Tips for Effective Data Visualization

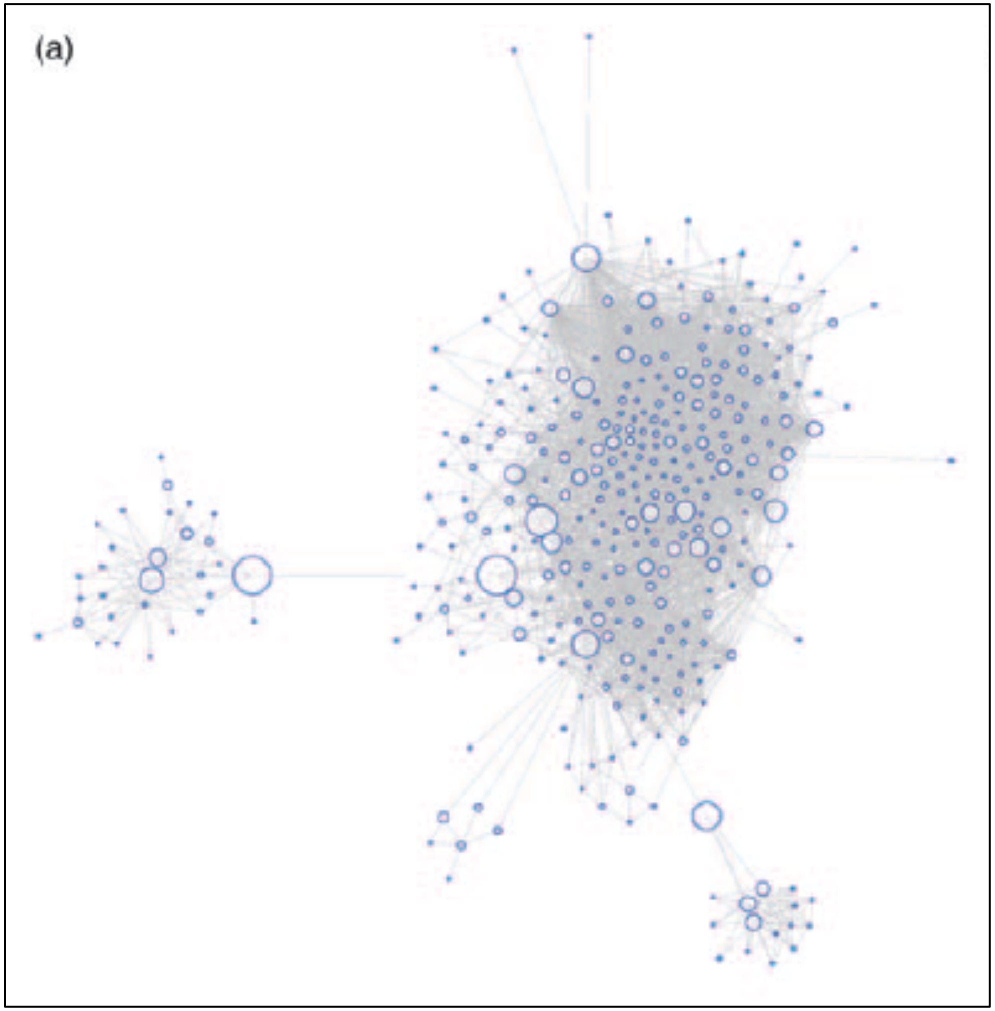
Angela Zoss · Eric Monson

Data and Visualization Services

STA 112FS · Fall 2017

Slides: http://bit.ly/STA112FSVisFall2017

Visual exploration can reveal data quality problems

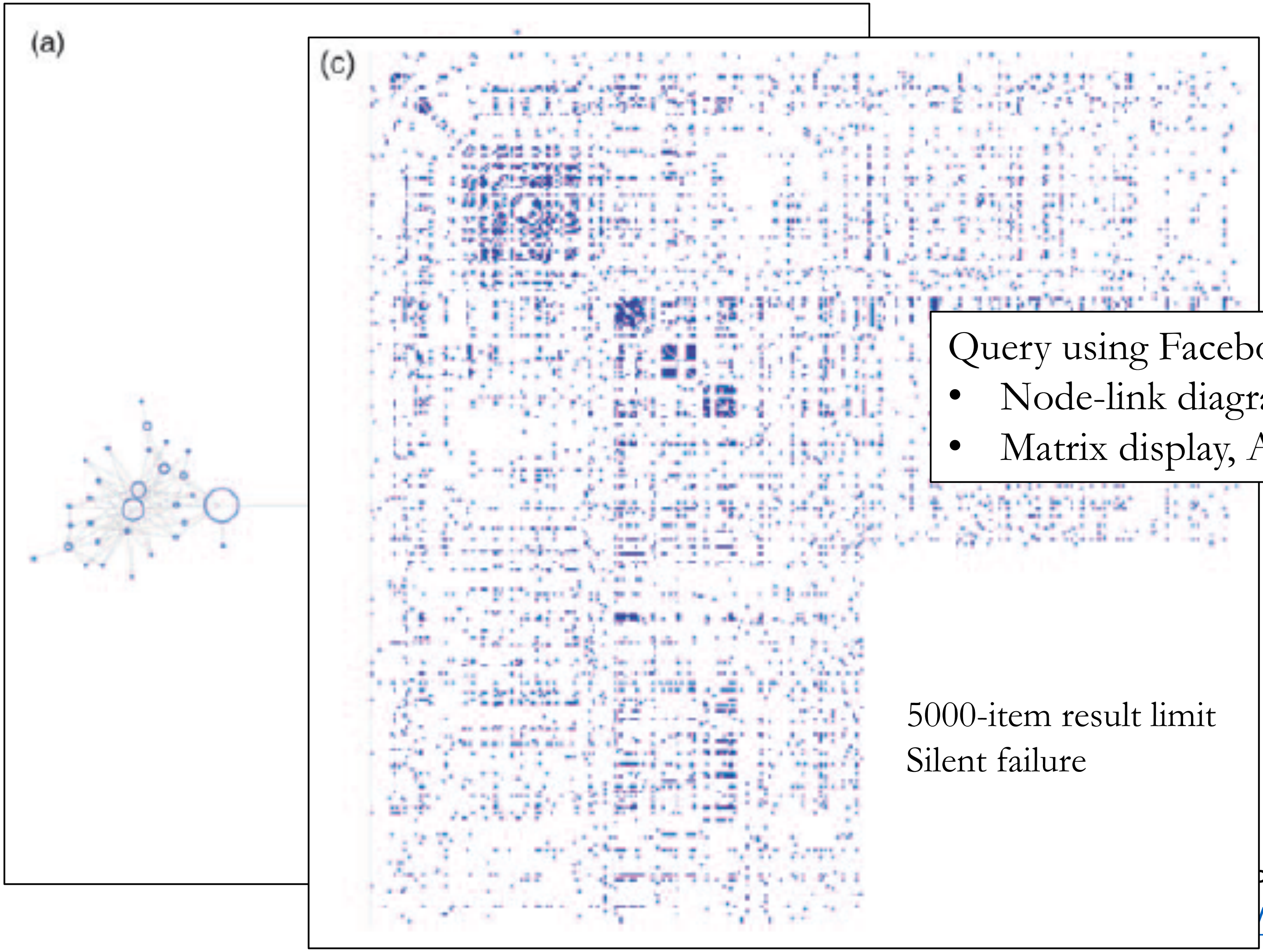


Query using Facebook API

- Node-link diagram

Kandel, Heer, Plaisant, et al. (2011)
http://dx.doi.org/10.1177/1473871611415994

Visual exploration can reveal data quality problems



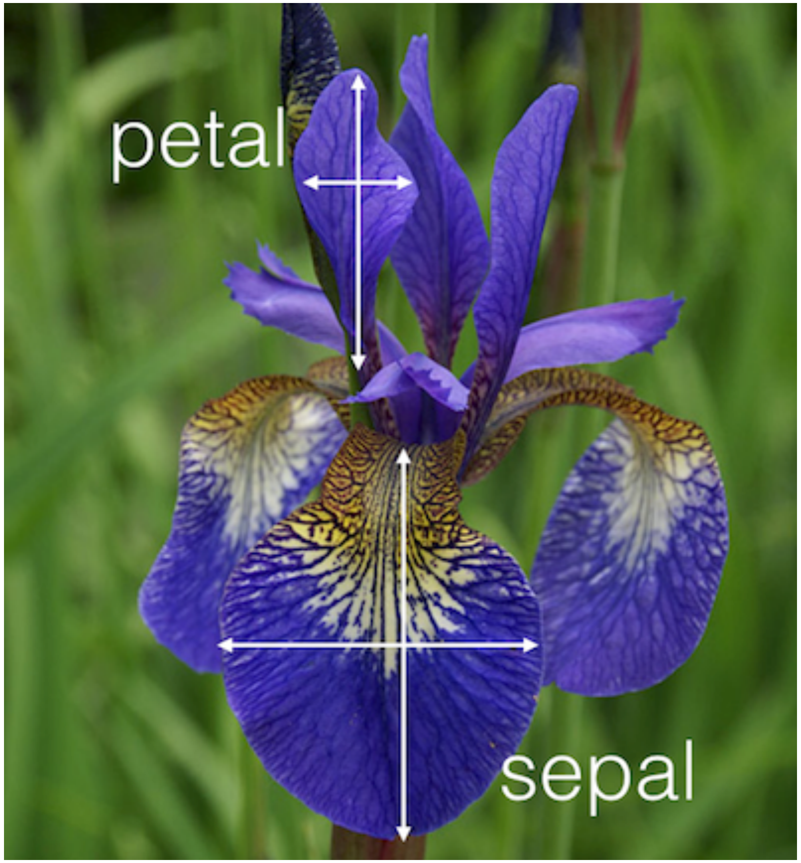
Query using Facebook API

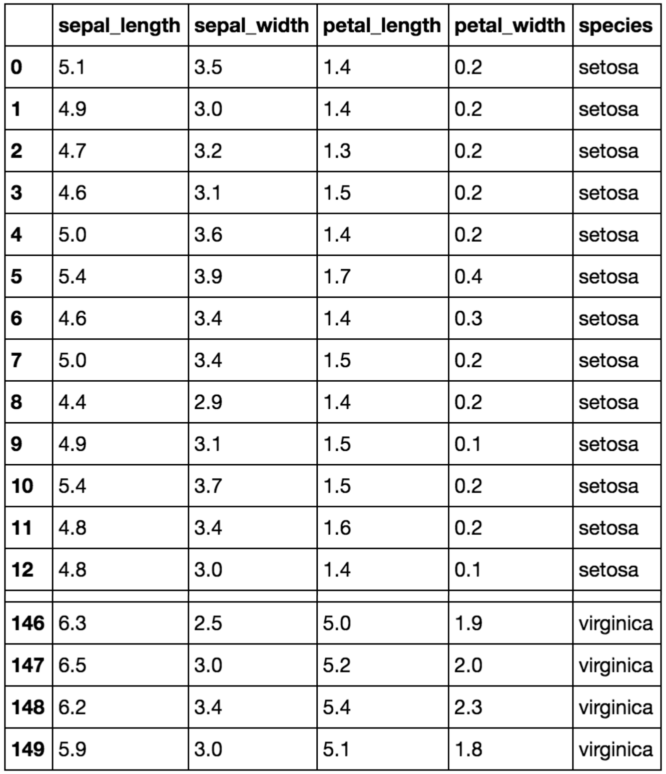
- Node-link diagram
- Matrix display, API return order

5000-item result limit
Silent failure

Plaisant, et al. (2011)
/1473871611415994

Fisher’s Iris data set (1936)

