IMT 573: Problem Set 4 - Data Analysis

Stephen V. Tucker

November 4, 2022

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 don't exist
# if you run this on its own it will 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(corrplot)
library(fmsb)
```

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. See here for more.

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

Answer

The data consist of quantitative statistical data pertaining to 50 states of america. This data has columns detailing population, Income, Illiteracy, Life Expectancy, murder rate, High school graduation success rate, Frost and Area. To understand this data fully I conducted research to get full context to what the numerical data meant.

Link: http://stats4stem.weebly.com/r-statex77-data.html

- 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.
- population: population estimate as of July 1, 1975
- Income: per capita income (1974)
- Illiteracy: per capita income (1974)
- 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

```
## Check for Answer under each H?eading tha says Answer

state_matrix <- state.x77
state_df <- data.frame(state_matrix)
str(state_df)
```

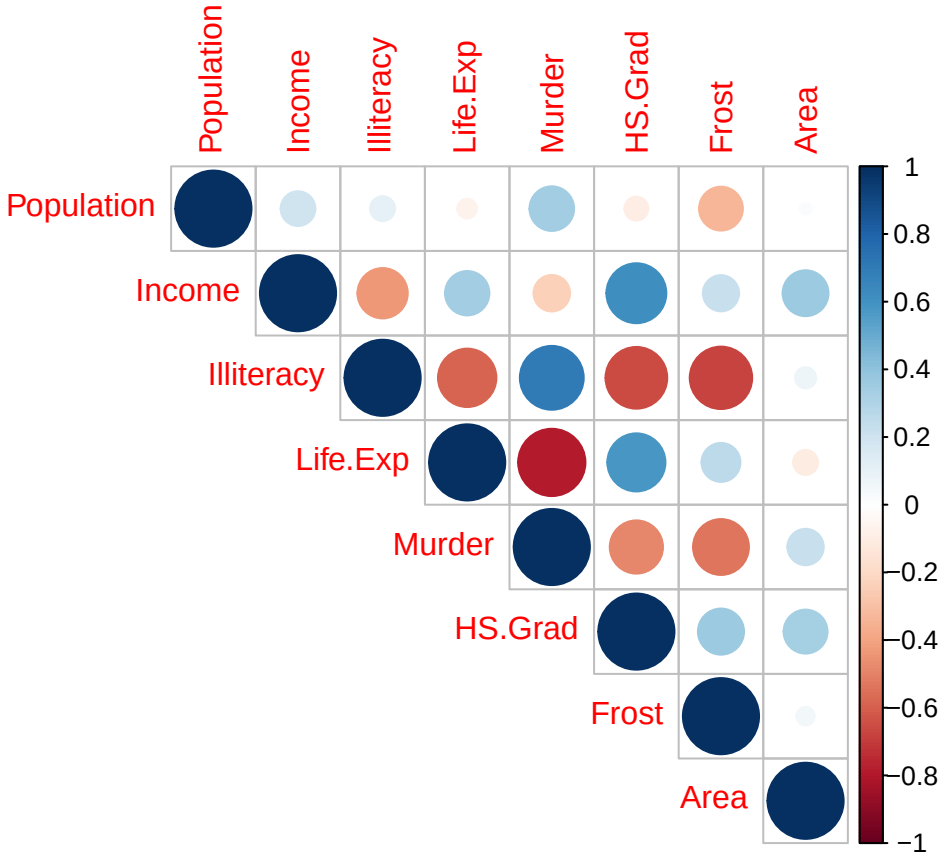
```
## 'data.frame':    50 obs. of  8 variables:
##  $ Population: num  3615 365 2212 2110 21198 ...
##  $ Income    : num  3624 6315 4530 3378 5114 ...
##  $ Illiteracy: num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
##  $ Life.Exp  : num  69 69.3 70.5 70.7 71.7 ...
##  $ Murder    : num  15.1 11.3 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 ...
```

(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 associated with `Murder` rate present in the data. What does your analysis suggest might be important variables to consider in building a model to explain variation in murder rates?

Answer & visual

According to the data some variables important to measure appear to be illiteracy, High Grad. I believe it would be be important to build a model with Illiteracy,population, & Area as our predictors.. Using correlation plot and then exploring more

