Estimating Damage from Natural Disasters in the USA

#### Visualization Link: <http://noaa-storms-viz.herokuapp.com>

#### Github Link:<https://github.com/ss-github-code/noaa_storm_analysis>

## Motivation

The National Oceanic and Atmospheric Administration (NOAA) in the USA releases the Storm Events Database that records the occurrence of storms and other significant weather phenomena having sufficient intensity to cause damage to property and/or crops. Our primary objective was to explore three major types of storms that cause widespread economic damage in the USA: tropical cyclones and floods (including hurricanes), severe local storms (including tornadoes), and wildfires and droughts. Our project focused on exploring varied independent variables like population, economic activity, and associated weather statistics that affect the total damage caused by these natural disasters in the USA.

While it is not possible to predict where and when the next natural disaster will strike, it is possible to develop models that can predict the amount of damage from such disasters. These predictive models can be useful to plan for better emergency management as the nation's increasing population grapples with the increase in both the frequency as well as the intensity of storms caused by global warming. In addition, local governments can use the models to plan for disruptions by using these models for what-if analysis.

This project aimed to answer the following questions:

* What are the various factors that contribute to the amount of damage from storms in the USA? These factors include population, economic activity in a county, and the weather related statistics such as precipitation and temperature.
* How to collect and visualize the input features related to natural disasters for the period 2000-2021.
* Is there a meaningful correlation between the input features and the damage caused by the storms?

## Data Sources

#### National Oceanic and Atmospheric Administration (NOAA) Storm Events Database

This database includes details on the occurrence of storms that cause loss of life, significant property damage, and disruption to commerce in the counties and forest zones in the USA.

1. **Name**: Storms Events Database
2. **Data location**: <https://www.ncdc.noaa.gov/stormevents/ftp.jsp>
3. **Format**: CSV files (one per year)
4. **Important variables**: State Name, County Name, State Federal Information Processing Standards (FIPS) ID and County FIPS ID), Event Type, Damage Property, Damage Crops, Begin Date and End Date.
5. **Time period used**: 2000-2021
6. **Size**: 1.5 GB (approximately before pre-processing)
7. **Total records**: 13,274 records, 3.3 MB (post pre-processing)
8. **Access method**: Files were downloaded from the FTP site, then records were preprocessed and stored in a relational database (AWS RDS).

#### US Census.gov

In order to explore the socio-economic factors that impact the damage from the natural disasters, we needed to access several secondary datasets. The Census Bureau’s Population Estimates Program dataset provided the population estimate in the county affected by the storm for the year of the event (between 2009-2019). The Decennial Census provided the population for 2000 and 2010 and we interpolated the data for the years 2001-2008. The Census Bureau’s County Business Patterns dataset provided the number of businesses, their number of employees and their total payroll in the county. Its Non-Employers dataset provided the number of small businesses and the total revenue generated by them. Finally, its Economic Census provided the economic activity for the top 3 industries in the county.

1. **Name**: Population Estimates Program, County Business Patterns, Economic Census
2. **Data location**: https://api.census.gov/data
3. **Format**: Python string processed into CSV
4. **Important variables**: Population estimate, number of establishments and their total number of employees and their total payroll, number of non-employer establishments and their total revenue
5. **Time period used**: 2000-2021
6. **Total records**: one per event in the storms database
7. **Access method**: Data was downloaded using the API, preprocessed and saved into the database (Amazon RDS)

#### US Drought Monitor (USDM)

The US Drought Monitor publishes a map showing the severity and location of drought throughout the country. We use the drought severity data for the county where the wildfire event occurred as input to the model for predicting the damage from wildfires.

1. **Name**: US Drought Monitor
2. **Data location**: <https://droughtmonitor.unl.edu/>
3. **Format**: Python string processed into CSV
4. **Important variables**: Drought intensity level
5. **Time period used**: 2000-2021
6. **Total records**: One map per wildfire event in the storms database
7. **Access method**: Data was downloaded from the website and saved as CSV files in Github.

#### Weather.gov

The primary dataset is missing the county FIPS, latitude and longitude of storm events that occur in a forest zone (a National Weather Service forest zone covers more than one county). We used the zone county correlation dataset to get the FIPS, latitude and longitude of a given zone FIPS ID for an event in the storm events dataset.

