Predicting Earthquakes

Using data science to calculate probabilities

Heather Ewton

Foundations of Data Science

Overview

- The Problem
- Method
  - Data
  - Exploratory Analysis
- Predictive Models
- Results & Discussion

The Problem

- There is no reliable method that exists to predict when and where a significant earthquake will occur OR how strong that earthquake will be

Background

- How are earthquakes measured?
  - Richter Scale
    - Created in 1935
    - Measures energy released in a quake on a log-based scale
      - A magnitude of 2 is 10 times stronger than a magnitude of 1
      - A magnitude of 3 is 100 times stronger than a magnitude of 1
    - Measured by seismographs around the world

The Data

- 3 data sets were used:
  - NOAA data set
  - USGS data set
  - GeoJSON file

NOAA data set

- Summary:
  - 23,400 observations
  - 21 columns
  - Includes earthquakes from 1965-2016 with a magnitude of 5.5 or higher

- Clean and tidy

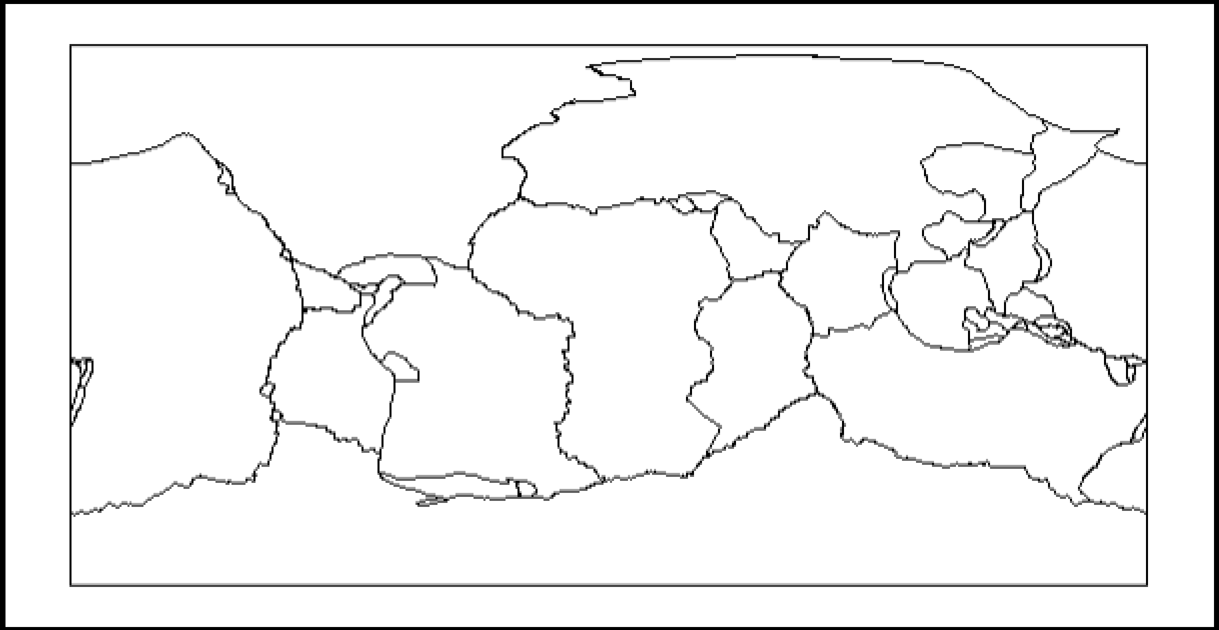
USGS data set

- Summary:
  - 6,047 observations
  - 23 columns
  - Sorted by region but does not include plate names

