1 Introduction

Natural disasters have devastating effects on human lives, properties, and the environment. Since 1980, the United States has witnessed over 360 weather and climate-related disasters, with a total cost reaching a staggering amount of 2.6 Trillion dollars (NCEI, 2023). These costs highlight the urgent need for effectively assessing risks and implementing hazard mitigation planning to minimize losses. Conventional means of risk analysis fall short and are unable to accurately forecast and aid in coming up with preventive measures to address damages caused by natural disasters like floods, wildfires, earthquakes, tornadoes, etc. The limitations of traditional methods have led to a need in exploring more innovative approaches and techniques that are capable of providing more precise insights into natural disasters and their impacts.

With advancements in technologies, particularly in the field of machine learning techniques that are capable of analyzing complex datasets, identifying patterns, and providing valuable insights, we now have the potential to inform better predictions and aid in decision-making, potentially saving lives and drastically reducing economic losses. Machine learning techniques such as Regression, Classification, and ensemble methods have proven to be extremely useful tools for analyzing key information related to natural hazards.

This research aims to leverage these advanced machine learning approaches to develop a robust predictive model for natural disasters to effectively strategies hazard mitigation. The project aims to design a predictive models that can forecast potential disasters based on historical patterns and weather. A visual tool - an interactive map is provided that translates these predictions into geographical representation. The results of this capstone will provide to be a useful tool for hazard mitigation authorities which will allow them to make data-driven decisions regarding emergency response strategies. Advancing the ability to forecast and respond to natural disasters represents a significant step forward to a more resilient future.

1.1 Problem Statement

With the increasing frequency and intensity of natural hazards over the years, traditional methods of mitigation are not enough in dealing with types of disasters occurring throughout the country. Existing approaches lack the ability to integrate different data sources and are more focused on handling a specific type of catastrophe. The primary focus of this report is to overcome these limitations by developing a machine learning model to predict the nature of potential incident types within specific counties in the U.S. Given the ready access of data on disasters, hazard initiatives, weather patterns, and census, there is a unique opportunity to conduct in-depth analysis and leverage machine learning. Such a solution can significantly improve prediction accuracy and provide invaluable insights to assist authorities in their decision-making process.

1.2 Goals and Objectives

Goal 1: Data Analysis and Model Development

Objective 1 : Define project scope and gather relevant historical disaster and weather-related data.
Objective 2 : Collect, analyze and select relevant features required by performing exploratory data analysis.
Objective 3 : Clean and preprocess data to ensure data usability for model training.
Objective 4 : Implement and evaluate machine learning algorithms (Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting)

Goal 2: Model Selection, Predictions, and Interactive Map

Objective 5 : Select the best model (Random Forest) and make predictions on disaster types.
Objective 6 : Create an interactive map using folium to display predictions and population for each county.
Objective 7 : Document code, and methodology, and generate a final report.

2 Related Work

Recent research studies (Linardos et al., 2022) have explored the application of machine learning(ML) and artificial intelligence(AI) to develop innovative methods for hazard mitigation. The ML methods consist of Support Vector Machines(SVM), Naive Bayes, Decision Trees, Random Forest, Logistic Regression, and K-nearest neighbor clustering algorithms which are leveraged to enhance risk assessment and improve decision-making in disaster management. (Sun et al., 2020) AI models have employed multiple machine-learning techniques for disaster management in different phases; mitigation, preparedness, response, and recovery.

In the mitigation phase, ML models have been explored to identify areas that are prone to natural disasters and access their vulnerabilities. ML models have been used to analyze historical data on past disasters, geographical data, and climate patterns. By selecting specific features and predicting algorithms, these models can provide insights into which area and resource should be prioritized(Arinta & Andi W.R., 2019). In the preparedness phase, ML and AI have been used to support the planning and forecasting intensity of natural disasters such as storms and tornadoes, thus enabling authorities to make decisions regarding evacuation plans. As an add-on, ML models are used to optimize the availability of necessary supplies during emergencies. The majority of ML systems are focused on the Response phase which provides in-depth real-time analyses of data from multiple sources such as social media, satellite, and sensor networks.

Figure 1. Risk Index Formula

(Drakaki et al., 2018) talks about how sentiment analysis and NLP (Natural Language Processing) is used to get an idea of public sentiment and identify critical needs in a given emergency.

Image recognition and other computer vision deep learning techniques are used to identify damaged buildings by developing a (Liu et al., 2020)state-of-the-art CNN(convolutional neural networks). Another Chaurasia et al., 2019 study shows a fast-paced approach for determining urban structure damage caused by Earthquakes using Elastic net regression. While (Huang et al., 2018) proposed an ensemble-based detection for structural damages and presented experimental results contributing to the advancement of structural health monitoring. (Khan et al., 2022)There is a significant amount of ongoing research focused on rapid assessment and remote sensing using UAVs to help in detecting, mitigating, responding, and preparing for disasters as well as rescue missions at sea. These systems are developed specifically to tackle the Recovery phase and aid in post-disaster analysis.

(Rahman et al., 2023) study focuses on the lack of comprehensive automated visualizations specifically for cyclone prediction and has designed a dashboard that serves to track cyclones and understand the formation, intensity, and direction. The study provides some valuable insights into an improved prediction model for cyclones. (Tian et al., 2020)Another study combines CNN to accurately estimate the severity of tropical cyclones using a hybrid model. (Bochenek & Ustrnul, 2022) explores the various applications of machine learning algorithms to enhance forecasting accuracy, and drastically improve climate analysis at the same time talks about the challenges faced in this field.

National Risk Index (Zuzak et al., 2022) offers a relative measurement of community level for natural hazards risk across 50 states integrating various factors. An expected annual loss estimate and risk index are calculated based on the following equation in Figure 1 expanding on the conventional methods.

Even though there have been significant advancements in disaster management, there are several challenges that still require attention to make these systems more robust and accurate.

Table 1. Description of Disaster Summaries Column

Column Name	Description
disasterNumber	Range from 1 to 5465
fyDeclared	Yearly data from 1953 to 2023
ihProgramDeclared	Binary, (0 or 1)
iaProgramDeclared	Binary, (0 or 1)
paProgramDeclared	Binary, (0 or 1)
pmProgramDeclared	Binary, (0 or 1)
fipsStateCode	FIPS identifier for state
fipsCountyCode	FIPS identifier for county

Table 2. Missing Values for Disaster Summaries

Column Name	Missing Values
incidentEndDate	605
disasterCloseoutDate	15114
lastIAFilingDate	46561

3 Methodology

The methodology consists of four main sections where the four datasets are explored; FEMA Natural Disaster Summaries, FEMA Hazard Mitigation Projects, NOAA Climate and US Census (Population). In-depth Data Analysis is performed to get an idea about the datasets. For the model, mainly the disaster and climate datasets are utilized. The two datasets are merged into a single file which is used to train and implement various machine learning algorithms. Each model is evaluated and compared to select the most optimal algorithm.

3.1 About the Data

The FEMA Disaster Summaries dataset contains records of federally declared disasters that have occurred in the United States. It contains information related to various disasters including their nature, location, declaration dates, and federal assistance programs((FEMA), 2023a). Key columns in the dataset such as incident type (e.g. Hurricane, Wildfire), fiscal year, incident begin and end date, geographical information like states and county codes, and specific programs invoked provide us with insights

Table 3. Description of NOAA Columns

Column Name	Description
year	yearly data from 1850 to 2023
month	January to December
precipitation	Amount of Rainfall (inches)
AverageTemprature	Mean Temperature (F)
MaxTemp	Highest temperature (F)
minTemp	Lowest temperature (F)
HeatingDays	Number of days heating required
CoolingDays	Number of days cooling required
StateCode	FIPS identifier for state
CountyCode	FIPS identifier for county

into understanding disaster trends and emergency responses. Multiple datasets from the National Oceanic and Atmospheric Administration (NOAA) were used to capture climate-related data such as precipitation, average temperature, maximum and minimum temperature, number of cooling days and number of heating days. Each of these metrics together formed a weather profile for different counties and states across the U.S. By integrating them, the influence of weather patterns on the nature and frequency of natural disasters. (Oceanic & (NOAA), 2023) The Hazard Mitigation Assistance Projects dataset focuses on mitigation projects that were initiated to minimize the risks associated with natural disasters. The dataset provides a detailed view of the program, project, and financial details of the assistance projects along with geographical information such as state, county, zip code, region, latitude, and longitude. ((FEMA), 2023b) The U.S. Census population dataset provides an overview of the nation's population. It offers essential insights into the demographic landscape across various counties in the U.S. This dataset plays a vital role in understanding the regional variations in population and can be leveraged for urban planning, resource management, and emergency strategies. (Bureau, 2023)

3.2 Exploratory Data Analysis

To develop a versatile machine learning model for forecasting the most probable disasters in various counties, Exploratory Data Analysis is the first step. EDA enables identifying underlying factors, patterns, and anomalies in the datasets. Visualization and statistical analysis of geographical information, population density, and historical data will help in formulating appropriate responses for a given disaster. Four datasets were analysed: FEMA Disaster Summaries, FEMA Hazard Mitigation Projects, NOAA

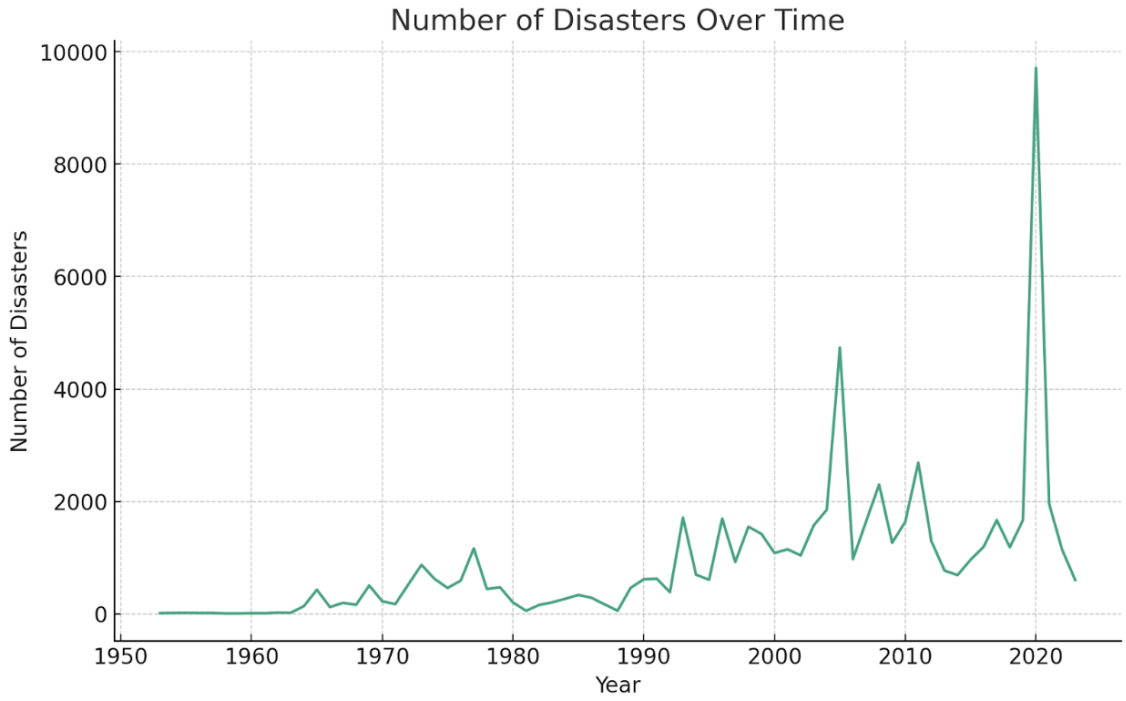

Figure 2. Frequency of disasters over the years

Climate data, and U.S. Census Population. From these, FEMA Disaster Summaries and NOAA Climate data were especially important for building the predictive machine learning model. By examining these datasets, regions that are more susceptible to specific disasters, owing to geographical or climate factors can be identified. By filtering and selecting most relevant features for the model, optimal accuracy in prediction is ensured. Rather than focusing on conventional ways of resource allocation, it enables the thoughtful orchestration of assistance and aiding the most vulnerable regions. This approach fosters more proactive, informed and responsive strategies for hazard mitigation.