1. **Name**: Zone County Correlation File
2. **Data location**: <https://www.weather.gov/gis/ZoneCounty>
3. **Format**: Pipe delimited text
4. **Important variables**:list of County FIPS in a zone, their latitude and longitude
5. **Access method**: File downloaded from the website and saved in Github.

#### National Center for Environmental Information (NCEI)

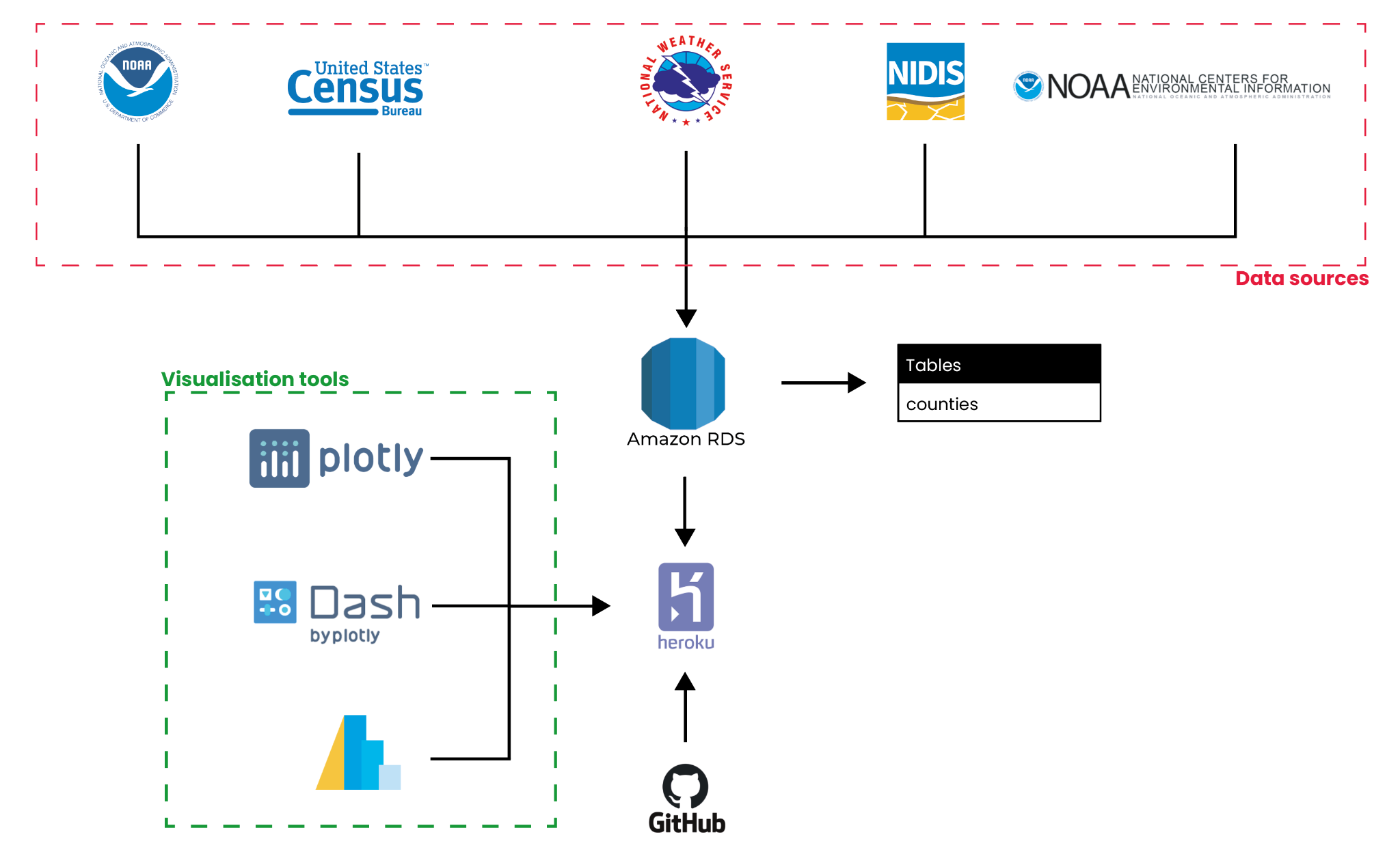
In order to get the weather records for a county, we needed the list of weather stations in and around an affected county. From the list, we find the nearest weather station that has daily weather summaries for the past 10 years from the date of the event. We obtain the precipitation, snowfall and temperature data for the affected county.

