Problem Set 1

Applied Stats/Quant Methods 1

September 30, 2024

Name: Ombeline Mussat
Student Number:24346050

Question 1: Education

A school counselor was curious about the average 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. Confidence Interval

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

Our confidence coefficient is 0.90 (90%). We will compute the sample mean, sample standard deviation, and standard error before we can calculate the confidence interval.

- **Sample mean/point estimate:** The sample mean is calculated using the formula:

$$\text{Sample Mean} = \frac{\sum_{i=1}^{n} y_i}{n}$$

In R we do:

```r
sample_mean <- mean(y) # Point estimate
```

We have a sample mean of 98.44.

- **Sample standard deviation:** The sample standard deviation is calculated using:

$$\text{Sample sd} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \text{Sample Mean})^2}{n-1}}$$

In R we do:

```
sample_sd <- sd(y) # Sample standard deviation
```

We have a sample standard deviation of 13.0929.

- **Standard error:** The standard error is calculated as:

$$\text{Standard Error} = \frac{\text{Sample sd}}{\sqrt{n}}$$

In R we do:

```
standard_error <- sample_sd/sqrt(length(y)) # Standard error
```

We have a standard error of 2.6186.

Next, we calculate the 90% confidence interval using the t-distribution since $n < 30$:

- **Sample size:** The sample size is given by:

$$n = \text{length}(y)$$

In R we do:

```
n <- length(y) #sample size
```

We have the sample size n=25.

- **Degree of freedom:** The degrees of freedom is:

$$df = n - 1$$

In R we do:

```
df <- n - 1 #degree of freedom
```

We have a degree of freedom of df=24.

- **t-score:** The t-score is obtained from the t-distribution:

$$t_{90} = qt\left(\frac{1-0.90}{2}, df\right)$$

  Here, qt is used to find the critical value from the t-distribution because we are working with a small sample size (less than 30).

In R we do:

```
t90 <- qt((1 - 0.90) / 2, df = df, lower.tail = FALSE)
```

We have a t-score of t90=1.7109.

- **Confidence interval:** The lower and upper bounds are calculated as:

$$\text{Lower} = \text{Sample Mean} - t_{90} \times \text{Standard Error}$$

$$\text{Upper} = \text{Sample Mean} + t_{90} \times \text{Standard Error}$$

In R we do:

```
lower_90 <- sample_mean - (t90 * standard_error)
upper_90 <- sample_mean + (t90 * standard_error)
confint90 <- c(lower_90, upper_90)
```

We have a confidence interval of [93.96 : 102.92]. This means that if we took multiple samples, 90% of the time, this interval would contain the true population parameter.

2. Hypothesis Test

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.

We will conduct a hypothesis test to determine whether the average IQ of the students in the school is greater than 100.

- **Step 1: Assumptions**
  - We have a random sampling of our data
  - The data is continuous
  - The variable is distributed normally
  - The sample is below 30 so we will be using a t-test

- **Step 2: State Hypotheses** We have the following hypotheses:
  - Null hypothesis: $H_0 : \mu = 100$
  - Alternative hypothesis: $H_1 : \mu > 100$

  This is a one-sided test (right-tailed) because we want to test if the mean is greater than the average.

- **Step 3: Calculate the test statistic** We calculate the test statistic with the following formula:

$$t = \frac{\text{Sample Mean} - \mu_0}{\frac{\text{Sample SD}}{\sqrt{n}}}$$

In R we do:

```
t_statistic <- (sample_mean - mu_0) / (sample_sd / sqrt(n))
```

We have a t-statistic of -0.5957.

- **Step 4: P-value** We can calculate the p-value:

$$\text{p-value} = \Pr\left(T \geq \left|\frac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}}\right|\right)$$

In R we use the function pt to calculate the p-value, as we use a t-test.

```
p_value <- pt(t_statistic, df, lower.tail = FALSE)
```

We have a p-value of 0.7215.

- **Step 5: Draw conclusions** We can draw conclusions based on our results
  - Error probability:

$$\alpha = 0.05$$

The p-value is compared to α: The p-value is 0.7215 and $\alpha = 0.5$, so p-value $> \alpha$. Therefore, we fail to reject the null hypothesis.
We cannot conclude that the average student has an IQ 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`.

