Earthquake Data

Description of data: Earthquake source parameters produced by contributing siesmic networks

- Data Source: ANSS Comprehensive Catalog hosted by the Northern California Earthquake Data Center
- Search URL: http://quake.geo.berkeley.edu/anss/catalog-search.html

Data extract parameters

- years: 1960 - 2014
- Minimum magnitude: 2
- Geographic Region: 120W to 130W,30N to 40N, primarily Northern California
- Events: All events excluding Acoustic Noise and Chemical events

Data

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    8907 obs. of  15 variables:
##  $ time     : chr  "2014-09-29T07:17:01.080Z" "2014-09-29T03:46:40.790Z" "2014-09-28T20:45:13.260Z" "2014-09-28T19:49:24.740Z" ...
##  $ latitude : num  38.2 36.6 36.6 36.6 38.2 ...
##  $ longitude: num  -122 -121 -121 -121 -122 ...
##  $ depth    : num  11.71 5.97 7.75 4.05 7.92 ...
##  $ mag      : num  2.56 2.71 4.43 2.61 2.29 2.3 2.49 2.42 2.03 2.02 ...
##  $ magType  : chr  "md" "md" "mw" "md" ...
##  $ nst      : int  69 55 100 48 49 44 46 14 53 13 ...
##  $ gap      : num  131 102 69 58 75 48 36 118 48 52 ...
##  $ dmin     : num  0.0568 0.0522 0.051 0.0566 0.0228 ...
##  $ rms      : num  0.13 0.11 0.13 0.12 0.11 0.04 0.09 0.11 0.05 0.08 ...
##  $ net      : chr  "nc" "nc" "nc" "nc" ...
##  $ id       : chr  "nc72316546" "nc72316316" "nc72316031" "nc72315991" ...
##  $ updated  : chr  "2014-09-29T08:19:27.583Z" "2014-09-29T06:17:07.438Z" "2014-09-29T07:53:10.424Z" "2014-09-29T00:28:08.997Z" ...
##  $ place    : chr  "3km WNW of American Canyon, California" "40km SSW of South Dos Palos, California" "41km SSW of South Dos Palos, California"
"41km SSW of South Dos Palos, California" ...
##  $ type     : chr  "earthquake" "earthquake" "earthquake" "earthquake" ...
```

```
## NULL
```

Descriptions of fields

Attributes

- time: Date - Time of event
- latitude: geographic latitude
- longitude: geographic longitude
- nst: The total number of number of seismic stations which reported P- and S-arrival times for this earthquake.
- gap: The largest azimuthal gap between azimuthally adjacent stations (in degrees). In general, the smaller this number, the more reliable is the calculated horizontal position of the earthquake typically [0 - 180]
- dmin: Horizontal distance from the epicenter to the nearest station (in degrees). 1 degree is approximately 111.2 kilometers. In general, the smaller this number, the more reliable is the calculated depth of the earthquake. Typically: [0.4, 7.1]
- magType: Scale used to measure magnitude, commonly:local magnitude (ML),surface-wave magnitude (Ms), body-wave magnitude (Mb), moment magnitude (Mw). All magnitude scales should yield approximately the same value for any given earthquake
- net: The ID of a data contributor. Identifies the network considered to be the preferred source of information for this event. Typical values: [ak, at, ci, hv, ld, …, pr, pt, se, us, uu, uw]
- id: A comma-separated list of event ids that are associated to an event
- updated: Time when the event was most recently updated.
- place: Textual description of named geographic region near to the event.

Variables

- depth: Depth of the event in kilometers, typically [0, 1000]
- mag: The magnitude for the event, typically [-1.0, 10]
- rms: The root-mean-square (RMS) travel time residual, in sec, using all weights. This parameter provides a measure of the fit of the observed arrival times to the predicted arrival times for this location. Smaller numbers reflect a better fit of the data. Typically: [0.13, 1.39]

```
countMissing <- function(input){
   d <- input[is.na(input) != FALSE]
   return(length(d))
}
countdfMissing <- function(input){
  z<- sapply(input, function(x) countMissing(x))
  return(z)
}

nst.miss <- df %>%select(nst,net)%>% filter(is.na(nst))%>%group_by(net)%>%summarise(count = n())
gap.miss <- df %>%select(gap,nst)%>% filter(is.na(gap))%>%group_by(nst)%>%summarise(count = n())
dmin.miss <- df %>%select(dmin,nst, net)%>% filter(is.na(dmin))%>%group_by(net)%>%summarise(count = n())
rms.miss <- df %>%select(rms, dmin)%>% filter(is.na(rms))%>%group_by(dmin)%>%summarise(count = n())
```

data summary and missing values

- Number of observations: 8907
- Number of complete rows: 587
- Fields with missing data:

```
countdfMissing(df)
```

```
##      time  latitude longitude     depth       mag   magType       nst       gap      dmin       rms       net        id
##         0         0         0         0         0         0      5413      4612      7084      6812         0         0
##   updated     place      type
##         0         0         0
```

There is a substantial amount of missing data, but the missing values may not impact some types of analysis. Missing values within the data are mostly in the descriptive attributes, and are all related to the network or system used for collecting the data, and may reflect limited reporting from the system. There are no missing values in the magnitude, depth, time, location, or type fields, so analysis involving the events and not concerned with the method of data collection would not be impacted by these missing values.

Missing values in nst are related to the network, most of the missing gap values are related to the missing nst value:

```
# nst missing
nst.miss
```

```
## Source: local data frame [5 x 2]
##
##     net count
## 1 atlas     7
## 2    ci     2
## 3    nc  1233
## 4   pde  4169
## 5    us     2
```

```
# gap missing
head(arrange(gap.miss, desc(count)))
```

```
## Source: local data frame [6 x 2]
##
##   nst count
## 1  NA  4177
## 2  12    31
## 3  14    30
## 4  15    30
## 5  10    28
## 6  11    28
```

Almost all of the missing dmin values are coming a particular network: pde. The missing data in rms is directly related to the missing data in dmin, which follows because rms is a measure of fitness of a model for dmin

```
# dmin missing
head(arrange(dmin.miss, desc(count)))
```

```
## Source: local data frame [4 x 2]
##
##          net count
## 1        pde  7064
## 2 centennial    12
## 3      atlas     7
## 4         us     1
```

```
#rms missing
rms.miss
```

```
## Source: local data frame [1 x 2]
##
##   dmin count
## 1   NA  6812
```

scope of this analysis

This analysis is primarily interested in the data related seismic events and their locations, and will select a subset of the data which excludes the missing values to evaluate.

Data transformations used

- city or place names were extracted from the place column
- latitude and longitude were rounded to the nearest whole value
- date and time were separated, only month and year included in the analysis
- magnitudes were grouped in 10 intervals from in = 2 to max = 6.9

```
pat <- "([A-Za-z ]* [A-Za-z0-9]*[, ]*[CaA-Za-z]*$)"
grups <- seq(min(df$mag), max(df$mag),.5)

df.quake <- select(df, which(countdfMissing(df)== 0))
df.quake <- df.quake %>% separate(time, c("year", "month", "day","D"), sep = c("[-T]")) %>%
  extract(place, "place2", pat) %>% mutate(place2 = gsub("^.*?of ","",place2)) %>%
  mutate(lat = round(latitude,digits =1), long = round(longitude, digits = 1), year = as.numeric(year), month = as.numeric(month)) %>%
```

```
  select(year, month, lat, long, depth, mag, place2, type, magType) %>%
  mutate(mag.group = grups[findInterval(mag, grups)], lat.rnd = round(lat,0), long.rnd = round(long,0))%>% arrange(-mag,year)

s4 <- df.quake %>% group_by(type) %>% summarise( avgmag = mean(mag), avgdepth = mean(depth), numevents = n()) %>% arrange(-avgmag)

s1 <- df.quake %>% group_by(place2) %>% summarise( avgmag = mean(mag), avgdepth = mean(depth), numevents = n()) %>% arrange(-avgmag)

s2 <- df.quake %>% group_by(month) %>% summarise( avgmag = mean(mag), avgdepth = mean(depth), numevents = n()) %>% arrange(-numevents)

s3 <- df.quake %>% group_by(year) %>% summarise( avgmag = mean(mag), avgdepth = mean(depth), numevents = n()) %>% arrange(-avgmag)

s7 <- df.quake %>% group_by(magType) %>% summarise( avgmag = mean(mag), avgdepth = mean(depth), numevents = n()) %>%
arrange(-numevents)
```

Data used in this analysis:

```
head(df.quake)
```

```
## Source: local data frame [6 x 12]
##
##   year month  lat   long depth mag                          place2       type magType mag.group lat.rnd long.rnd
## 1 1989    10 37.1 -121.8  11.4 6.9               Northern California earthquake       ms       6.5      37     -122
## 2 1989    10 37.0 -121.9  18.0 6.9               Northern California earthquake              6.5      37     -122
## 3 2003    12 35.6 -121.1  16.0 6.5                Central California earthquake      mw       6.5      36     -121
## 4 1983     5 36.2 -120.3  10.0 6.3               Central California earthquake       ms       6.0      36     -120
## 5 1983     5 36.2 -120.3  10.0 6.3               Central California earthquake             6.0      36     -120
## 6 1984     4 37.3 -121.7   8.8 6.1 San Francisco Bay area, California earthquake       ms       6.0      37     -122
```

Summary Statistics

```
(summary(df.quake))
```

```
##       year          month           lat            long          depth           mag          place2
##  Min.   :1966   Min.   : 1.00   Min.   :31.2   Min.   :-128   Min.   : 0.00   Min.   :2.00   Length:8907
##  1st Qu.:1989   1st Qu.: 4.00   1st Qu.:36.6   1st Qu.:-123   1st Qu.: 3.70   1st Qu.:2.50   Class :character
##  Median :2002   Median : 7.00   Median :37.2   Median :-122   Median : 6.00   Median :2.90   Mode  :character
##  Mean   :2000   Mean   : 6.54   Mean   :37.5   Mean   :-122   Mean   : 6.51   Mean   :2.94
##  3rd Qu.:2011   3rd Qu.: 9.00   3rd Qu.:38.8   3rd Qu.:-121   3rd Qu.: 8.70   3rd Qu.:3.20
##  Max.   :2014   Max.   :12.00   Max.   :40.0   Max.   :-120   Max.   :62.30   Max.   :6.90
##     type           magType           mag.group       lat.rnd        long.rnd
##  Length:8907        Length:8907        Min.   :2.00   Min.   :31.0   Min.   :-128
##  Class :character   Class :character   1st Qu.:2.50   1st Qu.:37.0   1st Qu.:-123
##  Mode  :character   Mode  :character   Median :2.50   Median :37.0   Median :-122
##                                        Mean   :2.75   Mean   :37.5   Mean   :-122
##                                        3rd Qu.:3.00   3rd Qu.:39.0   3rd Qu.:-121
##                                        Max.   :6.50   Max.   :40.0   Max.   :-120
```

```
# Summary of locations top 20 average magnitude:
head(s1,20)
```

```
## Source: local data frame [20 x 4]
##
##                                 place2 avgmag avgdepth numevents
## 1                North Pacific Ocean  4.275   10.000         4
## 2                         California  3.600   16.250         4
## 3                   Ferndale, California  3.400    4.300         1
## 4     Santa Barbara Channel, California  3.400    3.250         6
## 5           Channel Islands, California  3.300    7.000         1
## 6     Channel Islands region, California  3.230    6.930        27
## 7           offshore Central California  3.216    6.195       146
## 8                  Bolinas, California  3.200    7.600         2
## 9                   Central California  3.152    6.706      3057
## 10                  Southern California  3.150    8.050         2
## 11                  Northern California  3.104    5.882      2847
## 12 Vandenberg Air Force Base, California  2.980    6.965         4
## 13            San Pablo Bay, California  2.925    9.238         8
## 14               Saratoga, California  2.900   11.800         1
## 15         offshore Northern California  2.897    7.844       105
## 16                  Solvang, California  2.880    0.030         1
## 17    San Francisco Bay area, California  2.852    8.558       874
## 18                  Shandon, California  2.835    8.501        10
## 19                 Millbrae, California  2.800    3.900         1
## 20          South Dos Palos, California  2.794    5.971         8
```

```
# Summary of locations top 20 number of recorded events:
head(arrange(s1, -numevents),20)
```

```
## Source: local data frame [20 x 4]
##
##                                 place2 avgmag avgdepth numevents
## 1                   Central California  3.152    6.706      3057
## 2                  Northern California  3.104    5.882      2847
## 3  San Francisco Bay area, California  2.852    8.558       874
## 4             The Geysers, California  2.283    2.560       311
## 5                    Cobb, California  2.299    2.414       240
## 6           offshore Central California  3.216    6.195       146
## 7          offshore Northern California  2.897    7.844       105
## 8                 Soledad, California  2.345    6.621       104
## 9               Ridgemark, California  2.436    6.346       102
```

