Assignment 2

Exploratory data analysis

(Dataset: Earthquake dataset)

Submitted by,

Ponkothandaraman.S

Submitted to,

Dr.V.Bhuvaneshwari

Assumptions:

- This dataset explains the factors of the earthquake.
- Earthquake in seashore causes tsunami. The attribute that indicates the tsunami is included in the dataset.
- We assume that the type of algorithm used to predict the earthquake affects the accuracy of the estimation.
- We also verify whether the alert level justifies the depth of the earthquake.
- We will be performing EDA to verify these assumptions.

```
library(readr)
library(MASS)
library(dplyr)

library(lattice)
library(ggplot2)
dt <- read_csv("earthquake_data.csv")

## Rows: 782 Columns: 19

summary(dt)
```

```
##     title              magnitude      date_time             cdi        
##  Length:782         Min.   :6.500   Length:782         Min.   :0.000  
##  Class :character   1st Qu.:6.600   Class :character   1st Qu.:0.000  
##  Mode  :character   Median :6.800   Mode  :character   Median :5.000  
##                     Mean   :6.941                      Mean   :4.334  
##                     3rd Qu.:7.100                      3rd Qu.:7.000  
##                     Max.   :9.100                      Max.   :9.000  
##       mmi            alert              tsunami            sig        
##  Min.   :1.000   Length:782         Min.   :0.0000   Min.   : 650.0  
##  1st Qu.:5.000   Class :character   1st Qu.:0.0000   1st Qu.: 691.0  
##  Median :6.000   Mode  :character   Median :0.0000   Median : 754.0  
##  Mean   :5.964                      Mean   :0.3887   Mean   : 870.1  
##  3rd Qu.:7.000                      3rd Qu.:1.0000   3rd Qu.: 909.8  
##  Max.   :9.000                      Max.   :1.0000   Max.   :2910.0  
##      net                nst              dmin             gap        
##  Length:782         Min.   :  0.0   Min.   : 0.000   Min.   :  0.00  
##  Class :character   1st Qu.:  0.0   1st Qu.: 0.000   1st Qu.: 14.62  
##  Mode  :character   Median :140.0   Median : 0.000   Median : 20.00  
##                     Mean   :230.3   Mean   : 1.326   Mean   : 25.04  
##                     3rd Qu.:445.0   3rd Qu.: 1.863   3rd Qu.: 30.00  
##                     Max.   :934.0   Max.   :17.654   Max.   :239.00  
##    magType              depth           latitude         longitude      
##  Length:782         Min.   :  2.70   Min.   :-61.848   Min.   :-179.97  
##  Class :character   1st Qu.: 14.00   1st Qu.:-14.596   1st Qu.: -71.67  
##  Mode  :character   Median : 26.30   Median : -2.572   Median : 109.43  
##                     Mean   : 75.88   Mean   :  3.538   Mean   :  52.61  
##                     3rd Qu.: 49.75   3rd Qu.: 24.654   3rd Qu.: 148.94  
##                     Max.   :670.81   Max.   : 71.631   Max.   : 179.66  
##    location          continent           country         
##  Length:782         Length:782         Length:782         
##  Class :character   Class :character   Class :character   
##  Mode  :character   Mode  :character   Mode  :character   
##                                                           
##
```

```
str(dt)
```

```
## spc_tbl_ [782 × 19] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ title    : chr [1:782] "M 7.0 - 18 km SW of Malango, Solomon Islands"
"M 6.9 - 204 km SW of Bengkulu, Indonesia" "M 7.0 -" "M 7.3 - 205 km ESE of
Neiafu, Tonga" ...
##  $ magnitude: num [1:782] 7 6.9 7 7.3 6.6 7 6.8 6.7 6.8 7.6 ...
##  $ date_time: chr [1:782] "22-11-2022 02:03" "18-11-2022 13:37" "12-11-
2022 07:09" "11-11-2022 10:48" ...
##  $ cdi      : num [1:782] 8 4 3 5 0 4 1 7 8 9 ...
##  $ mmi      : num [1:782] 7 4 3 5 2 3 3 6 7 8 ...
##  $ alert    : chr [1:782] "green" "green" "green" "green" ...
##  $ tsunami  : num [1:782] 1 0 1 1 1 1 1 1 1 1 ...
##  $ sig      : num [1:782] 768 735 755 833 670 ...
##  $ net      : chr [1:782] "us" "us" "us" "us" ...
##  $ nst      : num [1:782] 117 99 147 149 131 142 136 145 175 271 ...
##  $ dmin     : num [1:782] 0.509 2.229 3.125 1.865 4.998 ...
##  $ gap      : num [1:782] 17 34 18 21 27 26 22 37 92 69 ...
##  $ magType  : chr [1:782] "mww" "mww" "mww" "mww" ...
##  $ depth    : num [1:782] 14 25 579 37 624 ...
##  $ latitude : num [1:782] -9.8 -4.96 -20.05 -19.29 -25.59 ...
##  $ longitude: num [1:782] 160 101 -178 -172 178 ...
##  $ location : chr [1:782] "Malango, Solomon Islands" "Bengkulu, Indonesia"
NA "Neiafu, Tonga" ...
##  $ continent: chr [1:782] "Oceania" NA "Oceania" NA ...
##  $ country  : chr [1:782] "Solomon Islands" NA "Fiji" NA ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   title = col_character(),
##   ..   magnitude = col_double(),
##   ..   date_time = col_character(),
##   ..   cdi = col_double(),
##   ..   mmi = col_double(),
##   ..   alert = col_character(),
##   ..   tsunami = col_double(),
##   ..   sig = col_double(),
##   ..   net = col_character(),
##   ..   nst = col_double(),
##   ..   dmin = col_double(),
##   ..   gap = col_double(),
##   ..   magType = col_character(),
##   ..   depth = col_double(),
##   ..   latitude = col_double(),
##   ..   longitude = col_double(),
##   ..   location = col_character(),
##   ..   continent = col_character(),
##   ..   country = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
#Removing two columns
dt = subset(dt, select = -c(latitude,longitude,title,date_time))
dt = subset(dt, select = -c(country,continent))
dt
```

```
## # A tibble: 782 × 13
##    magni…¹   cdi   mmi alert tsunami   sig net     nst  dmin   gap magType
depth
##      <dbl> <dbl> <dbl> <chr>   <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <chr>
<dbl>
##  1     7       8     7 green       1   768 us      117 0.509    17 mww
14
##  2     6.9     4     4 green       0   735 us       99 2.23     34 mww
25
##  3     7       3     3 green       1   755 us      147 3.12     18 mww
579
##  4     7.3     5     5 green       1   833 us      149 1.86     21 mww
37
##  5     6.6     0     2 green       1   670 us      131 5.00     27 mww
624.
##  6     7       4     3 green       1   755 us      142 4.58     26 mwb
660
##  7     6.8     1     3 green       1   711 us      136 4.68     22 mww
630.
##  8     6.7     7     6 green       1   797 us      145 1.15     37 mww
20
##  9     6.8     8     7 yell…       1  1179 us      175 2.14     92 mww
20
## 10     7.6     9     8 yell…       1  1799 us      271 1.15     69 mww
26.9
## # … with 772 more rows, 1 more variable: location <chr>, and abbreviated
## #   variable name ¹magnitude
```

```r
#Checking null values
colSums(is.na(dt))
```

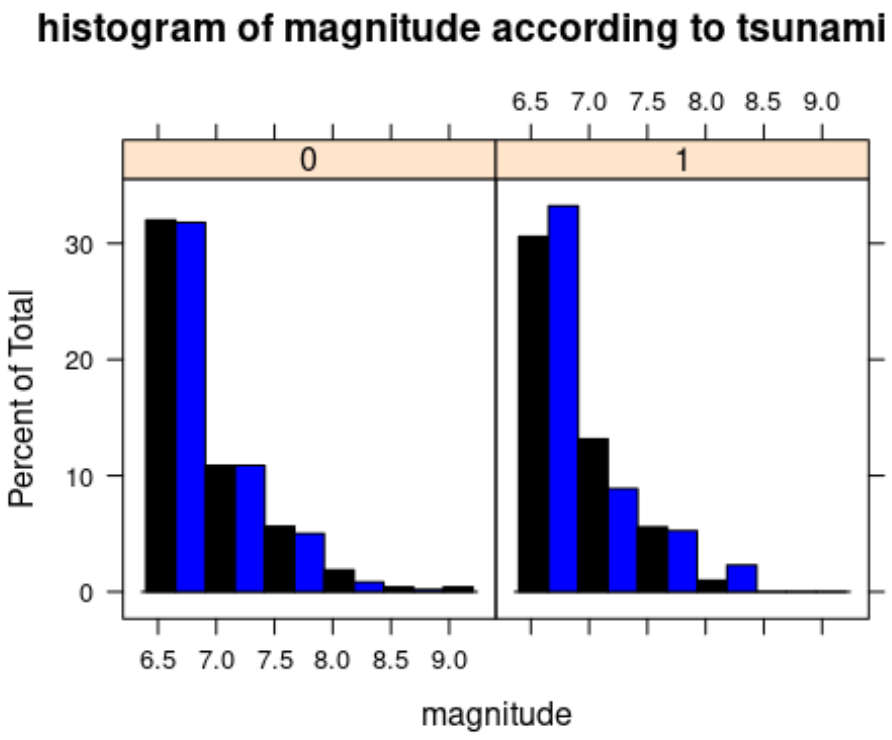
```
## magnitude       cdi       mmi     alert   tsunami       sig       net
nst
##         0         0         0       367         0         0         0
0
##      dmin       gap   magType     depth  location
##         0         0         0         0         5
```

```r
#Changing the specific variables to categorical
dt$alert=as.factor(dt$alert)
dt$tsunami=as.factor(dt$tsunami)
dt$magType=as.factor(dt$magType)
dt$net=as.factor(dt$net)
```

