PREDICTING FLOOD DAMAGE TO BUILDINGS

Based on FEMA Flood Insurance Policy Claim Data

Springboard Capstone Two

Summer 2020

CONTEXT

Flooding poses a risk to human health and causes significant damage to buildings and assets across the world every year.

QUESTION

Can we use building and site characteristics to help homeowners and property developers predict and possibly prevent severe flood damage?

DATA

Any property located within a FEMA-designated flood zone is required to purchase flood insurance.

In 2019, FEMA released a redacted dataset describing flood insurance policy claims submitted between the 1970s and today.

This dataset includes over 2 million flood damage claims.

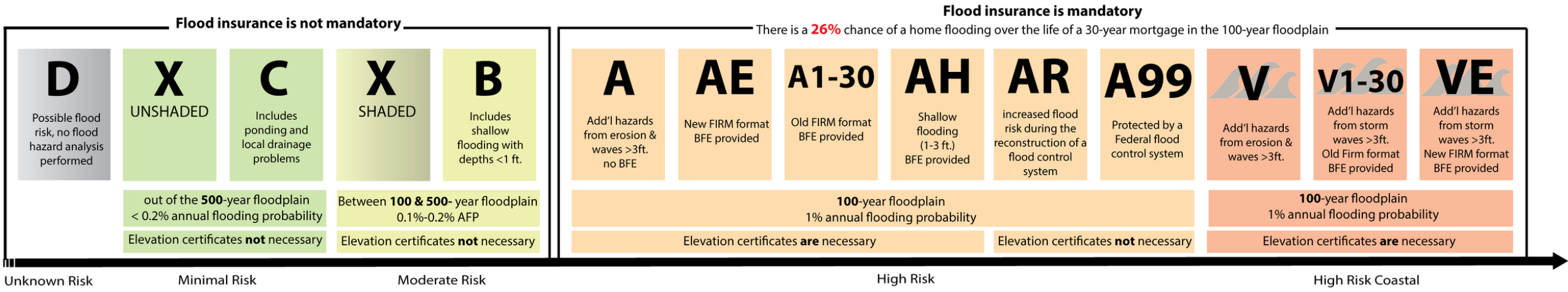
DATA CLEANING

Missing Values

No values were dropped. Indicator columns were created to capture missing values, then filled with either the median or mode of the column, based on the nature of the variable.

Flood Zone

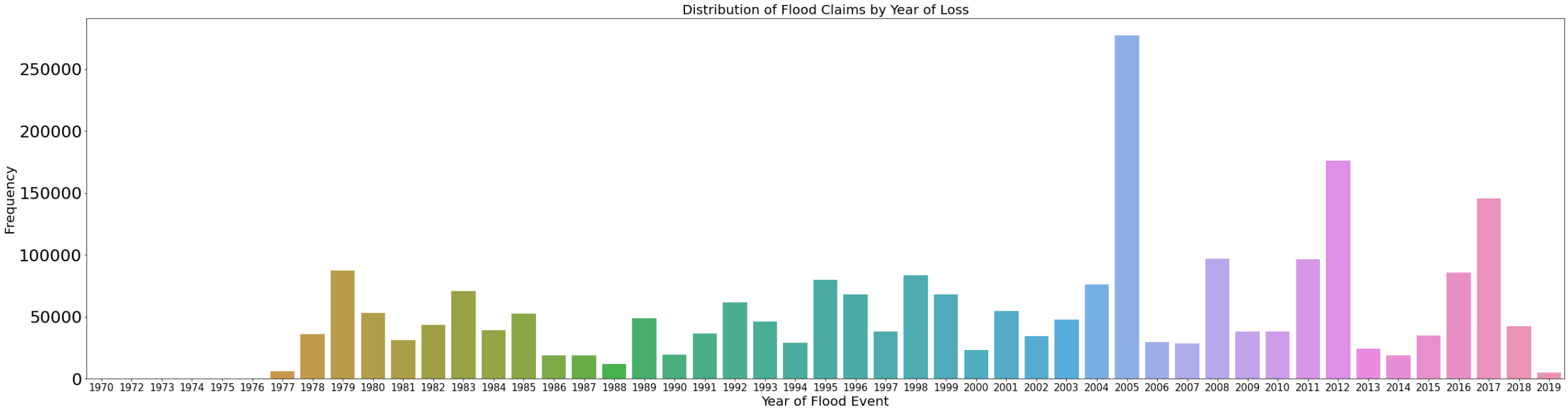
FEMA's flood zone designations have evolved over time. To simplify and make this variable relevant for today, a new flood zone variable was created based on the primary zone letter.



