Problem Set 1

Applied Stats/Quant Methods 1

Due: September 30, 2024

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 Monday September 30, 2024. No late assignments will be accepted.

Question 1: 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)
```

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

```
t <- qt(0.05, n-1, lower.tail = F)
# Step 2: Calculate lower and upper parts for the 90%
lower_CI <- mean(y)-(t*(sd(y)/sqrt(n)))
upper_CI <- mean(y)+(t*(sd(y)/sqrt(n)))
# print CIs with mean
```

```r
c(lower_CI, mean(y), upper_CI) #Confidence interval (93.95993 102.92007) mean value(98.44000)
# double check our answer
t.test(y, conf.level = 0.9)$"conf.int" #Use the t.test() function to directly calculate the 90% confidence interval and extract the confidence interval
```

```
Confidence interval (93.95993 102.92007) mean value(98.44000)
```

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 $\alpha = 0.05$.

```r
# Calculate the standard error
SE <- sd(y)/sqrt(n)
# Calculate the test statistic for this hypothesis testing of mean
t <- (mean(y) - 100)/SE
# Get the p-value from t-distribution
pvalue <- pt(t, n-1, lower.tail = F)
# Or another way to do this hypothesis testing is to use the function t.test directly
t.test(y, mu = 100, conf.level = 0.95, alternative = "greater")
#                One Sample t-test
#data:  y
#t = -0.59574, df = 24, p-value = 0.7215
#(The t-value is close to 0, indicating that there is not much difference between the sample mean and the assumed mean (100))
#(The p-value is much greater than 0.05, which means there is not enough evidence to reject the null hypothesis, i.e. there is no evidence to suggest that the sample mean is significantly greater than 100)
#alternative hypothesis: true mean is greater than 100
#(Indicating the hypothesis that the sample mean is greater than 100)
#95 percent confidence interval:
#  93.95993       Inf
#(The lower limit of the confidence interval is 93.95993.The upper limit of the confidence interval is infinite)
#sample estimates:
#mean of x
#   98.44
#(The sample mean is 98.44)
```

Question 2: Political Economy

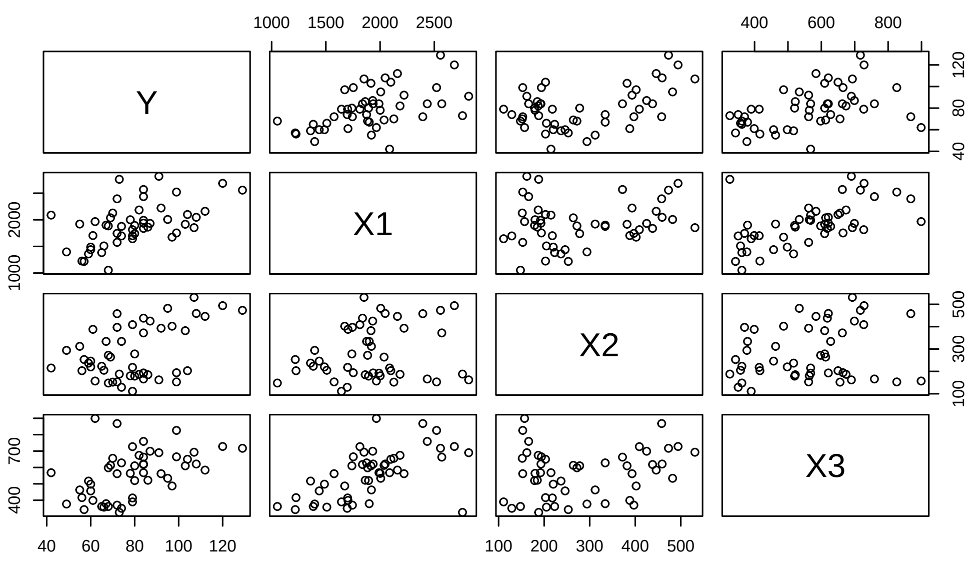
Researchers are curious about what affects the amount of money communities spend on addressing homelessness. The following variables constitute our data set about social welfare expenditures in the USA.

`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)?

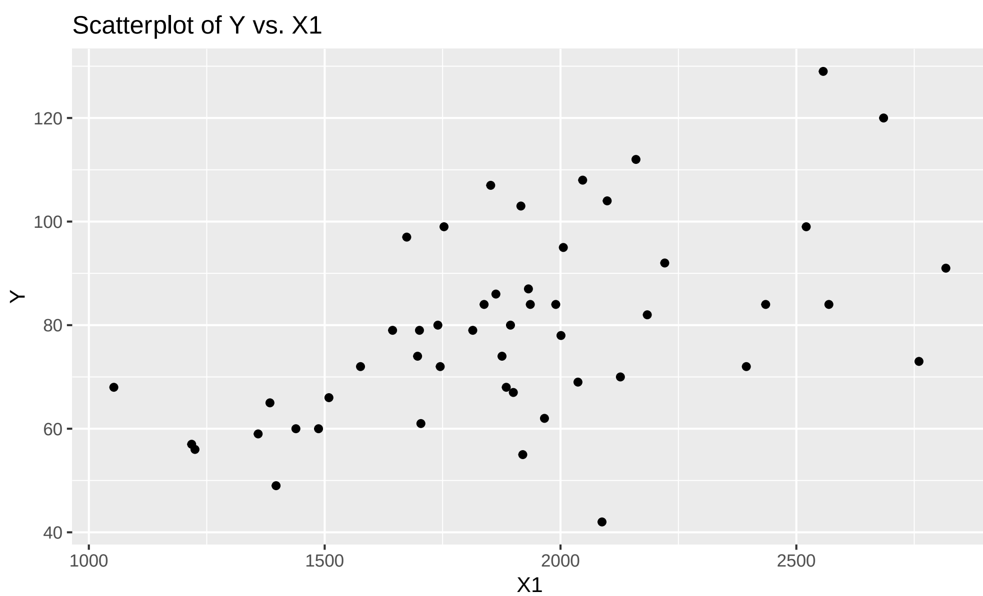
```r
pairs(expenditure[, c("Y", "X1", "X2", "X3")]) #Draw a scatter plot of Y
    with X1, X2, X3
```




```r
summary(expenditure) #Output the statistical results as a text file
```

```
   STATE                 Y                X1             X2              X3
 Length:50          Min.   : 42.00   Min.   :1053   Min.   :111.0   Min.   :326.0
 Class :character   1st Qu.: 67.25   1st Qu.:1698   1st Qu.:187.2   1st Qu.:426.2
 Mode  :character   Median : 79.00   Median :1897   Median :241.5   Median :568.0
                    Mean   : 79.54   Mean   :1912   Mean   :281.8   Mean   :561.7
                    3rd Qu.: 90.00   3rd Qu.:2096   3rd Qu.:391.8   3rd Qu.:661.2
                    Max.   :129.00   Max.   :2817   Max.   :531.0   Max.   :899.0
```

```r
pdf("plot.Y.X1_RJ.C.pdf")
plot(expenditure$X1,expenditure$Y)
dev.off()  #Complete the first question(Y/X1)
```




```
Y is positively correlated with x1,
 indicating that as personal income increases,
 per capita housing expenditure also increases
```

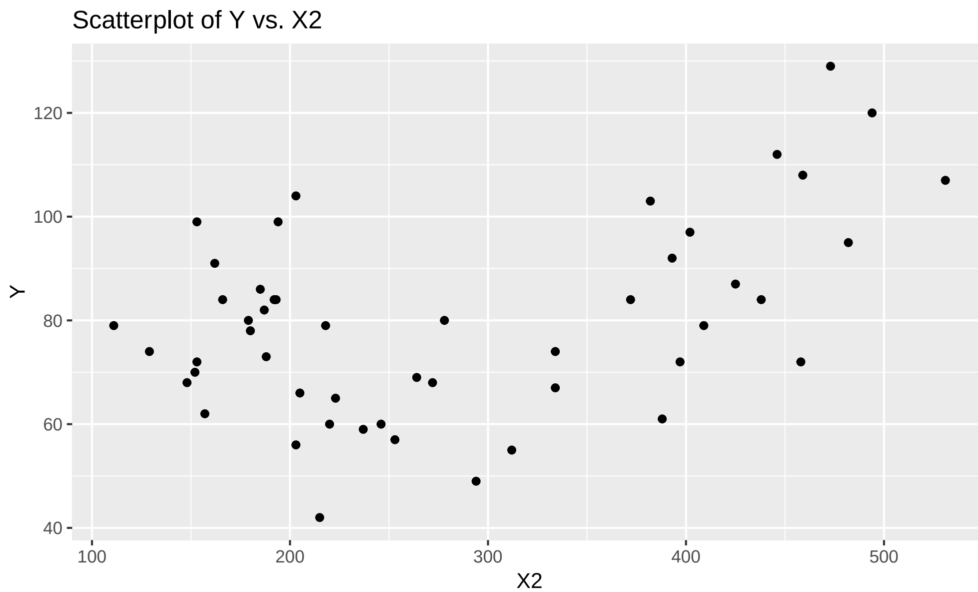
```r
output_stargazer("regression_output_RJ.C.tex", regression_model) #This
    will write the output of stargazer to the 'regression_output_RJ.C.tex'
     file #Complete the second question(Y/X1)
```

Table 1:

	Dependent variable:
	Y
X1	0.025^{***}
	(0.006)
Constant	32.546^{***}
	(11.034)
Observations	50
R^2	0.283
Adjusted R^2	0.268
Residual Std. Error	15.836 (df = 48)
F Statistic	18.920^{***} (df = 1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
The coefficient of X1 is 0.025 and has statistical significance (p-value<0.01),
indicating a positive correlation between X1 and Y.
Specifically, for every unit increase in X1, Y is expected to increase by 0.025 un
The constant term is 32.546 and has statistical significance (p-value<0.01).
This means that when X1 is zero, the expected value of Y is 32.546;
The adjusted R-squared is 0.268, slightly lower than R-squared,
indicating that there are other factors affecting housing expenditure;
as per capita income (X1) increases, housing expenditure (Y) will also increase
```

```
pdf("plot.Y.X2_RJ.C.pdf")
plot(expenditure$X2,expenditure$Y)
dev.off()  #Complete the first question(Y/X2)
```



There is a positive correlation between y and x2,
indicating that in continents with more economically unstable residents,
per capita housing expenditures will increase

```
output_stargazer("regression_output2_RJ.C.tex", regression_model) #
    Complete the second question(Y/X2)
```

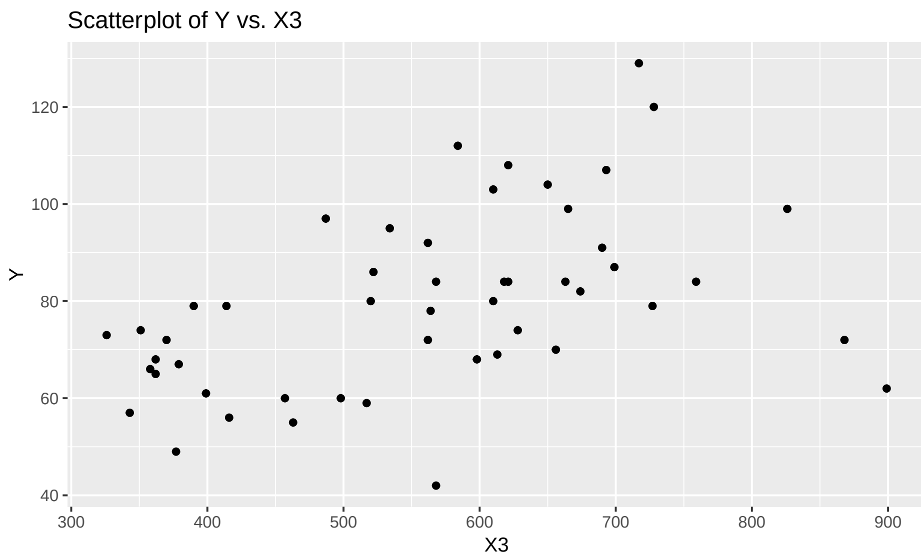
Table 2:

	Dependent variable:
	Y
X2	0.070***
	(0.020)
Constant	57.761***
	(6.164)
Observations	50
R^2	0.201
Adjusted R^2	0.184
Residual Std. Error	16.714 (df = 48)
F Statistic	12.072*** (df = 1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

The coefficient of X2 is 0.070 and has statistical significance (p-value<0.01),

indicating a positive correlation between X2 and Y;
The R-squared value is 0.201, indicating that X2 can explain 20.1% of Y variability
and suggesting that there are other factors affecting Y

```
pdf("plot.Y.X3_RJ.C.pdf")
plot(expenditure$X3,expenditure$Y)
dev.off()  #Complete the first question(Y/X3)
```



There is a positive correlation between y and x3,
indicating that in continents with higher urbanization,
per capita housing expenditures will also increase

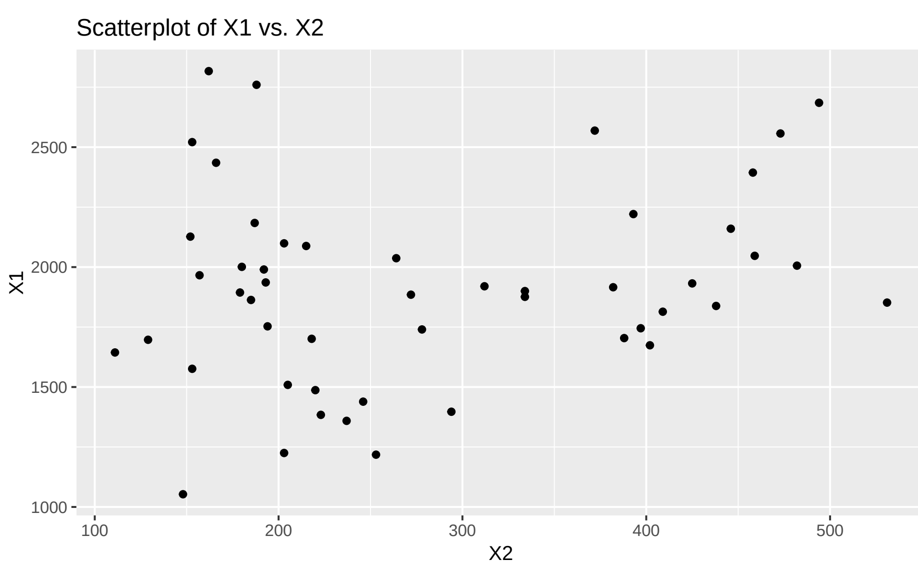
```
output_stargazer("regression_output3_RJ.C.tex", regression_model) #
   Complete the second question(Y/X3)
```

The coefficient of X3 is 0.059 and has statistical significance (p-value<0.01),
indicating a positive correlation between X3 and Y;
The R-squared value is 0.215, indicating that X3 can explain 21.5% of Y variability
This is a relatively low value,
indicating that there are other factors affecting Y

Table 3:

	Dependent variable:
	Y
X3	0.059***
	(0.016)
Constant	43.306***
	(9.461)
Observations	50
R^2	0.215
Adjusted R^2	0.199
Residual Std. Error	16.567 (df = 48)
F Statistic	13.1146*** (df = 1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
pdf("plot.X1.X2_RJ.C.pdf")
plot(expenditure$X2,expenditure$X1)
dev.off()#Complete the first question(X1/X2)
```




```
There is a negative correlation between x1 and x2.
Continents with high per capita income
fewer residents with unstable economies
```

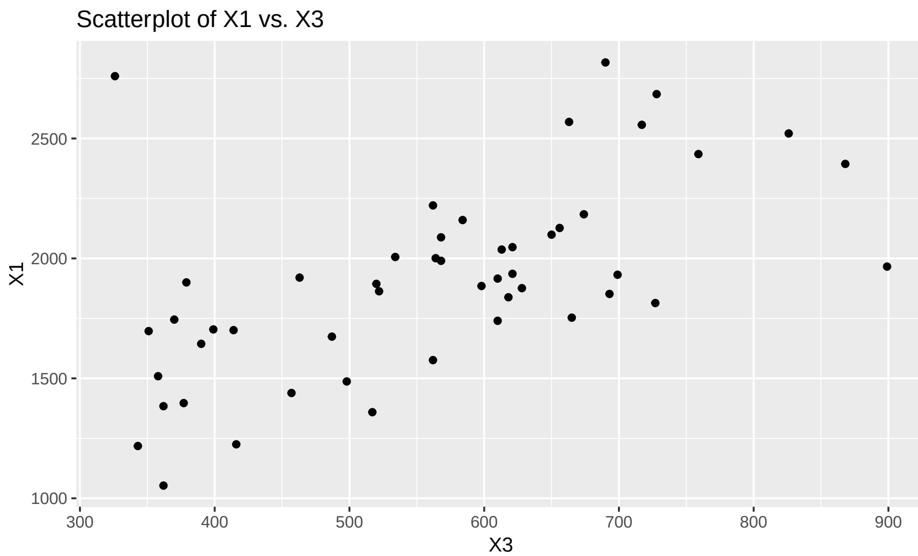
```
output_stargazer("regression_output4_RJ.C.tex", regression_model) #
   Complete the second question(X1/X2)
```

Table 4:

	Dependent variable:
	X1
X2	0.696***
	(0.478)
Constant	1715.655***
	(145.981)
Observations	50
R^2	0.042
Adjusted R^2	0.022
Residual Std. Error	395.854 (df = 48)
F Statistic	2.119*** (df = 1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
The coefficient of X2 is 0.696,
but this coefficient is not statistically significant (p-value>0.1),
which means there is not enough evidence to
conclude a significant linear relationship between X2 and X1;
The R-squared value is 0.042, indicating that X2 can explain 4.2% of X1's variabili
This is a very low value, indicating a very weak relationship between X2 and X1
```

```
pdf("plot.X1.X3_RJ.C.pdf")
plot(expenditure$X3,expenditure$X1)
dev.off()#Complete the first question(X1/X3)
```




```
There is a positive correlation between x1 and x3,
the higher the urbanization of the continent,
the higher the per capita income
```

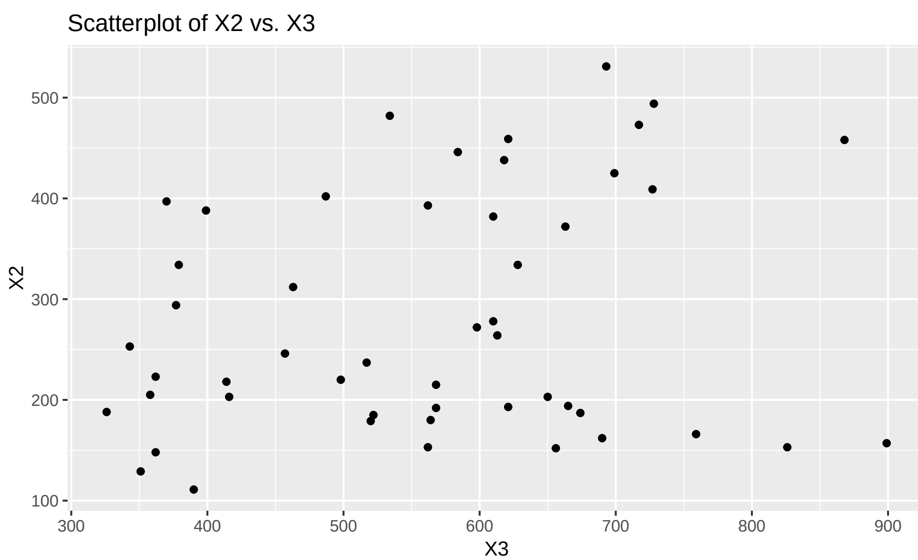
```
output_stargazer("regression_output5_RJ.C.tex", regression_model) #
   Complete the second question(X1/X3)
```

```
The coefficient of X3 is 1.643 and has statistical significance (p-value<0.01),
indicating a positive correlation between X3 and X1;
The R-squared value is 0.354,
indicating that X3 can explain 35.4% of X1's variability.
This is a moderate value,
indicating that X3 has a certain explanatory power for X1
```

Table 5:

	Dependent variable:
	X1
X3	1.643***
	(0.320)
Constant	988.947***
	(185.614)
Observations	50
R^2	0.354
Adjusted R^2	0.341
Residual Std. Error	325.029 (df = 48)
F Statistic	26.341*** (df = 1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
pdf("plot.X2.X3_RJ.C.pdf")
plot(expenditure$X3,expenditure$X2)
dev.off()#Complete the first question(X2/X3)
```




```
The correlation between X2 and X3 is weak
the degree of urbanization has little to do with economic instability
```

```
output_stargazer("regression_output6_RJ.C.tex", regression_model) #
   Complete the second question(X2/X3)
```

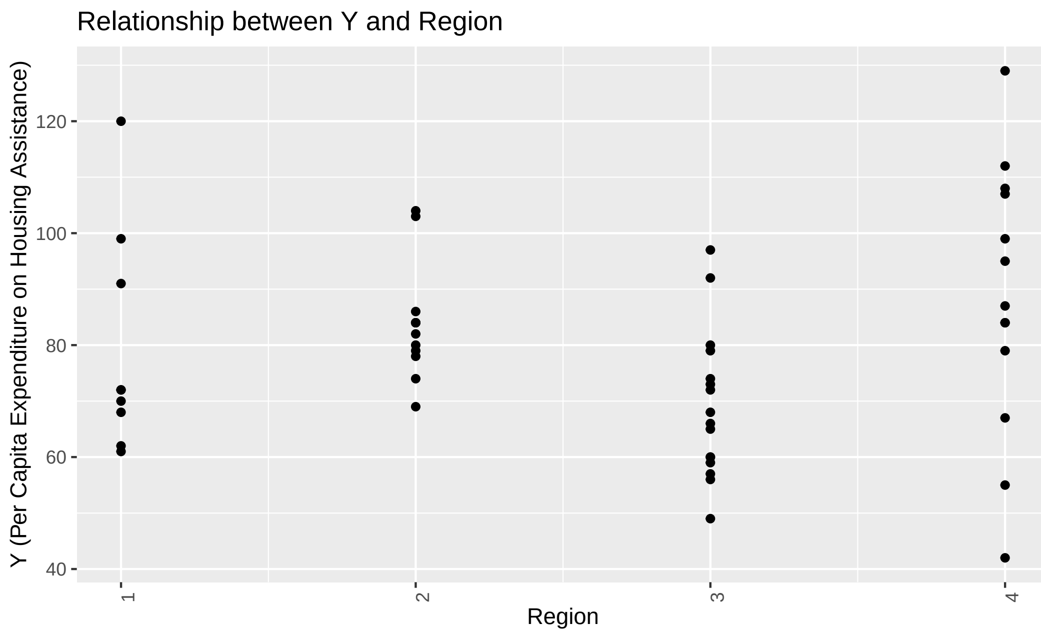
Table 6:

	Dependent variable:
	X2
X3	0.180^{***}
	(0.115)
Constant	180.609^{***}
	(66.509)
Observations	50
R^2	0.049
Adjusted R^2	0.029
Residual Std. Error	116.465 (df = 48)
F Statistic	2.465^{***} (df = 1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
The coefficient of X3 is 0.180,
but this coefficient is not statistically significant (p-value>0.1)
because there is no asterisk mark next to the coefficient,
which cannot prove a significant linear relationship between X3 and X2;
The R-squared value is 0.049,
indicating that X3 can explain 4.9% of X2 variability.
This is a very low value,
indicating a very weak relationship between X3 and X2,
suggesting that there are other important factors affecting X2
```

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

```
pdf("plot.Y.Region_RJ.C.pdf")
plot(expenditure$Region,expenditure$Y)
dev.off() #Complete the first question of the second question
```

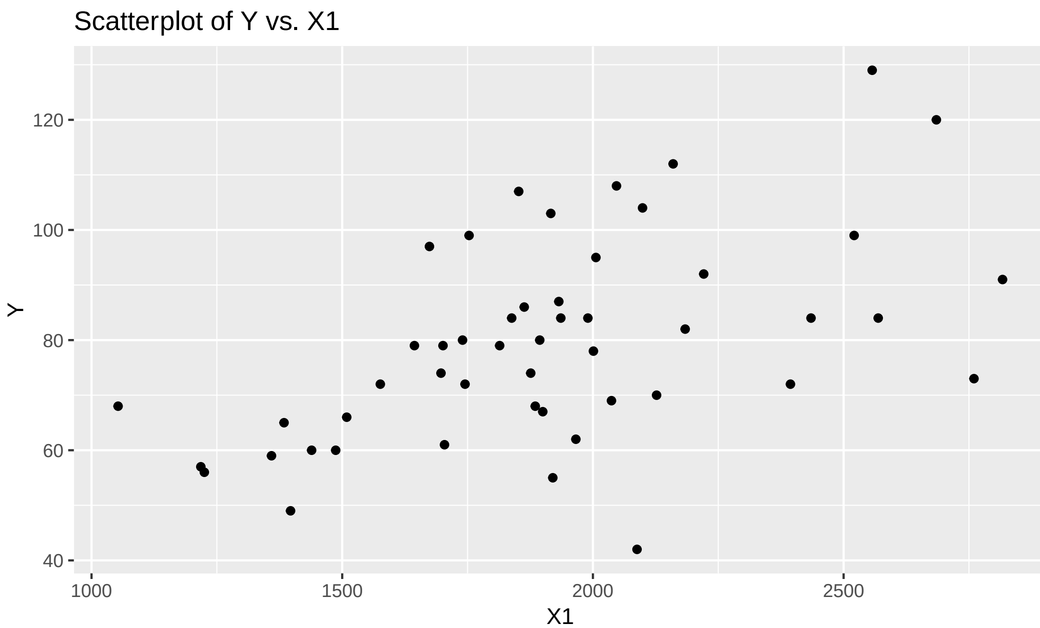



```
average_expenditure <- aggregate(Y ~ Region, data=expenditure, FUN=mean)
highest_region <- average_expenditure[which.max(average_expenditure$Y),]
```

```
Region        Y
  4       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.

```
pdf("plot.Y.X1_RJ.C.pdf")
plot(expenditure$X1,expenditure$Y)
dev.off()  #Complete the first question(Y/X1)
```




```
Y is strongly positively correlated with X1
as per capita income increases
per capita housing expenditure will also increase
```

```
output_stargazer("regression_output_RJ.C.tex", regression_model) #This will write the output of stargazer to the 'regression_output_RJ.C.tex' file #Complete the second question(Y/X1)
```

Table 7:

	Dependent variable:
	Y
X1	0.025***
	(0.006)
Constant	32.546***
	(11.034)
Observations	50
R^2	0.283
Adjusted R^2	0.268
Residual Std. Error	15.836 (df = 48)
F Statistic	18.920*** (df = 1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
pdf("plot.symbols.colors_RJ.C.pdf")
plot(expenditure$X1,expenditure$Y)
dev.off()    #Complete the third question(Y/X1)
```