3.2.1 FEMA Disaster Summaries

The line chart as seen in Figure 2 depicts the number of disaster declarations over time. Since the early 1950s, there has been a consistent and notable increase in the occurrence of disasters. The upward trend can be attributed to factors such as climate change, population growth, and urbanization. There is a large spike during 2020 which is because of COVID-19. The continuous rise in the graph emphasizes the growing requirement for disaster management and preparedness.

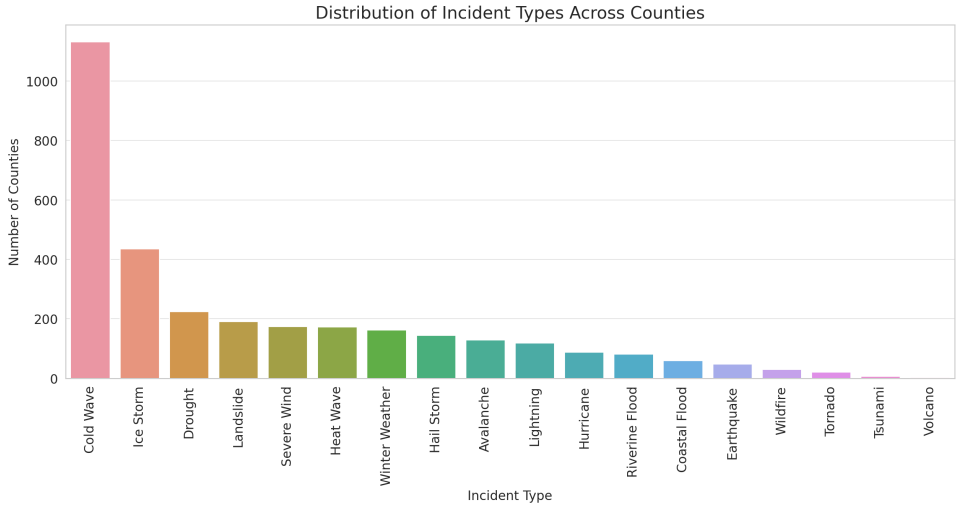

Figure 3. Most common type of incidents across the U.S.

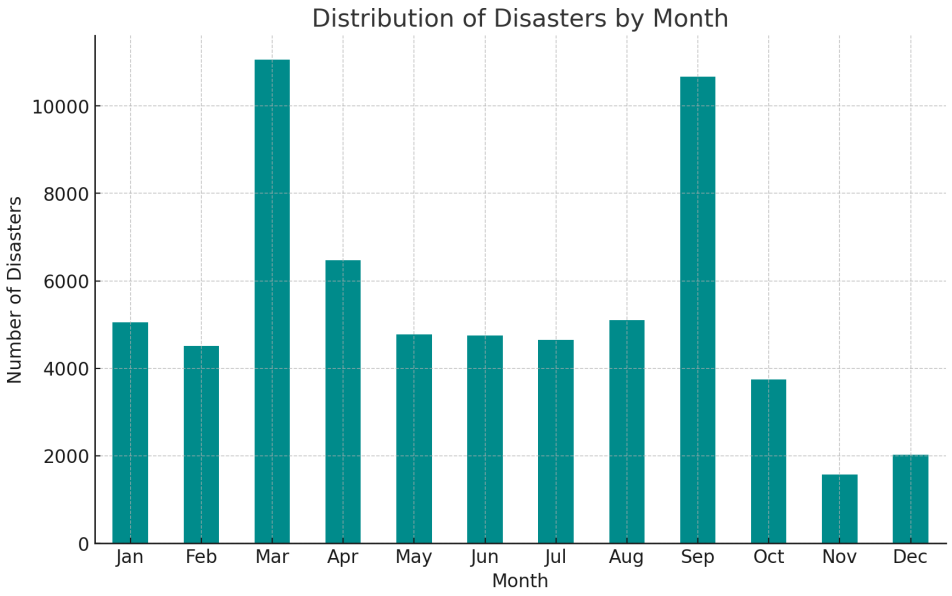

Figure 4. Frequency of Disasters per month in the U.S.

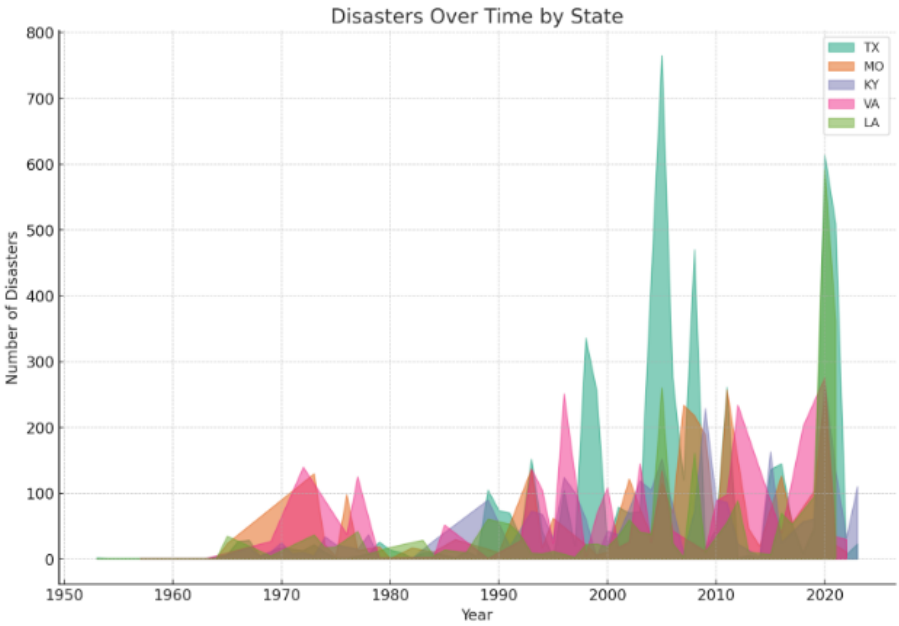

Figure 5. Number of disasters over time (Top 5 States)

In Figure 3, the top most occurring incidents are plotted. Cold Waves have the highest frequency of occurrence across all counties in the U.S. compared to other incidents by a large margin. Coastal flood frequency is lower as they predominantly affect regions that are along the ocean. While Volcanoes only affected certain counties that are present in Hawaii and are extremely rare.

