Connor Curtis

STT 180 - Section 001

April 29, 2025

What is the Magnitude of Magnitude?

Investigating how earthquake magnitude influences, and is influenced, by other variables

Introduction:

Though often a forgotten phenomenon, earthquakes still occur all the time around the world. These earthquakes vary in magnitude and location, and also have varying effects on their surroundings. For this project, the dataset that I will be using is from Kaggle and it is titled "Global Earthquake Data." The dataset contains 1137 entries and 43 columns of various variables related to earthquake occurrences that range from June 2023 to August 2024 which makes this a fairly recent dataset. However, only a select number of the variables present within the dataset will be used in later analysis as many of them don't provide insightful conclusions about the data or are generally rather messy. Finally, all of the data found in this dataset was collected using the EveryEarthquake API from RapidAPI.

The main goal of this study is to uncover the relationships that exist between earthquake magnitude and other variables in the dataset such as depth, location, whether the earthquake triggered a tsunami warning, and a few others. By investigating the relationships between these variables and magnitude, I will aim to develop a better understanding about how magnitude is influenced by some variables, and how it influences others. Additionally, I aim to use each of the investigated variables to create a linear model that could be used to predict the magnitude of an earthquake. Finally, in order to round out the understanding of how certain variables are connected to magnitude, I will conduct hypothesis testing for some variables in order to confidently conclude how they affect each other.

Research Questions and Methods:

1. What variables influence magnitude, and how?
    a. Data visualization
    b. Correlation analysis

2. What variables are influenced by magnitude, and how?
    a. Data visualization
    b. Correlation analysis

3. How can these variables be used to create a linear model that can predict magnitude?
    a. Linear regression
    b. P-value analysis

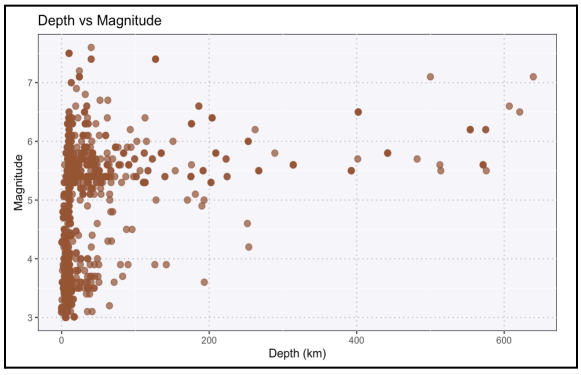
4. What conclusions can be solidified about some variable relationships via inference testing.
    a. Data visualization
    b. Hypothesis testing
    c. P-value analysis

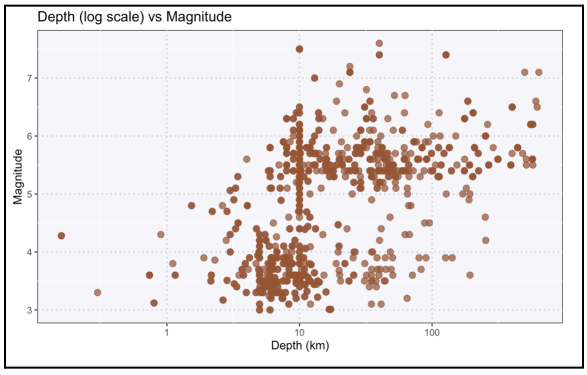
Stated above are the four questions that I will be answering within this project, and a brief description of the methods that I will be employing in order to answer each question. As can be seen, I plan to create many visuals using the 'ggplot2' package in which different variables are plotted against each other as this is an effective way to visibly see the relationships between variables rather than just describing the relationship via numbers. However, when used in tandem with visuals, numbers are also very helpful at quantifying the relationships between variables which is why I also plan to analyze correlation coefficients and p-value in order to gain a complete understanding of how magnitude is related to other variables within my dataset. Finally, many conclusions can be drawn via correlation and visualization, but I will also employ hypothesis testing in order to extract more solid conclusions about certain relationships from my dataset.

