Problem 1

Yajie Dong 23339234

September 29, 2023

Question 1: Education

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

Step 1: Calculate the sample mean

```
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)
y_bar <- mean(y)
y_bar
```

$\bar{y} = 98.44$

Step 2: Calculate the Sample Standard Deviation (S)

```
S <- sd(y)
S
```

$S = \sqrt{\frac{1}{24}\sum_{i=1}^{25}(y_i - 98.44)^2} \approx 13.09$

Step 3: Find the Z-Score

```
confidence_level <- 0.9
alpha <- 1 - confidence_level
z_score <- qnorm(1 - alpha / 2)
z_score
```

Z-score is approximately 1.645

Step 4:Calculate the Margin of Error (ME)

```
ME <- z_score * (S / sqrt(n))
ME
```

$ME = z \times \frac{S}{\sqrt{n}} \approx 1.645 \times \frac{13.09}{5} \approx 4.31$

Step 5: Calculate the confidence Interval

```
CI_lower <- y_bar - ME
CI_upper <- y_bar + ME
cat("90% Confidence Interval for Average Student IQ: [", CI_lower,
   ",", CI_upper, "]\n")
```

Lower limit: $\bar{y} - ME \approx 98.44 - 4.31 \approx 94.13$

Upper limit: $\bar{y} + ME \approx 98.44 + 4.31 \approx 102.75$

Therefore, the 90% confidence interval for the average student IQ in the school is approximately [94.13,102.75]

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.

Step 1: State Hypotheses

-Null Hypothesis$H_0 : \mu = 100$(the average IQ in the school is not different from 100)

-Alternative Hypothesis$H_\alpha : \mu > 100$(the average IQ in the school is greater than 100)

```
mu <- 100
alpha <- 0.05
```

The one-sided test is used in this question because the school counselor is specifically interested in finding out whether the average IQ of students in her school is "higher" than the average IQ of 100 for the general population.

Step 2: Calculate Test Statistic

```
Z <- (y_bar - mu) / (S / sqrt(n))
Z
```

$Z = \frac{\bar{y}-\mu}{\frac{S}{\sqrt{n}}} = \frac{98.44-100}{\frac{13.09}{\sqrt{25}}} \approx -0.60$

Step 3: Find the critical value

```
Z_alpha <- qnorm(1 - alpha)
Z_alpha
```

$Z_\alpha \approx 1.28$

Step 4: Calculate P-value

```
p_value <- 1 - pnorm(Z)
p_value
```

P-value ≈ 0.72

Step 5:Make a Decision.

```
if (Z > Z_alpha) {
  decision <- "Reject Null Hypothesis"
} else {
  decision <- "Fail to Reject Null Hypothesis"
}
decision
```

Neither the Z-test$Z < Z_\alpha$ or the P-value (P-value $> \alpha$) provide evidence to reject the null hypothesis.

Since we fail to reject the null hypothesis, we do not have enough evidence to suggest that the average IQ of the students in the school is greater than 100.

Question 2: Political Economy

1.Plot the relationships among Y, X1, X2, and X3? What are the correlations among them.

Step 1 :Import the data and summary

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

summary_stats <- summary(expenditure[, c("Y", "X1", "X2", "X3")])
cat("Dataset Summary:\n")
print(summary_stats)
```

```
       Y               X1             X2              X3
 Min.   : 42.00   Min.   :1053   Min.   :111.0   Min.   :326.0
 1st Qu.: 67.25   1st Qu.:1698   1st Qu.:187.2   1st Qu.:426.2
 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
```

Step2: Calculate and print the correlation matrix

```
cor_matrix <- cor(expenditure[, c("Y", "X1", "X2", "X3")])
cat("Correlation Matrix:\n")
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
```

Step 3 :Plot the relationships among Y, X1, X2, and X3

