Visualization for Data Science in R

Angela Zoss

Data Matters 2018

http://bit.ly/RVis2018

Objectives/Outline

Day 1: Static visualizations

- Visualization and data science
- Basic ggplot2 syntax
- Charts for exploration
- Charts for communication
- Advanced topics: mapping, saving charts out

Day 2: Interactivity

- Simple interactive plots
- Arranging charts into dashboards
- Incorporating Shiny elements into documents, dashboards
- Advanced topics: full Shiny apps

Set up environment

- R
- RStudio
- packages

Packages:

- tidyverse
- maps
- mapproj
- plotly
- flexdashboard
- shiny

Get workshop files

URL: https://github.com/amzoss/RVis-DM2018

With Git installed

In RStudio:

- Project → New project
- Version Control
- Git
  - Paste in GitHub URL
  - Project directory name: RVis-DM2018
  - Subdirectory: you choose
- Create Project

Without Git installed

- Click green button to download ZIP
- Unzip files on your laptop

In RStudio:

- Project → New project…
- Existing directory
- Select unzipped folder
- Create Project

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

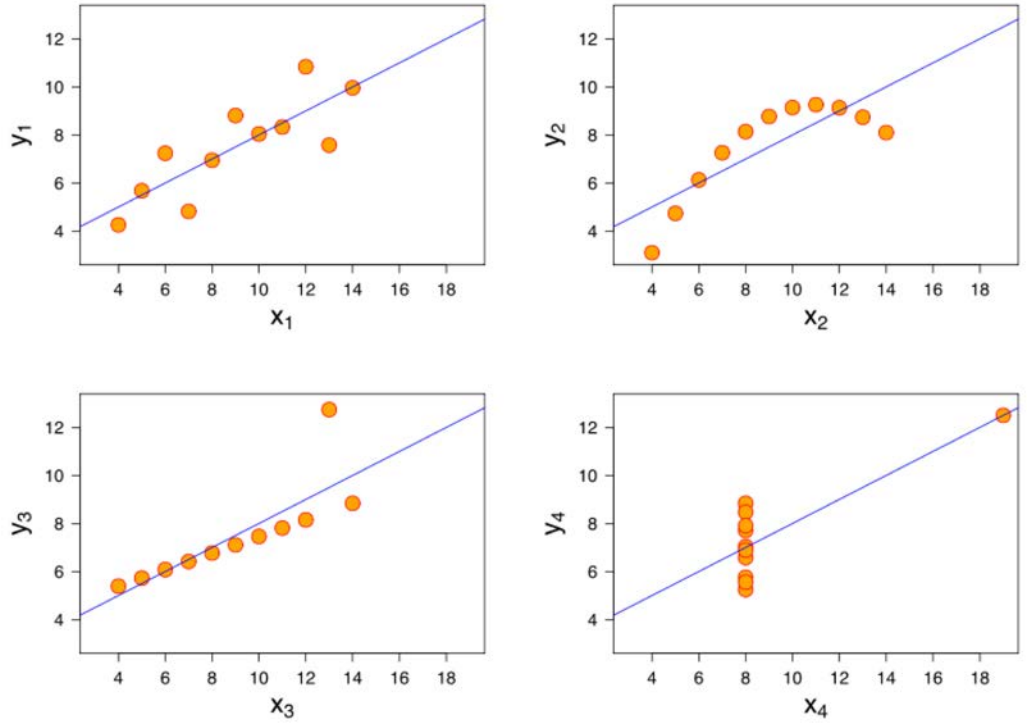
x & y variance

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

Why visualize in R?

- Quickly explore data
- Save time switching to another tool
- Use charts to inspire new analyses and vice versa
- Reproducibility

Why care about reproducibility?

- Open science makes review easier
- Increasingly a requirement
- Saves you a lot of time trying to figure out what you did last time!

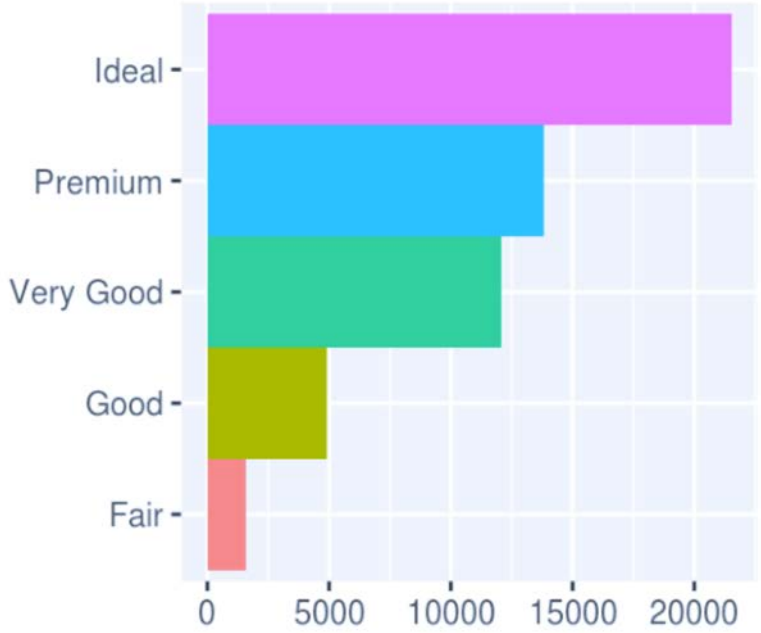
*“Your closest collaborator is **you** six months ago, but you don’t reply to emails.”*

- Mark Holder

ggplot2

What is ggplot2?

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



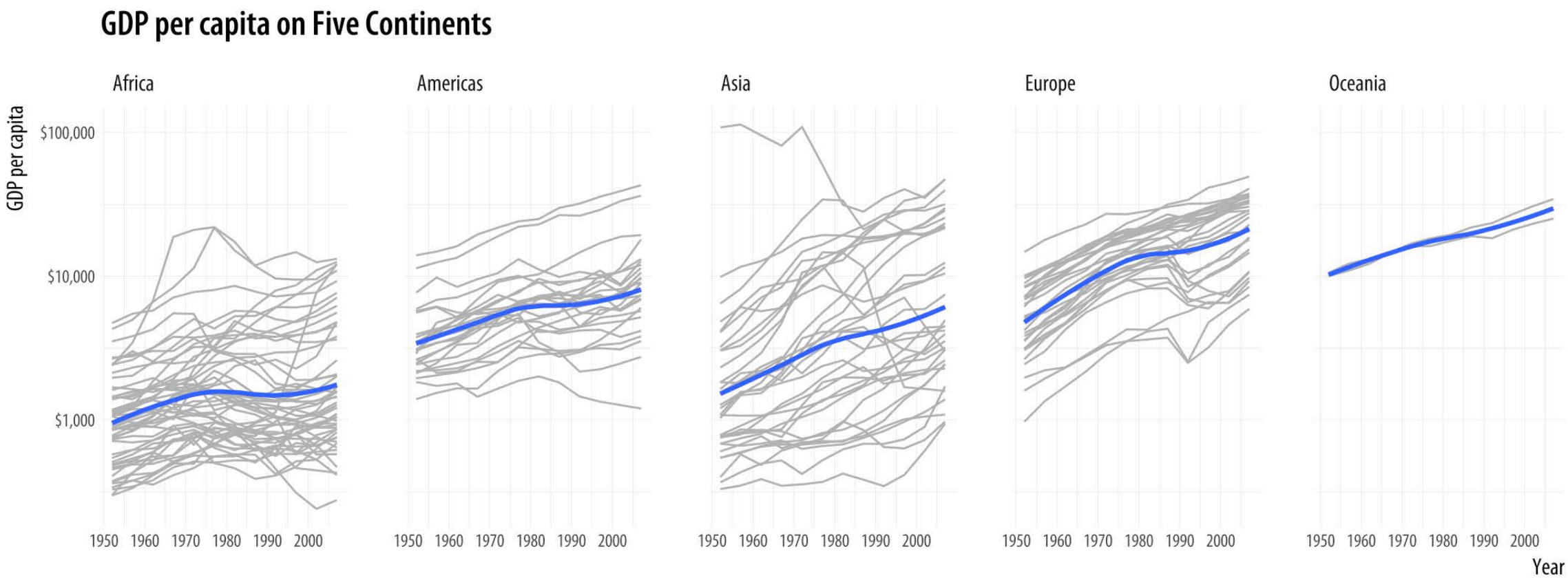
http://r4ds.had.co.nz/data-visualisation.html

Grammar of graphics

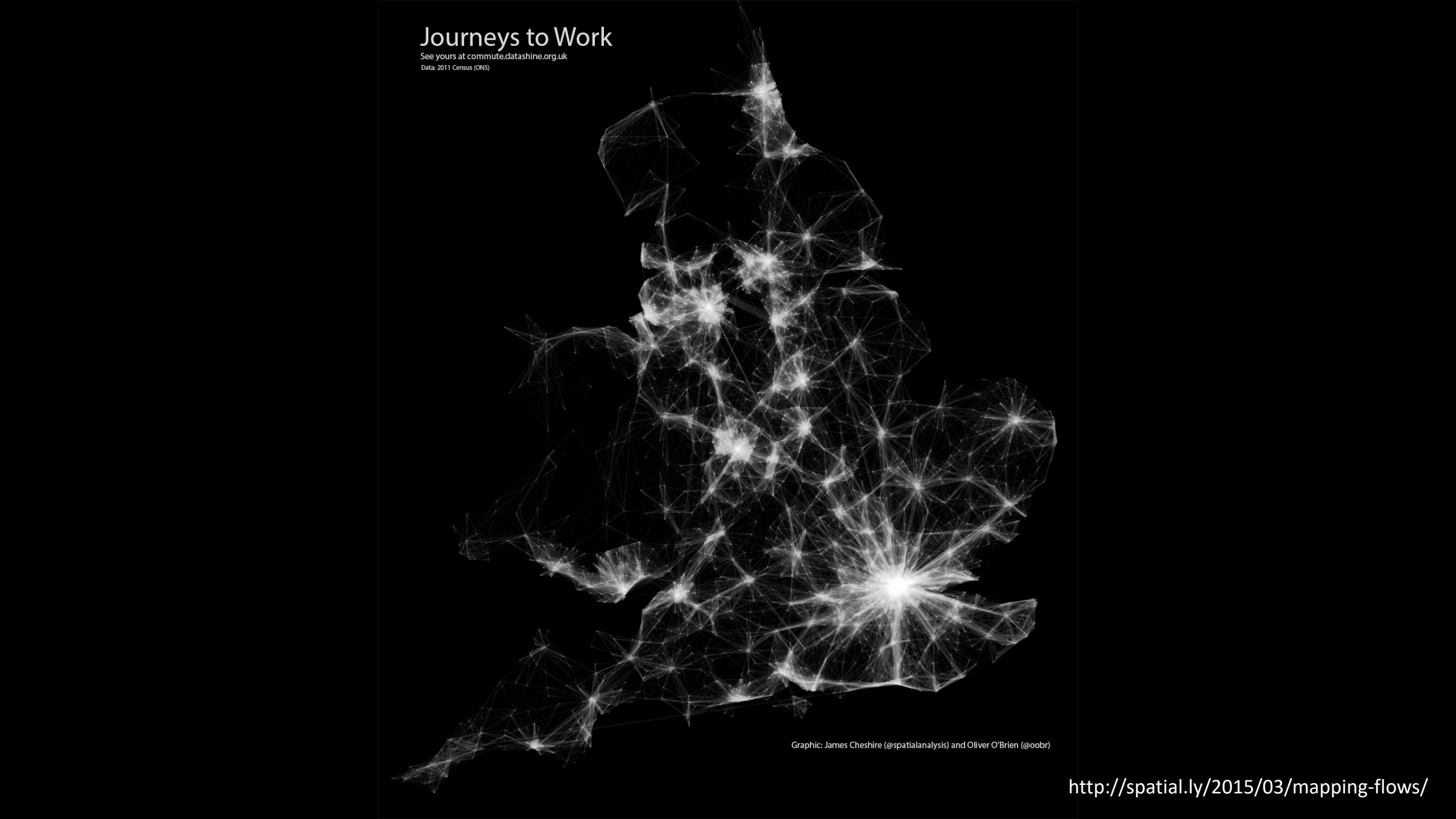
1. DATA: a set of data operations that create variables from datasets
2. TRANS: variable transformations (e.g., rank)
3. SCALE: scale transformations (e.g., log)
4. COORD: a coordinate system (e.g., polar)
5. ELEMENT: graphs (e.g., points) and their aesthetic attributes (e.g., color)
6. GUIDE: one or more guides (axes, legends, etc.).

Wilkinson, Leland. (2005). *The grammar of graphics (2nd ed)*. New York: Springer.

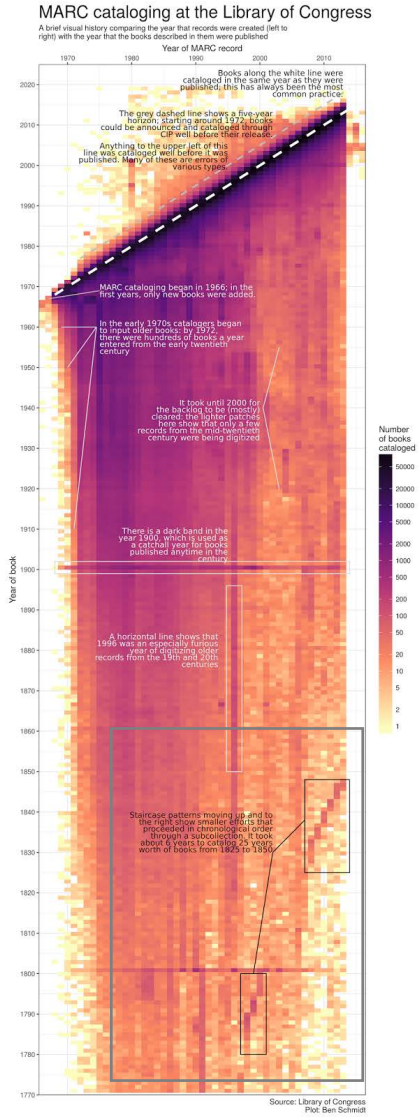
ggplot2 examples

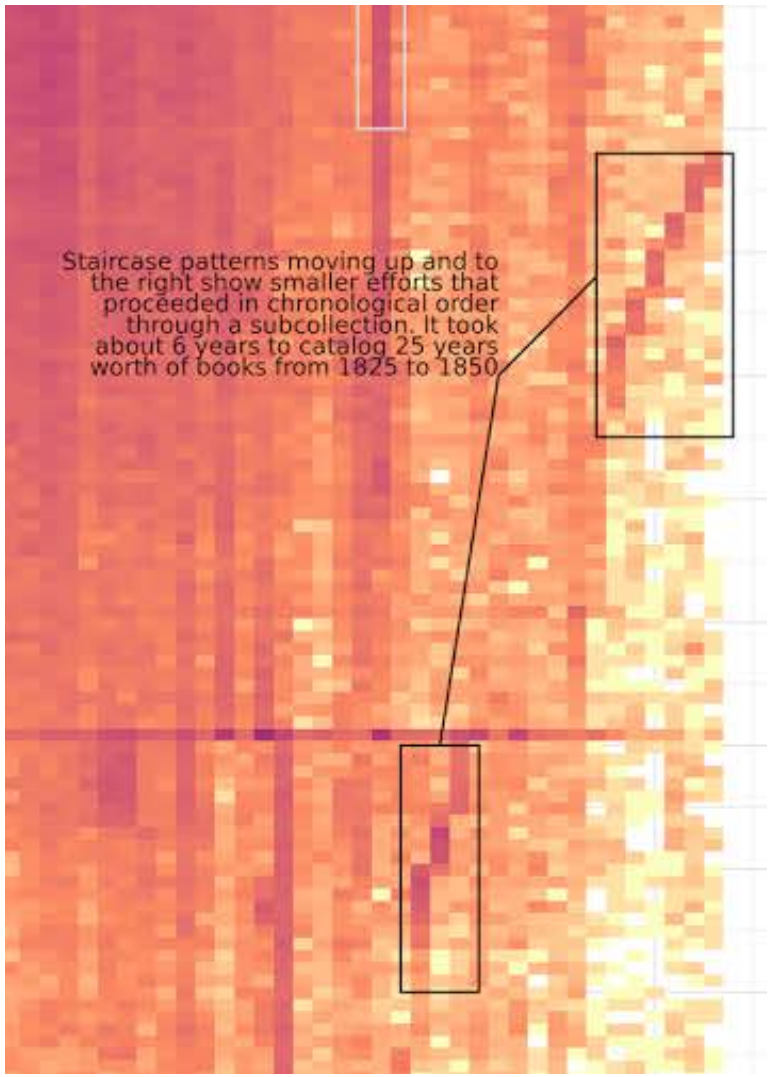


http://socviz.co/groupfacettx.html



http://spatial.ly/2015/03/mapping-flows/





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

Why ggplot2 instead of base R?

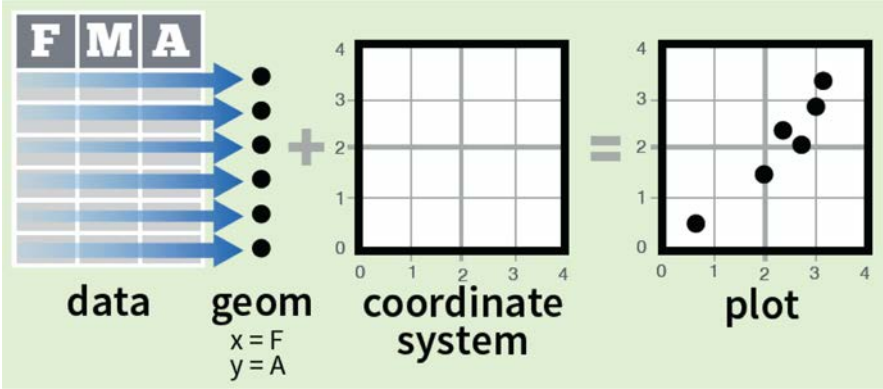
- nice defaults
- easy faceting
- (arguably) more natural syntax
- can switch chart types more easily

“Why I use ggplot2”, David Robinson
http://varianceexplained.org/r/why-I-use-ggplot2/

ggplot2: Elements

Basic elements in any ggplot2 visualization

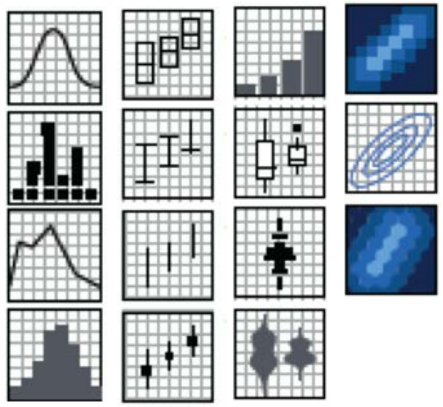
- **data**
- **aes**thetics
  (variable mappings)
- **geom**
  (chart type or shape)
- coordinate system
  (the arrangement of the marks;
  most geoms use default, cartesian)



http://bit.ly/ggplot2-cheatsheet

Types of geoms

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



http://bit.ly/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

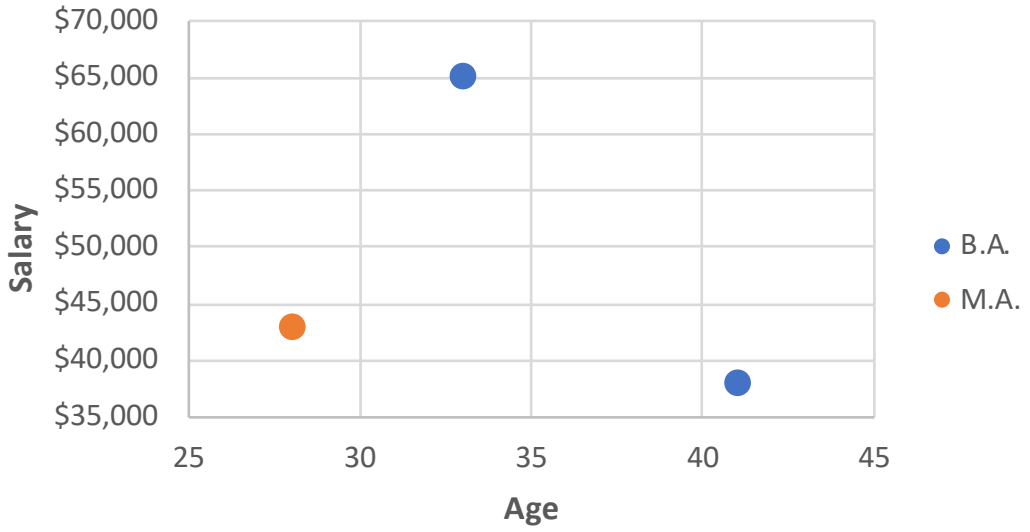
```
ggplot( data = data frame )
```

+

```
geom_...( aes(variable mappings) ,
          non-variable adjustments )
```

Aesthetic variable mappings

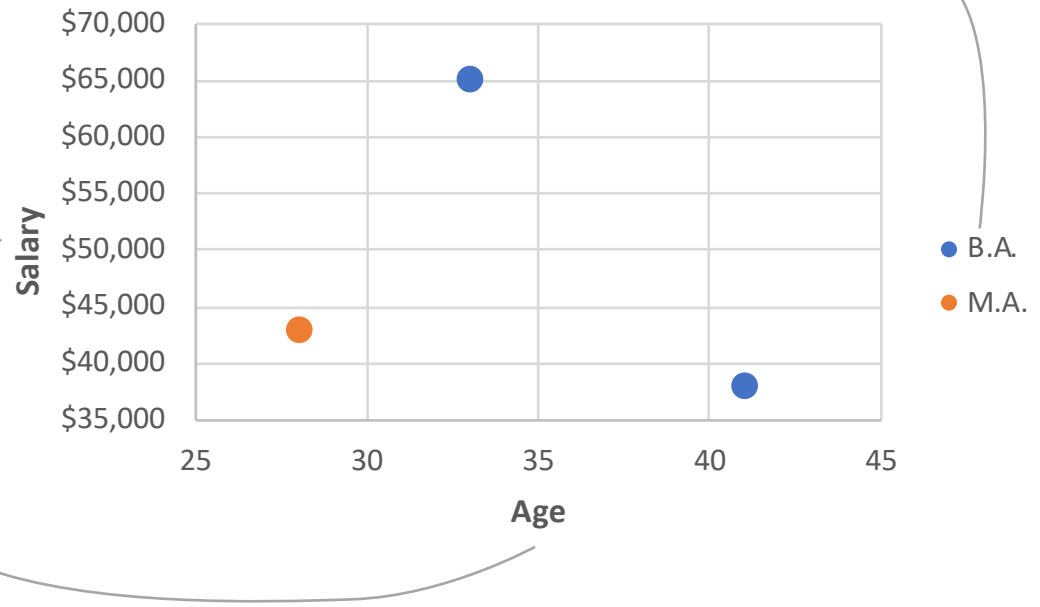
Name	Age	Salary	Highest Degree
Jane Smith	33	$65,000	B.A.
Abby Jones	28	$43,000	M.A.
Bridget Carden	41	$38,000	B.A.



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

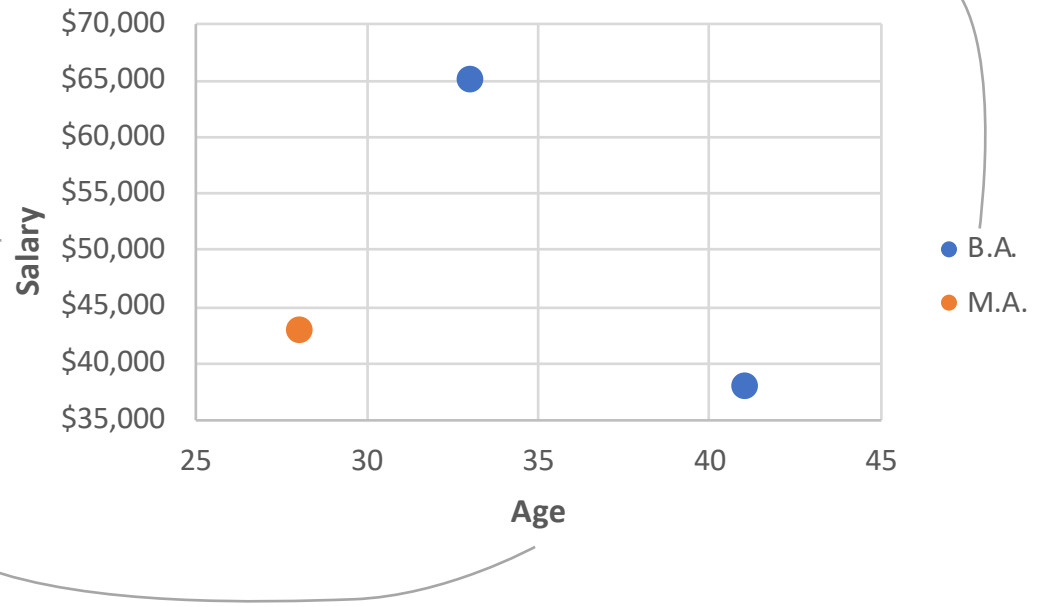


Aesthetic variable mappings

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

x position y position color

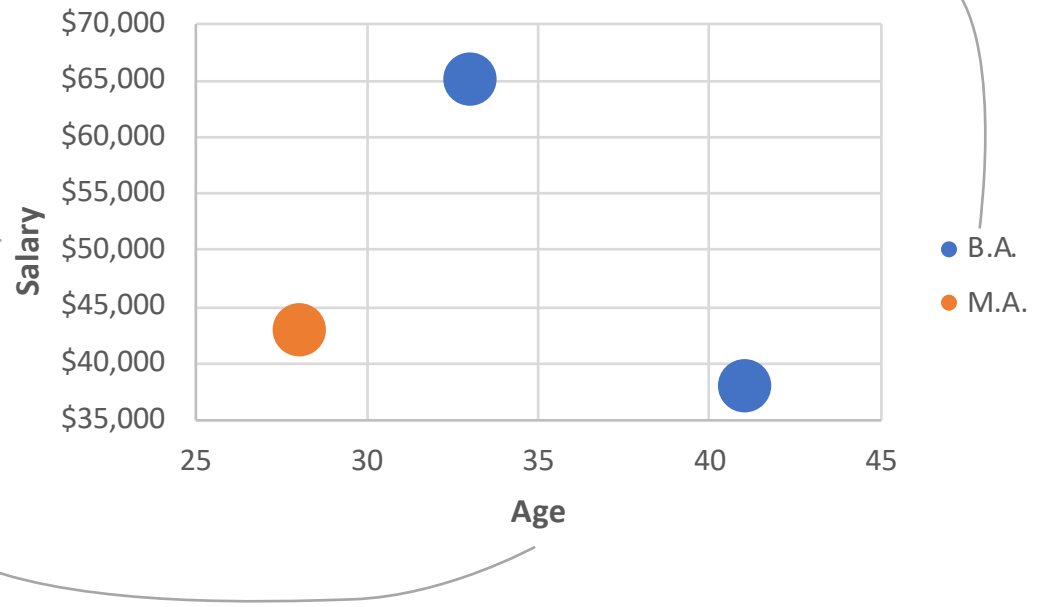


Non-variable adjustments

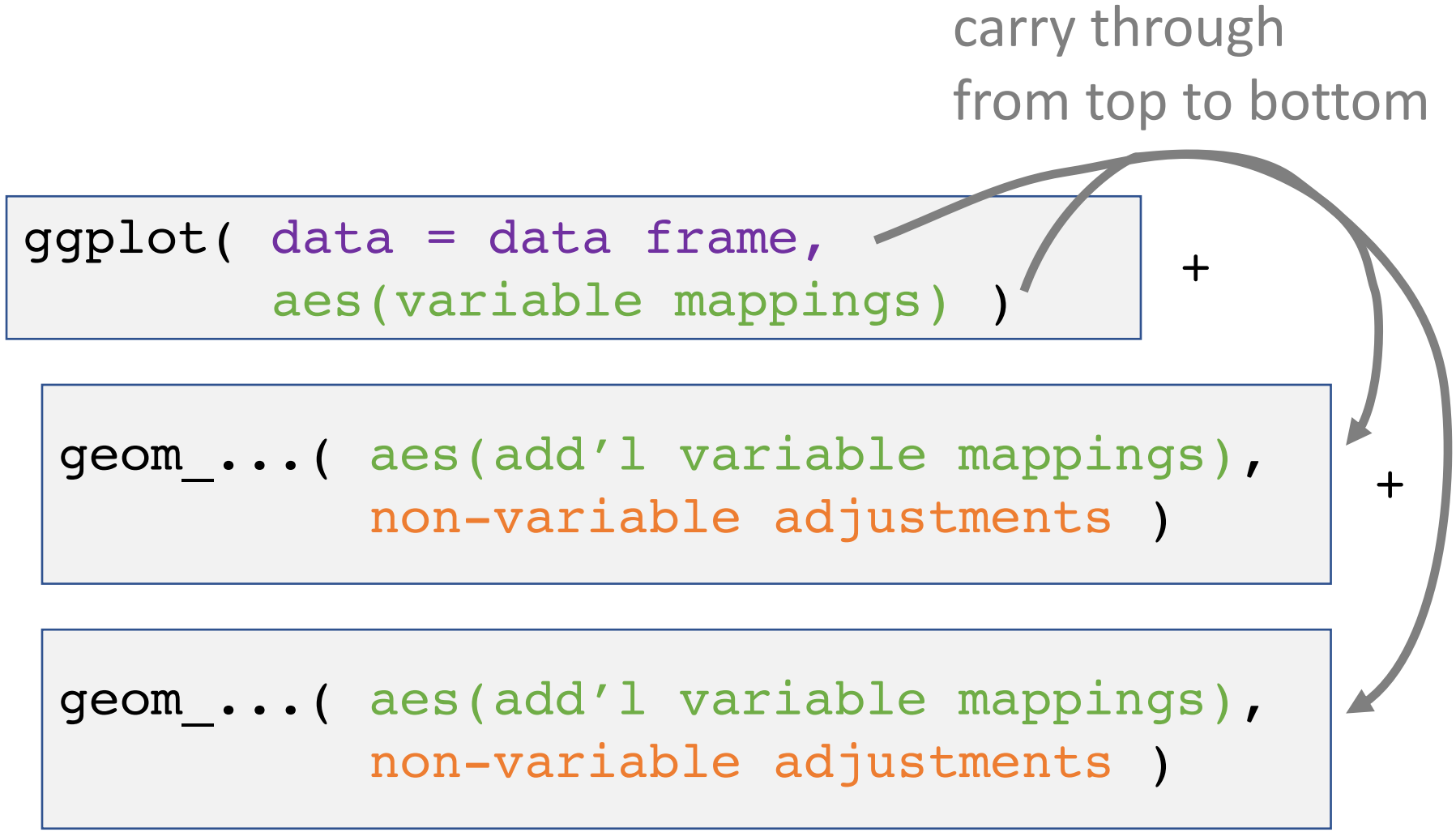
Name	Age	Salary	Highest Degree
Jane Smith	33	$65,000	B.A.
Abby Jones	28	$43,000	M.A.
Bridget Carden	41	$38,000	B.A.

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

x position y position color



Template for a more complex plot



Using RStudio

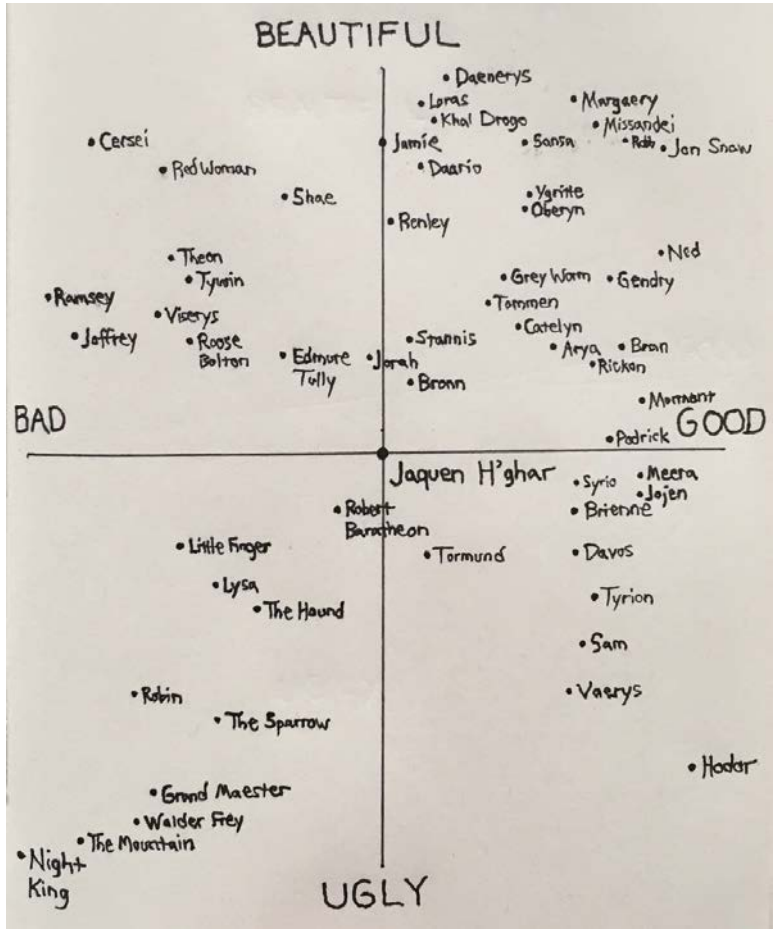
- Projects
- Rmarkdown
- Cheat sheets

https://www.rstudio.com/resources/cheatsheets/#rmarkdown

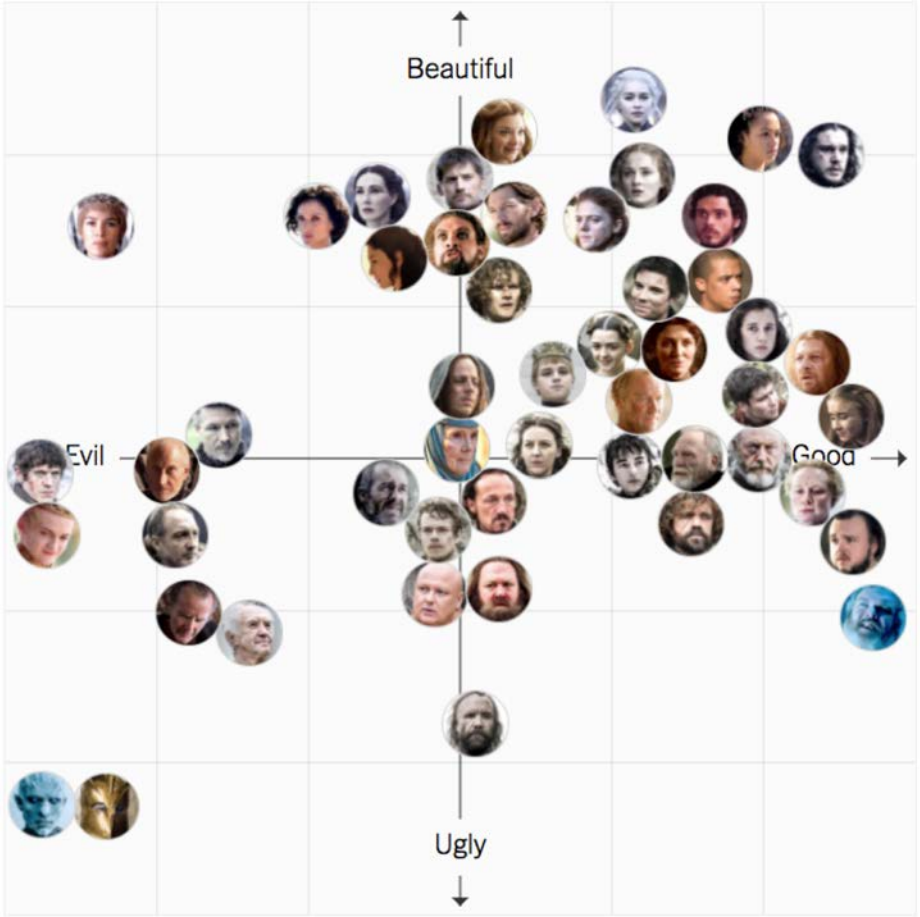
Dataset 1:
Game of Thrones character ratings

https://www.nytimes.com/interactive/2017/08/09/upshot/game-of-thrones-chart.html

Game of Thrones character ratings



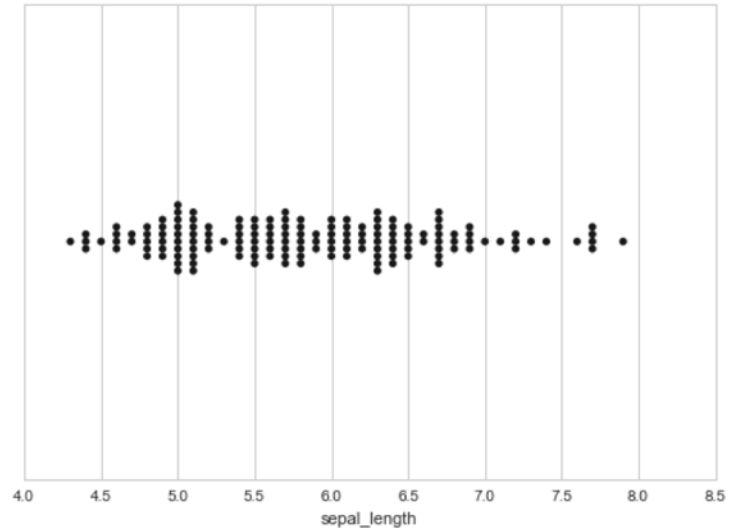
https://www.instagram.com/p/BWnn-YogX1n/

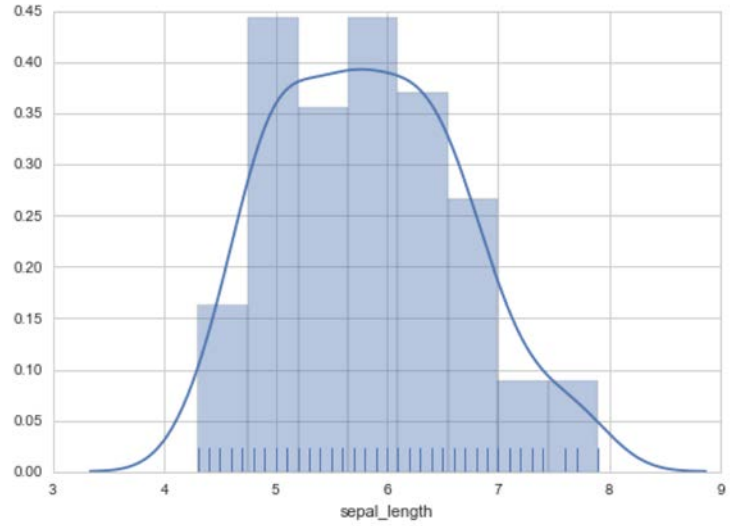


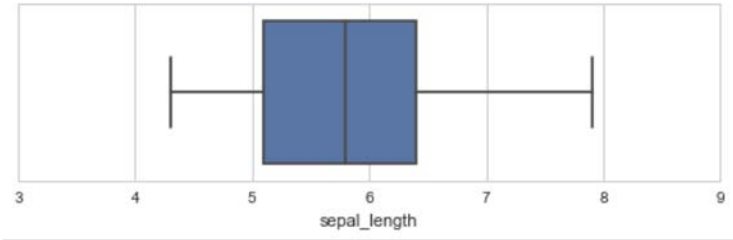
https://www.nytimes.com/interactive/2017/08/09/upshot/game-of-thrones-chart.html

Data exploration

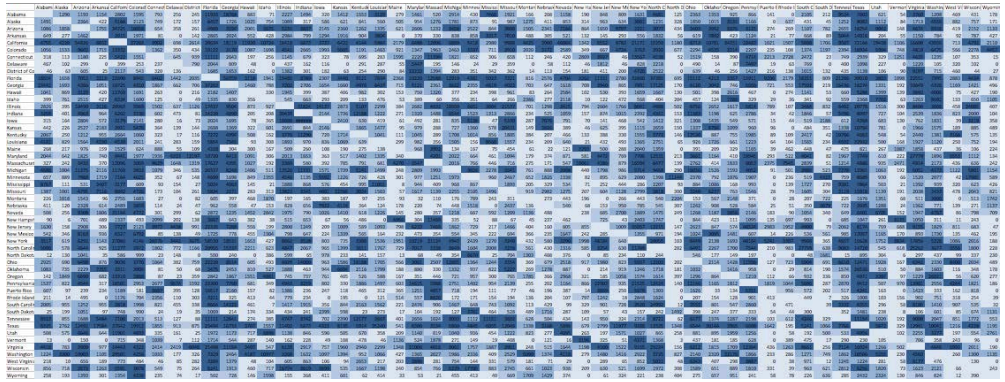
Exploring single variables

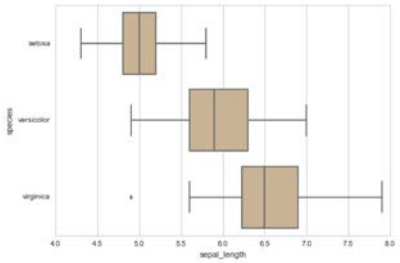


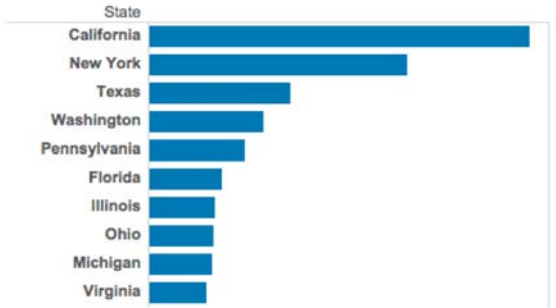


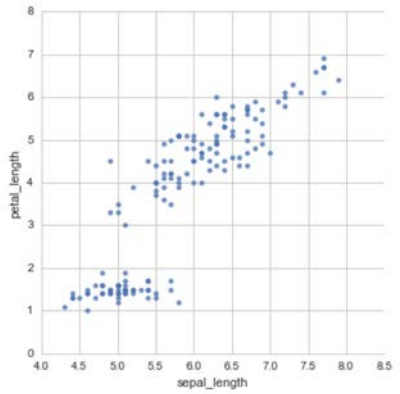


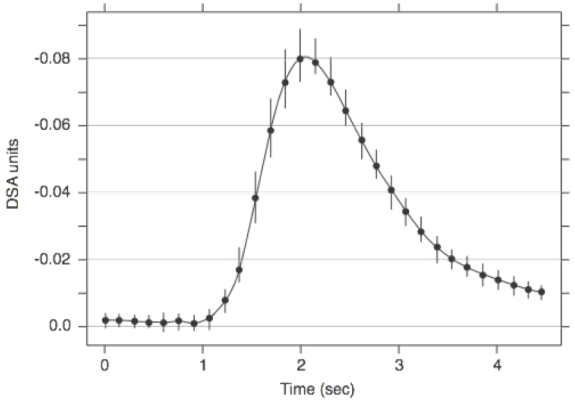
Exploring two variables

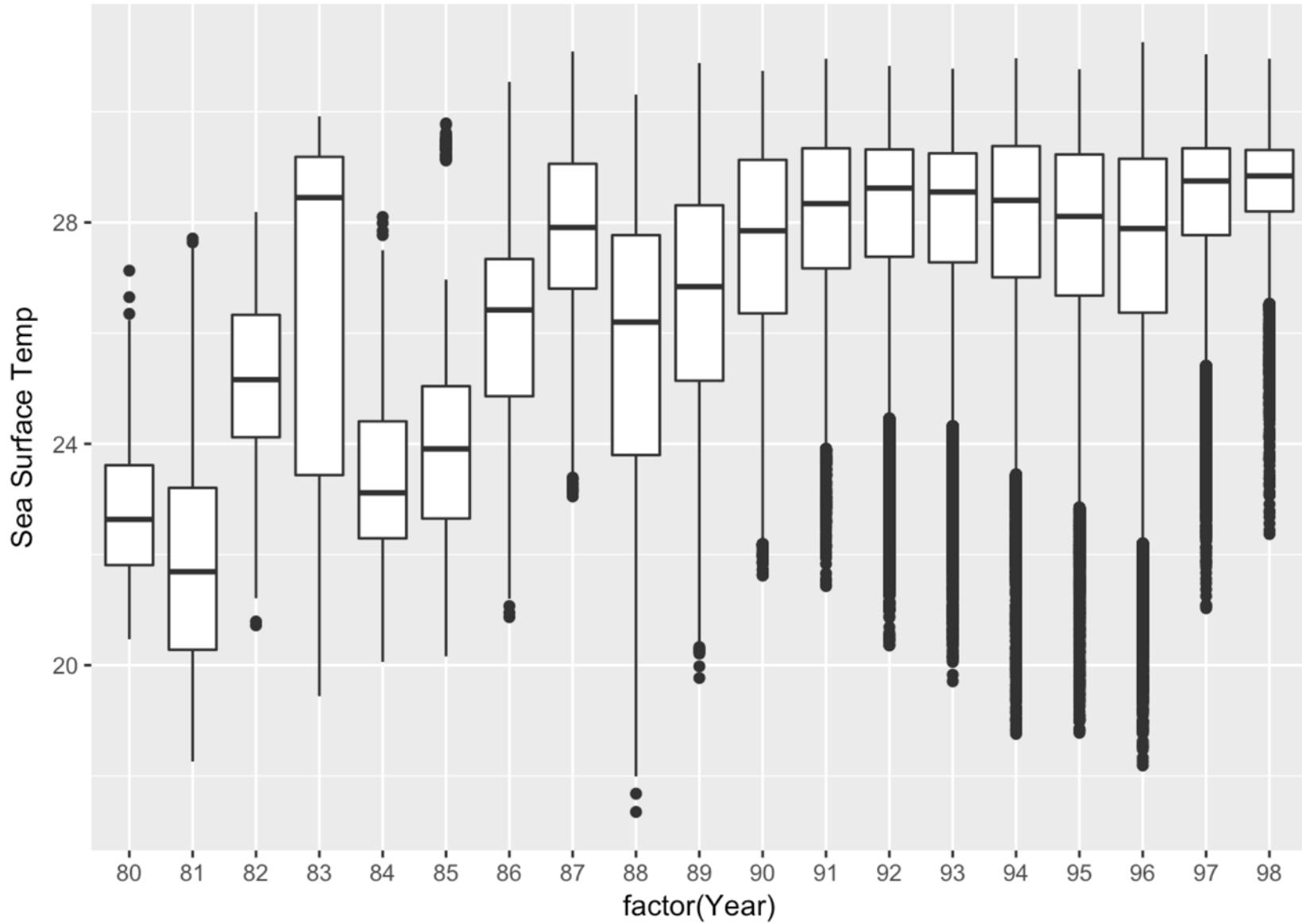


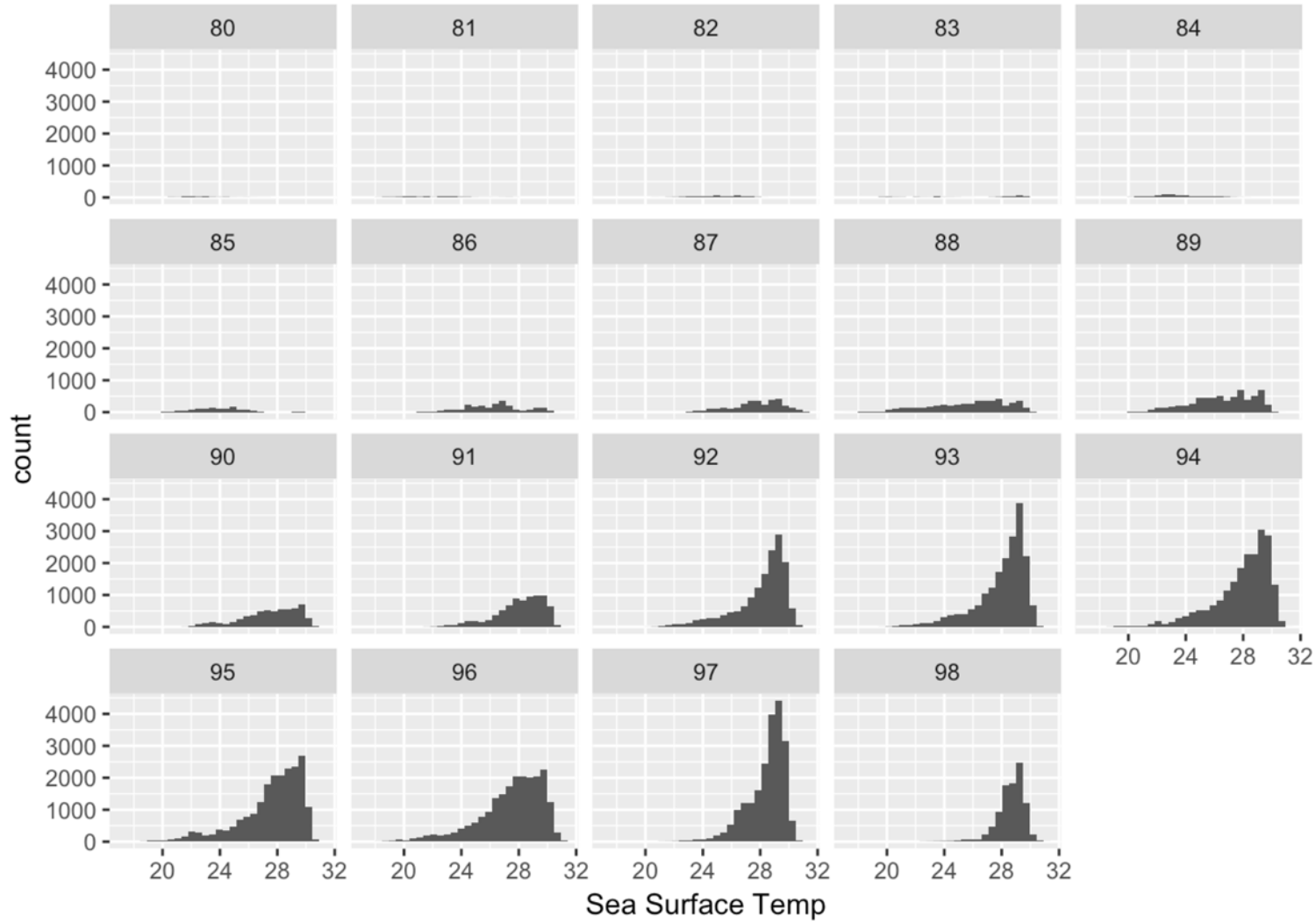






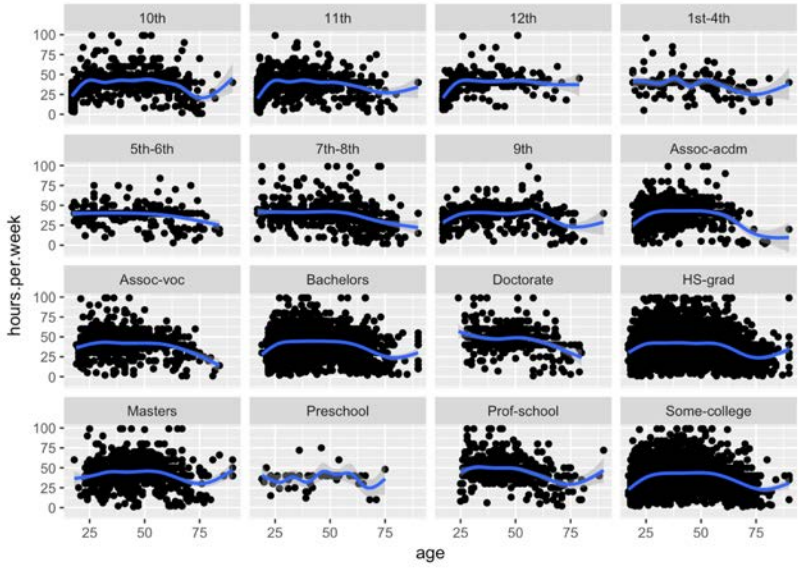


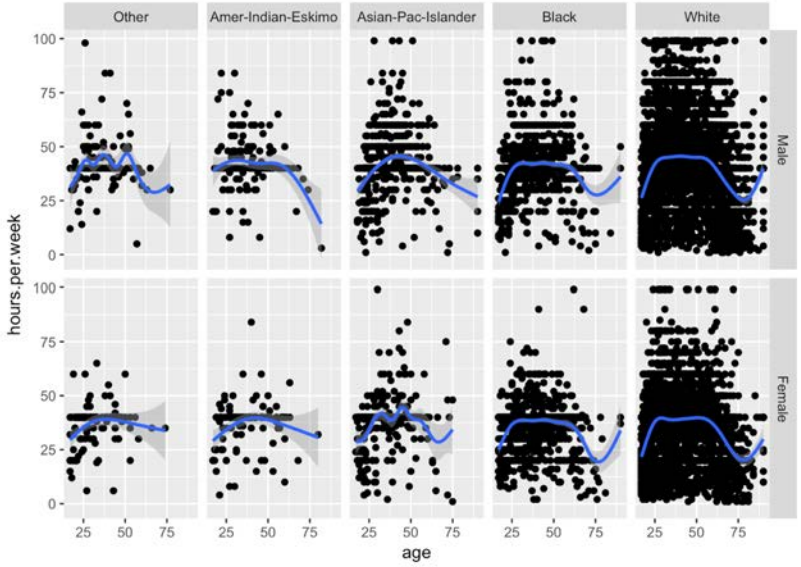

Sea Surface Temp
28
24
20
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
factor(Year)


80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
count
4000
3000
2000
1000
0
20
24
28
32
Sea Surface Temp

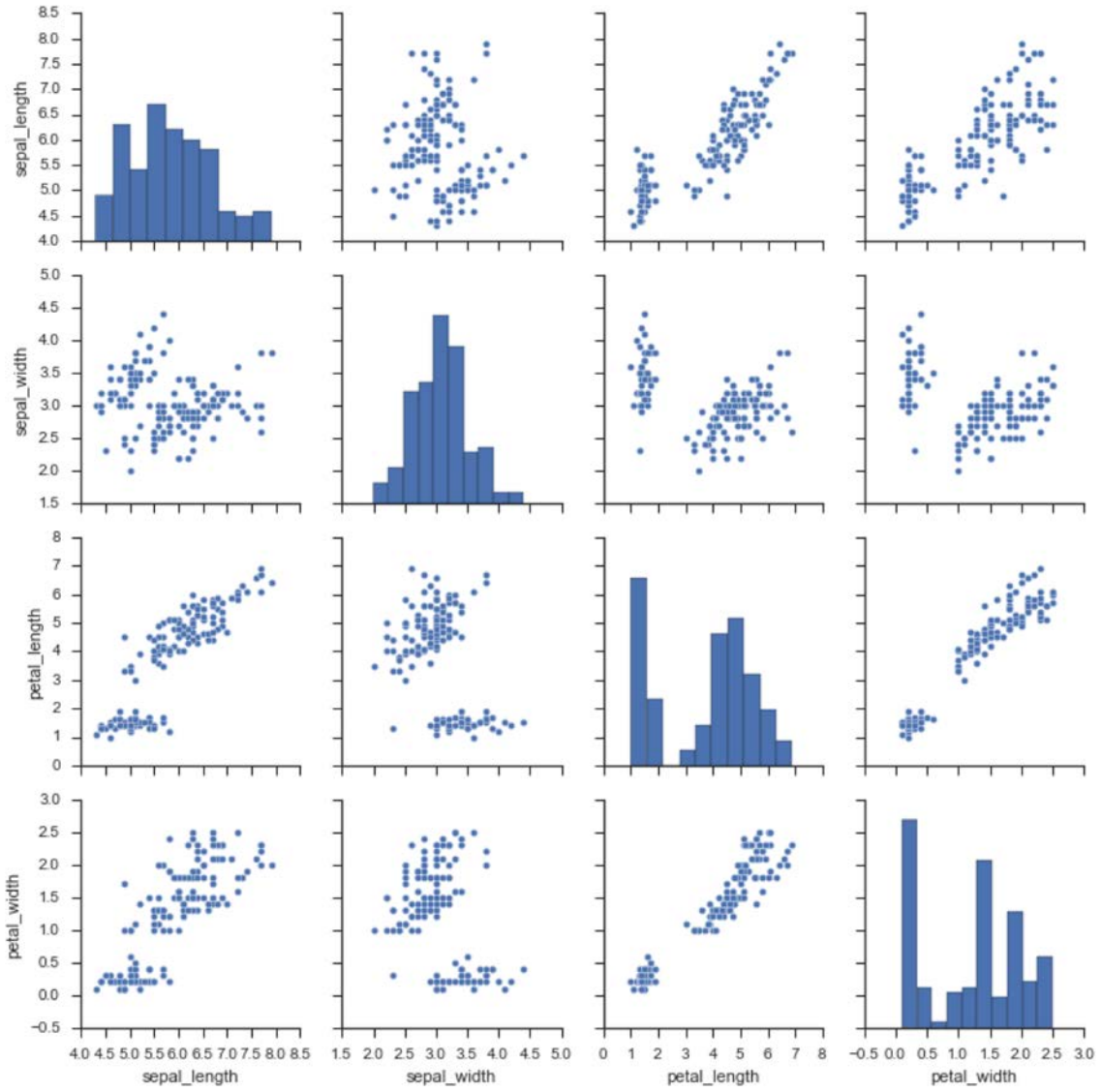
Facets

- grid vs. wrap

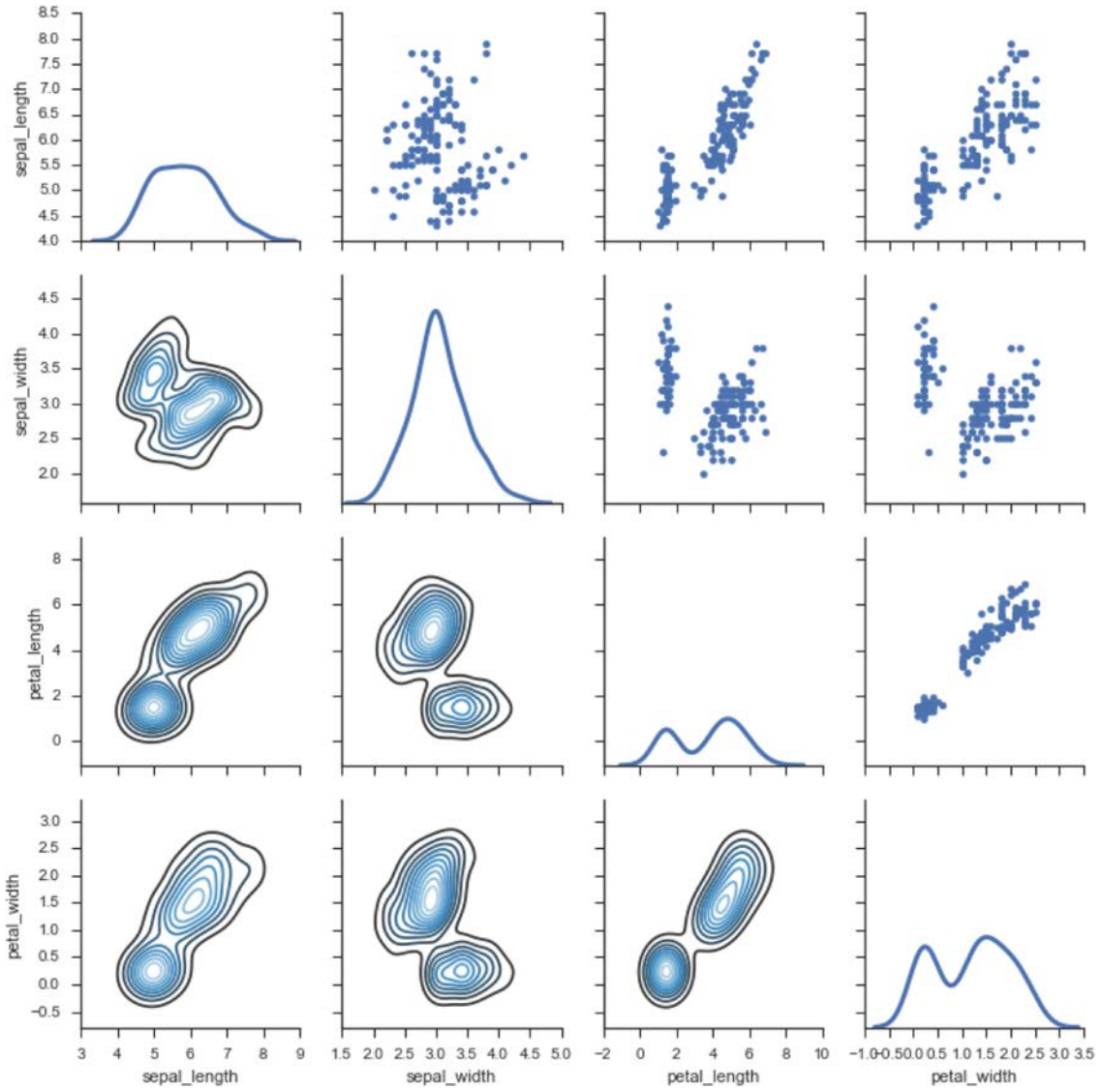




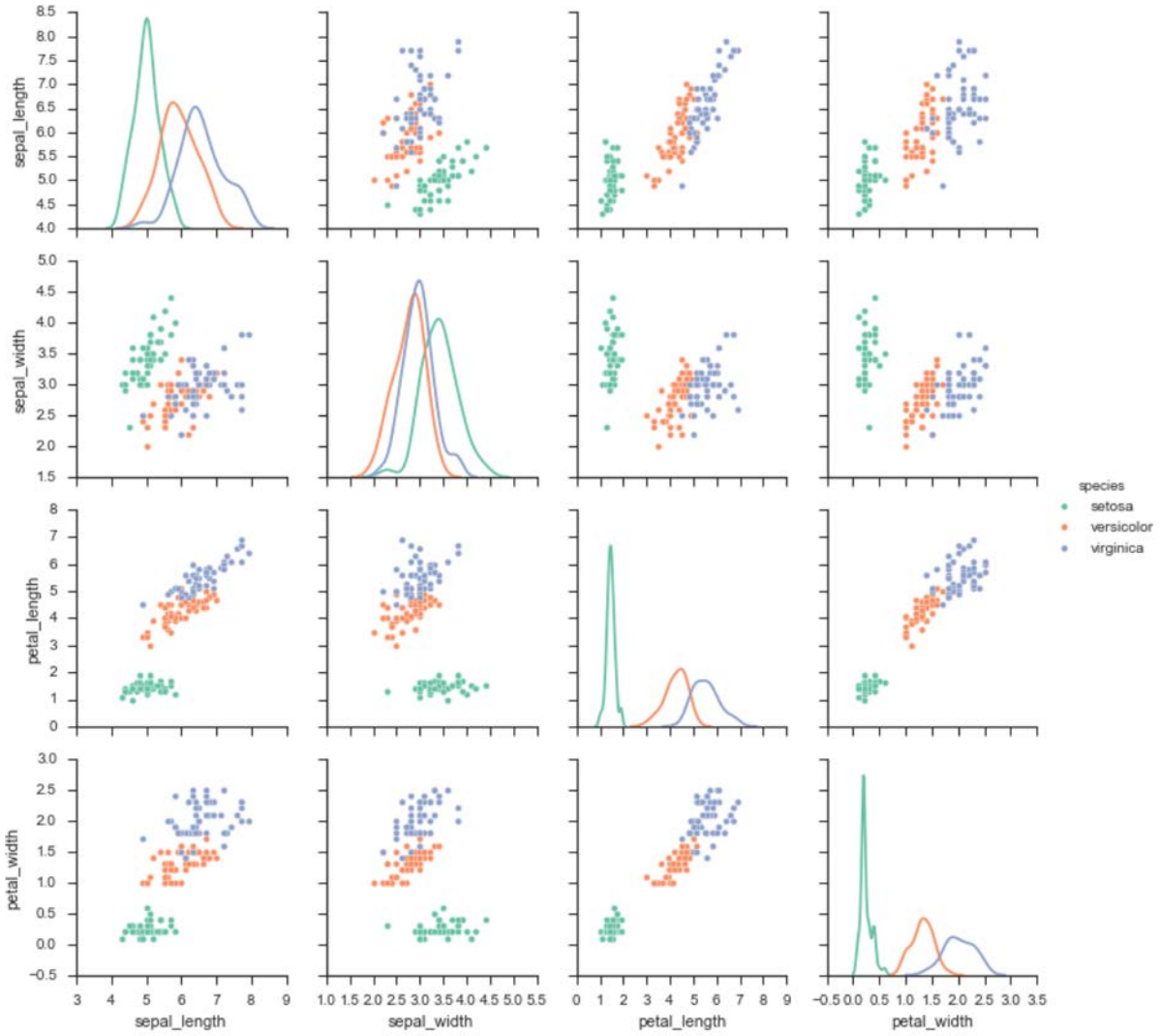
Exploring many variables, two at a time



Exploring many variables, two at a time



Exploring many variables, three at a time

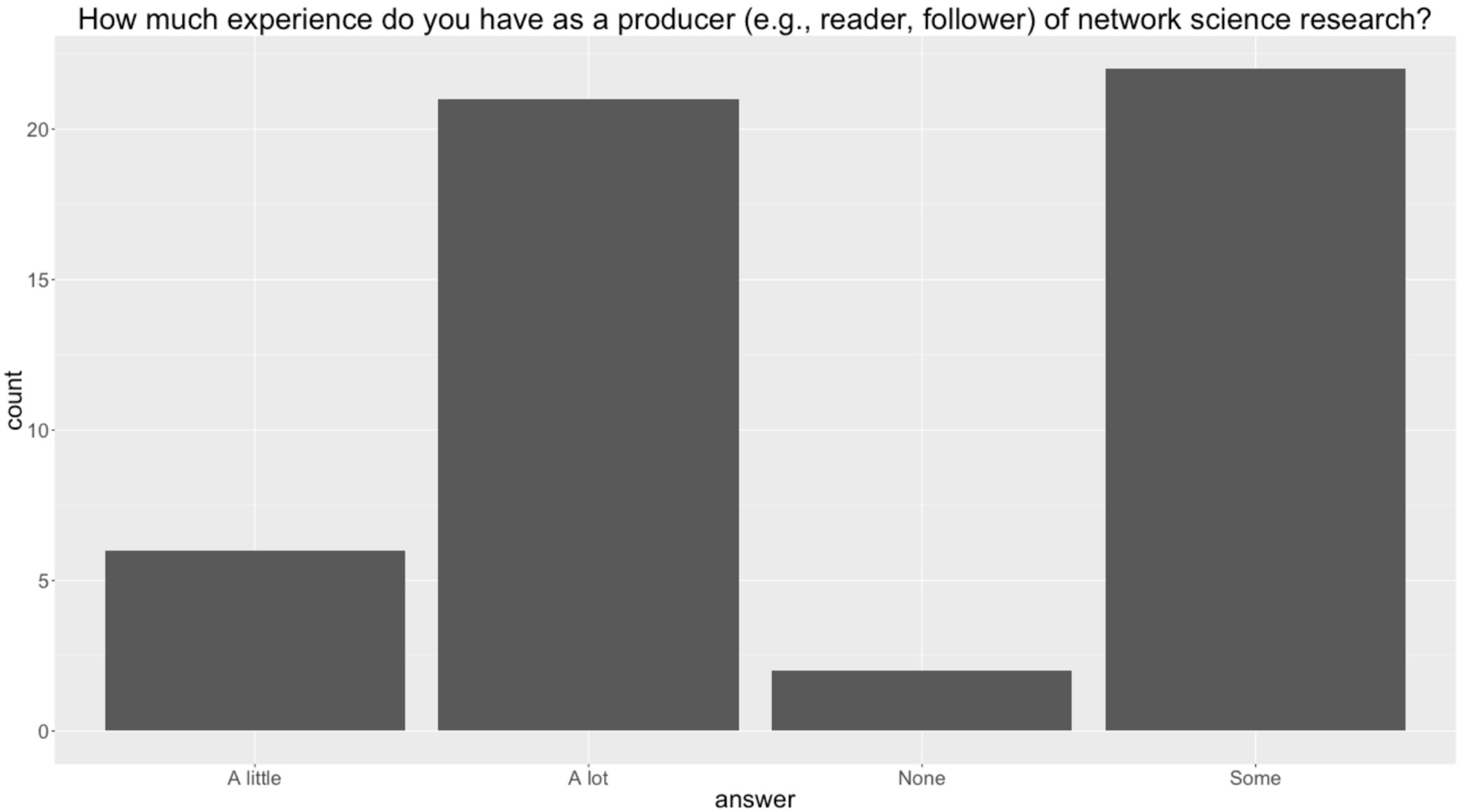


Dataset 2:
Star Wars character data

https://dplyr.tidyverse.org/reference/starwars.html

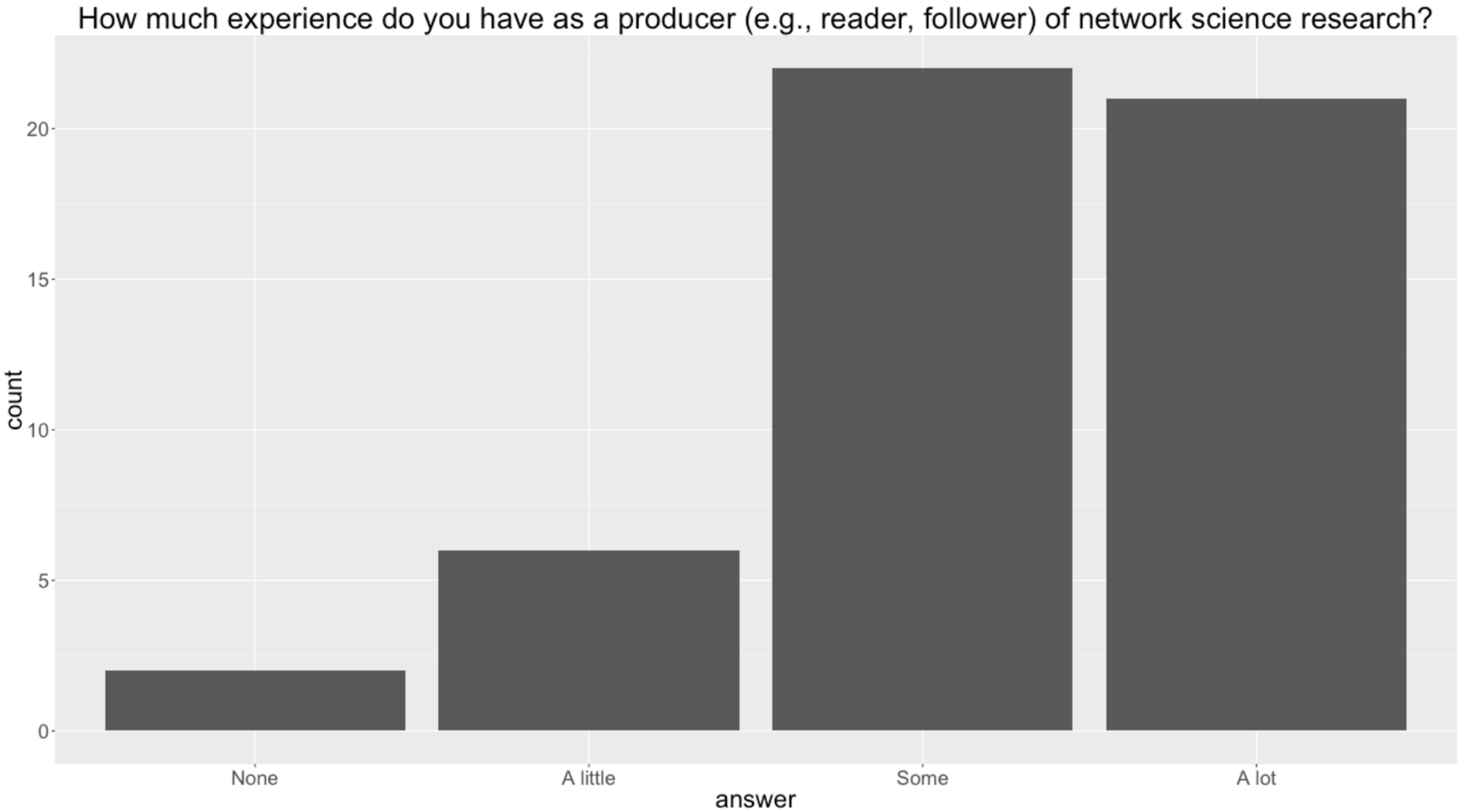
Principles for Effective Visualizations

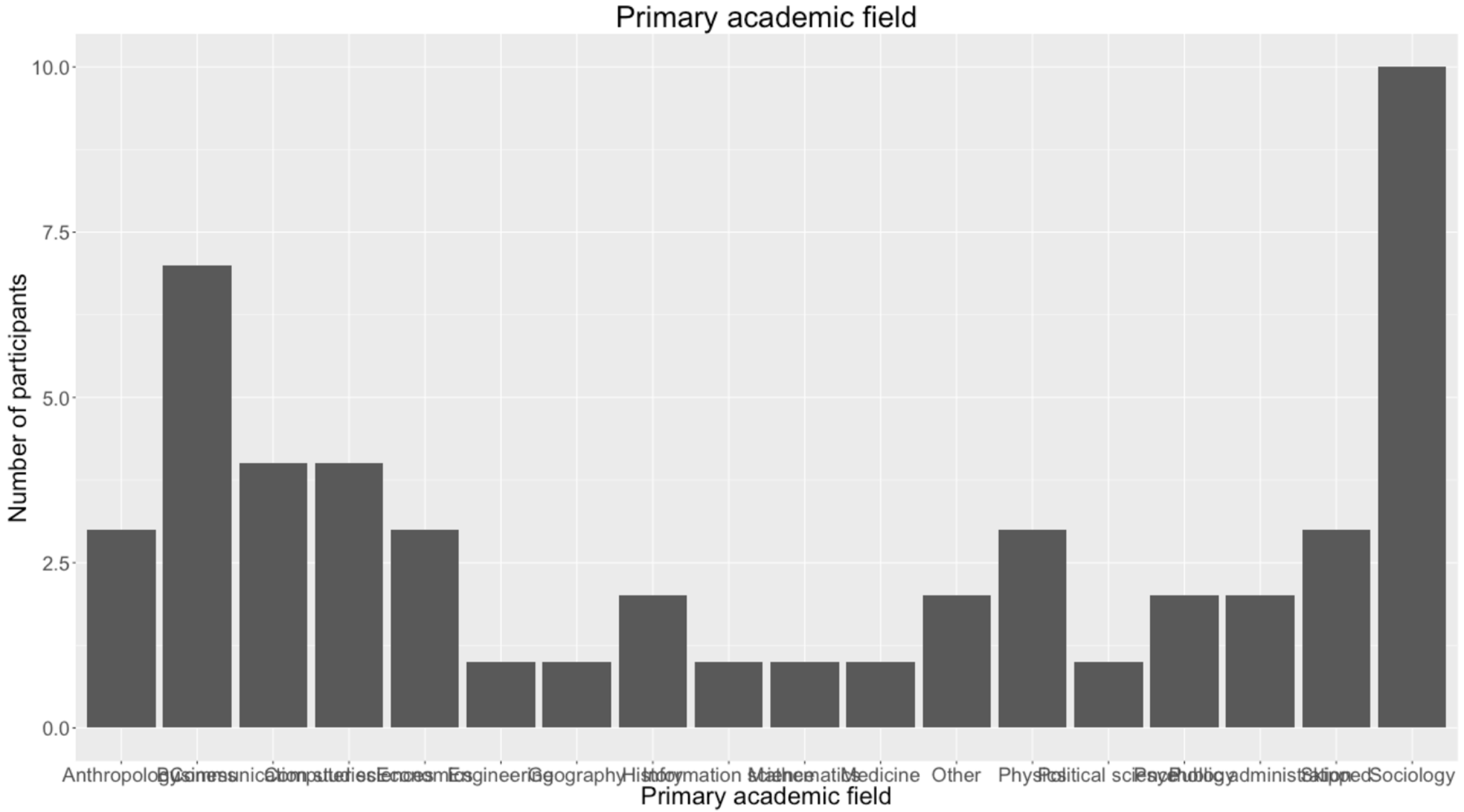
Principle 1: Order matters


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

Order by meaning

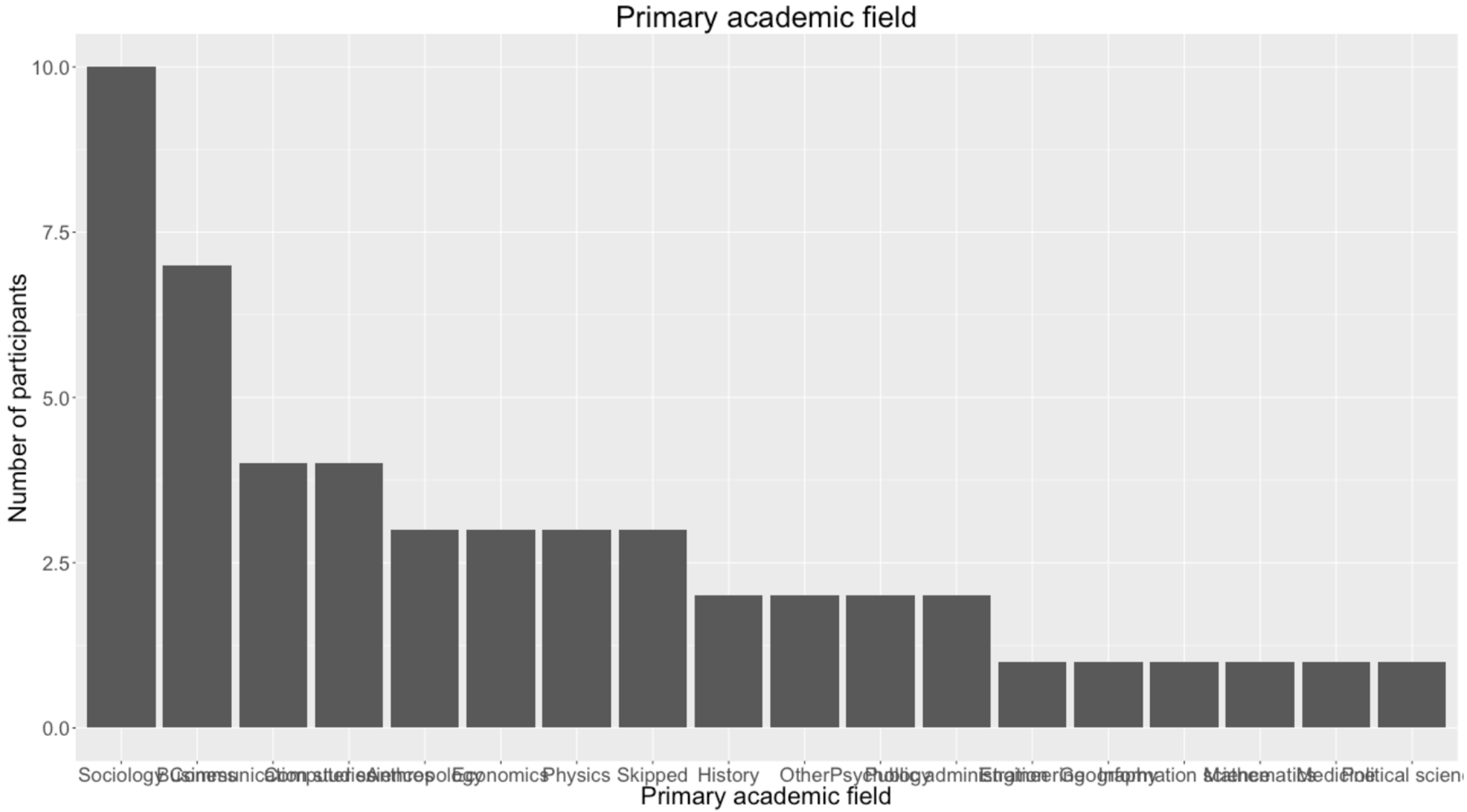
```
data$answer <-
    factor(data$answer,
           levels=c("None", "A little", "Some", "A lot"),
           ordered = TRUE)
```


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

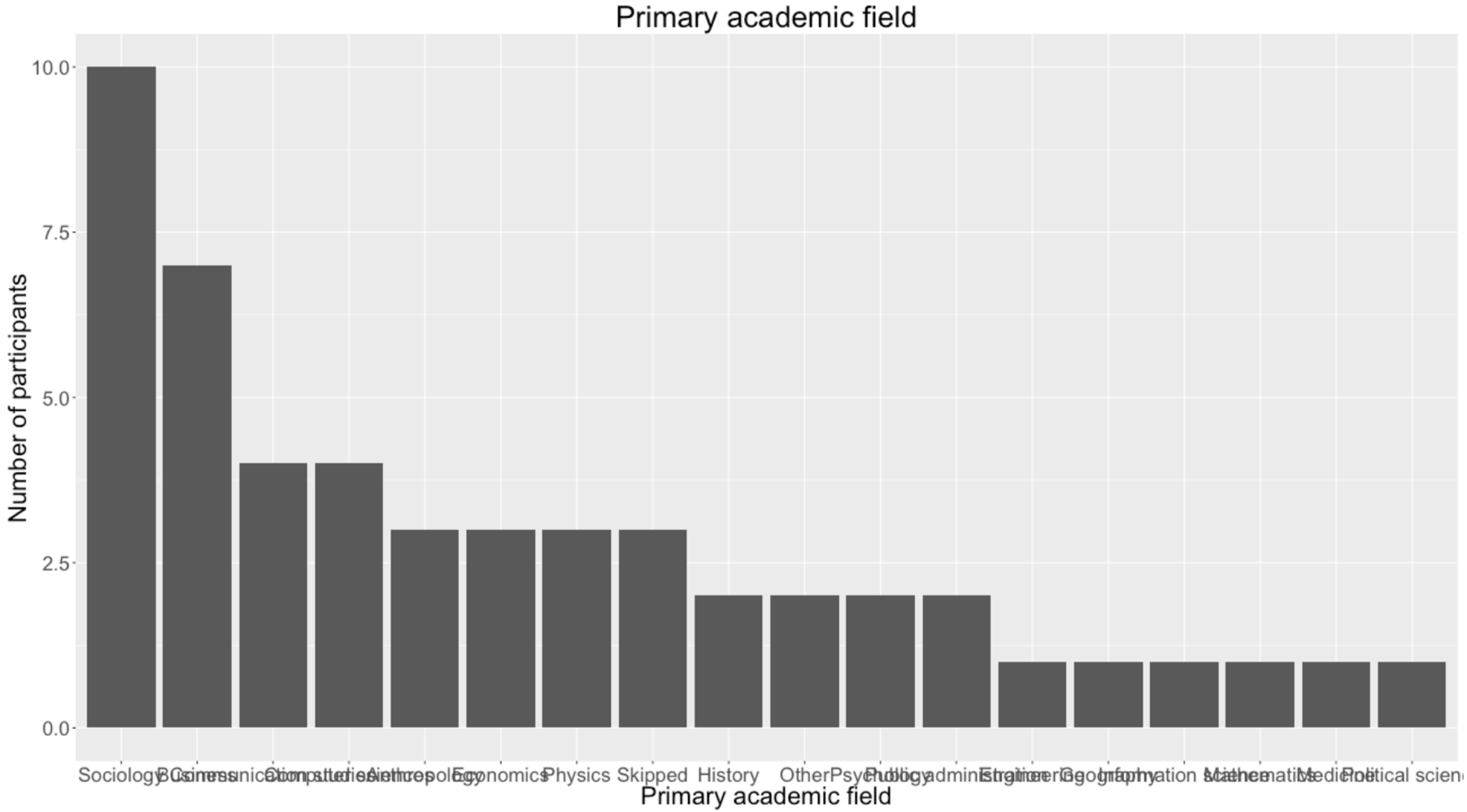

Primary academic field
Number of participants
10.0
7.5
5.0
2.5
0.0
Anthropology
Business
Communication
Computer science
Economics
Engineering
Geography
History
Information science
Mathematics
Medicine
Other
Physics
Political science
Psychology
Public administration
Social work
Sociology
Primary academic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

Order by value

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

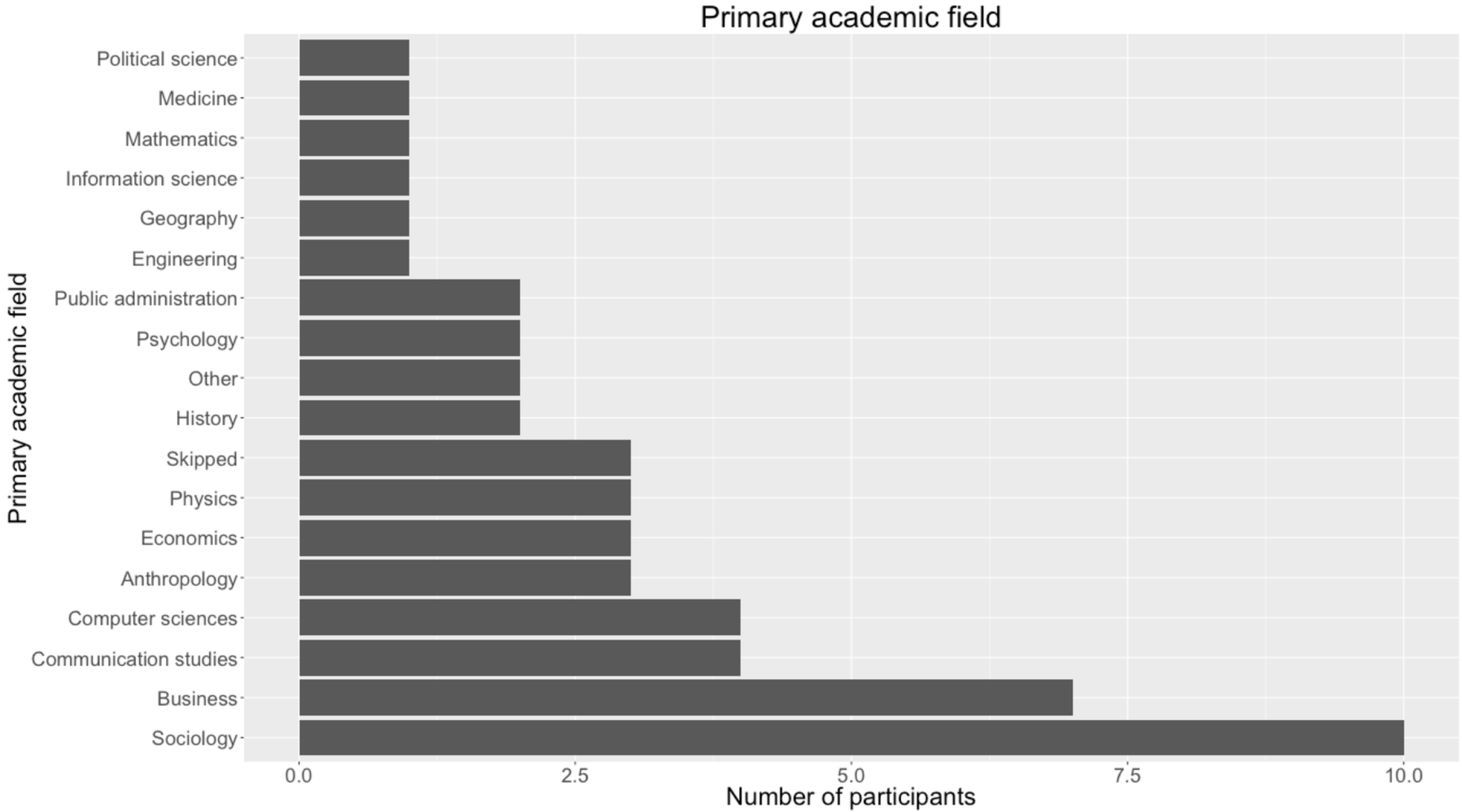

Primary academic field
Number of participants
10.0
7.5
5.0
2.5
0.0
Sociology
Business
Communication
Computer science
Anthropology
Economics
Physics
Skipped
History
Other
Psychology
Public administration
Engineering
Geography
Information science
Mathematics
Medicine
Political scien
Primary academic field

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

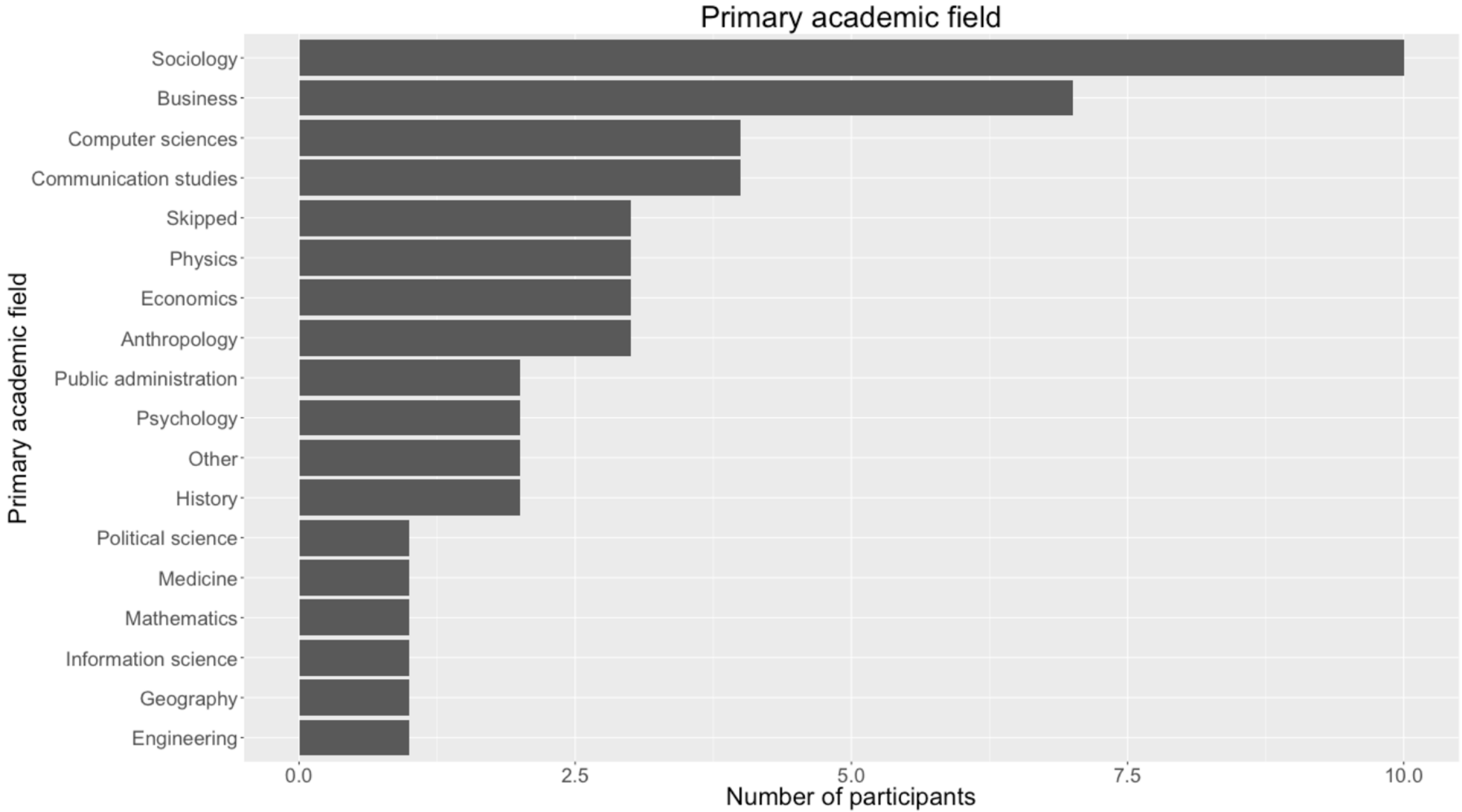
```
coord_flip()
```


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

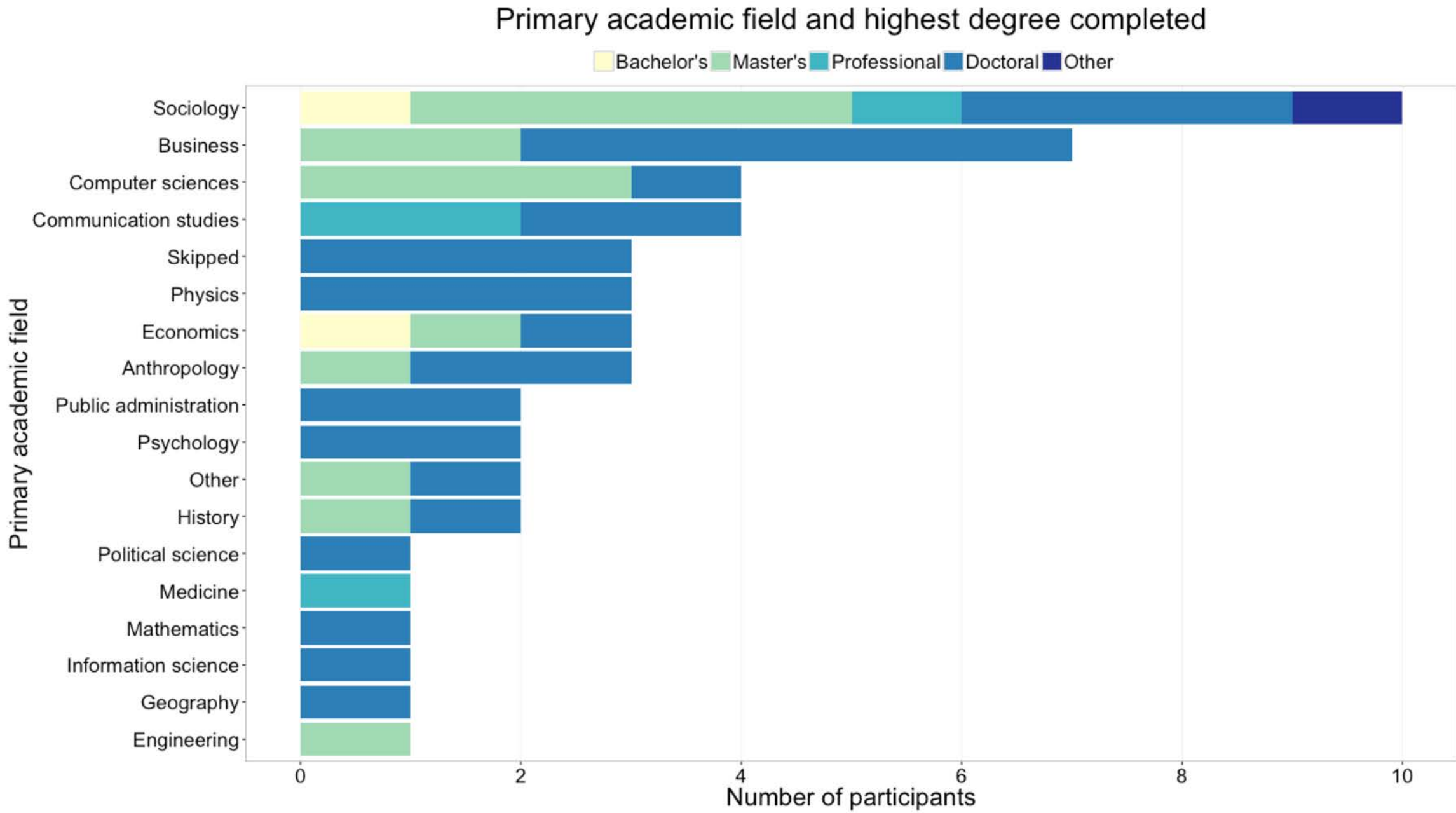
Oops!

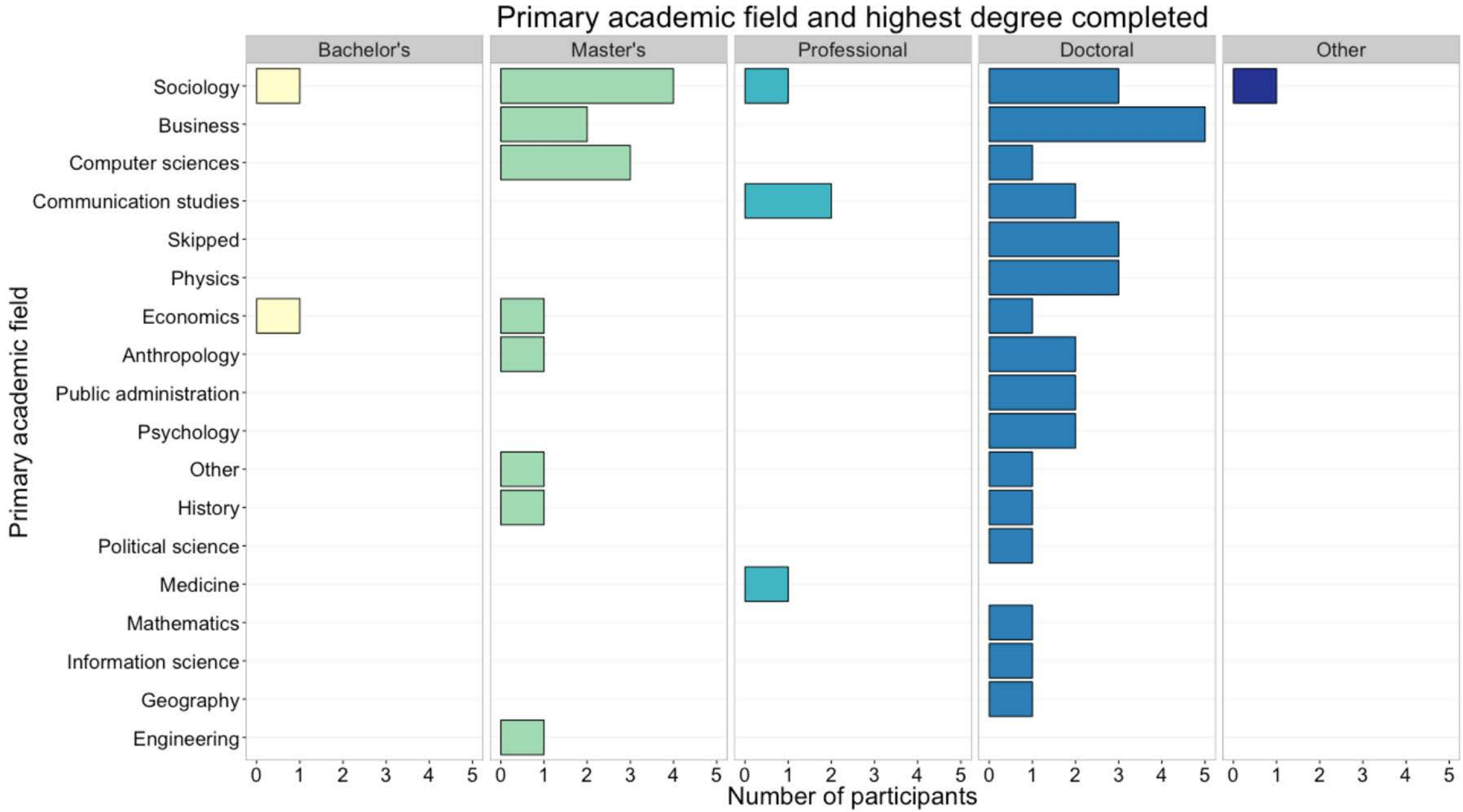
```
data$academic_field <-
     factor(data$academic_field,
            levels=names(
                sort(
                    table(data$academic_field),
                    decreasing=TRUE)))
```

```
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

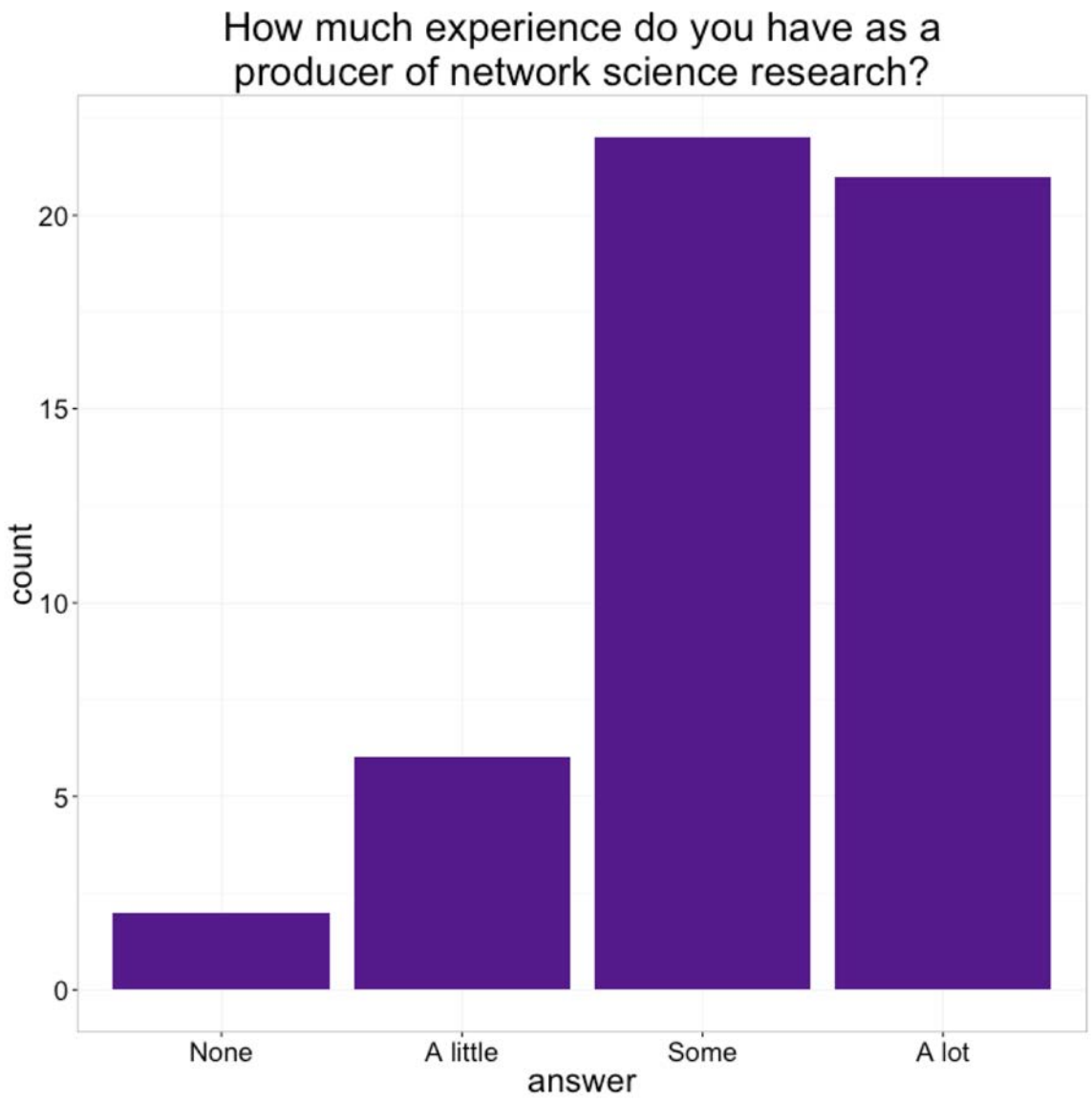

Primary academic field and highest degree completed
Bachelor's
Master's
Professional
Doctoral
Other
Primary academic field
Sociology
Business
Computer sciences
Communication studies
Skipped
Physics
Economics
Anthropology
Public administration
Psychology
Other
History
Political science
Medicine
Mathematics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

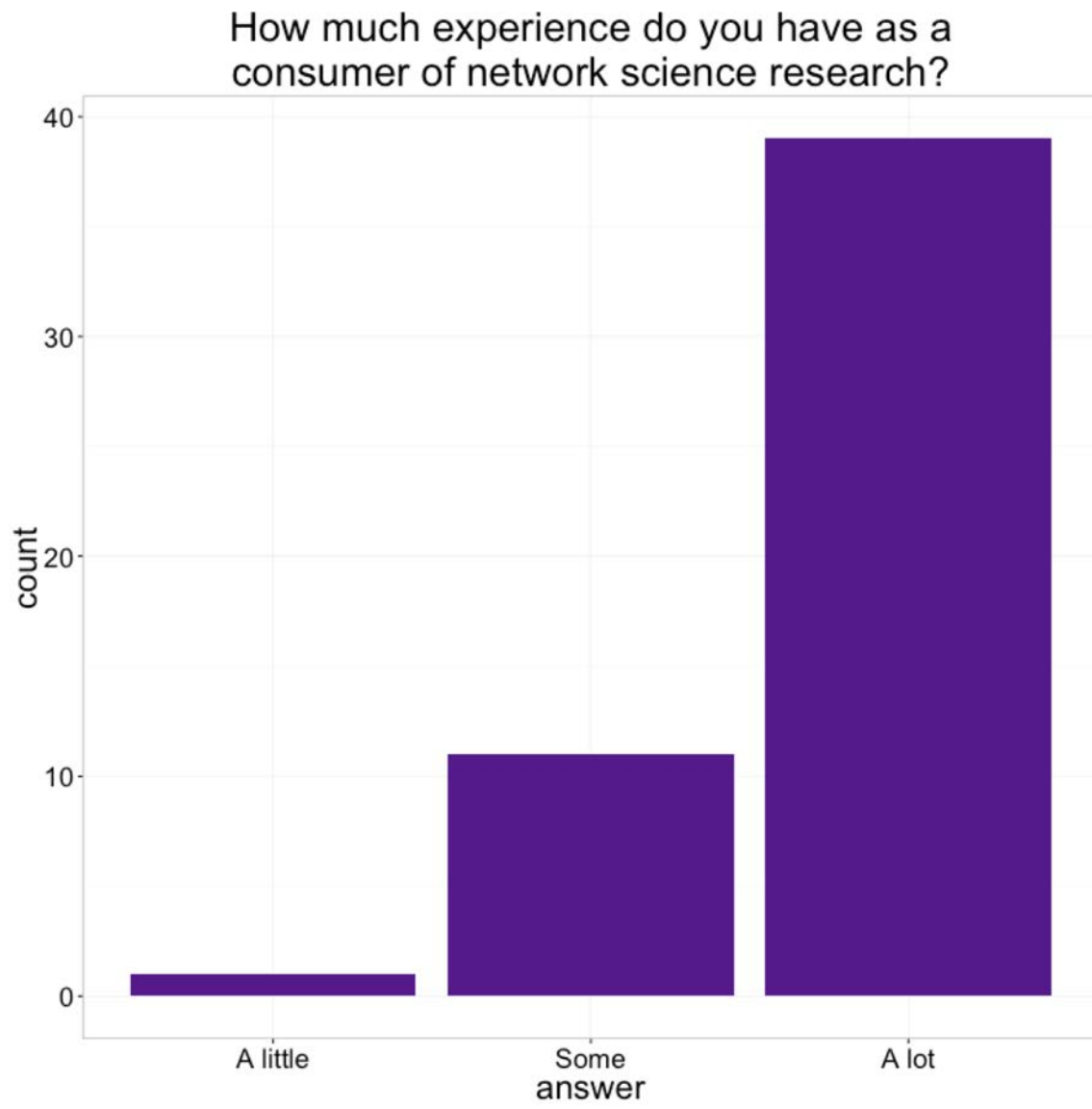
Different placement helps with different comparisons

```
fill=highest_degree
```

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

Principle 4:
Keep scales consistent

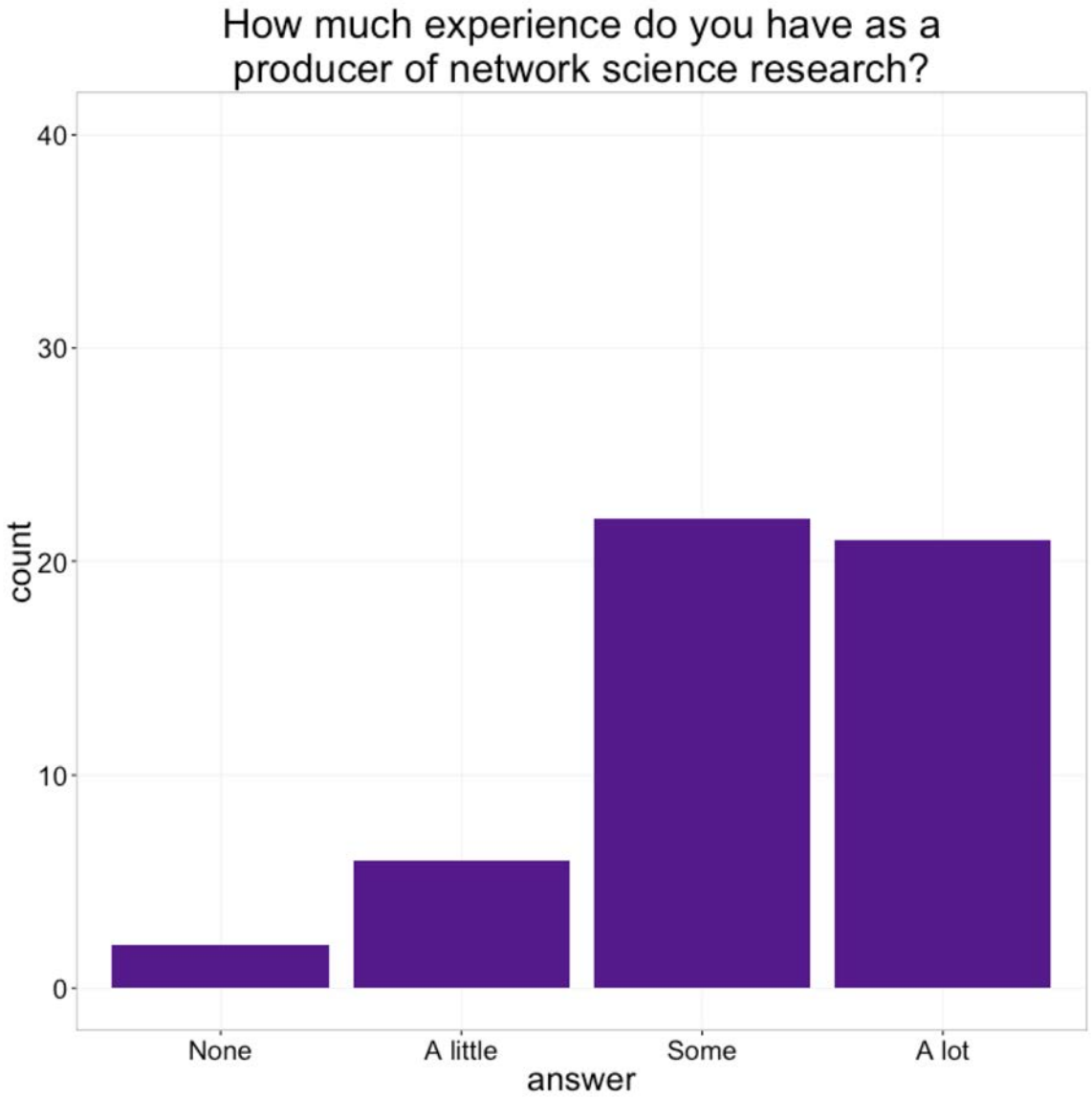

How much experience do you have as a producer of network science research?
count
0
5
10
15
20
None
A little
Some
A lot
answer

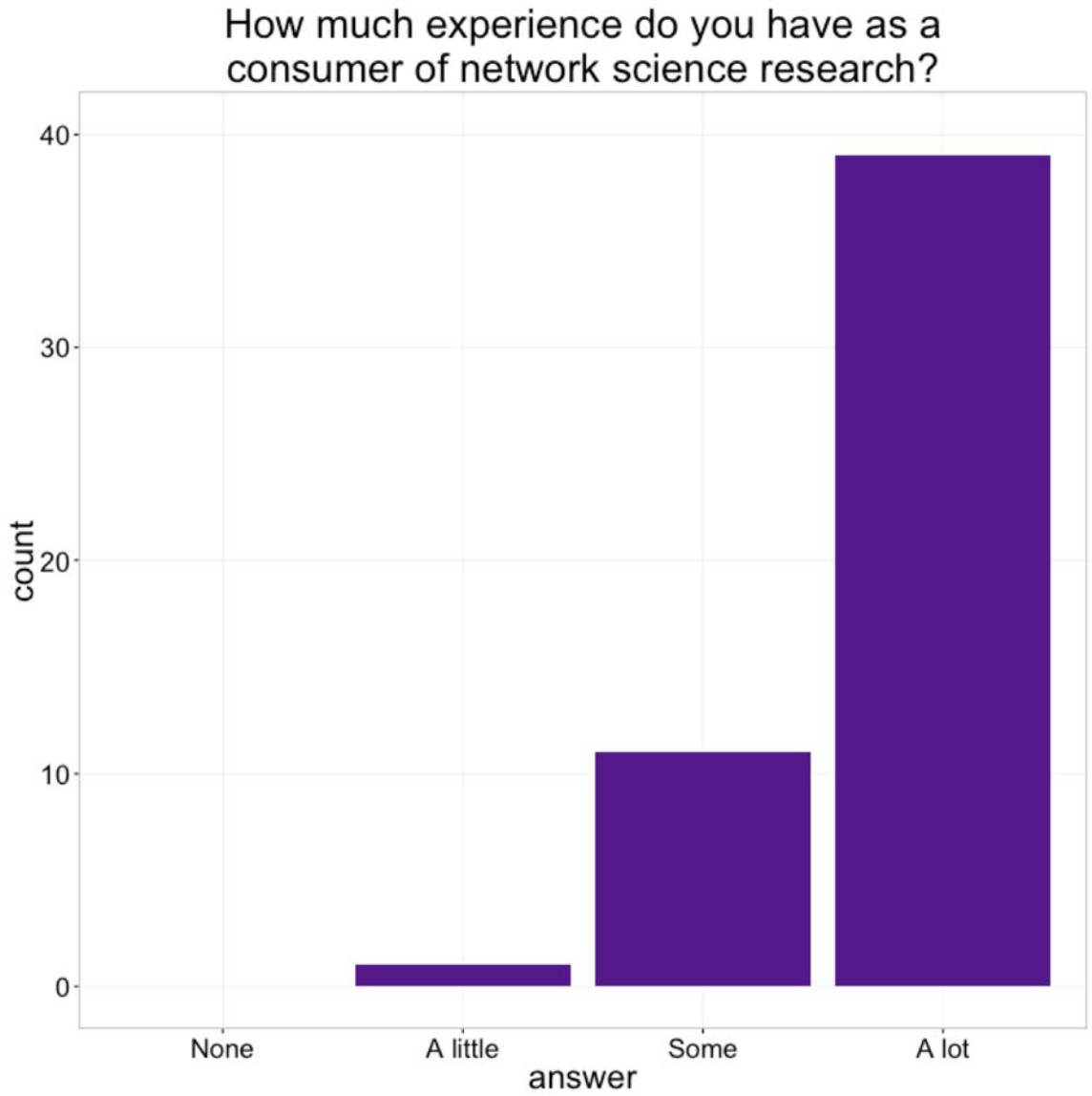

How much experience do you have as a consumer of network science research?
count
0
10
20
30
40
A little
Some
A lot
answer

Keep all categories, manually set axes

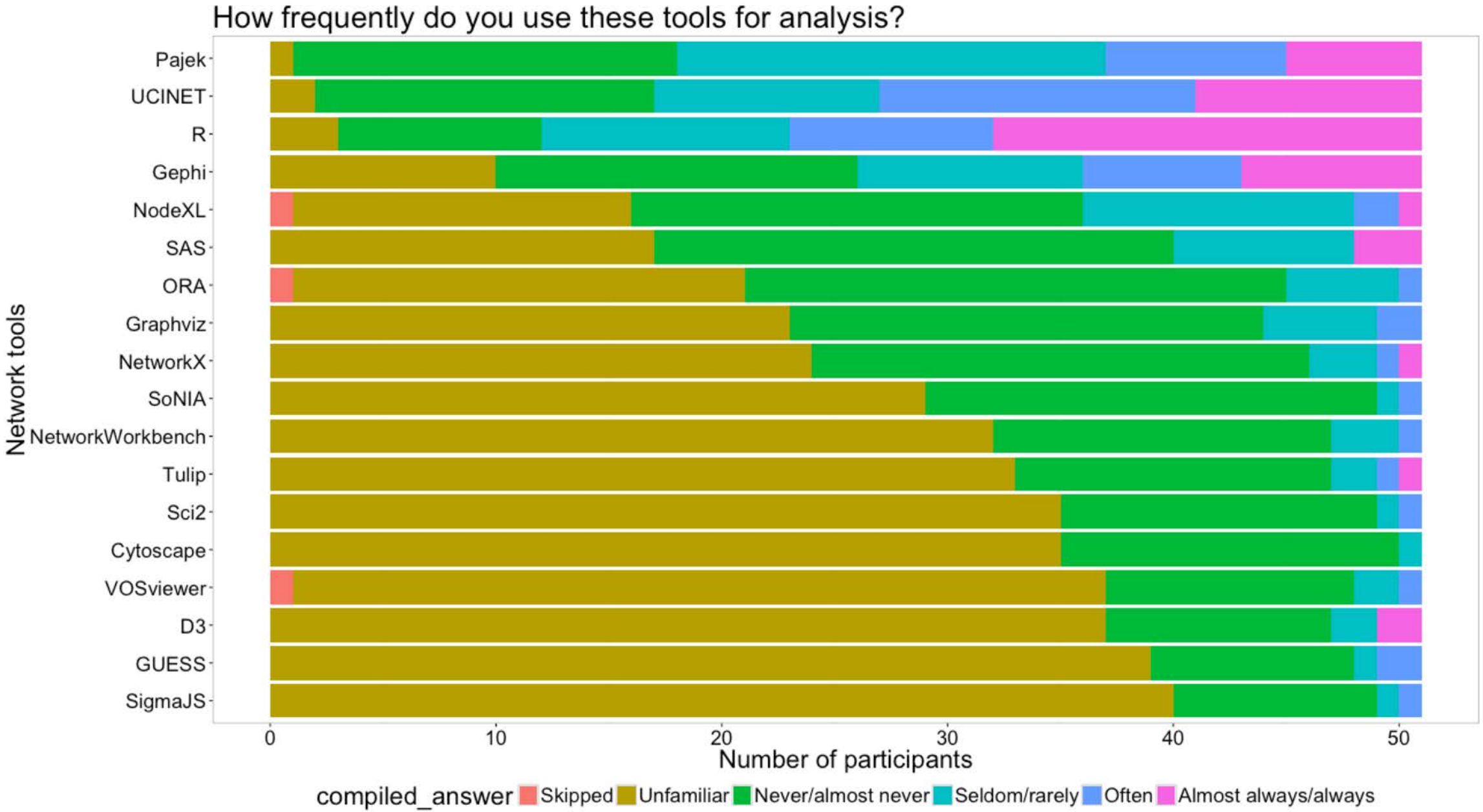
```
scale_x_discrete(drop=FALSE)

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

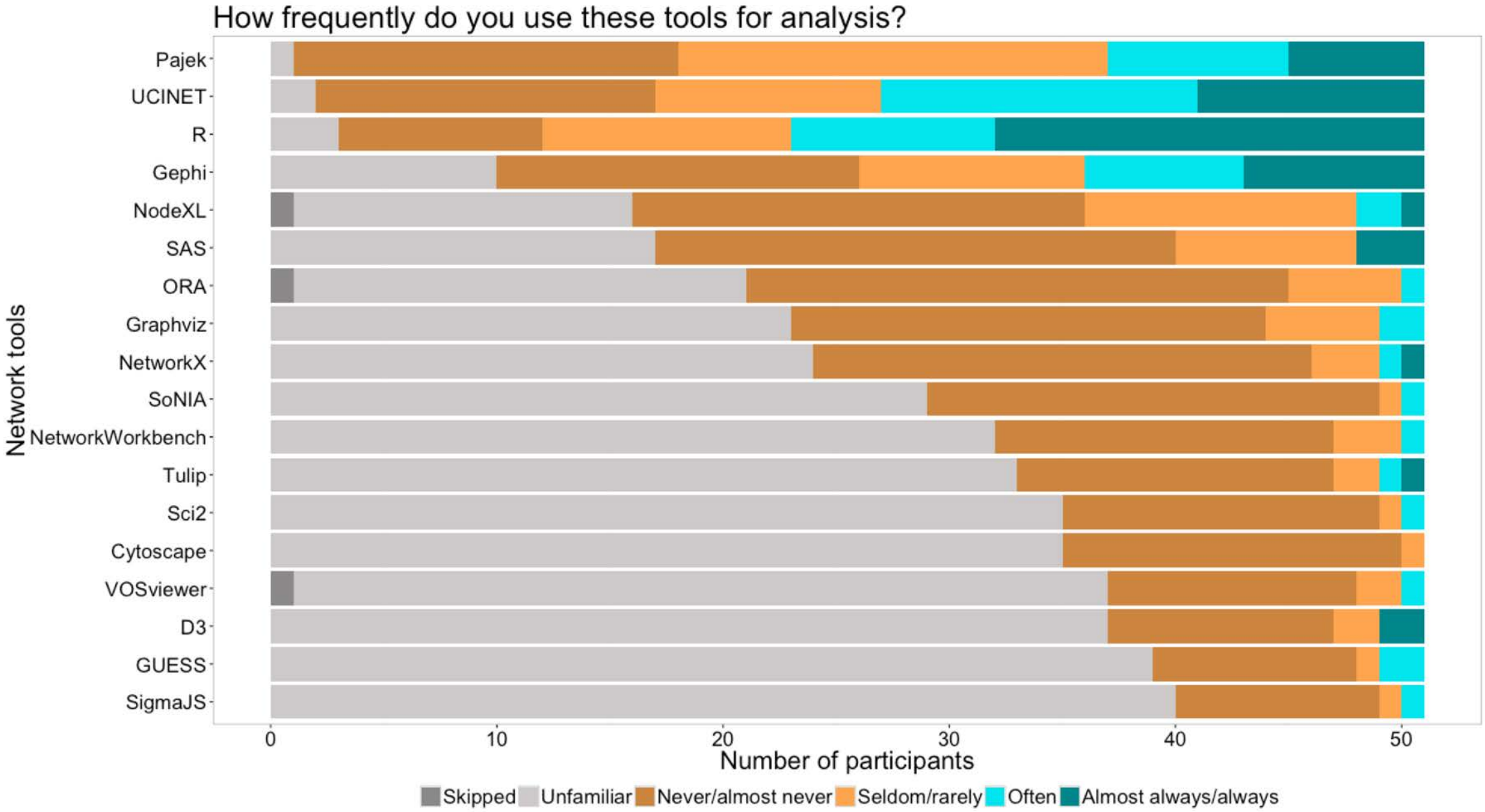

How frequently do you use these tools for analysis?
Network tools
Pajek
UCINET
R
Gephi
NodeXL
SAS
ORA
Graphviz
NetworkX
SoNIA
NetworkWorkbench
Tulip
Sci2
Cytoscape
VOSviewer
D3
GUESS
SigmaJS
0
10
20
30
40
50
Number of participants
compiled_answer
Skipped
Unfamiliar
Never/almost never
Seldom/rarely
Often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

Dataset 3:
Star Wars opinion survey

https://fivethirtyeight.com/features/americas-favorite-star-wars-movies-and-least-favorite-characters/

ggplot2: Chart quirks

See “templates” file

https://github.com/amzoss/RVis-DM2018/blob/master/Day%201/templates/templates.md

Chart components/slots

Bar chart, for example:

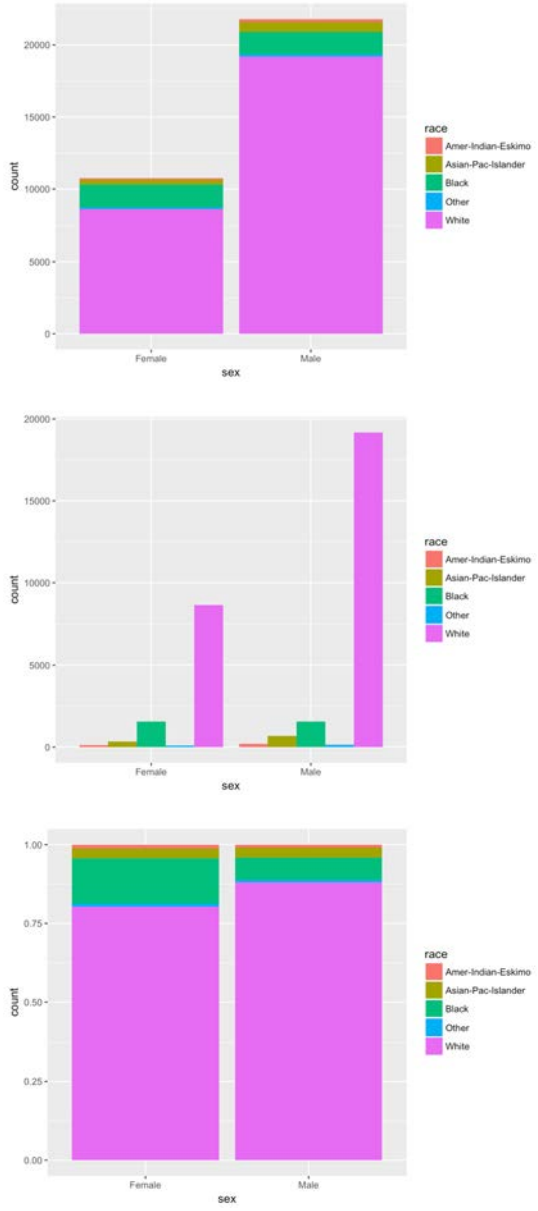
- x
  category (the names of the bars)

- y (optional)
  default is count, but can also specify a number (the length of the bars)

- color (optional)
  category (grouped or stacked bars)

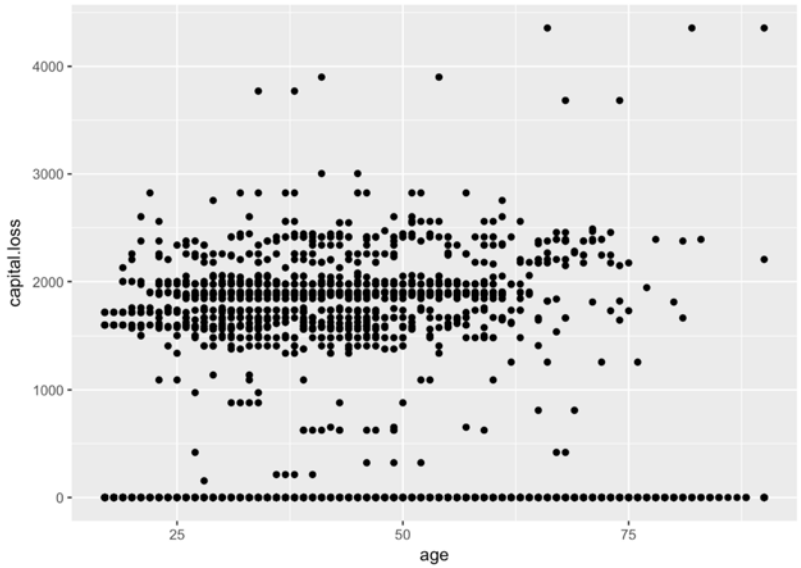
Bar chart

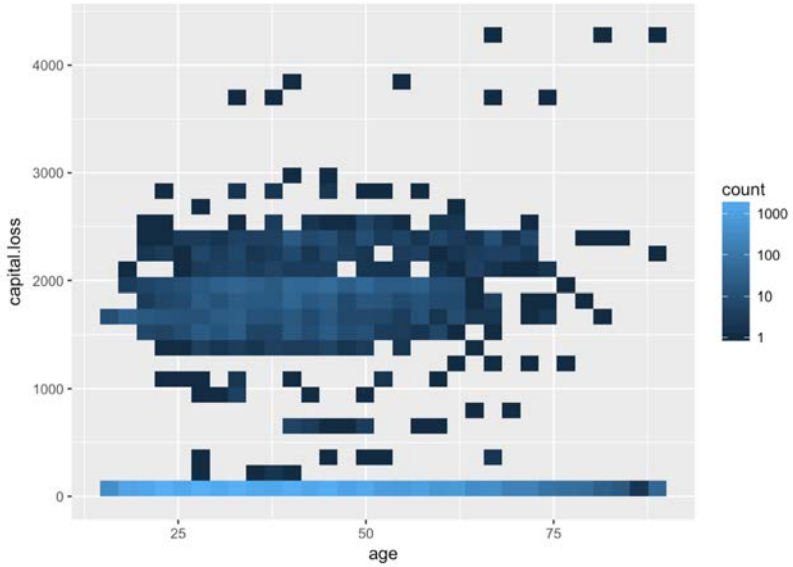
- geom_bar() vs. geom_col()
- count vs. identity vs. summary
- categorical vs. continuous
- stack vs. dodge vs. fill
- bar vs. pie



Scatter plot

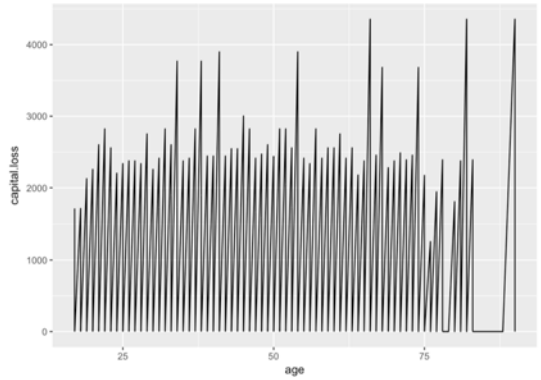
- Overplotting
- point vs. bin2d

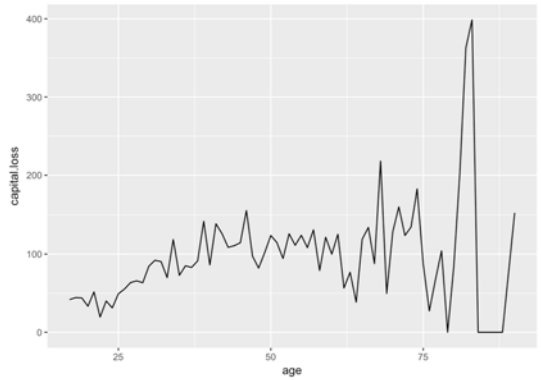


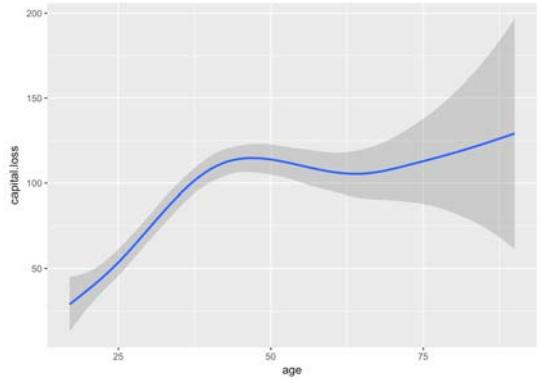


Line chart

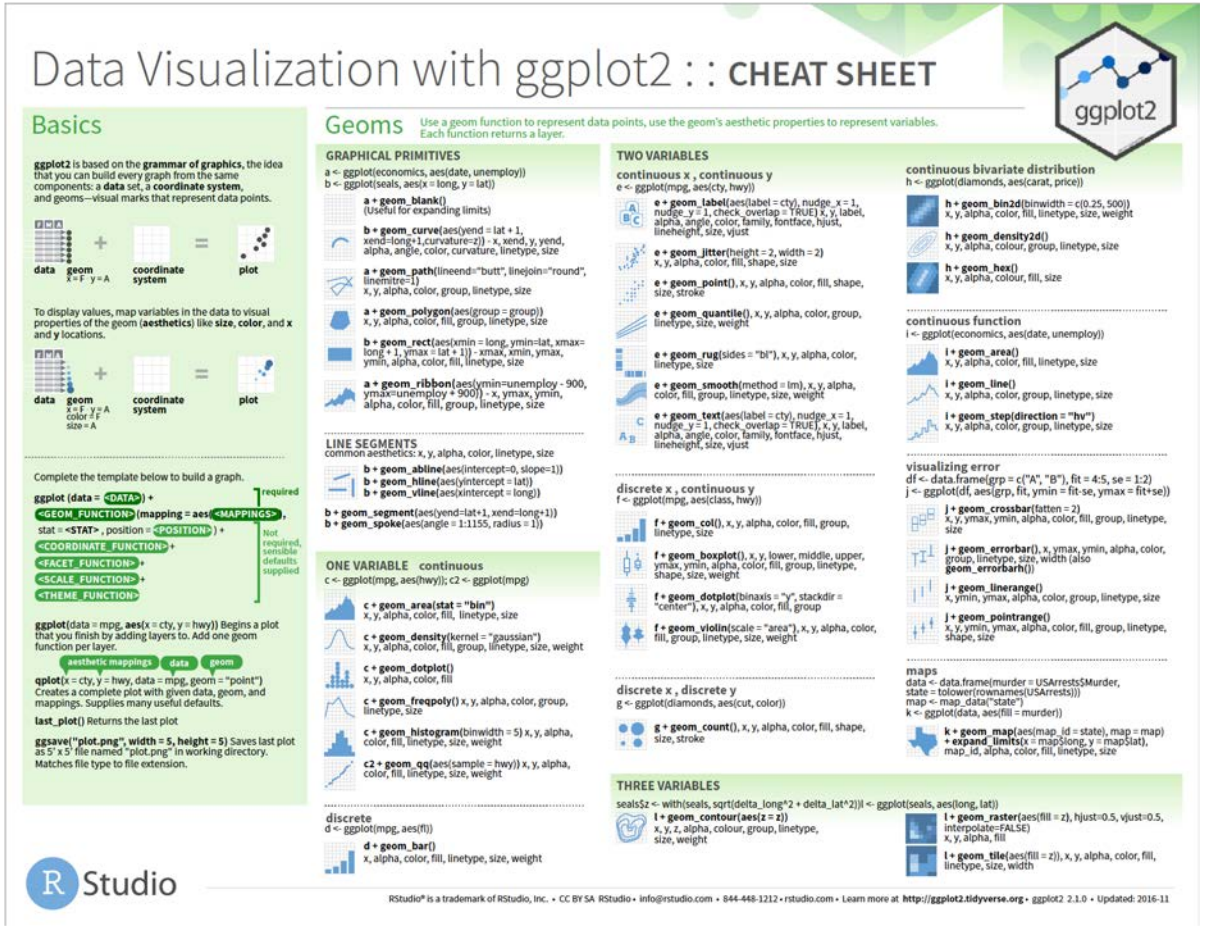
- identity vs. summary
- line vs. smooth







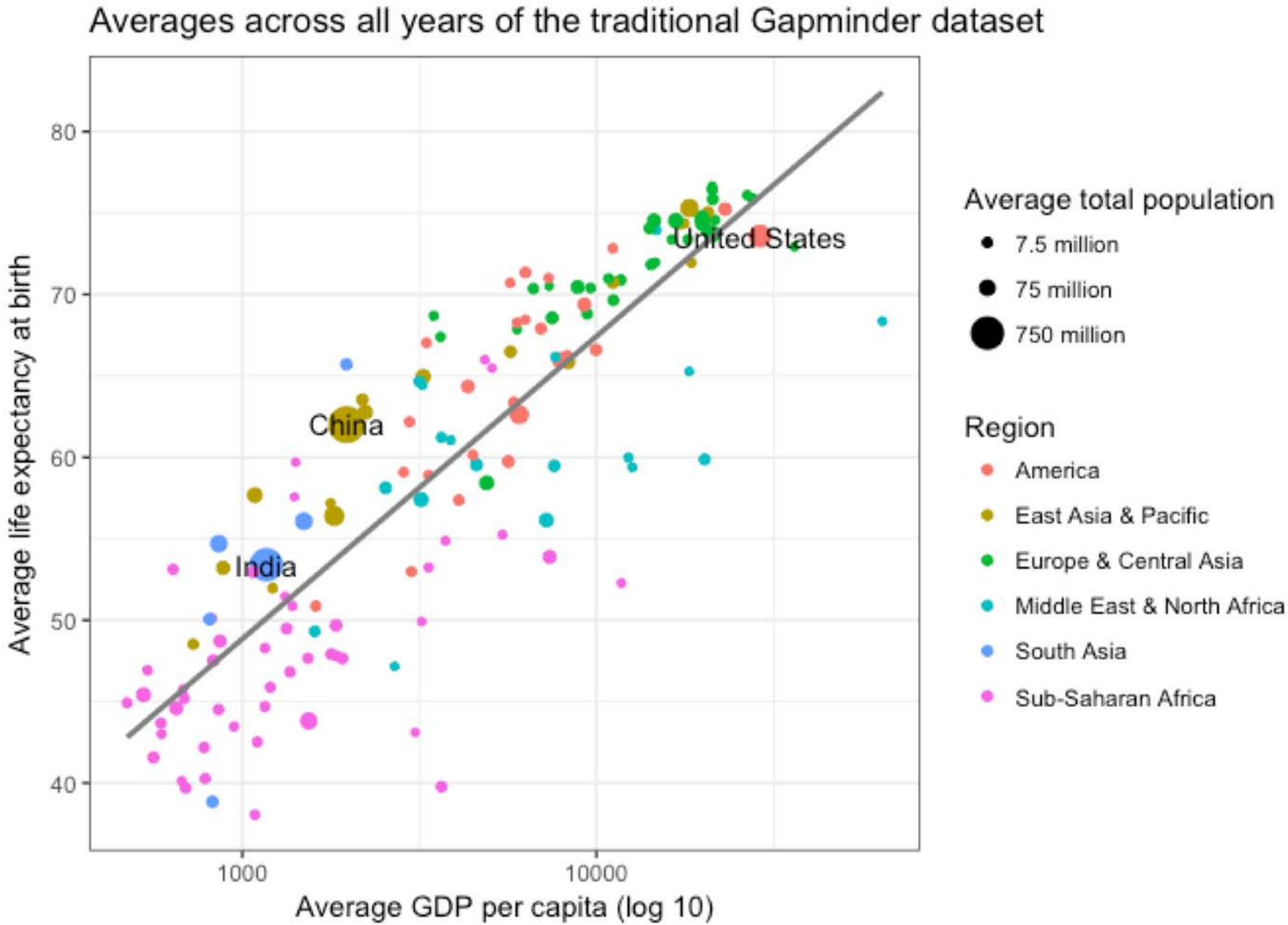
ggplot2 Cheat Sheet



https://www.rstudio.com/resources/cheatsheets/#ggplot2

Dataset 4: Gapminder Data

http://www.gapminder.org/

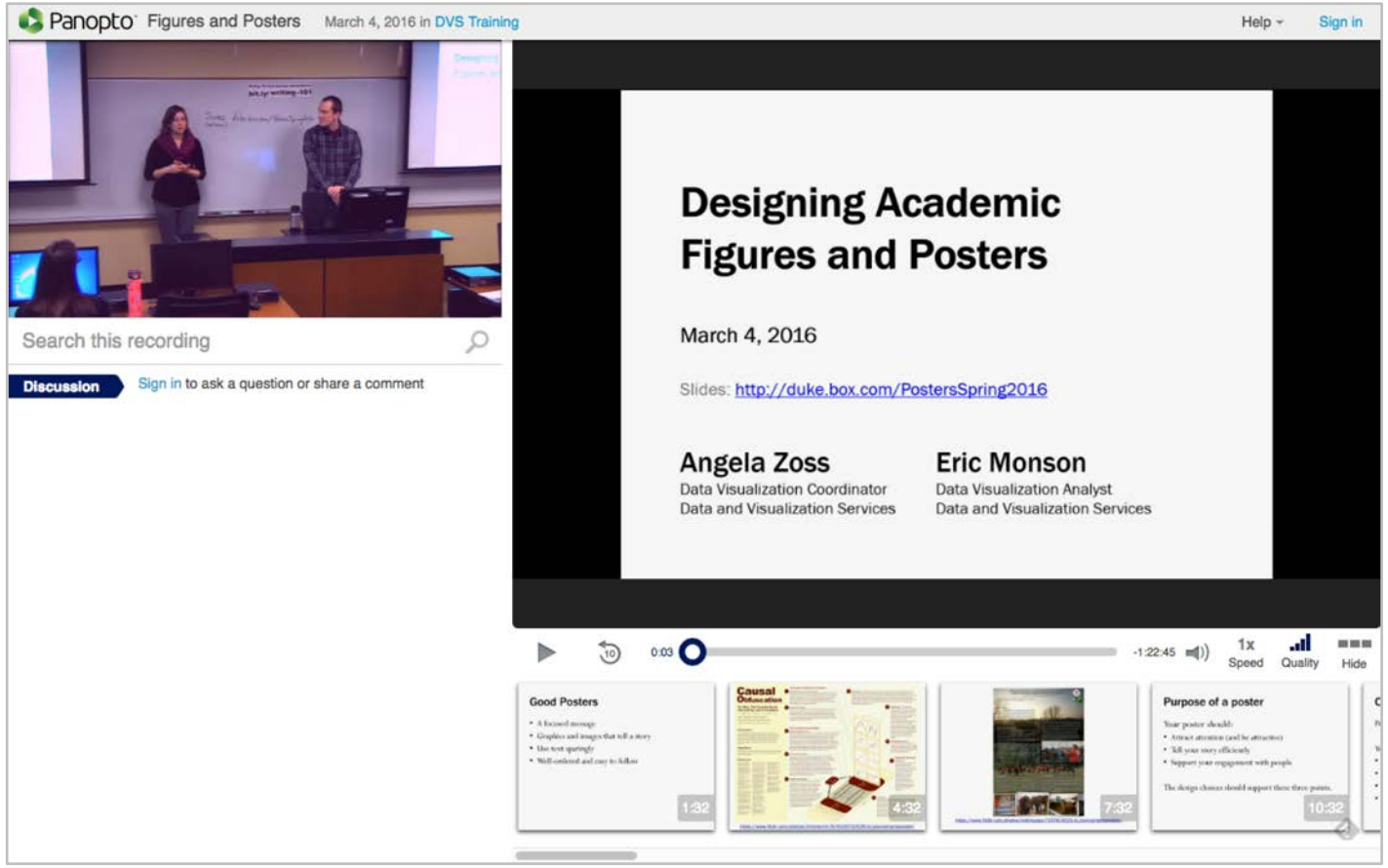

Averages across all years of the traditional Gapminder dataset
Average life expectancy at birth
80
70
60
50
40
1000
10000
Average GDP per capita (log 10)
China
India
United States
Average total population
7.5 million
75 million
750 million
Region
America
East Asia & Pacific
Europe & Central Asia
Middle East & North Africa
South Asia
Sub-Saharan Africa

Advanced topics:
Mapping, saving charts out

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

Videos of past workshops



http://bit.ly/DVSvideos

Thanks for your feedback!

angela.zoss@duke.edu