Problem Set 1

Anne Clarke

QTM 200

I did not know how to include output directly from code, so I have written in answers to questions instead, and for graphs uploaded screenshots from executed R code.

Question 1 (25 points)

A private 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:

```r
q1 <- 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)
```

Below is the code I used to to calculate the confidence interval:

```r
avgq1 <- mean(q1)

sdq1 <- sd(q1)

z90 <- qnorm((1-.90)/2, lower.tail = FALSE)

nq1 <- length(q1)

low_q1 <- avgq1 - (z90 * (sdq1/sqrt(nq1)))
up_q1 <- avgq1 + (z90 * (sdq1/sqrt(nq1)))

confintq1 <- c(low_q1, up_q1)
confintq1
```

Given repeated sampling, there is a 90% chance that the true avg IQ of students in the school, assuming a normal distribution, would lie between (94.13283 and 102.74717).

Question 2 (25 points)

A private school counselor was curious whether the average of IQ of the students in her school is higher than the average IQ score 100 among all the schools in the country. She took a random sample of 25 students' IQ scores. The following is the data set:

```
q2 <- 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)
```

Below is the code I used to to calculate the confidence interval:

```
avgq2 <- mean(q2)

sdq2 <- sd(q2)

z95 <- qnorm((1-.95)/2, lower.tail = FALSE)

nq2 <- length(q2)

low_q2 <- avgq2 - (z95 * (sdq2/sqrt(nq2)))
up_q2 <- avgq2 + (z95 * (sdq2/sqrt(nq2)))

confintq2 <- c(low_q2, up_q2)
confintq2
```

Given repeated sampling, there is a 95% chance that the true avg IQ of students in the school, assuming a normal distribution, would lie between (93.30769 and 103.57231).

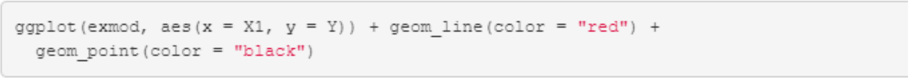
Question 3 (50 points)

Researchers are curious about what affects the education expenditure on public education. The following is available variables in a data set about the education expenditure.

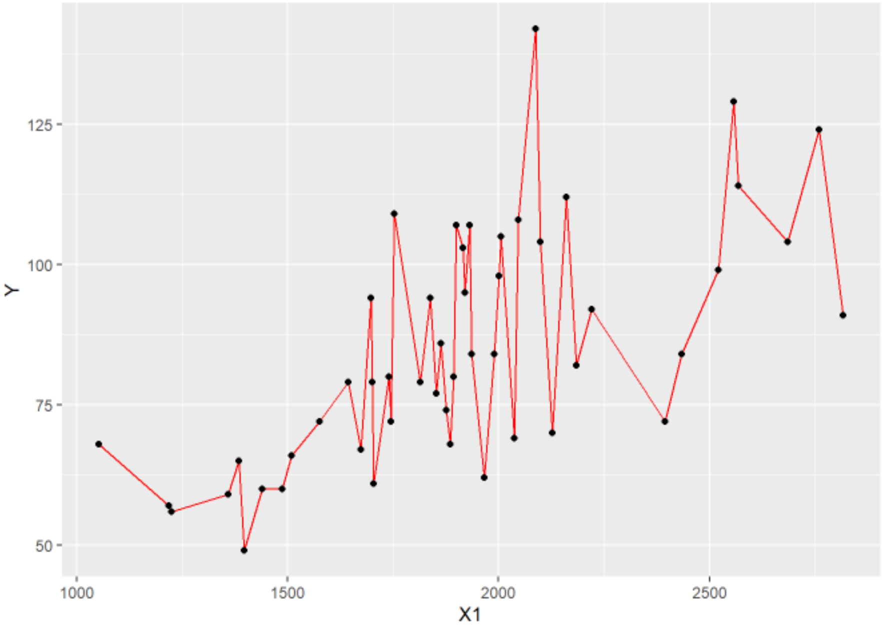
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?
  First, I imported the data into R:

  ```
  expenditure <- read.table("expenditure.txt", header = T)
  head(expenditure)
  ```

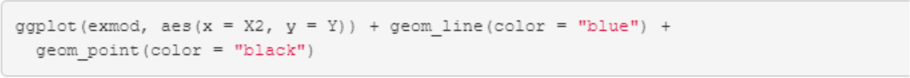
  Here is the first graph, in which I plot Personal Income per Capita vs Expenditure on Public Education per Capita. There seems to be a positive correlation between expenditure and personal income, with some outliers.



