Problem Set 1

Molly Marino

Due: October 9, 2025

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.
- This problem set is due before 23:59 on Thursday October 9, 2025. No late assignments will be accepted.

Question 1: 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:

1. Find a 90% confidence interval for the average student IQ in the school.

```r
average_student_IQ <- 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)
#I renamed ”y” as an object to average_student_IQ because it better represents the data set.
sample.mean <- mean(average_student_IQ)
# I found the mean of the sample because it is used as a point estimate for the unknown population mean.
sample.mean
#This step shows the mean in the console. The mean is 98.44.
```

```r
sample.n <-length(average_student_IQ)
# I did this step to calculate my sample size. The sample.n is 25L
sample.sd <-sd(average_student_IQ)
# The next step of calculating the confidence interval is to find the standard deviation or spread of the data. The sample.sd is 13.09.
sample.se <-sample.sd/sqrt(sample.n)
# I did this step to calculate the standard error and to see how much the mean will fluctuate between the samples.
sample.se
# This steps shows the standard error in my console. The sample.se is 2.618.
df <- sample.n-1
# I did this step to calculate the shape of the distribution. The df is 24.
t_critical <- qt(0.95,df=sample.n-1)
# This step is using a t distribution because the population size is less than 30. I adjusted for 90 percent confidence to find the t score. The t_critical value is 1.71.
margin_error <- t_critical * sample.se
# I did this step to find the discrepancy between the sample and the t value.The margin_error is 4.48.
lower_bound <-sample.mean-margin_error
#The lowest possible value within the 90 percent confidence level for the population parameter is 93.95.
upper_bound <-sample.mean +margin_error
# The highest possible value within the 90 percent confidence level for the population parameter is 102.92.
print(c(lower_bound,upper_bound))
# This shows the 90 percent confidence intervals  of the average student IQ in the school for upper and lower values. Meaning, if sampling was conducted numerous times 90 percent of the student's IQ scores would fall between 93.95-102.92
```

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 $\alpha = 0.05$.

```r
# H0:   =100
# Ha:   >100
mu0 <- 100
# muO is a way to store the value null hypothesis mean under a value name.
t_value <- (sample.mean - mu0) / sample.se
# The t value is used because the sample is less than 30 and it shows how far the sample mean is away from the null hypothesis. The t_value is -0.595.
t_value
# This shows the t value in the console
p_value <- pt(abs(t_value), df, lower.tail = FALSE)
```

```r
# the p value is used to compare against the a=0.05 to determine if we can reject the sample data based on the null hypothesis. I did tail=false because I am only interested in scores above 100 in the upper tail of the data.
p_value
# The p value is 0.27 and this means that is is greater than 0.05. The null hypothesis cannot be rejected.
alpha <- 0.05
if(p_value < alpha){
  print(”Reject H0: Average student IQ is greater than 100”)
} else {
  print(”Do not reject H0: Not enough evidence that average student IQ is greater than 100”)
}
# The null hypothesis cannot be rejected because the p value of 0.27 is greater than the alpha value of 0.05. This means that there is not enough evidence that the average student IQ score in the teacher's class is higher than the national average.
```

