ANALYSIS OF DISASTERS

Group Coursework - Disaster Management

Group Number - 26

Group Member:

Kishore Rajendra - 34812636

Sachin Suresh - 34812598

Umesh Uddar - 34884157

Vyom Khanna - 28965736

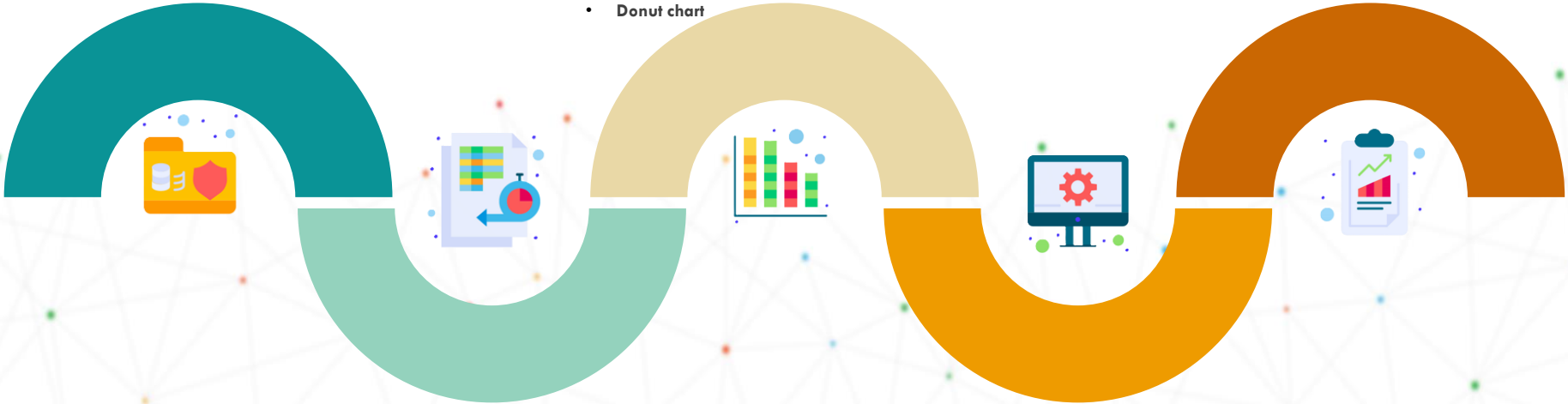
AGENDA

Data Extraction

Data Visualization

- Bar Plot
- Pie Chart
- Donut chart

Overview



Data Cleaning

- Imputation
- Data Cleaning

Analysis and Report

- Model Building
- Co - relation
- Classification

BACKGROUND

DesInventar stands at the forefront as a critical Disaster Information Management System, addressing the rising vulnerabilities of populations influenced by factors like population growth, urbanization, and inadequate infrastructure. Historically, the absence of systematic records obscured the understanding of small and medium-scale disasters. Originating from collaborative initiatives in Latin America, DesInventar became a globally applicable Disaster Inventory System, conceptualizing a unified framework for acquiring, collecting, and analyzing disaster information. With the backing of UNDP and UNDRR, this system empowers risk management dialogues, enabling the creation of National Disaster Inventories. As a comprehensive Disaster Information Management System, DesInventar facilitates the systematic documentation and analysis of disaster-induced losses. It serves as a powerful tool for informed decision-making, supporting strategic planning for effective prevention, mitigation, and preparedness measures.

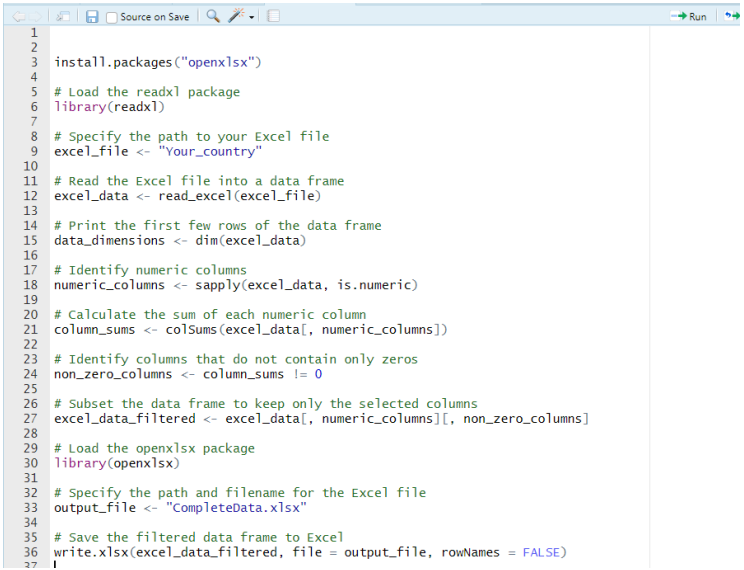
DATA EXTRACTION

Disaster Data from https://www.desinventar.net/ records various disasters across Cambodia, Ethiopia, Ghana, Indonesia, Jordan, Mali, Morocco, Mozambique, Myanmar, Nepal, Niger, Pakistan, and Iran.

DATA CLEANING

Steps Involved in Pre – processing

- Using R script the columns having all null values are removed.
- Applying count if formula across the column and the columns having 90% null data are removed.
- Manually the columns having similar combination are combined into single columns and the event combination are categorized based on the incident.



