## Main Instructions

1. Read about the California Schools data in "caschools-data-description.rtf"
2. Import the comma-delimited data file "caschools.csv" into SPSS (tip: File - Read Text Data)
3. Add variable labels in SPSS for all variables using the information in the data description file (tip: see Variable View)
4. For any variables that are showing too many decimal places, set decimals to 2 in the Variable View
5. Examine frequencies for each variable; categorize the variable type (e.g., numeric or non-numeric, binary, ordinal, etc.) (tip: Analyze - Descriptives - Frequencies)
6. Based on the variable type for each variable identify a couple of appropriate summary statistics
7. For all numeric variables, produce histograms and describe the distribution (e.g., how many modes, skew?, normality?, uniform?, etc.). Explain why the variable is distributed the way it is?
8. Examine a scatterplot of school reading and mathematics performance. Describe the scatterplot. What might cause this result?

## Additional Exercises

1. Create z-score versions of school mathematics and reading performance? (Tip: Analyze - Descriptives - Descriptives - Save standardized values)
2. Create an overall school academic performance variable that is the sum or mean of the mathematics and reading z-scores (Tip: Transform - compute) Which school performed best? Worst?
3. Try a few transformation to make the "students" variable (i.e., the number of students in the school) more normally distributed. (tip: Transform - compute; lg10(students), sqrt(students), etc.).

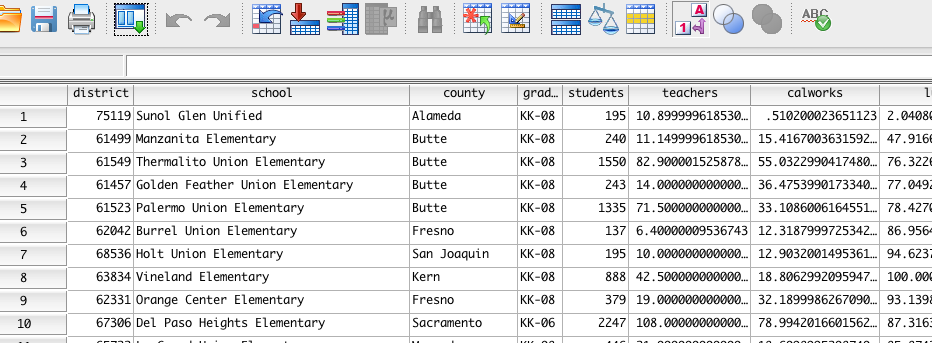
# Answers

## Import the comma-delimited data file "caschools.csv" into SPSS (tip: File - Read Text Data)

SPSS provides the Read Text Wizard for importing text data. In general, when you have your own data, you may have to try the options a few time until you get it right.

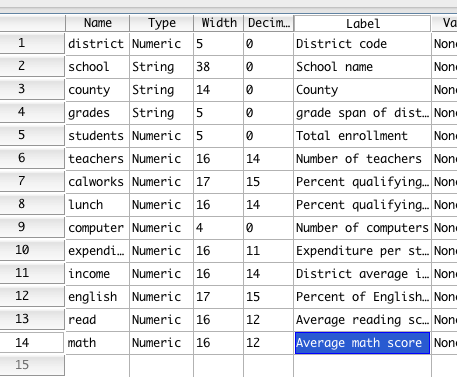
* Use the Read Text Wizard in SPSS to import the data
  + Go to File - Read Text Data or drag and drop the csv file into SPSS
  + How are your variables arranged? "delimited": The data is made up of one row per participant and a comma separates each variable. Thus, the data is delimited by a coma.
  + Are your variable names included at the top of the file? "yes"
  + Which delimiters appear between variables? just select "comma"

Here's a screen shot of the data after importing.



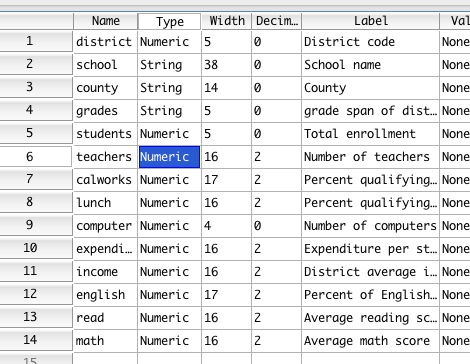
## Add variable labels in SPSS for all variables using the information in the data description file (tip: see Variable View)

Here is the view after copy and pasting in the variable labels



## For any variables that are showing too many decimal places, set decimals to 2 in the Variable View

This makes the display of numbers in tables and the data viewer more appropriate:



## Examine frequencies for each variable; categorize the variable type (e.g., numeric or non-numeric, binary, ordinal, etc.) (tip: Analyze - Descriptives - Frequencies)

Run Analyze - Descriptive Statistics - Frequencies

Put all variables into the variable list and click OK.

Frequencies is a great tool to get a quick overview of univariate features of the data. You can also use it to request descriptive statistics and basic charts.

* district: This is a unordered categorical ID variable
* school: This looks a unordered categorical variable that is a bit like an ID variable, but there are a couple of cases with the same name
* county: Unordered categorical grouping variable
* grades: binary categorical variable
* students: numeric count variable
* teachers: numeric fractional count variable
* calworks: numeric proportion (0 to 100)
* lunch: numeric proportion (0 to 100)
* computer: numeric count variable
* expenditure: numeric ratio variable
* income: numeric ratio variable
* english: numeric proportion (0 to 100)
* read: some form of numeric score; probably interval (i.e., no absolute zero)
* math: some form of numeric score; probably interval (i.e., no absolute zero)

## Based on the variable type for each variable identify a couple of appropriate summary statistics

* For all the numeric variables above the mean, SD, min, and max would be a good start. If the variable is highly skewed as many of the count variable are (e.g. number of computers), you might want to focus more on the median and interquartile ranges.
* District and school don't need to be summarised
* County can be summarised by percentages for each category perhaps; you could also flag the top few categories, or give a sense of the spread of cases per category.

## For all numeric variables, produce histograms and describe the distribution (e.g., how many modes, skew?, normality?, uniform?, etc.). Explain why the variable is distributed the way it is?

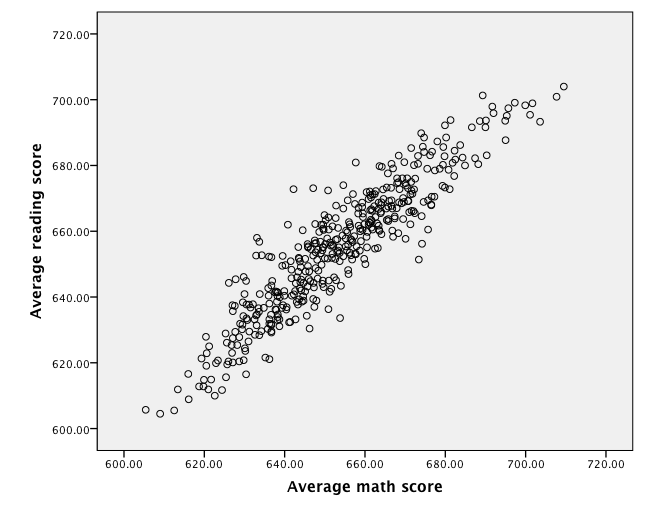
The point of this exercise is to get you thinking about the variables in a dataset. This process is not about saying do I or do I not have normality. It's about identifying the story in the variables and thinking about the mechanisms that might be driving the particular observed distributions.

* students: Highly postively skewed; It's just a general phenomena that there are more small schools but there are the occasional big schools.
* teachers: as per students
* calworks: Strong positive skew with some outliers. Most schools have low levels of income assistance, but it does vary and the proportion gradually declines as you move away from the modal amount which is close to zero. The outliers might be explained by particular areas or small school sizes.
* lunch: There is a spike around zero, and otherwise a slight curved distribution with another peak around 35%.
* computer: as per students; presumably this variable is largely driven by the number of students and the ratio of computers to students would be a more interesting variable.
* expenditure: Approximately normally distributed with slightly bigger tail to the right and a few outliers to the right.
* income: This appears to be a mixture distribution. There is one component that is roughly normal with a mild positive skew, and there is a long tail to the right and a few clear outliers to the right. This presumably reflects the distribution of income in the majority of areas in addition to the existence of particularly wealthy districts. And even when it comes to wealthy districts, there is wealthy and there is super-wealthy.
* english: Similar to calworks
* read: Basically normally distributed as would be expected from a well developed psychometric test of ability
* math: as per "read"

## Examine a scatterplot of school reading and mathematics performance. Describe the scatterplot. What might cause this result?

SPSS has a range of tools for graphing. I'll focus on the legacy tools. Scatterplots are great for learning about the relationship between your variables.

* Go to Graphs - Legacy - Scatter (Simple Scatter)
* X axis: math ; Y axis: read



**Scatterplot description**: The two variables are very strongly positively related. The relationship is linear. There are no major bivariate outliers, although there are a few cases that perhaps a little further away from the mass of data than others.

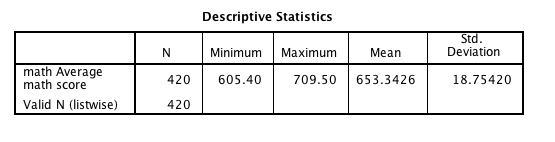
**Why this pattern?** Research shows that math and reading ability are correlated moderately in a given student. This relationship can be explained in terms of many things that are common in the individual particularly intelligence, academic aptitude, motivation to study, parental support, school adjustment, and so on. However, the data plotted is for average school performance. When variables are aggregated relationships often get a lot stronger than at the individual-level, and they can even change. Schools presumably differ in fairly reliable ways in terms of the academic aptitude of the average student that attends. Presumably students would still vary a lot within each school, but the average would be a reliable and meaningful aspect of the school. As for explanations for why this occurs, academic performance is strongly related to socio-economic status, ethnicity, and parental support. The school itself may also contribute.

## Additional Exercises

## Create z-score versions of school mathematics and reading performance? (Tip: Analyze - Descriptives - Descriptives - Save standardized values)

Z-scores involve centring the variable at zero and making the standard deviation equal one. They are a convenient way of making composite scores out of variables on different scales.

There are two main ways to create z-scores in SPSS. You can get the mean and SD for a variable and then use "Transform - compute" menu to write an equation. E.g., analyze - descriptives - descriptives (math)



COMPUTE zmath = (653.3426 - math) / 18.75420.

EXECUTE.

Alternatively, you can just use the "save as standardized variable" option in "Analyze - descriptives - descriptives".

That's what we'll do here.

* Analyze - descriptives - descriptives and place in math and read

## Create an overall school academic performance variable that is the sum or mean of the mathematics and reading z-scores (Tip: Transform - compute) Which school performed best? Worst?

For more information on computing scale scores for ability tests see: <http://jeromyanglim.blogspot.com.au/2009/03/calculating-composite-scores-of-ability.html>

* Transform - compute: Target variable = "academic"; numeric expression = "zmath + zread".

This will create a new variable called "academic" which represents average academic performance in reading and mathematics at the school.

To see the best and worst, you could sort the "academic variable" in SPSS data view by right clicking it.

* The best performing school was Los Altos Elementary.
* The worse performing school was Burrel Union Elementary.

## Try a few transformation to make the "students" variable (i.e., the number of students in the school) more normally distributed. (tip: Transform - compute; lg10(students), sqrt(students), etc.).

Use transform - compute to perform the transformation. Functions are sorted into various categories. In particular, the arithmetic category has many functions of relevance. The resulting command when pasted into syntax could like this:

COMPUTE students\_lg10=lg10(students).

The resulting histogram looks much more normally distributed than the raw variable does. Thus, this might be more appropriate when correlating with other normally distributed variables.

