Problem Set 1

Applied Stats/Quant Methods 1

Due: September 30, 2024

Name: Eimhin ONeill
Student Number: 20332107

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:

```
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)
```

Part 1 - Confidence Interval

Find a 90 per cent confidence-interval for the average student IQ in the school. The confidence interval is 90%. For calculating it, we need the point estimate, or sample mean, sample standard deviation and standard error. Then, once we have them, we can start calculating the confidence interval.

- **Sample mean/point estimate:** The point-estimate, or mean, can be calculated using the function in R as found below...

```
SchoolMean <- mean(y)    #Point estimate
```

This gives us a point estimate of 98.44.

- **Sample standard deviation:** The sample standard deviation can be calculated using the formula...

$$\text{Sample sd} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \text{Sample Mean})^2}{n-1}}$$

The function in R we use to calculate this is:

```
SchoolSD <- sd(y)          #Standard Deviation
```

This gives us a standard deviation of 13.09287.

- **Standard error:** The standard error can be calculated using the formula...

$$\text{Standard Error} = \frac{\text{Sample sd}}{\sqrt{n}}$$

The formula is different in R and we have to get the sqrt of the length of n. So the function we use to calculate this is:

```
SchoolError <- sd(y) / sqrt(length((y))) #Standard Error
```

This gives us a standard error of 2.618575.

Now that we have the point estimate, standard deviation and standard error, we need to calculate the 90% confidence interval.

However, we need to use a t-distribution for this because the sample doesn't seem to be normally distributed and n¿30. The sample size is actually 25, given by the length of y captured above, which means the degrees of freedom, n-1, is equal to 24.

The t-score we need is calculated from the t-distribution and the formula to calculate this is...

- **T-score:**

$$t - score_{90} = qt\left(\frac{1-0.90}{2}, df\right)$$

The function in R to calculate this is the qt function which will give us the critical value we need

```
SchoolT <- qt(1 - 0.05, SchoolDF) # Using this to get the critical value
```

This gives us a t-score of 1.710882

Next up, we need to calculate the lower and upper bounds which will serve as the two points of our confidence interval

- **Confidence interval:** The lower and upper bounds are calculated as:

$$\text{Upper Bound} = \text{Sample Mean} + tscore_{90} \times \text{Standard Error}$$

$$\text{Lower Bound} = \text{Sample Mean} - tscore_{90} \times \text{Standard Error}$$

The function in R to calculate the upper and lower bounds, and confidence intervals are as follows:

```
upper_90 <- (mean(y))+(SchoolT)*(sd(y)/sqrt(length(y))) # Upper Bound
lower_90 <- (mean(y))-(SchoolT)*(sd(y)/sqrt(length(y))) # Lower Bound

c(lower_90, upper_90) #The 90% confidence interval
CI90 <- c(lower_90, upper_90)
CI90
```

This gave us a confidence interval of [93.95993, 102.92007]. T

This is the range in which we expect the population parameter to fall given the 90% level of confidence we calculated for.

Part 2

1. 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, we will conduct the appropriate hypothesis test with $\alpha = 0.05$.

For conducting the hypothesis test, I am going to use the 5 Steps shown in class.

- **1. Assumptions about our data**
  - We can assume that the population is approximately normally distributed,
  - that the scores of the IQ test are independent of eachother and
  - that the previous sample is representive of the school and random.

- **2. Formulating our hypotheses**
  - H0 - The mean IQ of our sample school less than or equal to 100 (national average)
  - H1 - The mean IQ of our sample school is greater than 100

  Its going to be a one-sided test because we are trying to find out if the school average is *greater than* the national average.

- **3. Calculate the test statistic**

The function in R to calculate this is as follows...

```
NationalMean <- 100
# For finding the t-statistic, I'll use the formula
# (Mean of sample - assumed mean) / (Standard dev. / sqrt of N)
# But I'll use the standard error we already calculated as the numerator
TStat <- (SchoolMean - NationalMean) / (SchoolError)
TStat
```

This gave us a t-statistic of -0.5957439.

- **4. Calculate the p-value**

The function in R to calculate this is the pt function, where we use the previously calculated t-statistic and degrees of freedom.

```
P <- pt(TStat, SchoolDF, lower.tail = FALSE)
P
```

This gave us a p-value of 0.7215383.

- **5. Drawing a conclusion** Using the values previously calculated, I was able to draw the following conclusions

We fail to reject the null hypothesis as the p-value calculated of 0.7215383 is greater than our alpha of 0.05 and therefore, we fail to reject our Null Hypothesis of the mean IQ of our sample school less than or equal to the national average of 100.

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`.

Importing and reviewing the dataset

I imported the dataset using the read.table function provided and it produced this data set, provided by using the head function in R

```
STATE  Y   X1  X2  X3 Region
1    ME 61 1704 388 399      1
2    NH 68 1885 272 598      1
3    VT 72 1745 397 370      1
4    MA 72 2394 458 868      1
5    RI 62 1966 157 899      1
6    CT 91 2817 162 690      1
```

Then, I went on to view the structure and get a quick summary of the data using the str and summary functions. The structure was as follows..

```
'data.frame': 50 obs. of  6 variables:
$ STATE : chr  "ME" "NH" "VT" "MA" ...
$ Y     : int  61 68 72 72 62 91 120 99 70 82 ...
$ X1    : int  1704 1885 1745 2394 1966 2817 2685 2521 2127 2184 ...
$ X2    : int  388 272 397 458 157 162 494 153 152 187 ...
$ X3    : int  399 598 370 868 899 690 728 826 656 674 ...
$ Region: int  1 1 1 1 1 1 1 1 2 ...
```

While the summary allowed me to get a better overview of data including vital characteristics like the length, mean and median of different variables and more. The summary was as follows...

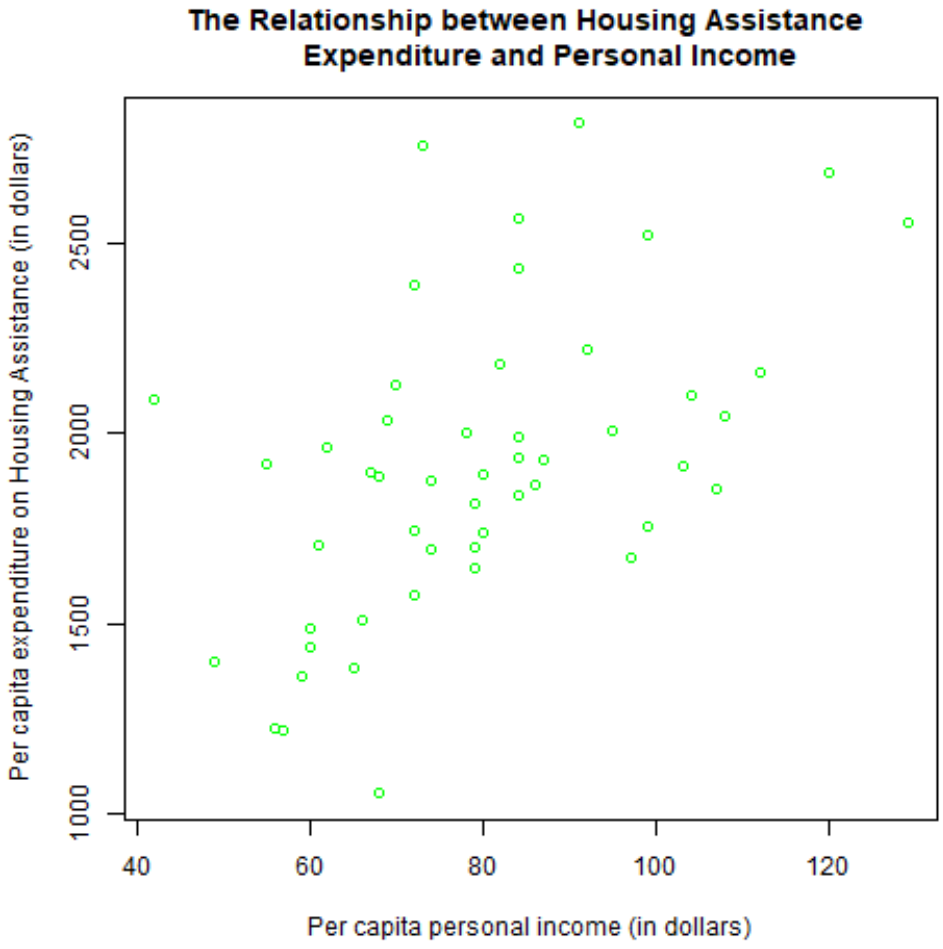
```
STATE                  Y                X1             X2
Length:50          Min.   : 42.00   Min.   :1053   Min.   :111.0
Class :character   1st Qu.: 67.25   1st Qu.:1698   1st Qu.:187.2
Mode  :character   Median : 79.00   Median :1897   Median :241.5
Mean   : 79.54   Mean   :1912   Mean   :281.8
3rd Qu.: 90.00   3rd Qu.:2096   3rd Qu.:391.8
Max.   :129.00   Max.   :2817   Max.   :531.0
X3            Region
Min.   :326.0   Min.   :1.00
1st Qu.:426.2   1st Qu.:2.00
Median :568.0   Median :3.00
Mean   :561.7   Mean   :2.66
3rd Qu.:661.2   3rd Qu.:3.75
Max.   :899.0   Max.   :4.00
```

- 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)?

I plotted each of the respective relationships between Y and X1,X2 and X3 variables using the plot() functions

Relationship between X1 and Y This was plot function I used for this relationship

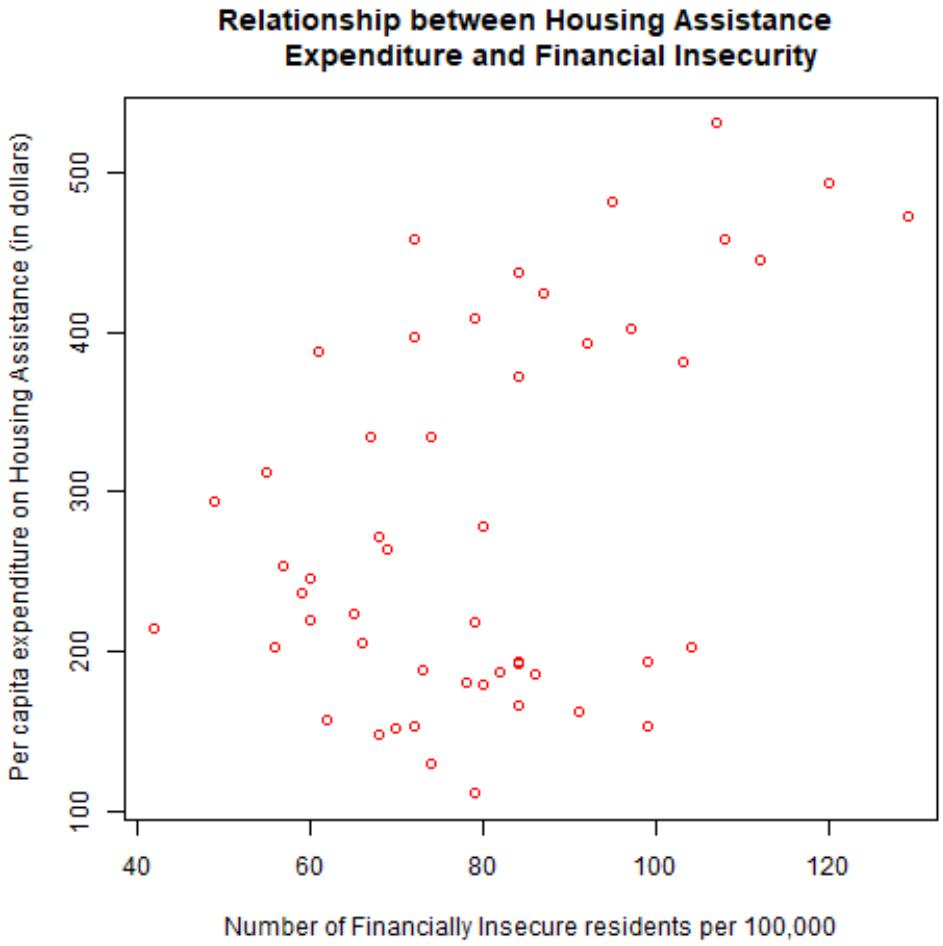
```
png(file="X1_Y_Plot.png")
plot(USexpenditure$X1 ~ USexpenditure$Y,
     main="The Relationship between Housing Assistance
     Expenditure and Personal Income",
     xlab="Per capita personal income (in dollars)",
     ylab="Per capita expenditure on Housing Assistance (in dollars)",
     col = "green")
```



This plot displays to us a positive linear relationship between X1 and Y, indicating that as personal income in a state increases, that expenditure of shelters and housing assistance increases too.

Relationship between X2 and Y This was plot function I used for this relationship

```
png(file="X2_Y_Plot.png")
plot(USexpenditure$X2 ~ USexpenditure$Y,
     main="Relationship between Housing Assistance
     Expenditure and Financial Insecurity",
      xlab="Number of Financially Insecure residents per 100,000",
      ylab="Per capita expenditure on Housing Assistance (in dollars)",
     col = "red")
```

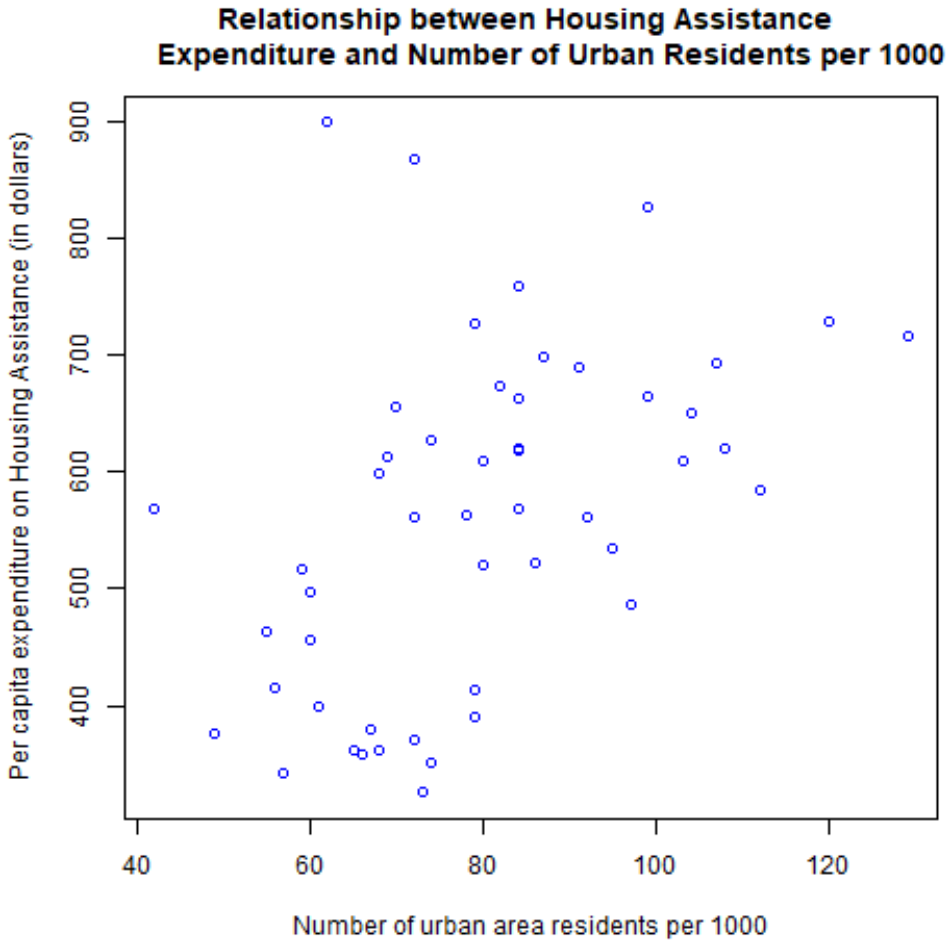


This plot displays to us a positive linear relationship between X2 and Y, but the dispersion of the data points would point to a weaker relationship between the variables. At-large, this indicates that as the number of financially insecure residents in a state increases, that expenditure of shelters and housing assistance increases would too. However, with the presence of outliers and general dispersion, this relationship could be less predictive than that between X1 and Y.

Relationship between X3 and Y This was plot function I used for this relationship

```
# Relationship between X3 and Y

png(file="X3_Y_Plot.png")
plot(USexpenditure$X3 ~ USexpenditure$Y,
      main="Relationship between Housing Assistance
     Expenditure and Number of Urban Residents per 1000",
      xlab="Number of urban area residents per 1000",
      ylab="Per capita expenditure on Housing Assistance (in dollars)",
     col = "blue")
```

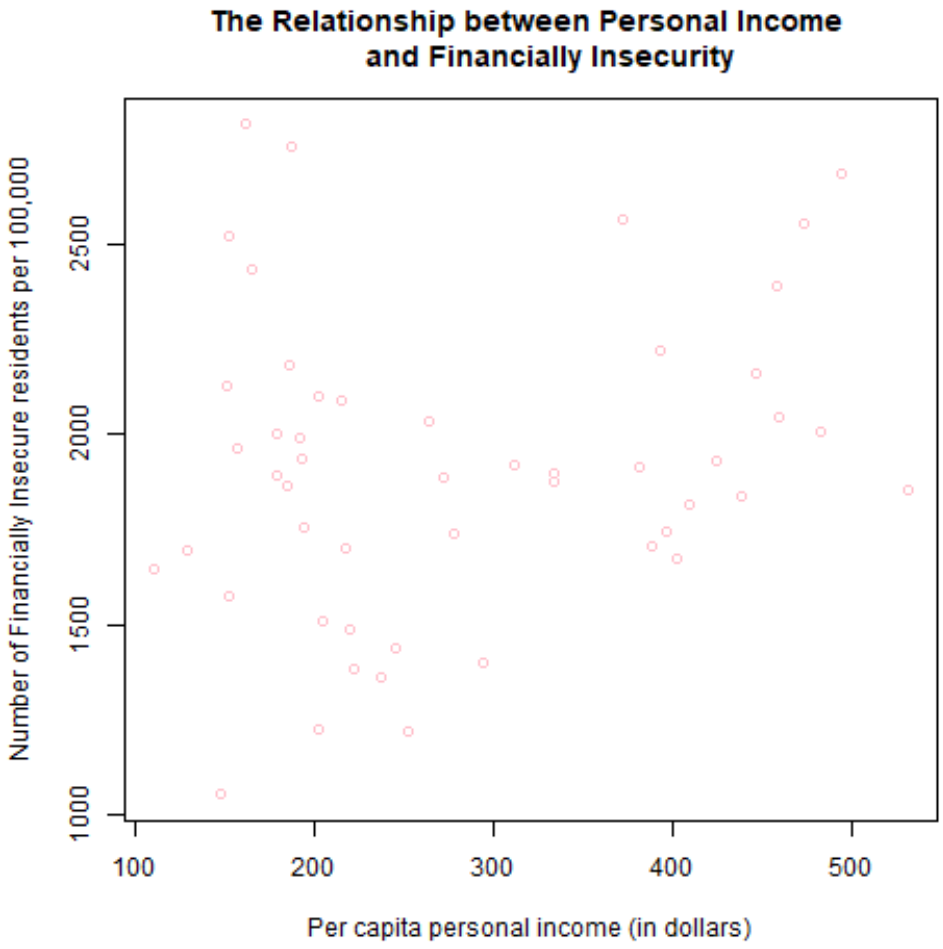


This plot displays to us a positive linear relationship between X3 and Y, like the past two plots. This generally indicates that as the number of urban residents in a state increases, that expenditure of shelters and housing assistance increases would too. However, once again, this is not as strong of a relationship as X1 and Y, perhaps suggesting other factors are influencing this relationship too.

Relationship between X1 and X2 This was plot function I used for this relationship

```
# Relationship between X1 and X2

png(file="X1_X2_Plot.png")
plot(USexpenditure$X1 ~ USexpenditure$X2,
     main="The Relationship between Personal Income
     and Financially Insecurity",
     xlab="Per capita personal income (in dollars)",
     ylab="Number of Financially Insecure residents per 100,000",
     col = "pink")
```



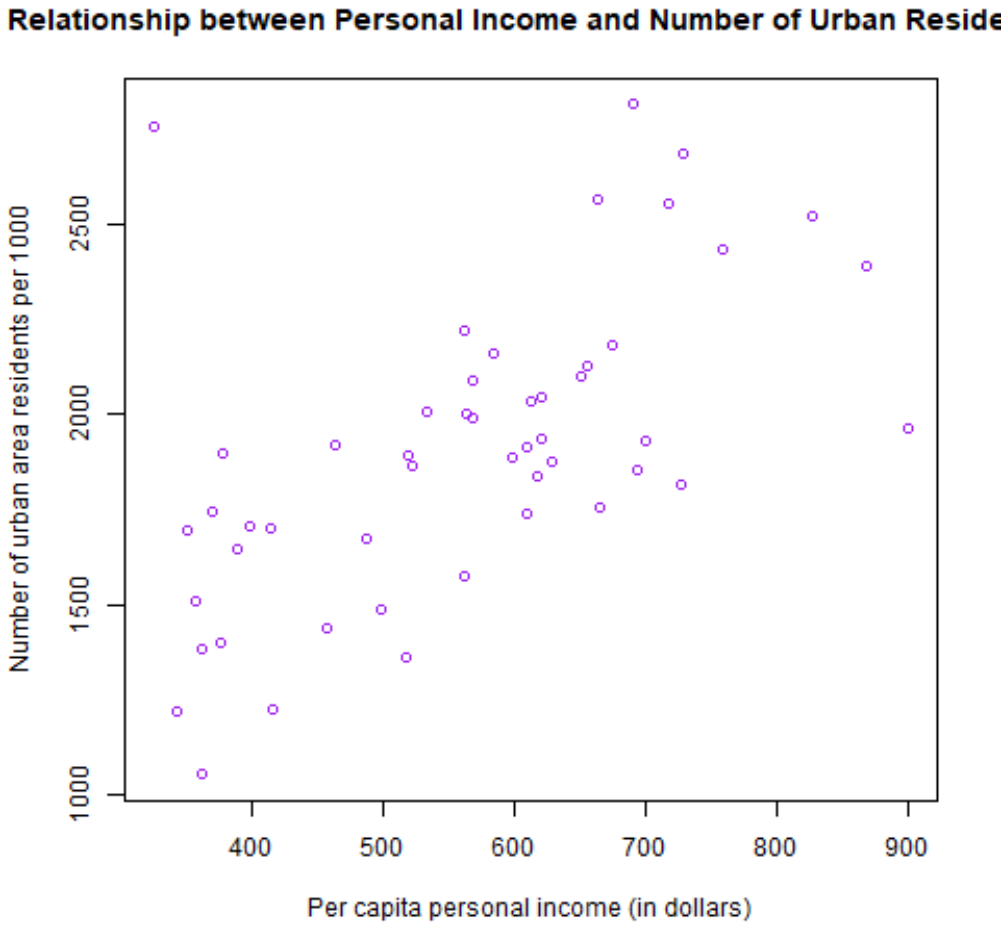
This plot displays a positive relationship between X1 and X2, but it is a very weak relationship between the variables. There is a slight trend upwards but with obvious outliers to this in states with lower levels of personal income, where one could contend the trend line resembling a U shape.

This U line indicates that as the number of personal income in a state increases, that the number of financially insecure individuals does too. However, with the presence

of the outliers in low income states, this relationship could predict as X1 decreases, X2 increases.

Relationship between X1 and X3 This was plot function I used for this relationship

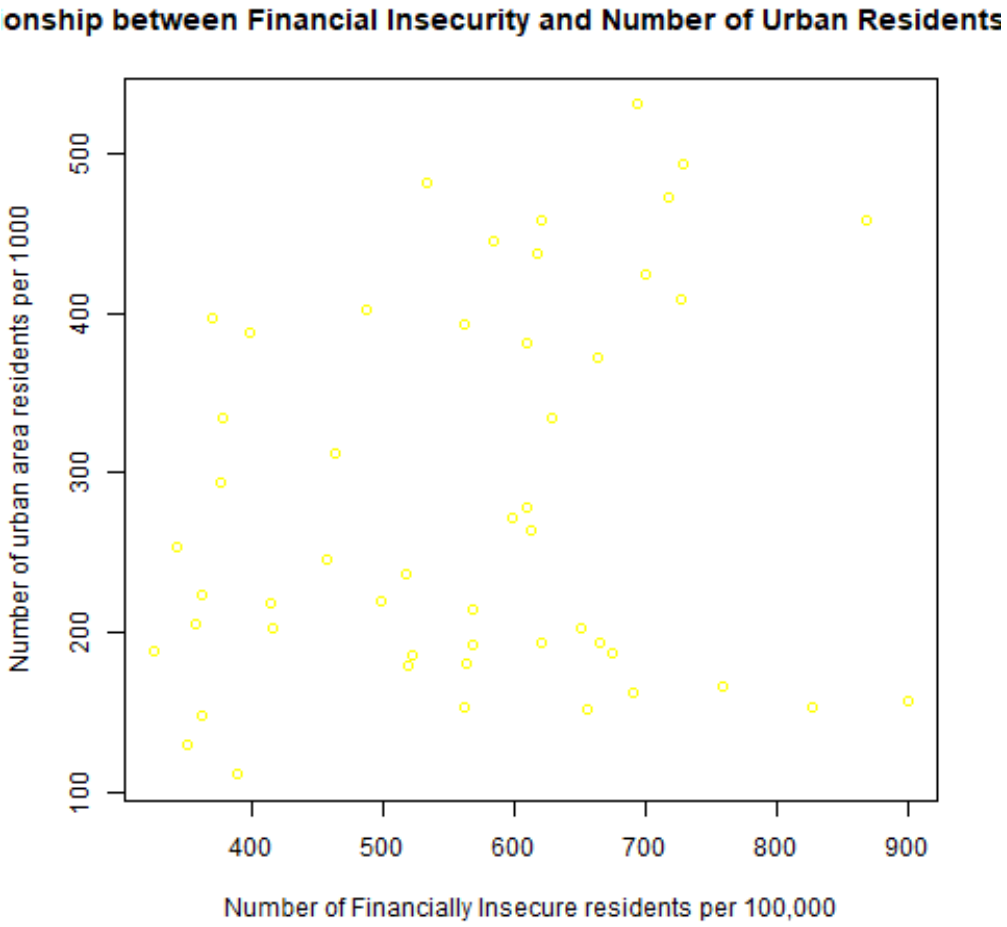
```
png(file="X1_X3_Plot.png")
plot(USexpenditure$X1 ~ USexpenditure$X3,
     main="Relationship between Personal Income and Number of Urban
   Residents",
     xlab="Per capita personal income (in dollars)",
     ylab="Number of urban area residents per 1000",
     col = "purple")
```



This plot displays a stronger positive relationship between X1 and X3, indicating that as personal income in a state increases, that the number of urban residents increases too. This could be down to more high paying jobs being located in capital cities and on the outskirts of them as well, a common phenomenon worldwide

Relationship between X2 and X3 This was plot function I used for this relationship

```
png(file="X2_X3_Plot.png")
plot(USexpenditure$X2 ~ USexpenditure$X3,
     main="Relationship between Financial Insecurity and Number of Urban
   Residents per 1000",
     xlab="Number of Financially Insecure residents per 100,000",
     ylab="Number of urban area residents per 1000",
     col = "yellow")
```

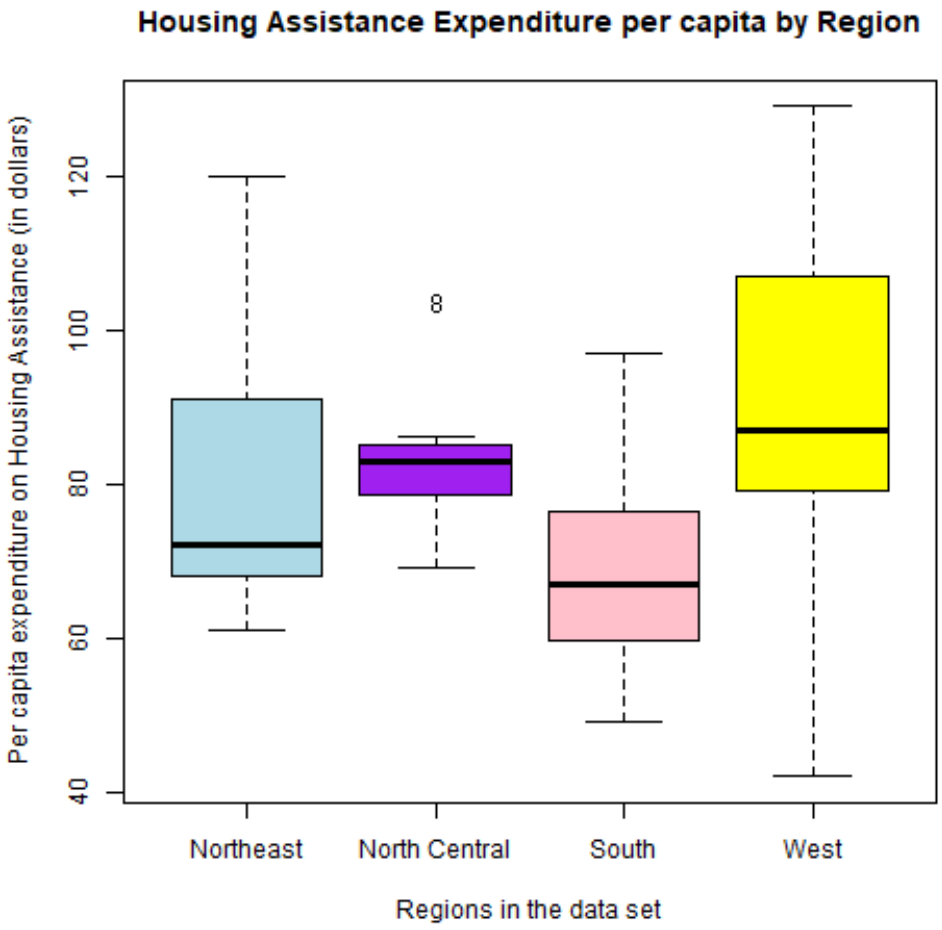


This plot does not display any concrete positive or negative relationship between X2 and X3,rather the data points are scattered and do not cluster around any given line. This would indicate to me that financial insecurity does not seem to be correlated with where the residents is located, in either urbanity or rural areas.

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

Relationship between Y and Region This was boxplot function I used for this relationship

```
png(file="Y_R_Boxplot2.png")
boxplot(USexpenditure$Y ~ USexpenditure$Region,
        main="Housing Assistance Expenditure per capita by Region",
        xlab="Regions in the data set",
        ylab="Per capita expenditure on Housing Assistance (in dollars)",
        col = c("lightblue", "purple", "pink","yellow"),
        names = c("Northeast", "North Central", "South", "West")
        )
```



This boxplot shows us that the region that spends the most on housing assistance and shelters is the West, as it enjoys the highest first and 3rd quartiles, and median, of around 90 dollars.

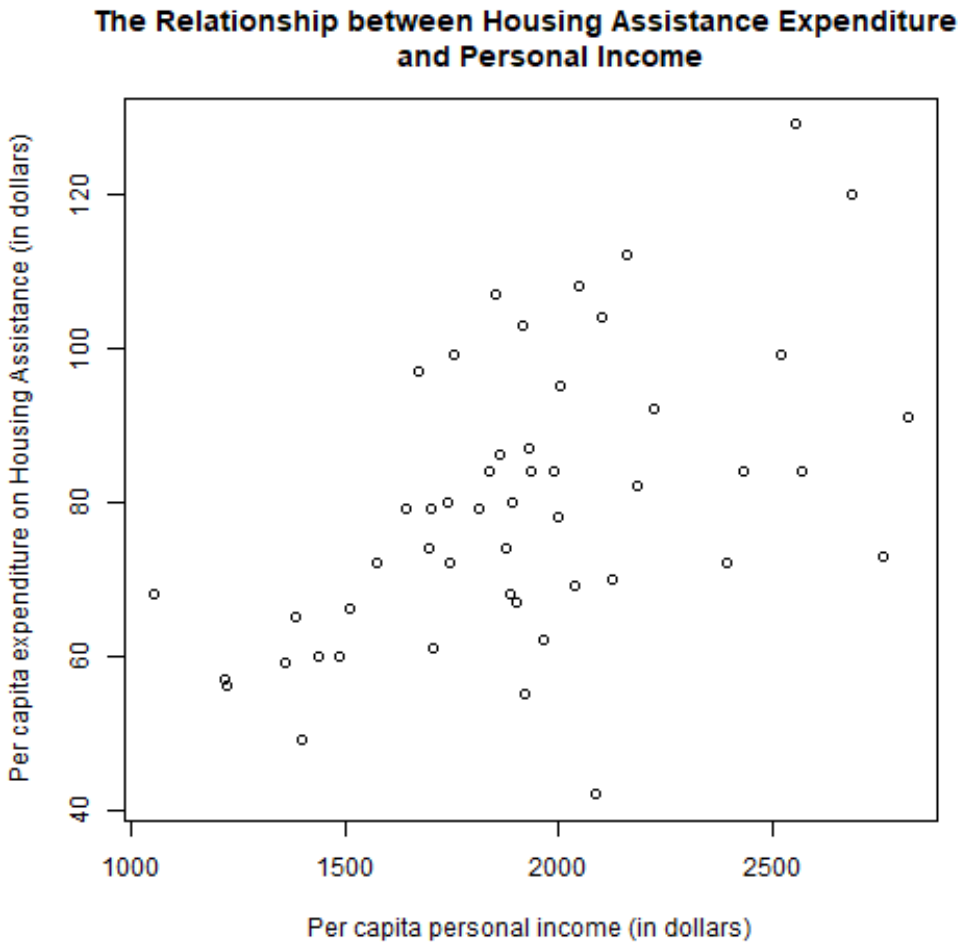
However, it has has the largest interquartile range, which indicates that there is a lot

of variation across the states in that region. We can compare this on the boxplot to North Central, which has a tighter interquartile range, signifying less variation between the region's different states' expenditure.

- 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 Y and X1 This was plot function I used for this relationship

```
png(file="X1_Y_Plot2.png")
plot(USexpenditure$Y ~ USexpenditure$X1,
     main="The Relationship between Housing Assistance Expenditure
     and Personal Income",
     xlab="Per capita personal income (in dollars)",
     ylab="Per capita expenditure on Housing Assistance (in dollars)",
     col = "black")
```

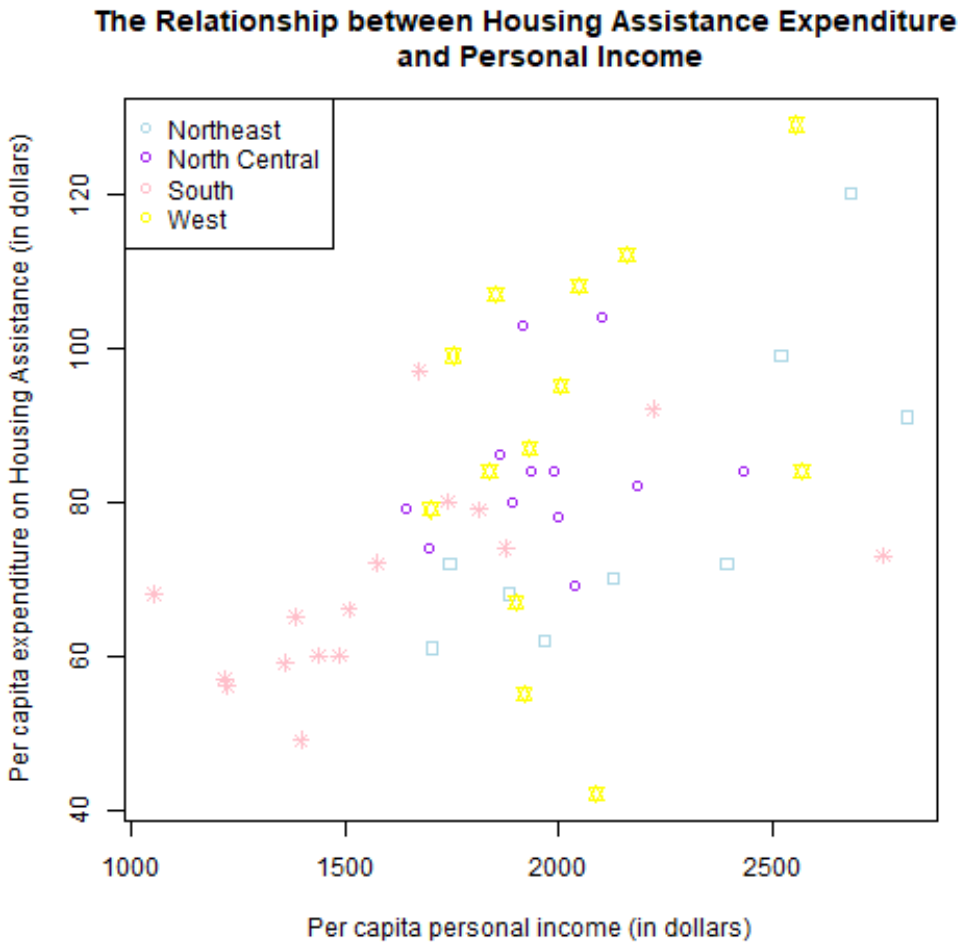


This plot displays to us, once again, a positive linear relationship between X1 and Y, indicating that as personal income in a state increases, that expenditure of shelters and housing assistance increases too.

Relationship between Y and X1, including the variable for Regions I then used this plot function to include Region as a variable, using different colours and variables, with a legend to distinguish each of the respective regions.

```
colours <- c("lightblue", "purple", "pink","yellow") #Same colours as the
    ones
# from the boxplot
symbols <- c(0, 1, 8, 11) #Symbols to easier distinguish each region

png(file="X1_Y_BetterPlot.png")
plot(USexpenditure$Y ~ USexpenditure$X1,
     main="The Relationship between Housing Assistance Expenditure
     and Personal Income",
     xlab="Per capita personal income (in dollars)",
     ylab="Per capita expenditure on Housing Assistance (in dollars)",
     col=colours[USexpenditure$Region],
     pch=symbols[USexpenditure$Region]
)
```



This reinforces what we found in the boxplot with the West having the highest

expenditures across all regions and while I previously noted North Central having a tight interquartile range compared to the West, on this plot, we see no real relationship between variables in the North Central region.