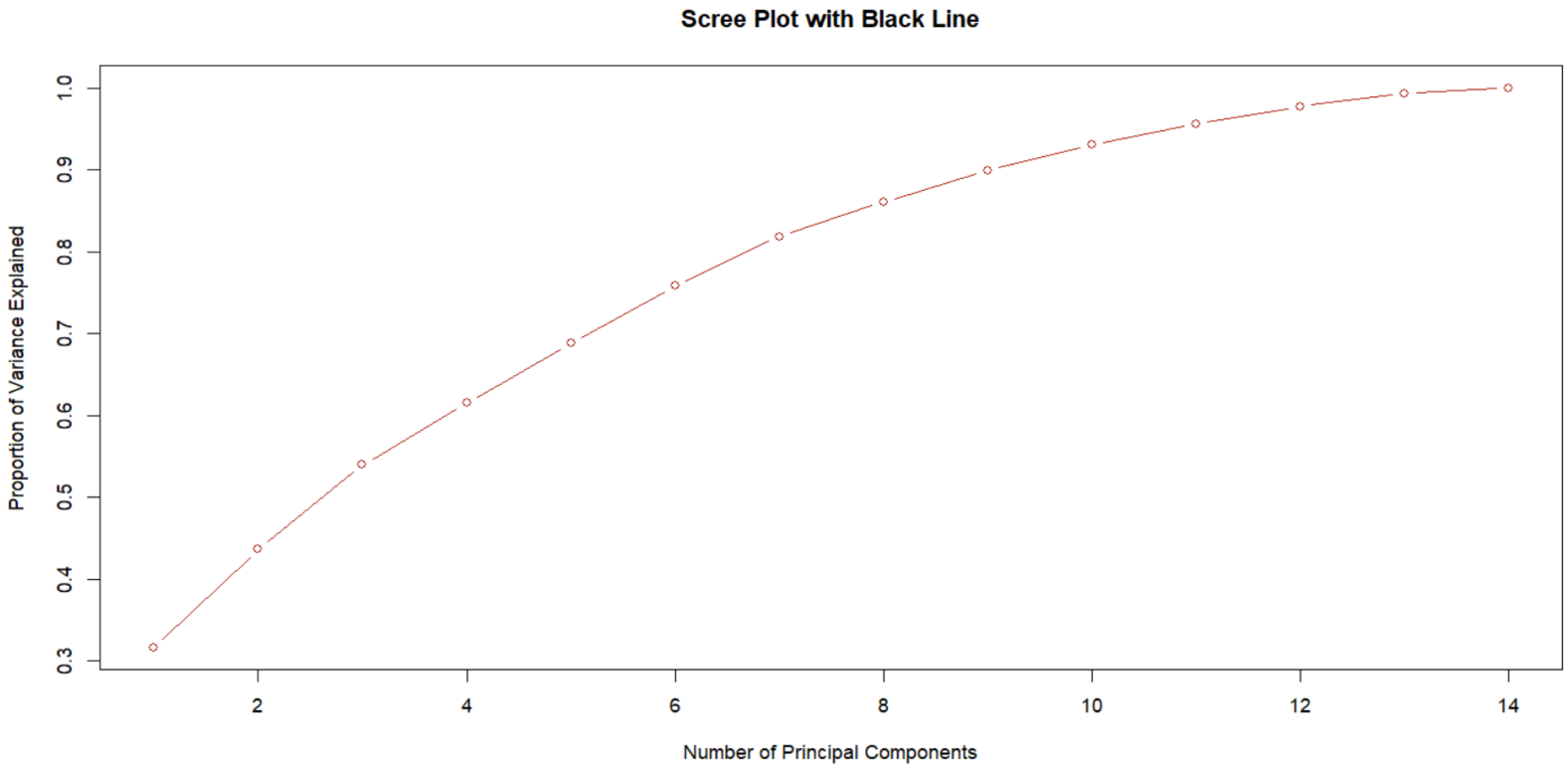
```r

install.packages("openxlsx")

# Load the readxl package
library(readxl)

# Specify the path to your Excel file
excel_file <- "Your_country"

# Read the Excel file into a data frame
excel_data <- read_excel(excel_file)

# Print the first few rows of the data frame
data_dimensions <- dim(excel_data)

# Identify numeric columns
numeric_columns <- sapply(excel_data, is.numeric)

# Calculate the sum of each numeric column
column_sums <- colSums(excel_data[, numeric_columns])

# Identify columns that do not contain only zeros
non_zero_columns <- column_sums != 0

# Subset the data frame to keep only the selected columns
excel_data_filtered <- excel_data[, numeric_columns][, non_zero_columns]

# Load the openxlsx package
library(openxlsx)

# Specify the path and filename for the Excel file
output_file <- "CompleteData.xlsx"

# Save the filtered data frame to Excel
write.xlsx(excel_data_filtered, file = output_file, rowNames = FALSE)
```

R Script to remove columns having all null values

Imputation Using Random Forest Method

1. Employed Random Forest imputation to predict and fill missing values in the dataset.
2. Leveraged relationships and patterns in existing data for accurate imputation.
3. Enhanced data completeness, ensuring robustness and reliability in the dataset.

