Data exploration & graphics with SAS

In this exercise, we will work with data on Canadian occupational prestige from 1971, in relation to education, income and % women for 102 occupational categories. The data is found in `N:\data\prestige.sas`, and `prestige.sas`. (You can examine this by opening it in notepad.)

Generally, we would be interested in predicting prestige from the other variables. The focus here is on using graphics (and my SAS macros) to explore variable distributions and relations among them.

1. Read the data into SAS using:
       %include **data**(prestige);
   Examine the data in the viewtable. Note: See the variable names via View -> Column names. Close the viewtable after you have looked at the data.

2. Examine the distribution of variables using boxplots and boxplots for symmetry.
       %***boxplot***(data=prestige, var=prestige, class=type, id=job);
       %***symbox***(data=prestige, var=prestige educ income, id=job);

   What do these options mean? Use **%webhelp(boxplot)** for details.
   Do any variables look obviously skewed, or contain possible outliers? You can also get boxplots using PROC BOXPLOT, e.g.,

   ```
   proc sort data=prestige; by type;
   proc boxplot data=prestige;
           plot prestige * type / boxstyle=schematicidfar;
           id job;
   run;
   ```

3. Examine the bivariate relations among the variables with a scatterplot matrix. Try the following, and observe the effect of the `group=` and `interp=` parameters.
       %***scatmat***(data=prestige, var=prestige educ income women);
       %***scatmat***(data=prestige, var=prestige educ income women, group=type);

       %***scatmat***(data=prestige, var=prestige educ income women, interp=rl);
       %***scatmat***(data=prestige, var=prestige educ income women, interp=rq);

   Note: You can also get scatterplot matrices from PROC CORR, but without my fancy options:

   ```
   proc corr plots=matrix  data=prestige nosimple;
   run;
   ```

4. From the above, consider what transformation(s) of educ and income might make the relation to prestige more nearly linear (ladder of powers and 'arrow rule'). Compose a data step to create new variables using SAS expressions; you can use any names for the new variables.
       data prestige;
          set prestige;
          educ2 = educ**2;
          loginc = log10(income);
       run;

Some possibly useful transformations are: `sqrt(x), log10(x), x**2`. How would you check whether the relations to prestige are more nearly linear?

5. Try a simple regression model or two. Note the use of the `id` statement and the `p` option. Which observations are not well predicted? Does it improve things to add %women to the model?

```
proc reg data=prestige;
    m1: model prestige=educ income /  p;
        * m2: model prestige= … ;   *-- try your transformed vars;
    id job; run;
```

6. In SAS 9.3, all output is shown by default as HTML in the Results Viewer. You can get output in other formats and styles using the Output Delivery System (ODS). Try the following for RTF (Word) output

```
ods rtf file="prestige-tut.rtf";
proc corr data=prestige plots=matrix;
run;

proc reg data=prestige;
    model prestige=educ income ;
              id job;
run;
ods rtf close;
```
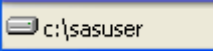
Note that some procedures (like PROC CORR) require an option (PLOTS) to produce ODS GRAPHICS, while others (like PROC REG) do them automatically.

All ODS output is displayed using a given "style". You can change the entire look of all the output using the option `style=` on the `ODS html/rtf/pdf` statement.

The **Default** style uses a grey background. Some other possible styles are **Analysis, Statistical, Journal**. For example, try the example above, using

```
ods rtf file="prestige-tut1.rtf" style=statistical;
… (rest same as above) …
ods rtf close;
```

Finally, note that, unless you specify a complete file path (e.g., file="f:\prestige-tut1.rtf"), ODS writes files and images to the "current directory", indicated at the bottom of the main SAS window, e.g., on my computers. Double-click on the disk icon to change that.