Note: For this midterm exam, you are not allowed to receive help from anyone except me on the exams. For example, you may not talk to other students about the exam problems, and you may not look at other students’ exams. Violations of this policy may result in a 0 on the exam, an F for the course, and/or punishment by the USC Office of Academic Integrity.

1. You are working as a consulting statistician for a company that has a contract with a medical researcher. She has gathered data on 60 adult female patients for a diabetes study. The variables measured include health and demographic variables for the females. The 7 variables she has are:
   X1 = Number of times pregnant
   X2 = Plasma glucose concentration (based on an oral glucose tolerance test)
   X3 = Diastolic blood pressure (mm Hg)
   X4 = Triceps skin fold thickness (mm)
   X5 = Two-Hour serum insulin (mu U/ml)
   X6 = Body mass index (weight in kg/(height in m)^2)
   X7 = Age (years)

The questions that the researcher would like answered include:
   1. Are there individual females who are highly unusual (in any way) based on the measured health variables, X2 through X6? If so, identify their numbers.
   2. Are there notable associations/relationships between some of the variables? (if so, describe them)
   3. Is there a way to graphically represent the raw data for the 60 patients and draw conclusions about the data set from such a graph?
   4. Can we find a few indices that describe the variation in the data set using a lesser dimension than the original set of variables? If so, what are those indices? Is there a convenient interpretation of any of the indices?
   5. Can we graphically display the data in a low number of dimensions using such indices? What conclusions about the patients (individual patients or groups of patients) can you draw from such a graph?
   6. What are any other potentially interesting aspects of the data set?

You will type a roughly 3-page report detailing your analysis of the data and your conclusions. Keep in mind that the report should be written for two audiences: the medical researcher, who has a sense for numbers but is not an expert in statistics; and your own supervisor at the statistical consulting company, who will be judging you and deciding on your possible promotion based on the statistical competency of the report. Your report should be understandable and meaningful to both audiences.

You may include graphs that illustrate and/or support your findings. (The graphs do not have to count as part of the roughly 3-page length.) Do NOT include computer code within the main body of your report. This will be incomprehensible to the researcher and would only annoy her. You may include such code in an appendix if you wish.

The data for this problem are given at the link “Diabetes Data 60” on the course web page. There is a link to a data file without patient ID numbers and a link to another file with patient ID numbers. Here is some R code that may be helpful in reading in and managing the data:

diab.full <- read.table("http://people.stat.sc.edu/hitchcock/diabetes60dataIDs.txt", header=T)
attach(diab.full)
diab.IDs <- as.character(diab.full[,1])
diab <- diab.full[,,-1]
2. You are working as a consulting statistician for an NFL football team that is preparing for roster planning by looking at a cohort of potential players who entered the NFL draft between 2011 and 2013. Data from the NFL draft combine (which measure certain physical characteristics of the players) have been gathered on 527 players (this data set, which is called nflcombinedata.recent in the R code below, is actually a subset of a larger data set called nflcombinedata.full). The 12 variables measured on each player include four identification-type variables and 8 numerical variables (the data analysis will be done on the numerical data). The four identifying variables are the Year the player was eligible for the draft; the Name of the player; the College the player went to; and the football Position (POS) that the player plays. The 8 numerical variables are: Height.in (player’s height in inches), Weight.lbs (player’s weight in pounds), Dash40Yard (the time in seconds the player needed to run a 40-yard dash), BenchPress (the number of times that a player could lift (in a bench-press exercise) 225 pounds of weight; this is a measure of chest and arm strength), VertLeap.in (the vertical height that the player could jump from a standing position, in inches), BroadJump.in (the horizontal distance that the player could jump from a standing position, in inches), ShuttleRun (the time in seconds that it takes a player to complete a “shuttle run”; this is a measure of quickness and agility), Run3Cone (the time in seconds that it takes a player to complete a “3 cone drill run”; this is a measure of quickness, agility, and speed while turning/changing directions).

For those who are not familiar with football positions, a “Position Abbreviations” glossary can be found at http://scores.nbcsports.com/fb/glossary.asp and some position descriptions can be found at https://en.wikipedia.org/wiki/American_football_positions (if it helps you with the report).

The questions that the team would like answered include:

1. Are there individual players who are highly unusual (in any way) based on the measured physical performance-related variables? If so, identify their names.
2. Are there notable associations/relationships between some of the variables? (if so, describe them)
3. Is there a way to graphically represent the raw data for the 527 players (or interesting subsets of those players) and draw conclusions about the data set from such a graph?
4. Are there a small number of underlying characteristics of players that the observed variables might be connected to? If so, determine how many latent characteristics there seem to be in this set of variables. Also, try to interpret them the best you can, with the aid of statistical techniques.
5. Can we graphically display the data (or interesting subsets of the data) in a low number of dimensions using such latent traits? What conclusions about the players (individual players or groups of players) can you draw from such graph(s)?
6. What are any other potentially interesting aspects of the data set?

You will type a roughly 3-page report detailing your analysis of the data and your conclusions. Keep in mind that the report should be written for two audiences: the football team’s client, who has a sense for numbers but is not an expert in statistics; and your own supervisor at the statistical consulting company, who will be judging you and deciding on your possible promotion based on the statistical competency of the report. Your report should be understandable and meaningful to both audiences.

You may include graphs that illustrate and/or support your findings. (The graphs do not have to count as part of the roughly 3-page length.) Do NOT include computer code within the main body of your report. This will be incomprehensible to the client and would only annoy him. You may include such code in an appendix if you wish.

The data for this problem are given at the link “NFL draft data” on the course web page. (The data are part of a data set originally collected by user daniel.ingham at the StatCrunch website.) On the next page is some R code that may be helpful in reading in and managing the data:
Grading Scale:

Each problem will be worth 30 points, for a total of 60 points. For each problem, your report will be graded based on Writing, Analysis, and Context. For example:

**Writing** (out of 10 points): How organized, clearly written, comprehensible, and grammatically correct is the report? Would the client reading this report be confident that it was written by an educated, well-trained statistical scientist?

**Analysis** (out of 10 points): Were the graphs and data analyses appropriate for the problem? Were the analyses carried out correctly? Were your statistical conclusions about the data set sensible and clearly justified by numerical or graphical evidence?

**Context** (out of 10 points): Were the questions answered in terms of the variables of the data set? Although you are not an expert in the field as your client is, have you attempted to frame your conclusions and interpretations in a subject-matter context rather than treating the data as simply a meaningless set of numbers?