A Comprehensive Guide to Creating and Interpreting Stem-and-Leaf Plots Using Stata

Authored by

Mohammed looti

November 8, 2025

Understanding the Stem-and-Leaf Plot

The Stem-and-Leaf Plot is an exceptionally powerful visualization technique foundational to Exploratory Data Analysis (EDA). Conceived by the eminent statistician John Tukey in the 1970s, this display offers a unique blend of visual data distribution and the preservation of all original, raw data values. Unlike the conventional histogram, which aggregates observations into bins and subsequently obscures the individual data points, the Stem-and-Leaf Plot maintains full fidelity to the source data. This characteristic makes it an invaluable diagnostic tool, particularly when conducting initial scrutiny of small to medium-sized datasets where every observation matters.

The mechanism of the plot relies on partitioning each numerical observation into two distinct parts: the **stem** and the **leaf**. As defined in its construction, the stem typically comprises the leading digits of the value, acting as a categorical bin, while the leaf represents the trailing, or last, significant digit. For clarity, a data point such as 25 would be separated into a stem of 2 and a leaf of 5. This systematic decomposition allows researchers to instantly grasp the concentration and spread of data, providing immediate insight into frequency and density across different ranges.

The visual representation organizes data efficiently. The stems are listed vertically in ascending order, usually on the left, and the leaves corresponding to that stem are listed horizontally to the right, also in ascending order. This structure immediately reveals the overall shape of the distribution, including its central tendency, potential skewness, and the presence of any significant gaps or outliers. The plot serves as a foundational step toward more complex statistical modeling by offering a robust, non-parametric view of the data's characteristics.

14, 14, 15, 16, 17, 28, 29, 30, 34, 35

GENERATE STEM AND LEAF PLOT

Stem	Leaf				
1	4	4	5	6	7
2	8	9			
3	0	4	5		

To efficiently generate such detailed diagnostic visualizations for empirical analysis, specialized statistical software is essential. The remainder of this guide will detail the precise, step-by-step procedures for constructing and customizing the Stem-and-Leaf Plot utilizing **Stata**, one of the most widely respected and utilized statistical packages across academic and professional research disciplines.

Setting Up the Analysis Environment in Stata

Generating a Stem-and-Leaf Plot within the **Stata** environment is accomplished using a simple, dedicated built-in command. However, before any statistical command can be executed, the prerequisite step is to ensure that the necessary data is correctly loaded into the program's active memory. For the purpose of this instructional walkthrough, we will leverage one of Stata's classic, readily available pre-installed datasets, thereby ensuring that these steps are completely reproducible by any user with access to the software.

The specific dataset we have chosen for demonstration is the *auto* dataset, a standard collection comprising various characteristics of automobiles, including critical metrics such as mileage per

gallon (mpg) and vehicle price. This particular dataset is ideal for our needs because it contains both small and large numerical variables, allowing us to thoroughly explore the challenges and solutions associated with visualizing different scales of quantitative data using the stem-and-leaf technique.

To initiate the session, the user must input the data loading command directly into the **Command window** and execute it by pressing the Enter key. This command is designed to fetch the dataset directly from the Stata Press website, preparing the workspace for the subsequent visualization steps.

Step 1: Load the data.

Load the *auto* dataset by typing the following command:

use http://www.stata-press.com/data/r13/auto

Once this command is successfully executed, **Stata** provides confirmation that the data has been loaded, signifying that the environment is fully prepared. The next crucial step involves applying the core **stem** command to examine the distribution of a key variable, starting with the straightforward *mpg* variable.

Generating the Basic Stem-and-Leaf Plot for MPG

With the *auto* dataset now active in memory, we can proceed directly to generate the foundational Stem-and-Leaf Plot for the fuel efficiency variable, *mpg* (miles per gallon). Since *mpg* values typically consist of two digits, the resulting plot under Stata's default parameters is generally simple and highly interpretable. The fundamental command required for this visualization is straightforward: simply the keyword `stem` followed by the name of the quantitative variable slated for analysis.

Step 2: Create a stem-and-leaf plot for the variable *mpg*.

Type the following into the Command window and click *Enter*:

stem mpg

Following execution, **Stata** immediately produces an ASCII-based output displayed in the Results window. This output provides a comprehensive initial overview of the *mpg* distribution, including essential summary statistics such as the total number of observations (N) and the plot's numerical representation. This initial visualization is instrumental for the rapid assessment of key distributional features, including symmetry, the modality (number of peaks), and the immediate identification of potential outliers.

A key characteristic of Stata's default output is its tendency to split the stems. For instance, a single stem value (e.g., 20s) might be represented across two or even more lines. This built-in splitting technique is designed to stretch the distribution vertically, effectively acting like using narrower bins in a histogram, thereby providing a more granular and detailed view of the data's inherent shape and density. This default approach helps prevent the distribution from becoming too condensed, ensuring fine details are not lost during the initial Exploratory Data Analysis.

```
. stem mpg

Stem-and-leaf plot for mpg (Mileage (mpg))

  1t | 22
  1f | 44444455
  1s | 66667777
  1. | 88888888899999999
  2* | 00011111
  2t | 22222333
  2f | 444455555
  2s | 666
  2. | 8889
  3* | 001
  3t |
  3f | 455
  3s |
  3. |
  4* | 1
```

Customizing Visualization: The lines() Option

While the automatic stem-splitting behavior in **Stata** is often beneficial for detailed diagnostic inspection, there are circumstances, such as presenting results or analyzing already compact data, where a more concise plot is highly desirable. When stems are split, each resulting stem line typically contains data points corresponding to a specific range of leaf digits (e.g., leaves 0 through 4 on one line, and leaves 5 through 9 on the subsequent line).

To override this automatic splitting and significantly condense the visualization, **Stata** offers the powerful `lines()` option. By specifying the syntax `lines(1)`, we explicitly instruct the program to aggregate all leaves corresponding to a single stem onto one line, irrespective of the leaf digit range. This parameter dramatically shortens the plot, enabling a quicker visual assessment of the entire range of values and their density distribution.

To implement this crucial customization for the *mpg* variable, the option is simply appended to the original command:

stem mpg, lines(1)

The resulting output is visibly shorter and far more compact than the default plot. Every individual stem now successfully aggregates all its associated leaves onto a single row. This feature is particularly effective for summarizing the data's distribution without unnecessary vertical elongation, facilitating a much faster comparison of data density across different stem values. The transformation below clearly demonstrates how the plot occupies significantly less screen space while retaining all original data points.

```
. stem mpg, lines(1)

Stem-and-leaf plot for mpg (Mileage (mpg))

  1* | 22444444556666777788888888899999999
  2* | 00011111222223334444555556668889
  3* | 001455
  4* | 1
```

Handling Large Values and Data Rounding with round()

The utility of the **stem** command extends far beyond small, two-digit numbers like *mpg*. It is equally applicable to variables that encompass much larger numerical scales, such as the *price* variable in the *auto* dataset, where values routinely reach into the thousands. When dealing with large numbers, the default Stem-and-Leaf Plot often becomes cumbersome and impractical; stems may stretch excessively long, or the leaves might end up representing insignificant trailing digits, obscuring the meaningful distribution shape.

Examining the default plot for the *price* variable illustrates this challenge:

stem price

```
Stem-and-leaf plot for price (Price)

  3***  | 291,299
  3***  | 667,748,798,799,829,895,955,984,995
  4***  | 010,060,082,099,172,181,187,195,296,389,424,425,453,482,499
  4***  | 504,516,589,647,697,723,733,749,816,890,934
  5***  | 079,104,172,189,222,379,397
  5***  | 705,719,788,798,799,886,899
  6***  | 165,229,295,303,342,486
  6***  | 850
  7***  | 140
  7***  | 827
  8***  | 129
  8***  | 814
  9***  |
  9***  | 690,735
 10***  | 371,372
 10***  |
 11***  | 385,497
 11***  | 995
 12***  |
 12***  | 990
 13***  | 466
 13***  | 594
 14***  |
 14***  | 500
 15***  |
 15***  | 906
```

To manage these large numerical values and ensure the creation of an easily interpretable visualization, **Stata** provides the highly effective `round()` option. This option grants the user precise control by allowing them to specify a rounding increment that is applied to the data before it is plotted. This process fundamentally dictates which digits will constitute the stem and which will form the leaf, which is a critical function for achieving visual clarity when analyzing data that spans multiple orders of magnitude.

For example, employing `round(100)` instructs Stata to round every price value to the nearest hundreds place. In this configuration, the stem effectively represents the hundreds and thousands digits, while the leaf represents the next significant digit (the tens place, after rounding). This strategic rounding simplifies the plot by filtering out minor fluctuations and focusing the visualization strictly on major price changes.

stem price, round(100)

```
. stem price, round(100)

Stem-and-leaf plot for price (Price)

price rounded to nearest multiple of 100
plot in units of 100

   3* | 33778889
   4* | 000011122223444555556677778 99
   5* | 11222447788899
   6* | 2233359
   7* | 18
   8* | 18
   9* | 77
  10* | 44
  11* | 45
  12* | 0
  13* | 056
  14* | 5
  15* | 9
```

Alternatively, if a slightly finer level of detail is required, one can specify rounding to the tens place using `round(10)`. This choice of granularity retains more specific information within the plot structure while still significantly cleaning up the appearance compared to the dense, unrounded default. The selection of the optimal rounding factor should always be guided by the level of precision deemed necessary for the specific goals of the exploratory analysis.

stem price, round(10)

```
. stem price, round(10)

Stem-and-leaf plot for price (Price)

price rounded to nearest multiple of 10
plot in units of 10

   3** | 29,30
   3** | 67,75,80,80,83,90,96,98
   4** | 00,01,06,08,10,17,18,19,20,30,39,42,43,45,48
   4** | 50,50,52,59,65,70,72,73,75,82,89,93
   5** | 08,10,17,19,22,38,40
   5** | 71,72,79,80,80,89,90
   6** | 17,23,30,30,34,49
   6** | 85
   7** | 14
   7** | 83
   8** | 13
   8** | 81
   9** |
   9** | 69,74
  10** | 37,37
  10** |
  11** | 39
  11** | 50
  12** | 00
  12** | 99
  13** | 47
  13** | 59
  14** |
  14** | 50
  15** |
  15** | 91
```

Refining the Display: Using the prune Option

When constructing a Stem-and-Leaf Plot, particularly after applying rounding or when the underlying data is inherently sparse across certain ranges, it is common to encounter stems that lack any corresponding leaves. These empty stems, while included by default to maintain the integrity of the scale and visually illustrate gaps in the data distribution, can unnecessarily lengthen the plot. This is especially true when the overall data range is wide, but the observations are heavily concentrated in just a few specific areas.

To resolve this issue and yield a more streamlined, concise visualization, **Stata** offers the `prune` option. This command instructs the program to systematically omit any stem line that does not contain a single leaf, effectively removing stem lines that represent value ranges where no

observations exist. Utilizing `prune` is an excellent strategy to focus visual attention exclusively on the ranges where the data is actually clustered, enhancing the clarity of the distribution.

For maximum impact regarding clarity and compactness, the `prune` option can be seamlessly combined with previously discussed options, such as `round()`. Let us apply this combination to the *price* variable, rounding to the tens place while simultaneously eliminating empty stems:

stem price, round(10) prune

The result is a highly efficient and targeted plot. By eliminating visually noisy blank lines, the plot immediately highlights the dense regions of the price distribution, making the identification of the mode and the overall shape of the data quicker and significantly more intuitive for Exploratory Data Analysis purposes.

```
. stem price, round(10) prune

Stem-and-leaf plot for price (Price)

price rounded to nearest multiple of 10
plot in units of 10

   3** | 29,30
   3** | 67,75,80,80,83,90,96,98
   4** | 00,01,06,08,10,17,18,19,20,30,39,42,43,45,48
   4** | 50,50,52,59,65,70,72,73,75,82,89,93
   5** | 08,10,17,19,22,38,40
   5** | 71,72,79,80,80,89,90
   6** | 17,23,30,30,34,49
   6** | 85
   7** | 14
   7** | 83
   8** | 13
   8** | 81
   9** | 69,74
  10** | 37,37
  11** | 39
  11** | 50
  12** | 00
  12** | 99
  13** | 47
  13** | 59
  14** | 50
  15** | 91
```

Summary of Stem-and-Leaf Plot Utility

The Stem-and-Leaf Plot remains a powerful and essential tool in the statistical analyst's arsenal, particularly when implemented efficiently using professional software like **Stata**. The enduring value of this visualization method lies in its dual functionality: it simultaneously provides a clear graphical representation of the data distribution while critically preserving the integrity of every raw data point. By mastering the core **stem** command and its essential options--including `lines()` for vertical compression, `round()` for effectively managing large scales, and `prune` for cleaning up sparse displays--analysts gain the ability to rapidly generate customized plots perfectly tailored to the specific characteristics of their dataset.

These plots serve as an indispensable first step in virtually any research project, offering immediate, actionable clues regarding data symmetry, overall spread, and the presence of potential outliers that necessitate further scrutiny. A thorough understanding of how to manipulate these options ensures that the visual output is not only statistically accurate but also maximally informative and readable, thereby greatly streamlining and enhancing the initial phase of data interpretation and research design.

Additional Resources for Data Visualization

For readers interested in expanding their knowledge of data visualization and Exploratory Data Analysis techniques within the Stata environment, we strongly recommend consulting the official Stata documentation. Specifically, documentation for related graphical commands such as `histogram` and `dotplot` offers further avenues for exploring quantitative data distributions.