Problem Set 1

Applied Stats/Quant Methods 1

Due: October 1, 2023 /// Wei Tang 23362496

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:

```
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.

```
n <- length(y)
sample_mean <- mean(y)
sample_sd <- sd(y)

#step-by-step method
t_score <- qt(1-(1-0.9)/2,df<- length(y)-1)
```

```
lower_90_t <- mean(y) - (t_score)*(sd(y)/sqrt(length(y)))
higher_90_t <- mean(y) + (t_score)*(sd(y)/sqrt(length(y)))

#quicker method
t.test(y,conf.level = 0.9 , alternative = "two.sided")

#answer
CI <- c(lower_90_t,higher_90_t)
CI
```

Answer: The 90% confidence interval for the average student IQ in this school is [93.95993 : 102.92007]

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 $\alpha = 0.05$.

```
se <- sample_sd/sqrt(n)
t_value <- (mean(y)-100)/(se)
p <- pt(t_value,df = n-1,lower.tail = FALSE)
t.test(y, mu = 100,  alternative = "greater")
```

Result of console:

```
One Sample t-test
data:  y
t = -0.59574, df = 24, p-value = 0.7215
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
 93.95993      Inf
sample estimates:
mean of x
    98.44
```

Answer: Because α == 0.05 < p-value ==0.7215,so we can NOT REJECT the null hypothesis that the average student IQ in this school is less than the average IQ score (100) among all the schools in the country. In other words, we can not support the counselor's hypothesis (but we can not reject it either).

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
pdf(file="Y-X1_X2_X3_Relationship_Plot.pdf")
plot(expenditure[c("Y", "X1", "X2", "X3")])
dev.off()
```

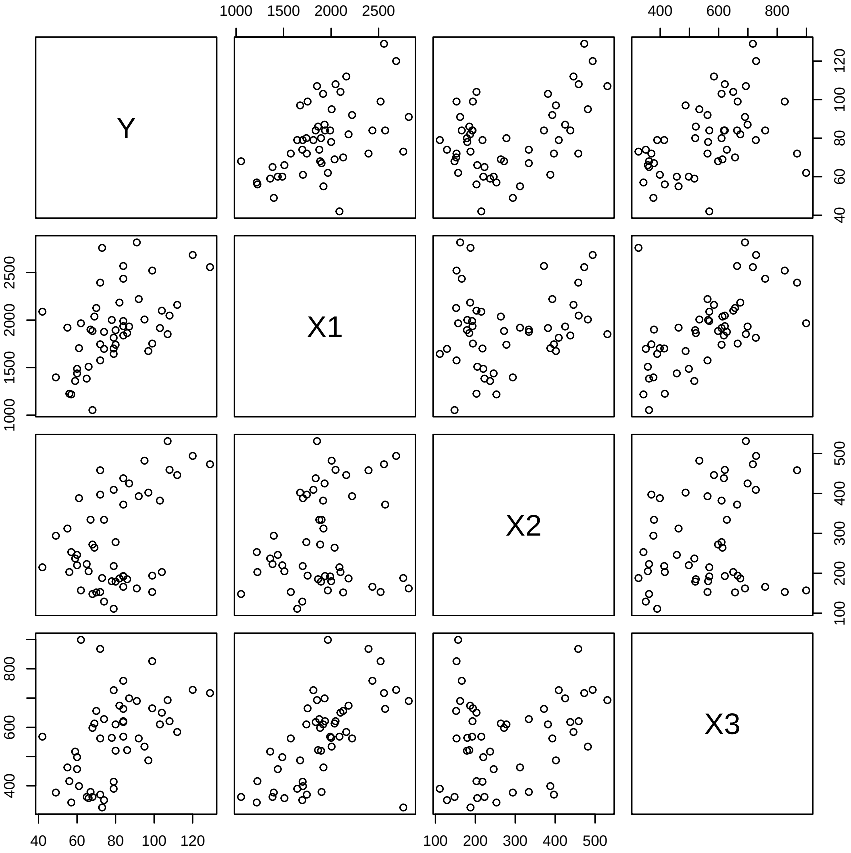
```
Answer: I think Y-X1 , Y-X2 , Y-X3 , X1-X3 are linearly correlated,
I use the methods to prove my idea.
```

```r
cor(expenditure$Y,expenditure$X1)
cor(expenditure$Y,expenditure$X2)
cor(expenditure$Y,expenditure$X3)
cor(expenditure$X2,expenditure$X3)
cor(expenditure$X1,expenditure$X3)
cor(expenditure$X2,expenditure$X3)
```

```
 Results of console are as follows:

 cor(expenditure$Y,expenditure$X1)
[1] 0.5317212
> cor(expenditure$Y,expenditure$X2)
[1] 0.4482876
> cor(expenditure$Y,expenditure$X3)
[1] 0.4636787
> cor(expenditure$X2,expenditure$X3)
```

```
[1] 0.2210149
> cor(expenditure$X1,expenditure$X3)
[1] 0.5952504
> cor(expenditure$X2,expenditure$X3)
[1] 0.2210149
```

Figure 1: Y-X1-X2-X3 Relationship Plot in R.

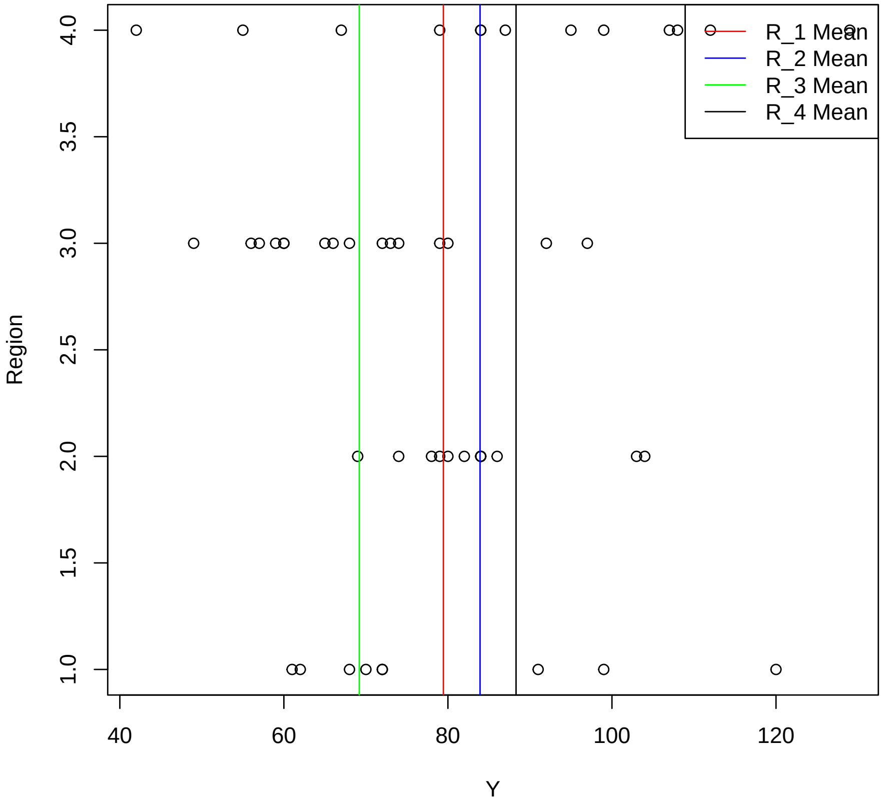


- Please plot the relationship between Y and *Region*? On average, which region has the highest per capita expenditure on housing assistance?

```
pdf(file="Y-Region_Relationship_Plot.pdf")
plot(expenditure[c("Y","Region")])
str(expenditure)
R_1 <- expenditure[expenditure$Region == 1, ]
R_2 <- expenditure[expenditure$Region == 2, ]
R_3 <- expenditure[expenditure$Region == 3, ]
R_4 <- expenditure[expenditure$Region == 4, ]

```

```
t.test(R_1$Y,conf.level = 0.95 , alternative = "two.sided")
t.test(R_2$Y,conf.level = 0.95 , alternative = "two.sided")
t.test(R_3$Y,conf.level = 0.95 , alternative = "two.sided")
t.test(R_4$Y,conf.level = 0.95 , alternative = "two.sided")

abline(v=mean(R_1$Y),col="red")
abline(v=mean(R_2$Y),col="blue")
abline(v=mean(R_3$Y),col="green")
abline(v=mean(R_4$Y),col="black")

legend("topright", legend = c("R_1 Mean", "R_2 Mean", "R_3 Mean", "R_4
    Mean"), col = c("red", "blue", "green", "black"),lty=1 )
dev.off()
```

Figure 2: Y-Region Relationship Plot in R.

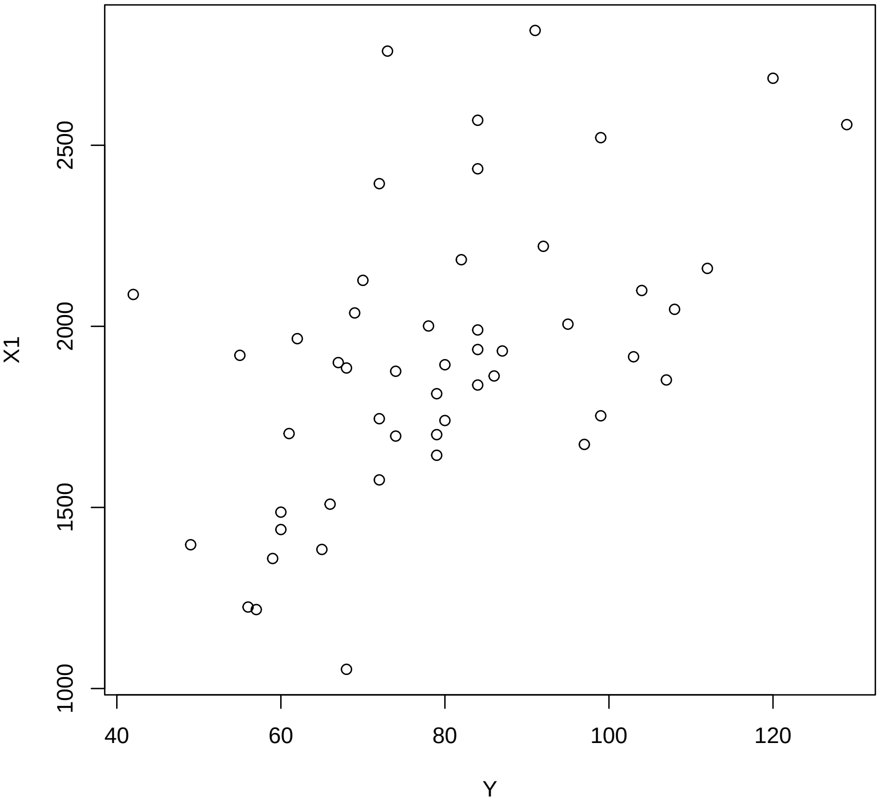


Answer: Obviously Region 4 has the highest per capita expenditure on housing assistance

- 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.

```
pdf(file="Y-X1_Relationship_Plot.pdf")
plot(expenditure[c("Y","X1")])
dev.off()
```

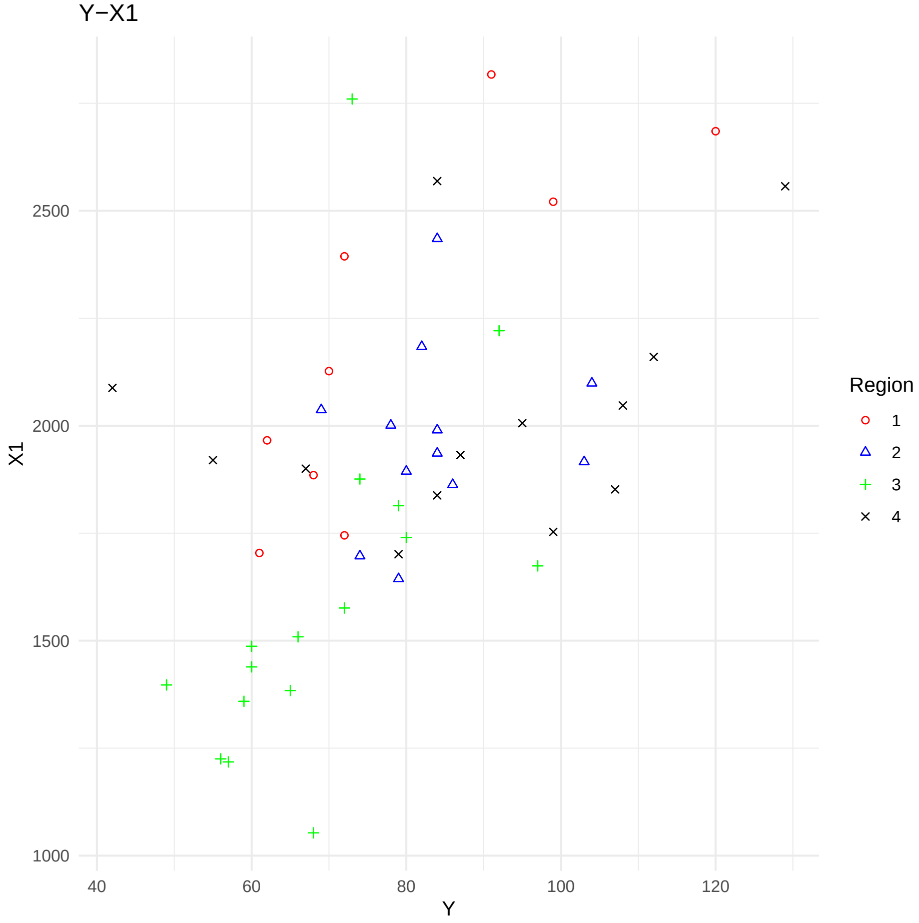
Figure 3: Y-X1 Relationship Plot in R.



`Answer: It looks like there is a linear correlation between Y and X1.`

```
expenditure$Region <- factor(expenditure$Region)
pdf(file="Y-X1-Region_Relationship_Plot.pdf")
ggplot(expenditure, aes(x = Y  , y = X1, color = Region ,shape = Region))
    +
  geom_point() +
  labs(x = "Y", y = "X1", title = "Y-X1") +
  scale_color_manual(values = c( "1" = "red", "2" = "blue", "3" = "green"
    , "4" = "black")) +
  scale_shape_manual(values = c(”1” = 1, ”2” = 2, ”3” = 3, ”4” = 4)) +
  theme_minimal()
dev.off()
```

Figure 4: Y-X1-Region Relationship Plot in R.


Y−X1
X1
Y
Region
○ 1
△ 2
× 4