Problem Set 1

Applied Stats/Quant Methods 1
Stefan Keel, 23366536

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:

```
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)
```

Solution: The average IQ of the school is: 98.44

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

   Solution: 90% of the true mean of the schools student IQ falls between 94.13283 and 102.74717.

```r
mean.y <- mean(y) #used to calculate the mean IQ of y
print(mean.y) #control

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

sum_errors_simple <- y-mean.y #easy way of the above code

sum_errors_sqrd <- sum_errors^2 #creating the sum squarred error
sum_errors_sqrd #control

variance <- (sum(sum_errors_sqrd))/(length(y)-1) #creating the variance
variance #control

std_devi <- sqrt(variance) #standard deviation
std_devi #control


#CI creation/calculation
z90 <- qnorm((1-0.90)/2, lower.tail = FALSE) #assigning and calculating
    the margin. For both tails use /2
CI_lower_95 <- mean(y) - (z90*(sd(y)/sqrt(length(y)))) #creating the
    lower and subseq. the upper CI with margin and standard dev.
CI_upper_95 <- mean(y) + (z90*(sd(y)/sqrt(length(y))))
CI90 <- c(CI_lower_95, CI_upper_95) #putting it into one c() for easy
    print

# control-test with hannah frankes precise code adjusted for 90% CI
CI_lower_95_test <- qnorm(0.05,
                    mean = mean(y),
                    sd = (sd(y)/sqrt(length(y))))

# Upper bound, 95 confidence level
CI_upper_95_test <- qnorm(0.95,
                    mean = mean(y),
                    sd = (sd(y)/sqrt(length(y))))

CI90_test <- c(CI_lower_95_test, CI_upper_95_test)
print(CI90_test)

#Solution
mean(y)
print(CI90)
#SOLUTION 1.1: 90% of the true mean of the schools student IQ falls
    between 94.13283 and 102.74717.
```

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 $\alpha = 0.05$.

Question: whether the average school IQ is higher than the average population IQ mean.

My hypothesis: whether school IQ is higher than 100.

Assumptions:

- Conducted IQ test at school is accurate for students IQs
- Average IQ scores in country was/is 100
- Sample group is fair

Stating the Hypothesis: The hypothesis states that the average IQ of school students is higher than the average country IQ.
H0 mu is maller qual than 100
HA mu is greater than 100

```r
t_stat <- (mean(y)-100)/(sd(y)/sqrt(length(y))) #calculating t statistic
P_value <- pt(t_stat, df = length(y)-1, lower.tail = FALSE)

print(P_value)

t_test <- t.test(y, mu = 100, alternative = 'greater') #control: doing a
    checkup with t.test function
t_test #gives same p-value
```

Solution: P-Value is 0.7215383 and therefore greater than a=0.05. As the P-Value is bigger than 0.05 we accept the Null-Hypotheses.

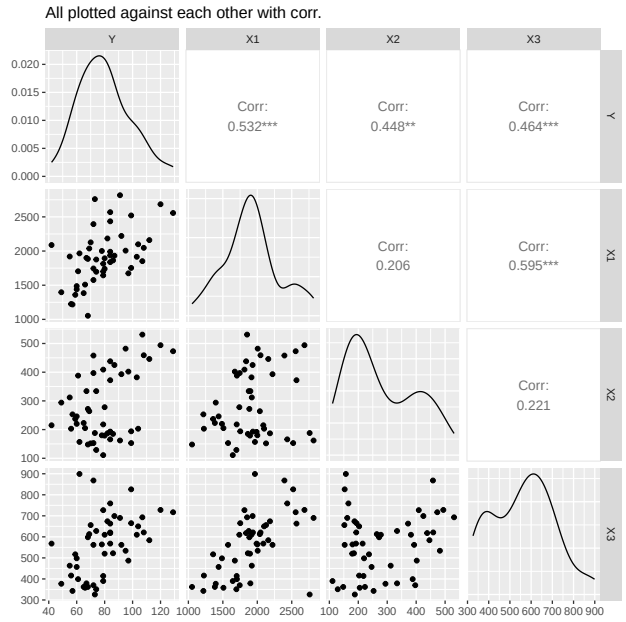
Question 2 (40 points): Political Economy

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)?