```
## 10            San Simeon, California  2.437    6.545       76
## 11              Coalinga, California  2.426    9.082       62
## 12             King City, California  2.299    8.553       41
## 13      San Juan Bautista, California  2.505    5.887       38
## 14                 Napa, California  2.464    8.346       38
## 15            Greenfield, California  2.214    6.204       34
## 16               Gilroy, California  2.633    5.658       31
## 17 Channel Islands region, California  3.230    6.930       27
## 18         East Foothills, California  2.398    7.205       24
## 19        American Canyon, California  2.595    9.701       23
## 20           Laytonville, California  2.290    6.663       23
```

```
# Summary of event type
s4
```

```
## Source: local data frame [3 x 4]
##
##          type avgmag avgdepth numevents
## 1   earthquake  2.944    6.539      8864
## 2      quarry  2.155    0.000        38
## 3 quarry_blast  2.116    0.134         5
```

```
#Summary of average number of events by month
s2
```

```
## Source: local data frame [12 x 4]
##
##    month avgmag avgdepth numevents
## 1      8  2.923    6.844       912
## 2      5  3.013    6.963       911
## 3     10  3.051    7.133       771
## 4     12  3.038    5.847       753
## 5      1  2.980    6.365       737
## 6      4  2.897    6.151       726
## 7      3  2.864    6.334       724
## 8      9  2.954    6.615       719
## 9      6  2.834    6.620       698
## 10     7  2.894    6.409       661
## 11    11  2.913    6.434       658
## 12     2  2.868    6.105       637
```

```
#Summary of top 10 average magnitude by year
head(s3, 10)
```

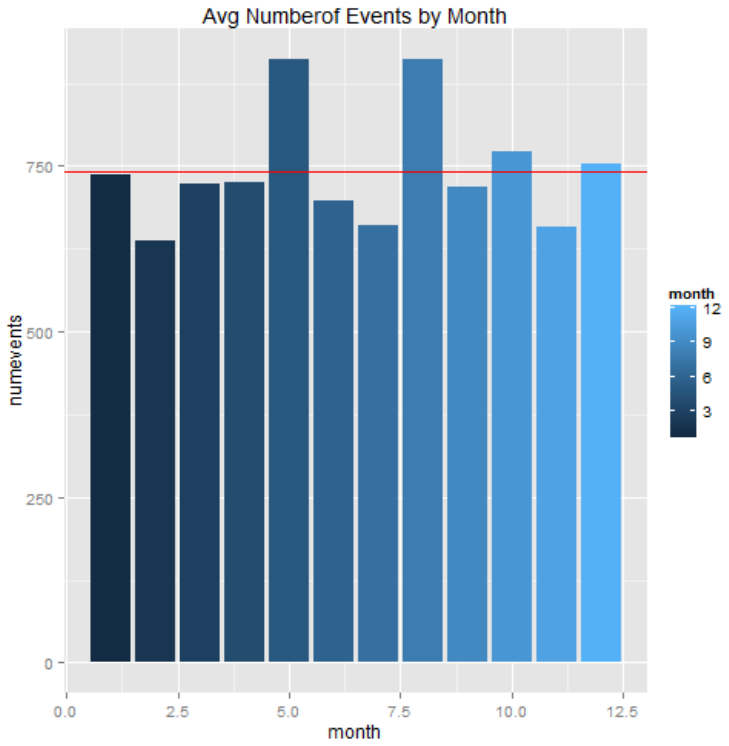
```
## Source: local data frame [10 x 4]
##
##    year avgmag avgdepth numevents
## 1  1966  5.800    1.600         2
## 2  1969  5.750    8.600         2
## 3  1975  3.791    7.558        89
## 4  1973  3.758    7.583        12
## 5  1974  3.610    7.600        30
## 6  1980  3.494    7.236        89
## 7  1983  3.445    8.218       380
## 8  1976  3.426    7.214        42
## 9  1977  3.422    7.282        78
## 10 1984  3.411    6.203       148
```

```
# Summary of magnitude and depth by magType
s7
```

```
## Source: local data frame [15 x 4]
##
##    magType avgmag avgdepth numevents
## 1       ml  3.121    6.522      4077
## 2       md  2.701    6.456      2397
## 3       Md  2.265    5.888      1476
## 4      mwr  3.734    6.882       347
## 5       mb  4.300    8.175       248
## 6       dr  3.185    8.531       146
## 7       Ml  2.976    6.351        71
## 8       mw  4.001    7.805        66
## 9       Mw  3.592    6.354        39
## 10      ms  5.600    8.957        14
## 11       H  2.712    6.612         8
## 12           5.943    9.586         7
## 13     mwc  5.100    9.057         7
## 14      uk  5.767    6.067         3
## 15     mww  5.400    9.400         1
```

Earthquake Weather Number of events by month. An anova model indicates it is unlikely different months have different means, and consequently the term 'earthquake weather' is shown to be meaningless.