Given below is a bar plot (Figure 4) of the number of disasters per month. It can be observed that March and September have the highest number of events declared compared to the other months. The spikes might correspond to the transition from winter to spring. This shift can lead to a mix of winter storms, early spring winds, tornadoes, and increased rainfall which can cause flooding. September is a prime month for hurricanes, storms, and cold waves due to the Atlantic Ocean. October, November, and December have the least number of disasters since they fall within the autumn and early winter months. Given this data, it would be beneficial for agencies to focus more resources during these months which could involve ensuring stockpiles of supplies and conducting general public awareness campaigns. With the changing climate patterns and increase in global temperatures, there may be an increase in extreme weather events and there is a need to continuously monitor and update the data.

In the following Figure 5, Texas stands out with the highest among the top five states with the most disasters. Louisiana and Virginia have frequent but less intense peaks compared to Texas. The pattern suggests that it might be a better idea to allocate more resources to states with more frequent and intense disasters. It is important to note that the visualizations are based on the number of disasters, not the severity or intensity

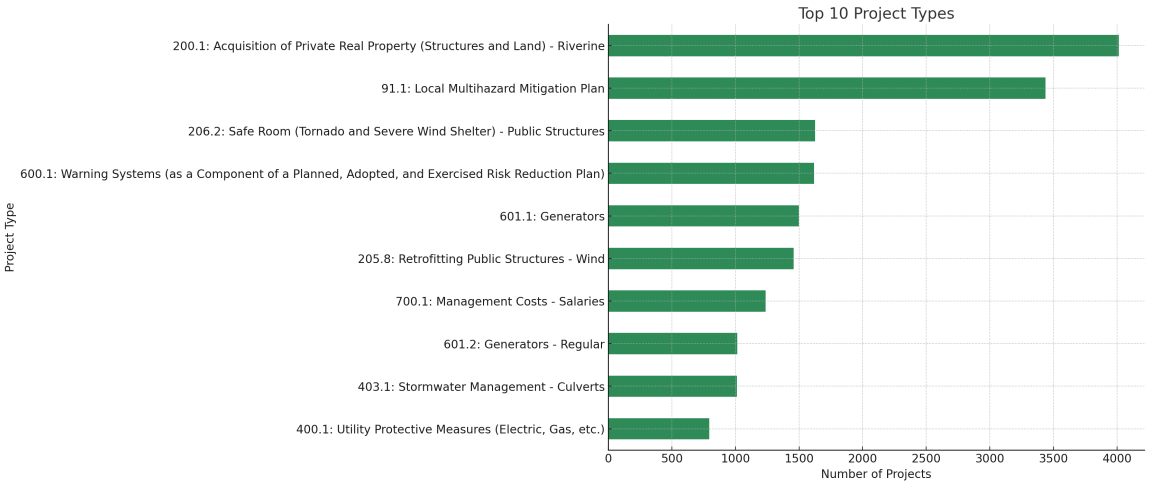

Figure 6. Average amount of funding over the years

of them. Some disasters like tsunamis or earthquakes have a significant impact but happen less frequently.

3.2.2 FEMA Hazard Mitigation Assistance Projects

Different types of projects have been set up throughout the U.S. In Figure 6 it can be observed that the largest number of projects are for the Acquisition of Private real estate related to Riverine - Structures and land, followed by Local multi-hazard mitigation plans.

Florida has the highest number of Hazard Mitigation Projects, followed by California, Alabama, and Texas as in Figure 7. Texas, California, and Oklahoma lead in the number of disaster declarations, yet the distribution of mitigation projects reveals a different pattern. The number of projects exhibits a drop from Florida to Texas reflecting a significant contrast between states. This could be due to the way the disasters are declared.

Looking at the amount of funding for the top five states in Figure 8, Texas has the most amount of funding. Florida on the other hand has lesser funding than Texas. In the previous graph, it can be seen that even though Florida has the highest number of hazard projects the amount of funding is not as high as that of Texas. The higher frequency of disasters in Texas and Florida is a prominent factor. Historically both these states have been more susceptible to a broader range of calamities when compared to New York, Iowa, and Mississippi. It seems there may be a higher intensity of disaster-related damage in Texas due to the size of the state, population, and number of structures. Another inference that emerges from the funding distribution is the coastal location of these states except Iowa. Specialized mitigation strategies are required to handle coastal areas due to the nature of the risks faced by them.

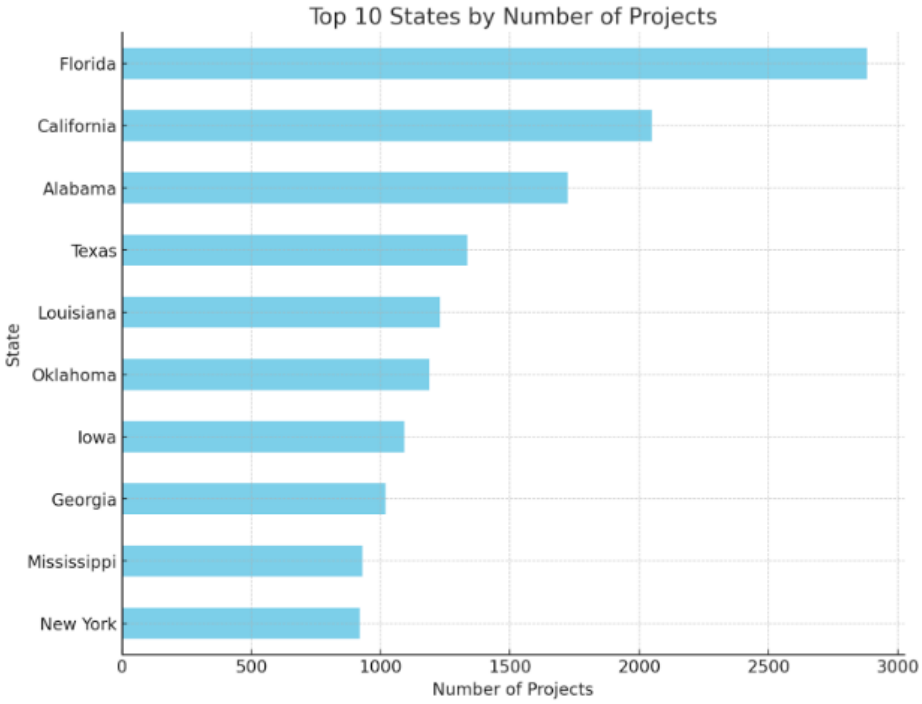

Figure 7. State by Number of Projects (Top 10 States)

