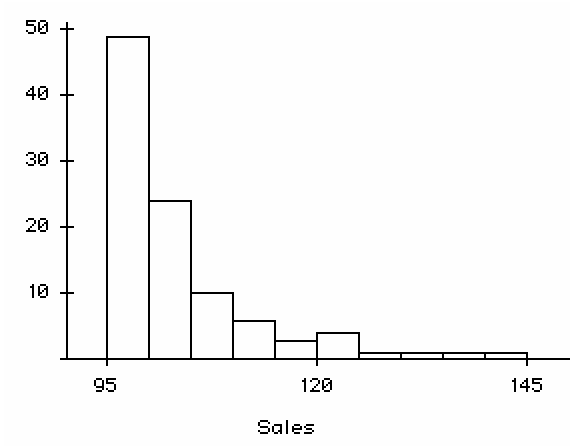
Stat 101 HW#4 Due October 2, 2008

1. **Bread** Clarksburg Bakery is trying to predict how many loaves to bake. In the last 100 days, they have sold between 95 and 140 loaves per day. Here is a histogram of the number of loaves they sold for the last 100 days.



a) Describe the distribution.

b) Which should be larger, the mean number of sales, or the median? Explain.

c) Here are the summary statistics for Clarksburg Bakery's bread sales. Use these statistics and the histogram above to create a boxplot by hand. You may approximate the values of any outliers.

Summary of Sales

Median	100
Min	95
Max	140
25th %tile	97
25th %tile	105.5

d) For these data the mean was 103 loaves sold per day, with a standard deviation of 9 loaves. Do these statistics suggest that Clarksburg Bakery should expect to sell between 94 and 112 loaves on about 68% of the days? Explain

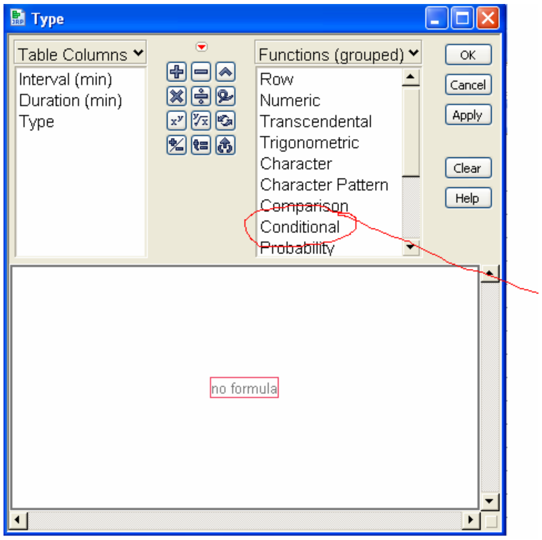
2. **Old Faithful?** Tourists at Yellowstone want to see the geyser Old Faithful erupt. The data set Old_Faithful.jmp shows the *Interval* between eruptions and the *Duration* of each eruption once it starts.

a) Summarize and describe the distribution of *Interval,* the time (in minutes) between eruptions. Be sure to use language that a park ranger can use to tell a tourist how long they'll have to wait until the next eruption.

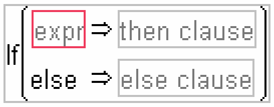
b) Once it starts, how long does the eruption last? Summarize and describe the distribution of *Duration* (in minutes).

c) Compare the distribution of how long you have to wait by whether the eruption is short or long. We'll define short as 3 minutes or less and long as more than 3 minutes. Here's how to create a new variable in JMP:

Double click on the top of the data sheet to the right of Duration. It should say Column 3 – which you can change to say, "Type" or any other name you want. Then *right click* on that top section where the name is and you should see *Formula* as an option (if that doesn't work, double click on the column and go to Column Info --- you can add a formula that way too). Now, once you have the formula dialog box up, it should look like this:

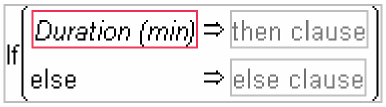


We want to create a categorical variable that labels *Durations* less than or equal to 3 minutes as *Short* and the rest as *Long*. So we first need a *Condition*. Click on *Conditional* and select *If* from the dialog. You'll see the following:



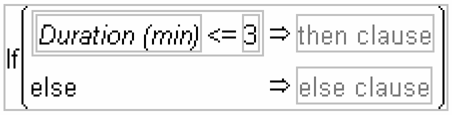
If expr ⇒ then clause
else ⇒ else clause

Now click on *Duration* in the upper left box and it will pop into the red expr box:



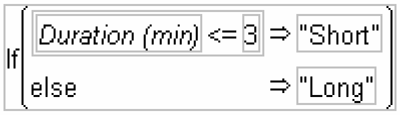
If Duration (min) ⇒ then clause
else ⇒ else clause

To *compare* duration to a number select *Comparison* (right above *Conditional*) and select a <=b. Now put 3 in the box:



If Duration (min) <= 3 ⇒ then clause
else ⇒ else clause

Almost there !! In the then clause box, put "Short" – be sure to put quotation marks around it or it won't understand. (It needs to know you're not making a mistake). Put "Long" in the else clause box.



If Duration (min) <= 3 ⇒ "Short"
else ⇒ "Long"

Click *Apply* and then *OK.*

You should see a column (called *Type*) with values Short and Long.

Now compare the distribution of *Interval* for *Long* and *Short* duration eruptions. Summarize what you see.

3. **Old Faithful Continued.** Continuing exercise 2, what we've just done is treat *Duration* as a categorical variable and looked at the relationship between *Duration* and *Interval* that way. Now, let's look at the relationship between *Duration* and *Interval* treating them both as quantitative. Make an appropriate plot and describe the relationship (remember, direction, form and strength).

   a) Find the regression line predicting how long you'll wait given how long the last eruption was. If you were the guide, how would you summarize that relationship to a tourist who asks the question? (Round the coefficients to make it easier to explain).

   b) How long would you predict you'll have to wait if the last eruption lasted 2 minutes? 1 minute? 10 minutes? Why might you have less faith in the last two predictions? (You can get these predictions in JMP by fitting the line, saving the predictions and then adding rows with *Durations* 1, 2 and 10 minutes.

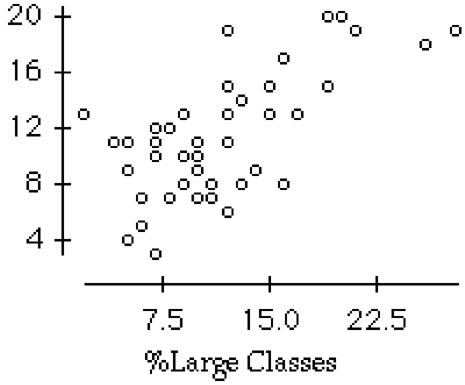
   c) Does this analysis mean you can use duration of the geyser to predict how long you'll have to wait? Why or why not? Does this analysis mean that longer lasting geysers cause the wait time to increase? Explain.

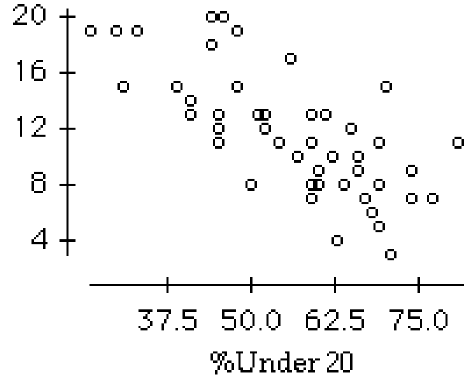
4. **Grades** A statistics professor created a linear regression equation to predict students' final exam scores from their midterm exam scores. The regression equation was: $\widehat{fin} = 10 + 0.9mid$

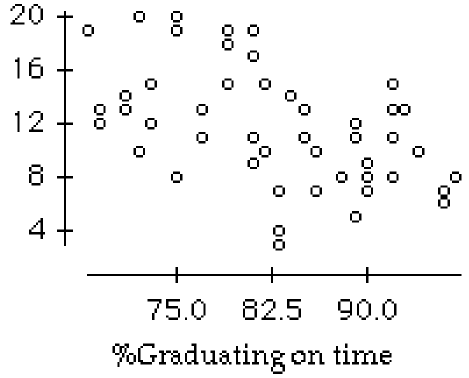
   a) If Susan scored a 70 on the midterm, what did the professor predict for her score on the final?
   b) Susan got an 80 on the final. How big is her residual?
   c) Suppose that the standard deviation of the final was 12 points and the standard deviation of the midterm was 10 points. What is the correlation between the two tests?
   d) How many points would someone need to score on the midterm to have a predicted final score of 100?
   e) Suppose someone scored 100 on the final. Explain why you cannot estimate this student's midterm score from the information given.
   f) One of the students in the class scored 100 on the midterm, but got overconfident, slacked off, and only scored 15 on the final exam. What is the residual for this student?
   g) No other student in the class "achieved" such a dramatic turnaround. If the professor decides not to include this student's scores when constructing a new regression model, will the R-squared value of the regression increase, decrease, or remain the same? Explain briefly.
   h) Will the slope of the new line increase or decrease?

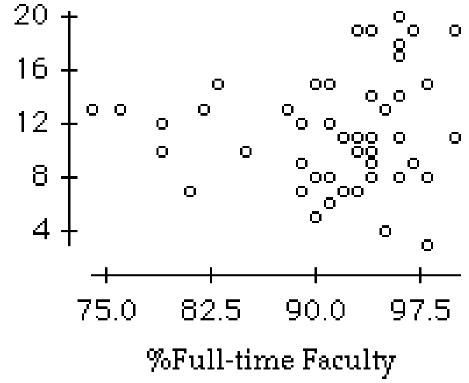
5. **French** Consider the association between a student's score on a vocabulary test and the weight of the student. What direction and strength of correlation would you expect in each of the following situations? Explain.
   a) The students are all in the third grade.
   b) The students are in third through twelfth grades in the same school district.
   c) Why are these answers different? Explain why the answer to b) changes? What's going on?

6. **College** Every year US *News and World Report* publishes a special issue on many US colleges and universities. The scatterplots below have Student/Faculty Ratio (number of students per faculty member) for the colleges and universities the *y*-axes plotted against 4 other variables. The correct correlations for these scatterplots appear in this list. Match them.
   -0.98 -0.71 -0.51 0.09 0.23 0.69









7. **Manatees** Marine biologists warn that the growing number of powerboats registered in Florida threatens the existence of manatees. The Feb 12, 2002 *NY Times* reported the data found in Manatees.jmp .
   (http://query.nytimes.com/gst/fullpage.html?res=9B00E4DD113CF931A25751C0A9649C8B63&scp=1&sq=Manatees&st=nyt)

   a) In this context, which do you think is the explanatory variable?
   b) Make a scatterplot of these data and describe the association you see.
   c) Find the correlation between boat registrations and manatee deaths.

d) Interpret the value of R-squared.

e) Does your analysis prove that powerboats are killing manatees?

8. **A Manatee Model** Continue your analysis of the manatee situation from the last exercise.

a) Create a linear model of the association between manatee deaths and powerboat registrations.

b) Interpret the slope of your model.

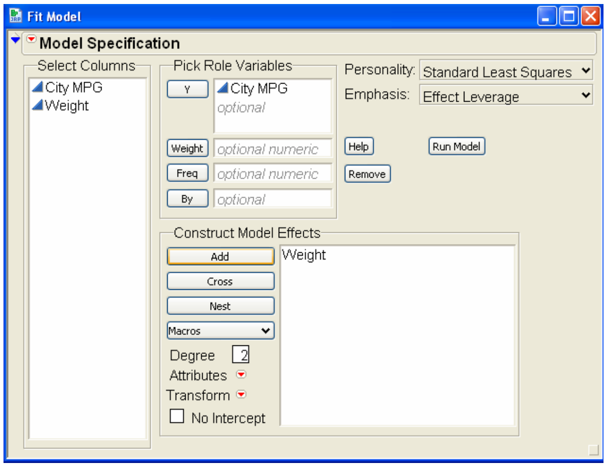
c) Interpret the *y*-intercept of your model.

d) How accurately did your model predict the high number of manatee deaths in 2001?

e) Which is better for the manatees, positive residuals or negative residuals? Explain.

f) What does your model suggest about the future for the manatee?

Transformations. We saw that reciprocal mpg worked much better for seeing the relationship between fuel efficiency and weight of cars. (I've put the *Science* article on the subject on Blackboard). How did I know reciprocal might work well? Because I used the Box-Cox transformation function. Here's how it works. Use **Fit Model** (not **Fit Y by X**) and put the response variable in *Y* and the predictor variable in ***Add.*** (see below). Click Run Model.



Y: City MPG

optional

Weight: optional numeric

Freq: optional numeric

By: optional

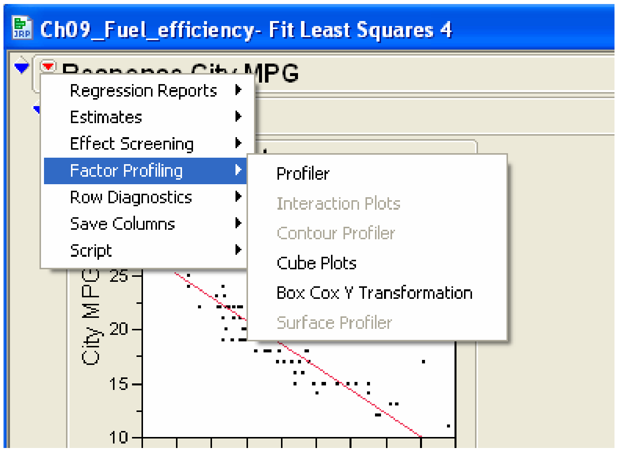
Personality: Standard Least Squares

Emphasis: Effect Leverage

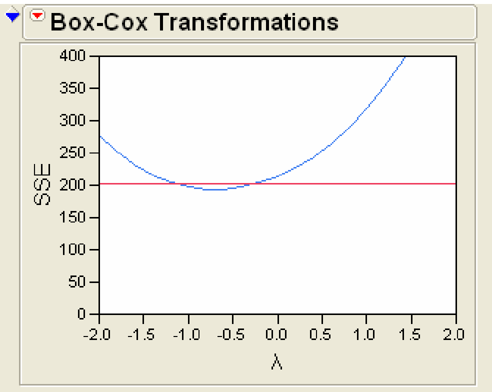
Model effects: Weight

Degree: 2

Now, under the top red triangle you'll see **Factor Profiling → Box Cox Y Transformation:**



Once you click that, you should see:



Below the red line shows all the transformations that should help

1) Straighten the relationship
2) Make the variation more constant over the range of *X*
3) Make the residuals more Normal

Notice that there's a range of transformations. Here from about 0 (log) to -1.0 (reciprocal). I chose reciprocal (instead of -0.5 – 1 over square root) because it's *easier* and makes sense. Typical transformations are square root (0.5) log (0.0 – either log 10 or ln both work) and -1.0 (reciprocal).

9. **Lobsters.** The state of Maine catches and sells the most lobster of any state in the United States. The values of the total lobster catch ($ million) (*Value Million)* for the state of Maine since 1990 are contained in the data set Lobster.jmp.

a) Fit a linear regression to the *Value (in millions of $)* by *Year*. Predict what the value will be in 2008.

b) Save both the residuals and the predicted values and plot the residuals *vs.* the predicted values. What problems with the residuals, if any do you see?

c) What does that say about your linear model?

10. **More lobster.** Using the Box Cox transformation, find a transformation of *Value* that straightens out the relationship with *Year*.

   a) What is the model now?

   b) Do a residual analysis and comment on the improvement.

   c) What is the prediction for 2008?

   d) What is that prediction in millions of $?