PCA Principal Component Analysis

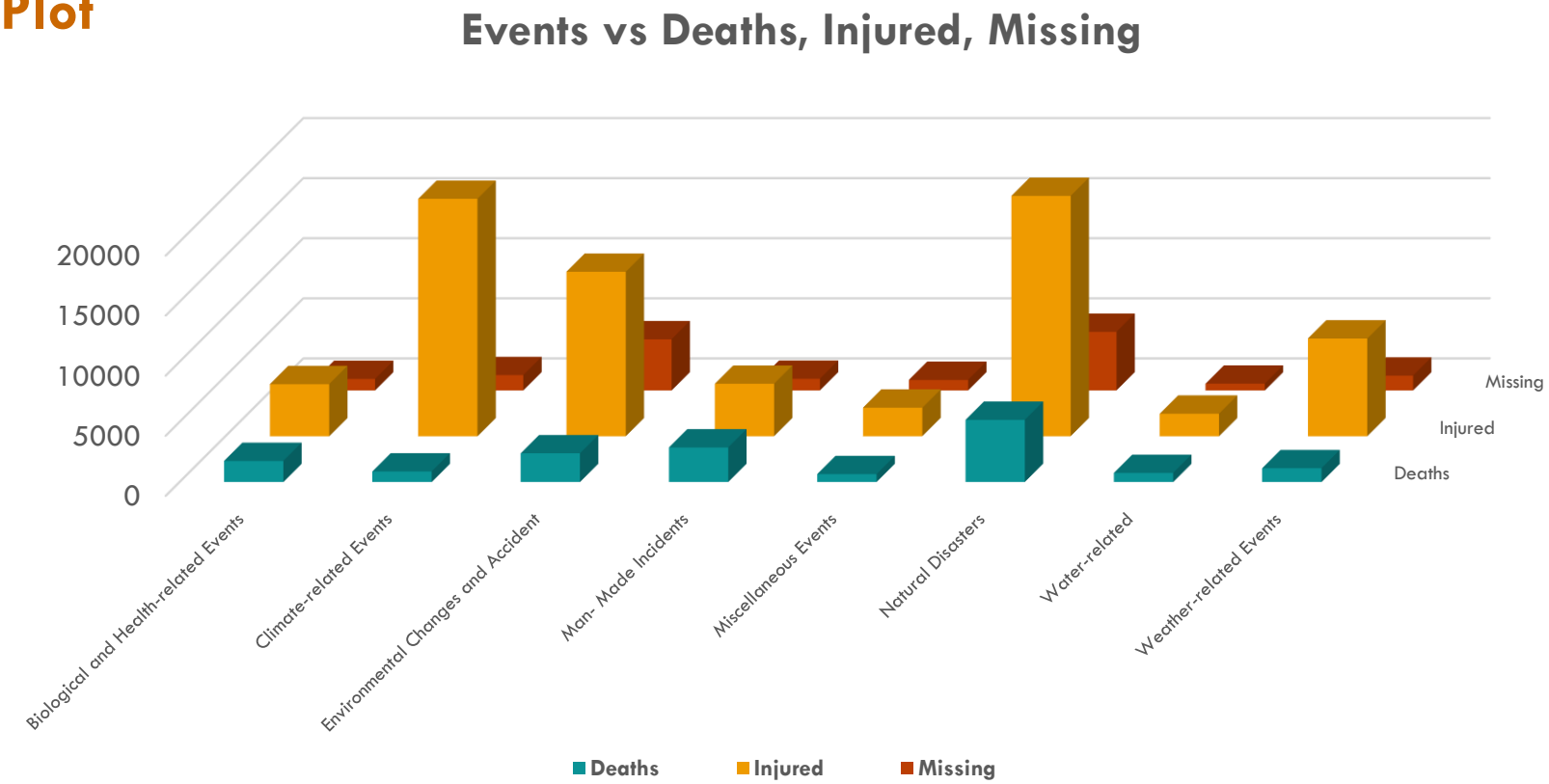


PCA is used here to reduce the dimensionality of the numerical data, retain the most significant information, and facilitate further analysis and interpretation of the dataset.

DATA VISUALIZATION

1. **Bar Graph -** Illustrates the frequency of death, missing, and injured occurrences across a range of different events that have occurred.
2. **Bar Graph -** Illustrates the number of Events occurred in each Decade.
3. **Area Plot –** Visualize the Event combination over the years.
4. **Pie Chart -** Distribution of infrastructure damage across number of events combined together.
5. **Donut Chart -** Distribution of infrastructure damage, deaths, and missing persons across two continents, Asia and Africa.

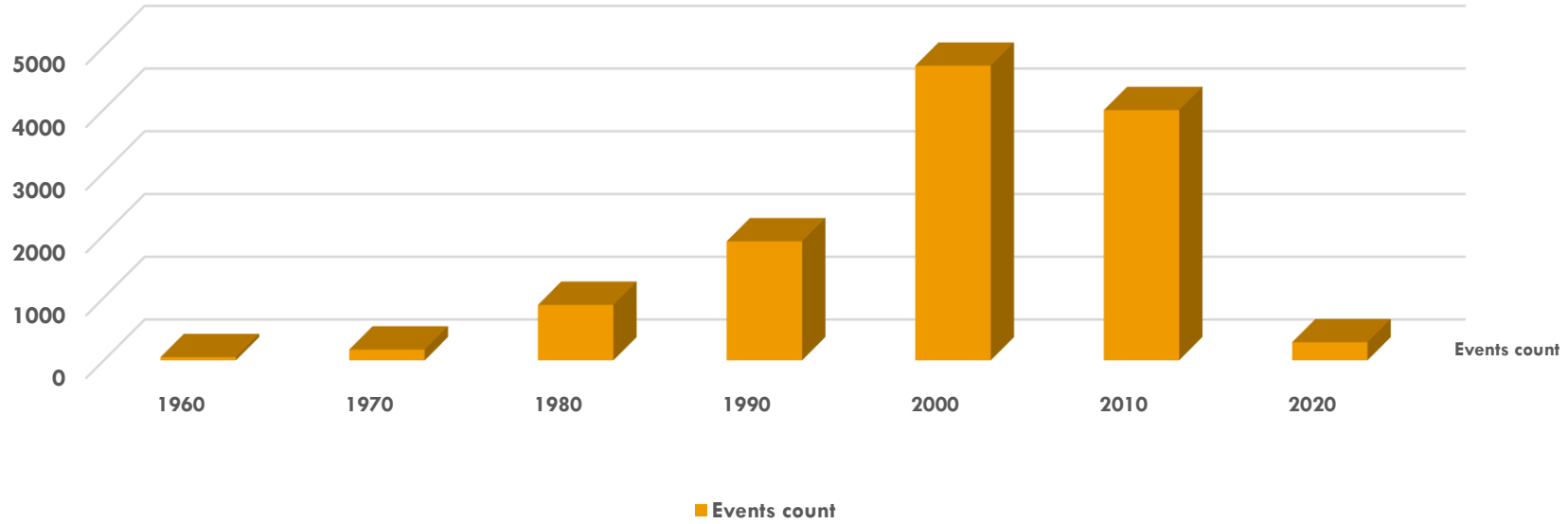
Bar Plot



Clear visual representation of the distribution of the above mention outcomes across different events.

