IMT 573: Problem Set 4 - Data Analysis

Miloni A Desai

Due: Tuesday, October 21, 2019

Collaborators:

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.

```
# Load standard libraries
library(tidyverse)
library(gridExtra)
library(psych)
```

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.

```
data(state)
```

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

```
df<-as.data.frame.matrix(state.x77)
#View(df)

describe(df)
```

```
##            vars  n     mean       sd   median  trimmed      mad     min
## Population    1 50  4246.42  4464.49  2838.50  3384.28  2890.33  365.00
## Income        2 50  4435.80   614.47  4519.00  4430.07   581.18 3098.00
## Illiteracy    3 50     1.17     0.61     0.95     1.10     0.52    0.50
## Life Exp      4 50    70.88     1.34    70.67    70.92     1.54   67.96
## Murder        5 50     7.38     3.69     6.85     7.30     5.19    1.40
## HS Grad       6 50    53.11     8.08    53.25    53.34     8.60   37.80
## Frost         7 50   104.46    51.98   114.50   106.80    53.37    0.00
## Area          8 50 70735.88 85327.30 54277.00 56575.72 35144.29 1049.00
##                 max     range  skew kurtosis       se
## Population  21198.0  20833.00  1.92     3.75   631.37
## Income       6315.0   3217.00  0.20     0.24    86.90
## Illiteracy      2.8      2.30  0.82    -0.47     0.09
## Life Exp       73.6      5.64 -0.15    -0.67     0.19
## Murder         15.1     13.70  0.13    -1.21     0.52
## HS Grad        67.3     29.50 -0.32    -0.88     1.14
## Frost         188.0    188.00 -0.37    -0.94     7.35
## Area       566432.0 565383.00  4.10    20.39 12067.10
```

```
summary(df)
```

```
##    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
```

```
head(df)
```

```
##            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
```

```
tail(df)
```

```
##               Population Income Illiteracy Life Exp Murder HS Grad Frost
## Vermont              472   3907        0.6    71.64    5.5    57.1   168
## Virginia            4981   4701        1.4    70.08    9.5    47.8    85
## Washington          3559   4864        0.6    71.72    4.3    63.5    32
## West Virginia       1799   3617        1.4    69.48    6.7    41.6   100
## Wisconsin           4589   4468        0.7    72.48    3.0    54.5   149
## Wyoming              376   4566        0.6    70.29    6.9    62.9   173
##                 Area
## Vermont         9267
## Virginia       39780
## Washington     66570
## West Virginia  24070
## Wisconsin      54464
## Wyoming        97203
```

```
dim(df)
```

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

#This dataset has data about the 50 states in the #United States like the population in each state as of July 1st 1975, income per capita, murder rate per 100,000 population, percent high school grads, illiteracy percent of population, life expectancy in years, mean number of days with min temp below freezing, land area in square miles. Most of the data gives us information about the different characteristics of the population, segmented by state

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

```
cor(df$Murder,df$Illiteracy)
```

```
## [1] 0.7029752
```

```
cor(df$Murder,df$Population)
```

```
## [1] 0.3436428
```

```
cor(df$Murder,df$Income)
```

```
## [1] -0.2300776
```

```
cor(df$Murder,df$`Life Exp`)
```

```
## [1] -0.7808458
```

```
cor(df$Murder, df$`HS Grad`)
```

```
## [1] -0.487971
```

```
cor(df$Murder,df$Frost)
```

```
## [1] -0.5388834
```

```
cor(df$Murder,df$Area)
```

```
## [1] 0.2283902
```

#From the correlation between the bivariate relationships, we can see that illiteracy has the most direct linear relationship with the murder rate. The higher the illiteracy percent in a state the higher is the murder rate. #Another indirect relationship that can be seen and holds true in all logic too is that of the murder rate and the life expectancy. They are negatively correlated. Thus we can see that higher the murder rate, lower is the life expectancy. Another strong negative correlation can be seen between the number of days with below freezing temps and murder rate. Thus those states which have more cold days in a year tend to have lower murder rates as well. #From our analysis we can conclude that the illiteracy percent in a state is the most important variable when considering building a model to better understand murder rates.