- Clean and tidy

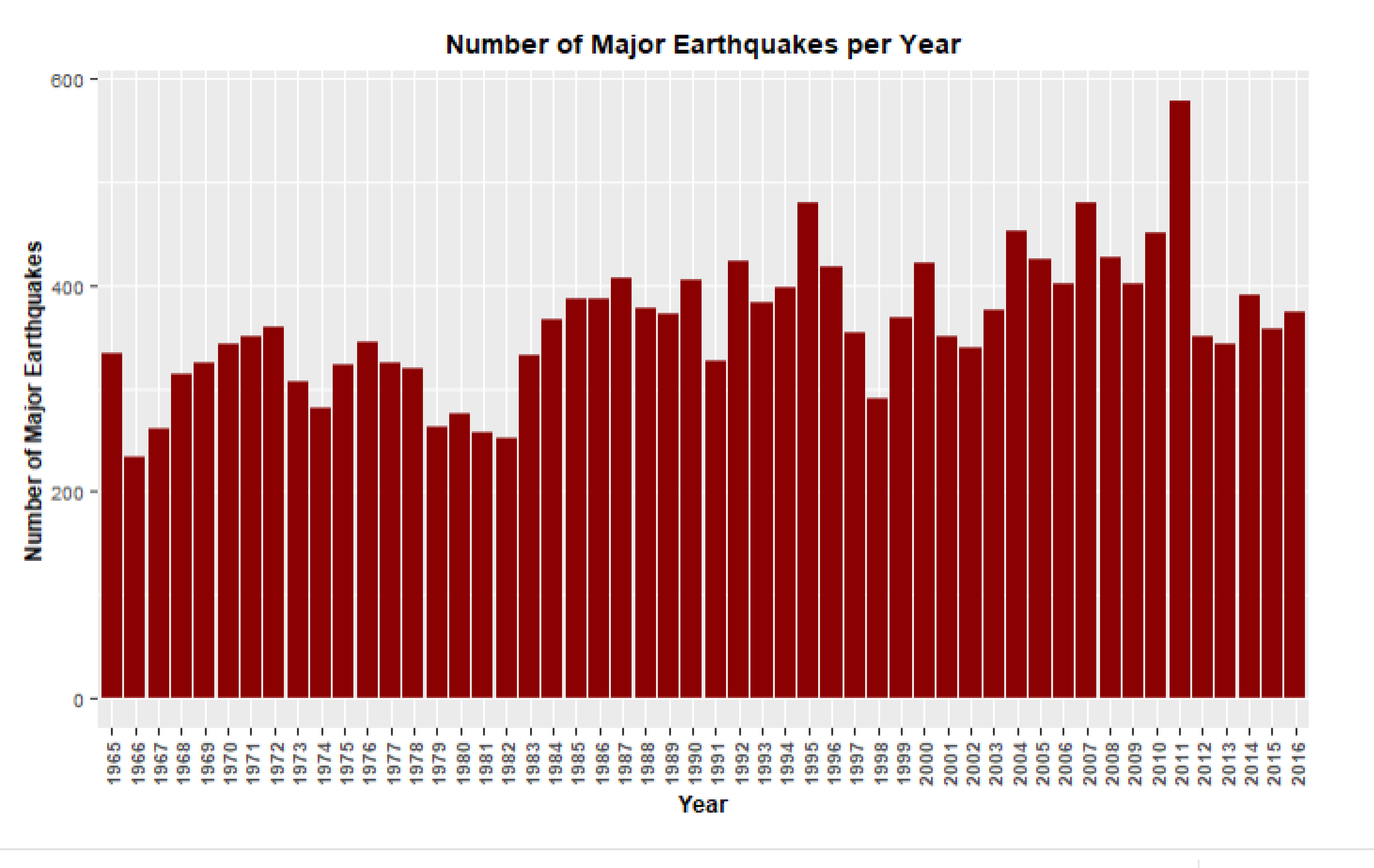
GeoJSON data set

- Contains geographic points on plate boundaries, connects each point with a line



