# Problem Set 1

## Applied Stats/Quant Methods 1

## Due: October 1, 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 1, 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:

```
1 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.

```
1 IQ_xbar <- mean(y)
2 IQ_s <- sd(y)
3 IQ_n <- length(y)
4 z <- qnorm(0.95) # 90% CI - alpha = .10 -> z(alpha/2) = z(0.05)
```

```
5  lower_limit <- IQ_xbar - z * (IQ_s /sqrt(IQ_n))
6  upper_limit <- IQ_xbar + z * (IQ_s/sqrt(IQ_n))
7  print(c(lower_limit, upper_limit))
```

Answer: 90% confidence interval for the average student IQ in the sample is (94.133, 102.747).

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 $\alpha = 0.05$.

```
1  test <- t.test(y, mu = 100, conf.level = 0.95, alternative="greater")
2
3  test$conf.int
4  test$p.value
```

Answer: We don't have evidence to reject the null hypothesis because p-value (-0.7215383) of the test is higher than the significance level (0.05).The average IQ score of the sample of 25 students is not significantly greater than the average IQ score among all the schools in the country (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`.

- 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)?
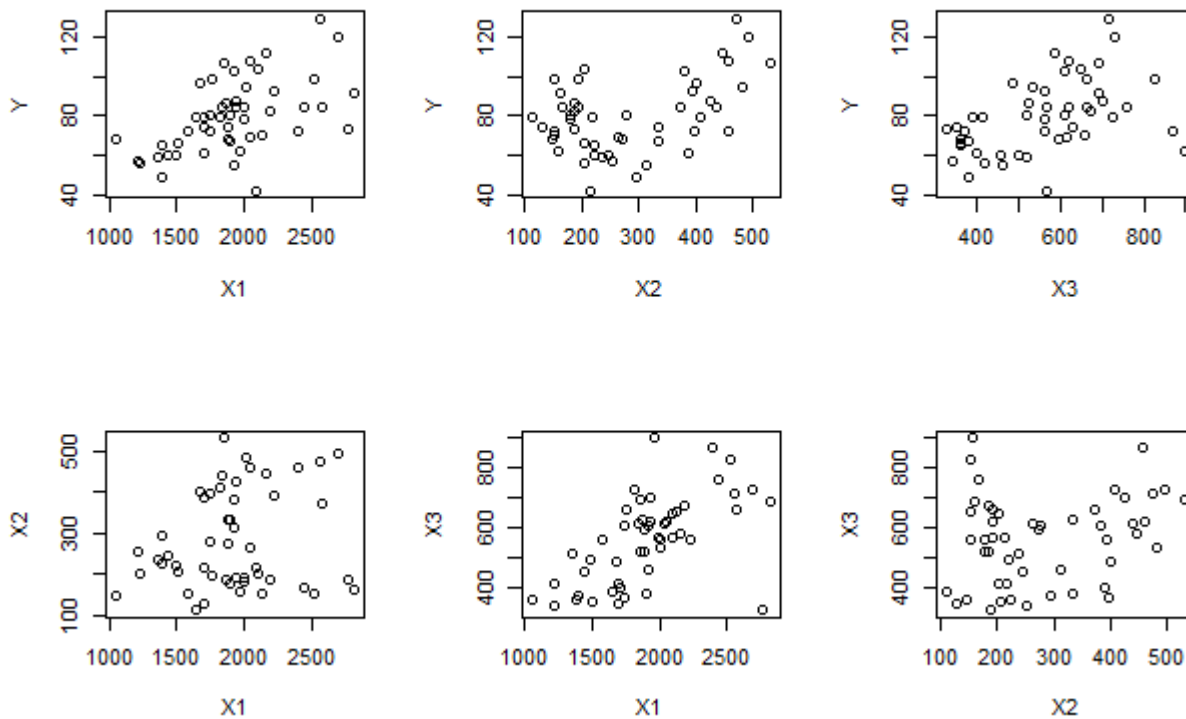
```r
#scatterplots
attach(expenditure)

p1 <- plot(X1, Y)
#There is a positive association, possibly linear and relatively strong,
    between the variables Y and X1, with potential outliers.