EXPLORATORY ANALYSIS

Year of Loss

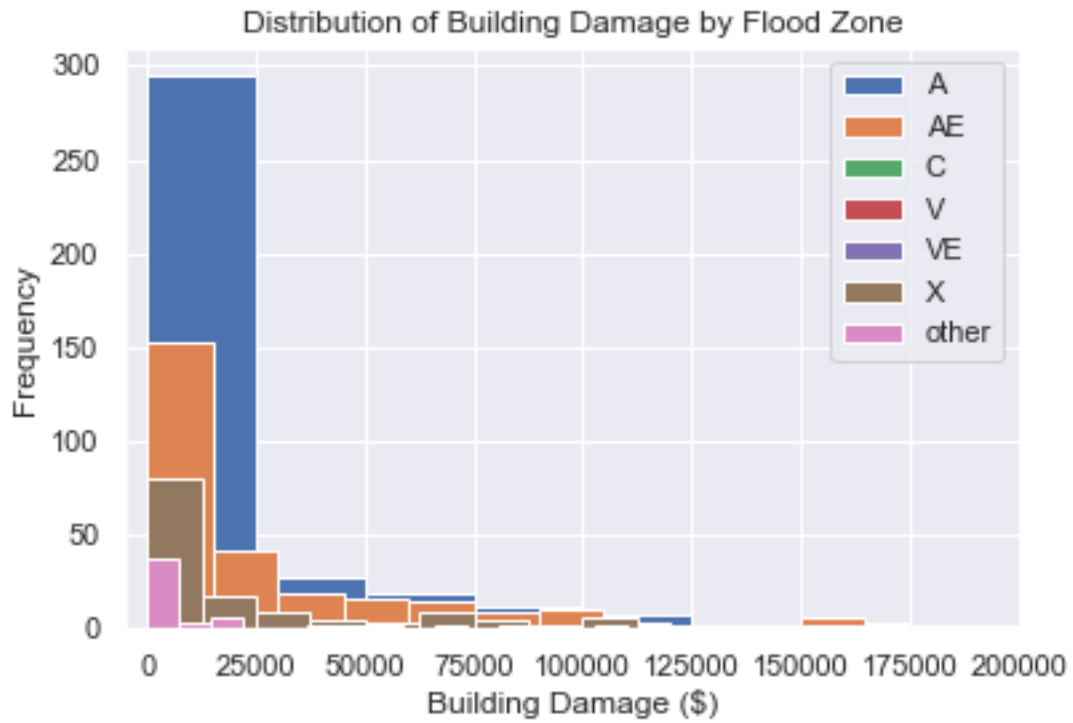
There is significant variation in the number of claims submitted each year. However, it is clear the overall trend is increasing with significant flooding years occurring in 2005 (Hurricane Katrina), 2012 (Super Storm Sandy), and 2017 (Hurricane Harvey).

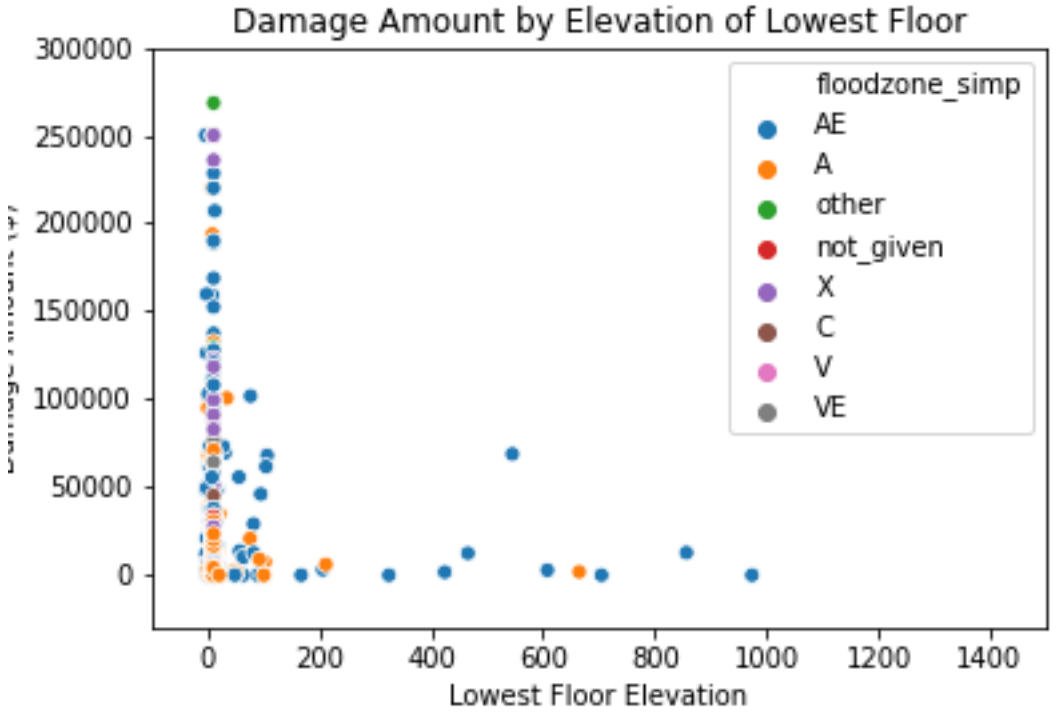


EXPLORATORY ANALYSIS

Floodzone

More flood insurance claims occur in the A and AE zones which represent the 100-year floodzone. However, the range of the amount paid on building damage is similar among the 100-year and 500-year zones.

