Problem Set 1

Philip Kruger
18328699
01/10/22

Question 1
Part 1

To find 90% confidence interval for mean student iQ at the school, we first need to find the mean and standard deviation of the sample. This is done with the following code.

```
y_bar = sum(y)/length(y)  #mean

sum_errors <- NULL   #sum of errors
for(i in 1:length(y))
{ sum_errors[i] <- y[i] - mean(y)}

sum_error_sq <- sum_errors^2   #sum of errors squared

variance <- (sum(sum_error_sq))/(length(y)-1)  #variance

st_dev <- sqrt(variance)   #standard deviation
```

from this we find that the mean is 98.44 and the standard deviation is 13.0928733795654.
alternatively the mean and standard deviation can be found using the following functions in r which verify the results found above:

```
y_bar = mean(y)
st_dev = st(y)
```

To calculate the confidence interval we need to find the margin away from mean on both sides. That is what the below line does.

```
conf_int <- 0.9  # assigning conf int value
margin <- qt((1-(1-conf_int)/2), df = length(y) -1)*sd(y)/sqrt(length(y))
```

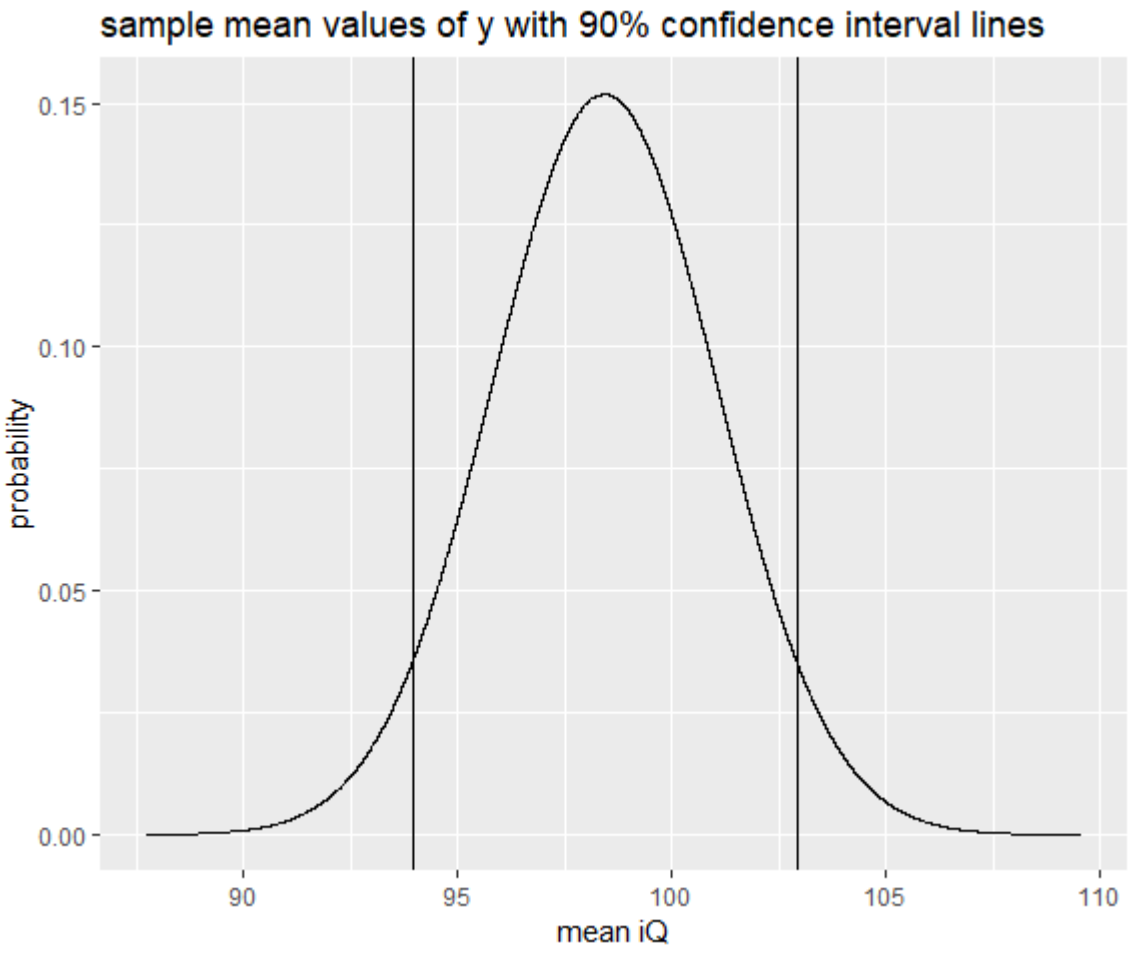
we then add and subtract this margin to/from the mean with the following lines:

```
lower_int = mean(y) - margin
upper_int = mean(y) + margin
```

which gives us the result:

Lower Interval	93.9599275120757
Upper Interval	102.920072487924

this certainly seems like a plausible result however just to see if these results make sense visually I graphed them against the mean of 100,000 samples taken from the normal distribution resulting from sample set y.



These certainly seem like plausible answers for a 90% confidence interval.

Part 2

When carrying out a hypothesis test we first need to be aware of the assumptions taken to begin to evaluate the null hypothesis. Some of the major assumptions are as follows:
The the IQ test completed by the students accurately determines their iQ.
The average iQ score of all the schools in the country was 100.
That the sample group was a roughly fair sampling of student iQs.

Our Null Hypothesis and alternative Hypothesis are as follows:

$$H_0 : \mu \leq 100$$
$$H_a| : \mu > 100$$

the following R code was then used to find the test statistic and it was determined that the test statistic is equal to -0.595743942057347.

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

Using this we calculate the p-value for a one sided t-test of the null hypothesis. This is done with the following code:

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

From this we get that the p-value is 0.278461658037606. This is greater than our α which is equal to 0.05. Thus we don not have sufficient evidence to dismiss our null Hypothesis.
To confirm our answer we can also use the *t.test()* function in R which does all the calculations for us:

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

This results in the same values as we calculated earlier.

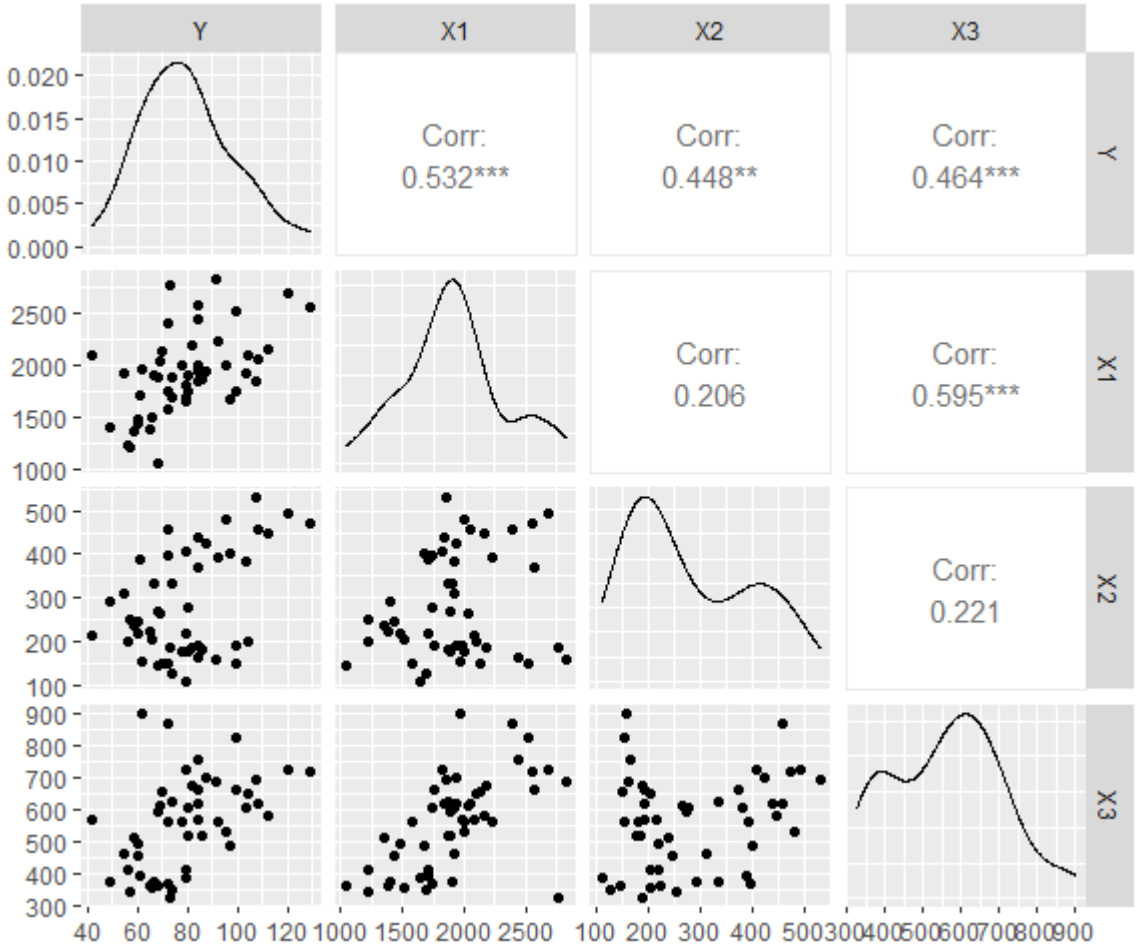
Question 2

Part 1

To plot all of relationships we use the following code which utilises the packages ggplot and GGally.

```
ggpairs(expenditure[,2:5], labels = c("Y", "X1", "X2", "X3"),
 main = "All colums plotted against eachotehr with the corresponding correlation")
```

which yields the following graph:



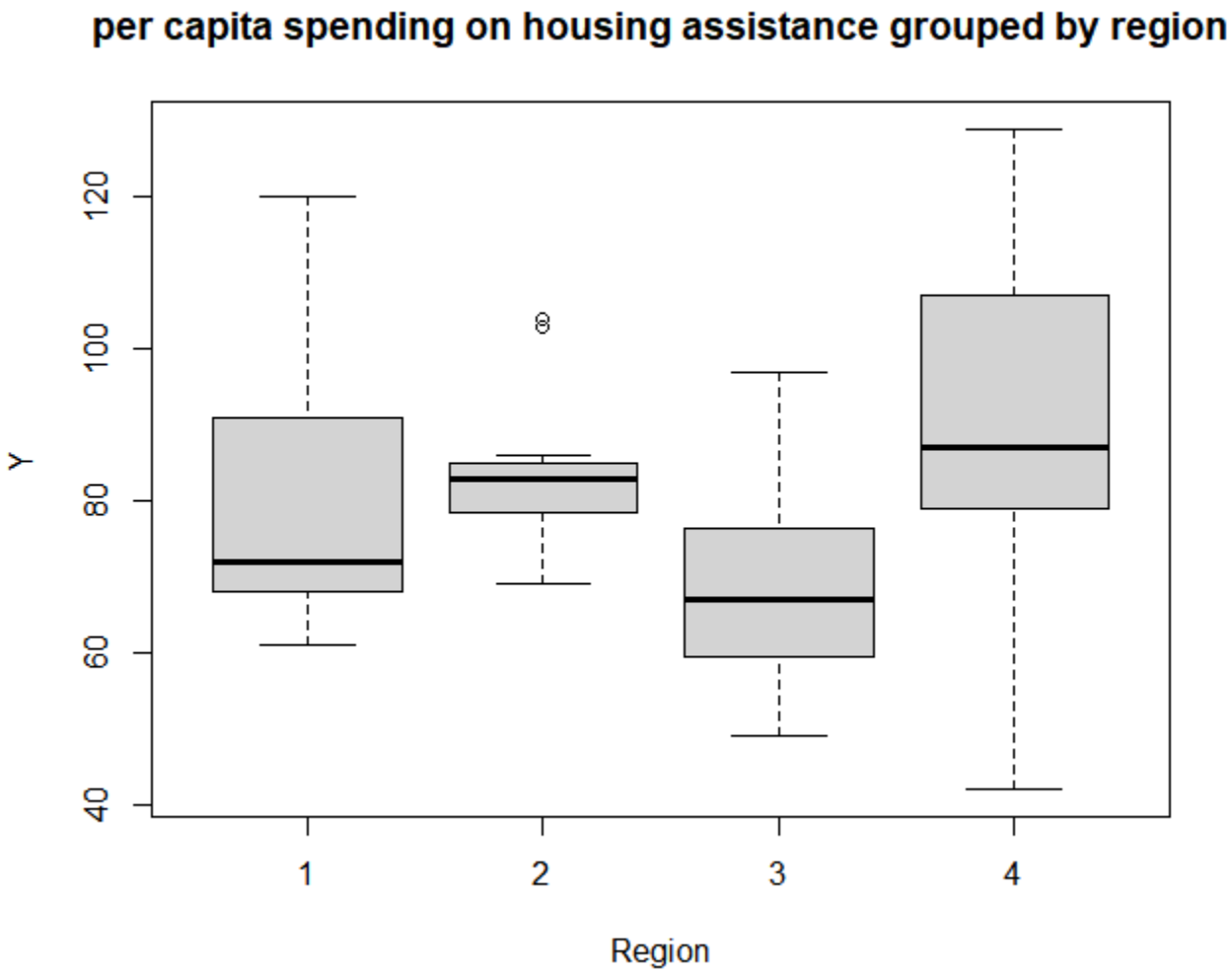
from eyeballing these plots we see that X1 vs X2 and X2 vs X3 appear to be almost random with barely any correlation. Y vs X2 and has moderate correlation with a general positive trend. Y vs X3 also has a general positive trend but there are multiple outliers which we would expect to effect the correlation coefficient. Y vs X1 and X1 vs X3 also appear to have a moderate positive relationship with fewer outliers than the previously mentioned graphs.
These guesses are backed by the correlation coefficients which show that X1 vs X2 and X2 vs X3 have very weak to random correlation. Y vs X2 and Y vs X3 have weak correlation and Y vs X1 and X1 vs X3 have moderate correlations. All have positive associations

Part 2

I used a point plot in this situation however a box plot would have also been The following code was used to graph Y vs Region:

```
boxplot(Y~Region,data = expenditure, main ="per capita spending on housing assistance grouped by region
```

This resulted in the following graph:



You could estimate the means from this graph. (however I did a point graph at first so here are the means calculated to five decimal places).
The following code was then used to find the mean:

```
for(i in 1:4)  #creating objects containing the section of the dataset which are from the same region
{
nam <- paste("Region_", i, sep = "")
assign(nam, expenditure[expenditure$Region == i,])
}
mean(Region_1$Y)  #printing the mean of Y
mean(Region_2$Y)
mean(Region_3$Y)
```

```
mean(Region_4$Y)
```

from this we get the Mean Y for each region being:

Region 1	79.44444
Region 2	83.91667
Region 3	69.1875
Region 4	88.30769

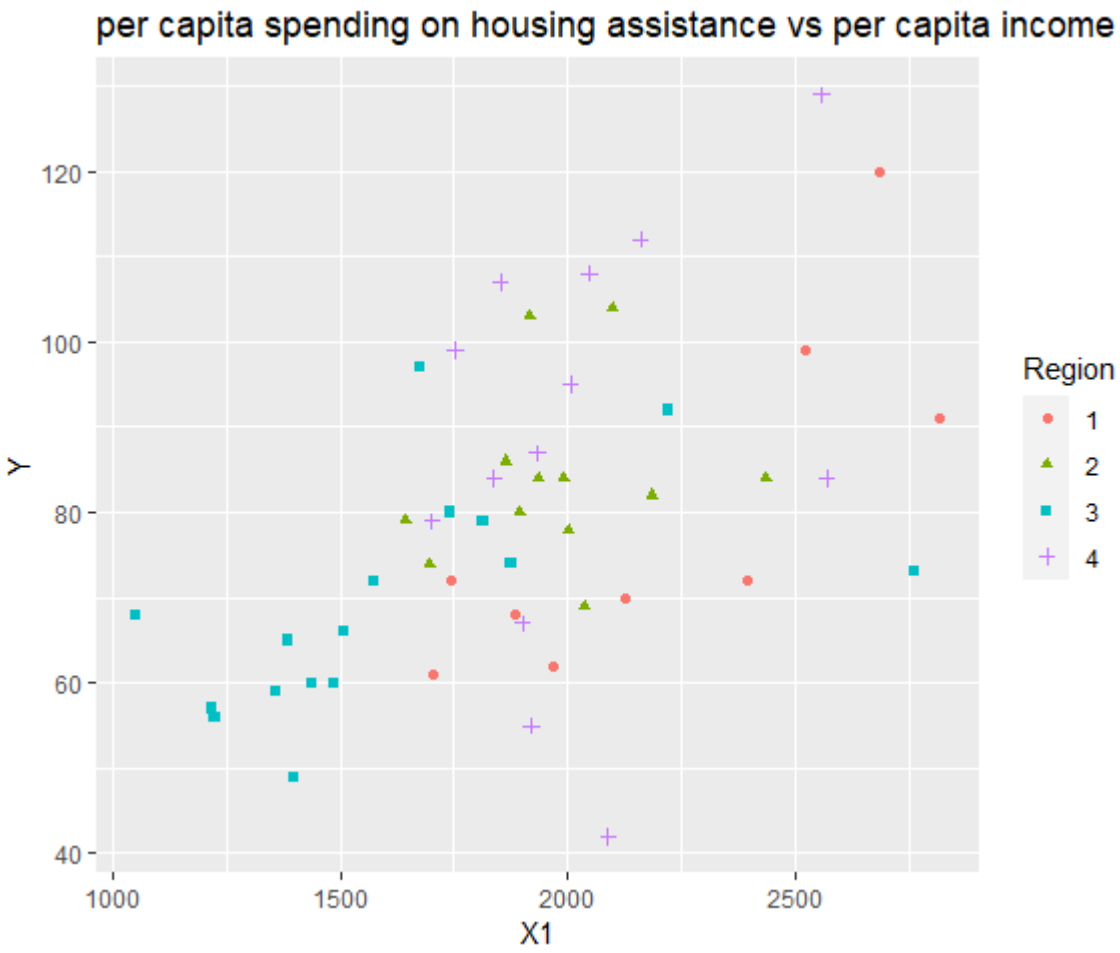
Thus we can see that Region 4 (West) has the highest per capita spending on social housing. With Region 2 (North Central) coming in second with Region 1 (North East) in third and Region 4 (South) having the lowest per capita spending on social housing.

Part 3

The graph is produced with the following code:

```
ggplot(data = expenditure) +  #its a simple plot
geom_point(mapping = aes(y = Y, x = X1, colour = as.factor(Region), shape = as.factor(Region))) +
labs(colour="Region", shape = "Region") +
ggtitle("per capita spending on housing assistance vs per capita income")
```

It yields the following graph:



It has a general moderate positive correlation. Thus, there is generally a correlation between the per capita income in the state and the per capita spending on housing assistance. We can also see that region 1 generally spends less on housing assistance than other states with similar income. In part 2, we saw that region 3 spends the least on housing assistance. From this graph we can see that a significant contributing factor may be household income as many of the lowest household income states are in region 3. We also see region 4 has a massive spread in the amount spent on housing assistance for states with similar income.