Basic Data Vizualization

Joshua F. Allen

March 21, 2021

Why This is Important

Why Can't I Just use Summary Stats?

- Measures of centeral tendency are vital for exploratory analysis

- They tell you a lot about your data

Why Can't I Just use Summary Stats?

- Measures of centeral tendency are vital for exploratory analysis

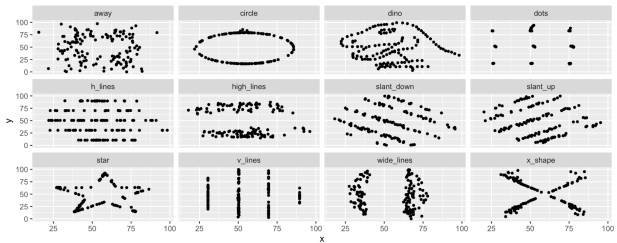
- They tell you a lot about your data

- However, they can not reveal patterns

Going Beyond Summary stats

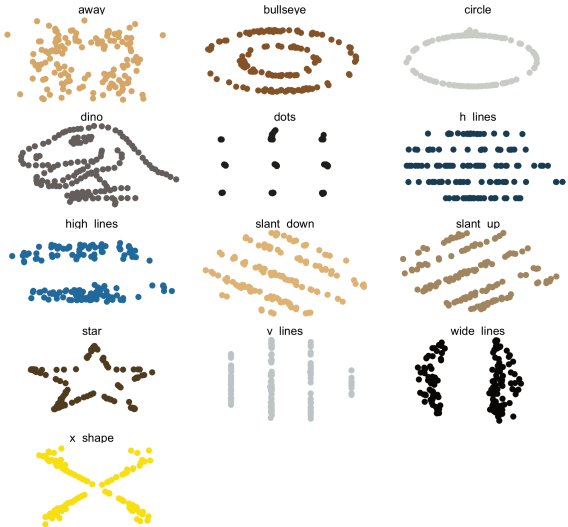
Raw data is not enough

Each of these has the same mean, standard deviation, variance, and correlation



18 / 37

Going Beyond Summary stats



What Makes a Good Vizualization?

- People have lots of opinions

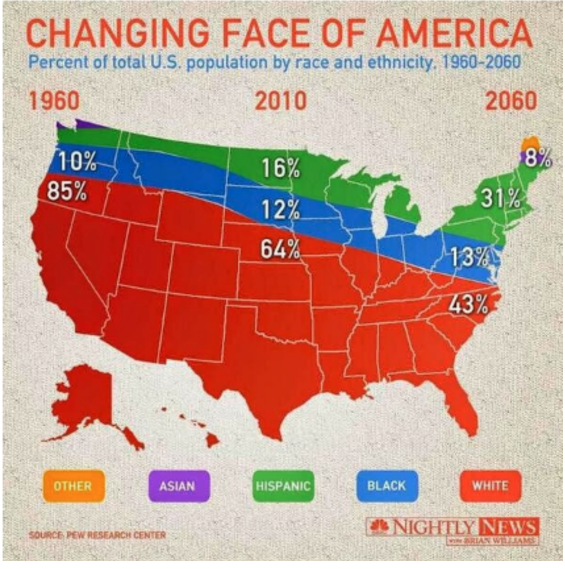
- Books are dedicated to this

- This is a legit topic scholars debate

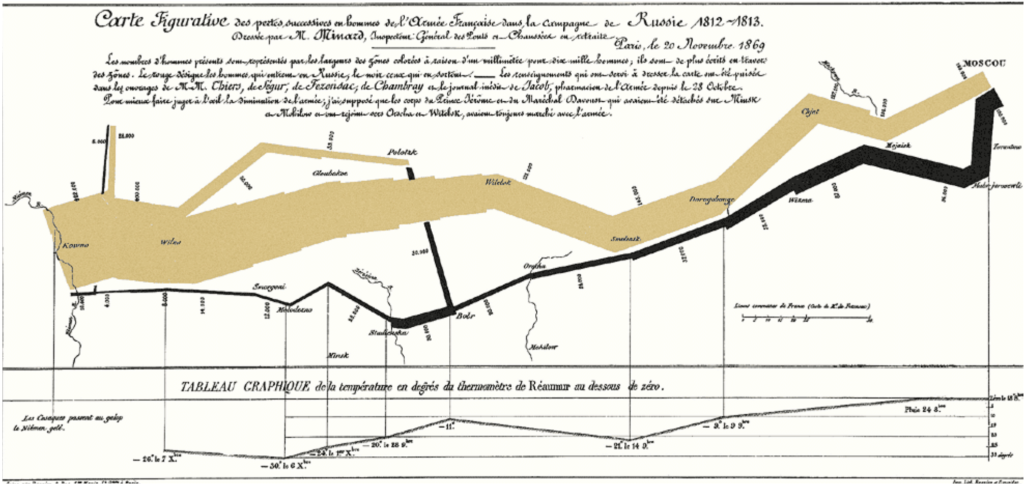
The Basics

- Faithful representation of the data
- Good Aesthics
- No perceptual issues
- Insightful

Bad Data Vizualization



Minard’s Map


Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869
MOSCOU
Polotzk
Wilna
Kowno
Smolensk
Dorogobouge
Mojaisk
Bobr
Minsk
Mohilow
Orscha
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.

Coding

Should I Learn to Code?

- STATA, SAS, Excel, and SPSS have dropdown menus

Should I Learn to Code?

- STATA, SAS, Excel, and SPSS have dropdown menus

- You can get away with never learning the basics

Should I Learn to Code?

- STATA, SAS, Excel, and SPSS have dropdown menus

- You can get away with never learning the basics

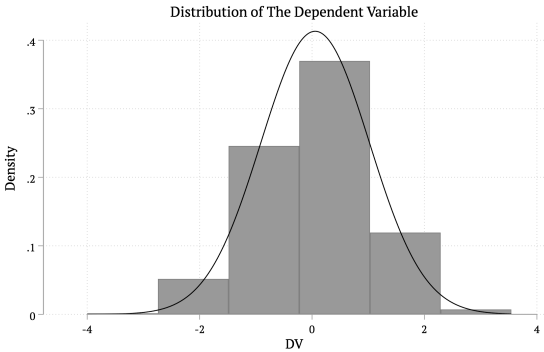
- It's really easy to do it that way..Kind of

Coding Continued

The big news in economics last week was the paper by a UMASS Grad Student showing that economists Kenneth Rogoff and Carmen Reinhart had made an Excel spreadsheet blunder in their famous paper arguing that as debt-to-GDP goes above 90%, growth slows dramatically.

- It is bad practice to do stuff in the console or by drop downs
- It leads to mistakes
- And you won't ever really remember how to do things

Try Recreating this Figure

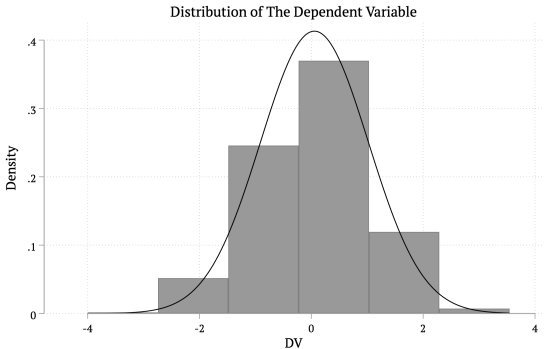



```
clear
*this will generate fake data
set seed 1994

set obs 1000

 g y = rnormal(0,1) /* mean = 0, sd
```

How I created this figure




```
hist y, graphregion(color(white))
bin(6) bcolor(gs8 start(-4)
normal /// xtitle("DV")
title("Distribution of The
Dependent Variable") ///
normopts(lcolor(black))
```

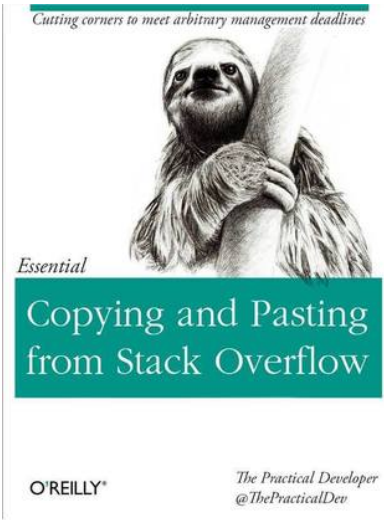
Learning To Code

- It is hard and often frustrating at first.

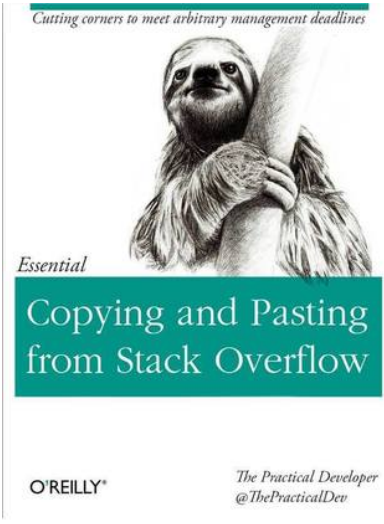
Learning To Code

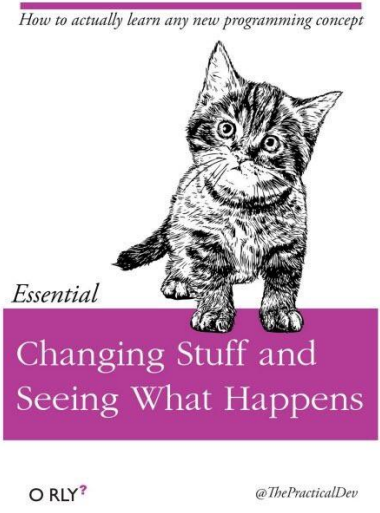
- It is hard and often frustrating at first.
- You got this
- Understanding the fundamentals will let you do a lot of things.

Resources


Cutting corners to meet arbitrary management deadlines
Essential
Copying and Pasting from Stack Overflow
O'REILLY
The Practical Developer
@ThePracticalDev

Resources


Cutting corners to meet arbitrary management deadlines
Essential
Copying and Pasting from Stack Overflow
O'REILLY
The Practical Developer
@ThePracticalDev


How to actually learn any new programming concept
Essential
Changing Stuff and Seeing What Happens
O RLY?
@ThePracticalDev

Stata Basics

Things to Keep in Mind When Making Graphs

- Stata has notoriously ugly defaults
- Stata® has lots of opinions about what you should not do
- All the graphing commands have similar syntax, but with slight tweaks
- These tweaks can cause you to get grumpy

Things to Keep in Mind When Making Graphs

- Stata has notoriously ugly defaults
- Stata® has lots of opinions about what you should not do
- All the graphing commands have similar syntax, but with slight tweaks
- These tweaks can cause you to get grumpy
- Like really grumpy

Basic Grammar

```
command varlist [if] [in], options
```

- "varlist" is a column in your dataset
- "if" is a set of conditional statements
- "in" is usually a some rows in your dataset or used for weights

Coding Time!

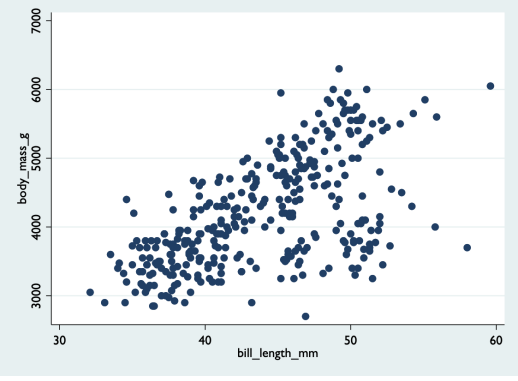
Very Basics

- The best practice is to have an individual folder for each project

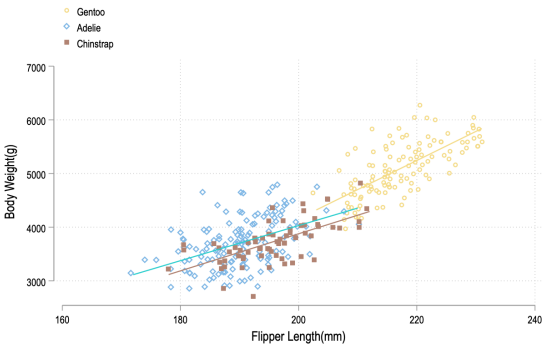
- you should have Stata pointed at that folder

- this is done by cd "path/to/your/file" on Mac

Default Stata Graph vs Highly Customized



Default Stata Graph vs Highly Customized



Why is it == and not =?

- Stata uses what are called booleans
- "&" is and
- "|" is or
- "!" is not.
- "==" is equal to
- "=" in most software languages is used for assignment

Example of Assignment in Stata

```
egen toy_var = mean(other_var)
```

Appendix

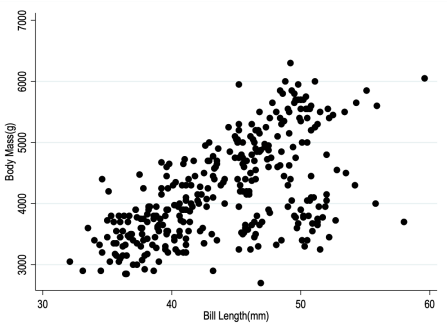
Resources

- Open Source Course in R

- Stata Cheat Sheet

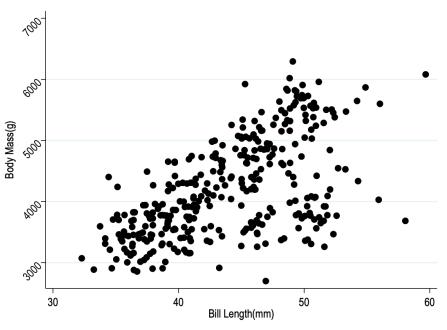
- Open Source Course for learning Stata

Manually Cleaning Up Stata Plots



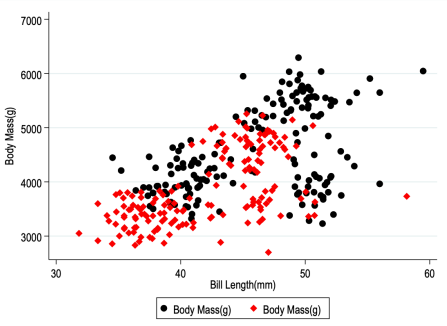

```
tw scatter body_mass_g
bill_length_mm,
graphregion(color(white)) ///
xtitle("Bill Length(mm)")
ytitle("Body Mass(g)")
mcolor(black)
```

Notice How the Points Overlap?



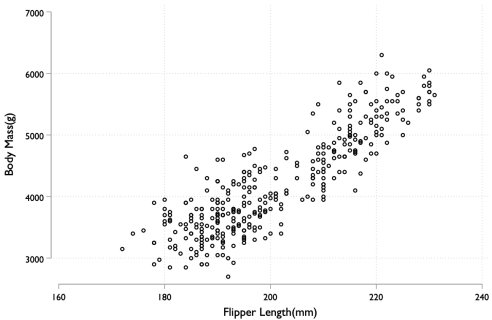

```
tw scatter body_mass_g
bill_length_mm,
graphregion(color(white)) ///
ylabel(,angle(45)) /// jitter(2)
jitterseed(1994) xtitle("Bill
Length(mm)") mclor(black)
```

Conditional Statements



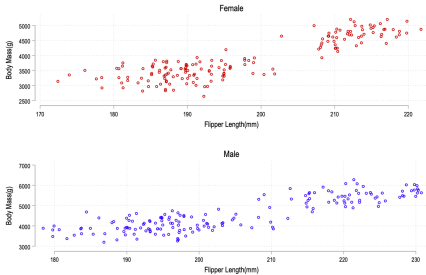

```
tw (scatter body_mass_g
bill_length_mm if
sex=="male",mcolor(black) ///
jitter(2) jitterseed(1994)) ///
(scatter body_mass_g bill_length_mm
if sex=="female", msymbol(d)
mcolor(red) /// jitter(2)
jitterseed(1994)), ///
graphregion(color(white)) ///
ylabel(,angle(360))
```

Schemes




```
tw scatter body_mass_g
flipper_length_mm, mcolor(black)
scheme(cleanplots)
set scheme cleanplots, perm /*sets
schemes permenately */
```

Combining Two Plots

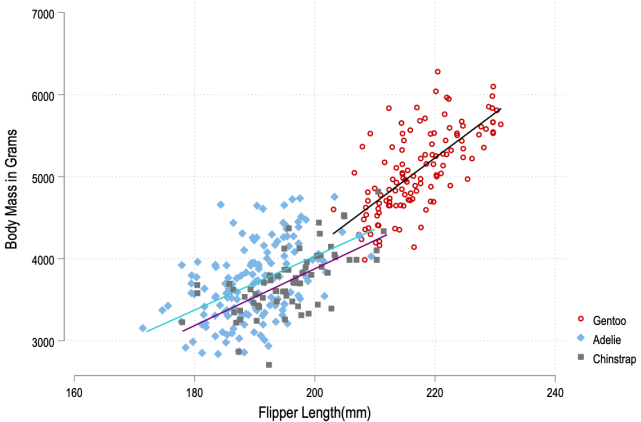



```
tw scatter body_mass_g
flipper_length_mm if sex
== "female", jitter(2)
/// jitterseed(1994)
name(female_scatter,replace)
tw scatter body_mass_g
flipper_length_mm if sex == "male",
jitter(2) /// jitterseed(1994)
name(male_scatter,replace)
gr combine female_scatter
male_scatter, col(1)
```

Combining Two Different Kinds of Plots

- We often want to convey other kinds of information
- You can add different kinds of plots to do this
- There are various ways to do this in Stata

Consider This Plot



Consider This Plot

```
tw (scatter body_mass_g flipper_length_mm if species== "Gentoo", ///
jitter(2) jitterseed(1994)) ///
(scatter body_mass_g flipper_length_mm if species== "Adelie", ///
jitter(2) jitterseed(1994) msymbol(d)) ///
(scatter body_mass_g flipper_length_mm if species== "Chinstrap", ///
jitter(2) jitterseed(1994) msymbol(s)) ///
(lfit body_mass_g flipper_length_mm if
species== "Gentoo") /// (lfit body_mass_g flipper_length_mm if
species== "Adelie") ///
(lfit body_mass_g flipper_length_mm if
species== "Chinstrap" ), ///
legend(order(1 "Gentoo" 2 "Adelie" 3 "Chinstrap")) ///
ytitle("Body Mass in Grams")
```