Note: For this final 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 statistician doing oversight for a food-safety and nutrition board, as part of a
governmental investigation. The agency has selected 43 types of cereal produced by three companies
(General Mills, Kellogg’s, Quaker). Eight numerical nutritional characteristics have been measured for
each cereal. These 8 variables are: Calories, Protein, Fat, Sodium, Fiber, Carbohydrates, Sugar,
Potassium. In addition, the data file gives the name of each cereal and the company that produced the
cereal.

The questions that the nutrition board would like answered include:

(1) Are there notable associations/relationships between some of the variables? (if so, describe them)
(2) Is there a way to graphically represent the raw data for the 43 cereals and draw conclusions about
the data set from such a graph?
(3) What are the basic underlying groups that the cereals form? Can you plot the data in a small
number of dimensions, showing the group separation of the cereals?
(4) Can you use the dissimilarities among the cereals to plot the cereals on a 2-dimensional map? If
so, how would you characterize those two dimensions?
(5) Would you say the three companies are the same in terms of their average values for the eight
variables? Are there interesting differences between the three companies’ cereals that can be
displayed using the variables (or combinations of variables) in the data set?
(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. Your
report should be understandable and meaningful to the general public, to the leaders of the cereal industry,
and to statisticians who will be reviewing your report.

You may include graphs that illustrate and/or support your findings (such graphs don’t count against the 3
page limit). Do NOT include computer code within the main body of your report. You may include such
code in an appendix if you wish.

The data for this problem are given at the link “Cereal Data” on the course web page. (Original Data
collected by Chad Dacus). The following code should read the data into R correctly:

cer <- read.table("http://people.stat.sc.edu/hitchcock/cereals.txt", header=T); attach(cer)
cer.data <- cbind(Calories, Protein, Fat, Sodium, Fiber, Carbohydrates, Sugar, Potassium)
cereal.names <- as.character(Brand); cereal.company <- as.character(Company)
2. It is 1983. The Cold War is on everyone’s mind, and Michael Jackson’s *Thriller* is rising up the charts. You are working as a statistical consultant for a conglomerate of American automobile manufacturers trying to use quantitative methods to better understand cars’ performance. The project is vital to prevent American manufacturers from falling behind the newly rising Japanese automobile industry. On 392 cars, the following variables have been measured: mpg (gas mileage in miles per gallon), displacement (in cubic inches), horsepower, weight (in pounds), and acceleration (number of seconds needed to go from 0 to 60 mph). There is a categorical variable describing the origin of the car (1=American, 2=European, 3=Japanese). In addition, the data file contains two columns strictly for identification (not analysis) with the year of the car and the name (make/model) of the car.

The questions that the conglomerate would like answered include:

1. Are there notable associations/relationships between some of the variables? (if so, describe them)
2. We wish to build a model to predict the gas mileage and acceleration based on the displacement, horsepower, and weight. Construct such a model and evaluate its appropriateness.
3. Use your model to predict mileage and acceleration for another car, which has displacement 160 cubic inches, horsepower of 120, and weight 3000 pounds. Can you characterize a degree of certainty for your prediction?
4. It is cheaper simply to measure the weight of a car rather than all of those predictor variables. Determine whether the model could predict the mileage and acceleration just as well using only weight as it could using all three predictors.
5. Would you say that the population of American cars and the population of non-American cars are the same in terms of their average values for the five numeric variables? If not, characterize the differences in the two populations in terms of the variables.
6. Based on this data set, could you come up with a method to classify a new car as American, European, or Japanese, based on the values of the five variables? How well does your method perform in predicting the (known) car origins for the given sample?
7. Consider a car of unknown origin having mileage 30 mpg, displacement 116 cubic inches, horsepower 130, weighing 2460 pounds, and having acceleration from 0 to 60 in 12 seconds. What origin would you predict for this car? How confident are you in the classification?
8. 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 auto manufacturers, who are not experts 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 (such graphs don’t count against the 3 page limit). Do NOT include computer code within the main body of your report. You may include such code in an appendix if you wish.

The data for this problem are given at the link “Car Data” on the course web page. (Original Data stored by Carnegie Mellon StatLib). The following R code may be useful in reading in the data and setting the dataset up for analysis:
cardata <- read.table("http://people.stat.sc.edu/hitchcock/carsdataorigin.txt", header=T);
attach(cardata)

car.data.numeric <- cbind(mpg, displacement, horsepower, weight, acceleration)
American.cars <- cardata[cardata$origin==1,]; Foreign.cars <- cardata[cardata$origin!=1,];
American.indicator <- ifelse(cardata$origin==1, 1, 0)

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?