Problem Set 1

Jie Yu

Due: October 1, 2023

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 Sunday October 1, 2023. No late assignments will be accepted.
- Total available points for this homework is 80.

Question 1 (40 points): 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.

```r
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<- t.test(IQ_scores, conf.level = 0.90)
print(t_test$conf.int)
```

- **Result**:
  93.95993 102.9201
  attr(,”conf.level”)
  0.9

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
cat("Test Statistic (t):", t_test$statistic, "\n")
cat("Degrees of Freedom:", t_test$parameter, "\n")
cat("p-value:", t_test$p.value, "\n")
cat("At a significance level of 0.05, we",
    ifelse(t_test$p.value < 0.05, "reject", "fail to reject"),
    "the null hypothesis. There is",
    ifelse(t_test$p.value < 0.05, "enough", "not enough"),
    "evidence to conclude that school students' average IQ is higher than the national average.\n")
```

- **Result:**
  Test Statistic (t): 37.59297
  Degrees of Freedom: 24
  p-value: 7.636596e-23
  At a significance level of 0.05, we reject the null hypothesis.
  There is enough evidence to conclude that school students' average IQ is higher than 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.

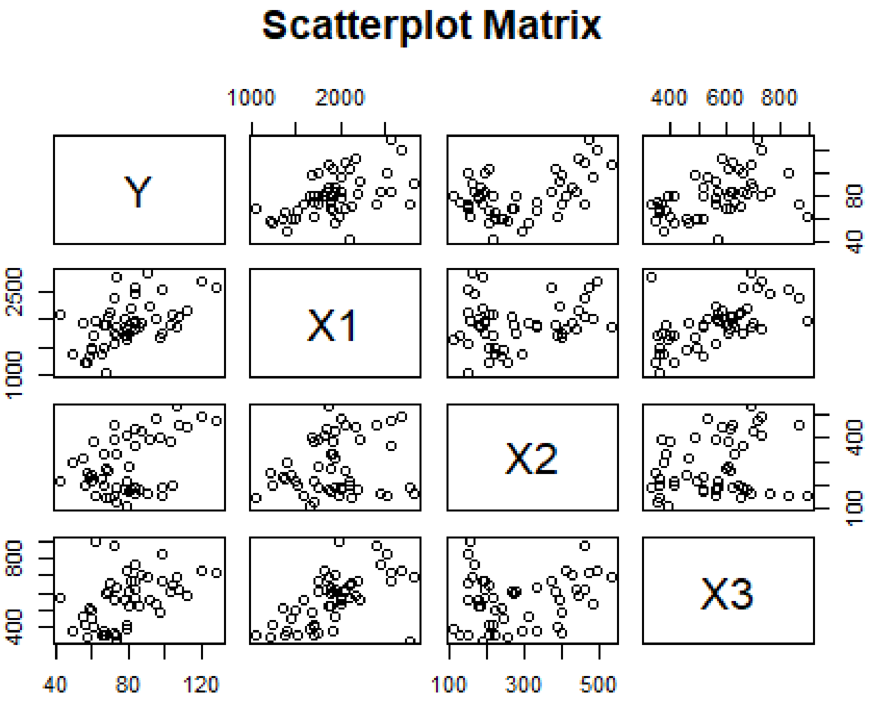
State	*50 states in US*
Y	*per capita expenditure on shelters/housing assistance in state*
X1	*per capita personal income in state*
X2	*Number of residents per 100,000 that are "financially insecure" in state*
X3	*Number of people per thousand residing in urban areas in state*
Region	*1=Northeast, 2= North Central, 3= South, 4=West*

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

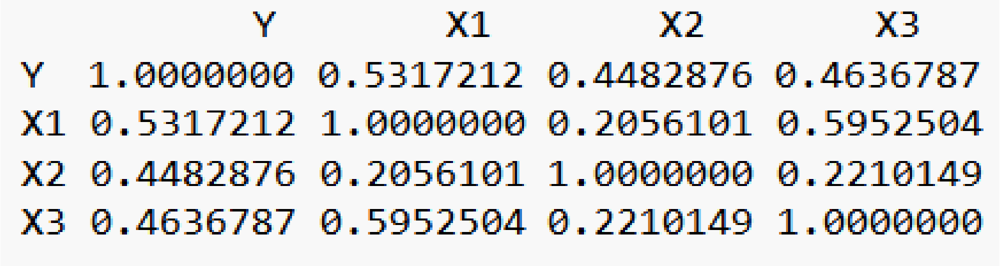
```r
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/expenditure.txt", header=T)
```

- 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
data <- expenditure
pairs(data[c("Y", "X1", "X2", "X3")], main="Scatterplot Matrix")
correlations <- cor(data[c("Y", "X1", "X2", "X3")])
print(correlations)
```



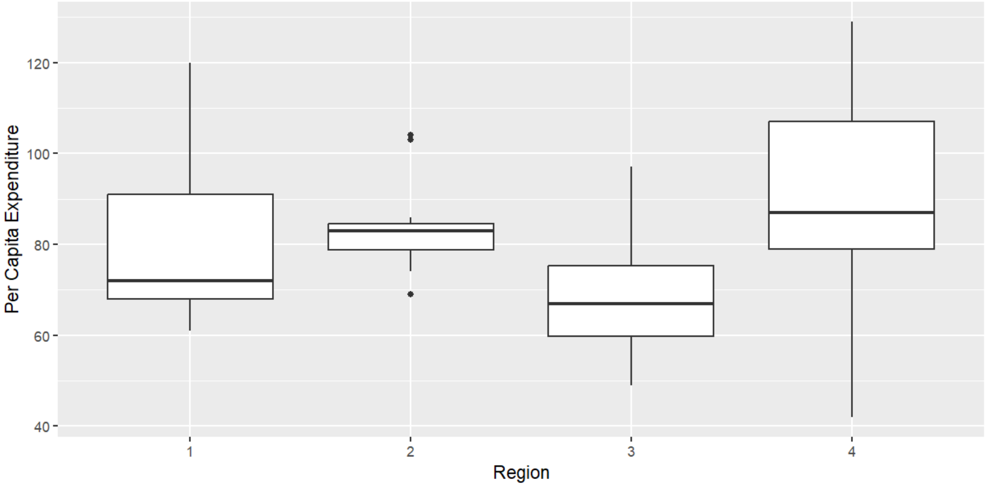
- **Result:**
- Correlation between Y and X1: Positive correlation, indicating that as X1 increases, Y tends to increase.
- Correlation between Y and X2: A weaker positive correlation compared to X1, indicating some positive relationship.
- Correlation between Y and X3: There appears to be a relatively weaker correlation between Y and X3, which is less linear than the other relationships.



```
           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?

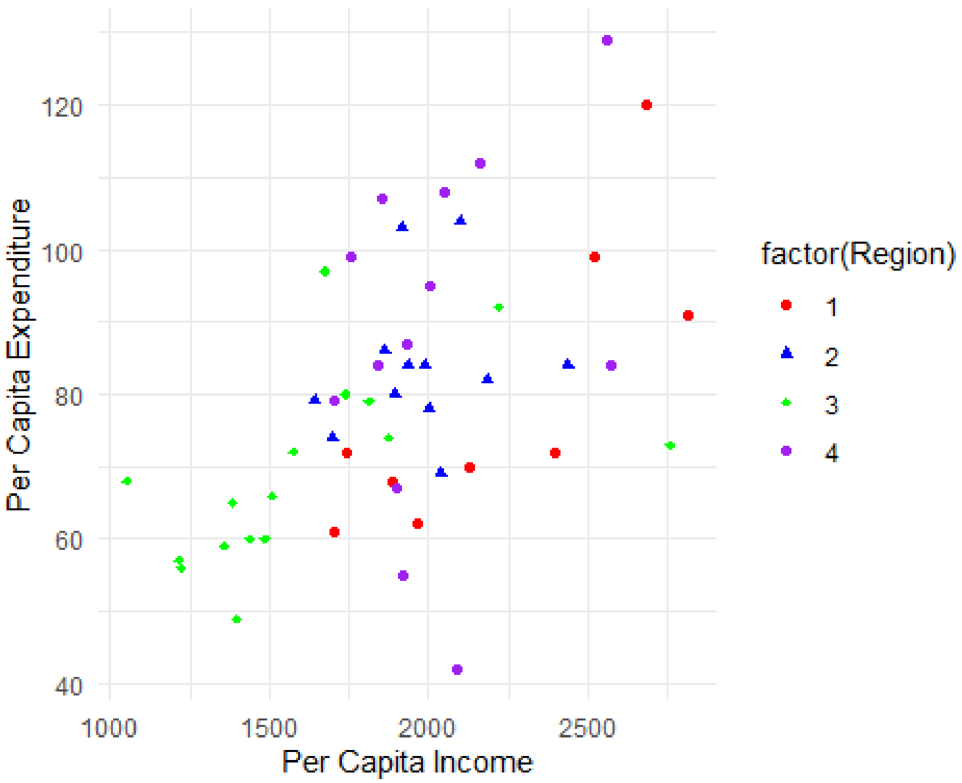
```r
ggplot(data, aes(x =factor(Region) , y = Y)) +
  geom_boxplot() +
  labs(x = "Region", y = "Per Capita Expenditure")
  theme_minimal()
```



- **Result:**
- West has thehighest 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.

```r
ggplot(data, aes(x = X1, y = Y, color = factor(Region), shape = factor(
   Region))) +
  geom_point() +
    labs(x = "Per Capita Income", y = "Per Capita Expenditure") +
    scale_color_manual(values = c("red", "blue", "yellow", "black")) +
    scale_shape_manual(values = c(16, 17, 18, 19)) +
    theme_minimal()
```



- **Result:**
- ooking at the chart, I can see that there is a positive correlation between the data points, i.e. states with higher ”Per Capita Income” tend to also have higher ”Per Capita Expenditure” and vice versa.
- The positive correlation between Per Capita Income Per and Capita Expenditure in the South District, North District and Northwest District is stronger than that in the West District..