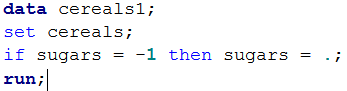
Lab 2

STA 216

Part 1

The first part of this lab will focus on the dataset about cereals from Lab 1. You should start by reading in the dataset using PROC IMPORT and applying labels and formats to the data as in Lab 1. (Note that the labels and formats will show up in your SAS output for this lab too – not just for PROC PRINT.)

1. Focus on the manufacturers of the cereals in the dataset.
   1. Provide SAS output that summarizes the manufacturers numerically. This should not include cumulative frequencies or cumulative percentages.
   2. It usually doesn’t make sense to include cumulative frequencies and cumulative percentages for nominal type variables like manufacturer. It makes a lot more sense to include them for ordinal type variables like letter grade (A, B, C, etc.). Describe the difference between nominal and ordinal type variables and why, of the two, cumulative frequencies and percentages are more appropriate for letter grade.
   3. Provide SAS output that summarizes the manufacturers graphically.
2. We noted in the Lab 1 that most of the cereals marketed to kids are in the middle shelf.
   1. Provide SAS output makes side-by-side boxplots comparing the sugar content of the cereals on each shelf.
   2. Provide SAS output that calculates the mean and 5-number summary of the sugar of the cereals on each shelf and displays only one decimal point.
   3. The boxplot for the middle shelf has no line in it. Why is this?
   4. Looking at (a) and (b), you should notice the cereal with sugar = -1, which we discovered in Lab 1. I think this means that the sugar value for this cereal was unknown, so we’ll change the value from -1 to “missing” (coded in SAS as a period). Use this code using an IF / THEN statement to make the change:  
        
      (The rest of your lab should use this updated dataset.) Then show the result of repeating (b) and write about something that shows that the change in the dataset was made.
3. We noted in the last lab that the cereals with higher ratings tended to have lower sugar tend and higher fiber content.
   1. Provide SAS output of graphs investigating these two relationships: that is,
      1. The relationship between rating and sugar
      2. The relationship between rating and fiber
   2. Show SAS output that calculates the correlations corresponding to both of these linear relationships.
   3. Using the correlation values, describe the strength and direction of these two linear relationships.
   4. Show SAS output that gives the equation of the regression line
      1. For the linear relationship between sugar and rating (rating should be the response variable). Filling in numbers from the output, write the equation of the regression line as rating(hat) = \_\_\_\_\_\_\_\_ + \_\_\_\_\_\_\_\_ \* sugar
      2. For the linear relationship between fiber and rating (rating should be the response variable). Write the equation of the regression line as   
         rating(hat) = \_\_\_\_\_\_\_\_ + \_\_\_\_\_\_\_ \* fiber
   5. Using numbers from the output, fill in the following:
      1. As sugar increases by 1 g, the rating decreases by \_\_\_\_\_\_\_\_\_\_\_\_ units, on average.
      2. As fiber increases by 1 g, the rating increases by \_\_\_\_\_\_\_\_\_\_\_\_ units, on average.

Part 2

Part 2 focuses on the dataset in the Excel file “lab 2 - mtcars.xlsx”. (You can start a new SAS program for this part if you like.) This data was taken from the 1974 *Motor Trend* magazine and contains information on 32 models of automobiles produced in 1973-1974. Descriptions of the variables include:

* model: name of model
* Origin: is car produced in America (domestic) or abroad (foreign)?
* Mpg: Miles per gallon
* Cyl: number of cylinders in the engine
* Disp: Displacement (in cubic inches) of the engine [the volume through which the cylinders move]
* Hp: Gross horsepower produced by the engine
* Drat: Rear axle ratio [ratio of number of gear teeth in axle to driveshaft]
* Wt: Weight in thousands of pounds
* Qsec: Time (in seconds) to go a quarter mile
* Vs: Engine shape (0 = V-shaped, 1 = straight)
* Am: Transmission type (0 = automatic, 1 = manual)
* Gear: Number of forward gears
* Carb: Number of carburetors

1. After using PROC IMPORT to load the dataset into SAS, add labels to each variable (except name, origin, and mpg), and add formats for the variables VS and AM.   
   Show output from PROC PRINT that displays the model name, engine shape, and transmission type for the domestic cars only with labels and formats.
2. Consider the relationship between the variables Origin and Transmission Type.
   1. Provide the SAS output for a two-way table that shows only frequencies (no percents).
   2. Answer the following questions:
      1. Out of all the cars, what percent are foreign?
      2. Out of all the cars, what percent are foreign with a manual transmission?
      3. Out of the domestic cars, what percent have an automatic transmission?
      4. Out of all the cars with a manual transmission, what percent are foreign?
   3. Provide the two-way table that shows all the default output (frequencies, percent, row pct, and col pct) and point out where it shows each of the percentages in part (b).
   4. Make a stacked bar graph for this data.
3. Cylinders is technically a quantitative variable, but it makes sense to treat it as a categorical one in this case because it only takes on three different values.
   1. Show SAS output that numerically summarizes the number of cars with different amounts of cylinders.
   2. Show SAS output of side-by-side boxplots of the displacement of engines with different numbers of cylinders.
   3. There is a strong relationship between number of cylinders and displacement, so much so that there is barely any overlap between the boxplots. Considering the definition of displacement, why do you think this is the case?
   4. Using appropriate numerical summaries of your choice, provide SAS output that compares the centers and spreads of the distributions of displacements for each number of cylinders.