```
correlations = cor(state_df)
corrplot(correlations, method="circle",type='upper',na.label = "o")
```



(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 1.c

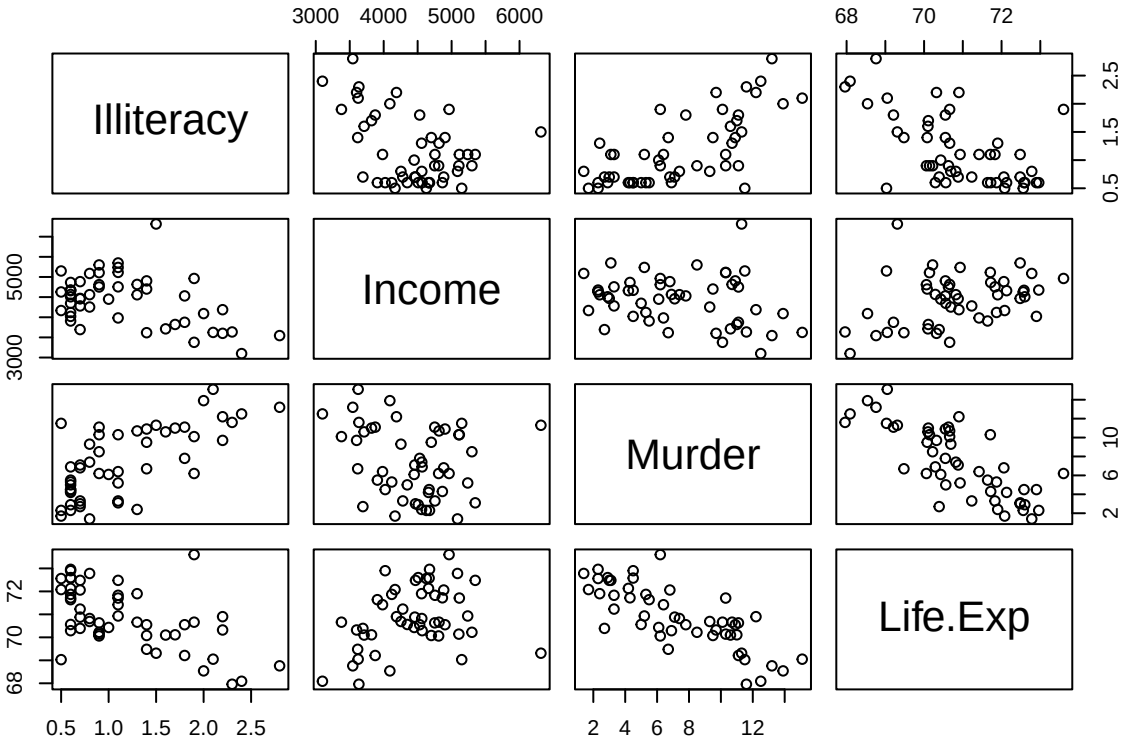
Q. How does illiteracy affect the quality of life for people across the nation?

running the model we see that illiteracy has a negative correlation. I wanted to fit a line with my work but could not do this properly.

```
state_regression = lm(Illiteracy~ Income + Murder + Life.Exp, data = state_df)
summary(state_regression)
```

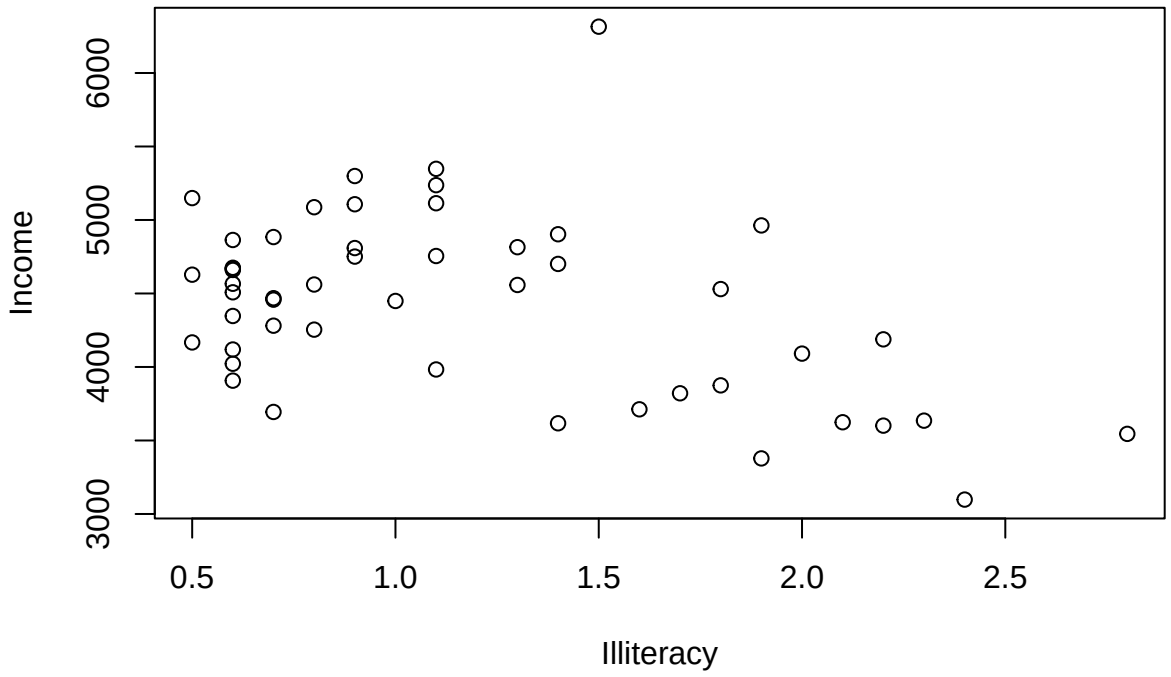
```
##
## Call:
## lm(formula = Illiteracy ~ Income + Murder + Life.Exp, data = state_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88832 -0.20434 -0.08169  0.26851  0.98608
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.0393698  5.1854787   0.200 0.842021
## Income      -0.0002917  0.0001017  -2.869 0.006195 **
## Murder       0.1074330  0.0254708   4.218 0.000114 ***
## Life.Exp     0.0089150  0.0724900   0.123 0.902657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4104 on 46 degrees of freedom
## Multiple R-squared:  0.5744, Adjusted R-squared:  0.5466
## F-statistic: 20.69 on 3 and 46 DF,  p-value: 1.243e-08
```

```
plot(state_regression$model)
```



regression model

```
model = lm(Illiteracy~ Income, data = state_df)
plot(model$model)
```



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 standard dataset will be available on canvas after the problem set 3 due date.

(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 data set you compiled in Problem Set 3.

Question 2a. (data science question)

Q1. What type of crime happens in certain areas and at what frequency?

This question is one that would take more time then the assignment allows for me to answer but something I believe would be nice to explore.

Q2. What is the impact of population density on quality of life. Quality of life is life expectancy, Murder rate, and income?

```
## analysis

## loading data
joined_Crime <- read_csv("~/Downloads/joined_Crime.csv")
```

```
## Rows: 389002 Columns: 36
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): Beat, Location.1, county_code, GEO_ID, Occurred.Date, Reported.Dat...
## dbl (25): Freq, Latitude, Longitude, census_tract, state_code, digital_code_...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#
crime_final <- joined_Crime

neighborhood <- unique(crime_final$Neighborhood)
neighborhood = data.frame(neighborhood)

ballard = crime_final %>%
  filter(Neighborhood ==c('BALLARD SOUTH','BALLARD NORTH'))

capitol_hill = crime_final %>%
  filter(Neighborhood =='CAPITOL HILL')

rainier_beach = crime_final %>%
  filter(Neighborhood =='RAINIER BEACH')
```

```
# Ballard
ballard_filter  = ballard %>%
  select(Primary.Offense.Description,Reported.Date,Neighborhood)

ballard_filter = as.factor(ballard_filter$Primary.Offense.Description)

ballard_filter = data.frame(ballard_filter) %>%
  group_by(ballard_filter) %>%
  count(ballard_filter)

ballard_filter = ballard_filter %>%
  filter(n > 65)

### Capitol Hill ### ### ### ### ### ### ### ### ### ### ### ### ###
library(dplyr)
capitol_hill_filter = crime_final %>%
  filter(Neighborhood =='CAPITOL HILL')


capitol_hill_filter = capitol_hill_filter %>%
  select(Primary.Offense.Description,Reported.Date,Neighborhood)

capitol_hill_filter = as.factor(capitol_hill_filter$Primary.Offense.Description)

# making a dataframe
capitol_hill_filter = data.frame(capitol_hill_filter)

# creating a count
capitol_hill_filter = capitol_hill_filter %>%
  group_by(capitol_hill_filter) %>%
  count(capitol_hill_filter)

capitol_hill_filter = capitol_hill_filter %>%
  filter(n > 75)

### Rainier Beach ### ### ### ### ### ### ### ### ### ### ### ### ###

rainier_beach = crime_final %>%
  filter(Neighborhood =='RAINIER BEACH')
```

```
rainier_beach_filter = rainier_beach %>%
  select(Primary.Offense.Description)

rainier_beach_filter = as.factor(rainier_beach_filter$Primary.Offense.Description)

rainier_beach_filter = data.frame(rainier_beach_filter)

rainier_beach_filter = rainier_beach_filter %>%
  group_by(rainier_beach_filter) %>%
  count(rainier_beach_filter)

rainier_beach_filter = rainier_beach_filter %>%
  filter(n > 75)
```

```
# ggplot(ballard_filter, aes (x = reorder(ballard_filter, n), y = n)) +
#   geom_col(colour = "black") +
#   xlab("Crimes commited in ballard") +
#   ylab("crime count")
```

```
# ggplot(capitol_hill_filter, aes (x = reorder(capitol_hill_filter, n), y = n)) +
#   geom_col(colour = "black") +
#   xlab("Crimes commited in Cap Hill") +
#   ylab("crime count")
```

```
# ggplot(rainier_beach_filter, aes (x = reorder(rainier_beach_filter, n), y = n)) +
#   geom_col(colour = "black", width = 0.6) +
#   xlab("Crimes commited in Rainier Beach") +
#   ylab("crime count")
```

(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.

data summary: The data comprises of various quantitative data that pertains to crimes, Additionally it comprises of categorical data to work with some things that can be fixed is more precise information on population in each neighborhood.

What crimes are committed in certain neighborhoods the most. Do to time constraints I wanted to have a straight forward question we can derive insight from.Understanding what crimes that are reported more frequently would allow organizations to deploy different. Resources to different areas that can be treated as such. For example if more property crimes are happening in a particular area they can strategize how to protect from these incidents.

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

Visual summary

I developed three graphs that tells one story on neighborhood in Seattle that is what crimes are being committed the most in each neighborhood and also at what different rates that are being committed at each particular location. In differentiating the locations it tells us a clearer picture of whats going on in silo's as well as whole. we see Theft is pervasive in all areas while different ones are in second car prowl is the leading in all three.

(d) Reflect and Question Comment on 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?

Answer

I was not able to answer all of my questions I had two specific ones i wanted to explore and one i was able to get valuable insight from. The other would of taken more time. This makes me reflect on why teams can be valuable in optimizing work. Teams allow many dynamic things that can benefit data analysis as well. The question's in my opinion were well defined and the data is available but in the second question relating to density depending on granularity we may be able to answer the question more broad pertaining to different states but coming from county perspective we would need to conduct further analysis.