There are 5 questions worth a total of 100 points. Your answers are due **Monday May 8 at 4:00pm**. You may turn in your answers to my mailbox in Education or my office in Education (rm 236C). If you would like to receive your graded exam (with comments) and course grade, include a self-addressed envelope with your exam.

The data for problems 1, 2, and 3 along with SAS code to create the datasets are on the course web-site.

Be sure to write legibly and/or type using font size 12pt or larger. Show your work and explain your reasoning. You should include an appendix to your exam that contains computer programs (SAS, R, Stata, and/or MATLAB — No SPSS). Partial credit will be given.

You are not to discuss these questions, your answers, any problems you encounter while working on the questions, or anything pertaining to this exam with your classmates or anyone else until after May 5th. All the work that you turn in, including any computer work, must be your own. If you have questions or problems on the final exam, you should ask the instructor (me).

Good luck & have fun!
1. (25 points) The data for this problem comes from Placed Rated Almanac by Richard Boyer and David Savageau copyrighted and published by Rand McNally. The data set and SAS to create the data are on the course web-site. The variables that you will use in this problem are

- Climate & Terrain
- Housing
- Health Care & Environment
- Crime
- Transportation
- Education
- The Arts
- Recreation
- Economics

(a) Also in the data set are the regions where a city is located. Do regions differ with respect to the 9 variables listed above?

(b) If they differ, how do they differ?

(c) Given values on these 9 variables, can you predict what region the city belongs to? Report how you did the classification and report errors of classification.

(d) What region would you like to live in?

Note that I created the region variable based on the following figure.
2. (25 points) In the state of Illinois (and other government bodies), civil servant positions have a number of different ranks or classes based on years in a position, passing tests, and such. The data for this problem consist of compensation for civil servant employees.
in the State of Illinois from 2013. The SAS code and data are on the course web-site. The forms of compensation in the dataset are

- Total Pay 2012
- Budgeted Salary 2013 (this is what employees was actually paid in 2013)
- Health Ins (health insurance)
- Grand Total Salaries 2013

Note that these have been re-scaled by 100,000 (see top of SAS code).

(a) Can you detect the different rankings based on forms of compensation? Describe the multivariate approach you took for this problem and why you choose this one. How many different ranks or classes did you find?

(b) How do the rankings differ in terms of the 4 variables given above?

(c) One of the variables in the data set is department location, which is basically the position (e.g., human services, fire, city council, etc). Within each rank, which positions have higher compensation packages and which have lower ones?
3. (25 points) The data for this problem come from Deborah Guber that I downloaded from the web. The data with SAS code to create a SAS dataset are available from the course web-site. Below is a description of the data that came with it.

**NAME:** Getting What You Pay For: The Debate Over Equity in Public School Expenditures  
**TYPE:** Census  
**SIZE:** 50 observations, 8 variables  
**DESCRIPTIVE ABSTRACT:** This dataset contains variables that address the relationship between public school expenditures and academic performance, as measured by the SAT.  
**SOURCE:** The variables in this dataset, all aggregated to the state level, were extracted from the 1997 Digest of Education Statistics, an annual publication of the U.S. Department of Education. Data from a number of different tables were downloaded from the National Center for Education Statistics (NCES) website (Is/was found at http://nces01.ed.gov/pubs/digest97/index.html) and merged into a single data file.  
**VARIABLE DESCRIPTIONS:** The first column contains the variable name used in the SAS that I included with the data.  

<table>
<thead>
<tr>
<th>SAS Columns</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>state</td>
<td>Name of state (in quotation marks)</td>
</tr>
<tr>
<td>exp-pp</td>
<td>Current expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of dollars)</td>
</tr>
<tr>
<td>ave-pt</td>
<td>Average pupil/teacher ratio in public elementary and secondary schools, Fall 1994</td>
</tr>
<tr>
<td>salary</td>
<td>Estimated average annual salary of teachers in public elementary and secondary schools, 1994-95 (in thousands of dollars)</td>
</tr>
<tr>
<td>taking</td>
<td>Percentage of all eligible students taking the SAT, 1994-95</td>
</tr>
<tr>
<td>ave-v</td>
<td>Average verbal SAT score, 1994-95</td>
</tr>
<tr>
<td>ave-m</td>
<td>Average math SAT score, 1994-95</td>
</tr>
<tr>
<td>ave-tot</td>
<td>Average total score on the SAT, 1994-95</td>
</tr>
</tbody>
</table>

(a) (20 points) Are characteristics of state expenditure, average pupil/teacher-pupil ratio, salary of teachers, and the percentage of students taking the SAT (i.e., exp-pp, ave-pt, salary, taking) related to the average verbal and math SAT scores (i.e., ave-v, ave-m)? If so, describe how they are related? Are these results surprising?  
(b) (5 points) What (if any) assumptions did you make in answering part (a)? If you made any assumptions, were they reasonable?
4. (15 points) We have considered a number of multivariate techniques for testing hypotheses that apply to the case where we have two sets of variables. Suppose that one set of variables consists of \( p \) continuous/numerical variables (e.g., various achievement test scores) and the other set consists of \( q \) dummy coded variables that define groups or populations (e.g., combinations of gender and high school program type).

(a) If you did a canonical correlation analysis using these two sets of variables and rejected the hypothesis that the first canonical correlation equals zero, does this tell you whether you would reject or retain each of the following two tests? Explain your reasoning for each.

- \( H_0 : \Sigma_{12} = 0 \) where \( \Sigma_{12} \) is the \((p \times q)\) covariance matrix between the two sets of variables.

- \( H_0 : \mu_1 = \mu_2 = \ldots = \mu_g \) where \( \mu_i \) is a \((p \times 1)\) vector of means for the \( i \) group and the number of groups defined by the variables in the second set equals \( g \).

(b) What does this tell you about the relationship between the corresponding multivariate techniques?

5. (10 points) Consider the multivariate hypothesis

\[ H_0 : \mu_1 = \mu_2 \]

where \( \mu_1 \) and \( \mu_2 \) are \((p \times 1)\) mean vectors from two independent groups.

**Fact:** To test this multivariate hypothesis you could use Hotelling’s \( T^2 \) or you could (equivalently) perform a single univariate \( t \) test for two independent groups.

Explain how this fact can be true and show the equivalence.

**Hint:** The union intersection principle is key here. The answer requires you to put together facts that we’ve covered in different parts of the course.