TIME SERIES ANALYSIS OF ALL NATURAL DISASTERS SINCE 1900

Individual Project in Statistical Methods and Applications I
(Spring 2023)

By
Vaishnavi Asuri
Student ID : 110544621
Department of Data Science
University of Colorado, Boulder

EXECUTIVE SUMMARY

This project aimed to explore the EM-DAT Natural Disasters Dataset, which contains data on natural disasters worldwide over the past 100 years. The short-term goals were to understand the usability of the dataset, conduct statistical tests to establish relationships between different attributes, and explore the overall importance of features. The long-term goals were to build a pipeline to make the data more usable for various applications and assess the possibility of building a model to predict the extent of damages that a particular disaster can cause in a particular part of the world.

It was found that the data has limited usability due to a large number of missing values in several variables, and data collection methods need improvement. However, damages were well assessed within certain timeframes, and Africa and Asia are the worst-hit continents when it comes to the frequency and extent of disasters. The project conducted various statistical tests to check and establish relationships between different attributes and successfully built a pipeline to help clean, sort, and make the data more usable for various applications.

Additionally, it was also found that a model to predict the extent of damages that a particular disaster can cause is possible only with better data. The project also assessed the possibility of predicting which continents are prone to what kind of disasters so that governments can build better precautionary and disaster management measures. Asia, Africa, and America seem to have the most amount of disasters, and floods, storms, and earthquakes top the charts.

Ultimately, these main findings showed that data collection has improved considerably over time, but most factors that help make data-backed conclusions are left with much ambiguity. It is important to disclose data manipulation while presenting results to determine the credibility of the said results. Prediction is possible only with better data, but this still does not guarantee that vulnerable nations will be better prepared. Governments have to actively take action and fund third-world nations as natural disasters cannot be curbed. The biggest thing learned from this dataset is the fact that even after 100 years, a significant amount of damage and loss of life still happens, which shows the truly unpredictable nature of natural disasters.

TABLE OF CONTENTS

I.INTRODUCTION

The UN Office for Disaster Risk Reduction recorded 7,348 natural disasters from 2010-2019, causing 1.23 million deaths and affecting 4.2 billion people. Web resources like EM-DAT, GHCN-D, and NOAA contain databases on climate, natural disasters, and their impact.

EM-DAT (Emergency Events Database) is an essential resource for disaster risk reduction and management. The database provides comprehensive information on the location, type, and impact of disasters globally. Analyzing EM-DAT data is crucial for identifying disaster patterns and trends, assessing vulnerability, evaluating interventions, and informing policy and decision-making. It is therefore important to analyze EM-DAT data for evidence-based disaster risk reduction strategies and efficient resource allocation.

The primary aim is to conduct a Statistical and experimental project based on three reasons

- To understand the usability and potential of this dataset to forecast damages to property, the extent of damage and the ability to assess and come up with feasible precautionary measures.
- To conduct various statistical tests to check and establish relationships between different attributes through correlation and causality
- To take an exploratory dive into the dataset and understand the overall importance of features

The big-picture question(s) we are trying to answer are:

- To build a pipeline that can help clean, sort and make the data more usable for various applications.
- To assess the possibility to build a model that can help predict the extent of damages that a particular disaster can cause in a particular part of the world.
- Is there a way to predict which continents are prone to what kind of disasters so that governments can build on it to implement better precautionary and disaster management measures

II. DATA SOURCE

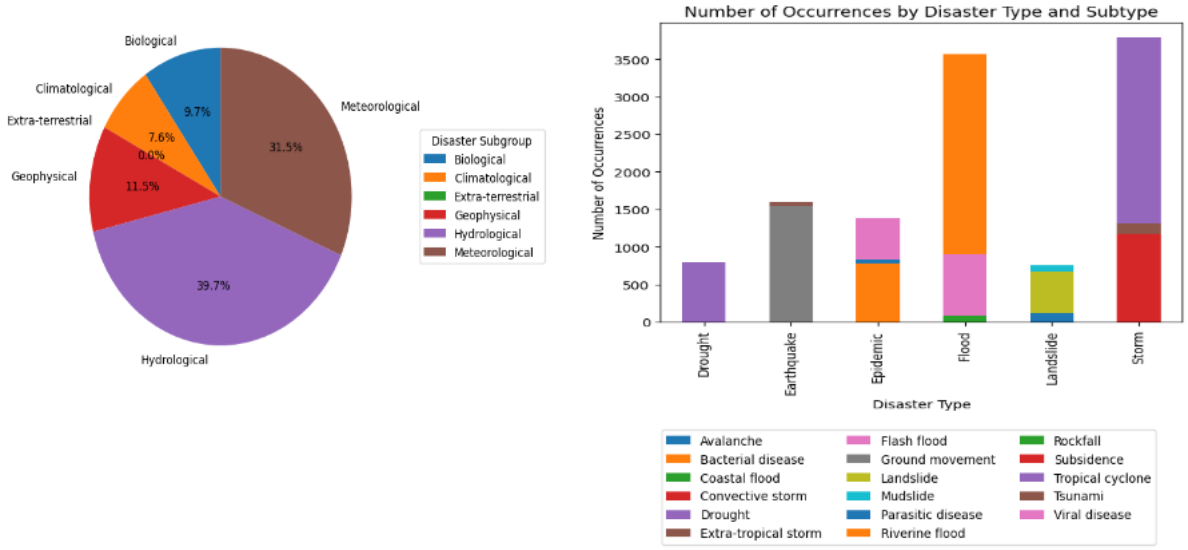
The current dataset from EM-DAT has a record of events starting from 1900 to 2023. The data includes:

- location (Country, Continent, Latitude, Longitude)
- time (Day, Month and Year of the beginning and end of the disaster)
- type (disaster type, subtype, group, sub-subtype)
- Damages (casualties, insured and total damages, injured people counts, etc) and

The dataset recorded close to 17,000 instances across 50 variables. It is raw and has many missing and null values. The plan was to try and clean this data and yield results with minimum possible manipulation.

