CSC334/424
Assignment #2 (DUE SUNDAY, April 15th by Midnight)

Deliverables: Turn in your answers in a single PDF file or Word Document. Use KnitR or Copy any R output relevant to your answer into your document and explain your answer thoroughly and include a copy of the full analysis in your report along with your conclusions. Also, provide your R code files.

1) (Due by Friday, April 13th) (10 Points) Post to the final project forum with the following:

- Subject Area
- Source of Data
- Specific dataset(s)
  description of its scope (# metric variables, #categorical variables, #samples, multiple related tables?)
- Group Members(s)
- Technology group plans to use for Project

In addition, as you are forming your groups, remember the following requirements for datasets and groups:
a. Your group should have 4-5 people in it.
b. Your group should have at least one **in-class** student. This helps me check in with each group if I have at least one in-class student in each group.
c. Your dataset should be a real and rich dataset with at least 15 to 20 variables mixed between categorical and metric. It should have at least (10 * #var, but better yet 20 * #var) samples (we will see that some techniques like PCA require this for significance/stability). You will need a large sample size if you have a large n umber of variables. See me if you have any doubts about your dataset.

Groups need to be finalized at this time with a chosen dataset(s).

2) (10 points) Answer each of the following questions:

a) What are the advantages and disadvantages of using ridge regression and lasso regression? How are these regressions different?

b) What are some causes of overfitting? How do we diagnose and treat overfitting in regression models?

c) What is multicollinearity? How do we diagnose and treat multicollinearity in regression models?

3) (Paper review 1) (10 Points) An academic paper from a conference or Journal will be posted to the Homework 2 content section of D2L. Review the paper and evaluate their usage of Principal Component Analysis. In particular address the following: **(See article on Patient Safety Questionnaire)**

- How suitable is their data for PCA?
- How are they applying PCA? Are they trying to extract interpretable underlying variables, or is their goal more along the lines of dimensionality reduction?
- What kind of factor rotation do they use if any?
- How many components do they concentrate on in their analysis?
- How do they evaluate the stability of the components (i.e. factorability)?
- What conclusions does PCA allow them to draw?

4) (Paper review 2) (10 Points) An academic paper from a conference or Journal will be posted to the Homework 2 content section of D2L. Review the paper and evaluate their usage of Factor Analysis. In particular address the following: **(See article on Social Media and Social Networking)**

- How are they applying Factoring Analysis?
- What kind of factor rotation do they use?
- How many factors do they concentrate on in their analysis? How did they arrive at these number of factors?
- Explain the breakdown of the factors and the significance of their names.
- How do they evaluate the stability of the components (i.e. factorability)?
- Do they use these factors in later analysis, such as regression? If so, what do they discover?
- What overall conclusions does Factor Analysis allow them to draw?

5) (Principal Component Analysis - 20 points): The data given in the file 'bfi.csv' is the16 multiple choice ability items taken from the Synthetic Aperture Personality Assessment (SAPA) web based personality assessment project. Techniques such as Principal Component Analysis (PCA) can be used to determine different types of personalities. There are 1,525 subjects in the file and 16 variable items as follows:

VarName	Item
A1	Am indifferent to the feelings of others.
A2	Inquire about others' well-being.
A3	Know how to comfort others.
A4	Love children.
A5	Make people feel at ease.
C1	Am exacting in my work.
C2	Continue until everything is perfect.
C3	Do things according to a plan.
C4	Do things in a half-way manner.
C5	Waste my time.
E1	Don't talk a lot.
E2	Find it difficult to approach others.
E3	Know how to captivate people.
E4	Make friends easily.
E5	Take charge.
N1	Get angry easily.
N2	Get irritated easily.
N3	Have frequent mood swings.
N4	Often feel blue.
N5	Panic easily.
O1	Am full of ideas.
O2	Avoid difficult reading material.
O3	Carry the conversation to a higher level.
O4	Spend time reflecting on things.
O5	Will not probe deeply into a subject.
gender	males=1, females=2
education	in HS, fin HS, coll, coll grad , grad deg
age	age in years

Run the data without gender, education, and age.

A) How many components are need to explain 100% of total variation for this data? How many components are determined from the scree plot? What number of components would you use in the model?

B) For the number of components in part A, give the formula for each component and a brief interpretation after rotating the components. What names might you give for each of the components?

C) What subjects have the highest and lowest values for each principal component (only include the number of components specified in part A. For each of those subjects, give the principal component scores (again only for the number of components specified in part A).

D) Finally, run a common factor analysis on the same data. What difference, if any, do you find? Does the factor analysis change your ability to interpret the results practically?

6) **(Principal Component Analysis - 20 points)** Begin with the "census2.csv" datafile, which contains census data on various tracts in a district. The fields in the data are

- Total Population (thousands)
- Professional degree (percent)
- Employed age over 16 (percent)
- Government employed (percent)
- Median home value (dollars)

a) Conduct a principal component analysis using the covariance matrix (the default for prcomp and many routines in other software), and interpret the results. How much of the variance is accounted for in the first component and why is this?

b) Try dividing the MedianHomeValue field by 100,000 so that the median home value in the dataset is measured in $100,000's rather than in dollars. How does this change the analysis?

c) Compute the PCA with the correlation matrix instead. How does this change the result and how does your answer compare (if you did it) with your answer in b)?

d) Analyze the correlation matrix for this dataset for significance, and also look for variables that are extremely correlated or uncorrelated. Discuss the effect of this on the analysis.

7) **(Principal Component Analysis - 20 Points)** Download the "GSS_2002_Health_PCA.xlsx" dataset and perform a principal component analysis on the data. In 2002, the General Social Survery by NORC provided a ballot of questions related to doctors and general healthcare answered by 2,765 respondents. In this dataset, there is a subset of the survey including 34 variables answered by the 2,765 respondents. Please find the data dictionary for the variables in the second datatab entitled DataDict. Note: Some variables are in DIFFERENT UNITS or in DIFFERENT SCALES!

Choose your PCA method carefully and give a reason for your choice. Your method should account for the differences in scales of the variables. Try different ways of formulating the analysis until you get a small set of components that are easy to interpret.

Finally, run a common factor analysis on the same data. What difference, if any, do you find? Does the factor analysis change your ability to interpret the results practically?