# Data exploration

Hanna Pylieva

Loading earthquakes data for year 2017.

```
dat <- read.csv("earthquakes.csv", stringsAsFactors = T)
head(dat)
```

```
##                  time latitude  longitude depth  mag magType nst    gap
## 1 2017-01-01 23:59:10 37.08550  -98.04067  6.36 1.56      ml  12 137.00
## 2 2017-01-01 23:50:58 62.79880 -149.46730  6.00 0.90      ml  NA     NA
## 3 2017-01-01 23:50:39 32.96683 -115.56283  9.88 2.56      ml  47  75.00
## 4 2017-01-01 23:50:07 38.42840 -118.88920  5.50 1.20      ml   6 210.16
## 5 2017-01-01 23:47:17 38.42840 -118.90130  6.70 1.90      ml  13  70.02
## 6 2017-01-01 23:46:45 38.40500 -118.91680  7.80 2.10      ml  29  57.64
##      dmin    rms        net                 id             updated
## 1 0.05775 0.0400 ismpkansas ismpkansas70216818 2017-03-02 15:40:54
## 2      NA 0.6300         ak         ak14881057 2017-01-06 02:34:10
## 3 0.06521 0.2400         ci         ci37778160 2017-02-08 19:29:28
## 4 0.07200 0.0885         nn         nn00571954 2017-03-14 05:05:00
## 5 0.07200 0.1497         nn         nn00575954 2017-03-14 05:05:02
## 6 0.05200 0.1966         nn         nn00571953 2017-03-07 00:13:01
##                          place       type horizontalError depthError
## 1      7km S of Anthony, Kansas earthquake            0.19       0.61
## 2 62km NNE of Talkeetna, Alaska earthquake              NA       0.40
## 3        3km WSW of Brawley, CA earthquake            0.21       0.52
## 4 25km WSW of Hawthorne, Nevada earthquake              NA       7.90
## 5 26km WSW of Hawthorne, Nevada earthquake              NA       1.90
## 6 28km WSW of Hawthorne, Nevada earthquake              NA       1.20
##   magError magNst   status locationSource magSource
## 1    0.128     13 reviewed           ismp      ismp
## 2       NA     NA reviewed             ak        ak
## 3    0.236     24 reviewed             ci        ci
## 4    0.250      2 reviewed             nn        nn
## 5    0.250      8 reviewed             nn        nn
## 6    0.370     18 reviewed             nn        nn
```

At first we need to understand the dataset. Let's look at the columns we have and their types. The description of columns is available here: https://earthquake.usgs.gov/earthquakes/feed/v1.0/csv.php

```
str(dat)
```

```
## 'data.frame':    126955 obs. of  22 variables:
##  $ time           : Factor w/ 126643 levels "2017-01-01 00:04:06",..: 643 642 641 640 639 638 637 636
##  $ latitude       : num  37.1 62.8 33 38.4 38.4 ...
##  $ longitude      : num  -98 -149 -116 -119 -119 ...
##  $ depth          : num  6.36 6 9.88 5.5 6.7 7.8 5.26 6.2 6.2 6.5 ...
##  $ mag            : num  1.56 0.9 2.56 1.2 1.9 2.1 1.06 0.9 1.4 1.3 ...
##  $ magType        : Factor w/ 14 levels "mb","Mb","mb_lg",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ nst            : int  12 NA 47 6 13 29 7 4 5 10 ...
##  $ gap            : num  137 NA 75 210 70 ...
##  $ dmin           : num  0.0578 NA 0.0652 0.072 0.072 ...
##  $ rms            : num  0.04 0.63 0.24 0.0885 0.1497 ...
##  $ net            : Factor w/ 14 levels "ak","ci","hv",..: 4 1 2 9 9 9 2 9 9 9 ...
```

```
##  $ id              : Factor w/ 126955 levels "ak14868407","ak14868411",..: 56416 72 44575 84253 85732
##  $ updated         : Factor w/ 92018 levels "2017-01-01 00:33:49",..: 10327 346 4982 13029 13030 1112
##  $ place           : Factor w/ 45844 levels "0km ENE of Bainbridge Island, Washington",..: 40126 3457
##  $ type            : Factor w/ 11 levels "chemical explosion",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ horizontalError: num  0.19 NA 0.21 NA NA NA 0.44 NA NA NA ...
##  $ depthError      : num  0.61 0.4 0.52 7.9 1.9 1.2 0.83 5 2.2 3.8 ...
##  $ magError        : num  0.128 NA 0.236 0.25 0.25 0.37 0.113 0.24 0.32 0.14 ...
##  $ magNst          : int  13 NA 24 2 8 18 5 3 4 7 ...
##  $ status          : Factor w/ 2 levels "automatic","reviewed": 2 2 2 2 2 2 2 2 2 2 ...
##  $ locationSource : Factor w/ 23 levels "ak","aust","bgs",..: 7 1 4 12 12 12 4 12 12 12 ...
##  $ magSource       : Factor w/ 25 levels "ak","bgs","buc",..: 8 1 5 13 13 13 5 13 13 13 ...
```

Find the absolute number of NAs in columns.

```
colSums(is.na(dat))
```

```
##            time        latitude       longitude           depth
##               0               0               0               0
##             mag         magType             nst             gap
##              16              16           55580           37947
##            dmin             rms             net              id
##           39084              49               0               0
##         updated           place            type horizontalError
##               0               0               0           50558
##      depthError        magError          magNst          status
##              45           41280           40060               0
##  locationSource       magSource
##               0               0
```

And the fractions in % of NAs for each column are:

```
colSums(is.na(dat)/nrow(dat)*100)
```

```
##            time        latitude       longitude           depth
##      0.00000000      0.00000000      0.00000000      0.00000000
##             mag         magType             nst             gap
##      0.01260289      0.01260289     43.77929188     29.89011855
##            dmin             rms             net              id
##     30.78571147      0.03859635      0.00000000      0.00000000
##         updated           place            type horizontalError
##      0.00000000      0.00000000      0.00000000     39.82355953
##      depthError        magError          magNst          status
##      0.03544563     32.51545823     31.55448781      0.00000000
##  locationSource       magSource
##      0.00000000      0.00000000
```

Now let's look at the data... General distribution of eathquakes on the Earth during the last year.

```
library(dplyr)
```

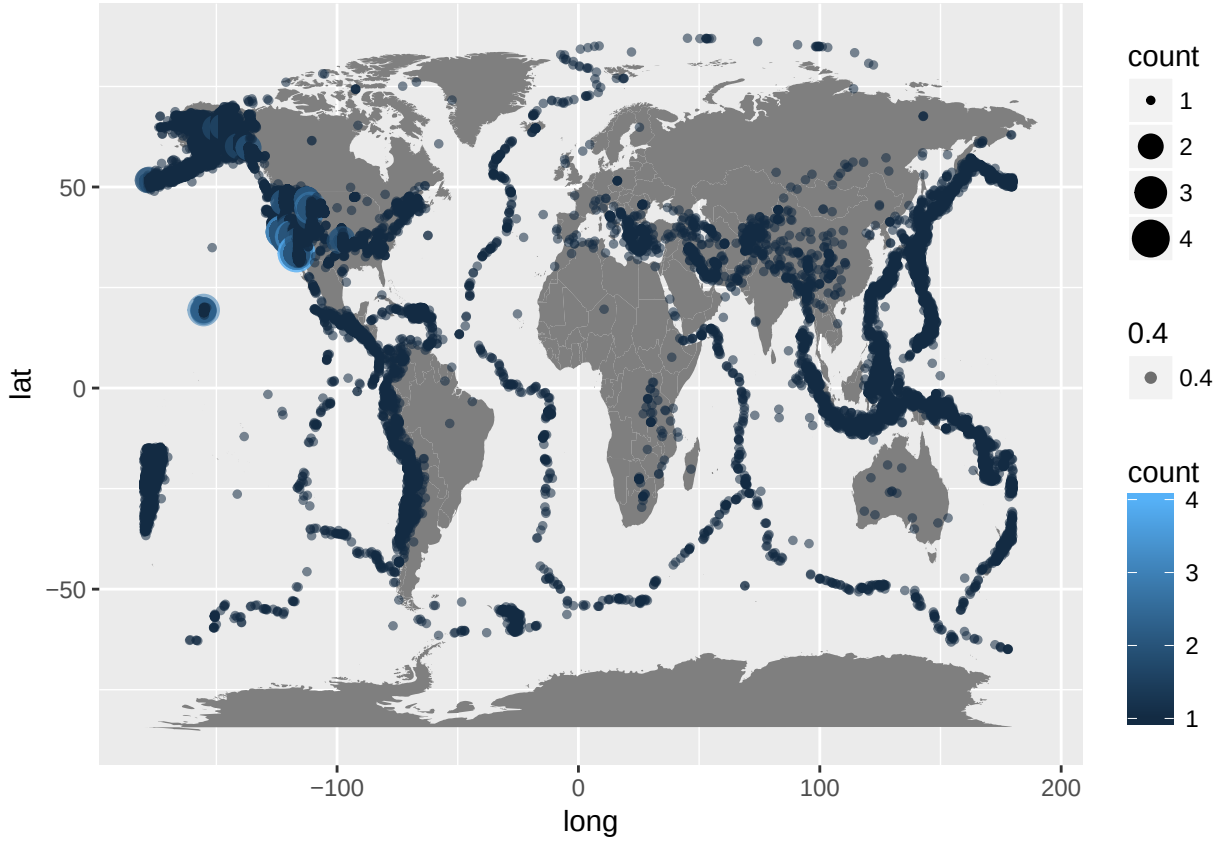
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
```

```
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(maps)
source("utils.R")
```

```
by_place <- dat %>% group_by(longitude, latitude, place) %>% summarize(count = n())
```

```
mdat <- map_data('world')