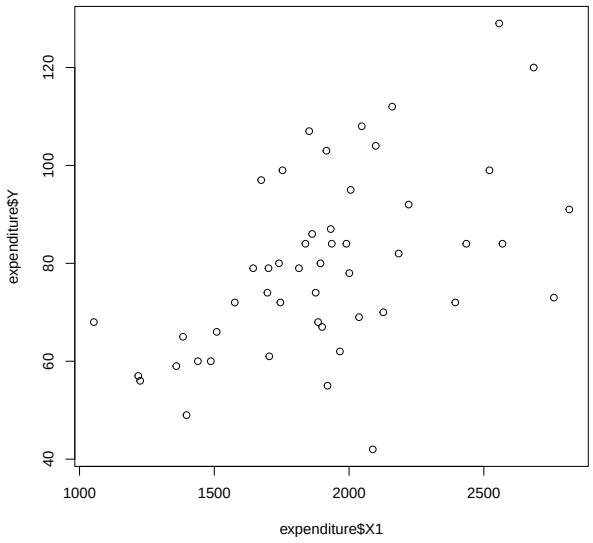


From visual interpretation:

- Strongest correlation X1 and X3 with some outliers
- Y and X1 also seem to have a good correlation but more outliers

- Y and X2 as well as X3 seem to have medium correlation
- X2 in combination with X2 and X3 seems to have very little correlation This visual interpretation is backed by the correlation numbers the plot calculated.

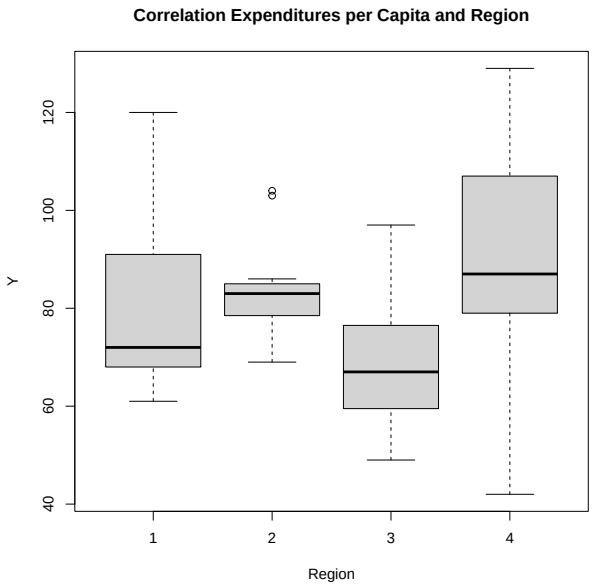
  It would also be possible to plot each combination seperately but it was too tiresome. Therefore I searched the internet for a code/package that combined it in one plot.
  Exmpaple of single plotting:




```r
set.seed(42) #setting my standard seed
library(tidyverse)
install.packages("ggplot2")
library(ggplot2)
library(GGally) #found after research on https://www.geeksforgeeks.org/how-to-create-and-interpret-pairs-plots-in-r/
#installed GGally like this with CRAN https://www.r-project.org/nosvn/pandoc/GGally.html

expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/expenditure.txt", header=T)

View(expenditure) #to get a visual overview of the dataframe

ggpairs(expenditure[,2:5]) + #this is the "cleaner"/"easier" overview that also gives correlations
  ggtitle("All plotted against each other with corr.")
#[,2:5] subsets the rows 2 until 5 which have the values for Y - X5

#alternative to GGpairs/Ggally-Plot
```

```r
plot(expenditure$Y, expenditure$X1)
plot(expenditure$Y, expenditure$X2)
plot(expenditure$Y, expenditure$X3)
#and then plot X1 vs X2 and X3 and so on. This is tiresome and would make the latex solution paper too big.
pdf("plot1.pdf")
plot(expenditure$X1, expenditure$Y)
dev.off()

pdf("plot2.pdf")
ggpairs(expenditure[,2:5]) +
  ggtitle("All plotted against each other with corr.")
dev.off()
```

