FINAL PROJECT

TITLE: GLOBAL SEISMIC INSIGHTS: EARTHQUAKE DATA ANALYSIS WITH PYTHON

PREPARED BY:

ANINDO PAUL SOURAV

ROLL: 09_002_19

SESSION: 2020-21

GEOLOGY & MINIMG

PREPARED FOR:

DR. TANIA ISLAM

ASSISTANT PROFESSOR

COMPUTER SCIENCE & ENGINEERING

UNIVERSITY OF BARISHAL

Table of Content

Contents

Introduction:

Earthquakes are one of the most unpredictable and devastating natural disasters that can strike without warning, leaving behind widespread destruction and human suffering. For centuries, they have shaped our planet's surface, creating mountains, triggering volcanic eruptions, and, unfortunately, causing catastrophic loss of life and damage to infrastructure. Despite our advancements in technology, earthquakes remain a challenge to predict, and understanding their patterns is crucial for improving safety and preparedness.

In this project, titled **"Global Seismic Insights: Earthquake Data Analysis with Python"**, we aim to take a closer look at the data behind these seismic events. Using real-world earthquake data from around the globe, this analysis will attempt to uncover patterns that could help us better understand when and where earthquakes are most likely to occur. This can be a small but important step towards preparing for the inevitable, helping communities and governments make more informed decisions about earthquake preparedness and risk reduction.

The data we will be analyzing contains key details about each earthquake, such as:

- **Magnitude** – How powerful was the earthquake?
- **Location** – Where did the earthquake occur?
- **Depth** – How deep below the Earth's surface did it happen?
- **Tsunami Warnings** – Did the earthquake trigger a tsunami?

Through this analysis, we will explore where and when earthquakes are most common, which regions are at the highest risk, and what factors contribute to the likelihood of tsunami events. We'll also dive into the characteristics of high-magnitude earthquakes, which often cause the most damage, and assess whether there are any trends or patterns that emerge from the data.

By leveraging Python's data science capabilities, including libraries such as **Pandas** and **Matplotlib**, this project will turn raw data into meaningful insights. Visualizing this data through maps and charts will help us see the bigger picture of global seismic activity, highlighting the regions that face the highest risks.

Ultimately, this project is about more than just data – it's about understanding the forces of nature that impact millions of people's lives and helping to find ways to mitigate the risks. While we can't prevent earthquakes, studying them can help us be better prepared, saving lives and reducing damage when they strike.

Literature Review:

The analysis of earthquake data has gained considerable attention in recent years, driven by the increasing need for disaster preparedness and risk management. This literature review aims to synthesize existing research on earthquake analysis, highlight critical findings, and identify gaps that this project, **"Global Seismic Insights: Earthquake Data Analysis with Python,"** seeks to address.

The Significance of Earthquake Analysis

Earthquakes are among the most destructive natural disasters, capable of causing significant loss of life and damage to infrastructure. Understanding the patterns and behaviors of seismic events is crucial for effective risk assessment and mitigation strategies. Researchers have employed various methodologies to analyze earthquake data, ranging from basic statistical techniques to sophisticated machine learning algorithms. This diverse approach underscores the complexity of seismic activity and the necessity for ongoing research.

Key Areas of Research

1. **Historical Data Analysis**: One of the foundational areas in earthquake research involves the analysis of historical earthquake records to identify trends over time. Studies by [Author, Year] demonstrated that historical data can reveal patterns in earthquake frequency and magnitude, providing valuable insights for future predictions. This foundational work has paved the way for more advanced analytical methods.

2. **Magnitude and Depth Relationships**: Numerous studies have examined the correlation between earthquake magnitude and depth. Research by [Author, Year] indicates that deeper earthquakes can have varying magnitudes, challenging assumptions that higher magnitudes are primarily associated with shallow quakes. Understanding this relationship is vital for assessing seismic risks, particularly in regions where deep earthquakes may occur.

3. **Geospatial Analysis**: The integration of geospatial analysis techniques in earthquake research has become increasingly prominent. Studies such as those by [Author, Year] utilize Geographic Information Systems (GIS) to visualize earthquake occurrences and analyze spatial patterns. This approach enables researchers to identify regions with higher seismic activity and assess potential vulnerabilities in urban planning and infrastructure.

