Problem Set 1 - Answers

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 (data set provided):

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

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 $\alpha = 0.05$.

Answers to Question 1:

Confidence Interval

- First, to calculate the confidence interval, since I do not know the SD for the population, and since n = 25 (while our threshold is 30), I must do a T test, rather than a Z test. I learned from this source that I can thus use qt instead of qnorm.
- I also know that my T statistic will be = (sample mean - pop mean) / (s/sqrt(n)) , and
- It will be between the following critical values: - T(n-1, alpha/2) ¡= T stat ¡= T(n-1, alpha/2)
- Solving for pop mean in my T statistic, the CI will have the following lower and upper bounds:
- Sample mean - T(n-1, alpha/2)(s/sqrt(n)) , sample mean + T(n-1, alpha/2)(s/sqrt(n))

In R, I make all the necessary calculations, and I obtain:

- mean(y) = 98.44
- sd(y) = 13.09
- sqrt(n) = 5

And I know that:

- alpha = 1 - confidence = 0.1
- And that degrees of freedom = n -1 = 24

After inputting alpha/2 and n-1 degrees of freedom into qt, I obtain a **critical value of 1.710882**.

Multiplying my critical value by s/sqrt(n), I obtain **4.480072**, which I must add and substract to the sample mean to get my upper and lower bounds.

This substraction and addition results in a **CI that goes from 93.95993 to 102.92007**.

Finally, I confirm this result with the function t.test() and the resulting 90 percent confidence interval is identical. I also wish to see what happens if I use qnorm instead of qt in this case, and the resulting CI is slightly different.

Interpretation: Now I can tell that, if multiple samples were drawn from the same population (in this case, the entire population of students in the school) and a 95% CI were calculated for each sample, I would expect the population mean to be found within 95

Relevant R Code for Confidence Interval

```
meany <- mean(y) # I see there is a sample mean of 98.44
sdy <- sd(y) # I see there is a sample sd of 13.09
sqrtn <- sqrt(25) # I calculate the sqrt of the sample size, which is = 5
alpha = 1-0.9

t90 <- qt(alpha/2,24,lower.tail = F) # Critical value T(n-1, alpha/2)
t90

T90 <- t90*sdy/sqrtn # T90 * (s/sqrt(n)). This I add and substract from sample
    mean
T90

lower_90 <- meany - T90
upper_90 <- meany + T90

confint_90 <- c(lower_90, upper_90)
confint_90 # Mean of school falls somewhere between these values

# Confirming with t.test()

test2 <- t.test(y,conf.level = .90)
test2 # The IC is identical

# I would also like to see what happens if I use qnorm instead of qt for a n =
    25

z90 <- qnorm(alpha/2,lower.tail = F)
lower_90_b <- meany - (z90 * (sdy/sqrtn))
upper_90_b <- meany + (z90 * (sdy/sqrtn))
confint90_b <- c(lower_90_b, upper_90_b)
confint90_b # It does change quite a bit using qnorm ...
```

Hypothesis Testing

Now I wish to conduct a hypothesis test to see if the school mean (the population mean) is greater than 100.

!! Assuming that, for the whole school, IQs are distributed normally, and that the sample was taken using randomization, and knowing that I am dealing with continuous data:

- I wish to test if the school mean is greater than 100, so I set up my Null (H0) and Alternative (Ha) Hypotheses accordingly:
- H0: mean school is lesser or equal to 100
- Ha: mean school is greater than 100

I must use T test, where

- alpha = 1 - Confidence = 0.05, and
- the degrees of freedom, like with the CI, are still 24.

If T is greater than T(n = 24, alpha = 0.05), I may reject H0 at the 95 confidence level.
And T = (mean(y) - 100) / (sd(y)/sqrt(25))

- Calculating the T statistic in R, using qt, I obtain -0.5957439
- And my critical value is 1.710882
- Since T is NOT greater than the critical value I may NOT reject H0 at the 95% confidence level

Alternatively, comparing my p-value with alpha:

- In R, using pt, I obtain a p-value of 0.7215383.
- This is much larger than my alpha = 0.05; thus I may not reject H0.

Interpretation: After conducting a hypothesis test at the 95% confidence level to see if the school mean (population mean) is greater than 100, not enough evidence was found to reject the null hypothesis that the mean is smaller or equal to 100; or to support the alternative hypothesis that the mean is greater than 100.

Relevant R Code for Hypothesis Testing

```r
# If T > T(n-1, alpha), I may reject H0 at the 95 confidence level

alpha2 <- 0.05

t95 <- qt(alpha2,24,lower.tail = F) # I am looking in the upper tail
t95 # Critical value of -1.71 (if T => -1.71, I may reject H0)

T95 <- (meany - 100)/(sdy/sqrtn)
T95 # T is NOT greater than -1.71, so I may not reject the H0 that mean school
    <= 100 at the 95 confidence level

# Calculating the P Value

p_val <- pt(abs(T95),df = 24, lower.tail = T)
p_val

# To confirm with t.test():

test <- t.test(y, mu = 100, alternative = "greater")
test
```

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)?

- Please plot the relationship between *Y* and *Region*? On average, which region 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.

Answers to Question 2:

Part 1

- When plotting X1 against Y, that is per capita personal income against per capita expenditure on shelters/housing assistance, a pattern does emerge; the data are a little concentrated along a line with a positive slope, showing a positive correlation between the two variables. The association may not be very strong, as the dots are quite dispersed with some outliers, but it is nonetheless visible.

- When plotting X2 against Y, that is the number of residents per 100,000 who are financially insecure, against per capita expenditure on shelters/housing assistance, the two variables do not seem to be associated linearly. Rather, from looking at the plot, it could be that, before X2 = 300, the number of residents who are financially insecure is negatively associated with per capita expenditure on shelters/housing assistance, but once the 300 threshold is passed on X2, the relationship between the number of financially insecure people and expenditure seems to become positive. This association would perhaps be best represented by a quadratic equation, and there does not seem to be a linear correlation between the two.

- When plotting X3 against Y, that is the number of people per 1000 who reside in urban areas, against per capita expenditure on shelters/housing assistance, there seems to be a positive correlation between the two variables. The pattern that emerges suggests the data are concentrated along a line with a positive slope, with a couple of considerable outliers in the lower-right corner.

- When plotting X1 against X2, that is per capita personal income against the number of residents per 100,000 who are financially insecure, the two variables seem to be unrelated in the scatterplot. No obvious pattern emerges; rather, the observations are dispersed all over the area of the plot.

- When plotting X1 against X3, that is per capita personal income against the number of residents per 1000 who live in urban areas, the two variables seem to be positively correlated, and the dots form a pattern that could adjust to a line with positive slope.

- Finally, when plotting X2 against X3, that is the number of residents per 100,000 who are financially insecure against the number of people per 1000 who live in urban areas, there seems to be no relationship between the two. The scatterplot reveals that observations are disperes all over the area of the plot.

Part 2

- When plotting the relationship between Y and Region, that is the per capita expenditure on shelters/housing assistance by region, the graph that emerges suggests that Region 4 (the West) has the highest per capita expenditure, on average, although from the plot it is also visible that this region has the largest variance in expenditure.
- Also, to confirm, we may use R to group the observations in the expenditure dataset by region, and then obtain the mean for each group. This exercise shows that Region 4 has an average per capita expenditure of $ 88.30 on shelters/housing assistance, which is higher than the mean values for the other three regions.

Part 3

- The relationship between X1 and Y has already been described in Part 1 above, and when adding the different colors and dot shapes by region, we gain some new insights:
- For Regions 1 and 3, there does seem to be a positive correlation between X1 and Y, as both the red and blue dots seem to gather around lines with positive slopes.
- For Regions 2 and 4, however, this is a lot less clear, as the green and purple dots are dispersed around a large area of the plot, forming no evident pattern.

