HW 1

SDS 322E

September 01, 2022

Ethan Chang - ehc586

Please submit as an HTML file on Canvas before the due date

For all questions, include the R commands/functions that you used to find your answer. Answers without supporting code will not receive credit.

How to submit this assignment

> All homework assignments will be completed using R Markdown. These `.Rmd` files consist of text/syntax (formatted using Markdown) alongside embedded R code. When you have completed the assignment (by adding R code inside codeblocks and supporting text outside codeblocks), create your document as follows:

> - Click the “Knit” button (above)
> - Fix any errors in your code, if applicable
> - Upload the HTML file to Canvas

Q1 (0.5 pts)

The dataset `quakes` contains information about earthquakes occurring near Fiji since 1964. The first few observations are listed below.

```
head(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
## 6 -19.68 184.31   195 4.0       12
```

How many observations are there of each variable (i.e., how many rows are there; show using code)? How many variables are there total (i.e., how many columns are in the dataset)? You can read more about the dataset here (https://stat.ethz.ch/R-manual/R-patched/library/datasets/html/quakes.html) *Do not forget to include the code you used to find the answer each question*

```
str(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 ...
```

There are 1000 observations of each variable and there are 5 variables total

---

Q2 (1 pts)

What are the minimum, maximum, mean, and median values for the variables `mag` and `depth` ? Note that there are many functions that can be used to answer this question. If you chose to work with each variable separately, recall that you can access individual variables in a dataframe using the `$` operator (e.g., `dataset$variable`). Describe your answer in words.

```
four_m <- function(x) {
    min <- min(x)
    max <- max(x)
    mean <- mean(x)
    med <- median(x)

    mmmm <- list(min = min, max = max, mean = mean,
        med = med)
    return(mmmm)
}

print("mag")
```

```
## [1] "mag"
```

```
four_m(quakes$mag)
```

```
## $min
## [1] 4
##
## $max
## [1] 6.4
##
## $mean
## [1] 4.6204
##
## $med
## [1] 4.6
```

```
print("depth")
```

```
## [1] "depth"
```

```
four_m(quakes$depth)
```

```
## $min
## [1] 40
##
## $max
## [1] 680
##
## $mean
## [1] 311.371
##
## $med
## [1] 247
```

The minimum, maximum, mean, and median for `mag` , respectively, are 4, 6.4, 4.6204, and 4.6. The minimum, maximum, mean, and median for `depth` , respectively, are 40, 680, 311.371, and 247

Q3

Recall how logical indexing of a dataframe works in R. To refresh your memory, in the example code below I ask R for the median magnitude for quakes whose longitude is greater than 175.

```
median(quakes$mag[quakes$long > 175])
```

```
## [1] 4.5
```

Breaking this down a bit, the above line of code is doing the following (*this is just for illustration, the code itself is unnecessarily verbose*):

```
mags <- quakes$mag
longs <- quakes$long
is_long_greater_175 <- longs > 175  ## Makes a logical vector
mags_where_long_is_greater_175 <- mags[is_long_greater_175]  ## Indexing using logical vec
tor

median(mags_where_long_is_greater_175)
```

```
## [1] 4.5
```

3.1 (0.5 pts)

Explain in words what the single line of code is doing. Remember that the `$` selects a single variable and that `[ ]` are used for indexing whatever object came before (either a single variable or a dataframe).

The single line of code is finding all of the values from the long variable from the quakes dataframe that are greater than 175, and taking the median of the values from the mag variable from the quakes dataframe that correspond with the previously found long variable values.

