Roisin Gallagher - Problem set 1

September 30, 2023

1 Question 1

1. For this question, the first step is to read the vector data. The next step is to find the confidence interval, for which I used the following formula: In this formula, 'average X' represents the sample mean, 't' is the critical value from the t-distribution for the desired confidence level, 's' is the sample standard deviation, and 'n' is the sample size. Then, I calculated the mean and standard deviation of the sample using the R code provided in the accompanying file.

To find the critical value, I consulted the t-table and found it to be 1.71. I considered the degrees of freedom as 'n-1,' which is equal to '25-1 = 24.' Since the question asked for a 90% confidence interval, I looked at the values for '1-0.9 = 0.1.' By plugging these values into the formula, I calculated the lower and upper bounds of the confidence interval, which are equal to 93.96 and 102.91, respectively.

Confidence Interval = average X $\pm t\left(\frac{s}{\sqrt{n}}\right)$

2. For this part of the question, I performed a one-sample t-test first to obtain the test statistic and p-value, which are both available in my R code. The test statistic is -0.5957439, and the p-value is 0.7215383, respectively. Then, I compared the p-value to 0.05 to determine whether to reject the null hypothesis. I found that we fail to reject the null hypothesis (H0), indicating that there is not enough evidence to conclude that the average student IQ in the school is greater than 100.

Null hypothesis: The average student IQ in her school is equal to the average IQ score of 100 among all schools in the country.

Alternative hypothesis: The average student IQ in her school is higher than the average IQ score of 100 among all schools in the country.

2 Question 2

1. To answer this question, I have read the expenditure dataset in R as a dataframe. Using the ggplot package, I have plotted the relationship between y and X1, X2, and X3. Additionally, I have calculated the correlation between each pair of variables using the cor() function. The R code for these analyses can be found in the R file.

The relationship between y and X1 shows a positive correlation between these two variables, with a correlation coefficient of 0.53.

The relationship between y and X2 exhibits a non-linear association between these two variables. However, when considering only the correlation value, it is positive, with a correlation coefficient of 0.44.

The relationship between y and X3 also demonstrates a positive correlation between these two variables, with a correlation coefficient of 0.46.

See Figures 1, 2 and 3.

2. To answer this question, we need to create a boxplot. To do that, we first need to convert the data type of the 'Region' column in the expenditure dataset to a factor. Then, by using the code provided in the R file,

