STAT 530 Exam 2 - Due by 5:00pm Wednesday, December 12th

- The exam should be turned into me, or the secretary in room 216; it should not be left in my mailbox.
- For this exam you may use your notes, any text or reference book, the course web page, and either SAS or R.
- You may not discuss the problems with anyone (especially your fellow students or other instructors) except me.
- You must turn in the code used to generate the output.
- The exam is worth 50 points total. All five questions are weighted equally.

The data set used for all of the following questions can be found at:
http://www.stat.sc.edu/~habing/courses/data/census.txt

The data is compiled from the U.S. Census web site at: http://www.census.gov/Press-Release/metrolis.htm and concerns 272 metropolitan statistical areas. (Yuma, Arizona was deleted because it was missing information on several of the variables.) The data set includes 20 variables for each of these areas.

1 - City - Name of the Metropolitan Statistical Area
2 - State - The State Associated With the Primary City in the MSA
3 - Population - The Estimated Population in 1997
4 - PopChange - The Percentage Change in Population from 1990 to 1997
5 - PopDens - The Population Density in Persons/Square Mile
6 - Under5 - The Percentage of the Population that was Under 5 Years of Age in 1996
7 - Over65 - The Percentage of the Population that was 65 Years of Age and Older in 1996
8 - Asian - The Percentage of the Population that is of Asian or Pacific Islander Ancestry
9 - Black - The Percentage of the Population that is of African Ancestry
10 - Hispanic - The Percentage of the Population that is of Hispanic Ancestry
11 - Birth - Births per 1,000 Population in 1994
12 - InfantD - Infant Deaths per 1,000 Live Births in 1994
13 - CarD - Motor Vehicle Accident Deaths per 100,000 Population in 1992
14 - HeartD - Cardiovascular Disease Deaths per 100,000 Population in 1992
15 - Income - Per Capita Income in 1994
16 - Poverty - Poverty Rate in 1993
17 - Unemploy - Unemployment Rate in 1996
18 - Grants - Federal Grant Dollars per Person in 1996
19 - BankpP - Bank Deposits per Person in 1996
20 - HousepP - Housing Starts per 1,000 People 1990-1996

One R command that may be useful is:

citynames<-substring(as.character(census[,1]),1,4)

This will create a vector containing the first four characters of the first column of the table called census. For purposes of generating graphics, it may be useful to replace the first column in the original data set with these abbreviated names instead.
1) Conduct an exploratory factor analysis (using the varimax rotation) on this data set to see what set of latent factors may possibly underlie the demographics of metropolitan statistical areas.

Say how you determined how many factors to use.

Label (describe) the factors to the best of your ability.

Which variables (if any) do not have high loadings on any of the factors?

Draw a path diagram showing the factors you found in the exploratory factor analysis and which of the variables they influence.

2) Consider the following two sets of variables:

Population Density and Distribution = Population, PopChange, PopDens, Under5, and Over65

and

Quality of Life = InfantD, CarD, HeartD, Income, Poverty, Unemploy, Grants, BankpP, HousepP

Perform the appropriate multivariate analysis to determine what the strongest linear relationships are between these two sets of variables.

Describe what those relationships are in terms of the variables, say how you measured if the relationships were statistically significant, and give a measure of the strength of the relationships.

3) Define a “Large Metropolitan Area” to be a metropolitan statistical area with a population of 1,500,000 or greater.

In many cases it is desirable to have several cities that are demographically similar so that studies concerning products and advertising may be carried out.

Find a group of four large metropolitan areas such that they do not differ too greatly on any of the 18 demographic variables.

Make a table showing how those four areas compare on each of the 18 variables.

Briefly justify your choice of method to select the group. Be sure to include a justification for any options that you had to specify in carrying out the analysis.
4) When attempting to show the degree of similarity or difference between a large number of observations, it is often important to display the results in as intuitive a manner as possible.

Construct a two-dimensional map of the large metropolitan areas such that the distances on the map reflect the dissimilarity of the metropolitan areas on the 18 demographic variables.

Use an appropriate statistic to determine how many dimensions are actually required to adequately represent the data.

Show that there is at least one pair of cities that appear to lie very near each other in two dimensions, but are evidently more distant when a third dimension is added.

5) Define a “Small Metropolitan Area” to be a metropolitan area with a population of 200,000 or less. In choosing cities in which to conduct research, it would be useful to know what the largest differences are between small metropolitan areas and large metropolitan areas besides simply the population.

Provide the best measure of how accurate classical linear discriminant analysis is for distinguishing between small metropolitan areas and large metropolitan areas using the demographic variables (except population).

Construct a classification and regression tree to distinguish between small metropolitan areas and large metropolitan areas using the demographic variables except population. Describe briefly how the tree should be read.

Why is the error rate reported for the CART tree misleading?