Problem Set 1

QTM 200: Applied Regression Analysis

Due: January 29, 2020

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 the course GitHub page in .pdf form.
- This problem set is due at the beginning of class on Wednesday, January 22, 2020. No late assignments will be accepted.
- Total available points for this homework is 100.

Question 1 (25 points)

A private 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)
```

Find a 90% confidence interval for the student IQ in the school assuming the population of IQ from which our random sample has been selected is normally distributed.

1 Question 1 answer

```
z90 <- qtnorm((1-0.90)/2, lower.tail = FALSE)
n <- length(y)
sample_mean <- mean(y)
```

```r
sample_sd <- sd(y)
lower_90 <- sample_mean - (z90 * (sample_sd/sqrt(n)))
upper_90 <- sample_mean + (z90 * (sample_sd/sqrt(n)))
confint90 <- c(lower_90, upper_90)
```

The confidence interval is (94.1, 102.7). There is a 90 percent probability that a confidence interval would contain the true average IQ of students in the school with repeated sampling. For example, if we gathered 100 samples, we would expect 90 of the CIs calculated to contain the population mean (student IQ in this school) and we would expect 10 of the CIs to not include the population mean.

Question 2 (25 points)

A private school counselor was curious whether the average of IQ of the students in her school is higher than the average IQ score 100 among all the schools in the country. She took a random sample of 25 students' IQ scores. The following is the data set:

```r
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)
```

Conduct a test with 0.05 significance level assuming the population of IQ from which our random sample has been selected is normally distributed.

2 Question 2 Answer

```r
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)
SE <- (sample_sd / sqrt(length(y)))
SE
t <- ((sample_mean - 100)/ SE)
1 - pt(abs(t), df=24, lower.tail = F)
t.test(y, mu = 100, alternative = "greater", df=24)
```

The p-value is 0.7215. Because the p-value is greater than the significance level of 0.05, I fail to reject the null hypothesis (that there are no significant differences between the sample IQ and the average IQ of 100 across all schools in the country). If the null hypothesis were true, there is a 72.15 percent probability that we would observe a test-statistic this extreme or more The school counselor's students do not have a significantly higher average IQ than the average IQ of 100 of students across the country.

Question 3 (50 points)

Researchers are curious about what affects the education expenditure on public education. The following is availabe variables in a data set about the education expenditure.

State	*50 states in US*
Y	*per capita expenditure on public education*
X1	*per capita personal income*
X2	*Number of residents per thousand under 18 years of age*
X3	*Number of people per thousand residing in urban areas*
Region	*1=Northeast, 2= North Central, 3= South, 4=West*

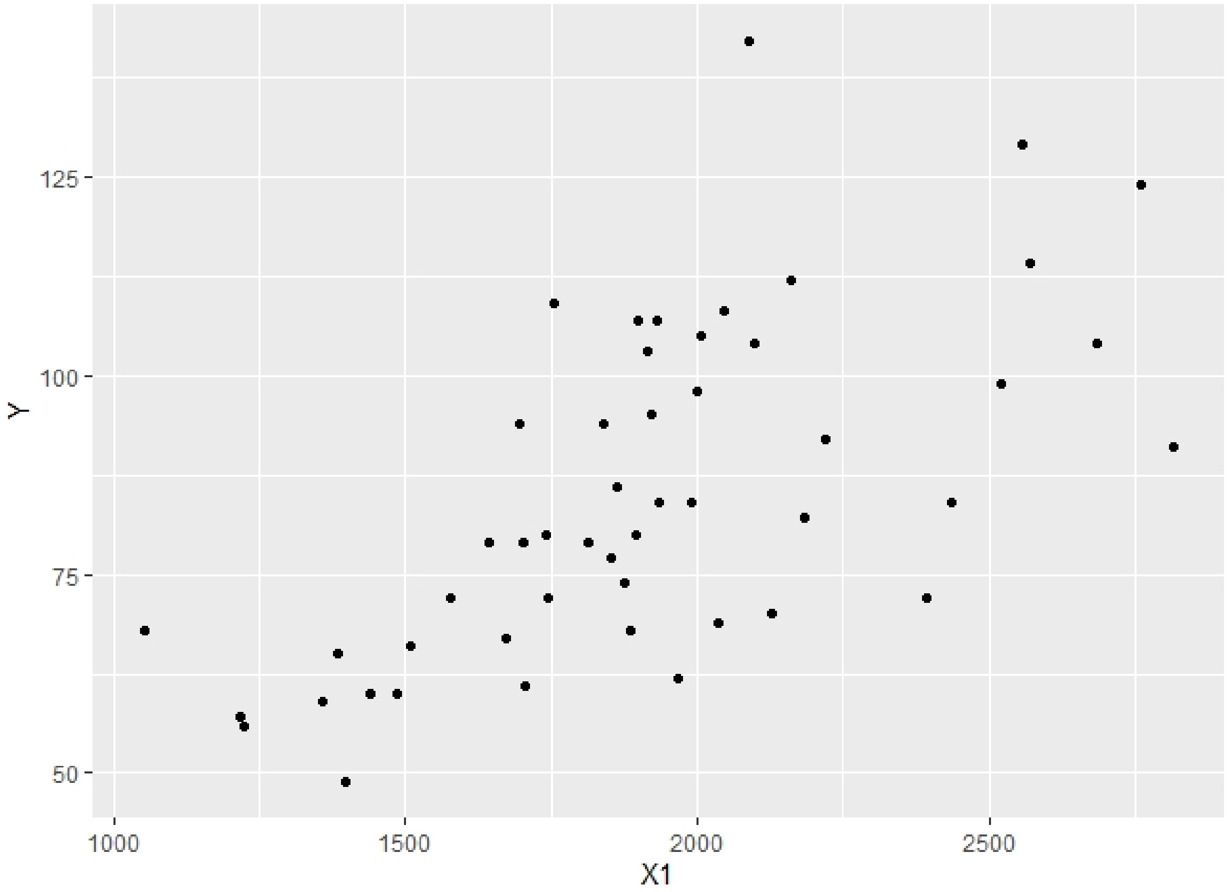
Explore the `expenditure` data set and import data into `R`.

```
expenditure <- read.table(”expenditure.txt”, header=T)
```

- 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)?

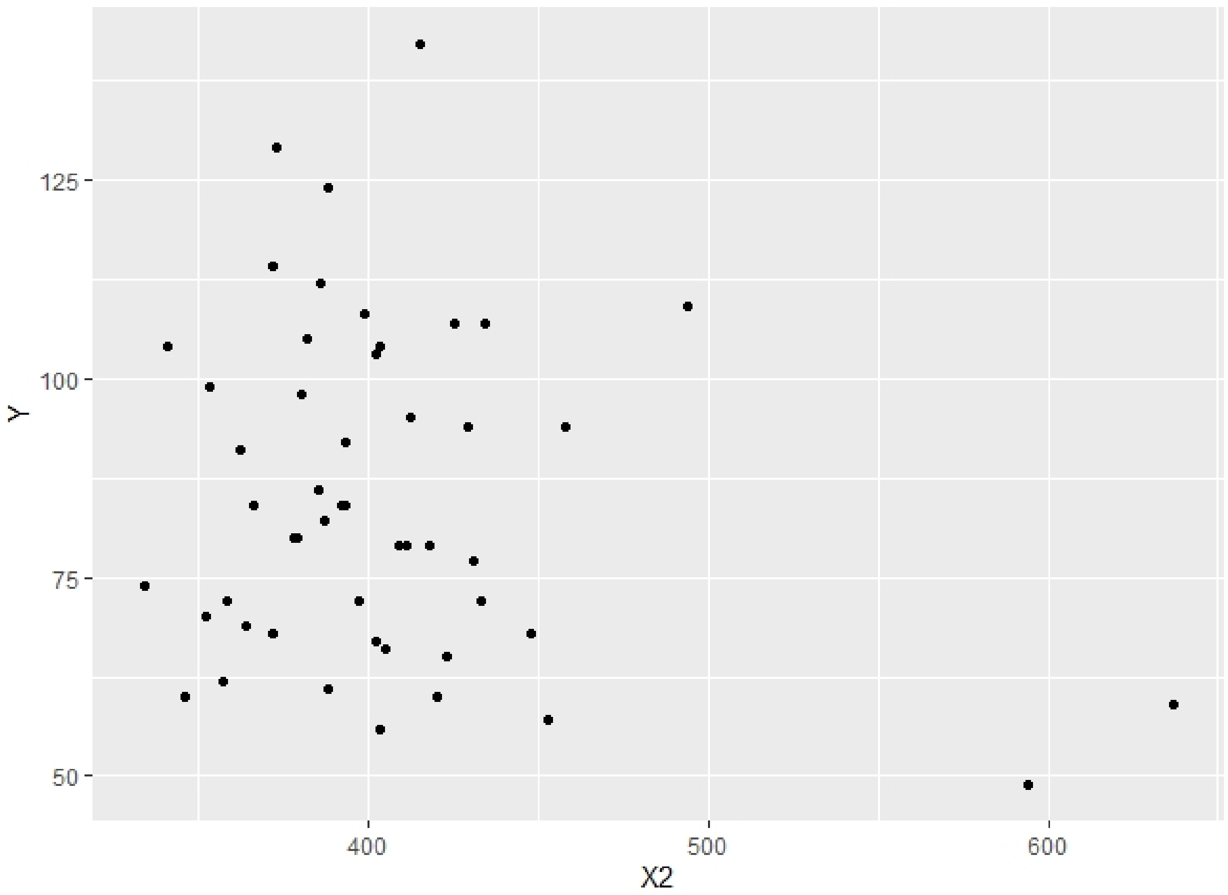
3 Question 3 Answers

```
expenditure <- read.table(”expenditure.txt”, header=T)
library(ggplot2)
ggplot(expenditure, aes(x=X1, y=Y)) + geom_point()
```



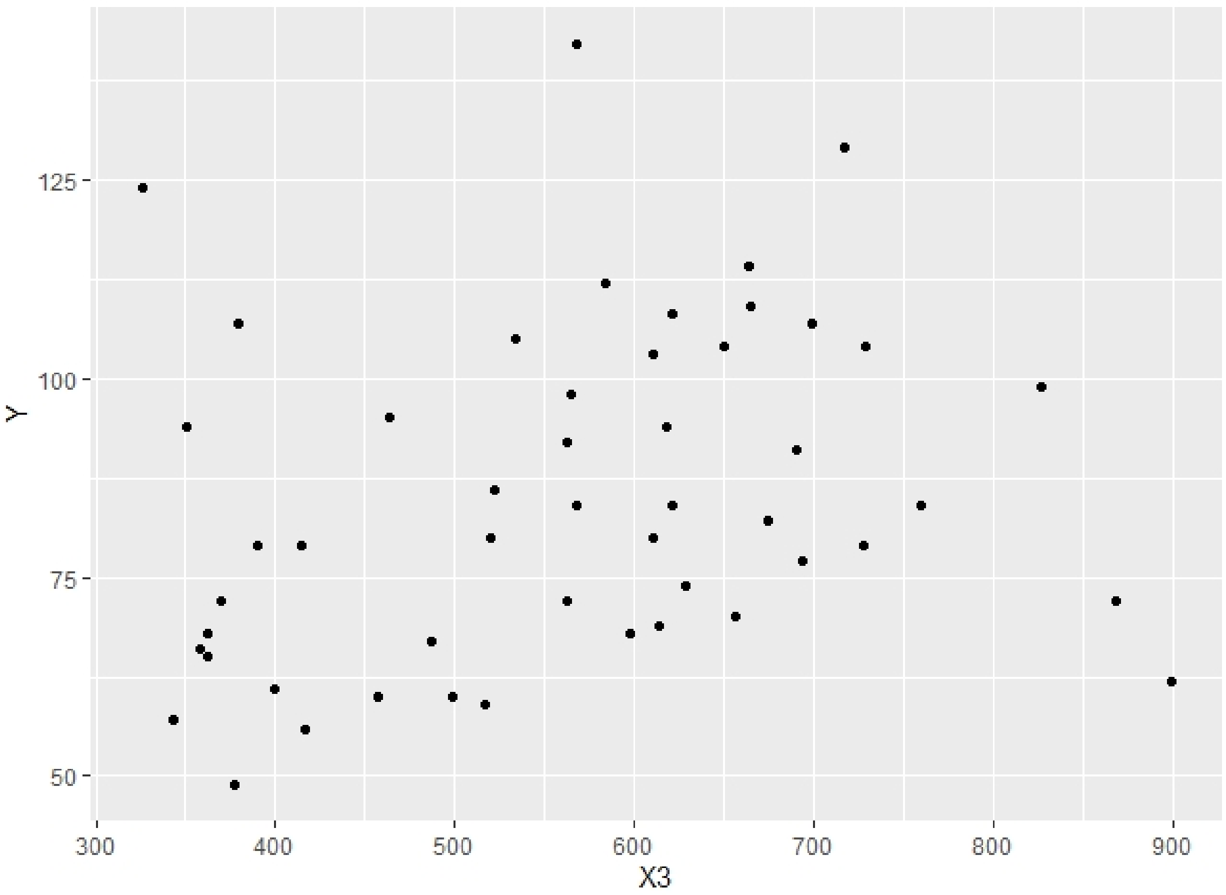
There is a linear, moderate, positive relationship between per capita personal income and per capita expenditure on public education

```
ggplot(expenditure, aes(x=X2, y=Y)) + geom_point()
```



There does not appear to be a relationship between number of residents per thousand under 18 years old and per capita expenditure on public education.

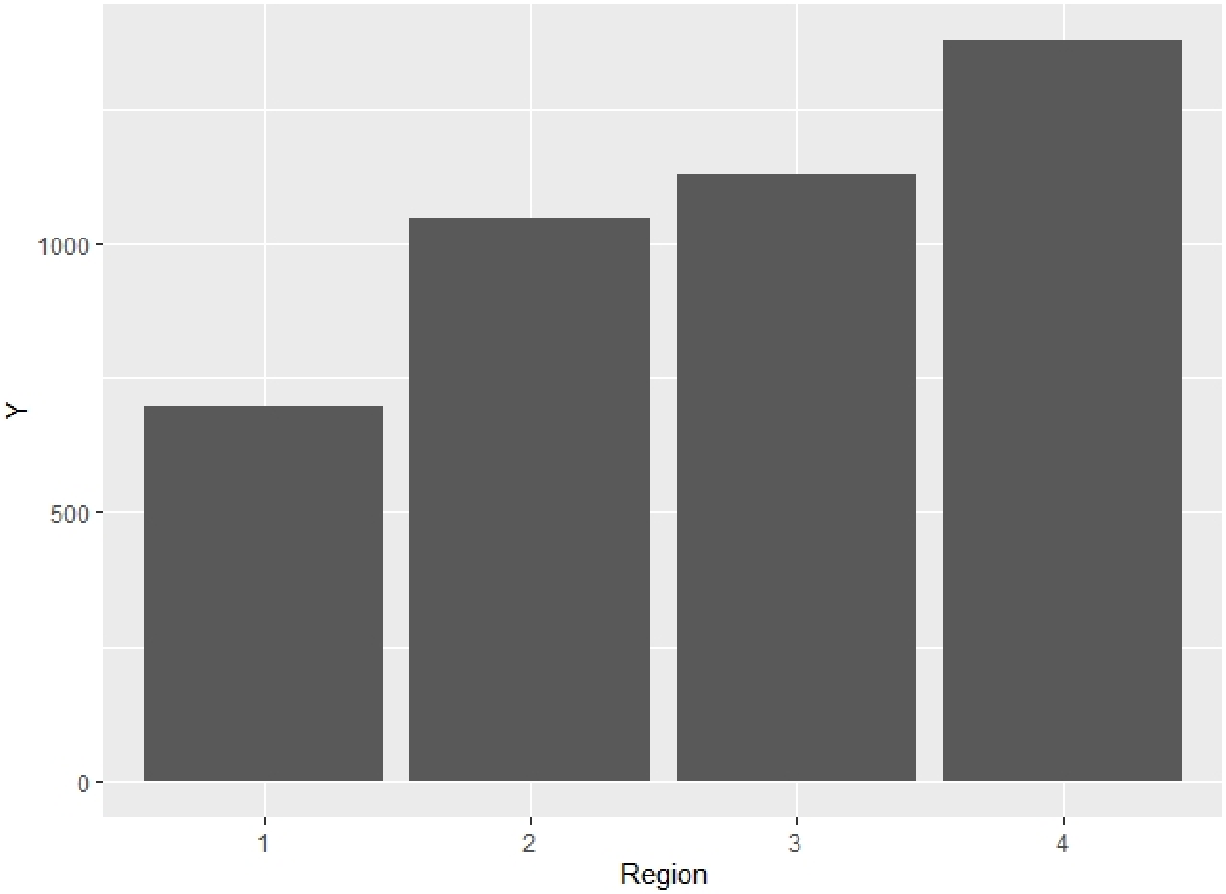
```
ggplot(expenditure, aes(x=X3, y=Y)) + geom_point()
```



There is a linear, weak, positive relationship between number of people per thousand residing in urban areas and per capita expenditure on public education

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

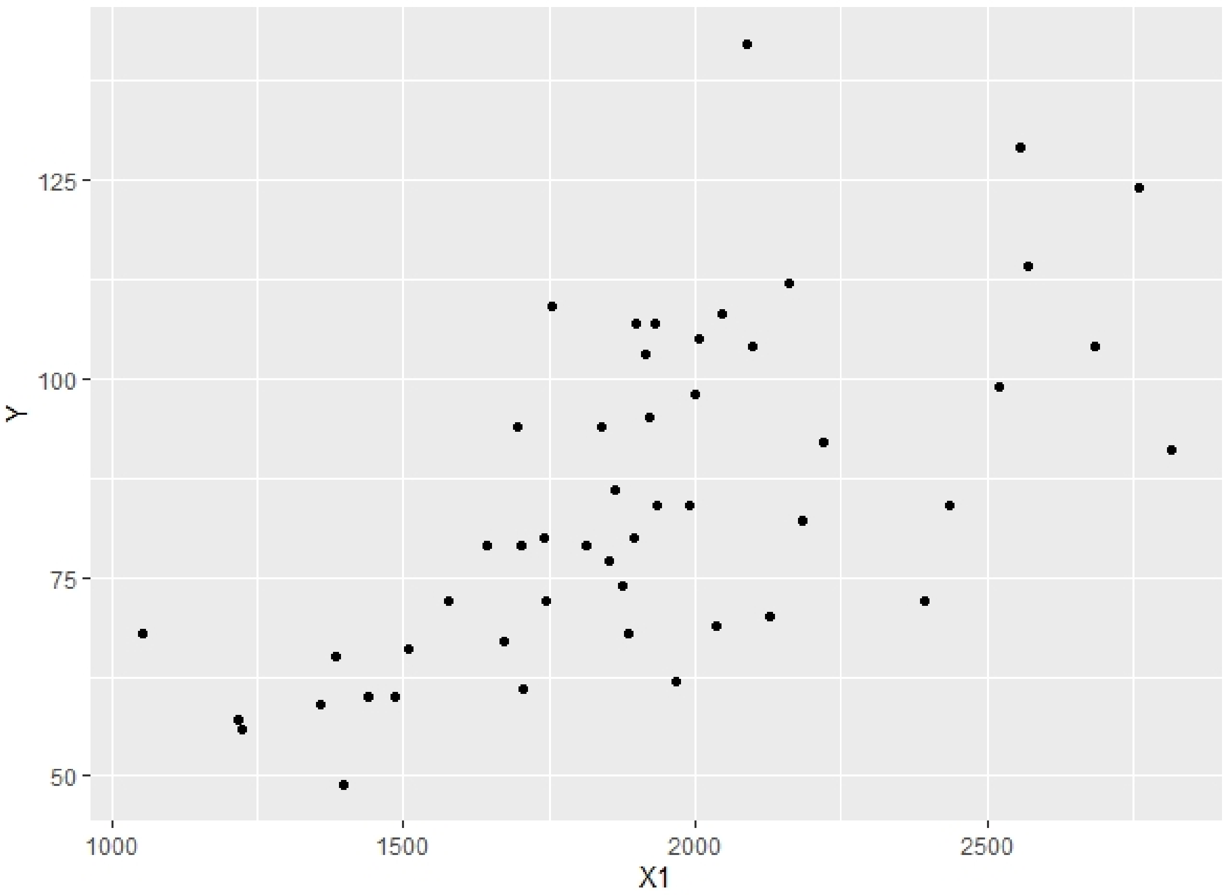
```
ggplot(data=expenditure, aes(x=Region, y=Y)) + geom_bar(stat="identity")
```



Region 4 (West) has the highest per capita expenditure on public education.

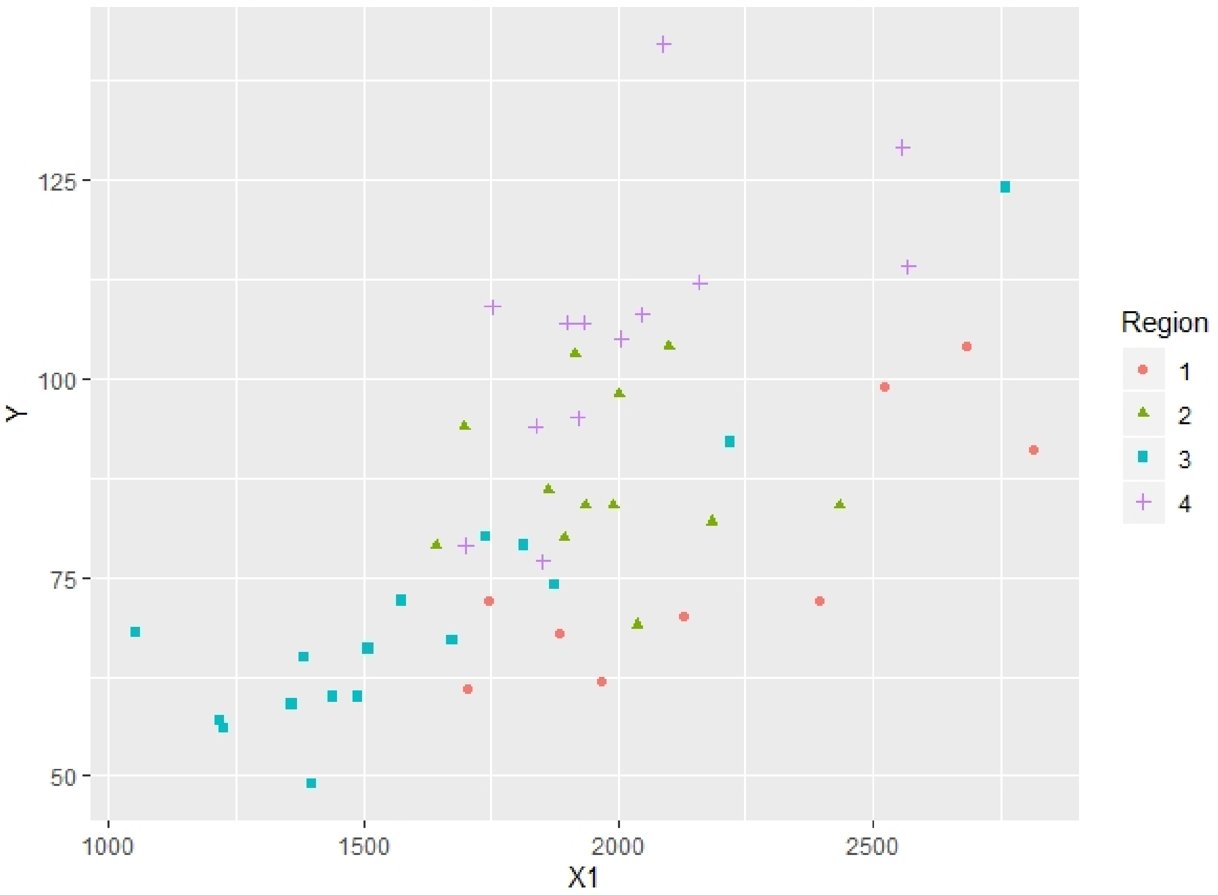
- Please plot the relationship between *Y* and *X1*? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.

```
expenditure <- read.table("expenditure.txt", header=T)
library(ggplot2)
ggplot(expenditure, aes(x=X1, y=Y)) + geom_point()
```



There is a linear, moderate, positive relationship between per capita personal income and per capita expenditure on public education

```r
expenditure$Region <- factor(expenditure$Region)
is.factor(expenditure$Region)
ggplot(expenditure, aes(x=X1, y=Y, col=Region, shape = Region)) + geom_point()
```


Region
1
2
3
4
125
100
75
50
Y
1000
1500
2000
2500
X1