(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.

```
model<-lm(Murder~Illiteracy, data=df)
#View(model)
summary(model)
```

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

#From our analysis we can see that residuals give us the difference between the actual observed response values and the response values that the model predicted.We look for them to have a certain symmetry in these values.The coefficients tell us that the average state has a murder rate of 2.3968 and that with an increase an of one percent in illiteracy rate, the murder rate goes up 4.2575 percent. #The coefficient Standard Error measures the average amount that the coefficient estimates vary from the actual average value of our response variable. We'd ideally want a lower number relative to its coefficients. Here our standard error is 0.617, which means that the murder rate could vary by this margin if we run the model again. Thus we would want this number to be very small. #Three stars represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis which allows us to conclude

that there is a relationship between murder rate and illiteracy in a state. #The multiple R squared value tells us how well the model is fitting the data. In our case it tells us that 49.42% of the variance in murder rate can be explained by illiteracy rate in a state alone.

(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(df$Income,df$`HS Grad`)
```

```
## [1] 0.6199323
```

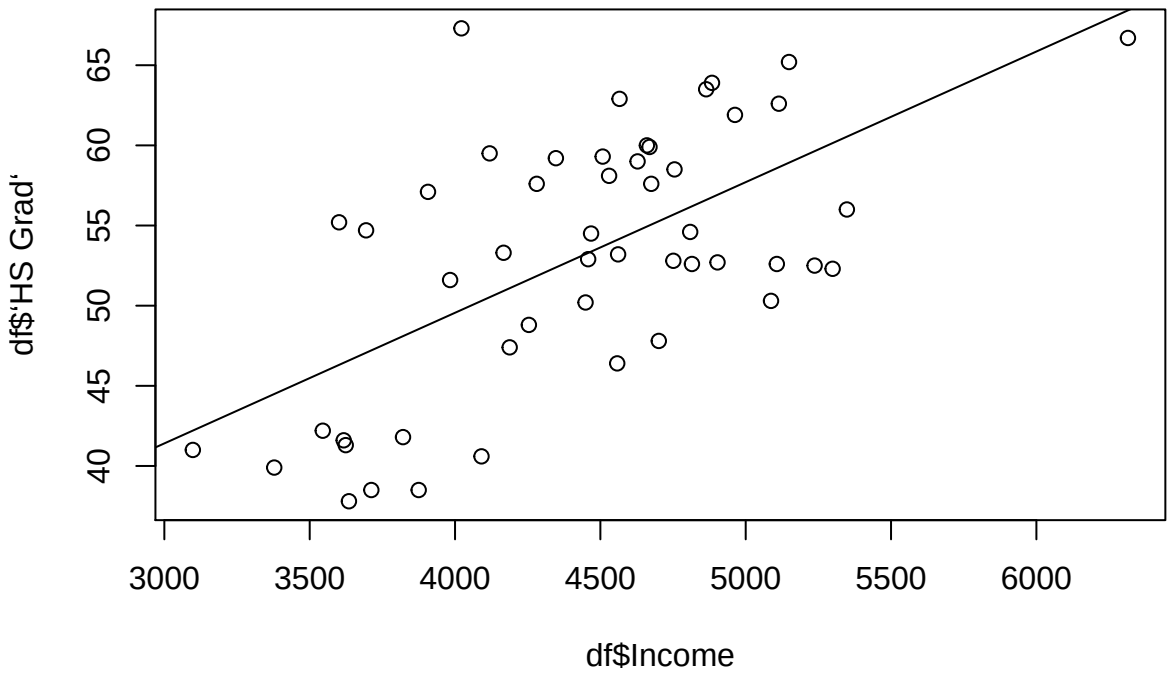
```
cor(df$Income,df$Illiteracy)
```

```
## [1] -0.4370752
```

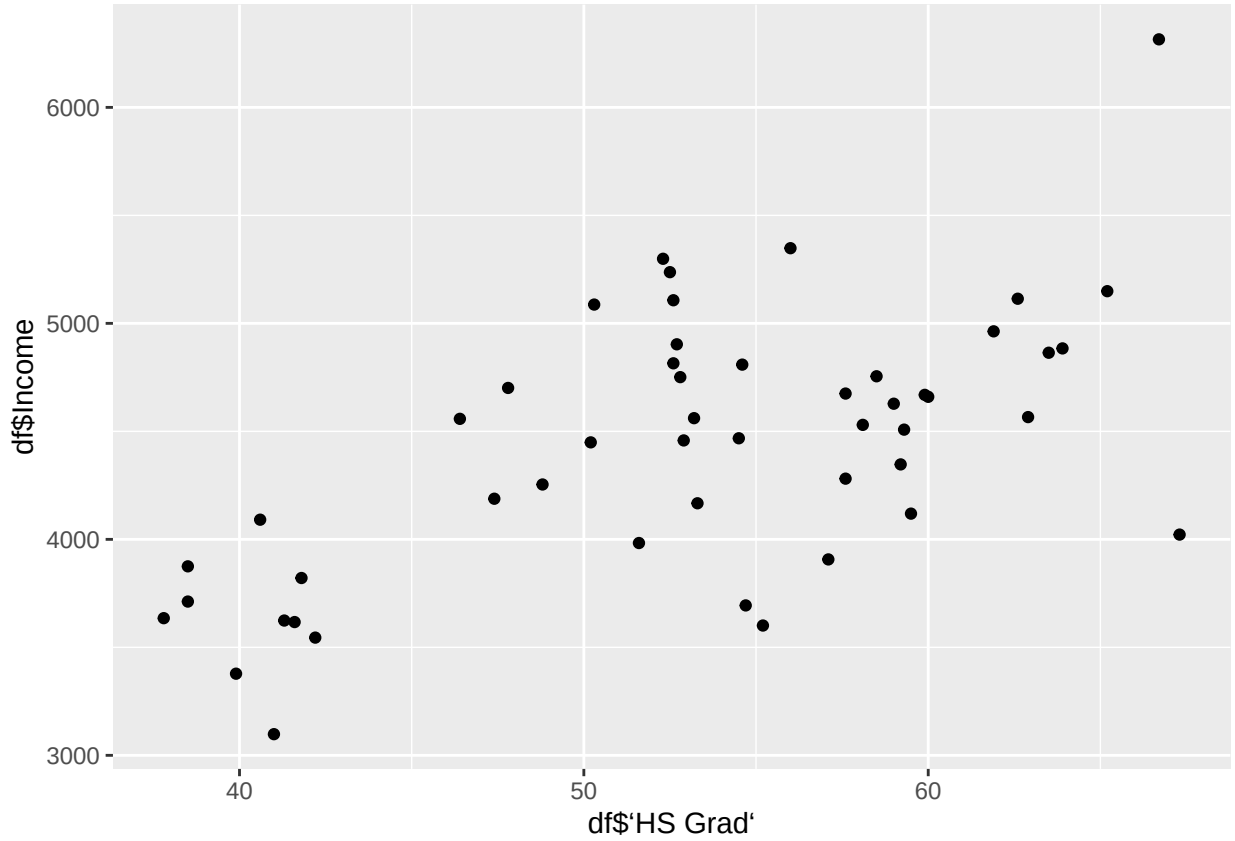
```
model2<-lm(df$`HS Grad`~ df$Income, data=df)
summary(model2)
```

```
##
## Call:
## lm(formula = df$`HS Grad` ~ df$Income, data = df)
##
## 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 *
## df$Income    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
```

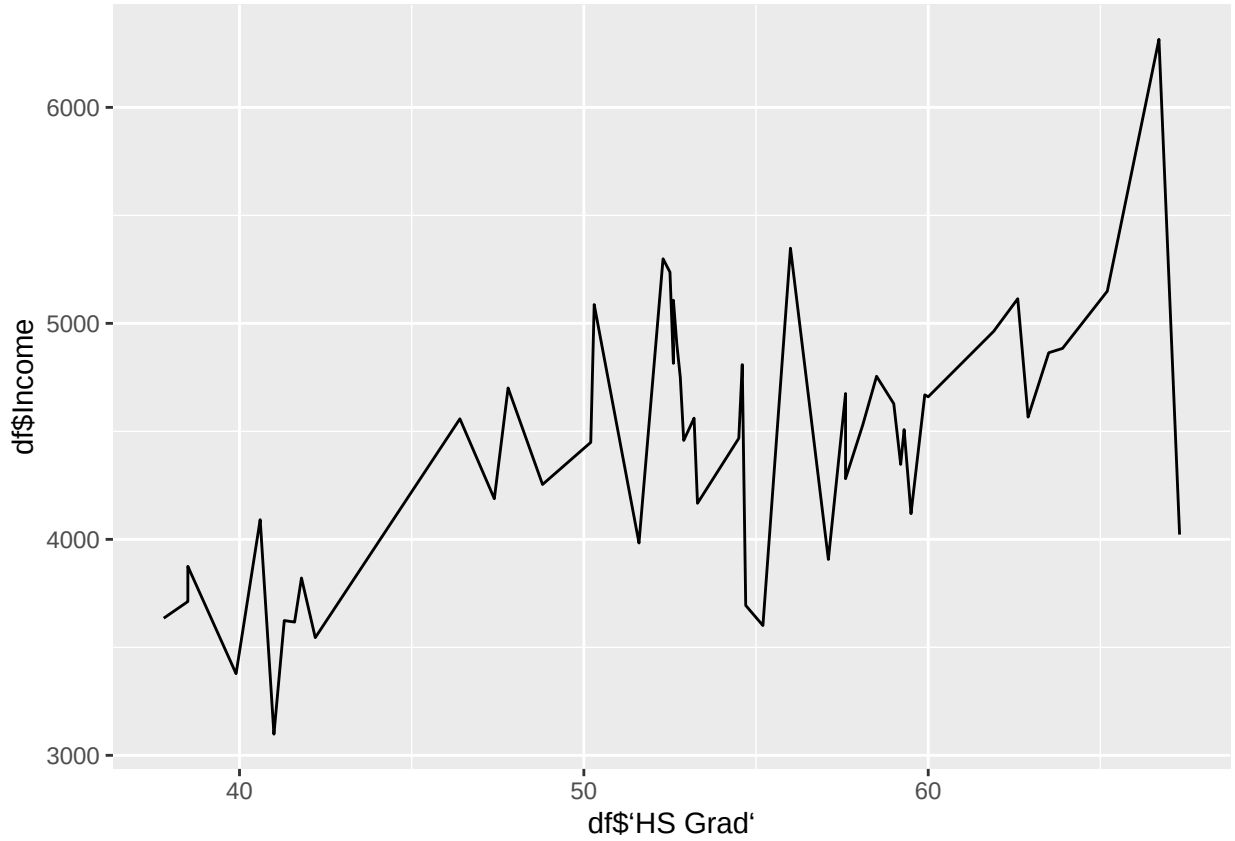
```
plot(df$`HS Grad` ~ df$Income, data = df)
abline(model2)
```




```
ggplot(data = df, aes(y=df$Income, x=df$`HS Grad`)) + geom_point()
```




```
ggplot(df, aes(df$‘HS Grad‘)) +
  geom_line(aes(y = df$Income))
```



#The data question that we can answer from this dataset is if there is a higher income per capita in those states that have a higher percentage of high school graduates. To examine this question, we will first see of these two variables are correlated.Once we find that Income is most closely correlated to HS grad, we further run a linear regression model using these two variables and visualize the same to see their relationship. It is clear from the graphs that as the percentage of high school grads in a state increases, the income is also likely to increase. There is also a strong negative correlation between illiteracy and income which helps us further validate the relationship between education and income in general.

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.

(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 crime dataset you compiled in Problem Set 3.

```
Beats_Dataset <- read_csv("imt573/Data/police_beat_and_precinct_centerpoints.csv")
```

```
## Parsed with column specification:
## cols(
##   Name = col_character(),
##   `Location 1` = col_character(),
##   Latitude = col_double(),
##   Longitude = col_double()
```

```
## )
```

```
Beats_Dataset['Census_Tract']<- 0
crime_data <- read_csv("imt573/Data/crime_data.csv")
```

```
## Parsed with column specification:
## cols(
##   `Report Number` = col_double(),
##   `Occurred Date` = col_character(),
##   `Occurred Time` = col_double(),
##   `Reported Date` = col_character(),
##   `Reported Time` = col_double(),
##   `Crime Subcategory` = col_character(),
##   `Primary Offense Description` = col_character(),
##   Precinct = col_character(),
##   Sector = col_character(),
##   Beat = col_character(),
##   Neighborhood = col_character()
## )
```

```
census_data <- read_csv("imt573/Data/census_edu_data.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   GEO.id = col_character(),
##   `GEO.display-label` = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
Beats_Dataset_Final<-extract(Beats_Dataset, Census_Tract, into = c("11_digit_code"),  "(.{11})", remove=
Beats_Census<- merge(Beats_Dataset_Final, census_data, by.x = "11_digit_code",by.y = "GEO.id2")
Crime_Beats_Census<- merge(crime_data, Beats_Census, by.x = "Beat",by.y = "Name")
dim(Crime_Beats_Census)
```

```
## [1]  0 43
```

```
summary(Crime_Beats_Census)
```

```
##      Beat           Report Number Occurred Date      Occurred Time
##  Length:0           Min.   : NA   Length:0           Min.   : NA
##  Class :character   1st Qu.: NA   Class :character   1st Qu.: NA
##  Mode  :character   Median : NA   Mode  :character   Median : NA
##                     Mean   :NaN                      Mean   :NaN
##                     3rd Qu.: NA                      3rd Qu.: NA
##                     Max.   : NA                      Max.   : NA
##  Reported Date      Reported Time Crime Subcategory
##  Length:0           Min.   : NA   Length:0
##  Class :character   1st Qu.: NA   Class :character
##  Mode  :character   Median : NA   Mode  :character
##                     Mean   :NaN
##                     3rd Qu.: NA
##                     Max.   : NA
##  Primary Offense Description   Precinct            Sector
##  Length:0                    Length:0           Length:0
##  Class :character            Class :character   Class :character
```

```
##  Mode  :character            Mode  :character   Mode  :character
##
##
##
##  Neighborhood       11_digit_code       Location 1           Latitude
##  Length:0           Length:0           Length:0           Min.   : NA
##  Class :character   Class :character   Class :character   1st Qu.: NA
##  Mode  :character   Mode  :character   Mode  :character   Median : NA
##                                                           Mean   :NaN
##                                                           3rd Qu.: NA
##                                                           Max.   : NA
##    Longitude    Census_Tract    GEO.id          GEO.display-label
##  Min.   : NA   Min.   : NA   Length:0           Length:0
##  1st Qu.: NA   1st Qu.: NA   Class :character   Class :character
##  Median : NA   Median : NA   Mode  :character   Mode  :character
##  Mean   :NaN   Mean   :NaN
##  3rd Qu.: NA   3rd Qu.: NA
##  Max.   : NA   Max.   : NA
##      total      no_schooling nursery_school  kindergarten   1st_grade
##  Min.   : NA   Min.   : NA   Min.   : NA    Min.   : NA   Min.   : NA
##  1st Qu.: NA   1st Qu.: NA   1st Qu.: NA    1st Qu.: NA   1st Qu.: NA
##  Median : NA   Median : NA   Median : NA    Median : NA   Median : NA
##  Mean   :NaN   Mean   :NaN   Mean   :NaN    Mean   :NaN   Mean   :NaN
##  3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA    3rd Qu.: NA   3rd Qu.: NA
##  Max.   : NA   Max.   : NA   Max.   : NA    Max.   : NA   Max.   : NA
##    2nd_grade     3rd_grade     4th_grade     5th_grade     6th_grade
##  Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA
##  1st Qu.: NA   1st Qu.: NA   1st Qu.: NA   1st Qu.: NA   1st Qu.: NA
##  Median : NA   Median : NA   Median : NA   Median : NA   Median : NA
##  Mean   :NaN   Mean   :NaN   Mean   :NaN   Mean   :NaN   Mean   :NaN
##  3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA
##  Max.   : NA   Max.   : NA   Max.   : NA   Max.   : NA   Max.   : NA
##    7th_grade     8th_grade     9th_grade     10th_grade    11th_grade
##  Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA
##  1st Qu.: NA   1st Qu.: NA   1st Qu.: NA   1st Qu.: NA   1st Qu.: NA
##  Median : NA   Median : NA   Median : NA   Median : NA   Median : NA
##  Mean   :NaN   Mean   :NaN   Mean   :NaN   Mean   :NaN   Mean   :NaN
##  3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA
##  Max.   : NA   Max.   : NA   Max.   : NA   Max.   : NA   Max.   : NA
##  12th_grade_no_diploma high_school_diploma ged_or_alternative_credential
##  Min.   : NA           Min.   : NA         Min.   : NA
##  1st Qu.: NA           1st Qu.: NA         1st Qu.: NA
##  Median : NA           Median : NA         Median : NA
##  Mean   :NaN           Mean   :NaN         Mean   :NaN
##  3rd Qu.: NA           3rd Qu.: NA         3rd Qu.: NA
##  Max.   : NA           Max.   : NA         Max.   : NA
##  some_college_less_than_1_year some_college_1_or_more_years_no_degree
##  Min.   : NA                   Min.   : NA
##  1st Qu.: NA                   1st Qu.: NA
##  Median : NA                   Median : NA
##  Mean   :NaN                   Mean   :NaN
##  3rd Qu.: NA                   3rd Qu.: NA
##  Max.   : NA                   Max.   : NA
##  associates_degree bachelors_degree masters_degree
```

```
##  Min.   : NA       Min.   : NA      Min.   : NA
##  1st Qu.: NA       1st Qu.: NA      1st Qu.: NA
##  Median : NA       Median : NA      Median : NA
##  Mean   :NaN       Mean   :NaN      Mean   :NaN
##  3rd Qu.: NA       3rd Qu.: NA      3rd Qu.: NA
##  Max.   : NA       Max.   : NA      Max.   : NA
##  professional_school_degree doctorate_degree
##  Min.   : NA                Min.   : NA
##  1st Qu.: NA                1st Qu.: NA
##  Median : NA                Median : NA
##  Mean   :NaN                Mean   :NaN
##  3rd Qu.: NA                3rd Qu.: NA
##  Max.   : NA                Max.   : NA
```

```
#describe(Crime_Beats_Census)
#View(Crime_Beats_Census)
```

#Using this dataset we will establish the relationship between education and crime. We will find out if the area in Seattle with the highest grads has a low crime rate.

(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.

```
df2 <- na.omit(Crime_Beats_Census)
summary(df2)
```

```
##      Beat           Report Number Occurred Date      Occurred Time
##  Length:0           Min.   : NA   Length:0           Min.   : NA
##  Class :character   1st Qu.: NA   Class :character   1st Qu.: NA
##  Mode  :character   Median : NA   Mode  :character   Median : NA
##                     Mean   :NaN                      Mean   :NaN
##                     3rd Qu.: NA                      3rd Qu.: NA
##                     Max.   : NA                      Max.   : NA
##  Reported Date      Reported Time Crime Subcategory
##  Length:0           Min.   : NA   Length:0
##  Class :character   1st Qu.: NA   Class :character
##  Mode  :character   Median : NA   Mode  :character
##                     Mean   :NaN
##                     3rd Qu.: NA
##                     Max.   : NA
##  Primary Offense Description   Precinct           Sector
##  Length:0                    Length:0           Length:0
##  Class :character            Class :character   Class :character
##  Mode  :character            Mode  :character   Mode  :character
##
##
##
##  Neighborhood       11_digit_code       Location 1          Latitude
##  Length:0           Length:0           Length:0           Min.   : NA
##  Class :character   Class :character   Class :character   1st Qu.: NA
##  Mode  :character   Mode  :character   Mode  :character   Median : NA
##                                                           Mean   :NaN
##                                                           3rd Qu.: NA
```

```
##                                                                 Max.   : NA
##    Longitude     Census_Tract    GEO.id          GEO.display-label
##  Min.   : NA   Min.   : NA   Length:0           Length:0
##  1st Qu.: NA   1st Qu.: NA   Class :character   Class :character
##  Median : NA   Median : NA   Mode  :character   Mode  :character
##  Mean   :NaN   Mean   :NaN
##  3rd Qu.: NA   3rd Qu.: NA
##  Max.   : NA   Max.   : NA
##      total      no_schooling nursery_school  kindergarten   1st_grade
##  Min.   : NA   Min.   : NA   Min.   : NA    Min.   : NA   Min.   : NA
##  1st Qu.: NA   1st Qu.: NA   1st Qu.: NA    1st Qu.: NA   1st Qu.: NA
##  Median : NA   Median : NA   Median : NA    Median : NA   Median : NA
##  Mean   :NaN   Mean   :NaN   Mean   :NaN    Mean   :NaN   Mean   :NaN
##  3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA    3rd Qu.: NA   3rd Qu.: NA
##  Max.   : NA   Max.   : NA   Max.   : NA    Max.   : NA   Max.   : NA
##    2nd_grade     3rd_grade     4th_grade     5th_grade     6th_grade
##  Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA
##  1st Qu.: NA   1st Qu.: NA   1st Qu.: NA   1st Qu.: NA   1st Qu.: NA
##  Median : NA   Median : NA   Median : NA   Median : NA   Median : NA
##  Mean   :NaN   Mean   :NaN   Mean   :NaN   Mean   :NaN   Mean   :NaN
##  3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA
##  Max.   : NA   Max.   : NA   Max.   : NA   Max.   : NA   Max.   : NA
##    7th_grade     8th_grade     9th_grade     10th_grade    11th_grade
##  Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA
##  1st Qu.: NA   1st Qu.: NA   1st Qu.: NA   1st Qu.: NA   1st Qu.: NA
##  Median : NA   Median : NA   Median : NA   Median : NA   Median : NA
##  Mean   :NaN   Mean   :NaN   Mean   :NaN   Mean   :NaN   Mean   :NaN
##  3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA
##  Max.   : NA   Max.   : NA   Max.   : NA   Max.   : NA   Max.   : NA
##  12th_grade_no_diploma high_school_diploma ged_or_alternative_credential
##  Min.   : NA           Min.   : NA         Min.   : NA
##  1st Qu.: NA           1st Qu.: NA         1st Qu.: NA
##  Median : NA           Median : NA         Median : NA
##  Mean   :NaN           Mean   :NaN         Mean   :NaN
##  3rd Qu.: NA           3rd Qu.: NA         3rd Qu.: NA
##  Max.   : NA           Max.   : NA         Max.   : NA
##  some_college_less_than_1_year some_college_1_or_more_years_no_degree
##  Min.   : NA                   Min.   : NA
##  1st Qu.: NA                   1st Qu.: NA
##  Median : NA                   Median : NA
##  Mean   :NaN                   Mean   :NaN
##  3rd Qu.: NA                   3rd Qu.: NA
##  Max.   : NA                   Max.   : NA
##  associates_degree bachelors_degree masters_degree
##  Min.   : NA       Min.   : NA      Min.   : NA
##  1st Qu.: NA       1st Qu.: NA      1st Qu.: NA
##  Median : NA       Median : NA      Median : NA
##  Mean   :NaN       Mean   :NaN      Mean   :NaN
##  3rd Qu.: NA       3rd Qu.: NA      3rd Qu.: NA
##  Max.   : NA       Max.   : NA      Max.   : NA
##  professional_school_degree doctorate_degree
##  Min.   : NA                Min.   : NA
##  1st Qu.: NA                1st Qu.: NA
##  Median : NA                Median : NA
```

```
##  Mean   :NaN                Mean   :NaN
##  3rd Qu.: NA                3rd Qu.: NA
##  Max.   : NA                Max.   : NA
```

```
#describe(df2)
names(df2)
```

```
##  [1] "Beat"
##  [2] "Report Number"
##  [3] "Occurred Date"
##  [4] "Occurred Time"
##  [5] "Reported Date"
##  [6] "Reported Time"
##  [7] "Crime Subcategory"
##  [8] "Primary Offense Description"
##  [9] "Precinct"
## [10] "Sector"
## [11] "Neighborhood"
## [12] "11_digit_code"
## [13] "Location 1"
## [14] "Latitude"
## [15] "Longitude"
## [16] "Census_Tract"
## [17] "GEO.id"
## [18] "GEO.display-label"
## [19] "total"
## [20] "no_schooling"
## [21] "nursery_school"
## [22] "kindergarten"
## [23] "1st_grade"
## [24] "2nd_grade"
## [25] "3rd_grade"
## [26] "4th_grade"
## [27] "5th_grade"
## [28] "6th_grade"
## [29] "7th_grade"
## [30] "8th_grade"
## [31] "9th_grade"
## [32] "10th_grade"
## [33] "11th_grade"
## [34] "12th_grade_no_diploma"
## [35] "high_school_diploma"
## [36] "ged_or_alternative_credential"
## [37] "some_college_less_than_1_year"
## [38] "some_college_1_or_more_years_no_degree"
## [39] "associates_degree"
## [40] "bachelors_degree"
## [41] "masters_degree"
## [42] "professional_school_degree"
## [43] "doctorate_degree"
```

```
str(df2)
```

```
## 'data.frame':    0 obs. of  43 variables:
##  $ Beat                                : chr
##  $ Report Number                       : num
```