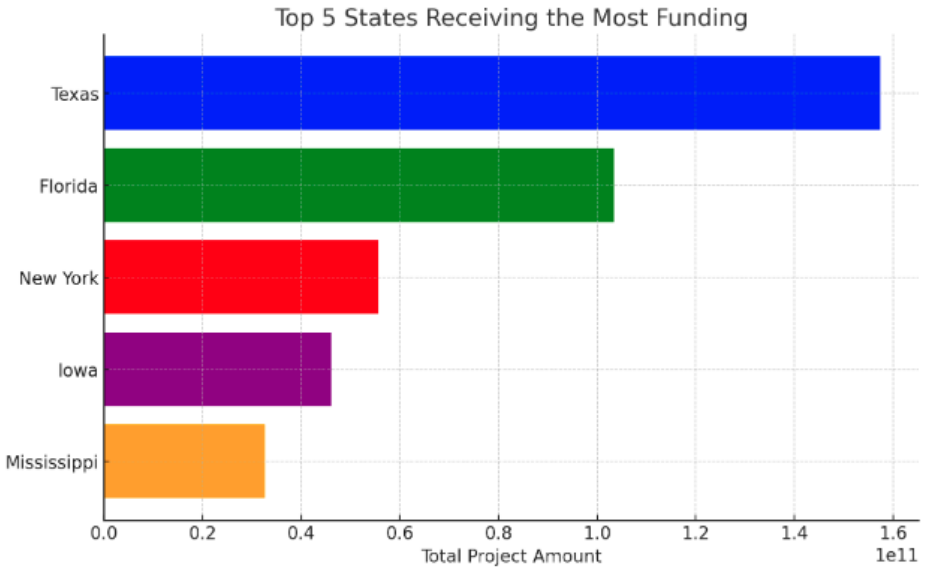

Figure 8. Highest funded States in the U.S.