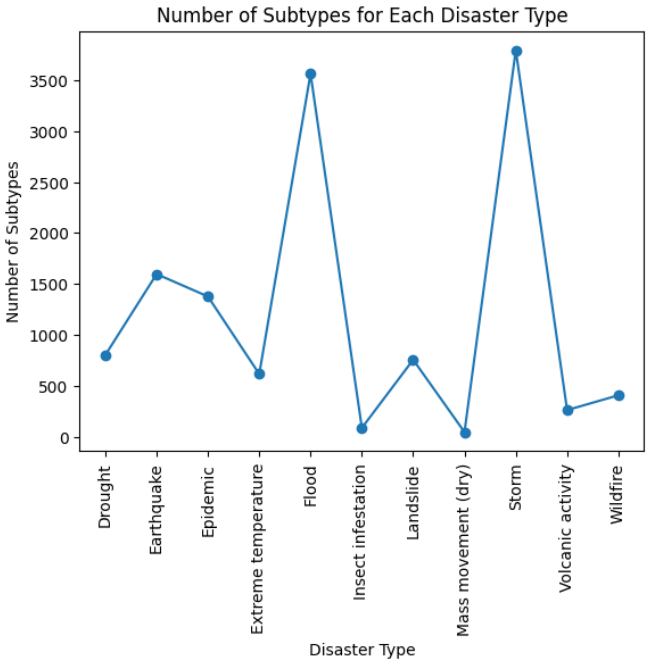
Variable Distributions

After some basic cleaning and pre-processing, I proceeded to perform EDA using grouping and value counts functions to get an idea of the data demographics

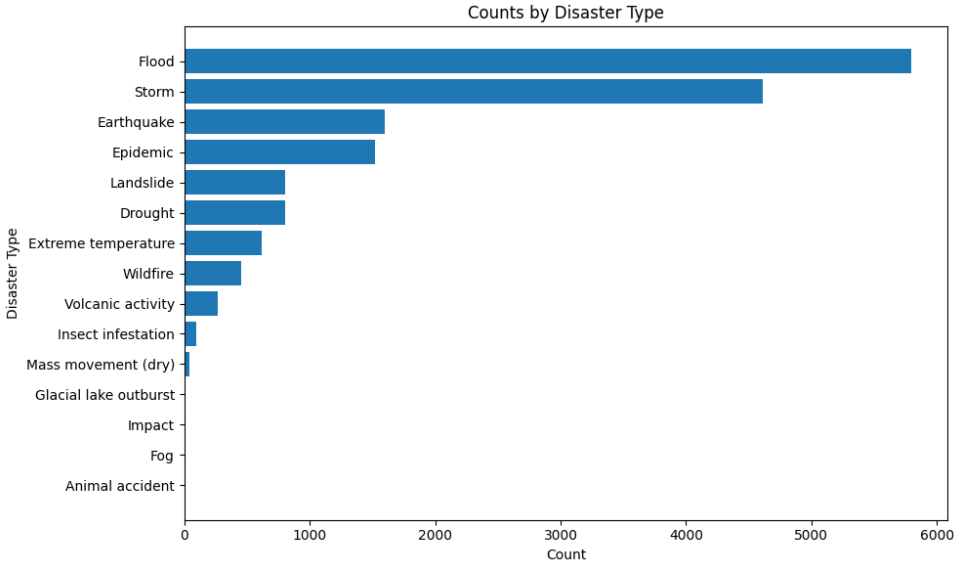


Fig(1) Overall distribution of Disaster types that are present in the current dataset.
Fig(2) Number of occurrence of disasters and their sub-types

A majority of these are Hydrological (40%), closely followed by Meteorological type of disaster. Out of these types of Disasters, I examined which sub-types are more frequently occurring than others. As fig.1 describes, Flash Floods and Bacterial Diseases seem to occupy a significant portion of our dataset. However, this might also be owed to the fact that floods and storms have the maximum data when it comes to subtypes as can be noticed in fig.2



