Problem Set 1

Applied Stats/Quant Methods 1

Name: Darragh McGee (18319331)

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:

```
Student_IQ <- 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.

90 Percent Confidence Interval = Point Estimate i.e. Mean +/- Margin of Error (Critical Value*Standard Error)

Step 1: Input Data Set of Student IQs and Create Vector

```
Student_IQ <- 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)
```

Step 2: Calculating Mean Student Height

```
mean_IQ <- mean(Student_IQ)
print(mean_IQ)
```

Mean Student IQ = 98.44

Step 3: Calculating Standard Error

```
standard_error_IQ <- sd(Student_IQ)/sqrt(length(Student_IQ))
print(standard_error_IQ)
```

Standard Error = 2.618575

Step 4: Calculate Test Statistic for 90 Percent Confidence Level

As the sample size is 25 (Less than 30) a T-Distribution should be used for Critical Value.

T-Statistic Formula requires Degrees of Freedom Degrees of Freedom (n-1)

```
df <- 25 - 1
print(df)
```

Degrees of Freedom = 24

T-Statistics for 90 Percent Confidence Interval is interested in the Critical Value for the First 5 Percent and Last 5 Percent of the Distribution

```
t_score_lower <- qt(0.05, df) # Critical Value for First 5% (Lower Bound)
print(t_score_lower)
t_score_upper <- qt(0.95, df) # Critical Value for Last 5% (Upper Bound)
print(t_score_upper)
```

t-score = +/- 1.710882

Step 5: Construct 90 Percent Confidence Interval

```
lower_bound <- (mean_IQ+(t_score_lower)*(standard_error_IQ))
print(lower_bound)
upper_bound <- (mean_IQ+(t_score_upper)*(standard_error_IQ))
print(upper_bound)
```

90 Percent Confidence Interval = 93.95933 to 102.9201

Validating Approach: Using in-built t.test() formula in R to check results

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

```
90 percent confidence interval:
93.95993 102.92007
sample estimates:
mean of x
98.44
```

The results from the in-built t.test function are equal to the step-by-step calculations (allowing for marginal rounding differences).

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: Assumptions about Data

- Sample Size is below 30. Therefore, T-Statistic should be used.
- Population IQ = 100
- IQ represents continuous data
- Random Sampling Conducted
- The data is approximately normally distributed.
- Observations are independent of one another.

Step 2: Setting Up Hypothesis

- **Null Hypothesis:** The average Student IQ in the sample school is less than or equal to the average IQ score (100) among all the schools in the country
- **Alternative Hypothesis:** The average student IQ in the sample school is greater than the average IQ score (100) among all the schools in the country

Steps 3: Calculate the Test-Statistic

```
Population_Mean_IQ <- 100
t_statistic <- ((mean_IQ - Population_Mean_IQ) / standard_error_IQ)
print(t_statistic)
```

T-Statistic = -0.5957439

Steps 4: Calculate the p-value

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

p-value = 0.7215383

Step 5: Conclusion

- Fail to reject the Null Hypothesis as the p-value is greater than 0.05 (exceeding the 5 percent significance level).
- There is insufficient evidence to conclude that the sample mean is greater 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*

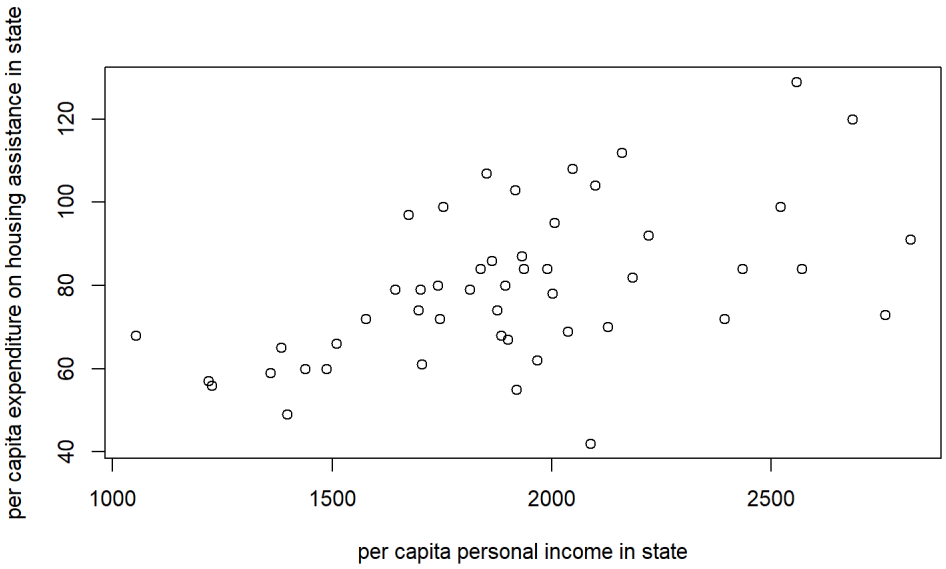
Explore the `expenditure` data set and import data into `R`.