# str(mdat)
ggplot() +
  geom_polygon(dat=mdat, aes(long, lat, group=group), fill="grey50") +
  geom_point(data=by_place, aes(x=longitude, y=latitude , size = count, col = count, alpha = 0.4))
```



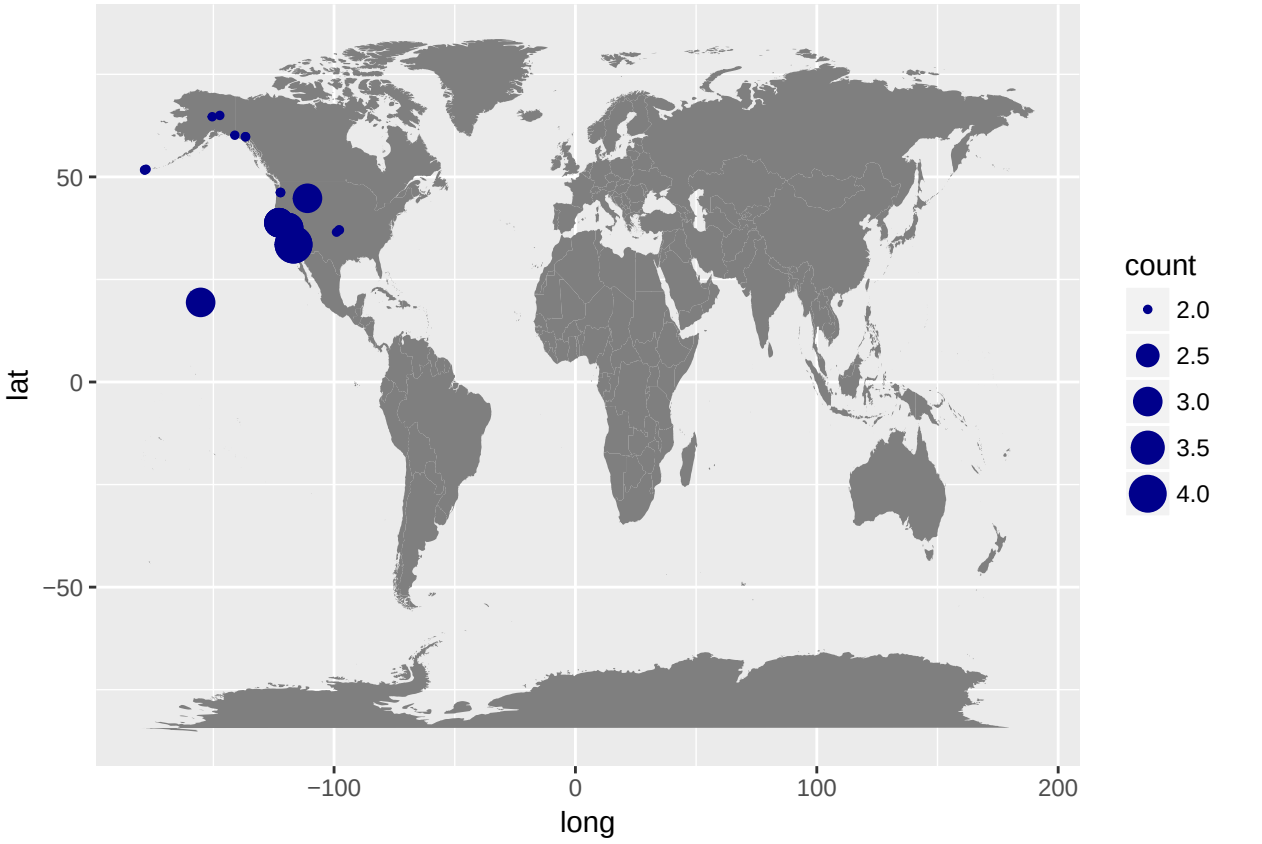

```
  # scale_fill_distiller( direction = 1)
```

The places where earthquakes occured more than 1 time.

```
by_place_filtered <- by_place %>% filter(count>1)
str(mdat)
```

```
## 'data.frame':    99338 obs. of  6 variables:
##  $ long     : num  -69.9 -69.9 -69.9 -70 -70.1 ...
##  $ lat      : num  12.5 12.4 12.4 12.5 12.5 ...
##  $ group    : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ order    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ region   : chr  "Aruba" "Aruba" "Aruba" "Aruba" ...
##  $ subregion: chr  NA NA NA NA ...
```

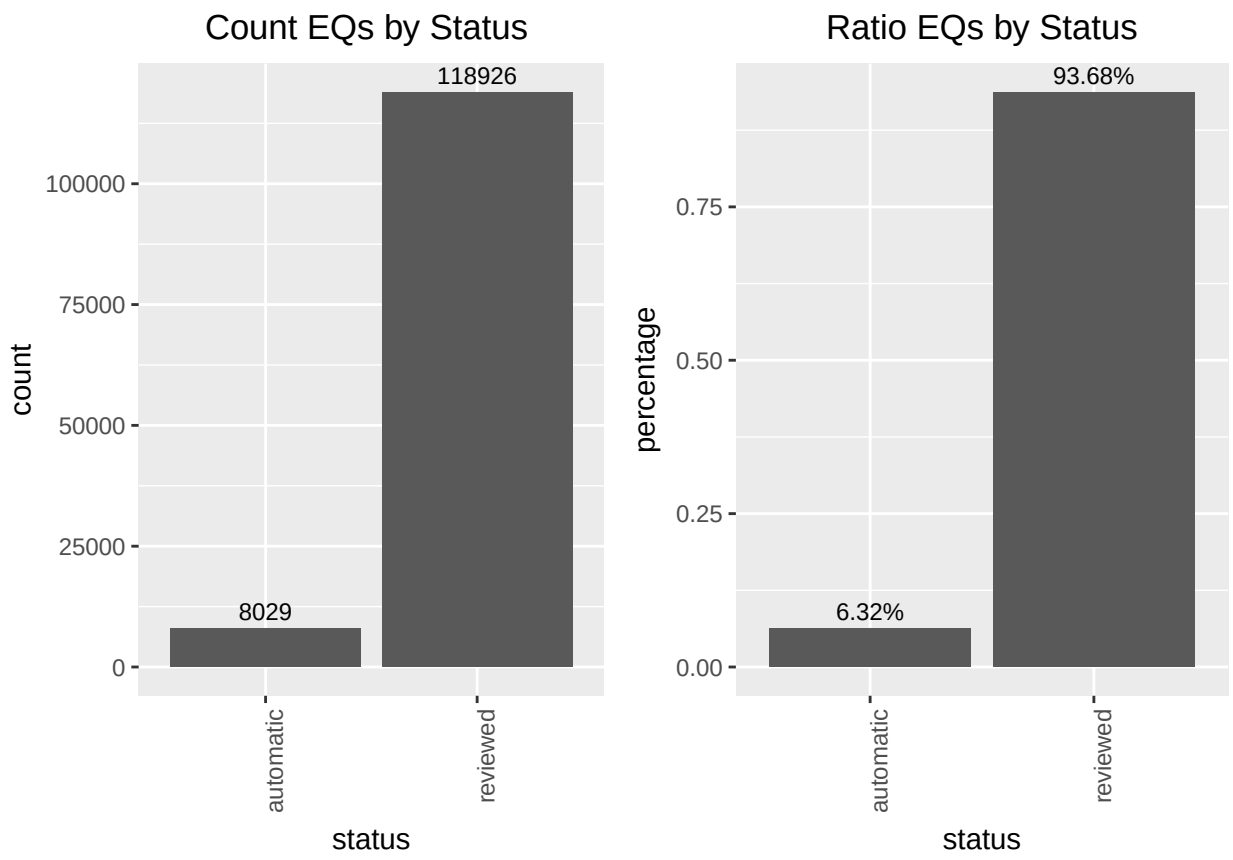
```
ggplot() +
  geom_polygon(dat=mdat, aes(long, lat, group=group), fill="grey50") +
  geom_point(data=by_place_filtered,
             aes(x=longitude, y=latitude , size = count), col="darkblue")
```



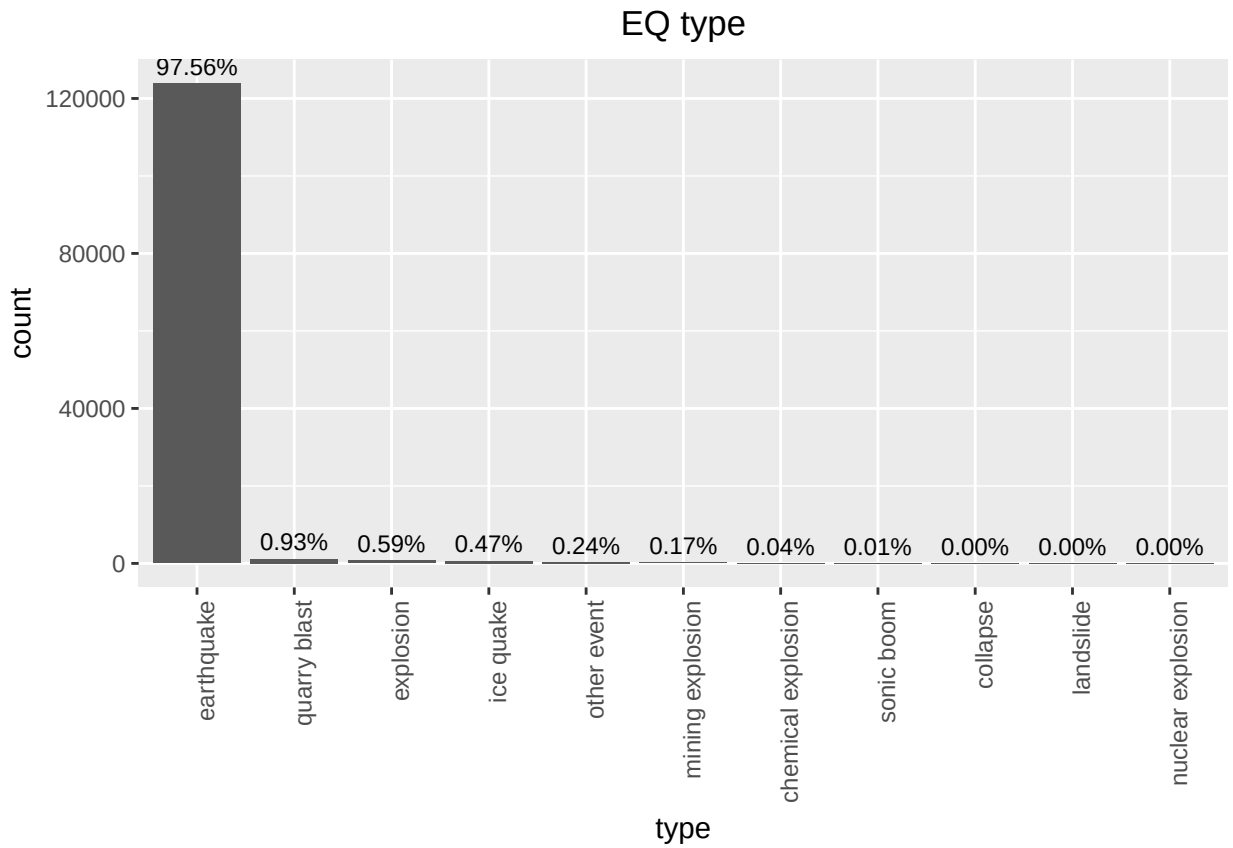

**Let's understand how our records are distributed by categorical variables.**