This is the `expenditure` data set that was imported:

```
STATE  Y   X1  X2  X3 Region
1    ME 61 1704 388 399      1
2    NH 68 1885 272 598      1
3    VT 72 1745 397 370      1
4    MA 72 2394 458 868      1
5    RI 62 1966 157 899      1
6    CT 91 2817 162 690      1
```

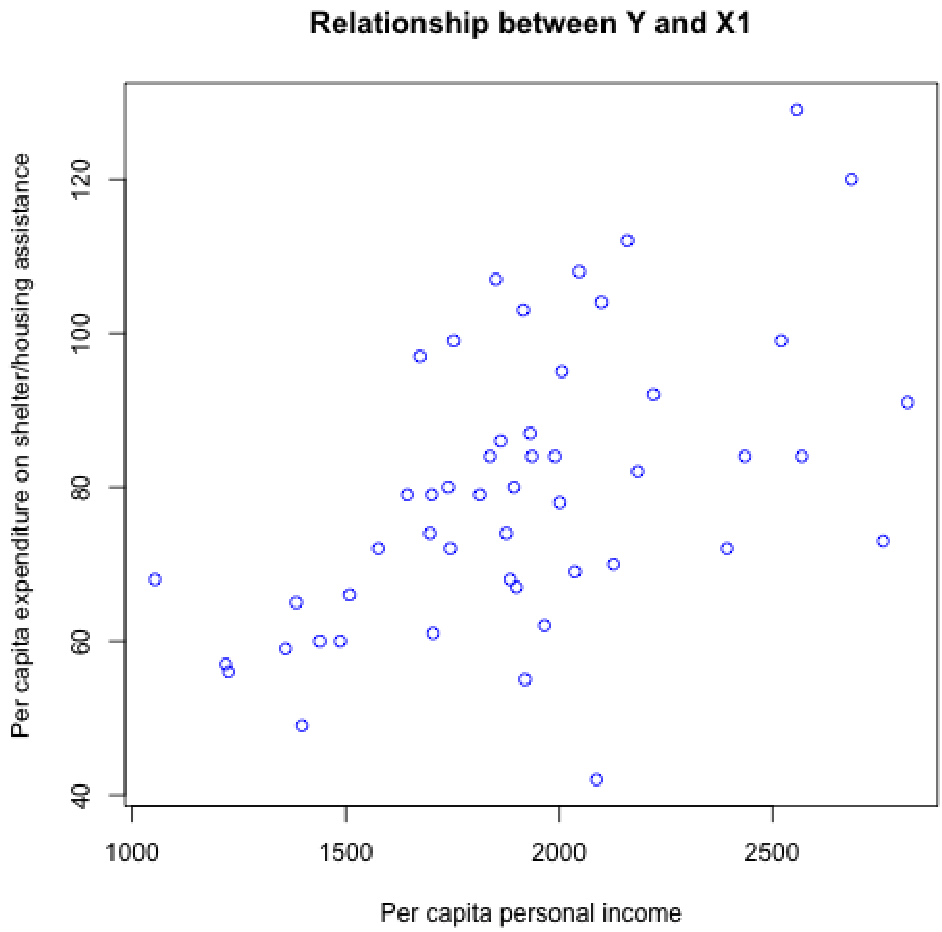
We can also look at the structure of the data, to see the length, the mean, median, minimum, maximum and 1st and 3rd quartile of each variable.

```
    STATE                     Y                 X1              X2
 Length:50          Min.   : 42.00   Min.   :1053   Min.   :111.0
 Class :character   1st Qu.: 67.25   1st Qu.:1698   1st Qu.:187.2
 Mode  :character   Median : 79.00   Median :1897   Median :241.5
 Mean   : 79.54   Mean   :1912   Mean   :281.8
 3rd Qu.: 90.00   3rd Qu.:2096   3rd Qu.:391.8
 Max.   :129.00   Max.   :2817   Max.   :531.0
 X3             Region
 Min.   :326.0   Min.   :1.00
 1st Qu.:426.2   1st Qu.:2.00
 Median :568.0   Median :3.00
 Mean   :561.7   Mean   :2.66
 3rd Qu.:661.2   3rd Qu.:3.75
 Max.   :899.0   Max.   :4.00
```

- 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)?

