Data visualization of built-in earthquake data `quakes`

Monika Baloda (mbalo005@ucr.edu)

Contents

This Rmarkdown does the followings:

1. Visualize using following plots: i)density histogram ii) box-plot iii) empirical cdf iv) Q-Q plot
2. Outlier detection
3. Writing code to calculate Minimum variance unbiased estimates for Normal distribution.
4. Visualize your model fitness by plotting standard normal curve over sample.
5. Multiple scatterplots in one plot in 2x2 layout.
6. Single scatterplot
7. Plotting on Map

- This work uses **R Base Graphics** to generate the figures, not `ggplot2`.

0. Load necessary packages

```
library(datasets) # for the `quakes` data set

## You need to install the maps package first, then you can comment out the following line.
# install.packages("maps")

library(maps) # for map visualization
## Warning: package 'maps' was built under R version 4.2.3
```

The `quakes` data set give the locations of 1000 seismic events of MB > 4.0. The events occurred in a cube near Fiji since 1964.

```
library(datasets) # we have already loaded the package but we can do it again
?quakes
```

```
## starting httpd help server ... done
```

Let's get a quick glimpse of the dataset.

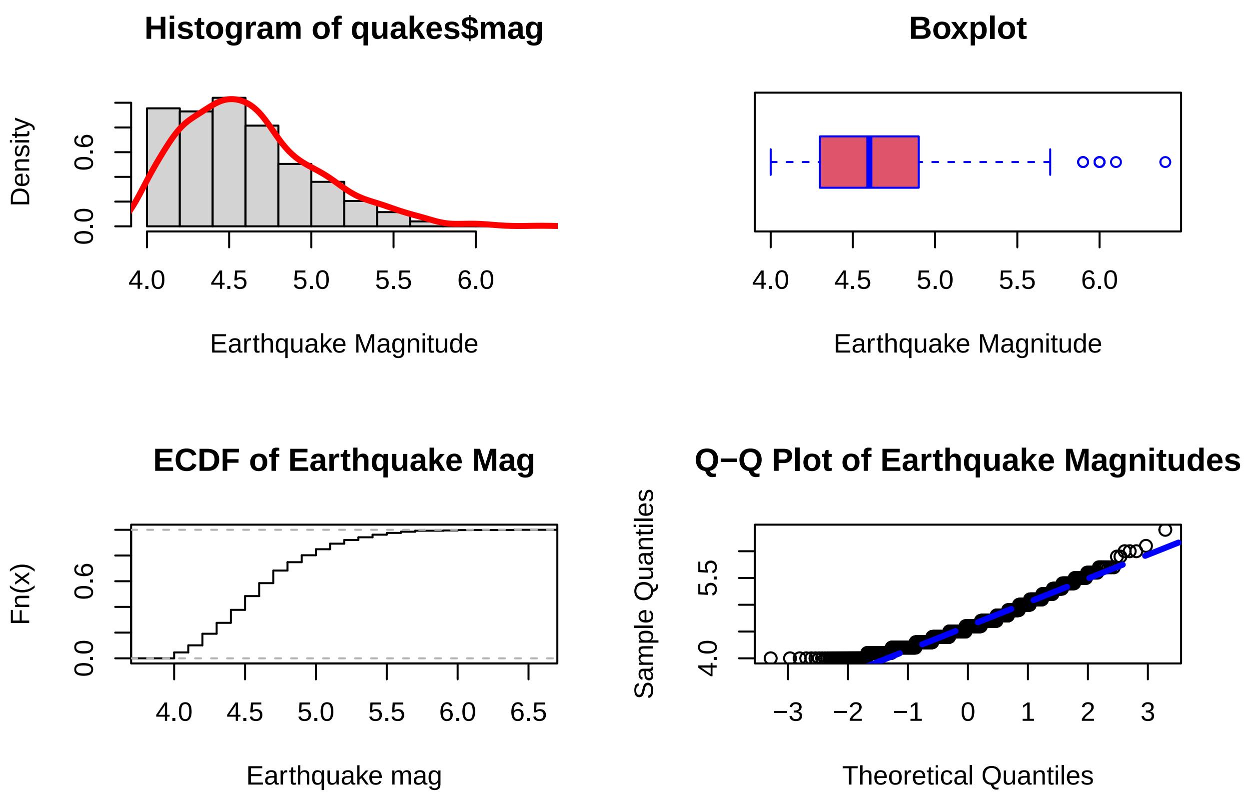
```
class(quakes)
## [1] "data.frame"
head(quakes, n=5) #print first 5 rows of quakes
##      lat   long depth mag stations
## 1 -20.42 181.62   562 4.8       41
## 2 -20.62 181.03   650 4.2       15
## 3 -26.00 184.10    42 5.4       43
## 4 -17.97 181.66   626 4.1       19
## 5 -20.42 181.96   649 4.0       11
dim(quakes)  # dimension of the table
## [1] 1000    5
names(quakes)  # list the variables in quakes
## [1] "lat"      "long"     "depth"    "mag"      "stations"
str(quakes)  # list the structures in quakes
## 'data.frame':    1000 obs. of  5 variables:
##  $ lat     : num  -20.4 -20.6 -26 -18 -20.4 ...
##  $ long    : num  182 181 184 182 182 ...
##  $ depth   : int  562 650 42 626 649 195 82 194 211 622 ...
##  $ mag     : num  4.8 4.2 5.4 4.1 4 4 4.8 4.4 4.7 4.3 ...
##  $ stations: int  41 15 43 19 11 12 43 15 35 19 ...
```