```
1
2  plot1 <- ggplot(expenditure, aes(x = X1, y = Y)) + geom_point() +
       geom_smooth(method = 'lm') + ggtitle("Y vs X1")
3  plot2 <- ggplot(expenditure, aes(x = X2, y = Y)) + geom_point() +
       geom_smooth(method = 'lm') + ggtitle("Y vs X2")
4  plot3 <- ggplot(expenditure, aes(x = X3, y = Y)) + geom_point() +
       geom_smooth(method = 'lm') + ggtitle("Y vs X3")
5  plot4 <- ggplot(expenditure, aes(x = X1, y = X2)) + geom_point() +
       geom_smooth(method = 'lm') + ggtitle("X1 vs X2")
6  plot5 <- ggplot(expenditure, aes(x = X1, y = X3)) + geom_point() +
       geom_smooth(method = 'lm') + ggtitle("X1 vs X3")
7  plot6 <- ggplot(expenditure, aes(x = X2, y = X3)) + geom_point() +
       geom_smooth(method = 'lm') + ggtitle("X2 vs X3")
8  print(plot1)
9  print(plot2)
10 print(plot3)
11 print(plot4)
12 print(plot5)
13 print(plot6)
```

Figure 1: Y and X1.

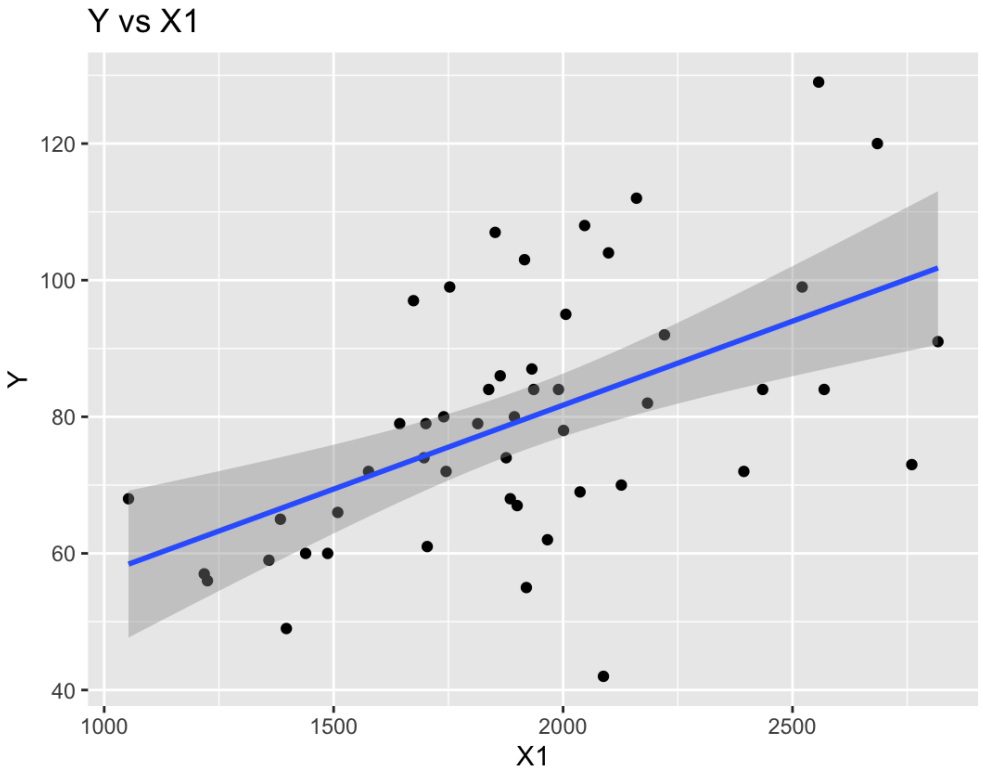


Figure 2: Y and X2.

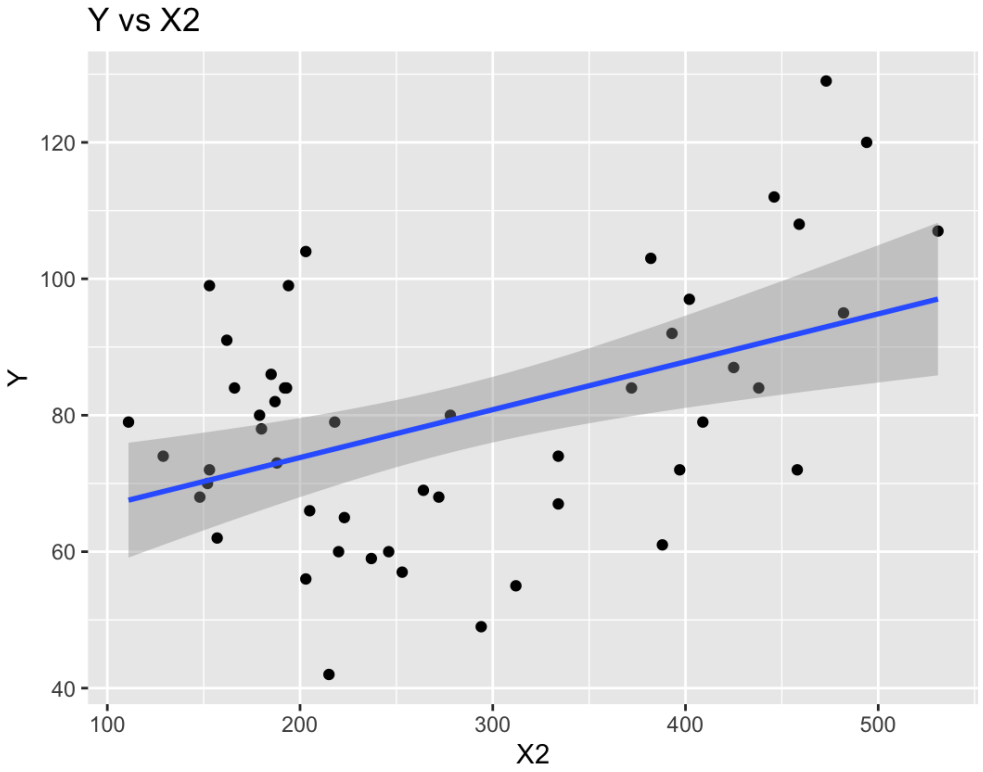


Figure 3: Y and X3.

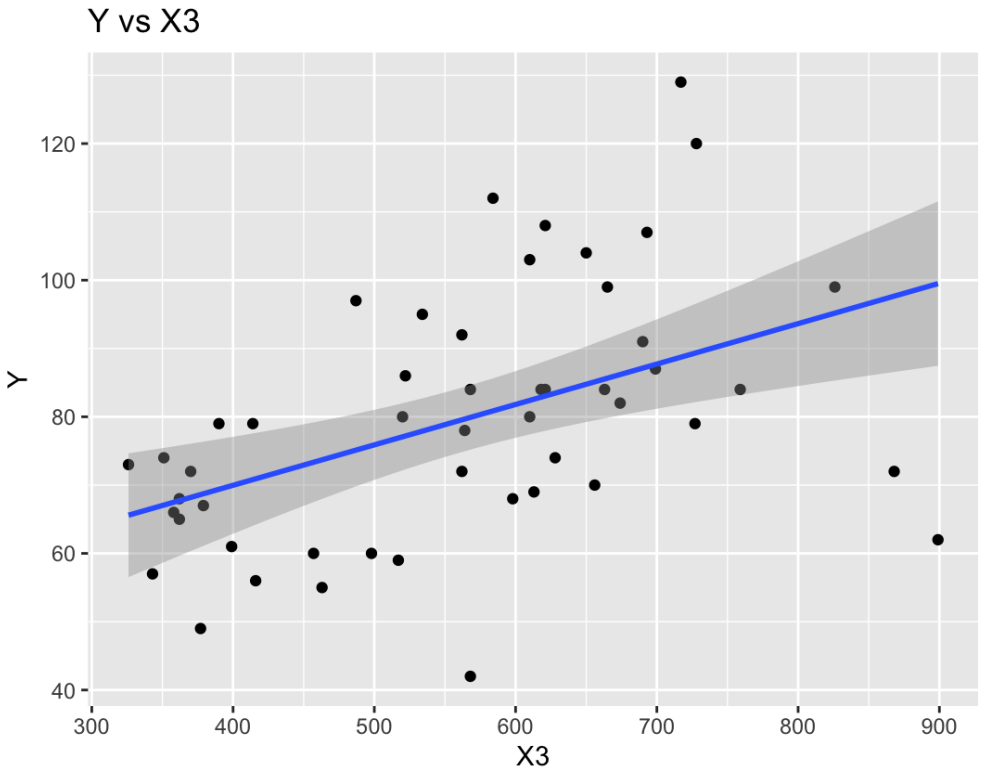