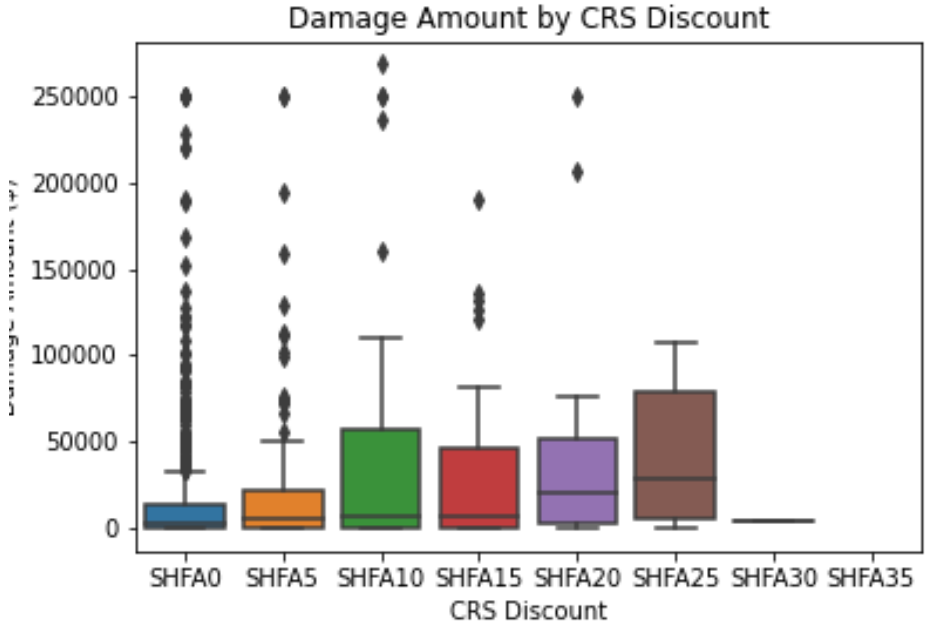




EXPLORATORY ANALYSIS

Community Rating System Discount

The Community Rating System (CRS) is a voluntary FEMA program that awards a discount on flood insurance premium to communities that employ flood hazard mitigation strategies on a community scale. The more floodzone management practices in place, the higher the discount on insurance premiums.



Unexpectedly, communities with a **higher CRS discount appear to have higher median flood damage costs**.

However, this characteristic may be driven by communities at more risk having more incentive to participate in the CRS program.

MODEL SELECTION

Best Model: Random Forest Regressor

Decision trees performed better than linear models. Ensemble method performed the best with the smallest RMSE of $38,946.

Primary Metrics: RMSE, R2

	Model	RMSE	MSE	MAE	R2
0	Initial Linear Regression	51637.851494	2.666468e+09	24587.314699	0.043978
1	Linear Regression_New Features	51509.955715	2.653276e+09	24575.968570	0.048708
2	Ridge Regression_alpha 0.1	51509.922542	2.653272e+09	24575.948201	0.048710
3	Ridge Regression_alpha 1.0	51509.674187	2.653247e+09	24575.801688	0.048719
4	Ridge Regression_alpha 10.0	51506.687010	2.652939e+09	24575.023541	0.048829
5	Decision Tree_no max depth	50940.682618	2.594953e+09	20628.905221	0.069619
6	Decision Tree_max depth 20	44352.663573	1.967159e+09	17738.136575	0.294705
7	Decision Tree_max depth 5	47845.999890	2.289240e+09	22923.252008	0.179228
8	Random Forest_estimators 10	40119.804842	1.609599e+09	17103.133113	0.422903
9	Random Forest_estimators 50	38946.508137	1.516830e+09	16968.584523	0.456163
10	Random Forest_estimators 100	39139.883893	1.531931e+09	16966.830714	0.450750

LIMITATIONS

Sample Bias

This dataset does not include properties that experience flooding but do not have flood insurance **OR** properties that have flood insurance but have never experienced a flood event.

Outliers/Invalid Data

This dataset includes many datapoints with questionable interpretation including very large and very small (negative) elevation values. This may indicate differences in units or database management standards that have evolved since 1970.

Size of Dataset

With over 2 million observations, this dataset was very large and therefore model fitting was very computationally intense. Further hyperparameter tuning may result in better model performance, if resources available.

KEY TAKEAWAYS

Feature Importances

The features with the most influence over building damage include the total amount of coverage of the insurance policy, the month of the flood event, and the year of construction of the building.

Feature	Importance
totalbuildinginsurancecoverage	0.1639693
monthofloss	0.0610176
constructionyear	0.0598852
policycount	0.0433211
basefloodelevation	0.0350354
lowestfloorelevation	0.034826
nbyear	0.0341901
state_MS	0.0215496

Base flood elevation and the **elevation of the lowest floor** are also within the top 10 important features suggesting some design considerations around building elevation may be a useful factor in reducing potential damage costs.