1. Visualizations of the earthquake magnitudes

Suppose we are interested in studying the distribution of the magnitude `mag` variable. Following R code makes a single figure with the following four subfigures in a 2-by-2 layout.
i) subfigure #1: plot a density histogram of the earthquake magnitudes, and then plot the estimated probability density curve in red color in the same plot ii) subfigure #2: plot a horizontal box plot of the earthquake magnitudes iii)subfigure #3: plot the empirical cdf of the earthquake magnitudes iv) subfigure #4: make a Q-Q plot to compare the observed earthquake magnitudes distribution with the theoretical Normal distribution. Add a *thick* qqline in blue color.

ANSWER

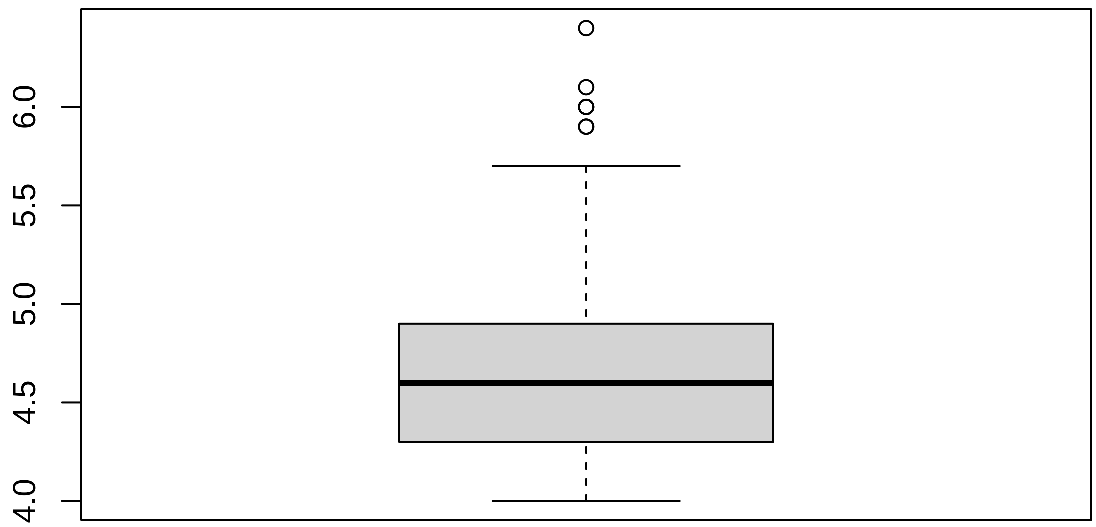
```
# part a)
par(mfrow=c(2,2))
hist(quakes$mag, freq=F, xlab="Earthquake Magnitude") # code for histogram
lines(density(quakes$mag), col="red", lwd=3) # superimposed edf

# part b)
boxplot(quakes$mag, horizontal=T, col=2, border="blue",
xlab="Earthquake Magnitude", main="Boxplot")

# part c)
plot.ecdf(quakes$mag, verticals=T, pch="",
xlab="Earthquake mag",
main="ECDF of Earthquake Mag") # empirical CDF plot

# part d)
qqnorm(quakes$mag, main = "Q-Q Plot of Earthquake Magnitudes")
qqline(quakes$mag, col="blue", lty=5, lwd=3)
```



2. Outlier detection

Following code prints out the indexes of the outliers and then prints out the outlier observations.

```
which(quakes$mag %in% boxplot(quakes$mag)$out)
```



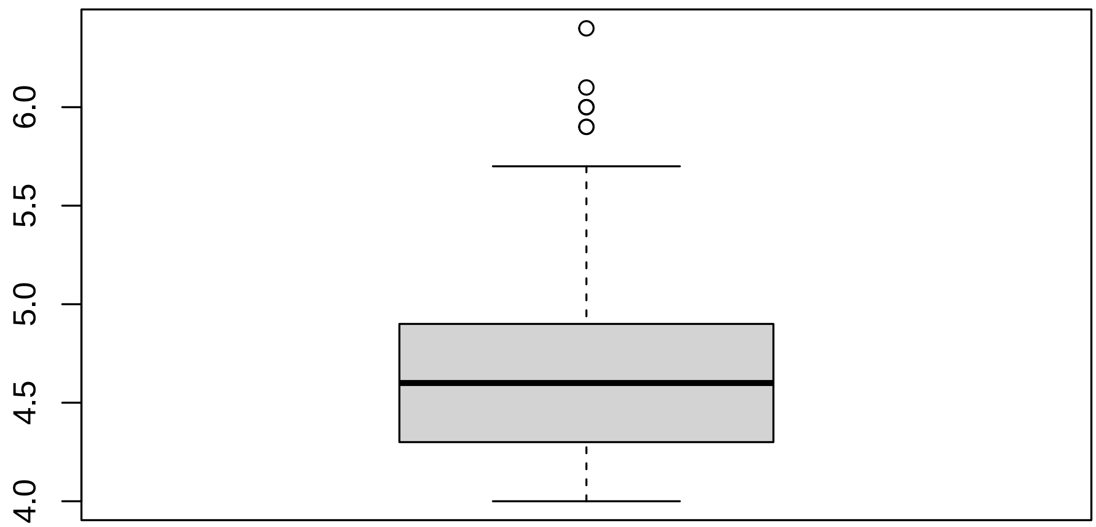

```
## [1]   15   17  152  558  753  870 1000
```

How many outliers have you found?

answer: There are 7 outliers

ANSWER

```r
outliersValue <- boxplot(quakes$mag)$out
```




```r
print(outliersValue)
## [1] 6.1 6.0 6.4 5.9 5.9 6.0 6.0
```

3. Minimum variance unbiased estimates for Normal distribution

Suppose we assume that the earthquake magnitude variable follows a Normal distribution and our 1000 earthquake magnitude observations are independent and identically distributed (iid).

That is, $X_i \sim N(\mu, \sigma^2)$, where X_i is the magnitude of the i-th earthquake observation and $i = 1, 2, \cdots, 1000$.

The **minimum variance unbiased estimators (MVUE)** for μ and σ^2 are:

$$\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \widehat{\sigma^2} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Using existing R functions to calculate the estimates of $\widehat{\mu}$ and $\widehat{\sigma^2}$.

ANSWER

```r
mean(quakes$mag)
## [1] 4.6204
(sd(quakes$mag))^2
## [1] 0.1622261
```

Creating our code to calculate the estimate of minimum variance unbiased estimators (MVUE) i.e. $\widehat{\mu}$ and $\widehat{\sigma^2}$.

ANSWER

```
x <- sum(quakes$mag)*(1/length(quakes$mag))
s2 <- (1/((length(quakes$mag)-1)))*sum(((quakes$mag - x)^2))
```

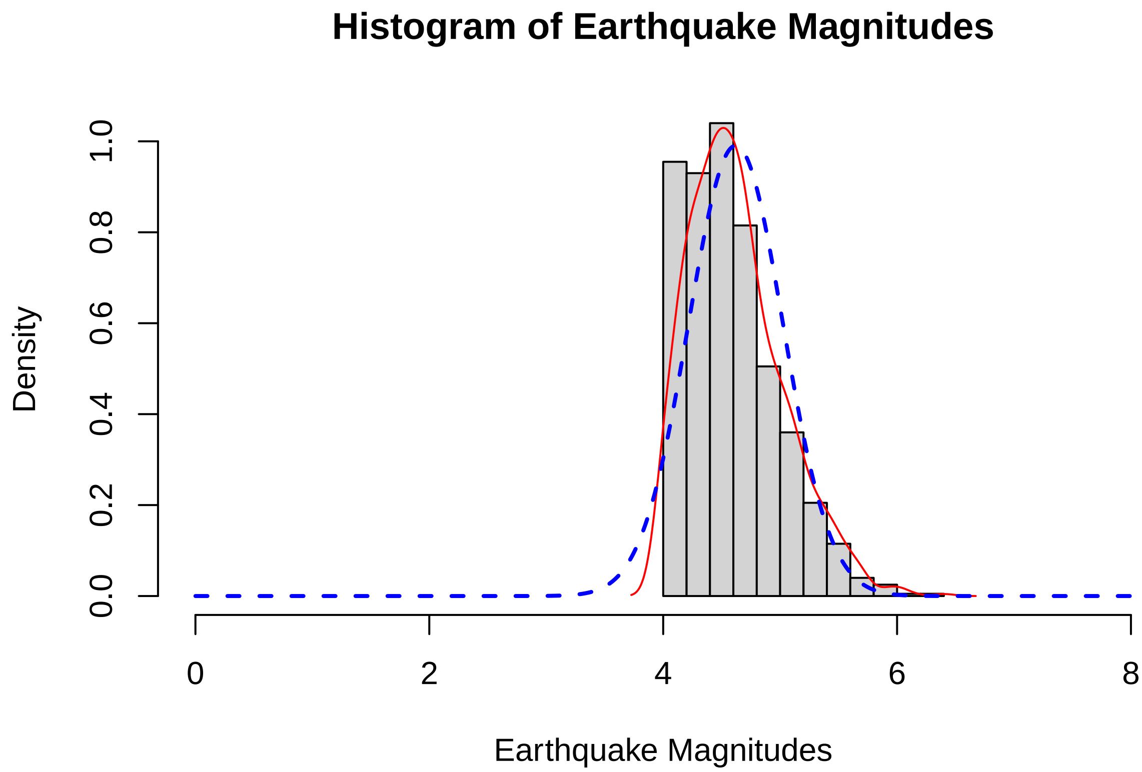
4 Visualize your model fitness

To visualize the model fitness, we can add the estimated Normal distribution curve to the histogram plot you have generated in part 1).

Our code generates the following figure.

- make a density histogram first, set `xlim` from 0 to 8.
- plot the empirical density curve in red color on the same figure.
- plot the estimated Normal distribution curve (that is, $X \sim N(\widehat{\mu}, \widehat{\sigma^2})$) as a blue dashed line on the same figure.

```
hist(quakes$mag, freq = FALSE, main = "Histogram of Earthquake Magnitudes", xlab = "Earthquake Magnitud
lines(density(quakes$mag), col="red")
z <- seq(0, 8, length=1000)
prob.z <- dnorm(z, mean=x, sd = sd(quakes$mag))
lines(z, prob.z, lwd=2, lty=2, col="blue")
```

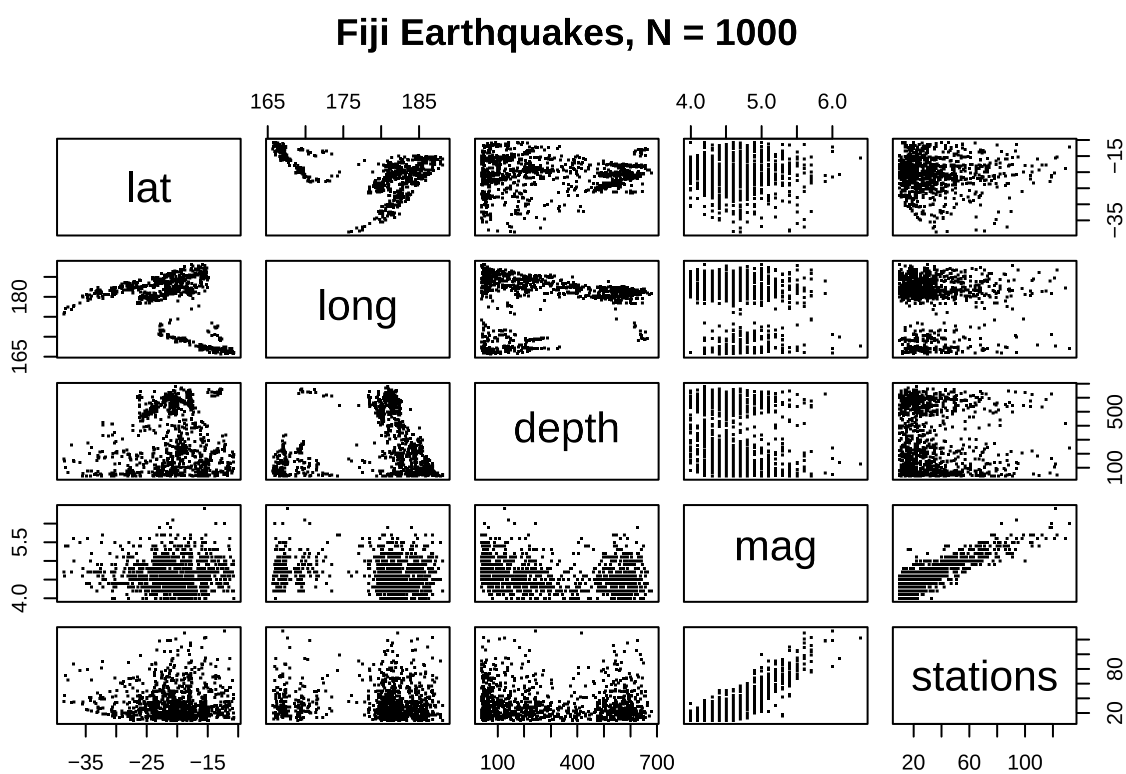


Questions How does your estimated normal curve differ from the empirical density curve? :Do you think the earthquake magnitude observations follow a Normal distribution? Does this result consist with the Q-Q plot you generated in part 1).

ANSWER The red curve and the blue line are quite similar, so the earthquake magnitude observations may follow a Normal distribution and it consist with the Q-Q plot.

5. Pairwise scatterplots This figure tells us about the relations of every two variables in the dataset

```
# require(graphics) # load the graphics library
library(graphics)
pairs(quakes, main = "Fiji Earthquakes, N = 1000", cex.main=1.2, pch=".")
```



Question: Do you think the `mag` variable and `stations` variables are positively correlated? Explain your answer. *Answer*: Yes, I think they are positively correlated because in the figures in the right bottom, the plots scatters in a left-bottom and top-right pattern.

6. Single scatterplot

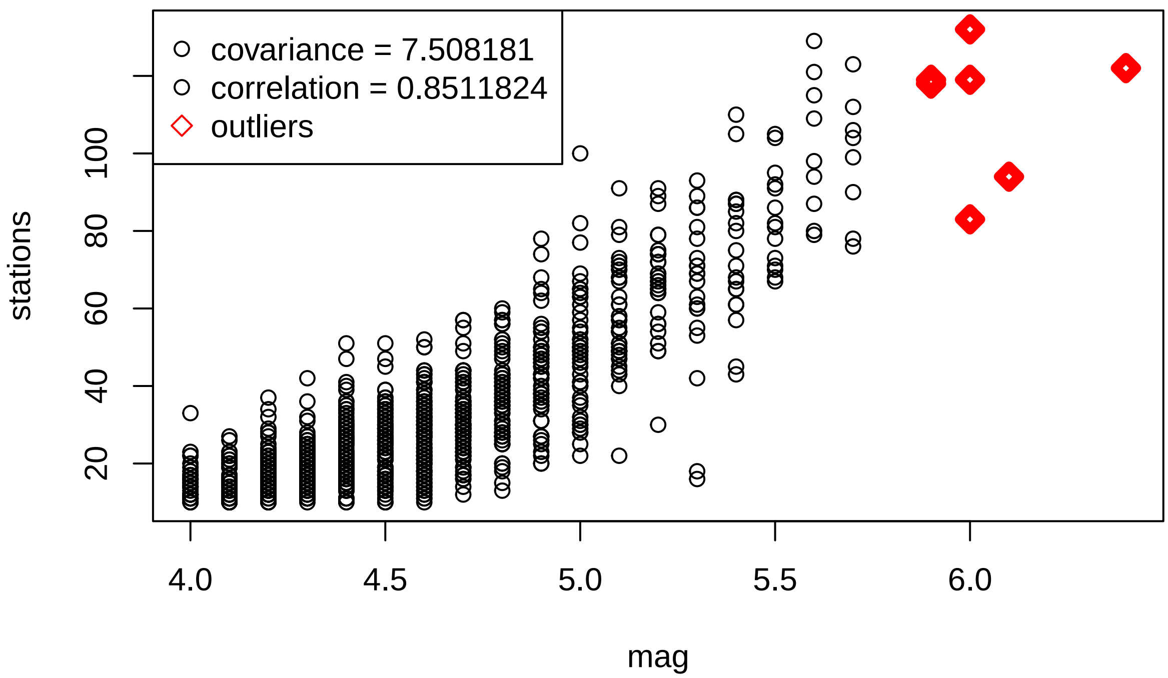
Following R code reproduces the scatterplot for the `mag` and `stations` variables.

