Problem Set 1

Applied Stats/Quant Methods 1
Dan Zhang 23335541

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:

```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
alpha_1<-0.1
n<-length(y)
mean_y <-mean(y)
```

```r

# standard error
se<-sd(y)/sqrt(n)

# critical value of t distribution
t_score<-qt(1-alpha_1/2,df=n-1)

#calculate confidence interval
lower_90<-mean_y-t_score*se
upper_90<-mean_y+t_score*se
```

```
The 90% confidence interval for the average student IQ in the school is :
[  93.96 , 102.92 ]
```

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 $\alpha = 0.05$.

```
step1: Build hypothesis:
  Null hypothesis: The average IQ of school students is equal to the national
average IQ.(mu=100)
  Alternative hypothesis: The average IQ of school students is higher than the
national average IQ.(mu>100)
step2:Calculate statistic value by the following code in R
```

```r
#Hypothesis test
mu<-100 #average IQ in all the country
alpha_2 <- 0.05
t_test_result<-t.test(y,mu=mu,alternative = "greater",alpha_2=0.05)
t_test_result
cat("t-statistic=", round(t_test_result$statistic,2),"\n")
cat("p_value=",round(t_test_result$p.value,2),"\n")

#determine whether to reject null hypothesis or not
if (t_test_result$p.value < alpha_2){
  cat("Reject null hypothesis, the school average IQ socre is higher than
    the average IQ score in the country.")
}else{
  cat("Failed to reject null hypothesis, there is no enough evidence to
   support that the school average IQ socre is higher than the average IQ
    score in the country.")
}
```

```
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
```

Conclusion:
As p-value(0.72) is greater than the significance level(0.05). So we can not reject null hypothesis. Which means based on the sample provided,there is no enough evidence to support that the school average IQ socre is higher than the average IQ score in the country.

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*

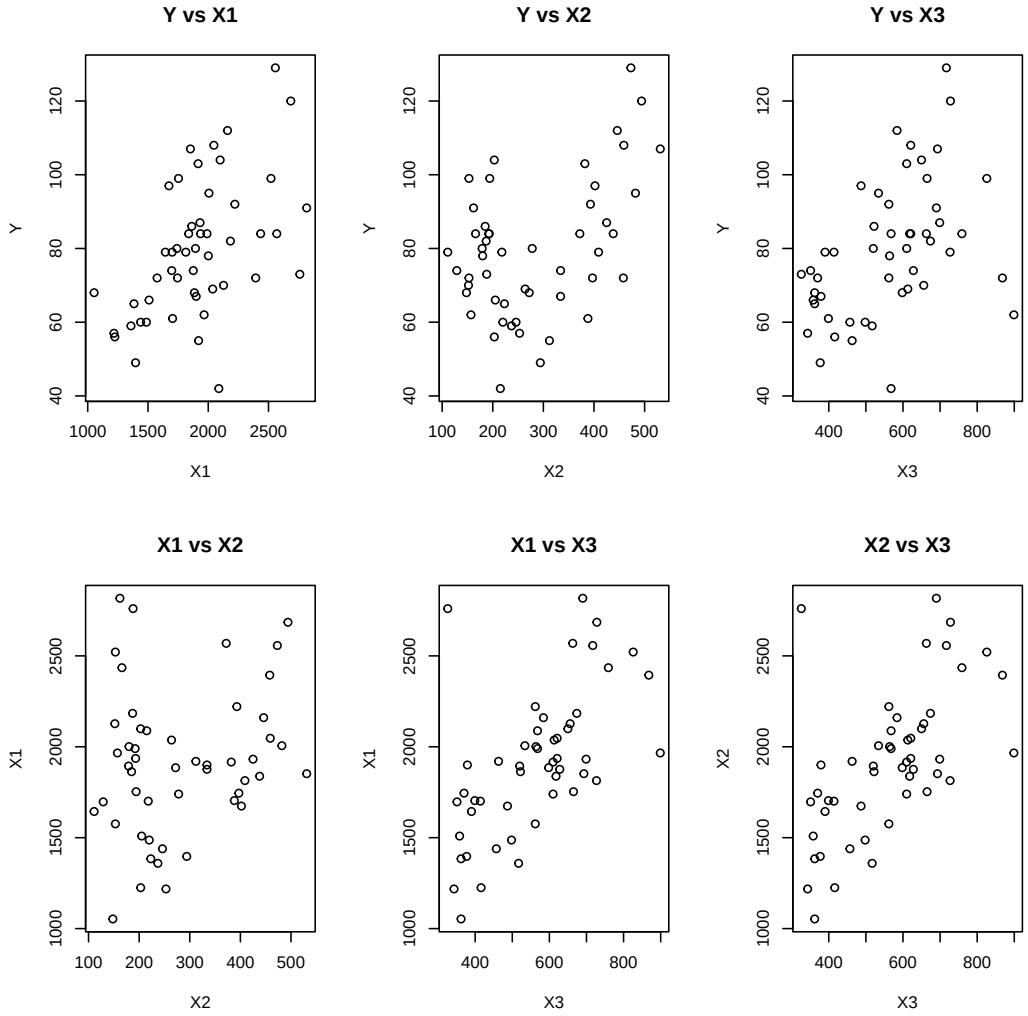
Explore the `expenditure` data set and import data into `R`.