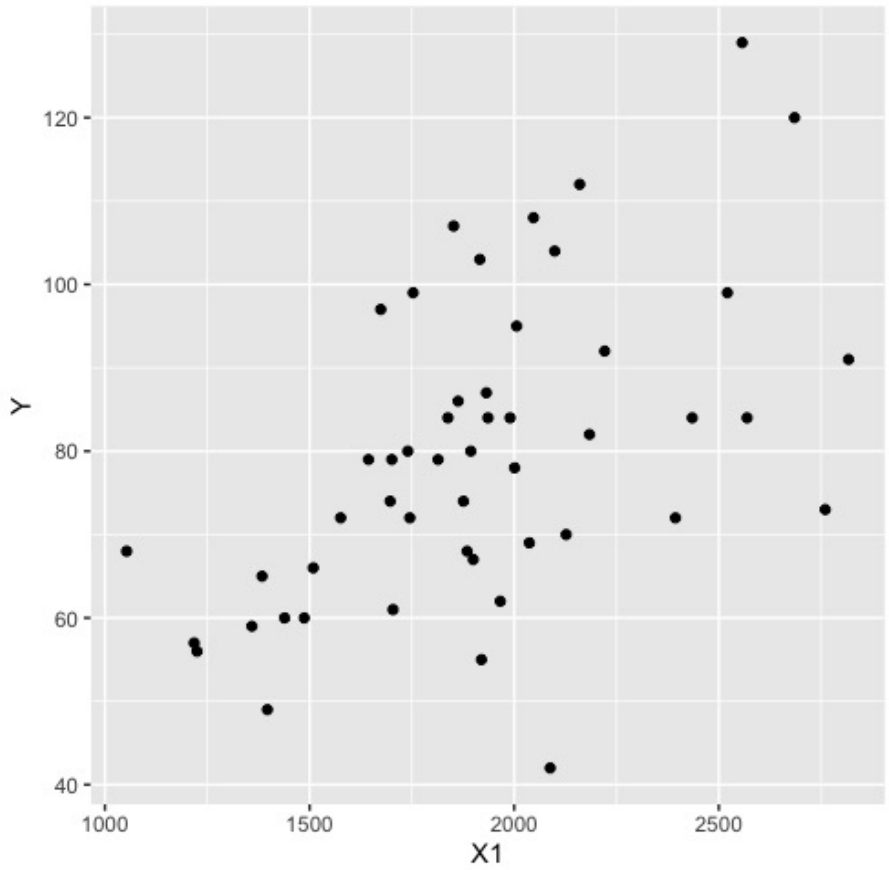

Figure 1: Y vs X1

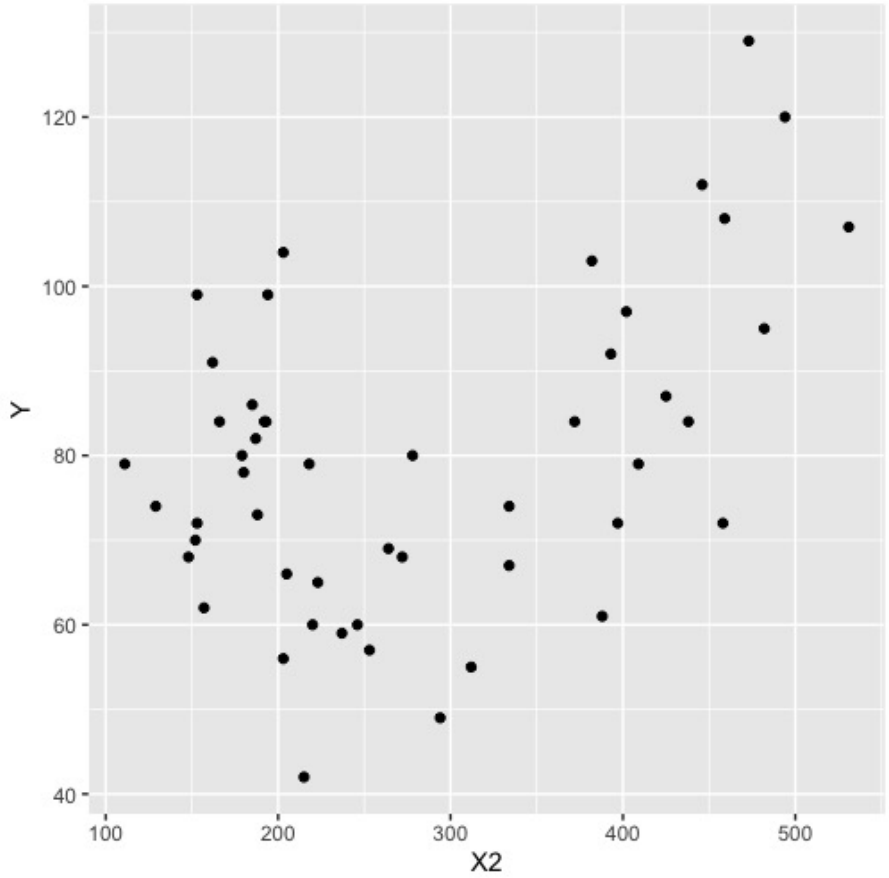

Figure 2: Y vs X2

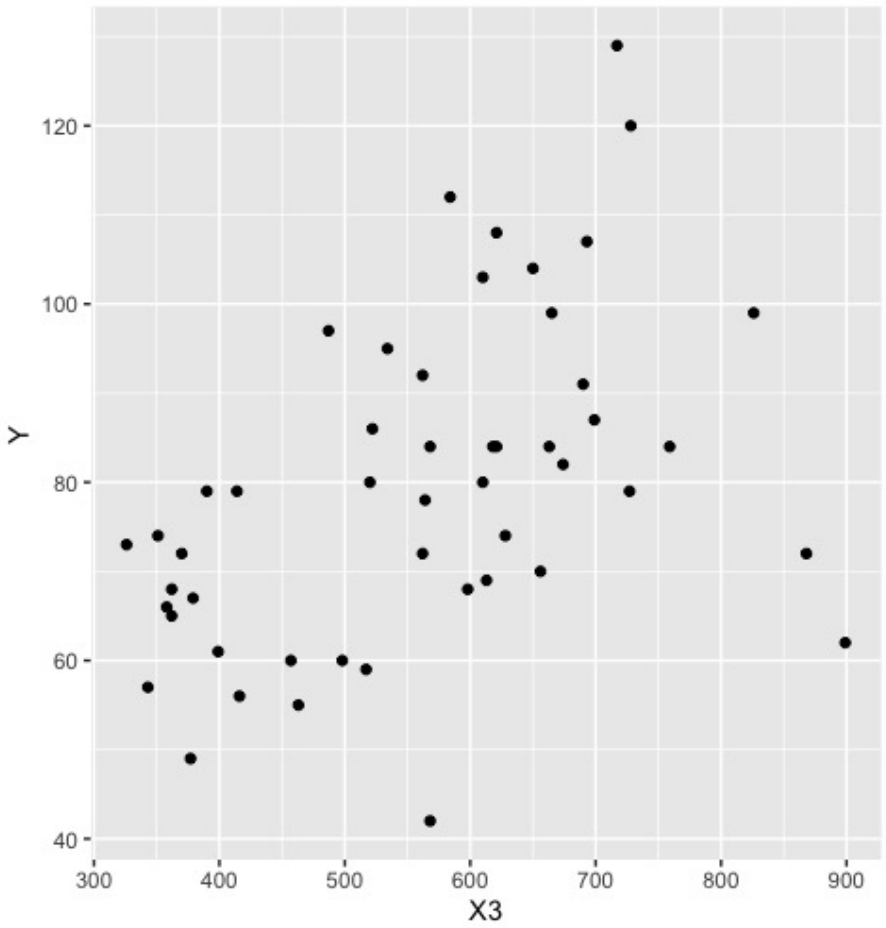

Figure 3: Y vs X3

we can generate the following plot. Based on this plot, it is evident that Region 4 has the highest average compared to the other regions. See Figure 4.

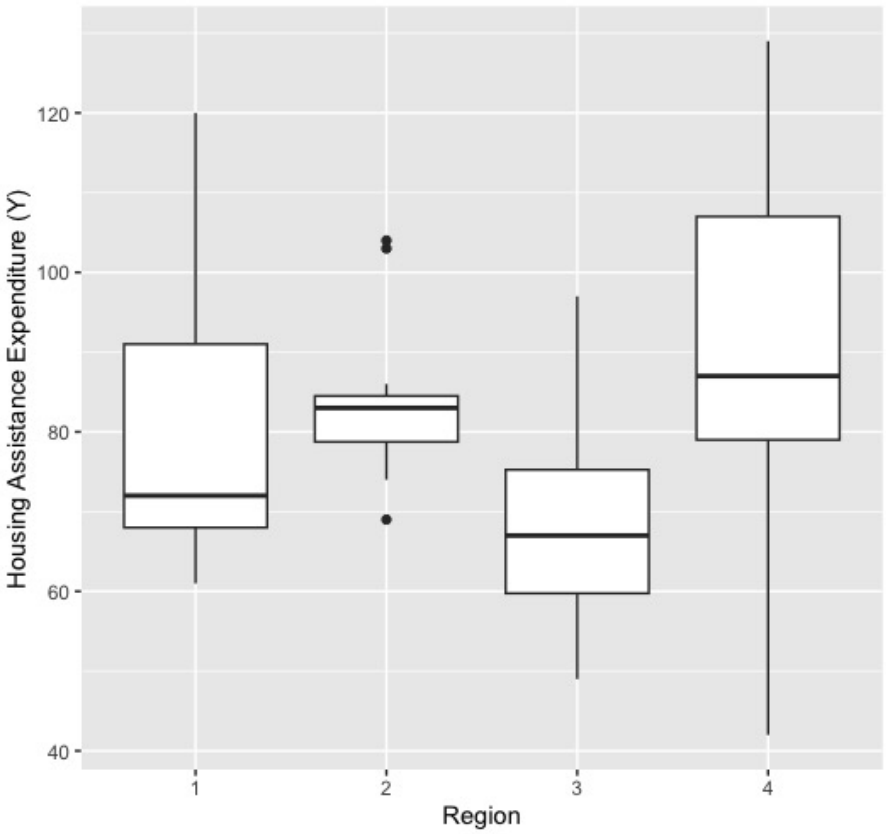

Figure 4: Boxplot

3. I have already plotted the first part of this question in the answer to the first part of question 2. Please see Figure 1.

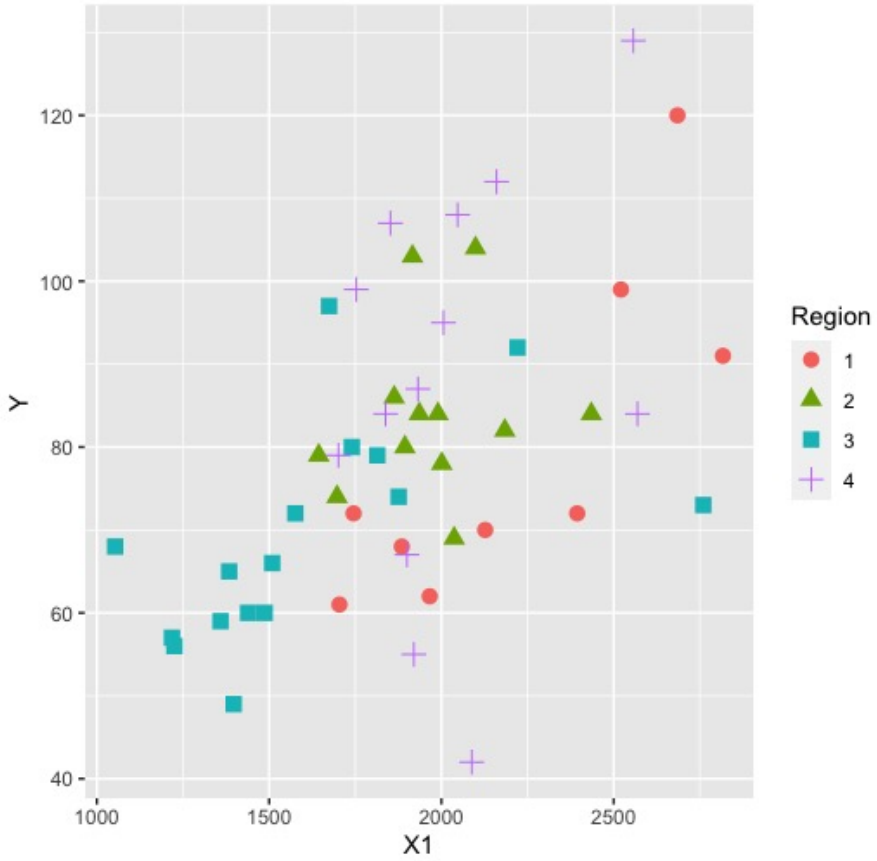

Figure 5: Yvs X1 with Regions