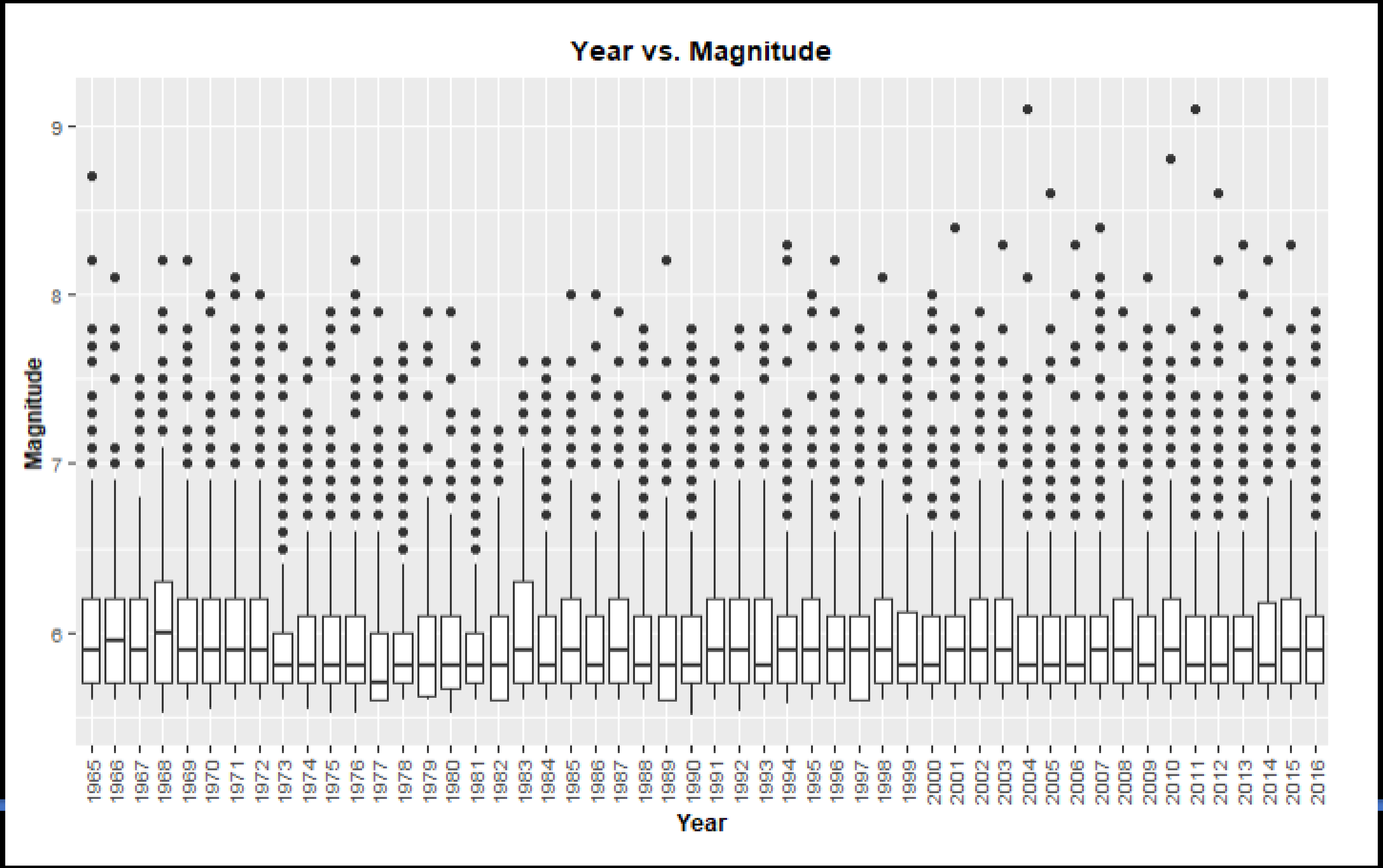
Exploratory Analysis

- Bar graph of significant earthquakes per year



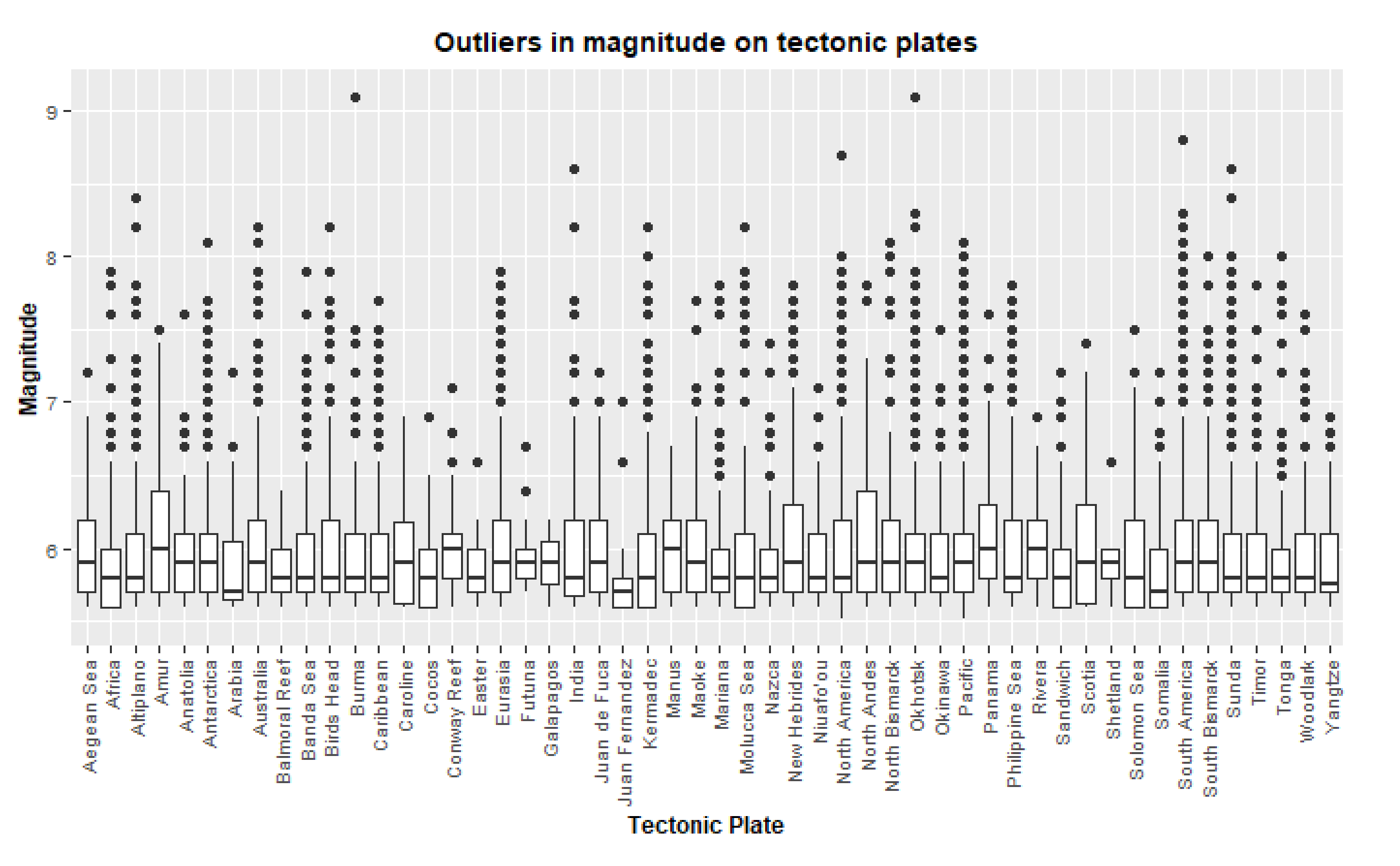
