

Creating and Customizing Box Plots for Data Analysis in Stata

Authored by
Mohammed loot

November 8, 2025

RECOMMENDED CITATION

Mohammed loot (2025). *Creating and Customizing Box Plots for Data Analysis in Stata*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=13690>

A **box plot**, commonly known as a box-and-whisker plot, is an indispensable visualization tool in **Stata** for conducting robust **exploratory data analysis** (EDA). This graphical representation offers a standardized and highly concise method for illustrating the distribution of a continuous dataset based fundamentally on its **five number summary**. By quickly analyzing the plot, researchers can immediately assess crucial statistical characteristics, including the central tendency, data variability (spread), symmetry, and the presence of potential statistical outliers across various samples or groups. Due to their compact nature, box plots are exceptionally effective when the goal is to visually compare the distributions of several variables or groups simultaneously, providing powerful visual evidence that complements traditional numerical summary statistics.

The **five number summary** forms the structural foundation of the box plot. These five specific points define the range and internal spread of the data, offering critical insights into the shape and characteristics of the underlying distribution:

The **Minimum**: The lowest observed value in the dataset, excluding any identified statistical **outliers**.

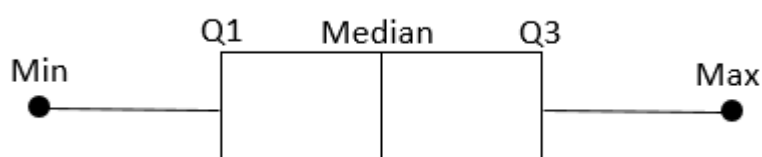
The **First Quartile (Q1)**: Representing the 25th percentile, meaning that one-quarter (25%) of the data observations fall below this specific point.

The **Median (Q2)**: Also known as the 50th percentile, this is the definitive midpoint of the data distribution, defining the central tendency.

The **Third Quartile (Q3)**: The 75th percentile, indicating that 75% of the observations have values less than this threshold.

The **Maximum**: The highest observed value in the dataset, again excluding any potential statistical outliers.

Visually, the central box encompasses the **interquartile range (IQR)**--the distance between Q1 and Q3--which captures the middle 50% of the observations. The "whiskers" extend from the box to the non-outlier minimum and maximum values, while any data points falling outside these whiskers are individually marked as potential outliers requiring further scrutiny. This comprehensive tutorial walks through the steps required to efficiently generate and aesthetically customize these essential visualizations within the Stata statistical environment.



Preparing the Stata Environment and Loading Sample Data

Prior to generating any graphical output, the first crucial step is to load an appropriate dataset into the active memory of [Stata](#). For instructional and demonstration purposes, we will rely on the proprietary, built-in *auto* dataset. This dataset is standard across Stata documentation and serves as an excellent reference point, containing a rich variety of vehicle characteristics. It is particularly well-suited for illustrating basic plotting commands and demonstrating how to handle both continuous variables (such as mileage per gallon, or *mpg*) and categorical variables (such as car origin, or *foreign*).

To ensure the reproducibility of the following steps, we must access and load this specific dataset consistently. The command below instructs Stata to retrieve the sample data directly from the official Stata Press online repository and load it into the active session, effectively replacing any data that might have been previously loaded. This guaranteed process ensures that every user following this guide operates with the exact, identical data structure, leading to consistent results.

To load the *auto* dataset and begin the analysis, execute the following command sequence in the Stata Command window:

use <http://www.stata-press.com/data/r13/auto>, clear

Once the data is successfully loaded, we can proceed directly to generating our visualizations. Our initial focus will be on the continuous variable *mpg* (miles per gallon), as its distribution is frequently subject to scrutiny in performance and efficiency analyses, making it a perfect candidate for box plot visualization.

Generating Essential Univariate Box Plots

Stata provides clear and efficient commands specifically designed for generating univariate box plots--visualizations that capture the distribution of a single continuous variable. Analysts must often choose between a vertical or horizontal orientation, a decision typically guided by presentation requirements, such as the total number of categories involved or the necessary length of variable labels to maintain readability.

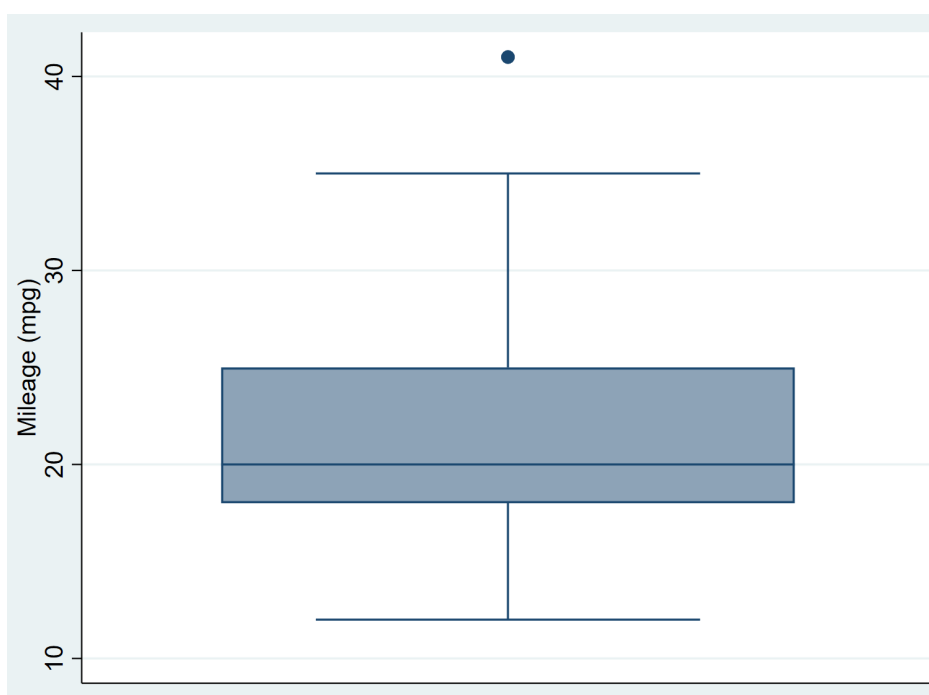
Vertical Box Plots: The Standard Orientation

The vertical orientation is the default setting and the format most frequently encountered in academic and statistical literature. In this arrangement, the measurement scale of the variable is positioned along the Y-axis. In Stata, the core command for producing this visualization is `graph box`, immediately followed by the name of the variable under investigation. To visualize the distribution of *mpg*, the command is straightforward and highly concise.

The resulting vertical graph offers an immediate, structured summary of the data: the central line inside the box precisely marks the [median](#) (Q2); the upper and lower edges of the box define the first (Q1) and third (Q3) [quartiles](#); and the extending whiskers delineate the non-outlier range. Any individual points plotted beyond these whiskers are flagged as potential outliers that may necessitate further detailed investigation.

To generate the standard vertical [box plot](#) for the variable *mpg*, utilize the following command syntax:

graph box mpg



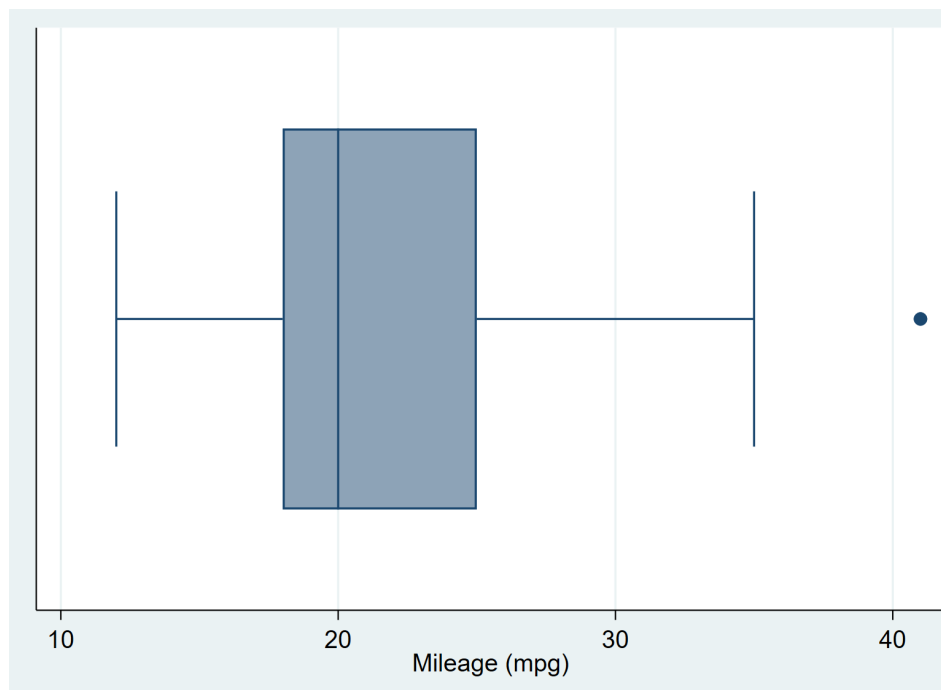
Horizontal Box Plots: Enhancing Readability

Although the vertical format is conventional, switching to a horizontal orientation can often significantly enhance graph readability, particularly when comparing numerous distributions side-by-side or when the categorical labels associated with the groups are particularly extensive. In the horizontal configuration, the variable scale is plotted along the X-axis.

Stata facilitates this alternative view through the dedicated command, `graph hbox`. This command operates identically to `graph box`, but crucially inverts the orientation of the final graph. When interpreting a horizontal plot, analysts must remember that the length of the box and the span of the whiskers still quantify the data's spread and variability, but this measure is now read along the horizontal axis. This presentation style can be particularly impactful in presentations, publications, or dashboards where a horizontal layout is preferred for aesthetic clarity or space management.

To construct a horizontal **box plot** for the *mpg* variable, execute the command below:

graph hbox mpg



Comparative Visualization: Grouping Box Plots by Category

A central strength of the box plot is its capacity to facilitate the comparison of a continuous variable's distribution across several distinct, predefined categories. This powerful technique permits analysts to visually compare key distributional metrics--such as medians, the spread represented by the interquartile range (IQR), and overall skewness--between groups. Such comparisons provide an immediate, clear understanding of how the categorical factor influences the measured continuous outcome.

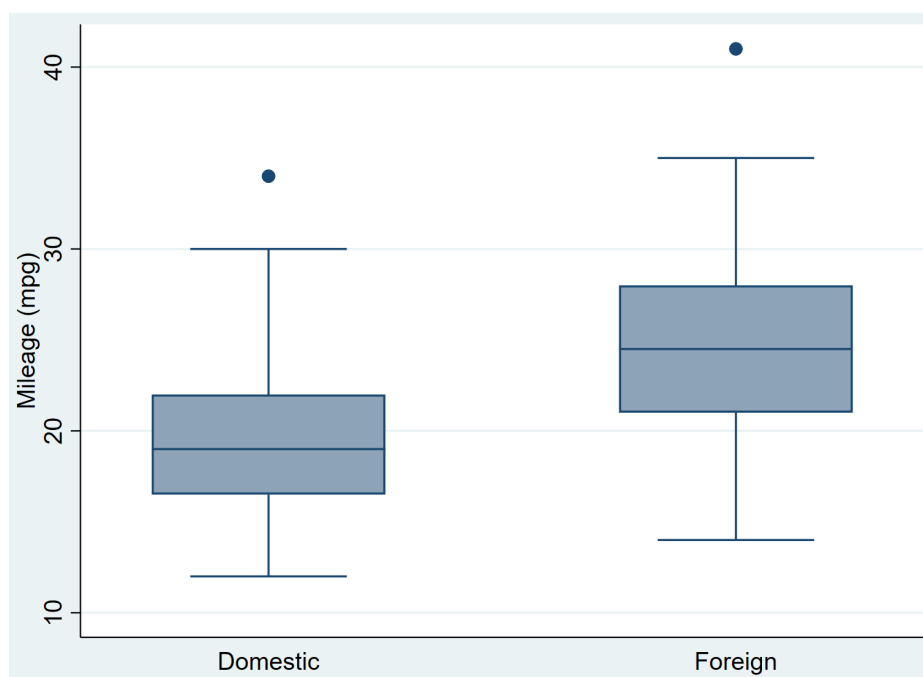
Within Stata, this critical grouping function is achieved efficiently using the `over()` option appended to the main `graph box` command. The variable designated inside the parentheses of the `over()` option must be a categorical (or nominal) variable. Stata then automatically renders a separate box plot for every unique value (or level) of the grouping variable, arranging them side-by-side to maximize comparative ease.

As an illustrative example using the *auto* dataset, we can investigate how the distribution of *mpg* varies depending on the car's origin, utilizing the binary categorical variable *foreign* (indicating domestic or imported status). This direct comparison is often essential for understanding inherent performance variations or differences in manufacturing philosophy between the two groups.

To generate comparative box plots for *mpg*, stratified by the categorical variable *foreign*, use the following syntax:

graph box mpg, over(foreign)

The resulting multi-panel graph instantly showcases the two distinct distributions, allowing for immediate visual assessment regarding which category displays a higher central tendency (median) or exhibits greater overall data variability (a wider IQR). This form of visual evidence often serves as a necessary preliminary step before conducting formal inferential statistical tests, such as [Analysis of Variance \(ANOVA\)](#) or various non-parametric comparisons.



Advanced Comparison: Displaying Multiple Continuous Variables Across Groups

In complex analytical scenarios, researchers frequently require the ability to compare the distributions of *several* continuous variables simultaneously, all of which are stratified by a single categorical grouping factor. [Stata](#) is designed to facilitate this complex, multivariate comparison by allowing analysts to list multiple continuous variables before the comma in the `graph box` command, while preserving the `over()` option for the critical categorical grouping.

This method generates a matrix-like structure of box plots, where the various continuous variables are clustered together and then segmented by the discrete levels of the categorical variable. This visualization is invaluable for multivariate analysis where the objective is to assess the consistent or divergent impact of a single grouping factor (e.g., treatment protocol, car type, or demographic

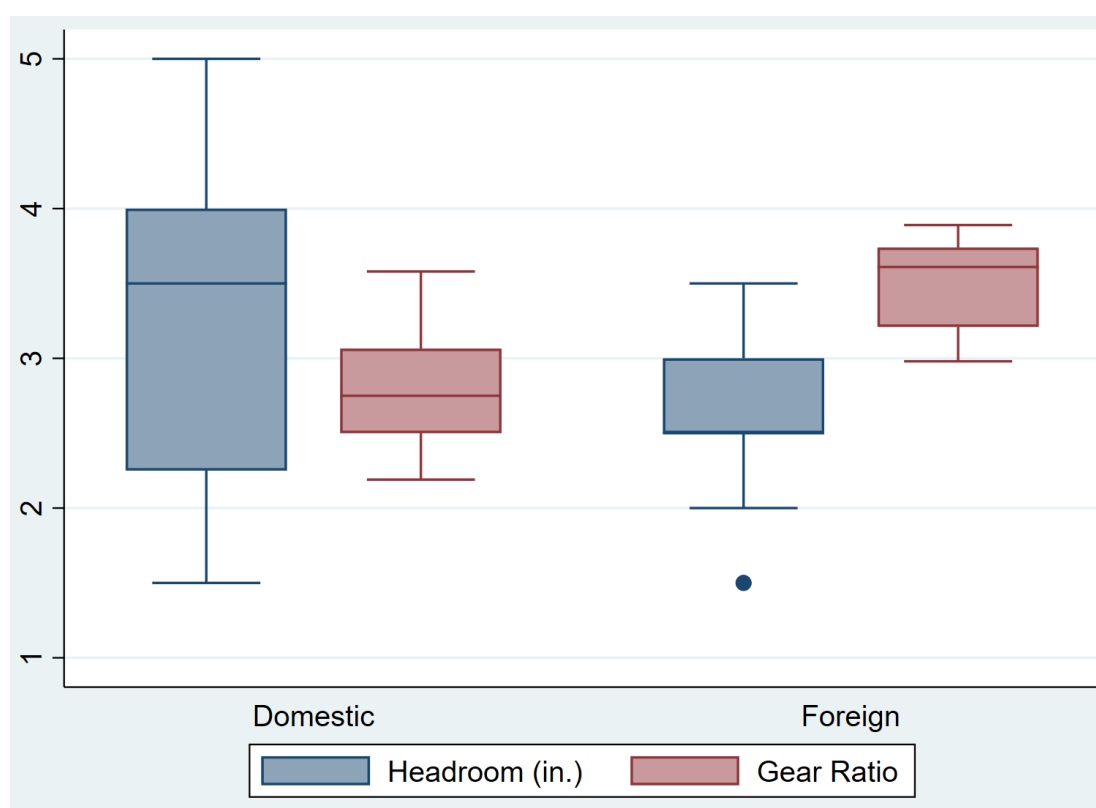
class) across several related outcome measures (e.g., price, weight, or fuel efficiency).

For instance, the following command generates box plots for both *headroom* and *gear_ratio*, with both variables being stratified by the categorical variable *foreign*. This configuration enables a direct, side-by-side comparison of how both interior space measurements and fundamental mechanical gearing characteristics differ systematically between domestic and imported automobiles in the sample.

To generate multiple box plots for more than one continuous variable based on the categorical variable *foreign*, use the following comprehensive syntax:

graph box headroom gear_ratio, over(foreign)

Effective interpretation of this multi-panel plot necessitates careful attention to the variable labels and axis scales. Analysts can quickly discern patterns, such as whether domestic cars exhibit, for example, a much wider spread in headroom measurements but possess a more narrowly controlled distribution of gear ratios compared to imported vehicles. This highly detailed, multi-panel approach maximizes the density of information conveyed within a single graphic, significantly streamlining the presentation and discussion of multivariate descriptive statistics.



Customizing Appearance and Enhancing Box Plot Annotation

While the default Stata graphics are inherently functional, refining their aesthetic appearance and incorporating clear textual annotations are essential steps for producing publication-quality figures ready for formal reports. Stata's command syntax for graphing is exceptionally flexible, granting users precise control over a wide array of aesthetic elements, including the main title, supporting subtitles, source notes, and color palettes. These systematic modifications are vital to ensure the plot is entirely self-explanatory and fully compliant with specific institutional or academic reporting standards.

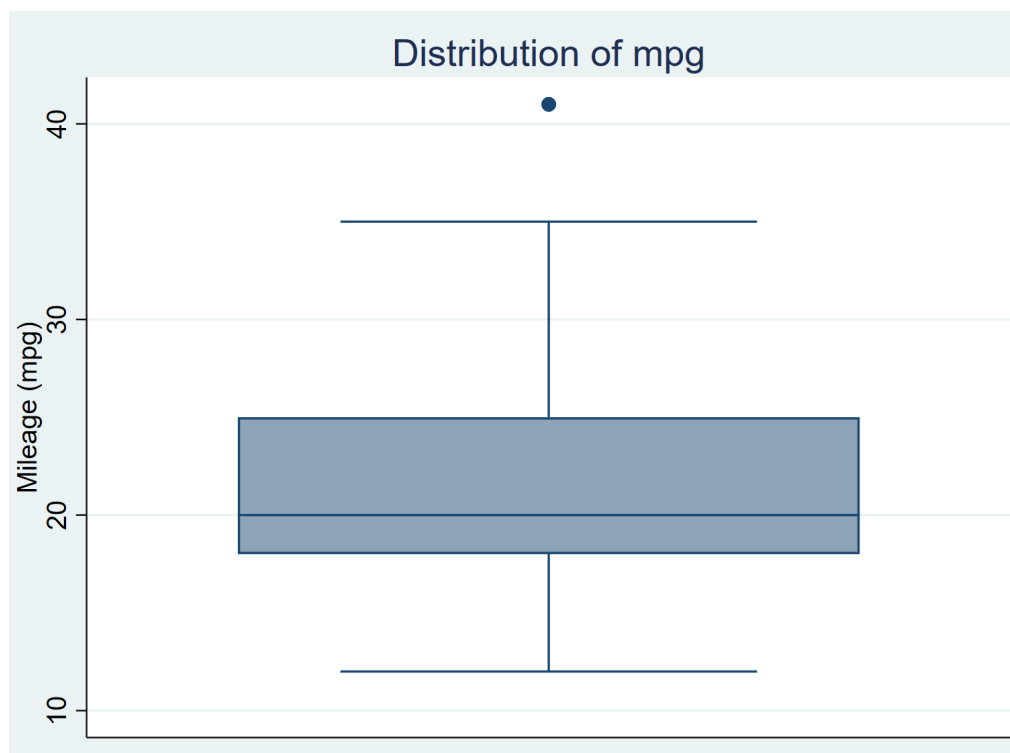
Crucially, all graphical customizations are applied through options that follow the comma in the primary graphing command (e.g., `graph box mpg,`). Mastering the effective use of these options fundamentally transforms the raw statistical output into a professional, highly interpretable visualization.

Incorporating Descriptive Titles and Subtitles

The inclusion of a descriptive title is arguably the most fundamental form of annotation. The `title()` option allows users to place a main heading on the plot, unequivocally stating the visualization's purpose and content. For all formal reports and academic submissions, a clear, informative title is mandatory for proper documentation and citation.

We can add a title to the plot using the `title()` command, as shown below:

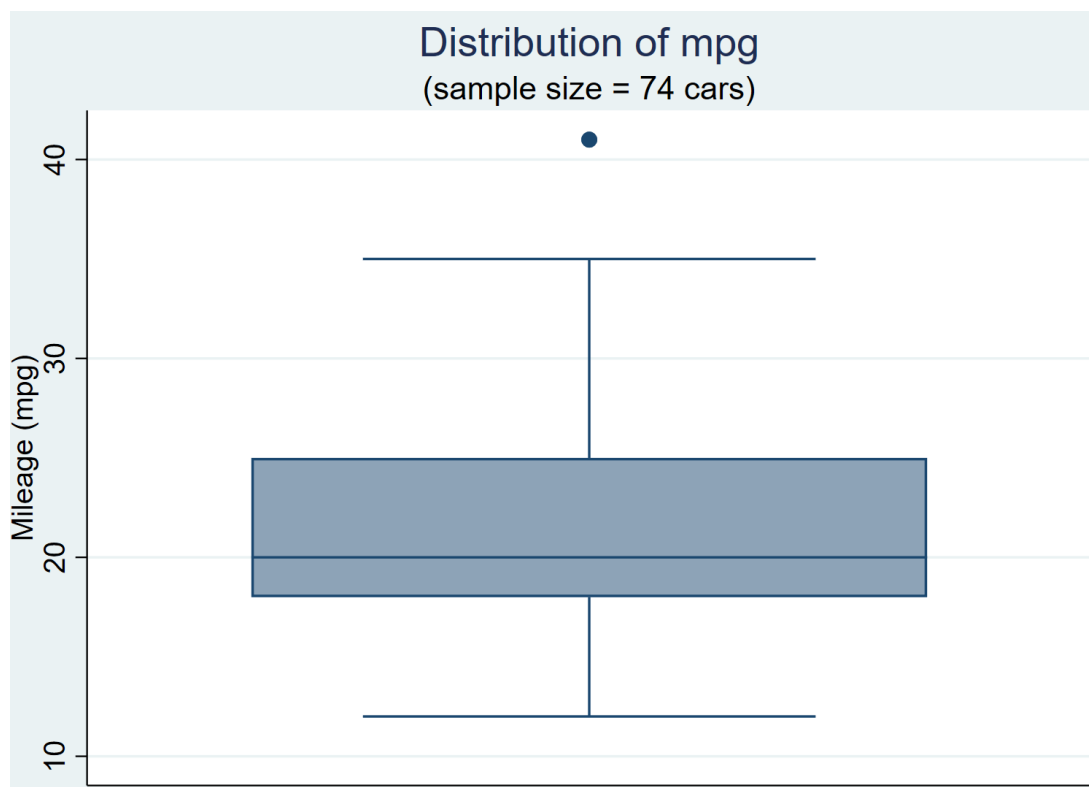
```
graph box mpg, title("Distribution of mpg")
```

For conveying supplementary details that support the main title but are not critical to the heading itself (e.g., specifics on sample size, cohort definitions, or the time frame of data collection), the `subtitle()` command proves invaluable. The subtitle is neatly positioned directly beneath the main title, offering necessary context without introducing clutter into the primary heading. This technique is particularly effective when juxtaposing visualizations derived from different subsets of the overall data population.

We can also add a subtitle underneath the main title using the `subtitle()` command:

```
graph box mpg, title("Distribution of mpg") subtitle("(sample size = 74 cars)")
```

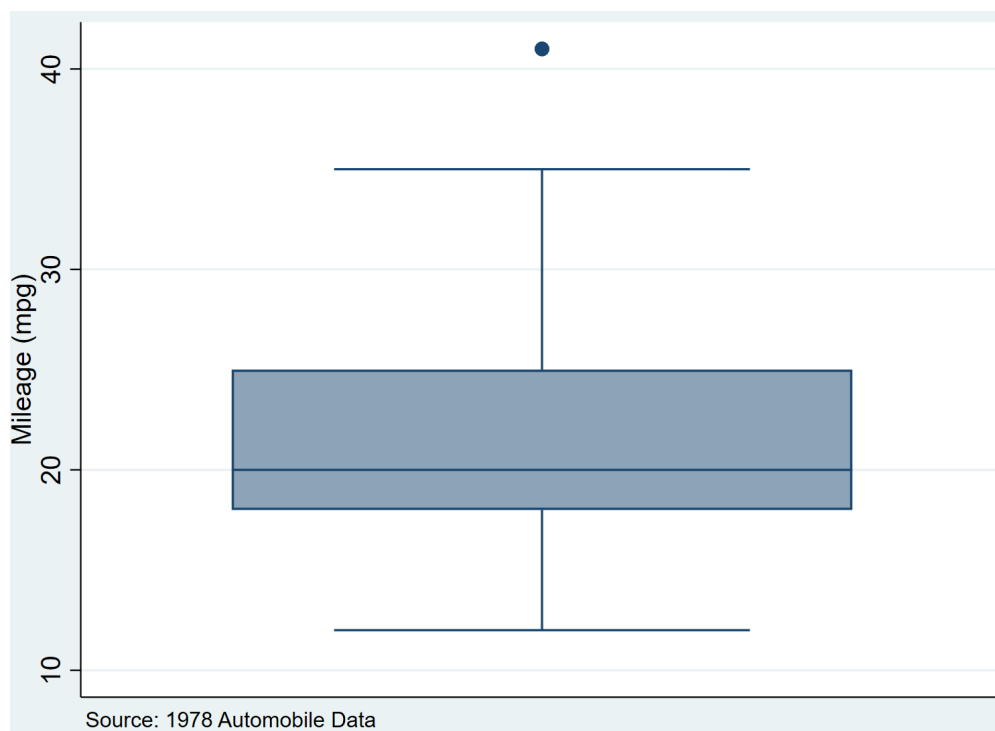


Including Source Notes and Attribution

For any scholarly or professional publication, proper attribution and methodological notes are essential components. The `note()` command is designed to place a comment, source citation, or brief methodology summary at the very bottom of the graph, typically rendered in a smaller font size. This location is ideal for specifying the original data sources, documenting specific data exclusions applied during cleaning, or including copyright details, thereby ensuring that the reader is fully informed about the provenance of the visualized data.

We can add a source note or methodological comment at the bottom of the graph by utilizing the `note()` command:

`graph box mpg, note("Source: 1978 Automobile Data")`



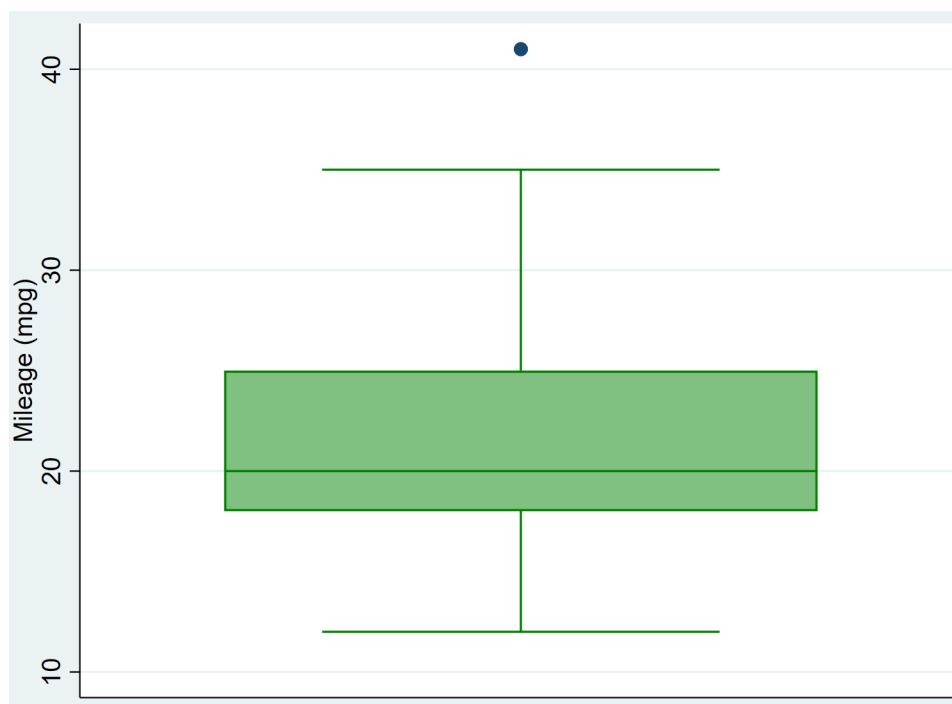
Customizing Box Colors for Visual Impact

Aesthetic customization extends fully to the color scheme used within the plot itself. Changing the color of the box plot elements can significantly aid in distinguishing groups, aligning the visualization with institutional or corporate branding guidelines, or enhancing visual contrast for improved accessibility. Stata utilizes an indexed system for plot elements, and the `box()` option provides granular, precise control over the visual properties of the box components.

The syntax `box(variable #, color(color_choice))` specifically targets a designated box group (indexed by the number, usually `1` for a single variable plot) and applies the chosen color. Stata supports a wide range of predefined color names (e.g., `teal`, `maroon`, `gold`) as well as exact RGB color codes, providing maximum flexibility for visual branding.

We can change the actual color of the box plot by applying the `box(variable #, color(color_choice))` command:

```
graph box mpg, box(1, color(green))
```



For analysts who require highly specific or complex color palettes, it is strongly recommended to consult the official [Stata Graphics Manual](#). This resource details the full inventory of available color schemes and advanced options for fine-tuning every aspect of the visual appearance. Mastery of these powerful modification commands ensures that all Stata outputs are not only statistically rigorous but also visually compelling and professionally rendered.

Beyond the Basics: Additional Resources for Stata Graphics

The process of generating and customizing box plots detailed here represents only the entry point into Stata's extensive and sophisticated graphics capabilities. Researchers who aim for more advanced visualizations--such as overlaying kernel density plots, customizing axis scales far beyond the standard options, or employing global scheme files for universal aesthetic control--should dedicate time to delving deeply into the official Stata documentation.

The Stata graphics manual provides exhaustive explanations of the command syntax, particularly focusing on the intricate interaction between core graphical commands and powerful options like `by()` for paneling, `ytick()` for detailed axis control, and advanced options for manipulating color schemes. Furthermore, actively engaging with the Stata community forums and exploring user-written commands often yields innovative and highly effective solutions for niche or extremely customized plotting requirements that extend significantly beyond the fundamental features covered in this guide. Exploring these comprehensive resources is the definitive key to unlocking and leveraging the full, robust potential of statistical visualization within the Stata ecosystem.