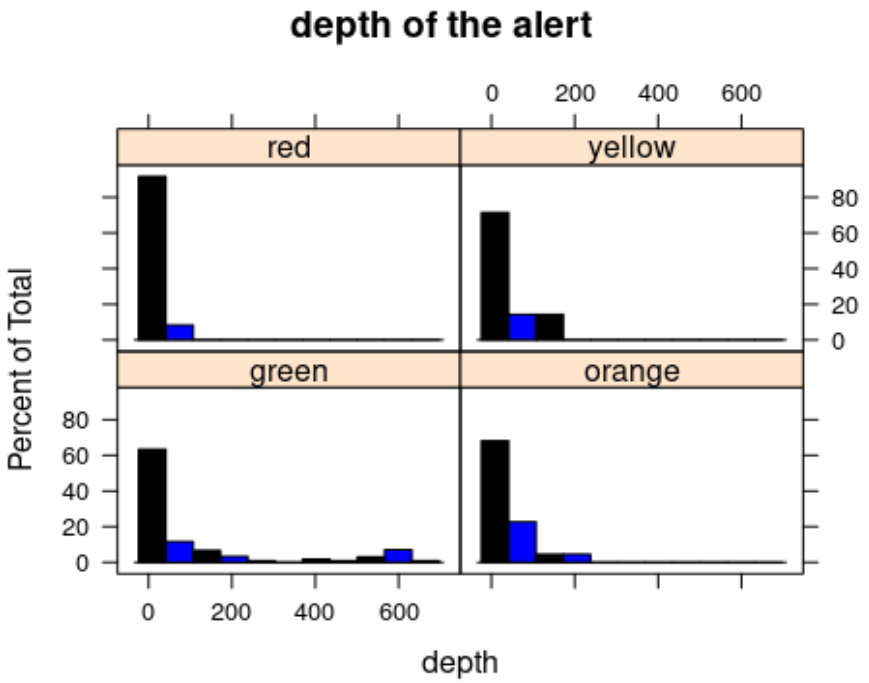
```
str(dt)
```

```
## tibble [782 × 13] (S3: tbl_df/tbl/data.frame)
##  $ magnitude: num [1:782] 7 6.9 7 7.3 6.6 7 6.8 6.7 6.8 7.6 ...
##  $ cdi      : num [1:782] 8 4 3 5 0 4 1 7 8 9 ...
##  $ mmi      : num [1:782] 7 4 3 5 2 3 3 6 7 8 ...
##  $ alert    : Factor w/ 4 levels "green","orange",..: 1 1 1 1 1 1 1 1 4 4
...
##  $ tsunami  : Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 2 2 ...
##  $ sig      : num [1:782] 768 735 755 833 670 ...
##  $ net      : Factor w/ 11 levels "ak","at","ci",..: 10 10 10 10 10 10 10
10 10 10 ...
##  $ nst      : num [1:782] 117 99 147 149 131 142 136 145 175 271 ...
##  $ dmin     : num [1:782] 0.509 2.229 3.125 1.865 4.998 ...
##  $ gap      : num [1:782] 17 34 18 21 27 26 22 37 92 69 ...
##  $ magType  : Factor w/ 9 levels "mb","md","Mi",..: 9 9 9 9 9 7 9 9 9 9
...
##  $ depth    : num [1:782] 14 25 579 37 624 ...
##  $ location : chr [1:782] "Malango, Solomon Islands" "Bengkulu, Indonesia"
NA "Neiafu, Tonga" ...
```

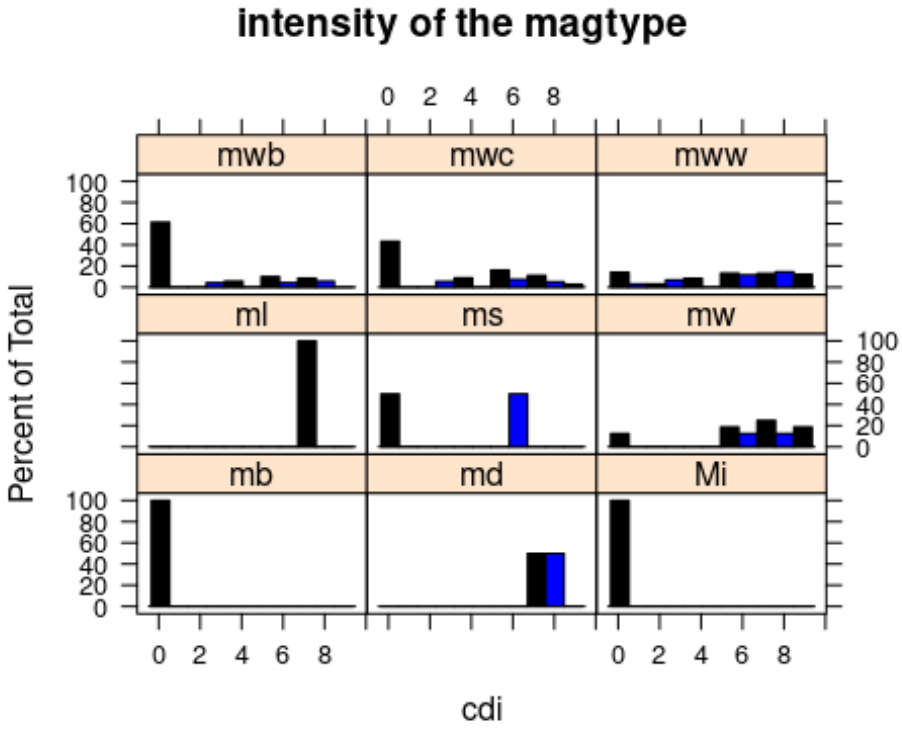
```
#histogram
histogram(~magnitude|tsunami,data=dt,col=c("black","blue"),main="histogram of
magnitude according to tsunami")
```



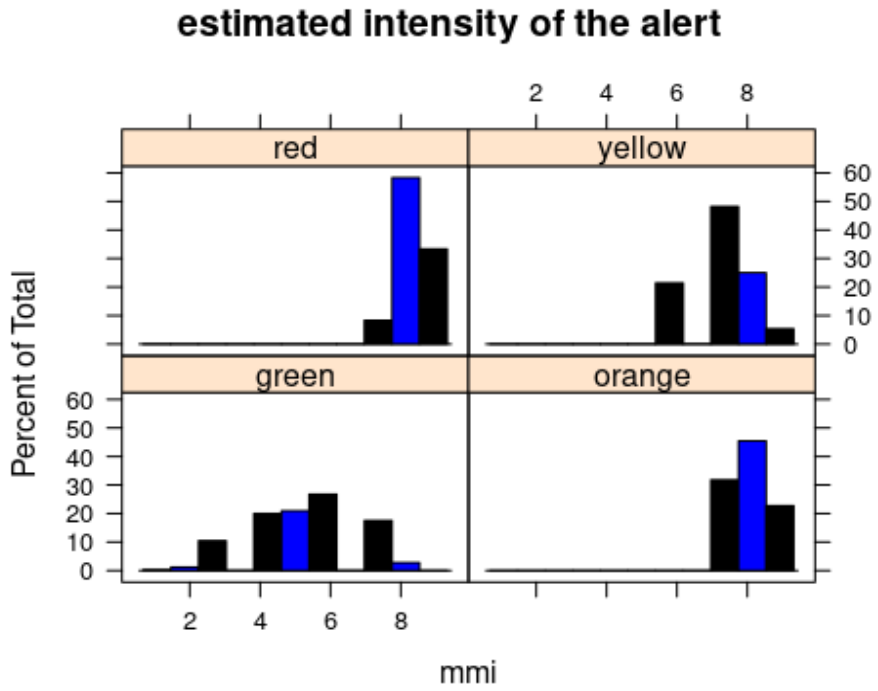
```
histogram(~depth|alert,data=dt,col=c("black","blue"),main="depth of the
alert")
```



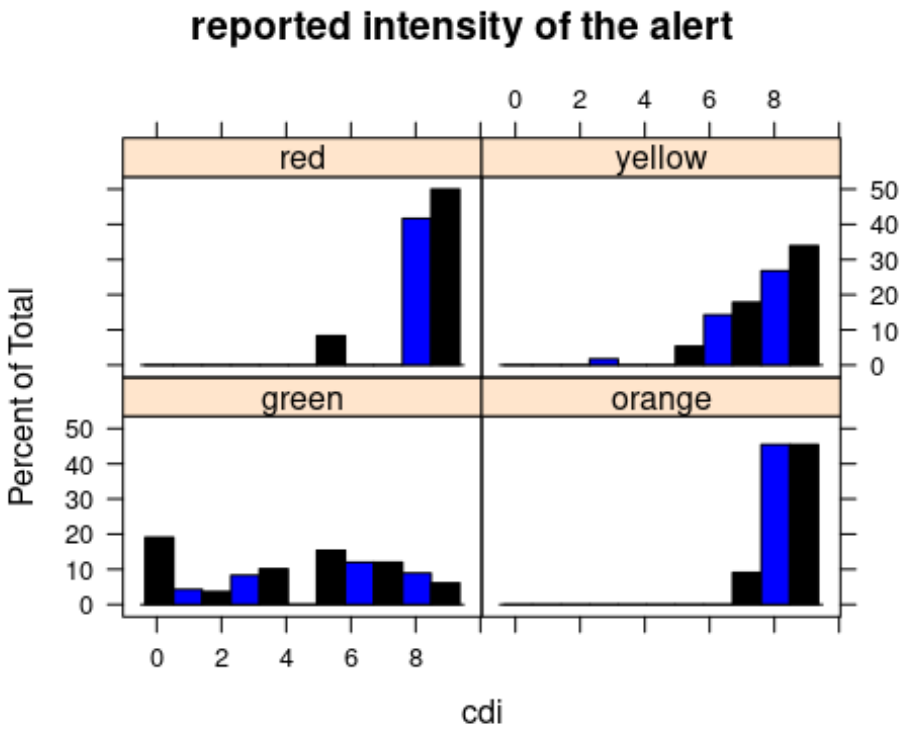

```
histogram(~cdi|magType,data=dt,col=c("black","blue"),main="intensity of the
magtype")
```



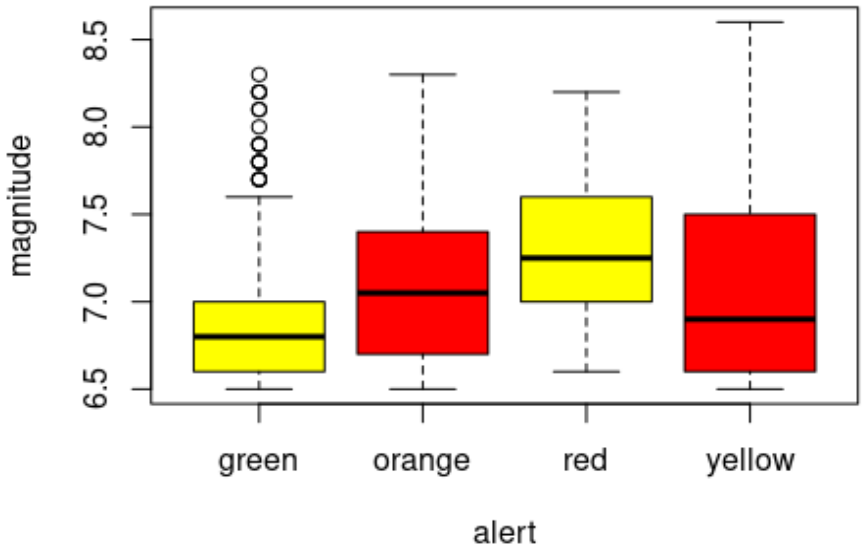
```
histogram(~mmi|alert,data=dt,col=c("black","blue"),main="estimated intensity
of the alert")
```



