**Los Angeles Earthquake Prediction**

**Using SAC**

Authors: Jasmine Pesantes, Christian Fierro, Kimberly Cervantes, Sze Wa Ng, Alejandro Ortiz Cazares

Professor: Jongwook Woo

California State University Los Angeles, FEMBA, BUS5100-93: Introduction to Business Analytics

[aortiz45@calstatela.edu](mailto:aortiz45@calstatela.edu), [cfierro8@calstatela.edu](mailto:cfierro8@calstatela.edu), [jdawson3@calstatela.edu](mailto:jdawson3@calstatela.edu), [kcerva40@calstatela.edu](mailto:kcerva40@calstatela.edu), [sng23@calstatela.edu](mailto:sng23@calstatela.edu)

**Abstract:** The purpose of our paper is to analyze the frequency and magnitude of earthquakes within Los Angeles between the years of 2012 and 2024 compiled by the Southern California Earthquake Data Center (SCEDC) maintained by the California Institute of Technology. The 6.77 MB dataset l LosAngeles\_Earthquake\_Dataset.csv obtained from the data set website Kaggle.com https://www.kaggle.com/datasets/batuhankalem/los-angeles-earthquake-dataset/data. The data helps us analyze and understand magnitude, depth, and geography using predictive modeling, including regression and time series analysis, to help forecast future trends, while clustering techniques can categorize areas more prone to seismic activity. Analysis and prediction of earthquakes is important to assist citizens to adequately prepare for the next large earthquake via retrofitting structures, having adequate emergency supplies on hand, and proper insurance coverage to rebuild if necessary.

# 1. Introduction

This dataset comprises extensive earthquake records for the Los Angeles region, obtained from the Southern California Earthquake Data Center (SCEDC) spanning from a twelve-year period from January 1, 2012, to September 1, 2024. It covers essential information such as magnitude categories (Ml, Mw, MLr), depth, and geographic locations within a 100-kilometer radius of Los Angeles. To guarantee uniformity, all magnitude categories were converted to Local Magnitude (ML) using SCEDC conversion methods, with non-standard magnitudes (Mh), which made up just 0.22% of the data, eliminated for accuracy. The dataset includes pre-processed variables such as rolling mean depth, Gutenberg-Richter b-values, and maximum magnitudes across certain time periods, making it ideal for machine learning applications. This data set was created to improve earthquake prediction algorithms and has already led to better forecasting accuracy. Researchers and data scientists may use it to examine patterns, create prediction algorithms, aid in risk assessment, and preparedness activities in earthquake-prone metropolitan areas.

# 2. Related Work

Predictive modeling is a critical tool in seismology. It helps forecast future seismic events by analyzing historical data. Regression analysis allows researchers to identify relationships between variables such as tectonic plate movement and earthquake magnitudes. Linear regression can be used to predict the likelihood of an earthquake based on stress accumulation along fault lines (Smith et al., 2020). Time series analysis, on the other hand, is particularly useful for understanding temporal patterns in seismic activity. By analyzing sequences of earthquake occurrences over time, researchers can identify cyclical trends or anomalies that may indicate increased seismic risk (Smith et al., 2020).

Clustering techniques and hierarchical clustering are widely used in seismology to categorize regions based on susceptibility to seismic activity. These methods group areas with similar seismic characteristics, such as frequency and magnitude of earthquakes, into distinct clusters. This categorization helps policymakers and urban planners prioritize disaster preparedness efforts in high-risk areas. By integrating clustering techniques with predictive modeling, researchers can not only forecast future seismic trends but also spatially delineate zones that require heightened monitoring and intervention (Johnson & Lee, 2019).

A study by Martinez et al. (2021) demonstrated the effectiveness of this integrated approach in analyzing seismic data from California. This combined methodology not only enhances the accuracy of seismic forecasts but also supports targeted mitigation strategies, reducing the impact of earthquakes on vulnerable communities.

# 3. Specifications

The dataset was retrieved from Kaggle, which is an online community where data sets can be published and examined. The data has been updated on a consistent basis with data from the Southern California Earthquake Data Center (SCEDC) to be inclusive from 2012 to 2024. We have a high level of confidence the data is adequate for our modeling as it is well documented and verified by governmental resources such as the United States Geological Survey. While this data can be used as a tool to predict seismic activity, earthquake strikes are mostly random in nature. In addition to known fault zones, unknown fault zones are present. Only once an earthquake erupts does it reveal itself. For example, the Puente Hills fault system was only recently discovered in 1999, and is thought to be potentially more severe and devasting than the San Andres fault due to the number of densely populated urban communities it is under. The most recent earthquake on this newer known fault was a 4.4 magnitude in August 2024 in El Sereno.

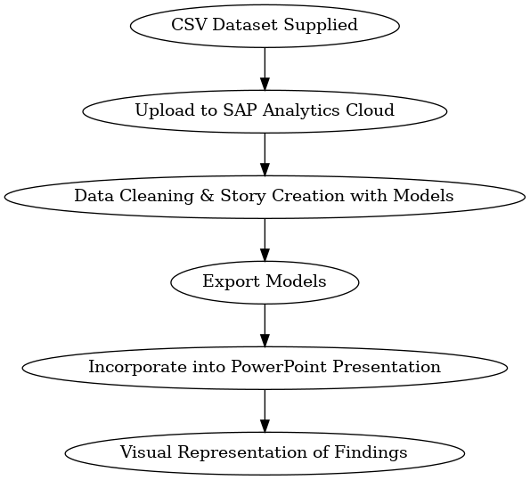
Table 1. Data Specifications – Los Angeles Earthquake

|  |  |
| --- | --- |
| Data Set | Size |
| Los Angeles Earthquake Dataset | 6.77 MB |

# 4. Implementation Flowchart

The raw dataset we downloaded from Kaggle provided valuable insights into earthquakes and the frequency of seismic events over a recent period. This data, originally scraped from the Southern California Earthquake Data Center (SCEDC) and subsequently posted on Kaggle, was pivotal in our analysis. A detailed flowchart below illustrates the data process followed for our analysis. The dataset was supplied in a single CSV file, which we then uploaded to SAP Analytics Cloud. Utilizing this platform, we cleaned the data and created comprehensive stories with models. These models were subsequently exported and incorporated into a PowerPoint presentation, enabling us to visually represent our findings effectively.

Figure 1. Implementation Flowchart



# 5. Data Cleaning

The data set download included latitude and longitude, forming a comprehensive Geo Mapping template. Our team imported the CSV file and uploaded it to SAP Analytics Cloud as a datasheet. SAP determined there were “no issues detected” in the sample data updated and we were at 98.38% confidence level. We identified the data and updated the measure aggregation type from sum to none for measures such as magnitude, elapsed time, and class. The b-value increment for each level was maintained as it was deemed crucial for projection and forecasting during regression analysis. The dataset was then ready for upload into the modeler, followed by story creation in SAP. Subsequently, we verified that the b-value is essential for interpreting the relationship between earthquake frequency and magnitude. A b-value greater than 1.0 indicates more small earthquakes and fewer severe ones, while a b-value less than 1.0 suggests more large earthquakes, indicating extreme seismic hazards. Further research on negative magnitude earthquakes was performed after the data revealed numerous earthquakes of this magnitude. A review of the usgs.gov website revealed that negative magnitude earthquakes, known as microquakes, are not felt by humans but only detectable with instrumentation.

# 6. Analysis and Visualization

Following the evaluation of the data and the selection of the most pertinent information for our analysis, we utilized SAP Analytics Cloud to develop a predictive model and narrative. This model enabled us to compare a variety of factors, including the elapsed time of occurrence, magnitude, and class, by providing a detailed visual representation of earthquake locations. By applying these insights, we were able to develop a more profound comprehension of the earthquake patterns and their implications.

## 6.1 Relationship between Earthquake Size and Frequency

The relationship between earthquake size and frequency is fundamental to understanding seismic activity patterns. Smaller earthquakes, typically those with a magnitude less than 3.0, occur far more frequently than larger ones. This relationship, known as the Gutenberg-Richter law, shows that for every tenfold decrease in earthquake size, the frequency of occurrence increases significantly. These minor tremors may happen daily and often go unnoticed by people. However, they play a crucial role in releasing accumulated stress in the Earth's crust, potentially preventing larger, more destructive earthquakes. By analyzing this relationship, scientists can better assess seismic hazards, predict future earthquake occurrences, and develop strategies to mitigate the impact of seismic events on communities.

Figure 2. Relationship between Earthquake Size and Frequency

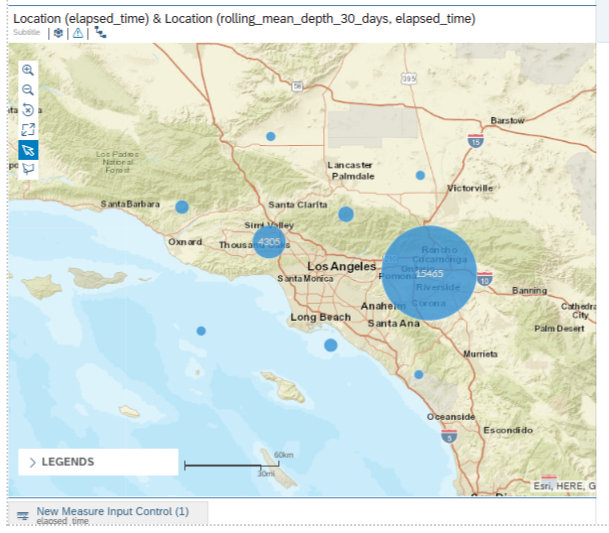
A map of the earth with a number of earthquake

AI-generated content may be incorrect.

## 6.2 GeoMapping of Duodecennial Earthquakes in the Los Angeles Area

The GeoMapping created in SAP demonstrates a substantial number of earthquakes that occurred within Los Angeles over the past 12 years. The data reveals over 15,000 earthquake occurrences in the eastern part of Los Angeles, with an additional 4,000 in the Thousand Oaks area. By zooming into the sample below, one can observe even more dispersed data points, offering a comprehensive view of the seismic activity in these regions.

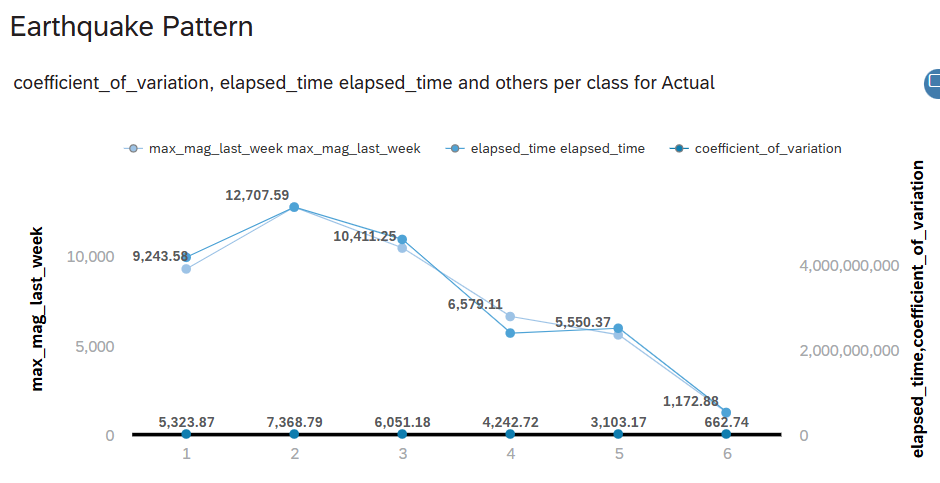
Figure 3. Recent earthquakes in Los Angeles area.



## 6.3 Coefficient of Variation and Earthquake Predictability

The coefficient of variation (CV) is a crucial metric in the study of earthquake patterns, providing insights into the predictability and regularity of seismic events. A high CV, greater than 1.0, indicates more unpredictable earthquake occurrences, reflecting significant variability in the timing and magnitude of earthquakes. This unpredictability poses a challenge for forecasting and hazard assessment, as it suggests irregular intervals between seismic events. Conversely, a low CV, <1.0, implies more regular earthquake cycles with events occurring at more consistent intervals. This regularity can aid in the development of predictive models and improve our understanding of seismic hazards. By analyzing the CV values, scientists can better assess the stability of a region and implement more effective strategies for earthquake preparedness and risk mitigation.

