Earthquake_Capstone_Report

H. Ewton

9/28/2018

Table of 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 [map of plates] (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

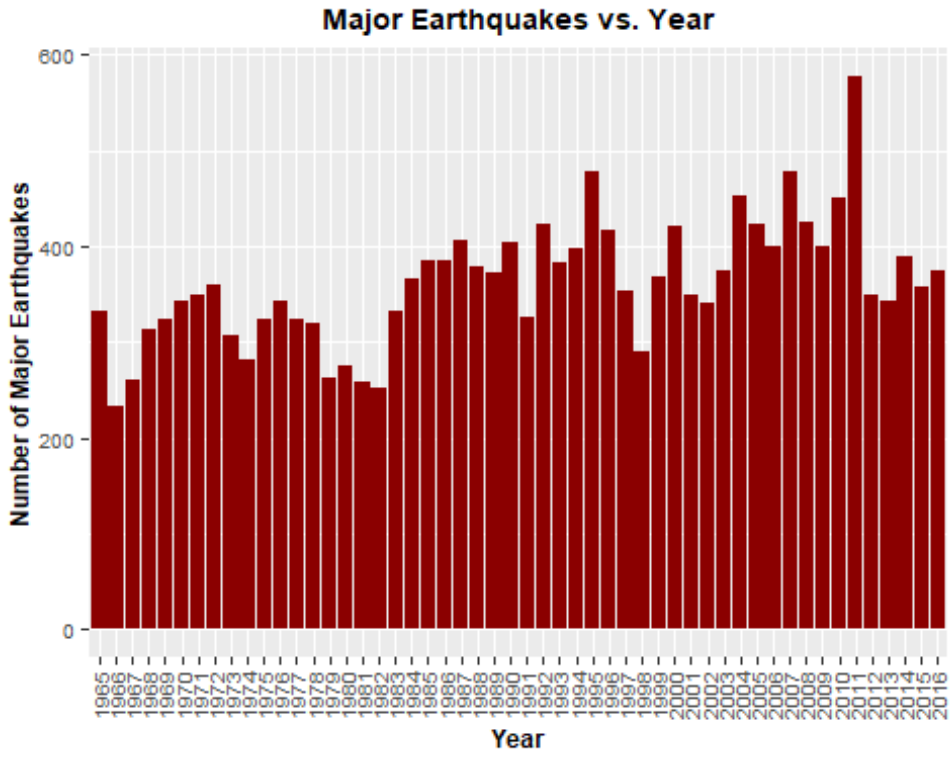
Once the data set was created, one more bit of information was needed for analysis- the tectonic plate data. Stored as a JSON file, this data set contained the plate boundaries of each plate by connecting a series of coordinates. Once loaded, the earthquakes data set was overlaid onto the plate boundaries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

year	month	day	time	focal_depth	magnitude	country	location_name	latitude	longitude	LAYER	Code	PlateName
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