1. **Name**: Climate Data Online
2. **Data location**: <https://www.ncdc.noaa.gov/cdo-web/>

<https://www.ncei.noaa.gov/access>

1. **Format**: Python string processed into CSV
2. **Important variables**: Date, precipitation, snow, snow depth, temperature max and min
3. **Time period used**: 2000-2021
4. **Total records**: approximately 3653 records for each of the 407 wildfire events
5. **Access method**: Data was downloaded using the NCDC and NCEI API and saved as CSV files in Github.

## Application Architecture

Below is the architecture of our live dashboard application:

# Data Manipulation

### Data helpers/packages (files are located in the /helper folder)

Since our datasets are quite large and not conducive towards local Jupyter manipulation at large, our data processing and manipulation jobs were chunked and we used several helpers and processes to move data for cleanup and visualization.

***How specifically did you need to manipulate the data?*** The total number of damage in the dataset was categorical in the dataset (using k to represent in 1000s). We had to change this to quantitative data. Also the damage was split into multiple columns. We calculated the total damage by adding all these columns. For the purpose of this project we look at events that only cross a certain threshold damage amount. We also had to manipulate the FIPS column in the dataset to match the plotly geojson data. Some events take place in multiple counties. The damage for these events are not available in the dataset. For the purpose of this project we divided the total damage equally for all counties in such cases.

***How did you handle missing, incomplete, or incorrect data?*** Some records had missing FIPS. We removed all such records. Any event for which the CZ\_Typecolumn as zorzonewould not have any fips information. This is because the event occurred in more than one county. In such cases the respective counties where the event occurred was found and the FIPS information for the county was added. In such cases the event will be duplicated for the total number of counties the event occurred. We also only worked with the data for the fifty States in the country. Any other region was ignored.

***How did you perform conversion or processing steps?*** We coded some helper functions to convert the total damage from categorical string columns to quantitative floating point values. The dataset only offers us the columns such as YEARMONTH**,** DAY and TIME. we had to use this information to create columns to get the datetime columns. Pandas offers us multiple methods to calculate this. We also calculated the total duration of the event. We first calculated the START\_DATETIME and END\_DATETIME and used these columns to calculate the duration of the event.

***What variables and steps did you use to join the data resources to perform your data analysis?*** For this dataset the FIPS information was the most useful column. We had to use the state FIPS (STATE\_FIPS) and county FIPS (CZ\_FIPS) to convert this into data similar to the geojson format. The new column will later be used for plotting information on a map. Also the YEARMONTH and TIME columns were used to find the starting and ending time of a given event.

***Briefly describe the workflow of your source code and what the main parts do.*** The main function in the dataset is the get\_storm\_data. This function will get the storm data from the NOAA website for a given year and manipulate some of the columns such as converting damage from string to floating point, calculating the duration of a given event in the data, calculating the FIPS information to match geojson data, and ignoring records that are out of scope for this project. Please take a look at the attached notebooks for detailed information on what was specifically done and how it was achieved.

***What challenges did you encounter and how did you solve them?*** One of the primary issues with the dataset was that a lot of manipulation was involved to get the data in a format that will be usable. Some information was not available like how to calculate the total damage if the event occurred in more than one county. For the purpose of this project, we have assumed that the damage will be equally divided between the counties. This assumption is because no information on how damage in such cases are calculated was available.

***How to access the data?*** The data can be visualised using the dashboard we created. The code to manipulate the data is available in the corresponding notebooks. Because all data is publicly available, one needs to only run the notebooks to get the data in the format we have used. The dataset is also available in a postgres database in AWS.

### US Census.gov

***How specifically did you need to manipulate the data?***  The census.gov website does not offer yearly data for years before 2010. One way around this problem is to use the 2000 data and populate the information for the years till 2010. When populating data this way, we will have to account for inflation. For this purpose, we used CPI information and calculated the census information that accounts for inflation.

***How did you handle missing, incomplete, or incorrect data?*** As mentioned above, census data between 2000 and 2010 was not available. We had to manually calculate this information and account for inflation. Census data for a few counties was also missing in the website. So while downloading all the data, we added the information for these FIPS records as NaNs.