Results:

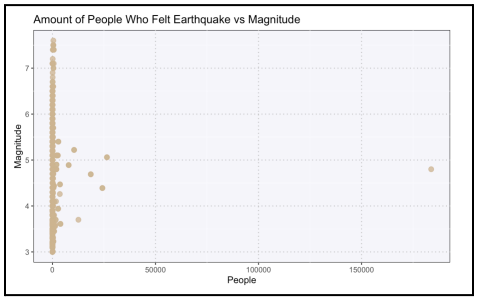
Question 1:

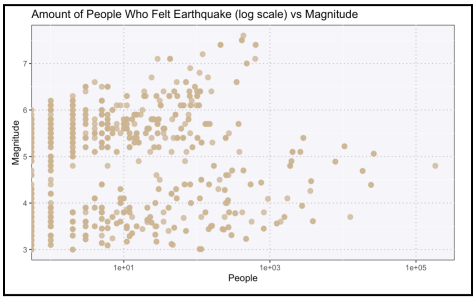
My first aim was to investigate which variables influence magnitude and how they do so. For this question, the variables that I chose to examine were earthquake depth, the amount of people that felt the earthquake, the latitude and longitude of an earthquake, and the continent in which an earthquake occurred. So, to start I created a visual depicting depth vs magnitude.



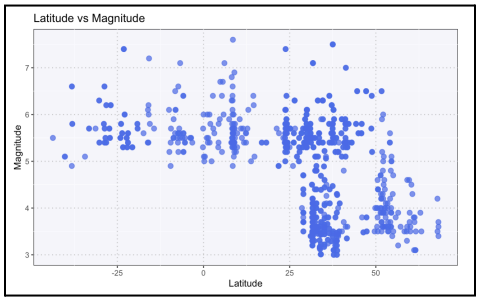


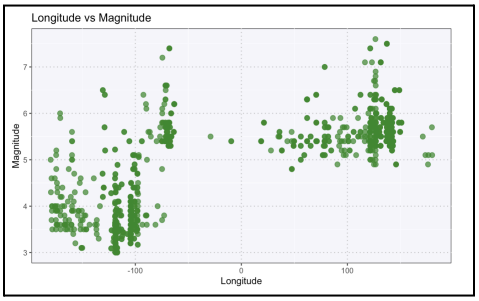
Initially, the plot of depth vs magnitude didn't reveal that much information as most of the data points were clustered along the left edge, but it still yielded a correlation coefficient of 0.310 which indicated there was some positive correlation between the two variables. However, as evident in the second figure, I then plotted depth vs magnitude but opted to use a logarithmic scale for the x-axis (depth). This yielded a much better looking plot with data points more spread out. There appears to be somewhat of a trend in which magnitude increases as depth increases according to this visualization which is supported by its correlation coefficient of 0.481 which is an increase compared to the first graph. After analyzing how depth affects magnitude, I then created a visual of the amount of people who felt an earthquake vs its magnitude.



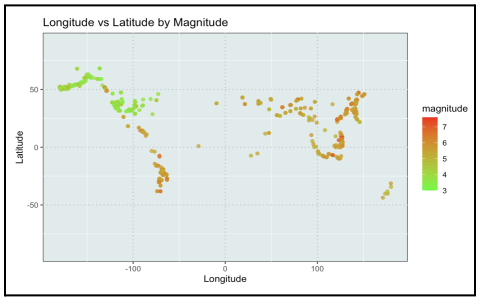


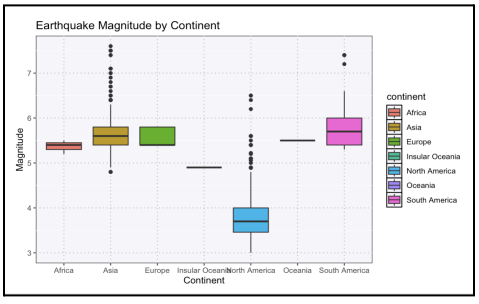
The methods I used for visualizing the amount of people who felt the earthquake vs the magnitude were identical to that of depth. The initial visual didn't reveal much information as all of the data was clustered on the left side of the graph, so in the second figure I once again employed a logarithmic scale. The respective correlation coefficients for these visuals were -0.008 and 0.015. So, although the correlation increased slightly after using a logarithmic scale on the amount of people who felt the earthquake, it still remained very close to zero indicating little correlation. I would attribute this to the fact that the amount of people who feel an earthquake likely relies on whether it occurs in a populated area, not its magnitude. After this, I then created visualizations depicting latitude vs magnitude and longitude vs magnitude.





Observing these plots, it would appear that magnitude is negatively correlated with latitude, and positively correlated with longitude. This is supported by their respective correlation coefficients of -0.410 and 0.747 with the coefficient of longitude indicating a strong positive correlation seeing its close proximity to 1. These coefficients support the idea that magnitude tends to decrease with latitude, and tends to increase with longitude. Finally, I then took a closer look into the influence of latitude and longitude as well as how continents influence magnitude.