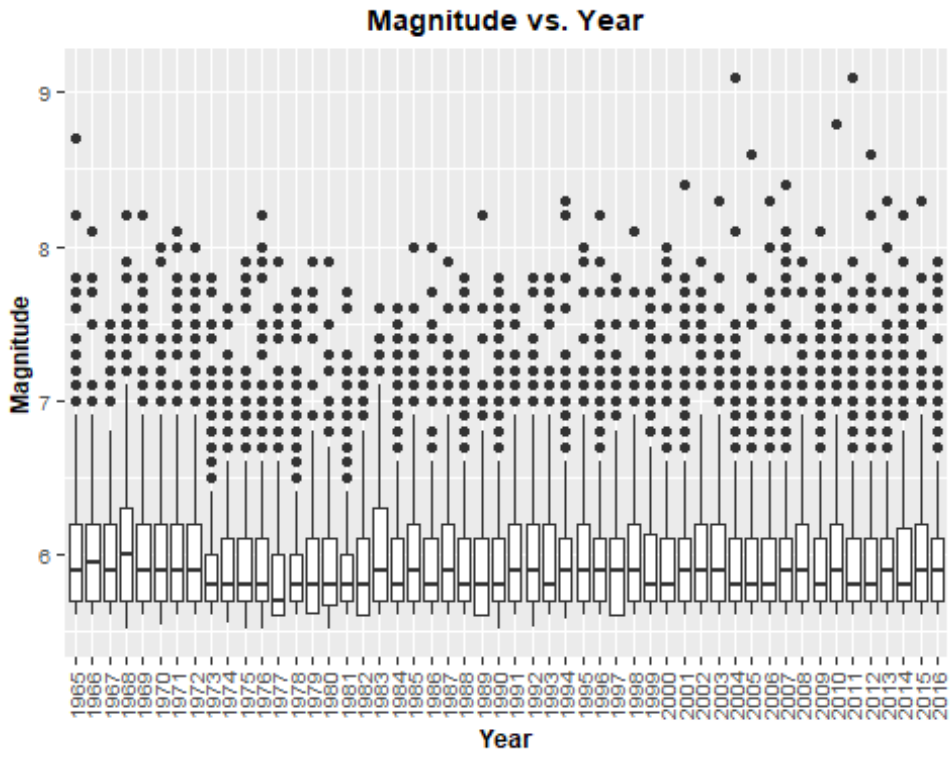
Exploratory Data Analysis

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

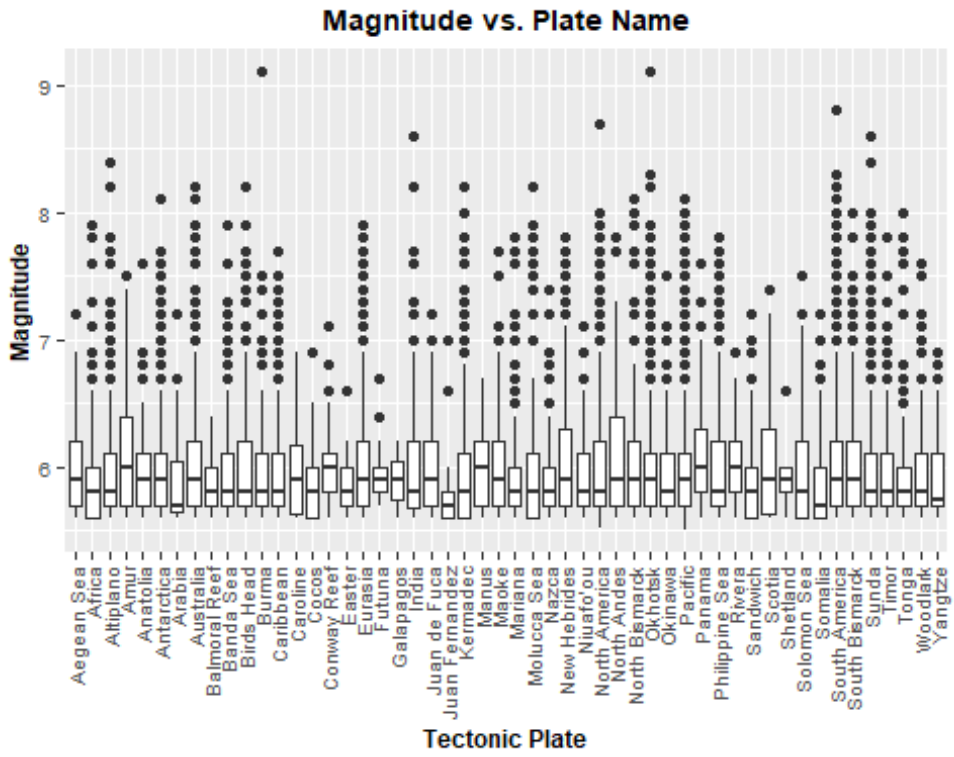
earthquakes from 1965 through 2016. The bar graph yielded the following results, with a trend of an increasing number of earthquakes worldwide. A peak number of earthquakes appears in 2011, which had nearly 100 more significant earthquakes than any other year in recorded seismic history.



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.

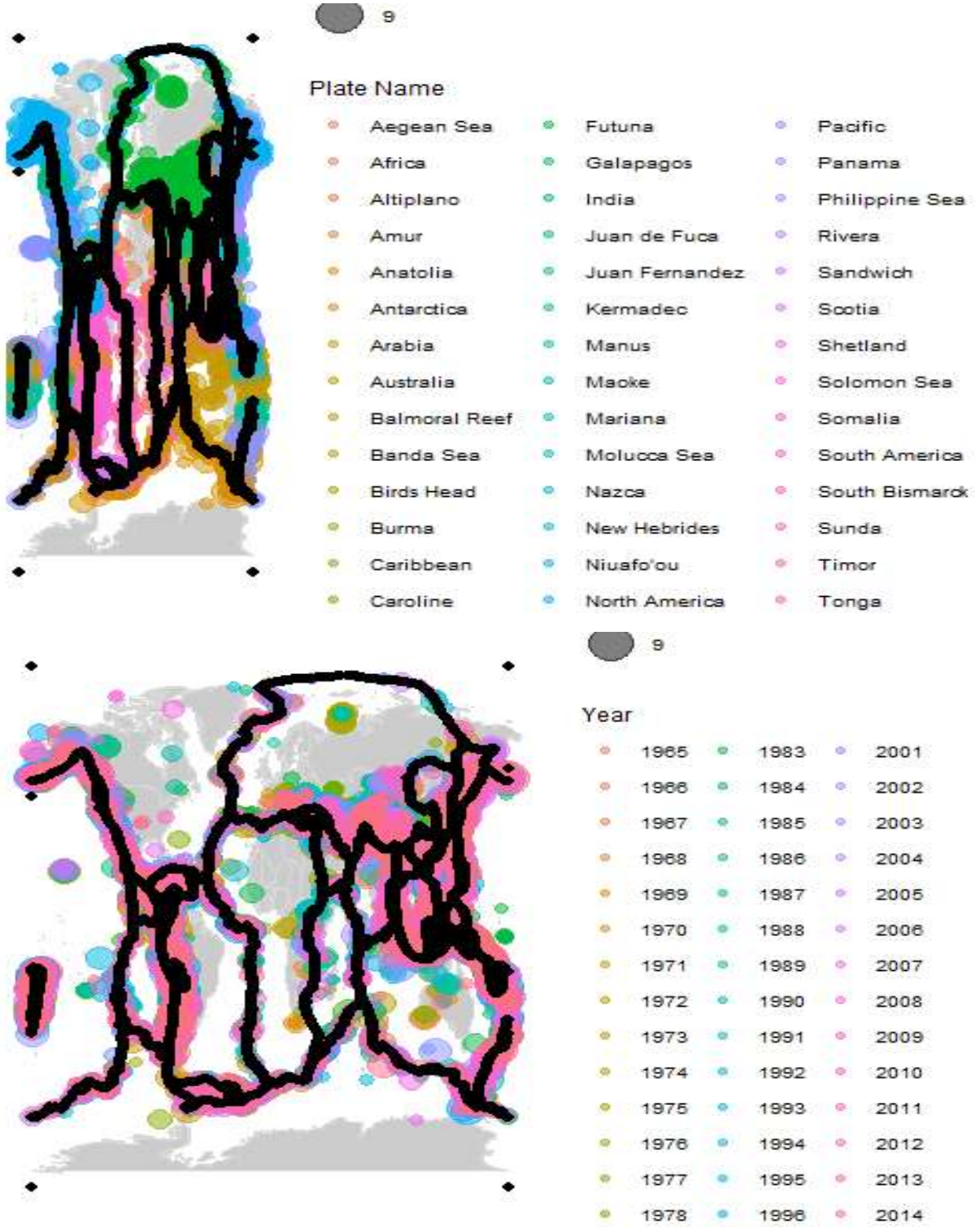


A boxplot of Plate name vs. magnitude was then created for all earthquakes in the data set. From this plot, it is easy to tell that the majority of significant earthquakes occuring on all plates register between a 5.5 and a 6.5 on the Richter scale. This tells us that the outlier events are above a 6.5. 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

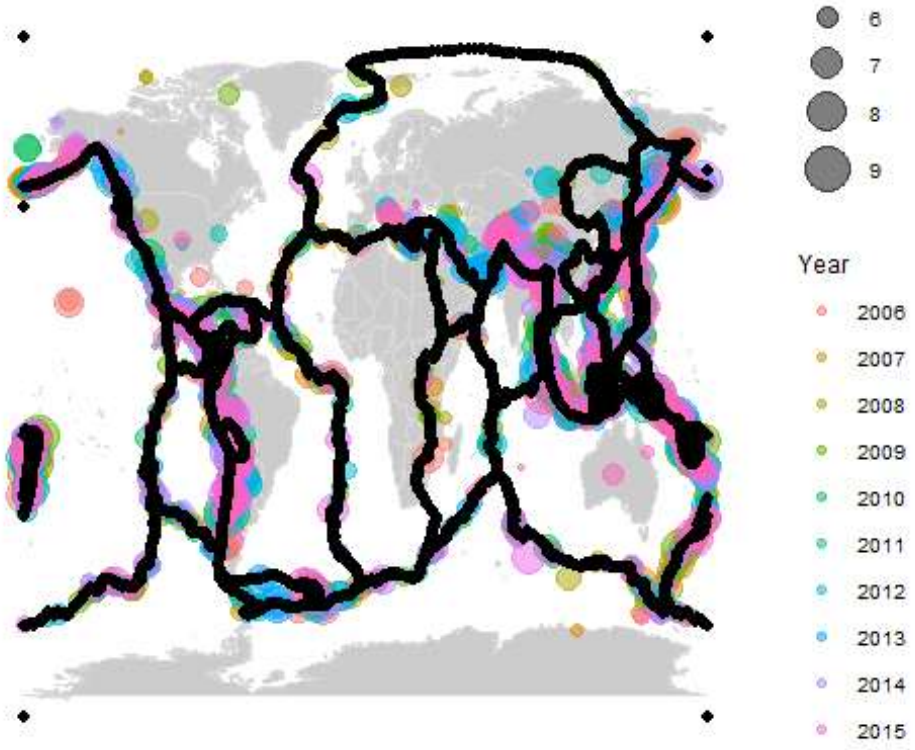
The next step taken in the data exploration was to plot the earthquake occurences on a world map. Once the map of the tectonic plates was established, the earthquakes were then charted by year to see if there was a recognizable pattern.