```
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
```

- 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)?

```r
plot(expenditure$X1, expenditure$Y, col =1,
     ylab = "per capita expenditure on housing assistance in state",
     xlab="per capita personal income in state",
     main="")
```

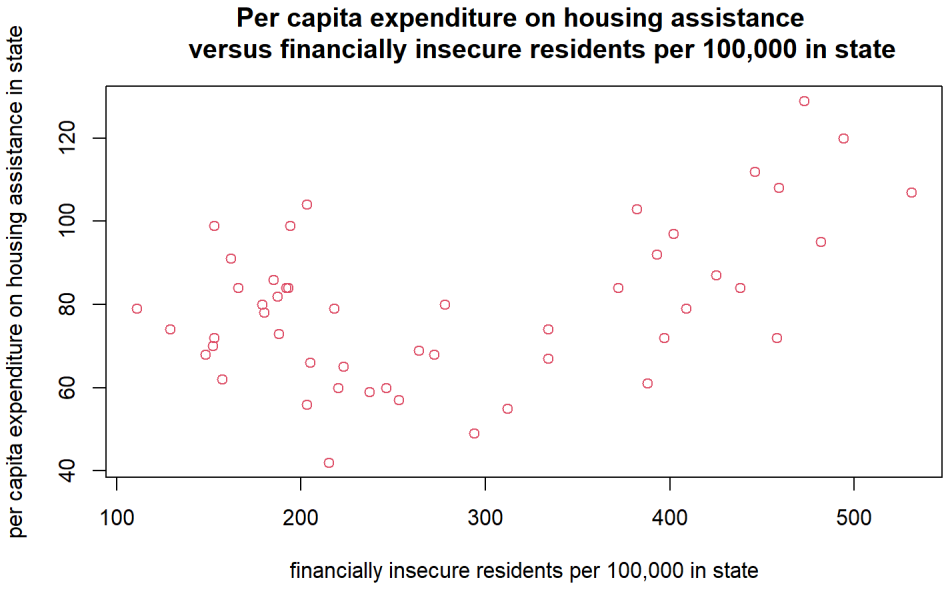
Figure 1: Scatterplot of relationship between per capita expenditure on housing assistance and personal income in state



- A moderate positive correlation is observable between per capita personal income in state and per capita expenditure on housing assistance in state.

