Submission for Problem Set 1

Applied Stats/Quant Methods 1

Duc Minh, VU
TCD StudentID: 22996761 / UCD StudentID: 19211157

Question 1: Education

Data will first be loaded or in this case manually constructed:

```
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. **Calculation of 90% confidence interval for student IQ**

   (a) *Step 1: Calculation of the sample mean:*
   • Calculation of sample mean manually:

   ```
   ybar_mnl <- sum(y)/length(y)
   ```

   The sample mean for the IQ score calculated manually is 98.44.
   • Calculation of sample mean using R:

   ```
   ybar_r_cal <- mean(y)
   ```

   The sample mean for the IQ score calculated using R is also 98.44.
   • Using R to double-check if the manual score is the same as the R-calculated score:

   ```
   ybar_r_cal == ybar_mnl
   ```

   ```
   [1] TRUE
   ```

   R output confirms that the manually caculated mean score is equal to/the same as the R-calculated mean score.

   (b) *Step 2: Calculation of the sample variance and standard deviation:*
   • Manual calculation:

   ```
   var_mnl <- (sum((y - ybar_mnl)^2))/(length(y)-1)
   sd_mnl <- sqrt(var_mnl)
   ```

The manually caclulated value for the sample variance is 171.42333, and for the standard deviation is 13.0929.

• Calculation using R:

```
var_r_cal <- var(y)
sd_r_cal <- sd(y)
```

The R-caclulated value for the sample variance and standard deviation is also 171.42333, and 13.0929 respectively.

• Using R to double-check the results between the manual and the R calculation:

```
ybar_r_cal == ybar_mnl
```

```
[1] TRUE
```

R output confirms that the manually caculated score is equal to/the same as the R-calculated score for variance and standard deviation.

(c) *Step 3: Finding the associated t-score :*
Since the number of observation in the dataset is 25 which is less than 30, t-distribution is more appropriated as it is more robust for small sample size and violations of the normal population assumptions. Hence, the t-score will be calculated using t-table.

The t-score at 90% confidence level is 1.71

(d) *Step 4: Caclculating the confidence interval*
The confidence interval will be calculated as:

$$\bar{y} \pm \text{t-value} \times \text{Standard Errors}$$

Based on the results from the two previous steps, the 25 observations from the IQ test scores are summarized by $\bar{y} = 98.44$ and $s = 13.0929$. The estimated standard error of the sampling distribution of $\bar{y}$ can be calculated as:

$$SE(StandardErrors) = \frac{s}{\sqrt{n}} = \frac{13.0929}{25}$$

```
SE <- sd_mnl/sqrt(length(y))
```

Which gives the result of 2.6186 for the standard error using R.

```
#T-score
lower_90_t <- ybar_mnl - (t10 * SE)
upper_90_t <- ybar_mnl + (t10 * SE)
conf_int_t <- c(round(lower_90_t,2), round(upper_90_t,2))
```

```
> conf_int_t
[1]  93.96 102.92
```

The confidence interval is then computed and rounded to 2 decmimal places, which provides the results of 93.96 for the lower bound and 102.92 for the upper bound of the 90% confidence interval.

2. **Hypothesis testing for the average student IQ versus the country average IQ score**

   (a) *Assumptions*
   It will be assumed that the school counselor's sample is randomly selected and normally distributed. The IQ scores itself are quantitative data.

   (b) *Generation of hypotheses*
   Let μ denote the population of the counselor's school average IQ scores. Since she is interested in whether her school's average score is HIGHER than the country's average score of 100, the alternative hypothesis will be one-sided:

$$\mathbf{H_1:}\ \mu > \mathbf{100}$$

   The alternative states that the population mean of the school's average IQ will be higher than 100. And thus the null hypothesis will look at whether the population mean would fall into the score of 100 or less, which will be:

$$\mathbf{H_0:}\ \mu \leq \mathbf{100}$$

   (c) *Calculation of test statistic*
   Based on the value of the sample mean and estimated standard error from the previous parts, the test statistic is

$$t = \frac{\bar{y} - \mu_0}{se} = \frac{98.44 - 100}{2.6186}$$

```
test_stat <- (ybar_mnl-100)/(SE)
```

   which gives the result of -0.5957.

   (d) *Calculation of P-value*
   For n = 25, the degree freedom is n - 1 = 24. Hence, the probability (P-value) of a t-score above the observed t-score or the right-tail probability abolve -0.5967 is:

```
p.value_t_val <- pt(test_stat,df=length(y)-1,lower.tail = F)
```

   which gives the result P = 0.7215.

   (e) *Conclusion*
   With sample mean $\bar{y} = 98.44$, the P-value is 0.7215 which is very large. If μ= 100, it would not be unusual to observe $\bar{y} = 98.44$. Since P-value is larger than α (0.7215 >0.05), we fail to reject the null hypothesis H_0. Students in the counselor's chool doesn't have higher average IQ scores than students among all the schools in the country. Check P.199 in textbook

Question 2: Political Economy

First, data is imported into R along with deploying the relevant packages to produce graphs and statistic.

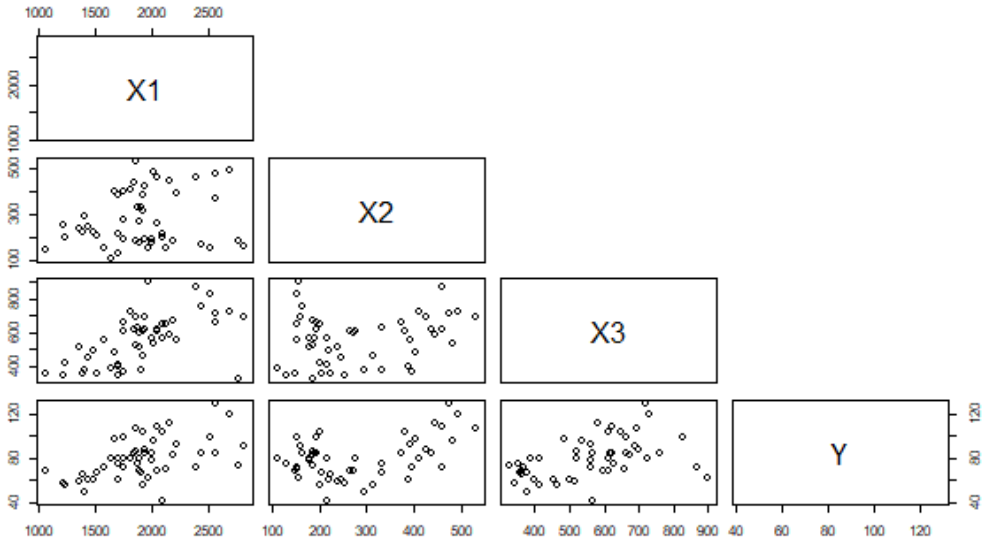
```
setwd("C:/Users/Admin/Documents/GitHub/StatsI_Fall2022/datasets")
data.exp <- read.table("expenditure.txt", header = T)
library(ggplot2)
library(dplyr)
```

1. **The relationships between Y, X1, X2 and X3**

```
pairs(~X1 + X2 + X3 + Y, data = data.exp, upper.panel = NULL)
```

Figure 1 contains the scatterplots for all of the relationships between Y, X1, X2 and X3. The subsequent sections will go into detail for each pair of the variables.

Figure 1: Scatterplots of all the relationships between Y, X1, X2 and X3 .



The correlation matrix is also obtained, so the observations from the graph can be cross-check and validated by the correlation coefficents.

```
cor(subset(data.exp,select = c(Y,X1,X2,X3))) #correlation matrix
```

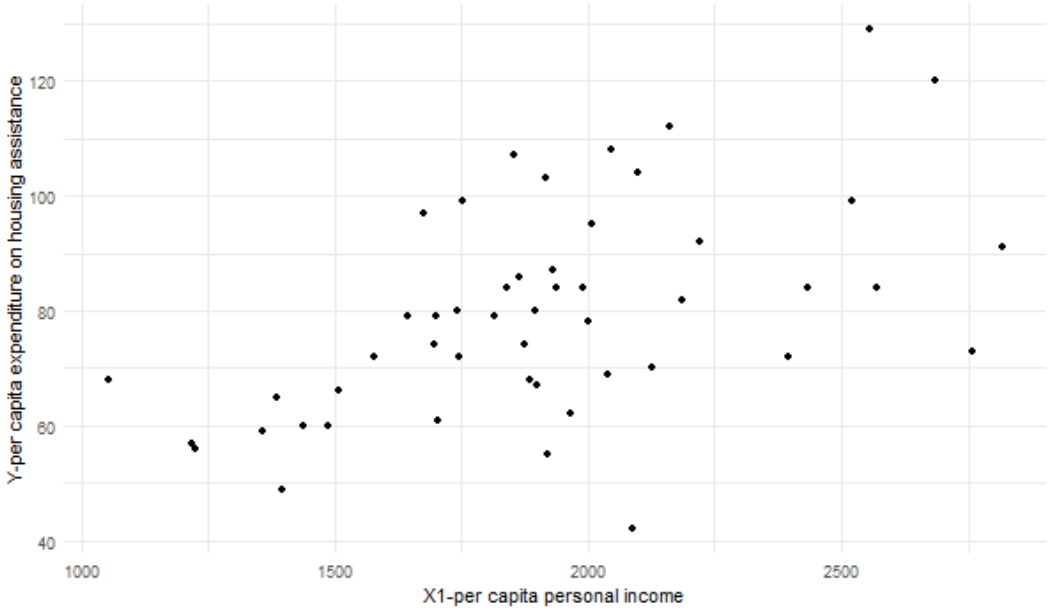
```
           Y        X1        X2        X3
Y  1.0000000 0.5317212 0.4482876 0.4636787
X1 0.5317212 1.0000000 0.2056101 0.5952504
X2 0.4482876 0.2056101 1.0000000 0.2210149
X3 0.4636787 0.5952504 0.2210149 1.0000000
```

(a) The relationship between Y and X1

```
ggplot(data.exp,aes(x = X1, y = Y)) +
  geom_point() +
  labs(x = "X1–per capita personal income",
       y = "Y–per capita expenditure on housing assistance")
```

Figure 2 displays the relationship between Per capita expenditure on housing assistance (Y) on the y-axis and Per capita income (X1) on the x-axis.

Figure 2: Scatterplot between Y and X1.



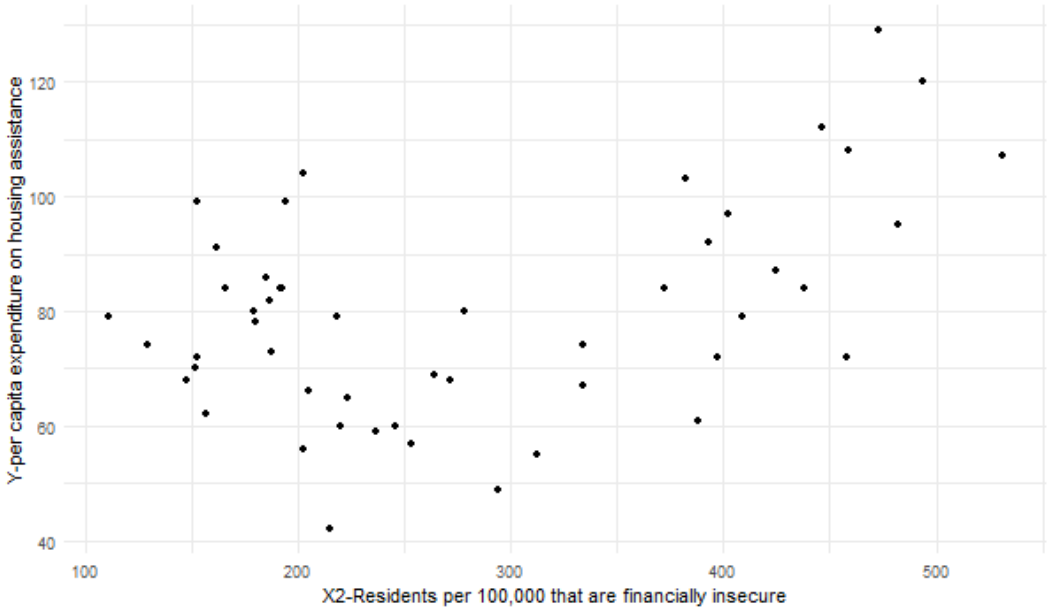
Visually, there is positive and linear relationship of weak-moderate strength between the two variables, in which Y tends to increase as X1 increases with a few potential outliers. So as Per capita income increases, Per capita expenditure also increases. This is confirmed by a moderate positive correlation coefficient of r = 0.53 from the above correlation matrix.

(b) The relationship between Y and X2

```
ggplot(data.exp,aes(x = X2, y = Y)) +
  geom_point() +
  labs(x = "X2–Residents per 100,000 that are financially insecure",
       y = "Y–per capita expenditure on housing assistance")
```

Figure 3 displays the relationship between Per capita expenditure on housing assistance (Y) on the y-axis and Number of residents per 100,000 that are "financially insecure" (X2) on the x-axis.

Figure 3: Scatterplot between Y and X2.



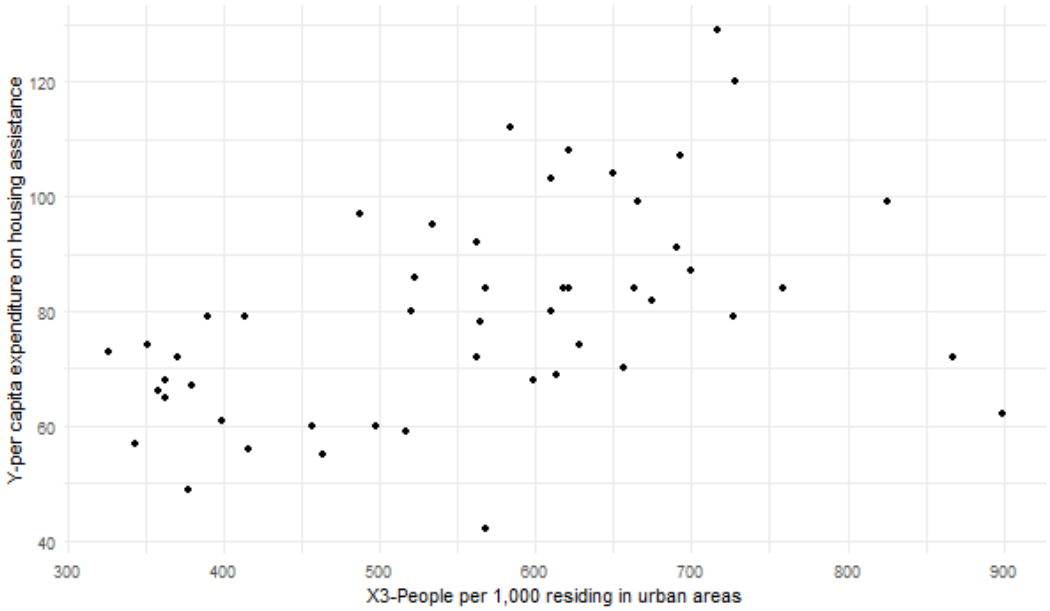
There seems to be a weak, positive and non-linear relationship between Y and X2. Per capita expenditure decrease up until around 300 residents per 100,000 before rising up again, which forms a U-shape concave up relationship. This is confirmed by a weak positive correlation coefficient of r = 0.45.

(c) The relationship between Y and X3

```
ggplot(data.exp,aes(x = X3, y = Y)) +
  geom_point() +
  labs(x = "X3–People per 1,000 residing in urban areas",
       y = "Y–per capita expenditure on housing assistance")
```

Figure 4 displays the relationship between Per capita expenditure on housing assistance (Y) on the y-axis and Number of people per thousand residing in urban areas (X3) on the x-axis.

Figure 4: Scatterplot between Y and X3.



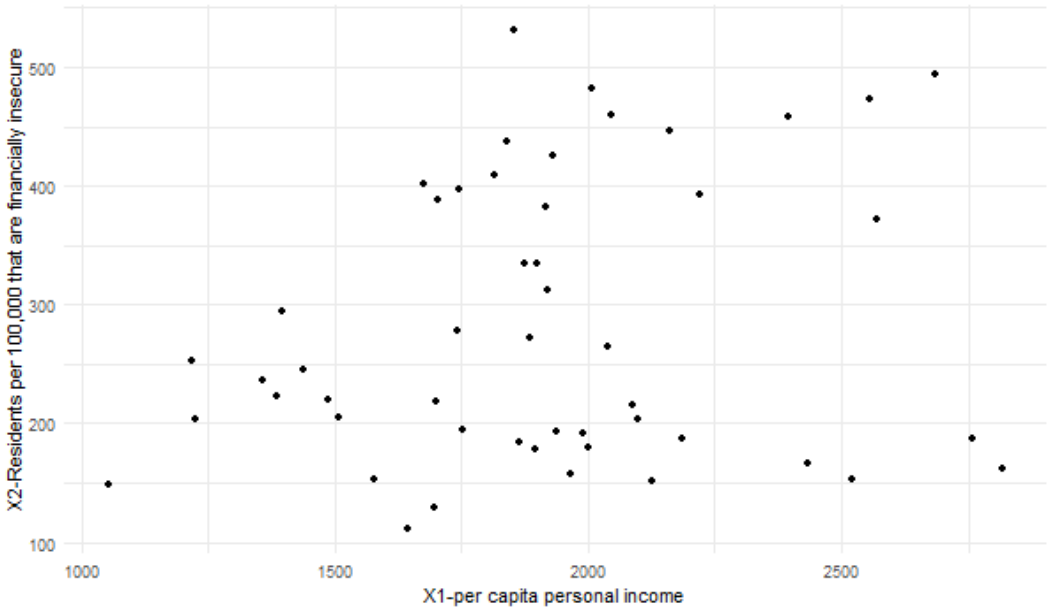
From the graph, we can observe a weak, positive and linear relationship between Y and X3, in which Y increases as X3 increases. This is confirmed by a weak positive correlation coefficient of 0.46. Moreover, there seems to be a presence of clusters in this scatter plot, with X3 = 500 (people per 1,000) as the dividing line. There is a cluster of urban areas with low level of people residing (smaller than 500 per 1,000), and a cluster of urban areas with high level of people residing (larger than 500 per 1,000).

(d) The relationship between X1 and X2

```
ggplot(data.exp,aes(x = X1, y = X2)) +
  geom_point() +
  labs(x = "X1–per capita personal income",
       y = "X2–Residents per 100,000 that are financially insecure")
```

Figure 5 displays the relationship between Per capita income (X1) on the y-axis and Number of residents per 100,000 that are "financially insecure" (X2) on the x-axis.

Figure 5: Scatterplot between X1 and X2.



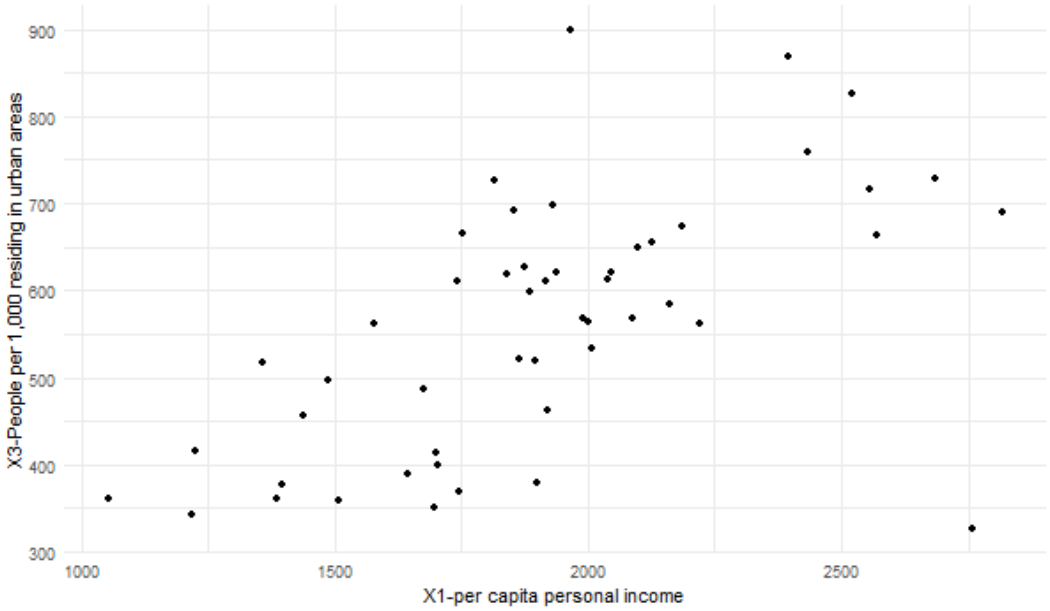
There is no clear direction for the points in this scatter plot, and no linear relationship can be detected. As such, there is no clear relationship between X1 and X2. This is confirmed by a very weak correaltion coefficient of 0.21.

(e) The relationship between X1 and X3

```
ggplot(data.exp,aes(x = X1, y = X3)) +
  geom_point() +
  labs(x = "X1–per capita personal income",
       y = "X3–People per 1,000 residing in urban areas")
```

Figure 6 displays the relationship between Per capita income (X1) on the y-axis and Number of people per thousand residing in urban areas (X3) on the x-axis.

Figure 6: Scatterplot between X1 and X3.



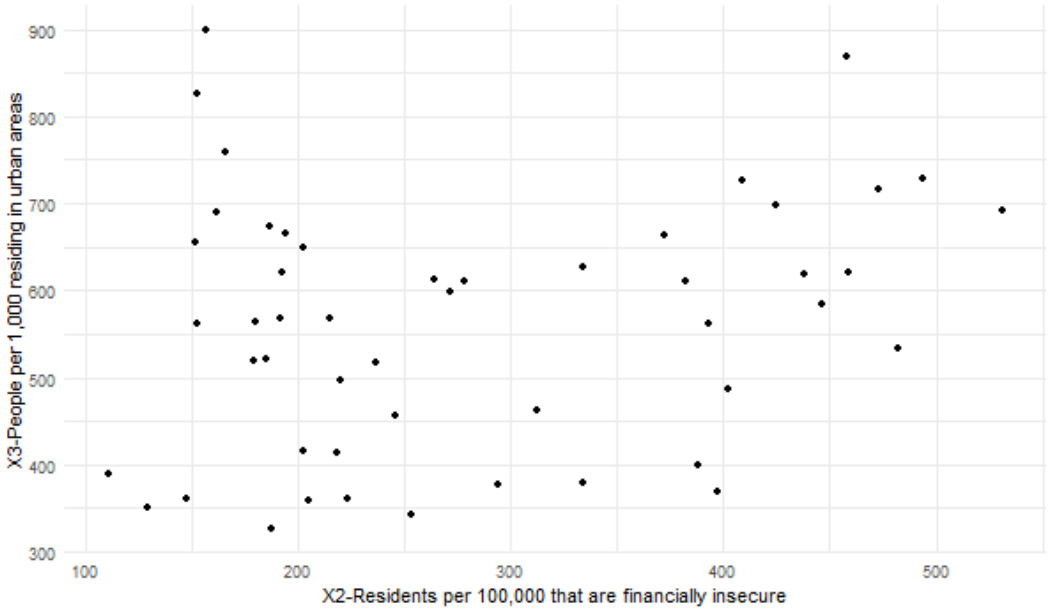
A positive, linear relationship of moderate strength can be depicted from Figure 6, as X1 increases when X3 increases with some potential outliers. So, people with higher capita income in a state tend to live more in the urban areas of that state. This relationship is confirmed by a moderate correlation coefficient of 0.59. Some potential clusters are also observed in this scatter plot, with 3 groups. The most visible group is state with "high" (in relation to this dataset) capita income, which is above $2,250 threshold. The seperation between the "middle" and "low" income group is less visible, with $1,750 as the separation line between these two clusters.

(f) The relationship between X2 and X3

```
ggplot(data.exp,aes(x = X2, y = X3)) +
  geom_point() +
  labs(x = "X2–Residents per 100,000 that are financially insecure",
       y = "X3–People per 1,000 residing in urban areas")
```

Figure 7 displays the relationship between Number of residents per 100,000 that are "financially insecure" (X2) on the y-axis and Number of people per thousand residing in urban areas (X3) on the x-axis.

Figure 7: Scatterplot between X2 and X3.



There is no clear relationship between X2 and X3 from this scatter plot which is confirmed by a low correlation coefficient of 0.221.

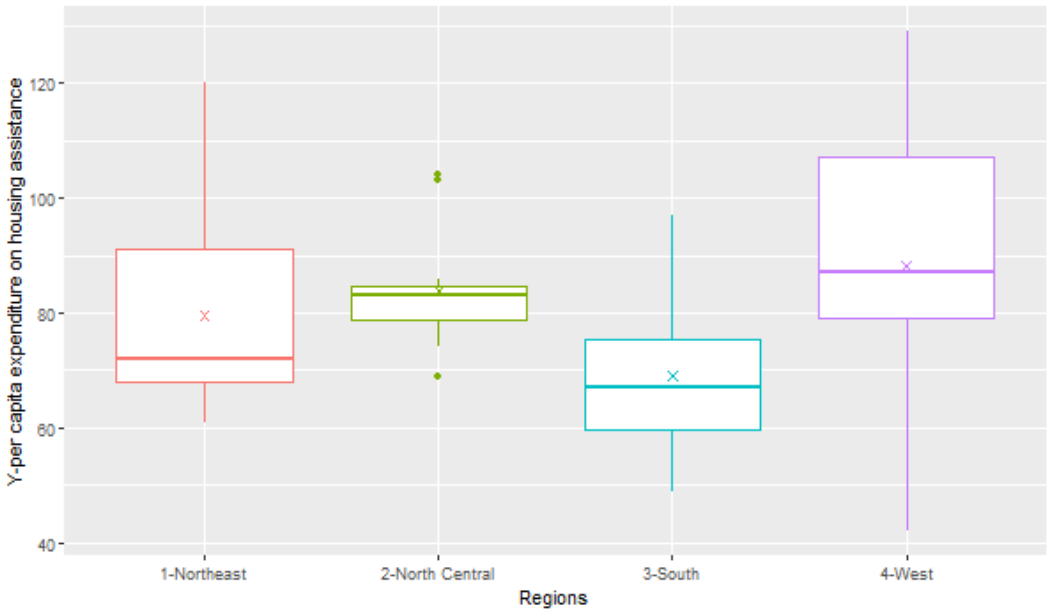
2. **The relationships between Y and different regions**

```
# Assigning the region name to the the region code
data.exp$Region_name <- ifelse(data.exp$Region == 1, "1-Northeast",
                               ifelse(data.exp$Region == 2, "2-North
    Central",
                                      ifelse(data.exp$Region == 3, "3-
    South",
                                             ifelse(data.exp$Region == 4,
     "4-West", "NA"))))
```

```
#Plotting the data
ggplot(data.exp,
       aes(x = Region_name, y = Y, group = Region_name, color = Region_
    name)) +
  geom_boxplot() + stat_summary(fun=mean,
                                geom='point',
                                shape=4, size = 2,
                                # show.legend = FALSE
                                aes( colour = "Mean")
                                ) +
  labs(x = "Regions",
       y = "Y-per capita expenditure on housing assistance",
       color = "Region name") +
       theme(legend.position="none")
```

Figure 8 displays the relationship between Y and different regions of states in the US in the form of a box plot which shows the distribution of per capita spending expenditure according to each region. The average per capita spending is also calculated and depicted as the symbol "x' in the plot.

Figure 8: Box plot between Y and different Regions.




```r
#Calculating summary for each region's per capita spending
region_summary <- data.exp %>%
  group_by(Region, Region_name) %>%
  summarise(ave_per_capita = mean(Y), max_per_cap = max(Y), min_per_cap =
    min(Y))
```

Region	Region_name	ave_per_capita	max_per_cap	min_per_cap
<int>	<chr>	<dbl>	<int>	<int>
1	1 1-Northeast	79.4	120	61
2	2 2-North Central	83.9	104	69
3	3 3-South	69.2	97	49
4	4 4-West	88.3	129	42

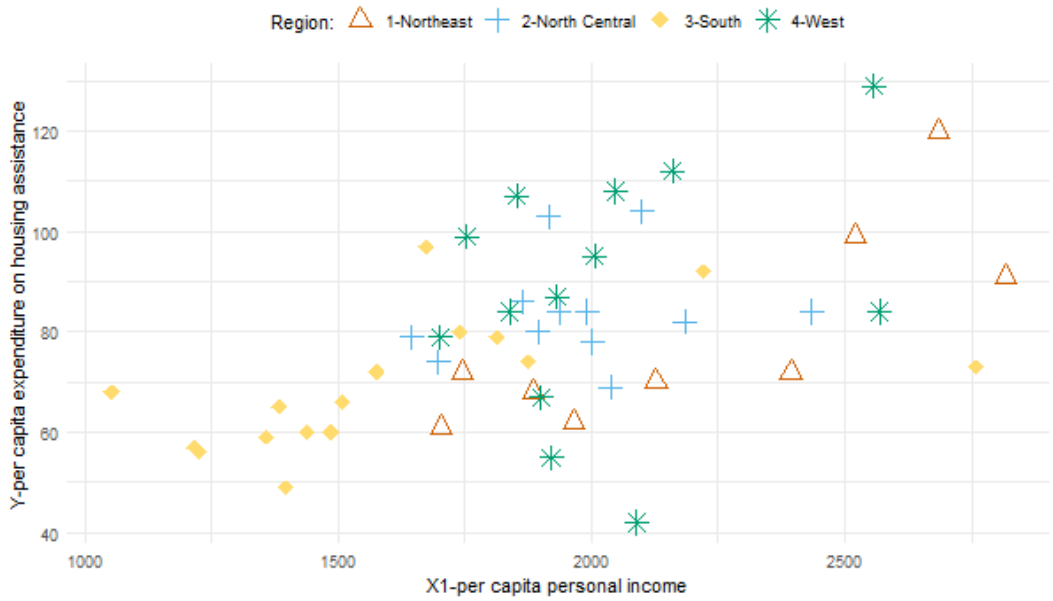
From the figure, we can then see that states in the **West** has the highest average above the three other states. This is confirmed by the calculated statistic as it has the highest mean for the per capita spending expenditure on housing assistance of 88.3.

3. **The relationships between Y and X1 with the inclusion of region**

```
ggplot(data.exp,aes(x = X1, y = Y)) +
  geom_point(aes(color = Region_name, shape = Region_name), size = 4) +
  theme(legend.position = "top") +
  scale_color_manual(values = c("#D16103", "#56B4E9", "#FFDB6D", "#009E73
    ")) +
  scale_shape_manual(values=c(2, 3, 18, 8)) +
  labs(colour = "Region:",
       shape = "Region:",
       x = "X1-per capita personal income",
       y = "Y-per capita expenditure on housing assistance")
```

As described earlier in section 1.(a) of this question, Y and X1 as a weak-moderate positve and linear relationship, in which the higher the per capita income in a state, the more that state will spend on its housing assistance per capita.

Figure 9: Scatterplot between Y and X1 for different regions.



With the inclusion of data on states' regions, we can then see how the relationships between Y and X1 differs for each region. For example: the linear positive relationship can be observerd more clearly for the South and Northeast regions, while no clear relationship can be seen for North Central and West regions.