Earthquakes

William Lovejoy

1/14/2022

Starting off with importing our data from the USGS earthquake database. I prefer using the file.choose() method for importing data because it saves on code and ensures I don't need to worry too much about my working directory. It also allows those that aren't familiar with coding practices to load in relevant datasets on their own. I hashed out the line here and inserted the full file path to make sure Markdown would knit correctly.

As always, check and make sure your data loads in correctly and get a feel for what kind of information you'll be working with.

```
#Data is earthquake data from the USGS ANNS for 1/1/2021 - 1/1/2022
#https://earthquake.usgs.gov/data/comcat/data-eventterms.php Contains informa
tion about what each df variable means

library(tidyverse)

library(RColorBrewer)


#df <- read.csv(file.choose())
df <- read.csv("C:/Users/[redacted]/Downloads/query.csv")
head(df)

##                       time latitude longitude     depth  mag magType nst g
ap
## 1 2021-12-31T22:56:50.593Z 31.64025 -104.4040  6.131567 2.70      ml  17
64
## 2 2021-12-31T21:24:28.100Z 40.47750 -124.3230 19.550000 2.59      md  24 1
24
## 3 2021-12-31T20:57:46.030Z 40.47783 -124.3288 19.050000 3.81      mw  30 1
21
## 4 2021-12-31T19:27:58.900Z 34.03467 -117.2080 14.430000 3.38      ml 162
17
## 5 2021-12-31T19:08:31.870Z 40.47683 -124.3200 19.350000 2.54      md  23 1
18
## 6 2021-12-31T13:52:50.485Z 31.21720 -103.3186  5.206079 2.60      ml  17
50
##         dmin  rms net         id                  updated
## 1 0.08725266 0.20  tx tx2021zqun 2022-01-07T12:24:44.019Z
## 2 0.01122000 0.08  nc nc73671466 2022-01-04T02:52:11.450Z
## 3 0.00697800 0.12  nc nc73671446 2022-01-12T19:54:24.588Z
## 4 0.07939000 0.20  ci ci39900159 2022-01-13T05:11:02.926Z
## 5 0.01340000 0.10  nc nc73670851 2022-01-04T02:38:11.364Z
```

```
## 6 0.05190920 0.20  tx tx2021zqcp 2022-01-07T12:29:30.127Z
##                                 place       type horizontalError depthError
## 1 59 km S of Whites City, New Mexico earthquake        1.064476   1.135905
## 2           12km SSW of Ferndale, CA earthquake        0.290000   0.290000
## 3           12km SSW of Ferndale, CA earthquake        0.410000   0.230000
## 4             3km SW of Redlands, CA earthquake        0.110000   0.230000
## 5           12km SSW of Ferndale, CA earthquake        0.350000   0.290000
## 6         24 km W of Coyanosa, Texas earthquake        1.158919   1.507164
##   magError magNst   status locationSource magSource
## 1    0.100     12 reviewed             tx        tx
## 2    0.128     20 reviewed             nc        nc
## 3       NA      3 reviewed             nc        nc
## 4    0.141    237 reviewed             ci        ci
## 5    0.082     19 reviewed             nc        nc
## 6    0.200      8 reviewed             tx        tx
```

Finally, look at the structure of the data. This is important for any work we do because it lets us make sure we're performing the right operations on the right variables in the dataframe.