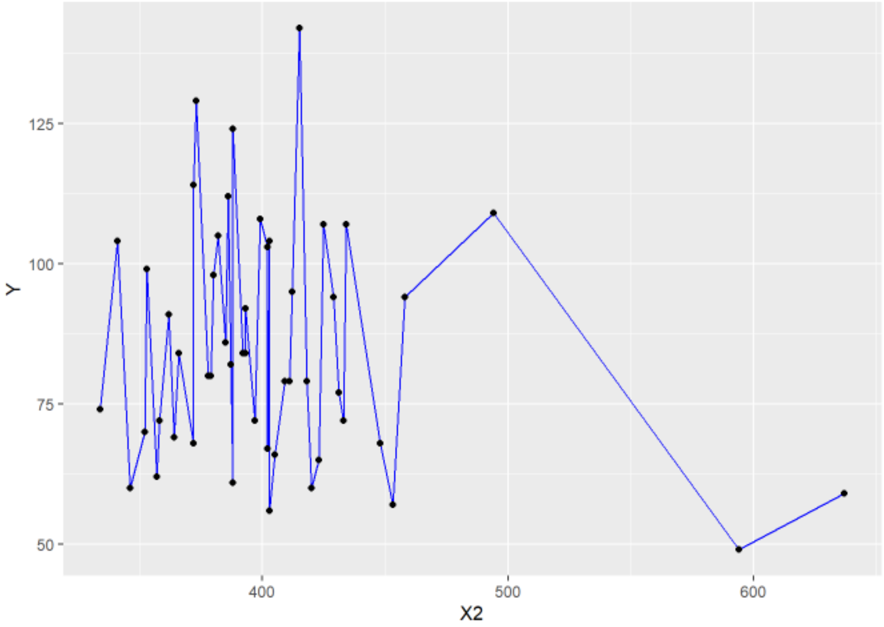
```
ggplot(exmod, aes(x = X1, y = Y)) + geom_line(color = "red") +
  geom_point(color = "black")
```



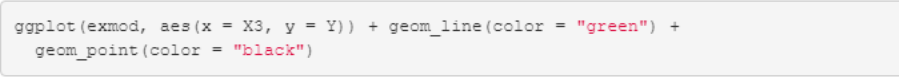
Here is the second graph, in which I plot The Number of Residents/1000 under 18 Years of Age vs Expenditure on Public Education per Capita. There seems to be almost no correlation btwn these two variables.



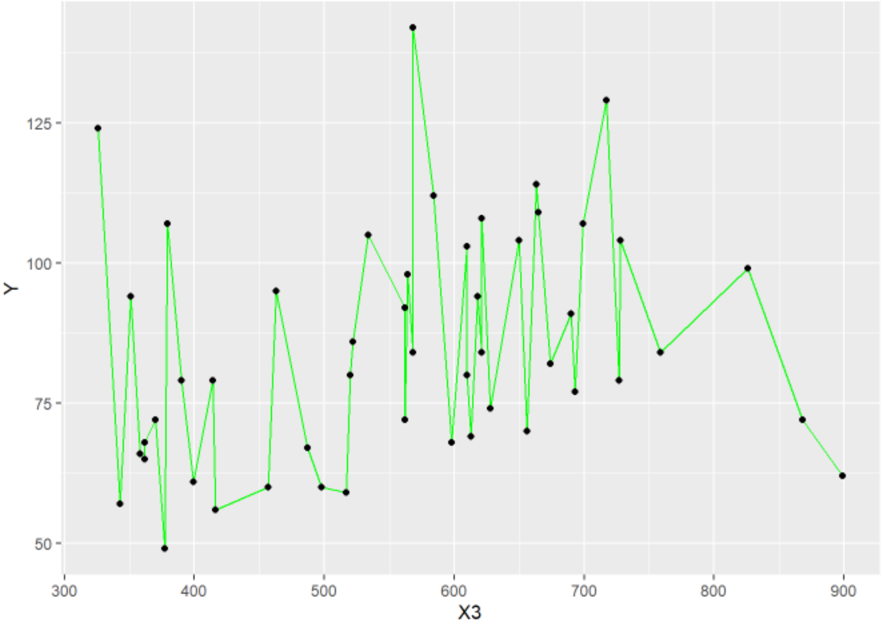
```
ggplot(exmod, aes(x = X2, y = Y)) + geom_line(color = "blue") +
  geom_point(color = "black")
```



Here is the first graph, in which I plot The Number of People/1000 Residing in Urban Areas vs Expenditure on Public Education per Capita. There seems to be a positive correlation between expenditure and people/1000 residing in urban areas, with some outliers.



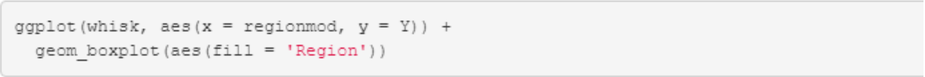
```
ggplot(exmod, aes(x = X3, y = Y)) + geom_line(color = "green") +
  geom_point(color = "black")
```



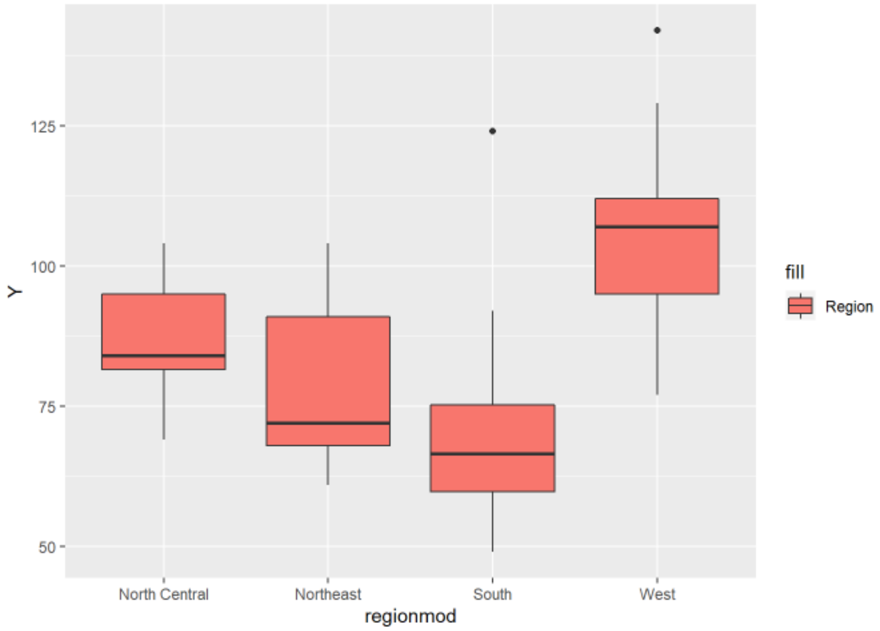
- Please plot the relationship between Y and *Region*? On average, which region has the highest per capita expenditure on public education? Here is the R code used to generate this graph:

```
whisk <- exmod %>%
  select(regionmod, Y)

NEwhisk <- whisk %>%
  filter(regionmod == 'Northeast')

NCwhisk <- whisk %>%
  filter(regionmod == 'North Central')

Swhisk <- whisk %>%
  filter(regionmod == 'South')

Wwhisk <- whisk %>%
  filter(regionmod == 'West')

ggplot(whisk, aes(x = regionmod, y = Y)) +
  geom_boxplot(aes(fill = 'Region'))
```

Here is the graph I produced, which is a box-and-whisker plot that shows the relationship between Per Capita Expenditure on Public Education and Region. It shows that on average the West spends the most on public education.



```
ggplot(whisk, aes(x = regionmod, y = Y)) +
  geom_boxplot(aes(fill = 'Region'))
```

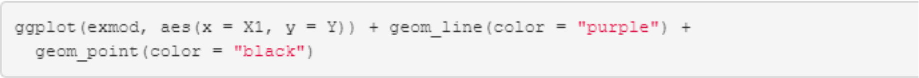


- Please plot the relationship between *Y* and *X1*? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.

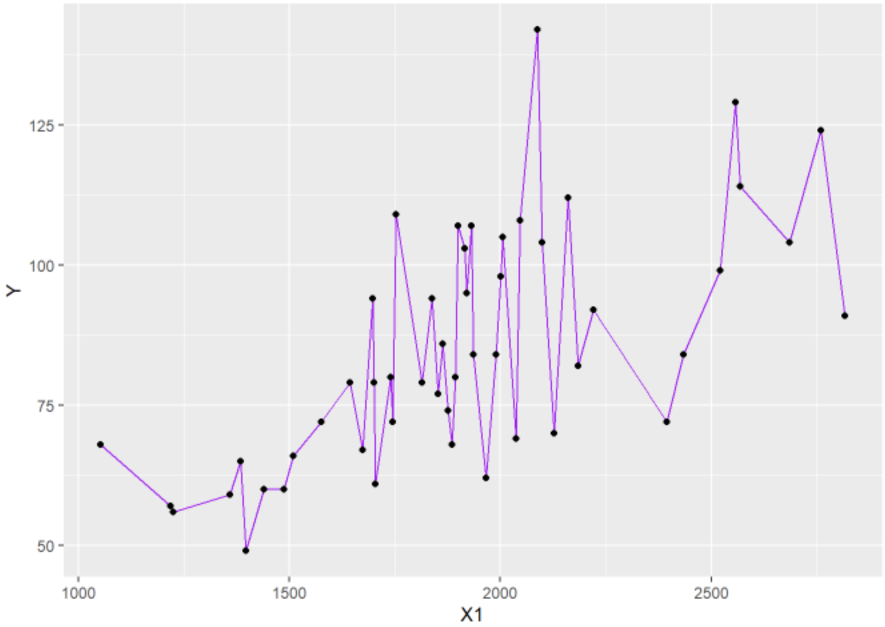
  Here is the R code used to generate this graph:

```
ggplot(exmod, aes(x = X1, y = Y)) + geom_line(color = "purple") +
  geom_point(color = "black")
```

Here is the first graph I produced which plots Personal Income per Capita vs Expenditure on Public Education per Capita. There seems to be a positive correlation between expenditure and personal income, with some outliers.

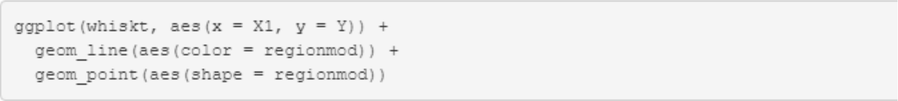


```
ggplot(exmod, aes(x = X1, y = Y)) + geom_line(color = "purple") +
  geom_point(color = "black")
```

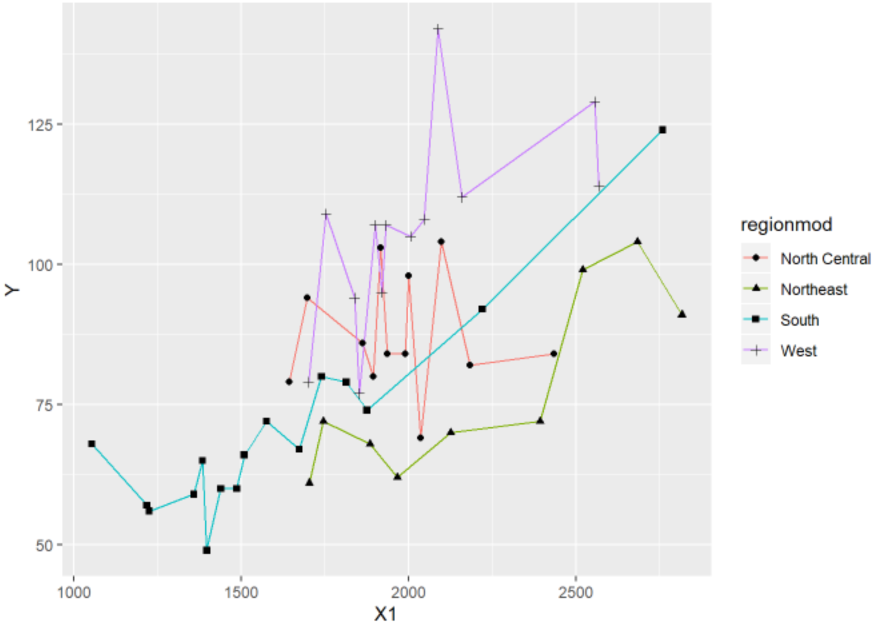


Next, I incorporate the Region variable into the same type of line graph, resulting in one line for each region. Below is the code and the resultant graph:

```
whiskt <- exmod %>%
  select(regionmod, Y, X1)

NEfin <- whiskt %>%
  filter(regionmod == 'Northeast')

NCfin <- whiskt %>%
  filter(regionmod == 'North Central')

Sfin <- whiskt %>%
  filter(regionmod == 'South')

Wfin <- whiskt %>%
  filter(regionmod == 'West')

ggplot(whiskt, aes(x = X1, y = Y)) +
  geom_line(aes(color = regionmod)) +
  geom_point(aes(shape = regionmod))
```



```
ggplot(whiskt, aes(x = X1, y = Y)) +
  geom_line(aes(color = regionmod)) +
  geom_point(aes(shape = regionmod))
```


Y
X1
regionmod
North Central
Northeast
South
West