Question 2: Political Economy

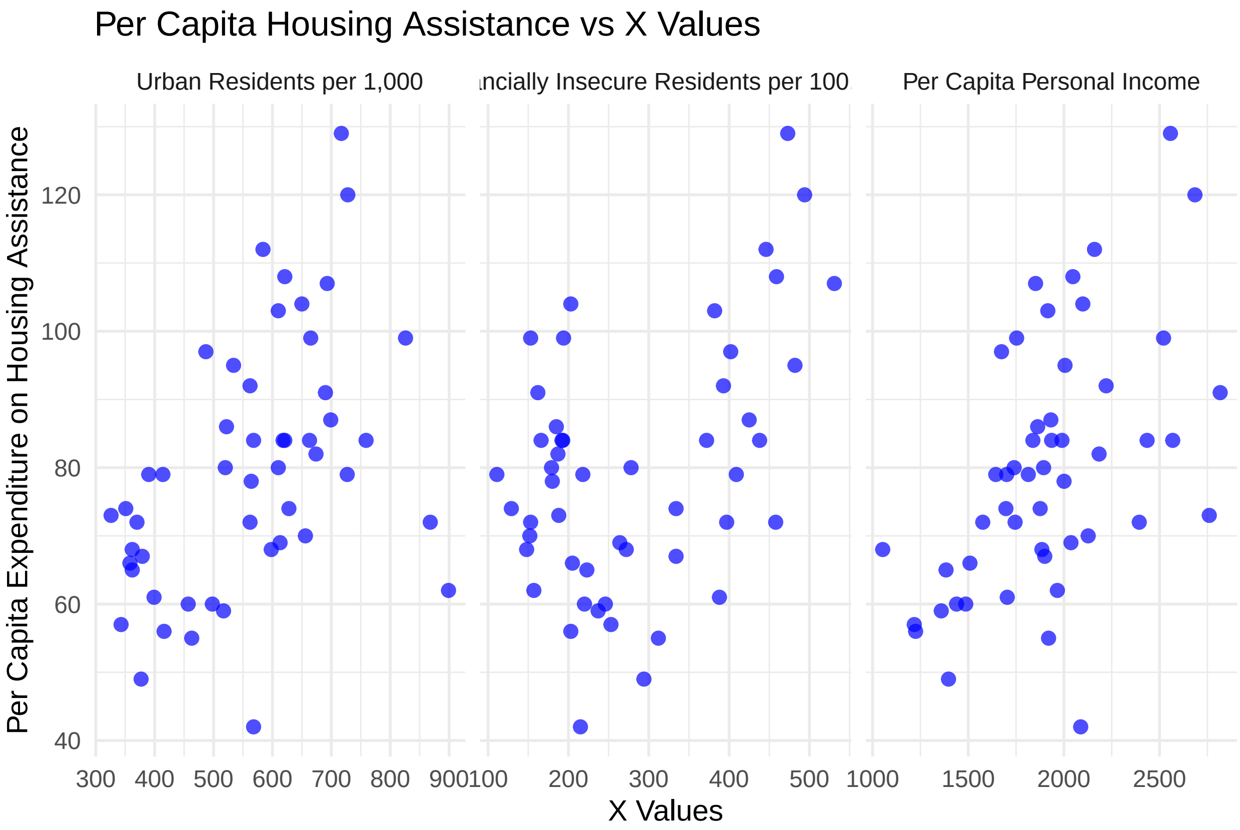
Researchers are curious about what affects the amount of money communities spend on addressing homelessness. The following variables constitute our data set about social welfare expenditures in the USA.

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`.

