Problem Set 1: Solution

Student: Shekhar Kedia (23351315)
Applied Stats/Quant Methods 1

Due: October 1, 2023

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 Sunday October 1, 2023. No late assignments will be accepted.
- Total available points for this homework is 80.

Notes:

- Please note, the responses are nested right after each problem in blue colour.
  For instance-
  `For problem 1.1`, the responses are mentioned right after the question for `1.1`.
- The format for response is in the following manner:
  - Steps to be followed
  - The `R` script
  - Interpretation of result
- The example `.R` file and `.tex` file that was provided, is modified to develop this document.
- My knowledge of library packages in both `R` and `LaTeX` is limited and so, I might have attached packages which are either redundant or unnecessary in doing this assignment.

Question 1 (40 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:

```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)
```

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

   Steps to be followed to find the 90% confidence interval:

   - Calculation of mean for the observations i.e. for the object (or vector), y
   - Calculation of standard error for the observations
   - As the number of observations are less than 30 and the population standard deviation in not known, we use the t-distribution to find the critical t-value using the in-built qt() function in R:

     qt((1-confidence coefficient)/2, df = degrees of freedom, lower.tail = FALSE)

     N.B.: For observations (n) greater than 30, the t-distribution and z-distribution is more or less similar and so, a z-distribution can also be used for n > 30.
   - Then, the confidence interval (both upper and lower bounds) are calculated using the formula:

     Confidence Interval = mean value of the obs. +/- t-value * standard error of the obs.

   R script used for calculations:

```r
len_y <- length(y) #Calculating the number of observations in object y
mean_y <- mean(y) #Calculating the mean value of the obs.
stand_dev_y <- sd(y) #Calculating the standard deviation of the obs.
stand_error_y <- stand_dev_y/sqrt(len_y) #Calculating the standard error

t90 <- qt((1-.90)/2, df = (len_y - 1), lower.tail = FALSE) #Calculating the t-value for given confidence coefficient (i.e. 90%)

lower_ci <- mean_y - (t90 * stand_error_y) #Calculating lower bounds
upper_ci <- mean_y + (t90 * stand_error_y) #Calculating upper bounds
confint90 <- round(c(lower_ci, upper_ci),2) #Combining both bounds of confidence interval rounded upto 2 decimal points

print(round(c(lower_ci, mean_y, upper_ci), 2)) #Printing the 90% confidence intervals

# Using the built-in function to cross verify the findings
t.test(y, conf.level = .90)
```

Interpretation of results:
The average IQ score of students of the school based on the given sample is 98, 90% CI[93.96, 102.92]. It is also cross-verified using the built-in t.test() function which gives the same result.

In other words, this means with repeated sampling there is a 90% probability that the average IQ score for all students in the school is between 93.96 and 102.92.
Or simply, there is a 90% probability that the average IQ score for students in the school lies 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 $\alpha = 0.05$.

   Steps to be followed to carry out the hypothesis testing:

   - Framing the hypothesis based on the research question.
     In this case, the counselor is interested if the average student IQ in her school is higher than 100 (which is the average IQ score among all the schools in the country).
     The null hypothesis,
     H0 = average IQ score of her school <= 100
     Ha = average IQ score of her school > 100

   - We calculate the t-statistics using the formula:

     t-stat = (mean of observed value - population mean) / standard error of the observations

   - Using the t-distribution, we then find the critical t-value. Again, the in-built qt() function in R is used with necessary degrees of freedom.
   - The t-statistics is then compared with the critical t-value. If the t-statistics is greater than the critical t-value, we have sufficient evidence to reject the null hypothesis and vice-versa.
   - We also use the built-in t.test() function in R to cross-verify the findings.
     In this case, we get a p-value and if the p-value is less than equal to $\alpha = 0.05$, we have sufficient evidence to reject the null hypothesis.

