Earthquake_Capstone_Report

H. Ewton

9/28/2018

Table of Contents

Contents

The Problem

One of the biggest problems that exists in geology involves the prediction of significant earthquakes. Earthquakes can be measured in several ways, but a significant earthquake is a tremor that measures 5.5 or higher on the Richter scale. Predicting such earthquakes is important as higher magnitude earthquakes often cause significant damage to infrastructure and loss of life. Ideally, a solution could be found using data to predict the probability of occurance of a significant earthquake for a given location.

To solve this problem, a significant amount of data would be needed to enable the creation of a probability prediction. The data would not only need to have all earthquakes registering above a 5.5 magnitude, but would also need the date the earthquake occurred, the geographical location (latitude/longitude), the focal depth of the quake (the origin of the tremor), and the plate that the tremor occurred on/along. This data would then be analyzed first for trends, then used to predict the probability of an earthquake occurring on each plate.

Methods

Data Sources

Three data sets were used in this analysis. The first data set, signif_earthquakes, was obtained from NOAA's Significant Earthquakes Database (https://www.ngdc.noaa.gov/nndc/struts/form?t=101650&s=1&d=1) and lists every recorded earthquake in history back to 2150 BC. The other data set used, USGS_df, was obtained from Kaggle (https://www.kaggle.com/usgs/earthquake-database#database.csv) and lists all recorded major earthquakes in the USGS data base from 1965 to 2016. Additionally, a map of tectonic plate points was used to assign the latitude and longitude points to a tectonic plate (https://github.com/fraxen/tectonicplates/tree/master/GeoJSON).

Signif_earthquakes clean-up

This file, signif_earthquakes, presented several challenges. After removing observations with missing magnitude values and data with estimated magnitude (pre-1935), the data was filtered to only include relevant columns: date, time, magnitude, and location information.

USGS_df Clean-up

The next step in the cleaning of data was to address the USGS data set from Kaggle. Like signif_earthquakes, the columns relevant to this analysis first had to be extracted. From USGS_df, selected columns were "Date", "Time", "Latitude", "Longitude", "Depth", and "Magnitude". The selection of these columns allows us to complete our data analysis as well as join the columns together. After selecting the relevant columns, the date column was reformatted to international date format and the “depth” column was renamed to “focal_depth” to better indicate what the values represent.

Joining the data frames

To best join the data frames together without losing data, a full_join function was used. In this process, the tables were joined by date, time, latitude, longitude, magnitude, and focal depth of the earthquakes.

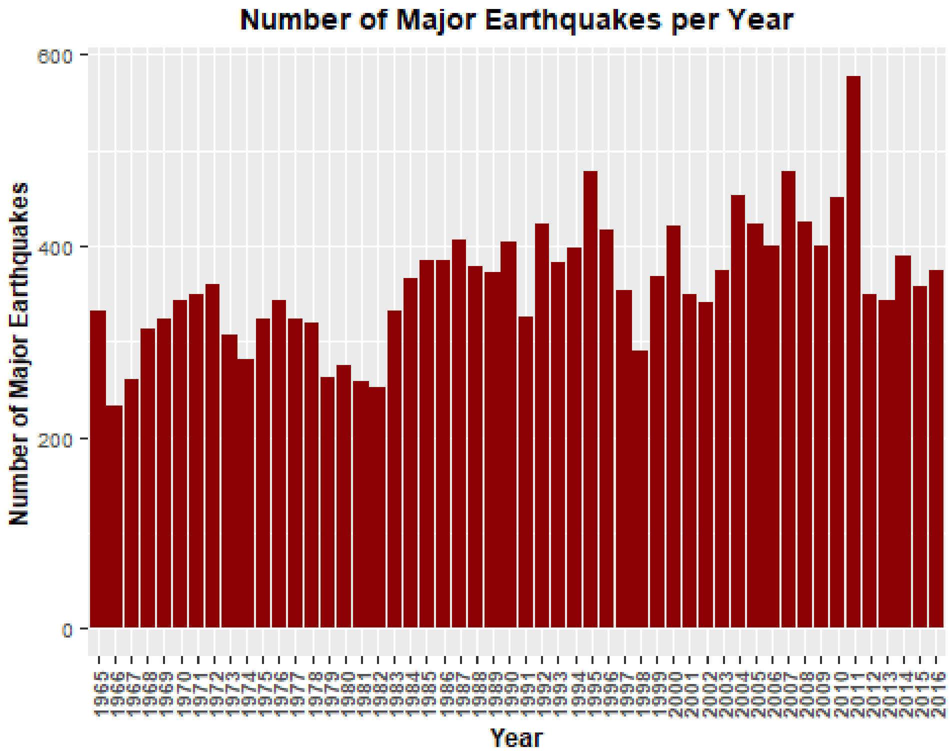
Adding plate information

Once the data set was created, one more bit of information was needed for analysis- the tectonic plate data. Stored as a JSON file, this data set contained the plate boundaries of each plate by connecting a series of coordinates. Once loaded, the earthquakes data set was overlaid onto the plate boundaries data set.