```
##  $ Occurred Date                          : chr
##  $ Occurred Time                          : num
##  $ Reported Date                          : chr
##  $ Reported Time                          : num
##  $ Crime Subcategory                      : chr
##  $ Primary Offense Description            : chr
##  $ Precinct                               : chr
##  $ Sector                                 : chr
##  $ Neighborhood                           : chr
##  $ 11_digit_code                          : chr
##  $ Location 1                             : chr
##  $ Latitude                               : num
##  $ Longitude                              : num
##  $ Census_Tract                           : num
##  $ GEO.id                                 : chr
##  $ GEO.display-label                      : chr
##  $ total                                  : num
##  $ no_schooling                           : num
##  $ nursery_school                         : num
##  $ kindergarten                           : num
##  $ 1st_grade                              : num
##  $ 2nd_grade                              : num
##  $ 3rd_grade                              : num
##  $ 4th_grade                              : num
##  $ 5th_grade                              : num
##  $ 6th_grade                              : num
##  $ 7th_grade                              : num
##  $ 8th_grade                              : num
##  $ 9th_grade                              : num
##  $ 10th_grade                             : num
##  $ 11th_grade                             : num
##  $ 12th_grade_no_diploma                  : num
##  $ high_school_diploma                    : num
##  $ ged_or_alternative_credential          : num
##  $ some_college_less_than_1_year          : num
##  $ some_college_1_or_more_years_no_degree: num
##  $ associates_degree                      : num
##  $ bachelors_degree                       : num
##  $ masters_degree                         : num
##  $ professional_school_degree             : num
##  $ doctorate_degree                       : num
```

#The given dataset has 43 coloumns. The beats, date and time of crime reported and occurance, short description of the crime that occurred, category of the crime, are all variables that decribe the crime and are used to answer questions about the what in this dataset. The precinct,sector, neighbourhood, location, 11 digit code and latitude and longitude give us the location details of the crime and help answer questions about the where in this dataset.The census data gives us information about the people of the area where the crime was committed,mainly their educational background. Using these variables we try to answer the why a crime occurs by establishing correlation between variables of the what, where and why.

#What mainly hinders our analysis is the fact that as many datasets have been combined, there is a lot of duplicate data, which is redundant. For a given neighbourhood or sector, there are mulitple rows, each describing one attribute which cohesively do not aid analysis and make it harder to distincly apply functions to analyse the dataset to get insights. There is also a lot of ambiguity in exactly what the education level columns indicate.

(c) Data Analysis

Use the dataset to provide empirical evidence that addressed your question from part (a). Discuss your results. Provide at least one visualization to support your narrative.

```
ggplot(data = df2, aes_(x=df2$Neighborhood,y=df2$bachelors_degree)) + geom_point()
```


```
ggplot(df2, aes(x = df2$Neighborhood, y = df2$bachelors_degree)) + geom_bar(stat = "identity")
```

#Thus using visuaizations we try to establish a relationship between education and crime, but as the data is diverse and is not tidy enough, the results are hard to read.

(d) Reflect and Question

Comment the questions (and answers) in this analysis. Were you able to answer all of these questions? Are all questions well defined? Is the data good enough to answer all these?

#I was able to partially answer the questions. While the questions were understood, they were not as well defined. The level of education that is assumed to influence crime/ prevention of crime was not defined. The data is good enough to answer the questions, once the questions are defined with very clear specifications. As the dataset is large and diverse with multiple duplicate datapoints it is important to filter down to be able to analyse the problem. Thus the questions need to specify what crimes, what level of education, what time frame etc needs to be considered for the quesiton to be well defined. In the quesition I generated, while I did narrow down the education level, I failed to provide more detail to the question which would help me establish the relationship between crime in a neighborhood and its education level.

Problem 3: Sampling with and without Replacement

In the following situations assume that half of the specified population wears hats every day and the other half does not wear hats.

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

```
cat("\nProbability of sampling two hat-wearing people in a row when sampling with replacement =",(5/10)*
```

```
##
## Probability of sampling two hat-wearing people in a row when sampling with replacement = 0.25
```

```
## Probability of sampling two hat-wearing people in a row when sampling with replacement = 0.25
cat("\nProbability of sampling two hat-wearing people in a row when sampling without replacement =",(5/
```

```
##
## Probability of sampling two hat-wearing people in a row when sampling without replacement = 0.222222
```

```
## Probability of sampling two hat-wearing people in a row when sampling without replacement = 0.222222
```

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

```
cat("\nProbability of sampling two hat-wearing people in a row when sampling with replacement =",0.5*0.5
```

```
##
## Probability of sampling two hat-wearing people in a row when sampling with replacement = 0.25
```

```
##
## Probability of sampling two hat-wearing peoplein a row when sampling with replacement = 0.25
cat("\nProbability of sampling two hat-wearing people in a row when sampling without replacement =",(0.5
```

```
##
## Probability of sampling two hat-wearing people in a row when sampling without replacement = 0.249975
```

```
##
## Probability of sampling two hat-wearing people in a row when sampling without replacement = 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.

```
#This assumption holds true and is demonstrated in the above parts. When the population was small (10 p

#Thus, it is reasonable to treat individuals who are sampled from a large population as independent.
```