# Problem Set 1

Applied Stats / Quant Methods 1

Lucas de Melo Prado / Due: October 3, 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 3, 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.*

   To find the confidence interval, first we calculate the sample mean ($\bar{y}$), the standard deviation ($s$), and the standard error ($se$):

```r
# Sample mean:
yBar <- mean(y)
```

$$\bar{y} = \frac{\sum y}{n} = \frac{105 + 69 + 86 + \dots}{25} = 98.4$$

```r
# Standard deviation:
s <- sqrt(sum((y-yBar)^2)/(length(y)-1))
```

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}} = \sqrt{\frac{(105 - 98.4)^2 + (69 - 98.4)^2 + \dots}{25 - 1}} = 13.1$$

```r
# Standard error:
se <- s/sqrt(length(y))
```

$$se = \frac{s}{\sqrt{n}} = \frac{13.1}{\sqrt{25}} = 2.6$$

   Then, considering the confidence interval of 90%, it is possible to find the z-score using the `qnorm` function in R:

```r
# Z-score:
z90 <- qnorm((1-0.9)/2, lower.tail = FALSE)
```

$$z = 1.64$$

   Finally, using these statistics, we can calculate the lower and upper values of the confidence interval:

```r
# Confidence interval lower value:
CIlower <- yBar-z90*se

# Confidence interval upper value:
CIupper <- yBar+z90*se
```

$$C.I. = \bar{y} \pm z \cdot se = 98.4 \pm 1.64 \cdot 2.6 = (94.1, 102.7)$$

   As a result, the 90% confidence interval for the average student IQ in the school is between **94.1** and **102.7**.

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 perform the hypothesis test, we need to follow *five* steps.

   **Step 1. Assumptions.** We are working with a quantitative variable. The 25 sample students' IQ scores were collected randomly.

   **Step 2. Hypotheses.** Our null hypothesis is the average IQ score among all the schools in the country:

```
# Null hypothesis:
mu <- 100
```

$$H_0 : \mu = 100$$

   Because the school counselor wants to check if the average student IQ in her school is *higher* than the average IQ score among all schools in the country, we perform a *one-sided* test, focusing on the right tail. So, our alternative hypothesis is:

$$H_a : \mu > 100$$

   **Step 3. Test statistic.** Taking into consideration the size of our sample ($n < 30$), we decide to use the *t-test*. For the sample mean ($\bar{y}$) and the standard error ($se$), we use the values we have already calculated in the previous question.

```
# T-test for alpha = 0.05:
t <- (yBar - mu)/se
```

$$t = \frac{\bar{y} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{y} - \mu}{se} = \frac{98.4 - 100}{2.6} = -0.62$$

   **Step 4. P-value.** As $H_a : \mu > 100$, we look at the right tail to establish p-value:

```
# P-value (right tail):
p <- pt(t, df = length(y)-1, lower.tail = FALSE)
```

$$p = 0.72$$

   **Step 5. Conclusion.** P-value ($p = 0.72$) is much higher than the significance level ($\alpha = 0.05$). Therefore, *we do not have enough evidence to reject the null hypothesis* ($H_0 : \mu = 100$).