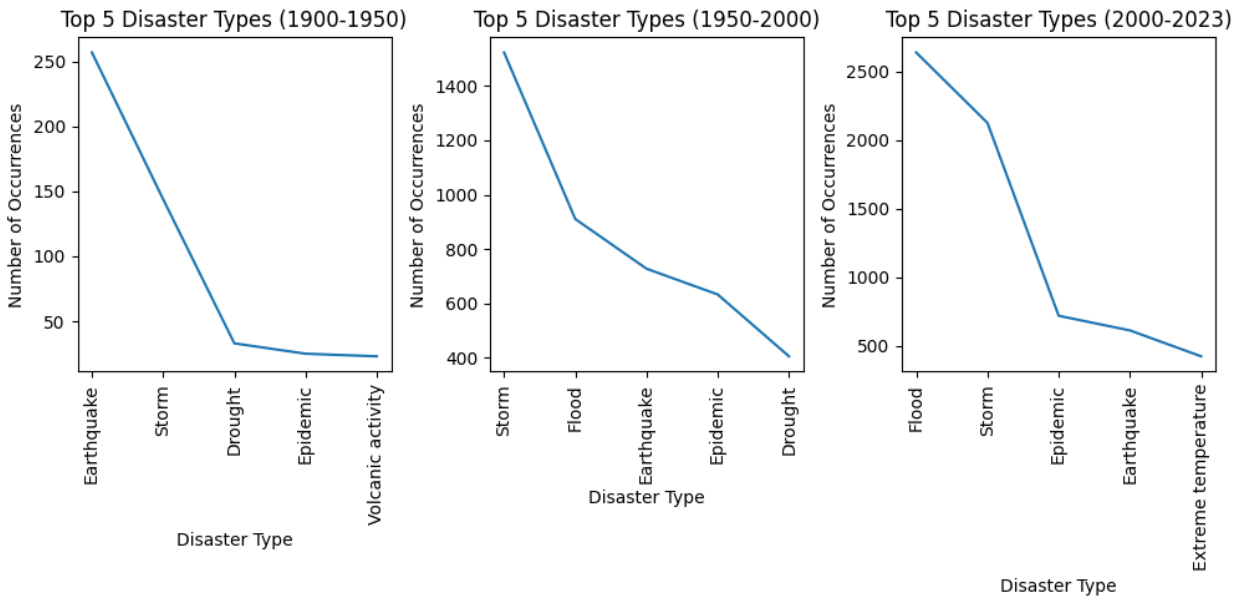
However, when we consider all the different sub-types of disasters by themselves without considering their types, we can notice that Floods and Storms take over the rest as shown in fig.3. This information is still consistent with what we inferred previously.



Fig(3) Counts of Various disaster types

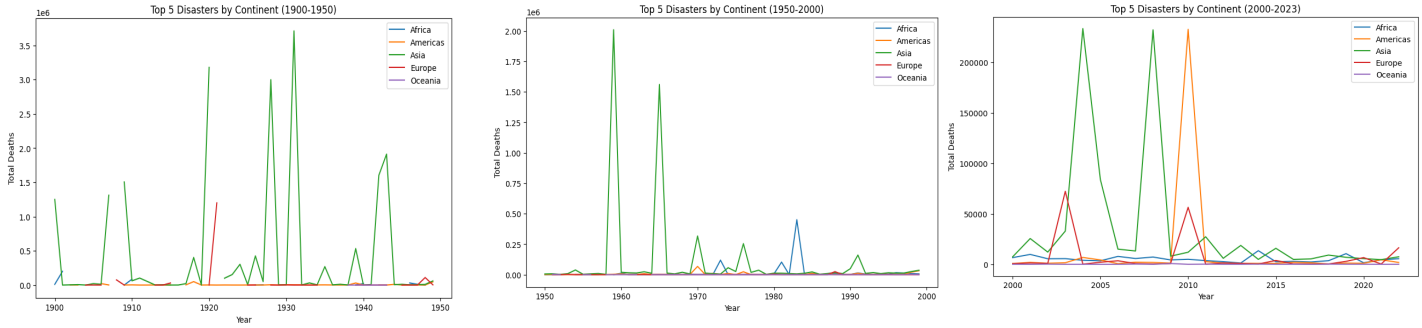
After this, I proceed to divide the dataset into three part with respect to the time periods. Since the data ranged from 1900 to 2023, I decided to divide it into 1900 to 1950, 1950 to 2000 and

2000 to 2023. Based on this, I performed some EDA to check for the top most documented disasters in each of those three periods.

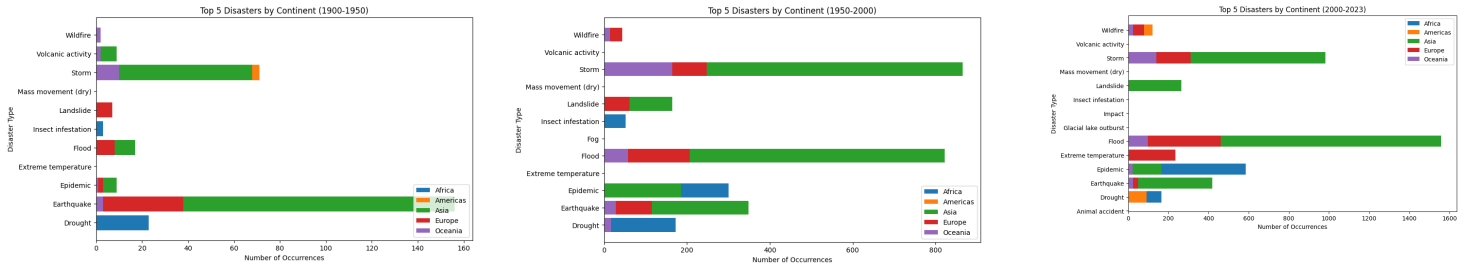


Fig(4) Counts of Various disasters through the years

This gives a sense of missing values and puts things into perspective for performing any further analysis. As you can notice from Fig. 4 Floods, Storms and Earthquakes dominate the records. However, since earthquakes are quasi-random phenomena, there has not been any significant increase or decrease in the past 100 years. Contrarily, fig.X shows that there has been an increase in earthquakes after 1950 but it pretty much remains the same after 2000. This points towards a clear lack of data as can be observed from the scaling of the plots. Yet again, this calls for a more completed dataset to make more concrete conclusions.



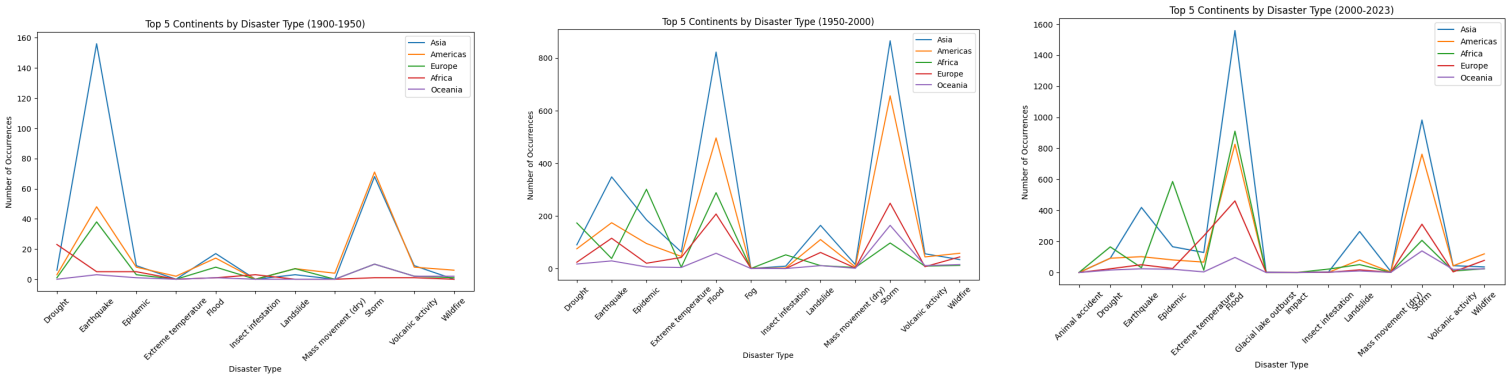
Fig(5) Counts of Various deaths through the three time periods



Fig(6) Top 5 disasters by Continent through the three time periods

With the same three-time period groups, I decided to perform a continent-wise analysis to check for fatalities(fig.5), injuries and an overall number of people affected. Asia remains to consistently top the charts. This can be owed to the fact that Asia is one of the most populous continents in the world.

Jumping into Continents-based Analysis, a closer look at the top 10 most recurring disasters through the three time periods(fig. 6,7) reveals that Asia and Africa are almost always the most hit continents by when it comes to Storms, Floods and Earthquakes.

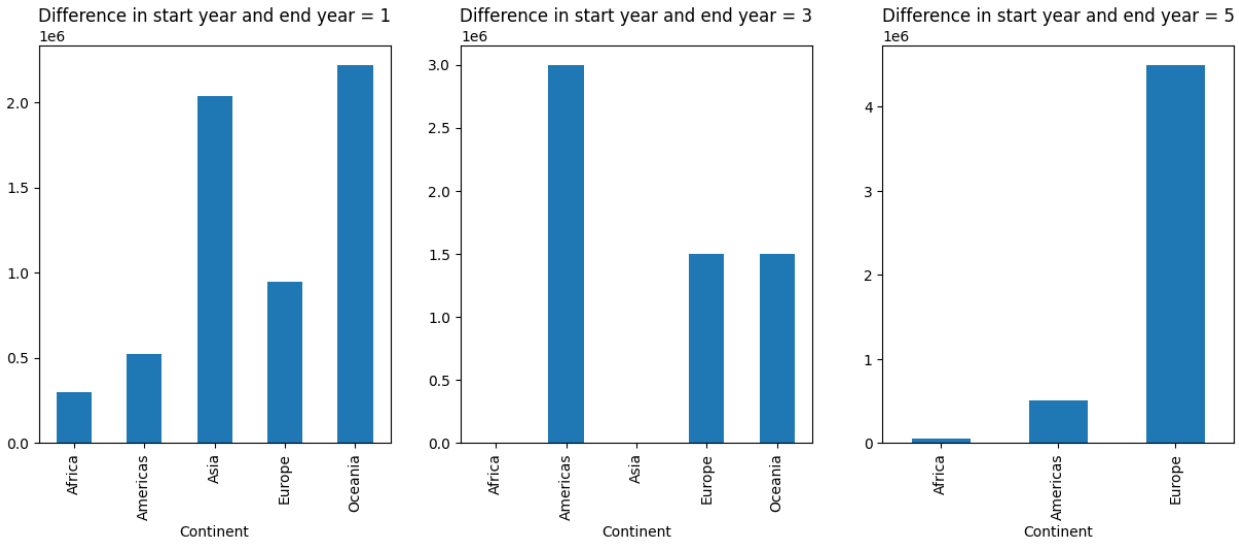


Fig(7) Top 10 disasters by Continent through the three time periods

Another interesting pattern that can be observed is when we take a close look at what kind of disasters are most experienced by each continent, the Americas seem to be closely following Asia. The lack of data on Africa caused might be the reason why it follows a different trend in the first 50 years. However, in the next two periods, it closely follows Asia and the Americas. Oceania and Europe seem to be significantly less affected than other continents.

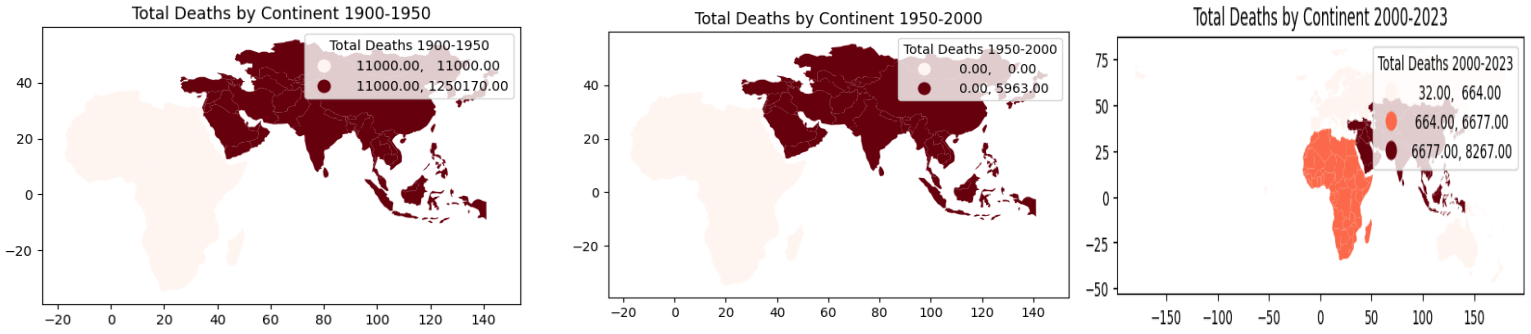
One of the assumptions I made was - The longer a disaster lasts, the higher the damages are going to cost. So I plotted three graphs which have the 3 durations - Disasters that ended within the same year, those that lasted 3 years and those that lasted 5 years. What few continents did

experience disasters for more than a year, certainly showed a higher amount of total estimated damages(fig. 8)



Fig(8) Damages in accordance with the duration of disasters

However, when we assess the total number of deaths over the years, Africa shows a significant decrease in the numbers. (fig. 9)



Fig(9) Deaths in various continents through the years

Variable Relationships

To get an idea about the correlation between the total number of people who have been affected I tried to calculate the correlation coefficient and the result was 0.13. It is very low so this might mean that the number of people who are dying is correlated to the number of people being affected. This makes complete sense because for every death there are 300 people being affected in other ways.

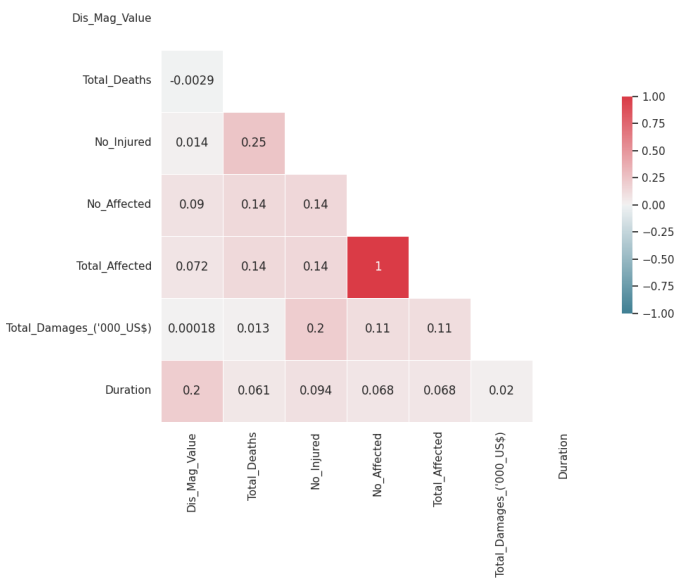


Fig: Correlation matrix showing the correlations between the 'effects' dataset

III. MAIN METHODOLOGY

One of the biggest ambitions behind this project was to build a times-series-based prediction model. For this, the data in the confederation has to span over a considerable amount of time to check for seasonality and trends.

1. *Data Collection & Preprocessing*

   First, all the columns that were filled entirely with null values were eliminated from the dataset. This included the four columns that had Reconstruction Costs, Insure Damages, and Administrative level Codes. Although some of this data might have been useful, the other part was peripheral to the aim, so we can continue.

   I tried to take a subset of the original dataframe called effects which consisted of variables that describe casualties, property damages, homelessness, injuries and people affected overall due to various disasters. Here I implemented a 95% confidence interval test to make sure the certainty of this subset of data since a big chunk of my objective was to perform a damage assessment.