R script used for calculations:

```r
t_stat <- (mean_y - 100)/stand_error_y #Calculating the t-statistics

t95 <- qt((1-.95)/2, df = (len_y - 1), lower.tail = FALSE) #Calculating the t-value for given confidence coefficient (i.e. 95%)

compare <- t_stat > t95 #Comparing the observed t-statistics with the critical t-value and storing the results
print(compare)  #Printing the result of the comparison to draw conclusion

# Performing the one-sample t-test using the built-in function to cross verify the findings
t.test(y, alternative = c("greater"), mu = 100)
```

Interpretation of results:

As, the t-statistics is found to be -0.59 which is less than the critical t-value which is 2.06, we have sufficient evidence to not reject the null hypothesis or in other words we accept the null hypothesis which states that the average IQ score of students in the counselor's school is less than equal to the average IQ score (100) among all the schools in the country.

The findings are also cross-verified using the in-built one-sample t.test() function. The output produces a p-value = 0.7215 which is greater than the $\alpha = 0.05$, suggesting we have sufficient evidence to accept the null hypothesis.

In simple terms, we can say there is sufficient evidence that the average IQ score of students in the counselor's school is not higher than 100 (which is the average IQ score among all the schools in the country).

Question 2 (40 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`.

`R` script used for importing the `expenditure` data and viewing/exploring it:

```r
# Reading the expenditure data
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/expenditure.txt", header=T)

View(expenditure) #Viewing the expenditure data to visually understand the structure and spread
```

- 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)?

  Steps to be followed to plot the relationships:

  - We look at relationship between two variables at one point in time i.e., between Y and X1, then between Y and X2, Y and X3, X1 and X2, X1 and X3, and finally between X2 and X3.
  - The plot() function in `R` with proper labels in both the axes is used to plot the scatter graph depicting the pattern of relationship between two variables.
  - The cor() function in `R` is also used to calculate the correlation value indicating the strength and direction of relationship between the two variables.
  - The correlation value is appended in the main title of the graph for ease of referencing.
  - For inserting and formatting the graphs using *LaTeX*, I referred to this website: `https://overleaf.libguides.com/c.php?g=711515&p=5224773`

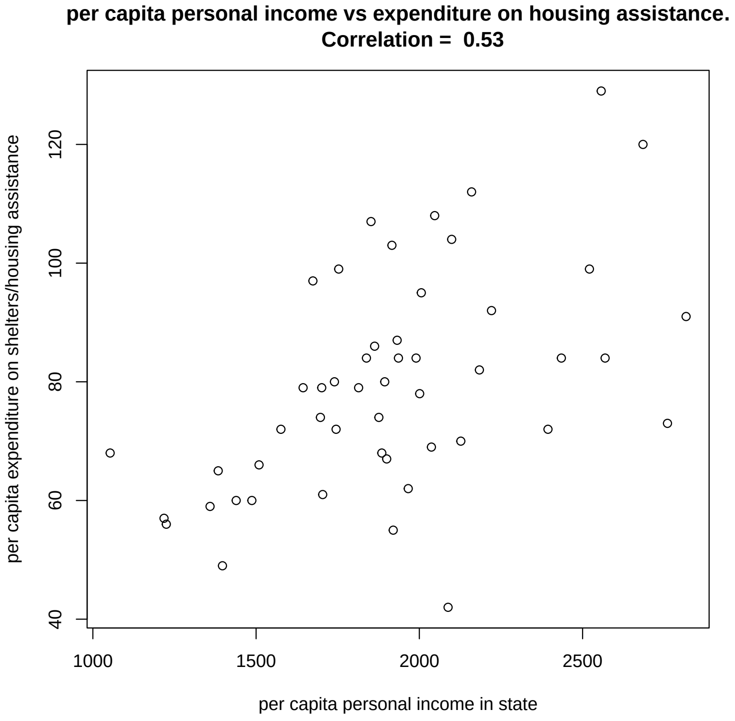
R script used for calculating the correlation values and plotting the relationship between Y and X1:

```
# Creating scatterplot & calculating correlation value to show relationship between Y and X1
cor_Y_X1 <- round(cor(expenditure$Y, expenditure$X1),2) #Shows the default pearson correlation value rounded upto 2 decimals

pdf("plot_Y_X1.pdf") #Creates a pdf file where we can then input/write the scatter plot

plot(expenditure$X1, expenditure$Y,
     xlab = "per capita personal income in state",
     ylab = "per capita expenditure on shelters/housing assistance",
     main = paste("per capita personal income vs expenditure on housing assistance.
     Correlation = ", cor_Y_X1)) #The scatter plot also has the correlation value mentioned in the title
dev.off()
```