***How did you perform conversion or processing steps?*** All of the data required for the project was available through the census.gov APIs. When downloading, we put all the information such as the population, business establishments, employee statistics etc in a single column separated by a ‘|’. This column will then later be split when needed for analysis and visualization purposes. Multiple pandas methods such as apply and concat were used to get the data in a format we needed. To look at all methods used, take a look at the accompanying notebooks.

***What variables and steps did you use to join the two data resources to perform your data analysis?***  Our primary aim was to show the amount of damage the disasters cause on a county level. For this purpose we needed the population information, Total number of business establishments in the county, Employee statistics and other economic information. The FIPS ID from the primary dataset is used to compare the effect of a disaster in a certain county which is available from the census data.

***Briefly describe the workflow of your source code and what the main parts do.*** We wrote multiple helper functions to get the separate information such as population and employee statistics with the API. These functions are then called in a main loop to get all the information for a given year. All of this data was then uploaded to a postgres database hosted in AWS. Please take a look at the notebook for complete details on each helper function.

***What challenges did you encounter and how did you solve them?*** One of the main issues we faced was the missing data for years between 2000 and 2010. We had to manually calculate this information ensuring that we account for inflation. For this purpose we had to get the CPI information for each year and apply that to the inflation formula. Another challenge was to put this data in a format that we can easily use when we need them for visualization or analysis. Another issue we faced was that sometimes the total number of the primary business sectors will differ from county to county. We put all this information in a single column separated by a ‘|’ which can then easily be split later on.

***How to access the data?*** The data can be visualised using the dashboard we created. The code to manipulate the data is available in the corresponding notebooks. Because all data is publicly available, one needs to only run the notebooks to get the data in the format we have used. The dataset is also available in a postgres database in AWS.

### U.S. Drought Monitor (USDM)

How specifically did you need to manipulate the data?

Get wildfire info from NOAA dataset

Get lat lon data from National weather service

How did you handle missing, incomplete, or incorrect data?

How did you perform conversion or processing steps?

What variables and steps did you use to join the two data resources to perform your data analysis?

Briefly describe the workflow of your source code and what the main parts do.

What challenges did you encounter and how did you solve them?

### Weather.gov

How specifically did you need to manipulate the data?

How did you handle missing, incomplete, or incorrect data?

How did you perform conversion or processing steps?

What variables and steps did you use to join the two data resources to perform your data analysis?

Briefly describe the workflow of your source code and what the main parts do.

What challenges did you encounter and how did you solve them?

### National Climatic Data Center (NCDC)

How specifically did you need to manipulate the data?

How did you handle missing, incomplete, or incorrect data?

How did you perform conversion or processing steps?

What variables and steps did you use to join the two data resources to perform your data analysis?

Briefly describe the workflow of your source code and what the main parts do.

What challenges did you encounter and how did you solve them?

# Analysis and Visualization

##### Analysis Steps

For our datasets, finding data within the appropriate location was very important, and most of our extract-transform-load routines were used in order to have working datasets that could be merged with each other.

For the float data (whether they be from Argo, GTSPP, or DFO Quebec):

* The raw datasets needed to be filtered to find data located within the Estuary and Gulf of St. Lawrence (“Gulf”). The shapefile was used as the filter.
* Once we had this subset of measurements from the Gulf, then further averaging of temperature and salinity by depth was done, with depth bins of 100 meters by day.
* This final dataset was averaged by the year and returned, which in turn was joined with the fish population dataset.

Fish data needed very little additional cleanup, as it came in a single CSV file. Averages were calculated by the year per fish species. Because every latitude/longitudinal measurement has a fish population measurement (even if 0), we made the decision that 0 figures did not count towards population measurements and therefore did not count towards the averages.

##### What Didn’t Work

### Because Argo data generally captures ocean data and the Gulf location is sheltered away from the Atlantic, very few float data points were available for the Gulf. Of the 620,000 initial data points filtered from Argo data in the ocean\_data table, only 528 measurements were ultimately found in the Gulf. We found acceptable supplements to the Argo dataset via GTSPP and DFO Quebec data, however.

##### What interesting relationships or insights did you get from your analysis?

# Statement of Work