Relevant R Code for Question 2 (Plotting)

```
# Exercise 1. Initial plots

plot_X1Y <- plot(expenditure$X1,expenditure$Y,
                 xlab = "Per capita personal income in state",
                  ylab = "Number of people per thousand residing in urban
    areas in state")
plot_X1Y

plot_X2Y <- plot(expenditure$X2,expenditure$Y,
                 xlab = "Number of residents per 100,000 that are
       financially     insecure  in state",
                 ylab = "Per capita expenditure in shelters/housing assistance
     in state")
plot_X2Y

plot_X3Y <- plot(expenditure$X3,expenditure$Y,
                 xlab = "Per capita personal income in state",
                 ylab = "Number of people per thousand residing in urban areas
     in state")
plot_X3Y

plot_X1X2 <- plot(expenditure$X1,expenditure$X2,
                  xlab = "Per capita personal income in state",
                  ylab = "Number of residents per 100,000 that are
       financially     insecure  in state")
plot_X1X2

plot_X1X3 <- plot(expenditure$X1,expenditure$X3,
                  xlab = "Per capita personal income in state",
                  ylab = "Number of people per thousand residing in urban
    areas in state")
plot_X1X3

plot_X2X3 <- plot(expenditure$X2,expenditure$X3,
                  xlab = "Number of residents per 100,000 that are
       financially     insecure  in state",
                  ylab = "Number of people per thousand residing in urban
    areas in state")
plot_X2X3
# Exercise 2. Region plot #

# I think it makes sense to have Region as factor, otherwise R will do a dots
    plot for this:

expenditure$Region <- factor(expenditure$Region)
plot_Reg <- plot(expenditure$Region,expenditure$Y,
                 xlab = "Region",
                 ylab = "Per capita expenditure in shelters/housing assistance
     in state")
plot_Reg
```

```r

# Exercise Confirming

reg_highest_exp <- expenditure %>% group_by(Region) %>% summarise(MeanExp =
    mean(Y)) # Region 4 has the highest

# 3. X1 and Y, by Region #

plotX1Y_Reg <- ggplot(expenditure,aes(x = X1, y = Y, shape = Region, colour =
    Region)) +
  geom_point()+
  labs(x = "Per capita personal income in state",
       y = "Number of people per thousand residing in urban areas in state")+
  theme_minimal()
plotX1Y_Reg
```

Please note that I was not able to make the scatterplot do what I needed it to do with the colors and dot shapes per group, so I did it with ggplot. I already knew ggplot from before, but I did consult two sources to remember what exactly I needed to input to make the plot I needed. There are more details about this and these sources in my R script.

Plots that resulted from Question 2

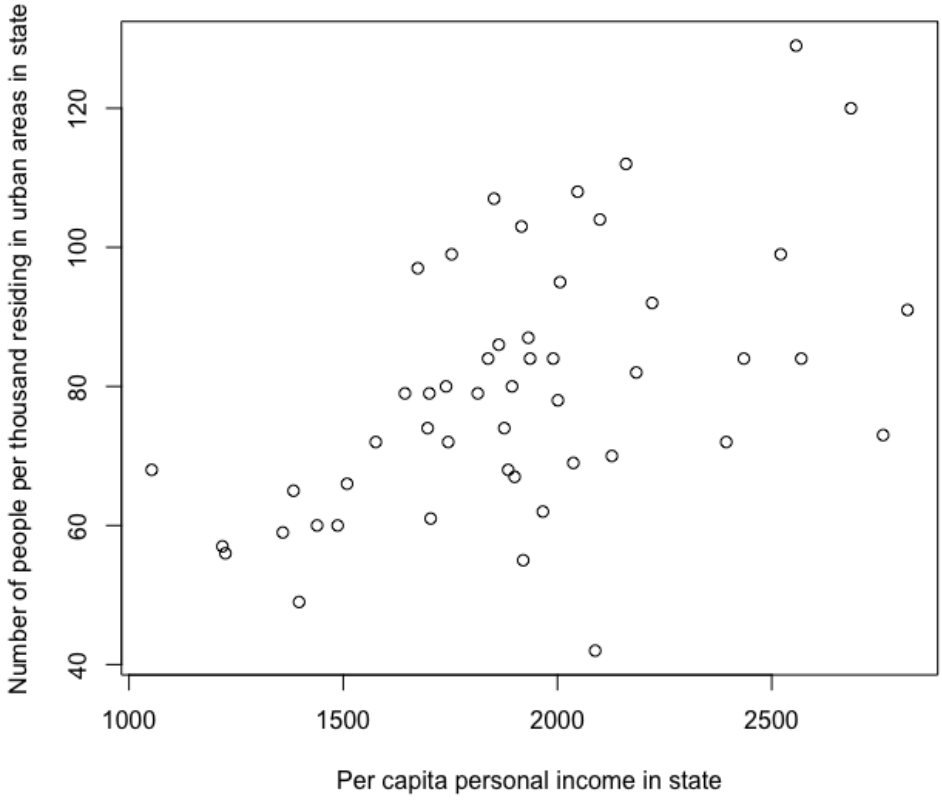

Figure 1: Plot1

Please also note that I consulted overleaf.com several times throughout this assignment to fix issues with LaTeX. For example, I learned here how to include URLs in the document. I also consulted overleaf.com because I was having trouble making the first of the plots above go where I wanted it to go; that is where I saw that including [htbp] next to "begin figure" makes it go where we mark it, since the default is something different!

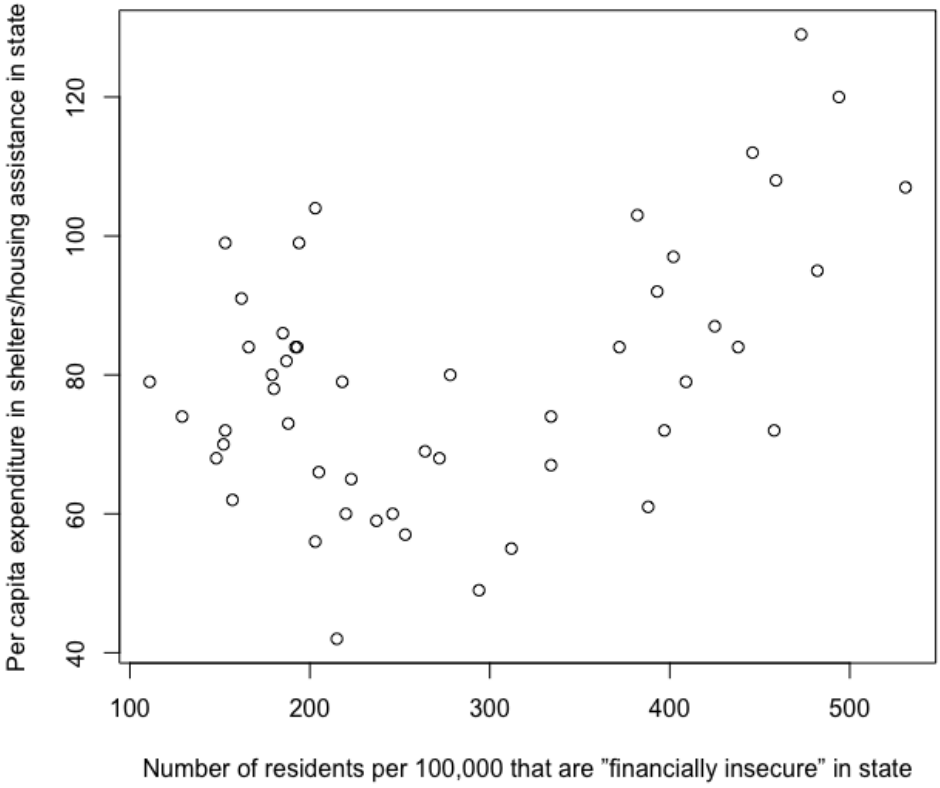

Figure 2: Plot2

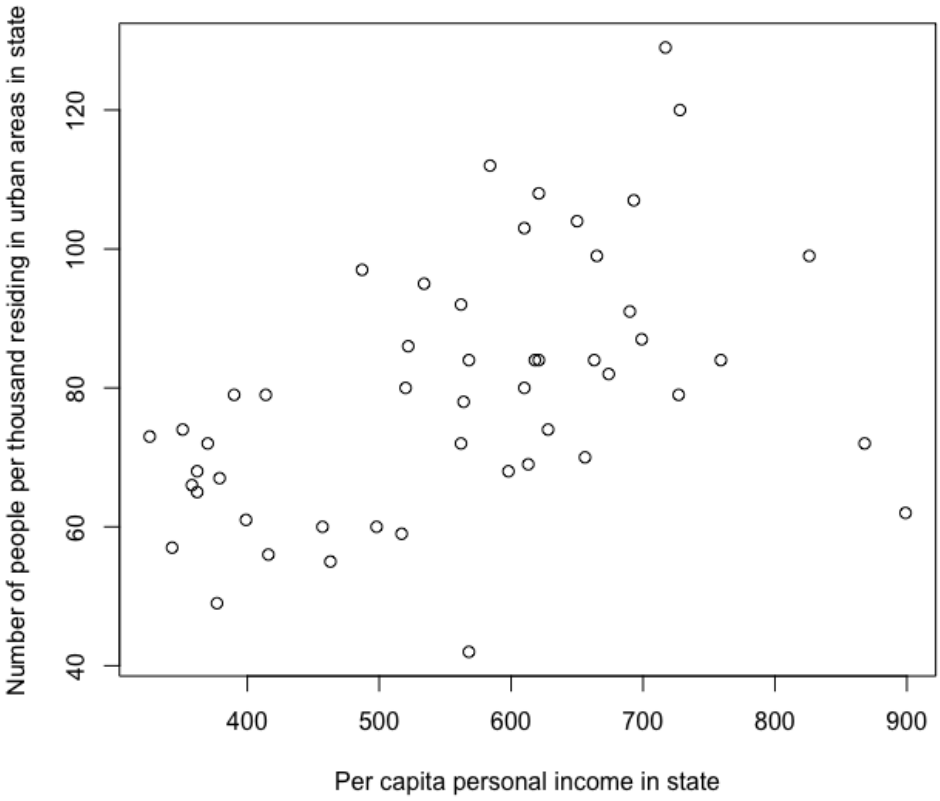

