Volcanic Eruption Analysis: From History to Insight

Application of the KDD Process (Knowledge Discovery in Databases)

Course: Data Science Programming

Leader: Isabela MORA | Members: Mathieu MAURY, Paul MILLIEN, Yannix MICHOUX

Data Source: NOAA Volcanic Eruptions Database and external set World Population Data (2022) from Kaggle

Problem Statement & Objectives

Problem Statement

How can we transform a significant volume of historical and heterogeneous data on volcanic eruptions into actionable insights for risk prevention and understanding?

KDD Objective

Implement a modular processing pipeline to construct an analytical dashboard based on four key indicators.

Key Methodologies

Temporal Trend Analysis

(Regression Models)

Spatial Clustering

(K-Means)

Data Discretization & Enrichment

(Feature Engineering)

Phase I: Data Acquisition & Cleaning

1

Modular Functions

Our initial phase involved developing modular load_data and data_cleaning functions. This approach ensures reusability, simplifies debugging, and allows for easy integration into future workflows.

2

Specific Cleaning Procedures

Critical cleaning steps included managing missing values, particularly for geographical coordinates where imputation or strategic removal was necessary. We also standardized various units to ensure consistency across the dataset.

Historical Challenge: Data Inconsistency

We encountered a significant historical challenge: data prior to 1800 is notably incomplete and inconsistent, presenting a clear recording bias. To ensure the statistical robustness of our trend analyses, we applied a targeted temporal filter, while still retaining the full historical context for general understanding.

Indicator 1 & 2: Trends and Severity

Indicator 1: Temporal Trend

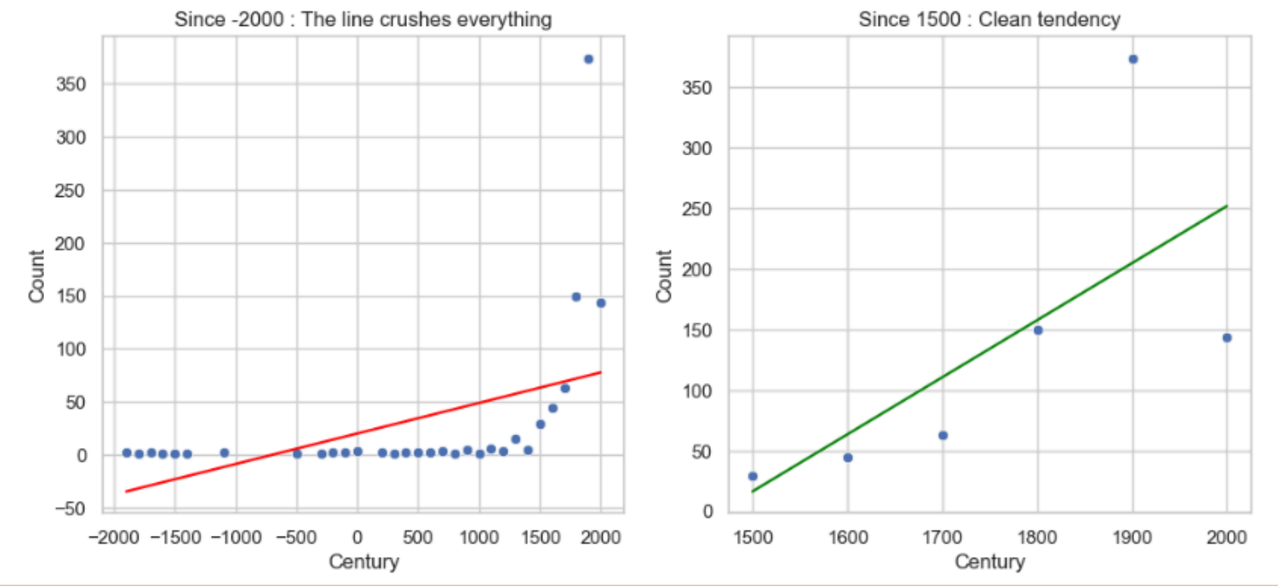
This indicator aims to determine if there is a discernible variation in the frequency or severity of volcanic eruptions over time. Understanding long-term patterns is crucial for forecasting and risk assessment.