   Some other data pre-processing steps included:

   a. Making sure the column names had no spaces in them to make the coding process easier.
   b. Subsetting the bigger dataset for multiple smaller datasets to perform a more efficient EDA and take a closer look at the extent of the effects of missing values.
   c. Checking both measures of central tendencies and variability

d. Performing standardization on multiple variables to compensate for extremely high variance
e. Since the occurrence of events was collected in three separate columns as Month, Day, and Year, I had to convert them to a standard format to improve their usability in terms of EDA etc.
f. Outlier detection was performed. Some outliers that are out of range even after standardization have been eliminated.
g. A new column called 'duration' was constructed out of the start and end dates of disasters to perform related EDA and other analyses

	Total_Deaths	No_Injured	No_Affected	No_Homeless	Total_Affected	Total_Damages_('000_US$)
count	11821.00	4135.00	9650.00	2466.00	12118.00	5372.00
mean	2752.76	2553.51	871002.74	72303.08	709195.79	795008.74
std	67414.81	33469.64	8403059.13	519239.98	7577177.27	5033559.29
min	1.00	1.00	1.00	3.00	1.00	2.00
25%	6.00	13.00	1250.00	555.25	658.50	9933.50
50%	19.00	50.00	10000.00	3000.00	6000.00	62550.00
75%	61.00	200.00	91941.00	17117.75	58890.50	350000.00
max	3700000.00	1800000.00	330000000.00	15850000.00	330000000.00	210000000.00

	Total_Deaths	No_Injured	No_Affected	No_Homeless	Total_Affected	Total_Damages_('000_US$)
range	3699999.00	1799999.00	329999999.00	15849997.00	329999999.00	209999998.00
variance	1488335015722.05	352218708104.59	1182168061016676 8.00	27201489320748.09	11830830273710900.00	4785030012525179.00
mode	11821.00	4135.00	9650.00	2466.00	12118.00	5372.00
iqr	25703.70	11427.91	2746466.84	182048.99	2421526.54	1845853.25

2. *Statistical Analysis*

a. **T-TEST:** A t-statistic test was performed for each continent to compare the total number of deaths in two time periods: 1950-1999 and 2000-2023. Specifically, the null hypothesis being tested is that there is no significant difference in the mean number of deaths between the two time periods within each continent. The alternative hypothesis is that there is a significant difference in the mean number of deaths between the two time periods within each continent.
b. **Correlation:** The sub-dataset 'effects' (Pg. X) which consist of the number of deaths, duration of the disaster, and damages were assessed for correlation and when a heatmap was plotted some variables turned out to have significantly higher correlation, Based on

this, I proceeded to perform Multi-variate analysis to further understand how these variables are affecting each other.

c. **Multivariate analysis using regression:** The relationship between the dependent variable "Total_Damages_('000_US$)" and the independent variables "Dis_Mag_Value", "Total_Deaths", "No_Injured", and "Duration". The model summary provides various statistics to evaluate the quality of the model fit, including R-squared, F-statistic, and p-values for each independent variable.
d. **Chi-square test:** Another hypothesis that I wanted to test was to check if there is any relationship between continents and the type of disasters that are occurring.
e. **Stationarity test:** For this project, I conducted two stationarity tests- ADF (Augmented Dickey-Fuller) test and KPSS (Kwiatkowski-Phillips-Schmidt-Shin) test is performed. The former gave unsatisfactory results. In this case, the ADF test is performed on both the 'Start_Year' and 'End_Year' columns.
f. **KPSS:** I proceeded with KPSS which by default assumes that the dataset is stationary and looks for parameters that can reject the null hypothesis.
g. **Granger Causality:** Since one of the tests showed favourable results, I proceeded further to perform a time series analysis of the data. However, before that, I wanted to check if the time in which the disasters were occurring had any causal relationship with the causalities. The output showed the results for a maximum lag of 1. The F-statistic is a measure of the fit of the regression model and the p-value tests the null hypothesis that the lagged values of "Start_Year" do not have a significant effect on the current value of "Total_Deaths".

3. Linear Regression Analysis and LSTM for forecasting and Time Series Analysis:

Linear Regression Analysis and LSTM for forecasting: I performed two machine learning tasks: regression with Lasso regularization and time series forecasting with an LSTM model.

Ultimately my aim was to conduct a model that could perform a time series analysis and forecast if the continent you live on will have any effect on when a disaster was going to occur. Two models namely SARIMAX and LSTM (Long Short Term Memory) were used in this process.

For the time series forecasting task, we first prepared the data by scaling it using the MinMaxScaler and splitting it into training and testing sets. We then created sequences of data points with a length of 60, where the input sequence was 59 data points, and the output was the next data point in the sequence. We trained an LSTM model on the training data and made predictions on the test data. Finally, we calculated the mean squared error, root mean squared error, and mean absolute error of the predictions. The output of this code was the MSE, RMSE, and MAE values calculated for the LSTM model's predictions.

IV. FINDINGS

The insights are listed section-wise in the same order as section II.

1. *Data Collection & Preprocessing*

An examination of the dataset revealed in most columns, an average of 10 per cent null values was observed, with more than half of columns having around 70 per cent missing values. Needless to say, the data was not reliable to present accurate conclusions without imputing the missing values. After the 95% confidence interval test it turned out that for all columns the confidence level was great. So this is suggestive of the fact that this subset of the data is definitely balanced and representational of the true population. For the outlier detection, it was found that since the data is quite large, there are a huge number of outliers, however, standardization also aided with this issue.