Figure 3: Plot3

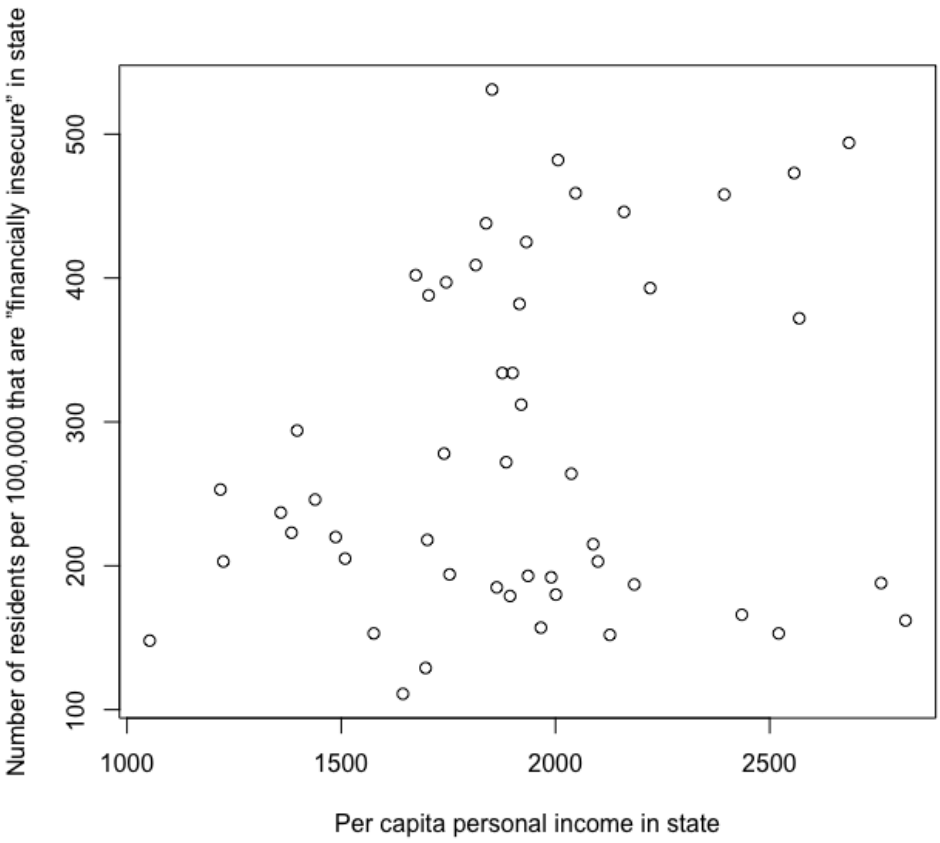

Figure 4: Plot4

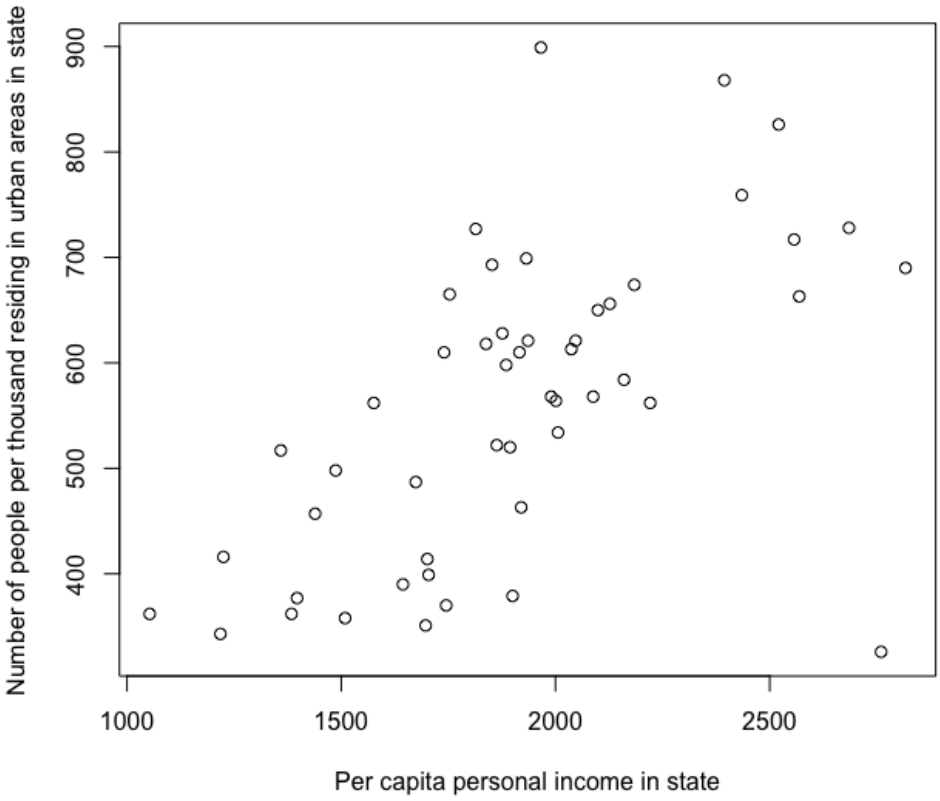

Figure 5: Plot5

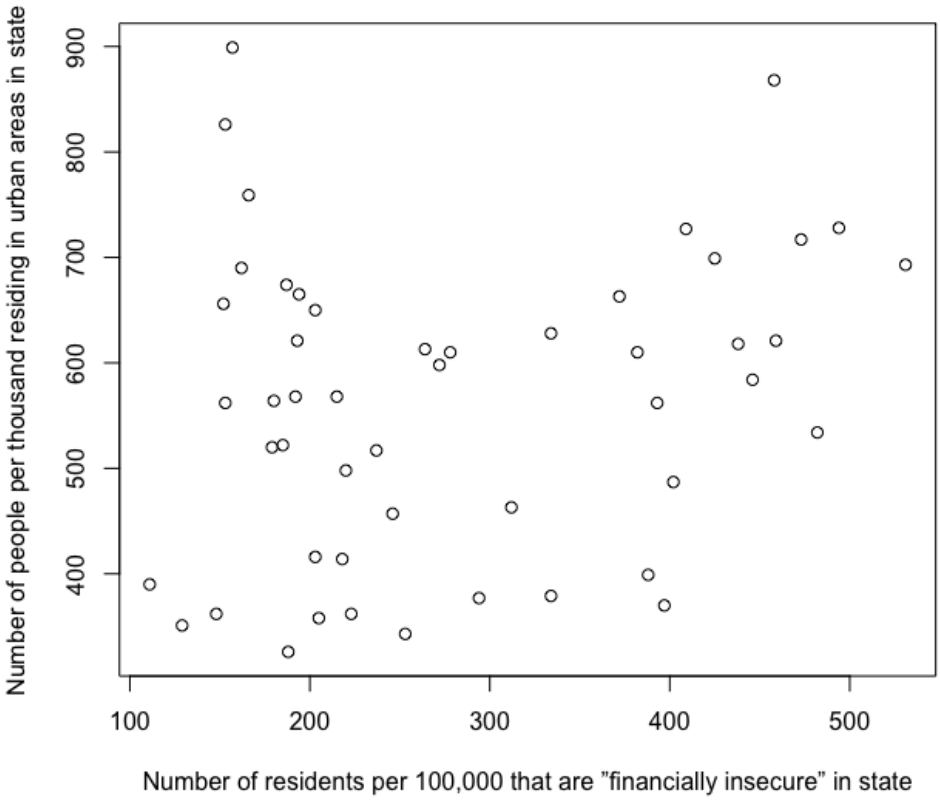

Figure 6: Plot6

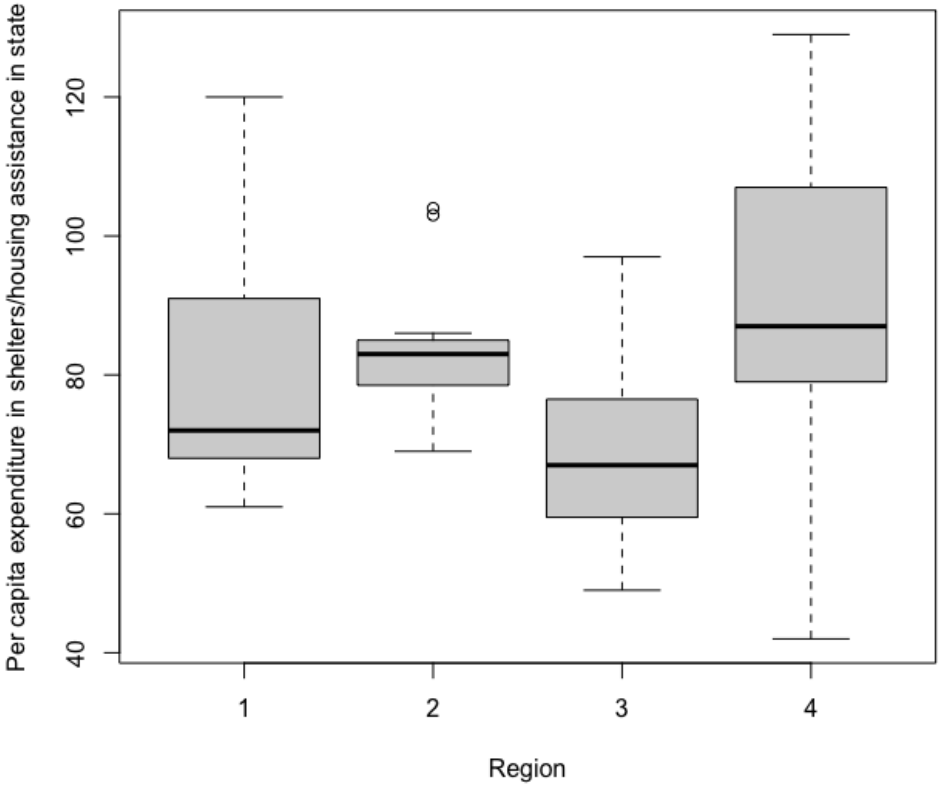

Figure 7: Plot7

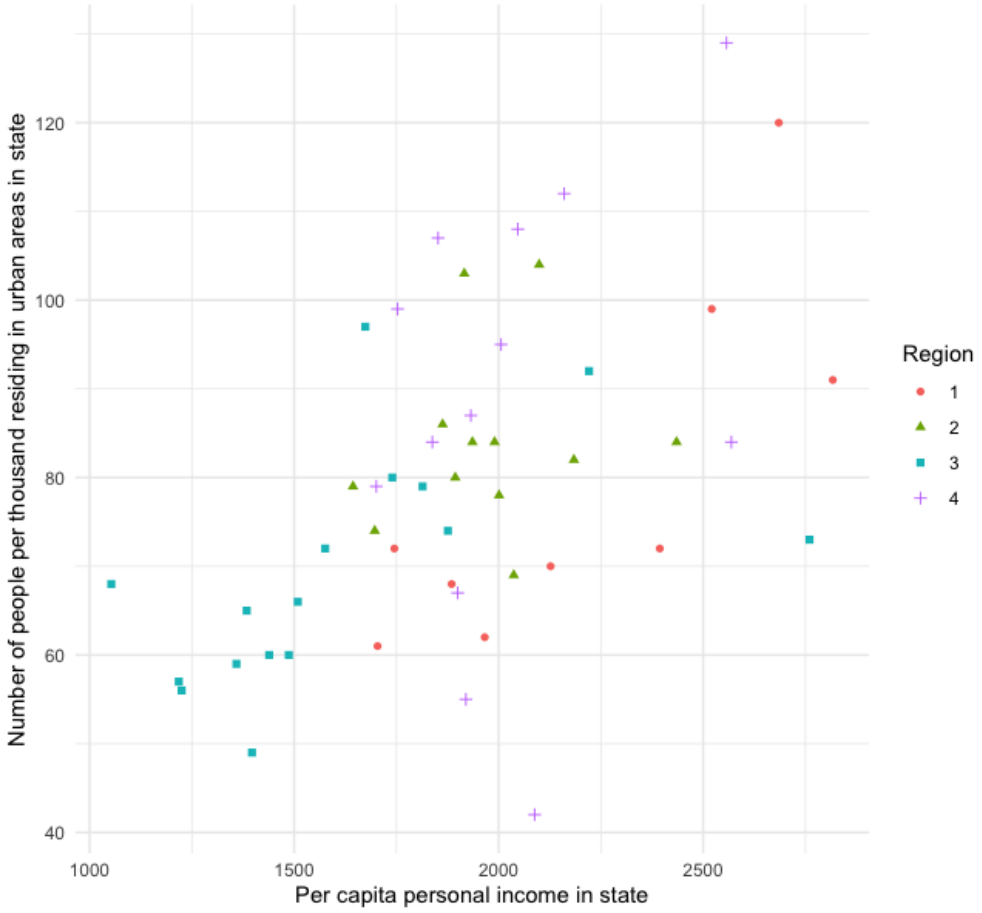

Figure 8: Plot8