Figure 4: X1 and X2.

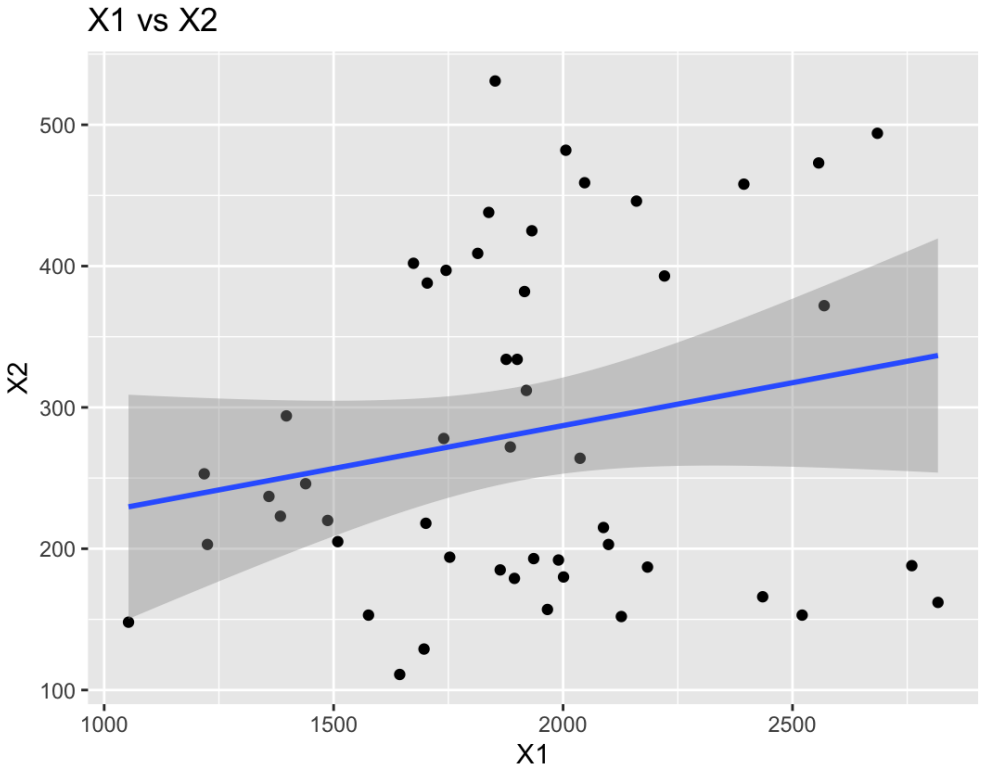


Figure 5: X1 and X3.

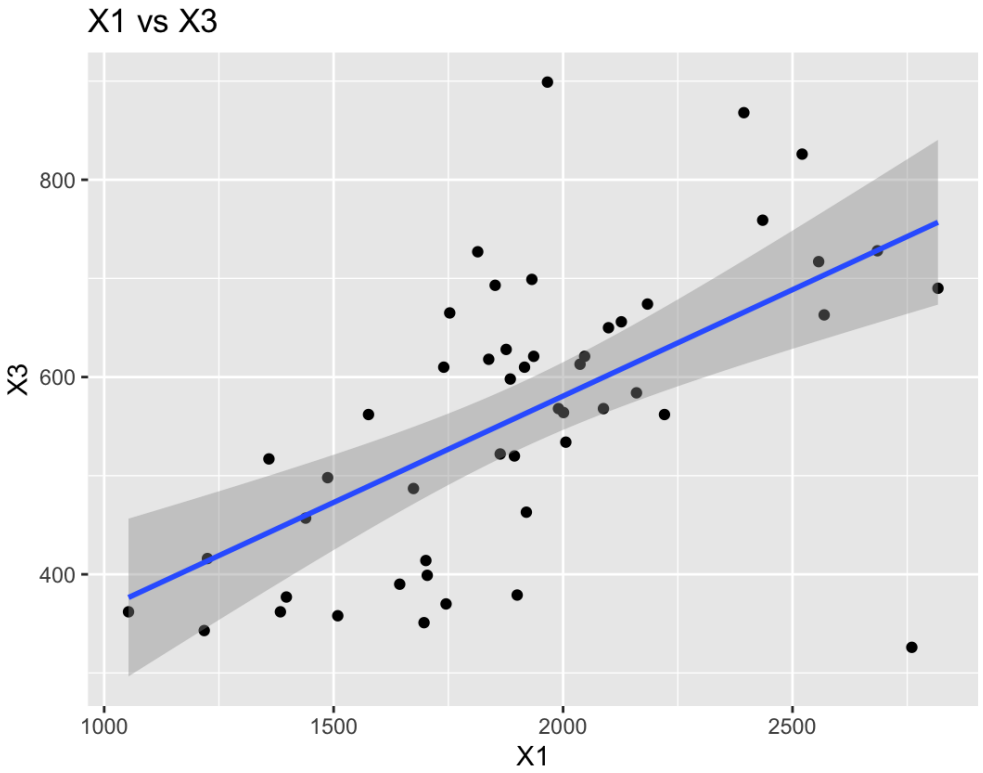
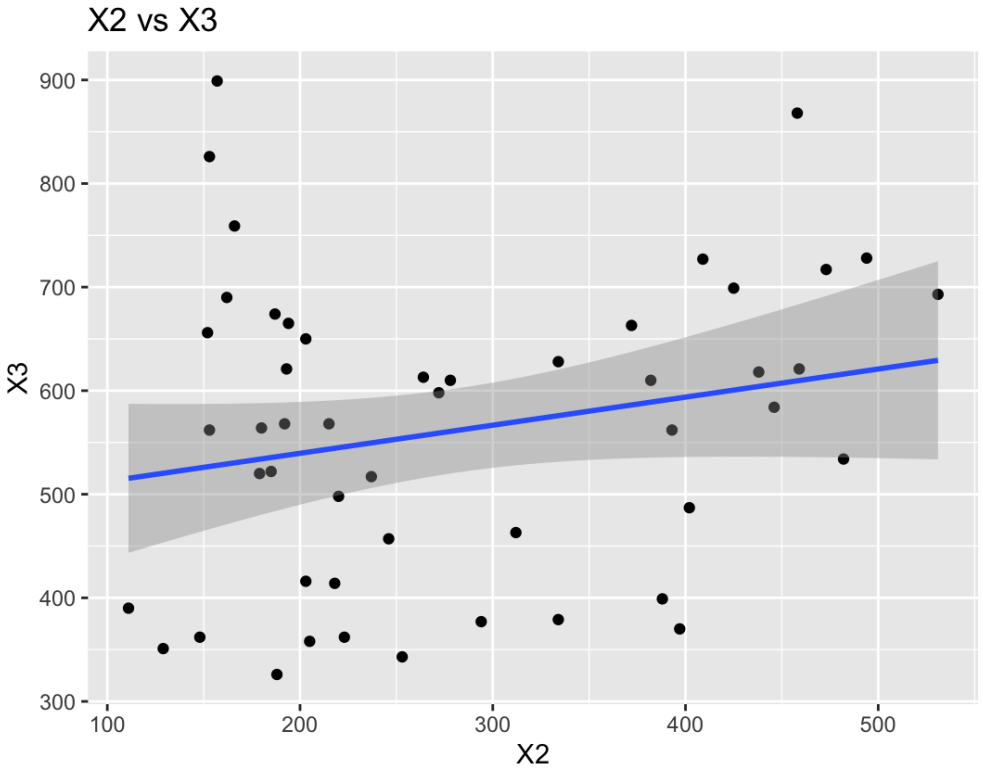


Figure 6: X2 and X3.



Conclusion Based on the correlation matrix and graphs:

1. Y and X1 are positively related. When X1 goes up, Y tends to also go up.

2. Y and X2 have a moderate positive relationship. When X2 increases, Y generally goes up, but not as consistently as with X1.

3. Y and X3 also have a moderate positive relationship. Similar to X2, when X3 rises, Y often goes up, but not all the time.

4. X1 and X2 have a low positive relationship. They somewhat move in the same direction, but it's not very consistent.

5. X1 and X3 are positively related. When one increases, the other usually does too.

6. X2 and X3 have a low positive relationship. They somewhat move in the same direction, but it's not very consistent.

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

Step 1:Calculate the average per capita expenditure on housing assistance for each region