2. *Statistical Analysis*

a. **T-TEST:** It appears that there is a statistically significant difference in the means of the variable across Africa and the other regions of the world, as the p-value for Africa is below the commonly used significance level of 0.05. However, for the other regions, there is insufficient evidence to reject the null hypothesis of no difference, as the p-values are greater than 0.05.

b. **Correlation:** For correlation when a heatmap was plotted, some variables turned out to have significantly higher correlation

c. **Multivariate analysis using regression:** In this case, the R-squared value is 0.072, indicating that the model explains only 7.2% of the variance in the dependent variable. The F-statistic is 19.13 with a corresponding p-value of 3.59e-15, indicating that the overall model is statistically significant. The p-values for each independent variable show that "Total_Deaths" and "No_Injured" are significant predictors of "Total_Damages_('000_US$)" ($p < 0.05$), while "Dis_Mag_Value" and "Duration" are not significant ($p > 0.05$). Some interesting inferences that can be gathered from this test are as follows.

- For every one-unit increase in "Total_Deaths", "Total_Damages_('000_US$)" is estimated to decrease by $64,554.3.

- Conversely, the coefficient for "No_Injured" is 141.5192, indicating that for every one-unit increase in "No_Injured", "Total_Damages_('000_US$)" is estimated to increase by $141,519.2.

d. **Chi-Square Test:** The chi-square test results show a chi-square statistic of 4312.38 and a p-value of 0.0, indicating strong evidence against the null hypothesis of independence between the two variables. In other words, the test suggests that there is a significant relationship between the type of disaster and the continent on which it occurs. However, the number of missing values must be noted while considering this inference.

e. **Stationarity Test:** The ADF statistics for both columns are negative, and their corresponding p-values are very low, which suggests that we can reject the null hypothesis and conclude that both the 'Start_Year' and 'End_Year' time series are stationary. This might be because the ADF test checks whether the data can be transformed into a stationary series by testing the null hypothesis that the series has a unit root.

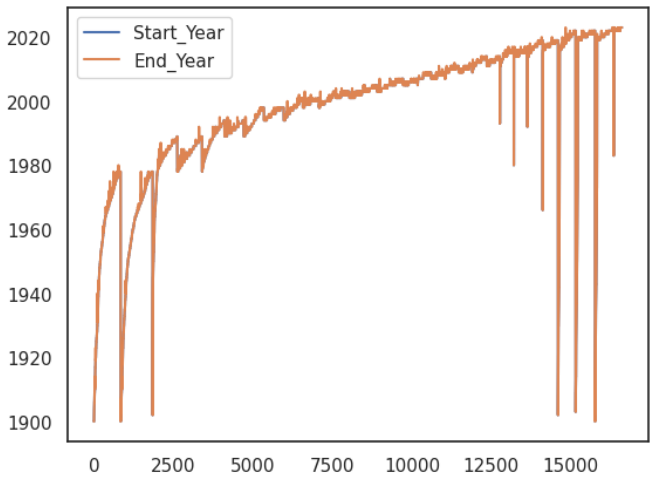


Fig: Graph of ADF stationarity test

f. **KPSS:** In this case, both the start year and end year columns have KPSS test statistics greater than their respective critical values, indicating that we can reject the null hypothesis of stationarity at the 1% level of significance. This suggests that there is a trend in the data that needs to be accounted for when modelling these time series.

g. Granger Causality: In this case, the F-statistic is 153.50 and the p-value is 0.00, indicating strong evidence against the null hypothesis and suggesting that the lagged values of "Start_Year" do have a significant effect on the current value of "Total_Deaths".

3. Linear Regression Analysis and LSTM for forecasting and Time Series Analysis:

SARIMAX Results: The output shows that the p-value for the Ljung-Box test is 0.98, which suggests that there is no significant autocorrelation in the residuals, while the p-value for the Jarque-Bera test is 0.00, which suggests that the residuals are not normally distributed. The Heteroskedasticity (H) test indicates that there is some evidence of heteroskedasticity in the residuals.

For the regression task, we used the sci-kit-learn library to perform cross-validated Lasso regression on a subset of a dataset containing information on natural disasters. We used TimeSeriesSplit as our cross-validation strategy, and we tried different values for the regularization parameter alpha and the l1_ratio. We then selected the model with the best mean cross-validation score and reported its parameters. The output for the regression task was the best Lasso model that was fitted to the data and its parameters.

The time series analysis results were not feasible, since the code ran for 5 epochs only which is not enough for the model to learn. Although more epochs might have yielded better results, the computational power was insufficient to go through 16000 rows in desired time.

4. Correspondings of findings to the big picture

a. Findings' correspondence to short-term goals:

i) To understand the usability and potential of this dataset to forecast damages to property, the extent of damage and the ability to assess and come up with feasible precautionary measures.

The usability of the data when taken as it is, is extremely limited. There must be a lot of improvement in Data collection methods to allow a broader stage for research and analysis. With more than half the data points missing in several variables, obtaining results without manipulation would be next to impossible.

That being said, damages were well assessed within certain timeframes with abundant information in the dataset. Some valuable inferences in this regard are as follows:

As the duration of disasters increases, so do the damages. One might expect that due to better-coping mechanisms over a period of time, there might be reduced damages, but that's not true.

Africa and Asia seem to be the worst-hit continents when it comes to the frequency and extent of disasters. According to the UN Population Fund, "Asia and the Pacific is the most disaster-prone region in the world. Nearly 45 per cent of the world's natural disasters occur in the region, and more than 75 per cent of those affected by natural disasters globally live in the region". Thus, our inferences are consistent with this information.

ii) To conduct various statistical tests to check and establish relationships between different attributes through correlation and causality

This was done successfully. Most variables that had a correlation with each other were assessed for casualties and statistically, most of the output obtained seemed reasonable. However, due to the extremely high number of outliers, some of the results might be skewed.

iii) To take an exploratory dive into the dataset and understand the overall importance of features.

It was relatively easy to carry out this part of the project. The data after processing makes visualizations easier to implement.

b. Findings' correspondence to long-term goals:

i) To build a pipeline that can help clean, sort and make the data more usable for various applications.

The project definitely involved a lot of pre-processing however, a definite pipeline has not been established. This is because of the nature of missing values. Most of the values could not be imputed as it would mean under-representation or over-representation of certain types of variables. The improvement in data collection is one clear trend that was observed since the more recent years saw better amounts of data.

ii) To assess the possibility to build a model that can help predict the extent of damages that a particular disaster can cause in a particular part of the world.

One crucial conclusion made from this project was the amount of damage that occurs due to a disaster is 300 times more than the number of deaths. Variables like the recovery period were entirely filled with null values. This was the reason why we had to eliminate that direction of analysis. Although some analysis with regards to overall damages was done and explored up to an extent, it did not reach its full potential due to the absence of required data.

iii) Is there a way to predict which continents are prone to what kind of disasters so that governments can build on it to implement better precautionary and disaster management measures?

Asia, Africa and America(up to some extent) seem to be having the most amount of disasters consequently requiring more endurance. Floods, Storms and Earthquakes top the charts. This might be because many areas in the Asia-Pacific region are coastal and lack tools that help prepare for such emergencies.

As they are quasi-random phenomena, the number of disasters remained constant over the years however this did not make nations stronger in any way. Continents like America and Europe have better infrastructure than most Asian countries but end up with fewer disasters annually.

V. CONCLUSIONS

Below are a few things that we clearly realize through this project:

1. Data Collection has improved considerably over time, however, most factors that help make data-backed conclusions are left with much ambiguity.
2. It is very important to disclose data manipulation while presenting results as this determines the credibility of the said results. This is the primary reason why we did not impute missing values in the dataset, but it came at a cost of incomplete and unanswered questions(such as recovery time after disasters, etc)

3. Prediction is possible and can be implemented only with better data but this still does not guarantee the fact that vulnerable nations will be better prepared. Governments have to actively take action and fund third-world nations as Natural disasters cannot be curbed.
4. Data scaling is very difficult in this case since the outliers are accounted for and cannot be easily eliminated. The values are extreme because of the unpredictability of the extent of disasters but that is what makes projects like these very valuable.
5. The biggest thing I learned from this dataset is the fact that even after 100 years, a significant amount of damage and loss of life still happens. This shows the truly unpredictable nature of Natural disasters. One can never be prepared for such an inevitable and uncontrolled phenomenon however more analysis and research in this direction amplifies preparedness among vulnerable populations to curb the consequences up to an extent.

VI. REFERENCES

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

2. McKinney, W., & others. (2010). Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference, 51-56.

3. Brownlee, J. (2017). How to Develop LSTM Models for Time Series Forecasting. Retrieved from https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/

4. Hyndman, R.J., & Athanasopoulos, G. (2018). Forecasting: principles and practice (2nd ed.). OTexts.

5. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer.

6. United Nations Office for Disaster Risk Reduction. (2015). Sendai Framework for Disaster Risk Reduction 2015-2030. United Nations.

7. International Federation of Red Cross and Red Crescent Societies. (2017). World disasters report 2017. International Federation of Red Cross and Red Crescent Societies.

8. Emergency Management Institute. (2017). Introduction to incident command system (ICS 100). Federal Emergency Management Agency.

9. Federal Emergency Management Agency. (2014). National response framework (2nd ed.). Federal Emergency Management Agency.

10. United Nations. (2015). Transforming our world: the 2030 Agenda for Sustainable Development. United Nations.

We used R programming language and Python to analyze the Brand presence of various Fast food chains across America, health inspection data provided by the County of LA Public Health for Los Angeles and Chicago. We used several R packages and libraries for data cleaning, wrangling, and visualization, including dplyr, tidyr, ggplot2, and reshape2.

Our methodology involved analyzing the total number of violations for each brand, the number of violations per store, and comparing the performance of each brand with their respective number of stores. The analysis provided insights into which brands are struggling to maintain the health code standards despite having a smaller store count.

In our conclusion, we highlighted the significance of compliance with health code standards in the fast food industry and the importance of prioritizing the health and safety of customers. Our findings can help a private equity firm to make informed decisions about potential investment opportunities in the fast food industry and support brands in improving their operations.

The presentation of our work includes a detailed report on our methodology, findings, and insights. We have also provided visualizations in the form of charts and graphs to support our analysis and highlight the key findings. The report is organized in a clear and concise manner, with headings and subheadings to improve readability and comprehension. We have also included a list of references and citations to acknowledge our sources of information and data.