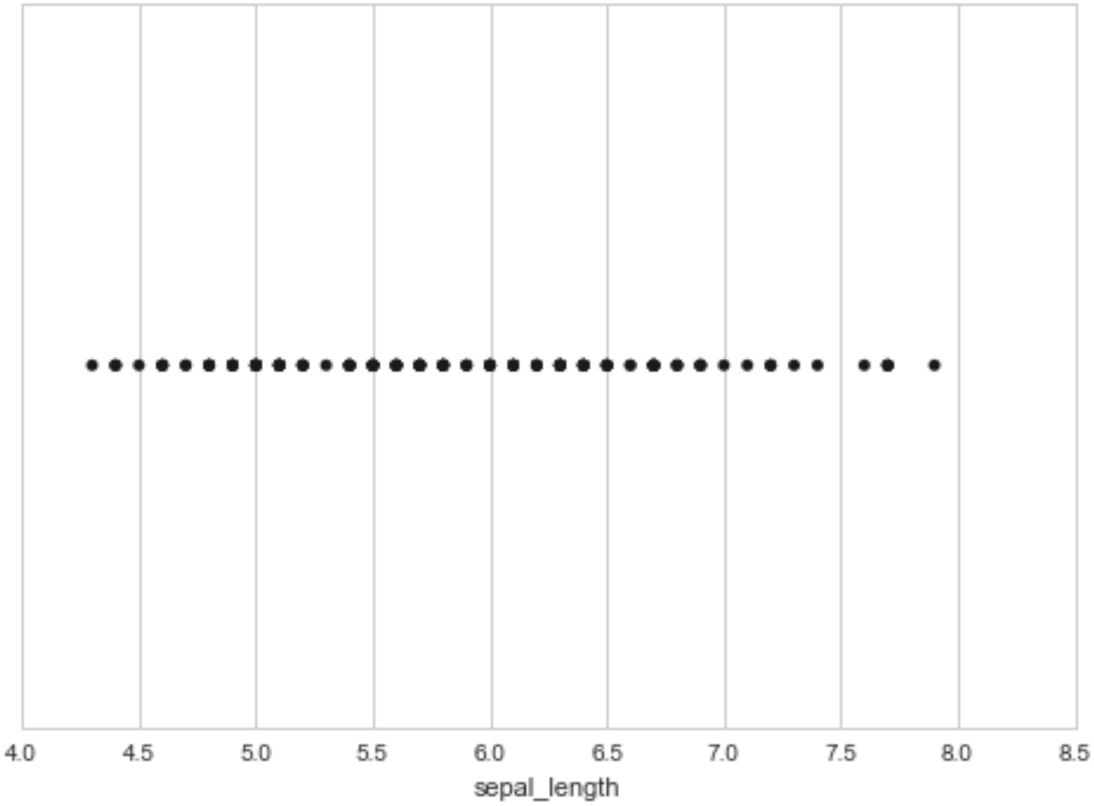




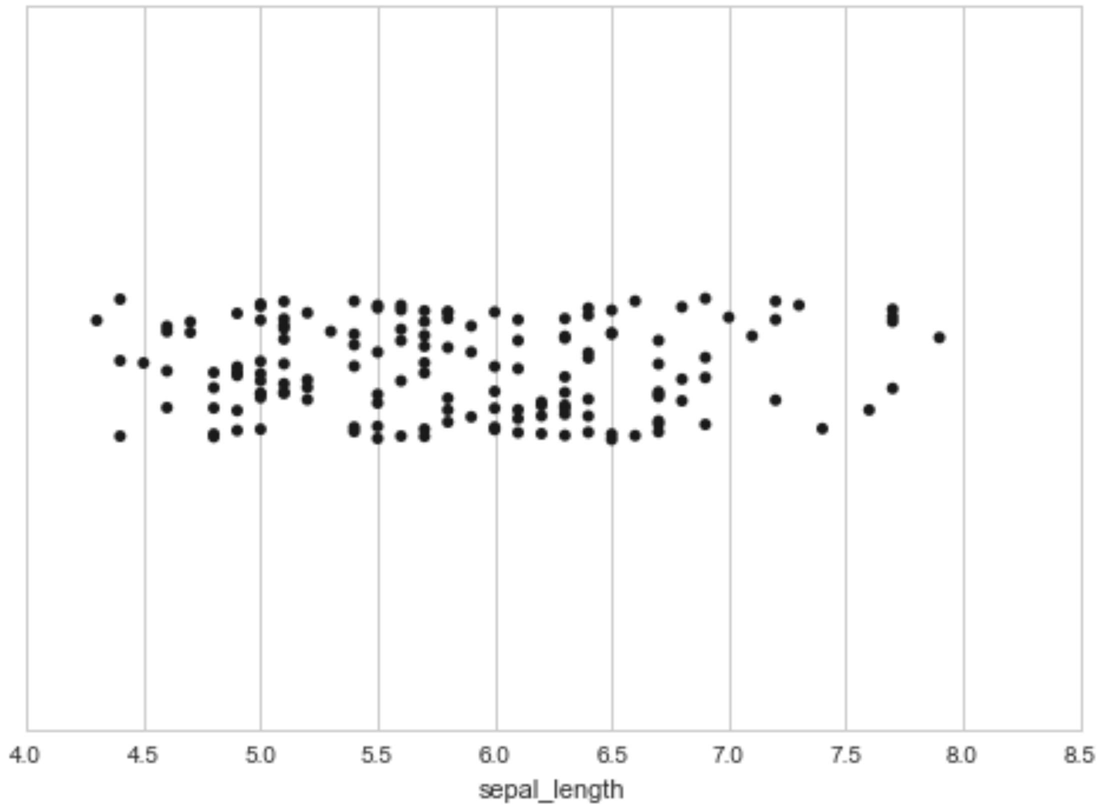
	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
5	5.4	3.9	1.7	0.4	setosa
6	4.6	3.4	1.4	0.3	setosa
7	5.0	3.4	1.5	0.2	setosa
8	4.4	2.9	1.4	0.2	setosa
9	4.9	3.1	1.5	0.1	setosa
10	5.4	3.7	1.5	0.2	setosa
11	4.8	3.4	1.6	0.2	setosa
12	4.8	3.0	1.4	0.1	setosa
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

http://sebastianraschka.com/images/blog/2014/linear-discriminant-analysis/iris_petal_sepal.png

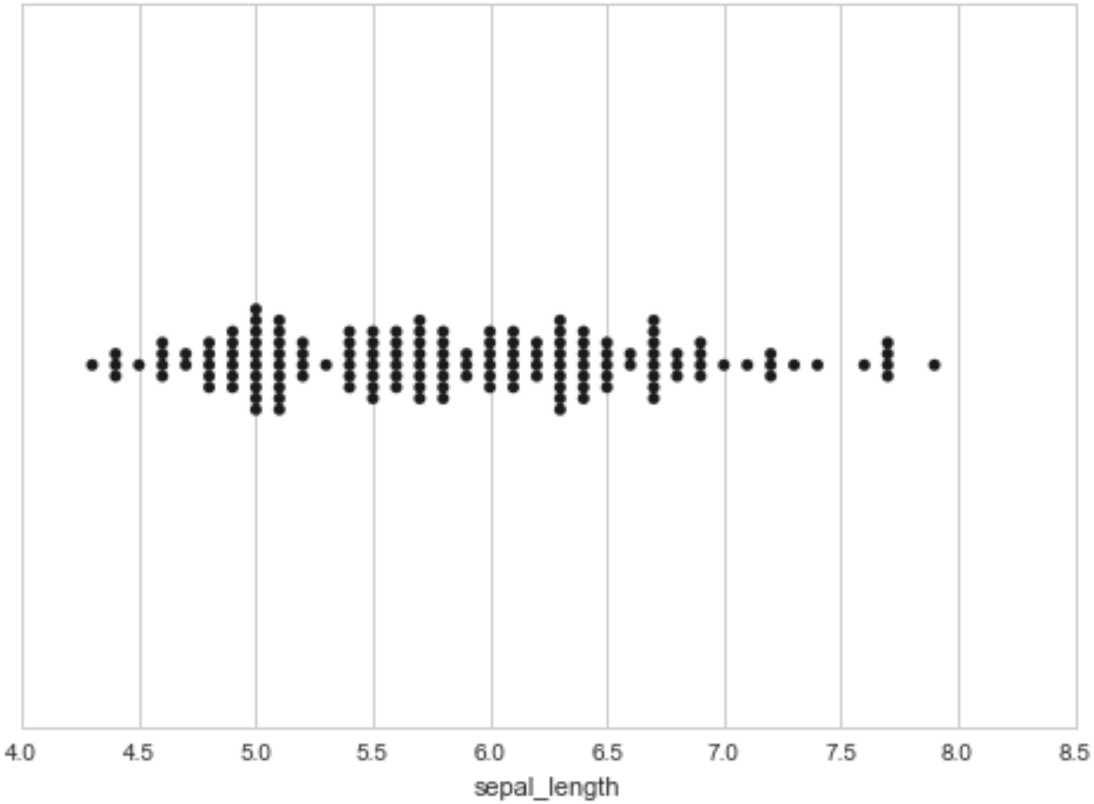
You can see a variable distribution by just plotting all the points



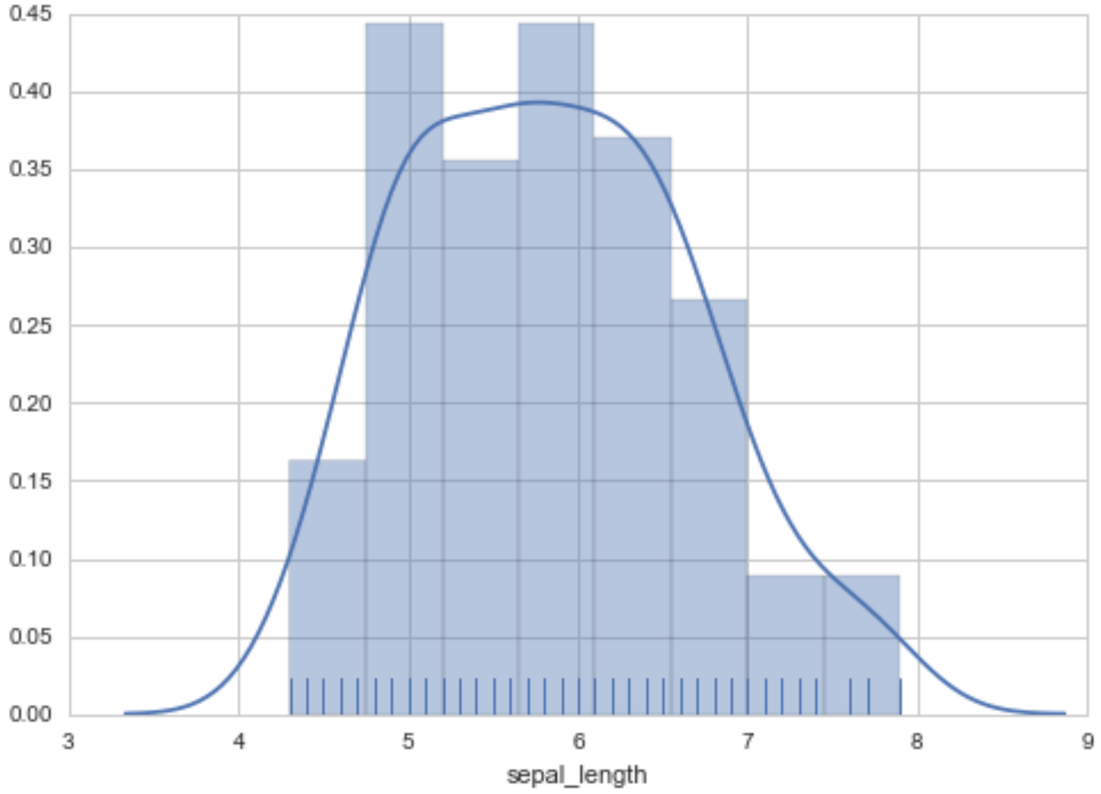
You can see a variable distribution by just plotting all the points (+jitter)



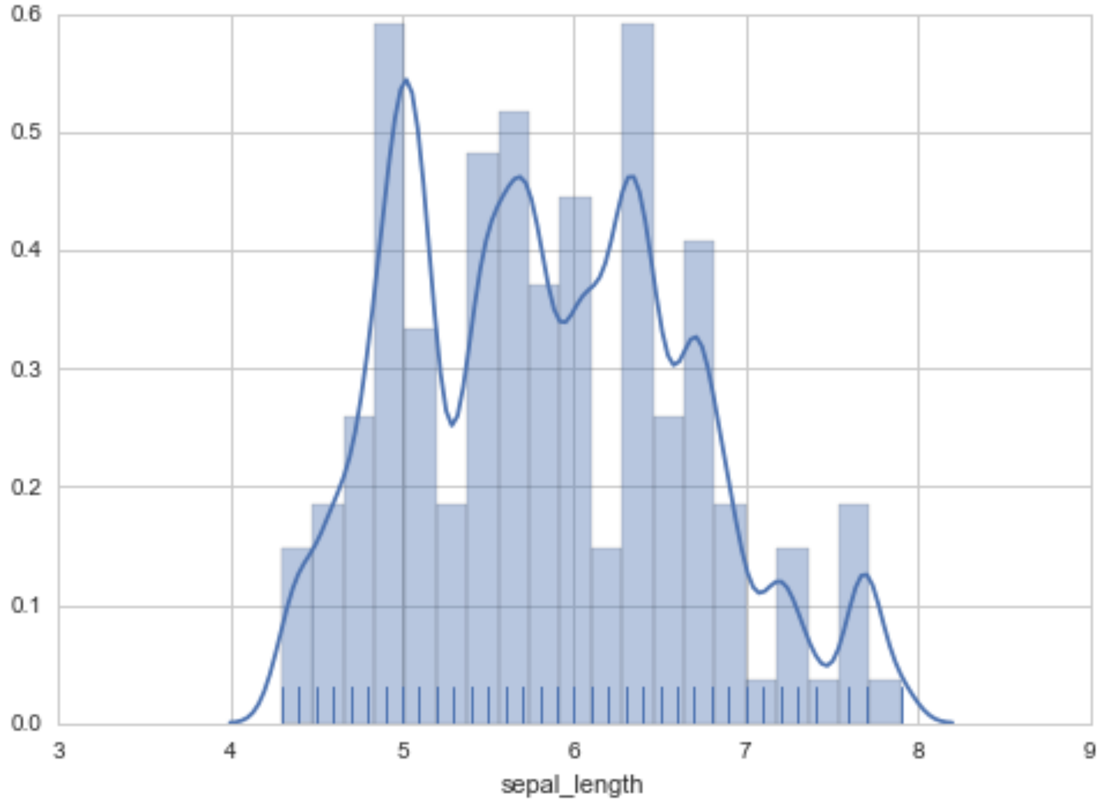
You can see a variable distribution by just plotting all the points (swarm)



