Problem Set 1

Applied Stats/Quant Methods 1

Due: October 1, 2021

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 in `.pdf` form.
- This problem set is due before 8:00 on Friday October 1, 2021. No late assignments will be accepted.
- Total available points for this homework is 100.

Question 1 (50 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:

```
# Problem 1
#######################

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.

   Answer:

```r
n <- 25   ## sample size
SAMMEAN <- mean(y)    ## calculate sample mean
SAMSD <- sd(y)   ## calculate sample standard deviation
MARGIN <- qt(0.95,df=n-1)*SAMSD/sqrt(n)
LL <- SAMMEAN - MARGIN ## Lower Limit
LL

HL <- SAMMEAN + MARGIN ## Upper Limit
HL
```

Confidence interval for the average student IQ: lower Limit= 93.96, Upper Limit = 102.92

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 $\alpha = 0.05$.

   Step 1: Assumptions: the type of data is quantitative data; the sample is generated by random sampling; the population is distributed normally.

   Step 2: state hypothesis: Null hypothesis: the average student IQ in the school is lower than or equal to the average IQ score (100), (H0 100)
   Alternative hypothesis: Ha ¿ 100

```r
SAMMEAN  ## sample mean
IQMEAN <- 100  ## average IQ score among all the schools in the country
SAMSD  ## standard deviation of the sample
SAMSE <- SAMSD/sqrt(n)      #sample standard error

H0 <- IQMEAN
```

   Step 3: Calculate a test statistics:

```r
TS <- (SAMMEAN - H0)/SAMSE
TS
```

   report TS = -0.1191488

   Step 4: P-value:

```r
ZS <- (SAMMEAN-H0)/2.618    ## Z-score
ZS
PV <- 1-pnorm(-abs(-0.596))
PV
```

   Report P-value = 0.72, larger than 0.05
   or

```r
t.test(y, mu = 100, alternative = "greater")
```

Report P-value = 0.72, larger than 0.05
Step 5: Draw a conclusion
The p-value is higher than 0.05, so we cannot deny the null hypothesis that the average IQ of students in this school is lower than or equal to the average IQ of students in all schools in this country.

Question 2 (50 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*

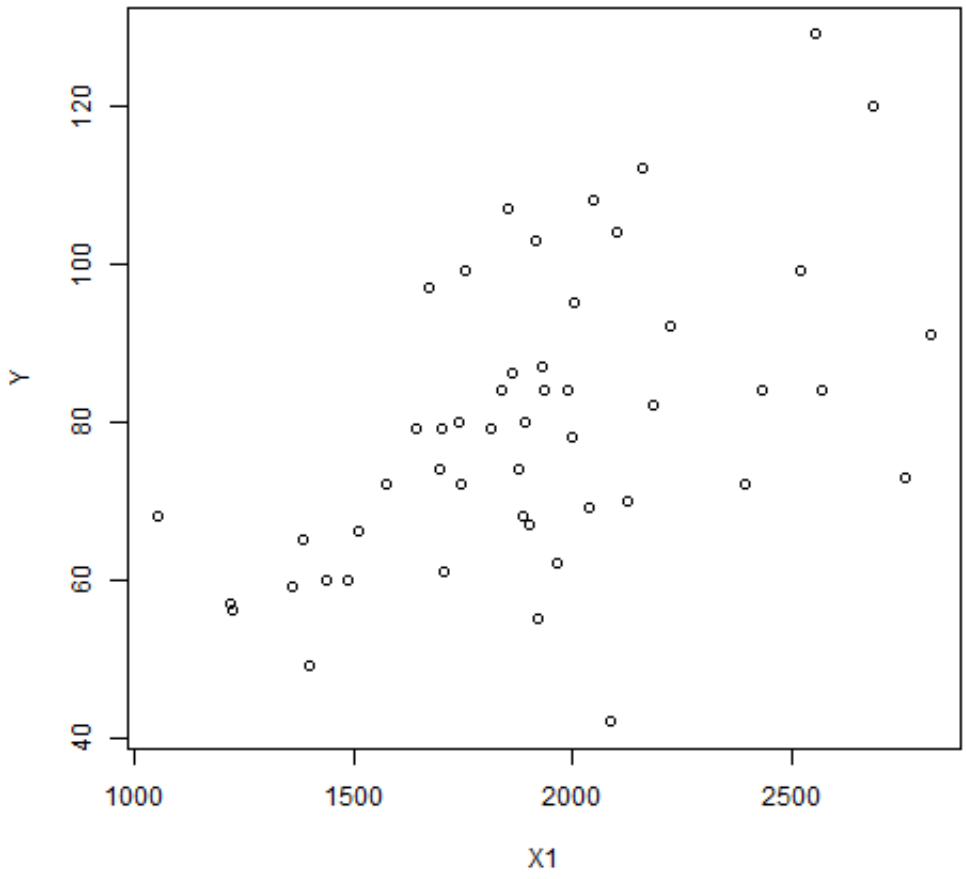
Explore the `expenditure` data set and import data into `R`.

```
#######################
# Problem 2
#######################

expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2021/main/datasets/expenditure.txt", header=T)
## this link does not work. I also tried read.csv() but all columns
# fall into one, so I loaded the local file


expenditure <- read.table("C:/Users/Caesar/Documents/GitHub/StatsI_Fall2022/datasets/expenditure.txt")


# 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.


## Check the structure, class and type of data
str(expenditure) ## The Structure of data is a data frame of character vectors
class(expenditure) ## The Class of data is data frame
typeof(expenditure) ## The type of data is list.
head(expenditure) ## The first row should be removed

## change first row into headers
names(expenditure) <- expenditure[1, ]
expenditure <- expenditure[-1, ]
head(expenditure)


## save as csv. file
write.csv(expenditure, "C:/Users/Caesar/Documents/GitHub/StatsI_Fall2022/
```

```
    problemSets/PS01/My_Answers/exenditure.csv")

expenditure_csv <- read.csv("C:/Users/Caesar/Documents/GitHub/StatsI_Fall2022/
    problemSets/PS01/My_Answers/exenditure.csv")
```

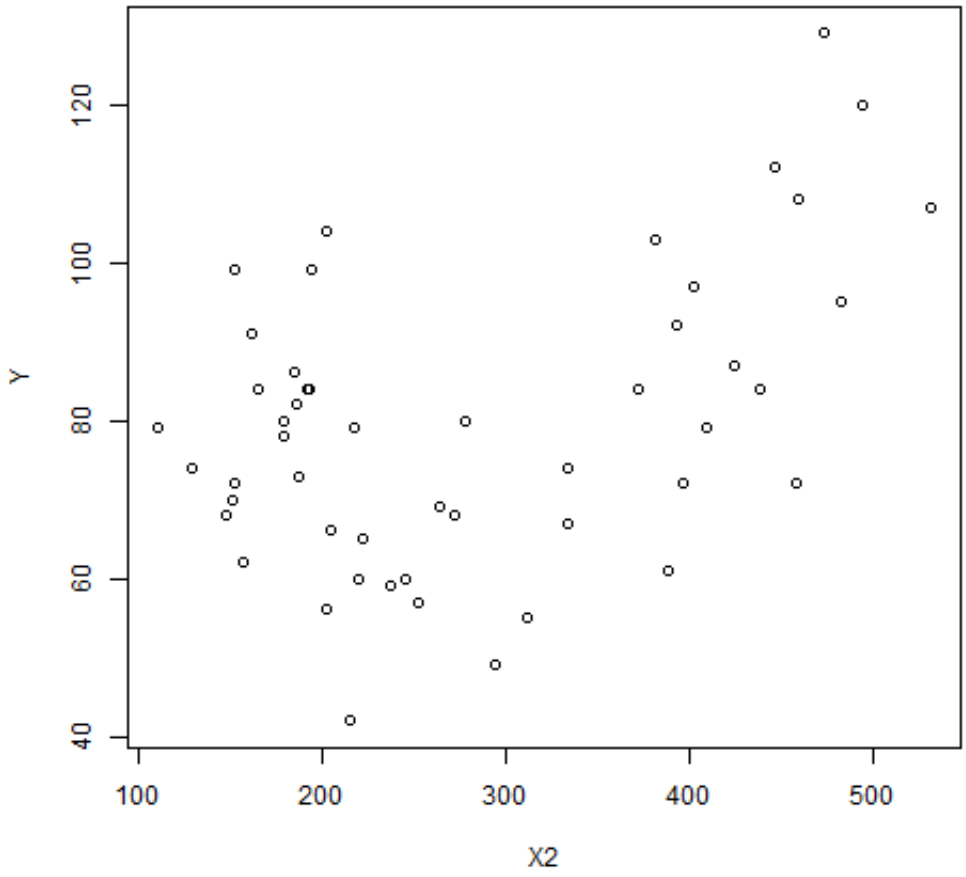
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?

```
png("Y ~ X1.png")
plot(Y ~ X1, data = expenditure_csv)
# Preliminary analysis of the graph: When X1 increases, we expect to see
    an
# increase in Y at the same time.
dev.off()
REGYX1 <- lm(Y ~ X1, data = expenditure_csv)
summary(REGYX1)
```



Preliminary analysis of the graph: When X1 increases, we expect to see an increase in Y at the same time. The coefficient of X1 is 0.025, the intercept is 32.546, which means when X1, per capita personal income in state increases by 1 US dollar, we expect to see an increase in Y, per capita expenditure on shelters/housing assistance in state, by 0.025 US dollar; When per capita personal income in state is 0 dollar, the per capita expenditure on shelters/housing assistance in state is 32.546 US dollar. The p value of the coefficient of X1 is 7.08e-05, which is smaller than 0.001. Therefore, We can reject the null hypothesis that there is no correlation between Y and X1 at 0.001 level.

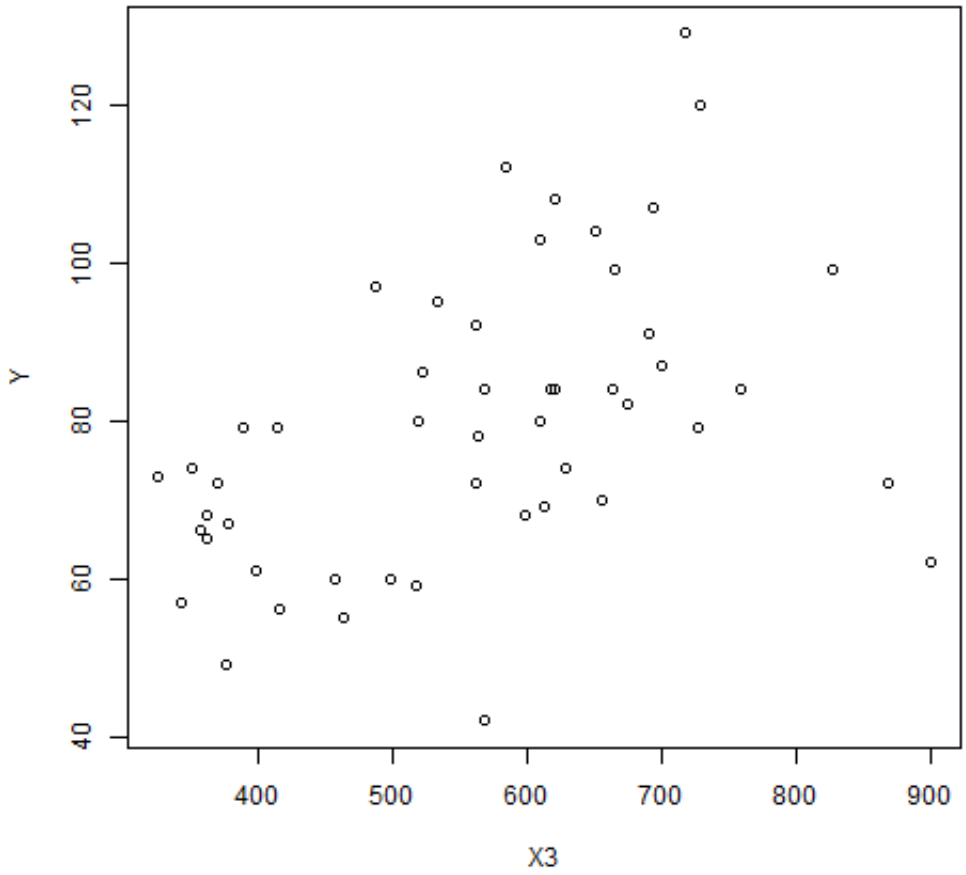
```
png("Y ~ X2.png")
plot(Y ~ X2, data = expenditure_csv)
# Preliminary analysis of the graph: When X2 increases, we expect to see
    an
# increase in Y at the same time.
dev.off()
REGYX2 <- lm(Y ~ X2, data = expenditure_csv)
summary(REGYX2)
```



Preliminary analysis of the graph: When X2 increases, we expect to see an increase in Y at the same time.

The coefficient of X2 is 0.070, the intercept is 59.761, which means when X2, the number of residents per 100,000 that are "financially insecure" in state increases by 1, we expect to see an increase in Y, per capita expenditure on shelters/housing assistance in state, by 0.070 US dollar; When Number of residents per 100,000 that are "financially insecure" in state is 0, the per capita expenditure on shelters/housing assistance in state is 59.761 US dollar. The p value of the coefficient of X2 is 0.001, which is smaller than 0.01. Therefore, We can reject the null hypothesis that there is no correlation between Y and X2 at 0.01 level.

```
png("Y ~ X3.png")
plot(Y ~ X3, data = expenditure_csv)
# Preliminary analysis of the graph: When X3 increases, we expect to see
    Y
# increase at the same time.
dev.off()
REGYX3 <- lm(Y ~ X3, data = expenditure_csv)
summary(REGYX3)
```

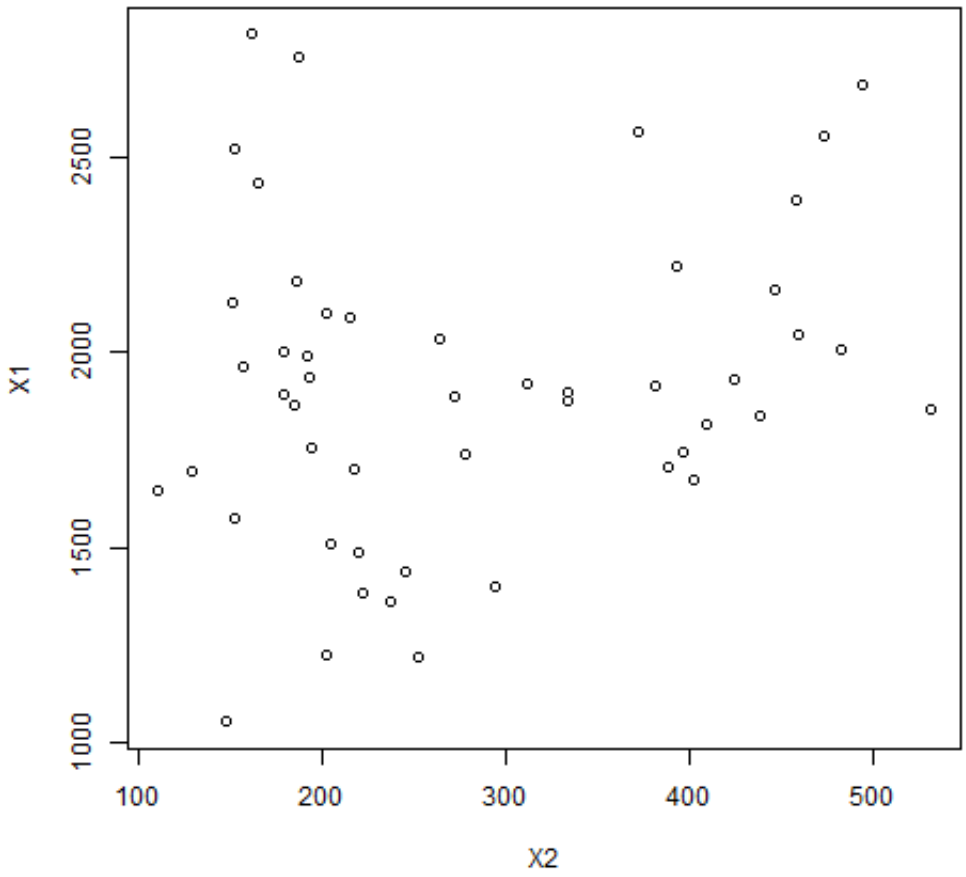


Preliminary analysis of the graph: When X3 increases, we expect to see Y increase at the same time.

The coefficient of X3 is 0.059, the intercept is 46,306, which means when X3,the number of people per thousand residing in urban areas in state increases by 1, we expect to see an increase in Y, per capita expenditure on shelters/housing assistance in state, by 0.059 US dollar; When Number of people per thousand residing in urban areas in state is 0, the per capita expenditure on shelters/housing assistance in state is 46.306 US dollar. The p value of the coefficient of X3 is 0.000695, which is smaller than 0.001. Therefore, We can reject the null hypothesis that there is no correlation between Y and X3 at 0.001 level.

```
png("X1 ~ X2.png")
plot(X1 ~ X2, data = expenditure_csv)
# Preliminary analysis of the graph: there is no obvious correlation
# between X1 and X2.
dev.off()
REGX1X2 <- lm(X1 ~ X2, data = expenditure_csv)
```

```
summary(REGX1X2)
```

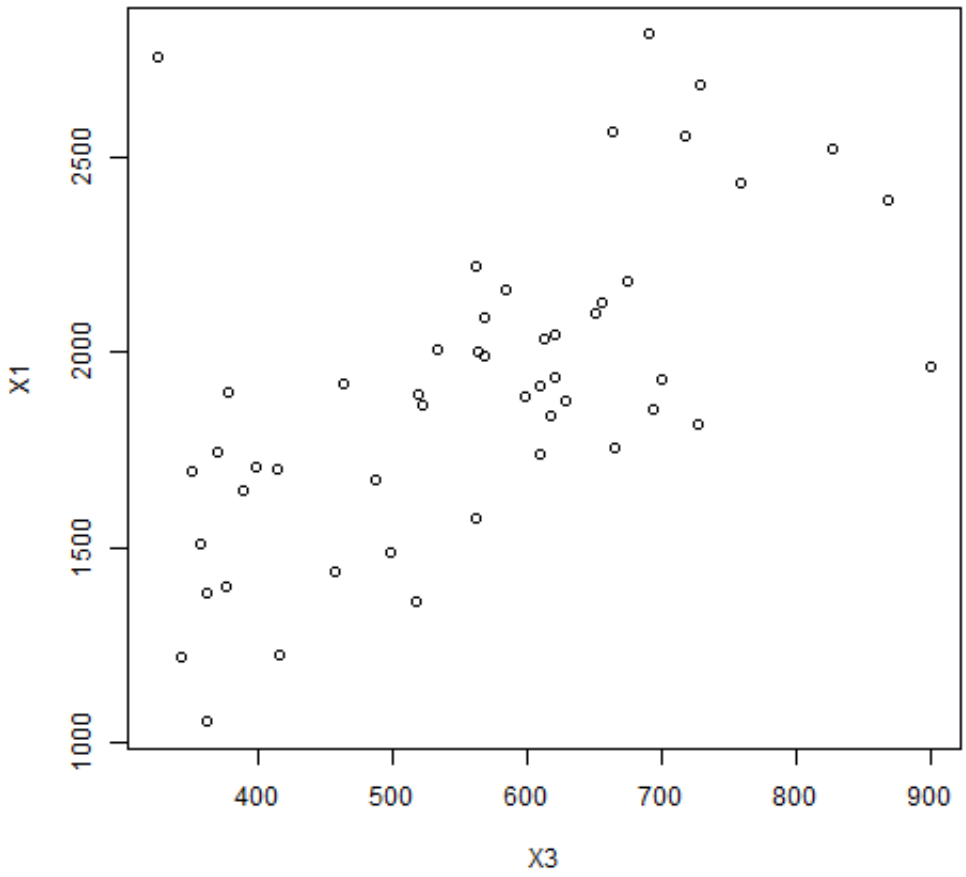


Preliminary analysis of the graph: there is no obvious correlation between X1 and X2.

The coefficient of X2 is 0.696, the intercept is 1715.655, which means when X2, number of residents per 100,000 that are "financially insecure" in state increases by 1, we expect to see an increase in X1, per capita personal income in state, by 0.696 US dollar; When the number of residents per 100,000 that are "financially insecure" in state is 0, the per capita personal income in state is 1715.655 US dollar. The p value of the coefficient of X2 is 0.152, which is larger than 0.05. Therefore, We cannot reject the null hypothesis that there is no correlation between X1 and X2.

```
png("X1 ~ X3.png")
plot(X1 ~ X3, data = expenditure_csv)
# Preliminary analysis of the graph: when X3 increases, X1 is expected to
# increase at the same time.
dev.off()
REGX1X3 <- lm(X1 ~ X3, data = expenditure_csv)
```

```
summary(REGX1X3)
```

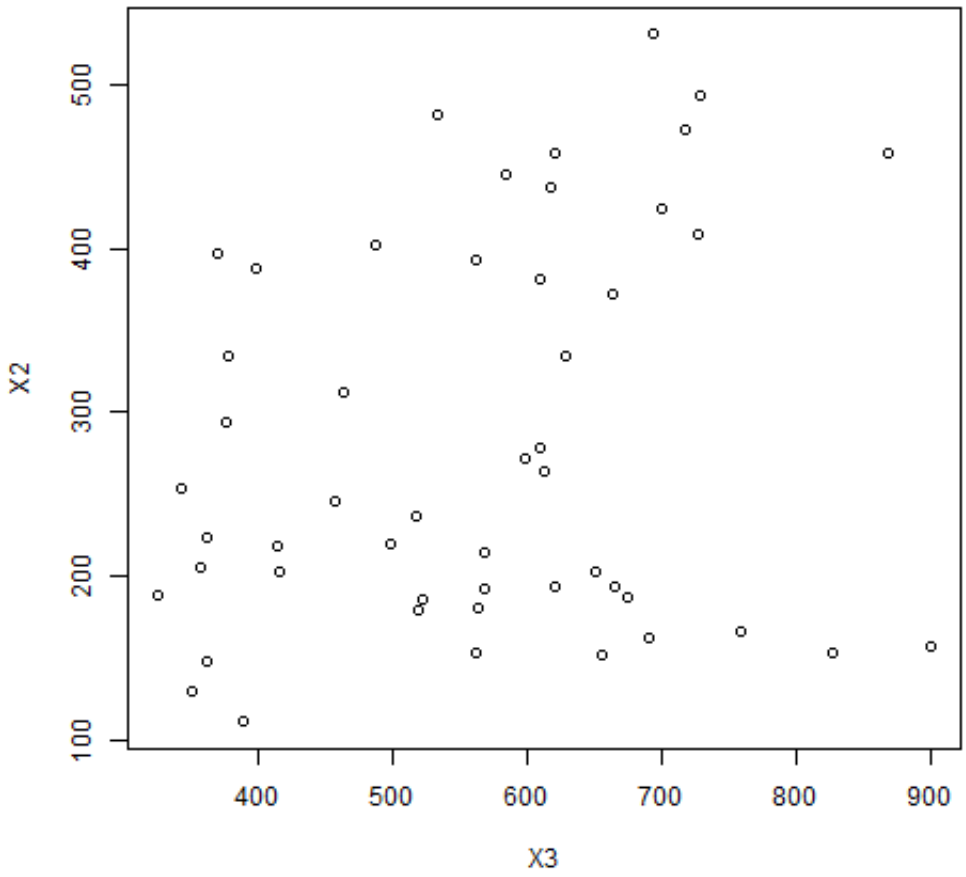


Preliminary analysis of the graph: when X3 increases, X1 is expected to increase at the same time.

The coefficient of X3 is 1.643, the intercept is 988.947, which means when X3, the number of people per thousand residing in urban areas in state increases by 1, we expect to see an increase in X1, per capita personal income in state, by 1.643 US dollar; When the Number of people per thousand residing in urban areas in state is 0, the per capita personal income in state is 988.947 US dollar. The p value of the coefficient of X3 is 5.13e-06, which is smaller than 0.001. Therefore, We can reject the null hypothesis that there is no correlation between X1 and X2 at 0.001 level.

```
png("X2 ~ X3.png")
plot(X2 ~ X3, data = expenditure_csv)
# Preliminary analysis of the graph: There is no obvious correlation
# between X2 and X3.
dev.off()
```

```r
REGX2X3 <- lm(X2 ~ X3, data = expenditure_csv)
summary(REGX2X3)
```

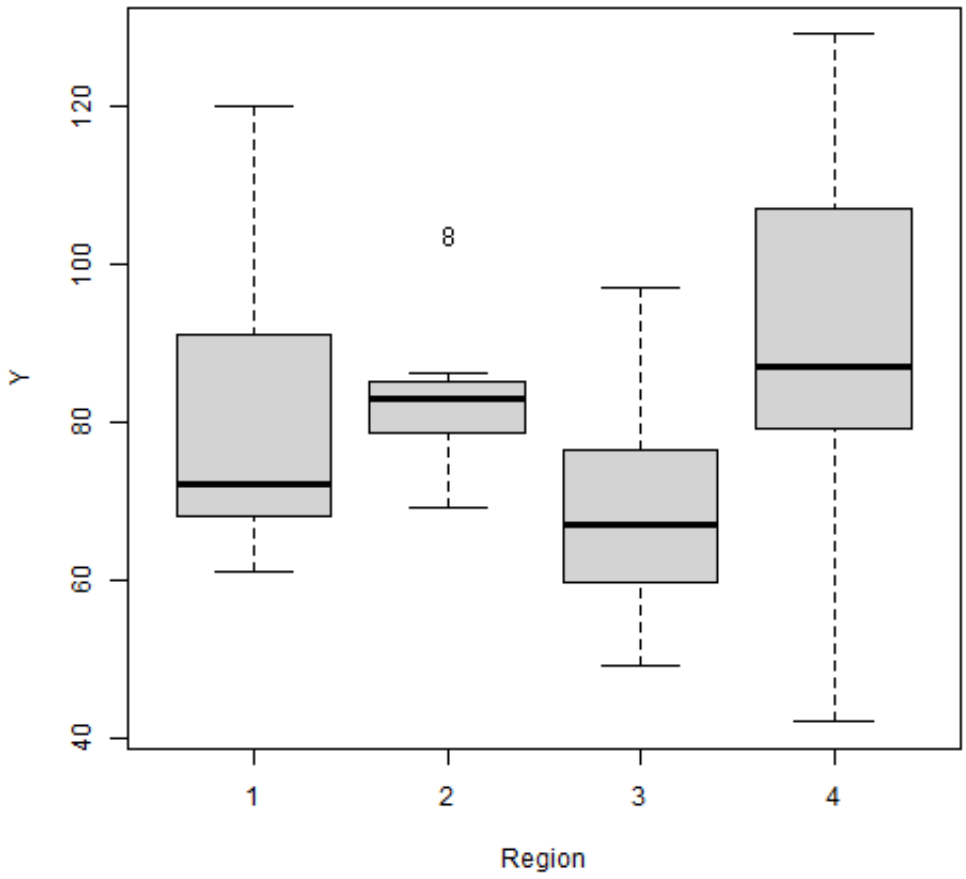


Preliminary analysis of the graph: There is no obvious correlation between X2 and X3.

The coefficient of X3 is 0.180, the intercept is 180.609, which means when X3, the number of people per thousand residing in urban areas in state increases by 1, we expect to see an increase in X2, Number of residents per 100,000 that are "financially insecure" in state by 0.180; When the number of people per thousand residing in urban areas in state is 0, the number of residents per 100,000 that are "financially insecure" in state is 180.609. The p value of the coefficient of X3 is 0.123, which is larger than 0.05. Therefore, We cannot reject the null hypothesis that there is no correlation between X2 and X3.

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

```
png("Y ~ Region.png")
boxplot(Y ~ Region, data = expenditure_csv)
## on average, West has the highest per capita expenditure on housing
    assistance.
dev.off()
```

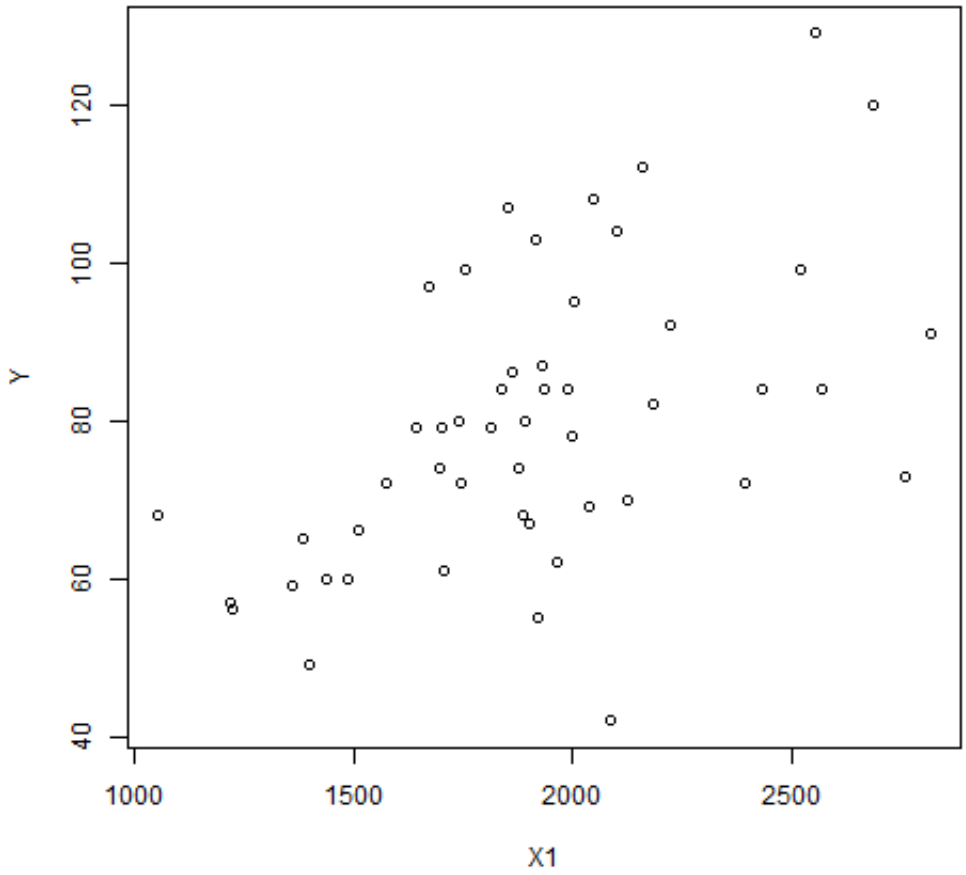


On average, West has the highest per capita expenditure on housing assistance.

- 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.

```
png("Y ~ X1.png")
plot(Y ~ X1, data = expenditure_csv)
# Preliminary analysis of the graph: When X1 increases, we expect to see
    an
# increase in Y at the same time.
```

```
dev.off()
REGYX1 <- lm(Y ~ X1, data = expenditure_csv)
summary(REGYX1)
```

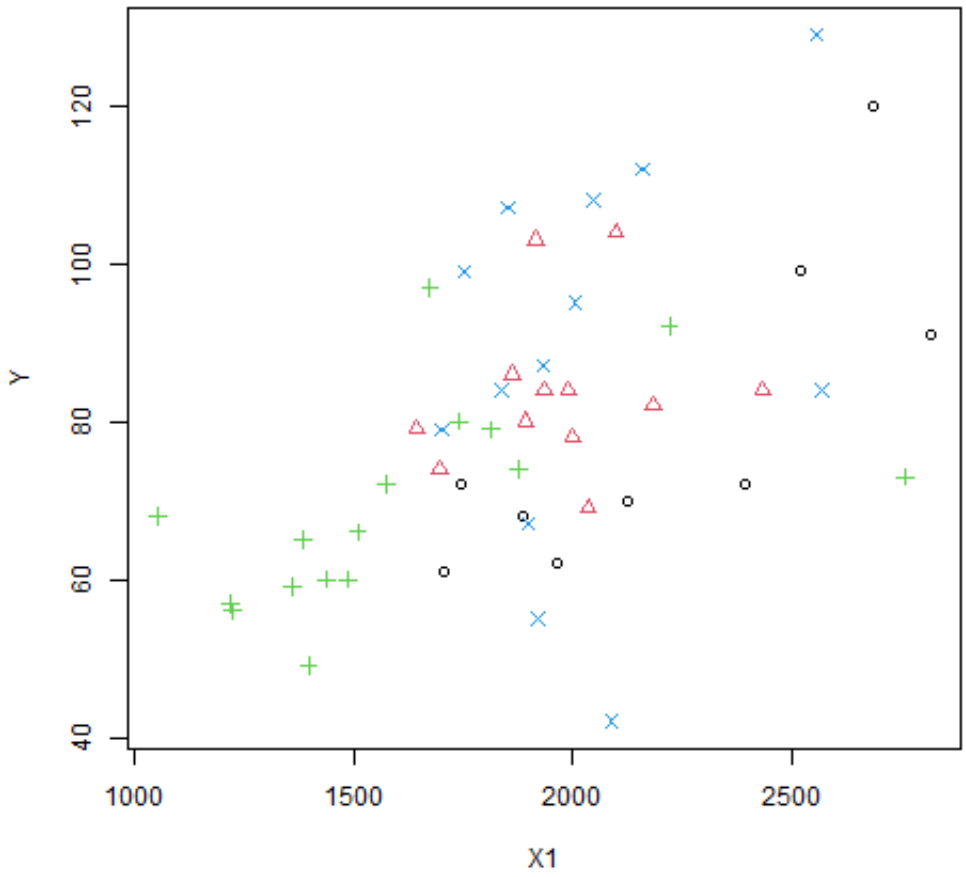


Preliminary analysis of the graph: When X1 increases, we expect to see an increase in Y at the same time.

The coefficient of X1 is 0.025, the intercept is 32.546, which means when X1, per capita personal income in state increases by 1 US dollar, we expect to see an increase in Y, per capita expenditure on shelters/housing assistance in state, by 0.025 US dollar; When per capita personal income in state is 0 dollar, the per capita expenditure on shelters/housing assistance in state is 32.546 US dollar. The p value of the coefficient of X1 is 7.08e-05, which is smaller than 0.001. Therefore, We can reject the null hypothesis that there is no correlation between Y and X1 at 0.001 level.

```
png("Y ~ X1, colour, symbol.png")
plot(Y ~ X1, data = expenditure_csv, col = Region, pch = Region)

```

```
dev.off()
REGYX1 <- lm(Y ~ X1, data = expenditure_csv)
summary(REGYX1)
REGYX1REG <- lm(Y ~ X1 + Region, data = expenditure_csv)
summary(REGYX1REG)
```



The coefficient of X1 is 0.027, the coefficient of Region is 3.333, which means when X1, per capita personal income in state increases by 1 US dollar, we expect to see an increase in Y, per capita expenditure on shelters/housing assistance in state, by 0.027 US dollar; when the region is different, we expect to see an average difference in the per capita expenditure on shelters/housing assistance in state of 3.333 US dollar. The p value of the coefficient of X1 is 2.77e-05, which is smaller than 0.001. Therefore, We can reject the null hypothesis that there is no correlation between Y and X1 at 0.001 level.

The p value of the coefficient of Region is 0.128, which is larger than 0.05. Therefore, we cannot reject the null hypothesis that there is no correlation between Y and Region.