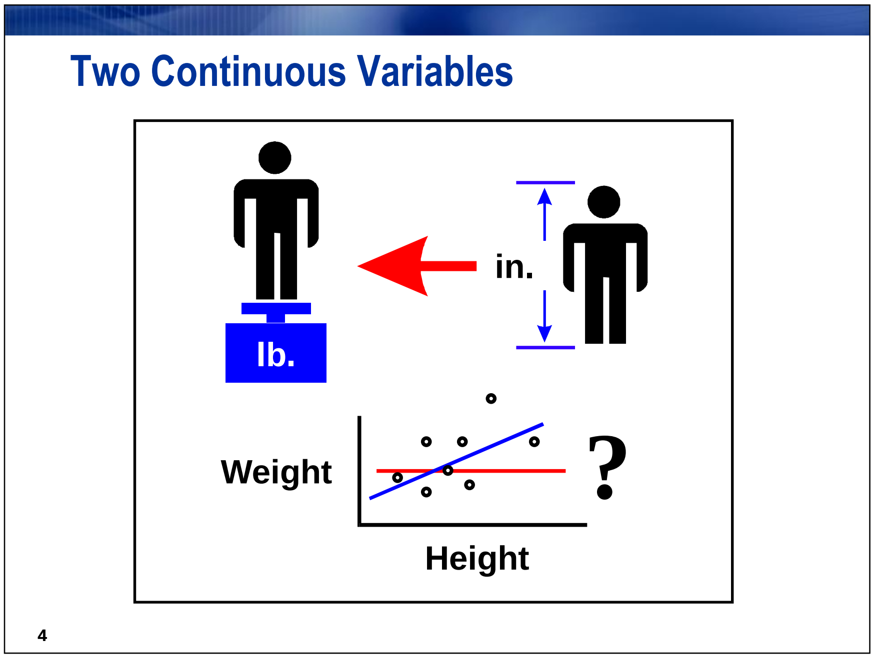
SASEG 8B - Correlation Analysis

<u>Datasets</u> – BodyFat2; Fitness; NewFitness; AbdomenPred

(Fall 2015)

<u>Sources</u> (adapted with permission)**-**

T. P. Cronan, Jeff Mullins, Ron Freeze, and David E. Douglas Course and Classroom Notes
Enterprise Systems, Sam M. Walton College of Business, University of Arkansas, Fayetteville
Microsoft Enterprise Consortium
IBM Academic Initiative
SAS® Multivariate Statistics Course Notes & Workshop, 2010
SAS® Advanced Business Analytics Course Notes & Workshop, 2010
Microsoft® Notes
Teradata® University Network

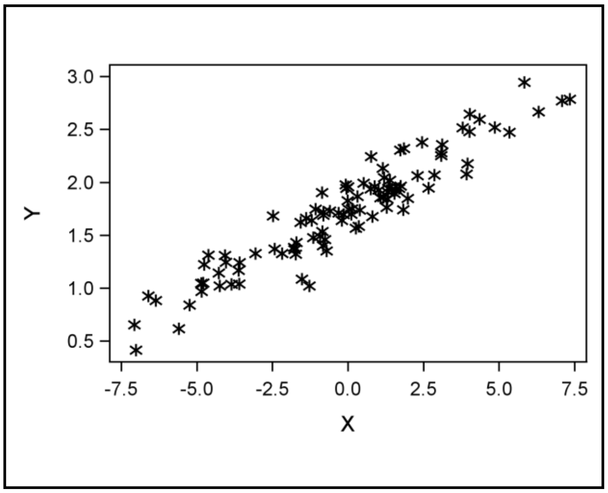


Previously, you learned that when you have a discrete predictor variable and a continuous outcome variable you use ANOVA to analyze your data. In this section, you have two continuous variables.

You use correlation analysis to examine and describe the relationship between two continuous variables. However, before you use correlation analysis, it is important to view the relationship between two continuous variables using a scatter plot.

Example: A random sample of high school students is selected to determine the relationship between a person's height and weight. Height and weight are measured on a numeric scale. They have a large, potentially infinite number of possible values, rather than a few categories such as short, medium, and tall. Therefore, these variables are considered to be continuous.

Scatter Plots

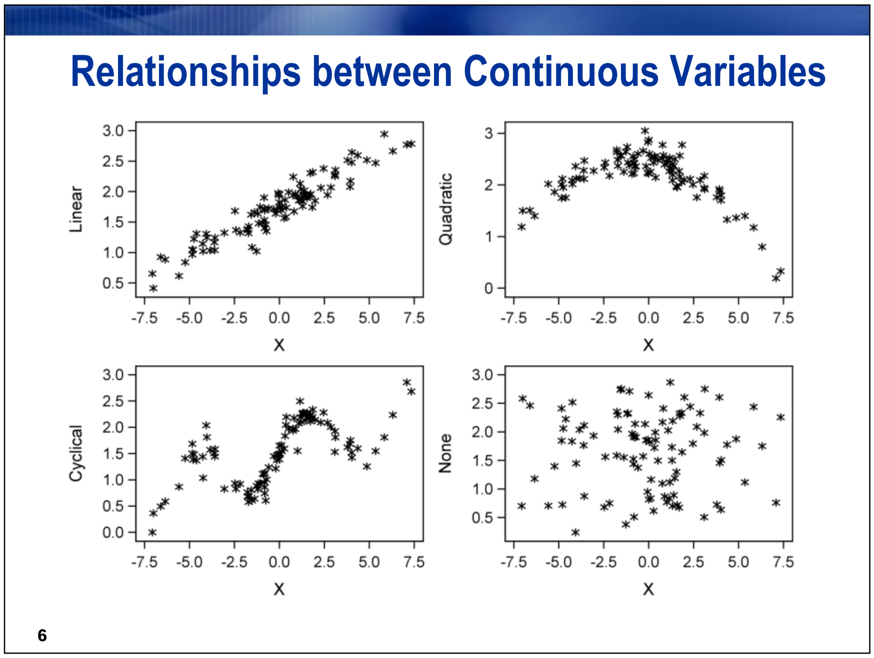


5

Scatter plots are two-dimensional graphs produced by plotting one variable against another within a set of coordinate axes. The coordinates of each point correspond to the values of the two variables.

Scatter plots are useful to

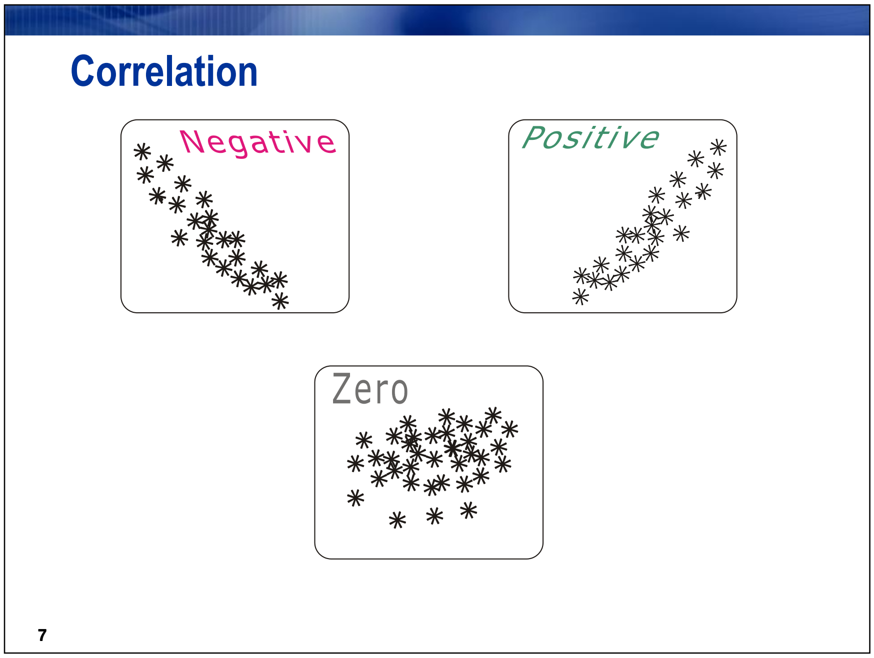
- explore the relationships between two variables
- locate outlying or unusual values
- identify possible trends
- identify a basic range of Y and X values
- communicate data analysis results.



Describing the relationship between two continuous variables is an important first step in any statistical analysis. The scatter plot is the most important tool you have in describing these relationships. The diagrams above illustrate some possible relationships.

1. A straight line describes the relationship.
2. Curvature is present in the relationship.
3. There could be a cyclical pattern in the relationship. You might see this when the predictor is time.
4. There is no clear relationship between the variables.

Correlation between two variables …



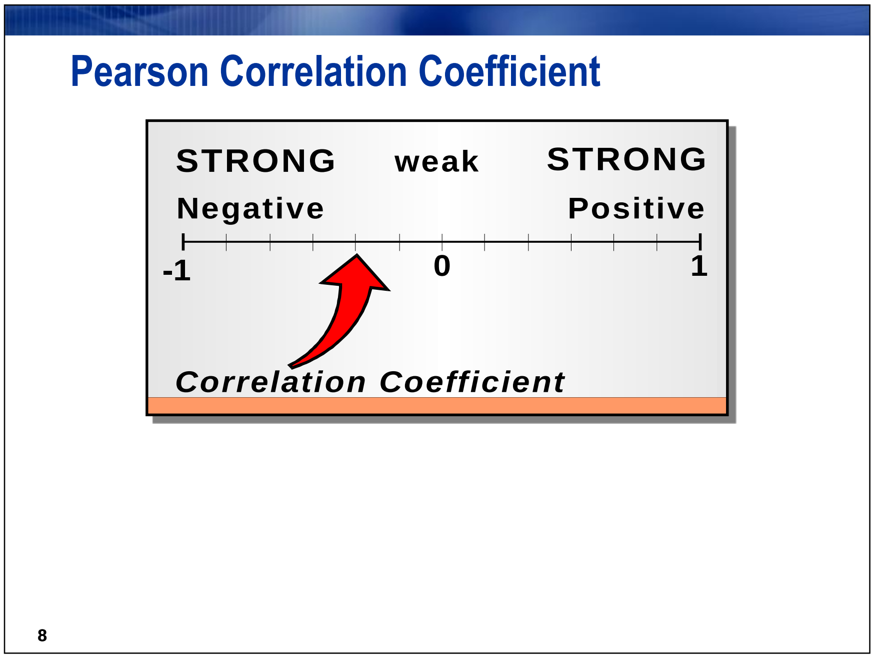
As you examine the scatter plot, you can also quantify the relationship between two variables with correlation statistics. Two variables are correlated if there is a **linear association** between them. If not, the variables are uncorrelated.

You can classify correlated variables according to the type of correlation:

positive one variable tends to increase in value as the other variable increases in value

negative one variable tends to decrease in value as the other variable increases in value

zero no linear relationship between the two variables (uncorrelated)



Correlation statistics measure the degree of linear association between two variables. A common correlation statistic used for continuous variables is the Pearson correlation coefficient. Values of correlation statistics are

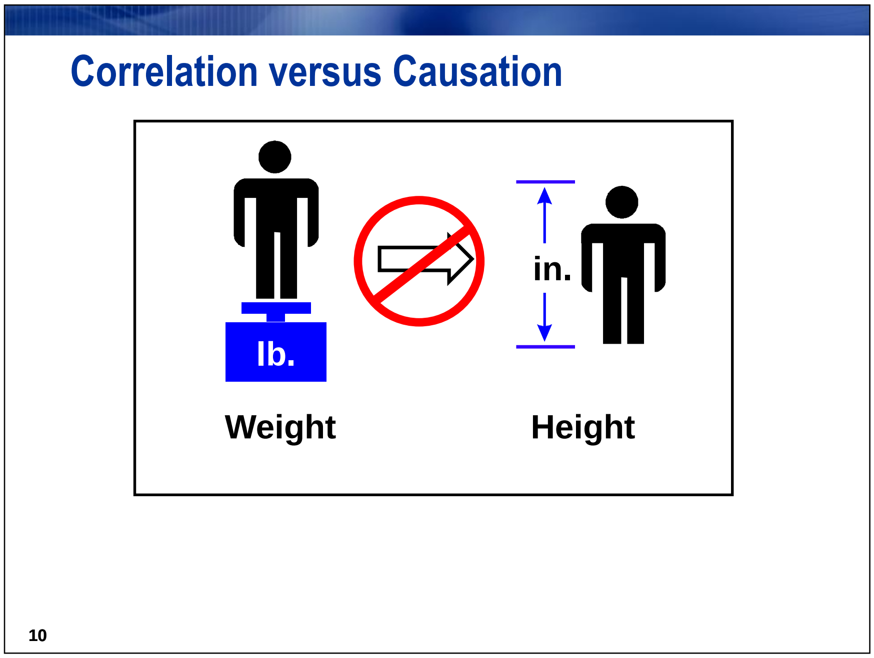
- between –1 and 1
- closer to either extreme if there is a high degree of linear association between the two variables
- close to 0 if there is no linear association between the two variables
- greater than 0 if there is a positive linear association
- less than 0 if there is a negative linear association.

Hypothesis Test for a Correlation

- The parameter representing a correlation is ρ.
- ρ is estimated by the sample statistic r.
- H_0: $\rho=0$
- Rejecting H_0 indicates only great confidence that ρ is not exactly zero.
- A *p*-value does not measure the magnitude of the association.
- Sample size affects the *p*-value.

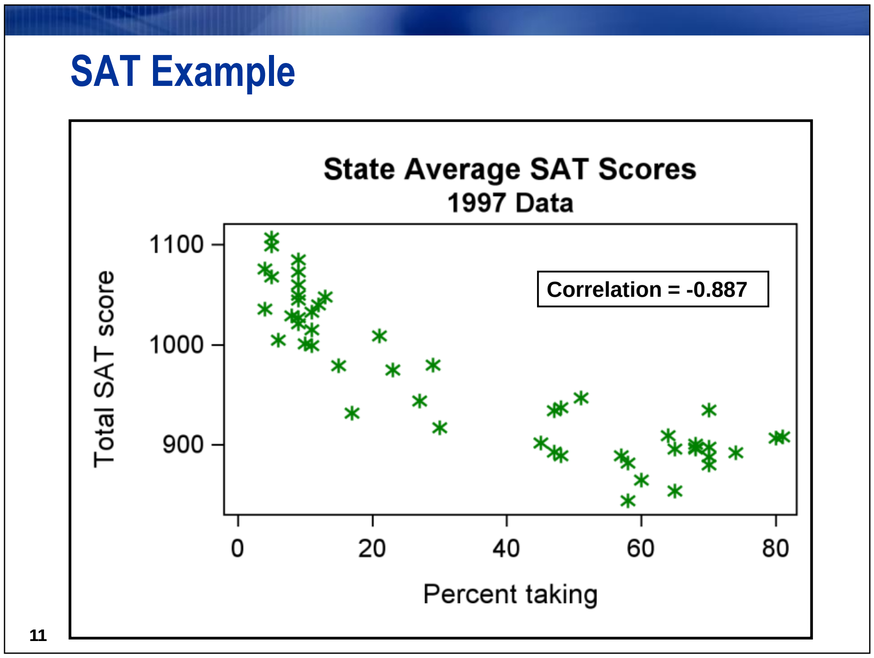
9

The null hypothesis for a test of a correlation coefficient is $\rho=0$. Rejecting the null hypothesis only means that you can be confident that the true population correlation is not 0. Small *p*-values can occur (as with many statistics) because of very large sample sizes. Even a correlation of 0.01 can be statistically significant with a large enough sample size. Therefore, it is important to also look at the value of r itself to see if it is a meaningfully large correlation.



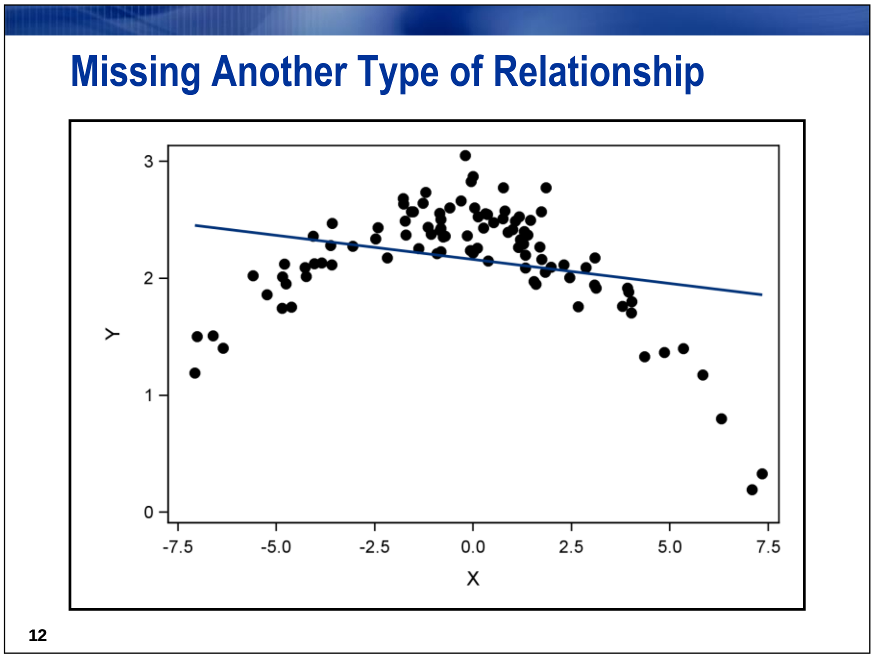
Common errors can be made when interpreting the correlation between variables. One example of this is using correlation coefficients to conclude a cause-and-effect relationship.

- A strong correlation between two variables does not mean change in one variable causes the other variable to change, or vice versa.
- Sample correlation coefficients can be large because of chance or because both variables are affected by other variables.
- “Correlation does not imply causation.”



An example of improperly concluding a cause-and-effect relationship is illustrated using data from the Scholastic Aptitude Test (SAT) from 1997. The scatter plot shown above plots each state's average total SAT score versus the percent of eligible students in the state who took the SAT. The correlation between the two variables is –0.88712. Looking at the plot and at this statistic, an eligible student for the next year can conclude, "If I am the only student in my state to take the SAT, I am guaranteed a good score."

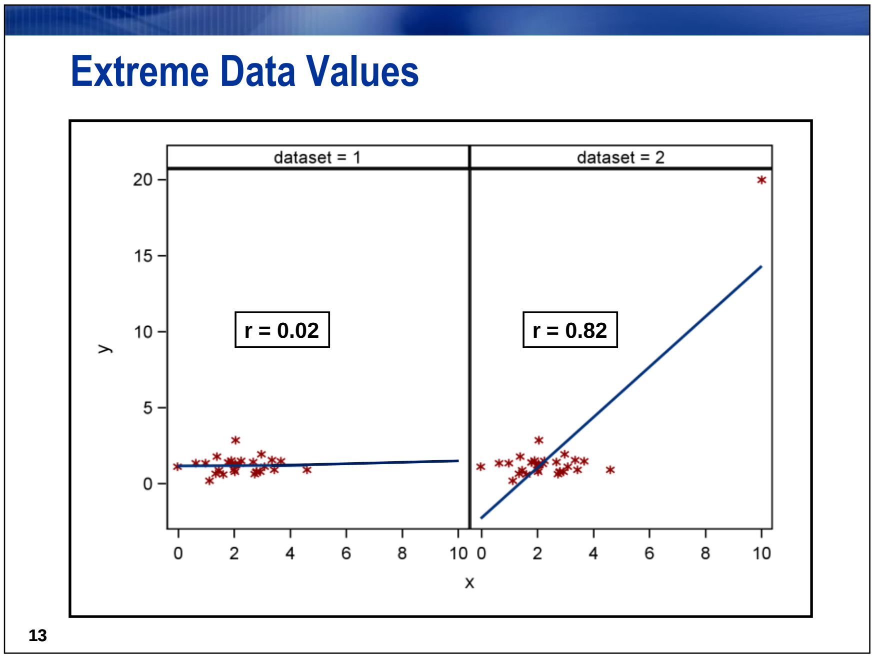
Clearly this type of thinking is faulty. Can you think of possible explanations for this relationship?



In the scatter plot above, the variables have a fairly low Pearson correlation coefficient. Why?

- Pearson correlation coefficients measure linear relationships.
- A Pearson correlation coefficient close to 0 indicates that there is not a strong linear relationship between two variables.
- A Pearson correlation coefficient close to 0 does not mean there is no relationship of any kind between the two variables.

In this example, there is a curvilinear relationship between the two variables.

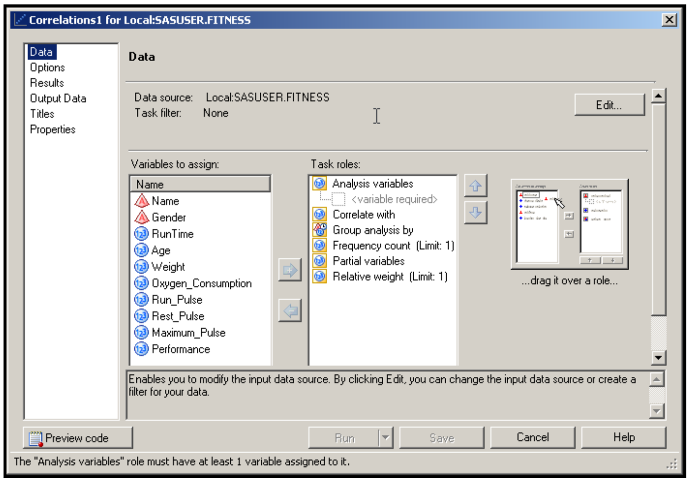


Correlation coefficients are highly affected by a few extreme values of either variable. The scatter plots above shows that the degree of linear relationship is mainly determined by one point. If you include the unusual point in the data set, the correlation is close to 1. If you do not include it, the correlation is close to 0.

In this situation, follow these steps:

1. Investigate the unusual data point to make sure it is valid.
2. If the data point is valid, collect more data between the unusual data point and the group of data points to see whether a linear relationship unfolds.
3. Try to replicate the unusual data point by collecting data at a fixed value of *x* (in this case, *x*=10). This determines whether the data point is unusual.
4. Compute two correlation coefficients, one with the unusual data point and one without it. This shows how influential the unusual data point is in the analysis. In this case, it is greatly influential.

The Correlations Task



14

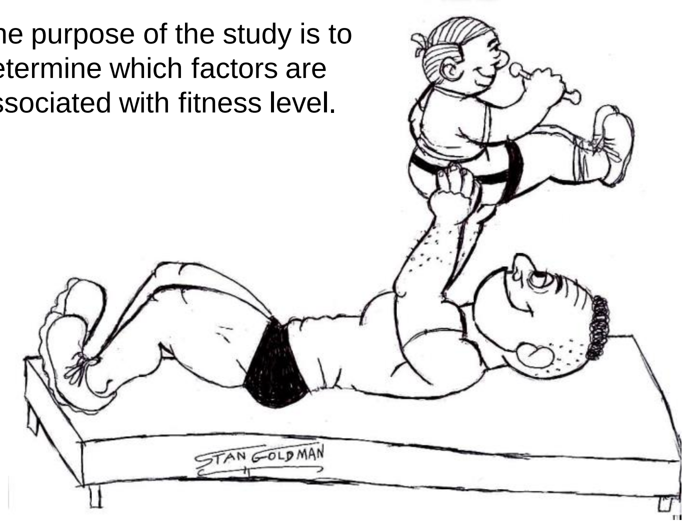
You can use the Correlations task to produce correlation statistics and scatter plots for your data.

Exploratory analysis in preparation for multiple regression often involves looking at bivariate scatter plots and correlations between each of the predictor variables and the response variable. It is not suggested that exclusion or inclusion decisions be made on the basis of these analyses. The purpose is to explore the shape of the relationships (because linear regression assumes a linear shape to the relationship) and to screen for outliers. You will also want to check for multivariate outliers when you test your multiple regression models later.

The Correlations task provides bivariate correlation tables. These tables are accompanied by ODS Statistical Graphics.

Fitness Example

The purpose of the study is to determine which factors are associated with fitness level.



15

In exercise physiology, an objective measure of aerobic fitness is how fast the body can absorb and use oxygen (oxygen consumption). Subjects participated in a predetermined exercise run of 1.5 miles. Measurements of oxygen consumption as well as several other continuous measurements such as age, pulse, and weight were recorded. The researchers are interested in determining whether any of these other variables can help predict oxygen consumption. This data is found in Rawlings (1998) but certain values of **`Maximum_Pulse`** and **`Run_Pulse`** were changed for illustration. **`Name`**, **`Gender`**, and **`Performance`** were also contrived for illustration.

The **`fitness`** data set contains the following variables:

`Name`	name of the member
`Gender`	gender of the member
`Runtime`	time to run 1.5 miles (in minutes)
`Age`	age of the member (in years)
`Weight`	weight of the member (in kilograms)
`Oxygen_Consumption`	a measure of the ability to use oxygen in the blood stream
`Run_Pulse`	pulse rate at the end of the run
`Rest_Pulse`	resting pulse rate
`Maximum_Pulse`	maximum pulse rate during the run
`Performance`	a measure of overall fitness on a 0-100 scale.

Data Exploration, Correlations, and Scatter Plots

Correlation

Describing the Relationships between Continuous Variables

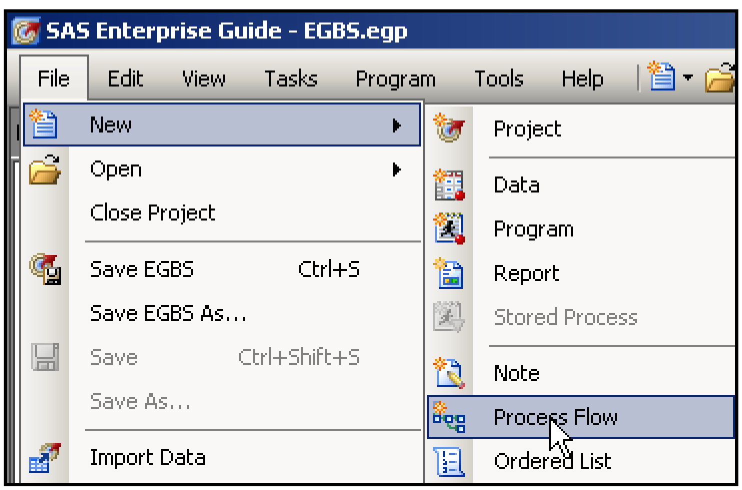
a. Generate scatter plots and correlations for the variables **Age**, **Weight**, **Height** and the circumference measures with the variable, **PctBodyFat2**.

Important! Graphics in the Correlations task limits you to 10 variables at a time, so for this exercise, run the Correlations task on all variables once without plots. Then, look at the relationships with **Age**, **Weight**, **Height** separately from the circumference variables (**Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**).

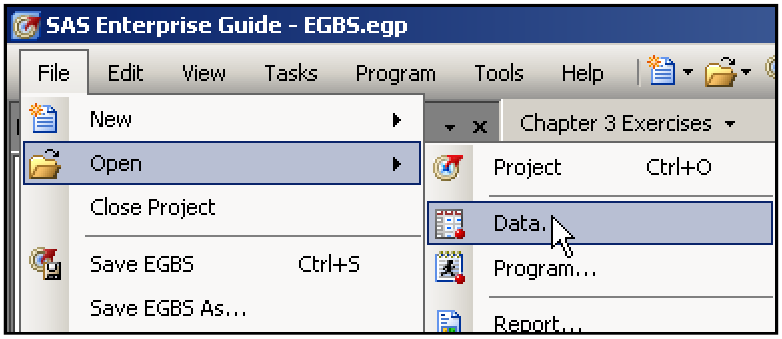
Correlation tables can be created using more than 10 variables at a time.

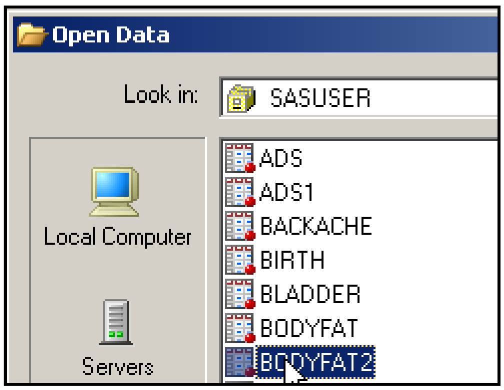
1) What variable has the highest correlation with **PctBodyFat2**?

- Create a new Process Flow and rename it **SASEG8B**.

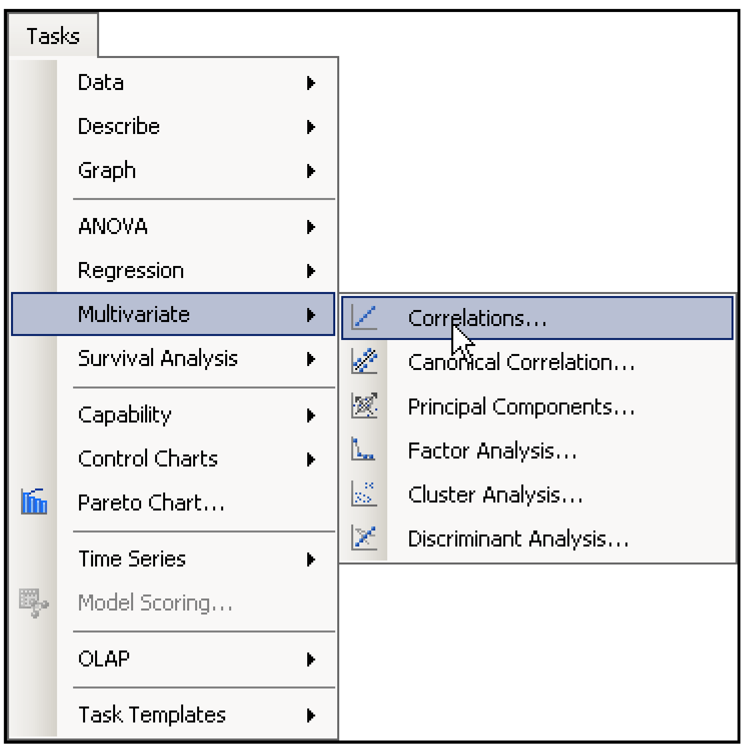


- Open the **`BodyFat2`** data set.

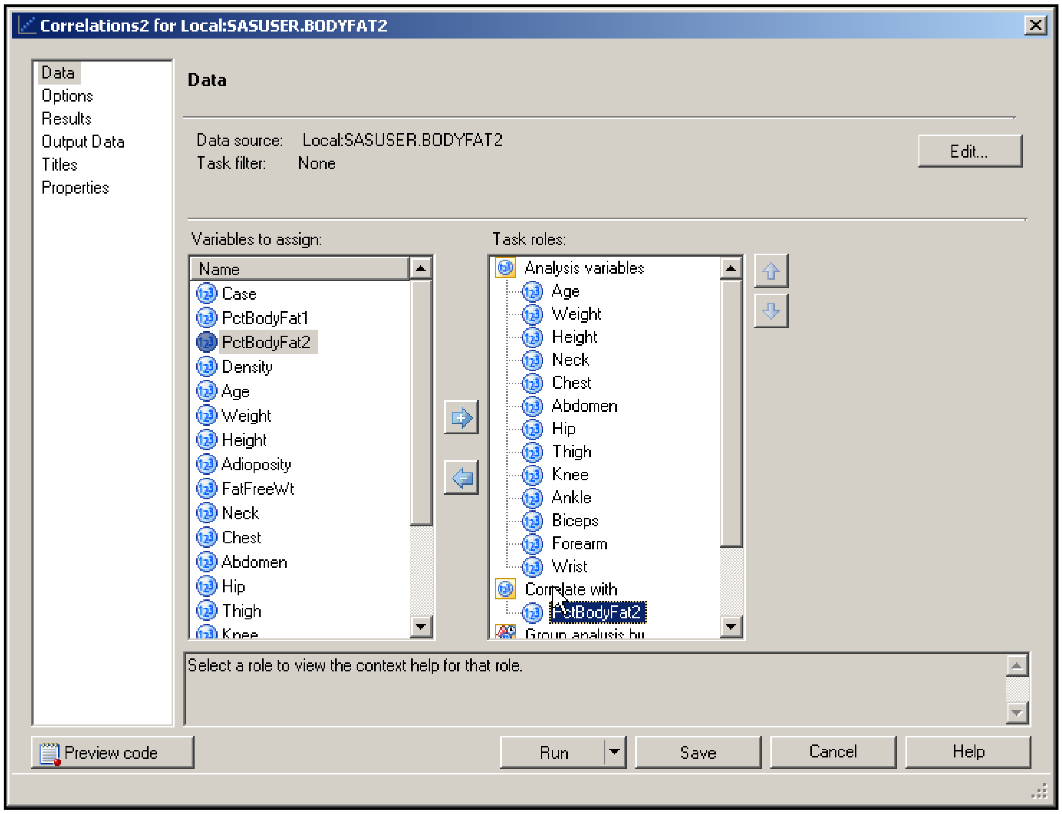




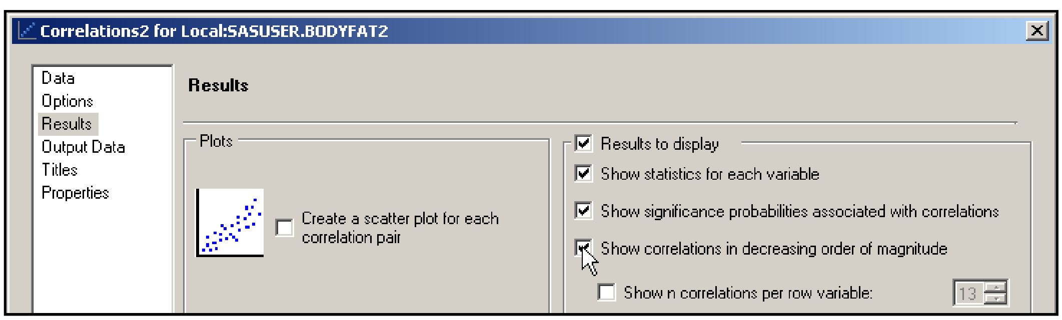
- Select <u>**Tasks**</u> ⇨ <u>**Multivariate**</u> ⇨ <u>**Correlations…**</u>.



- With **Data** selected at the left, assign **`Age`**, **`Weight`**, **`Height`**, **`Neck`**, **`Chest`**, **`Abdomen`**, **`Hip`**, **`Thigh`**, **`Knee`**, **`Ankle`**, **`Biceps`**, **`Forearm`**, and **`Wrist`** to the analysis variables task role and **`PctBodyFat2`** to the role of correlate with.



- With **Results** selected at the left, check the box for **Show correlations in decreasing order of magnitude**.



- Click Run .

Correlation Analysis

The CORR Procedure

1 With Variables:	PctBodyFat2												
13 Variables:	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist

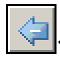
Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
PctBodyFat2	252	19.15079	8.36874	4826	0	47.50000
Age	252	44.88492	12.60204	11311	22.00000	81.00000
Weight	252	178.92440	29.38916	45089	118.50000	363.15000
Height	252	70.30754	2.60958	17718	64.00000	77.75000
Neck	252	37.99206	2.43091	9574	31.10000	51.20000
Chest	252	100.82421	8.43048	25408	79.30000	136.20000
Abdomen	252	92.55595	10.78308	23324	69.40000	148.10000
Hip	252	99.90476	7.16406	25176	85.00000	147.70000
Thigh	252	59.40595	5.24995	14970	47.20000	87.30000
Knee	252	38.59048	2.41180	9725	33.00000	49.10000
Ankle	252	23.10238	1.69489	5822	19.10000	33.90000
Biceps	252	32.27341	3.02127	8133	24.80000	45.00000
Forearm	252	28.66389	2.02069	7223	21.00000	34.90000
Wrist	252	18.22976	0.93358	4594	15.80000	21.40000

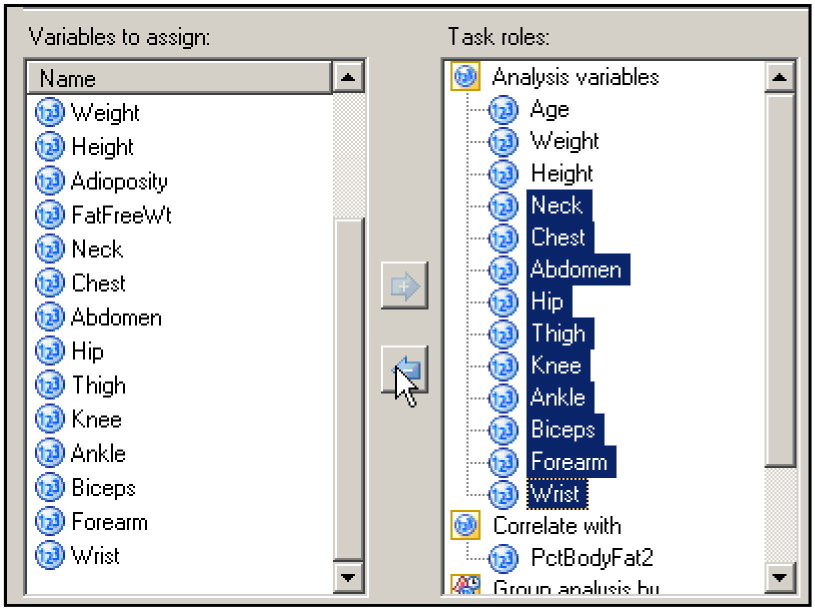
Pearson Correlation Coefficients, N = 252 Prob > |r| under H0: Rho=0													
	Abdomen	Chest	Hip	Weight	Thigh	Knee	Biceps	Neck	Forearm	Wrist	Age	Ankle	Height
PctBodyFat2	0.81343	0.70262	0.62520	0.61241	0.55961	0.50867	0.49327	0.49059	0.36139	0.34657	0.29146	0.26597	-0.02529
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.6895

The variable with the highest correlation with **`PctBodyFat2`** is **`Abdomen`** (measured diameter of the abdomen).

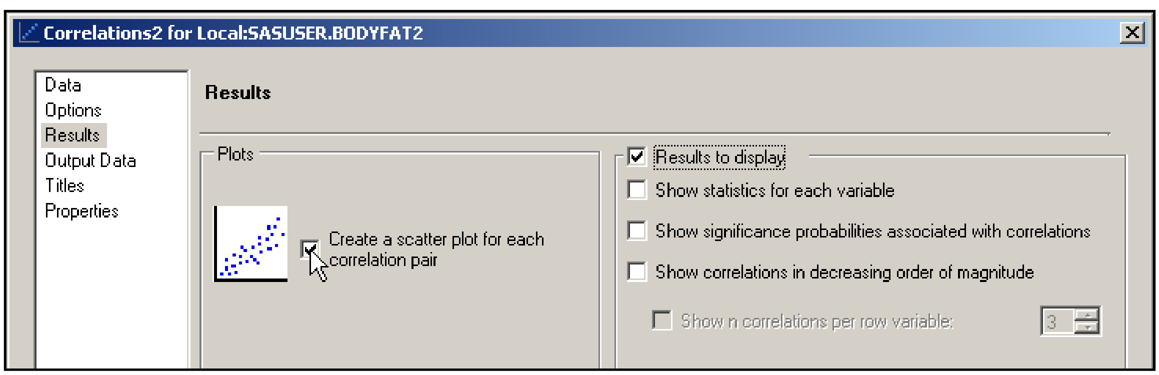
2) Can straight lines adequately describe the relationships?

3) Are there any outliers you should investigate?

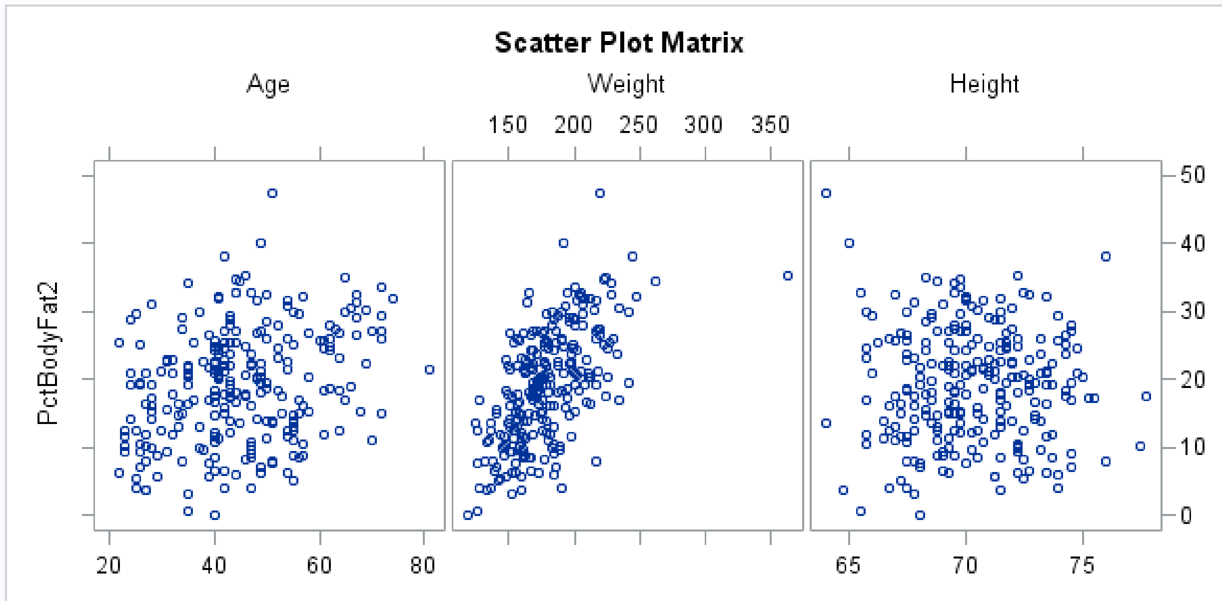
- Reopen the last Correlations task by right-clicking the icon for it and selecting **Modify…** from the drop-down menu.
- In Data, remove the circumference variables from the analysis variables task role by highlighting them and clicking .



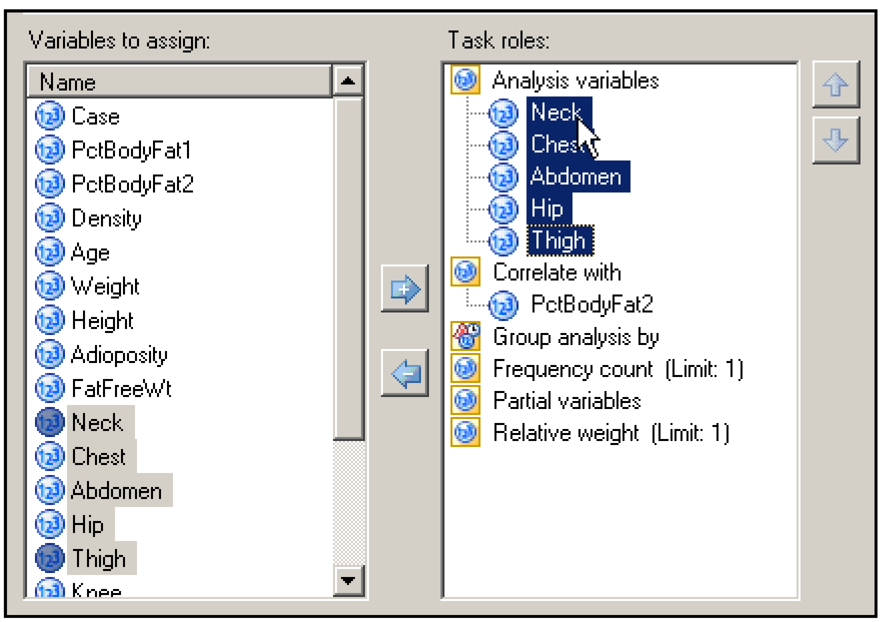
- With **Results** selected at the left, uncheck all boxes at the right except for **Results to display** and then check the box at the left for **Create a scatter plot for each correlation pair**.

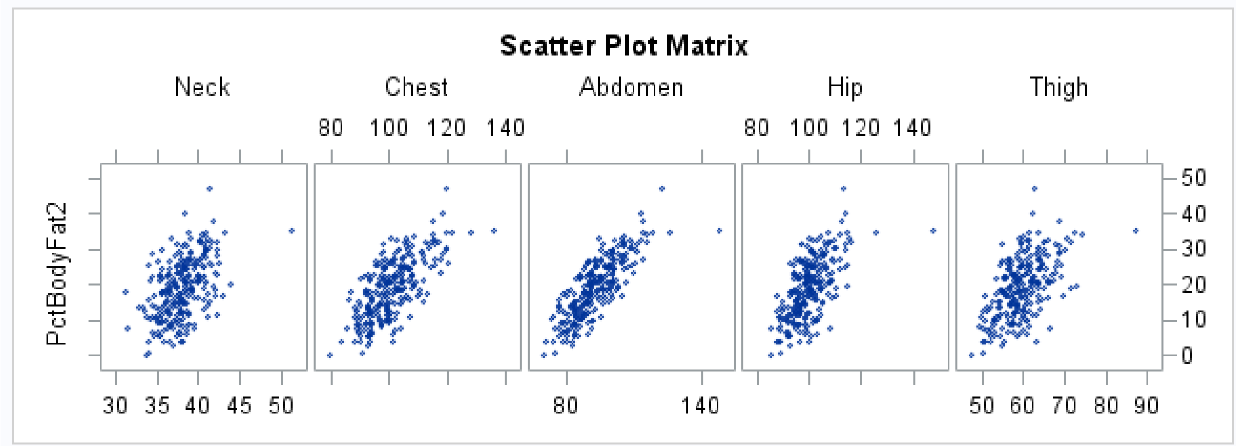


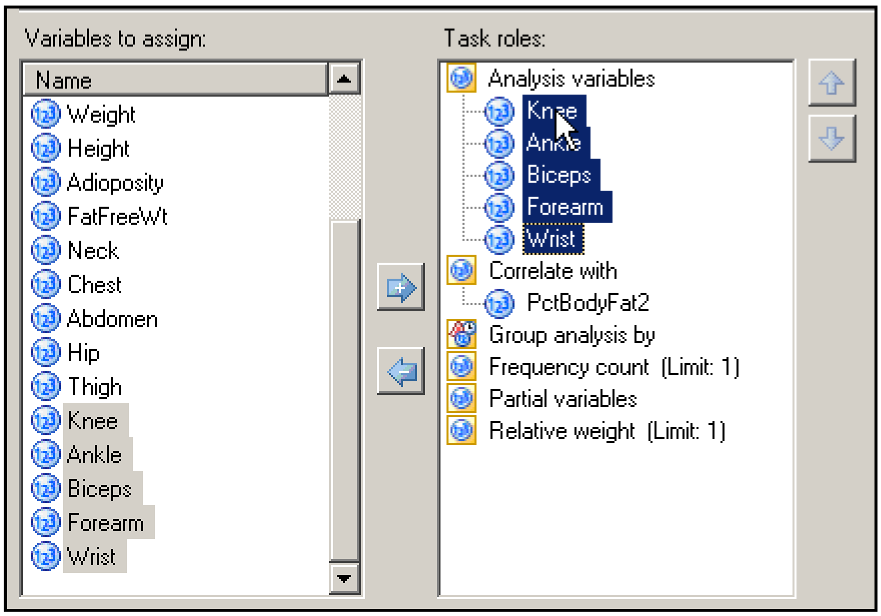
- Click Run.
- Do not replace the results from the previous run.

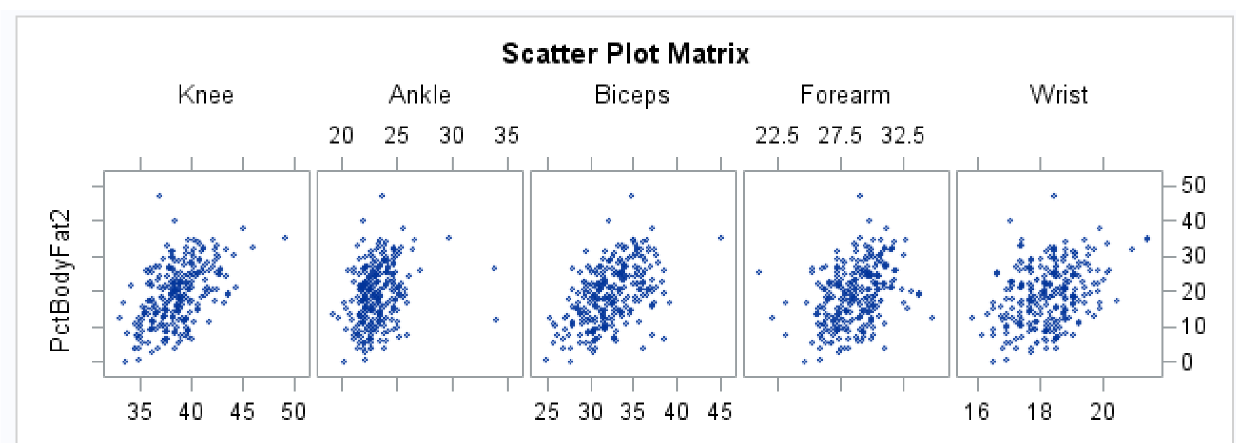


- Repeat the process for the first set of variables for the first 5 circumference variables and then for the last 5.
- Do not replace results from previous runs.







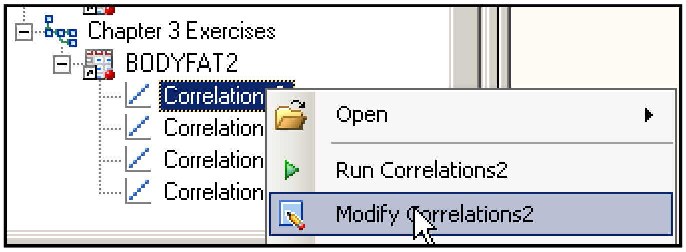


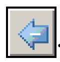
Most variables seem to have at least a bit of linear association with **`PctBodyFat2`**.

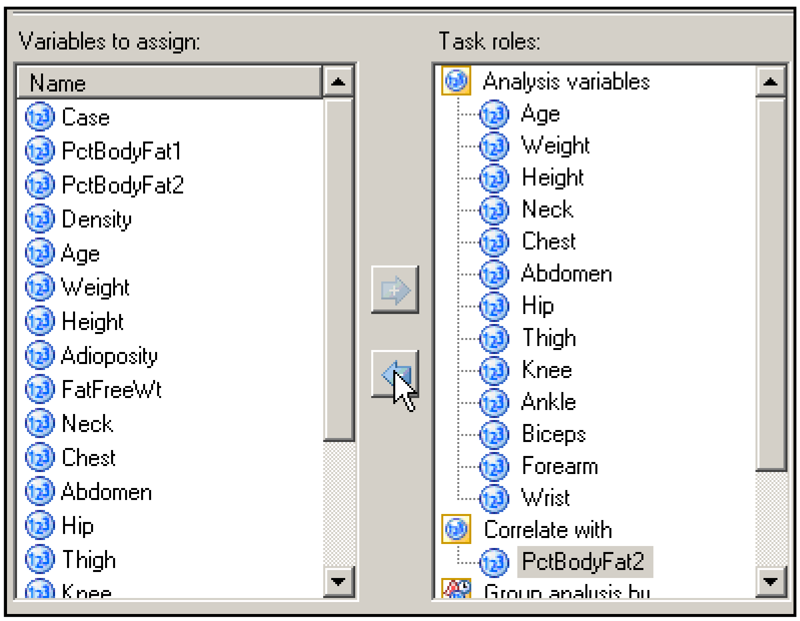
There seems to be at least one outlier for **`Weight`** (over 350 pounds). This same person is likely the outlier for several other circumference measures.

b. Generate correlations among all of the potential predictor variables **`Age`**, **`Weight`**, **`Height`**, and the circumference measures. Are there any notable relationships?

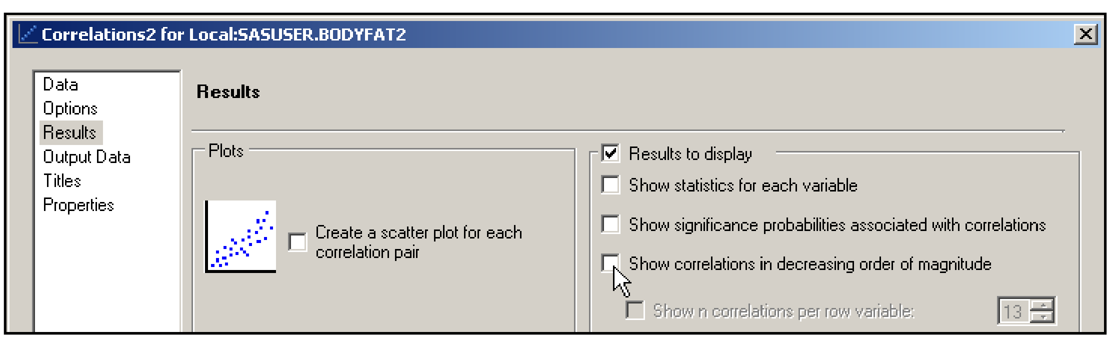
- Re-open the first Correlation task by right-clicking the icon for it and selecting **<u>Modify Correlations2</u>** from the drop-down menu.



- In Data, remove **`PctBodyFat2`** from the correlate with task role by highlighting it and clicking .



- Under Results, uncheck all boxes except for **<u>Results to display</u>**.



- Click .
- Do not replace the results from the previous run.

Correlation Analysis

The CORR Procedure

13 Variables:	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist

Pearson Correlation Coefficients, N = 252													
	Age	**Weight**	**Height**	**Neck**	**Chest**	**Abdomen**	**Hip**	**Thigh**	**Knee**	**Ankle**	**Biceps**	**Forearm**	**Wrist**
Age	1.00000	-0.01275	-0.24521	0.11351	0.17645	0.23041	-0.05033	-0.20010	0.01752	-0.10506	-0.04116	-0.08506	0.21353
Weight	-0.01275	1.00000	0.48689	0.83072	0.89419	0.88799	0.94088	0.86869	0.85317	0.61369	0.80042	0.63030	0.72977
Height	-0.24521	0.48689	1.00000	0.32114	0.22683	0.18977	0.37211	0.33856	0.50050	0.39313	0.31851	0.32203	0.39778
Neck	0.11351	0.83072	0.32114	1.00000	0.78484	0.75408	0.73496	0.69570	0.67240	0.47789	0.73115	0.62366	0.74483
Chest	0.17645	0.89419	0.22683	0.78484	1.00000	0.91583	0.82942	0.72986	0.71950	0.48299	0.72791	0.58017	0.66016
Abdomen	0.23041	0.88799	0.18977	0.75408	0.91583	1.00000	0.87407	0.76662	0.73718	0.45322	0.68498	0.50332	0.61983
Hip	-0.05033	0.94088	0.37211	0.73496	0.82942	0.87407	1.00000	0.89641	0.82347	0.55839	0.73927	0.54501	0.63009
Thigh	-0.20010	0.86869	0.33856	0.69570	0.72986	0.76662	0.89641	1.00000	0.79917	0.53980	0.76148	0.56684	0.55868
Knee	0.01752	0.85317	0.50050	0.67240	0.71950	0.73718	0.82347	0.79917	1.00000	0.61161	0.67871	0.55590	0.66451
Ankle	-0.10506	0.61369	0.39313	0.47789	0.48299	0.45322	0.55839	0.53980	0.61161	1.00000	0.48485	0.41905	0.56619
Biceps	-0.04116	0.80042	0.31851	0.73115	0.72791	0.68498	0.73927	0.76148	0.67871	0.48485	1.00000	0.67826	0.63213
Forearm	-0.08506	0.63030	0.32203	0.62366	0.58017	0.50332	0.54501	0.56684	0.55590	0.41905	0.67826	1.00000	0.58559
Wrist	0.21353	0.72977	0.39778	0.74483	0.66016	0.61983	0.63009	0.55868	0.66451	0.56619	0.63213	0.58559	1.00000

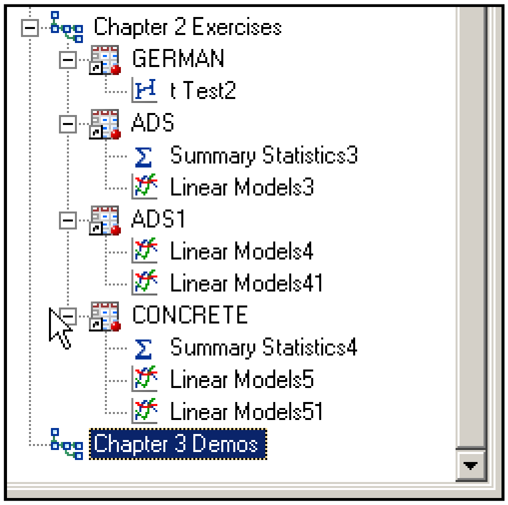
The highest correlation (0.94088) is between **`Hip`** and **`Weight`**.

There is a pattern of high correlations among all of the diameter measures and Weight. None of this should be surprising given a lay intuition about the consistency of dimensions of the human body.

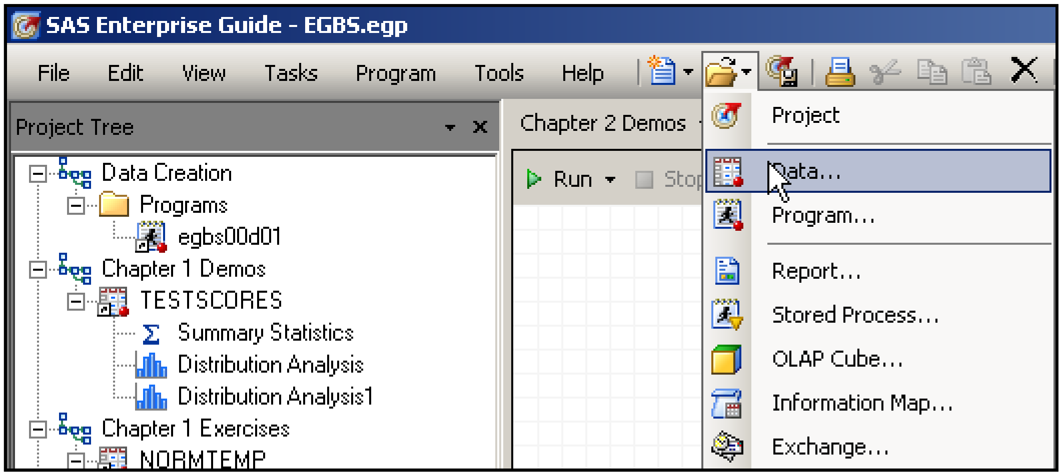
Possible predictors of Oxygen Consumption

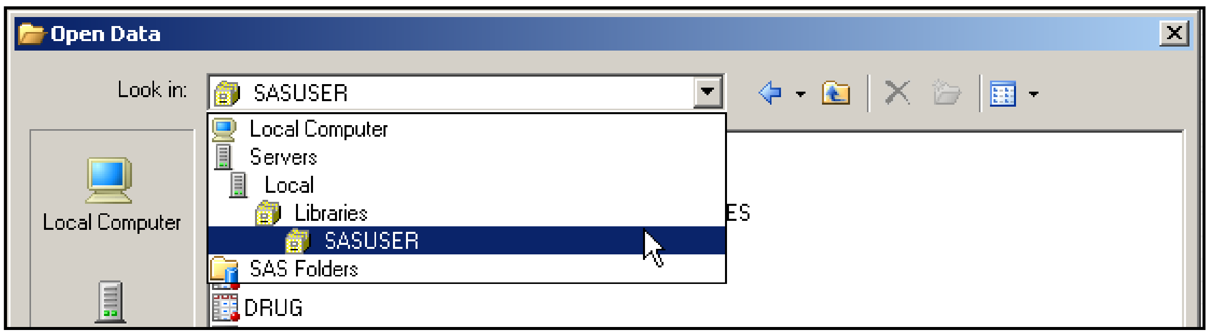
Perform a correlation analysis relating seven potential predictors of oxygen consumption with the variable, **`Oxygen_Consumption`**, using the **`Fitness`** data set.

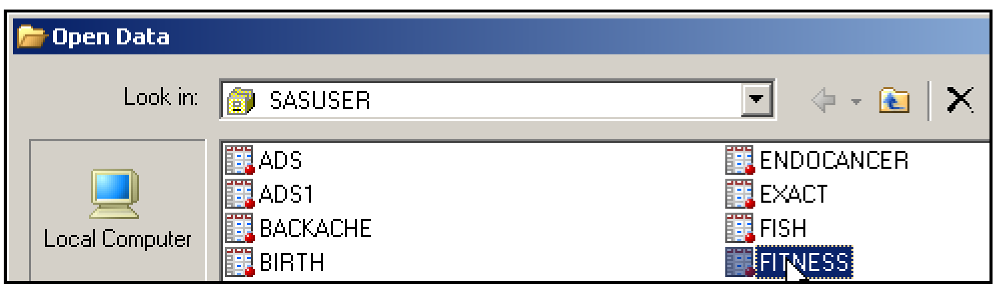
1. Create a new Process Flow and rename it **`SASEG8B`**.

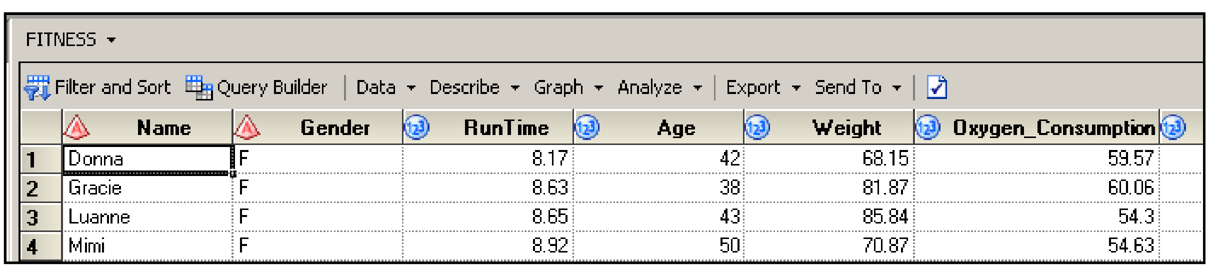


2. Open the **`Fitness`** data set.

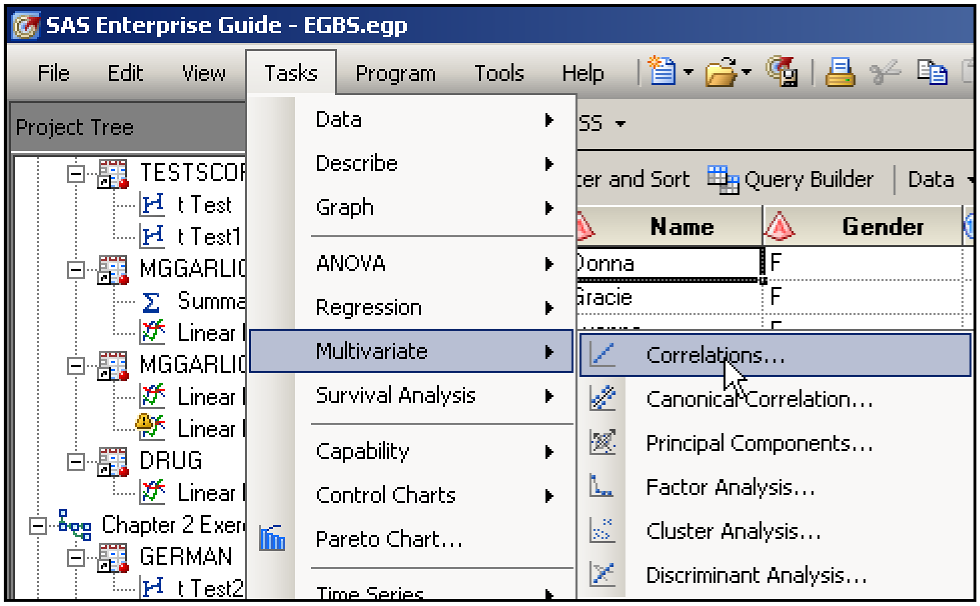


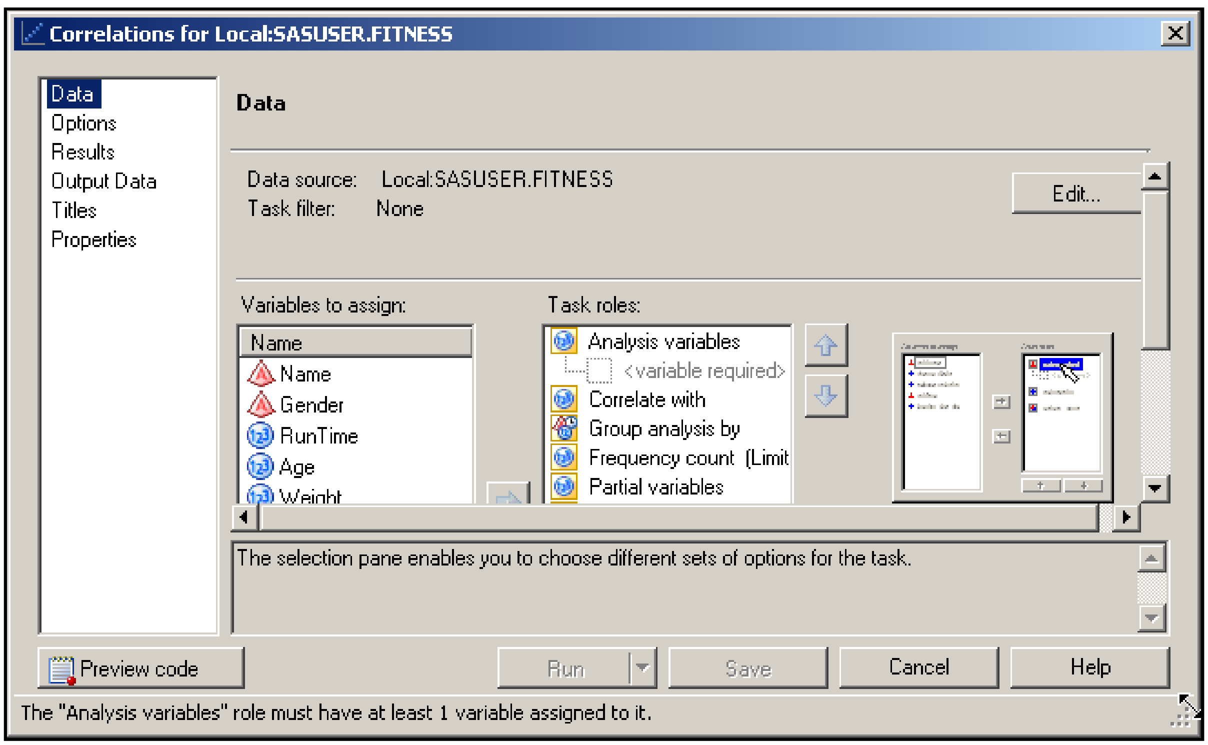






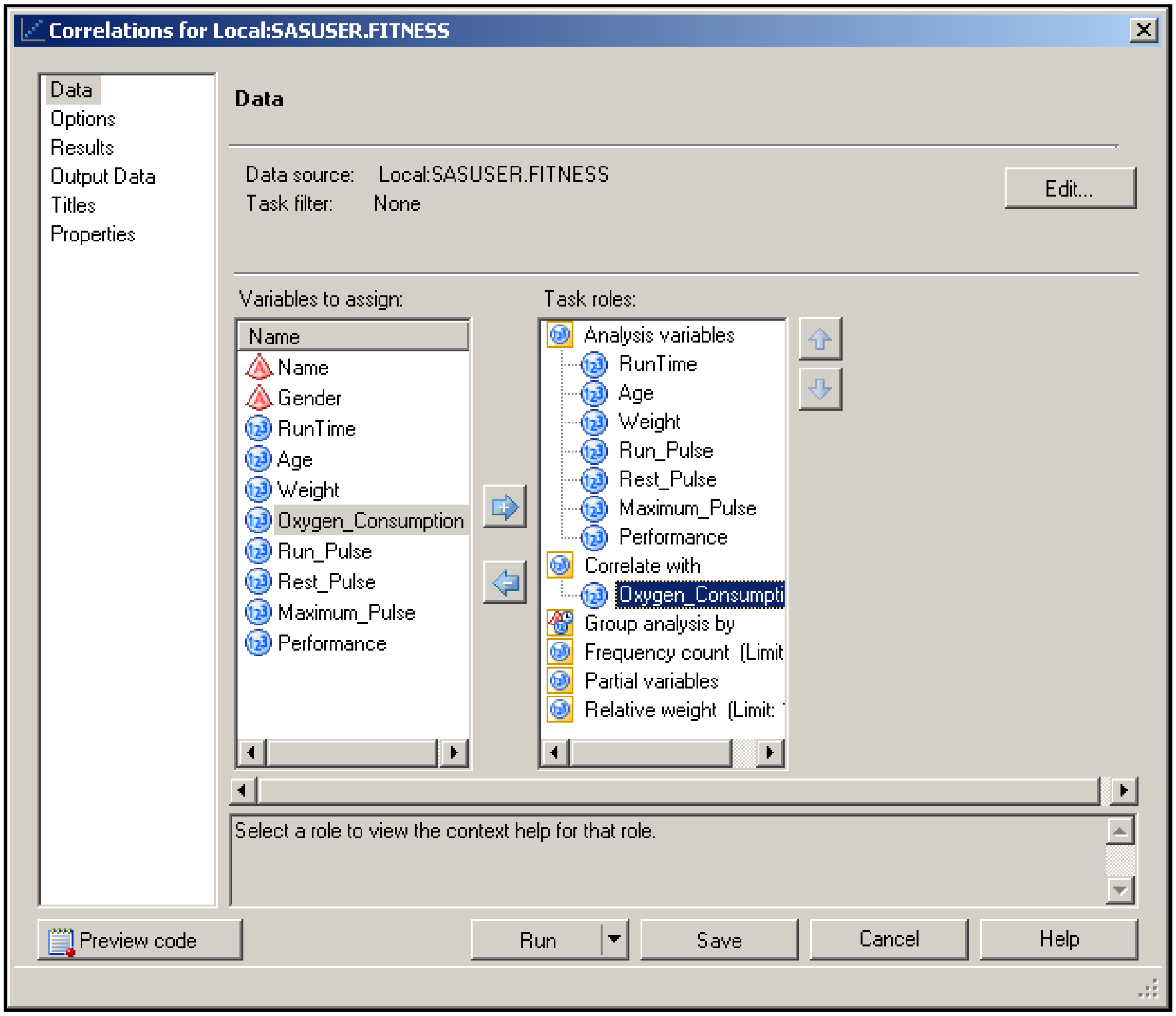
3. Select **Tasks** ⇨ **Multivariate** ⇨ **Correlations…**.



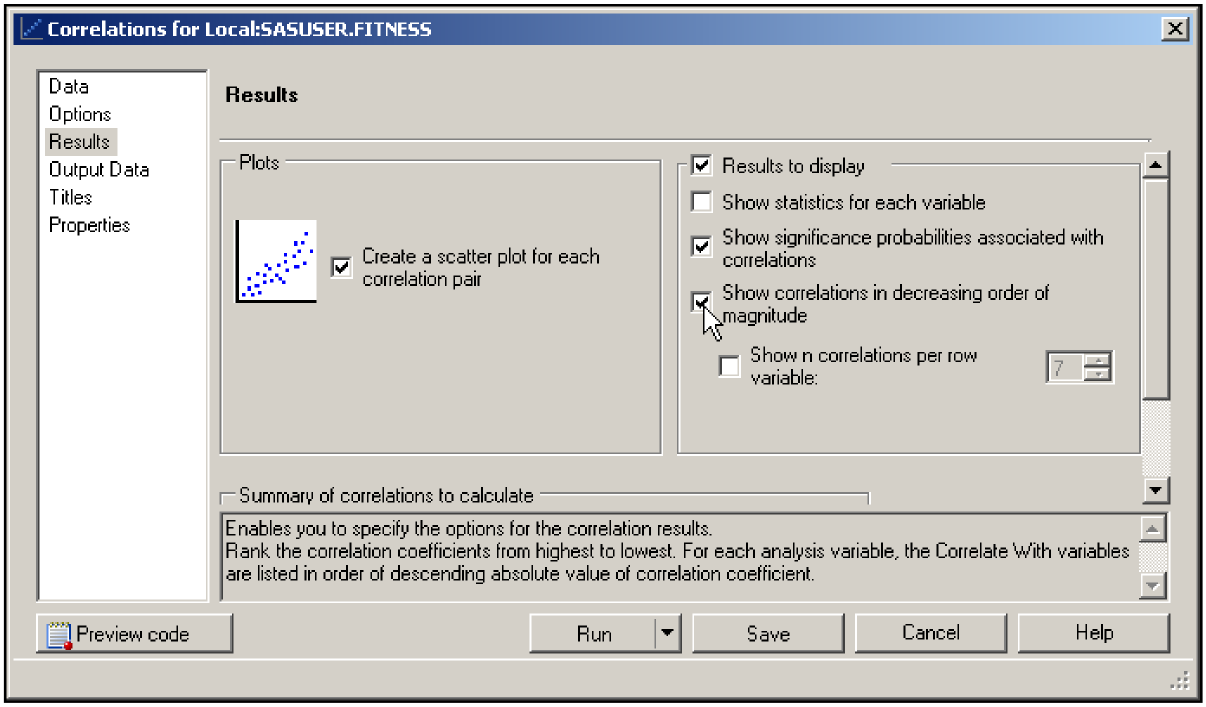


You should also investigate the univariate statistics of continuous variables in the data set, just as you did in the earlier chapters, using the Distribution Analysis task to explore distributions, measure central tendency and spread, and look for outliers.

4. With **Data** selected at the left, assign **`Runtime`**, **`Age`**, **`Weight`**, **`Run_Pulse`**, **`Rest_Pulse`**, **`Maximum_Pulse`**, and **`Performance`** to the task role of analysis variables and **`Oxygen_Consumption`** to the role of correlate with.



5. In Results, check the box for **Create a scatter plot for each correlation pair**. Also, check the box at the right for **Show correlations in decreasing order of magnitude** and uncheck the box for **Show statistics for each variable**.



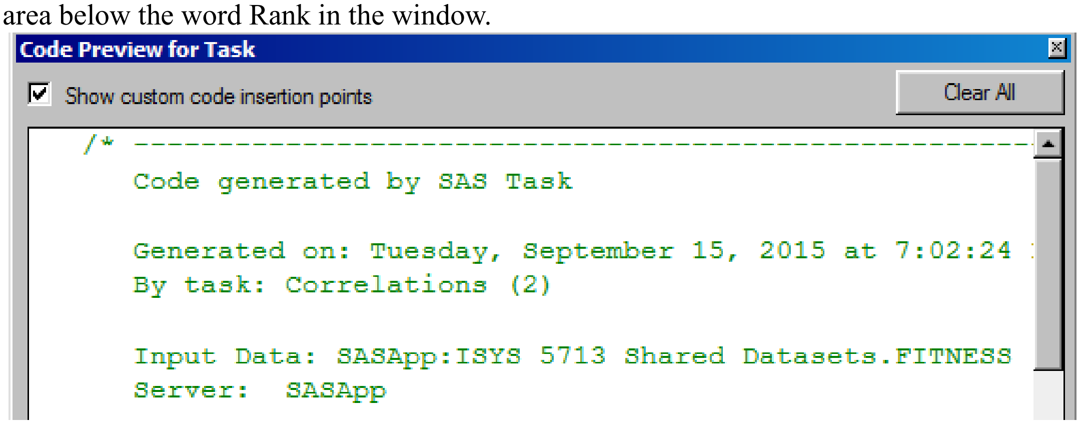
By default, there will be a maximum of 5 scatter plots produced. If you want to allow for the creation of more than 5 scatter plots, then the code will need to be modified. The next steps explain how to do that.

6. Click 

.

7. Check the Show custom code insertion points box in the window that opens.

8. Scroll down and type the statement **`PLOTS=(MATRIX(NVAR=ALL))`** in <insert custom code here> area below the word Rank in the window.



9. Click ☒ to close the insert custom code window.

10. Click Run.

Code Preview for Task

☑ Show custom code insertion points   Clear All

```
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME,

/* Start of custom user code (BeforeProcCorr) */
<insert custom code here>
/* End of custom user code (BeforeProcCorr) */

PROC CORR DATA=WORK.SORTTempTableSorted
    PLOTS=(SCATTER MATRIX)
    PEARSON
    VARDEF=DF
    NOSIMPLE
    RANK

/* Start of custom user code (InProcCorrBeforeSemicolon)
PLOTS=(MATRIX(NVAR=ALL))
/* End of custom user code (InProcCorrBeforeSemicolon) *

    ;
```

Correlation Analysis

The CORR Procedure

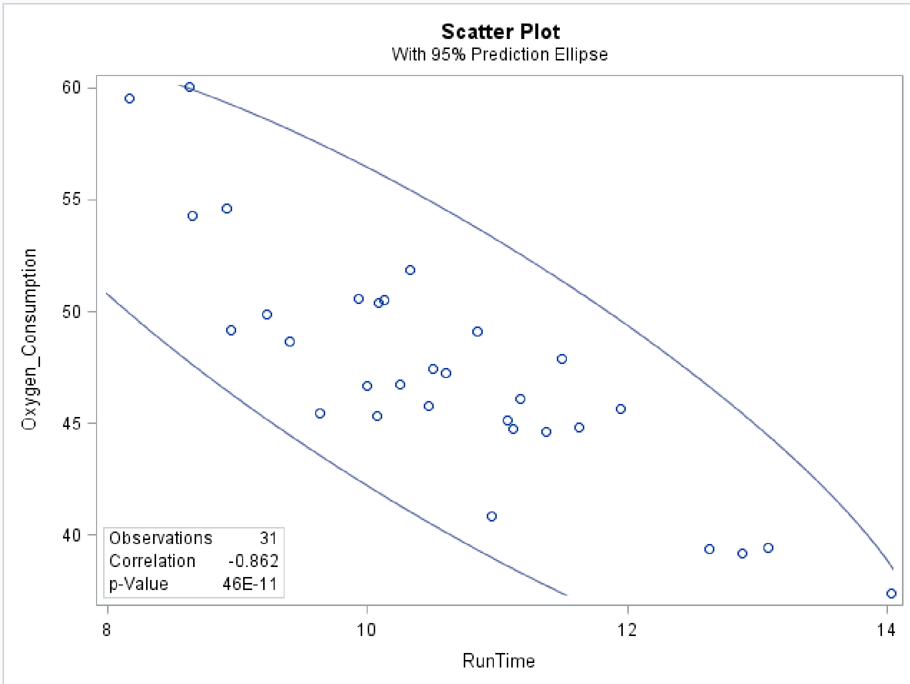
1 With Variables:	Oxygen_Consumption
7 Variables:	RunTime Age Weight Run_Pulse Rest_Pulse Maximum_Pulse Performance

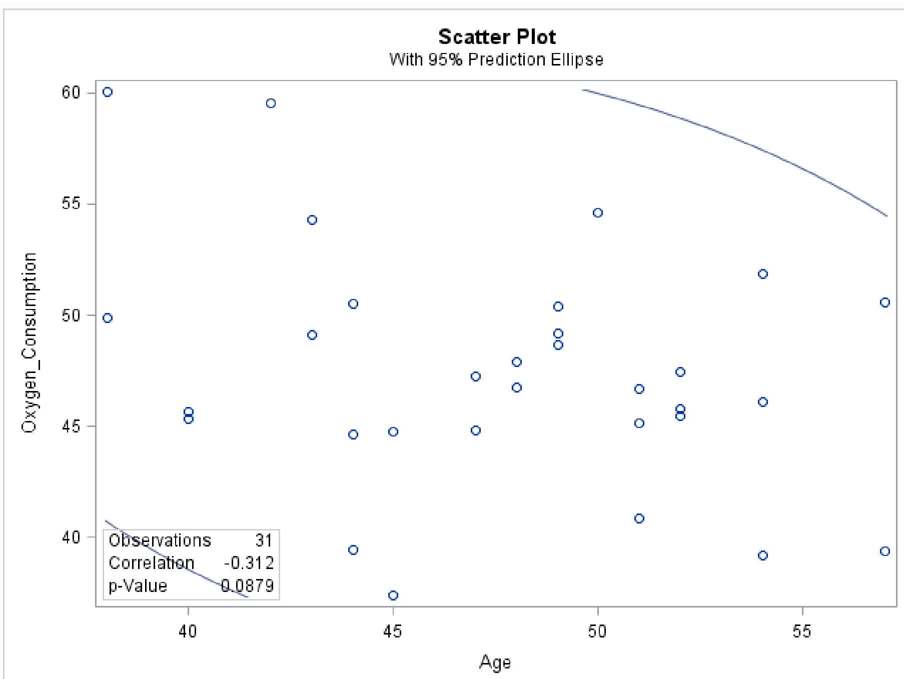
Pearson Correlation Coefficients, N = 31
Prob > \|r\| under H0: Rho=0

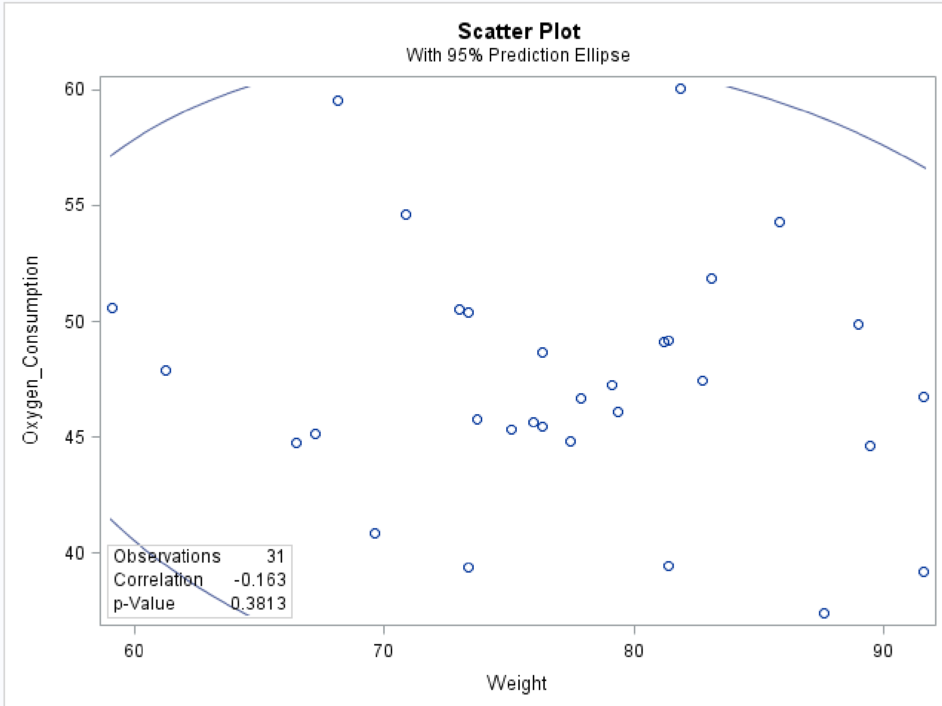
	RunTime	Performance	Rest_Pulse	Run_Pulse	Age	Maximum_Pulse	Weight
Oxygen_Consumption	-0.86219	0.77890	-0.39935	-0.39808	-0.31162	-0.23677	-0.16289
	<.0001	<.0001	0.0260	0.0266	0.0879	0.1997	0.3813

The correlation coefficient between **`Oxygen_Consumption`** and **`RunTime`** is -0.86219. The *p*-value is small, which indicates that the population correlation coefficient (Rho) is significantly different from 0. The second largest correlation coefficient, in absolute value, is **`Performance`**, at 0.77890.

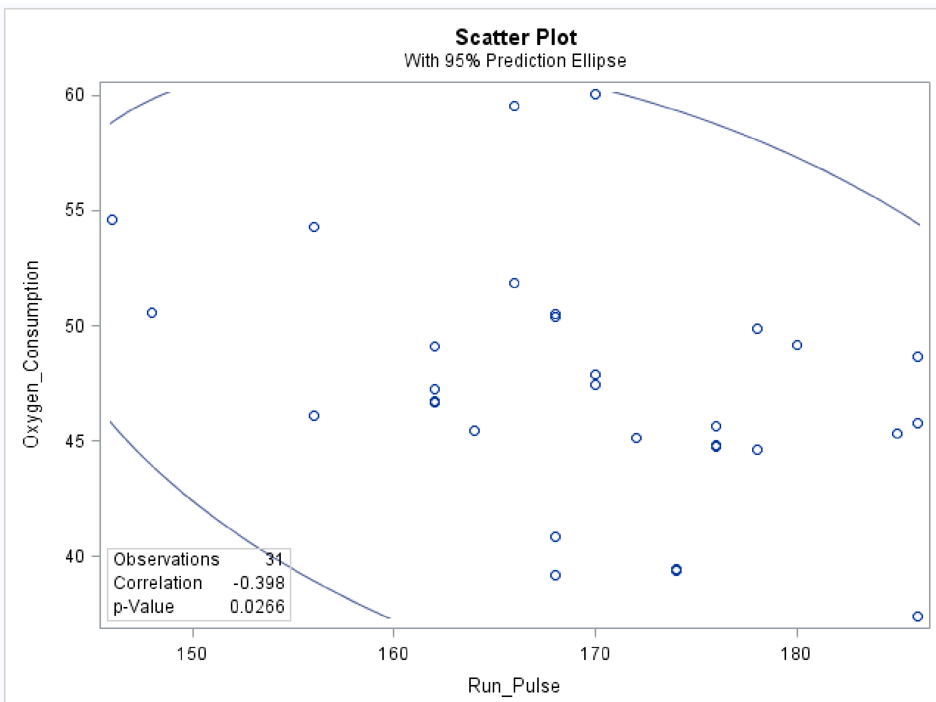
Scatter plots associated with these correlations are shown below.



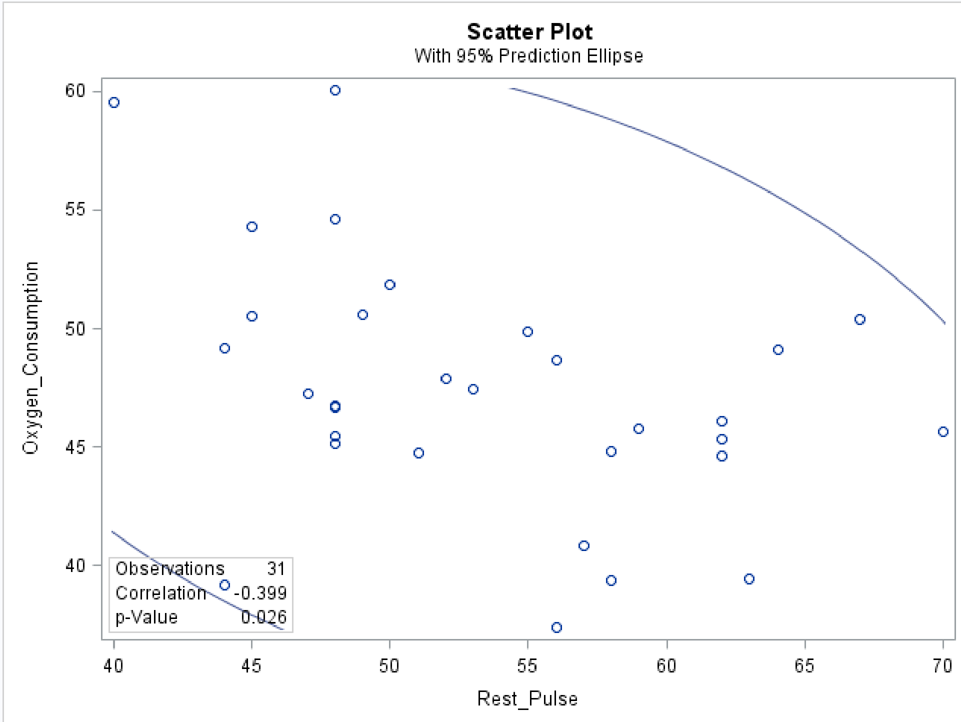


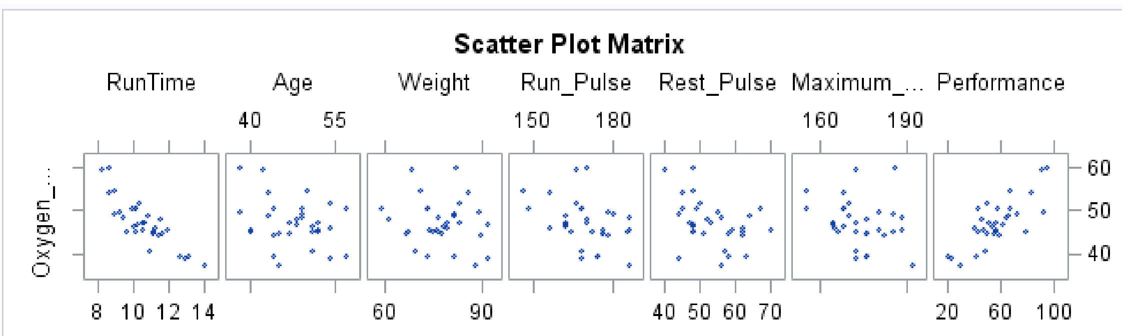


Scatter Plot
With 95% Prediction Ellipse
Oxygen_Consumption
Weight
60
55
50
45
40
60
70
80
90
Observations 31
Correlation -0.163
p-Value 0.3813



Scatter Plot
With 95% Prediction Ellipse
Oxygen_Consumption
Run_Pulse
60
55
50
45
40
150
160
170
180
Observations 31
Correlation -0.398
p-Value 0.0266

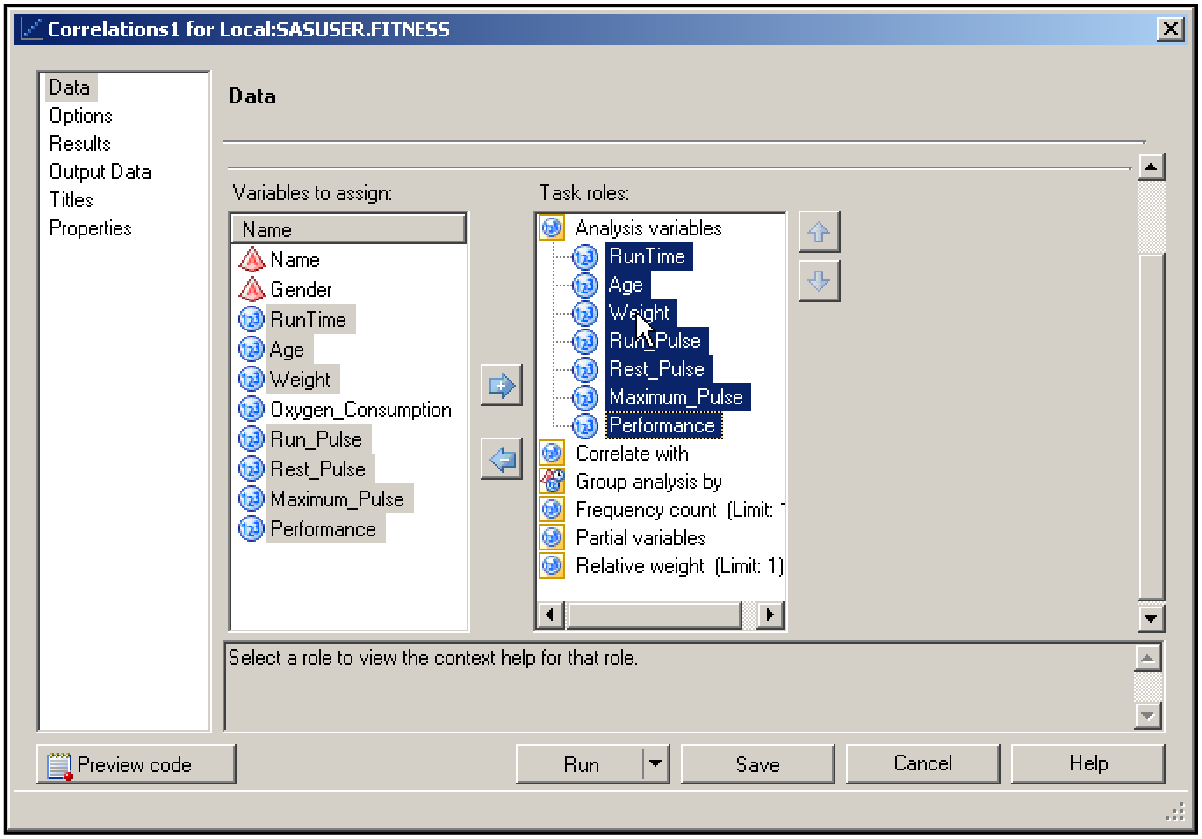




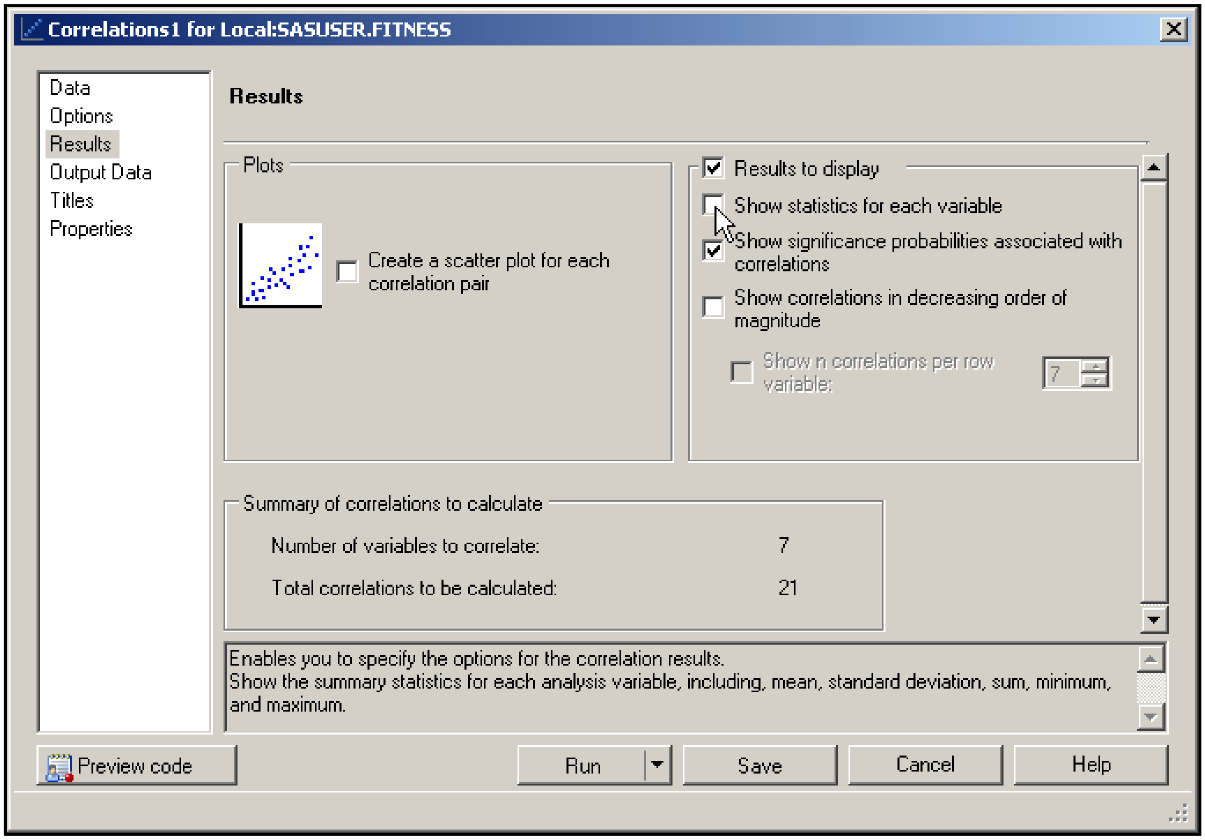
The correlations and scatter plots indicate that several variables might be good predictors for **`Oxygen_Consumption`**.

When you prepare to conduct a regression analysis, it is always good practice to examine the correlations among the potential predictor variables.

1. With the **Fitness** data set selected, open the Correlations task again.
2. Assign all numeric variables except for `Oxygen_Consumption` to the analysis variables task role.



3. In Results uncheck the box for **Show statistics for each variable**. If you would like scatter plots, then check the box for it and repeat the code modifications from the previous task.



4. Click .

Correlation Analysis

The CORR Procedure

7 Variables:	RunTime	Age	Weight	Run_Pulse	Rest_Pulse	Maximum_Pulse	Performance

Pearson Correlation Coefficients, N = 31
Prob > |r| under H0: Rho=0

	RunTime	Age	Weight	Run_Pulse	Rest_Pulse	Maximum_Pulse	Performance
RunTime	1.00000	0.19523	0.14351	0.31365	0.45038	0.22610	-0.82049
		0.2926	0.4412	0.0858	0.0110	0.2213	<.0001
Age	0.19523	1.00000	-0.24050	-0.31607	-0.15087	-0.41490	-0.71257
	0.2926		0.1925	0.0832	0.4178	0.0203	<.0001
Weight	0.14351	-0.24050	1.00000	0.18152	0.04397	0.24938	0.08974
	0.4412	0.1925		0.3284	0.8143	0.1761	0.6312
Run_Pulse	0.31365	-0.31607	0.18152	1.00000	0.35246	0.92975	-0.02943
	0.0858	0.0832	0.3284		0.0518	<.0001	0.8751
Rest_Pulse	0.45038	-0.15087	0.04397	0.35246	1.00000	0.30512	-0.22560
	0.0110	0.4178	0.8143	0.0518		0.0951	0.2224
Maximum_Pulse	0.22610	-0.41490	0.24938	0.92975	0.30512	1.00000	0.09002
	0.2213	0.0203	0.1761	<.0001	0.0951		0.6301
Performance	-0.82049	-0.71257	0.08974	-0.02943	-0.22560	0.09002	1.00000
	<.0001	<.0001	0.6312	0.8751	0.2224	0.6301	

There are strong correlations between **`Run_Pulse`** and **`Maximum_Pulse`** (0.92975) and between **`Runtime`** and **`Performance`** (-0.82049).

The following correlation table was created from the matrix by choosing small *p*-values. The table is in descending order, based on the absolute value of the correlation. It provides a summary of the correlation analysis of the independent variables.

Row Variable	Column Variable	Pearson's r	Prob > |r|
`Run_Pulse`	**`Maximum_Pulse`**	0.92975	<.0001
`Runtime`	**`Performance`**	-0.82049	<.0001
`Performance`	**`Age`**	-0.71257	<.0001
`Runtime`	**`Rest_Pulse`**	0.45038	0.0110
`Age`	**`Maximum_Pulse`**	-0.41490	0.0203
`Run_Pulse`	**`Rest_Pulse`**	0.35246	0.0518