Problem Set 1

Applied Stats/Quant Methods 1
Stefan Keel

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 0.1: The average IQ of the students in her school is 98.44. This was calculated with following line of code in R:

```
mean(y)
```

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

   Solution 1.1: 90% of the true mean of the schools student IQ falls between 93.95993 and 102.92007.

As N ¡ 30, I opted for a t.distribution and therefore the qt function to find the standard error and subsequently the confidence interval. In the R code I calculated the length, standard deviation and sum of error, and standard error by hand. The fast way to find the confidence interval can be found below.

```
st_error <- qt(0.950, df = length(y) - 1) * (sd(y) / sqrt(length(y)))
CI_lower <- mean(y) - st_error
CI_higher <- mean(y) + st_error
print(c(CI_lower, mean(y), CI_higher))
93.95993  98.44000 102.92007
```

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 $\alpha = 0.05$.

   Solution 1.2: P-Value is 0.7215383 and therefore greater than $\alpha = 0.05$. As the P-Value is bigger than 0.05 we accept the Null-Hypotheses.

   Assumptions:

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

   My hypothesis: whether school IQ is higher than 100. Therefore it is a one-sided test.

   Stating the Hypothesis: The hypothesis states that the average IQ of school students is higher than the average country IQ.
   H0 mu is smaller equal than 100
   HA mu is greater than 100
   Following code was used to calculate P-Value and t.test conducted for proof of P-Value:

```
t_stat <- (mean(y)-100)/(sd(y)/sqrt(length(y)))
P_value <- pt(t_stat, df = length(y)-1, lower.tail = FALSE)
print(P_value)
P_Value = 0.7215383

t_test <- t.test(y, mu = 100, alternative = 'greater')
print(t_test)
data:  yt = -0.59574, df = 24, p-value = 0.7215
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:  93.95993      Inf
sample estimates: mean of x     98.44
```

Question 2 (40 points): Political Economy

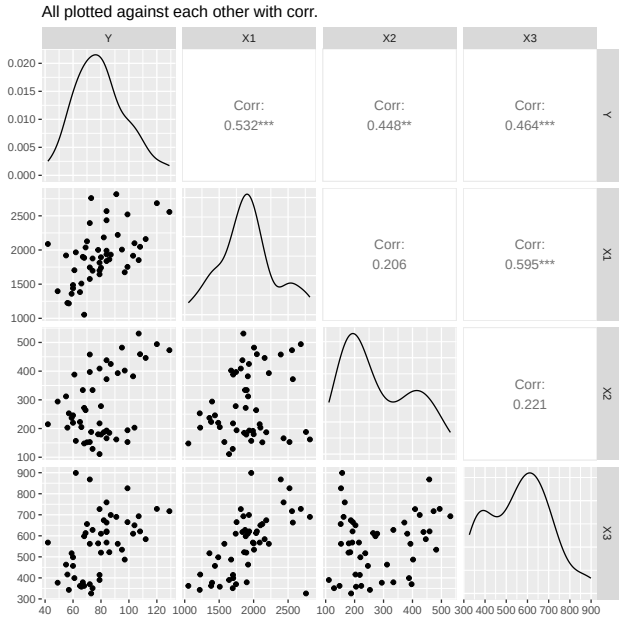
Researchers are curious about what affects the amount of money communities spend on addressing homelessness. The following variables constitute our data set about social welfare expenditures in the USA.

`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`.

```
sum_errors_sqrd #control
```

- 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)?

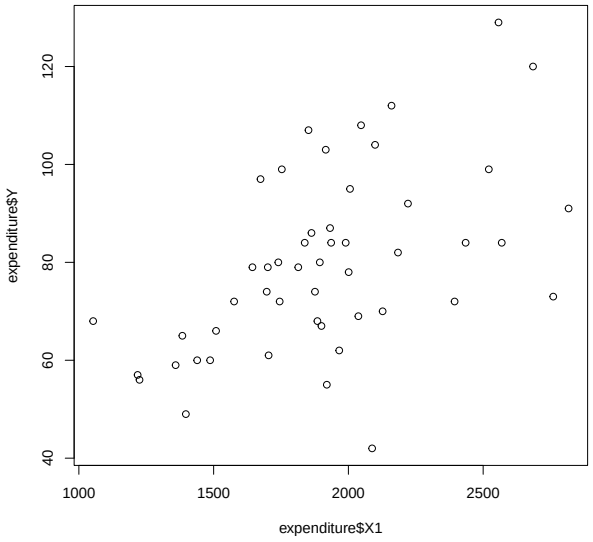


Solution 2.1: From visual interpretation:

- 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.

