IMT 573: Problem Set 4 - Data Analysis

Xinyi Yang

Due: Tuesday, November 3, 2020

Collaborators:

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

1. Download the `problemset4.Rmd` file from Canvas. 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. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment. In particular, note that Stack Overflow is licenses as Creative Commons (CC-BY-SA). This means you have to attribute any code you refer from SO.
4. Partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. But please **DO NOT** submit pages and pages of hard-to-read code and attempts that is impossible to grade. That is, avoid redundancy. Remember that one of the key goals of a data scientist is to produce coherent reports that others can easily follow. 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 without errors you can do so with the `eval=FALSE` option as follows:

```
a + b # these object dont' exist
# if you run this on its own it with give an error
```

6. 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.
7. Collaboration is often fun and useful, 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.

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

```
# Load standard libraries
library(tidyverse)
library(gridExtra)
library(reshape2)
```

Problem 1: 50 States in the USA In this problem we will use the `state` dataset, available as part of the R statistical computing platform. 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.

```r
data(state)
```

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

RESPONSE: **data description**

1. state.abb: character vector of 2-letter abbreviations for the state names.
2. state.area: numeric vector of state areas (in square miles).
3. 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.
4. 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).
5. state.name: character vector giving the full state names.
6. state.region: factor giving the region (Northeast, South, North Central, West) that each state belongs to.
7. state.x77: matrix with 50 rows and 8 columns giving the following statistics in the respective columns.
    - Population: population estimate as of July 1, 1975
    - Income: per capita income (1974)
    - Illiteracy: illiteracy (1970, percent of population)
    - Life Exp: life expectancy in years (1969–71)
    - Murder: murder and non-negligent manslaughter rate per 100,000 population (1976)
    - HS Grad: percent high-school graduates (1970)
    - Frost: mean number of days with minimum temperature below freezing (1931–1960) in capital or large city
    - Area: land area in square miles

```r
state <- data.frame(state.x77)

names(state) <- c("Population", "Income", "Illiteracy", "LifeExp", "Murder", "HSGrad", "Frost", "Area")
state$abb <-state.abb
#area not the same

state$name <- state.name
state$region <- state.region
state$division <- state.division
```

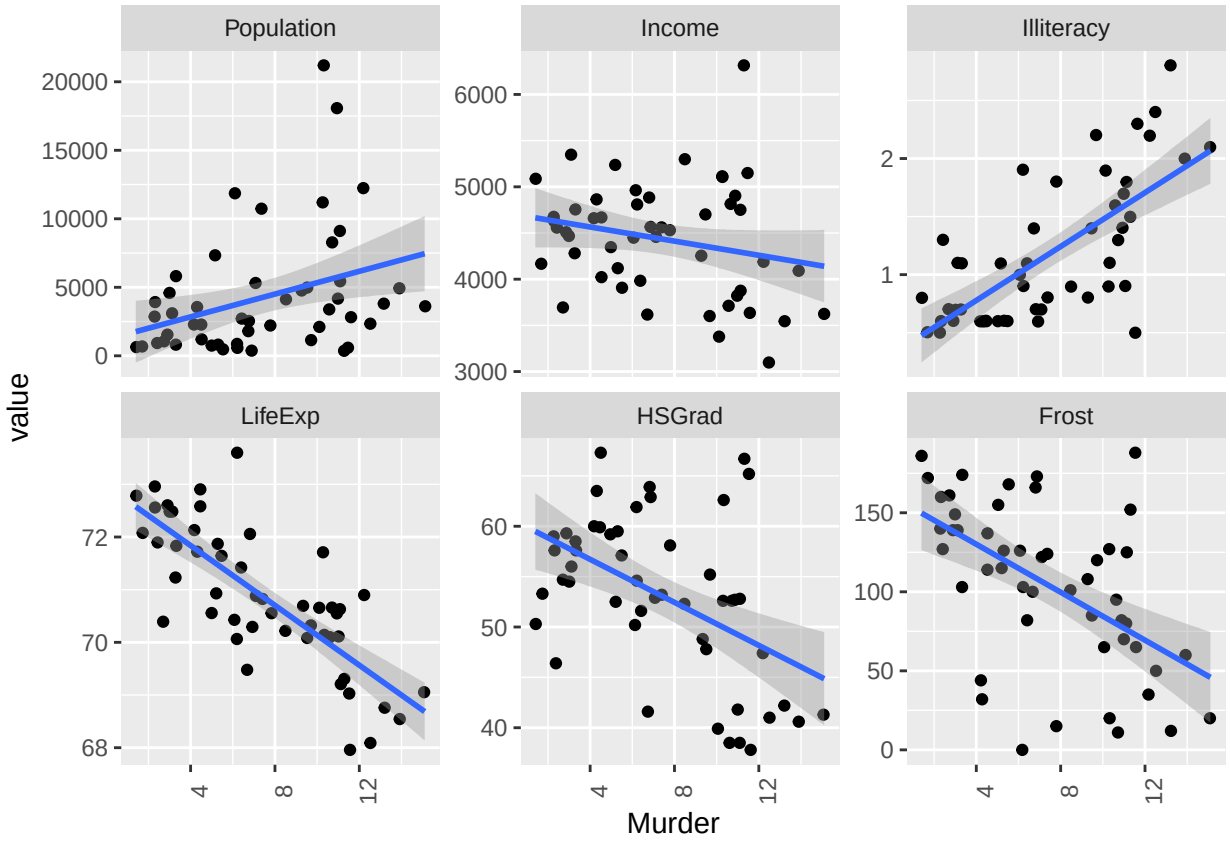
...

