INFX 573: Problem Set 3 - Data Analysis

Amol Surve

Due: Monday, October 18, 2016

Collaborators: Abhishek Gupta

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(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.

Loading the Dataset

```
?flights
?weather
flightInfo<-flights
#storing flights data in flightInfo
wData<-weather
#storing weather data in wData
```

Inspecting flights dataset to get delays

```
str(flightInfo)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    336776 obs. of  19 variables:
##  $ year          : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
##  $ month         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ day           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ dep_time      : int  517 533 542 544 554 554 555 557 557 558 ...
##  $ sched_dep_time: int  515 529 540 545 600 558 600 600 600 600 ...
##  $ dep_delay     : num  2 4 2 -1 -6 -4 -5 -3 -3 -2 ...
##  $ arr_time      : int  830 850 923 1004 812 740 913 709 838 753 ...
##  $ sched_arr_time: int  819 830 850 1022 837 728 854 723 846 745 ...
##  $ arr_delay     : num  11 20 33 -18 -25 12 19 -14 -8 8 ...
##  $ carrier       : chr  "UA" "UA" "AA" "B6" ...
##  $ flight        : int  1545 1714 1141 725 461 1696 507 5708 79 301 ...
##  $ tailnum       : chr  "N14228" "N24211" "N619AA" "N804JB" ...
##  $ origin        : chr  "EWR" "LGA" "JFK" "JFK" ...
##  $ dest          : chr  "IAH" "IAH" "MIA" "BQN" ...
##  $ air_time      : num  227 227 160 183 116 150 158 53 140 138 ...
##  $ distance      : num  1400 1416 1089 1576 762 ...
##  $ hour          : num  5 5 5 5 6 5 6 6 6 6 ...
##  $ minute        : num  15 29 40 45 0 58 0 0 0 0 ...
##  $ time_hour     : POSIXct, format: "2013-01-01 05:00:00" "2013-01-01 05:00:00" ...
```

```
#Describing the dataset
flightInfo$DelaySum<-rowSums(flightInfo[,c("dep_delay","arr_delay")],na.rm=TRUE)
#creating a new column called "DelaySum" where
#you can see the total delay as the addition of
#arrival as well as departure delays
```

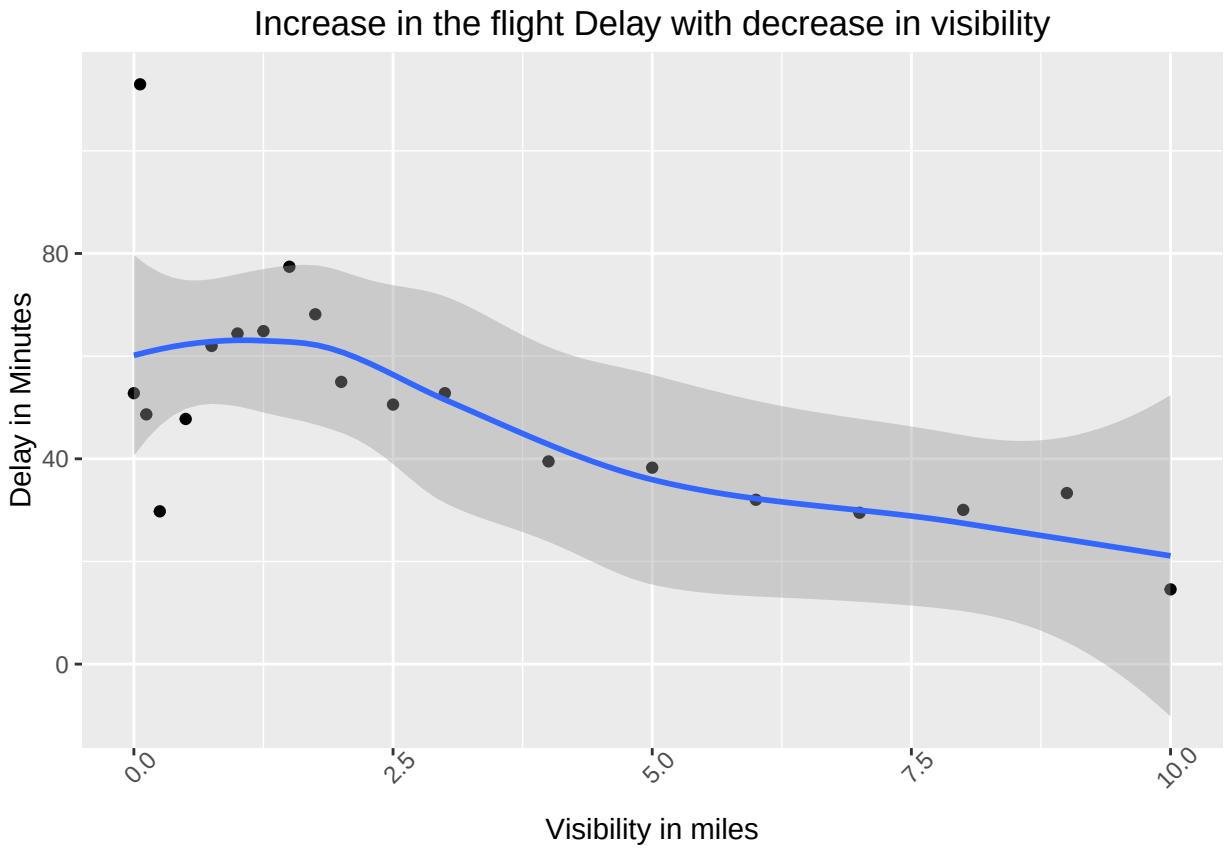
In order to inspect how weather conditions impact flights delays rom NYC, I first created the column called DelaySum which will store the total of the arrival as well as the departure delay in order to examine the effect of weather conditions on this delay.

```
flightInfo<-merge(flightInfo,wData,by="time_hour")
#merging with the weather dataset by "time_hour variable"
```

Now, in order to link flight dataset with weather dataset, the common variable in both the datasets is "time_hour", which is therefore used to merge. The new dataset is called flightInfo with total 34 variables and 1006717 observations.

1. Increase in Visibility Decreases the flight's Delay

```
flightInfo %>%
  dplyr::group_by(visib) %>%
   #grouping based on visibility
   dplyr::summarise(Delay = mean(DelaySum)) %>%
  # Accouting for Mean delay
  ggplot(aes(x = visib, y = Delay)) + geom_point() +
  geom_smooth()+
  theme(axis.text.x = element_text(angle = 45))+
  ggtitle("Increase in the flight Delay with decrease in visibility")+
  labs(x="Visibility in miles",y="Delay in Minutes")
```

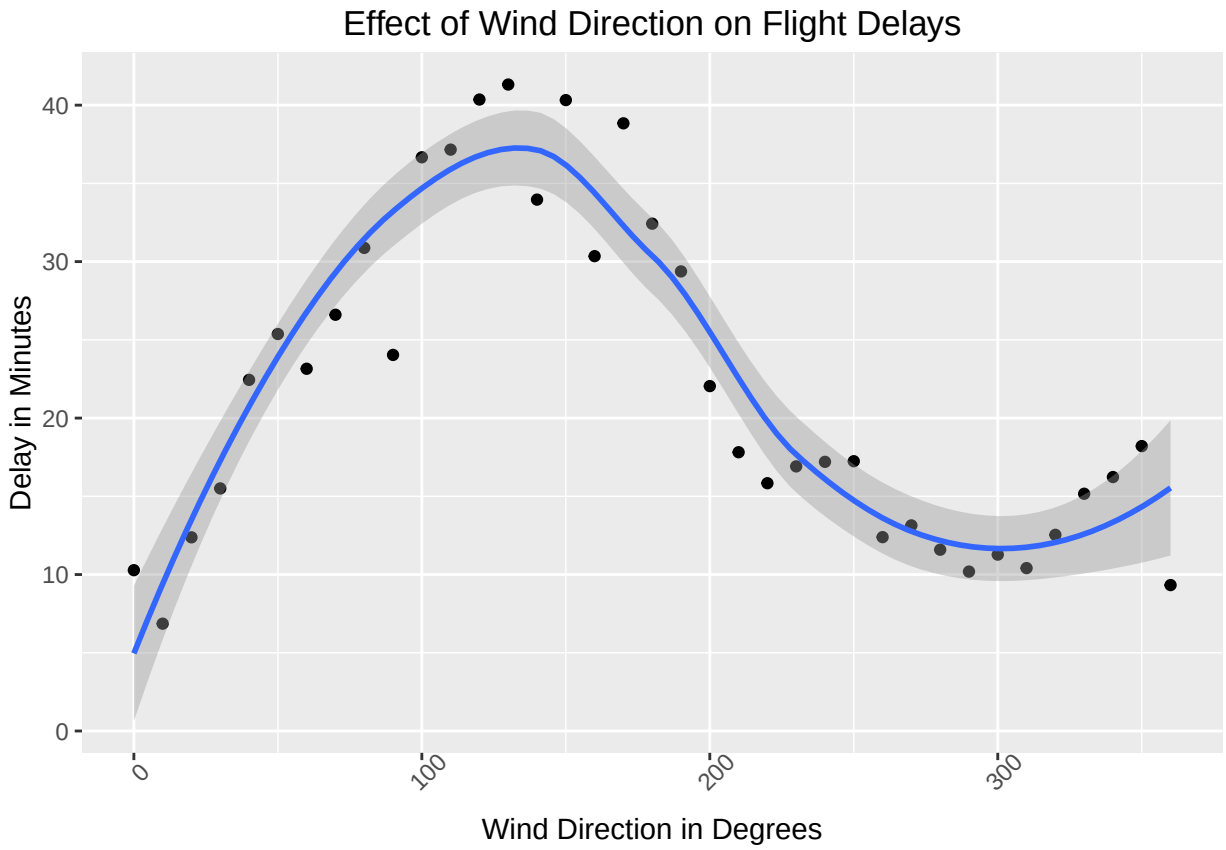



```
#plotting the mean delay against visibility
```

As you can see in the plot, with the increase in the visibility, the delay is decreasing.

2. Effect of wind direction

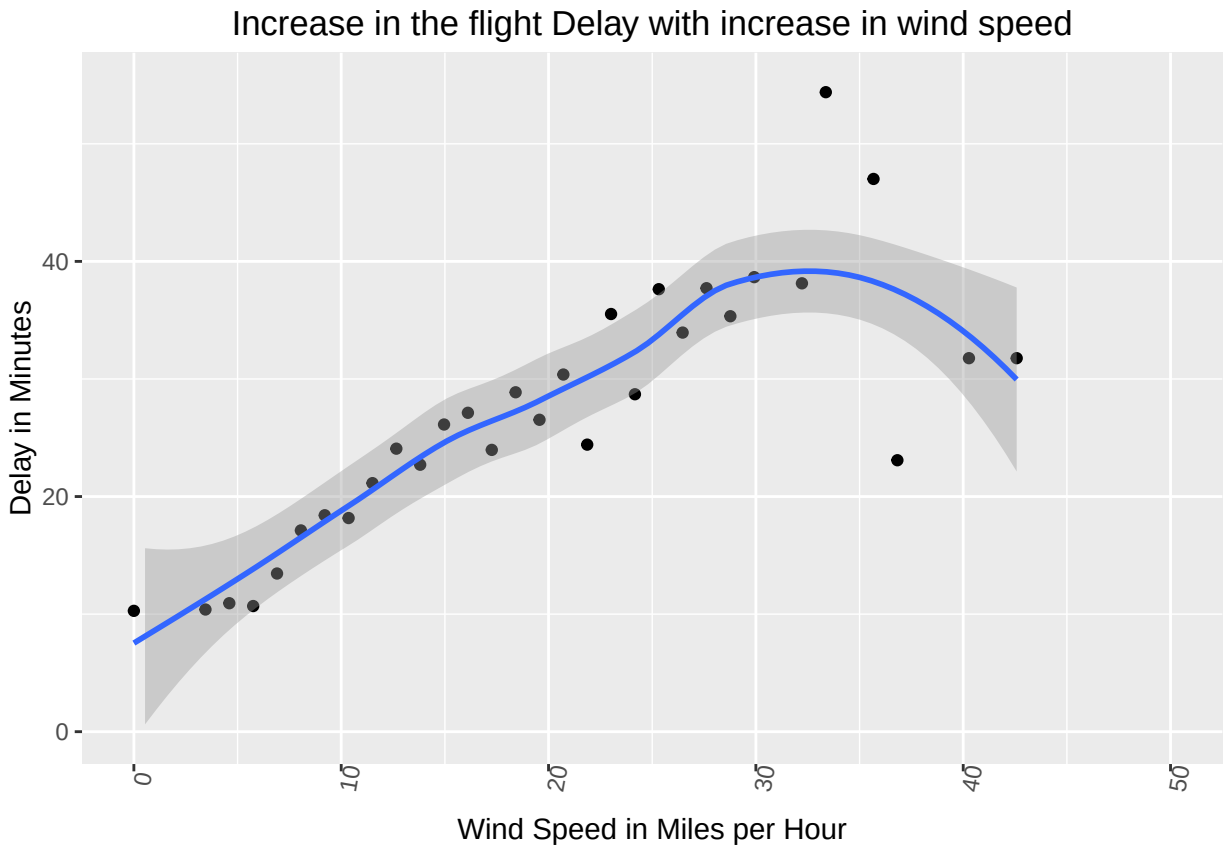
```
flightInfo %>%
  dplyr::group_by(wind_dir) %>%
   #grouping based on wind direction
   dplyr::summarise(Delay = mean(DelaySum)) %>%
  # Accouting for Mean delay
  ggplot(aes(x = wind_dir, y = Delay)) + geom_point() + geom_smooth()+
  theme(axis.text.x = element_text(angle = 45))+
  ggtitle("Effect of Wind Direction on Flight Delays")+
  labs(x="Wind Direction in Degrees",y="Delay in Minutes")
```



There is an interesting pattern in the relationship between flight delays and the wind direction. As you can see, the delay increases when the direction is approaching from zero to approximately 150 degrees but beyond that, for the next cycle of 150 degrees i.e. till 300 degrees, the delay is decreasing and then beyond 300 degrees it is again increasing.

3. Effect of wind speed

```
flightInfo %>%
  dplyr::group_by(wind_speed) %>%
   #grouping based on wind_speed
   dplyr::summarise(Delay = mean(DelaySum)) %>%
  # Accouting for Mean delay
  ggplot(aes(x = wind_speed, y = Delay))+geom_point()+xlim(0,50)+
  geom_smooth()+theme(axis.text.x = element_text(angle = 80))+
  ylim(0,55)+
  ggtitle("Increase in the flight Delay with increase in wind speed")+
  labs(x="Wind Speed in Miles per Hour",y="Delay in Minutes")
```




```
#plotting the mean delay against wind speed
```

As you can see, the increase in wind_speed increases the delay. For better visibility of the results by removing the outliers, I have restricted the x and y range.

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.

```
StateInfo<-data.frame(state.x77)
#converting into StateInfo Dataframe
str(StateInfo)
```

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

```
#Describing the dataframe
StateDetails <- cbind(States = rownames(StateInfo), StateInfo)
#Adding First Column as States
rownames(StateDetails) <- NULL
#removing the rownames since we have State name column
```

1. State - Data sets related to the 50 states of United States describing the facts and figures of the each state. Following data sets are contained: state.abb: character vector of 2-letter abbreviations for the state names.

state.area: numeric vector of state areas (in square miles).

state.center: list with components named x and y giving the approximate geographic center of each state in negative longitude and latitude. Alaska and Hawaii are placed just off the West Coast.

state.division: factor giving state divisions (New England, Middle Atlantic, South Atlantic, East South Central, West South Central, East North Central, West North Central, Mountain, and Pacific).

state.name: character vector giving the full state names.

state.region: factor giving the region (Northeast, South, North Central, West) that each state belongs to.

state.x77: matrix with 50 rows and 8 columns giving the following statistics in the respective columns.

Out of the above information, the matrix state.x77 is of my interest since it gives insights about Population count, Income, Illiteracy, Life Expectancy, Murder rate, HS grad, Frost and Land Area.

2. Data Tidying:
3. Since, it is a matrix, I have first converted it into the StateInfo Dataframe.
4. Also, each row is being referenced using the state name. Hence, for tidying purpose, I have added a column in the beginning called states that takes all fifty state's names and put it under the first column.
5. Hence, the final data frmae is stored as StateDetails is with
   9 variables and 50 observations
6. Variables:
7. States: Factor. %0 levels for 50 states in United states.
8. Population: num. Gives the population Estimate as of July 1, 1975
9. Income: num. Per capita income (1974)
10. Illiteracy: num. Percentage of the illiterate population
11. Life.Exp: num. Life expectancy in years (1969–71)
12. Murder: num. Murder and non-negligent manslaughter rate per 100,000 population (1976)
13. HS.Grad: num. Percent high-school graduates (1970)
14. Frost: num. Mean number of days with minimum temperature below freezing (1931–1960) in capital or large city
15. 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 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?

```
cor(StateInfo)
```

```
##             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
```

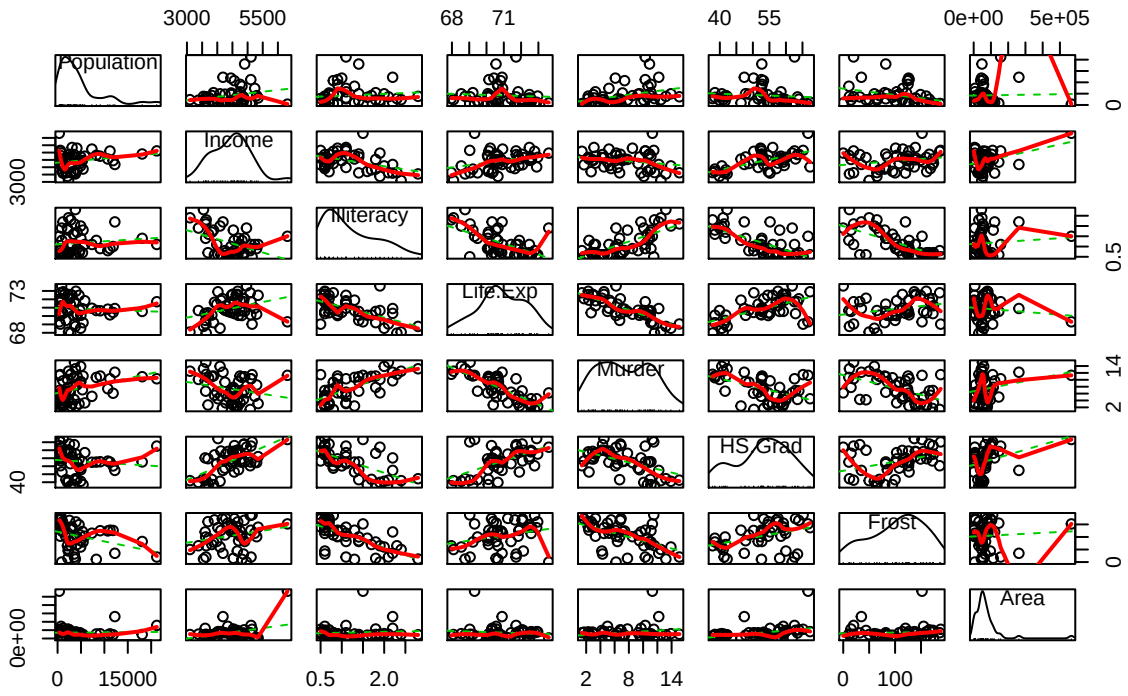
```
# Using Correlation function to find out
#the correlation between different variables in
#StatesInfo Dataset

library(car)
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
#for Scatter Plot Matrix
scatterplotMatrix(StateInfo,spread=FALSE,lty=2,main="Scatter Plot Matrix")
```




```
#Scatter plot matrix for every pair of variables
```

Here, we got the correlation matrix using cor() and we want to explore the relationships between different variables of the dataset.

Based on the Correlation table, the Murder rate is most closely related to the Life Expectancy variable since it shows the highest correlation with the negative value of (-0.7808458). However, on the positive correlation side, it is most closely correlated with Illiteracy. Life Expectancy and Murder rate would be logically less sound compared to Illiteracy vs Murder Rate to identify how illiteracy actually affects the murder rate. Whether educational background of the people impacts the mentally and criminal instincts which would lead us to valuable insights about the data.

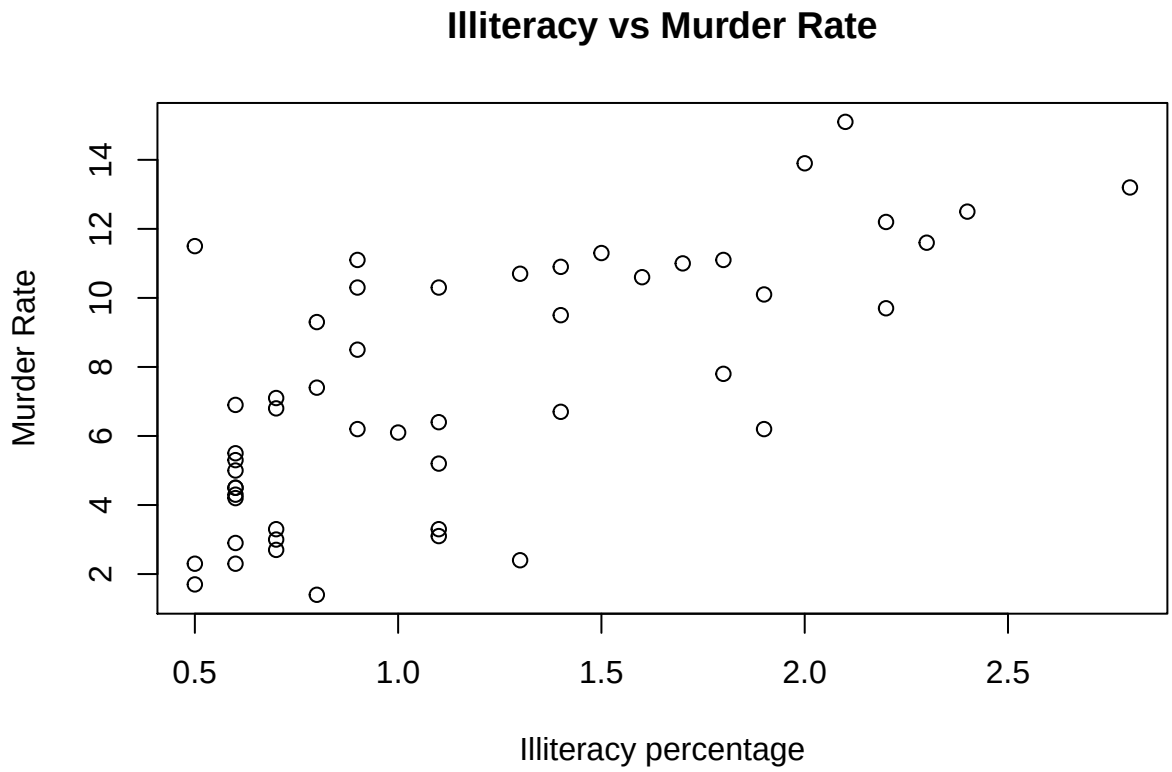
(c) Choose one variable and fit a simple linear regression model, $Y = \beta_1 X + \beta_0$, using the `lm()` function in R. Describe your results.

```
LinearR<-lm(formula=Murder~Illiteracy,data = StateInfo)
#Simple Linear Regression with Murder as the response variable
#and  Illiteracy as the Explanatory variable
summary(LinearR)
```

```
##
## Call:
## lm(formula = Murder ~ Illiteracy, data = StateInfo)
##
## 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 **
## 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
```

```
#finding the p value
plot(StateDetails$Illiteracy,
     StateDetails$Murder,xlab="Illiteracy percentage",
     ylab="Murder Rate",main="Illiteracy vs Murder Rate")
```



Here, murder is the response variable and Illiteracy is the explanatory variable. The p value is less than 0.05 (1.26e-08) which shows that there is a correlation between variables Murder and Illiteracy, concluding the statistical significance in the relationship.

(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.

```
cor(StateInfo)
```

```
##             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
```

```
#Finding the correlation between each pair of variables
LinearR2<-lm(formula=Income~HS.Grad,data = StateInfo)
#Simple Linear Regression with Income as the response
#variable and HS>Grad as the Explanatory variable
summary(LinearR2)
```

```
##
## Call:
## lm(formula = Income ~ HS.Grad, data = StateInfo)
##
## 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
```

```
#finding the p value
```

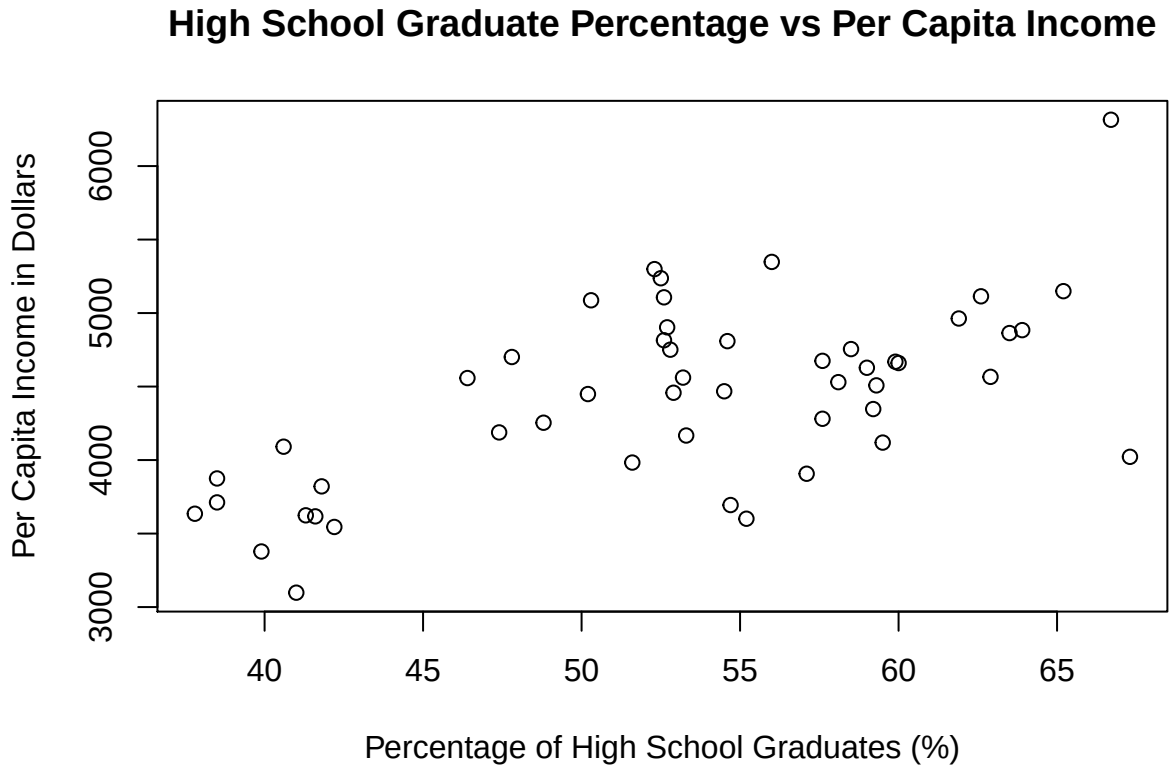
The p value is less than 0.05 (1.58e-06) shows that the relationship between Income and High School Graduate Percentage is statistically significant.

Research Question: How the per capita income changes as the percentage of the High School Graduate changes?

Reason for selecting the question is to identify how education affected the per capita income of the people. Wehther the convention that high education yields more income can be apply here or not, is the reason for selecting this research question.

Plotting the Per Capita Income vs High School Graduate Percentage

```
plot(StateDetails$HS.Grad,StateDetails$Income,
     xlab="Percentage of High School Graduates (%)",
     ylab="Per Capita Income in Dollars",
     main="High School Graduate Percentage vs Per Capita Income")
```



As you can see in the plot, as the percentage of High School Graduates increases, there is increase in the income as well. Thus, showing the significant impact of education on income.

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. The scatterplot given shows the Percentage of People with Bachelors Degree and the Per Capita income in thousands.

(a) What are the explanatory and response variables?

Explanatory Variable: Percent with Bachelors degree is the explanatory variable here since more number of people with bachelors degree would increase the per capita income. But, the inverse is not true.

Response Variable: Per Capita income in thousands is the Response variable since it depends on the change in the value of the Percentge of the people with Bachelor's Degree. It is directly proportional to the change in the explanatory variable.

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

Based on the scatterplot's visualization, we can see the strong positive correlation between the variables Percentage with Bachelors Degree and Per Capita Income. The relation is strong especially between the window of 10 to 30 percent approximately.

The unusual observation accoridng to me is beyond the 30-35 percentage value of the explanatory variable where the observations are not strongly correlated and scattering all over the graph. For example, you can see the high per capita income value for lower percentage of bacheor's degree and vice versa. There are several outliers as well in this range.

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

Correlation is not equal to the causation. In addition to that, we cannot calculate the p value since the data in insufficient. Hence, we cannot determine whether the percentage change in the Bachelors Degree would affect the per capita income in a statistically significant way.