Problem Set 1

Daniel Murray (13303981)

Due: October 3, 2022

Question 1 (50 points): Education

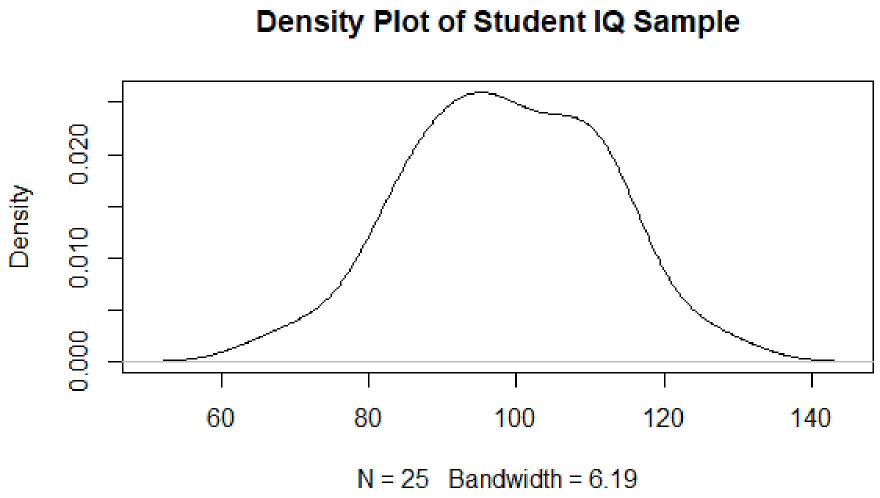
A school counsellor was curious about the average of IQ of the students in her school and took a random sample of 25 students' IQ scores. The following is the data set:

```r
IQ <- 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)
```

Part (1)

We are interested in finding a 90% confidence interval for the average student IQ in the school. An inspection of the data revealed that it is approximately normally distributed, as shown in the table below.

Figure 1: Density plot of student IQ sample



Following this inspection the mean, standard deviation, and standard error of the sample data was calculated.

A t-distribution was then used to construct a 90% confidence interval for the average IQ of the population of all students in the school, as the sample size is relatively small and the exact standard error of the population is unknown. The sample standard deviation was substituted for the population standard deviation to calculate the estimated standard error, which introduces extra error and necessitates replacing the z-score with a t-score.

Assuming that the sampling distribution of the sample mean is approximately normal, a t-score was calculated for a 90% confidence interval with df = 24 using the following code:

```r
tscore_IQ <- qt(0.95, 24, lower.tail = TRUE)
```

From this t-score, the margin of error was calculated. A confidence interval was constructed by adding and subtracting the margin of error value from our point estimate, the sample mean.

The average student IQ in the school was found to be 98.4, 90% CI [94.0, 102.9].

Part (2)

We are also interested in investigating whether the average student IQ in the school is higher than the average IQ score (100) among all schools in the country.

The sample data from Part 1 was used to represent the average student IQ in the school. This data was obtained through random sampling, and is normally distributed as determined above.

A hypothesis test with $\alpha = 0.05$ was conducted as per below:

$$H_0 : \mu = \bar{y}$$

$$H_a : \mu < \bar{y}$$

The value for the test statistic (t-score) was calculated as -0.59. As we are testing the probability that the average IQ of students in our sample is higher than the average IQ score, we find the right-tailed P-score for our test statistic with degrees of freedom df = 24. The below code was used, yielding a P-value of 0.72.

```r
p_value <- pt(abs(tscore_null), n_IQ-1, lower.tail = T)
```

The P-value of P = 0.72 is not below our threshold of 0.05, therefore we do not reject the null hypothesis that the average IQ of students in our sample is equal to the average IQ score among all schools in the country.

Question 2 (50 points): Political Economy

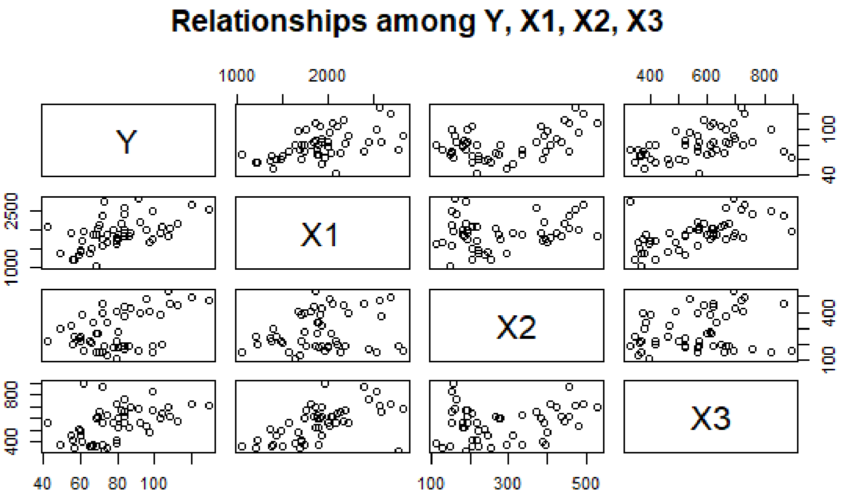
Researchers are curious about what affects the amount of money communities spend on addressing homelessness. The following variables constitute our data set about social welfare expenditures in the USA.

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*

Part (1)

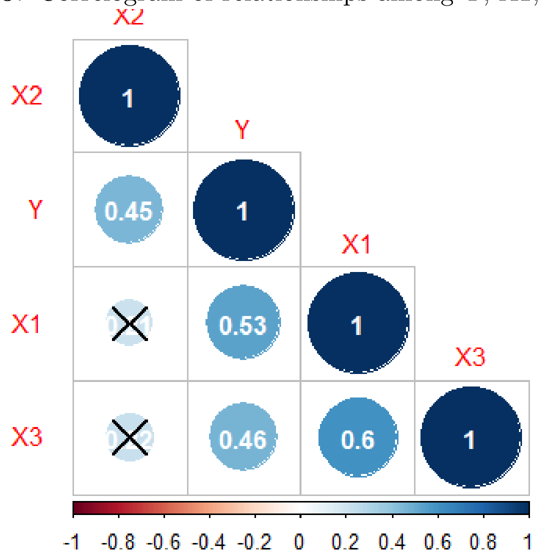
Below is a multiple scatter plot illustrating the relationships among *Y*, *X1*, *X2*, and *X3*. The graphs show that all variables have a positive, broadly linear relationship, with the strongest correlations appearing to be between *X1* and *X3*, and *Y* and *X1*.

Figure 2: Multiple scatter plot of relationships among Y, X1, X2 and X3



The same relationships can also be illustrated using a correlogram, as below. This shows that *X1*/*X3* and *Y*/*X1* are indeed the variable pairs with the strongest correlations. It also shows that while there is correlation between *X1*/*X2* and *X2*/*X3*, these relationships are not statistically significant at a 95% confidence level.

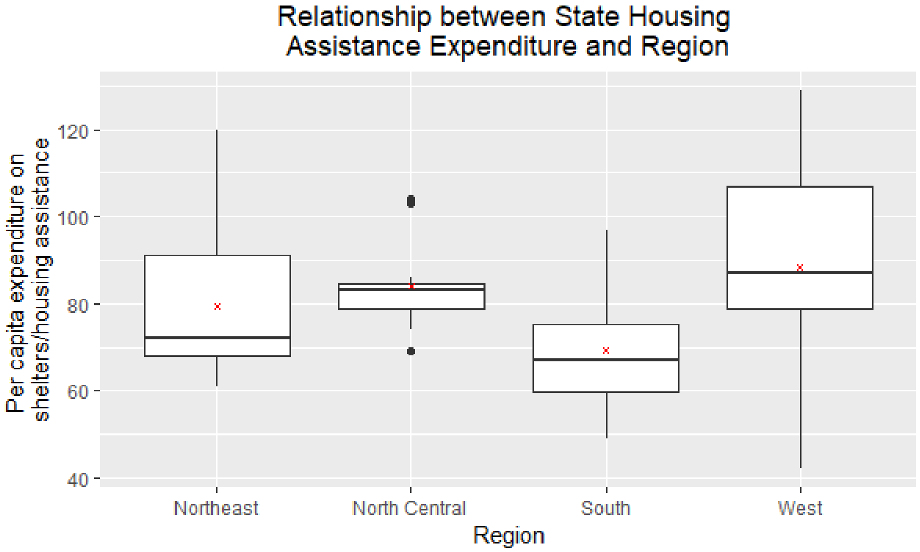
Figure 3: Correlogram of relationships among Y, X1, X2 and X3



Part (2)

Below is a boxplot illustrating the relationship between Y and *Region*. It shows that, on average, the West is the region the highest per capita expenditure on housing assistance.