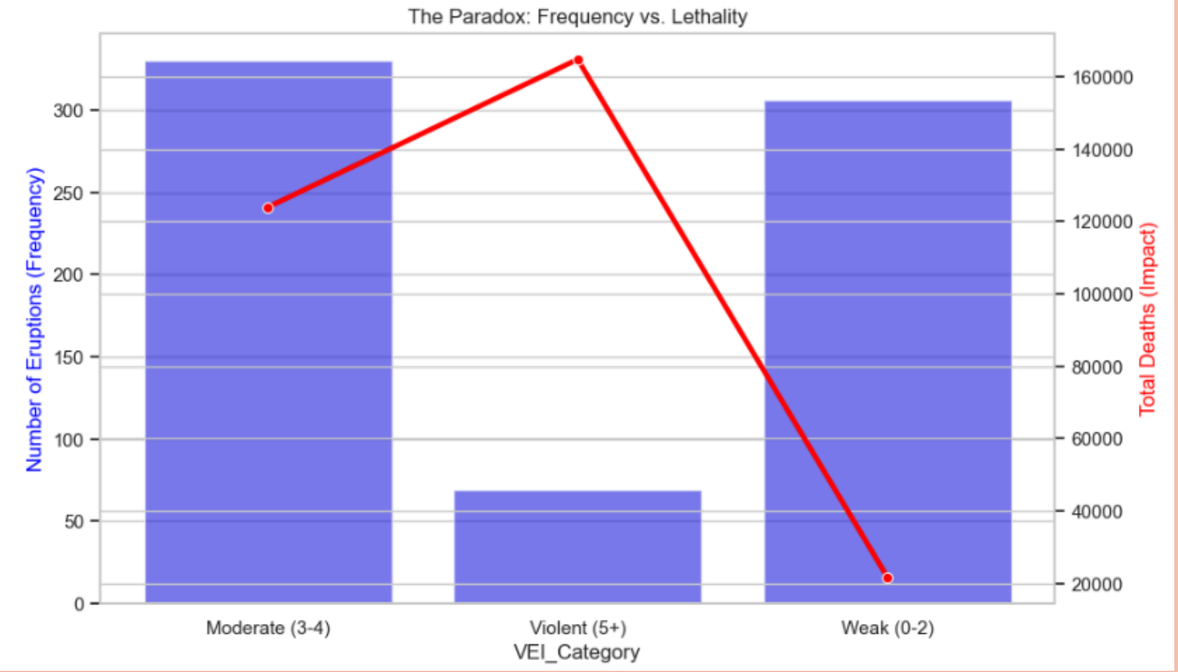
- **Method:** Application of a Linear Regression model on the chronological series of eruption events.
- **Insight:** Identify upward or downward trends in volcanic activity.

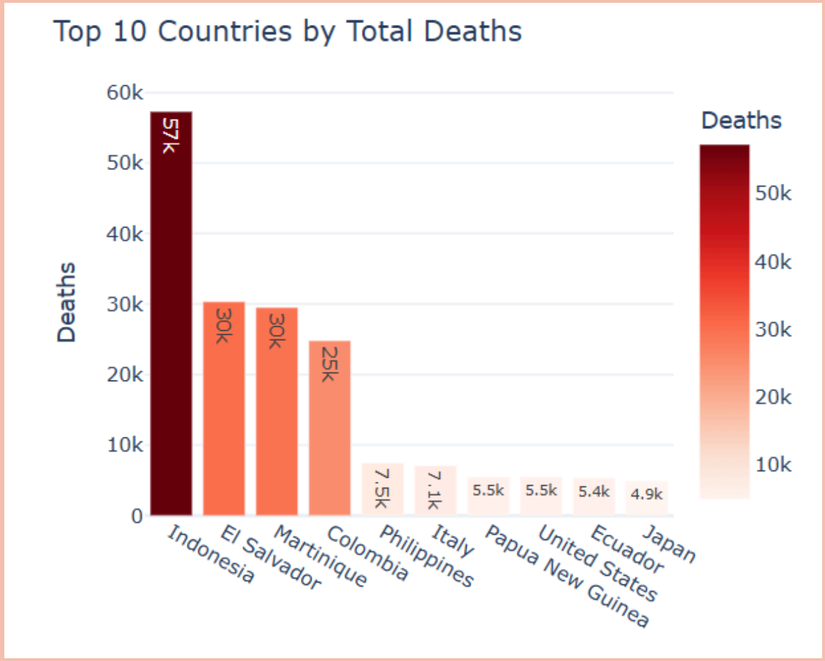
Indicator 2: VEI Distribution

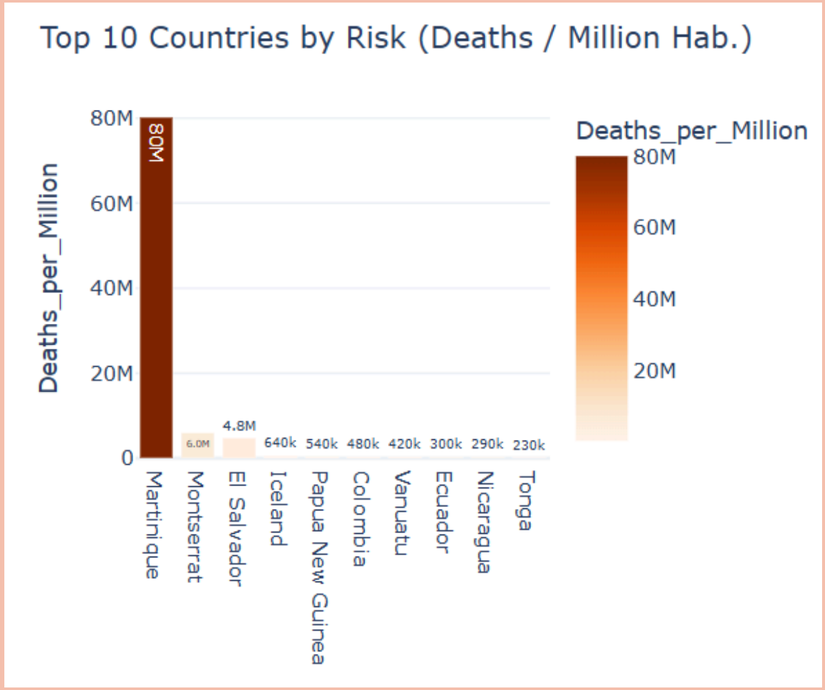
The Volcanic Explosivity Index (VEI) provides a relative measure of the explosiveness of volcanic eruptions. Analyzing its distribution helps us understand the typical magnitude of events.

- **Method:** Discretization of the raw VEI values into interpretable categories (Low/Moderate/High) for enhanced visual risk assessment.
- **Insight:** Characterize the proportion of different eruption magnitudes.









Indicator 3: Human Risk - Mortality Impact

Assessing the Direct Impact on Human Lives

Objective

Evaluate the direct impact of volcanic eruptions on human life by analyzing mortality rates associated with each event.

Method

Data Enrichment (external_data): Raw mortality data is cross-referenced with external datasets to contextualize the impact.

Beyond VEI: Incorporating Socio-Economic Vulnerability

To enrich our analysis, we didn't just cross-reference mortality with the VEI. We specifically integrated socio-economic data (development levels and population density of affected areas) to correlate mortality with local vulnerability. This goes beyond the mere strength of the eruption, providing a more nuanced understanding of human risk.

