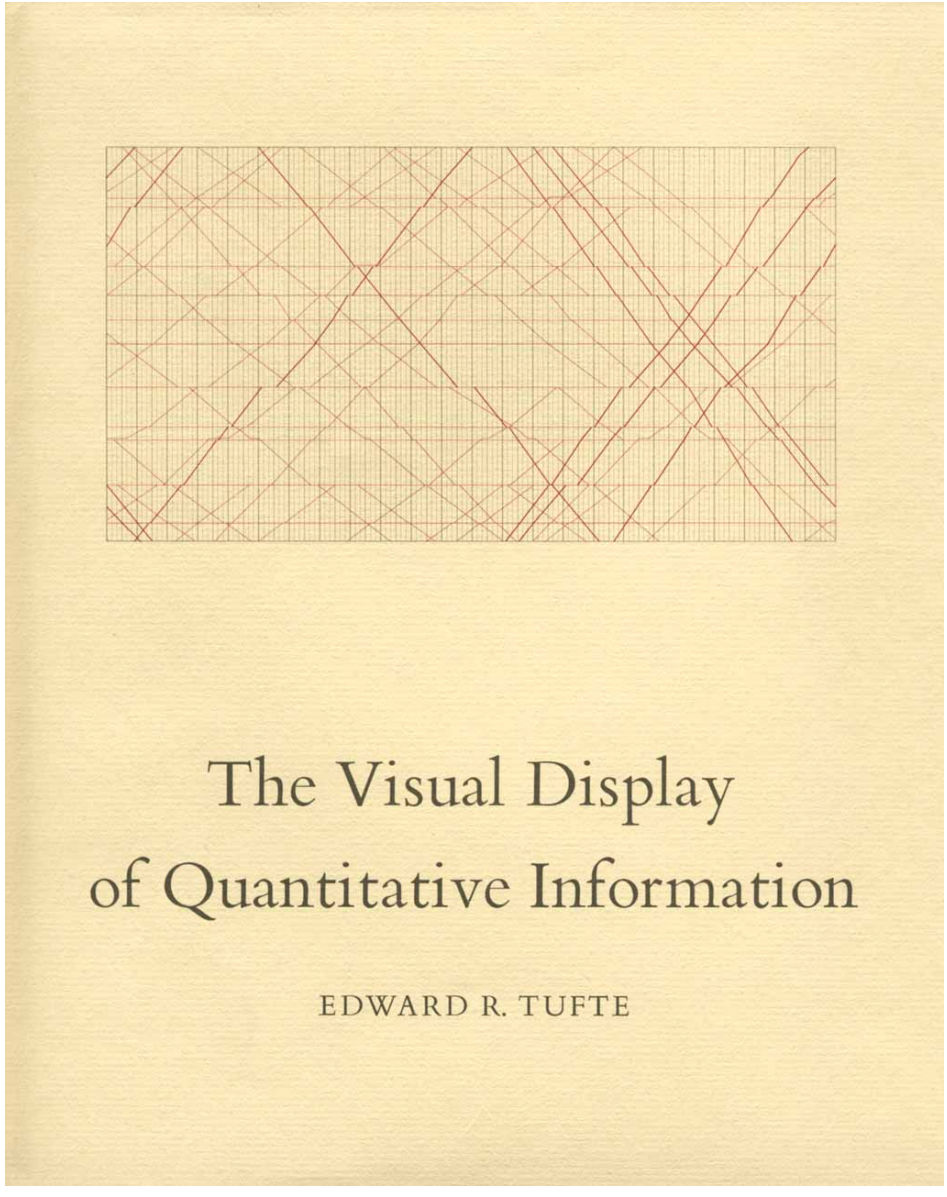
Visualization

Andrew Stokes

April 9, 2019

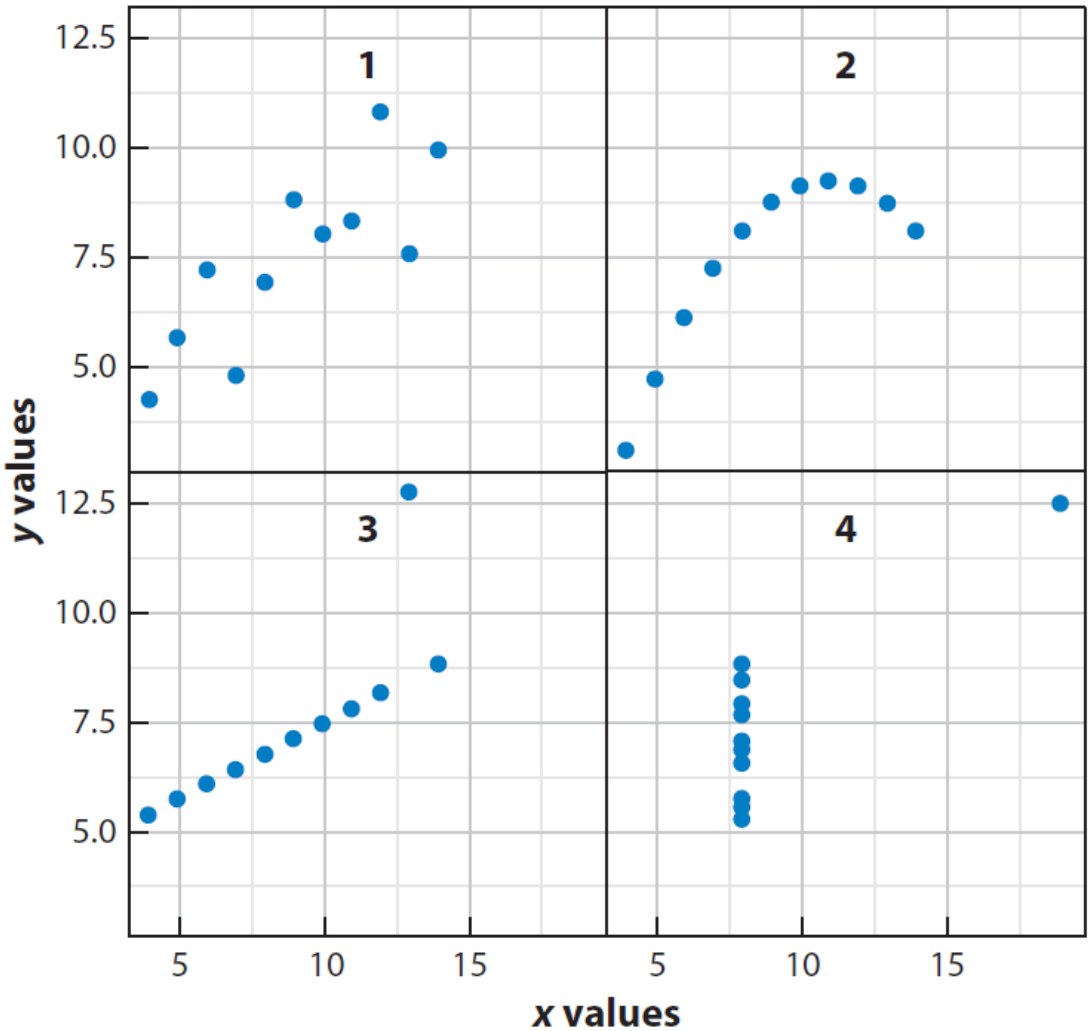
The classic text on visualization



Tufte on visualization

Graphical excellence is the well-designed presentation of interesting data a matter of substance, of statistics, and of design. . . . [It] consists of complex ideas communicated with clarity, precision, and efficiency. . . . [It] is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space. . . . [It] is nearly always multivariate. . . . And graphical excellence requires telling the truth about the data. (Tufte 1983, p. 51, via Healy and Moody 2014)

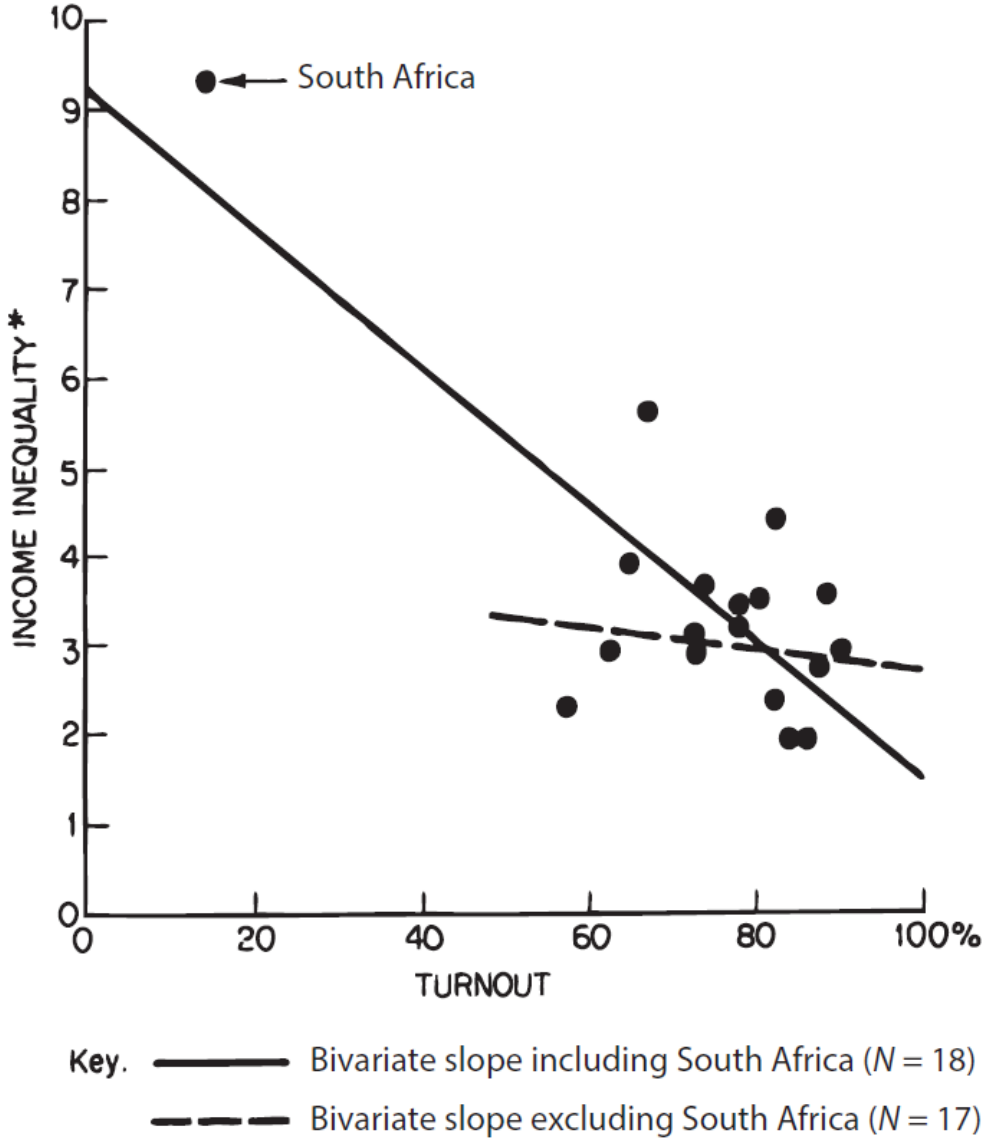
Example[1]



For all panels, $N = 11$; mean = 7.5; regression: $Y = 3 + 0.5(X)$; $r = 0.82$.
SE of slope estimate: 0.118, $t = 4.24$; sum of squares $(X - X)$: 100

[1] Healy & Moody 2014

Example[2]



[2]Healy & Moody 2014

Design Principles: Big Picture[3]

- Generate interest
- Provoke thought
- Motivate readers

Know your audience!

[3] Following slides adapted from Sullivan 2011

Design Principles: Specifics

- Clarity: titles, labels, axes
- Objectivity: fair scaling, appropriate comparisons
- Sound statistical practice
- Minimize “chart junk”: extraneous features that clutter tables & figures

When to use what

Text

- Appropriate for small amounts of quantitative data.
- Can be used when data are part of a sensitivity analysis.

Tables

- Many data points to present and values are important
- Useful for presenting main findings (readers will often to refer to tables before reading text)

When to use what (continued)

Figures

- Complex relationships among variables
- Trends over time
- Geographic variation
- Main findings (useful for disseminating results)

Presenting data in tables

- Consider your audience (technical experts vs lay persons)
- Consider context, time, place, situation
- Follow relevant style guidelines for papers, posters, reports

Components of a table

Table number. Table title.

Description of what follows	Column spanner
	Heading 1 Heading 2 . . . Heading x
Rows (variables and units)	Data

Example

Table 12-1. Association Between BMI categories and Incident Cardiovascular Disease After Adjustment for Clinical Risk Factors*

	Odds Ratio* (OR) (95% Confidence Interval)		
	Normal Weight	Overweight	Obese
Incident Myocardial Infarction	1.00 (Reference)	1.01 (0.69-1.29)	1.14 (1.01-1.50)
Incident Cardiovascular Disease	1.00 (Reference)	1.21 (0.89-1.37)	1.36 (1.13-2.54)
Incident Stroke	1.00 (Reference)	0.99 (0.82, 1.08)	1.18 (1.09-1.23)
*Note: Adjusted for age, sex, systolic and diastolic blood pressure, total serum cholesterol, high density lipoprotein and smoking; normal weight (body mass index (BMI) < 25.0), overweight ($25.0 \leq$ BMI < 30.0) and obese (BMI ≥ 30.0).			

Decluttered

Table 12-3. Association Between BMI categories and Incident Cardiovascular Disease After Adjustment for Clinical Risk Factors*

	Odds Ratio* (OR) (95% Confidence Interval)		
	Normal Weight	Overweight	Obese
Incident MI*	1.00 (Reference)	1.01 (0.69-1.29)	1.14 (1.01-1.50)
Incident CVD	1.00 (Reference)	1.21 (0.89-1.37)	1.36 (1.13-2.54)
Incident Stroke	1.00 (Reference)	0.99 (0.82, 1.08)	1.18 (1.09-1.23)

*Note: Adjusted for age, sex, systolic and diastolic blood pressure, total serum cholesterol, high density lipoprotein and smoking; normal weight (body mass index (BMI) < 25.0), overweight ($25.0 \leq$ BMI < 30.0) and obese (BMI ≥ 30.0); MI=Myocardial infarction; CVD=cardiovascular disease

Summarizing statistical results

Summary statistics

- provide measures of central tendency and variability for continuous variables
- n(%) for dichotomous, categorical and ordinal variables

Measures of effect

- provide estimates and standard errors or confidence limits

Example: Descriptive Statistics

Table 12-7. Background Characteristics of Study Participants by Intervention Group

	Intervention Group		
Characteristic*	Self-Help (n=100)	Group Therapy (n=90)	Individual Therapy (n=80)
Age, years	78.2 (6.2)	79.6 (5.9)	81.4 (5.7)
Male Sex, n (%)	46 (46%)	38 (42%)	28 (35%)
Education, years	9.3 (4.2)	10.7 (3.9)	8.6 (4.1)
Marital status			
Single, never married, n (%)	9 (9%)	11 (12%)	5 (6%)
Married or domestic partnership, n (%)	36 (36%)	36 (40%)	23 (29%)
Widowed, n (%)	43 (43%)	33 (37%)	43 (54%)
Divorced or separated, n (%)	12 (12%)	10 (11%)	9 (11%)

*Note: Means (standard deviations) are shown for continuous measures and n(%) are shown for categorical measures.

Example: Multivariable Results

Table 12-12. Association Between Racial/Ethnic Background, Maternal Age, Gestational Age and Birthweight

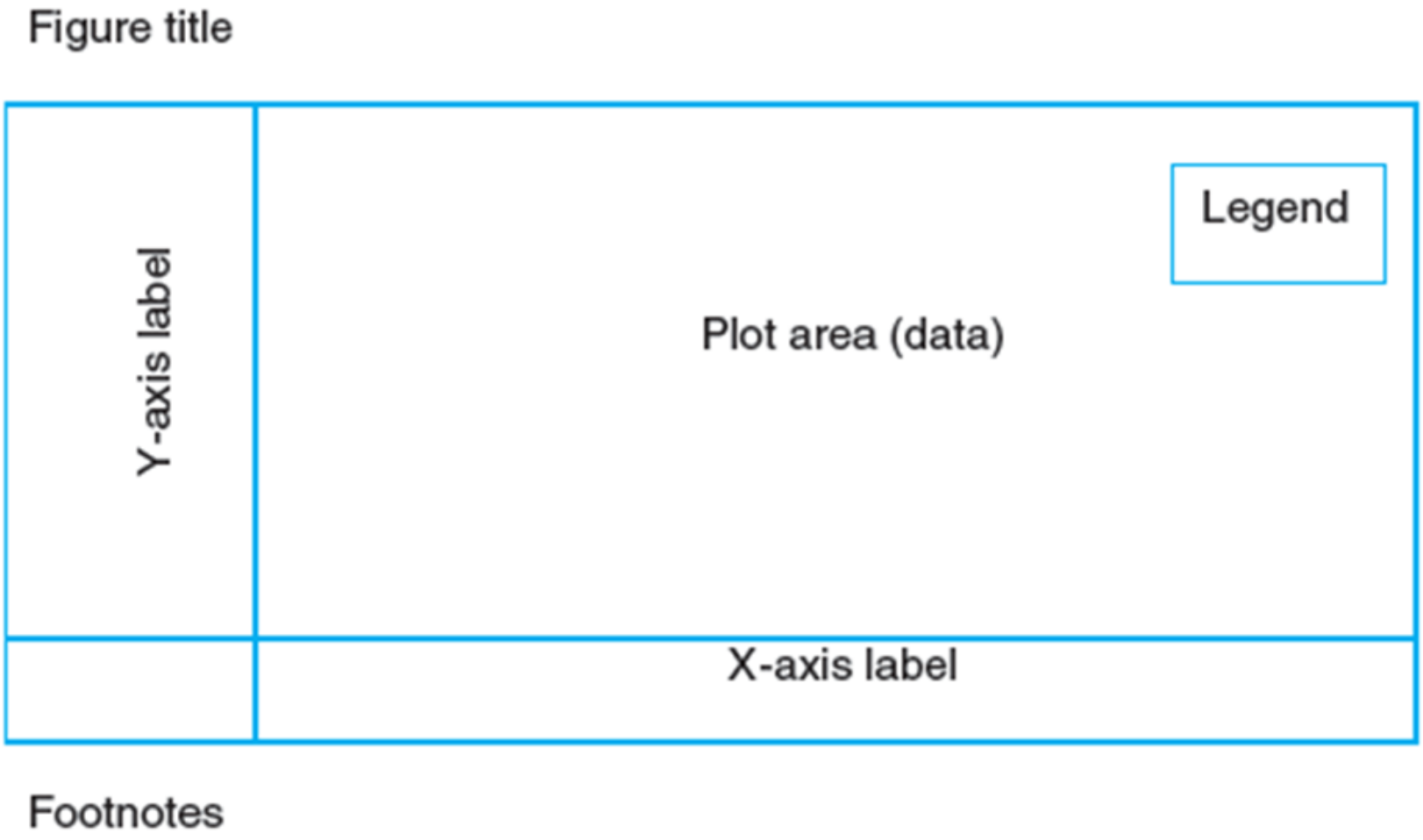
Characteristic	Regression Coefficient*	Standard Error	p-value
Intercept	-4366.5	188.3	<0.01
Racial/ethnic group			
White	Reference	-	
Black	-46.0	47.0	0.33
Hispanic	46.7	47.6	0.32
Maternal age, years	-0.27*	2.8	0.92
Gestational age, weeks	193.6*	4.7	<0.01

*Note: Regression coefficients are based on multiple linear regression analysis and are relative to a one year change in maternal age and one week change in gestational age.

Presenting data in figures

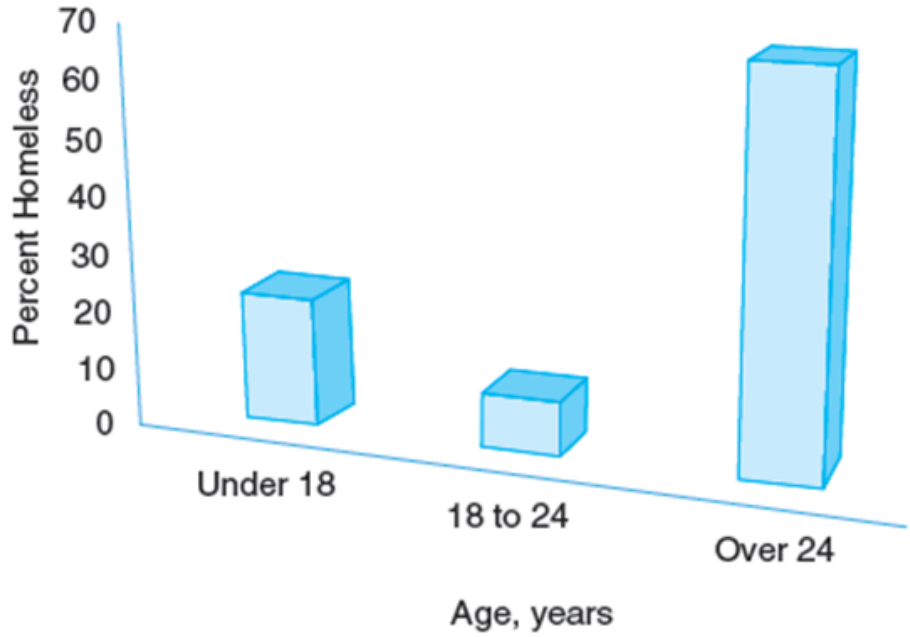
- Consider your audience (technical experts vs lay persons)
- Consider context, time, place, situation
- Follow relevant style guidelines for papers, posters, reports

Components of a figure



Example of "Chart Junk"

FIGURE 12-13 Homelessness in the United States by Age, 2015: 3D Chart Type

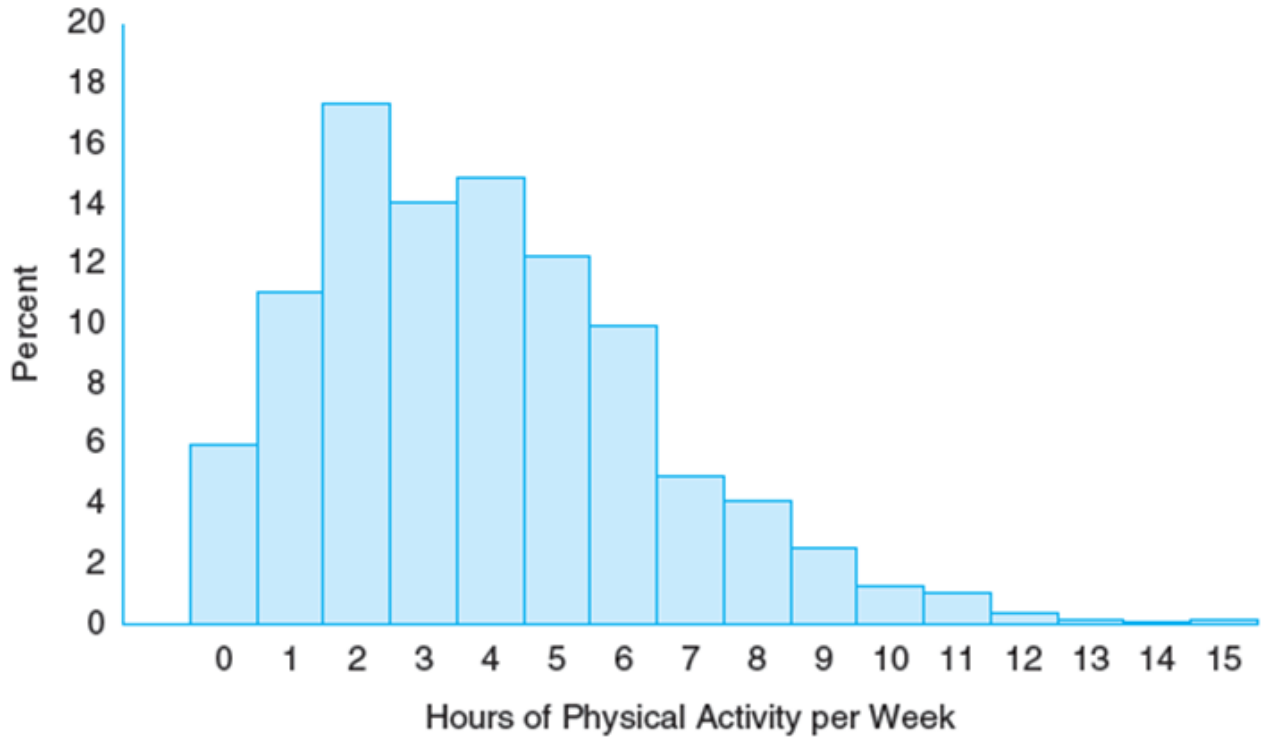


Data from US Department of Housing and Urban Development, Office of Community Planning and Development. Point-in-Time Estimates of Homelessness: The 2015 Annual Homeless Assessment Report (HAR) to Congress. https://www.hudexchange.info/resources/documents/2015-AHAR-Part-1.pdf.

Displaying data and distributions

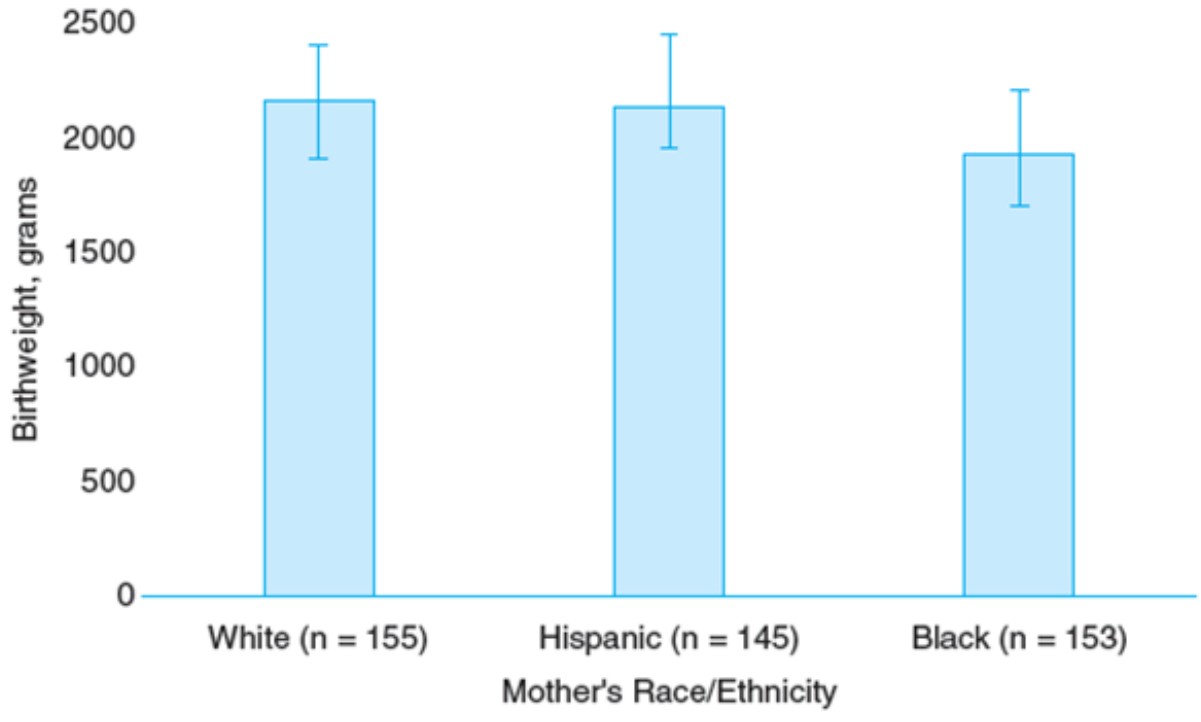
- Histograms for ordinal variables, with clear title, axis labels

FIGURE 12-23 Hours of Physical Activity per Week

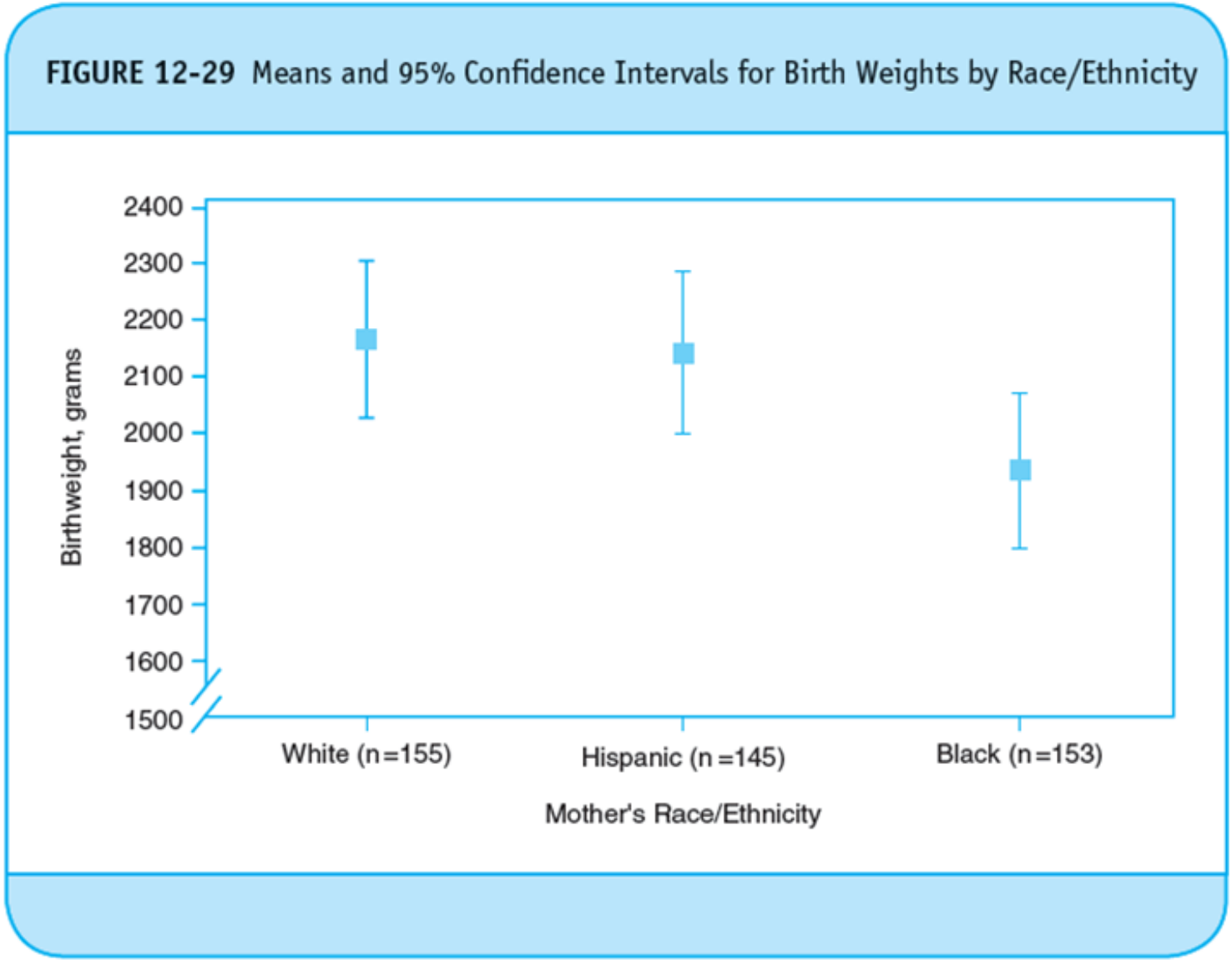


Displaying means with uncertainty: Example

FIGURE 12-28 Means and 95% Confidence Intervals for Birth Weights by Race/Ethnicity

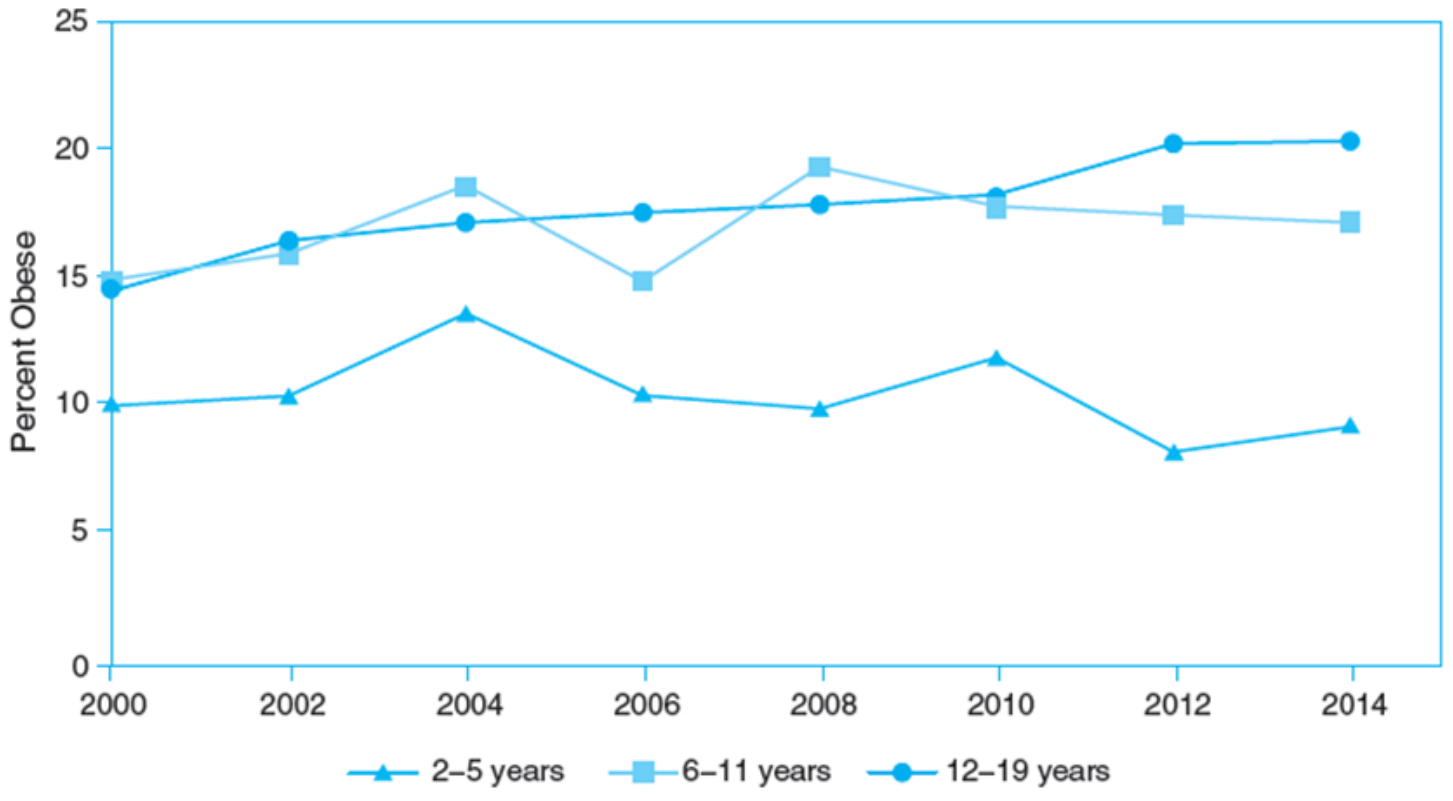


Displaying means with uncertainty: Better example



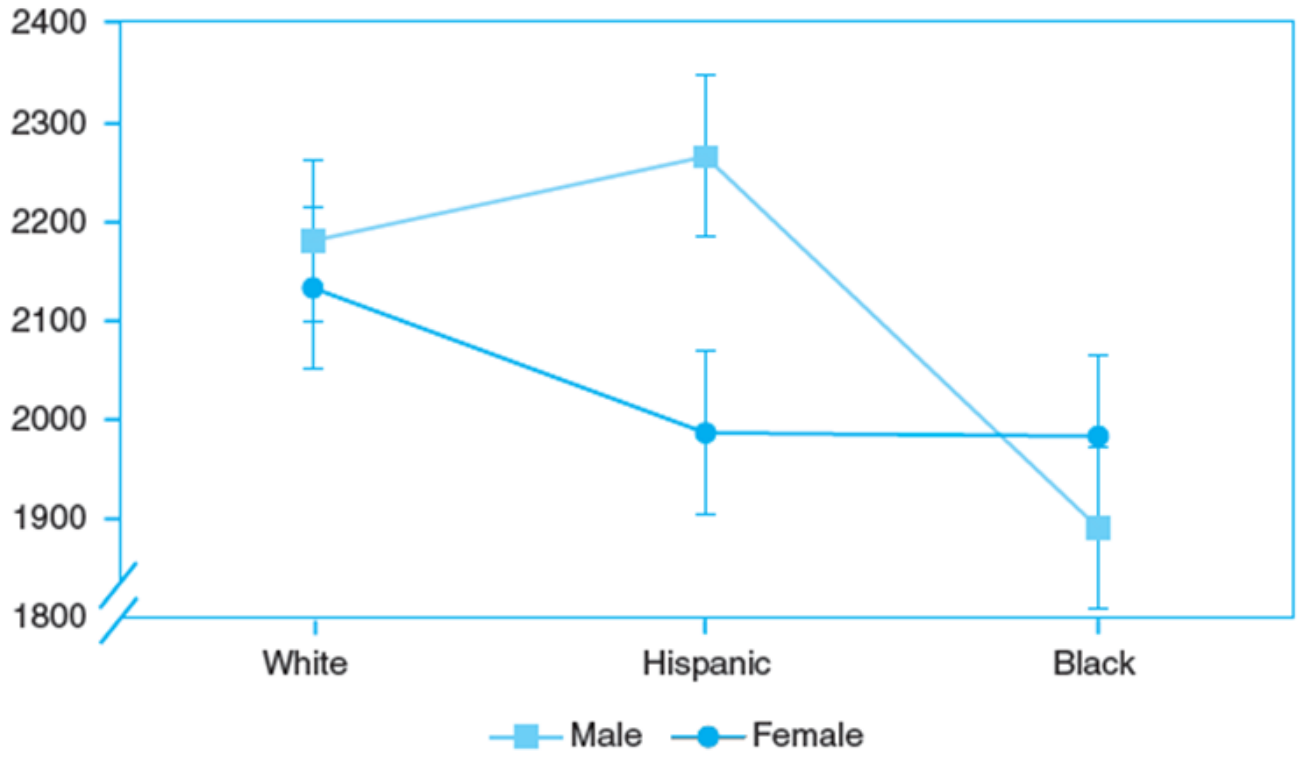
Trends & Line Charts

FIGURE 12-35 Prevalence of Obesity Among Children and Adolescents in the United States, 2000–2014

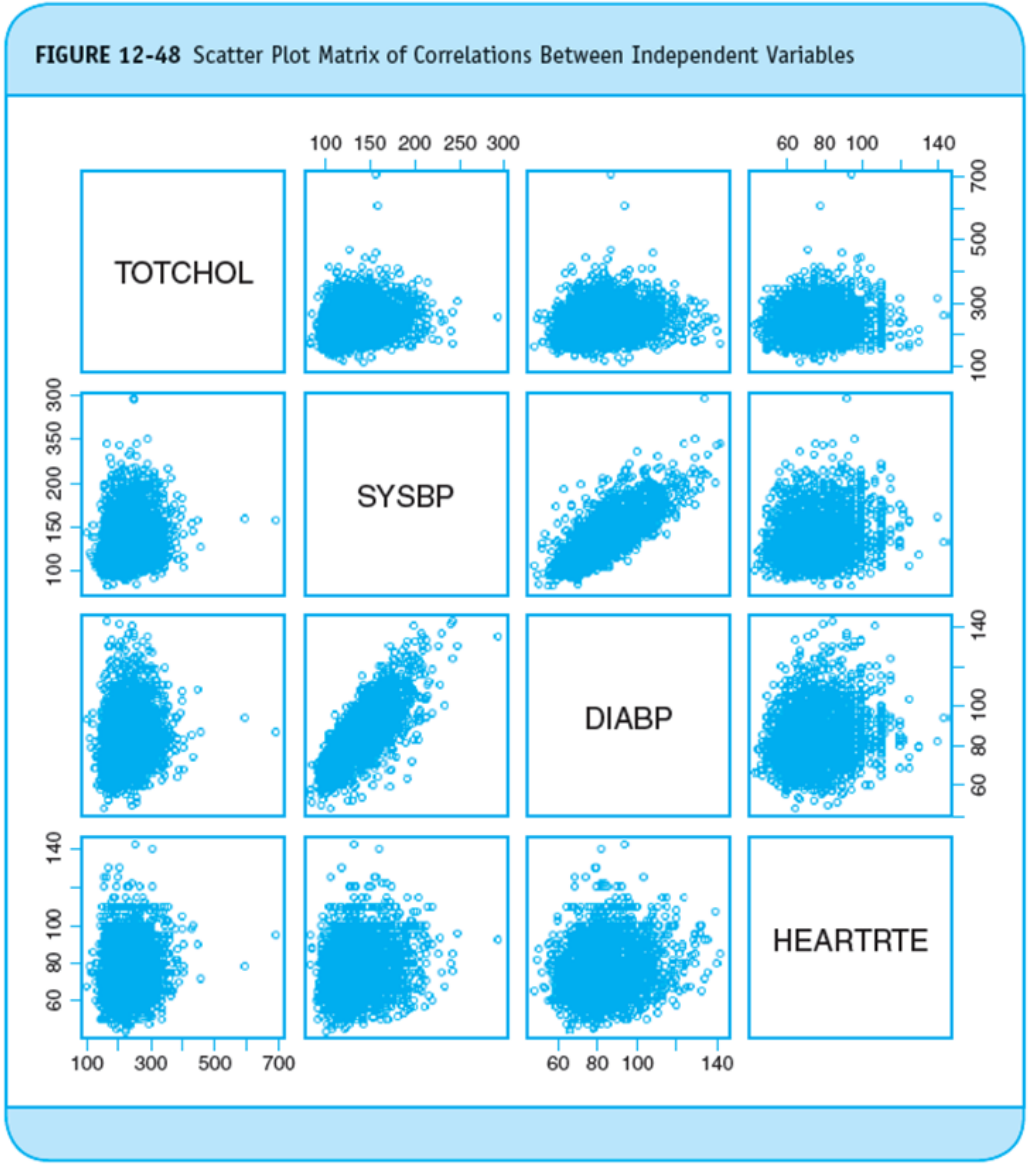


Incorrect use of trend lines

FIGURE 12-34 Mean and 95% Confidence Intervals for Birth Weights by Race/Ethnicity and Infant Sex: Incorrect Use of Trend Lines

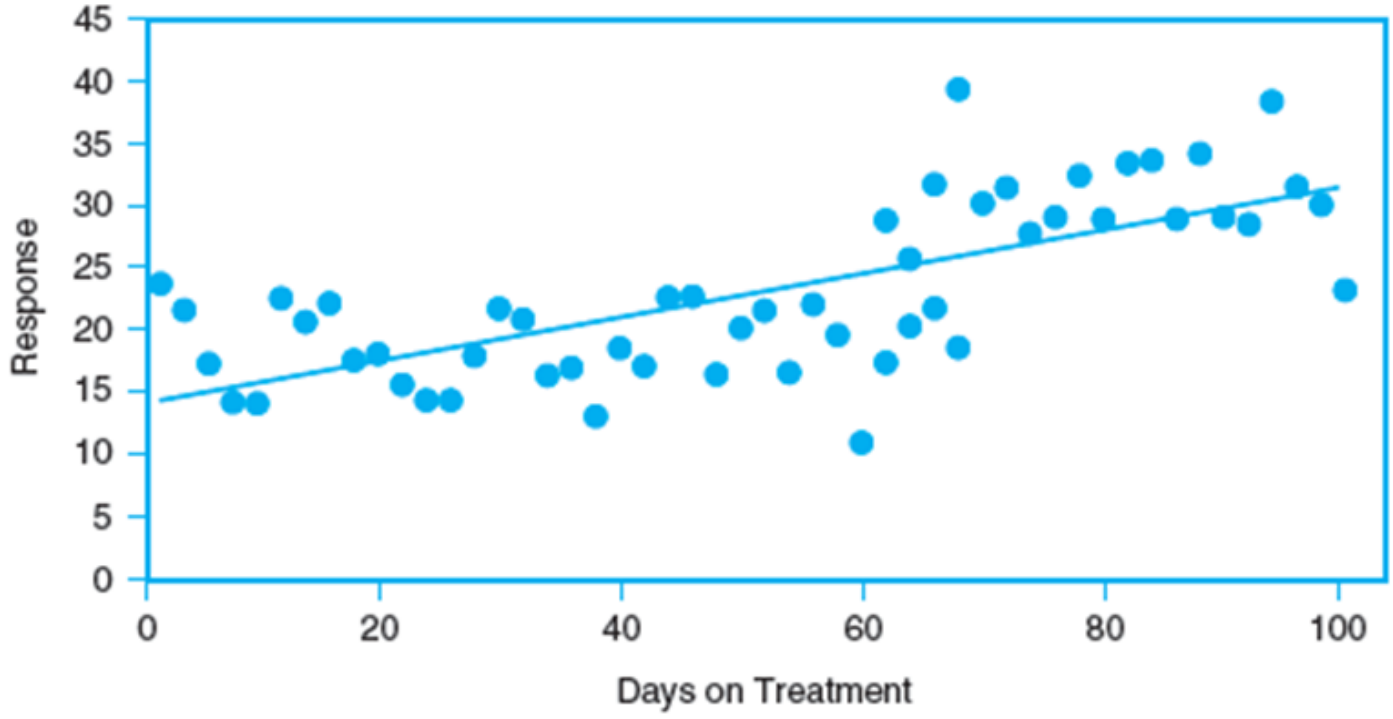


Relationships between continuous variables



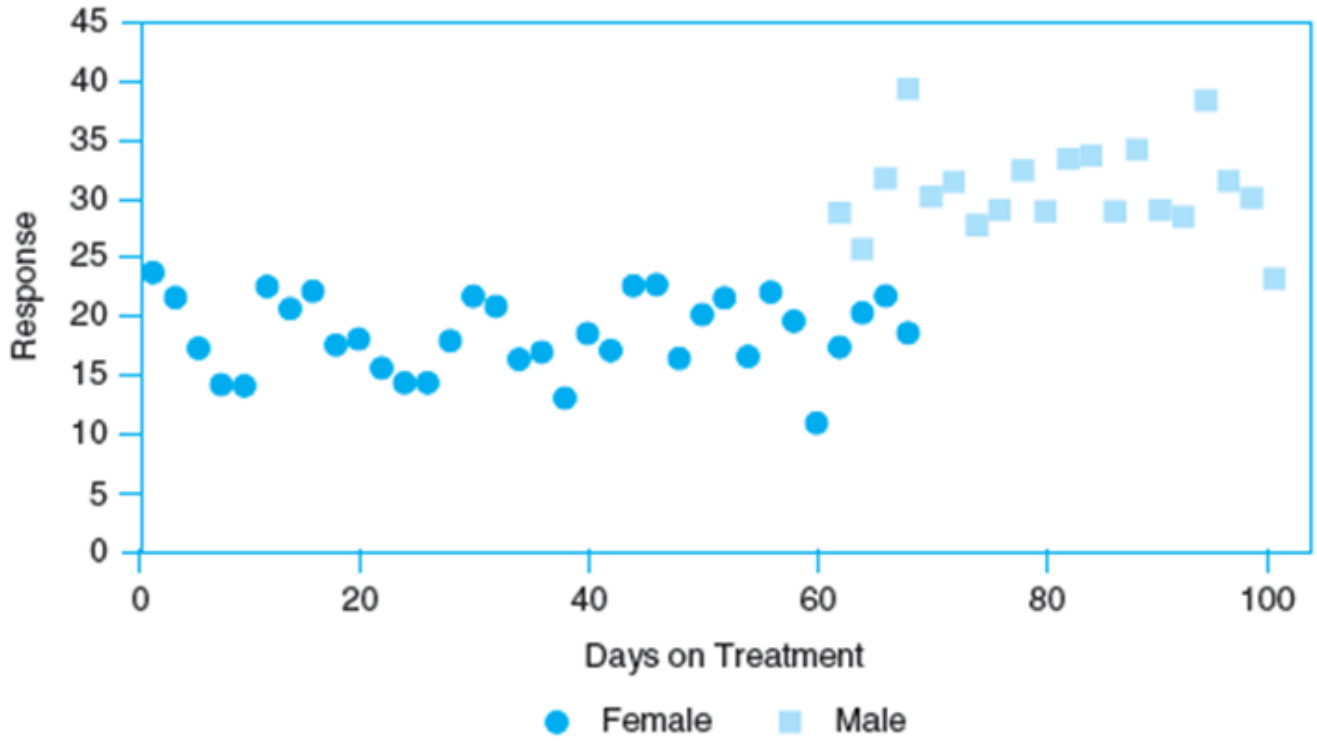
Scatter between two variables

FIGURE 12-43 Regression of Response on Days on Treatment (Total Sample)



Use scatter to explore heterogeneity

FIGURE 12-44 Association Between Days on Treatment and Responses in Men and Women

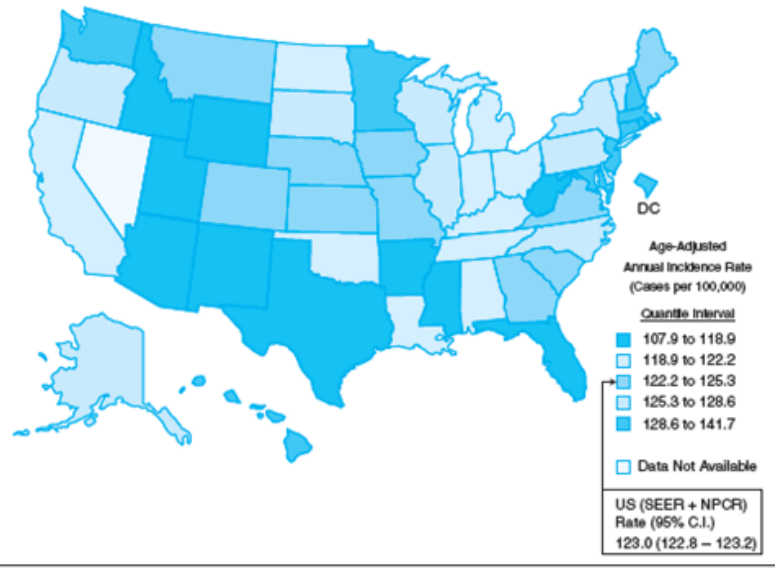


Geographic variation

- Using maps to display distributions of key health indicators
- Must include clear titles, description of measures, and clarity of geographic subunit of interest
- Chloropleth maps use shading to reflect the magnitude of measures

Geographic variation: Example

FIGURE 12-51 Variation in Breast Cancer Incidence by State, 2008–2012



Notes:
Created by statecancerprofiles.cancer.gov on 06/27/2016 2:37 pm.
Data for the United States does not include data from Nevada.
State Cancer Registries may provide more current or more local data.
Data presented on the State Cancer Profiles Web Site may differ from statistics reported by the State Cancer Registries (for more information).

† Incidence rates (cases per 100,000 population per year) are age-adjusted to the 2000 US standard population (19 age group: <1, 1–4, 5–9,..., 80–84, 85+). Rates are for invasive cancer only (except for bladder which is invasive and in situ) or unless otherwise specified, Rates calculated using SEER*Stat. Population counts for denominators are based on Census populations as modified by NCI.

◊ The 1969–2013 US Populatiuon Data File is used for SEER and NPCR incidence rates.
Data not available for this combination of geography, statistic, age and race/ethnicity.

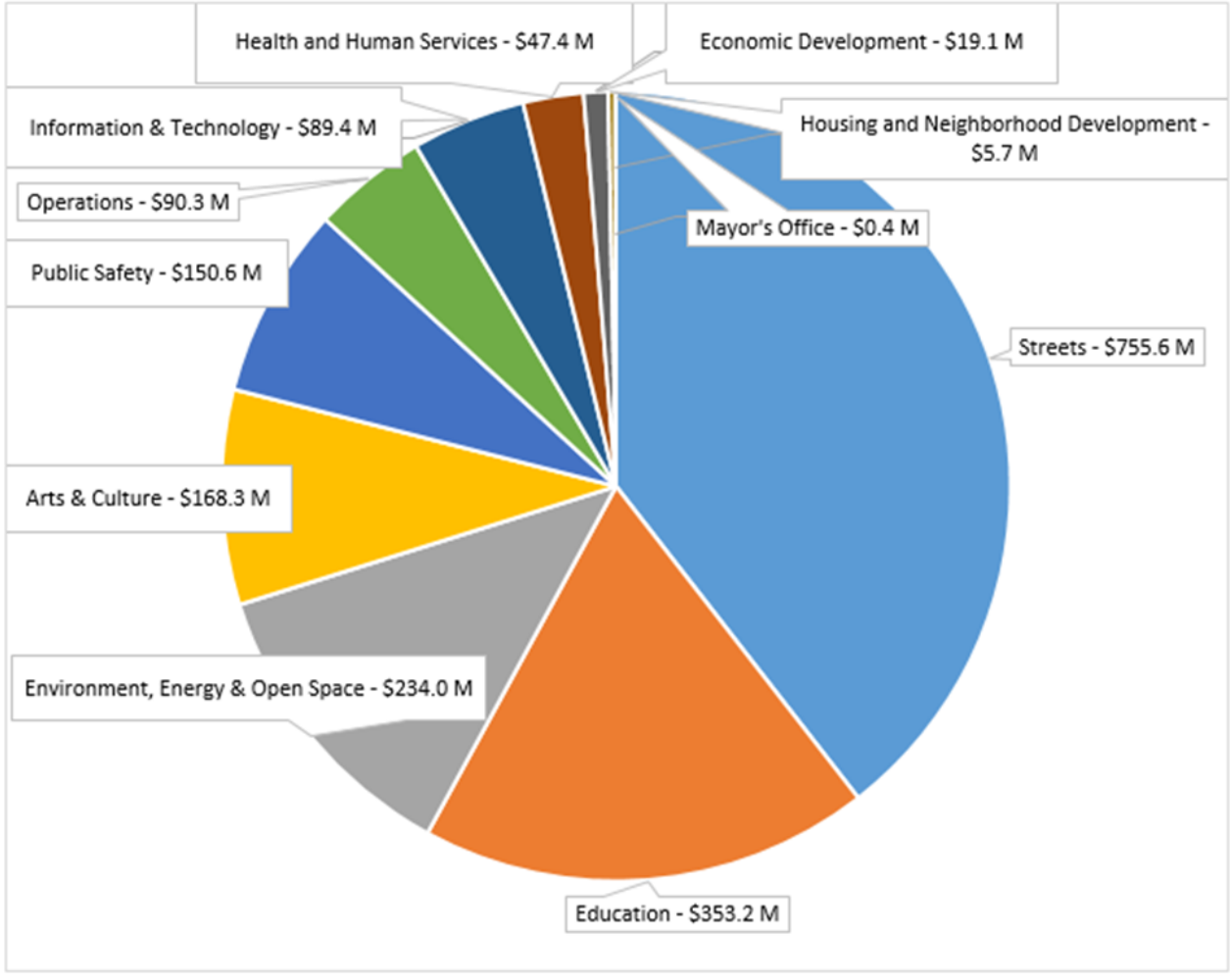
Reproduced from National Cancer Institute. State Cancer Profiles. http://statecancerprofiles.cancer.gov/map/map.withimage.php?00&001&055&00&2&01&0&1&5&0#results.

Pie Charts

- Popular displays to represent component parts of whole
- Can be challenging for readers to interpret
- Should be use sparingly, if at all (other displays often more effective)

Pie Chart or Table?

Figure 12-54. Capital Budget for the City of Boston Fiscal Year 2017 - Using a Pie Chart



Source: City of Boston Open Budget Application, http://budget.data.cityofboston.gov/#/

Pie Chart or Table (continued)?

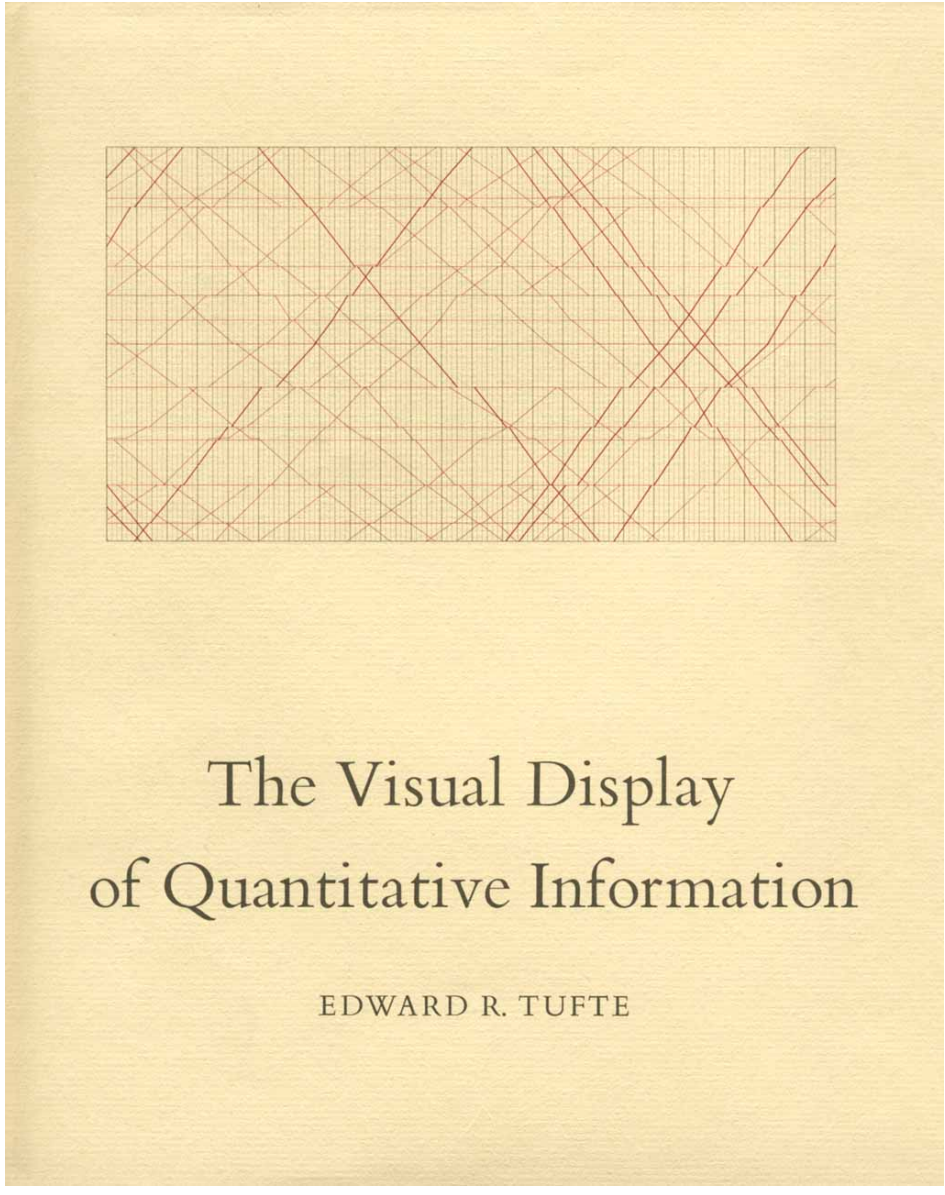
Table 12-17. Capital Budget for the City of Boston Fiscal Year 2017

Budget Category	Millions	% of Total
Streets	$755.6	39%
Education	$353.2	18%
Environment, Energy & Open Space	$234.0	12%
Arts & Culture	$168.3	9%
Public Safety	$150.6	8%
Operations	$90.3	5%
Information & Technology	$89.4	5%
Health and Human Services	$47.4	2%
Economic Development	$19.1	1%
Housing and Neighborhood Development	$5.7	0%
Mayor's Office	$0.4	0%
Total capital budget	$1,914.0	100%

Summary

- The right approach to present data and statistical results depends on the audience and the nature of the data and statistical results to be displayed
- Effective communication requires clarity and accuracy
- Must adhere to sound statistical practice and effective design principles

A valuable resource

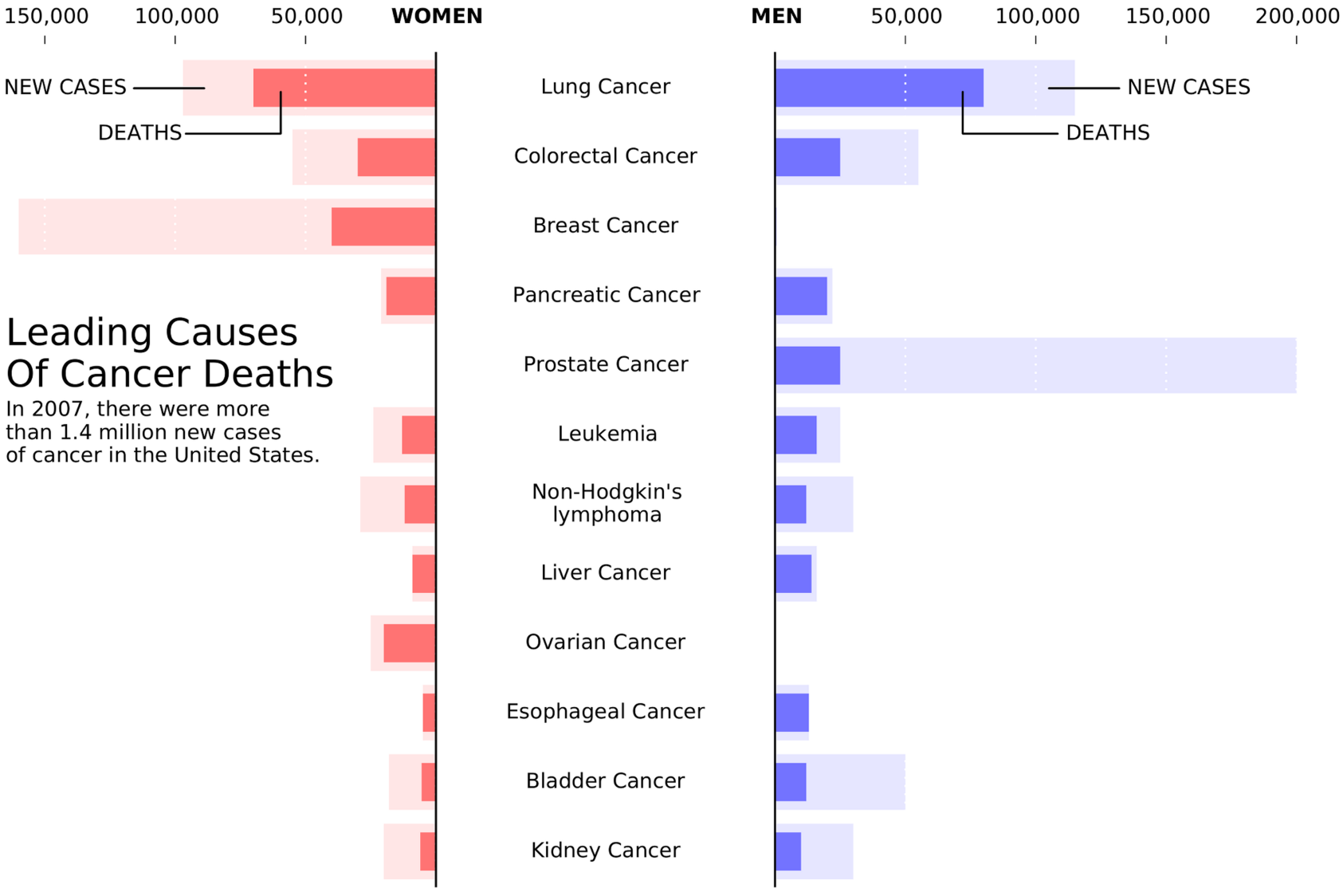


Simple rules for making compelling visualizations[4]

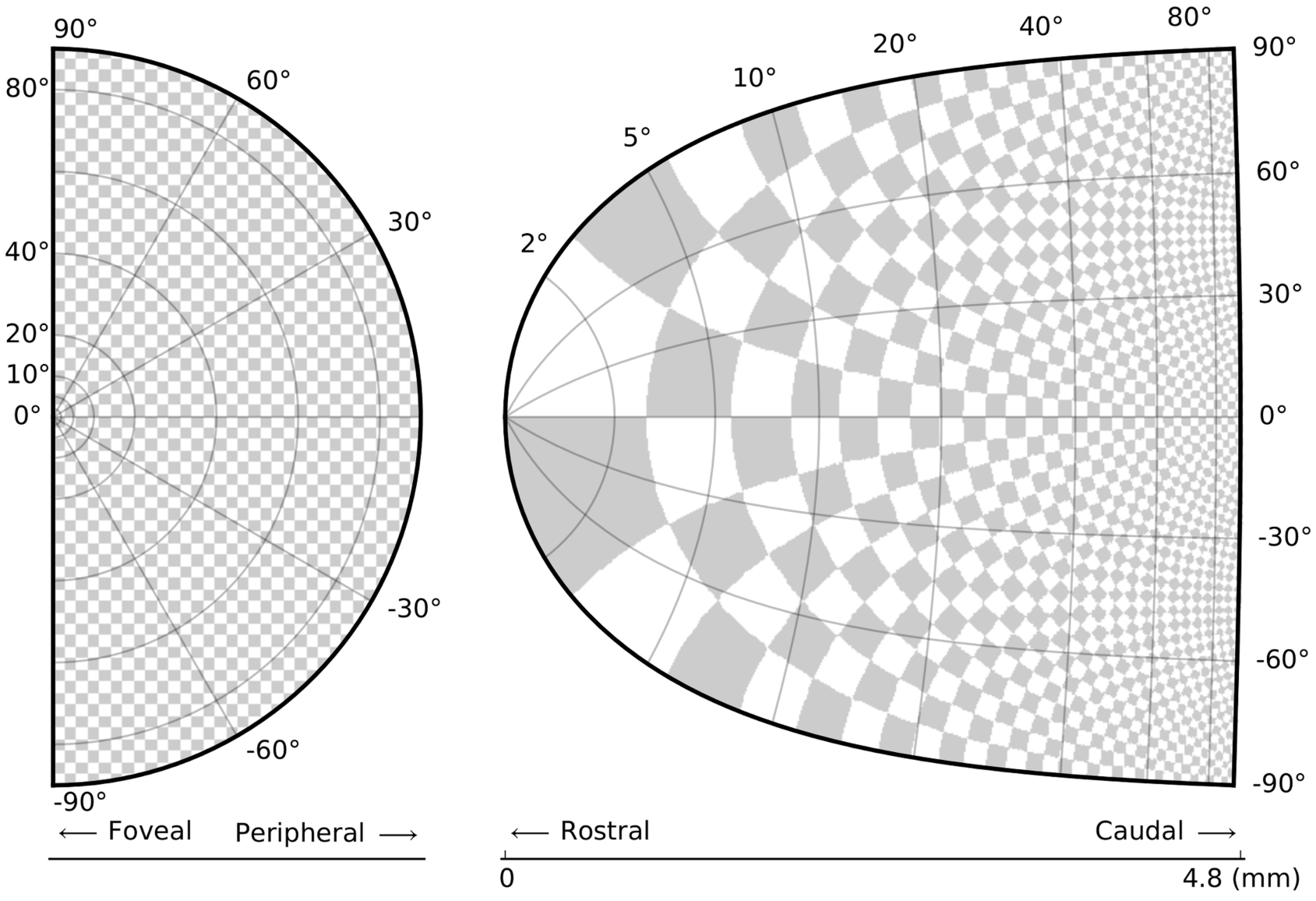
- Consider the audience
- Identify the key message
- Consider the medium
- Use captions to make graph free standing
- Defaults aren't always best
- Use color to your advantage
- Be honest
- Keep it simple
- Message is more important than aesthetics

[4] From Rougier, Droettboom & Bourne 2014

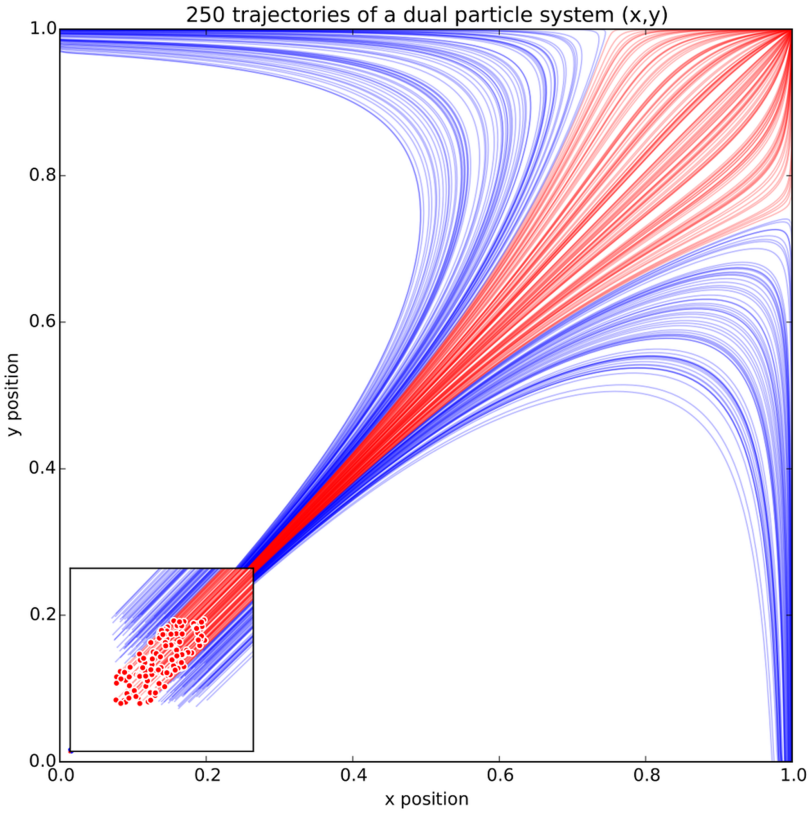
Consider the audience

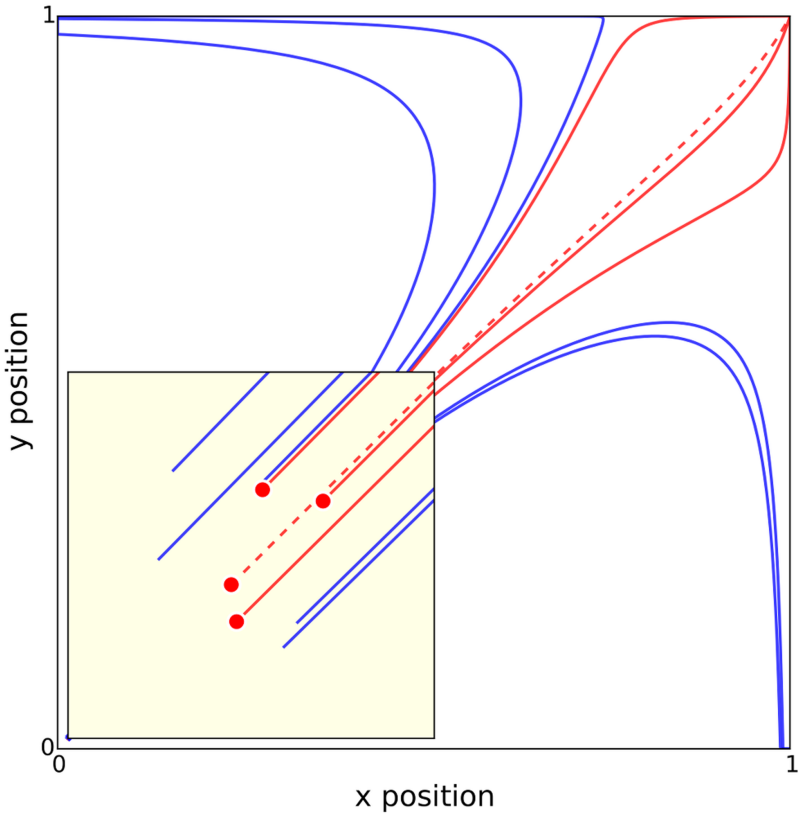


Identify the key message

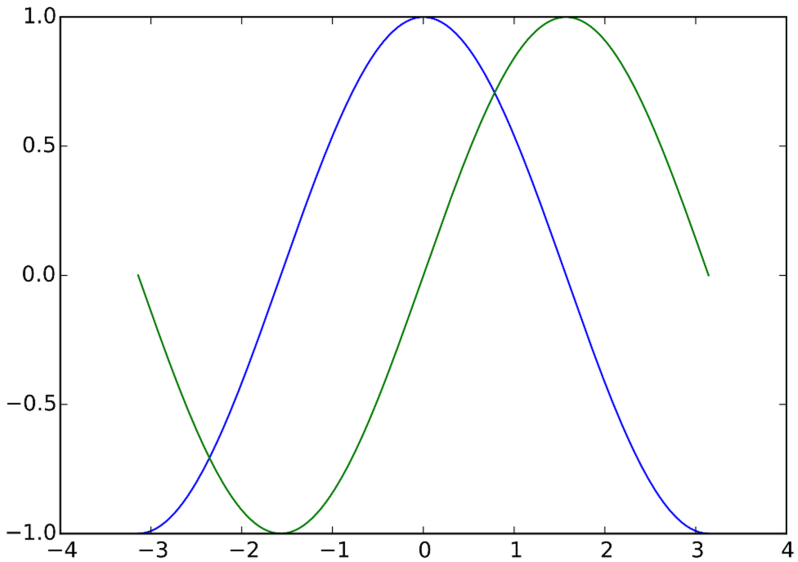


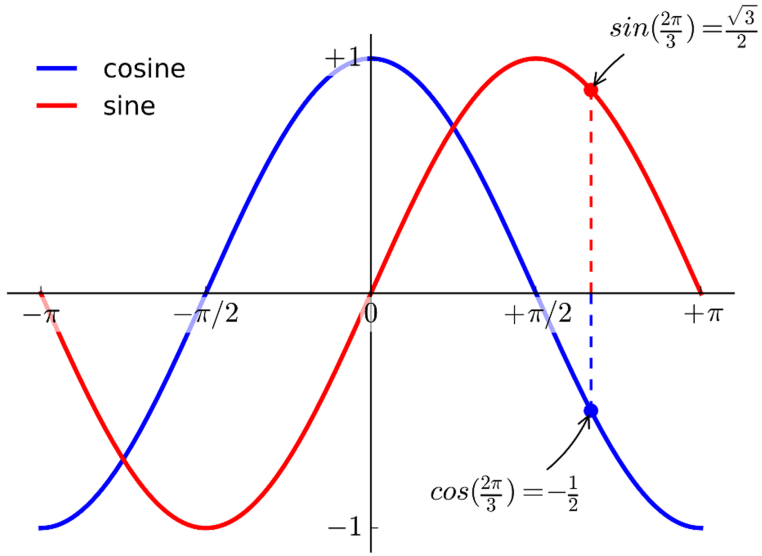
Consider the medium



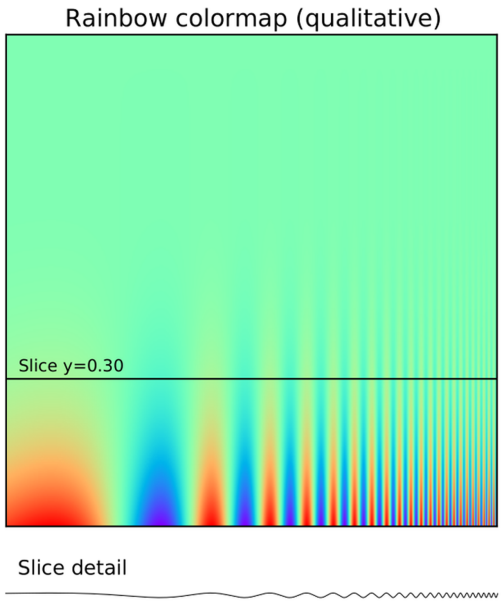


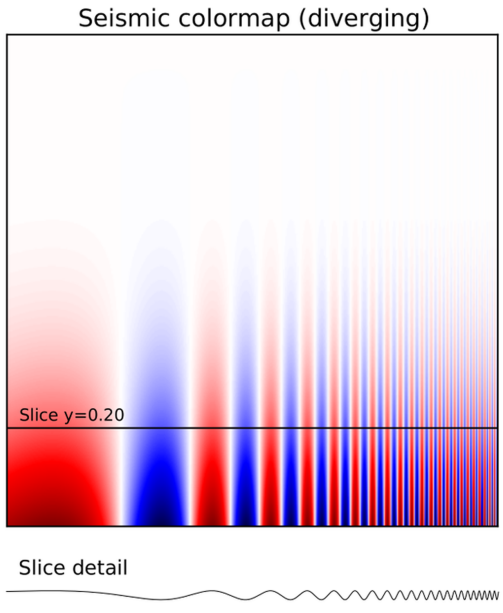
Defaults aren't always best

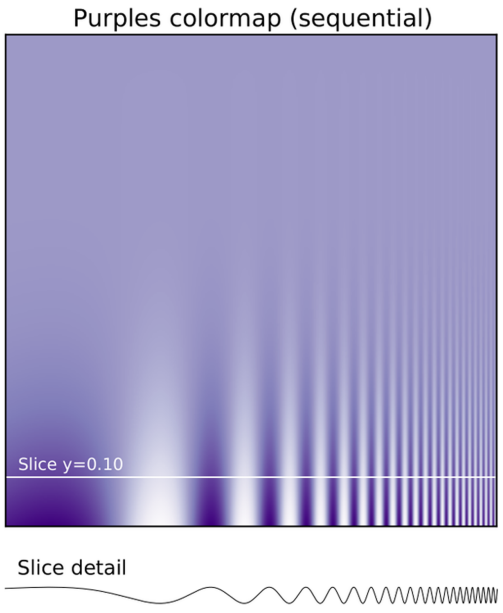




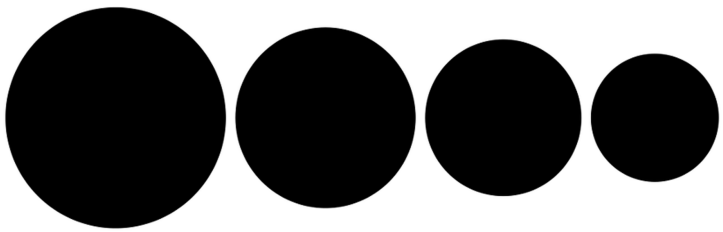
Use color to your advantage





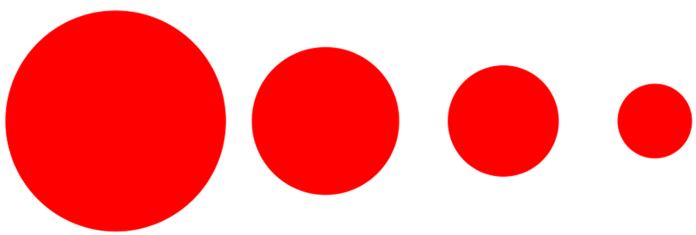


Be honest



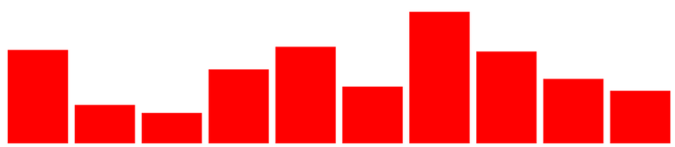
Relative size using disc area

Relative size using disc radius

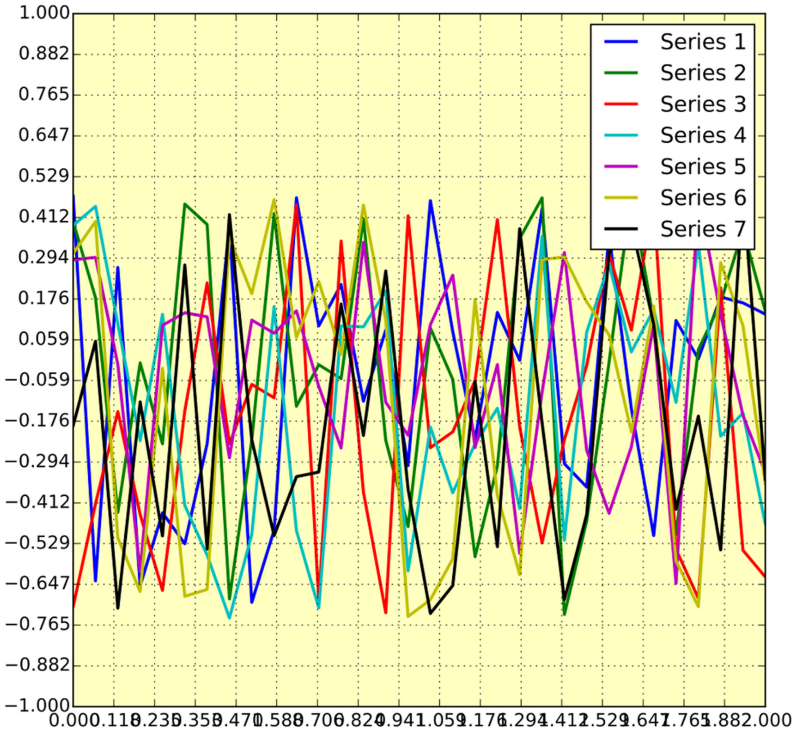


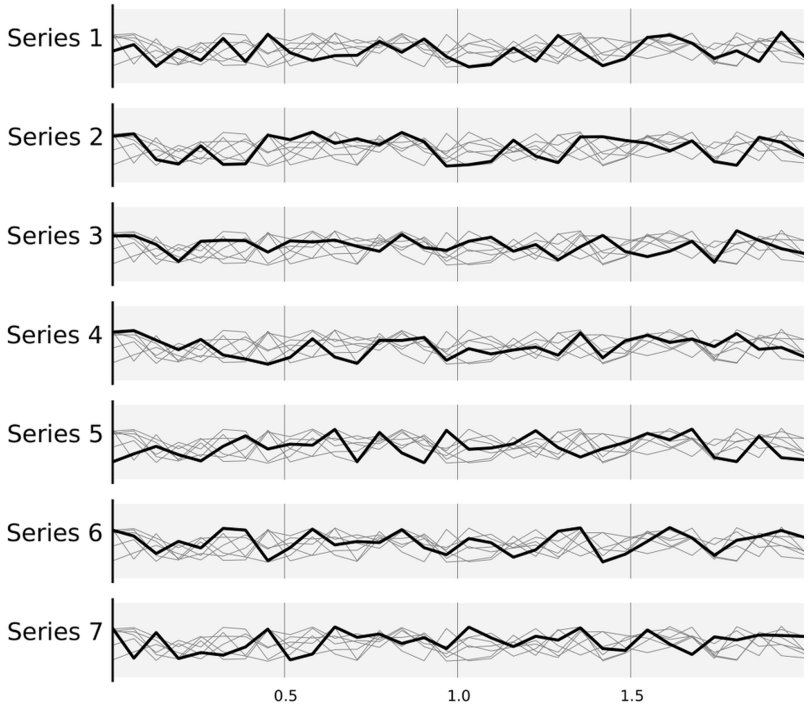
Relative size using full range

Relative size using partial range

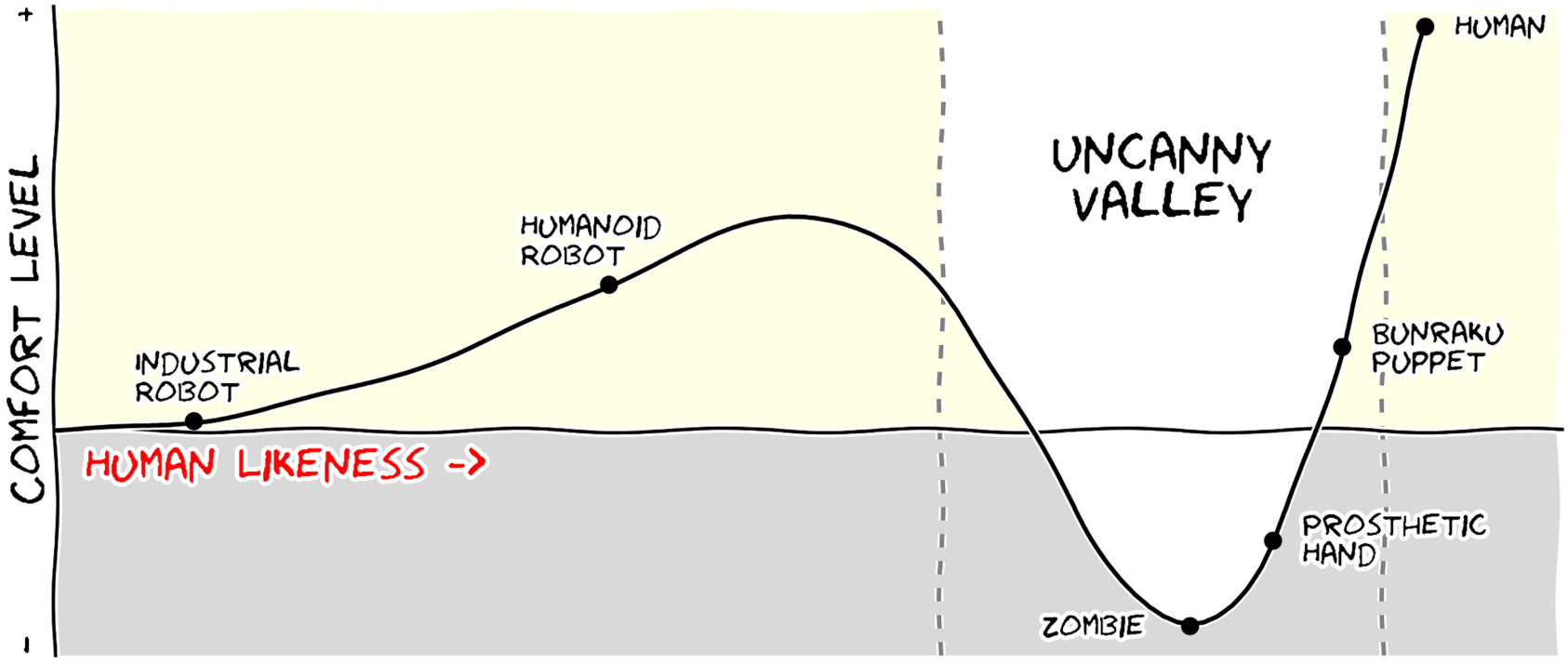


Keep it simple

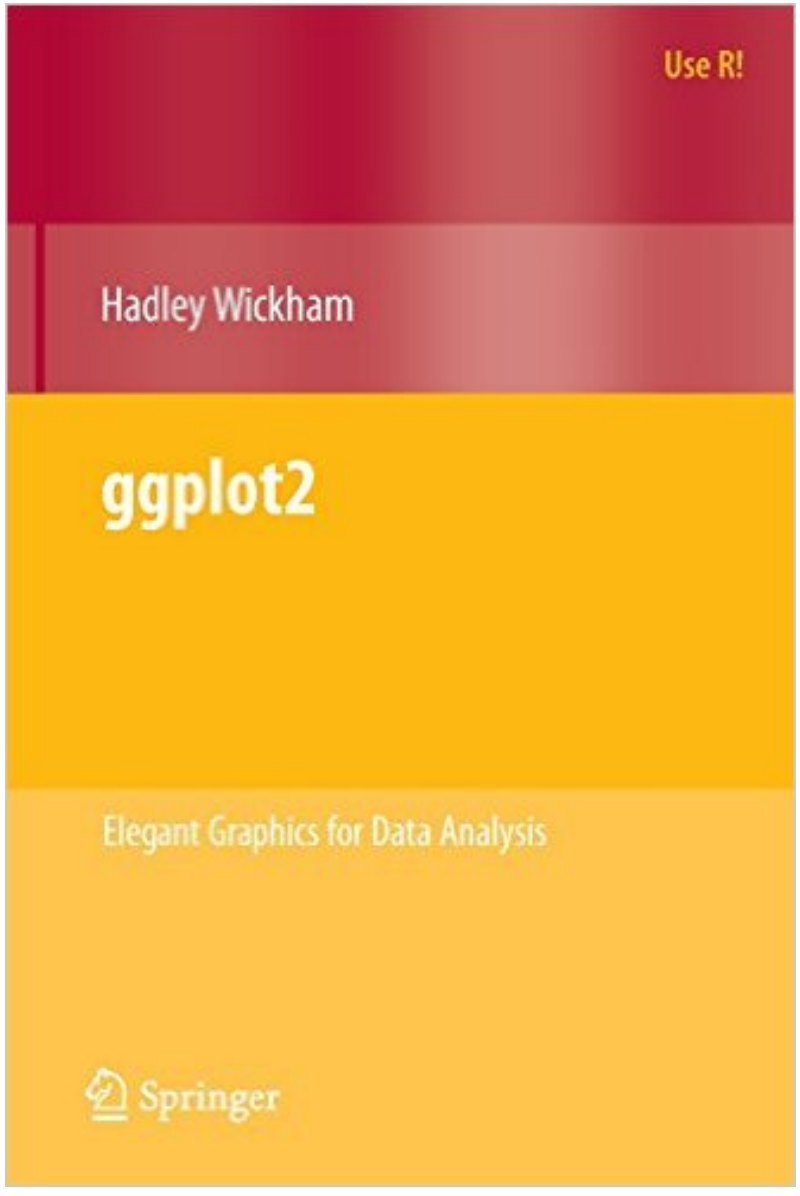




Message is more important than aesthetics



ggplot2: Elegant Graphics for Data Analysis



What is ggplot2?

- R package for producing graphics designed by Hadley Wickham
- Based on the Grammar of Graphics (Wilkinson, 2005)
- Enables you to produce publication-quality graphics quickly and efficiently
- ggplot takes care of the aesthetics, allowing you to focus on what's most important: creating graphs that most effectively communicate your data

What is a grammar of graphics?

Def.: In linguistics, **grammar** is the set of structural rules governing the composition of clauses, phrases, and words in any given natural language.

- Similarly, ggplot2 is composed of independent components that can be combined in a variety of ways
- Unlike MS Excel which constrains you to a small number of existing graphics, ggplot2 allows you to create new graphics specific to your problem

Components of the ggplot2 grammar

The basic idea of ggplot is that a graph is built up in layers

- raw data
- annotations
- statistics

Components of the ggplot2 grammar

- The **data** that you want to graph. Must be stored in a data frame
- **Aesthetic mappings** are the rules you set for translating data into aesthetic attributes such as color and size
- Geometric objects or **geoms** such as points, lines and shapes
- statistical transformations or **stats** for applying statistical transformations to the data (smoothing)
- **scales** set how values of your variables will appear on the graph, whether color, size or shape
- **faceting** creates multiple plots stratified on a third variable

Demonstration

- Data are drawn from the NHANES 1988-2011
- Sample consists of adults ages 50-74
- Is there a relationship between lifetime maximum BMI and hemoglobin A1c?
- Does this relationship differ by sex?

Load libraries

```
suppressMessages(library(ggplot2))
```

```
## Warning: package 'ggplot2' was built under R version 3.4
```

```
suppressMessages(library(gdata))
```

What do the data look like?

seqn	survey	age	male	hisp	black	other	bmiM	bmimax	
352	0	50	1	1	0	0	25.1	26.39343	25
363	0	66	0	1	0	0	23.6	28.24927	23
3124	0	63	0	0	0	0	23.6	23.60000	22
3130	0	55	1	0	1	0	33.6	34.50113	31
3168	0	65	1	1	0	0	25.2	25.20000	20

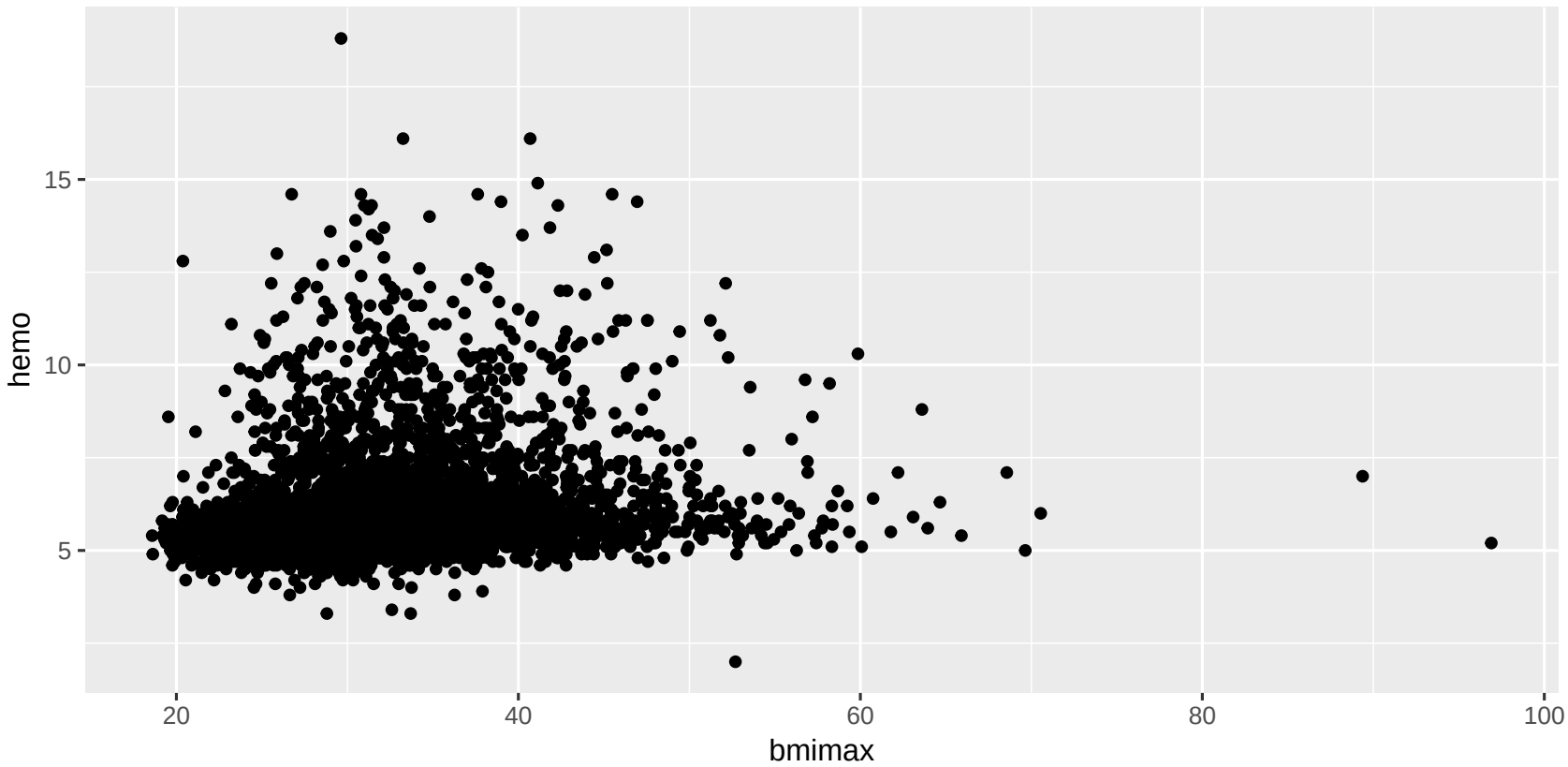
Add first layer of plot

```
g <- ggplot(data, aes(bmimax, hemo))
summary(g)
```

```
## data: seqn, survey, age, male, hisp, black, other, bmiM
##   bmiSR, hemo, smoke [6026x12]
## mapping:  x = ~bmimax, y = ~hemo
## faceting: <ggproto object: Class FacetNull, Facet, gg>
##     compute_layout: function
##     draw_back: function
##     draw_front: function
##     draw_labels: function
##     draw_panels: function
##     finish_data: function
##     init_scales: function
##     map_data: function
##     params: list
##     setup_data: function
##     setup_params: function
```

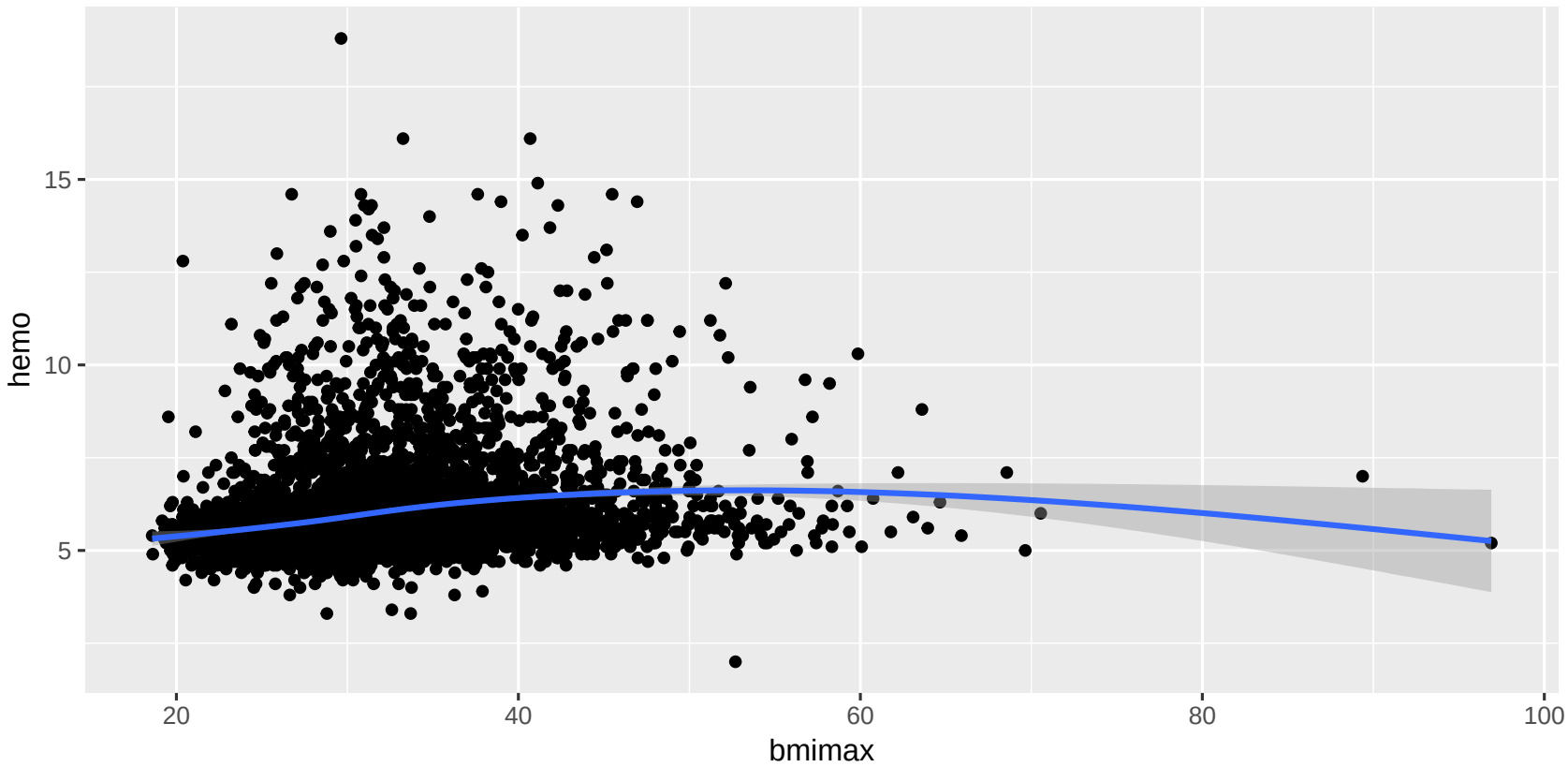
Add points

```
p <- g + geom_point()
print(p)
```



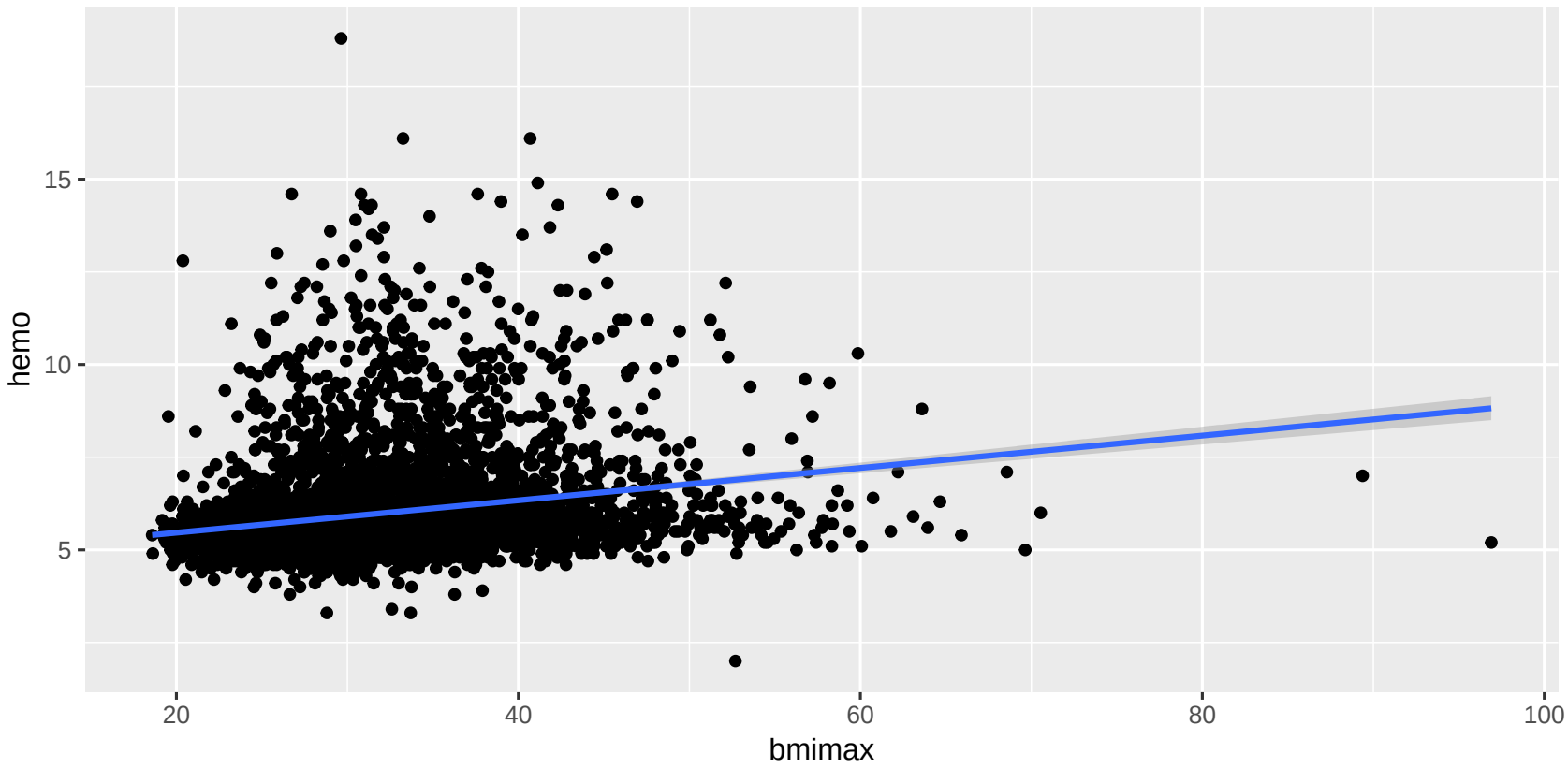
Add statistic

```
p <- g + geom_point() + geom_smooth()
print(p)
```



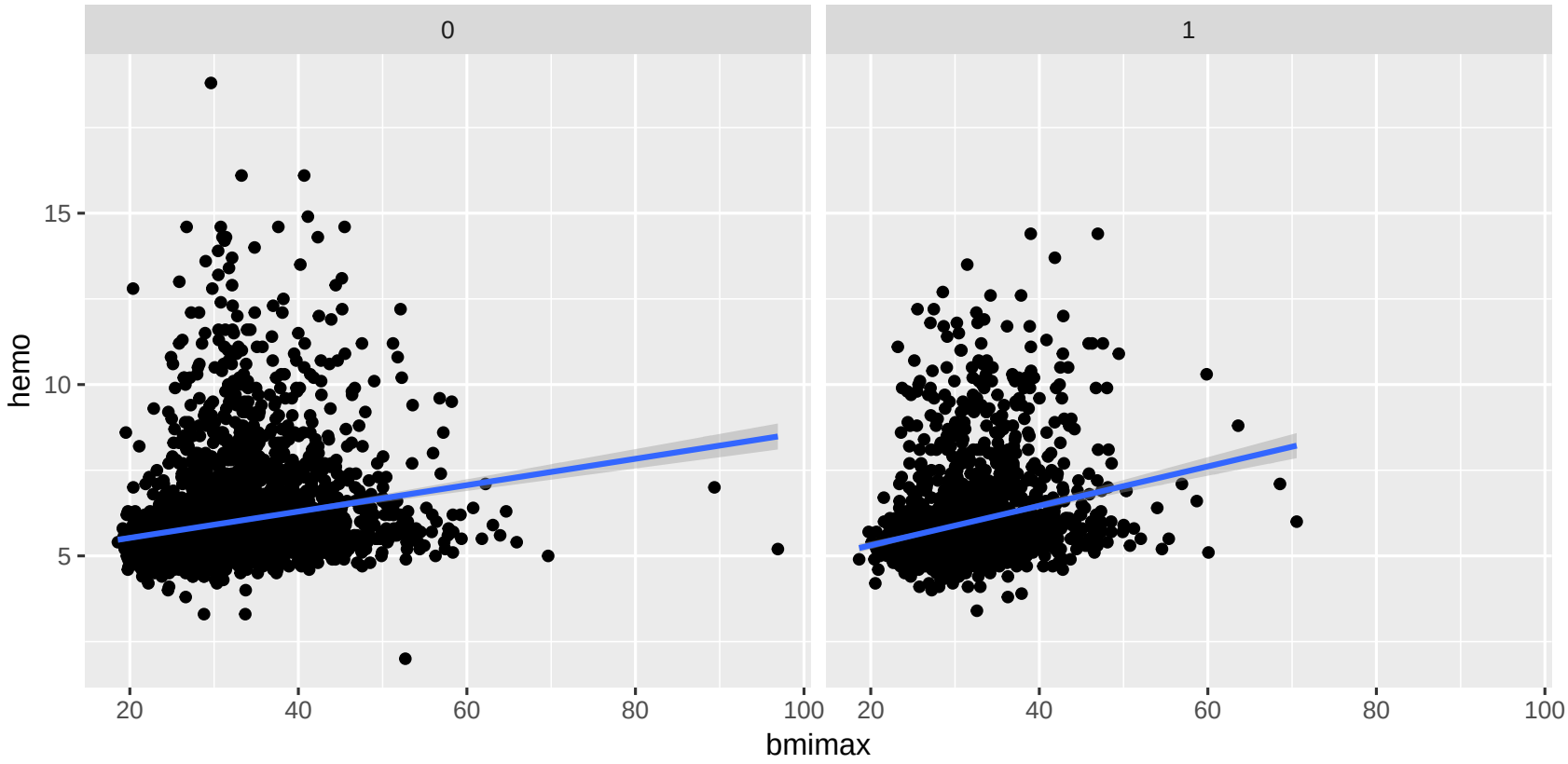
Change smoothing method

```
p <- g + geom_point() + geom_smooth(method="lm")
print(p)
```



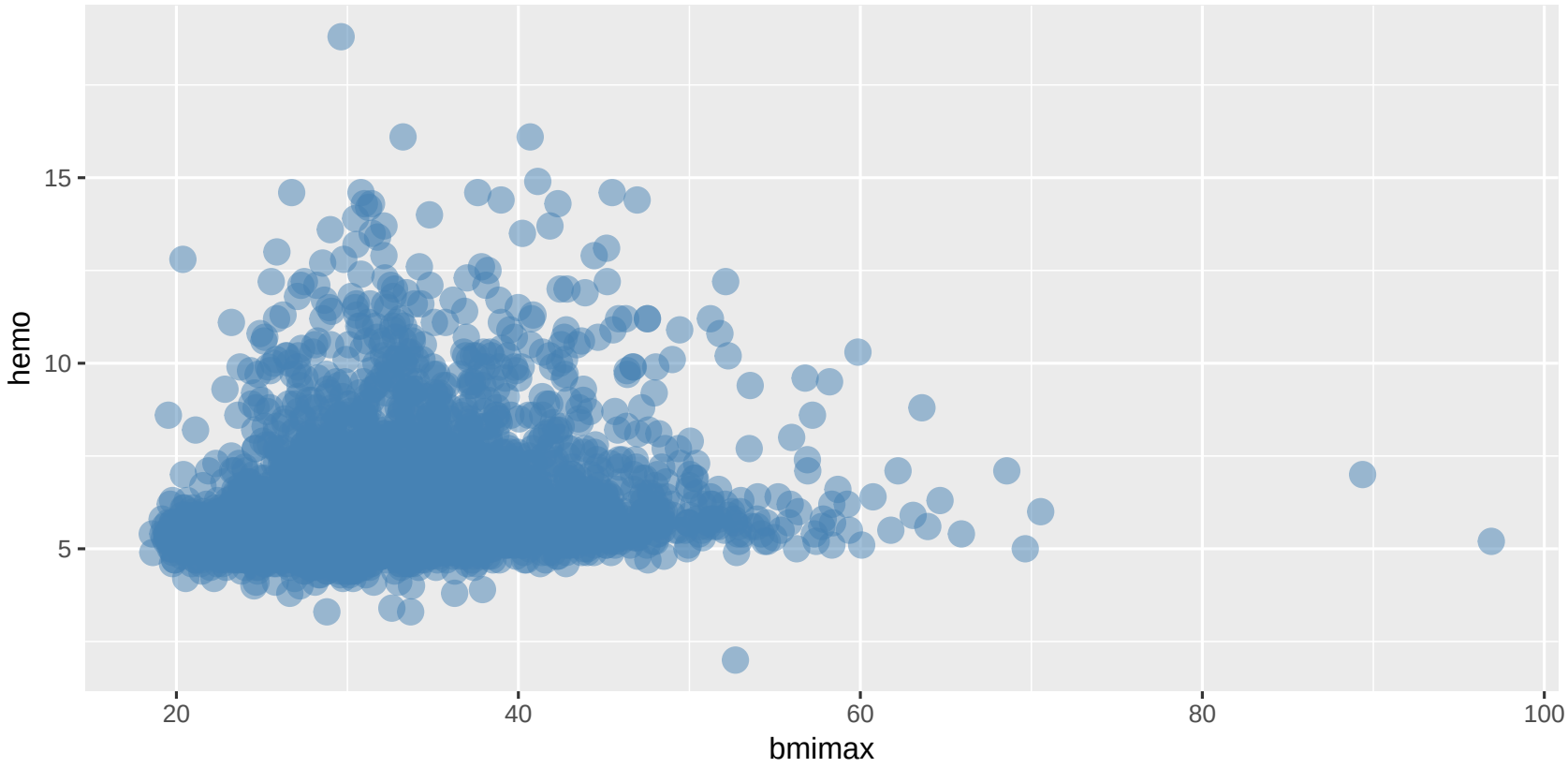
Stratify plot by third variable

```
p <- g + geom_point() + geom_smooth(method="lm")
p + facet_grid(. ~ male)
```



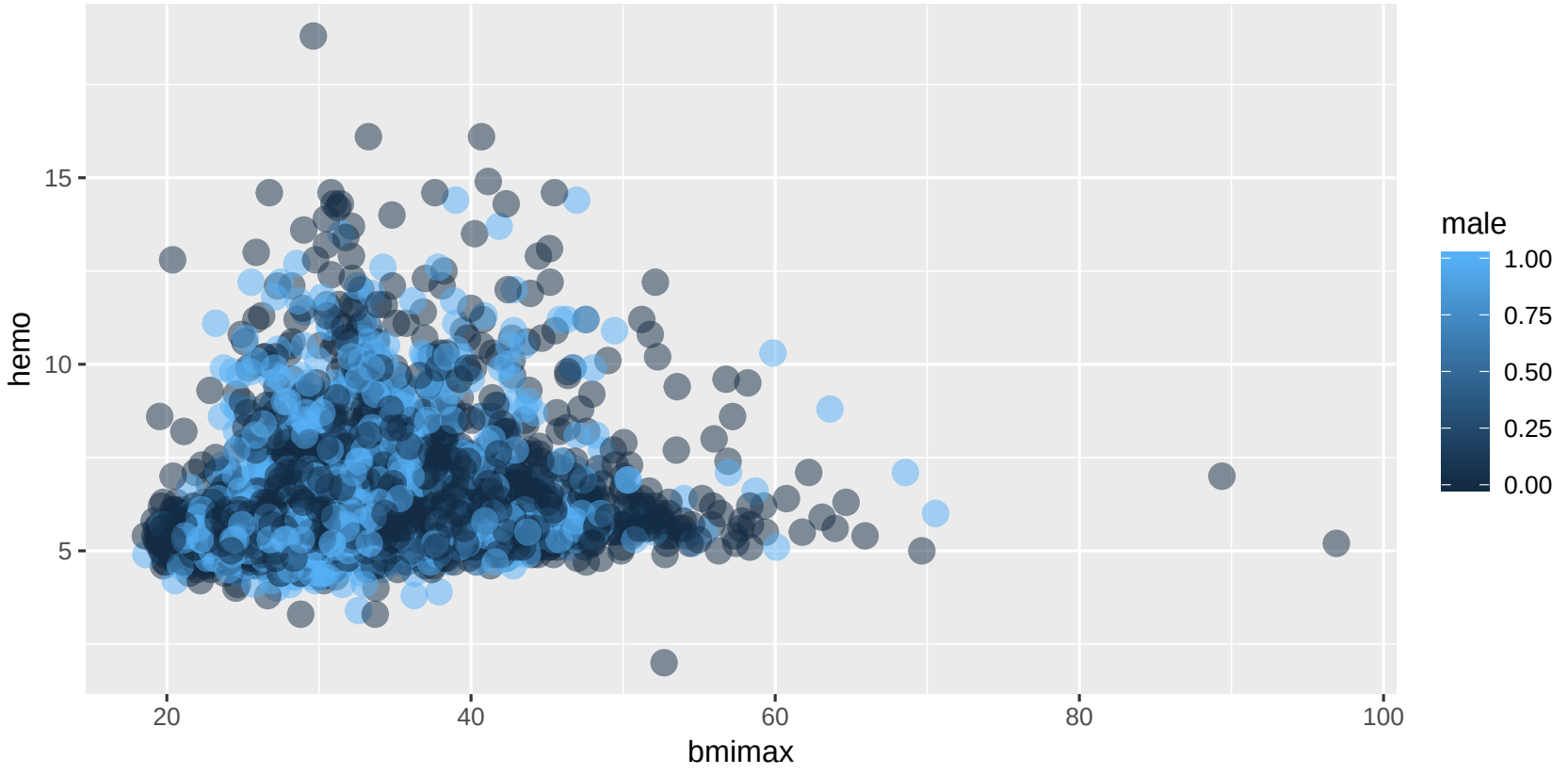
Make a global change to the plot

```
g + geom_point(color = "steelblue", size = 4, alpha = 1/2)
```



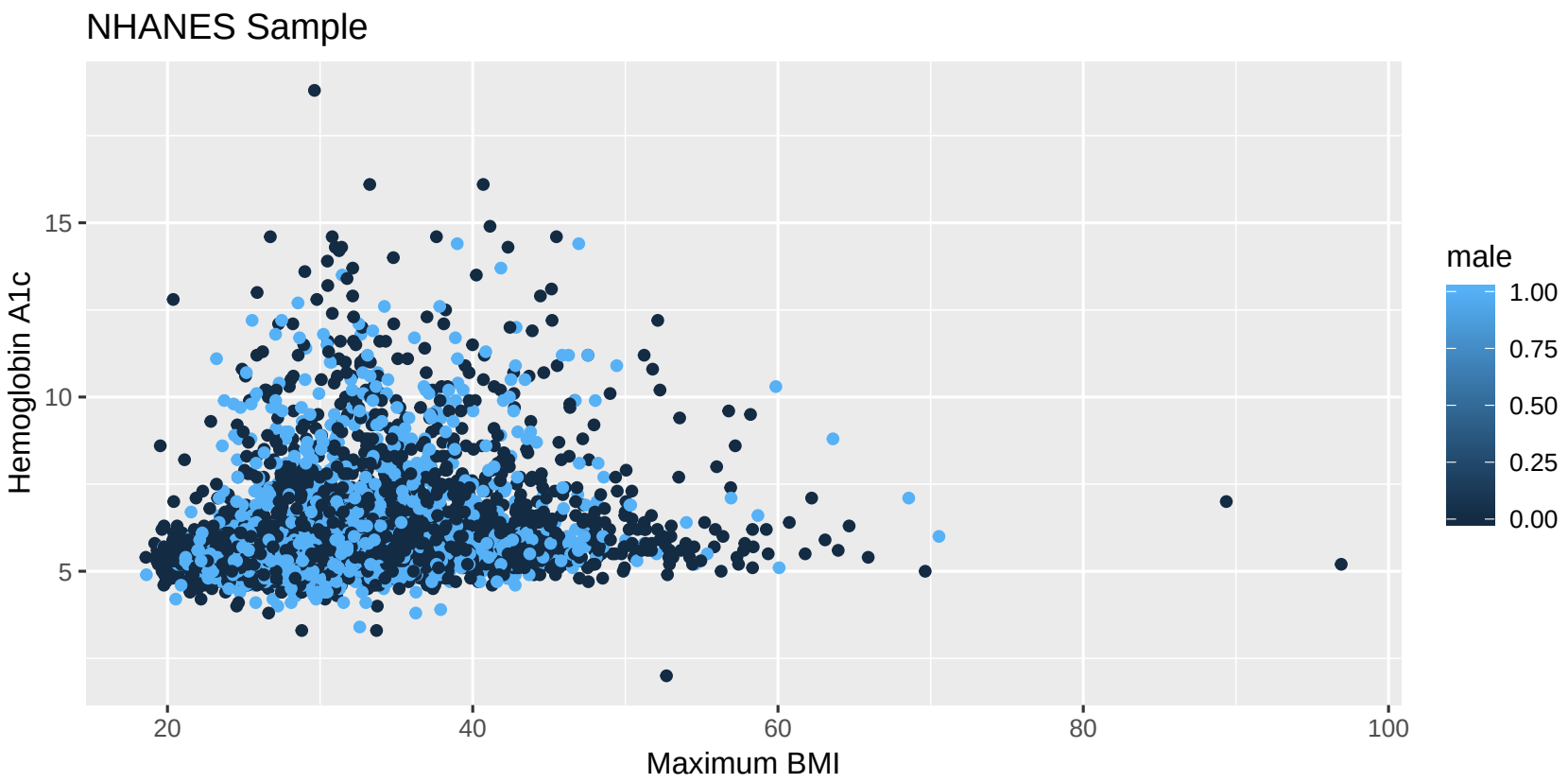
Modify by values

```
g + geom_point(aes(color = male), size = 4, alpha = 1/2)
```



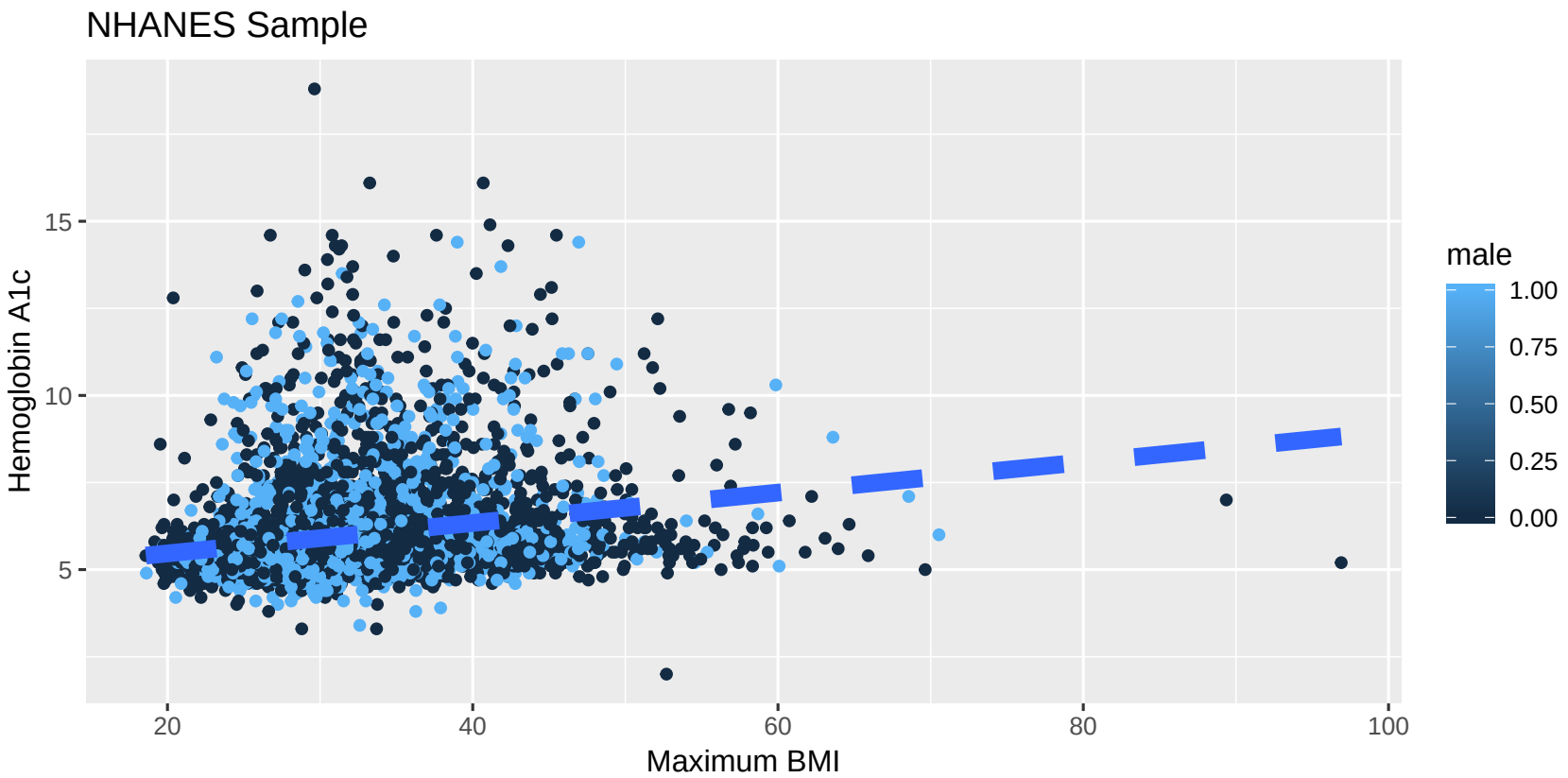
Add labels

```
p <- g + geom_point(aes(color = male))
p <- p + labs(title = "NHANES Sample")
p <- p + labs(x="Maximum BMI", y="Hemoglobin A1c")
print(p)
```



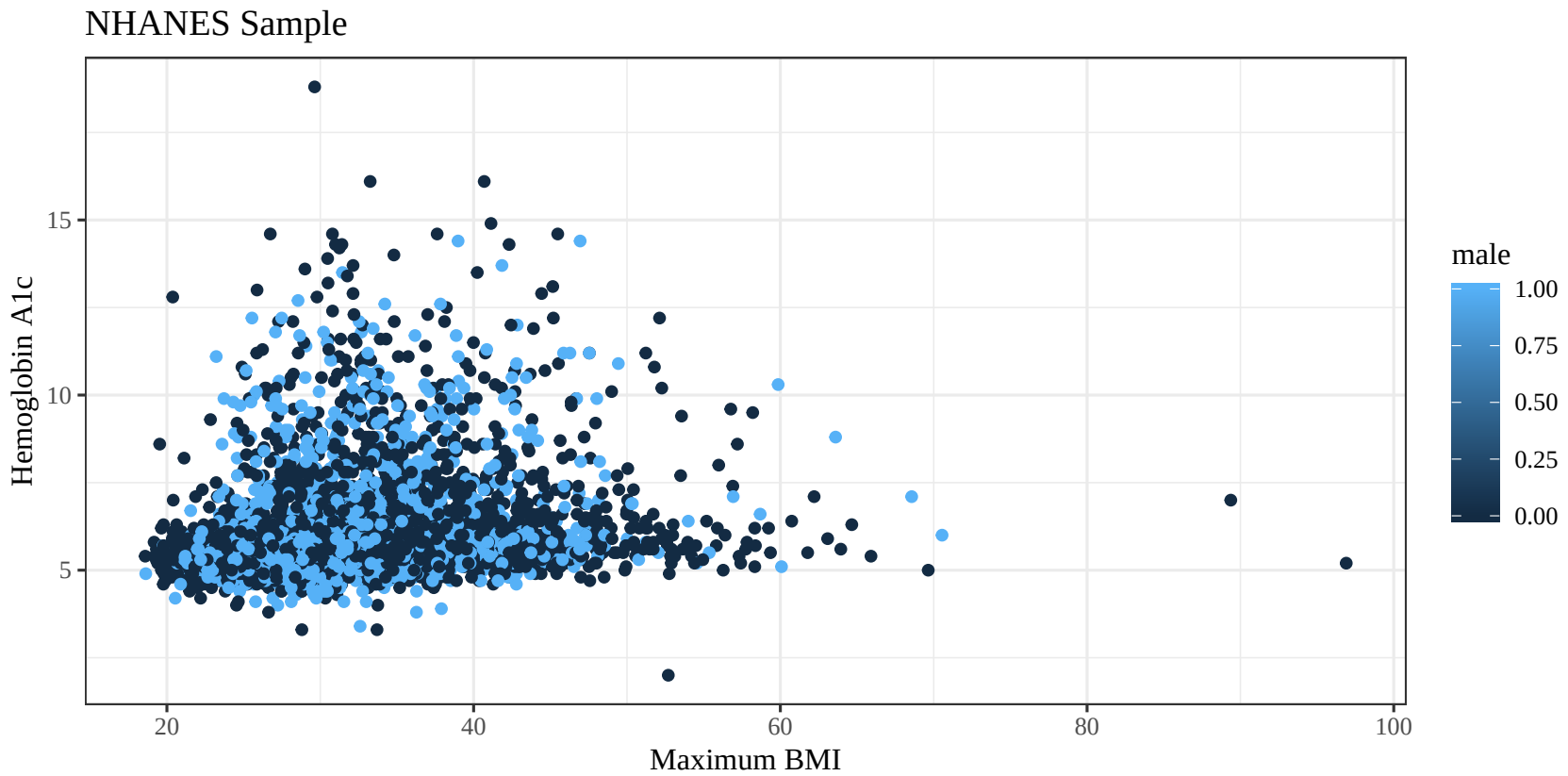
Change default options for the smoother

```
p + geom_smooth(size=3, linetype = 2, method = "lm", se = F
```



Change the theme of the plot

```
p + theme_bw(base_family = "Times")
```



Upcoming deadlines

- Descriptive & bivariate tables (Sunday, April 14 at 5 pm)
- Peer review of descriptive tables (Tuesday, April 16 at 2 pm)
- Problem set 3 (Tuesday, April 16 at 2 pm)