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:

```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.

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 $\alpha = 0.05$.

Solution - Question 1: Education

1. Confidence Interval for the Average Student IQ in the School

Since the sampling size is smaller than 30, we cannot assume that there is a normal sampling distribution. Instead, we will assume a t-distribution with 25-1 = 24 degrees of freedom (because the sample has n = 25 observations).

The process for determining the confidence interval is outlined below:

- Calculate the mean

```
mean_y = mean(y)
mean_y
```

```
[1] 98.44
```

- Calculate the sum of the squared errors to get the variance, and then the standard deviation (the square root of the variance)

```
demeanedSum <- y - mean_y
squaredError <- demeanedSum ^ 2
variance <- sum(squaredError)/(length(y)-1)
stdev = sqrt(variance)
stdev
```

```
[1] 13.09287
```

- Calculate the standard error

```
sterror = stdev/sqrt(length(y))
sterror
```

```
[1] 2.618575
```

- Find the t-score for the desired confidence level, with 24 degrees of freedom
  I calculated the t-score instead of a Z-score because the assumed sampling distribution is a t-distribution with 24 degrees of freedom.

```
t = qt(p = (1 - 0.9)/2 , df = 24, lower.tail = FALSE)
t
```

```
[1] 1.710882
```

- Construct the confidence interval by calculating its lower and upper limits

```
lowerlimit = mean_y - t * sterror
upperlimit = mean_y + t * sterror
lowerlimit
upperlimit
```

```
> lowerlimit
[1] 93.95993
> upperlimit
[1] 102.9201
```

- **Conclusion:** The interval [93.96, 102.92] contains the true average of student IQs at the school (the true population mean) at least 90% of the time, with repeated random sampling.

2. Hypothesis Testing

- **Step 1:** assumptions
  - the data is quantitative and continuous
  - the sampling method is random
  - the sample size is smaller than 30, so we assume it follows a t-distribution with 24 degrees of freedom and not a normal distribution; the test statistic that was calculated is a t-score
- **Step 2:** formulate hypotheses
  - **Null hypothesis:** the average student IQ in the school (the mean) is lower than or equal to 100
  - **Alternative hypothesis:** the average student IQ in the school (the mean) is higher than 100
- **Step 3:** calculate the test statistic : t-score

```
ts = (mean_y - 100)/sterror
ts
```

```
[1] -0.5957439
```

- **Step 4:** calculate the p-value

  Because the hypothesis test that is carried out is one-sided, and the alternative hypothesis is that the true population mean is greater than 100, we are interested in the upper tail of the distribution. Hence, the lower.tail argument was set to FALSE.

