Earthquake Analysis

Team: CommUnify

Members

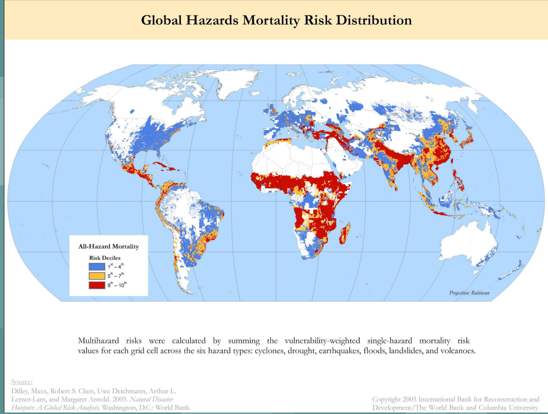
Panth Patel - ppate429@uic.edu - 676440358
Krish Patel - kpate400@uic.edu - 667120827

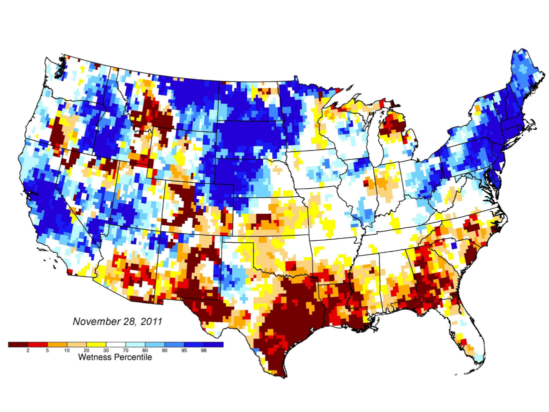
Problem: Big Idea

- The **Big Idea** for our project was is to analyse the data for a natural disaster like **Earthquake**

- Find the **Relationship** between the **Location** and **Number of Deaths** that occured in that location.

- Know the **location** and **intensity** in order to **save valuable lives** of the people



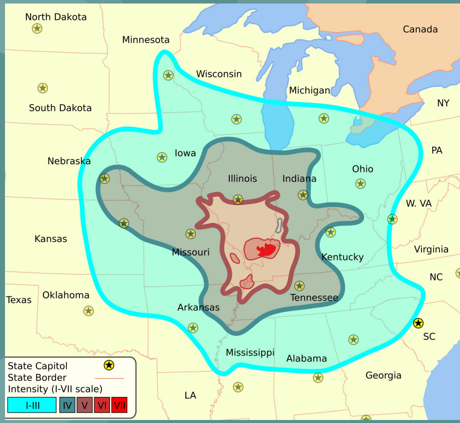


Problem: Research Questions

- **The question that we want to answer here with this analysis are:**

1. What is the most likely latitudes and longitudes of the upcoming major earthquakes that will occur in USA?

2. What is the potential number of lives that can be saved by accurately predicting an earthquake before it occurs in USA?



Problem: Problem Selection

- Natural disasters cause a **significant** number of deaths globally.
- The project aims to explore if there is a way to **predict** the occurrence of natural disasters.
- The prediction of natural disasters could **help save lives**.
- The recent natural disasters in Turkey and Syria **inspired** the idea for this **project**.
- The **death toll** from these disasters was **high** and prompted the need for a **solution**.

Data:

- We plan to use two datasets that are relevant to the project.

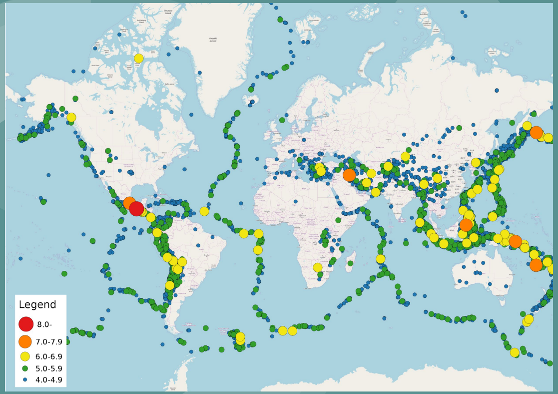
- EM-DAT – The International Disaster Database

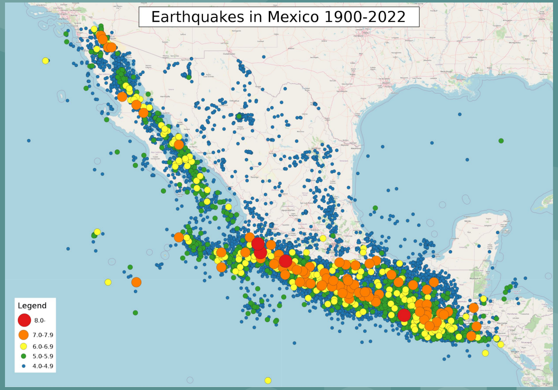
- **Data Collection:** Collected by downloading data from EmDat resources.

