Data Visualization & Exploratory Data Analysis

Statistics can be ugly

```
> summary(model)

Call:
lm(formula = y ~ x + z, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8090 -0.7421  0.0217  0.6816  3.7718

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.035250   0.032817  -1.074    0.283
x            0.021991   0.033189   0.663    0.508
z           -0.003659   0.031512  -0.116    0.908

Residual standard error: 1.036 on 997 degrees of freedom
Multiple R-squared: 0.0004585, Adjusted R-squared: -0.001547
F-statistic: 0.2287 on 2 and 997 DF,  p-value: 0.7956
```

Numbers are scary

OLS: Actual Class Size - gktmaths

VARIABLES	(1) 1	(2) 2	(3) 3	(4) 4
1.gkclasst	7.73**	8.93***	9.01***	8.84***
	(3.76)	(2.38)	(2.33)	(2.32)
3.gkclasst	-0.40	0.28	0.63	0.42
	(3.87)	(2.18)	(2.15)	(2.14)
st_whiteasian			16.82***	16.91***
			(2.38)	(2.40)
st_girl			6.53***	6.46***
			(1.12)	(1.12)
freelunch			-20.15***	-20.08***
			(1.32)	(1.33)
t_whiteasian				-1.01
				(3.80)
gktyears				0.42**
				(0.20)
teacher_MA				-2.20
				(2.08)
Constant	483.20***	482.60***	477.63***	475.52***
	(2.80)	(1.57)	(2.37)	(4.49)
Observations	5,871	5,871	5,852	5,809
R-squared	0.01	0.01	0.07	0.07
Number of gkschid		79	79	79

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Walls of numbers

Table 4: Simulation results
N=4000; 500 runs per variation and motivation; standard deviation in parentheses.

ive_prevalence	objective_value	dv	baseline_social	market_social	costless_social	with_cost_social	baseline_individual	market_individual	costless_indi
	High	a1	0.373 (0.112)	0.93 (0.069)	1 (0)	0.955 (0.054)	0.372 (0.112)	0.649 (0.114)	0.865 (0.098)
	Low	a2	0.368 (0.113)	0.999 (0.006)	1 (0)	0.611 (0.131)	0.378 (0.113)	0.784 (0.107)	0.863 (0.103)
	High	b1	0.38 (0.208)	1 (0)	1 (0)	0.909 (0.129)	0.375 (0.213)	0.993 (0.038)	0.994 (0.035)
	Low	b2	0.38 (0.21)	1 (0)	1 (0)	0.664 (0.222)	0.375 (0.209)	0.996 (0.033)	0.994 (0.036)
	High	c1	0.128 (0.077)	0.519 (0.1)	1 (0)	0.872 (0.089)	0.13 (0.078)	0.26 (0.094)	0.455 (0.117)
	Low	c2	0.128 (0.079)	0.133 (0.088)	1 (0)	0.463 (0.133)	0.121 (0.077)	0.388 (0.095)	0.449 (0.121)
	High	d1	0.127 (0.147)	0.995 (0.031)	1 (0)	0.781 (0.19)	0.13 (0.145)	0.612 (0.21)	0.558 (0.219)
	Low	d2	0.123 (0.14)	0.961 (0.112)	1 (0)	0.385 (0.221)	0.121 (0.144)	0.68 (0.202)	0.549 (0.215)

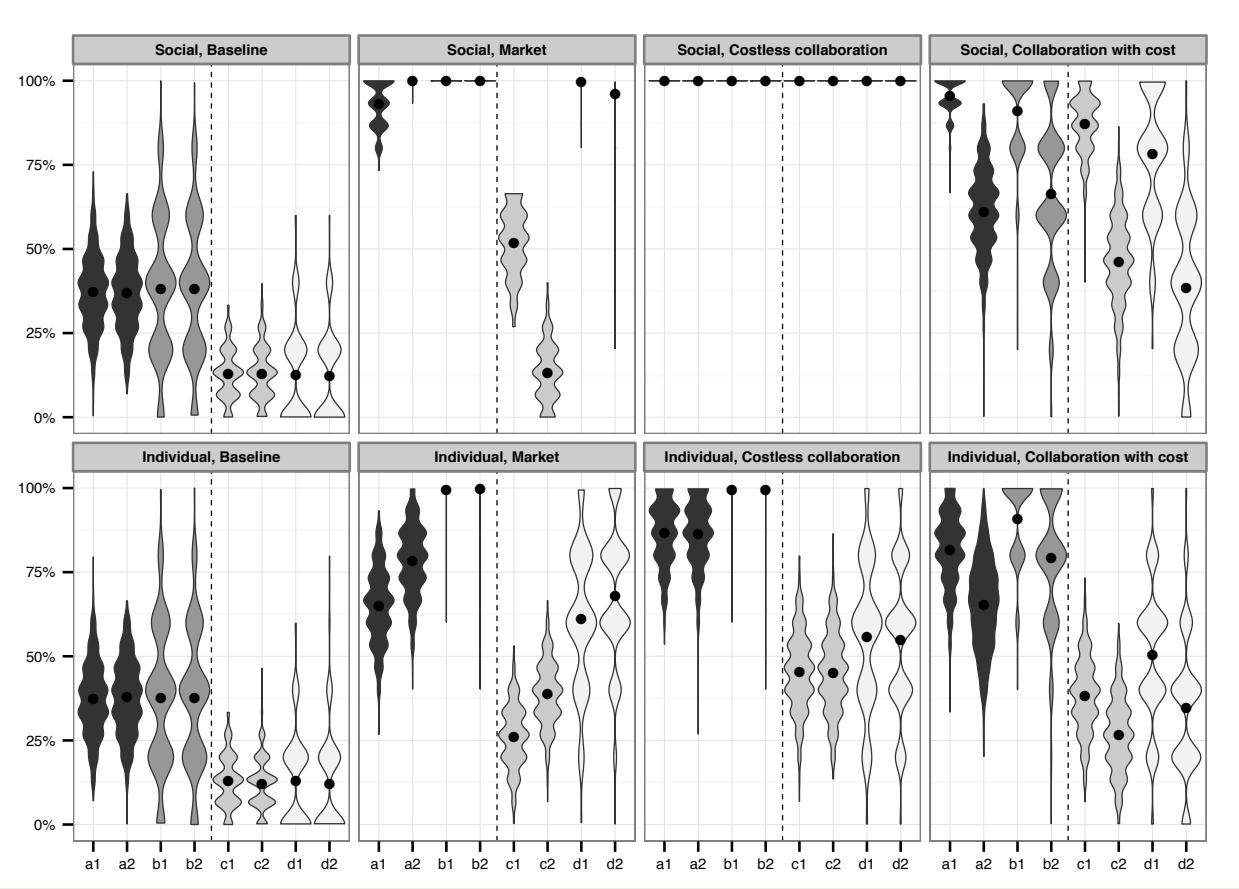
Exploratory Data Analysis

"Interocular traumatic impact"

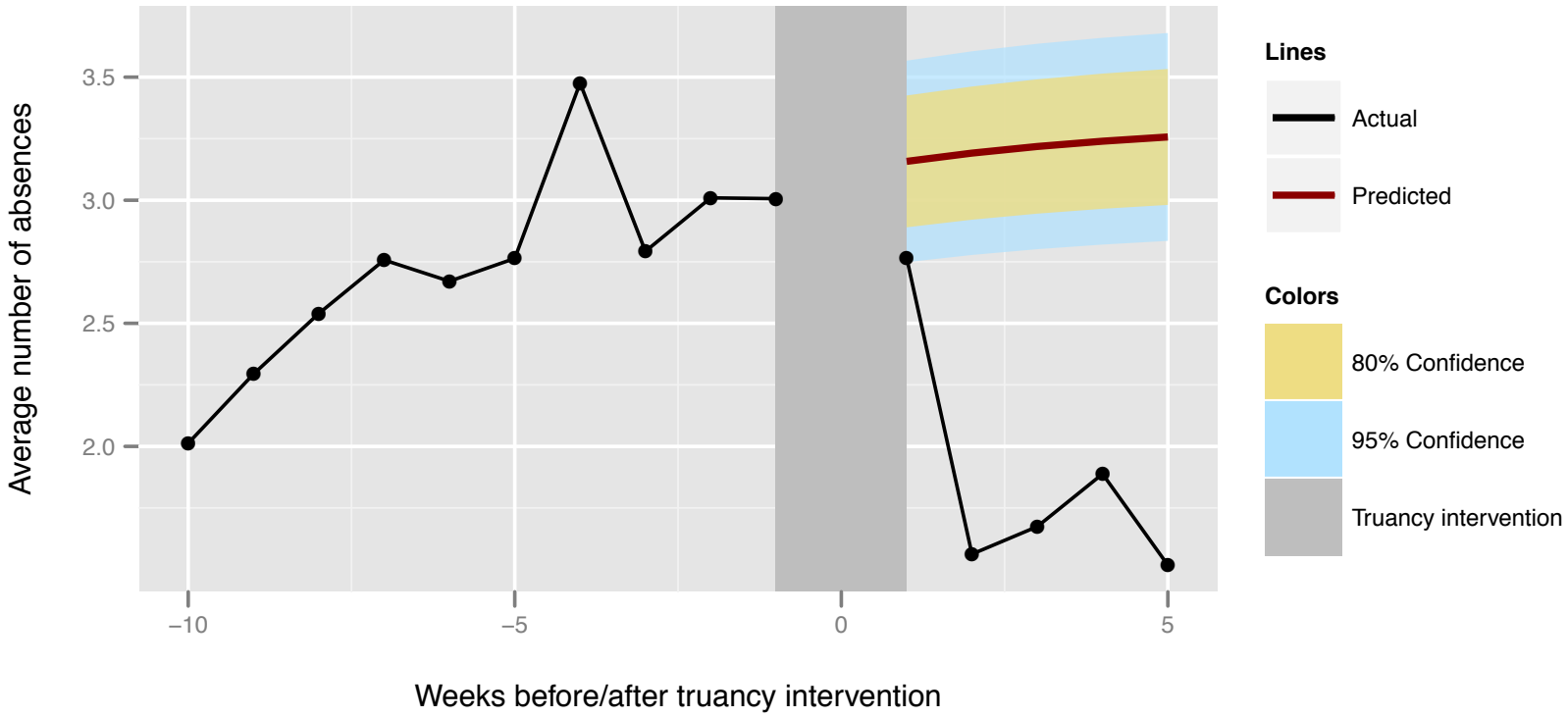
Enhance probabilistic analysis

Check shape and assumptions

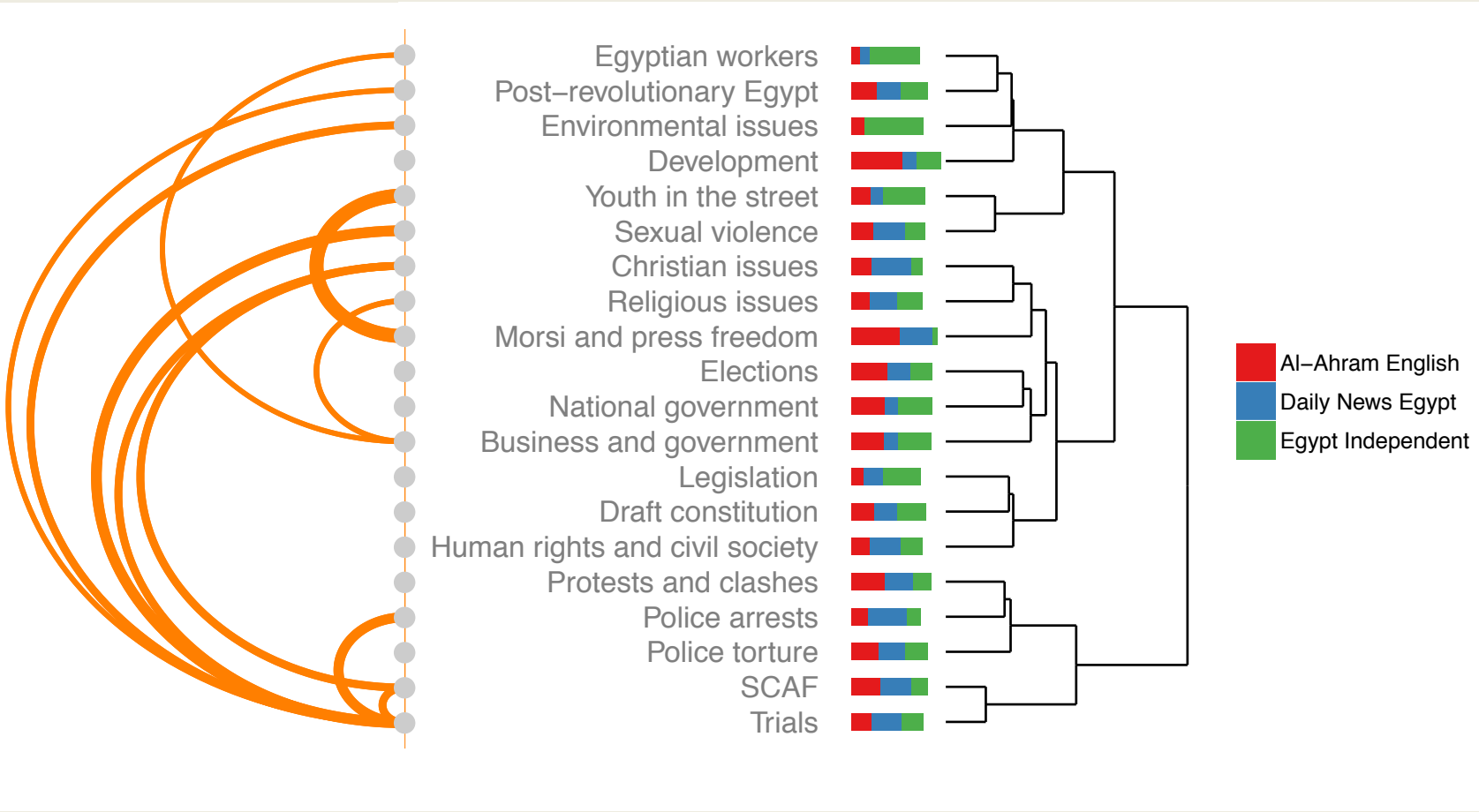
Right between the eyes…



Right between the eyes…



Right between the eyes…



Understand your data

Visualize every variable individually

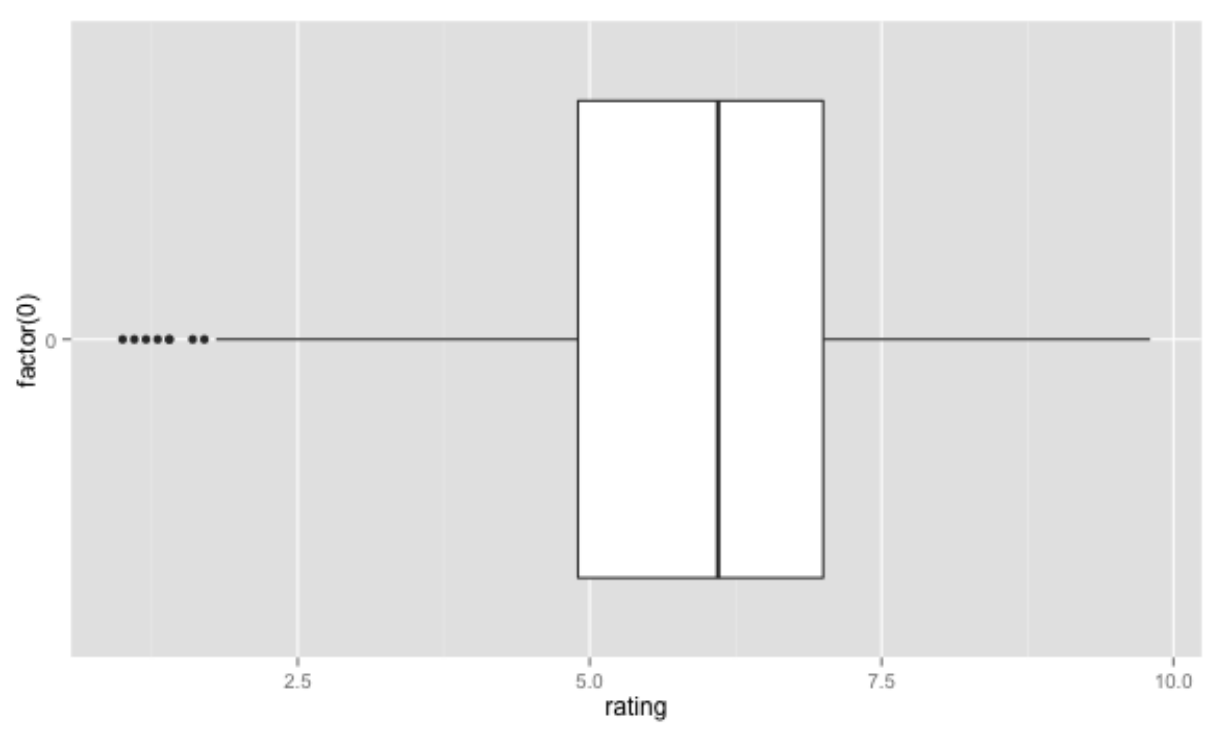
Visualize relationships between variables

Visualize models

Boxplots

Univariate visualization

Continuous data

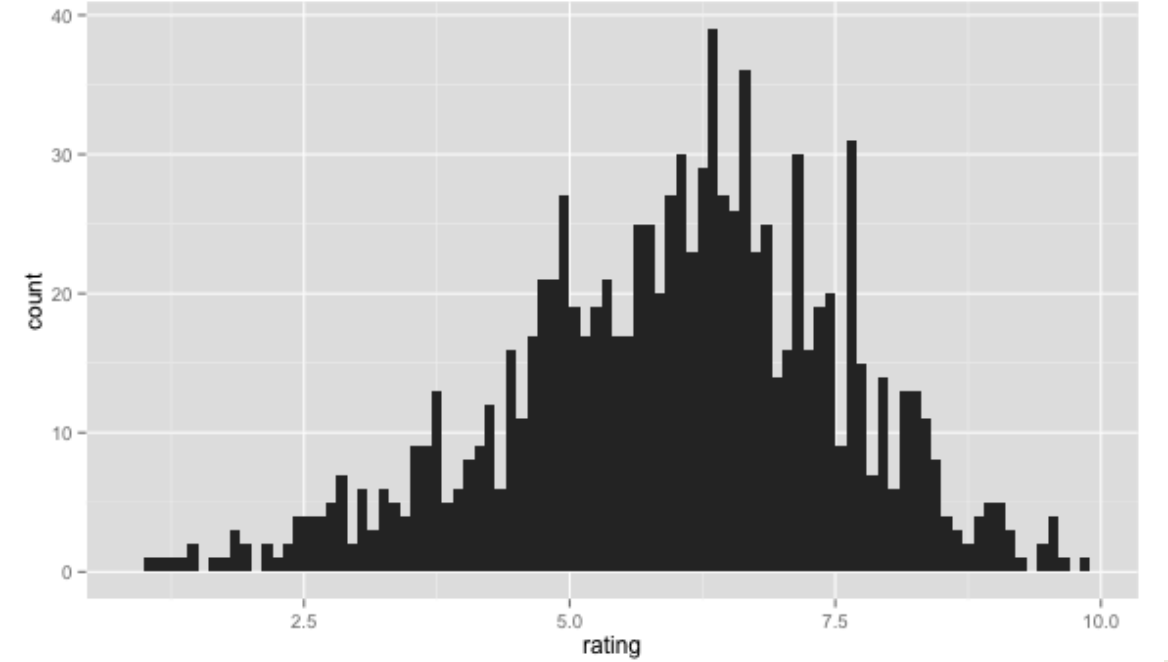



```
graph box variable
```

Histograms

Univariate visualization

Continuous data

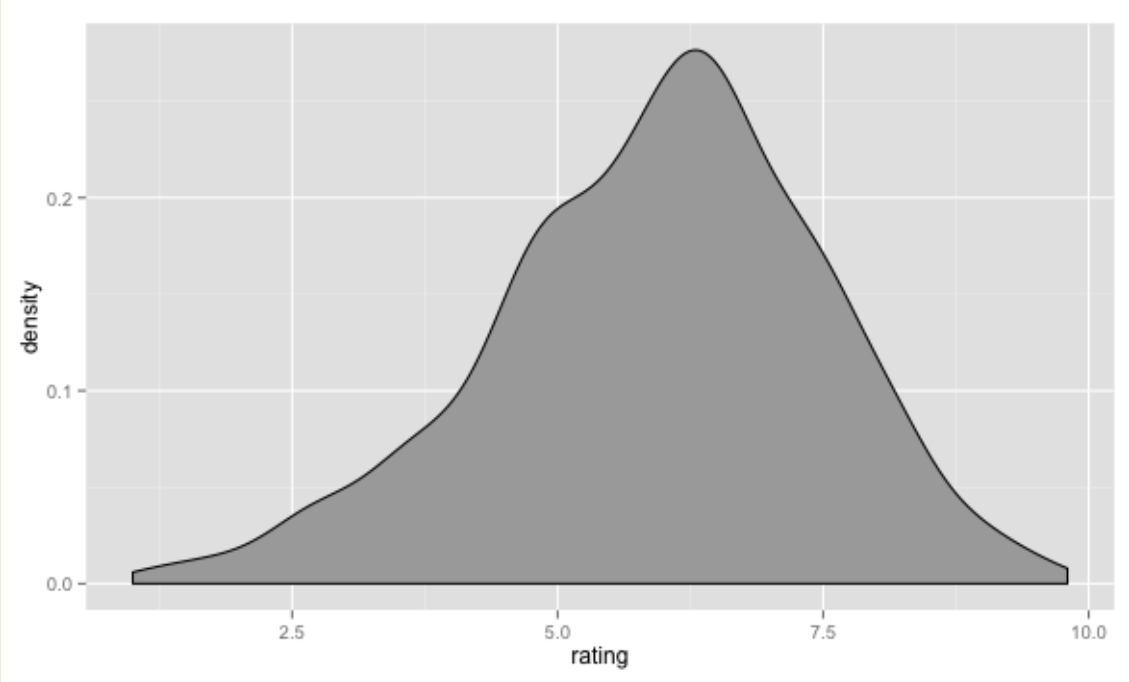



```
histogram variable
```

Density plots

Univariate visualization

Continuous data

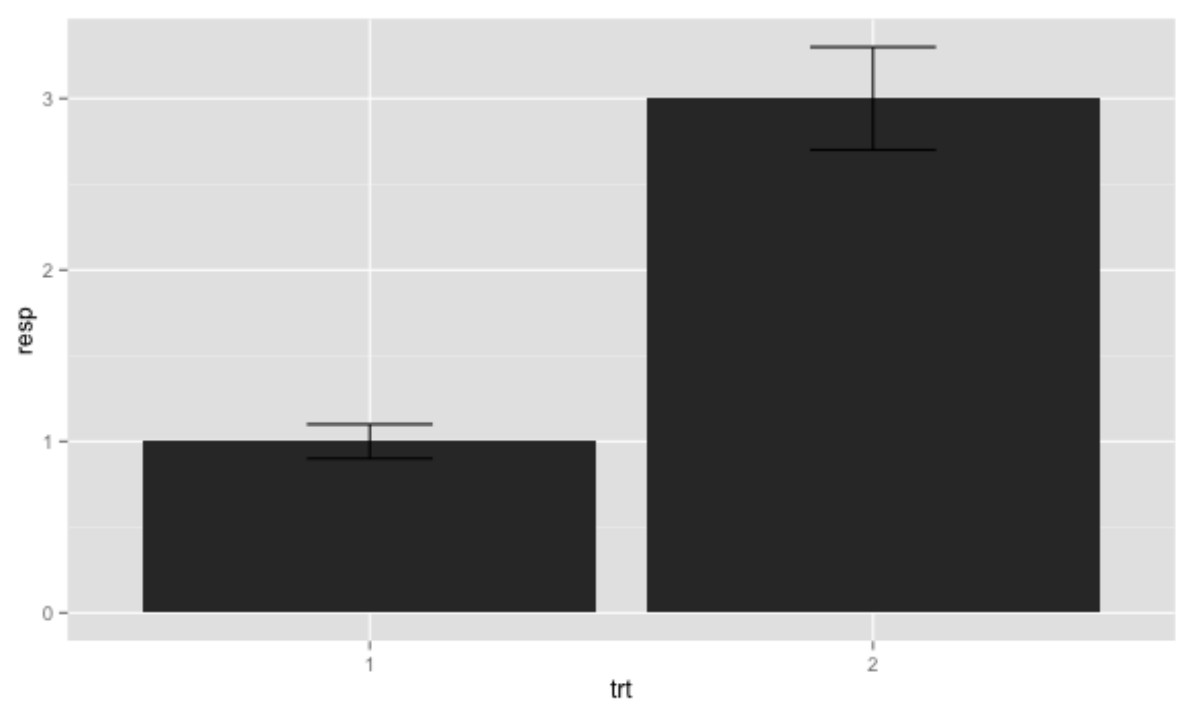



```
kdensity variable
```

Bar charts

Univariate visualization

Categorical data



Bivariate visualization

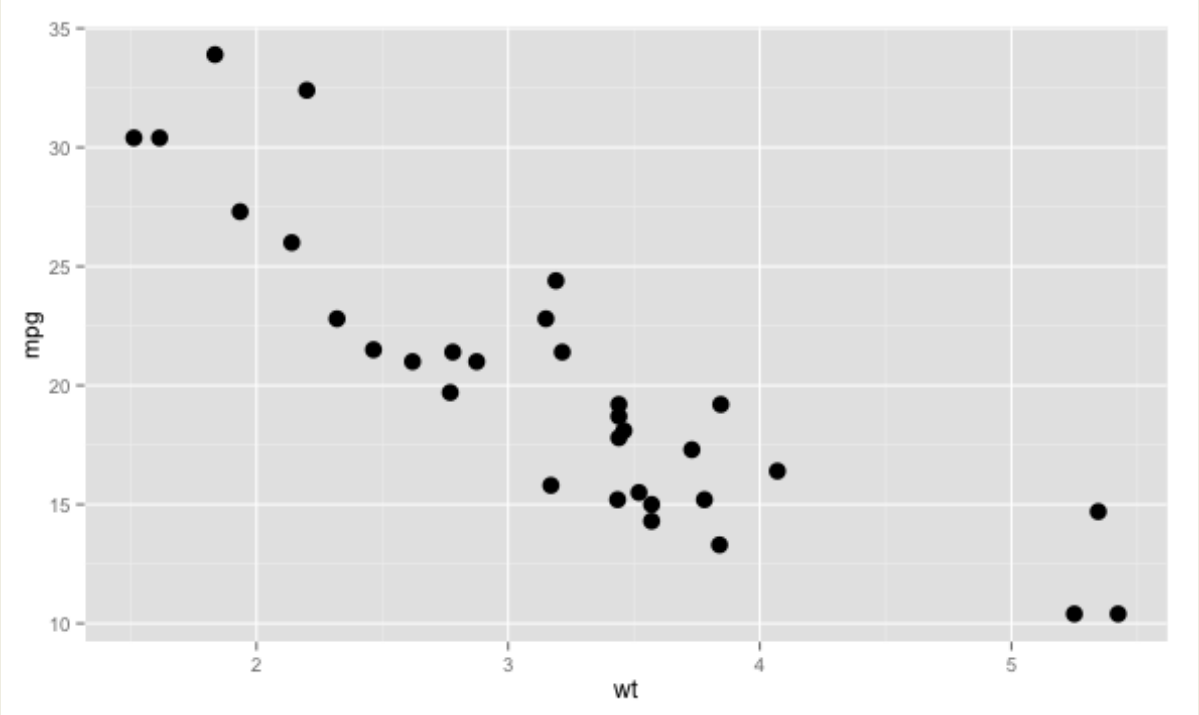
	Continuous	Categorical
Continuous	Scatterplots	Grouped plots
Categorical	—	Mosaic plots, grouped plots

Pack in as much data as you can

Scatterplots

Bivariate visualization

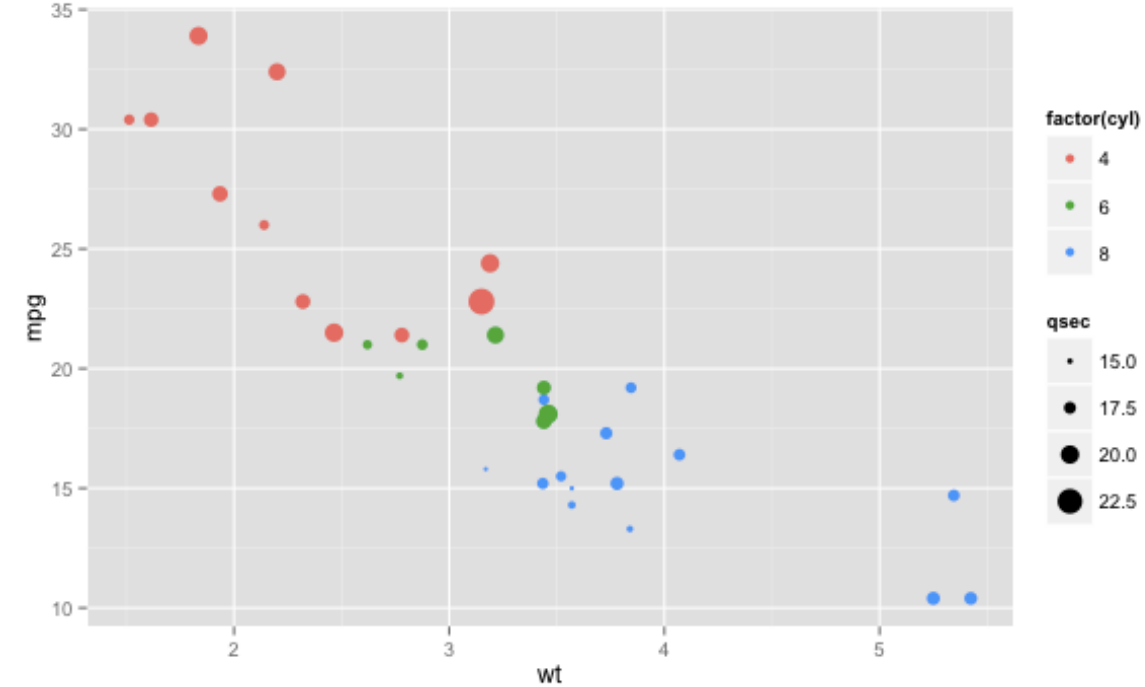
Continuous + continuous



MOAR DATA!!1!

Bivariate visualization

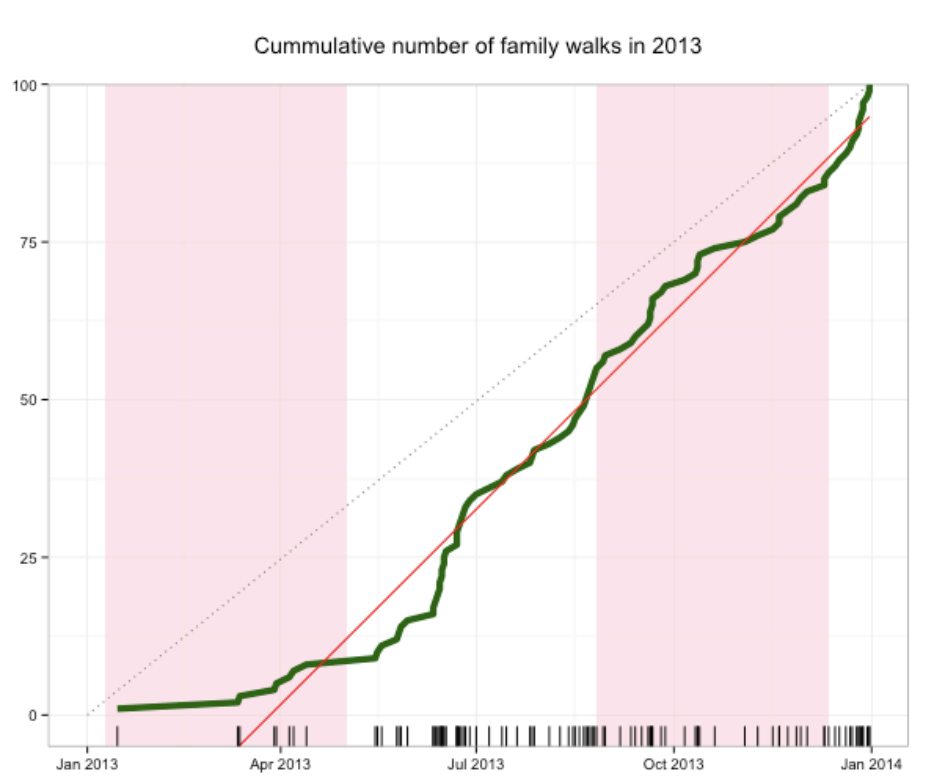
Continuous + continuous



Lines

Bivariate visualization

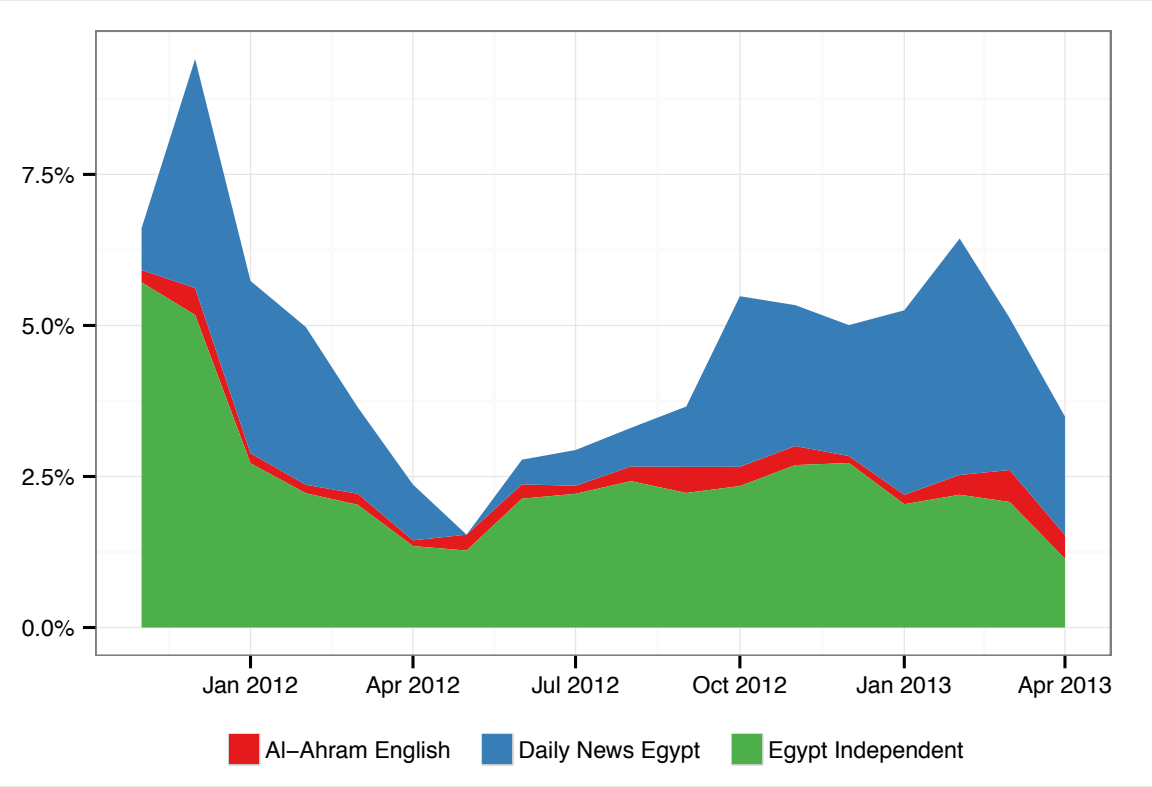
Continuous + continuous



Filled lines

Bivariate visualization

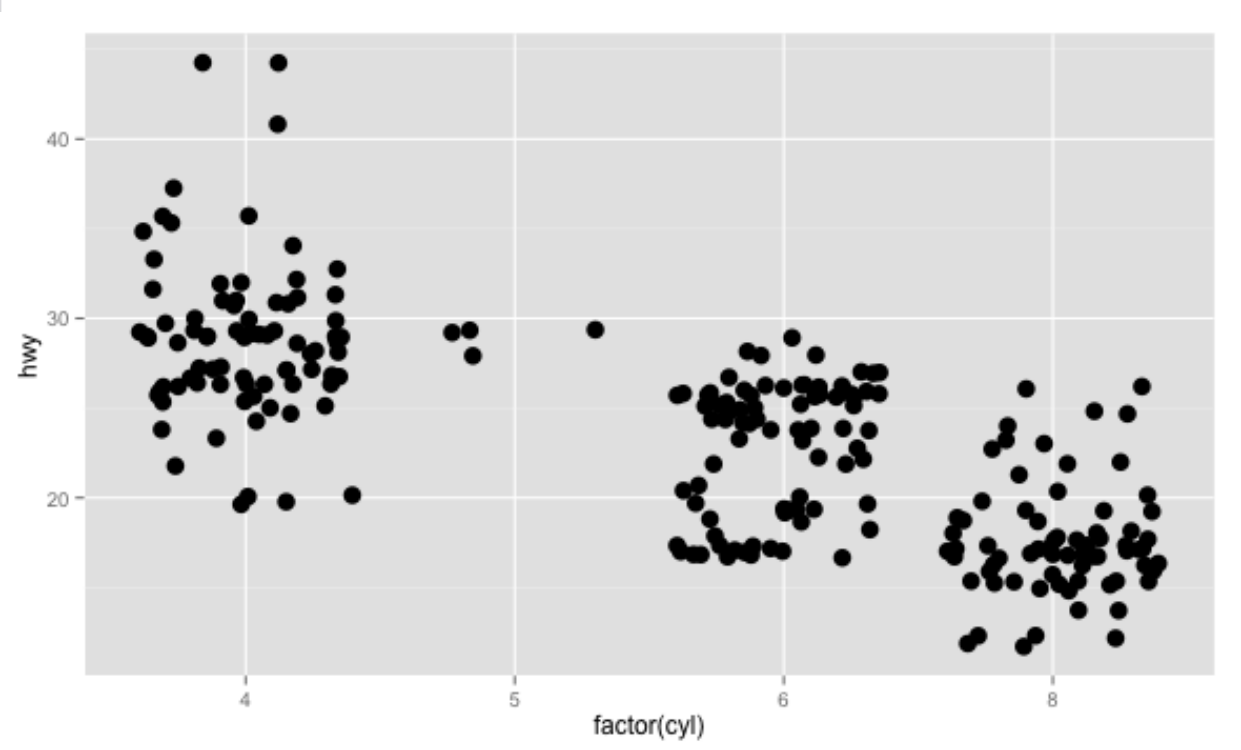
Continuous + continuous



Grouped points

Bivariate visualization

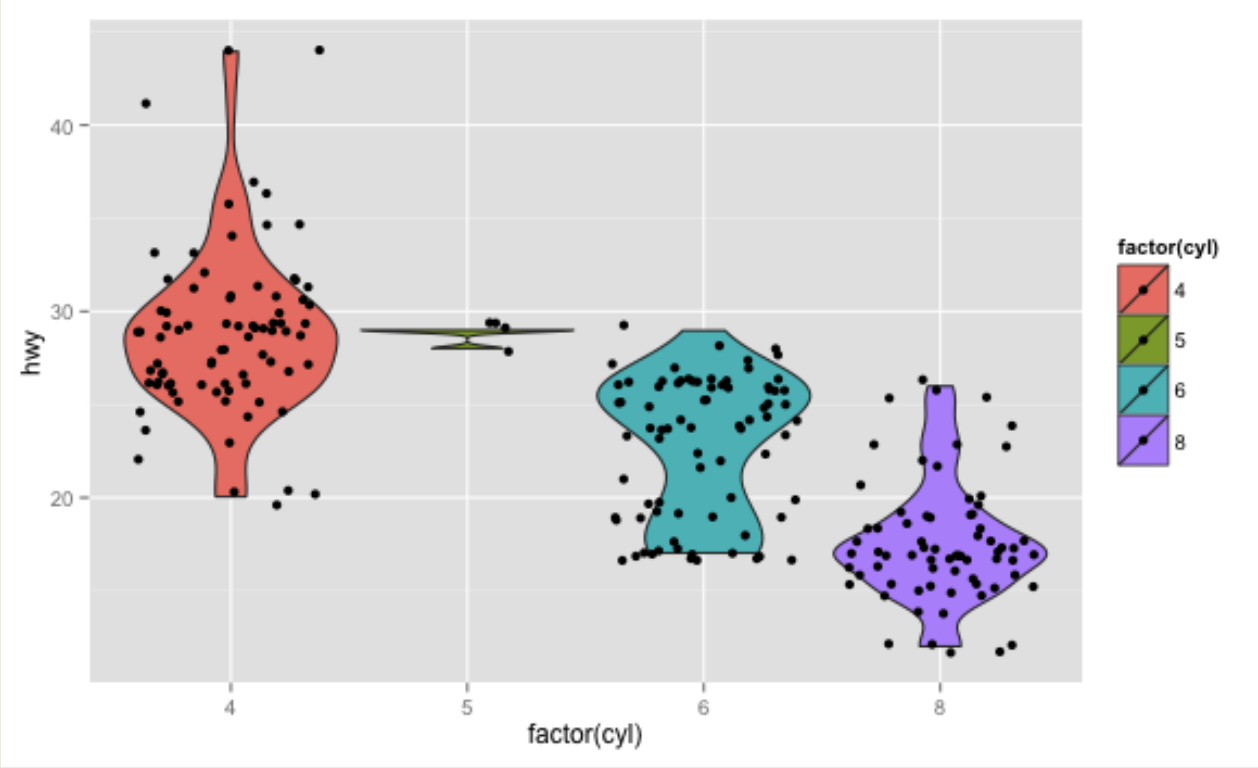
Continuous + categorical



Violin plots

Bivariate visualization

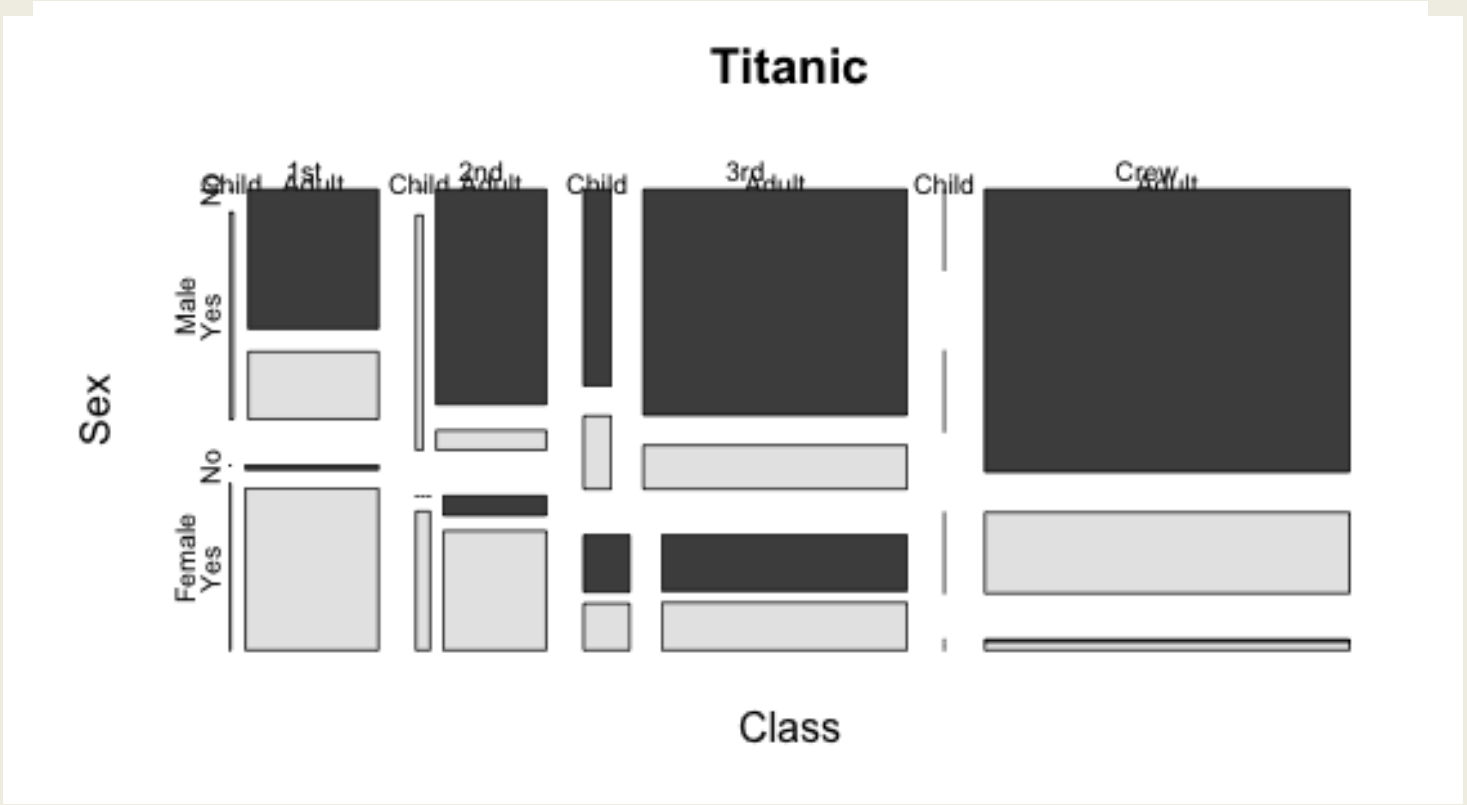
Continuous + categorical



Mosaic plots

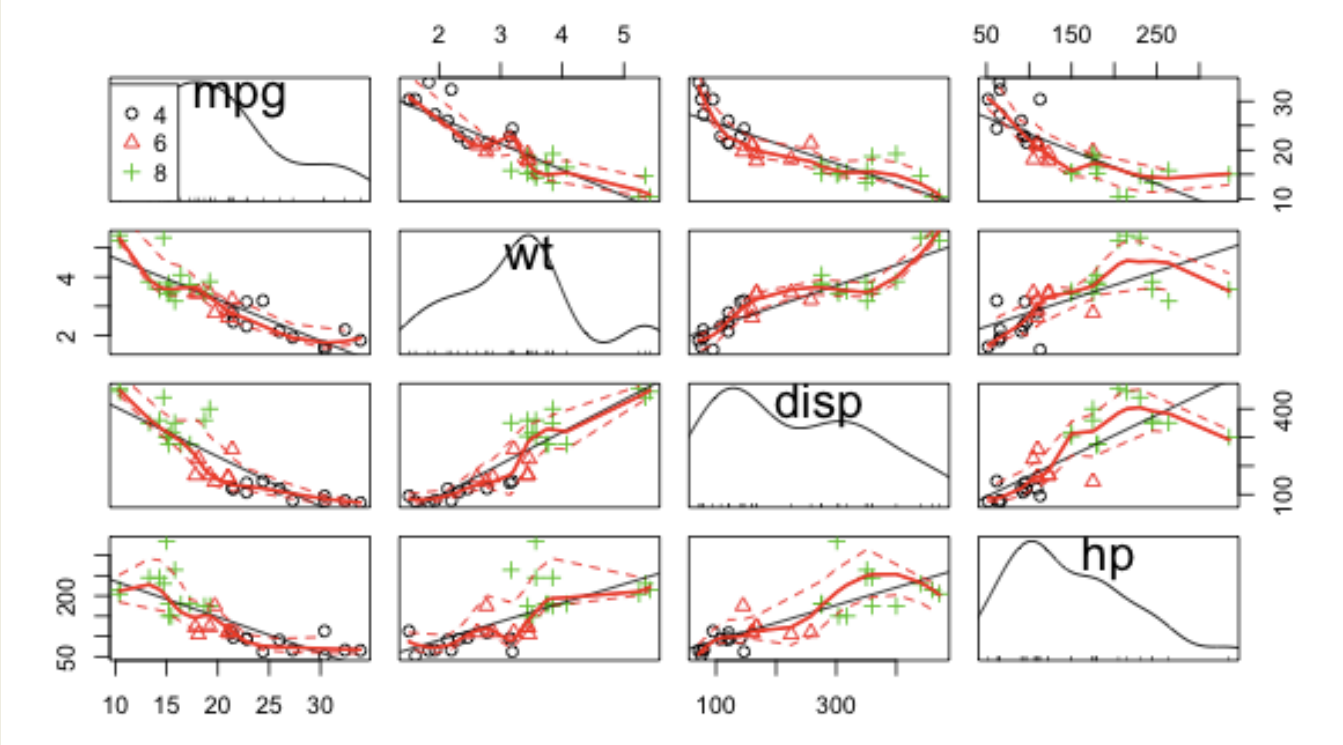
Bivariate visualization

Categorical + categorical



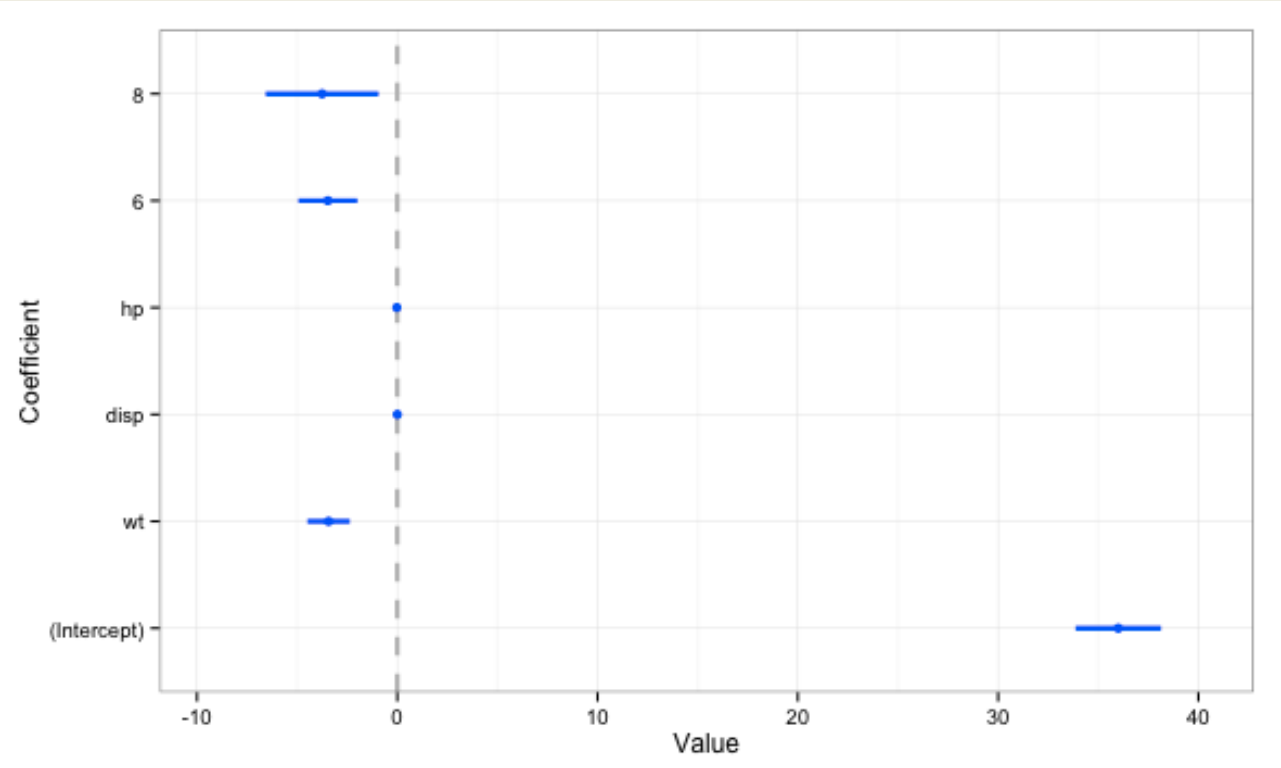
Scatterplot matrices

Model diagnostics | **Regression**



Coefficient plots

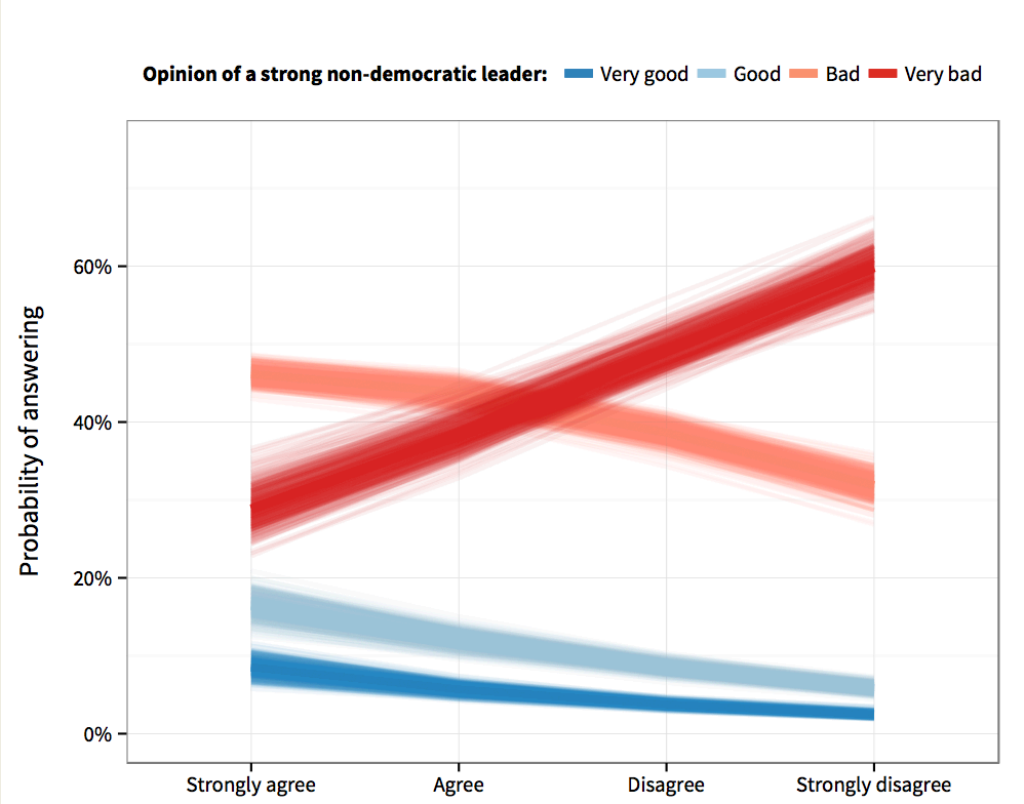
Model visualization | **Regression**



Predicted probabilities

Model visualization

Logit/Ologit



Other models

Diff-in-diff

(link to file)

Regression discontinuity

(link to file)

How do I do all this?

Stata

```
sysuse auto
graph box mpg, by(rep78, cols(2))

install ssc vioplot
vioplot mpg, over(rep78) horizontal

install ssc catplot
catplot rep78, by(foreign) percent(foreign)
```

R + ggplot2

How do I do all this?

Ask for help!

Go make
pretty pictures.