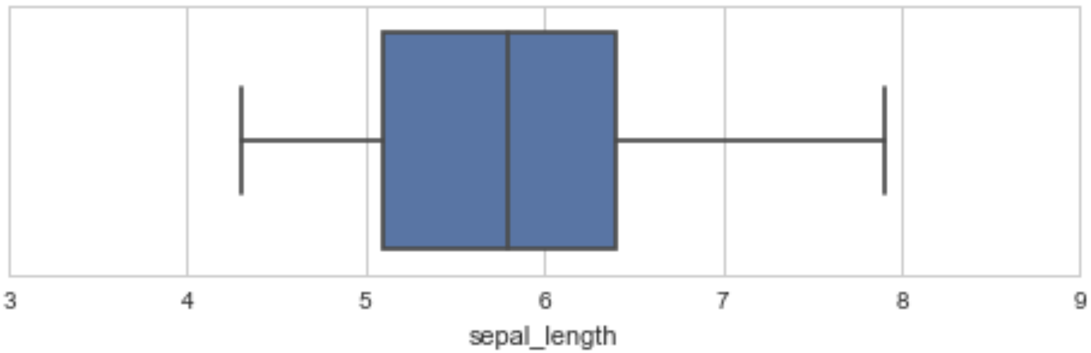
Histograms can show the distribution of one variable

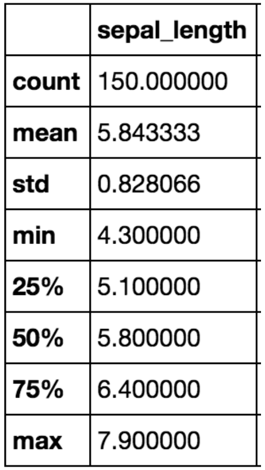


…but the results will depend on the bin width



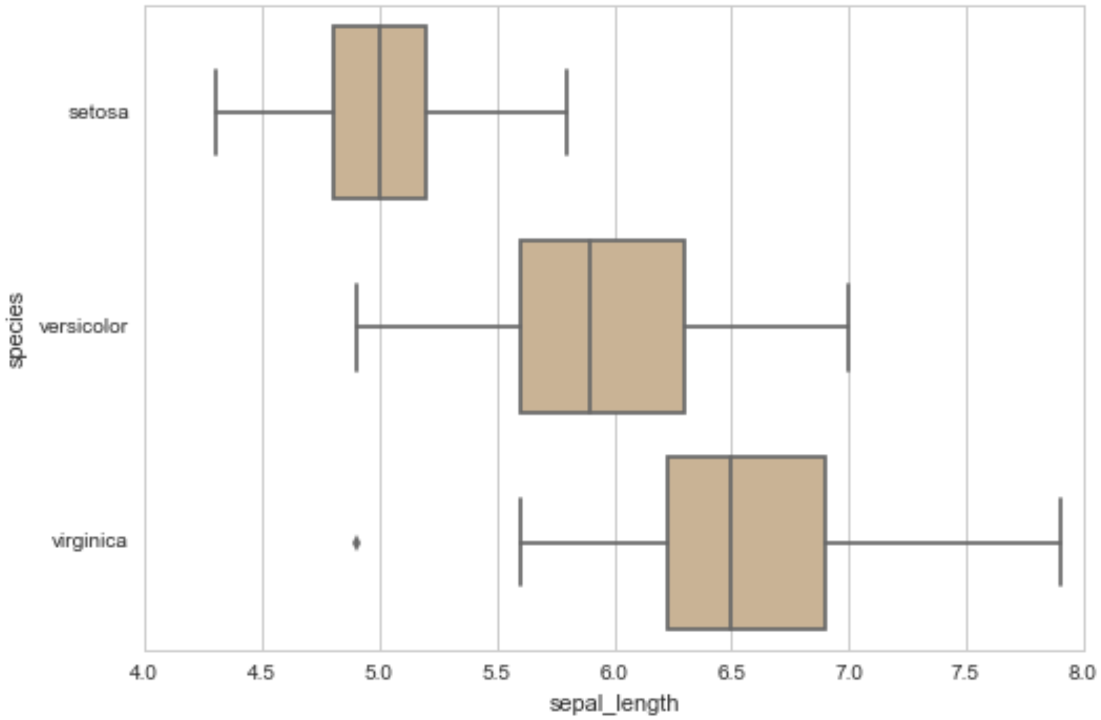
Box plots can summarize the distribution of one variable



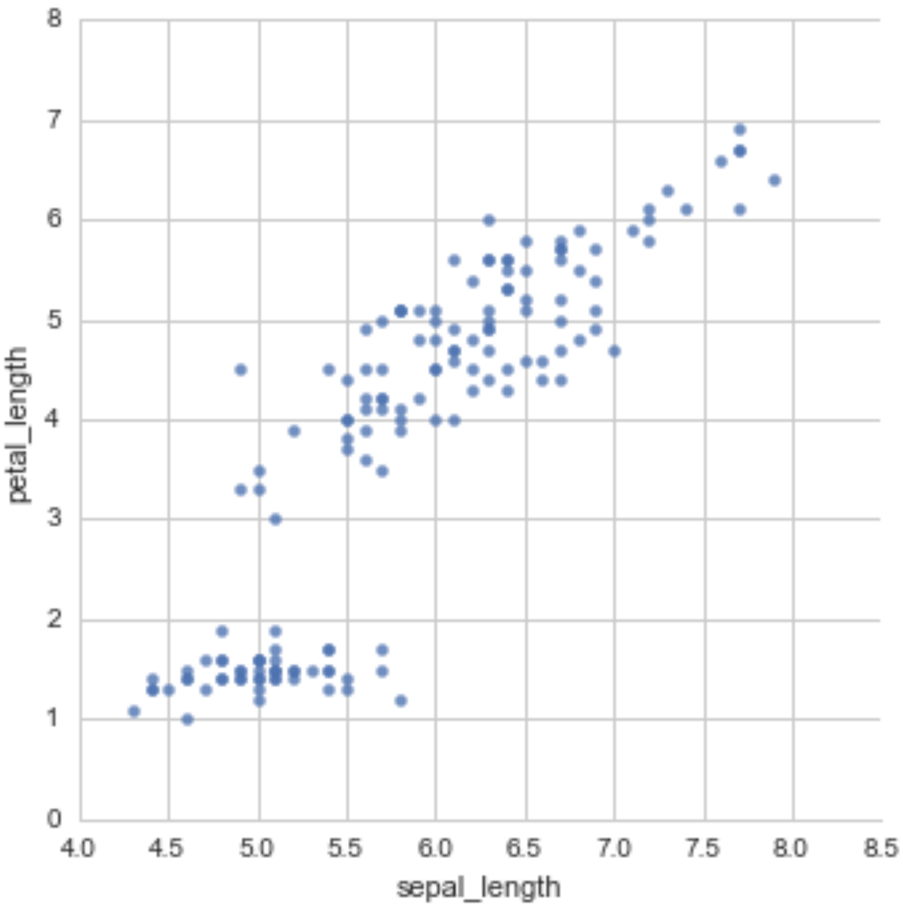


	sepal_length
count	150.000000
mean	5.843333
std	0.828066
min	4.300000
25%	5.100000
50%	5.800000
75%	6.400000
max	7.900000

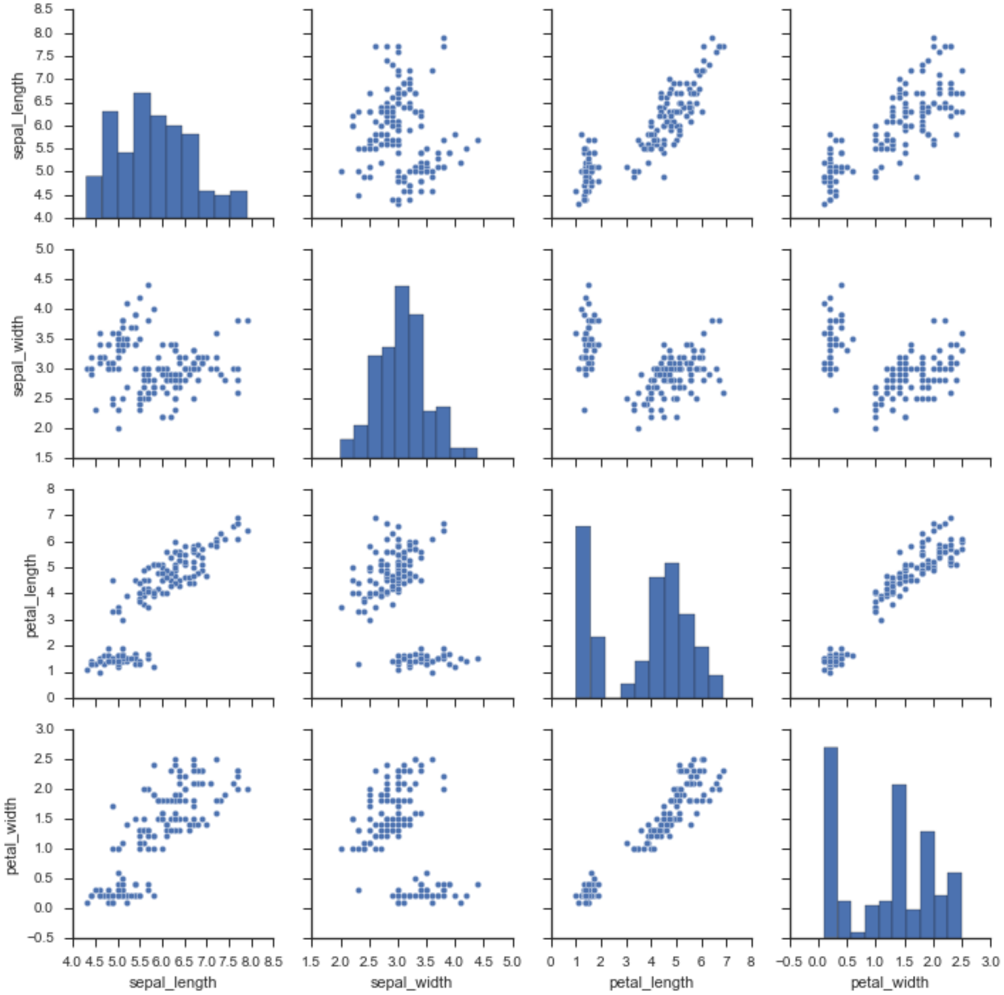
…and are great for comparing distributions across categories



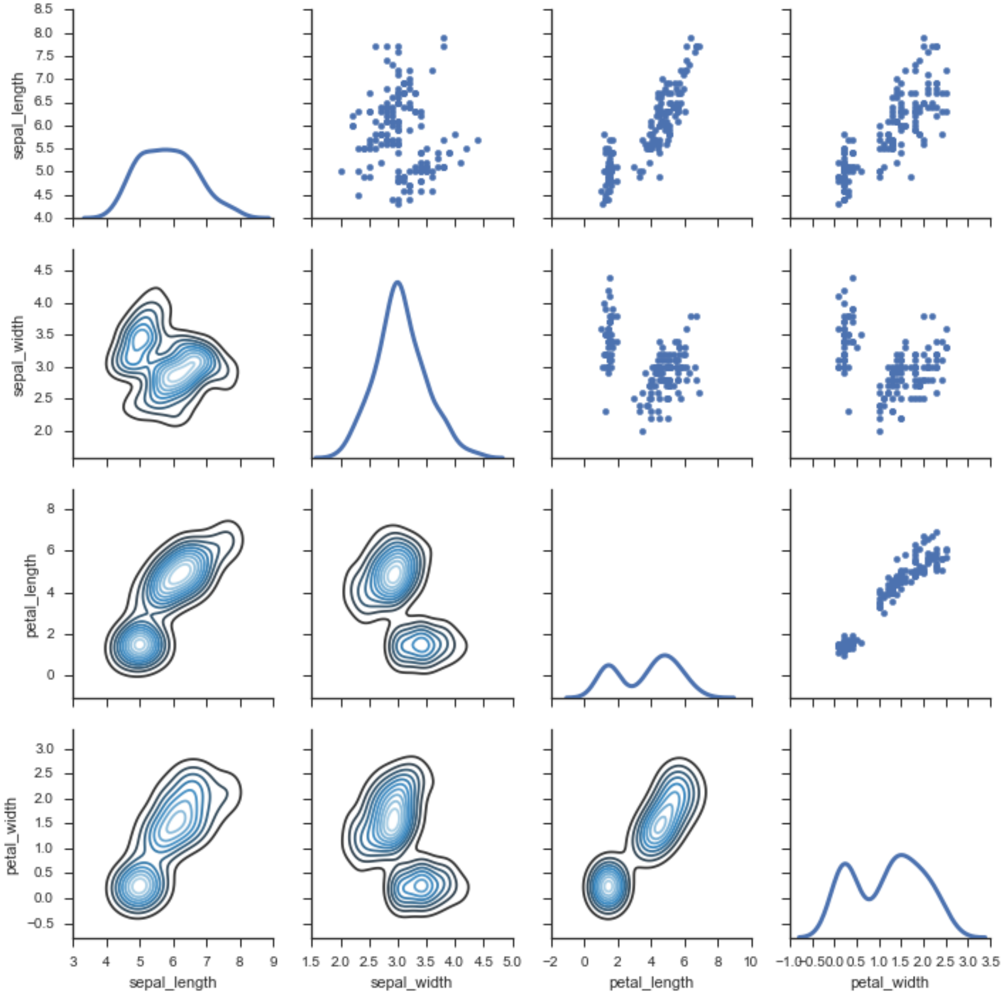
Scatter plots explore relationships between variable pairs



