STAT167 HW1 - Spring 2025

Ethan Choi

Contents

Homework #1 instructions . . . 2
Homework submission guideline . . . 2
Acknowledgments . . . 2
Question 1 [70pt] Analysis of the quakes dataset . . . 3
(b) [10pt] Outlier detection . . . 4
(c) [10pt] Minimum variance unbiased estimates for Normal distribution . . . 5
(d) [10pt] Visualize your model fitness . . . 6
(e) [10pt] Pairwise scatterplots . . . 7
(f) [10pt] Single scatterplot . . . 8
(g) [10pt] Earthquake maps . . . 9
Question 2 [30pt] Introductory Survey . . . 12
(a) What is your major and concentration (if applicable)? . . . 12
(b) What is your minor (if any)? . . . 12
(c) What motivated you to enroll in STAT167? . . . 12
(d) What are you hoping to learn or achieve in this course? . . . 12
(e) What operating system(s) do you use on your primary computer(s)? . . . 13
(f) Which programming languages have you learned or used in the past? . . . 13
(g) What is your primary programming language? . . . 13
(h) How often did you write code before taking this class? . . . 13
(i) Overall, how comfortable are you with programming? . . . 13
(j) What probability and statistics courses have you completed, if any? . . . 13
(k) What computer science courses have you completed, if any? . . . 14

Homework #1 instructions

Review the lecture notes on "Review of R and base graphics" before answering the homework questions.

This homework contains 2 questions, each with multiple parts, 100 points in total.

Replace **INSERT_YOUR_ANSWER** with your own answers.

- First open this `rmd` file in RStudio and click `Knit -> Knit to PDF` to render it to PDF format. You need to have `LaTex` installed on the computer to render it to PDF format. If not, you can also render it to HTML format.
- It is best to read this `rmd` file and the rendered `pdf`/`html` file side-by-side, while you are working on this homework.
- If the question asks you to write some R code, remember to put your code into a **R code chunk**. Make sure both your R code chunk and its output are visible in the rendered `pdf`/`html` file.
- For this homework, use **R Base Graphics** to generate the figures. Do **NOT** use `ggplot2` for this homework.
- **Please comment your R code thoroughly, and follow the R coding style guideline (https://google.github.io/styleguide/Rguide.xml). Partial credit will be deducted for insufficient commenting or poor coding styles.**
- If you have any question about this homework assignment, we encourage you to post it on **Piazza**.

Homework submission guideline

- **This homework is DUE at *11:59 PM* on *Sunday April 13, 2025*.**
- Late submission penalties.
  - Submissions up to 24 hours late will incur a 10% deduction.
  - Submissions up to 48 hours late will incur a 30% deduction.
- **If you are using one or two of your free late days, please state here: INSERT_YOUR_ANSWER**
- After you complete all questions, save your `rmd` file to `FirstnameLastname-SID-HW1.rmd` and save the rendered pdf file to `FirstnameLastname-SID-HW1.pdf`. If you can not knit it to pdf, knit it to html first and then print/save it to pdf format.
- Submit **BOTH your source `rmd` file and the knitted `pdf` file** to **GradeScope**. Do NOT create a zip file. For the `pdf` submission, please tag specific pages that correspond with each question in the assignment.
- You can submit multiple times, you last submission will be graded.

Acknowledgments

Please list all the help you have received for completing this homework.

INSERT_YOUR_ANSWER

Install necessary packages

In order to use a package, it needs to be installed on your computer, but it only needs to be installed once.

```
## If needed, you can install the `maps` package first, then comment out the line below.
# install.packages("maps")
```

Load necessary packages

It is a recommended best practice to load all of your R packages and datasets at the beginning of your file in one chunk.

```
library(datasets) # for the `quakes` dataset
library(maps) # for map visualization
```

Question 1 [70pt] Analysis of the quakes dataset

The quakes dataset provides the locations of 1000 seismic events of magnitudes (MB) > 4.0. These events occurred in a cube near Fiji since 1964.

```
?quakes
```

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

```
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)  # get the dimension of the quakes dataset
## [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 ...
```

