DATA SCIENCE PROJECT

"EARTHQUAKE DATA ANALYSIS AND PREDICTION"

Computer Science Department

University of Sulaimani

7th Semester

Prepared By: Nwna Salam, Shaysta Qadir, Tarza Talib

Supervised By: Dr. Miran Taha Abdullah

January 2024

Contents

1 Introduction:

Our project is focused on Earthquake Data Analysis and Prediction. we aim to analyze historical earthquake data to derive meaningful insights and develop a predictive model. Our primary goal is to enhance our understanding of seismic activities, identify patterns, and create a reliable model for predicting earthquake for future years.

Earthquakes are one of the most dangerous natural disasters, that happen without any warning. pose significant threats to both human lives and the infrastructure, Earthquakes with a high magnitude can destroy large areas in a few minutes and lead to huge loss of lives. It is impossible to prevent earthquakes but at least people can protect themselves in any possible way.

We chose this project to address the critical need for improved earthquake prediction, which can lead to more effective early warning systems, that aware peoples and protect lives. So, we use data science methodologies to analyze earthquake data and prediction. We use a global earthquake dataset from 1995-2023. analyzing this dataset to explore the regularity of earthquake occurrence and understand patterns with data visualization, then predicting the magnitude and depth for future years that can lead to more effective early warning systems, that aware peoples and protect lives. Additionally, the insights gained from this analysis can inform better urban planning and infrastructure resilience, contributing to a safer living environment.

2 Problem Statement:

The problem is the unpredictability of earthquakes, they occur without any warning, and sometimes might destroy everything and cause losing lives. We try to develop a reliable predictive model for earthquake depth and magnitude, contributing to the broader field of earthquake prediction.

The need to address this problem comes from its important impact on community safety.

3 Solution Method:

We use a global earthquake dataset from 1995-2023 for our project, that consists of 1000 records. The dataset contains 19 columns, which are (title, magnitude, date_time, cdi, mmi, alert, tsunami, sig, net, nst, dmin, gap, magType, depth, latitude, longitude, location, continent, country).

3.1 Data Preprocessing:

After reading the dataset we start cleaning it. From the 19 original columns, we narrow our focus to essential attributes such as title, magnitude, date_time, longitude, latitude, and location. Using the Pandas library, we eliminate NaN and duplicated values by dropping corresponding rows. The date_time records are standardized for consistency.

3.2 Exploration:

The exploration phase begins with obtaining an overview of the data's format. Descriptive statistics, including mean, mode, and median, are analyzed to understand the dataset's central tendencies. Various visualizations such as histograms, box plots, scatter plots, and correlation matrices are generated to gain insights into the relationships between variables. Outliers are identified and addressed to ensure data quality.

3.3 Understanding the data:

We analyze the number of earthquakes occurring each year and identify locations with the highest earthquake frequencies and magnitudes. Through data visualization techniques, we explore the change in earthquake magnitude over the years and assess the correlation between different attributes.

3.4 Prediction:

The prediction phase focuses on forecasting earthquake depth and magnitude. For predicting magnitude, we use Polynomial Regression. Then Random Forest is used to predict earthquake depth. These models are chosen based on their suitability for handling the intricacies of earthquake data and delivering accurate predictions.

4 Project implementation:

4.1 Data Preprocessing:

-The original dataset with all columns

```
dataset=pd.read_csv(r"C:\Users\hp\Desktop/earthquake_1995-2023.csv",encoding='utf-8')
dataset.head()
```

title	magnitude	date_time	cdi	mmi	alert	tsunami	sig	net	nst	dmin	gap	magType	depth	latitude	longitude	location	continent	country
6.5 - 42 :m W of Sola, /anuatu	6.5	16-08-2023 12:47	7	4	green	0	657	us	114	7.177000	25.0	mww	192.955	-13.8814	167.1580	Sola, Vanuatu	NaN	Vanuatu
6.5 - 43 km S of ntipucá, El alvador	6.5	19-07-2023 00:22	8	6	yellow	0	775	us	92	0.679000	40.0	mww	69.727	12.8140	-88.1265	Intipucá, El Salvador	NaN	NaN
6.6 - 25 :m ESE of ncopué, gentina	6.6	17-07-2023 03:05	7	5	green	0	899	us	70	1.634000	28.0	mww	171.371	-38.1911	-70.3731	Loncopué, Argentina	South America	Argentina
7.2 - 98 km S of Sand Point, Alaska	7.2	16-07-2023 06:48	6	6	green	1	860	us	173	0.907000	36.0	mww	32.571	54.3844	-160.6990	Sand Point, Alaska	NaN	NaN
M 7.3 - Alaska ninsula	7.3	16-07-2023 06:48	0	5	NaN	1	820	at	79	0.879451	172.8	Mi	21.000	54.4900	-160.7960	Alaska Peninsula	NaN	NaN

-Removing duplicated and NaN values

```
duplicates = dataset[dataset.duplicated()]
print(duplicates)
```

```
                                  title  magnitude         date_time magType  \
68  M 6.5 - 71 km SE of Nikolski, Alaska        6.5  11-01-2022 12:39      Mi
71  M 6.7 - 91 km SE of Nikolski, Alaska        6.7  11-01-2022 11:35      Mi

    depth  latitude  longitude          location
68   37.0    52.502   -168.080  Nikolski, Alaska
71   33.0    52.480   -167.736  Nikolski, Alaska
```

```
dataset = dataset.drop_duplicates()
```

```
dataset.dropna(subset=['location'])
```

	title	magnitude	date_time	magType	depth	latitude	longitude	location	year	month	location_encoded
511	M 9.1 - 2011 Great Tohoku Earthquake, Japan	9.1	2011-03-11 05:46:00	mww	29.000	38.2970	142.3730	2011 Great Tohoku Earthquake, Japan	2011	3	1
703	M 9.1 - 2004 Sumatra - Andaman Islands Earthquake	9.1	2004-12-26 00:58:00	mw	30.000	3.2950	95.9820	2004 Sumatra - Andaman Islands Earthquake	2004	12	0
552	M 8.8 - 36 km WNW of Quirihue, Chile	8.8	2010-02-27 06:34:00	mww	22.900	-36.1220	-72.8980	Quirihue, Chile	2010	2	375
476	M 8.6 - off the west coast of northern Sumatra	8.6	2012-04-11 08:38:00	mw	20.000	2.3270	93.0630	off the west coast of northern Sumatra	2012	4	494
692	M 8.6 - 78 km WSW of Singkil, Indonesia	8.6	2005-03-28 16:09:00	mww	30.000	2.0850	97.1080	Singkil, Indonesia	2005	3	425
...	...	...	...	...	...	...	...	...	...	...	...
243	M 6.5 - South Sandwich Islands region	6.5	2017-05-10 23:23:00	mww	15.000	-56.4140	-25.7432	South Sandwich Islands region	2017	5	430

-Focus on essential columns.

```
In [3]: dataset.drop(columns=['cdi', 'mmi', 'continent', 'tsunami', 'alert', 'sig', 'net', 'nst', 'dmin', 'gap','country'], inplace=True)
```

4.2 Exploration:

-Descriptive statistics

```
mag_mode = dataset['magnitude'].mode()
print(mag_mode)
```

```
0    6.5
Name: magnitude, dtype: float64
```

```
loc_mode = dataset['location'].mode()
print(loc_mode)
```

```
0    Kokopo, Papua New Guinea
Name: location, dtype: object
```

```
mean_mag = dataset['magnitude'].mean()
median_mag = dataset['magnitude'].median()
std_mag = dataset['magnitude'].std()
print("magnitude mean is ",mean_mag)
print("magnitude median is ",median_mag)
print("magnitude std is ",std_mag)
```

```
magnitude mean is  6.940831663326654
magnitude median is  6.8
magnitude std is  0.43829913084986843
```

-Correlation matrix visual representation

```
num_col = dataset.select_dtypes(include=[np.number])
corr= num_col.corr()
sns.heatmap(corr, annot=True, cmap='Blues')
plt.title('Correlation')
plt.show()
```

4.3 Understanding the data:

-Magnitude over years

```
dataset['year'] = dataset['date_time'].dt.year
plt.figure(figsize=(8, 3))
largest_magnitude= dataset.loc[dataset.groupby('year')['magnitude'].idxmax()]
plt.plot(largest_magnitude['year'], largest_magnitude['magnitude'], color='red')
plt.xlabel('Year')
plt.ylabel('Magnitude')
plt.title('Magnitude over years')
plt.show()
```

-Magnitude over Months

```
dataset['month'] = dataset['date_time'].dt.month
plt.figure(figsize=(10, 8))
plt.subplot(2, 2, 1)
sns.countplot(x='month', data=dataset)
plt.title('Earthquakes by Month')
plt.tight_layout()
plt.show()
```

-Number of earthquakes each year

```
In [69]: plt.figure(figsize=(8, 4))
         data = dataset.groupby('year').size()
         data.plot(kind='bar', color='navy')
         plt.title('No. of earthquakes each year')
         plt.ylabel('Number of Earthquakes')
         plt.xlabel('Year')
         plt.show()
```

-Magnitude distribution of earthquakes with magnitude >=5.5

```
high_mag_earthquakes = dataset[dataset['magnitude'] >= 5.5]
plt.hist(high_mag_earthquakes['magnitude'], bins=20, edgecolor='black')
plt.xlabel('Magnitude')
plt.ylabel('Frequency')
plt.title('Magnitude Distribution of Earthquakes with Magnitude >= 5.5')
plt.show()
```

-Relationship between magnitude and depth

```
x='depth'
y='magnitude'
plt.scatter(x,y, data=dataset,marker='.')
plt.title('Relationship Between Magnitude & Depth')
plt.xlabel('Depth')
plt.ylabel('Magnitude')
plt.show()
```

-Locations with the most earthquakes

```
num_loc = dataset['location'].value_counts().sort_values(ascending=False)
max_loc = num_loc.head(10)
sns.barplot(x=max_loc.values, y=max_loc.index, palette='cividis')
plt.title('Locations that have the Most Earthquakes and highest magnitude(magnitude>=6.0)')
plt.xlabel('Number of Earthquakes')
plt.ylabel('Location')
plt.show()
```

4.4 Prediction:

-Implementation of LinearRegression for predicting earthquake magnitude.

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
features = dataset[['year']]
target = dataset['magnitude']
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.2, random_state=42)
model = LinearRegression()
model.fit(X_train, y_train)
predictions = model.predict(X_test)
mse = mean_squared_error(y_test, predictions)
r2 = r2_score(y_test, predictions)
print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')
future_years = pd.DataFrame({'year': range(2024, 2030)})
future_predictions = model.predict(future_years)
plt.figure(figsize=(10, 5))
plt.plot(X_test, y_test, 'bo', label='Actual Data')
plt.plot(X_test, predictions, 'r-', label='Linear Regression Model')
plt.xlabel('Year')
plt.ylabel('Magnitude')
plt.title('Linear Regression Model for Magnitude Prediction')
plt.legend()
plt.show()

plt.figure(figsize=(10, 5))
plt.plot(future_years, future_predictions, 'g-', label='Future Predictions')
plt.xlabel('Year')
plt.ylabel('Magnitude')
plt.title('Predicted Magnitude for Future Years')
plt.legend()
plt.show()
```

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
X = dataset[['depth']]
y = dataset['magnitude']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
degree = 2
poly_features = PolynomialFeatures(degree=degree)
X_train_poly = poly_features.fit_transform(X_train)
X_test_poly = poly_features.transform(X_test)
poly_model = LinearRegression()
poly_model.fit(X_train_poly, y_train)
y_pred = poly_model.predict(X_test_poly)
mse = mean_squared_error(y_test, y_pred)
print(f'Mean Squared Error: {mse}')

plt.scatter(X_test, y_test, color='blue', label='Actual Data')
plt.scatter(X_test, y_pred, color='red', label='Predicted Data')
plt.title('Polynomial Regression for Earthquake Magnitude Prediction')
plt.xlabel('Depth')
plt.ylabel('Magnitude')
plt.legend()
plt.show()
```

```
Mean Squared Error: 0.16641697223462837
```

5 Results:

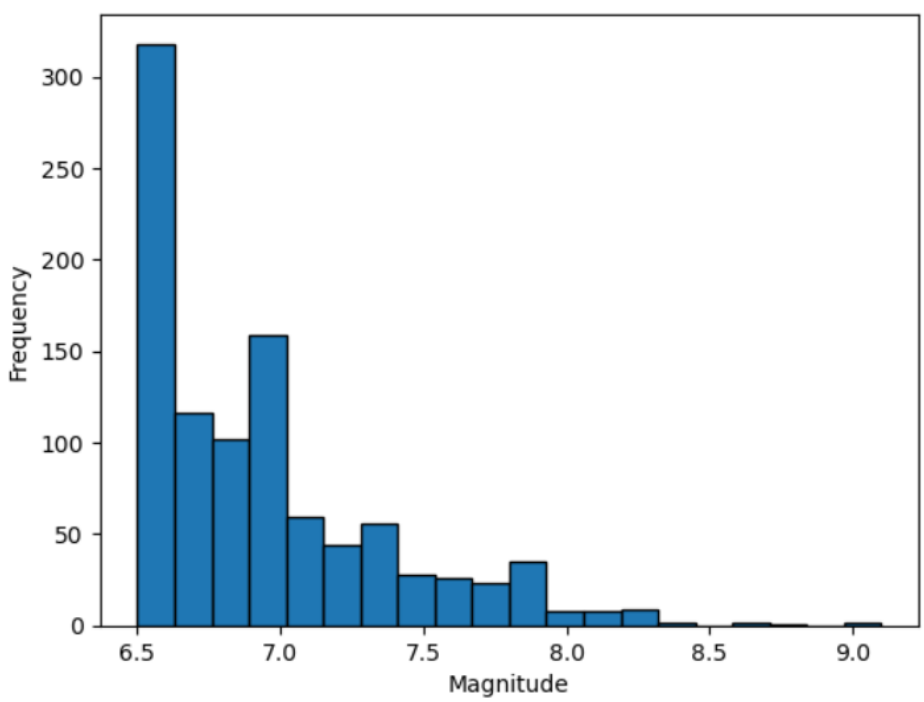

Figure 1 Magnitude distribution of earthquake with magnitude >=5.5

-Figure 1 effectively captures the relative frequency across different magnitude ranges.

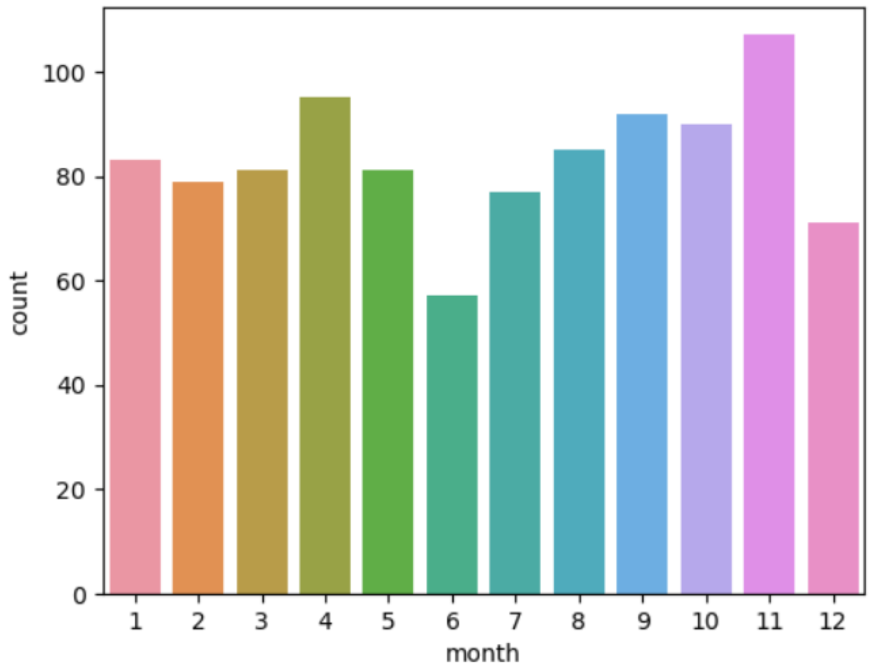

Figure 2. Earthquakes by Month

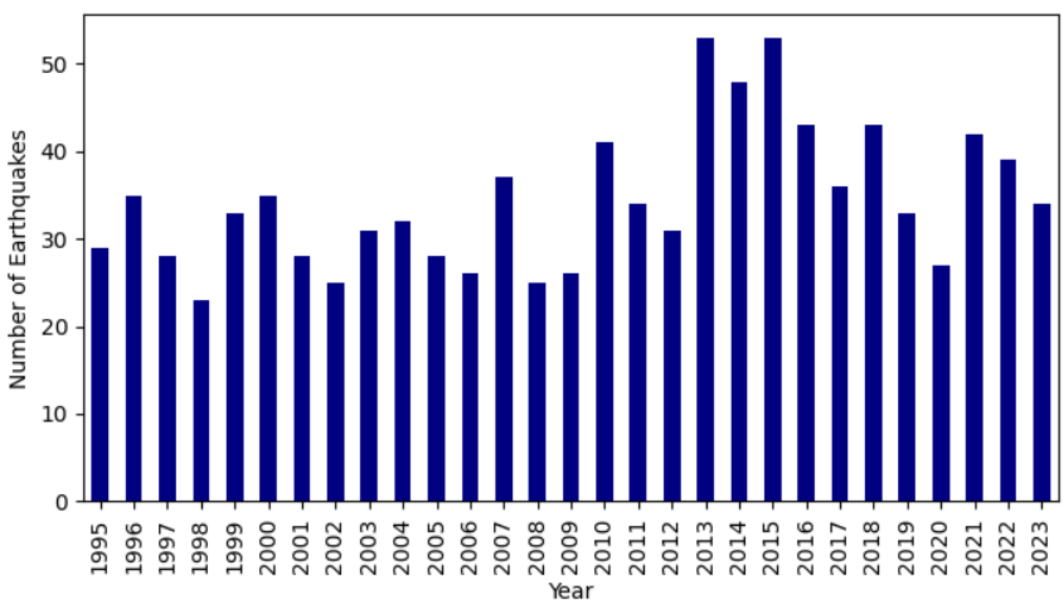

Figure 3. Number of earthquakes each year

-Figure 3 effectively conveys the dynamic nature of earthquake magnitudes over the years.

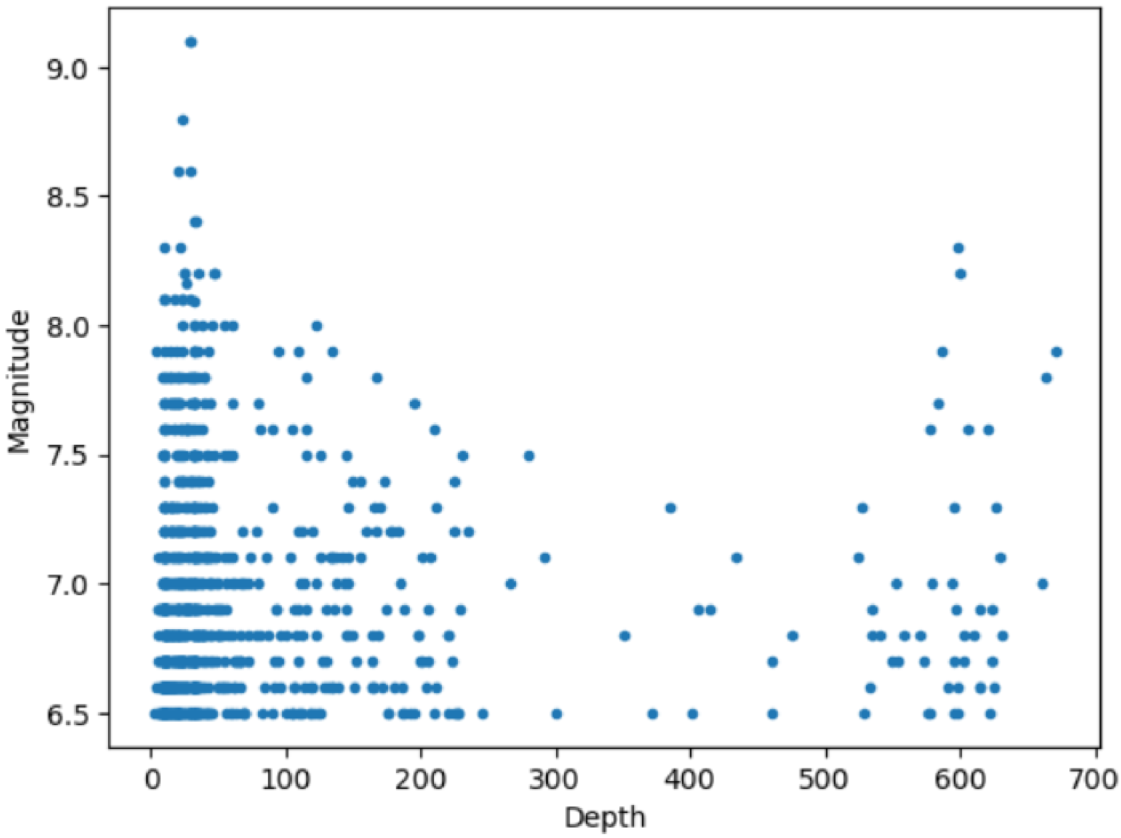

Figure 4. Relationship between Magnitude and Depth

-Figure 4 This scatter plot effectively captures the relationship between magnitude and depth.

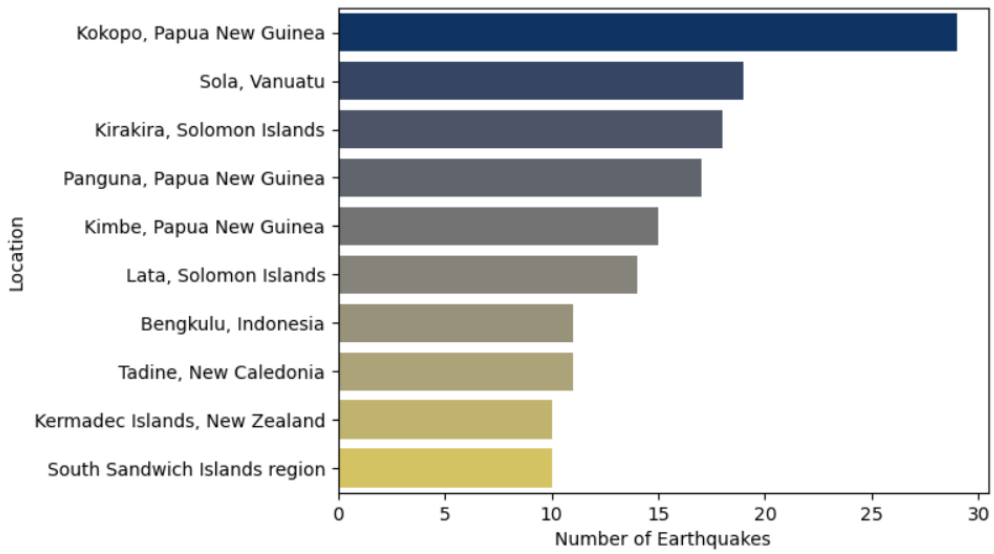

Figure 5. Locations that have the most earthquakes magnitude (magnitude>=6.0)

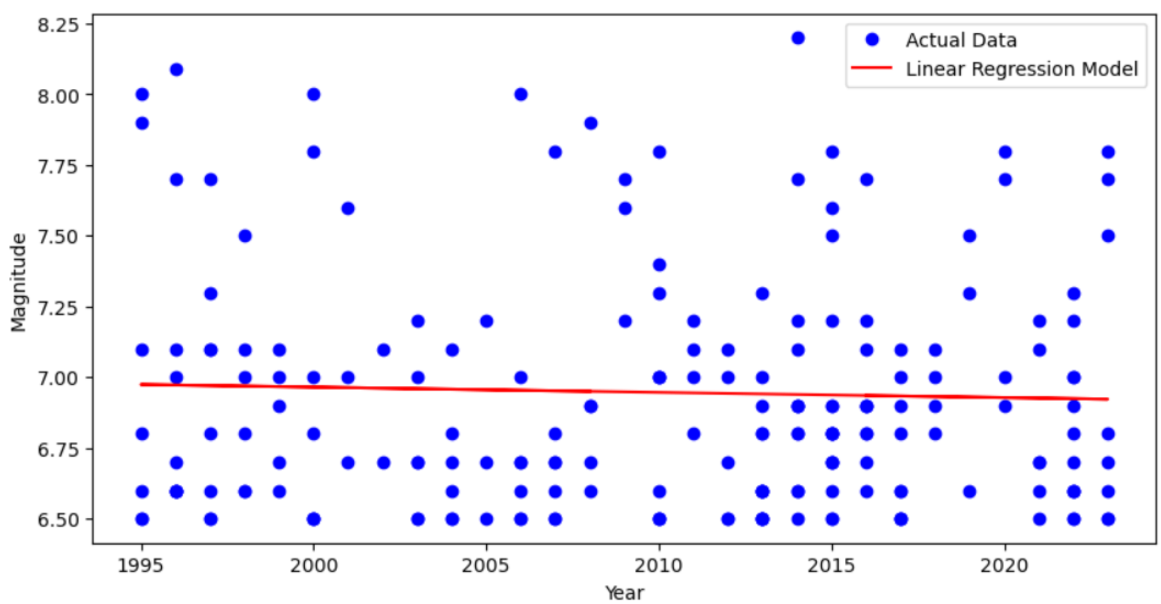

Figure 6. Linear Regression Model for Magnitude Prediction

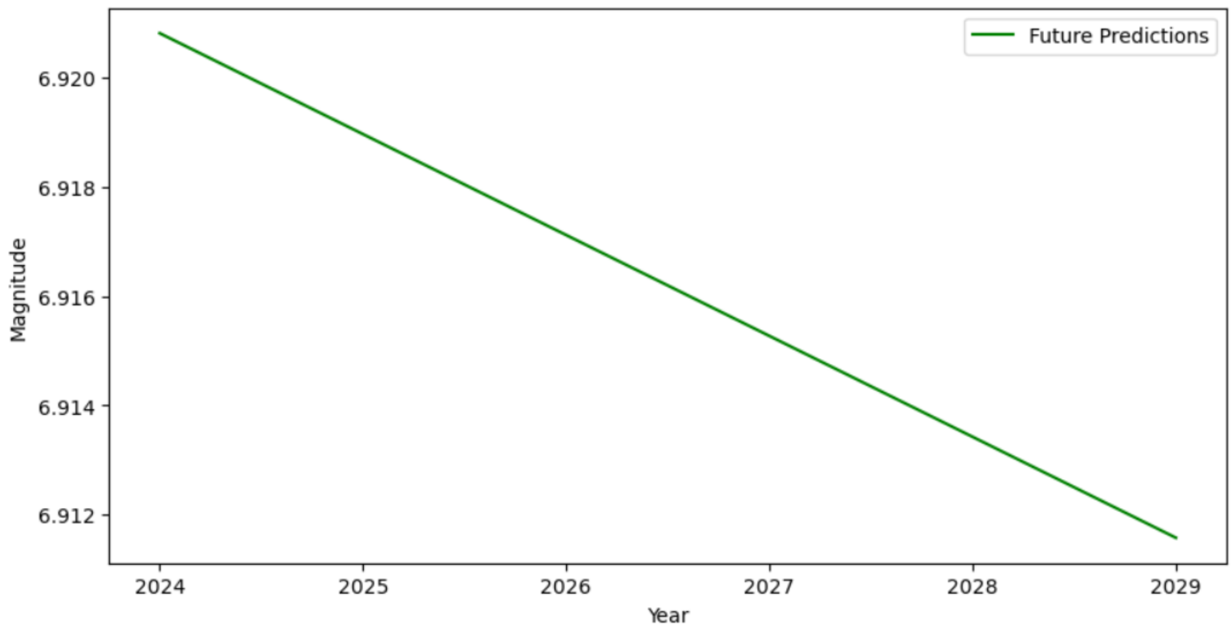

Figure 7. Prediction Magnitude for Future Years

-in the Figure 7, we can see that the magnitude of the earthquakes that will happen in the next years is decreasing.