Exploratory Analysis

- Year vs. Magnitude box plot



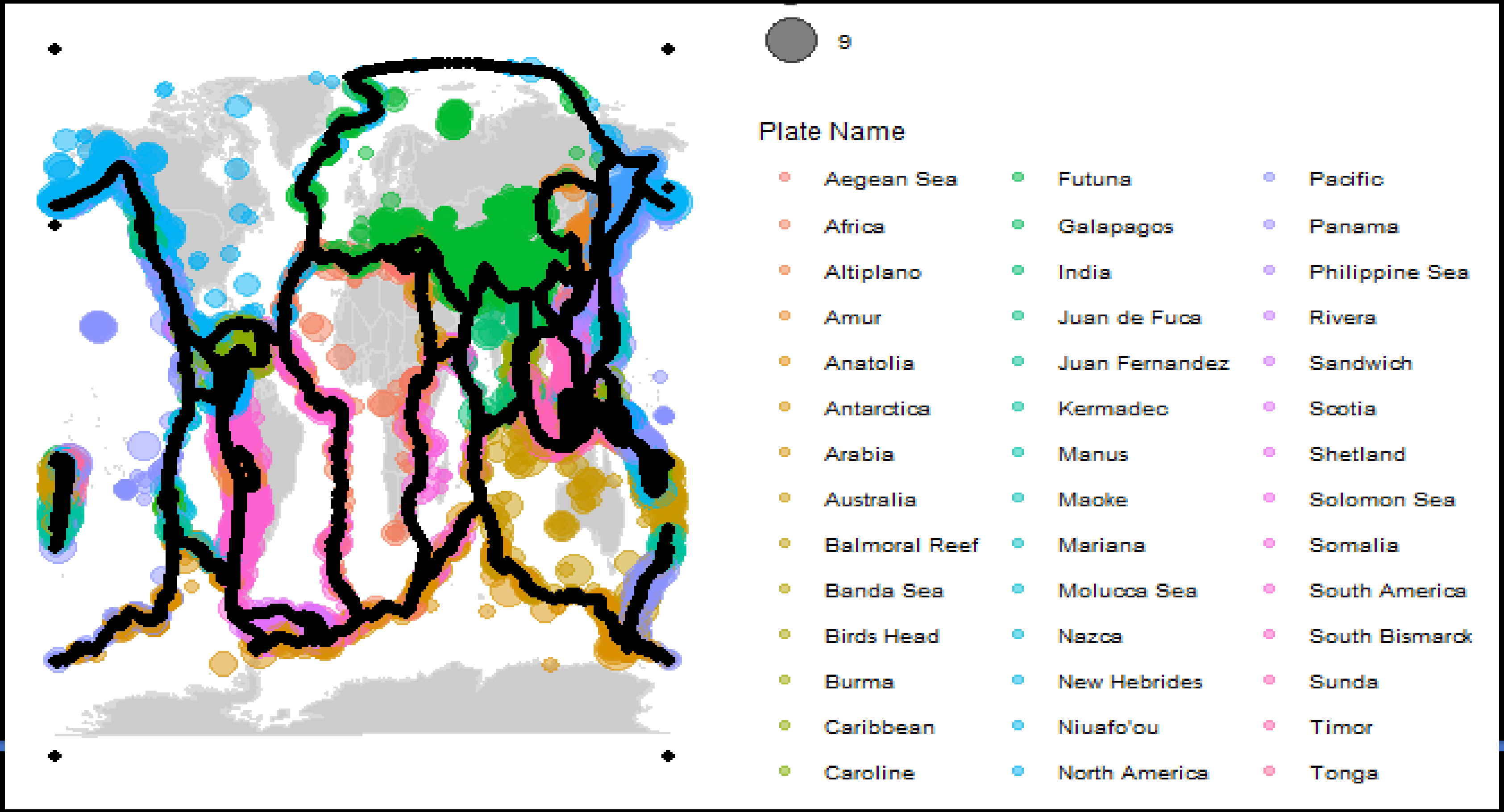
Exploratory Analysis

- Plate Name v. Magnitude



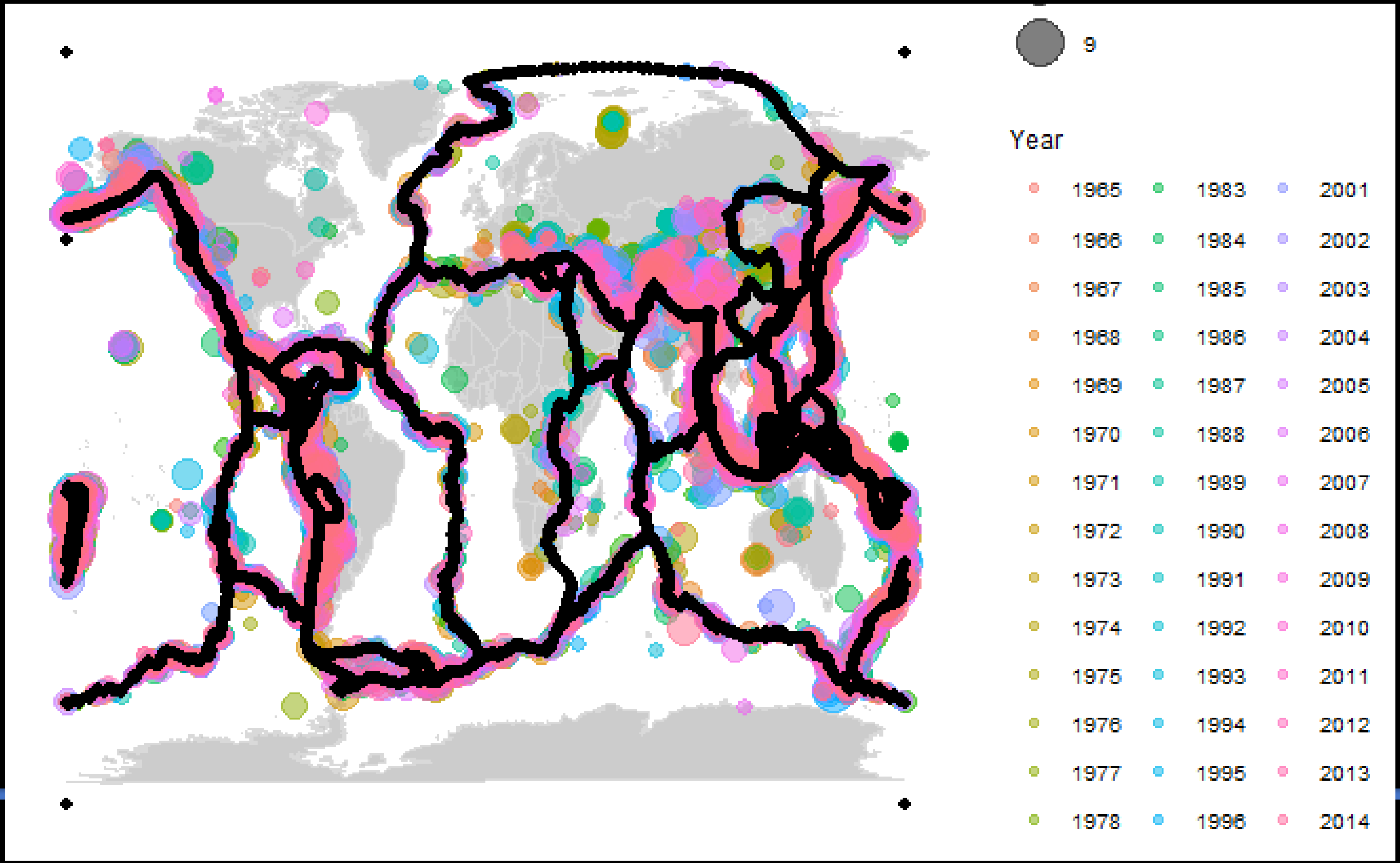
Exploratory Analysis

- Quake plot by plate name, magnitude



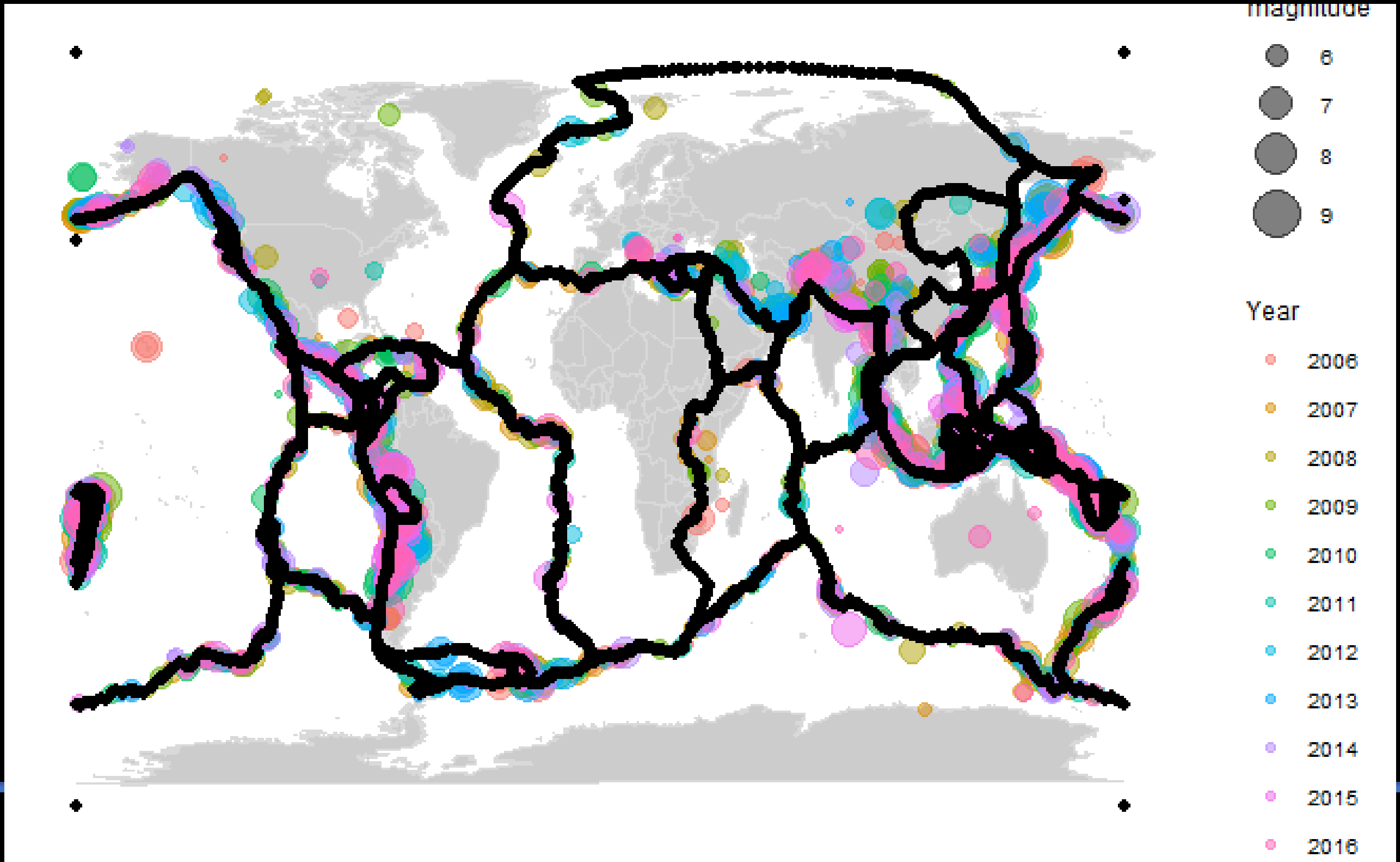
