Final Project
STA 4291/5291

Spring 2018

The project requires you to synthesize the material from the course. It is not a regular Homework and it should be treated differently. It is the best way to solidify your understanding of the material that you learned in this course. You will present your findings in class and in a written report . You should explain what you did using simple words (just like in the challenge problems). You do not need to explain the terminology in details. No formulas. Your can choose your data or select a subset of the data set that I am posting using the technique I describe below. The final report should be clear and readable. The maximum number of pages allowed for the report is 5 (both sides). All figures and tables that are included should be readable, relevant and well labeled. Figures can be added in an appendix (not part of the 5 pages). Include only relevant plots. The written report should have a summary, an introduction (description of the data and why is the analysis is important for the client, what were you trying to find out and what are the findings of your initial data exploration), results (what results did the analysis produce and how do you interpret those results), and conclusion (what is the significance of your results, what is the answer to the research question).

Problem: Officials in Kings county (Brooklyn) wish to determine which factors influence the number of serious crimes per county. The goal is to implement policies that will lead to the reduction of the number serious crimes in their county. Suppose that you were hired to help with this objective using the county demographic information (CDI) data set from Applied Linear Statistical Models, 5th edition, by Kutner, Nachtsheim, Neter, and Li., Appendix C2 (APPENC02.txt). This data set provides selected county demographic information for 440 of the most populous counties in the United States (each one of you is going to analyze a subset of this dataset) . Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set.

Variable Number	Variable Name	Description
1	Identification Number	1-440
2	County Name	County name
3	State	Two-letter state abbreviation
4	Land area	Land area (square miles)
5	Total population	Estimated 1990 population
6	Percent of population aged 18-34	Percent of 1990 CDI population aged 18-34
7	Percent of the population 65 or older	Percent of 1990 CDI population aged 65 years old or older
8	Number of active physicians	Number of professionally active nonfederal physicians during 1990
9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor's degree	Percent of adult population (persons 25 years old or older) with bachelor's degree
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI labor force that is unemployed
15	Per capita income	Per capita income of 1990 CDI population (dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	Geographic region	Geographic region classification is that used by the U.S. Bureau of the Census, where: 1 = NE, 2 = NC, 3 = S, 4 = W

The method to select your data set is described next. Using the following R code and the numbers in your UNI number select a a random sample of 300 observations to use as your data set. I am going to show you what I would do to get my dataset, My UNI is eh2336 (my dataset was saved on the desktop).

```
data<-read.table("/Users/HElbarmi/Desktop/APPENC02.txt", sep="", header=FALSE)

UNI<-2336

set.seed(UNI)

index <-sample(c(1:440))

mydata<-data[index[1:300],]
```

the dataset that you need to analyze is mydata.