Natural Disaster Analysis

Babita Chaini (x21139211)
Sasikumar Jayapal (x21153272)
Sangeeta Kumari (x21141088)
Sai Manoj Thirunahari (x21146730)
School of computing
National College of Ireland
Dublin, Ireland

Abstract—**Around the world, thousands of lives are lost each year, in addition to extensive damage to property, animals, and so on, because of natural disasters (e.g., earthquakes, floods, tsunamis, hurricanes, landslides, cloudburst, heat waves, forest fires). Our paper examines the application of data mining and analytical techniques so far designed for (i) disaster forecasting, (ii) disaster detection, and (iii) disaster management strategies developed from disaster data. An extensive description of the data that can be obtained from seismic and hydrological observatories, satellites, remote sensing, and social networking sites, such as Twitter, is provided. A comprehensive and in-depth literature review has been conducted on current techniques for disaster prediction, detection, and management, with outcomes summarized according to the different types of disasters. According to the study, India and China are among the top five countries when it comes to the number of deaths by accident.** ***Index Terms*—**
Keywords — natural disaster, timeseries, Kmeans, risk rate, death rate, forecasting, ARIMA, prediction

I. Introduction

There are some developing nations where natural disasters took place frequently and there is lack of restrictive measures to tackle the disaster harm. The damages are mainly causing in the higher developing countries where earthquakes, floods, volcanic eruptions, wild fire, tsunamis are common and these large scale events are sometimes cause massive disruption. In order to analyze and to see the effectiveness of the environmental calamities as well as the victims of disasters, the effective data has been collected from various different source systems and various formats like web scraping, CSV, API and JSON, and various machine learning models to built to analysis and forecasting natural disaster data.

In this research paper, we intent to analysis and visualize various disaster occurred from 1900 to 2021 which consist of parameters like number of disasters, disaster types and death rates using machine learning algorithms like Kmeans clustering and ARIMA Time series analysis.

we intent to predict and cluster the number of countries based on risk scores using Kmeans clustering machine learning algorithm from the data collected and analyze death tolls of various disasters occurred from 1920 to 2020.

A. Research Questions

Based on the obtained datasets, some specific research questions will be solved.

1) Dataset 1:

- How accurate is the Kmeans clustering algorithm for stratifying countries by risk score?
- Analyze and visualize the disasters that caused the most deaths by country, year and risk category.

2) Dataset 2:

- To find the states in the US with maximum disaster risk by analysing the occurrence of disasters over the course of time.
- To visualize the trend of disaster in US states from 2007 to 2018

3) Dataset 3:

- Forecasting natural disaster using time series analysis with ARIMA model.

4) Dataset 4:

- what is the reported magnitude and depth of the earthquake in different locations?
- what is the time period in which the locations have experienced an earthquake?

5) Integrated Data sets:

- Disaster management organizations should employ metrics to inform cost-benefit decisions for investment in the countries that are more prone to disasters. In order to take such decisions it is required to identify top 5 countries with maximum risk and then analyze the most prominent disasters events and their course over the period of time in these 5 counties.

II. Related Works

A lot of research in data and records of natural disasters are available online.

The book ”Natural Disaster Hotspots, A global risk analysis” written by Maxx dilley, Robert S chen.. was described about risk analysis of disaster related outcomes such as mortality and economic losses based risk levels[11].

In the article ”A review of data mining techniques to combat natural disasters”, insights were given into various fields of natural disaster phase such as prediction, detection, and disaster management strategy, as well as how machine learning models have been effective in predicting and forecasting natural disasters[1].

In the paper by Prihandoko, Bertalya published in 2016, they describe how K-means have been implemented for clustering analysis.The results of the study outline the crucial conditions associated with the occurrence of natural disasters; however, the geographic conditions are the major determinants of the problem[2].

There is a lot of research work is going on to find economical impact. Further studies can be performed on the data regarding if appropriate actions are taken by the state in the US according to the emergency plans or not. Some machine learning algorithms can also be implemented to do further analysis of the data set.

III. METHODOLOGY

There are many existing methods including CRISP-DM (cross-industry standard process for data mining), KDD(knowledge discovery in databases) and SEMMA that have already been applied to solving data mining problems.The CRISP-DM methodology was chosen to solve our problems based on our research questions. There are six stages are in CRISP-DM method which are described in the Fig.1.

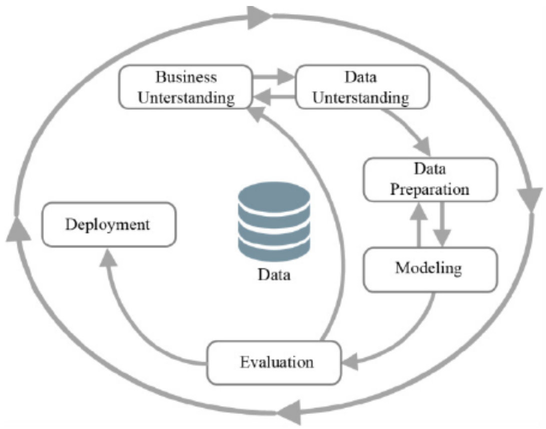


Fig. 1: CRISP-DM

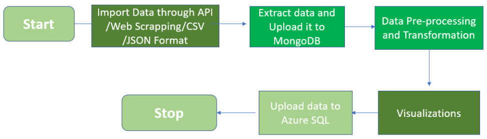

Start
Import Data through API /Web Scrapping/CSV /JSON Format
Extract data and Upload it to MongoDB
Data Pre-processing and Transformation
Stop
Upload data to Azure SQL
Visualizations

Fig. 2: Phases of Data Analysis

IV. Terminologies

A. Beautiful Soup

Beautiful Soup is a Python library primarily used for parsing HTML, XML files. It provides well defined structure for extracting information contained within HTML tags in a website.

B. Mongo DB

Mongo DB is a non-relational database system, and it is used to store semi structure or unstructured data like HTML, JSON and XML. Mongo DB is storing data in the form of collections as it is document oriented.

C. Microsoft Azure SQL Database

Microsoft Azure SQL Database is a managed cloud database (PaaS) that is delivered as part of Microsoft Azure.The cloud database is a database that runs on a cloud computing platform and is accessible as a service.

D. Plotly

Plotly is a python's graphic library makes interactive graphs such as line plots, scatter plots, area charts, bar charts, box plots, histograms and bubble charts.

V. Data Selection and Preprocessing

For our study, we have collected data from four different data sources in four different formats such as web, API, JSON and CSV. The programming language python is used to extract, clean and load into Microsoft azure sql server.

A. Design Architecture

The Fig.3 describes the entire flow of the data processing and loading stages. It gives high level structure of natural disaster data flow in this research.

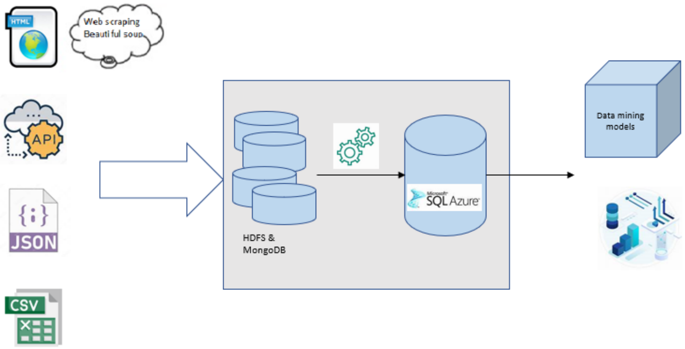


Fig. 3: Project Design Architecture

B. Datasets

1) Web scraping Dataset: The data collected via web scrapping is termed as Dataset1. As part of this analysis there are two datasets were scrapped from the web. First piece of data was scrapped off natural disaster vulnerable countries list based on World Risk Index, calculated by the United Nations University Institute for Environment and Human Security, this list has around 171 countries. The other dataset contains details about death tolls by various disaster events from the year 1992 to 2020.

- Source 1: https://en.wikipedia.org/wiki/List_of_countries_by_natural_disaster_risk
- Source 2: https://en.wikipedia.org/wiki/List_of_natural_disasters_by_death_toll

- Source 3: https://www.kaggle.com/da andradaolteanu/country-mapping-iso-continent-reg

2) API Dataset: FEMA disaster summary dataset tracted using API which is freely available in FEM website. All details are specific to US states and for range from 2007 to 2018. Data Description: Data cc 1000 rows and 16 columns.

- Source: https://www.fema.gov/api/op FemaWebDisasterDeclarations

3) CSV Dataset: The EM-DAT dataset contains the d data from the years 1900 to 2021. We chose two dataset one dataset contains the data from 1900 to 2021 with records and the other dataset contains the data from 1! 2021 with 14644 records.

- Source 1: https://www.kaggle.com/datasets/brsc all-natural-disasters-19002021-eosdis

4) CSV Dataset: The National Earthquake Inforr Centre publishes a dataset containing information on v aspects of earthquakes, such as the date, time, depth, r tude, and source of all earthquakes with a reported magnitude of 5.5 or higher since 1965 to 2021.

- https://www.usgs.gov/programs/earthquake-hazards/data

C. Database Schema

The database schema diagram, also known as an entity-relationship diagram, depicts the relationships between the entities/tables contained in the database. The constructed tables have a ”many to one” connection. Figure 4 depicts a database schema for this research subject. There are following six tables are created in Microsoft Azure sql database to store all our pre-processed data.

