Problem Set 1

Ana Sofia Jesus 19327602

01/10/2023

Question 1

Part 1

In order to calculate a confidence interval I first performed a descriptive analysis on the data: mean(y): The mean of the IQ score is the sum of all scores divided by the number of students

var(y): The variance informs about the degree of dispersion of the data points around the mean.

sd(y): The standard deviation measures how much the IQ scores deviate from the mean

sd(y)/sqrt(length(y)): The standard error tells us how likely is one sample to change from one random sample to the other; it is the measure of uncertainty

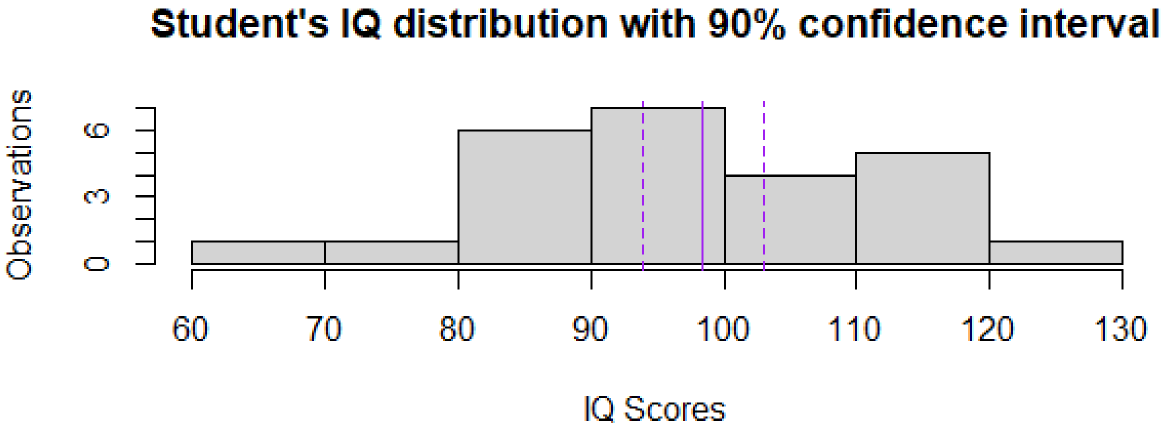
length(y): Number of observations

From these codes, I found that the mean = 98.44; the standard deviation = 13.09, and the sample size (n) = 25

Confidence intervals are used to quantify the uncertainty of estimating parameters in the population based on samples of data. To calculate it here I require the Sample Mean, the critical value and the standard error (SE)

```
qt(0.010, df=length(y)-1)  critical value for first 10%
qt(0.90, df=length(y)-1)  last 10%
qt(0.010, df=length(y)-1, lower.tail=FALSE)  last 10%

t_score <- qt(0.95, df=length(y)-1)
lower_90_t <- mean(y)-(t_score)*(sd(y)/sqrt(length(y)))
upper_90_t <- mean(y)+(t_score)*(sd(y)/sqrt(length(y)))


These bounds contain the lower and upper limits of the 90% confidence interval for the
average IQ in my sample.

lower_90_t
upper_90_t

The 90% confidence interval for the population mean IQ score is (93.95993, 102.9201).
Therefore, I can conclude with 90% confidence, that the true population mean IQ score is
likely to fall within the range of 93.95993 to 102.9201.
```

I have created a histogram to display these results:



0.1 Part 2

Before beginning a hypothesis test, it is necessary to formulate the associated hypotheses:

- Null Hypothesis (H0): The average student IQ in the school is lesser than or equal to the national average IQ score.

- Alternative Hypothesis (Ha): The average student IQ in the school is higher than the national average IQ score.

Next, I calculate the t-statistic using the following code:

```
t_stat <- (mean(y)-100)/(sd(y)/sqrt(length(y)))
```

I obtain this result -0.595743942057347. The t-statistic measures the standard errors between the sample mean and the hypothesized population mean (100) under the assumption that the null hypothesis is true. This negative result indicates that the sample mean is lower than the hypothesized population mean.
Next, I calculate the p-value using the following R code:

```
P_value <- pt(abs(t_stat), df = length(y)-1, lower.tail = FALSE)
```

The p-value is a measure of the evidence against a null hypothesis in my statistical hypothesis test. This argument is set to FALSE because I am interested in calculating the probability in the upper tail of the t-distribution.

Since p-value (0.278461658037606) is superior to alpha (0.05), I do not possess sufficient evidence to reject the null hypothesis. A large p-value suggests weak evidence against H0.