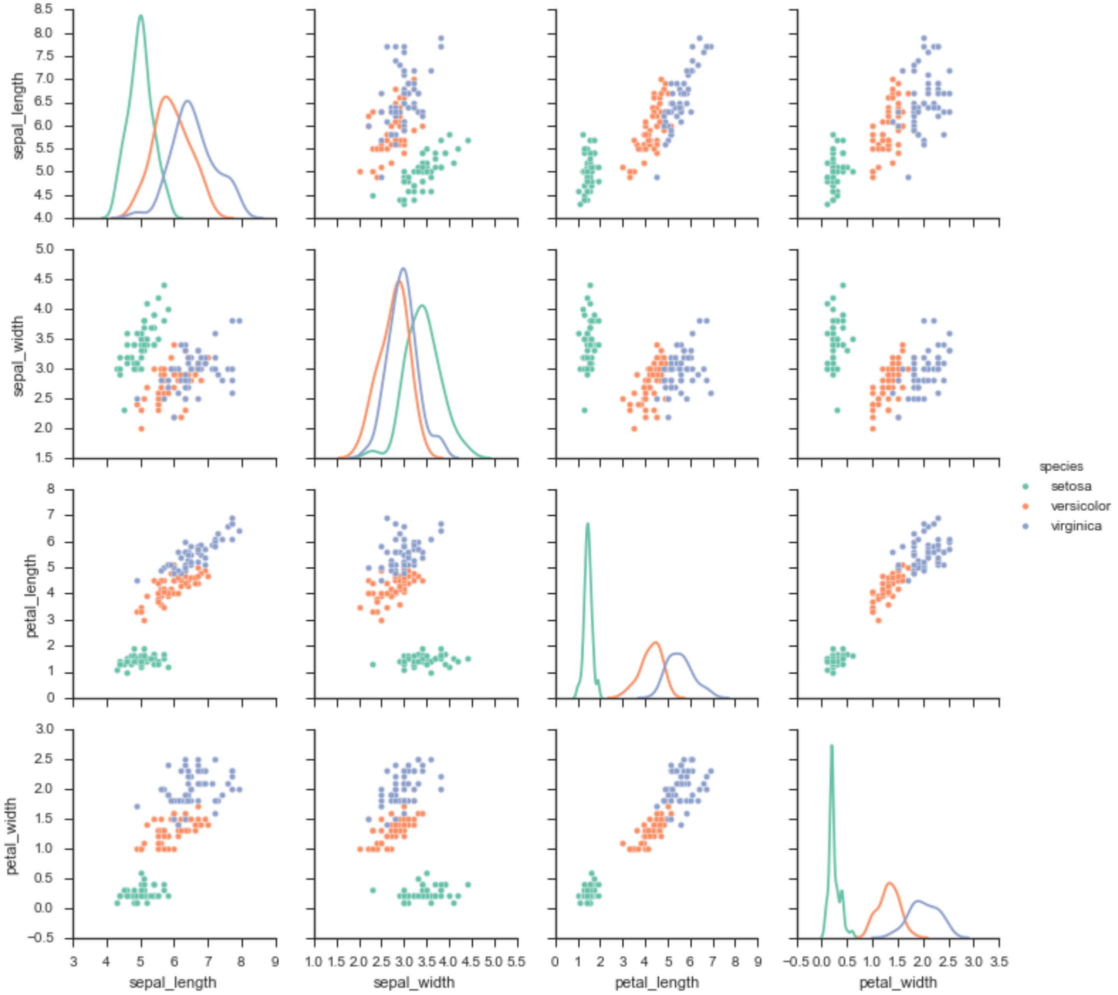
Pairs plots can help you explore relationships between variables

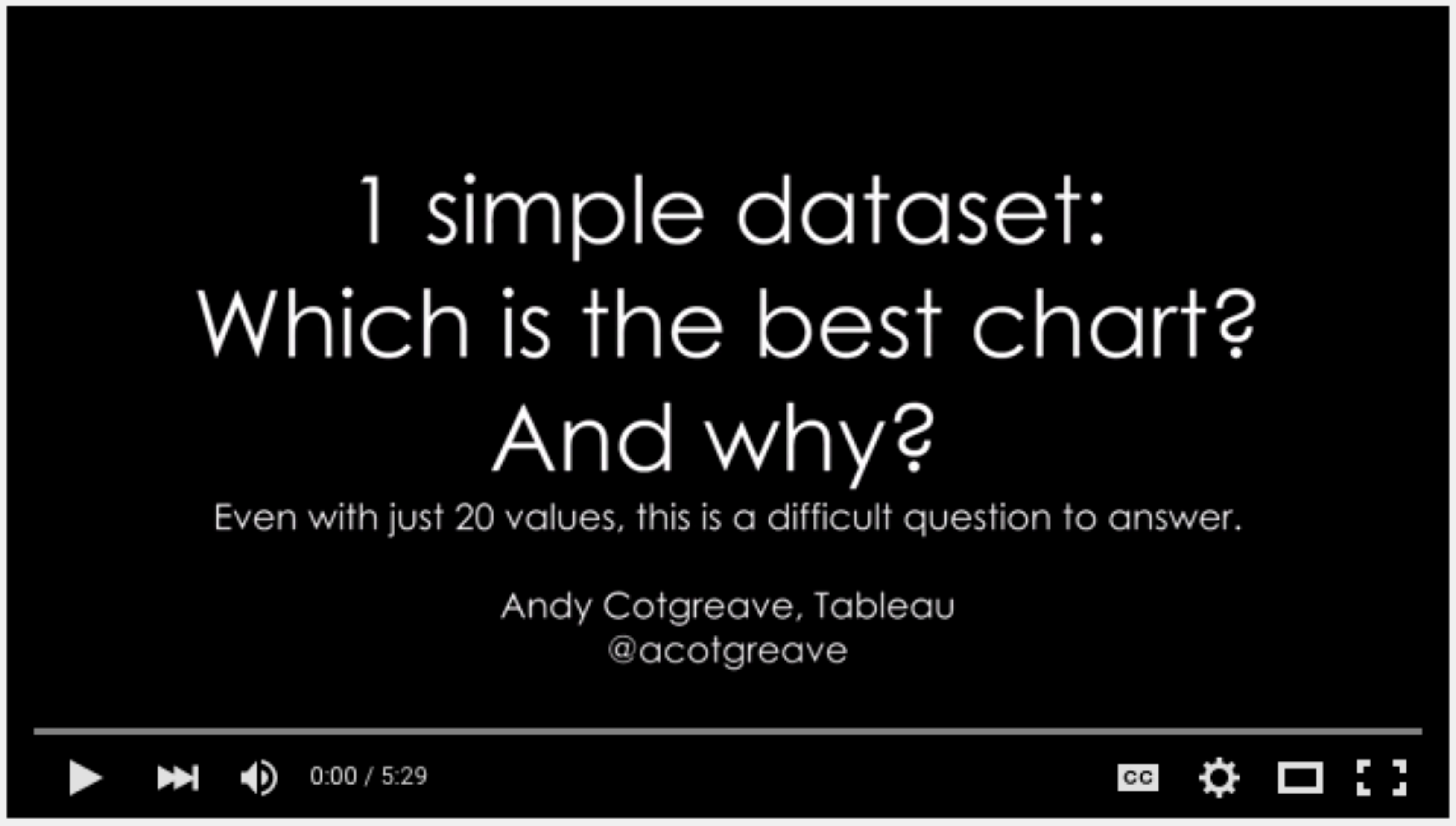


Pairs plots can help you explore relationships between variables



Pairs plots can help you explore relationships between variables





1 simple dataset:
Which is the best chart?
And why?

Even with just 20 values, this is a difficult question to answer.

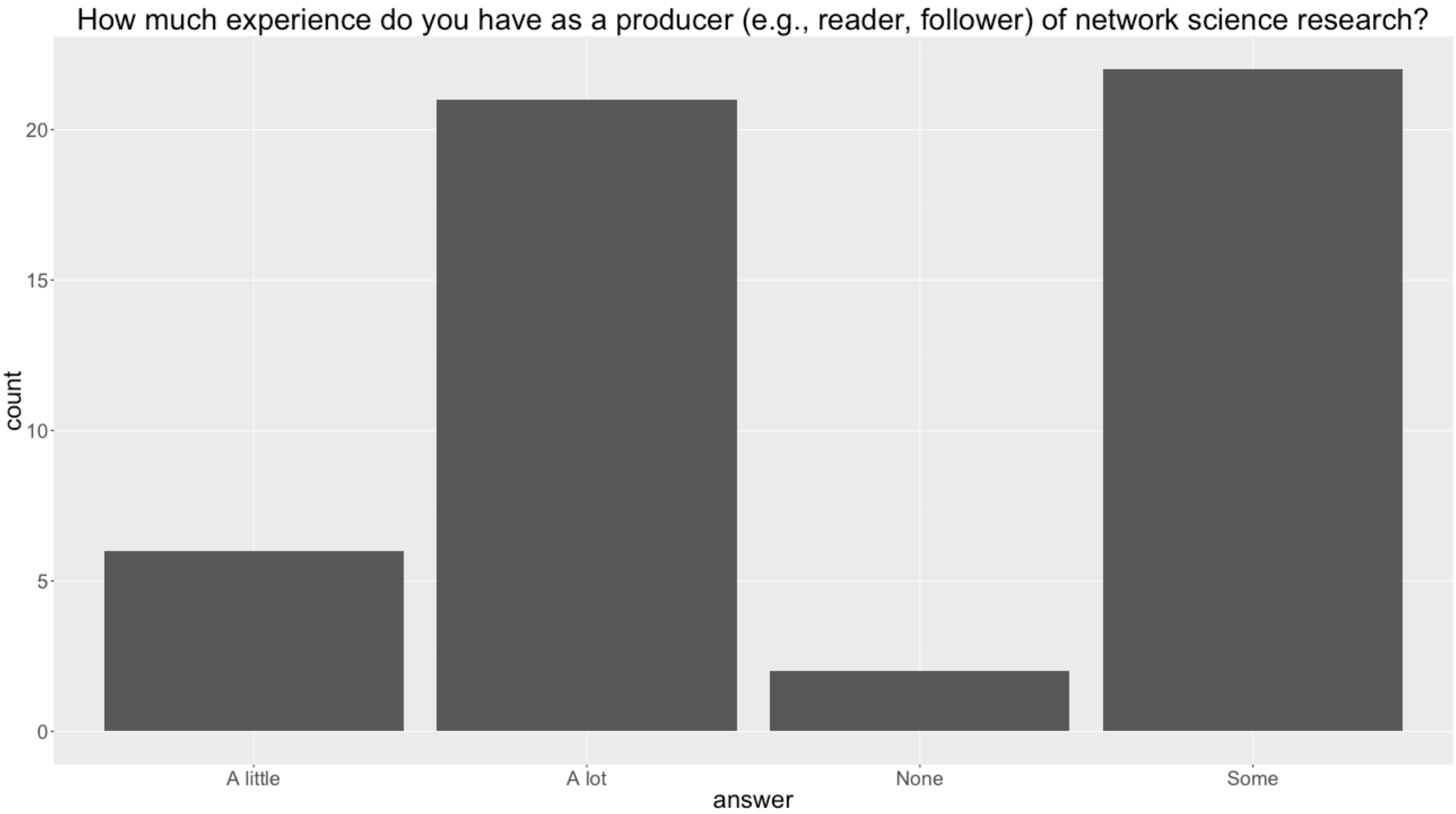
Andy Cotgreave, Tableau
@acotgreave

https://www.youtube.com/watch?v=AuJFuEq-qD8

ggplot2

Principles for Effective Visualizations

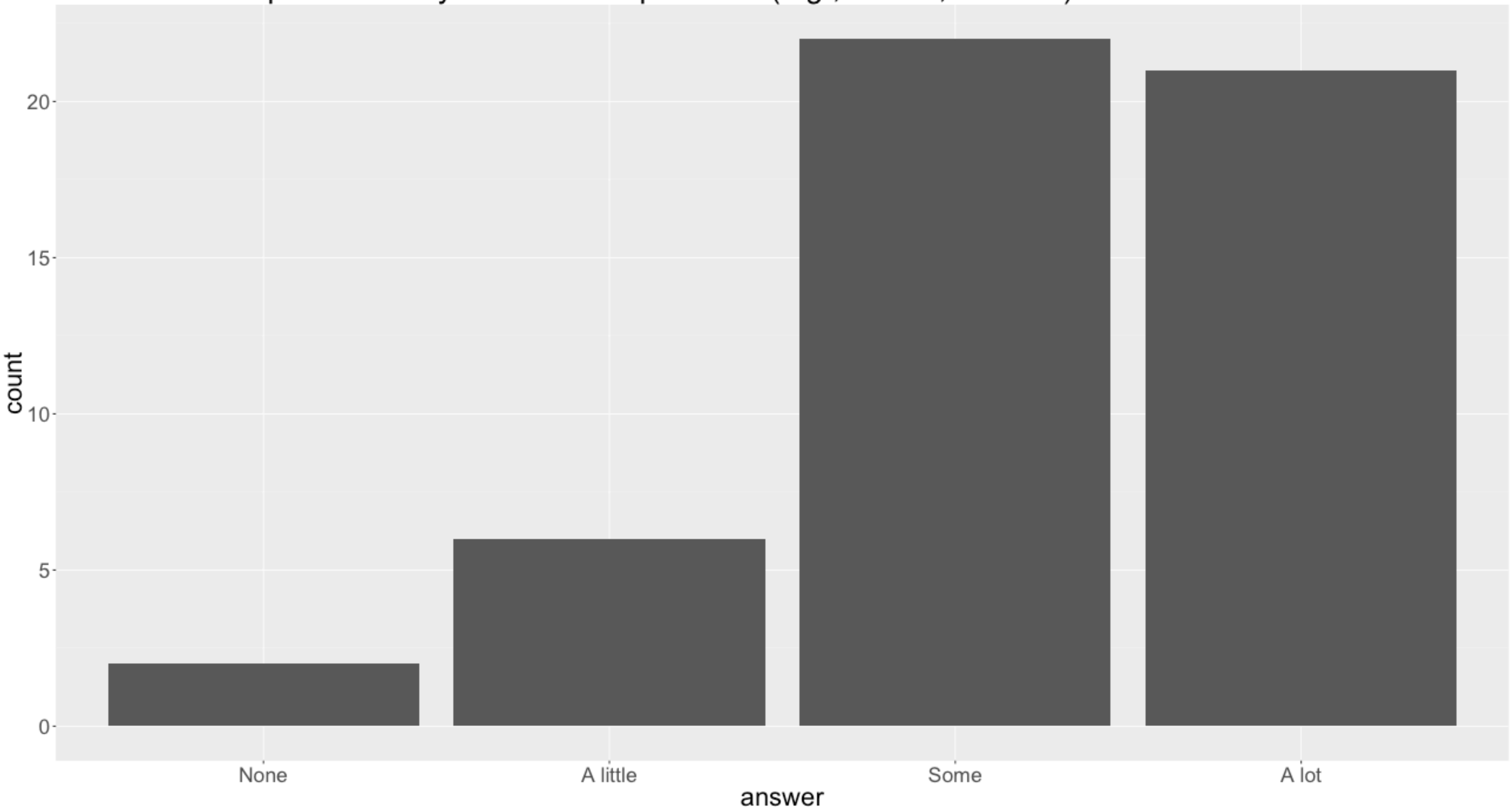
Principle 1: Order matters

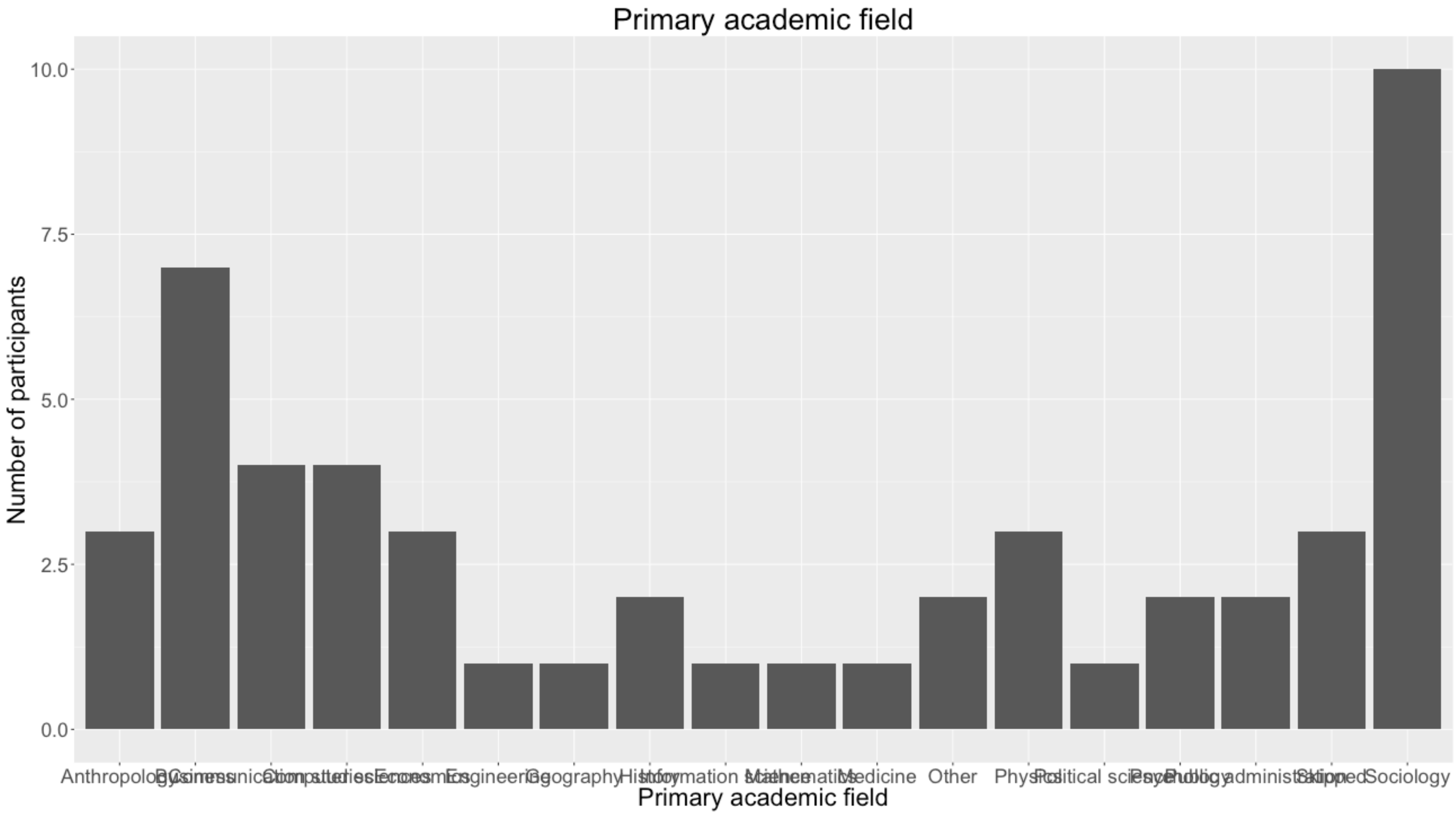

How much experience do you have as a producer (e.g., reader, follower) of network science research?
count
0
5
10
15
20
A little
A lot
None
Some
answer

Order by meaning

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

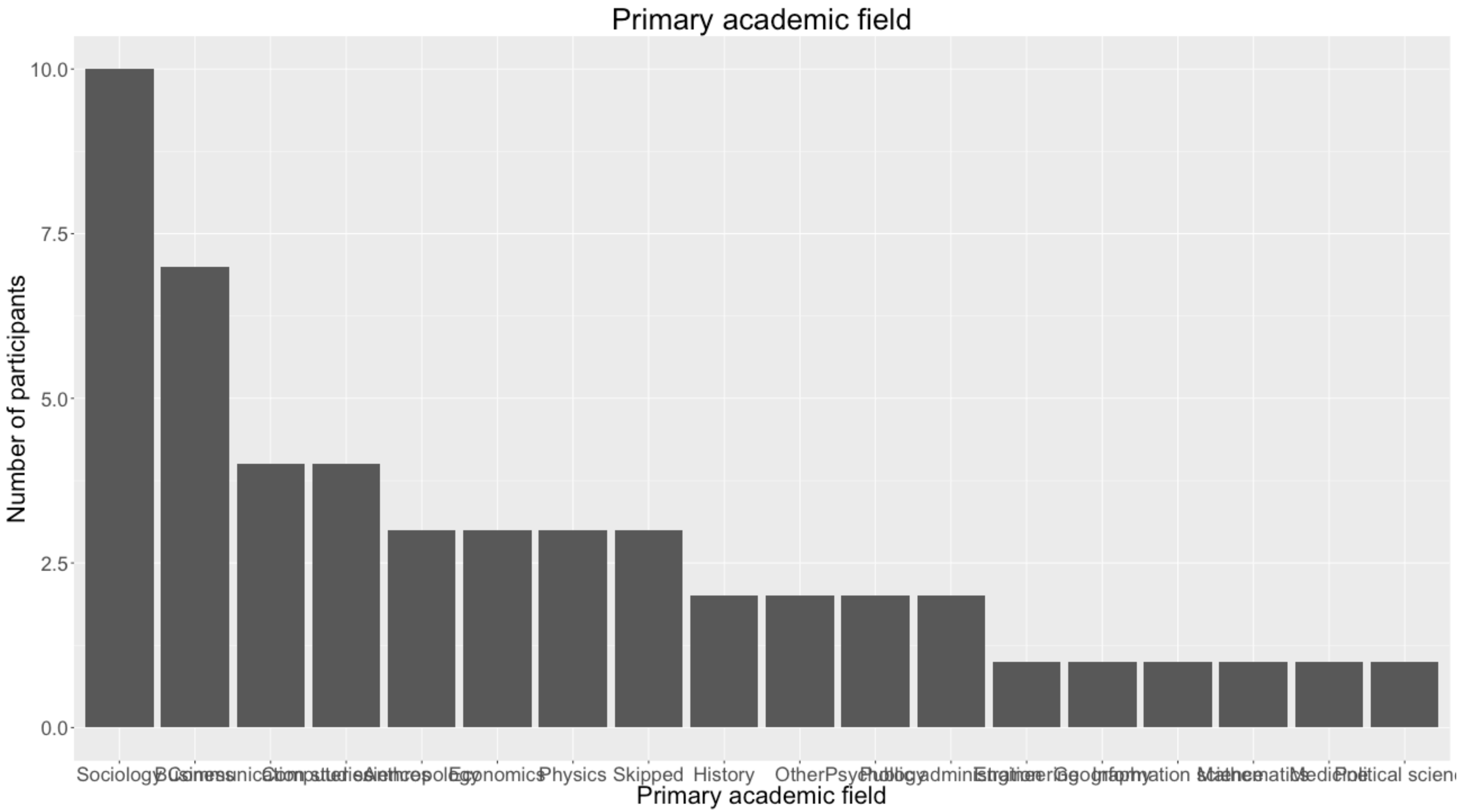
How much experience do you have as a producer (e.g., reader, follower) of network science research?



