Motivation

- **For what purpose was the dataset created?**
    - The Storm Data was created as a historical record for storm events in the United States. It included data such as property damage, loss of life, type of storm, and storm intensity, and is used to analyze past storms and their behavior.
    - Climate Data is a combination of datasets which results in combined global land and ocean temperature anomalies. It is specifically designed for the sake of studying climate change and variability.
- **Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
    - Storm Data is a publication of the National Oceanic and Atmospheric Association (NOAA), and was compiled via data from the National Weather Service, which in turn received information from a variety of sources including county, state, and federal emergency officials, law enforcement officials, damage surveys, newspapers, and the general public.
    - Climate Data obtained its data from both the Global Historical Climatology Network-Monthly (GHCN-M) dataset, and the International Comprehensive Ocean-Atmosphere Data Set (ICOADS).
    - They are both published on the NOAA website.
- **Who funded the creation of the dataset?**
    - Both datasets are published by NOAA, which is funded by the US government.

Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**
    - The storm dataset instances represent each storm event.
    - The climate dataset instances represent each year for each state.
- **How many instances are there in total (of each type, if appropriate)?**
    - There are a total of 1,538,979 unfiltered instances for storm data.
    - There are a total of 6,046 unfiltered instances of climate data per state and year.
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
    - The datasets contain all possible instances gathered from the source.
- **What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.**
    - Each instance consists of unprocessed text.
    - For Storm Data, property damage and crop damage represent money in dollars. Deaths signify people. Storm intensity is based upon the Fujita tornado scale, with six categories of storm intensity. Year represents the year the event took place, and state represents the US state the event occurred in.
    - For Climate Data, the year and state represent the year and state the anomaly and average temperature were taken. The average temperature represents the temperature at the specified time and state, and the anomaly represents how far the average temperature is from the climate average.

- **Is there a label or target associated with each instance? If so, please provide a description.**
    - There are multiple columns with a varying number of available instances.
    - For storms, there is the year it took place in, property damage, direct event deaths, crop damage, and the state the event occurred in.
    - For climate, there is the year, average temperature, state, and anomaly.
- **Is any information missing from individual instances?**
    - There is information missing from individual instances. Not all instances have known damage to property, crops, or human lives. Not all storms fall under the Fujita tornado scale either.
- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**
    - There are no known relationships between individual instances.
- **Are there recommended data splits (e.g., training, development/validation, testing)?**
    - There are no data splits.
- **Are there any errors, sources of noise, or redundancies in the dataset?**
    - There are very few errors where values are unclear or not clearly specified. For example, there was one invalid character 'H' instead of a number in the field for property damage. There was no explanation provided for this 'H'.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**
    - Both datasets are self-contained.
- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?**
    - Neither datasets contain any information that is confidential.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
    - No, neither dataset contains offensive, threatening, or anxiety causing data.
- **Does the dataset relate to people?**
    - Storm Data identifies deaths and injuries per storm event. Climate Data does not relate to people.
- **Does the dataset identify any subpopulations (e.g., by age, gender)?**
    - No subpopulations are identified.
- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**
    - It is impossible to identify individuals from the datasets.
- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**
    - The datasets do not contain data that is sensitive

Collection Process

- **How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly**

inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
    - Storm Data dataset was created through public knowledge, US weather stations and centers, newspapers, and government officials.
    - Climate Data was created from two combined climate datasets, reported by national and international climate centers.
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?**
    - All data was collected from the NOAA website.
    - For temperature data, we wrote a script that would go to the NOAA website and automatically select each state and download the temperature data in a text file. The text file would then be turned into a CSV.
        - We used Selenium, BeautifulSoup, Requests and CSV python library to automate browser interactions.
    - For the storms data set, we wrote a script that auto-clicks links and saves the file to a respective folder. We did this because there were 300+ links.
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
    - The datasets are not samples from a larger set.
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**
    - Unknown
- **Over what timeframe was the data collected?**
    - The temperature data ranges from 1850 to 2018
    - the storms data rages from 1950 to 2019
- **Were any ethical review processes conducted (e.g., by an institutional review board)?**
    - This information is unknown.
- **Does the dataset relate to people?**
    - The storm data contains basic information on casualties and injuries.
- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
    - Temperature data was collected from https://www.ncdc.noaa.gov/cag/statewide/time-series/41/tavg/12/12/1895-2019?base_prd=true&firstbaseyear=1896&lastbaseyear=2000
    - The storms data was collected from https://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/
    - Distances between individual states were taken from https://www.mapdevelopers.com/distance_from_to.php
- **Were the individuals in question notified about the data collection?**
    - Information is unknown.
- **Did the individuals in question consent to the collection and use of their data?**
    - No, as this information is publicly known and was reported to government entities.

- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis been conducted?**
    - Information is unknown.

Cleaning

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**
    - The datasets were cleaned and split into easier to manage parts.
    - For Storm Data:
        - The Fujita tornado scale was changed to a numerical system for easier calculations.
        - If the crop damage or property damage raw values had money characters such as k, m, b, or t (kilo, million, billion, trillion), the characters were removed and the numbers were multiplied with the corresponding base 10 value.
        - When collecting data for a specific column, if a row had an invalid or no value for that column, the row was manually removed for that dataset.
    - For Climate Data:
        - Each file had the US state name on top of the data table. This state name was cut and added as a column in the data table, with each row having that state name as another field.
        - Each file had the starting rows skipped, as they were not part of the data table, and were interfering with data collection.
        - There was one file with empty fields in the dataset, which had to be manually removed.
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**
    - The raw, original data was saved.
- **Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.**
    - All commands, software, and packages used to clean these datasets

Uses

- **Has the dataset been used for any tasks already?**
    - The article and notebook using these datasets will be included with this datasheet.
- **Is there a repository that links to any or all papers or systems that use the dataset?**
    - The article and notebook using these datasets will be included with this datasheet.
- **What (other) tasks could the dataset be used for?**
    - These datasets could be used to judge property damage and how climate change affects certain states more compared to other states. It could also be used to identify clusters of damage over time, and the clusters that have the biggest growth.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
    - No, collection and processing will not impact future uses.