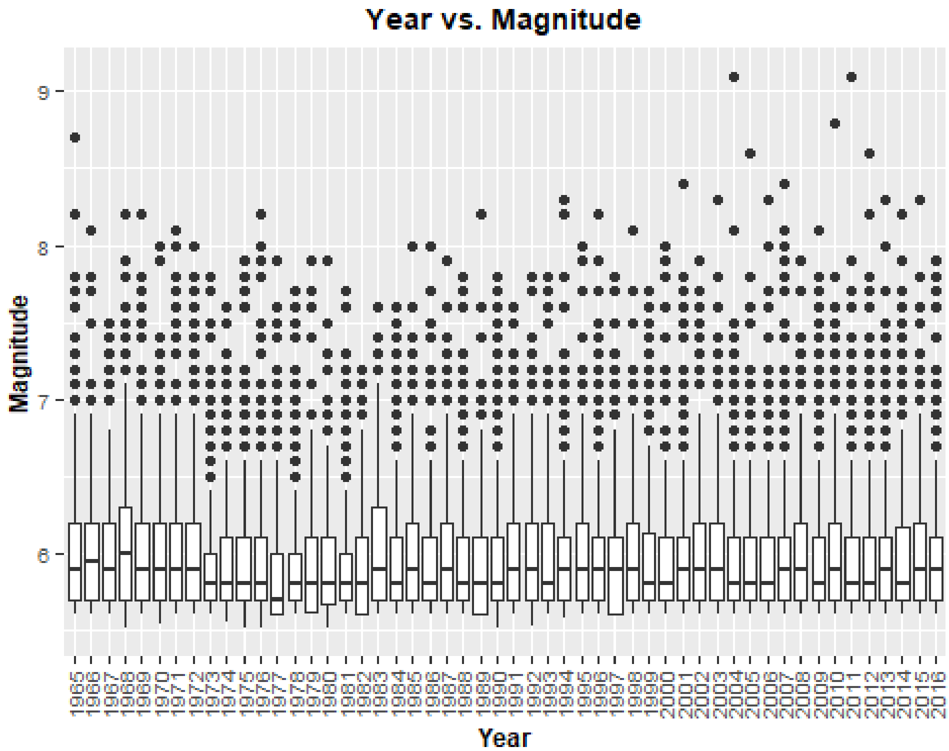
year	month	day	time	Focal depth	magnitude	country	Location name	Lati-tude	Long-itude	LAYER	Code	Plate Name
1965	3	28	16:33:00.0	61	7.3	CHILE	CHILE: CENTRAL	-32.4	-71.2	plate	SA	South America
1965	3	31	09:47:00.0	78	7.1	GREECE	GREECE	38.6	22.4	plate	EU	Eurasia
1965	4	29	15:28:43.7	59	6.5	USA	WASHINGTON: SEATTLE	47.4	-122.3	plate	NA	North America
1965	6	21	00:21:00.0	40	6.0	IRAN	IRAN: HADJIABAD, SARKHUN, SARCHAHAN	28.1	55.9	plate	EU	Eurasia
1966	3	7	01:16:00.0	38	6.0	TURKEY	TURKEY: VARTO, MUS	39.1	41.6	plate	EU	Eurasia
1966	8	15	02:15:00.0	53	5.6	INDIA	INDIA: N	28.7	78.9	plate	IN	India

Exploratory Data Analysis

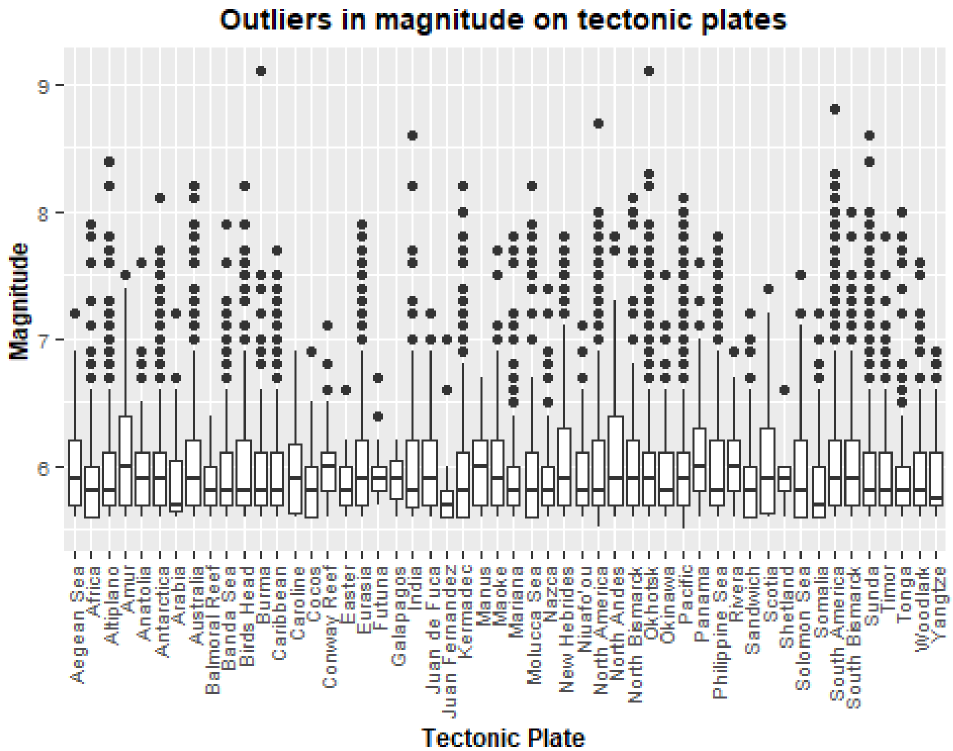
After combining the data sets, an initial exploration was completed to identify any possible trends. Using the earthquakes, data set, a bar graph was created for year vs # of major earthquakes from 1965 through 2016. The bar graph yielded the following results, with a trend of an increasing number of earthquakes worldwide. A peak number of earthquakes appears in 2011, which had nearly 100 more significant earthquakes than any other year in recorded seismic history.



To get more detail on the magnitude of earthquakes that occurred by year, a boxplot was created for year against magnitude. Major earthquake outliers above a 9.0 magnitude appeared in both 2004 and 2011. However, as the data reveals, although there were more earthquakes in 2011, the majority of the earthquakes were within the 5.5 to 6.5 magnitude range.



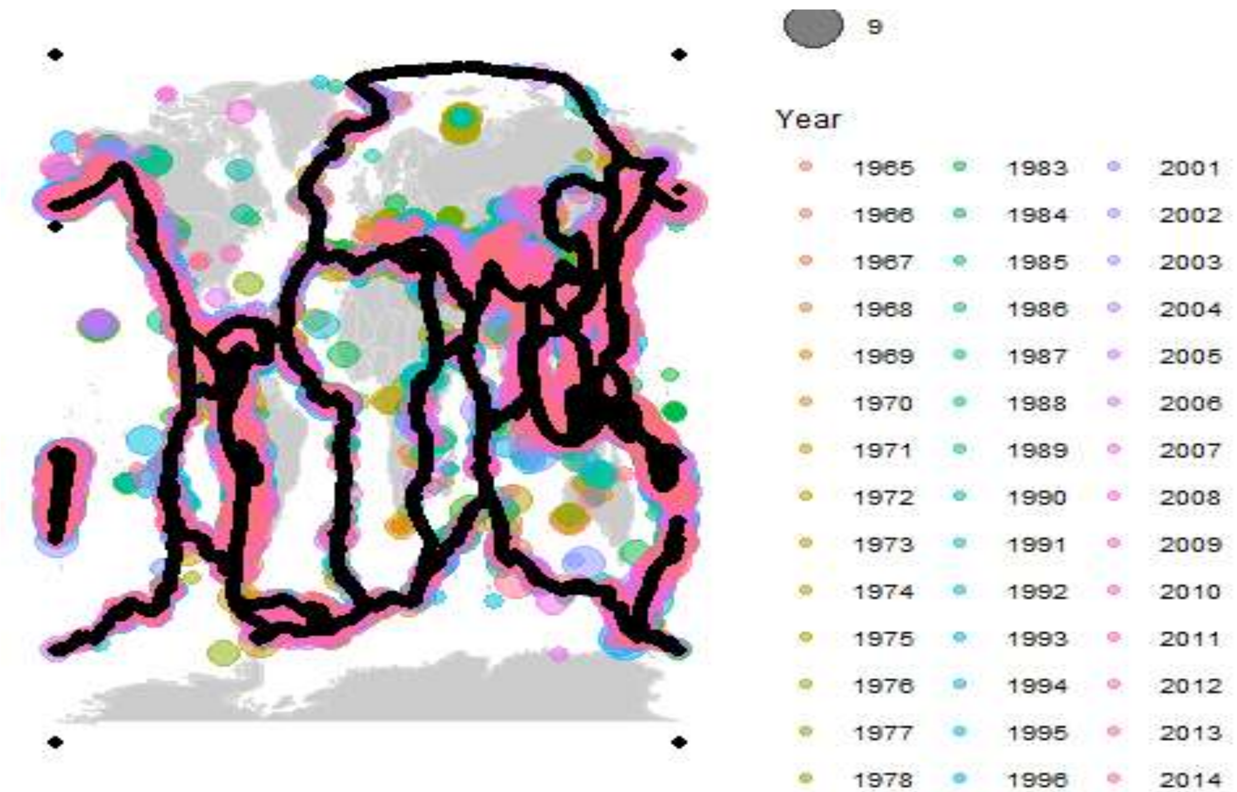
A boxplot of Plate name vs. magnitude was then created for all earthquakes in the data set. From this plot, it is easy to tell that the majority of significant earthquakes occuring on all plates register between a 5.5 and a 6.5 on the Richter scale. This tells us that the outlier events are above a 6.5. The plates that have extreme outliers (India, Burma, North America, Okhotsk) are plates that sit on top of convergent subduction zones, where pressure would built until one plate slips under the other, creating a large seismic event.



Geographic Plots

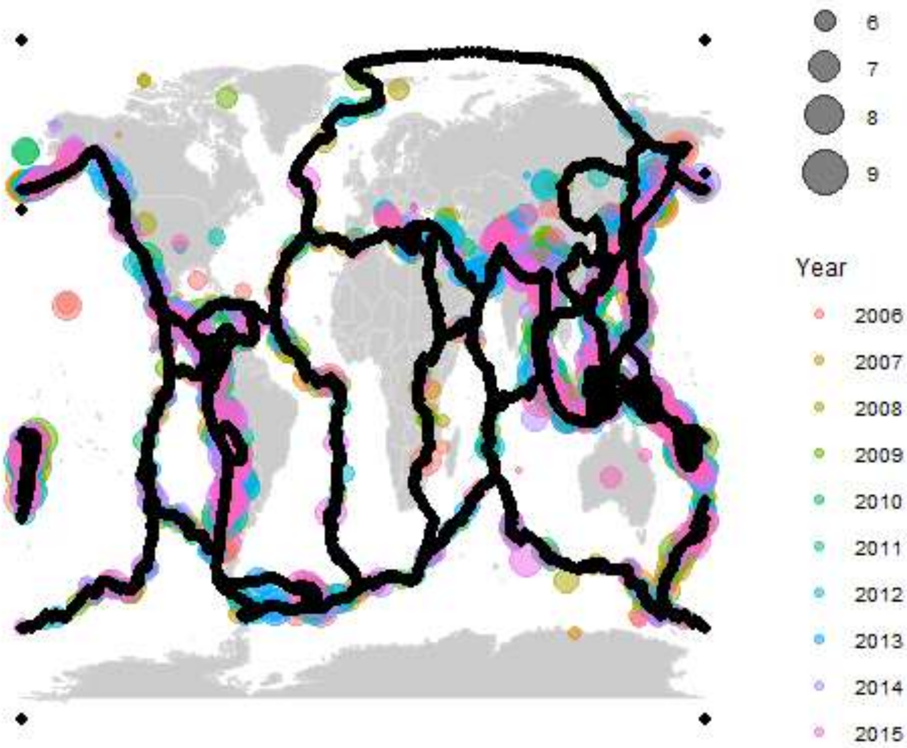
The next step taken in the data exploration was to plot the earthquake occurences on a world map. Once the map of the tectonic plates was established, the earthquakes were then charted by year to see if there was a recognizable pattern.


9
Plate Name
Aegean Sea
Africa
Altiplano
Amur
Anatolia
Antarctica
Arabia
Australia
Balmoral Reef
Banda Sea
Birds Head
Burma
Caribbean
Caroline
Cocos
Conway Reef
Easter
Eurasia
Futuna
Galapagos
India
Juan de Fuca
Juan Fernandez
Kermadec
Manus
Maoke
Mariana
Molucca Sea
Nazca
New Hebrides
Niuafo'ou
North America
North Andes



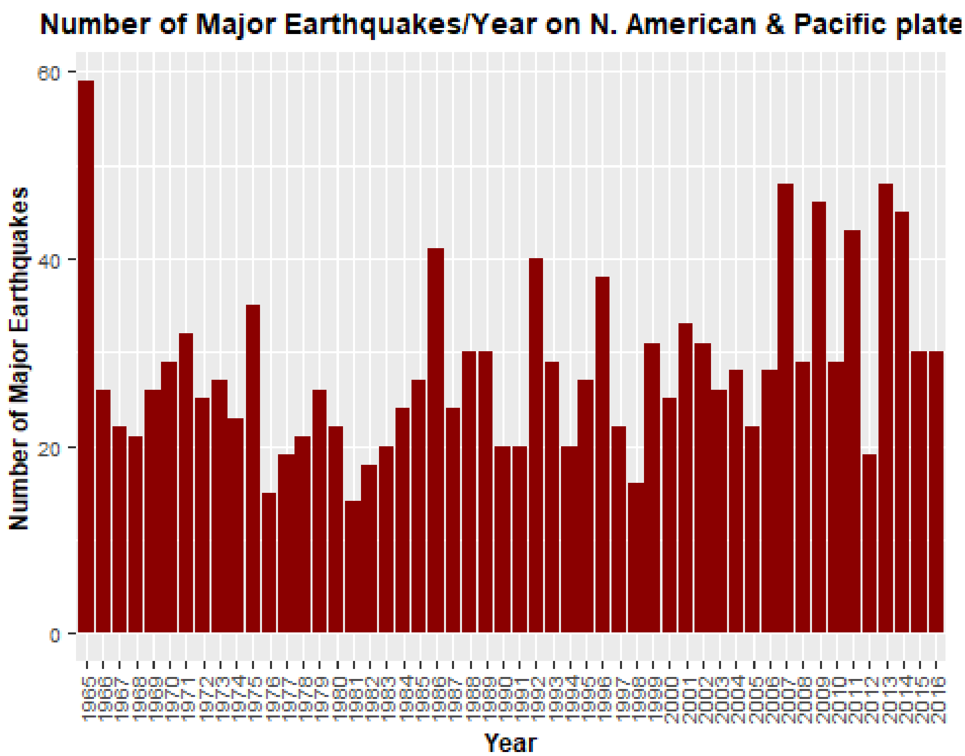
As the year plate map revealed, there were many significant earthquakes within the last ten years of the data set. To get a better look at the data, the map was restricted to just the data from 2006 through 2016. When that data was charted, the results were mixed and not quite clear. It became obvious that the most recent major earthquakes were along subduction zones such as the western edge of the South American continent and along the Aleutian islands of Alaska. Places where multiple plates met along the ring of fire (the western, northern, and eastern edges of the Pacific plate) experienced strong annual seismic events.

