Applied Stats - Problem Set #1 - Luke Duggan.

Question 1.

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

We can do this using a quick command in R:

```
results <- t.test(y, mu = 100, conf.level=.90)

print(results)
```

Alternatively, we can compute the confidence interval manually as follows:

```
sample_mean <- mean(y)
sample_sd <- sd(y)
t90 <- qt(0.05, 24, lower.tail=FALSE)

lower <- sample_mean - t90*(sample_sd/sqrt(25))
upper <- sample_mean + t90*(sample_sd/sqrt(25))
CI <- c(lower, upper)

print(CI)
```

The result is as follows:

```
> print(CI)
[1]  93.95993 102.92007
```

Consulting the attached R script file, we see that exactly the same result is obtained using either method.

2. *Conduct a hypothesis test that the average IQ is 100 at the 0.05 significance level.*

We can do this using the t.test command or we can construct the test statistic manually:

```
t <- (sample_mean - 100)/(sample_sd/sqrt(25))

print(t)

> print(t)
[1] -0.5957439
```

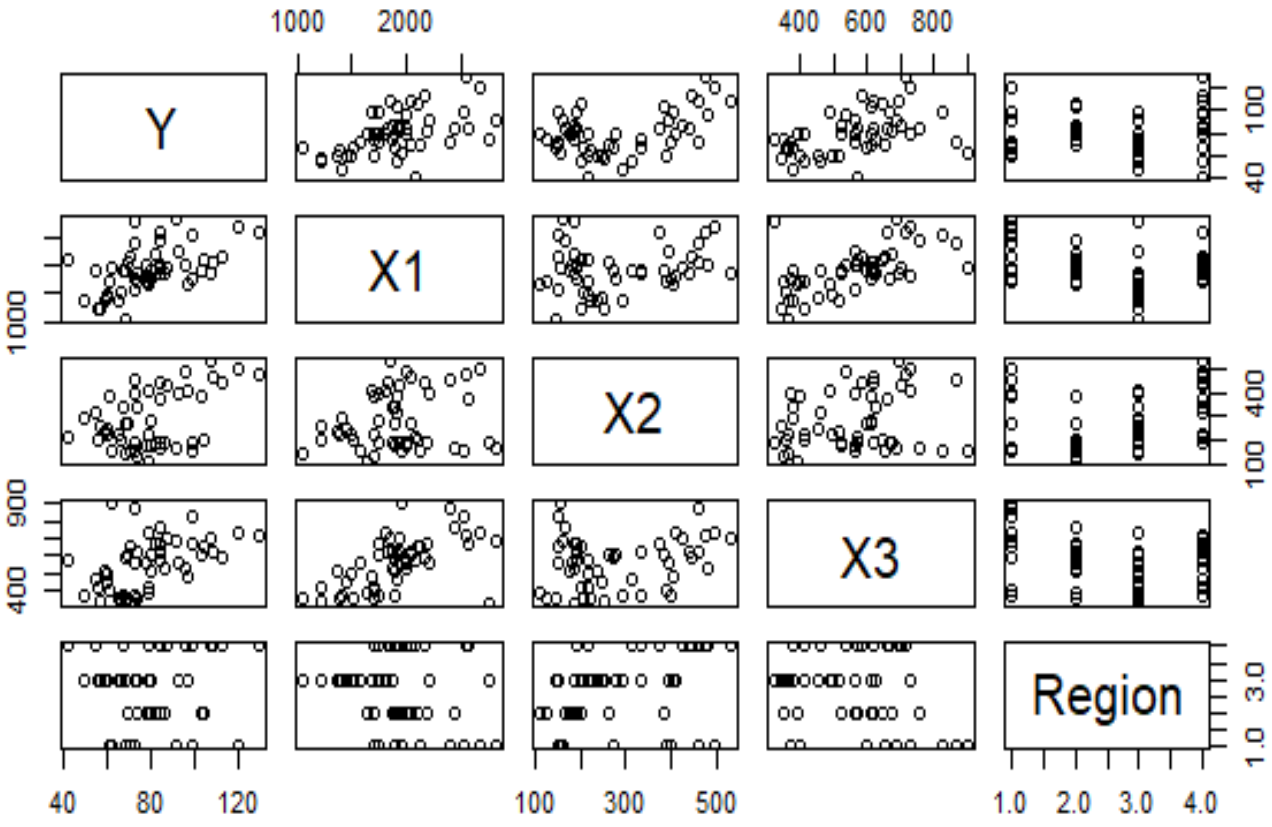
This gives us the value of the test statistic; *under the assumption that the null hypothesis is true*, this statistic follows Student's t-distribution with 23 degrees of freedom. To conduct a hypothesis test of the null hypothesis that the average IQ is 100 against the *two-tailed* alternative hypothesis that the average IQ is not 100, we find the "critical values" at the 0.05 significance level.

```
> print(qt(0.025, 24))
[1] -2.063899
> print(qt(0.975,24))
[1] 2.063899
```