```
plot(expenditure$X2, expenditure$Y, col =2,
     ylab = "per capita expenditure on housing assistance in state",
     xlab="financially insecure residents per 100,000 in state",
     main="Per capita expenditure on housing assistance
     versus financially insecure residents per 100,000 in state")
```

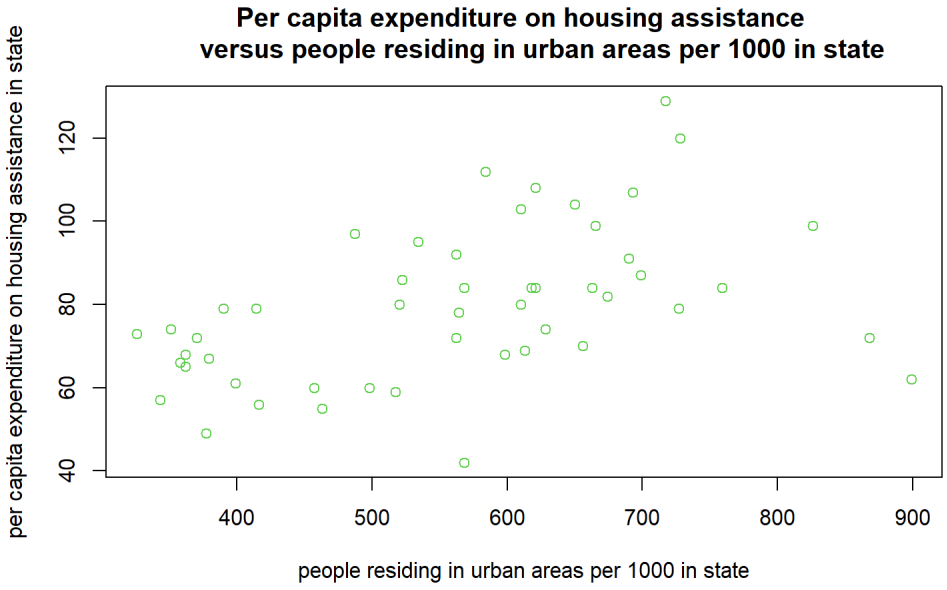
Figure 2: Scatterplot of relationship between per capita expenditure on housing assistance and financially insecure residents per 100,000 in state



- A moderate positive correlation is observable between per capita expenditure on housing assistance and financially insecure residents per 100,000 in the state.

```r
plot(expenditure$X3, expenditure$Y, col =3,
     ylab = "per capita expenditure on housing assistance in state",
     xlab="people residing in urban areas per 1000 in state",
     main="Per capita expenditure on housing assistance
     versus people residing in urban areas per 1000 in state")
```

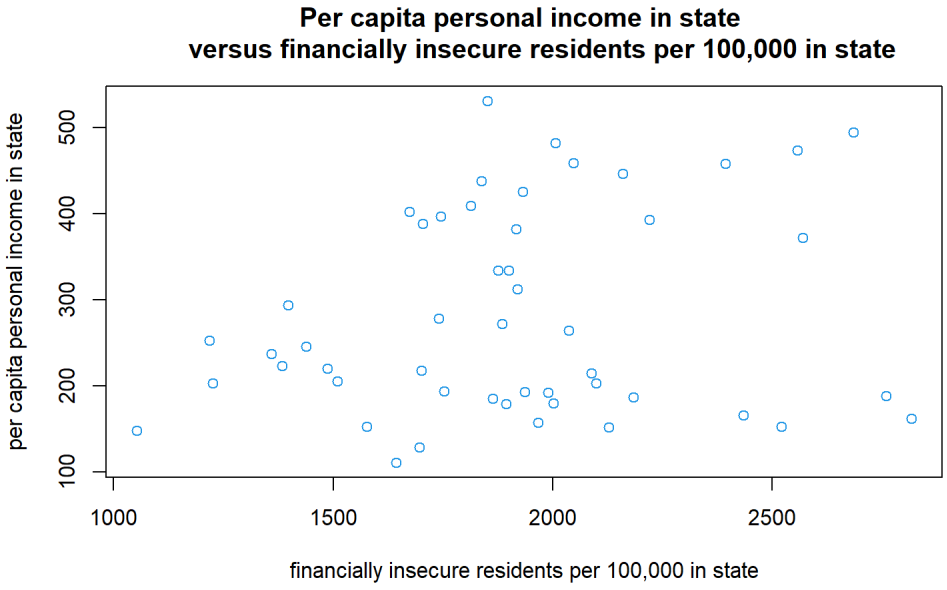
Figure 3: Scatterplot of relationship between per capita expenditure on housing assistance and people residing in urban areas per 1000 in state



- A moderate positive correlation is observable between people residing in urban areas per 1000 in state and per capita expenditure on housing legal assistance in state.

```r
plot(expenditure$X1, expenditure$X2, col =4,
     ylab = "per capita personal income in state",
     xlab="financially insecure residents per 100,000 in state",
     main="Per capita personal income in state
     versus financially insecure residents per 100,000 in state")
```

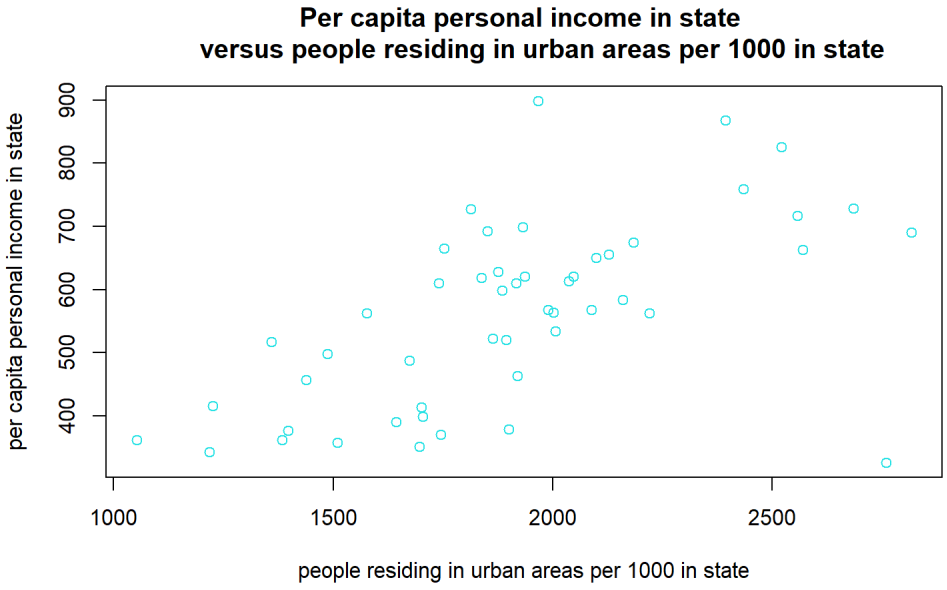
Figure 4: Scatterplot of relationship between per capita personal income in state and financially insecure residents per 100,000 in state



- There is no clear linear correlation observable between the per capita personal income in state and financially insecure residents per 100,000 in state.

```
png("Figure_1_5.png", width = 1500, height = 950, res = 200)
plot(expenditure$X1, expenditure$X3, col =5,
     ylab = "per capita personal income in state",
     xlab="people residing in urban areas per 1000 in state",
     main="Per capita personal income in state
```

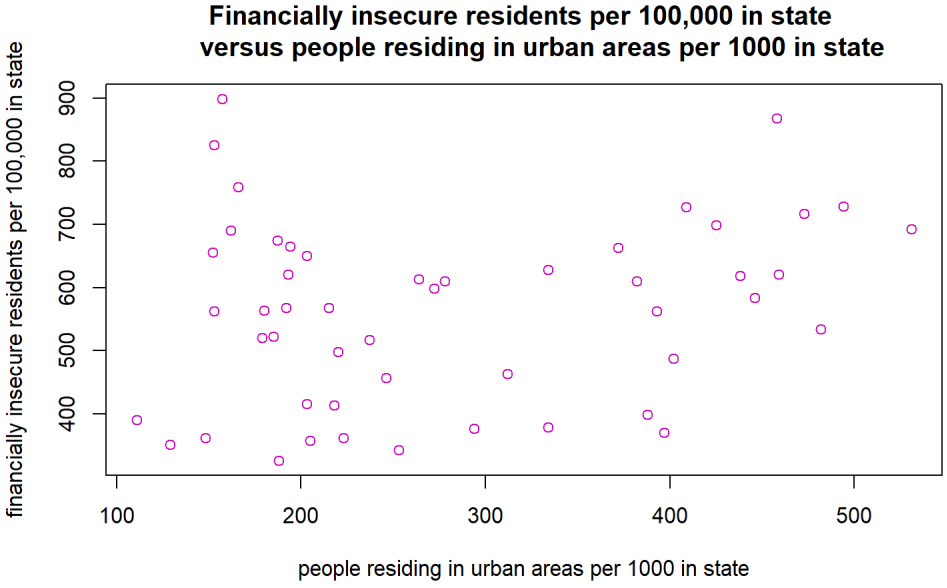
Figure 5: Scatterplot of relationship between per capita personal income in state and people residing in urban areas per 1000 in state



- A moderate positive correlation is observable between people residing in urban areas per 1000 in state and per capita personal income in state.