- dbo.Risk_Country
- dbo.Death_Toll_by_Disaster
- dbo.country
- US_Disaster_Events
- Natural_Disaster_Events
- Earthquake
- Master_Disaster_Events

D. Data Cleaning and Transformation

Data selection and preprocessing is the very crucial step in every data mining problem. It includes data cleaning, preprocessing and feature selection process using python libraries such as pandas, numpy, pyodbc, sklearn, matplotlib and plotly .

1) Dataset 1 - Death rates/Vulnerable Countries by risk rates: This involves loading the necessary packages into a Python environment and scraping data from the web and loading it into MongoDB.Data retrieved from MongoDB will be loaded into a Python pandas dataframe, and the following data pre-processing operation will be performed.

- Pandas data frame was analyzed to identify and remove unnecessary rows, columns and rename every columns according to a proper naming convention.

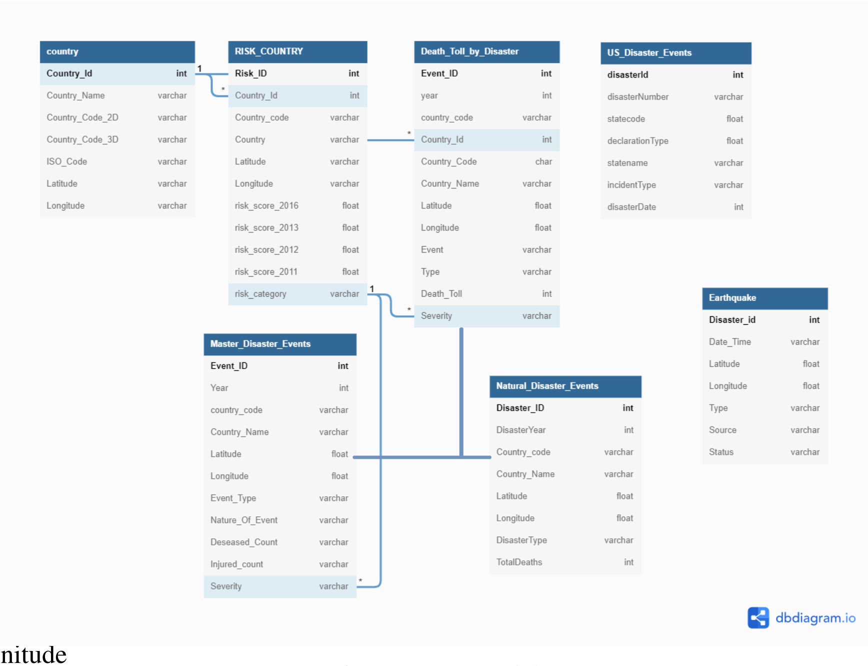


Fig. 4: Data Model

- Removed unwanted characters such as ’%’ from the risk_score attributes and converted them into float data times as the machine learning model accepts only numerical values.
- A new column Risk_Score was derived from the previous risk scores.
- Our data was formatted to comply with the first normal form.Then, we merged with country dataset to get a few attributes such as country code, longitude, and latitude.
- All the above data cleaning and pre-processing steps will be applied on death toll dataset as well.
- Finally, pre-processed data will be loaded into Azure Sql server table dbo.country,dbo.RISK_COUNTRY, dbo.Death_Toll_by_Disaster for the further analysis.

2) Dataset 2 - API Dataset:

- First dataset is checked for null values but there was no Null Values.
- Dates in the date-related fields were not in proper format. For analysis of data only year is required so for all date-related field year is fetched using the to_datetime function in pandas
- Some of the columns were also renamed so as to have some meaningful insights while carrying out analysis.
- Some of the not useful columns were removed that were not required for further analysis.
- All the texts inside the data frame were converted to a similar format for clear visualization.

The cleansed and preprocessed data will get loaded into dbo.US_Disaster_Events.

3) Dataset 3 - Time series analysis using ARIMA model:

- The first step of the analysis is to import the required libraries into the python environment and load the dataset

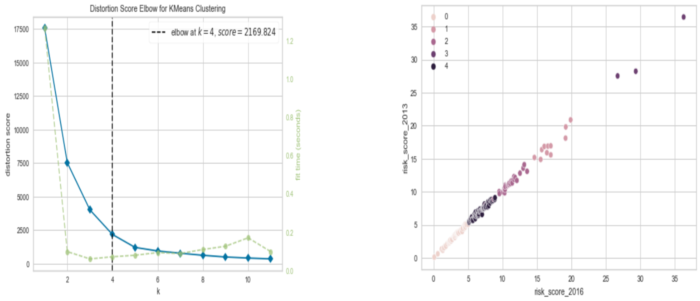


Fig. 5: Kmeans cluster

into a Pandas data frame.

- The type conversion function converts the original column datatypes into best possible datatypes for the respective columns.
- The dataset has contains some missing values. This treatment was done in four ways – remove the rows containing missing values, remove the columns containing missing values, imputing the missing values in a numerical column with its mean or median and imputing the missing values in a categorical column with its mode.
- The next step is to transform and scale the features in the dataset.
- Finally, the preprocessed data will get loaded into Natural_Disaster_Events table.

4) Dataset 4 - Earthquake Dataset:

- We find some missing values in the dataset. We have parsed the date column into date time format.
- Apart from earthquake, we find there is information related to other disasters like nuclear explosion, rock burst and explosion. So, eliminating these records will make the analysis easier and more precise.
- Then for missing value analysis, we will drop the records, where more than 50% of the values are not available.
- The root mean square column also has around one third of the values missing in it. So, we have to assign something to proceed with the analysis further
- A total of 447 values have been found to be outliers when earthquakes of higher magnitude have occurred.
- The processed data will be loaded into dbo.Earthquake table.

VI. Model Building

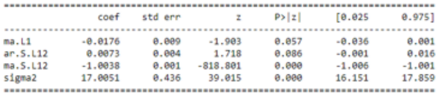
A. Kmeans clustering on Dataset 1

Kmeans clustering is an unsupervised classification algorithm. The Kmeans algorithm clusters countries into 5 different groups, with labels like 'Very Low Risk, Low Risk, Medium Risk, High Risk and Very High Risk. Fig:5 The result indicates that kmeans clustering is not optimal for classifying countries according to risk scores.The dataset is imbalanced and the volume of the data is low.

B. Time series analysis using ARIMA model on Dataset 3

ARIMA stands for autoregressive integrated moving average model and is specified by three order parameters: (p, d, q). There are different types of ARIMA model which are stated below.

- ARIMA – Non seasonal Autoregressive Integrated Moving Averages.
- SARIMA – Seasonal ARIMA.
- SARIMAX – Seasonal ARIMA with exogenous variables.
- Fig.6, Statistical summary of ARIMA model:



```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.0176      0.009     -1.903      0.057      -0.036       0.001
ar.S.L12       0.0073      0.004      1.718      0.086      -0.001       0.016
ma.S.L12      -1.0038      0.001   -818.801      0.000      -1.006      -1.001
sigma2        17.0051      0.436     39.015      0.000      16.151      17.859
==============================================================================
```

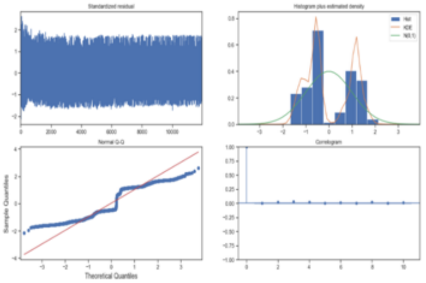


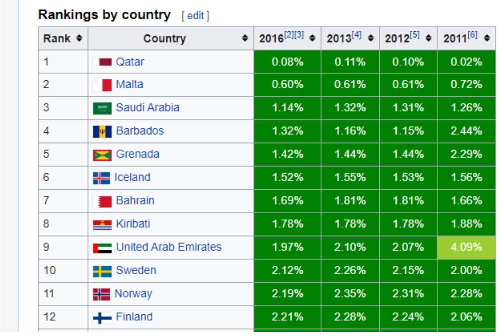
Fig. 6: ARIMA model

VII. Results and Evaluations

Data analysis on the natural disaster dataset has been carried out for the various data sources and presented interactive visualizations below.

A. Dataset 1:Death rates/Vulnerable Countries by risk rates

1) : In Fig.7, the web scraping dataset 1 shows risk score rankings for all the countries.

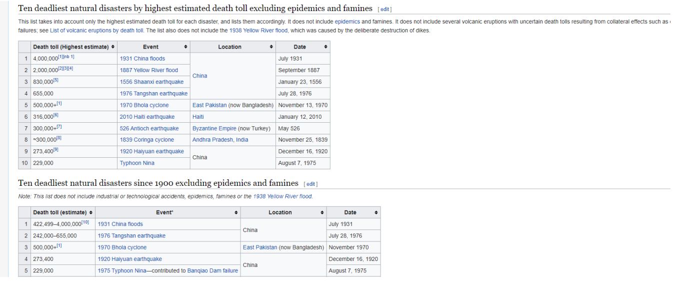


Rankings by country [edit]

Rank ⇕	Country ⇕	2016[2][3] ⇕	2013[4] ⇕	2012[5] ⇕	2011[6] ⇕
1	Qatar	0.08%	0.11%	0.10%	0.02%
2	Malta	0.60%	0.61%	0.61%	0.72%
3	Saudi Arabia	1.14%	1.32%	1.31%	1.26%
4	Barbados	1.32%	1.16%	1.15%	2.44%
5	Grenada	1.42%	1.44%	1.44%	2.29%
6	Iceland	1.52%	1.55%	1.53%	1.56%
7	Bahrain	1.69%	1.81%	1.81%	1.66%
8	Kiribati	1.78%	1.78%	1.78%	1.88%
9	United Arab Emirates	1.97%	2.10%	2.07%	4.09%
10	Sweden	2.12%	2.26%	2.15%	2.00%
11	Norway	2.19%	2.35%	2.31%	2.28%
12	Finland	2.21%	2.28%	2.24%	2.06%

Fig. 7: Web scraping Dataset1

2) : In Fig 8, the web scraping dataset 2 shows death tolls of countries based on each disaster types.



