Problem Set 1

Applied Stats/Quant Methods 1

Due: September 30, 2024

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 Monday September 30, 2024. 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.

```r
sample_mean <- mean(y)
sample_sd <- sd(y)
n <- length(y)
t_score <- qt(0.95, df = n-1)
lower_CI <- sample_mean - (t_score * sample_sd / sqrt(n))
upper_CI <- sample_mean + (t_score * sample_sd / sqrt(n))
confint90 <- c(lower_CI, upper_CI)
confint90
```

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
#H0: Average IQ of students in schools <= 100  H1: Average IQ of students in schools > 100
t_test <- t.test(y, mu = 100, alternative = "greater")
print(t_test)
#t = -0.59574, df = 24, p-value = 0.7215
#alternative hypothesis: true mean is greater than 100
#p>0.05,so the null hypothesis cannot be rejected.
#there is insufficient statistical evidence to support that the average IQ of the school's students is significantly higher than the national average IQ.
```

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)?

```
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
pdf("plot.relationships_YuFan.pdf")
pairs(expenditure[, c("Y", "X1", "X2", "X3")], main = "Scatterplot Matrix")
dev.off()

matrix <- cor(expenditure[, c("Y", "X1", "X2", "X3")])
print(matrix)
summary(expenditure)
sink("summary.txt")
print(summary(expenditure) )
sink()
```

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

```
# Create correlation matrix
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
cor_matrix <- cor(expenditure[, c("Y", "X1", "X2", "X3")])
print(cor_matrix)
```