```
mean_expenditure_by_region <- aggregate(Y ~ Region, data =
    expenditure, FUN = mean)

print(mean_expenditure_by_region)

highest_region <- mean_expenditure_by_region[which.max(mean_
    expenditure_by_region$Y), ]
cat("The region with the highest per capita expenditure on housing
    assistance is:", highest_region$Region)
```

```
  Region        Y
1      1 79.44444
2      2 83.91667
3      3 69.18750
4      4 88.30769
```

```
The region with the highest per capita expenditure on housing assistance is: 4
```

Step 2: Plot the relationship between Y and Region

```r
highest_region <- mean_expenditure_by_region[which.max(mean_
   expenditure_by_region$Y), ]
cat("The region with the highest per capita expenditure on housing
   assistance is:", highest_region$Region)

plot7<- ggplot(expenditure, aes(x = as.factor(Region), y = Y, fill
   = as.factor(Region))) +
  geom_boxplot() +
  ggtitle("Y vs Region") +
  scale_fill_manual(values = c("red", "blue", "green", "purple"))
print(plot7)
```

Description

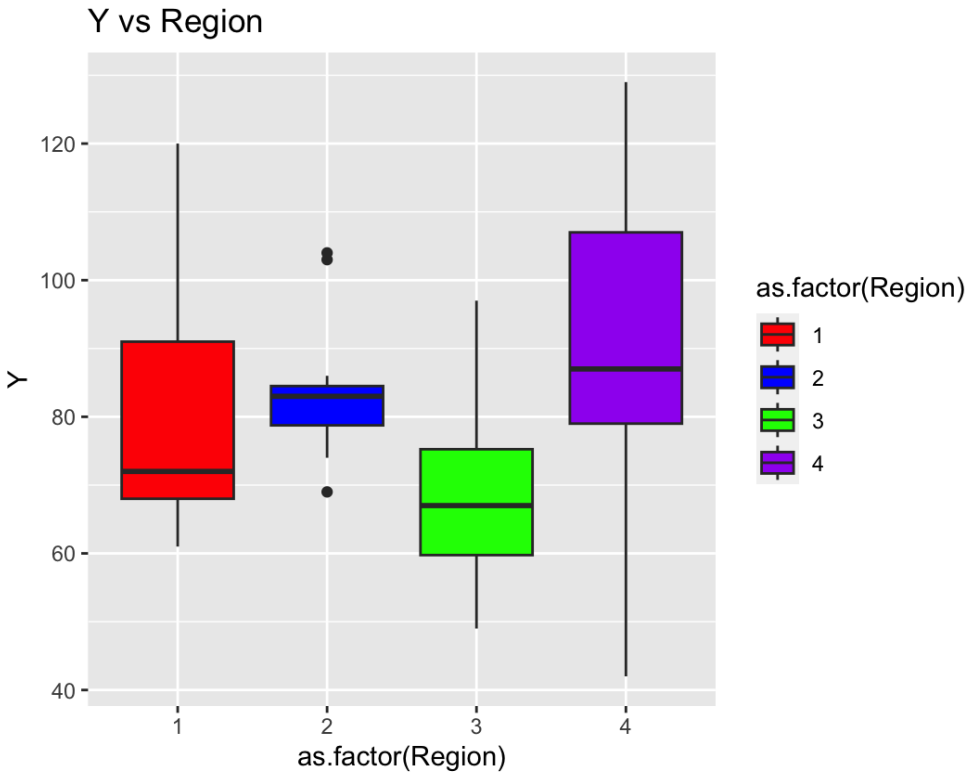
1. Northeast Region (Red): This region has relatively stable and moderately high expenditure on average.

2. North Central Region (Blue): Expenditure levels are moderate and relatively consistent in this region.

3. South Region (Green): Expenditure levels vary significantly, with some areas having low expenditure and others having high expenditure.

4. West Region (Purple): On average, this region has the highest expenditure, but there is also some variation in expenditure levels.

Figure 7: Y and Regions.



In summary, the West region(4) has the highest average expenditure, the South region(3) shows the most variation, and the Northeast(1) and North Central regions(2) have moderate expenditure levels.

3.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.

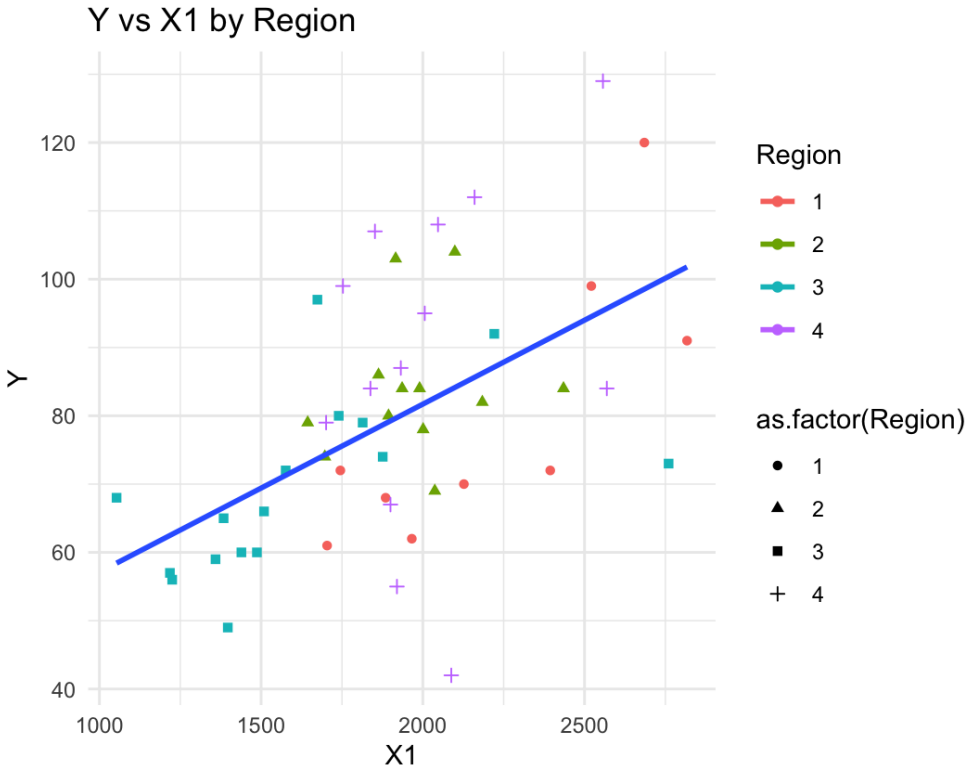
Plot the relationship between Y, as shown in Figure 1.

Reproduce the graph of Y and X1.