Ten deadliest natural disasters by highest estimated death toll excluding epidemics and famines [edit]

This list takes into account only the highest estimated death toll for each disaster, and lists them accordingly. It does not include epidemics and famines. It does not include several volcanic eruptions with uncertain death tolls resulting from collateral effects such as
failures; see List of volcanic eruptions by death toll. The list also does not include the 1938 Yellow River flood, which was caused by the deliberate destruction of dikes.

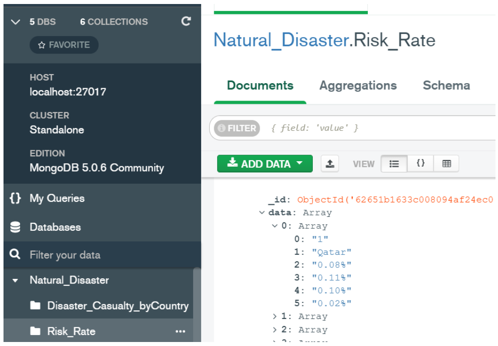
	Death toll (Highest estimate)	Event	Location	Date
1	4,000,000[1][nb 1]	1931 China floods	China	July 1931
2	2,000,000[2][3][4]	1887 Yellow River flood		September 1887
3	830,000[5]	1556 Shaanxi earthquake		January 23, 1556
4	655,000	1976 Tangshan earthquake		July 28, 1976
5	500,000+[1]	1970 Bhola cyclone	East Pakistan (now Bangladesh)	November 13, 1970
6	316,000[6]	2010 Haiti earthquake	Haiti	January 12, 2010
7	300,000+[7]	526 Antioch earthquake	Byzantine Empire (now Turkey)	May 526
8	≈300,000[8]	1839 Coringa cyclone	Andhra Pradesh, India	November 25, 1839
9	273,400[9]	1920 Haiyuan earthquake	China	December 16, 1920
10	229,000	Typhoon Nina		August 7, 1975

Ten deadliest natural disasters since 1900 excluding epidemics and famines [edit]

Note: This list does not include industrial or technological accidents, epidemics, famines or the 1938 Yellow River flood.

	Death toll (estimate)	Event*	Location	Date
1	422,499–4,000,000[10]	1931 China floods	China	July 1931
2	242,000–655,000	1976 Tangshan earthquake		July 28, 1976
3	500,000+[1]	1970 Bhola cyclone	East Pakistan (now Bangladesh)	November 1970
4	273,400	1920 Haiyuan earthquake	China	December 16, 1920
5	229,000	1975 Typhoon Nina—contributed to Banqiao Dam failure		August 7, 1975

Fig. 8: Web scraping Dataset2



```
_id: ObjectId('62651b1633c008094af24ec0
data: Array
  0: Array
    0: "1"
    1: "Qatar"
    2: "0.08%"
    3: "0.11%"
    4: "0.10%"
    5: "0.02%"
  1: Array
  2: Array
```

Fig. 9: MongoDB Connection

3) : The Fig. 9 and Fig.10 shows the MongoDB and Azure sql database connectivity.

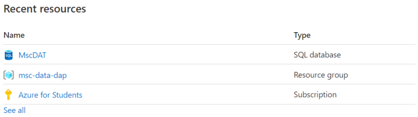
4) : The Fig.11 and Fig.12 depict the tables that were loaded into the Azure SQL database with pre-processed data.

5) : The following Fig .13 represents number of countries in each risk category defined by the world risk index.

6) : The geographical map shown below in Fig.14 is visualizing countries by each risk category.

7) : The Fig.16 and Fig.17 shows the death counts of each natural disaster category from 1920 to 2020.

8) : According to the Fig. 18 and Fig.19, China and India are the two countries that are hardly affected by natural disaster.The death toll is also very high in those countries.

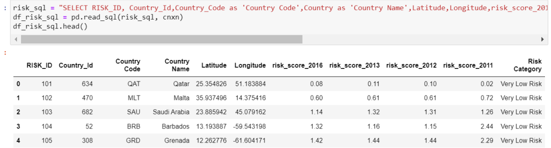


Recent resources

Name	Type
MscDAT	SQL database
msc-data-dap	Resource group
Azure for Students	Subscription

See all

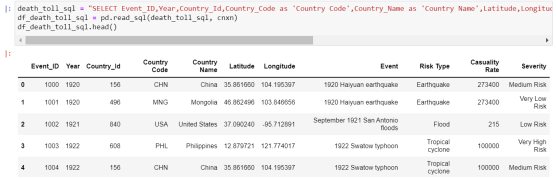
Fig. 10: Azure sql Connection



```
risk_sql = "SELECT RISK_ID, Country_Id,Country_Code as 'Country Code',Country as 'Country Name',Latitude,Longitude,risk_score_201
df_risk_sql = pd.read_sql(risk_sql, cnxn)
df_risk_sql.head()
```

	RISK_ID	Country_Id	Country Code	Country Name	Latitude	Longitude	risk_score_2016	risk_score_2013	risk_score_2012	risk_score_2011	Risk Category
0	101	634	QAT	Qatar	25.354826	51.183884	0.08	0.11	0.10	0.02	Very Low Risk
1	102	470	MLT	Malta	35.937496	14.375416	0.60	0.61	0.61	0.72	Very Low Risk
2	103	682	SAU	Saudi Arabia	23.885942	45.079162	1.14	1.32	1.31	1.26	Very Low Risk
3	104	52	BRB	Barbados	13.193887	-59.543198	1.32	1.16	1.15	2.44	Very Low Risk
4	105	308	GRD	Grenada	12.262776	-61.604171	1.42	1.44	1.44	2.29	Very Low Risk

Fig. 11: Risk Country Table



```
death_toll_sql = "SELECT Event_ID,Year,Country_Id,Country_Code as 'Country Code',Country_Name as 'Country Name',Latitude,Longituc
df_death_toll_sql = pd.read_sql(death_toll_sql, cnxn)
df_death_toll_sql.head()
```

	Event_ID	Year	Country_Id	Country Code	Country Name	Latitude	Longitude	Event	Risk Type	Casuality Rate	Severity
0	1000	1920	156	CHN	China	35.861660	104.195397	1920 Haiyuan earthquake	Earthquake	273400	Medium Risk
1	1001	1920	496	MNG	Mongolia	46.862496	103.846656	1920 Haiyuan earthquake	Earthquake	273400	Very Low Risk
2	1002	1921	840	USA	United States	37.090240	-95.712891	September 1921 San Antonio floods	Flood	215	Low Risk
3	1003	1922	608	PHL	Philippines	12.879721	121.774017	1922 Swatow typhoon	Tropical cyclone	100000	Very High Risk
4	1004	1922	156	CHN	China	35.861660	104.195397	1922 Swatow typhoon	Tropical cyclone	100000	Medium Risk