```
colnames <- as.data.frame(colnames(dat))
```

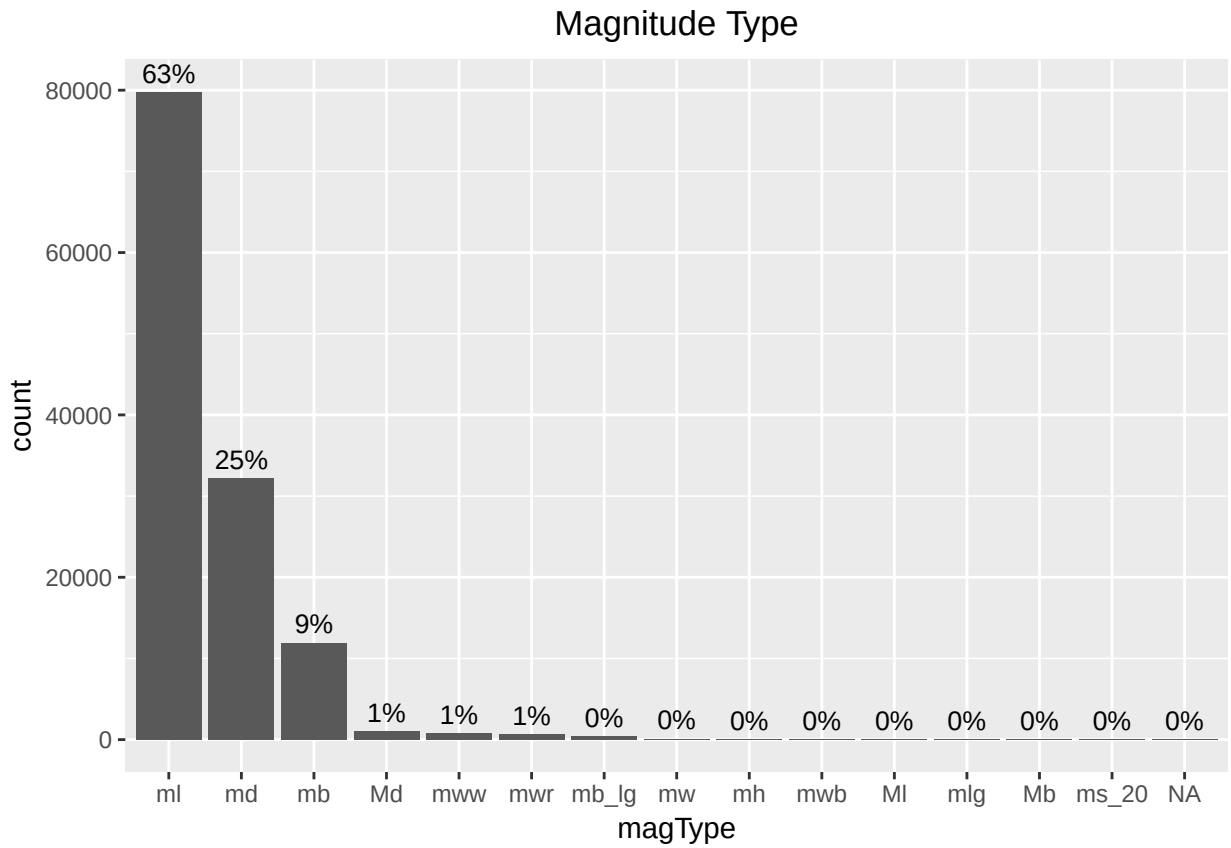
```
# alternatively
# ggplot(dat, aes(status)) + geom_bar(aes(y = (..count..)/sum(..count..))) + ggtitle("Status") + labs(y
  # theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(angle = 90, hjust = 1))

p1 <- ggplot(dat, aes(status)) + geom_bar() + ggtitle("Count EQs by Status ") +
  geom_text(stat='count', aes(label=..count..),position = position_dodge(width = 1), vjust = -0.5, size
  theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(angle = 90, hjust = 1))

by_status_perc <- dat %>% group_by(status) %>% summarise(percentage = n() / nrow(dat))
p2 <- ggplot(by_status_perc, aes(x = status, y = percentage, label=sprintf("%0.2f%%", round(percentage*
    geom_bar(stat = "identity") + ggtitle("Ratio EQs by Status") +
    geom_text(position = position_dodge(width = 1), vjust = -0.5, size = 3) +
    theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(angle = 90, hjust = 1))
multiplot(p1, p2, cols = 2)
```



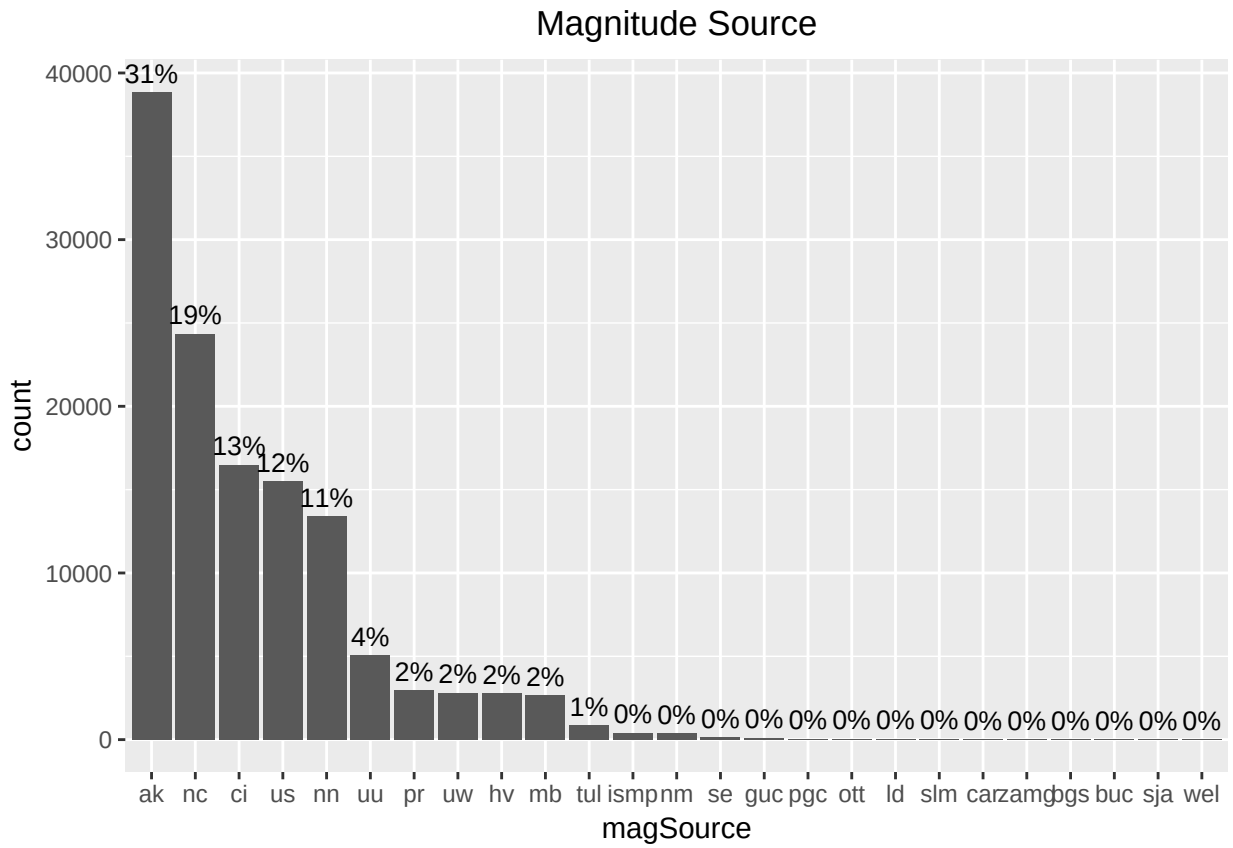

```
by_type<- dat %>% group_by(type) %>% summarize(count = n(), percentage = n() / nrow(dat))
by_type$type <- factor(by_type$type, levels = by_type$type[order(-by_type$count)] )

ggplot(by_type, aes(x = type, y = count, label=sprintf("%0.2f%%", round(percentage*100, digits = 2)))) +
  geom_bar(stat = "identity" ) + ggtitle("EQ type") +
  geom_text(aes(y = count) ,position = position_dodge(width = 1), vjust = -0.5, size = 3) +
  theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(angle = 90, hjust = 1))
```



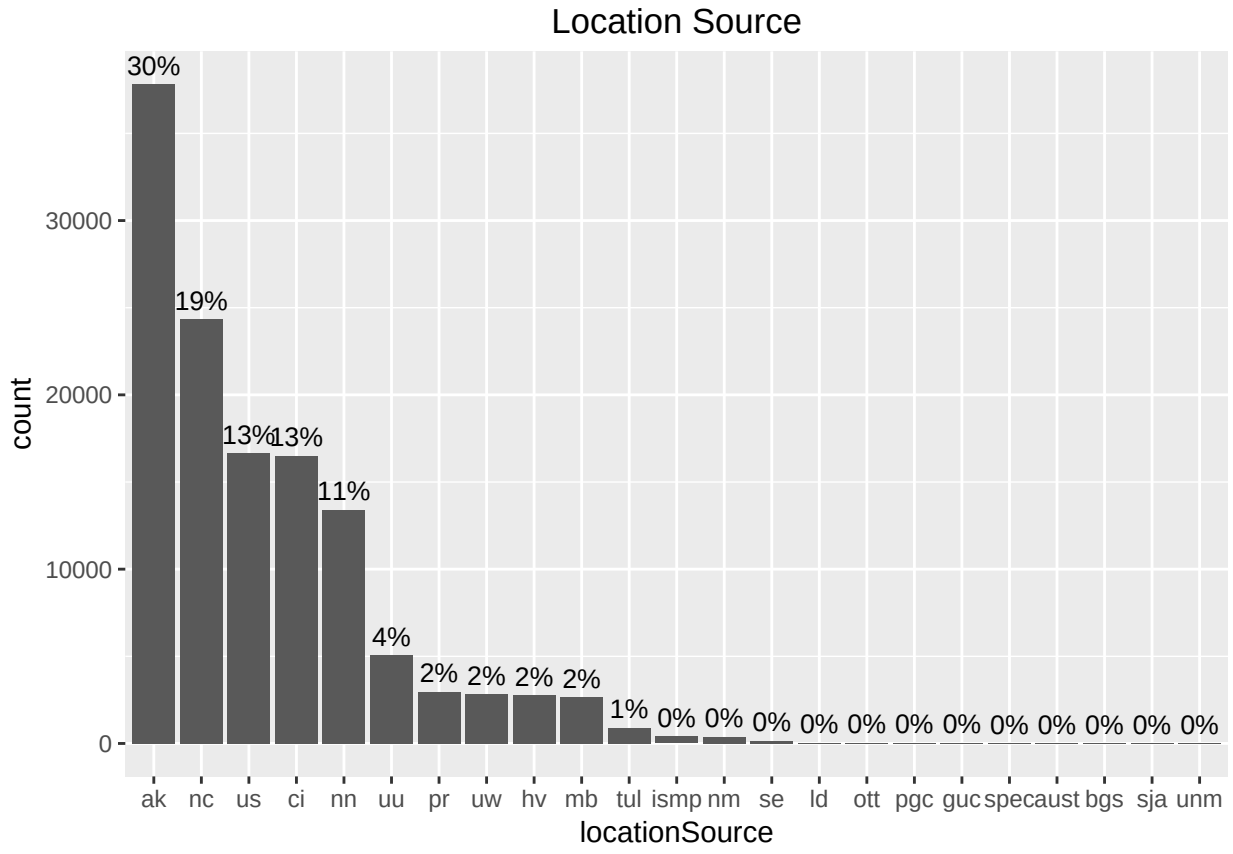