```
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
```

The correlation coefficient between Y and X1 is 0.5317212, indicating a moderate positive correlation.

The correlation coefficient between Y and X2 is 0.4482876, indicating a slight to moderate positive correlation.

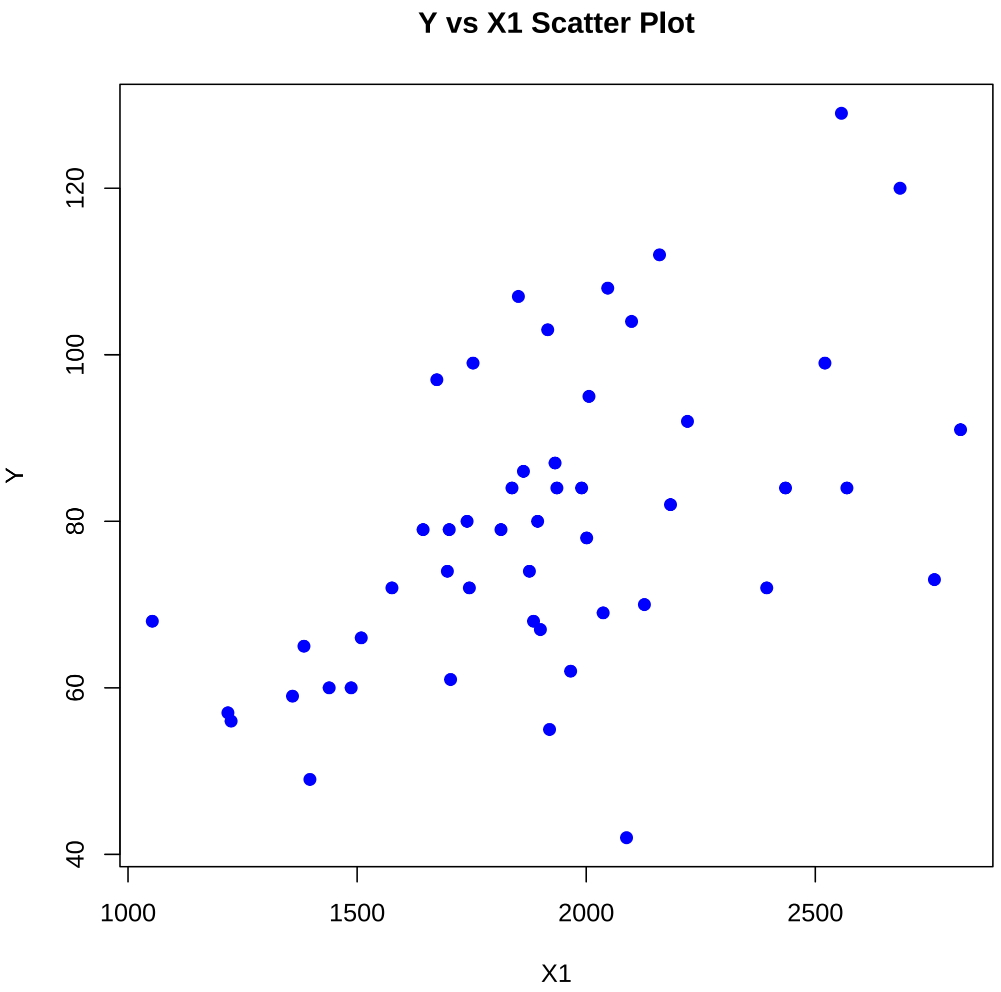
The correlation coefficient between Y and X3 is 0.4636787, indicating that they have a slightly higher positive correlation than Y and X2.

The correlation coefficient between X1 and X2 is 0.2056101, indicating a weak positive correlation.