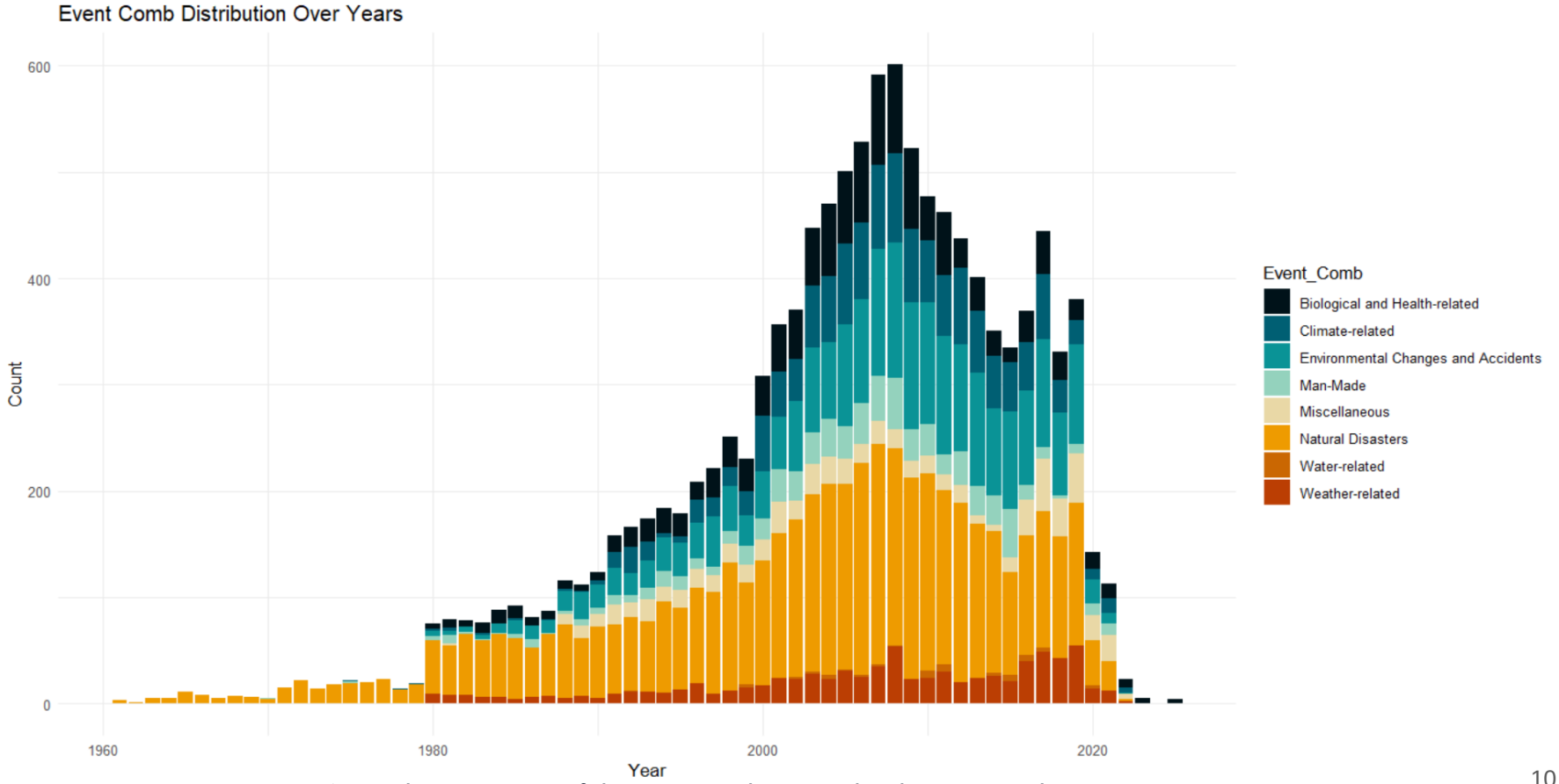
Bar Plot

Number of events in each decade.



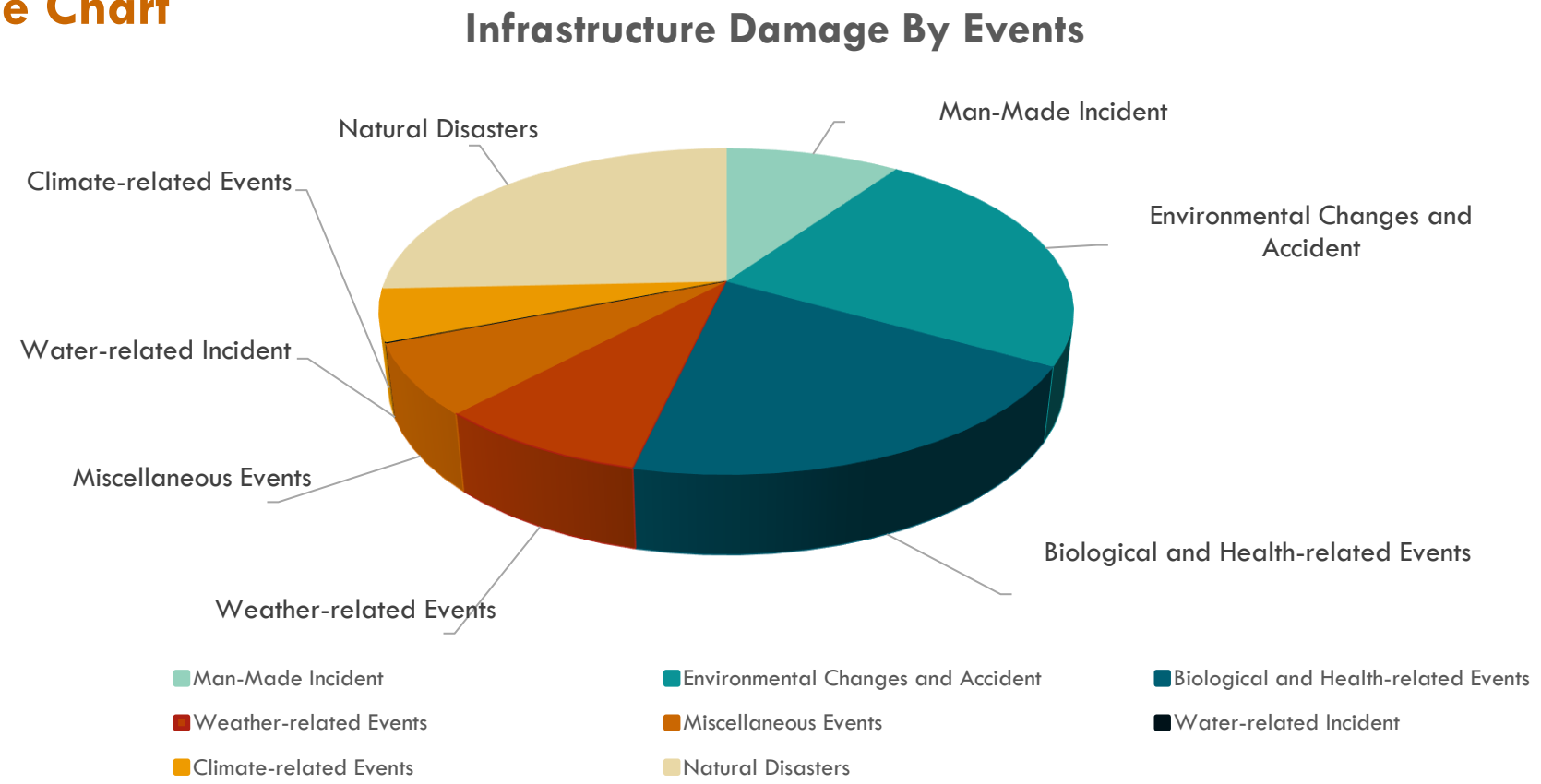
A visual examination of the event distribution across decades reveals a prominent peak in the 2000s, indicating that this decade experienced the highest number of Disaster events.