- calculate the covariance and correlation coefficient.
- redraw the outlier points you identified in part (2) using red color filled diamond symbol.
- add a three-line legend to your plot. The first legend line reports the covariance value; the second legend line reports the correlation coefficient value; and the third legend line indicates the red diamonds are likely outliers.

ANSWER

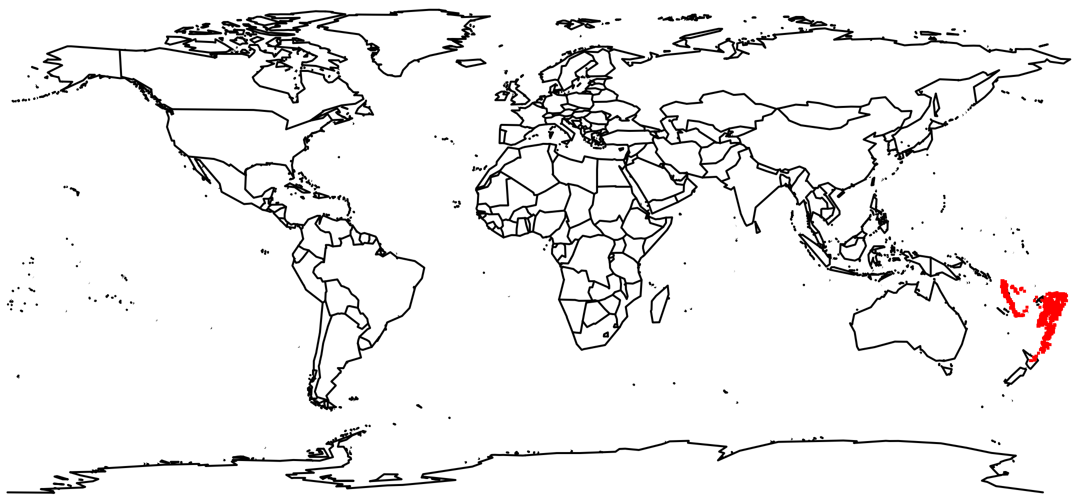
```
attach(quakes)
cor(mag, stations)
## [1] 0.8511824
cov(mag, stations)
## [1] 7.508181
```

```
plot(mag, stations)
points(mag[mag > 5.8], stations[mag > 5.8], col="red", pch=5, lwd=5)
legend("topleft", legend=c("covariance = 7.508181","correlation = 0.8511824", "outliers"), col=c("black
```



7. Plotting Earthquakes using maps We can plot our earthquake records on a world map using the `maps` package. Look at the following map. Where is Fiji?

```
library(maps)
map()
points(quakes$long, quakes$lat, pch=".", col="red")
```



We can also zoom in and make local map of Fiji and its neighborhood area.

```
long.mean <- mean(quakes$long)
lat.mean <- mean(quakes$lat)
#orient <- c(lat.mean, long.mean, 0)
xlim <- c(min(quakes$long)/2, max(quakes$long)*1.5)
ylim <- c(min(quakes$lat)-10, max(quakes$lat)+10)
map(database="world", xlim=xlim, ylim=ylim, col="grey80", fill=T)
```

Add our earthquake locations into the above plot as red color dots. Used blue color filled rectangles to mark locations of the outliers identified in part (2).

```
map(database="world", xlim=xlim, ylim=ylim, col="grey80", fill=T)
points(quakes$long, quakes$lat, pch=".", col="red")
points(quakes$long[quakes$mag > 5.8], quakes$lat[quakes$mag > 5.8], col="blue", pch=15, lwd=5)
```