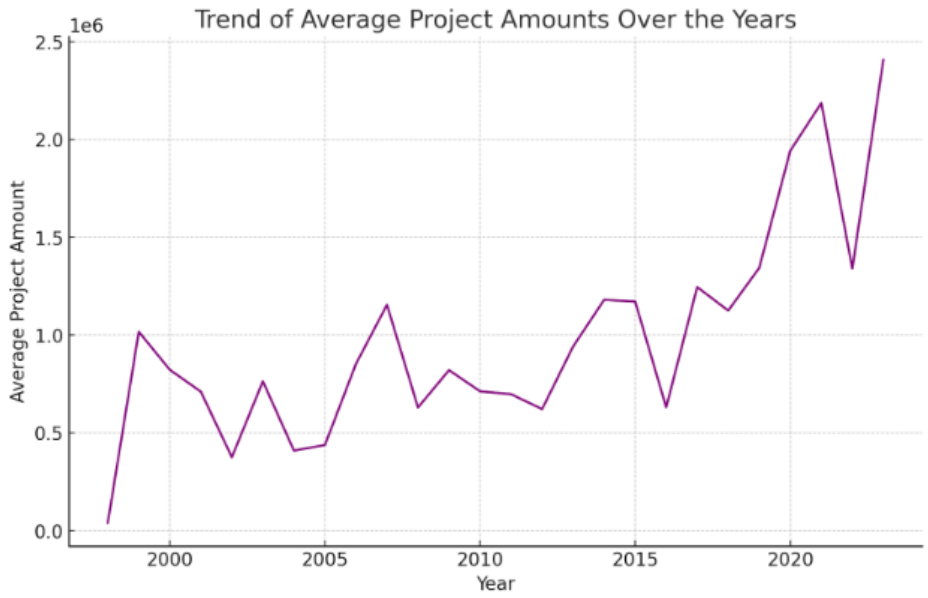

Figure 9. Average amount of funding over the years

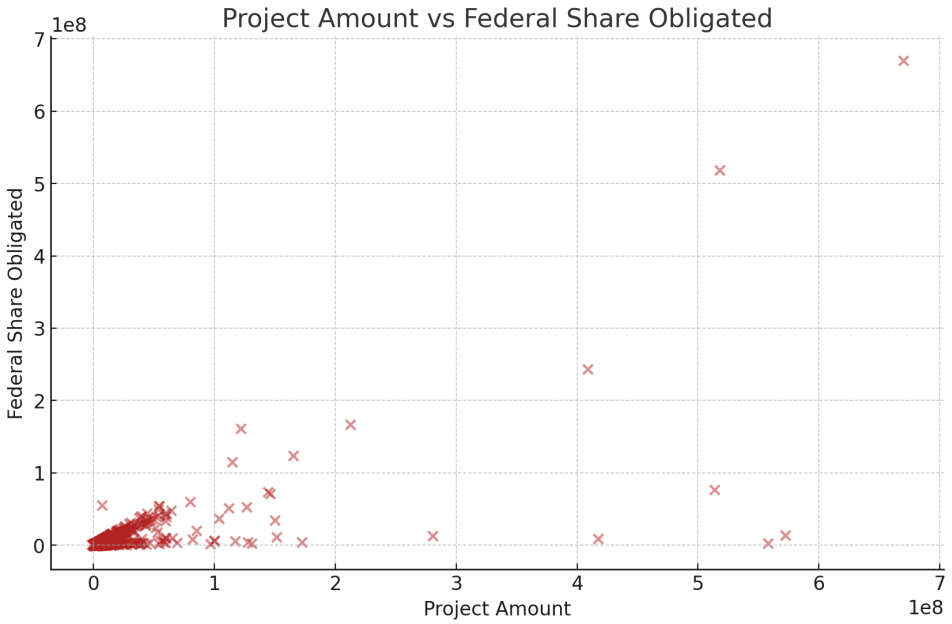

Figure 10. Federal shares and project amount scatterplot

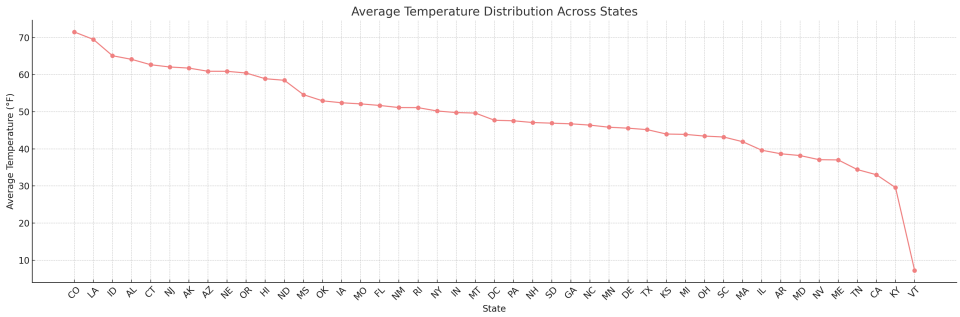

Figure 11. Average Temperature Distribution across U.S.

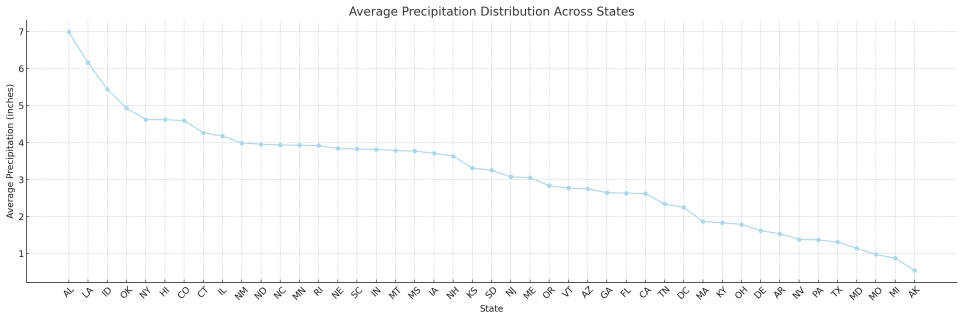

Figure 12. Average Precipitation Distribution across U.S.

The analysis of average project amounts for hazard mitigation over the years in Figure 9 reveals an upward trend. Several factors contribute to the increase in cost like inflation, increased complexity of projects, and the scale of natural disasters. The magnitude of resources required has drastically increased over time. A closer look at the trend depicts the number of disasters that have also increased over time in turn driving up the funding as well. The severity and complexity of disasters are escalating and require more compensation and effective resource management. In the scatter plot (Figure 10), the amount of funding each project receives along with the federal shares that are obligated for it. With the increase in funding more federal shares are provided and they seem to be related exponentially. The government might be aiding in providing additional incentives for larger projects to ensure that the benefits are received by a larger subsection of the population. The bigger the project the greater the risks which signal the federal government's responsibility to share the larger burden in case the project fails. Understanding this relationship is essential for project planners and authorities that are responsible for setting budgets. It is crucial to ensure that smaller projects are not overlooked and require a balance for equal distribution of resources.

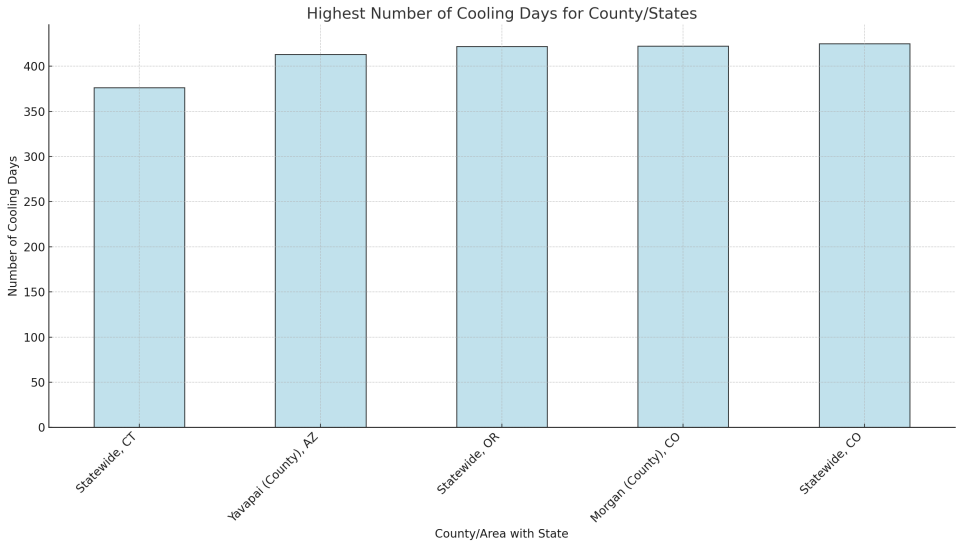

Figure 13. Highest Number of Cooling Days across U.S.

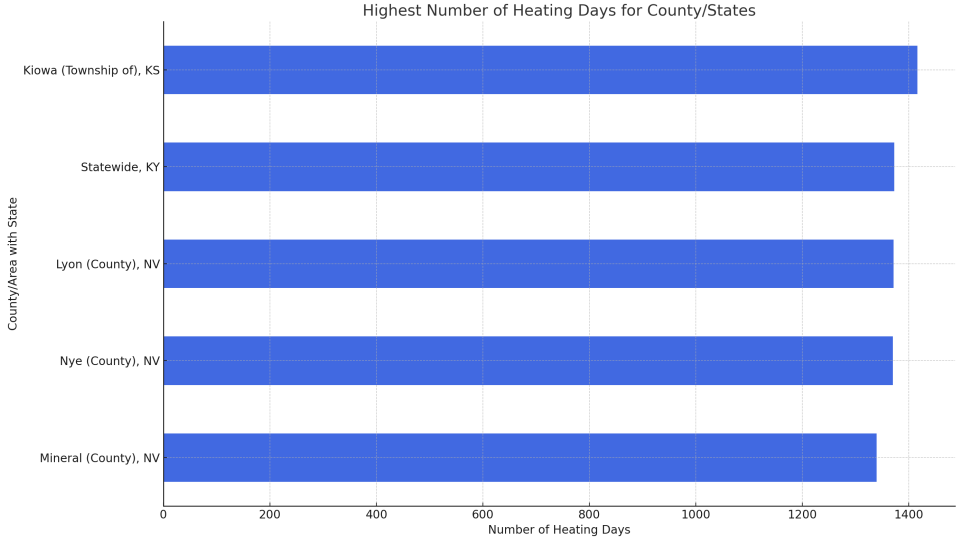

Figure 14. Highest Number of Heating Days across U.S.