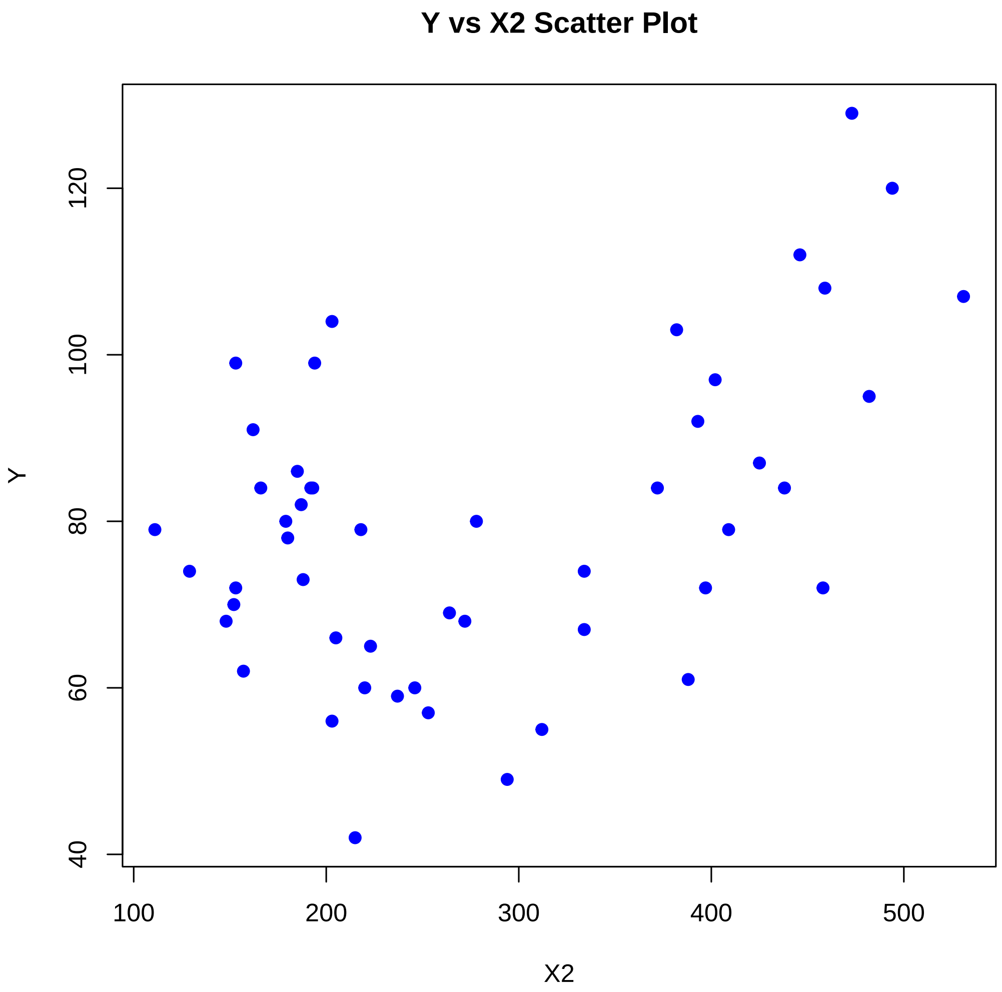
The correlation coefficient between X1 and X3 is 0.5952504, indicating a moderate positive correlation.

The correlation coefficient between X2 and X3 is 0.2210149, indicating a weak positive correlation.

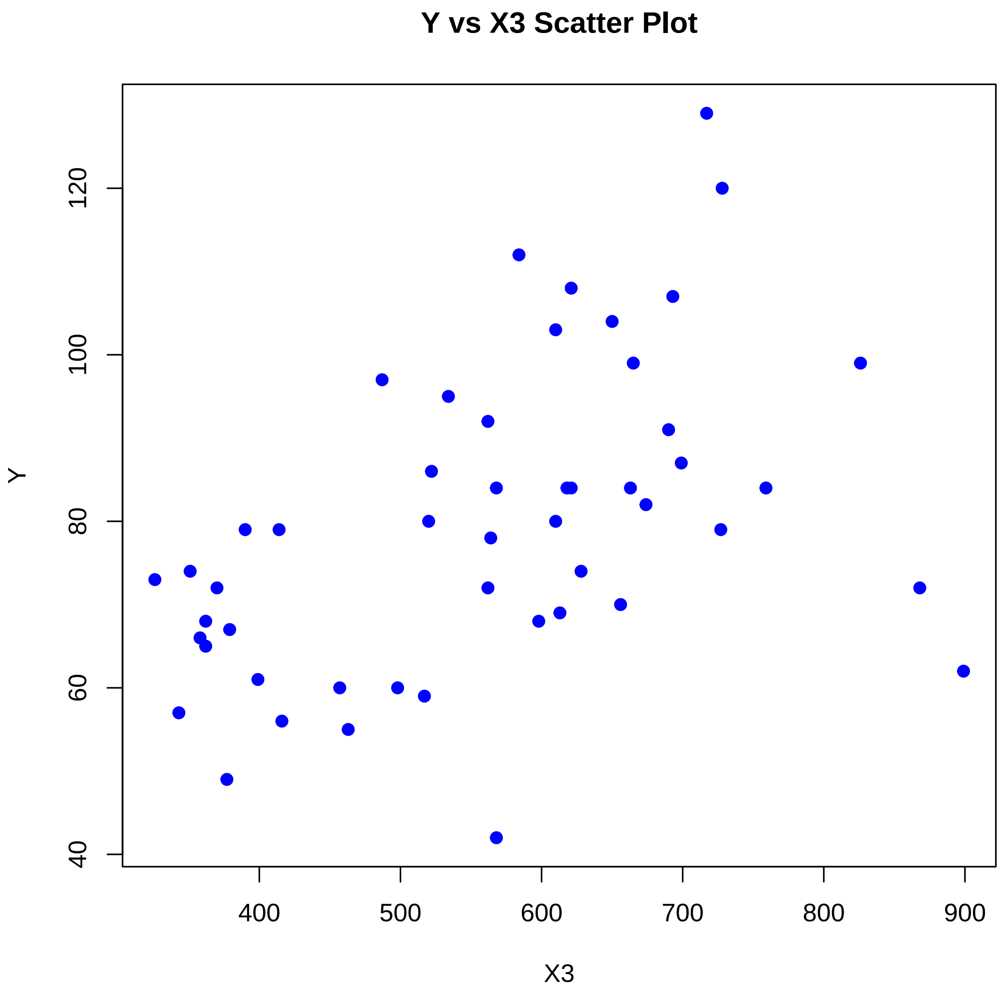
```r
#Y and X1
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
pdf("plot.Y.X1_YuFan.pdf")
plot(x = expenditure$X1, y = expenditure$Y, main = "Y vs X1 Scatter Plot", xlab = "X1", ylab = "Y", pch = 19, col = "blue")
dev.off()
```



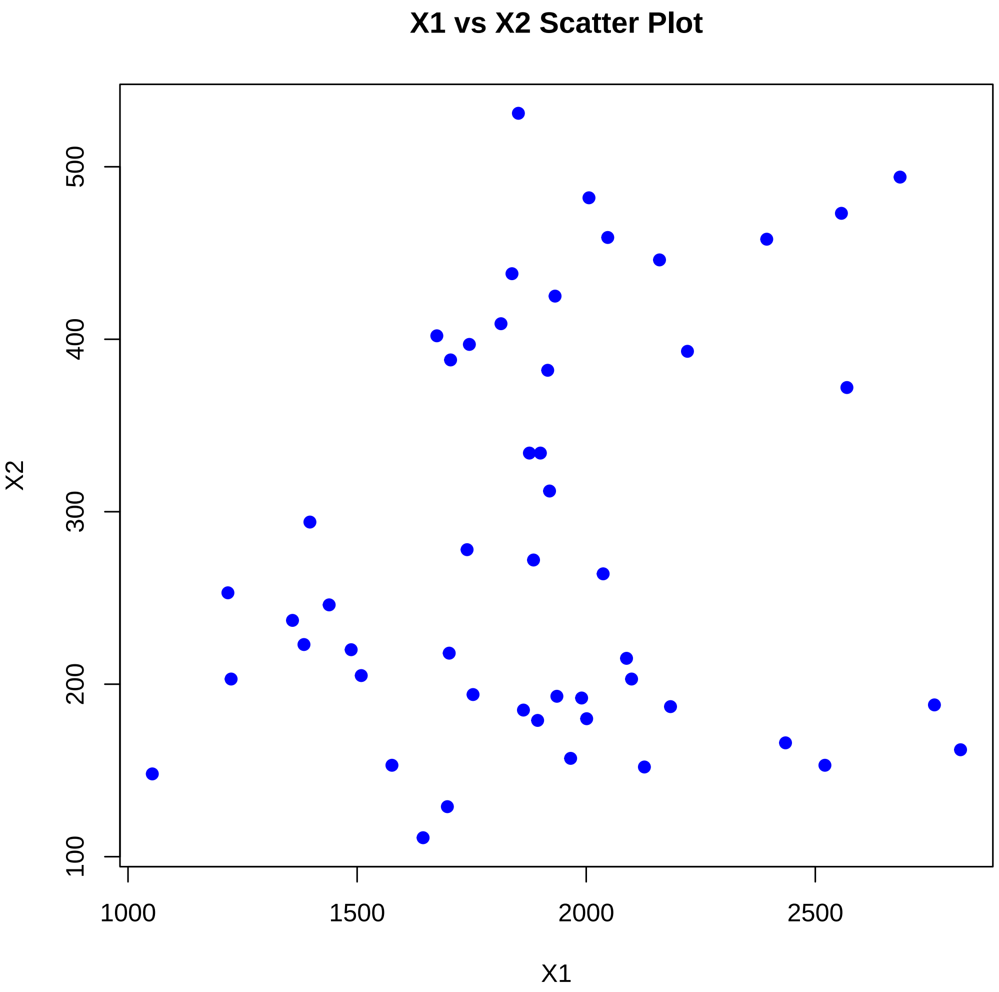
```
#Y and X2
expenditure <− read.table(”https://raw.githubusercontent.com/ASDS−TCD/StatsI_Fall2024/main/datasets/expenditure.txt”, header=T)
pdf(”plot.Y.X2_YuFan.pdf”)
plot(x = expenditure$X2, y = expenditure$Y, main = ”Y vs X2 Scatter Plot”, xlab = ”X2”, ylab = ”Y”, pch = 19, col = ”blue”)
dev.off()
```



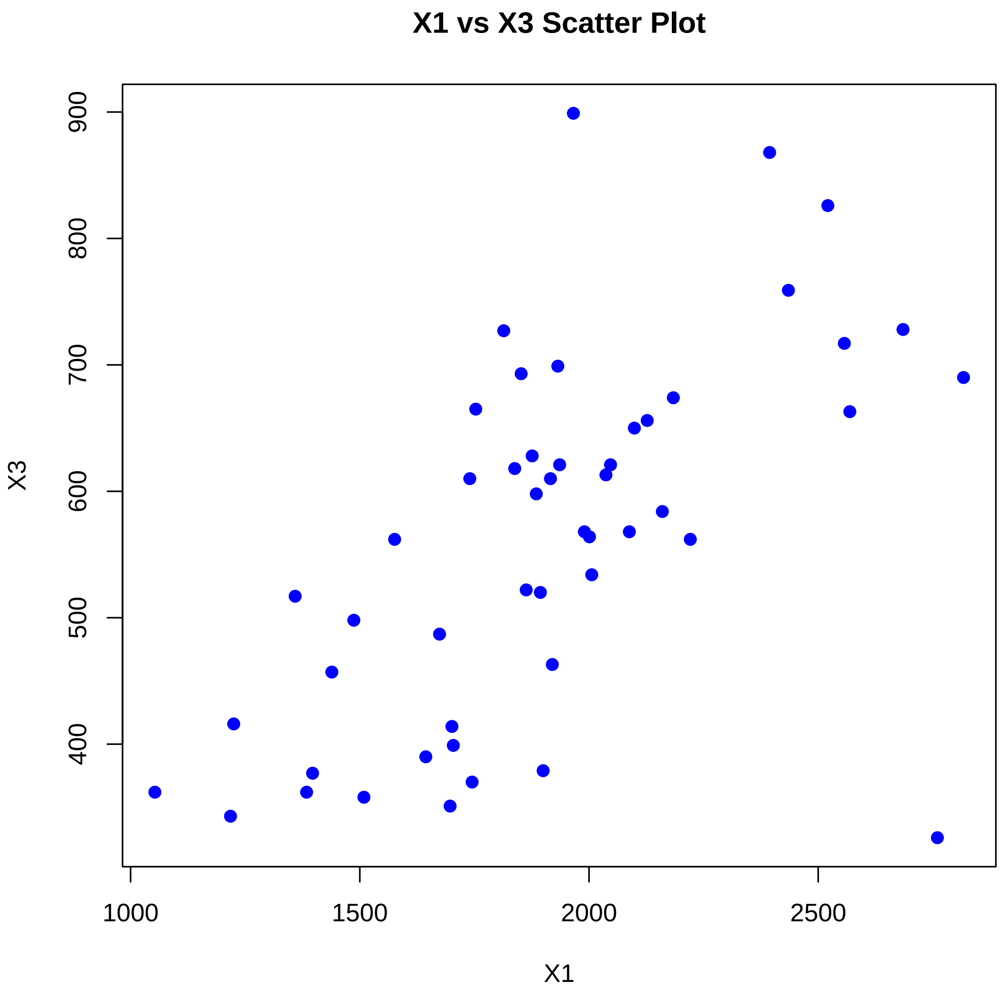
```
#Y and X3
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
pdf("plot.Y.X3_YuFan.pdf")
plot(x = expenditure$X3, y = expenditure$Y, main = "Y vs X3 Scatter Plot", xlab = "X3", ylab = "Y", pch = 19, col = "blue")
dev.off()
```



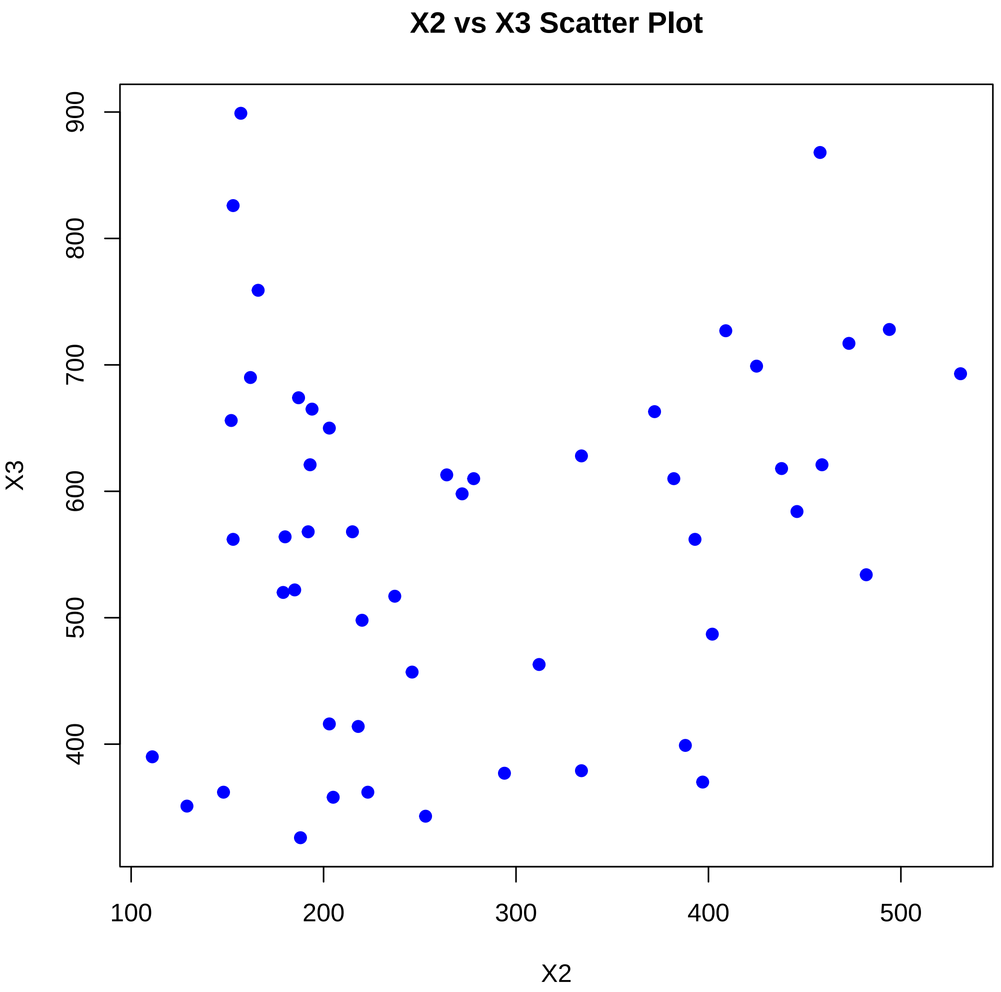
```
#X1 and X2
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
    StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
pdf("plot.X1.X2_YuFan.pdf")
plot(x = expenditure$X1, y = expenditure$X2, main = "X1 vs X2 Scatter
    Plot", xlab = "X1", ylab = "X2", pch = 19, col = "blue")
dev.off()
```