Fig. 12: Casualty Rate Table

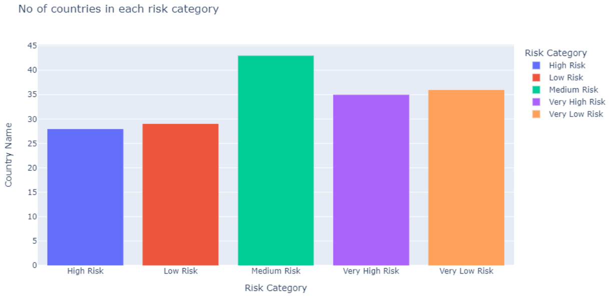


Fig. 13: Countries by risk

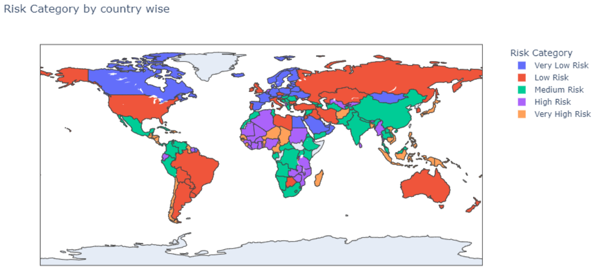


Fig. 14: Geological map by risk

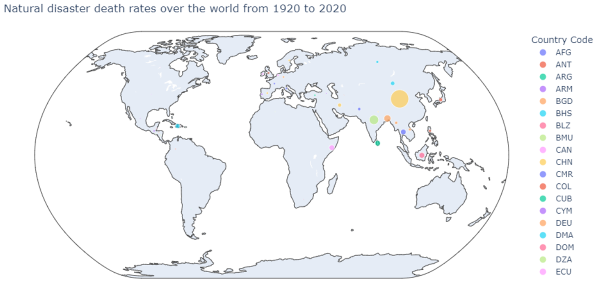


Fig. 15: World map by death counts

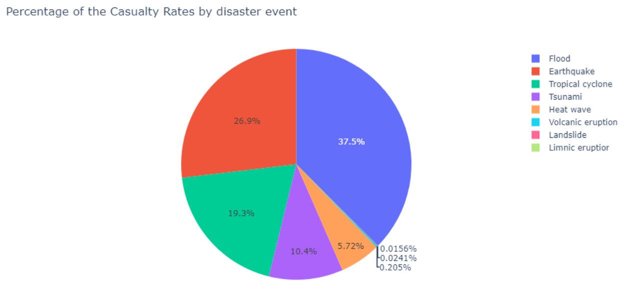


Fig. 16: Casualty rate by disaster category

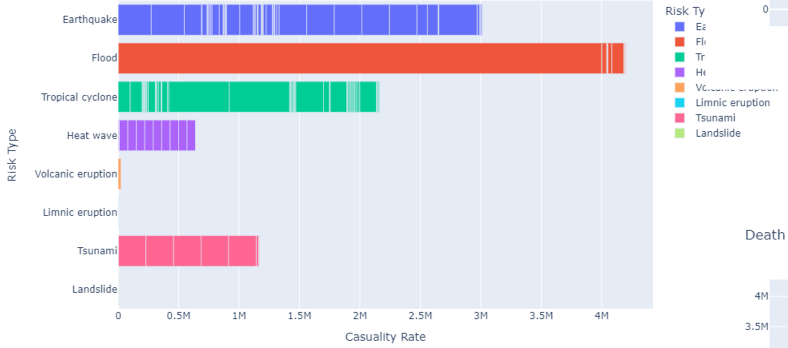


Fig. 17: Death counts by each disaster category

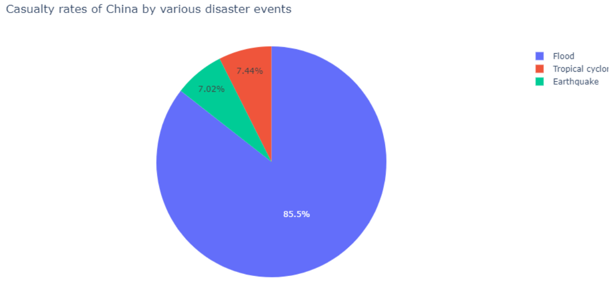


Fig. 18: Casualty rate in China

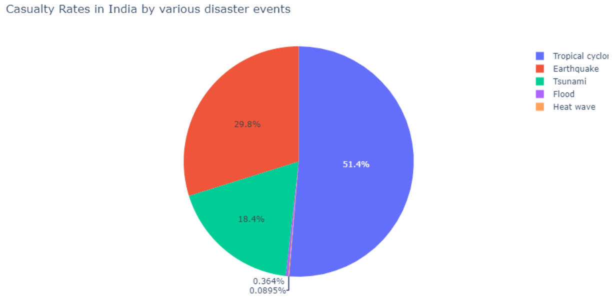


Fig. 19: Casualty rate in India

9) : As per the Fig.20 and Fig.21, A new report explains the trends of death counts by country and risk type, and confirms that China suffered from a deadly flood in 1931, which resulted in more than 4 million deaths.

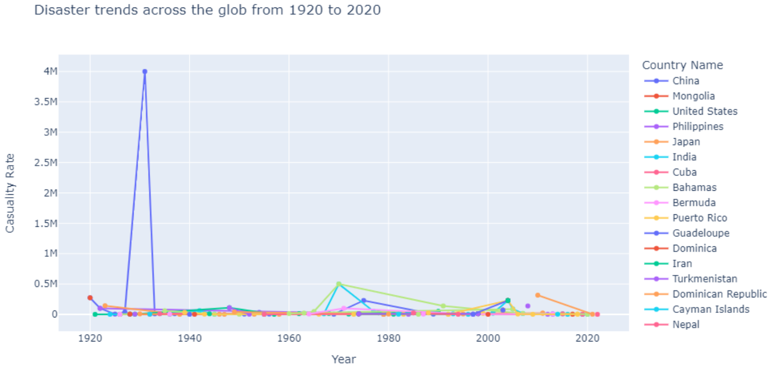


Fig. 20: Death count by Country from 1920 to 2020

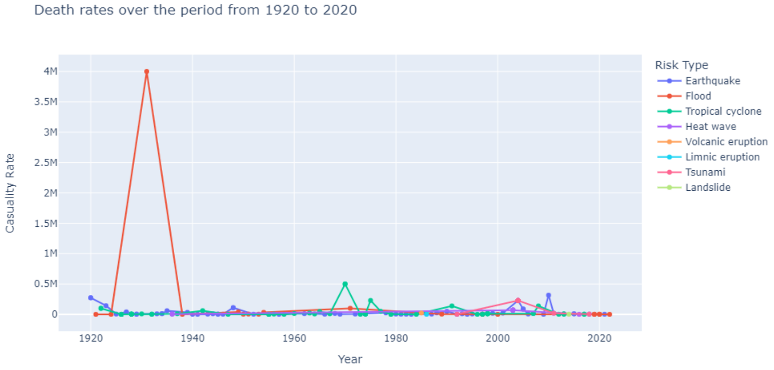


Fig. 21: Death count by disaster from 1920 to 2020

10) : Due to natural disasters, China has suffered a large number of casualties than other countries as shown in Fig.22

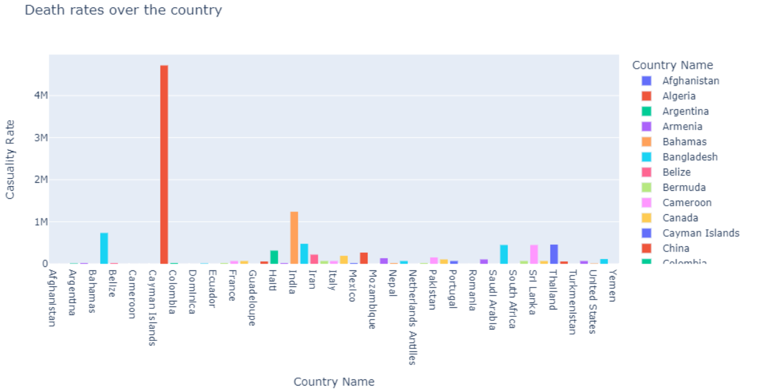


Fig. 22: Death counts by country

B. Dataset 2:US natural disaster

1) : Figure 23 shows Number of Disasters happened in U states from 2007 to 2018

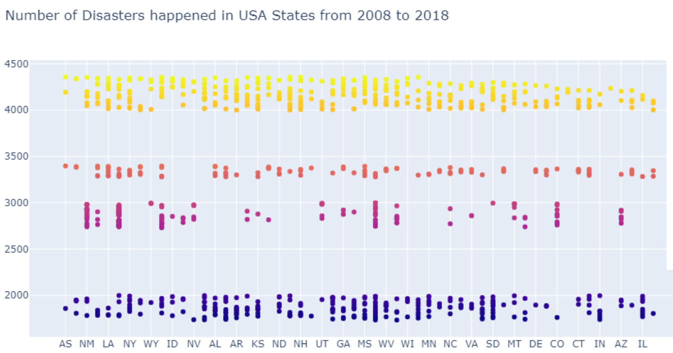


Fig. 23: Disaster numbers from 2007- 2018

2) : Figure 24 shows different Declaration Types for disasters according to state laws.

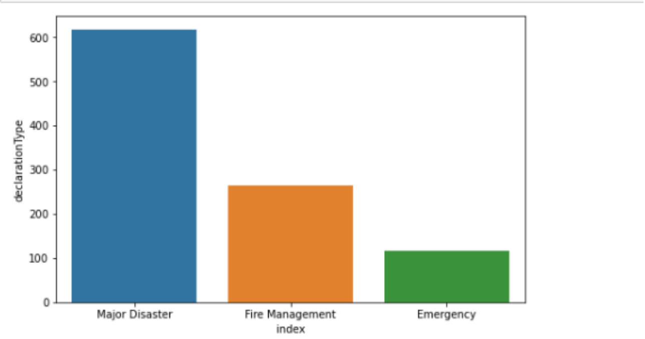


Fig. 24: Declaration Types

3) : Figure 25 shows the heatmap of all the disasters tl has happened in respective states over the period of time. gives a clear picture of the disaster happening in US states

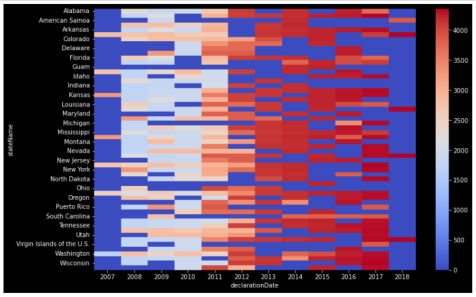


Fig. 25: Number of Disasters from 2007- 2018

4) : Figure 26 and 27 shows the top ten states in US where number of Disasters are more over the period of time. Texas tops the list. Data shows that Texas has reported maximum disasters in US. So exploring more about Texas.

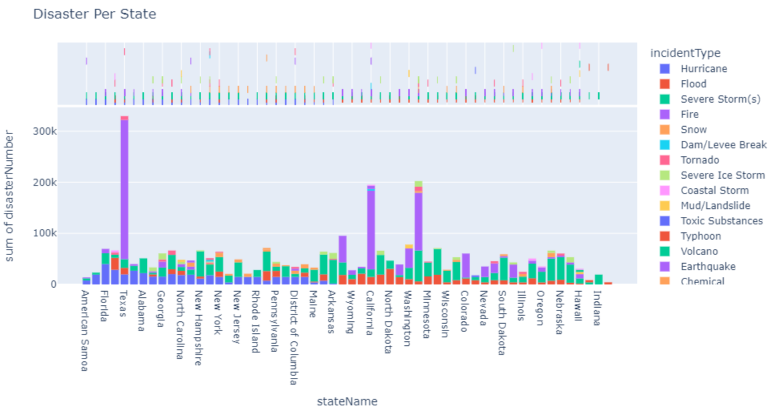


Fig. 26: Disaster per State

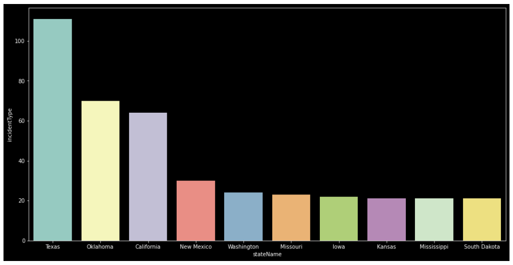


Fig. 27: Top ten states in US prone to disasters

5) : Figure 28 shows that Hurricane is the most prominent disaster that Texas has reported.