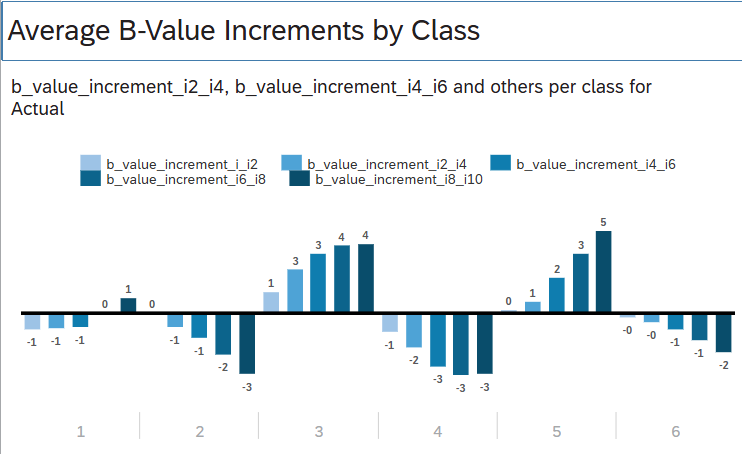
Figure 4. Earthquake Predictability



## 6.4 Impact of B-Value on Earthquake Frequency and Seismic Hazard

An increase in the B-value indicates a stronger impact of the independent variable on the correlation with the dependent variable. As the B-value rises, the variable's influence on the outcome becomes more pronounced. This is particularly significant in the context of seismic studies. For instance, a high B-value (greater than 1.0) implies that smaller earthquakes are more frequent, with fewer large earthquakes occurring. Conversely, a low B-value (less than 1.0) suggests that larger earthquakes are more common, leading to an increased seismic hazard. Understanding these intervals is crucial for accurately interpreting the relationship between earthquake frequency and magnitude, and for developing effective strategies to mitigate seismic risks. The average B-value intervals are illustrated below, providing clear definitions and implications for both high and low B-values. As below illustrates, just as the probability bell curve, the normal distribution focus within the middle session, bound around class 3-4, and flatter out at the two extremes end, class 1 and 6.

Figure 5. Earthquake Frequency and Seismic Hazard



## 6.5 Regression Analysis

The GeoMapping data from SAP revealed significant seismic activity in Los Angeles over the past decade, with over 15,000 earthquake events recorded in the eastern part of the region and an additional 4,000 occurrences near Thousand Oaks. By zooming in on the sample data, a more detailed and dispersed distribution of these earthquake events becomes visible, offering a deeper understanding of the seismic patterns across the region. The model used for predicting these occurrences demonstrated a high level of accuracy, with a root mean square error (RMSE) indicating a 98.38% prediction confidence. This suggests that the model’s predictions closely align with the actual recorded data, effectively capturing the seismic trends with these areas. This combination of comprehensive visualization and accurate prediction modeling emphasizes the intensity of seismic activity in the Los Angeles and underscores the critical need for ongoing monitoring and preparedness to mitigate potential risk.

Figure 6. RMSE & Prediction Confidence

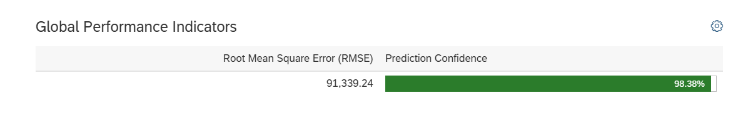
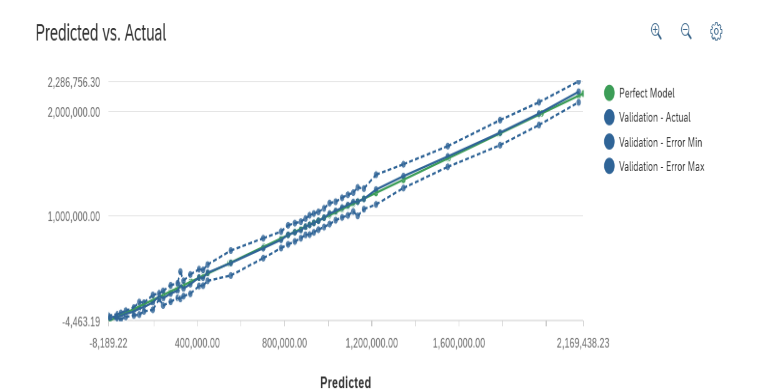
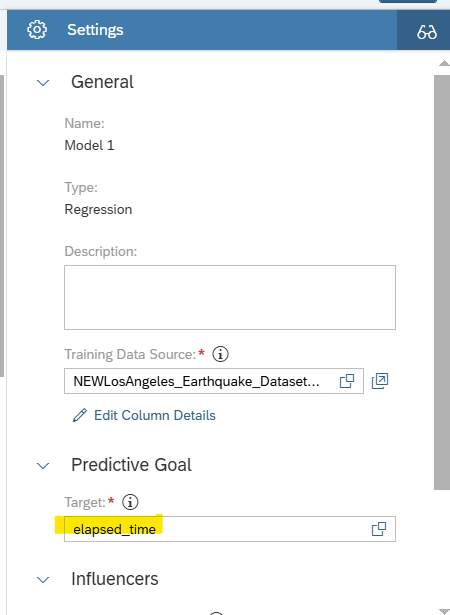
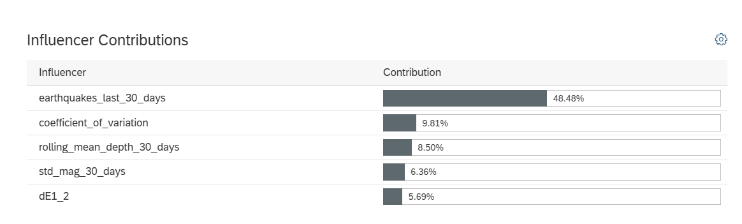


Figure 7. Predicted vs Actual



The data analysis reveals that the actual observed value was 1,795,365.06, which is very close to the model’s perfect predictions of 1,788,394.24, indicating that the model is generally performing with a high degree of accuracy. However, the model exhibited variability in its predictions, with the maximum error recorded at 1,915,213.42 and the minimum error at 1,675,576.70, resulting in an overall error range of 239,636.72. This spectrum implies that although the model usually forecasts with accuracy, occasionally it deviates greatly from the real values.

Figure 9. Influencer Contributions 

The influencer’s contributions for the earthquake activity over the past thirty days point to the main elements influencing seismic patterns. Elapsed time was set as the target to forecast the time gap between earthquakes. With 48.48% of the effect, earthquake\_last\_30\_days is the most important factor; so, the sheer quantity of resent seismic occurrences determines the direction of current trends most substantially. With a contribution of 9.81% the coefficient of variation shows that the outputs of the model are strong influenced by variability in earthquake data. Emphasizing the value of average earthquake depth over the last month in determining seismic behavior, the rolling\_mean\_depth\_30\_days add 8.50%. Furthermore, the standard deviation of magnitudes std\_mag\_30\_days takes 6.63% into consideration, therefore representing the function of magnitude variance in seismic activity forecasting. Finally, dEl1\_2 has a little influence on the analysis of the model with 5.69%. With the frequency of recent earthquakes most dominating, these influences taken together offer a complete awareness of the elements causing seismic activity.

**7. Conclusion**

Using the data from the Southern California Earthquake Data Center (SCEDC), this study offers an in-depth study of Los Angeles area earthquake activity between 2012 and 2024. Using predictive modeling approaches including time series forecasting and regression analysis, along with clustering strategies we found important trends and patterns in seismic activities. With the B-value as a vital indication of seismic dangers, the study highlighted the strong correlation between earthquake frequency, magnitude, and depth. With a 98.38% confidence rating, the prediction model showed a great degree of accuracy, therefore supporting the dependability of the results. In addition, underlining in the study as main drivers of seismic trends were recent earthquake events, variability, and risk analysis among other aspects of earthquake readiness depends on these realizations. In the end, data driven methods can help to increase resilience and reduce the effects of next earthquakes in Los Angeles area even if seismic occurrences are always unexpected.

For a copy of this paper and relevant work, please visit https://github.com/cfier001/Earthquake-data.git

**References**

Kalem, B. (2023). *Los Angeles earthquake dataset* [Data set]. Kaggle. <https://www.kaggle.com/datasets/batuhankalem/los-angeles-earthquake-dataset/data>

Johnson, R., & Lee, S. (2019). Clustering seismic data for risk assessment: A case study of the Pacific Ring of Fire. Journal of Geophysical Research, 124(5), 789-801. <https://doi.org/10.1029/2018JB016789>

Martinez, P., Gonzalez, F., & Rivera, L. (2021). Integrating predictive modeling and clustering for seismic risk analysis in California. Seismological Research Letters, 92(3), 1456-1468. <https://doi.org/10.1785/0220200321>

Smith, T., Brown, A., & Davis, K. (2020). Predictive modeling in seismology: Regression and time series analysis. Earth and Planetary Science Letters, 530, 115-123. https://doi.org/10.1016/j.epsl.2020.115123

ABC7. (2024, August 14). Recent earthquakes raise concern over lesser-known—but dangerous—Puente Hills fault. ABC7 Los Angeles. <https://abc7.com/post/recent-earthquakes-raise-concerns-lesser-known-dangerous-puente-hills-fault/15190473>

U.S. Geological Survey. (n.d.). How can an earthquake have a negative magnitude? U.S. Department of the Interior. <https://www.usgs.gov/faqs/how-can-earthquake-have-negative-magnitude>