Predicting Earthquakes

Using data science to calculate probabilities

Heather Ewton

Foundations of Data Science

Overview

- The Problem
- Background
- Data
- Exploratory Analysis
- Model Building
- 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

- 2 data sets were used:
  - NOAA data set
  - USGS data set

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

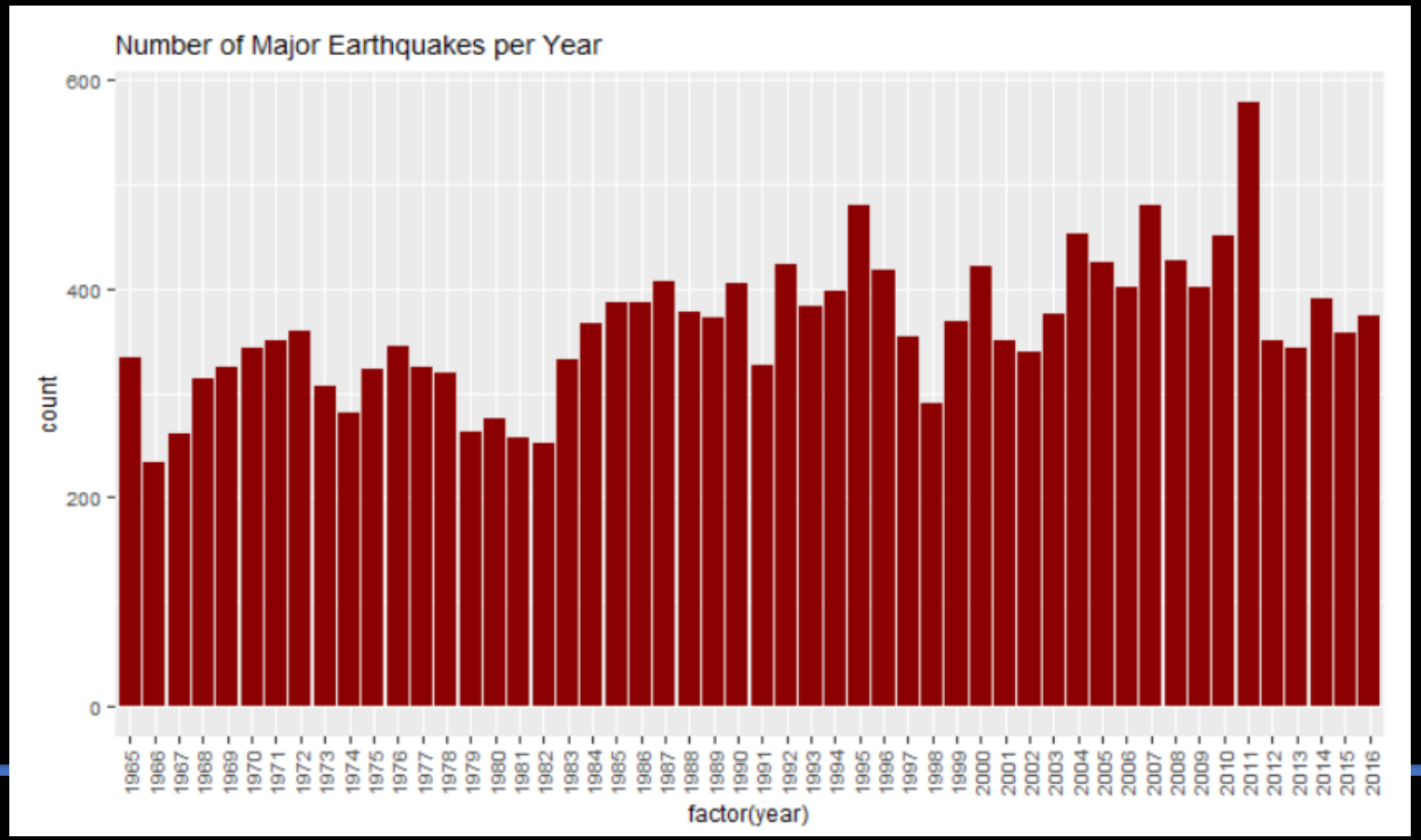
USGS data set

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

- Clean and tidy

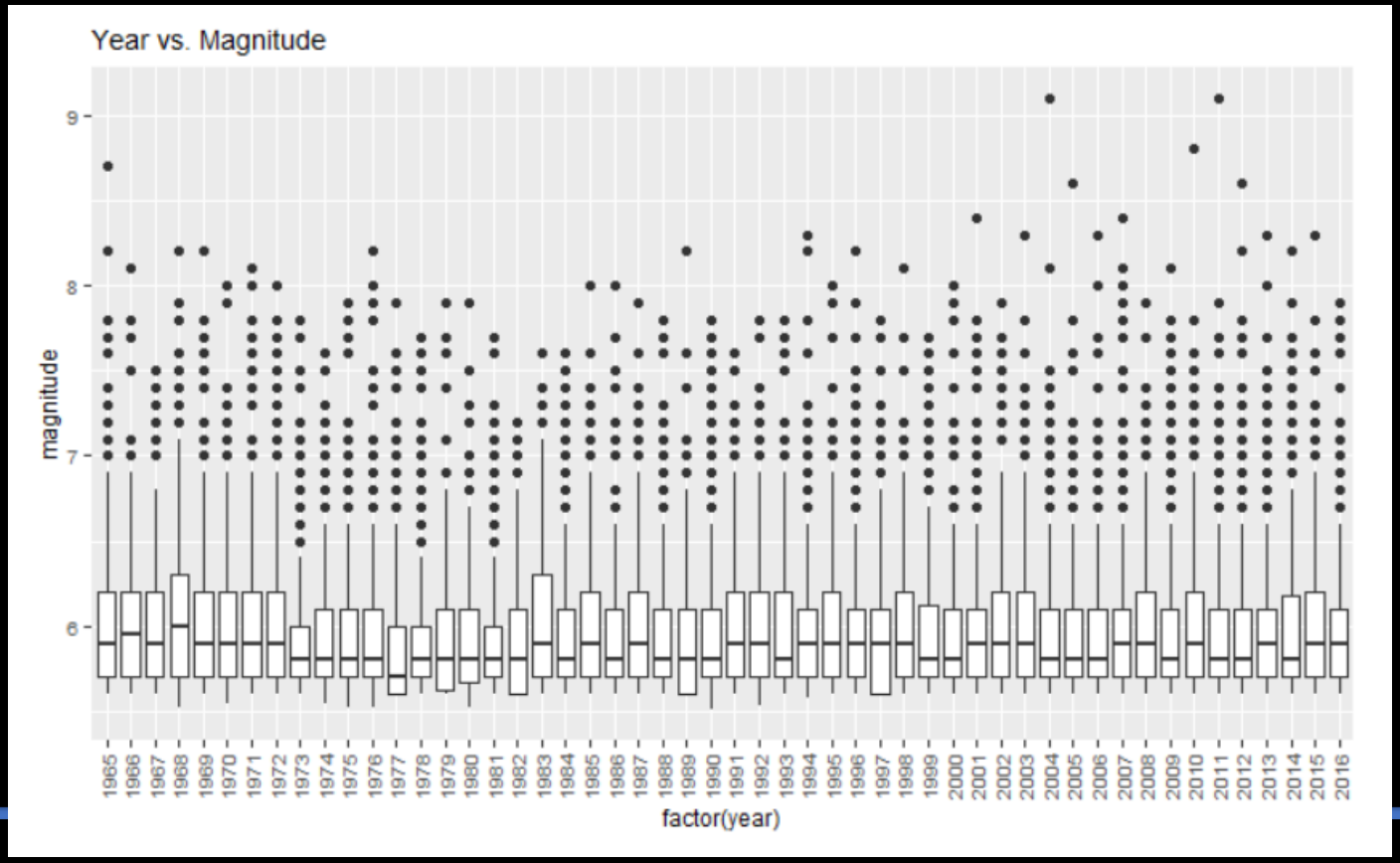
Exploratory Analysis

- Bar graph of significant earthquakes per year



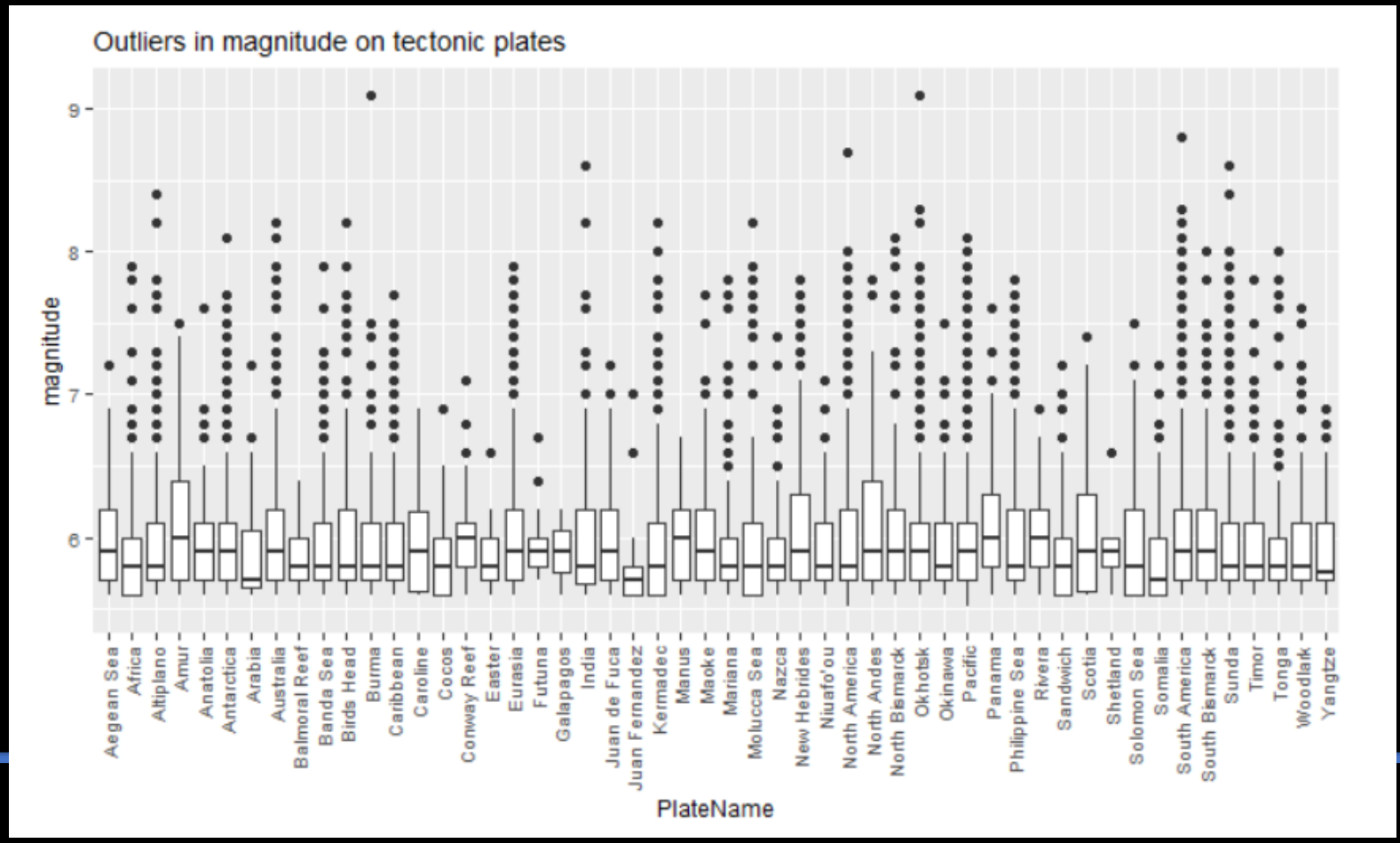
Exploratory Analysis

- Year vs. Magnitude box plot



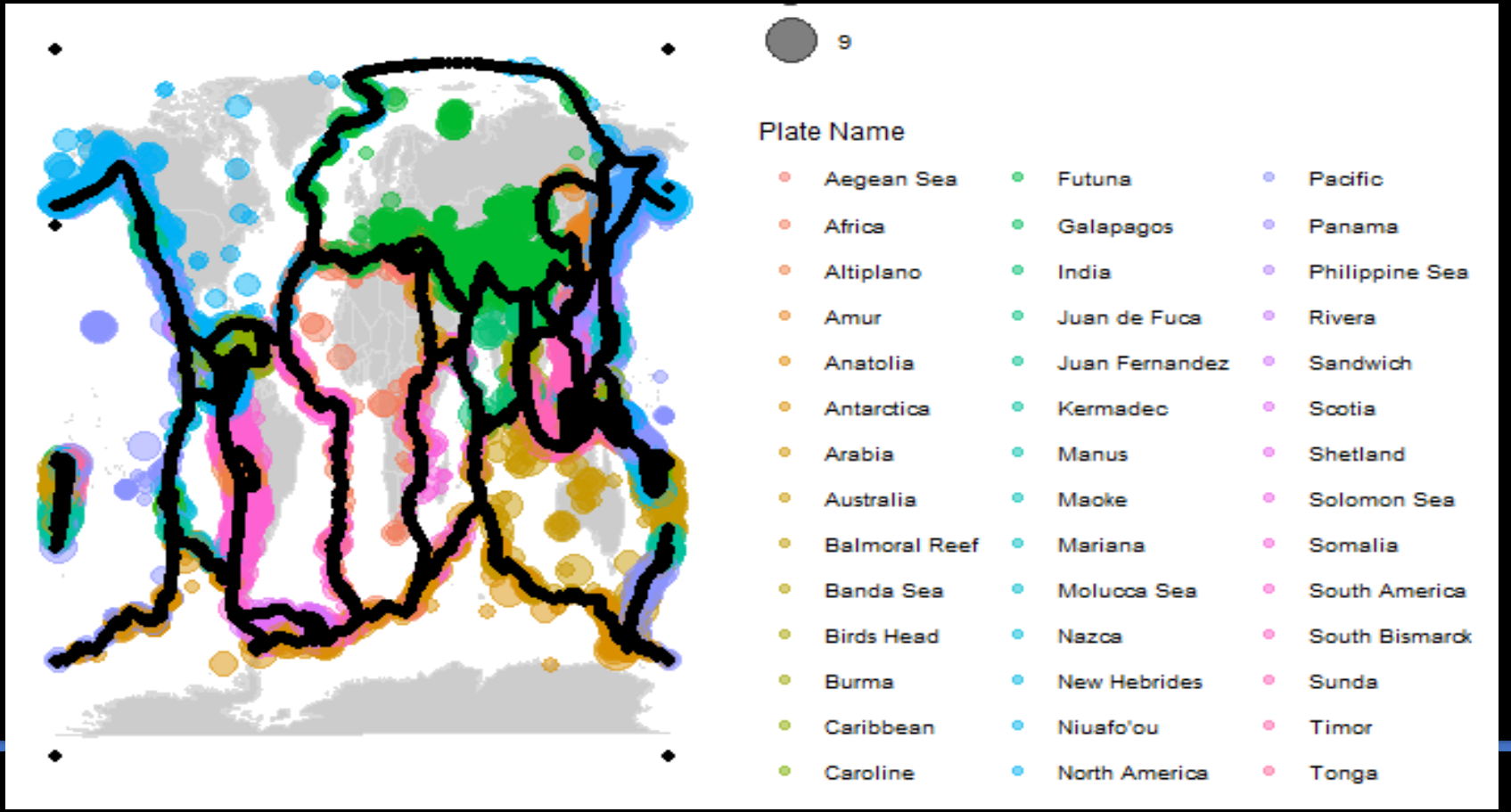
Exploratory Analysis

- Plate Name v. Magnitude



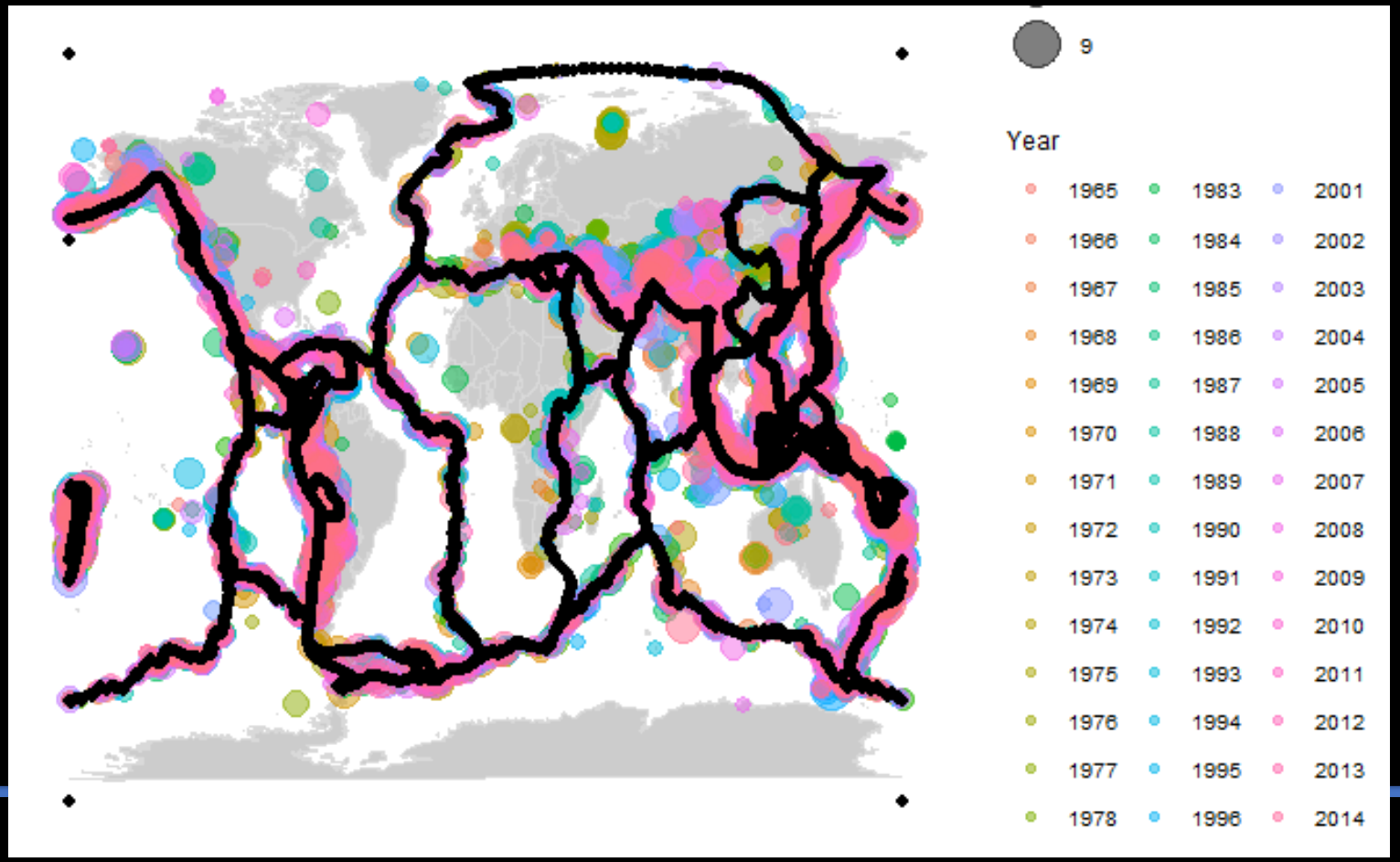
Exploratory Analysis

- Quake plot by plate name, magnitude



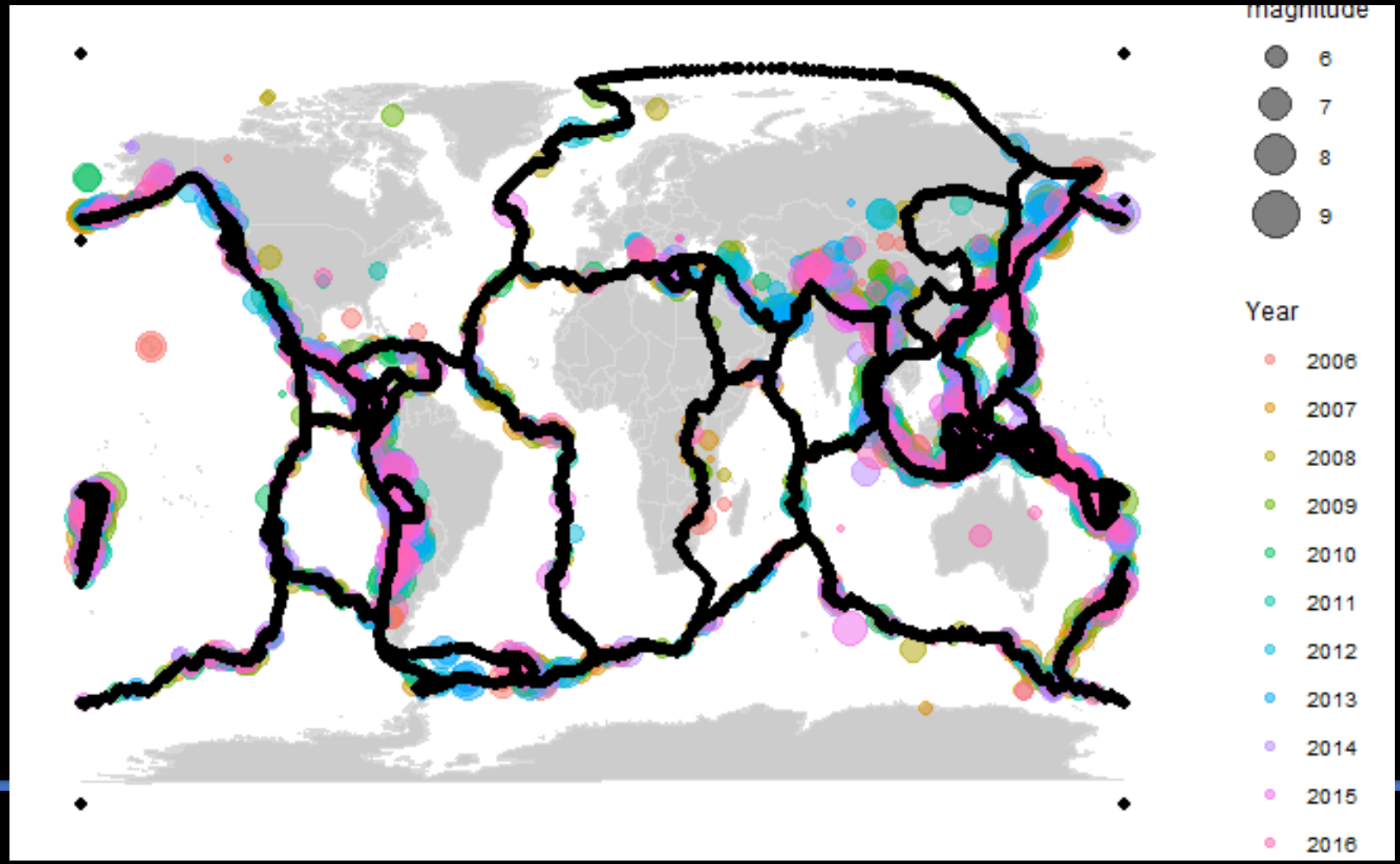
Exploratory Analysis

- Quake plot by year, size = magnitude



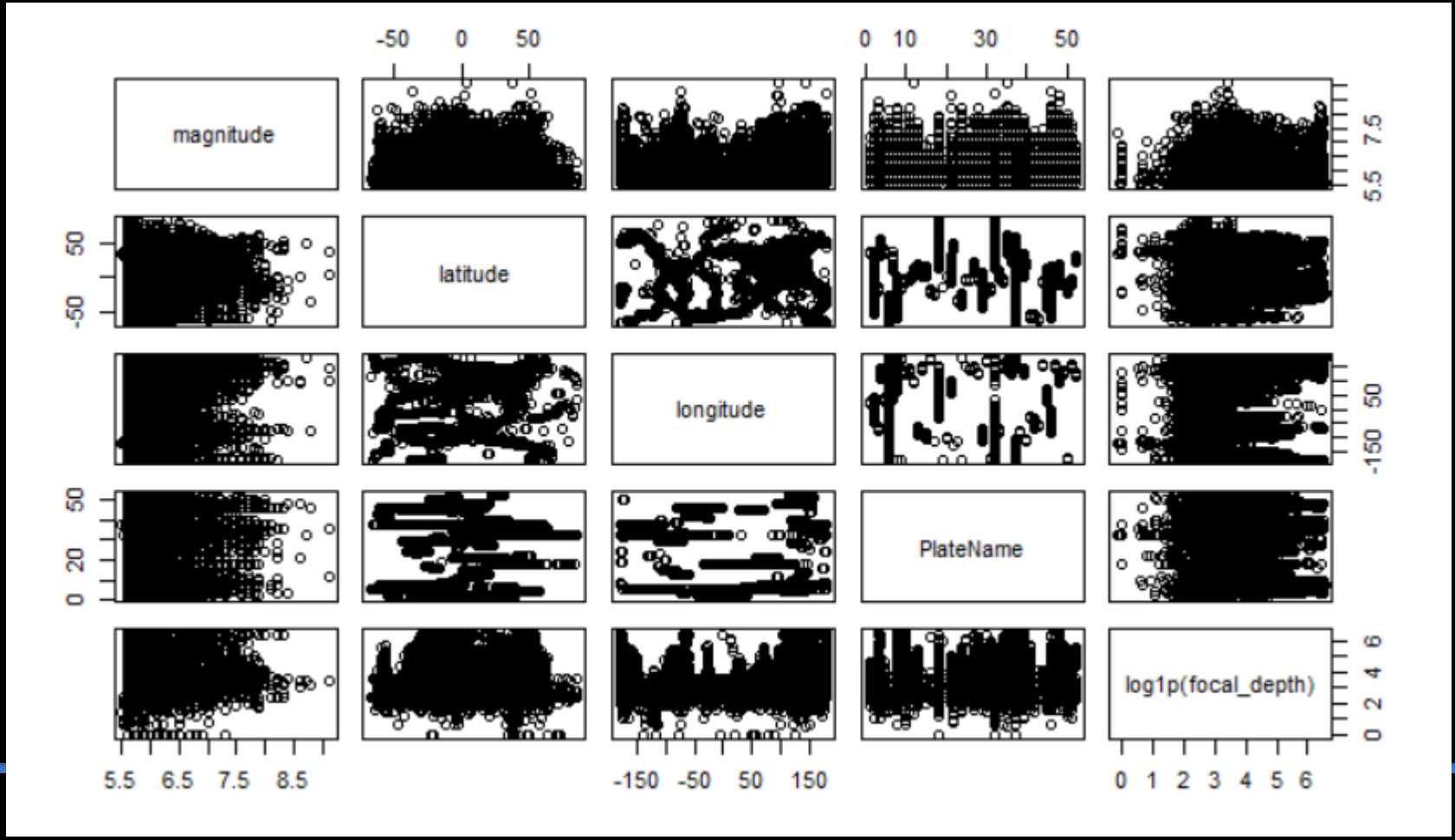
Exploratory Analysis

- Exploring the last ten years of data in detail



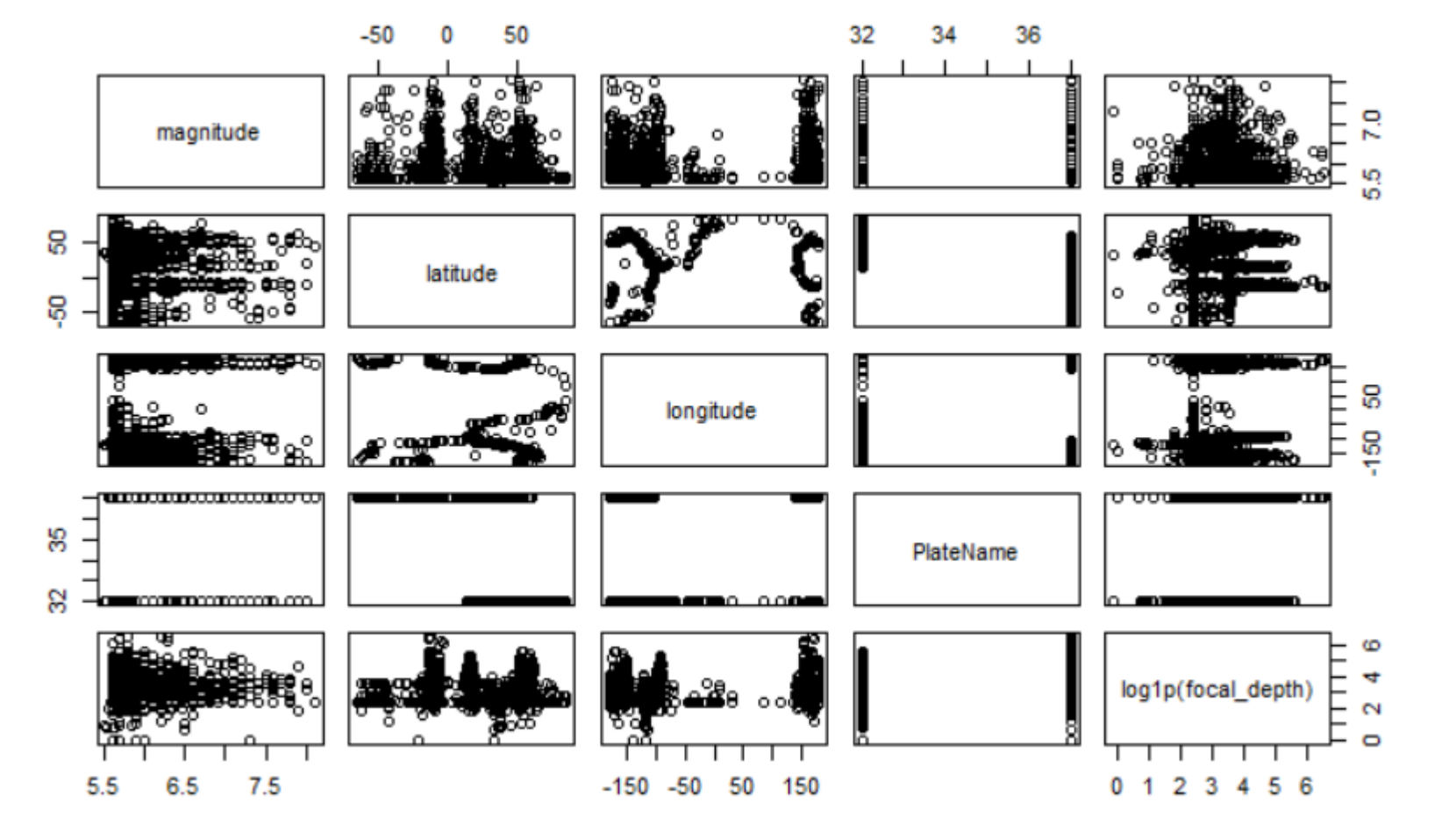
Exploratory Analysis

- 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

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

Results & Discussion

- Relationship found between magnitude and longitude and magnitude and focal_depth
- Poisson model is best for predicting probability of a major earthquake (magnitude 7.0 or higher)
- Per Poisson model, the Pacific plate has the highest rate of probability of a major earthquake, at 10%

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