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
#Question 1 - find the CI for the same student IQ scores 'y' given a 90% confidence interval
ci<-t.test(y, conf.level = 0.9)
ci
#The 90% confidence interval for the sample student IQ scores with a confidence level of 90% 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 $\alpha = 0.05$.

```r
#Question 2 - considering the national average IQ is 100, determine if the average student IQ score in the school is greater than the national average, using the sample set y, with alpha=0.05
# Null Hypothesis: The average student IQ score in the school is 100 (H0: u=100)
# Alternate hypothesis: The average student IQ score in the school is greater than 100 (HA: u>100)
#As determining if IQ is greater than 100, use a one-tailed (right, so positive) t-test
hyp<-t.test(y, mu=100, alternative="greater", conf.level=0.95)
hyp
#P-value is 0.7215, the data does not have any statistical importance at the 95% confidence level as p>0.05, we cannot reject the null hypothesis as the dataset of sample student IQ scores is insufficient evidence to determine if the average student IQ score of the school is greater than the national average of 100.
```

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.

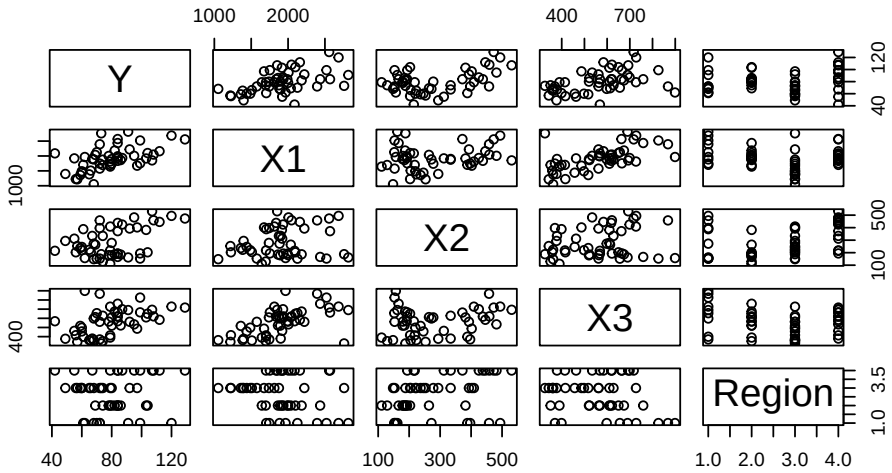
`State`	*50 states in US*
`Y`	*per capita expenditure on shelters/housing assistance in state*
`X1`	*per capita personal income in state*
`X2`	*Number of residents per 100,000 that are "financially insecure" in state*
`X3`	*Number of people per thousand residing in urban areas in state*
`Region`	*1=Northeast,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

```
library(corrplot)
str(expenditure)
numExp<-expenditure[sapply(expenditure, is.numeric)]
corRel<-cor(numExp)
corrplot(corRel)
#We can see the strongest variable relationship correlation between X1(per capita personal income in state) and X3(Number of people per thousand residing in urban areas in state), with a correlation of 0.4 or higher. Y(per capita expenditure on shelters/housing assistance in state) has a not as strong but still notable correlation with X1, X2(Number of residents per 100,000 that are 'financially insecure' in state), and X3. We expect to see rough trajectories of linearly increasing relationships when the scatterplots of these variables are created. Region has very weak correlational relationships with all variables except X2, with which it has a midly positive relationship.
print(plot(numExp))
#From the scatterplots generated, the relationships are all reflective of what the correlation plot matrix inferred. Region has parallel lines of point densities, which is to be expected as regions are numerical natural numbers - no decimal values were possible. (Parallel in that when region is on the X-axis we see horizontal parallel lines when all other variables are held individually against it on the y-axis, and vertical when it is repeated but with region on the y-axis and all other variables compared to it on the x-axis). The cleasrest trendline can be seen as between X1 and X3, with the densest areas of points following along an increasning linear slope (only one notable outlier can be seen on the top left corner when X1 is on the y-axis
```

```
and X3 on the x-axis). Linearly increasing relationships are visible
when Y is held against X1, X2 and X3, but not to the same degree of
clarity as that between X1 and X3.
```



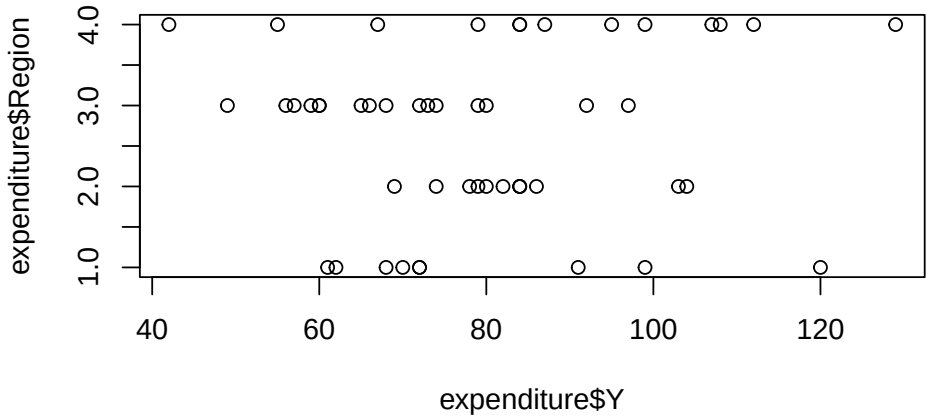
- Please plot the relationship between Y and *Region*? On average, which region has the highest per capita expenditure on housing assistance?

```
#Question 2 - Plotting the relationship between Y and Region:
plot(expenditure$Y,expenditure$Region)

#Which region has the highest per capita expenditure on housing
    assistance?
avRegion<-tapply(expenditure$Y, expenditure$Region, mean)
avRegion
#from this, we can see that region 4 has the highest average per capita
    expenditure on housing assistance.
```

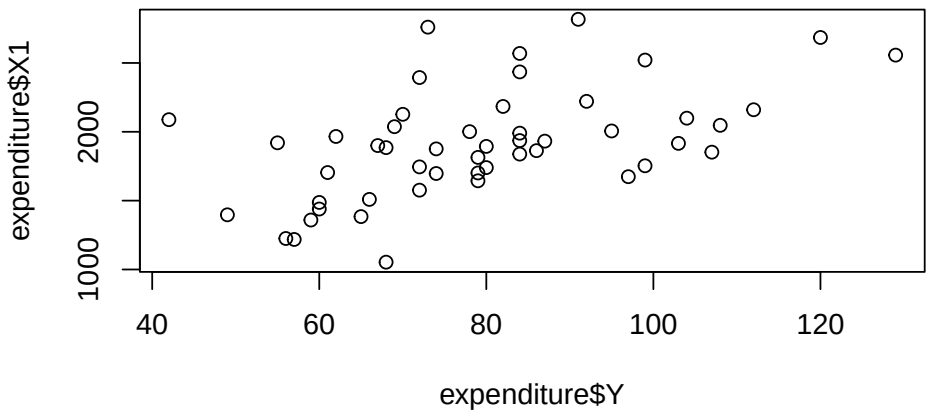
Please plot the relationship between Y and *X1*? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.

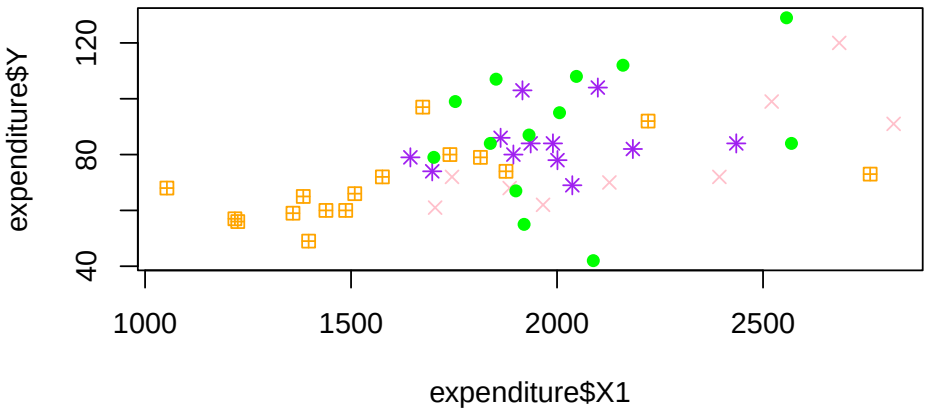
```
#Question 3 - Plot the graph between Y and X1,and describe the
    relationship
plot(expenditure$Y,expenditure$X1)
#There is an increasing linear relationship between X1 and Y. We can see
    that in general that as Y increases, so too does X1. This is clearly
    not a hard and fast rule, as there are several outliers, one in
    particular appearing on the left hand side of the graph with one value
    appearing slightly above 2000 on the X1 axis with it's corresponding
    Y value being only slightly beyond 40 (if this point were to follow
    the general trend of the other points on the graph, we would expect to
```




```
    see it surpassing 80 on Y, roughly double it's actual value). The
    points while increasing also do not have the level of density the X1-
    X3 graph depicted around the central increasing slope.

#Reproduce graph but include variable region, with it colour-coded and
    ustilising symbols to differentiate between each of the four regions
colour<-c("pink","purple","orange","green")[expenditure$Region]
symbol<-c(4,8,12,16)[expenditure$Region]
plot(expenditure$X1,expenditure$Y,
     col=colour, pch=symbol)
```




expenditure$Y
40
80
120
1000
1500
2000
2500
expenditure$X1