Exploratory Analysis

- Quake plot by year, size = magnitude



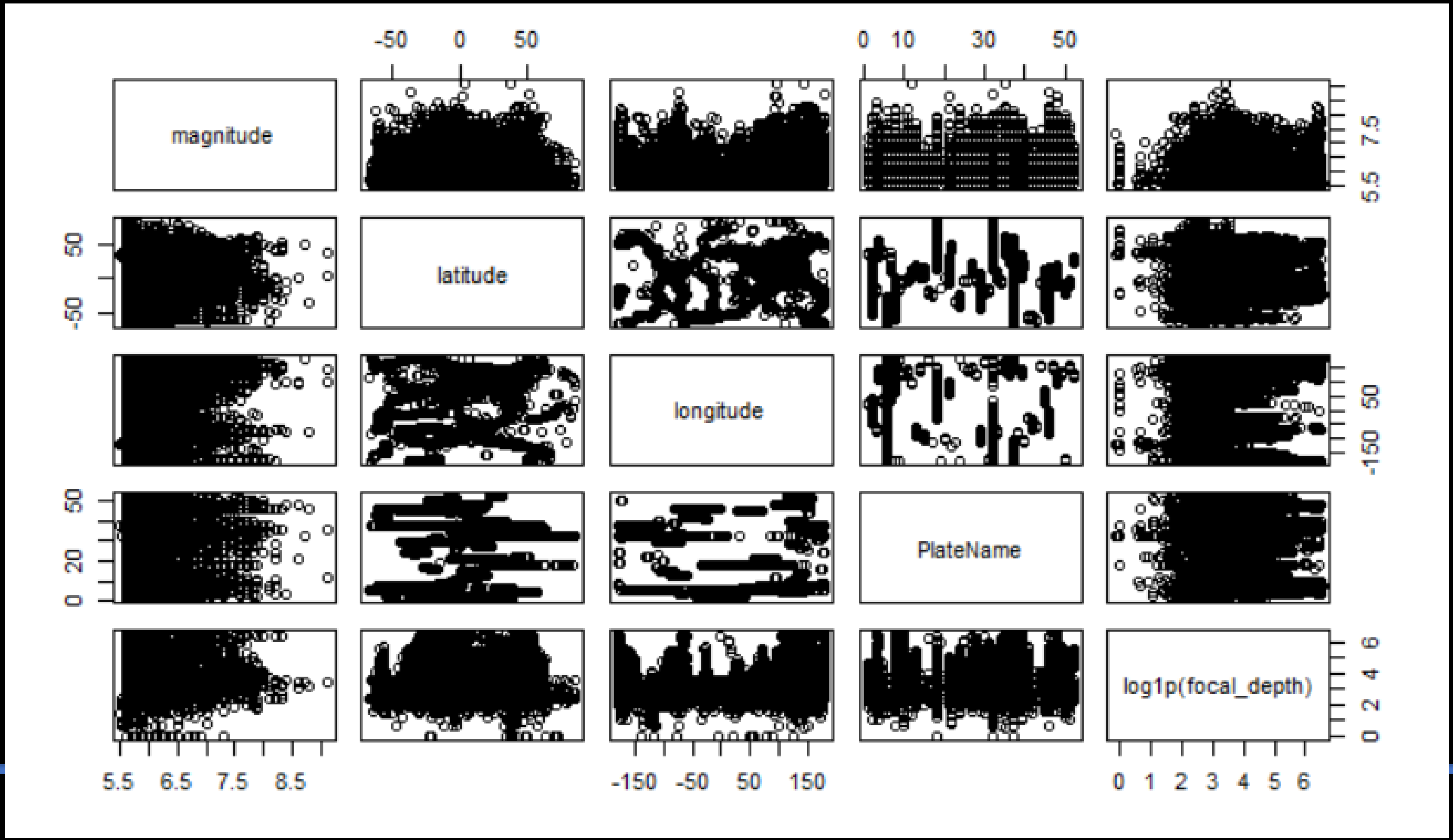
Exploratory Analysis

- Exploring the last ten years of data in detail



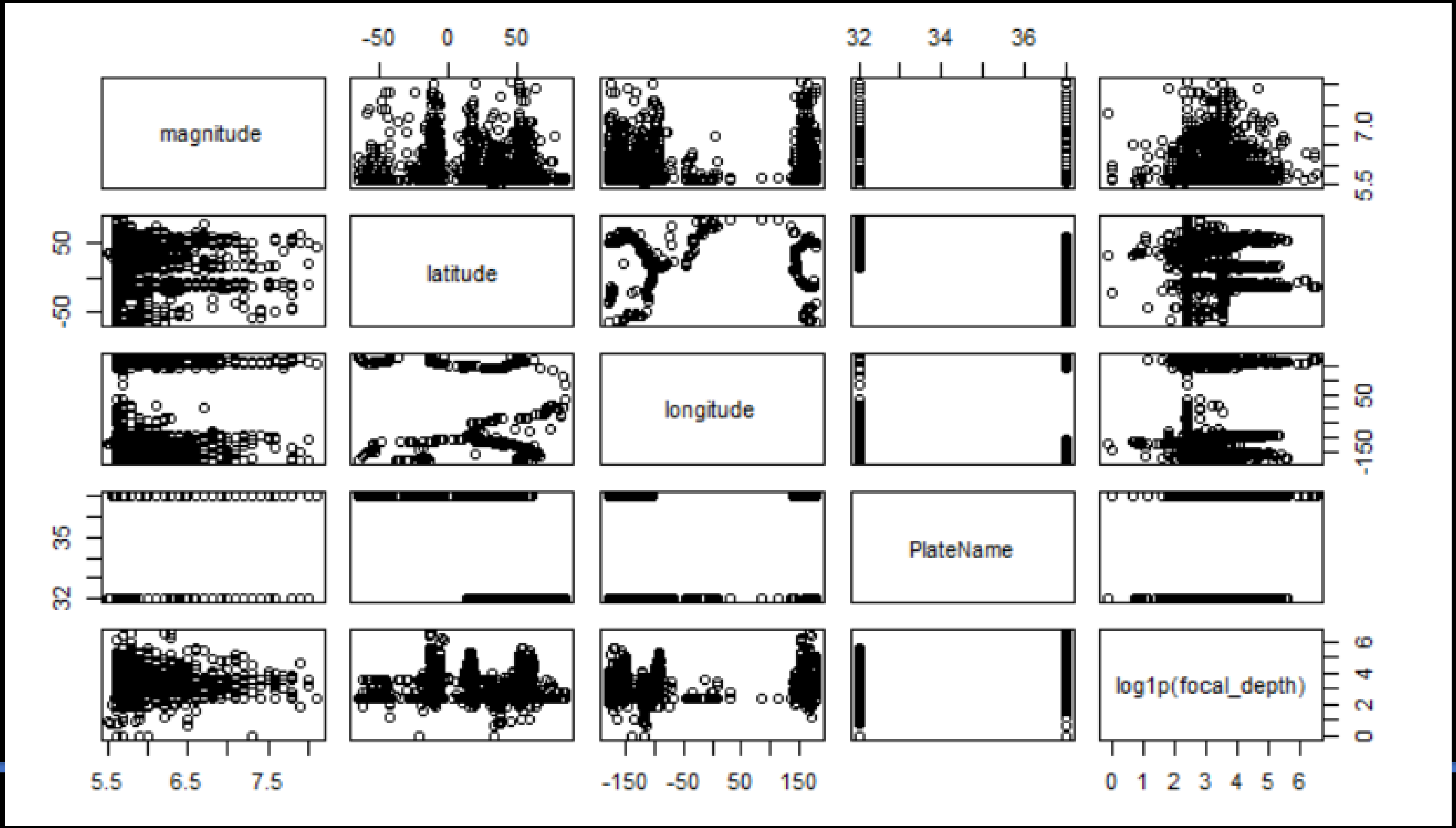
Exploratory Analysis

- Interactions between variables in Earthquakes data set (global)