3.2 (1.5 pts)

What is the mean of the variable `mag` when `depth` is *greater than* the median depth? What is the mean of the variable `mag` when `depth` is *less than* the median depth? What does this suggest about the relationship between an earthquake's depth and its magnitude?

```
print("depth > median")
```

```
## [1] "depth > median"
```

```
mean(quakes$mag[quakes$depth > median(quakes$depth)])
```

```
## [1] 4.5232
```

```
print("depth < median")
```

```
## [1] "depth < median"
```

```
mean(quakes$mag[quakes$depth < median(quakes$depth)])
```

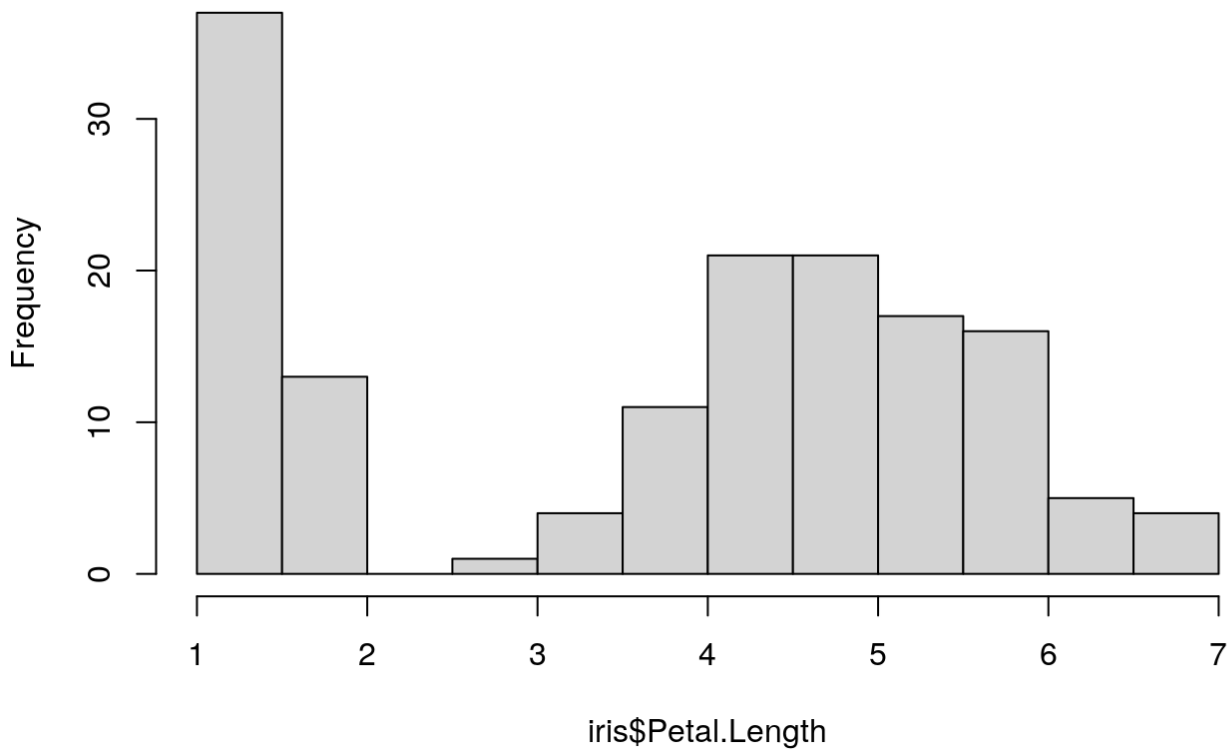
```
## [1] 4.7176
```

The mean of the variable `mag` when `depth` is *greater than* the median depth is 4.5232, while the mean of the variable `mag` when `depth` is *less than* the median depth is 4.7176. This suggests that the relationship between an earthquakes depth and its magnitude is that on average, the smaller the depth, the greater the magnitude.

