Assignment 5 - Team Analysis Assignment Using Cognos Analytics

Sandra Dufour, Christopher Heiner, Vidya Koduri, and Tanushree Kumar
DATA 610 - Fall 2023
Dr. George Cross

Introduction and Data Exploration

This paper aims to explore and analyze an earthquake dataset developed as part of the CORGIS Dataset Project. Earthquakes can potentially cause a significant loss of life and economic damage. Given these impacts, data analytics can be leveraged to predict and minimize these events. The dataset being explored contains enough data points and features to provide a starting point for how analytics can be used in this field. Using Cognos Analytics, this paper will explain how the dataset was prepared for analytics, what visuals were created to explore it, the predictive models developed, how this information was collected into a dashboard and story for presentation, and potential practical applications.

The dataset was by and large usable prior to cleaning, so it did not require extensive data preparation. Earthquakes included in the dataset took place from July to August 2016. There are 8,394 cases with 18 features. The features are composed of a few categories; time data, location data, information about the earthquake, and administrative information. The dataset's columns and the first few rows can be seen in Figure 1. All of the columns have readily understandable information except for gap, significance, and distance. The gap is a measure of the accuracy of the position of the earthquake while the significance is related to the impact of the event related to magnitude, felt reports, etc. Distance is measured in degrees with one degree being equal to approximately 111.2 km (Whitcomb, 2016). Given other columns are measured in kilometers the distance was converted using the aforementioned relationship. No columns contained missing values but distance depth did include negative values. These values do not make logical sense, so it was recalculated to form the corrected depth in kilometers. Thus, if the depth was between -.25 km and 0 km, it was rounded to 0 km. If the depth value measured below -.25 km, they were filtered out of the data set. Significance had a large range, 0-1000, so values were put together in groups of 50. Magnitude was also grouped in a similar way using naming similar to that found in Encyclopedia Britannica (Rafferty, 2018). The last calculations created were two-time columns to allow for visuals that summarized information by day and by hour.

Two filters were created for potential use with the visualizations and dashboards that were to be created. A filter was created to only include data from the three locations with the most earthquakes. This ended up being California, Nevada, and Alaska. Filtering the top two results serves two purposes. By only having the top results a significant amount of noise will be eliminated from subsequent visuals, allowing the viewer to more easily interpret the data being presented. Additionally, the target audience for this paper is American, making these top three results particularly relevant. A second filter was created to only show data points with a gap value below 180. The gap is a measure of the station's accuracy on the earthquake's position and depth. The data source mentioned that values greater than 180 have large location and depth uncertainty so that value was selected as a cutoff (Whitcomb, 2016). Inaccurate data points can significantly impact a predictive model so it was important to remove them.

Data Visualization Process & Results

According to the United States Geological Survey, "the strength of shaking from an earthquake diminishes with increasing distance from the earthquake's source, so the strength of shaking at the surface from an earthquake that occurs at 500 km deep is considerably less than if the same earthquake had occurred at 20 km depth (USGS, 2011)". This is a direct reflection of our line graph 'Magnitude' vs. 'Corrected Depth' (km), seen in Figure 2 because it shows at a 95% confidence interval the summed values of 'Corrected Depth' range from a minimum of 0.42 (when the magnitude is 2.65) to a maximum of 12,759.85 (when the magnitude is 4.4). According to the Earthquake Magnitude Scale (Michigan Tech, 2023), a 4.4 magnitude earthquake will cause minor damages at the maximum depth (km) away from the earth's mantle at 12,759.85 km. We can also determine from the graph that at a magnitude of 6.3 and a closer distance to the earth's mantle at 510 (km), it may cause more significant damage in populated areas.

For significance, Alaska and California are the most important categories of State/Country with a total value of 169,597 (36.4 % of the total). The summed values of significance range from a minimum of 29 (when the State/Country is Pennsylvania) to a maximum of 96,052 (when the State/Country is Alaska). The value of significance is unusually high when the values of State/Country are Alaska and California. This map visualization can be seen in Figure 3.

Month trend analysis

The details of the graph of 'Magnitude by Month', seen in Figure 4, show that the aggregate sum of 'magnitude' across all months is 12,367.32, which is the total value we're analyzing. The 'Range of Summed Values by Month' is given by the minimum and maximum values. The 'Minimum Value' is the lowest summed value of magnitude, which is 2,356.57, and occurs in month 7 (July). The 'Maximum Value' is the highest summed value of magnitude significantly larger, at 10,010.75, and it occurs in month 8 (August).

One can interpret that there is a stark contrast between the values for July and August, suggesting a significant variation in magnitude between these two consecutive months. Given that the total for August is very close to the overall total (10,010.75 out of 12,367.32), it indicates that August holds a disproportionately large share of the total magnitude. The analysis is that the large difference between July and August might indicate a seasonal or monthly trend. August's magnitude is over four times that of July, which could suggest that events or factors in August significantly contribute to the overall magnitude. Investigating what occurs in August compared to other months could provide insights. This could involve exploring external factors, special events, or other variables that are unique to or peak in August.

Due to the unexpected discrepancy between the two months, additional analysis was needed to understand the full picture. Figure 5 is a bubble chart that shows the average magnitude by day against the date with the size of the bubbles indicating the number of earthquakes. This visual reveals that the apparent trend is due to a combination of factors. The dataset begins on July 27th meaning there were significantly fewer days included in the sum of the magnitude. This in combination with August 4th having a seemingly disproportionate amount of small ~1.3 magnitude earthquakes, would make it appear on other visuals that there were either more earthquakes or strong earthquakes in August. This analysis reveals how susceptible data is to being manipulated to tell a particular story. For example, imagine one had a company that could increase a home's resilience to earthquakes and they wanted to create a marketing campaign in July. They could use the graph in Figure 4, which was created with real data, to say that one should buy their products before 'earthquake season' arrives, which would be a misleading narrative.

Average Depth and Average Distance by Magnitude

Through our analysis, we gained many insights about the 'depth'. The dataset contains 6,490 records with depth measurements with a 'gap < 180' filter. The average depth varies significantly across magnitude categories, with Micro earthquakes being the shallowest on average (14.649 km) and Strong earthquakes being the deepest on average (156.747 km). This variation suggests that more minor earthquakes (like Micro) tend to occur closer to the surface, while more significant earthquakes (like Strong) often originate at greater depths.

We also gained many insights into 'distance'. The average distance shows a considerable range, being shortest for Micro earthquakes (7.333 km) and longest for Moderate earthquakes (511.397 km). This indicates that more significant magnitude earthquakes (Moderate and Strong) are detected at greater distances, possibly due to their stronger seismic waves being detectable over a larger area. The Richter Scale Name moderately drives the distance_km variable, with a 37% influence. This means the magnitude of an earthquake has a significant but not exclusive impact on how far away its effects are felt or detected. The unusually high values of distance_km for Moderate and Strong earthquakes could be attributed to their greater energy release, allowing seismic waves to travel further distances. Applying a filter for a seismic gap below 180 will focus the analysis on seismographic earthquakes that are better located. This filter may reveal more accurate trends in depth and distance relative to earthquake magnitudes.

Overall, we deduced two observations. First**,** microearthquakes, while frequent, are generally shallower and detected at shorter distances. In contrast, Strong and Moderate earthquakes tend to occur at greater depths and are detected over longer distances. Second, the relationship between earthquake magnitude and these spatial variables is significant but moderated by other factors, such as geological conditions and seismic detection capabilities. The bar chart for average distance and average depth can be seen in Figure 6. A table that correlates the magnitudes to their respective category and effects can be seen in Table 1.

Average Gap and Average Significance for Richter Scale Name

There are 6,490 records for both the ‘gap’ and ‘significance’ variables. The 'Micro' category is the most frequent, making up 78.9% of the total earthquakes. When analyzing ‘significance’, we found that the fact that the Richter Scale Name drives significance with a strength of 90% indicates a robust correlation between the magnitude of an earthquake and its overall significance. This means that as the magnitude increases, the likelihood of an earthquake being more significant in its impact also increases substantially. The average significance varies significantly across the magnitude categories, with the lowest average in the 'Micro' category (20.509) and the highest in the 'Major' category (871). The Richter Scale Name strongly influences significance, with a 90% impact, suggesting that the significance score increases substantially as the magnitude increases.

When analyzing ‘gap’, we observed that the average gap varies, being the smallest for 'Major' earthquakes (18) and most significant for 'Light' earthquakes (98.778). This suggests that 'Major' earthquakes are generally better located (smaller gap), whereas 'Light' earthquakes are less accurately located (larger gap). The bar chart for the average gap and average significance can be seen in Figure 7.

Decision Tree Models & Results

Two decision tree models were created, one for ‘significance’ and one for ‘magnitude’ which can be seen in Figure 8 and Figure 9 respectively. The predictive strength for the magnitude model was 74.9% and the strength for the significance model was 87.0%. After removing drivers that do not make logical sense for such a model, such as the second the earthquake occurred, both models were straightforward. In both instances, the state and country was the first node. This makes sense given the significant role that fault lines play in creating earthquakes.

The earthquake magnitudes and significance data correlate these two factors. Earthquakes with the highest magnitudes, which comprise 20% of the cases, tend to have the highest average significance. Conversely, those with very low significance, have the lowest magnitude. This pattern makes logical sense as a stronger earthquake is much more likely to cause damage, be felt by more people, and likely influence other factors tied to the significance rating. Another pattern revealed by the decision trees is lower gap scores produced higher magnitude and significance. This is likely due to larger earthquakes being felt by more measuring stations. Presumably, if more stations recorded the earthquake, the accuracy of its position and depth would be better and therefore the gap would be lower.

Prioritizing areas prone to higher magnitude quakes for enhanced monitoring and resource allocation is essential. Strengthening building codes and infrastructure in these regions can significantly reduce potential damage. Ongoing research and data analysis are crucial for improving earthquake prediction and early warning systems. International cooperation and preparedness for aid are necessary, especially for resource-limited regions. Tailored insurance and financial planning can help mitigate economic impacts in high-risk areas. Additionally, key strategies include building community resilience through local initiatives and incorporating earthquake risk in environmental and land-use planning. These measures collectively aim to reduce risks and enhance preparedness for regions prone to significant earthquakes.

Developed Displays

To present our findings and visualizations cohesively, we made a dashboard within Cognos Analytics. The primary goal was to organize the data and visualizations in a cohesive manner that enabled the audience to fully understand the dataset. The first tab of the dashboard gives a general overview of the dataset. This allows the viewer to be introduced to the dataset and gain an idea of what topics will be explored. The following tab covers various visuals related to the dataset's magnitudes. Given magnitude is the most frequently discussed aspect of earthquakes, it makes sense for it to follow the overview. The next two tabs cover the data presented on a map. In both instances the data is not complex; however, the volume of information being presented makes the maps strong candidates to be presented in the middle. As the decision trees are the most complex visual to understand they are the last visuals to be presented. This allows the audience to use the knowledge they have gained from the preceding tabs to understand the decision trees themselves. A slide from our dashboard can be seen in Figure 10.

Storybook

Using our dashboard, we developed and recorded a storybook, where we describe this project in its entirety. The video of this storybook can be viewed at this link: https://youtu.be/UE4NJUs4QDE. Our objective for the storybook is to tell a story about the data, and what conclusions can be drawn from the analysis. We also explain and discuss how this story could be of value, which can be read more about in detail in the following section.

Organizational Approach

The analysis of this dataset can offer valuable insights that can profoundly impact many organizations in fields such as seismology, geology, disaster management, and urban planning to name a few. Through predictive modeling of seismic activity, organizations can leverage historical and current earthquake data to develop models that can predict the likelihood of future earthquakes and monitor seismic activity. These models help take more proactive measures and allow for better resource allocation, for example, forecasting high-risk areas. Considering infrastructure planning and risk mitigation, the insights derived from this dataset, along with other similar earthquake datasets, are integral when it comes to making decisions in the urban planning sector. City planners and infrastructure developers can identify high-risk zones, guiding the design of structures that can withstand seismic events, ultimately minimizing damage and loss. Emergency response planning can benefit from special visualizations that highlight areas with a high frequency of earthquakes. This can aid teams in resource allocation, evacuation route planning, communication management, and overall preparedness when disaster strikes.

Although this analysis is rudimentary and surface level, scientific research and publications in the field of seismology can be enriched by using Cognos Analytics as they will be able to perform preliminary analyses on the magnitudes, depths, and frequencies of the earthquakes. This initial insight would guide researchers to extract valuable findings, contributing to scientific publications and advancing their understanding using specialized tools and software such as SYNthetic DEPTH Phase Modeling, which uses models made in Python for example (Earthquake Hazards - Software | U.S. Geological Survey, n.d.). Public awareness, outreach, and education initiatives can utilize the straightforward and easy-to-understand visualizations in Cognos Analytics to inform communities about earthquake-prone areas and safety measures, empowering residents to take proactive steps in disaster preparedness. Insurance companies can also use historical earthquake data and Cognos Analytics to assess risk and set premiums for properties in earthquake-prone regions. This data-driven approach ensures that insurance policies align with the actual risk levels, benefiting both insurers and property owners.

The integration of Cognos Analytics into decision-making processes extends beyond just visualizations. The predictive models derived can assist organizations in understanding the correlation between earthquake characteristics and significance. These insights can be crucial for various applications, such as those mentioned above. For risk assessment and urban planning, the decision tree can inform urban planners about the correlation between earthquake magnitude and significance, guiding decisions related to infrastructure development in earthquake-prone areas. Emergency response planning would benefit from understanding the relationship between earthquake characteristics and significance scores aids in allocating resources and planning effective emergency responses for different magnitudes. Researchers in seismology can leverage the decision tree insights to refine their studies, focusing on the correlation between earthquake magnitudes and significance. Insurance Premium Calculation: Insurance companies can use the predictive model to assess risk more accurately, setting premiums based on the correlation between earthquake characteristics and significance scores.

Furthermore, there is potential for international collaboration to be facilitated by sharing earthquake data insights across borders. The dashboards made in Cognos Analytics serve as a common platform for visualizing and sharing data, fostering collaboration in disaster response, aid, and scientific research. Overall, the multidisciplinary nature of insights gained from the earthquake dataset, coupled with the capabilities of Cognos Analytics, highlights the pivotal role of data-driven decision-making in enhancing safety, planning, research, and response strategies related to earthquakes across diverse organizations.

Conclusion

In conclusion, this paper has delved into the exploration and analysis of an earthquake dataset, revealing valuable insights about the data and practical applications for Cognos Analytics. By leveraging Cognos Analytics, this study prepared the dataset for analysis, created insightful visuals, developed predictive models, and created a dashboard for effective presentation. The dataset encompassed earthquakes from the late summer of 2016 and was generally free of missing and erroneous data. Various types of visual analysis were performed to identify relationships between an earthquake’s magnitude, depth, distance from a recording station, etc. These relationships were then fully established through the creation of decision trees. The narrow timeframe the dataset covers opens up potential opportunities for future time-based analysis. A future analysis would also benefit from additional data attributes such as the type of earthquake e.g. volcanic. This paper demonstrates the critical role that data analytics can play in areas where lives are at stake and the importance of communicating these results effectively.

References

Earthquake Hazards - Software | U.S. Geological Survey. (n.d.). www.usgs.gov. Retrieved December 9, 2023, from https://www.usgs.gov/programs/earthquake-hazards/software

Michigan Tech. (2023). *Earthquake Magnitude Scale | UPSeis*. Michigan Technological University. https://www.mtu.edu/geo/community/seismology/learn/earthquake-measure/magnitude/

Rafferty, J. P. (2018). Richter scale | Definition & Facts. In Encyclopædia Britannica. https://www.britannica.com/science/Richter-scale

Russell, J. (2020, May 12). Exploring Earthquake Magnitude and Depth. Joshua Russell. https://jbrussell.github.io/eilive2020/part1a3_magnitudedepth/