```
#X1 and X3
expenditure <- read.table(”https://raw.githubusercontent.com/ASDS-TCD/
    StatsI_Fall2024/main/datasets/expenditure.txt”, header=T)
pdf(”plot.X1.X3_YuFan.pdf”)
plot(x = expenditure$X1, y = expenditure$X3, main = ”X1 vs X3 Scatter
    Plot”, xlab = ”X1”, ylab = ”X3”, pch = 19, col = ”blue”)
dev.off()
```

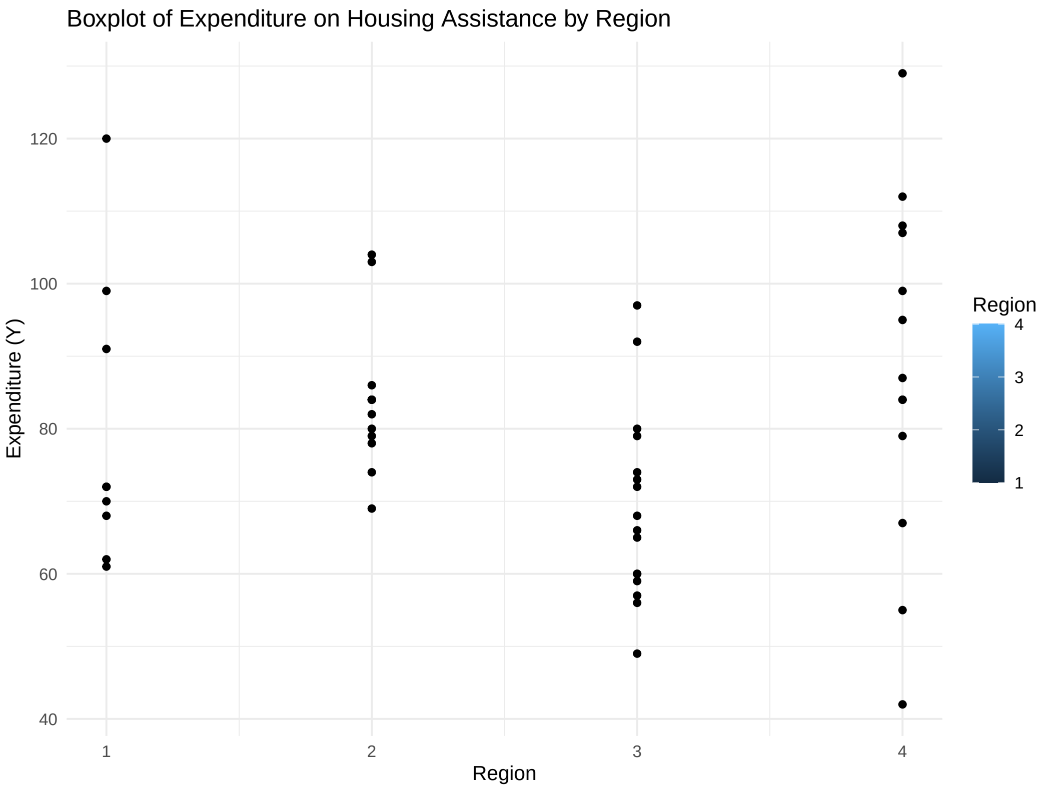


