Data Visualization With Stata

Andy Grogan-Kaylor

2024-08-30

Table of contents

1 Introduction

Stata is a powerful and intuitive data analysis program. Learning how to graph in Stata is an important part of learning how to use Stata. Yet, until recently, the default graphs in Stata have been less than optimal. However, recent versions of Stata have a very professional looking and aesthetically appealing default graph scheme.

This document is an introduction to (a) basic graphing ideas in Stata; and (b) a quick note on the use of schemes to customize your Stata graphs.

2 What are Variables?

- By variables, I simply mean the columns of data that you have.
- For our purposes, you may think of variables as synonymous with questionnaire items, or columns of data.

	Column 1	Column 2	Column 3
Row 1			
Row 2			
Row 3			

3 Variable Types

- *Categorical variables* represent unordered categories like *race*, *ethnicity*, *neighborhood*, *religious affiliation*, or *place of residence*.
- *Continuous variables* represent a continuous scale like *income*, a *mental health scale*, or a *measure of life expectancy*.

4 A Data Visualization Strategy

Once we have discerned the type of variable that have, there are two followup questions we may ask before deciding upon a graphing strategy:

- Is our graph about **one thing at a time**?
  - How much of x is there?
  - What is the distribution of x?
- Is our graph about **two things at a time**?
  - What is the relationship of x and y?
  - How are x and y associated?

5 Data Source

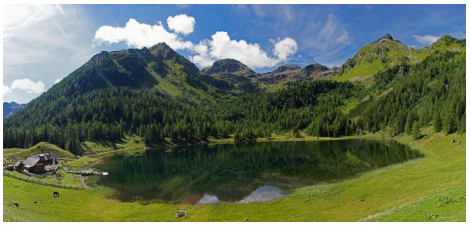

Figure 1: Norway Spruce and Larch Forest in Austrian Alps

Image Source: https://ec.europa.eu/jrc/en/research-topic/forestry/qr-tree-project/norway-spruce

The data used in this example are derived from the R package *Functions and Datasets for "Forest Analytics with R"*.

According to the documentation, the source of these data are: "von Guttenberg's Norway spruce (Picea abies [L.] Karst) tree measurement data."

Figure 2: Old Tjikko, a 9,550 Year Old Norway Spruce in Sweden

The documentation goes on to further note that:

> "The data are measures from 107 trees. The trees were selected as being of average size from healthy and well stocked stands in the Alps."

```
use "https://github.com/agrogan1/Stata/raw/main/data-visualization-with-Stata/
gutten.dta", clear
```

6 Variables

`site` Growth *quality* class of the tree's habitat. 5 levels.

`location` Distinguishes tree *location*. 7 levels.

`tree` An identifier for the tree within location.

`age_base` The tree age taken at ground level.

`height` Tree height, m.

`dbh_cm` Tree diameter, cm.

`volume` Tree volume.

`age_bh` Tree age taken at 1.3 m.

`tree.ID` A factor uniquely identifying the tree.

7 Graphs

7.1 One Continuous Thing At A Time (`histogram x`)

```
histogram height, title("Tree Height")

graph export myhistogram.png, width(1000) replace
```

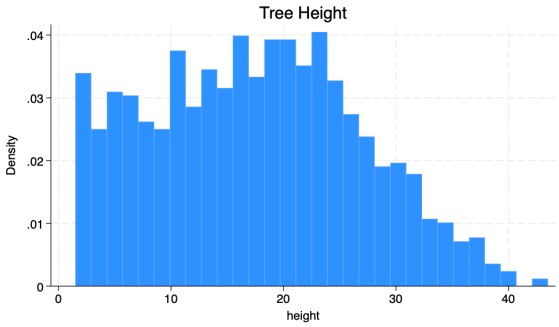

Figure 3: Histogram Of Tree Height

7.2 One Categorical Thing At A Time (`graph bar, over(x)`)

```
graph bar, over(location) title("Tree Location")

graph export mybargraph.png, width(1000) replace
```

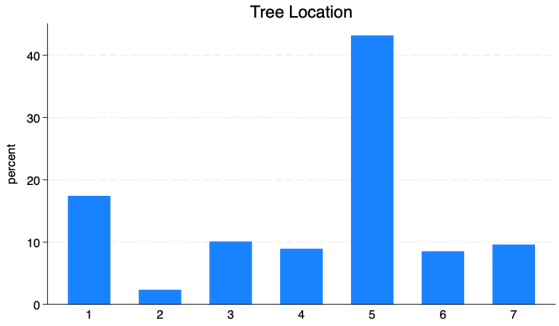

Figure 4: Bar Graph Of Tree Location

7.3 Continuous by Continuous (`twoway scatter y x`)

```
twoway scatter height age_base, title("Tree Height by Age")

graph export myscatter.png, width(1000) replace
```

Figure 5: Scatterplot Of Tree Height By Age

7.4 Categorical by Categorical (`graph bar, over(x) over(y)`)

```
graph bar, over(site) over(location) title("Tree Site Growth Quality by
Location")

graph export mybargraph2.png, width(1000) replace
```

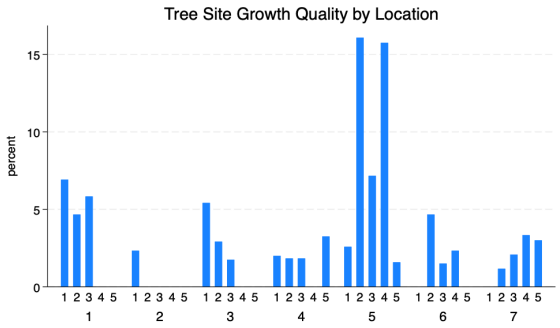

Figure 6: Bar Graph Of Tree Site By Location

7.5 Continuous by Categorical (`graph bar y, over(x)`)

```
graph bar height, over(location) title("Tree Height by Location")

graph export mybargraph3.png, width(1000) replace
```

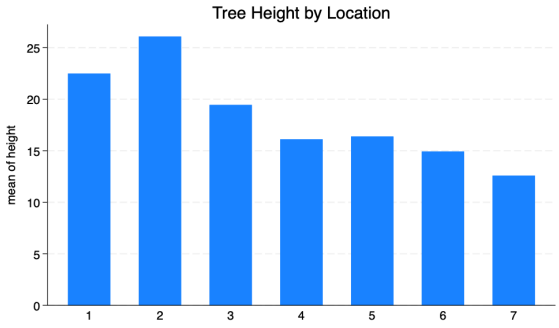

Figure 7: Bar Graph Of Mean Tree Height By Location

8 Schemes (`,scheme(...)`)

Stata *graph schemes* can substantially change the look of a graph. Built in graph schemes include `s1color`, the new default scheme `stcolor`, the older default scheme `s2color`, `sj`, `economist` and `s1rcolor`.

`lean2` (type `findit lean2` in the Stata Command Window) is a user written scheme that is very helpful when preparing graphics for publication. I have written a Stata Michigan graph scheme that can be installed. `cleanplots` and `modern` are excellent graph schemes that can be installed directly into Stata from GitHub. `burd` is another user written graph scheme that *somewhat* replicates the look of `ggplot`. Asjad Naqvi has written an excellent and comprehensive set of Stata graph schemes.

8.1 Continuous by Continuous (`twoway scatter y x, scheme(...)`)

```
twoway scatter height age_base, title("Tree Height by Age") scheme(michigan)

graph export myscatterM.png, width(1000) replace
```

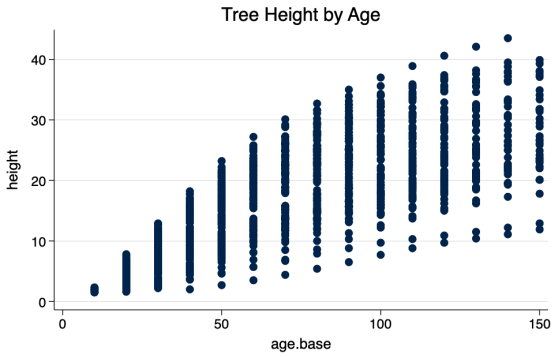

Figure 8: Scatterplot Of Tree Height By Age With Michigan Graph Scheme

```
twoway scatter height age_base, title("Tree Height by Age") scheme(lean2)
msymbol(o)
```

```
graph export myscatterL.png, width(1000) replace
```

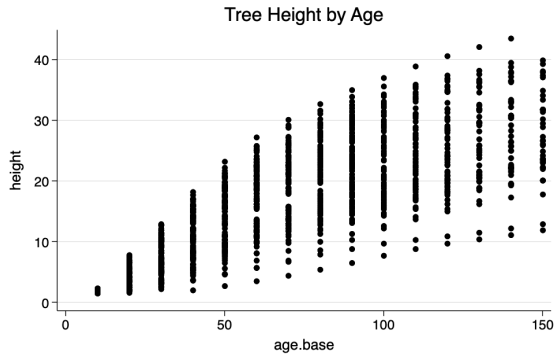

Figure 9: Scatterplot Of Tree Height By Age With lean2 Graph Scheme

```
twoway scatter height age_base, title("Tree Height by Age") scheme(s1color)

graph export myscatterS.png, width(1000) replace
```

Figure 10: Scatterplot Of Tree Height By Age With s1color Graph Scheme

```
twoway scatter height age_base, title("Tree Height by Age") scheme(burd) ///
  msymbol(o) graphregion(lcolor(none))

graph export myscatterB.png, width(1000) replace
```

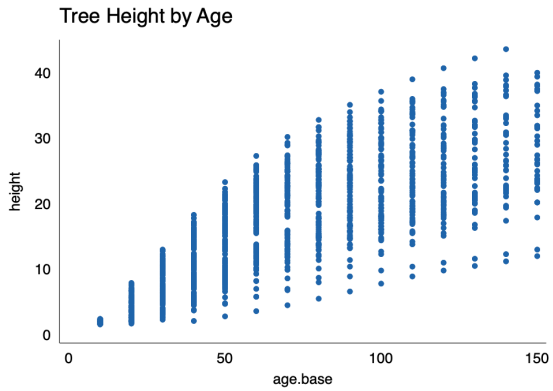

Figure 11: Scatterplot Of Tree Height By Age With burd Graph Scheme

8.2 Continuous by Categorical (`graph bar y, over(x) scheme(...)`)

Note that in the graph below, I have used the `asyvars` option to give different colors to the different bars.

```
graph bar height, over(location) asyvars title("Tree Height by Location")
scheme(michigan)

graph export mybarM.png, width(1000) replace
```

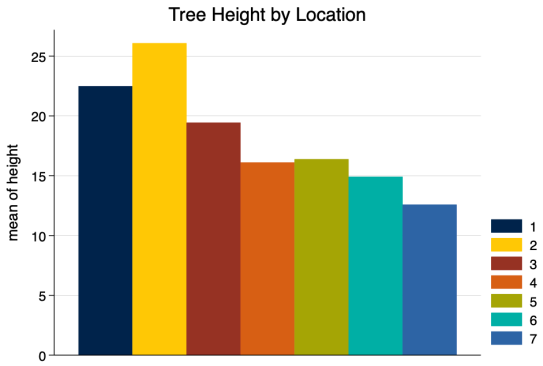

Figure 12: Bar Graph Of Mean Tree Height By Location With Michigan Graph Scheme

```
graph bar height, over(location) asyvars title("Tree Height by Location")
scheme(lean2)

graph export mybarL.png, width(1000) replace
```

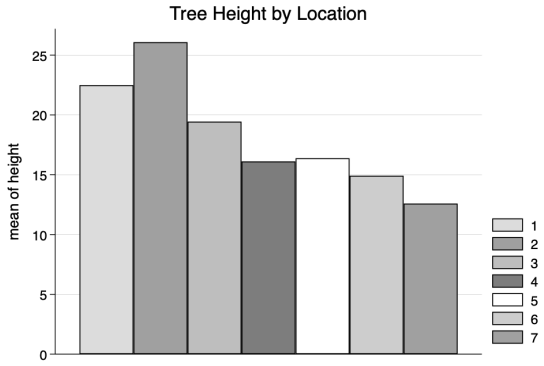

Figure 13: Bar Graph Of Mean Tree Height By Location With lean2 Graph Scheme

```
graph bar height, over(location) asyvars title("Tree Height by Location")
scheme(s1color)

graph export mybarS.png, width(1000) replace
```

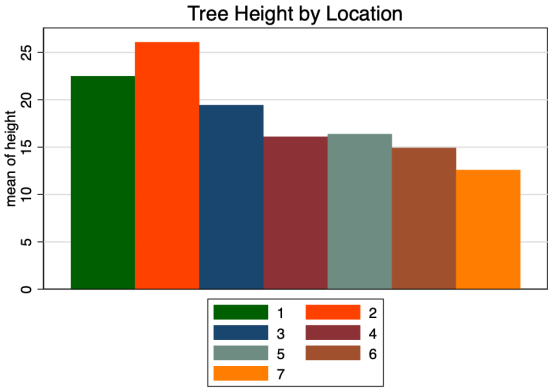

Figure 14: Bar Graph Of Mean Tree Height By Location With s1color Graph Scheme

```
graph bar height, over(location) asyvars title("Tree Height by Location")
scheme(burd) graphregion(lcolor(none))

graph export mybarB.png, width(1000) replace
```

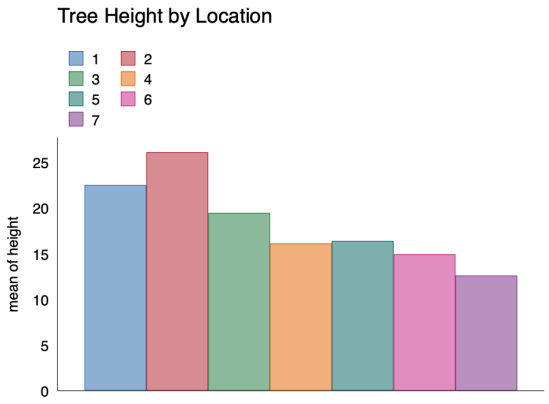

Figure 15: Bar Graph Of Mean Tree Height By Location With burd Graph Scheme