

HW 1

SDS 322E

August 31, 2022

Contents

Q1 (0.5 pts)	1
Q2 (1 pts)	2
Q3	3
Q4 (1 pts)	6
Q5	6

Name: Simon Xiang. EID: szx58 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](#) 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 ...
```

Your typed answer goes here. 1-2 sentences only.

There are 1000 observations and 5 variables as can be seen in the output of `str(quakes)`.

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.

```
min(quakes$mag)
```

```
## [1] 4
```

```
max(quakes$mag)
```

```
## [1] 6.4
```

```
mean(quakes$mag)
```

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

```
median(quakes$mag)
```

```
## [1] 4.6
```

```
min(quakes$depth)
```

```
## [1] 40
```

```
max(quakes$depth)
```

```
## [1] 680
```

```
mean(quakes$depth)
```

```
## [1] 311.371
```

```
median(quakes$depth)
```

```
## [1] 247
```

Your typed answer goes here. 1-2 sentences only.

For `mag`: minimum = 4, maximum = 6.4, mean = 4.6204, median = 4.6. For `depth`: minimum = 40, maximum = 680, mean = 311.371, median = 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 vector

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).

This line takes the median magnitude of quakes, and only returns those whose longitude is greater than 175.

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?

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

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

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

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

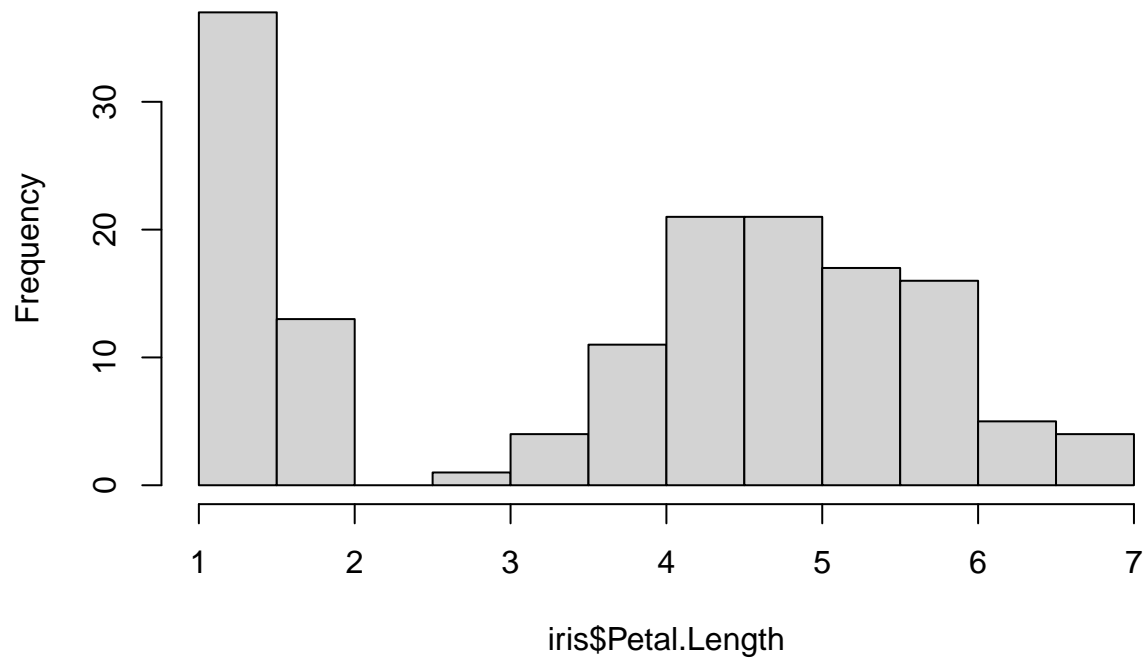
The mean of mag when depth is greater than median depth is 4.5232, and the mean of mag when depth is less than median depth is 4.7176. This implies that in general, the greater the depth the lower the magnitude.

3.3 (1 pts)

The 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).

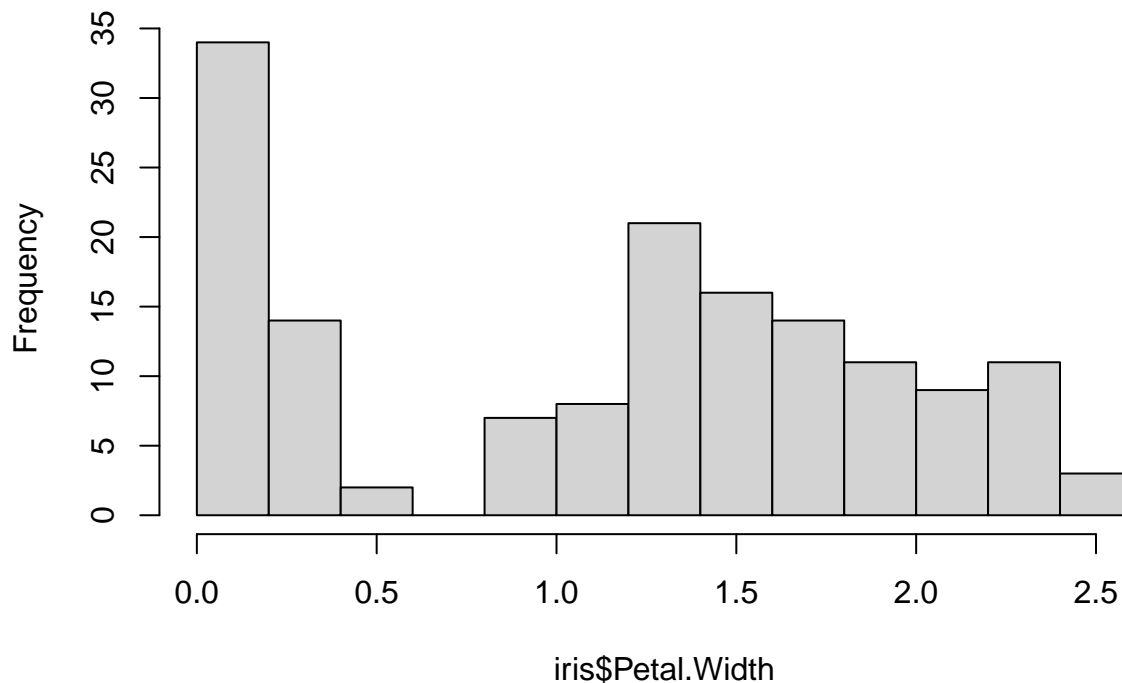
```
hist(iris$Petal.Length)
```

Histogram of iris\$Petal.Length



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

Histogram of 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?*

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

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

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

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

The standard deviation of latitude when depth is greater than the median is 3.57, and the standard deviation of latitude when depth is less than the median is 6.15. From this we deduce that the latitude is more variable when depth is above average.

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 NAs 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
```