Area Plot

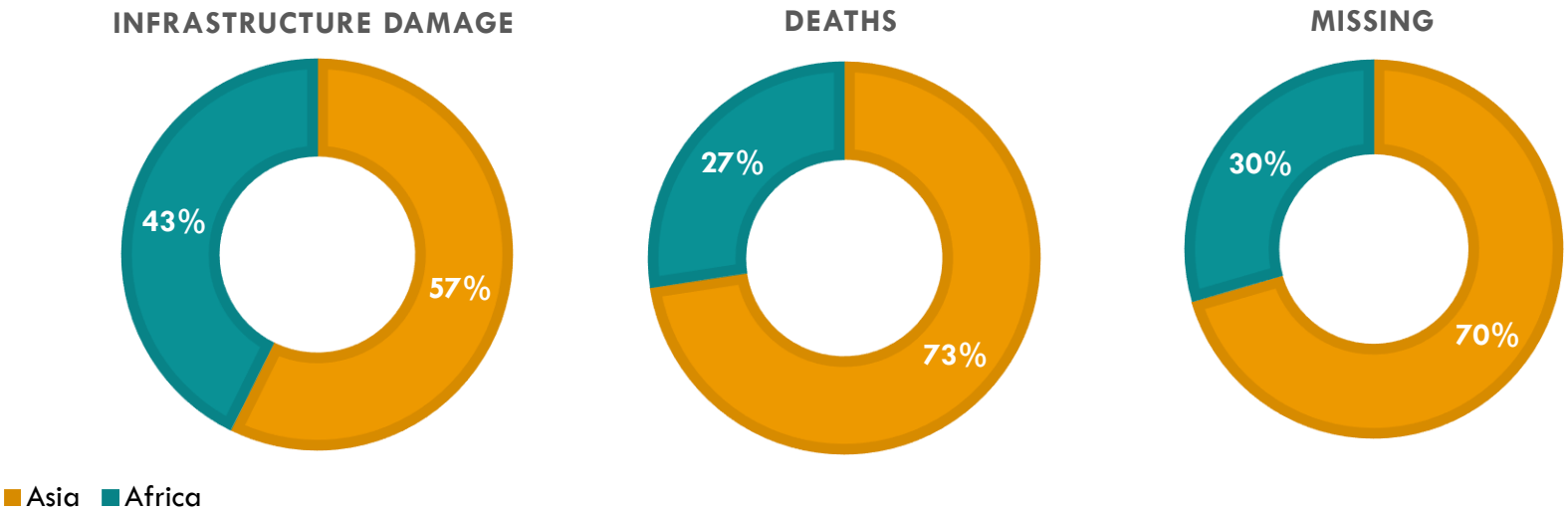


A visual examination of the event combination distribution over the years.

Pie Chart


Infrastructure Damage By Events
Man-Made Incident
Environmental Changes and Accident
Biological and Health-related Events
Weather-related Events
Miscellaneous Events
Water-related Incident
Climate-related Events
Natural Disasters
Man-Made Incident
Environmental Changes and Accident
Biological and Health-related Events
Weather-related Events
Miscellaneous Events
Water-related Incident
Climate-related Events
Natural Disasters

Donut Chart



An intuitive and quick way to visualize the impact of disasters on Asia and Africa is through the pie chart, which provides a visual snapshot of how infrastructure damage, deaths, and missing persons are spread across the two continents.

ANALYSIS AND REPORT

Multi linear regression and XG - Boost model for Losses_USD

Adjusted R-squared (Adjusted R^2) measures the goodness of fit for regression models, considering the number of predictors and penalizing irrelevant ones.

- Multilinear Regression, with an Adjusted R^2 of 0.735, explains about 73.5% of the variability in Losses_USD, indicating it captures some of the relationships but could benefit from enhancements.
- XG Boost, with an Adjusted R-squared of 0.8924, performs better than multilinear regression in understanding and predicting Losses_USD. It excels at capturing intricate patterns in the data and explaining a larger portion of the variability in financial losses.
- In simpler terms, the XG Boost method does a better job of predicting or understanding Losses in USD compared to the multilinear regression model with an effectiveness score of 0.735.

Multi linear regression and XG - Boost model for Deaths

- Multiple Linear Regression (MLR) Results (Adjusted $R^2 = 0.5037$) explains approximately 50.37% of Deaths' variability. Captures some relationships but may have limitations.
- XG Boost Results (Adjusted $R^2 = 0.720$) outperforms Multiple Linear Regression , explaining 72% of Deaths' variability. Its complexity captures intricate patterns for a superior fit. Overall **XG-Boost model** Predicts best model for the given disaster Data.

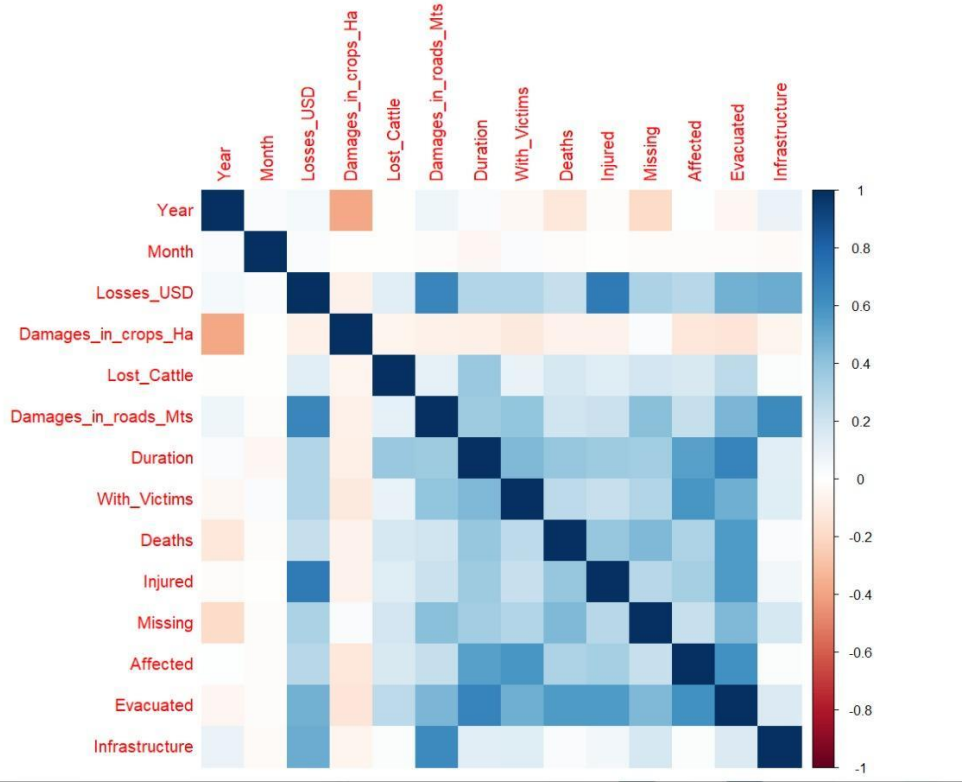
Finding Co – Relation using Heatmap

A correlation heatmap is a graphical representation of the correlation matrix, where each cell in the matrix is color-coded to represent the correlation between two variables.