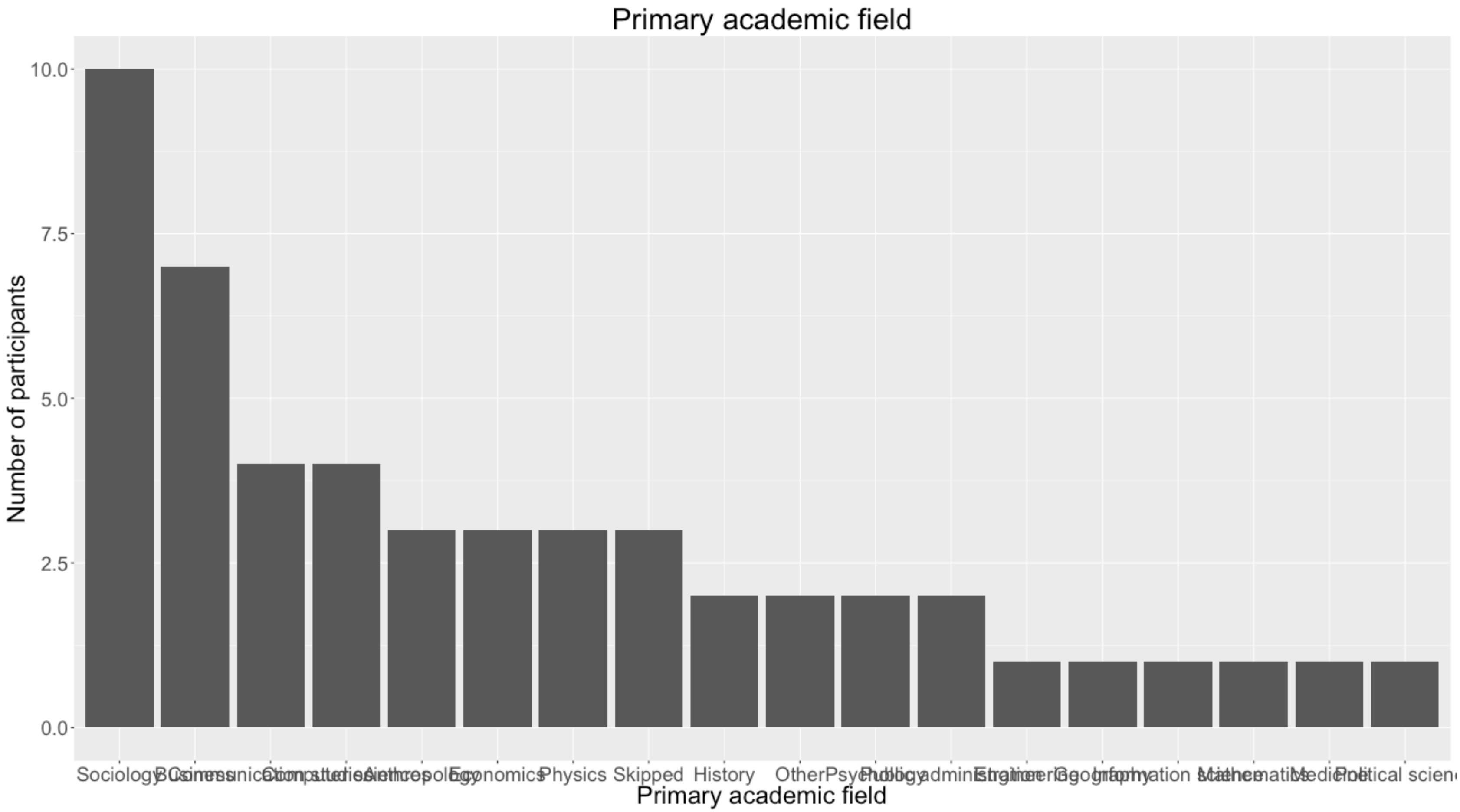

Primary academic field
Number of participants
10.0
7.5
5.0
2.5
0.0
Anthropology
Business
Communication studies
Computer science
Economics
Engineering
Geography
History
Information science
Mathematics
Medicine
Other
Physics
Political science
Psychology
Public administration
Skipped
Sociology
Primary academic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

Order by value

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

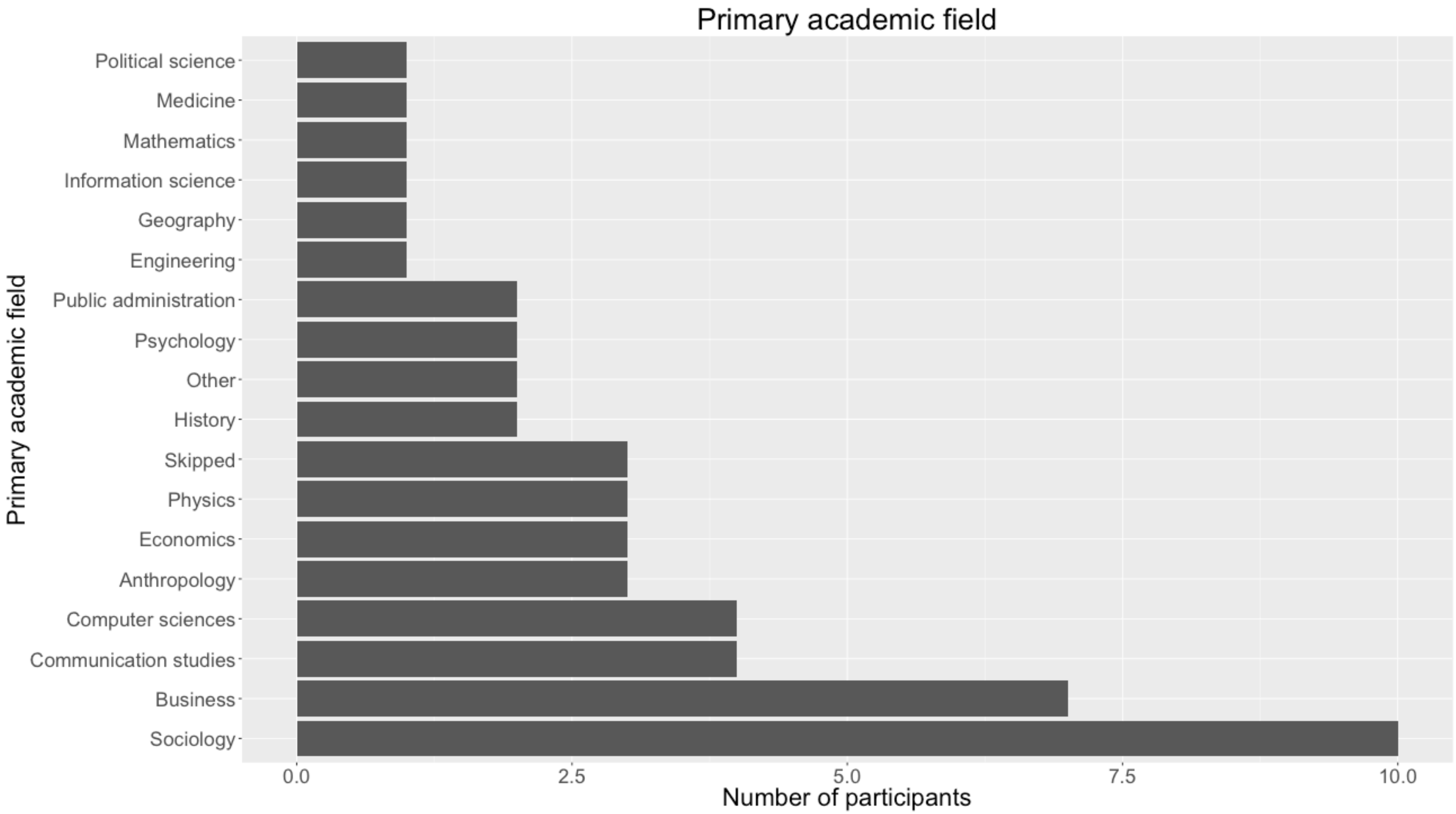

Primary academic field
Number of participants
10.0
7.5
5.0
2.5
0.0
Sociology
Business
Communication studies
Computer science
Anthropology
Economics
Physics
Skipped
History
Other
Psychology
Public administration
Engineering
Geography
Information science
Mathematics
Medicine
Political scienc
Primary academic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

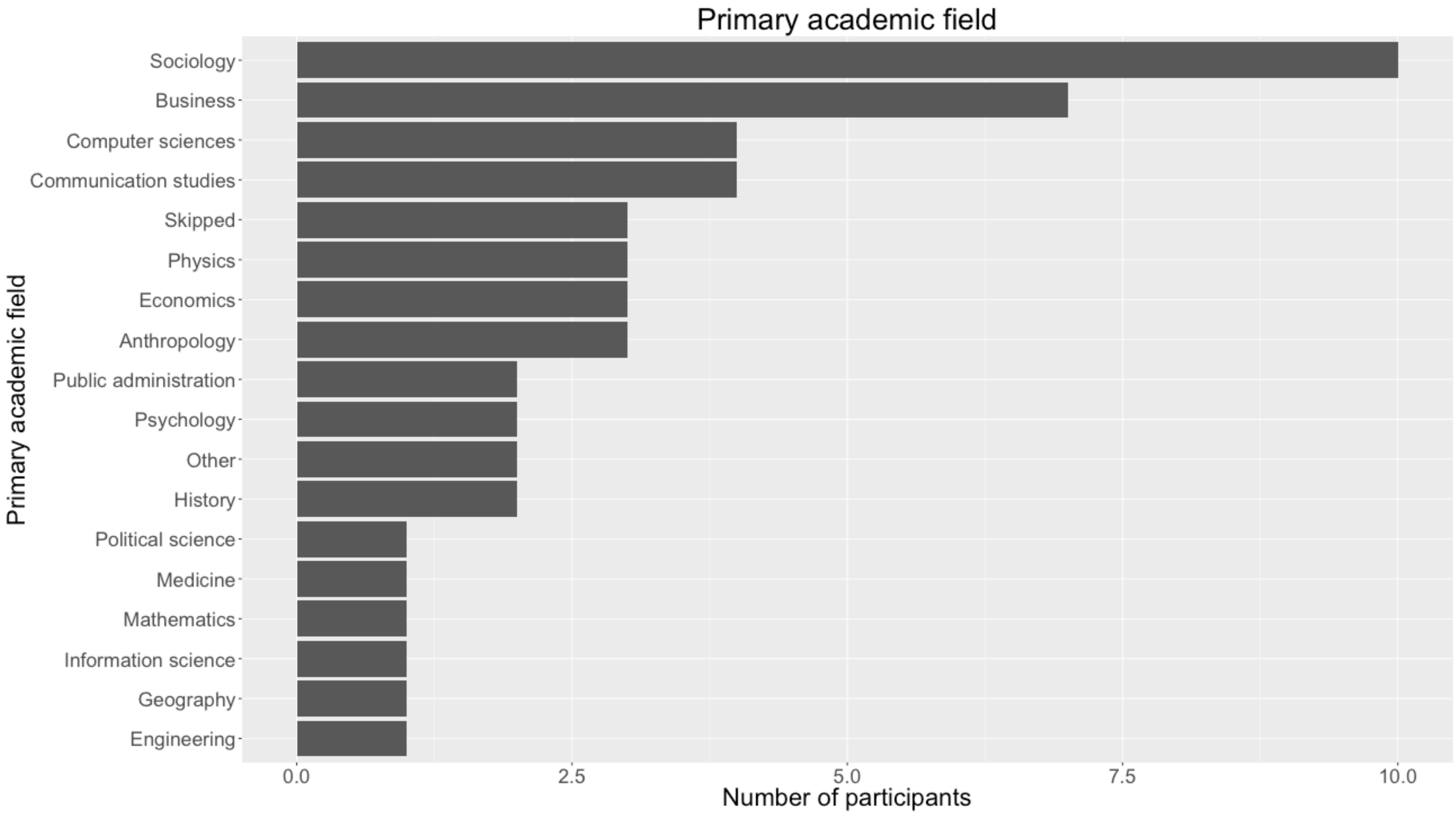
```
coord_flip()
```


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

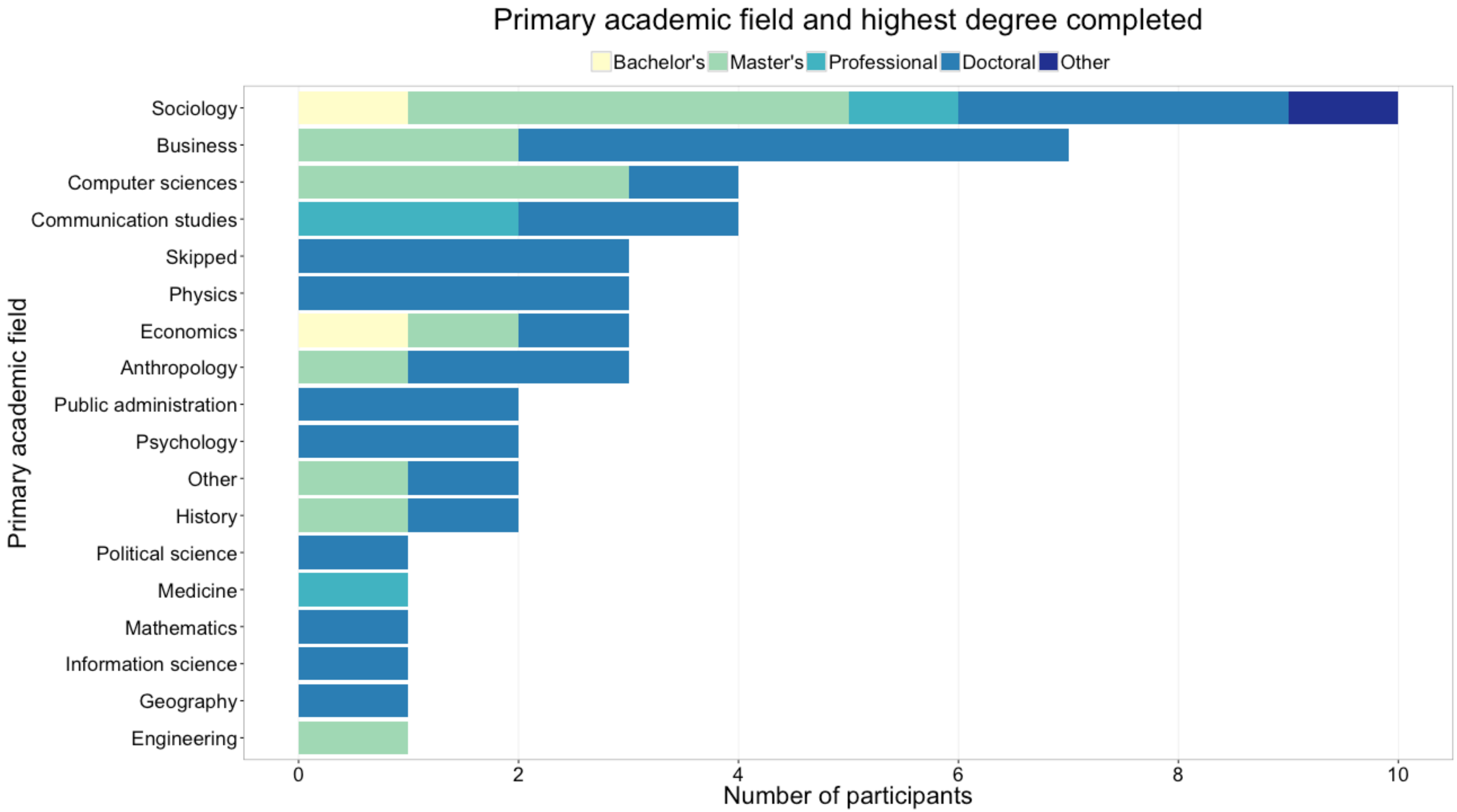
Oops!