The following graph shows the relationship between Y and X1:



Interpretation of results:

We can see that the variables are associated and the correlation value is also moderately high (0.53). This means, the per capita personal income has some relation with the per capita expenditure on housing assistance in state.

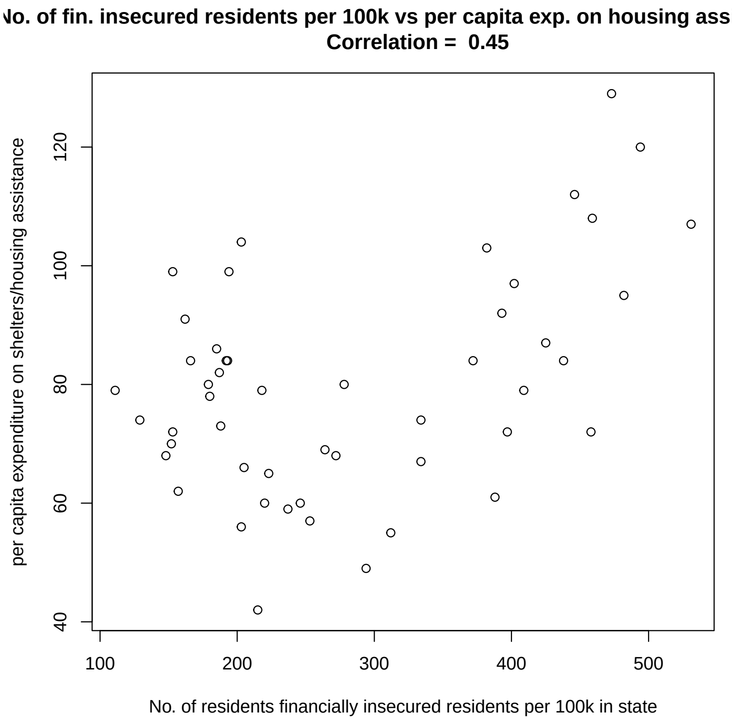
R script used for calculating the correlation values and plotting the relationship between Y and X2:

```r
# Creating scatterplot & calculating correlation value to show
  relationship between Y and X2
cor_Y_X2 <- round(cor(expenditure$Y, expenditure$X2),2)

pdf(”plot_Y_X2.pdf”)

plot(expenditure$X2, expenditure$Y,
     xlab = ”No. of residents financially insecured residents per 100k in
    state”,
     ylab = ”per capita expenditure on shelters/housing assistance”,
     main = paste(”No. of fin. insecured residents per 100k vs per capita
    exp. on housing assistance.
     Correlation = ”, cor_Y_X2))
dev.off()
```

The following graph shows the relationship between Y and X2:



Interpretation of results:

We can see that the variables are mildly associated and the correlation value is also moderate (0.45). However, we also see a lot of dispersions in per capita expenditure on housing assistance with lower values of number of residents financially insecured per 100k in state.

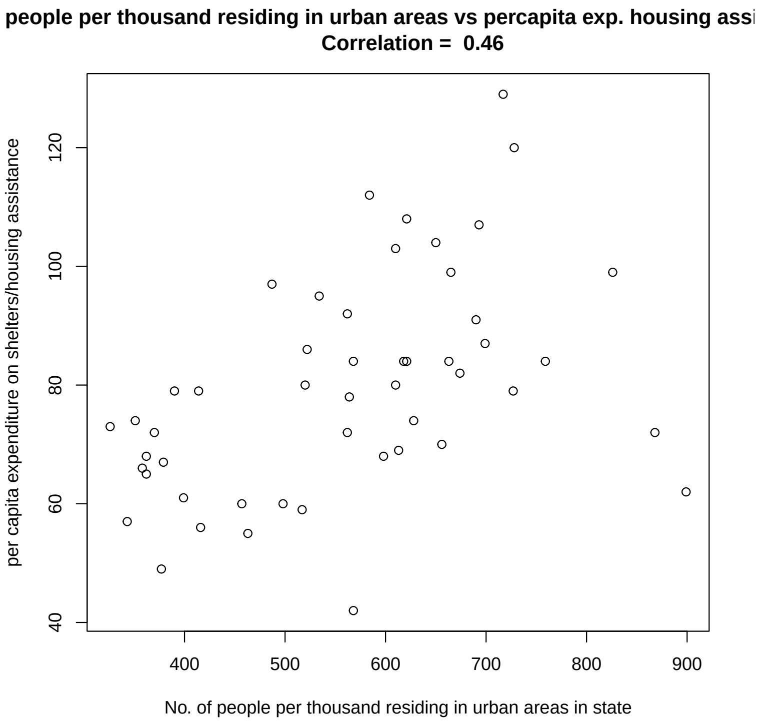
R script used for calculating the correlation values and plotting the relationship between Y and X3:

```r
# Creating scatterplot & calculating correlation value to show relationship between Y and X3
cor_Y_X3 <- round(cor(expenditure$Y, expenditure$X3),2)

pdf("plot_Y_X3.pdf")

plot(expenditure$X3, expenditure$Y,
     xlab = "No. of people per thousand residing in urban areas in state",
     ylab = "per capita expenditure on shelters/housing assistance",
     main = paste("No. people per thousand residing in urban areas vs percapita exp. housing assistance.
     Correlation = ", cor_Y_X3))
dev.off()
```

The following graph shows the relationship between Y and X3:



Interpretation of results:

We can see that the variables are mildly associated and the correlation value is also moderate (0.46). However, we see that there are some outliers in the data especially with higher values of number of people per thousand residing in urban areas in state.

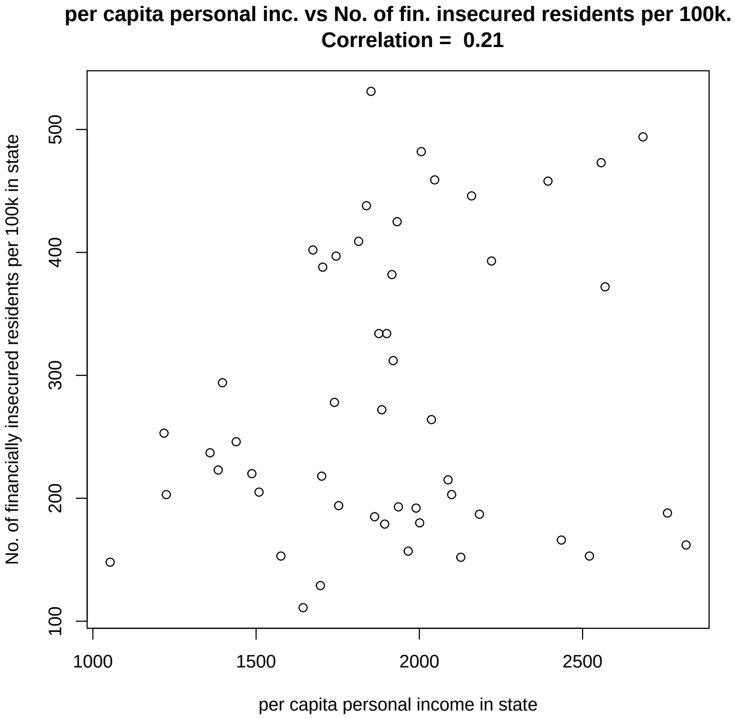
R script used for calculating the correlation values and plotting the relationship between X1 and X2:

```r
# Creating scatterplot & calculating correlation value to show relationship between X1 and X2
cor_X1_X2 <- round(cor(expenditure$X1, expenditure$X2),2)

pdf(”plot_X1_X2.pdf”)

plot(expenditure$X1, expenditure$X2,
     xlab = ”per capita personal income in state”,
     ylab = ”No. of financially insecured residents per 100k in state”,
     main = paste(”per capita personal inc. vs No. of fin. insecured residents per 100k.
      Correlation = ”, cor_X1_X2))
dev.off()
```

The following graph shows the relationship between X1 and X2:



Interpretation of results:

We can see that the variables are not associated and the correlation value is also very low (0.21). This means, the per capita personal income has no relation with number of financially insecured residents per 100k in state.

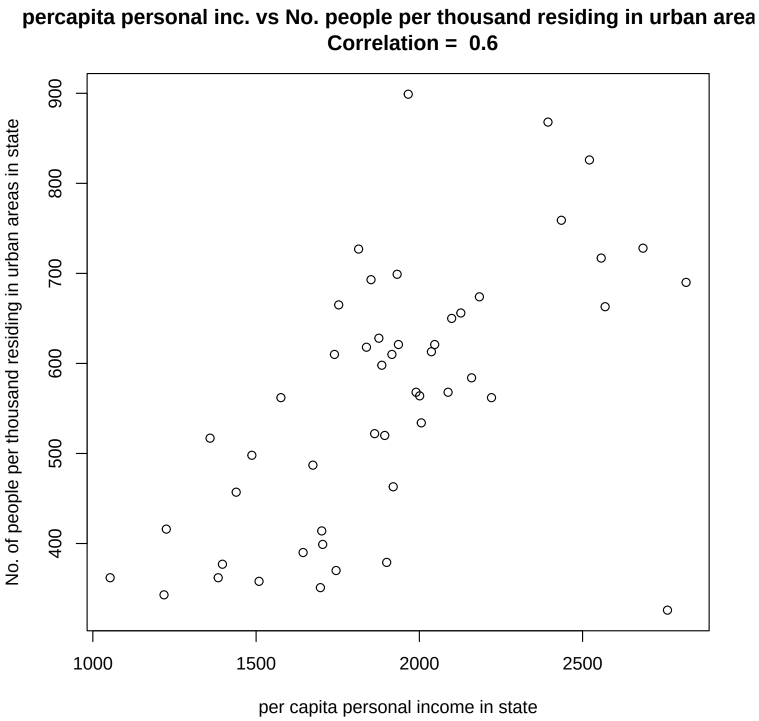
R script used for calculating the correlation values and plotting the relationship between X1 and X3:

```
# Creating scatterplot & calculating correlation value to show
  relationship between X1 and X3
cor_X1_X3 <- round(cor(expenditure$X1, expenditure$X3),2)

pdf(”plot_X1_X3.pdf”)

plot(expenditure$X1, expenditure$X3,
     xlab = ”per capita personal income in state”,
     ylab = ”No. of people per thousand residing in urban areas in state”
  ,
     main = paste(”percapita personal inc. vs No. people per thousand
  residing in urban areas.
     Correlation = ”, cor_X1_X3))
dev.off()
```

The following graph shows the relationship between X1 and X3:



Interpretation of results:

We can see that the variables are associated and the correlation value is also moderately high (0.60). This means, the per capita personal income has some relation with number of people per thousand residing in urban areas in state.

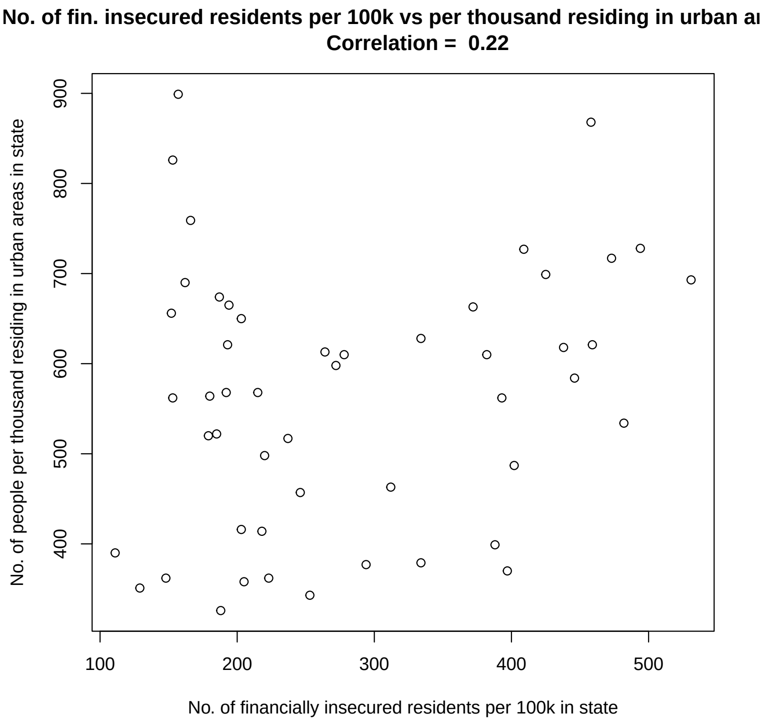
R script used for calculating the correlation values and plotting the relationship between X2 and X3:

```r
# Creating scatterplot & calculating correlation value to show relationship between X2 and X3
cor_X2_X3 <- round(cor(expenditure$X2, expenditure$X3),2)

pdf("plot_X2_X3.pdf")

plot(expenditure$X2, expenditure$X3,
     xlab = "No. of financially insecured residents per 100k in state",
     ylab = "No. of people per thousand residing in urban areas in state",
     main = paste("No. of fin. insecured residents per 100k vs per thousand residing in urban areas.
     Correlation = ", cor_X2_X3))
dev.off()
```

The following graph shows the relationship between X2 and X3:



Interpretation of results:

We can see that the variables are not associated and the correlation value is also very low (0.22). This means, the number of financially insecured residents per 100k has no relation with number of people per thousand residing in urban areas in state.

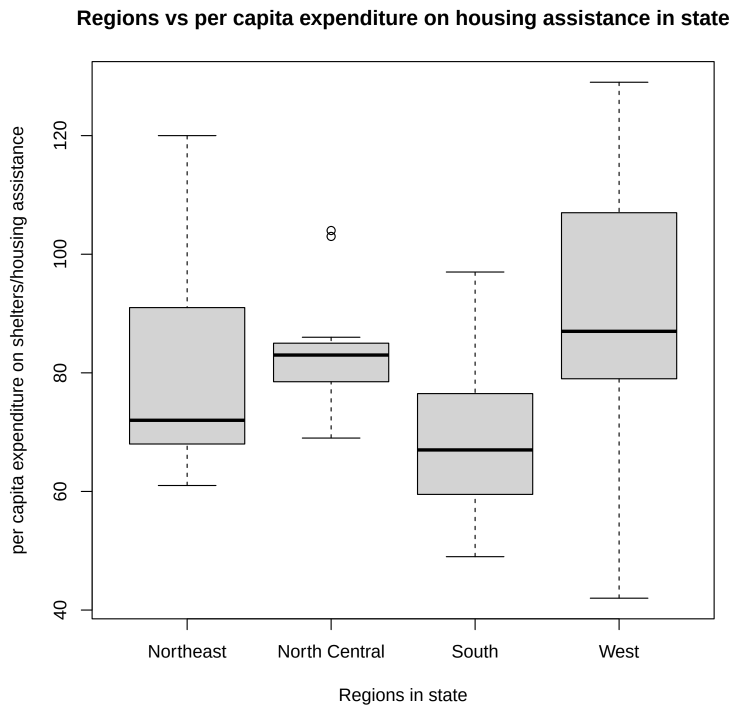
- Please plot the relationship between Y and *Region*? On average, which region has the highest per capita expenditure on housing assistance?

R script used for plotting the relationship between Y and Region:

```
# Plotting relationship between Y and Region
pdf(”plot_Y_Region.pdf”)

boxplot(Y~Region,
      xlab = ”Regions in state”,
      ylab = ”per capita expenditure on shelters/housing assistance”,
      data = expenditure,
      main = ”Regions vs per capita expenditure on housing assistance in state”,
      names = c(”Northeast”, ”North Central”, ”South”, ”West”)
      )
dev.off()
```

The following graph shows the relationship between Y and Region:



Interpretation of results:

From the above boxplot visually we see that the median for *Region* 4 or *West* is greater than other regions. We also calculated the mean (mentioned in the R script) and found that on average, *Region* 4 (*West*) has the highest per capita expenditure on housing assistance with 88.30769.

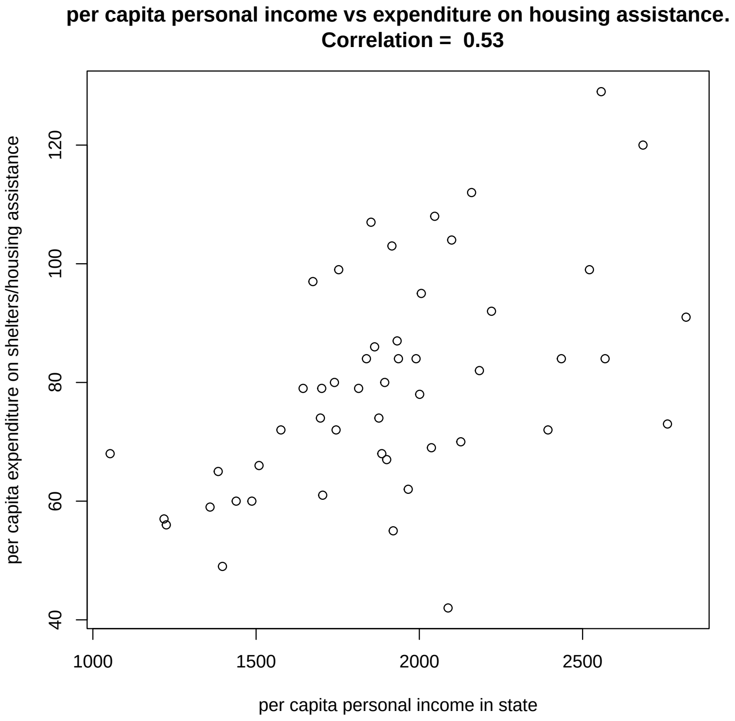
- Please plot the relationship between *Y* and *X1*? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.

  R script used for calculating the correlation values and plotting the relationship between Y and X1:

```
# Plotting relationship between Y and X1
  # Already covered in 2.1 and so, repeating the same here
pdf("plot_Y_X1.pdf")

plot(expenditure$X1, expenditure$Y,
     xlab = "per capita personal income in state",
     ylab = "per capita expenditure on shelters/housing assistance",
     main = paste("per capita personal income vs expenditure on housing assistance.
     Correlation = ", cor_Y_X1))
dev.off()
```

  The following graph shows the relationship between Y and X1:



  Interpretation of results:
  As discussed in 2.1, we can see that the variables are associated and the correlation value is also moderately high (0.53). Indicating, the per capita personal income has some relation with the per capita expenditure on housing assistance in state.

Reproducing the graph including the *Region* variable and displaying the regions with different types of symbols and colours.

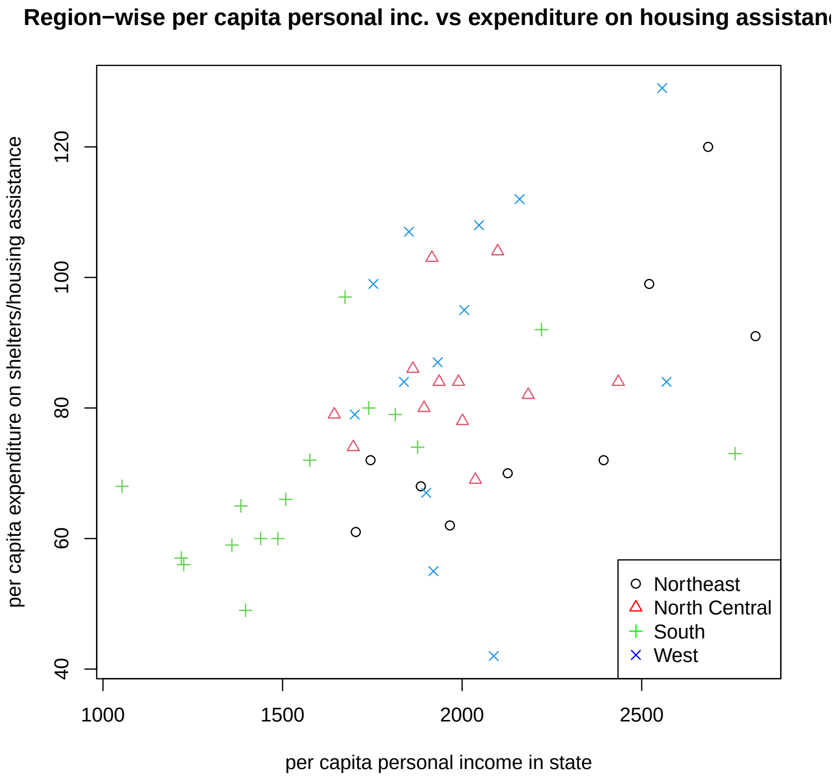
Following R script is used:

```r
# Adding Region variable to the plot and adding legends to depict different colours and shapes
pdf("plot_Y_X1_new.pdf")

plot(expenditure$X1, expenditure$Y, col = expenditure$Region, pch = as.numeric(factor(expenditure$Region)),
     xlab = "per capita personal income in state",
     ylab = "per capita expenditure on shelters/housing assistance",
     main = "Region-wise per capita personal inc. vs expenditure on housing assistance")
legend("bottomright",
     legend = c("Northeast", "North Central", "South", "West"),
     col = c("black","red","green","blue"),
     pch = 1:4)

dev.off()
```

The following graph shows the relationship between Y and X1 alongwith the *Region* variable:



Interpretation of results:

- Firstly, as explained above, we can see the per capita personal income has some relation (correlation value = 0.53) with the per capita expenditure on housing assistance in state. In terms of region, we see *Region* 4 or the *West* region takes on an average greater value for per capita expenditure on housing assistance for any given value of per capita personal income in state.
- *Region* 2 or the *North Central* region is mostly clustered around the middle of the graph with the least dispersion.
- Furthermore, the *Region* 1 or *Northeast* and *Region* 3 or *South* region are widely dispersed with *Region* 1 mostly taking higher values of per capita personal income compared to *Region* 3 which takes mostly lower values of per capita personal income. Also, the mean per capita expenditure on housing assistance for *Region* 1 is higher than *Region* 3.

End of document

Last edits made on: October 1, 2023