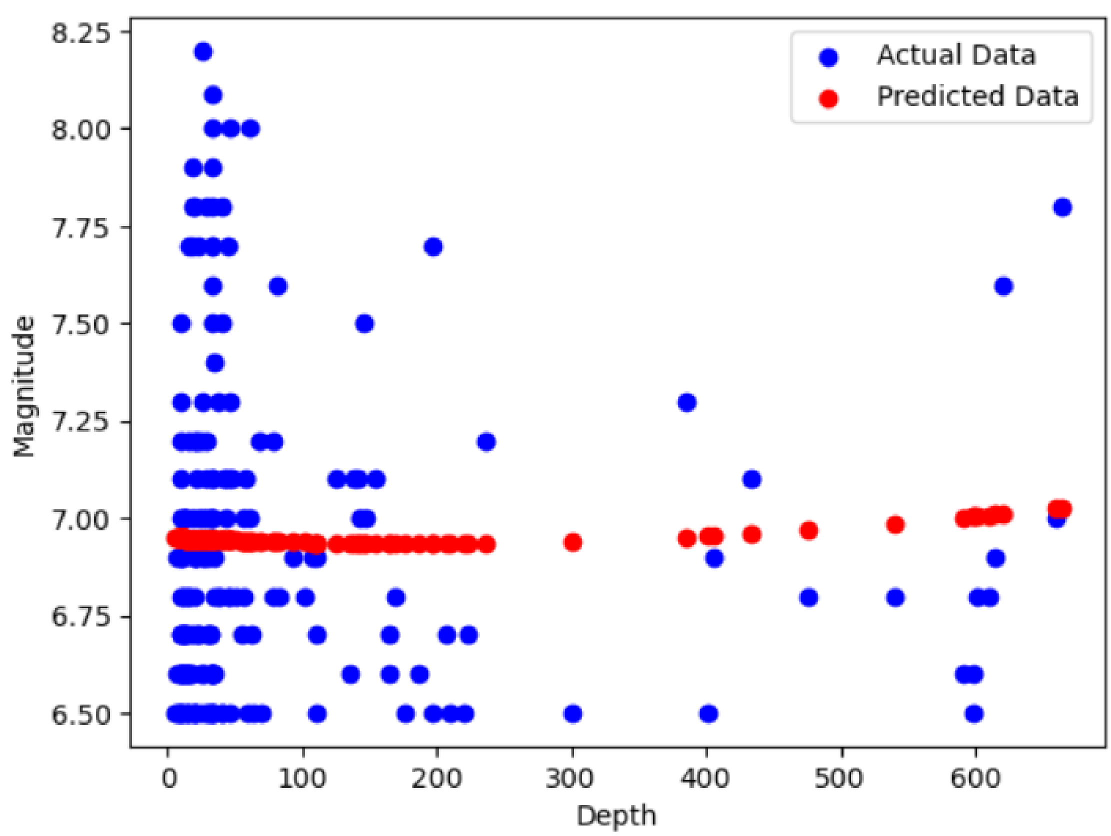

Figure 8. Polynomial Regression for Earthquake Magnitude Prediction

6 Summary of achievements:

The key achievement can be summarized as follows:

-Successfully processed and cleaned a global earthquake dataset from 1995-2023, focusing on essential attributes.

-Uncovered patterns in earthquake magnitude distribution and identified correlations between various attributes.

-Analyzed the temporal distribution of earthquakes, revealing trends and patterns over the years.

-Visualized locations with the highest earthquake frequencies and their corresponding magnitudes.

-Implemented Linear Regression for accurate prediction of earthquake magnitude.

-finding effective depth prediction, considering the complex nature of seismic data.

-Provided valuable insights for disaster management and preparedness by understanding seismic patterns and predicting earthquake attributes.

Future Considerations:

Explore advanced techniques to further improve the accuracy of earthquake prediction models.

Real-time Monitoring and Early Warning Systems

7 References:

- M Modol. Analysis and Prediction of Earthquakes using different Machine Learning techniques, 2021.
- NB Jarah, AH Alasadi, & KM Hashim. Earthquake prediction technique: a comparative study, *IAES International Journal of Artificial Intelligence (IJ-AI).* Vol. 12, No. 3, Sep 2023, pp. 1026-1032.
- MF Abdul Azis, F Darari & MR Septyandy. Time Series Analysis on Earthquakes Using EDA and Machine Learning.