Assignment1-1 Chen

Chloe Chen

2023-09-08

Turn in this assignment as an HTML or PDF file to ELMS. Make sure to include the R Markdown or Quarto file that was used to generate it. You should include the questions in your solutions. You may use the qmd file of the assignment provided to insert your answers.

Git and GitHub

1) Provide the link to the GitHub repo that you used to practice git from Week 1. It should have:

- Your name on the README file.
- At least one commit with your name, with a description of what you did in that commit.
- https://github.com/yqyhhh/Yang-Lou-Chen.git

======= ## Reading Data

Download both the Angell.dta (Stata data format) dataset and the Angell.txt dataset from this website: https://stats.idre.ucla.edu/stata/examples/ara/applied-regression-analysis-by-fox-data-files/

2) Read in the .dta version and store in an object called `angell_stata`.

```
library(haven)
angell_stata<-read_dta("D:/111/SURV727/Yang-Lou-Chen/angell.dta")
head(angell_stata)
```

```
## # A tibble: 6 × 5
##   city       morint ethhet geomob region
##   <chr>       <dbl>  <dbl>  <dbl> <chr>
## 1 Rochester    19     20.6   15   E
## 2 Syracuse     17     15.6   20.2 E
## 3 Worcester    16.4   22.1   13.6 E
## 4 Erie         16.2   14     14.8 E
## 5 Milwaukee    15.8   17.4   17.6 MW
## 6 Bridgeport   15.3   27.9   17.5 E
```

3) Read in the .txt version and store it in an object called `angell_txt`.

```
angell_txt<-read.table("https://stats.oarc.ucla.edu/wp-content/uploads/
2016/02/angell.txt")
head(angell_txt)
```

```
##           V1   V2   V3   V4 V5
## 1  Rochester 19.0 20.6 15.0  E
## 2   Syracuse 17.0 15.6 20.2  E
## 3  Worcester 16.4 22.1 13.6  E
## 4       Erie 16.2 14.0 14.8  E
## 5  Milwaukee 15.8 17.4 17.6 MW
## 6 Bridgeport 15.3 27.9 17.5  E
```

4) What are the differences between `angell_stata` and `angell_txt`? Are there differences in the classes of the individual columns?

- There are certain variable names in angell_stata, but the column names in angell_txt are simply V1…V5

5) Make any updates necessary so that `angell_txt` is the same as `angell_stata`.

```
colnames(angell_txt)<-c("city","morint","ethhet","geomob","region")
head(angell_txt)
```

```
##         city morint ethhet geomob region
## 1  Rochester   19.0   20.6   15.0      E
## 2   Syracuse   17.0   15.6   20.2      E
## 3  Worcester   16.4   22.1   13.6      E
## 4       Erie   16.2   14.0   14.8      E
## 5  Milwaukee   15.8   17.4   17.6     MW
## 6 Bridgeport   15.3   27.9   17.5      E
```

6) Describe the Ethnic Heterogeneity variable. Use descriptive statistics such as mean, median, standard deviation, etc. How does it differ by region?

```
mean(angell_stata$ethhet)

## [1] 31.37209

median(angell_stata$ethhet)

## [1] 23.7

sd(angell_stata$ethhet)

## [1] 20.41149

#1. if use the "split" function
table(angell_stata$region)

##
##  E MW  S  W
##  9 14 14  6

by_region<-split(angell_stata, angell_stata$region)
by_region$E

## # A tibble: 9 × 5
##   city       morint ethhet geomob region
##   <chr>       <dbl>  <dbl>  <dbl> <chr>
```

```
## 1 Rochester    19     20.6   15   E
## 2 Syracuse     17     15.6   20.2 E
## 3 Worcester    16.4   22.1   13.6 E
## 4 Erie        16.2   14     14.8 E
## 5 Bridgeport   15.3   27.9   17.5 E
## 6 Buffalo      15.2   22.3   14.7 E
## 7 Reading      14.2   10.6   19.4 E
## 8 Trenton      13     32.5   15.8 E
## 9 Baltimore    12     45.8   12.1 E

#2. if use the "piping operator" & group_by()
library(dplyr)    # also can "library(tidyverse)"

##
## package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

result <- angell_stata %>%
  group_by(region) %>%
  summarise(
    Mean_ethhet = mean(ethhet, na.rm = TRUE),
    Median_ethhet = median(ethhet, na.rm = TRUE),
    SD_ethhet = sd(ethhet, na.rm = TRUE)
  )

print(result)

## # A tibble: 4 × 4
##   region Mean_ethhet Median_ethhet SD_ethhet
##   <chr>        <dbl>         <dbl>     <dbl>
## 1 E             23.5          22.1     10.8
## 2 MW            21.7          19.2      9.08
## 3 S             52.5          53.8     21.4
## 4 W             16.5          16.1      4.16
```

Describing Data

R comes also with many built-in datasets. The “MASS” package, for example, comes with the “Boston” dataset.

7) Install the “MASS” package, load the package. Then, load the Boston dataset.

```
library(MASS)
```

```
##
## package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

data(Boston)
?Boston   #check details and codebooks of the data "Boston"
```

8) What is the type of the Boston object?

```
head(Boston)

##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
 lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7

typeof(Boston)

## [1] "list"
```

- The type of the object is a list.

9) What is the class of the Boston object?

```
class(Boston)

## [1] "data.frame"
```

- The class of the Boston object is data frame.

10) How many of the suburbs in the Boston data set bound the Charles river?

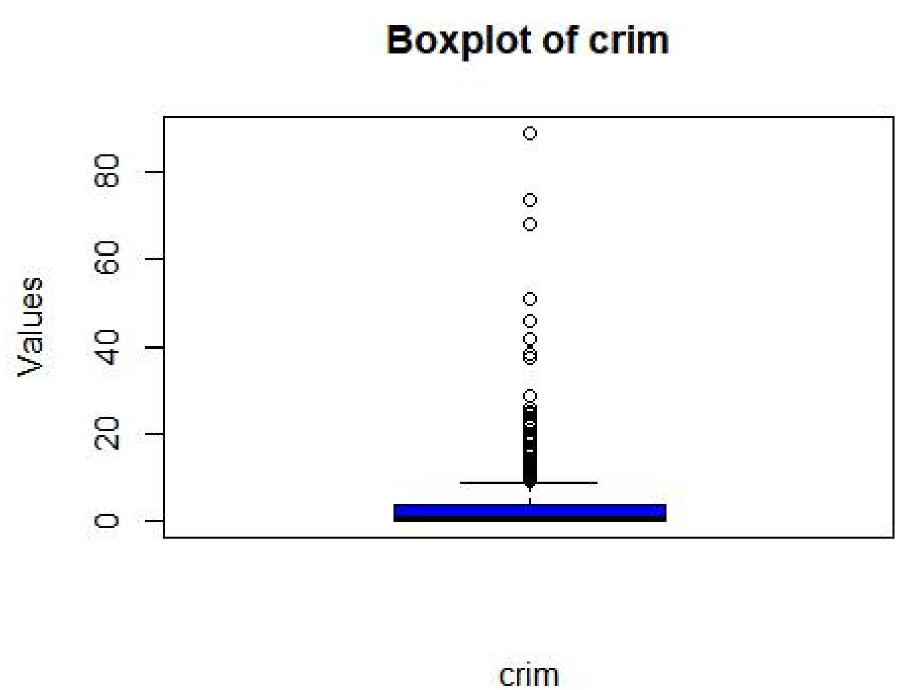
```
sumtable<-summary(Boston$chas)
tabB<- dim(Boston)
NBos<-tabB[1]
```

```
nsub<-NBos*mean(Boston$chas)
nsub

## [1] 35
```

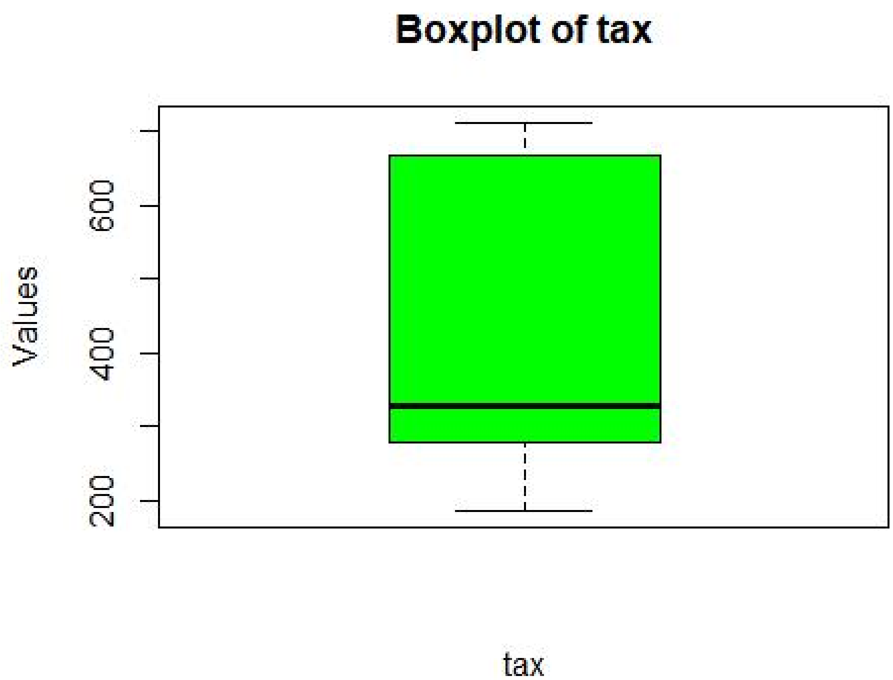
- There *35* suburbs set bound the Charles river.

11) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each variable.

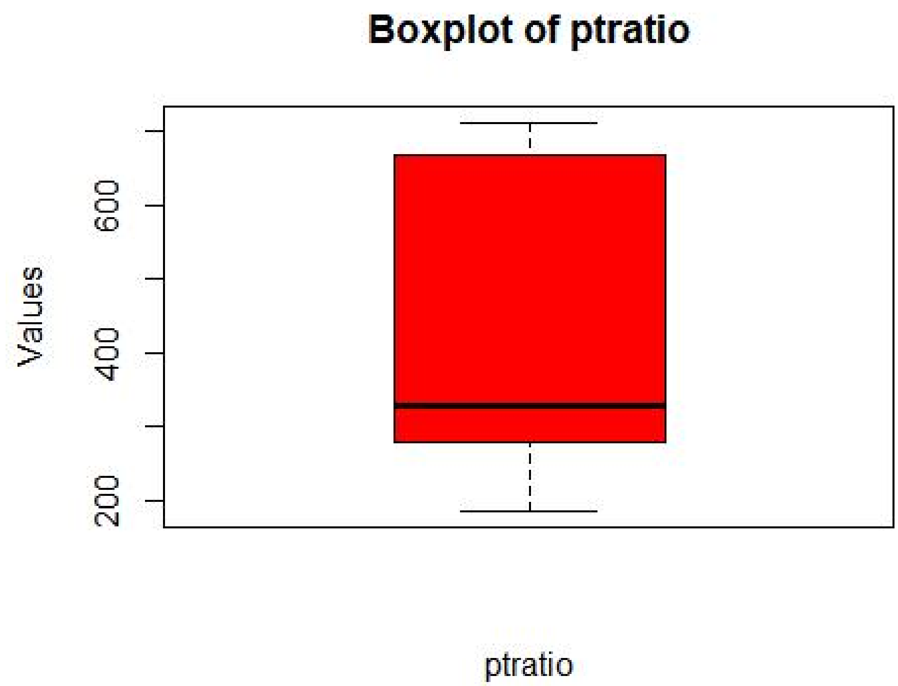
```
range_crim <- range(Boston$crim)
cat("Range of crim:", range_crim, "\n")

## Range of crim: 0.00632 88.9762

range_tax <- range(Boston$tax)
cat("Range of tax:", range_tax, "\n")

## Range of tax: 187 711

range_ptratio <- range(Boston$ptratio)
cat("Range of ptratio:", range_ptratio, "\n")

## Range of ptratio: 12.6 22

boxplot(Boston$crim, main="Boxplot of crim",
        xlab="crim", ylab="Values",
        col=c("blue"))
```



```r
boxplot(Boston$tax, main="Boxplot of tax",
        xlab="tax", ylab="Values",
        col=c("green"))
```




```r
boxplot(Boston$tax, main="Boxplot of ptratio",
        xlab="ptratio", ylab="Values",
        col=c("red"))
```

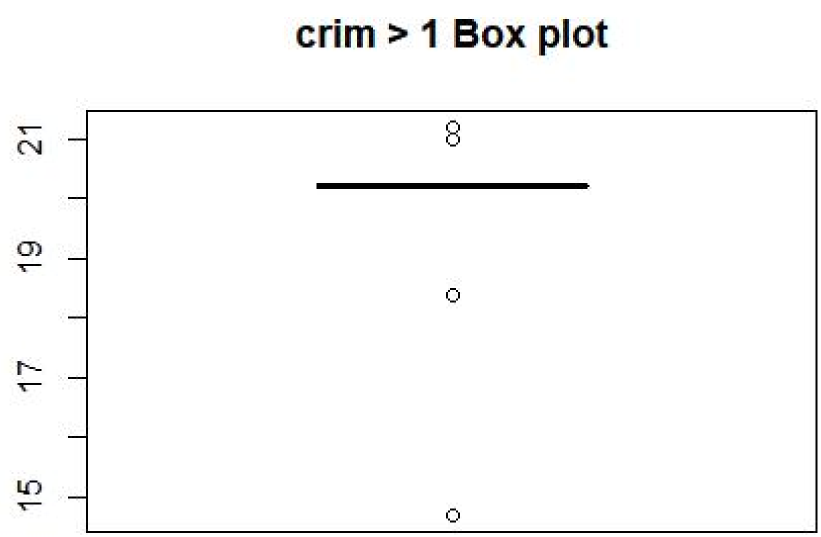


- The range of crime is $[0.00632, 88.9762]$; the range of tax is $[187, 711]$; the range of ptratio is $[12.6, 22]$.
- As we can see from the ranges and box-plots, there are particularly high crime rates in several suburbs.
- That doesn't exist in the other two variables. The range of tax is larger than ptratio, but there are no outliers occur. ptratio is the most compactly distributed data, with little difference between suburbs on this variable.