- **Size:** 12470 rows × 50 columns

- **Type of Data:** Float, Int and Strs.

- **Features of Data:** The Data contains impact(significance, magnitude), Type of natural disaster(subtype, etc.), the financial loss due to the natural disaster, it also contains time, and death count for the natural disaster that occured.





Data:

- [CORGIS Dataset – USA Earthquake Database]

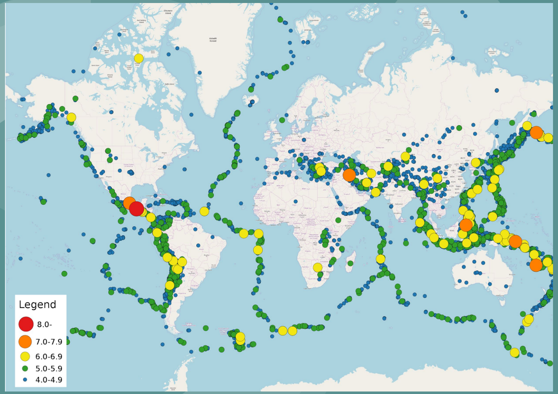
- **Data Collection:** earthquakes.csv file

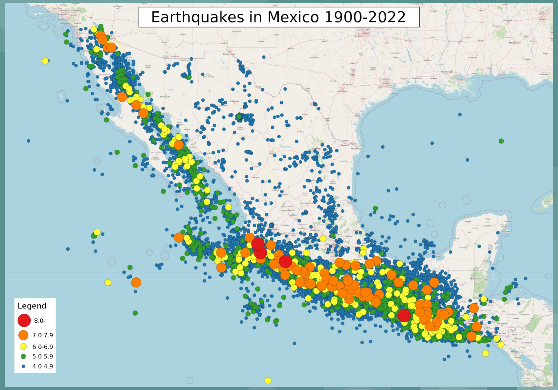
- **Size:** 8395 rows and 18 columns

- **Data:** Impact (magnitude, significance, depth), Location(city, state, latitude, longitude), Time(date, time, month, year) of each Earthquake

- **Type of Data** - Datetime, text, float

- **Other Relevant Information** - Data from 27 July 2016 to 25 August 2016 only





Data Cleaning:

1. <u>EM-DAT – The International Disaster Database</u>

- Change the **column names** to to the ones in the dataset.
- Drop the columns that are not necessary for the analysis and make a new data set with **Year, Disaster Type, Latitude, Longitude and Total Deaths**.
- Convert total **Deaths** into **int** data type.
- Drop all **NAN values** using the dropna function.
- Keep the rows only where the **disaster type** is **Earthquake.**
- Now clean the **Latitude and Longitude** values and change the type from **string to float.**
- Analyze the **Latitude, Longitude and Total Deaths** and see if all the values are legitimate and store the new cleaned dataset in a **new pandas dataframe**.
- In the process of data cleaning, the new dataframe has shrinked form **50 columns to 5 columns** and from **12470 rows to 581 rows**.

	Source:	EM-DAT, CRED / UCLouvain, Brussels, Belgium	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	...	Unnamed: 40	Unnamed: 41	Unnamed: 42	Unnamed: 43	Unnamed: 44
5	Dis No	Year	Seq	Glide	Disaster Group	Disaster Subgroup	Disaster Type	Disaster Subtype	Disaster Subsubtype	Event Name	...	Reconstruction Costs, Adjusted ('000 US$)	Insured Damages ('000 US$)	Insured Damages, Adjusted ('000 US$)	Total Damages ('000 US$)	Total Damages, Adjusted ('000 US$)
6	1990-9579-CMR	1990	9579	NaN	Natural	Climatological	Drought	Drought	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
7	1990-0230-ECU	1990	0230	NaN	Natural	Geophysical	Earthquake	Ground movement	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
8	1990-0361-AUT	1990	0361	NaN	Natural	Meteorological	Storm	Convective storm	Severe storm	NaN	...	NaN	28200	63164	NaN	NaN
9	1990-9059-BOL	1990	9059	NaN	Natural	Climatological	Drought	Drought	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
12470	2023-0080-ZAF	2023	0080	NaN	Natural	Hydrological	Flood	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
12471	2023-0110-ZMB	2023	0110	NaN	Natural	Hydrological	Flood	Flash flood	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
12472	2023-0068-ZMB	2023	0068	EP-2023-000013	Natural	Biological	Epidemic	Bacterial disease	NaN	Cholera	...	NaN	NaN	NaN	NaN	NaN
12473	2023-0095-ZWE	2023	0095	NaN	Natural	Meteorological	Storm	Tropical cyclone	NaN	Tropical cyclone 'Freddy'	...	NaN	NaN	NaN	NaN	NaN
12474	2023-0022-SRB	2023	0022	NaN	Natural	Hydrological	Flood	Riverine flood	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN

12470 rows × 50 columns

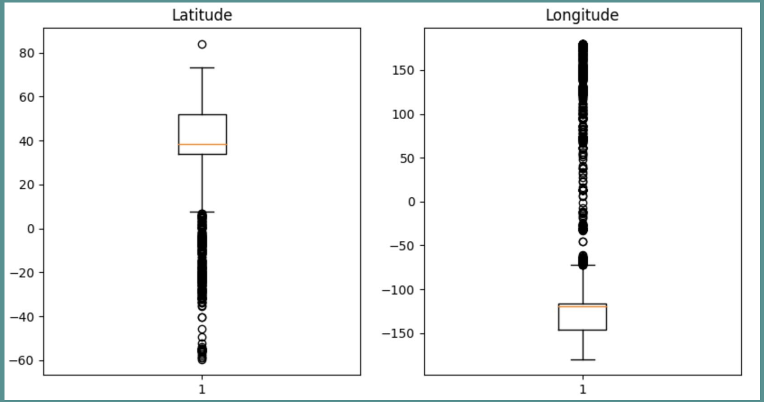
5	Year	Disaster Type	Latitude	Longitude	Total Deaths
7	1990	Earthquake	-0.259	-78.449	4
36	1990	Earthquake	45.841	26.668	1
52	1990	Earthquake	37.001	103.863	1
53	1990	Earthquake	35.986	100.245	126
74	1990	Earthquake	9.869	-84.302	1
...	...	...	...	...	...
12413	2022	Earthquake	40.847	30.967	2
12416	2022	Earthquake	23.119	121.422	1
12425	2022	Earthquake	40.525	124.423	2
12453	2023	Earthquake	38.055	36.510	38044
12466	2023	Earthquake	38.055	36.510	4500

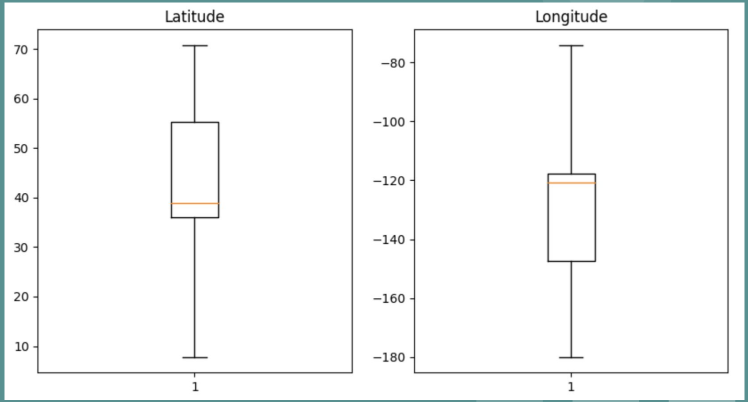
581 rows × 5 columns

Data Cleaning:

2. <u>CORGIS Dataset – USA Earthquake Database</u>

- Dropped Rows with **NaN values**
- Drop all columns other than **latitude, longitude, and time** as they are the only ones required to predict locations of **future earthquakes**
- Created **box plots** for latitude and longitude columns to detect and **remove outliers**. Below are the boxplots before and after removing outliers.
- During Data Cleaning the number of columns reduced from **18 to 3**. And the Number of Rows r**educed by 974 from 8394 to 7420**





Exploratory Data Analysis:

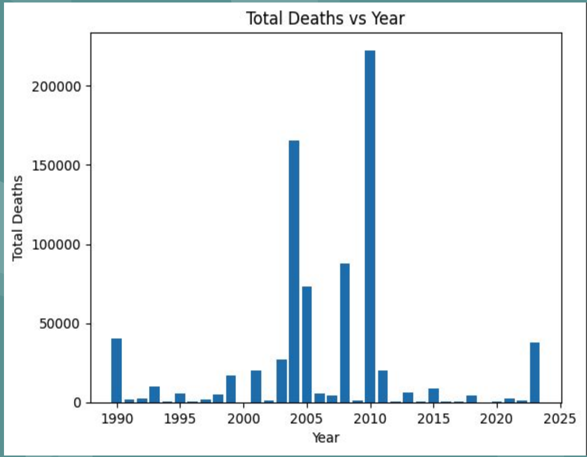
1. <u>EM-DAT – The International Disaster Database</u>

- The hypothesis that frequency and severity of natural disasters have increased over time resulting in a higher number of people affected each year tends to be an interesting one.

