Assignment 1, Due 31 March 5PM.

You will use the file Ozone3.csv. This is the ozone data we discussed in class with Detroit and Des Moines removed. It is arranged so that the sites are rows; we treat these as the observations. The columns are days; we treat these as the variables.

Please upload a file produced by kniting an R markdown file (any of word, html, or pdf are acceptable) showing your code, your plots, and your answers to the questions below. I strongly suggest "knitting" the file after every piece of analysis. Do not wait until the submission day to try knitting!

1) (4 marks for the plot and code, 2 for comments) Make a boxplot showing all the days, in chronological order (this is the order they are listed). Comment on any trends or outliers. Make sure you have appropriate title and axis labels.
2) (4 marks for the plot and code, 2 for comments) One way to get a sense of skewness would be to look at the difference between median and mean. Compute this for each day, and make a histogram of the 89 resulting numbers. Comment on any interesting features
3) (4 marks for the plot and code) Plot the mean for each day as a time series (joined up with lines).
4) (5 marks) Should we compute the principal components of this dataset based on the correlation matrix or covariance matrix (ie with or without scaling the data)? Explain your reasoning.
5) (4 marks for plot and code, 2 for comments) Make an image plot of your chosen matrix. Comment on any features.
6) (4 marks for the plot and code, 2 for explanation, 1 for computation of % variability) Compute the principal components. Show the scree plot. How many components do you think we should use? Explain your reasoning. How much of the variability does this number of components represent?
7) (2 marks for plot/code , 2 for comments) Plot the first two principal component scores. Comment on interesting features.
8) (6 marks for the plots—there will be at least 3—6 marks for comments) If there are any outliers, plot the time series for them and comment on how they differ from the mean. If there are groups (or arms of a horseshoe), plot a representative from the center of each group.

Total : 50 marks.