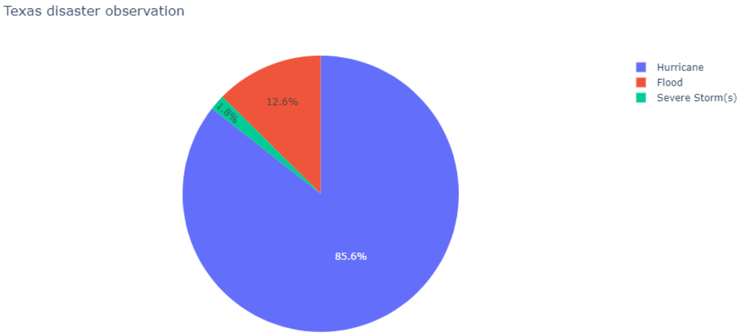


Fig. 28: Texas disaster observation

6) : Figure 29 shows that Texas was hit by maximum number of disasters in 2011 followed by 2008.

7) : Figure 30 shows that over the course of time disasters number has been increasing in the US.

8) : Finally after all visualization and analysis is done data is stored into Azure SQL for merging with other datasets to have a final observation. Below figure 31 shows the data fetched from Aure SQL.

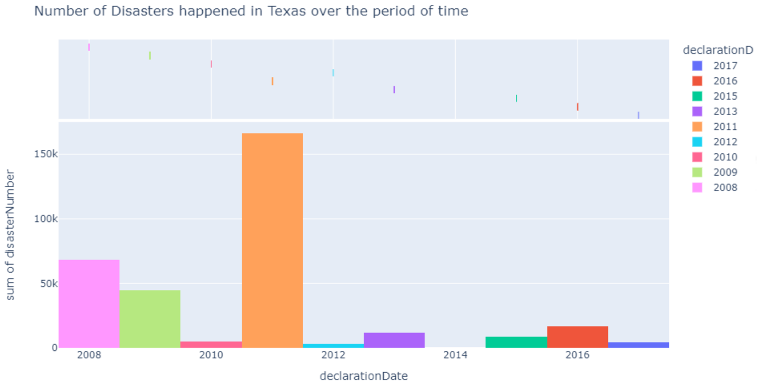


Fig. 29: Disasters happened in Texas

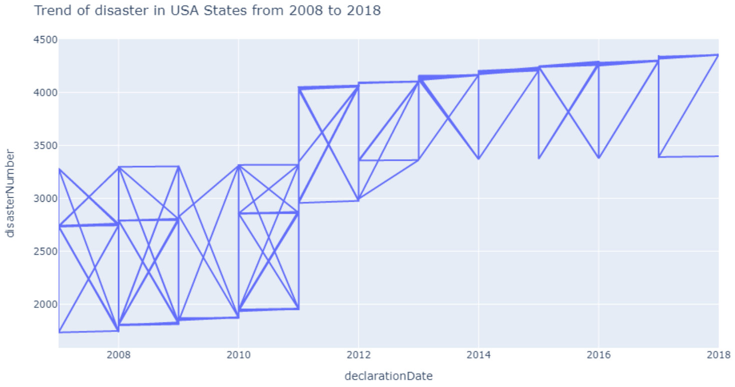


Fig. 30: Disaster Trend

C. Time series analysis on Dataset 3

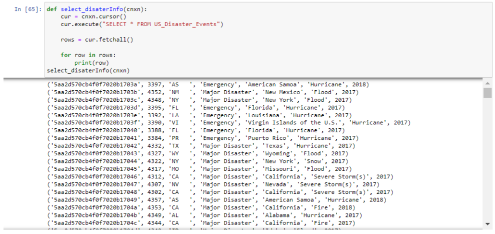
The below graphs gives us some useful insights about the natural disasters

D. Dataset 4: Earthquake

The following images describes the analysis of earthquake data.

E. Analysis and Visualizations on Integrated Datasets

Obtained datasets are cleaned, pre-processed and integrated in the Microsoft Azure Sql Server.Analysis have been made on Master_Disaster_Event table and visualizations are presented below.



```
In [65]: def select_disaterInfo(cnxn):
             cur = cnxn.cursor()
             cur.execute("SELECT * FROM US_Disaster_Events")

             rows = cur.fetchall()

             for row in rows:
                 print(row)
         select_disaterInfo(cnxn)
```

```
('5aa2d570cb4f0f7020b1703a', 3397, 'AS   ', 'Emergency', 'American Samoa', 'Hurricane', 2018)
('5aa2d570cb4f0f7020b1703b', 4352, 'NM   ', 'Major Disaster', 'New Mexico', 'Flood', 2017)
('5aa2d570cb4f0f7020b1703c', 4348, 'NY   ', 'Major Disaster', 'New York', 'Flood', 2017)
('5aa2d570cb4f0f7020b1703d', 3395, 'FL   ', 'Emergency', 'Florida', 'Hurricane', 2017)
('5aa2d570cb4f0f7020b1703e', 3392, 'LA   ', 'Emergency', 'Louisiana', 'Hurricane', 2017)
('5aa2d570cb4f0f7020b1703f', 3390, 'VI   ', 'Emergency', 'Virgin Islands of the U.S.', 'Hurricane', 2017)
('5aa2d570cb4f0f7020b17040', 3388, 'FL   ', 'Emergency', 'Florida', 'Hurricane', 2017)
('5aa2d570cb4f0f7020b17041', 3384, 'PR   ', 'Emergency', 'Puerto Rico', 'Hurricane', 2017)
('5aa2d570cb4f0f7020b17042', 4332, 'TX   ', 'Major Disaster', 'Texas', 'Hurricane', 2017)
('5aa2d570cb4f0f7020b17043', 4327, 'WY   ', 'Major Disaster', 'Wyoming', 'Flood', 2017)
('5aa2d570cb4f0f7020b17044', 4322, 'NY   ', 'Major Disaster', 'New York', 'Snow', 2017)
('5aa2d570cb4f0f7020b17045', 4317, 'MO   ', 'Major Disaster', 'Missouri', 'Flood', 2017)
('5aa2d570cb4f0f7020b17046', 4312, 'CA   ', 'Major Disaster', 'California', 'Severe Storm(s)', 2017)
('5aa2d570cb4f0f7020b17047', 4307, 'NV   ', 'Major Disaster', 'Nevada', 'Severe Storm(s)', 2017)
('5aa2d570cb4f0f7020b17048', 4302, 'CA   ', 'Major Disaster', 'California', 'Severe Storm(s)', 2017)
('5aa2d570cb4f0f7020b17049', 4357, 'AS   ', 'Major Disaster', 'American Samoa', 'Hurricane', 2018)
('5aa2d570cb4f0f7020b1704a', 4353, 'CA   ', 'Major Disaster', 'California', 'Fire', 2018)
('5aa2d570cb4f0f7020b1704b', 4349, 'AL   ', 'Major Disaster', 'Alabama', 'Hurricane', 2017)
('5aa2d570cb4f0f7020b1704c', 4344, 'CA   ', 'Major Disaster', 'California', 'Fire', 2017)
```