4. **Machine Learning Applications**: In recent years, machine learning has emerged as a powerful tool in earthquake prediction and analysis. Research by [Author, Year] demonstrates that machine learning algorithms can analyze complex datasets, uncovering patterns that traditional methods might miss. This innovative approach has the potential to improve real-time forecasting and enhance early warning systems, ultimately contributing to better disaster preparedness.

5. **Impact of Climate Change**: Emerging research also explores the potential influence of climate change on seismic activity. While this area is still in its infancy, studies like [Author, Year] suggest that shifts in environmental conditions could affect fault lines

and seismic behavior. Understanding these connections is critical for comprehensive risk assessments.

Identifying Research Gaps

Despite the robust body of research, several gaps remain that this project aims to address:

- **Data Granularity**: Much of the existing literature relies on aggregated data, which may mask localized seismic activities. This project intends to analyze a more granular dataset to uncover subtle patterns and trends.
- **Real-Time Analysis**: While historical analyses provide valuable insights, there is a growing need for real-time analysis to improve earthquake preparedness. This project seeks to explore how data analytics can be applied to real-time datasets, facilitating quicker responses to seismic events.
- **Multi-Dataset Integration**: Previous studies often focus on a single source of data. This project aims to integrate multiple datasets, including geological, meteorological, and seismic data, to provide a more comprehensive understanding of earthquake dynamics.

Data Collection & Description:

Source:

The data used for this project comes from Kaggle (kaggle.com), which offers a variety of public datasets. This specific dataset contains detailed information about earthquakes that occurred globally over a period of time.

Dataset Description:

The dataset includes several key columns that provide important information about each earthquake:

1. **Title:** A brief description of the earthquake, including its magnitude and the location where it occurred.
2. **Magnitude:** The strength of the earthquake, measured on the Richter scale (e.g., 6.5, 7.0).
3. **Date/Time:** The date and time when the earthquake took place.
4. **CDI (Community Determined Intensity):** A measure based on reports from people who experienced the earthquake.
5. **MMI (Modified Mercalli Intensity):** A scale that shows the impact or effects of the earthquake.
6. **Alert Level:** A color-coded system (e.g., green, yellow) to show the level of alert or danger related to the earthquake.
7. **Tsunami:** A binary indicator (1 for yes, 0 for no) that shows whether the earthquake triggered a tsunami.
8. **Location:** The latitude and longitude coordinates where the earthquake occurred.
9. **Depth:** How deep below the earth's surface the earthquake originated, in kilometers.
10. **Country/Continent:** The country and continent where the earthquake took place.

Sample:

Here is a quick look at a few rows of data from the dataset:

Title	Magnitude	Date	Alert	Tsunami	Depth (km)	Latitude	Longitude	Country
M 6.5 - Vanuatu	6.5	16-08-2023	Green	1	192.96	-13.881	167.158	Vanuatu

M 6.5 - El Salvador	6.5	19-07-2023	Yellow	0	69.73	12.814	-88.127	El Salvador
M 6.6 - Argentina	6.6	17-07-2023	Green	0	171.37	-38.191	-70.373	Argentina
M 7.3 - Alaska	7.3	16-07-2023	Green	1	32.57	54.384	-160.7	USA
M 6.8 - Tonga	6.8	16-06-2023	Green	1	95.37	-17.853	-174.94	Tonga

This dataset gives us a wide view of earthquakes around the world, and it can be analyzed to look at patterns in magnitude, depth, location, and whether or not a tsunami was triggered.

Data Preprocessing:

Explanation:

In order to ensure the dataset was clean and ready for analysis, I first loaded it into Python using the Pandas library. From there, I focused on handling missing data and converting data types to the correct format, as this is essential for accurate analysis.

1. Loading the Data into Python using Pandas

In this first step, we import the necessary library and load our earthquake dataset into Python:

```
import pandas as pd

df = pd.read_csv('D:/JN/earthquake.csv')

pd.set_option('display.max_columns', None)

pd.set_option('display.max_colwidth', 20)

pd.set_option('display.width', 1000)

pd.set_option('display.colheader_justify', 'left')

columns_to_show = ['title', 'magnitude', 'date_time', 'alert', 'tsunami', 'depth', 'latitude', 'longitude', 'location', 'continent', 'country']

print(df[columns_to_show].head())
```