Exploratory Analysis

- Interactions between variables in NAP Earthquakes data set



Model Building

- Linear Regression Results

Model	Data set	Factors	Adjusted R^2
LRModel1	Earthquakes	Lat + long	0.002018
LRModel2	Earthquakes	lat*long	0.001433
LRModel3	Earthquakes	Lat * long + focal_depth	0.002827
LRModel4	Earthquakes	Lat*long* + focal_depth + PlateName	0.009274
LRModel5	Earthquakes	Lat*long + focal_depth*PlateName	0.01287
LRModel6	Earthquakes	Lat*long*PlateName	0.01507
LRModel7	Nap_Earthquakes	Lat*long + focal_depth	0.001786
LRModel8	Nap_Earthquakes	Lat*long*focal_depth + PlateName	0.009719
LRModel9	Nap_Earthquakes	Lat*long*focal_depth*PlateName	0.008226

Model Building

- Logistic Regression Results

Model	Data set	Factors	df	Residual Deviance
QuakeLog1	Earthquakes	Lat + long + focal_depth	18789	6369.7
QuakeLog2	Earthquakes	Long * focal_depth	18840	6501.7
QuakeLog3	Earthquakes	Focal_depth	18842	6511.9
QuakeLog4	Earthquakes	Lat*long	18842	6524.8
NapQuakeLog1	Nap_Earthquakes	Lat*long	1475	639.23
NapQuakeLog2	Nap_Earthquakes	Lat*long*focal_depth	1471	629.69
NapQuakeLog3	Nap_Earthquakes	Lat + long + focal_depth	1475	638.14
NapQuakeLog4	Nap_Earthquakes	Long*focal_depth	1475	637.64
NapQuakeLog5	Nap_Earthquakes	Focal_depth	1478	645.01

Model Building

- Binomial Regression
- Relative vs. absolute effects

Model	Data set	Residual Deviance	Plogis(x)
bd_model	Earthquakes	640.90	0.0225
Nap_model	Nap_earthquakes	-2.15×10^{-13}	0.0562

Quake Probabilities

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

Model Building

- Poisson Distributions

Model	Data set	Residual Deviance	Prob.
p_model	Earthquakes	5.58×10^{-10}	Varies per plate
Nap_p_model	Nap_earthquakes	4.44×10^{-15}	1.1

Predicted Earthquake counts

	x
(Intercept)	2.0
PlateNameAfrica	3.0
PlateNameAltiplano	5.0
PlateNameAmur	5.5
PlateNameAnatolia	0.5
PlateNameAntarctica	6.0
PlateNameArabia	0.5
PlateNameAustralia	21.0
PlateNameBalmoral Reef	0.0
PlateNameBanda Sea	6.5
PlateNameBirds Head	6.5
PlateNameBurma	4.5
PlateNameCaribbean	6.5
PlateNameCaroline	0.0
PlateNameCocos	0.0
PlateNameConway Reef	0.5
PlateNameEaster	0.0
PlateNameEurasia	30.5
PlateNameFutuna	0.0
PlateNameGalapagos	0.0
PlateNameIndia	5.0
PlateNameJuan de Fuca	1.5
PlateNameJuan Fernandez	0.5
PlateNameKermadec	12.5
PlateNameManus	0.0
PlateNameMaoke	3.5
PlateNameMariana	3.5
PlateNameMolucca Sea	7.0
PlateNameNazca	1.0
PlateNameNew Hebrides	22.5
PlateNameNiuafo'ou	0.5
PlateNameNorth America	31.5
PlateNameNorth Andes	5.0
PlateNameNorth Bismarck	7.0
PlateNameOkhotsk	36.0
PlateNameOkinawa	2.0
PlateNamePacific	42.0
PlateNamePanama	3.0
PlateNamePhilippine Sea	13.5
PlateNameRivera	0.0
PlateNameSandwich	1.0
PlateNameScotia	1.5
PlateNameShetland	0.0
PlateNameSolomon Sea	2.5
PlateNameSomalia	1.0
PlateNameSouth America	38.5
PlateNameSouth Bismarck	11.0
PlateNameSunda	31.5
PlateNameTimor	3.0
PlateNameTonga	4.5
PlateNameWoodlark	4.0
PlateNameYangtze	0.0

Results & Discussion

- Relationship found between magnitude and longitude and magnitude and focal_depth
- Binomial model is best for predicting probability of a major earthquake (magnitude 7.0 or higher) on a given plate
- Poisson model most successful for predicting counts of major earthquakes

Results & Discussion

- Future expansions to analysis:
  - Analyze using a time-series model
  - Add data set with fault zone classifications and analyze for impact of fault zone/stress types
  - Mixed-effects model