Analyzing and Predicting Environmental Disasters

Manovay Sharma

Project Goals:

The goal of this project was to utilize NASA's EONET API in tandem with Plotly, Pandas, Seaborn, and Sklearn to analyze environmental disaster trends in North America over the last decade and predict further occurences.

The 3 main questions this project centered around were:

1. How does the frequency of certain natural disasters change over time, on a yearly basis, and are there trends?
2. Where do certain disasters occur most often and can they be highlighted?
3. Can future disasters be predicted/modeled and then visualized geospatially?

The workflow was as follows:

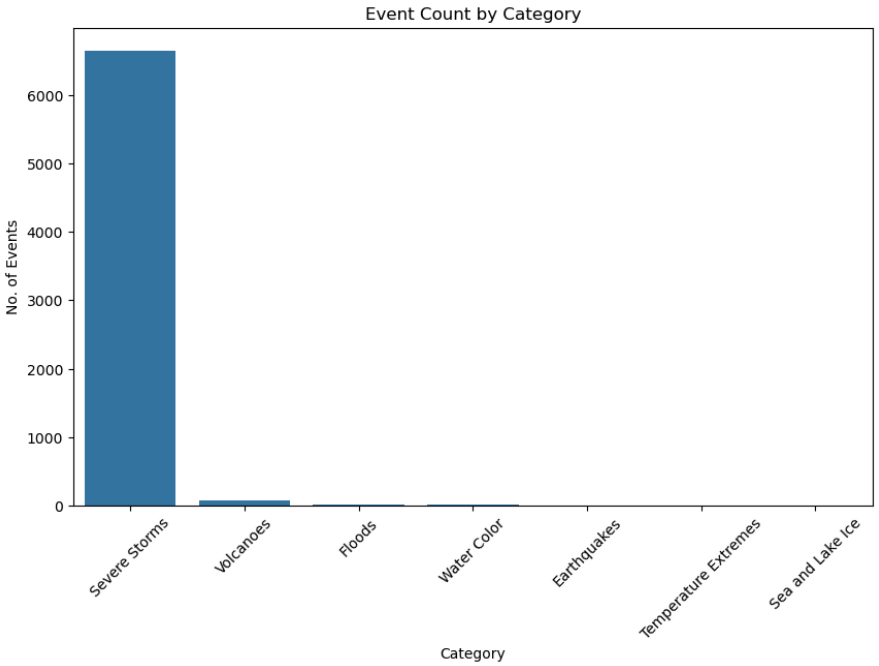
a. Data Collection and Wrangling: Fetching disaster event data, cleaning it, saving it to a local csv file for analysis.
b. Statistical Summaries: Exploratory data analysis of distributions of data and temporal patterns.
c. Geospatial Visualization: Visualizing hotspots of certain events using mapboxes.
d. Predictive Modeling: Utilizing a K Nearest Neighbors (KNN) model to classify and predict future events based on previous history.

Data Collection and Preprocessing:

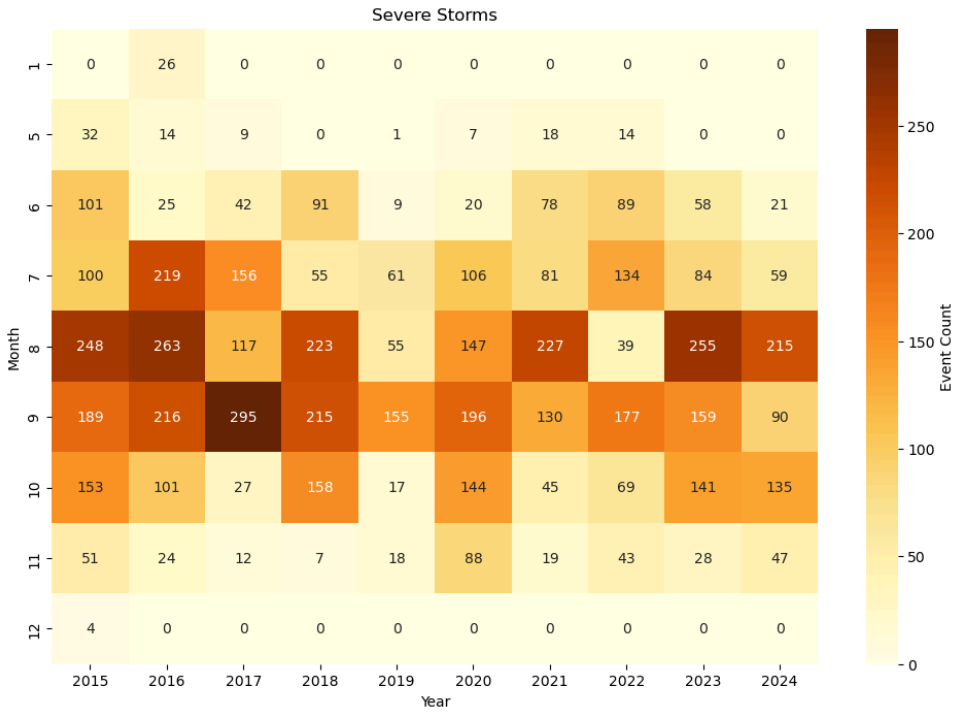
Using NASA' EONET API I fetched the previous 10 years of data regarding droughts, earthquakes, floods, severe storms, "manmade", seaLakeIce, snow, tempExtremes, volcanoes, and waterColor. I first used a function *fetch_events* to fetch data using parameters regarding the disaster category, the time period, and the location. This data was then cleaned and normalized using the *process_event_data* function, based on the various geometry types that the API returned. For example, polygons' centers were calculated and the corresponding latitude/longitude was stored. The result of this step was a cleaned dataset *all_disasters_10_years.csv*, which contained a total of 6,757 events.

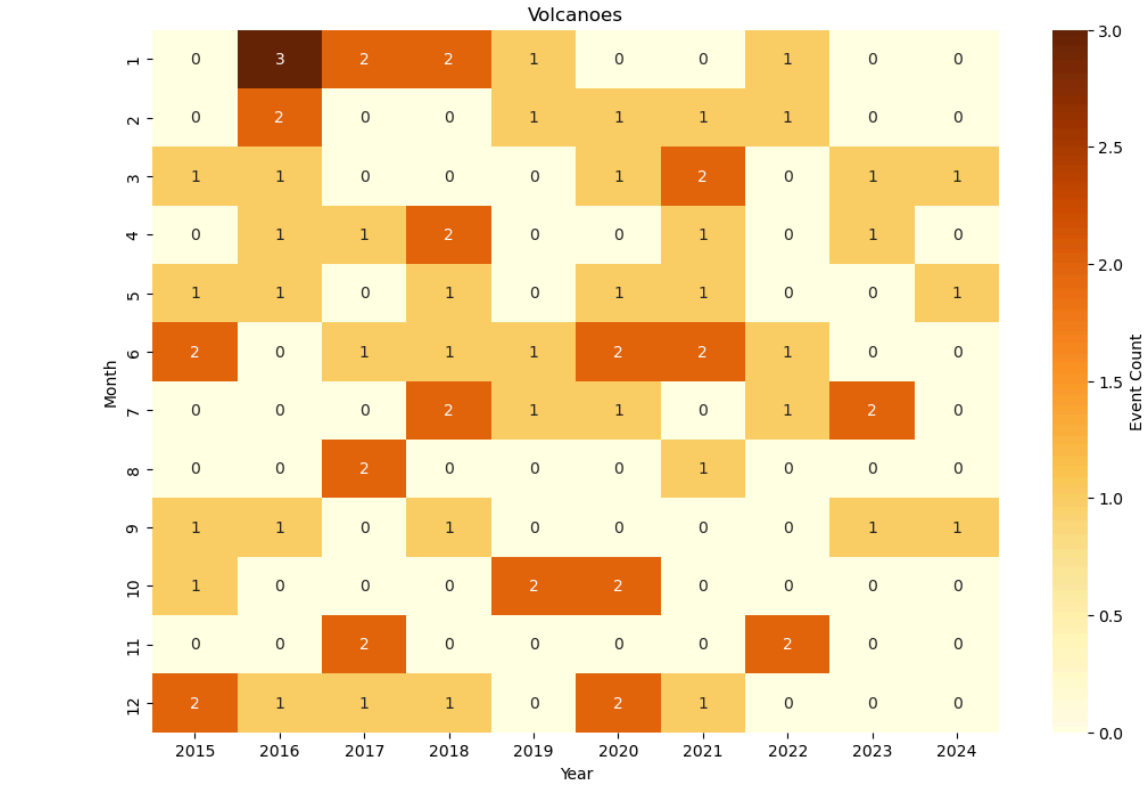
***Exploratory Data Analysis - Statistical Summaries*:**

