INFX 573: Problem Set 3 - Data Analysis

Shuyang Wu

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.

```
# Load standard libraries
library(tidyverse)
library(ggplot2)
library(reshape2)
library(nycflights13)
```

Problem 1: Flight Delays

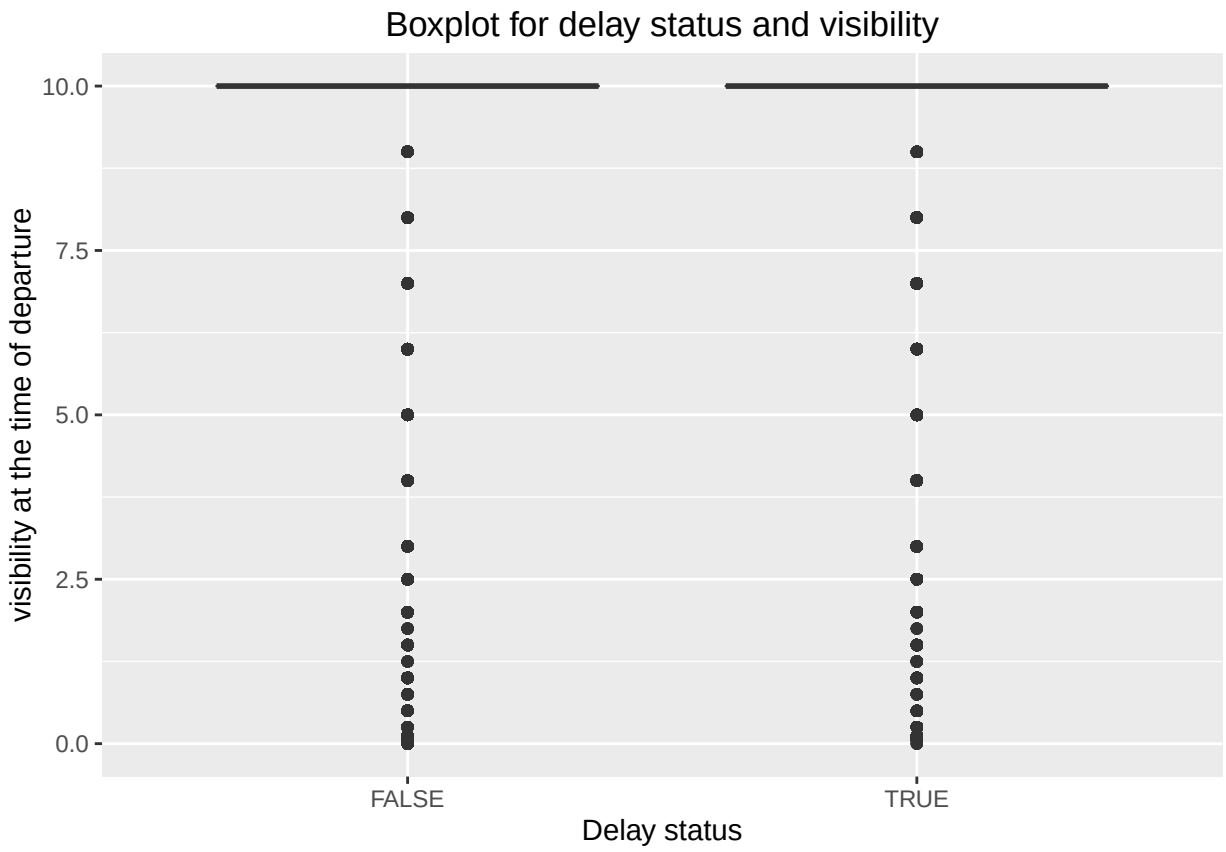
Flight delays are often linked to weather conditions. How does weather impact flights from NYC? Utilize both the `flights` and `weather` datasets from the `nycflights13` package to explore this question. Include at least two visualizations to aid in communicating what you find.

```
#recode delayed flights to be departure_delay >= 30 min
delay<- flights
delay$dep_delay[delay$dep_delay >= 30] <- "T"
delay$dep_delay[delay$dep_delay < 30] <- "F"
delay$dep_delay <- as.logical(delay$dep_delay)
prop.table(table(delay$dep_delay)) #16% flights were delayed
```

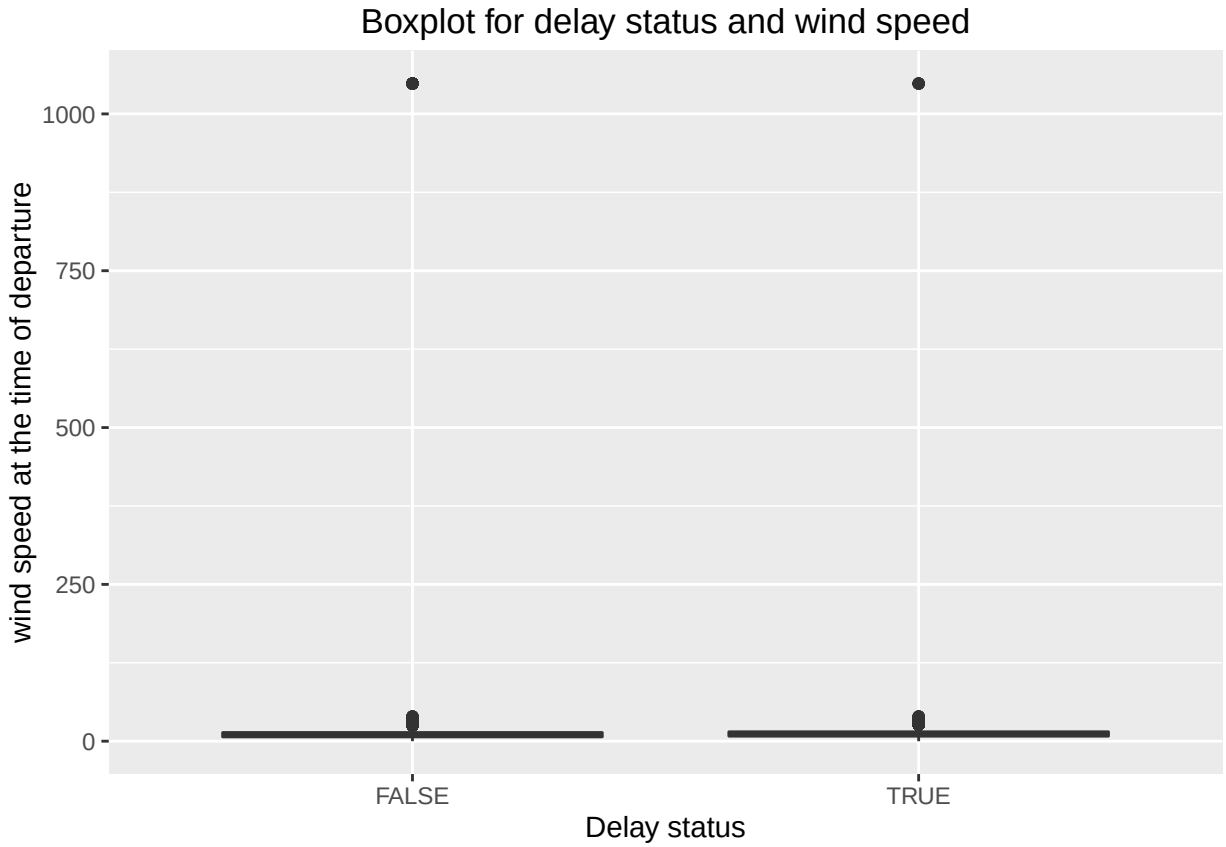
```
##
##     FALSE      TRUE
## 0.8382633 0.1617367
```

```r
#merge flights with weather based on the same time_hour and origin
merged <- merge(delay, weather, by = c("time_hour", "origin"))
merged <- na.omit(merged) # remove rows containing NAs

#relationship between visibility and delay departure delay
ggplot(data=merged, aes(x = merged$dep_delay, y = merged$visib)) +
        geom_boxplot() +
        labs(x = "Delay status", y = "visibility at the time of departure",
             title = "Boxplot for delay status and visibility")
```




```r
#relationship between wind speed and delay departure time
ggplot(data=merged, aes(x = merged$dep_delay, y = merged$wind_speed)) +
        geom_boxplot() +
        labs(x = "Delay status", y = "wind speed at the time of departure",
             title = "Boxplot for delay status and wind speed")
```



From the boxplot of visibiliy and delay status, data does not show enough evidence that visibility and delay status is associated in any way as I had hypothesized. Visibility is mosly centered around 10 with a few outliners, thus probably not a good explanatory variable to use, despite of my intuition that low visibility would cause delayed departure. From the boxplot of wind speed and delay status, there are a few outliers for the very high wind speed, the rest of the wind speeds are mostly below 50 which makes the boxplot looks very skewed. This boxplot also does not suggest any correlation between wind speed and delay status.

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.

```
state <- as.data.frame(state.x77)
str(state)#look at data type
```

```
## '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(state) #look at variables and distribution of the data
```

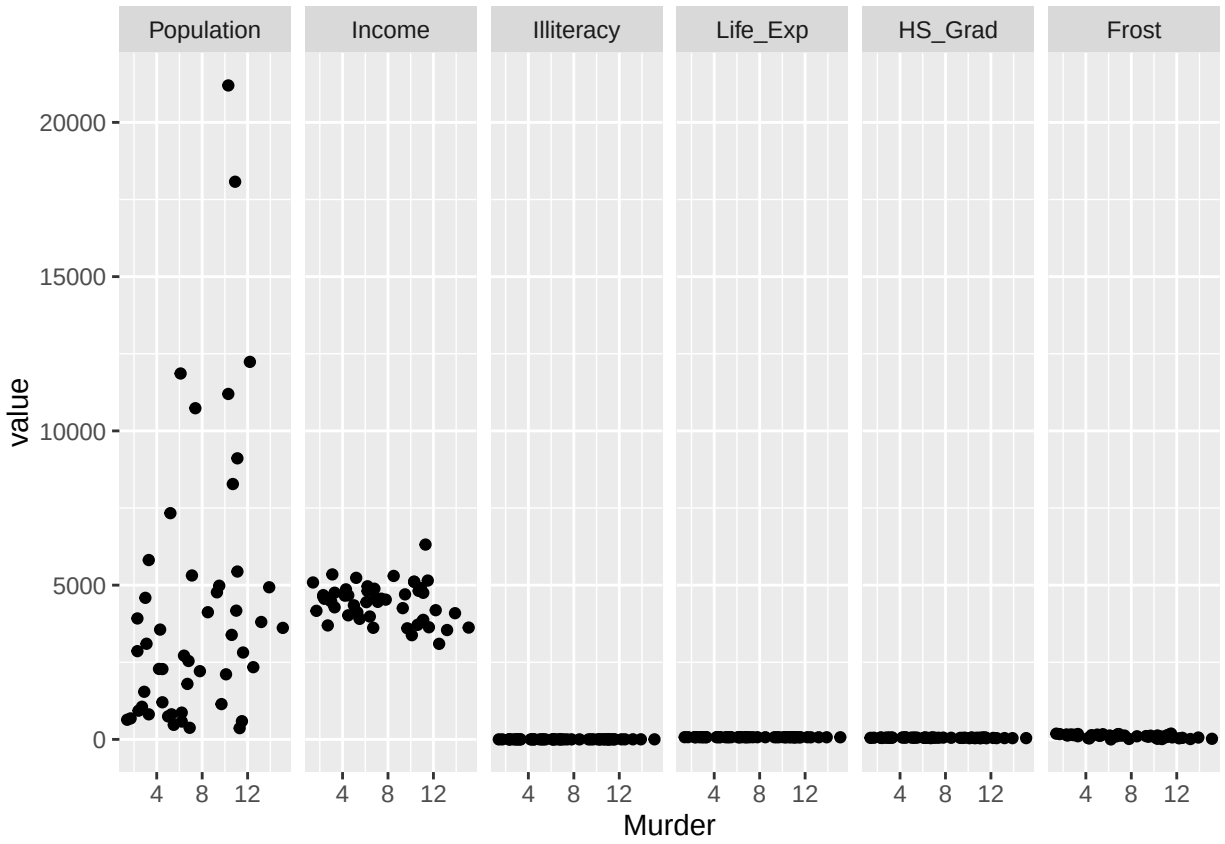
```
##    Population        Income       Illiteracy       Life Exp
##  Min.   :  365   Min.   :3098   Min.   :0.500   Min.   :67.96
##  1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12
##  Median : 2838   Median :4519   Median :0.950   Median :70.67
##  Mean   : 4246   Mean   :4436   Mean   :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.: 81162
##  Max.   :15.100   Max.   :67.30   Max.   :188.00   Max.   :566432
```

```
names(state)[names(state) == 'Life Exp'] <- 'Life_Exp' #renaming columns
names(state)[names(state) == 'HS Grad'] <- 'HS_Grad'
```

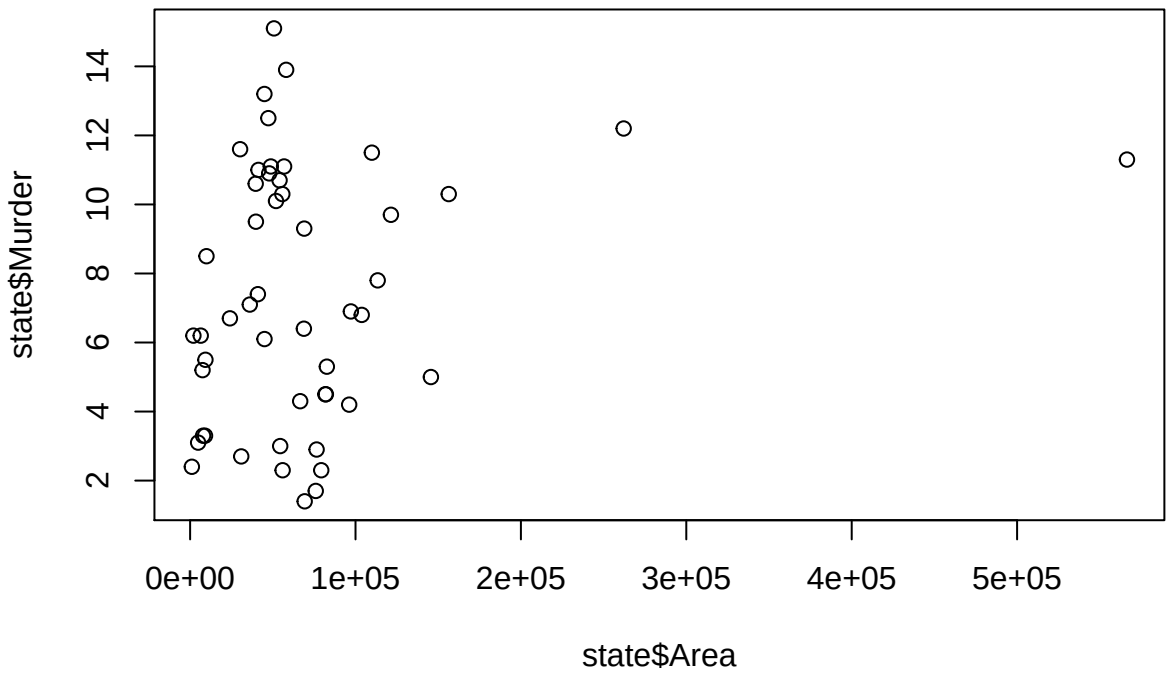
The state dataset contains statistics of the 50 states of the United States of America. Specifically, population estimate, per capita income, percent of the population with illiteracy, life expectancy in years, murder and non-negligent manslaughter rate per 100,000 population, percentage of high-school graduates, mean number of days with minimum temperature below freezing in capital or large city, and 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 examine the bivariate relationships present in the data. What does your analysis suggest might be important varibles to consider in building a model to explain variation in murder rates?

```
state.1 <- state
state.1$Area <- NULL #subset the state dataset because area has a higher mean than all other variables
df_melt <- melt(state.1,"Murder")
ggplot(df_melt,aes(Murder,value)) + geom_point() + facet_grid(.~variable) #correlation plots of Murder
```




```
plot(state$Area, state$Murder) #correlation plot of Murder and all other variable cont.d
```




```
cor(state) #correlation coefficient
```

```
##            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
```

Based on the correlation coefficients, Illiteracy and life expectancy are most correlated (positively and negatively) with Murder rates. Frost could also be a explanatory variable but the correlation is not as strong as the first two.

(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.

```r
# Linear regression model of high school graduation rates vs income
fit <- lm(Income ~ HS_Grad, data = state)
summary(fit)
```

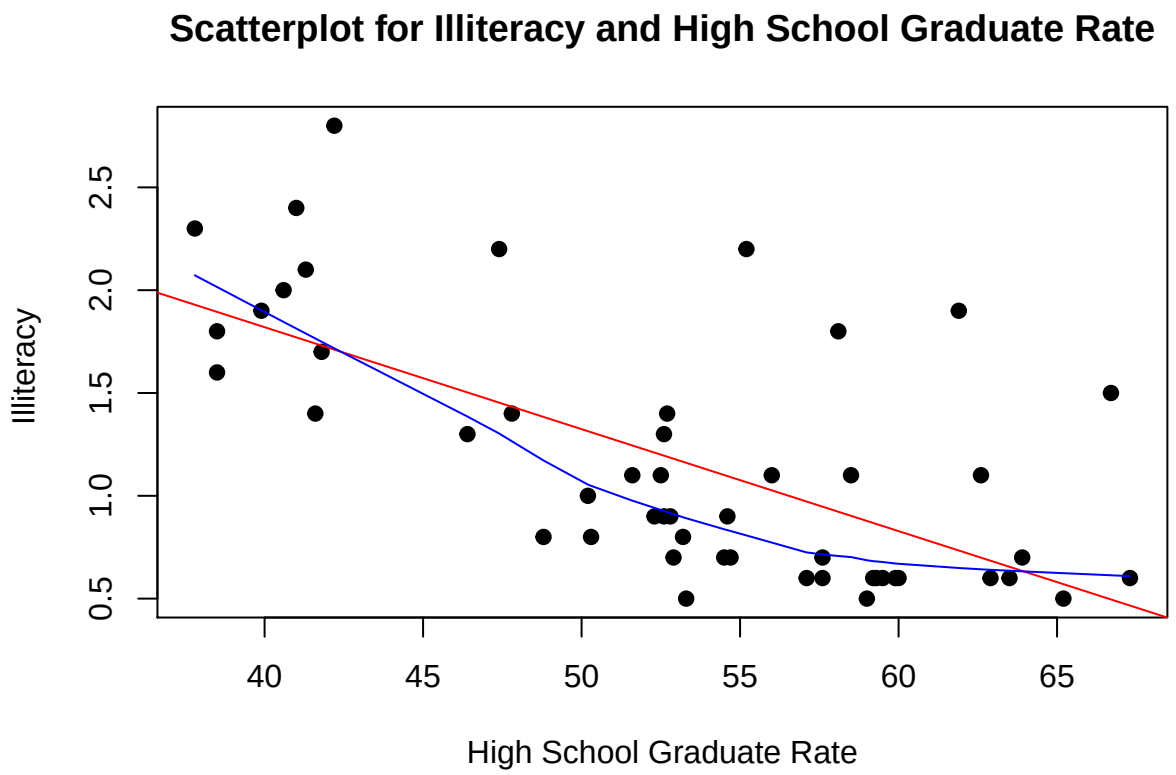
```
##
## Call:
## lm(formula = Income ~ HS_Grad, data = state)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1083.13  -277.41   -34.15   241.46  1238.17
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1931.105    462.739   4.173 0.000125 ***
## HS_Grad       47.162      8.616   5.474 1.58e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 487.1 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
```

I used high school graduate rate to explain per capita income, linear regression model fitted a linear line with the following expression: Income = 47.162 X High School Graduation Rate + 1931.105. P-value is way smaller than 0.05, thus the suggested correlation is not random.

(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 question: Hypothesis: Higher High School Graduation rate is correlated with a lower Illiteracy rate in the population.

```
attach(state)
plot(HS_Grad, Illiteracy, main="Scatterplot for Illiteracy and High School Graduate Rate",
    xlab="High School Graduate Rate", ylab="Illiteracy ", pch=19)
abline(lm(Illiteracy~HS_Grad), col="red") # add regression line
lines(lowess(HS_Grad,Illiteracy), col="blue") # add lowess line
```




```
fit1 <- lm(Illiteracy~HS_Grad)
summary(fit1)
```

```
##
## Call:
## lm(formula = Illiteracy ~ HS_Grad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6605 -0.3064 -0.1225  0.1815  1.1660
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.80389    0.44093   8.627 2.53e-11 ***
## HS_Grad     -0.04960    0.00821  -6.041 2.17e-07 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4642 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
```

The linear regression of High School Graduation rate and Illiteracy rate: Illiteracy = -0.0496 x HS_Grad + 3.80389, with a p value << 0.05. Thus we can reject the null hypothesis that the observed correlation is by chance. The negative slope of the linear regression also suggests that the two variables are negatively correlated, thus supports the hypothesis.

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?

Percent of population with a bachelor's degree is the explanatory variable, per capita income is the response variable.

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

The two variables are positivly correlated, which means a higher percent of population with a bacherlor's degree is associated with a higher per capita income.

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

We can't conclude causal relationships between the two variables. Because it is not a randomized control trial, the relationship is purely based on observation, there might be confounding variables present affecting both variables.