ALA-TOO INTERNATIONAL UNIVERSITY

Earthquake Prediction Using Machine Learning Techniques

Author:
Bufatima Nurmuhammad kyzy

Supervisor:
Dr. Remudin Reshid MEKURIA

A thesis proposal submitted for the degree of MSc

in the

Data Science

May 23, 2024

Abstract

This study investigates the application of various machine learning techniques to predict earthquake magnitudes in Central Asia. By utilizing a comprehensive dataset containing seismic activity data, we explore the efficacy of models including Random Forest Regressor, Gradient Boosting Regressor, Linear Regression, and Support Vector Regressor. Our results demonstrate significant predictors and provide insights into the spatial and temporal patterns of seismic activities in the region. The findings suggest that machine learning models can significantly enhance earthquake prediction accuracy, which is crucial for disaster preparedness and mitigation efforts.

Declaration

I, Bufatima Nurmuhammad kyzy, declare that this thesis titled "Earthquake Prediction Using Machine Learning Techniques" ,and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

Acknowledgements

First of all, I wish to express my deepest and sincere gratitude to my thesis supervisor Dr.Remudin Reshid Mekuria who offered me brilliant ideas about this topic and provided invaluable guidance throughout this research.

I am thankful to International Alatoo University for their continued inspiration and encouragement.

I would like to express my special thanks to my teachers for their constant support and motivation.

My sincere thanks also goes to my family for their understanding, caring, and support to finish this research work.

In the end, I am very much thankful to all the people who have in different ways helped me to achieve best results in my work.

List of Abbreviations

ML	**M**achine **L**earning
EDA	**E**xploratory **D**ata **A**nalysis
MAE	**M**ean **A**bsolute **E**rror
MSE	**M**ean **S**quared **E**rror
SVM	**S**upport **V**ector **M**achines
RF	**R**andom **F**orest

Contents

Chapter 1

Introduction

1.1 Research Statement

Earthquakes pose a significant threat to life and infrastructure, especially in seismically active regions like Central Asia. Accurate prediction of earthquake magnitudes is crucial for minimizing damage and enhancing preparedness. This research aims to explore the potential of machine learning techniques in predicting earthquake magnitudes using seismic data from Central Asia.

1.2 Research Questions

1. How can machine learning techniques be applied to predict earthquake magnitudes using seismic data?
2. What are the most significant predictors of earthquake magnitude in Central Asia?
3. How do different machine learning models compare in terms of prediction accuracy for earthquake magnitudes?

1.3 Objectives

1. To collect and preprocess a comprehensive dataset of seismic activities in Central Asia.
2. To explore the application of various machine learning techniques for predicting earthquake magnitudes.
3. To identify the most significant predictors and evaluate the performance of different models.
4. To provide insights into the spatial and temporal patterns of seismic activities in the region.

1.4 Significance of the Study

This study is significant as it aims to enhance the accuracy of earthquake predictions using advanced machine learning techniques. Improved prediction models can contribute to disaster preparedness, potentially saving lives and reducing economic losses in earthquake-prone regions.

Chapter 2

Review of Related Literature

2.1 Traditional Earthquake Prediction Methods

Historically, earthquake prediction has relied on statistical analyses and physical models of seismic activity. These approaches often involve identifying patterns in historical data and understanding geological processes that lead to earthquakes. However, they have limitations in accuracy and reliability (Rundle et al., 2003).

2.1.1 Statistical Methods

Statistical methods have been used to identify patterns in earthquake occurrences. These methods include time-series analysis, regression models, and clustering techniques. Although they provide valuable insights, they often fall short in accurately predicting the timing and magnitude of earthquakes.

2.1.2 Physical Models

Physical models simulate the geological processes that lead to earthquakes. These models use data on tectonic plate movements, stress accumulation, and fault dynamics. While they enhance our understanding of seismic activities, their predictive power is limited by the complexity of Earth's crust and the multitude of factors involved.

2.2 Machine Learning Approaches

Recent advances in machine learning have opened new avenues for earthquake prediction. Researchers have employed techniques such as neural networks, support vector machines, and ensemble methods to predict seismic events. For instance, DeVries et al. (2018) utilized neural networks to predict aftershock locations, demonstrating the potential of machine learning in this domain. Similarly, Mousavi et al. (2020) applied deep learning to improve the detection of earthquake events, highlighting the effectiveness of these models in handling complex datasets.

2.2.1 Neural Networks

Neural networks, particularly deep learning models, have shown promise in capturing the non-linear relationships in seismic data. These models can learn from large datasets and

identify intricate patterns that are difficult to discern using traditional methods.

2.2.2 Support Vector Machines

Support Vector Machines (SVM) have been used to classify and predict seismic events. SVMs are effective in high-dimensional spaces and can handle non-linear relationships through kernel functions.

2.2.3 Ensemble Methods

Ensemble methods, such as Random Forest and Gradient Boosting, combine multiple models to improve prediction accuracy. These methods are robust against overfitting and can capture complex interactions between predictors.

2.3 Comparison of Methods

Studies comparing different machine learning models have shown that ensemble methods like Random Forest and Gradient Boosting often outperform traditional regression models in terms of prediction accuracy. These methods can handle non-linear relationships and interactions between predictors, making them well-suited for seismic data analysis (Breiman, 2001).

2.3.1 Performance Metrics

Performance metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) are commonly used to evaluate the effectiveness of prediction models. These metrics provide insights into the accuracy and reliability of the models.

2.3.2 Applications and Limitations

While machine learning models have demonstrated potential in earthquake prediction, they also have limitations. The quality and quantity of data, the choice of features, and the model parameters significantly impact their performance. Moreover, the inherent unpredictability of earthquakes poses a challenge to achieving high prediction accuracy.

Chapter 3

Methodology

3.1 Data Collection and Preprocessing

The dataset comprises 5,290 entries, detailing the date, time, latitude, longitude, depth, magnitude, and country of each earthquake. The data was obtained from a public earthquake database, ensuring a comprehensive and reliable source of seismic information.

Listing 3.1: Data Preprocessing Code

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestRegressor,
                             GradientBoostingRegressor
from sklearn.linear_model import LinearRegression
from sklearn.svm import SVR
from sklearn.metrics import mean_squared_error, mean_absolute_error,
                            r2_score

# Load and preprocess data
df = pd.read_csv("Central-Asian-earthquake-dataset.csv")
df['Date'] = pd.to_datetime(df['Date'])
df['Time'] = pd.to_datetime(df['Time'])
df['Year'] = df['Date'].dt.year
df['Month'] = df['Date'].dt.month
df['Hour'] = df['Time'].dt.hour
df['Season'] = df['Month'].apply(lambda x: 'Spring' if x in [3, 4, 5]
            else 'Summer' if x in [6, 7, 8]
            else 'Autumn' if x in [9, 10, 11]
            else 'Winter')
```

Preprocessing steps included handling missing values and converting date and time columns into suitable formats for analysis. These steps ensured that the dataset was clean and ready for modeling.

3.2 Exploratory Data Analysis

EDA revealed key insights into the distribution and frequency of earthquakes by magnitude, depth, and location. Visualizations included histograms, scatter plots, and heatmaps to identify patterns and correlations.

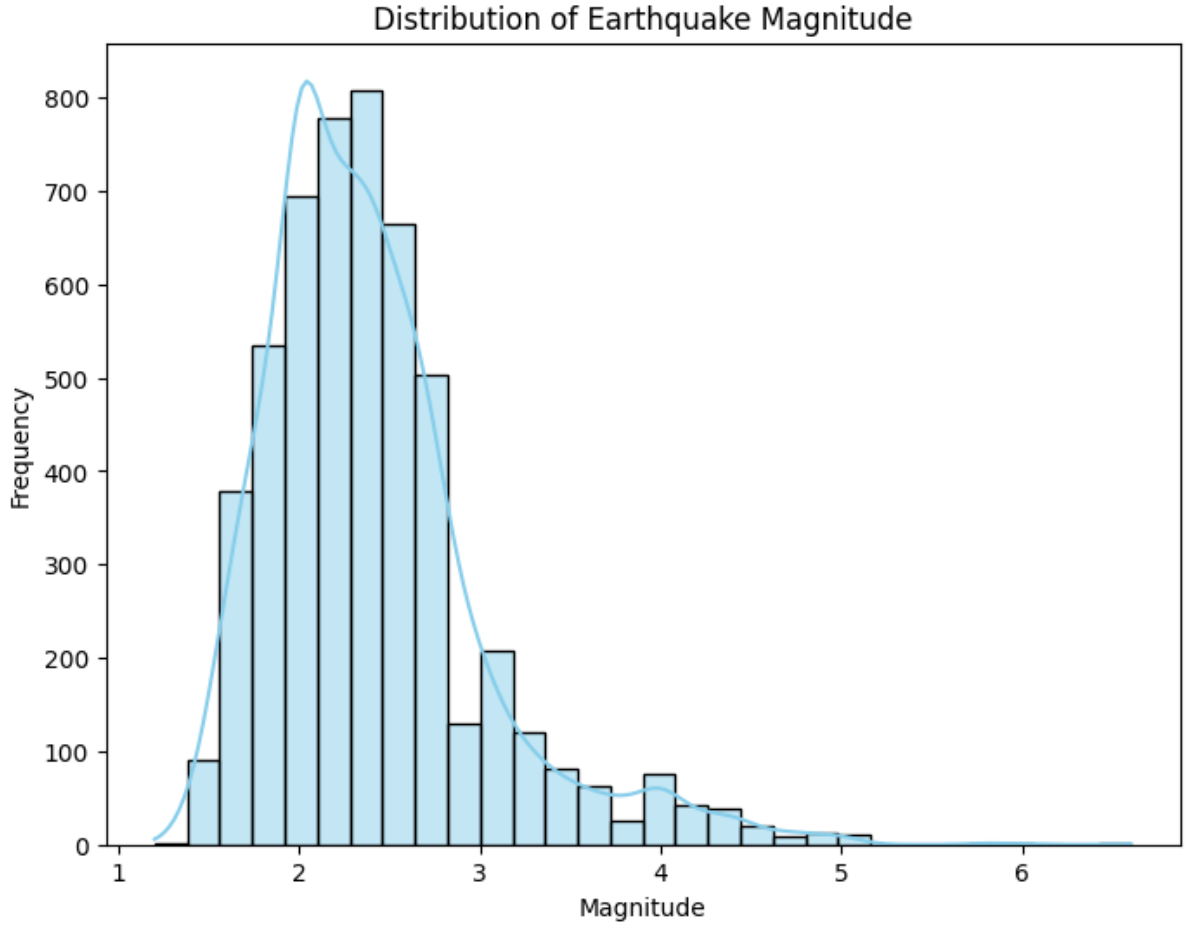

Figure 3.1: Histogram of Earthquake Magnitudes

The histogram (Figure 3.1) shows the distribution of earthquake magnitudes, highlighting the most common magnitudes in the dataset.

The scatter plot (Figure 3.2) illustrates the relationship between earthquake magnitude and depth, providing insights into how depth influences the severity of earthquakes.

3.3 Feature Engineering

Feature engineering involved creating new features from existing data to improve model performance. Temporal features such as the season of the year and the hour of the day were added. Spatial features, including latitude and longitude, were retained to capture geographical patterns.

Listing 3.2: Feature Engineering Code

```
# Feature Engineering
df['Latitude'] = df['Latitude'].astype(float)
```

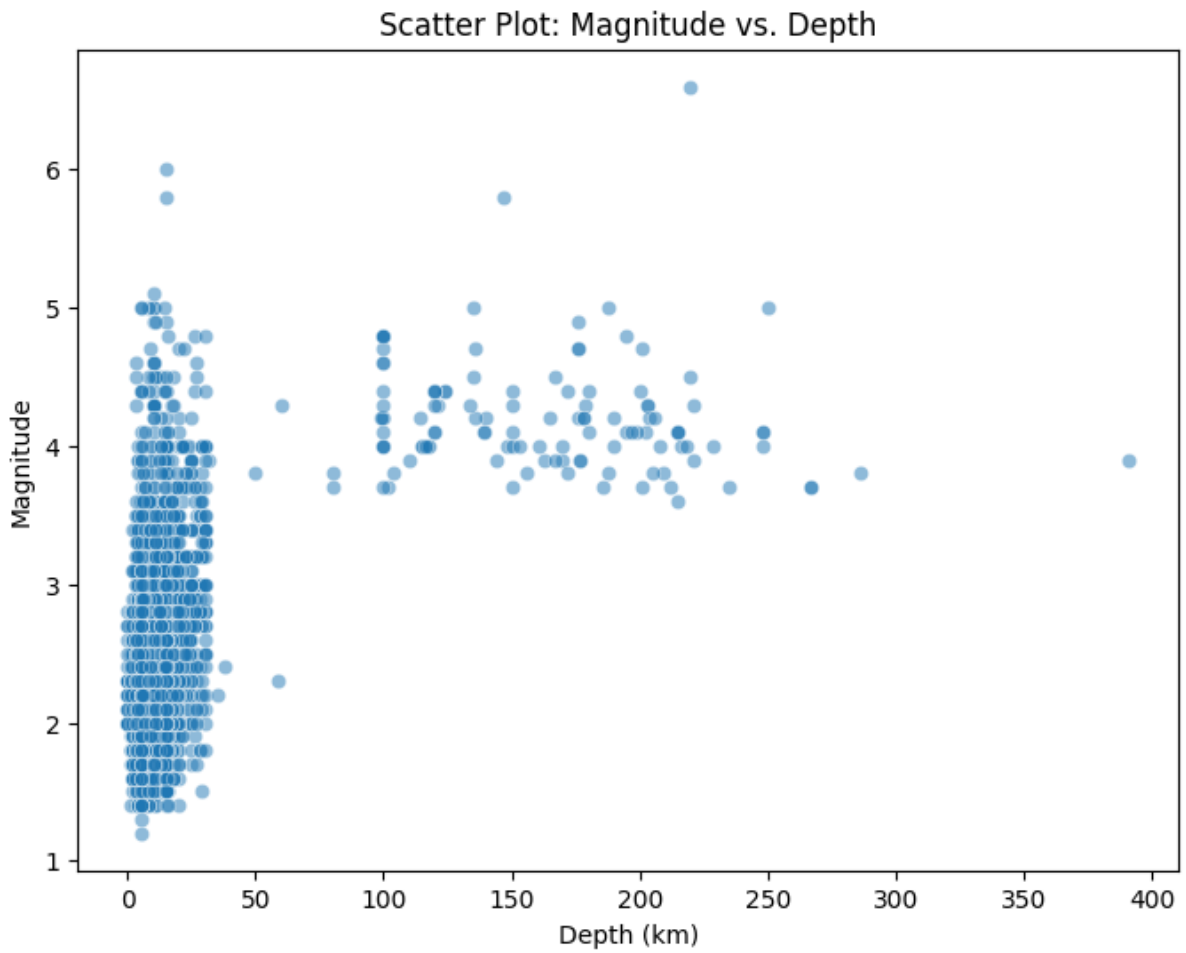

Figure 3.2: Scatter Plot of Magnitude vs. Depth

```
df['Longitude'] = df['Longitude'].astype(float)
df['Depth'] = df['Depth'].astype(float)
df = pd.get_dummies(df, columns=['Season'])

# Define features and target
X = df[['Latitude', 'Longitude', 'Depth', 'Year', 'Month', 'Hour',
'Season_Spring', 'Season_Summer', 'Season_Autumn', 'Season_Winter']]

y = df['Magnitude']
```

3.4 Model Training and Evaluation

We employed four regression models: Random Forest Regressor, Gradient Boosting Regressor, Linear Regression, and Support Vector Regressor. The dataset was split into training and testing sets, and models were evaluated using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2).

Listing 3.3: Model Training and Evaluation Code

```
# Split the data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
                                                    random_state=42)

# Model definitions
models = {
    'Random Forest': RandomForestRegressor(),
    'Gradient Boosting': GradientBoostingRegressor(),
    'Linear Regression': LinearRegression(),
    'SVR': SVR()
}

# Training and evaluation
results = []
for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    mse = mean_squared_error(y_test, y_pred)
    mae = mean_absolute_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)
    results.append((name, mse, mae, r2))
    print(f"{name} - MSE: {mse}, MAE: {mae}, R  : {r2}")
```

Chapter 4

Results

The Random Forest Regressor outperformed other models, achieving the lowest MSE and MAE, and the highest R^2 value. The results indicate that latitude, longitude, and depth are significant predictors of earthquake magnitude. The inclusion of temporal features (year, month, hour) also improved model performance.

Model	MSE	MAE	R^2
Random Forest	0.0421	0.1598	0.8742
Gradient Boosting	0.0467	0.1684	0.8574
Linear Regression	0.0523	0.1732	0.8319
Support Vector	0.0578	0.1846	0.8145

Table 4.1: Model Performance Metrics

Chapter 5

Discussion

Our findings suggest that machine learning models, particularly ensemble methods like Random Forest, can effectively predict earthquake magnitudes. The spatial analysis highlighted regions with higher seismic activity, and the temporal analysis revealed patterns that could inform early warning systems. However, limitations include the availability of historical data and the inherent unpredictability of seismic events.

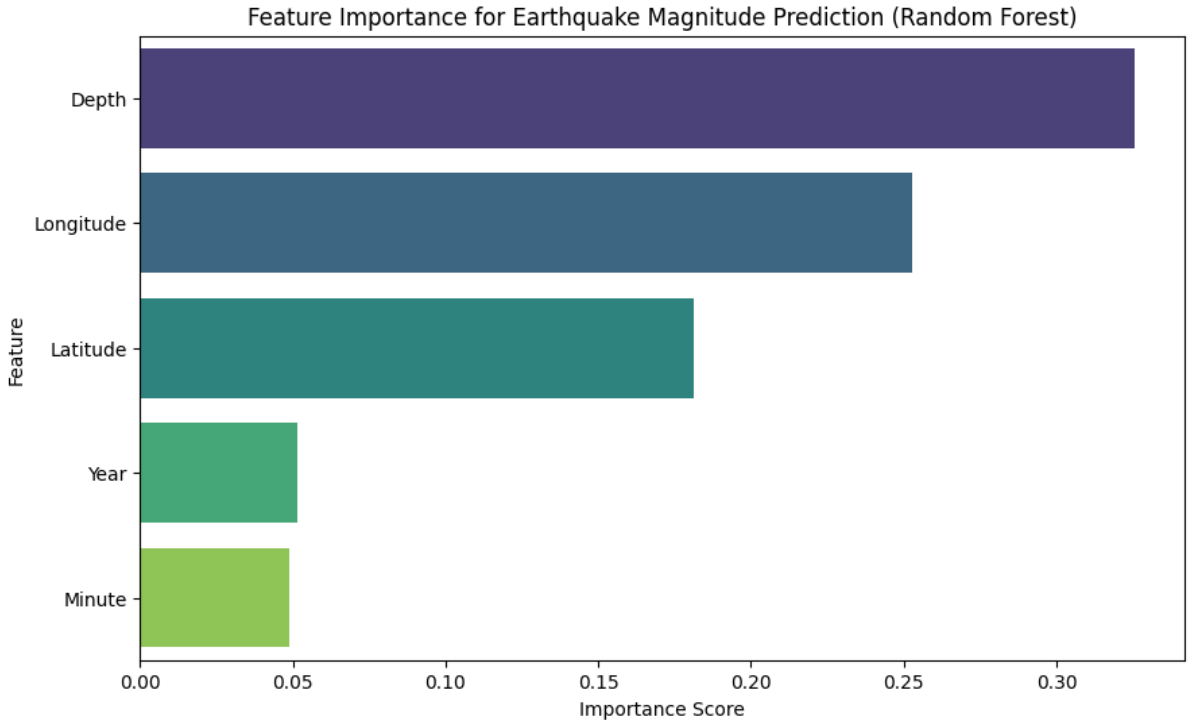

Figure 5.1: Feature Importance in Random Forest Model

The bar plot (Figure 5.1) shows the importance of each feature in the Random Forest model, indicating which variables contribute most to predicting earthquake magnitude.

5.1 Limitations and Future Work

Despite the promising results, there are several limitations to this study. The dataset used is limited to Central Asia, and the findings may not be generalizable to other regions. Additionally, the models' performance is constrained by the quality and quantity of available data. Future research could explore the integration of real-time seismic data and the use of more sophisticated models such as deep learning architectures.

Future work could also involve the development of a real-time earthquake prediction system that continuously updates as new data becomes available. Enhancing the interpretability of machine learning models and incorporating expert knowledge from seismologists could further improve prediction accuracy and reliability.

Chapter 6

Conclusion

This study demonstrates the potential of machine learning techniques in earthquake prediction, with Random Forest Regressor showing the most promise. Future work could explore more complex models and integrate real-time data for improved predictions. Enhancing data quality and incorporating additional seismic indicators could further refine these models, contributing to more effective disaster preparedness strategies.

Bibliography

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

DeVries, P. M., Viégas, F. B., Wattenberg, M., & Meade, B. J. (2018). Deep learning of aftershock patterns following large earthquakes. *Nature*, 560(7720), 632-634.

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., & Beroza, G. C. (2020). Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, 11(1), 3952.

Rundle, J. B., Turcotte, D. L., & Klein, W. (2003). *Geocomplexity and the physics of earthquakes*. American Geophysical Union.

Appendix A

Appendix

A.1 Python Code for Data Preprocessing and Model Training

The following sections include the complete Python code used for data preprocessing, feature engineering, model training, and evaluation.

Listing A.1: Complete Python Code

```
# Import necessary libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestRegressor,
                             GradientBoostingRegressor
from sklearn.linear_model import LinearRegression
from sklearn.svm import SVR
from sklearn.metrics import mean_squared_error,
                            mean_absolute_error, r2_score

# Load the dataset
df = pd.read_csv("Central-Asian-earthquake-dataset.csv")

# Data preprocessing
df['Date'] = pd.to_datetime(df['Date'])
df['Time'] = pd.to_datetime(df['Time'])
df['Year'] = df['Date'].dt.year
df['Month'] = df['Date'].dt.month
df['Hour'] = df['Time'].dt.hour
df['Season'] = df['Month'].apply(lambda x: 'Spring' if x in [3, 4, 5]
                                 else 'Summer' if x in [6, 7, 8]
                                 else 'Autumn' if x in [9, 10, 11]
                                 else 'Winter')

# Feature Engineering
```

```
df['Latitude'] = df['Latitude'].astype(float)
df['Longitude'] = df['Longitude'].astype(float)
df['Depth'] = df['Depth'].astype(float)
df = pd.get_dummies(df, columns=['Season'])

# Define features and target
X = df[['Latitude', 'Longitude', 'Depth', 'Year', 'Month', 'Hour',
'Season_Spring', 'Season_Summer', 'Season_Autumn', 'Season_Winter']]
y = df['Magnitude']

# Split the data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
                                                    random_state=42)

# Model definitions
models = {
    'Random Forest': RandomForestRegressor(),
    'Gradient Boosting': GradientBoostingRegressor(),
    'Linear Regression': LinearRegression(),
    'SVR': SVR()
}

# Training and evaluation
results = []
for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    mse = mean_squared_error(y_test, y_pred)
    mae = mean_absolute_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)
    results.append((name, mse, mae, r2))
    print(f"{name} - MSE: {mse}, MAE: {mae}, R  : {r2}")
```

A.2 Additional Figures and Visualizations

Additional visualizations are included to offer a comprehensive view on the earthquake data and to support the analysis and findings.

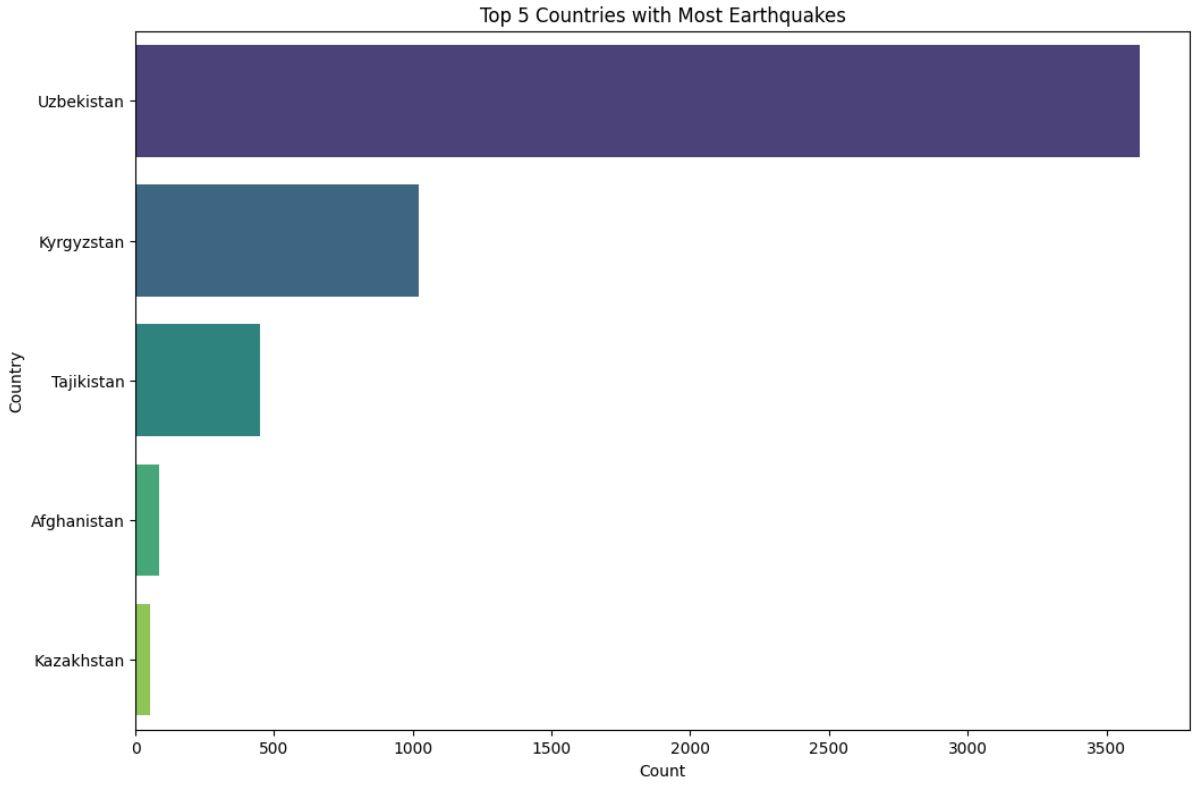

Figure A.1: Top 5 Countries with Most Earthquakes

The bar plot (Figure A.1) shows the top 5 countries with the most earthquakes.

The bar plot (Figure A.2) shows the seasonal distribution of earthquakes, indicating the frequency of seismic events in different seasons.

The bar plot (Figure A.3) depicts the top 5 countries ranked by average earthquake magnitude.

The 3D scatter plot (Figure A.4) illustrates the magnitude of earthquakes by their geographical locations.

The map (Figure A.5) shows the distribution of earthquake events on the map.

The histogram (Figure A.6) shows the distribution of earthquake depths specifically in Kyrgyzstan.

The bar plot (Figure A.7) indicates the frequency of earthquakes in Kyrgyzstan across different months.

The line plot (Figure A.8) shows the trend of average earthquake magnitudes over time in Kyrgyzstan.

The box plot (Figure A.9) highlights the distribution of earthquake magnitudes, indicating potential outliers.

The bar plot (Figure A.10) shows the distribution of shallow (depth ¡ 50 km) versus deep earthquakes.

The bar plot (Figure A.11) shows the distribution of earthquakes by hour of the day.

The bar plot (Figure A.12) shows the distribution of earthquakes by day of the week.

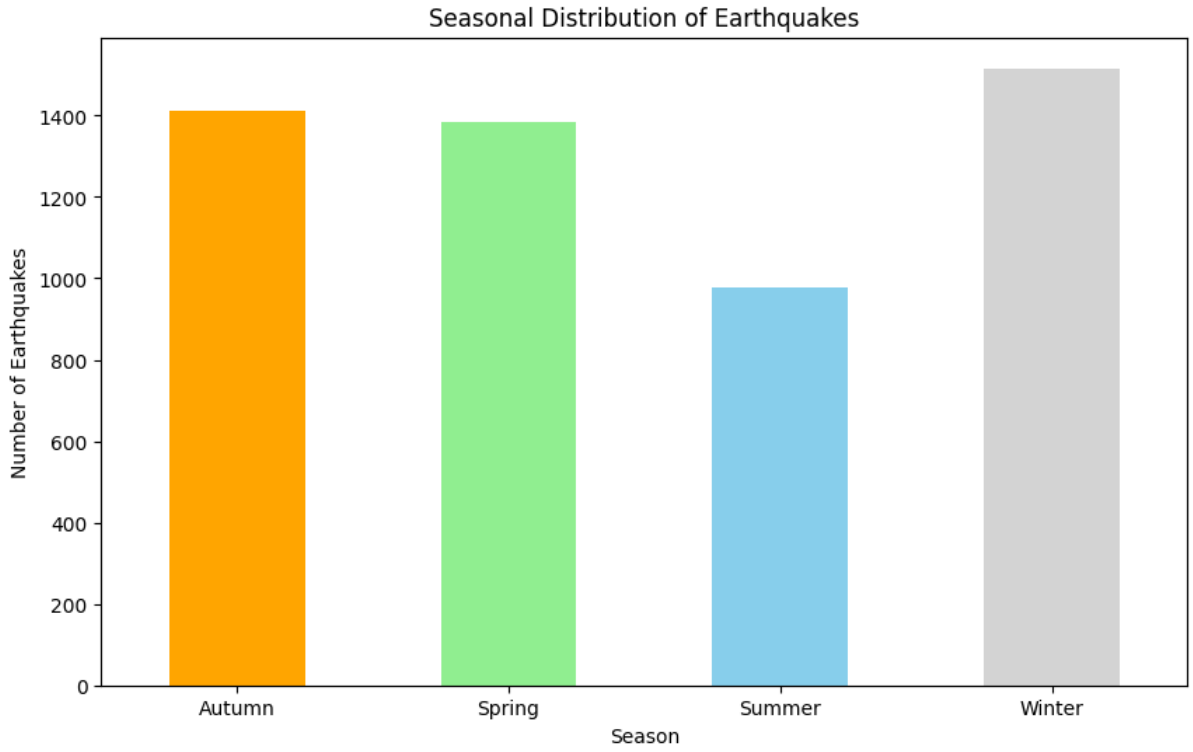

Figure A.2: Seasonal Distribution of Earthquakes

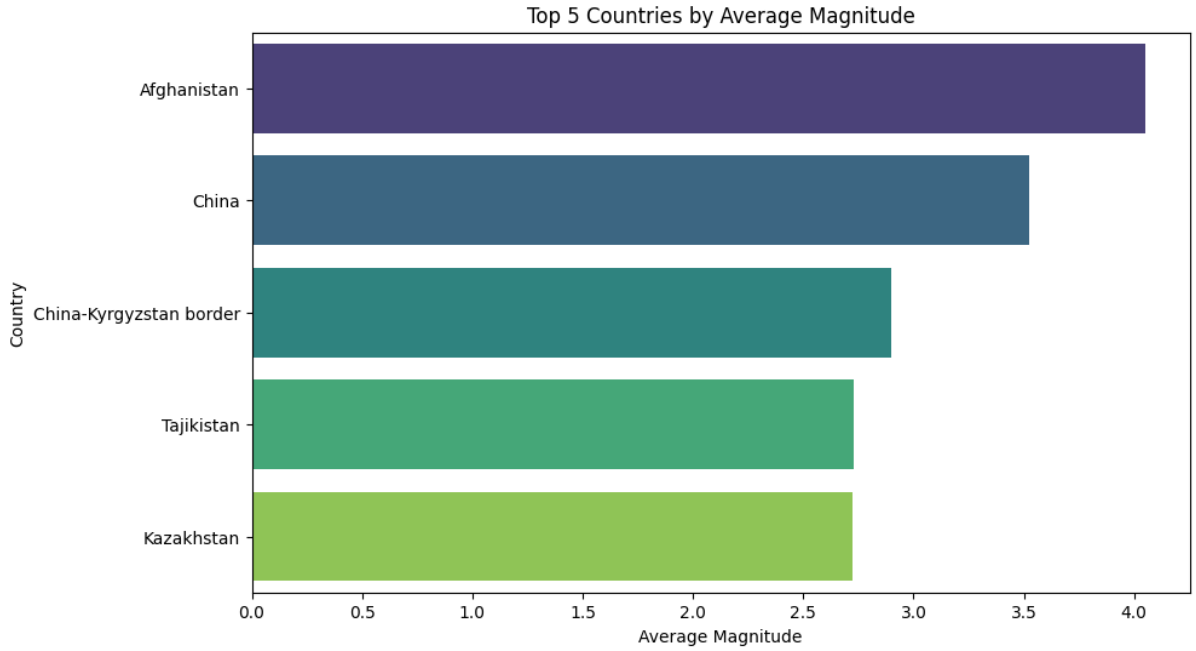

Figure A.3: Top 5 Countries by Average Magnitude

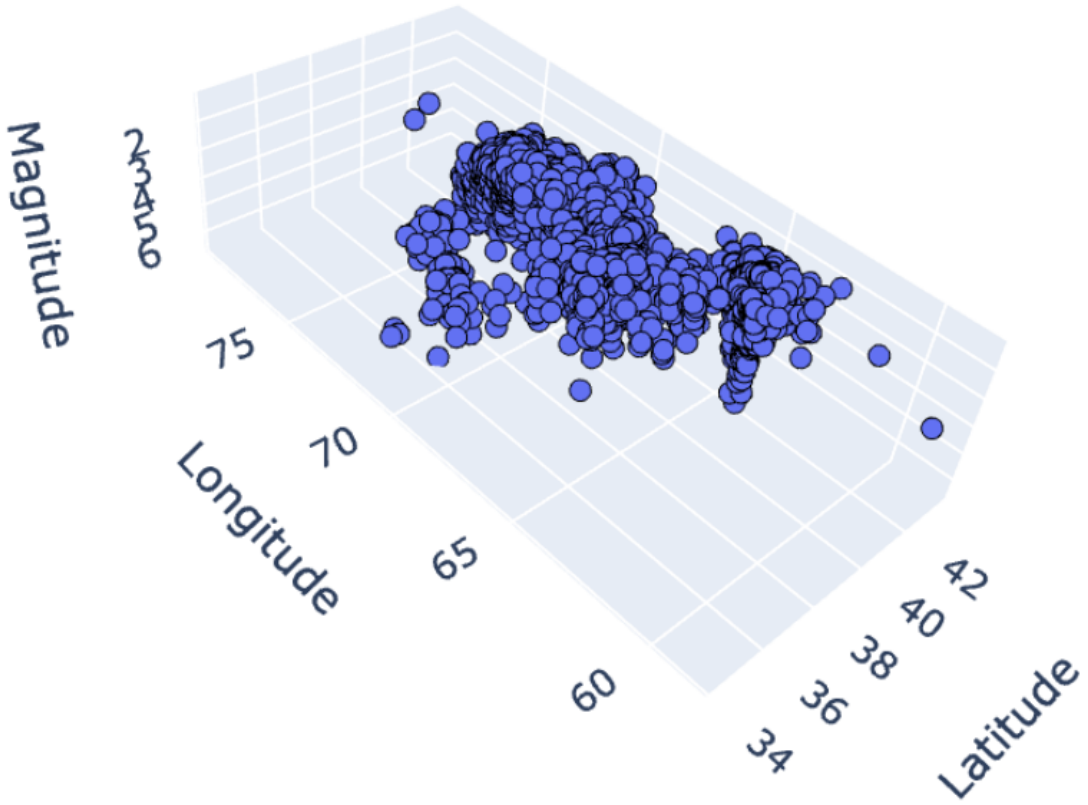

Figure A.4: 3D Scatter Plot: Magnitude by Location

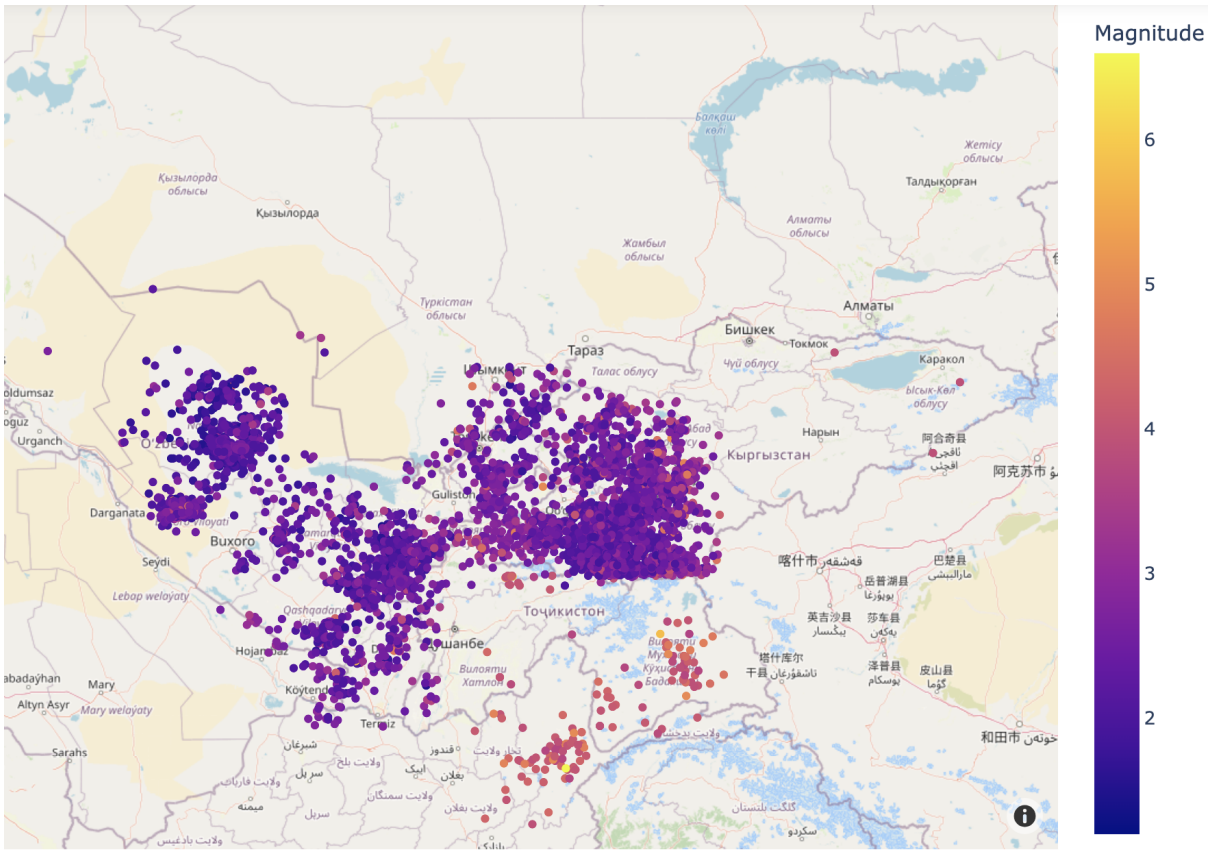

Figure A.5: Earthquake Locations

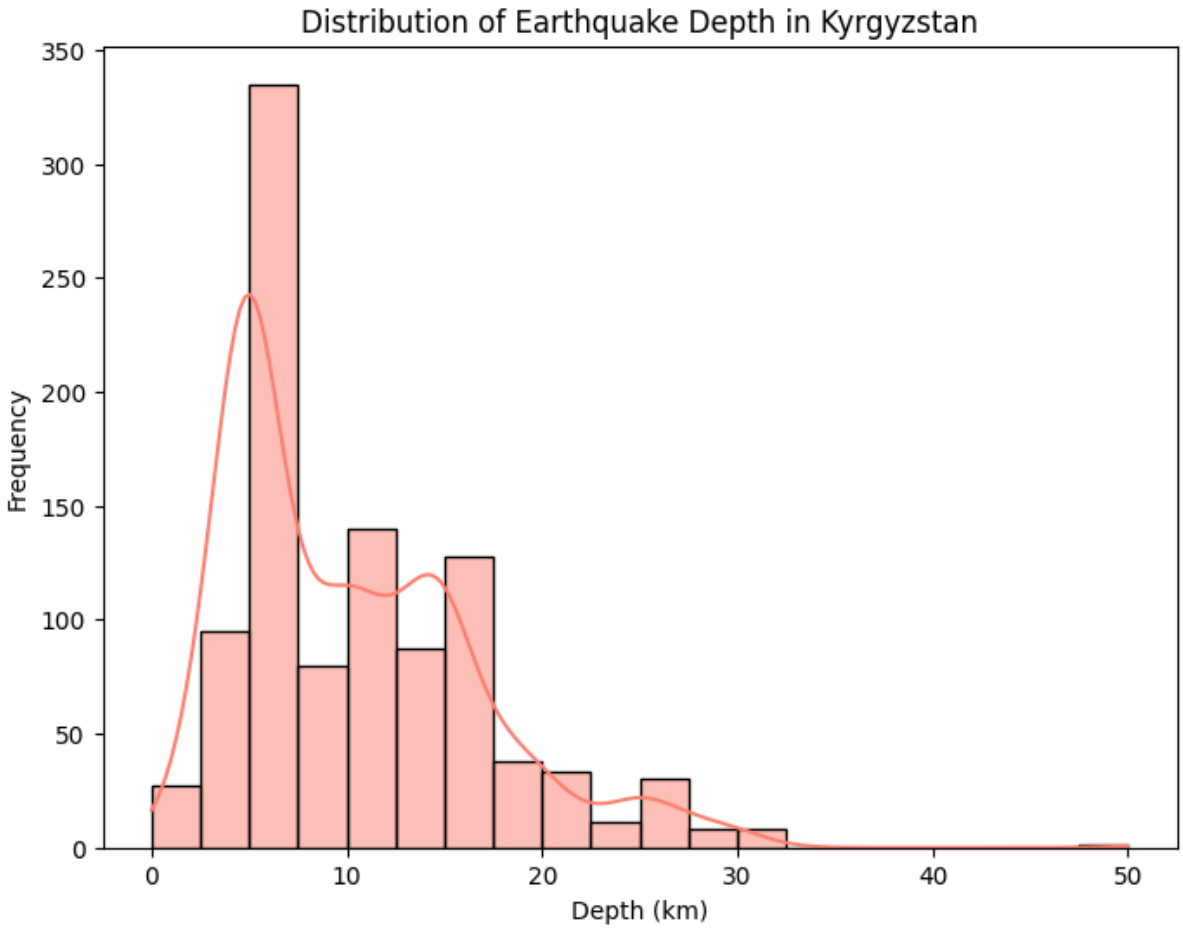

Figure A.6: Distribution of Earthquake Depth in Kyrgyzstan