I have confirmed my result through an alternative code:

```
t.test(y, mu = 100, alternative = 'less')
```

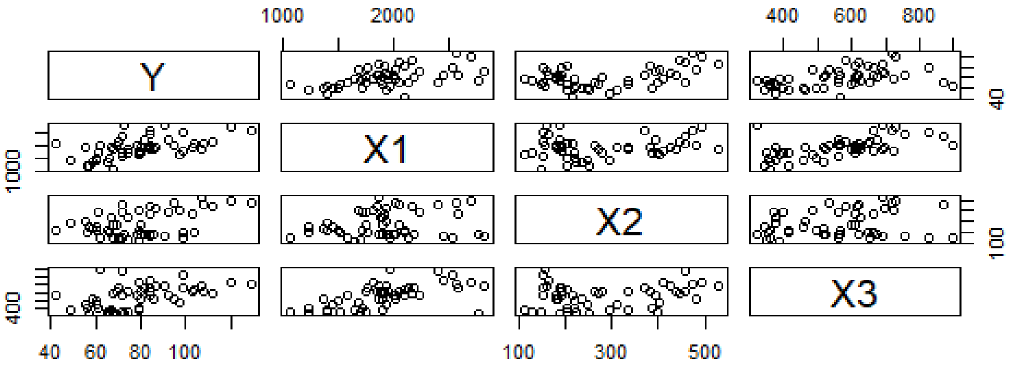
Question 2

Part 1

In order to plot the relationship between variables Y, X1, X2, and X3, I created a scatter plot matrix using the following code:

```
pairs(expenditure[, c("Y", "X1", "X2", "X3")])
```

I obtained the following plot:



Next, I calculated the correlations between all of these variables using the following code:

```
correlations <- cor(expenditure[, c("Y", "X1", "X2", "X3")])
```

To visualise the strength and direction of correlations I created a heat map. In this map, red represents positive correlations, blue negative correlations, and white for no correlation. The intensity of the color represents the strength of the correlation.

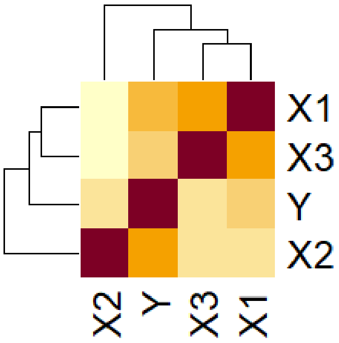


Figure 1: Heat Map of Correlations between Y, X1, X2 and X3

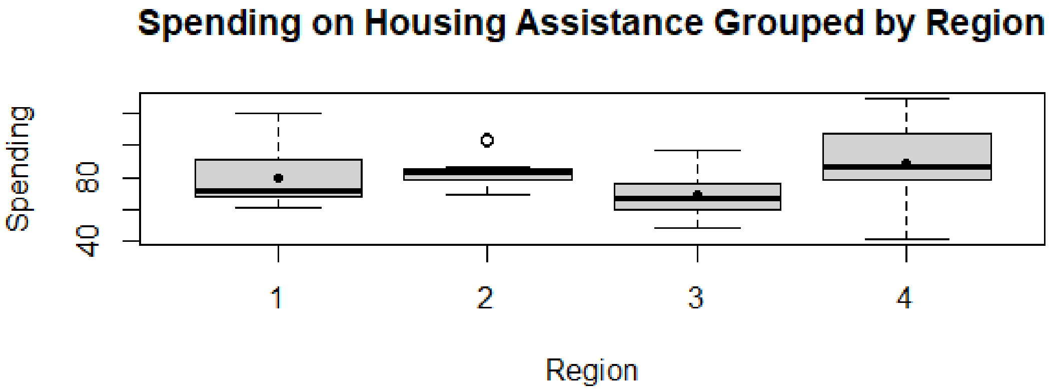
By analysing the scatter plots side by side with the correlation value and the heat map, it is possible to conclude that correlations between Y, X1, X2, and X3 are all positive and weak to moderate. The strongest correlation is a moderate positive correlation between X1 and X3, followed by the positive moderate correlation between Y and X1. Following those, there are other two moderate positive correlations between Y vs. X2 and Y vs. X3. The remaining correlations are all weak or random, being situated below 0.40. Some plots have many outliers, such as Y vs. X3, which could be distorting the correlation.

Part 2

For this exercise, I created a box plot to visualise the relationship between per capita spending on housing assistance (Y) and regions (Region). I used the following code:

```
boxplot(Y ~ Region, data = expenditure, main = "Spending on Housing Assistance Grouped by Region",
ylab = "Spending")
means <- tapply(expenditure$Y, expenditure$Region, mean)
points(means, pch = 20) # add means as circles to each boxplot
```

I obtained the following plot:



Even though the boxplot provides a visual aid to answering the question at hand, I will compare the mean of the 4 regions. For this, I use the following code to first create a list to store mean expenditures for each region.

```
region_means <- list()
```

Next, I use the following code to calculate the mean expenditure for each region using a loop function:

```
  for (region_id in unique(expenditure$Region)) {
region_data <- expenditure[expenditure$Region == region_id, ]
mean_expenditure <- mean(region_data$Y, na.rm = TRUE)
region_means[[as.character(region_id)]] <- mean_expenditure}
```

I obtained the following mean expenditures for each region:

-Region 1 79.44444

-Region 2 83.91667

-Region 3 69.1875

-Region 4 88.30769

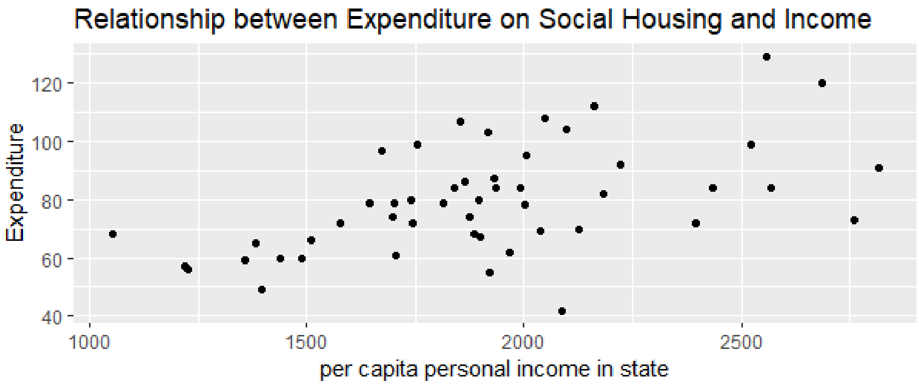
From these results it is possible to conclude that Region 4 (West) has the highest per capita spending on social housing. With Region 2 (North Central) coming in second place, with Region 1 (North East) in third and Region 3 (South) having the lowest per capita spending on social housing.

Part 3

First, I created a plot of Expenditure per personal income in state, using ggplot:

```
ggplot(data = expenditure, aes(y = Y, x = X1)) +
  geom_point() +
  ggtitle("Relationship between Expenditure on Social Housing and Income") +
  xlab("per capita personal income in state")+
  ylab("Expenditure")
```

I obtained the following plot:



From this plot it is possible to anticipate a positive correlation between the two variables. Thus, as personal income per capita increases so does the expenditure on shelters and social housing.

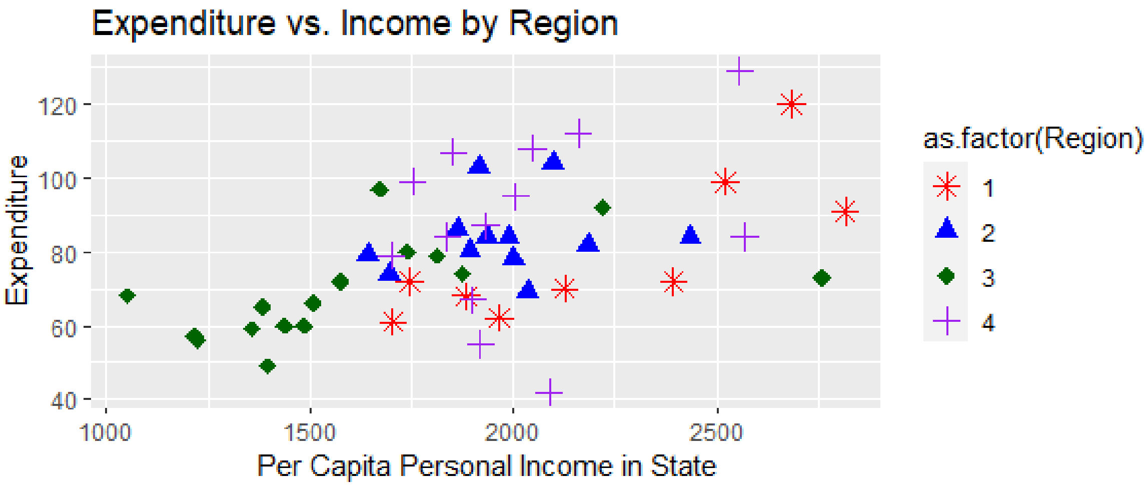
Second, I created a ggplot of the three variables combined:

```
ggplot(data = expenditure, aes(x = X1, y = Y, color = as.factor(Region), shape = as.factor(Region))
  geom_point(size = 3) +
  ggtitle("Expenditure vs. Income by Region") +
  xlab("Per Capita Personal Income in State") +
  ylab("Expenditure") +
```

And I added a customisation of color and shape:

```
scale_color_manual(values = c("1" = "orange", "2" = "blue", "3" = "darkgreen", "4" = "purple")) +
scale_shape_manual(values = c("1" = 8, "2" = 17, "3" = 18, "4" = 3))
```

I have obtained the following customised visualisation of the interaction between the 3 variables: Expenditure, Income and Region



From this plot, there seems to be a moderate positive correlation between the per capita income in the state and the per capita spending on housing assistance. It is clear from the graph that Region 4 is the highest spending region in housing assistance within its range of income. Adversely, it is also possible to note that region 1 generally spends less on housing assistance than other regions with similar income. In this graph it also becomes clear that a possible explanation for region 3 poor expenditure on housing assistance can be a low income per capita.