IMT 573: Problem Set 4 - Data Analysis

Vighnesh Misal

Due: Tuesday, October 29, 2019

Collaborators: Ashish Anand

Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Server.

1. Download the `problemset4.rmd` file from Canvas or save a copy to your local directory on RStudio Server. Open `problemset4.rmd` in RStudio and supply your solutions to the assignment by editing `problemset4.rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures, and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and we encourage it, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. Remember partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run withouth errors you can do so with the `eval=FALSE` option.
7. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the knitted PDF file to `ps4_ourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

Setup

In this problem set you will need, at minimum, the following R packages.

```r
# Load standard libraries
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.1

## Warning: package 'purrr' was built under R version 3.6.1

## Warning: package 'dplyr' was built under R version 3.6.1

## Warning: package 'stringr' was built under R version 3.6.1
```

```r
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.6.1
```

Problem 1: 50 States in the USA

In this problem we will use the `state` dataset, available as part of the R statistical computing platforms. This data is related to the 50 states of the United States of America. Load the data and use it to answer the following questions.

(a) Describe the data and each variable it contains. Tidy the data, preparing it for a data analysis.

```
dataset1 <- tbl_df(state.x77)
nrow(dataset1)
```

```
## [1] 50
```

```
str(dataset1)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    50 obs. of  8 variables:
##  $ Population: num  3615 365 2212 2110 21198 ...
##  $ Income    : num  3624 6315 4530 3378 5114 ...
##  $ Illiteracy: num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
##  $ Life Exp  : num  69 69.3 70.5 70.7 71.7 ...
##  $ Murder    : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
##  $ HS Grad   : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6 ...
##  $ Frost     : num  20 152 15 65 20 166 139 103 11 60 ...
##  $ Area      : num  50708 566432 113417 51945 156361 ...
```

```
summary(dataset1)
```

```
##    Population        Income       Illiteracy       Life Exp    
##  Min.   :  365   Min.   :3098   Min.   :0.500   Min.   :67.96  
##  1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12  
##  Median : 2838   Median :4519   Median :0.950   Median :70.67  
##  Mean   : 4246   Mean   :4436   Mean   :1.170   Mean   :70.88  
##  3rd Qu.: 4968   3rd Qu.:4814   3rd Qu.:1.575   3rd Qu.:71.89  
##  Max.   :21198   Max.   :6315   Max.   :2.800   Max.   :73.60  
##      Murder          HS Grad          Frost             Area       
##  Min.   : 1.400   Min.   :37.80   Min.   :  0.00   Min.   :  1049  
##  1st Qu.: 4.350   1st Qu.:48.05   1st Qu.: 66.25   1st Qu.: 36985  
##  Median : 6.850   Median :53.25   Median :114.50   Median : 54277  
##  Mean   : 7.378   Mean   :53.11   Mean   :104.46   Mean   : 70736  
##  3rd Qu.:10.675   3rd Qu.:59.15   3rd Qu.:139.75   3rd Qu.: 81163  
##  Max.   :15.100   Max.   :67.30   Max.   :188.00   Max.   :566432  
```

```
?state.x77
```

```
## starting httpd help server ... done
```

The data contains 50 rows and 8 columns in its matrix.

Populaion (num) : Number of people

Income (num): Per capita income

Illeteracy (num) : Percentage of illeterate population

Life Exp (num) : Life expectancy in years.

Murder (num) : Murder rate per 100,000 of the population

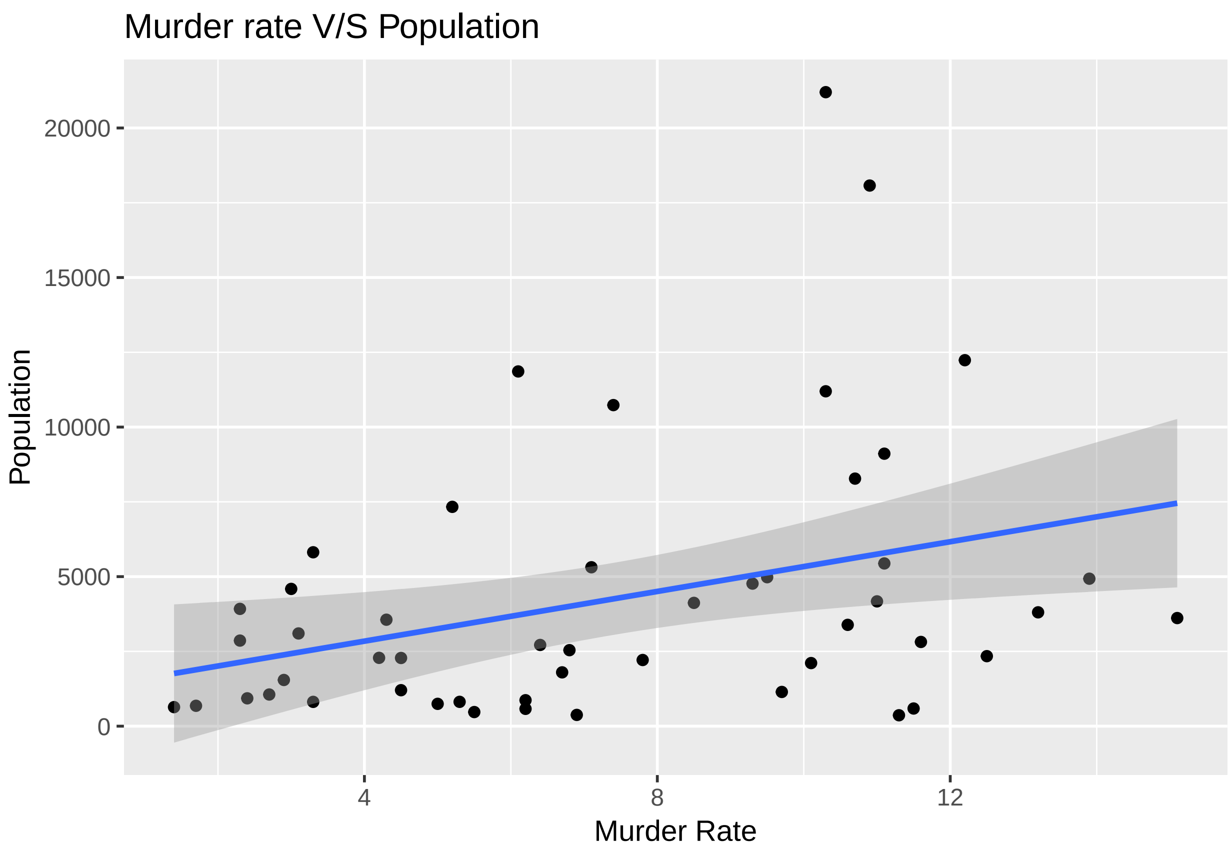
HS Grad (num) : Percentage of high school graduates

Frost (num) : mean number of days when temperature was below the freezing point in a large city/capital.

Area (num) : Land area in square miles.