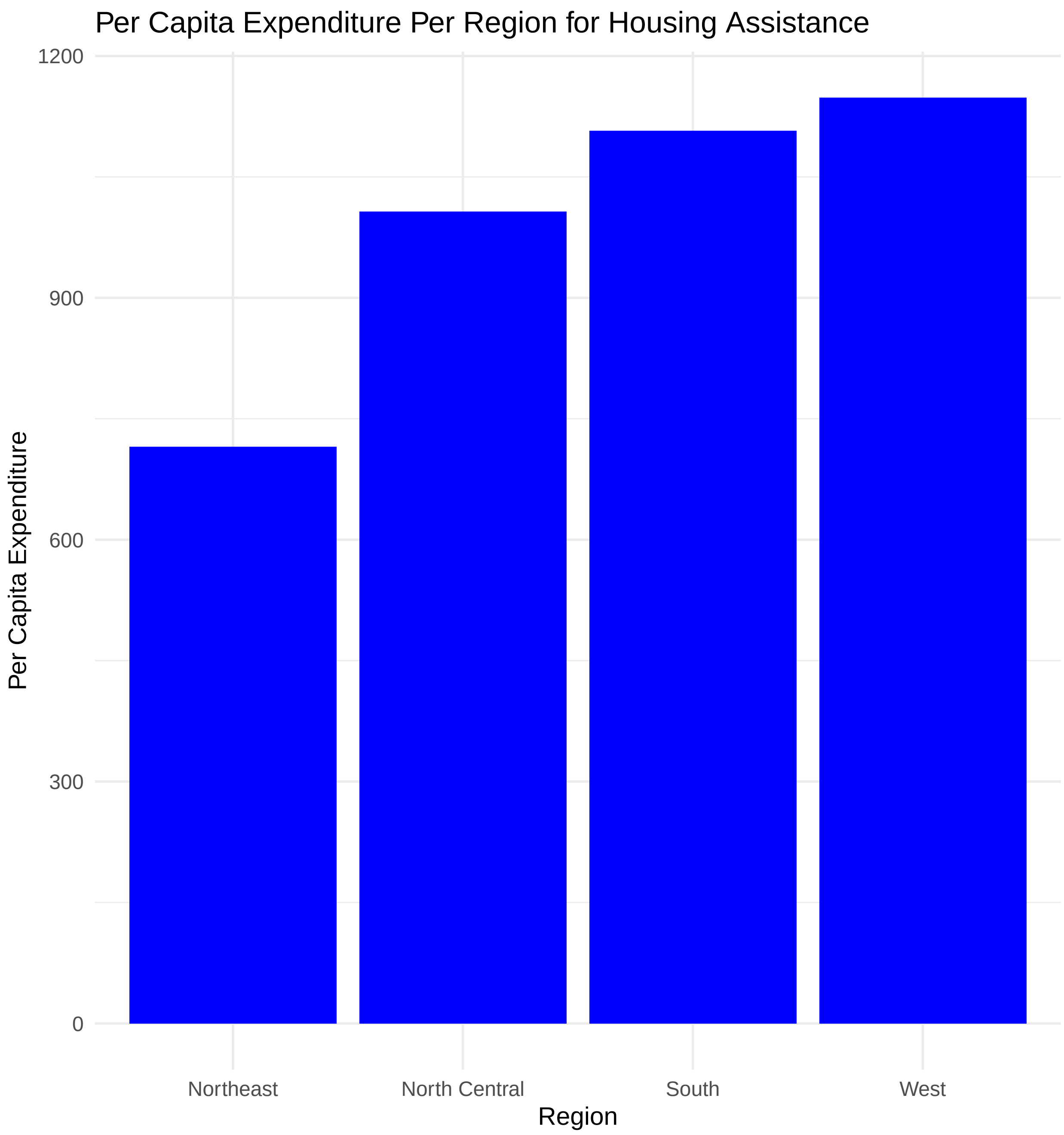
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?

```r
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_2025/main/datasets/expenditure.txt", header=T)
expenditure$Per_Capita_Expenditure_Housing <-expenditure$Y
# I am creating a new column within the expenditure data frame for Y
expenditure$Per_Capita_Personal_Income_State <-expenditure$X1
# I am creating a new column within the expenditure data frame for X1.
expenditure$Number_Residents_Financially_Insecure_Per_Hundred_Thousand <-expenditure$X2
# I am creating a new column within the expenditure data frame for X2.
expenditure$Number_People_Residing_Urban_Areas_Per_Thousand <-expenditure$X3
# I am creating a new column within the expenditure data frame for X3.
Per_Capita_Expenditure_Housing <-expenditure$Y
# I am pulling out the new vector so that R recognizes it as Y.
Per_Capita_Personal_Income_State <-expenditure$X1
# I am pulling out the new vector so that R recognizes it as X1.
Number_Residents_Financially_Insecure_Per_Hundred_Thousand <-expenditure$X2
# I am pulling out the new vector so that R recognizes it as X2.
Number_People_Residing_Urban_Areas_Per_Thousand <-expenditure$X3
#I am pulling out the new vector so that R recognizes it as X3.
library(tidyverse)
# I am working in ggplot and I need to restructure the data in the long format. This is because I am comparing multiple x variables to one y variable.
df_long <- pivot_longer(
  df,
  cols = c(Per_Capita_Personal_Income_State,
           Number_Residents_Financially_Insecure_Per_Hundred_Thousand,
           Number_People_Residing_Urban_Areas_Per_Thousand),
  names_to = "Variable",
  values_to = "X_value"
)

# I did the pretty labels to get rid of the dashes between the titles.
pretty_labels <- c(
  Per_Capita_Personal_Income_State = "Per Capita Personal Income",
  Number_Residents_Financially_Insecure_Per_Hundred_Thousand = "Financially Insecure Residents per 100,000",
  Number_People_Residing_Urban_Areas_Per_Thousand = "Urban Residents per 1,000"
)

# I am creating side by side scatter plots using ggplot.I used facet wrap because each x variable has a different scale it is measured by. I used the df long from the step above to be able to compare the multiple x variables against one y variable.
ggplot(df_long, aes(x = X_value, y = Per_Capita_Expenditure_Housing)) +
  geom_point(color = "blue", size = 2, alpha = 0.7) +
  facet_wrap(~Variable, scales = "free_x",
             labeller = labeller(Variable = pretty_labels)) +
  labs(
    title = "Per Capita Housing Assistance vs X Values",
    x = "X Values",
    y = "Per Capita Expenditure on Housing Assistance"
  ) +
  theme_minimal()
ggsave("C:/Users/molly/OneDrive/Documents/GitHub/StatsI_2025/problemSets/PS01/my_answers/Per_Capita_Housing_X_Values.pdf")
# Per capita personal income and urban residents per 1,000 are positively related to each other, meaning states with higher incomes also tend to have more urban residents. Both of these predictors also show a positive relationship with per capita housing assistance: higher income and more urban residents correspond to higher housing expenditure. For financially insecure residents per 100,000, there is no clear relationship either with the other predictors or with housing assistance.
```



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

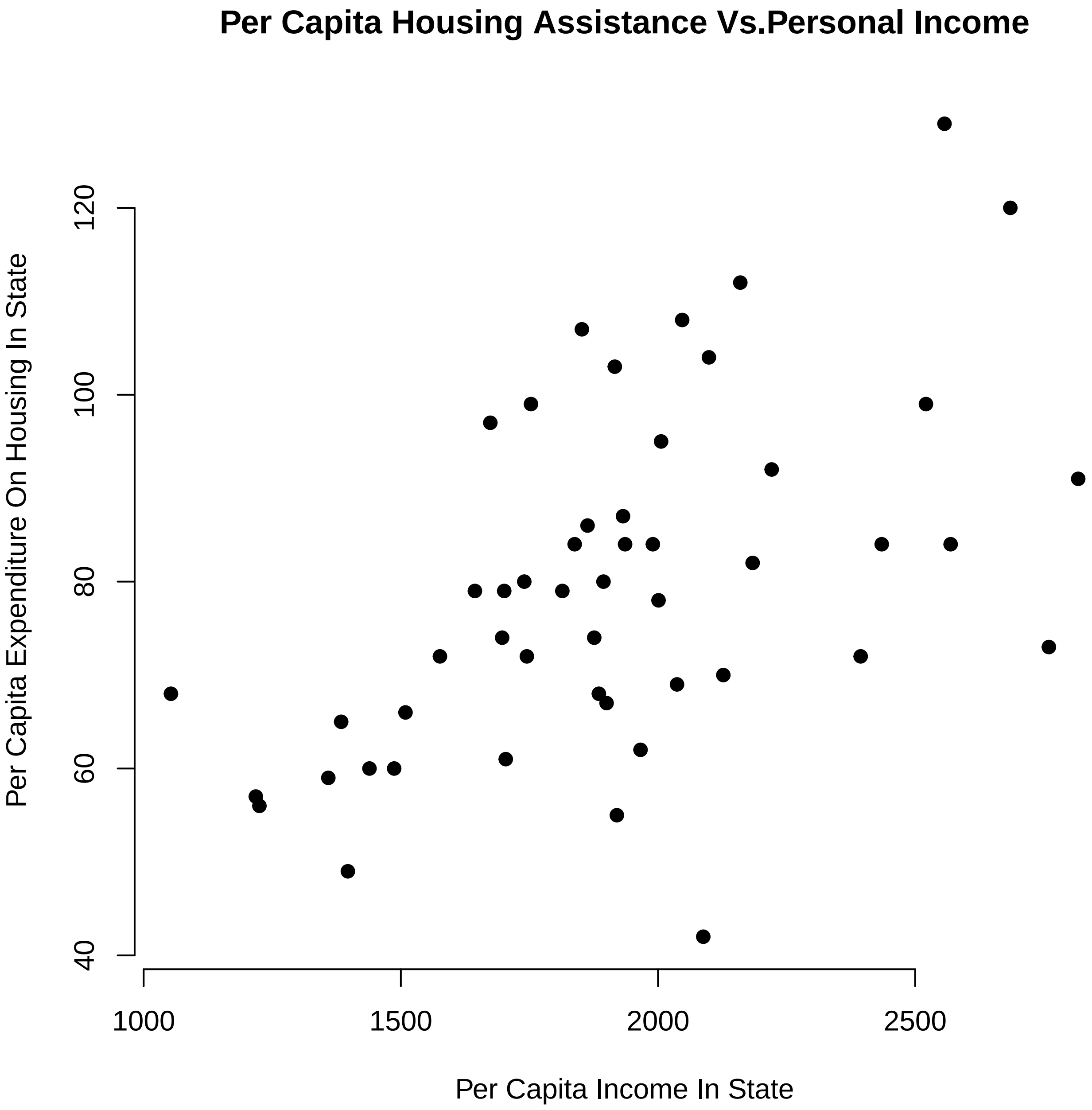
```
expenditure$Region <- factor(expenditure$Region,
                                   levels = c(1, 2, 3, 4),
                                   labels = c("Northeast", "North Central", "South", "West"))
# I did this step to create the region value into a factor. I needed to do this change the numbers into labels to be used in the graph.
expenditure$Per_Capita_Expenditure <- expenditure$Y
# I did this step to create a new value called Per Capita Expenditure that is part of the expenditure data set.
Region <-expenditure$Region
# I created region as a new value as part of the expenditure data set.
avg_expenditure <- tapply(expenditure$Per_Capita_Expenditure, expenditure$Region, mean, na.rm = TRUE)
# I did this step to get the aggregate of the per capita expenditure per region for the graph by the mean.This created a new variable called avg_expenditure.
avg_expenditure
Per_Capita_Expenditure <- expenditure$Y

ggplot(expenditure, aes(x = Region, y = Per_Capita_Expenditure)) +
  geom_col(fill = "blue") +
  labs(
    title = "Per Capita Expenditure Per Region for Housing Assistance",
    x = "Region",
    y = "Per Capita Expenditure"
  ) +
  theme_minimal()
ggsave("C:/Users/molly/OneDrive/Documents/GitHub/StatsI_2025/problemSets/PS01/my_answers/Per_Capita_Expenditure.pdf")
# I needed to create a barplot because one of the variables is categorical and one is numerical.
# The region that has the highest per capita expenditure on housing is the West.
```

