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:

1. Find a 90% confidence interval for the average student IQ in the school.

   Hello Jeff, here are my answers. I know this PDF may not look the best. However, it's not due to lack of effort I invested a serious amount of time and effort into this assignment and worked on it until almost the deadline. Thank you for reviewing it. To find the Confidence Interval I will use the formula mean +/- t-statistic * standard error. I will begin by showing how to produce each element in the fromula from the data

Step 1 Mean

Mean is the sum of all data points/number of values. In this case it is 98.44. One could also use the function mean() in R. This is also the point estimate.

Step 2 Standard Error

To find the standard error, we first need to calculate the variance and then the standard deviation.

To calculate the variance, we must find the sum of all squared differences (i.e., the difference of each data point from the mean squared and added together) and divide it by n-1. We use n-1 because this is a sample of less than 30. In this case, the variance is 171.4233. Alternatively, one could use the function var() in R.

To calculate the standard deviation, I took the square root of the variance. Again, you could use the function sd() in R. The standard deviation is 13.09287.

Now, we can calculate the standard error using the formula: standard deviation / sqrt(n) or 13.09287 / sqrt(25). The standard error is 2.618575.

Step 3 t-Statistic We use t statistic instead of z/normal because we have a small sample size (less than 30). First we must find degrees of freedom (n-1=24) We used .95 rather than .90 because to account for .05 on each side. The formula for t-statistic is hypothesized value – sample mean/standard error One could also use this function in R code

```
qt(.95,df=(length(y)-1))
```

Step 4 Confidence Interval The formula is mean plus or minus t-score * se (98.44 +/- 1.317 * 2.618)

A good way to find this in R is with the following lines of code where ”sey” signifies standard error of y.

```
lower_ci <- mean(y) - qt(.95, df = (length(y) - 1)) * sey
upper_ci <- mean(y) + qt(.95, df = (length(y) - 1)) * sey
```

Alternativley, one could use this line of code, which gives the confidence interval as well as other information about the dataset.

```
t.test(y, conf.level = 0.90, alternative = "two.sided")
```

The confidence intervals are 93.95(lower) and 120.84 (upper)

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 $\alpha = 0.05$.

   The first step is to look at data. I used the following code

   ```
   View(y)
   str(y)
   ```

   Step 2 Form the Null and Alternate Hypothesis The Alternate Hypotheses is: Students from school y will on average have IQ scores ¿ to the average IQ score of students from all other schools in the county. The Null Hypotheses is: Students from school y will on average have IQ scores less than or = to the average IQ score of students from all other schools in the county.

   Step 3 Find the critical value (again, we use t not z because we have a sample less than 30) and t-value. We can use the same mean (98.44) and standard deviation (13.09287). We still need to re-calculate the t-stat because we are using 95 percent confidence. (alpha-1).

   We can also use this code in R.

   ```
   qt(.975,df=(length(y)-1))
   ```

   We used .975 in the code above because a=.05. The accounts for .025 on each side. We used .975 because a=.05 The critical value is 2.063899

   The t-value is the mean of sample-population mean/standard error. In this code sey is the standard error of y.

   ```
   (mean(y)-100)/sey
   ```

   The t-value is -0.5957439 By comparing the critical value to the t-value, we can see that the t-value is smaller. This indicates that we cannot reject the null hypothesis. We can also draw a conclusion by examining the p-value. Since this is a one-tailed test, where we only want to determine if the mean IQ of students at school Y is greater than the population mean, we focus solely on the upper tail.

   Step 4 P Value. The P Value is used to determine the level of supprise. The smaller p-value signifies a higher level of significance. The p-value in this case is .73 Since this value is larger than .05 we fail to reject the Null Hypothesis.

   Step 5
   Conclusion: We fail to reject the null hypothesis that students at school Y have an average IQ less than or equal to the average IQ of students in the county.

   Therefore, we cannot support the alternative hypothesis that students at school Y have an average IQ greater than the average IQ of students from across the country.

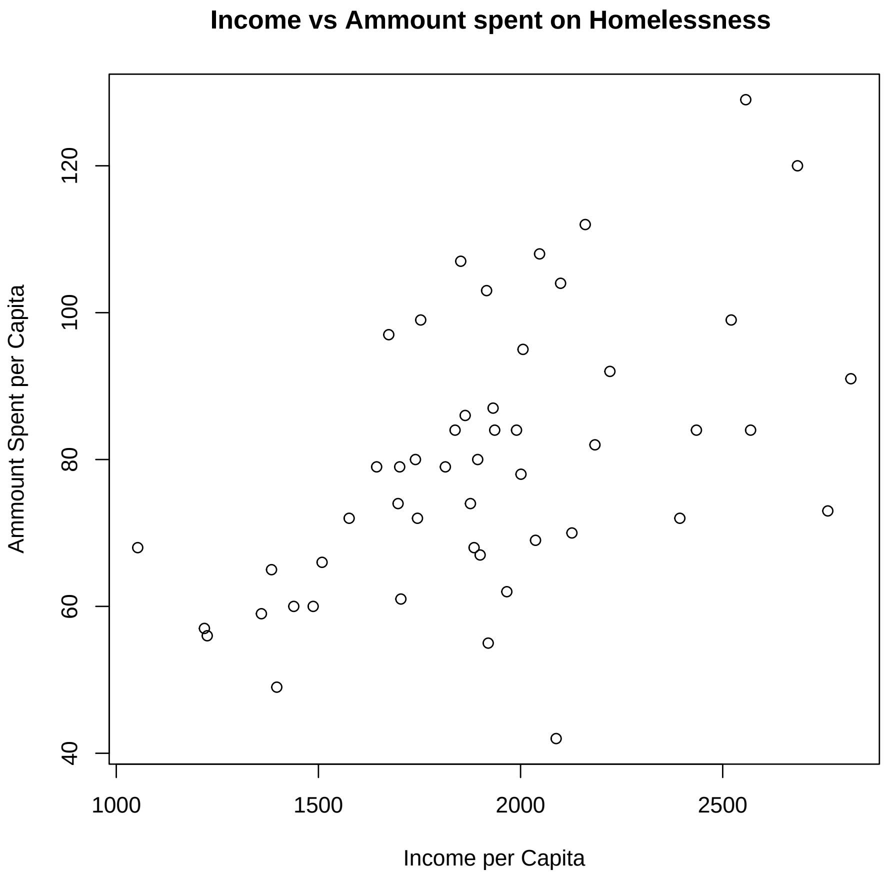
Question 2 (40 points): Political Economy

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.

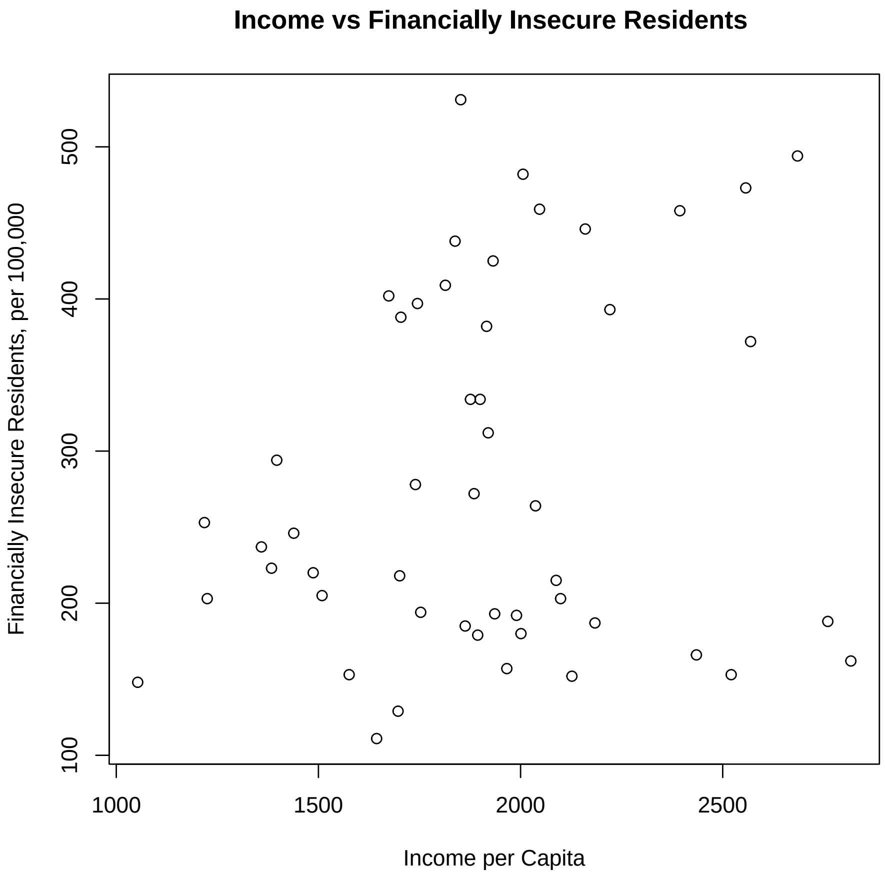
`State`	*50 states in US*
`Y`	*per capita expenditure on shelters/housing assistance in state*
`X1`	*per capita personal income in state*
`X2`	*Number of residents per 100,000 that are "financially insecure" in state*
`X3`	*Number of people per thousand residing in urban areas in state*
`Region`	*1=Northeast,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

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

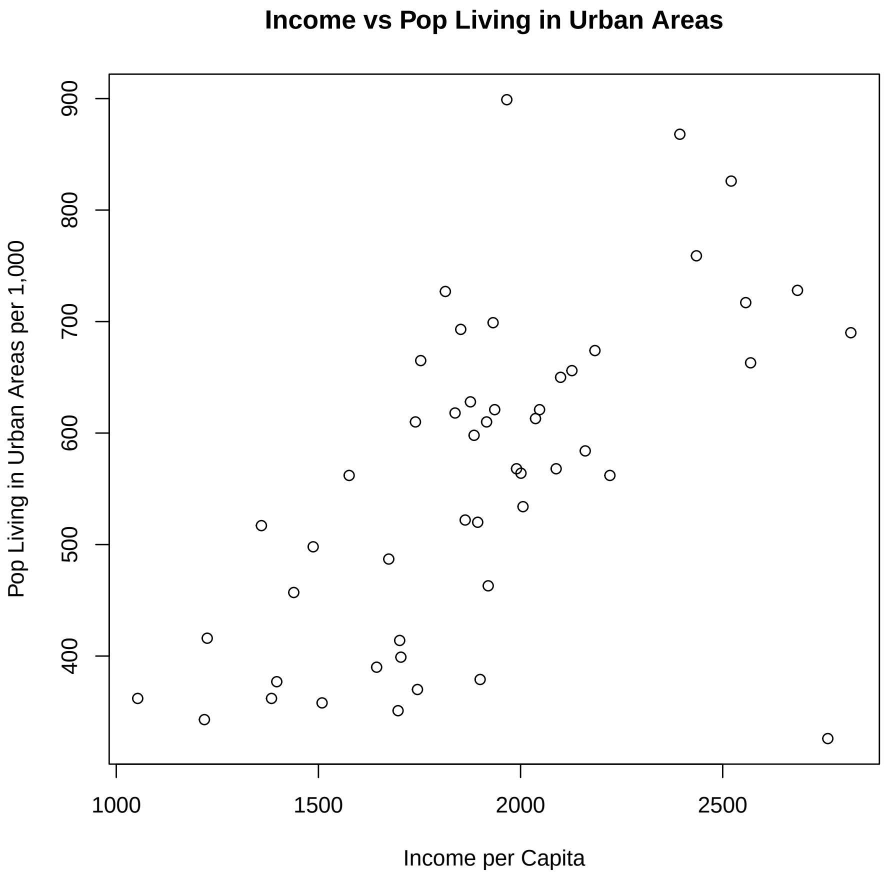
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?



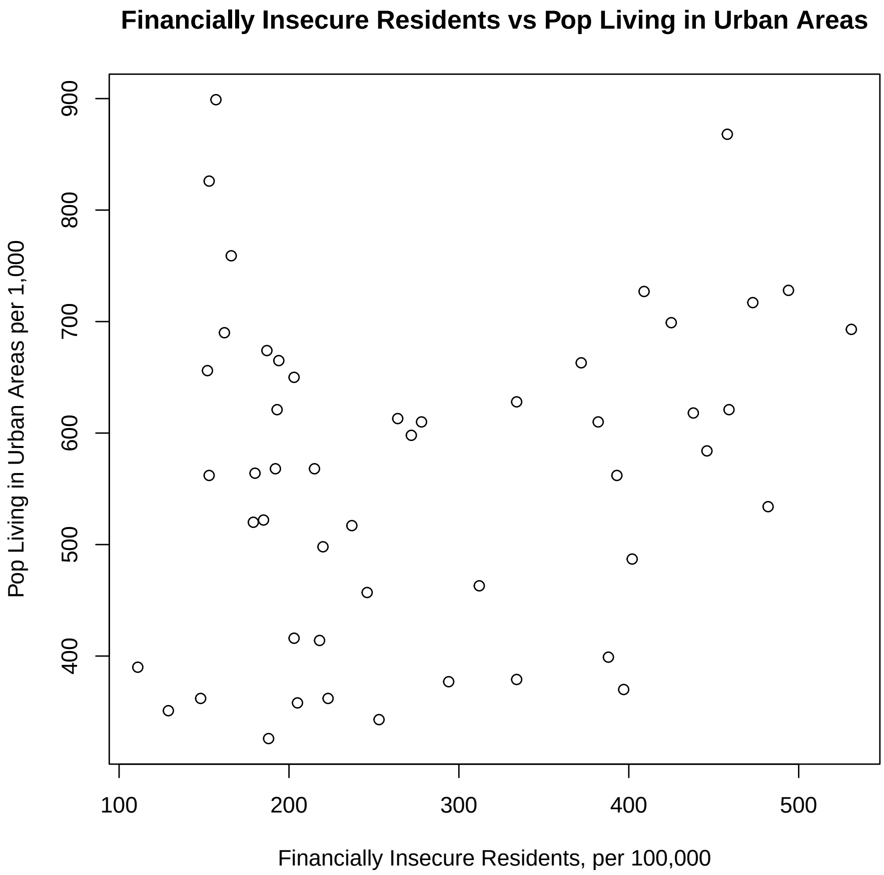
The relationship between X1 and Y is positive.



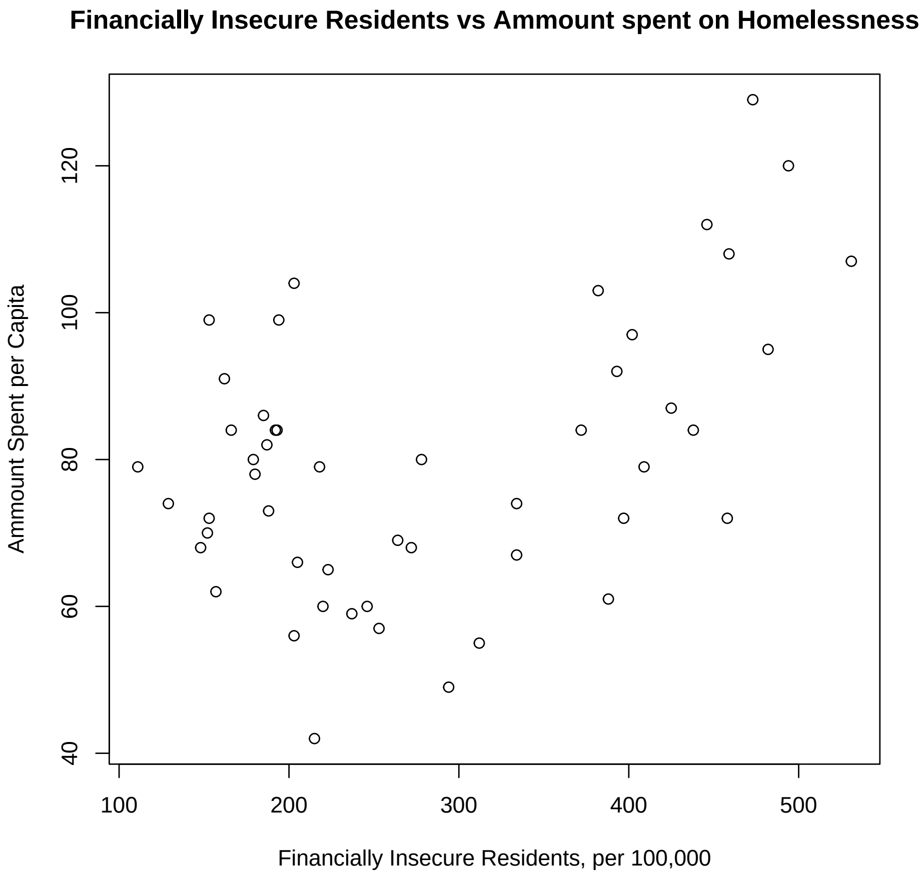
There is no obvious relationship between X1 and X2.



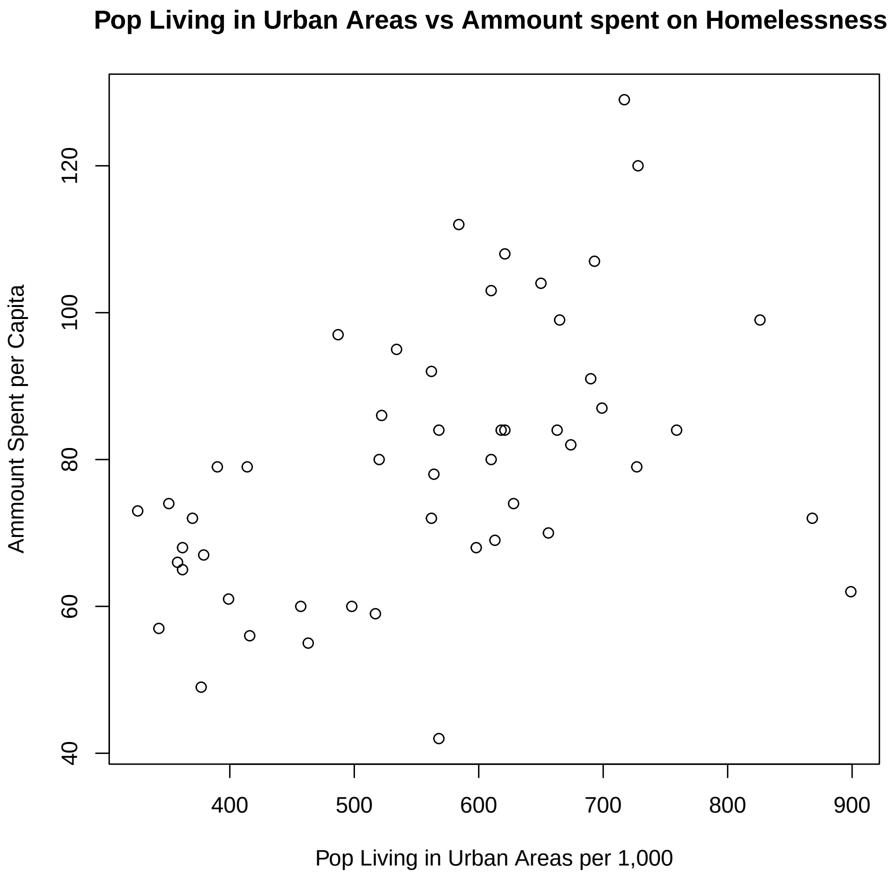
The relationship between X1 and X3 is positive.



There is no obvious relationship between X2 and X3.

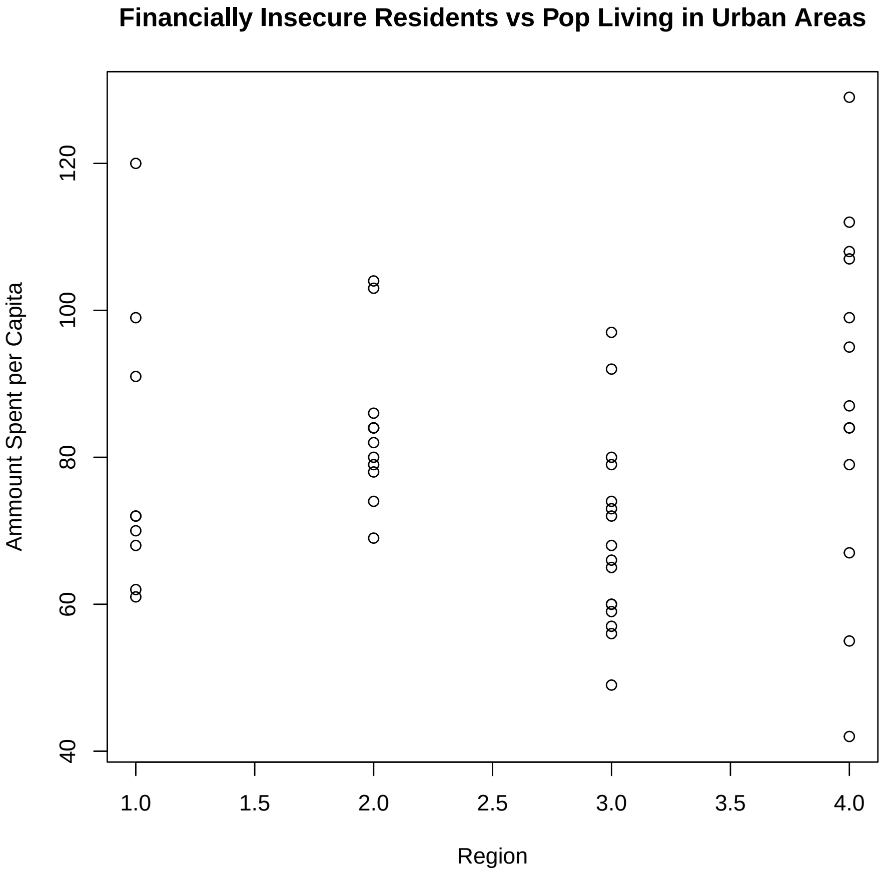


There is no obvious relationship between X2 and XY.



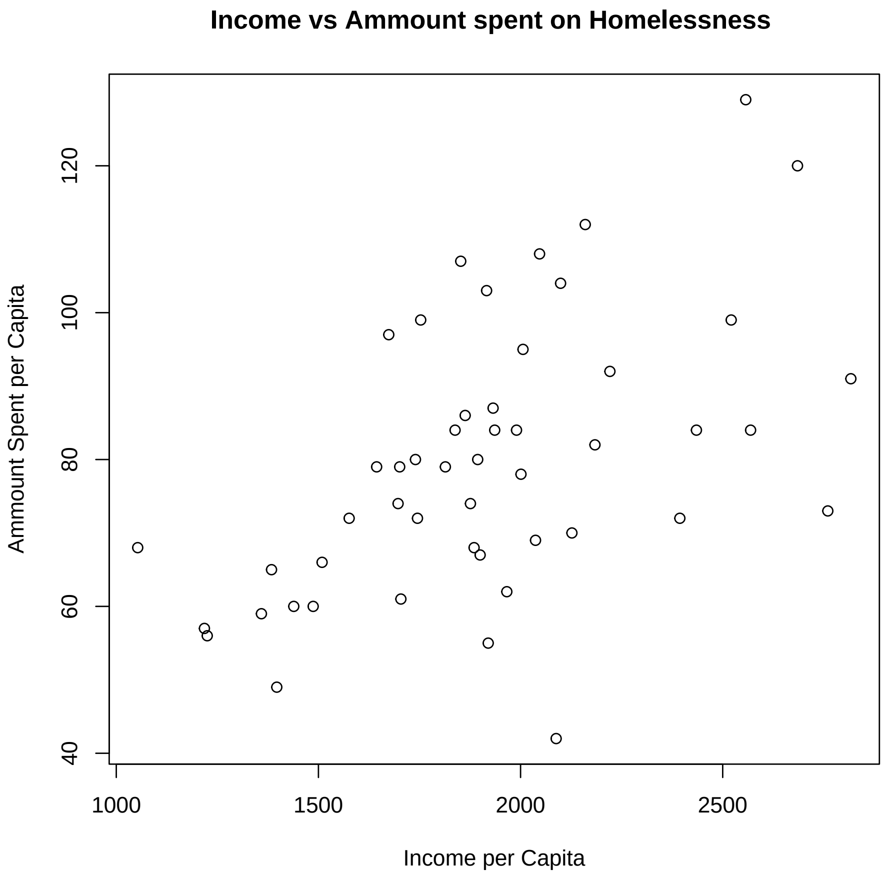
There is a positive relationship between X3 and Y.

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

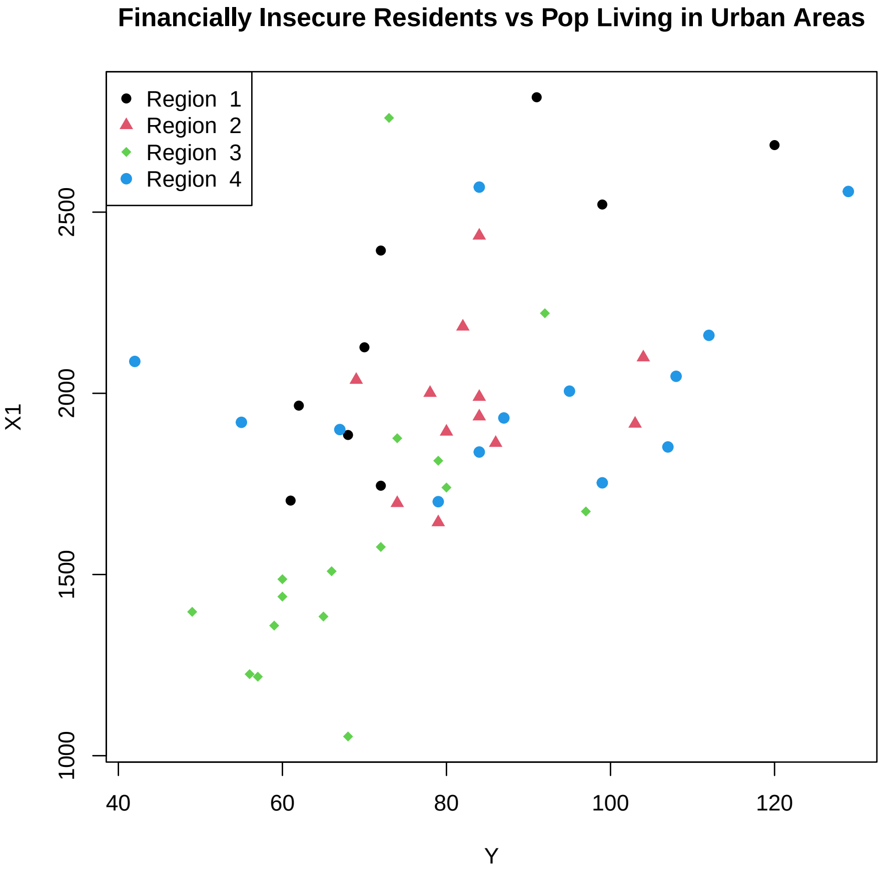


On average Region 4 has the highest per capita expenditure on housing assistance.

- 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.



There is a positive relationship between X1 and Y.


Financially Insecure Residents vs Pop Living in Urban Areas
Region 1
Region 2
Region 3
Region 4
X1
1000
1500
2000
2500
Y
40
60
80
100
120