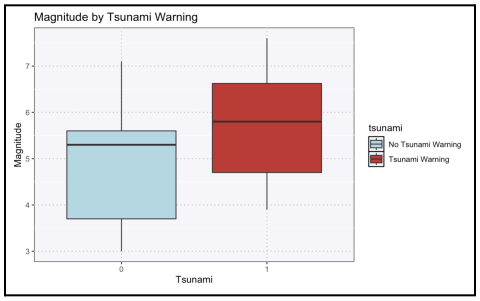


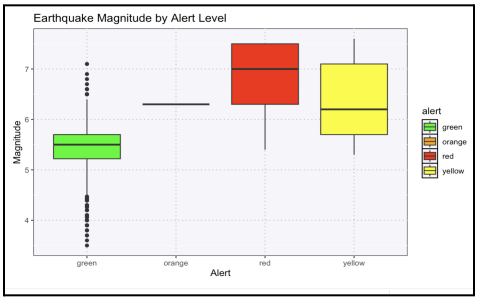


By plotting longitude and latitude against each other, a visual that resembles the world map is produced. Furthermore, by adjusting the color of each earthquake data point based on its magnitude, we visually see where the highest magnitude earthquakes occur. The figure indicates that the highest magnitude earthquakes occur in South America and parts of Asia, with the earthquakes of the lowest magnitude occurring in North America. These observations are further supported by the boxplot in the second figure which highlights that South America and Asia do indeed see highest magnitude earthquakes with Europe, Africa, and Oceania seeing slightly lower magnitudes, and finally Insular Oceania and North America seeing earthquakes of the lowest magnitude.

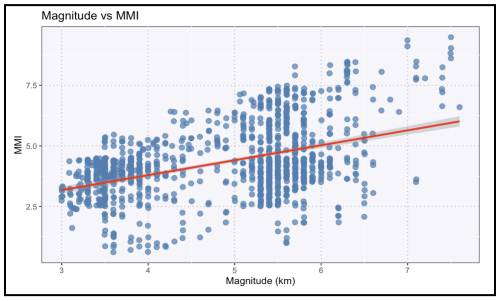
Question 2:

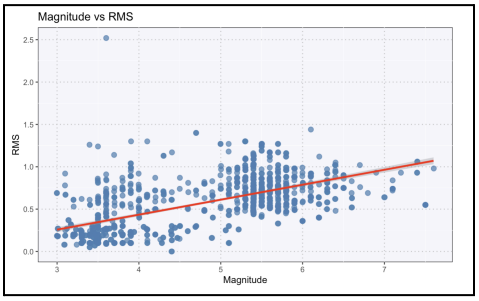
My second aim was to determine how magnitude influences other variables in the dataset which was done through similar methods of visualizations and correlation analysis. To start, I looked at how magnitude influences whether tsunami warnings occur, and also how magnitude impacts and earthquake's alert level.





As can be seen from the first figure, the box plot indicates that earthquakes which triggered a tsunami warning do tend to have higher magnitudes which is evident from the higher median line and the higher quartiles. Looking at the second figure, it's important to note that the severity order of earthquake alerts is actually green, yellow, orange, and red. Knowing this, it can be seen that as earthquake magnitude increases, we tend to see more severe alert levels with green having the lowest magnitude and red having the highest magnitude. The median for orange alert levels is higher than yellow alert levels which makes sense, but it's hard to compare orange with anything as its box is simply a line due to the significantly lower amount of orange occurrences within the dataset. After this, I plotted magnitude vs MMI and magnitude vs RMS. MMI and RMS are both different ways to measure the seismic activity of an earthquake. Specifically, MMI estimates the shaking intensity of an earthquake and RMS measures the average magnitude of the seismic signal over a period of time.





From these graphs, we can see a general positive correlation for both MMI and RMS as they tend to increase as magnitude increases. The included line of best fit, which was created using the 'geom_smooth' function, supports this relationship, and the correlation coefficients of each figure support this relationship as well being 0.445 and 0.600 respectively.

Question 3:

Having analyzed the individual relationships that magnitude has with each of the aforementioned variables, I then use linear regression in order to create a model that could be used to predict magnitude using these variables. To do so, I used the *lm( )* function which is part of base R, and the *step( )* function, which is also a part of base R, in order to find the most optimal linear regression model. The *step( )* function creates the best model using the available variables based on a specific set of criteria known as the Akaike Information Criterion (AIC).