(a) [10pt] Visualizations of the earthquake magnitudes

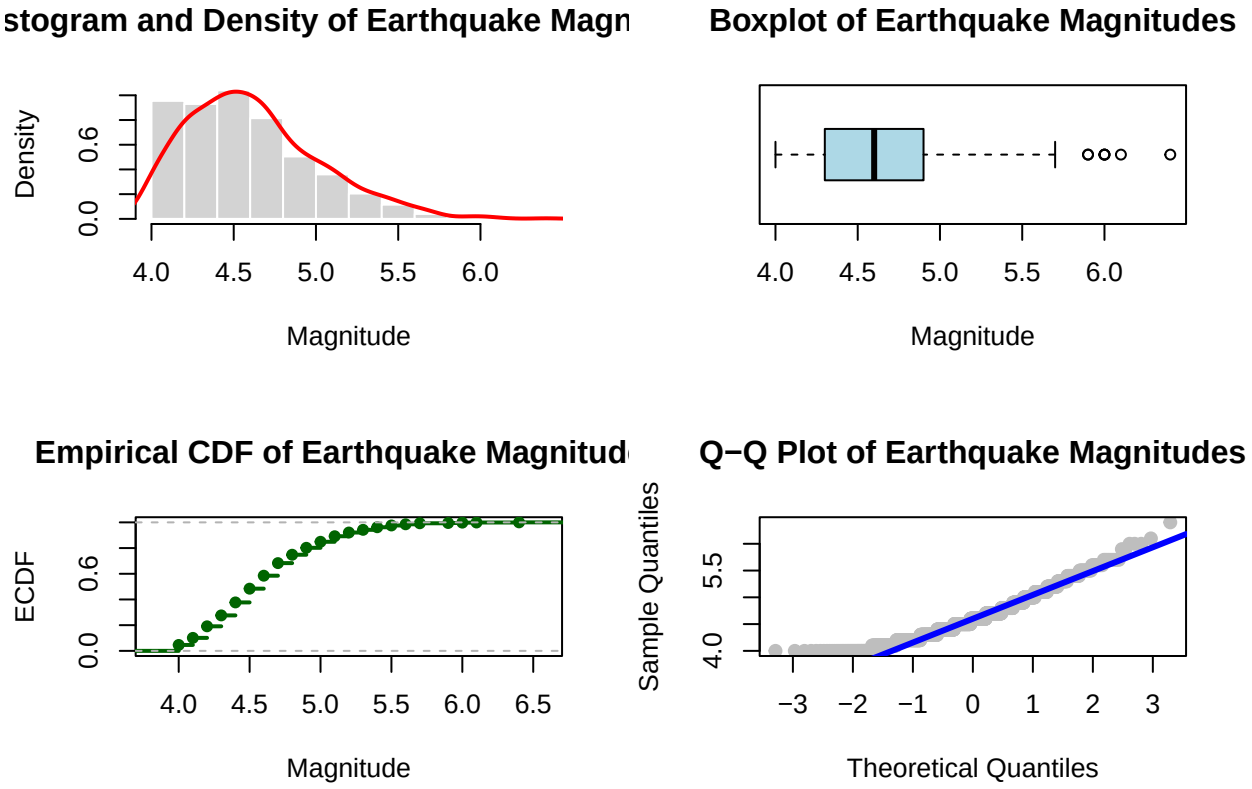
Suppose we are interested in studying the distribution of the magnitude mag variable. Write your own R code to create a single figure with the following four subfigures in a 2-by-2 layout.

- subfigure #1: plot a density histogram of earthquake magnitudes, and then plot the estimated probability density curve in red color in the same plot
- subfigure #2: plot a horizontal boxplot of earthquake magnitudes
- subfigure #3: plot the empirical cumulative distribution function (CDF) of the earthquake magnitudes
- subfigure #4: make a Q-Q plot to compare the observed distribution of earthquake magnitudes with the theoretical normal distribution. Add a *thick* qqline in blue color.

Note: make sure each subfigure includes proper axis labels and a figure title.

