**AML 612 Homework #3**

**For all questions, please submit your R code, and a doc file with a copy of the screen output and plots (where applicable), and all other relevant files to** [**smtowers@asu.edu**](mailto:smtowers@asu.edu)**.**

**Due Wed Feb 21st at noon.**

**The code in your R files should exhibit all of the good coding practices mentioned in** [**http://sherrytowers.com/2012/12/14/good-programming-practices-in-any-language/**](http://sherrytowers.com/2012/12/14/good-programming-practices-in-any-language/)

**Please do not just copy and paste code from the examples given in class into your code. Plots should exhibit all of the good plotting practices mentioned in** [**http://sherrytowers.com/2013/01/04/good-practices-in-producing-plots/**](http://sherrytowers.com/2013/01/04/good-practices-in-producing-plots/)

**Please submit your files in a format hwk3\_<last name>.xxx where xxx is the format of the document (R, doc, tex, pdf, bib, etc)**

**Question 1)**

By now you have read several papers in this course related to topics that interest you.

Prepare a short prospectus describing an idea for a term project that involves some source(s) of data (disease data, crime, socioeconomic, population data, etc… whatever you like), and describes a hypothesis you believe you would like to test using the sources of data. You can either extract the source of data using DataThief from one or more of the papers you’ve read, or look for sources of data online, or perhaps get ideas from the course web page describing lots of sources of free online sources of data <http://sherrytowers.com/2012/04/03/finding-sources-of-data-free-online-data/>

In your prospectus (written in Latex, with bibtex references), give a few sentences **motivating** your proposed project based on literature you have read. Cite at least three prior studies related to the topic.

Cite the literature sources in your prospectus, and provide me with the PDF of the papers you cite in your homework submission. Then, go on to describe your proposed **objective** (remember that “motive” is a description of why someone should be interested in your project, and why what has been done in the past is insufficient to really solve or understand the issue… “objective” is what you plan to do). Discuss whether your proposed objective has been studied before. You don’t have to go into detail of the statistical tests that might be used. You only need to talk about the hypothesis you would like to examine at this point.

**Thoroughly** describe the details of what your data consist of, and describe your sources of data giving links and/or references.

**Do not include your name within the prospectus document (ie; keep it anonymous).**

Note that you are not required to do any statistical analysis of the data at this point.

After I’ve gone through them all once you’ve handed them in, the PDF files of these prospectus proposals will be circulated to all the other students in the class. Students will review the prospectuses, and also rank their top four choices of projects to work on. Based on student rankings (combined with my own opinions of which projects are likely to be feasible based on the data and hypothesis) I will assign you all to project groups. There is no guarantee that your project proposal will be chosen for a project group, but if it is, you will of course be in that group.

**Question 2)**

I have created a file summarizing the number of various types of crime in Chicago by day and hour between 2001 and 2017:

<https://www.dropbox.com/s/gkojluyeq8exbi0/chicago_crimes_by_day_and_hour_2001_to_2017.csv?dl=0>

The data were downloaded from the City of Chicago data portal

<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data>

from their Citizen Law Enforcement Analysis and Reporting

(CLEAR) data base. This data set represents over 6 million reported incidents of crime and is one of the largest such data sets publicly available. It is very richly detailed, including, amongst other things, both temporal and geo-spatial data.

a) From the web page, read the description of the CLEAR data, and answer the following questions

1. Does the data in the CLEAR system reflect details of arrests for crimes? If so, does the CLEAR system give any information about the arrestees? If not, what does the data in the CLEAR system reflect?
2. What sources of data in the CLEAR system did I not include in the summarized file linked above?
3. What steps do you think I had to take in order to obtain the summarized file? (you don’t need to do the steps, just indicate what you think they were). The reason I provide this summarized file is because the file including all information available for the crimes from 2001 onwards would be almost unmanageable in size.

b) Create an R script that reads this data into a data table (call it vdat).

The data table contains a column called “jul” which is the number of days that date is from January 1, 1970 (aka “Julian days”). This was calculated using the julian() function in the R chron library, which takes as its arguments julian(month,day,year).

There is a column called “hour” which is the hour of the day during which the crimes occurred. Hour “0” is midnight to 1am, and hour “23” is 11pm to midnight.

Calculate a new column for the data frame that is the batteries that are not related to domestic violence. This can be obtained from the total number of batteries, minus the batteries that are domestic.

c) Install the chron library in R using install.packages().

A list of the dates of past Super Bowls can be found at <http://www.espn.com/nfl/superbowl/history/winners>

From this list, create a vector called vjul that is filled with the Julian dates of all the past Super Bowls using the julian() function in the chron library

d) From vdat, subset a new data frame that selects the dates that were Super Bowl Sundays. Also, subset another data frame that is exactly one week before, and another data frame that is exactly one week after. The latter two datasets will assess our “baseline” crime that we would expect for an average Sunday that time of year.

e) Do a loop over crime types:

"batteries\_not\_domestic"

"batteries\_domestic"

"burglaries"

"thefts"

"mv\_thefts"

"prostitution"

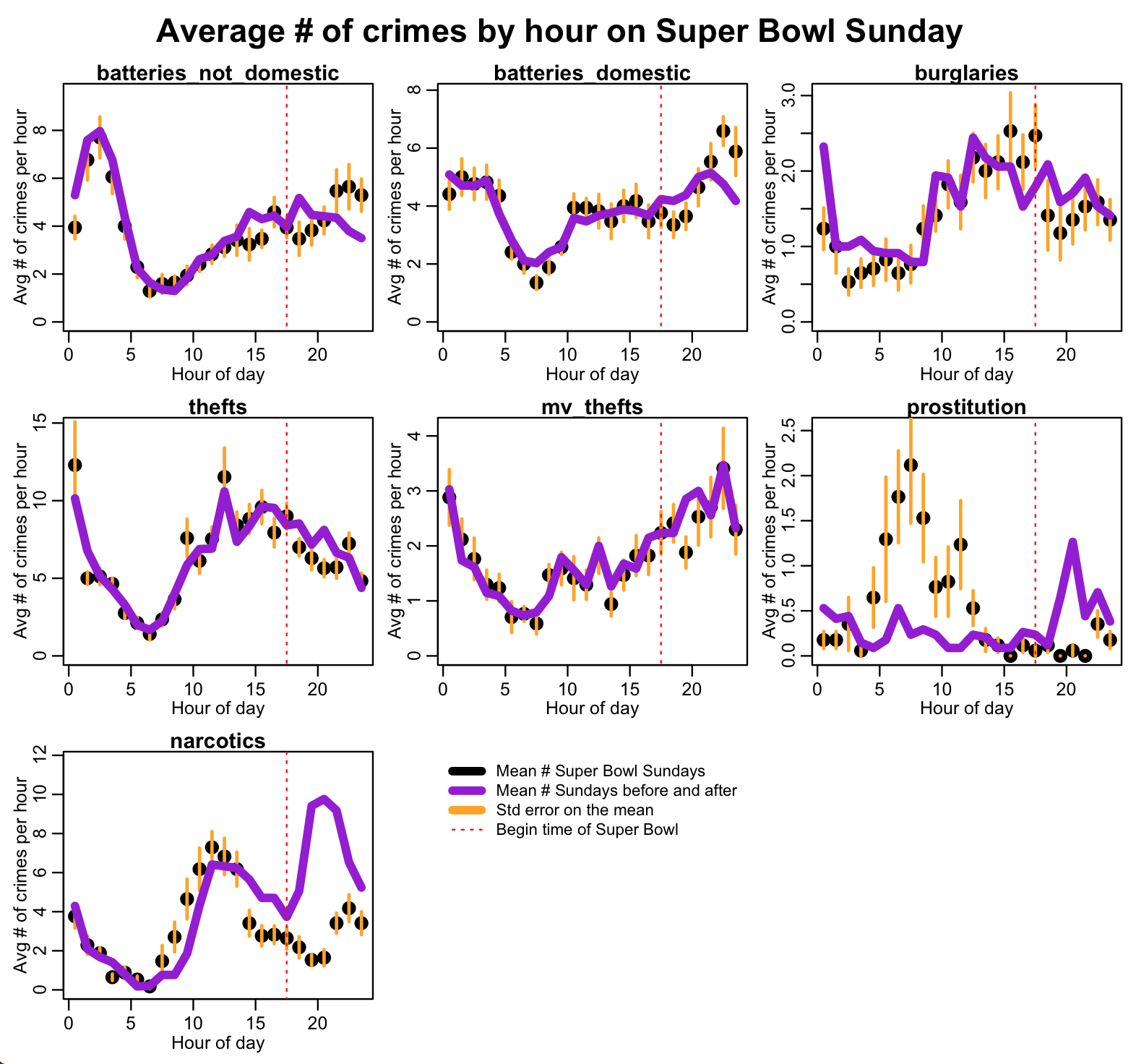
"narcotics"

For each crime type, use the R aggregate() function to get the mean number of crimes by hour for Super Bowl Sundays, and also the mean number by hour for the week before and week after. Average the results for the week before and week after.

Also calculate the standard error on the mean for the number of crimes by hour on Super Bowl Sunday. The base package of R does not include a standard error function. Either write one yourself using the formula SE=sigma/sqrt(N) (where N is the number of data points, and sigma is the standard deviation), or use a function like the [std.error() function in the R plotrix package](https://www.rdocumentation.org/packages/plotrix/versions/3.7/topics/std.error). Recall that the standard error is the uncertainty on the sample estimate of the mean.

Note that the Super Bowl starts close to 6:30 pm (18:30) EST, which is 17:30 Chicago time. Super Bowl parties tend to start between around 2pm to 5pm.

f) Using all of the above, re-create the following plot (using different colours than what I’ve used here):



Some hints:

In order to make a blank plot in R over which you can overlay a legend, use the R command

plot(0,axes=F,xaxt="n",yaxt="n",col=0,xlab="",ylab="")

(Sometimes it is nice to have the legend by itself, rather than overlaid on one of the plots)

To overlay the lines denoting the standard deviation, calculate the minimum and maximum of the ends of the line using the mean-SE and mean+SE. Loop over the hours, plotting each of the lines using the R lines() function.

Other comments:

This isn’t a statistical analysis yet, per se, but this “meet and greet” with our data shows compelling evidence that some types of crimes show statistically significant unique patterns on Super Bowl Sundays. Particularly because we see patterns by hour where several consecutive points are systematically above or below the baseline average.

Later in this course, we’ll discuss ways that we can assess the statistical significance of these patterns.

Also, note that we have shown the uncertainty on the mean of the number of crimes by hour on Super Bowl Sunday, but haven’t shown the uncertainty on our baseline estimate.

g) Using either Google Scholar or Google itself, find commentary in the literature or anecdotally in the news about patterns of various types of crime on Super Bowl Sunday (particularly for the types of crime where we appear to see what are likely significantly different patterns compared to the baseline expectation).

If the source of information is a website, provide the URL of the website, and summarize the commentary, and the sources of data upon which the website draws its conclusions.

If the source of information is a published paper, summarize the key findings of the paper using the Lacum et al rubric.

Are the patterns we see in the Chicago data consistent with the patterns suggested by the information you’ve found?

**Question 3)**

By making minor changes to the R code in Question 2, make the following plots for Mothers’ Day and Fathers’ Day, using different colours than what I’ve used here.

Comment on the similarities and differences in what appear to be significant deviations in some patterns of crime on Mothers’ Day compared to Father’s Day, and what factors you think might underlie the patterns.

5 bonus points if you can think of a plausible explanation for the patterns of prostitution early in the day on Mothers’ day… because I can’t.

