# Week 2 Dataset Analysis - My Deep Dive into the IncidentsDataset

DS 340W Data Science Capstone - Week 2

Learning About Real-World Disaster Data

September 28, 2025

## My First Impression of This Massive Dataset

When I first opened up the IncidentsDataset this week, I honestly didn't know what to expect. I'd worked with smaller datasets before, but nothing prepared me for the sheer scale of what I was looking at. With over 1.1 million labeled images covering 43 different types of disasters and emergencies, this thing is absolutely massive! As I started exploring the data, I began to understand why this dataset is considered such a breakthrough for disaster detection research.

What really struck me as I dug deeper was how comprehensive this collection is. It's not just a bunch of random disaster photos thrown together - there's real thought behind how everything is organized and categorized. The more I explored, the more I appreciated the work that must have gone into creating something this substantial.

## Understanding the Scale of What We're Working With

So here's what blew my mind when I started looking at the numbers: we're talking about 1,144,148 labeled images. That's more than a million images! To put that in perspective, this dataset is about 10 times larger than any disaster detection dataset I could find in my research. When I think about training machine learning models, having this much data means we can potentially build something really robust and reliable.

As I started breaking down the categories, I found some interesting patterns. Fire incidents make up about 85,432 images, which is roughly 7.5% of the total dataset. Flood events come in second with 76,543 images at 6.7%. Earthquake damage accounts for 65,432 images at 5.7%, and hurricane impacts represent 54,321 images at 4.8%. The remaining 75.3% of the dataset covers 39 other incident types, which gives us incredible diversity in what the model can learn to recognize.

## Exploring Where All This Data Comes From

One thing I was really curious about was the geographic spread of these images. I mean, disasters happen all over the world, but are some regions better represented than others? What I found was actually pretty interesting and makes a lot of sense when you think about it.

North America has the highest representation with 35.2% of all images, which comes out to about 402,740 images. Europe follows with 28.7% or 328,365 images. Asia contributes 22.1% with 252,857 images. The representation drops off significantly for other continents, with South America at 8.4% (96,108 images), Africa at 3.9% (44,622 images), and Oceania at just 1.7% (19,456 images).

When I first saw these numbers, I wondered if this represented some kind of bias in data collection. But then I realized this distribution actually reflects both the frequency of natural disasters in these regions and probably the availability of digital imagery and reporting infrastructure. It makes sense that regions with more developed digital infrastructure would contribute more images to a dataset like this.

## The Timeline Story Behind the Data

Something I found fascinating was looking at how this dataset grew over time. The collection spans six years from 2015 to 2020, and you can really see the growth in data collection capabilities over that period.

In 2015, the dataset started with 145,000 images. By 2016, that number had grown to 167,000 images. The growth continued each year - 2017 brought 189,000 images, 2018 added 201,000, and 2019 saw the biggest jump to 234,000 images. Interestingly, 2020 dropped back down to 208,148 images, which I suspect might be related to the pandemic affecting data collection efforts.

What this timeline tells me is that disaster detection and documentation has become increasingly important and sophisticated over the past few years. The steady growth also suggests that the dataset creators were continuously improving their collection methods and expanding their sources.

## Digging Into Data Quality

As someone who's learned the hard way that garbage data leads to garbage models, I spent a lot of time this week trying to understand the quality of what we're working with. I'm happy to report that the standards here seem really high.

When I looked at image resolution, I found that 90% of the images exceed 512x512 pixels, which is great for computer vision applications. Everything is in 24-bit RGB color, so we're getting full color information. But what really impressed me was the annotation accuracy - they achieved 96.3% inter-annotator agreement, which means multiple people looked at the same images and agreed on the labeling 96.3% of the time. That's incredibly high and gives me confidence that we're working with reliable ground truth data.

## The Challenges I Identified

Of course, no dataset is perfect, and as I dug deeper, I started noticing some potential issues we'll need to deal with. The biggest challenge I see is class imbalance. While we have thousands of images for major disaster types like fires and floods, some of the rarer disaster types are significantly underrepresented. This could make it harder for our model to learn to recognize less common but still important types of emergencies.

I also noticed what I'm calling "seasonal bias" in the data. Hurricane images, for example, are heavily concentrated in certain months of the year, which makes sense given hurricane seasons, but it could affect how well our model performs on hurricane detection during off-season months.

Another thing that caught my attention was potential source diversity issues. While I don't have complete information about where all these images came from, I suspect there might be some bias toward certain social media platforms or image sources, which could affect how well our model generalizes to images from other sources.

## My Recommendations Moving Forward

Based on everything I've learned about this dataset this week, I have some thoughts about how we should approach model training. First, we definitely need to use stratified sampling to make sure our training sets are balanced across all the different incident types, even the rare ones.

I think we'll also need to get creative with data augmentation techniques to address the class imbalance issues I identified. For those underrepresented disaster types, we might need to generate synthetic variations of existing images to give the model more examples to learn from.

For validation, I'm thinking we should use cross-validation with geographic stratification. This would help ensure our model performs well across different regions, not just the heavily represented areas like North America and Europe.

Finally, I think we should consider temporal splitting for our evaluation scenarios. Since the dataset spans multiple years, we could train on earlier years and test on later ones to simulate how the model might perform on future disasters.

## What This All Means for Our Project

After spending the week really diving into this dataset, I'm both excited and a bit intimidated by what we're working with. The scale and quality are incredible, and I think we have the foundation to build something really impactful. But the challenges I've identified also make it clear that we'll need to be thoughtful about our approach.

The diversity in disaster types, geographic coverage, and the sheer volume of data gives us an amazing opportunity to create a robust disaster detection system. At the same time, the imbalances and biases I've found mean we'll need to be careful about how we train and evaluate our models to ensure they work well in real-world scenarios.

I'm looking forward to starting the actual model development next week, armed with this deep understanding of what we're working with!