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:

```
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.

   My answer: The 90 per cent confidence interval for the average student IQ in the school lies between 93.96 and 102.92. I found this using the following R code:

   t.test(y, conf.level = 0.9)

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 $\alpha = 0.05$.

   Part 2 Answer:

   The null hypothesis is that they are the same or lower- the alternative hypothesis is that they are higher than the average IQ

   100 is the average IQ score among all the schools in the country.

   We gained the following information from the function that enabled us to find the confidence interval:

   The mean of x is 98.44.

   t= 37.593 ,which is the test statistic

   p-value is less than 2.2e-16

   This p value is really low, so we would be very surprised if the data was significant.

   In order to run the hypothesis test, we found the standard deviation using the following function:

   sd(y)

   answer is 13.09287

   We then used this function to run the hypothesis test:

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

   We made the hypothesis test one-sided by inputting "greater" as the alternative hypothesis. The p value we get from the above hypothesis test is 0.7215 which is really large, meaning we cannot reject the null hypothesis that the mean for the school is less than or equal to 100,i.e. her students are not better than the national average as she'd hoped.

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.

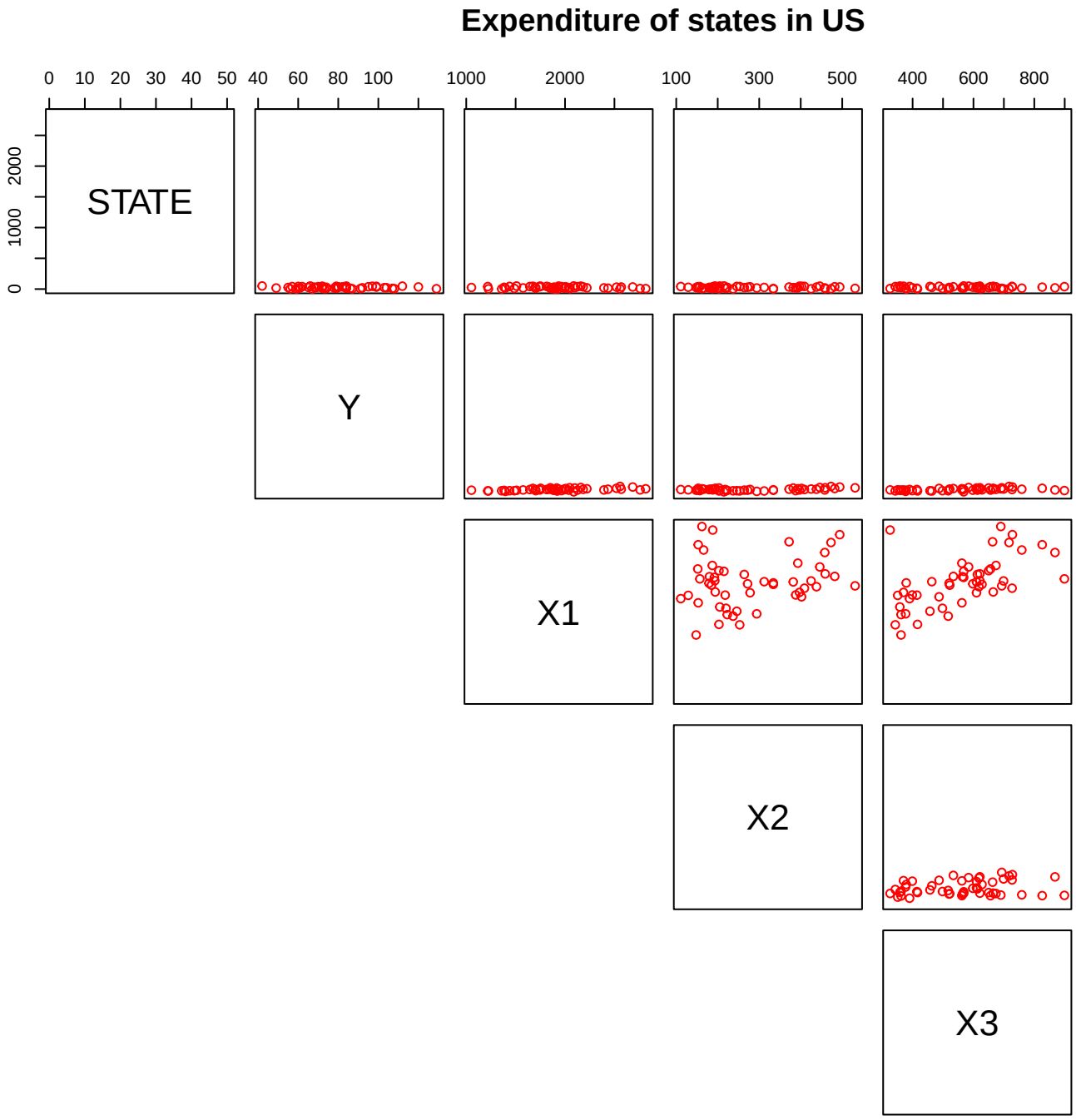
`State`	*50 states in US*
`Y`	*per capita expenditure on shelters/housing assistance in state*
`X1`	*per capita personal income in state*
`X2`	*Number of residents per 100,000 that are "financially insecure" in state*
`X3`	*Number of people per thousand residing in urban areas in state*
`Region`	*1=Northeast,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

  I used the following code to get the graph below of the relationships between the vectors (I needed a lot of help from other students to get to this stage):

  str (expenditure) lines(expenditure$Y$)$lines(expenditure$X1) lines(expenditure$X2$)$lines(expenditure$X$
  plot(expenditure, ylim=range(expenditure$Y$, $expenditure$X1,expenditure$X2$, $expenditure$X3),
  col='red', main = "Expenditure of states in US", lower.panel = NULL)

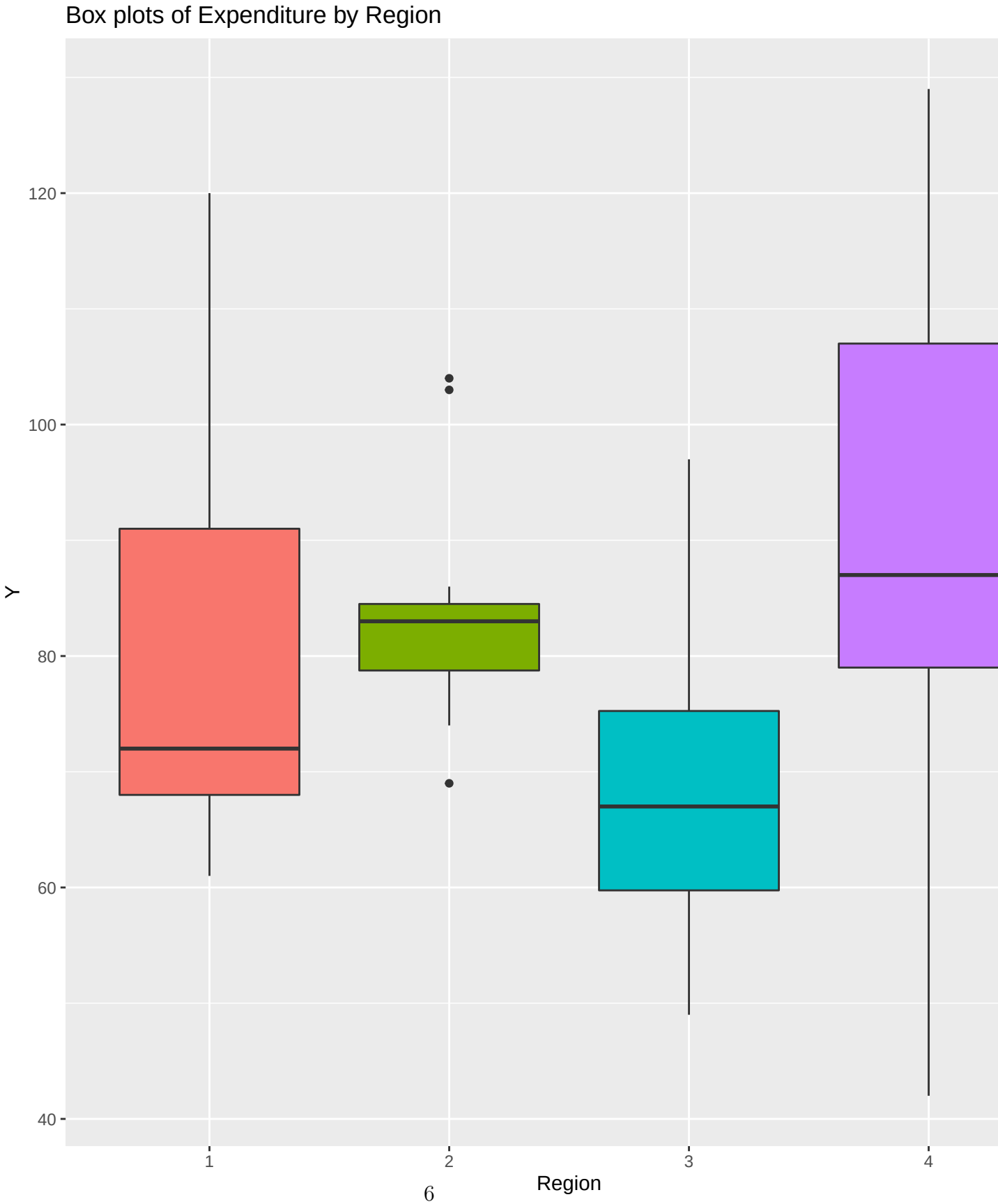

Expenditure of states in US
STATE
Y
X1
X2
X3

My interpretation of this graph is that there is no relationship between the state and any of the other variables. There also appears to be no relationship between housing assistance and the other variables. There is a strong positive relationship between X1 and X2.

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

  Here is the code I used to get the next graph that shows the relationship between Y and Region (again, help was needed):

  install.packages("ggplot2") library(ggplot2) data=as.data.frame(expenditure[,c(2,6)])
  data$Region = as.factor(data$Region) mode(data$Region)Icreatedaboxplotgraphcomparingexpenditu
  $Y, x = Region, fill = Region), data = data)$+$geom_boxplot()$+$ggtitle("BoxplotsofExpenditurebyRe$

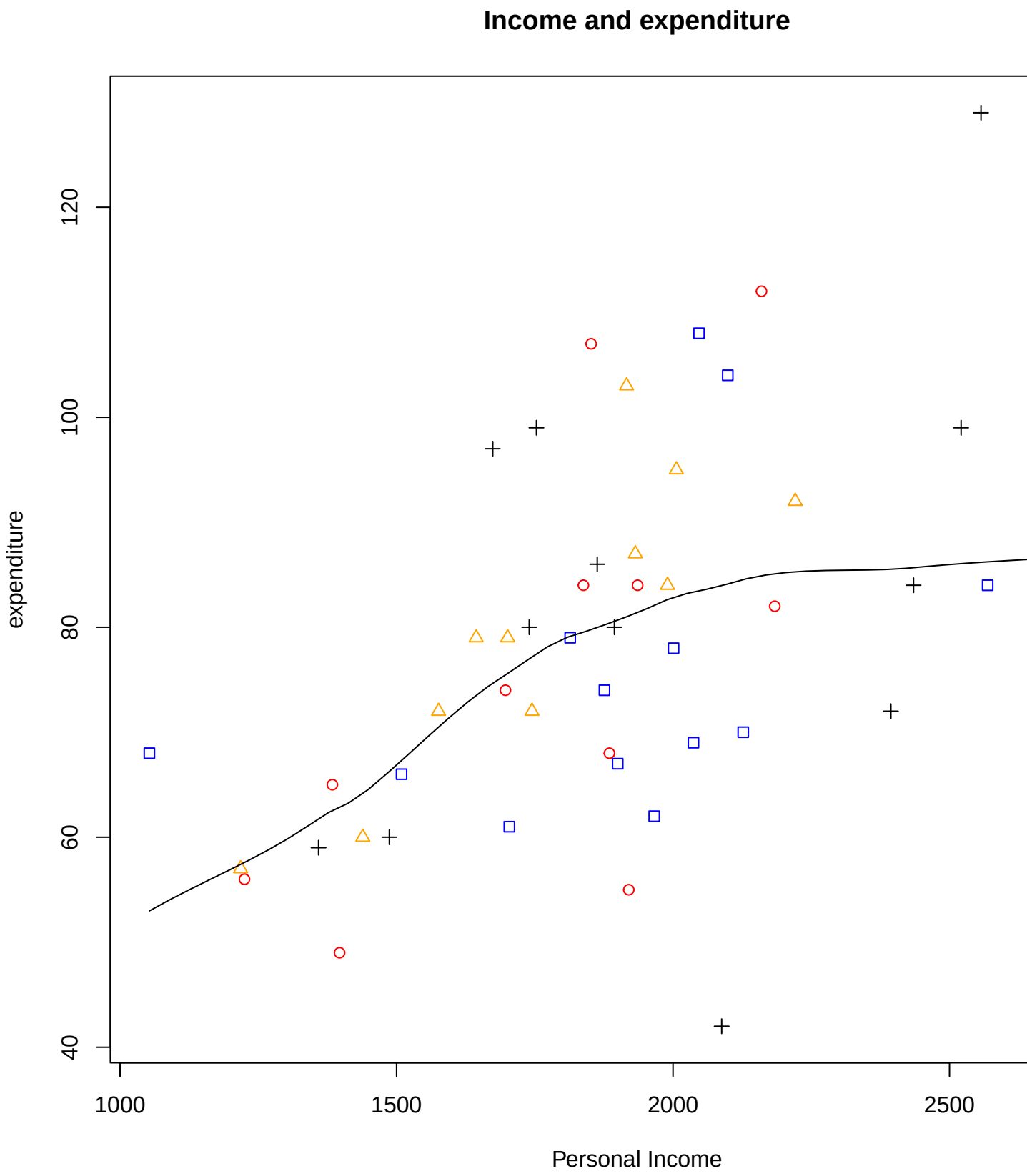

Box plots of Expenditure by Region
120
100
80
60
40
Y
1
2
3
4
Region

It is clear that number 4, the purple bar, has the highest expenditure.

- 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.

  Below is the code I used to create the graph to answer part 3 of this question, displayed above the graph:

  scatter.smooth(expenditure$X1, expenditure$Y, xlab = 'Personal Income', ylab = 'expenditure', main = 'Income and expenditure', col = c("blue", "red", "orange", "black"), pch = c(0,1,2,3))


Income and expenditure
expenditure
120
100
80
60
40
1000
1500
2000
2500
Personal Income

From the graph we can read the following information:

It appears that when personal income increases the per capita expenditure on shelters and housing also slightly increases, but once personal income reaches 2000 the increase in per capita expenditure ceases to increase as sharply, levels off and stagnates.