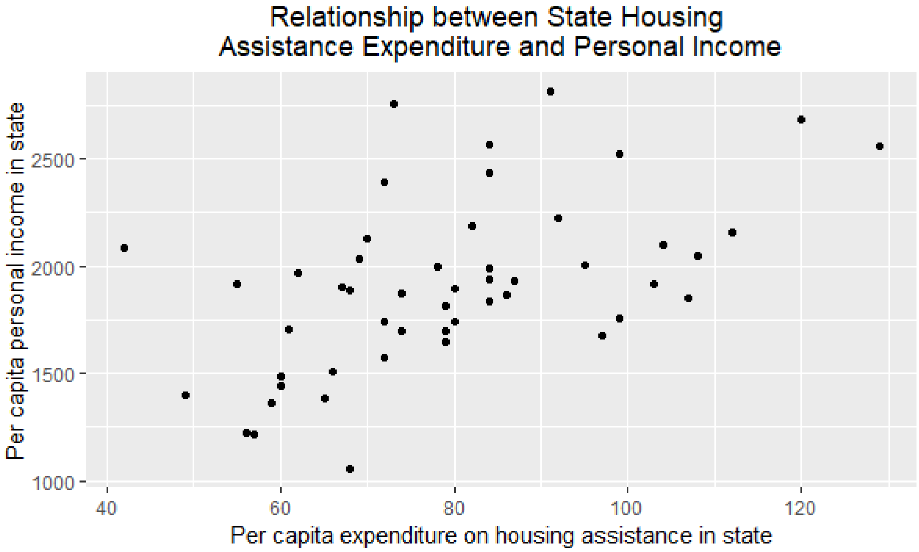
Figure 4: Boxplot of relationship between Y and Region



Part (3)

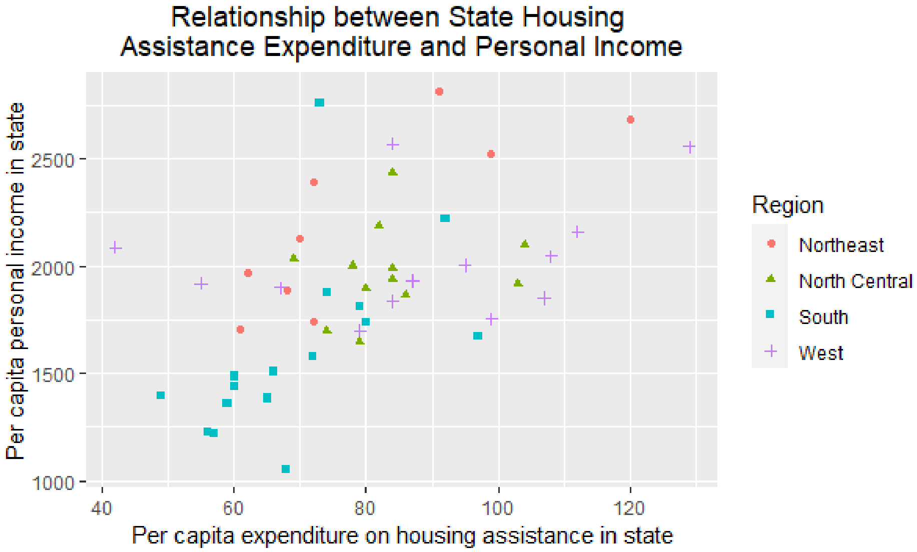
Below is a scatter plot illustrating the relationship between *Y* and *X1*. It indicates that these two variables have a positive, linear relationship of moderate strength, with the presence of several apparent outliers.

Figure 5: Boxplot of relationship between Y and X1



Below is the same graph as Figure 5 above, modified to include one more variable *Region*.

Figure 6: Boxplot of relationship between Y, X1 and Region


Relationship between State Housing
Assistance Expenditure and Personal Income
Per capita personal income in state
Per capita expenditure on housing assistance in state
Region
Northeast
North Central
South
West