(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? ...

```r
# draw scatter point plots to observe the bivariate relationship
# print the correlation between the two variates on the plot

# trash
# draw <- function(x){
#   ggplot(mapping = aes(x,Murder))+
```

```r
#   geom_jitter()+
#   geom_smooth(method = 'glm')+
#   annotate("text", y=max(Murder),x=max(x),label = round(cor(x, Murder),3))
# }
# draw(Population)
# draw(Income)
# draw(Frost)
# draw(LifeExp)

df <- state %>% select(c("Population", "Income", "Illiteracy", "LifeExp", "HSGrad", "Frost","Murder"))

# create a table to store the correlation value
corr <- tibble(
  varialbe = names(df),
  correlation = rnorm(length(names(df)))
)

# use a loop to output the correlation
for (i in seq_along(df)){
  corr[i,2] <- cor(df[i],df$Murder)
}


# to plot these plots in grid area
# reference: https://stackoverflow.com/questions/21653269/use-columns-as-facets-in-graph

df.mlt <- melt(df, id.vars=c("Murder"))

ggplot(df.mlt, aes(x=Murder, y=value)) +
  geom_jitter() +
  geom_smooth(method = 'glm')+
  facet_wrap(~ variable, scales="free_y") +
  theme(axis.text.x=element_text(angle=90))
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
corr
```

