Problem Set 1

Applied Stats/Quant Methods 1

Due: October 3, 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 3, 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:

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

   Q1.1 Answer:

   To find the 90% Confidence interval for the average student IQ in the school, I first loaded the data set into R Studio by executing the below command. This created an object, y, that was this 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)
```

Next, I calculated the sample mean for this data set with the below command in R Studio:

```
Mean_y <- sum(y)/length(y)
```

This produces the value of 98.44, meaning the avg IQ in the sample of 25 students is 98.44. This is the "sample mean".

To check I have calculated the sample mean correctly, I input the following command:

```
mean(y)
```

This confirms that my initial calculation of the sample mean is correct.

Next, I calculate the sample's standard deviation, denoted by sd_y. To do so I must take the square root of the sum of all values in the sample minus the sample mean squared, divided by the sample size minus 1.

I calculate all of these values with the following input into R Studio:

```
n <- length(y)

nominator <- sqrt(sum((y-Mean_y)*(y-Mean_y)))

denominator <- sqrt(length(y)-1)

SD_y <- nominator/denominator
```

I then calculate the Sample Standard Deviation using the sd() function in R just to confirm my "by hand" calculation was correct:

```
sd(y)
```

Both calculations of the sample Standard Deviation produce the outcome of 13.09287, confirming my "by hand" calculation is correct.

Finally, I will use the sample mean and sample standard deviation to calculate the 90% Confidence Interval (CI) for the avg student IQ in the school. CI is the point estimate plus and minus the margin of error. Because the Sample Size is small and we are assessing means, we will use a t-distribution to calculate the margin of error. The formula I will be putting through R Studio is Mean_y + and - (t_0.05)(SE_y) SE_y is the estimated standard error, which I will calculate first below:

```
SE_y <- SD_y/sqrt(n)
```

t_0.05 is 1.711 (according to the student's t-distribution table).

Finally, I will input the above values to the below equations to identify the upper and lower bands of the 90% confidence interval for the average IQ of students at the school:

```
Upper_CI_y <- Mean_y+1.711*SE_y
Lower_CI_y <- Mean_y-1.711*SE_y
```

This input outputs the Upper_CI_y as 102.92 and the Lower_CI_y as 93.96 (when rounded to two decimal places). I then check my "by hand" calculation using the below formula:

```
t.test(y, conf.level = 0.9, alternative = "two.sided")
```

This confirms I have calculated the 90% Confidence Interval for this sample correctly.

CONCLUSION: 90% of the time, when random sampling from the population of students at the school, the average (or sample mean) IQ will be between 93.96 and 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 $\alpha = 0.05$.

   STEP 1 of conducting my hypothesis test = Outline Assumptions: I assume the data is continuous, sample size is 25, and that the sample was randomly selected.

   STEP 2 = Formulate Hypotheses:

   Null Hypothesis H_0: mu_y is less than or = to 100 (the schools population mean (Mu_y) is less than or equal to country-wide students' mean.

   Alternative Hypothesis H_a: is that mu_y greater than 100 (the population mean for y is greater than the average student IQ country-wide, 100).

   STEP 3 = Calculate the Test Statistic "TS". The Test Statistic summarises how much our data differs from what we would have expected if H_0 is true. I am going to use a t-table test statistic (TS) because I was working with a t-distribution previously (because the sample size is small). To calculate TS I input the following into R Studio:

```
TS <- (Mean_y-100)/SE_y
```

   This outputs a TS of -.5957 (if rounding to 4 decimal points).

   STEP 4 = Calculate the p-value. This is one sided because we are only concerned about whether or not mu_y is greater than the country-wide student IQ, and not if it is greater or less than. The degrees of freedom are n-1 because we are using t-table. To calculate the p-value I input the following into R Studio:

```
p <- pt(abs(TS),df = n-1)
```

   This outputs a p-value of 0.72 (if rounding to 2 decimal points).

   STEP 5 = Make Conclusions.

   The p value = .72, and the level of test (alpha) = .05,(so the p value is greater than alpha). Because p is greater than alpha, we fail to reject the null hypothesis that the students in the school's average IQ is less than or equal to the average IQ of students country-wide. It is therefore unlikely that the average student IQ in her school is higher than the average IQ score of students country-wide.

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`.

I import the expenditure datatset into R with the following inputs:

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

And I explore this data with the following inputs:

```
summary(expenditure)
head(expenditure)
```

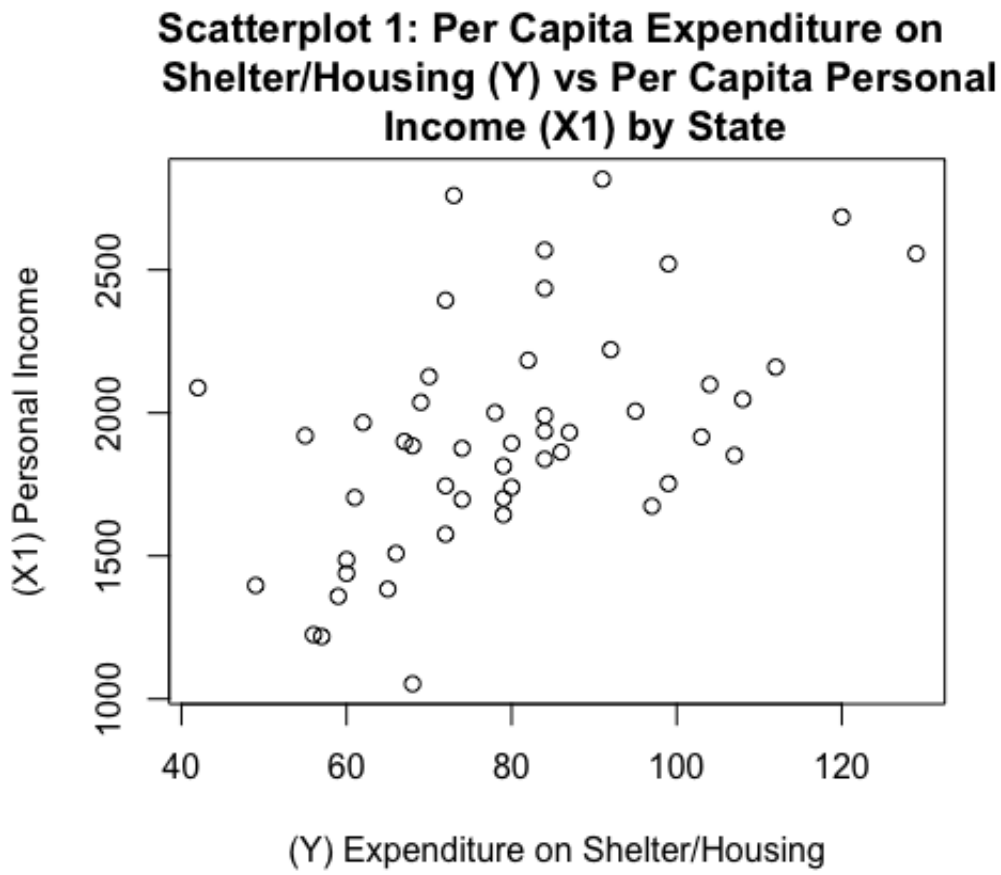
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?

My first step in answering this quesion is to create a vector for each variable in the data set that I am required to plot. I do this with the following inputs, creating 4 separate objects:

```
Y_HExpenditure <- c(expenditure$Y)
X1_Income <- c(expenditure$X1)
X2_FInsecurity <- c(expenditure$X2)
X3_UResidents <- c(expenditure$X3)
```

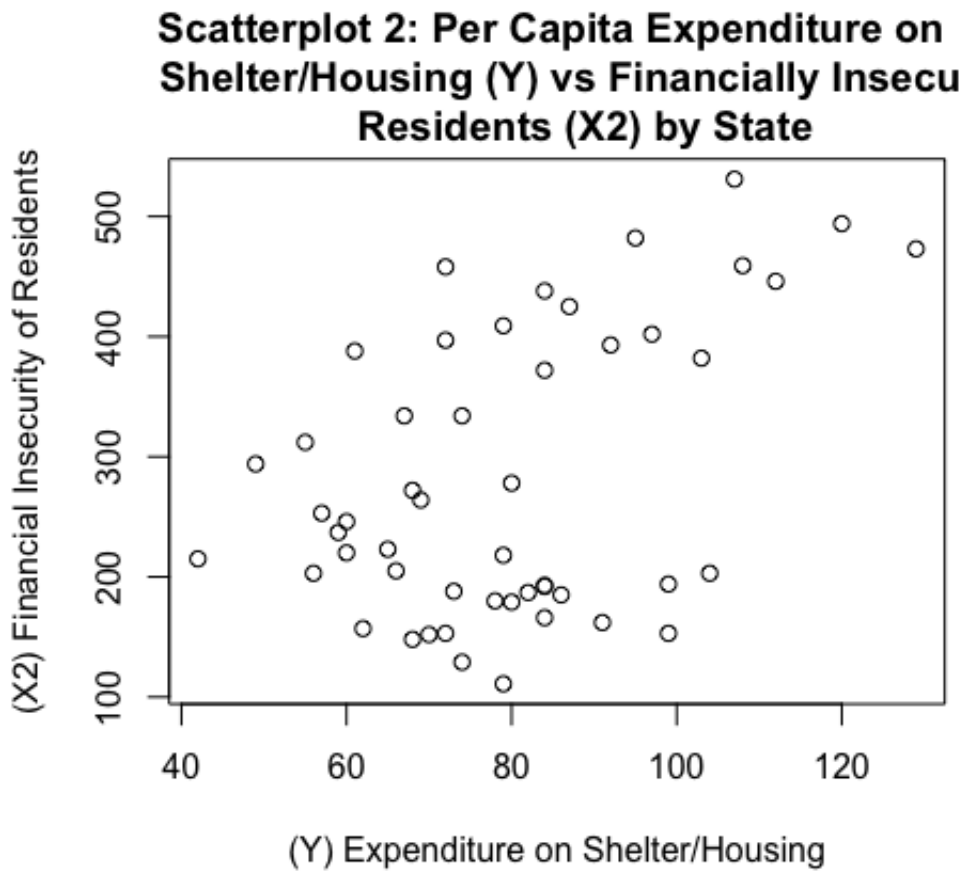
Then, I plot the variables against each other using the inputs listed below. I produce 6 scatterplots total (the R code used to produce each plot is listed above its respective plot on the following pages), so that the relationships between each variable with one another is plotted.

```r
plot(expenditure$Y, expenditure$X1,
     main = "Scatterplot 1: Per Capita Expenditure on
     Shelter/Housing (Y) vs Per Capita Personal
     Income (X1) by State",
     ylab = "(X1) Personal Income",
     xlab = "(Y) Expenditure on Shelter/Housing")
```



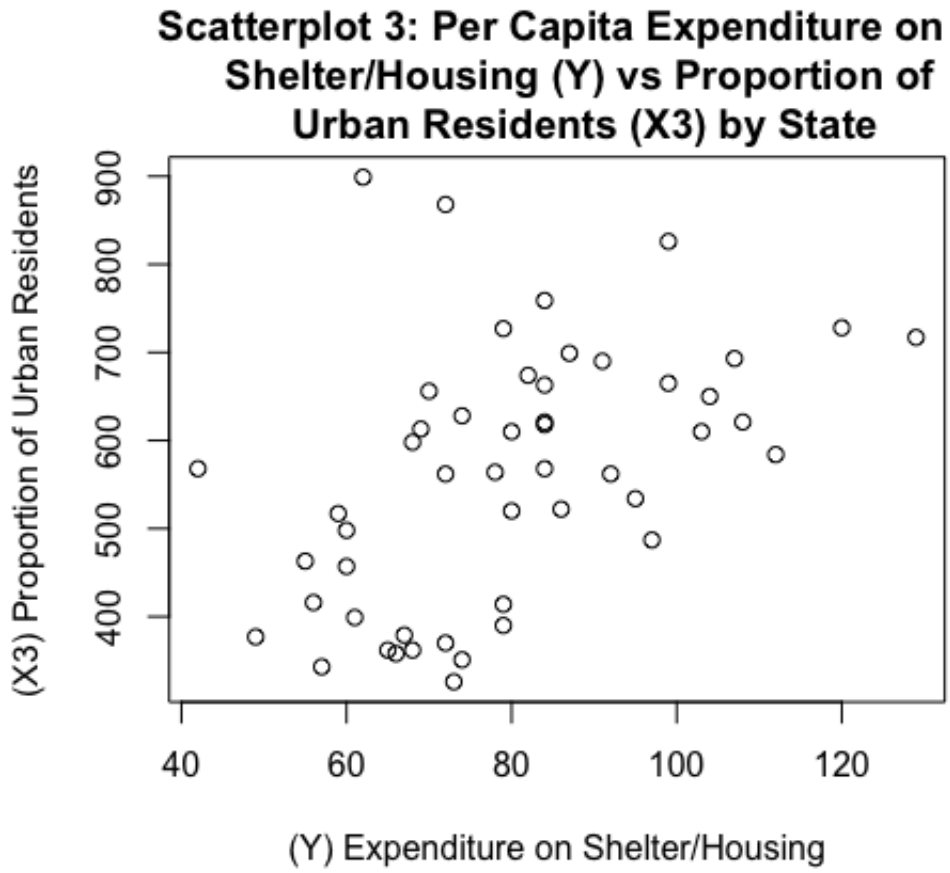
Description of Scatterplot 1: It looks like there is a slight positive, linear relationship between Per Capita Income and Expenditure on Housing based on this graph, because as y increases, x tends to increase as well.