(Note that the t-distribution is symmetric, so each critical value is simply the other times negative one). Thus, the null hypothesis is rejected if the test statistic lies outside the region [-2.063899, 2.063899]; that is, lies in the “rejection region”. Since the test statistic does not lie in the rejection region, we cannot reject the null hypothesis that the average IQ is 100 (this is, of course, not the same thing as accepting the hypothesis).

Question 2.

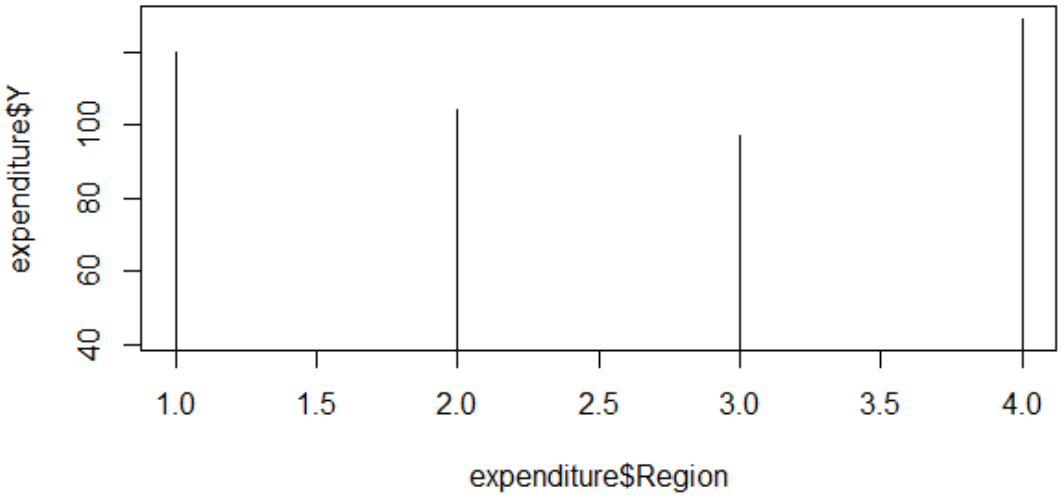
A. Firstly, to generate all possible two-way scatterplots with the variables in the data frame, we use “plot(expenditure)”. This yields the following:



If we were just to eyeball these graphs and render a first impression of the relationship between the variables, we might say the following. Y is moderately positively correlated with X1 and X3 and weakly positively correlated with X2. X1 is apparently uncorrelated with X2, and moderately positively correlated with X3.

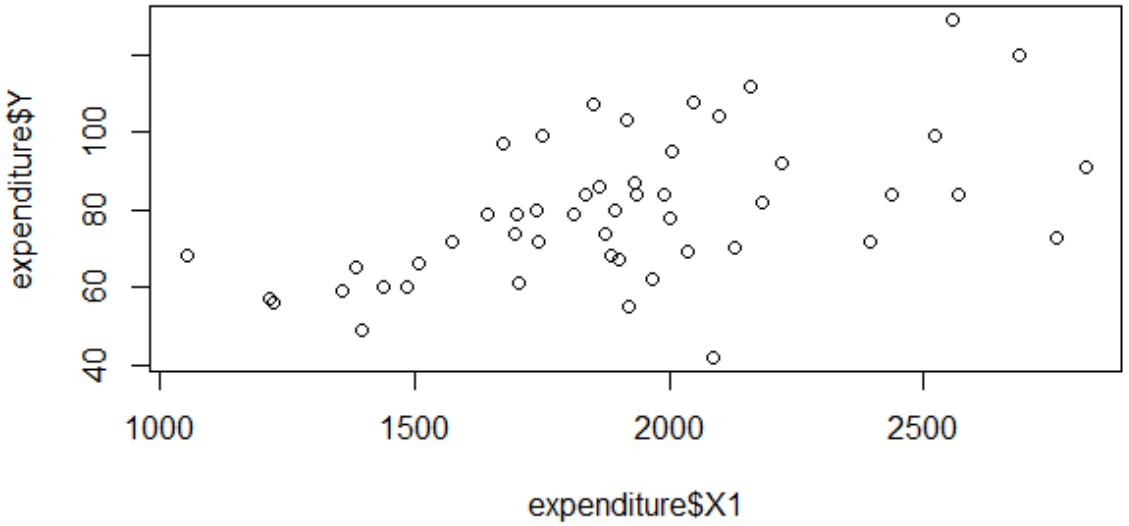
Note two things. Firstly, Region is not really a numerical quantity, the numbers being mere placeholders for geographic regions; so it doesn't really make sense to talk about correlations (although if, e.g., larger numbers meant a region was further West it would make some sense). Secondly, correlation is intuitively (and, for formal measures like the Pearson correlation coefficient, rigorously) a symmetric notion; so having considered whether, e.g., Y and X1 are correlated, we don't have to separately consider whether X1 and Y are correlated.