Indicator 4: Spatial Clustering for Risk Zones

Identifying Regions with Similar Activity Profiles

1

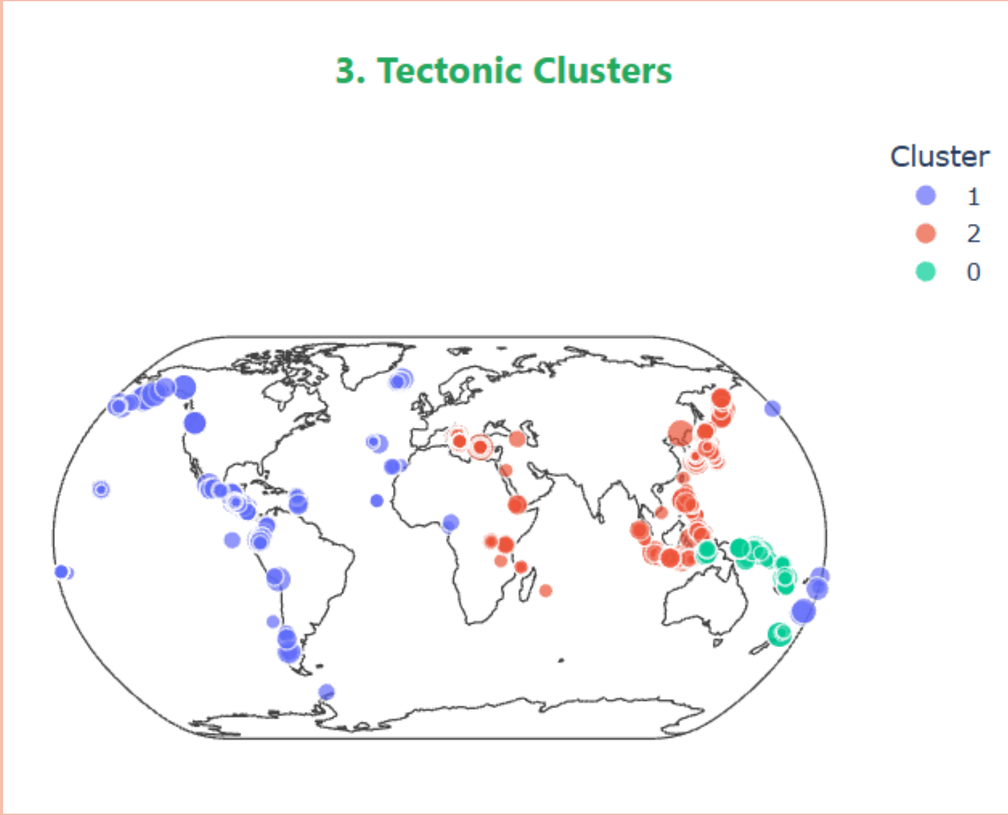
Objective

Group volcanoes based on their geographic coordinates and activity patterns to identify distinct risk regions.

2

Method

K-Means Clustering applied to latitude and longitude data. This algorithm effectively partitions data points into a predefined number of clusters.



Key Insight: The selection of the number of clusters (*k*) was not arbitrary. It was empirically optimized using the Elbow Method to ensure the best inertia and a significant separation of risk zones. This rigorous approach validates the distinctness and relevance of each identified cluster.

Architecture & Interactive Dashboard

1

Code Architecture

Our entire codebase is encapsulated within modular functions, promoting maintainability and testability. This clear separation of KDD stages ensures a robust and scalable solution.

2

Visualization Tool

The analytical dashboard is built using an interactive framework (e.g., Plotly Dash or Streamlit), designed for dynamic data exploration and insight generation.

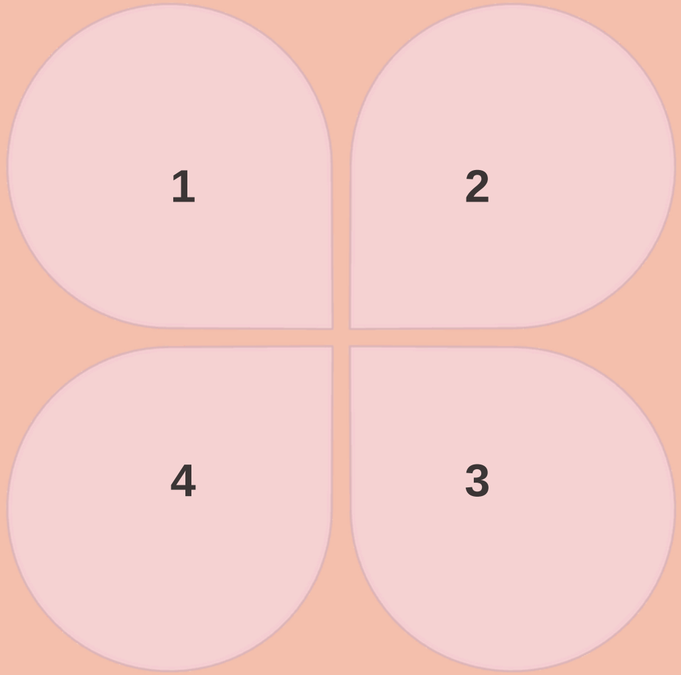
Dashboard Key Features

- **Interactive Filters:** Enable drill-down analysis by period, country, or VEI.
- **Synchronized Visualization:** Connects cluster maps (K-Means) and trend graphs (Regression) for coherent global analysis, offering a holistic view of volcanic activity.

Conclusion & Future Perspectives

Conclusion

This project effectively demonstrates the power of the KDD process in transforming raw data into a strategic decision-making tool for volcanic risk analysis.



Predictive Modeling

Implement Time Series models for forecasting eruption frequency.

Geological Integration

Cross-reference with tectonic plate data to contextualize global volcanic risk.

Advanced Clustering

Test DBSCAN algorithm for non-spherical and complex spatial clusters.

Thank you for your attention.

Questions & Answers