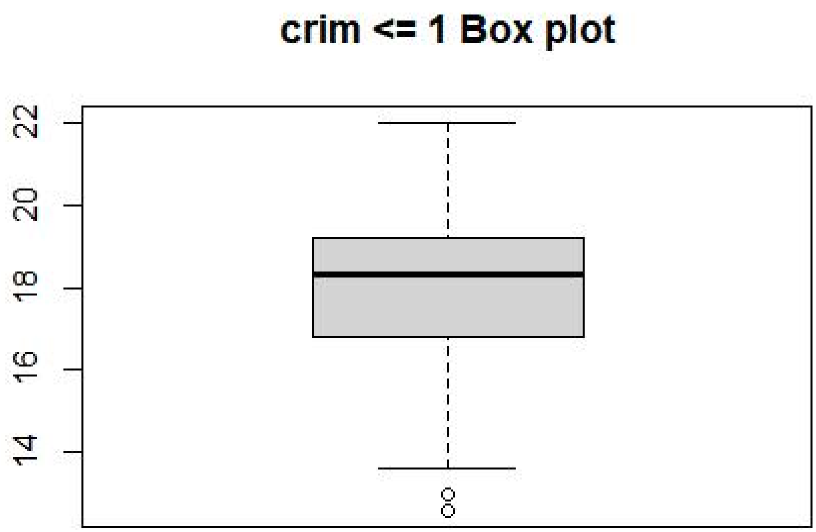
12) Describe the distribution of pupil-teacher ratio among the towns in this data set that have a per capita crime rate larger than 1. How does it differ from towns that have a per capita crime rate smaller than 1?

```
subset1 <- subset(Boston, Boston$crim > 1)
subset2 <- subset(Boston, Boston$crim <= 1)

#summary
summary(subset1$ptratio)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   14.70   20.20   20.20   19.29   20.20   21.20

summary(subset2$ptratio)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   12.60   16.80   18.30   18.02   19.20   22.00

var(subset1$ptratio)
```

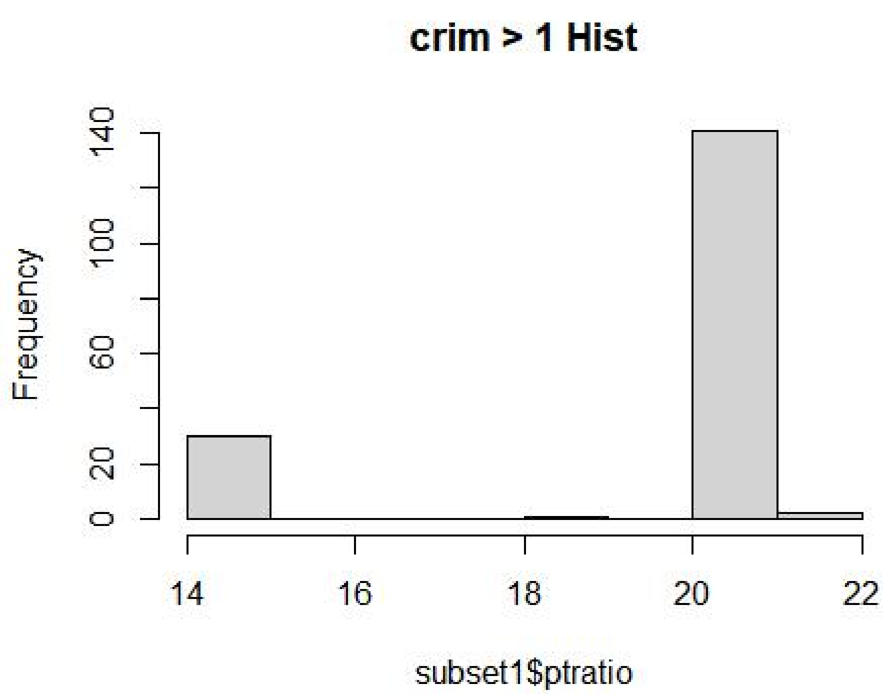
```
## [1] 4.484129

var(subset2$ptratio)

## [1] 4.243581

#Box plot
boxplot(subset1$ptratio, main="crim > 1 Box plot")
```



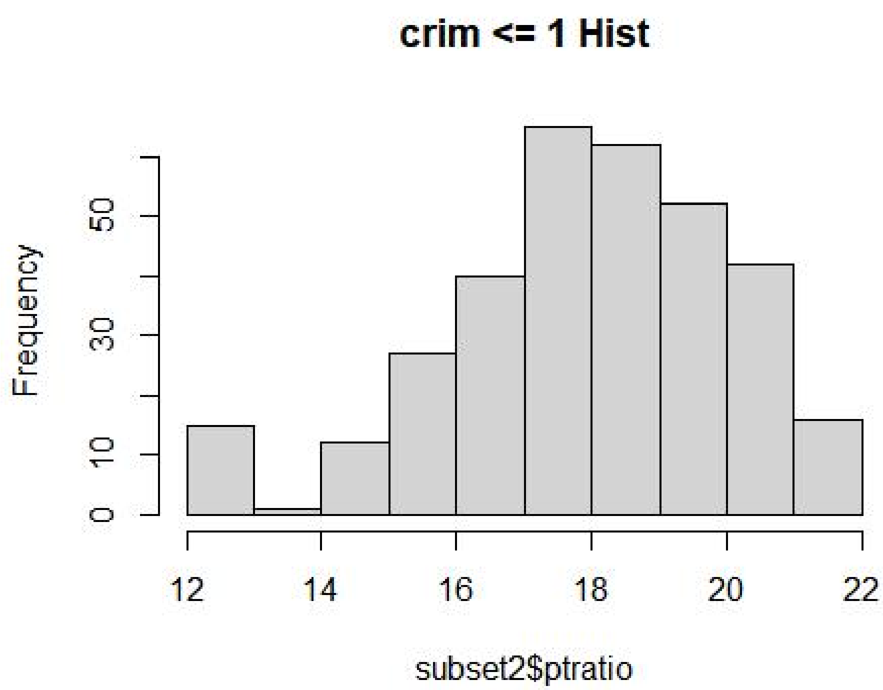

```
boxplot(subset2$ptratio, main="crim <= 1 Box plot")
```



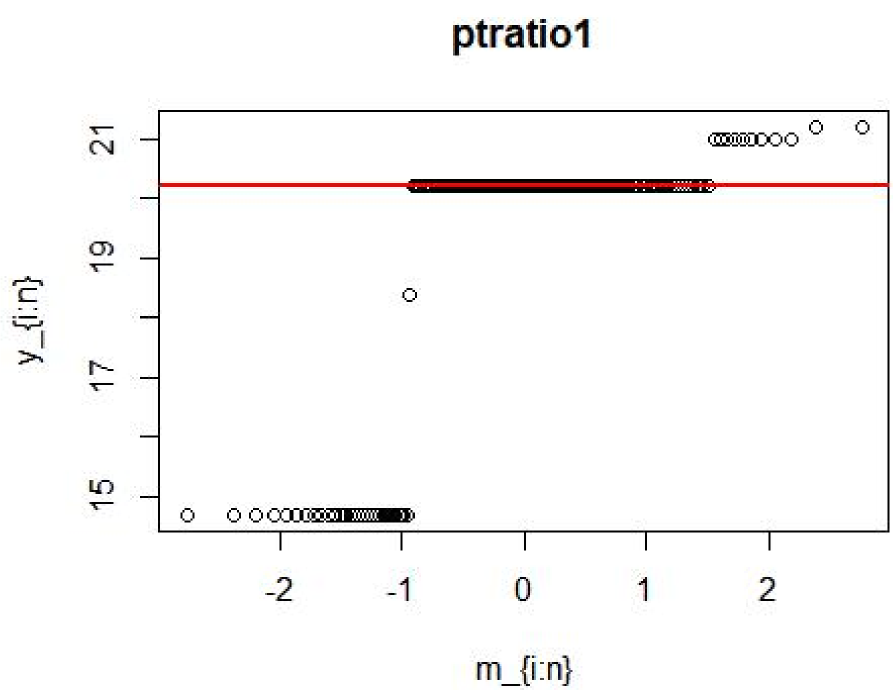

```
#Histogram
hist(subset1$ptratio, main="crim > 1 Hist")
```



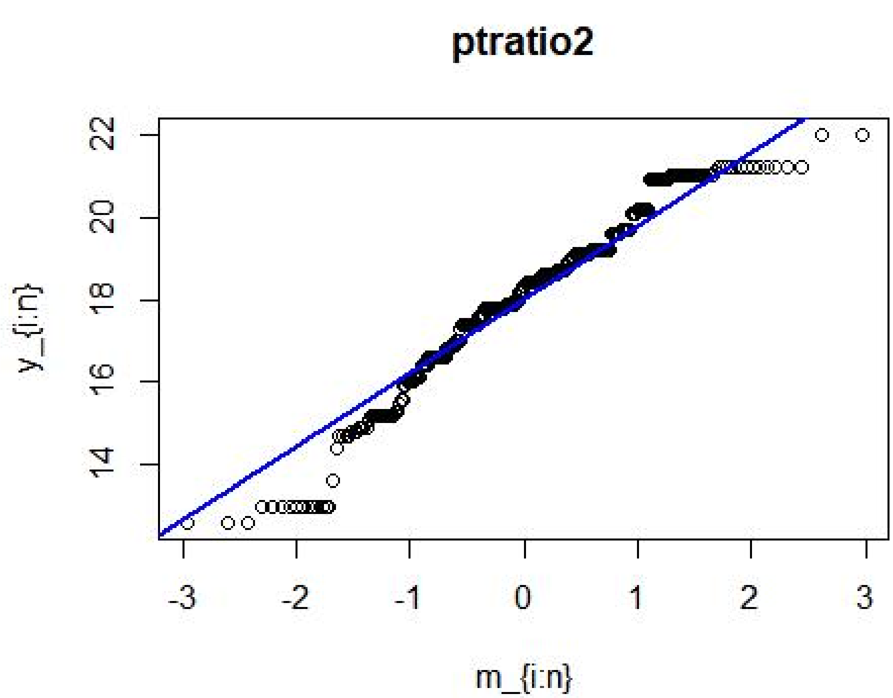
```
hist(subset2$ptratio, main="crim <= 1 Hist")

#Q-Q plot & Conduct statistical tests
library(ggplot2)
```




```
qqnorm(subset1$ptratio, main="ptratio1", ylab="y_{i:n}", xlab="m_{i:n}")

qqline(subset1$ptratio, col="red",lwd=2)
```




```
qqnorm(subset2$ptratio, main="ptratio2", ylab="y_{i:n}", xlab="m_{i:n}")

qqline(subset2$ptratio, col="blue",lwd=2)
```



```
shapiro.test(subset1$ptratio)

##
##  Shapiro-Wilk normality test
##
## data:  subset1$ptratio
## W = 0.51756, p-value < 2.2e-16

shapiro.test(subset2$ptratio)

##
##  Shapiro-Wilk normality test
##
## data:  subset2$ptratio
## W = 0.96226, p-value = 1.453e-07
```

Measures of Central Tendency The pupil-teacher ratio among the towns in this data set that have a per capita crime rate larger than 1 has a mean of *19.29* and a median of *20.20*. *Measures of Dispersion* The pupil-teacher ratio among the towns in this data set that have a per capita crime rate larger than 1 has a variance of *4.484129*. *Distribution Shape* In Shapiro-Wilk normality test, p-value < 0.05, and from the histogram, as well as the qq-plot, homeprice does not appear to follow a normal distribution. *Outliers* From the box plot we can see that there are *4* outliers in the distribution.

Writing Functions

13) Write a function that calculates 95% confidence intervals for a point estimate. The function should be called `my_CI`. When called with `my_CI(2, 0.2)`, the function should print out "The 95% CI upper bound of point estimate 2 with standard error 0.2 is 2.392. The lower bound is 1.608."

```
my_CI<- function(point_estimate,se){
  lower_bound<-point_estimate-1.96*se
  upper_bound<-point_estimate+1.96*se
  text <- paste("The 95% CI upper bound of point estimate", point_estima
te, "with standard error", se,"is", upper_bound, ". The lower bound is",
 lower_bound)
  text
}

ci<-my_CI(2,0.2)
ci

## [1] "The 95% CI upper bound of point estimate 2 with standard error 0.
2 is 2.392 . The lower bound is 1.608"
```

*Note: The function should take a point estimate and its standard error as arguments. You may use the formula for 95% CI: point estimate +/- 1.96*standard error.*

*Note: The function should take a point estimate and its standard error as arguments. You may use the formula for 95% CI: point estimate +/- 1.96*standard error.*

Hint: Pasting text in R can be done with: `paste()` *and* `paste0()`

14) Create a new function called `my_CI2` that does that same thing as the `my_CI` function but outputs a vector of length 2 with the lower and upper bound of the confidence interval instead of printing out the text. Use this to find the 95% confidence interval for a point estimate of 0 and standard error 0.4.

```
my_CI2<- function(point_estimate,se){
  lower_bound<-point_estimate-1.96*se
  upper_bound<-point_estimate+1.96*se
  c(lower_bound,upper_bound)
}

ci<-my_CI2(0,0.4)
ci

## [1] -0.784  0.784
```

15) Update the `my_CI2` function to take any confidence level instead of only 95%. Call the new function `my_CI3`. You should add an argument to your function for confidence level.

```
my_CI3 <- function(point_estimate, se, confidence_level) {
  if (confidence_level <= 0 || confidence_level >= 1) {
    stop("Confidence level must be between 0 and 1")
  }

  z_value <- qnorm(1 - (1 - confidence_level) / 2)

  lower_bound <- point_estimate - z_value * se
  upper_bound <- point_estimate + z_value * se

  c(lower_bound, upper_bound)
}
#can also do the same thing by function(point_estimate, se, z-value), wi
thout calculating the z-value within the function

ci_90 <- my_CI3(0, 0.4, 0.90)
print(ci_90)

## [1] -0.6579415  0.6579415

ci_99 <- my_CI3(0, 0.4, 0.99)
print(ci_99)

## [1] -1.030332  1.030332
```

Hint: Use the `qnorm` *function to find the appropriate z-value. For example, for a 95% confidence interval, using* `qnorm(0.975)` *gives approximately 1.96.*

16) Without hardcoding any numbers in the code, find a 99% confidence interval for Ethnic Heterogeneity in the Angell dataset. Find the standard error by dividing the standard deviation by the square root of the sample size.

```
se_ethhet <- sd(angell_stata$ethhet)/sqrt(nrow(angell_stata))
mean_ethhet <- mean(angell_stata$ethhet)
ethhetCI<-my_CI3(mean_ethhet, se_ethhet, 0.99)
ethhetCI

## [1] 23.35425 39.38993
```

- The 99% confidence interval for Ethnic Heterogeneity is [*23.35425*,*39.38993*].

17) Write a function that you can apply to the Angell dataset to get 95% confidence intervals. The function should take one argument: a vector. Use if-else statements to output NA and avoid error messages if the column in the data frame is not numeric or logical.

```
my_CI4 <- function(column) {
  if (is.numeric(column)==TRUE |is.logical(column)==TRUE) {
    mean_value <- mean(column, na.rm = TRUE)
    se <- sqrt(var(column, na.rm = TRUE) / length(column))
    z_value <- qnorm(0.975)  # 95% confidence interval

    lower_bound <- mean_value - z_value * se
    upper_bound <- mean_value + z_value * se

    return(c(lower_bound, upper_bound))
  } else {
    return(NA)
  }
}


result <- lapply(angell_stata, my_CI4) ## Apply this function to each co
lumn of Agell

result

## $city
## [1] NA
##
## $morint
## [1] 10.13242 12.26758
##
## $ethhet
## [1] 25.27127 37.47292
##
## $geomob
## [1] 24.67187 30.52347
```

```
##
## $region
## [1] NA
```