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:

```
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.

   I used a t-test because n is less than 30.

```
t.test(y,
       conf.level = 0.90,) # Run t-test with 90% confidence interval
```

With a t-test, the 90 percent confidence interval for the average student IQ is [93.95993, 102.92007].

If I were to use a normal distribution, I would do the following:

```
z90 <- qnorm((1 - .90)/2, lower.tail = FALSE) # Establish z-score with
    90% CI
n <- length(y) # Establish n
y_mean <- mean(y) # Establish mean of y
y_sd <- sd(y) # Establish standard deviation of y
lower_90 <- y_mean - (z90 * (y_sd/sqrt(n))) # Lower CI
upper_90 <- y_mean + (z90 * (y_sd/sqrt(n))) # Upper CI
confint90 <- c(lower_90, upper_90) # CI
print(confint90) # Print CI
```

With a normal distribution, the 90 percent confidence interval for the average student IQ would be [94.13283, 102.74717].

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 $\alpha = 0.05$.

```
average_iq <- 100 # I set the average IQ at 100.

y_null <- t.test(y, mu = average_iq,
        alternative = "greater") # I conduct a t-test to see the p-value
    for the school having an above-average IQ.

print(y_null) # Display the t-test.
```

   The p-value is very high (.7215), much higher than the .05 significance level. We cannot reject the null hypothesis that the school's average IQ is equal or less than the average school IQ score (100).

   Source for one-sided t-test: http://www.sthda.com/english/wiki/one-sample-t-test-in-r

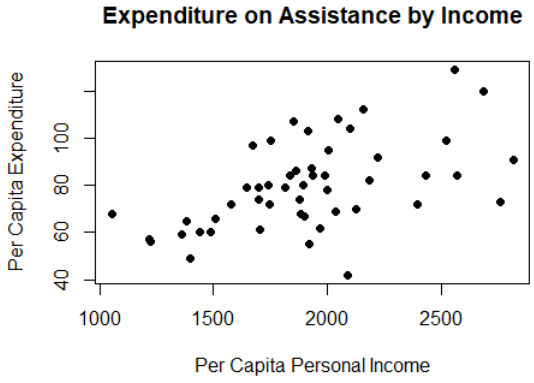
Question 2 (50 points): Political Economy

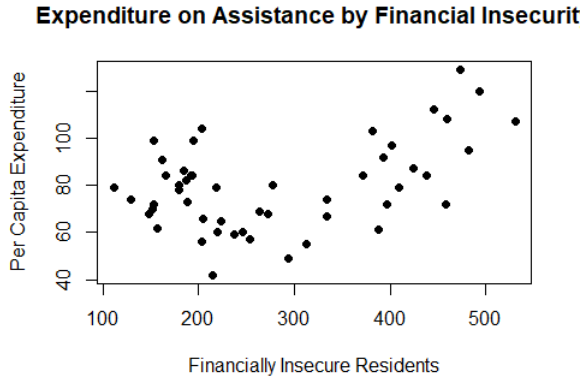
Researchers are curious about what affects the amount of money communities spend on addressing homelessness. The following variables constitute our data set about social welfare expenditures in the USA.

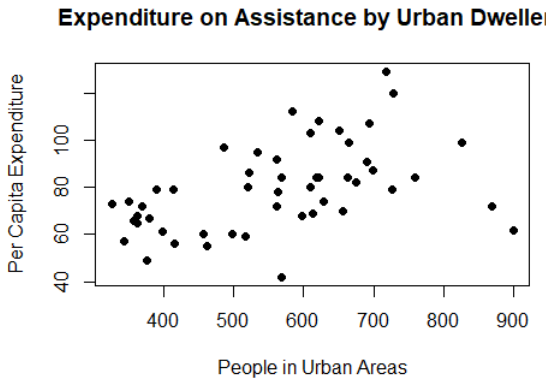
`State`	*50 states in US*
`Y`	*per capita expenditure on shelters/housing assistance in state*
`X1`	*per capita personal income in state*
`X2`	*Number of residents per 100,000 that are "financially insecure" in state*
`X3`	*Number of people per thousand residing in urban areas in state*
`Region`	*1=Northeast,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

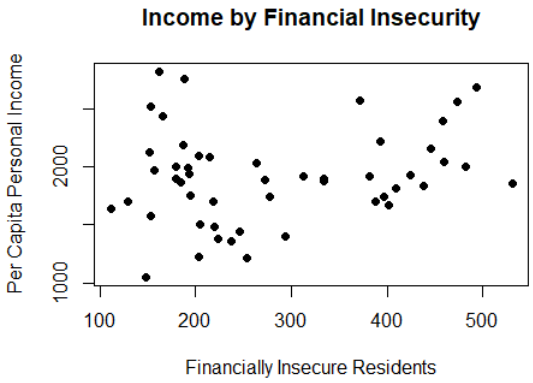
Explore the `expenditure` data set and import data into `R`.

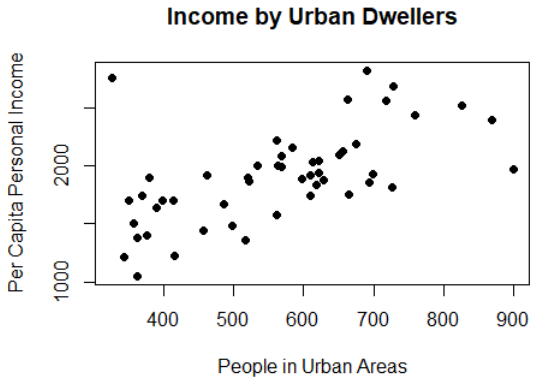
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?

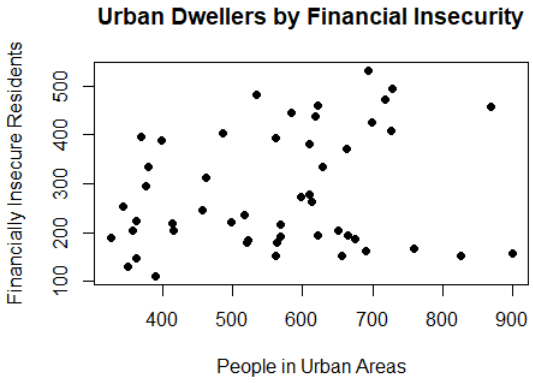













