Problem Set 1

Applied Stats/Quant Methods 1

Due: October 3, 2021

Linette Lim

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:

```
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.

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 $\alpha = 0.05$.

- Answer for Question 1.1
  First, load packages such as Tidyverse and set the working directory.
  Assign the dataset of 25 students' IQ scores to an object, "y".
  Next, check the data using functions like summary(), mean(), sd(). We get:

```
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  69.00   89.00   98.00   98.44  110.00  126.00
> sd <- sd(y)
> sd
```

```
[1] 13.09287
> y.bar <- mean(y)
> y.bar
[1] 98.44
> se <- sd(y)/sqrt(length(y))
> se
[1] 2.618575
```

Check the distribution of the data and calculate the lower and upper ends of the 90 percent confidence interval using qnorm():

```r
CI_lower <- qnorm(0.05,
                  mean = mean(y),
                  sd = (sd(y)/sqrt(length(y)))
)

CI_upper <- qnorm(0.95,
                  mean = mean(y),
                  sd = (sd(y)/sqrt(length(y)))
)
```

We get:

```
    Lower    Upper
 94.13283 102.7472
```

We can check this answer with t.test():

```r

# A way to check the working
t.test(y, conf.level = 0.9, alternative = "two.sided")
```

We get:

```
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
```

Therefore we can conclude that with 90 percent confidence, the average student IQ in the school is between 94 and 103.

- Answer for Question 1.2
  We are told the average IQ score among all the schools in the country is 100. Therefore, let mu = 100. The null and alternative hypotheses are formulated as follows:

```r
# H0: y.bar <= mu
# Ha: y.bar > mu
```

Since the alternative hypothesis is $y.bar > 100$, we compute the one-sided p-value as follows:

```r
upper <- pnorm(y.bar, mean = 100, sd = se, lower.tail = FALSE)
```

The p-value calculated is:

```
[1] 0.7243269
```

We can check this answer using the z score and pnorm() formula:

```r
z.score <- (y.bar - 100) / se
pnorm(z.score, lower.tail = FALSE)
```

We get the same answer. The p-value is 0.7243269. This is larger than the significance level alpha = 0.05, so we cannot reject the null hypothesis. There is not enough statistical evidence to support Ha, that is, that the average IQ in the school is higher than the average IQ score (100) among all the schools in the 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`.

- 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.

- Answer for Question 2.1

First we import the data:

```
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
    StatsI_Fall2022/main/datasets/expenditure.txt", header=T)
```

Next, we plot each pair of variables and generate 12 graphs using the following formula:

