Principal Components

Summary

The **Principal Components** procedure is designed to extract *k* principal components from a set of *p* quantitative variables *X*. The principal components are defined as the set of orthogonal linear combinations of *X* that have the greatest variance. Determining the principal components is often used to reduce the dimensionality of a set of predictor variables prior to their use in procedures such as multiple regression or cluster analysis. When the variables are highly correlated, the first few principal components may be sufficient to describe most of the variability present.

Sample StatFolio: *pca.sgp*

Sample Data:

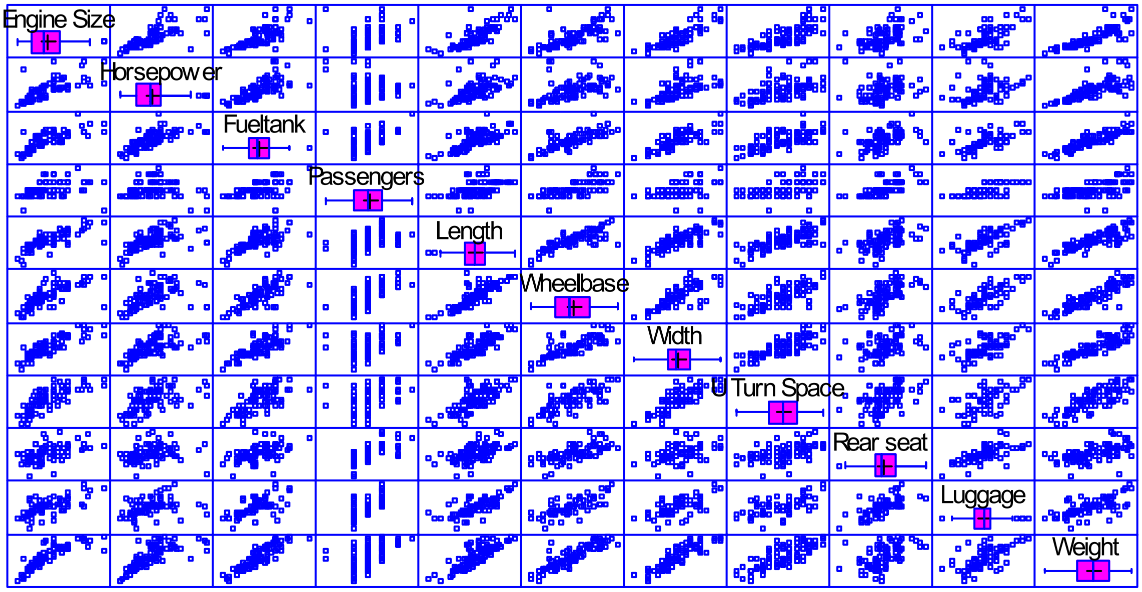
The file *93cars.sf6* contains information on 26 variables for *n* = 93 makes and models of automobiles, taken from Lock (1993). The table below shows a partial list of the data in that file:

Make	*Model*	*Engine Size*	*Horsepower*	*Fuel Tank*	*Passengers*	*Length*
Acura	Integra	1.8	140	13.2	5	177
Acura	Legend	3.2	200	18	5	195
Audi	90	2.8	172	16.9	5	180
Audi	100	2.8	172	21.1	6	193
BMW	535i	3.5	208	21.1	4	186
Buick	Century	2.2	110	16.4	6	189
Buick	LeSabre	3.8	170	18	6	200
Buick	Roadmaster	5.7	180	23	6	216
Buick	Riviera	3.8	170	18.8	5	198
Cadillac	DeVille	4.9	200	18	6	206
Cadillac	Seville	4.6	295	20	5	204
Chevrolet	Cavalier	2.2	110	15.2	5	182

It is desired to extract the principal components from the following variables:

Engine Size
Horsepower
Fueltank
Passengers
Length
Wheelbase
Width
U Turn Space
Rear seat
Luggage
Weight

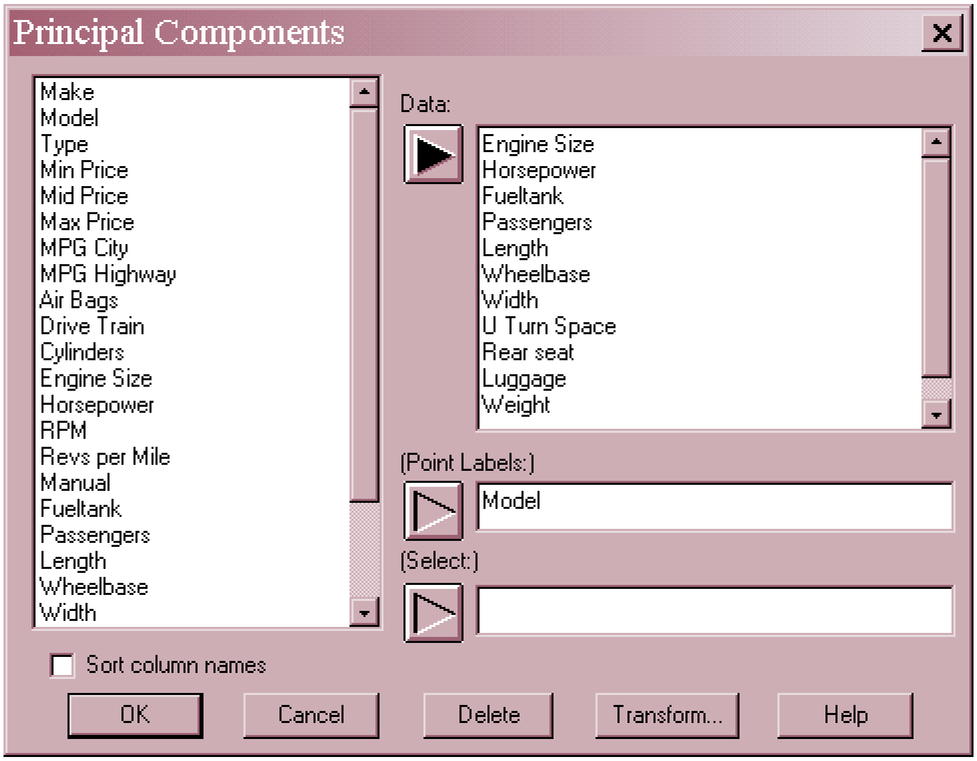
A matrix plot of the data is shown below:



As might be expected, the variables are highly correlated, since most are related to vehicle size.

Data Input

The data input dialog box requests the names of the columns containing the data:


Principal Components

Make
Model
Type
Min Price
Mid Price
Max Price
MPG City
MPG Highway
Air Bags
Drive Train
Cylinders
Engine Size
Horsepower
RPM
Revs per Mile
Manual
Fueltank
Passengers
Length
Wheelbase
Width

Data:
Engine Size
Horsepower
Fueltank
Passengers
Length
Wheelbase
Width
U Turn Space
Rear seat
Luggage
Weight

(Point Labels:)
Model

(Select:)

□ Sort column names

- **Data:** either the original observations or the sample covariance matrix $\hat{\Sigma}$. If entering the original observations, enter p numeric columns containing the n values for each column of X. If entering the sample covariance matrix, enter p numeric columns containing the p values for each column of $\hat{\Sigma}$. If the covariance matrix is entered, some of the tables and plots will not be available.

- **Point Labels**: optional labels for each observation.
- **Select:** subset selection.

Statistical Model

The goal of a principal components analysis is to construct *k* linear combinations of the *p* variables *X* that contain the greatest variance. The linear combinations take the form

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p \\ &\dots \\ Y_k &= a_{k1}X_1 + a_{k2}X_2 + ... + a_{kp}X_p \end{aligned} \qquad (1)$$

The first principal component is that linear combination that has maximum variance, subject to the constraint that the coefficient vector has unit length, i.e.,

$$\sum_{i=1}^{p} a_{ip}^2 = 1 \qquad (2)$$

If the covariance matrix of *X* equals Σ, then the variance of Y_1 is

$$Var(Y_1) = a_1'\Sigma a_1 \qquad (3)$$

The second principal combination is that linear combination that has the next greatest variance, such to the same constraint on unit length and also to the constraint that it be uncorrelated with the first principal component. Subsequent components explain as much of the remaining variance as possible, while being uncorrelated with all of the other components.

Under this model, the coefficients *a* correspond to the eigenvectors of Σ, while the variances of the *Y's* are equal to the eigenvalues:

$$Var(Y_j) = \lambda_j \qquad (4)$$

The population variance equals the sum of the eigenvalues

$$\text{Total population variance} = \lambda_1 + \lambda_2 + \dots + \lambda_p \qquad (5)$$

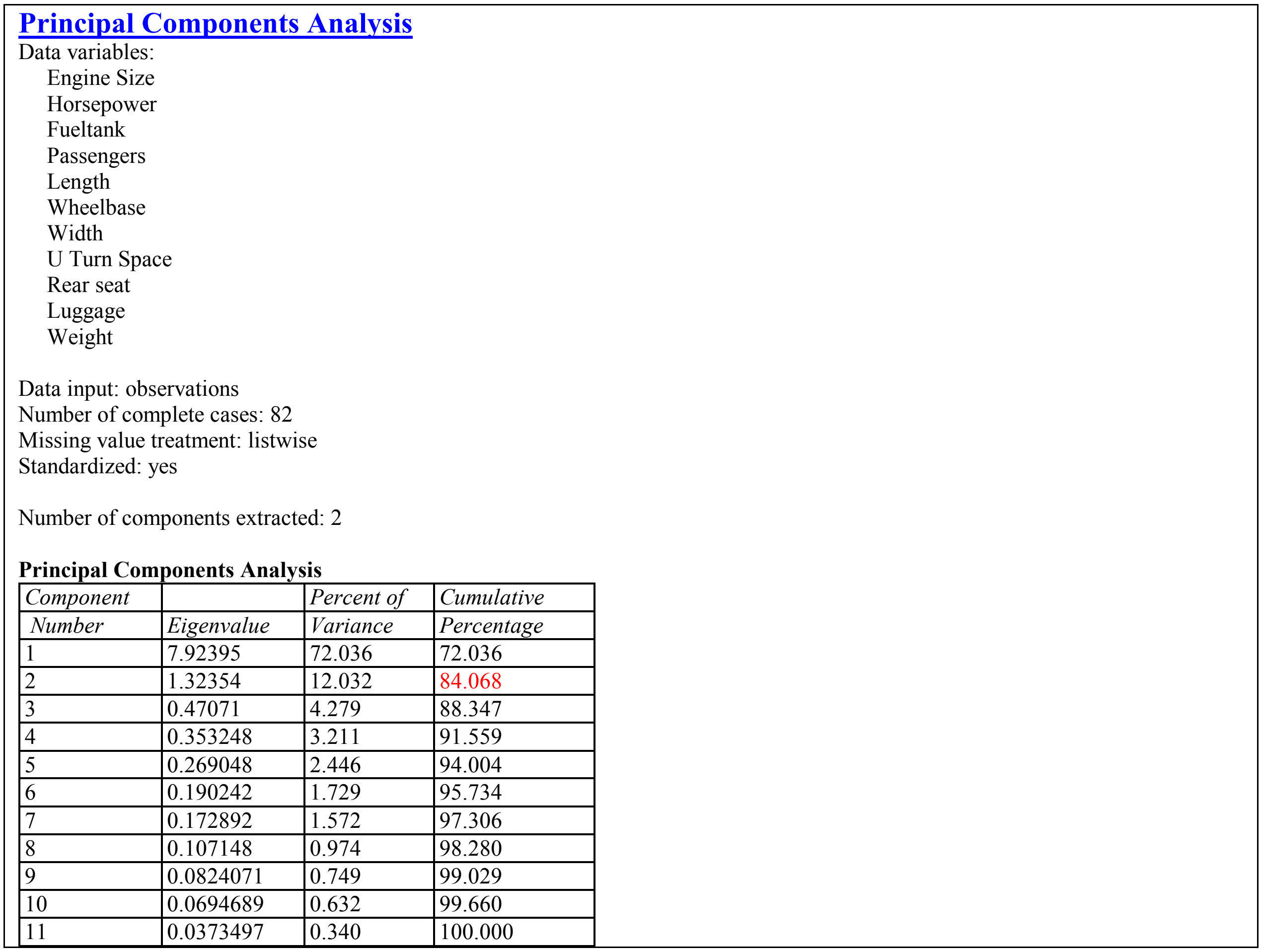
One criterion for selecting the number of principal components to extract is to select all components for which the corresponding eigenvalue is at least 1, implying that the component represents at least a fraction 1/*p* of the total population variance.

STATGRAPHICS gives the option of extracting principal components based on either the covariance matrix Σ or the correlation matrix ρ, depending on the *Standardize* setting on the

Analysis Options dialog box. When the variables are in different units, it is usually best to base the analysis on the correlation matrix (which is the default).

Analysis Summary

The *Analysis Summary* table is shown below:



Principal Components Analysis

Data variables:
  Engine Size
  Horsepower
  Fueltank
  Passengers
  Length
  Wheelbase
  Width
  U Turn Space
  Rear seat
  Luggage
  Weight

Data input: observations
Number of complete cases: 82
Missing value treatment: listwise
Standardized: yes

Number of components extracted: 2

Principal Components Analysis

Component		*Percent of*	*Cumulative*
Number	*Eigenvalue*	*Variance*	*Percentage*
1	7.92395	72.036	72.036
2	1.32354	12.032	84.068
3	0.47071	4.279	88.347
4	0.353248	3.211	91.559
5	0.269048	2.446	94.004
6	0.190242	1.729	95.734
7	0.172892	1.572	97.306
8	0.107148	0.974	98.280
9	0.0824071	0.749	99.029
10	0.0694689	0.632	99.660
11	0.0373497	0.340	100.000

Displayed in the table are:

- **Data variables**: the names of the *p* input columns.

- **Data input**: either *observations* or *matrix*, depending upon whether the input columns contain the original observations or the sample covariance matrix.

- **Number of complete cases**: the number of cases *n* for which none of the observations were missing.

- **Missing value treatment**: how missing values were treated in estimating the covariance or correlation matrix. If *listwise*, the estimates were based on complete cases only. If *pairwise*, all pairs of non-missing data values were used to obtain the estimates.

- **Standardized**: *yes* if the analysis was based on the correlation matrix. *No* if it was based on the covariance matrix.

- **Number of components extracted**: the number of components k extracted from the data. This number is based on the settings on the *Analysis Options* dialog box.

A table is also displayed showing information for each of the p possible principal components:

- **Component number**: the component number j, from 1 to p.

- **Eigenvalue**: the eigenvalue of the estimated covariance or correlation matrix, $\hat{\lambda}_j$.

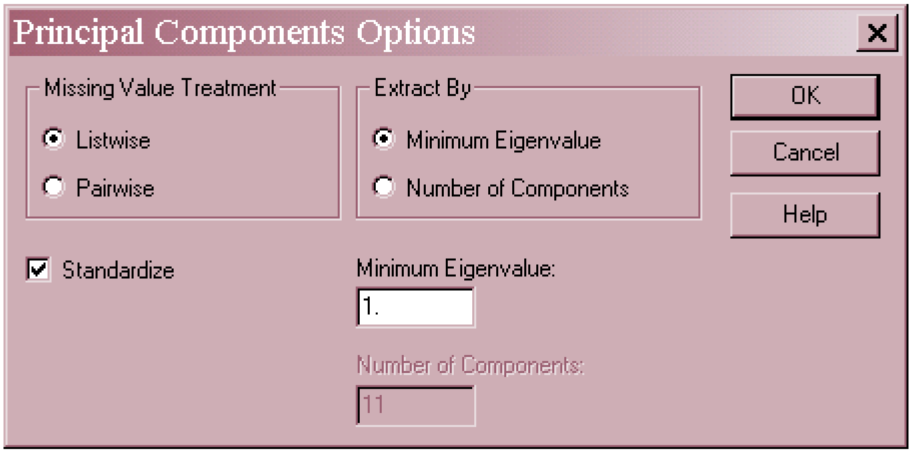
- **Percentage of variance**: the percent of the total estimated population variance represented by this component, equal to

$$100\left(\frac{\hat{\lambda}_j}{\hat{\lambda}_1+\hat{\lambda}_2+\ldots+\hat{\lambda}_k}\right)\% \tag{6}$$

- **Cumulative percentage**: the cumulative percentage of the total estimated population variance accounted for by the first j components.

In the example, the first k = 2 components account for over 84% of the overall variance amongst the 11 variables.

Analysis Options



Principal Components Options

Missing Value Treatment

● Listwise

○ Pairwise

Extract By

● Minimum Eigenvalue

○ Number of Components

☑ Standardize

Minimum Eigenvalue: 1.

Number of Components: 11

- **Missing Values Treatment**: method of handling missing values when estimating the sample covariances or correlations. Specify *Listwise* to use only cases with no missing values for any of the input variables. Specify *Pairwise* to use all pairs of observations in which neither value is missing.

- **Standardize**: check this box to base the analysis on the sample correlation matrix rather than the sample covariance matrix. This corresponds to standardizing each input variable before calculating the covariances, by subtracting its mean and dividing by its standard deviation.

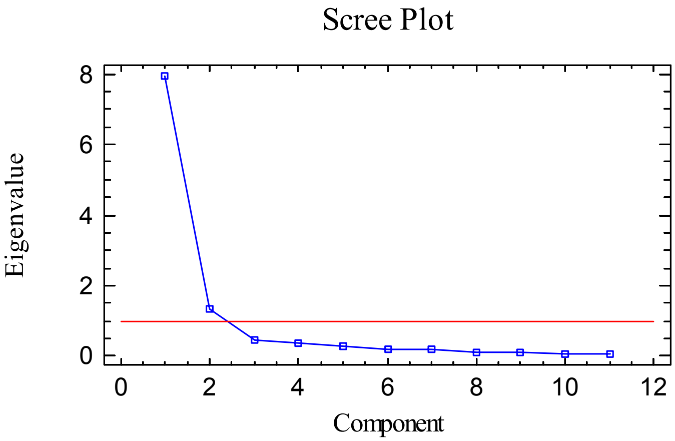
- **Extract By**: the criterion used to determine the number of principal components to extract.

- **Minimum Eigenvalue**: if extracting by magnitude of the eigenvalues, the smallest eigenvalue for which a component will be extracted.

- **Number of Components**: if extracting by number of components, the number *k*.

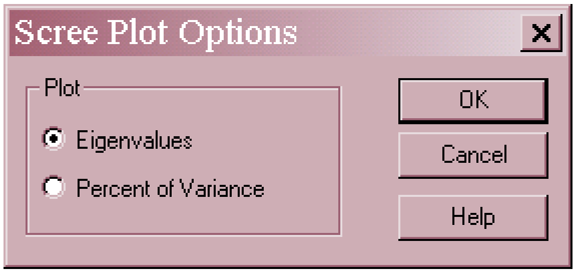
Scree Plot

The *Scree Plot* can be very helpful in determining the number of principal components to extract. By default, it plots the size of the eigenvalues corresponding to each of the *p* possible components:



A horizontal line is added at the minimum eigenvalue specified on the *Analysis Options* dialog box. In the plot above, note that only the first 2 components have large eigenvalues.

Pane Options


Scree Plot Options

Plot

● Eigenvalues

○ Percent of Variance

- **Plot**: value plotted on the vertical axis.

Component Weights

The *Component Weights* pane shows the estimated value of the coefficients *a* for each component extracted:

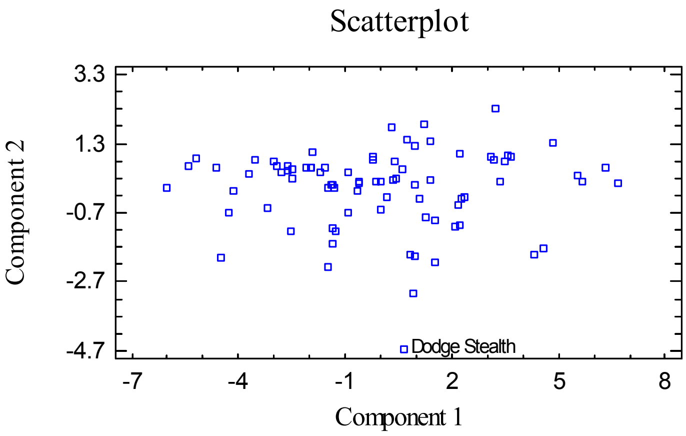
Table of Component Weights

	Component	*Component*
	1	2
Engine Size	0.332726	-0.133891
Horsepower	0.268123	-0.442852
Fueltank	0.311244	-0.210124
Passengers	0.238683	0.530291
Length	0.335379	0.02122
Wheelbase	0.335386	0.0610323
Width	0.324896	-0.134248
U Turn Space	0.299218	-0.0830471
Rear seat	0.231256	0.53351
Luggage	0.276494	0.322776
Weight	0.337017	-0.206599

The weights within each column often have interesting interpretations. In the example, note that the weights in the first column are all approximately the same. This implies that the first component is basically an average of all of the input variables. The second component is weighted most heavily in a positive direction on the number of *Passengers*, the *Rear Seat* room, and the amount of *Luggage* space, and in a negative direction on *Horsepower*. It likely differentiates amongst the different types of vehicles.

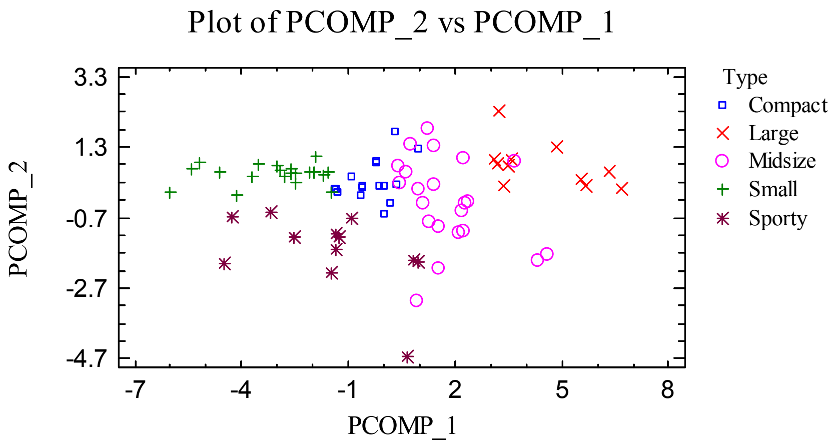
2D and 3D Component Plots

These plots display the values of 2 or 3 selected principal components for each of the *n* cases.



It is useful to examine any points far away from the others, such as the highlighted *Dodge Stealth*, which has a very low value for the second component.

An interesting variation of this plot is one in which the variables are coded according to another column, such as the type of vehicle:

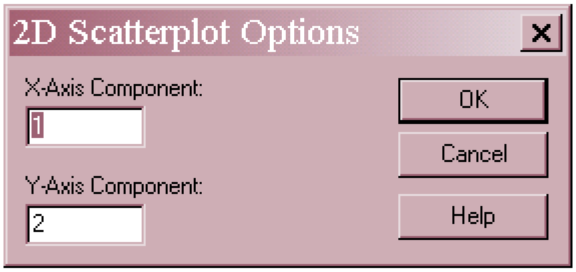


To produce the above plot:

1. Press the *Save Results* button and save the *Principal Components* to new columns of the datasheet.
2. Select the *X-Y Plot* procedure from the top menu and input the new columns.
3. Select *Pane Options* and specify *Type* in the *Point Codes* field.

It is now clear that the first component is related to the size of the vehicle, while the second component separates the sporty cars from the others.

Pane Options


2D Scatterplot Options

X-Axis Component: 1

Y-Axis Component: 2

Specify the components to plot on each axis.

Data Table

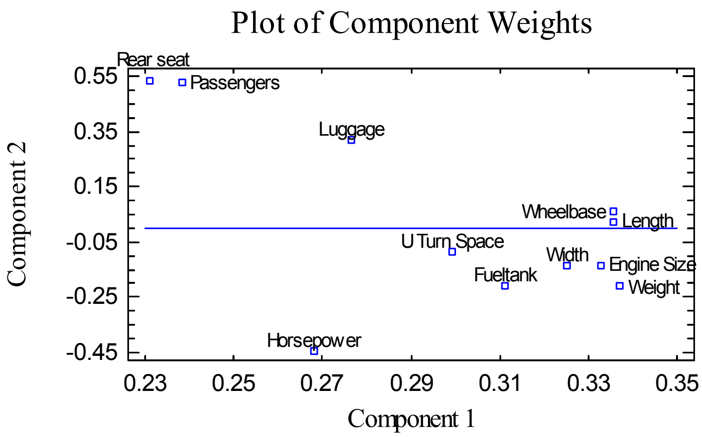
The *Data Table* display the values of the principal components for each of the *n* cases.

Table of Principal Components

		Component	*Component*
Row	*Label*	*1*	*2*
1	Integra	-1.49203	0.00673575
2	Legend	2.37408	-0.247278
3	90	0.165636	-0.261873
4	100	2.23212	1.01524
5	535i	1.52815	-2.15174
6	Century	0.723227	1.39817
7	LeSabre	3.46805	0.778351
8	Roadmaster	6.6603	0.133406
9	Riviera	2.24466	-1.07736

2D and 3D Component Plots

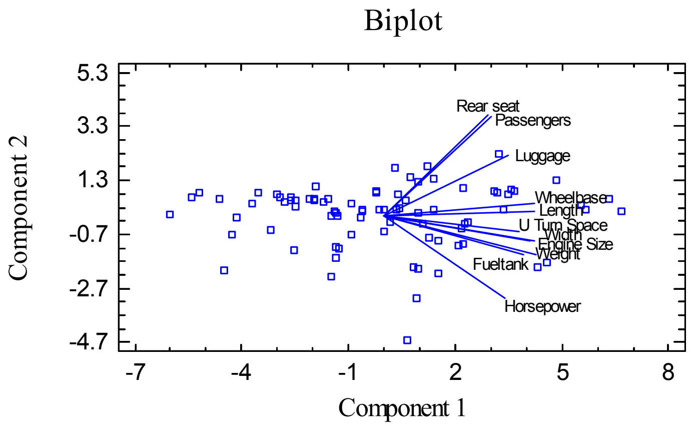
The *Component Plots* show the location of each variable in the space of 2 or 3 selected components:



Variables furthest from the reference lines at 0 make the largest contribution to the components.

2D and 3D Biplots

The *Biplots* display both the observations and the variables on a single plot.



The point symbols correspond to the observations. The ends of the solid lines correspond to the variables.

Save Results

The following results may be saved to the datasheet:

1. *Eigenvalues* – the *k* eigenvalues.

2. *Component weights* – *k* columns, each containing *p* estimates of the coefficients *a*.

3. *Principal Components* – *k* columns, each containing *n* values corresponding to the extracted principal components.