Problem Set 1

V. Dringelyte

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:

```
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.

   To calculate the confidence interval when the confidence coefficient = .90, I first found the mean:

```
sum(y)/length(y)
mean(y)
```

This returned a mean of 98.44

Then, I found the standard deviation:

```
sd_y <- sd(y, na.rm = FALSE)
```

Which returned a value of 13.09

So I could find the standard error.

```
se_y <- sd_y/sqrt(n)
```

Value of 2.62

Then I found the t-score

```
t_score <- qt(0.95, df=n-1)
```

I could use this to find the upper bounds of confidence interval

```
upper_90 <- mean(y)+(t_score)*(sd(y)/sqrt(length(y)))
```

And the lower bounds of confidence interval

```
lower_90 <- mean(y)-(t_score)*(sd(y)/sqrt(length(y)))
```

This showed that the 90% confidence interval for the the average student IQ is 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 $\alpha = 0.05$.

   To find the answer to this, I used the 5 steps of a statistical significance test.

   Step 1: Assumptions.

   1. The national average is 100.

```
meana <- 100
```

   2. The national sample has a normal distribution and random sampling, so we will use a t-score.

   Step 2: State hypothesis.

   The average IQ of the students in this school is higher than the average IQ score among all the schools in the country. The null hypothesis, then, is that the average IQ in the school is lower or equal to the average of schools nationally.

   Step 3: Calculate a test statistic. To do this, I used the formula:

$$t = \frac{(\bar{y}-\mu)}{se}$$

In R:

```r
n <- length(y)
sd_y <- sd(y)
se_y <- sd_y/sqrt(n)
y_ts <- (mean(y) - meana)/(se_y)
```

or, a faster way

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

t-statistic = -0.59574

Step 4: Calculate a P-value

```r
p_value <- pt(y_ts,n-1,lower.tail = FALSE)
```

P-value = 0.722

Step 5: Draw a conclusion As the P-value is larger than 0.05, we fail to reject the null hypothesis. So, the average IQ in the school is less than or equal to the average of schools in 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.

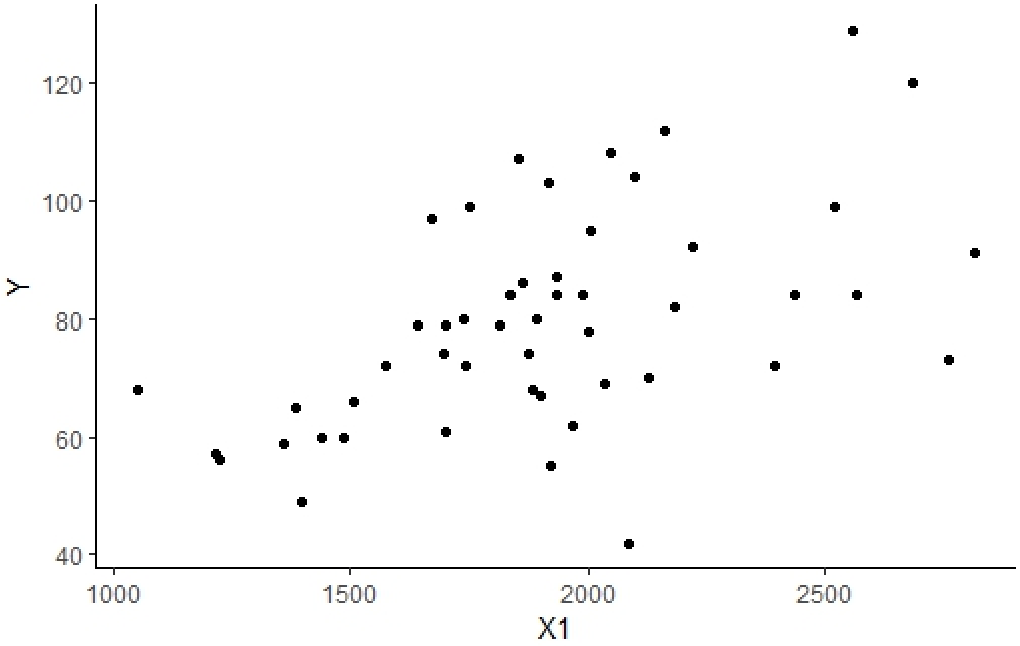
`State`	*50 states in US*
`Y`	*per capita expenditure on shelters/housing assistance in state*
`X1`	*per capita personal income in state*
`X2`	*Number of residents per 100,000 that are "financially insecure" in state*
`X3`	*Number of people per thousand residing in urban areas in state*
`Region`	*1=Northeast,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

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

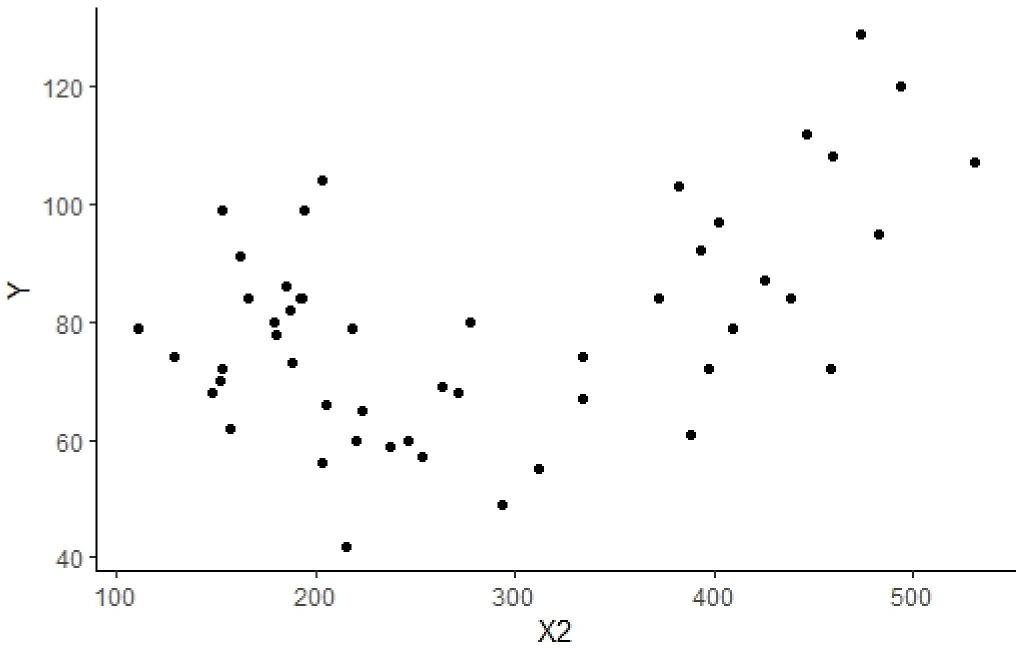
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?

  I used the code we learnt in coding camp to plot these graphs.

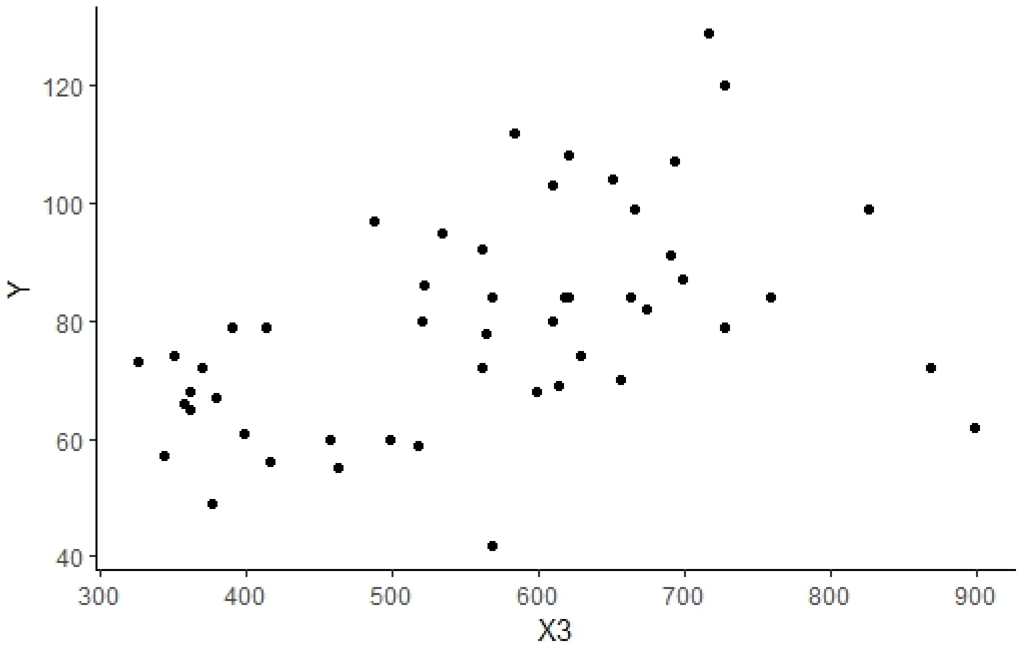
  ```
  ggplot(expenditure, aes(x, y))+
  geom_point() +
  theme_classic()
  ```

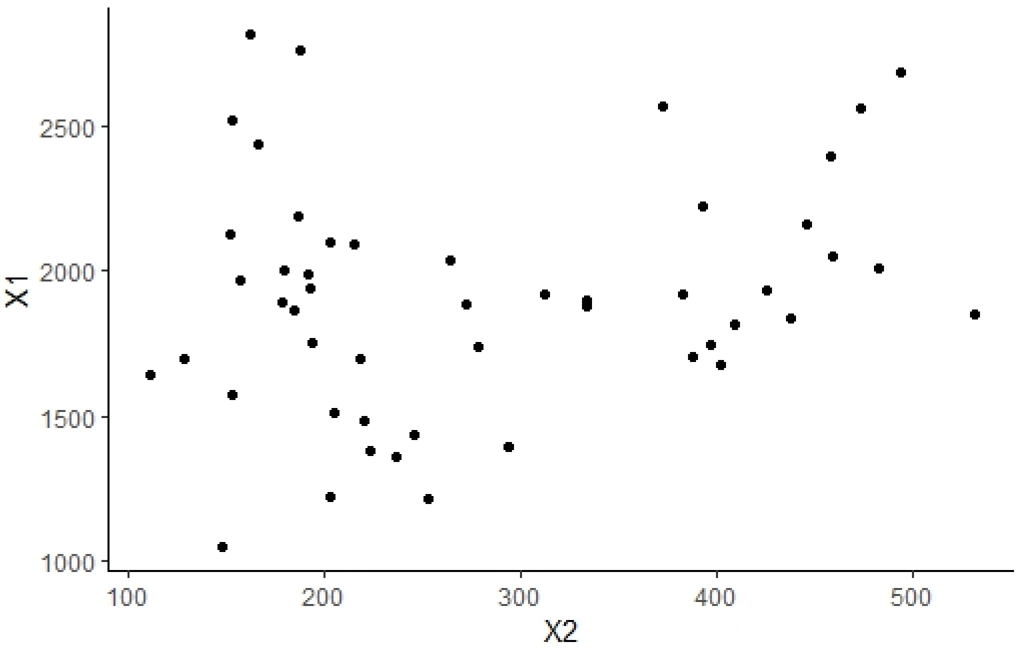
  Y and X1. There is a slight positive relationship between Y and X1, on the lower end, the relationship is stronger, and gets weaker as values increase.



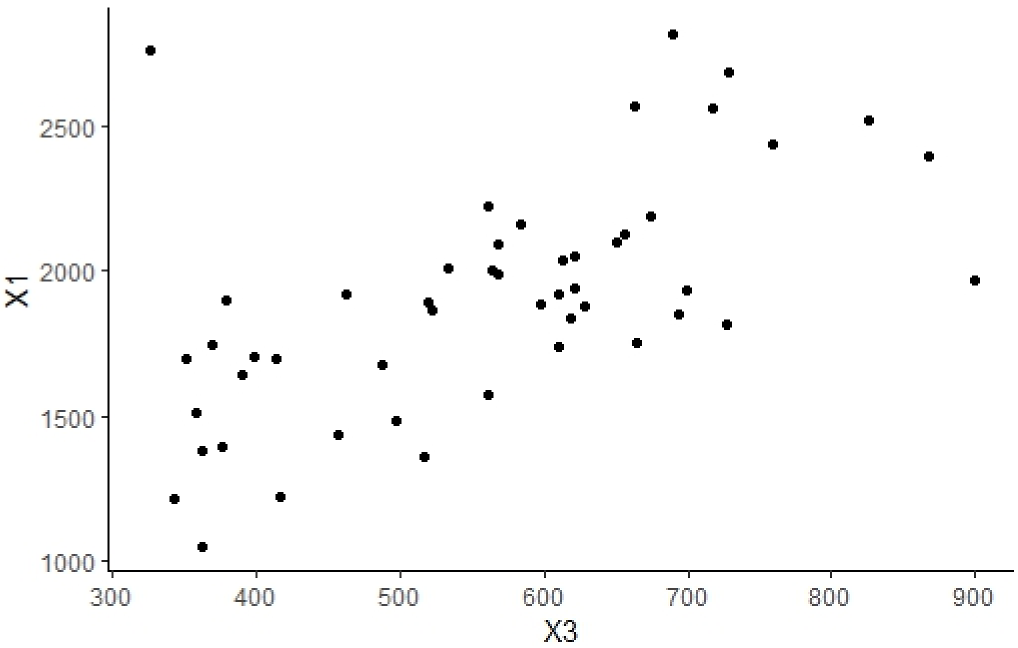
Y and X2. The scatter plot forms a U shape. It seems that the relationship between Y and X2 is negative at first, but changes to a positive one near the middle of the graph.



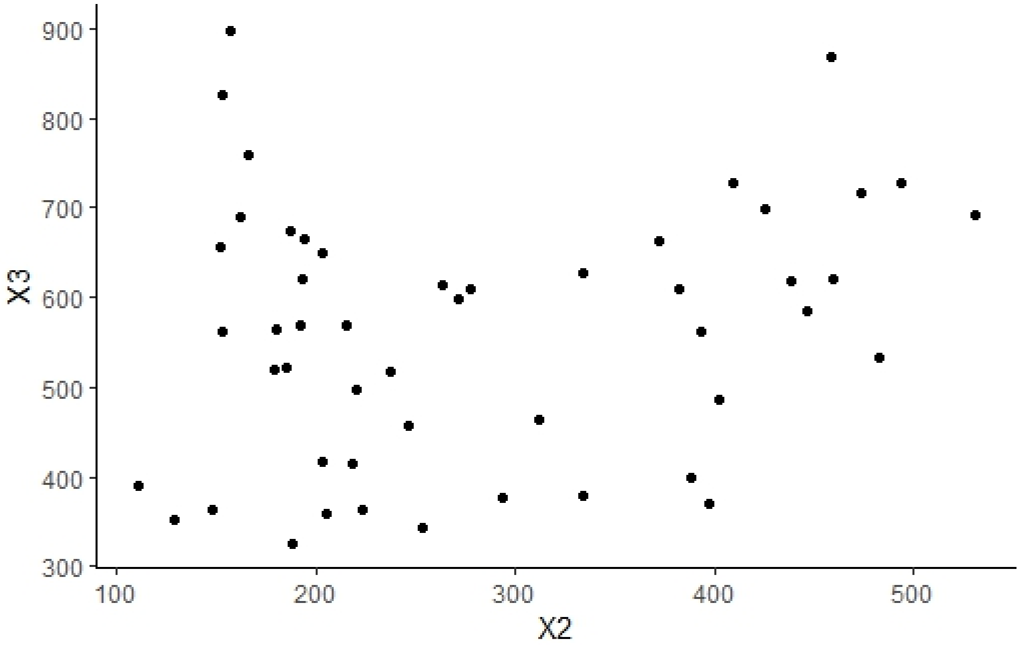
Y and X3. There seems to be a slight positive relationship between Y and X3, as one increases, so does the other.



X1 and X2. The scatter plot forms a U shape. X2 values show little correlation to Y values.



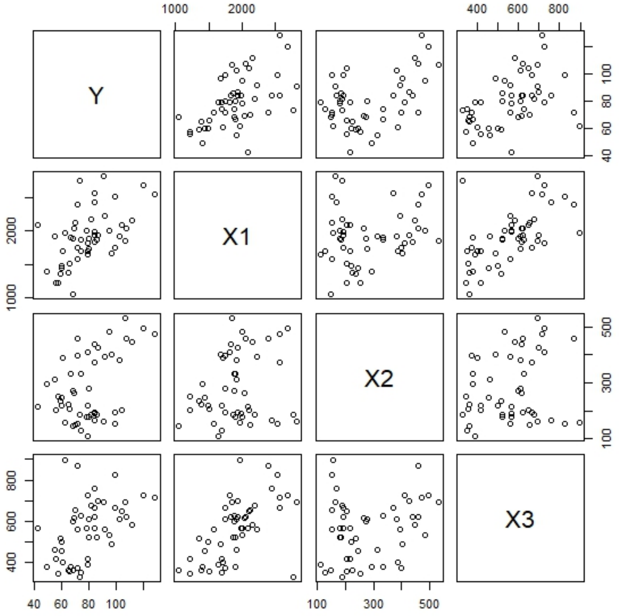
X1 and X3. There seems to be a positive relationship between X1 and X3 values. With the exception of some outliers, a lower X1 value correlates to a lower X3 value and vice versa.



X3 and X2. The scatter plot shows a slight U shape. There seems to be very little correlation between X2 and X3.

I used this code to show all of the plots together:

```
ggplot(expenditure, aes(X1, Y))+
```



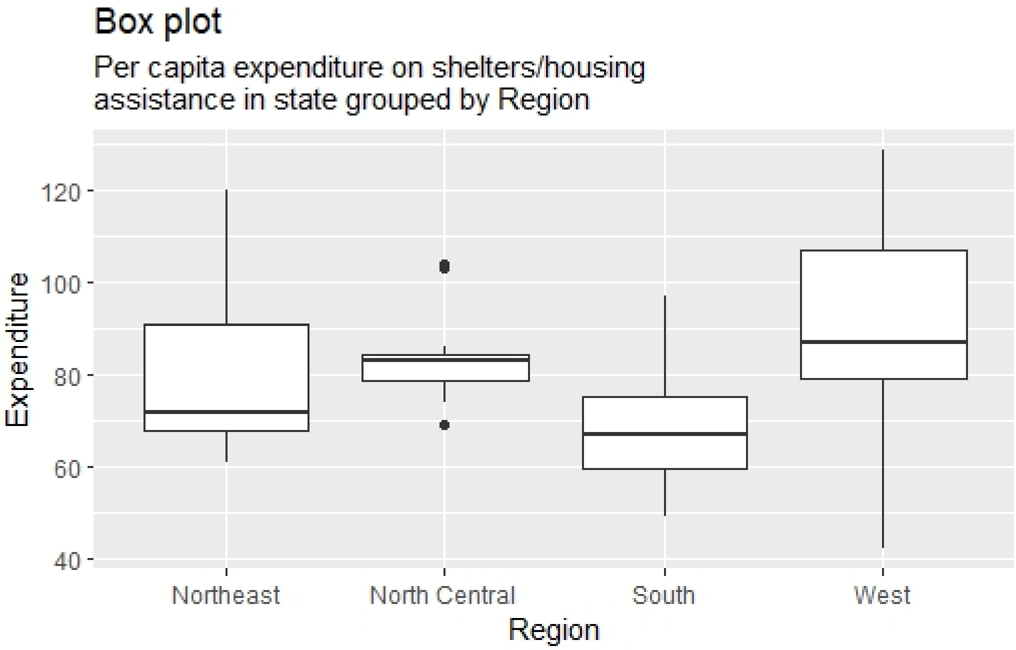
- Please plot the relationship between Y and *Region*? On average, which region has the highest per capita expenditure on housing assistance?

  I thought the best way to show this relationship was using a box plot.

  I used this code to plot this relationship:

```
ggplot(expenditure, aes(x=factor(Region), y=Y)) +
  geom_boxplot() +
  labs(title=”Box plot”,
        subtitle=”Per capita expenditure on shelters/housing
assistance in state grouped by Region”,
        x=”Region”,
        y=”Expenditure”) +
  scale_x_discrete(labels=c(”Northeast”, ”North Central”, ”South”, ”West”))
```

Resulting in this box plot.



In order to get the means, I had to take the four subsets out of the data, and turn them into separate objects:

```r
ne_1 <- expenditure[expenditure$Region == 1,]
nc_2 <- expenditure[expenditure$Region == 2,]
so_3 <- expenditure[expenditure$Region == 3,]
we_4 <- expenditure[expenditure$Region == 4,]
```

I used the code we learnt in coding camp for this.

Then, I found their respective means:

```r
mean1 <- mean(ne_1$Y)
mean2 <- mean(nc_2$Y)
mean3 <- mean(so_3$Y)
mean4 <- mean(we_4$Y)
```

This showed the following means:

North-East: 79.44

North-Central: 83.92

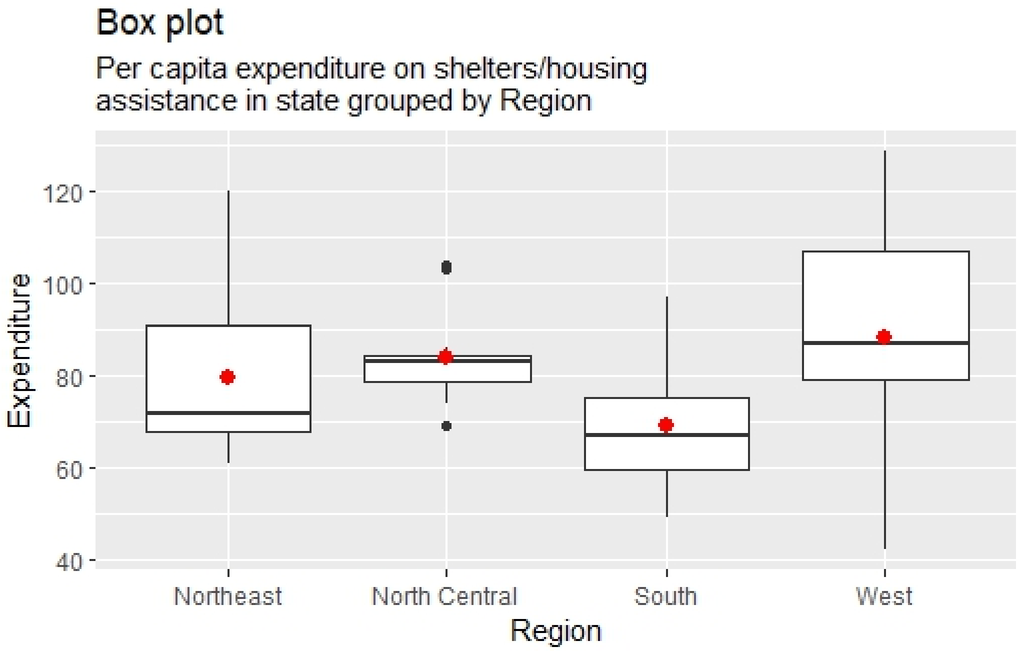
South: 69.19

West: 88.31

I used this script to plot a new box plot:

```
ggplot(expenditure, aes(x=factor(Region), y=Y)) +
  geom_boxplot() +
  stat_summary(fun.y=”mean”, color=”red”) +
  labs(title=”Box plot”,
        subtitle=”Per capita expenditure on shelters/housing
assistance in state grouped by Region”,
        x=”Region”,
        y=”Expenditure”) +
  scale_x_discrete(labels=c(”Northeast”, ”North Central”, ”South”, ”West”
    ))
```

I came upon the issue that R saw the values for region as continuous, so I had to use the code `x = factor` to ensure that these values were categorical when plotting the graph. And I also used the code `stat_summary(fun.y="mean", color="red")` to show the means. I adapted this from code I found online.



This plot shows the distribution of expenditure in different regions as well as their respective means. I chose to use ggplot, as I found it easier to work with than R's built in plot feature. As we can see, region 4, the West, has the highest mean expenditure on housing assistance.

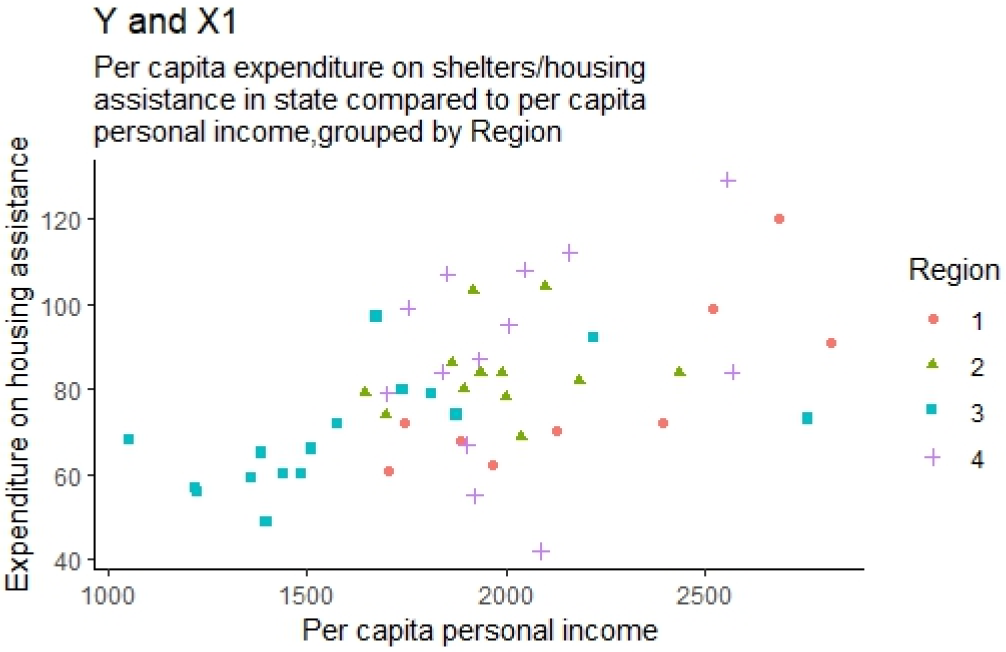
- Please plot the relationship between *Y* and *X1*? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.

A simple plot shows us a positive correlation between X1 and Y, as, in general, as Y increases, so does X1. This correlation gets weaker as the values get bigger.

```
ggplot(expenditure, aes(X1, Y)) +
  geom_point()
```

Plotting the same graph while grouping the results according to region.

```
expenditure %>%
   filter(Region %in% c(1,2, 3, 4)) %>%
  group_by(Region) %>%
ggplot(aes(x = X1, y = Y, color = factor(Region), shape = factor(Region)))+
  geom_point() +
  theme_classic() +
  labs(title=”Y and X1”,
         subtitle=”Per capita expenditure on shelters/housing
assistance in state compared to per capita
personal income,grouped by Region”,
        x=”Per capita personal income”,
        y=”Expenditure on housing assistance”,
        col = ”Region”,
        shape = ”Region”)
```


Y and X1
Per capita expenditure on shelters/housing
assistance in state compared to per capita
personal income,grouped by Region
Expenditure on housing assistance
Per capita personal income
Region
● 1
▲ 2
■ 3