```r
plot8 <- ggplot(expenditure, aes(x = X1, y = Y, color = as.factor(
    Region), shape = as.factor(Region))) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE, aes(group = 1)) +
  ggtitle("Y vs X1 by Region") +
  xlab("X1") +
  ylab("Y") +
  labs(color = "Region") +
  theme_minimal()

print(plot8)
```

Figure 8: Y and X1.



From the graph, it appears that there is a linear relationship between Y and X1. Therefore,I proceed to conduct a statistical test in R to confirm this.

```r
save_lm_summary <- function(model, file_name) {
  summary_text <- capture.output(summary(model))
  writeLines(summary_text, con = file_name)
}
# Run the linear regression model between Y and X1
linear_regression_Y_X1 <- lm(Y ~ X1, data = expenditure)

linear_regression_summary <- summary(linear_regression_Y_X1)

linear_regression_summary_text <- capture.output(linear_regression_summary)
# Save the summary to a .tex file
writeLines(linear_regression_summary_text, "/Users/miadong/Desktop/linear_regression_Y_X1_summary.tex")
# Load the stargazer package
library(stargazer)

# Run the regression
regression1 <- lm(Y ~ X1, data = expenditure)

# Define the output_stargazer function
output_stargazer <- function(outputFile, ...) {
  output <- capture.output(stargazer(..., type = "latex"))
  cat(paste(output, collapse = "\n"), "\n", file = outputFile, append = TRUE)
}

# Use the function to save the LaTeX-formatted table to a .tex file
output_stargazer("regression_output11.tex", regression1)
getwd()
```

Table 1:

	Dependent variable:
	Y
X1	0.025^{***}
	(0.006)
Constant	32.546^{***}
	(11.034)
Observations	50
R^2	0.283
Adjusted R^2	0.268
Residual Std. Error	15.836 (df = 48)
F Statistic	18.920^{***} (df = 1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

Based on the data presented in the table, a linear regression model is evident between Y and $X1$, represented by the equation $Y = 32.546 + 0.025 \cdot X1$.