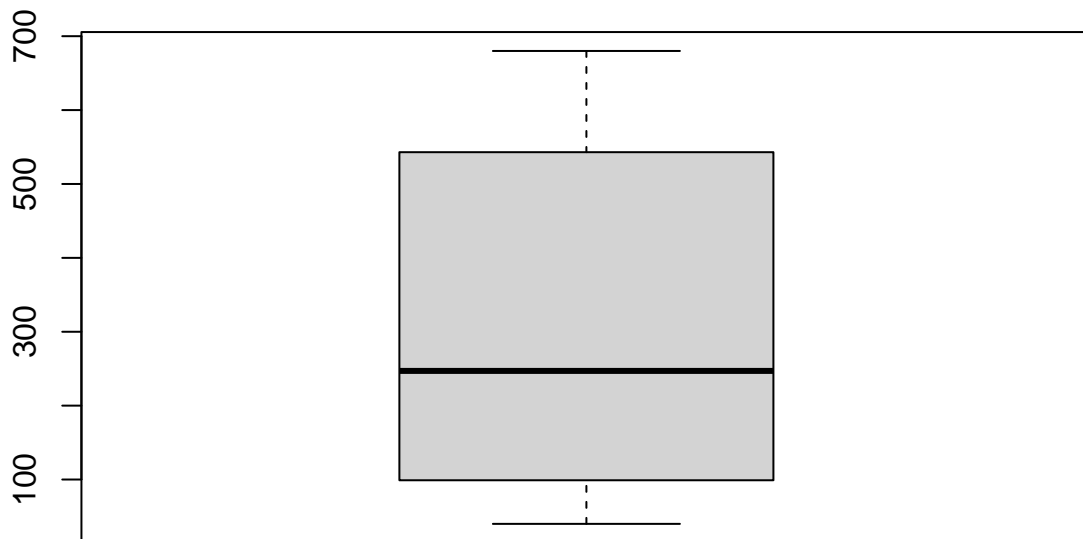
Q5

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)
```

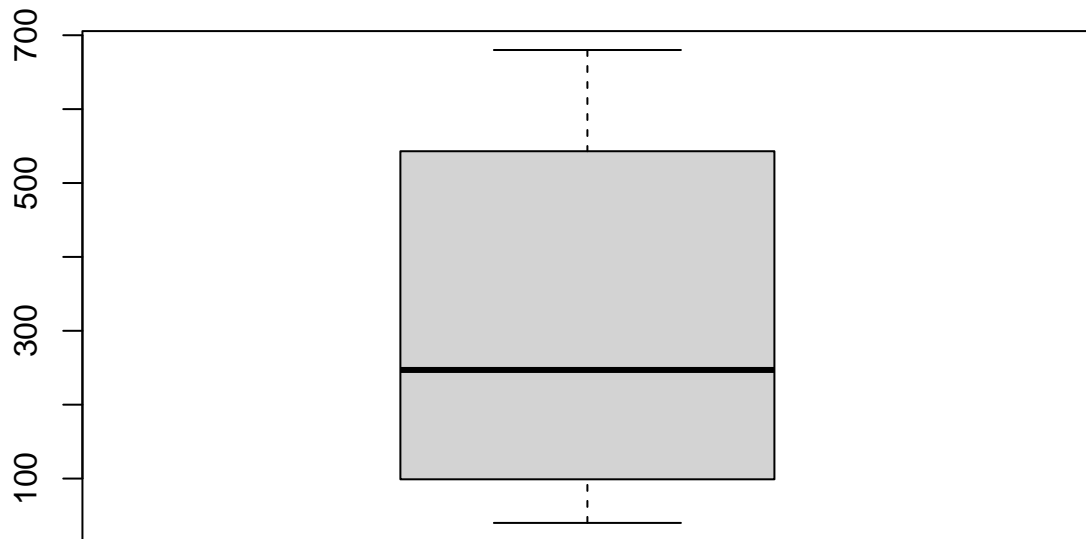


The min is the bottom line, the max is the top line, and the median is the thick black line that runs through the middle of the grey box.

5.2 (1 pts)

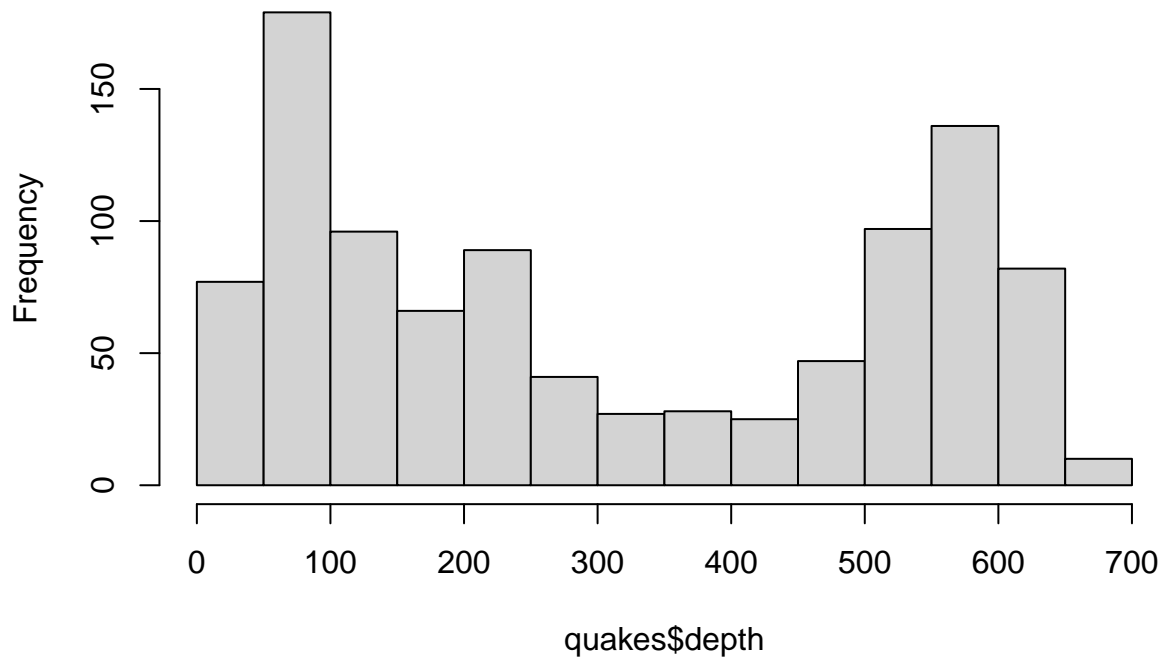
Create a histogram of depth using the `hist()` function. What important information does the histogram provide that the boxplot does not?

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



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

Histogram of quakes\$depth

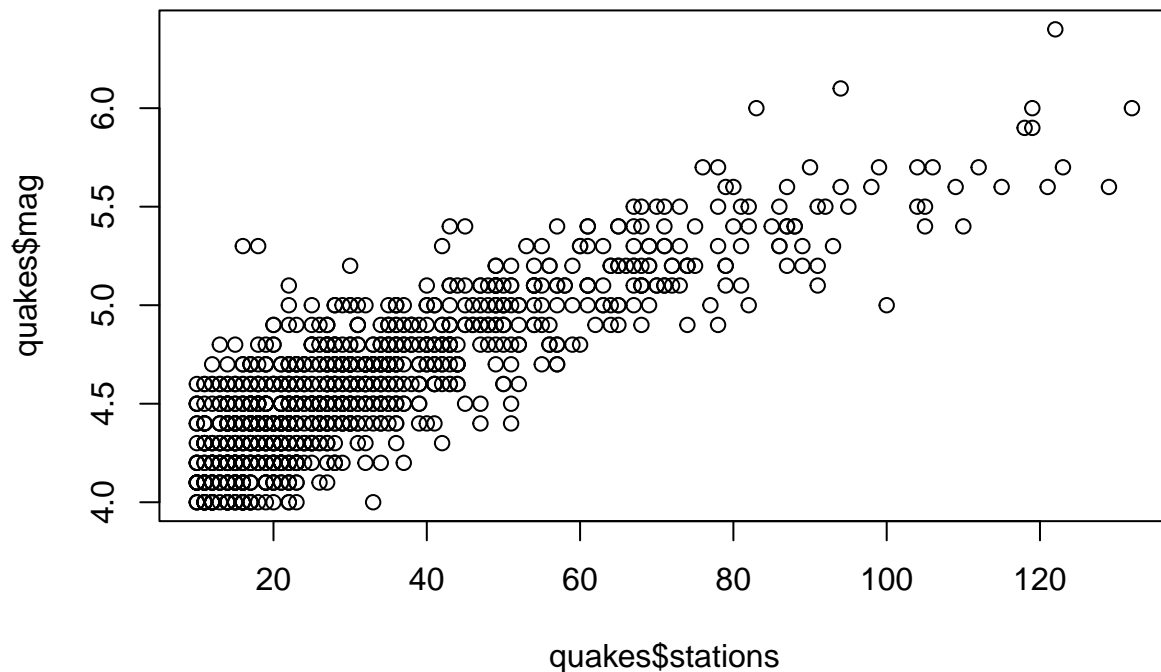


The histogram provides data about each element of `quakes$depth` itself, while boxplot aggregates the data and only displays overall information like median, min/max, etc.

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$stations, quakes$mag)
```

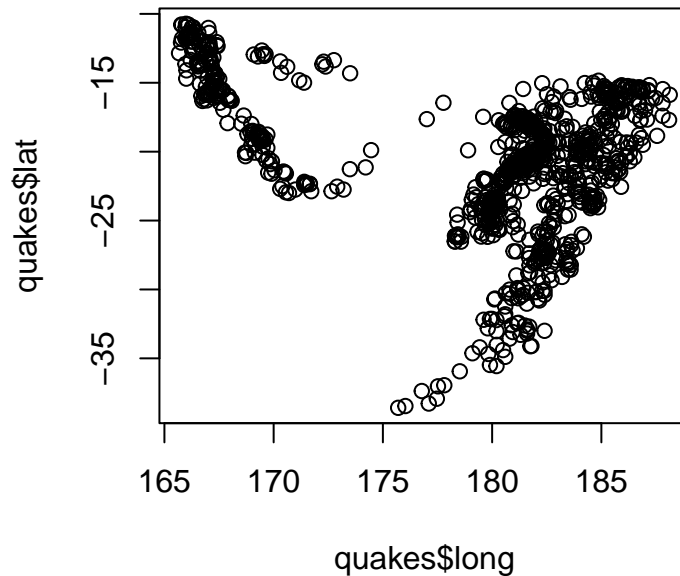
There seems to be a positive correlation with number of stations reporting the earthquake and magnitude. This makes sense because the larger the earthquake, the more likely stations will report it.

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?

Removed images attached since PdfLaTeX refused to compile them

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



Based off of the map and the scatterplot, earthquakes seem to be more likely to occur to the east of Fiji.

```
## R version 4.2.1 (2022-06-23)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Arch Linux
##
## Matrix products: default
## BLAS:   /usr/lib/libblas.so.3.10.1
## LAPACK: /usr/lib/liblapack.so.3.10.1
##
## 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] compiler_4.2.1  magrittr_2.0.1  fastmap_1.1.0  cli_3.3.0
##  [5] tools_4.2.1     htmltools_0.5.3 rstudioapi_0.13 yaml_2.3.5
##  [9] stringi_1.7.8   rmarkdown_2.15  knitr_1.39     stringr_1.4.0
## [13] xfun_0.32       digest_0.6.28   rlang_1.0.4    evaluate_0.16
## [1] "2022-08-31 17:44:46 CDT"
```

```
##          sysname
##          "Linux"
##          release
##          "5.15.54-1-lts"
##          version
## "#1 SMP Tue, 12 Jul 2022 18:08:24 +0000"
##          nodename
##          "automorphism"
##          machine
##          "x86_64"
##          login
##          "simon"
##          user
##          "simon"
##          effective_user
##          "simon"
```