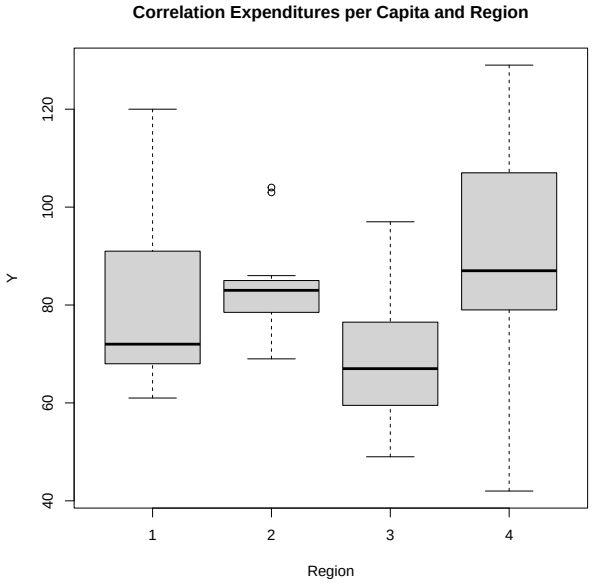
  It would also be possible to plot each combination seperately but it was too tiresome. Therefore I searched the internet (URL in R file) for a code/package that combined it in one plot.
  Exmpaple of single plotting:



  The code for the 4 by 4 plot:

```
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?



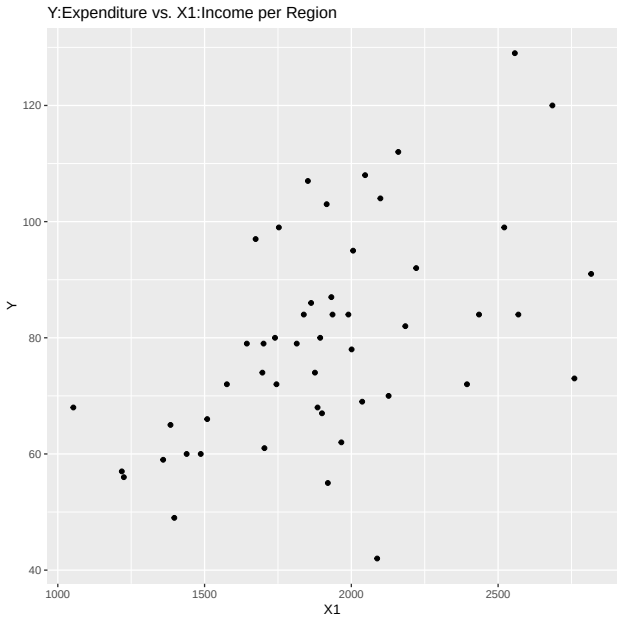
Solution 2.2: Looking at the boxplot that also indicates the mean, we can see that Region 4 has the highest per capita expenditure on housing assistance. After calculating the means of each Region it can be stated that **on average Region 4 has with 88.30769 the highest per capita expenditure on housing assistance.**
Used code in R:

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

for(i in 1:4)
{
nam <- paste("Region", i, sep = "")  assign(nam,
expenditure[expenditure$Region == i,])
}
mean(Region1$Y)
[1] 79.44444
mean(Region2$Y)
[1] 83.91667
mean(Region3$Y)
[1] 69.1875
mean(Region4$Y)
[1] 88.30769
```

- 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

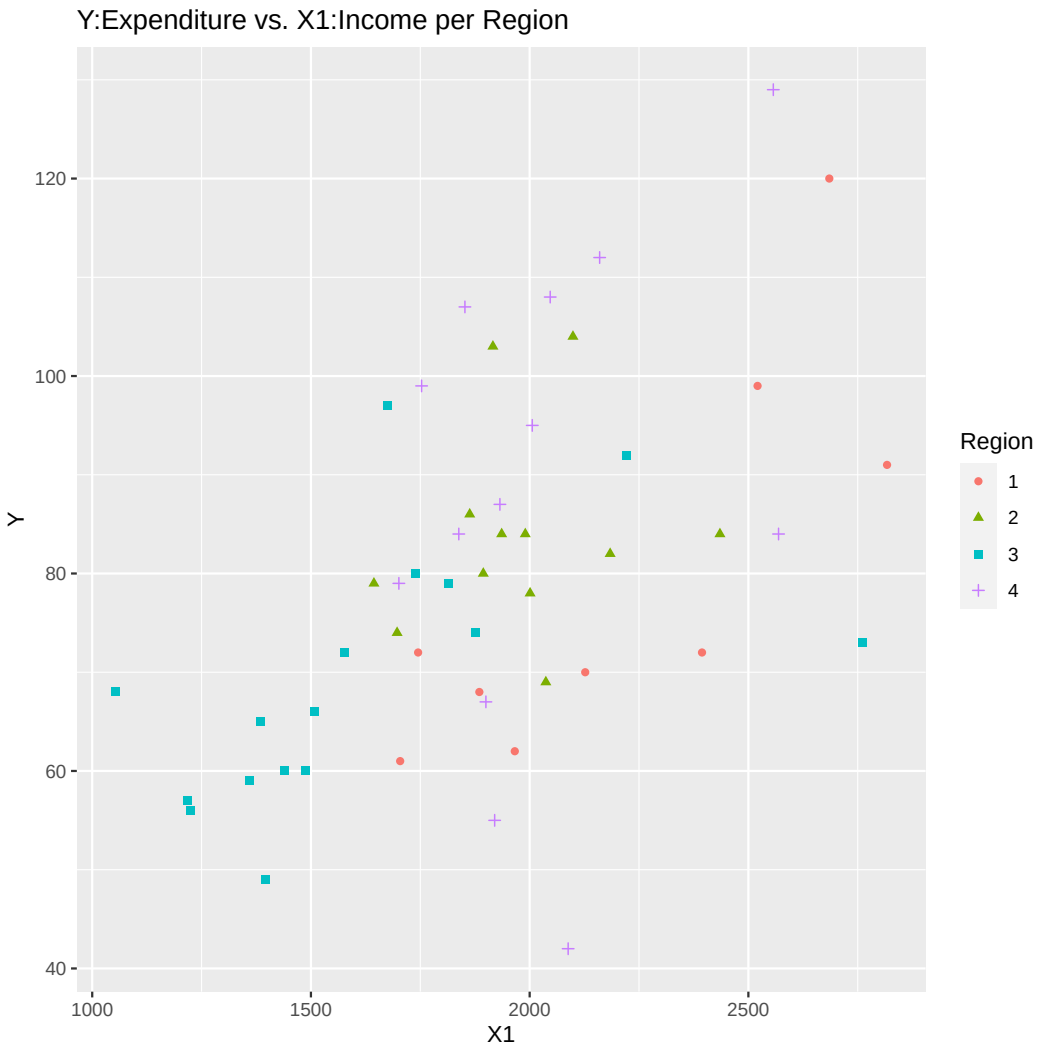
different regions with different types of symbols and colors.



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

This code was used to visualise the correltaion Y and X1:

```
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 2.3.2

- Region 1 (Mean Expenditure: 79.44) is spread the furthest although most of the points are with low income and low expenditure.
- Region 2 (Mean Expenditure: 83.91) could be described as the best balanced between income and expenditure
- Region 3 (Mean Expenditure: 69.18) is spread the widest but most values are concentrated in the lower left corner.
- 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)

This code was used to enhance the previous code with colours and shapes for a better distinction.

```
pdf("plot5.pdf")
ggplot(data = expenditure) +
geom_point(mapping = aes(y = Y, x = X1, colour = as.factor(Region),
shape = as.factor(Region))) +
labs(colour="Region", shape = "Region") +
ggtitle("Y:Expenditure vs. X1:Income per Region")
dev.off()
```