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

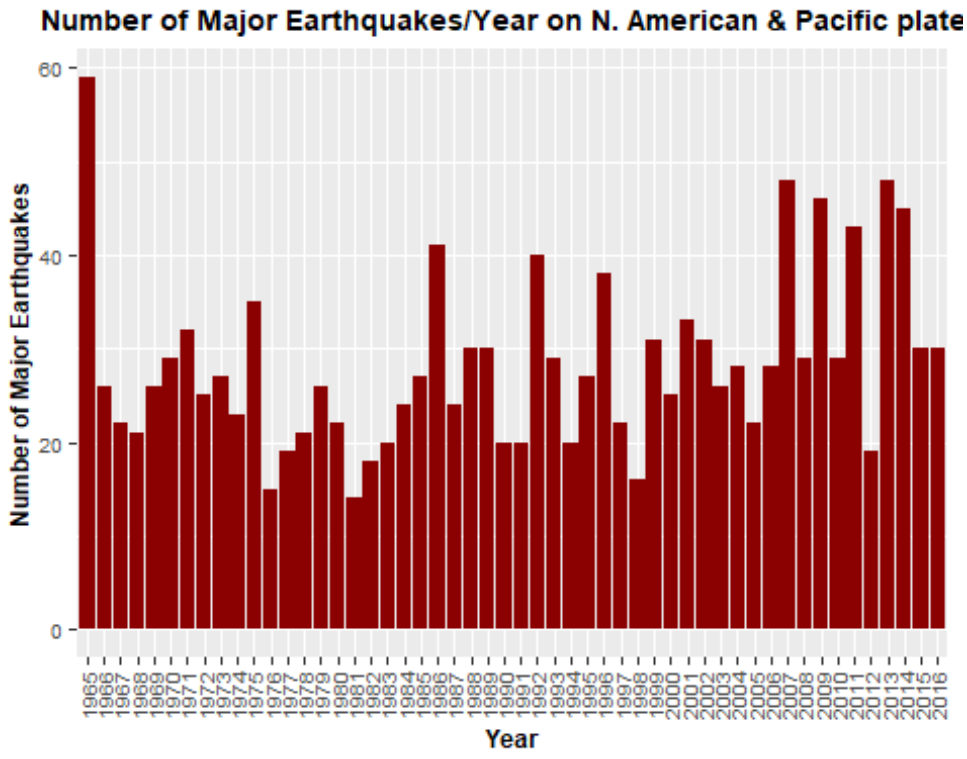
Aleutian islands of Alaska. Places where multiple plates met along the ring of fire (the western, northern, and eastern edges of the Pacific plate) experienced strong annual seismic events.

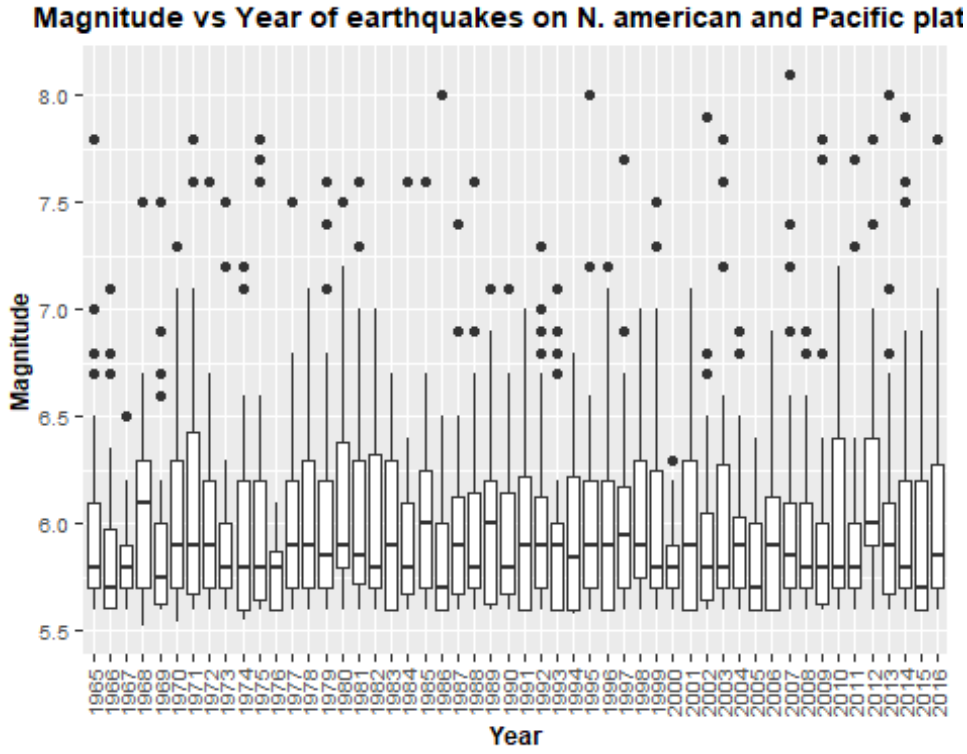
Interestingly, the decade map also picked up increased seismic activity that was recorded in the middle of the tectonic plates. Some of this, such as the Hawaiian Islands, can be caused by 'hot spots' or thin, weak areas in the Earth's crust that allow magma to push through, forming a volcano. However, 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
0
20
40
60
Year


Magnitude vs Year of earthquakes on N. american and Pacific plat
Magnitude
5.5
6.0
6.5
7.0
7.5
8.0
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.