```r
plot(expenditure$Y, expenditure$X2,
     main = "Scatterplot 2: Per Capita Expenditure on
     Shelter/Housing (Y) vs Financially Insecure
     Residents (X2) by State",
     ylab = "(X2) Financial Insecurity of Residents",
     xlab = "(Y) Expenditure on Shelter/Housing")
```



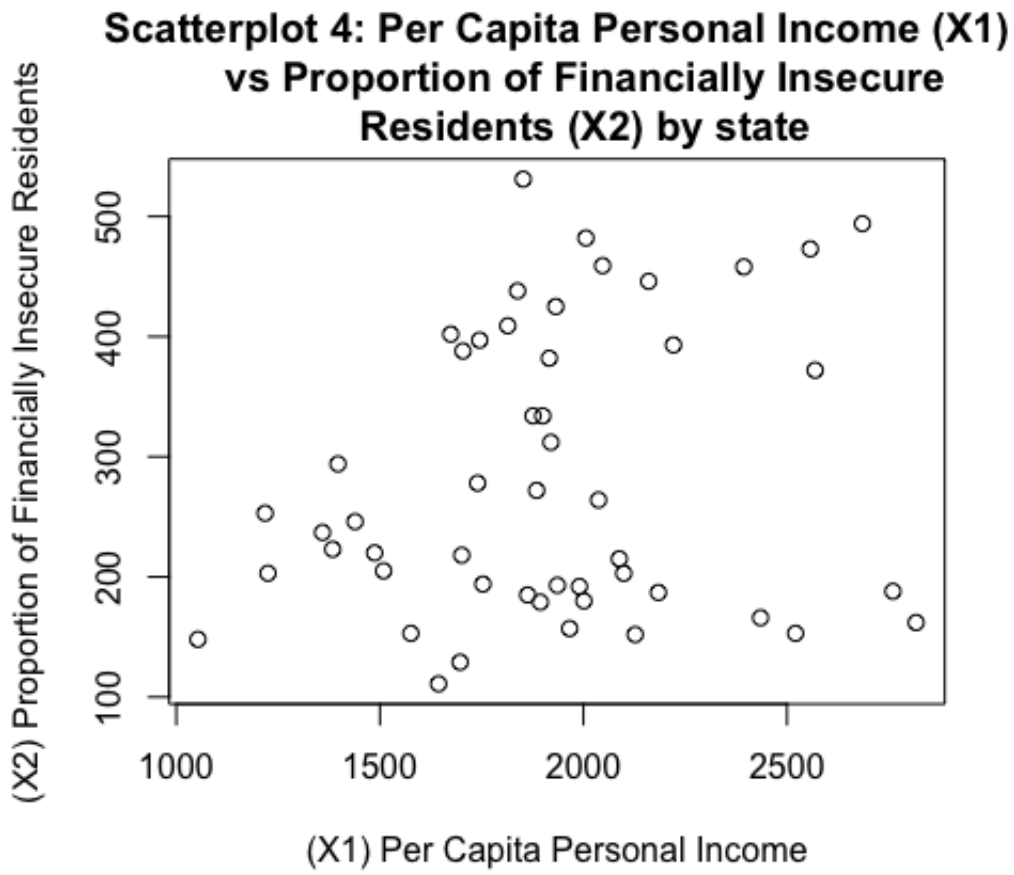
Description of Scatterplot 2: It looks like there is less correlation between Financial security of residents and expenditure on shelter/housing based on this graph (especially if compared to the relationship observed in Scatterplot 1), but there could still be a weak positive correlation between the two variables.

```r
plot(expenditure$Y, expenditure$X3,
     main = "Scatterplot 3: Per Capita Expenditure on
     Shelter/Housing (Y) vs Proportion of
     Urban Residents (X3) by State",
     ylab = "(X3) Proportion of Urban Residents",
     xlab = "(Y) Expenditure on Shelter/Housing")
```



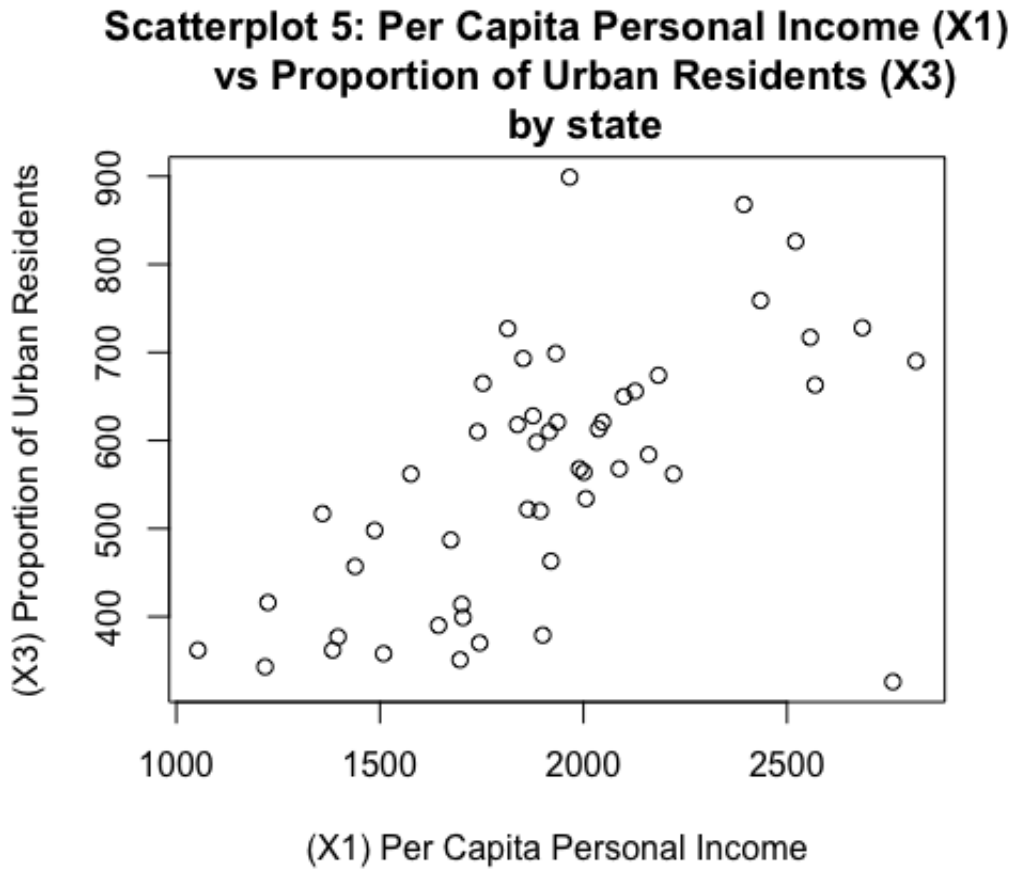
Description of Scatterplot 3: It looks like there is a positive, relatively linear relationship between the proportion of urban residents and Expenditure on Housing in a state based on this graph, because as y increases, x tends to increase as well.