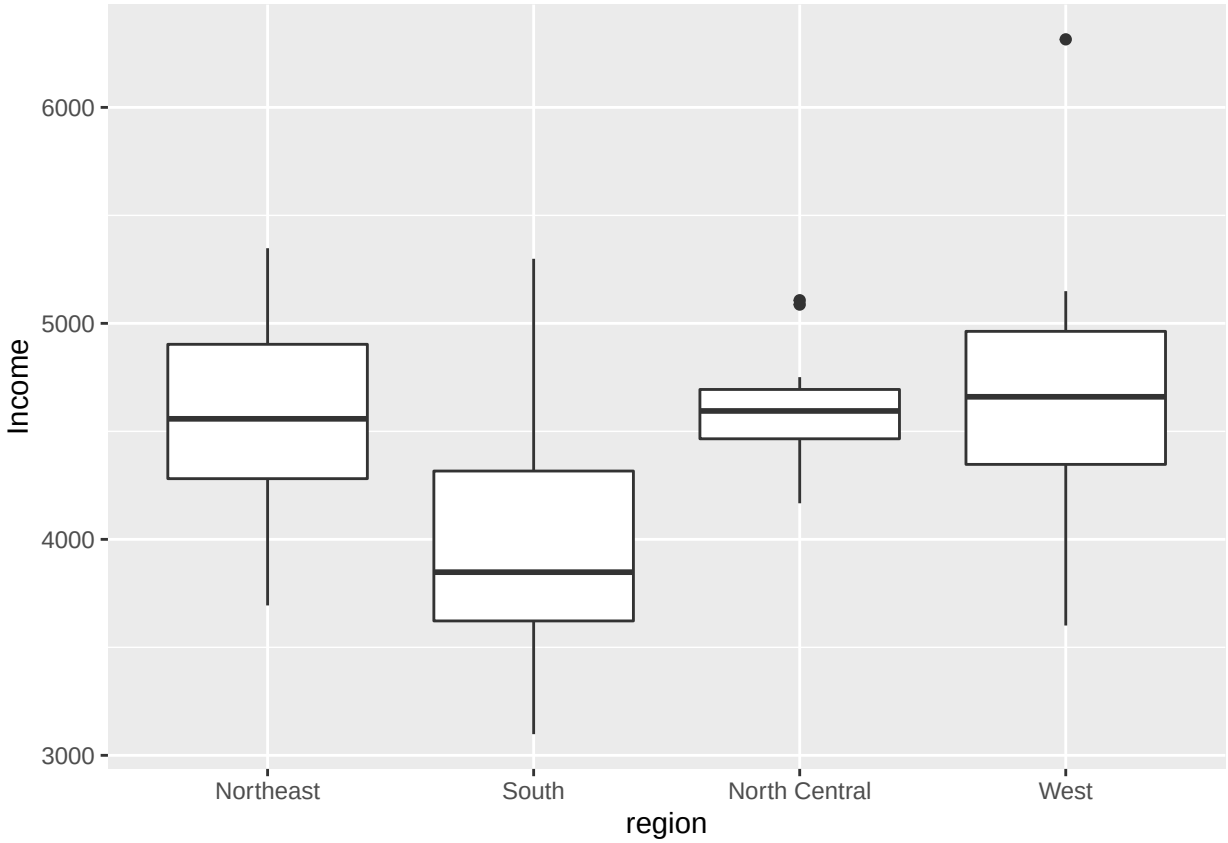
```
## # A tibble: 7 x 2
##   varialbe   correlation
##   <chr>            <dbl>
## 1 Population       0.344
## 2 Income          -0.230
## 3 Illiteracy       0.703
## 4 LifeExp         -0.781
## 5 HSGrad          -0.488
## 6 Frost           -0.539
## 7 Murder           1.00
```

Response: I found two variables that might be important to consider in building a model to explain variation in murder rates, and they are `LifeExp`(life expectancy in years) and `Illiteracy`. The respective correlation shown on the plots is closer to 1 or -1.

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

What does the income distribution looks like in each region across the states? Do different cities in the same region have obivous gaps in their income?

```
# income distribution by region
ggplot(state,aes(region, Income)) +
  geom_boxplot()
```




```
# how spread apart of the income of cities in each region
state %>%
  group_by(region) %>%
  summarise(var_income = var(Income), std_income = sd(Income))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 4 x 3
##   region        var_income std_income
##   <fct>              <dbl>      <dbl>
## 1 Northeast        312567.       559.
## 2 South            366570.       605.
## 3 North Central     80136.       283.
## 4 West             440764.       664.
```

```
# just curious about that single point in West
state %>%
  filter(region == 'West') %>%
  slice_max(Income)
```

```
##        Population Income Illiteracy LifeExp Murder HSGrad Frost   Area abb
## Alaska        365   6315        1.5   69.31   11.3   66.7   152 566432  AK
##          name region division
## Alaska Alaska   West  Pacific
```

RESPONSE:

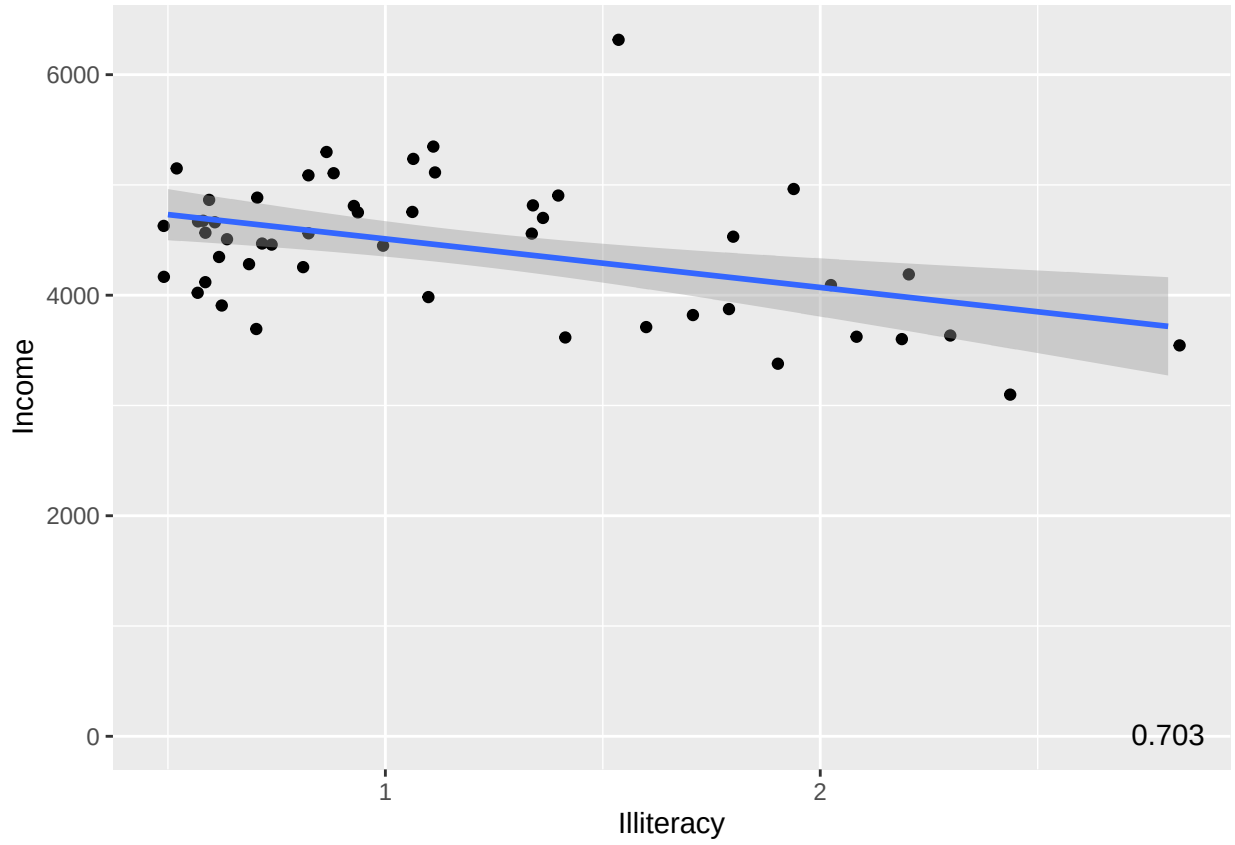
I wish I could do this analysis on newer data. Within the four regions, the west region has the widest spread of income. And the gap between different cities in the region of north central is the smallest.

TO SATISFY MY CURIOSITY

Does people's income have a relationship with their Illiteracy?

```
attach(state)
ggplot(mapping = aes(Illiteracy,Income))+
  geom_jitter()+
  geom_smooth(method = 'glm')+
  annotate("text", y=max(Murder),x=max(Illiteracy),label = round(cor(Illiteracy, Murder),3))
```

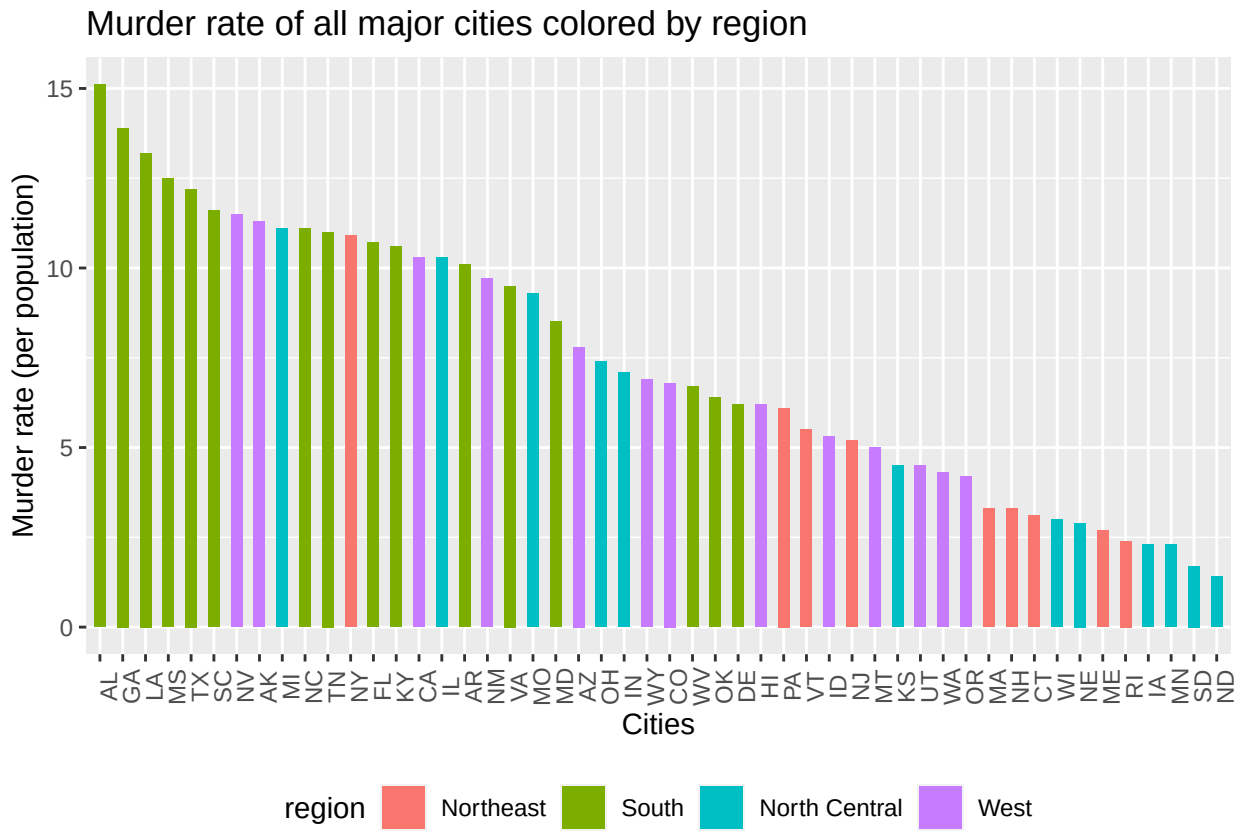
```
## `geom_smooth()` using formula 'y ~ x'
```



RESPONSE: Although the correlation between illiteracy and income is not very close to 1, there's a vague trend that cities with higher illiteracy earn less income.

Just want to make a visualization of the murder rate of all the cities in a descending order and colored by regions.

```
df <- state %>% arrange(desc(Murder))
ggplot(df, aes(abb,Murder))+
  geom_bar(aes(fill=region) ,stat = 'identity', width = 0.5)+
  xlim(df$abb)+
  labs(title="Murder rate of all major cities colored by region")+
  xlab("Cities")+
  ylab("Murder rate (per population)")+
  theme(axis.text.x = element_text(angle = 90),
        legend.position = 'bottom')
```



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.

QUESTION: Do beats with more people who have been to college have less murders occur?

(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.

The dataset is about the reports of crimes in the Seattle area, and the corresponding education information in the area. It has the time, category, place of the crime, and how many people are to what level educated.

```
# load the crimes dataset
crime <- read.csv("crime_beats_census.csv")
# remove the original index column
crime <- crime %>%  select(-c('X'))

# the columns
names(crime)
```

```
##  [1] "Report.Number"
##  [2] "occur_day"
```

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

```
# the data type of each column
str(crime)
```

```
## 'data.frame':    347980 obs. of  49 variables:
##  $ Report.Number                        : num  2.01e+13 2.01e+13 2.01e+13 2.01e+13 2.01e+13 ...
##  $ occur_day                            : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ occur_month                          : int  2 2 2 2 2 2 2 2 2 2 ...
```

```
##  $ occur_year                            : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ..
##  $ Occurred.Time                         : int  2040 2100 1930 2144 2218 2229 2230 2015 2256 2255 ..
##  $ Reported.Date                         : chr  "04/03/2012" "04/02/2012" "04/02/2012" "04/02/2012"
##  $ Reported.Time                         : int  28 2103 2126 2144 2218 2229 2356 2312 2256 2300 ...
##  $ Crime.Subcategory                     : chr  "NARCOTIC" "ROBBERY-COMMERCIAL" "MOTOR VEHICLE THEFT
##  $ Primary.Offense.Description           : chr  "NARC-POSSESS-MARIJU" "ROBBERY-BUSINESS-GUN" "VEH-TH
##  $ Precinct                              : chr  "WEST" "NORTH" "NORTH" "EAST" ...
##  $ Sector                                : chr  "K" "B" "J" "E" ...
##  $ Beat                                  : chr  "K2" "B2" "J1" "E3" ...
##  $ Neighborhood                          : chr  "PIONEER SQUARE" "BALLARD SOUTH" "BALLARD NORTH" "CA
##  $ beat_fct                              : chr  "K2" "B2" "J1" "E3" ...
##  $ Location.1                            : chr  "(47.5998930290529, -122.326813620856)" "(47.67905219
##  $ Latitude                              : num  47.6 47.7 47.7 47.6 47.6 ...
##  $ Longitude                             : num  -122 -122 -122 -122 -122 ...
##  $ geo_code                              : num  5.3e+14 5.3e+14 5.3e+14 5.3e+14 5.3e+14 ...
##  $ state                                 : int  53 53 53 53 53 53 53 53 53 53 ...
##  $ county                                : int  33 33 33 33 33 33 33 33 33 33 ...
##  $ eleven_digit                          : num  5.3e+10 5.3e+10 5.3e+10 5.3e+10 5.3e+10 ...
##  $ GEO.id                                : chr  "1400000US53033009200" "1400000US53033003200" "14000
##  $ GEO.id2                               : num  5.3e+10 5.3e+10 5.3e+10 5.3e+10 5.3e+10 ...
##  $ GEO.display.label                     : chr  "Census Tract 92, King County, Washington" "Census Tr
##  $ total                                 : int  2529 6896 2806 2477 7239 215 5123 4414 5355 6414 ...
##  $ no_schooling                          : int  56 26 17 100 27 0 135 0 164 59 ...
##  $ nursery_school                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ kindergarten                          : int  0 0 0 44 1 0 0 0 0 0 ...
##  $ X1st_grade                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X2nd_grade                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X3rd_grade                            : int  37 0 0 0 27 0 79 0 24 0 ...
##  $ X4th_grade                            : int  5 0 0 0 0 0 87 0 19 0 ...
##  $ X5th_grade                            : int  17 0 0 15 0 0 4 0 0 102 ...
##  $ X6th_grade                            : int  156 0 0 0 0 0 43 0 83 0 ...
##  $ X7th_grade                            : int  4 0 0 0 0 0 35 0 36 1 ...
##  $ X8th_grade                            : int  100 15 26 53 0 0 31 0 0 8 ...
##  $ X9th_grade                            : int  49 0 0 47 1 0 88 8 46 0 ...
##  $ X10th_grade                           : int  19 0 0 0 0 0 18 7 56 0 ...
##  $ X11th_grade                           : int  14 15 23 0 0 0 63 72 109 0 ...
##  $ X12th_grade_no_diploma                : int  63 0 4 31 28 41 90 17 109 70 ...
##  $ high_school_diploma                   : int  354 348 120 104 286 5 653 372 1020 302 ...
##  $ ged_or_alternative_credential         : int  88 102 4 110 26 11 165 153 255 83 ...
##  $ some_college_less_than_1_year         : int  134 205 128 53 224 10 338 155 260 315 ...
##  $ some_college_1_or_more_years_no_degree: int  503 776 266 243 840 42 1289 788 615 703 ...
##  $ associates_degree                     : int  114 444 106 136 570 5 501 220 804 254 ...
##  $ bachelors_degree                      : int  536 3000 1175 936 3256 70 1062 1613 1131 2992 ...
##  $ masters_degree                        : int  172 1433 659 365 1252 26 282 699 410 1005 ...
##  $ professional_school_degree            : int  71 374 144 130 458 0 97 189 94 286 ...
##  $ doctorate_degree                      : int  37 158 134 110 243 5 63 121 120 234 ...
```

```
# check the contents of Sector
unique(crime$Sector)
```

```
##  [1] "K"    "B"    "J"    "E"    "C"    "U"    "F"    "S"    "Q"    "G"
## [11] "M"    "D"    "N"    "R"    "L"    "O"    "W"    ""     "6804"
```

```
# clean the data
crime <- crime[!(is.na(crime$Sector) | crime$Sector=="" |crime$Sector=='6804'), ]
```

(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.

```
# create a vector of all the high-education columns
colnms=c('bachelors_degree','masters_degree',   'professional_school_degree','doctorate_degree')

crime_df <- crime %>%
  # sum the number of the high-education people by rows, remove NAs
  mutate(HighEdu = rowSums(crime[,colnms],na.rm = TRUE)) %>%
  # select the columns I might use to reduce runtime
  select(Report.Number,Sector,Beat, total,HighEdu) %>%
  # calculate the proportion of the high-education people among the whole population
  mutate(HighEduperPpl=HighEdu/total)

# create another dataframe grouped by Beat
crime_df2 <- crime_df %>%
  group_by(Beat) %>%
    # calculate the mean proportion of high-education people in each Beat
  summarise(meanEdu=mean(HighEduperPpl))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

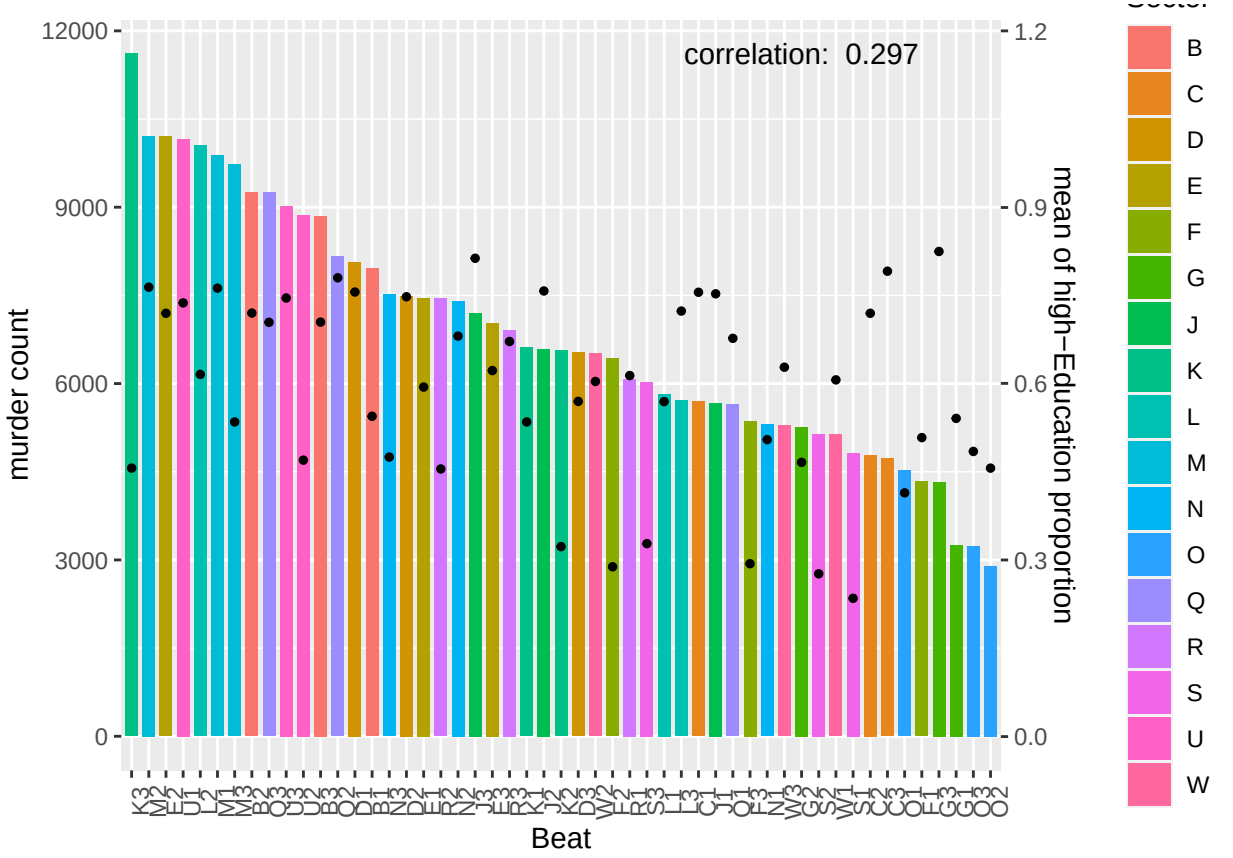
```
# create a small dataframe with Beat and number of murders occurred
df <- crime %>% count(Beat) %>% arrange(desc(n)) %>%
  # match the education variable
  left_join(crime_df2)
```