```
by_magtype <- dat %>% group_by(magType) %>% summarize(count = n(), percentage = n() / nrow(dat))
by_magtype$magType <- factor(by_magtype$magType, levels = by_magtype$magType[order(-by_magtype$count)] )
ggplot(by_magtype, aes(x = magType, y = count, label=sprintf("%0.0f%%", round(percentage*100, digits = 1
  geom_bar(stat = "identity") +ggtitle("Magnitude Type") +
  geom_text(aes(y = count) ,position = position_dodge(width = 1), vjust = -0.5, size = 3.5) +
  theme(plot.title = element_text(hjust = 0.5))
```



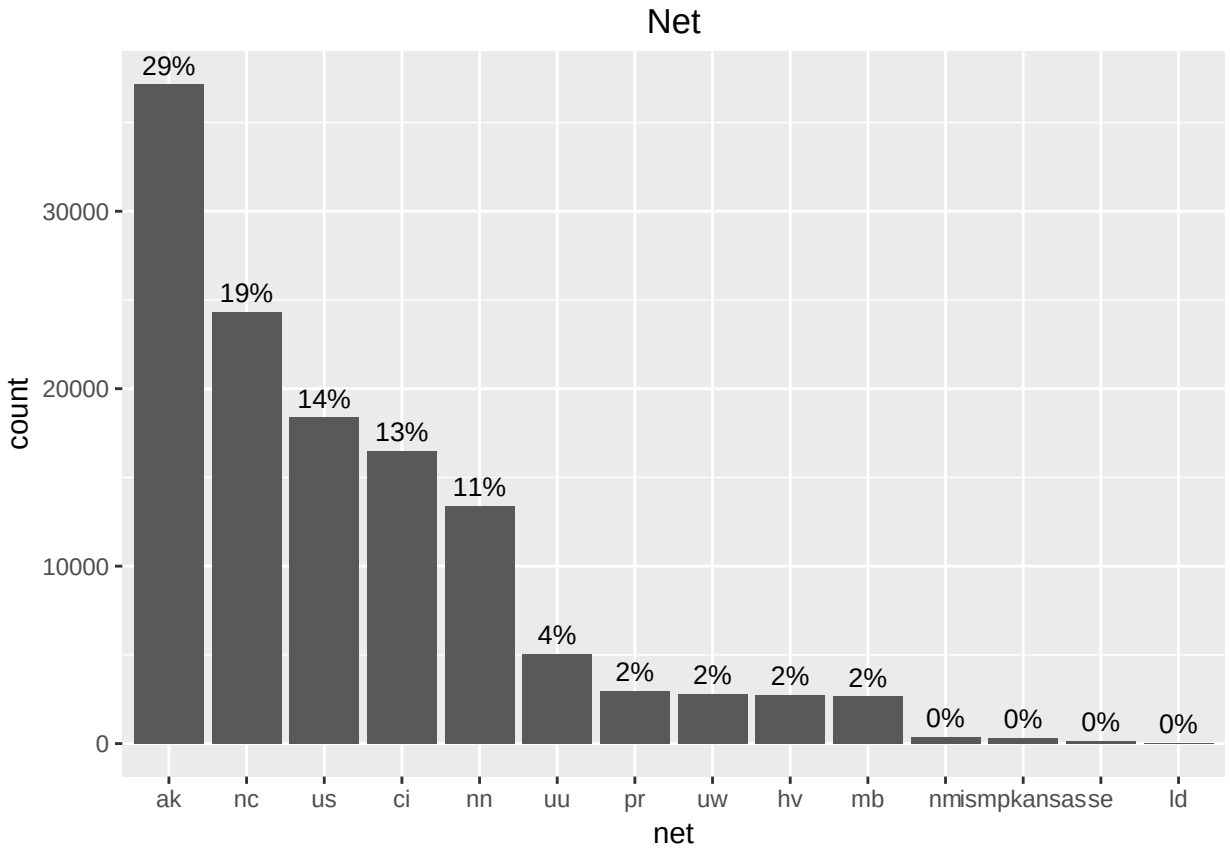

```
by_magSource <- dat %>% group_by(magSource) %>% summarize(count = n(), percentage = n() / nrow(dat))
by_magSource$magSource <- factor(by_magSource$magSource, levels = by_magSource$magSource[order(-by_magS
ggplot(by_magSource, aes(x = magSource, y = count, label=sprintf("%0.0f%%", round(percentage*100, digit
  geom_bar(stat = "identity") +ggtitle("Magnitude Source") +
  geom_text(aes(y = count) ,position = position_dodge(width = 1), vjust = -0.5, size = 3.5) +
  theme(plot.title = element_text(hjust = 0.5))
```




```
by_locationSource <- dat %>% group_by(locationSource) %>% summarize(count = n(), percentage = n() / nrow
by_locationSource$locationSource <- factor(by_locationSource$locationSource,
                                           levels = by_locationSource$locationSource[order(-by_locationS
ggplot(by_locationSource, aes(x = locationSource, y = count, label=sprintf("%0.0f%%", round(percentage*1
  geom_bar(stat = "identity") +ggtitle("Location Source") +
  geom_text(aes(y = count) ,position = position_dodge(width = 1), vjust = -0.5, size = 3.5) +
  theme(plot.title = element_text(hjust = 0.5))
```

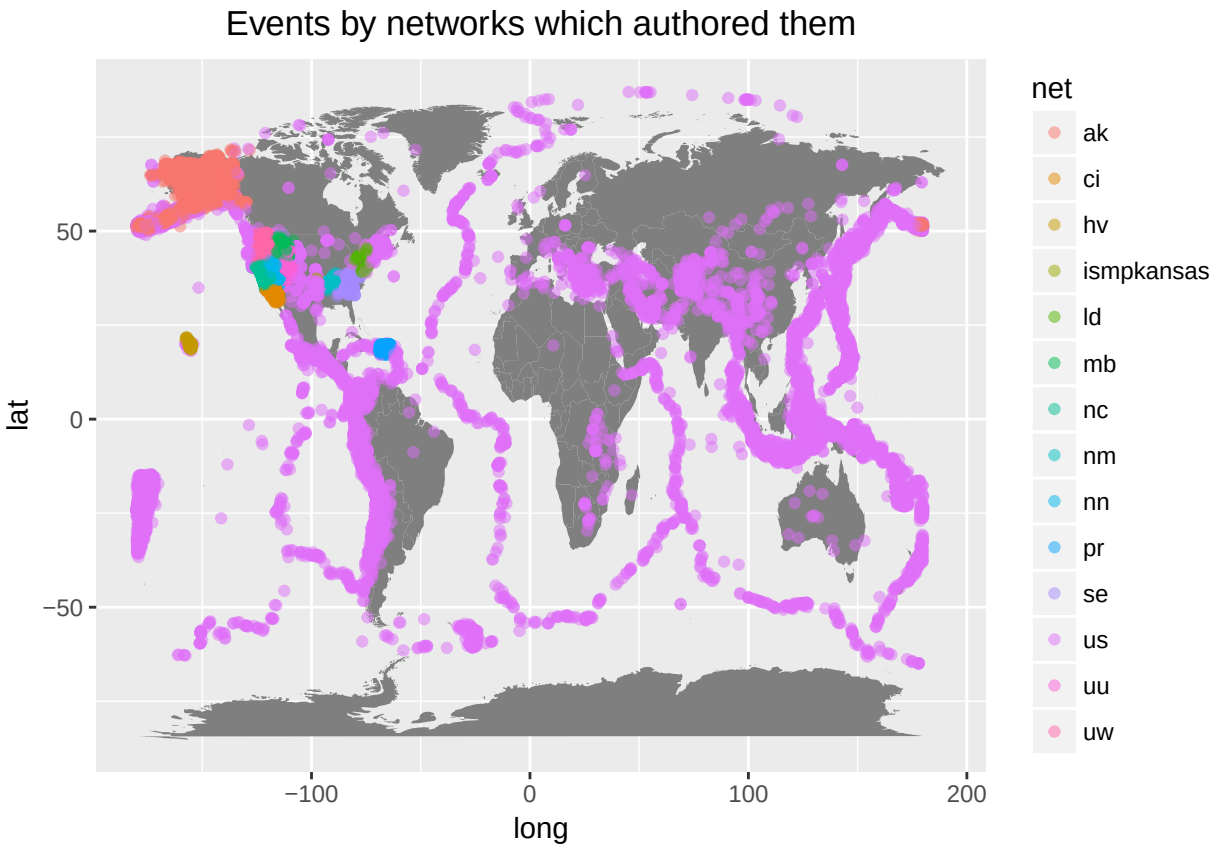