  I plotted the relationship between *Y* and *X1*:

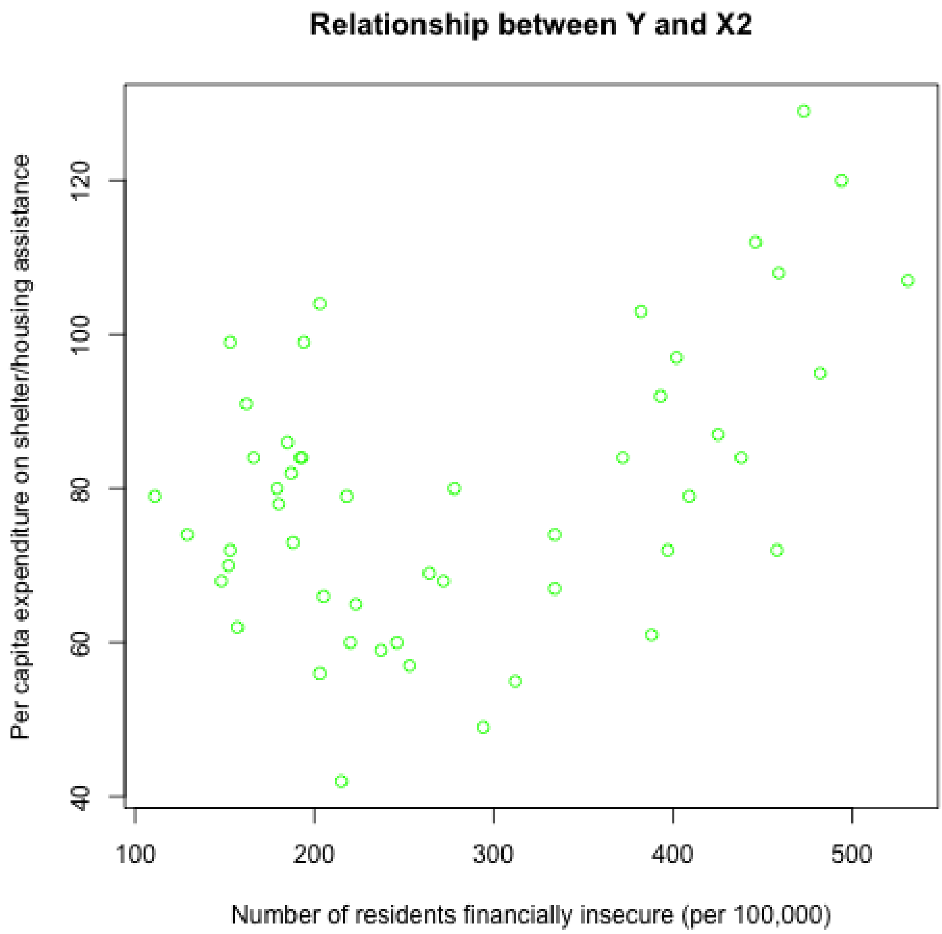
```r
plot(expenditure$X1, expenditure$Y,
     xlab = "Per capita personal income",
     ylab = "Per capita expenditure on shelter/housing assistance",
     main = "Relationship between Y and X1",
     col = "blue")
```



  This graph shows a positive linear relationship between per capita personal income and per capita expenditure on shelter/housing assistance. It suggests that as income increases, expenditure on shelter/housing assistance also tends to increase, and vice versa.

I plotted the relationship between Y and $X2$:

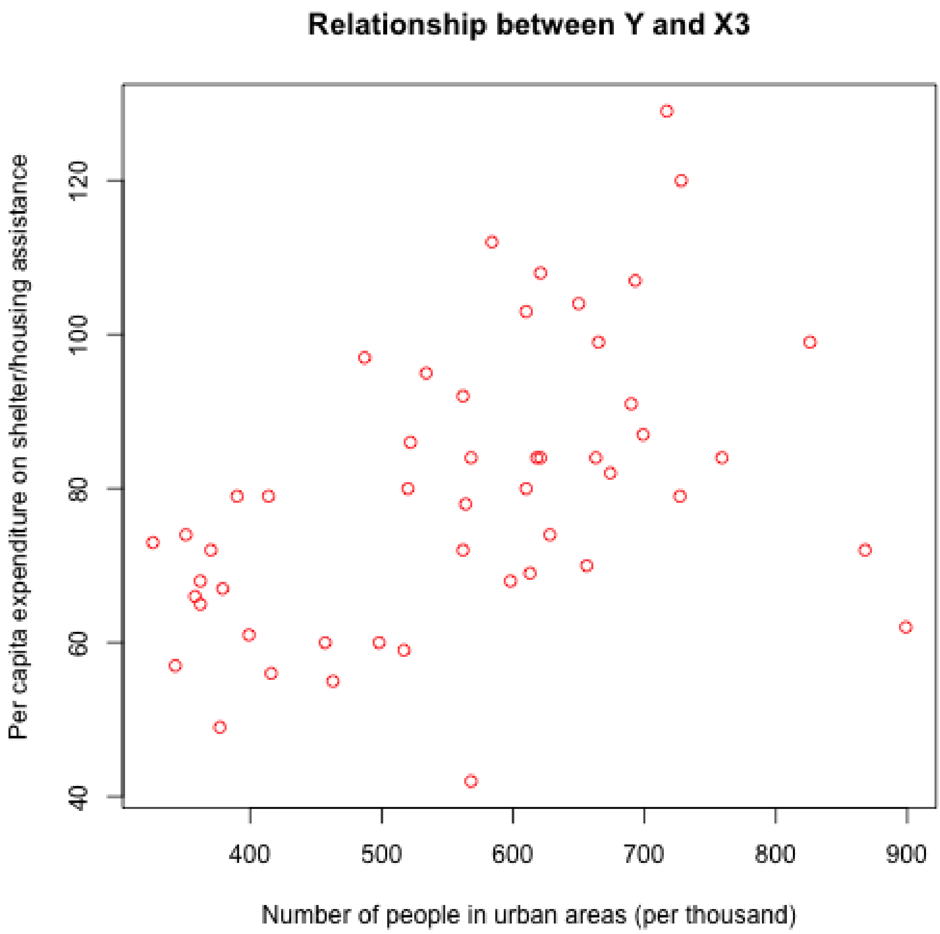
```
plot(expenditure$X2, expenditure$Y,
     xlab = "Number of residents financially insecure (per 100,000)",
     ylab = "Per capita expenditure on shelter/housing assistance",
     main = "Relationship between Y and X2",
     col = "green")
```



This graph shows a weak positive linear relationship between the number of residents financially insecure and the per capita expenditure on shelter/housing assistance. It suggests that as the number of residents financially insecure increases, the per capita expenditure on shelter/housing assistance also tends to increase slightly, and vice versa.

I plotted the relationship between Y and $X3$:

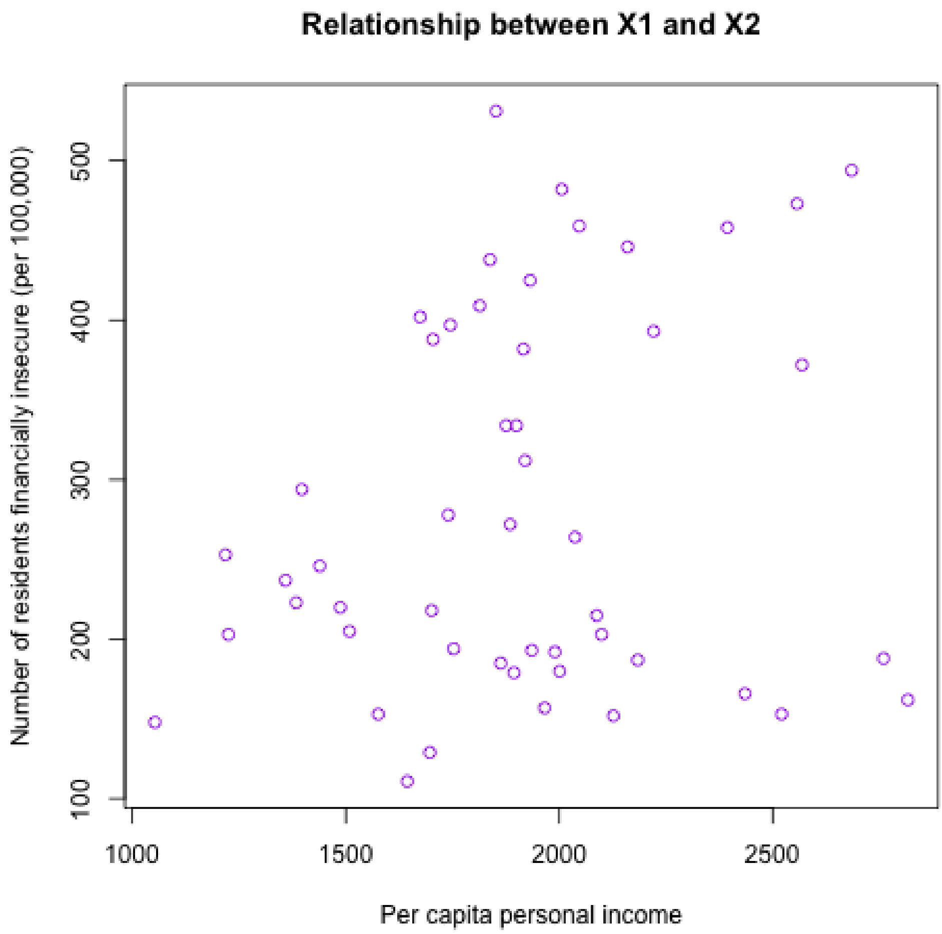
```
plot(expenditure$X3, expenditure$Y,
     xlab = "Number of people in urban areas (per thousand)",
     ylab = "Per capita expenditure on shelter/housing assistance",
     main = "Relationship between Y and X3",
     col = "red")
```



This graph shows a positive linear relationship between the number of people in urban areas and per capita expenditure on shelter/housing assistance. It suggests that as the number of people in urban areas increases, expenditure on shelter/housing assistance also tends to increase, and vice versa.

I plotted the relationship between *X1* and *X2*:

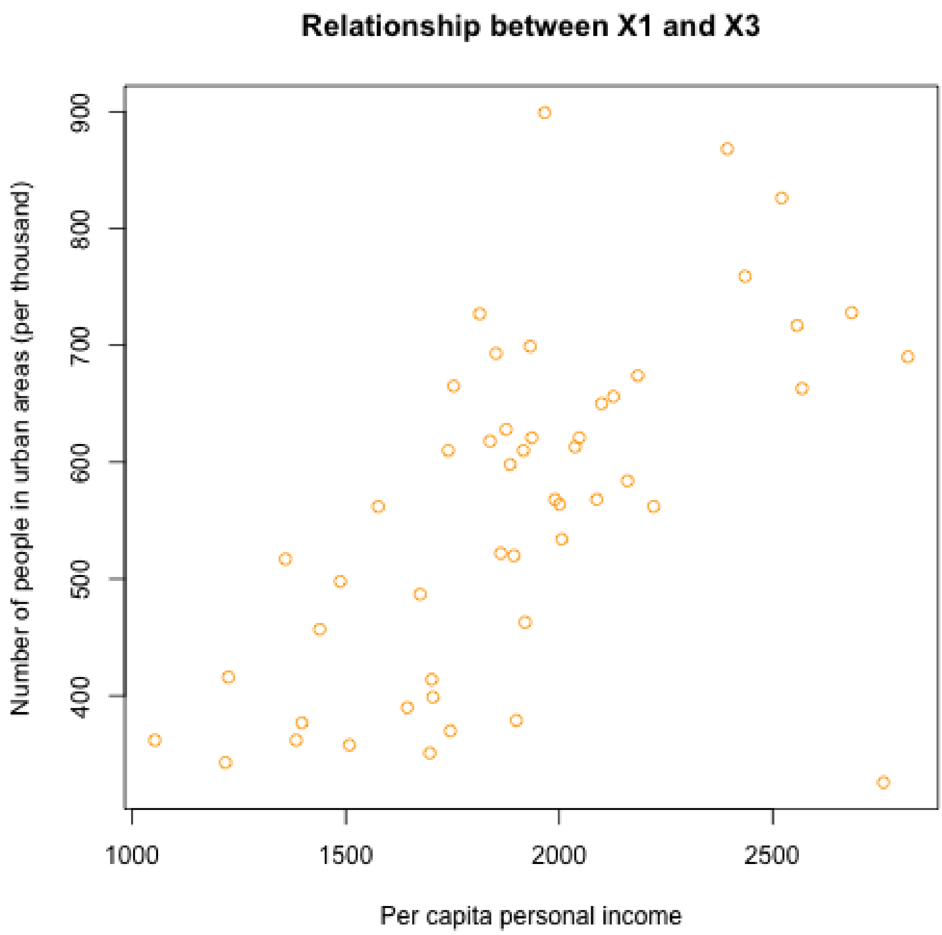
```
plot(expenditure$X1, expenditure$X2,
     xlab = "Per capita personal income",
     ylab = "Number of residents financially insecure (per 100,000)",
     main = "Relationship between X1 and X2",
     col = "purple")
```



