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
dataset_edu<-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.
   The 90% confidence interval: [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 $\alpha = 0.05$.

H1: higher than average

H0: lower than or equal to average

p-value = 0.7215

0.7215 is a higher value than 0.05, therefore we reject our alternative hypothesis (alternative hypothesis= the average student IQ in her school is higher than the average IQ score among all the schools in the country.)

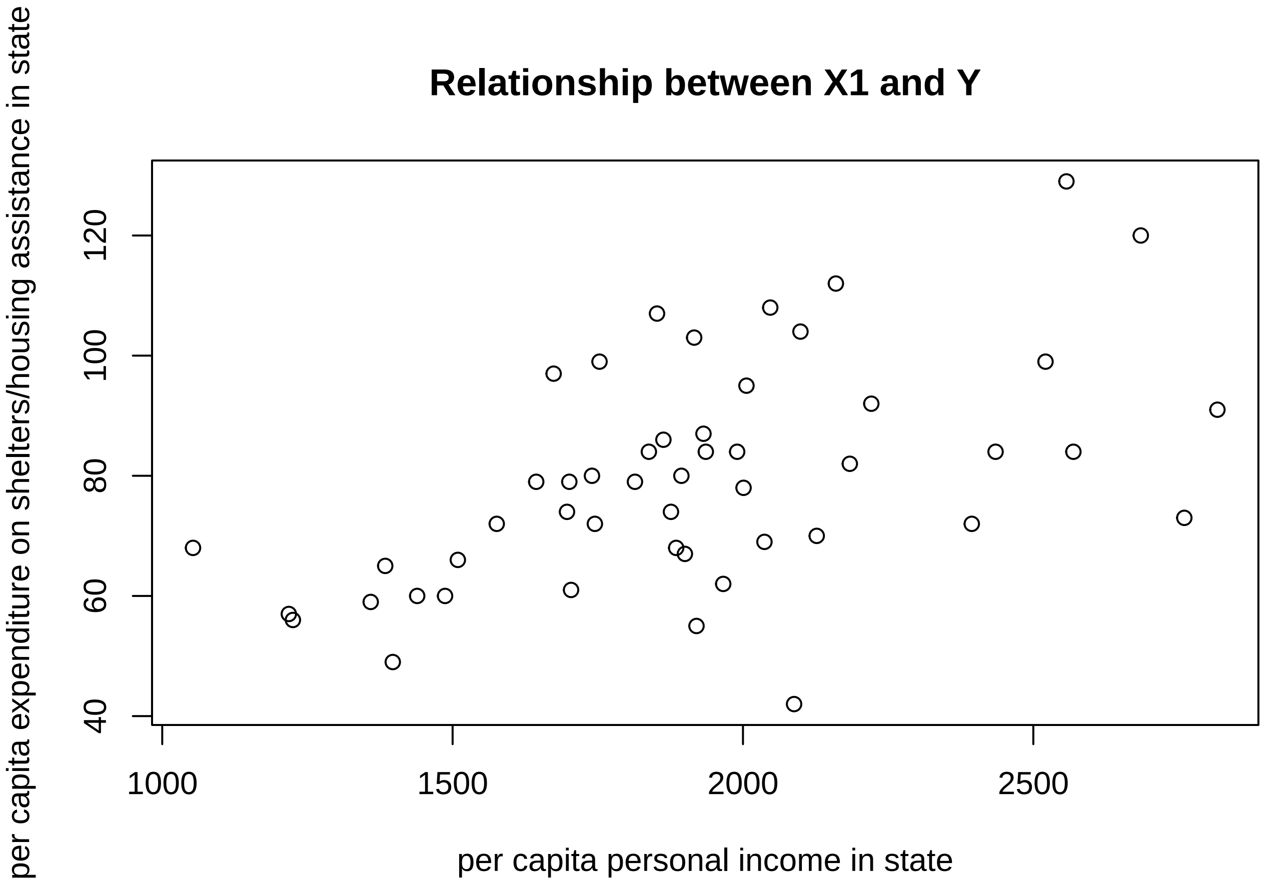
Question 2: Political Economy

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.

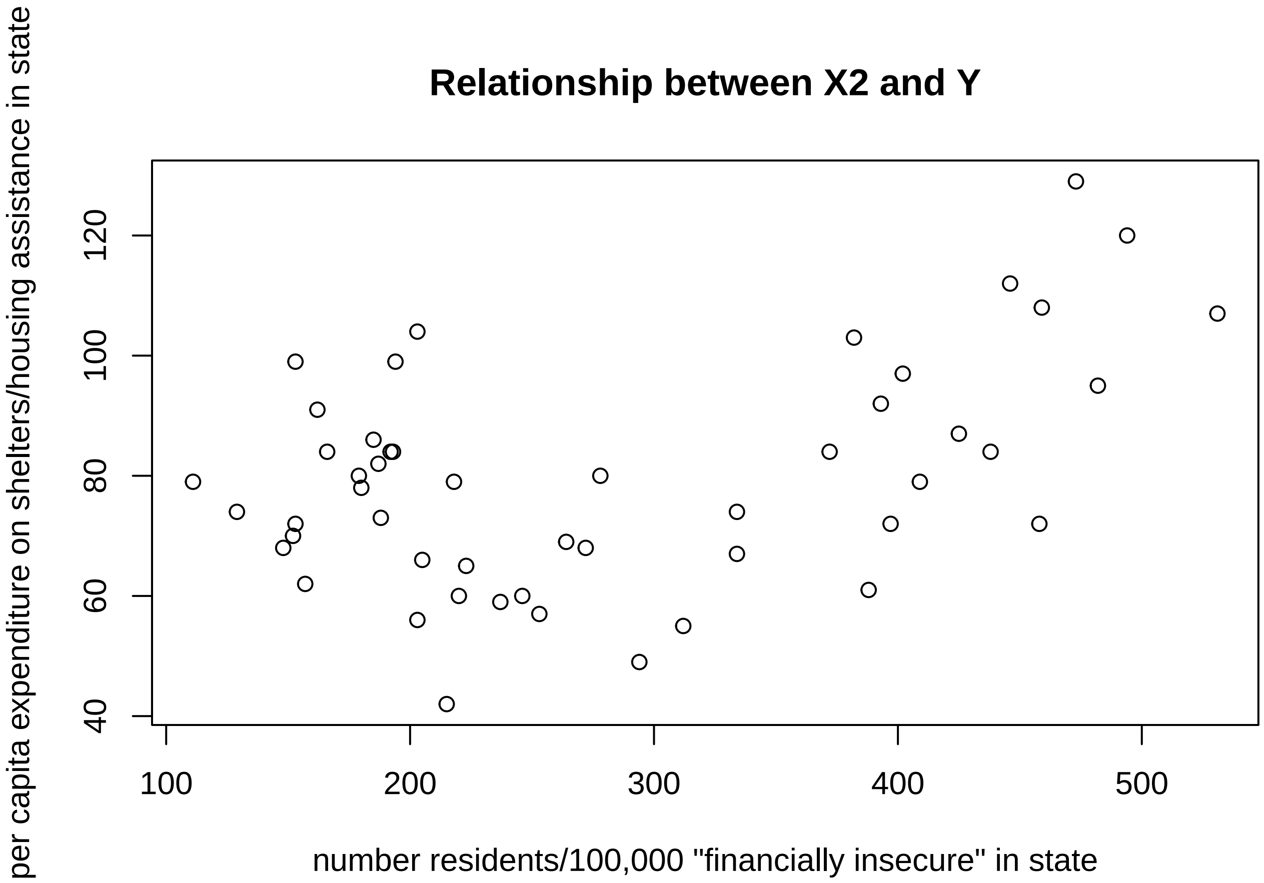
`State`	*50 states in US*
`Y`	*per capita expenditure on shelters/housing assistance in state*
`X1`	*per capita personal income in state*
`X2`	*Number of residents per 100,000 that are ”financially insecure” in state*
`X3`	*Number of people per thousand residing in urban areas in state*
`Region`	*1=Northeast, 2= North Central, 3= South, 4=West*

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

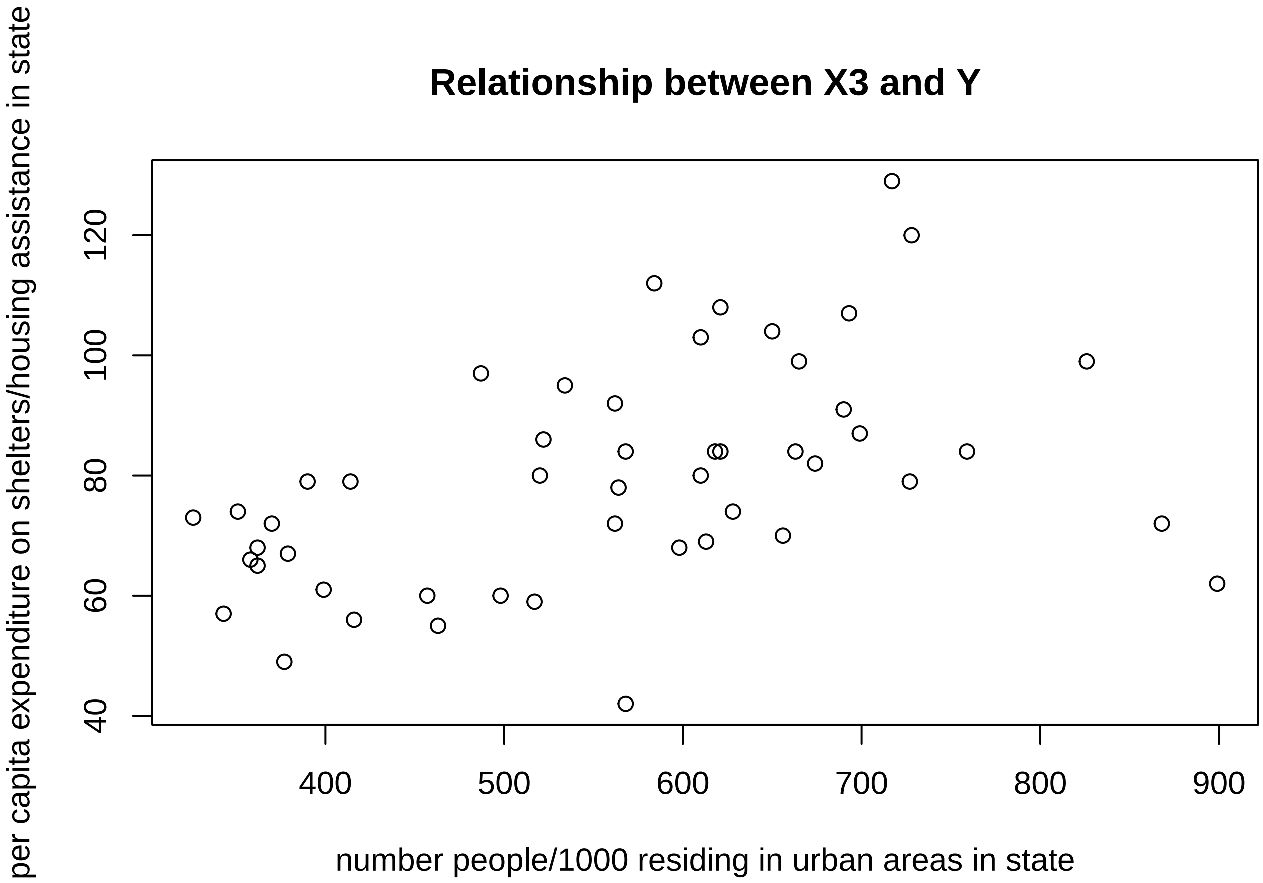
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?



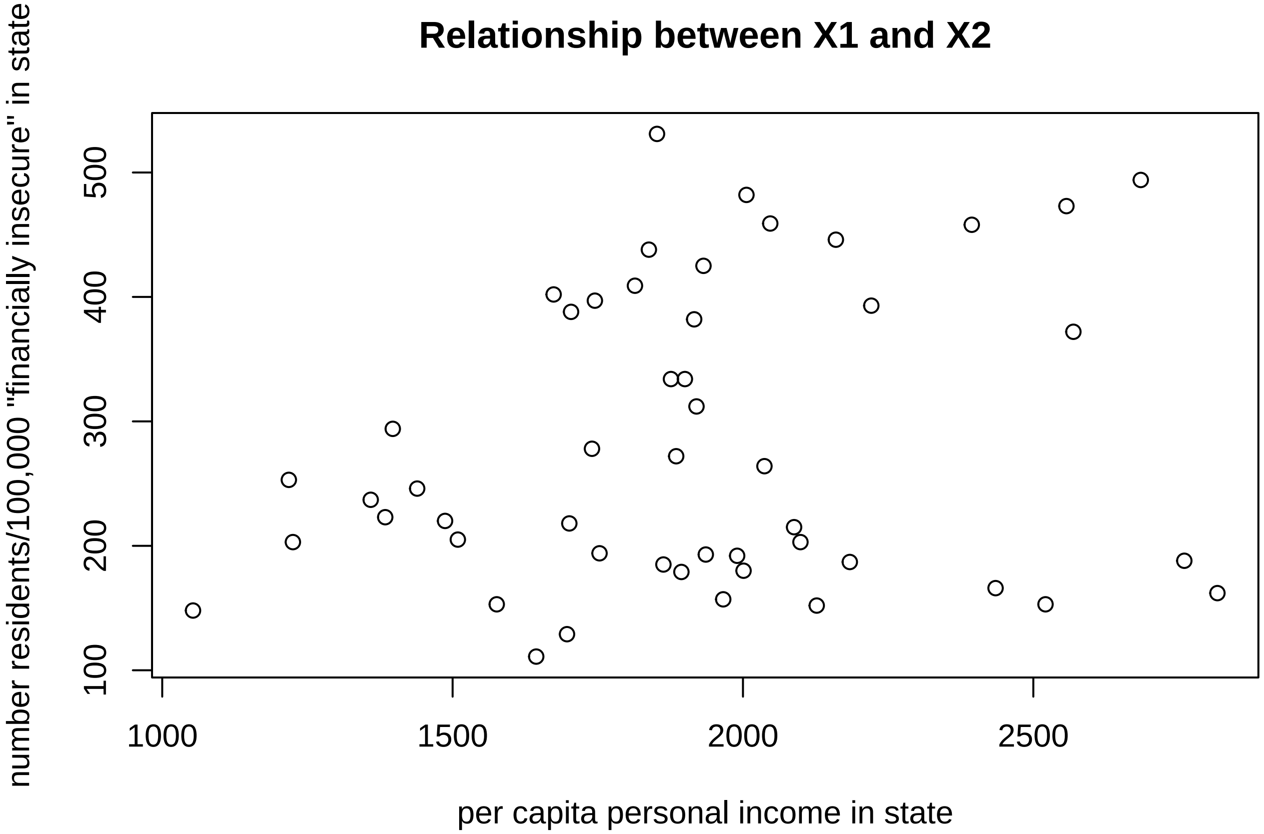
visible linear correlation (cor = 0.53)



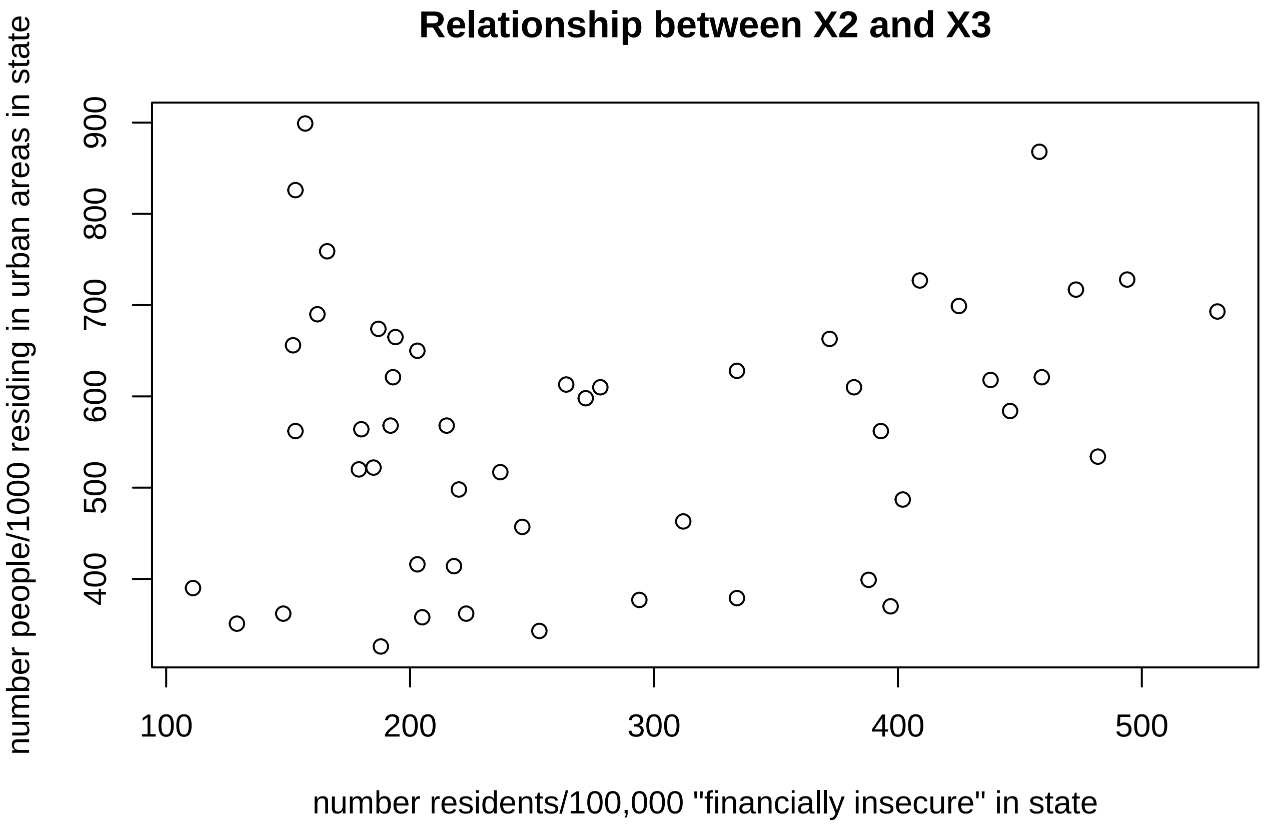
also slight linear correlation (cor = 0.45), however graph shows almost a U-shape along the lower x-axis (X2 = 100-200)



slight linear correlation (cor =0.46), not many obs for higher X3 values (X3=700-900), no remarkable differences in the skewing of variable on the y- or x-axis at any certain point in the graph

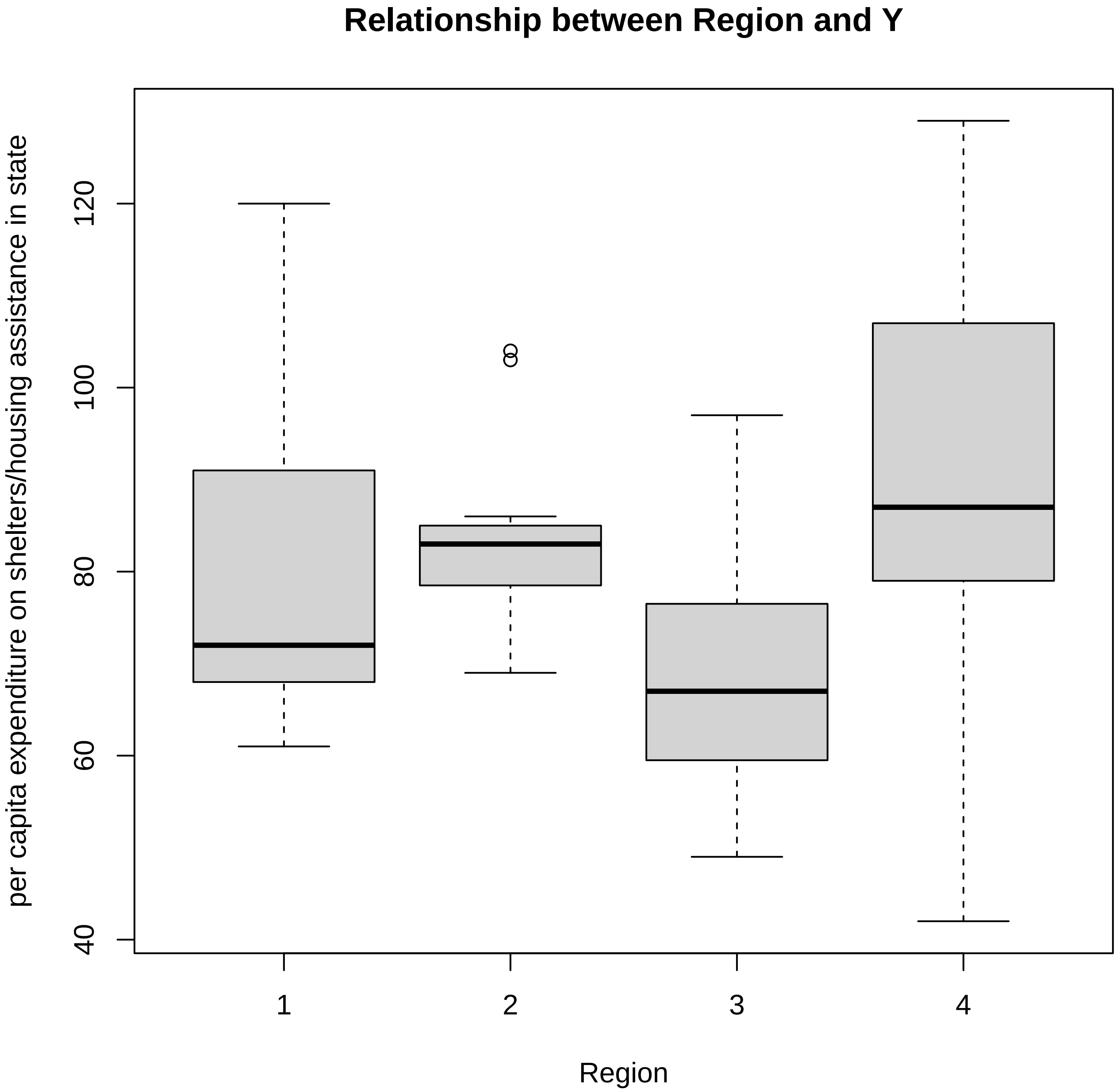


no clear linear correlation to see (cor =0.21), almost equal distribution on Y-axis at mid-point of x-axis (X1=2000), only extreme values (X2=100-200 or X2=400-500) on upper side of x-axis (X1=2500+)



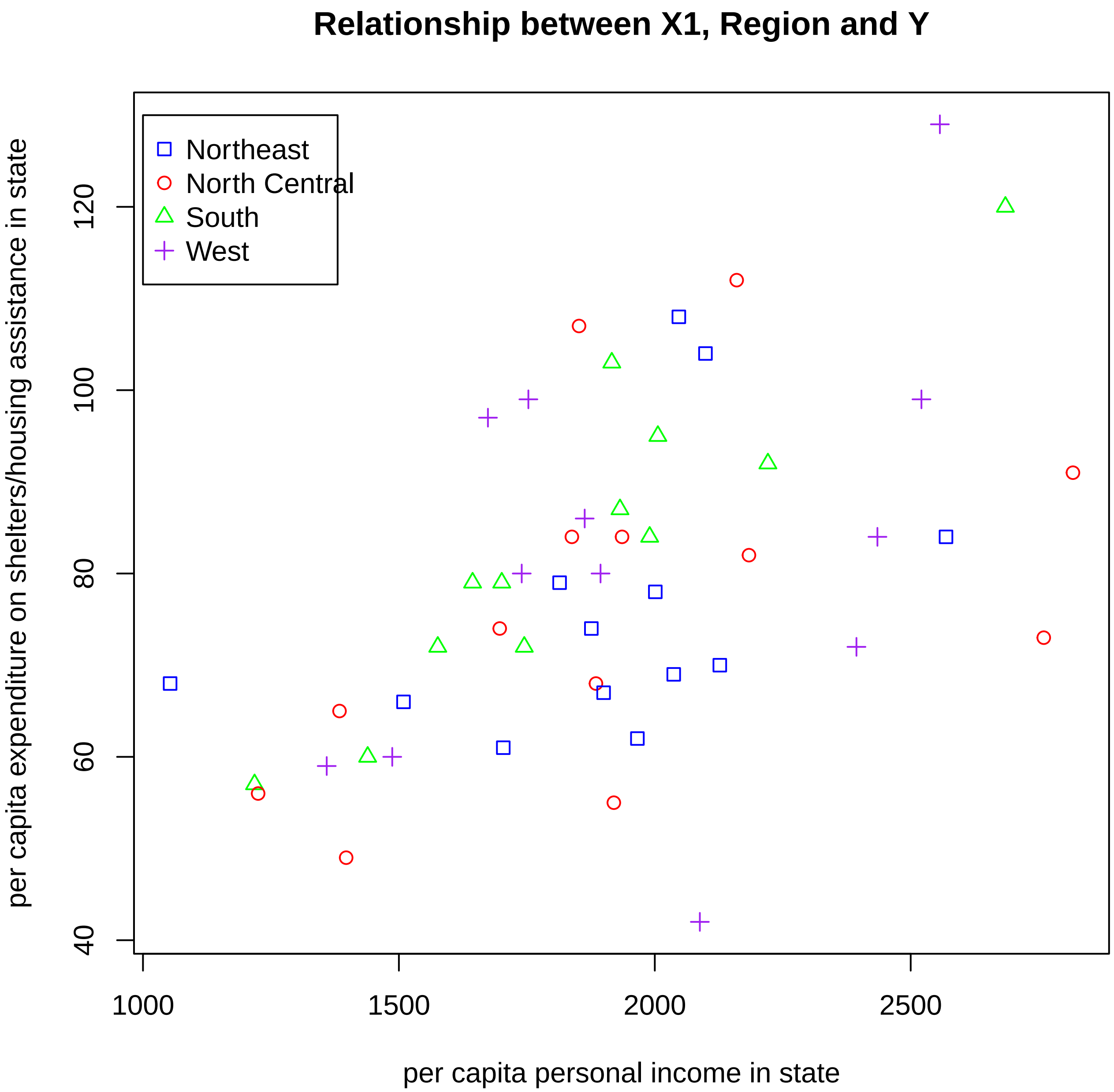
again no visible clear correlation (cor = 0.22), almost equal distribution of X2-values on y-axis at X2=200

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



Region 4 (West) seems to have the highest median (shown by the black line), followed by Region 2 (North Central). However the maximum values of Y can be found in Region 4 and Region 1 (Northeast), not Region 2.

- 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.


Relationship between X1, Region and Y
□ Northeast
○ North Central
△ South
+ West
per capita expenditure on shelters/housing assistance in state
40
60
80
100
120
1000
1500
2000
2500
per capita personal income in state