Fig. 31: US_Disaster_Events table

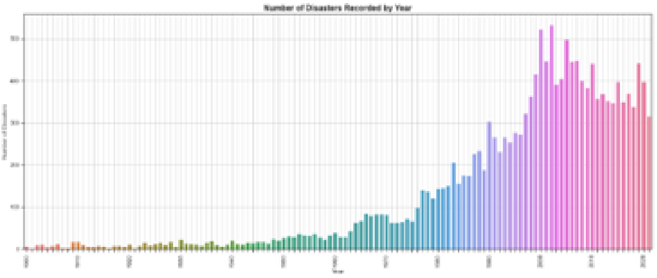


Fig. 32: Number of Disasters recorded per Year

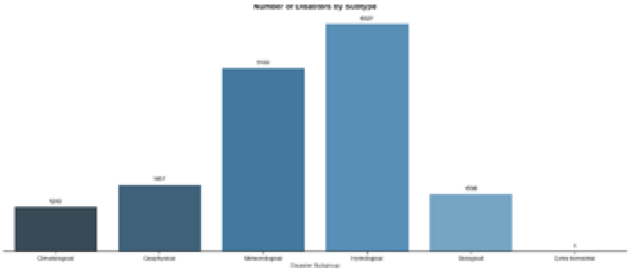


Fig. 33: Number of Disasters by Subtype

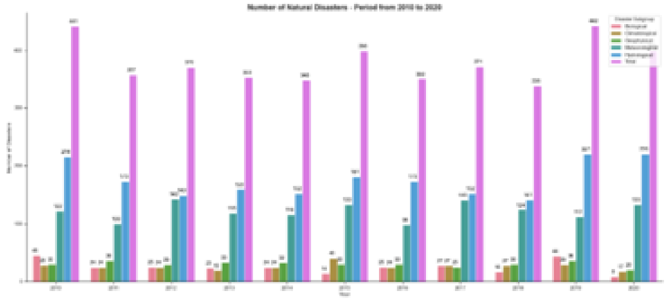


Fig. 34: Number of Disasters from 1900 – 2020

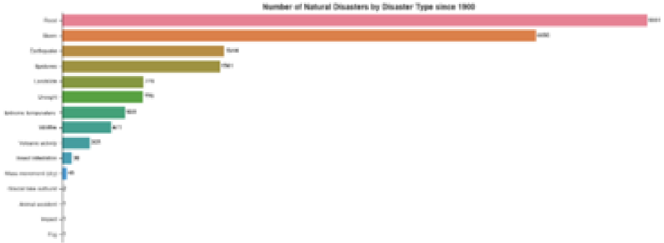


Fig. 35: Number of Disasters by Type

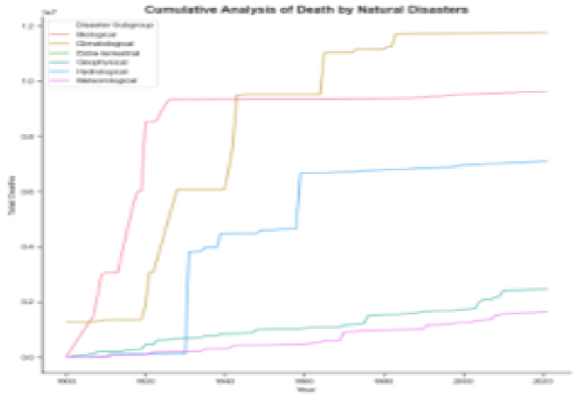


Fig. 36: Cumulative Analysis of Death by Natural Disasters

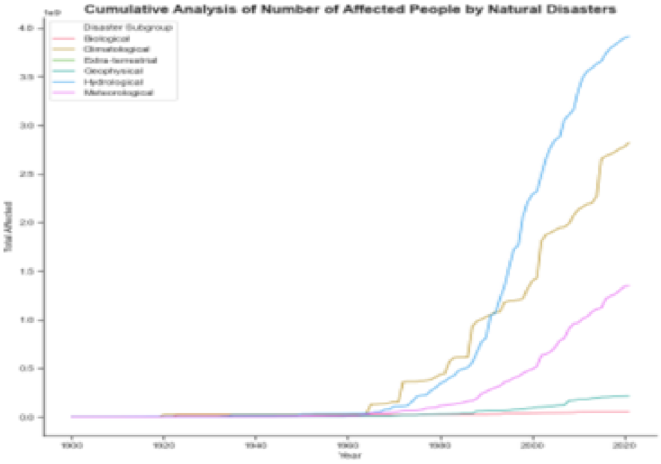


Fig. 37: Cumulative Analysis of Number of Affected People by Natural Disasters

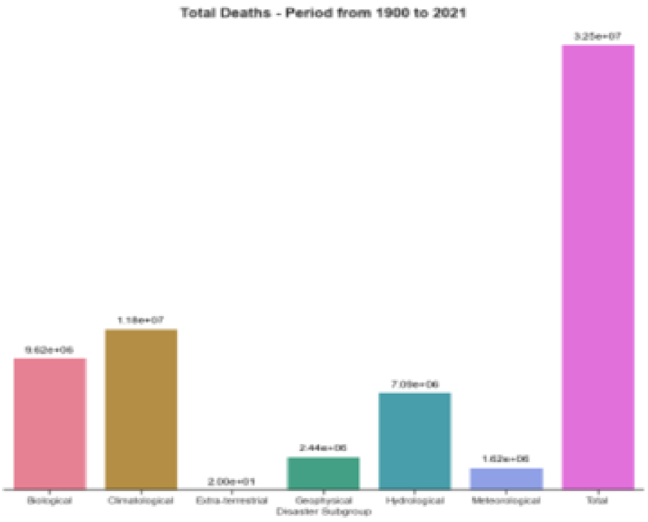


Fig. 38: Total Death

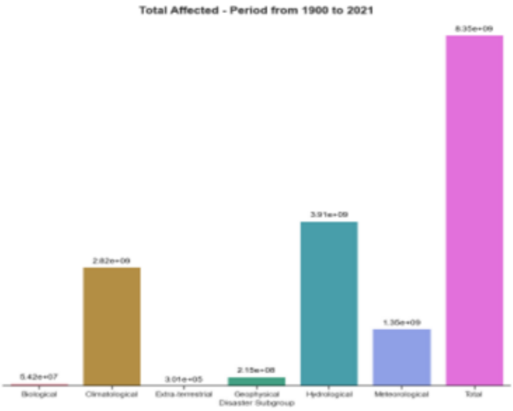


Fig. 39: Total Affected

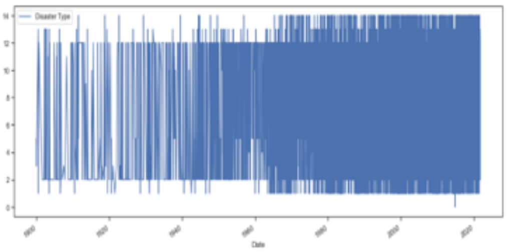


Fig. 40: Visualizing Time Series data

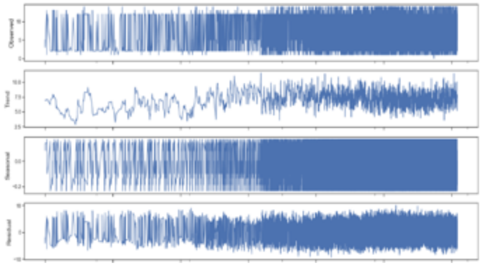


Fig. 41: Seasonal Decomposition of data

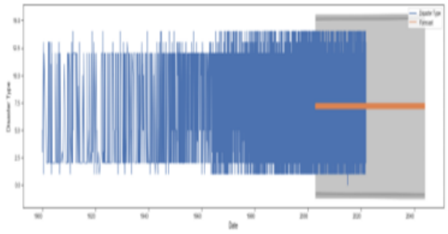


Fig. 42: Visualizing forecasts for future years

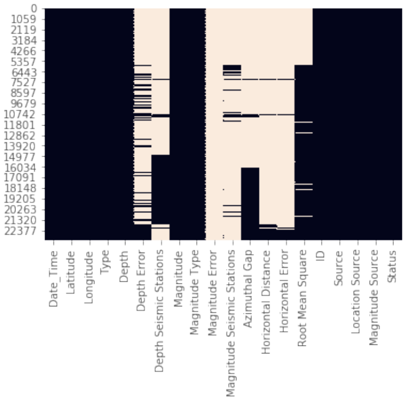


Fig. 43: Missing value analysis

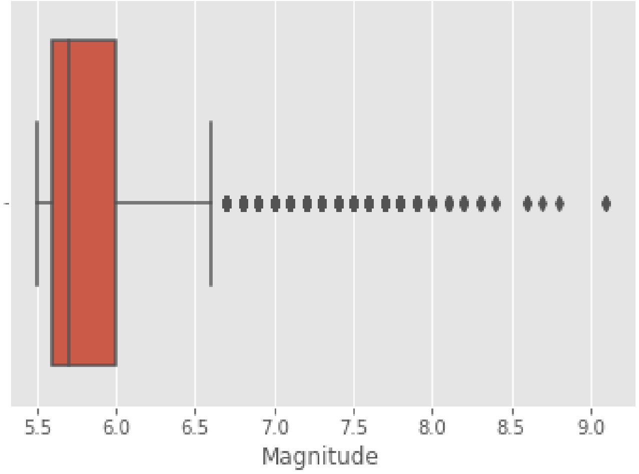


Fig. 44: Outlier analysis

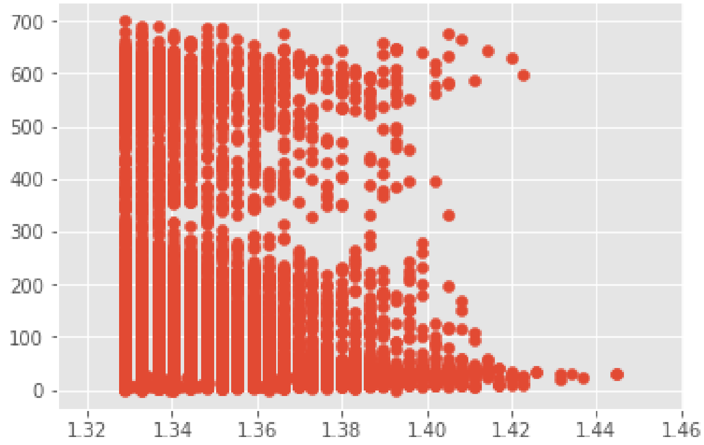


Fig. 45: Correlation

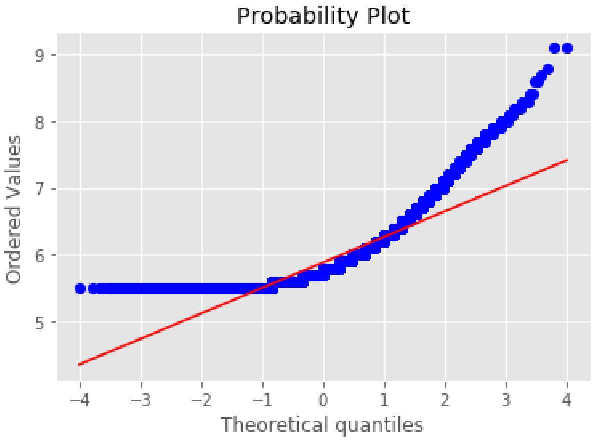


Fig. 46: Probability Plot

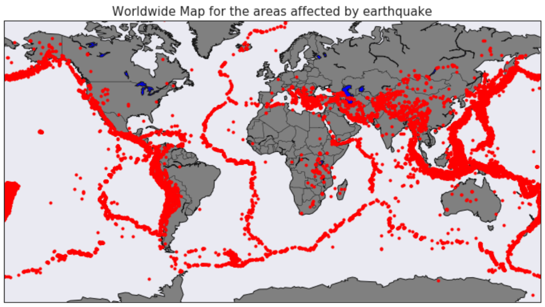
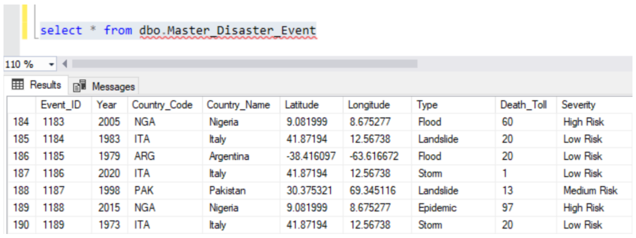