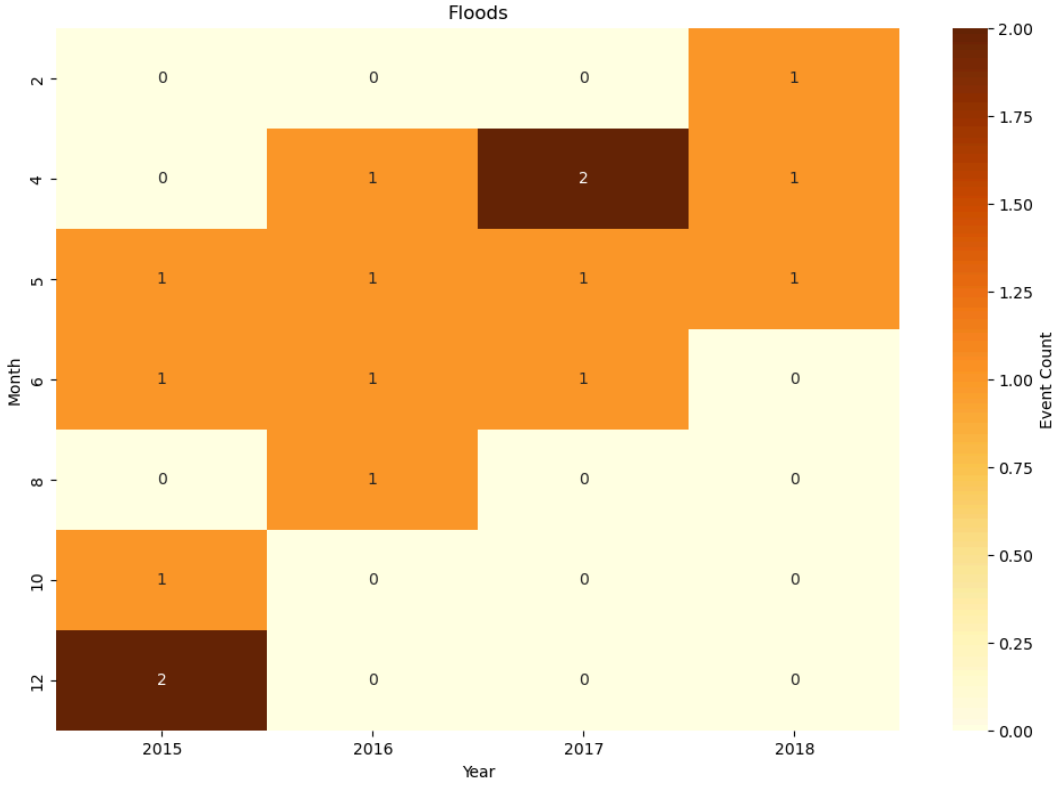
The next step was conducting EDA on the dataset, which revealed significant imbalances across the disaster categories. A large amount of events were of the Severe Storm category, as 6,652 of the data points were Severe Storms. This imbalance can be visualized using a bar plot:



Seeing as though this was not something that could be remedied, it simply was the data set, I set forward to explore temporal trends. Categories with more than 10 events were analyzed using heat maps to showcase yearly trends. This yielded important data as Severe Storms showcased clear seasonal trends, with high activity shown in months from July - September, holding in line with scientific basis. In contrast, events like volcanic events were sparse and irregular in activity, with no temporal trends to extract. These heatmaps (shown below) are a detailed view of disaster frequencies over months and years.

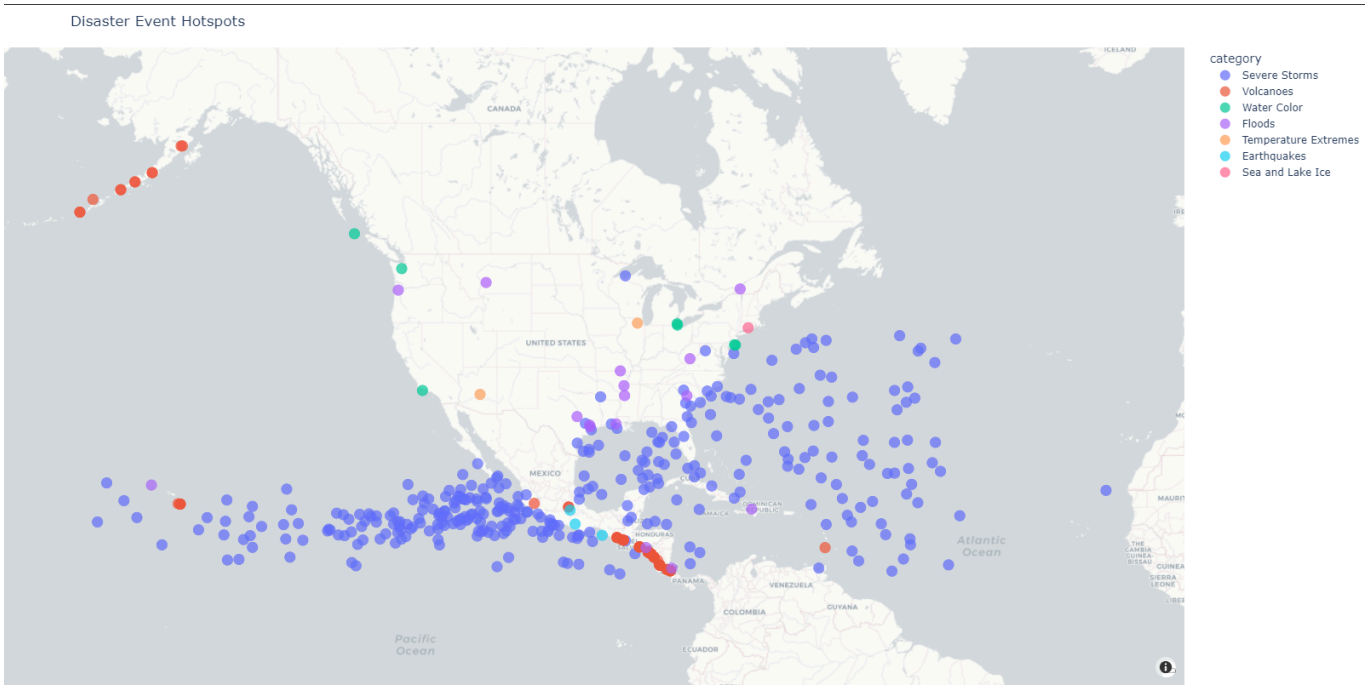



Volcanoes
Month
Year
Event Count
1 0 3 2 2 1 0 0 1 0 0
2 0 2 0 0 1 1 1 1 0 0
3 1 1 0 0 0 1 2 0 1 1
4 0 1 1 2 0 0 1 0 1 0
5 1 1 0 1 0 1 1 0 0 1
6 2 0 1 1 1 2 2 1 0 0
7 0 0 0 2 1 1 0 1 2 0
8 0 0 2 0 0 0 1 0 0 0
9 1 1 0 1 0 0 0 0 1 1
10 1 0 0 0 2 2 0 0 0 0
11 0 0 2 0 0 0 0 2 0 0
12 2 1 1 1 0 2 1 0 0 0
2015 2016 2017 2018 2019 2020 2021 2022 2023 2024
3.0 2.5 2.0 1.5 1.0 0.5 0.0


Floods
Month
Year
Event Count
2 0 0 0 1
4 0 1 2 1
5 1 1 1 1
6 1 1 1 0
8 0 1 0 0
10 1 0 0 0
12 2 0 0 0
2015 2016 2017 2018
2.00 1.75 1.50 1.25 1.00 0.75 0.50 0.25 0.00

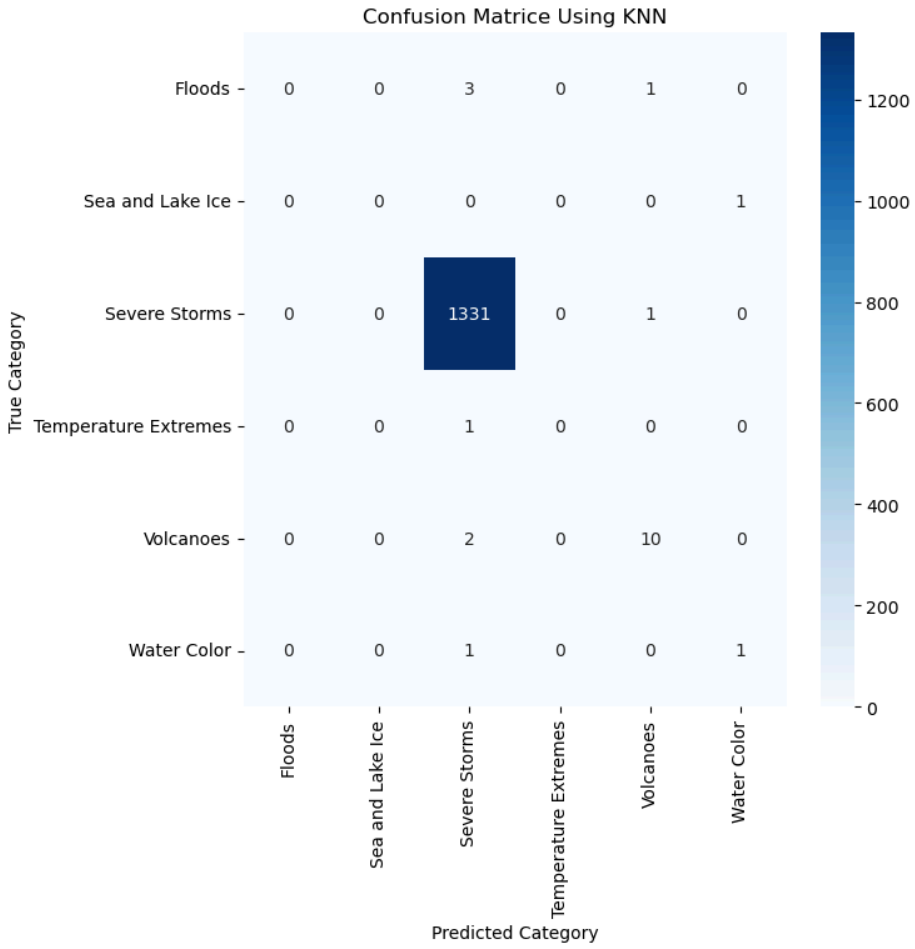
GeoSpatial Analysis:

Geospatial visualizations were created to identify disaster event hotspots across North America. Duplicate events, such as those found in severe storms, were removed and the average coordinates were calculated to normalize this data. The Mapbox (shown below) effectively shows the hotspots, as Severe Storms dominate the central and southeastern United States. Other events such as volcanic activity are shown in Alaska, following the Ring of Fire. This visualization was important for showcasing the concentration of each type of event, even if they occurred more sparsely than Severe Storms.