```
p = pt(ts, df = 24, lower.tail = FALSE)
p
```

```
[1] 0.7215383
```

- **Step 5:** draw a conclusion

  The p-value is much higher than the alpha value (0.05). Thus, we cannot reject the null hypothesis that the average student IQ in the school is lower than or equal to 100. The test result is not statistically significant.

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)?

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

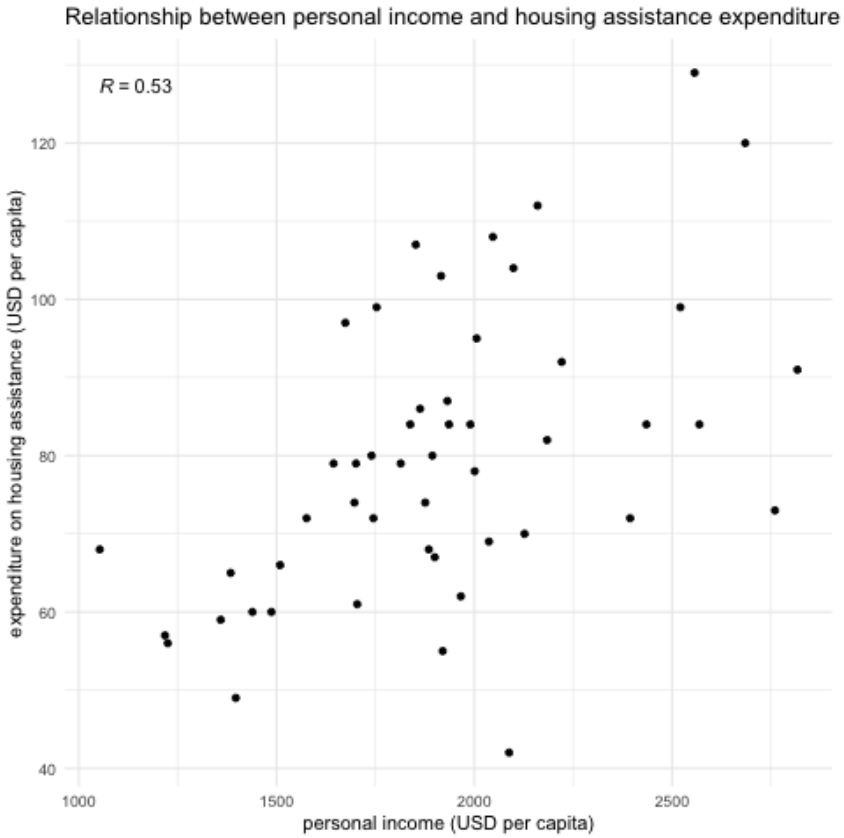
- Please plot the relationship between *Y* and *X1*? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.

Solution - Question 2: Political Economy

Plot the relationships among X1, X2, X3 and Y

```
ggplot(expenditure, aes(x = X1, y = Y)) +
  geom_point() +
  labs(title = "Relationship between personal income
          and housing assistance expenditure",
       x = "personal income (USD per capita)",
       y = "expenditure on housing assistance (USD per capita)") +
  stat_cor(aes(label = ..r.label..)) +
  theme_minimal()
```

Figure 1: Scatterplot of X1 and Y

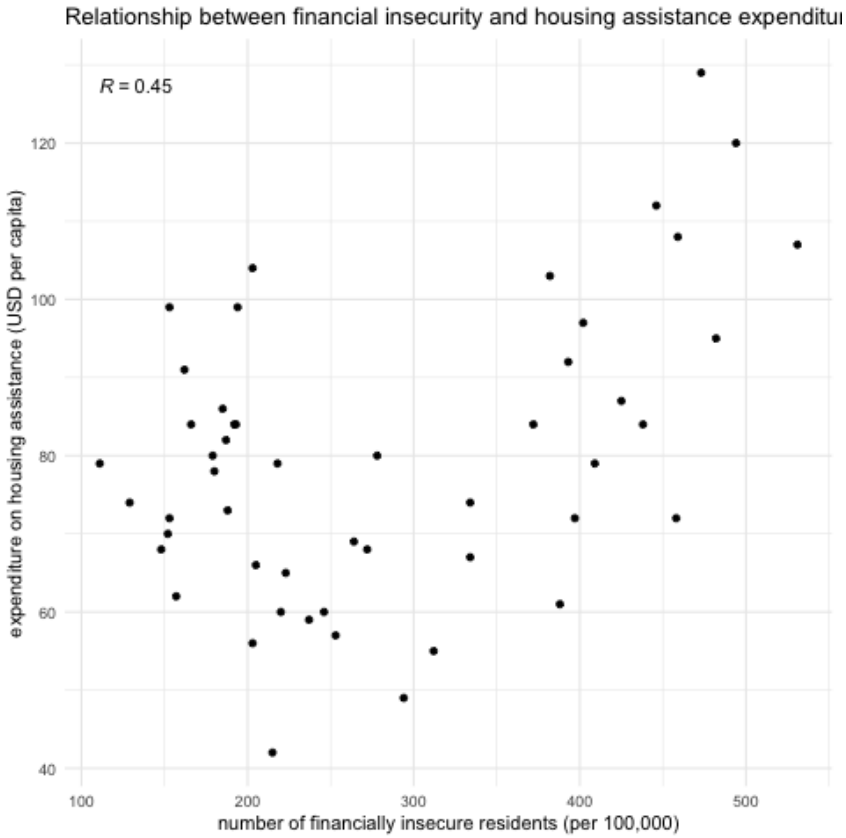


Interpretation:

There is a moderate positive association between the personal income and the expenditure on housing assistance per capita in a state. The Pearson correlation coefficient points towards the same conclusion, as its value is above 0, but not close enough to 1 to suggest a strong association between the two variables. A positive linear pattern can also be distinguished by looking at the plot.

```r
ggplot(expenditure, aes(x = X2, y = Y)) +
  geom_point() +
  labs(title = "Relationship between financial insecurity
          and housing assistance expenditure",
       x = "number of financially insecure residents (per 100,000)",
       y = "expenditure on housing assistance (USD per capita)") +
  stat_cor(aes(label = ..r.label..)) +
  theme_minimal()
```

Figure 2: Scatterplot of X2 and Y

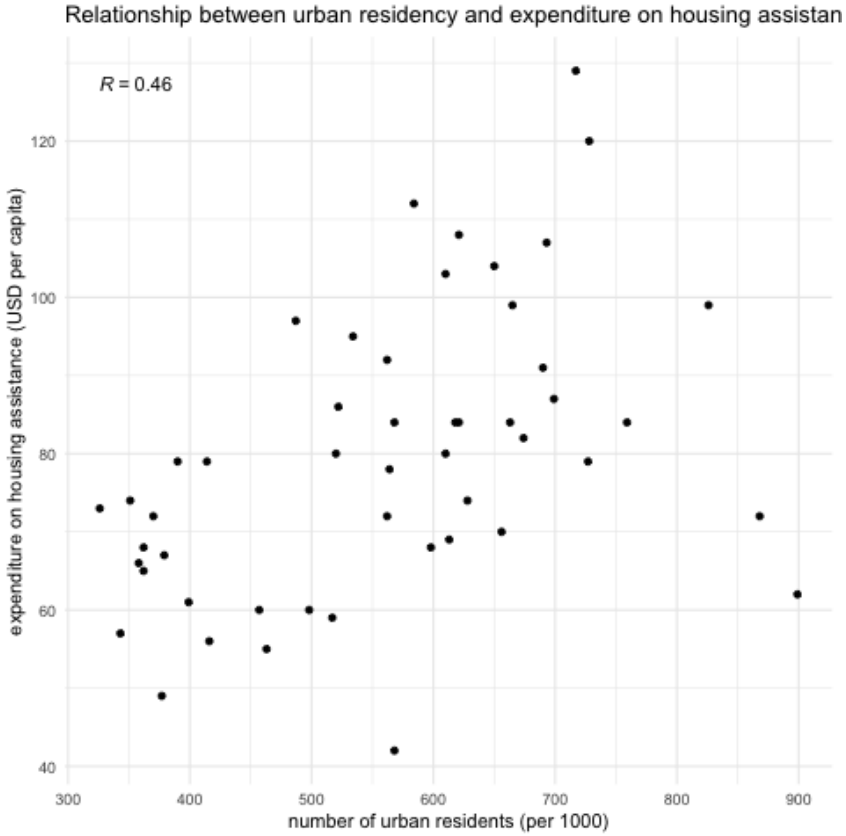


Interpretation:

There is a moderate positive association between the number of "financially insecure" residents and the expenditure on housing assistance per capita in a state. The Pearson correlation coefficient suggests the same, as its value is above 0, but not above 0.5. Visually analysing the plot, there is an upward linear trend that can be identified, but there are many dots that fall outside of it, which suggests that the association is fairly weak.

```r
ggplot(expenditure, aes(x = X3, y = Y)) +
  geom_point() +
  labs(title = "Relationship between urban residency and
          expenditure on housing assistance (in state)",
       x = "number of urban residents (per 1000)",
       y = "expenditure on housing assistance (USD per capita)") +
  stat_cor(aes(label = ..r.label..)) +
  theme_minimal()
```

Figure 3: Scatterplot of X3 and Y

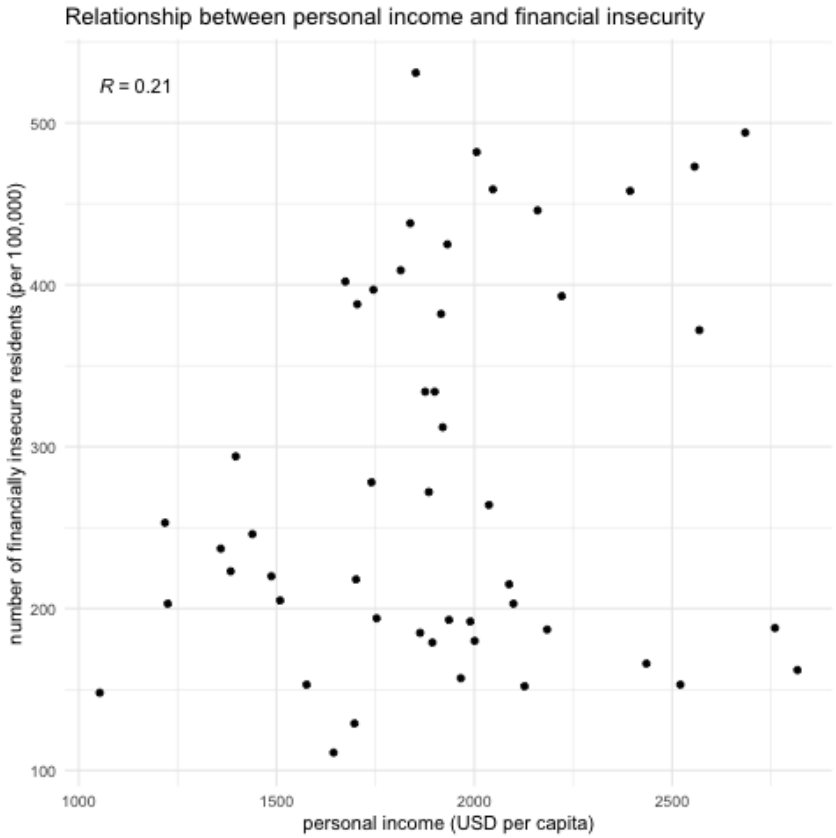


Interpretation:

There is a mild positive association also between the number of residents in urban areas and the expenditure on housing assistance per capita in a state. The value of the Pearson correlation coefficient is 0.46, suggesting a weak association between the two variables. There is a faint upward linear trend that can be observed from the plot, but the dots are very loosely grouped around it.

```r
ggplot(expenditure, aes(x = X1, y = X2)) +
  geom_point() +
  labs(title = "Relationship between personal income
          and financial insecurity",
       x = "personal income (USD per capita)",
       y = "number of financially insecure residents (per 100,000)")+
  stat_cor(aes(label = ..r.label..)) +
  theme_minimal()
```

Figure 4: Scatterplot of X1 and X2

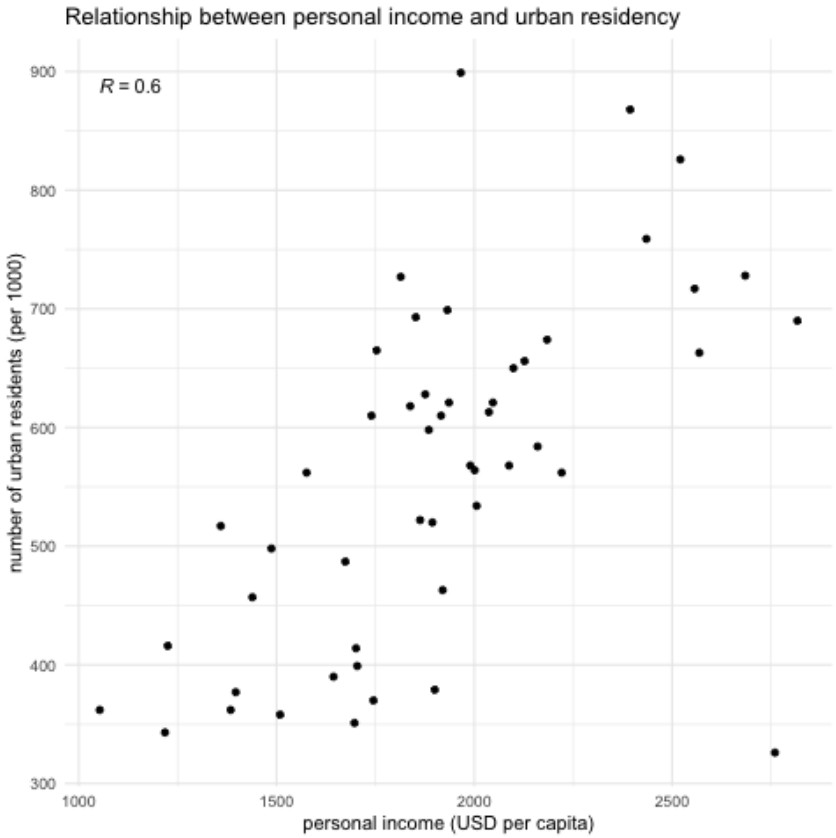


Interpretation:

There doesn't seem to be a linear association between personal income per capita and the number of "financially insecure" residents within a state. The data points appear to be quite scattered in the graph, showing no clear linear trend. The Pearson correlation coefficient has a value of 0.21, indicating that we can only observe a very weak, if any, association between the two variables.

```r
ggplot(expenditure, aes(x = X1, y = X3)) +
  geom_point() +
  labs(title = "Relationship between personal income
          and urban residency",
       x = "personal income (USD per capita)",
       y = "number of urban residents (per 1000)") +
  stat_cor(aes(label = ..r.label..)) +
  theme_minimal()
```

Figure 5: Scatterplot of X1 and X3

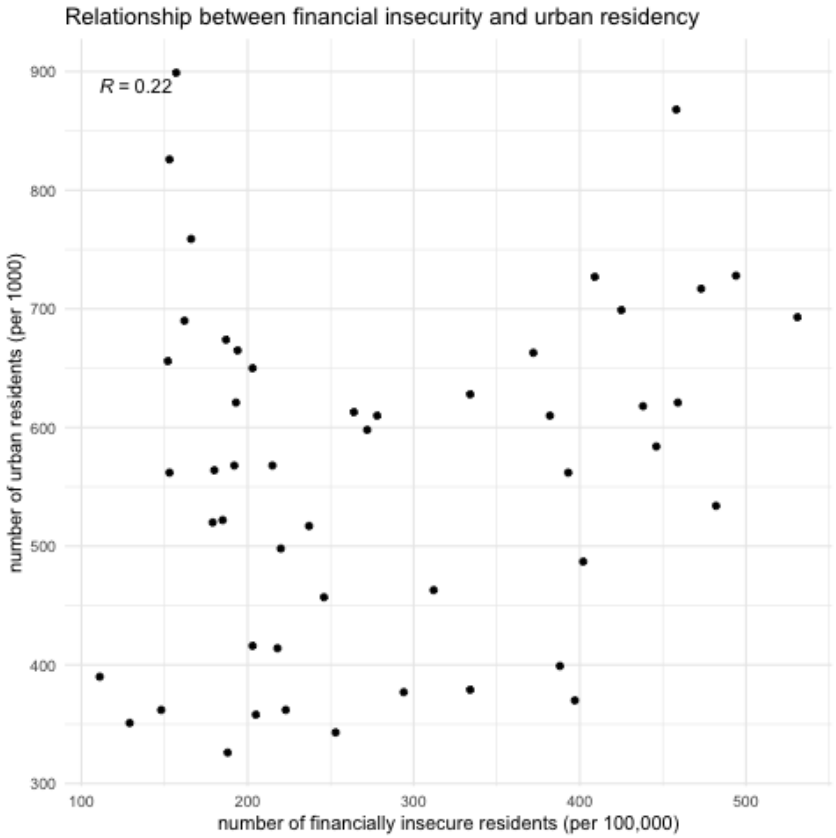


Interpretation:

There is a moderate positive association between the personal income per capita and the number of urban residents in a state. The value of the Pearson correlation coefficient shows that these two variables have the strongest correlation within the dataset, but the value is still not high enough to indicate a strong association. Visually inspecting the plot, it can be observed that the dots follow an upward linear trend, but they are not very tightly grouped around it.

```r
ggplot(expenditure, aes(x = X2, y = X3)) +
  geom_point() +
  labs(title = "Relationship between financial insecurity
          and urban residency",
       x = "number of financially insecure residents (per 100,000)",
       y = "number of urban residents (per 1000)") +
  stat_cor(aes(label = ..r.label..)) +
  theme_minimal()
```

Figure 6: Scatterplot of X2 and X3



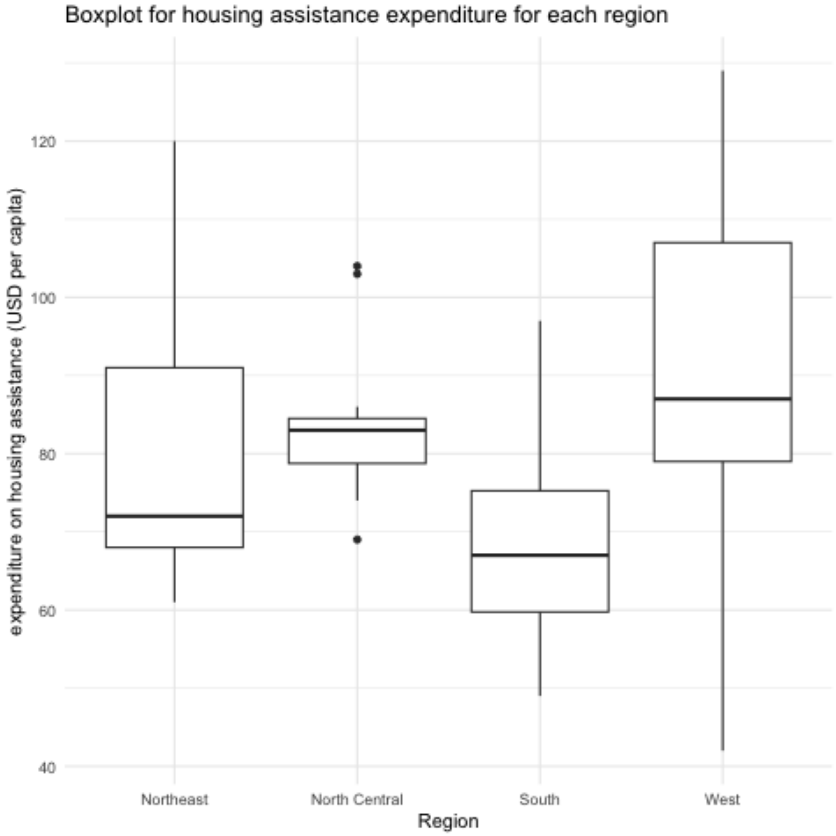
Interpretation:

There doesn't seem to be a linear association between the number of "financially insecure" residents and the number of urban residents in a state. The dots are spread out quite evenly within the plot, without any distinguishable pattern. The Pearson correlation coefficient is quite close to 0, further indicating the absence of a meaningful association between the two variables.

Plot the relationship between Y and Region

```
ggplot(expenditure, aes(x=Region, y=Y, group=Region)) +
   geom_boxplot() +
   theme(
      legend.position="none",
      plot.title = element_text(size=11)) +
      labs(title = "Boxplot for housing assistance expenditure for each
    region",
           x = "Region",
           y = "expenditure on housing assistance (USD per capita)") +
  theme_minimal()
```

Figure 7: Boxplot of Y and Region



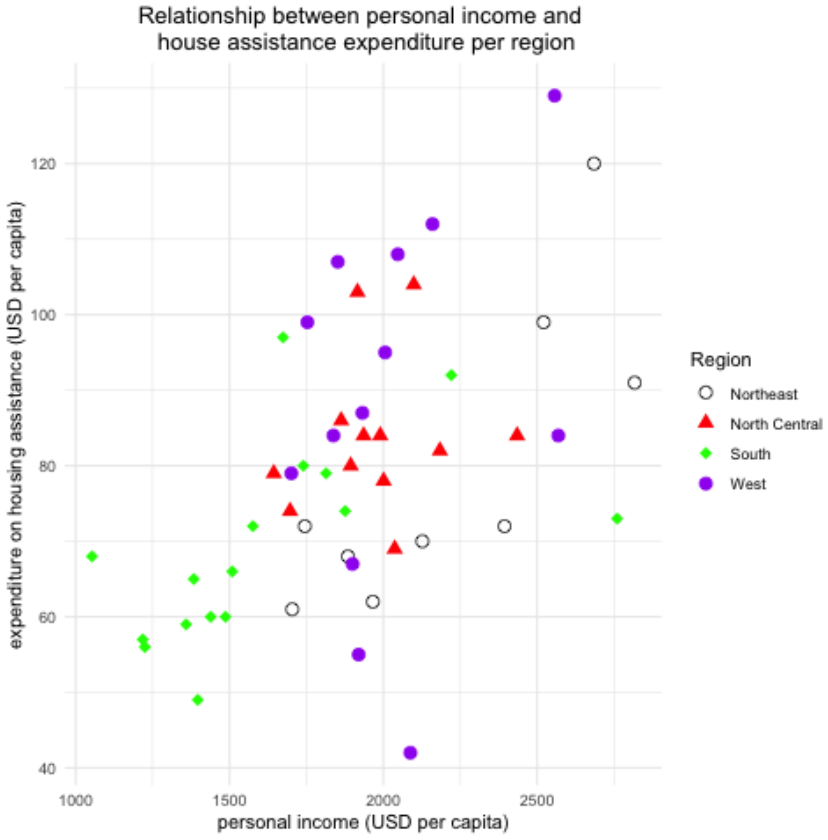
Interpretation:

In order to visualize the average expenditure on housing assistance in each region, I generated a boxplot. This type of plot was chosen because it shows the distribution of data for each region, indicating the median but also the variability of the values.
It can be observed that the West region has the highest median for expenditure on housing assistance per capita and its interquartile range starts and ends at higher values than the others. Thus, the West spends on average the most out of the regions to provide housing assistance.

Plot the relationship between Y, X1 and Region

```
ggplot(expenditure, aes(x = X1, y = Y, shape = factor(Region), color =
    factor(Region))) +
  geom_point(size = 3) +  # increase size to distinguish between the shapes
  scale_color_manual(values = c("black", "red", "green", "purple")) +
  scale_shape_manual(values = c(21, 17, 18, 19)) +  # each value
    corresponds to a shape
  labs(title = "Relationship between personal income and
          \n house assistance expenditure per region",
       x = "personal income (USD per capita)",
       y = "expenditure on housing assistance (USD per capita)",
       shape = "Region",
       color = "Region") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) # center title
```

Figure 8: Scatterplot of X1, Y and Region



Interpretation:

In the previous plot of personal income and expenditure on housing assistance per capita in a state (Figure 1), a moderate positive association between the two was identified. This plot highlighting the values from different regions adds a new layer of interpretation. Most notably, the South region shows on average both low income and low expenditure on housing assistace. The North Central region seems to also have mostly values for personal income that are proportional with the expenditure on housing assistance, the observations

being concetrated towards the centre of the plot. Nevertheless, the West region shows some contrast in terms of expenditure on housing assistance for similar personal income values. This suggests that factors at the regional level might be correlated with both the housing assistance expenditure and the personal income per capita, and the association between the two of them might require further consideration and assessment.