PDRI Data
Assignment

**Background**

In January 2012, the Cook County State's Attorney's Office established a program intended to reduce re-arrest among people on bail awaiting trial. The program ran through October 2013. The SA's Office asked you to evaluate the effectiveness of the program and provided the data described below.

Use of R or Python is strongly encouraged. Use of Excel is discouraged. Using multiple different programs is also discouraged. Please provide back your code and answers to the questions below within 48 hours.

**Data**

There are four datasets: *case.csv*, *demo.csv*, *prior_arrests.csv*, and *grades.csv*. *case.csv* is the main dataset and reflects dates of arrest and disposition (trial or court appearance) during the period in which the program operated. The file also contains an indicator of whether the arrestee was referred to the intervention program for that arrest (i.e. whether they were treated), whether the person was re-arrested while awaiting trial, the number of prior arrests at the time of program entry, and the arrest location. *demo.csv* contains demographic information about arrestees, including some who were not included in the program evaluation. *prior_arrests.csv* reflects pre-period arrests among individuals in *case.csv*; the pre-period ran from 2008-2011. Finally, *grades.csv* includes 9th and 10th grade course grades for a subset of individuals in *case.csv*. Further description of the datasets is included at the end of this document.

**Part 1: Data Management**

1. The demographic data were extracted from a system that inconsistently coded *gender*. Recode it so that males are consistently coded as "M" and females are consistently coded as "F".
2. Merge the *case* and *demo* datasets together so that each row in the *case* dataset also contains the demographics of the defendant. Keep in mind that the populations in the *case* and *demo* data may not be 100% aligned.
3. While the program was mostly rolled out to defendants in Chicago, the State's Attorney's Office also ran a pilot serving a small number of individuals arrested in other parts of Cook County. For the purpose of this analysis, please restrict the data to only individuals who were arrested in Chicago.

**Part 2: Variable Creation**

1. Create an *age* variable equal to the defendant's age at the time of arrest for each case.
2. The State's Attorney is interested in pursuing a partnership with the Chicago Public Schools to investigate the relationship between high school achievement and criminal justice outcomes in early adulthood. To that end, the State's Attorney's Office has requested 9th and 10th grade course grade data from defendants between the ages of 18 and 24. These data are included in *grades.csv*. Please construct measures for 9th and 10th grade GPA for this target population. When constructing GPA, please use a 4 point scale, where: A=4, B=3, C=2, D=1, and F=0.

3.

a. The provided *case.csv* file includes a variable that indicates the number of arrests prior to that case for each individual. Please reconstruct the variable using the *prior_arrests.csv* file. Assume that all of the individual's arrests prior to the study period are contained in *prior_arrest.csv*. If someone is not included in *prior_arrests.csv*, assume they had zero arrests at the start of the study period. Also note that some individuals were arrested multiple times during the study period and that this should be accounted for in your prior arrest count. For example, if individual A was arrested 5 times prior to the study period and appears twice in the *case* file, their first arrest in the *case* file should have a prior arrest count of "5" and their second arrest should have a prior arrest count of "6". One final note, some people really do get arrested multiple times on the same day. Count each arrest separately, regardless of whether another arrest occurred on the same day.
b. The *case* file also includes a variable *re_arrest* which indicates whether individuals were arrested during their case period (i.e. after the case's arrest date and before the case's disposition date). Please reconstruct this indicator. Assume that all arrests during the study period are reflected in the *case* file.
c. Please show that the variables you reconstructed are equal to the versions in the provided datasets.

**Part 3: Statistical Analysis**

Help the State's Attorney's Office determine if the program should be continued/expanded by estimating the program's effect on re-arrests prior to disposition. Because we only have grades data for young adults, please do not use these data to inform your statistical analysis. To draw conclusions about this program's effect, answer the following questions.

1. Describe the demographic characteristics of the study population based on the data available to you. (Hint: the study population has 25,000 subjects).
    a. Are the treatment and control groups balanced (on *race*, *gender*, etc.), or are there differences in the composition of the two groups? Please present your answer in the form of a table.
    b. Choose one observable characteristic and visualize the difference between those who were enrolled in the program and those who were not.
3. Did participating in the program reduce the likelihood of re-arrest before disposition? Explain your answer and your methodology.
4. The State's Attorney's Office is interested in expanding the program if it is shown to reduce re-arrest. However, they do not have the budget to serve every individual on bail awaiting trial. In order to make best use of their restricted budget, they would like to target the individuals most likely to benefit from the program. Using the data available to you, what recommendation would you make regarding who to serve?