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	3.521876e+00	2.433319e-01	14.4735490	2.656170e-40
log10(depth)	5.321066e-01	4.889706e-02	10.8821797	4.733673e-25
felt	-3.222540e-06	1.879838e-06	-1.7142646	8.705931e-02
longitude	5.409444e-03	6.748314e-04	8.0159936	6.905733e-15
latitude	9.739364e-04	1.292111e-03	0.7537561	4.513273e-01
continentAsia	-3.464113e-01	2.127180e-01	-1.6285002	1.040075e-01
continentEurope	6.980420e-02	2.471425e-01	0.2824452	7.777114e-01
continentInsluar Oceania	-1.041157e+00	4.117334e-01	-2.5287158	1.173443e-02
continentInsular Oceania	-1.411381e+00	4.128133e-01	-3.4189337	6.765972e-04
continentNorth America	6.460668e-02	2.261792e-01	0.2856438	7.752615e-01

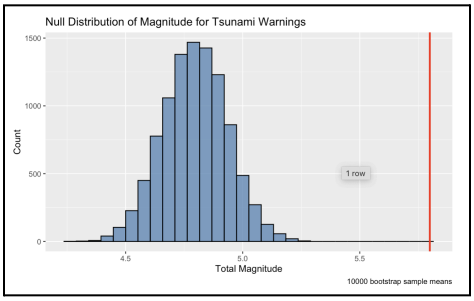
term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
continentOceania	-8.180016e-01	4.120061e-01	-1.9854112	4.760884e-02
continentSouth America	6.649365e-01	2.239136e-01	2.9696123	3.115527e-03
mmi	1.755583e-01	1.722110e-02	10.1943727	2.016541e-22
rms	1.628428e-01	6.968771e-02	2.3367509	1.982025e-02
tsunami1	7.084819e-01	6.695999e-02	10.5806750	6.938996e-24
alertorange	7.100169e-01	1.517841e-01	4.6778071	3.677005e-06
alertred	6.425871e-01	1.291590e-01	4.9751634	8.800286e-07
alertyellow	2.848135e-01	6.998735e-02	4.0694994	5.422169e-05

The coefficients of each variable and their corresponding p-value for the final model are depicted in the two figures above. Interestingly, there are many variables that the *step( )* function included in its final model that have a p-value greater than 0.05 which would indicate that they are not statistically significant. However, a majority of the included variables have a p-value that is lower than 0.05, with some variables having an extremely low p-value. These variables include depth, longitude, MMI, and tsunami warning with each of their respective p-values being 4.733e-25, 6.906e-15, 2.017e-22, and 6.939e-24. The coefficients for these variables also hold meaning. For depth, each unit that the log10 of depth increases magnitude increases by 0.5321. For longitude, as longitude increases by one degree, magnitude also increases by .005. Additionally, for each unit that MMI increases by, magnitude increases by 0.176. Finally, when there is a tsunami warning present, magnitude increases by 0.708. The coefficients for every variable in the model are able to be interpreted, but for the sake of being concise, I will only be personally interpreting these four rather than all 18. In the end, this linear model had an adjusted r-squared value of 0.785 which indicates that this model explains 78.5% of the variance in earthquake magnitude within the data.

Question 4:

The final question of this project aimed to evaluate the statistical significance of the relationships between some variables using hypothesis testing at a 0.05 significance level and subsequent p-value analysis. I chose to further examine the relationships between magnitude and tsunami warnings, as well as how earthquake magnitude varied across continents. The null hypothesis, alternative hypothesis, and null distribution for my first test are shown below. My reasoning for using a one-tailed test is due to the fact that the tsunami warning boxplot shown in question 2 highlighted that earthquakes that resulted in a tsunami warning tended to have a higher magnitude.

- **Null Hypothesis:** Earthquakes that do result in a tsunami warning have the same average magnitude compared to earthquakes that do not result in a tsunami warning.
- **Alternative Hypothesis:** Earthquakes that do result in a tsunami warning have a greater average magnitude than those that do not result in a tsunami warning.



The above figure displays the null distribution of earthquakes which resulted in a tsunami centered at the average magnitude of earthquake that didn't result in a tsunami warning. Furthermore, the red line shows the average magnitude of earthquakes that did result in a tsunami warning. After carrying out the hypothesis test, the resultant p-value was 0. This value is obviously lower than the significance level of 0.05, which allows us to reject the null hypothesis. Therefore, we can conclude that any variation that is present between the magnitude of earthquakes that do result in a tsunami warning and earthquakes that don't result in a tsunami warning is statistically significant. Having concluded this test, the null and alternative hypothesis are as follows:

- **Null Hypothesis:** The average magnitude of earthquakes in South America is equal to the mean magnitude of earthquakes in other continents.

- **Alternative Hypothesis:** Earthquakes in South America have a greater average magnitude compared to earthquakes in the other continents.