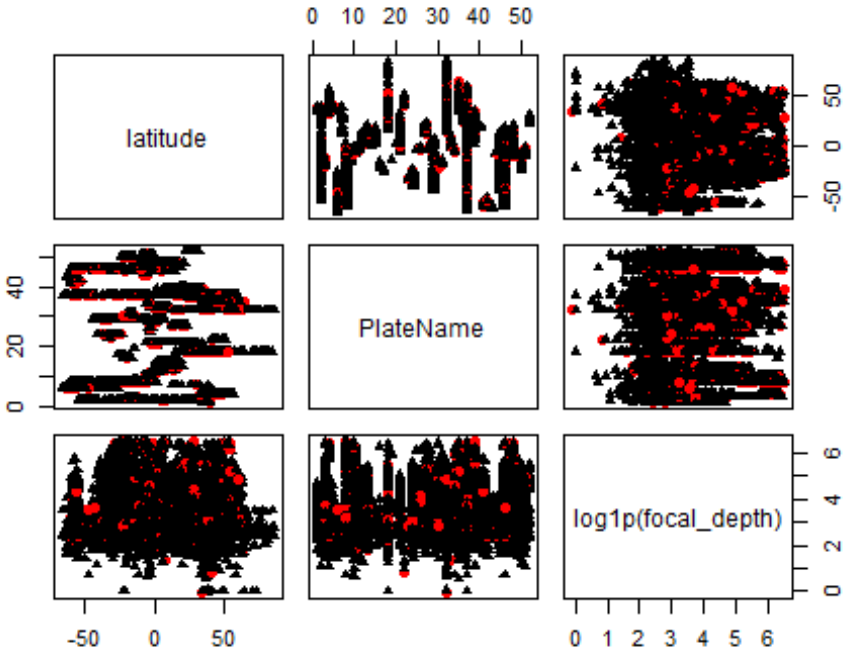
Interactions between variables

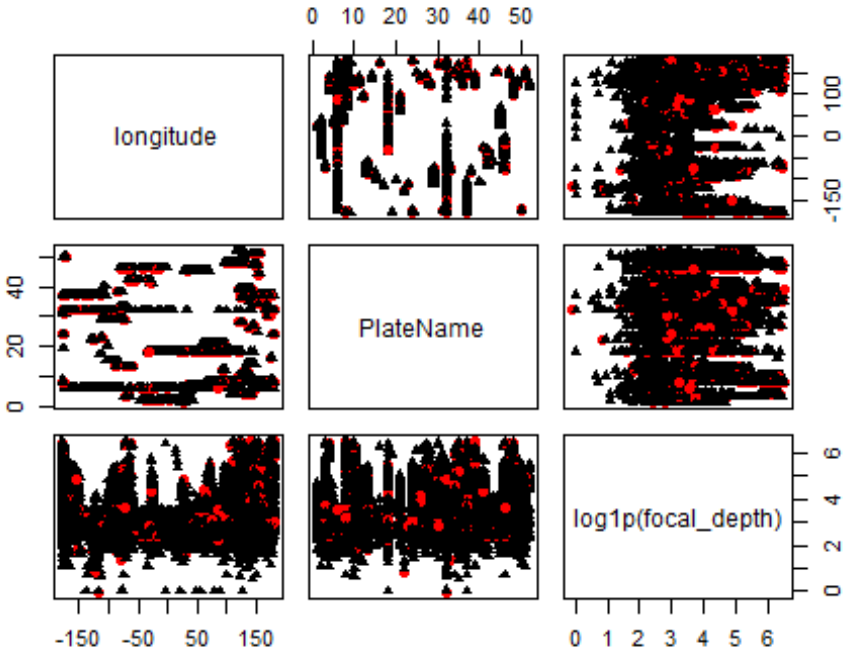
After checking for skew, interactions between variables were analyzed across the entire data set, earthquakes, and across the restricted North American/Pacific Plate earthquakes (nap_earthquakes). In the pairs analysis, the red circles represent any earthquake with a magnitude of 7.0 or higher.

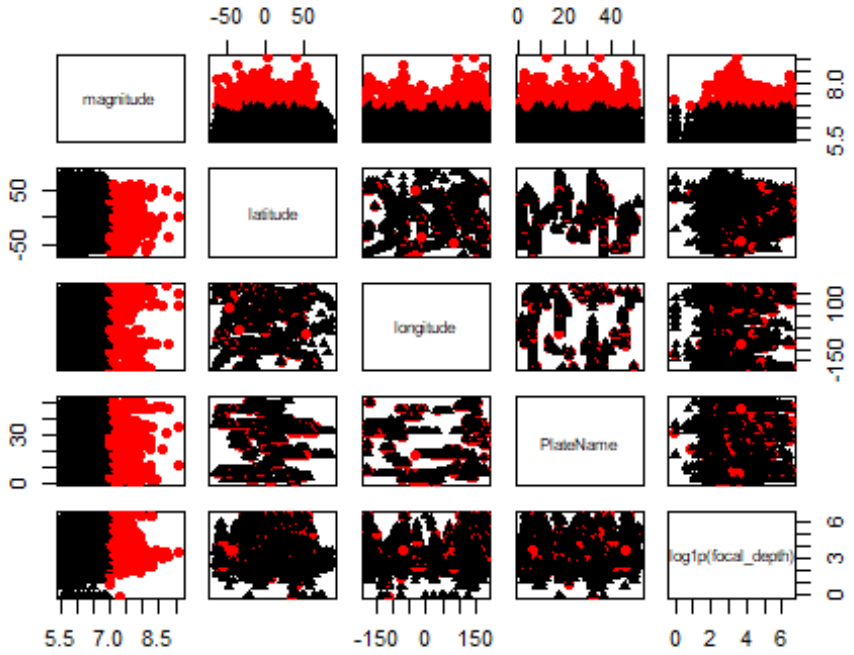
For the full earthquakes data set, the first pairs chart did not reveal a discernible relationship between latitude, plate name, and focal depth. However, from the second and third pairs plots, there does appear to be a relationship between plate name and longitude, as well as plate name and focal depth. Additionally, there appears to be a relationship between magnitude and focal_depth, with the strongest quakes occurring within a narrow depth range.

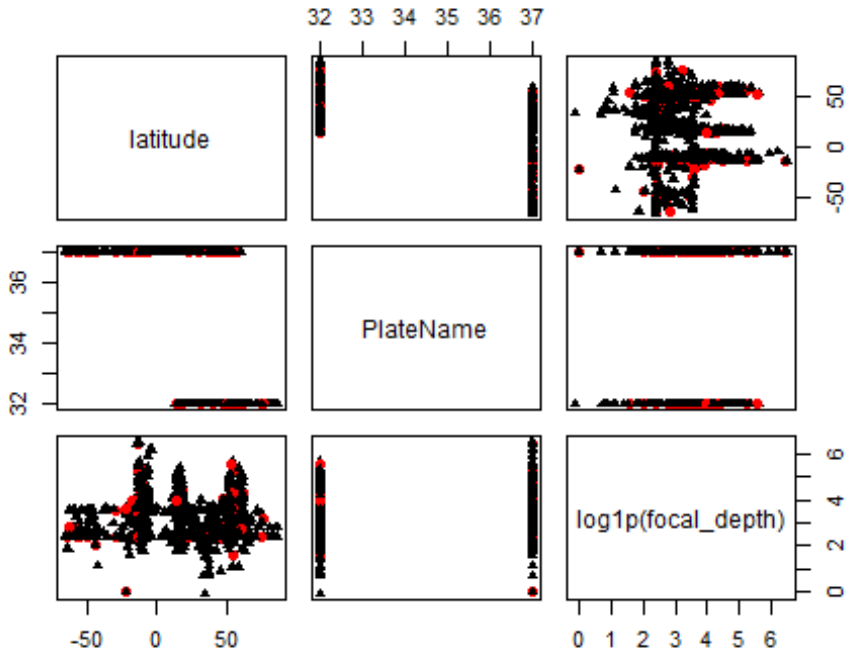
In the pairs plot for the North America/Paccific data set, the majority of quakes are clustered around specific latitudes and longitudes. These plots reveal that the majority of quakes occur along the edge of the plate, with few significant quakes towards the center of the plate. This plot also shows that the majority of earthquakes with a magnitude of 7.0 or above occur within a fairly narrow focal depth range.

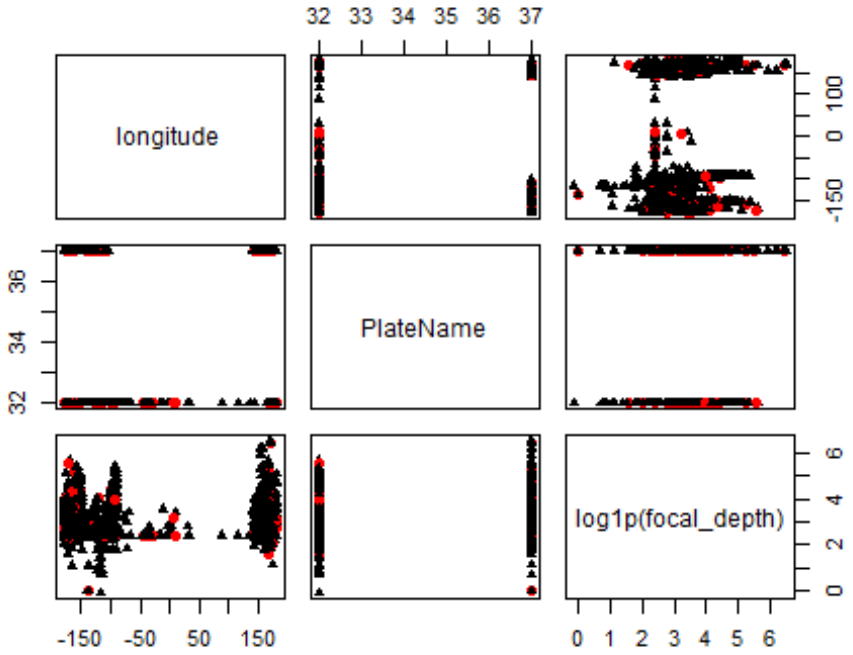
```
## Warning in log1p(focal_depth): NaNs produced
```

