**<Data Science Tool Box: Python Programming >**

**PROJECT REPORT**

(Project Semester January-April 2025)

***(Landslide Dataset Analysis)***

Submitted by

Abhay Chaudhary

Registration No: 12315216

Programme and Section BTech CSE

Course Code: INT375

Under the Guidance of

**(Name of faculty coordinator U.Id and designation)**

**Discipline of CSE/IT**

**Lovely School of: Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Abhay Chaudhary bearing Registration no. 12315216 has completed INT 375 project titled, **“*Landslide Dataset Analysis*”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of …………………………………………….**

Lovely Professional University

Phagwara, Punjab.

Date:

**DECLARATION**

I, Abhay Chaudhary, student of BTech Computer Science Engineering under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: Signature

Registration No:12315216 Abhay Chaudhary

Landslide Dataset Analysis Report

# 1. Introduction

Landslides are one of the most dangerous natural disasters, often resulting in human casualties, injuries, and property damage. This project focuses on performing Exploratory Data Analysis (EDA) on the Global Landslide Catalog to uncover patterns, trends, and insights. By analyzing fatalities, injuries, geographical data, and triggers, we aim to better understand landslide behavior across the globe.

# 2. Source of Dataset

- Dataset Name: Global Landslide Catalog  
- Source: NASA's Global Landslide Catalog (GLC)  
- Format: CSV  
- Content: Includes fields like fatality\_count, injury\_count, country\_name, landslide\_trigger, admin\_division\_population, gazeteer\_distance, and geographic details.

# 3. EDA Process

The EDA process included:  
- Importing and cleaning the data  
- Handling missing values  
- Detecting and removing outliers  
- Analyzing correlations between variables  
- Visualizing data using plots like bar charts, box plots, histograms, and heatmaps

## Analysis 1: Most Common Landslide Triggers

i. Introduction  
Understanding what commonly causes landslides helps in planning preventive measures.

ii. General Description  
The dataset contains a column 'landslide\_trigger' that specifies causes like rainfall, earthquakes, mining, etc.

iii. Specific Requirements, Functions and Formulas  
Used .value\_counts() to count occurrences. Visualized with seaborn.barplot.

iv. Analysis Results  
Rainfall and downpour are among the most common causes of landslides globally.

v. Visualization  
Top 10 triggers were visualized using a bar chart.

## Analysis 2: Distribution of Fatalities and Injuries

i. Introduction  
Understanding how deadly the landslides are in various locations.

ii. General Description  
Numeric columns fatality\_count and injury\_count are analyzed.

iii. Specific Requirements, Functions and Formulas  
Handled missing values with .fillna(0). Visualized with sns.histplot.

iv. Analysis Results  
Most landslides cause fewer than 10 fatalities or injuries, but a few extreme cases exist.

v. Visualization  
Two histograms showing frequency distributions.

## Analysis 3: Outlier Detection

i. Introduction  
Outliers can skew analysis and need to be detected or treated.

ii. General Description  
Columns analyzed: fatality\_count, injury\_count, admin\_division\_population, gazeteer\_distance.

iii. Specific Requirements, Functions and Formulas  
Used IQR method: Outlier = value < Q1 - 1.5\*IQR or > Q3 + 1.5\*IQR. Visualized using seaborn.boxplot.

iv. Analysis Results  
Outliers were detected in all 4 fields, especially in fatalities and population.

v. Visualization  
Multiple boxplots highlighting the outliers.

## Analysis 4: Correlation Between Numeric Features

i. Introduction  
To understand relationships like whether population affects fatalities.

ii. General Description  
Correlation computed between numeric fields.

iii. Specific Requirements, Functions and Formulas  
Used .corr(). Visualized with sns.heatmap.

iv. Analysis Results  
Low to moderate correlation was found between population and fatalities/injuries.

v. Visualization  
Heatmap with correlation matrix.

## Analysis 5: Spatial Distribution by Country

i. Introduction  
Some countries experience more landslides than others.

ii. General Description  
Analyzed country\_name column to find frequency.

iii. Specific Requirements, Functions and Formulas  
Used .value\_counts(). Visualized with barplot.

iv. Analysis Results  
Countries like India, Nepal, and the Philippines reported the highest number of landslides.

v. Visualization  
Bar chart showing top 10 affected countries.

# 5. Conclusion

The EDA helped uncover meaningful patterns in the landslide dataset:  
- Rainfall is the leading cause.  
- Fatalities and injuries are low in most cases but can be extreme.  
- Remote areas have different characteristics compared to urban ones.  
- Country-wise trends help identify regions needing better disaster management.

# 6. Future Scope

- Time-Series Analysis: Analyze changes over time.  
- Geospatial Mapping: Use folium or plotly for interactive maps.  
- Predictive Modeling: Train machine learning models to predict fatalities or risk levels.  
- Risk Index Development: Combine population, distance, and trigger types to assess risk.

# 7. References

- NASA Global Landslide Catalog  
- Python Libraries: pandas, matplotlib, seaborn  
- https://data.nasa.gov  
- https://seaborn.pydata.org  
- https://matplotlib.org