# 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`.

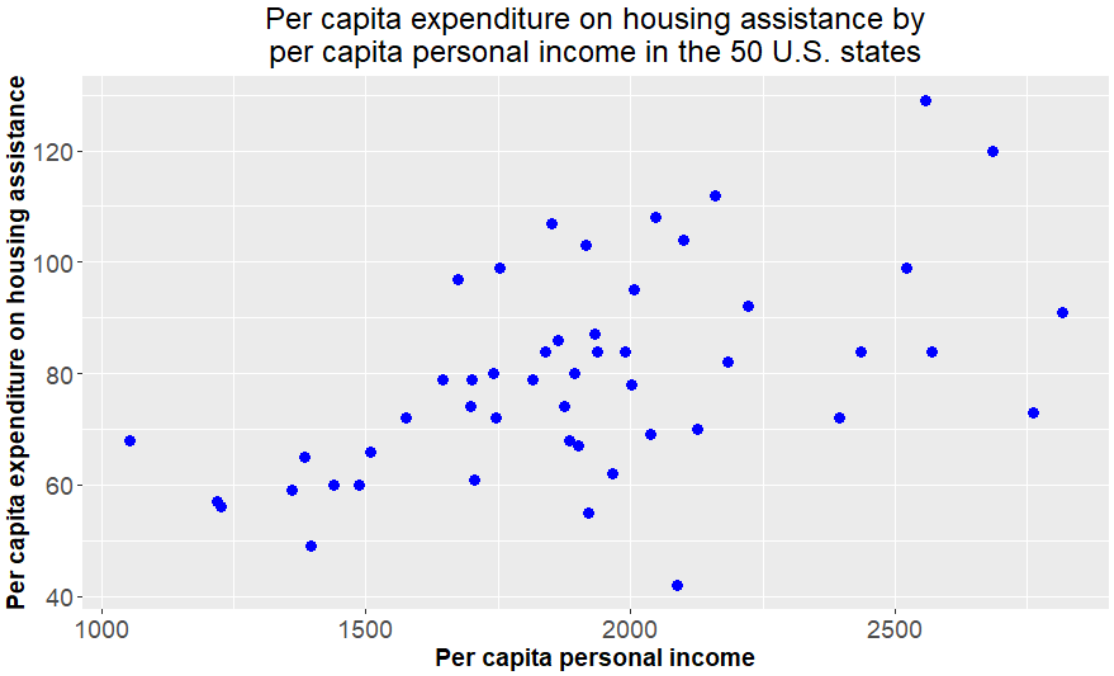
```
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2021/main/datasets/expenditure.txt", header=T)
```

- *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)?*

### 1. Relationship between $Y$ and $X1$:

```
# (1) Scatterplot Y–X1:
ggplot(data=expenditure, mapping = aes(x = X1, y = Y)) +
  geom_point(color="blue", size=3) +
  theme(plot.title = element_text(hjust = 0.5, size = 18),
        axis.text = element_text(size = 14), axis.title = element_text(
    size = 14, face = "bold")) +
  ggtitle("Per capita expenditure on housing assistance by
per capita personal income in the 50 U.S. states") +
  xlab("Per capita personal income") +
  ylab("Per capita expenditure on housing assistance")
```

Figure 1:

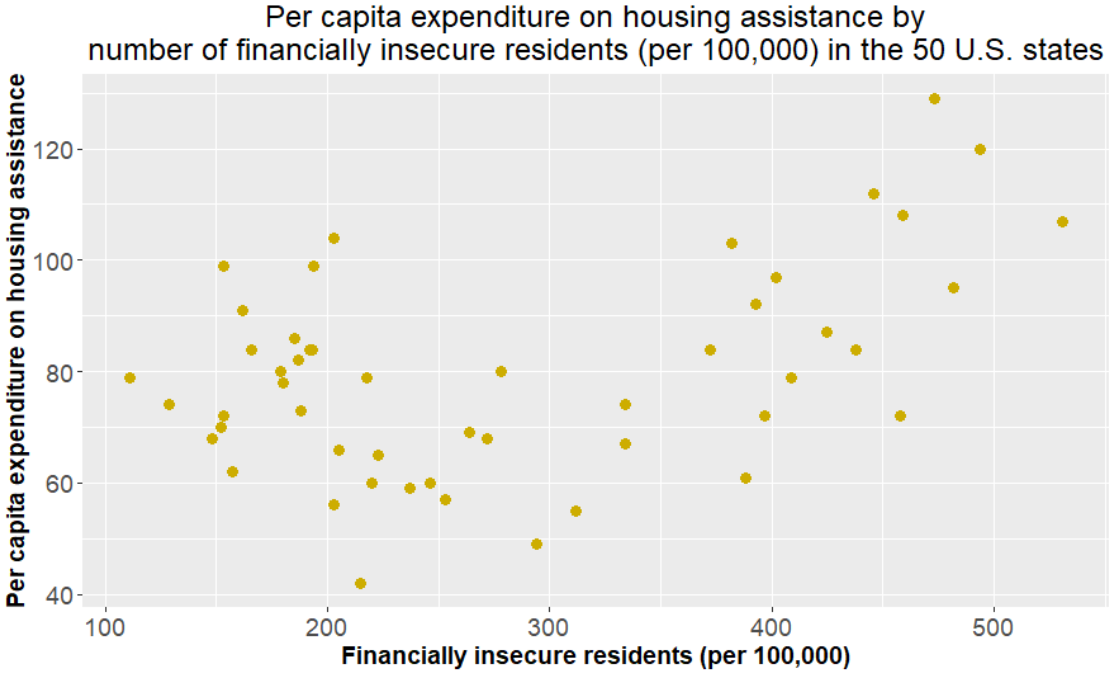



In figure 1, data points scatter from the bottom-left corner of the graphic to the top-right corner. There seems to be some outliers both in the top and in the bottom of the graphic. Even so, the pattern suggests a positive linear relationship between $Y$ and $X1$.