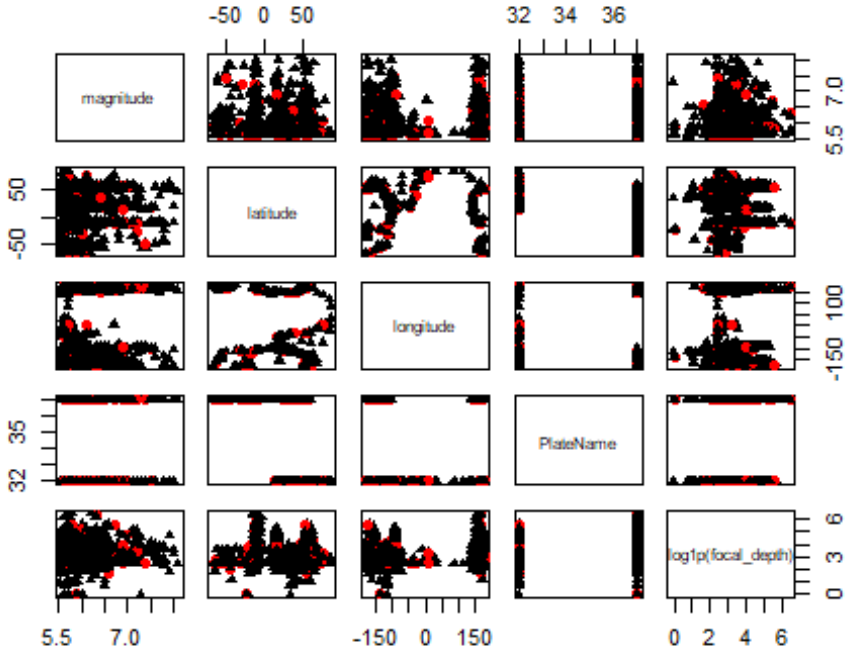

latitude
PlateName
log1p(focal_depth)


longitude
PlateName
log1p(focal_depth)


magnitude
latitude
longitude
PlateName
log1p(focal_depth)


latitude
PlateName
log1p(focal_depth)


longitude
PlateName
log1p(focal_depth)


magnitude
latitude
longitude
PlateName
log1p(focal_depth)

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 and if that relationship could be used to predict the magnitude of a quake. However, after completing the linear regressions,it was determined that this model was not successful in predicting magnitude of a quake as the likelihood and errors given by the models are not normally distributed.

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 from the Wald Type II Chi-sq test that longitude is 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 fairly 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, the results are similar. The deviances are approximately 1/10 of the values found in the larger set: deviances hover around 600 as do AIC values. This makes sense as the data for the Pacific plate is roughly 1/10 of the total data from the earthquakes data set.

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, as compared to longitude in the full data set.

Binomial distributions

Relative Effects	x		
(Intercept)	0.0229885	PlateNameBalmoral Reef	0.0000000
PlateNameAfrica	1.0609756	PlateNameBanda Sea	2.0196429
PlateNameAltiplano	1.7058824	PlateNameBirds Head	1.9635417
PlateNameAmur	3.6250000	PlateNameBurma	1.6659574
PlateNameAnatolia	0.7631579	PlateNameCaribbean	2.1750000
PlateNameAntarctica	0.8169014	PlateNameCaroline	0.0000000
PlateNameArabia	1.0357143	PlateNameCocos	0.0000000
PlateNameAustralia	1.6979554	PlateNameConway Reef	1.1756757
		PlateNameEaster	0.0000000

PlateNameEurasia	2.4188696	PlateNameOkinawa	0.7665198
PlateNameFutuna	0.0000000	PlateNamePacific	2.4689189
PlateNameGalapagos	0.0000000	PlateNamePanama	3.1445783
PlateNameIndia	3.0633803	PlateNamePhilippine Sea	1.9097561
PlateNameJuan de Fuca	1.6730769	PlateNameRivera	0.0000000
PlateNameJuan Fernandez	3.6250000	PlateNameSandwich	0.3031359
PlateNameKermadec	1.6527356	PlateNameScotia	3.3461538
PlateNameManus	0.0000000	PlateNameShetland	0.0000000
PlateNameMaoke	3.1718750	PlateNameSolomon Sea	3.8839286
PlateNameMariana	1.5378788	PlateNameSomalia	0.7190083
PlateNameMolucca Sea	3.4406780	PlateNameSouth America	2.8170732
PlateNameNazca	0.3140794	PlateNameSouth Bismarck	1.5688525
PlateNameNew Hebrides	2.5688976	PlateNameSunda	1.7149562
PlateNameNiuafo'ou	0.3020833	PlateNameTimor	1.2920792
PlateNameNorth America	2.0512725	PlateNameTonga	0.6503322
PlateNameNorth Andes	3.5655738	PlateNameWoodlark	2.3200000
PlateNameNorth Bismarck	2.2723881	PlateNameYangtze	0.0000000
PlateNameOkhotsk	2.0167418		

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

PlateNameNorth Bismarck	0.0496341	PlateNameSolomon Sea	0.0819492
PlateNameOkhotsk	0.0442976	PlateNameSomalia	0.0162563
PlateNameOkinawa	0.0173119	PlateNameSouth America	0.0608078
PlateNamePacific	0.0536963	PlateNameSouth Bismarck	0.0348021
PlateNamePanama	0.0674007	PlateNameSunda	0.0379202
PlateNamePhilippine Sea	0.0420464	PlateNameTimor	0.0288395
PlateNameRivera	0.0000000	PlateNameTonga	0.0147265
PlateNameSandwich	0.0069188	PlateNameWoodlark	0.0506214
PlateNameScotia	0.0714127	PlateNameYangtze	0.0000000
PlateNameShetland	0.0000000		

Given the sample size, binomial regression was also used to predict the probability of a significant earthquake. This type of model was selected because each plate has a different number of total earthquakes and major earthquakes; a binomial regression preserves thesesample sizes and the variability. 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. From the relative effect model, we can see the Scotia plate, the South American plate, the North Bismark plate, the Pacific plate, the North Andes plate, and the New Hebrides plate all significantly increase 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: the Amur plate, the Molucca Sea plate, the Maoke plate, the India plate, the North Andes plate, the Juan Fernandez plate, the Scotia plate, the Panama Plate, and the Solomon Sea plate all have probabilities higher than 6.5%. Of these, the Solomon Sea plate has the highest probability of experiencing a major earthquake at 8.19%.

Poisson Models

	x		
(Intercept)	2.0	PlateNameBanda Sea	6.5
PlateNameAfrica	3.0	PlateNameBirds Head	6.5
PlateNameAltiplano	5.0	PlateNameBurma	4.5
PlateNameAmur	5.5	PlateNameCaribbean	6.5
PlateNameAnatolia	0.5	PlateNameCaroline	0.0
PlateNameAntarctica	6.0	PlateNameCocos	0.0
PlateNameArabia	0.5	PlateNameConway Reef	0.5
PlateNameAustralia	21.0	PlateNameEaster	0.0
PlateNameBalmoral Reef	0.0	PlateNameEurasia	30.5
		PlateNameFutuna	0.0

PlateNameGalapagos	0.0	PlateNamePacific	42.0
PlateNameIndia	5.0	PlateNamePanama	3.0
PlateNameJuan de Fuca	1.5	PlateNamePhilippine Sea	13.5
PlateNameJuan Fernandez	0.5	PlateNameRivera	0.0
PlateNameKermadec	12.5	PlateNameSandwich	1.0
PlateNameManus	0.0	PlateNameScotia	1.5
PlateNameMaoke	3.5	PlateNameShetland	0.0
PlateNameMariana	3.5	PlateNameSolomon Sea	2.5
PlateNameMolucca Sea	7.0	PlateNameSomalia	1.0
PlateNameNazca	1.0	PlateNameSouth America	38.5
PlateNameNew Hebrides	22.5	PlateNameSouth Bismarck	11.0
PlateNameNiuafo'ou	0.5	PlateNameSunda	31.5
PlateNameNorth America	31.5	PlateNameTimor	3.0
PlateNameNorth Andes	5.0	PlateNameTonga	4.5
PlateNameNorth Bismarck	7.0	PlateNameWoodlark	4.0
PlateNameOkhotsk	36.0	PlateNameYangtze	0.0
PlateNameOkinawa	2.0		

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. From the binomial regression, the Solomon Sea plate was found to have the highest probability of experiencing a major earthquake. The plates around the Solomon sea plate, such as the New Hebrides plate and the Molucca Sea plate are also found to have high

probabilities of a major quake. From the Poisson model, it was predecited that the Pacific plate would have the highest number of major earthquakes (42). However, even with this probability, there is significant room to improve the possibility of earthquake prediction, as the Binomial regression and Poisson models do 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. This model would be used to complete a pairwise comparison for the plates and would be used to maximize the data by better accounting for variances in plate measurements, geological activy, and geological formations not captured directly in the measurement values.

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.