This graph indicates a weak positive relationship between the income per capita and the number of residents financially insecure. There may be a tendency for the number of financially insecure residents to increase as well, and vice versa, but the relationship is weak.

I plotted the relationship between *X1* and *X3*:

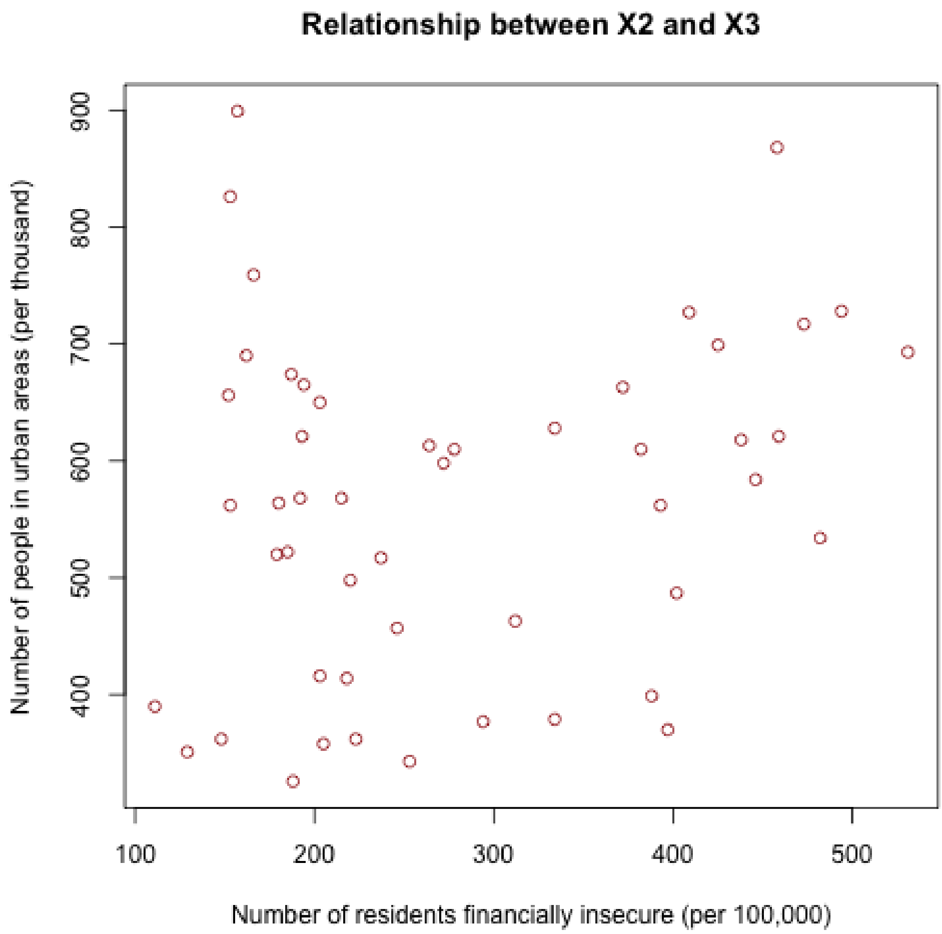
```
    xlab = "Per capita personal income",
    ylab = "Number of people in urban areas (per thousand)",
    main = "Relationship between X1 and X3",
    col = "orange")
```



This graph shows a strong positive linear relationship between the number of people in urban areas and income per capita. This suggests a strong correlation, indicating that as the population in urban areas increases, income per capita tends to increase as well, and vice versa.

