PS01_Answers_KG_PDF

Kathryn Glen

9/30/2021

PROBLEM ONE

Question 1

Load dataframe

```
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)
```

Run t.test function to find the confidence interval, specify to .9 for 90 instead of 95 per cent. Found online.

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

##
##  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
```

Answer: 90 percent confidence interval for average student IQ is between 93.95993 and 102.92007

Question 2

Write out the Null Hypothesis and the Alternative Hypothesis Null hypothesis is that the average IQ in the school is equal to or lower than the average IQ in the country Alternative hypothesis is that the average IQ in the school is higher than the average IQ in the country

Personal Comment: I really struggled with this bit of code, as I felt I only knew how to do a T-Test by hand, and I struggled to use Google as I did not know which was the right question to be asking

```
my_test=t.test(y, mu=100, alternative="greater")
my_test$p.value

## [1] 0.7215383
```

Answer: p-value = 0.7215383, alpha = 0.05, therefore we have failed to reject the null hypothesis. The average IQ in the school is lower than the average IQ in the country

Personal Comment: I didn't end up needing any of the below code, but I calculated it because those were the elements of the formulas to do hypothesis testing by hand however I couldn't figure out how to use them in R

```
mean(y)

## [1] 98.44

sd(y)

## [1] 13.09287

length(y)

## [1] 25

t= 37.593
alpha=0.05
```

PROBLEM 2

Question 1

```
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-
TCD/StatsI_Fall2021/main/datasets/expenditure.txt", header=T)
str("expenditure")

##  chr "expenditure"

View(expenditure)
```

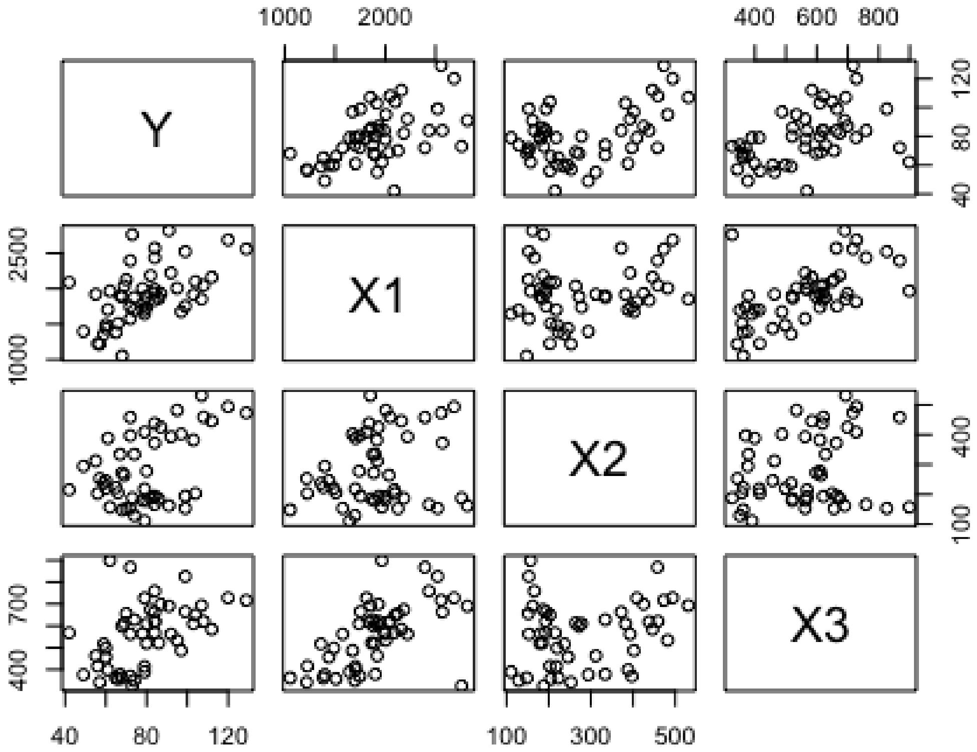
I asked a friend who has a statistics degree to help at this point, as I was entirely lost and didn't know how to plot graphs

Created a new list using just the variables I want for this question

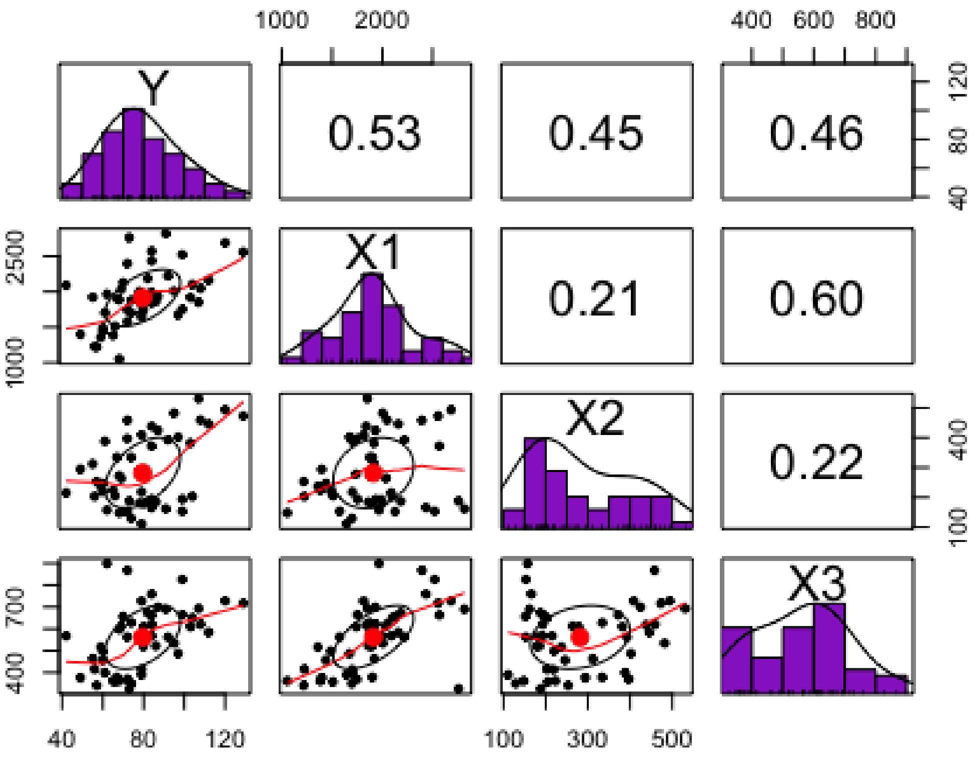
```
new.data = expenditure[,c(2,3,4,5)]
```

Friend advised trying the pairs method

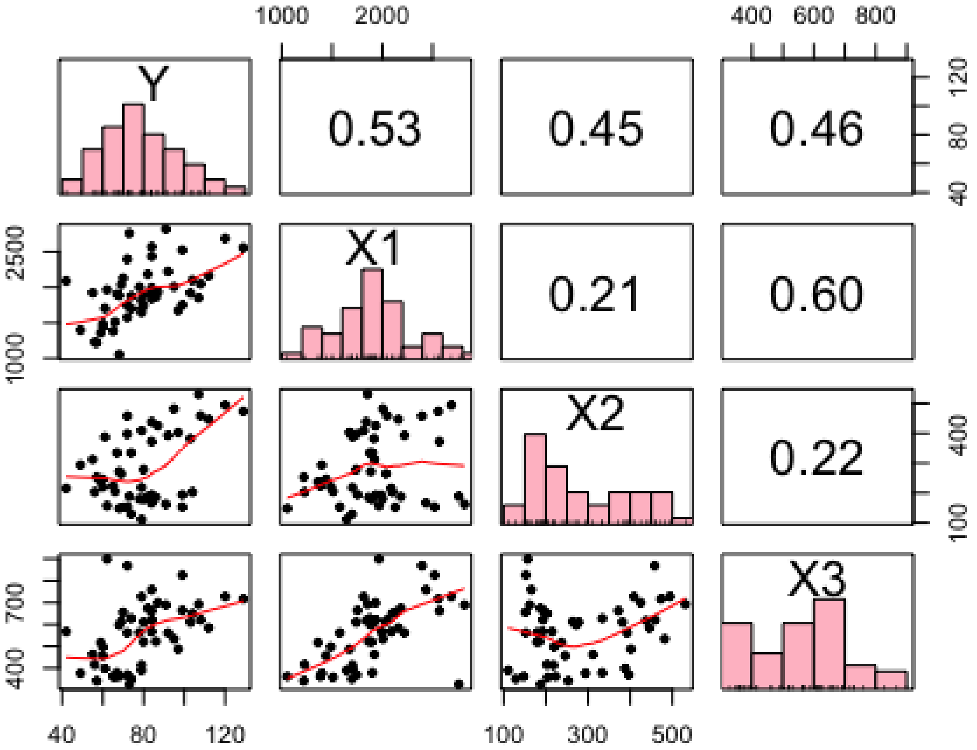
```
pairs(new.data)
library(psych) #to use pairs
```



```
pairs.panels(new.data,
             method = "pearson", # correlation method
             hist.col = "darkorchid"
)
```



```
pairs.panels(new.data,
             method = "pearson", # correlation method
             hist.col = "pink",
             density = FALSE,
             ellipses = FALSE
)
```



Answer: Expenditure on housing assistance mildly affects personal income in the state it affects financial insecurity to a lesser extent than X1 the strongest relationship is X1 and X3, however nothing was above 0.6. I googled the correlation ranges and so I don't think .6 is considered a significantly strong correlation.

Question 2

Plot relationship between Y and Region I need to make sure R recognises Region numbers as factors not integers

```
expenditure$Region = as.factor(expenditure$Region)
```

barplot(table(expenditure$Region)) This did not work, just turned up as bars

Tried ggplot, download ggplot2

```
library(ggplot2)

##
## Attaching package: 'ggplot2'
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

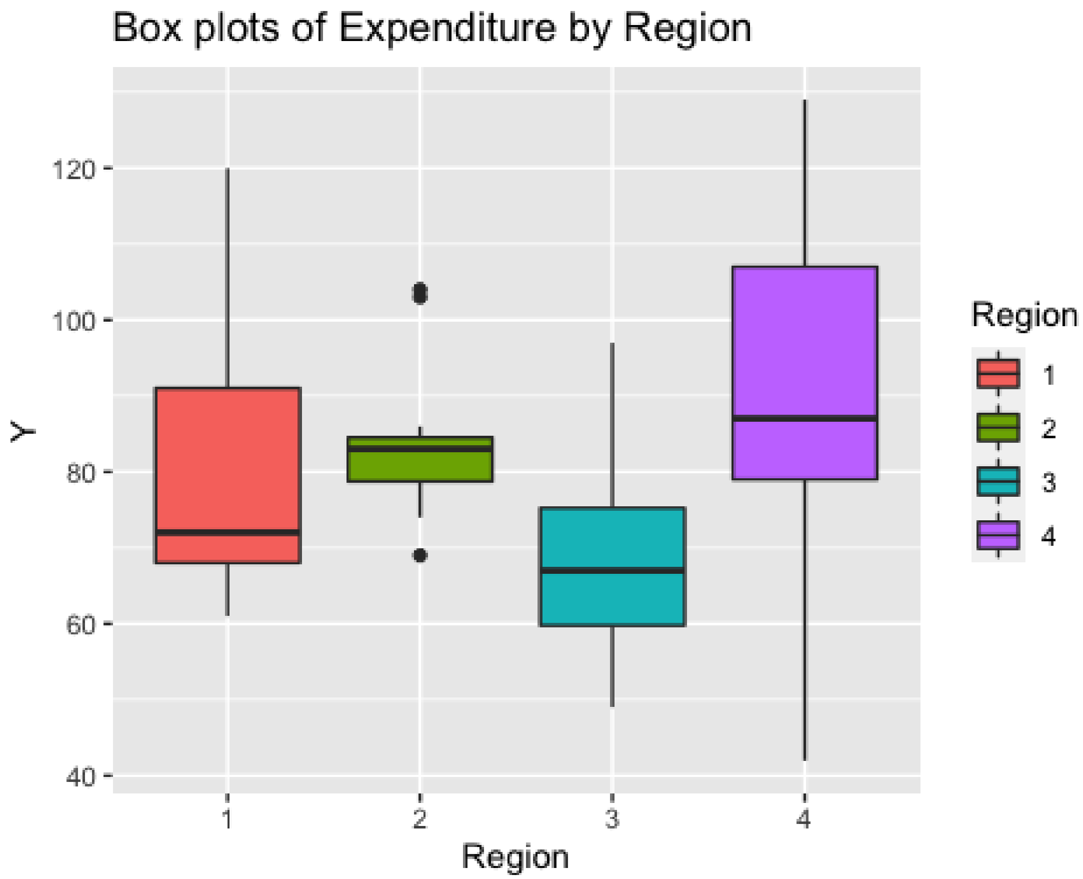
Get variables I need from Expenditure by making data frame of Y and Region

```
data=as.data.frame(expenditure[,c(2,6)])
data$Region = as.factor(data$Region)
mode(data$Region)

## [1] "numeric"
```

Create a box plot graph comparing expenditure and region I would not have thought of a box plot by myself without my friend as this is very new territory for me

```
ggplot(aes(y = Y, x = Region, fill=Region), data = data)+
geom_boxplot()+ggtitle("Box plots of Expenditure by Region")
```



I then googled how to read box plot graphs.

Answer: Region 4 (West), has both the widest range of expenditure as well as the concentration of the most wealth. The West has the highest per capita expenditure on housing assistance.

Question 3

Plot the relationship between Y and X1

This is everything I tried based on Google before once again reaching out to my friend to teach me how to create plots with more than 2 variables

barplot(table(expenditure$X1)) ggplot(aes(y = Y, x = X1, fill=X1), data = data)+ geom_boxplot()+ggtitle(“Box plots of Expenditure by Income”)

plot(hist(expenditure$Y), main=“Distribution of Y”, xlab="")

hist(expenditure$Y)hist(expenditure$Y, probability=TRUE) lines(density(expenditure$Y))plot(density(expenditure$Y), main = “Expenditure”)

plot(density(expenditure$Y), main="Distribution of Y and X1", xlab="", col="red", xlim=c(0,100)) lines(density(expenditure$X1), lty=2, col=“blue”)

data2=as.data.frame(expenditure[,c(2,3)]) p3<-ggplot(aes(y = Y, x = X1, fill=X1), data = data2)+ geom_boxplot()+ggtitle(“Box plots of Expenditure by Income”) p3

Personal Comment: None of the above attempts are working for me to plot a graph which can then have a third variable added in easily, and I have no idea what is going wrong with that final box plot. I think there must be a simpler way to make a line graph but I am struggling to find it.

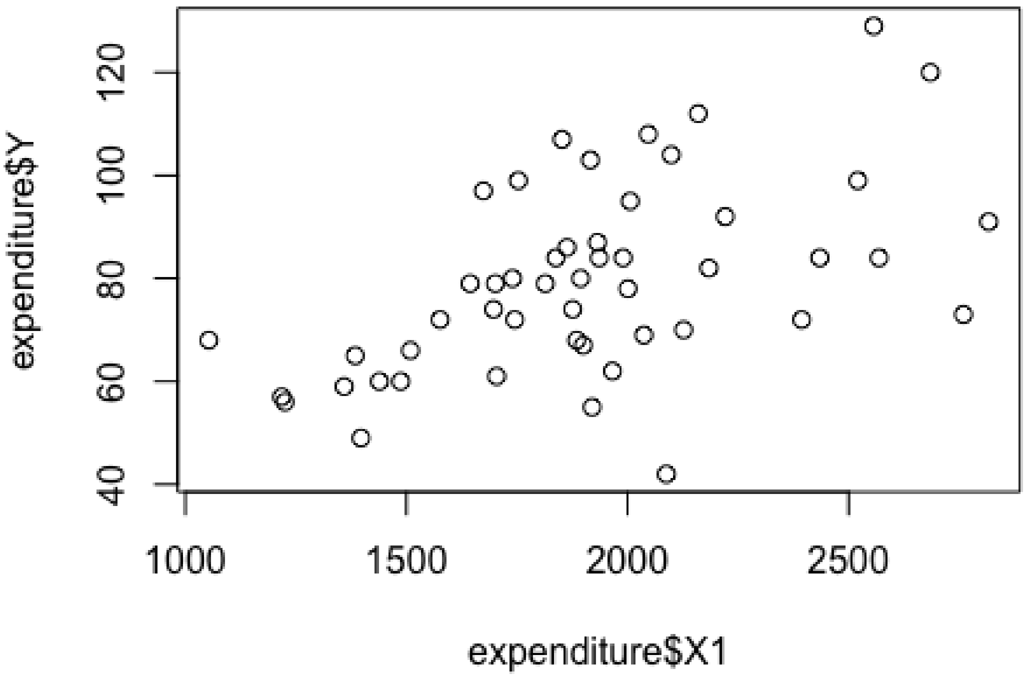
I then tried a line plot based on group work

plot(expenditure, ylim=range(expenditure$Y, expenditure$X1), col=‘black’, main = “Expenditure vs Income”) plot(expenditure$Y)plot(expenditure$Y, type=“l”, col=“blue”, main=“Expenditure by Income”)

After hours of work on this I realised I was not meant to be making a line graph

The below is the material I produced after my friend explained these functions in R

```
plot(expenditure$X1, expenditure$Y)
```



This created a scatter plot which showed that Y and X1 are positively correlated

Override data to get a new data frame with the 3 variables needed in this question and make new data frame. Need to make sure Region is read as a factor.

```
data=as.data.frame(expenditure[,c(2,3)])
region = as.factor(expenditure$Region)
data = cbind(data, region)
data = as.data.frame(data)

region

##  [1] 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 4
## [39] 4 4 4 4 4 4 4 4 4 4 4 4
## Levels: 1 2 3 4
```

Prints out levels , which means region is a factor

Put in Colours by creating new vector with colour names

```
colour.vector=c("chartreuse", "brown", "blue", "goldenrod")
```

Label the colour vectors by the region they represent

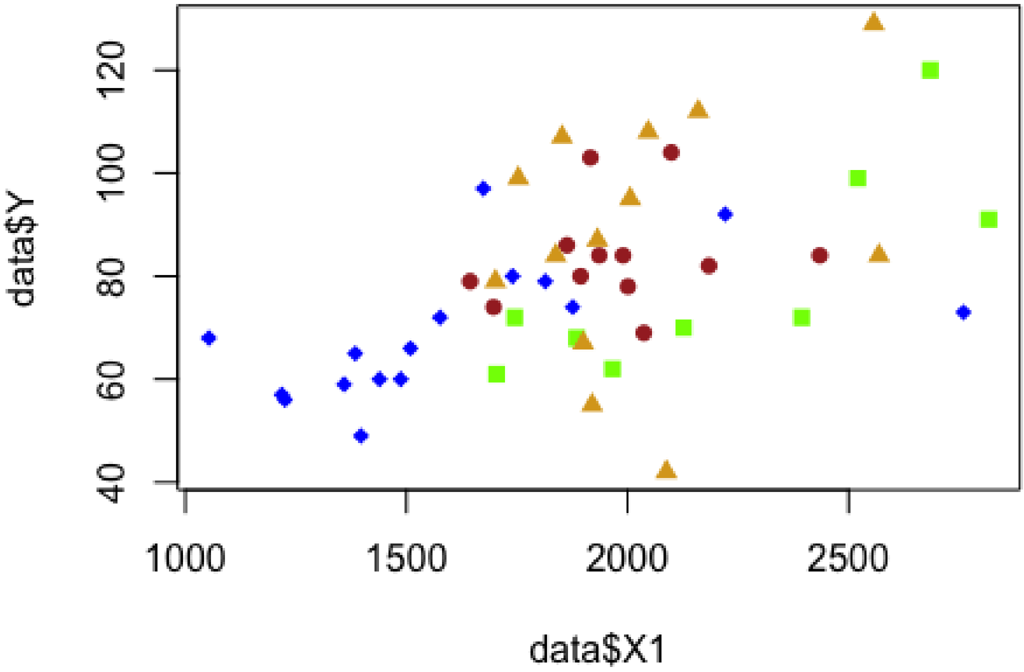
```
region.types = c("1", "2", "3", "4")
names(colour.vector) = region.types
```

Pick the point shapes - found their numbers online

```
pch.vector = c(15, 16, 18, 17)
names(pch.vector) = pch.vector
```

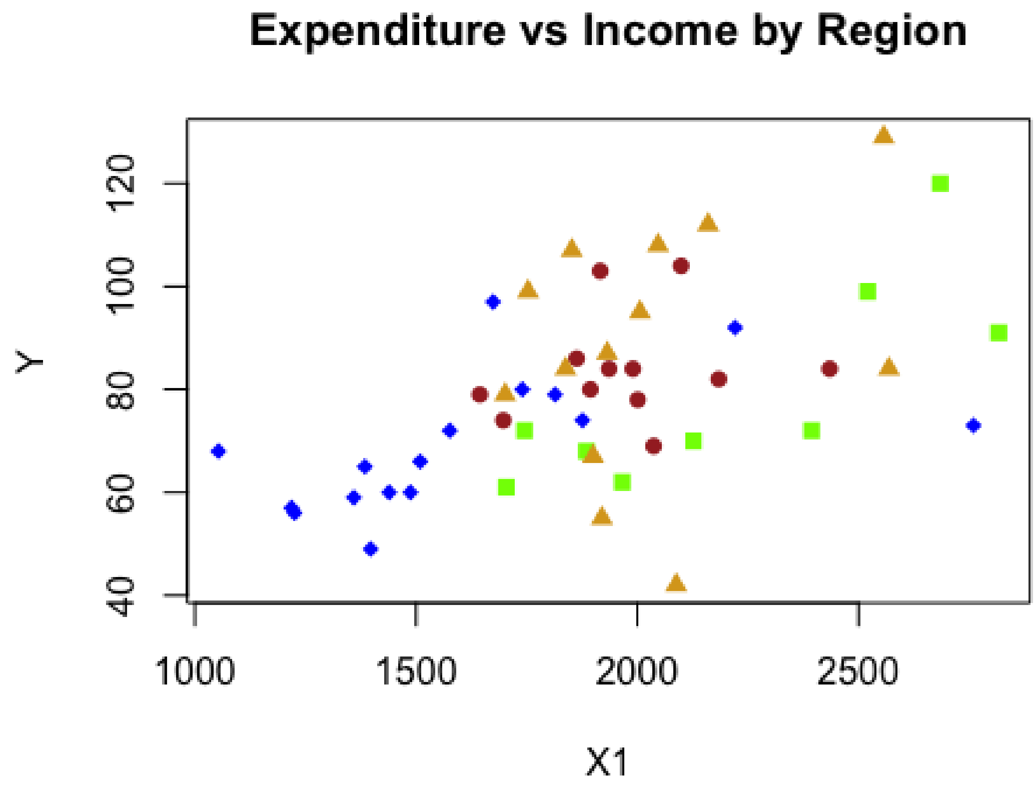
Adding colours and shapes to graph

```
plot(data$X1, data$Y,  col=colour.vector[region], pch = pch.vector[region])
```



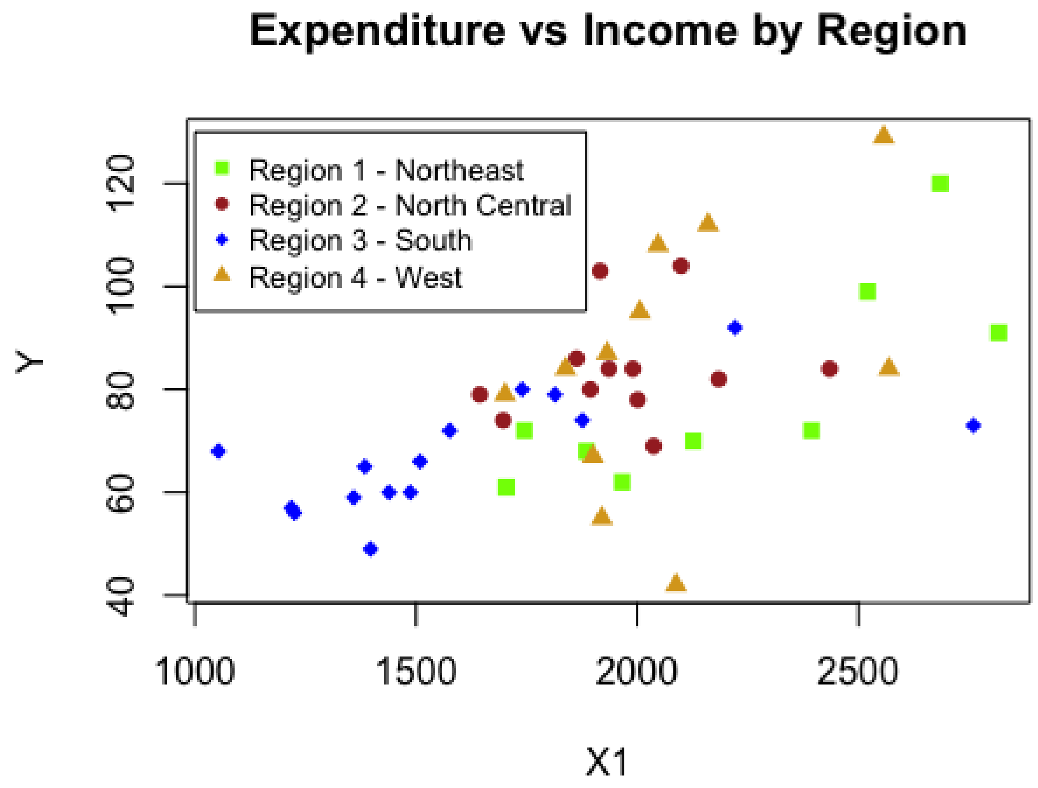
Added further details including proper labeling, title

```
plot(data$X1, data$Y,  col=colour.vector[region], pch = pch.vector[region],
main = "Expenditure vs Income by Region", xlab ="X1", ylab="Y")
```



I added a legend to make the meaning of the symbols clearer but the legend was too big so I needed to rerun the first bit and decrease the cex

```
plot(data$X1, data$Y,  col=colour.vector[region], pch = pch.vector[region],
main = "Expenditure vs Income by Region", xlab ="X1", ylab="Y")
legend(1000, 130, legend=c("Region 1 - Northeast", "Region 2 - North
Central", "Region 3 - South", "Region 4 - West"), col=colour.vector,
pch=pch.vector, cex=0.8)
```



Personal Comment: Overall this was a very difficult assignment for me which took at least a dozen hours and I would not have been able to complete it without help from somebody well versed in R. We created a study group within the course to work together and we all really struggled. I think there is a misconception of how advanced our skills are, I feel very lost, but I even needed to show classmates how to right click on a file to find the full name of their location so they could set the working directory, one girl was typing all of her code into the console of RStudio. We need very basic instruction in how to use R, not just the mathematics behind why we are running this.

I learned a lot from this assignment but it was very stressful and I do not see how I can do this every week since I needed to neglect other classes to do this. Hopefully in the future we will be more versed in R while doing these assignments but extra instruction on the basics of this program would be hugely beneficial to the class as a whole.