Problem Set 1

Applied Stats/Quant Methods 1

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:

```
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.

Input

```r
# way 1: mathematical method
t_score <- qt(0.95, df=length(y)-1)
lower_90_y <- mean(y)-(t_score)*(sd(y)/sqrt(length(y)))
upper_90_y <- mean(y)+(t_score)*(sd(y)/sqrt(length(y)))

# combine two ends into a confident interval and print result in a
    formated way
ci_y <- c(round(lower_90_y, 2), round(upper_90_y, 2))
ci_y
cat("The 90% confidence interval for the average student IQ in the school
     is:",
     paste(ci_y, collapse = "~"))


# way 2: encapsulated function, an easier method
T <- t.test(y, conf.level = 0.90, alternative = "two.sided")
str(T)
# subset the lower and upper end of the confident interval
T_lower_90_y <- T$conf.int[1]
T_upper_90_y <- T$conf.int[2]

# combine two ends into a confident interval and print result in a
    formated way
T_ci_y <- c(round(T_lower_90_y, 2), round(T_upper_90_y, 2))
cat("The 90% confidence interval for the average student IQ in the school
     is:",
     paste(T_ci_y, collapse = "~"))
```

Output

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 $\alpha = 0.05$.

   Research Hypothesis

   - Null Hypothesis: The average student IQ in the school is the same as the average IQ score among all the schools in the country $\rightarrow \mu = 100$
   - Alternative Hypothesis: The average student IQ in the school is higher than the average IQ score among all the schools in the country $\rightarrow \mu > 100$

Input

```r
# run a one-sided t.test function with a significant level of 0.05
t.test(y, alternative = "greater", mu = 100, conf.level = 0.95)
```

Output

```
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

$p = 0.7215 > \alpha = 0.05$

Therefore, the evidence cannot reject null hypothesis and accept alternative hypothesis. We are unable to conclude that the average student IQ in the school is higher than the average IQ score among all the schools in the country.

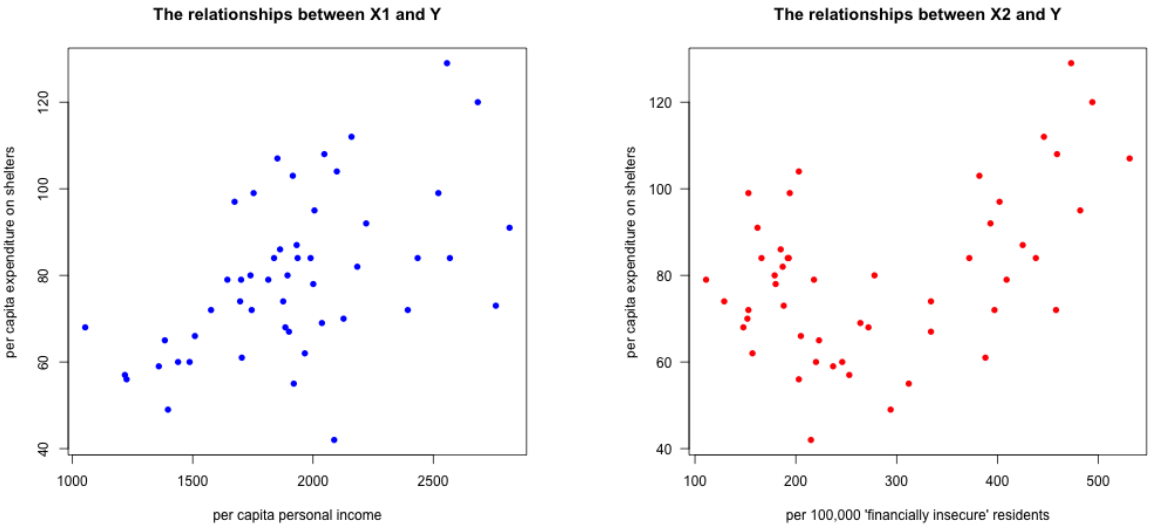
Question 2 (40 points): Political Economy

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*

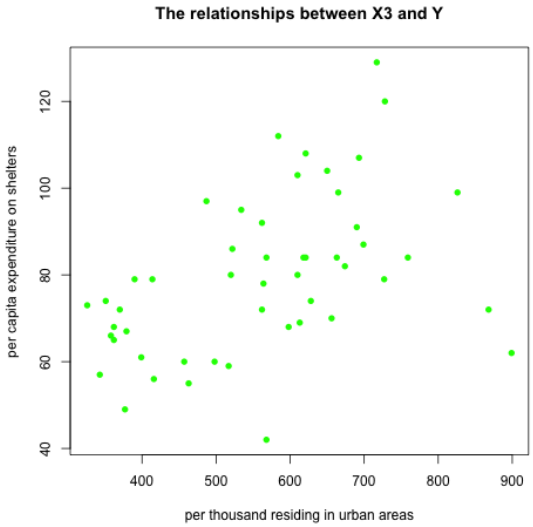
Explore the `expenditure` data set and import data into `R`.

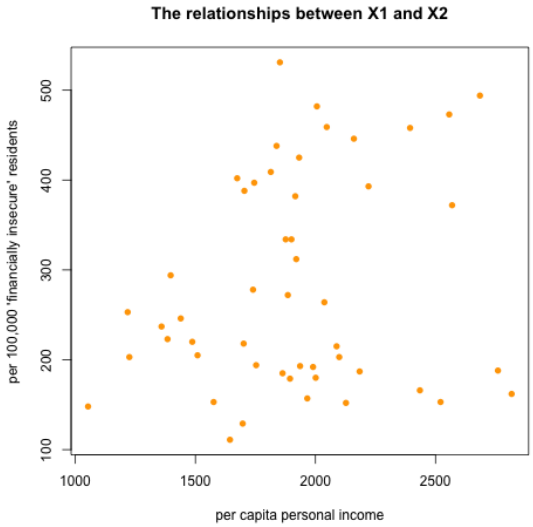
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?



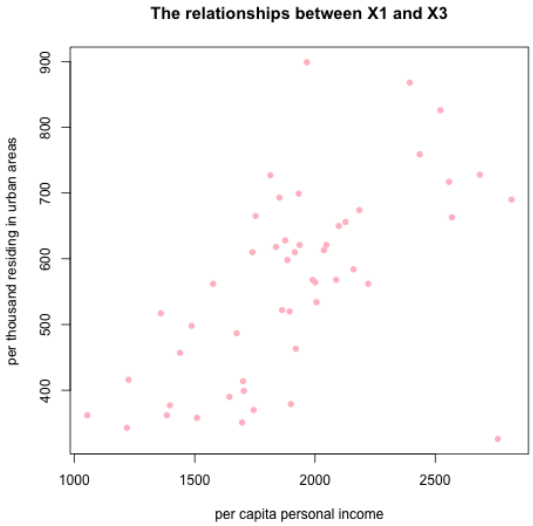
X1 and Y : From the graph, it can be observed that there is a positive correlation between per capita personal income and per capita expenditure on shelters: as per capita personal income increases, per capita expenditure on shelters also increases. Most of the data points are distributed within the range of 1600 to 2200 on the x-axis and 60 to 110 on the y-axis.
X2 and Y : There is no obvious linear correlation between 'per 100,000 financially insecure residents' and 'per capita expenditure on shelters'. However, the data shows

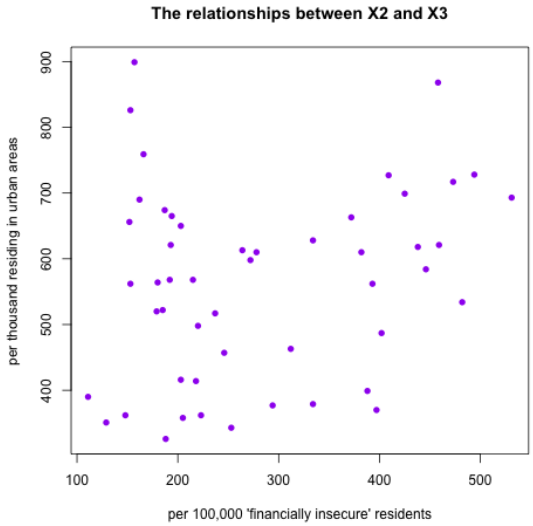
a downward trend followed by an upward trend, with a turning point around 300 on the x-axis. There might be a nonlinear correlation that needs further investigation.





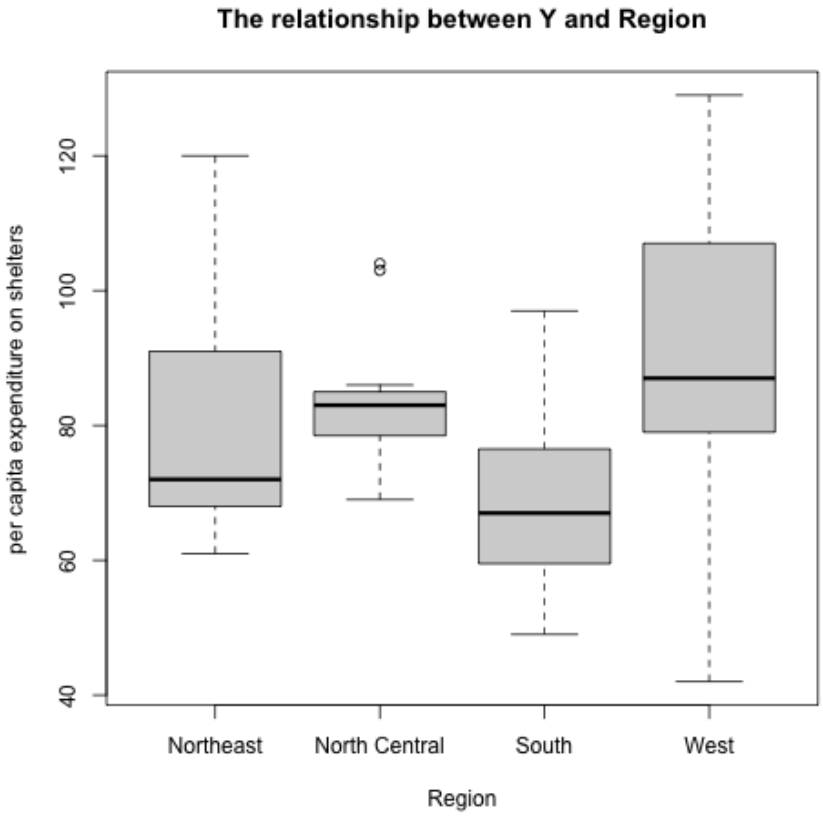
X3 and Y : There is a positive correlation between "per thousand residing in urban areas" and "per capita expenditure on shelters". Most of the data points are distributed in the range of 350 to 700 on the x-axis and 60 to 110 on the y-axis. One outlier is located around 570 on the x-axis, and the other two around 900 on the x-axis.
X1 and X2 : There is no correlation between per capita personal income and per 100,000 financially insecure residents.





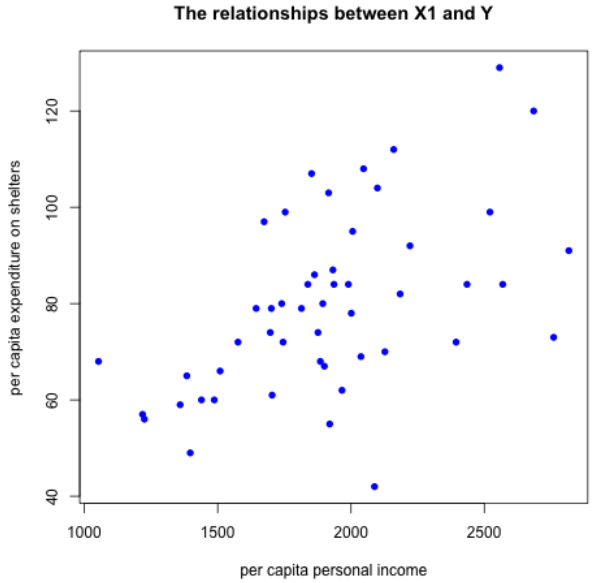
X1 and X3 : There is a positive correlation between per capita personal income and per thousand residing in urban areas. The majority of data points are distributed in the range of 1500 to 2200 on the horizontal axis and 350 to 700 on the vertical axis. There are two outliers, one at 2000 on the horizontal axis and 900 on the vertical axis, and another at 2700 on the horizontal axis and 310 on the vertical axis. These outliers might require further handling.
X2 and X3 : There is no correlation between per thousand residing in urban areas and per 100,000 financially insecure residents.

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



From this boxplot, the region of *West* 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.



X1 and Y : In general, the correlation between per capita personal income and per capita expenditure on shelters is positive. However, However, after the income surpasses 2000, with the increase in x, the central tendency begins to weaken and the data becomes more discrete.