Problem Set 1

Conor Twomey Student No. 22996168 Applied Stats/Quant Methods 1

Due: October 3, 2021

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 in `.pdf` form.
- This problem set is due before 8:00 on Friday October 3, 2021. No late assignments will be accepted.
- Total available points for this homework is 100.

Question 1 (50 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:

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

```r
#load data
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)

#calculate mean
ybar <- mean(y)
```

```r

#visualize y
hist(y, xlab = "IQ scores")

# get confidence intervals
CI_lower <- qnorm(0.05,
                  mean = mean(y),
                  sd = (sd(y)/sqrt(length(y))) # the equation for the standard error of the mean
)

CI_upper <- qnorm(0.95,
                  mean = mean(y),
                  sd = (sd(y)/sqrt(length(y)))
)

matrix(c(CI_lower, CI_upper), ncol = 2,
       dimnames = list("",c("Lower", "Upper")))
```

```
     Lower :  94.13283  Upper : 102.7472
```

```r
#get confidence intervals using t distribution
se <- sd(y)/sqrt(length(y))
```

```
 2.618575
```

```r
t_score <- qt(.05, df = length(y)-1, lower.tail = FALSE)
```

```
1.710882
```

```r
CI_lower_t <- mean(y) - (se * t_score)
```

```
93.95993
```

```r
CI_upper_t <- mean(y) + (se * t_score)
```

```
102.9201
```

1. 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 $\alpha = 0.05$.

```
Hypothesis testing

Step 1 : I assuming that this is a normal distribution

Step 2: The null hypothesis is that ybar is less than or equal to 100
The alternative hypothesis is that ybar is greater than 100

Step 3: Calculating the t-statistic
```

```r
# t.score = ybar - mu/ se
ts <- (ybar - 100)/se
```

```
-0.5957439
```

```r

#check t.score result with R t.test()
t.test(y,
       mu = 100,
       alternative = "greater",
       conf.level = .95)
```

```
One Sample t-testdata:  yt = -0.59574, df = 24, p-value = 0.7215
Alternative hypothesis: true
Mean is greater than 100
95 percent confidence interval: 93.95993
Infsample estimates:mean of x     98.44
```

Question 2 (50 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)?

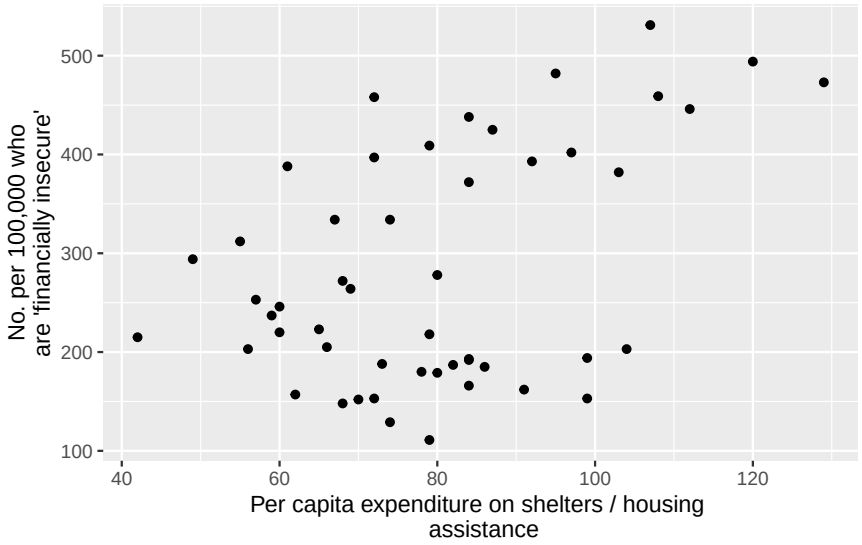

Figure 1: Y and X1 plot

`The relationship between these variables is broadly linear however there are a few outliers with a cluster at the bottom of the plot`

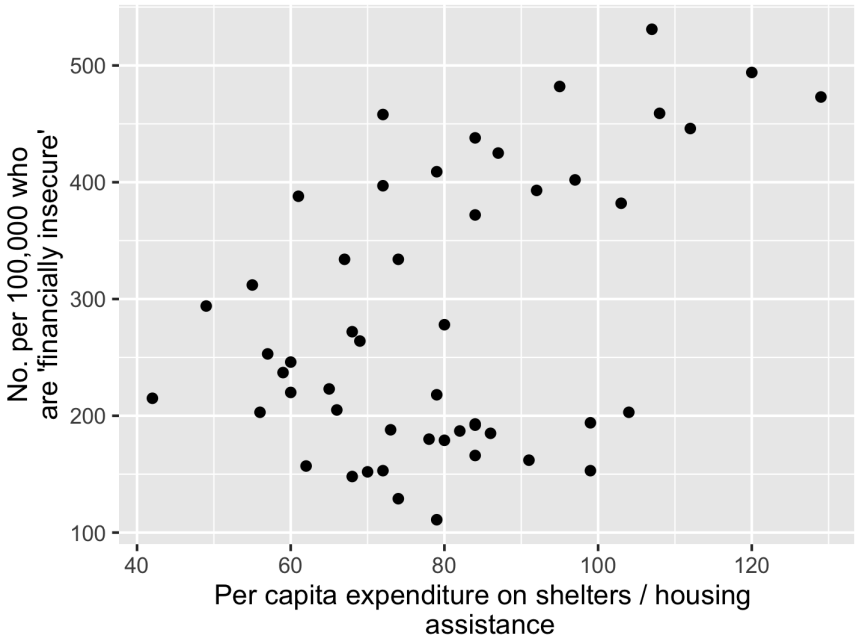

Figure 2: Y and X2 plot

`The relationship between these variables is broadly linear however there are a few outli`

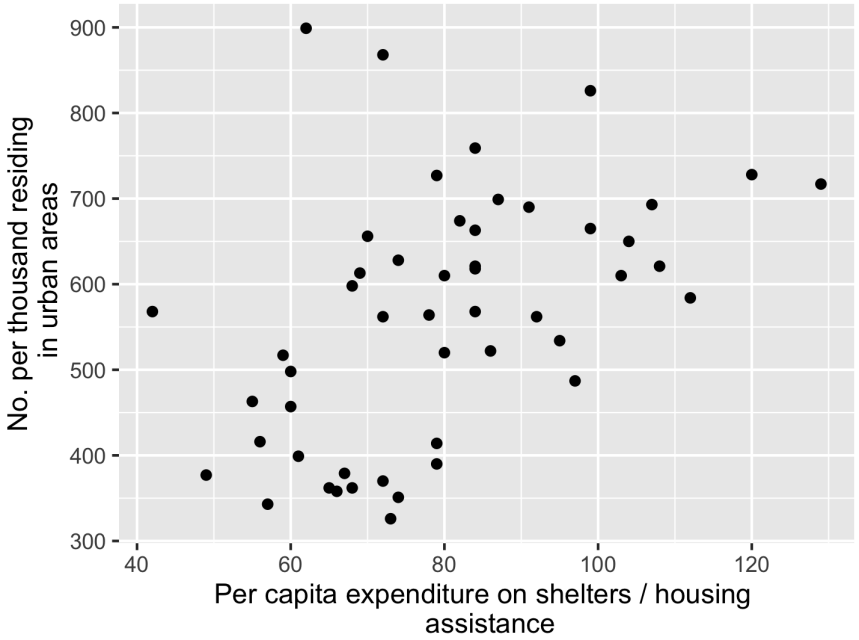

Figure 3:

There is a somewhat linear relationship between these variables

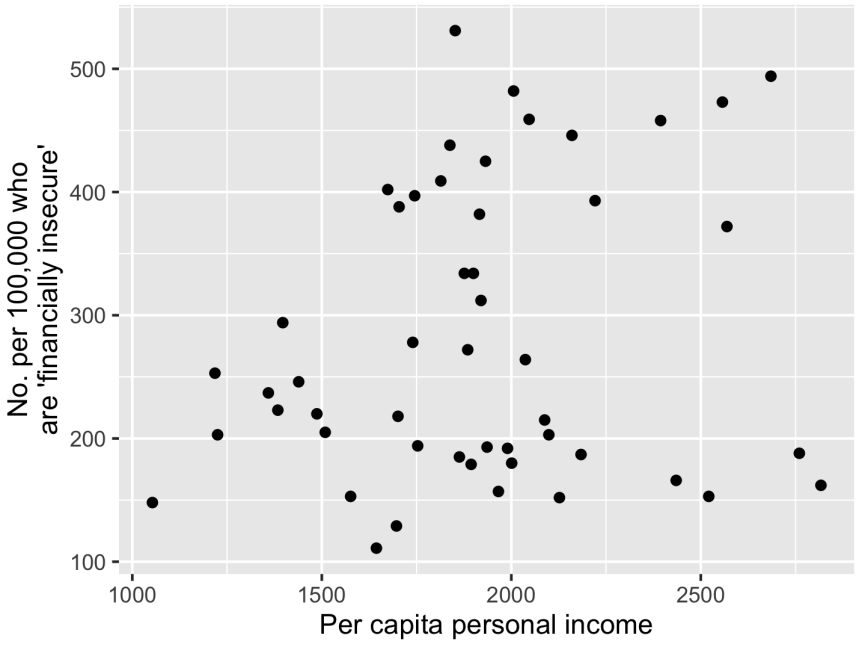

Figure 4: X1 and X2 plot

There is not a linear relationship between these variables

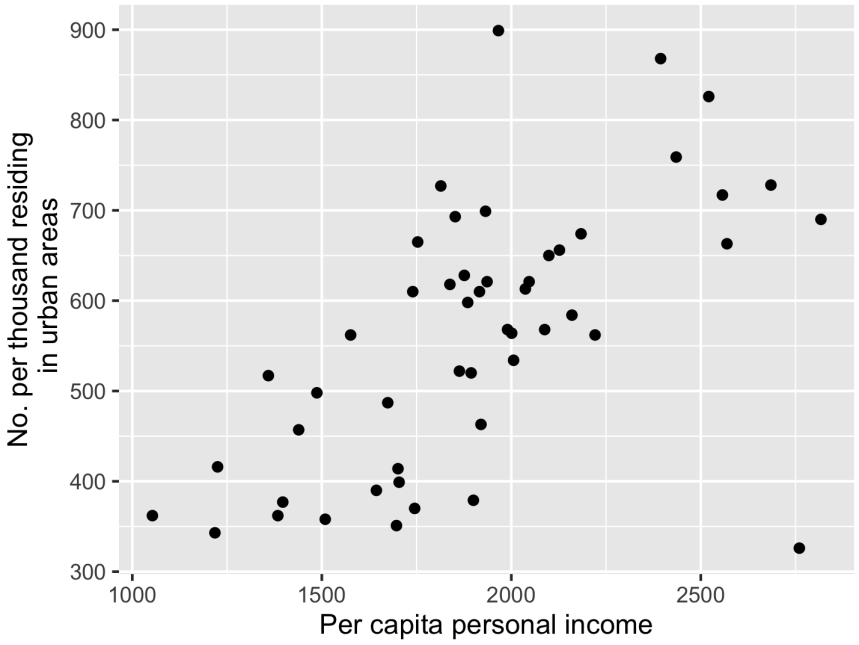

Figure 5:

```
There is a linear relationship between these variables
```

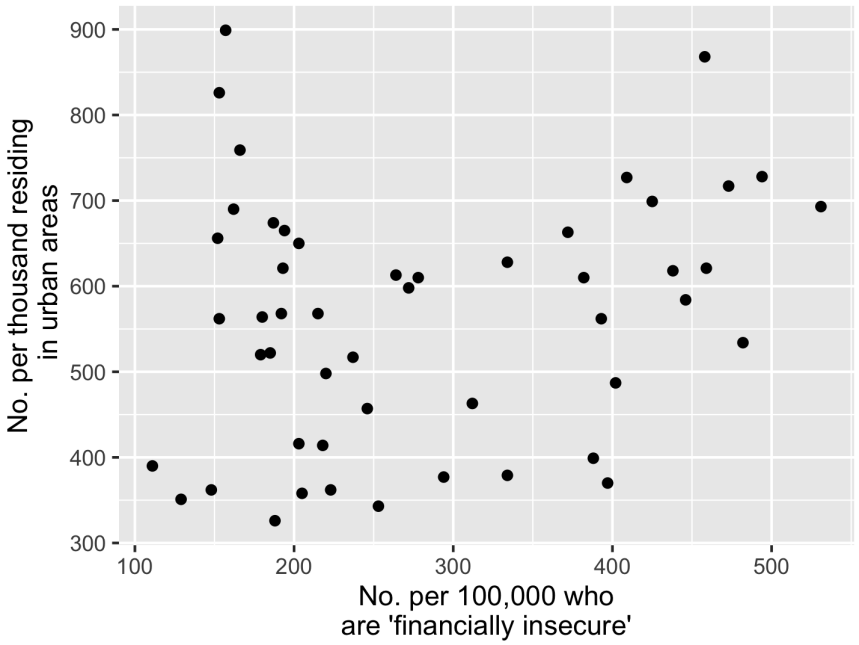

Figure 6:

```
There is no relationship between these variables
```

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

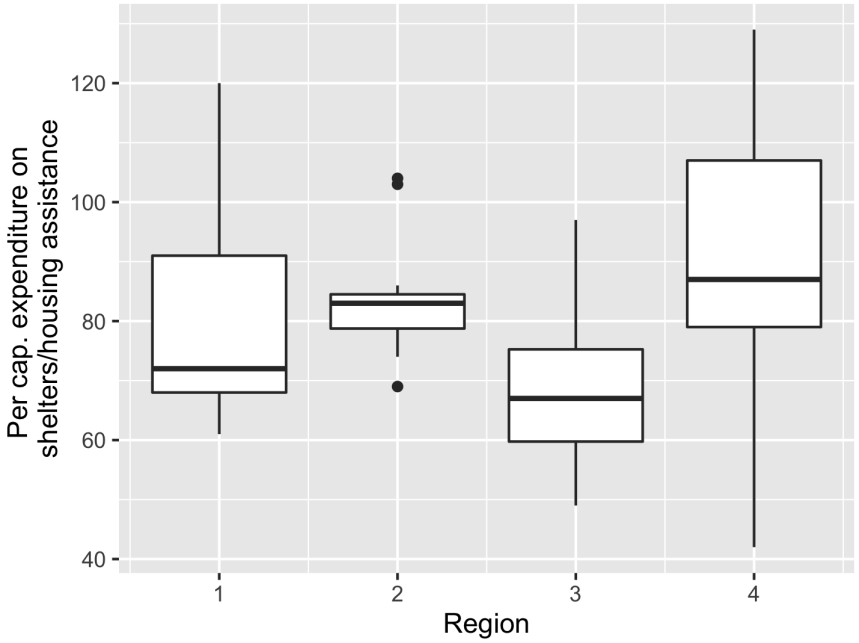

Figure 7:

```
Region 4 (West) spends the most on average per capita on shelters/ housing assistance
```

- 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.

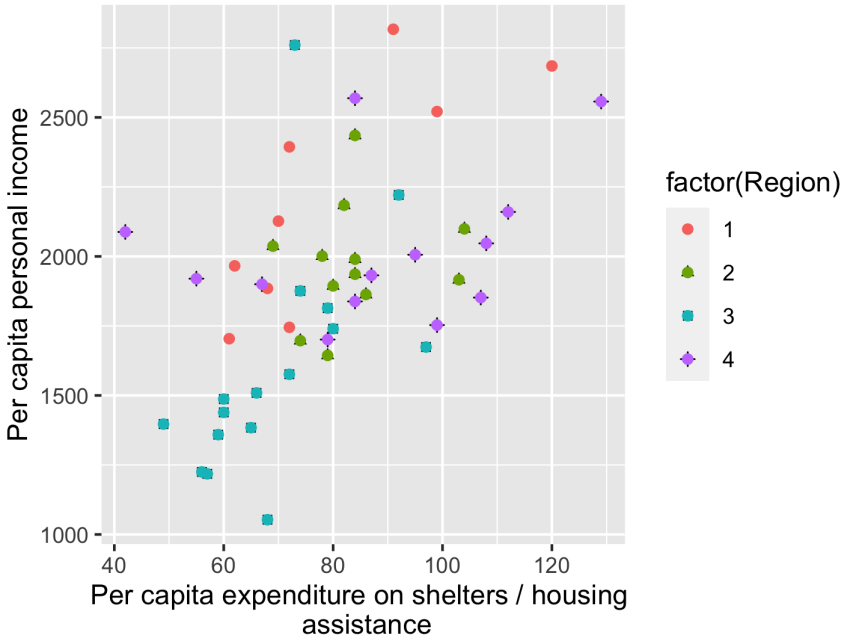

Figure 8: Y and X1 plot with Region variable