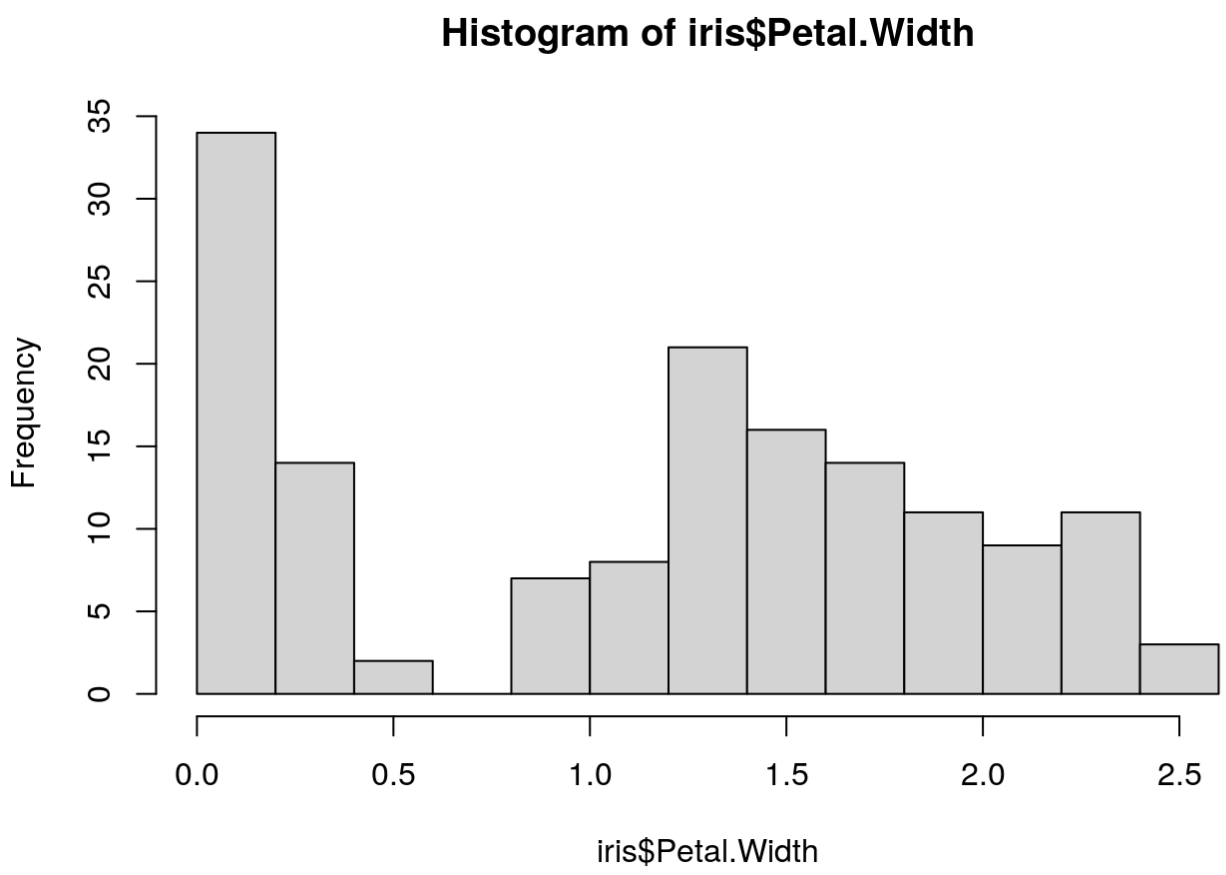
3.3 (1 pts)

The standard deviation (https://en.wikipedia.org/wiki/Standard_deviation) of a quantity is a measure of variable that quantity is. For example, the following plot gives histograms () of two variables (petal length and petal width from the `iris` dataset).

```
hist(iris$Petal.Length)
```




```
hist(iris$Petal.Width)
```



We see that the petal length is more variable than the petal width, which can be measured using the standard deviation (computed using the `sd` function):

```
print(sd(iris$Petal.Length))
```

```
## [1] 1.765298
```

```
print(sd(iris$Petal.Width))
```

```
## [1] 0.7622377
```

What is the standard deviation of the variable `lat` when `depth` is *greater than **the median depth? What is the standard deviation of the variable `lat` when `depth` is *less than* the median depth? What does this suggest about the relationship between an earthquake's latitude and it's depth?**

```
sd(quakes$lat[quakes$depth > median(quakes$depth)])
```

```
## [1] 3.577252
```

```
sd(quakes$lat[quakes$depth < median(quakes$depth)])
```

```
## [1] 6.1501
```

The standard deviation of the variable `lat` when `depth` is *greater than* the median depth is 3.577252, while the standard deviation of the variable `lat` when `depth` is *less than* the median depth is 6.1501. This suggests that the relationship between an earthquake's latitude and it's depth is, on average, the smaller the depth, the greater the standard deviation (variability) of the latitude.

Q4 (1 pts)

The variable `depth` is measured in kilometers. **Create a new variable called `depth_m` that gives depth in meters rather than kilometers and add it to the dataset `quakes` .** To help get you started, I have given you code that creates the new variable but fills it with `NA` values. Overwrite the `NA` s below by writing code on the right-hand side of the assignment operator (`<-`) that computes the requested transformation. Print out the first few rows of the updated dataset using `head()` .

```
# update the code below by replacing the NA with
# the correct expression to convert to meters
quakes$depth_m <- quakes$depth * 1000
head(quakes$depth_m)
```

```
## [1] 562000 650000  42000 626000 649000 195000
```

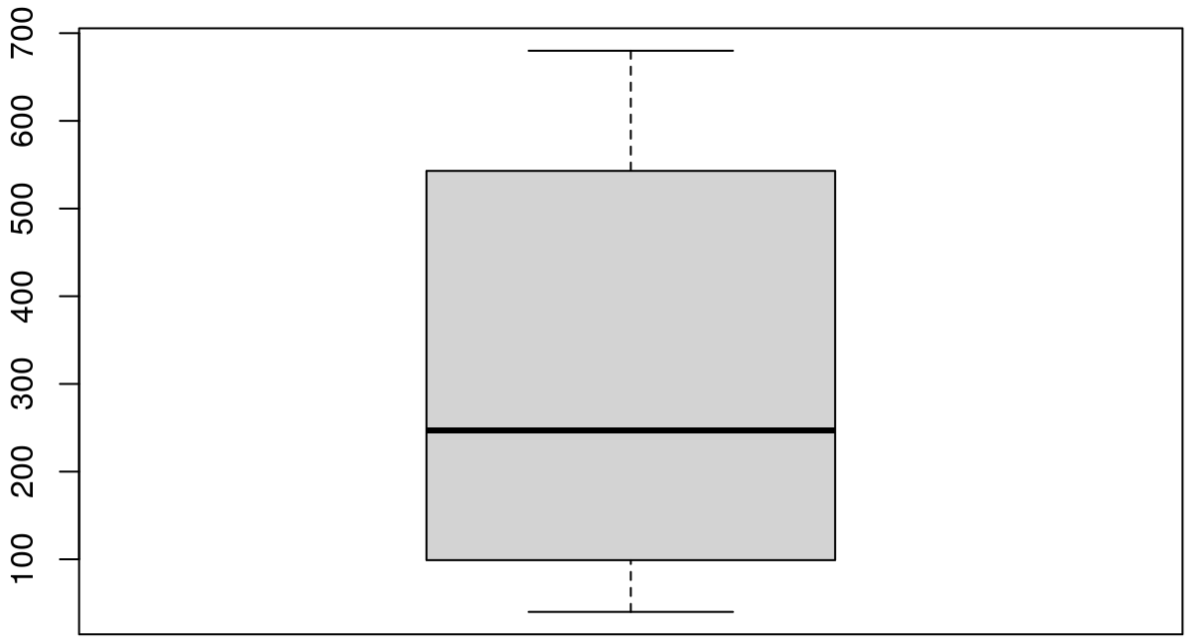
Q5

Let's make some plots in base R.

5.1 (1 pts)

Create a boxplot of `depth` using the `boxplot()` function. Describe where you see the min, max, and median (which you calculated in question 2) in this plot.

```
boxplot(quakes$depth)
```

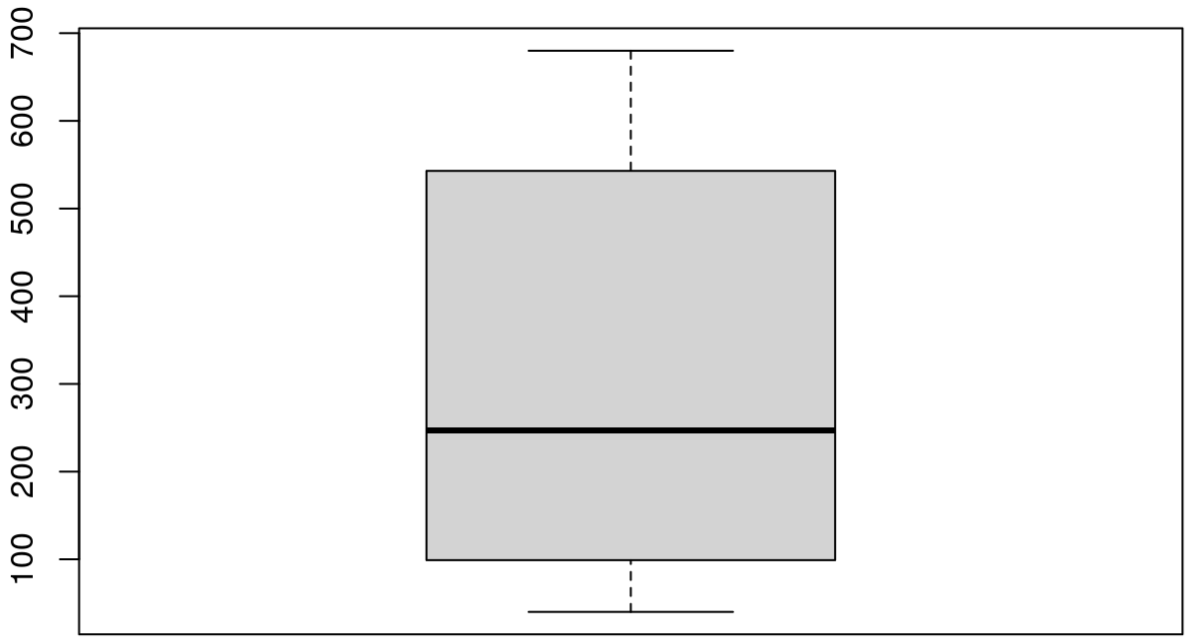


In this boxplot, you see the min in the bottom line (near 40), the median in the dark line in the middle of the box (near 247), and the max in the topmost line (near 680).

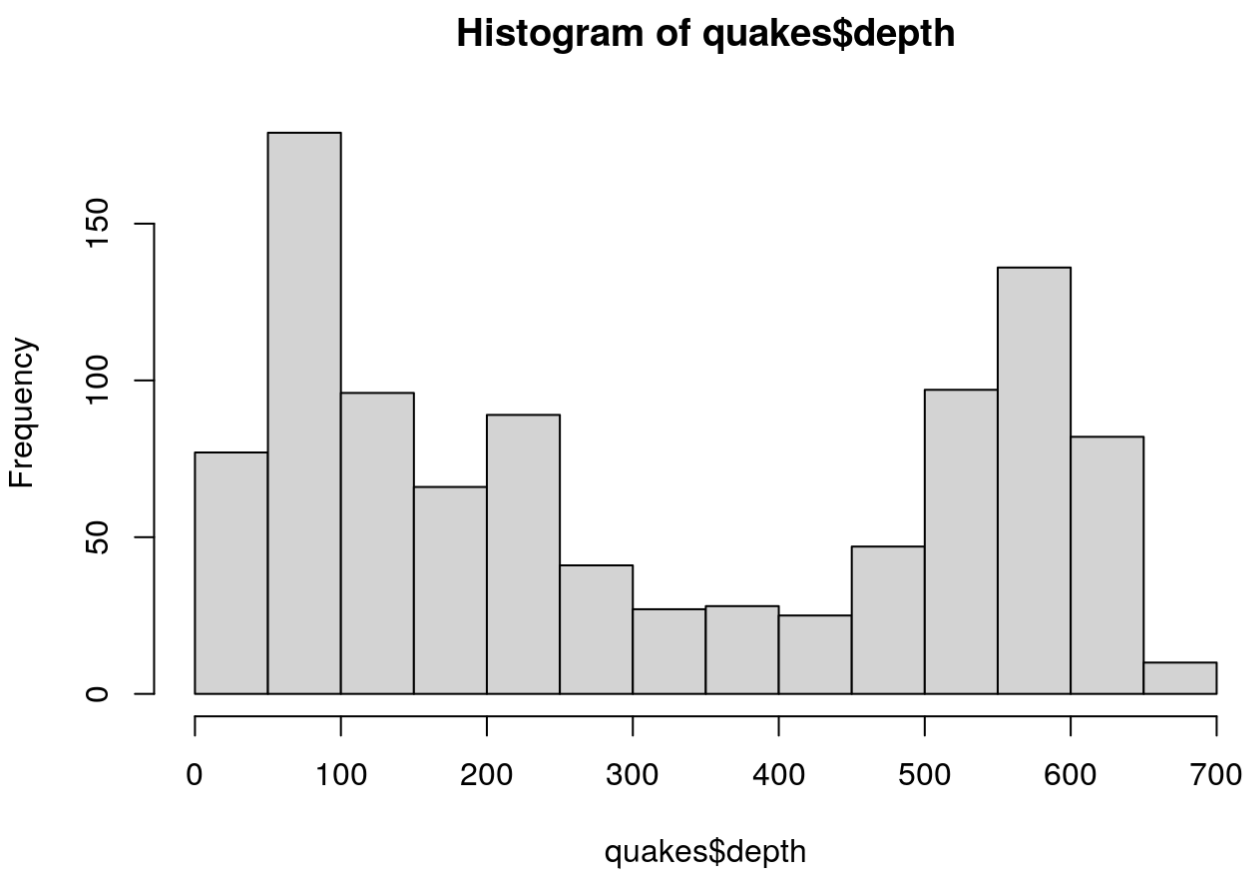
5.2 (1 pts)

Create a histogram of `depth` using the `hist()` function. What important information does the histogram provide that the boxplot does not?

```
# your code here
boxplot(quakes$depth)
```




```
hist(quakes$depth)
```

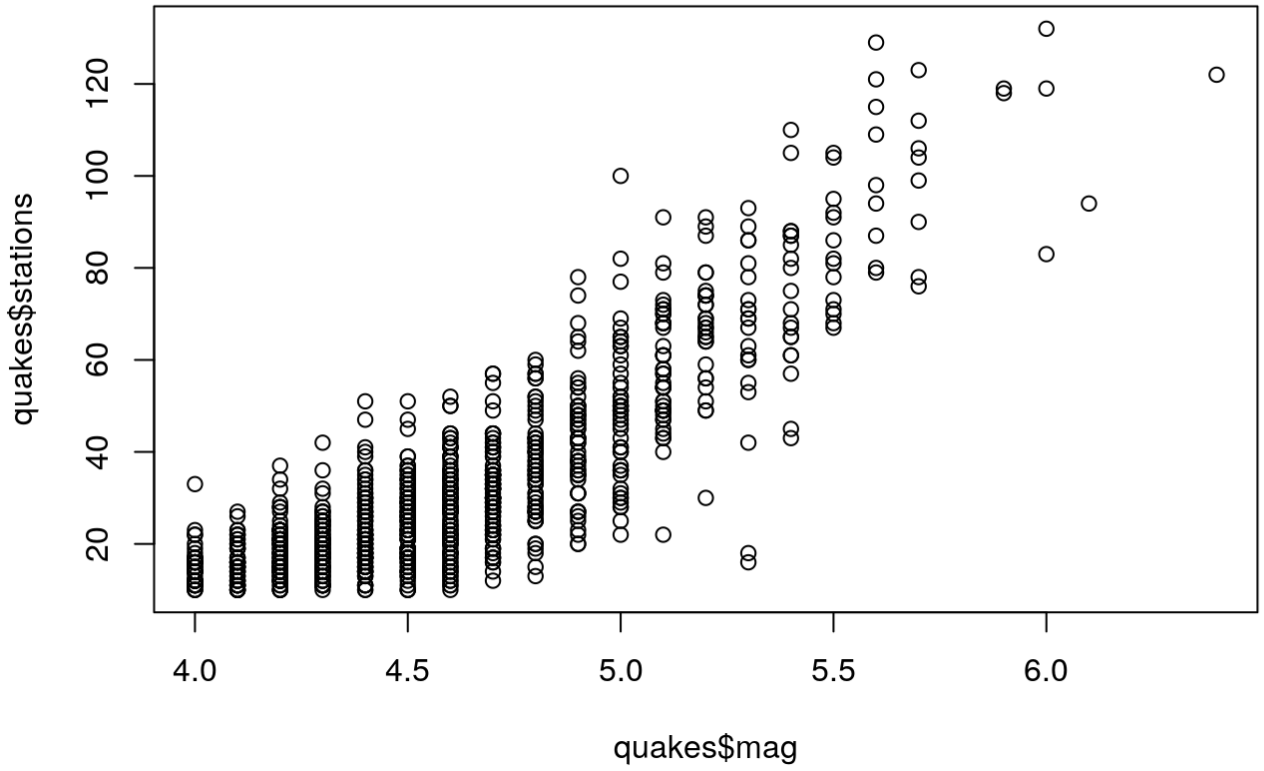


The important information that the histogram provides that the boxplot does not is that the histogram provides frequencies for each depth so we know how many times each depth tends to occur relative to the other depths. It makes its peaks more clear as a result, the boxplot mostly highlights the main 5 number statistics.

5.3 (1 pts)

Create a scatterplot by plotting variables `mag` and `stations` against each other using the `plot()` function. Note that to generate a scatterplot, the `plot()` takes two arguments: the x-axis variable and the y-axis variable. Describe the relationship between the two variables.

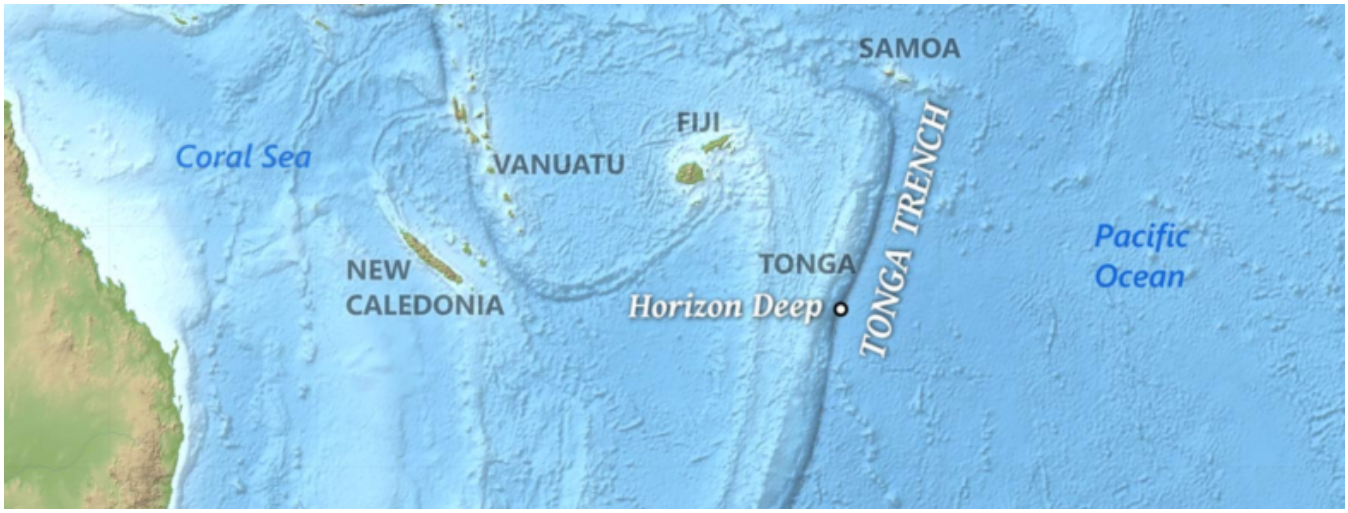
```
plot(quakes$mag, quakes$stations)
```



There appears to be a positive relationship between `mag` and `stations`. As `mag` increases, so does `stations`.