```
## Joining, by = "Beat"
```

```
# create a Sector-Beat sheet
StoB <- unique(crime_df %>% select(Sector,Beat))

# match the Sector
dff <- df %>% left_join(StoB)
```

```
## Joining, by = "Beat"
```

```
# set the xais and annotation before-hand
xaxis <- crime %>% count(Beat) %>% arrange(desc(n))
corr <- paste("correlation: " ,round(cor(dff$n,dff$meanEdu),3))

ggplot(dff,aes(x=Beat)) +
  geom_bar(aes(y=n,fill=Sector),stat = 'identity',width = 0.7) +
  geom_point(aes(y=meanEdu*10000),color='black',size=1) + # multiple by 10000 to make them visible on t
  theme(axis.text.x = element_text(angle = 90),
        legend.position = 'right') +
  scale_y_continuous(
    name = "murder count",
    sec.axis = sec_axis(~.*0.0001,name = "mean of high-Education proportion")
    # adjust the axis to show the real value of the education variable
  ) +
  scale_x_discrete(limits=xaxis$Beat) +
  annotate("text",y=max(dff$n),x=40,label = corr)
```



RESPONSE: Beats with more people who have been to college don't have less murders occur. And the correlation between the proportion of people whose education level are higher than bachelor's degree and the number of murders is low. We can also observe the results from the visualization above. Although the bar chart is ordered by the number of murders, the points are splattered on the plot messily. Maybe people just don't do such malicious things in their neighborhood. I should look more into places where the illiteracy is higher.

(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?

RESPONSE: When answering the question of bivariate relationships, I felt a little bit confused because I don't know how to observe the relationship. I don't think we did this in class or in lab. The crime and education dataset is not organized enough to answer my question. I need to preprocess the dataset first, adding several columns together.