**2. Relationship between $Y$ and $X2$:**

```
# (2) Scatterplot Y–X2:
ggplot(data=expenditure, mapping = aes(x = X2, y = Y)) +
  geom_point(color="gold3", size=3) +
  theme(plot.title = element_text(hjust = 0.5, size = 18),
        axis.text = element_text(size = 14), axis.title = element_text(
    size = 14, face = "bold")) +
  ggtitle("Per capita expenditure on housing assistance by
number of financially insecure residents (per 100,000) in the 50 U.S.
    states") +
  xlab("Financially insecure residents (per 100,000)") +
  ylab("Per capita expenditure on housing assistance")
```

Figure 2:

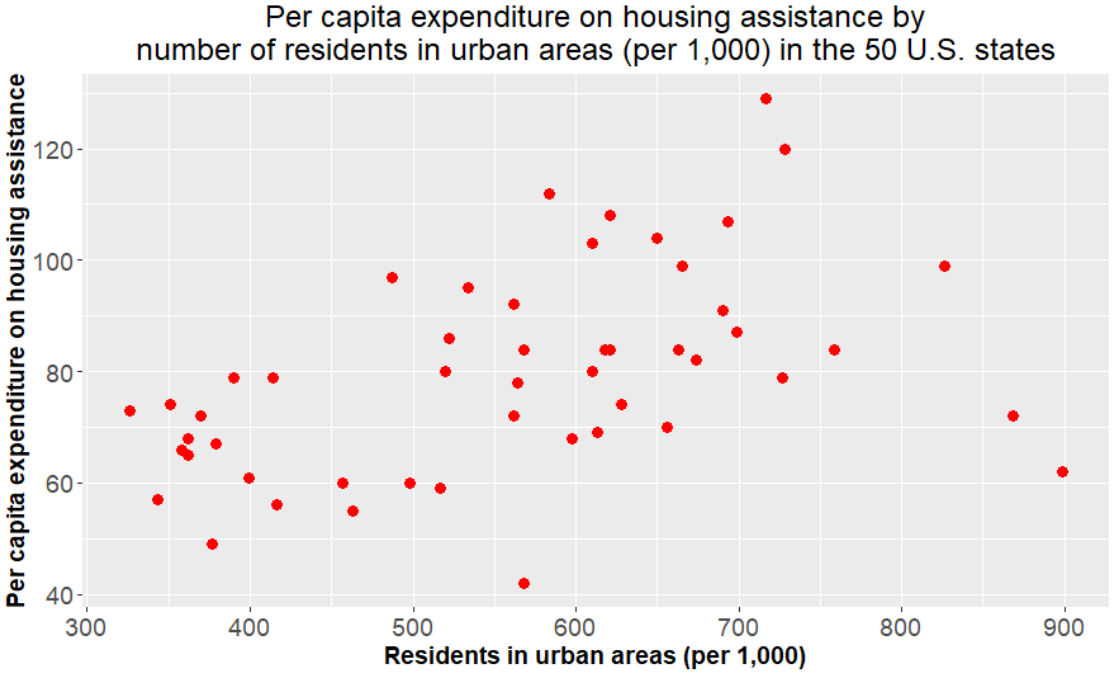



Data points in figure 2 are scattered in a U-shaped form, suggesting a non-linear relationship between $Y$ and $X2$.

### 3. Relationship between $Y$ and $X3$:

```
# (3) Scatterplot Y–X3:
ggplot(data=expenditure, mapping = aes(x = X3, y = Y)) +
  geom_point(color="red", size=3) +
  theme(plot.title = element_text(hjust = 0.5, size = 18),
        axis.text = element_text(size = 14), axis.title = element_text(
    size = 14, face = "bold")) +
  ggtitle("Per capita expenditure on housing assistance by
number of residents in urban areas (per 1,000) in the 50 U.S. states") +
  xlab("Residents in urban areas (per 1,000)") +
  ylab("Per capita expenditure on housing assistance")
```

Figure 3:

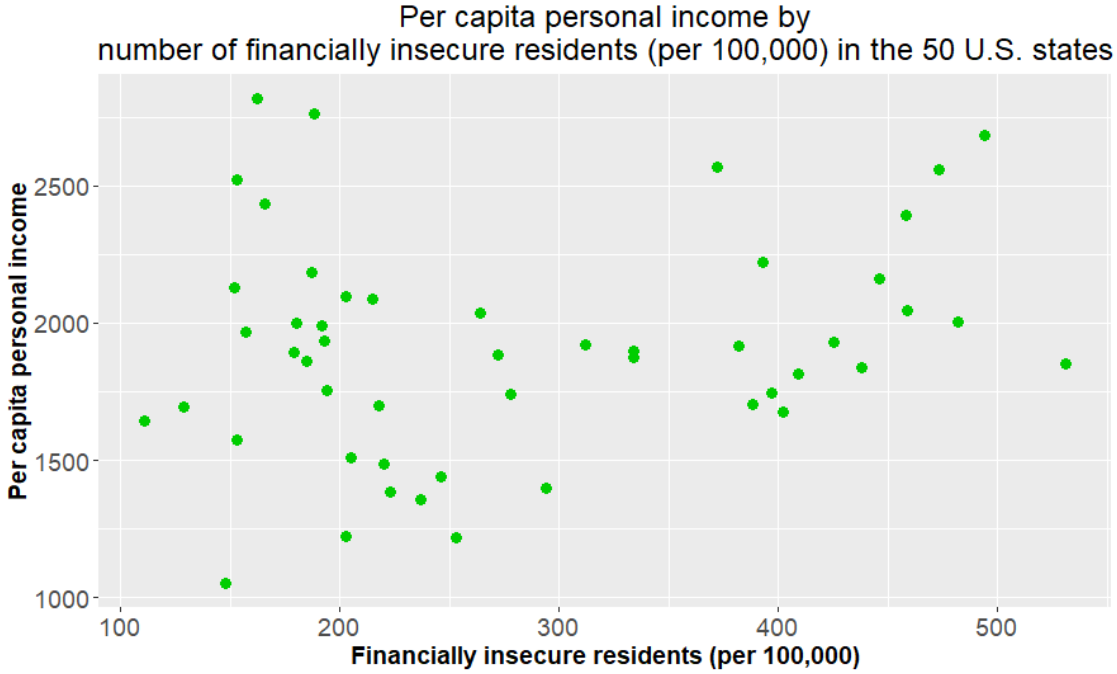



In figure 3, most data concentrate in the bottom-left corner and the center of the graphic. The pattern suggests a positive linear relationship between $Y$ and $X3$.

**4. Relationship between $X1$ and $X2$:**

```
# (4) Scatterplot X1–X2:
ggplot(data=expenditure, mapping = aes(x = X2, y = X1)) +
  geom_point(color="green3", size=3) +
  theme(plot.title = element_text(hjust = 0.5, size = 18),
        axis.text = element_text(size = 14), axis.title = element_text(
    size = 14, face = "bold")) +
  ggtitle("Per capita personal income by
number of financially insecure residents (per 100,000) in the 50 U.S.
    states") +
  xlab("Financially insecure residents (per 100,000)") +
  ylab("Per capita personal income")
```

Figure 4:

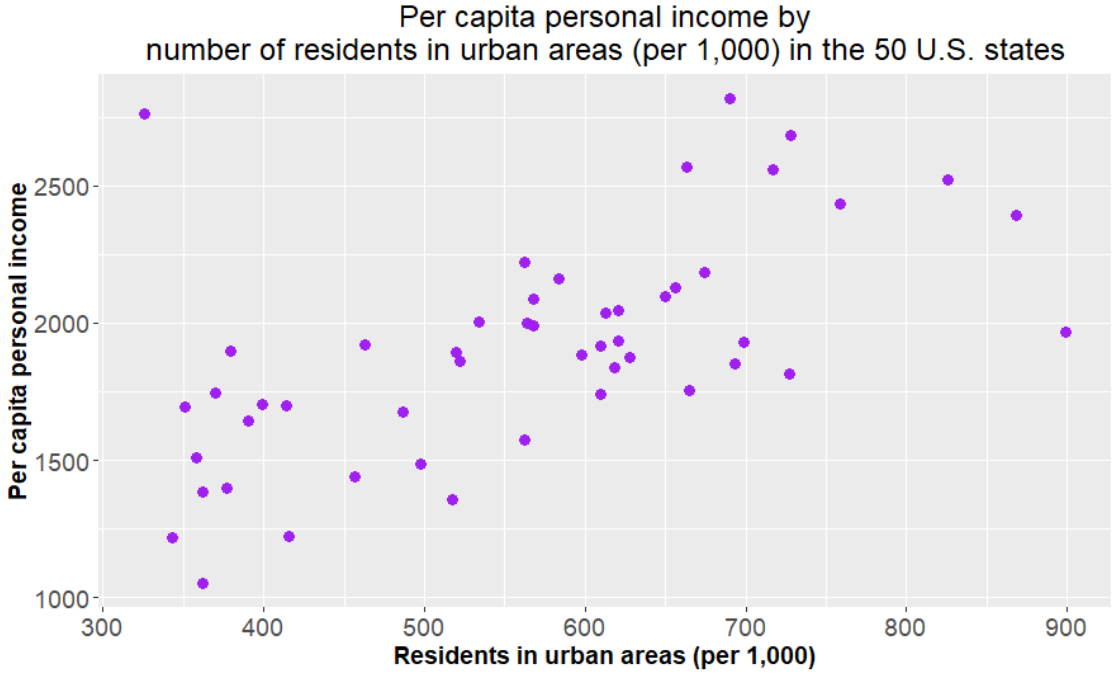



Data points are scattered all over figure 4, with no distinctive shape. The graphic does not allow any inference about the relationship between $X1$ and $X2$.

**5. Relationship between $X1$ and $X3$:**

```
# (5) Scatterplot X1–X3:
ggplot(data=expenditure, mapping = aes(x = X3, y = X1)) +
  geom_point(color="purple", size=3) +
  theme(plot.title = element_text(hjust = 0.5, size = 18),
        axis.text = element_text(size = 14), axis.title = element_text(
    size = 14, face = "bold")) +
  ggtitle("Per capita personal income by
number of residents in urban areas (per 1,000) in the 50 U.S. states") +
  xlab("Residents in urban areas (per 1,000)") +
  ylab("Per capita personal income")
```

Figure 5:

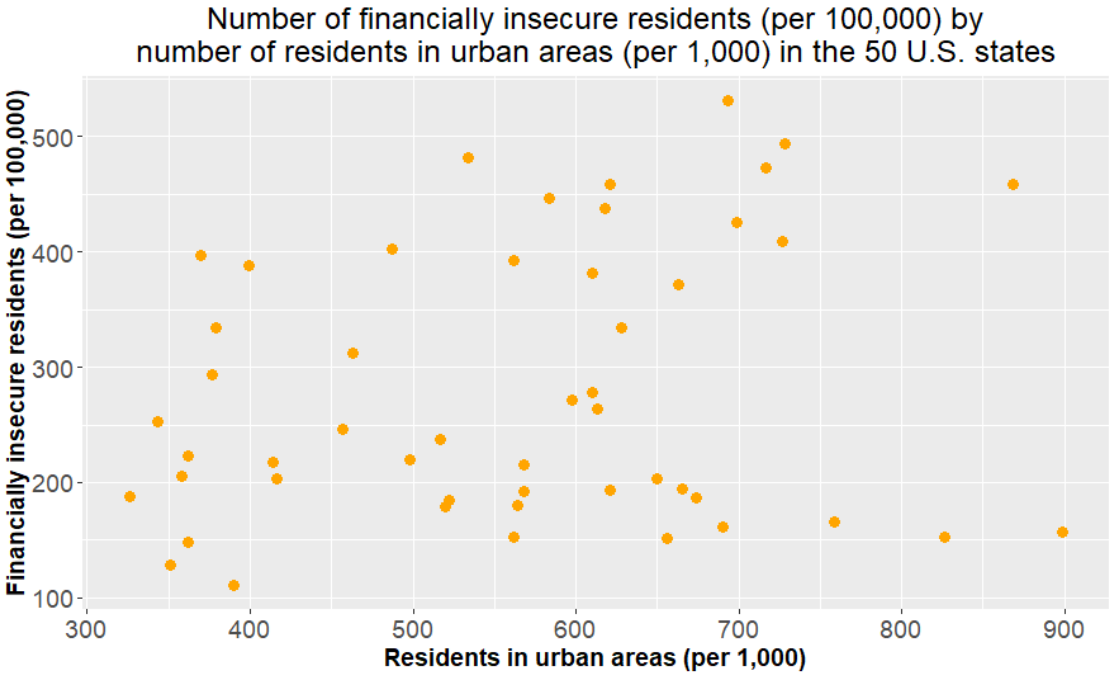



Data in figure 5 scatter from the bottom-left corner of the graphic to the top-right one. Apart from the outlier in the top-left corner, the scatter pattern suggests a positive linear relationship between $X1$ and $X3$.

**6. Relationship between $X2$ and $X3$:**

```
# (6) Scatterplot X2–X3:
ggplot(data=expenditure, mapping = aes(x = X3, y = X2)) +
  geom_point(color="orange", size=3) +
  theme(plot.title = element_text(hjust = 0.5, size = 18),
        axis.text = element_text(size = 14), axis.title = element_text(
    size = 14, face = "bold")) +
  ggtitle("Number of financially insecure residents (per 100,000) by
number of residents in urban areas (per 1,000) in the 50 U.S. states") +
  xlab("Residents in urban areas (per 1,000)") +
  ylab("Financially insecure residents (per 100,000)")
```

Figure 6:

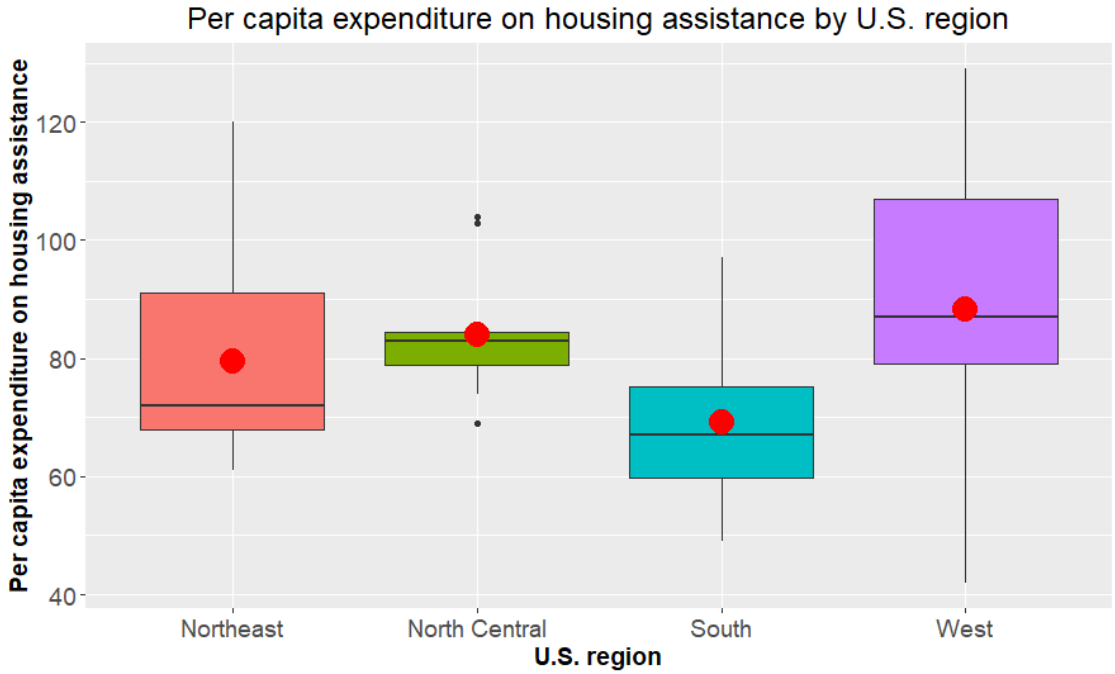



Data are scattered all over figure 6, with no distinctive shape. The graphic does not allow any inference about the relationship between $X2$ and $X3$.

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

```
# (7) Boxplot Y by Region:
ggplot(data=expenditure, mapping = aes(x = RegionFactor, y = Y, fill=
    RegionFactor)) +
  geom_boxplot() +
  stat_summary(fun.y=mean, geom="point", shape=20, size=10, color="red",
    fill="red") +
  theme(plot.title = element_text(hjust = 0.5, size = 18),
        axis.text = element_text(size = 14), axis.title = element_text(
    size = 14, face = "bold")) +
  ggtitle("Per capita expenditure on housing assistance by U.S. region")
   +
  xlab("U.S. region") +
  ylab("Per capita expenditure on housing assistance") +
  guides(fill=F)
```

Figure 7:

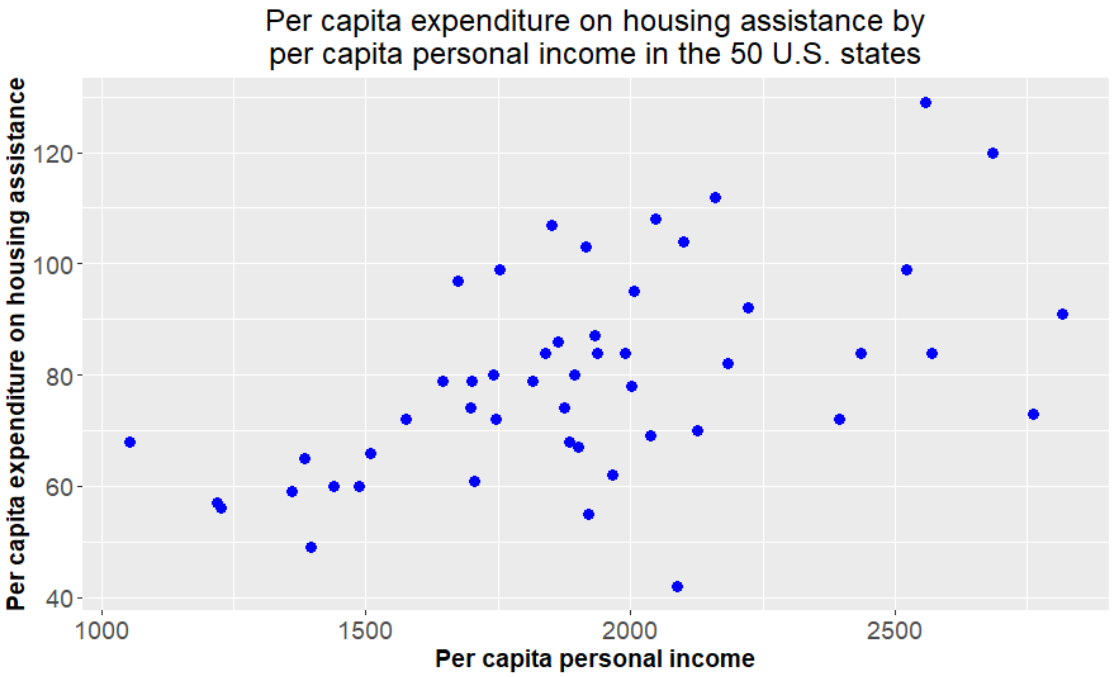



The West presents the highest mean (red dots), the highest median (horizontal bars) and the highest interquartile interval (boxes) among the four regions. Therefore, by those three measures, data indicate that **the West has, on average, the highest *per capita* expenditure on housing assistance**.

- *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.*

```
# (8) Scatterplot Y–X1:
ggplot(data=expenditure, mapping = aes(x = X1, y = Y)) +
  geom_point(color="blue", size=3) +
  theme(plot.title = element_text(hjust = 0.5, size = 18),
        axis.text = element_text(size = 14), axis.title = element_text(
    size = 14, face = "bold")) +
  ggtitle("Per capita expenditure on housing assistance by
per capita personal income in the 50 U.S. states") +
  xlab("Per capita personal income") +
  ylab("Per capita expenditure on housing assistance")
```

Figure 8:




In figure 8, data points scatter from the bottom-left corner of the graphic to the top-right corner. There seems to be some outliers both in the top and in the bottom of the graphic. Even so, the pattern suggests a positive linear relationship between these two variables.

Figure 9 displays the same scatter plot, but different regions in the U.S. are represented as different types of symbols and colors.

```r
# (9) Scatterplot Y–X1–Region:
ggplot(data=expenditure, mapping = aes(x = X1, y = Y, group =
    RegionFactor)) +
  geom_point(size=5, aes(shape = RegionFactor, color = RegionFactor)) +
  theme(plot.title = element_text(hjust = 0.5, size = 18),
        axis.text = element_text(size = 14), axis.title = element_text(
    size = 14, face = "bold"),
        legend.text = element_text(size = 14), legend.title = element_
    text(size = 14, face = "bold"),
        legend.position = "bottom") +
  ggtitle("Per capita expenditure on housing assistance by
per capita personal income in the 50 U.S. states classified by region") +
  xlab("Per capita personal income") +
  ylab("Per capita expenditure on housing assistance") +
  labs(color = "Region", shape = "Region")
```

Figure 9: