Problem Set 1

Applied Stats/Quant Methods 1

Han Li

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub.
- This problem set is due before 23:59 on Sunday October 1, 2023. No late assignments will be accepted.
- Total available points for this homework is 80.

Question 1 (40 points): Education

A school counselor was curious about the average of IQ of the students in her school and took a random sample of 25 students' IQ scores. The following is the data set:

```
y <- c(105, 69, 86, 100, 82, 111, 104, 110, 87, 108, 87, 90, 94, 113, 112, 98,
    80, 97, 95, 111, 114, 89, 95, 126, 98)
```

1. Find a 90% confidence interval for the average student IQ in the school.

```
y <- c(105, 69, 86, 100, 82, 111, 104, 110, 87, 108, 87, 90, 94, 113,
    112, 98, 80, 97, 95, 111, 114, 89, 95, 126, 98)
t.test(y,conf.level = 0.9, alternative = "two.sided")
```

```
One Sample t-test
data:  y
t = 37.593, df = 24, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:  93.95993 102.92007
sample estimates:
mean of x     98.44
```

To sanity check, we can also do the calculate the CI from mean:

```r
t_score <- qt(0.95, df=length(y)-1)
lower_90_t <- mean(y)-(t_score)*(sd(y)/sqrt(length(y)))
upper_90_t <- mean(y)+(t_score)*(sd(y)/sqrt(length(y)))
```

```
cat(lower_90_t, upper_90_t)
93.95993 102.9201
```

2. Next, the school counselor was curious whether the average student IQ in her school is higher than the average IQ score (100) among all the schools in the country.

   Using the same sample, conduct the appropriate hypothesis test with $\alpha = 0.05$.

```r
#H0:school iq <= 100 H1:school_iq> 100 (one sided)
t.test(y,mu=100, alternative = "greater")
```

```
One Sample t-test
data:  yt = -0.59574, df = 24,
p-value = 0.7215
alternative hypothesis: true mean is greater than 100
95 percent confidence interval: 93.95993      Inf
sample estimates:mean of x     98.44
```

   As the p-value is 0.72, way greater than the alpha value, we have more evidence in favor of the null hypothesis, which is the IQ of the school is not higher than the average of 100.

Question 2 (40 points): Political Economy

Researchers are curious about what affects the amount of money communities spend on addressing homelessness. The following variables constitute our data set about social welfare expenditures in the USA.

`State`	*50 states in US*
`Y`	*per capita expenditure on shelters/housing assistance in state*
`X1`	*per capita personal income in state*
`X2`	*Number of residents per 100,000 that are "financially insecure" in state*
`X3`	*Number of people per thousand residing in urban areas in state*
`Region`	*1=Northeast, 2= North Central, 3= South, 4=West*

Explore the `expenditure` data set and import data into `R`.

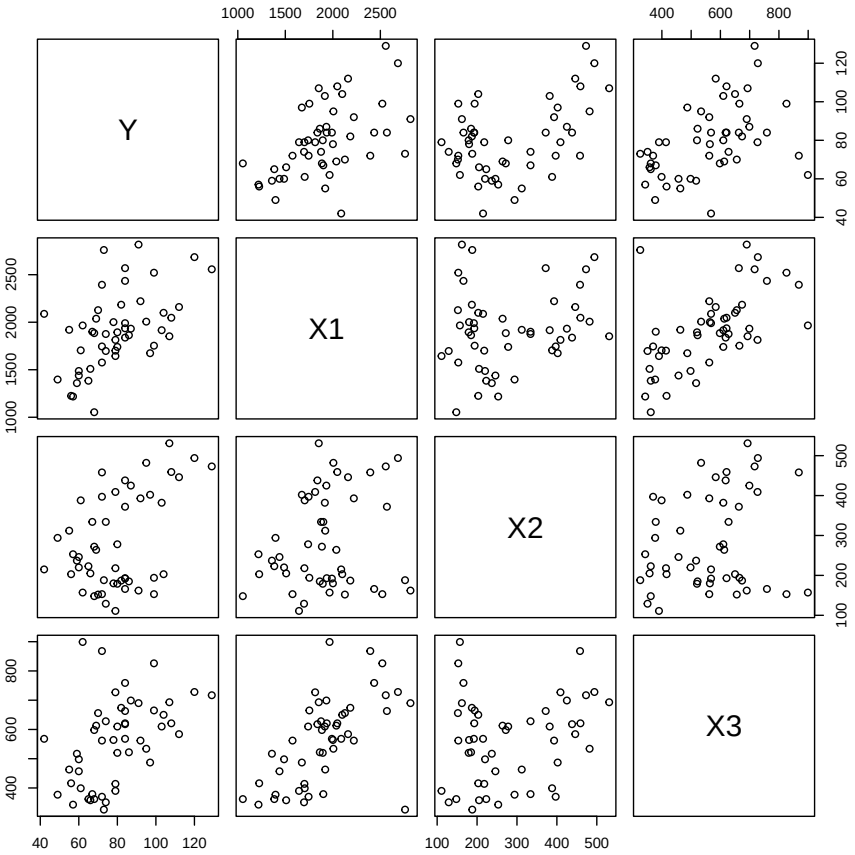
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?

- Please plot the relationship between *Y* and *Region*? On average, which region has the highest per capita expenditure on housing assistance?

- Please plot the relationship between *Y* and *X1*? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.

```
pdf("relationships.pdf")
pairs(expenditure[,c('Y','X1','X2','X3')])
dev.off( )
```

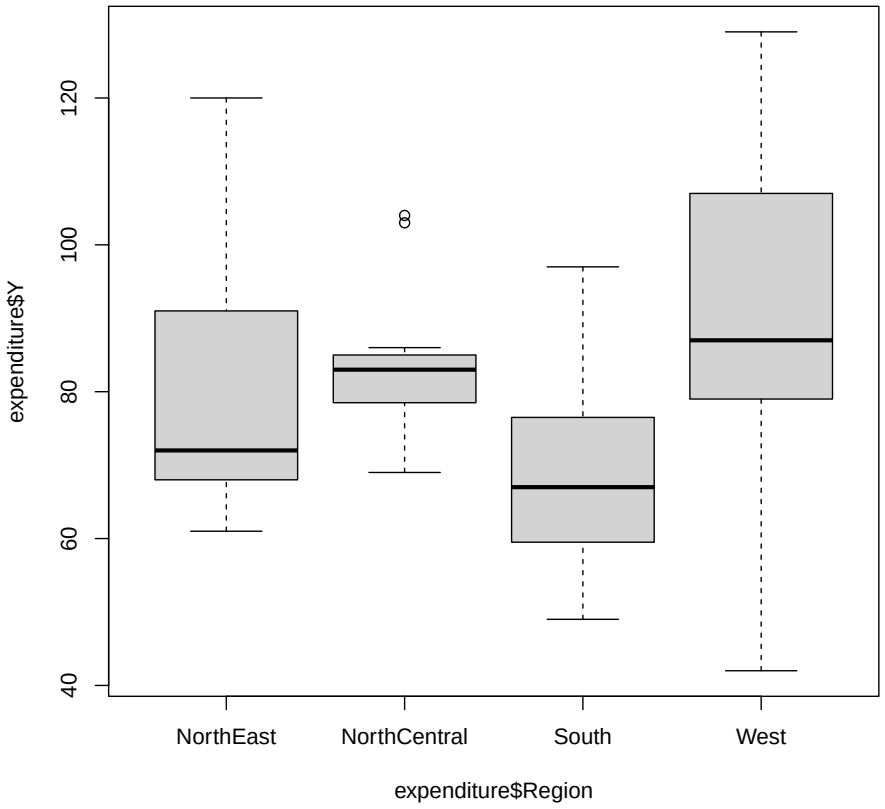
Figure 1: Scatter plot mattix Y, X1, X2,X3



There is a visually linear positve correlation between Y and X1, with seemingly increased variance as X1 increase.
There is a visually linear strong positve correlation between Y and X2.
There is a visually weak linear positve correlation between Y and X3.
There is a visually strong linear positve correlation between X1 and X2
There is unclear (no correlation) or weak positive correlation between X2 and X3 visually, more analysis is needed to decide whether correlation exists.
There is a visually strong linear positve relationship between X1 and X3, with some outliers.

```
pdf("ExpenditurebyRegion.pdf")
boxplot(expenditure$Y~ expenditure$Region,names =c("NorthEast", "NorthCentral"
    , "South","West"))
dev.off()
```

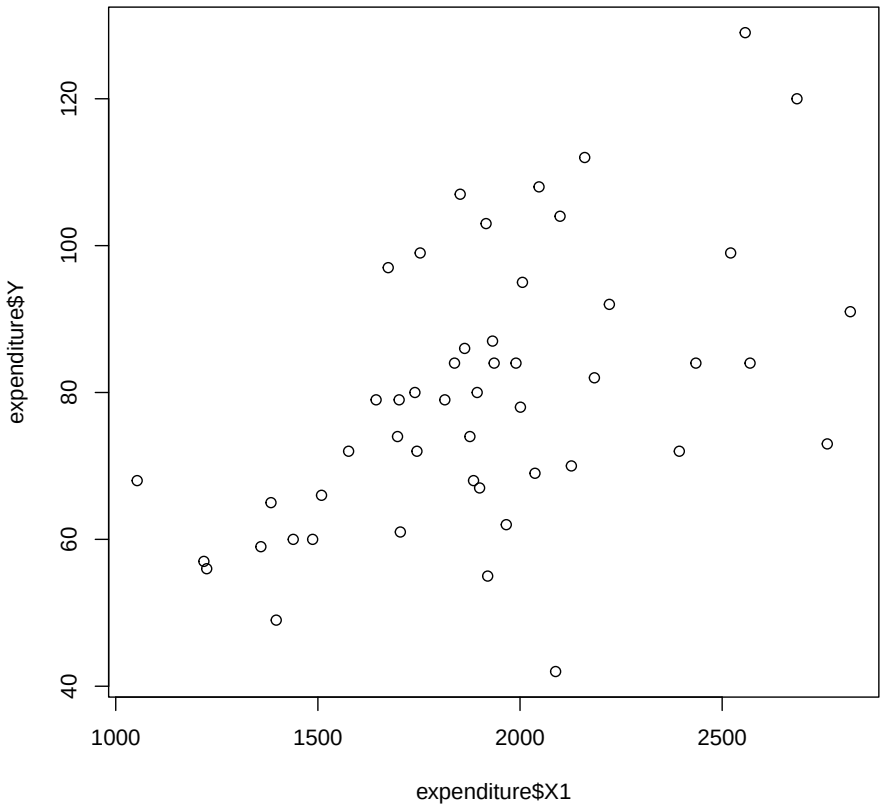
Figure 2: Expenditure by Region



On average the West region has the highest per capita expenditure.

```
pdf("plotyx1.pdf")
plot(expenditure$X1,expenditure$Y)
dev.off()
```

Figure 3: Scatter plot Y and X1



There is a positive linearcorrelation between X1,per capita personal income in state and Y, the per capita expenditure. When X1 increases, Y also increases.

```
pdf("YbyX1Region.pdf")
plot(expenditure$X1, expenditure$Y, col=expenditure$Region, pch=expenditure$
    Region)
legend(1000,120,
legend=c("NorthEast", "NorthCentral", "South","West"),
col = c('black','red','green','blue'),
pch=c(1,2,3,4))
dev.off()
```

Figure 4: Scatter plot Y and X1 by Region