BUSINESS DATA MINING (IDS 572)

HOMEWORK 1
DUE DATE: WEDNESDAY, SEPTEMBER 26 AT 3:00 PM

- Please provide succinct answers to the questions below.
- You should submit an electronic pdf or word file in blackboard.
- Please include the names of all team-members in your write up and in the name of the file.
- One submission is sufficient for the entire group.
- You should include all the R functions you use in your pdf file.
- Please make sure your graphs have titles, labels, legends (if necessary), appropriate colors and etc.

The goal of the first two questions is to practice visualization and exploring data in R.

Problem 1. FiveThirtyEight, a data journalism site devoted to politics, sports, science, economics, and culture, recently published a series of articles on gun deaths in America. Gun violence in the United States is a significant political issue, and while reducing gun deaths is a noble goal, we must first understand the causes and patterns in gun violence in order to craft appropriate policies. As part of the project, FiveThirtyEight collected data from the Centers for Disease Control and Prevention, as well as other governmental agencies and non-profits, on all gun deaths in the United States from 2012-2014. You can find this dataset, called "gun_deaths.csv", on blackboard.

(a) Generate a data frame that summarizes the number of gun deaths per month. Print the data frame as a formatted kable() table.

(b) Generate a bar chart with labels on the x-axis. That is, each month should be labeled "Jan", "Feb", "Mar" and etc.

(c) Generate a bar chart that identifies the number of fun deaths associated with each type of intent cause of death. The bars should be sorted from highest to lowest values.

(d) Generate a boxplot visualizing the age of gun death victims, by sex. Print the average age of female gun death victims.

Answer the following questions. Generate appropriate figures/tables to support your conclusions.

(e) How many white males with at least a high school education were killed by guns in 2012?

(f) Which season of the year has the most gun deaths? Assume that

  – Winter = January - March
  – Spring = April - June
  – Summer = July - September
  – Fall = October - December
  – Hint: You need to convert a continuous variable into a categorical variable.

(g) Are whites who are killed by guns more likely to die because of suicide or homicide? How does this compare to blacks and Hispanics?