Math 124 Fall 2004
Assignment 1
Due: Sept 27, 2004

Dr Ben Bolstad

bolstad_math124@bmbolstad.com

http://math124sfsu.bmbolstad.com

This assignment is intended to give you practice using excel to perform some EDA techniques as applied to both real and simulated data. You should submit your solutions to this assignment as a brief written report. Do not submit print outs of raw Excel spreadsheets alone.

Exploratory Data Analysis of cereals dataset

Current research states that adults should consume no more than 30% of their calories in the form of fat, they need about 50 grams (women) or 63 grams (men) of protein daily, and should provide for the remainder of their caloric intake with complex carbohydrates. One gram of fat contains 9 calories and carbohydrates and proteins contain 4 calories per gram. A "good" diet should also contain 20-35 grams of dietary fiber.

Download the datafile *cereals.dat* from the course webpage. Note the datafile is a tab delimited text file and the first row contains column names. This datafile consists of measurements of 10 different variables on 77 different types of breakfast cereals. In particular, the variables measured are:

- Name: Name of cereal

- calories: calories per serving

- protein: grams of protein

- fat: grams of fat

- sodium: milligrams of sodium

- fiber: grams of dietary fiber

- carbo: grams of complex carbohydrates

- sugars: grams of sugars

- potass: milligrams of potassium

- shelf: display shelf (1, 2, or 3, counting from the floor)

Note that a value of -1 for any nutrient indicates a missing value.

Use the exploratory data analysis techniques discussed in class to learn more about this data. In particular, you should consider using histograms or boxplots to explore the data. You may also wish to use scatterplots to look at relationships between variables. Also be sure to report any summary statistics you might compute. Your goal is to discover any interesting features of this data. Some questions you might wish to explore

1. Look at the distributions of the Calories, protein, fat, sodium, fiber and carbo, sugars and potass variables. Comment on each distribution. Do you see any interesting features or outliers?

2. Is there any relationship between the shelf and the grams of sugar? Can you suggest a reason for this? Consider checking this out using boxplots.

3. Investigate the relationship between sugars and calories. If possible also show shelf on your plot. Hint: scatterplot.

Hints: How to draw a boxplot in Excel

Sadly, Excel does not provide a built in function for drawing boxplots. However, with a little bit of work you can create something close (handling the outliers is not possible).

Steps for creating a boxplot in Excel

1. Create a table where the first column contains names min, LQ, median, UQ, max, then each additional column in the table should consist of the name of the data series (or value of the categorical variable) and the values of min, LQ, median, UQ and max (in that order). You can use the functions QUARTILE(,1), MIN, MEDIAN, MAX and QUARTILE(,3) to compute these values.

2. Highlight the whole table, including figures and series labels, then click on the Chart Wizard.

3. Select a Line Chart.

4. At Step 2 plot by Rows, (the default is Columns), then Finish.

5. Select each data series in turn and use Format Data Series to remove the connecting lines.

6. Select any of the data series and Format Data Series; select the Options tab and switch on the checkboxes for High-Low lines and Up-Down bars.

Other Excel hints

You may wish to make sure that you have the *Analysis ToolPak* installed. This will give you more plotting options (including better histograms).

Using Excel on campus

If you do not have access to Excel at home or elsewhere you may use it in a lab on campus.

Name	Location
Math Dept Computer Lab `http://userwww.sfsu.edu/~ralex/drafts/mathlab.htm`	TH 404
John F. True Computer Lab `http://www.sfsu.edu/~doit/labs/24hour.htm` 415.338.1490	Main Floor - J. Paul Leonard Library
Media Access Center `http://www.sfsu.edu/~doit/labs/mac.htm` 415.338.2991	3rd Floor Library