```r
plot(expenditure$X1,expenditure$X2,
     main = "Scatterplot 4: Per Capita Personal Income (X1)
     vs Proportion of Financially Insecure
     Residents (X2) by state",
     ylab = "(X2) Proportion of Financially Insecure Residents",
     xlab = "(X1) Per Capita Personal Income")
```



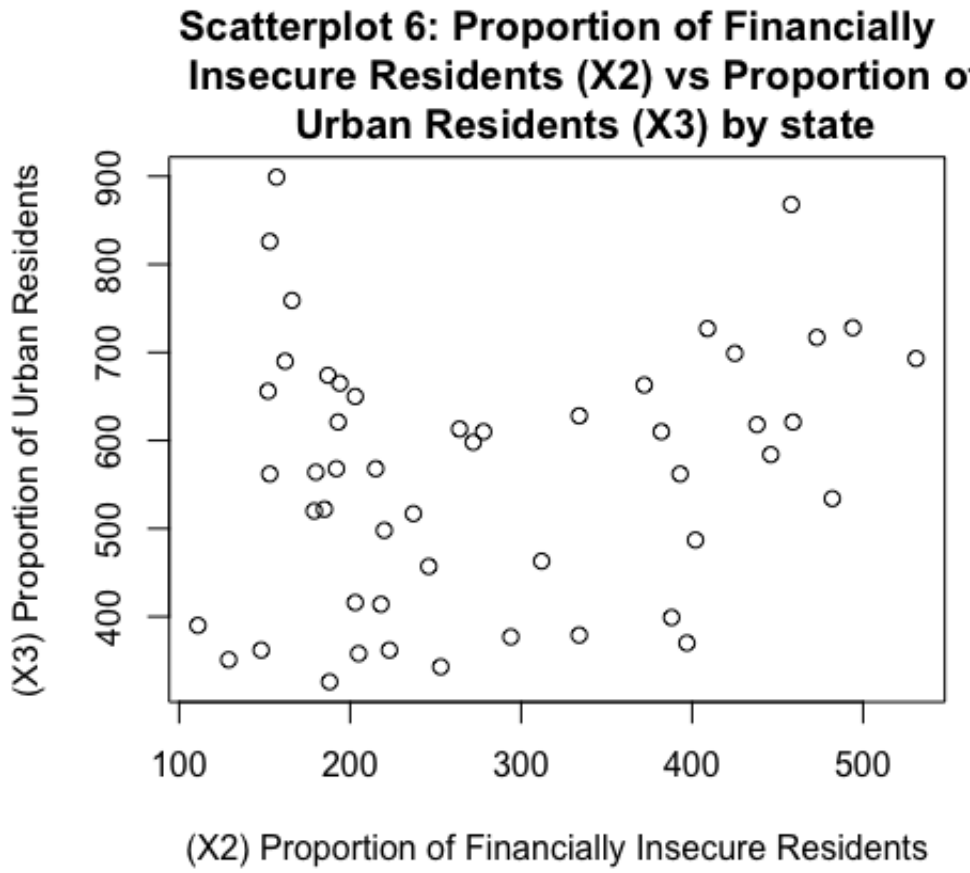
Description of Scatterplot 4: It looks like there is no correlation (or at the very least, a very weak positive relationship) between Per Capita Income and Financially Insecure Residents in a state based on this graph.

```r
plot(expenditure$X1, expenditure$X3,
     main = "Scatterplot 5: Per Capita Personal Income (X1)
     vs Proportion of Urban Residents (X3)
     by state",
     ylab = "(X3) Proportion of Urban Residents",
     xlab = "(X1) Per Capita Personal Income")
```



Description of Scatterplot 5: It looks like there is a strong (compared to the other graphs) positive, linear relationship between Per Capita Income and proportion of urban residents in a state based on this graph, because as y increases, x tends to increase as well, and these points are closer to one another in a linear shape, with less outliers (though there are a few potential outliers observable).

```
plot(expenditure$X2, expenditure$X3,
     main = "Scatterplot 6: Proportion of Financially
     Insecure Residents (X2) vs Proportion of
     Urban Residents (X3) by state",
     ylab = "(X3) Proportion of Urban Residents",
     xlab = "(X2) Proportion of Financially Insecure Residents")
```



Description of Scatterplot 6: It looks like there is little to no correlation between the proportion of urban residents and proportion of financially insecure residents in a state based on this graph, as there does not seem to be any linear relationship in any direction. In comparison to the other 5 Scatterplots, these two variables seem to have the least / weakest correlation with each other.

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

  My first step in answering this question is to create the object "Region" in R. I do this with the following inputs:

```
Region <- expenditure$Region
```

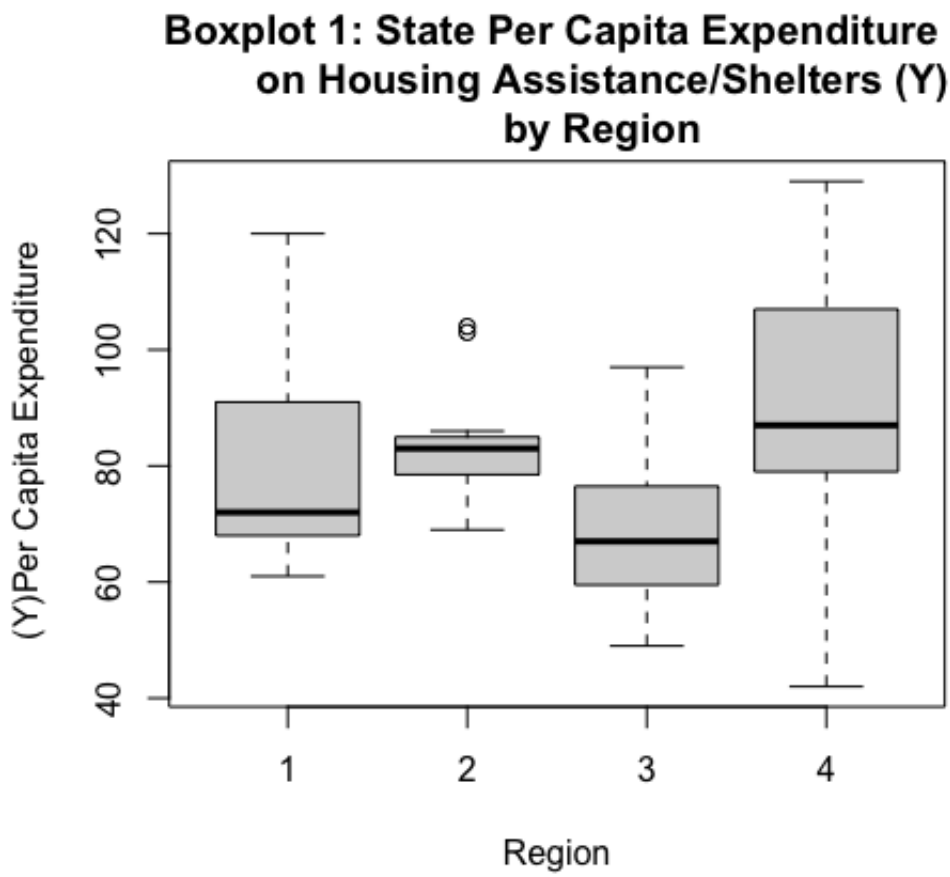
  I then create a data.frame object in R that binds the two variables (Region and Y) with the following inputs:

```
DF_Y_Region <- data.frame(cbind(Y_HExpenditure, Region))
```

  Now, I create the box plot that shows the variation in State level per-capita expenditure on housing assistance/shelters in each Region (1-4), by inputting the following code into R:

```
boxplot(DF_Y_Region$Y_HExpenditure~DF_Y_Region$Region,
        main = "Boxplot 1: State Per Capita Expenditure
        on Housing Assistance/Shelters
        by Region",
        ylab = "Per Capita Expenditure ",
        xlab = "Region")
```

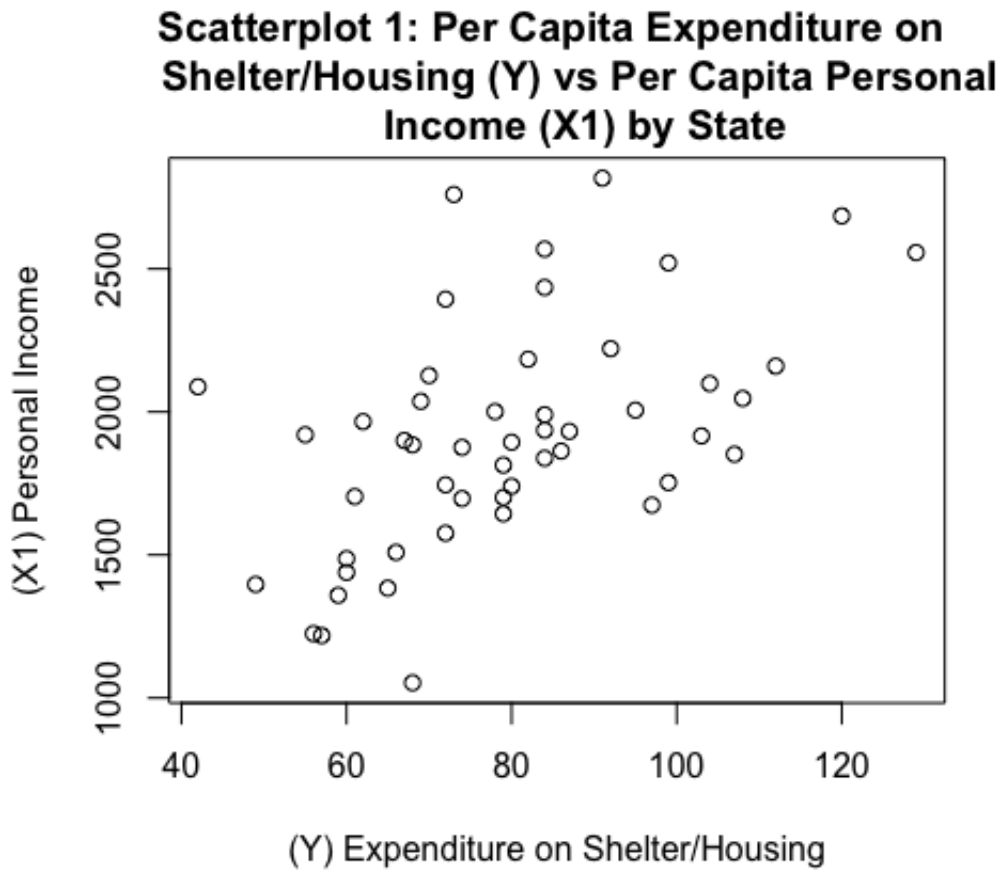
  And this produces Boxplot 1 (on next page):



On average, Region 4 (West) has the highest per capita expenditure on housing assistance. This is observable in Boxplot 1, as we can see the mean (black line) is close to 90, and higher than the means of all other regions. Because the mean = the average, we can see that the average expenditure on assistance is highest in Region 4 (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.

  This relationship between Y (Expenditure on Housing Assistance/Shelter)and X1 (Per Capita Personal Income) by state was plotted in part 1 of this question. The scatterplot that was produced (Scatterplot 1) is below:



  And based on this graph, it seems Y and X1 have a positive, linear relationship. This is because we can see that when Per Capita Income (X1) increases, Expenditure on Housing (Y) also seems to increase on the graph. To understand if there are differences in these values / trends that occur depending on the Region a state is in, we add the Region variable to this plot.

  My first step in adding the Variable, Region, is to turn the Region section of the expenditure data into a Factor with 4 levels (1, 2, 3, and 4). I do this so that each Region will be a different colour on the graph, rather than each region being a gradation of the same colour. My inputs to turn Region into a Factor with 4 levels is below:

```
expenditure$Region <- as.factor(expenditure$Region)
```

Then, I produce Scatterplot 7 with this code, which will separate the initial scatterplot (Scatterplot 1) by Region, using 4 different colours to indicate the 4 different regions:

```
ggplot(expenditure,aes(x=Y_HExpenditure,y=X1_Income, col=Region))+ggtitle
    ("Scatterplot 7: State Expenditure on
       Shelters/Housing Assistance vs Personal
       Income, Differentiated by Region")+geom_point()
```

This code produces Scatterplot 7, as seen below: