

# Data Visualization With Stata

Andy Grogan-Kaylor

24 Jan 2023

## 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, the default graphs in Stata can sometimes be less than optimal.
- This document is an introduction to (a) basic graphing ideas in Stata; and (b) a quick note on the use of schemes to make your Stata graphs look more professional.

When this document is presented in *slide show format*, some slides may be long, and you may need to *scroll down* to see the full slide. In slide show format use the left and right arrow keys to navigate through the slides. **b** will make the text bigger. **s** will make the text smaller.

## 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		

## 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*.

## 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?

## 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/master/data-visualization-with-Stata/gutten.d
> ta", clear
```

## 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.

For some purposes, it might be best to use a centered age variable, centered at the grand mean of tree age:

```
. egen ageMEAN = mean(age_base)
. generate ageCENTERED = age_base - ageMEAN
```

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.

## Graphs

### One Continuous Thing At A Time (histogram x)

```
. histogram height, title("Tree Height")
(bin=30, start=1.5, width=1.4)

. graph export myhistogram.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/Stata/data-visualization-with-Stata/myhistogram.png
  saved as PNG format
```

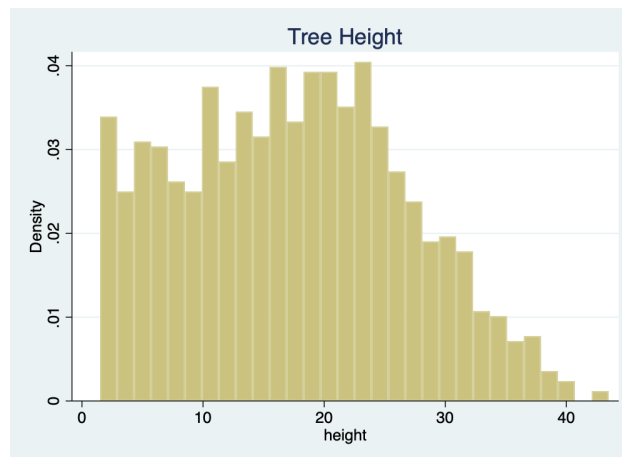


Figure 3: Histogram Of Tree Height

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

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

. graph export mybargraph.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/Stata/data-visualization-with-Stata/mybargraph.png
  saved as PNG format
```

### Continuous by Continuous (tway scatter y x)

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

. graph export myscatter.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/Stata/data-visualization-with-Stata/myscatter.png
  saved as PNG format
```

### 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
```

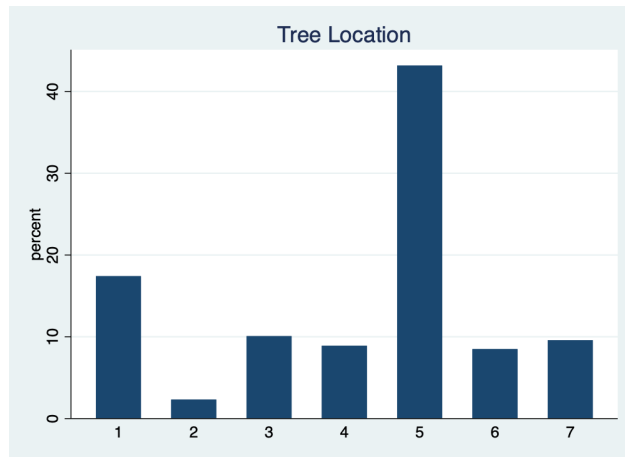


Figure 4: Bar Graph Of Tree Location

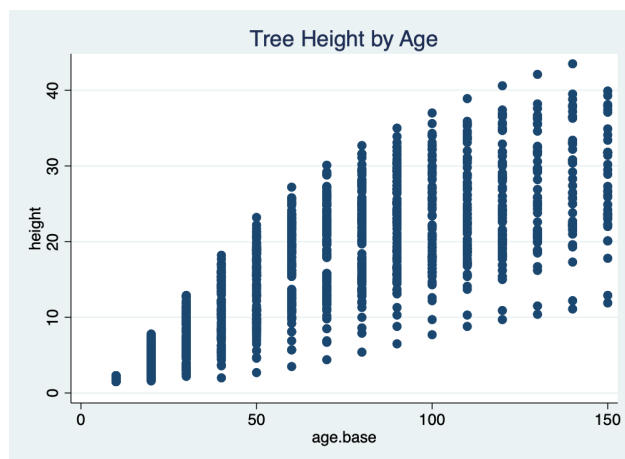


Figure 5: Scatterplot Of Tree Height By Age

```
file /Users/agrogan/Desktop/GitHub/Stata/data-visualization-with-Stata/mybargraph2.png
saved as PNG format
```

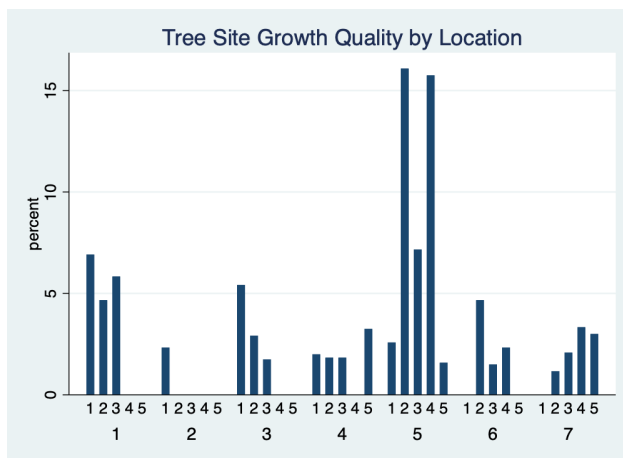


Figure 6: Bar Graph Of Tree Site By Location

## 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
file /Users/agrogan/Desktop/GitHub/Stata/data-visualization-with-Stata/mybargraph3.png
saved as PNG format
```

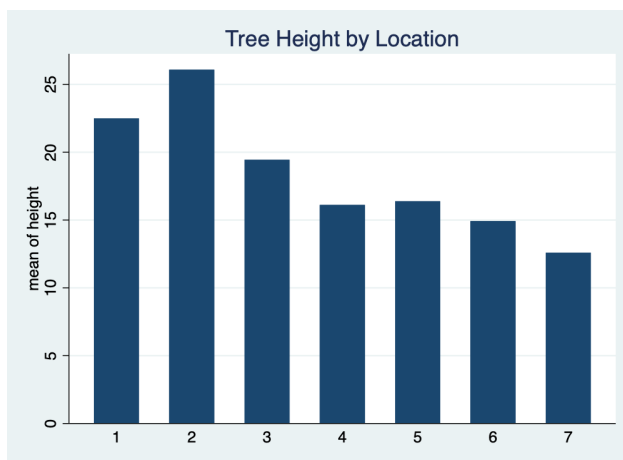


Figure 7: Bar Graph Of Mean Tree Height By Location

## Schemes (`,scheme(...)`)

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

(Notably, as a *quick fix*, the default Stata graph scheme of `s2color` can be substantially improved by adding the option `,graphregion(fcolor(white))` which changes the background fill color of this scheme to *white*.)

`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.

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

```
. twoway scatter height age_base, title("Tree Height by Age") mcolor(%25) scheme(michigan)

. graph export myscatterM.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/Stata/data-visualization-with-Stata/myscatterM.png
  saved as PNG format
```

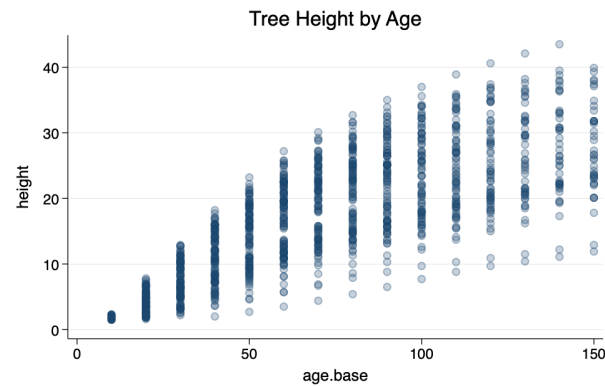


Figure 8: Scatterplot Of Tree Height By Age With Michigan Graph Scheme and Transparent Color

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

. graph export myscatterL.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/Stata/data-visualization-with-Stata/myscatterL.png
  saved as PNG format
```

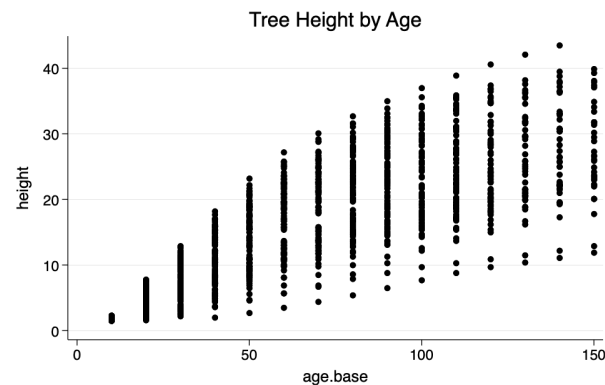


Figure 9: Scatterplot Of Tree Height By Age With lean2 Graph Scheme

```
. twoway scatter height age_base, title("Tree Height by Age") mcolor(%25) scheme(s1color)

. graph export myscatterS.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/Stata/data-visualization-with-Stata/myscatterS.png
  saved as PNG format
```



Figure 10: Scatterplot Of Tree Height By Age With slcolor Graph Scheme and Transparent Color

```
. twoway scatter height age_base, title("Tree Height by Age") scheme(burd) msymbol(o) grap
> hregion(lcolor(none))

. graph export myscatterB.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/Stata/data-visualization-with-Stata/myscatterB.png
  saved as PNG format
```

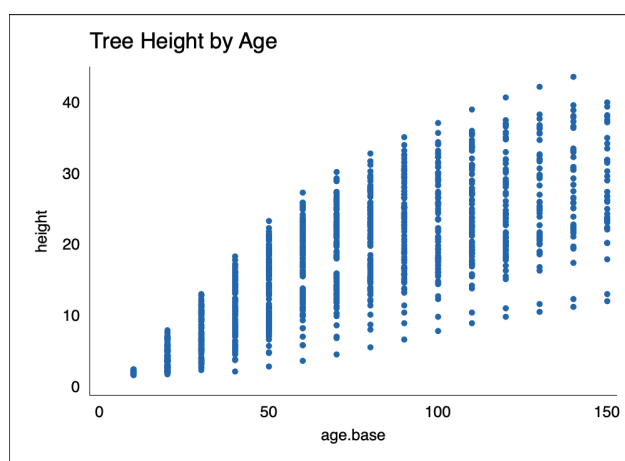


Figure 11: Scatterplot Of Tree Height By Age With burd Graph Scheme

## 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(michiga
> n)

. graph export mybarM.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/Stata/data-visualization-with-Stata/mybarM.png saved
  as PNG format

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

. graph export mybarL.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/Stata/data-visualization-with-Stata/mybarL.png saved
```

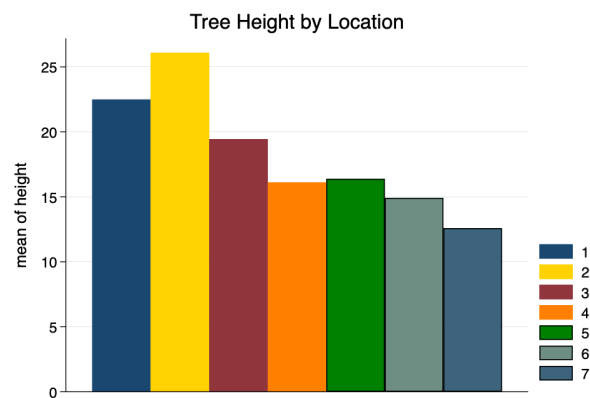


Figure 12: Bar Graph Of Mean Tree Height By Location With Michigan Graph Scheme

as PNG format

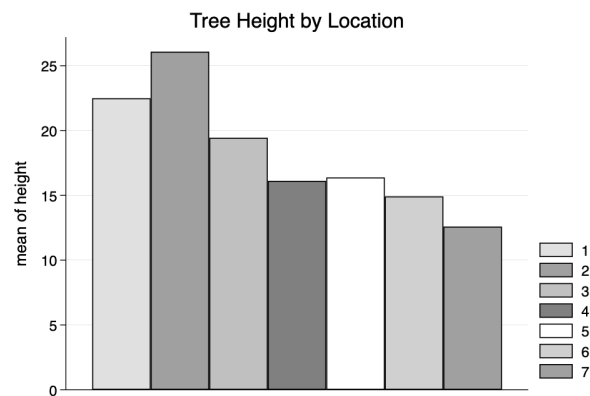


Figure 13: Bar Graph Of Mean Tree Height By Location With lean2 Graph Scheme

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

. graph export mybarS.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/Stata/data-visualization-with-Stata/mybarS.png saved
as PNG format

. graph bar height, over(location) asyvars title("Tree Height by Location") scheme(burd) g
> raphregion(lcolor(none))

. graph export mybarB.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/Stata/data-visualization-with-Stata/mybarB.png saved
as PNG format
```





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

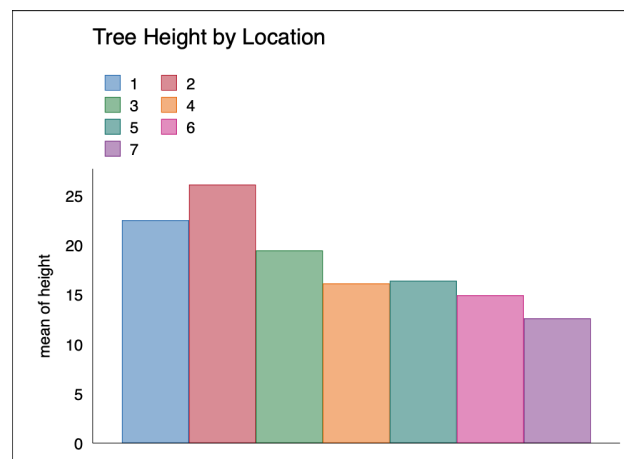


Figure 15: Bar Graph Of Mean Tree Height By Location With burd Graph Scheme