Fig. 47: Probability Plot

1) : As shown in the Fig. 48, Integrated data sets are loaded into Master_Disaster_Event table.



```
select * from dbo.Master_Disaster_Event
```

	Event_ID	Year	Country_Code	Country_Name	Latitude	Longitude	Type	Death_Toll	Severity
184	1183	2005	NGA	Nigeria	9.081999	8.675277	Flood	60	High Risk
185	1184	1983	ITA	Italy	41.87194	12.56738	Landslide	20	Low Risk
186	1185	1979	ARG	Argentina	-38.416097	-63.616672	Flood	20	Low Risk
187	1186	2020	ITA	Italy	41.87194	12.56738	Storm	1	Low Risk
188	1187	1998	PAK	Pakistan	30.375321	69.345116	Landslide	13	Medium Risk
189	1188	2015	NGA	Nigeria	9.081999	8.675277	Epidemic	97	High Risk
190	1189	1973	ITA	Italy	41.87194	12.56738	Storm	20	Low Risk

Fig. 48: Master_Disaster_Event

2) : A Fig.49 shows, In 1931, there is around 4 millions death reported worldwide so digging more about disasters that happened in 1931.

3) : Figure. 50 shows that China's Flood , China's Earthquake and volcanic activity in Indonesia has caused huge loss of life in 1931

4) : Fig. 51 shows that China Followed by United States , India , Indonesia and Philippines are the top 5 countries effected by disasters.

5) : The Fig 52 and 53 shows that China has mostly effected by Storm. 1931 is the worst year in China from disaster perspective when lot of deaths were reported due to flood.

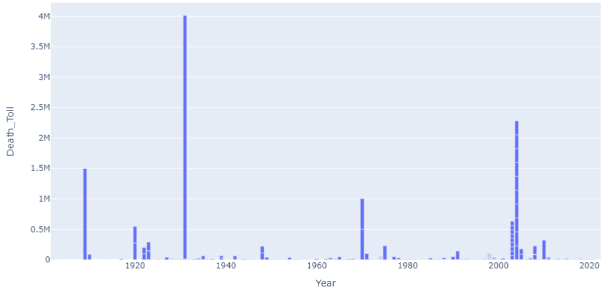


Fig. 49: Disaster by Year

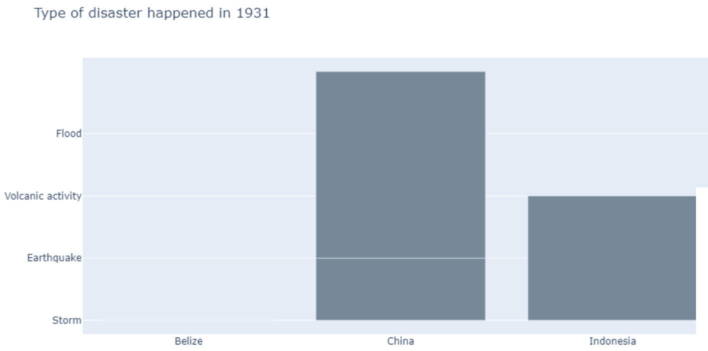


Fig. 50: Disasters in 1931

6) : Fig 54 and 55 shows that United States has mostly be effected by Storm. 1936 is the worst year in United States fro disaster perspective when lot of deaths were reported due heat wave.

7) : Fig 56 and 57, shows that India has mostly been affected by Flood. 1970 is the worst year in India from disaster perspective when lot of deaths were reported due to Tropical Cyclone.

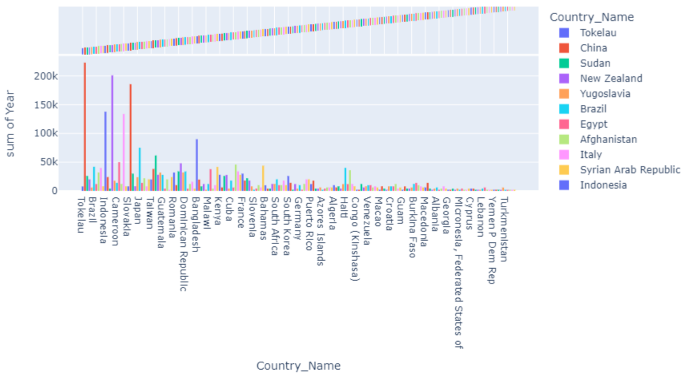


Fig. 51: Disaster by Country

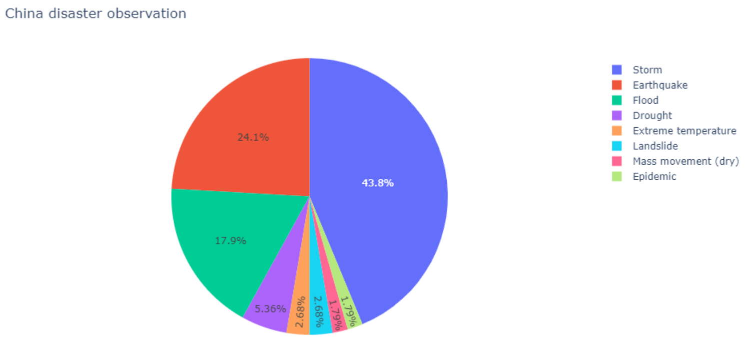


Fig. 52: Disasters in China

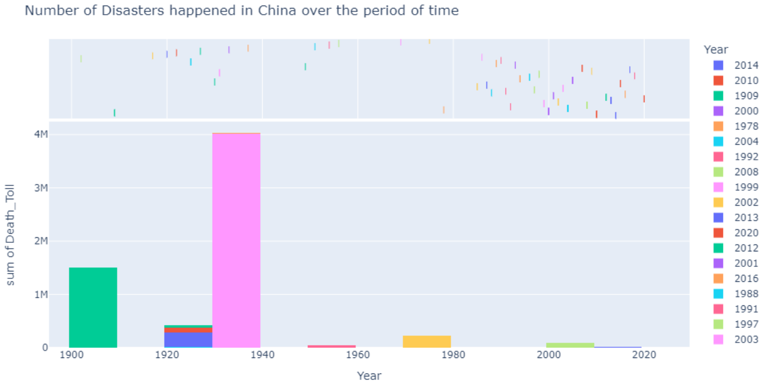


Fig. 53: Disasters in China by Year

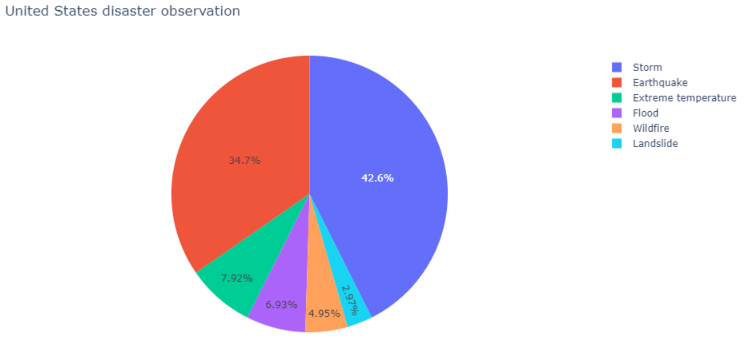


Fig. 54: Disasters in USA

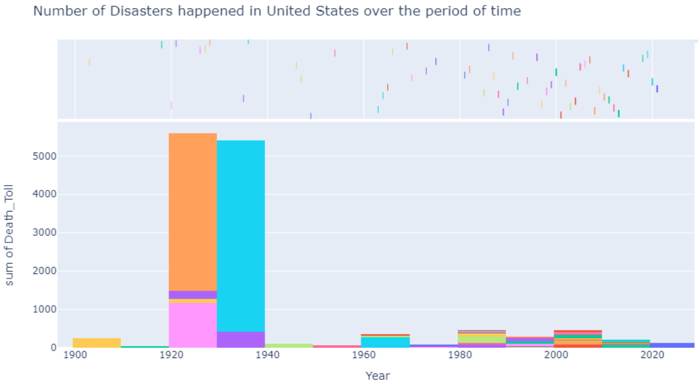


Fig. 55: Disasters in USA by Year

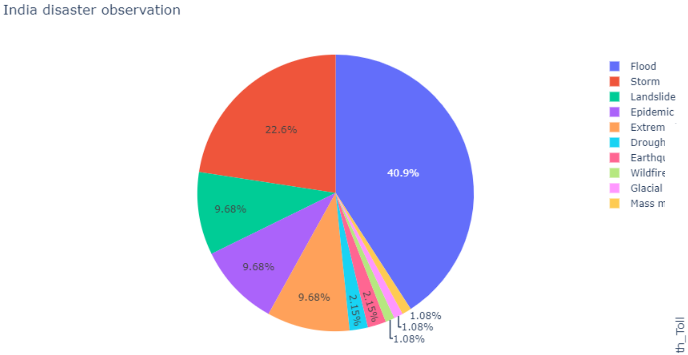


Fig. 56: Disasters in India

8) : Fig 58 and 59 shows that Indonesia has been mos effected by Earthquake. 2004 is the worst year in Indonesia from disaster perspective when lot of deaths were reported due to Tsunami.

