Problem Set 1

Hanyu Li (Student ID: 25346841)

October 9, 2025

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)
```

1. Find a 90% confidence interval for the average student IQ in the school.

```
# calculate sample size
n<-length(y)
n

# since sample size <30,using t score for small sample
t90 <- qt((1-0.9)/2, df=n-1, lower.tail = FALSE)

# calcluate sample mean and standard deviation
y_mean <- mean(y)
y_mean
y_sd <- sd(y)

# calculate the upper and lower side of confidence interval
upper_90 <- y_mean + t90 * (y_sd/sqrt(n))
lower_90 <- y_mean - t90 * (y_sd/sqrt(n))

# build the 90% confidence interval
confint90 <- c(lower_90,upper_90)
confint90
```

The 90% confidence interval is (93.95993, 102.92007); sample mean (98.44).

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 $\alpha = 0.05$.

```r
# conduct one-sample and one-tailed t-test under the confidence level of 95%
# null hypothesis:y_mean <= 100
# alternative hypothesis:y_mean > 100
t.test(y, mu = 100, conf.level = 0.95, alternative = 'greater')
```

```
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
```

The outcome shows: t value is rather close to 0, indicating there is no apparent difference between the observed mean and 100 and p-value is obviously greater than 0.05, which means the null hypothesis can't be rejected. In other words, the average student IQ in the school can't be seen as higher than the average IQ score (100) among all the schools in the country.

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)?

```
# view and grasp data structure
str(expenditure)

# display relationships among Y,X1,X2,X3 in a plot as a whole
pdf("plot_all.pdf")
pairs(expenditure[,c('Y','X1','X2','X3')])
dev.off()
# explore correlations between Y,X1,X2 and X3
cor(expenditure[, c("Y", "X1", "X2", "X3")])
```

The correlations between Y,X1,X2 and X3 are shown as below:

```
           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
```

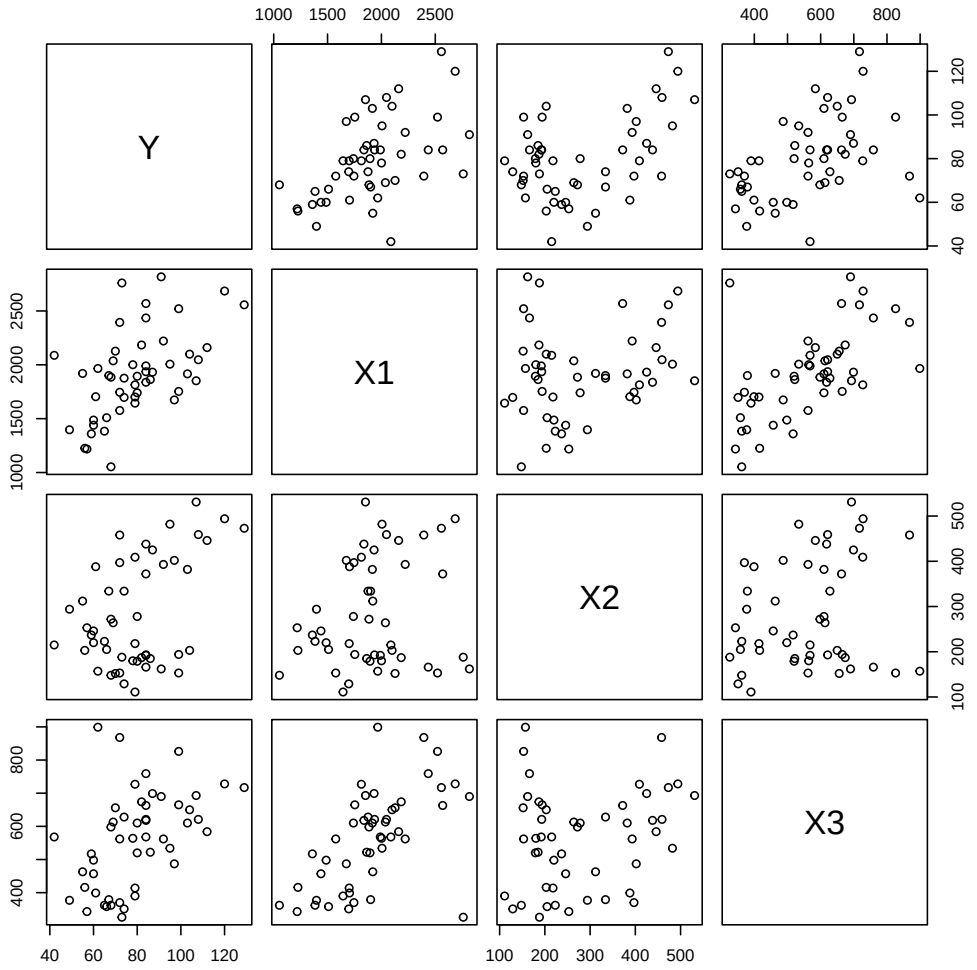

Figure 1: Relationships Among Y, X1, X2, X3

Figure 1 and correlation analysis outcomes indicates that:
A moderate and positive correlation between per capita expenditure on shelters/housing assistance (Y) and per capita personal income (X1), with a correlation coefficient of approximately 0.53.
Y also shows moderate positive associations with both financial insecurity (X2, r≈0.45) and urban population density (X3, r≈0.46).
Among the independent variables, per capita personal income (X1) and urban population density (X3) exhibited a strong positive correlation (r≈0.60), while the relationships involving financial insecurity (X2) and the other predictors were found to be weaker.

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