I plotted the relationship between *X2* and *X3*:

```
plot(expenditure$X2, expenditure$X3,
     xlab = "Number of residents financially insecure (per 100,000)",
     ylab = "Number of people in urban areas (per thousand)",
     main = "Relationship between X2 and X3",
     col = "brown")
```

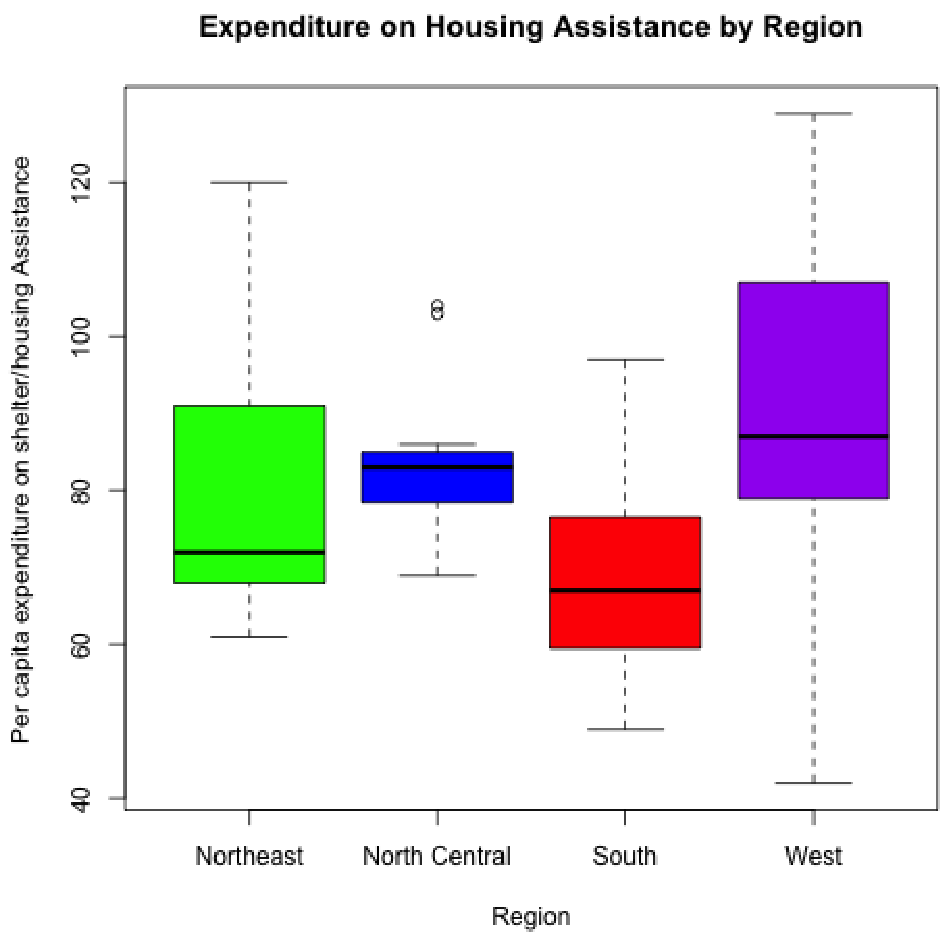


This graph shows that there is not particular relationship between the number of residents that are financially insecure and the number of people in urban areas. We cannot tell from this graph that these two variables are particularly correlated.

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

  I plotted the relationship between Y and *Region*.

```
boxplot(expenditure$Y ~ expenditure$Region,
        main = "Expenditure on Housing Assistance by Region",
        xlab = "Region",
        ylab = "Per capita expenditure on shelter/housing Assistance",
        names=c('Northeast','North Central', 'South', 'West'),
        col = c("green", "blue", "red", "purple"))
```

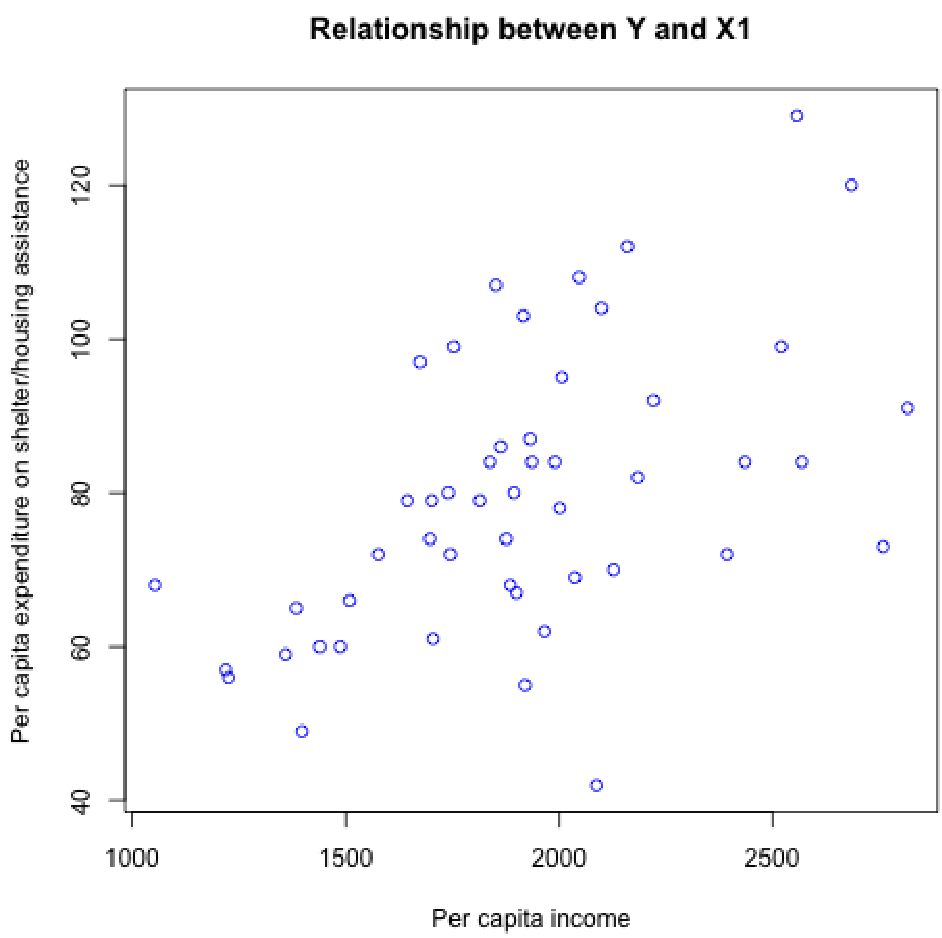


On average, the region with the highest per capita expenditure on housing assistance is the West. The box plot shows that the highest value of expenditure is in the West, with the highest 1st quartile, median, and 3rd quartile compared to the other regions.

- 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.

  I plotted the relationship between Y and $X1$:

```
plot(expenditure$X1, expenditure$Y,
     xlab = "Per capita income",
     ylab = "Per capita expenditure on shelter/housing assistance",
     main = "Relationship between Y and X1",
     col = "blue")
```



This graph shows a positive linear correlation between Y and $X1$, where an increase in per capita personal income corresponds with an increase in per capita expenditure on housing assistance and vice versa.

I reproduced the above graph including the variable *Region* which was displayed with different types of symbols and colors:

```
plot(expenditure$X1, expenditure$Y,
     xlab = "Per capita income",
     ylab = "Per capita expenditure on shelter/housing assistance",
     main = "Relationship between Y and X1 by Region",
     col = expenditure$Region,    # Display Regions in different colours
     pch = expenditure$Region)   # Display Regions in different symbols
#We need to add a legend
legend("topright",
       legend = c("Northeast", "North Central", "South", "West"),  # Region labels
       col = 1:4, #As we did not specify in the graph which colour to use, it will be the 1st to the 4th
       pch = 1:4) #As we did not specify in the graph which symbol to use, it will be the 1st to the 4th
```