```r
plot(expenditure$X1, expenditure$Y, main="Expenditure on Assistance by Income",
     xlab="Per Capita Personal Income", ylab="Per Capita Expenditure", pch=19) # Plot Y and X1

plot(expenditure$X2, expenditure$Y, main="Expenditure on Assistance by Financial Insecurity",
     xlab="Financially Insecure Residents", ylab="Per Capita Expenditure", pch=19) # Plot Y and X2

plot(expenditure$X3, expenditure$Y, main="Expenditure on Assistance by Urban Dwellers",
     xlab="People in Urban Areas", ylab="Per Capita Expenditure", pch=19) # Plot Y and X3

plot(expenditure$X2, expenditure$X1, main="Income by Financial Insecurity",
     xlab="Financially Insecure Residents", ylab="Per Capita Personal Income", pch=19) # Plot X1 and X2

plot(expenditure$X3, expenditure$X1, main="Income by Urban Dwellers",
     xlab="People in Urban Areas", ylab="Per Capita Personal Income", pch=19) # Plot X1 and X3
```

```
15
16 plot(expenditure$X3, expenditure$X2, main="Urban Dwellers by Financial
       Insecurity",
17      xlab="People in Urban Areas", ylab="Financially Insecure Residents",
        pch=19) # Plot X2 and X3
```

Y and X1: These exhibit a positive correlation that has medium strength (r = .53). The slope is not very steep. Datapoints are distributed more at the lower values of Y and X1.

Y and X2: These exhibit a positive correlation that is not very strong (at least linearly) (r = .45). Datapoints are distributed more at medium values of Y and low values of X2. This is a non-linear correlation. As X2 increases, Y decreases and then increases. There is an overall slight positive relationship.

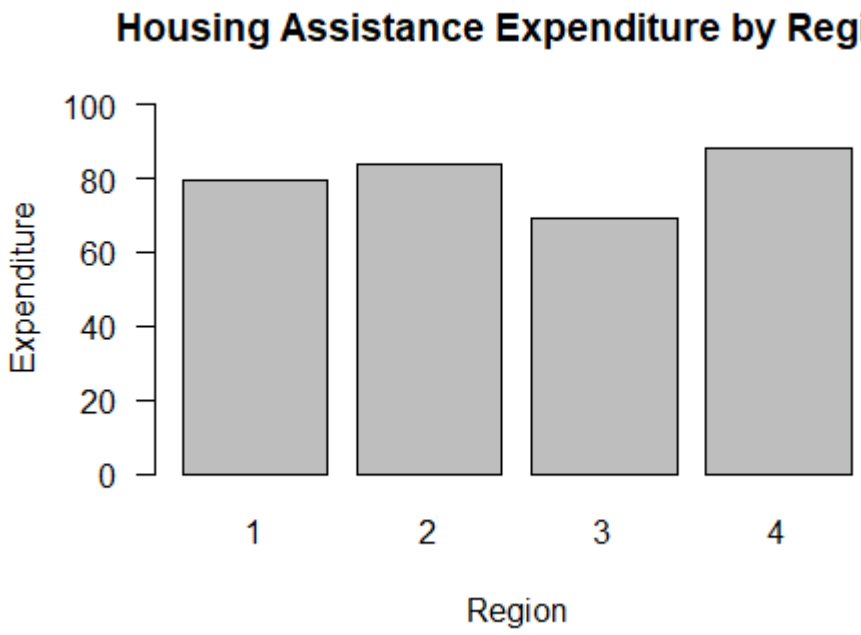
Y and X3: These exhibit a positive correlation that is not very strong (r = .46). The slope is moderate. Datapoints are distributed relatively evenly, but with fewer at high values of Y and X3.

X1 and X2: These exhibit a positive correlation that is weak (r = .21). The slope is quite flat. Datapoints are distributed relatively evenly for X1 and at the lower values for X2.

X1 and X3: These exhibit a positive correlation that is moderately strong (r = .60). The slope is not very steep. Datapoints are distributed relatively evenly for X1 and are sparser at high values of X3.

X2 and X3: These exhibit a positive correlation that is not very strong (r = .22). There is only a slight slope. Datapoints are distributed relatively evenly for X2 and are sparser at high values of X3.

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



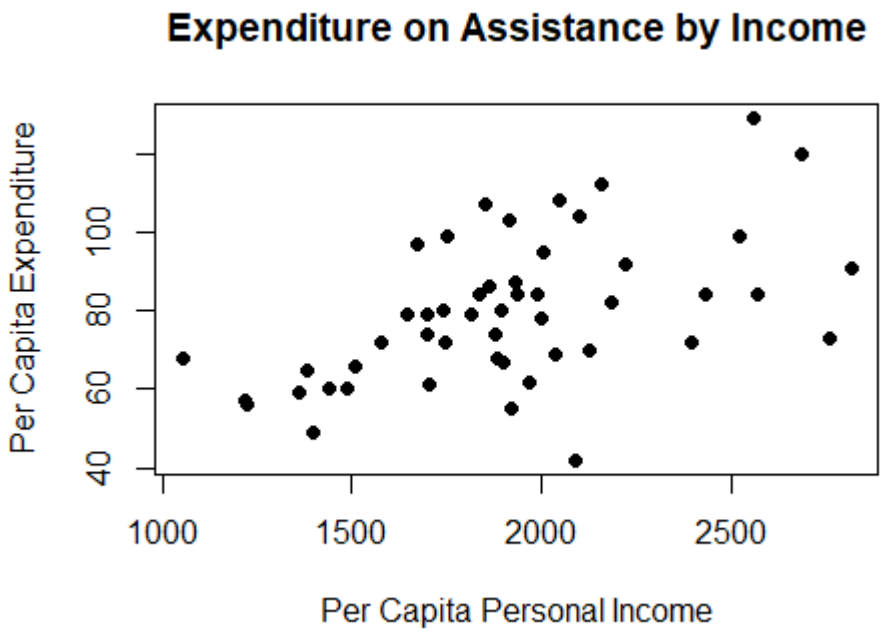