B. We now plot the relationship between the amount given per capita to the homeless and the region in which a state lies.



We see that states in region 4 (the West) give the most, followed by the Northeast, North, and South in descending order. The heights of the lines gives us a guide to the relative quantity of charitable contributions between regions.

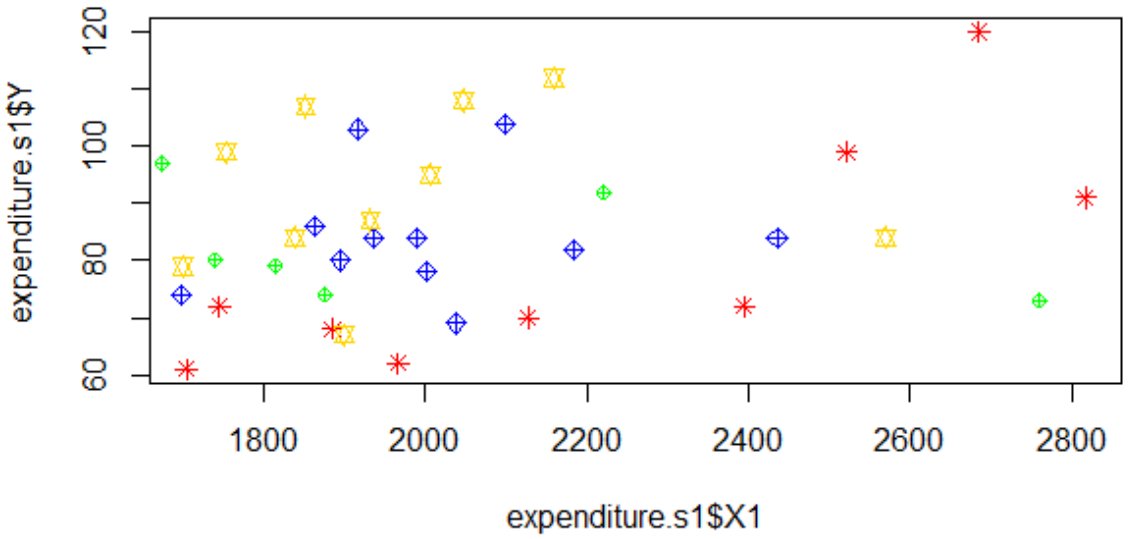
C. We plot the per capita income in a state to per capita expenditures on homelessness relief in that state (it should be borne in mind that the per capita incomes for each state have been invented for this exercise, since the figures given do not correspond to reality).



It is a plausible conjecture that there should be a positive correlation here: very roughly and intuitively, a person with a larger income has, among other things, more money to give to homeless relief. In more detail, a person with a larger income has more than enough to satisfy their own needs and can therefore afford to give money to help other people satisfy theirs, while a person with a small income has less money left over after satisfying their own needs with which to give charitably.

We see something like that in this graph: the states with a per capita income of 1500 or less all give quite a small amount to the homeless, and the states with middling per capita incomes cluster around the centre of the graph. The conjecture is not unambiguously true of higher income states, however; although the two largest givers are high-income states, as a whole these states give a similar amount to middle-income states. Thus, it seems there is a moderate positive correlation (in fact, the Pearson coefficient is a little over 0.5).

We now want to reproduce this plot but with different colours and symbols for each state corresponding to the region in which it lies. The easiest way to do this is to plot the states from each region separately and then overlay the graphs. To do this, we need to “subset” our dataframe, i.e., create dataframes for each region. The specific code is found in the attached R script file. The result of doing this is:



Where the red asterisks represent Northeastern states, the blue kites represent North Central states, the green kites represent Southern states, and the yellow stars represent Western states. It is possible to add a legend to the graph using R's base graphics but as this wasn't requested I did not.