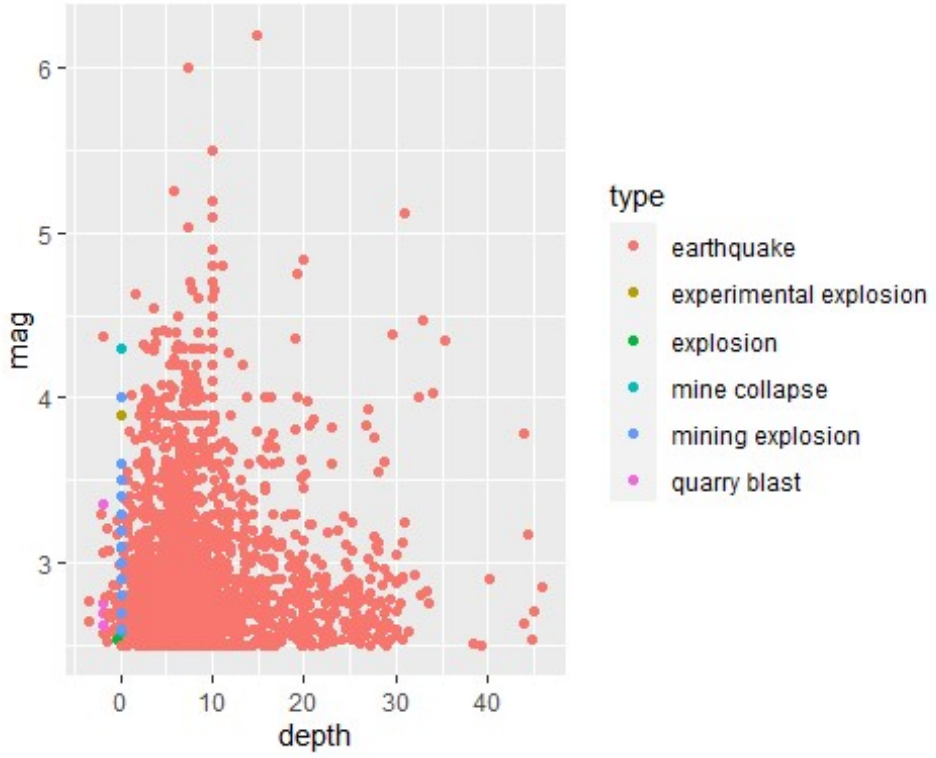
```
str(df)
```

```
## 'data.frame':    3327 obs. of  22 variables:
##  $ time           : chr  "2021-12-31T22:56:50.593Z" "2021-12-31T21:24:28.1
00Z" "2021-12-31T20:57:46.030Z" "2021-12-31T19:27:58.900Z" ...
##  $ latitude       : num  31.6 40.5 40.5 34 40.5 ...
##  $ longitude      : num  -104 -124 -124 -117 -124 ...
##  $ depth          : num  6.13 19.55 19.05 14.43 19.35 ...
##  $ mag            : num  2.7 2.59 3.81 3.38 2.54 2.6 2.55 2.62 2.95 2.6 ..
.
##  $ magType        : chr  "ml" "md" "mw" "ml" ...
##  $ nst            : int  17 24 30 162 23 17 40 78 59 24 ...
##  $ gap            : num  64 124 121 17 118 50 67 25 52 66 ...
##  $ dmin           : num  0.08725 0.01122 0.00698 0.07939 0.0134 ...
##  $ rms            : num  0.2 0.08 0.12 0.2 0.1 0.2 0.11 0.23 0.2 0.2 ...
##  $ net            : chr  "tx" "nc" "nc" "ci" ...
##  $ id             : chr  "tx2021zqun" "nc73671466" "nc73671446" "ci3990015
9" ...
##  $ updated        : chr  "2022-01-07T12:24:44.019Z" "2022-01-04T02:52:11.4
50Z" "2022-01-12T19:54:24.588Z" "2022-01-13T05:11:02.926Z" ...
##  $ place          : chr  "59 km S of Whites City, New Mexico" "12km SSW of
Ferndale, CA" "12km SSW of Ferndale, CA" "3km SW of Redlands, CA" ...
##  $ type           : chr  "earthquake" "earthquake" "earthquake" "earthquak
e" ...
##  $ horizontalError: num  1.06 0.29 0.41 0.11 0.35 ...
##  $ depthError     : num  1.14 0.29 0.23 0.23 0.29 ...
##  $ magError       : num  0.1 0.128 NA 0.141 0.082 0.2 0.257 0.133 0.149 0.
2 ...
##  $ magNst         : int  12 20 3 237 19 8 44 26 157 15 ...
##  $ status         : chr  "reviewed" "reviewed" "reviewed" "reviewed" ...
```

```
##  $ locationSource : chr  "tx" "nc" "nc" "ci" ...
##  $ magSource      : chr  "tx" "nc" "nc" "ci" ...
```