```
#factor region becasuse region here is a categorical variable
class(expenditure$Region)
expenditure$Region <- factor(expenditure$Region,
                             levels = c(1:4),
                             labels = c("Northwest","North Central","
    South","West"))
#calculate Y_mean of each region
mean_rg <- aggregate(Y ~ Region, data = expenditure, mean, na.rm = TRUE)

#plot the relationship between Y and Region
pdf("plot_Y_RG.pdf")
boxplot(expenditure$Y ~ expenditure$Region,data = expenditure,
        xlab = "Region",
        ylab = "Expenditure On Assistance Per Capita")

#show positions of mean by region on the boxplot
points(1:length(mean_rg$Y), mean_rg$Y, pch = 19 ,col = "blue"")
dev.off()
```

To compare the expenditure across four regions **averagely**,means of expenditure(Y) should be calculated and displayed on the plot.
Figure 2 shows: of the four regions, the **West** exhibits the highest expenditure: its median (black line) and mean(blue point) are both the highest.Therefore, it is reasonable to conclude that on average,the West region has the highest level of expenditure.

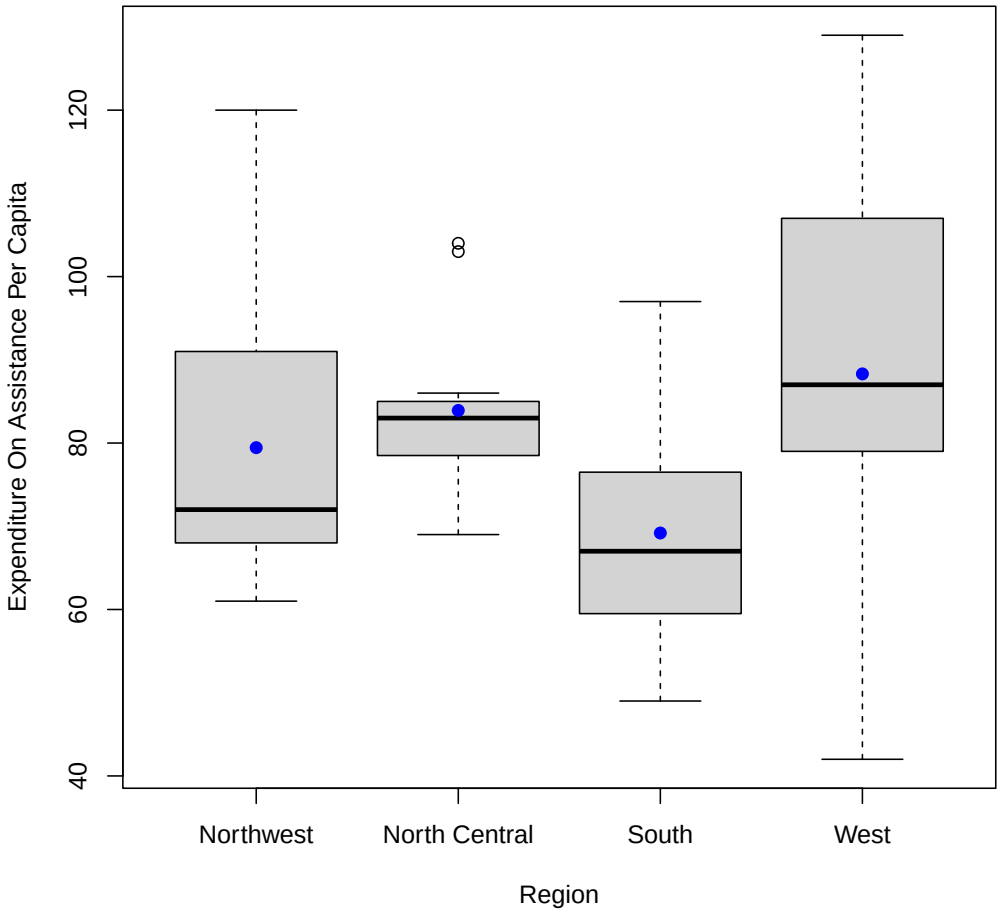

Figure 2: Relationship Between Y And Region

- 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.

```
# create scatter plot of Y and X1
pdf("plot_Y_X1.pdf")
plot(expenditure$X1,expenditure$Y,
     xlab = "Personl Income Per Captia",
     ylab = "Expenditure On Assistance Per Capita")
     abline(lm(Y ~ X1, data = expenditure), col = "grey", lwd = 1)
dev.off()
```

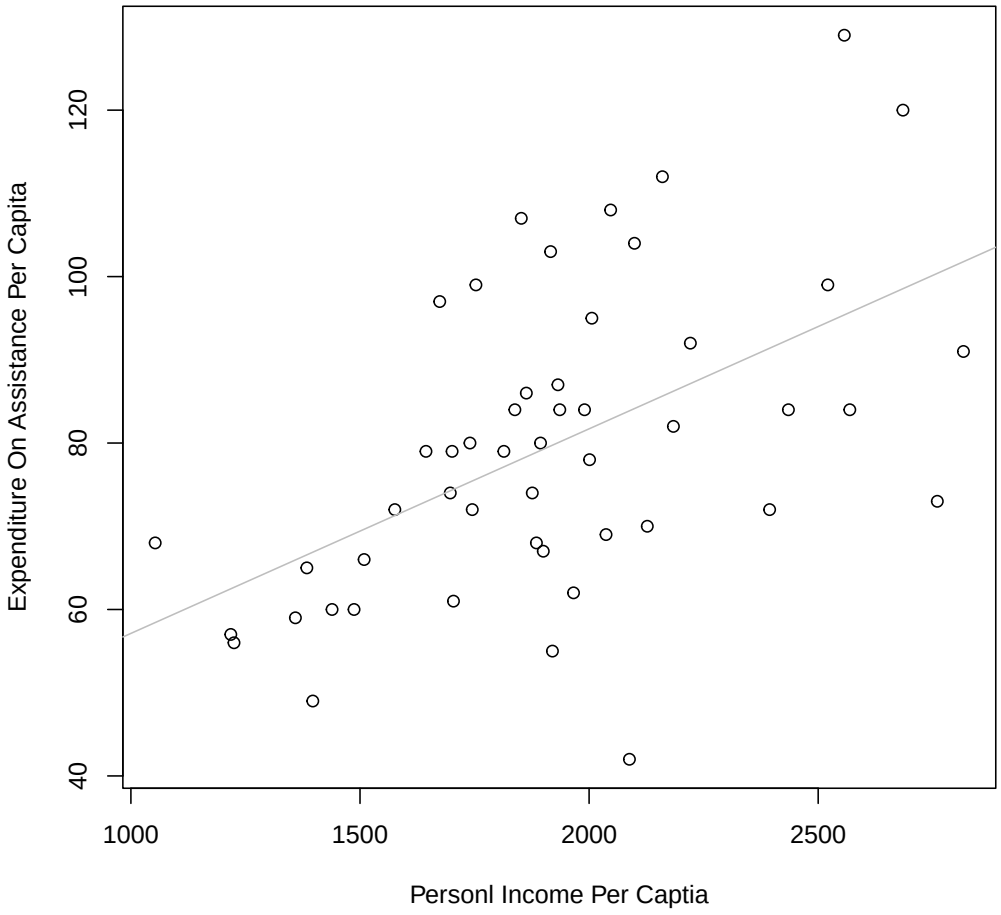

Figure 3: Relationship Between Y And X1

Figure 3 indicates that per capita expenditure (Y) is positively associated with per capita personal income (X1), which means as the state's per capita personal income increases, shelters/housing assistance spending per capita grows as well.

Adding variable *Region* to the plot and display different regions with different types of symbols and colors.

```
# plot relationship between Y and X1 based on Region
# and display different regions with different types of symbols and
    colors
pdf("plot_Y_X1_byRG.pdf")
ggplot(expenditure, aes(x = X1, y = Y, color = Region, shape = Region)) +
```

```
  geom_point()+
  geom_smooth(aes(group = 1),method = "lm", se = FALSE, linetype = "solid
    ",col= "grey") +
  xlab("Personl Income Per Captia") +
  ylab("Expenditure On Assistance Per Capita")
dev.off()
```

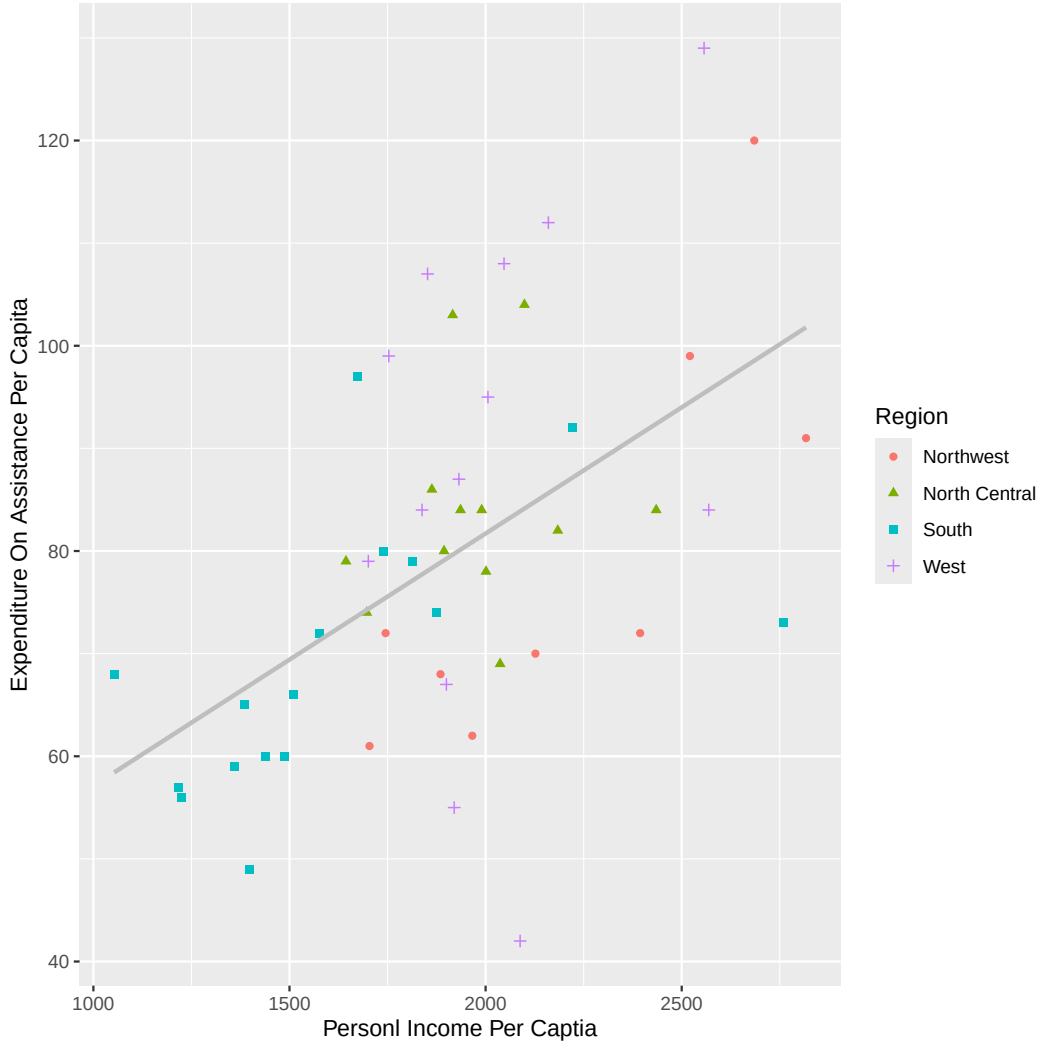

Figure 4: Relationship Between Y And X1 By Region