```
#  number of events by month
c <- ggplot(data = s2, aes(x = month, y = numevents, fill = month))
c + geom_bar(stat = "identity") + geom_hline(aes(yintercept = mean(s2$numevents)),colour = "red")+ ggtitle("Avg Numberof Events by Month")
```

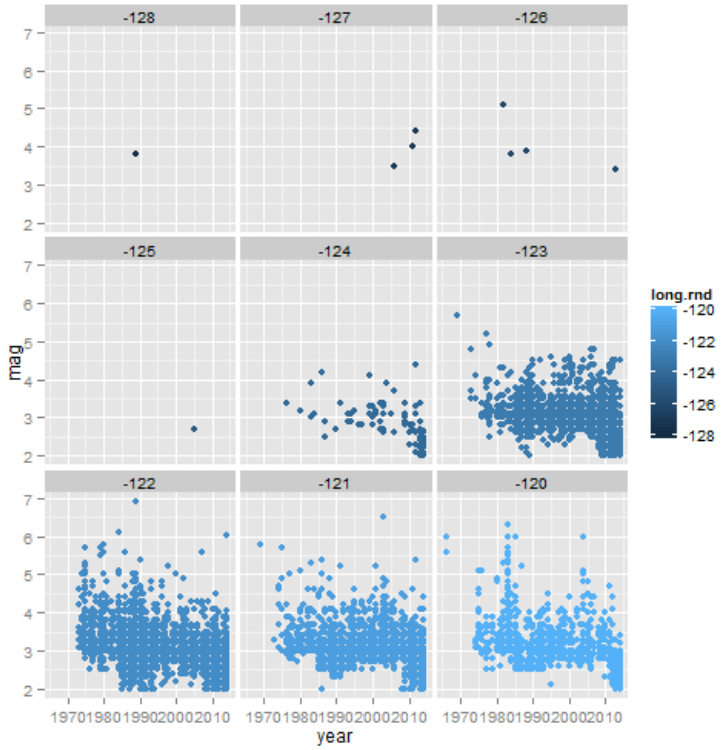



```
aov.out <- aov(numevents ~ month, data = s2)
(summary.lm(aov.out))
```

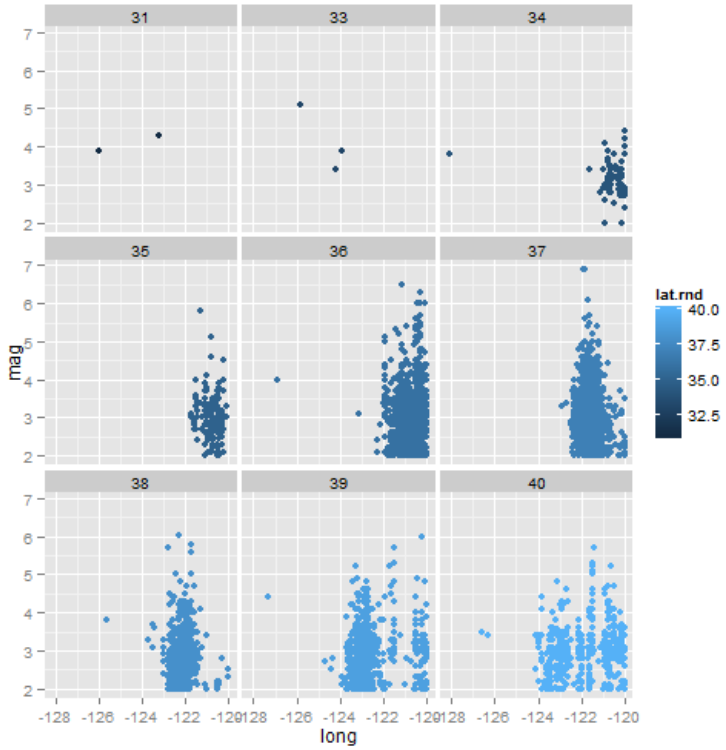
```
##
## Call:
## aov(formula = numevents ~ month, data = s2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -95.4  -53.0  -10.7   10.3  172.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   728.05      56.89   12.80  1.6e-07 ***
## month           2.19       7.73    0.28     0.78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 92.4 on 10 degrees of freedom
## Multiple R-squared:  0.00793,    Adjusted R-squared:  -0.0913
## F-statistic: 0.0799 on 1 and 10 DF,  p-value: 0.783
```

*** magnitude by year or location *** The plots of magnitude by year, broken out by longitude and by longitude, show, as would be expected, a higher frequency and magnitude within certain geographic areas.

```
# magnitude by year, split by longitude
(m1 <- qplot(data = df.quake, x = year, y = mag, colour = long.rnd, facets = ~ long.rnd))
```

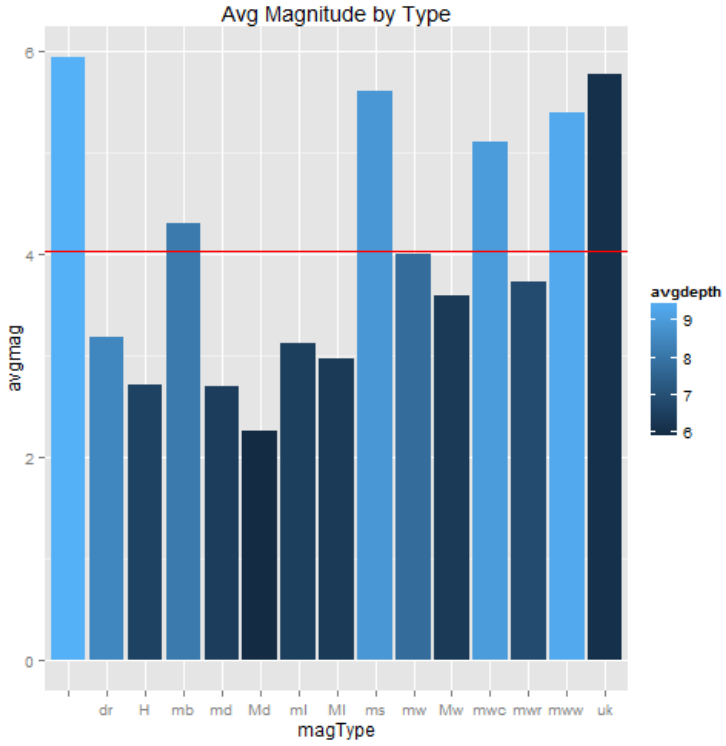



```
# magnitude by longitude, split by latitude
(m2 <- qplot(data = df.quake,  x = long, y = mag, colour = lat.rnd, facets = ~ lat.rnd))
```



*** magnitude related to magnitude type *** Different methods (magType) are used for calculating the magnitude, apparently directly related to the magnitude and depth

```
c <- ggplot(data = s7, aes(x = magType, y = avgmag, fill = avgdepth))
c + geom_bar(stat = "identity") + geom_hline(aes(yintercept = mean(s7$avgmag)),colour = "red")+ ggtitle("Avg Magnitude by Type")
```




```
aov.out <- aov(mag ~ magType, data = df.quake)
summary(lm(aov.out))
```

```
##
## Call:
## lm(formula = aov.out)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4000 -0.2327 -0.0012  0.1988  2.4988
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.943      0.140   42.48  < 2e-16 ***
## magTypedr     -2.758      0.143  -19.26  < 2e-16 ***
## magTypeH      -3.230      0.192  -16.86  < 2e-16 ***
## magTypemb      -1.643      0.142  -11.58  < 2e-16 ***
## magTypemd      -3.242      0.140  -23.14  < 2e-16 ***
## magTypeMd      -3.678      0.140  -26.23  < 2e-16 ***
## magTypeml      -2.822      0.140  -20.15  < 2e-16 ***
## magTypeMl      -2.967      0.147  -20.23  < 2e-16 ***
## magTypems      -0.343      0.171   -2.00    0.045 *
## magTypemw      -1.942      0.147  -13.20  < 2e-16 ***
## magTypeMw      -2.351      0.152  -15.47  < 2e-16 ***
## magTypemwc     -0.843      0.198   -4.26  2.1e-05 ***
## magTypemwr     -2.209      0.141  -15.63  < 2e-16 ***
## magTypemww     -0.543      0.396   -1.37    0.170
## magTypeuk      -0.176      0.255   -0.69    0.490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.37 on 8892 degrees of freedom
## Multiple R-squared:  0.615,  Adjusted R-squared:  0.614
## F-statistic: 1.01e+03 on 14 and 8892 DF,  p-value: <2e-16
```

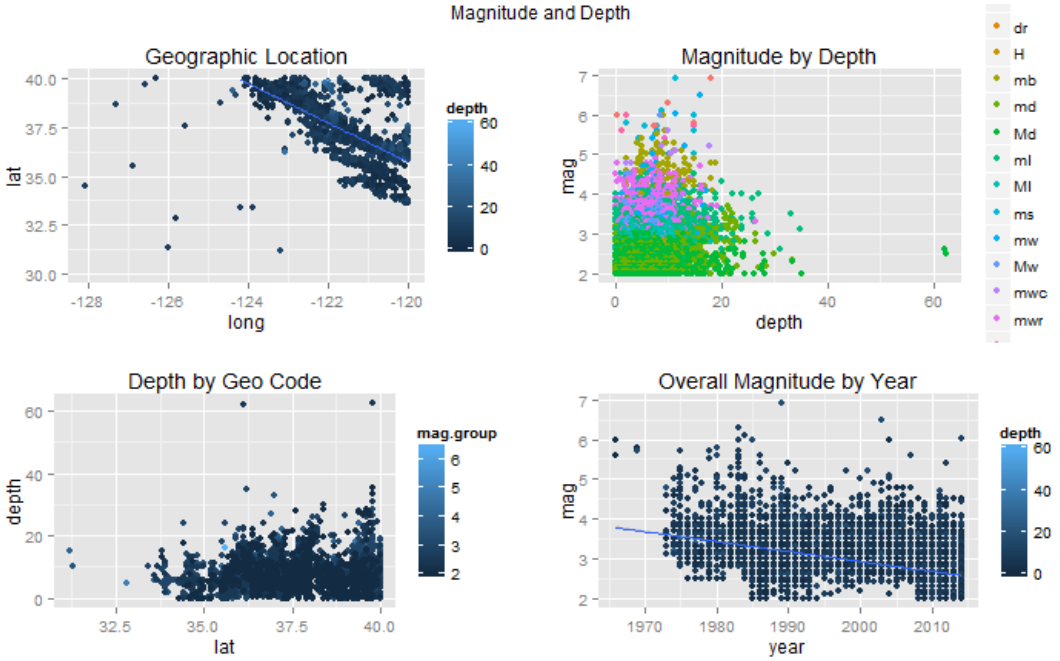
Additional Comparisons

- Amapping geographic coordinates shows the expected correlation between frequency and geographic location.
- Higher magnitude are clustered chiefly between depth of 0 and 20, methods of measuring magnitude is related to the magnitude.
- Depth appears to vary slightly across different latitudes
- Based on the data provided, there appears to be an inverse relationship between overall magnitude and year. This should be interpreted cautiously, as the increased sensitivity and prevalence of data collection over time would increase the number small magnitude observations.

```
# fequency
m3 <- qplot(data = df.quake, x = long, y = lat, colour = depth)+ stat_smooth(method = "lm") + scale_y_continuous(limits= c(30,40)) +
ggtitle("Geographic Location")
m4 <- qplot(data = df.quake, x = depth, y = mag, colour = magType) + ggtitle("Magnitude by Depth")
m5 <- qplot(data = df.quake, x = lat, y = depth, colour = mag.group)+ ggtitle("Depth by Geo Code")
m6 <- qplot(data = df.quake, x = year, y = mag, colour = depth)+stat_smooth(method=lm) + ggtitle("Overall Magnitude by Year")
```

```
grid.arrange(m3,m4,m5,m6, ncol=2, main = "Magnitude and Depth")
```

```
## Warning: Removed 38 rows containing missing values (geom_path).
```


Magnitude and Depth
Geographic Location
lat
long
depth
Magnitude by Depth
mag
depth
dr
H
mb
md
Md
ml
MI
ms
mw
Mw
mwc
mwr
Depth by Geo Code
depth
lat
mag.group
Overall Magnitude by Year
mag
year
depth