Problem Set 1

Liu Yuanyuan

Due: October 1, 2023

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.

```
IQ_scores <- 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)
t_test_result <- t.test(IQ_scores,conf.level = 0.90)

min <- t_test_result$conf.int[1]
max <- t_test_result$conf.int[2]
cat("1. 90% Confidence Interval for School Students' Average IQ:\n")
cat(min,max)
```

- **Result**:
  90% Confidence Interval for School Students' Average IQ:
  93.95993 102.9201

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$\alpha = 0.05$.

```
t_test_result <- t.test(IQ_scores,mu=100,alternative="greater")
t_test_result
cat("\nHypothes Test Results:\n")
cat("Test Statistic (t):", t_test_result$statistic , "\n")
cat("Degrees of Freedom:", t_test_result$parameter, "\n")
```

```r
cat("p-value:", t_test_result$p.value, "\n")

if (t_test_result$p.value < 0.05) {
  cat("we reject the null hypothesis. School students' average IQ is higher than the national average.\n")
} else {
  cat("we fail to reject the null hypothesis. School students' average IQ is equal to the national average.\n")
}
```

- **Result:**
  Hypothesis Test Results:
  Test Statistic (t): -0.5957439
  p-value: 0.7215383
  we fail to reject the null hypothesis. School students' average IQ is equal to the national average.

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*

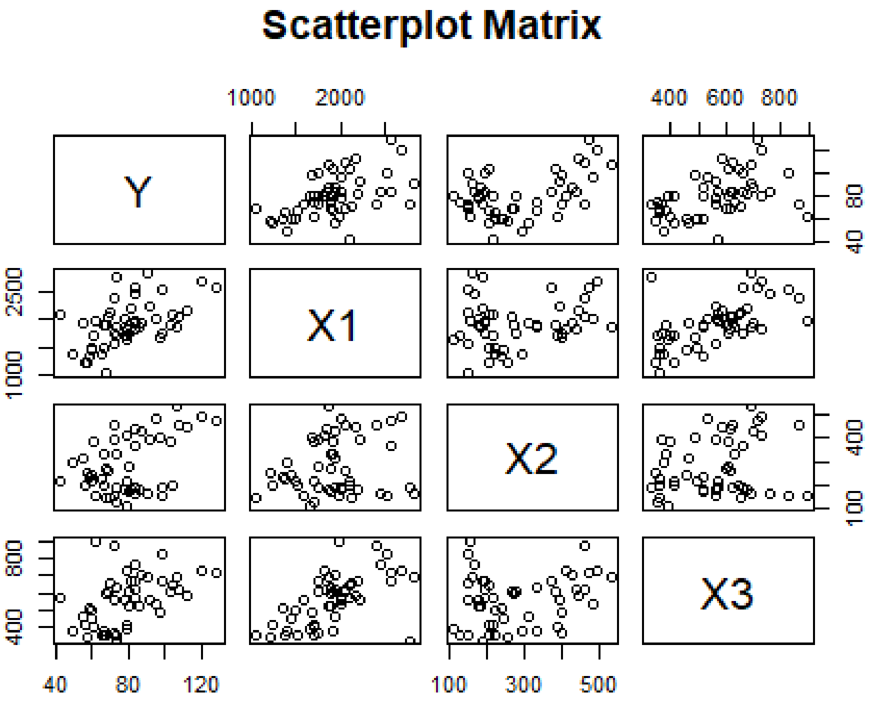
Explore the `expenditure` data set and import data into `R`.

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

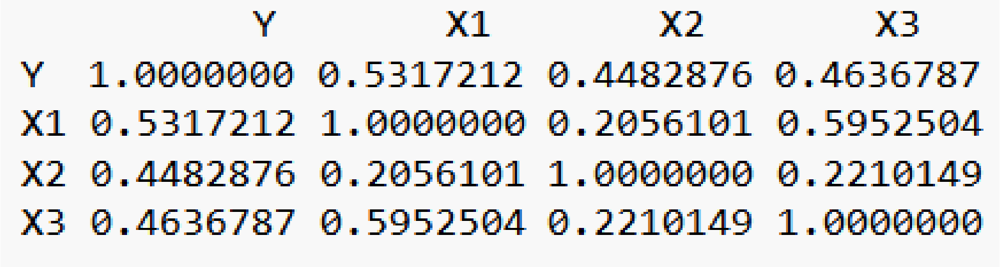
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?

```r
pairs(data_expenditure[c("Y", "X1", "X2", "X3")], main="Scatterplot Matrix")

correlations <- cor(data_expenditure[c("Y", "X1", "X2", "X3")])
print(correlations)
```



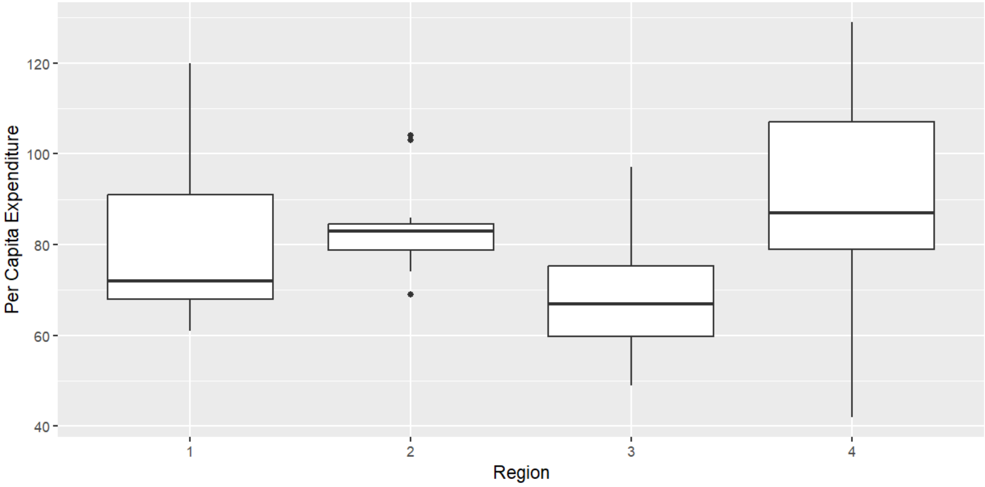
- **Result:**
- relationship between Y and X1 : Most of the data points are located near a trend line in the scatterplot matrix, indicating that the correlation between Y and X1 may be strong. In the scatterplot matrix, there appears to be some kind of positive linear relationship between Y and X1, with the data points showing a roughly positive correlation trend.
- Relationship between Y and X2 : Most of the data points also lie near a trend line in the scatterplot matrix, suggesting that the correlation between Y and X2 may be strong. There appears to be some sort of negative linear relationship between Y and X2 in the scatterplot matrix, with the data points showing a roughly negative correlation trend.
- Relationship between Y and X3 : The data points are widely distributed in the graph with no apparent linear relationship, possibly indicating a weak correlation between Y and X3. The relationship between Y and X3 in the scatterplot matrix looks much more scattered, with no clear linear trend.



```
           Y        X1        X2        X3
Y  1.0000000 0.5317212 0.4482876 0.4636787
X1 0.5317212 1.0000000 0.2056101 0.5952504
X2 0.4482876 0.2056101 1.0000000 0.2210149
X3 0.4636787 0.5952504 0.2210149 1.0000000
```

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

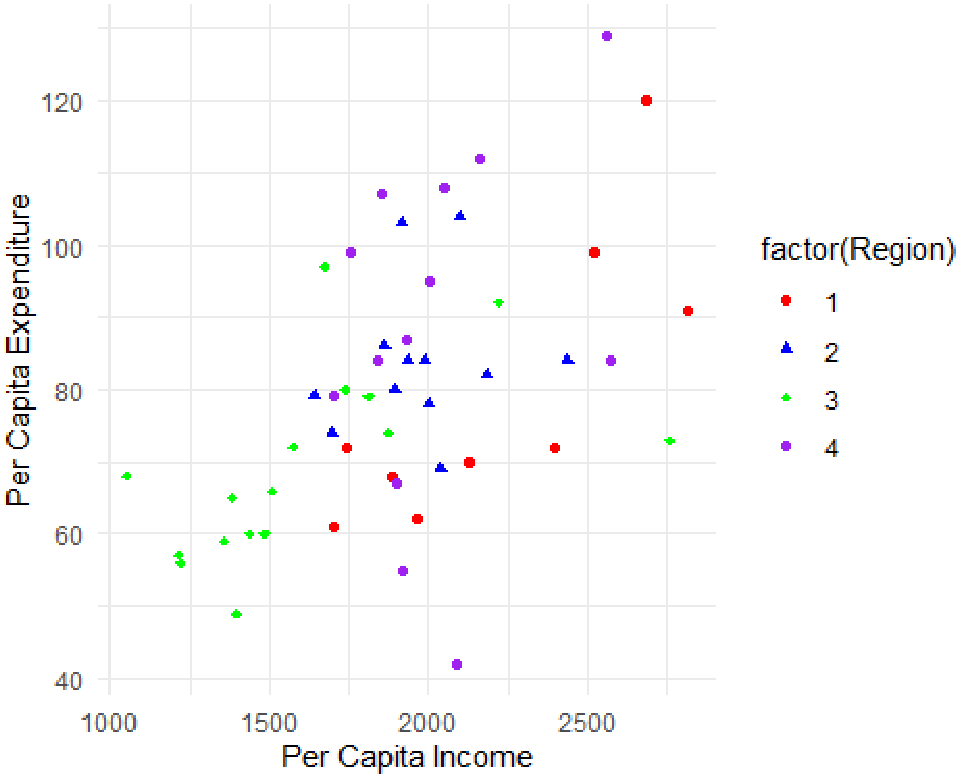
```
ggplot(data_expenditure, aes(x = factor(Region), y = Y)) +
  geom_boxplot() +
  labs(x = "Region", y = "Per Capita Expenditure") +
  ggtitle("relationship:Region VS Y")
```



  - **Result:**
  - On average, region 4 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.

```
ggplot(data_expenditure, aes(x = X1, y = Y, color = factor(Region), shape
    = factor(Region))) +
  geom_point(size=3) +
  labs(x = "Per Capita Income", y = "Per Capita Expenditure") +
  ggtitle(”relationship:X1 VS Y”)
scale_color_manual(values = c(”1” = ”red”, ”2” = ”blue”, ”3” = ”green”, ”4” = ”purple”)) +
  scale_shape_manual(values = c(”1” = 16, ”2” = 17, ”3” = 18, ”4” = 19)) +
  theme_minimal()
```



- The x-axis represents the per capita personal income in the state, while the y-axis represents the per capita expenditure on shelters/housing assistance in the state. Different colors and shapes are used to distinguish regions (1 = Northeast, 2 = North Central, 3 = South, 4 = West).
  - **Result:**
  - Focusing on Region 3 (South), characterized by its comparatively lower per capita income, it is evident from the chart that a positive correlation exists between per capita personal income (X1) and per capita housing/housing assistance expenditures (Y). Specifically, higher per capita personal income (X1) is associated with higher per capita housing/housing assistance expenditures (Y). Notably, the South region exhibits the lowest per capita personal income (X1) and the lowest per capita housing/housing assistance expenditures (Y).