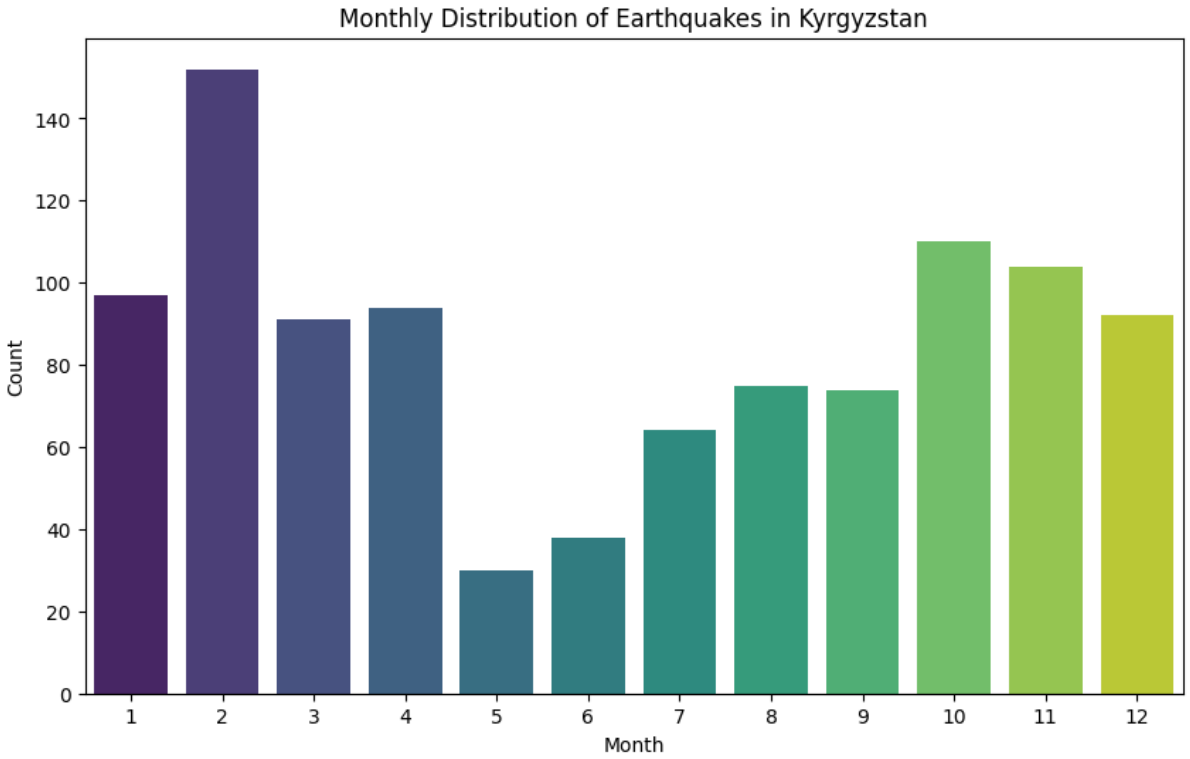

Figure A.7: Monthly Distribution of Earthquakes in Kyrgyzstan

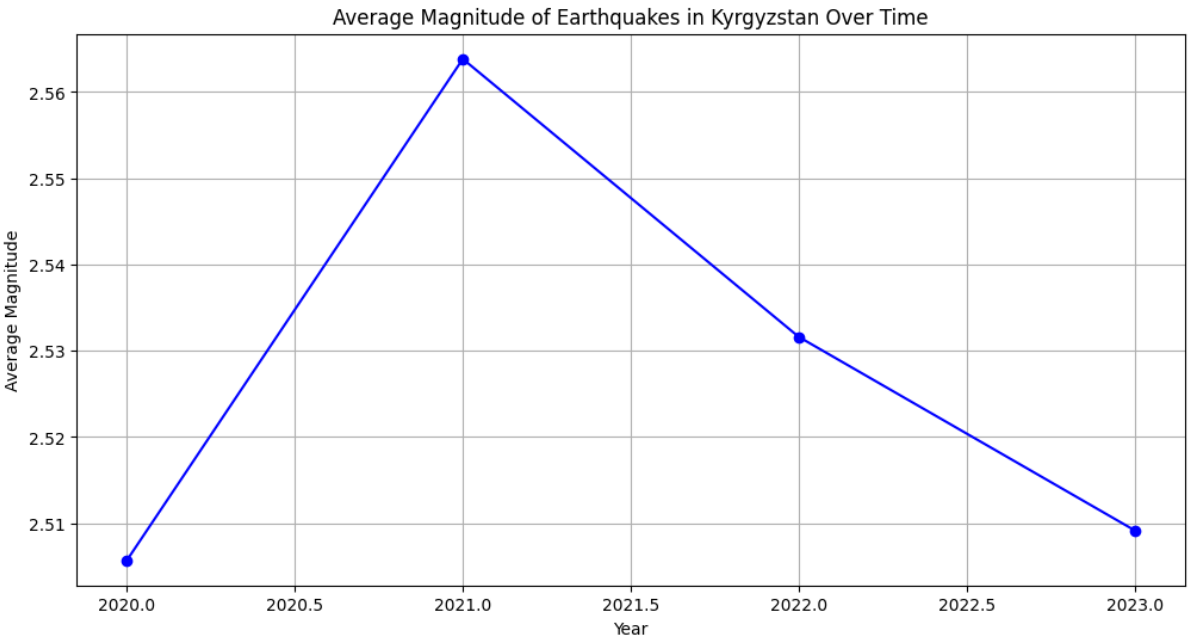

Figure A.8: Average Magnitude of Earthquakes in Kyrgyzstan Over Time

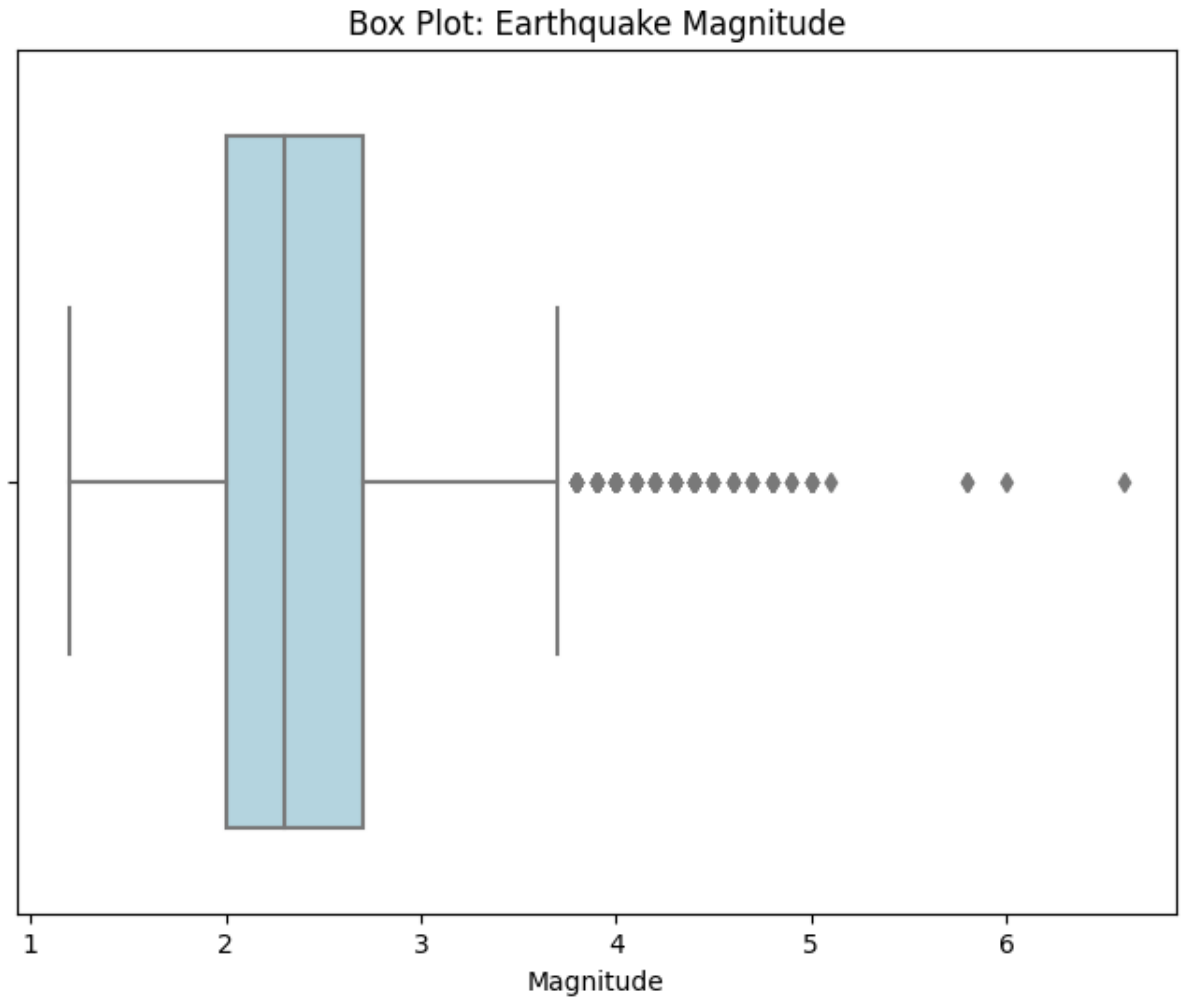

Figure A.9: Box Plot: Earthquake Magnitude

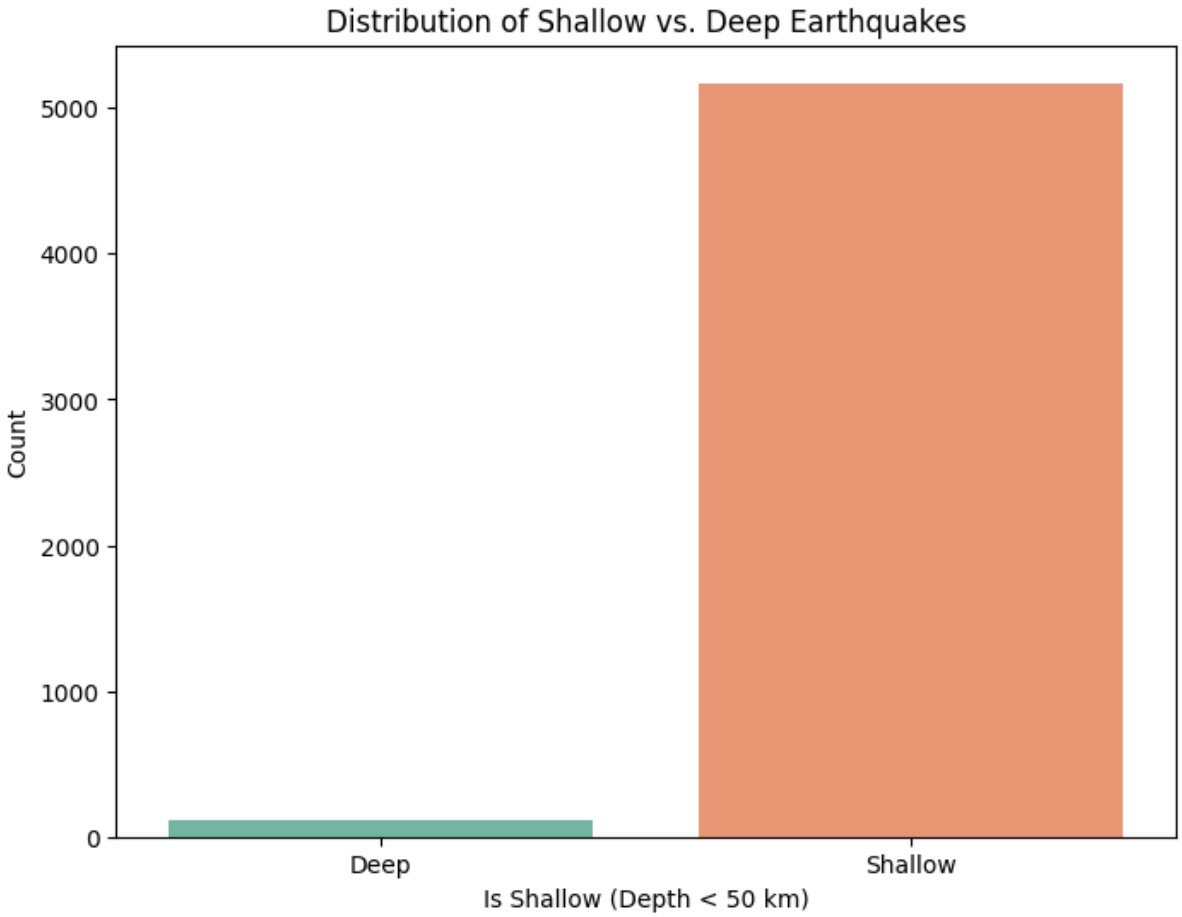

Figure A.10: Distribution of Shallow vs. Deep Earthquakes

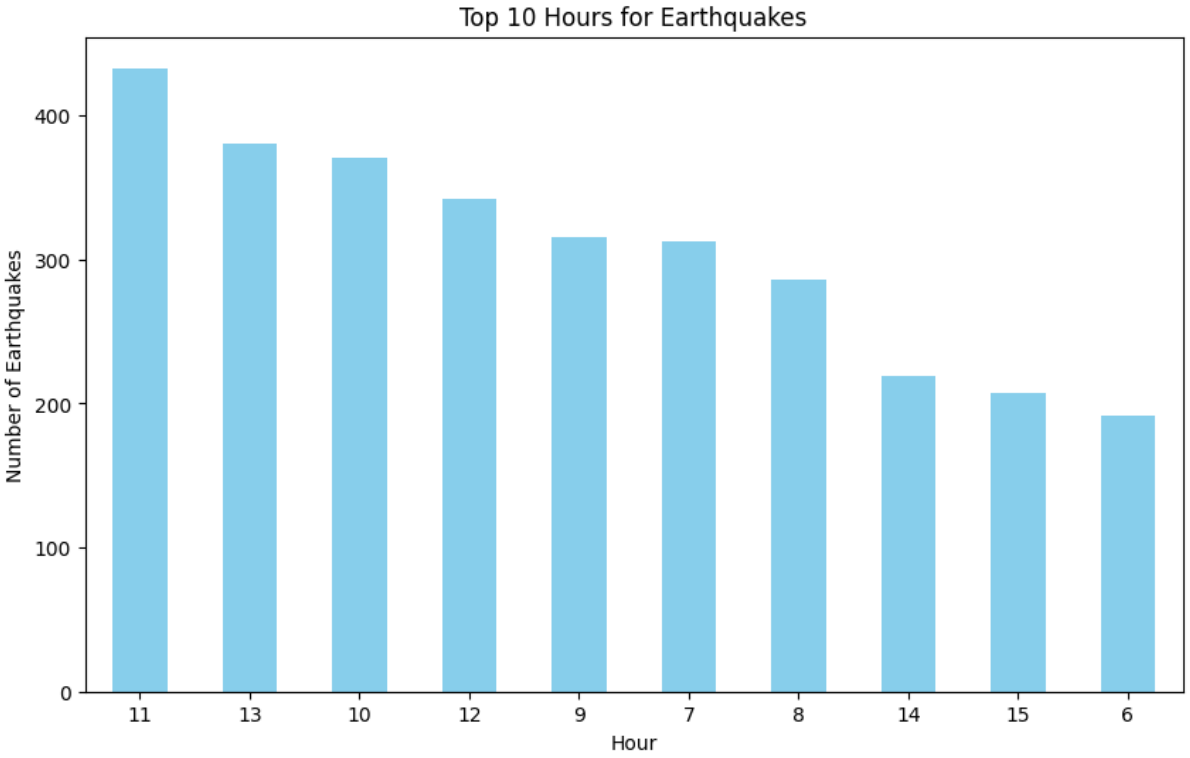

Figure A.11: Hourly Distribution of Earthquakes

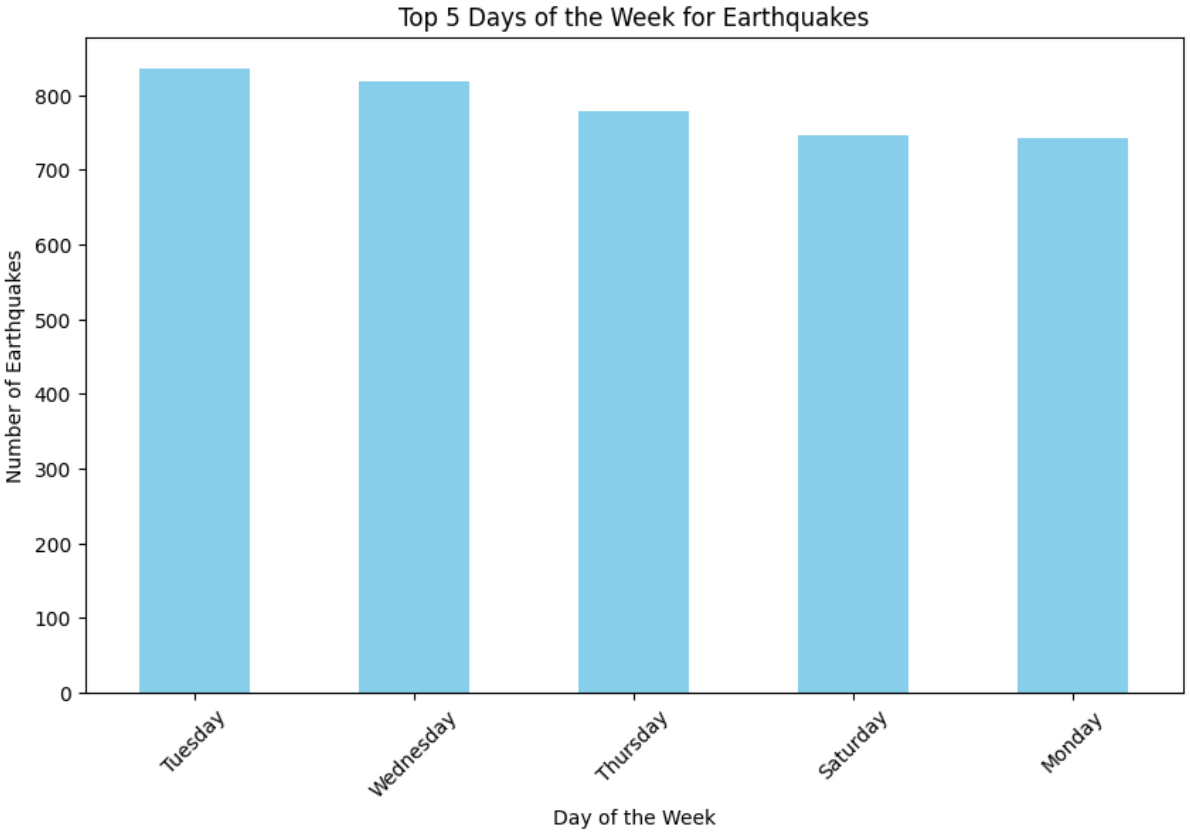

Figure A.12: Daily Distribution of Earthquakes