focused on a Random Forest model, knowing that this model would serve as a good baseline to improve upon. Knowing the data had a high number of both zeroes and large outliers, I experimented with different permutations of these two factors. I found that the model generally performed better with the zeroes included in the dataset, while performance also improved as large outliers were removed.
I then performed the same initial check on LightGBM, XGBoost, and Tweedie algorithms, and these as well performed best with the zeroes included and outliers removed.

I was also able to improve model performance by removing features that had a large number of null values, and features that displayed little predictive value in early runs. See the Pre-processing and Modeling Notebooks for full details.

Model Selection
I chose to implement the following algorithms in my pursuit of the most effective model for this regression problem.

- Random Forest
- LightGBM
- XGBoost
- Tweedie Regression (useful for special cases where there are many zero and high outlier predictor values)

Metric Selection

The below table shows the performance of the models. Each of the algorithms were relatively close in performance. **LightGBM performed the best, followed by XGBoost, Random Forest, and Tweedie.**

In the below table, it is also evident that increasing the number of removed outliers from 513 to 5,000 improved model performance by over 70%.

Model Metrics - Initial Baseline With 513 Outliers Removed

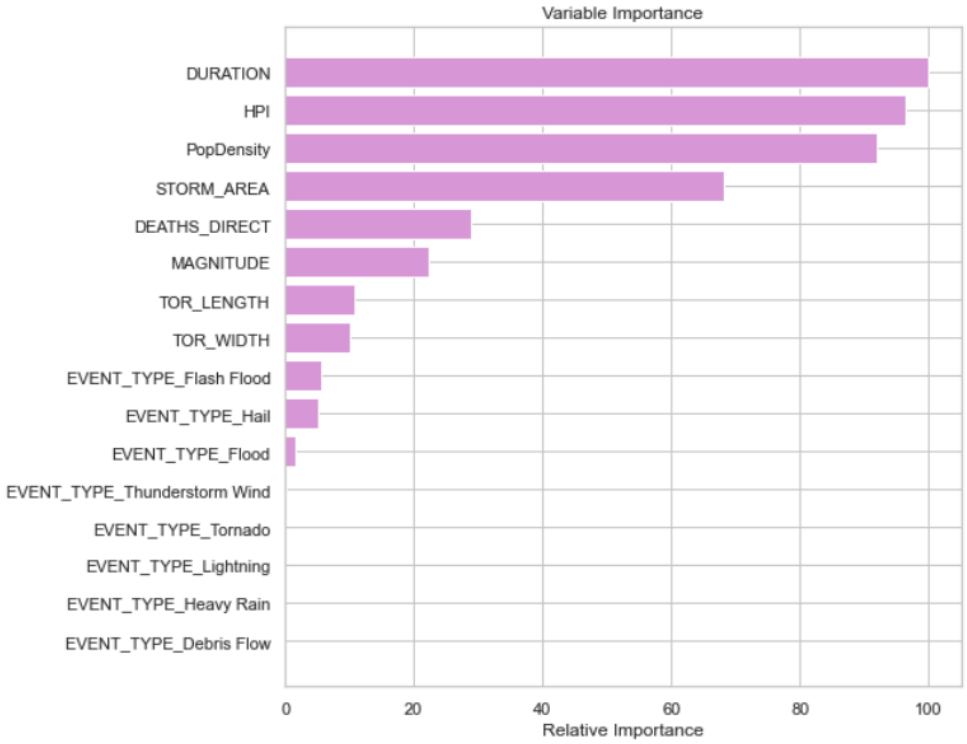
Algorithm	MAE		MSE		RMSE		MAE	
	Training	Test	Training	Test	Training	Test	DIFF	Comments
Random Forest								
Baseline	39,776	122,166	290,466,891,901	2,183,893,842,696	538,950	1,477,800	207 %	MAE difference of 207% between Train and Test implies there could be overfitting
Tuned	111,400	120,107	1,820,368,568,871	2,200,730,037,941	1,349,210	1,483,486	8%	This seems to be a solid model; only 8% difference between Train and Test scores.
Light GBM								
Baseline	100,749	112,986	1,523,342,294,233	2,050,794,462,465	1,234,238	1,432,060	12%	Suitable
Tuned	104,462	115,672	1,596,650,240,189	2,105,435,306,080	1,263,586	1,451,012	11%	Suitable
XGBoost								
Baseline	98,687	116,693	1,316,754,848,247	2,155,716,691,789	1,147,499	1,468,236	18%	Suitable
Tuned	96,141	115,470	1,274,231,448,321	2,130,958,777,782	1,128,819	1,459,780	20%	Suitable
Tweedie Regression								
Baseline	118,398	124,829	1,982,242,118,815	2,272,426,963,815	1,407,921	1,507,457	5%	Suitable
Tuned	120,540	125,133	2,370,604,253,978	2,301,905,785,478	1,539,677	1,517,203	4%	Suitable; used power = 3
MIN	39,776	112,986	290,466,891,901	2,050,794,462,465	538,950	1,432,060	0	

Model Metrics - With 5,000 Outliers Removed

Algorithm	MAE		MSE		RMSE		MAE	
	Training	Test	Training	Test	Training	Test	DIFF	Comments
Random Forest								
Baseline	10,192	27,985	2,908,505,925	19,156,222,167	53,931	138,406	175 %	Evidence of overfitting
Tuned	27,157	27,739	17,531,346,006	19,479,915,784	132,406	139,571	2%	Suitable
Light GBM								

Baseline	25,751	26,708	15,935,606,039	18,901,718,418	126,236	137,484	4%	Suitable
Tuned	25,862	26,861	16,013,959,656	18,950,807,358	126,546	137,662	4%	Suitable
XGBoost								
Baseline	25,001	26,917	14,684,962,256	19,102,140,608	121,182	138,211	8%	Suitable
Tuned	25,275	26,860	15,048,715,958	19,034,888,273	122,673	137,967	6%	Suitable
Tweedie Regression								
Baseline	29,731	30,585	18,761,667,052	20,554,164,254	136,973	143,367	3%	Suitable
Tuned	28,632	29,326	21,214,757,723	22,392,103,136	145,653	149,640	2%	Suitable; power = 1
MIN	10,192	26,708	2,908,505,925	18,901,718,418	53,931	137,484	0	
IMPROVEMENT FROM BASELINE								
Random Forest								
Baseline	-74%	-77%	-99%	-99%	-90%	-91%		
Tuned	-76%	-77%	-99%	-99%	-90%	-91%		
Light GBM								
Baseline	-74%	-76%	-99%	-99%	-90%	-90%		
Tuned	-75%	-77%	-99%	-99%	-90%	-91%		
XGBoost								
Baseline	-75%	-77%	-99%	-99%	-89%	-91%		
Tuned	-74%	-77%	-99%	-99%	-89%	-91%		
Tweedie Regression								
Baseline	-75%	-75%	-99%	-99%	-90%	-90%		
Tuned	-76%	-77%	-99%	-99%	-91%	-90%		

Feature Importance from LightGBM Model



Future Improvements, Opportunities for Pursuit

- It would be interesting to attempt to separate the impact of increasing property prices and storm severity on total damage. With property prices increasing substantially in the United States over the past few decades, this would have a meaningful influence.

- Explore additional datasets that could provide predictive value

Further Reading / Related Articles
Hurricane Costs
Here's why the US has more tornadoes than any other country
2022 U.S. billion-dollar weather and climate disasters in historical context
Where the Most Weather Warnings Are Issued in the U.S.
National Centers for Environmental Information
Climate Change Indicators: Weather and Climate