9) : Fig 60 and 61 shows that Philippines has mostly been effected by Storm. 1922 is the worst year in Philippines from disaster perspective when lot of deaths were reported due to Tropical Cyclone.

10) : Azure final sql tables are in the following pic.

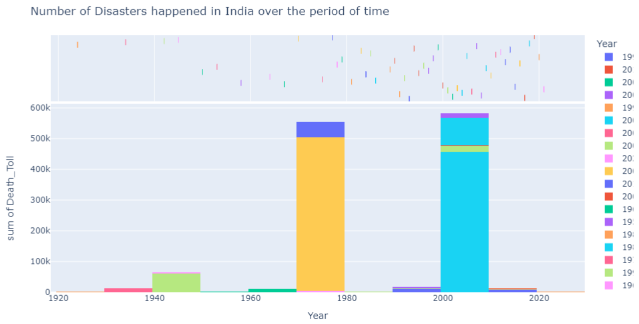


Fig. 57: Disasters in India by Year

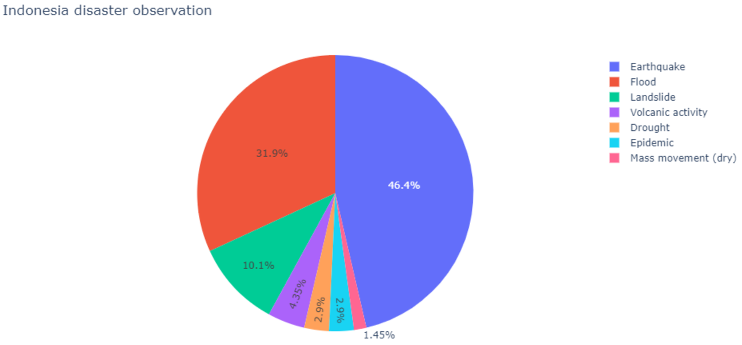


Fig. 58: Disasters in Indonesia

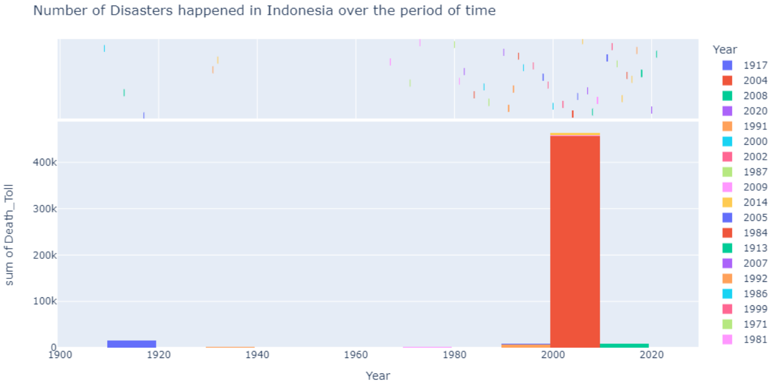


Fig. 59: Disasters in Indonesia by Year

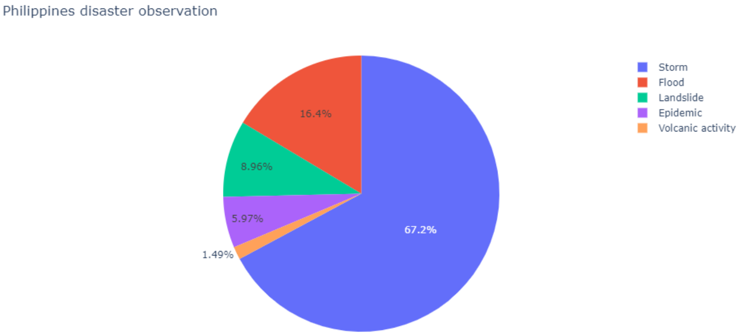


Fig. 60: Disasters in Philippines

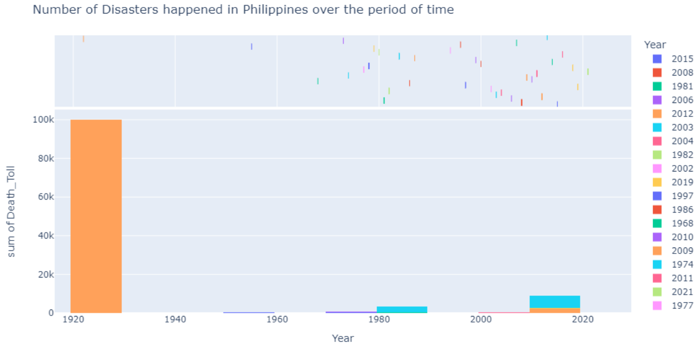
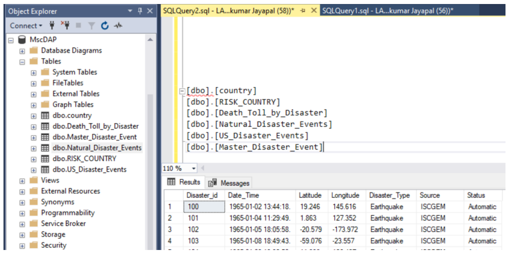


Fig. 61: Disasters in Philippines by Year



```
[dbo].[country]
[dbo].[RISK_COUNTRY]
[dbo].[Death_Toll_by_Disaster]
[dbo].[Natural_Disaster_Events]
[dbo].[US_Disaster_Events]
[dbo].[Master_Disaster_Event]
```

	Disaster_id	Date_Time	Latitude	Longitude	Disaster_Type	Source	Status
1	100	1965-01-02 13:44:18.	19.246	145.616	Earthquake	ISCGEM	Automatic
2	101	1965-01-04 11:29:49.	1.863	127.352	Earthquake	ISCGEM	Automatic
3	102	1965-01-05 18:05:58.	-20.579	-173.972	Earthquake	ISCGEM	Automatic
4	103	1965-01-08 18:49:43.	-59.076	-23.557	Earthquake	ISCGEM	Automatic

Fig. 62: Database Objects

VIII. Conclusion and Future Work

The natural disasters are frequently happened due to the geographical position of the country. Thus, natural disasters mostly occurred as an impact of natural condition. This research has completed an analysis to the data collected from various source systems to find out the correlation between the natural disasters happened, the number of victims. We also found that 1931 was the worst-hit year, 4 million deaths have been reported and the geographical disaster (Flood, Storm, Earthquake) and Asian countries are more frequently impacted zones. In America, Storm is the prominent disaster and the worst-hit year is 1936 when heat waves and storms were reported. The machine learning algorithm Kmeans clustering and ARIMA time series model were used to cluster countries according to risk scores and to forecast disasters using ARIMA time series models respectively. In the future, the research will focus on forecasting the disaster type in each country and continent.

References

[1] Ain Shams Engineering Journal ”A review on application of data mining techniques to combat natural disasters”,Saptarsi Goswami, Sanjay Chakraborty, Sanhita Ghosh, Amlan Chakrabarti ,Basabi Chakraborty,Volume 9, Issue 3, September 2018, Pages 365-378[online], Available:https://www.sciencedirect.com/science/article/pii/S2090447916000307

[2] Nick Brooks, W. Neil Adger, Country level risk measures of climate-related natural disasters and implications for adaptation to climate change, Tyndall Centre for Climate Change (2013), pp. 1-30

[3] ”Ontology based data warehouse modeling and mining of earthquake data: prediction analysis along Eurasian–Australian continental plates.Nimmagadda Shastri L, Dreher Heinz.

[4] Prihandoko, Bertalya, “A Data Analysis Of The Impact Of Natural Disaster Using K-Means Clustering Algorithm”, vol. 8, no. 4, December 2016.

[5] W. I. D. Mining, “Data Mining: Concepts and Techniques,” Morgan Kaufinann, 2006.K. Elissa, “Title of paper if known,” unpublished.

[6] Natural disaster risk analysis for critical infrastructure systems: An approach based on statistical learning theory, Author:Seth D.Guikema,Reliability Engineering & System Safety,Volume 94, Issue 4, April 2009, Pages 855-860.

[7] ”Big Data in Natural Disaster Management: A Review” by Manzhu Yu,Chaowei Yang and Yun Li,NSF Spatiotemporal Innovation Center, George Mason University, 4400 University Drive, Fairfax, VA 22030, USA

[8] Mathew C. Schmidtlein, Christina Finch& Susan L. Cutter. “Disaster Declarations and Major Hazard Occurrences in the United States. ”.Published online: 31 Jan 2008, Pages 1-14

[9] Cummins, D.J., M. Suher, et al. 2010. Federal Financial Exposure to Natural Catastrophe Risk, in Measuring and Managing Federal Financial Risk. D. Lucas (editor). Chicago, IL: University of Chicago Press Books: 61–96.

[10] Carolyn Kousky, Brett Lingle, and Leonard Shabman.” FEMA Public Assistance Grants: Implications of a Disaster Deductible” ”.Published online: APR 2016, Pages 1-17

[11] Disaster risk management series,Book:”Natural Disaster Hotspots, A global risk analysis”, Author:Maxx dilley, Robert S chen.