CSC334/424
Assignment #2 (DUE SUNDAY, June 24th by Midnight)

Deliverables: Turn in your answers in a single PDF file. 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.

Problem 1 (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?

Problem 2 (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 Clustering Goal-Driven Security Factors for Protecting Data in Cloud Storage using EFA)**

- 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?

Problem 3 (10 points-Data Ethics or Data Integrity): Using Google Scholar, locate **a journal article**, which discusses data ethics or data integrity in terms of big data in your field of interest. **Write a summary** of the journal article and how it utilizes data ethics or data integrity in **two to three paragraphs**. **Cite the paper in APA format.**

Problem 4 (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 2,800 subjects in the file and 28 variable items as follows:

- ***Use a complete dataset (remove missing values)***
- ***Subset for variables A1-O5***

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?