```
pdf("plot1_Y_X1.pdf")
plot(expenditure$X1, expenditure$Y)
dev.off()
```

We plot the 12 graphs (see Figures 1 - 12 from page 7 to page 18) to examine the correlations, using Y, followed by X1, then X2, and finally X3, as the response variable.

Fig 1 shows a weak, positive, linear relationship between Y, per capita expenditure on shelters/housing assistance in state, and X1, the per capita personal income in state. There are a number of outliers in the data that need to be investigated.

Fig 2 shows a weak, positive, linear relationship between Y, per capita expenditure on shelters/housing assistance in state, and X2, the number of residents per 100,000 that are "financially insecure" in state. There are a number of outliers in the data that need to be investigated.

Fig 3 shows a moderately strong, positive, linear relationship between Y, per capita expenditure on shelters/housing assistance in state, and X3, the number of people per thousand residing in urban areas in state. This suggests it is likely that the greater the number of urban residents in state, the greater the expenditure on shelters/housing assistance in state. There are a few outliers in the data.

Fig 4 shows a weak, positive, linear relationship between Y, per capita expenditure on shelters/housing assistance in state, and X1, the per capita personal income in state. There are a number of outliers in the data that needs to be investigated.

Fig 5. The points on the scatter plot seem to be scattered randomly, suggesting that there is no relationship or no correlation between X2, the number of residents per 100,000 that are "financially insecure" in state, and X1, the per capita personal income in state.

Fig 6 shows a moderately strong, positive, linear association between X3, the number of people per thousand residing in urban areas in state, and X1, the per capita personal income in state. This suggests it is likely that the greater the number of people living in urban areas in state, the greater the per capita personal income in state. There are a few outliers that need to be investigated.

Fig 7 shows a weak, positive, linear relationship between Y, per capita expenditure on shelters/housing assistance in state, and X2, the number of residents per 100,000 that are "financially insecure" in state. There are a number of outliers in the data that

need to be investigated.

Fig 8. The points on the scatter plot seem to be scattered randomly, suggesting that there is no relationship or no correlation between X1, the per capita personal income in state, and X2, the number of residents per 100,000 that are "financially insecure" in state.

Fig 9. The points on the scatter plot seem to be scattered randomly, suggesting there is no relationship or no correlation between X3, the number of people per thousand residing in urban areas in state, and X2, the number of residents per 100,000 that are "financially insecure" in state.

Fig 10 shows a weak, positive, linear relationship between Y, per capita expenditure on shelters/housing assistance in state, and X3, the number of people per thousand residing in urban areas in state. There are a number of outliers in the data.

Fig 11 shows a moderately strong, positive, linear association between X1, the per capita personal income in state, and X3, the number of people per thousand residing in urban areas in state. This suggests it is likely that the greater the per capita personal income in state, the greater the number of people living in urban areas in state. There are a number of outliers in the data that need to be investigated.

Fig 12. The points on the scatter plot seem to be scattered randomly, suggesting there is no relationship or no correlation between X2, Number of residents per 100,000 that are "financially insecure" in state, and X3, the number of people per thousand residing in urban areas in state.

Graphs for Question 2.1

Figure 1: Relationship between Y and X1.

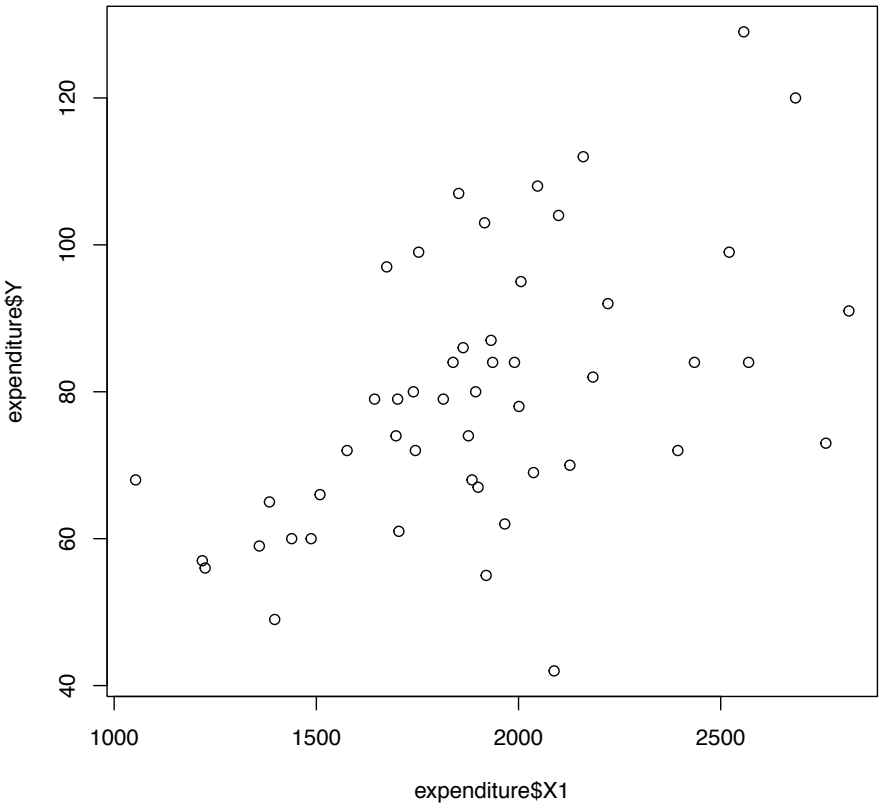


Figure 2: Relationship between Y and X2.

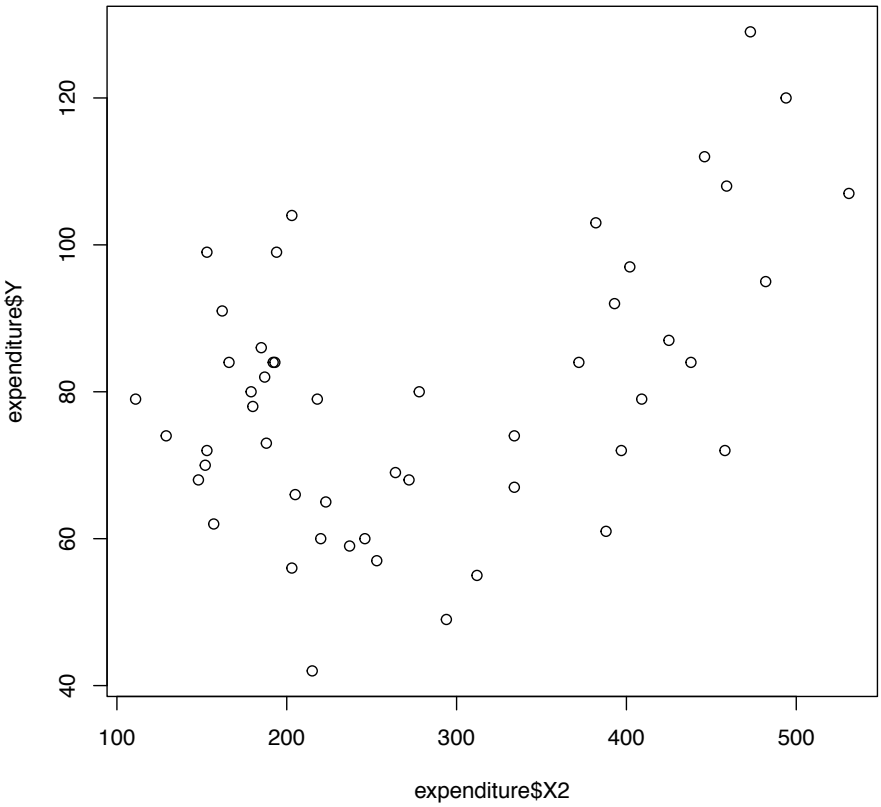


Figure 3: Relationship between Y and X3.

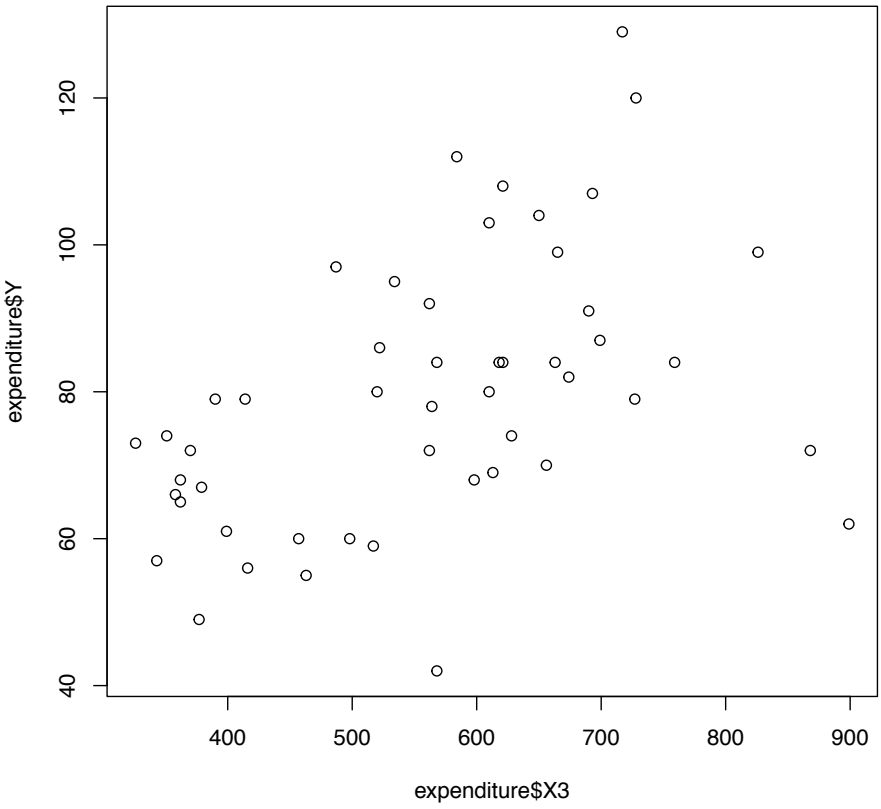


Figure 4: Relationship between X1 and Y.

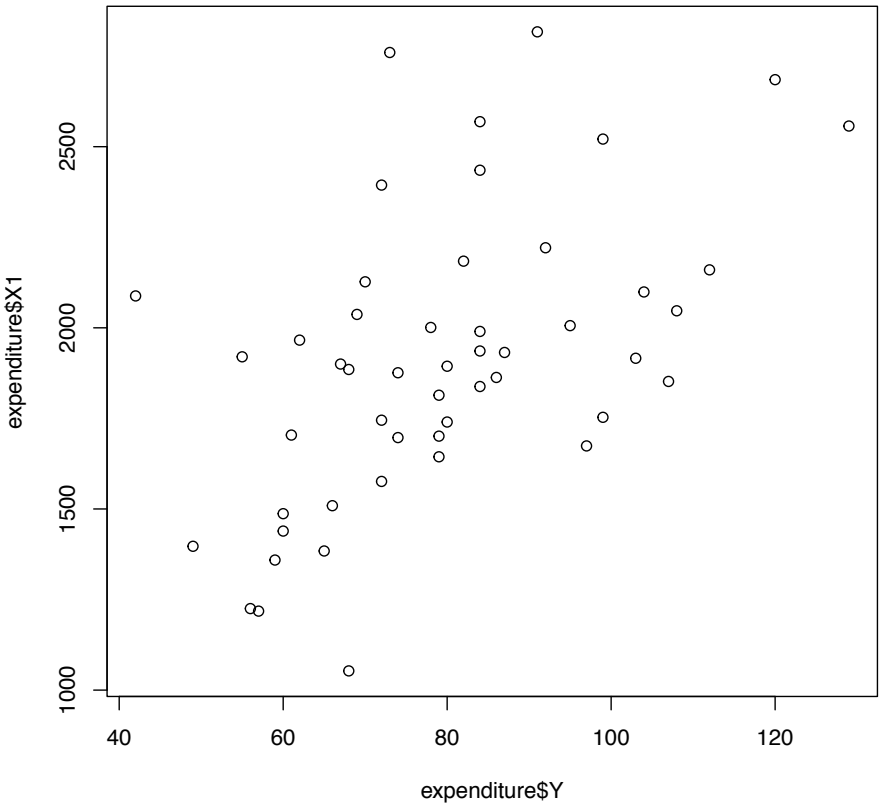


Figure 5: Relationship between X1 and X2.

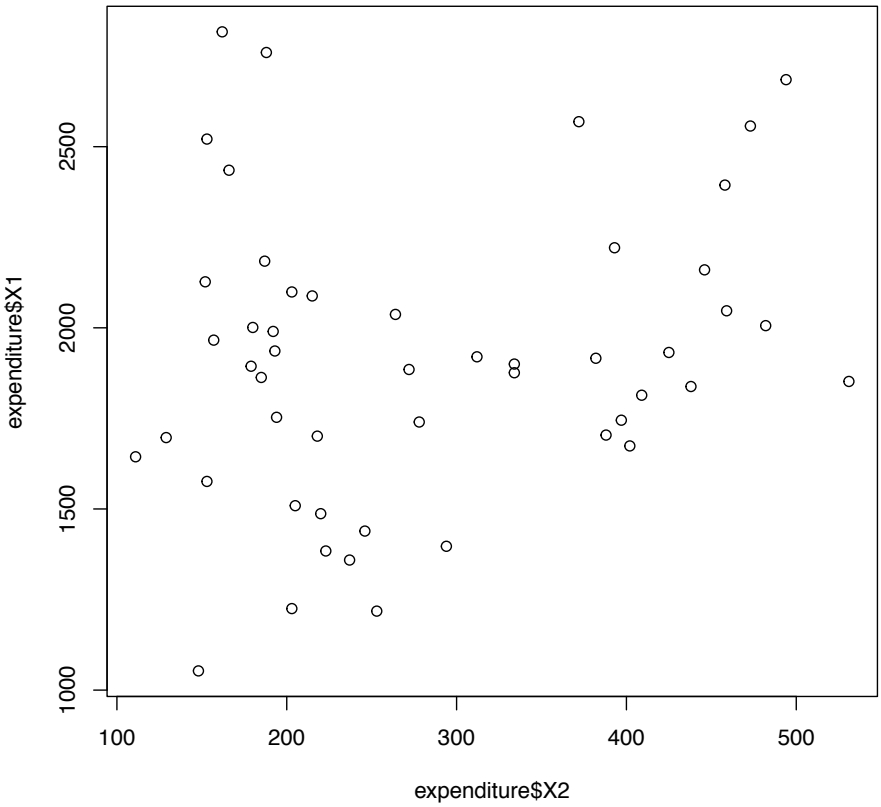


Figure 6: Relationship between X1 and X3.

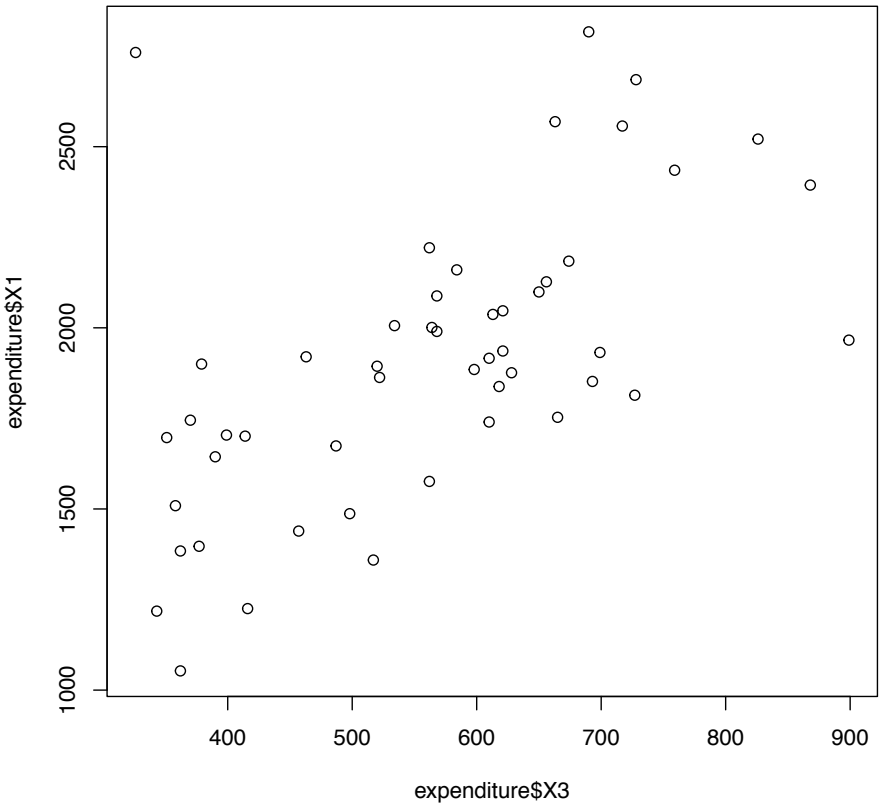


Figure 7: Relationship between X2 and Y.

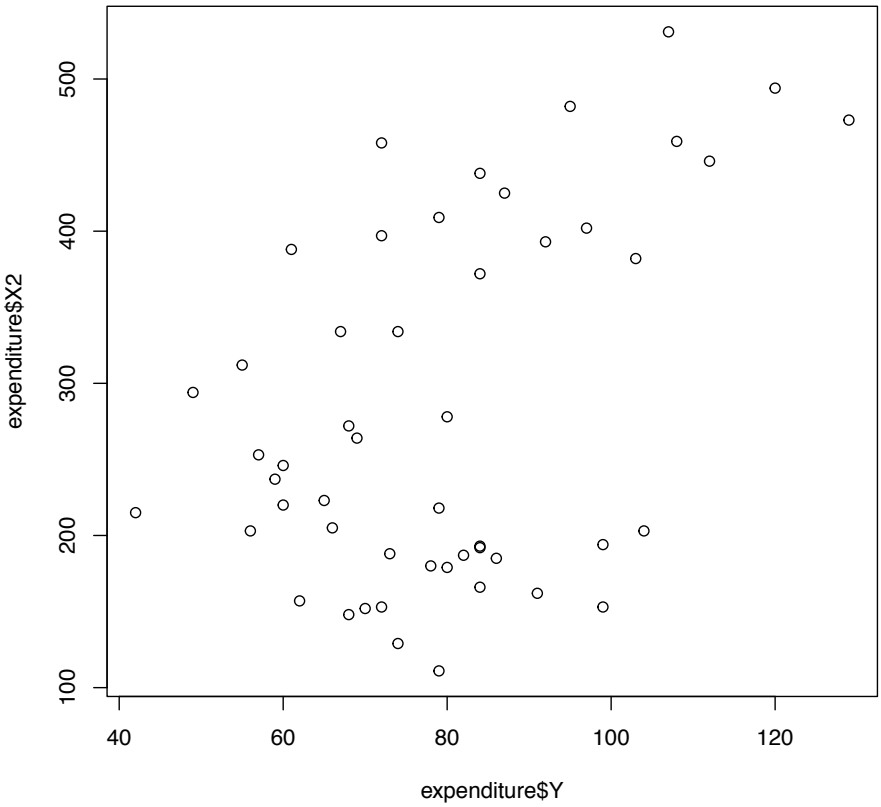


Figure 8: Relationship between X2 and X1.

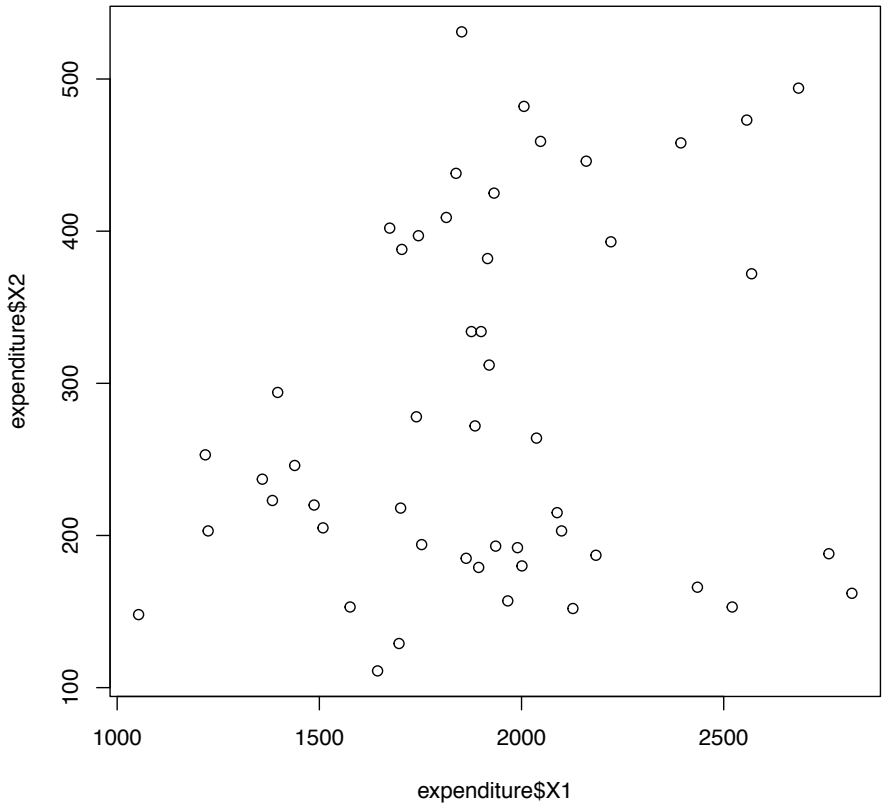


Figure 9: Relationship between X2 and X3.

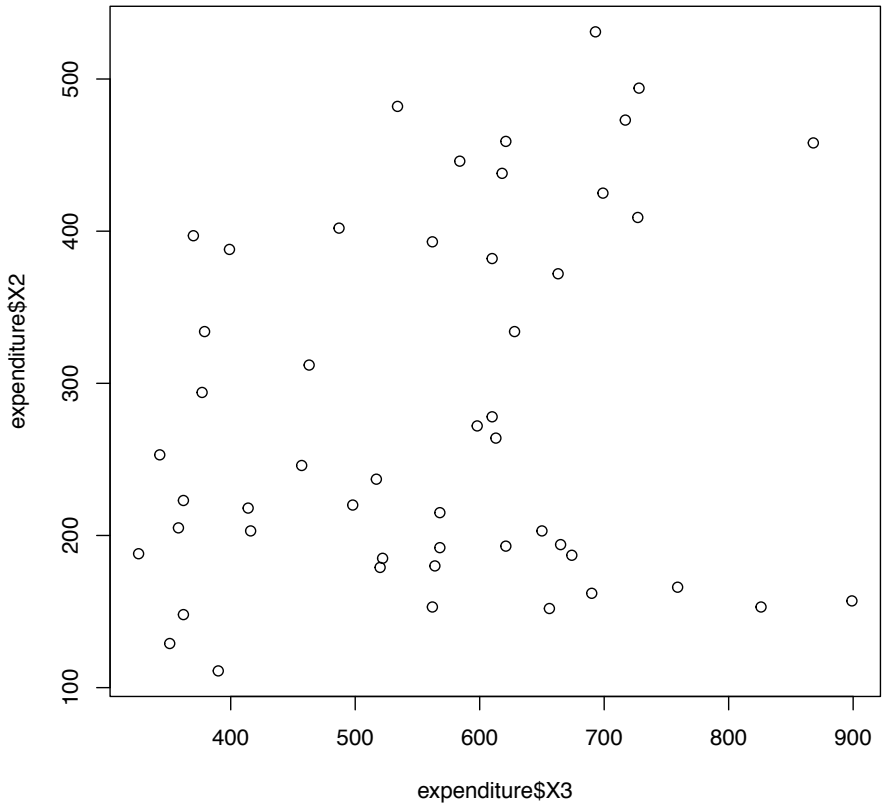


Figure 10: Relationship between X3 and Y.

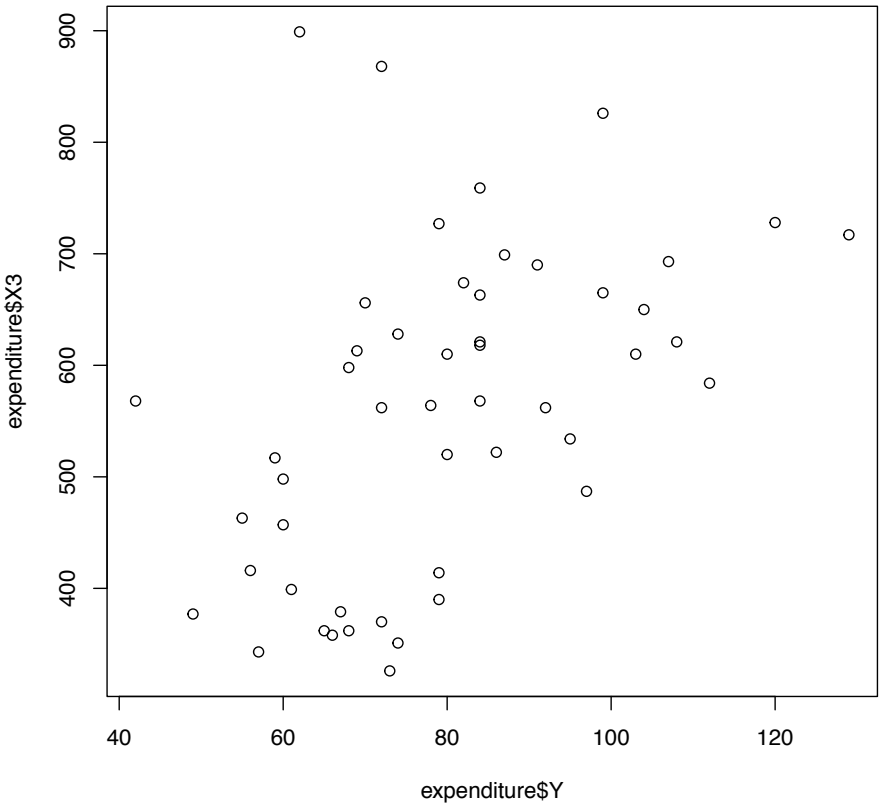


Figure 11: Relationship between X3 and X1.

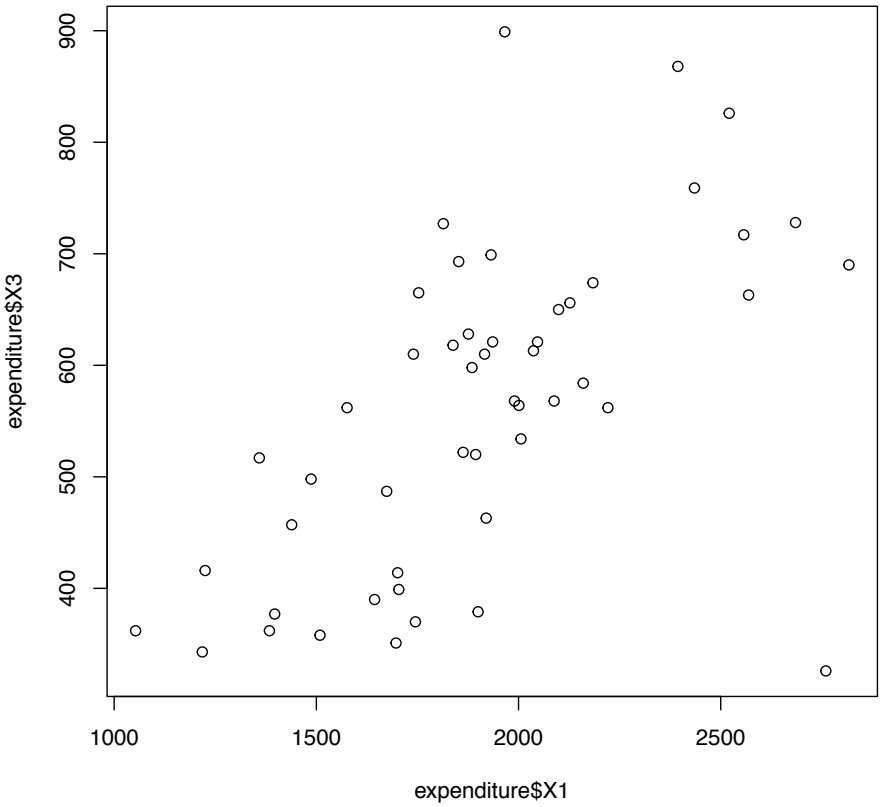
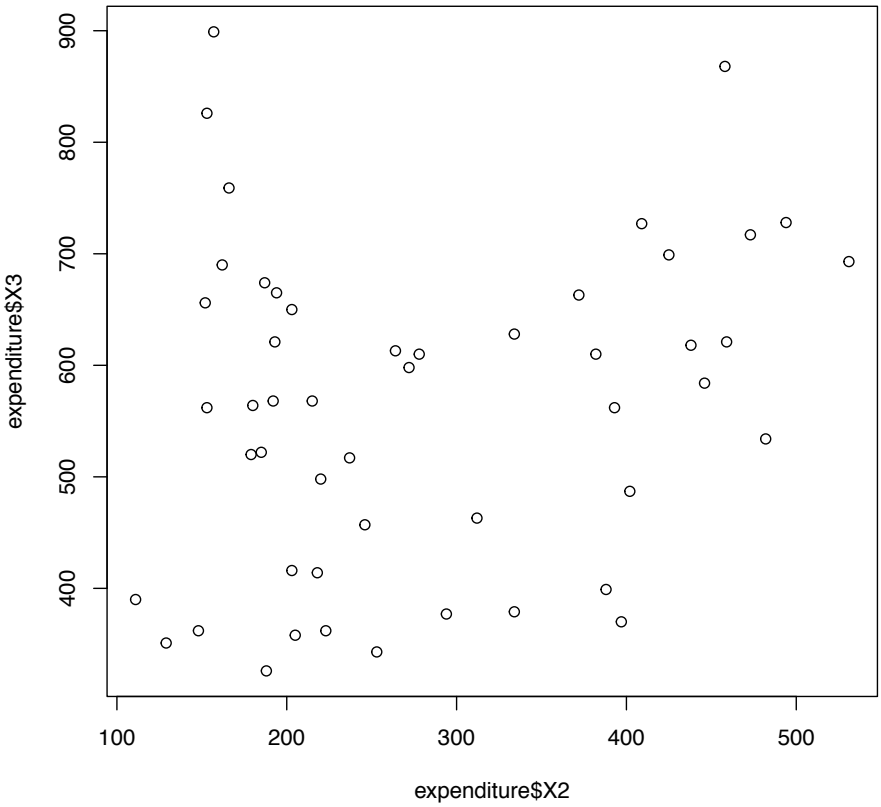


Figure 12: Relationship between X3 and X2.



- Answer for Question 2.2

  We begin by wrangling the data. We want to find the average per capita expenditure on housing assistance for each of the four regions.

