CSE3020 Data Visualization
Project Report
Detecting and Visualizing Earthquake Data

Team members:

- R. Tharun Murugesh, 18BCE1059
- Pravalikka Putha, 18BCE1079
- Adrian Nirmal Andrew, 18BCE1164

Abstract

Earthquakes cause large amounts of destruction to life and property every single year across the world. Damage that isn't directly caused by seismic activity can also occur due to other events like floods, landslides, tsunamis and surface rupture. Despite technological advances that allow structures to withstand an earthquake, it still remains a feared natural disaster.

Hence, in order to be better informed on past earthquakes, as well as potential future occurrences, we aim to create a set of informational material composed of historical earthquake data, associated with its location, as well as a set of predicted seismic activity data for the locations.

Introduction

Through this project, analysis on the patterns of historical data on seismic activity was done with the intention of asserting the occurrence of earthquakes and in doing so, the approach taken involved the study on the close relation of tectonic plates which are massive irregularly shaped slabs of solid rock, primarily composed of continental and oceanic lithosphere with relation to the past data of earthquakes. The tectonic plates slide past each other and at times, get stuck which leads to large amounts of built-up energy, eventually being released, causing an earthquake. An earthquake is generally characterized by the depth and the magnitude with the depth being significant as the depth of the earthquake gives us insights into the information on the Earth's structure, more specifically the mechanics and characteristics of the deformation of the Earth's surface and the tectonic setting where the earthquake is occurring by plotting the location and the depth. The magnitude or the size of the earthquake measured by a seismograph besides giving an insight into the intensity of the earthquake may be able to explain the roughness of the fault surfaces.
This paper explores the relationship between the tectonic plates, taking into account the depth, magnitude and its location to allow us to form a prediction model to help in the

understanding of earthquake data both statistically and graphically. Our first priority was in recognizing the most common epicenters or the location below the earth's surface where the earthquake occurs, for us to understand which of the regions are most susceptible to earthquakes as well as measure its severity.

We used the following Datasets in our analysis:
Dataset 1: USGS Historical earthquake data: This dataset contains the consolidated datat from registered earthquakes around the world from 1970 to March 2018.
The columns are: time, latitude ,longitude ,depth ,mag ,magType ,nst ,gap ,dmin ,rms ,et ,id ,updated ,place ,Type , horizontalError , depthError magError, magNst ,Status, locationSource and magSource.
Dataset 2: Tectonic Plates Data : This data contains latitude and longitude data that completely encloses 56 tectonic plates.
The columns are: longitude, latitude and border IDs

Literature Survey

Sampling historical data has always been considered for the prediction of earthquakes, specifically, data related to the topology of a specific location, such as soil density, altitude, and data of previous earthquakes in the region.

Earthquake magnitude prediction in Hindukush region using machine learning techniques by K. M. Asim1 • F. Martı´nez-A´ lvarez2 • A. Basit3 • T. Iqbal1in / Published online: 8 September 2016 Springer Science+Business Media Dordrecht 2016 :

This research paper employed four learning techniques (PRNN, RNN, Random Forest and LPBoost encumber) which were used to predict earthquakes in the subducting Hindukush region out of which mathematically calculated 8 parameters were chosen to be used as input classifiers. The results showed that Linear Programming Boost ensemble classifiers had better results considering sensitivity while Pattern Recognition has the least number of false alarms. The crux of this paper describes how earthquake occurrence prediction can be based on geographical facts and sophisticated modeling and learning of Machine Learning Algorithms.

Earthquake Prediction Using Expert Systems: A Systematic Mapping Study by Rabia Tehseen, Muhammad Shoaib Farooq * and Adnan Ab published on 19 March 2020

This paper combed through 70 systematically selected peer reviewed high quality research articles pertaining to earthquake prediction using ES and discussed the different tools used, datasets employed, and various methods (rule based, fuzzy, and machine learning) used in seismic prediction. Research papers were analyzed and put against all the different Algorithms defined on variety of parameters (Deterministic, Probabilistic, Analytical Work, Global Approximation, Numerical Experiment, Exploration with actual forecasts, Success Achieved, Characteristic Earthquake, Precursors Zone

Studied) =. This paper found that there is an increased trend in applying FES and ML in long term Earthquake prediction as well as that while older papers focused on precursors to finding the predictions, newer papers took into account various other factors such as natural calamities like volcanic eruptions, hurricanes , etc.

Earthquake Prediction: An Overview by Hiroo Kanamori of California Institute of Technology, Pasadena, California, (2003). 72 Earthquake prediction: An overview. International Handbook of Earthquake and Engineering Seismology, 1205–1216. doi:10.1016/s0074-6142(03)80186-9

This article looked into the scientific aspect of earthquake prediction research such as long term forecast. This constituted in understanding the seismic gap method and the stress transfer with various examples of earthquakes and how they reacted. The article also ventured into the Short term prediction in specifying the time, place, and magnitude of the earthquake in question with sufficiently high reliability and probability by observing the precursors and anomalous phenomena and finally looking into short term and long term strategies which constituted of understanding the various mitigation strategies in place in various countries like, USA , Taiwan and Mexico.

Methods/Techniques/Algorithms

For this project, we will be using two datasets, one containing information about historical earthquake intensity, and the other containing information about Earth's tectonic plates.

First, the data is combined to get information for every tectonic plate border or intersection. This is done by first taking the Tectonic plate dataset, then finding every single unique point in it. Then, every Plate ID associated with each individual point is combined to create a Border ID. Finally, the centroid for each individual Border is found by taking the average of every single point associated with it.

Next, for each individual data point in the Earthquake activity dataset, the coordinates are classified by finding the Euclidean distance between it and each Border category centroid. The category with the lowest distance gets chosen for each data point. Following this, the data is segregated on the basis of the border category.

Using this pre-processed data, we are creating two Time Series models for each border region, i.e. for Magnitude and Depth of the earthquake.

Datasets Used

- Historical earthquake data obtained from USGS: https://www.kaggle.com/danielpe/earthquakes
- Tectonic Plate Boundaries: https://www.kaggle.com/cwthompson/tectonic-plate-boundaries

Preliminary Data Analysis

We visualized the data for better understanding and for observing how the the magnitude and depth correlated and act against each other and with other variables

The R code is as below:

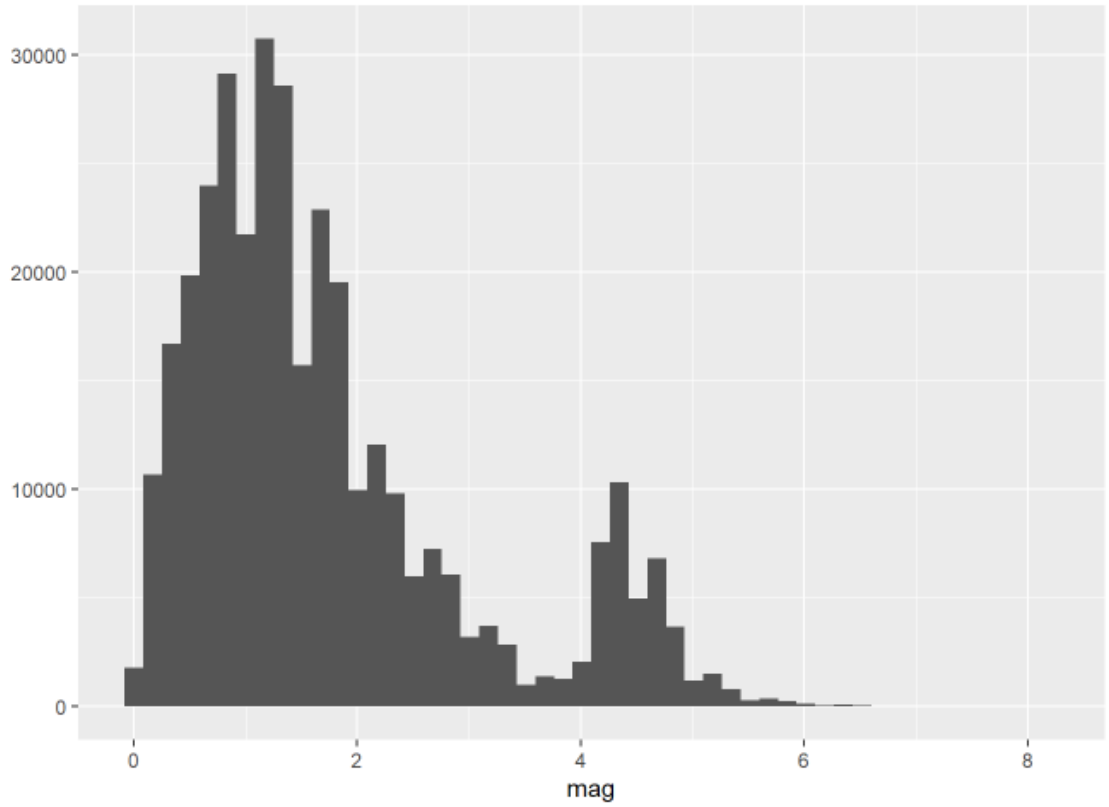
Loading the data:

```
df <-
read.csv('C:/Users/anith/OneDrive/Desktop/classified_3L_new.csv',strin
gsAsFactors = FALSE)
summary(df)
```

```
     time              latitude        longitude          depth
 Length:345971      Min.   :-79.98   Min.   :-180.0   Min.   :  0.001
 Class :character   1st Qu.: 33.67   1st Qu.:-150.0   1st Qu.:  4.600
 Mode  :character   Median : 38.81   Median :-121.5   Median :  9.970
                    Mean   : 40.03   Mean   :-110.2   Mean   : 28.025
                    3rd Qu.: 59.70   3rd Qu.:-116.5   3rd Qu.: 22.480
                    Max.   : 87.00   Max.   : 180.0   Max.   :679.120
      mag            border
 Min.   :0.010   Length:345971
 1st Qu.:0.830   Class :character
 Median :1.400   Mode  :character
 Mean   :1.728
 3rd Qu.:2.200
 Max.   :8.200
```

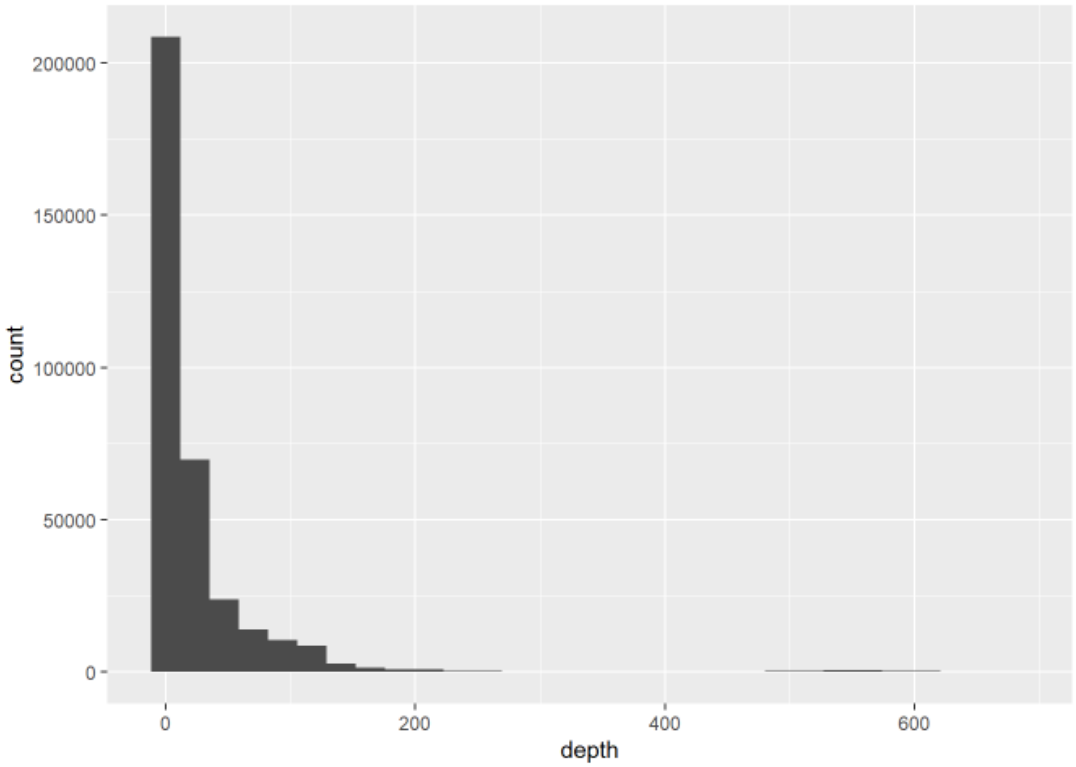
Histogram of the Magnitudes of the Earthquakes

```
qplot(mag, data = df, bins = 50)
```



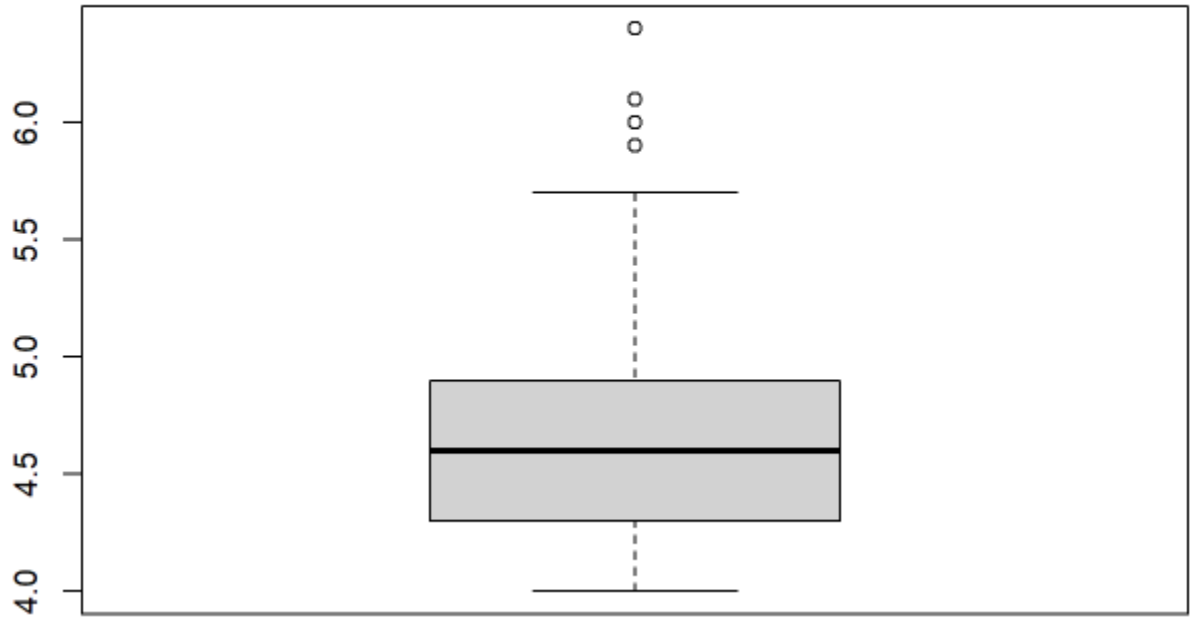
Histogram with frequency of depth

```
ggplot(df) + geom_histogram(aes(x = depth), fill = 'grey30')
```



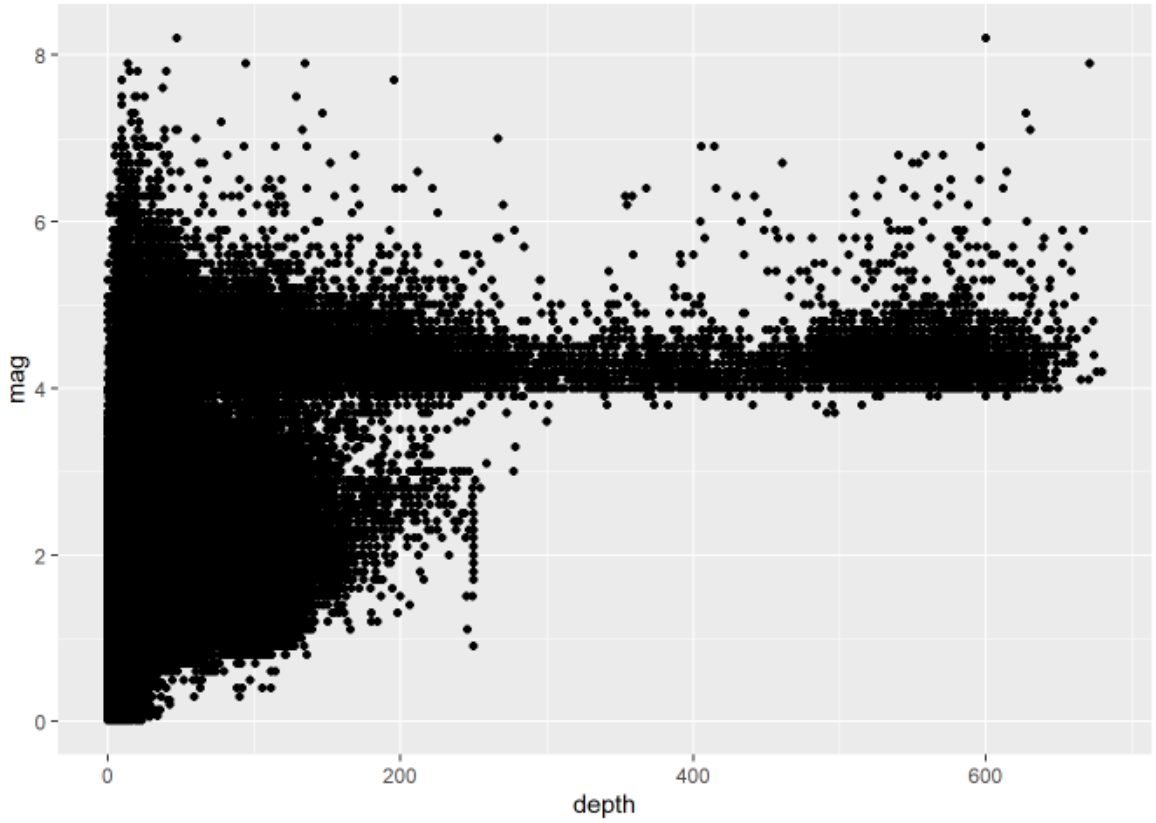
Boxplot of Earthquake magnitudes

```
boxplot(quakes$mag)
```



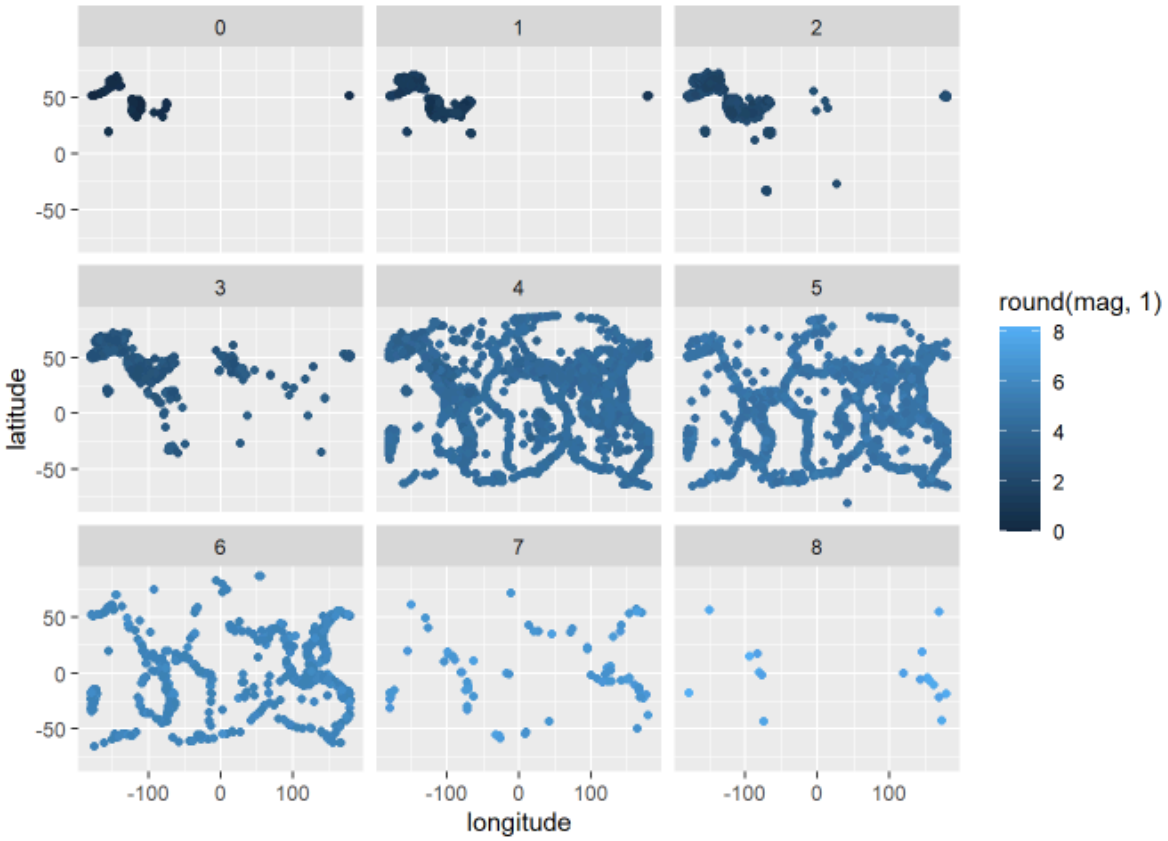
Scatter plot of quake depth and magnitude

```
ggplot(df) + geom_point(aes(x = depth, y = mag))
```



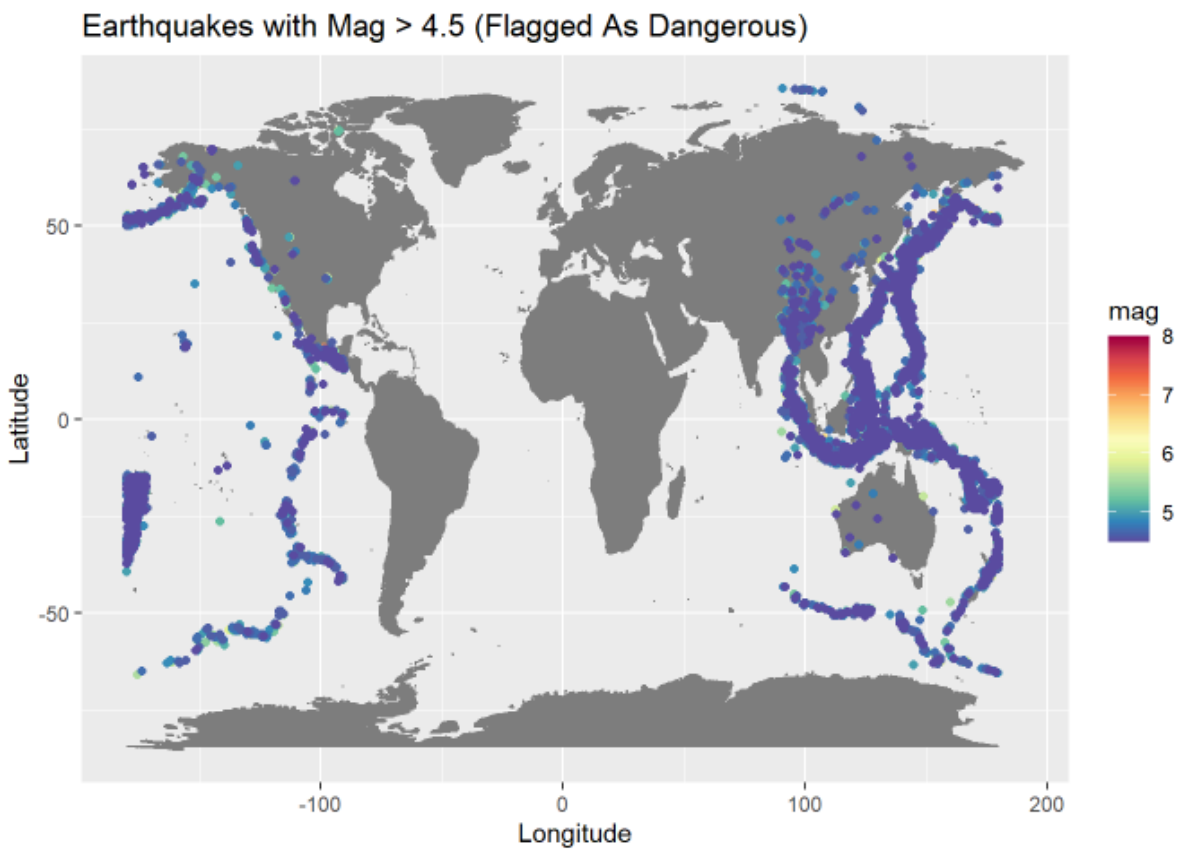
Faceted scatter plot of magnitude by longitude and latitude

```
ggplot(df) + geom_point(aes(x = longitude, y = latitude, color =
round(mag,1))) + facet_wrap(~round(df$mag,0))
```



Earthquakes with Mag>=4.5 (Flagged As Dangerous) - Plotting

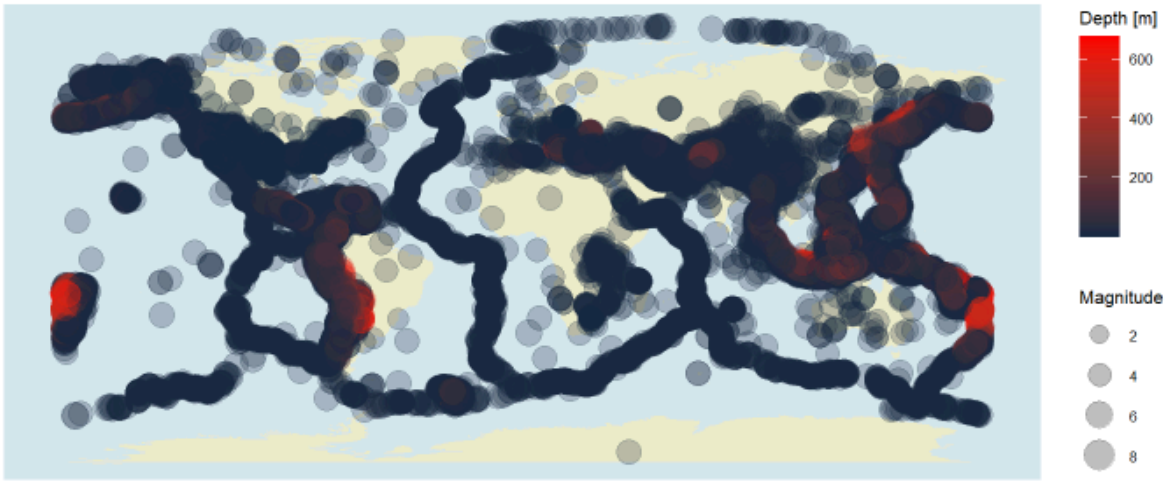
```
eq45 <- subset(df, df$mag >= 4.5)

eq45.sorted <- arrange(eq45, desc(mag))
eq45.mod <- select(eq45.sorted, latitude, longitude, mag)
inside <- filter(eq45.mod, between(longitude, -90, 90),
between(latitude, -180, 180))
eq45.mod <- setdiff(eq45.mod, inside)
wmap <- borders("world", colour = "gray50", fill = "gray50")
eq45_map <- ggplot() + wmap
eq45_map <- eq45_map + geom_point(data = eq45.mod, aes(x =
as.numeric(longitude),
    y = as.numeric(latitude), colour = mag)) + ggtitle("Earthquakes
with Mag > 4.5 (Flagged As Dangerous)") +
    xlab("Longitude") + ylab("Latitude")
myPalette <- colorRampPalette(rev(brewer.pal(11, "Spectral")))
sc1 <- scale_colour_gradientn(colours = myPalette(100), limits =
c(4.5, 8))
eq45_map + sc1
```



Plotting location of the earthquakes as points, with magnitudes indicated by the sizes of the points and depths given by their colour

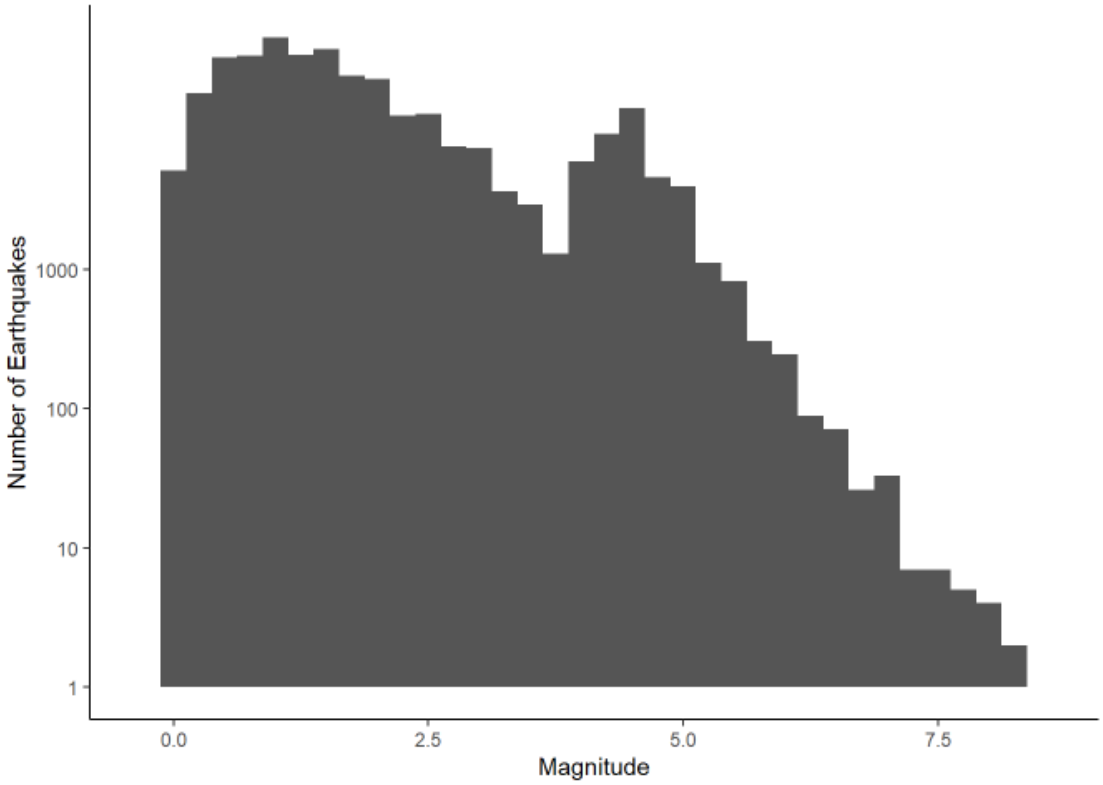
```
world.map <- map_data("world")

 ggplot() +
   geom_polygon(data = world.map, aes(x = long, y = lat, group =
group),
                fill = "#EEEECC") +
   geom_point(data = df, alpha = 0.25,
              aes(x = longitude, y = latitude, size = mag, colour =
depth)) +
   labs(x = NULL, y = NULL) +
   scale_colour_gradient("Depth [m]", high = "red") +
   scale_size("Magnitude") +
   coord_fixed(ylim = c(-82.5, 87.5), xlim = c(-185, 185)) +
   theme_classic() +
   theme(axis.line = element_blank(), axis.text = element_blank(),
```

```
    axis.ticks = element_blank(),
    plot.margin=unit(c(3, 0, 0, 0),"mm"),
    legend.text = element_text(size = 6),
    legend.title = element_text(size = 8, face = "plain"),
    panel.background = element_rect(fill='#D6E7EF'))
```



Distribution of Earthquake Magnitudes using a log scale

```
ggplot(df, aes(x = mag)) +
  xlab("Magnitude") + ylab("Number of Earthquakes") +
  stat_bin(pad = TRUE, binwidth = 0.25) +
  scale_y_log10(breaks = c(1, 10, 100, 1000)) +
  theme_classic()
```



Correlation Between Magnitude and Two Depth Curves depth curve:

splitting depth curves to study them both

```
low_depth_quakes <- df[which(df$depth > 400), ]
high_depth_quakes <- df[which(df$depth <= 400), ]
```

pearson correlation between magnitude high depth curve

```
cor(high_depth_quakes$depth, high_depth_quakes$mag)
```

```
## [1] -0.3296122
```

pearson correlation between magnitude and low depth curve

```
cor(low_depth_quakes$depth, low_depth_quakes$mag)
```

```
## [1] 0.001745948
```

Proposed Model, Algorithm and Implementation

Our project's main focus is on creating an ARIMA model that can forecast the magnitude and depth of future earthquakes in each specific border region. Preprocessing involved associating each historical data point (for magnitude and depth) with a border plate ID, separating the data on the basis of this ID, then performing time series prediction using an ARIMA model to predict future seismic activity data.

The R code for these steps is as follows:

Loading data:

```
earthquakes <- na.omit(read.csv("earthquakes.csv")) %>%
filter(mag >= 0 & depth >= 0) %>% select(time, latitude,
longitude, depth, mag)
```

	time	latitude	longitude	depth	mag
1	1970-01-01T00:00:00.0Z	37.00350	-117.9968	0.000	0.00
2	1970-01-01T00:00:00.0Z	35.64279	-120.9336	5.000	1.99
3	1970-01-01T00:00:00.0Z	34.16452	-118.1850	0.000	0.00
4	1970-01-01T00:00:00.0Z	33.83649	-116.7819	0.000	0.00
5	1970-01-01T00:00:00.0Z	33.20848	-115.4770	5.000	0.00
6	1970-01-01T00:00:00.0Z	32.66356	-116.1050	0.000	0.00
7	1970-01-01T00:00:00.0Z	35.35445	-115.4849	0.000	0.50
8	1970-01-01T15:13:21.040Z	32.70717	-115.4170	6.000	2.75
9	1970-01-01T17:11:00.000Z	-29.40000	-177.1690	35.000	5.60
10	1970-01-01T19:49:24.730Z	37.43333	-118.7435	6.000	3.69
11	1970-01-02T08:58:51.550Z	46.74950	-119.3715	1.869	1.50
12	1970-01-02T10:45:20.570Z	34.20600	-119.6957	6.000	3.14

```
tectonic <- unique(read.csv("tectonic_plates.csv"))
```

	plate	lat	lon
1	am	30.754	132.824
2	am	30.970	132.965
3	am	31.216	133.197
4	am	31.515	133.500
5	am	31.882	134.042
6	am	32.200	134.691
7	am	32.326	135.026
8	am	32.460	135.356

Identify each unique tectonic plate border:

```
uniquePoints <- data.frame(row.names = 1:nrow(tectonic))

uniquePoints$lat <- tectonic$lat

uniquePoints$lon <- tectonic$lon

uniquePoints <- unique(uniquePoints)

uniquePoints$plates <- rep(NA, nrow(uniquePoints))

#Add plate border points to categories:

for(i in c(1:nrow(uniquePoints))) {

    tempU <- uniquePoints[i,]

    tPlates <- as.list(tectonic %>% filter(lat == tempU$lat &
lon == tempU$lon) %>% select(plate))

    uniquePoints[i,]$plates = tPlates

}

#Find mean centroids for each border:

borderCentroids <- uniquePoints %>% group_by(plates) %>%
summarise(c_lat = mean(lat), c_lon = mean(lon))
```

	plates	c_lat	c_lon
1	am ON ps	30.7540000	132.82400
2	am ps	32.6338667	135.76473
3	am OK ps	35.0340000	138.67400
4	am OK	43.6044500	139.64415
5	am eu OK	54.0350000	142.40900
6	am eu	47.3978462	118.13395
7	am eu yz	31.4710000	123.36400
8	am yz	31.3231111	126.36622
9	am ON yz	31.7930000	129.44200
10	am ON	31.6766250	131.33363

Classify Data points:

```
#Euclidean distance function:
eucliDist <- function(lat1, lon1, lat2, lon2) {
    sqrt((lat1 - lat2) ^ 2 + (lon1 - lon2) ^ 2)
}
#Classify points:
earthquakes$border <- rep(NA, nrow(earthquakes))
for(i in c((nrow(earthquakes) - 71354):1)) {
    minDist <- Inf
    minCat <- NA
    tEQ <- earthquakes[i,]
    for(j in c(1:nrow(borderCentroids))) {
      tBC <- borderCentroids[j,]
      tDist <- eucliDist(tEQ$latitude, tEQ$longitude, tBC$c_lat,
tBC$c_lon)
      if(tDist < minDist) {
```

```
            minDist <- tDist
            minCat <- tBC$plates
        }
    }
    earthquakes[i,]$border <- minCat
}
classified2 <- earthquakes[which(!is.na(earthquakes$border)),]
#Convert border lists to strings:
#This is required before storing the data in CSV format:
p1 <- function(ent) {
    paste(unlist(ent), collapse = " ")
}
classified2$border = lapply(classified2$border, p1)
```

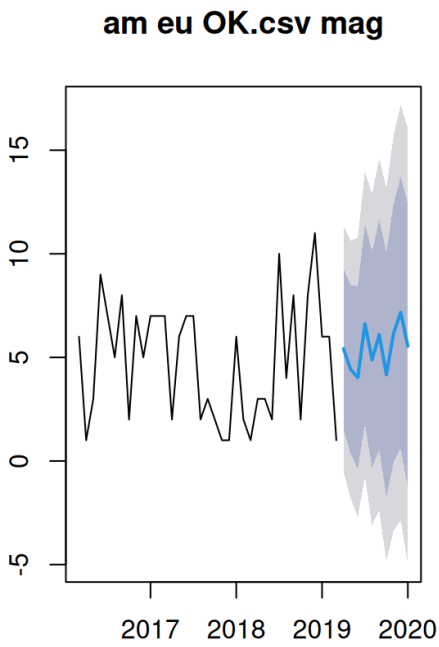
	time	latitude	longitude	depth	mag	border
1	2016-03-22 22:05:56	63.49090	-150.81780	6.000	0.80	jf pa
2	2016-03-22 22:13:51	53.23020	-166.94430	31.100	1.80	jf pa
3	2016-03-22 22:14:47	-2.87700	129.80230	10.000	4.60	BH BS
4	2016-03-22 22:20:00	38.46590	-118.37710	4.700	0.50	jf na
5	2016-03-22 22:25:39	38.46710	-118.37740	4.700	1.30	jf na
6	2016-03-22 22:30:57	38.46760	-118.37750	5.300	0.40	jf na
7	2016-03-22 22:31:12	36.00333	-117.80067	1.870	0.29	jf na
8	2016-03-22 22:34:49	61.99210	-150.72410	67.900	1.30	jf pa
9	2016-03-22 22:35:11	65.38780	-144.71980	11.900	1.40	jf pa
10	2016-03-22 22:35:42	58.68490	-137.92200	6.800	1.70	jf pa

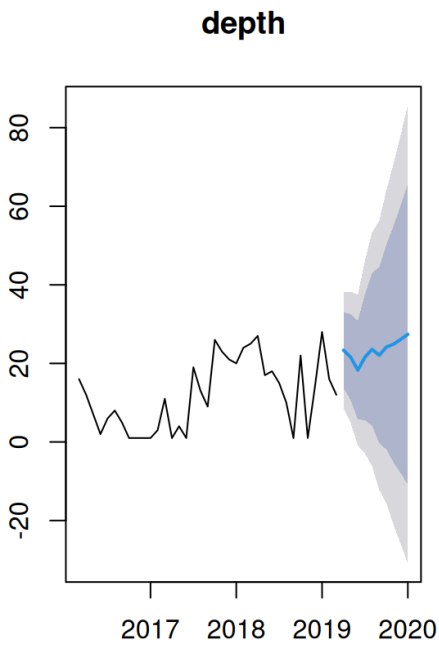
After saving the classified data, plotting ARIMA model predictions of depth and magnitude for all border regions:

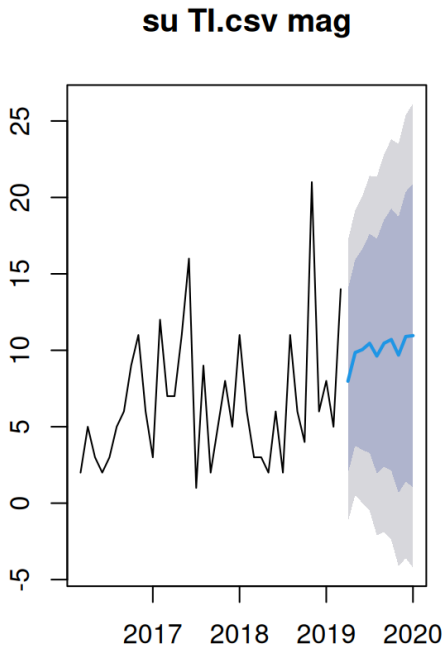
```
#Load forecast library:
```

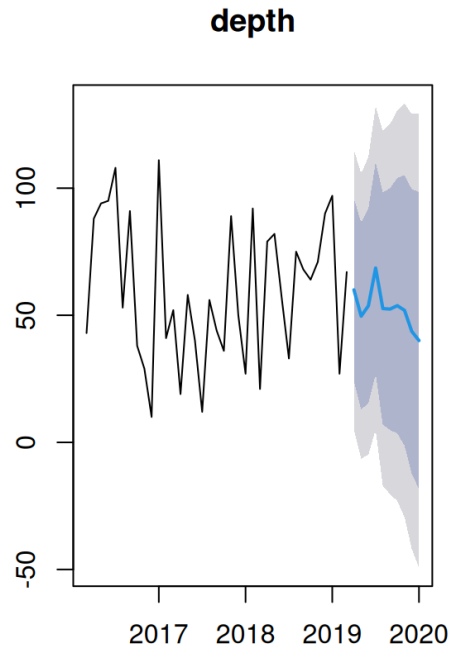
```
library(forecast)

#Iterate through files:

for(tFile in list.files("./sorted3.5L/")) {

    tryCatch({

      #Read data:

      tData <- read.csv(paste("./sorted3.5L/", tFile, sep = ""))

      #Create time series:

      magTS <- ts(factor(tData$mag), start = c(2016, 3), end =
c(2019, 3), frequency = 12)

      depthTS <- ts(factor(tData$depth), start = c(2016, 3), end
= c(2019, 3), frequency = 12)

      #Create ARIMA model:

      magFit <- arima(magTS, order = c(3, 2, 1), seasonal = c(1,
0, 0), method = "ML")

      depthFit <- arima(depthTS, order = c(3, 2, 1), seasonal =
c(1, 0, 0), method = "ML")

      #Forecast:

      magFc <- forecast(magFit, 10)

      depthFc <- forecast(depthFit, 10)

      #Plot graph:

      layout(matrix(1:2, ncol = 2))

      plot(magFc, main = paste(tFile, "mag"))

      plot(depthFc, main = "depth")

    }, error = function(e){})

}
```

Examples:


am eu OK.csv mag
15
10
5
0
-5
2017
2018
2019
2020


depth
80
60
40
20
0
-20
2017
2018
2019
2020


su TI.csv mag
25
20
15
10
5
0
-5
2017
2018
2019
2020


depth
100
50
0
-50
2017
2018
2019
2020

Finally, plotting ARIMA model predictions for the coordinates for the locations of future earthquakes, along with their corresponding Plate border region:

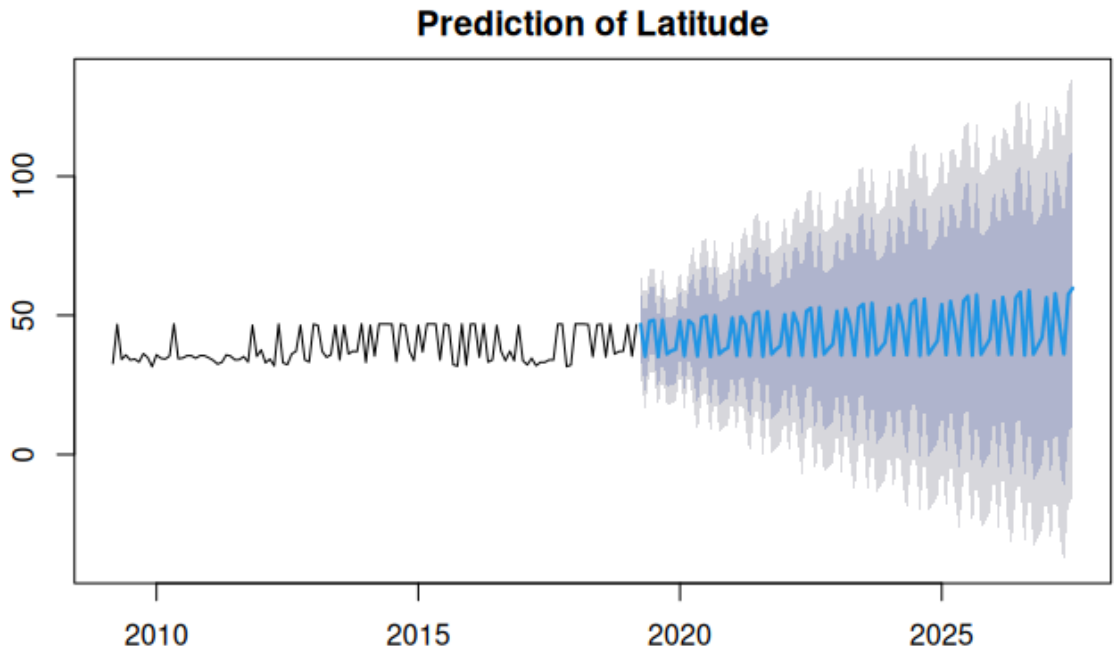
Setup:

```
# Load Libraries:

library(dplyr)

library(forecast)

library(ggplot2)

#Read border centroid data:

borderCentroids <- read.csv("borderCentroids.csv")

#Euclidean distance:

eucliDist <- function(lat1, lon1, lat2, lon2) {

    sqrt((lat1 - lat2) ^ 2 + (lon1 - lon2) ^ 2)

}

#Function to classify points:

getCentroid <- function(lat, lon) {

    minDist <- Inf

    minCat <- NA

    for(i in c(1:nrow(borderCentroids))) {

      tBC <- borderCentroids[i,]

      tDist <- eucliDist(lat, lon, tBC$c_lat, tBC$c_lon)

      if(tDist < minDist) {

           minDist <- tDist

           minCat <- tBC$plates

      }

    }

    return(minCat)
```

```
}
```

Latitude prediction:

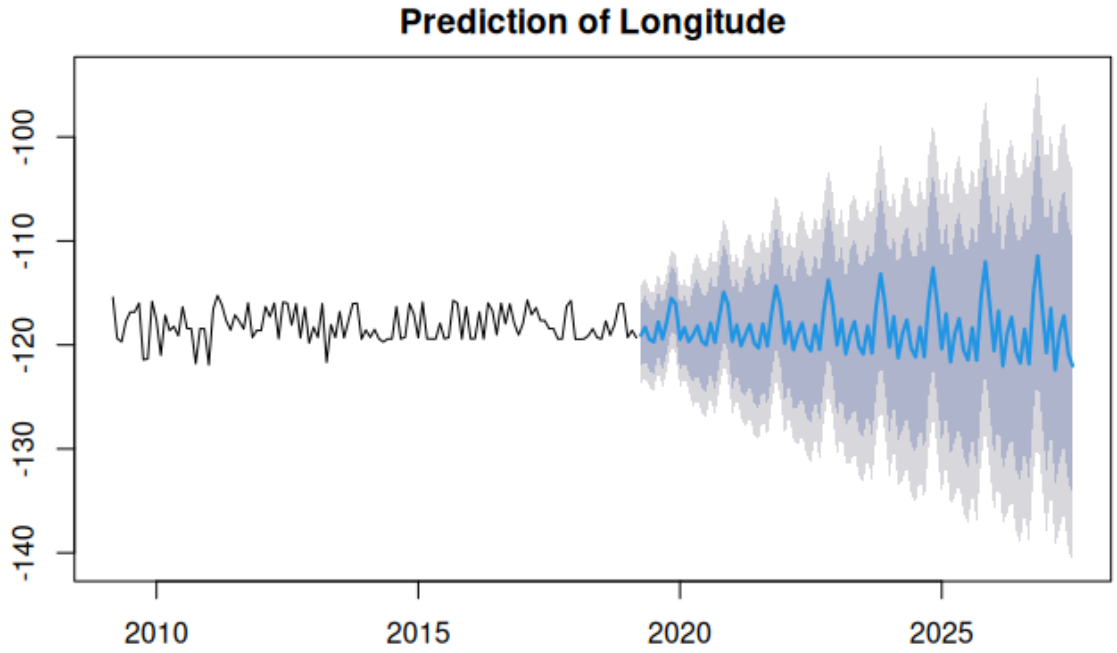
```
latitudeTS <- ts(earthquakes$latitude, start = c(2009, 3), end =
c(2019, 3), frequency = 12)

latitudeFit <- arima(latitudeTS, order = c(0, 0, 0), seasonal =
c(0, 2, 1), method = "ML")

latitudeFc <- forecast(latitudeFit, 100)

latSumm <- summary(latitudeFc)

plot(latitudeFc, main = "Prediction of Latitude")
```



Longitude prediction:

```
longitudeTS <- ts(earthquakes$longitude, start = c(2009, 3), end
= c(2019, 3), frequency = 12)

longitudeFit <- arima(longitudeTS, order = c(0, 0, 0), seasonal
= c(0, 2, 1), method = "ML")

longitudeFc <- forecast(longitudeFit, 100)
```

```
lonSumm <- summary(longitudeFc)

plot(longitudeFc, main = "Prediction of Longitude")
```



Finally, labelling the plate borders:

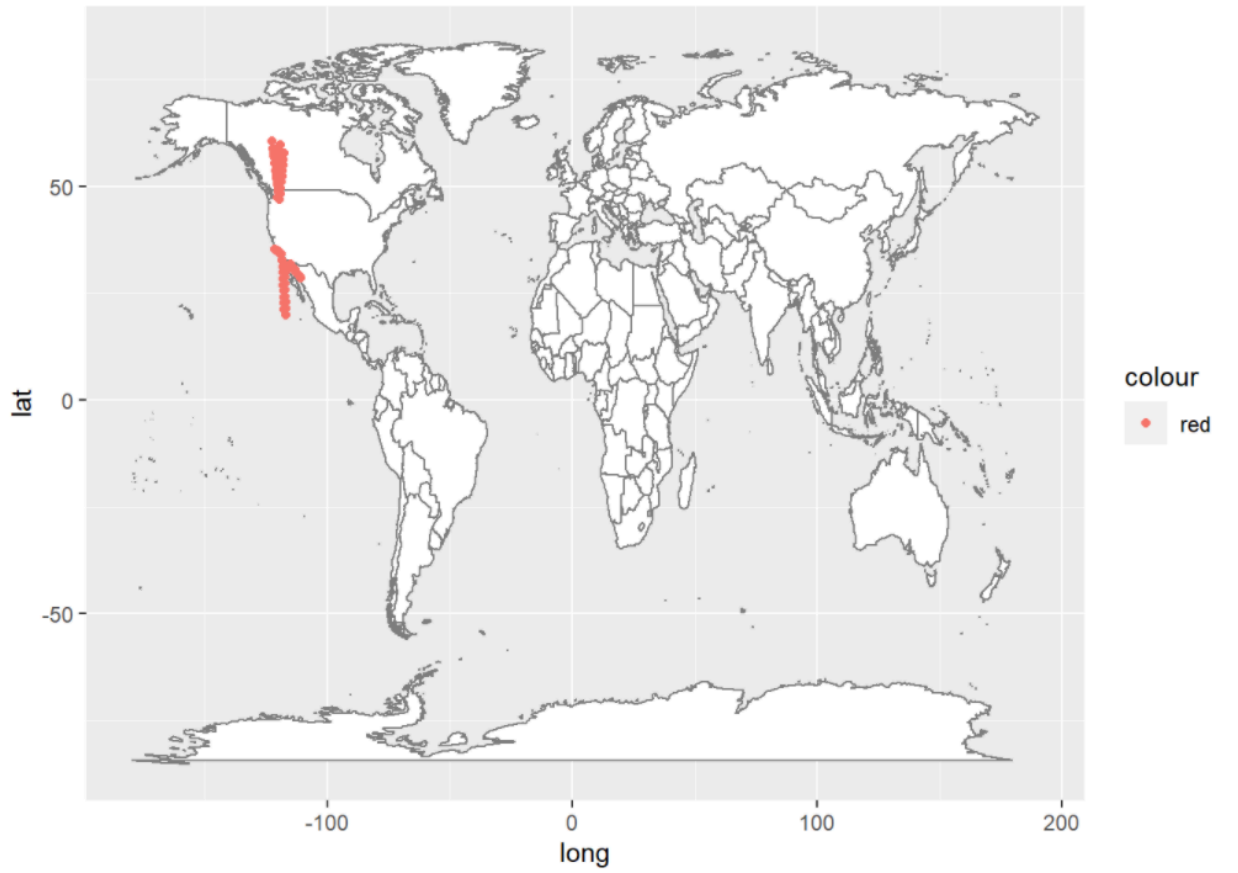
```
predBorders <- as.data.frame(cbind(latSumm$`Point Forecast`,
lonSumm$`Point Forecast`))

predBorders$Date <- rownames(lonSumm)

colnames(predBorders) <- c("Latitude", "Longitude", "Date")

predBorders$Border <- rep(NA, nrow(predBorders))

for(i in 1:nrow(predBorders)) {

    tPb <- predBorders[i,]

    predBorders[i,]$Border <- getCentroid(tPb$Latitude,
tPb$Longitude)

}

predBorders
```

Latitude <dbl>	Longitude <dbl>	Date <chr>	Border <chr>
46.71989	-119.0832	Apr 2019	jf na
35.26268	-118.3107	May 2019	jf na
47.85703	-119.4630	Jun 2019	jf na
48.30644	-119.7178	Jul 2019	jf na
35.12592	-117.7991	Aug 2019	jf na
48.33562	-119.4173	Sep 2019	jf na
36.10186	-117.6815	Oct 2019	jf na
37.26818	-115.5083	Nov 2019	na pa
37.75610	-116.0612	Dec 2019	na pa
47.88751	-119.4693	Jan 2020	jf na

Finally, plotting the predicted coordinates on a world map:

```
z <- read.csv("C:/Users/Muthug/Desktop/DATA
VISUAL/PROJECT/maps/predBorders1.csv")

world <- map_data('world')

a <- ggplot() + geom_map(data = world, map = world, aes(x=long,

y=lat, group=group, map_id=region),

fill="white",colour="#7f7f7f",size=0.5)

a <- a + geom_point(data=z, aes(x=Longitude, y=Latitude,
colour='red'))
```

Output:



Results and Discussion

Firstly, the basic plots and data visualization has been completed, through which one can gather insights about the dataset, its attributes, various trends and essential data analysis. Corrections & data cleaning and preprocessing was carried out as well.

Secondly, we classified the earthquake epicenters on the basis of their location in the various tectonic plates, in order to understand the impact, and trends of earthquakes with respect to each tectonic plate. We have classified around 3.5 lakh data points around the world based on the above-mentioned parameters.

After classification, we found out that the tectonic plates which cover most coastlines and islands, tend to have earthquakes with higher magnitude and depth, which in-turn have caused other natural disasters such as tsunami, landslides, volcanic eruptions etc. (Countries like - Indonesia, Japan etc. which are part of the Eurasian & North American plates).

We also developed and implemented an improvised ARIMA model which can be used to forecast and predict future earthquakes, with respect to latitude and longitude, and

also the magnitude and depth as well. The predictions are carried out on a monthly interval till 2022.

For testing the model, we began to predict the earthquakes which are part of the North American plate ie. Alaska & Western USA. We observed that the earthquakes if occured might have a moderate magnitude between 4.5 - 5.5 and most of the predicted epicenters lie in the western USA & Alaska coasts. We also found out that it similarly matched with the epicenters of the past quakes in the same region. We predicted around 100 upcoming earthquakes in the region. The same can be carried out to predict earthquake epicenters, their magnitude and other attributes with respect to other existing tectonic plates.

Conclusion

Predicting earthquakes using location-related information is possible, given a sufficient volume of data. Each tectonic plate has its own patterns of seismic activity, which were analyzed and used to predict future earthquakes.

References

1.Earthquake magnitude prediction in Hindukush region using machine learning techniques by K. M. Asim1 • F. Martı´nez-A´ lvarez2 • A. Basit3 • T. Iqbal1in / Published online: 8 September 2016 Springer Science+Business Media Dordrecht 2016

2.Earthquake Prediction Using Expert Systems: A Systematic Mapping Study by Rabia Tehseen, Muhammad Shoaib Farooq * and Adnan Ab published on 19 March 2020

3.Earthquake Prediction: An Overview by Hiroo Kanamori of California Institute of Technology, Pasadena, California, (2003). 72 Earthquake prediction: An overview. International Handbook of Earthquake and Engineering Seismology, 1205–1216. doi:10.1016/s0074-6142(03)80186