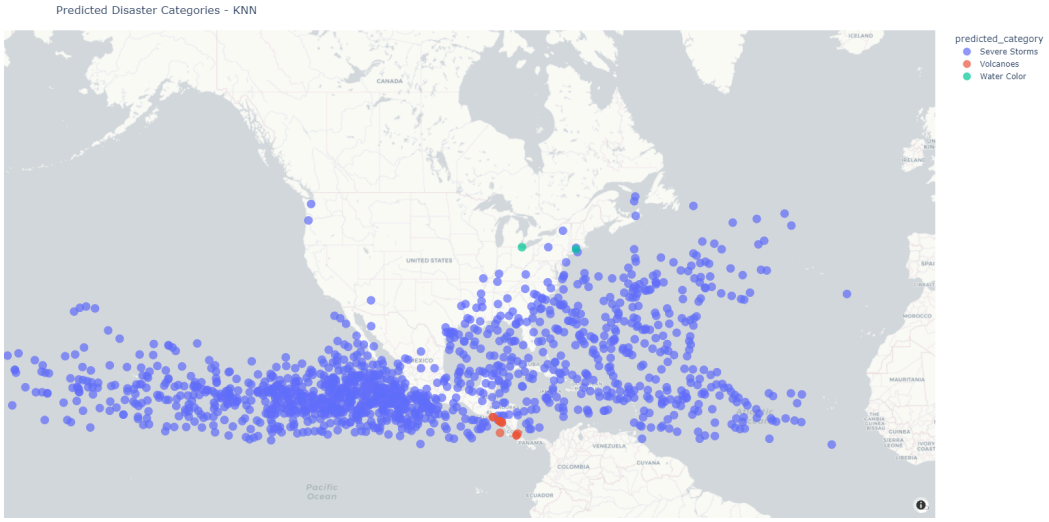


Predictive Modeling:

To predict future events, the K-Nearest Neighbor Algorithm (KNN) was applied, taking in features such as *latitude, longitude and geometry count*. First, disaster categories were numerically encoded using *LabelEncoder*. The feature values were then standardized to ensure equal weights, which was done to address the imbalance in the dataset. Finally, the dataset was split 80% training and 20% testing, and the model was trained with a K-value of 5. The model did obtain an overall accuracy of 99.3%, however this high accuracy was misleading as it was a result of the large amount of Severe Storm data points. While Severe Storms were classified near perfectly, categories like Floods suffered. The model was a reflection of the dataset, as seen in the confusion matrix (shown below). This matrix visualizes the severe imbalance, as the predictors were heavily skewed towards Severe Storms.



Finally, a scatter plot was generated using MapBox(shown below) to visualize the KNN predictions. Here we can see that the visualization reinforces the model's bias towards Severe Storms, but does have a few minority predictions that seem accurate based off of previous data. In all, the predictive model did the best it could based on the imbalance in the training data, shown in both visualizations.



Results:

Circling back to the 3 main questions, we can now answer them based on the results.

1. How does the frequency of certain natural disasters change over time, on a yearly basis, and are there trends?
    a. The seasonal analysis revealed that Severe Storms follow yearly trends. Using heatmaps, we were able to extract a peak between the months of July - September, aligning with the idea of the "storm season." Other disaster types, such as Volcanoes or Floods did not yield any significant temporal data, likely due to their sporadic nature.
2. Where do certain disasters occur most often and can they be highlighted?
    a. Geospatial analysis using Mapbox visualizations was able to successfully highlight disaster hotspots across North America. Severe Storms primarily occurred in the Central and South East areas of North America. Volcanic hotspots were also properly visualized, as shown with Alaska and the Ring of Fire and parts of Hawaii. Thus, we were able to successfully geospatially model disaster hotspots in North America based on our dataset.
3. Can future disasters be predicted/modeled and then visualized geospatially?
    a. Using the KNN model, disaster events were classified based on features such as latitude and longitude. The model achieved a high accuracy rate of 99.3%, primarily due to the imbalance in the dataset. While predictions for Severe Storms were reliable, the model was unable to properly predict minority disaster categories. The resulting predictions were visualized geospatially, but the lack of diversity in events is clear. Thus, we were not able to accurately predict future events for the vast variety of categories.

Challenges:

The project faced several challenges, starting with the API rate limits. Wildfires were excluded from the data analysis as there were so many points that I was unable to scrape them for 3 months, making 10 years of datapoints impossible. Secondly, the wide variety of geometry types in disasters forced me to normalize them. I used some basic geometry to just find the average of the coordinates. However, this may have caused locations to shift for certain events, and my calculations may not have accurately reflected the nature of each disaster type. Finally, the dataset was imbalanced, a recurring and severe issue. This heavily affected the predictive modeling step, and I tried to address it by scaling the data and using the confusion matrix for a more holistic look at the resulting model.

Future Work:

If I were to continue this analysis, I would address the dataset's class imbalance by oversampling the data, ensuring that the minority classes had more training data for the model to work off of. To build off this, I would use a more advanced predictive model, such as Neural

Networks, as they can be further fine tuned and result in better classification performance. Finally, I would look further into ensuring that the geometry of the data points was preserved, ensuring that the mapboxes can properly showcase the paths/areas which different events affect in unique ways.