I like to start off by making a bunch of quick graphs of the data. It's a great way to find some of the more easily accessible trends and can give you a great opening to exploring the data easier. This graph looks at the relationship between an earthquakes depth and it's magnitude. By coloring the plot points by the type of event as well, we can get a rough idea of how much of our data is in each event.

```
e <-  ggplot(data = df)

e + geom_point(mapping = aes(x = depth, y = mag, color = type))
```



Already we can see that almost all the data in this set is an earthquake. So let's look at more exact numbers

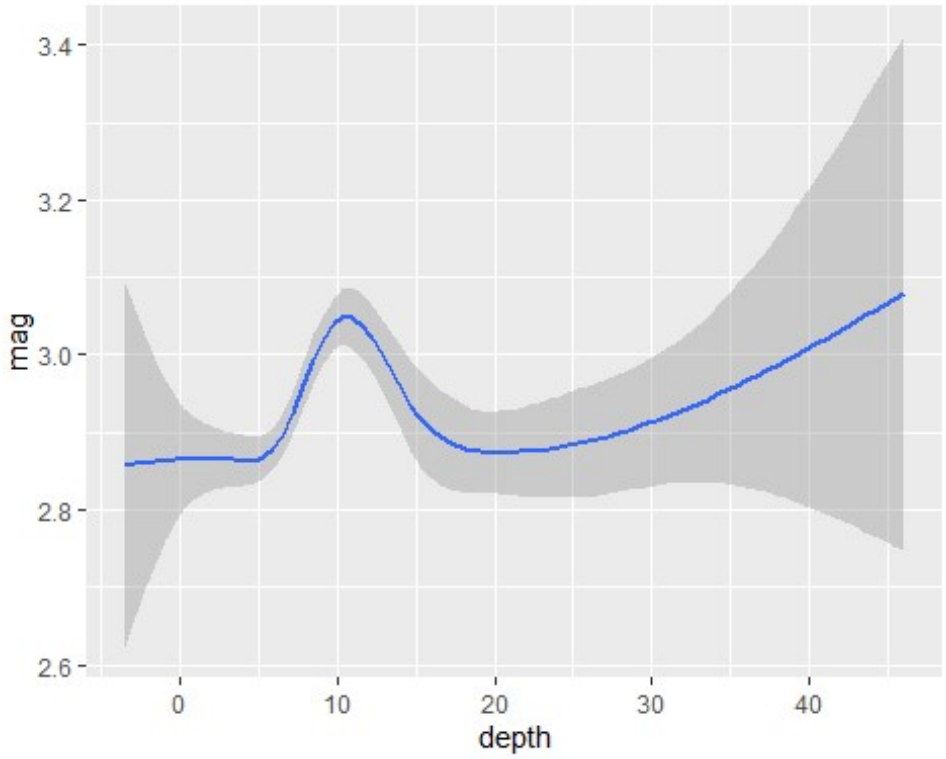
```
table(df$type)

##
##             earthquake experimental explosion              explosion
##                   3117                      3                      1
##          mine collapse       mining explosion           quarry blast
##                      3                    198                      5
```

As we thought. Of over 3300 data points, 3117 of them were from actual quakes. So we'll be narrowing down our data by cutting out the excess. This also lets us name our data frame something more relevant which makes the code more readable in the future.