- 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)?

```
#Create scatter plots
pdf("Problem2_Question1_ScatterPlots.Pdf")
par(mfrow=c(2,3))
#Y vs X1
plot(expenditure$X1,expenditure$Y,xlab = "X1",
     ylab = "Y",
     main = "Y vs X1")
#Y vs X2
plot(expenditure$X2,expenditure$Y,xlab = "X2",
     ylab = "Y",
     main = "Y vs X2")
#Y vs X3
plot(expenditure$X3,expenditure$Y,xlab = "X3",
     ylab = "Y",
     main = "Y vs X3")
#X1 vs X2
plot(expenditure$X2,expenditure$X1,xlab = "X2",
     ylab = "X1",
     main = "X1 vs X2")
#X1 vs X3
plot(expenditure$X3,expenditure$X1,xlab = "X3",
     ylab = "X1",
     main = "X1 vs X3")
#X2 vs X3
plot(expenditure$X3,expenditure$X1,xlab = "X3",
     ylab = "X2",
```

```
      main = "X2 vs X3")
par(mfrow=c(1,1))
dev.off()
```

Figure 1: The relationships between Y/X1/X2/X3 in R.



Conclusion:
As we can see from the scatter plots between Y,X1,X2 and X3. There are positive correlations between Y and X1, Y and X3, X1 and X3, X2 and X3,respectively. When the former variable increases, the latter variable also shows an increasing trend. The relationships between X1 and X3, X2 and X3, are stronger than the relationships between Y and X1, Y and X3.

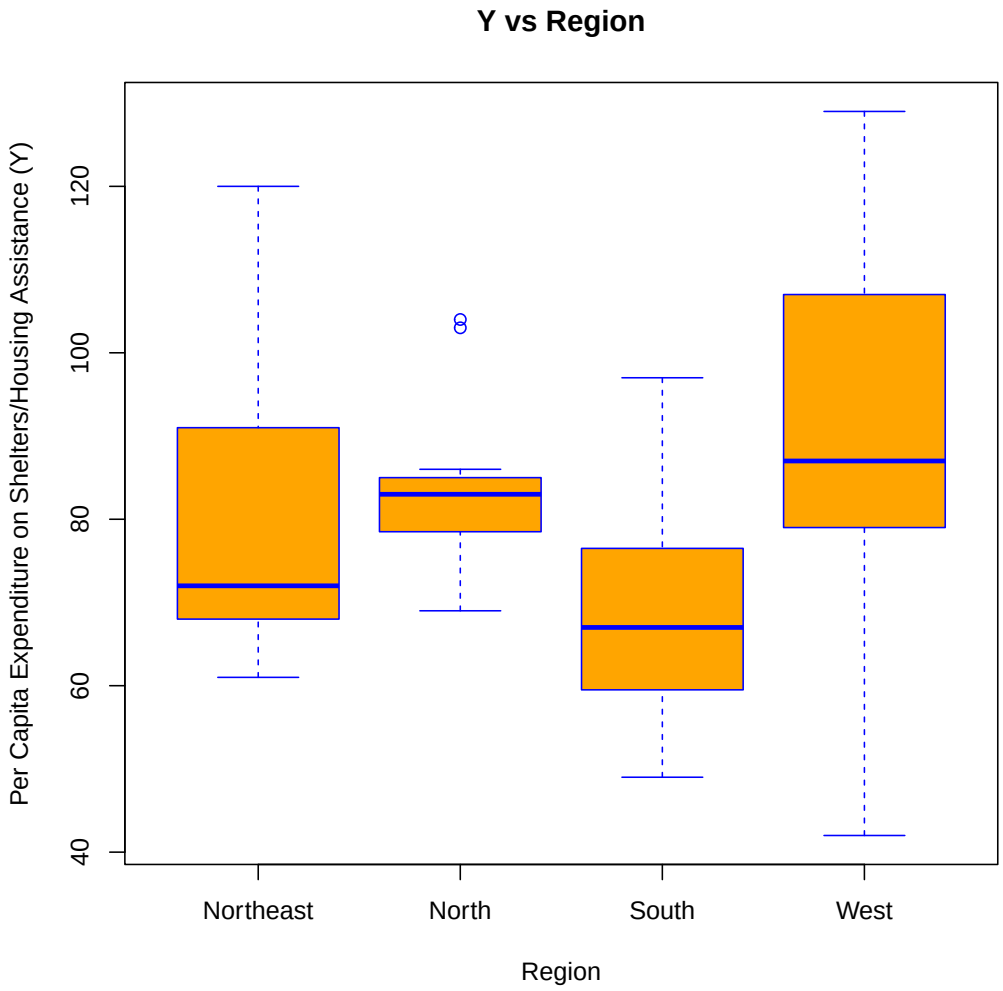
Besides, the correlations between Y and X2, X2 and X2 seem to be

non-linear. When X2 is less than about 300, there are negative relationships between Y and X2,X1 and X2. While X2 is greater than about 300, there are positive relationships between Y and X2,X1 and X2.

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

```
#Create box plot
pdf("Problem2_Question2_BoxPlot.pdf")
boxplot(expenditure$Y~expenditure$Region,xlab = "Region",
        ylab = "Per Capita Expenditure on Shelters/Housing Assistance (Y)
    ",
        xaxt="n",
        main="Y vs Region",
        col="orange",
        border="blue")
axis(1, at = 1:4, labels =c ("Northeast", "North", "South", "West"))
dev.off()
```

Figure 2: The relationships between Y and Region in R.



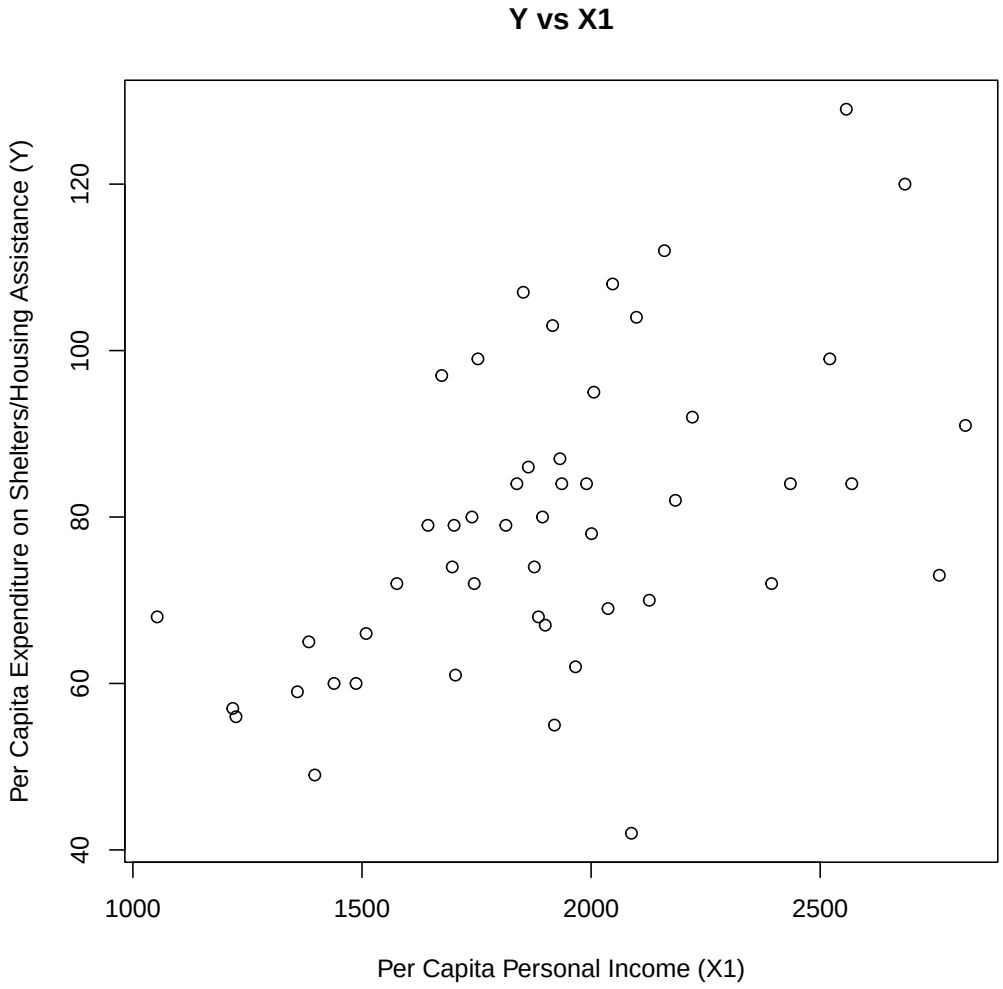

```
Conclusion:
According to the boxplot,we can see that region 4 has the highest per
capita expenditure on housing assistance.
```

- 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.

```
#Plot Y vs X1 again
pdf("Problem2_Question3_Plot_Y_vs_X1.pdf")
plot(expenditure$X1,expenditure$Y,xlab = "Per Capita Personal Income (X1)
    ",
     ylab = "Per Capita Expenditure on Shelters/Housing Assistance (Y)",
```

```
5       main = "Y vs X1")
6 dev.off()
```

Figure 3: The relationships between Y and X1 in R.



As can be observed from the scatter plot, the X-axis represents variable Per Capita Personal Income (X1) and the Y-axis represents variable Per Capita Expenditure on Shelters/Housing Assistance (Y).

We can see that here is a positive correlation between Y and X1, as X1 increases, Y also shows an upward trend. The strength of this relationship is weak and not very close. There might be some sorrelation between Y and X1. Further statistical analysis may help quantify the strength and significance of this relationship.

```r
#Reproduce a plot include region variable
pdf("Problem2_Question3_Plot_Y_X1_Region.pdf")
plot(expenditure$X1,expenditure$Y,xlab = "Per Capita Personal Income (X1)
    ",
     ylab = "Per Capita Expenditure on Shelters/Housing Assistance (Y)",
     main = " ")

#create symbol and color vectors for different regions
regions<-unique(expenditure$Region)
symbols<-c("A","B","C","D")
colors<-c("red","green","purple","blue")

#Using different symbol and color to plot scatter points for each region
for (i in 1:length(regions)) {
  region_points<-expenditure[expenditure$Region==regions[i] , ]
  points(region_points$X1,region_points$Y,
         pch=symbols[i],
         col=colors[i])

}
legend("topright",legend=(names(regions)<-c ("Northeast", "North", "South
    ", "West")),col=colors,pch=symbols,title="Region")
title(main=" Scatter Plot of Y vs X1 by Region")
dev.off()
```

Figure 4: The relationships between Y X1 and Region in R.