```r
#X2 and X3
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
pdf("plot.X2.X3_YuFan.pdf")
plot(x = expenditure$X2, y = expenditure$X3, main = "X2 vs X3 Scatter Plot", xlab = "X2", ylab = "X3", pch = 19, col = "blue")
dev.off()
```

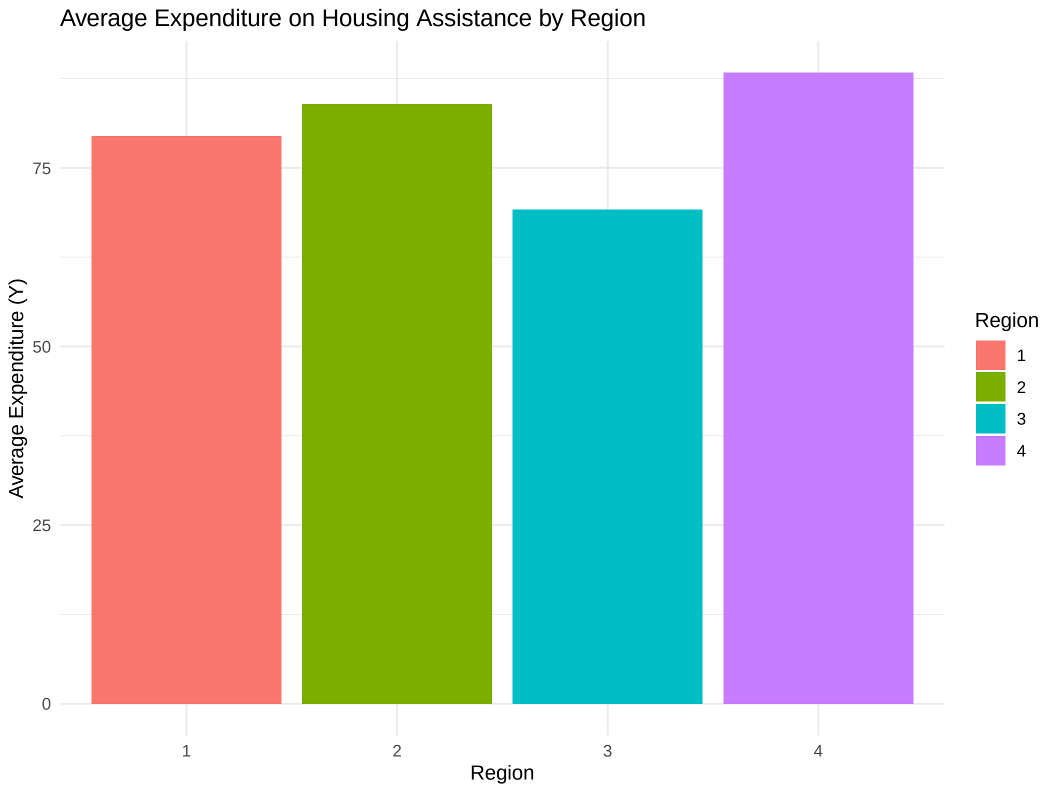

X2 vs X3 Scatter Plot
X3
900
800
700
600
500
400
X2
100
200
300
400
500

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

```r
library(ggplot2)
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
boxplotYregion <- ggplot(expenditure, aes(x = Region, y = Y, fill = Region)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Boxplot of Expenditure on Housing Assistance by Region",
       x = "Region",
       y = "Expenditure (Y)",
       fill = "Region")
ggsave("boxplot.Y.Region_YuFan.pdf", plot = boxplotYregion, width = 8, height = 6)
```



```r
barplotYregion <- ggplot(expenditure, aes(x = Region, y = Y, fill = Region)) +
  geom_bar(stat = "summary", fun = "mean") +
  theme_minimal() +
  labs(title = "Average Expenditure on Housing Assistance by Region",
       x = "Region",
       y = "Average Expenditure (Y)",
       fill = "Region")
```

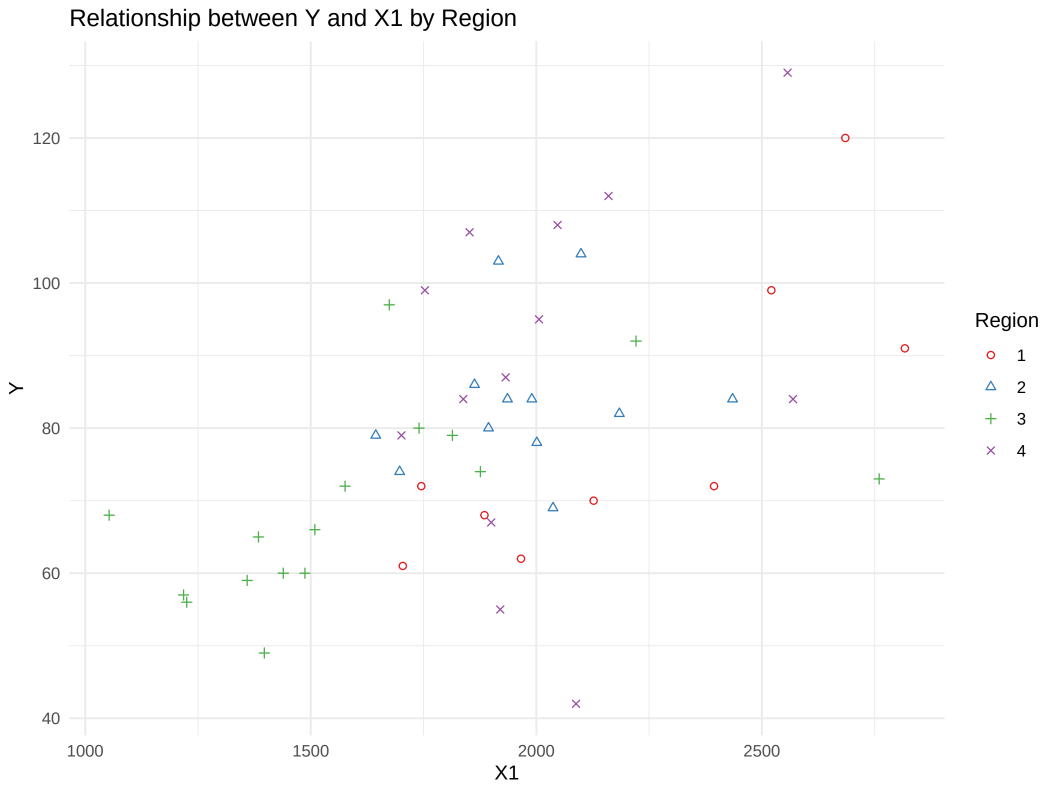


```
ggsave("barplot.Y.Region_YuFan.pdf", plot = barplotYregion, width = 8,
    height = 6)
```

As can be seen from the bar chart, per capita expenditure on housing benefit is highest in region 4

- 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.

```
library(ggplot2)
expenditure$Region <- as.factor(expenditure$Region)
YX1region <- ggplot(expenditure, aes(x = X1, y = Y, color = Region, shape
    = Region)) +
  geom_point() +
  scale_color_brewer(palette = "Set1") +
  scale_shape_manual(values = c(1, 2, 3, 4)) +
  theme_minimal() +
  labs(title = "Relationship between Y and X1 by Region",
       x = "X1",
       y = "Y",
       color = "Region")
```



```
ggsave("plot_Y_X1_by_Region.pdf", plot = YX1region, width = 8, height =
    6)
```