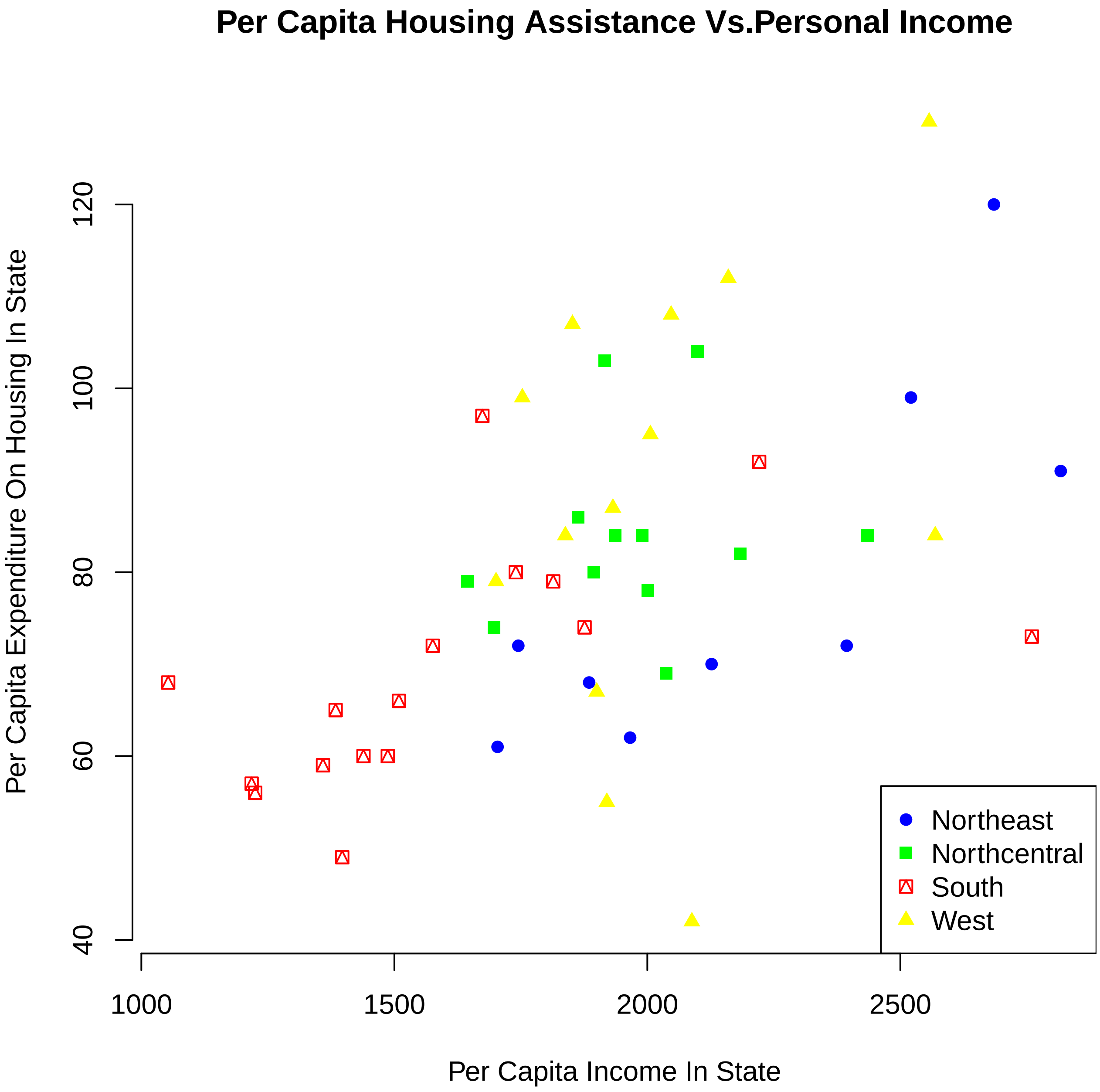


- Please plot the relationship between Y and $X1$? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.

```
expenditure$Per_Capita_Personal_Income_State <-expenditure$X1
Per_Capita_Personal_Income_State <-expenditure$X1
```

```r
# I created a new variable name to name X1 to, "Per Capita Personal Income State."
pdf("C:/Users/molly/OneDrive/Documents/GitHub/StatsI_2025/problemSets/PS01/my_answers/Per_Capita_Housing_Assistance.pdf")
plot(Per_Capita_Personal_Income_State,Per_Capita_Expenditure,main="Per Capita Housing Assistance Vs.Personal Income",xlab="Per Capita Income In State",ylab="Per Capita Expenditure On Housing In State",
      pch=19, frame=FALSE)
dev.off()
# I needed to do this part because this is not ggplot and has to be saved a special way.
# The relationship that is displayed by the graph is that as per capita income in the state goes up, the per capita expenditure on housing also goes up. This is a positive correlation.
colors<-c("Northeast"="blue", "Northcentral"="green","South"="red","West"="yellow")
symbols<-c("Northeast"=16,"Northcentral"=15,"South"=14,"West"=17)
# I am assigning colors and symbols to the different regions.
region_colors<-colors[Region]
# I am assigning a variable called region_colors to represent the colors for the region variable.
region_symbols<-symbols[Region]
# I am assigning a variable called region_symbols to represent symbols for the region variable.
pdf("C:/Users/molly/OneDrive/Documents/GitHub/StatsI_2025/problemSets/PS01/my_answers/Per_Capita_Housing_Assistance_Colors.pdf")
plot(Per_Capita_Personal_Income_State,Per_Capita_Expenditure,main="Per Capita Housing Assistance Vs.Personal Income",xlab="Per Capita Income In State",ylab="Per Capita Expenditure On Housing In State",
     pch=region_symbols,col=region_colors,frame=FALSE)
legend("bottomright",legend=names(colors),col=colors,pch=symbols)
# I am rewriting this code over again to include the newly created variables for color and the symbol. I also made a legend to show what the different symbols mean.
dev.off()
#I did this because this is not ggplot and has to be saved a special way.
```


Per Capita Housing Assistance Vs.Personal Income
Per Capita Expenditure On Housing In State
40
60
80
100
120
1000
1500
2000
2500
Per Capita Income In State


Per Capita Housing Assistance Vs.Personal Income
Per Capita Expenditure On Housing In State
40
60
80
100
120
Per Capita Income In State
1000
1500
2000
2500
Northeast
Northcentral
South
West

-