```
Region_1 <- expenditure[expenditure$Region == "1",]
Region_2 <- expenditure[expenditure$Region == "2",]
Region_3 <- expenditure[expenditure$Region == "3",]
Region_4 <- expenditure[expenditure$Region == "4",]

Region_1_Expend <- mean(Region_1$Y)
Region_2_Expend <- mean(Region_2$Y)
Region_3_Expend <- mean(Region_3$Y)
Region_4_Expend <- mean(Region_4$Y)
```

  We get the following:

```
> Region_1_Expend
[1] 79.44444
> Region_2_Expend
[1] 83.91667
> Region_3_Expend
[1] 69.1875
> Region_4_Expend
[1] 88.30769
```

  Based on r's calculations, we see that on average Region 4 has the highest per capita expenditure on housing assistance, at 88.3 (assumed to be) USD. Let us use a boxplot to check the minimum and maximum points, as well as the median, and the first and third quartile for the four regions. The formula is as follows:

```
library(ggplot2)
expenditure %>%
  filter(Region %in% c("1", "2", "3", "4")) %>%
  group_by(Region) %>%
  ggplot(aes(factor(Region), Y)) +
  geom_boxplot()
```

Figure 13: Per capita expenditure on housing assistance by Regions.

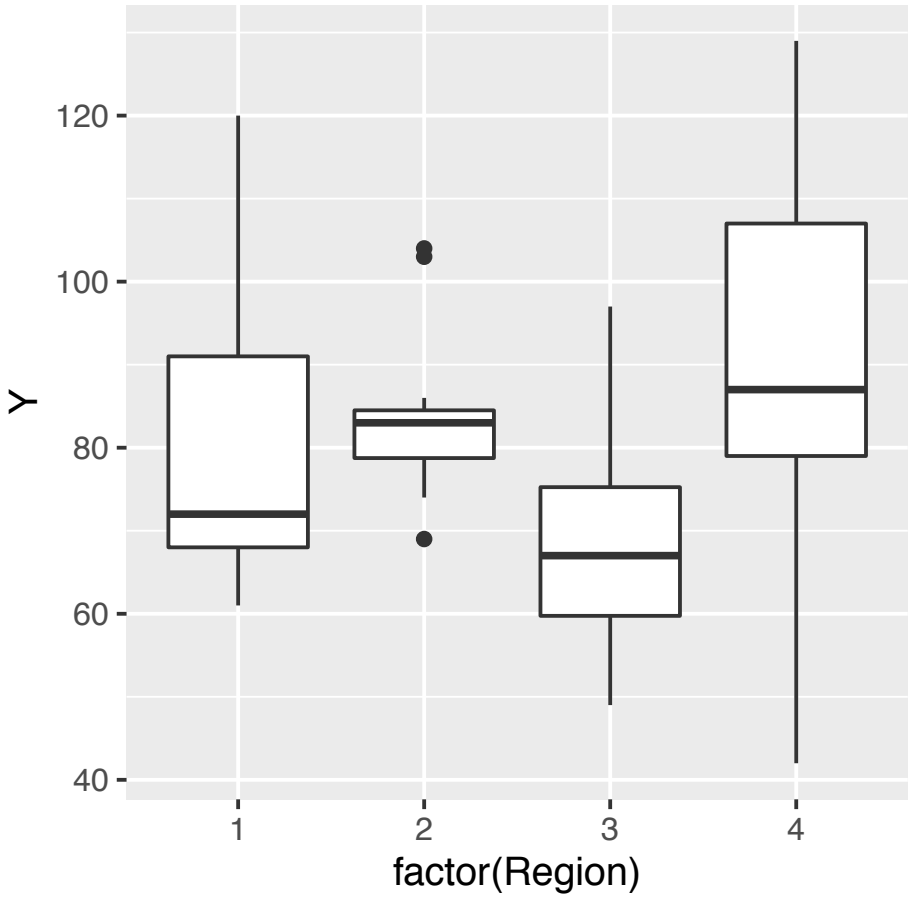


Fig 13 shows that the boxplot for Region 4 (West), is observably higher than that of the other three regions, suggesting that it has on average, the highest per capita expenditure on housing assistance. Looking at the boxplots, 75 percent of the states in the West have a per capita expenditure on housing assistance of between 79 (assumed to be) and 129 USD. This is compared to between 68 and 120 for Region 1 (North-east), between 60 and 98 for Region 3 (South), and between 79 and 85 for Region 2 (North Central). While Region 4 has on average, the highest per capita expenditure on housing assistance, it also has the longest boxplot and the longest whiskers, suggesting that the states within the region also have the largest disparity, or distribution, in the level of housing assistance expenditure per capita. Region 2 has the shortest boxplot and shortest whiskers, suggesting it has the smallest disparity in the level of housing assistance expenditure per capita. Region 2 is also interesting as it is the only region with outliers - it has three outliers.

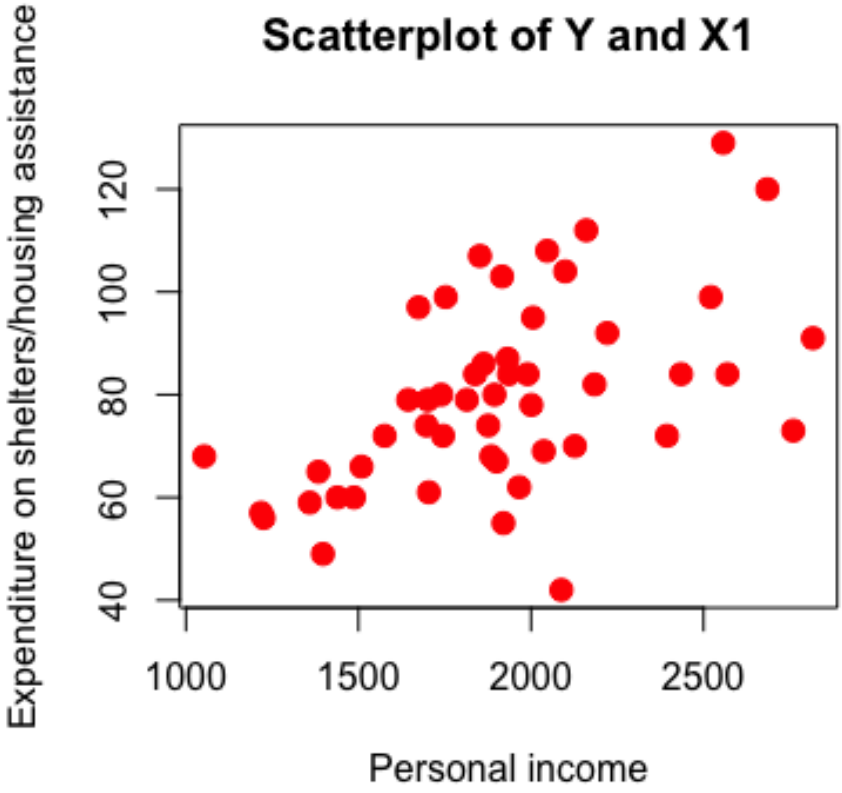
- Answer for Question 2.3

We replot Y and X1, this time with more bells and whistles, using the below code:

```
plot(expenditure$X1, expenditure$Y, col='red', pch=19, cex=1.3,
     xlab='Personal income', ylab='Expenditure on shelters/housing
   assistance', main='Scatterplot of Y and X1')
```

The scatterplot below (Fig 14) shows a weak to medium positive, linear correlation between Y, the per capita expenditure on shelters/housing assistance in state, and X1, the per capita personal income in state. This suggests there is a likelihood that with increasing personal capita personal income in state, there is increasing per capita expenditure on shelters/housing in state. There are a number of outliers in the data that need to be investigated.

Figure 14: Per capita expenditure on housing assistance by Regions.



Next, we plot the above graph with a third variable, Region. The r code is reproduced as follows:

```
pdf("plot_3_variables.pdf")
ggplot(data=expenditure, mapping = aes(x = X1, y = Y)) +
  geom_point(aes(color = factor(Region))) +
  theme_light() +
  scale_color_discrete(labels = c("Northeast", "North Central", "South",
    "West")) +
  ggtitle("Relationship between Expenditure on shelters and Personal
    income by Region") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Per capita personal income") +
  ylab("Per capita expenditure on shelters/housing")
dev.off()
```

From the graph (Fig 15), we can observe more clearly the differences between the four regions. There are generally lower levels of per capita personal income and per capita expenditure on shelters/housing in the South. The Northeast tends to have higher levels of per capita personal income, but overall, per capita expenditure on shelters/-housing is not particularly high, even though there are three states that are outliers in terms of shelter expenditure. The North Central lies somewhat in between the spread of the South and the Northeast. The points for the West look rather peculiar - with the majority of the data points concentrated between 1600 and 2100 on the x-axis. This seems to suggest there might be some form of basic income or minimum wage policy in many states in the West. In the West, unlike the other three regions, personal income levels do not seem to have a strong correlation with spending on housing/shelters.

Figure 15: Per capita expenditure on housing assistance by Regions.