INSERT_YOUR_ANSWER

```r
# 2-by-2 layout
par(mfrow=c(2,2))

# Figure 1: Density Histogram
hist(quakes$mag, probability = TRUE,
     main = "Histogram and Density of Earthquake Magnitudes",
     xlab = "Magnitude", col = "lightgray", border = "white")
lines(density(quakes$mag), col = "red", lwd = 2)

# Figure 2: Horizontal Boxplot
boxplot(quakes$mag, horizontal = TRUE,
        main = "Boxplot of Earthquake Magnitudes",
        xlab = "Magnitude", col = "lightblue")

# Subfigure 3: Empirical CDF
plot(ecdf(quakes$mag), main = "Empirical CDF of Earthquake Magnitudes",
     xlab = "Magnitude", ylab = "ECDF", col = "darkgreen", lwd = 2)

# Subfigure 4: Q-Q Plot
qqnorm(quakes$mag, main = "Q-Q Plot of Earthquake Magnitudes",
       xlab = "Theoretical Quantiles", ylab = "Sample Quantiles", pch = 19, col = "gray")
qqline(quakes$mag, col = "blue", lwd = 3)
```



(b) [10pt] Outlier detection

There are several outliers indicated as open circles in the boxplot. Write R code to print out the indexes of these outliers and then print out the outlier observations.

Hint: you can use the `summary()` function to find out the first quartile (Q_1) and the third quartile (Q_3).

INSERT__YOUR__ANSWER

```
summary_stats <- summary(quakes$mag)
Q1 <- summary_stats["1st Qu."]
Q3 <- summary_stats["3rd Qu."]
IQR <- Q3 - Q1

lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

outlier_range <- which(quakes$mag < lower_bound | quakes$mag > upper_bound)

outlier_range
```

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

```
quakes[outlier_range, ]
```

```
##         lat   long depth mag stations
## 15   -20.70 169.92   139 6.1       94
## 17   -13.64 165.96    50 6.0       83
## 152  -15.56 167.62   127 6.4      122
## 558  -22.91 183.95    64 5.9      118
## 753  -21.08 180.85   627 5.9      119
## 870  -12.23 167.02   242 6.0      132
## 1000 -21.59 170.56   165 6.0      119
```

```
length(outlier_range)
```

```
## [1] 7
```

How many outliers do you find?

INSERT__YOUR__ANSWER Seven outliers

(c) [10pt] Minimum variance unbiased estimates for Normal distribution

Suppose we assume that the earthquake magnitudes follow 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:

$$\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

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

INSERT__YOUR__ANSWER

```
mu_hat <- mean(quakes$mag)
sigma2_hat <- var(quakes$mag)

print(mu_hat)
```

```
## [1] 4.6204
```

```r
print(sigma2_hat)
```

```
## [1] 0.1622261
```

Do NOT use any built-in R functions, write your own R code to calculate the estimates $\widehat{\mu}$ and $\widehat{\sigma^2}$.

INSERT_YOUR_ANSWER

```r
n <- length(quakes$mag)
mu_hat_nofunc <- sum(quakes$mag) / n

sigma2_hat_nofunc <- sum((quakes$mag - mu_hat_nofunc)^2) / (n - 1)

print(mu_hat_nofunc)
```

```
## [1] 4.6204
```

```r
print(sigma2_hat_nofunc)
```

```
## [1] 0.1622261
```

(d) [10pt] Visualize your model fitness

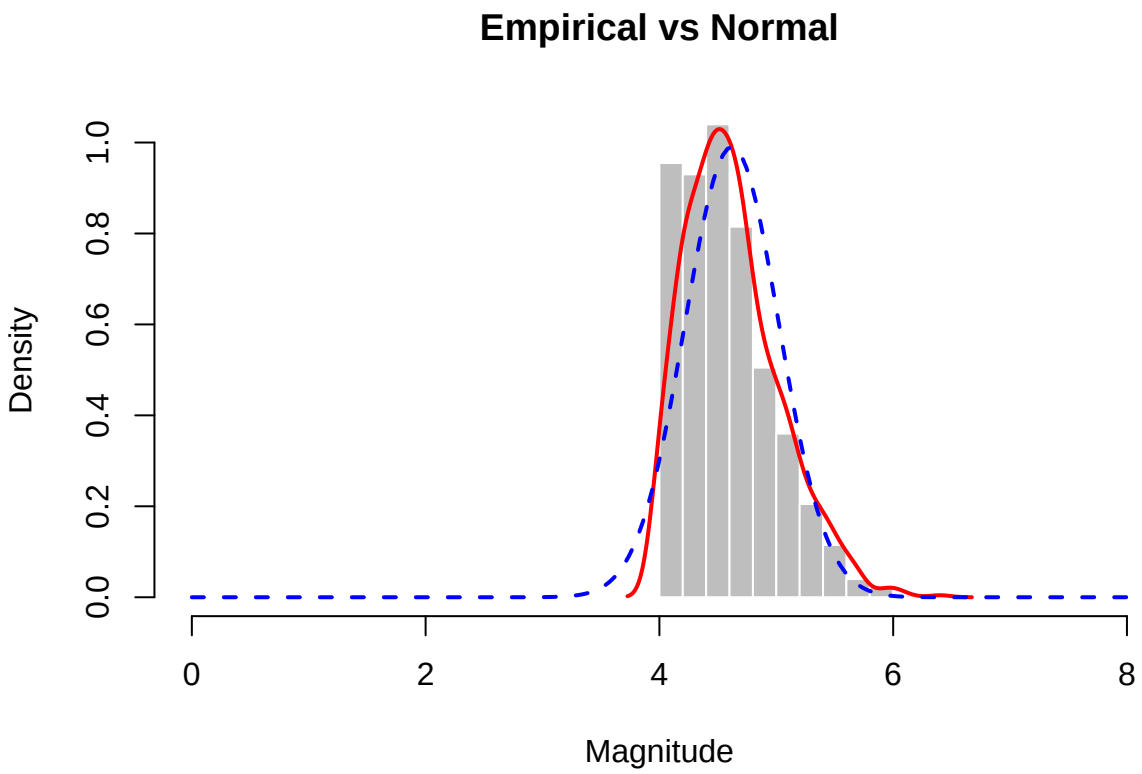
To visualize the model fitness, you can add the estimated normal distribution curve to the histogram plot you have generated in part (a).

Write your own code to generate the following figure.

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

INSERT_YOUR_ANSWER

```r
hist(quakes$mag, probability = TRUE, xlim = c(0, 8),
     main = "Empirical vs Normal",
     xlab = "Magnitude", col = "gray", border = "white")

lines(density(quakes$mag), col = "red", lwd = 2)

x_vals<-seq(0, 8, length.out = 200)
normal_density <- dnorm(x_vals, mean = mu_hat_nofunc, sd = sqrt(sigma2_hat_nofunc))
lines(x_vals, normal_density, col = "blue", lwd = 2, lty = 2)
```



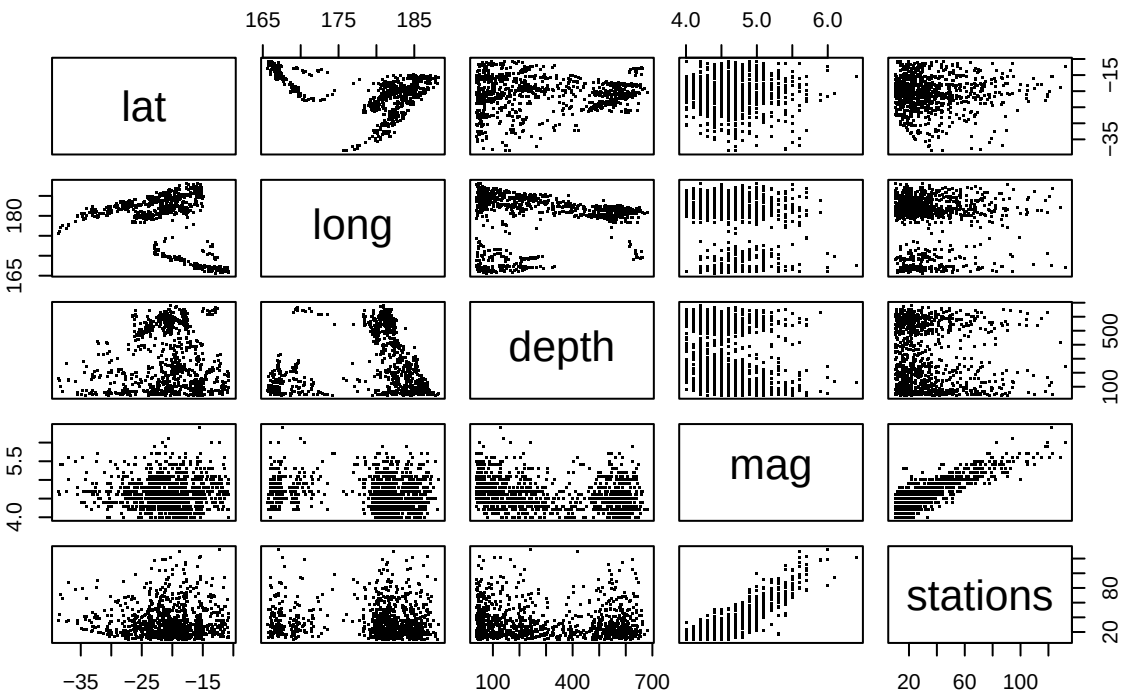
How does your estimated normal curve differ from the empirical density curve? Do you think the earthquake magnitude observations follow a normal distribution? Does this result align with the Q-Q plot you have generated in part (a).

INSERT_YOUR_ANSWER The estimated normal curve differs from the ED curve by being more bell-shaped and does not align perfectly with the empirical density. The tails on each end of the empirical curve appear to be slightly heavier. This suggests a deviation from normality. The observation does align with the Q-Q plot from part A, which showed that there are likely deviations in the tails.

(e) [10pt] Pairwise scatterplots

Run the following code.

```
pairs(quakes, main = "Fiji Earthquakes, N = 1000", cex.main=1.2, pch=".")
```



Describe the output figure.

INSERT_YOUR_ANSWER The output figure shows relationships between all listed variables. Some variable pairs show moderate patterns or clustering, more so than others. An example of this is that latitude and longitude show very clear grouping.

Do you think the `mag` variable and `stations` variables are positively correlated? Explain your answer.

INSERT_YOUR_ANSWER As magnitude increases, the number of reporting stations tends to increase. This is logical, since more noticeable, and stronger, quakes are detected by more stations. Therefore, the mag and stations variables are positively correlated.

(f) [10pt] Single scatterplot

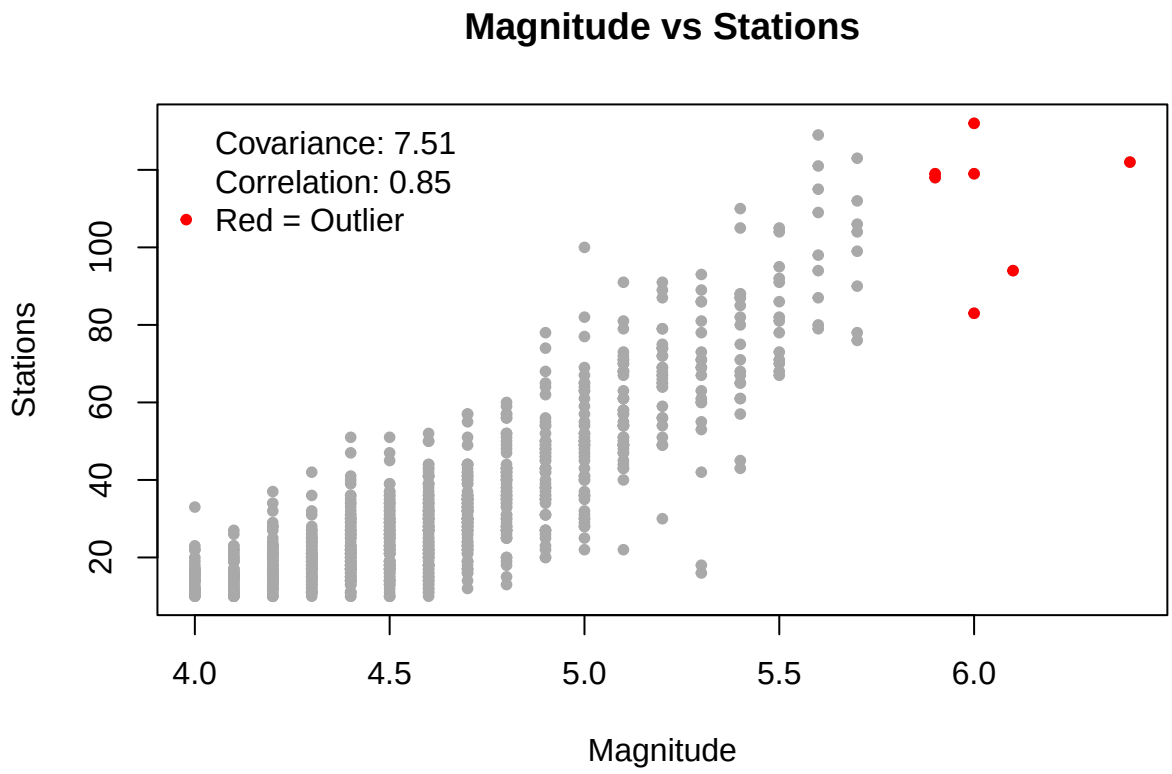
Write your own R code to reproduce the scatterplot for the `mag` and `stations` variables.

- calculate the covariance and correlation coefficient. **Hint**: look up the `cov` and `cor` functions.
- highlight the outlier points in part (b) using red-filled diamond symbols.
- 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 represent likely outliers.

INSERT_YOUR_ANSWER

```
#Cov and Corr
cov_magstat <- cov(quakes$mag, quakes$stations)
cor_magstat <- cor(quakes$mag, quakes$stations)

#Outlier indices from part B, use L and U bounds to determine what the outliers are
summary_stats <- summary(quakes$mag)
Q1 <- summary_stats["1st Qu."]
Q3 <- summary_stats["3rd Qu."]
IQR <- Q3 - Q1
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR
```

```r
outlier_range <- which(quakes$mag < lower_bound | quakes$mag > upper_bound)

#Plot
plot(quakes$mag, quakes$stations,
     xlab = "Magnitude", ylab = "Stations",
     main = "Magnitude vs Stations",
     pch = 20, col = "darkgray")

#Highlihgt
points(quakes$mag[outlier_range], quakes$stations[outlier_range],
       pch = 20, col = "red", bg = "red")

#Legend
legend("topleft", legend = c(
  paste("Covariance:", round(cov_magstat, 2)),
  paste("Correlation:", round(cor_magstat, 2)),
  "Red = Outlier"),
  pch = c(NA, NA, 20), col = c("black", "black", "red"), pt.bg = c(NA, NA, "red"),
  bty = "n")
```

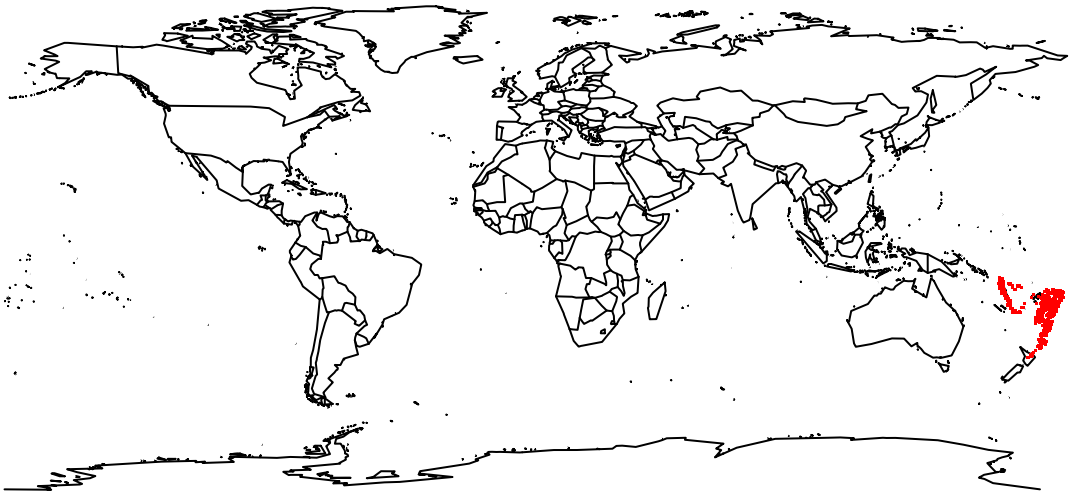


(g) [10pt] Earthquake maps

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

INSERT_YOUR_ANSWER Fiji is roughly northeast of New Zealand and east of Australia and is in the Southeast area of the Pacific Ocean.

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



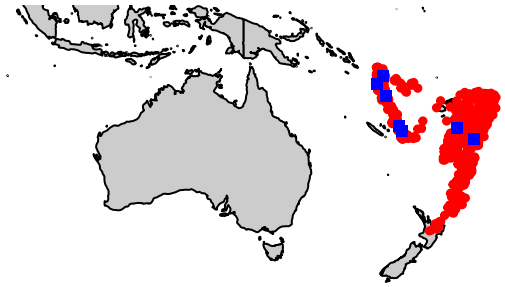
We can also zoom in and make local map of Fiji and its neighborhood area.

```
long.mean <- mean(quakes$long)
lat.mean <- mean(quakes$lat)
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 dots. Use blue-filled rectangles to mark locations of the outliers identified in part (b).

INSERT_YOUR_ANSWER

```
map(database="world", xlim=xlim, ylim=ylim, col="grey80", fill=TRUE)

points(quakes$long, quakes$lat, pch = 19, col = "red", cex = 0.5)

points(quakes$long[outlier_range], quakes$lat[outlier_range],
       pch = 15, col = "blue", bg = "blue", cex = 0.8)
```



Question 2 [30pt] Introductory Survey

Please complete the following short survey. *For multiple-choice questions, mark your answers using* ***bold*** *text.* Note that as long as your complete this survey, you will receive full credit for this question.

(a) What is your major and concentration (if applicable)?

INSERT_YOUR_ANSWER My major is Data Science

(b) What is your minor (if any)?

INSERT_YOUR_ANSWER N/A

(c) What motivated you to enroll in STAT167?

INSERT_YOUR_ANSWER Since I am a data science major, I wanted to take a class for it. I am a sophomore that is quite tired of learning lower division coursework.

(d) What are you hoping to learn or achieve in this course?

INSERT_YOUR_ANSWER I am hoping to gain rudimentary skills (or just practice) on data science tools.

(e) What operating system(s) do you use on your primary computer(s)?

Mark all that apply using **bold** text.

- **Windows**
- Mac OS X
- Linux
- Other: please specify

(f) Which programming languages have you learned or used in the past?

Mark all that apply using **bold** text.

- BASIC
- C
- **C++**
- C#
- Java
- JavaScript
- **HTML / CSS**
- LISP (lisp, scheme, clojure, etc)
- Perl
- PHP
- **Python**
- Ruby
- **SQL**
- VB / VBScript
- **R**
- SAS
- **Matlab**
- Julia
- Scala
- **Other: Verse**

(g) What is your primary programming language?

INSERT_YOUR_ANSWER C++ or Python

(h) How often did you write code before taking this class?

Mark your answer using **bold** text.

- daily
- two or more time per week
- **once per week**
- less than once per week
- little to no programming experience

(i) Overall, how comfortable are you with programming?

Mark your answer using **bold** text.

- 1 (not comfortable)
- 2
- 3
- **4**
- 5 (very comfortable)

(j) What probability and statistics courses have you completed, if any?

INSERT_YOUR_ANSWER STAT156A/B, STAT107, STAT010

(k) What computer science courses have you completed, if any?

INSERT_YOUR_ANSWER CS010A/B/C, CS141, CS111