(b) Suppose you want to explore the relationship between a state's `Murder` rate and other characteristics of the state, for example population, illiteracy rate, and more. Begin by examining the bivariate relationships present in the data. Present and interpret numeric value to describe the linear relationships as well as plots to contextualize these numeric values. What does your analysis suggest might be important varibles to consider in building a model to explain variation in murder rates? Are linear relationships appropriate to assume for all bivariate relationships? Why or why not?

```
dependent_var <- dataset1$Murder
independent_var <- dataset1$Population
model <- lm(dependent_var ~ independent_var)

ggplot(mapping = aes(x = dependent_var, y = independent_var)) + geom_point() + geom_smooth(method = "lm
```



```
model
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Coefficients:
##     (Intercept)  independent_var
##       6.1713934        0.0002841
```

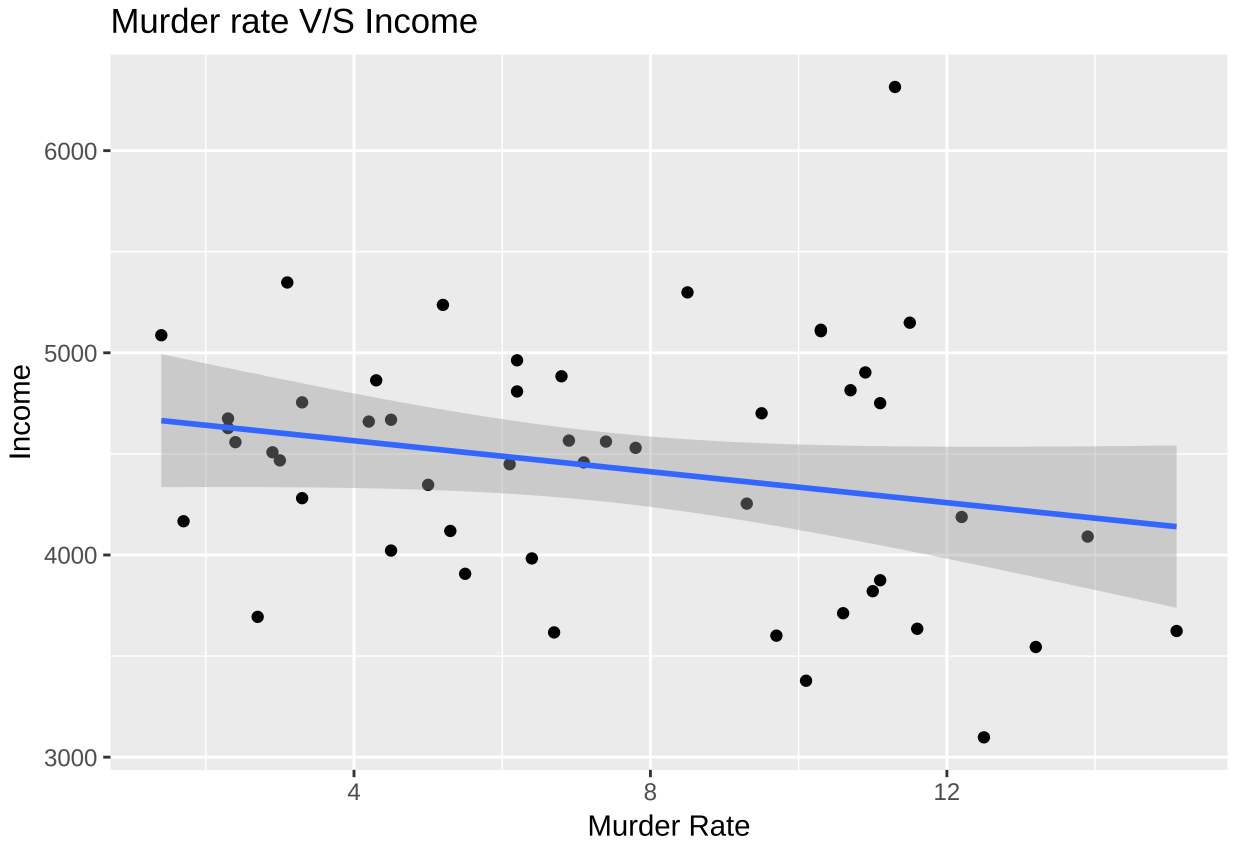
```
summary(model)
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9855 -3.0119 -0.3128  2.4986  7.9014
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.1713934  0.6869410   8.984 7.49e-12 ***
## independent_var 0.0002841  0.0001121   2.535   0.0146 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.503 on 48 degrees of freedom
## Multiple R-squared:  0.1181, Adjusted R-squared:  0.09972
## F-statistic: 6.427 on 1 and 48 DF,  p-value: 0.01455
```

We can observe a linearly increasing relationship between Murder rate and population although the slope has a very low value 0.0002841 which indicates a very small change. The t value is relatively not that large, and although it states that there might be a relationship between the two data points, this relationship is not that strong. Also the p-value of 0.01455 is significantly low.

```
independent_var <- dataset1$Income
model <- lm(dependent_var ~ independent_var)

ggplot(mapping = aes(x = dependent_var, y = independent_var)) + geom_point() + geom_smooth(method = "lm
```



```
model
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Coefficients:
```

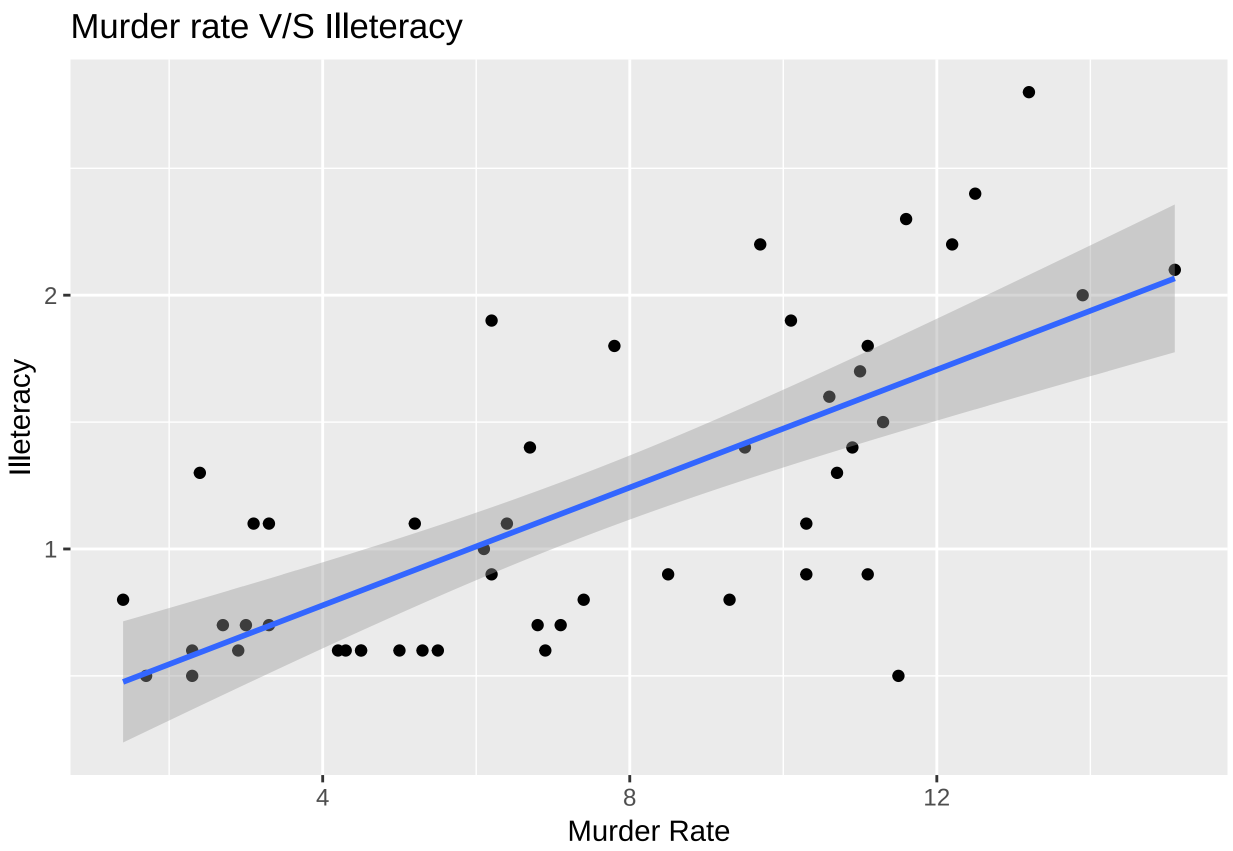
```
##     (Intercept)  independent_var
##       13.509309        -0.001382
```

```r
summary(model)
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0495 -2.8033 -0.2727  3.0730  6.5999
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     13.5093092  3.7782753   3.576  0.00081 ***
## independent_var -0.0013822  0.0008439  -1.638  0.10797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.63 on 48 degrees of freedom
## Multiple R-squared:  0.05294,    Adjusted R-squared:  0.03321
## F-statistic: 2.683 on 1 and 48 DF,  p-value: 0.108
```

We can observe a linearly decreasing relationship between Murder rate and Income although the slope has a very low value -0.001382 which indicates a very small change. The t value of -1.638 is relatively not that large, and although it states that there might be an inverse relationship between the two data points, this relationship is not that strong. Also the p-value of 0.108 not that low and might have us reject the alternative hypothesis.

```r
independent_var <- dataset1$Illiteracy
model <- lm(dependent_var ~ independent_var)

ggplot(mapping = aes(x = dependent_var, y = independent_var)) + geom_point() + geom_smooth(method = "lm
```



```
model
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Coefficients:
##     (Intercept)  independent_var
##           2.397            4.257
```

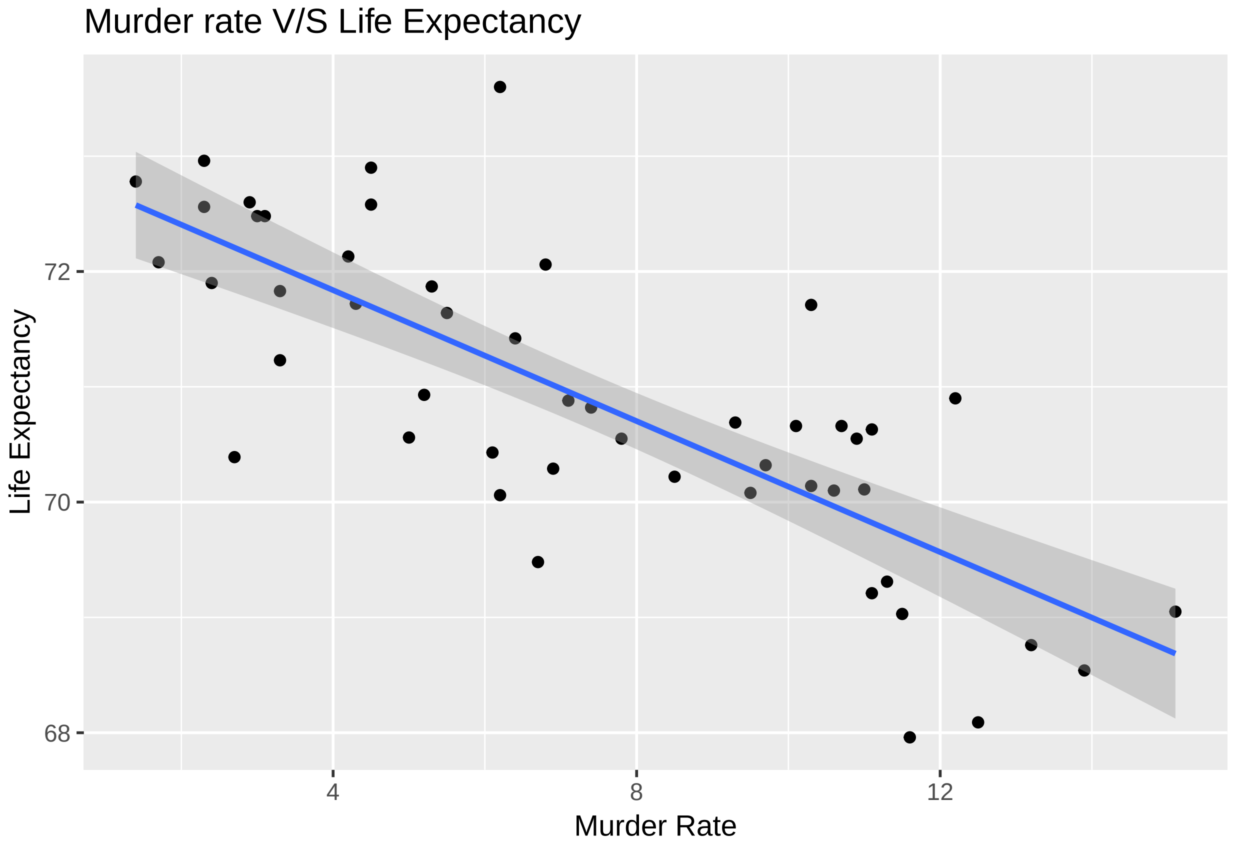
```
summary(model)
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5315 -2.0602 -0.2503  1.6916  6.9745
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.3968     0.8184   2.928   0.0052 **
## independent_var   4.2575     0.6217   6.848 1.26e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.653 on 48 degrees of freedom
## Multiple R-squared:  0.4942, Adjusted R-squared:  0.4836
## F-statistic: 46.89 on 1 and 48 DF,  p-value: 1.258e-08
```

We can observe a linearly increasing relationship between Murder rate and Illiteracy although the slope has a very high which indicates a very small change. The t value of 6.848 is relatively large, and it states that there might be a strong correlation between the two data points. Also the p-value is quite low and of the order 10^-8.

```
independent_var <- dataset1$`Life Exp`
model <- lm(dependent_var ~ independent_var)

ggplot(mapping = aes(x = dependent_var, y = independent_var)) + geom_point() + geom_smooth(method = "lm
```



```
model
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Coefficients:
```