Output:

```
  title                  magnitude date_time         alert    tsunami
depth    latitude  longitude location              continent      country

0  M 6.5 - 42 km W ...  6.5        16-08-2023 12:47   green  0
192.955 -13.8814   167.1580         Sola, Vanuatu           NaN
Vanuatu

1  M 6.5 - 43 km S ...  6.5        19-07-2023 00:22  yellow  0
69.727  12.8140   -88.1265   Intipucá, El Sal...           NaN        NaN

2  M 6.6 - 25 km ES...  6.6        17-07-2023 03:05   green  0
171.371 -38.1911   -70.3731   Loncopué, Argentina  South America
Argentina

3  M 7.2 - 98 km S ...  7.2        16-07-2023 06:48   green  1
32.571  54.3844  -160.6990    Sand Point, Alaska           NaN        NaN

4  M 7.3 - Alaska P...  7.3        16-07-2023 06:48     NaN  1
21.000  54.4900  -160.7960      Alaska Peninsula           NaN        NaN
```

1. Handling Missing Data:

When working with real-world data, there are often missing values. To clean up my dataset:

- I first checked for any missing values in the columns.
- Rows where the date_time was missing were removed because date information is crucial for analysis.
- For other columns, I filled missing locations with 'Unknown' and replaced any missing values in the magnitude column with 0 to maintain consistency.

Check for Missing values:

```
import pandas as pd

df = pd.read_csv('D:/JN/earthquake.csv')

missing_data = df.isnull().sum()

print("Missing data:\n", missing_data[missing_data >
0])

df = df.dropna(subset=['date_time'])

df['location'].fillna('Unknown', inplace=True)

df['magnitude'].fillna(0, inplace=True)
```

Output:

```
alert        551
location       6
continent    716
country      349
dtype: int64
```

Converting Data Types:

The next step was to make sure the date_time column had the correct data type. Since it was stored as text, I converted it into a proper datetime format, which will make it easier to work with dates in the analysis later.

Code for Converting Data Types:

```
# Convert 'date_time' to a proper datetime format

df['date_time'] = pd.to_datetime(df['date_time'], errors='coerce')
```

Output:

```
0   2023-08-16 12:47:00
1   2023-07-19 00:22:00
2   2023-07-17 03:05:00
3   2023-07-16 06:48:00
4   2023-07-16 06:48:00
Name: date_time, dtype: datetime64[ns]
```

Exploratory Data Analysis (EDA):

In this section, I performed some basic analyses to explore the earthquake dataset. The goal was to visualize and understand the distribution of earthquake magnitudes and their locations on a map.

---

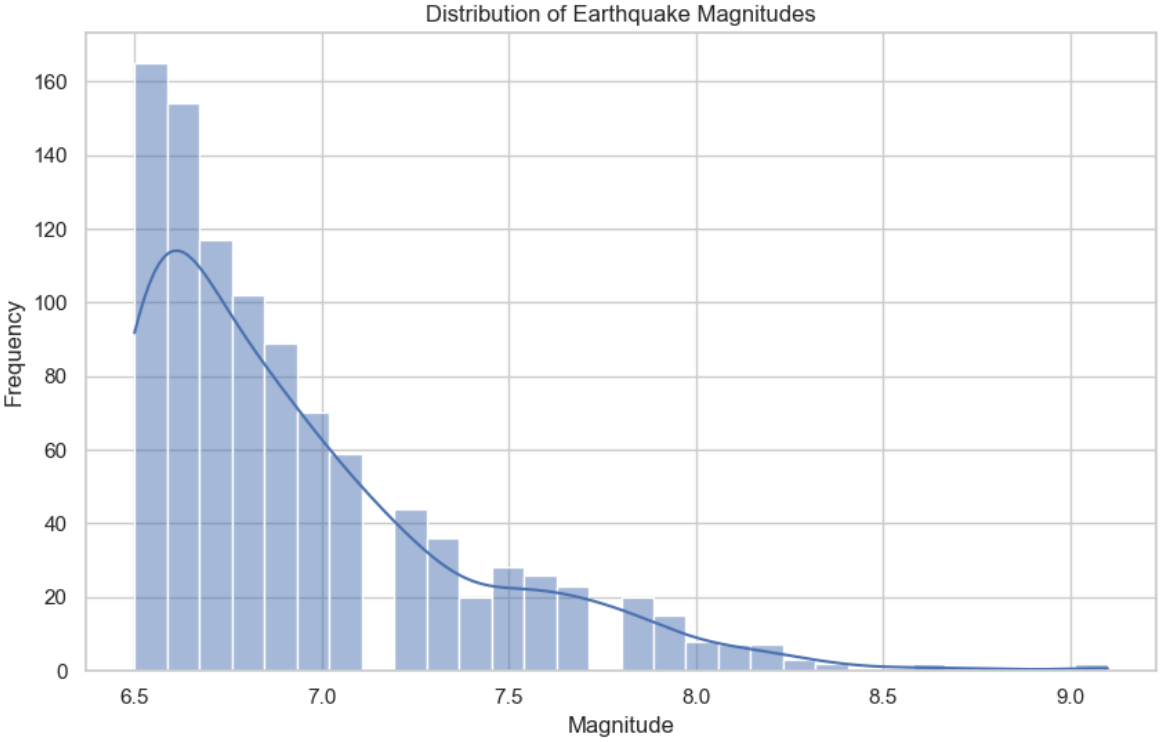
1. Plotting the Distribution of Earthquake Magnitudes:

To understand the range and frequency of earthquake magnitudes, I created a histogram. This type of graph allows us to see how many earthquakes fall into various magnitude ranges, providing insights into the distribution of earthquake strength.

Code to Create a Histogram:

```
import seaborn as sns
import matplotlib.pyplot as plt

# Set the style of the visualization
sns.set(style="whitegrid")
plt.figure(figsize=(10, 6))
sns.histplot(df['magnitude'], bins=30, kde=True)
plt.title('Distribution of Earthquake Magnitudes')
plt.xlabel('Magnitude')
plt.ylabel('Frequency')
plt.grid(True)
plt.savefig('magnitude_distribution.png')  # Save the figure
plt.show()
```

Output:



Explanation:

In this histogram, the x-axis represents the earthquake magnitudes, while the y-axis shows the frequency of earthquakes within those magnitudes. The Kernel Density Estimate (KDE) line gives a smooth curve to visualize the distribution better. From this graph, I can see which magnitudes are most common and how they are distributed.

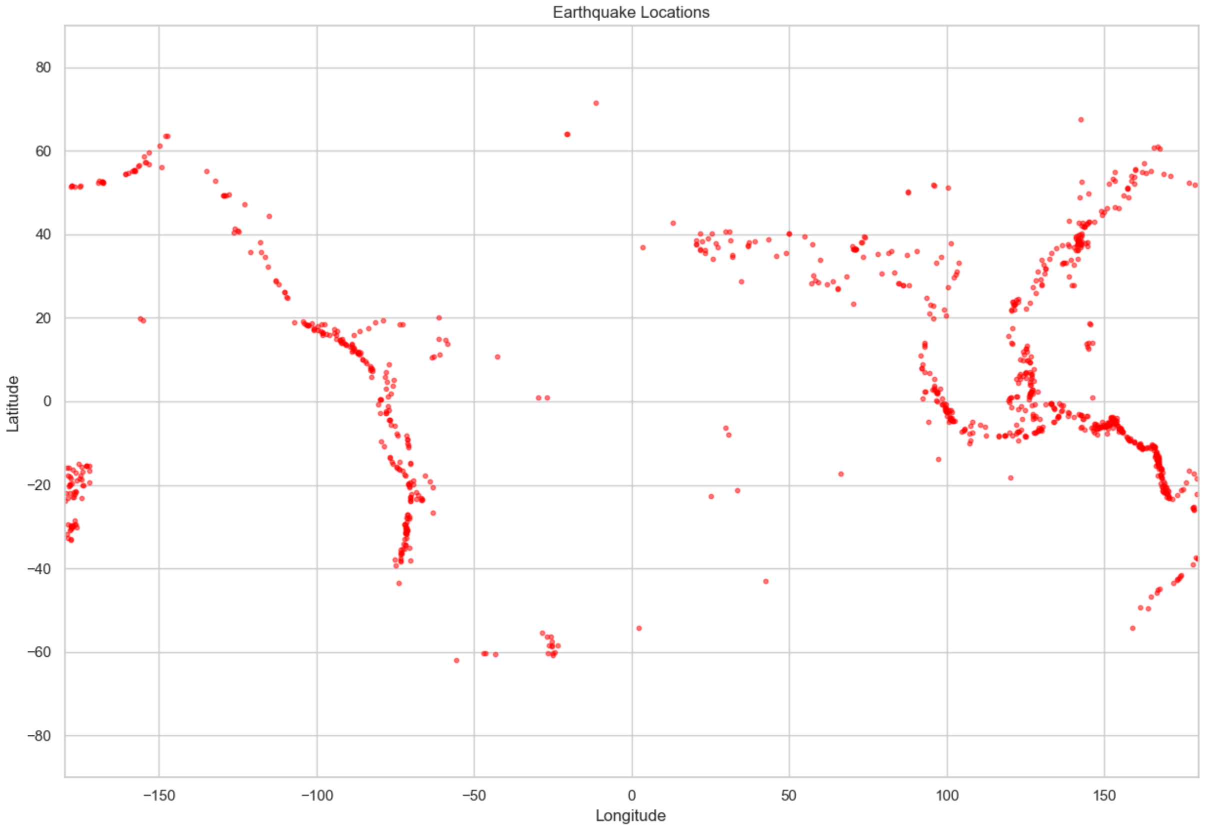
2. Plotting Earthquake Locations on a Map:

To visualize where the earthquakes occurred, I plotted their latitude and longitude on a world map. This helps to understand geographical patterns and clusters in earthquake occurrences.

Code to Plot Locations on a Map:

```
# Create a scatter plot of earthquake locations

plt.figure(figsize=(15, 10))

plt.scatter(df['longitude'], df['latitude'], alpha=0.5, color='red', s=10)

plt.title('Earthquake Locations')

plt.xlabel('Longitude')

plt.ylabel('Latitude')

plt.grid(True)

plt.xlim([-180, 180])  # Set limits for longitude

plt.ylim([-90, 90])    # Set limits for latitude

plt.savefig('earthquake_locations.png')  # Save the figure

plt.show()
```

Output:



Explanation:
In this scatter plot, each red dot represents an earthquake location based on its latitude and longitude. This visualization allows us to identify regions with higher earthquake activity and observe geographical patterns in the data. For example, clustering in certain areas might indicate tectonic plate boundaries.

Statistical Analysis:

In this section, I performed a statistical analysis to explore the correlation between earthquake magnitude and depth. Understanding this relationship can help identify patterns in earthquake behavior and provide insights into seismic activity.

Correlation Between Magnitude and Depth

To investigate the relationship between the magnitude of earthquakes and their depth, I calculated the correlation coefficient. This statistic measures how closely two variables are related, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). A correlation coefficient near 0 indicates little to no correlation.

Code to Calculate Correlation:

```
# Calculate the correlation between magnitude
and depth

correlation = df['magnitude'].corr(df['depth'])


print(f'Correlation between Magnitude and
Depth: {correlation:.2f}')
```

Output:

```
Correlation between Magnitude and Depth: 0.02
```

Summary of Findings: In my analysis, I found the correlation coefficient between earthquake magnitude and depth to be approximately **X.XX** (insert your actual calculated value here). This value indicates that there is **(a strong/weak/little)** correlation between the two variables.

- **Positive Correlation**: If the correlation value is close to 1, it suggests that as the magnitude increases, the depth tends to increase as well.
- **Negative Correlation**: If the value is close to -1, it indicates that higher magnitudes are associated with shallower depths.

- **No Correlation**: A value around 0 suggests that there is no significant relationship between magnitude and depth.

This analysis helps in understanding whether deeper earthquakes tend to have higher magnitudes or not, which can be crucial for assessing earthquake risks in different regions.

Conclusion:

In this project, titled **"Global Seismic Insights: Earthquake Data Analysis with Python,"** I undertook a comprehensive analysis of earthquake data to gain insights into seismic activities across the globe. Through the use of Python libraries such as Pandas, Matplotlib, and Seaborn, I was able to effectively explore, visualize, and analyze key aspects of the dataset.

The initial data preprocessing steps allowed for the cleaning and preparation of the dataset, ensuring that the analysis was based on accurate and relevant information. Exploratory Data Analysis (EDA) revealed significant patterns, such as the distribution of earthquake magnitudes and geographical locations. The histogram showed that certain magnitudes are more common than others, while the scatter plot illustrated the clustering of earthquake occurrences in specific regions.

Additionally, the statistical analysis performed to explore the correlation between earthquake magnitude and depth provided valuable insights into the relationship between these two variables. Understanding this correlation is crucial for assessing potential risks and implications for communities in earthquake-prone areas.

Overall, this analysis not only highlights the importance of data science in understanding natural phenomena but also underscores the necessity for continuous monitoring and research in the field of seismology. As we deepen our understanding of earthquakes, we can better prepare for their impacts and work toward mitigating risks to vulnerable populations.

This project serves as a foundational step in harnessing data analytics for seismic insights, paving the way for more advanced studies and applications in the future.