```
by_net <- dat %>% group_by(net) %>% summarize(count = n(), percentage = n() / nrow(dat))
by_net$net <- factor(by_net$net, levels = by_net$net[order(-by_net$count)] )
ggplot(by_net, aes(x = net, y = count, label=sprintf("%0.0f%%", round(percentage*100, digits = 2)))) +
  geom_bar(stat = "identity") + ggtitle("Net") +
  geom_text(aes(y = count) ,position = position_dodge(width = 1), vjust = -0.5, size = 3.5) +
  theme(plot.title = element_text(hjust = 0.5))
```



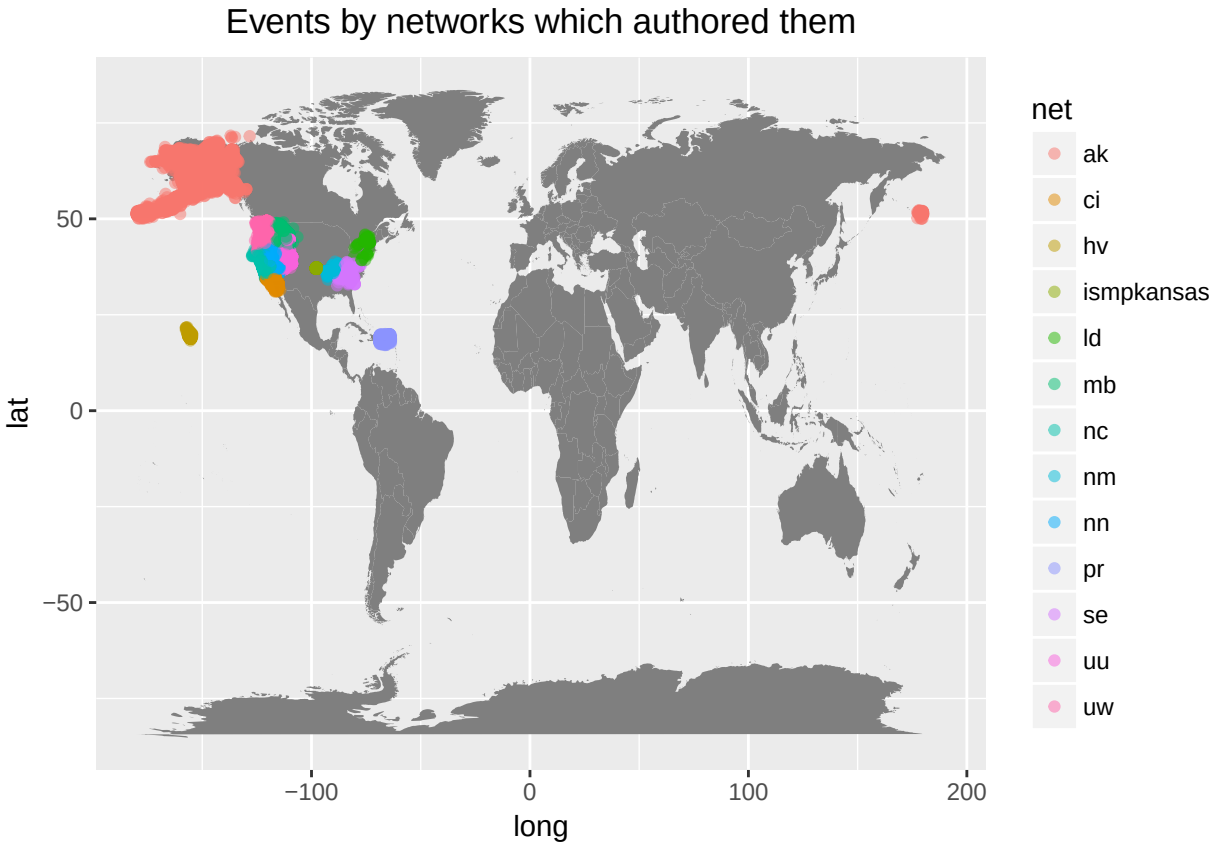

Summary:
1. Most of EQs (94%) were reviewded by people.
2. Most of events (97.5%) are eqrthquakes, all the rest types of events have ration less than 1% in the total number of events.
3. 63% of events have local(ml) magnitude type (predumably with magnitude range 2 - 7.5).
25% of events were of magnitude type Duration (magnitude range <4) and 9% of events were of type Short-period surface wave (magnitude range 3.5-7)
4. It looks like it is the same network which originally authored both magnitude and location of the event (as distributions of magSource/locationSource/net by networks are similar). The networks which authored the most of events (magSource/locationSource/net) is ak (31% of events). It is interesting, events of which location did each network authored.

```
# hjust - horizontal justification
ggplot() +
  geom_polygon(dat=mdat, aes(long, lat, group=group), fill="grey50") +
  geom_point(data=dat, aes(x=longitude, y=latitude ,  col = net), alpha = 0.5) +
  ggtitle("Events by networks which authored them") +
  theme(plot.title = element_text(hjust = 0.5))
```




Looks like ak network is a local one for Alaska region, whereas us networ reports about event worldwide. Let's look at the picture without us net.

```
dat_no_us <- dat %>% filter(net != 'us')
ggplot() +
  geom_polygon(dat=mdat, aes(long, lat, group=group), fill="grey50") +
  geom_point(data=dat_no_us, aes(x=longitude, y=latitude ,  col = net), alpha = 0.5) +
  ggtitle("Events by networks which authored them") +
  theme(plot.title = element_text(hjust = 0.5))
```



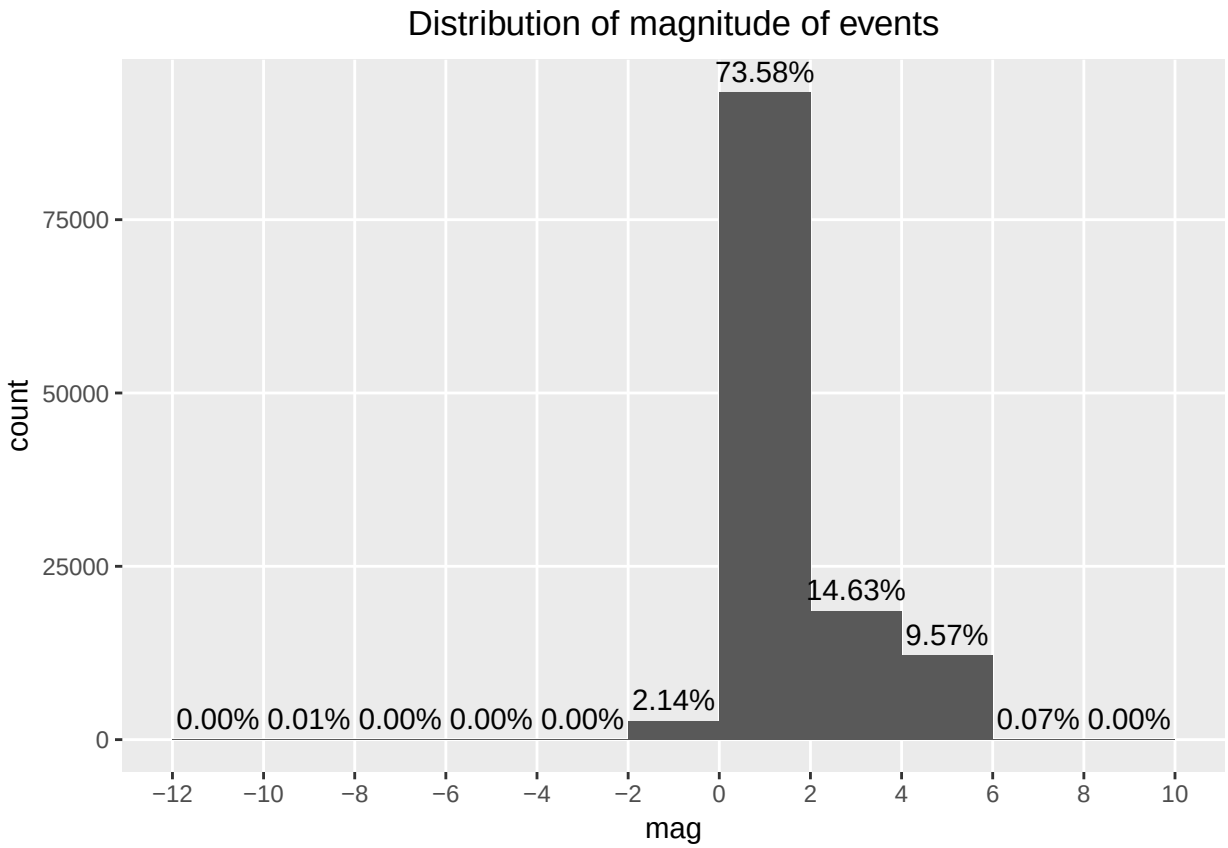

So **us** is the only network which reports about events worldwide. All the rest of networs are local. It's also interesting that that us network has reported only 14% of events, but as they are very distributed on the map one can think from the first glance that us network reported 80% of events. This also tells us that there are areas which need special attention (like Alaska), and there local networks were established. And the number of events on the rest of world is relatively smaller.

**Let's not look at float features and their distribution.**

```
breaks = seq(-12,10,by=2)
ggplot(dat, aes(x=mag)) + geom_histogram(breaks = breaks) +
  stat_bin(geom="text", aes(label=sprintf("%0.2f%%",round(..count../sum(..count..)*100,2))),
           position = position_dodge(width = 1), vjust = -0.5,
           breaks = breaks) +
  scale_x_continuous(breaks = breaks) +
  ggtitle("Distribution of magnitude of events") +
  theme(plot.title = element_text(hjust = 0.5),panel.grid.minor = element_blank())
```

```
## Warning: Removed 16 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 16 rows containing non-finite values (stat_bin).
```

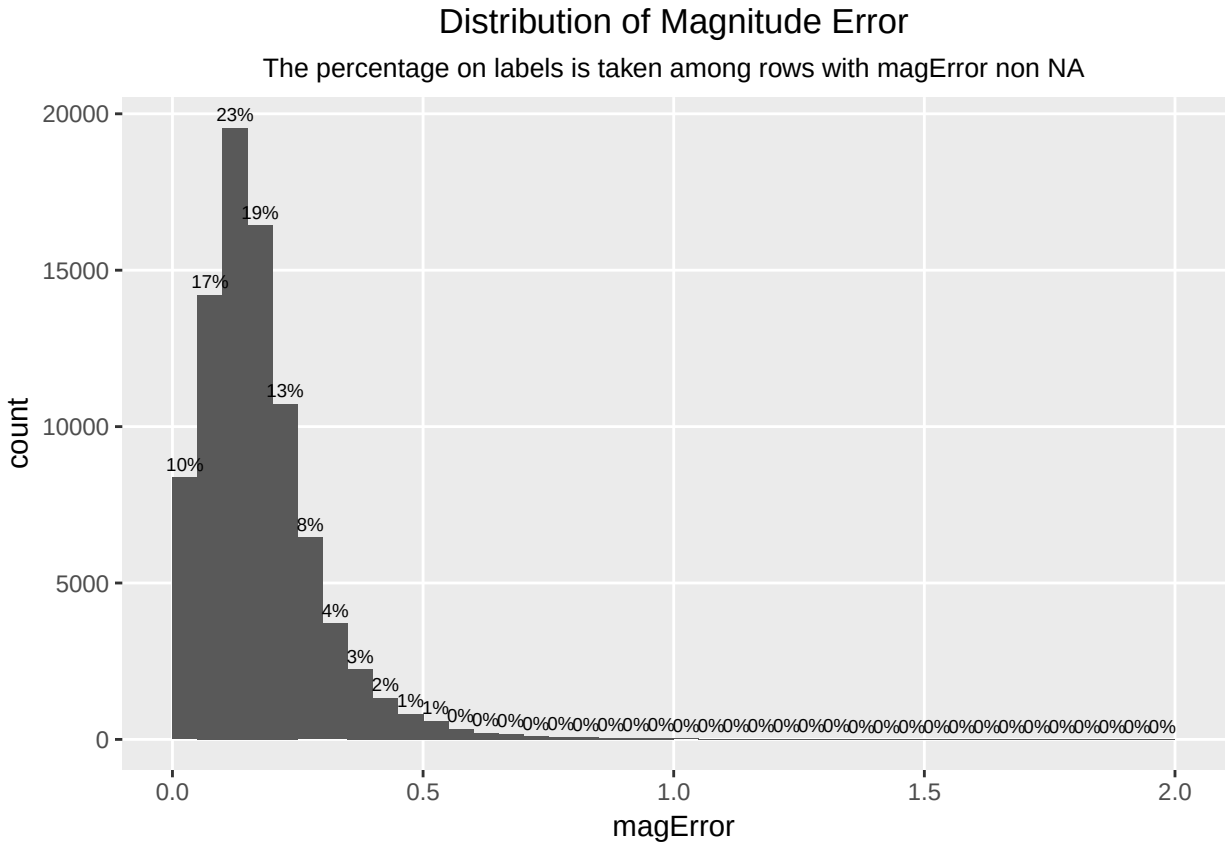



Let's look how trustfull are measurements of magnitude in general.

```
breaks = seq(0,2,by=0.05)
ggplot(dat, aes(x=magError)) + geom_histogram(breaks=breaks) +
  stat_bin(geom="text", aes(label=sprintf("%0.0f%%",round(..count../sum(..count..)*100,2))),
           position = position_dodge(width = 1), vjust = -0.5,
           breaks = breaks, size = 2.5) +
  ggtitle("Distribution of Magnitude Error", subtitle = "The percentage on labels is taken among rows wi
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5), panel.grid.mi
```

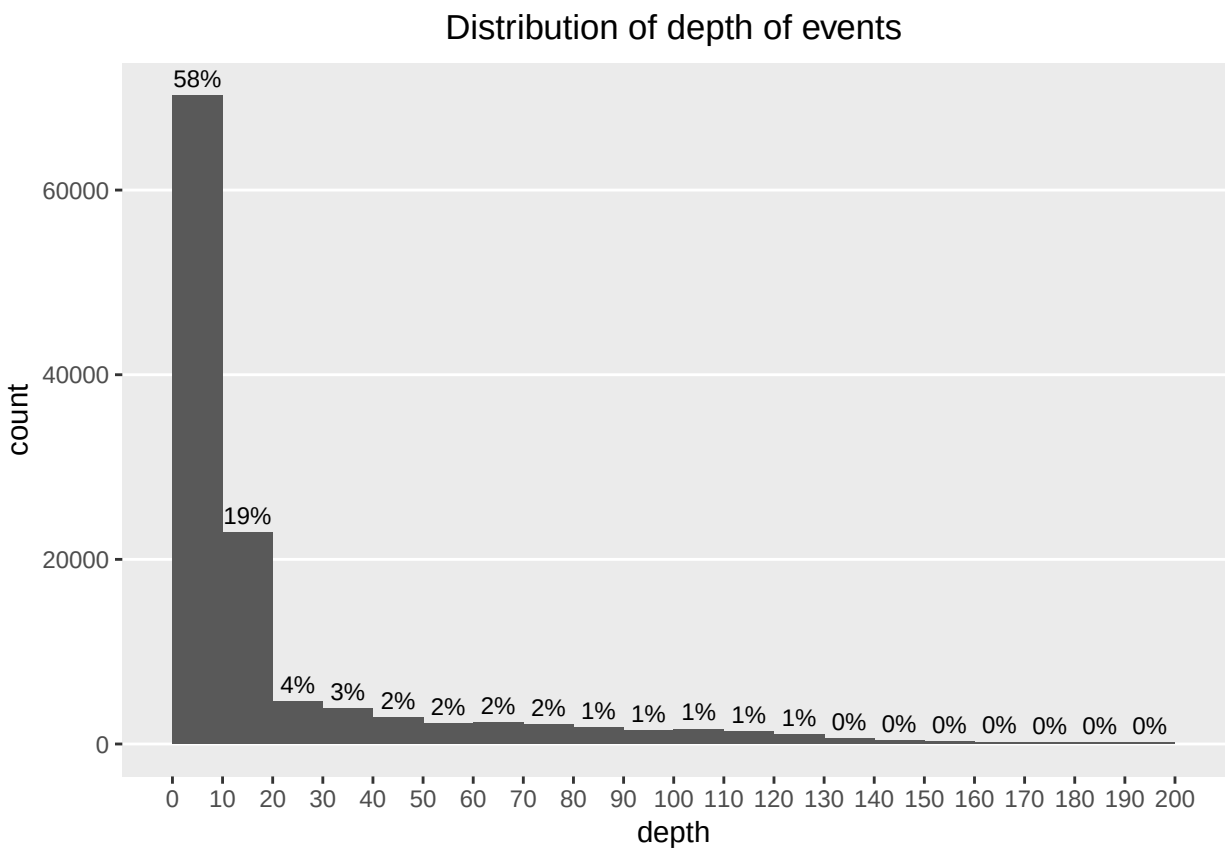
```
## Warning: Removed 41280 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 41280 rows containing non-finite values (stat_bin).
```




Pay attention that 41280 rows (32.5% of dataset) contain NA values and are not depicted on the plot above. The NA values mean that the contributing seismic network does not supply uncertainty estimates.

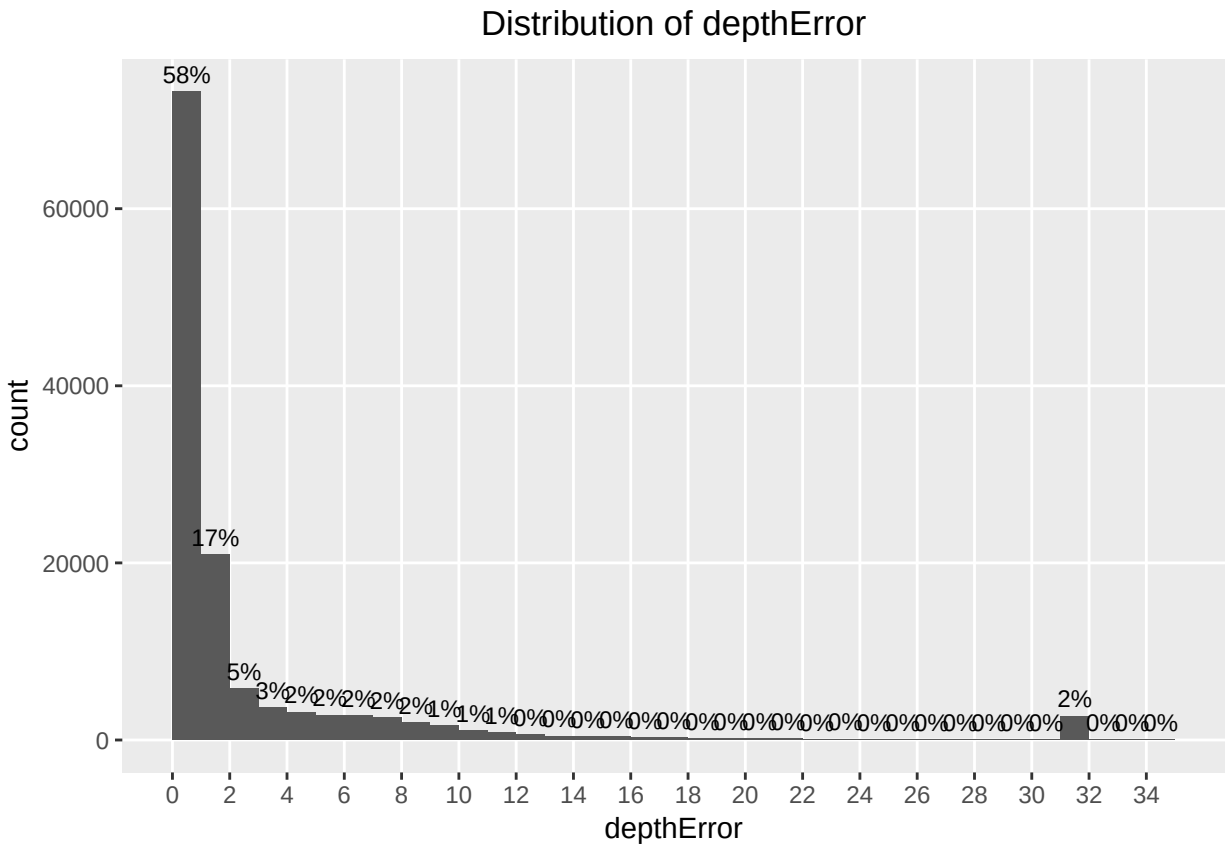
```
breaks = seq(0,200,by=10)
ggplot(dat, aes(x=depth)) + geom_histogram(breaks = breaks) +
  stat_bin(geom="text", aes(label=sprintf("%0.0f%%",round(..count../sum(..count..)*100,2))),
           position = position_dodge(width = 1), vjust = -0.5,
           breaks = breaks, size = 3) +
  scale_x_continuous(breaks = breaks) +
  ggtitle("Distribution of depth of events") +
  theme(plot.title = element_text(hjust = 0.5), panel.grid.major.x = element_blank(), panel.grid.minor =
```




Let's underestand how trustfull are measurements of depth in general.

```
breaks = seq(0,35,by=1)

ggplot(dat, aes(x=depthError)) + geom_histogram(breaks = breaks) +
  stat_bin(geom="text", aes(label=sprintf("%0.0f%%",round(..count../sum(..count..)*100,2))),
           position = position_dodge(width = 1), vjust = -0.5,
           breaks = breaks, size = 3) +
  scale_x_continuous(breaks = seq(0,35,by=2) ) +
  ggtitle("Distribution of depthError") +
  theme(plot.title = element_text(hjust = 0.5),panel.grid.minor = element_blank())
```

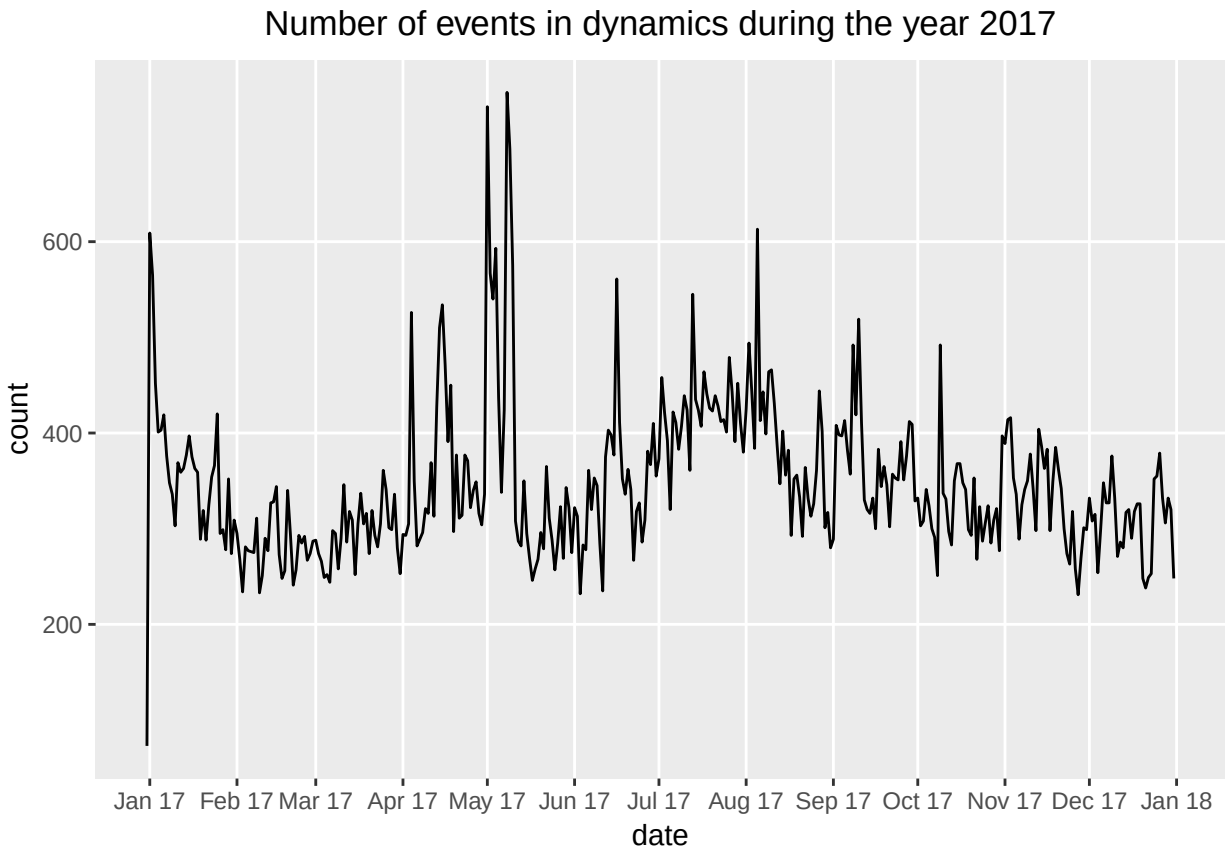
```
## Warning: Removed 45 rows containing non-finite values (stat_bin).

## Warning: Removed 45 rows containing non-finite values (stat_bin).
```




Most depth is identified as less than 10 (almost 60% of data), whereas from data description we can see that the depth could have been set to a fixed value and 5 or 10 km are often used in mid-continental areas and on mid-ocean ridges. But anyway we can see that most of the EQs are shallow (depth less than 60km).
The error in depth up to 1km mostly makes depth trustful enough from my point of view.

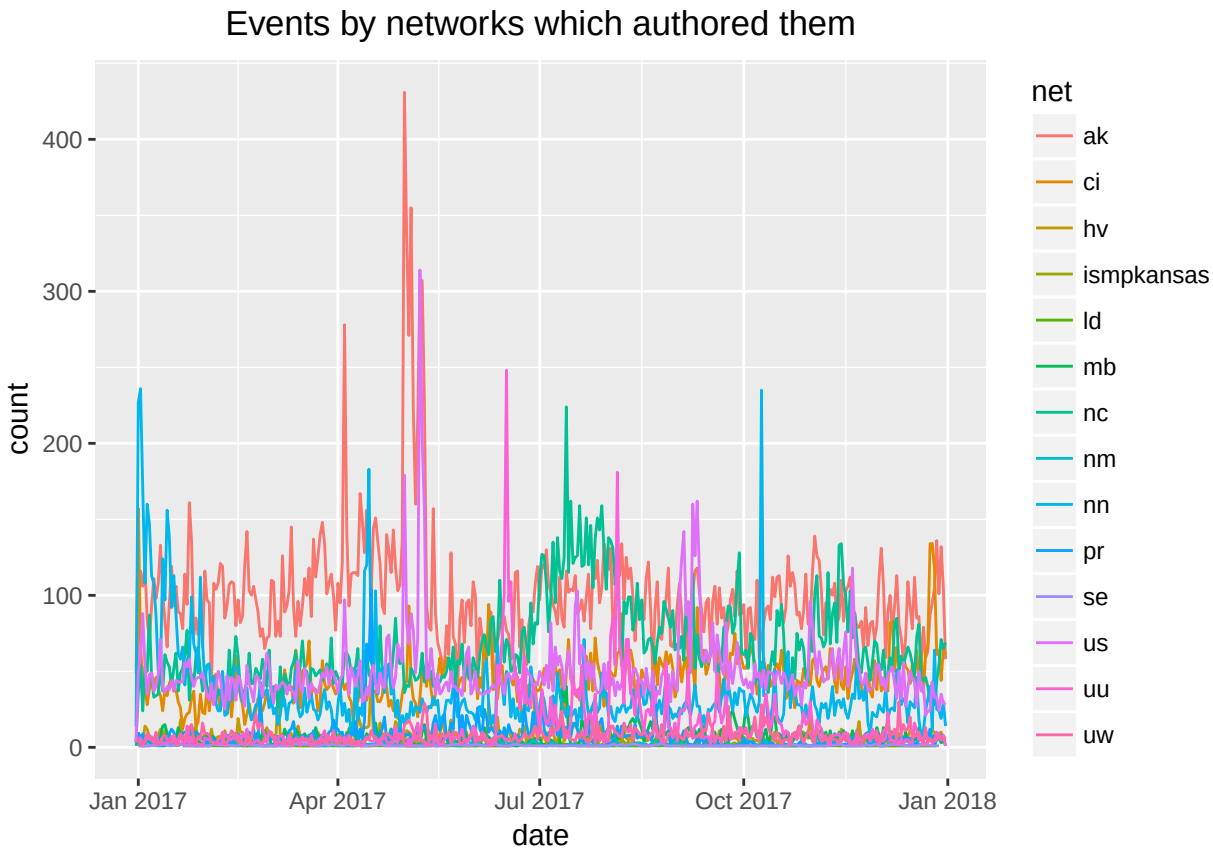
**Now I want to look at data in dynamics**

```
dat$time <- as.POSIXct(dat$time)
by_day <- dat %>% group_by(date = as.Date(time)) %>% summarise(count = n())

ggplot(by_day, aes(x = date, y = count)) +  geom_line() +
  ggtitle("Number of events in dynamics during the year 2017") +
  scale_x_date(name = 'date', date_breaks = '1 month', date_labels = '%b %y') +
  theme(panel.grid.minor = element_blank(), plot.title = element_text(hjust = 0.5))
```

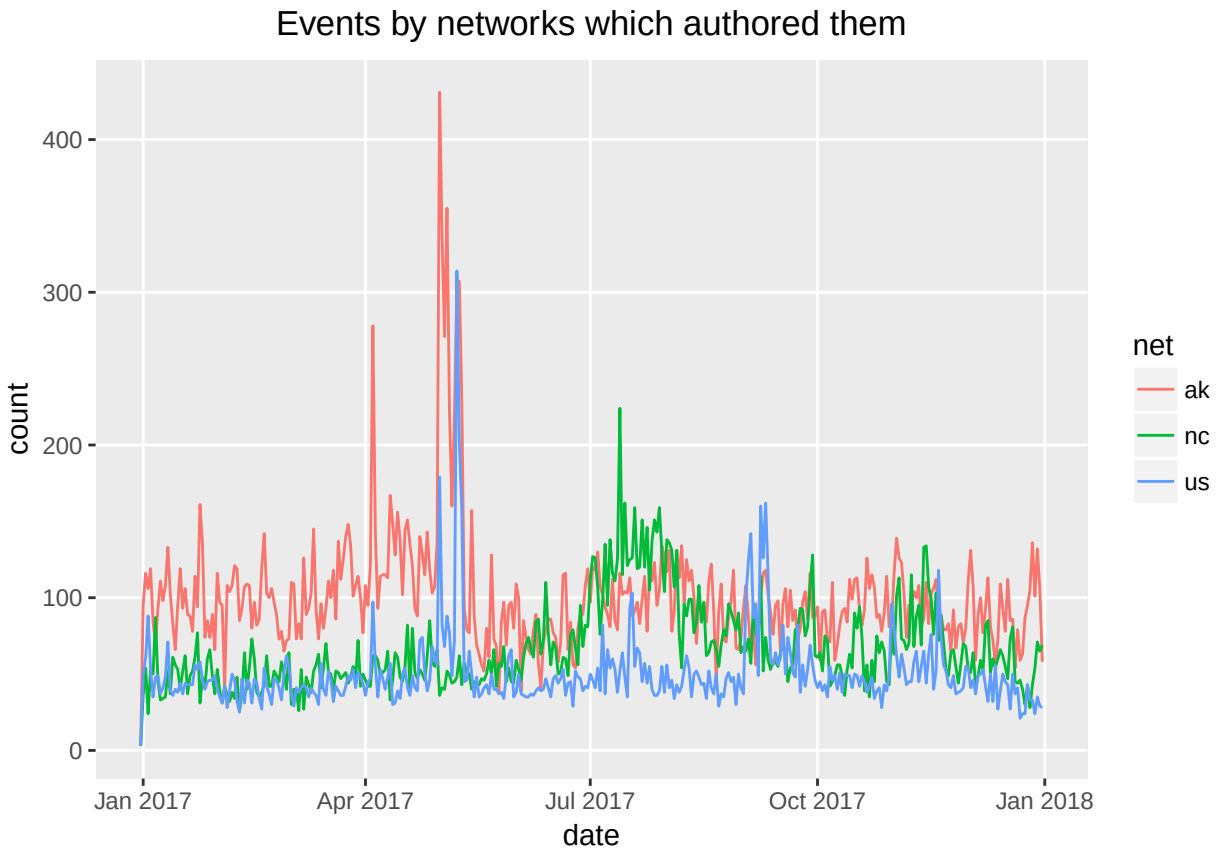



So during the year 2017 the highest number of EQs observed in total on the Earth was on the first half of May. The relatively small number was on Feb-Mar, second half of May, beginning of June, Dec.
It is interesting to look at distribution of EQs on earth om monthly basis. But my computer wasn't able to do that on facets:) That's why let's look at least how the number in distributed by networks who fixed the event.

```
by_day_net<- dat %>% group_by(date = as.Date(time), net) %>% summarise(count = n())
ggplot() +
  #geom_polygon(dat=mdat, aes(long, lat, group=group), fill="grey50") +
  geom_line(data=by_day_net, aes(x=date, y = count, col = net)) +
  ggtitle("Events by networks which authored them") +
  theme(plot.title = element_text(hjust = 0.5))
```



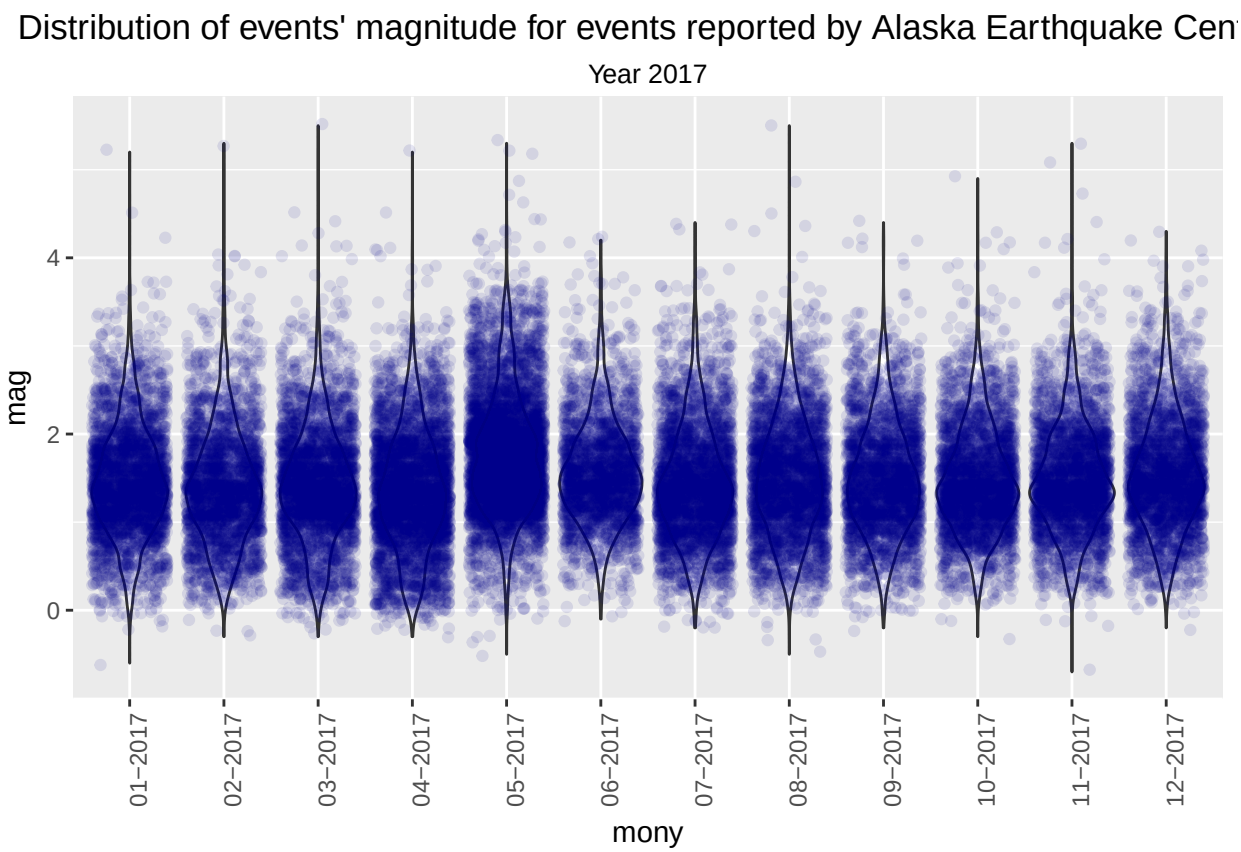

```
by_day_net<- by_day_net %>% filter(net %in% c('ak','us','nc'))
ggplot() +
  #geom_polygon(dat=mdat, aes(long, lat, group=group), fill="grey50") +
  geom_line(data=by_day_net, aes(x=date, y = count, col = net)) +
  ggtitle("Events by networks which authored them") +
  theme(panel.grid.minor = element_blank(), plot.title = element_text(hjust = 0.5))
```



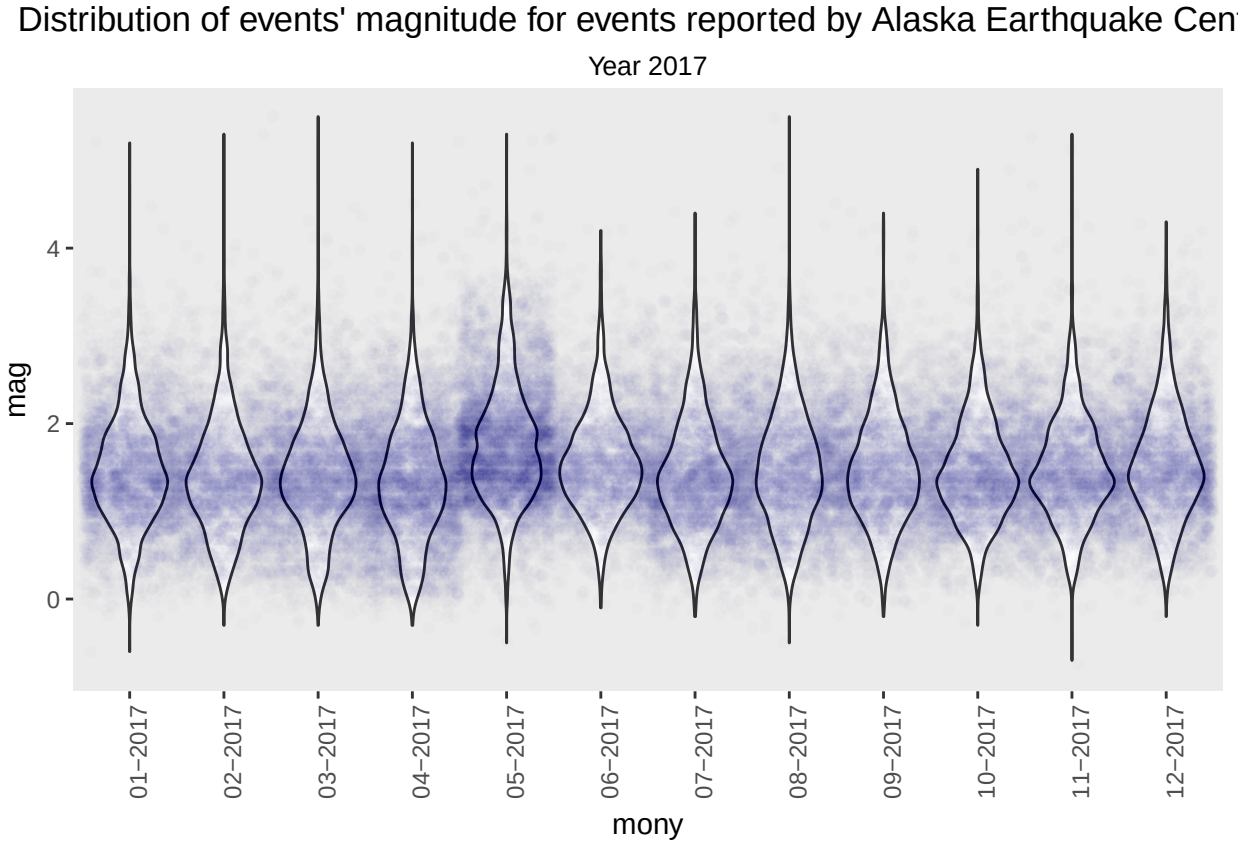

So now we can see that that extreme number of EQs in the first half of may was on Alaska region. The growth of number of EQs in July-August is due to intensive EQs reported by **nc** network which corresponds to Nothern California Seismic System.

Just curious how the magnitude changed on the events fixed by **ak** station (will filter by magSOurce as we are interested exactly in magnitude distribution for now).

```
ak_magSource <- dat %>% filter(magSource == 'ak') %>%
          mutate(mony= paste(format(time, "%m"), year = format(time, "%Y"), sep = "-"))

ggplot(ak_magSource, aes(x = mony, y = mag)) + geom_violin() +
  geom_point( col="darkblue", alpha = 0.1, position = "jitter") +
  ggtitle("Distribution of events' magnitude for events reported by Alaska Earthquake Center", subtitle
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5), axis.text.x =
```




```
ggplot(ak_magSource, aes(x = mony, y = mag)) + geom_violin() +
  #geom_point(binaxis='y', stackdir='center', dotsize=0.1, col="darkblue", alpha = 0.5) +
   geom_jitter( alpha = 0.01, col="darkblue", width = 0.5) +
   ggtitle("Distribution of events' magnitude for events reported by Alaska Earthquake Center", subtitl
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(), plot.title = element_te
```




As we can see the magnitude fixed by **ak** network is more or les similarly distributed on all months across the year 2017 except for May w No obvious specific behaviour on May 2017.