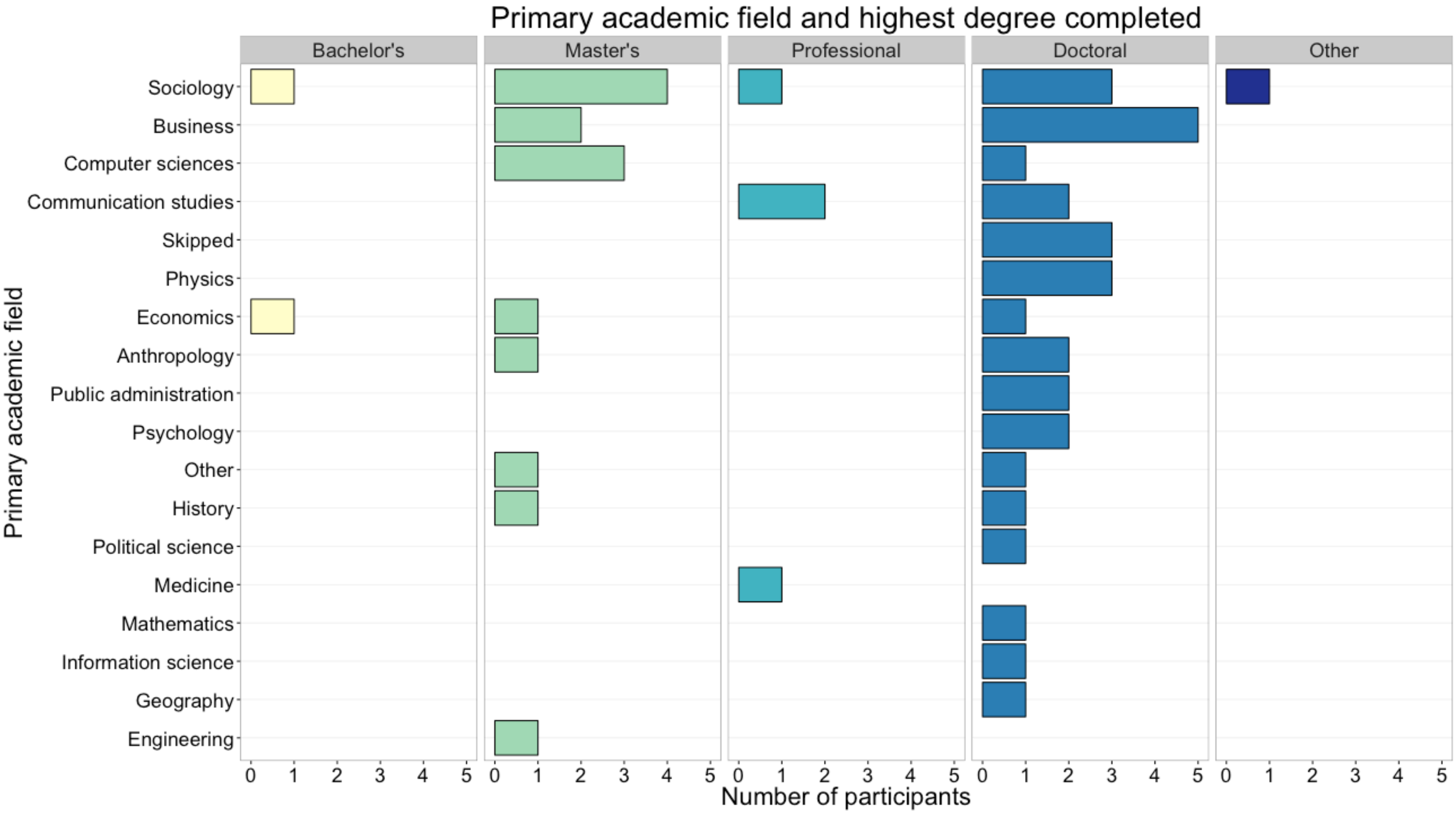
```
data$academic_field <-
     factor(data$academic_field,
           levels=names(
               sort(
                   table(data$academic_field),
                   ~~decreasing=TRUE~~)))
```

```
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

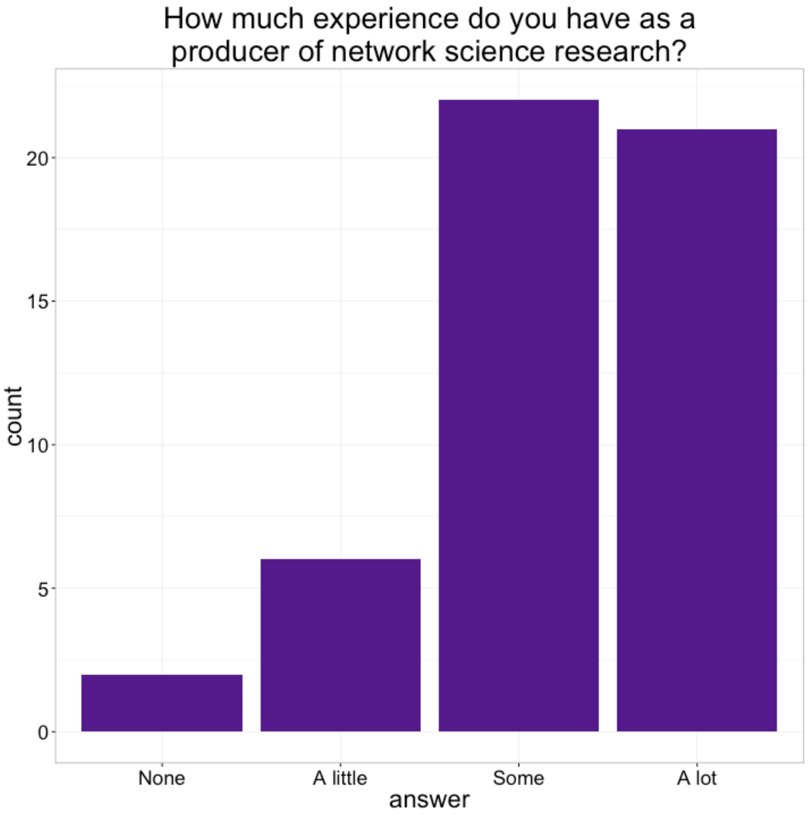
Other

Different placement helps with different comparisons

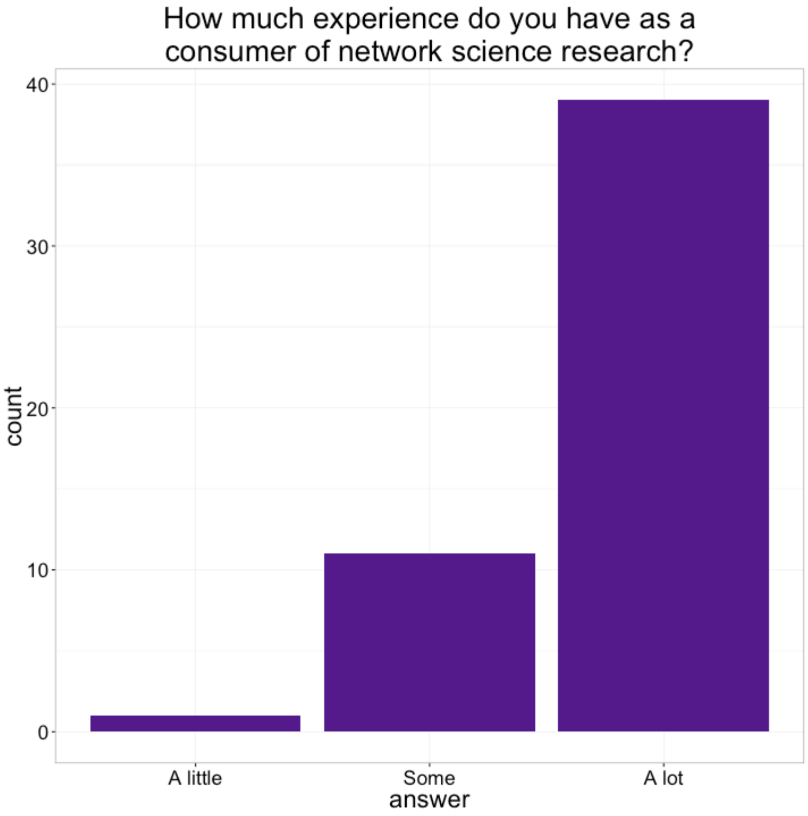
```
fill=highest_degree
```

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

Principle 4: Keep scales consistent



How much experience do you have as a producer of network science research?

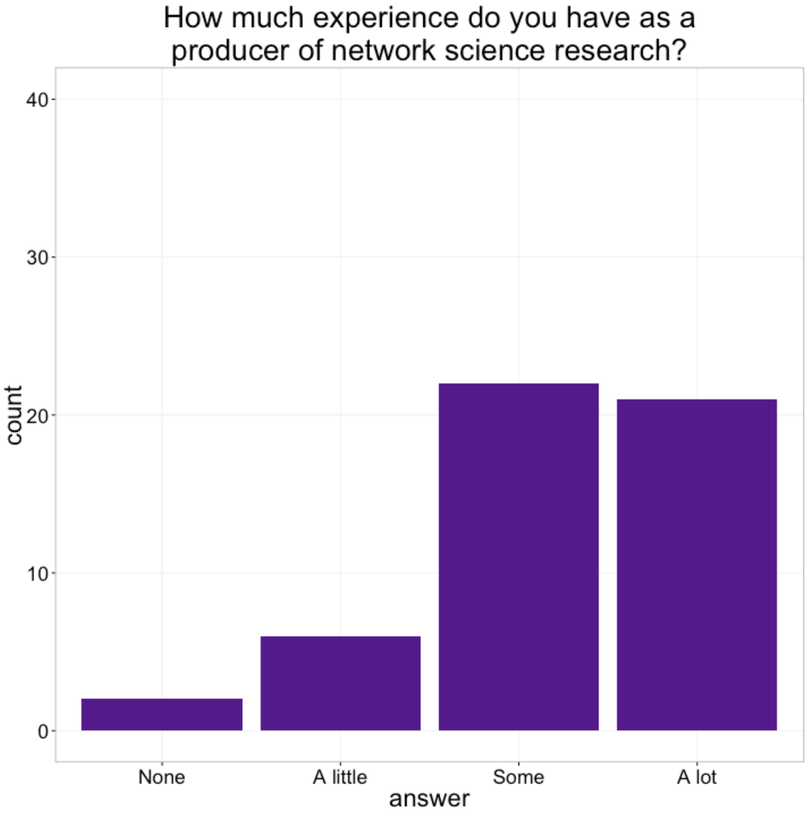


How much experience do you have as a consumer of network science research?

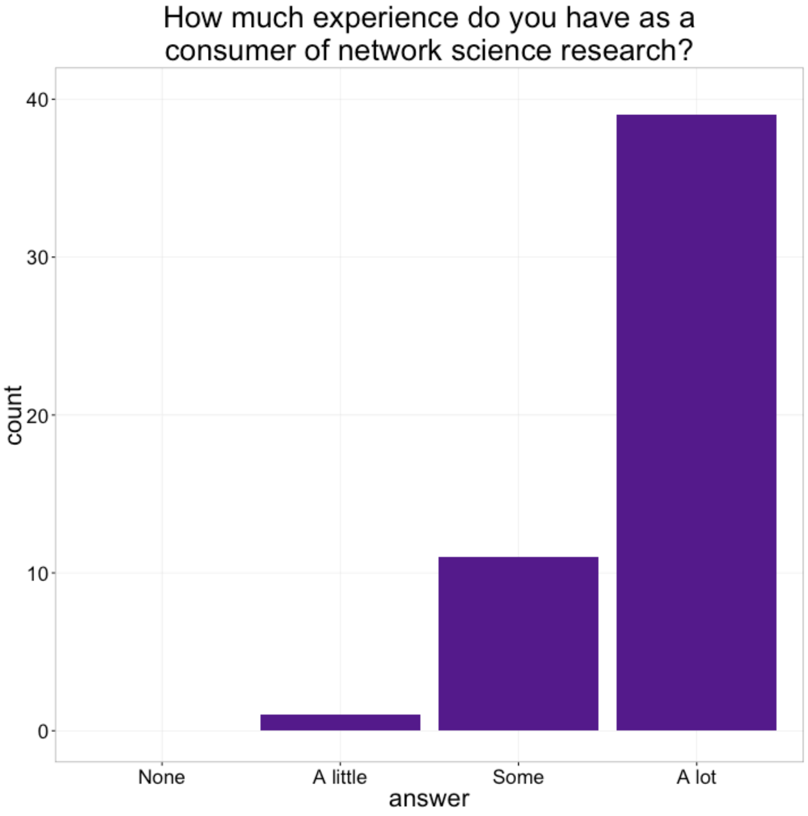
Keep all categories, manually set axes

```
scale_x_discrete(drop=FALSE)

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

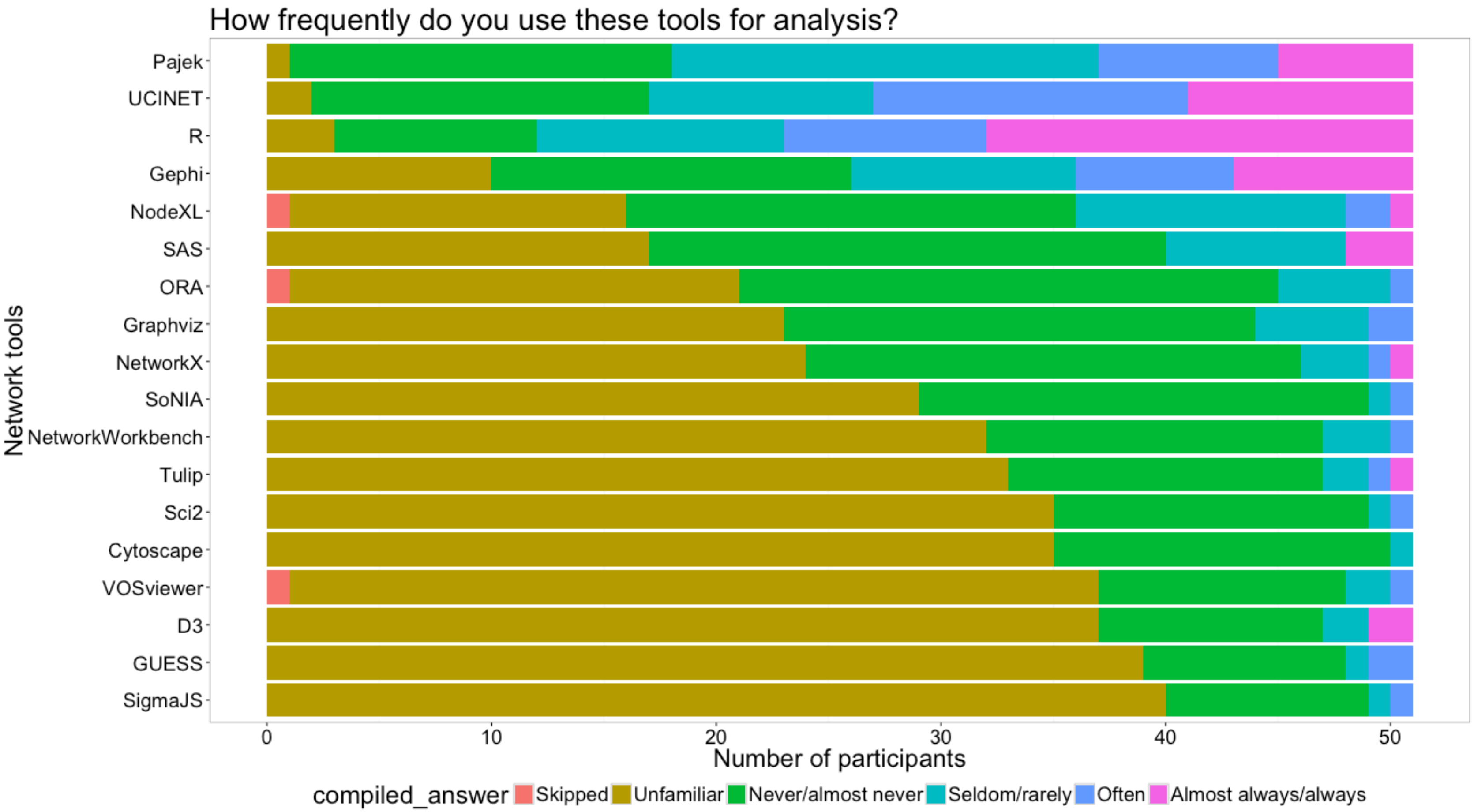


How much experience do you have as a producer of network science research?



How much experience do you have as a consumer of network science research?

Principle 5: Select meaningful colors

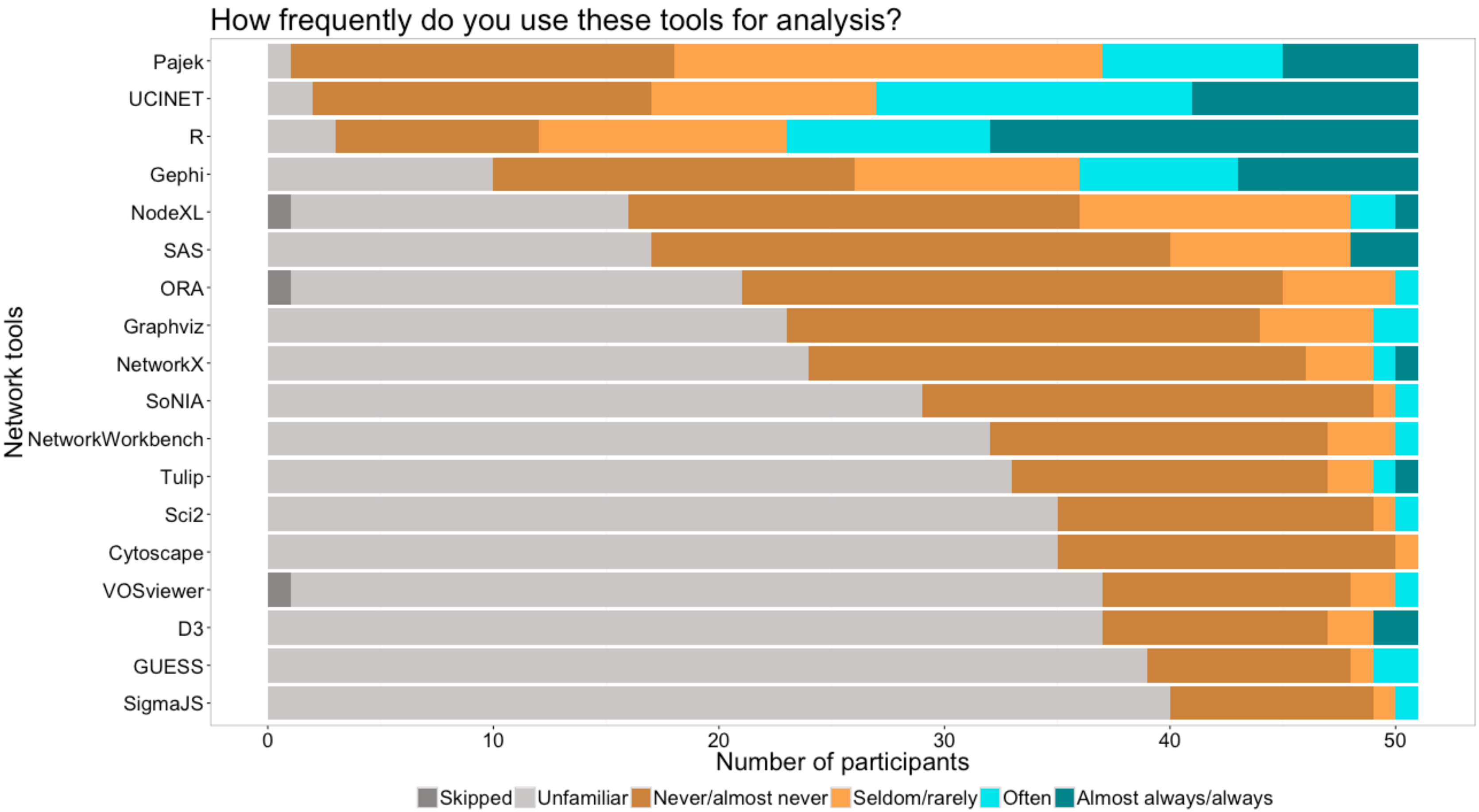

How frequently do you use these tools for analysis?
Network tools
Pajek
UCINET
R
Gephi
NodeXL
SAS
ORA
Graphviz
NetworkX
SoNIA
NetworkWorkbench
Tulip
Sci2
Cytoscape
VOSviewer
D3
GUESS
SigmaJS
0
10
20
30
40
50
Number of participants
compiled_answer
Skipped
Unfamiliar
Never/almost never
Seldom/rarely
Often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

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

Resources

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

Support Areas

Data Sources

Data Management

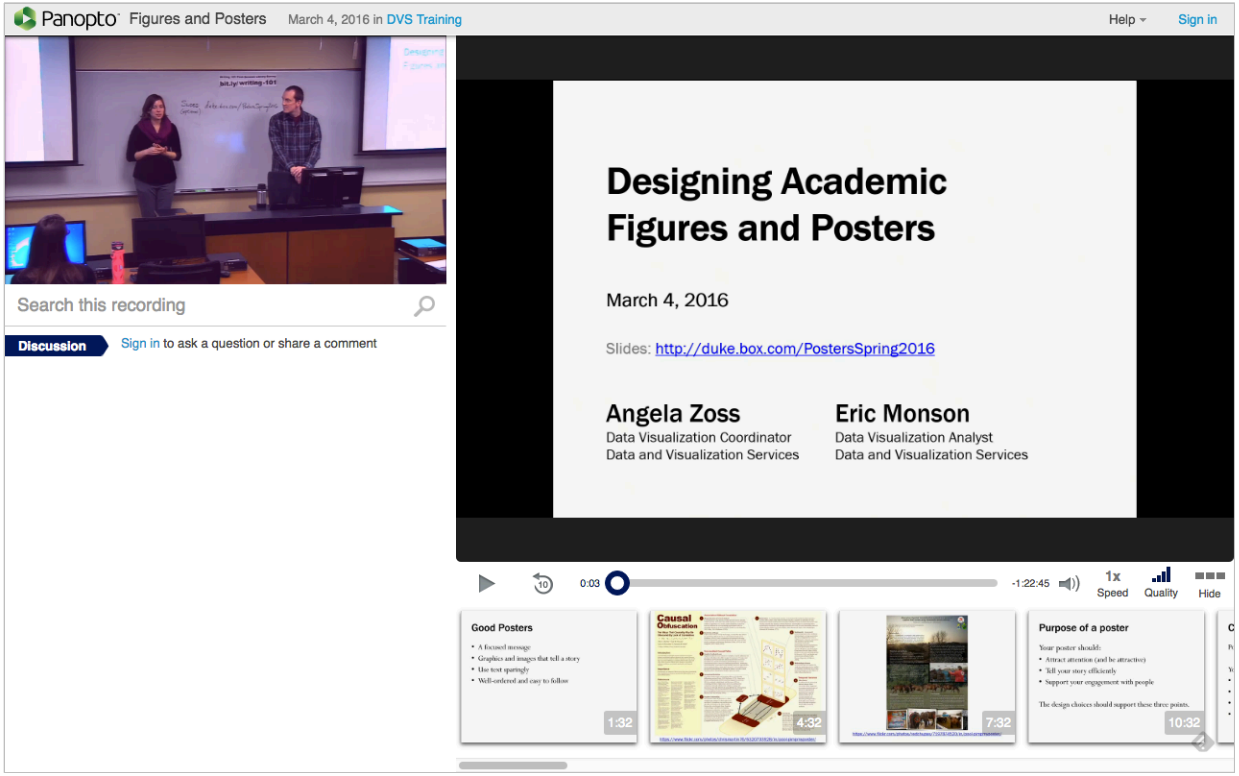
Data Cleaning

Data Analysis

Mapping and GIS

Data Visualization

Videos of past workshops



Sign in to ask a question or share a comment

Designing Academic Figures and Posters

March 4, 2016

Slides: http://duke.box.com/PostersSpring2016

Angela Zoss
Data Visualization Coordinator
Data and Visualization Services

Eric Monson
Data Visualization Analyst
Data and Visualization Services

Good Posters

- A focused message
- Graphics and images that tell a story
- Use text sparingly
- Well-ordered and easy to follow

Purpose of a poster

Your poster should:

- Attract attention (and be attractive)
- Tell your story efficiently
- Support your engagement with people

The design choices should support these three points.

http://bit.ly/DVSvideos

Questions?

askdata@duke.edu