Problem Set 1

Tolga Bag

Due: October 1, 2023

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in R, please include the code you used to get your answers. Please also include the .R file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub.
- This problem set is due before 23:59 on Sunday October 1, 2023. No late assignments will be accepted.
- Total available points for this homework is 80.

Question 1 (40 points): Education

A 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
y <- 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)
```

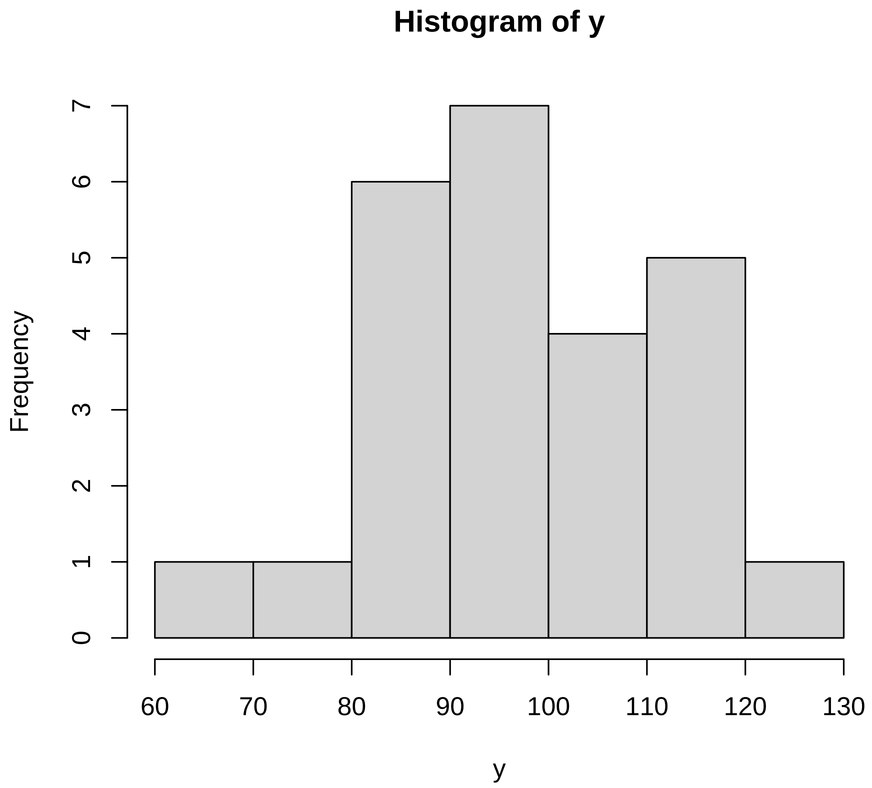
1. Find a 90% confidence interval for the average student IQ in the school.

   I calculated the mean, variability, standard deviation and standard error. I made them vectors for convenience.

```r
mean(y)
meany <- mean(y)
vary <- var(y)
sdv <- sd(y)
sde <- sdv/sqrt(length(y))
```

I check the distribution of the y. It is normal distribution per Figure 1. I calculate the quantile percentages per the 90% confidence interval.

Figure 1: Distribution of Y



```r
qnorm(0.05) # value for first 5%
qnorm(0.95) # value for last 5%
```

I calculate the lower and upper bound for 90% confidence level.

```r
print(lower_90)
print(meany)
print(upper_90)
```

Based on a 90% confidence interval calculation for y, the CI ranges from 94.13283 to 102.7472 and the mean is 98.44.

2. Next, the school counselor was curious whether the average student IQ in her school is higher than the average IQ score (100) among all the schools in the country.
   Using the same sample, conduct the appropriate hypothesis test with $\alpha = 0.05$.

   Our hypothesis is average student IG is higher than the one in the country (100) and our null hypothesis is it is lower. P is lower or equals to 0.05. I already have mean (meany) and standard deviation (sde) as vectors. Before the p score, I need to find the z score.

```
yzscore <- (meany-100)/1
print (yzscore)
```

   I have the zscore as -1.56. I calculate the p score:

```
ypscore <- 2-pnorm(-abs(yzscore))
print (ypscore)
```

   Our null hypothesis is correct since p is 1.94062 and lower than 5. We reject our hypothesis and can say that the average student IQ in the counselor‘s school is lower than the average IQ score (100) among all the schools in the country.

Question 2 (40 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*

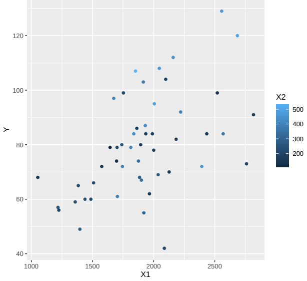
Explore the `expenditure` data set and import data into `R`.

- 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)?

  After reviewing the table and adding Y, X1, X2, X3 as vectors for convenience, I researched online to plot all 4 variables. I used the ggplot approach detailed in 'https://smin95.github.io/dataviz/basics-of-ggplot2-and-correlation-plot.html' and added the relevant packages. I use a variation of the below code to plot the relationships.

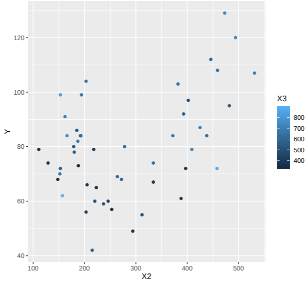
```

ggplot(data = expenditure) +
```

Figure 2: Relationship between Y, X1 and X2



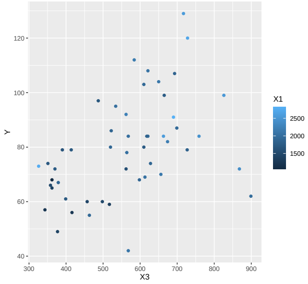
Per Figure 2, there is positive correlation between Y and X1, but X2 is independent from both. It might be more meaningful to check the Y and X2.

Figure 3: Relationship between Y, X2 and X3



Per Figure 3, there is no correlation between Y and X2, which is interesting.

Figure 4: Relationship between Y, X3 and X1



Per Figure 4, there is a positive correlation between Y and X1.

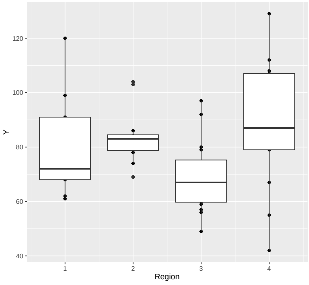
- Please plot the relationship between Y and $Region$? On average, which region has the highest per capita expenditure on housing assistance?

  After some trial and error, and online research through these webpages, I created a ggplot. References: https://ggplot2.tidyverse.org/articles/faq-axes.html and https://stackoverflow.com/questions/39628480/plotting-median-of-the-points-in-r-ggplot

```
ggplot(expenditure, aes(x = Region, y = Y, group = Region)) +
  geom_point() +
  scale_x_continuous(labels = label_number(accuracy = 1)) +
  scale_y_continuous(labels = label_number(accuracy = 1)) + geom_
    boxplot()
```

  Per Figure 5, region 4 has the highest per capita expenditure on housing assistance on average. On average, the highest per expenditure on housing assistance

Figure 5: Relationship between Y, X3 and X2



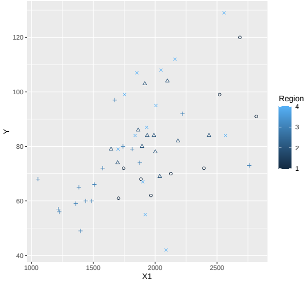
is ranked from highest: West, North Central, North East and South.

- 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.

  On average, there is a positive correlation between per capita expenditure and per capita personal income in a state as seen on ggplot5. Again, I use https://smin95.github.io/datav of-ggplot2-and-correlation-plot.html as a reference to enrich my code and add the regions Figure 6.

```
ggplot(data = expenditure) +
  geom_point(mapping = aes(x = X1, y = Y, color = Region, shape=
    Region)) +
  scale_shape_identity()
```

Figure 6: Relationship between Y, and X1 with Regions



I tried making the regions more readable on the shape but couldn't resolve the error. Nevertheless, o shape is Region 1, triangle is region 2, + is region 3 and x is Region 4 in Figure 6.