Problem Set 1

Applied Stats/Quant Methods 1

Due: September 30, 2024

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:

```r
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.

```r
### Code
n <- length (na.omit (y))
t90 <- qt((1 - 0.90) / 2, df = n - 1, lower.tail = FALSE)
sample_mean <- mean(y, na.rm = TRUE )
sample_sd <- sd (y, na.rm = TRUE )
lower_90 <- sample_mean - (t90 * (sample_sd/sqrt(n)))
```

```
upper_90 <- sample_mean + (t90 * (sample_sd/sqrt(n)))
confint90 <- c ( lower_90 , upper_90)
confint90

### Results are: 93.95993 102.92007
```

The 90% confidence interval for the average student IQ in the school is: 93.95993 and 102.92007. This means that this confidence interval range contains the true parameter at least 90% of the time if we were to repeat the experiment or sampling process a large number of times.

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 test the school counselor assumption, we are going to follow the *five step null hypothesis testing*. We start off by exploring our data. From running the previous code, we observe that our data is numerical and discrete. We also see that our data contains 25 observations. This will determine what method of hypothesis testing we will need. We will also assume that our sampling method is determined by randomisation.

```
n
[1] 25
y
[1] 105  69  86 100  82 111 104 110  87 108  87  90  94
113 112  98  80  97  95 111 114  89
[23]  95 126  98
```

   The second step is establishing our hypothesis. The school counselor believes the average student IQ in her school might be **higher** than the average IQ score. That is H = 1. Our null hypothesis is, therefore, that the average student IQ in that school is **not higher** than the average IQ score (100).

   To test whether the school conselor intuition is true, we are going to follow the *proof by contradiction* procedure. We need to disprove our null hypothesis in order to show that the data we are observing would rarely occur if H = 0 were true. Given that our sample data has less than 30 observations, we run a t-test. Since we only want to test whether students' IQ is lower than the average student IQ, we run a one-sided t-test. We specify that we are interested in knowing if the mean of students' IQ is higher than the average mean; hency why we write ”alternative = ’greater’”.

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

```
One Sample t-test

   data:  y
   t = -0.59574, df = 24, p-value = 0.7215
   alternative hypothesis: true mean is greater than 100
   95 percent confidence interval:
   93.95993      Inf
   sample estimates:
   mean of x
   98.44
```

Our significance level is 0.05. The p-value (0,7215) we get from the t-test, however, is larger than our significance level. That is, our p-value is not statistically relevant. Therefore, we do not have enough evidence to reject H = 0, which stated that the students' IQ from the school counselor was not higher than the average IQ (100). We cannot, thus, sustain the school counselor's assumption that average student IQ in her school is higher than the average IQ score.

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.

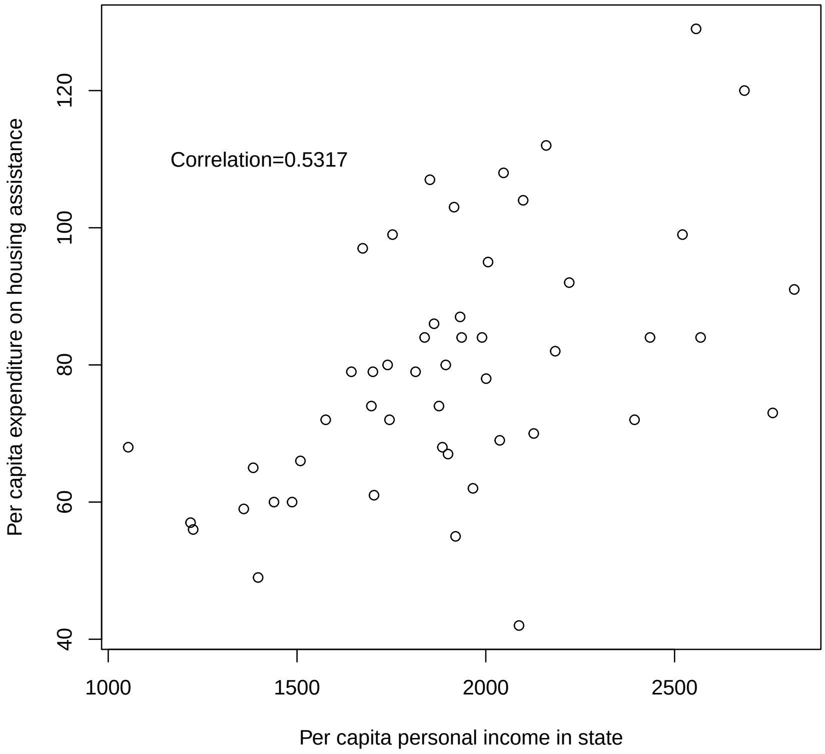
`State`	*50 states in US*
`Y`	*per capita expenditure on shelters/housing assistance in state*
`X1`	*per capita personal income in state*
`X2`	*Number of residents per 100,000 that are ”financially insecure” in state*
`X3`	*Number of people per thousand residing in urban areas in state*
`Region`	*1=Northeast,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

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

- 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)?

```
pdf(”plot_1.pdf”)
plot(expenditure$X1,
     expenditure$Y,
     xlab=”Per capita personal income in state”,
     ylab=”Per capita expenditure on housing assistance”)
text(1400, 110, sprintf(”Correlation=%s”, round(cor(expenditure$Y,
   expenditure$X1), 4)))
dev.off()
```

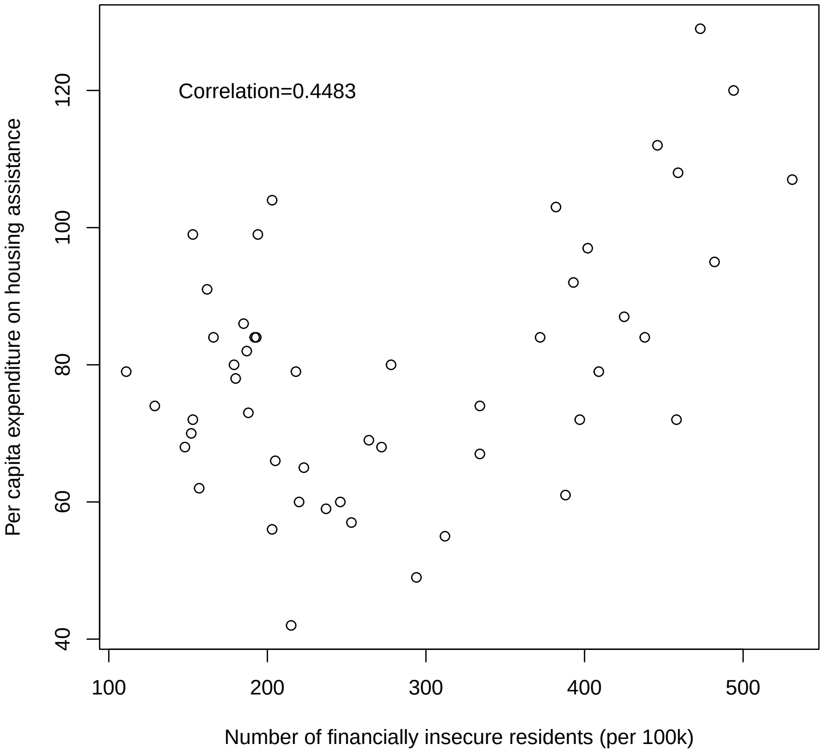
Figure 1: The relationship between personal income (X1) and expenditure on housing assistance (Y).



From this first figure, in which we visualise the relationship between the personal income in a state (per capita) and the expenditure on housing assistance (per capita), we could gather there is a slightly relevant correlation between these two variables. We can see that, generally, the observations on our data draw a positive correlation: in broad terms, increase on X1 seems to relate with increase on Y. The correlation value we get is 0.5317, which denotes there is a moderate positive correlation between variables, as we firstly assessed. That confirms us that when one variable increases (X1), the other one also tends to increase (Y). However, this does not necessarily mean one causes the increase of the other; rather, it shows us that the variation of X1 is relatively correlated with the variation in Y.

```r
pdf("plot_2.pdf")
plot(expenditure$X2,
     expenditure$Y,
     xlab="Number of financially insecure residents (per 100k)",
     ylab="Per capita expenditure on housing assistance")
text(200, 120, sprintf("Correlation=%s", round(cor(expenditure$Y,
   expenditure$X2), 4)))
dev.off()
```

Figure 2: The relationship between financially insecure residents (X2) and expenditure on housing assistance (Y).

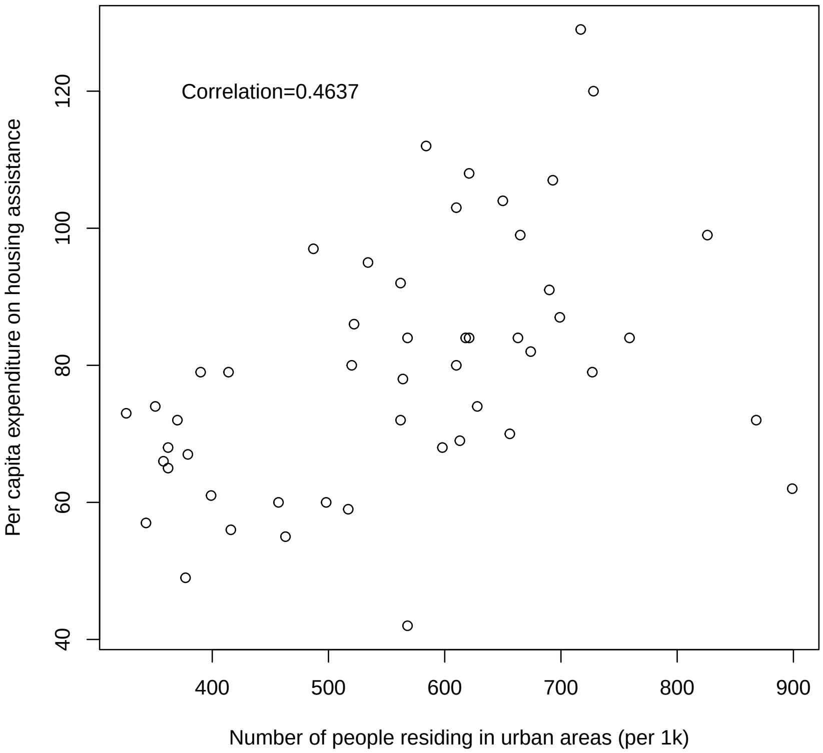


As per this second figure, in which we plot the relationship between the number of residents that feel ”financially insecure” (per 100k habitants) in a state (X2) and the expenditure on housing assistance in a state (per capita) (Y), we observe that the correlation is not as strong as the previous figure. In this case, we can also consider there is a positive correlation, i.e., when one variable increases, the other does so too, but the strenght of this correlation is slightly weaker than the previous one. Our correlation value (0.4483) asserts this observation: the observed variation of one variable positively associates with observed variation of the other, although more moderate than the previous figure.

```
pdf("plot_3.pdf")
plot(expenditure$X3,
      expenditure$Y,
      xlab="Number of people residing in urban areas (per 1k)",
      ylab="Per capita expenditure on housing assistance")
text(450, 120, sprintf("Correlation=%s", round(cor(expenditure$Y,
    expenditure$X3), 4)))
```

```
dev.off()
```

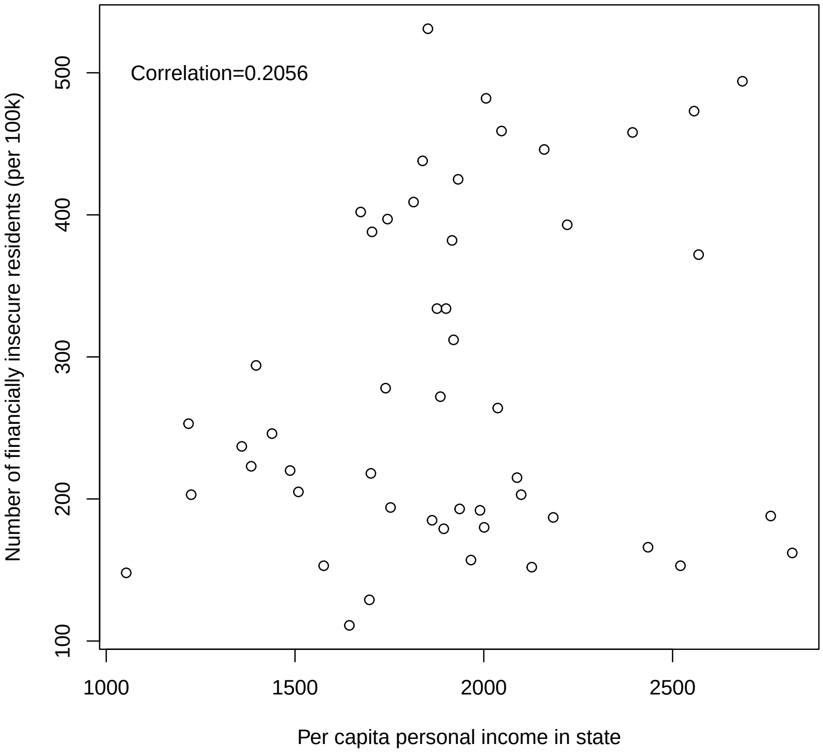
Figure 3: The relationship between number of residents in urban areas (X3) and expenditure on housing assistance (Y).



In this third figure, in which we plot the relationship between the number of people residing in urban areas (per 100k) and the expenditure per capita spent on housing assistance, we also observe a positive correlation between these variables. The correlation value is 0.4637, which indicates that the variation of one variable positvely relates and co-varies with the variation of the other. We could say that the variation we see in number of people living in urban areas positively correlates/associates with the expenditure spent on housing assistance in a state.

```
pdf("plot_4.pdf")
plot(expenditure$X1,
     expenditure$X2,
     xlab="Per capita personal income in state",
     ylab="Number of financially insecure residents (per 100k)")
text(1300, 500, sprintf("Correlation=%s", round(cor(expenditure$X1,
   expenditure$X2), 4)))
dev.off()
```

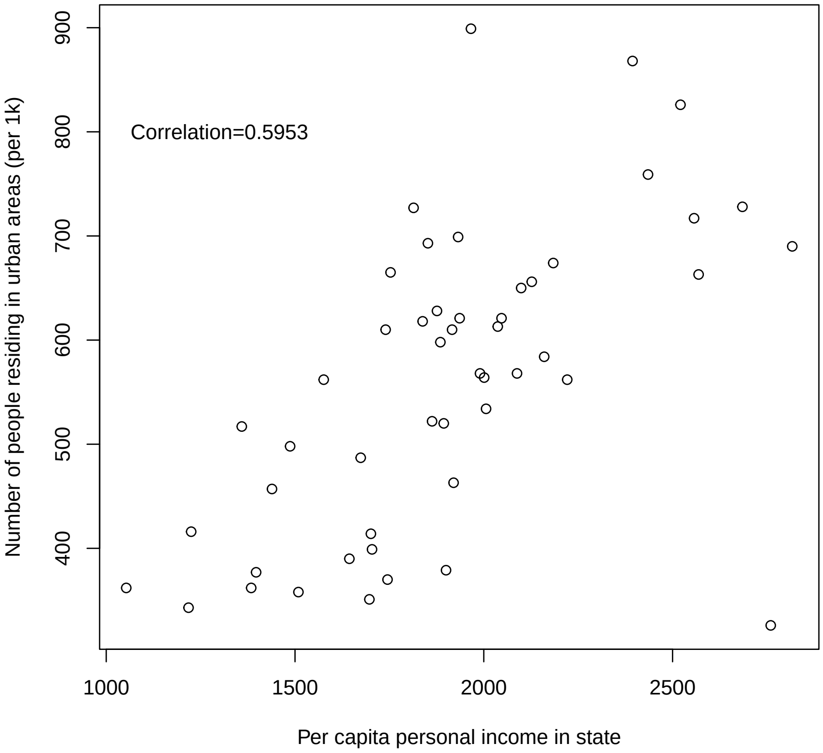
Figure 4: The relationship between personal income (per capita) (X1) and the number of financially insecure residents (X2).



In the fourth figure, we plot the correlation between personal income and the number of people who are financially insecure. According to the correlation value we get (0.2056), these two variables do not share a very strong association relationship; the variance we observe in one variable slightly and weakly relates to the variation we observe in the other one, but this is not strong enough to assert there is a strong association between the two.

```
pdf("plot_5.pdf")
plot(expenditure$X1,
     expenditure$X3,
     xlab="Per capita personal income in state",
     ylab="Number of people residing in urban areas (per 1k)")
text(1300, 800, sprintf("Correlation=%s", round(cor(expenditure$X1,
   expenditure$X3), 4)))
dev.off()
```

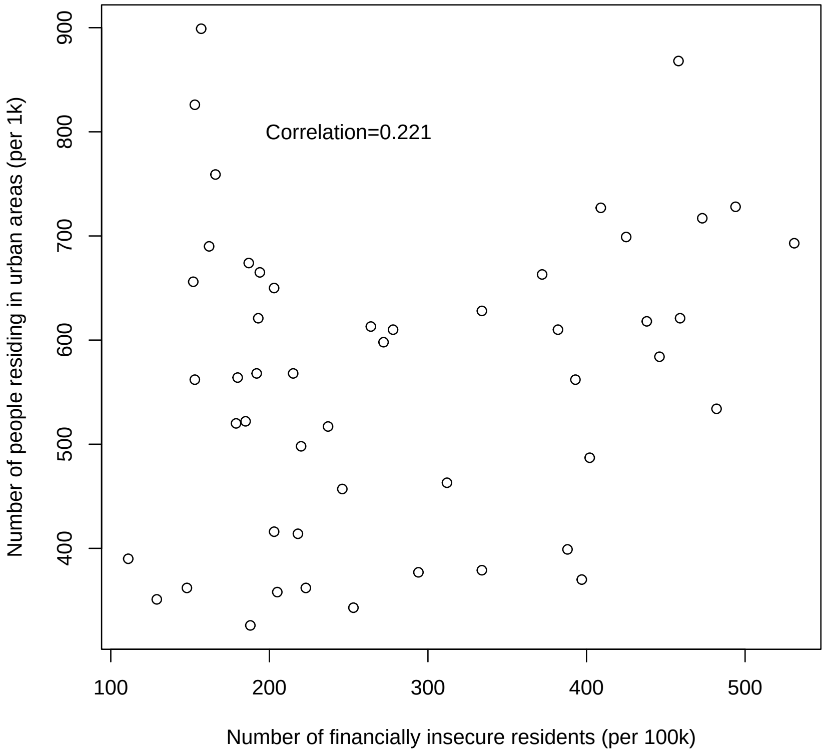
Figure 5: The correlation between personal income (per capita) (X1) and number of people residing in urban areas (X3)



In the fifth figure, in which we plot the relation between the personal income (X1) and the number of people residing in urban areas (X3), we observe a much stronger correlation between the two variables. It is, in fact, the strongest correlation we have plotted so far. This means that every variation we observe on the personal income variable is positively associated with the variation observed in the number of people residing in urban areas.

```r
pdf("plot_6.pdf")
plot(expenditure$X2,
     expenditure$X3,
     xlab="Number of financially insecure residents (per 100k)",
     ylab="Number of people residing in urban areas (per 1k)")
text(250, 800, sprintf("Correlation=%s", round(cor(expenditure$X2,
    expenditure$X3), 4)))
dev.off()
```

Figure 6: The correlation between number of ’financially insecure’ people (X2) and number of people residing in urban areas (X3)



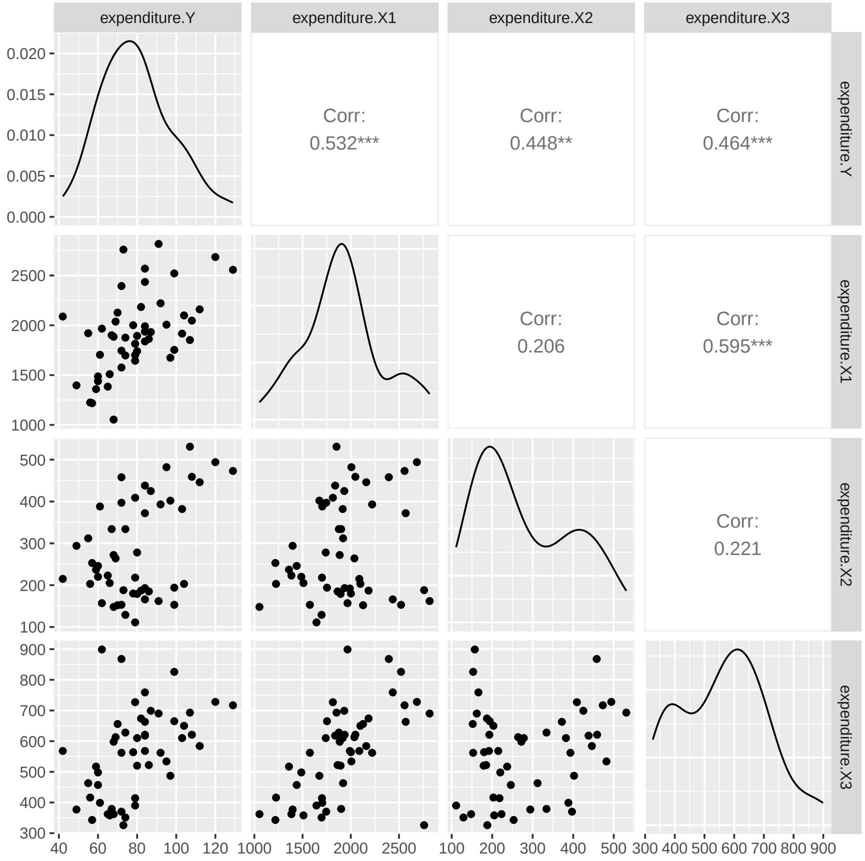
This sixth figure shows the correlation between the number of people who are financially insecure and the number of people who reside in urban areas. Unlike the other correlation, these variables show a very weak correlation of 0.221: the variation seen in one variable associates, although very weakly, to the variation observed in the other variable. However, this association is very frail.

In order to put into perspective all the plots shown previously, we have elaborated the following matrix. As previously stated, the stronger correlations are between variable X1 and X3 and X1 and Y. The weakest are between variables X1 and X2, and X2 and X3. Visualising and understanding the level of correlation between variables is important to assess which variables are necessary if we are interested in running a regression. By avoiding variables that are too closely correlated, we can avoid problems of multicollinearity (King, Keohane, and Verba, 1994).

```
pdf(”plot_7.pdf”)
ex_mat <- data.frame(expenditure$Y, expenditure$X1, expenditure$X2,
    expenditure$X3)
```

```r
cor_matrix <- cor(ex_mat)
cor_matrix
ggpairs(ex_mat)
dev.off()
```

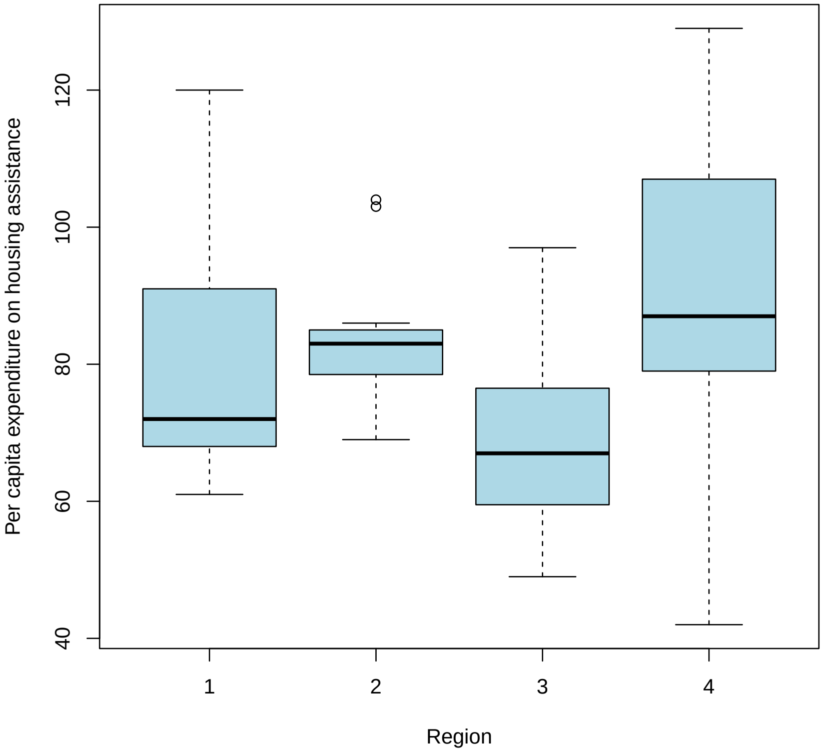
Figure 7: Matrix of all plotted variables (Y, X1, X2, X3)



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

```r
pdf("plot_8.pdf")
boxplot(Y ~ Region,
        data = expenditure,
        xlab = "Region",
        ylab = "Per capita expenditure on housing assistance",
        col = "lightblue")
dev.off()
```

Figure 8: Boxplot of relationship between per capita expenditure and housing assistance by region



The previous boxplot provides us with information regarding the mean, distribution of data, the maximum and minimum values, etc. We observe that the region with the highest average per capita expenditure on housing assistance is region 4, which corresponds with *the West*. This region has a mean slightly below 90. We also see that this region has a larger spread data; it has the highest value of expenditure; and the lowest, in comparison with other regions. The second region with the highest second average expenditure would be region 2, which corresponds with the *North Central* region. This region, however, shows the lowest dispersion in data; and it has two outliers. The region with the lowest average expenditure on housing assistance is region 3, which corresponds to *the South*.

- 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.

```
pdf(”PLOT_9.pdf”)
plot(expenditure$X1,
     expenditure$Y,
     col=expenditure$Region,
     xlab="Per capita personal income in state",
     ylab="Per capita expenditure on housing assistance in state")

legend("topleft",
       legend=c("South", "Northeast", "North Central", "West"),
       col=c("green", "black","red", "blue"),
       pch=1) # Marker type (1 is default)
dev.off()

# Or
expenditure$Region_names <- factor(expenditure$Region,
                             levels = c(1, 2, 3, 4),
                             labels = c("Northeast", "North Central", "South", "West"))

region_colors <- c("Northeast" = "blue",
                   "North Central" = "green",
                   "South" = "red",
                   "West" = "orange")

pdf("plot_10.pdf")
ggplot(expenditure, aes(x = X1, y = Y, color = Region_names, shape = Region_names)) +
  geom_point(size = 2) +
  geom_text(aes(label = STATE), hjust = 0, vjust = 0, size = 2.5) +
  scale_color_manual(values = region_colors) +
  scale_shape_manual(values = c(16, 17, 18, 19)) +
  labs(x = "Per capita personal income in state",
       y = "Per capita expenditure on housing assistance in state",
       color = "Region",
       shape = "Region") +
  theme_minimal()
dev.off()
```

Figure 9: The relationship between personal income and expenditure on housing assistance

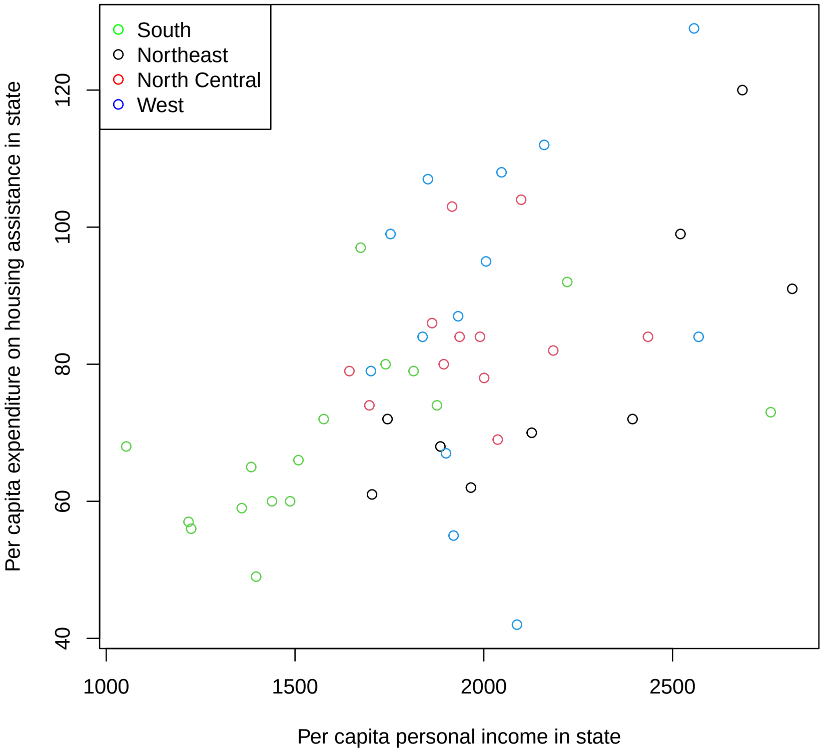
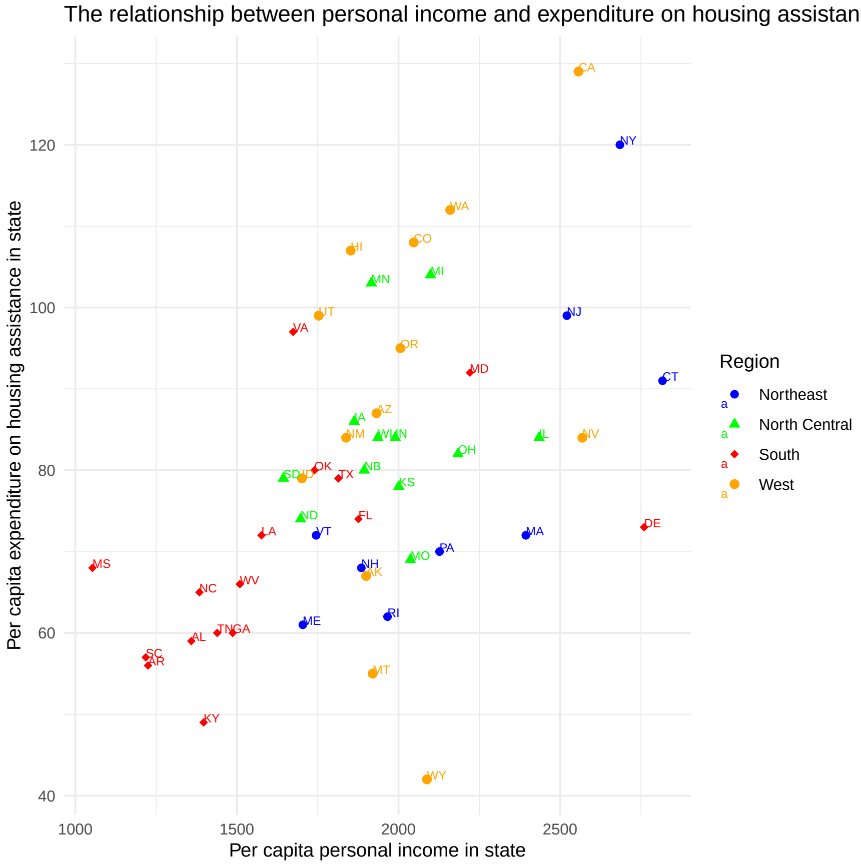


Figure 10: Relationship between personal income and expenditure on housing assistance (with ggplot library)



Overall, we see that those regions with the highest per capita personal income also have the highest expenditure on housing assistance. That is, the higher the personal income in a state, the more is spent on housing assistance. As we mentioned before, there is a relatively strong correlation between these two phenomena. In addition, we observe that the Northeast regions, such as New York, Massachusetts and Connecticut, have the highest personal capita and highest expenditure on housing assistance. On the other hand, Southern regions have the lowest per capita personal income and, at the same time, spent the lowest on housing assistance. It is also interesting to report the behaviour of Western states: as we mentioned before, this region has a very highly spread data; a proof of this is that we see states, such as California, which have very high values in both personal income and expenditure, and other Western states, such as Montana, that present more moderate values. Lastly, the North Central region, as we mentioned in the previous plot, presents very clustered data, centred at the middle of the values. However, this region also presents some outliers.