Correlation coefficients range from -1 to 1.

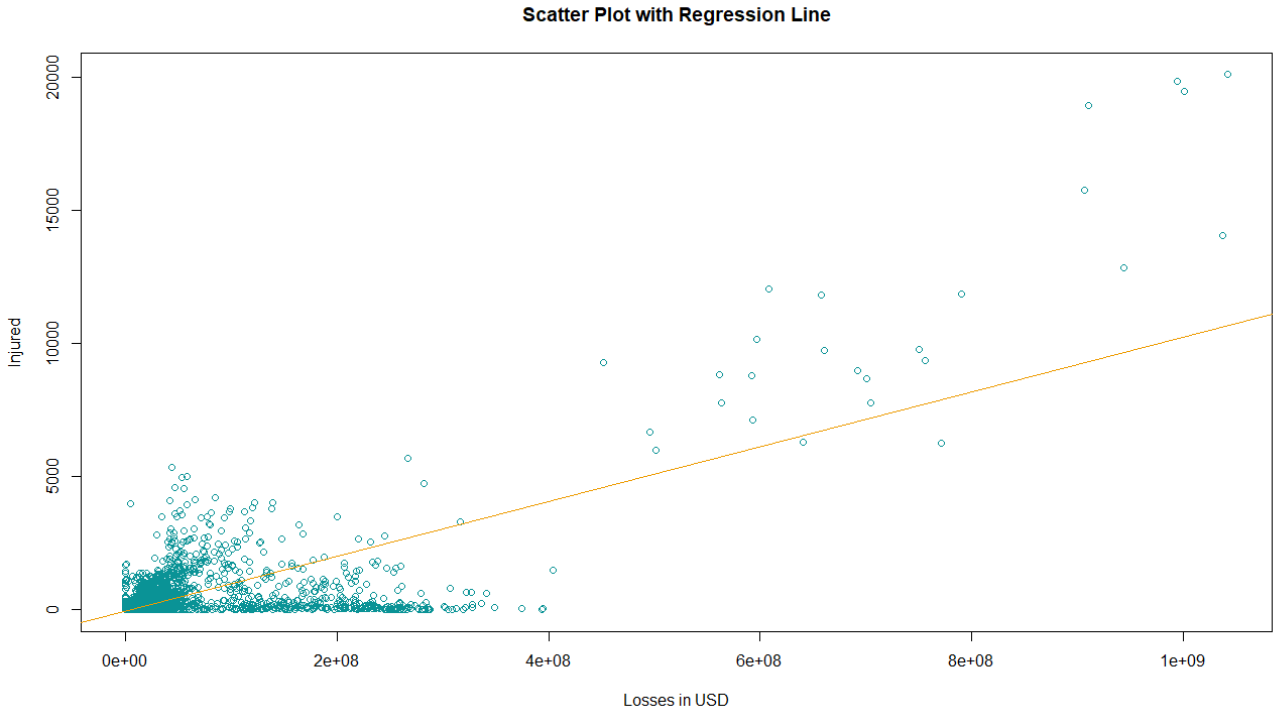
- A value of 1 indicates a perfect positive correlation (as one variable increases, the other variable increases proportionally).
- A value of -1 indicates a perfect negative correlation (as one variable increases, the other variable decreases proportionally).
- A value of 0 indicates no linear correlation.

The outcome of a correlation graph is a visual representation of the relationships among variables in a dataset. The graph illustrates correlations using color-coded indicators, highlighting positive and negative correlations.



Linear Regression model to predict Injured

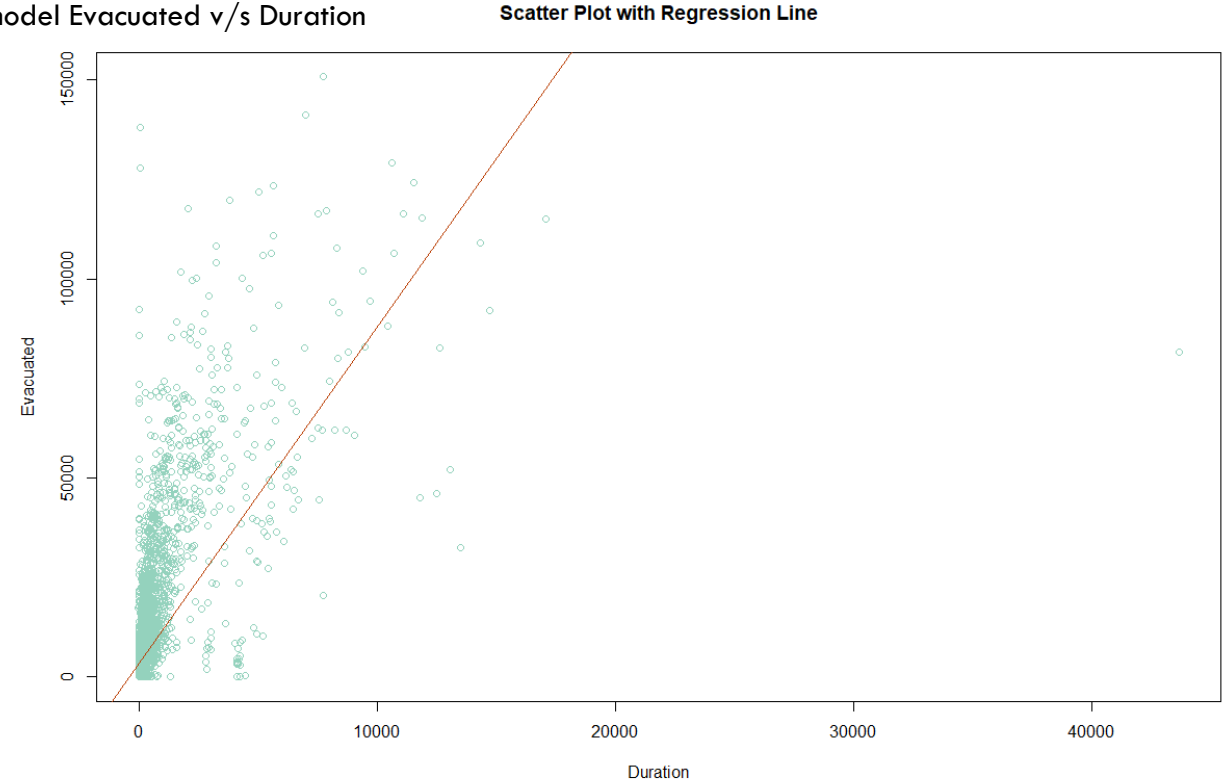
linear regression model injured v/s losses_usd



An adjusted R-squared of 0.50 suggests that half of the variability in the Injured variable is explained by the linear regression model with Losses_USD as the predictor.