Interestingly, the decade map also picked up increased seismic activity that was recorded in the middle of the tectonic plates. Some of this, such as the Hawaiian Islands, can be caused by ‘hot spots’ or thin, weak areas in the Earth’s crust that allow magma to push through, forming a volcano. However, other seismic events, such as the ones recorded in Arkansas, Virginia, and the Gulf of Mexico, may be caused by human activity. As a result, the focus is on major earthquakes occurring at plate boundaries.

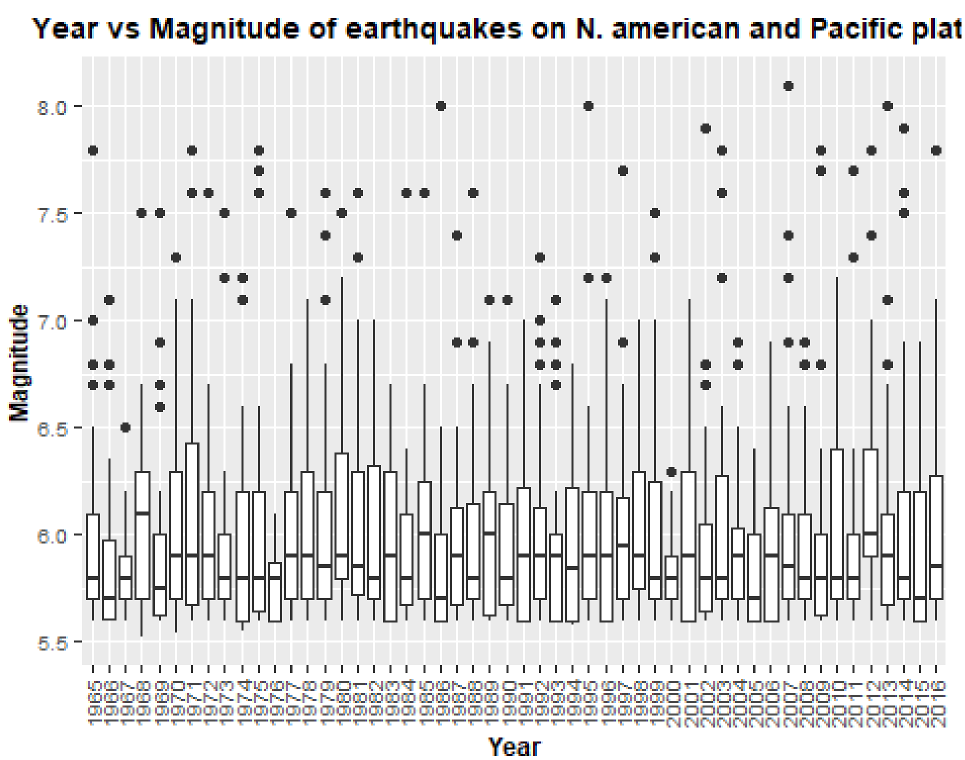


Creation of NAP data set

The next step in the data exploration was to narrow the data. This was done by restricting the data to just one or two plates and repeating the bar graph and the box plots. An animation was also added to better visualize the data. The plates that were tested individually were the Pacific plate, the North American Plate, the South American Plate, the Eurasian Plate, and the North American and Pacific Plates combined. Of these, the only restricted data to show a pattern was the North American and the Pacific plates. These two plates were combined because as the Pacific plate subducts under the North American plate, the seismic events are recorded on the North American plate. As a result of this analysis, the North American/Pacific plate data was isolated and analyzed in addition to the full data set, earthquakes.



Number of Major Earthquakes/Year on N. American & Pacific plate
Number of Major Earthquakes
Year



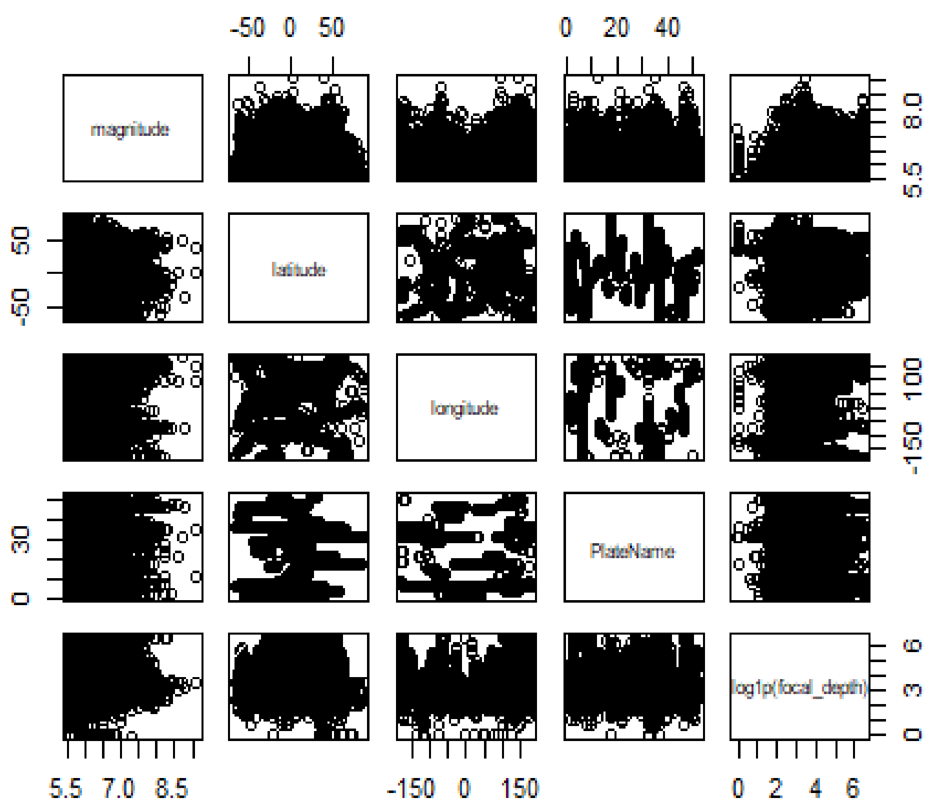
Year vs Magnitude of earthquakes on N. american and Pacific plat
Magnitude
Year

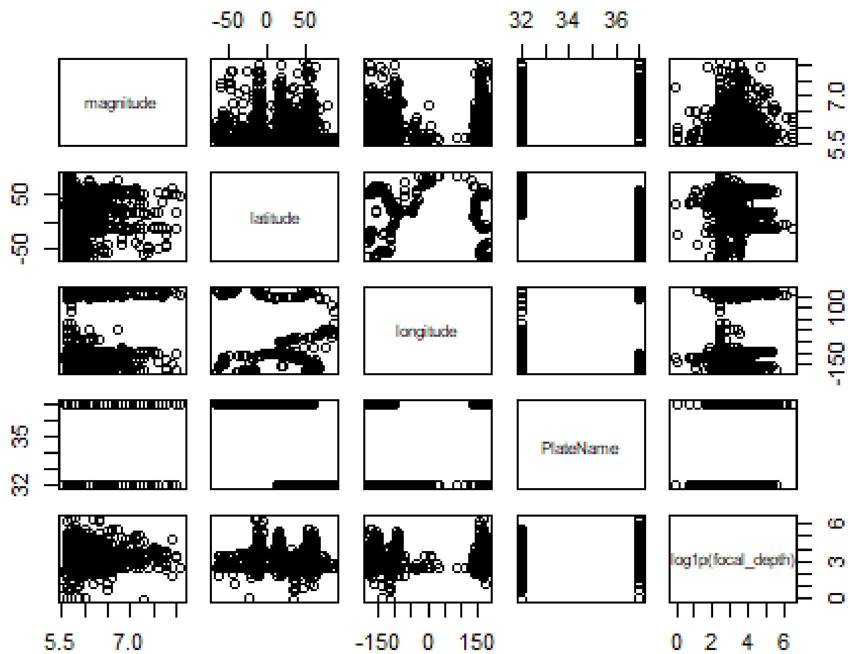
Checking for skew of variables

After looking at the data and narrowing down interactions to the Pacific and North American tectonic plates, the next step was to look for interactions between variables. During this analysis, it was found that the data for focal depth is slightly skewed to the right, so the log1p() function was used to correct this during data analysis.

Interactions between variables

After checking for skew, interactions between variables were analyzed across the entire data set, earthquakes, and across the restricted North American/Pacific Plate earthquakes (nap_earthquakes). From the pairs analysis, there appears to be an interaction between longitude and focal depth in both the full data set (first chart) as well as in the filtered North American/Pacific plate data set (second chart).





Creating predictive models

After analyzing the above interactions, several predictive models were created to attempt to better predict the magnitude of major earthquakes. The models that were used included linear regression, logistic regression, binomial regressions, and the Poisson model.

Linear Regression

As was discovered earlier, there are multiple variables that can impact an earthquake's magnitude. Due to this, multiple linear regression was used to attempt to find if there is a linear relationship between the variables. However, after completing the linear regressions, it was determined this model is not successful in predicting the probability of earthquakes as it does not follow a normal distribution, the coefficients vary substantially in value, and the variables produce a t-value that is close to zero, indicating that there is not a linear relationship between variables.

Logistic Regression

	LR Chisq	Df	Pr(>Chisq)
latitude	0.3878971	1	0.5334068
longitude	2.7782866	1	0.0955503
log1p(focal_depth)	21.7411449	1	0.0000031

PlateName	128.4982456	51	0.0000000

	LR Chisq	Df	Pr(>Chisq)
longitude	10.1898193	1	0.0014122
log1p(focal_depth)	20.8374870	1	0.0000050
longitude:log1p(focal_depth)	0.0001287	1	0.9909489

	LR Chisq	Df	Pr(>Chisq)
log1p(focal_depth)	21.05975	1	4.5e-06

	LR Chisq	Df	Pr(>Chisq)
latitude	2.652064	1	0.1034155
longitude	8.128204	1	0.0043582
latitude:longitude	1.686511	1	0.1940615

Reviewing the models for logistic regression for all earthquakes, it becomes clear that a logistic regression does not accurately predict magnitude for earthquakes. However, from the Wald Type II Chi-sq test, we find longitude to be highly significant in determining whether an earthquakes will have a magnitude of 7.0 or above. The tests also reveal focal depth and plate name to be significant predictors.

This process was then repeated by restricting the plates to the NAP plate data.

	LR Chisq	Df	Pr(>Chisq)
latitude	1.2460819	1	0.2643021
longitude	0.7247357	1	0.3945949
latitude:longitude	2.7282562	1	0.0985869

	LR Chisq	Df	Pr(>Chisq)
latitude	2.5501608	1	0.1102830
longitude	0.0606905	1	0.8054079
log1p(focal_depth)	2.8730611	1	0.0900734
latitude:longitude	0.8196292	1	0.3652886
latitude:log1p(focal_depth)	2.3995346	1	0.1213714
longitude:log1p(focal_depth)	4.5915414	1	0.0321301
latitude:longitude:log1p(focal_depth)	2.0016404	1	0.1571291

	LR Chisq	Df	Pr(>Chisq)

latitude	1.7466164	1	0.1863027
longitude	0.1379187	1	0.7103590
log1p(focal_depth)	3.8169406	1	0.0507368

	LR Chisq	Df	Pr(>Chisq)
longitude	1.017728	1	0.3130586
log1p(focal_depth)	3.316406	1	0.0685916
longitude:log1p(focal_depth)	2.243741	1	0.1341561

	LR Chisq	Df	Pr(>Chisq)
log1p(focal_depth)	4.110623	1	0.0426147

Looking at the results of the logistic regression for the North American and Pacific plates, it is clear that a logistic regression model better fits this restricted data set rather than the entire earthquakes set. The deviances are 1/10 of the values found in the larger set: deviances hover around 600 as do AIC values, indicating that this is a better fit than when the model is applied to all earthquake data.

Interestingly, when the Wald type II test is run across each of the NAP models, focal depth of the earthquake becomes the most significant predictor of when an earthquake's magnitude will exceed 7.0 on the Richter scale.

Binomial distributions

Relative Effects	x
(Intercept)	0.0229885
PlateNameAfrica	1.0609756
PlateNameAltiplano	1.7058824
PlateNameAmur	3.6250000
PlateNameAnatolia	0.7631579
PlateNameAntarctica	0.8169014
PlateNameArabia	1.0357143
PlateNameAustralia	1.6979554
PlateNameBalmoral Reef	0.0000000
PlateNameBanda Sea	2.0196429
PlateNameBirds Head	1.9635417
PlateNameBurma	1.6659574
PlateNameCaribbean	2.1750000
PlateNameCaroline	0.0000000
PlateNameCocos	0.0000000
PlateNameConway Reef	1.1756757
PlateNameEaster	0.0000000
PlateNameEurasia	2.4188696
PlateNameFutuna	0.0000000
PlateNameGalapagos	0.0000000
PlateNameIndia	3.0633803
PlateNameJuan de Fuca	1.6730769
PlateNameJuan Fernandez	3.6250000
PlateNameKermadec	1.6527356
PlateNameManus	0.0000000
PlateNameMaoke	3.1718750
PlateNameMariana	1.5378788
PlateNameMolucca Sea	3.4406780
PlateNameNazca	0.3140794
PlateNameNew Hebrides	2.5688976

PlateNameNiuafo'ou	0.3020833
PlateNameNorth America	2.0512725
PlateNameNorth Andes	3.5655738
PlateNameNorth Bismarck	2.2723881
PlateNameOkhotsk	2.0167418
PlateNameOkinawa	0.7665198
PlateNamePacific	2.4689189
PlateNamePanama	3.1445783
PlateNamePhilippine Sea	1.9097561
PlateNameRivera	0.0000000
PlateNameSandwich	0.3031359
PlateNameScotia	3.3461538
PlateNameShetland	0.0000000
PlateNameSolomon Sea	3.8839286
PlateNameSomalia	0.7190083
PlateNameSouth America	2.8170732
PlateNameSouth Bismarck	1.5688525
PlateNameSunda	1.7149562
PlateNameTimor	1.2920792
PlateNameTonga	0.6503322
PlateNameWoodlark	2.3200000
PlateNameYangtze	0.0000000

Absolute Effects	x
(Intercept)	0.0005281
PlateNameAfrica	0.0238040
PlateNameAltiplano	0.0377272
PlateNameAmur	0.0769061
PlateNameAnatolia	0.0172373
PlateNameAntarctica	0.0184289
PlateNameArabia	0.0232504
PlateNameAustralia	0.0375584
PlateNameBalmoral Reef	0.0000000
PlateNameBanda Sea	0.0443585
PlateNameBirds Head	0.0431795
PlateNameBurma	0.0368768
PlateNameCaribbean	0.0476082
PlateNameCaroline	0.0000000
PlateNameCocos	0.0000000
PlateNameConway Reef	0.0263097
PlateNameEaster	0.0000000
PlateNameEurasia	0.0526651
PlateNameFutuna	0.0000000
PlateNameGalapagos	0.0000000
PlateNameIndia	0.0657748
PlateNameJuan de Fuca	0.0370285
PlateNameJuan Fernandez	0.0769061
PlateNameKermadec	0.0365948
PlateNameManus	0.0000000
PlateNameMaoke	0.0679460
PlateNameMariana	0.0341385
PlateNameMolucca Sea	0.0732822
PlateNameNazca	0.0071668
PlateNameNew Hebrides	0.0557495
PlateNameNiuafo'ou	0.0068949
PlateNameNorth America	0.0450219
PlateNameNorth Andes	0.0757408
PlateNameNorth Bismarck	0.0496341
PlateNameOkhotsk	0.0442976

PlateNameOkinawa	0.0173119	PlateNameSomalia	0.0162563
PlateNamePacific	0.0536963	PlateNameSouth America	0.0608078
PlateNamePanama	0.0674007	PlateNameSouth Bismarck	0.0348021
PlateNamePhilippine Sea	0.0420464	PlateNameSunda	0.0379202
PlateNameRivera	0.0000000	PlateNameTimor	0.0288395
PlateNameSandwich	0.0069188	PlateNameTonga	0.0147265
PlateNameScotia	0.0714127	PlateNameWoodlark	0.0506214
PlateNameShetland	0.0000000	PlateNameYangtze	0.0000000
PlateNameSolomon Sea	0.0819492		

In the binomial regression, both relative and absolute effects were tested to give probabilities for each plate. In the relative effect, values represent how adding plate information increases the odds of a major earthquake.

In the absolute effect, the probabilities of a plate being affected by a major earthquakes were calculated. From this calculation, it is obvious which plates have the highest probability of experiencing a major tremor.

Poisson Models

	x		
(Intercept)	2.0	PlateNameConway Reef	0.5
PlateNameAfrica	3.0	PlateNameEaster	0.0
PlateNameAltiplano	5.0	PlateNameEurasia	30.5
PlateNameAmur	5.5	PlateNameFutuna	0.0
PlateNameAnatolia	0.5	PlateNameGalapagos	0.0
PlateNameAntarctica	6.0	PlateNameIndia	5.0
PlateNameArabia	0.5	PlateNameJuan de Fuca	1.5
PlateNameAustralia	21.0	PlateNameJuan Fernandez	0.5
PlateNameBalmoral Reef	0.0	PlateNameKermadec	12.5
PlateNameBanda Sea	6.5	PlateNameManus	0.0
PlateNameBirds Head	6.5	PlateNameMaoke	3.5
PlateNameBurma	4.5	PlateNameMariana	3.5
PlateNameCaribbean	6.5	PlateNameMolucca Sea	7.0
PlateNameCaroline	0.0	PlateNameNazca	1.0
PlateNameCocos	0.0	PlateNameNew Hebrides	22.5
		PlateNameNiuafo’ou	0.5

PlateNameNorth America	31.5	PlateNameShetland	0.0
PlateNameNorth Andes	5.0	PlateNameSolomon Sea	2.5
PlateNameNorth Bismarck	7.0	PlateNameSomalia	1.0
PlateNameOkhotsk	36.0	PlateNameSouth America	38.5
PlateNameOkinawa	2.0	PlateNameSouth Bismarck	11.0
PlateNamePacific	42.0	PlateNameSunda	31.5
PlateNamePanama	3.0	PlateNameTimor	3.0
PlateNamePhilippine Sea	13.5	PlateNameTonga	4.5
PlateNameRivera	0.0	PlateNameWoodlark	4.0
PlateNameSandwich	1.0	PlateNameYangtze	0.0
PlateNameScotia	1.5		

From the Poisson model, it becomes clear which plates are likely to have the highest number of earthquakes: the Pacific plate is projected to have 42 major earthquakes while the South American plate comes in with a close 38.5 projected major quakes. The Sunda plate, under southeast Asia and western Indonesia is also predicted to experience a large volume of significant tremors, with a predicted 31.5 earthquakes.

Conclusion

After gathering and analyzing two data sets covering global earthquakes with a magnitude of 5.5 or higher (as measured by the Richter scale), a relationship was found to exist between magnitude and longitude as well as magnitude and focal depth. Given what is known about earthquakes, this information is not surprising as subduction zones often occur along the longitudinal lines that create a border along the Pacifc tectonic plate, a part of the Ring of Fire.

Out of the four types of models created, the most successful models were the Binomial disribution and the Poisson model. The binomial distribution was able to predict the probability of a major quake occurring on a given plate while the Poisson model predicted the number of tremors with a magnitude of 7.0 or greater, depending on the tectonic plate location. However, even with this probability, there is significant room to improve the possibility of earthquake prediction, as the Poisson model does not indicate where the quake may occur.

For future analysis, I would propose a few additions. The first modification to this analysis would be to analyze the data using a time series model as this may show a relationship between the time different earthquakes occurred on a given plate, as well as their locations. This type of analysis would likely work best on a plate that has similar to characteristics the Pacific plate: large numbers of significant quakes (magnitude 5.5 or larger), constantly moving, a variety of types of fault zones. Another addition would include adding a data set that classifies the type of fault zone and to run an analysis on how fault zone impacts magnitude. Current geologic knowledge indicates that subduction zones cause the largest quakes; however, the amount of time that earthquakes have been able to be accurately measured is small when compared to the geologic timeline and it is possible that a strike-slip fault may also trigger less-frequent major quakes. The last addition would be a mixed effects model to maximize the data.

References

Ahlenius, Hugo. "Fraxen/Tectonicplates." Tectonicplates/GeoJSON, Github, 2 Oct. 2014, github.com/fraxen/tectonicplates/tree/master/GeoJSON.

National Oceanic and Atmospheric Administration. "NCEI/WDS Global Significant Earthquake Database." 10 June 2017. https://www.ngdc.noaa.gov/nndc/struts/form?t=101650&s=1&d=1

US Geological Survey. "Significant Earthquakes, 1965-2016." Significant Earthquakes, 1965-2016, 26 Jan. 2017, www.kaggle.com/usgs/earthquake-database.