Problem Set 1

Marcus Ó Faoláin, 16327268. Applied Stats/Quant Methods 1.

Due: October 1, 2021

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 in .pdf form.
- This problem set is due before 8:00 on Friday October 1, 2021. No late assignments will be accepted.
- Total available points for this homework is 100.

Question 1 (50 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.

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 $\alpha = 0.05$.

Question 1 Part 1:

Since n is less than 30, we must use a t-score instead of a Z score. As the confidence coefficient is 0.90, we are looking for a t-value corresponding with (1-0.90)/2 = 0.05 using one-tailed test scores or 0.1 for two-tailed test scores.

Since degrees of freedom = (n-1), we calculate the degrees of freedom = (25-1) df = 24. The T-value at these points is 1.711

We calculate the sample mean to be ¯x = 98.44.

```r
mean1 <- mean(y)
```

We create an empty vector to hold the demeaned sum of yn.

```r
demeaned_sum_y <- NULL
```

We add the demeaned sum of each element of y to the demeaned sum y vector. This is yn minus the mean of y (98.44).

```r
for(i in 1: length(y)){
  demeaned_sum_y[i] <- y[i] - mean(y)
}
```

We calculate the Squared Error for each element of demeaned sum y by squaring each element.

```r
squaredError_y <- demeaned_sum_y^2
```

We then find the sum of the elements of the vector. Sum = 4114.16.

```r
sum(squaredError_y)
```

We find the variance of y by dividing the sum of the squared errors of y by the number of elements minus 1. Variance = 171.4233.

```r
variance_y <- sum(squaredError_y)/(length(y)-1)
```

We find the standard deviation by finding the square root of the variance. SD = 13.09287.

```
sd1 <- sqrt(variance_y)
```

We calculate the standard error using by dividing the standard deviation by the square root of the number of elements in y. SE = 2.618575.

```
se1 <- sd1/sqrt(length(y))
```

We now multiply the standard error by the t-value we found earlier = 4.480381.

```
t_by_se1 <- 1.711*se1
```

We now find the lower limit and the upper limit for the confidence interval by subtracting t-by-se1 from the mean and by adding it to it.

```
lower_limit1 <- mean1 - t_by_se1
upper_limit1 <- mean1 + t_by_se1
confidence_interval1 <- c(lower_limit1, upper_limit1)
confidence_interval1
```

We calculate that the confidence interval is [93.95962, 102.92038]

Question 1 Part 2.

5 steps to Hypothesis testing

Step 1: Assumptions. Since n is greater than 30, we will use a t-test to test the hypotheses. According to the question, the data is a random sample

Because we want to know whether the mean of the average school student is higher than the average among all schools in the country, we will perform a one-tailed test.

Step 2: We set up our null and alternative hypotheses.

H0: µ is less than or equal to 100.
Null Hypothesis: The average IQ of the students in the school is less than or equal to 100

```
# H0 <- mu =< 100
```

Ha: µ is greater than 100.
Alternative Hypothesis: The average IQ of the students in the school is greater than 100.

```
# Ha <- mu > 100
```

Step 3: We calculate a test statistic.
Test statistic formula = (¯x-mu0)/(s/ the square root of n)

```
ts1 <- (mean(y)-100)/(sd(y)/sqrt(25))
ts1
```

We calculate the test statistic to be -0.5957439. The degrees of freedom (df) are (n-1) = (25-1) = 24.

Step 4 - we must calculate the p-value.

We use the following formula in R to compute the probability of this test statistic occuring with degrees of freedom = 24.

```
p1 <- pt(-0.5957439, 24)
p1
```

The p-value is equal to 0.2784617.

Step 5: Draw a conclusion

Since the p-value is greater than the alpha (0.2784617 is greater than 0.05), the result is not extreme enough to be statistically significant. There is not enough evidence allowing us to reject the null hypothesis and to accept the alternative hypothesis.

Therefore, the school counselor cannot conclude that the average student IQ in her school is higher than the average IQ score (100) among all the schools in the country.

Question 2 (50 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*

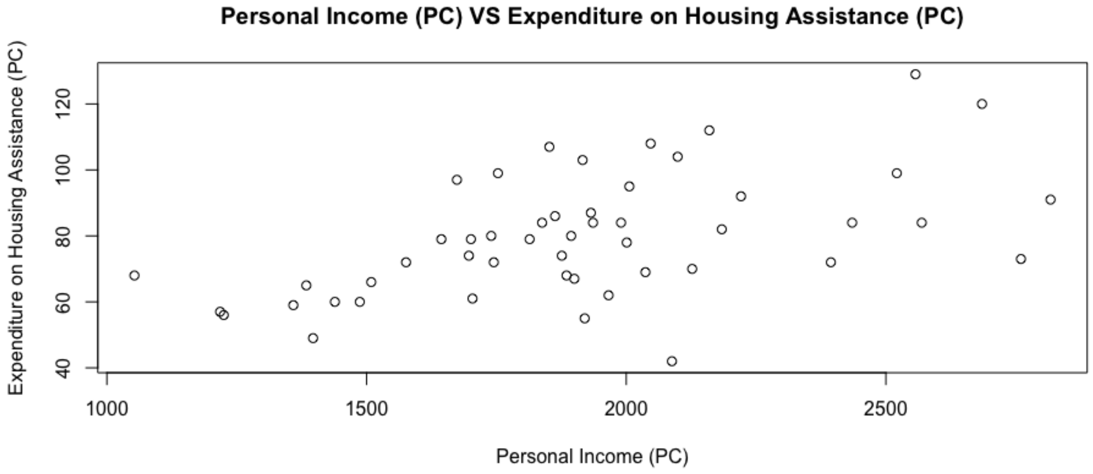
Explore the `expenditure` data set and import data into `R`.

- 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?

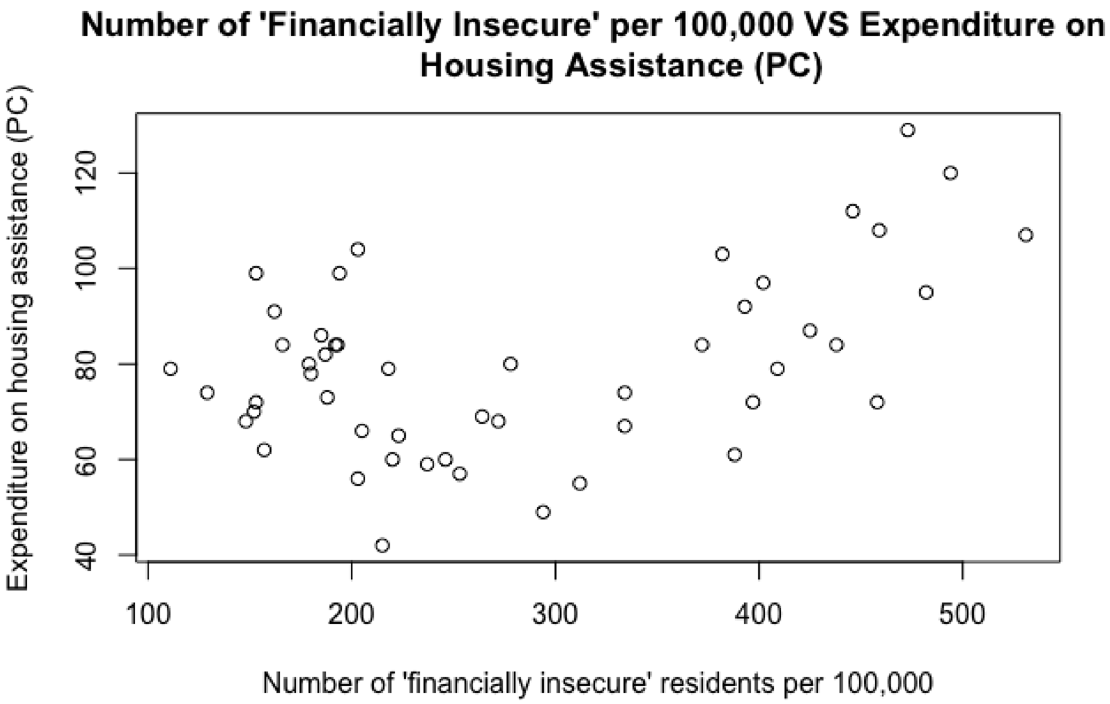
- Please plot the relationship between *Y* and *X1*? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.

Question 2 Part 1:



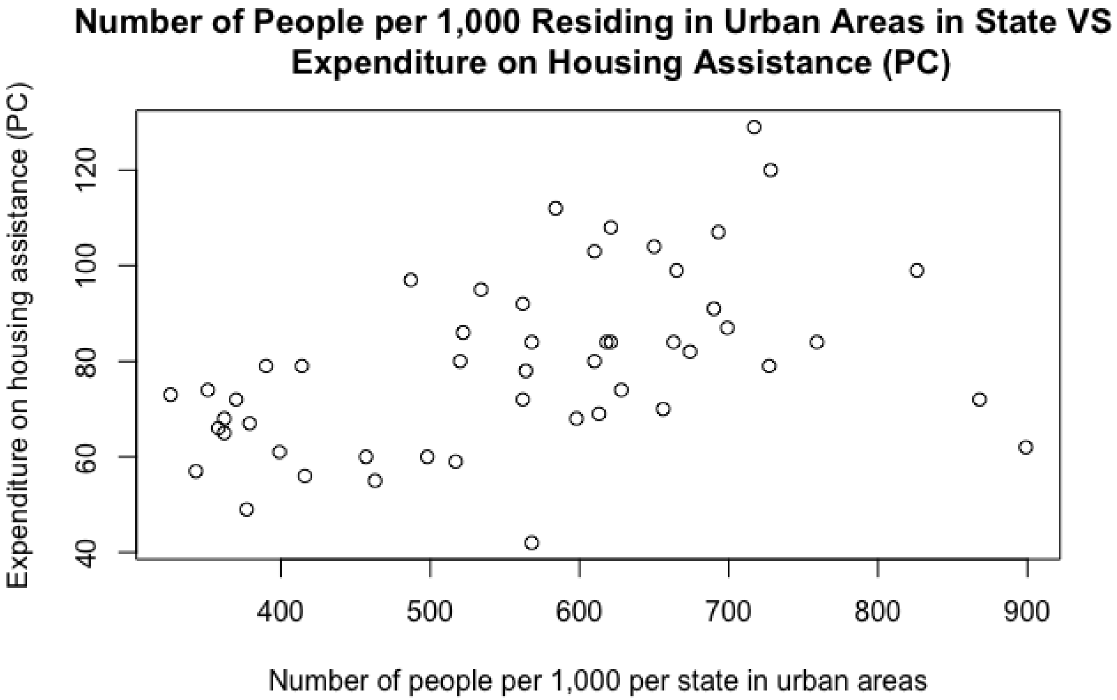
The relationship between personal income per capita and expenditure on housing assistance per capita appears to be positively correlated. As one variable increases, so does the other. This means that states with higher personal income per capita also appear to have higher spending per capita on housing assistance.

```
plot(x = expenditure$X1,
      y = expenditure$Y,
      xlab ="Personal income (PC)",
      ylab = "Expenditure on housing assistance (PC)",
      main = "Personal Income (PC) VS Expenditure on Housing Assistance (PC)")
```



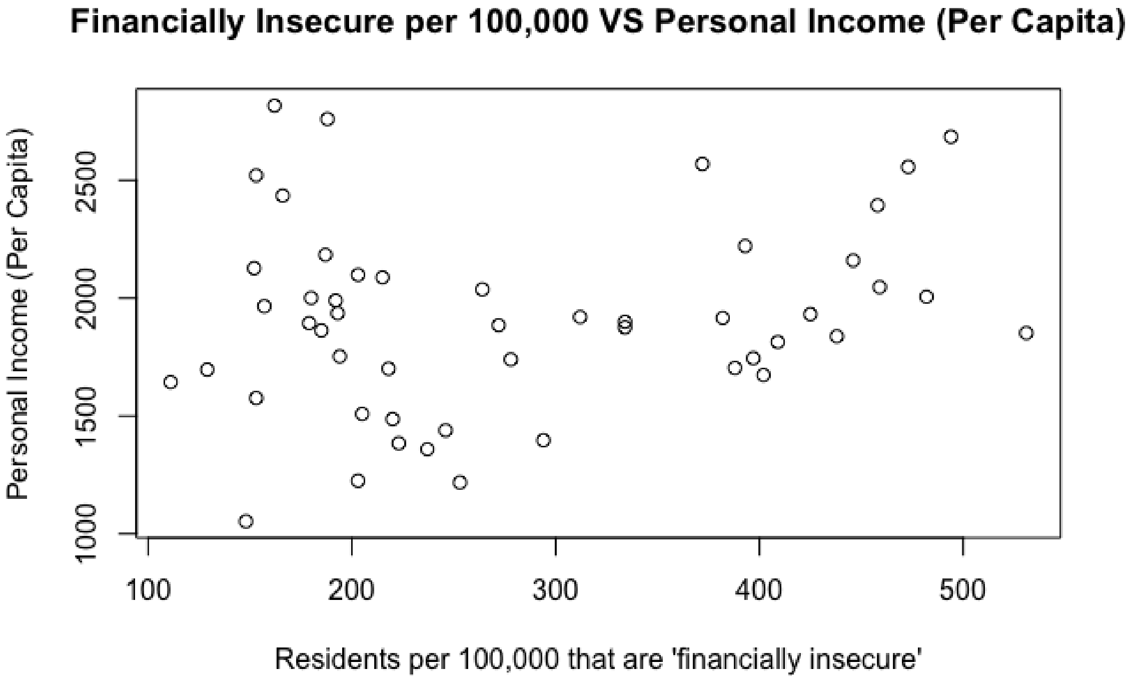
The relationship between the number of ’financially insecure’ residents per 100,000 and expenditure on housing assistance per capita appears to somewhat form a u-shaped curve. This would indicate a non linear relationship between the two variables, but rather a quadratic relationship. This would perhaps indicate that a squared function could describe the graph better than a linear one.

```
plot(x = expenditure$X2,
     y = expenditure$Y,
     xlab =”Number of ’financially insecure’ residents per 100,000”,
     ylab = ”Expenditure on housing assistance (PC)”,
     main = ”Number of ’Financially Insecure’ per 100,000 VS Expenditure on
     Housing Assistance (PC)”)
```



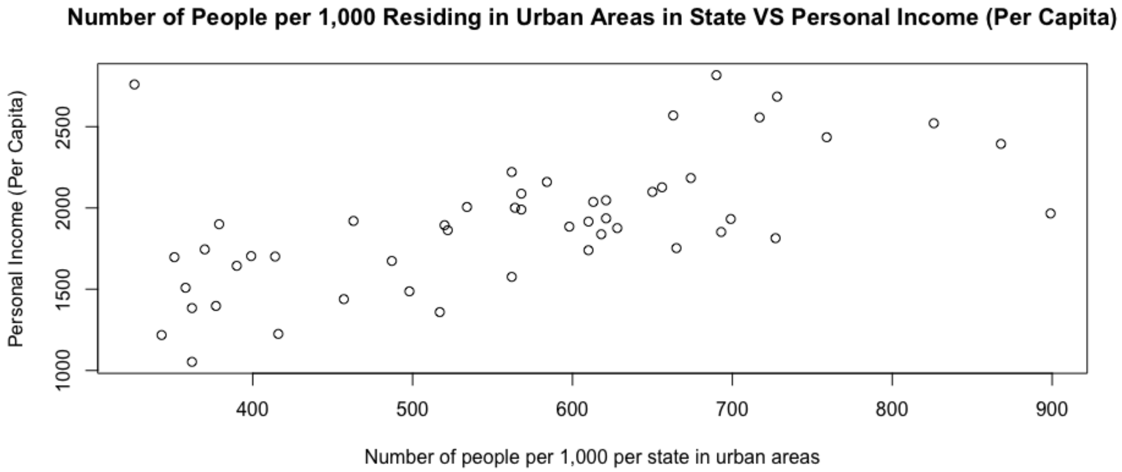
The number of people residing in urban areas per thousand in the state appears to be somewhat positively correlated with expenditure on housing assistance, though it does not appear to be as strongly correlated as personal income per capita. States with a greater proportion of the population living in urban areas appear to in general have a greater amount of spending on housing assistance per capita.

```
plot(x = expenditure$X3,
      y = expenditure$Y,
      xlab =”Number of people per 1,000 per state in urban areas”,
      ylab = ”Expenditure on housing assistance (PC)”,
      main = ”Number of People per 1,000 Residing in Urban Areas in State VS
      Expenditure on Housing Assistance (PC)”)
```



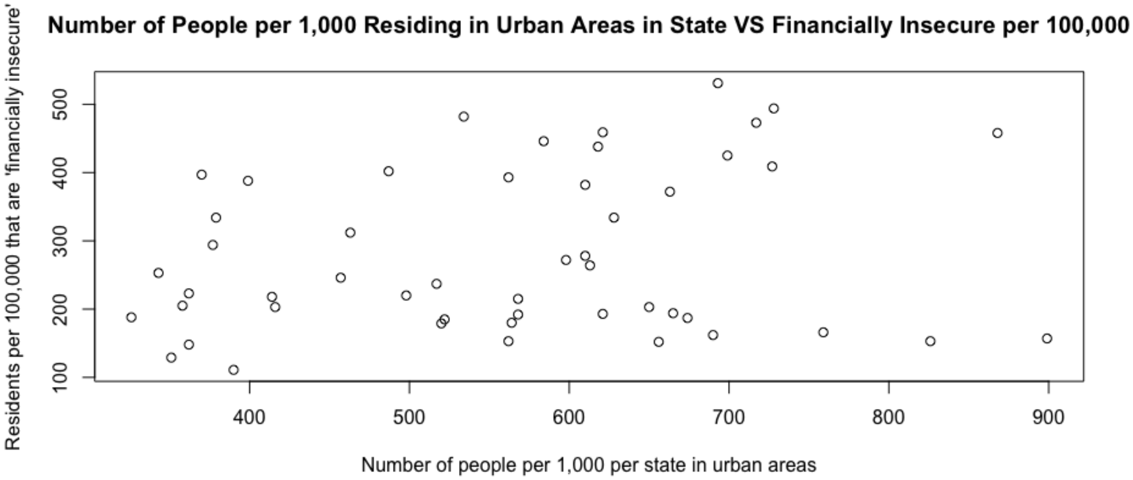
There appears to be a weak relationship between the number of financially insecure’ per 100,000 in the state and personal income per capita. They do not appear to be strongly correlated overall. There may however be a positive correlation between the two variables when the number of residents per 100,000 that are ’financially insecure’ exceeds 400.

```
plot(x = expenditure$X2,
     y = expenditure$X1,
     xlab =”Residents per 100,000 that are 'financially insecure'”,
     ylab = ”Personal Income (Per Capita)”,
     main = ”Financially Insecure per 100,000 VS Personal Income (Per Capita)”
   )
```



There appears to be a very strongly positively correlated relationship between the number of people in urban areas per thousand in the state and the personal income per capita of the people in these states. As one of these variables goes up, it is very likely that the other will as well. This means that states with higher income per capita also have a higher proportion of people living in urban areas, though we do not have enough data yet to know whether one causes the other.

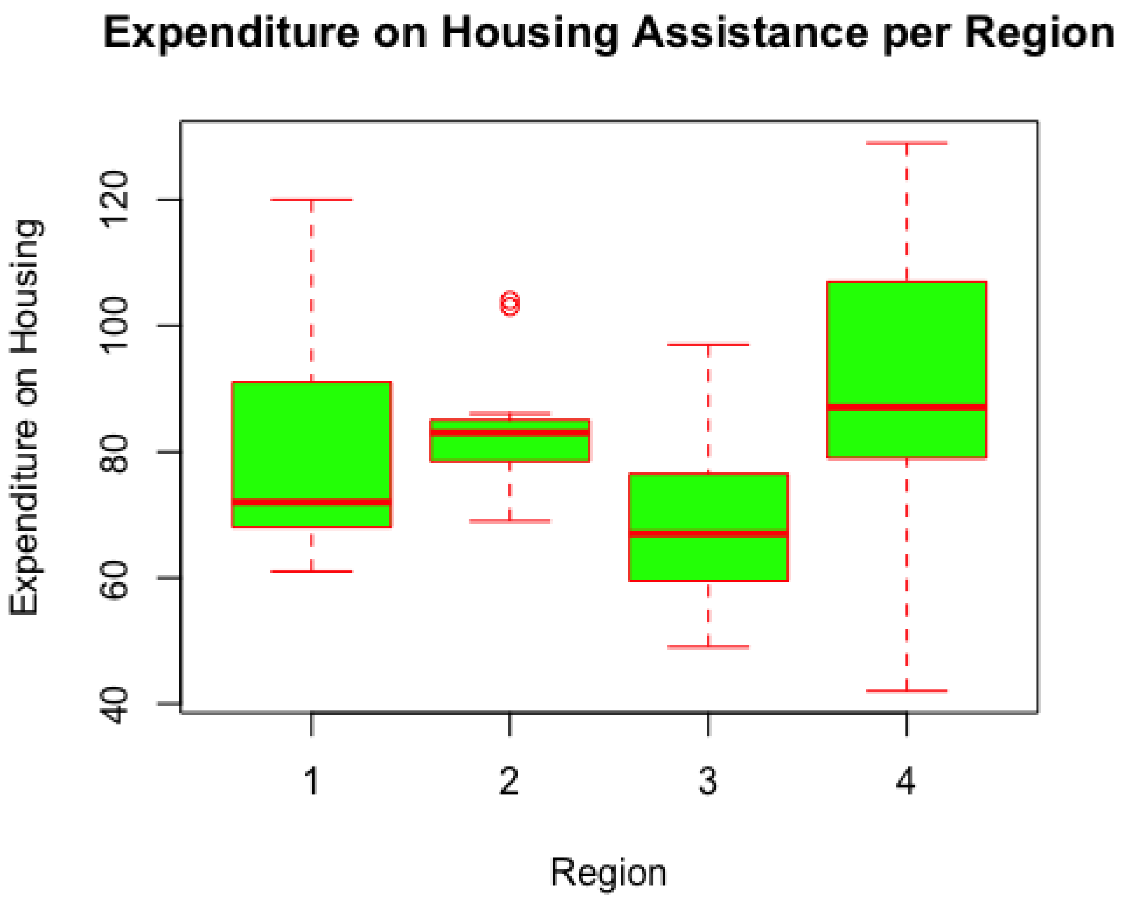
```
plot(x = expenditure$X3,
     y = expenditure$X1,
     xlab =”Number of people per 1,000 per state in urban areas”,
     ylab = ”Personal Income (Per Capita)”,
     main = ”Number of People per 1,000 Residing in Urban Areas in State VS Personal Income (Per Capita)”)
```



At first glance, there does not appear to be any correlation between the number of people residing in urban areas per 1000 people in the state and the number of ’financially insecure’ people per 100,000 in the state. The data points appear to be scattered around the plot without any clear relationship. The two variables seem to be neither positively nor negatively correlated.

```
plot(x = expenditure$X3,
      y = expenditure$X2,
      xlab =”Number of people per 1,000 per state in urban areas”,
      ylab = ”Residents per 100,000 that are ’financially insecure’”,
      main = ”Number of People per 1,000 Residing in Urban Areas in State VS Financially Insecure per 100,000”)
```

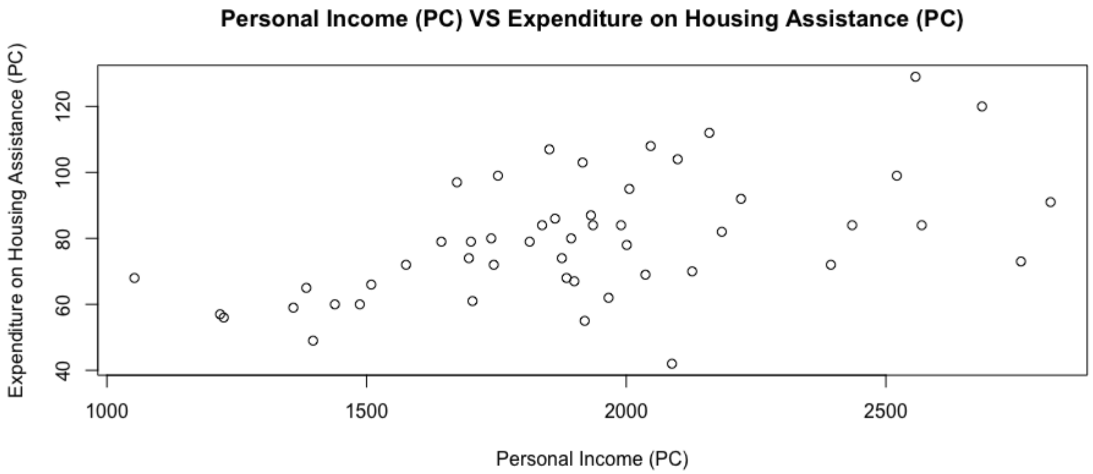
Question 2 Part 2:



On average, Region 4 (West) has the highest per capita expenditure on housing assistance. Its median, the red line in its box plot, is above that of all the others.

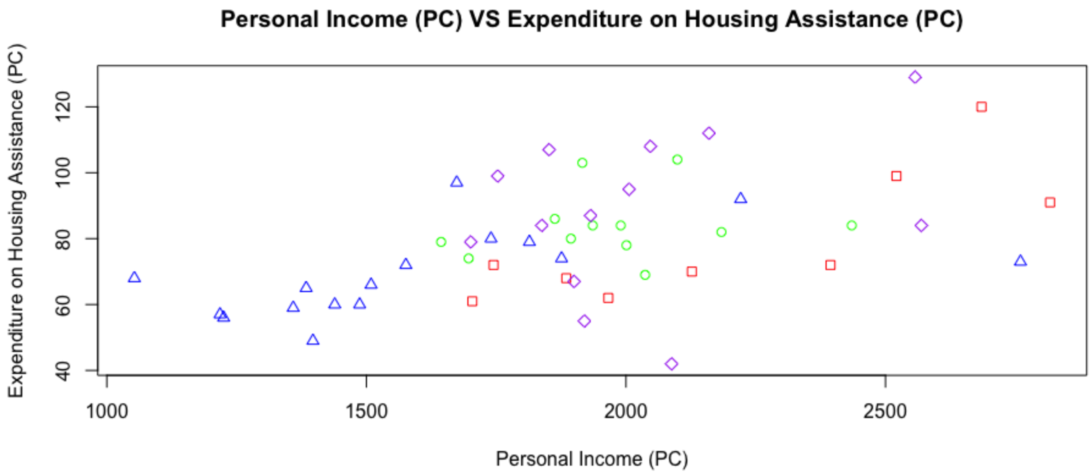
```
boxplot(expenditure$Y ~ expenditure$Region,
        data =expenditure,
        main = ”Expenditure on Housing Assistance per Region”,
        xlab=”Region”,
        ylab=”Expenditure on Housing”,
        col = ”green”,
        border = ”red”)
```

Question 2 Part 3:



The relationship between personal income per capita and expenditure on housing assistance per capita appears to be positively correlated. As one variable increases, so does the other. This means that states with higher personal income per capita also appear to have higher spending per capita on housing assistance.

```
plot(x = expenditure$X1,
     y = expenditure$Y,
     xlab =”Personal Income (PC)”,
     ylab = ”Expenditure on Housing Assistance (PC)”,
     main = ”Personal Income (PC) VS Expenditure on Housing Assistance (PC)”)
```



We see the same graph here with each region represented by a different colour and and symbol. It is immediately noticeable that there is a grouping of blue triangles with low personal income per capita and low expenditure on housing assistance per capita. These blue triangles are all in the ’South’ region.

1 = Northeast is represented by red squares, 2 = North Central is represented by green circles, 3 = South is represented by blue triangles and 4 = West is represented by purple diamonds.

```
plot(x = expenditure$X1,
     y = expenditure$Y,
     xlab =”Personal Income (PC)”,
     ylab = ”Expenditure on Housing Assistance (PC)”,
     main = ”Personal Income (PC) VS Expenditure on Housing Assistance (PC)”,
     col = c(’red’, ’green’, ’blue’, ’purple’)[expenditure$Region],
     pch  = c(0, 1, 2, 5)[expenditure$Region]
     #1Northeast, 2NorthCentral, 3South, 4West
```