USGS. (2011). *At what depth do earthquakes occur? What is the significance of the depth?* Usgs.gov. https://www.usgs.gov/faqs/what-depth-do-earthquakes-occur-what-significance-depth

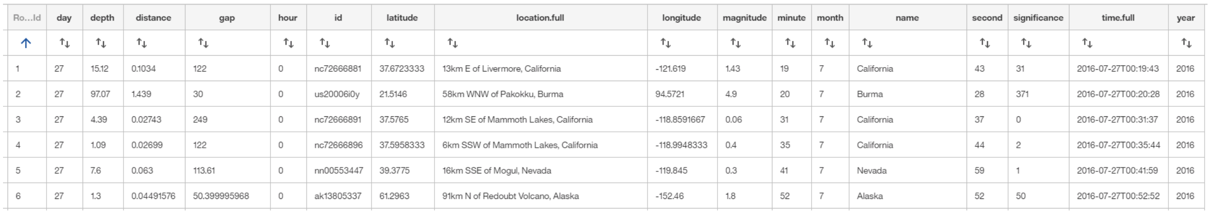
Whitcomb, R. (2016, June 7). *Earthquakes CSV File*. Think.cs.vt.edu; CORGIS Dataset Project. https://think.cs.vt.edu/corgis/csv/earthquakes/

Appendix A

All tables and figures mentioned throughout the paper can be found here.

Figure 1

A Snippet of the Earthquake Dataset



Ro...Id	day	depth	distance	gap	hour	id	latitude	location.full	longitude	magnitude	minute	month	name	second	significance	time.full	year
↑	↑↓	↑↓	↑↓	↑↓	↑↓	↑↓	↑↓	↑↓	↑↓	↑↓	↑↓	↑↓	↑↓	↑↓	↑↓	↑↓	↑↓
1	27	15.12	0.1034	122	0	nc72666881	37.6723333	13km E of Livermore, California	-121.619	1.43	19	7	California	43	31	2016-07-27T00:19:43	2016
2	27	97.07	1.439	30	0	us20006i0y	21.5146	58km WNW of Pakokku, Burma	94.5721	4.9	20	7	Burma	28	371	2016-07-27T00:20:28	2016
3	27	4.39	0.02743	249	0	nc72666891	37.5765	12km SE of Mammoth Lakes, California	-118.8591667	0.06	31	7	California	37	0	2016-07-27T00:31:37	2016
4	27	1.09	0.02699	122	0	nc72666896	37.5958333	6km SSW of Mammoth Lakes, California	-118.9948333	0.4	35	7	California	44	2	2016-07-27T00:35:44	2016
5	27	7.6	0.063	113.61	0	nn00553447	39.3775	16km SSE of Mogul, Nevada	-119.845	0.3	41	7	Nevada	59	1	2016-07-27T00:41:59	2016
6	27	1.3	0.04491576	50.399995968	0	ak13805337	61.2963	91km N of Redoubt Volcano, Alaska	-152.46	1.8	52	7	Alaska	52	50	2016-07-27T00:52:52	2016

Figure 2

Scatterplot of Average Magnitude v. Average Corrected Depth

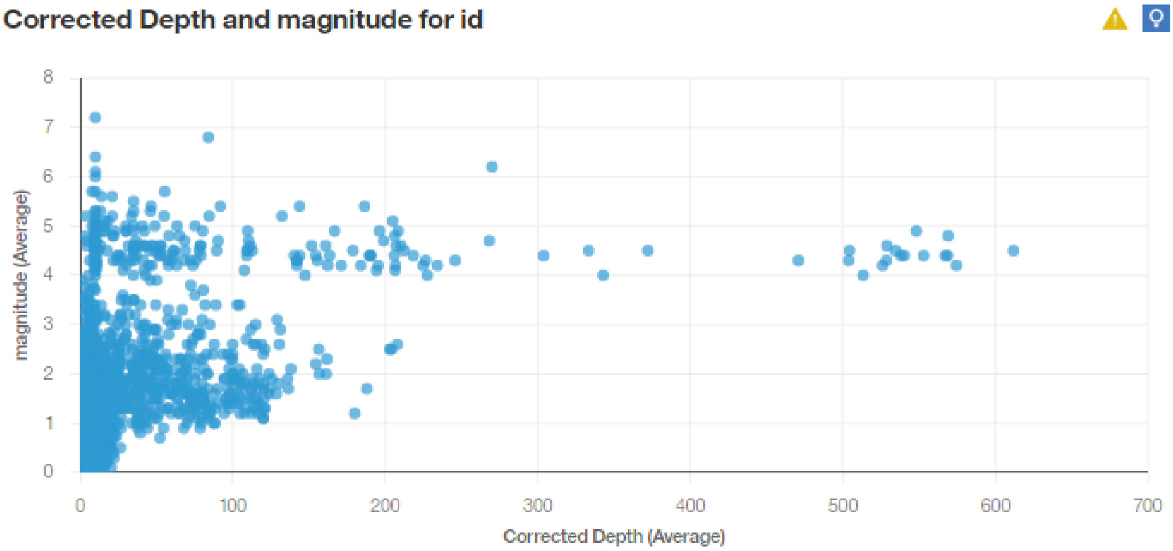


Figure 3

Map Visualization of the Three Most Common States/Countries

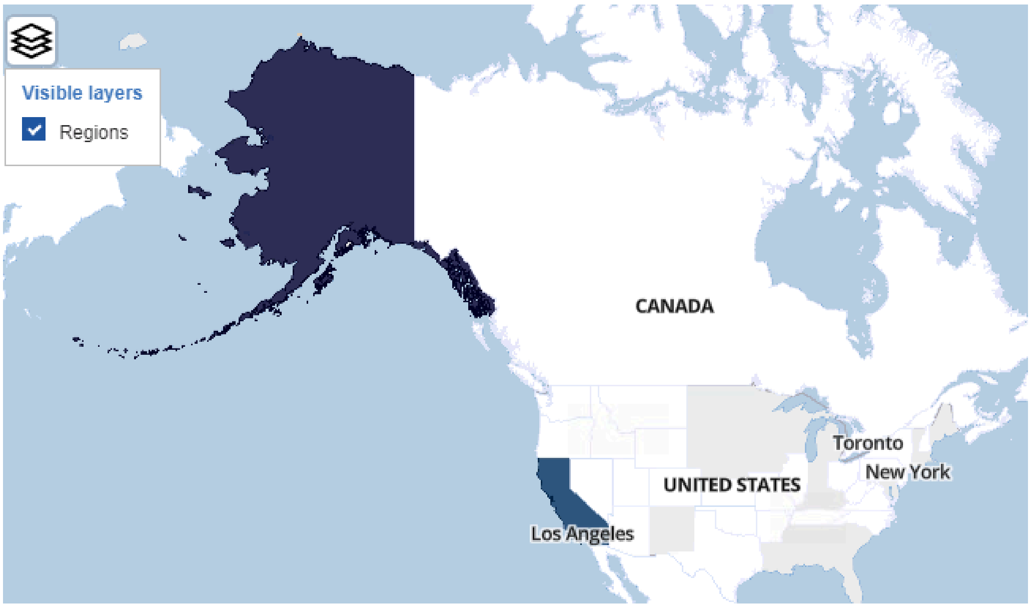
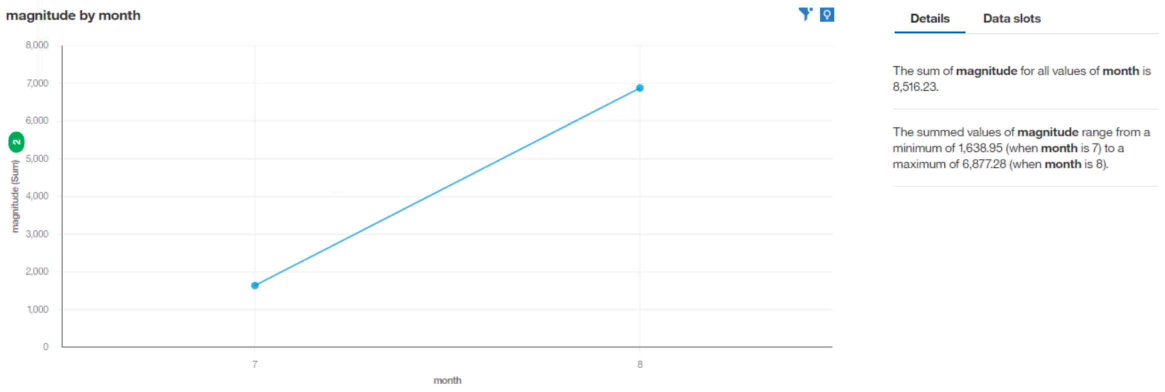


Figure 4

Line Graph of Magnitude v. Month



The sum of **magnitude** for all values of **month** is 8,516.23.

The summed values of **magnitude** range from a minimum of 1,638.95 (when **month** is 7) to a maximum of 6,877.28 (when **month** is 8).

Figure 5

Bubble Chart of Average Magnitude and Number of Earthquakes per Day

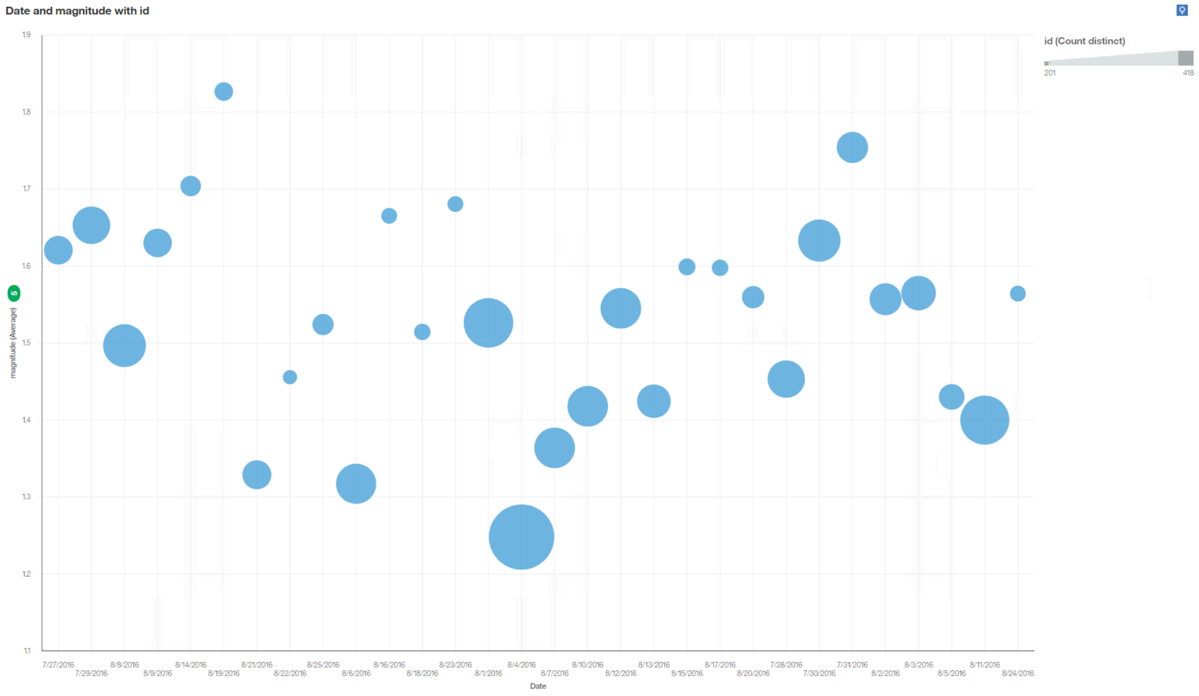


Figure 6

Bar Chart of Earthquake Distance and Depth v. Richter Scale Name

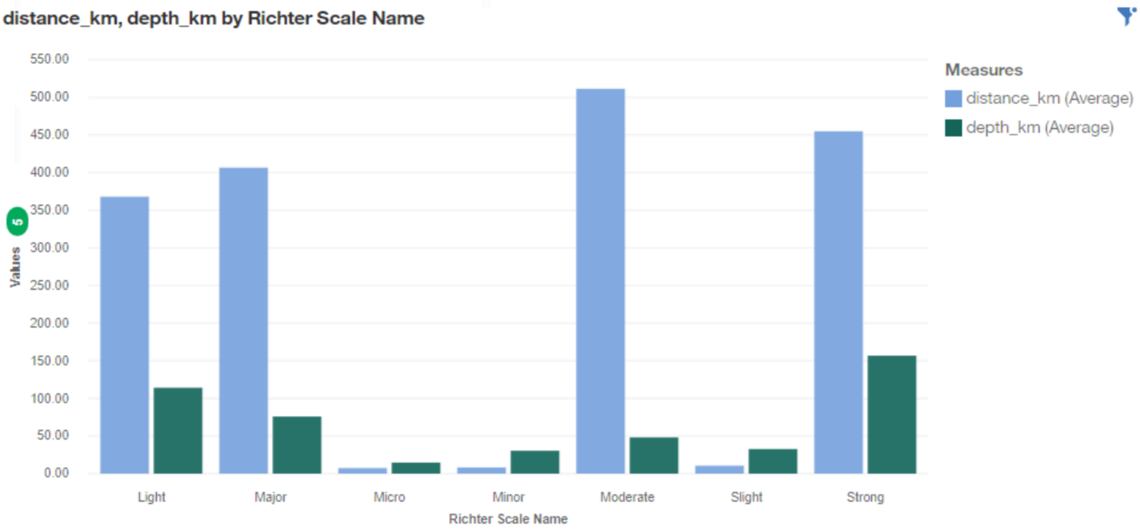


Figure 7

Visual of Average Significance and Average Gap v. Richter Scale Name

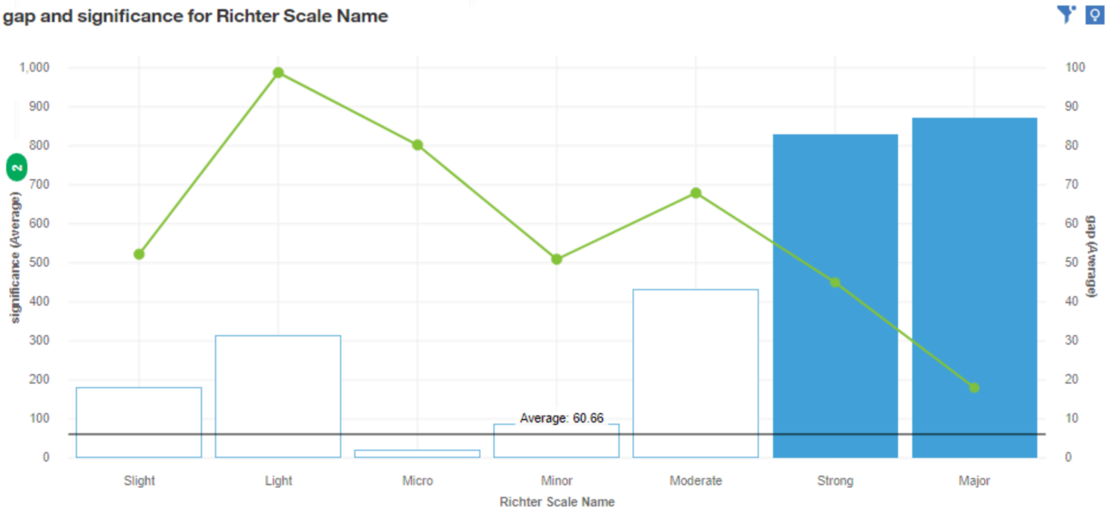
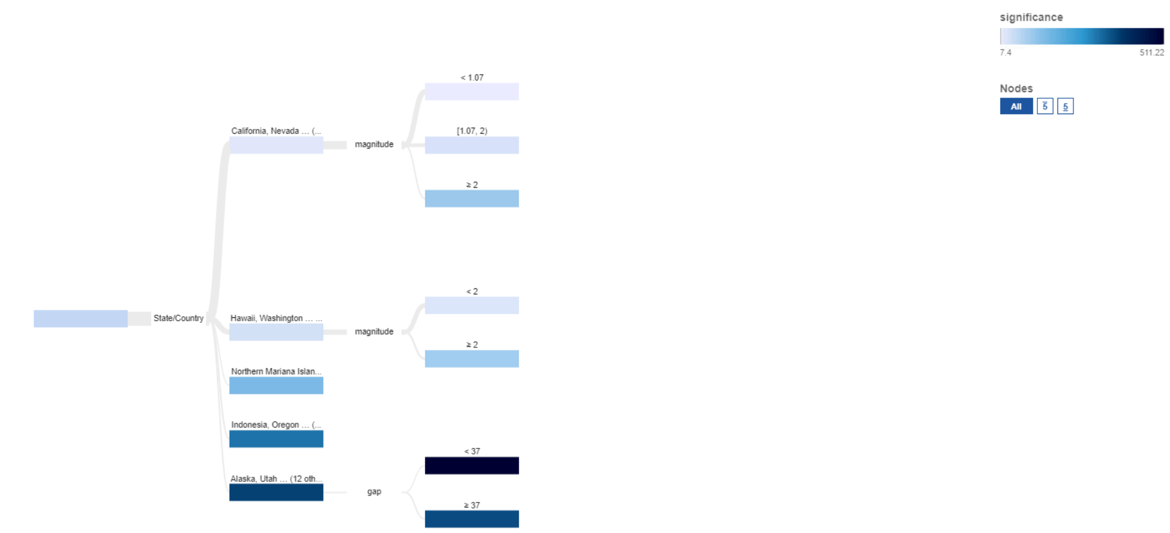


Figure 8

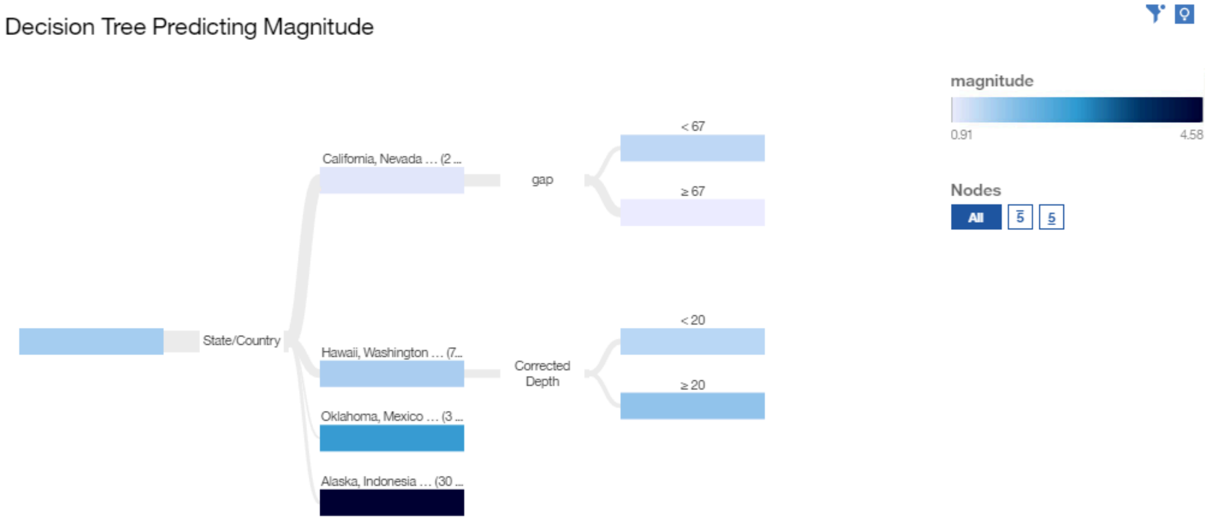
Significance Decision Tree



significance
7.4
511.22
Nodes
All
5̄
5̲
State/Country
California, Nevada ... (...
magnitude
< 1.07
[1.07, 2)
≥ 2
Hawaii, Washington ... ...
magnitude
< 2
≥ 2
Northern Mariana Islan...
Indonesia, Oregon ... (...
Alaska, Utah ... (12 oth...
gap
< 37
≥ 37

Figure 9

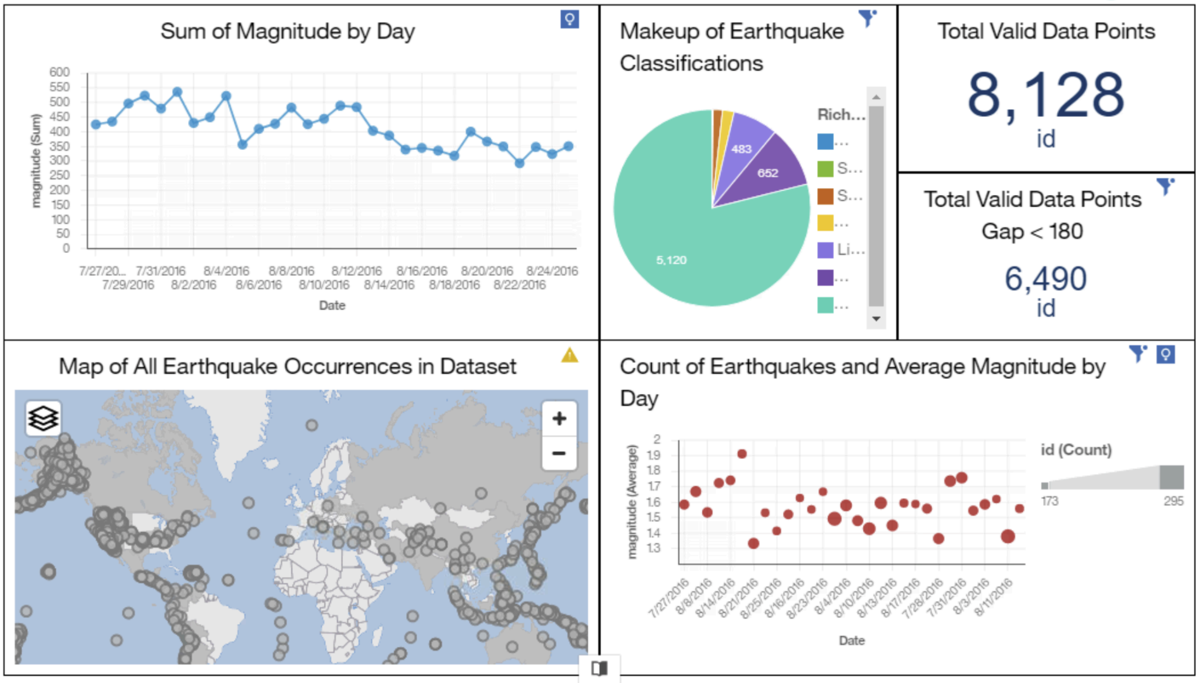
Magnitude Decision Tree



Decision Tree Predicting Magnitude
magnitude
0.91
4.58
Nodes
All
5̄
5̲
State/Country
California, Nevada ... (2 ...
gap
< 67
≥ 67
Hawaii, Washington ... (7...
Corrected Depth
< 20
≥ 20
Oklahoma, Mexico ... (3 ...
Alaska, Indonesia ... (30 ...

Figure 10

Overview Tab from the Developed Dashboard



Sum of Magnitude by Day

Makeup of Earthquake Classifications

Total Valid Data Points

8,128

id

Total Valid Data Points Gap < 180

6,490

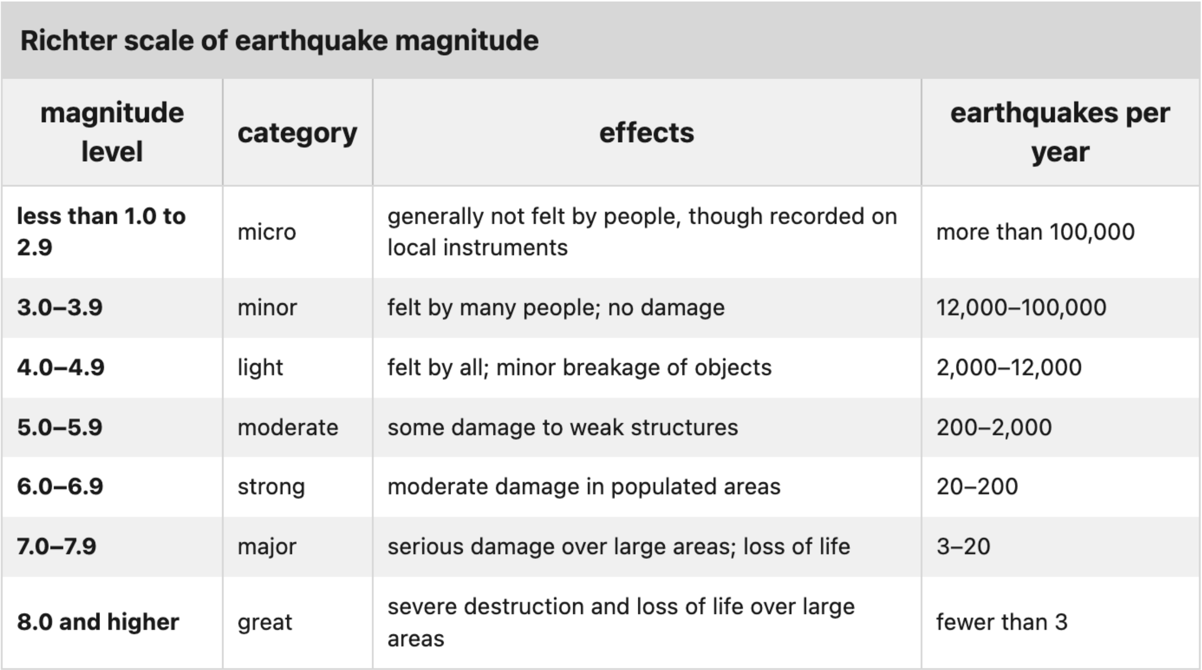
id

Map of All Earthquake Occurrences in Dataset

Count of Earthquakes and Average Magnitude by Day

Table 1

Table of the Richter scale Magnitude Mapped to its Respective Category



Richter scale of earthquake magnitude

magnitude level	category	effects	earthquakes per year
less than 1.0 to 2.9	micro	generally not felt by people, though recorded on local instruments	more than 100,000
3.0–3.9	minor	felt by many people; no damage	12,000–100,000
4.0–4.9	light	felt by all; minor breakage of objects	2,000–12,000
5.0–5.9	moderate	some damage to weak structures	200–2,000
6.0–6.9	strong	moderate damage in populated areas	20–200
7.0–7.9	major	serious damage over large areas; loss of life	3–20
8.0 and higher	great	severe destruction and loss of life over large areas	fewer than 3

Note. (Rafferty, 2018)