Visualization for Data Science in R

Angela Zoss

Data Matters 2017

http://bit.ly/RVisDay1

Set up environment

- R?
- RStudio?
- tidyverse?
- Project files?

Visual Encoding of Data

What does it mean to represent data visually?

Why do it?

Math is hard

1		2		3		4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Almost identical summary statistics:

x & y mean

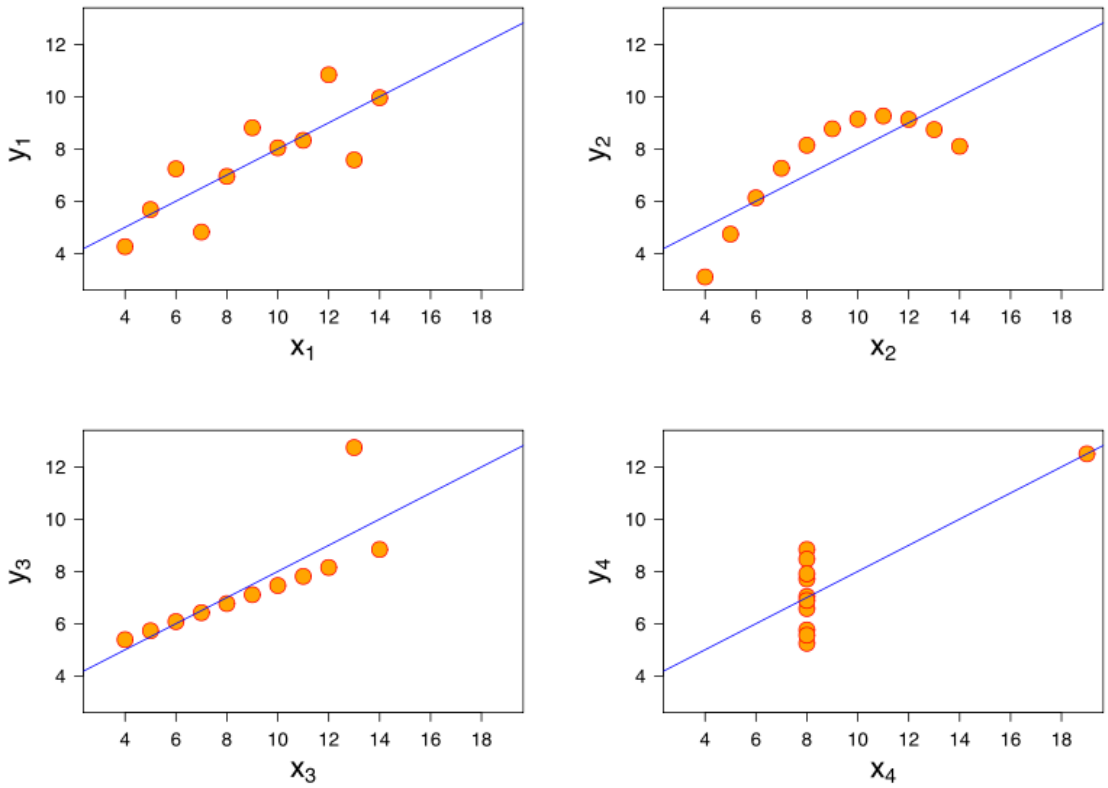
x & y variance

x-y correlation

x-y linear regression

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Shapes are much easier



Anscombe's Quartet

http://en.wikipedia.org/wiki/Anscombe%27s_quartet

Primary kinds of variables

Numbers

- Ratio
- Interval

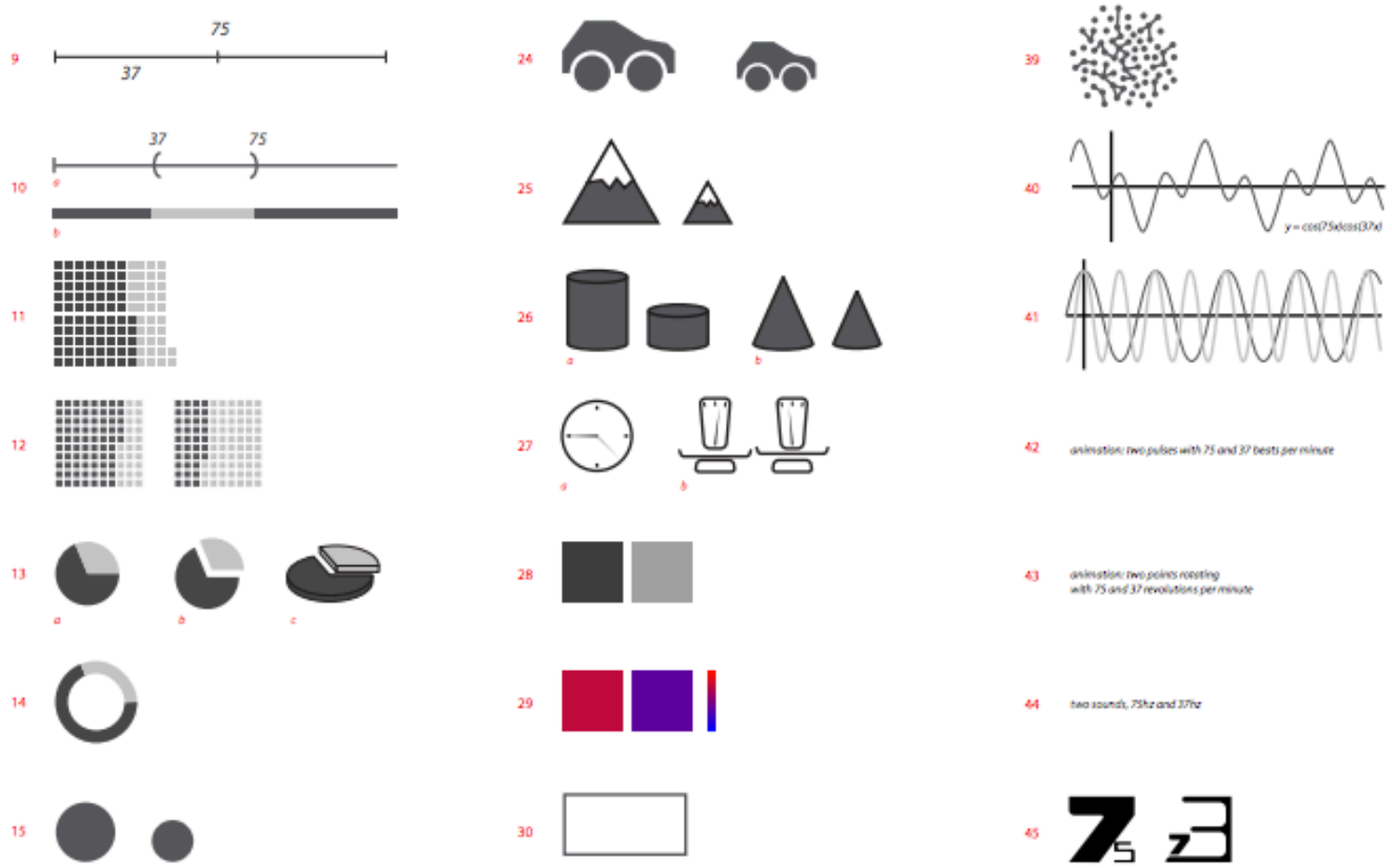
Text/Categorical

- Ordinal
- Nominal

What are some ways to represent **numbers** visually?

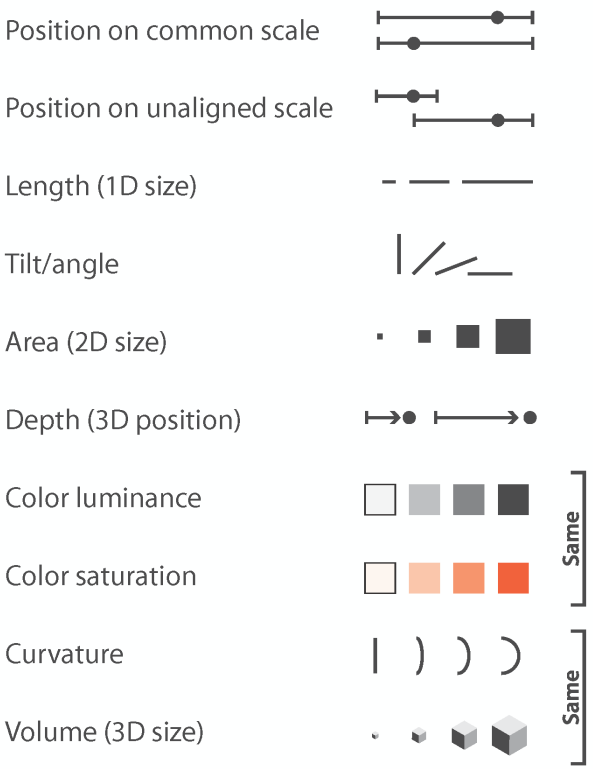
37 75

Think creatively about possibilities!



http://blog.visual.ly/45-ways-to-communicate-two-quantities/

Encodings for Numerical Data

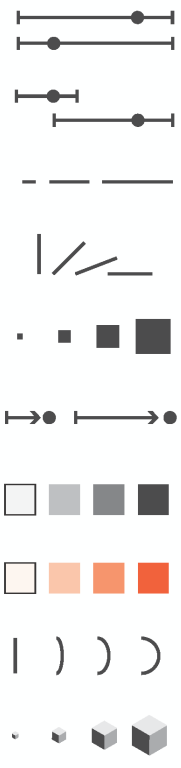


Tamara Munzner

http://www.cs.ubc.ca/~tmm/talks/minicourse14/vad16act.pdf

Encodings for Numerical Data

Twice as far	Position on common scale	
Twice as far	Position on unaligned scale	
Twice as long	Length (1D size)	
Twice as tilted	Tilt/angle	
Twice as large	Area (2D size)	
Twice as deep	Depth (3D position)	
Twice as dark	Color luminance	Same
Twice as red	Color saturation	
Half the diameter(?)	Curvature	Same
Twice as much	Volume (3D size)	



Tamara Munzner

http://www.cs.ubc.ca/~tmm/talks/minicourse14/vad16act.pdf

What are some ways to represent **categories** visually?

Encodings for Categorical Data

Tamara Munzner

http://www.cs.ubc.ca/~tmm/talks/minicourse14/vad16act.pdf

Encodings for Categorical Data

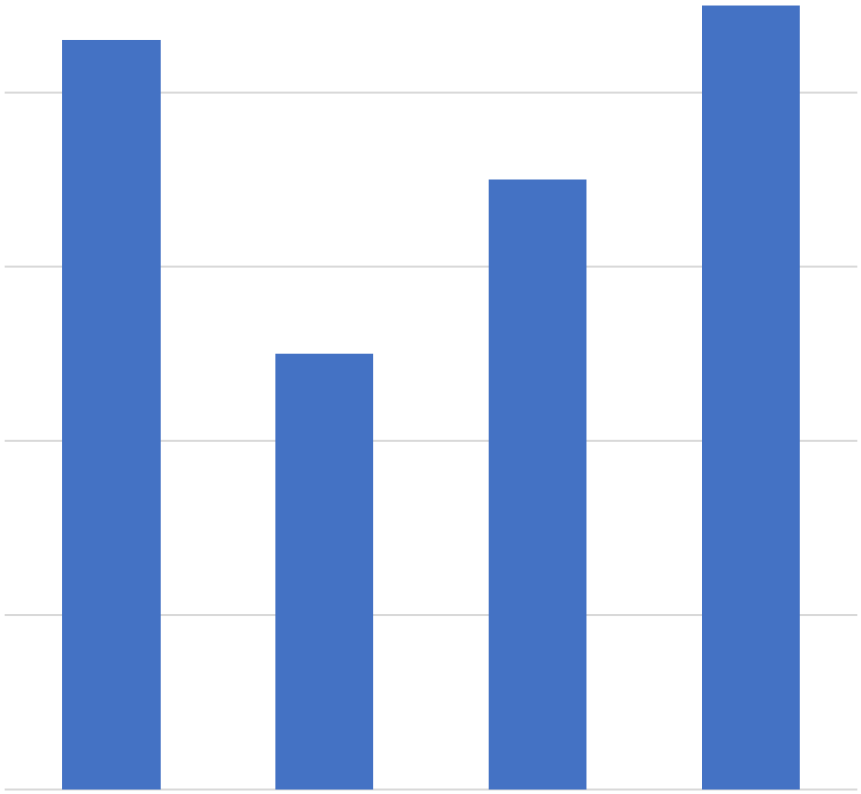
Different place	Spatial region	
Different color	Color hue	
Different path	Motion	
Different shape	Shape	

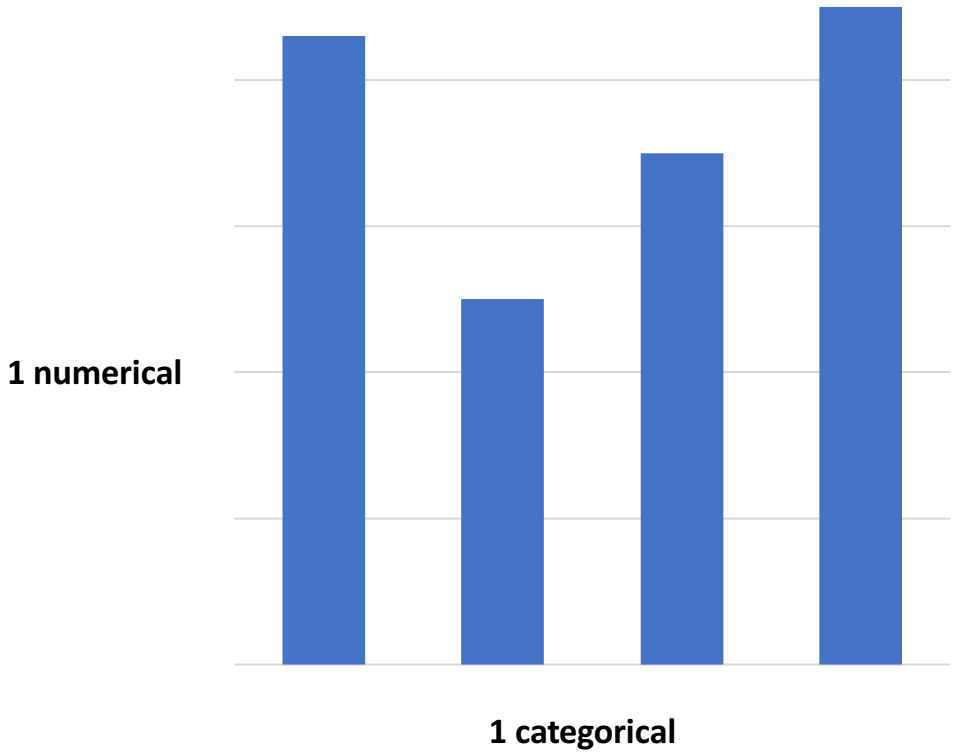
Tamara Munzner

http://www.cs.ubc.ca/~tmm/talks/minicourse14/vad16act.pdf

Visual Encodings in Charts

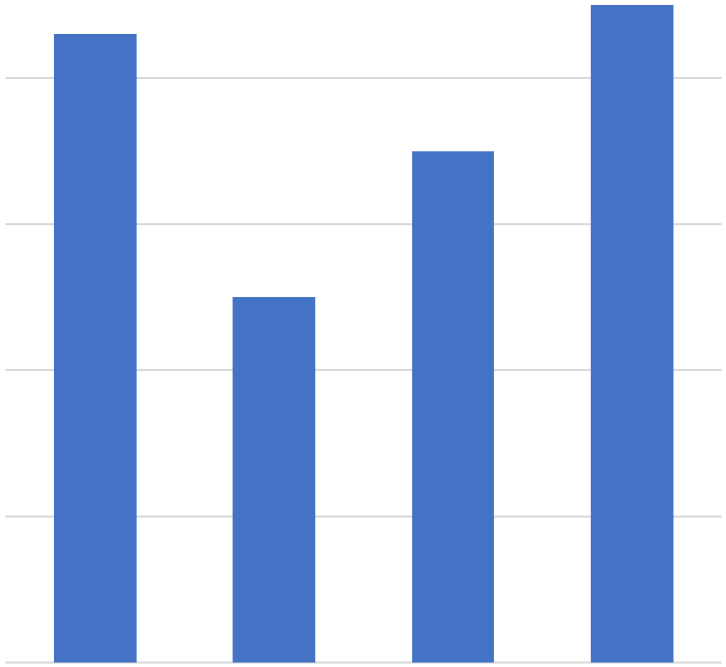
What types of data go in a bar chart?





Now, what type of encoding is each axis using?

1 numerical:
length/ position on common scale

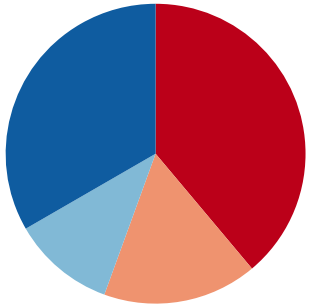


1 categorical:
spatial region/position

Different chart types, different encodings

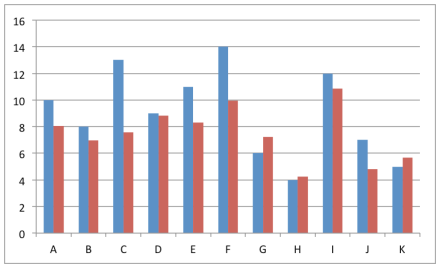
categorical (color/position)

numerical (area)



categorical (color)

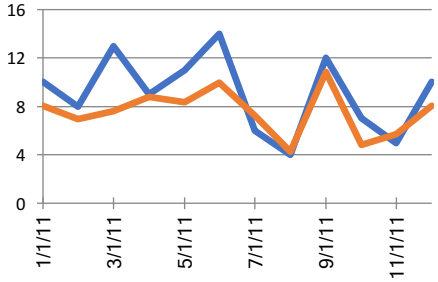
numerical (length)



categorical (position)

categorical (color)

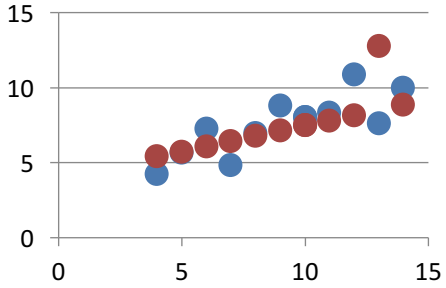
numerical (position)



date/numerical (position)

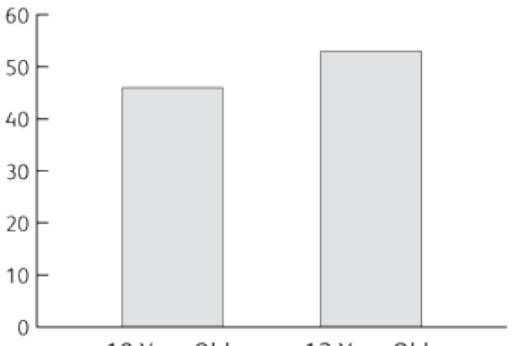
categorical (color)

numerical (position)

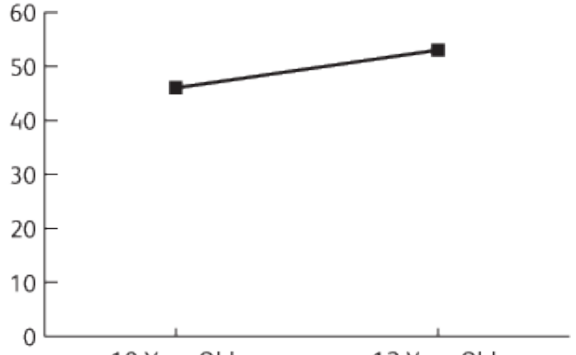


numerical (position)

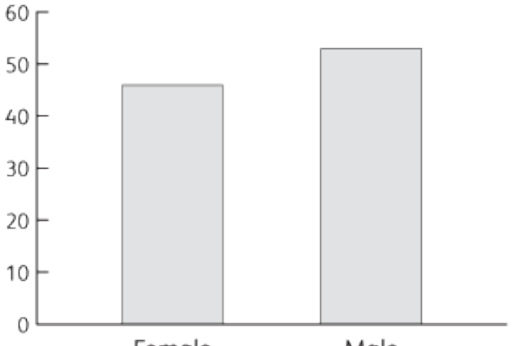
Not all encodings make sense



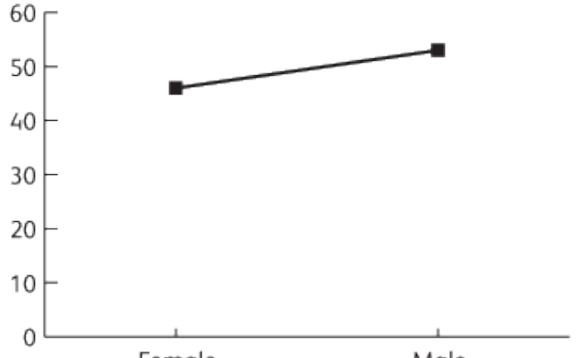
"Twelve year olds are taller than ten year olds."



"As you get older, you get taller."



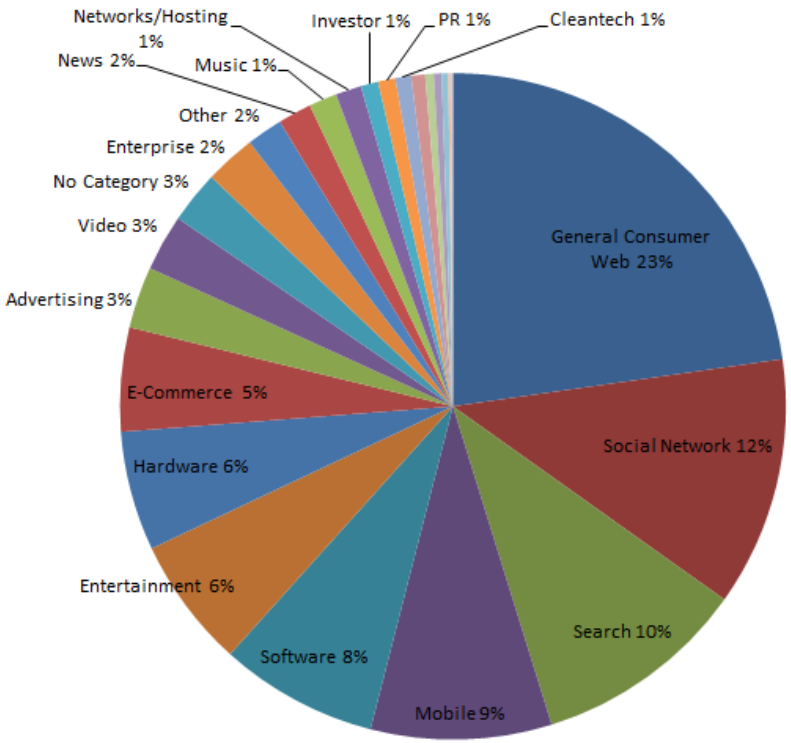
"Males are taller than females."

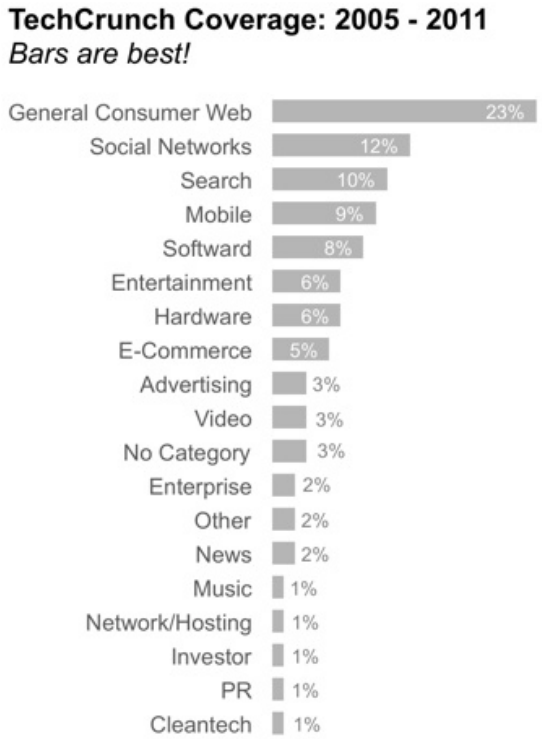


"As you get more male, you get taller."

http://www.scribblelive.com/blog/2011/12/05/a-line-walks-into-a-bar/

Even the “right” data may not work



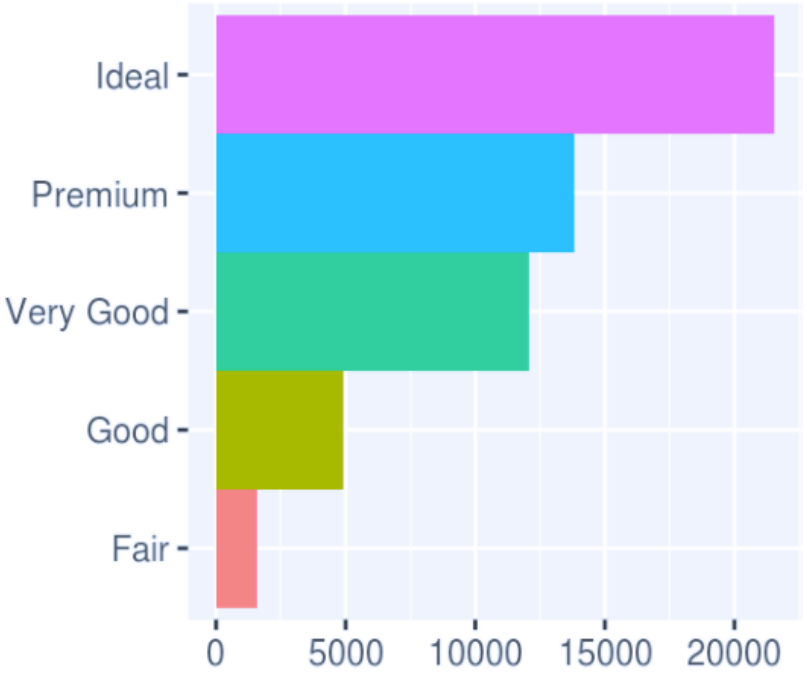


http://bit.ly/r3NUCs

ggplot2

What is ggplot2?

an R package designed to create plots based on a theory of the grammar of graphics.



[Data visualization chapter, R4DS book]

Why ggplot2 instead of base R?

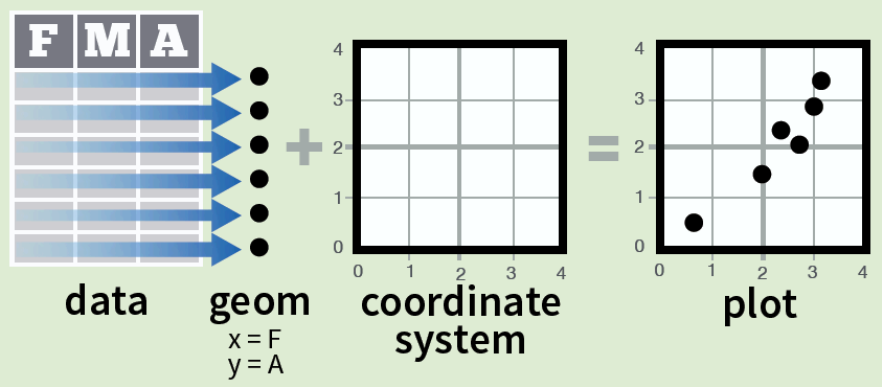
- nice defaults
- easy faceting
- (arguably) more natural syntax
- can switch chart types more easily

"Why I use ggplot2", David Robinson

ggplot2: Elements

Basic elements in any ggplot2 visualization

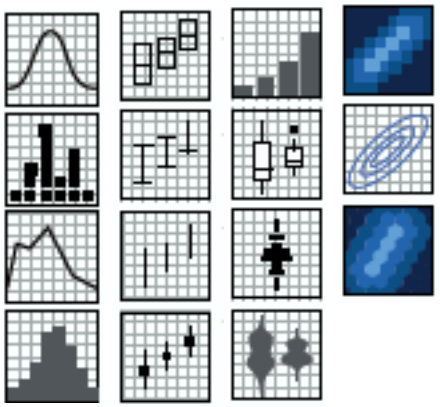
- **data**
- **aes**thetics
  (variable mappings)
- **geom**
  (chart type or shape)
- coordinate system
  (the arrangement of the marks;
  most geoms use default, cartesian)



ggplot2 cheatsheet

Types of geoms

- geom_bar()
- geom_point()
- geom_histogram()
- geom_map()
- etc.



ggplot2 cheatsheet

Note: some geoms also include data summary functions.
e.g., the “bar” geom will count data points in each category.

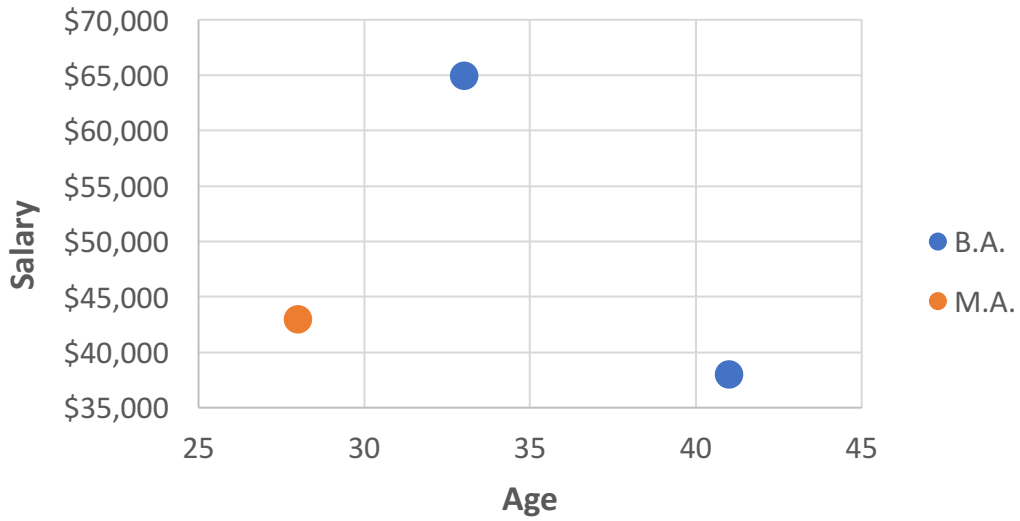
ggplot2: Basic syntax

Template for a simple plot

```
ggplot()  +

    geom_...( [data = data frame]
              [aes(variable mappings)]
              [non-variable adjustments] )
```

Aesthetic variable mappings

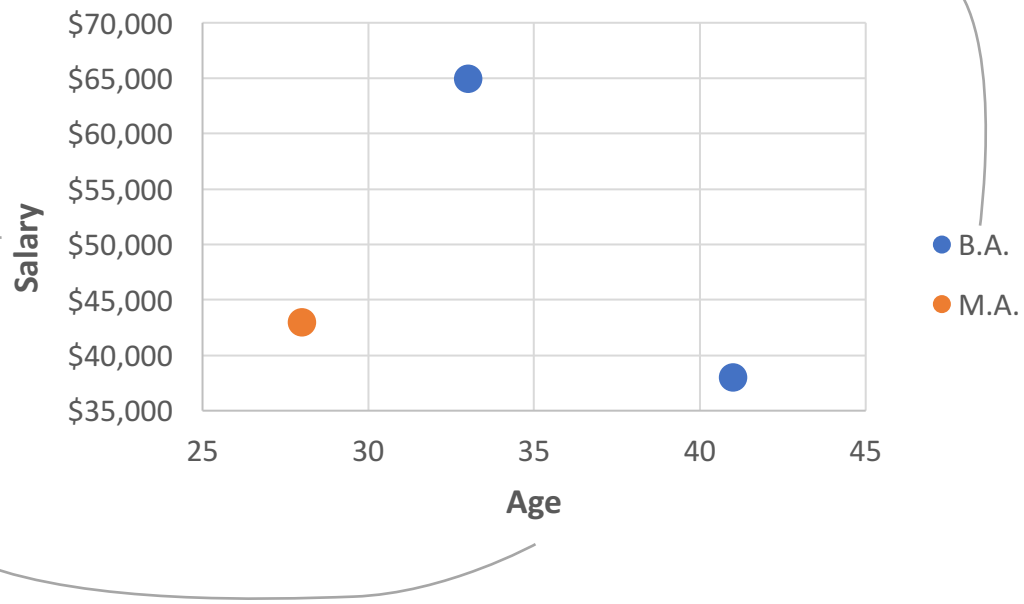
Name	Age	Salary	Highest Degree
Jane Smith	33	$65,000	B.A.
Abby Jones	28	$43,000	M.A.
Bridget Carden	41	$38,000	B.A.



Aesthetic variable mappings

Name	Age	Salary	Highest Degree
Jane Smith	33	$65,000	B.A.
Abby Jones	28	$43,000	M.A.
Bridget Carden	41	$38,000	B.A.

x position y position color

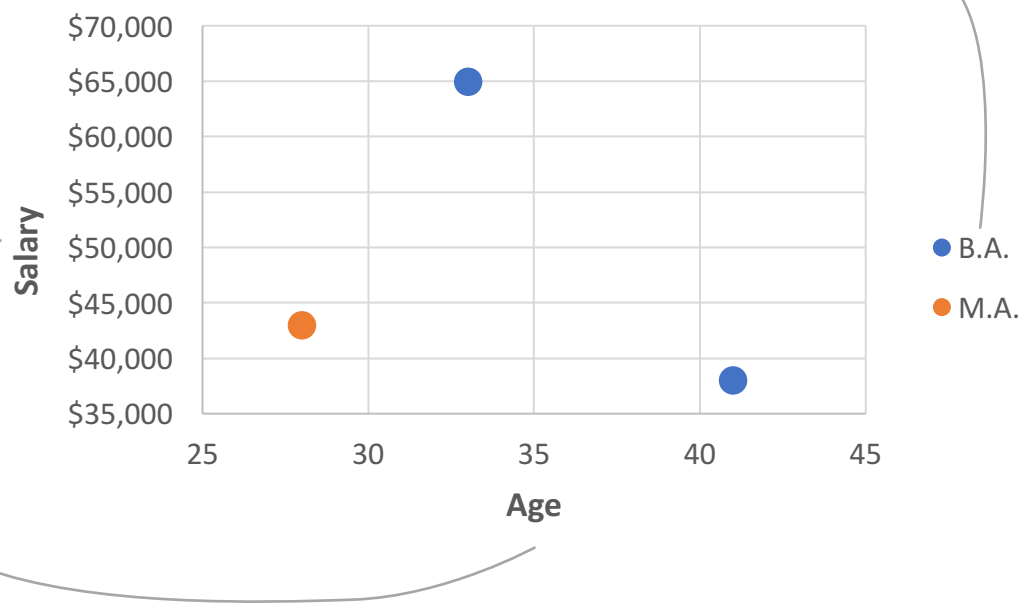


Aesthetic variable mappings

Name	Age	Salary	Highest Degree
Jane Smith	33	$65,000	B.A.
Abby Jones	28	$43,000	M.A.
Bridget Carden	41	$38,000	B.A.

```
ggplot() +
    geom_point(data,
        aes(x=age,
            y=salary,
            fill=degree))
```

x position y position color

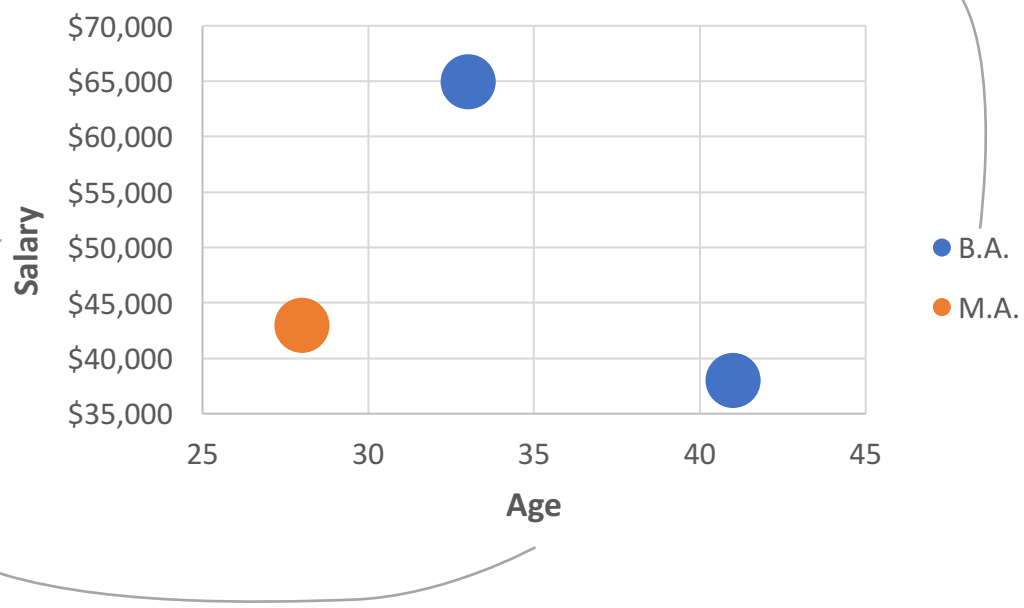


Non-variable adjustments

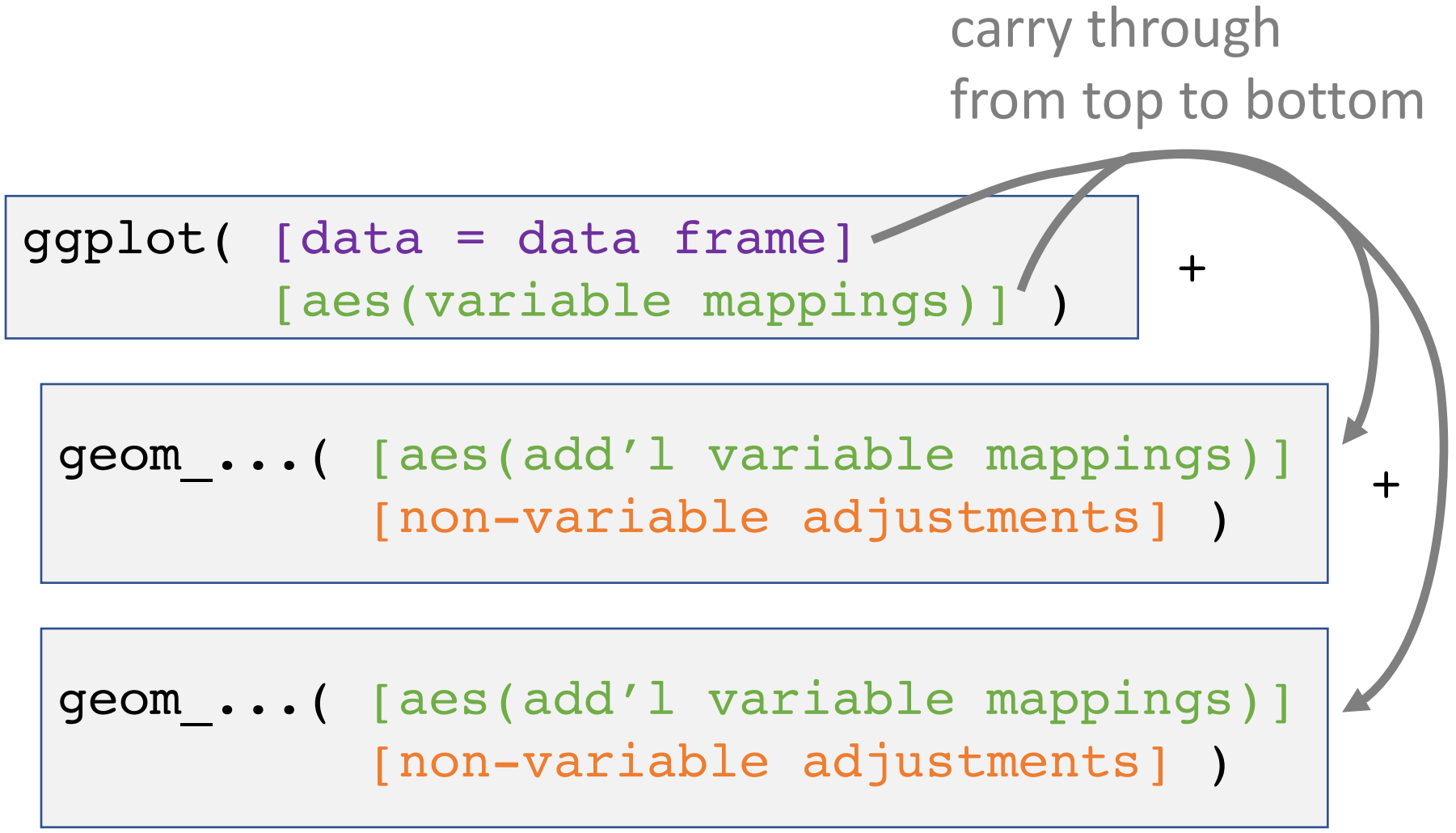
Name	Age	Salary	Highest Degree
Jane Smith	33	$65,000	B.A.
Abby Jones	28	$43,000	M.A.
Bridget Carden	41	$38,000	B.A.

```
ggplot() +
    geom_point(data,
        aes(x=age,
            y=salary,
            fill=degree),
        size=10)
```

x position y position color



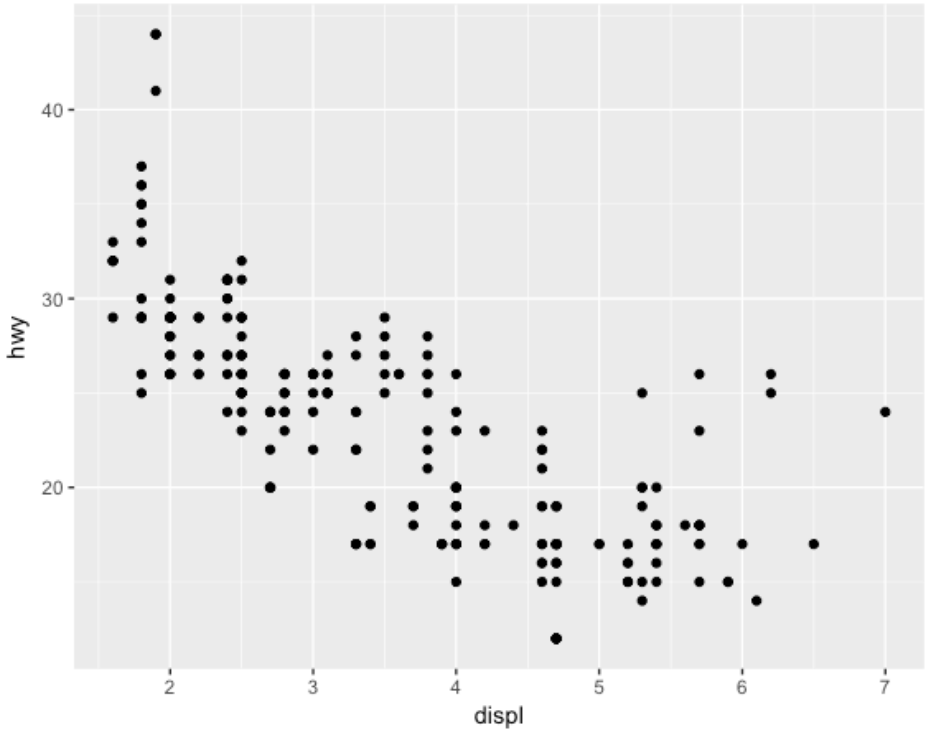
Template for a more complex plot



ggplot2: Building a plot

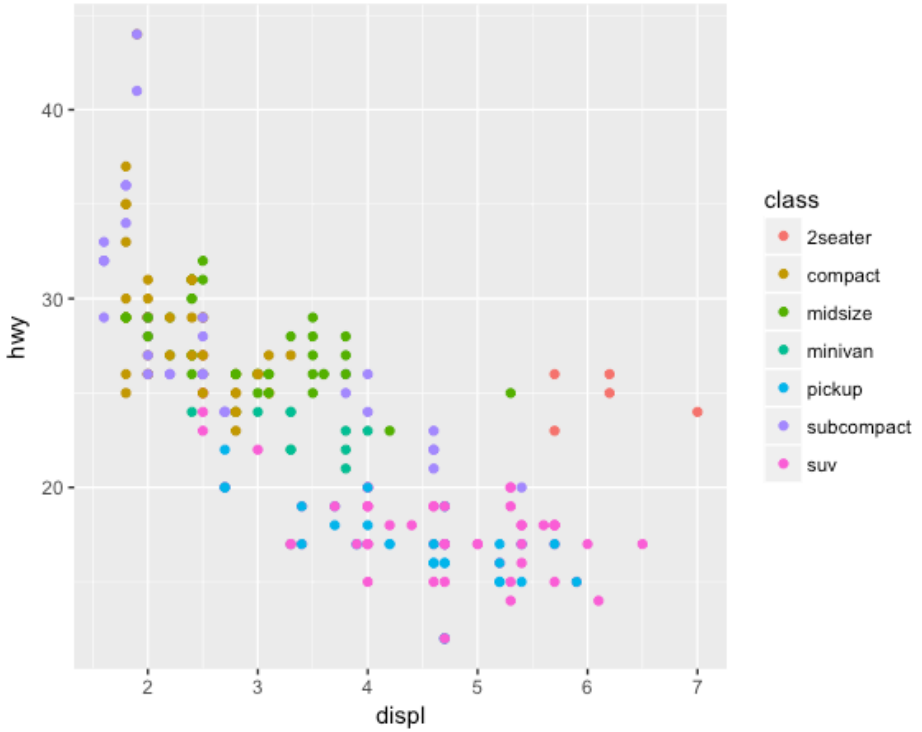
Follow along in an empty R script

```
library(ggplot2)


ggplot(mpg, aes(displ, hwy)) +
   geom_point()
```



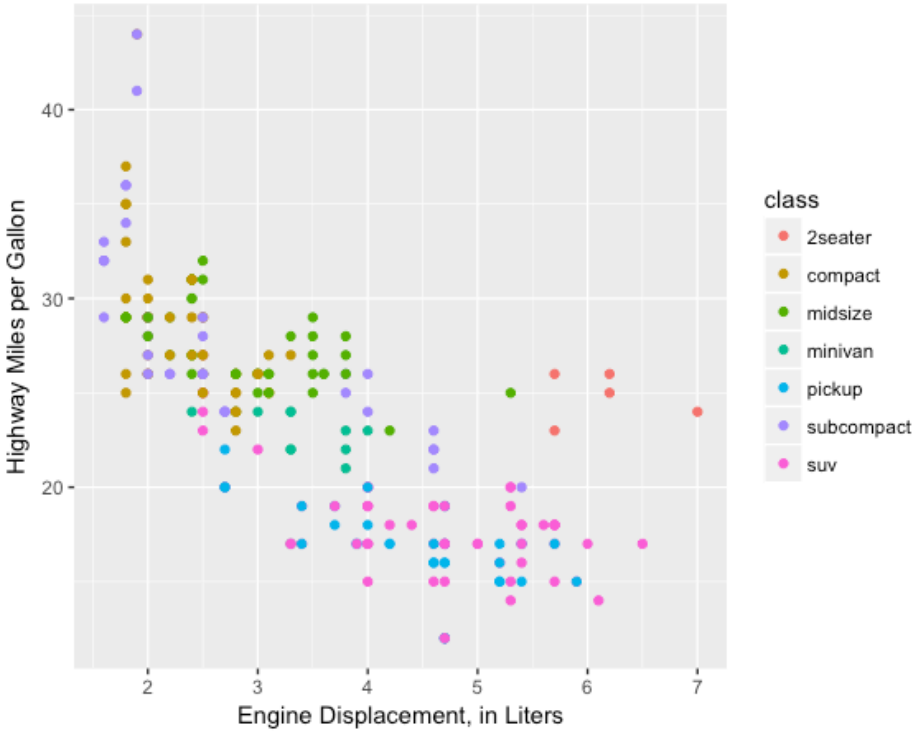
Graphics for Communication,
R for Data Science

```
library(ggplot2)


ggplot(mpg, aes(displ, hwy)) +
   geom_point(aes(color = class))
```



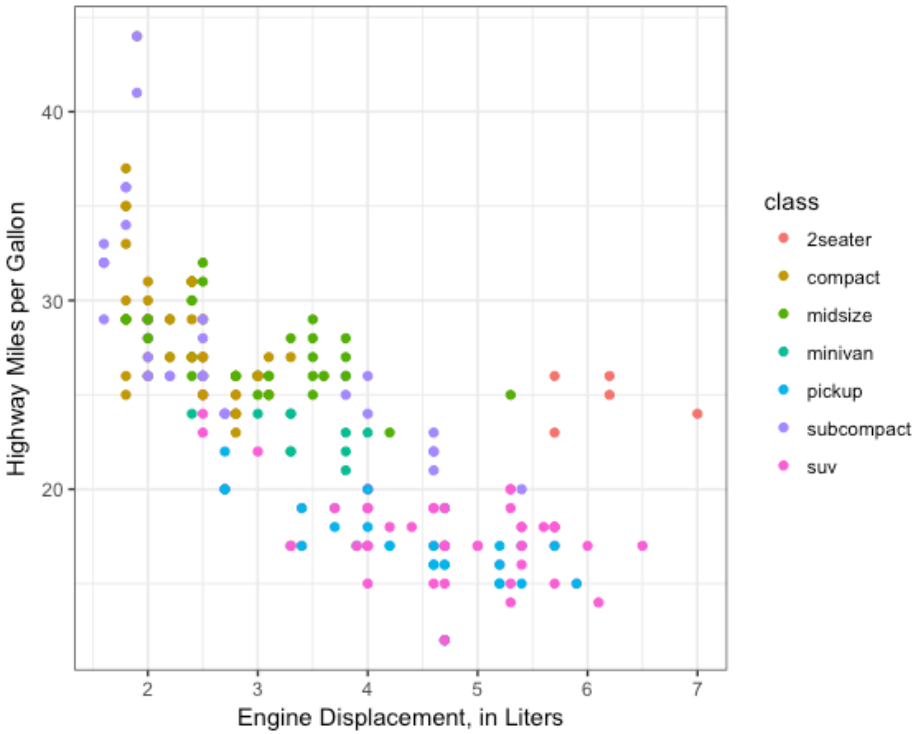
Graphics for Communication, R for Data Science

```
library(ggplot2)


ggplot(mpg, aes(displ, hwy)) +
   geom_point(aes(color = class)) +
   labs(x = "Engine Displacement,
        in Liters", y="Highway
        Miles per Gallon")
```



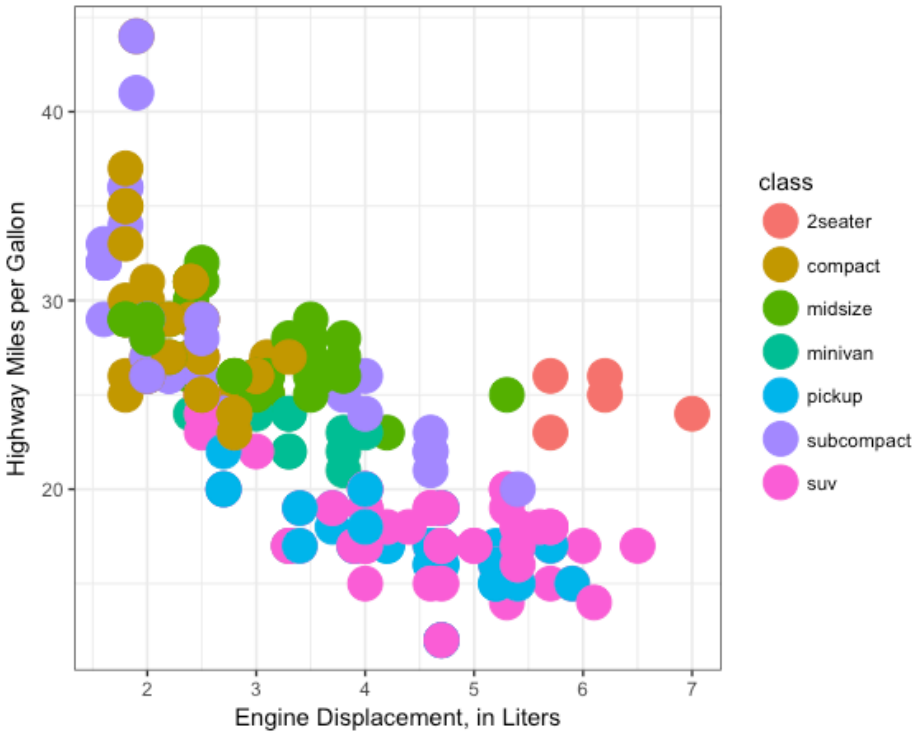
[Graphics for Communication, R for Data Science]

```
library(ggplot2)


ggplot(mpg, aes(displ, hwy)) +
   geom_point(aes(color = class)) +
   labs(x = "Engine Displacement,
        in Liters", y="Highway
        Miles per Gallon") +
   theme_bw()
```



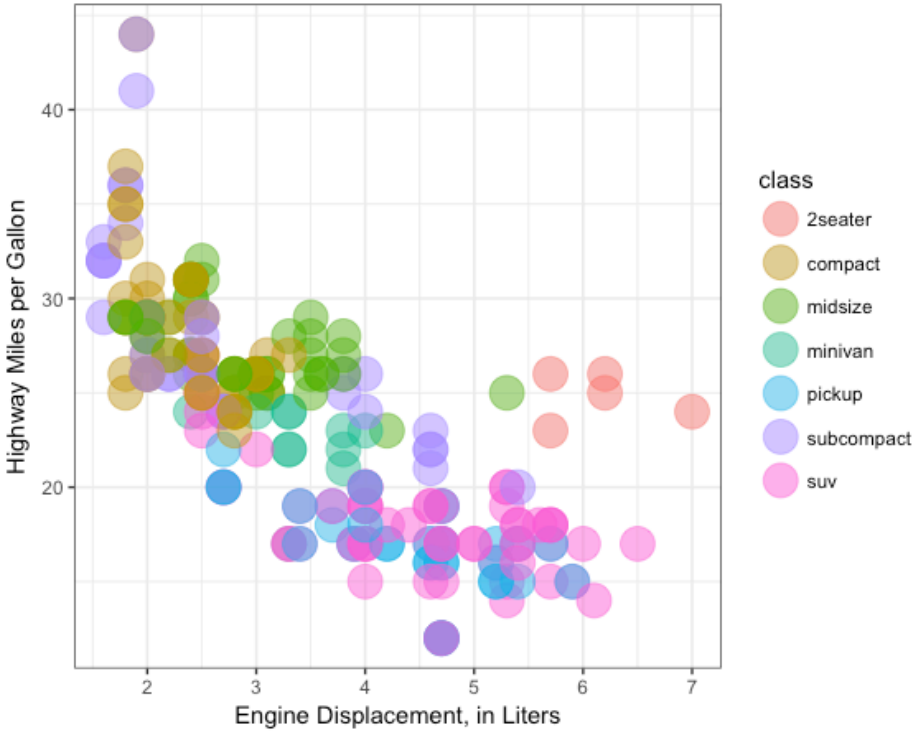
Graphics for Communication, R for Data Science

```
library(ggplot2)


ggplot(mpg, aes(displ, hwy)) +
   geom_point(aes(color = class),
                  size = 7) +
   labs(x = "Engine Displacement,
        in Liters", y="Highway
        Miles per Gallon") +
   theme_bw()
```



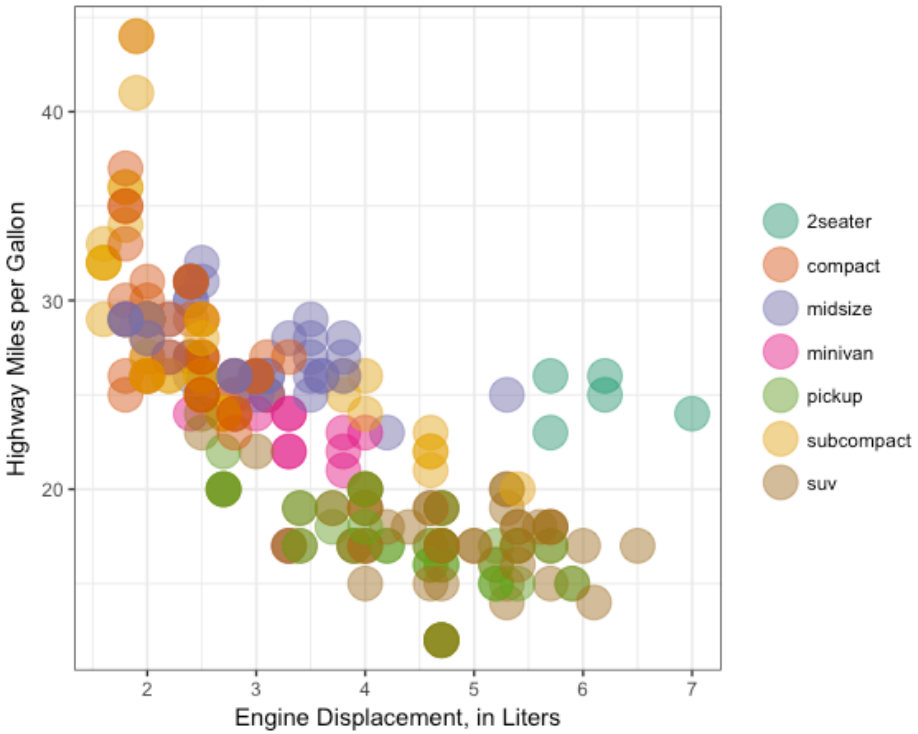
[Graphics for Communication, R for Data Science]

```
library(ggplot2)


ggplot(mpg, aes(displ, hwy)) +
   geom_point(aes(color = class),
                  size = 7,
                  alpha = 0.5) +
   labs(x = "Engine Displacement,
        in Liters", y="Highway
        Miles per Gallon") +
   theme_bw()
```



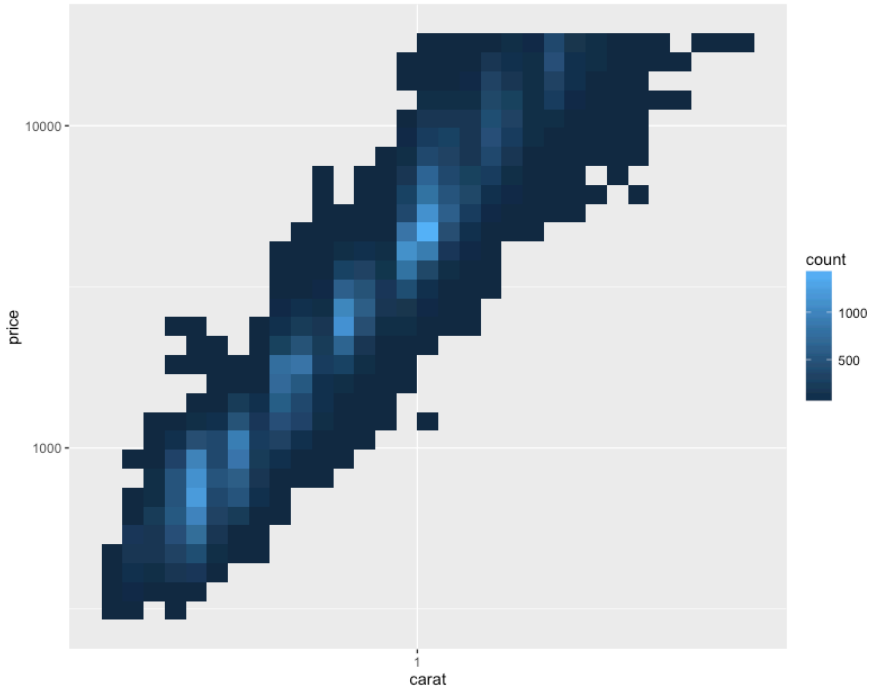
Graphics for Communication, R for Data Science

```
library(ggplot2)


ggplot(mpg, aes(displ, hwy)) +
   geom_point(aes(color = class),
                  size = 7,
                  alpha = 0.5) +
   labs(x = "Engine Displacement,
        in Liters", y="Highway
        Miles per Gallon") +
   scale_color_brewer(
        palette="Dark2", name="") +
   theme_bw()
```



Graphics for Communication,
R for Data Science

```
# geom_bin2d will aggregate points
for you

# using scale_?_log10 will change the
axis spacing but leave labels
comprehensible

ggplot(diamonds, aes(carat, price)) +
   geom_bin2d() +
   scale_x_log10() +
   scale_y_log10()
```



[Graphics for Communication, R for Data Science]

Morning break

ggplot2: Chart quirks

Chart components/slots

Bar chart, for example:

- x
  category (the names of the bars)
- y (optional)
  default is count, but can also specify a number (the length of the bars)
- color (optional)
  category (grouped or stacked bars)

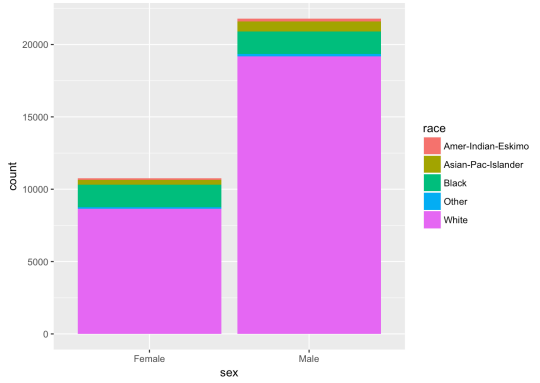
Adult Census Income data

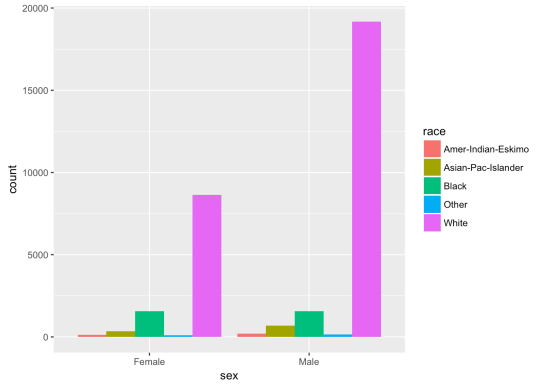
- From https://www.kaggle.com/uciml/adult-census-income
- 32k records, 15 variables

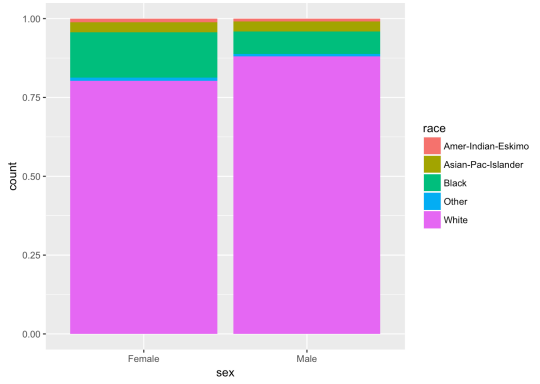
age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex
90	NA	77053	HS-grad	9	Widowed	NA	Not-in-family	White	Female
82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female
66	NA	186061	Some-college	10	Widowed	NA	Unmarried	Black	Female
54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female
41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female
34	Private	216864	HS-grad	9	Divorced	Other-service	Unmarried	White	Female
38	Private	150601	10th	6	Separated	Adm-clerical	Unmarried	White	Male
74	State-gov	88638	Doctorate	16	Never-married	Prof-specialty	Other-relative	White	Female
68	Federal-gov	422013	HS-grad	9	Divorced	Prof-specialty	Not-in-family	White	Female
41	Private	70037	Some-college	10	Never-married	Craft-repair	Unmarried	White	Male

Bar chart

- geom_bar() vs. geom_col()
- count vs. identity vs. summary
- categorical vs. continuous
- stack vs. dodge vs. fill
- bar vs. pie

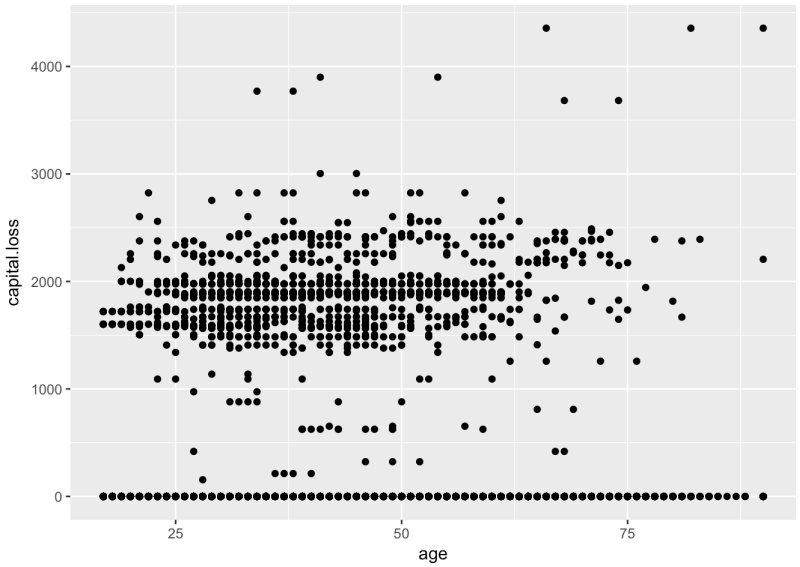


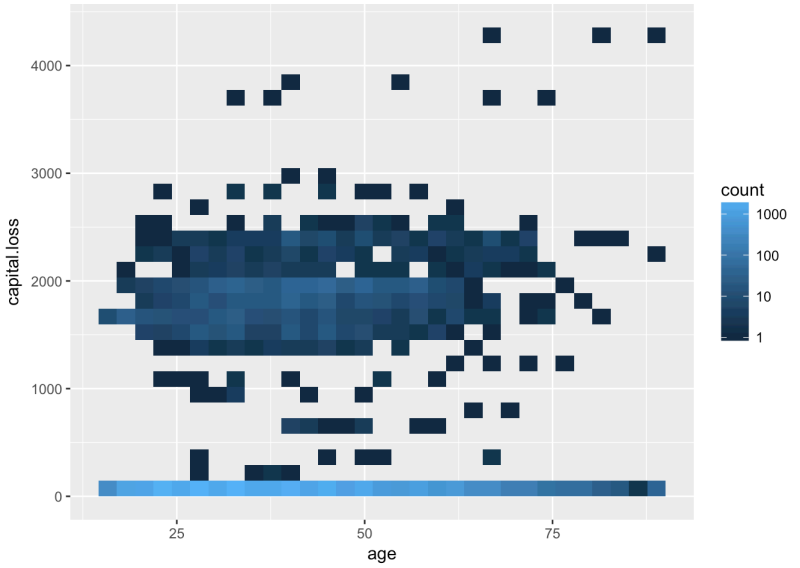




Scatter plot

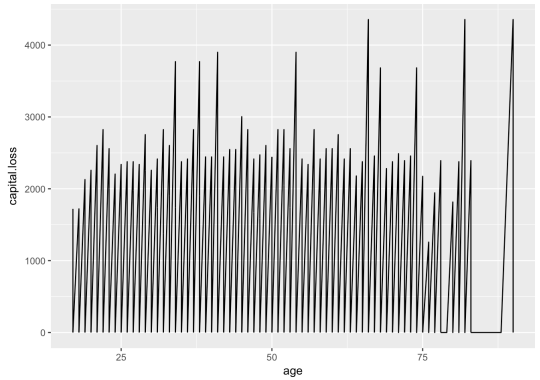
- Overplotting
- point vs. bin2d

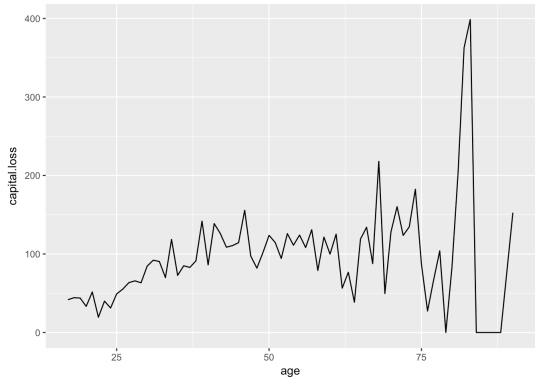


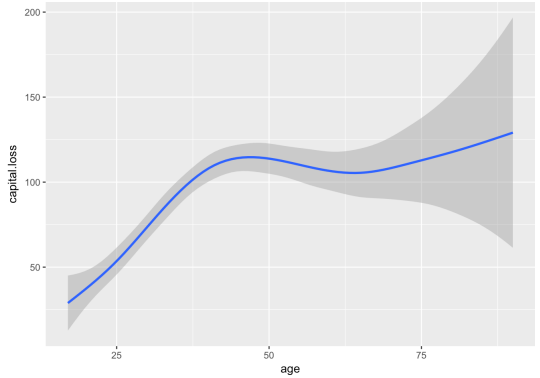


Line chart

- identity vs. summary
- line vs. smooth



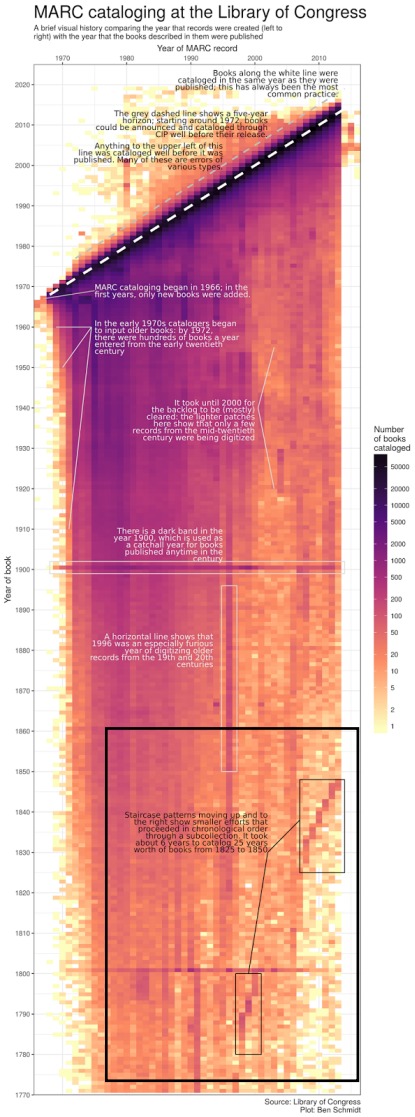


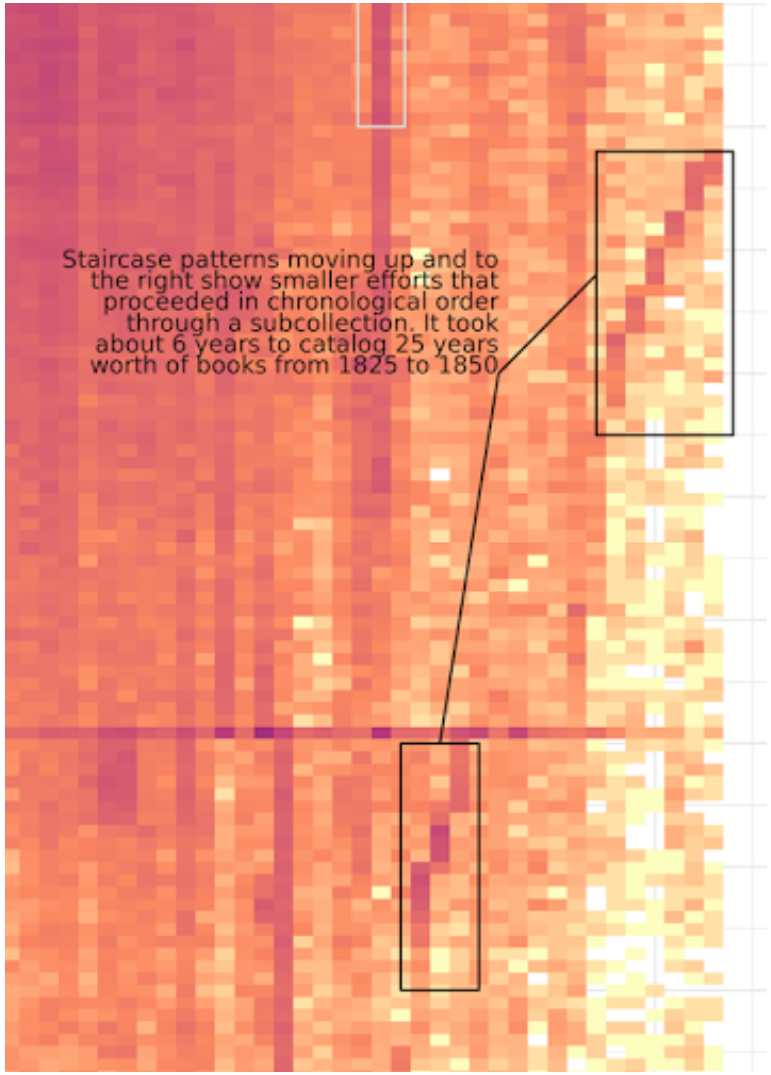


Plot exercise: Survey

Lunch

ggplot2: Design elements

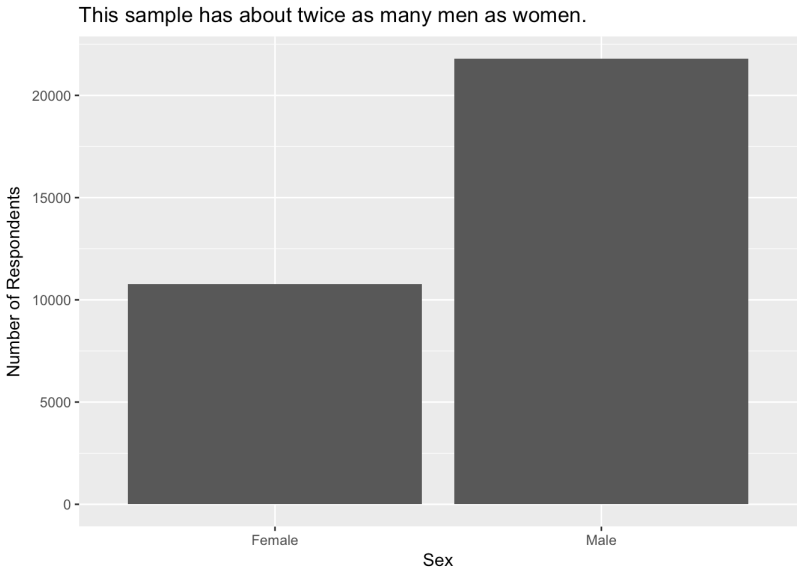


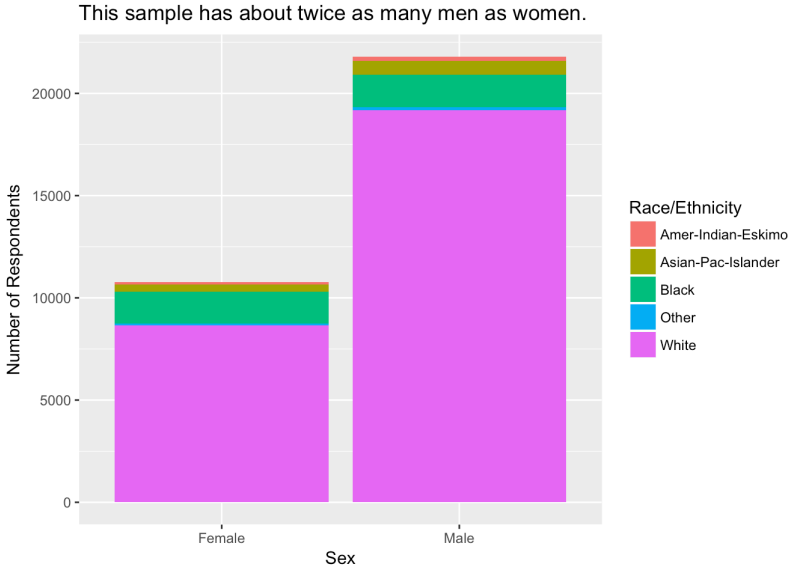


http://sappingattention.blogspot.com/2017/05/a-brief-visual-history-of-marc.html

Titles

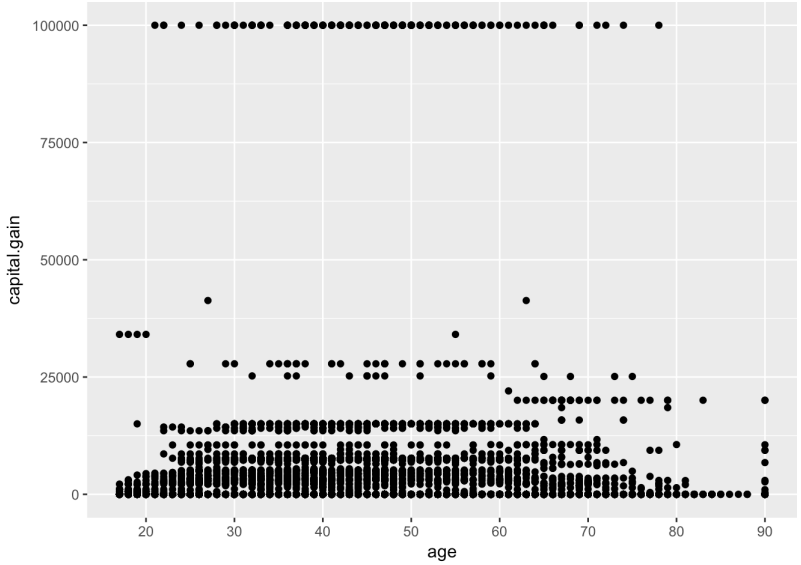
- Main title
- x-axis
- y-axis
- legend

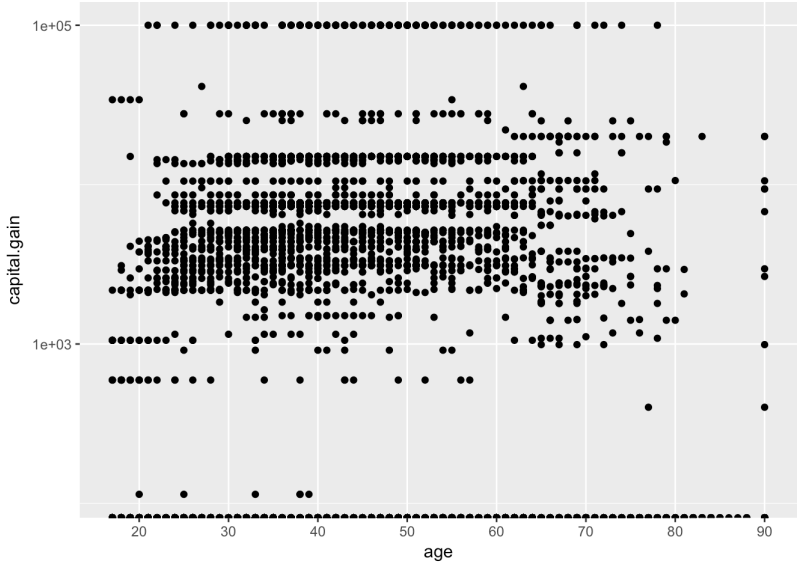




Axes

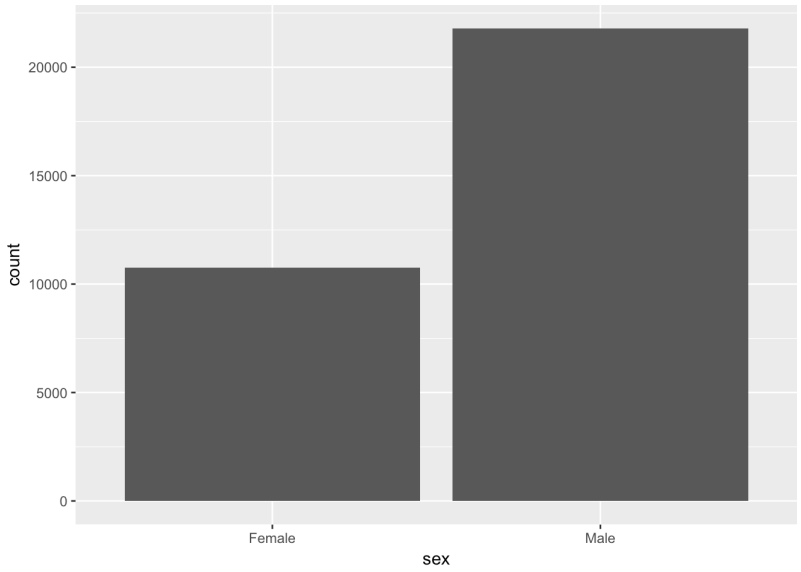
- limits
- major breaks
- minor breaks
- gridlines
- scales vs. coordinates

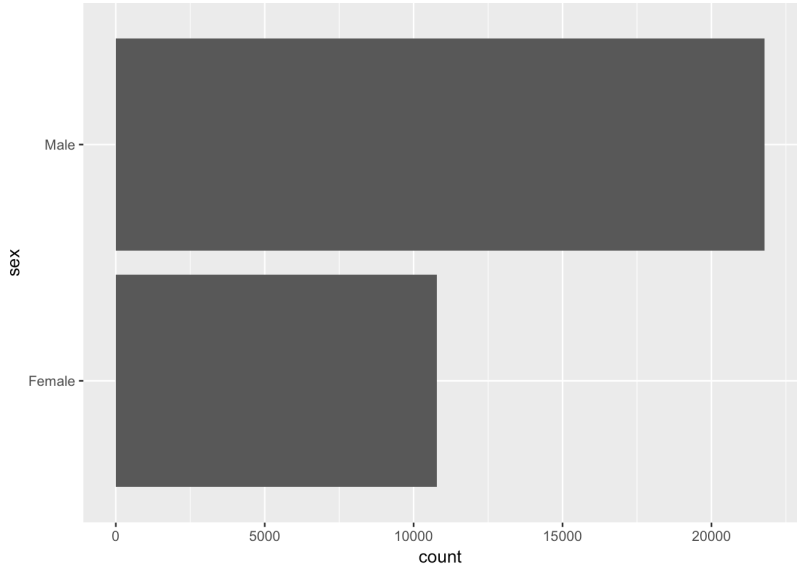




Coordinate systems

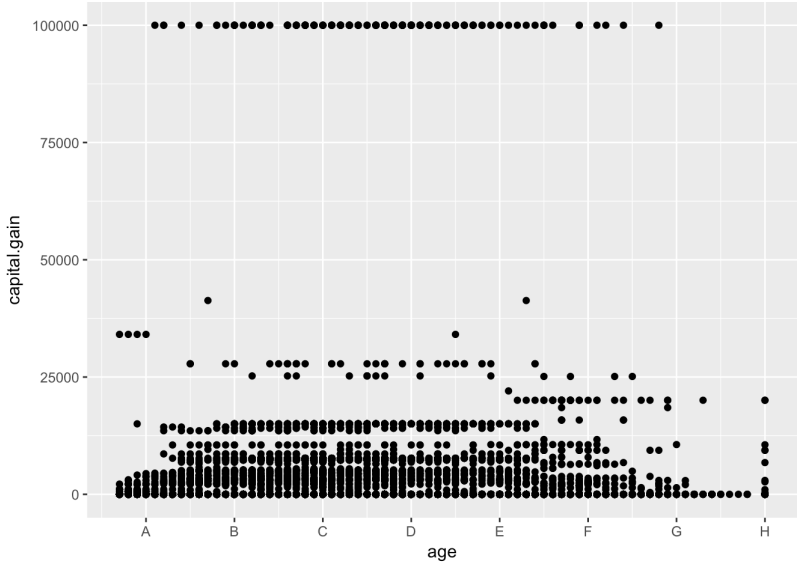
- fixed
- flip
- trans

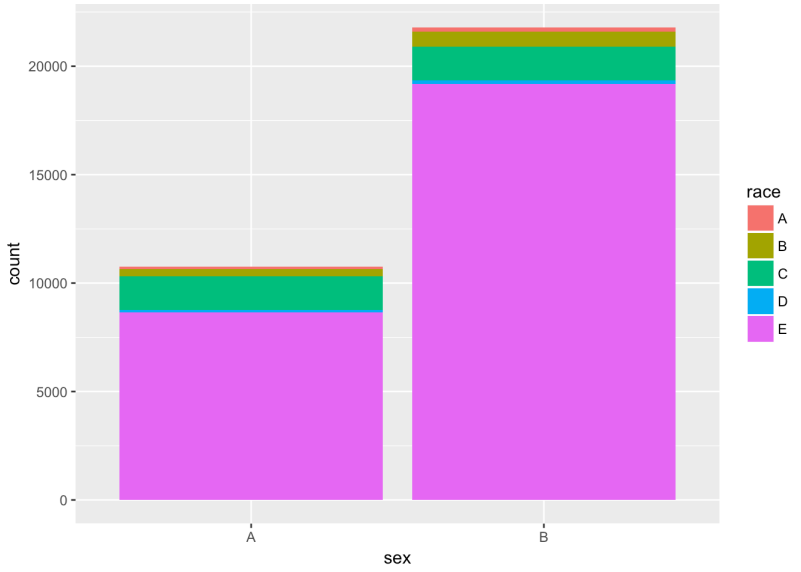




Labeling

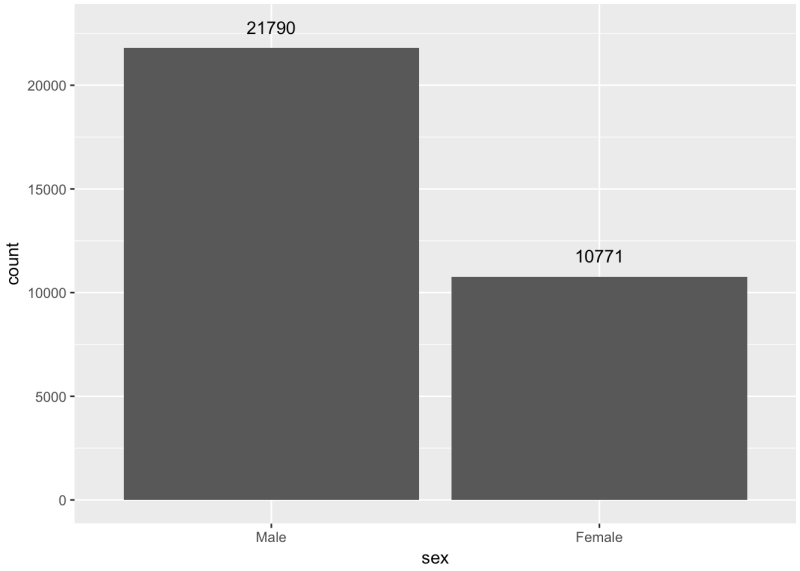
- axis/legend labels
- factors

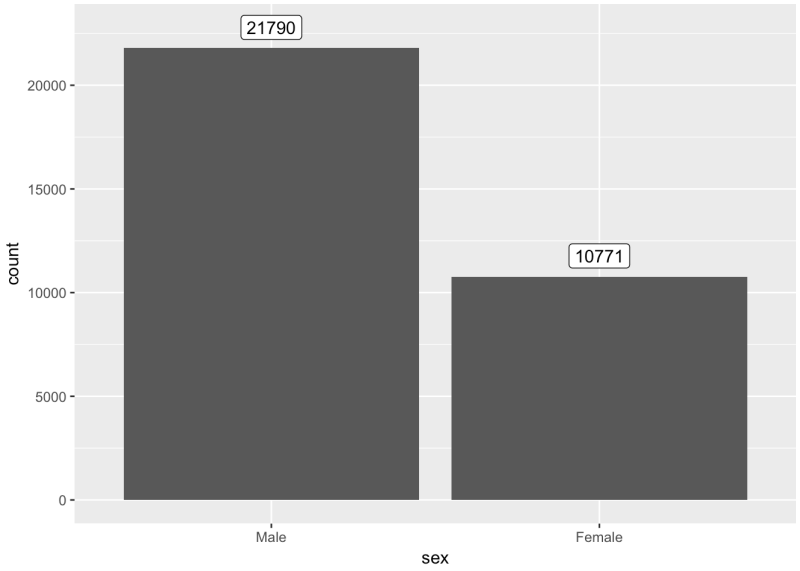




Data labels

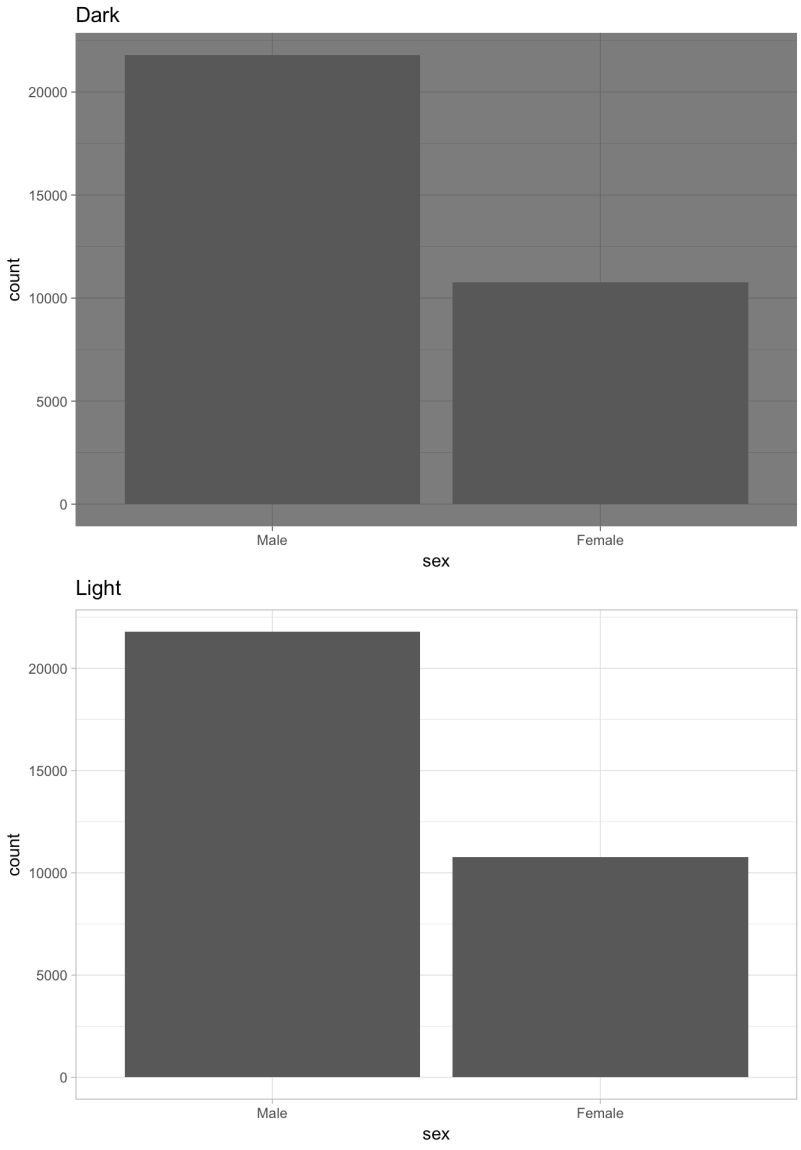
- position
- calculated values
- geom_text vs. geom_label





Themes

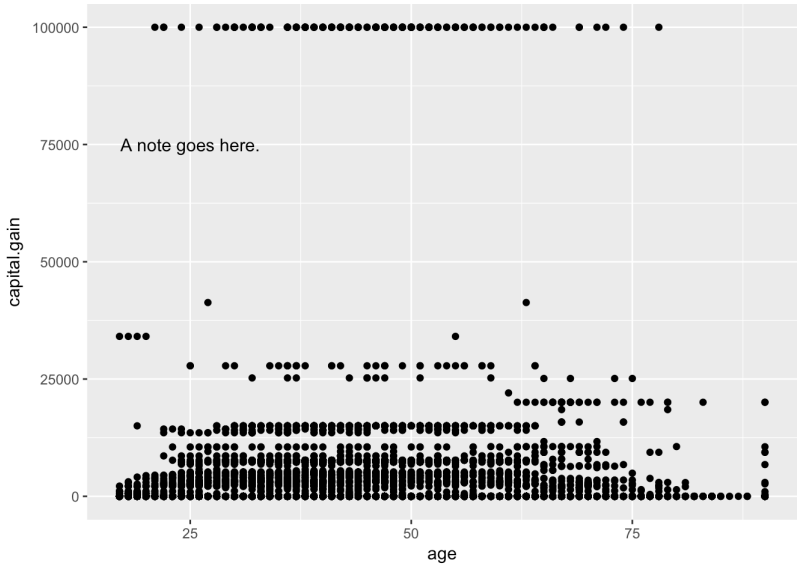
- Built-in themes
- Custom themes

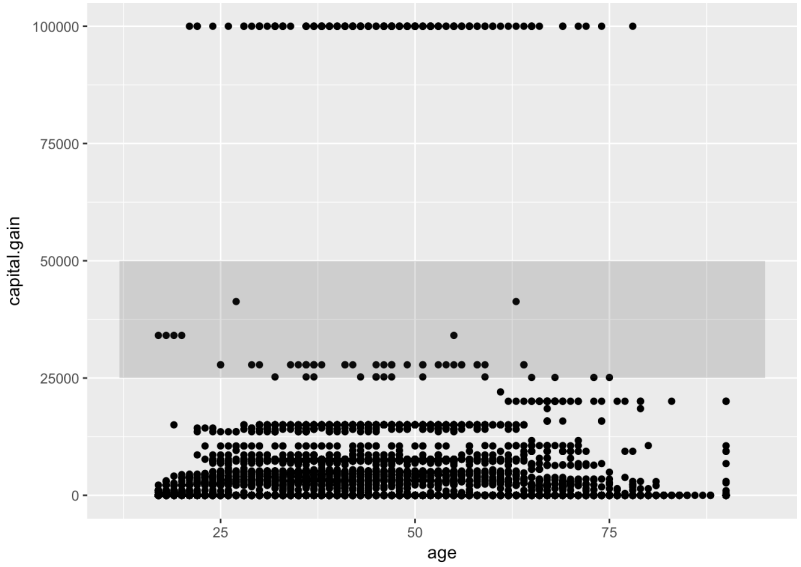


http://ggplot2.tidyverse.org/reference/theme.html
http://ggplot2.tidyverse.org/reference/#section-themes

Annotation

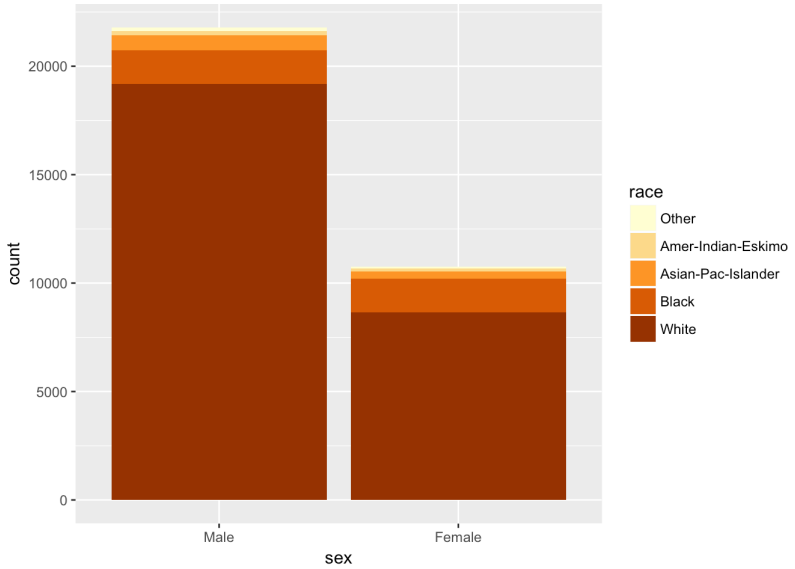
- text vs. annotate

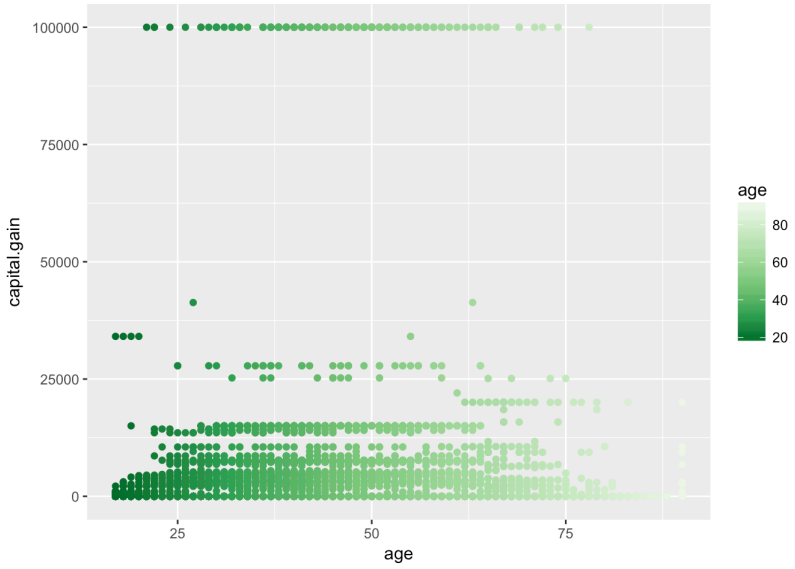




Colors

- continuous vs. discrete

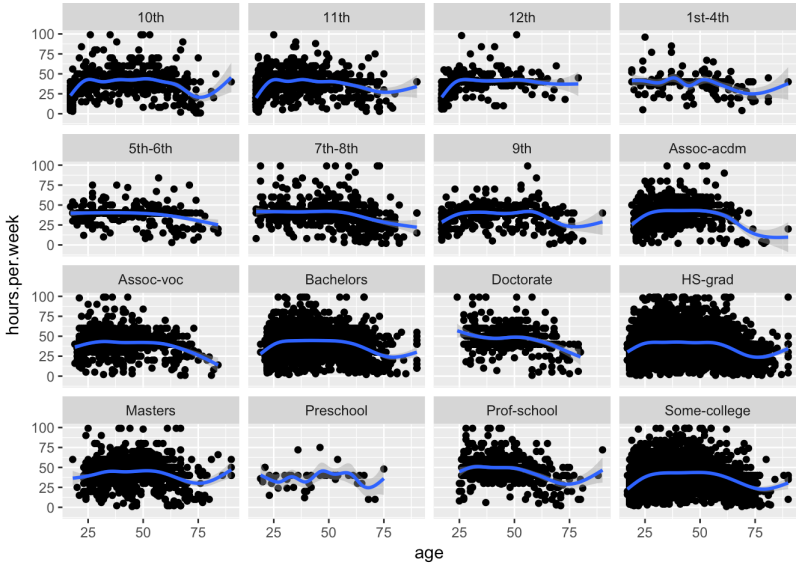


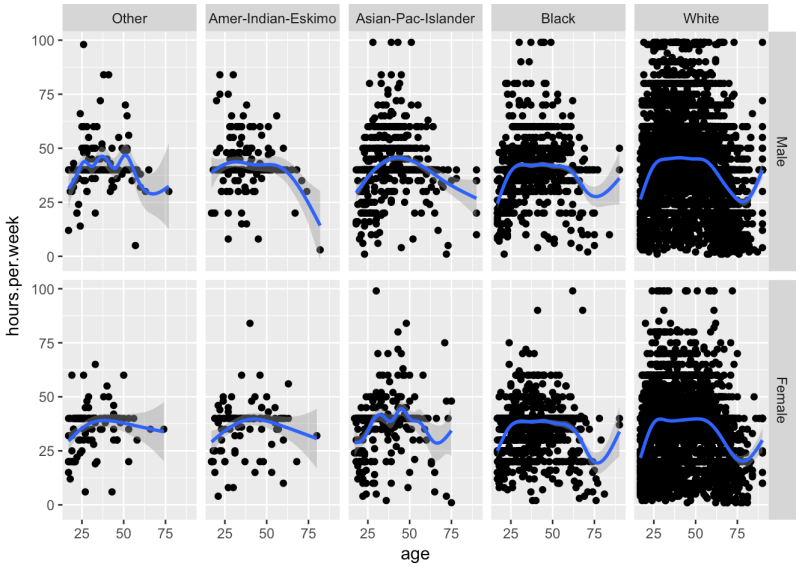


http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/
http://ggplot2.tidyverse.org/reference/scale_brewer.html

Facets

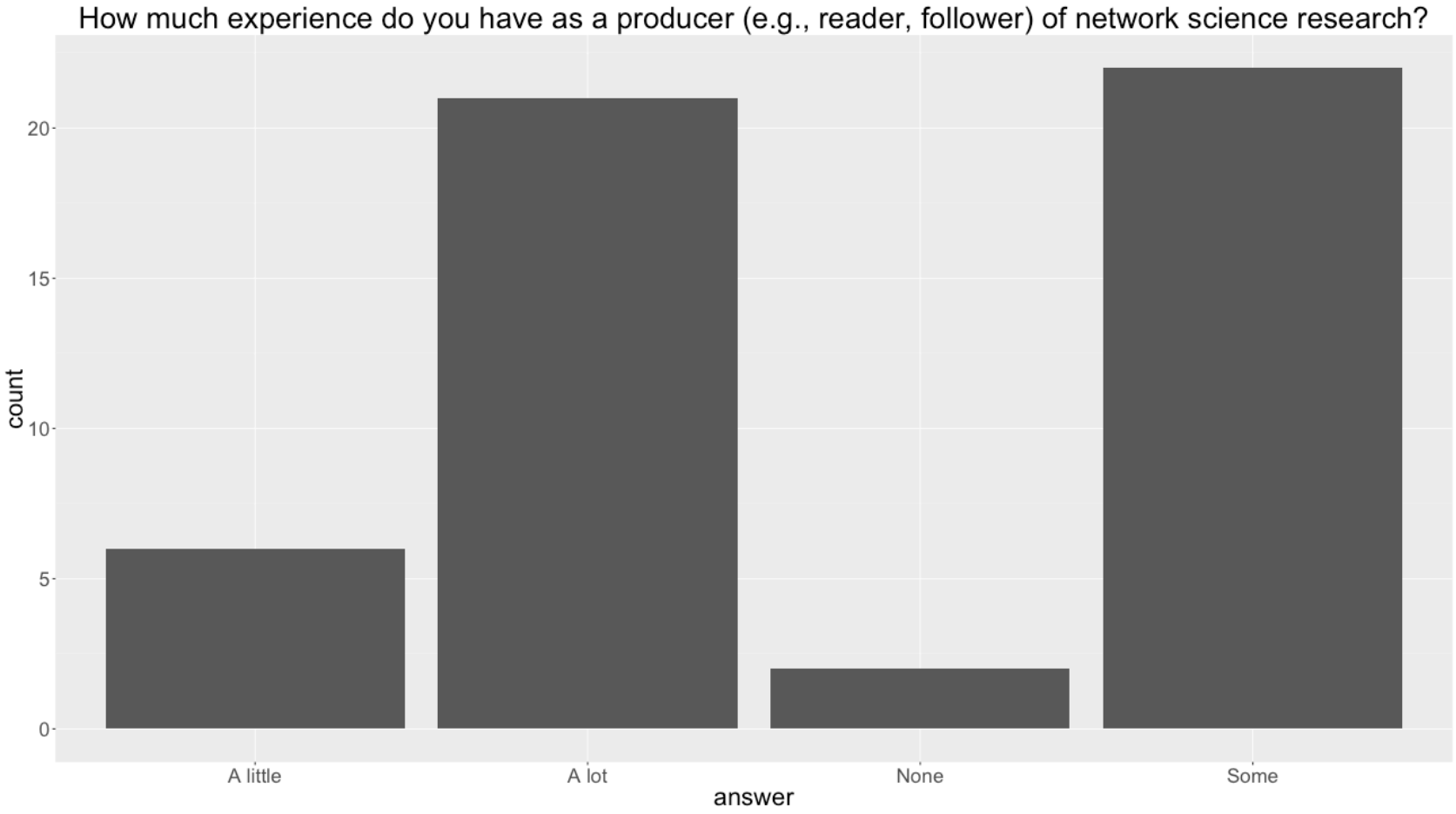
- grid vs. wrap





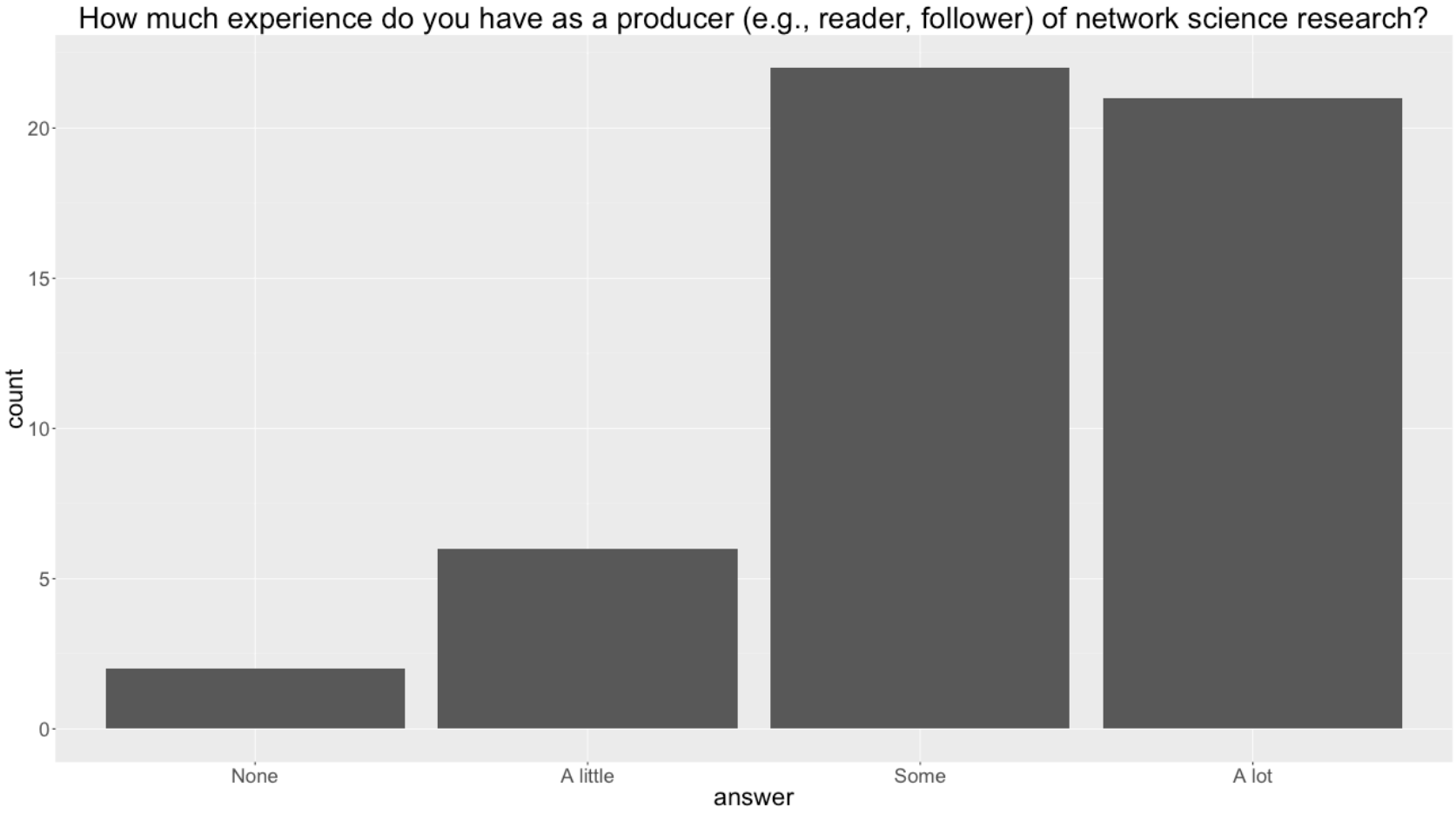
Principles for Effective Visualizations

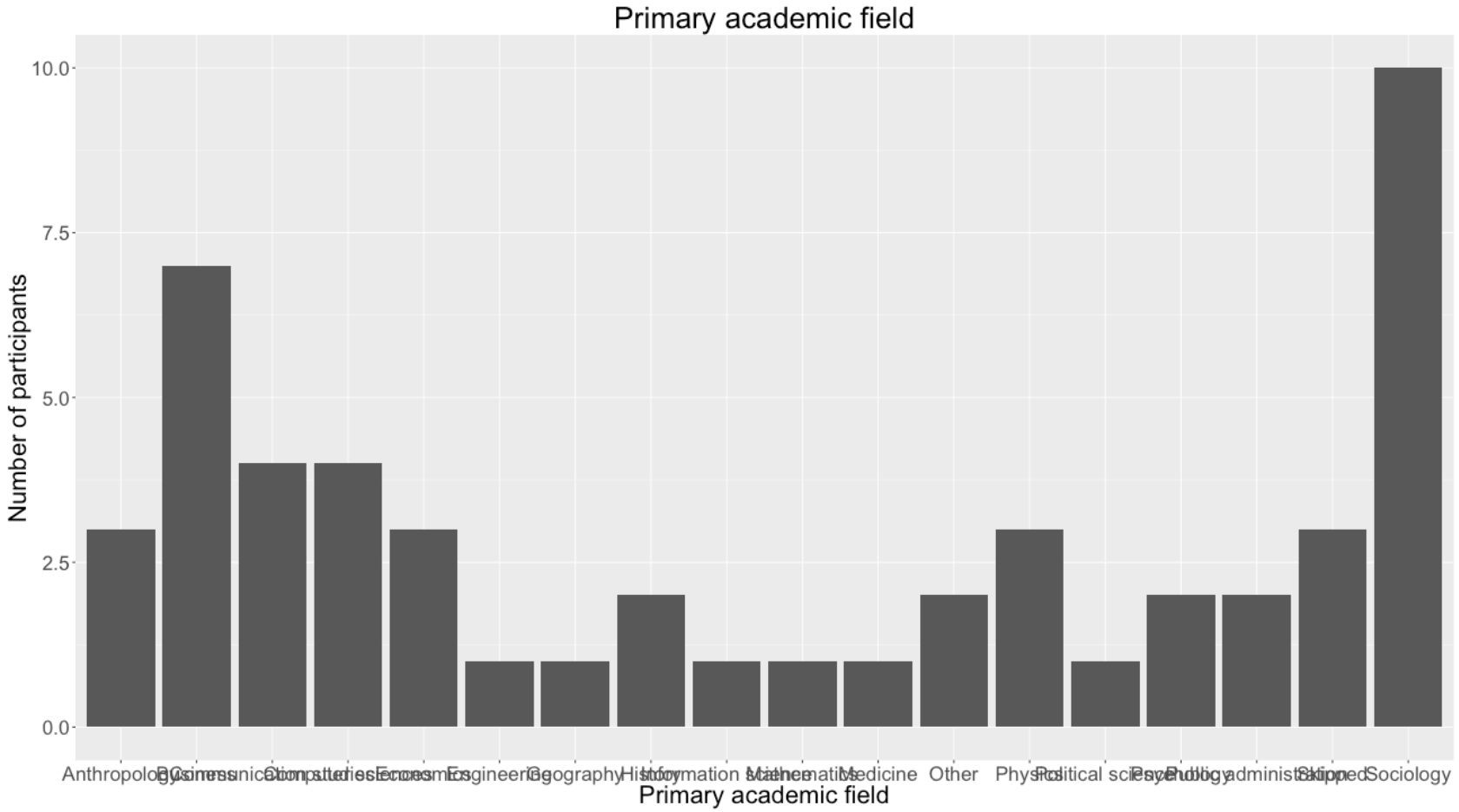
Principle 1: Order matters


How much experience do you have as a producer (e.g., reader, follower) of network science research?
count
20
15
10
5
0
A little
A lot
None
Some
answer

Order by meaning

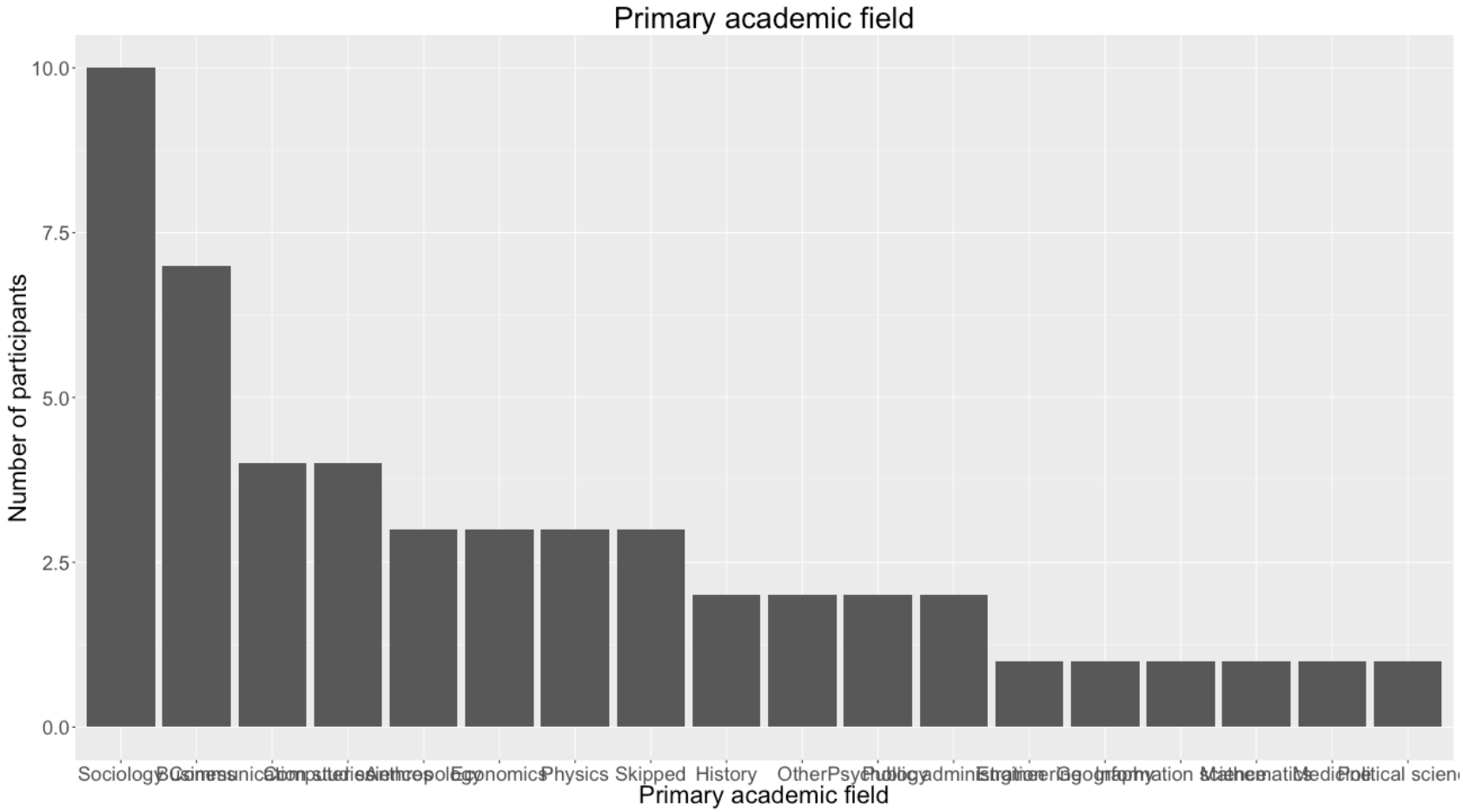
```
data$answer <-
    factor(data$answer,
           levels=c("None", "A little", "Some", "A lot"),
           ordered = TRUE)
```


How much experience do you have as a producer (e.g., reader, follower) of network science research?
count
20
15
10
5
0
None
A little
Some
A lot
answer

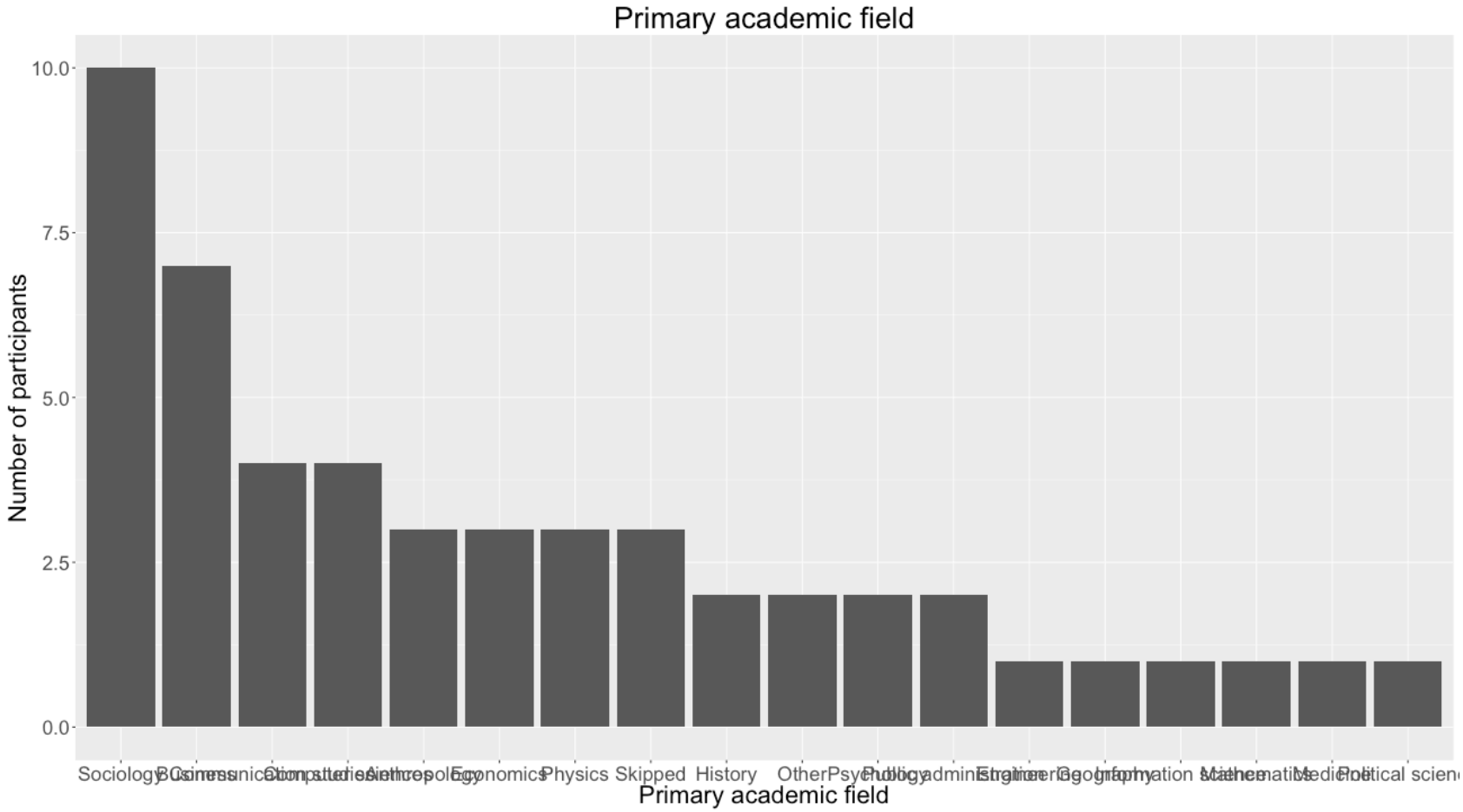

Primary academic field
Number of participants
10.0
7.5
5.0
2.5
0.0
Anthropology
Business
Communication
Computer science
Economics
Engineering
Geography
History
Information science
Mathematics
Medicine
Other
Physics
Political science
Psychology
Public administration
Sociology
Primary academic field

Order by value

```
data$academic_field <-
     factor(data$academic_field,
           levels=names(
               sort(
                   table(
                       data$academic_field),decreasing=TRUE)))
```

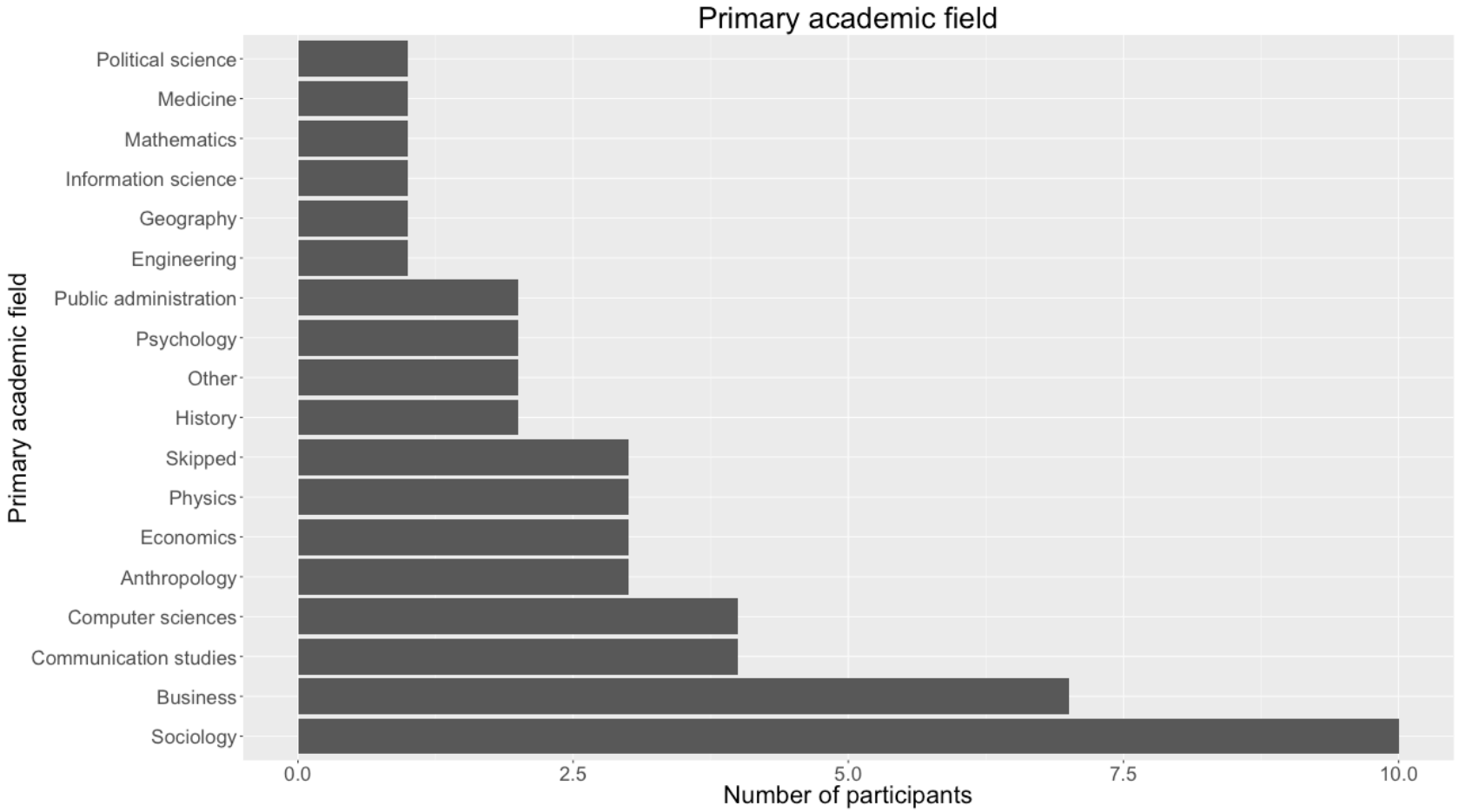

Primary academic field
Number of participants
10.0
7.5
5.0
2.5
0.0
Sociology
Business
Communication
Computer science
Anthropology
Economics
Physics
Skipped
History
Other
Psychology
Public administration
Engineering
Geography
Information science
Mathematics
Medicine
Political scien
Primary academic field

Principle 2: Put long categories on y-axis


Primary academic field
Number of participants
10.0
7.5
5.0
2.5
0.0
Sociology
Business
Communication
Computer science
Anthropology
Economics
Physics
Skipped
History
Other
Psychology
Public administration
Engineering
Geography
Information science
Mathematics
Medicine
Political scien
Primary academic field

Flip the axes

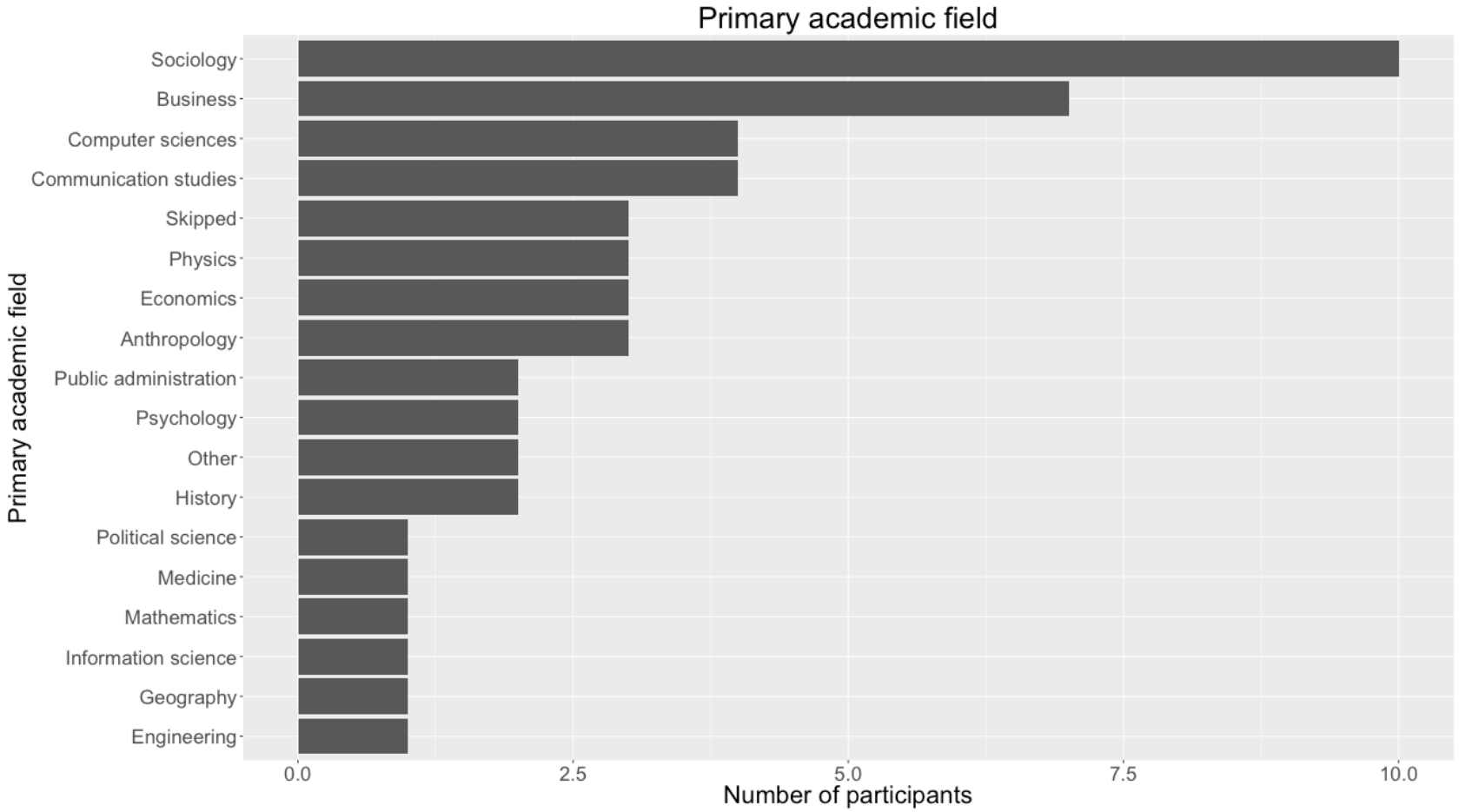
```
coord_flip()
```


Primary academic field
Political science
Medicine
Mathematics
Information science
Geography
Engineering
Public administration
Psychology
Other
History
Skipped
Physics
Economics
Anthropology
Computer sciences
Communication studies
Business
Sociology
Primary academic field
0.0
2.5
5.0
7.5
10.0
Number of participants

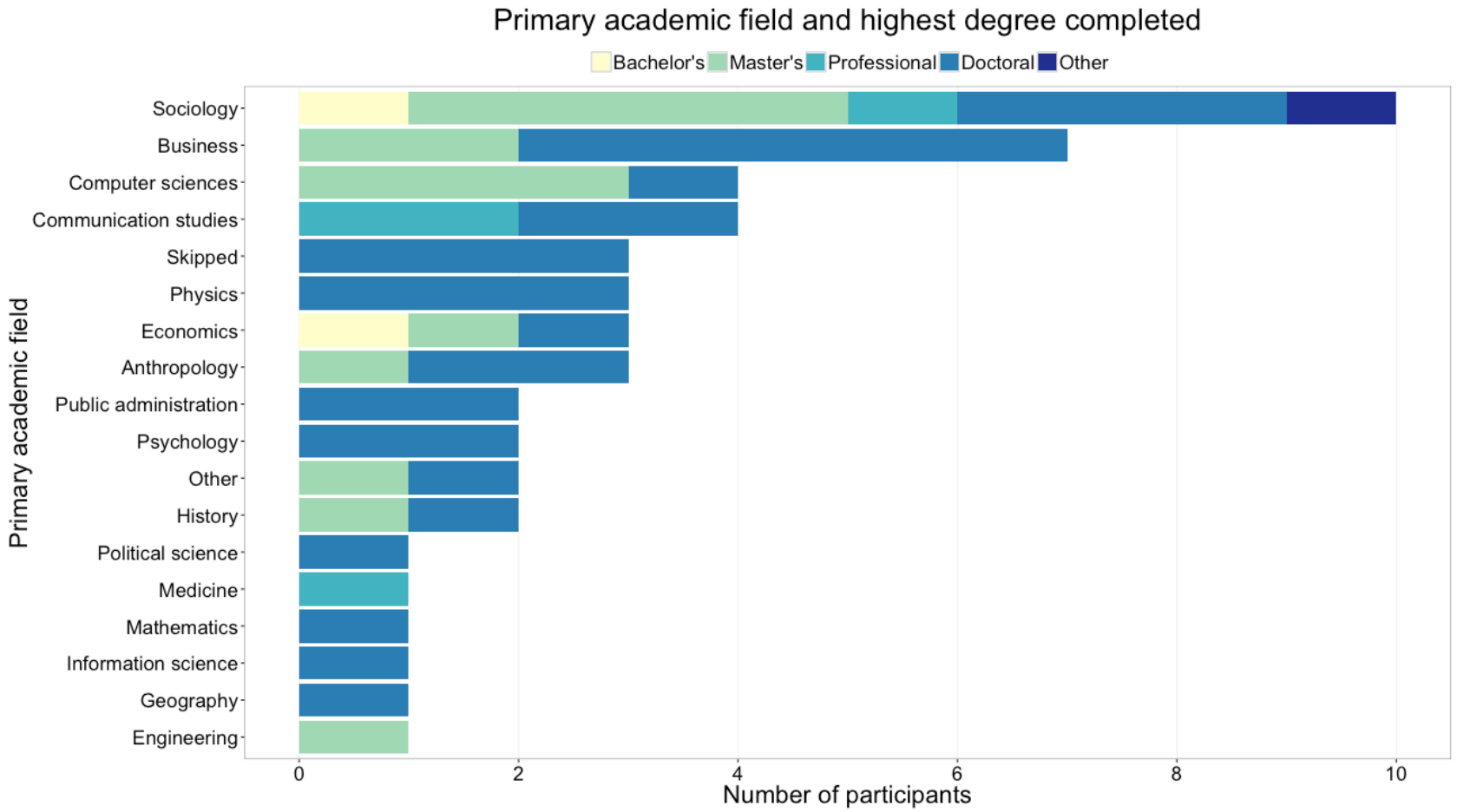
Oops!

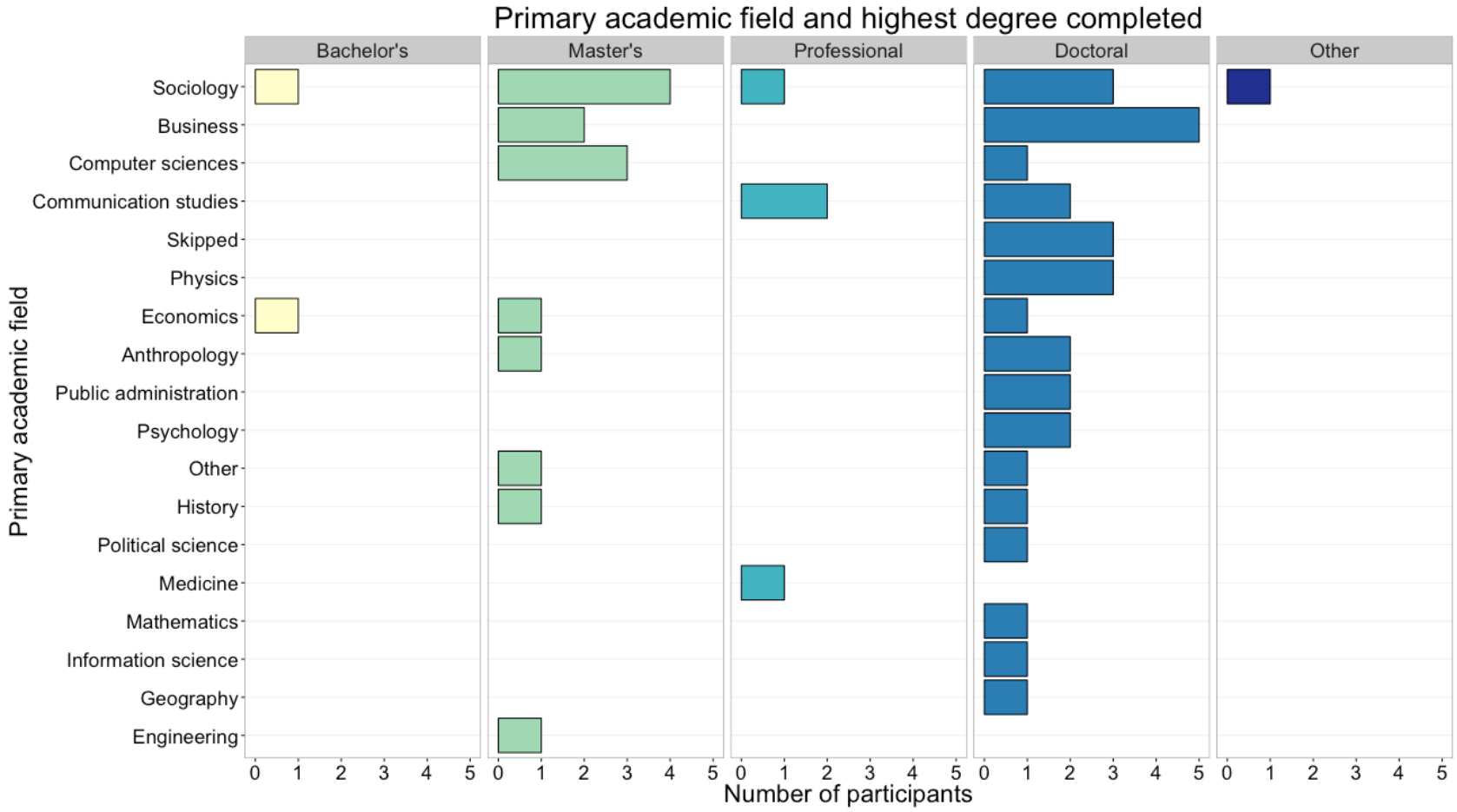
```
data$academic_field <-
     factor(data$academic_field,
           levels=names(
               sort(
                   table(data$academic_field),
                   decreasing=TRUE)))
```

```
data$academic_field <-
     factor(data$academic_field,
           levels=names(
               sort(
                   table(data$academic_field))))
```


Primary academic field
Sociology
Business
Computer sciences
Communication studies
Skipped
Physics
Economics
Anthropology
Public administration
Psychology
Other
History
Political science
Medicine
Mathematics
Information science
Geography
Engineering
Primary academic field
0.0
2.5
5.0
7.5
10.0
Number of participants

Principle 3: Pick a purpose


Primary academic field and highest degree completed
Bachelor's
Master's
Professional
Doctoral
Other
Primary academic field
Sociology
Business
Computer sciences
Communication studies
Skipped
Physics
Economics
Anthropology
Public administration
Psychology
Other
History
Political science
Medicine
Mathematics
Information science
Geography
Engineering
0
2
4
6
8
10
Number of participants

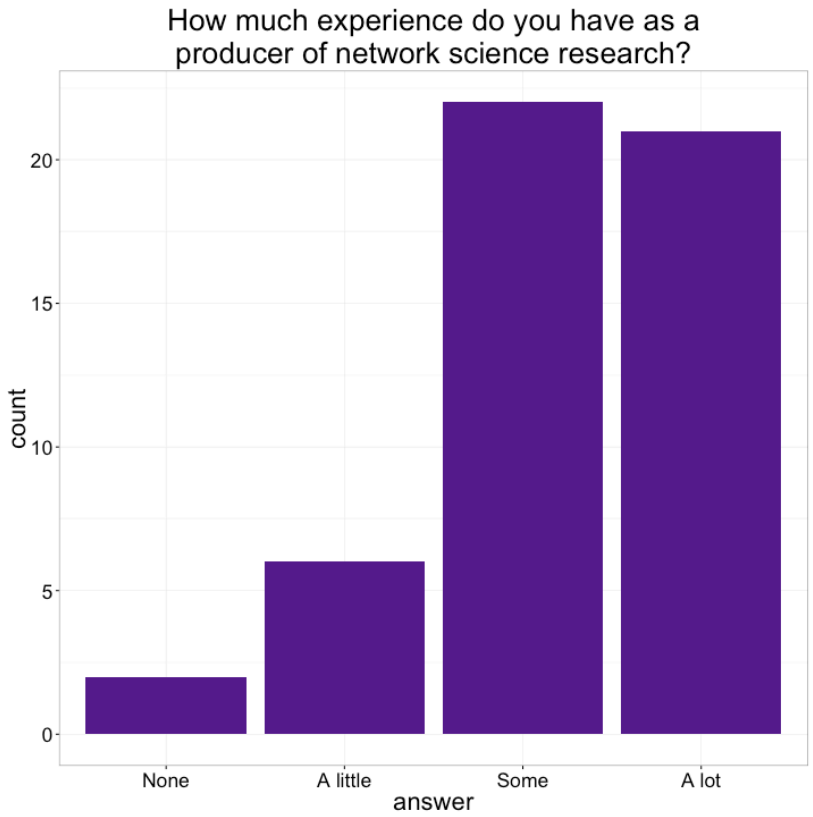

Primary academic field and highest degree completed
Bachelor's
Master's
Professional
Doctoral
Other
Primary academic field
Sociology
Business
Computer sciences
Communication studies
Skipped
Physics
Economics
Anthropology
Public administration
Psychology
Other
History
Political science
Medicine
Mathematics
Information science
Geography
Engineering
0 1 2 3 4 5
Number of participants

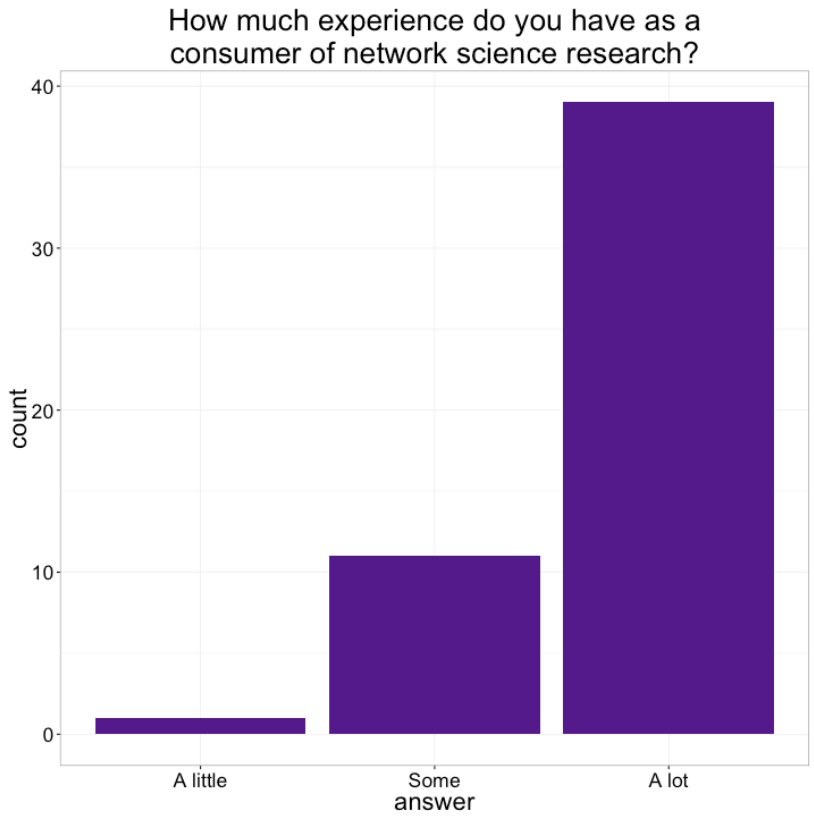
Different placement helps with different comparisons

```
fill=highest_degree
```

```
facet_grid(.~highest_degree)
```

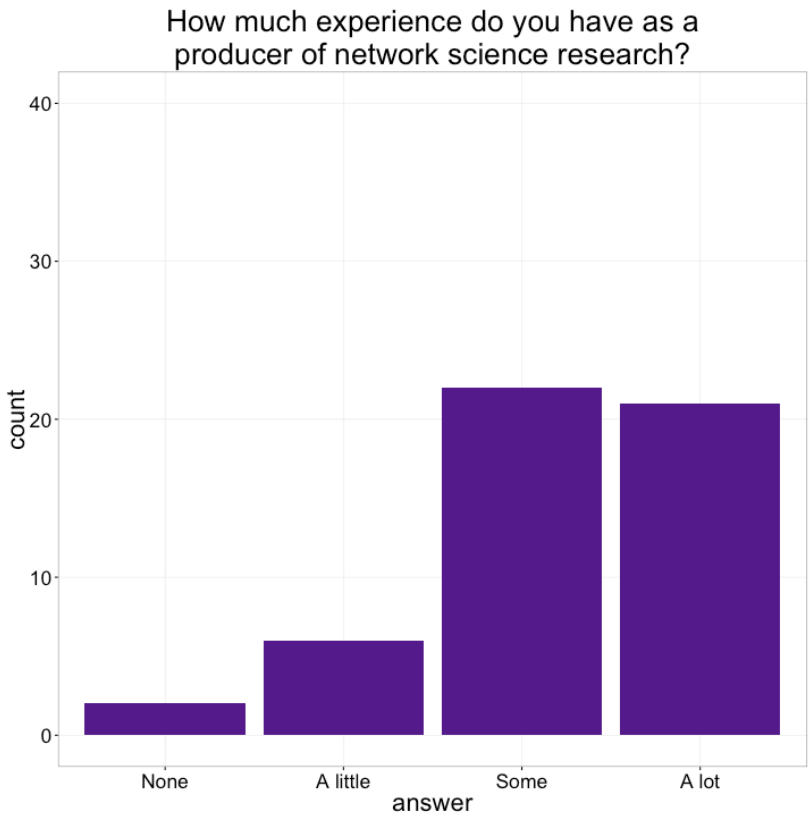
Principle 4: Keep scales consistent

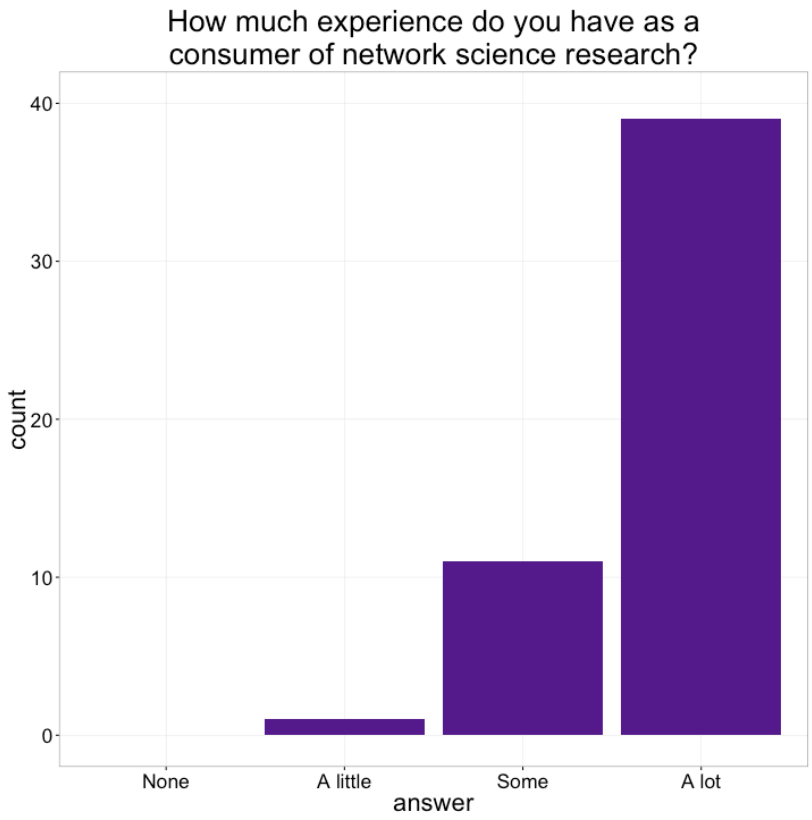

How much experience do you have as a producer of network science research?
count
0
5
10
15
20
None
A little
Some
A lot
answer


How much experience do you have as a consumer of network science research?
count
0
10
20
30
40
A little
Some
A lot
answer

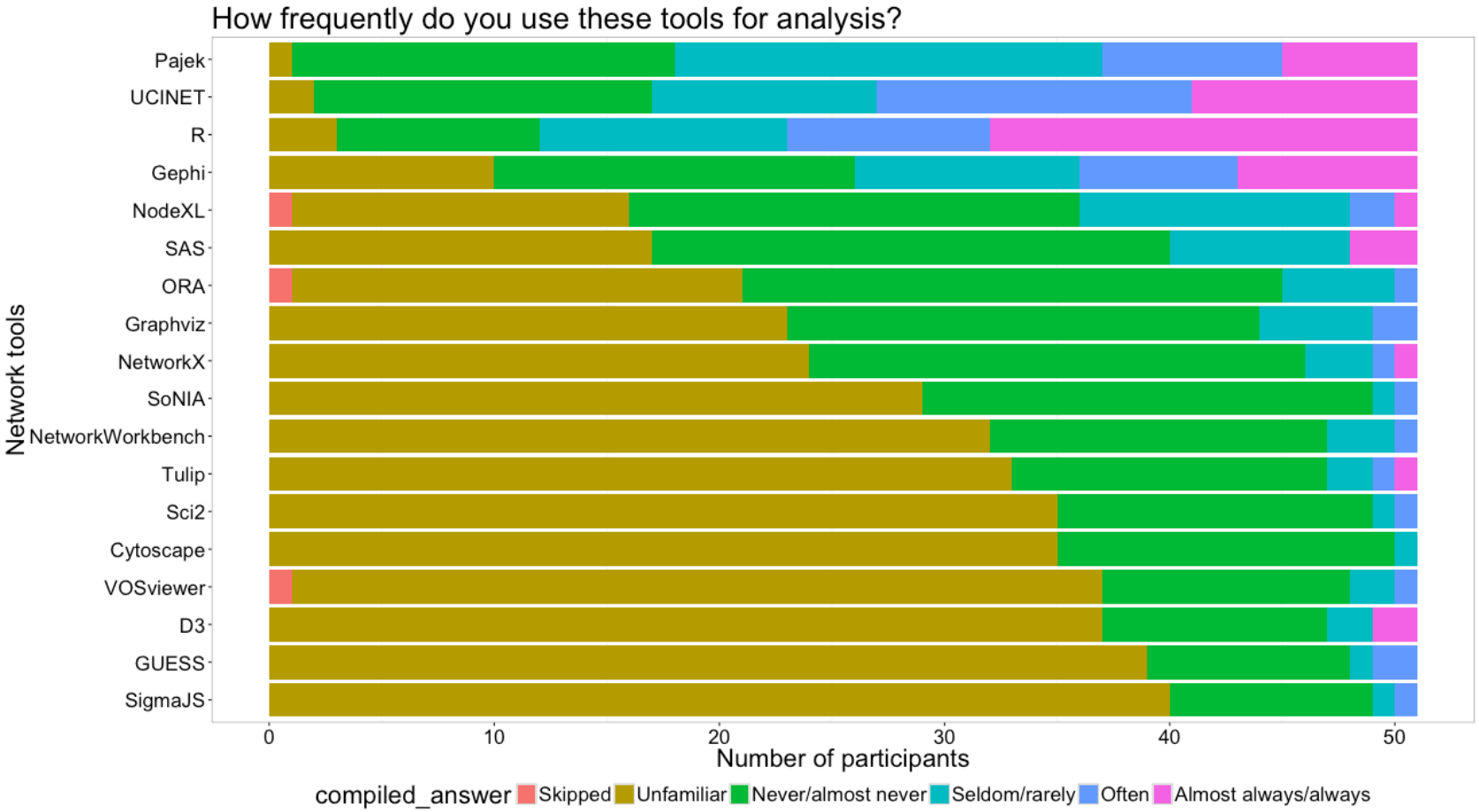
Keep all categories, manually set axes

```
scale_x_discrete(drop=FALSE)

scale_y_continuous(limits=c(0,40),
                   breaks=c(0,10,20,30,40),
                   minor_breaks=NULL)
```


How much experience do you have as a producer of network science research?
count
40
30
20
10
0
None
A little
Some
A lot
answer

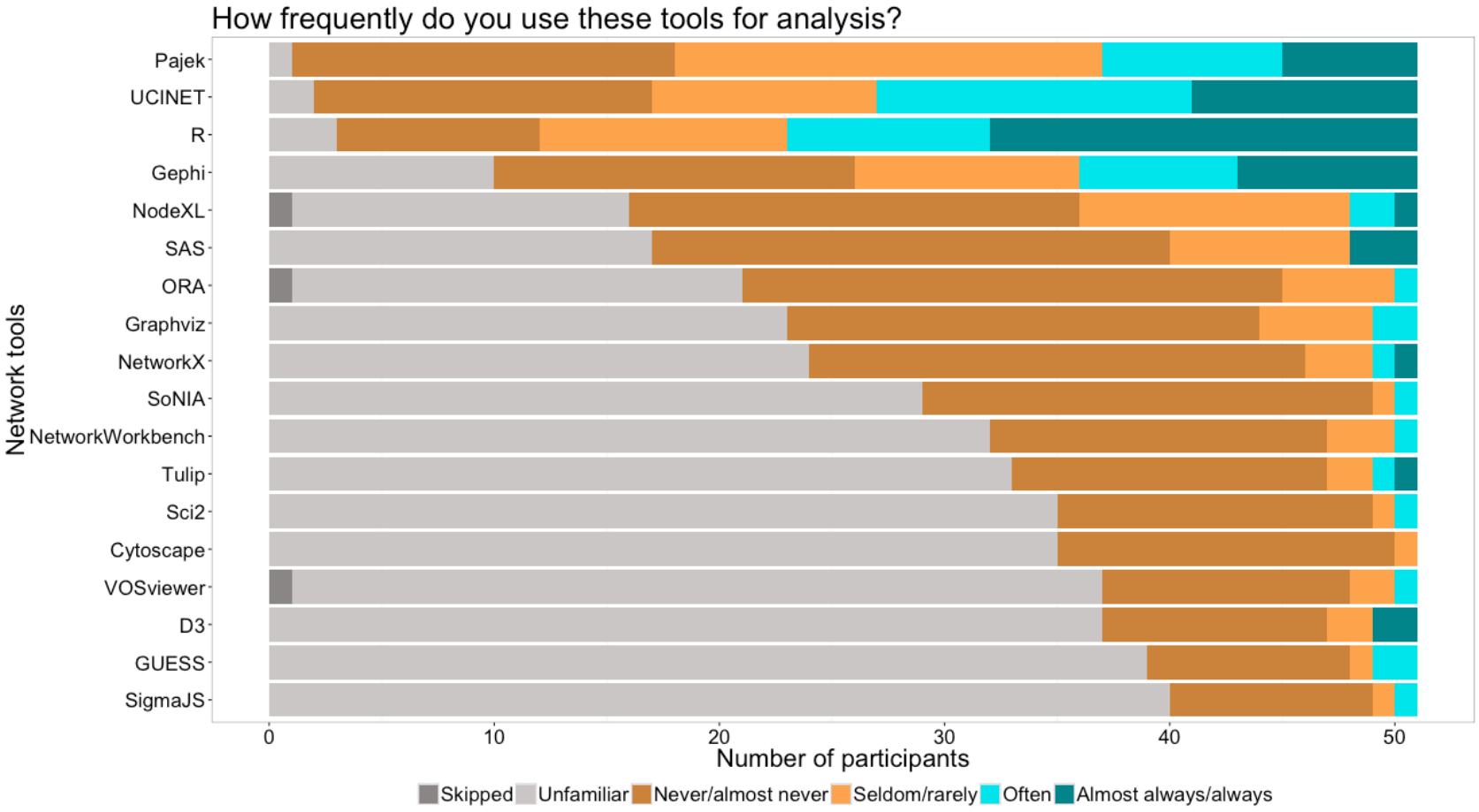

How much experience do you have as a consumer of network science research?
count
40
30
20
10
0
None
A little
Some
A lot
answer

Principle 5: Select meaningful colors


How frequently do you use these tools for analysis?
Network tools
Pajek
UCINET
R
Gephi
NodeXL
SAS
ORA
Graphviz
NetworkX
SoNIA
NetworkWorkbench
Tulip
Sci2
Cytoscape
VOSviewer
D3
GUESS
SigmaJS
0
10
20
30
40
50
Number of participants
compiled_answer
Skipped
Unfamiliar
Never/almost never
Seldom/rarely
Often
Almost always/always

Select colors manually, or use alternate palette

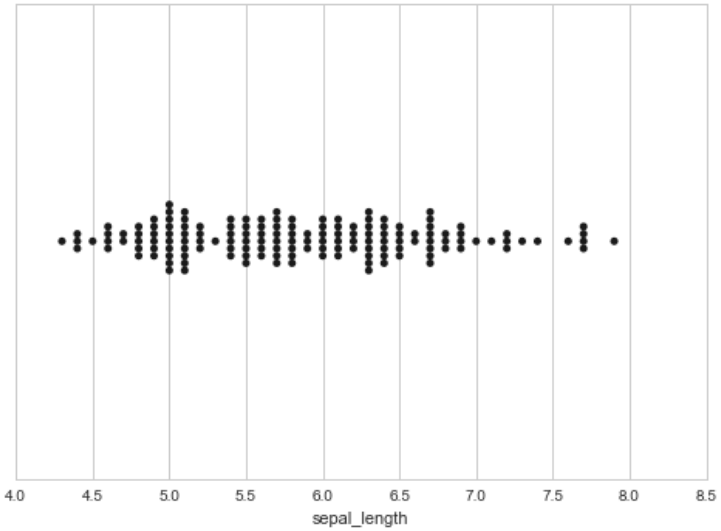
```
scale_fill_manual(
    values=c("snow4","snow3",
             "tan3","tan1",
             "turquoise2","turquoise4"))

scale_fill_manual(
    values=c("#fee391","#fe9929", "#cc4c02"))

# Also see package RColorBrewer
scale_fill_brewer(palette="BrBG")
```

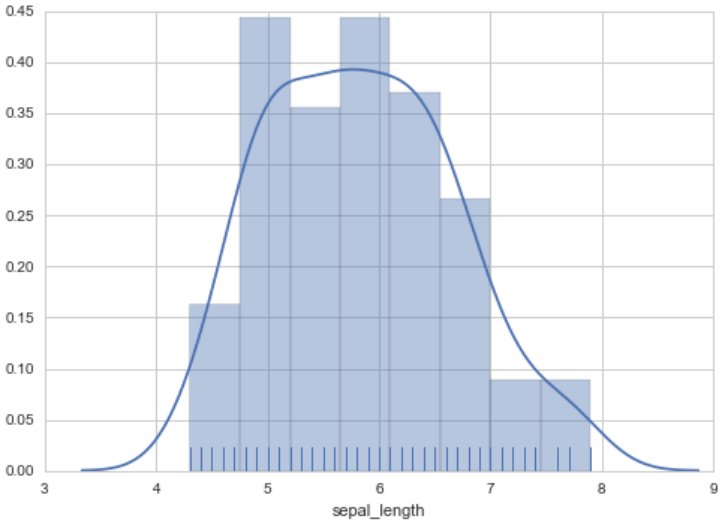

How frequently do you use these tools for analysis?
Network tools
Pajek
UCINET
R
Gephi
NodeXL
SAS
ORA
Graphviz
NetworkX
SoNIA
NetworkWorkbench
Tulip
Sci2
Cytoscape
VOSviewer
D3
GUESS
SigmaJS
0
10
20
30
40
50
Number of participants
Skipped
Unfamiliar
Never/almost never
Seldom/rarely
Often
Almost always/always

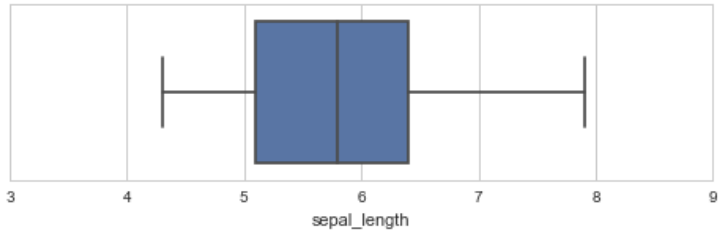
Afternoon break

Data exploration

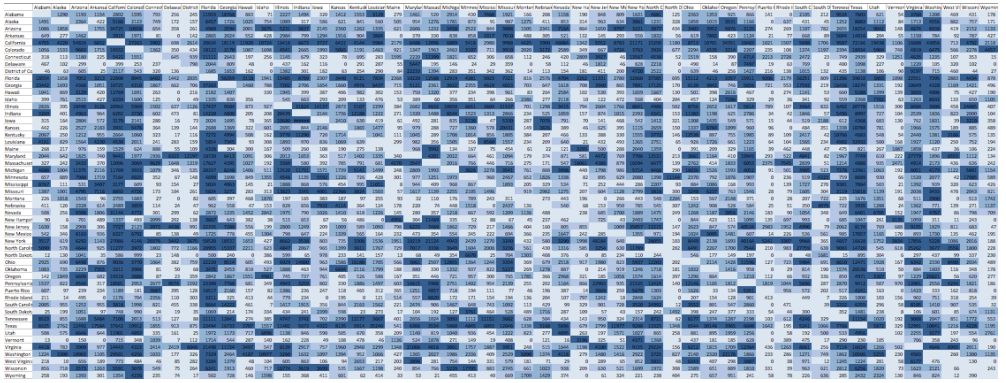
Exploring single variables

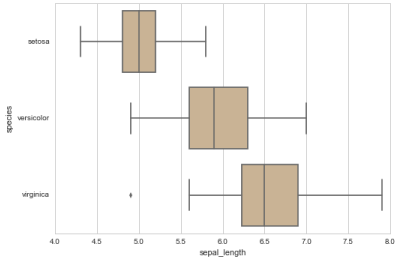


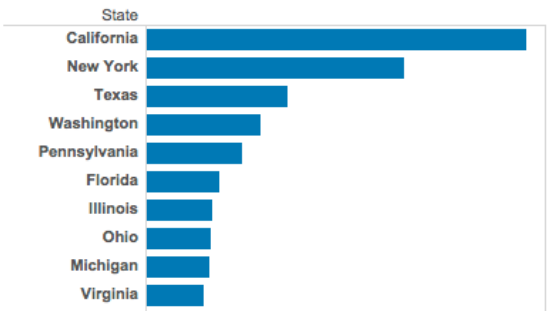


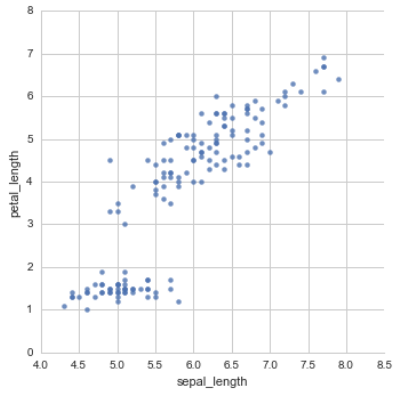


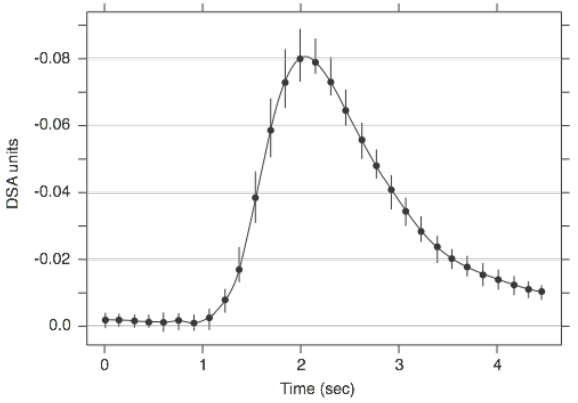
Exploring two variables



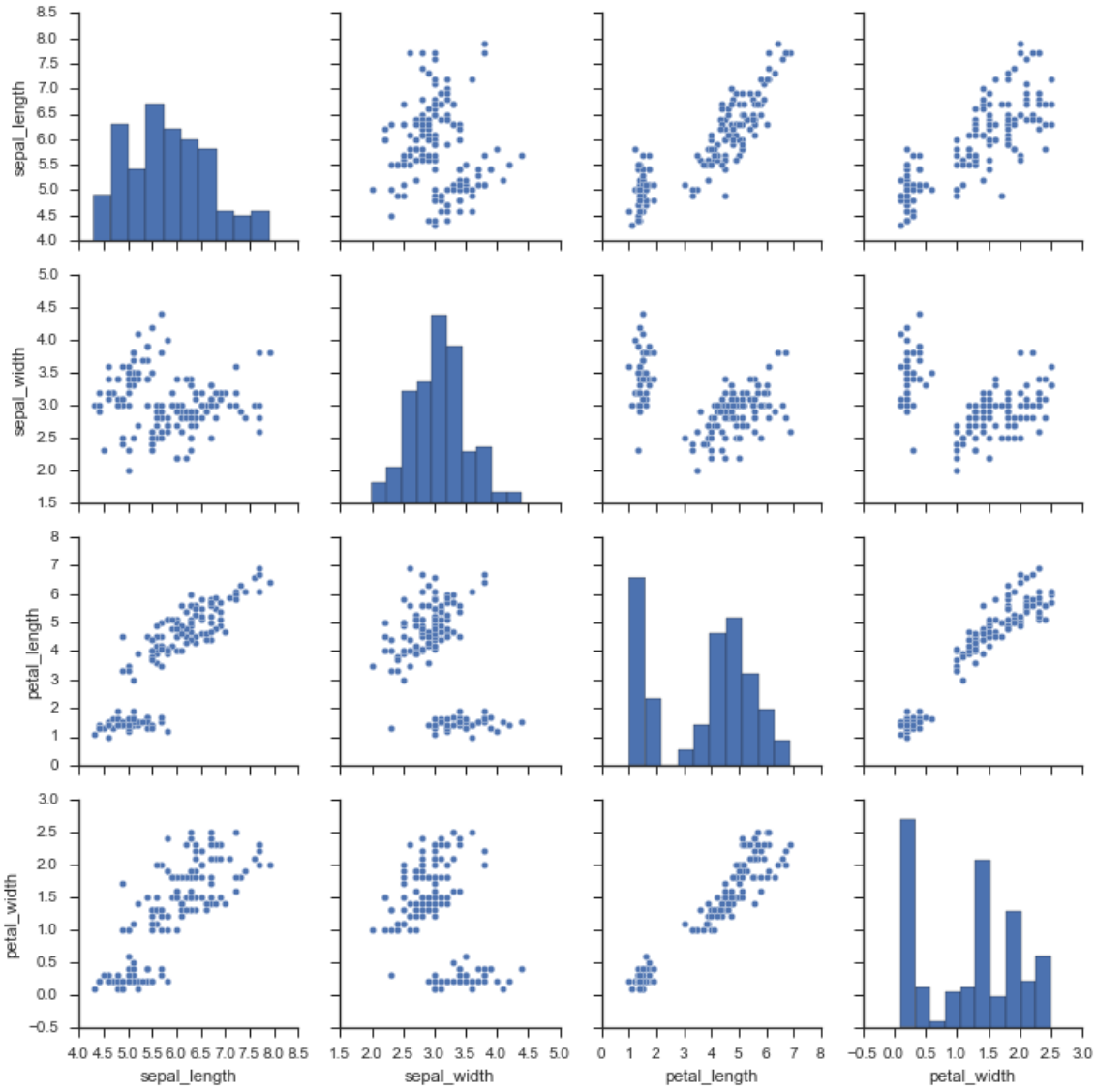




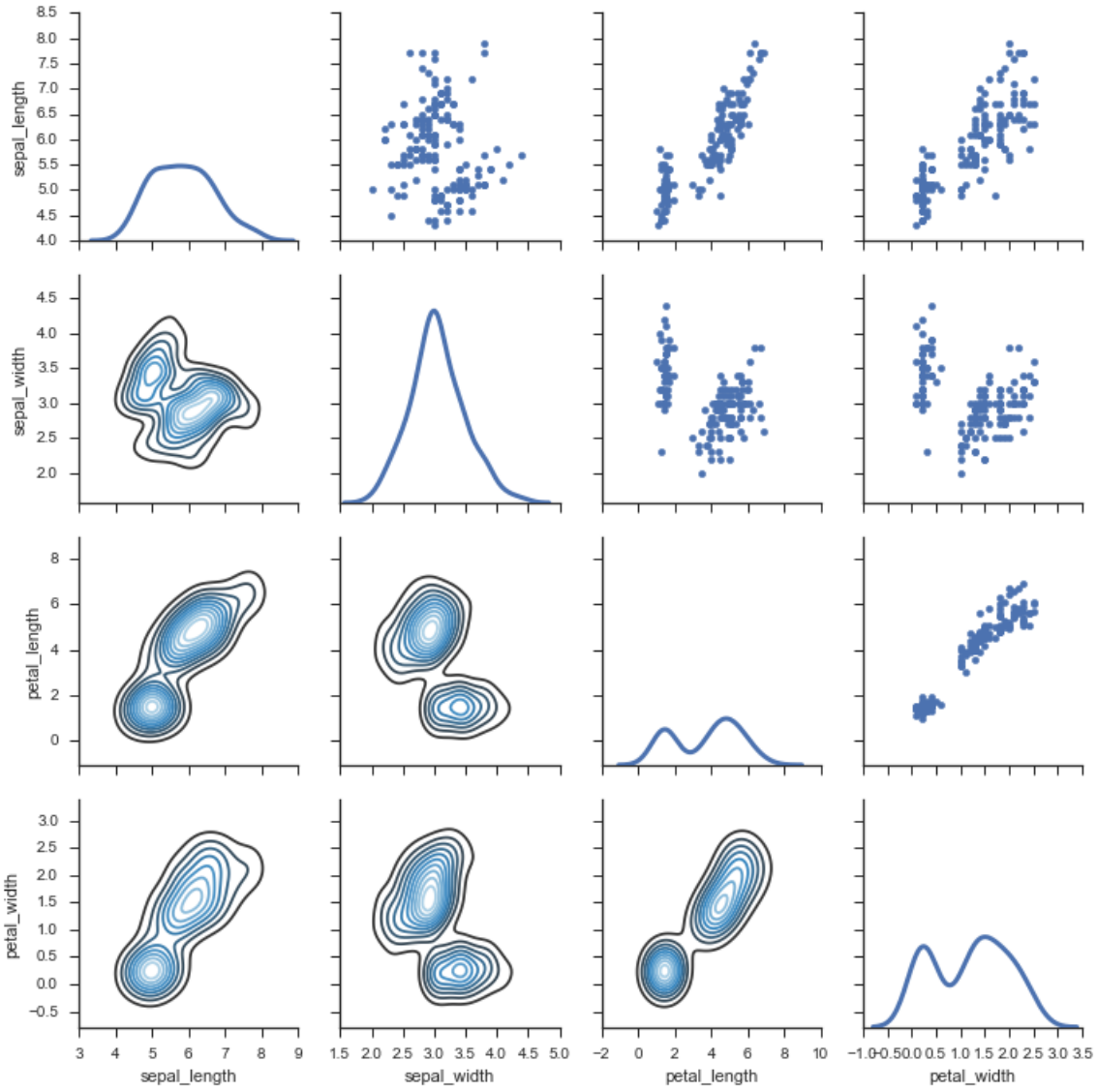




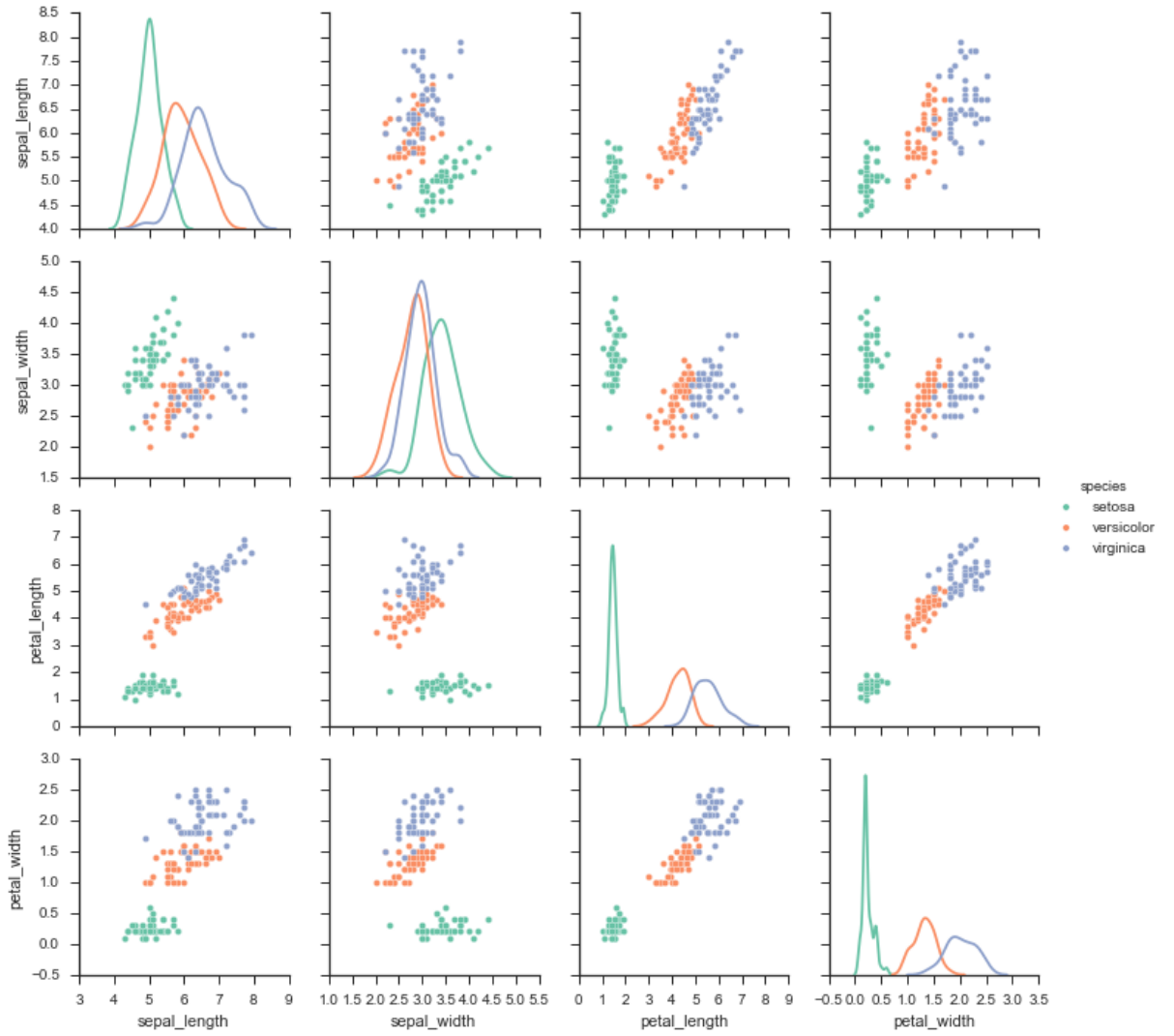
Exploring many variables, two at a time



Exploring many variables, two at a time



Exploring many variables, three at a time



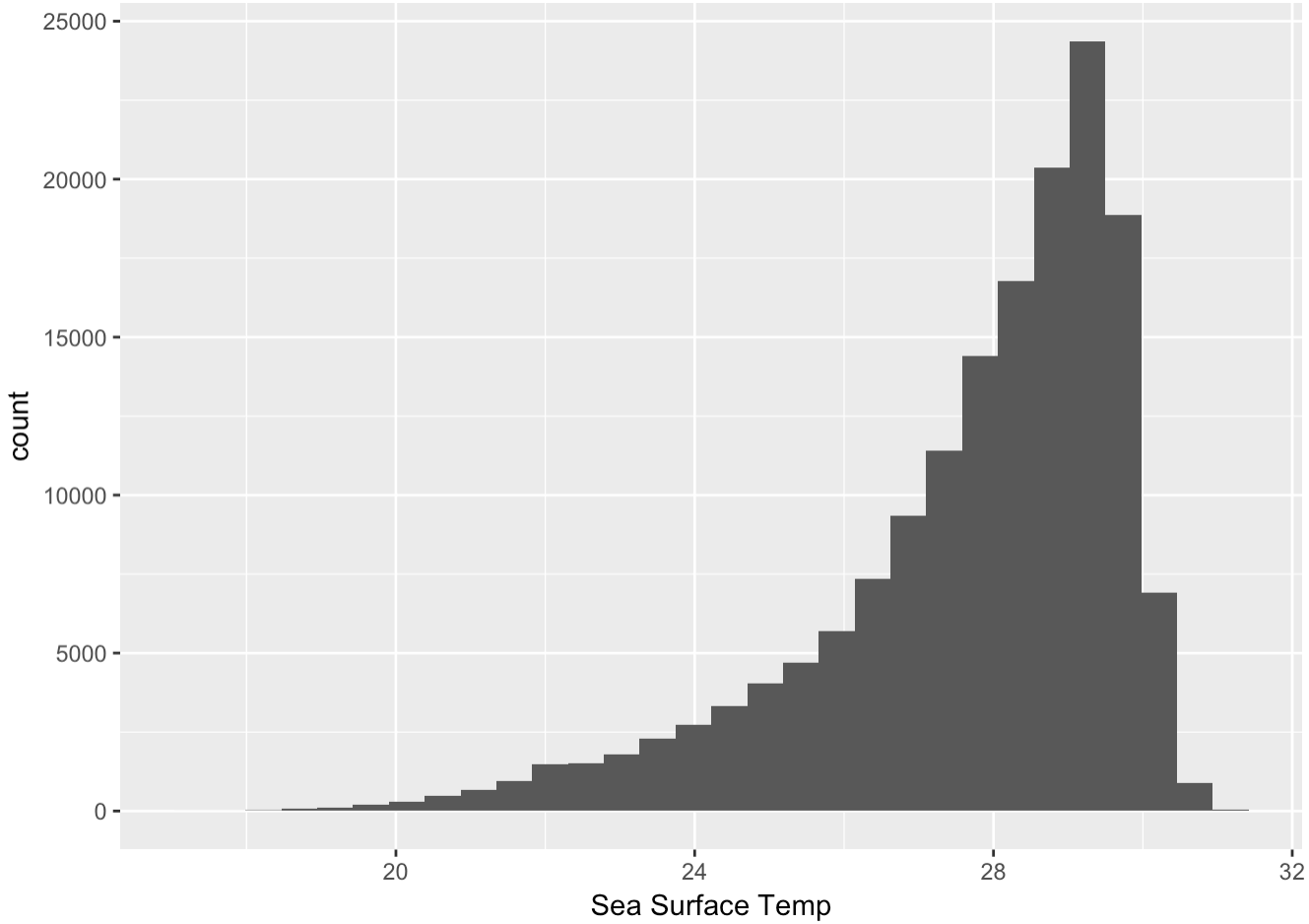
ggplot2: Explore a dataset

El Niño data

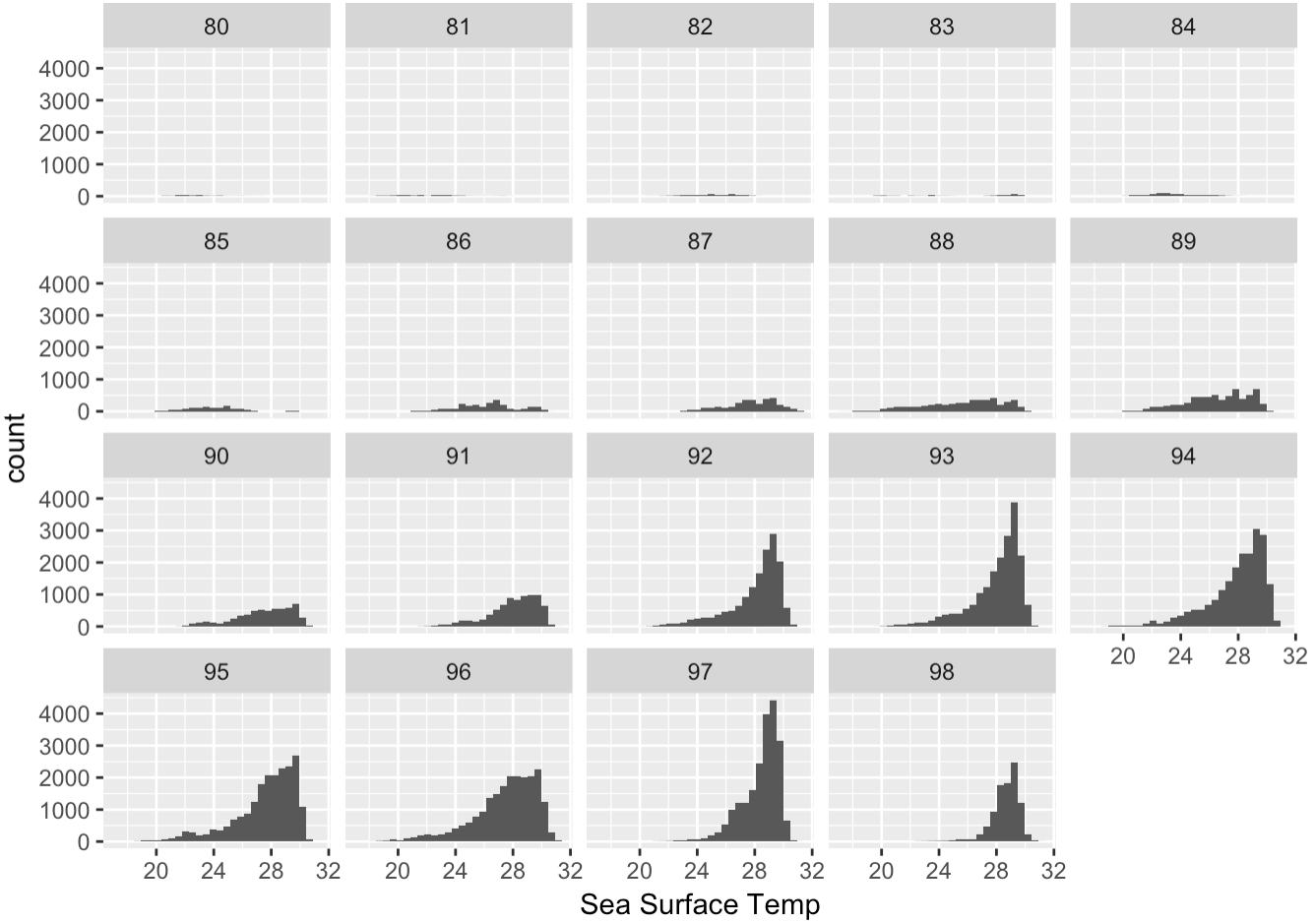
- From https://www.kaggle.com/uciml/el-nino-dataset
- 178k records, 12 variables

age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex
90	NA	77053	HS-grad	9	Widowed	NA	Not-in-family	White	Female
82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female
66	NA	186061	Some-college	10	Widowed	NA	Unmarried	Black	Female
54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female
41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female
34	Private	216864	HS-grad	9	Divorced	Other-service	Unmarried	White	Female
38	Private	150601	10th	6	Separated	Adm-clerical	Unmarried	White	Male
74	State-gov	88638	Doctorate	16	Never-married	Prof-specialty	Other-relative	White	Female
68	Federal-gov	422013	HS-grad	9	Divorced	Prof-specialty	Not-in-family	White	Female
41	Private	70037	Some-college	10	Never-married	Craft-repair	Unmarried	White	Male

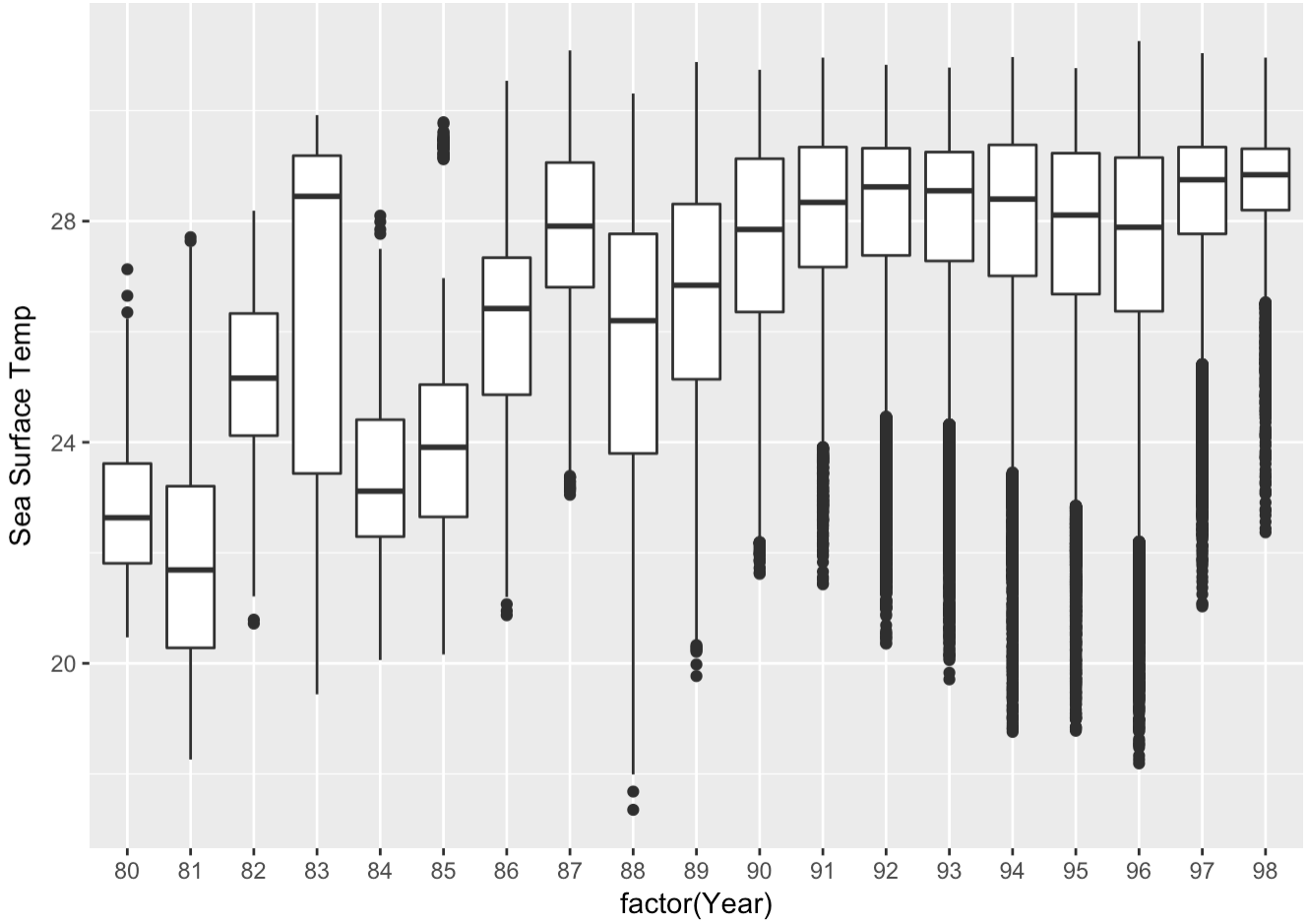
Distributions



Pairs of variables



Larger combinations



Design exercise: El Niño

ggplot2 Resources

- General ggplot2 information
  http://ggplot2.tidyverse.org/
- R Graphics Cookbook (recipes for plots)
  http://www.cookbook-r.com/Graphs/index.html
- R for Data Science (online book that includes ggplot2)
  http://r4ds.had.co.nz/
- ggplot2: Elegant Graphs for Data Analysis (book by Hadley Wickham)
  http://ggplot2.org/book/
- ggplot2 cheatsheet (also in RStudio)
  http://bit.ly/ggplot2-cheatsheet

Thanks for your feedback!

angela.zoss@duke.edu