p2 <- plot (X2, Y)
#There is an apparent scattered relation around a smooth curve (non-
    linear) and not very strong relationship between variables Y and X2.

p3 <- plot (X3, Y)
#There is no apparent relationship between the variables Y and X3,
    possibly a positive non-linear.

p4 <- plot(X1, X2)
#There is no apparent relationship between the variables X1 and X2, data
    points are scattered.

p5 <- plot(X1, X3)
#There is a stronger positive between variables X1 and X3 relation with
    potential outliers.

p6 <- plot(X2, X3)
#There is no apparent association between the variables X2 and X3, the
    data points are scattered distributed.
```
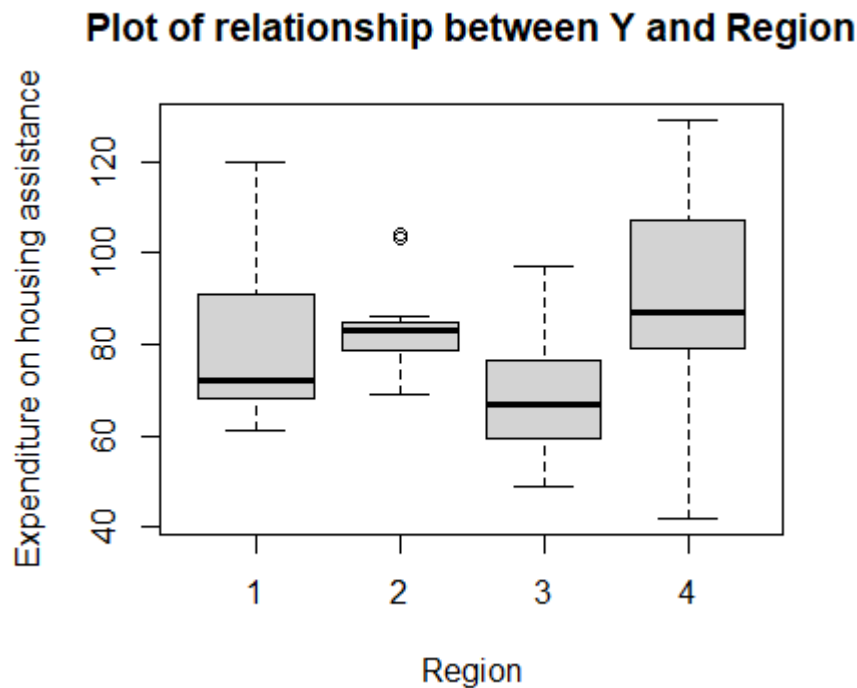
- Please plot the relationship between *Y* and *Region*? On average, which region has the highest per capita expenditure on housing assistance?

```r
plot(expenditure$Region, expenditure$Y,
     main = "Plot of relationship between Y and Region",
     xlab = "Region",
     ylab = "Expenditure on housing assistance")
```

**Plot of relationship between Y and Region**

Answer: On average, the states of the region 4 (West) have a higher per capita expenditure on shelters/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.

```r
library(ggplot2)
plot_1 <- ggplot(expenditure, aes(x = X1, y = Y))+
  geom_point() +
  theme_minimal() +
  labs(title="Plot of relationship between Y and X1",
       x ="Per capita personal income in state", y = "Per capita expenditure
  on housing")
plot_1
```

## Plot of relationship between Y and X1



There is an apparent positive relation between the variables, the higher the per capita personal income in the state, the higher the per capita expenditure on housing assistance by the state.

```
1  plot_2 <- ggplot(expenditure, aes(x = X1, y = Y))+
2    geom_point(aes(color = Region, shape = Region))+
3    theme_minimal()+
4    labs(title="Plot of relationship between Y and X1 by Region",
5        x ="Per capita personal income in state", y = "Per capita expenditure
     on housing")
6  plot_2
```

Plot of relationship between Y and X1 by Region