```
plot(expenditure$X2, expenditure$X3, col =6,
     ylab = "financially insecure residents per 100,000 in state",
     xlab="people residing in urban areas per 1000 in state",
     main="Financially insecure residents per 100,000 in state
     versus people residing in urban areas per 1000 in state")
```

Figure 6: Scatterplot of relationship between financially insecure residents per 100,000 in state and people residing in urban areas per 1000 in state

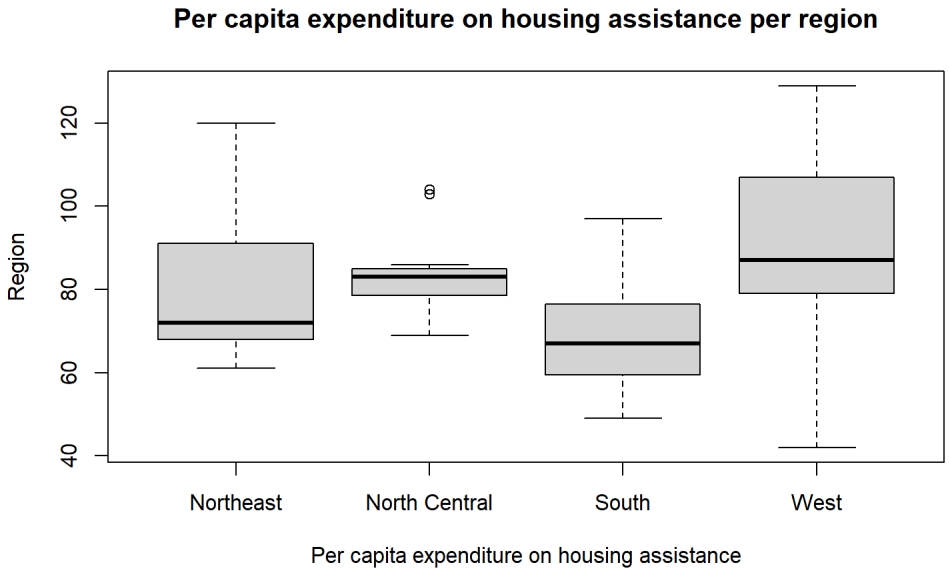


- There is no clear linear correlation observable between people residing in urban areas per 1000 in state and financially insecure residents per 100,000 in state.

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

```
boxplot(expenditure$Y ~ expenditure$Region,
        main="Per capita expenditure on housing assistance per region",
        ylab="Region",
        xlab="Per capita expenditure on housing assistance",
        names=c("Northeast", "North Central", "South", "West"))
```

Figure 7: Boxplot of per capita expenditure on housing assistance per region

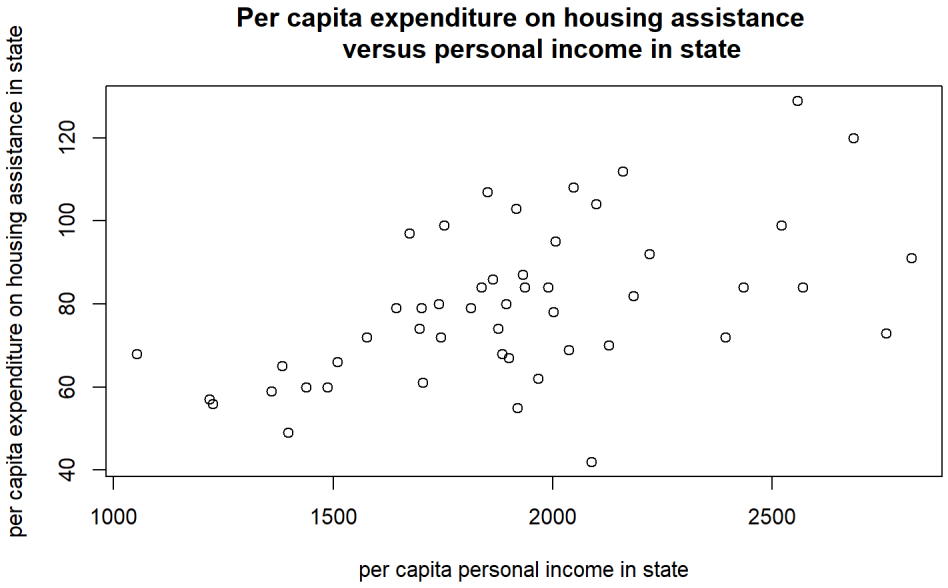


- The box plot visualises the median and interquartile range for per capita expenditure on housing assistance in state.
- The West Region has the highest median per capita expenditure on housing assistance. There is wide variation in the data for this region, as evidenced by the largest Interquartile Range across 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.

```
plot(expenditure$X1, expenditure$Y, col =1,
     ylab = "per capita expenditure on housing assistance in state",
     xlab="per capita personal income in state",
     main="Per capita expenditure on housing assistance
     versus personal income in state")
```

Figure 8: Scatterplot of relationship between Per Capita Personal Income and Housing Assistance Expenditure



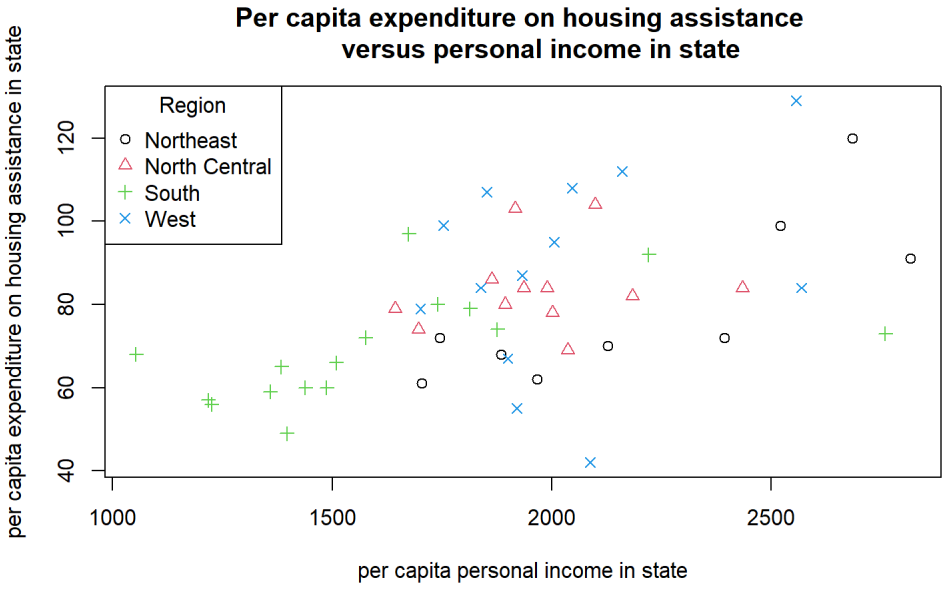
- A moderate positive correlation is observable between per capita personal income in state and per capita expenditure on housing assistance in state.
- The data points are relatively spread out, indicating moderate variability in expenditure at similar income levels.
- There are some outliers in the data, particularly at higher end of per capita personal income in state.

```r
# Creating vector of colours for identification
colours <- c("Northeast" = 1, "North Central" = 2, "South" = 3, "West" = 4)
icons <- c("Northeast" = 1, "North Central" = 2, "South" = 3, "West" = 4)

plot(expenditure$X1, expenditure$Y, col = colours[expenditure$Region],
     pch = icons[expenditure$Region],
     ylab = "per capita expenditure on housing assistance in state",
     xlab = "per capita personal income in state",
     main = "Per capita expenditure on housing assistance
     versus personal income in state")

# Adding in a Legend to Identify which Icons and colours relate to each Region
legend("topleft", legend =c("Northeast", "North Central", "South", "West"),
       col = c(1, 2, 3, 4), pch = c(1, 2, 3, 4), title = "Region")
```

Figure 9: Scatterplot of relationship between Per Capita Personal Income and Housing Assistance Expenditure by Region



- Overall, there is a general upward trend, indicating that states with higher per capita income tend to spend more on housing assistance per capita.
- The Northeast and West Regions show that higher per capita income states spend more per capita on housing assistance.
- North Central states show moderate expenditures on housing assistance at moderate income levels, with some variability.
- Southern States tend to have lower per capita income in state and lower per capita expenditure on housing assistance in state.