Visualization in R using ggplot2

Angela Zoss

September 25, 2017

https://github.com/amzoss/ggplot2-F17

Set up environment

- R?
- RStudio?
- tidyverse?

If you haven’t installed RStudio, try Docker:

https://vm-manage.oit.duke.edu/containers

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!

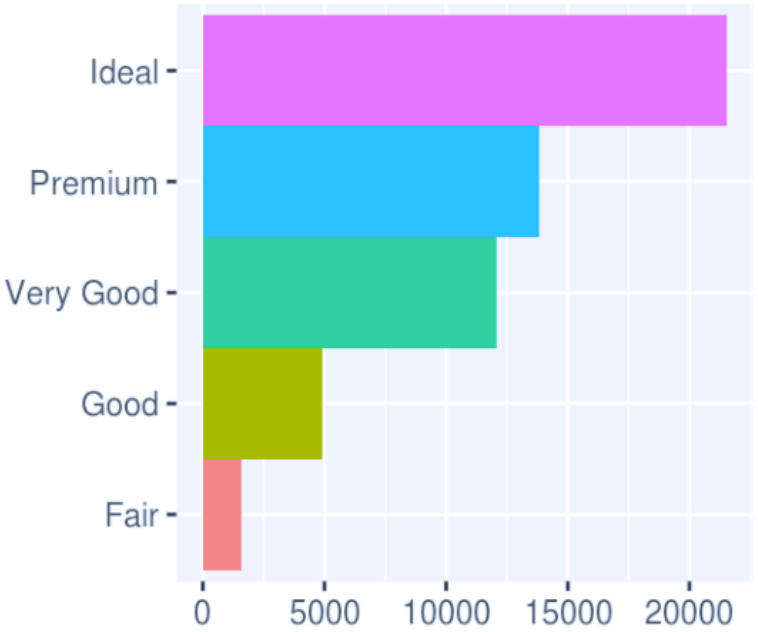
*"Your closest collaborator is **you** six months ago, but you don't reply to emails."*

- 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

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/

Get workshop files

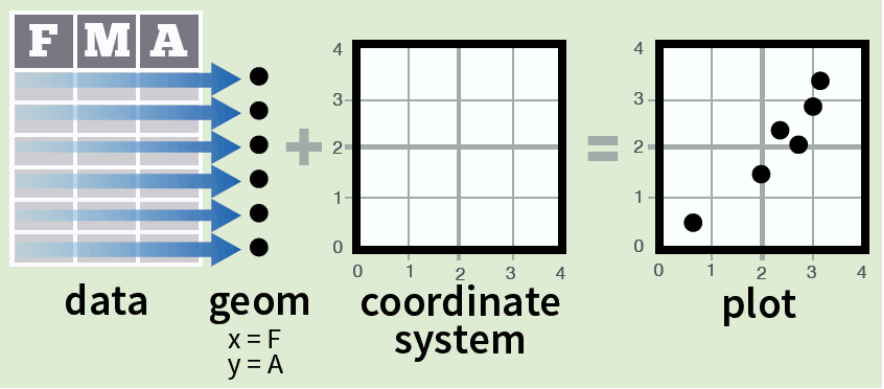
In RStudio:

- Project → New project
- Version Control
- Git
  - URL: https://github.com/amzoss/ggplot2-F17
  - Project directory name: ggplot2-F17
  - Subdirectory: you choose
- Create Project

ggplot2: Elements

Basic elements in any ggplot2 visualization

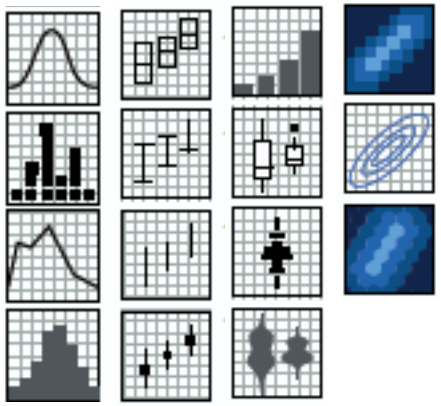
- **data**
- **aes**thetics
  (variable mappings)
- **geom**
  (chart type or shape)
- coordinate system
  (the arrangement of the marks;
  most geoms use default, cartesian)



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.

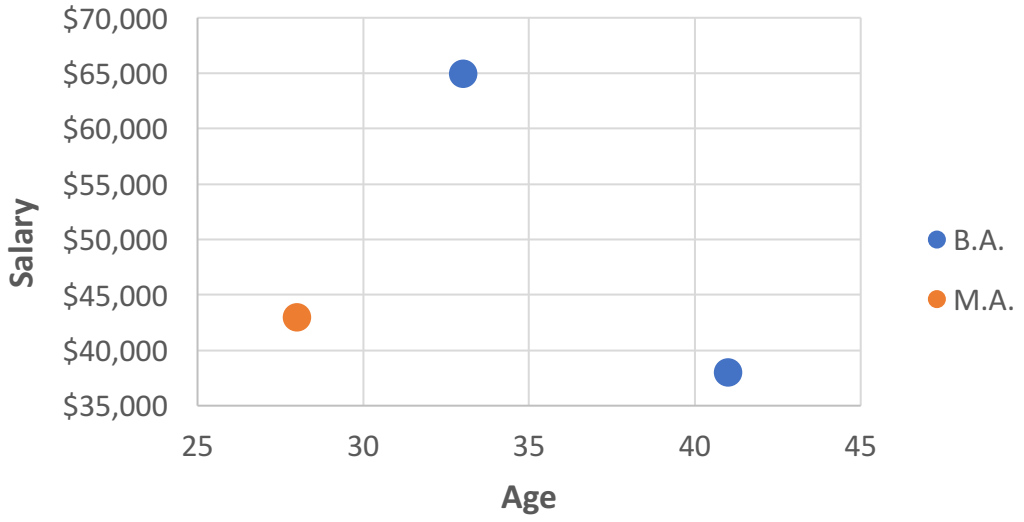
ggplot2: Basic syntax

Template for a simple plot

```
ggplot()  +

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

Aesthetic variable mappings

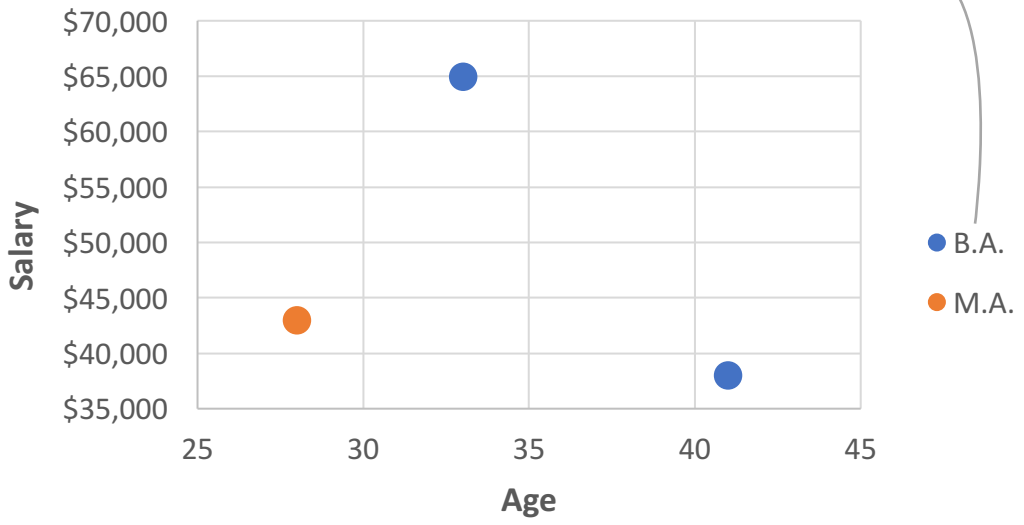
Name	Age	Salary	Highest Degree
Jane Smith	33	$65,000	B.A.
Abby Jones	28	$43,000	M.A.
Bridget Carden	41	$38,000	B.A.



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

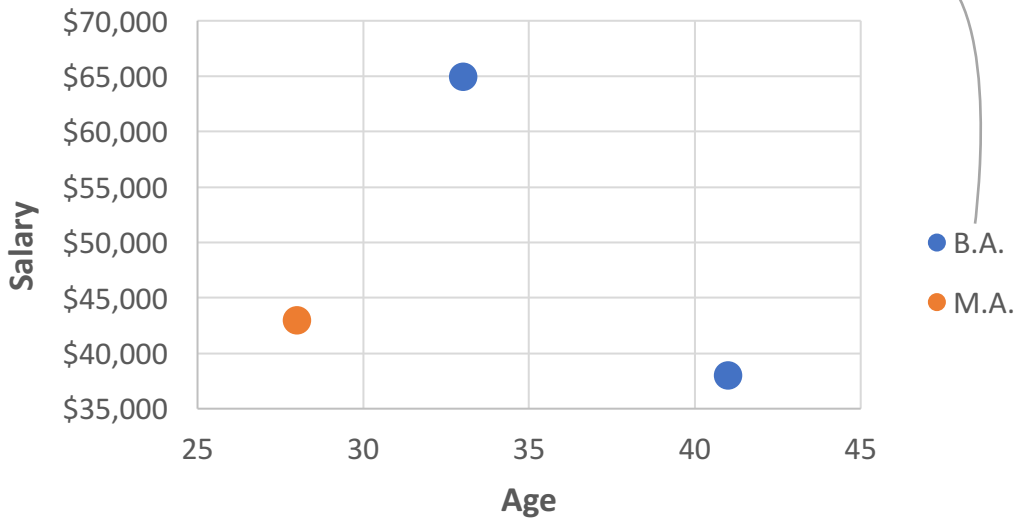


Aesthetic variable mappings

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

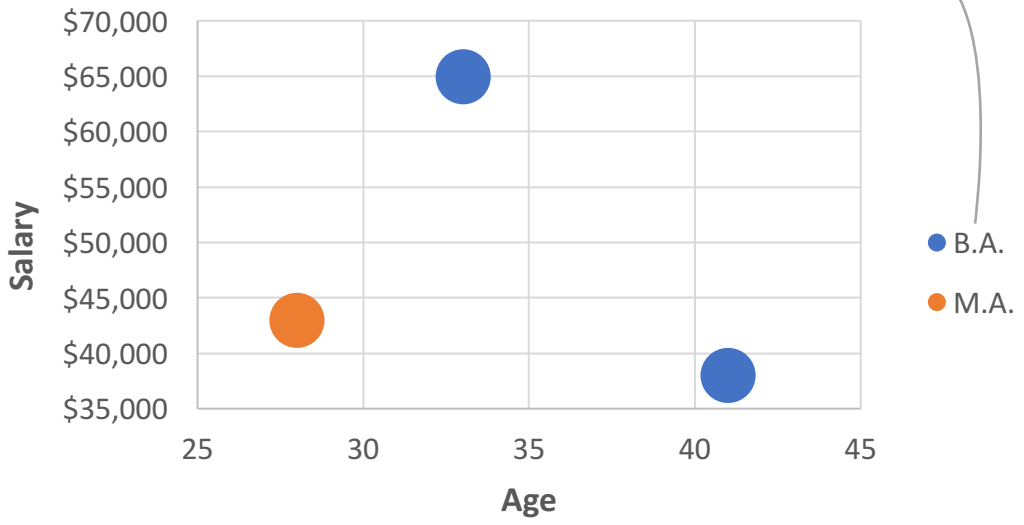


Non-variable adjustments

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

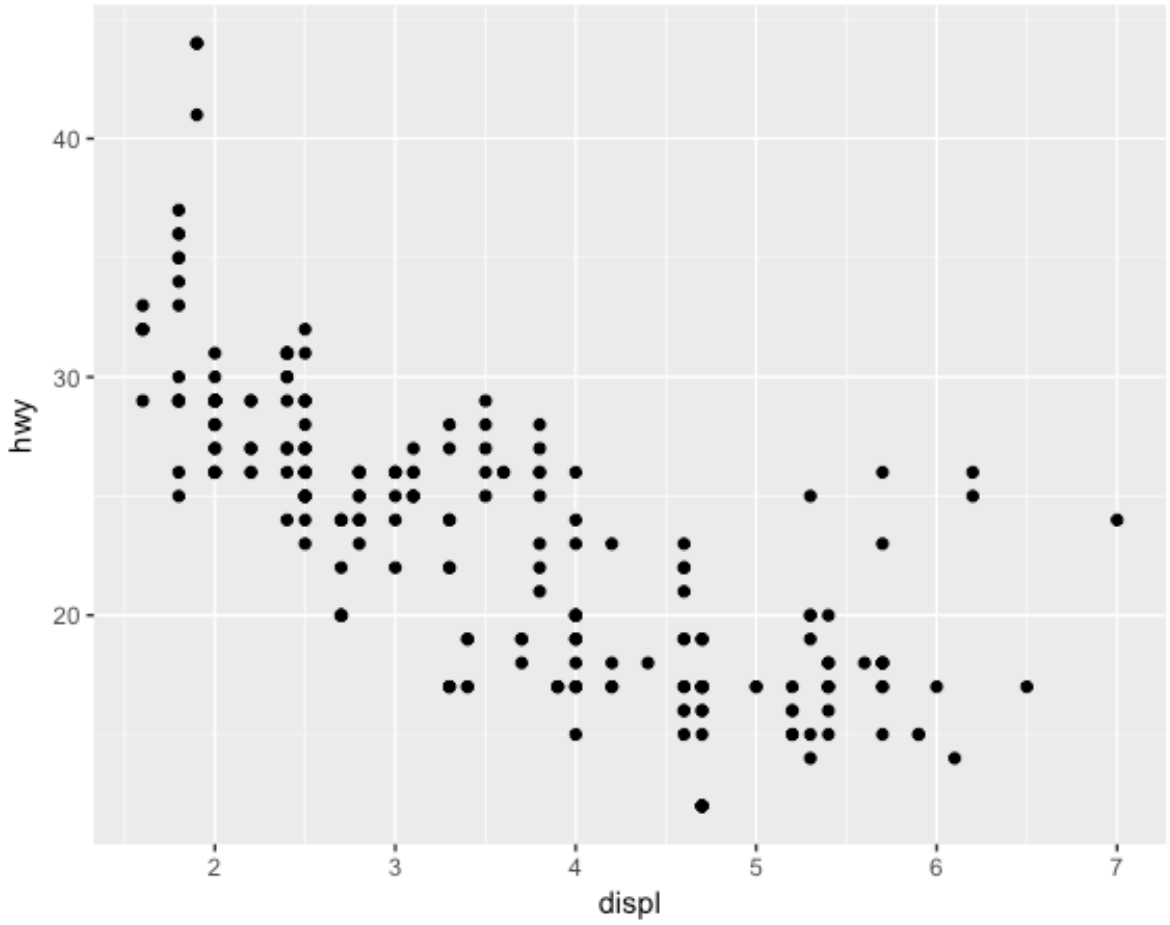
carry through
from top to bottom

```
ggplot( [data = data frame]
        [aes(variable mappings)] )
+
geom_...( [aes(add'l variable mappings)]
          [non-variable adjustments] )
+
geom_...( [aes(add'l variable mappings)]
          [non-variable adjustments] )
```

ggplot2: Building a plot

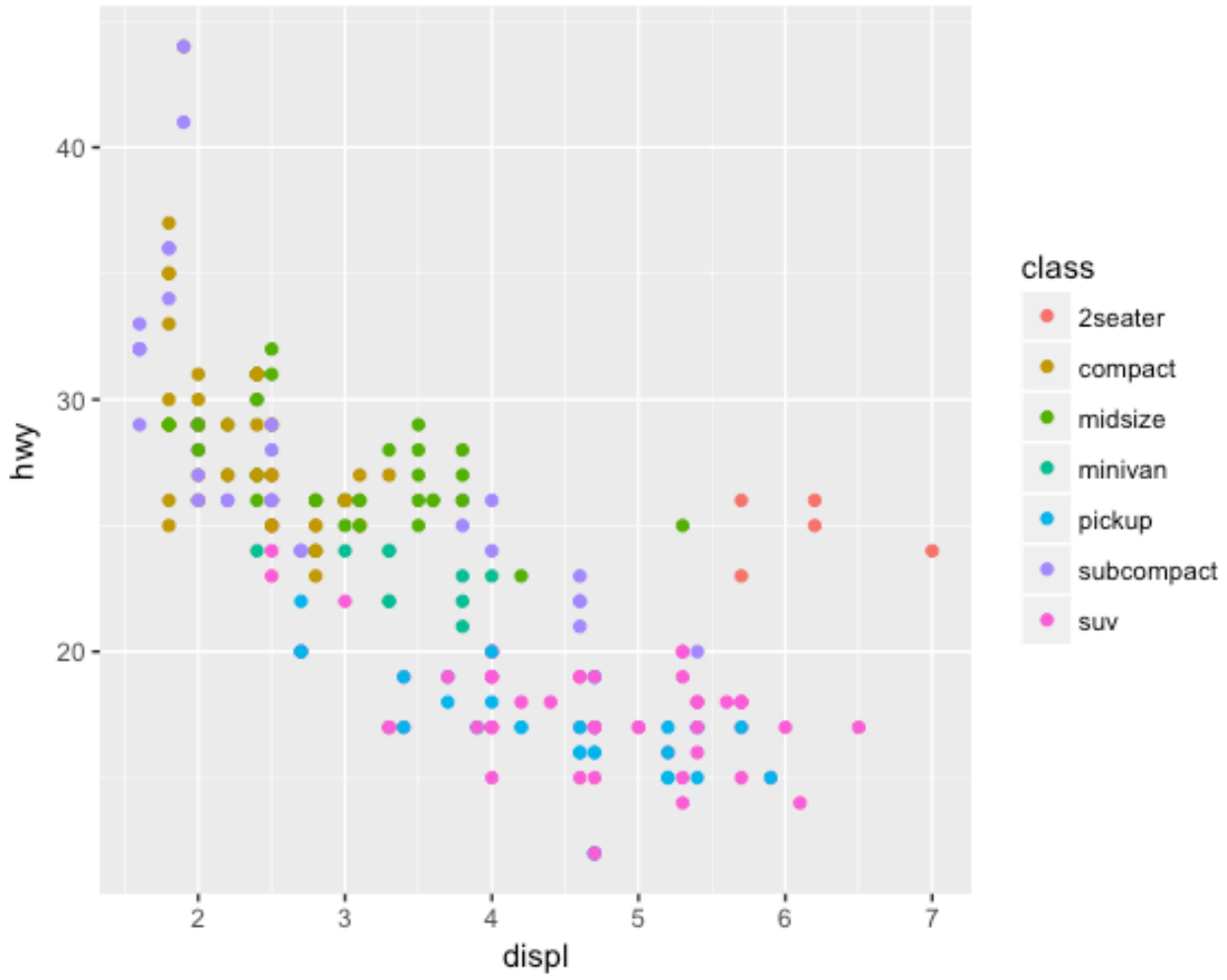
Follow along in an empty R script

```
library(ggplot2)


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



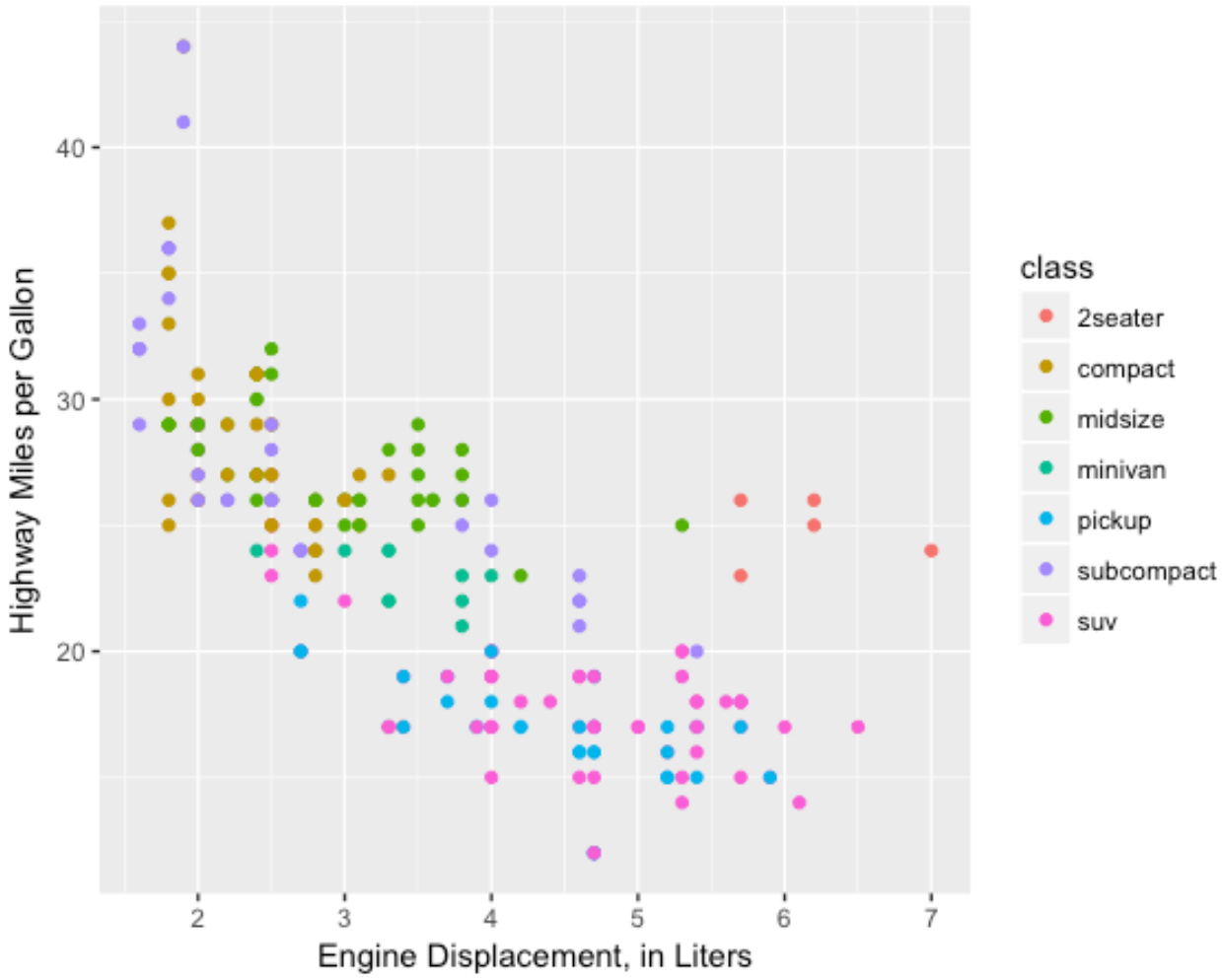
http://r4ds.had.co.nz/graphics-for-communication.html

```
library(ggplot2)


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



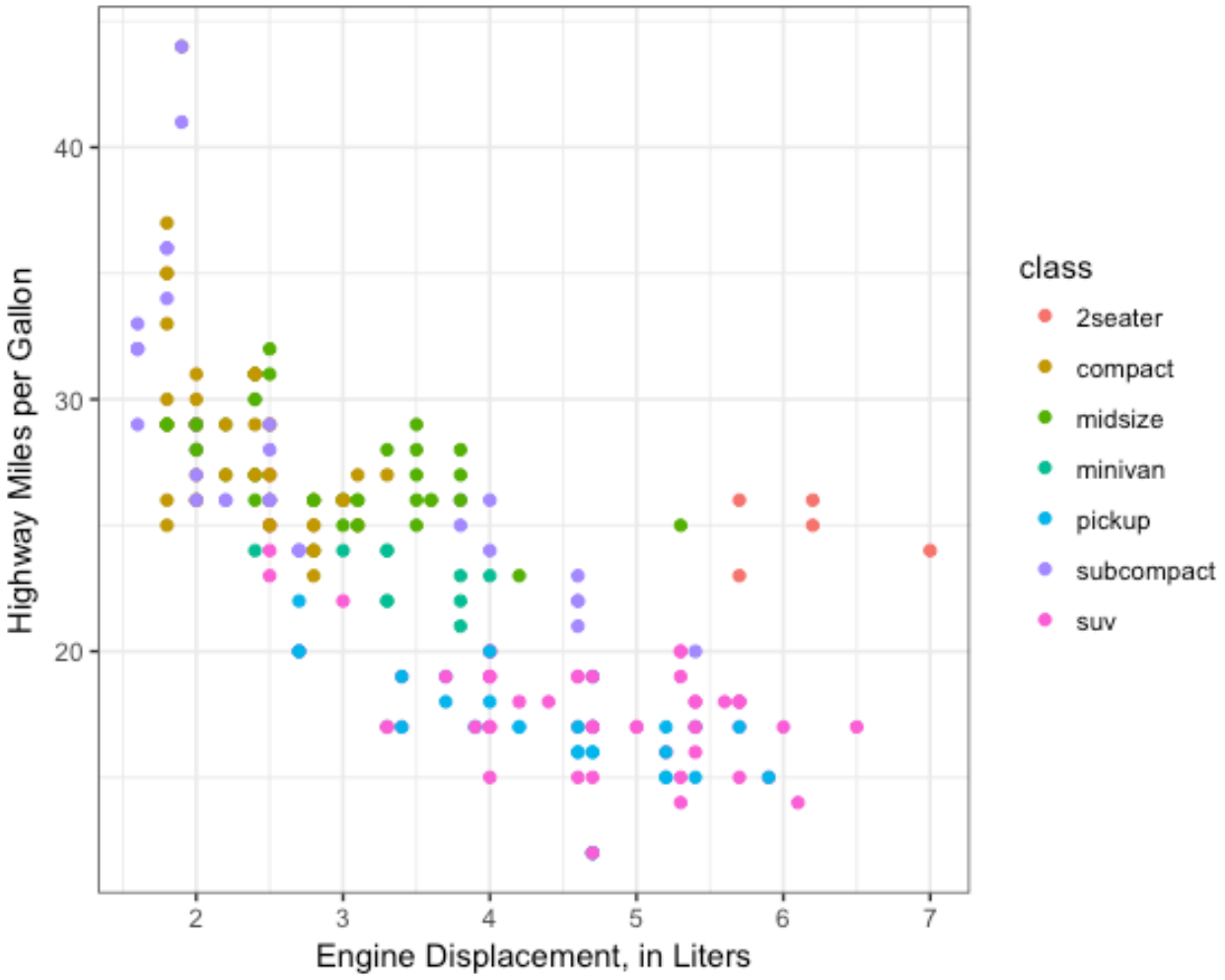
http://r4ds.had.co.nz/graphics-for-communication.html

```
library(ggplot2)


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



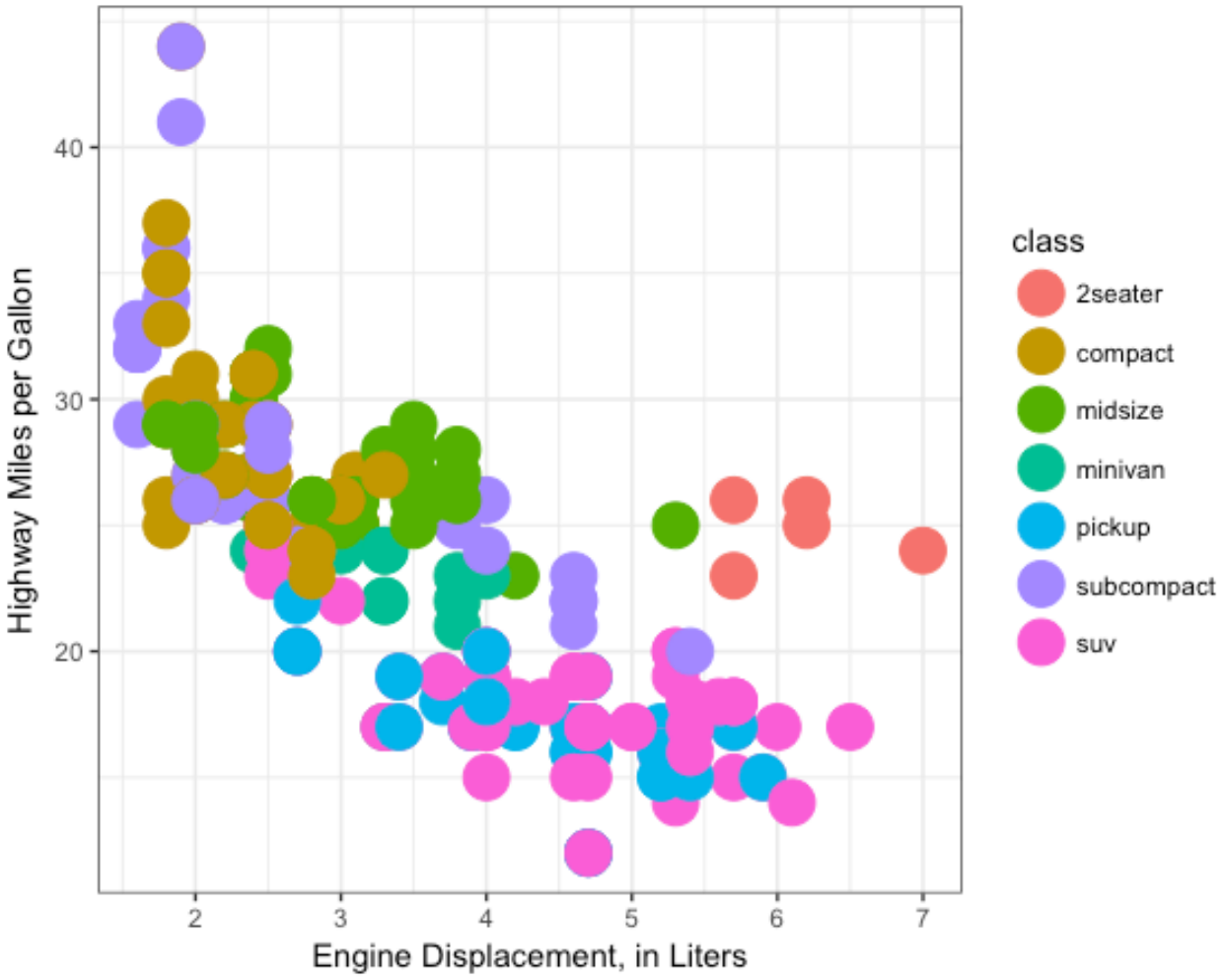
http://r4ds.had.co.nz/graphics-for-communication.html

```r
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()
```



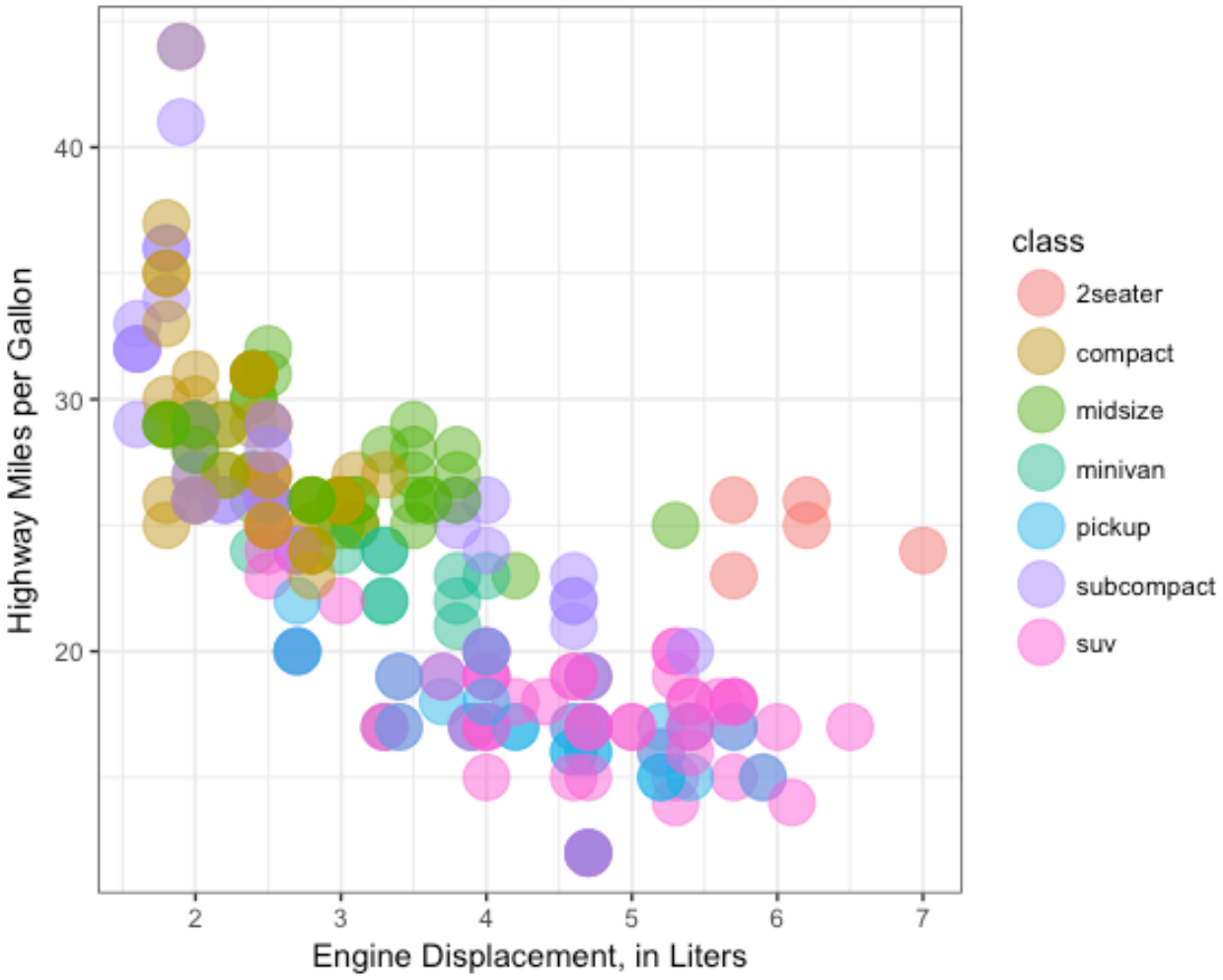
http://r4ds.had.co.nz/graphics-for-communication.html

```
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()
```



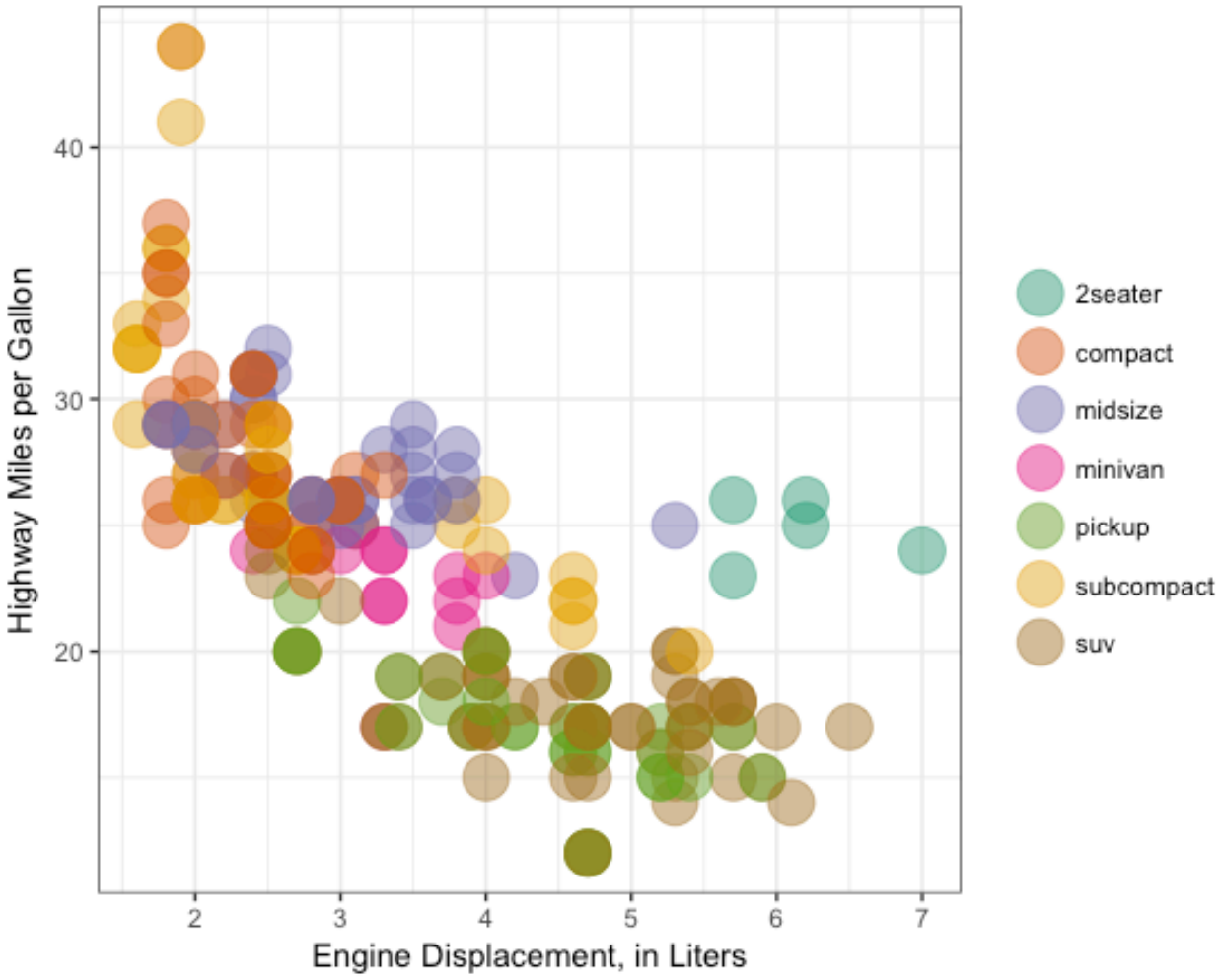
http://r4ds.had.co.nz/graphics-for-communication.html

```
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()
```



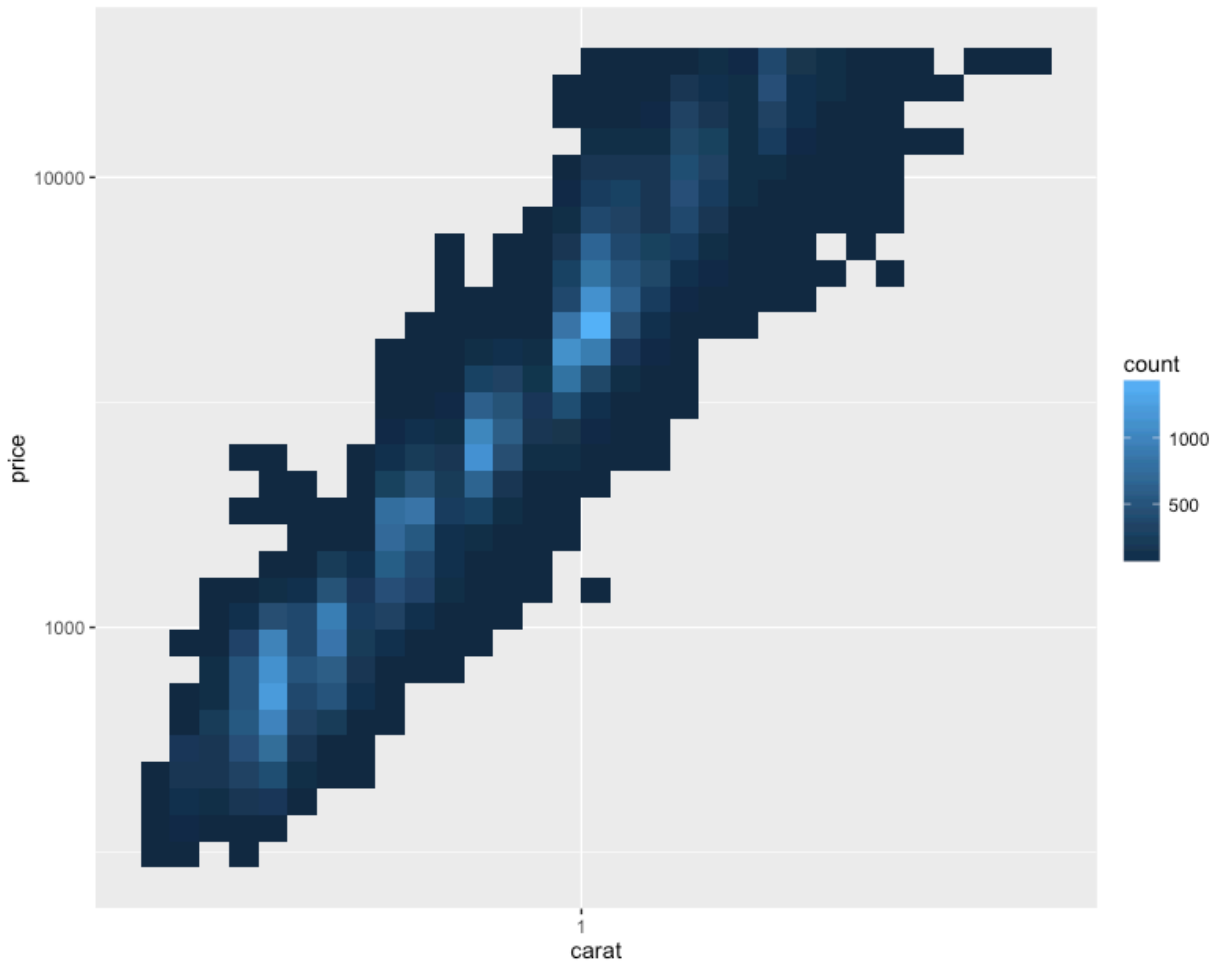
http://r4ds.had.co.nz/graphics-for-communication.html

```r
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()
```



http://r4ds.had.co.nz/graphics-for-communication.html

```r
# 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()
```



http://r4ds.had.co.nz/graphics-for-communication.html

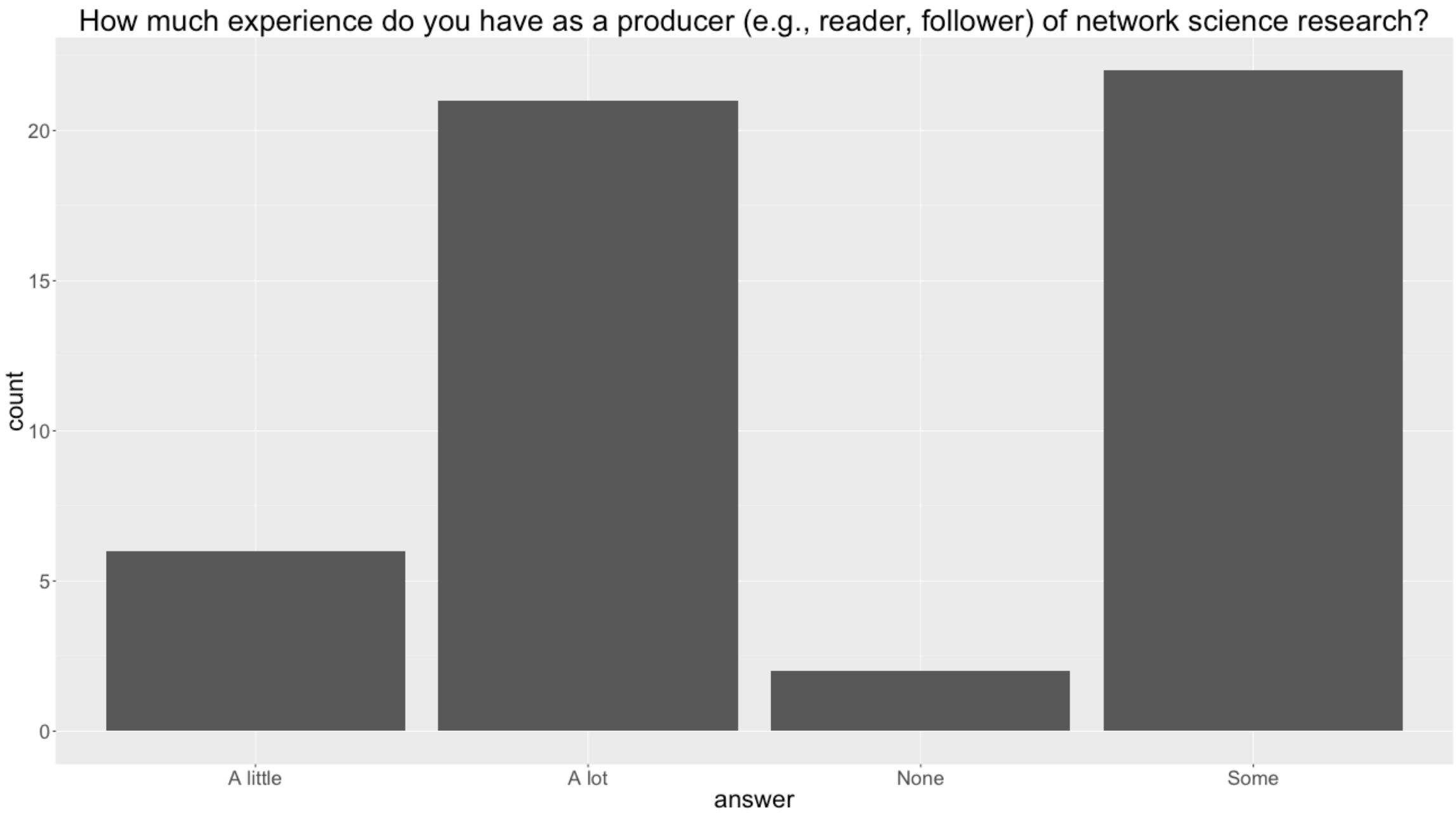
Exercises

Data files

- Game of Thrones character ratings
- Time to Statham Punch
- Gapminder

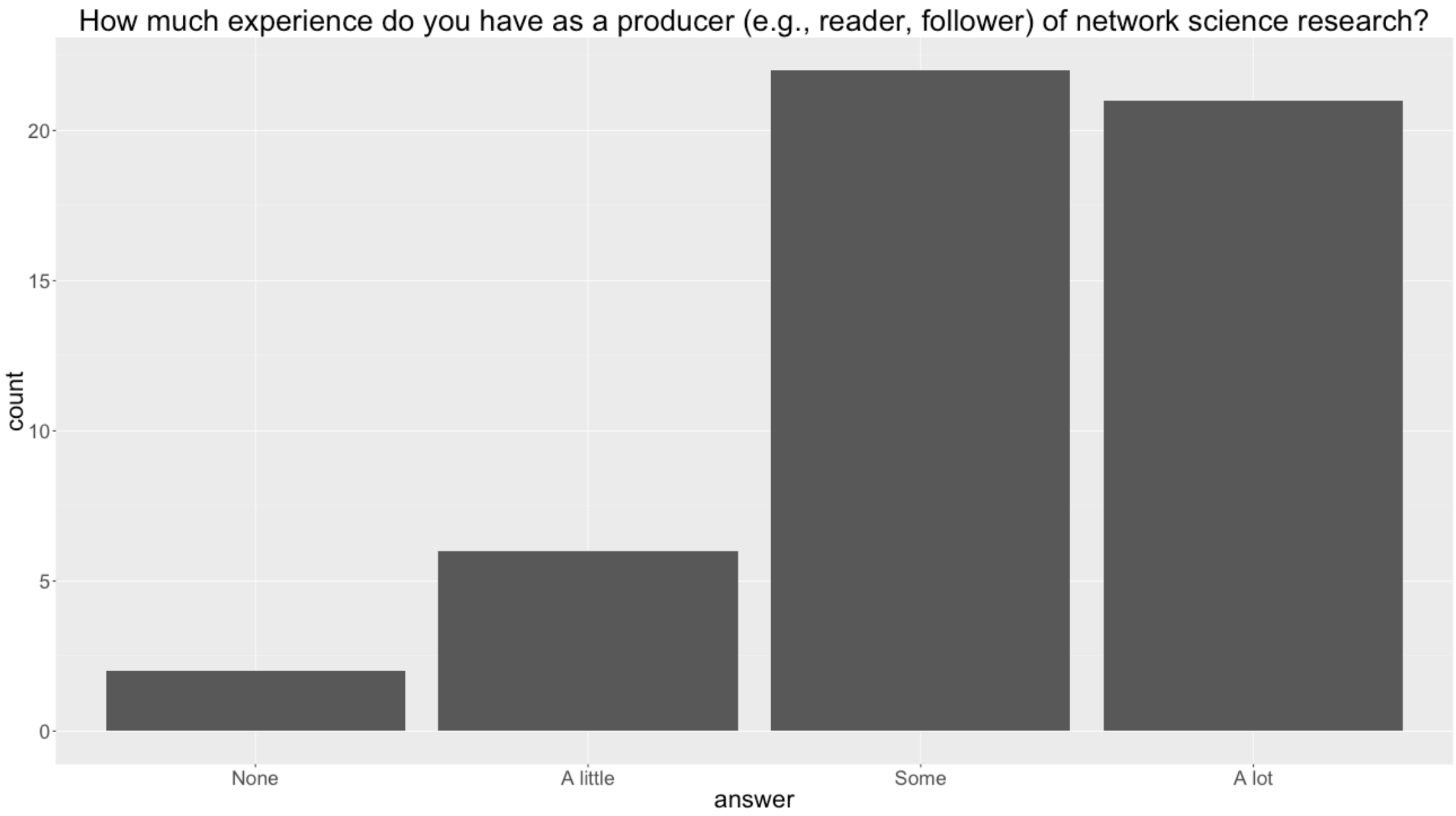
Principles for Effective Visualizations

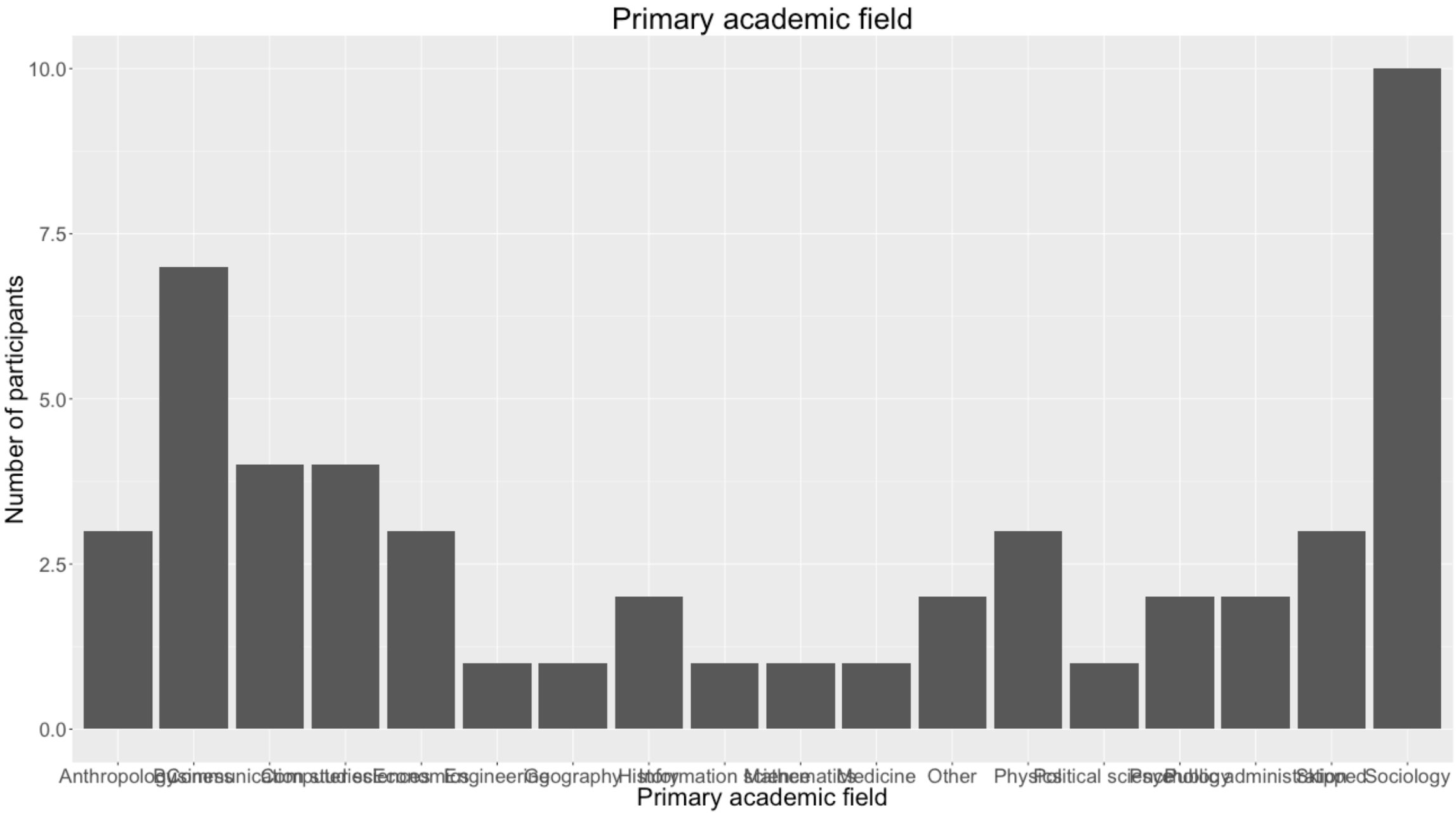
Principle 1: Order matters


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

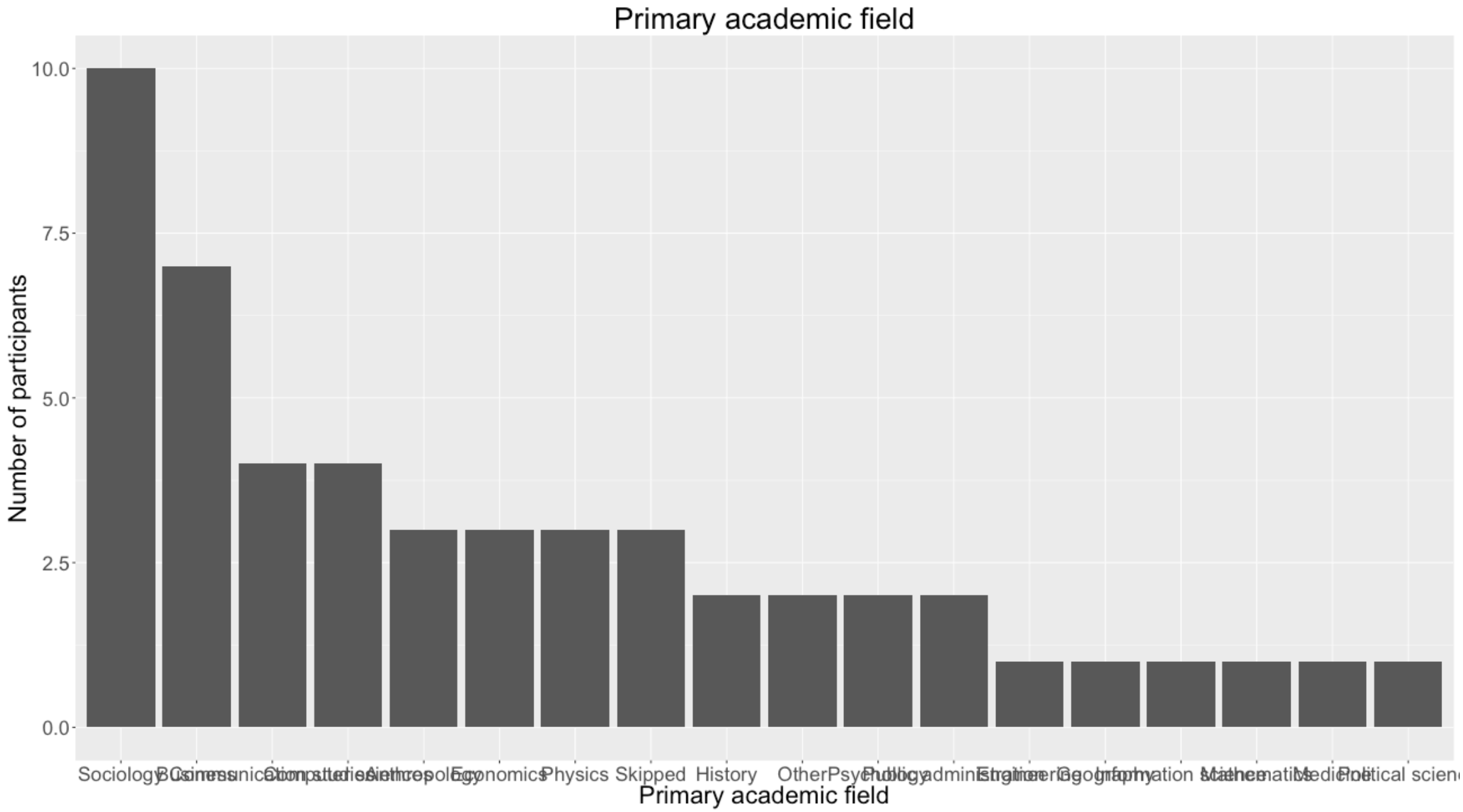
```
data$answer <-
    factor(data$answer,
           levels=c("None", "A little", "Some", "A lot"),
           ordered = TRUE)
```


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

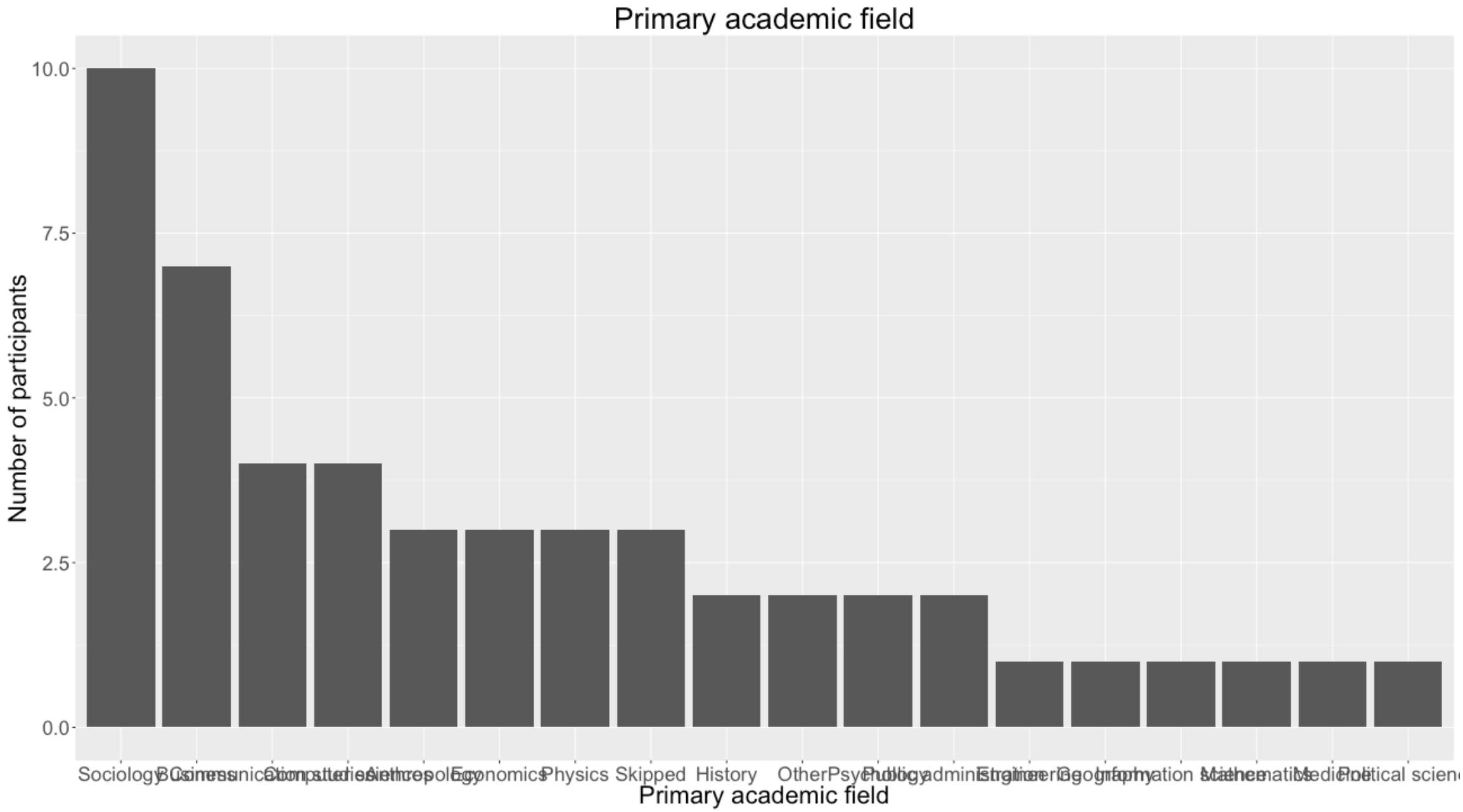

Primary academic field
Number of participants
10.0
7.5
5.0
2.5
0.0
Anthropology
Business
Communication
Computer sciences
Economics
Engineering
Geography
History
Information sciences
Mathematics
Medicine
Other
Physics
Political science
Psychology
Public administration
Skipped
Sociology
Primary academic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

Order by value

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

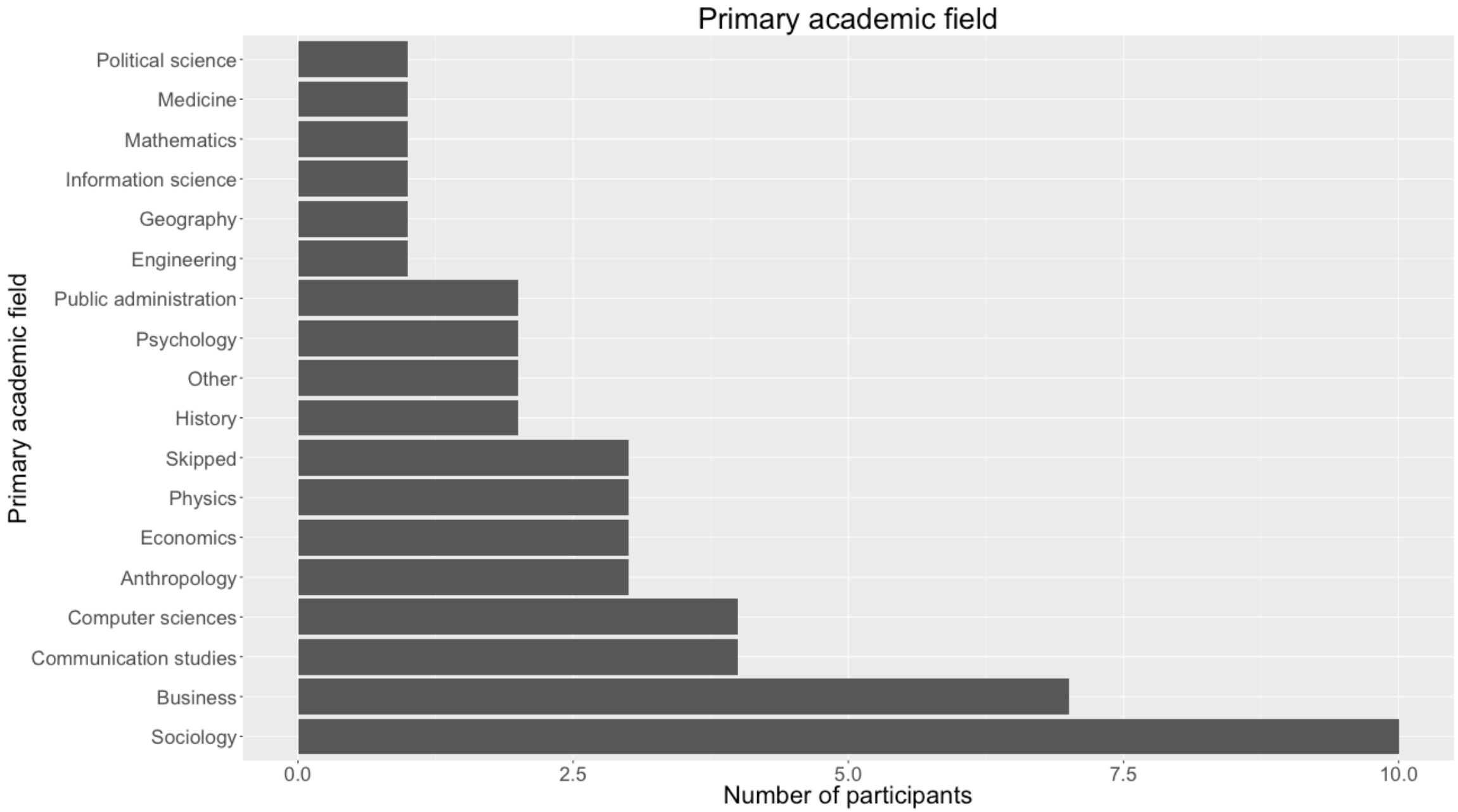

Primary academic field
Number of participants
10.0
7.5
5.0
2.5
0.0
Sociology
Business
Communication
Computer science
Anthropology
Economics
Physics
Skipped
History
Other
Psychology
Public administration
Engineering
Geography
Information science
Mathematics
Medicine
Political scien
Primary academic field

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

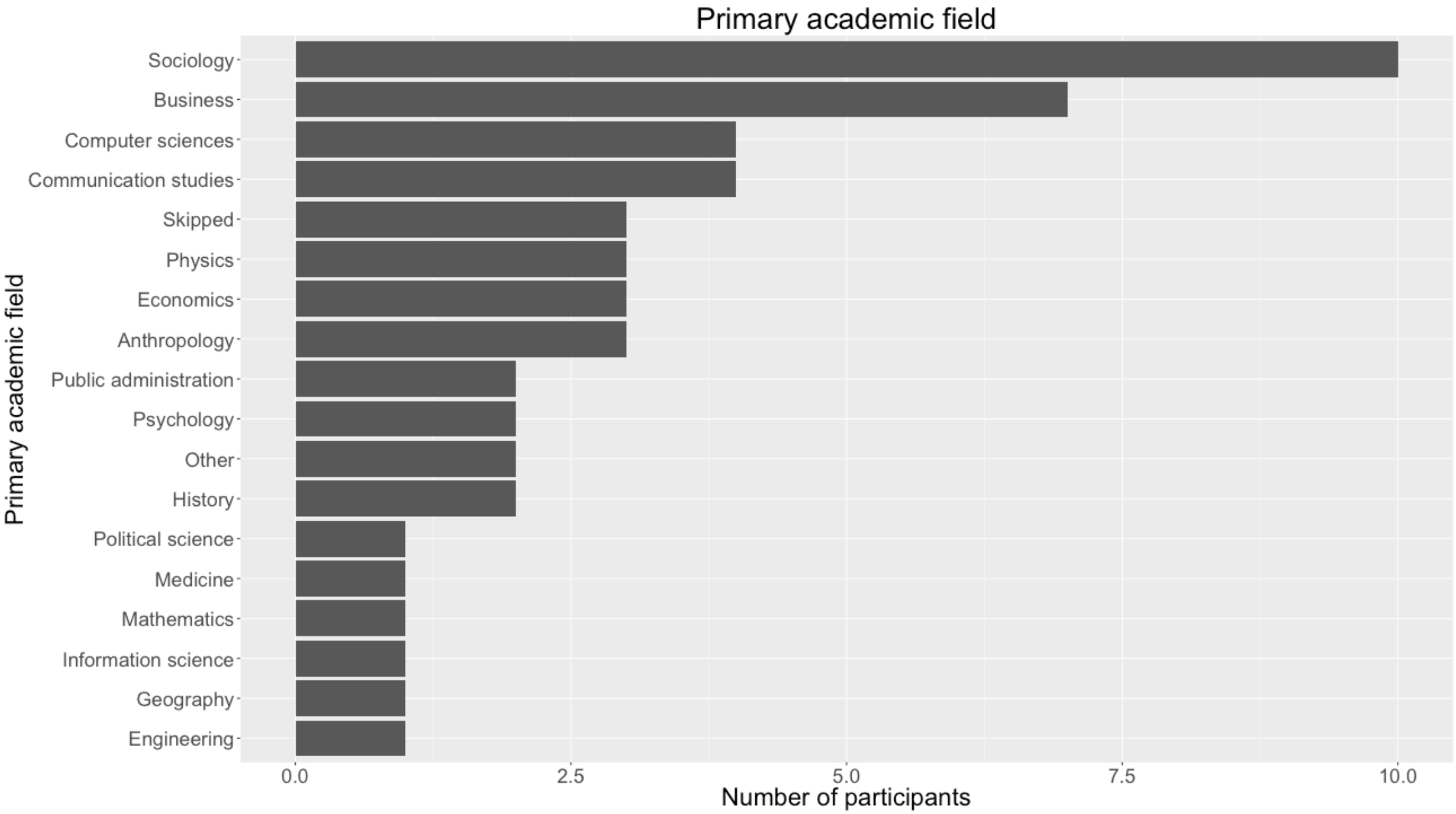
```
coord_flip()
```


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

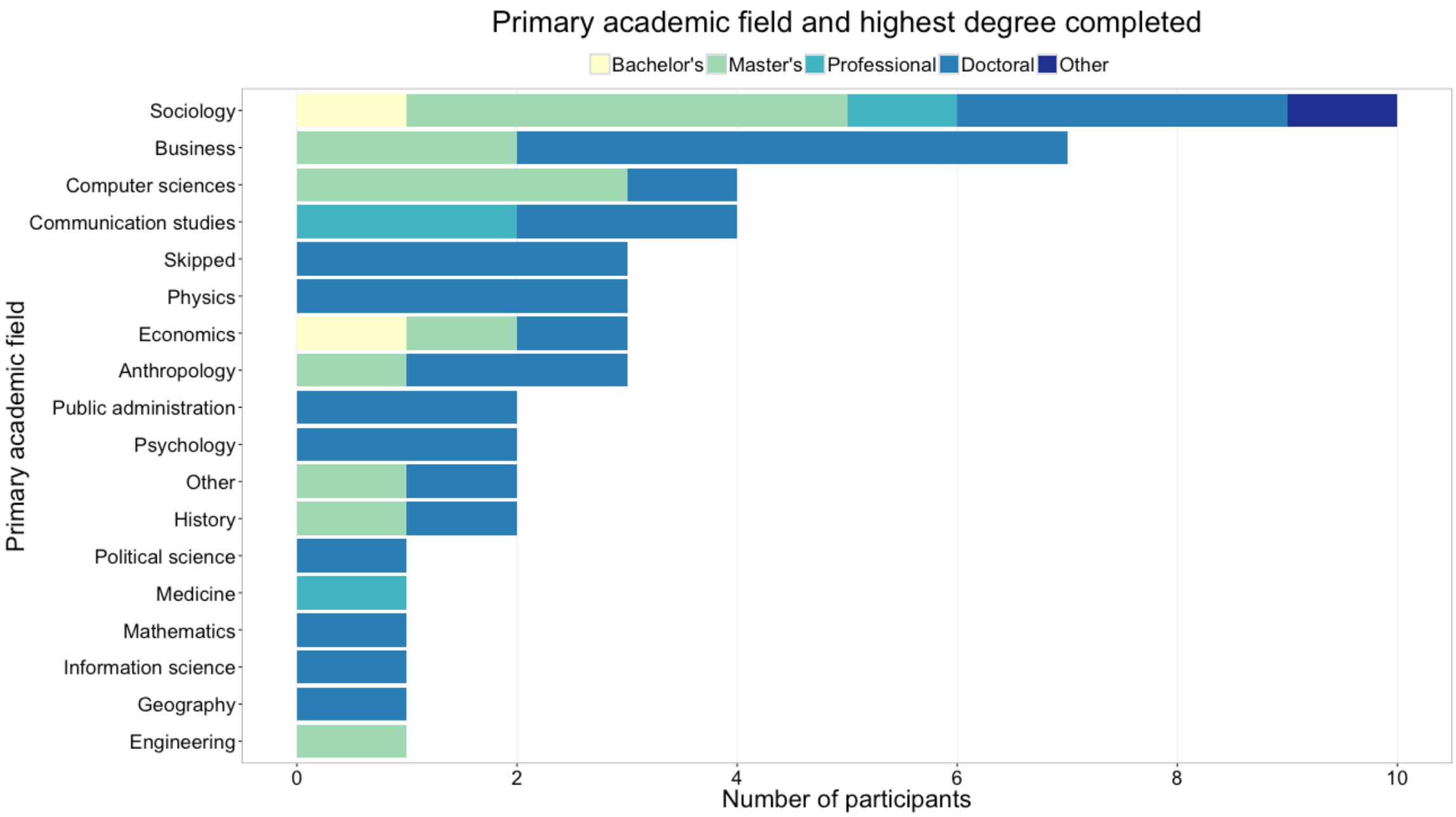
Oops!

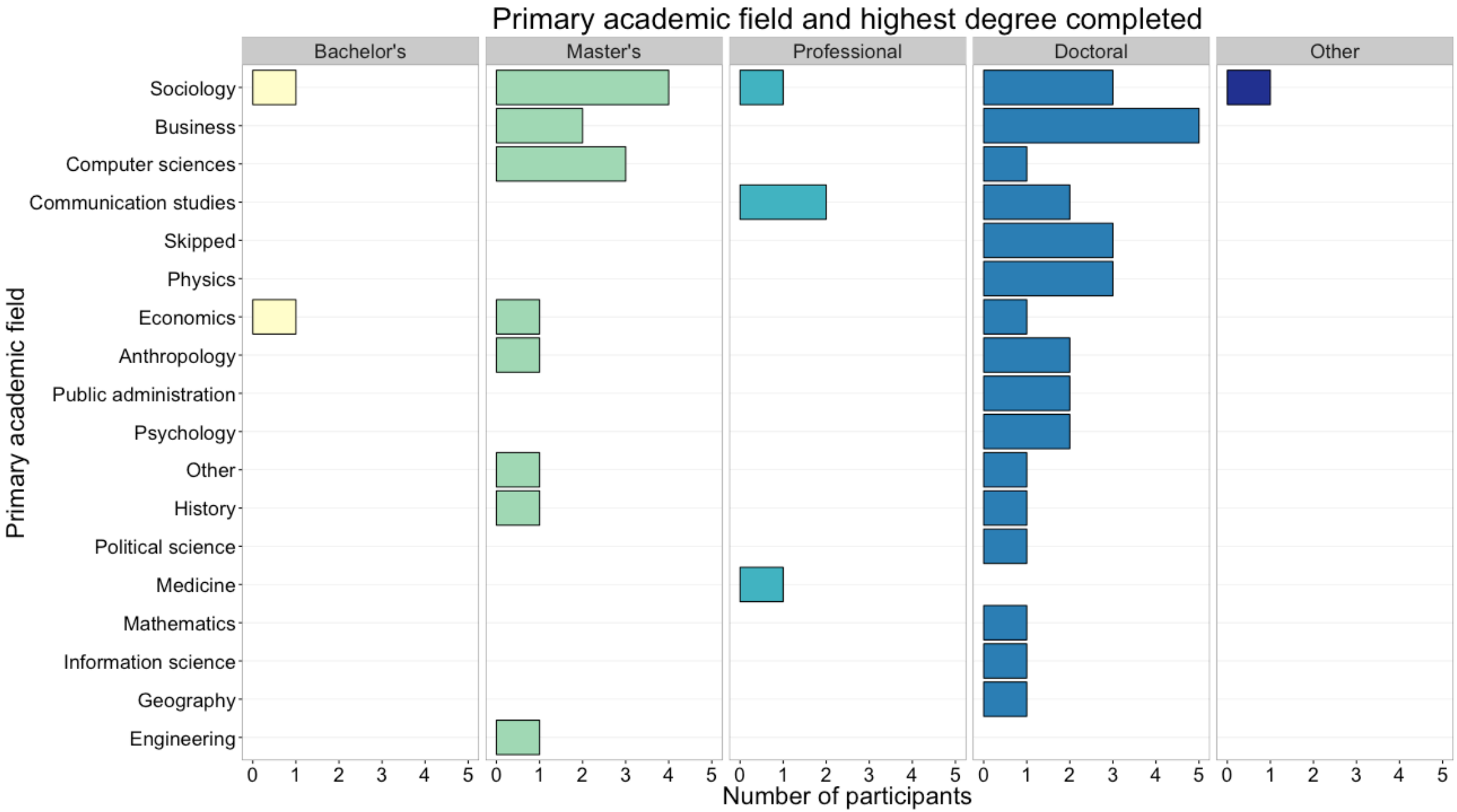
```
data$academic_field <-
     factor(data$academic_field,
           levels=names(
               sort(
                   table(data$academic_field),
                   decreasing=TRUE)))
```

```
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
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
2
4
6
8
10
Number of participants

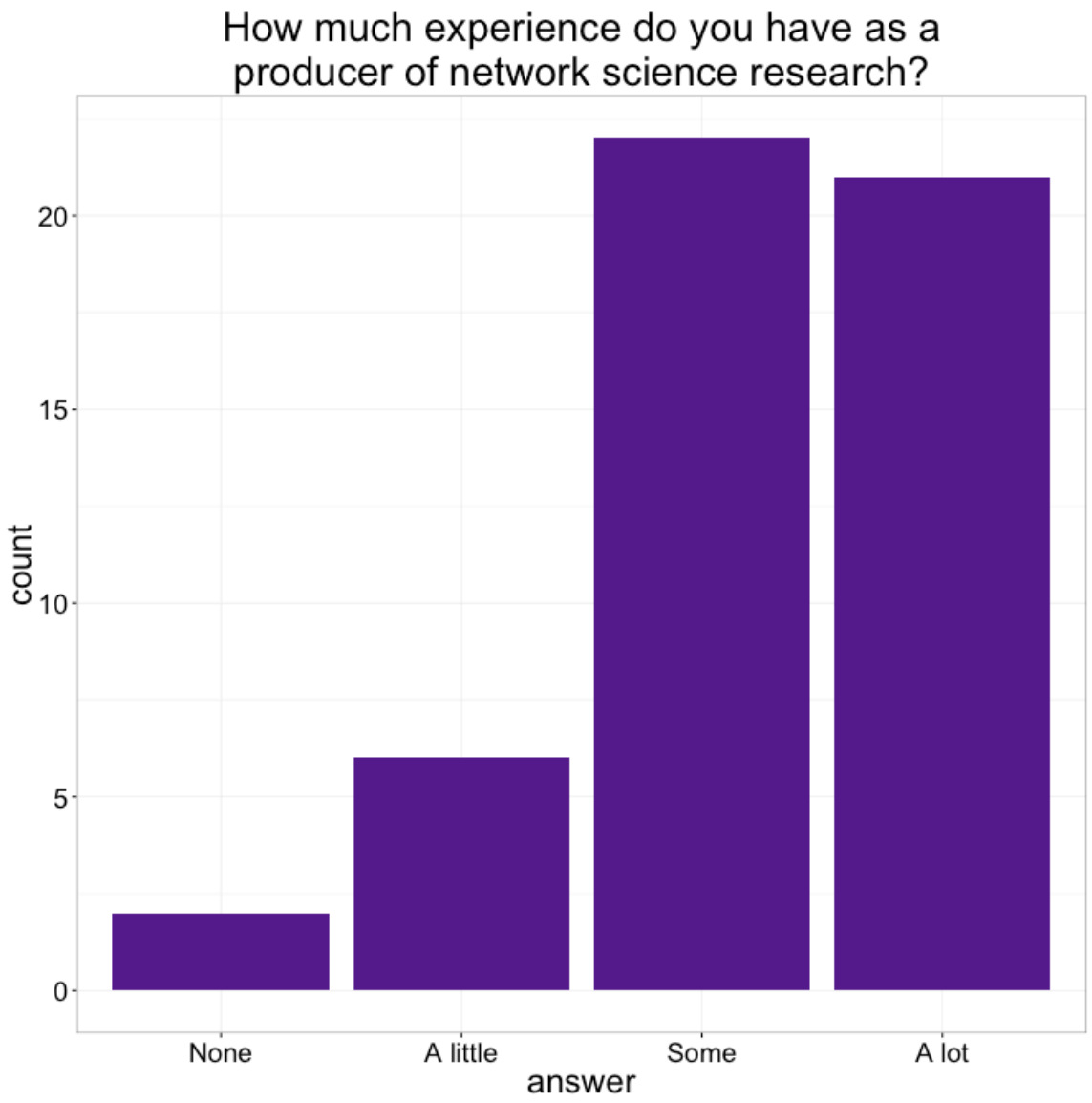

Primary academic field and highest degree completed
Bachelor's
Master's
Professional
Doctoral
Other
Primary academic field
Sociology
Business
Computer sciences
Communication studies
Skipped
Physics
Economics
Anthropology
Public administration
Psychology
Other
History
Political science
Medicine
Mathematics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

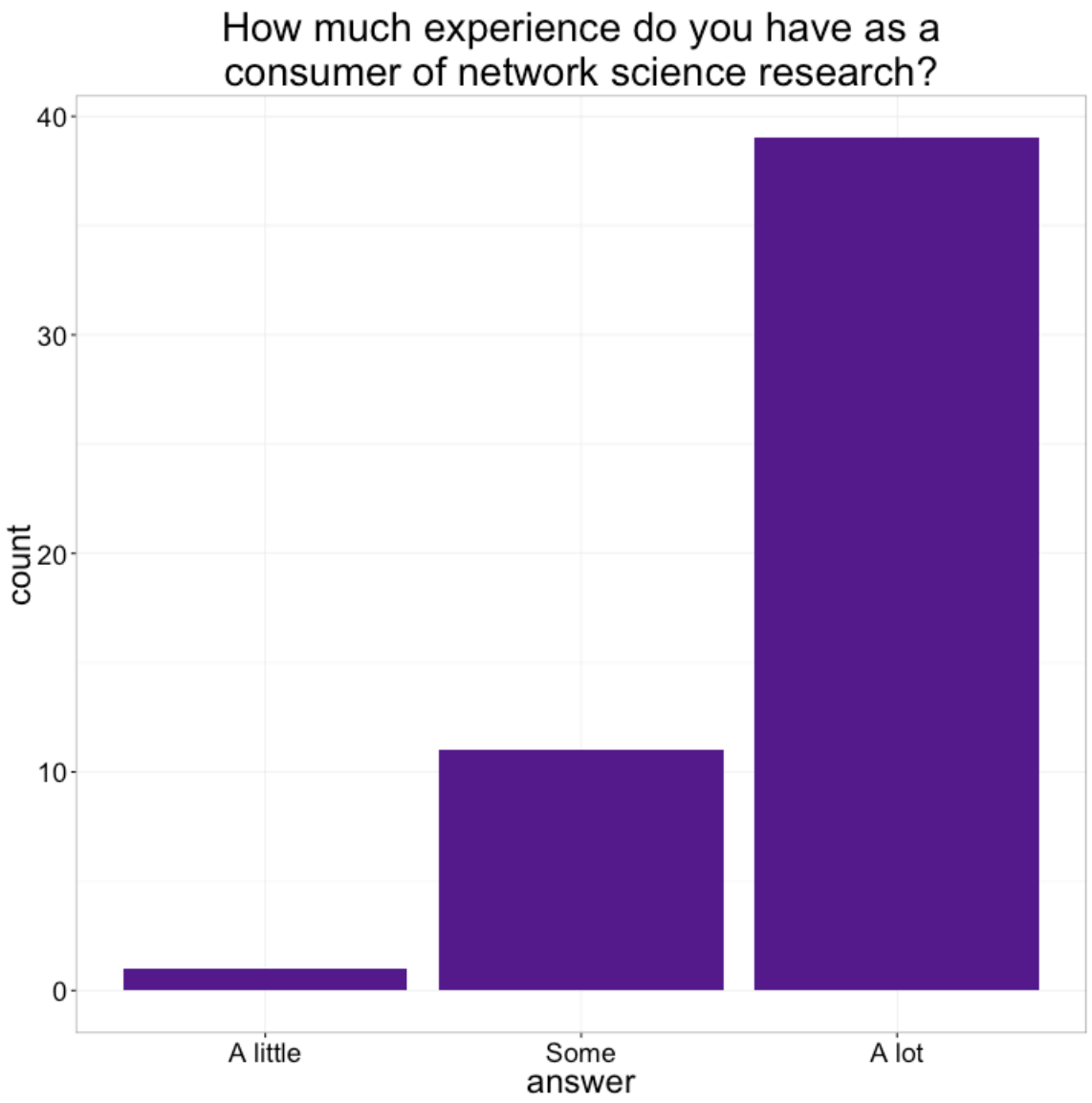
Different placement helps with different comparisons

```
fill=highest_degree
```

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

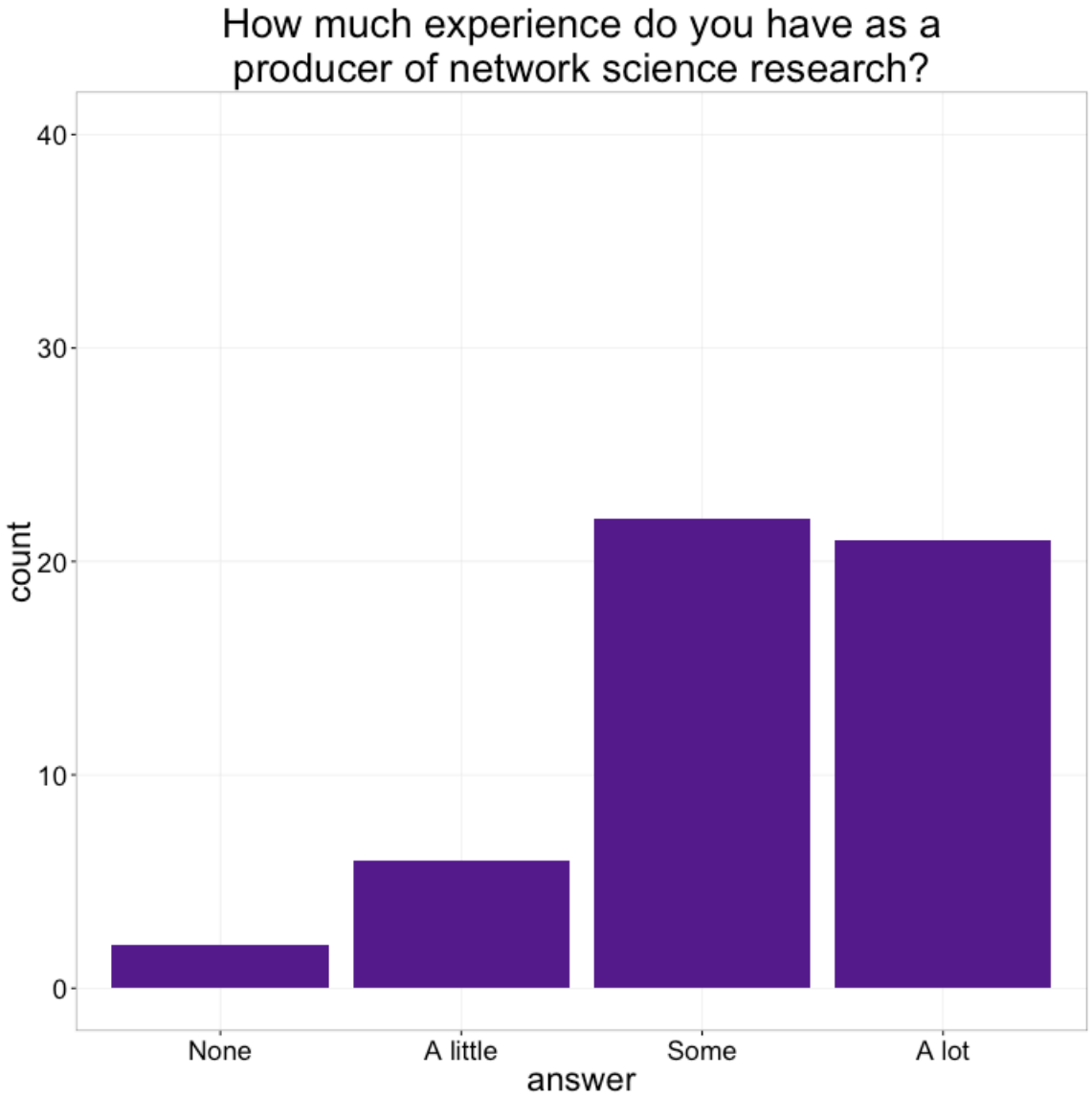
Principle 4:
Keep scales consistent

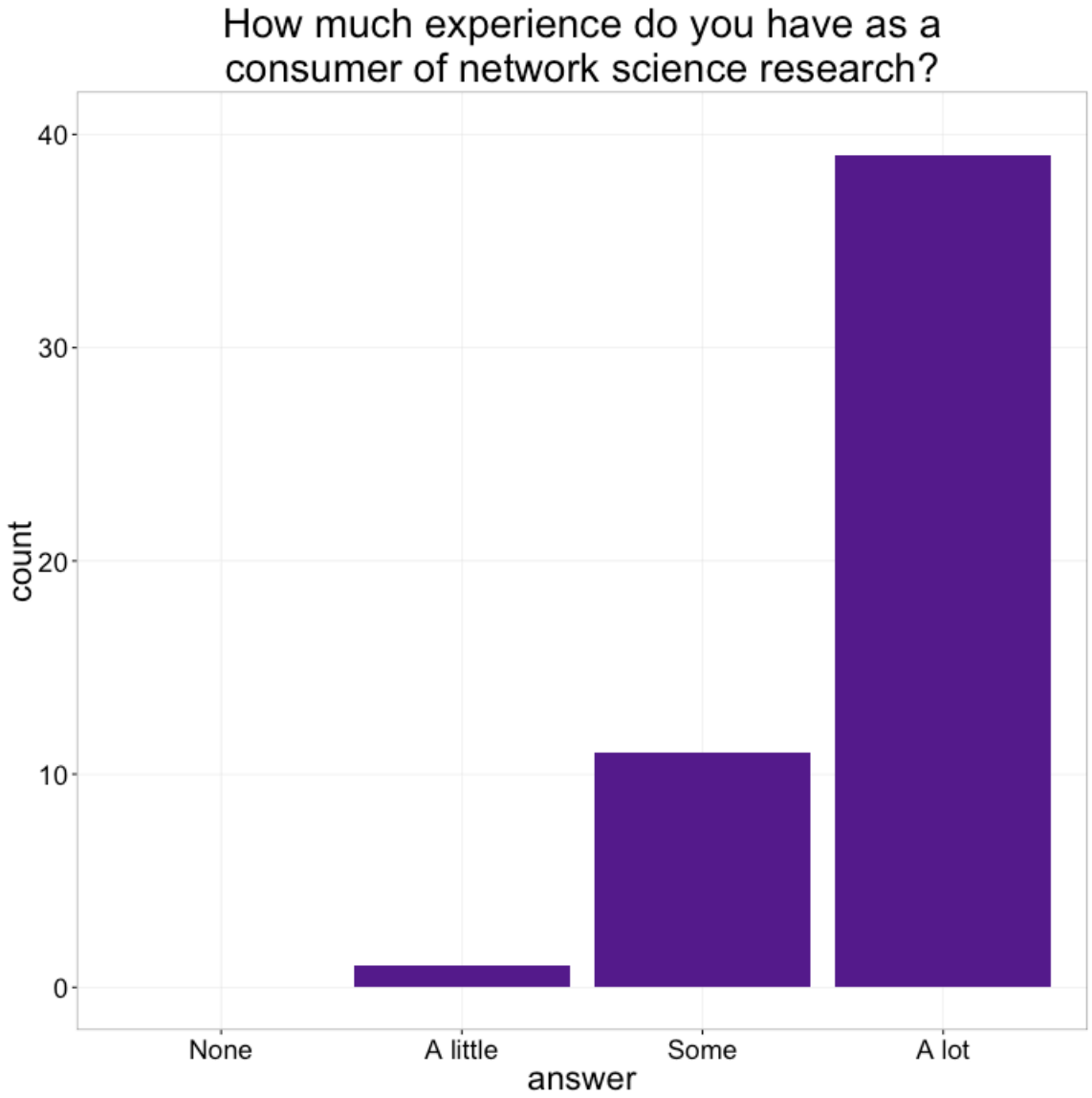

How much experience do you have as a producer of network science research?
count
0
5
10
15
20
None
A little
Some
A lot
answer


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

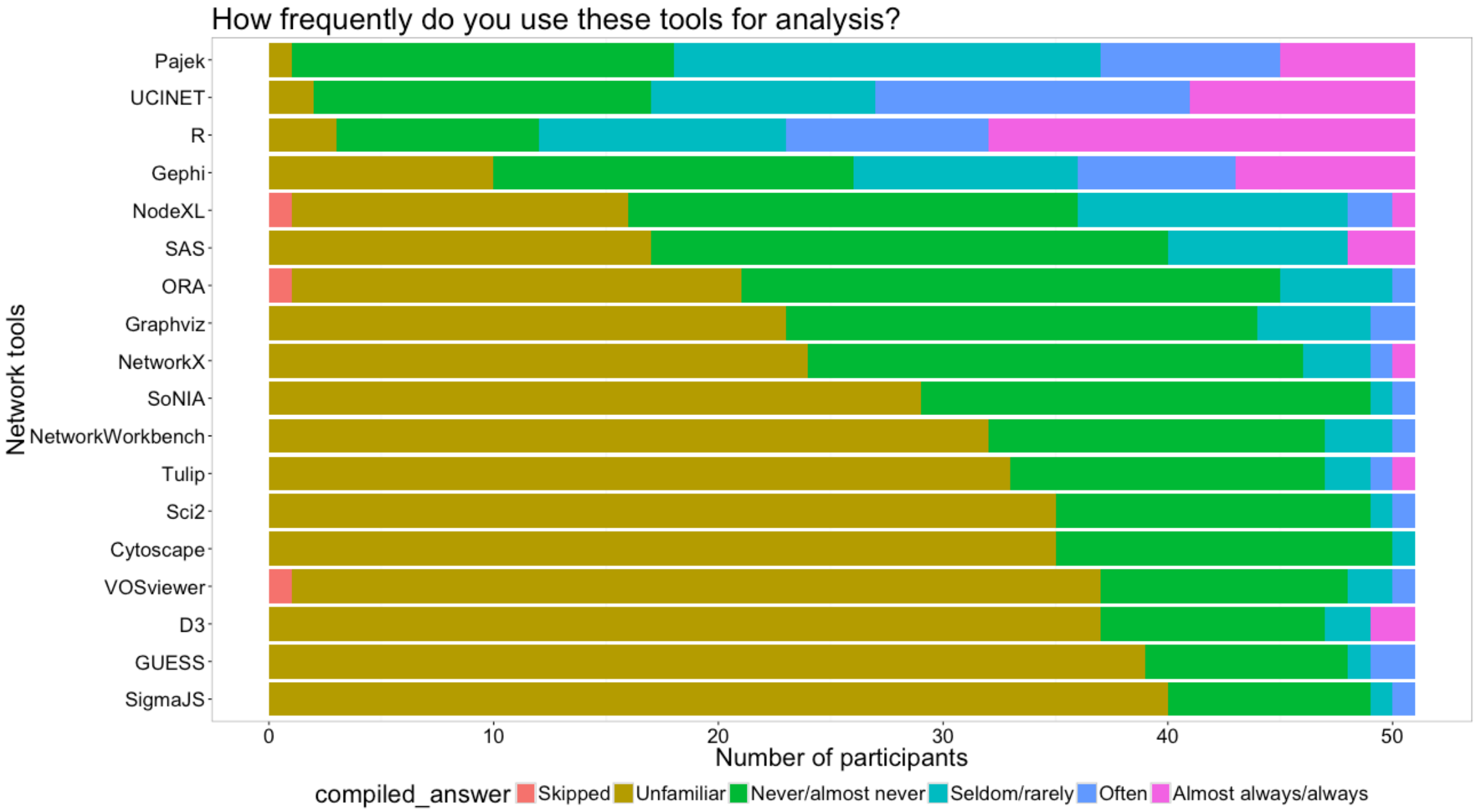
Keep all categories, manually set axes

```
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

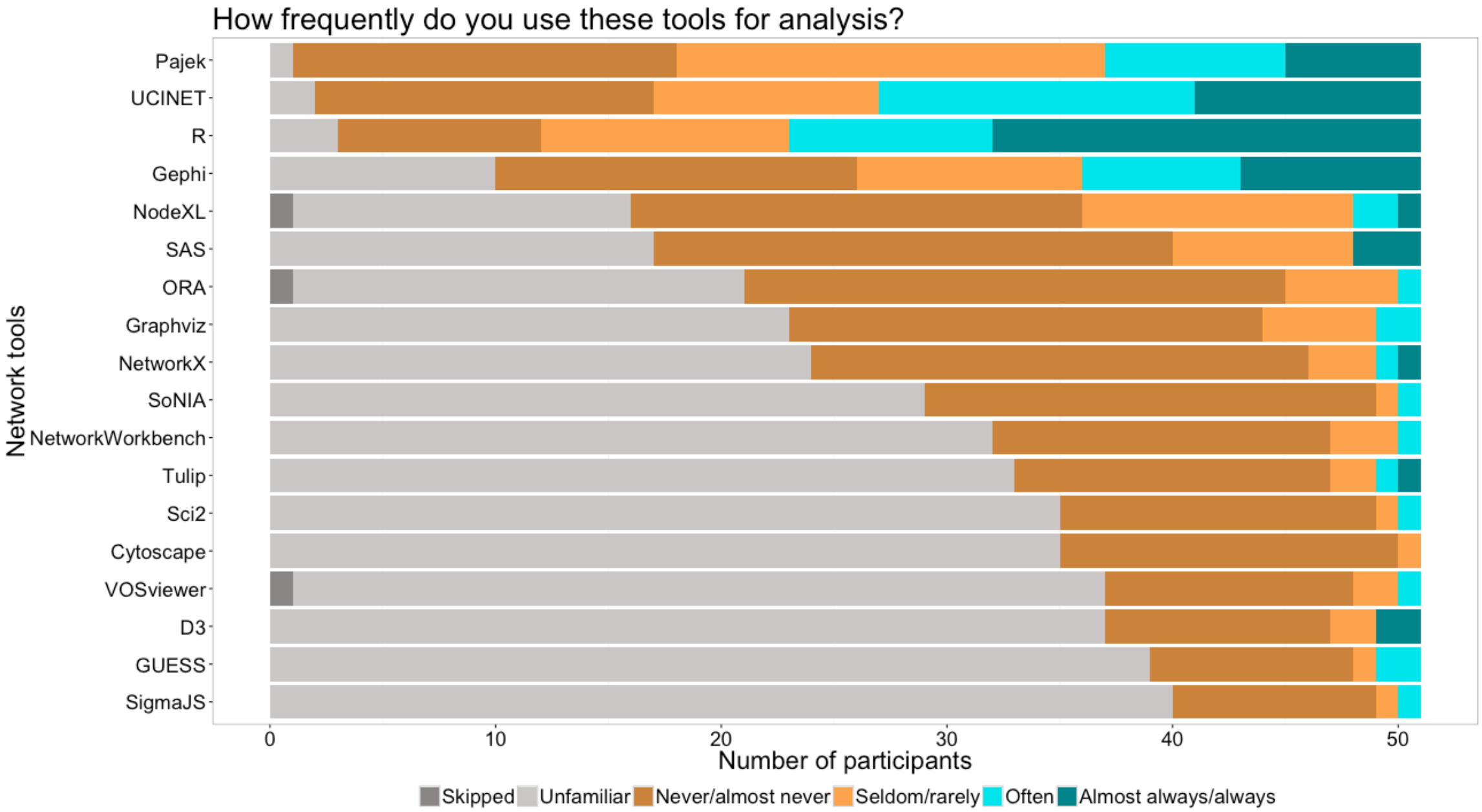

How much experience do you have as a consumer of network science research?
count
40
30
20
10
0
None
A little
Some
A lot
answer

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")
```

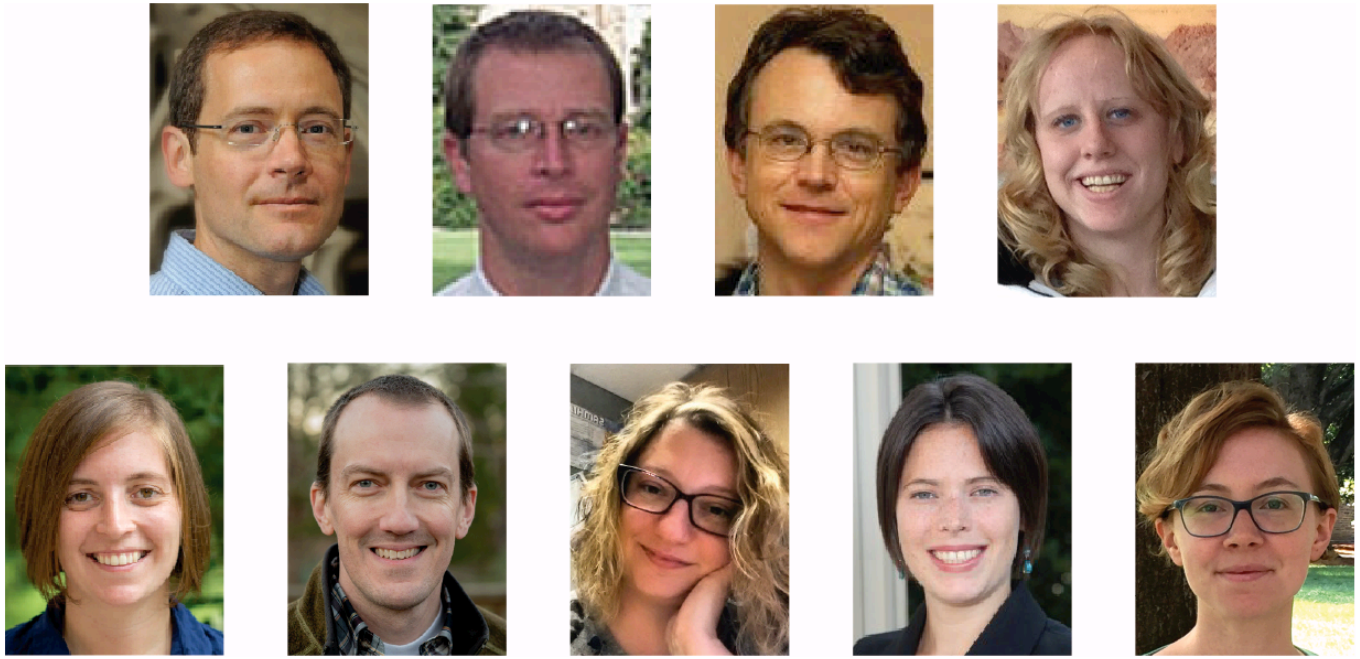

How frequently do you use these tools for analysis?
Network tools
Pajek
UCINET
R
Gephi
NodeXL
SAS
ORA
Graphviz
NetworkX
SoNIA
NetworkWorkbench
Tulip
Sci2
Cytoscape
VOSviewer
D3
GUESS
SigmaJS
0
10
20
30
40
50
Number of participants
Skipped
Unfamiliar
Never/almost never
Seldom/rarely
Often
Almost always/always

Getting help

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

Data and Visualization Services



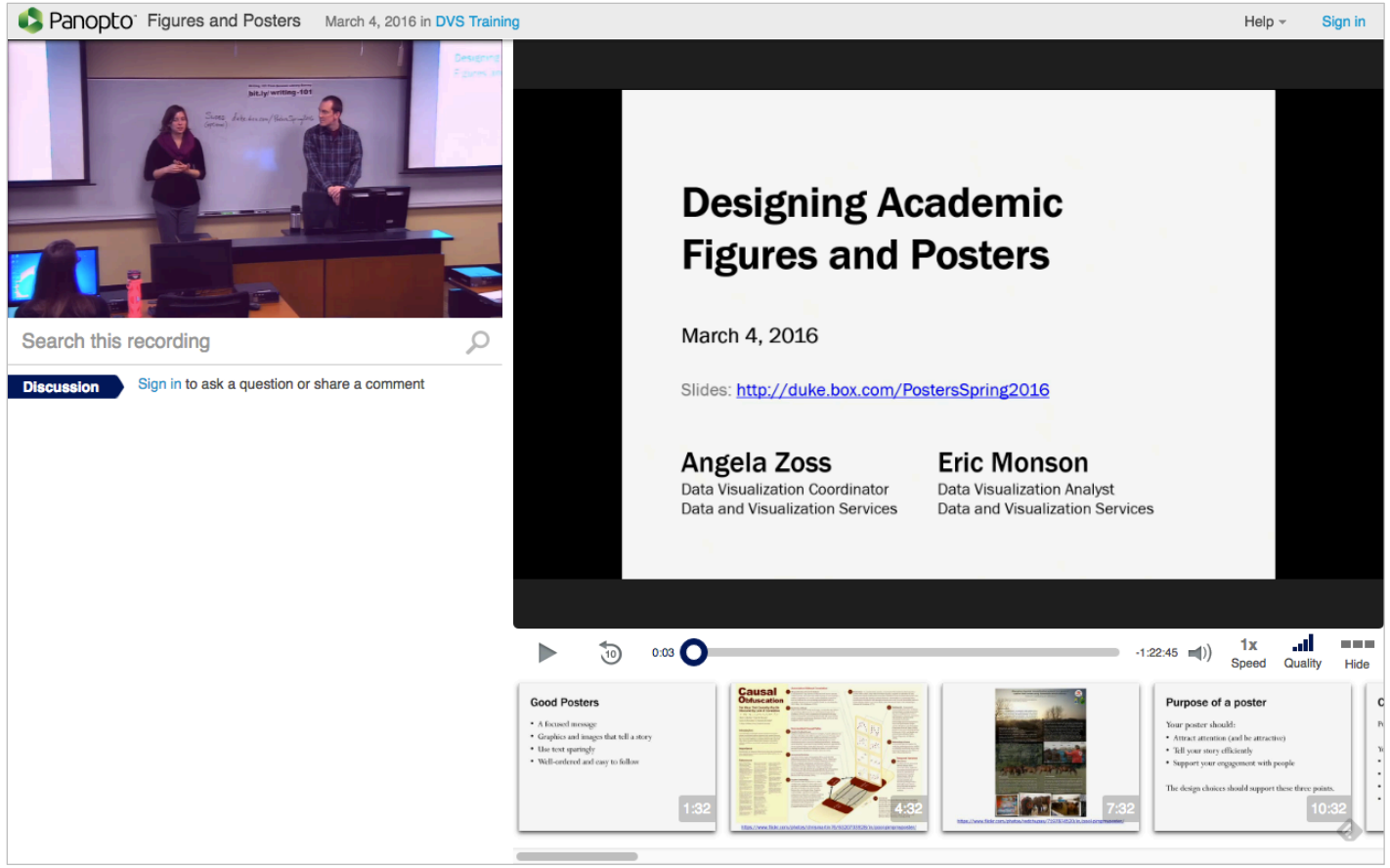
Data and Visualization Services Department

http://library.duke.edu/data
askdata@duke.edu

Information about DVS

- Data collections, LibGuides, etc.
  http://library.duke.edu/data/
- Blog (tutorials, announcements, etc.)
  http://blogs.library.duke.edu/data/
- E-mail consultations
  askdata@duke.edu
- Mailing list for announcements:
  https://lists.duke.edu/sympa/subscribe/dvs-announce
- Twitter accounts
  @duke_data, @duke_vis

Videos of past workshops



http://bit.ly/DVSvideos

Questions?

askdata@duke.edu