Chapter 29
Details and Examples

Chapter Table of Contents

Chapter 29
Details and Examples

Details

This chapter provides details on the following topics:

- terminology
- labels for chart features
- scaling the cumulative percent curve
- creating output data sets
- constructing effective Pareto charts
- missing values
- the role of variable formats
- large data sets

The "Examples" section illustrates these topics with several detailed examples.

Terminology

Basic Pareto Charts

A basic Pareto chart (see Figure 26.1 on page 801) analyzes the unique values of a *process variable*, which are referred to as *Pareto categories* or *levels*. These values typically represent problems encountered during some phase of a manufacturing or service activity.

A basic vertical Pareto chart (as produced by the Pareto procedure's VBAR statement) has one horizontal and two vertical axes. The horizontal (or *category*) axis is displayed at the bottom of the chart and lists the Pareto categories. The *primary vertical axis* (or *frequency axis*) is displayed on the left. The relative frequency of each Pareto category is represented by a vertical bar whose height is measured on the primary vertical axis. You can use the SCALE= option to scale this axis in percent, count, or weight units. The *secondary vertical axis* (or *cumulative percent axis*) is displayed on the right. This axis is scaled in cumulative percent units and is used to read the *cumulative percent curve*. The height of each point on the curve represents the percent of the total frequency accounted for by the Pareto categories to the left of the point.

In a horizontal Pareto chart (as produced by the HBAR statement), the category axis is displayed vertically on the left. Categories appear in order of decreasing relative frequency from top to bottom. The frequency axis appears at the top of the chart and the cumulative percent axis is at the bottom. The relative frequencies of the Pareto categories are represented by horizontal bars. A point on the cumulative percent curve represents the percent of the total frequency accounted for by the Pareto categories above that point.

Note: For the sake of brevity, in this chapter the term *height* is used to refer to the size of a bar as measured along the frequency axis, whether the Pareto chart is oriented vertically or horizontally.

Restricted Pareto Charts

A *restricted Pareto chart* (see Figure 26.6 on page 805) displays only the n most frequently occurring categories in a data set that contains N categories, where $N > n$. The remaining $N - n$ categories are dropped or are merged into a single "other" category created with the OTHER= option. The MAXCMPCT=, MAXNCAT=, and MINPCT= options provide alternative methods for specifying n. See the entries for these options in "Dictionary of Options" on page 813.

Weighted Pareto Charts

A *weighted Pareto chart* (see Example 29.8 on page 914) displays bars whose heights represent the weighted frequencies of the categories. Typical weights are the cost of repair or the loss incurred by the customer.

The weight W_i for the i^{th} Pareto category is computed as

$$W_i = \sum_{u \in \mathcal{C}_i} w(u) f(u)$$

where $\mathcal{C}_i$ is the set of observations that make up the i^{th} category, $w(u)$ is the value of the weight variable in the u^{th} observation, and $f(u)$ is the value of the frequency variable in the u^{th} observation (taking $f(u) \equiv 1$ if a FREQ= variable is not specified). If SCALE=WEIGHT is specified, the height of the bar for the i^{th} category is W_i. If SCALE=PERCENT is specified, the height of this bar is

$$\frac{100 W_i}{\sum_{j=1}^{N} W_j}$$

where N is the total number of categories.

Comparative Pareto Charts

A *comparative Pareto chart* combines two or more Pareto charts for the same process variable. The component charts are displayed with uniform axes to facilitate comparison. The observations represented by a components chart are referred to as a *cell*. The framed areas for the component charts are referred to as *tiles*.

In a *one-way comparative Pareto chart*, each component chart corresponds to a different level of a single classification variable specified with the CLASS= option. The component charts are arranged in a stack or a row, as illustrated in Output 29.1.3 (page 897), Output 29.2.2 (page 902), and Output 29.2.3 (page 903). In a *two-way comparative Pareto chart*, each component chart corresponds to a different combination of levels of two classification variables specified with the CLASS= option. The component charts are arranged in a matrix, as illustrated in Output 29.2.4 on page 904.

In any comparative Pareto chart there is a *key cell*, in which the bars are in decreasing order and whose order is imposed on all the other cells to achieve a uniform category axis. By default, the key cell is the cell in the upper left corner, but you can use the CLASSKEY= option to designate any other cell as the key cell. In this case, the rows and columns of the comparative chart will be rearranged so that the key cell appears in the upper left. However, if you require the rows and columns in a particular order, you can specify the NOKEYMOVE option in conjunction with the CLASSKEY= option to suppress the rearrangement.

If you are creating your chart with a graphics device, you can use the NROWS= and NCOLS= options to specify the numbers of rows and columns in a comparative Pareto chart. By default, NROWS=2 and NCOLS=1 for a one-way comparison and NROWS=2 and NCOLS=2 for a two-way comparison. There is no upper limit to the number of rows or columns that you can specify, but in practice the limit is determined by the display area of your graphics device. If the numbers of classification variable levels exceed the NROWS= and NCOLS= values, the chart is created on multiple screens or pages.

If the same set of Pareto categories does not occur in each cell of a comparative Pareto chart, the categories are said to be *unbalanced*. In this case, the procedure uses the following convention to construct the uniform category axis. First, the categories that occur in the key cell are arranged on the category axis from left to right (top to bottom for a horizontal chart), sorted in decreasing order of frequency, with tied levels arranged in order of their formatted values. The categories not in the key cell are assigned frequencies of zero in the key cell, and they are arranged at the right (bottom) of the category axis, where they are ordered by their formatted values. This arrangement is simply a convention of the procedure and should not be interpreted to mean that one category is more important than another.

Whether the categories in the input data set are balanced or not, the categories in the OUT= data set are always balanced. The procedure balances this data set by assigning values of zero to the _COUNT_ and _PCT_ variables as necessary.

Unbalanced categories present a special problem when the MAXNCAT= option is used to restrict the number of categories displayed on the chart. For instance, suppose that you specify MAXNCAT=12 and there are 15 categories in all, 10 of which occur in the key cell. Since there is no unambiguous method for selecting two of the remaining five categories to complete the restricted list, the procedure reduces the restricted list to the categories that occur in the key cell and displays only those 10 categories. A warning message is issued in the SAS log.

Labels for Chart Features

The following table summarizes methods for labeling the features of Pareto charts.

Table 29.1. Labeling Features of Pareto Charts

Feature	Method for Specifying Label
titles	TITLE*n* statements
footnotes	FOOTNOTE*n* statements
category axis	*process variable* label
primary vertical axis (VBAR only)	VAXISLABEL= option
secondary vertical axis (VBAR only)	VAXIS2LABEL= option
primary horzontal axis (HBAR only)	HAXISLABEL= option
secondary horzontal axis (HBAR only)	HAXIS2LABEL= option
bars	BARLABEL= option
points on cumulative percent curve	CMPCTLABEL= option
rows and columns	CLASS= variable labels
cells	NLEGEND option or NLEGEND= variable
category legend	CATLEGLABEL= option
high/low bar legend	HLLEGLABEL= option
bar color legend	BARLEGLABEL= option
tile legend	TILELEGLABEL= option
annotation	ANNOTATE= and ANNOTATE2= data sets

Scaling the Cumulative Percent Curve

Pareto charts shown in textbooks typically scale the cumulative percent curve so that it is anchored at the top right corner of the leftmost bar. The upper end of the primary vertical axis is then extended to accommodate the curve. For an illustration, see Output 29.2.1 on page 901. By default, the PARETO procedure uses the top right corner as the anchor position on a vertical chart and the bottom right corner of the topmost bar as the anchor position on a horizontal chart. You can override the default with the ANCHOR= option.

This method of scaling is not feasible if the number of categories is very large and if the Pareto distribution is uniform. In this case, the bars are excessively compressed relative to the curve. Conversely, this method excessively compresses the curve relative to the bars when you use a count scale for the frequency axis in a comparative Pareto chart and the tallest bar does not occur in the key cell. In either situation, the procedure overrides the textbook scaling method and balances the scales of the bars and the curve.

You can use the AXISFACTOR= option to specify the extent to which the frequency axis should be extended. Alternatively, you can extend the frequency axis by using the VBAR statement VAXIS= option or HBAR statement HAXIS= option to specify the tick mark values for the axis.

Another scaling anomaly is illustrated by the comparative Pareto chart in Output 29.1.4 on page 898. Here, the cumulative percent curve in the bottom chart is not anchored due to the combination of a uniform count scale and different sample sizes in the two cells.

Output Data Sets

The OUT= data set saves the information displayed on a Pareto chart. If you specify CLASS= variables, the OUT= data set contains one block of observations for each combination of levels of the CLASS= variables, and within each block there is an observation. The observations are sorted in the order in which the categories are displayed on the chart. The following variables read from a DATA= data set are saved in an OUT= data set:

- process variables
- CLASS= variables
- BY variables
- WEIGHT= variables
- the CTILES= variable
- the TILELEGEND= variable
- the NLEGEND= variable
- CBARS= variables
- PBARS= variables
- BARLEGEND= variables

In addition, the OUT= data set contains the following variables that are created during the analysis:

- _COUNT_, which saves the frequency count for each Pareto category
- _WCOUNT_, which saves the weighted count for each category. This variable is created only when you specify the WEIGHT= option.
- _PCT_, which saves the percent of the total count for each category. If you specify the WEIGHT= option, the variable _PCT_ saves the percent of the total weighted count.
- _CMPCT_, which saves the cumulative percent for each category

See Output 29.8.2 on page 915 for an example of an OUT= data set.

If you specify the MAXNCAT=, MAXCMPCT=, or MINPCT= option, the OUT= data set saves only the categories displayed on the chart. If you create an OTHER=

category that merges the remaining categories, an additional observation is saved with the new category. Since the OTHER= value is defined as a formatted value of the process variable, you should also specify a corresponding internal value, as follows:

- If the process variable is a character variable, specify the internal value with the OTHERCVAL= option. If you do not specify this value, the OTHER= value is saved as the internal value.
- If the process variable is a numeric variable, specify the internal value with the OTHERNVAL= option. If you do not specify this value, an internal missing value is saved.

Constructing Effective Pareto Charts

The following are recommendations for improving the visual clarity of Pareto charts:

- Decide carefully how the bars should be scaled. The default percent scale is not always the best choice. For instance, a count scale may be more appropriate in a comparative Pareto chart where the total count per cell varies widely from cell to cell and where you want to compare Pareto distributions on an *absolute* scale rather than a *relative* scale. You can request a count scale by specifying SCALE=COUNT. In other situations, it may be more appropriate to use a weighted percent scale or a weighted count scale (specify a WEIGHT= variable and either SCALE=PERCENT or SCALE=WEIGHT).
- Use a weight variable if the counts are dependent on a factor such as exposure or opportunity that varies from one category to another. For instance, suppose that you are creating a Pareto chart for the number of medical claims submitted by company employees categorized by job title. The counts can be weighted to adjust for the fact that there are more individuals in some jobs than in others and for the fact that some jobs may be associated with greater health risks than others.
- Use the NOCURVE option to eliminate the cumulative percent curve in situations where the curve reveals little information about the data. In general, the bars should be more prominent than the curve.
- Maximize the space used for the bars by eliminating unnecessary labels and visual clutter. This is particularly important for comparative Pareto charts. The NOHLABEL and NOVLABEL options are useful for this purpose. You can also use the NOVLABEL2, NOVTICK, and NOVTICK2 options with a VBAR statement or the NOHLABEL2, NOHTICK and NOHTICK2 options with an HBAR statement.
- Make legends more informative by specifying legend labels.

- Avoid filling bars with multiple types of cross-hatched patterns; solid color fills are less distracting. Use color sparingly to emphasize important features (such as the "vital few" categories), and choose bar colors that provide good visual discrimination.
- If you are working with a large data set involving many categories, limit the number displayed to achieve visual clarity.
- If your application involves classification effects, construct more than one Pareto chart for the data using various combinations of classification variables (this approach is illustrated in Example 29.2 on page 899.
- Provide reference lines on comparative Pareto charts to aid visual comparison.

Refer to Chapter 2 of Cleveland (1985) for a general discussion of the principles of statistical graphics.

Missing Values

By default, observations with missing values of a process variable are not processed. If you specify the MISSING option, then missing values are treated as a Pareto category.

Likewise, observations with missing values of the CLASS= variables are not processed by default. Missing values of the first CLASS= variable are treated as a level if the MISSING1 option is specified, and missing values of the second CLASS= variable are treated as a level if the MISSING2 option is specified.

Role of Variable Formats

The categories of a Pareto chart are always determined using formatted values of the process variable, and the format is used to label the categories.

On the chart, the categories are displayed in decreasing order of frequency. If there are multiple categories with the same count, the tied categories are displayed in order of their formatted values.

When you create a comparative Pareto chart, the formatted levels of the CLASS= variables are used to group the data into cells. There is a cell for each level of the CLASS= variable in a one-way comparative chart, and there is a cell for each combination of levels of the CLASS= variables in a two-way comparative chart.

You can specify the order of the rows and columns corresponding to the classification levels with the ORDER1= and ORDER2= options. The default value of these options is INTERNAL, which means that the order is determined by the internal values of the CLASS= variables. It is possible for a particular formatted value to correspond to more than one internal value. To resolve this ambiguity, the internal value that determines the position of the row or column is the value that occurs first in the input data set.

Other values that you can specify for the ORDER1= and ORDER2= options are FORMATTED, FREQ, and DATA. Detailed descriptions of these options are provided on pages 826–827.

Large Data Sets

While there is no limit to the number of observations that can be read from an input data set, the maximum number of Pareto categories that can be read is 32,767. This limit is a practical issue only if you are creating a restricted Pareto chart from a large data set, since the number of categories that can be displayed is limited by the resolution of your graphics device. The number of categories that can be read is limited by the amount of memory available, since the levels are stored in memory. If you run out of memory, you should first reduce the data with the FREQ procedure.

Examples

Example 29.1. Creating Before-and-After Pareto Charts

See PARETO7 in the SAS/QC Sample Library

During the manufacture of a metal-oxide semiconductor (MOS) capacitor, causes of failures were recorded before and after a tube in the diffusion furnace was cleaned. This information was saved in a SAS data set named FAILURE3.

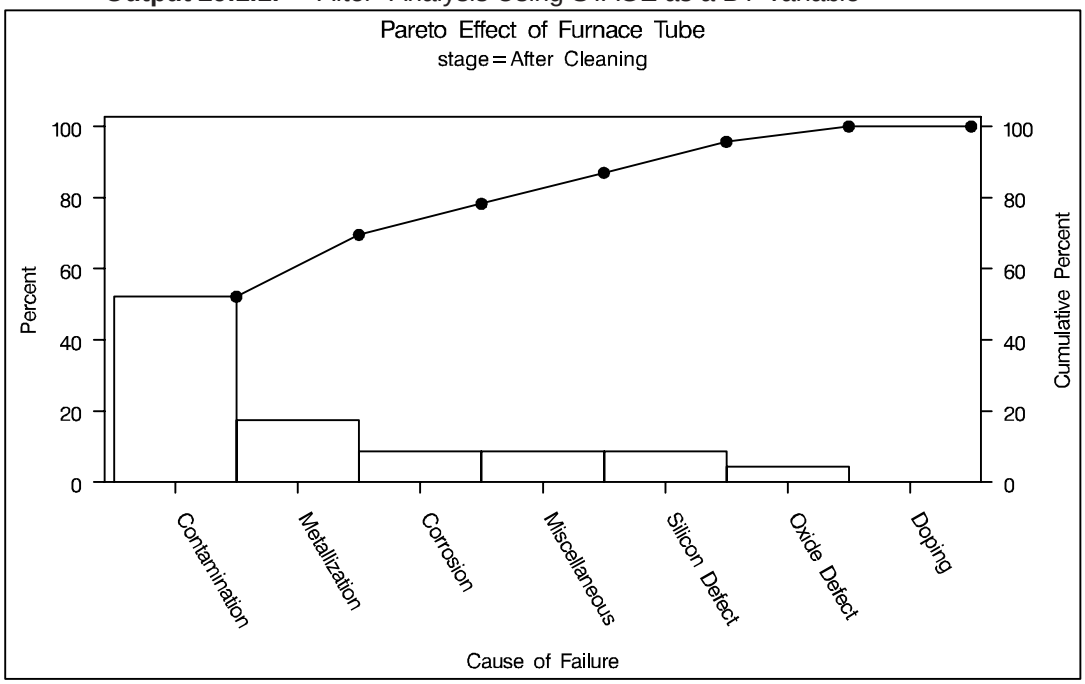
```
data failure3;
   length cause $ 16 stage $ 16 ;
   label  cause = 'Cause of Failure' ;
   input  stage $ 1-16 cause $ 19-34 counts;
   datalines;
Before Cleaning   Contamination    14
Before Cleaning   Corrosion         2
Before Cleaning   Doping            1
Before Cleaning   Metallization     2
Before Cleaning   Miscellaneous     3
Before Cleaning   Oxide Defect      8
Before Cleaning   Silicon Defect    1
After Cleaning    Doping            0
After Cleaning    Corrosion         2
After Cleaning    Metallization     4
After Cleaning    Miscellaneous     2
After Cleaning    Oxide Defect      1
After Cleaning    Contamination    12
After Cleaning    Silicon Defect    2
;
```

To compare distribution of failures before and after cleaning, you can create two separate Pareto charts, one for the observations in which STAGE is equal to `Before Cleaning` and one for the observations in which STAGE is equal to `After Cleaning`. You can do this with the BY statement.

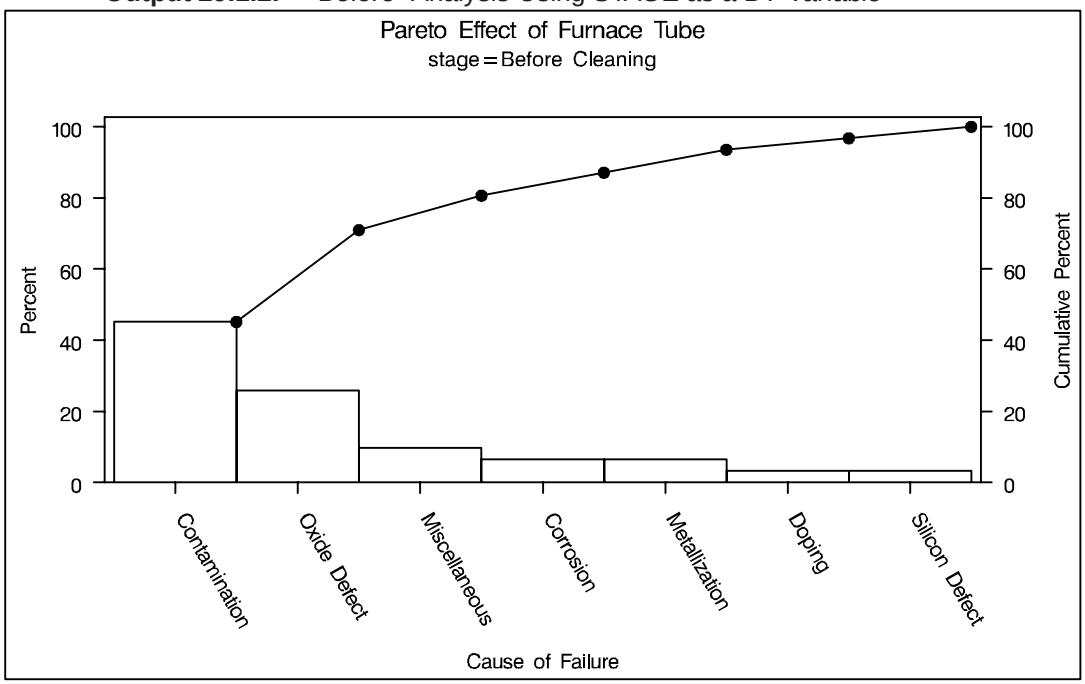
```
proc sort data=failure3;
   by stage;

title 'Pareto Effect of Furnace Tube' ;
symbol v=dot;
proc pareto data=failure3;
   vbar cause / freq  = counts
                angle = -60 ;
   by stage;
run;
```

The SORT procedure sorts the observations in order of the values of STAGE. It is not necessary to sort by the values of CAUSE since this is done by the PARETO procedure. The two charts, displayed in Output 29.1.1 and Output 29.1.2, reveal a reduction in oxide defects after the tube was cleaned. This is a relative reduction, since the primary axes are scaled in percent units.

Output 29.1.1. "After" Analysis Using STAGE as a BY Variable



Output 29.1.2. "Before" Analysis Using STAGE as a BY Variable



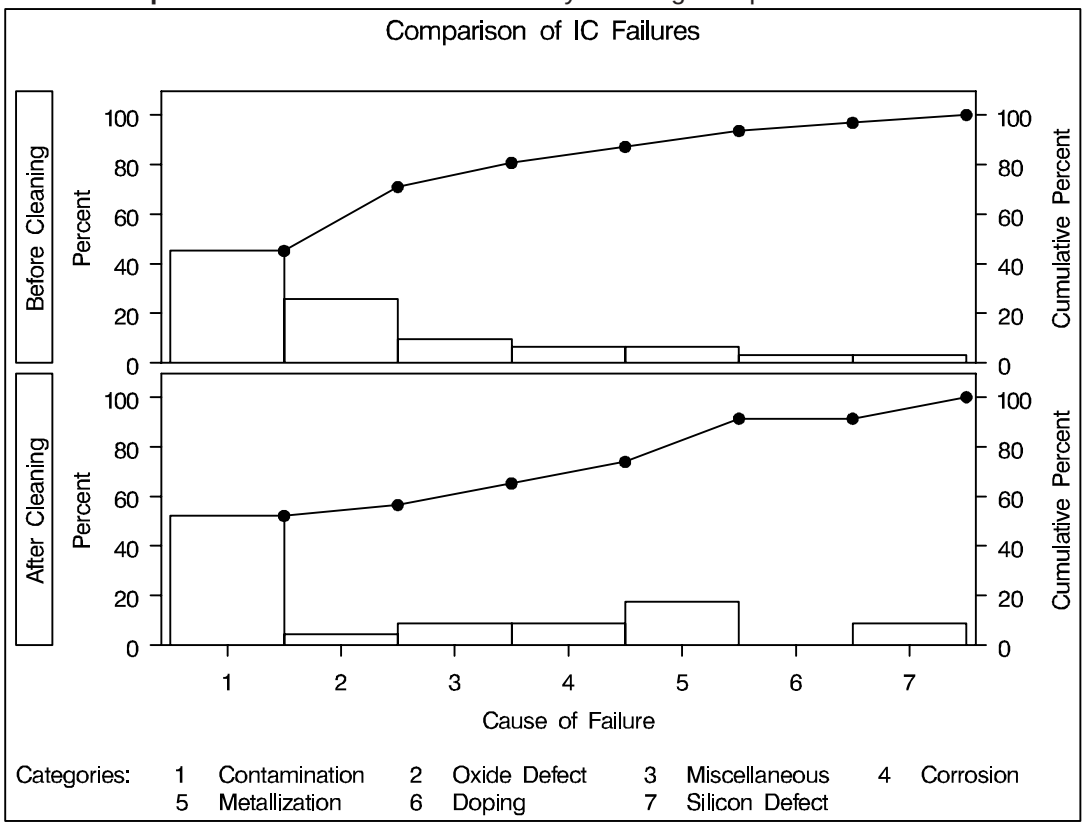
In general, it is difficult to compare Pareto charts created with BY processing because their axes are not necessarily uniform. A better approach is to construct a comparative Pareto chart, as illustrated by the following statements:

```
title 'Comparison of IC Failures' ;
proc pareto data=failure3;
   vbar cause / class      = stage
                freq       = counts
                intertile  = 1.0
                classkey   = 'Before Cleaning' ;
run;
```

See PARETO8 in the SAS/QC Sample Library

The CLASS= option designates STAGE as a classification variable, and this directs the procedure to create the one-way comparative Pareto chart, shown in Output 29.1.3, that displays a component chart for each level of STAGE.

Output 29.1.3. Before-and-After Analysis Using Comparative Pareto Chart



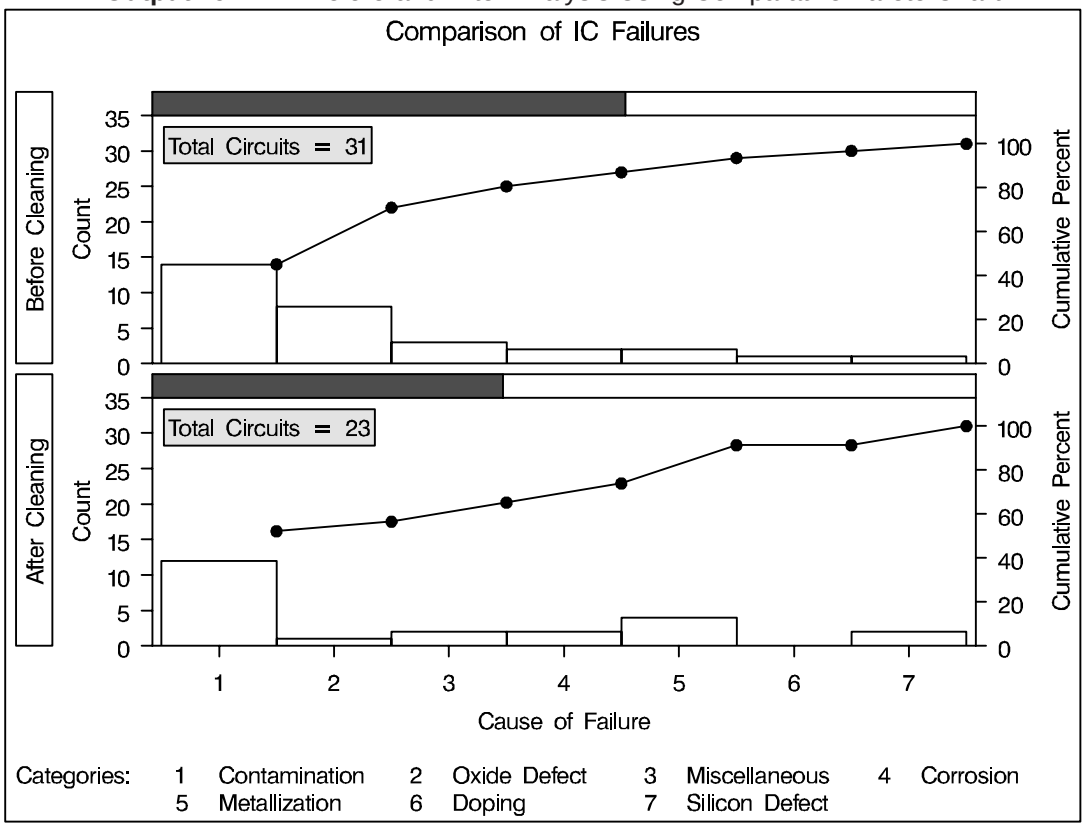
In a comparative Pareto chart, there is always one special cell, called the *key cell*, in which the bars are displayed in decreasing order, and whose order determines the uniform horizontal axis used for all the cells. The key cell is positioned at the top of the chart. Here, the key cell is the set of observations for which STAGE equals `Before Cleaning`, as specified by the CLASSKEY= option. By default, the levels are sorted in the order determined by the ORDER1= option, and the key cell is the the level that occurs first in this order.

In many applications, it may be more revealing to base comparisons on counts rather than percents. The following statements construct a chart with a frequency scale:

```
title 'Comparison of IC Failures' ;
proc pareto data=failure3;
   vbar cause / class      = stage
                freq       = counts
                scale      = count
                intertile  = 1.0
                nlegend    = 'Total Circuits'
                cframenleg = yellow
                cprop      = red
                classkey   = 'Before Cleaning' ;
run;
```

The chart is shown in Output 29.1.4.

Output 29.1.4. Before-and-After Analysis Using Comparative Pareto Chart



Specifying SCALE=COUNT scales the primary vertical axis in frequency units. The NLEGEND= option adds a sample size legend, and the CFRAMENLEG= option frames the legend. The CPROP= option adds bars that indicate the proportion of total frequency represented by each cell. The INTERTILE= option separates the tiles with a small offset.

Note that the lower cumulative percent curve in Output 29.1.4 is not anchored to the first bar. This is a consequence of the uniform frequency scale and of the fact that the number of observations in each cell is not the same.

Example 29.2. Creating Two-Way Comparative Pareto Charts

See PARETO9 in the SAS/QC Sample Library

During the manufacture of a MOS capacitor, different cleaning processes were used by two manufacturing systems operating in parallel. Process A used a standard cleaning solution, while Process B used a different cleaning mixture that contained less particulate matter. The failure causes observed with each process for five consecutive days were recorded and saved in a SAS data set called FAILURE4.

```
data failure4;
   label cause = 'Cause of Failure' ;
   input process $ 1-9 day $ 13-19 cause $ 23-36 counts 40-41;
   datalines;
Process A   March 1   Contamination    15
Process A   March 1   Corrosion         2
Process A   March 1   Doping            1
Process A   March 1   Metallization     2
Process A   March 1   Miscellaneous     3
Process A   March 1   Oxide Defect      8
Process A   March 1   Silicon Defect    1
Process A   March 2   Contamination    16
Process A   March 2   Corrosion         3
Process A   March 2   Doping            1
Process A   March 2   Metallization     3
Process A   March 2   Miscellaneous     1
Process A   March 2   Oxide Defect      9
Process A   March 2   Silicon Defect    2
Process A   March 3   Contamination    20
Process A   March 3   Corrosion         1
Process A   March 3   Doping            1
Process A   March 3   Metallization     0
Process A   March 3   Miscellaneous     3
Process A   March 3   Oxide Defect      7
Process A   March 3   Silicon Defect    2
Process A   March 4   Contamination    12
Process A   March 4   Corrosion         1
Process A   March 4   Doping            1
Process A   March 4   Metallization     0
Process A   March 4   Miscellaneous     0
Process A   March 4   Oxide Defect     10
Process A   March 4   Silicon Defect    1
Process A   March 5   Contamination    23
Process A   March 5   Corrosion         1
Process A   March 5   Doping            1
Process A   March 5   Metallization     0
Process A   March 5   Miscellaneous     1
Process A   March 5   Oxide Defect      8
Process A   March 5   Silicon Defect    2
Process B   March 1   Contamination     8
Process B   March 1   Corrosion         2
Process B   March 1   Doping            1
Process B   March 1   Metallization     4
Process B   March 1   Miscellaneous     2
Process B   March 1   Oxide Defect     10
```

```
Process B   March 1   Silicon Defect    3
Process B   March 2   Contamination     9
Process B   March 2   Corrosion         0
Process B   March 2   Doping            1
Process B   March 2   Metallization     2
Process B   March 2   Miscellaneous     4
Process B   March 2   Oxide Defect      9
Process B   March 2   Silicon Defect    2
Process B   March 3   Contamination     4
Process B   March 3   Corrosion         1
Process B   March 3   Doping            1
Process B   March 3   Metallization     0
Process B   March 3   Miscellaneous     0
Process B   March 3   Oxide Defect     10
Process B   March 3   Silicon Defect    1
Process B   March 4   Contamination     2
Process B   March 4   Corrosion         2
Process B   March 4   Doping            1
Process B   March 4   Metallization     0
Process B   March 4   Miscellaneous     3
Process B   March 4   Oxide Defect      7
Process B   March 4   Silicon Defect    1
Process B   March 5   Contamination     1
Process B   March 5   Corrosion         3
Process B   March 5   Doping            1
Process B   March 5   Metallization     0
Process B   March 5   Miscellaneous     1
Process B   March 5   Oxide Defect      8
Process B   March 5   Silicon Defect    2
;
```

In addition to the process variable CAUSE, there are two classification variables in this data set: PROCESS and DAY. The variable COUNTS is a frequency variable.

This example creates a series of displays that progressively use more of the classification information.

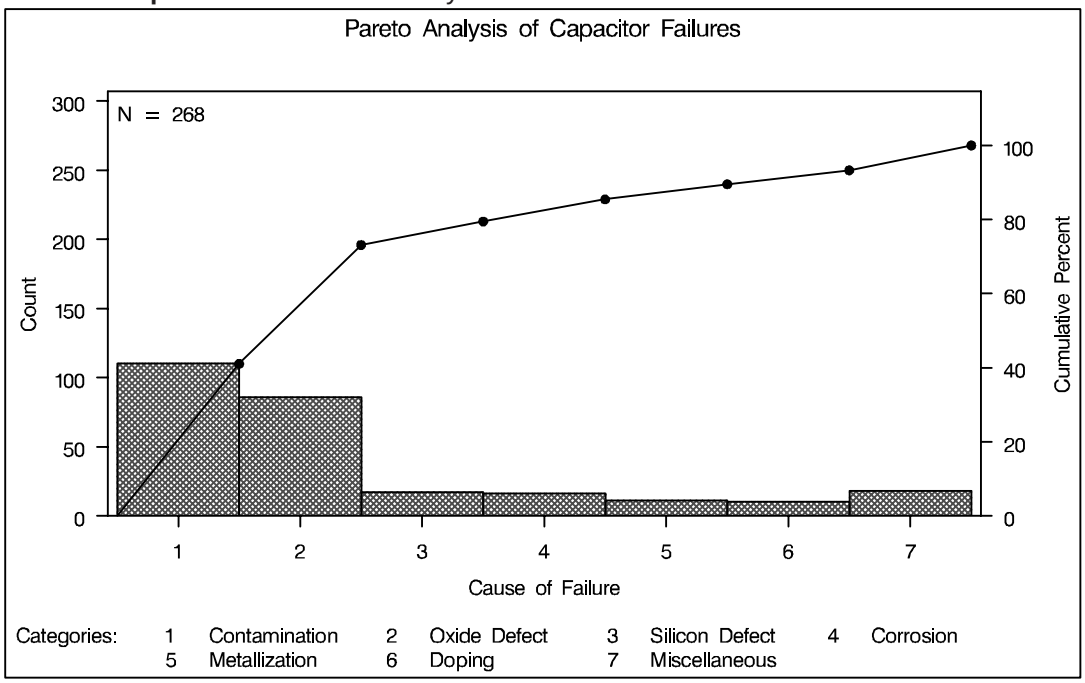
Basic Pareto Chart

The first display, created with the following statements, analyzes the process variable without taking into account the classification variables.

```
title 'Pareto Analysis of Capacitor Failures' ;
symbol v=dot;
proc pareto data=failure4;
   vbar cause / freq        = counts
                last        = 'Miscellaneous'
                scale       = count
                cbars       = green
                pbars       = m5x45
                anchor      = bl
                nlegend ;
run;
```

The chart, shown in Output 29.2.1, indicates that contamination is the most frequently occurring problem.

Output 29.2.1. Pareto Analysis without Classification Variables



The color and pattern for the bars are specified with the CBARS= and PBARS= options. The pattern M5X45 is a particular type of crosshatching (refer to *SAS/GRAPH Software: Reference* for a pattern selection guide). If you specify a color but not a pattern, the bars are filled with a solid color.

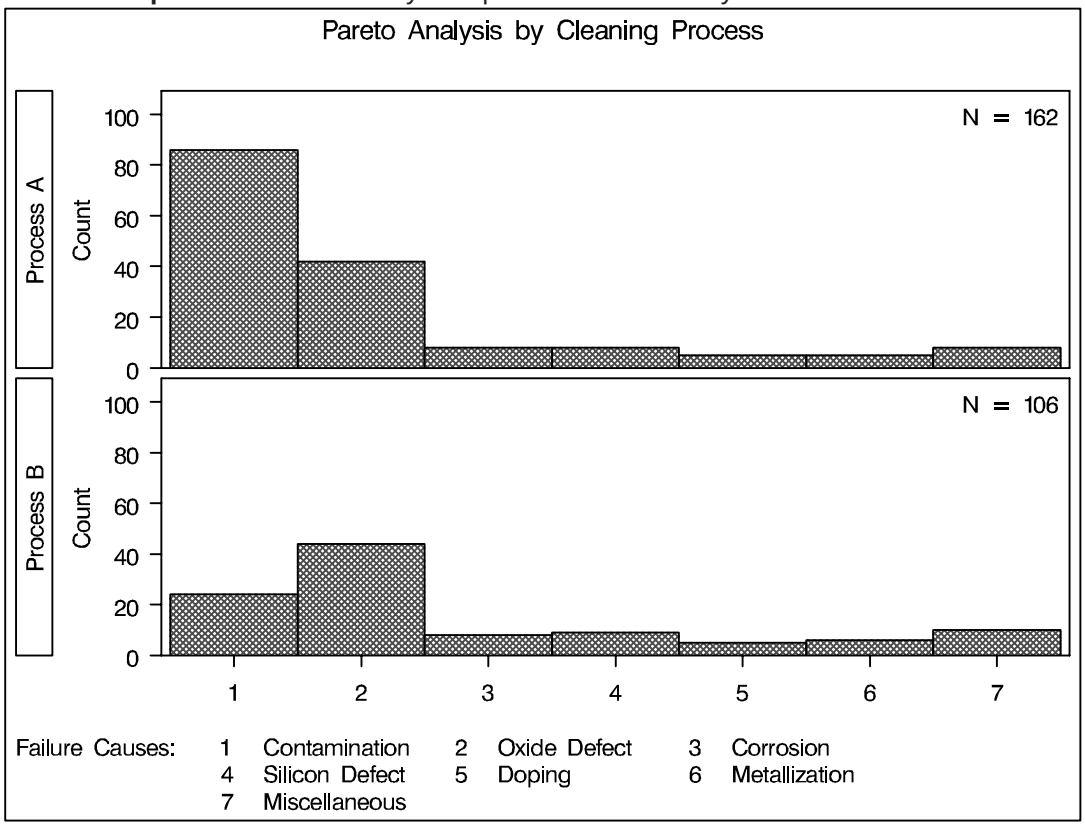
The option ANCHOR=BL anchors the cumulative percent curve at the bottom left (BL) of the first bar. The NLEGEND option adds a sample size legend.

One-Way Comparative Pareto Chart for PROCESS

The following statements specify PROCESS as a classification variable to create the comparative Pareto chart displayed in Output 29.2.2:

```
title 'Pareto Analysis by Cleaning Process';
proc pareto data=failure4;
   vbar cause / class       = process
                freq        = counts
                last        = 'Miscellaneous'
                scale       = count
                cbars       = green
                pbars       = m5x45
                catleglabel = 'Failure Causes:'
                intertile   = 1.0
                nohlabel
                nocurve
                nlegend ;
run;
```

Output 29.2.2. One-Way Comparative Pareto Analysis with CLASS=PROCESS



Each cell corresponds to a level of the CLASS= variable (PROCESS). By default, the cells are arranged from top to bottom in alphabetical order of the formatted values of PROCESS, and the key cell is the top cell. The main difference in the two cells is a drop in contamination using Process B.

The CATLEGLABEL= option specifies the category legend label *Failure Causes:*. The NOHLABEL option suppresses the horizontal axis labels. The NOCURVE option suppresses the cumulative percent curve.

One-way Comparative Pareto Chart for DAY

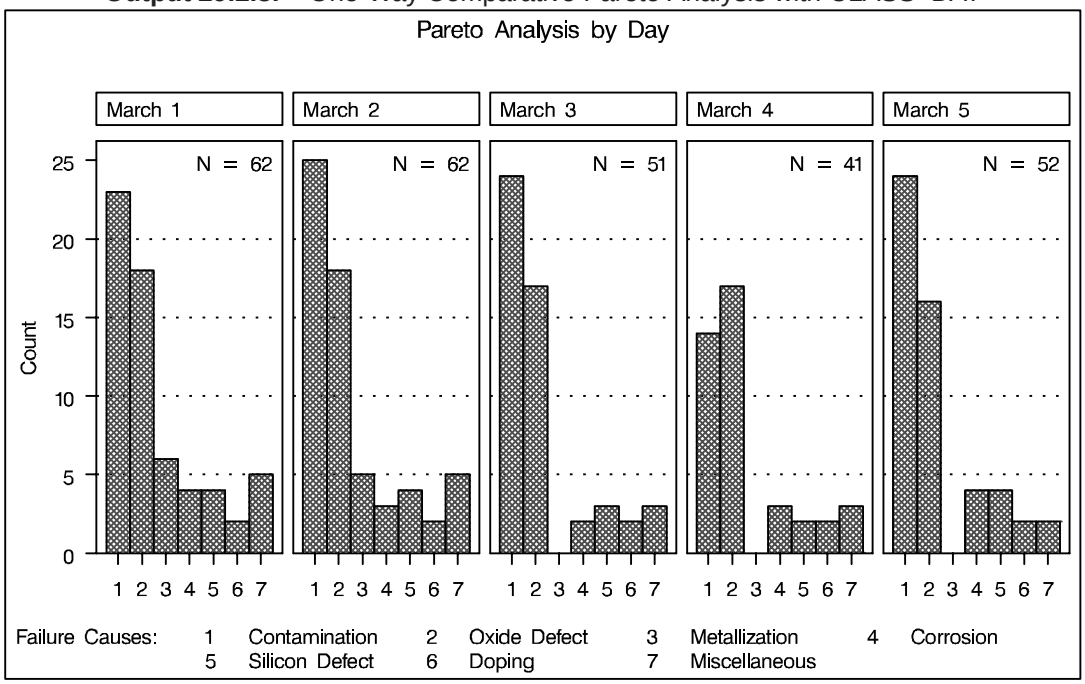
The following statements specify DAY as a classification variable:

```
title 'Pareto Analysis by Day';
proc pareto data=failure4;
   vbar cause / class       = day
                freq        = counts
                last        = 'Miscellaneous'
                scale       = count
                cbars       = green
                pbars       = m5x45
                catleglabel = 'Failure Causes:'
                intertile   = 1.0
                nrows       = 1
                ncols       = 5
                vref        = 5 10 15 20
```

```
                    lvref       = 34
                    nohlabel
                    nocurve
                    nlegend ;
run;
```

The NROWS= and NCOLS= options display the cells in a side-by-side arrangement. The VREF= and LVREF= options add reference lines. The chart is displayed in Output 29.2.3.

Output 29.2.3. One-Way Comparative Pareto Analysis with CLASS=DAY



By default, the key cell is the leftmost cell. There were no failures due to Metallization starting on March 3 (in fact, process controls to reduce this problem were introduced on this day).

Two-way Comparative Pareto Chart for PROCESS and DAY

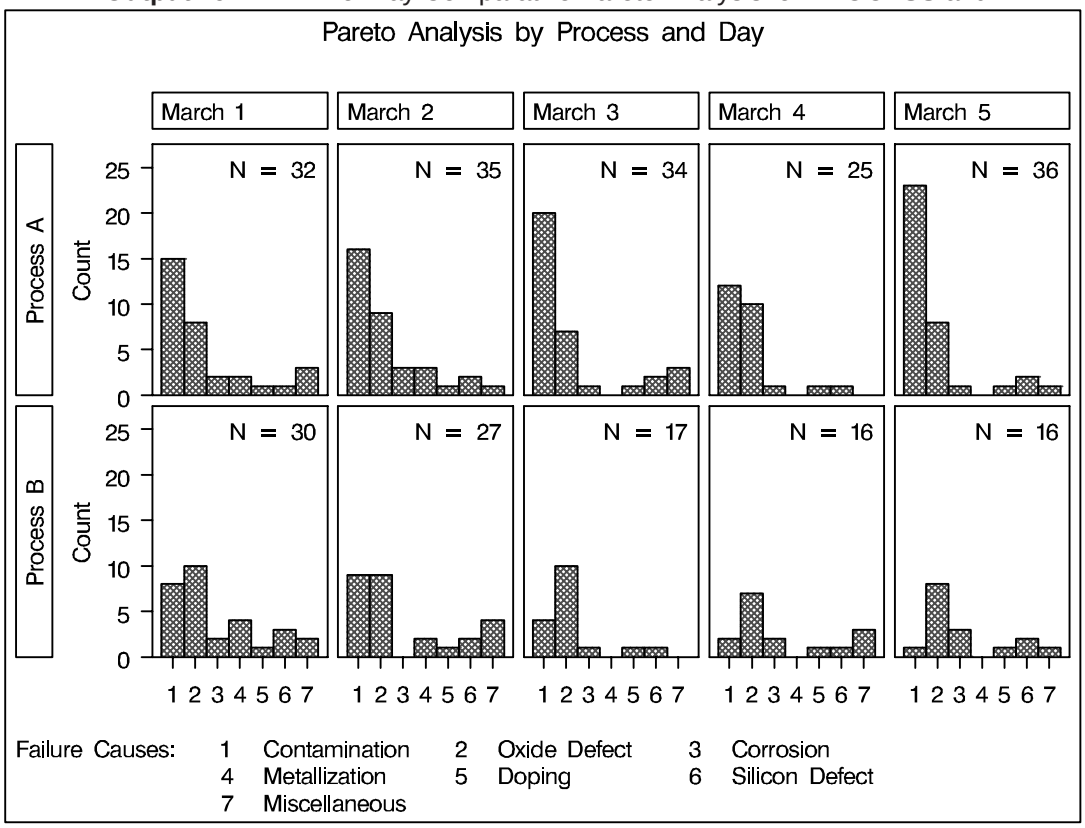
The following statements specify both PROCESS and DAY as CLASS= variables to create a two-way comparative Pareto chart:

```
title 'Pareto Analysis by Process and Day' ;
proc pareto data=failure4;
   vbar cause / class       = ( process day )
                freq        = counts
                nrows       = 2
                ncols       = 5
                cbars       = green
                pbars       = m5x45
                last        = 'Miscellaneous'
                scale       = count
                catleglabel = 'Failure Causes:'
```

```
                    intertile   = 1.0
                    nohlabel
                    nocurve
                    nlegend ;
run;
```

The chart is displayed in Output 29.2.4.

Output 29.2.4. Two-Way Comparative Pareto Analysis for PROCESS and DAY



The cells are arranged in a matrix whose rows correspond to levels of the first CLASS= variable (PROCESS) and whose columns correspond to levels of the second CLASS= variable (DAY). The dimensions of the matrix are specified with the NROWS= and NCOLS= options. The key cell is in the upper left corner.

The chart reveals continuous improvement with Process B.

Example 29.3. Highlighting the “Vital Few”

See PARETO10 in the SAS/QC Sample Library

This example is a continuation of Example 29.2.

In some applications you may want to use colors and patterns to highlight the bars corresponding to the most frequently occurring categories, which are referred to as the “vital few.”

The following statements highlight the two most frequently occurring categories in each cell of the comparative Pareto chart shown in Output 29.2.4:

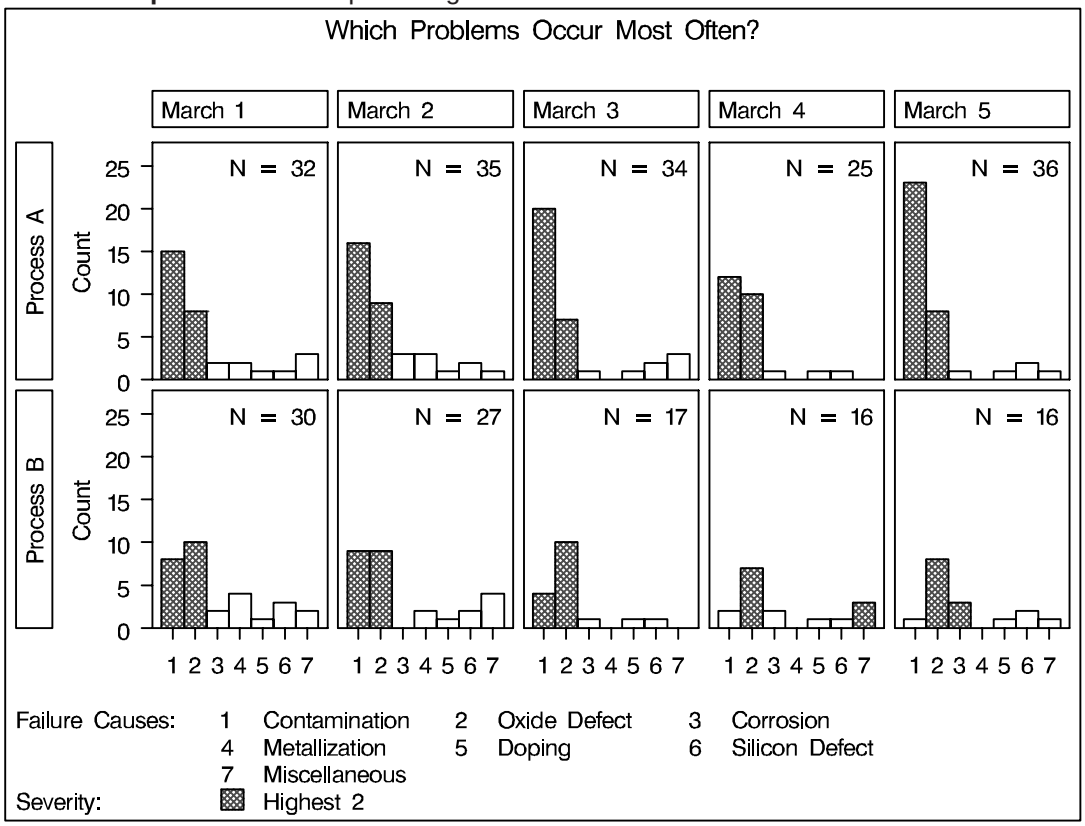
```
title 'Which Problems Occur Most Often?';
proc pareto data=failure4;
   vbar cause / class       = ( process day )
                freq        = counts
                nrows       = 2
                ncols       = 5
                last        = 'Miscellaneous'
                scale       = count
                chigh(2)    = green
                phigh(2)    = m5x45
                hlleglabel  = 'Severity:'
                catleglabel = 'Failure Causes:'
                intertile   = 1.0
                nohlabel
                nocurve
                nlegend ;
run;
```

Specifying CHIGH(2)=GREEN and PHIGH(2)=M5X45 causes the two highest bars in each cell to be filled in green with the pattern M5X45 (refer to *SAS/GRAPH Software: Reference* for a pattern selection guide). If you omit the PHIGH(2)= option, a solid green fill is used.

The new chart is displayed in Output 29.3.1. In all but two of the cells, the two vital problems are Contamination and Oxide Defect.

You can also highlight the “trivial many” categories (also referred to as the “useful many”) with the CLOW(m)= and PLOW(m)= options. You can use these options in conjunction with the CHIGH(n)=, PHIGH(n)=, CBARS=, and PBARS= options. For further details, see the entries for these options in the “Dictionary of Options” on page 813.

Output 29.3.1. Emphasizing the “Vital Few”



Example 29.4. Highlighting Combinations of Categories

See PARETO11 in the SAS/QC Sample Library

In some applications, it is useful to classify the categories into groups that are not necessarily related to frequency. This example, which is a continuation of Example 29.2, shows how you can display this classification with a bar legend.

Suppose that Contamination and Metallization are high priority problems, Oxide Defect is a medium priority problem, and all other categories are low priority problems. Begin by adding this information to the data set FAILURE4.

```
data failure4;
   length color $ 8 pattern $ 8 priority $ 16 ;
   set failure4;
   if cause='Contamination' or cause='Metallization' then do;
      color='red';    pattern='s';     priority='High';   end;
   else if cause='Oxide Defect' then do;
      color='yellow'; pattern='m5x45'; priority='Medium'; end;
   else do;
      color='white';  pattern='s';     priority='Low' ;   end;
run;
```

The variable PRIORITY indicates the priority, and the variables COLOR and PATTERN (character variables of length eight) provide colors and patterns corresponding to the levels of PRIORITY. The pattern values S and M5X45 correspond to a solid fill and a crosshatched fill, respectively.

The following statements specify PRIORITY as a BARLEGEND= variable, COLOR as a CBARS= variable, and PATTERN as a PBARS= variable:

```
title 'Which Problems Take Priority?';
proc pareto data=failure4;
   vbar cause / class       = ( process day )
                freq        = counts
                nrows       = 2
                ncols       = 5
                last        = 'Miscellaneous'
                scale       = count
                cbars       = ( color )
                pbars       = ( pattern )
                barlegend   = ( priority )
                barleglabel = 'Priority:'
                catleglabel = 'Failure Causes:'
                intertile   = 1.0
                nohlabel
                nocurve
                nlegend ;
run;
```

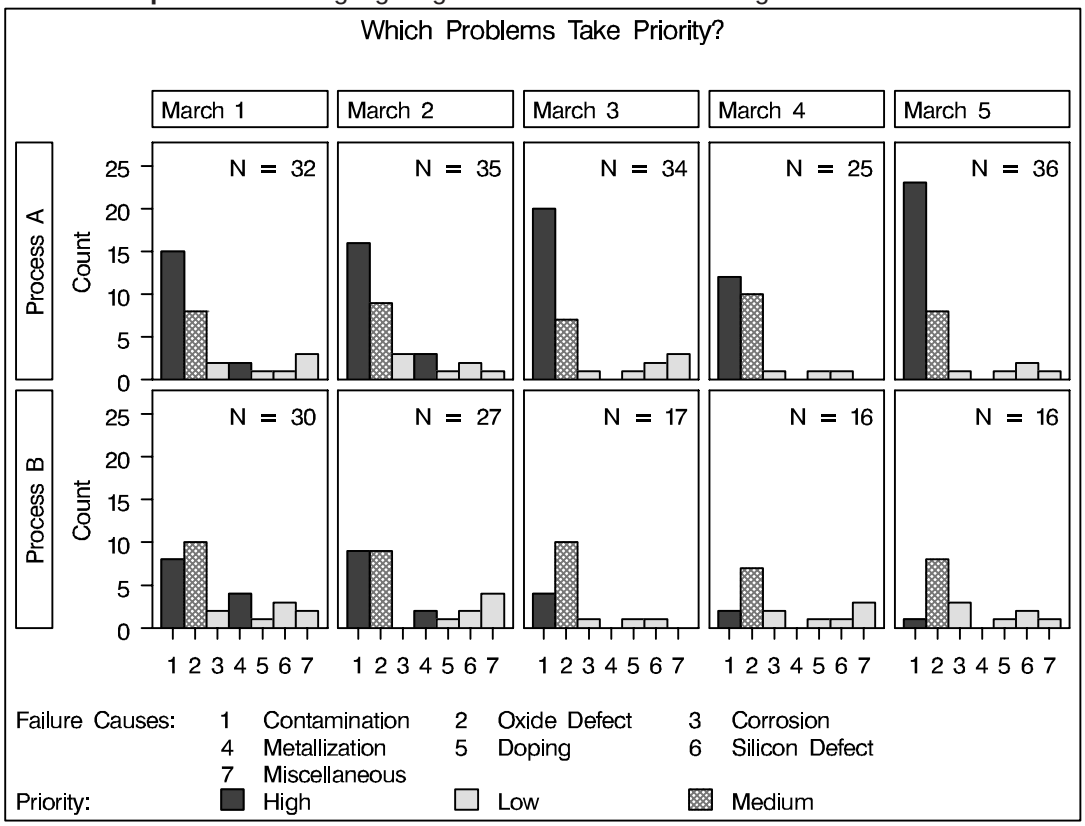
Note that the BARLEGEND=, CBARS=, and PBARS= variable names are enclosed in parentheses. (Parentheses are not used when you specify fixed colors and patterns with the CBARS= and PBARS= options, as in Example 29.2.)

The chart is displayed in Output 29.4.1. The levels of the BARLEGEND= variable are the values displayed in the legend labeled *Priority:* at the bottom of the chart.

In general, when you create CBARS=, PBARS=, and BARLEGEND= variables, their values must be consistent and unambiguous. You must assign distinct color and pattern values to the CBARS= and PBARS= variables for each level of the BARLEGEND= variable. It is not necessary to specify a PBARS= variable to accompany a BARLEGEND= variable, and if a PBARS= variable is omitted, the bars are filled with solid colors.

For further details, see the entries for the BARLEGEND=, CBARS=, and PBARS= options in "Dictionary of Options" on page 813.

Output 29.4.1. Highlighting Selected Subsets of Categories



Example 29.5. Highlighting Combinations of Cells

This example is a continuation of Example 29.4.

See PARETO1 in the SAS/QC Sample Library

In some applications involving comparative Pareto charts, it is useful to classify the cells into groups. This example shows how you can display this type of classification by coloring the tiles and adding a legend.

Suppose that you want to enhance Output 29.4.1 by highlighting the two cells for which PROCESS=`Process B` and DAY=`March 4` and `March 5` to emphasize the improvement displayed in those cells. Begin by adding a tile color variable (TILECOL) and a tile legend variable (TILELEG) to the data set FAILURE4.

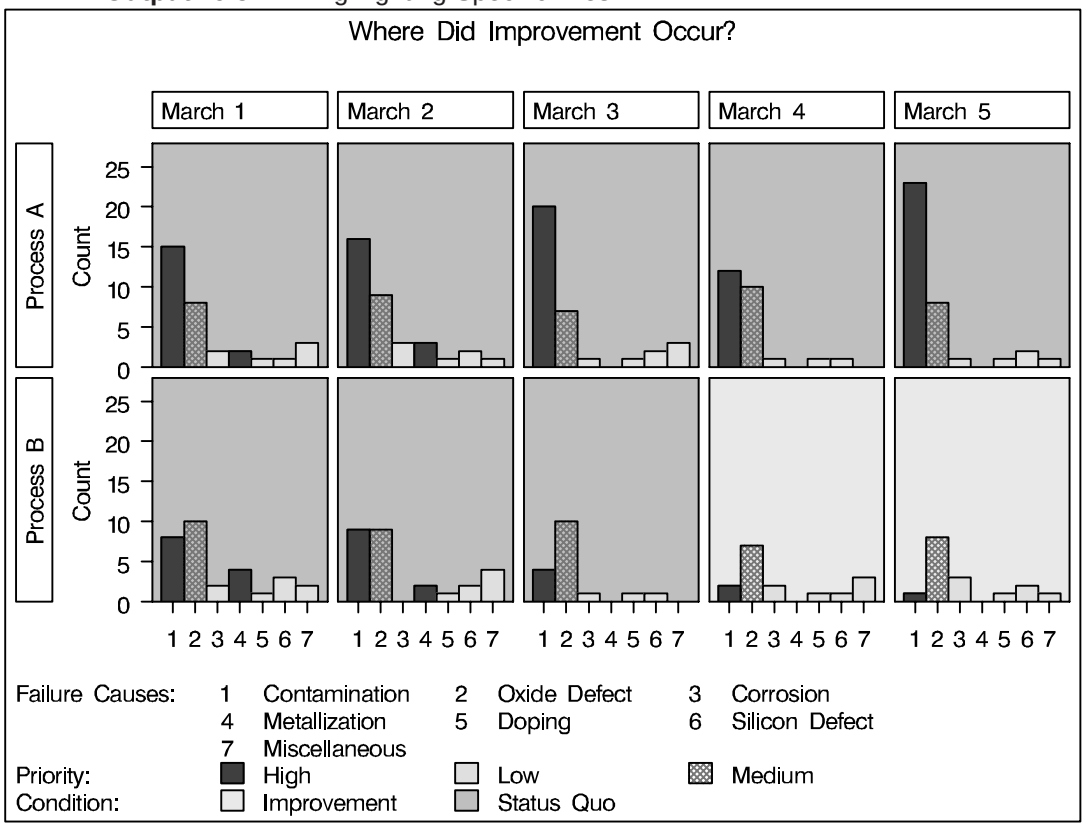
```
data failure4;
   length tilecol $ 8 tileleg $ 16 ;
   set failure4;
   if (process='Process B') and (day='March 4' or day='March 5')
   then do; tilecol='orange'; tileleg = 'Improvement'; end;
   else do; tilecol='empty' ; tileleg = 'Status Quo' ; end;
run;
```

The following statements specify TILECOL as a CTILES= variable and TILELEG as a TILELEGEND= variable. Note that the variable names are enclosed in parentheses.

```
title 'Where Did Improvement Occur?' ;
proc pareto data=failure4;
   vbar cause / class        = ( process day )
                freq         = counts
                nrows        = 2
                ncols        = 5
                last         = 'Miscellaneous'
                scale        = count
                catleglabel  = 'Failure Causes:'
                cbars        = ( color )
                pbars        = ( pattern )
                barlegend    = ( priority )
                barleglabel  = 'Priority:'
                ctiles       = ( tilecol )
                tilelegend   = ( tileleg )
                tileleglabel = 'Condition:'
                intertile    = 1.0
                nohlabel
                nocurve ;
run;
```

In the chart, shown in Output 29.5.1, the values displayed in the legend labeled *Condition:* are the levels of the TILELEGEND= variable.

Output 29.5.1. Highlighting Specific Tiles



Example 29.6. Ordering Rows and Columns in a Comparative Pareto Chart

See PARETO13 in the SAS/QC Sample Library

This example illustrates methods for controlling the order of rows and columns in a comparative Pareto chart.

The following statements create a data set named FAILURE7:

```
proc format;
   value procfmt   1   = 'Process A'
                   2   = 'Process B' ;
   value dayfmt    1   = 'Monday'
                   2   = 'Tuesday'
                   3   = 'Wednesday'
                   4   = 'Thursday'
                   5   = 'Friday'    ;

data failure7;
   length cause $16 ;
   format process procfmt. day dayfmt. ;
   label  cause   = 'Cause of Failure'
          process = 'Cleaning Method'
          day     = 'Day of Manufacture' ;
   input  process day cause $16. counts @@;
   datalines;
1   1   Contamination    15   1   1   Corrosion         2
1   1   Doping           1   1   1   Metallization     2
1   1   Miscellaneous    3   1   1   Oxide Defect      8
1   1   Silicon Defect   1   1   2   Contamination    16
1   2   Corrosion        3   1   2   Doping            1
1   2   Metallization    3   1   2   Miscellaneous     1
1   2   Oxide Defect     9   1   2   Silicon Defect    2
1   3   Contamination   20   1   3   Corrosion         1
1   3   Doping           1   1   3   Metallization     0
1   3   Miscellaneous    3   1   3   Oxide Defect      7
1   3   Silicon Defect   2   1   4   Contamination    12
1   4   Corrosion        1   1   4   Doping            1
1   4   Metallization    0   1   4   Miscellaneous     0
1   4   Oxide Defect    10   1   4   Silicon Defect    1
1   5   Contamination   23   1   5   Corrosion         1
1   5   Doping           1   1   5   Metallization     0
1   5   Miscellaneous    1   1   5   Oxide Defect      8
1   5   Silicon Defect   2   2   1   Contamination     8
2   1   Corrosion        2   2   1   Doping            1
2   1   Metallization    4   2   1   Miscellaneous     2
2   1   Oxide Defect    10   2   1   Silicon Defect    3
2   2   Contamination    9   2   2   Corrosion         0
2   2   Doping           1   2   2   Metallization     2
2   2   Miscellaneous    4   2   2   Oxide Defect      9
2   2   Silicon Defect   2   2   3   Contamination     4
2   3   Corrosion        1   2   3   Doping            1
2   3   Metallization    0   2   3   Miscellaneous     0
2   3   Oxide Defect    10   2   3   Silicon Defect    1
2   4   Contamination    2   2   4   Corrosion         2
```

```
2   4   Doping           1   2   4   Metallization    0
2   4   Miscellaneous    3   2   4   Oxide Defect     7
2   4   Silicon Defect   1   2   5   Contamination    1
2   5   Corrosion        3   2   5   Doping           1
2   5   Metallization    0   2   5   Miscellaneous    1
2   5   Oxide Defect     8   2   5   Silicon Defect   2
;
```

Note that FAILURE7 is similar to the data set FAILURE4 created in Example 29.2. Here, the classification variables PROCESS and DAY are numeric formatted variables, and the formatted values of DAY are `Monday` through `Friday`. In Example 29.2, PROCESS and DAY are character variables, and the values of DAY are `March 1` through `March 5`.

The following statements create a two-way comparative Pareto chart for CAUSE in which the rows represent levels of PROCESS and the columns represent levels of DAY:

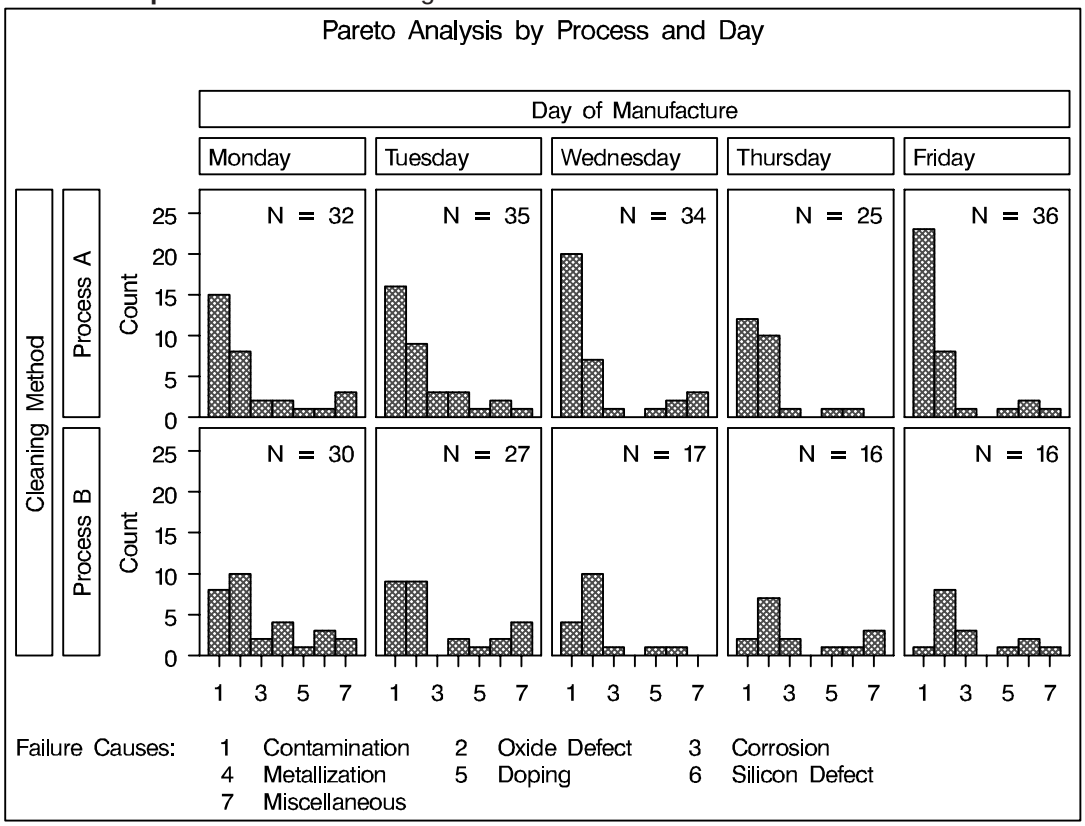
```
title 'Pareto Analysis by Process and Day' ;
proc pareto data=failure7;
   vbar cause / class       = ( process day )
                freq        = counts
                nrows       = 2
                ncols       = 5
                cbars       = green
                pbars       = m5x45
                last        = 'Miscellaneous'
                scale       = count
                catleglabel = 'Failure Causes:'
                intertile   = 1.0
                nohlabel
                nocurve
                nlegend ;
run;
```

The chart is shown in Output 29.6.1. The levels of the classification variables are determined by their formatted values. The default order in which the rows and columns are displayed is determined by the internal values of the classification variables, and, consequently, the columns appear in the order of the days of the week.

If DAY had been defined as a character variable with values `Monday` through `Friday`, the columns in Output 29.6.1 would have appeared in alphabetical order.

You can override the default order with the ORDER1= and ORDER2= options , which are described on pages 826–827.

Output 29.6.1. Controlling Row and Column Order



Example 29.7. Merging Columns in a Comparative Pareto Chart

See PARETO14 in the SAS/QC Sample Library

This example is a continuation of Example 29.4 and illustrates a method for merging the columns in a comparative Pareto chart.

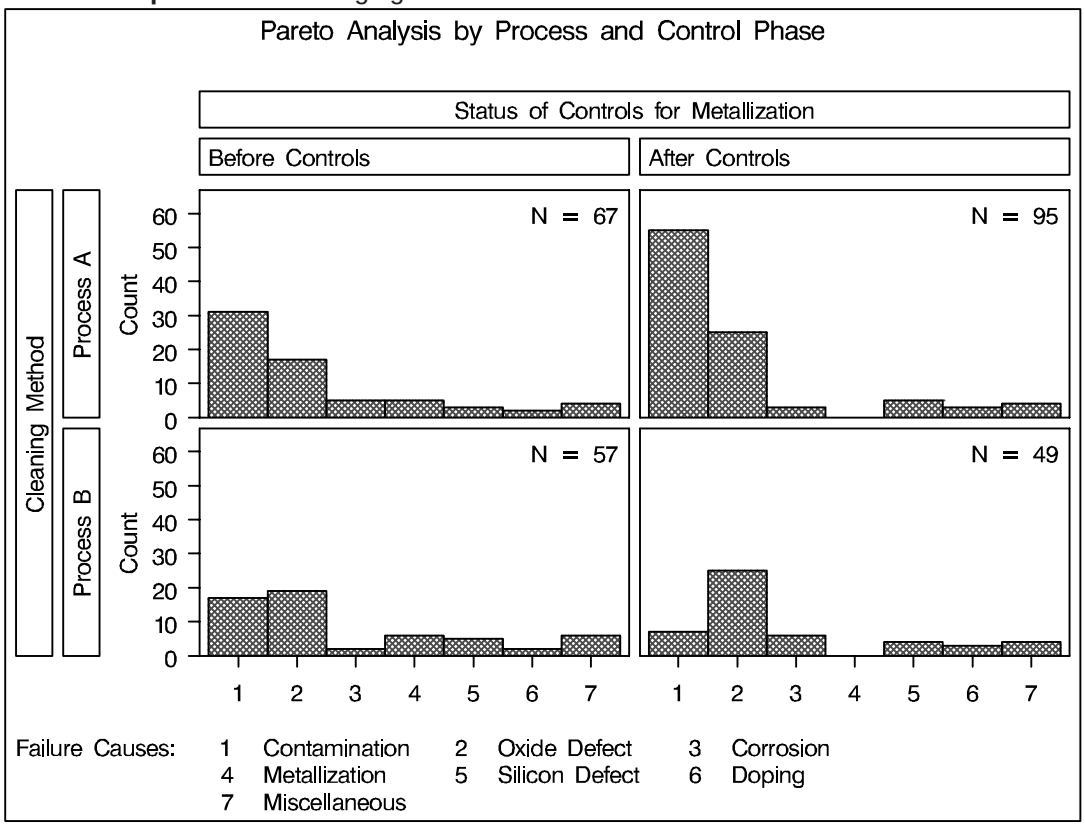
Suppose that controls for metallization were introduced on Wednesday. To show the effect of the controls, the columns for *Monday* and *Tuesday* are to be merged into a column labeled *Before Controls*, and the remaining columns are to be merged into a column labeled *After Controls*. The following statements introduce a format named CNTLFMT that merges the levels of DAY:

```
proc format;
   value cntlfmt   1-2 = 'Before Controls'
                   3-5 = 'After Controls' ;
run;
```

The following statements create the chart shown in Output 29.7.1:

```
title 'Pareto Analysis by Process and Control Phase' ;
proc pareto data=failure7;
   vbar cause / class       = ( process day )
                freq        = counts
                cbars       = green
                pbars       = m5x45
                last        = 'Miscellaneous'
                scale       = count
                catleglabel = 'Failure Causes:'
                intertile   = 1.0
                nohlabel
                nocurve
                nlegend ;
   format day cntlfmt. ;
   label  day = 'Status of Controls for Metallization';
run;
```

Output 29.7.1. Merging Classification Levels



The levels of DAY are determined by its formatted values, `Before Controls` and `After Controls`. By default, the order in which the columns are displayed is determined by the internal values. In this example, there are multiple distinct internal values for each level, and the procedure uses the internal value that occurs first in the input data set.

Example 29.8. Creating Weighted Pareto Charts

See PARETO12 in the SAS/QC Sample Library

In many applications, you can quantify the priority or severity of a problem with a measure such as the cost of repair or the loss to the customer expressed in man-hours. This example shows how to analyze such data with a weighted Pareto chart that incorporates the cost.

Suppose that the cost associated with each of the problems in data set FAILURE7 (see Example 29.6 on page 910) has been determined and that the costs have been converted to a relative scale. The following statements add the cost information to the data set:

```
data failure7;
   length analysis $ 16 ;
   label analysis = 'Basis for Analysis' ;
   set failure7;
   analysis = 'Cost' ;
      if      cause = 'Contamination'  then cost = 3.0 ;
      else if cause = 'Metallization'  then cost = 8.5 ;
      else if cause = 'Oxide Defect'   then cost = 9.5 ;
      else if cause = 'Corrosion'      then cost = 2.5 ;
      else if cause = 'Doping'         then cost = 3.6 ;
      else if cause = 'Silicon Defect' then cost = 3.4 ;
      else                                  cost = 1.0 ;
      output;
   analysis = 'Frequency' ;
      cost = 1.0 ;
      output;
run;
```

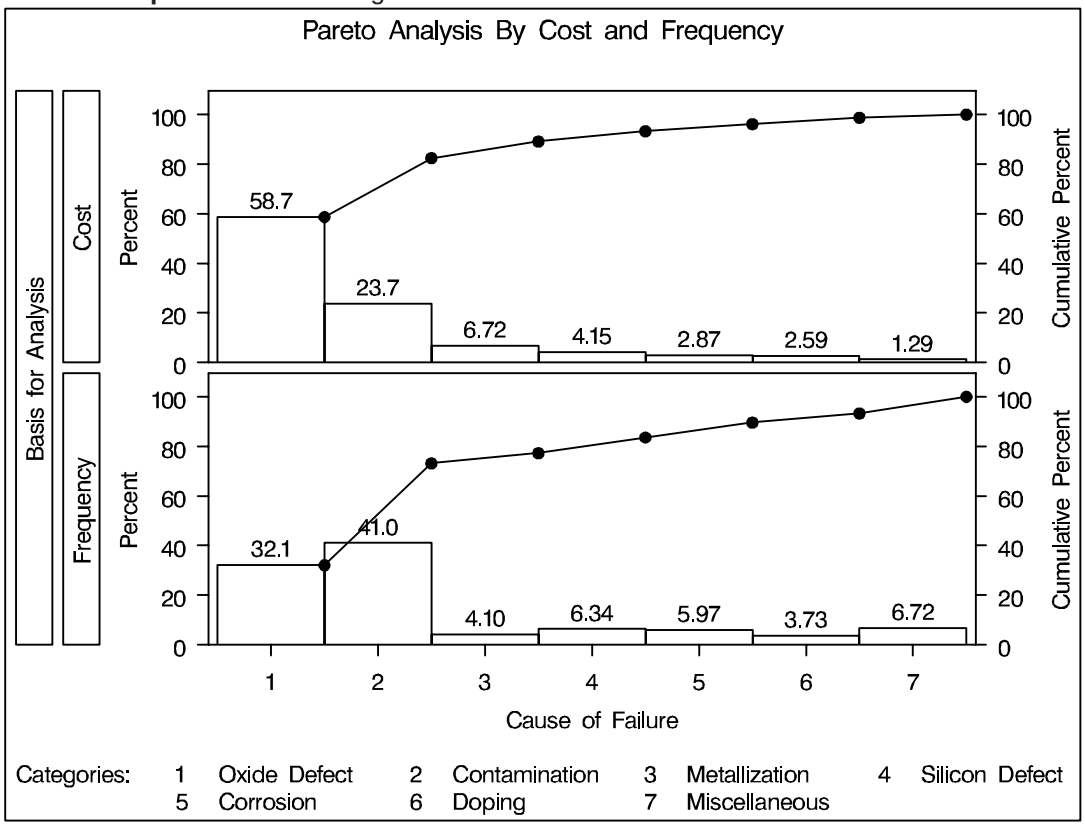
The classification variable ANALYSIS has two levels, `Cost` and `Frequency`. For ANALYSIS=`Cost`, the value of COST is the relative cost, and for ANALYSIS=`Frequency`, the value of COST is one.

The following statements create a one-way comparative Pareto chart with ANALYSIS as the classification variable, in which the cells are weighted Pareto charts with COST as the weight variable:

```
title 'Pareto Analysis By Cost and Frequency' ;
proc pareto data=failure7;
   vbar cause / class     = ( analysis )
                freq      = counts
                weight    = cost
                barlabel  = value
                out       = summary
                intertile = 1.0 ;
run;
```

The display is shown in Output 29.8.1.

Output 29.8.1. Taking Cost into Account



Within each cell, the height of a bar is the frequency of the category multiplied by the value of COST, expressed as a percent of the total across all categories. Thus, for the cell in which ANALYSIS is equal to `Frequency`, the bars simply indicate the frequencies expressed in percent units. This display shows that the most commonly occurring problem (Contamination) is not the most expensive problem (Oxide Defect). The output data set SUMMARY is listed in Output 29.8.2.

Output 29.8.2. The Output Data Set SUMMARY

Obs	analysis	cause	cost	_COUNT_	_WCOUNT_	_PCT_	_CMPCT_
1	Cost	Oxide Defect	9.5	172	1634.0	58.6799	58.680
2	Cost	Contamination	3.0	220	660.0	23.7018	82.382
3	Cost	Metallization	8.5	22	187.0	6.7155	89.097
4	Cost	Silicon Defect	3.4	34	115.6	4.1514	93.249
5	Cost	Corrosion	2.5	32	80.0	2.8729	96.122
6	Cost	Doping	3.6	20	72.0	2.5856	98.707
7	Cost	Miscellaneous	1.0	36	36.0	1.2928	100.000
8	Frequency	Oxide Defect	1.0	172	172.0	32.0896	32.090
9	Frequency	Contamination	1.0	220	220.0	41.0448	73.134
10	Frequency	Metallization	1.0	22	22.0	4.1045	77.239
11	Frequency	Silicon Defect	1.0	34	34.0	6.3433	83.582
12	Frequency	Corrosion	1.0	32	32.0	5.9701	89.552
13	Frequency	Doping	1.0	20	20.0	3.7313	93.284
14	Frequency	Miscellaneous	1.0	36	36.0	6.7164	100.000

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/QC® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999. 1994 pp.

SAS/QC® User's Guide, Version 8

ISBN 1–58025–493–4

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999