```r
expenditure_means <- aggregate(expenditure$Y, by = list(expenditure$
    Region), FUN = mean) # Establish expenditure means by region

subset(expenditure_means, Group.1 %in% c(1, 2, 3, 4)) # Subset
    expenditure means by region

expenditure_bp <- barplot(expenditure_means$x,
        ylim=c(0,100),
        main = "Housing Assistance Expenditure by Region",
        names.arg = expenditure_means$Group.1,
        xlab = "Region",
        ylab = "Expenditure",
        las = 1) # Create bar plot for expenditure means by region
```

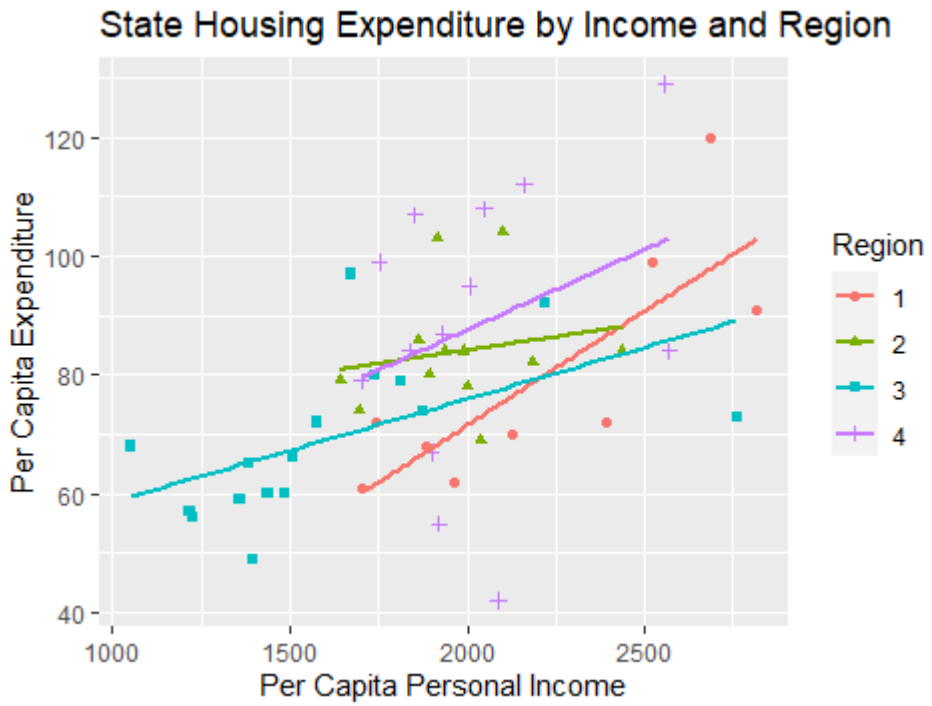
The West (4) has the highest per capita expenditure on housing assistance.

- 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$X1, expenditure$Y, ylim=c(0,150), main="Expenditure on
    Assistance by Income",
      xlab="Per Capita Personal Income", ylab="Per Capita Expenditure",
    pch=19) # Plot Y and X1 relationship again
```

Above is a plot of the bivariate relationship between Y and X1. These exhibit a positive correlation that has medium strength (r = .53). The slope is not very steep. Datapoints are distributed more at the mid-to-low values of Y and X1.




```
expenditure$Region <- as.factor(expenditure$Region) # Turn Region into a
    categorical variable
head(expenditure) # View small subset of data

ggplot(expenditure, aes(x=X1, y=Y, shape=Region, color=Region)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE) +
  labs(title="State Housing Expenditure by Income and Region",
    x="Per Capita Personal Income", y = "Per Capita Expenditure") #
    Create scatter plot, differentiated by region
```

Above is a graph with the bivariate relationship separated by region. Higher per capita income has a positive correlation with housing assistance expenditure for all four regions. However, the regions also have different average expenditures. Furthermore, the slope is higher for some regions (e.g., Northeast) than others (e.g., North Central). This indicates that both income and region (and an interaction effect between the two) may be related to expenditure, unless there are other causal variables involved that nullify this.

Source for ggplot: http://www.sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization