Problem Set 1

Samanta Nedzinskaite

02 October 2022

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:

```
iqscores <- 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.

```
#confidence interval of a mean = mean +- z-score * standard error
mean <- mean(iqscores)
#finding standard error = standard deviation / sqrt of sample size
n <- length(iqscores)
sd <- sd(iqscores)
se <- sd/sqrt(n)

#z-score for 90% CI
conf.level <- 0.9
z <- qt((1+conf.level)/2, df=n-1)
ci <-z*se

conf_interval <- c(mean-ci, mean+ci)
```

Results : [1] 93.95993 102.92007

Therefore, according to our confidence interval, we can say that in ninetey out of a hundred instances, we are confident that the average IQ score would fall between the range of about 94 and 103.

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 $\alpha = 0.05$.

   - Null hypothesis: the average IQ score of the students in this particular school is not significantly higher than the average population IQ score.
   - Alternate hypothesis: the average IQ score of the students in this particular school is higher than the average of all the other schools in the country (represented by a smaple of 100)

```
#sample size = 25
#sample mean = 98.44
#sample standard deviation = 13.0928733795654
#Level of significance =    =0.05
popmean <- 100
z_score <- (mean - popmean) / se
z_score
```

   Results: [1] -0.5957439

   Our z-score is 0.6...which is smaller than our critical value of 2 when $\alpha = 0.05$. Therefore, we fail to reject the null hypothesis.

Question 2 (50 points): Political Economy

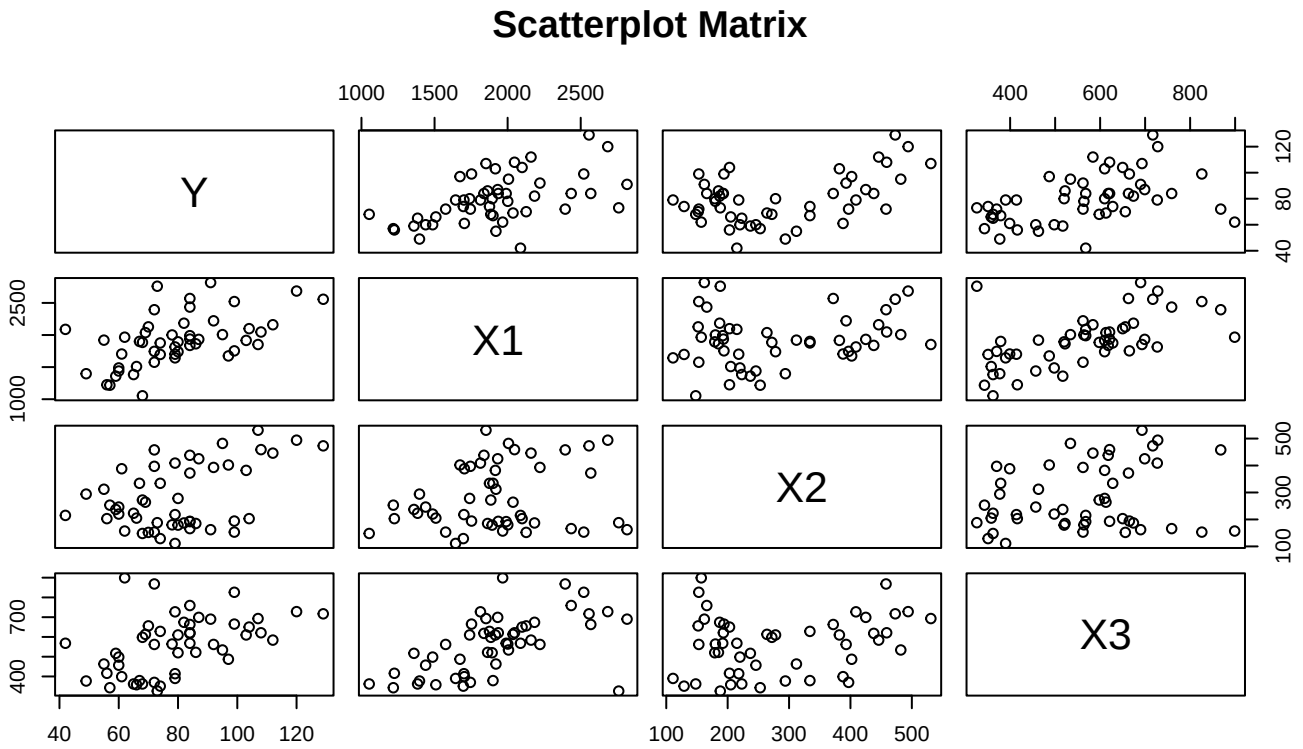
Researchers are curious about what affects the amount of money communities spend on addressing homelessness. The following variables constitute our data set about social welfare expenditures in the USA.

`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`.

- 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)?

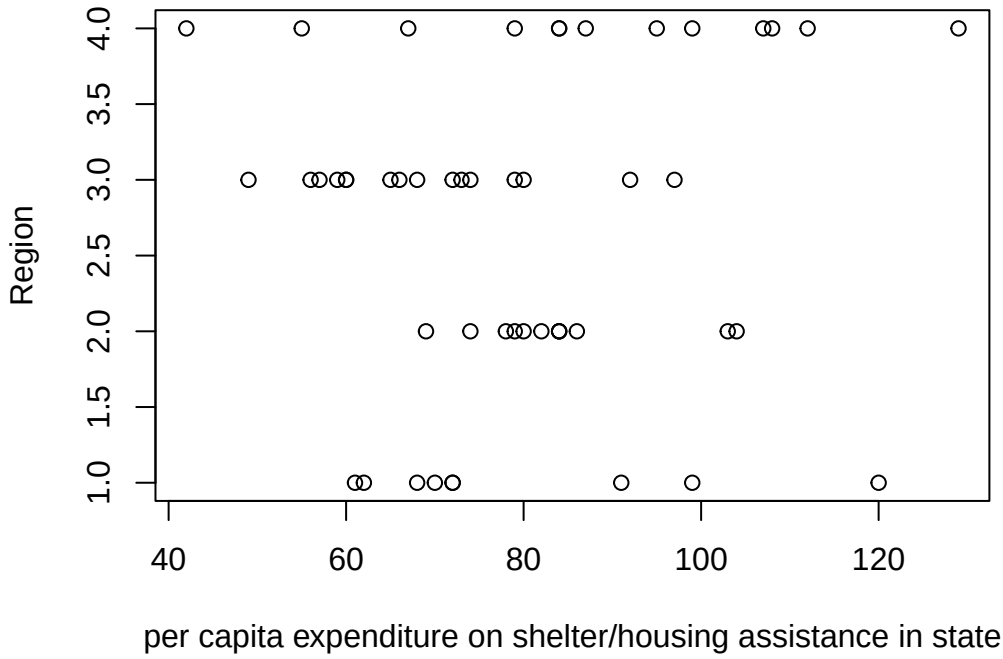
```
expenditure
#first plot
y1 <- as.data.frame(expenditure$Y)
min(expenditure$Y)
max(expenditure$Y)
pairs(~Y+X1+X2+X3,data=expenditure,
      main="Scatterplot Matrix")
```



From the matrix scatterplot above, there is no clear observable correlation between any of the variables. That is, none of the relationshios represented above in the scatterplots demonstrate clear linearity. The relationship between Y and X1 observable in the designated scatterplot resembles a linear distribution the most, but not perfectly.

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

```
y <- expenditure$Y
x <- expenditure$Region
plot(y,x, xlab="per capita expenditure on shelter/housing assistance in
    state", ylab="Region")
```

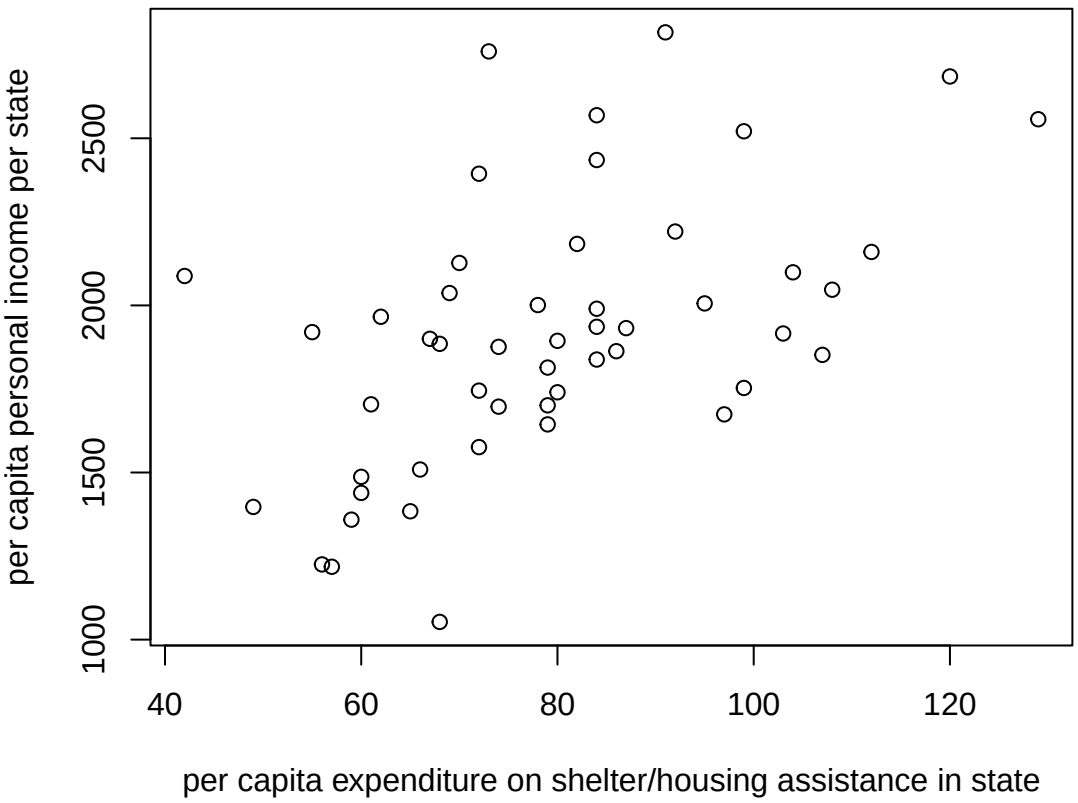


From the scatterplot abvove, we can say that on average, Region 4 (West) - has the highest per capita expenditure on housing assistance.

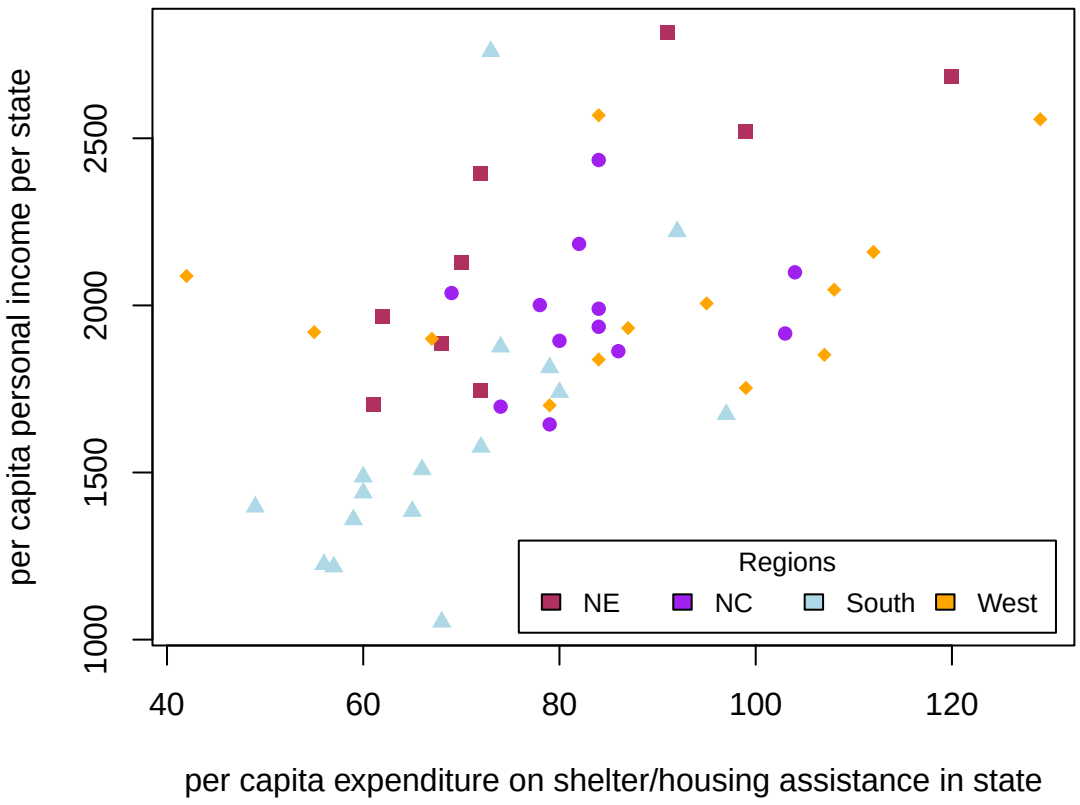
- Please plot the relationship between Y and $X1$? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.

```
plot(expenditure$Y, expenditure$X1, xlab="per capita expenditure on
    shelter/housing assistance in state", ylab="per capita personal income
     per state")


expenditure <- read.delim("/Users/samantanedzinskaite/Documents/GitHub/
    StatsI_Fall2022/datasets/expenditure.txt")
cols <- c("maroon","purple","light blue","orange")
pchs <- c(pch=15, pch=16, pch=17, pch=18)
plot(expenditure$Y, expenditure$X1, col= cols[expenditure$Region], pch=
    pchs[expenditure$Region],  xlab="per capita expenditure on shelter/
    housing assistance in state", ylab="per capita personal income per
    state")
legend("bottomright", inset=.02, title="Regions",
       c("NE","NC","South", "West"), fill=c(cols, pchs), horiz=TRUE, cex
    =0.8)
```



There is a moderate and weak, positive, linear association between the two variables. However, the large number of outliers weakens the association - as demonstrated on the scatterplot above.


per capita personal income per state
1000
1500
2000
2500
40
60
80
100
120
per capita expenditure on shelter/housing assistance in state
Regions
NE
NC
South
West