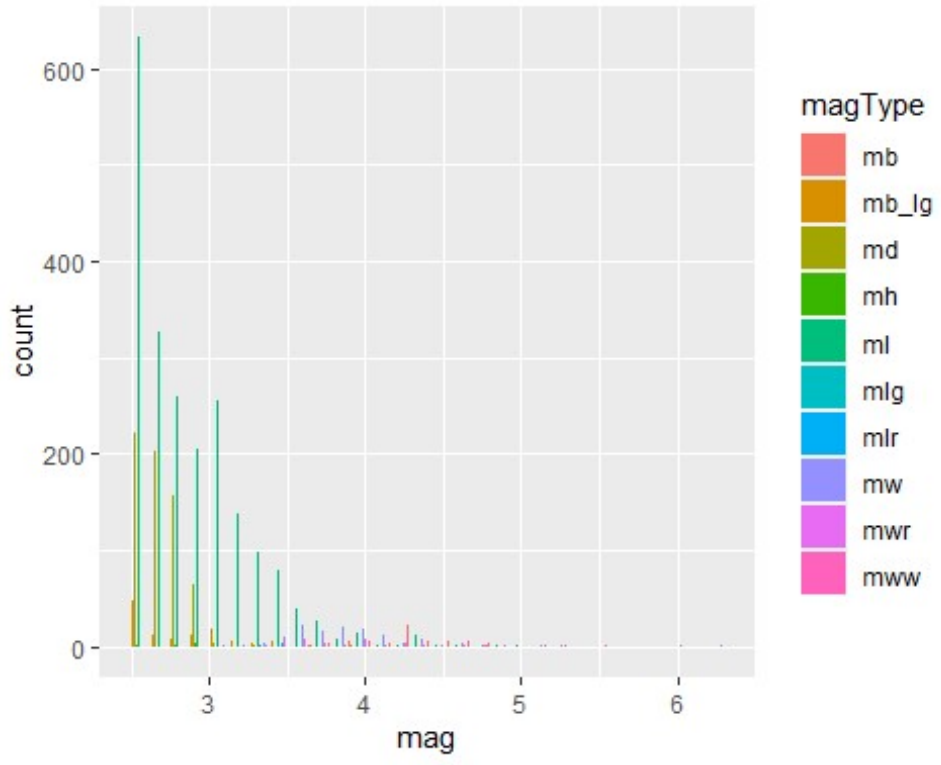
```
earthquakes <-  subset(df, type == "earthquake")
```

Now we'll keep graphing things that might have a relationship.

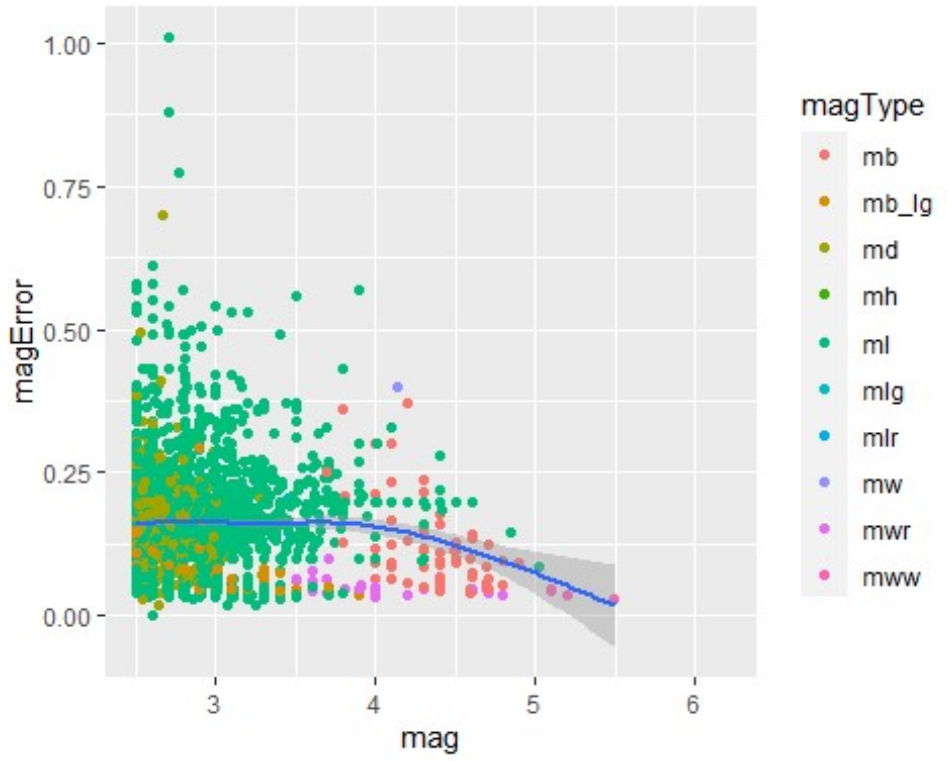
```
g <- ggplot(earthquakes)

g + geom_smooth(mapping = aes(x = depth, y = mag))

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
g + geom_histogram(mapping = aes(x = mag, fill = magType), position = "dodge"
)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
g + geom_point(mapping = aes(x = mag, y = magError, color = magType), positio
n = "jitter") + geom_smooth(mapping = aes(x = mag, y = magError))

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 153 rows containing non-finite values (stat_smooth).

## Warning: Removed 153 rows containing missing values (geom_point).
```



We can immediately tell through our histogram that most earthquakes tend to be in the 1 - 3.5 magnitude range.

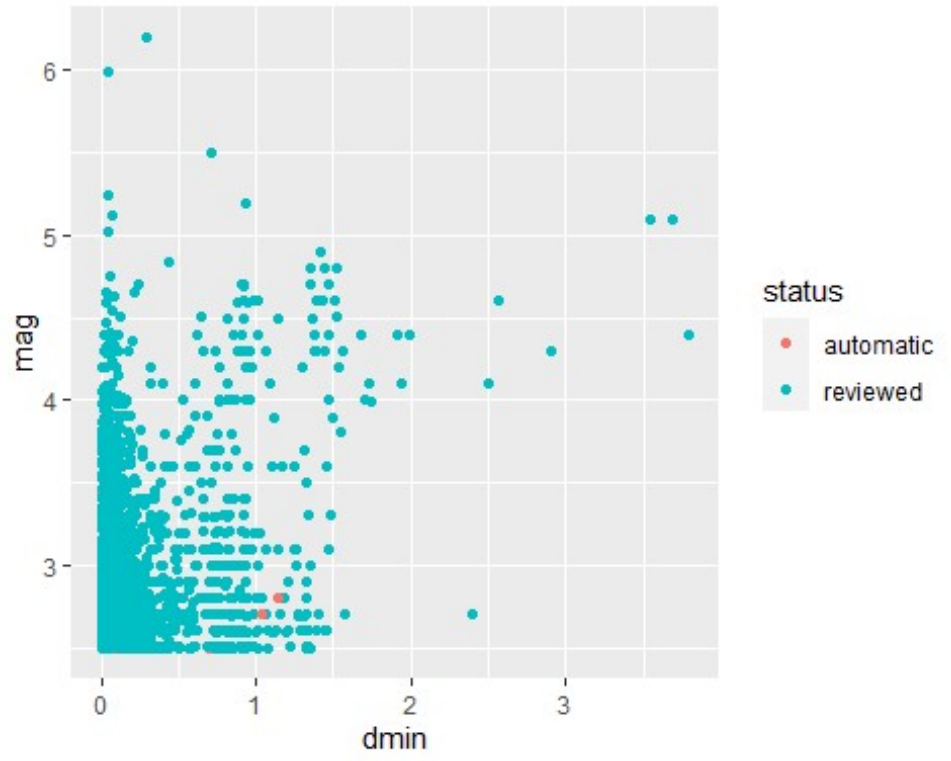
When comparing magnitude and magnitude error scores, it seems that larger earthquakes tend to have less error in their magnitude assessment, and tend to have more “mb”, “mwr”, and “mww” waveforms. This could mean that earthquakes with more magnitude are assessed more accurately because they pose a greater risk, however the standard error shown on the geom_smooth() line on the graph gets wider as the magnitude gets bigger. There’s no certainty here other than saying that “larger magnitudes tend to have more accurate magnitude assessments” and “large earthquakes tend to have certain waveforms”.