In order to carry out this test, I conducted individual hypothesis tests comparing average earthquake magnitude in South America to every other continent. My reasoning for using a one-tailed test was from the boxplot show in question 1 that depicted South America having earthquakes of the greatest magnitude. The resulting p-values for each test are as follows:

continent <chr>	p_value <dbl>
North America	0.0000
Africa	0.0000
Europe	0.0001
Asia	0.0294
Oceania	0.0000

As you can see from the above table, the p-value for each test was lower than the significance level of 0.05 which allows us to reject the null hypothesis. Therefore, we can conclude that any variation in magnitude between the earthquakes in South America and other continents is statistically significant.

Conclusion:

Findings Summary:

Overall, I found some decent correlations between many of the variables and magnitude. Some of those include the positive correlations of depth, longitude, MMI, and RMS at 0.481, 0.747, 0.445, and 0.6. Also some negative correlations were found such as latitude at -0.410. Despite this, each relationship that was explored provided valuable information no matter what the correlation coefficient was as they gave insight into how magnitude is influenced by other variables, and influences other variables. All of these variable relations were used to create a linear model that can be used to predict the magnitude of an earthquake. This model yielded an adjusted r-squared value of 0.785 which is quite good as this concludes the model explains 78.5% of variation in earthquake magnitude within the dataset. Finally, after conducting two hypothesis tests we were able to gain an even deeper understanding of some variable relationships. Specifically, since every test resulted in a p-value less than 0.05 we were able to concurrently conclude that any variation that is present between the magnitude of earthquakes that do result in a tsunami warning and earthquakes that don't result in a tsunami warning is statistically significant, and any variation in magnitude between the earthquakes in South America and other continents is also statistically significant.

Limitations and Method Critiques:

The main limitation of this dataset is simply the fact that its entries are only from June, 2023 to August, 2024. Though there was certainly a decent amount of data to work with, it would have been beneficial to have even more that dated back further into the past as this would provide richer insights. However, this shorter time frame could also be viewed as an advantage as any insights we found are likely more applicable to present day earthquakes. Additionally, some variables were also limited due to how they were represented in the data. For example, the way that earthquake dates and titles are structured contain various assortments of numbers and letters making them hard to use without some serious data cleaning.

Overall, I believe my methods were straightforward and sound. I don't see many things that I would change in the future as I think my methods helped gather many insights from the data in a logical, easy-to-understand manner. One thing that I could improve would be planning out such projects a bit better so I know exactly what variables I will touch on, but this wasn't a big issue and it didn't affect my end result.

Future Questions:

There are many other questions pertaining to this dataset that could be investigated in the future due to its expansive number of variables. The questions that I investigated revolved around magnitude; however, in the future I could repeat the questions I used but apply them to any other variable that is present in the data. For example, could I investigate how variables such as earthquake location, earthquake depth, and number of earthquake detection stations affect the accuracy of earthquake detection? Additionally, I could investigate how earthquake damage varies depending on the location.

References:

- Shreya Sur965. (2024). Global Earthquake Data. Kaggle.com. https://www.kaggle.com/datasets/shreyasur965/recent-earthquakes
- finnstats. (2022, July 14). How to Change Background Color in ggplot2? | R-bloggers. https://www.r-bloggers.com/2022/07/how-to-change-background-color-in-ggplot2-3/
- Awesome List Of 657 R Color Names You Need to Know. (2018, November 17). Datanovia. https://www.datanovia.com/en/blog/awesome-list-of-657-r-color-names/
- Jitter points to avoid overplotting — position_jitter. (n.d.). Ggplot2.Tidyverse.org. https://ggplot2.tidyverse.org/reference/position_jitter.html
- Create your own discrete scale — scale_manual. (n.d.). Ggplot2.Tidyverse.org. https://ggplot2.tidyverse.org/reference/scale_manual.html
- Norhther. (2017, December 19). Logarithmic scale plot in R. Stack Overflow. https://stackoverflow.com/questions/47890742/logarithmic-scale-plot-in-r
- Background color in ggplot2. (2021, February 6). R CHARTS | a Collection of Charts and Graphs Made with the R Programming Language. https://r-charts.com/ggplot2/background-color/
- Soetewey, A. (2020, May 28). Correlation coefficient and correlation test in R. Stats and R. https://statsandr.com/blog/correlation-coefficient-and-correlation-test-in-r/
- Making Sense of the Modified Mercalli Intensity Scale (MMI) -A Measure of Shaking. (n.d.). https://abag.ca.gov/sites/default/files/making_sense_of_the_modified_mercalli_intensity_scale.pdf
- RMS amplitude - SEG Wiki. (n.d.). Wiki.seg.org. https://wiki.seg.org/wiki/RMS_amplitud