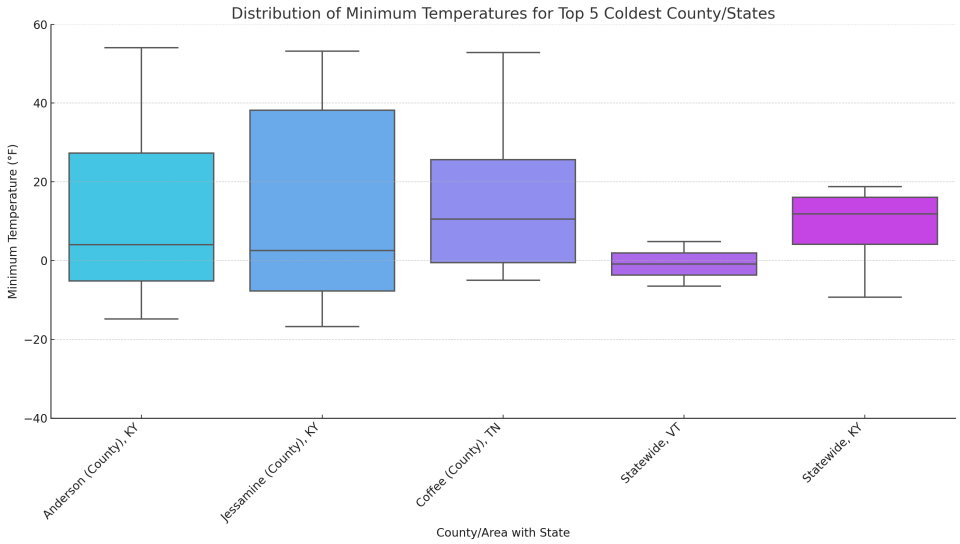

Figure 15. Lowest Minimum Temperature across U.S.

3.2.3 NOAA Climate

12 shows a detailed perspective of average precipitation level across states in the U.S. providing insights into the rainfall trends across the country. Rainfall can be associated with disasters such as Severe storms, flooding and other events. Alabama and Louisiana and Idaho have the highest number of rainfall while Hawaii and Alaska have the lowest. Louisiana is more prone to hurricane and tropical storm due to the coastal location while Idaho is more prone to flash floods and landslides due to the soil and hilly regions in the state. The amount of precipitation is dependent on a lot of factors such as geographical location, wind patters, and proximity to water bodies.

In Figure 11, average temperature distribution of states are given which is essential climate parameter that influences a variety of factors as it effects the length and timing of seasons along with other environmental implications. Colorado, Louisiana and Idaho have been identified to have the highest average temperature among the rest of the states. From a disaster management point of view, these insights are extremely valuable since a distinct link between elevated temperature and potential natural hazards can be found.

The bar chart in Figure 13, illustrates the areas that experience the highest number of cooling days annually. Cooling days refers to the days when the temperature is high and the use of air conditioning or cooling is required. Counties in states such as Colorado and Orlando experience more intense and prolonged warmer climate. This could help in identifying incidents such as heatwaves and the potential challenges that need to be addressed.

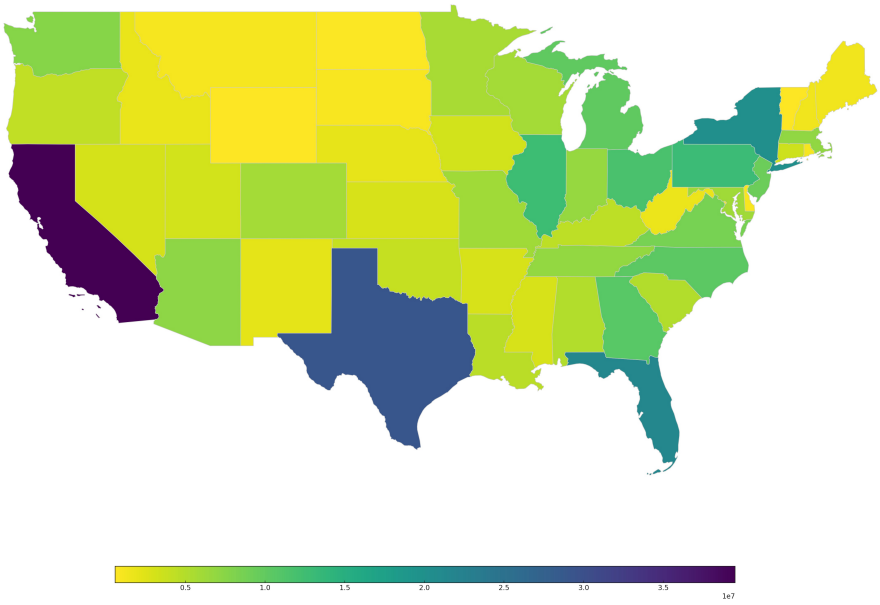

Figure 16. Highest Populated States in the U.S.

Figure 14, showcases the top five counties that experience the highest average number of heating days. Heating days refers to the number of days in a year where the temperature is low enough to necessitate heating in building to maintain the temperature. From the chart, it's evident that the counties tend to have cooler or extended winter seasons which require more indoor heating. This provides valuable insights about regional climate variation and insights for planning energy and heating demands.

The boxplot in Figure 15, provides a detailed view on the distribution of minimum temperature across five of the coldest places. There's noticeable variability in the temperature across these regions throughout the year. The median temperature is represented by the center line of the boxplot. It can observed that Coffee, TN and the entire state of Kentucky have a higher median range than the rest of the group. The coldest places seem to be in Kentucky State which emphasizes its characteristic cold climate indicating that the state often experiences natural disasters related to the cold weather.