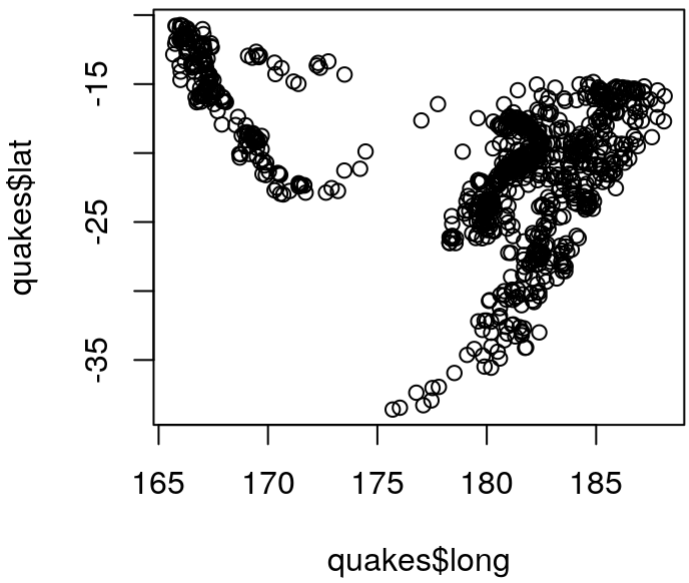
5.4 (1.5 pts)

Create scatterplot of the quakes' geographic locations by plotting `long` on the x-axis and `lat` on the y-axis. Using this plot, and the map/link below (note the two trenches), and some of the techniques you practiced above, are deeper quakes more likely to originate east or west of Fiji?



Link to location on Google maps
(https://www.google.com/maps/@-20.1679389,175.7587479,3513560m/data=!3m1!1e3)

```
plot(quakes$long, quakes$lat)
```



Based on this plot and the map provided above, deeper quakes are more likely to originate East of Fiji, along the Tonga Trench.

```
## R version 4.0.3 (2020-10-10)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.6 LTS
##
## Matrix products: default
## BLAS:   /stor/system/opt/R/R-4.0.3/lib/R/lib/libRblas.so
## LAPACK: /stor/system/opt/R/R-4.0.3/lib/R/lib/libRlapack.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] digest_0.6.29   R6_2.5.1        jsonlite_1.8.0  formatR_1.12
##  [5] magrittr_2.0.3  evaluate_0.15   highr_0.9       stringi_1.7.8
##  [9] cachem_1.0.6    rlang_1.0.4     cli_3.3.0       rstudioapi_0.13
## [13] jquerylib_0.1.4 bslib_0.4.0     rmarkdown_2.14  tools_4.0.3
## [17] stringr_1.4.0   xfun_0.31       yaml_2.3.5      fastmap_1.1.0
## [21] compiler_4.0.3  htmltools_0.5.3 knitr_1.39      sass_0.4.2
```

```
## [1] "2022-09-01 15:12:02 CDT"
```

```
##                                     sysname
##                                     "Linux"
##                                     release
##                           "4.15.0-191-generic"
##                                     version
## "#202-Ubuntu SMP Thu Aug 4 01:49:29 UTC 2022"
##                                    nodename
##                   "educcomp01.ccbb.utexas.edu"
##                                     machine
##                                    "x86_64"
##                                       login
##                                   "unknown"
##                                        user
##                                    "ehc586"
##                                effective_user
##                                    "ehc586"
```