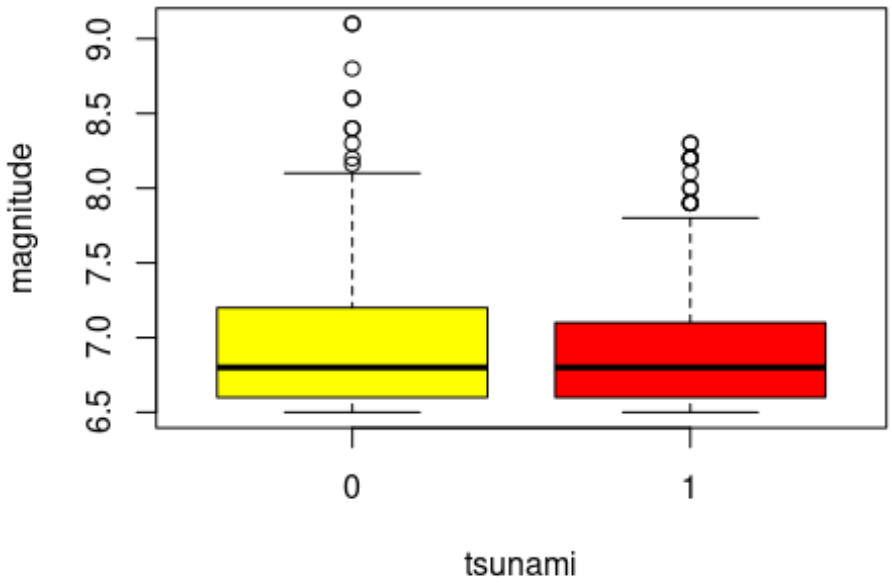

```
histogram(~cdi|alert,data=dt,col=c("black","blue"),main="reported intensity
of the alert")
```



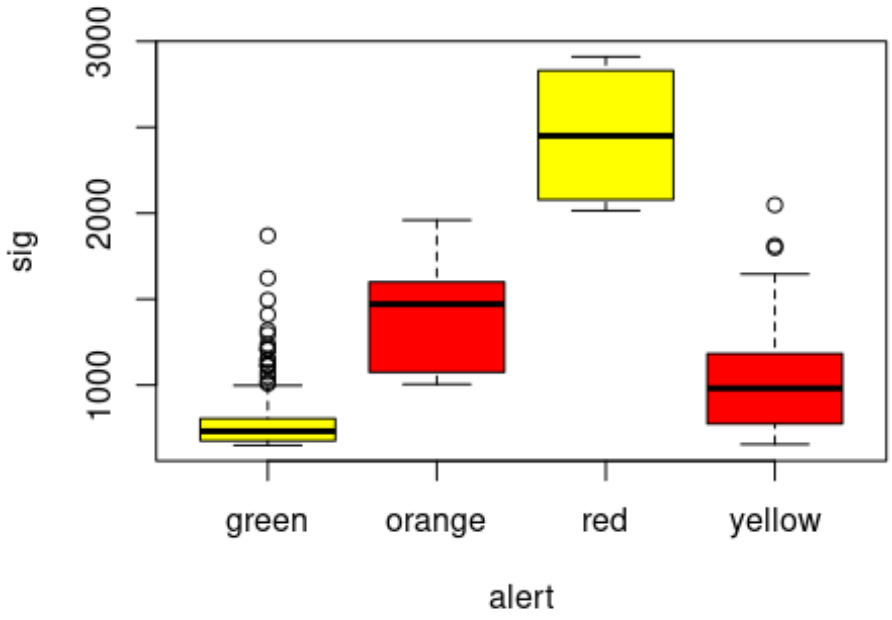
```r
#boxplot
boxplot(magnitude~alert,data=dt,col=c("yellow","red"))
```



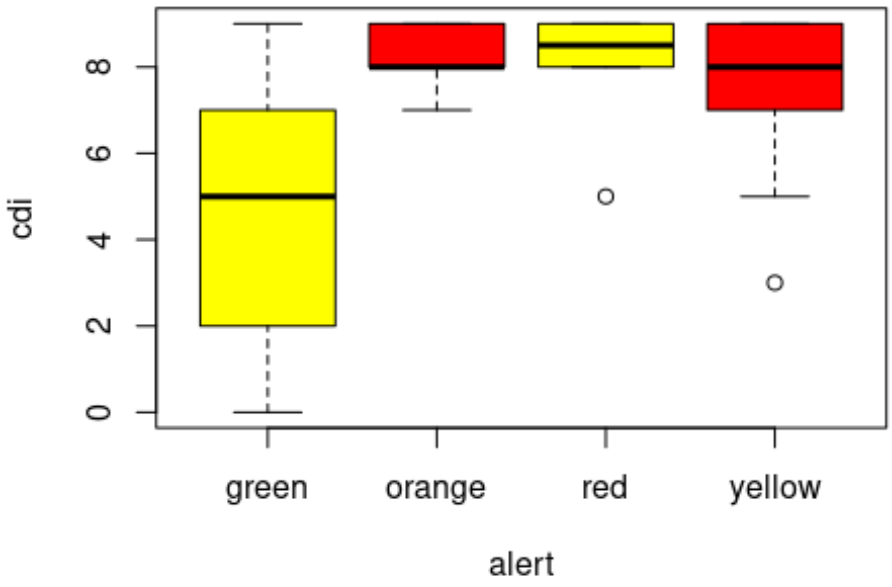

```r
boxplot(magnitude~tsunami,data=dt,col=c("yellow","red"))
```



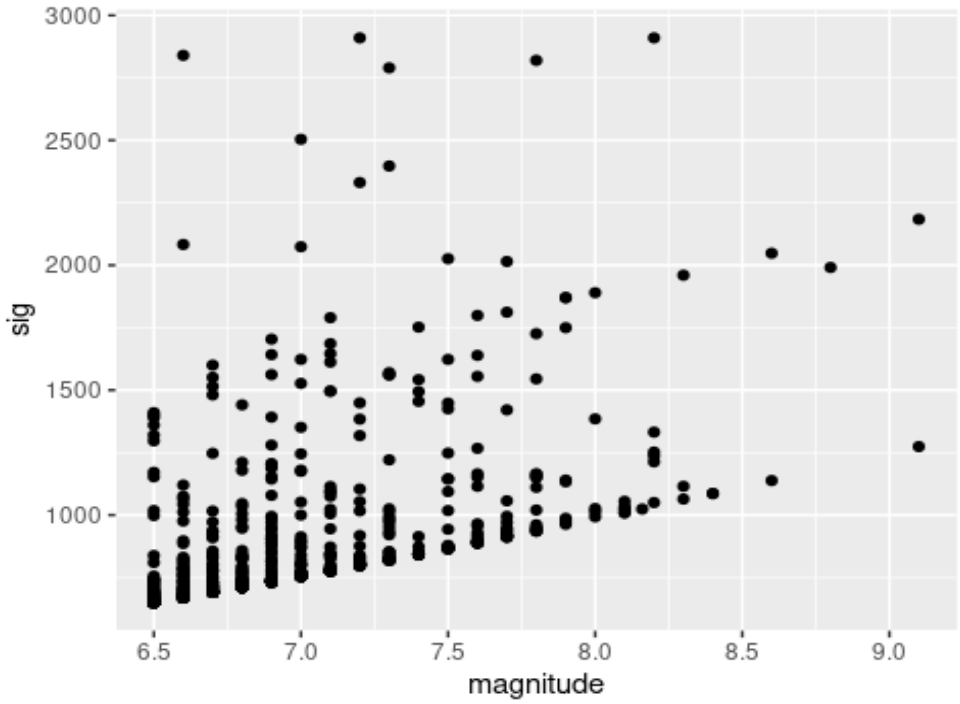
```
boxplot(sig~alert,data=dt,col=c("yellow","red"))
```



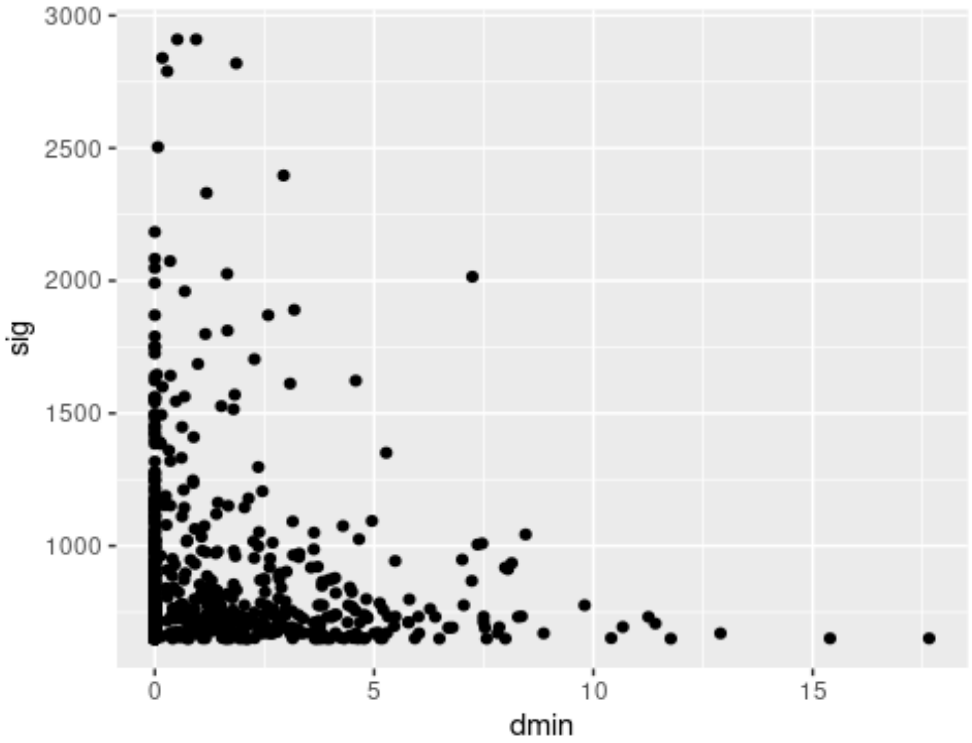

```
boxplot(cdi~alert,data=dt,col=c("yellow","red"))
```



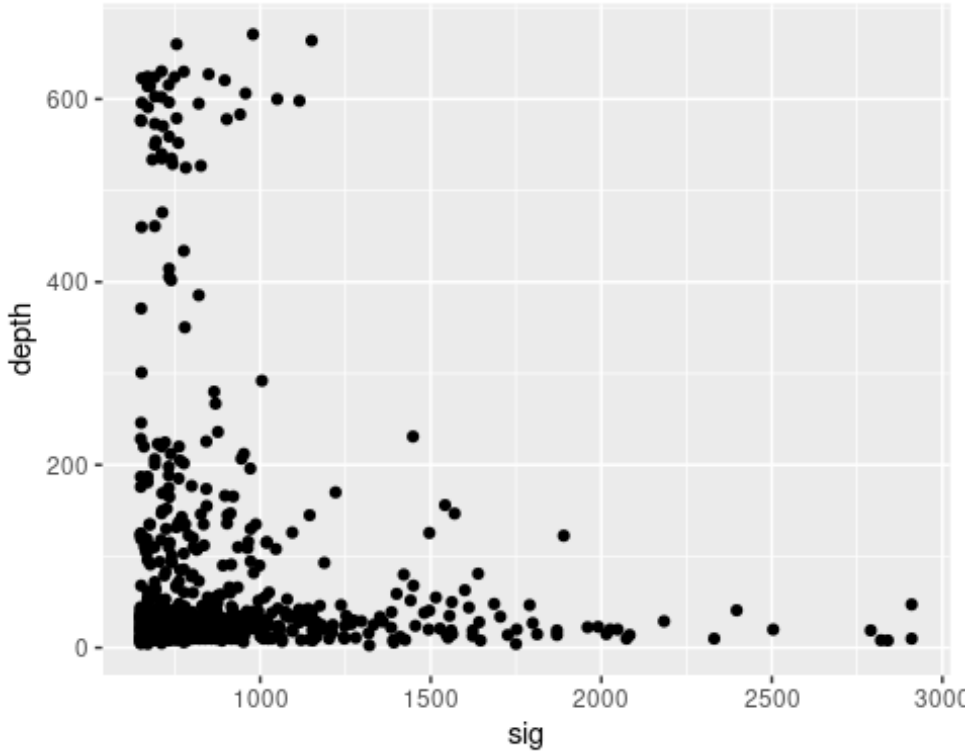
```
#scatterplot
ggplot(dt, aes(x=magnitude, y=sig)) + geom_point()
```



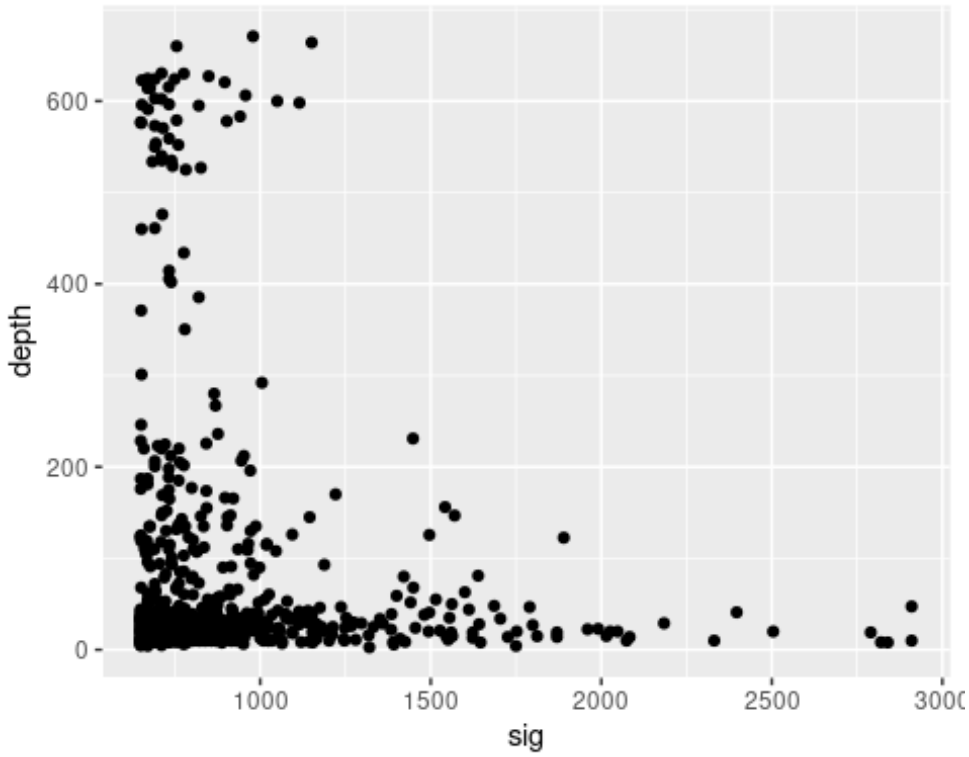

```
ggplot(dt, aes(x=dmin, y=sig)) + geom_point()
```



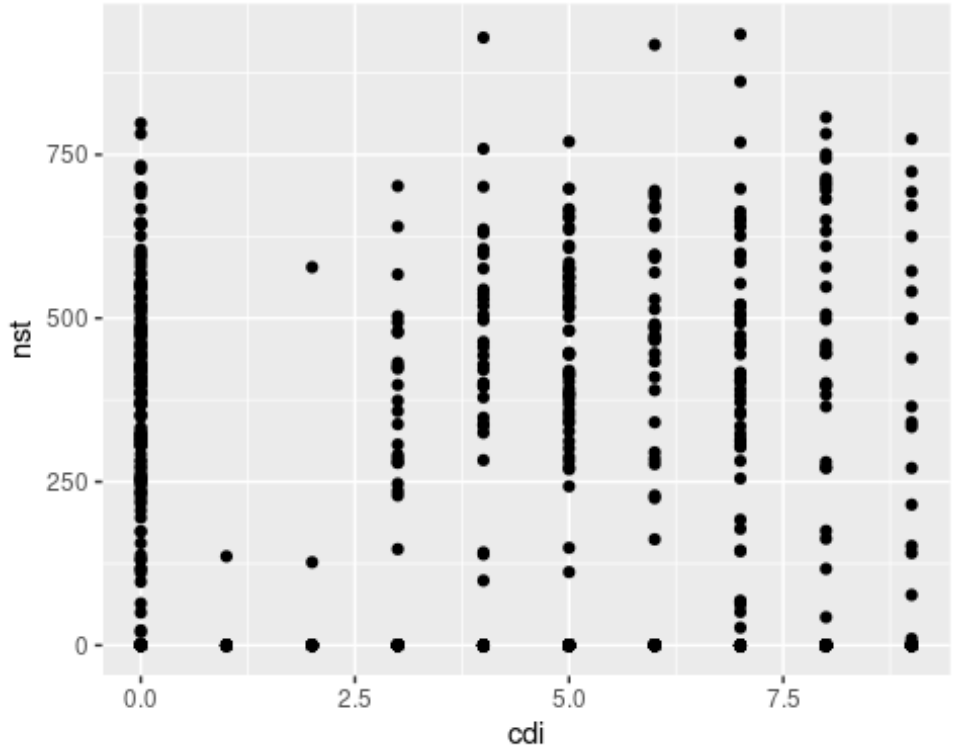
```
ggplot(dt, aes(x=sig, y=depth)) + geom_point()
```



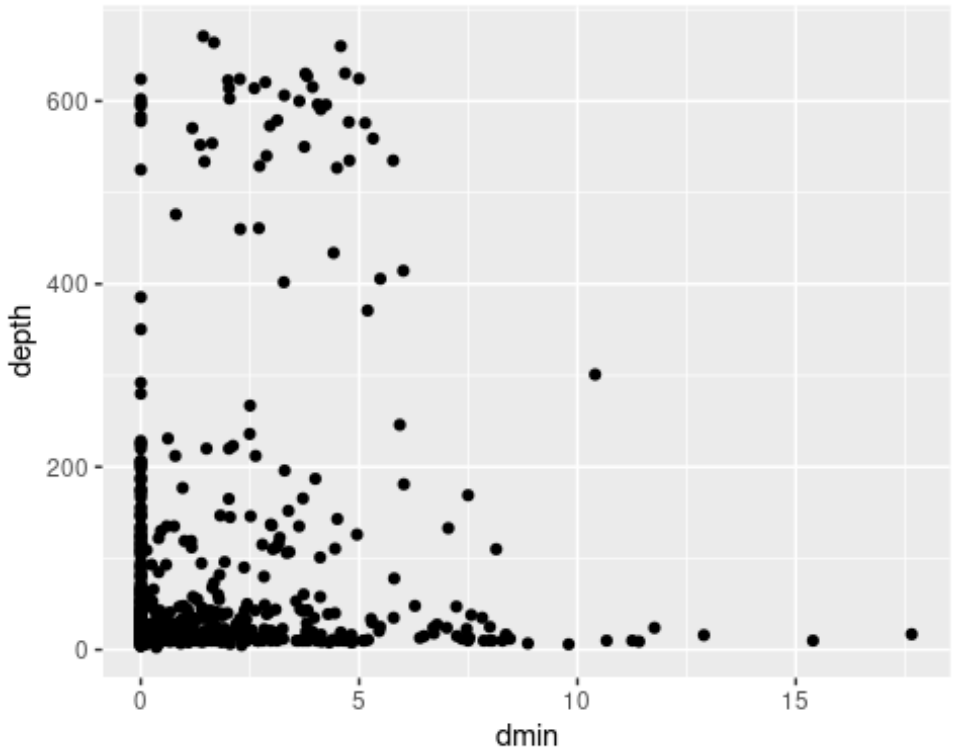

```
ggplot(dt, aes(x=sig, y=depth)) + geom_point()
```



```
ggplot(dt, aes(x=cdi, y=nst)) + geom_point()
```




```
ggplot(dt, aes(x=dmin, y=depth)) + geom_point()
```



INFERENCE:

Histogram:

- The magnitude of the tsunami ranges from 6.5-8.0 but the earthquake reported on land has been extended till 6.5-9.0.
- Red alert has the range of 0-50 depth but green alert has the highest range of 0-600 depth
- Yellow and Orange alert has moderate range of 0-200.
- The intensity in mww, mwc and mwb, the range has been spreaded from 0-8.
- The green alert has the intensity of 0-8 , and the other alert has the range of 5-9 intensity.

Boxplot:

- Green and orange alert has maximum and minimum values and the median lies in the center so it has no skewness. Green alert has outliers.
- In yellow and red alert the median lies nearer to the Q1 so it is positively skewed.
- In tsunami 0 and 1 there are maximum and minimum values , the median lies near Q1 so it is positively skewed. It has outliers.
- Comparing sig and alert it has maximum and minimum values, the Green alret has more outliers.
- In Orange alert , the median lies near Q3 so it is negatively skewed and in other the median lies in the center.
- By Comparing the intensity and the alert ,only green has both maximum and minimum values the median lies in near Q3 so it is negatively skewed ,orange and yellow has only minimum values and red has no maximum and minimum values the median lies in the center it has no skewness.

Scatterplot:

- X axis is plotted as the magnitude and Y axis is plotted as the sig.
- No Correlation because the points are scattered all over the plot so it is difficult to conclude whether it is increasing or decreasing.

INSIGHTS:

- The magnitude of the tsunami is slightly higher than the earthquake reported on land.
- There were earthquakes with very high magnitude but there is no tsunami which have magnitude higher than 8.5
- Surprisingly earthquakes with high depth are reported in 'Green' alert.
- Earthquakes happened on red alert are seem to have low depth.
- The estimated intensity of red is the highest among the alerts but also yellow and orange have slightly same estimated intensity levels.
- The reported intensity is alike for Red and Orange alert.
- Yellow alert have earthquake events with intensity lower than 3.While green alert have events with intensity all along 0-8.
- The green alert have many outliers of magnitude. we've already mentioned that green have reported with high magnitude that indicates the reason for these outliers.
- Yellow and Orange alert have high magnitude.
- There are earthquakes with high magnitude than tsunami. And earthquakes have many outliers compared to tsunami.
- Green alerts are supposed to have low significance but they have many high significant outliers.
- Red alert have the highest significance among the alerts.