Problem Set 1

Applied Stats/Quant Methods 1

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.

```
n <- length(na.omit(student_iqs))
sample_mean <- mean (student_iqs, na.rm = TRUE)
sample_sd <- sd(student_iqs, na.rm = TRUE)
t90 <- qt((1 - .90)/2, df = 24, lower.tail = FALSE)
lower_90 <- sample_mean- (t90 * (sample_sd/sqrt(n)))
upper_90 <- sample_mean + (t90 * (sample_sd/sqrt(n)))
confint_90 <- c(lower_90, upper_90)
```

We are 90 percent confident that the true population mean of student IQ in the school lies around 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 $\alpha = 0.05$.

```r
#Ho <= 100
#Ha > 100
t.test(student_iqs, mu = 100, alternative = "greater")
```

For this hypothesis test, the null hypothesis is that the average of student IQ among all schools in the country is less than or equal to 100.The alternative hypothesis states that the sample school's average IQ is higher than 100. Therefore, this question is looking for an upper tail, one sided t test. Using the sample information for the part before, the t statistic can be calculated with this equation:

t = (sample mean - hypothesized population mean)/(standard deviation/square root of the sample size) = (98.44 - 100) / (13.093/sqrt(25)) = -0.59574.

In order to discover if the null hypothesis can be rejected, we need to find the p value. To find it outside of coding, you use the degrees of freedom (24), the t value (-0.59574), and type of test (one-sided)to look up in a t table for the approx p-value (it tends to be a range). However, in r, the mathematical function above can calcuate the p value given the degrees of freedom, t value, and one-sided test to get the p value of approx. 0.7215. Since the calculated p value is higher than our alpha of 0.05, we fail to reject the null hypothesis. This means that there is insufficent evidence to conclude that the sample school's average IQ is higher than 100.

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.

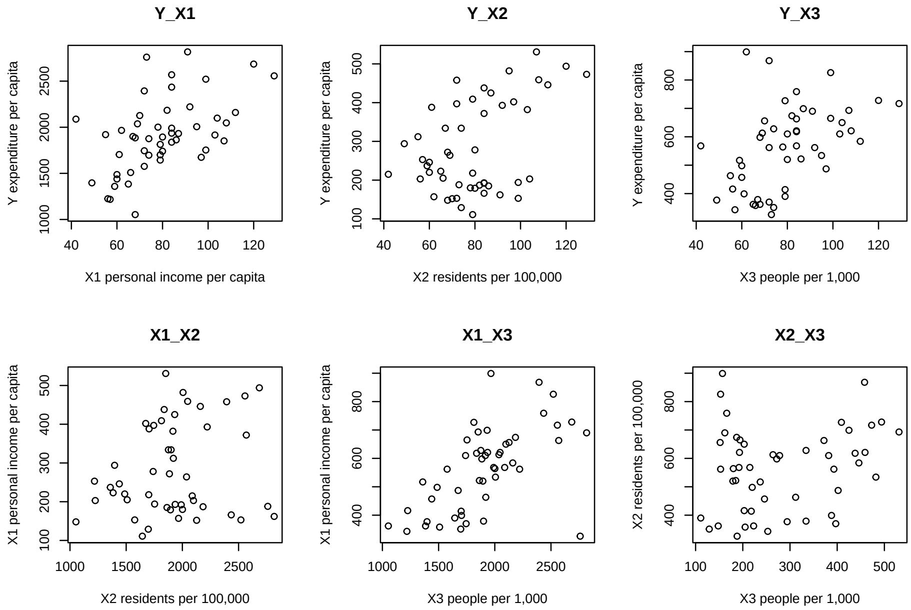
`State`	*50 states in US*
`Y`	*per capita expenditure on shelters/housing assistance in state*
`X1`	*per capita personal income in state*
`X2`	*Number of residents per 100,000 that are "financially insecure" in state*
`X3`	*Number of people per thousand residing in urban areas in state*
`Region`	*1=Northeast, 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)?

```
pdf("combined_plots.pdf", width = 12, height = 8)
par(mfrow = c(2,3))
plot(expenditure$Y,expenditure$X1,
     main = "Y_X1",
     ylab = "Y expenditure per capita",
     xlab = "X1 personal income per capita"
)

plot(expenditure$Y,expenditure$X2,
     main = "Y_X2",
     ylab = "Y expenditure per capita",
     xlab = "X2 residents per 100,000"
)

plot(expenditure$Y,expenditure$X3,
     main = "Y_X3",
     ylab = "Y expenditure per capita ",
     xlab = "X3 people per 1,000"
)

plot(expenditure$X1,expenditure$X2,
     main = "X1_X2",
     ylab = "X1 personal income per capita",
     xlab = "X2 residents per 100,000"
)

plot(expenditure$X1,expenditure$X3,
     main = "X1_X3",
     ylab = "X1 personal income per capita",
     xlab = "X3 people per 1,000"
)

plot(expenditure$X2,expenditure$X3,
     main = "X2_X3",
     ylab = "X2 residents per 100,000",
     xlab = "X3 people per 1,000"
)
```

Figure 1: Relationships between Y, X1, X2, and X3.



YX1 The relationship between expentiture on shelter/housing assistance and personal income shows a moderate positive linear relationship. As personal income increases, expenditure on shelter/housing assistance also tends to increase with less points scattered across the graph.

YX2 The relationship between expentiture on shelter/housing assistance and residents that are "financially insecure" shows a weak positive linear relationship. Below the 300 mark on the Y axis, the values are more clumped together, however, above 300, as the number of residents increase the expenditure tends to also increase. Therefore, there might be some variable that indicates more of a correlation when the baseline of expenditures is at 300, however, there is no evidence provided to support this claim.

YX3 The relationship between expentiture on shelter/housing assistance and people residing in urban areas shows a moderate positive linear relationship. As the number of people living in urban areas increases, the expenditure also tends to increase.

X1X2 The relationship between personal income per capita and residents that are "financially insecure" per 100,000 shows weak to no correlation. The plot points are more randomly scattered around the graph with no clear pattern.

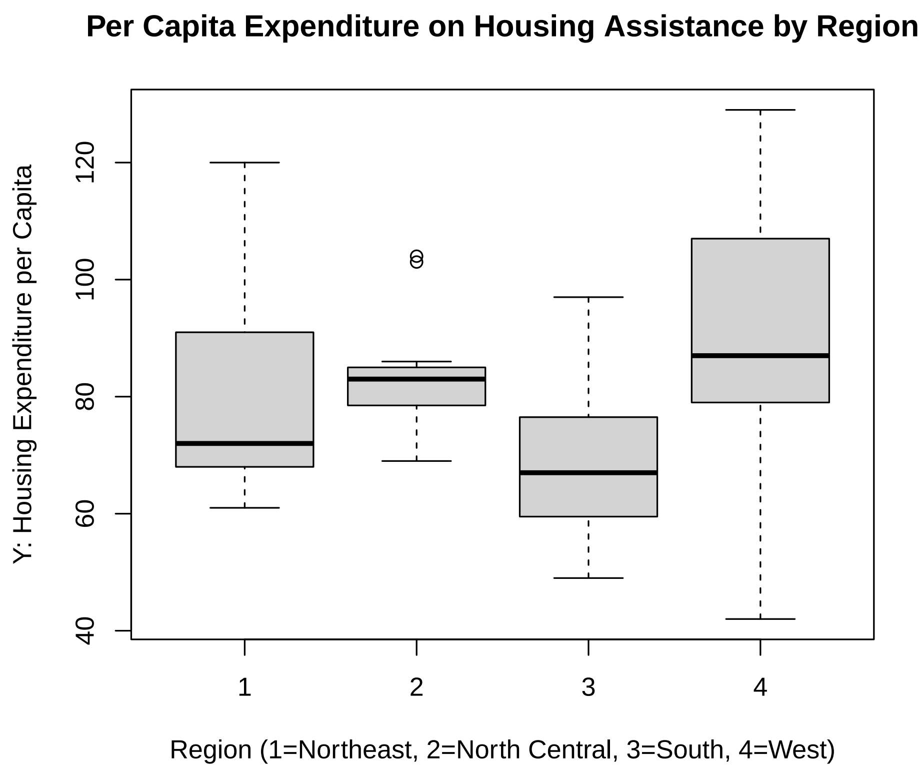
X1X3 The relationship between personal income per capita and people residing in urban areas shows a moderate to strong positive linear correlation. Therefore, as the amount of people residing in urban areas increase, the personal income also tends to increase with less points scattered.

X2X3 The relationship between residents that are "finacially insecure" and people residing in urban area shows a weak positive correlation. The points show that as the number of people residing in urban areas increase, the nummber of residents that are "finacially insecure" also tends to increase, but with more points scattered on the graph.

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

```
par(mfrow = c(1,1))
boxplot(Y ~ Region, data = expenditure,
        main = "Per Capita Expenditure on Housing Assistance by Region",
        xlab = "Region (1=Northeast, 2=North Central, 3=South, 4=West)",
        ylab = "Y: Housing Expenditure per Capita")
```

Figure 2: Relationships Y and Region.

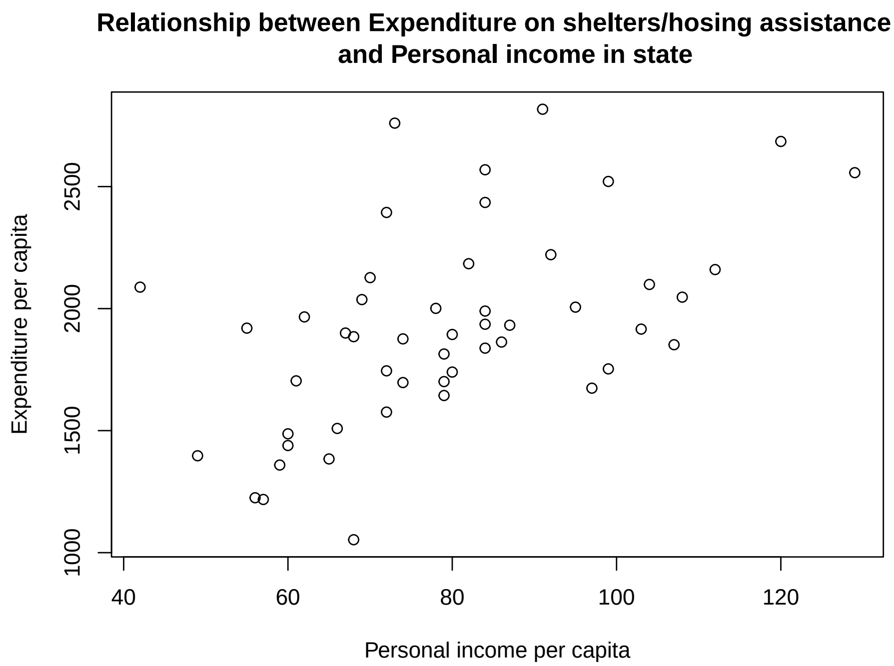


For this question, the best way to visualize the different averages for each region is by using box plots. By looking at the box plots, the West region has the highest per capita expenditure on housing/shelter assistance average at approximately 85 .

- 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.

```
plot(expenditure$Y,expenditure$X1,
     main = "Relationship between Expenditure on shelters/hosing
  assistance
     and Personal income in state",
     ylab = "Expenditure per capita",
     xlab = "Personal income per capita"
)
```

Figure 3: Relationship between Y and Region.



The relationship between expenditure on shelter/housing assistance and personal income per capita in state shows a moderate positive linear relationship. The points are less scattered and show as the personal income increase, the expenditure on housing/shelter assistance also tends to increase.

```r
regions <- unique(expenditure$Region)
colors <- c("forestgreen","skyblue","salmon","rosybrown")[1:length(
    regions)]
shapes <- c(16, 17, 18, 15)[1:length(regions)]
point_colors <- colors[match(expenditure$Region, regions)]
point_shapes <- shapes[match(expenditure$Region, regions)]

plot(expenditure$Y,expenditure$X1,
      col = point_colors,
      pch = point_shapes,
      main = "Relationship between Expenditure on shelters/hosing
    assistance
      and Personal income in state",
      ylab = "Expenditure per capita",
      xlab = "Personal income per capita"
)
legend(
  "bottomright",
  legend = region_names,
  col = colors,
  pch = shapes,
  inset = 0.0
  )
```

Figure 4: Relationship between Y and Region (color/shape).