Earthquake\_Capstone\_Report

H. Ewton

9/28/2018

## Data Sources

Two 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.

## signif\_earthquakes clean-up

This file, signif\_earthquakes, presented several challenges. The file first needed to be reduced to only include relevant variables from the original original data set; these relevant columsn included date information, time, magnitude, and location information. Next, the data needed to be filtered to include observations that included reliable magnitude measurements. The chosen measurement scale for this analysis is the Richter scale as it has been shown to be reliable and provides substantial amounts of data; therefore, all data prior to the creation and regular use of the Richter scale (1935) was removed. Earthquakes that were documented without a measurement were also filtered and removed from the table. These processes left us with the table below:

head(signif\_earthquakes)

## year month day hour minute second focal\_depth eq\_mag\_unk country  
## 1 1935 1 4 14 41 <NA> 7 6.2 TURKEY  
## 2 1935 2 25 2 51 <NA> NA 6.7 GREECE  
## 3 1935 3 5 10 27 <NA> NA 6.0 IRAN  
## 4 1935 4 11 23 15 <NA> 14 6.3 IRAN  
## 5 1935 5 1 10 24 <NA> NA 6.1 TURKEY  
## 6 1935 7 11 18 35 <NA> 10 6.3 JAPAN  
## location\_name latitude longitude  
## 1 TURKEY 40.5 27.5  
## 2 GREECE: NEAPOLIS-ANOGNIA (CRETE) 35.8 25.0  
## 3 IRAN: ALBORZ 36.3 53.3  
## 4 IRAN: KEVSUT, ALBORZ, SARI 36.3 53.5  
## 5 TURKEY: KIGI 39.3 40.6  
## 6 JAPAN: HONSHU 35.0 138.0

## 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. The cleaned table for USGS\_df appears below:

#Print head(USGS\_df)  
head(USGS\_df)

## date time latitude longitude focal\_depth magnitude  
## 1 1965-01-02 13:44:18 19.246 145.616 131.6 6.0  
## 2 1965-01-04 11:29:49 1.863 127.352 80.0 5.8  
## 3 1965-01-05 18:05:58 -20.579 -173.972 20.0 6.2  
## 4 1965-01-08 18:49:43 -59.076 -23.557 15.0 5.8  
## 5 1965-01-09 13:32:50 11.938 126.427 15.0 5.8  
## 6 1965-01-10 13:36:32 -13.405 166.629 35.0 6.7

## Joining the data frames

To best join the data frames together without losing data, a full\_join function was used. However, for the full\_join function to work best, the data from signif\_earthquakes needed to be slighly altered to match the format found in USGS\_df. This alteration was done by assigning a zero value to time measurements, then by combining hours, minutes, and seconds into one column marked “time” in hh:mm:ss format. This zero value was assigned as some earthuakes were measured in hh:mm and some in hh:mm:ss.

Another set of columns that needed to be combined in the signif\_earthquakes data frame were the date columns (month, day, year). These were united into one column labelled “date” and put into international date format YYYY-MM-DD. These changes led to the new format of the signif\_earthquake data frame:

Once the two tables were in the correct formats, full\_join was used to consolidate the data into one data frame:

#Join tables together via date, time, latitude, and longitude  
earthquakes <- full\_join(signif\_earthquakes, USGS\_df, by = c("date", "time", "latitude", "longitude", "magnitude", "focal\_depth"))  
  
summary(earthquakes)

## date time focal\_depth magnitude   
## Length:23777 Length:23777 Min. : -1.10 Min. :3.200   
## Class :character Class1:hms 1st Qu.: 14.60 1st Qu.:5.600   
## Mode :character Class2:difftime Median : 33.00 Median :5.800   
## Mode :numeric Mean : 70.61 Mean :5.892   
## 3rd Qu.: 54.30 3rd Qu.:6.100   
## Max. :700.00 Max. :9.100   
## NA's :79   
## country location\_name latitude longitude   
## Length:23777 Length:23777 Min. :-77.080 Min. :-180.00   
## Class :character Class :character 1st Qu.:-18.404 1st Qu.: -76.11   
## Mode :character Mode :character Median : -3.406 Median : 102.42   
## Mean : 1.971 Mean : 39.29   
## 3rd Qu.: 27.312 3rd Qu.: 144.74   
## Max. : 86.005 Max. : 180.00   
##

However, one more data set was needed to complete the analysis. Stored as JSON file, this data set contained the plate bounddaries of each plate by connecting a series of coordinates. This file was uploaded and read using the geojsonio package. Once loaded, the earthquakes data set was overlaid onto the plate boundaries data set.

library(sp)  
library(geojsonio)

##   
## Attaching package: 'geojsonio'

## The following object is masked from 'package:base':  
##   
## pretty

# Import data for tectonic plate boundaries  
plate\_data <- "PB2002\_plates.json"  
plates <- geojson\_read(plate\_data, what = "sp")   
  
  
# convert list of earthquake points into a SpatialPointsDataFrame  
coordinates(earthquakes) <- ~ longitude + latitude  
  
# convert earthquakes to use the same coordinate system as plates (needed for overlay)  
proj4string(earthquakes) <- proj4string(plates)  
  
# Create overlay  
earthquakes\_plates <- over(earthquakes, plates)

The final step in tidying the data was to combine the columns from the two resulting tables and save the new result as a data frame. This gives us the following result:

# Attach the resulting columns that we got from over to the rows of eqs  
earthquakes$LAYER <- earthquakes\_plates$LAYER  
earthquakes$Code <- earthquakes\_plates$Code  
earthquakes$PlateName <- earthquakes\_plates$PlateName  
  
earthquakes <- as.data.frame(earthquakes)  
  
  
head(earthquakes)

## date time focal\_depth magnitude country  
## 1 1935-1-4 14:41:00 7 6.2 TURKEY  
## 2 1935-2-25 02:51:00 NA 6.7 GREECE  
## 3 1935-3-5 10:27:00 NA 6.0 IRAN  
## 4 1935-4-11 23:15:00 14 6.3 IRAN  
## 5 1935-5-1 10:24:00 NA 6.1 TURKEY  
## 6 1935-7-11 18:35:00 10 6.3 JAPAN  
## location\_name latitude longitude LAYER Code  
## 1 TURKEY 40.5 27.5 plate AT  
## 2 GREECE: NEAPOLIS-ANOGNIA (CRETE) 35.8 25.0 plate AS  
## 3 IRAN: ALBORZ 36.3 53.3 plate EU  
## 4 IRAN: KEVSUT, ALBORZ, SARI 36.3 53.5 plate EU  
## 5 TURKEY: KIGI 39.3 40.6 plate AT  
## 6 JAPAN: HONSHU 35.0 138.0 plate AM  
## PlateName  
## 1 Anatolia  
## 2 Aegean Sea  
## 3 Eurasia  
## 4 Eurasia  
## 5 Anatolia  
## 6 Amur