- The graph below shows that the number of deaths tends to be higher between 2004 and 2010 and decreases between 2010 and 2021.

- The visualization of total affected vs the year tends to be a useful tool in explaining that the hypothesis is false.

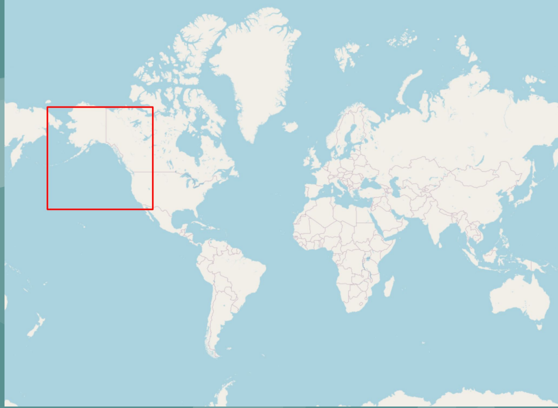
- Therefore, the frequency and severity of natural disasters may not have increased over time as previously thought.

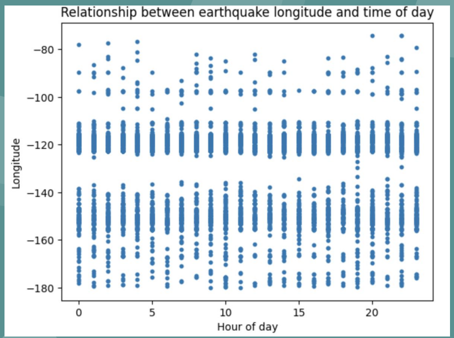


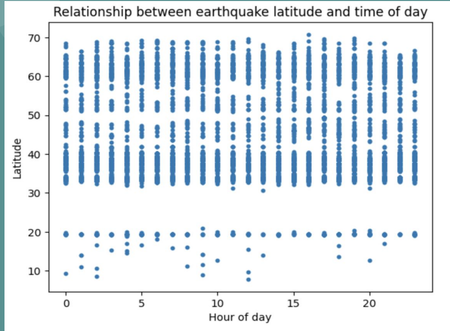
Exploratory Data Analysis:

2. <u>CORGIS Dataset – USA Earthquake Database</u>

- One interesting hypothesis to investigate based on the given EDA would be to check if there is a **relationship** between the **latitude/longitude** of the **earthquake** epicenters and the **time of day** when they occur.
- These **scatter plots** will show if there is any **trend or pattern** in the **occurrence** of earthquakes based on the time of day. If there is a noticeable pattern, it could suggest that certain times of day are more likely to have earthquakes and could be useful for predicting upcoming earthquakes. If there is no pattern, it could suggest that the time of day is not a useful predictor for earthquake occurrence.
- There is **no relation** between the **time of the day and longitude**. However we can notice that most earthquakes took place between **latitudes -110 to -180** and **longitudes 30 to 70**. The region is shown on the right. (Western Part of United States)







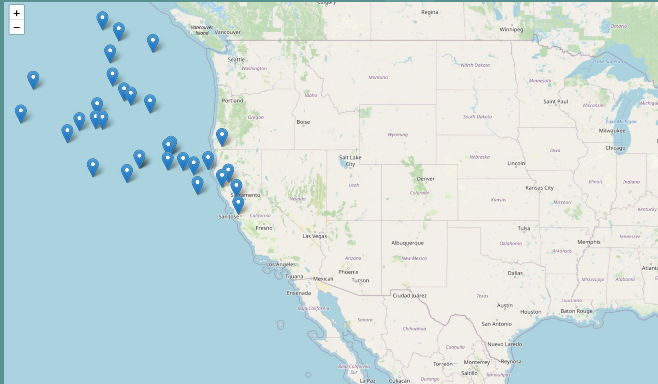
ML/Stats - Random Forest:

2. <u>CORGIS Dataset – USA Earthquake Database</u>

The aim was to **predict** the possible **locations** of future **earthquakes** for each day in the **month of May 2023** using the Random Forest Regression model. We performed the following steps:

1. We extracted the **'year', 'month', 'day', and 'hour'** features from the **'time.full' column** in the earthquake data using the 'dt' attribute of pandas dataframe.
2. We split the data into training and testing sets using '**train_test_split**' function from **scikit-learn** library.
3. We trained a **Random Forest Regression** model on the training data using '**RandomForestRegressor**' function from scikit-learn library.
4. We created a new dataframe '**new_data**' with all days of **May 2023** using '**pd.date_range**' function from pandas library.
5. We predicted the locations for each day of May 2023 using the trained Random Forest Regression model and the 'predict' function from scikit-learn library.
6. We stored the predicted locations in a new dataframe 'predicted_data' with columns 'latitude' and 'longitude' and we set the index as the dates for each day in May 2023.

On plotting the predicted locations on map we found out that all the locations were on the Western Part of USA



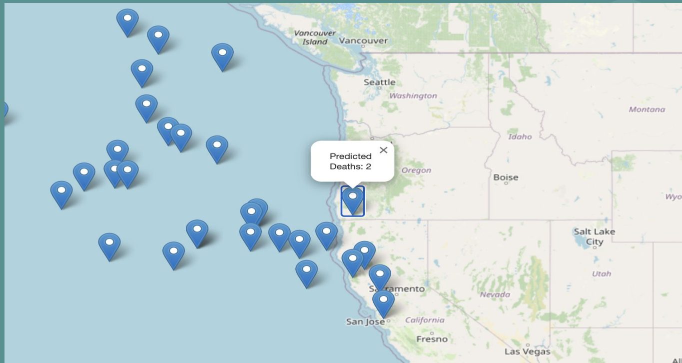
ML/Stats - KNN:

1. EM-DAT – The International Disaster Database

In this we wanted to predict the **total number of deaths** that can be caused by an **Earthquake**, which will help us to answer the second question where we needed to report the death **counts/lives** that can be **saved** with the analysis. Below are the steps to show how we implemented the model and got the predictions.

1. We started by extracting the necessary information like **latitude and longitude** in a dataframe and called it **data_x**, we then used the **death counts** as the **target** value.
2. After that we split the data into **training and testing** set using the **train_test_split module**. Where training set was **20%**
3. We then trained our **KNN model** on the **training set** and used to it get **predictions** for the **testing set.**
4. Once we got the prediction we measured the accuracy for various **k values**, and used the value with the **highest accuracy** to **predict the data.**
5. Once we predicted the data we used it to plot a graph between the **latitude and longitude** and the **number of deaths** that can occur at those **location**.

In this we uncovered that all the data that was used from Em-Dat dataset contained a lot of values that had deaths as 1, due to which we received a lot of prediction values as 1.



Lessons Learned:

The main takeaways from the project are:

- It is hard to predict the exact location as well as the exact number of death counts.
- The accuracy of prediction is really low and due to which the reliability of the prediction tends to be low.
- We will need to take into focus a lot of other factors such as location in urban areas/rural areas, population density, other disaster that can occur due to an Earthquake(eg. Tsunami).
- The process of formulating a question as well as journey from data collection to prediction has taught us the importance and impact of data science.
- Our model was an underfit for the problem that we wanted to analyze due to which we were not able to receive accurate predictions.

Teamwork:

- Each of the task performed in the project were distributed equally to each of the team member:

Krish Patel

- **Data Collection** - EMDat Dataset
- **Data Cleaning** - EMDat Dataset
- **Exploratory Data Analysis and Visualizations.**
- **Model Planning:** KNN, Linear Regression, and Association Rule Mining.
- **ML/Stat:** Perform the KNN on the data collected.

Panth Patel

- **Data Collection** - CORGIS Dataset
- **Data Cleaning** - CORGIS Dataset
- **Exploratory Data Analysis and Visualizations.**
- **Model Planning:**ARMA, K-Means Clustering, and Random Forest Classification.
- **ML/Stat:** Perform the Random Forest Classification on the data collected.

References:

- https://sherbold.github.io/intro-to-data-science
- https://regenerativetoday.com/learn-to-formulate-good-research-question-for-efficient-statistical-analysis/
- https://corgis-edu.github.io/corgis/csv/earthquakes/
- https://public.emdat.be/
- https://medium.com/analytics-vidhya/ml-algorithms-pros-cons-and-suitable-usages-b377c3c09f1b
- https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
- https://docs.google.com/presentation/d/1FzOIvXhKBhWwsZ0Hi6-jnYn0SYEYdVop/edit#slide=id.p1
- https://pypi.org/project/folium/

Thank You