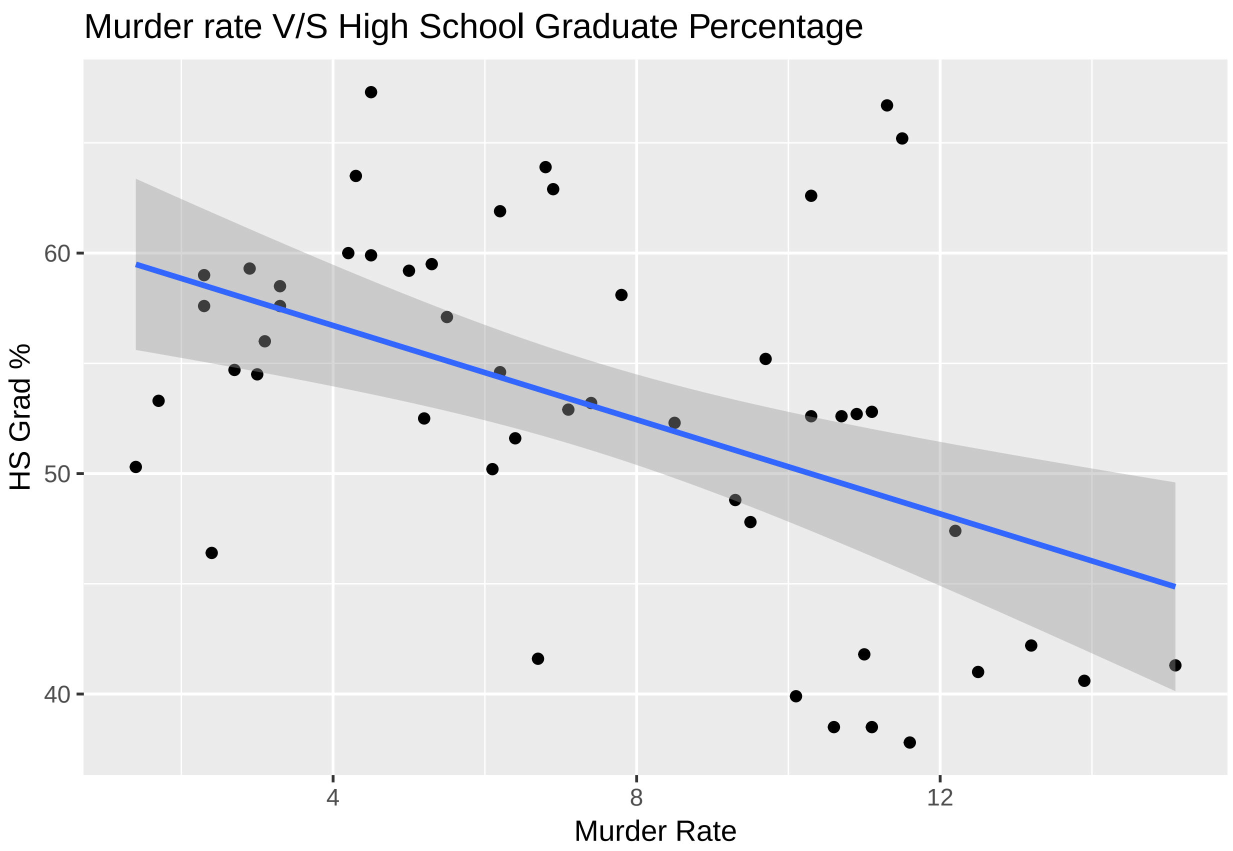
```
##     (Intercept)  independent_var
##         159.576           -2.147
```

```
summary(model)
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7272 -1.6733 -0.1734  1.4909  4.8680
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      159.576     17.579   9.078 5.45e-12 ***
## independent_var   -2.147      0.248  -8.660 2.26e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.33 on 48 degrees of freedom
## Multiple R-squared:  0.6097, Adjusted R-squared:  0.6016
## F-statistic: 74.99 on 1 and 48 DF,  p-value: 2.26e-11
```

We can observe a linearly decreasing relationship between Murder rate and Life Expectancy although the slope has a high value of -2.147 that suggests a large change. The t value of -8.660 is large, it states that there might be a strong inverse relationship between the two data points,. Also the p-value is quite low and of the order 10^-11.

```
independent_var <- dataset1$'HS Grad'
model <- lm(dependent_var ~ independent_var)

ggplot(mapping = aes(x = dependent_var, y = independent_var)) + geom_point() + geom_smooth(method = "lm
```



```
model
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Coefficients:
##     (Intercept)  independent_var
##          19.222           -0.223
```

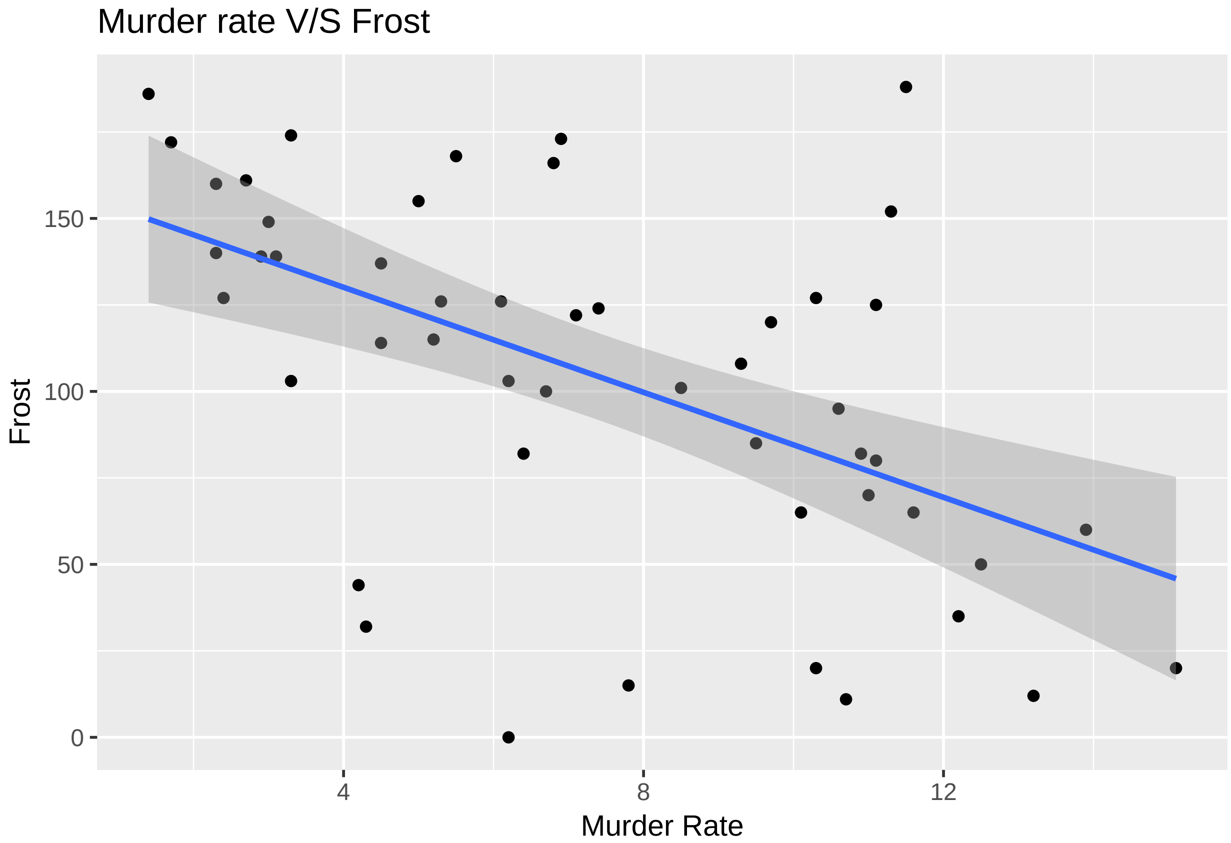
```
summary(model)
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6043 -2.2168  0.0033  2.2734  6.9533
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     19.22236    3.09249   6.216 1.17e-07 ***
## independent_var -0.22302    0.05758  -3.873 0.000325 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.256 on 48 degrees of freedom
## Multiple R-squared:  0.2381, Adjusted R-squared:  0.2222
## F-statistic:    15 on 1 and 48 DF,  p-value: 0.0003248
```

We can observe a linearly decreasing relationship between Murder rate and HS Grad rate and the slope has a relatively low value pf -0.223 which indicates a small change. The t value of -3.873 is relatively not that large, and although it states that there might be an inverse relationship between the two data points, this relationship is not that strong. Also the p-value of 0.0003248 is quite low.

```
independent_var <- dataset1$Frost
model <- lm(dependent_var ~ independent_var)

ggplot(mapping = aes(x = dependent_var, y = independent_var)) + geom_point() + geom_smooth(method = "lm
```



```
model
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Coefficients:
```

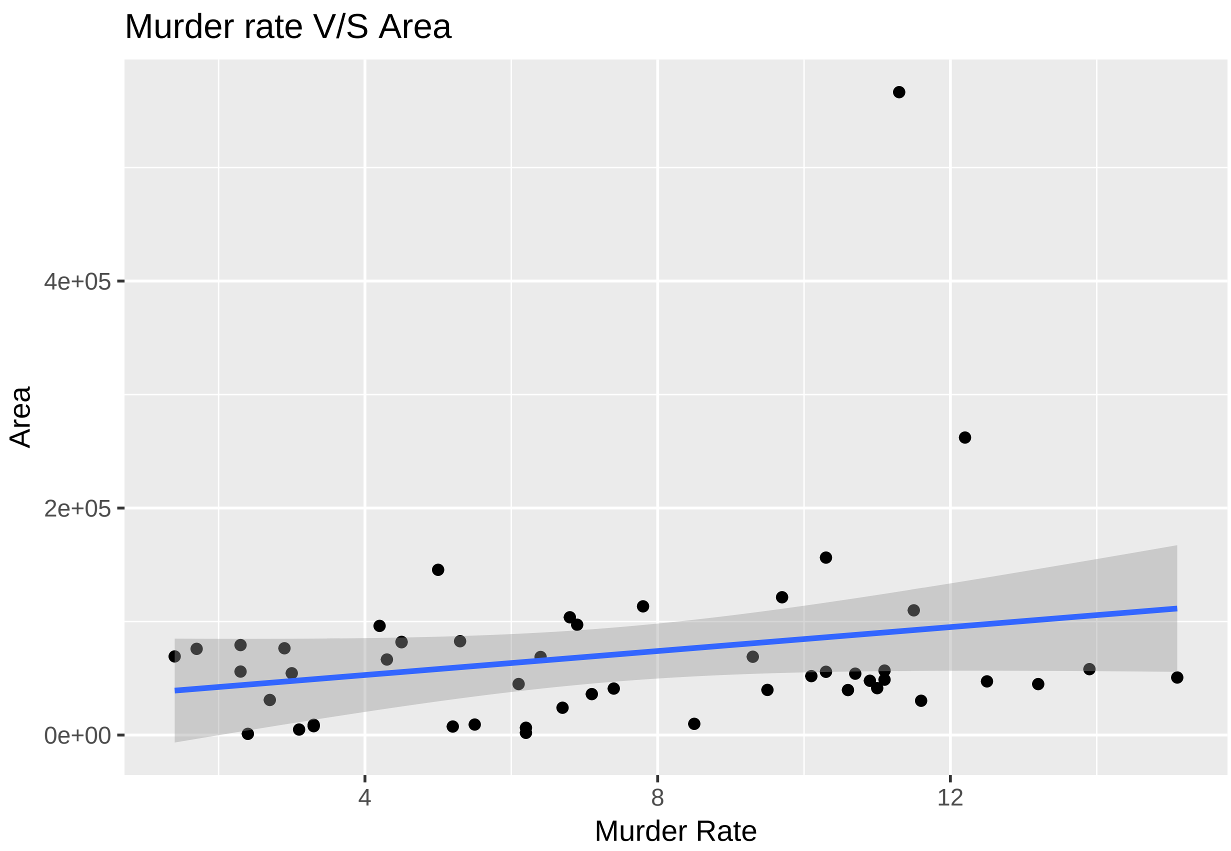
```
##     (Intercept)  independent_var
##        11.37569         -0.03827
```

```
summary(model)
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8510 -2.6336 -0.2825  2.2983  7.3191
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     11.375689   1.005494  11.314 3.83e-15 ***
## independent_var -0.038270   0.008635  -4.432 5.40e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.142 on 48 degrees of freedom
## Multiple R-squared:  0.2904, Adjusted R-squared:  0.2756
## F-statistic: 19.64 on 1 and 48 DF,  p-value: 5.405e-05
```

We can observe a linearly decreasing relationship between Murder rate and HS Grad rate and the slope has a relatively low value of -0.03827 which indicates a small change. The t value of -4.432 is relatively not that large, and although it states that there might be an inverse relationship between the two data points, this relationship is not that strong. Also the p-value is quite low.

```
independent_var <- dataset1$Area
model <- lm(dependent_var ~ independent_var)

ggplot(mapping = aes(x = dependent_var, y = independent_var)) + geom_point() + geom_smooth(method = "lm
```



```
model
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Coefficients:
##     (Intercept)  independent_var
##       6.679e+00        9.881e-06
```

```
summary(model)
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9635 -3.0974 -0.5206  3.0353  7.9199
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.679e+00  6.698e-01   9.972 2.78e-13 ***
## independent_var 9.881e-06  6.079e-06   1.625    0.111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.631 on 48 degrees of freedom
## Multiple R-squared:  0.05216,    Adjusted R-squared:  0.03242
## F-statistic: 2.642 on 1 and 48 DF,  p-value: 0.1106
```

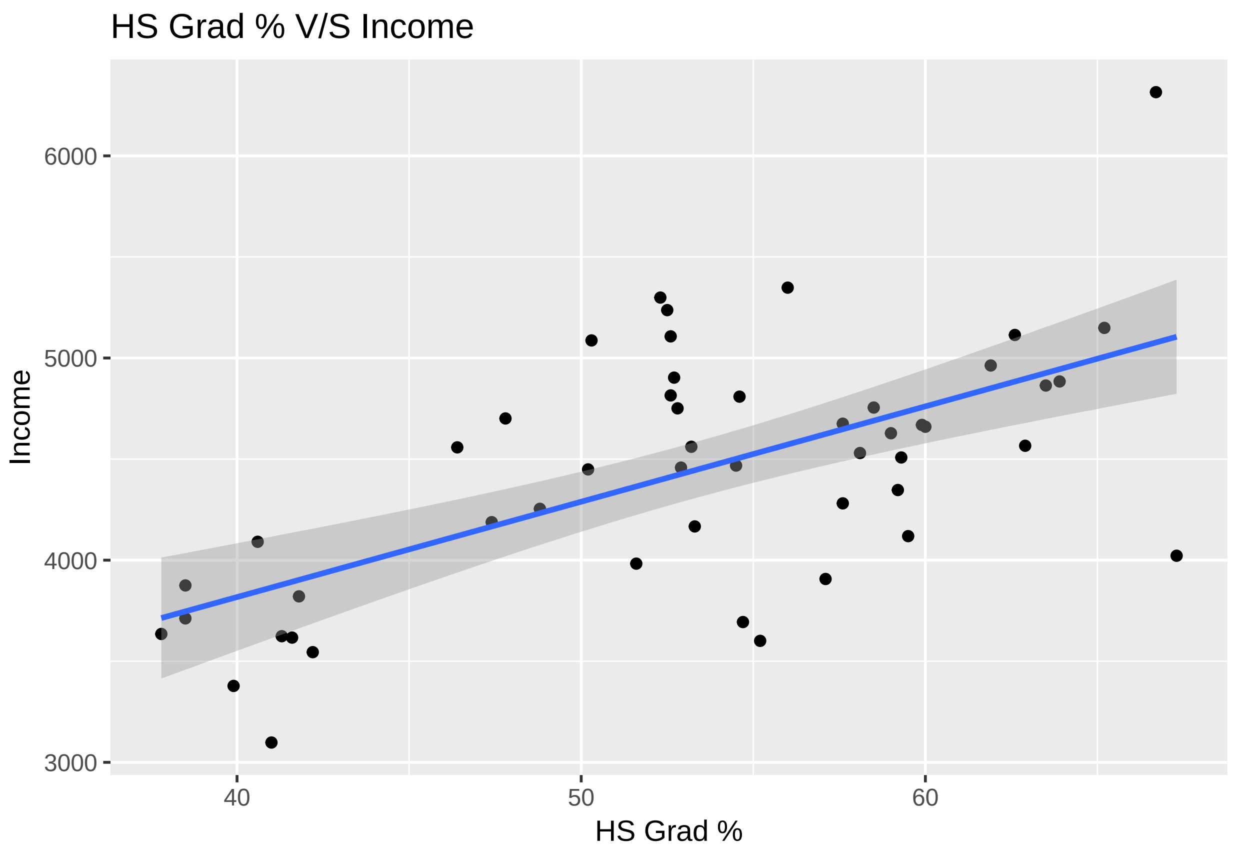
We can observe a linearly increasing relationship between Murder rate and Area size and the slope has a relatively low value which indicates a very very small change. The t value of 1.625 is low, and although it states that there might be an inverse relationship between the two data points, this relationship is not all that strong. Also the p-value is quite high att 0.1106.

We can't always assume that a linear relationship exists between bivariate data as the relationship might be quadratic in nature and depends on the type of data sample that is used for the test.

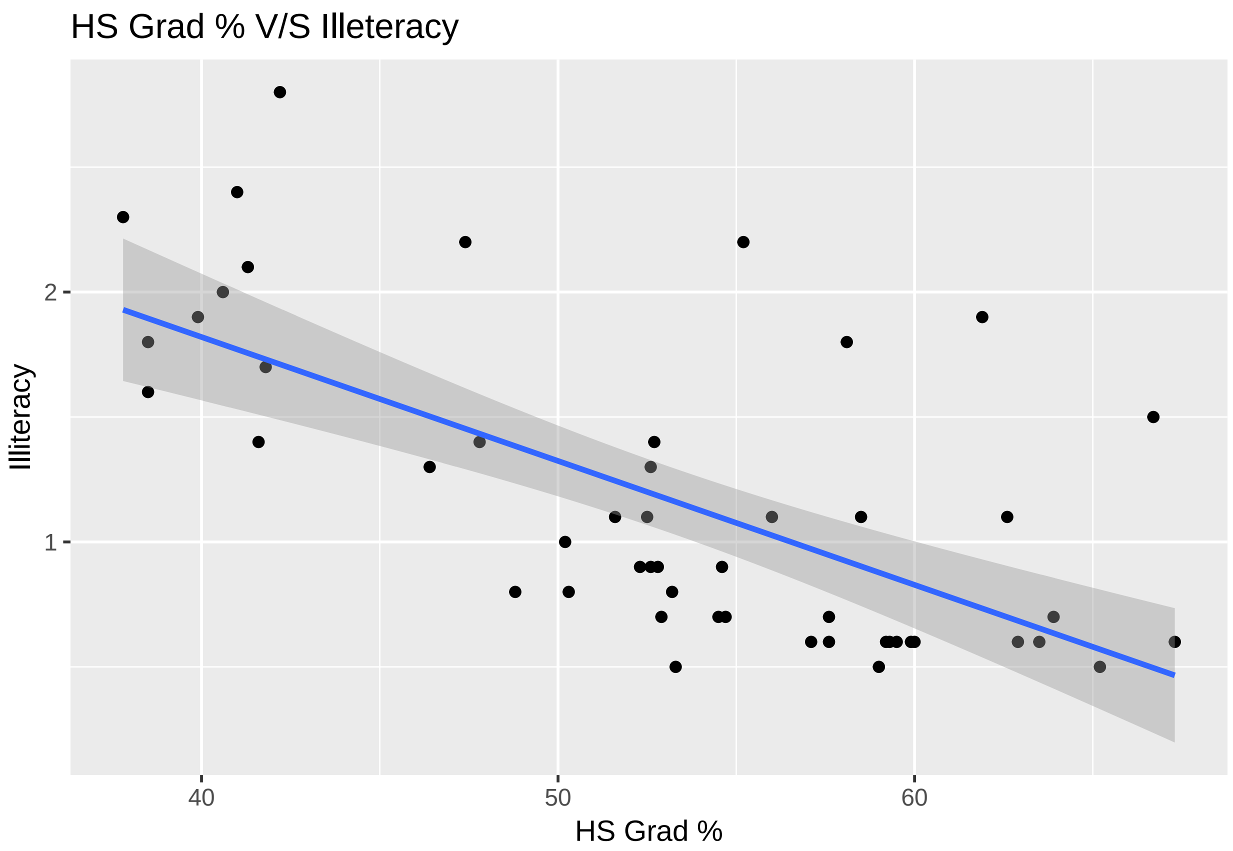
(c) Develop a new research question of your own that you can address using the `state` dataset. Clearly state the question you are going to address. Provide at least one visualization to support your exploration of this question. Discuss what you find in your exploration.

QUESTION: Do people with high school degree have a higher income?

```r
dependent_var <- dataset1$'HS Grad'
independent_var <- dataset1$Income
independent_var2 <- dataset1$Illiteracy

model <- lm(dependent_var ~ independent_var)
model2 <- lm(dependent_var ~ independent_var2)

ggplot(mapping = aes(x = dependent_var, y = independent_var)) + geom_point() + geom_smooth(method = "lm
```



```
ggplot(mapping = aes(x = dependent_var, y = independent_var2)) + geom_point() + geom_smooth(method = "l
```



```
model
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Coefficients:
##     (Intercept)  independent_var
##       16.961557         0.008149
```

```
summary(model)
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.038  -4.774  -1.067   5.022  17.564
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     16.961557   6.665384   2.545   0.0142 *
## independent_var  0.008149   0.001489   5.474 1.58e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 6.403 on 48 degrees of freedom
## Multiple R-squared:  0.3843, Adjusted R-squared:  0.3715
## F-statistic: 29.96 on 1 and 48 DF,  p-value: 1.579e-06
```

```r
model2
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var2)
##
## Coefficients:
##      (Intercept)  independent_var2
##           63.297            -8.708
```

```r
summary(model2)
```

```
##
## Call:
## lm(formula = dependent_var ~ independent_var2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8634  -4.1530  -0.9156   3.0169  16.4658
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         63.297      1.898  33.353  < 2e-16 ***
## independent_var2    -8.708      1.442  -6.041 2.17e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.151 on 48 degrees of freedom
## Multiple R-squared:  0.4319, Adjusted R-squared:  0.4201
## F-statistic: 36.49 on 1 and 48 DF,  p-value: 2.172e-07
```

There is some relationship between the data but we can’t say with absolute certainty that the two variables have a very strong correlation between the data points.

Problem 2: Asking Data Science Questions: Crime and Educational Attainment

In Problem Set 3, you joined data about crimes and educational attainment. Here you will use this new combined dataset to examine questions around crimes in Seattle and the educational attainment of people living in the areas in which the crime occurred. The combined state dataset is available on the course Canvas website (note: this will be available after all students submit Problem Set 3).

(a) Develop a Data Science Question

Develop your own question to address in this analysis. Your question should be specific and measurable, and it should be able to be addressed through a basic analysis of the dataset from Problem Set 3. This analysis must involve at least one hypothesis test. Clearly state what the question is and the suitable null and alternative hypotheses.

#QUESTION: Do locations with a high crime rate have a high number of individuals who are uneducated?

H_o : Mean count of people who have never attended school = 50

H_a : Mean count of people who have never attended school is not 50

(b) Describe and Summarize

Briefly summarize the dataset, describing what data exists and its basic properties. Comment on any issues that need to be resolved before you can proceed with your analysis. Provide descriptive statistics of variables of interest.

```
dataset2 <- read.csv("C:/Users/User/Downloads/combinedCrimeDataset/combinedCrimeDataset.csv")
```

```
nrow(dataset2)
```

```
## [1] 347980
```

```
colnames(dataset2)
```

```
##  [1] "X"
##  [2] "Beat"
##  [3] "Report.Number"
##  [4] "Occurred.Date"
##  [5] "Occurred.Time"
##  [6] "Reported.Date"
##  [7] "Reported.Time"
##  [8] "Crime.Subcategory"
##  [9] "Primary.Offense.Description"
## [10] "Precinct"
## [11] "Sector"
## [12] "Neighborhood"
## [13] "Year"
## [14] "censusId"
## [15] "Location.1"
## [16] "Latitude"
## [17] "Longitude"
## [18] "CensusCode"
## [19] "state"
## [20] "county"
## [21] "GEO.id"
## [22] "GEO.id2"
## [23] "GEO.display.label"
## [24] "total"
## [25] "no_schooling"
## [26] "nursery_school"
## [27] "kindergarten"
## [28] "X1st_grade"
## [29] "X2nd_grade"
## [30] "X3rd_grade"
## [31] "X4th_grade"
## [32] "X5th_grade"
## [33] "X6th_grade"
## [34] "X7th_grade"
## [35] "X8th_grade"
## [36] "X9th_grade"
## [37] "X10th_grade"
## [38] "X11th_grade"
## [39] "X12th_grade_no_diploma"
## [40] "high_school_diploma"
## [41] "ged_or_alternative_credential"
## [42] "some_college_less_than_1_year"
## [43] "some_college_1_or_more_years_no_degree"
## [44] "associates_degree"
## [45] "bachelors_degree"
## [46] "masters_degree"
## [47] "professional_school_degree"
## [48] "doctorate_degree"
```

```r
str(dataset2)
```

```
## 'data.frame':    347980 obs. of  48 variables:
##  $ X                                     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Beat                                  : Factor w/ 51 levels "B1","B2","B3",..: 1 1 1 1 1 1 1 1 1 1
##  $ Report.Number                         : num  2.02e+13 2.02e+13 2.02e+13 2.01e+12 2.01e+12 ...
##  $ Occurred.Date                         : Factor w/ 2684 levels "01/01/2012","01/01/2013",..: 2075 2
##  $ Occurred.Time                         : int  1449 230 2200 1200 1247 1900 2345 2110 2100 2100 ...
##  $ Reported.Date                         : Factor w/ 2684 levels "01/01/2012","01/01/2013",..: 2075 2
##  $ Reported.Time                         : int  1449 1517 2320 1445 1247 4 106 927 937 606 ...
##  $ Crime.Subcategory                     : Factor w/ 30 levels "AGGRAVATED ASSAULT",..: 29 6 8 25 11
##  $ Primary.Offense.Description           : Factor w/ 137 levels "ADULT-VULNERABLE-FINANCIAL",..: 125
##  $ Precinct                              : Factor w/ 7 levels "","EAST","NORTH",..: 3 3 3 3 3 3 3 3 3
##  $ Sector                                : Factor w/ 19 levels "","6804","B",..: 3 3 3 3 3 3 3 3 3 3
##  $ Neighborhood                          : Factor w/ 59 levels "ALASKA JUNCTION",..: 4 4 4 4 4 4 4 4
##  $ Year                                  : int  2018 2018 2015 2012 2012 2017 2016 2015 2018 2014 ..
##  $ censusId                              : num  5.3e+10 5.3e+10 5.3e+10 5.3e+10 5.3e+10 ...
##  $ Location.1                            : Factor w/ 51 levels "(47.5093533353672, -122.259542630385)
##  $ Latitude                              : num  47.7 47.7 47.7 47.7 47.7 ...
##  $ Longitude                             : num  -122 -122 -122 -122 -122 ...
##  $ CensusCode                            : num  5.3e+14 5.3e+14 5.3e+14 5.3e+14 5.3e+14 ...
##  $ state                                 : int  53 53 53 53 53 53 53 53 53 53 ...
##  $ county                                : int  33 33 33 33 33 33 33 33 33 33 ...
##  $ GEO.id                                : Factor w/ 48 levels "1400000US53033000200",..: 5 5 5 5 5 5
##  $ GEO.id2                               : num  5.3e+10 5.3e+10 5.3e+10 5.3e+10 5.3e+10 ...
##  $ GEO.display.label                     : Factor w/ 48 levels "Census Tract 100.01, King County, Was
##  $ total                                 : int  4155 4155 4155 4155 4155 4155 4155 4155 4155 4155 ..
##  $ no_schooling                          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nursery_school                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ kindergarten                          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X1st_grade                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X2nd_grade                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X3rd_grade                            : int  15 15 15 15 15 15 15 15 15 15 ...
##  $ X4th_grade                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X5th_grade                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X6th_grade                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X7th_grade                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X8th_grade                            : int  33 33 33 33 33 33 33 33 33 33 ...
##  $ X9th_grade                            : int  18 18 18 18 18 18 18 18 18 18 ...
##  $ X10th_grade                           : int  110 110 110 110 110 110 110 110 110 110 ...
##  $ X11th_grade                           : int  20 20 20 20 20 20 20 20 20 20 ...
##  $ X12th_grade_no_diploma                : int  34 34 34 34 34 34 34 34 34 34 ...
##  $ high_school_diploma                   : int  472 472 472 472 472 472 472 472 472 472 ...
##  $ ged_or_alternative_credential         : int  100 100 100 100 100 100 100 100 100 100 ...
##  $ some_college_less_than_1_year         : int  245 245 245 245 245 245 245 245 245 245 ...
##  $ some_college_1_or_more_years_no_degree: int  536 536 536 536 536 536 536 536 536 536 ...
##  $ associates_degree                     : int  310 310 310 310 310 310 310 310 310 310 ...
##  $ bachelors_degree                      : int  1301 1301 1301 1301 1301 1301 1301 1301 1301 1301 ..
##  $ masters_degree                        : int  760 760 760 760 760 760 760 760 760 760 ...
##  $ professional_school_degree            : int  64 64 64 64 64 64 64 64 64 64 ...
##  $ doctorate_degree                      : int  137 137 137 137 137 137 137 137 137 137 ...
```

The dataset contains the information regarding crimes. It has fields like beat to show which police cars were in the area, the area that it took place, the type of crime that was commited. THe time that it was actually committed and the time that the crime was actually reported. It also has information about the education levels of the people that reside in the area. Each education field column contains the total number of people who have the level of education that is mentioned as part of the column name. We should be able to predict from the data the tracts that have a high rate of crime and whether or not the people within that area are well educated.

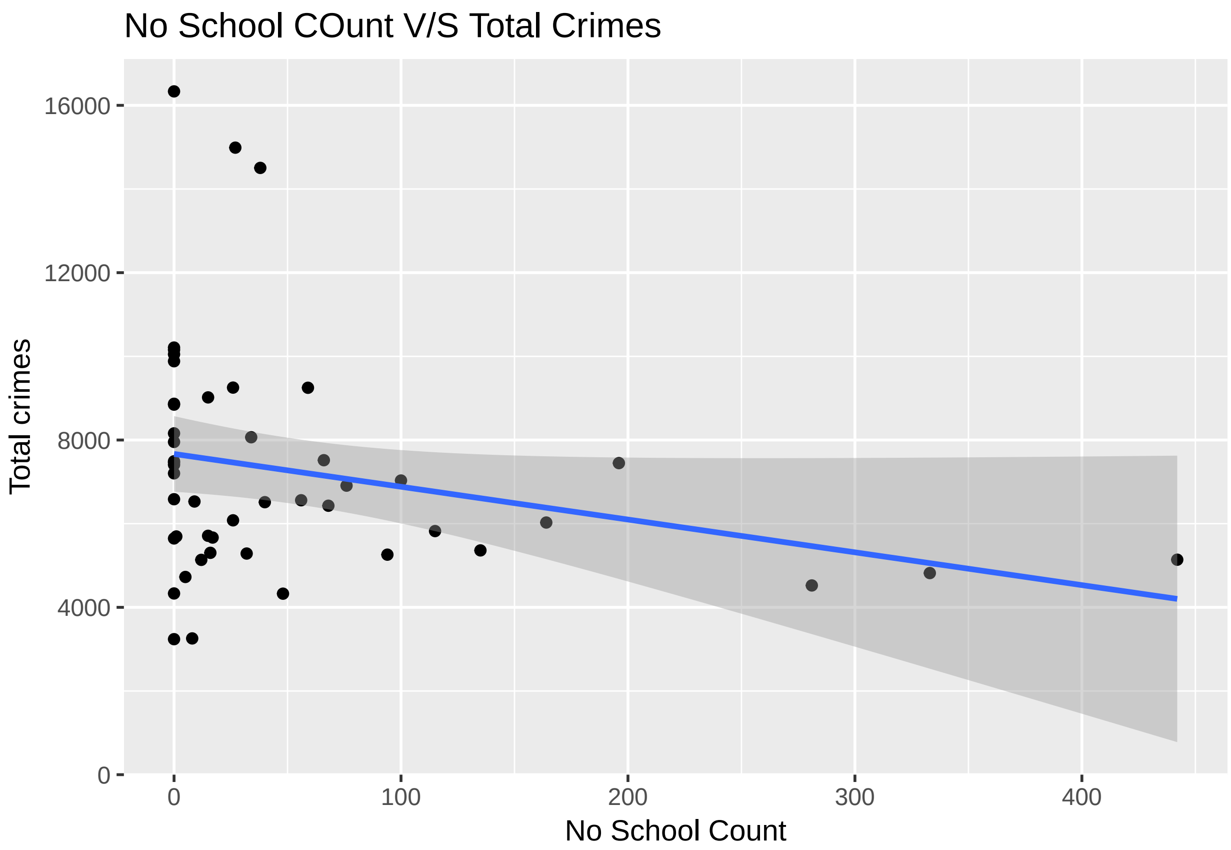
(c) Data Analysis

Use the dataset to provide empirical evidence to answer your question from part (a). Discuss your results. Provide at least one visualization to support your narrative. (NOTE: you will not be graded on whether you see statistically significant results but rather on your interpretation of findings)

```r
res <- dataset2 %>%
  group_by(GEO.display.label) %>%
  summarise(total_no_school = mean(no_schooling))

res2 <- dataset2 %>%
  group_by(GEO.display.label) %>%
  summarise(total_crimes = n())

joined_res <- inner_join(res,res2,by = 'GEO.display.label')
joined_res
```

```
## # A tibble: 48 x 3
##    GEO.display.label                                 total_no_school total_crimes
##    <fct>                                                       <dbl>        <int>
##  1 Census Tract 100.01, King County, Washingt~                   196         7448
##  2 Census Tract 102, King County, Washington                      76         6909
##  3 Census Tract 105, King County, Washington                      40         6514
##  4 Census Tract 108, King County, Washington                       0         4332
##  5 Census Tract 109, King County, Washington                       0         3239
##  6 Census Tract 11, King County, Washington                        0        10049
##  7 Census Tract 110.01, King County, Washingt~                   333         4819
##  8 Census Tract 113, King County, Washington                     135         5361
##  9 Census Tract 114.01, King County, Washingt~                    68         6429
## 10 Census Tract 116, King County, Washington                      32         5286
## # ... with 38 more rows
```

```
model <- lm(joined_res$total_no_school ~ joined_res$total_crimes)

ggplot(mapping = aes(x = joined_res$total_no_school, y = joined_res$total_crimes)) + geom_point() + geom
```



```
model
```

```
##
## Call:
## lm(formula = joined_res$total_no_school ~ joined_res$total_crimes)
##
## Coefficients:
##             (Intercept)  joined_res$total_crimes
##              116.605031                -0.008745
```

```
summary(model)
```

```
##
## Call:
## lm(formula = joined_res$total_no_school ~ joined_res$total_crimes)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -88.28 -51.45 -29.45  23.30 370.33
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              116.605031  36.807674   3.168  0.00273 **
## joined_res$total_crimes  -0.008745   0.004753  -1.840  0.07224 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 89.68 on 46 degrees of freedom
## Multiple R-squared:  0.06855,    Adjusted R-squared:  0.0483
## F-statistic: 3.385 on 1 and 46 DF,  p-value: 0.07224
```

```r
mean = mean(res$total_no_school)
standard_dev = sd(res$total_no_school)
len = length(res$total_no_school)
df = len - 1

t.test(res$total_no_school, mu = 50)
```

```
##
##  One Sample t-test
##
## data:  res$total_no_school
## t = 0.24179, df = 47, p-value = 0.81
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  26.51441 79.90226
## sample estimates:
## mean of x
##  53.20833
```

```r
test_stat <- (mean - 50)/(standard_dev/sqrt(len))

tstat <- qt(0.975, df)

errBound <- tstat*standard_dev/sqrt(len)

c(mean - errBound, mean + errBound)
```

```
## [1] 26.51441 79.90226
```

```r
t.test(res$total_no_school, mu = 50, alternative = 'g')
```

```
##
##  One Sample t-test
##
## data:  res$total_no_school
## t = 0.24179, df = 47, p-value = 0.405
## alternative hypothesis: true mean is greater than 50
## 95 percent confidence interval:
##  30.94381      Inf
## sample estimates:
## mean of x
##  53.20833
```

```r
tstat1 <- qt(0.95, df)
```

My p-value of 0.81 is relatively large and it indicates weak evidence against the null-hypothesis, so we fail to reject the null hypothesis.

(d) Reflect and Question

Comment on the questions (and answers) in this analysis. Were you able to adequately answer your question? Is there additional data that would help provide a more clear picture of the problem you are analyzing?

Based on the preliminary analysis and data visualization conducted by me to reach a conclusion regarding the hypothesis posited, it appears that there seems to be no real correlation between the number of uneducated/illiterate people and the number of crimes commited in each location. There are certain regions with 0 people with no schooling but have extremely high crime rate exceeding the 16000 mark. It would be better to supplement this information with employment rates because being illiterate doesn't equate to being broke as crimes are generally a result of a lack of disposable income. Information regarding employment, average wages and disposable income would help me confirm or deny my hypothesis over the data that is currently available.

Problem 3: Sampling with and without Replacement

In the following situations assume that half of the specified population wears glasses and the other half does not.

(a) Suppose you're sampling from a room with 10 people. What is the probability of sampling two people wearing glasses in a row when sampling with replacement? What is the probability when sampling without replacement?

```
P_glass_replaced <- 5/10 * 5/10
P_glass_not_replaced <- 5/10 * 4/9
P_glass_replaced
```

```
## [1] 0.25
```

```
P_glass_not_replaced
```

```
## [1] 0.2222222
```

(b) Now suppose you're sampling from a stadium with 10,000 people. What is the probability of sampling two people wearing glasses in a row when sampling with replacement? What is the probability when sampling without replacement?

```
P_glass_replaced_2 <- 5000/10000 * 5000/10000
P_glass_not_replaced_2 <- 5000/10000 * 4999/9999
P_glass_replaced_2
```

```
## [1] 0.25
```

```
P_glass_not_replaced_2
```

```
## [1] 0.249975
```

(c) We often treat individuals who are sampled from a large population as independent. Using your findings from parts (a) and (b), explain whether or not this assumption is reasonable.

The assumption would be correct. We get a value of 0.249975 for the sample that consists of 10000 people and this is quite close in value to the 0.25 that we get from sampling with replacement.