3.2.4 US Census

The map in Figure 16 visually emphasizes areas with higher population density. Coastal states tend to have higher populations including states like California, Texas, Florida, and New York. California is the most populous state as seen on the map. From Figure 17, Los Angeles, CA has a population of over 10 million followed by Maricopa, AZ with 4.4 million. Several counties in CA such as San Diego, Orange, and Riverside are in the top ten most populated counties. Most of these counties are coastal and prone to

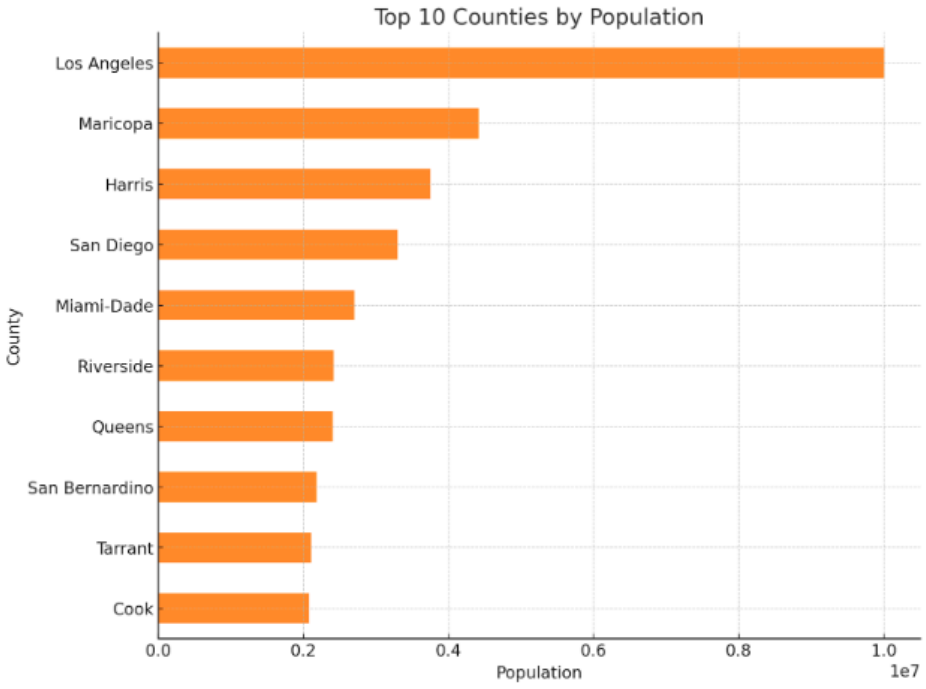

Figure 17. Highest Populated Counties in the U.S.

Table 4. Combined FIPS Code

StateCode	CountyCode	CombinedFIPS	Name
01	001	1001	Autauga, AL
06	000	6000	Statewide, CA
48	113	48113	Dallas, TX
11	001	11001	District of Columbia

hurricanes and coastal storms, while Los Angeles is prone to earthquakes and extreme heat along with landslides. The high populations underline the importance of mitigation planning to reduce the risks and safeguard human lives. Ensuring that buildings, roads, and utilities are a necessity to ensure that these regions are able to sustain natural disasters. There is also a need to position emergency services in disaster-prone areas to ensure rapid response when needed. Protecting densely populated areas goes beyond immediate responses and includes long-term planning.

3.3 Data Cleaning and Preprocessing

Data Preprocessing and cleaning is a vital step while implementing machine learning algorithms. During the data cleaning phase, columns were identified and dropped to focus on more relevant features for predicting incident types for each county. In the first dataset, several columns were removed; 'femaDeclarationString', 'state', 'declarationTitle', 'incidentEndDate', 'tribalRequest', 'placeCode', 'designatedArea', 'lastIAfilingDate', 'lastRefresh', 'hash' and 'id'. These columns were redundant information or non-essential for the analysis and model design. From the NOAA dataset, columns such as 'Precipitation', 'AverageTemperature', 'HeatingDays','CoolingDays', 'MaxTemp', 'minTemp' were selected. The state and county information in both datasets is represented using FIPS code which is a numerical representation of the respective state and county names. Year, and month are extracted from the incident begin date. After which the resultant columns will be converted into numerical features if required.
Missing values and inconsistent data types were addressed :

1. Rows with empty disaster numbers were removed.
2. County and State FIPS code were merged into a single column called 'combinedFIPS'.
3. Dates were converted to appropriate date time formats.
4. Duplicate rows were checked and removed.
5. Month were consolidated into one column and converted into numerical data.

3.3.1 Merging Datasets

The integration of FEMA Disaster Summaries and NOAA Climate data proved to be efficient for developing the machine learning model. The FEMA disaster data consists of information about various incidents occurred across the U.S. and was combined with NOAA's climate data which encompasses metrics such as precipitation, temperature and other vital weather-related elements. The merging process was based on three features: 'combinedFIPS', 'year', and 'month'. The combined FIPS code represents the geographical information about the state and county in a single feature. By aligning the data based on 'year' and 'month' the merged dataset shows the exact temporal and weather related information for the exact time frame of a disaster event with the concurrent climate conditions. The final dataset after being combined, offers deeper insights with prevailing weather conditions into the potential cause of natural disasters. This enhances the predictive capabilities of the machine learning model ensuring they have information about both historical incidents and relevant environmental features.

3.3.2 GeoJson Data for State and County

Since the aim of the project is to generate an interactive map for hazard mitigation authorities to view and get insightful information regarding the predictions made by the machine learning model about the high-risk counties and the type of disaster that are most prone to. There is a necessity to use a Geo Data Frame that consists of geographical information including names, FIPS code, land and water areas, and geometric shapes about the States and Counties in the U.S. The dataset will be combined with the predictions made by the model to create a visual representation of the incidents in each county. The file contains 3220 entries representing different counties in the U.S. along with their latitude and longitude, and geographical boundaries.

3.3.3 Encoding Incident Type

Before developing the model, 'incidentType' column needs to be converted from a categorical feature. Since machine learning models require a numerical input features are required to be transformed into a numerical format. There are various encoding techniques like one-hot encoding, label encoding, ordinal encoding, etc. that can be utilized. In this project, label encoding was implemented to transform 'incidentType'. Label encoding involves assigning a unique integer to each category, for example, hurricane is associated '1' and flood is associated with '7' as shown in Table 2. This retains the categorical information of the column which will be utilized as our target variable.