Linear Regression model to predict People Evacuated

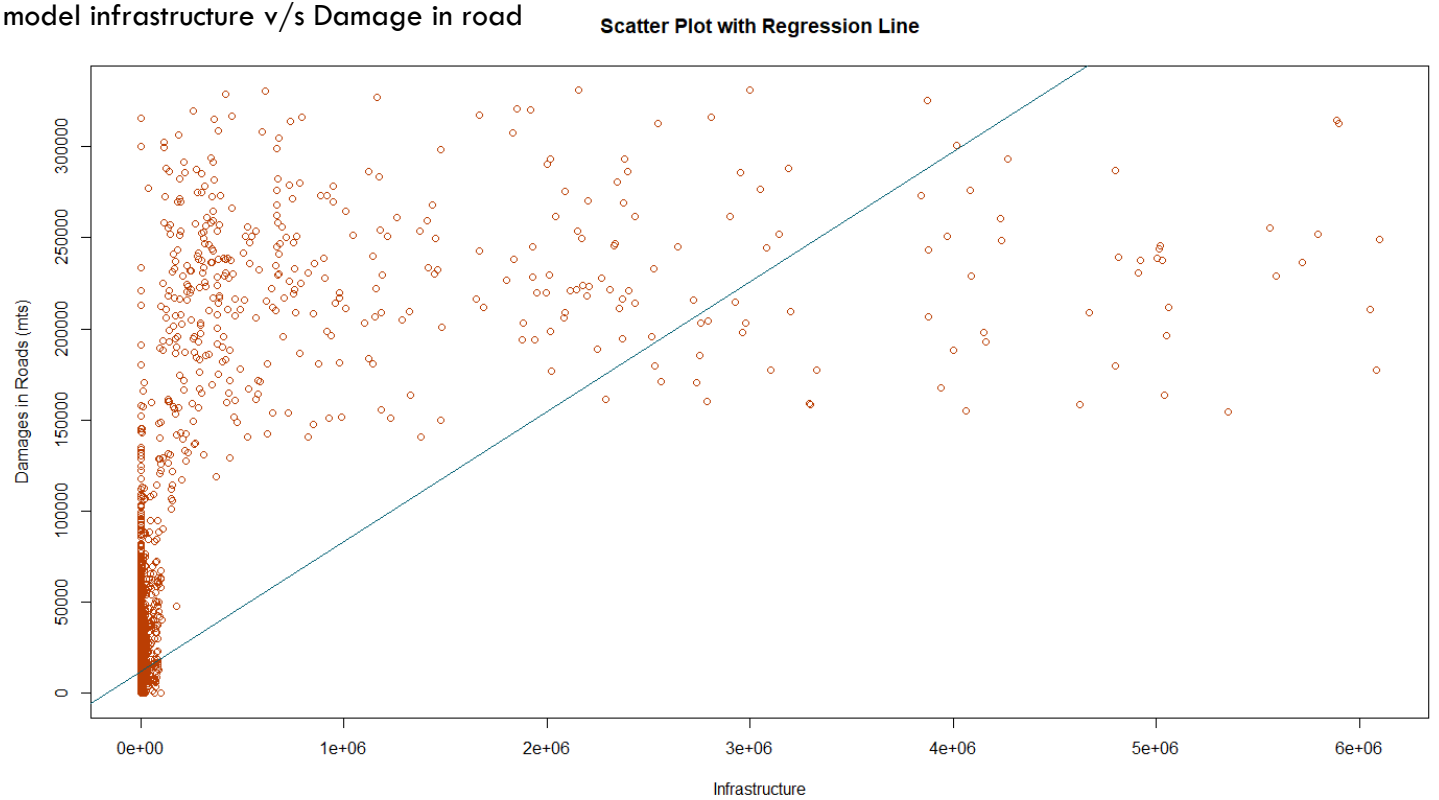
linear regression model Evacuated v/s Duration



An adjusted R-squared of 0.453 suggests that half of the variability in the Evacuated variable is explained by the linear regression model with Duration as the predictor.

Linear Regression model to predict Infrastructure Damage

linear regression model infrastructure v/s Damage in road



An adjusted R-squared of 0.3992 suggests that half of the variability in the Damages in road (mts) variable is explained by the linear regression model with Infrastructure damage as the predictor.

CLASSIFICATION

1. Random Forest: Random Forest, an ensemble learning algorithm, excels in classification and regression tasks, offering robustness, high accuracy, and adeptness in handling large and complex datasets. Known for its minimal hyperparameter tuning, it belongs to decision tree-based methods, providing reliable predictions.

