# INFX 573: Problem Set 3 - Data Analysis

*Pierre Augustamar*

*Due: Monday, October 18, 2016*

### Collaborators:

### Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset3.Rmd` file from Canvas. Open `problemset3.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset3.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.
4. Collaboration on problem sets is acceptable, and even encouraged, but each student must turn in an individual write-up in his or her own words and his or her own work. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. 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 R Markdown file to `YourLastName_YourFirstName_ps3.Rmd`, knit a 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)

# Load ny flights dataset
library(nycflights13)
```

### Problem 1: Flight Delays

Flight delays are often linked to weather conditions. How does weather impact flights from NYC? Utilize both the `flights` and `weather` datasets from the `nycflights13` package to explore this question. Include at least two visualizations to aid in communicating what you find.

*Response 1 - To validate how weather affects flights, I will analyze the relationship between departure delay and Wind's speed.*

Each airplane's models has specific guidance on the minimum requirement for an aircraft to take off under various weather systems including wind speed. That said, in this analysis, I will use a baseline of 25 knots as the minimum requirements for an airplane to delay departure.

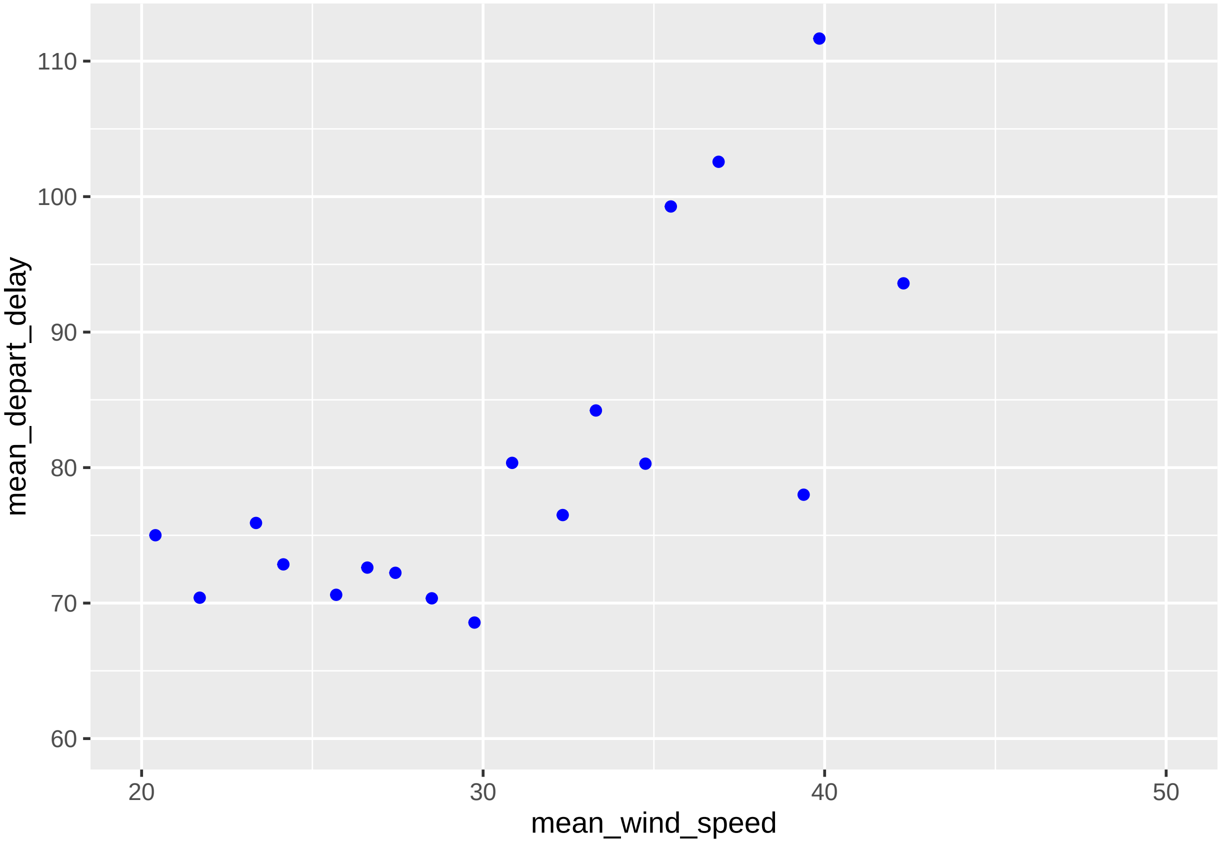
```
head(flights) # get the top 6 flights records in order to understand the data
```

```
## # A tibble: 6 × 19
##    year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1  2013     1     1      517            515         2      830
## 2  2013     1     1      533            529         4      850
## 3  2013     1     1      542            540         2      923
## 4  2013     1     1      544            545        -1     1004
## 5  2013     1     1      554            600        -6      812
## 6  2013     1     1      554            558        -4      740
## # ... with 12 more variables: sched_arr_time <int>, arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,
## #   time_hour <dttm>
```

```
head(weather) # get the top 6 weather records in order to understand the data
```

```
## # A tibble: 6 × 15
##   origin  year month   day  hour  temp  dewp humid wind_dir wind_speed
##    <chr> <dbl> <dbl> <int> <int> <dbl> <dbl> <dbl>    <dbl>      <dbl>
## 1    EWR  2013     1     1     0 37.04 21.92 53.97      230   10.35702
## 2    EWR  2013     1     1     1 37.04 21.92 53.97      230   13.80936
## 3    EWR  2013     1     1     2 37.94 21.92 52.09      230   12.65858
## 4    EWR  2013     1     1     3 37.94 23.00 54.51      230   13.80936
## 5    EWR  2013     1     1     4 37.94 24.08 57.04      240   14.96014
## 6    EWR  2013     1     1     6 39.02 26.06 59.37      270   10.35702
## # ... with 5 more variables: wind_gust <dbl>, precip <dbl>,
## #   pressure <dbl>, visib <dbl>, time_hour <dttm>
```

```
#tidy data
flights_data <- flights %>%
             # remove records that contain na records
             filter(!is.na(arr_delay)) %>%
             # pick out flights that have arrival delays greater than 15 minutes
             filter(arr_delay > 15) %>%
             # remove records that contan na records
             filter(!is.na(dep_delay)) %>%
             # pick out flights that have departure delays greater than 15 minutes
             filter(dep_delay > 15) %>%
            # join flights with weather based on year, month, day, origin, hour and time_hour
            inner_join(weather, by = c("year", "month", "day", "origin", "hour", "time_hour")) %>%
           # select origin, depature delay, month and wind_speed for analysis
           select(origin, dep_delay, month, wind_speed)

# use average departure delay and average wind speed to
#plot out the relationship between departure delays and wind's speed.
flights_delay <- flights_data %>%
        group_by(wind_speed) %>%
        summarize(mean_depart_delay=mean(dep_delay, na.rm=TRUE), mean_wind_speed=mean(wind_speed, na.rm=

#plot graph showing relationship between average wind speed and departure delays
ggplot(flights_delay, aes(x=mean_wind_speed, y=mean_depart_delay)) + geom_jitter(color="blue") + xlim(20
## Warning: Removed 18 rows containing missing values (geom_point).
```



Wind's speed over 25 knots could affect the way planes would have to take off. Pilots would have to perform extremes maneuver to ensure that the airplane can stay stabilize at high wind's speed. The graph between departure delay and wind speed shows that there is a correlation between these two observations. Though this correlation is moderate with a non-linear form, there is still enough evidence to confirm that has the wind's speed is higher than 25 knots, more flights are delayed from taking off.

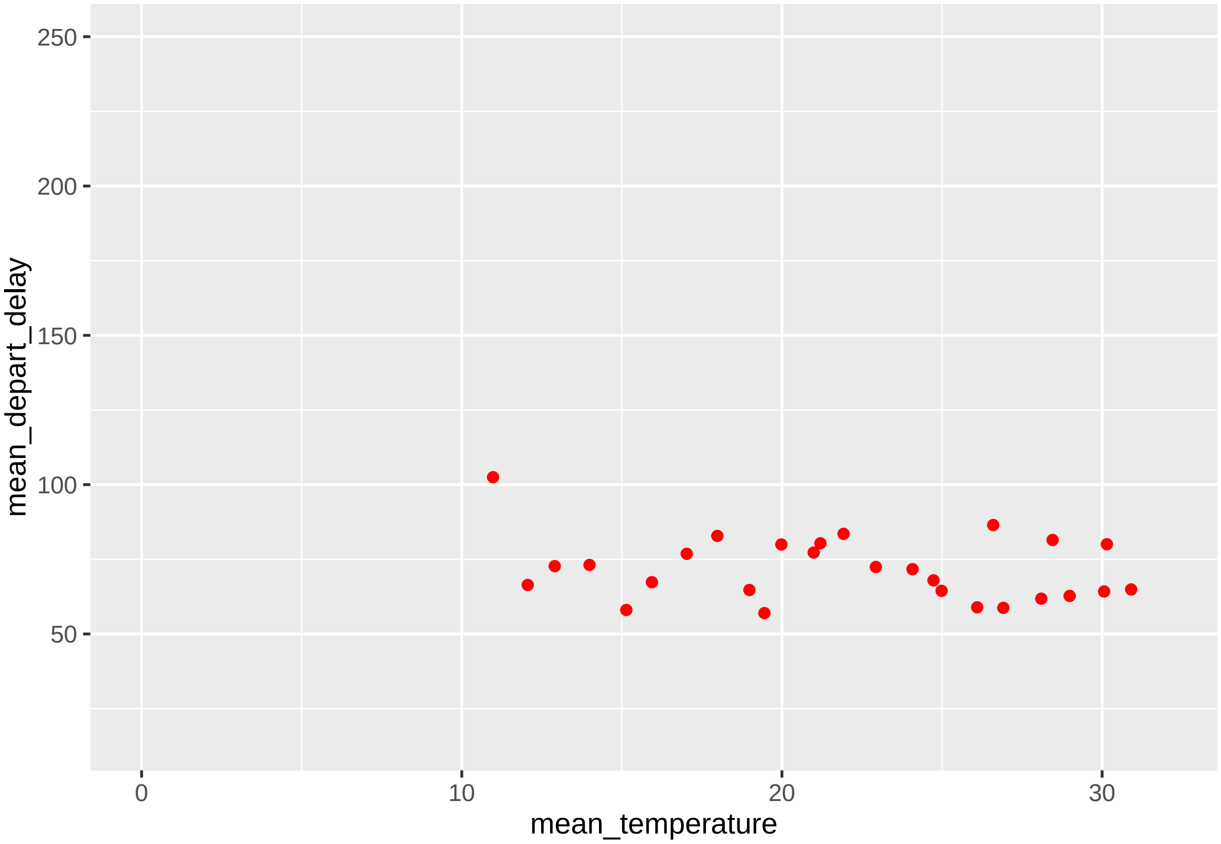
*Response 2 - Departure delay in the 3 airports due to extrem weather*

Extreme weather can be any conditions that include tornadoes, hurricanes, blizzards, etc. For this analysis, I will investigate how freezing temperature could force flights departure to be delayed.

```
#tidy data
flights_data <- flights %>%
            # remove records that contain na records
            filter(!is.na(arr_delay)) %>%
            # pick out flights that have arrival delays greater than 15 minutes
            filter(arr_delay > 15) %>%
            filter(!is.na(dep_delay)) %>%  # remove records that contan na records
            # pick out flights that have departure delays greater than 15 minutes
            filter(dep_delay > 15) %>%
            # join flights with weather based on year, month, day, origin, hour and time_hour
            inner_join(weather, by = c("year", "month", "day", "origin", "hour", "time_hour")) %>%
            # select origin, depature delay, month and temp for analysis
            select(origin, dep_delay, month, temp)
```

```
# use average departure delay and average wind speed to plot out
#the relationship between departure delays and wind's speed.
flights_delay <- flights_data %>%
        group_by(temp) %>%
        summarize(mean_depart_delay=mean(dep_delay, na.rm=TRUE), mean_temperature=mean(temp, na.rm=TRUE

#plot between the average temperature and average departure delay
ggplot(flights_delay, aes(x=mean_temperature, y=mean_depart_delay)) + geom_jitter(color="red") + xlim(0
```

```
## Warning: Removed 80 rows containing missing values (geom_point).
```



When the temperature is below freezing level (< 32F) then many flights are delayed. Interestingly, the data show that almost the same number of flights are delayed with some variations here and there. We also noticed that no flights took off under 10F. The data show a non-linear form with no particular direction for either a positive or negative trend. However, we can confirm a correlation within the same range of flights being delayed when the temperature falls below freezing level. However, it should be noted that we noticed an outlier of over 100 delays. These delays can be explained with possible snow on the ground. Thus, planes needed to be de-iced before taking off.

### Problem 2: 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.**

```r
str(state.x77) # check the type of object
```

```
##  num [1:50, 1:8] 3615 365 2212 2110 21198 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##   ..$ : chr [1:8] "Population" "Income" "Illiteracy" "Life Exp" ...
```

```r
dim(state.x77) # Determine the number of rows and columns
```

```
## [1] 50  8
```

```r
head(state.x77) # view the top 6 records
```

```
##            Population Income Illiteracy Life Exp Murder HS Grad Frost
## Alabama          3615   3624        2.1    69.05   15.1    41.3    20
## Alaska            365   6315        1.5    69.31   11.3    66.7   152
## Arizona          2212   4530        1.8    70.55    7.8    58.1    15
## Arkansas         2110   3378        1.9    70.66   10.1    39.9    65
## California      21198   5114        1.1    71.71   10.3    62.6    20
## Colorado         2541   4884        0.7    72.06    6.8    63.9   166
##              Area
## Alabama     50708
## Alaska     566432
## Arizona    113417
## Arkansas    51945
## California 156361
## Colorado   103766
```

```r
summary(state.x77) # view the full data set in tabular form
```

```
##    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
```

*Population:* estimated polpulation as of July 1, 1975 *Income:* income per capita *Illiteracy:* percent of the population unable to read and write *Life Exp:* life expectancy in years *Muder:* murder rate per 100,000 population *HS Grad:* percentage of high-school graduates *Frost:* minimum temperature below freezing in capita *Area:* land area in square miles

```r
stateInfo <- as.data.frame(state.x77) #convert the matrix data as a data frame

str(stateInfo) # check the internal structure of the state data
```

```
## '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 ...
```

```r
summary(stateInfo) #
```

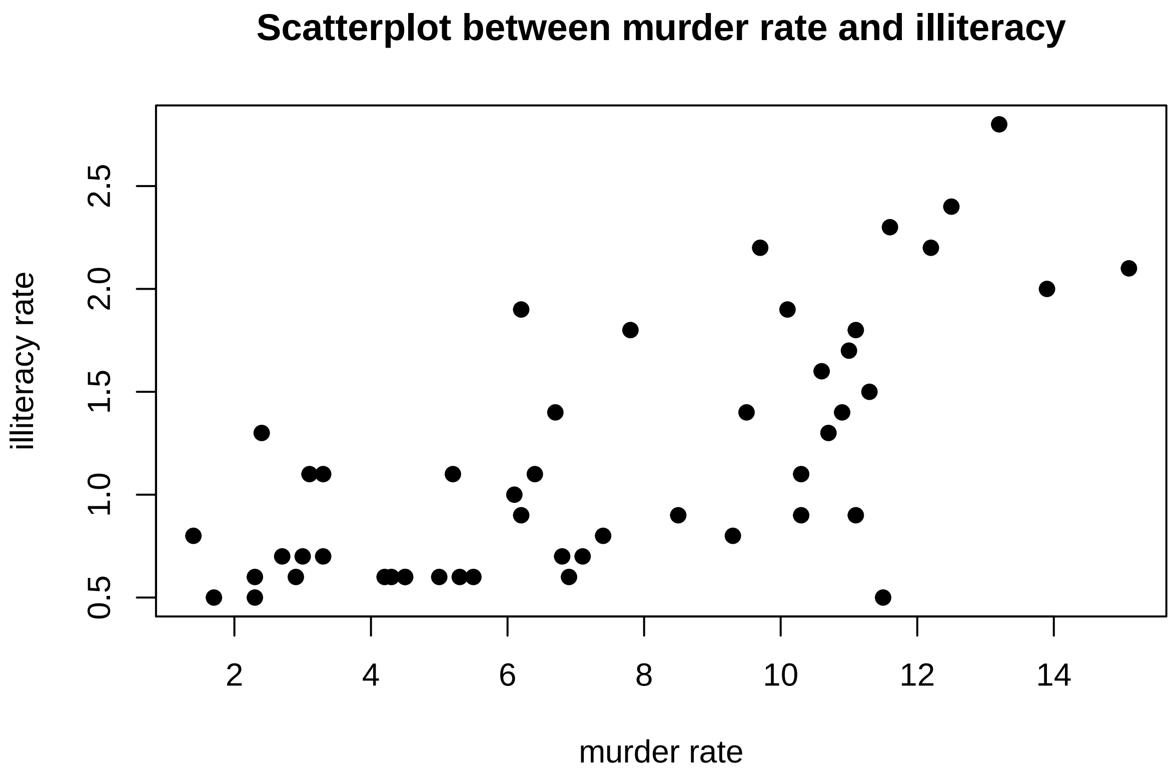
```
##    Population        Income       Illiteracy       Life Exp    
##  Min.   :  365   Min.   :3098   Min.   :0.500   Min.   :67.96  
##  1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12  
##  Median : 2838   Median :4519   Median :0.950   Median :70.67  
##  Mean   : 4246   Mean   :4436   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
```

```r
# check that all the values are numeric by calculating the mean.
# If any calculation returns an error, then we would have to remove any of the non-numeric values
stateInfo %>% summarize(population_mean = mean(stateInfo$Population),
                       income_mean= mean(stateInfo$Income),
                       illiteracy_mean = mean(stateInfo$Illiteracy),
                       life_Exp_mean=mean(stateInfo$`Life Exp`),
                       murder_mean=mean(stateInfo$Murder),
                       hS_grad_mean=mean(stateInfo$`HS Grad`),
                       frost_mean=mean(stateInfo$Frost),
                       area_mean=mean(stateInfo$Area))
```

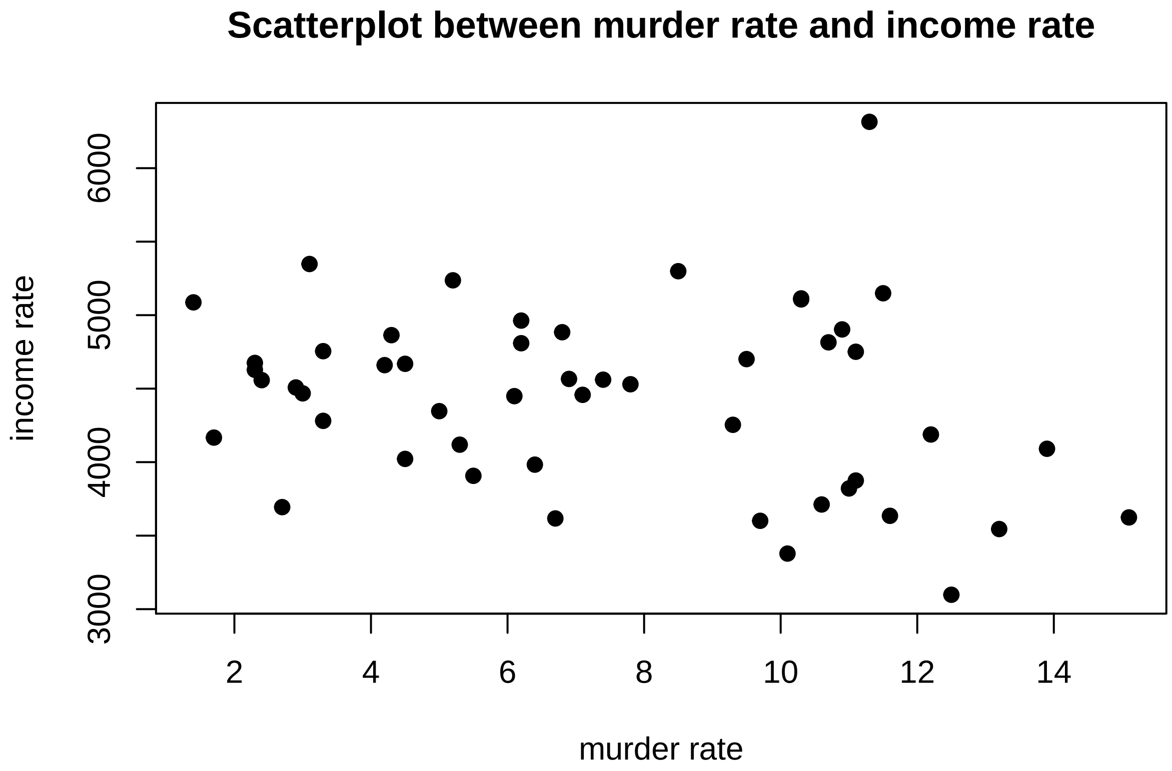
```
##   population_mean income_mean illiteracy_mean life_Exp_mean murder_mean
## 1         4246.42      4435.8            1.17       70.8786       7.378
##   hS_grad_mean frost_mean area_mean
## 1       53.108     104.46  70735.88
```

**(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. What does your analysis suggest might be important variables to consider in building a model to explain variation in murder rates?**

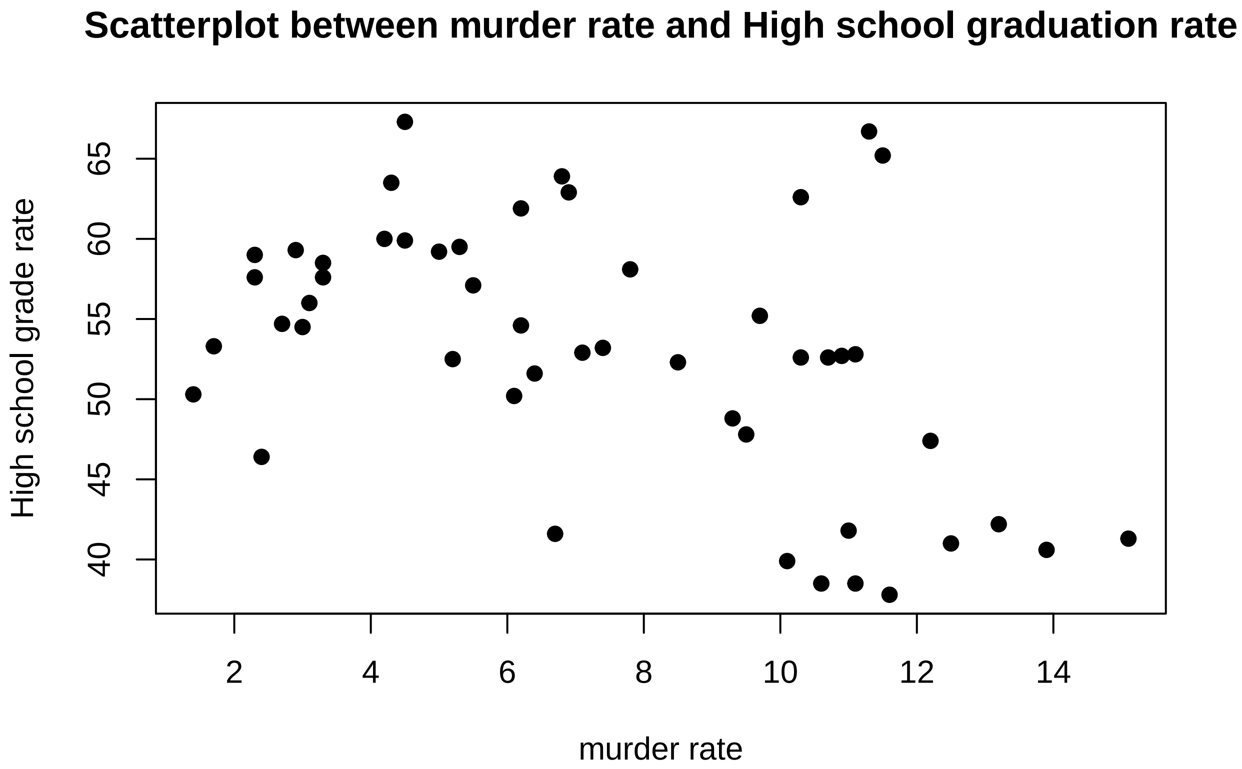
```
attach(stateInfo)

# case 1 - scatter plot between murder rate and illiteracy rate
plot(stateInfo$Murder, stateInfo$Illiteracy, main="Scatterplot between murder rate and illiteracy", xlab
```



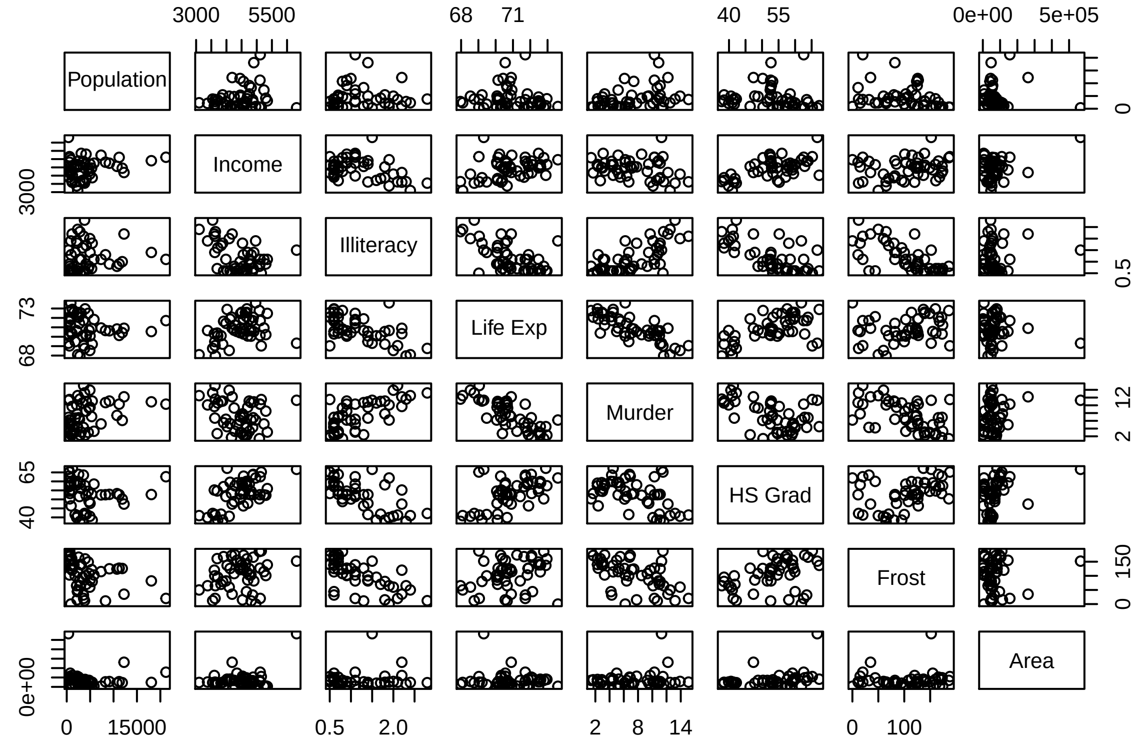
```
# case 2 - scatter plot between murder rate and income rate
plot(stateInfo$Murder, stateInfo$Income, main="Scatterplot between murder rate and income rate", xlab =
```



```
#case 3 - scatter plot murder rate in relation to high school graduation
plot(stateInfo$Murder, stateInfo$`HS Grad`, main="Scatterplot between murder rate and High school gradu
```



```
#generate a matrix of scattered plots for all of the variables.
#This gives a general idea of all the possible relationships between all the variables
pairs(stateInfo)
```



```
cor(stateInfo) #calculate correlation coefficient for further analysis of the the model
```

```
##              Population     Income  Illiteracy    Life Exp     Murder
## Population   1.00000000  0.2082276  0.10762237 -0.06805195  0.3436428
## Income       0.20822756  1.0000000 -0.43707519  0.34025534 -0.2300776
## Illiteracy   0.10762237 -0.4370752  1.00000000 -0.58847793  0.7029752
## Life Exp    -0.06805195  0.3402553 -0.58847793  1.00000000 -0.7808458
## Murder       0.34364275 -0.2300776  0.70297520 -0.78084575  1.0000000
## HS Grad     -0.09848975  0.6199323 -0.65718861  0.58221620 -0.4879710
## Frost       -0.33215245  0.2262822 -0.67194697  0.26206801 -0.5388834
## Area         0.02254384  0.3633154  0.07726113 -0.10733194  0.2283902
##                 HS Grad      Frost        Area
## Population  -0.09848975 -0.3321525  0.02254384
## Income       0.61993232  0.2262822  0.36331544
## Illiteracy  -0.65718861 -0.6719470  0.07726113
## Life Exp     0.58221620  0.2620680 -0.10733194
## Murder      -0.48797102 -0.5388834  0.22839021
## HS Grad      1.00000000  0.3667797  0.33354187
## Frost        0.36677970  1.0000000  0.05922910
## Area         0.33354187  0.0592291  1.00000000
```

case 1: The scatter plot depecting murder rate in contrast to illiteracy shows that murder rate tends to increase for states with high rate of illiteracy. The graph shows a moderate positive linear trend with a few outliers.

Case 2: The scatter plot depecting murder rate in comparison to income rate shows that murder rate tends to increase as income rate decreases. The grap shows a weak negative linear terend with a few outliers.

case 3: the scatter plot depecting murder rate in comparison to high school graduration rate shows no particular trend either positve or negative. The graph has a number of clustered data here and there, but no specific form or direction can be found.

Analysis for a model: I will be using the Pearson's correlation measurement to come up with a set of variables that would provide sufficient evidence that can be used to show how dependent variables are affected by the independent variable, murder rate.

First, I generated a matrix that contains all the possible combination of scattered plots between each of the variables. Looking at the row that contains the murder variable it shows a clear picture of possible patterns with other variables. It shows the followings: - Possible negative correlation between murder and life expectancy - Possible negative correlation between murder and freezing temperature. - Possible positive correlation between murder and illiteracy. - No deterministic relationship between murder and income

Second, I calculated the Pearson's r value to investigate the possible relationships between murder rate and other dependent variables. The results can be summarized as follows: - When comparing murder to illiteracy, the coefficient value = 0.70. This means that there is a strong positive linear correlation. - When comparing murder to life expectancy, the coefficient value = -0.78. This means that there is a strong negative correlation between murder rate and life expectancy rate. - Both the rate of high school graduation and freezing have a coefficient of -0.48 and -0.53 respectively. These values tell us that there is a moderate negative correlation between the rate of murder and the rate fo freezing temperature as well as the rate of high-school graduation.

Conclusion, based on the analysis above, illiteracy and life expectancy could be used to model either a strong positive correlation or a strong negative correlation. In other words, states with higher illiteracy tend to have a higher murder rate. On the other hand, states with a lower rate of murders tend to have people with higher life expectancy.

### (c) Choose one variable and fit a simple linear regression model, $Y = \beta_1 X + \beta_0$, using the `lm()` function in R. Describe your results.

```
library(ISLR) # load statistical package for R
data("Auto")
linear_fit = lm(stateInfo$Murder ~ stateInfo$Illiteracy, data=Auto) # linear model to show relationship
summary(linear_fit) # generate a summary of the linear model
```

```
##
## Call:
## lm(formula = stateInfo$Murder ~ stateInfo$Illiteracy, data = Auto)
##
## 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 **
## stateInfo$Illiteracy   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
```

In this linear regression, I will be representing the relationship between murder's rate and the illiteracy rate. Based on the summary result, my analysis of the result is as follows:

Residuals: The five summary results for residuals access how well the model fit the data. It shows that the distribution of the residuals does not appear to be strongly symmetrical. It signifies that the model will have certain points that fall away from the observed points.

Coefficients:

a. Estimate It takes an average illiteracy of 2.39% to see a relationship between murder rate and illiteracy.

b. Standard Error The required level of illiteracy to relate to a murder can vary by 0.81.

c. t-value Ideally, we would like the t-value to be as far away as possible from zero to strongly reject the null hypothesis. Our data set shows a t-value = 2.928. It means that we will be able to reject the null hypothesis, which is there is a relationship between murder rate and illiteracy rate.

d. Pr(>|t|) A small value that's close to zero for the intercept indicates that we can reject the null hypothesis. Our data sets show a p-value = 0.0052 for the intercept. Thus, we can reject the null hypothesis which tells us that there is a relationship between murder rate and illiteracy rate.

e. Significance stars This component shows a significance stars that are between 2 to 3 stars. In this scenario, three stars indicate that it's likely that a relationship exists between illiteracy rate and murder rate.

Residual standard error/Degrees of freedom: The actual illiteracy percentage required to have a relationship with the average murder rate will be approximately around 2.65 percent on average. Our data sets show that the mean percentage for illiteracy rate to relate to murder rate is 2.39 percent. Based on this, we can say that the percentage error is 110 percent. The fact percent error is greater than 100% signifies that the expected value might be off from the actual value. The degree of freedom is 48 because we have 50 data points that are analyzed against two parameters, intercept, and slope.

Multiple R-squared: Our results set shows an R-squared of 0.49 which means that 49% of the variance found in the response variable can be explained by the predictor variable which is the illiteracy rate.

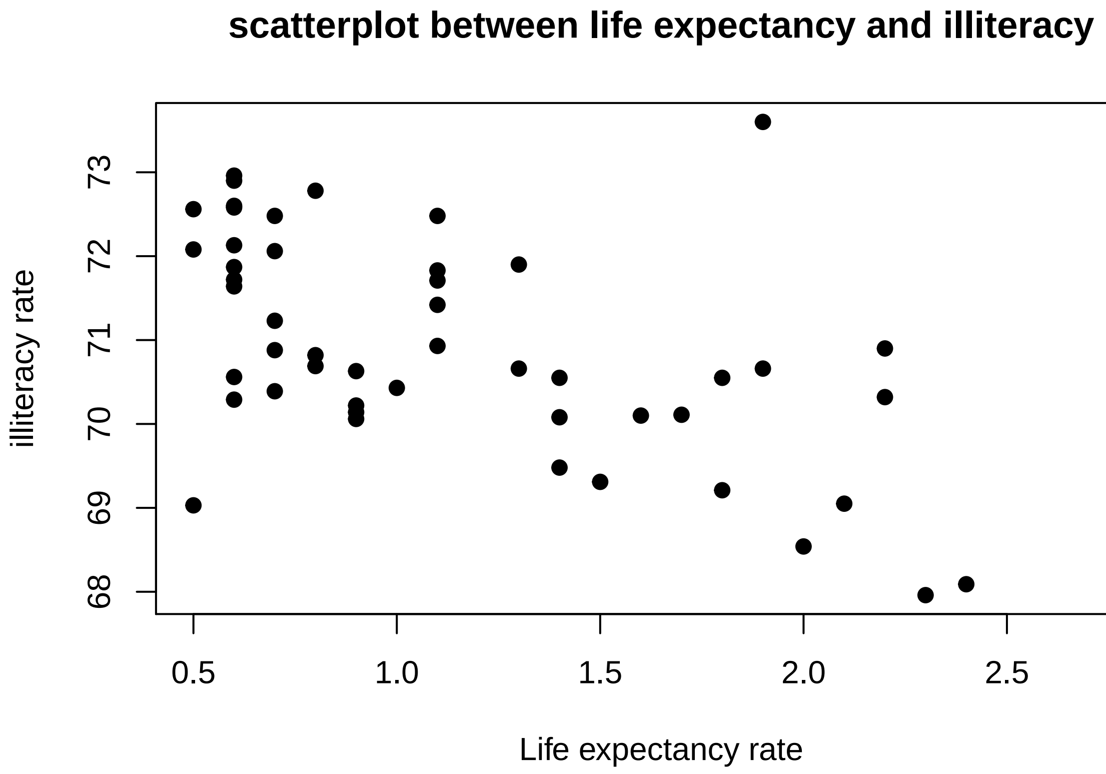
F-Statistics: Our data set shows an F-Statistics = 46.89 which is relatively large in regards to our data points. This F-statistics is greater than 1. Thus, in this case, we can reject the null hypothesis (H0: there is no relationship between murder rate and illiteracy rate)

**(d) 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 visualizations to support your exploration of this question. Discuss what you find.**

Research: Does illiteracy affect someone's life expectancy?

Previously we have seen that illiteracy negatively impacts life expectancy based on the graphs generated from executing the pairs function. In other words, a literate person will tend to live longer than someone who is illiterate. I will attempt to accept or reject this hypothesis by plotting the related data and running the coefficient values.

```r
library(ISLR) #statistical package
data("Auto")

# case 1 - scatterplot life expectancy and illiteracy
plot(stateInfo$Illiteracy, stateInfo$`Life Exp`, main="scatterplot between life expectancy and illitera
```



```r
# case 2 -
research_fit = lm(stateInfo$`Life Exp` ~ stateInfo$Illiteracy, data=Auto)
summary(research_fit)
```

```
##
## Call:
## lm(formula = stateInfo$`Life Exp` ~ stateInfo$Illiteracy, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7169 -0.8063 -0.0349  0.7674  3.6675
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           72.3949     0.3383 213.973  < 2e-16 ***
## stateInfo$Illiteracy  -1.2960     0.2570  -5.043 6.97e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.097 on 48 degrees of freedom
## Multiple R-squared:  0.3463, Adjusted R-squared:  0.3327
## F-statistic: 25.43 on 1 and 48 DF,  p-value: 6.969e-06
```

The scattered plot shows a moderate negative trend as expected. It appears that the more illiterate your are, the most likely you will have a shorter lifespan. It pays to be educated!

lm analysis:

The coefficient t-value equals to 213.973 is largely away from 1. Thus, it means a strong relationship exists between illiteracy and life expectancy.

The significance stars show that there are three stars for the intercept. It means that there is a strong possibility that a relationship exists between someone's level of literacy and life expectancy.

The F-statistics is a bit larger than one which is sufficient to conclude that we can reject the null hypothesis(H0: There is no relationship between illiteracy and life expectancy)

Conclusion: Both the scattered plot and the lm analysis show that there is enough evidence to conclude that a relationship exists between illiteracy and life expectancy. Thus, I would reject the null hypothesis and state that the more educated you are, the higher likelihood you have to live a longer life.

### Problem 3: Income and Education

The scatterplot below shows the relationship between per capita income (in thousands of dollars) and percent of population with a bachelor's degree in 3,143 counties in the US in 2010.

### (a) What are the explanatory and response variables?

Explanatory variable = "Percent with Bachelor's degree" Response variable = "Per capita income (in thousands)"

### (b) Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.

There is a clustering of data at the lower left hand corner of the graph. Thus, there is an association between the two variables, but it appears that the association is strongly shown in that lower conner. Even with the clustering data, it does show a weak positive association between the two variables.

### (c) Can we conclude that having a bachelor's degree increases one's income? Why ior why not?

No, we cannot conclude that income will increase by just having a bachelor's degree. This is because the graph shows that income barely increases for those above 50 percent with a bachelor's degree.
The data shows a strong correlation for just a few bachelor's holders, but then the relationship tends to fade as the rate of bachelor's holder's increases. We would need to see a stronger relationship or an increase in income as the percentage of bachelor's holder's increase to state any firm conclusion.

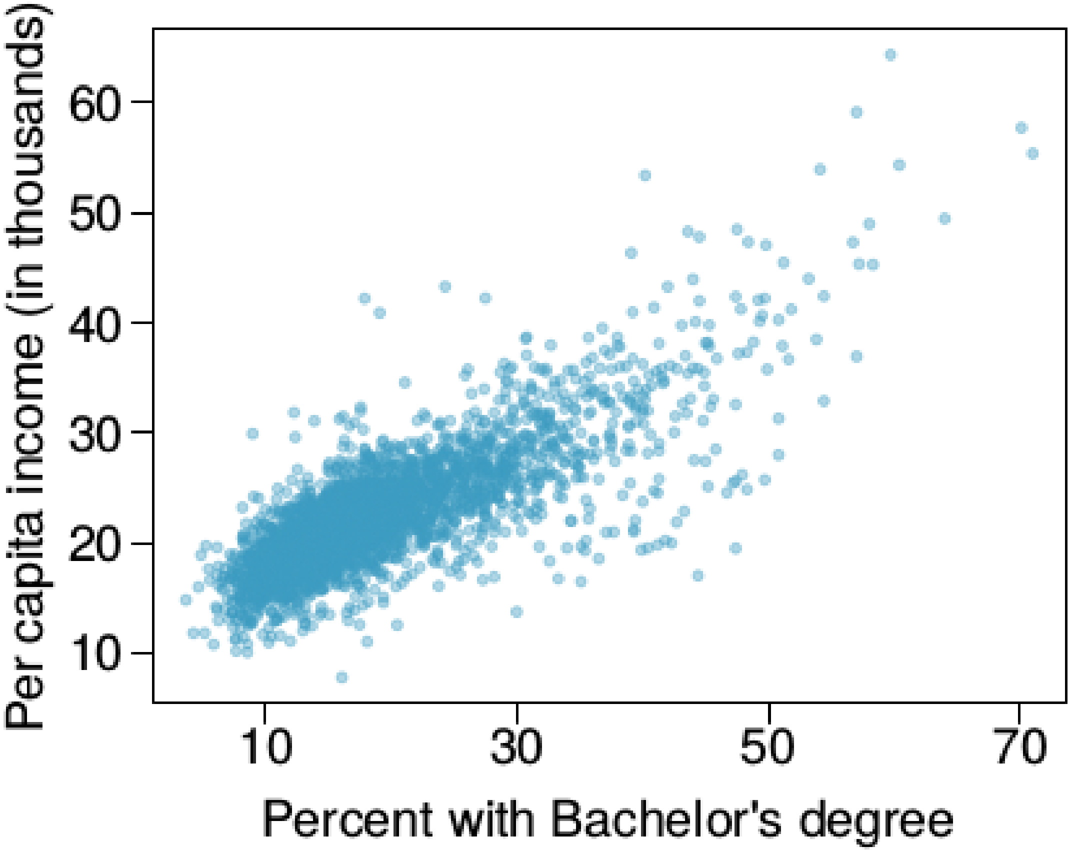


Figure 1: Income and Education in the US.