If it’s true that larger earthquakes tend to have one of a a couple waveforms, then assessing the movement as an earthquake starts can be used to estimate the potential magnitude and prepare relief supplies earlier. Unfortunately, our data doesn’t have enough of these waveforms or larger magnitude quakes to make that kind of assessment. However, this is good data to pass along to researchers in this field to experiment with.

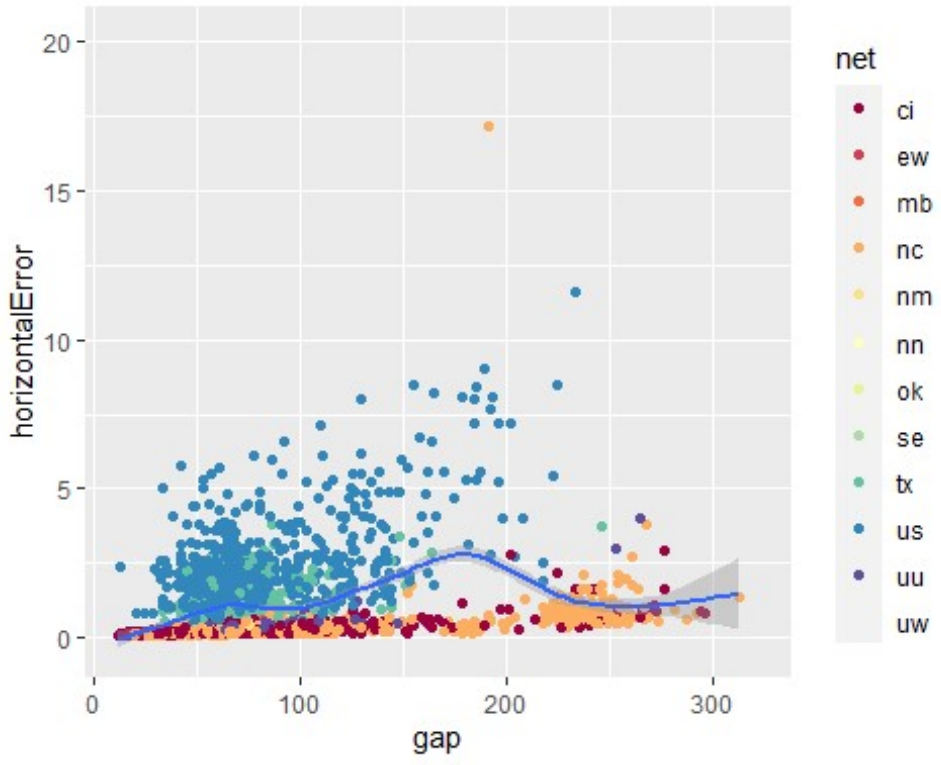
Next up we’ll compare dmin and magnitude as well as gap and horizontal error. According to the USGS: dmin is how far the epicenter was from the nearest station, with 1 degree ~= 111.2 Km, and gap is the largest azithmuthally adjacent stations (in degree). It’s used to calculate horizontal earthquake position and lower numbers are better. Gap’s >180 have large location and depth uncertainties.

```
g + geom_point(mapping = aes(x = dmin, y = mag, color = status), position = "
jitter")
```

```
## Warning: Removed 12 rows containing missing values (geom_point).
```

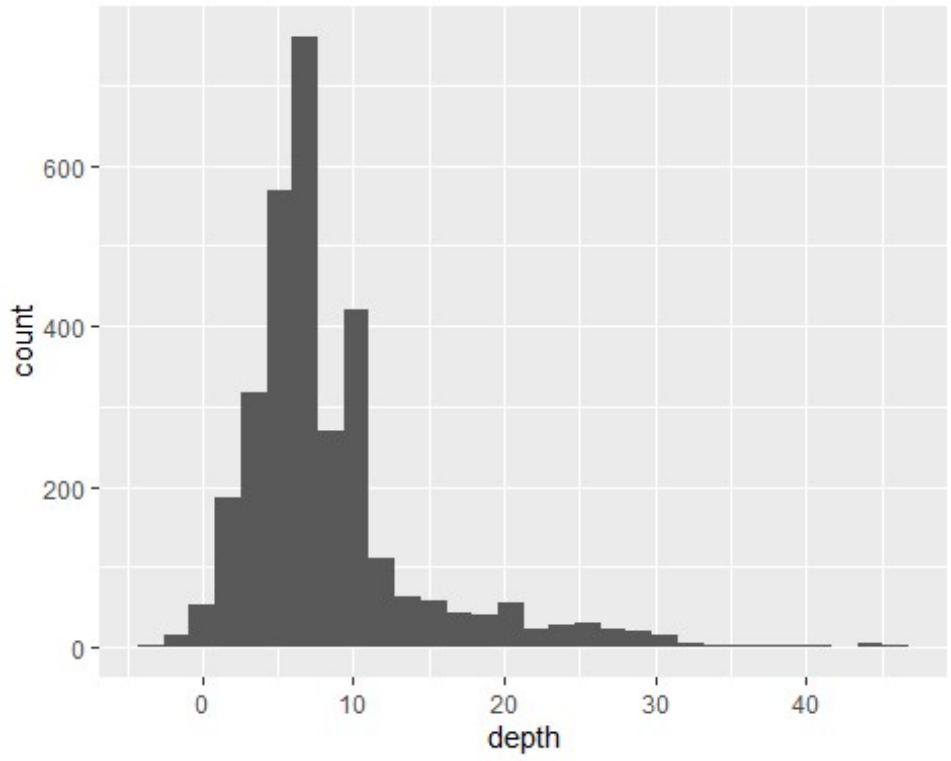



```
g + geom_point(mapping = aes(x = gap, y = horizontalError, color = net)) + ge
om_smooth(mapping = aes(x = gap, y = horizontalError), se = TRUE) +
  scale_color_brewer(palette = "Spectral")

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 320 rows containing non-finite values (stat_smooth).

## Warning in RColorBrewer::brewer.pal(n, pal): n too large, allowed maximum
for palette Spectral is 11
## Returning the palette you asked for with that many colors

## Warning: Removed 347 rows containing missing values (geom_point).
```



3 notable things about these graphs: 1) Dmin score tend to be low, meaning that most earthquakes happen fairly close to earthquake stations. This implies that there is a good spread of stations allowing for good coverage of seismic events. 2) There's a noticeable uptick in horizontal error measurements for gap scores between 150 - 200. This might be something worth digging into a bit more. 3) There is a very visible difference in horizontal error reporting between networks. Networks such as: ci, ew, mb, nc, nm, mn, and ok all seem to have consistently low horizontal error score. While these scores tend to grow at larger gaps, it's not as bad as the scores reported by other networks. The other networks (namely se, tx, us, uu, and uw) all seem to report horizontal error scores above the formulated line. There might be an issue with these networks.

```
g + geom_histogram(mapping = aes(x = depth))

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Since there are differences between the networks and their horizontal error reporting, we should try to find out why. Unfortunately, the ANNS system doesn't have all the data I need to do so. So we're left with lingering questions. Since my own knowledge about seismic activities is limited, I can't hazard more than surface level guesses about the cause. Maybe the monitoring stations are at different elevations? Could underground artifacts (such as pockets or substrate differences) be causing variation in the seismic waves recorded?