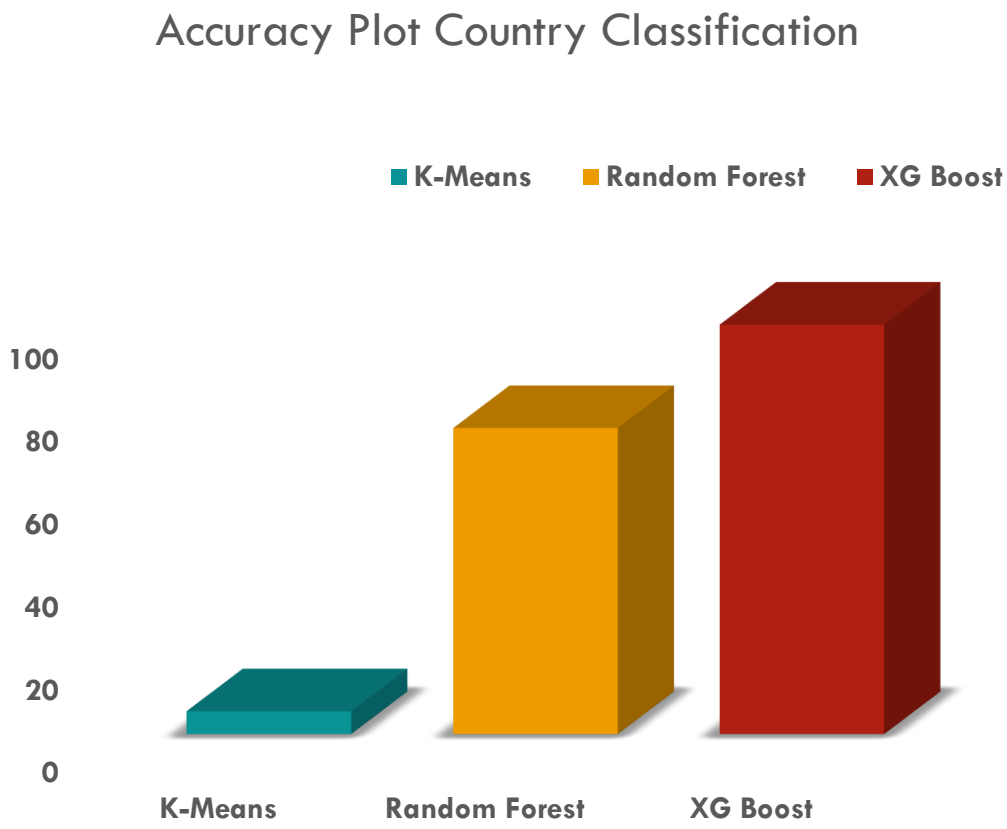
2. K-Means : K-Means is a widely used clustering algorithm for unsupervised partitioning of datasets into K distinct clusters. It iteratively minimizes within-cluster variance, ensuring data points are grouped to minimize the sum of squared distances to their assigned cluster centroids.

3. XG-Boost : XG-Boost is a versatile and robust machine learning algorithm, excelling in classification and regression tasks. By constructing a sequence of decision trees and iteratively correcting errors, XG-Boost minimizes the gradient of the loss function, showcasing its effectiveness in improving predictive accuracy.

Classification

1. Country

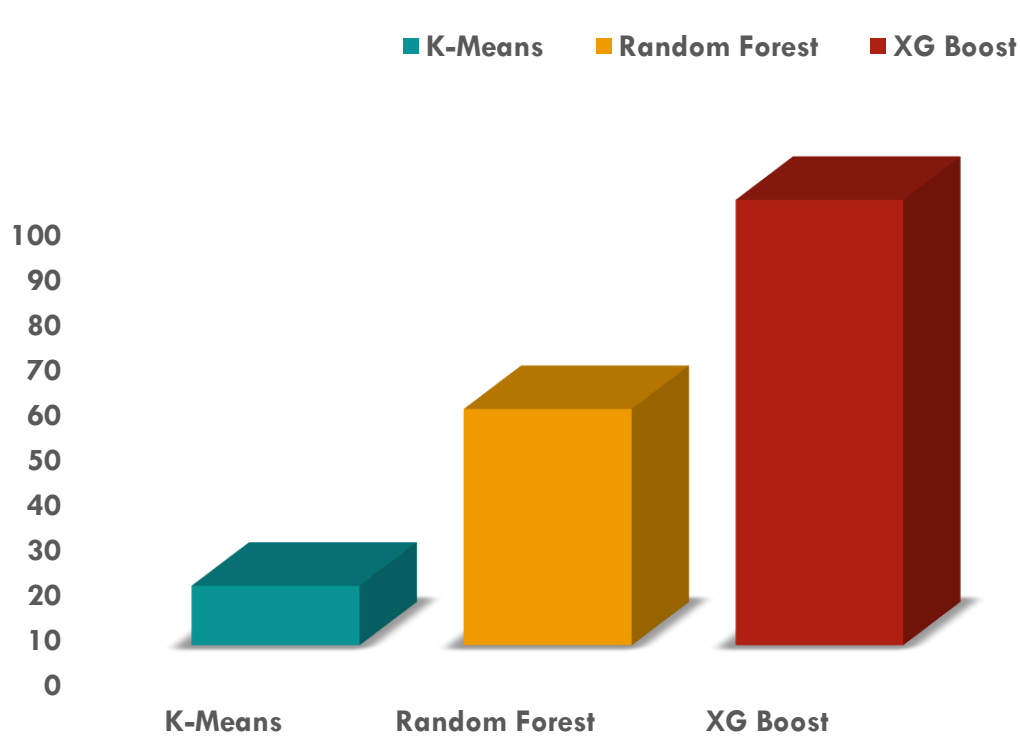
The Random Forest model achieved an accuracy of 73.76%, indicating its ability to correctly predict the country of a disaster nearly three-fourths of the time. In contrast, K-Means, primarily used for clustering, reported 5.56%, which might not directly represent classification accuracy. XG-Boost outperformed both with an accuracy of 86.2%, demonstrating its effectiveness in predicting the country of a disaster in the dataset.



Classification

2. Event_comb

Random Forest (52.42% accuracy) predicts disaster types by combining insights from multiple decision trees. K-Means (13.20%) clusters data without traditional accuracy, while XG-Boost (98.86% accuracy) excels in precise disaster classification, making it the top performer.



OVERVIEW

The exploration of Principal Component Analysis (PCA) further contributed to the analysis by reducing the dimensionality of numerical data, facilitating more manageable interpretation and analysis.

Visual representations, such as bar graphs depicting the frequency of death, missing, and injured occurrences across different events and bar graphs illustrating the number of events in each decade, provided clear insights into the distribution and trends in the dataset.

Additionally, pie and donut charts were employed to visually convey the distribution of infrastructure damage, deaths, and missing persons across continents (Asia and Africa). The analysis also delved into the effectiveness of multiple linear regression and XG Boost models for predicting losses in USD and deaths, with XG Boost outperforming in both cases.

The correlation heatmap offered a comprehensive overview of relationships among variables, and the linear regression model further explored the connection between injuries and losses in USD.

The classification algorithms, namely Random Forest, K-Means, and XG Boost, demonstrated their unique strengths in predicting country and event types, showcasing the versatility of these models in addressing different aspects of disaster classification.

In summary, this analysis presents a robust and multi-faceted exploration of disaster data, utilizing various statistical and machine learning techniques to extract meaningful insights and patterns. The findings provide a foundation for informed decision-making and further research in disaster management and mitigation strategies.

THANK YOU