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

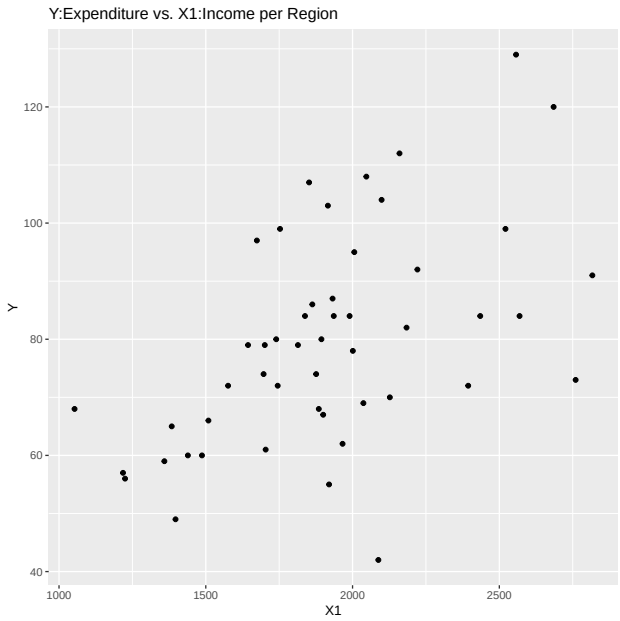


Looking at the boxplot that also indicates the mean, we can see that Region 4 has the highest per capita expenditure on housing assistance. After calculating the means of each Region it can be stated that **on average Region 4 has with 88.30769 the highest per capita expenditure on housing assistance.** Here to code in R:

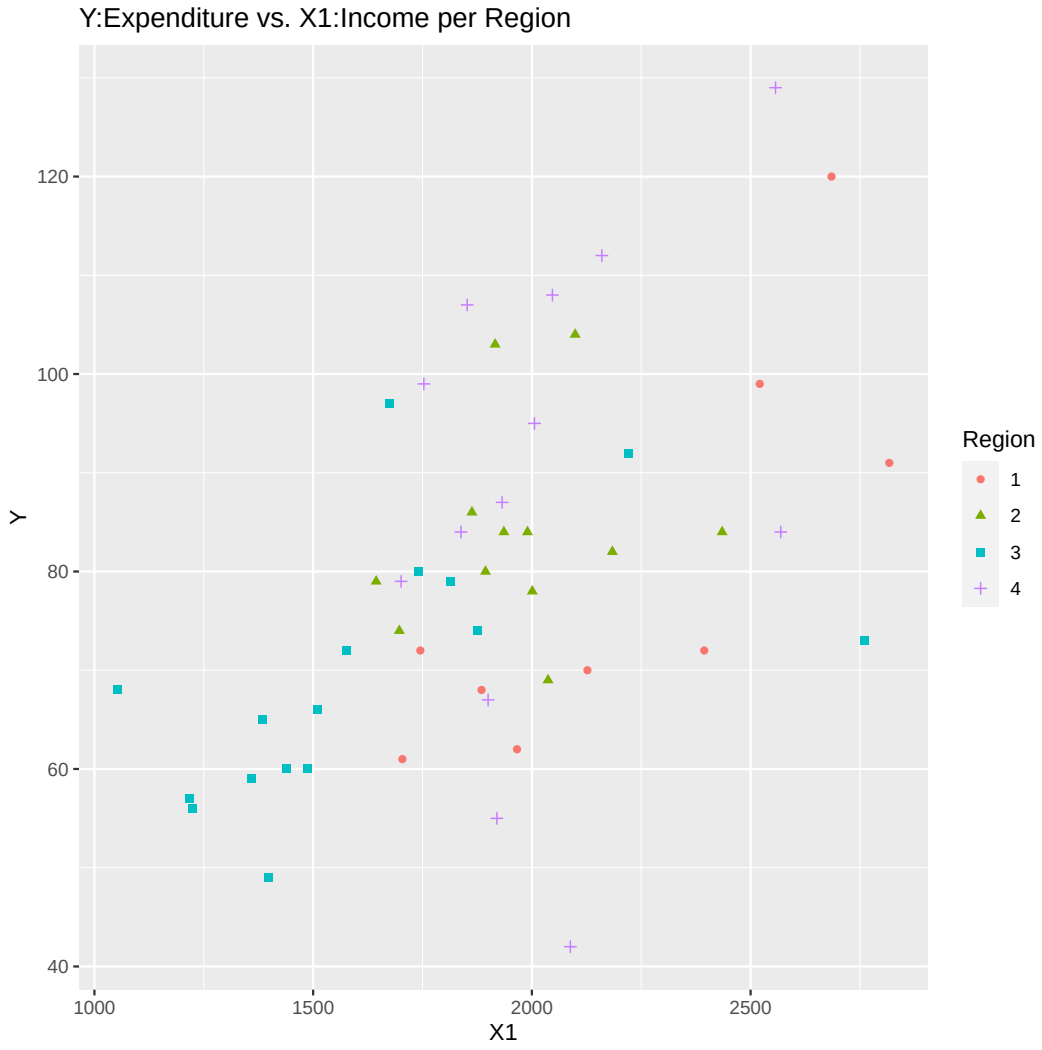
```r
pdf("plot3.pdf")
boxplot(Y~Region,data = expenditure, main ="Correlation Expenditures per Capita and Region")
dev.off()
```

```r
#Looking at the boxplot that also indicates the mean, we can see that Region 4 has the highest per capita expenditure on housing assistance. Below the R calculation for this.

#calculating the mean
for(i in 1:4) #objects of the section of expenditure of region. Of 4 Regions
{
  nam <- paste("Region", i, sep = "")
  assign(nam, expenditure[expenditure$Region == i,])
}

str(Region1) #looking at the structure of the regions to test and subset afterwards correctly

mean(Region1$Y)  #accessing and calculating the mean of Region 1 Line Y
mean(Region2$Y)
mean(Region3$Y)
mean(Region4$Y)

#could also be done for X1 Income and all other
mean(Region1$X1)
```

- 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.



As mentioned before Y and X1 have a moderate correlation with some outliers. Most values lie between Y: 60 to 90 and X1: 1500 - 2000.



3. Region 1 (Mean Expenditure: 79.44) is spread the furthest although most of the points are with low income and low expenditure.

4. Region 2 (Mean Expenditure: 83.91) could be described as the best balanced between income and expenditure

5. Region 3 (Mean Expenditure: 69.18) is spread the widest but most values are concentrated in the lower left corner.

6. Region 4 (Mean Expenditure: 88.30) has the highest expenditure. This can be seen on the graph as most income is centered between 1700 and 2200. In comparison with other regions, region 4 has a big spread in the expenditures compared to other regions with similar income (which are not as far spread)

Please find the R code below:

```
pdf("plot4.pdf")
ggplot(data = expenditure) +
  geom_point(mapping = aes(y = Y, x = X1)) +
  ggtitle("Y:Expenditure vs. X1:Income per Region")
dev.off()

#Solution 2.3.1
#As mentioned before Y and X1 have a moderate correlation with some
    outliers. Most values lie between Y: 60 to 90 and X1: 1500 - 2000.

pdf("plot5.pdf")
ggplot(data = expenditure) +
  geom_point(mapping = aes(y = Y, x = X1, colour = as.factor(Region),
    shape = as.factor(Region))) +#adding the region with different shapes
     and colors for better distinction.
  labs(colour="Region", shape = "Region") +
  ggtitle("Y:Expenditure vs. X1:Income per Region")
dev.off()
```