Categorical data

`slides/06_categorical.pdf`

Numeric data

```
'data.frame':   15 obs. of  20 variables:
 $ a1 : num  18.6 37.6 71.6 94.2 100.2 ...
 $ a2 : num  17 38.2 67.8 106.8 64.2 ...
 $ a3 : num  19 36.2 90.4 110.9 83.4 ...
 $ a4 : num  6 48.6 77 115.5 94.1 ...
 $ a5 : num  15.8 43.6 81.6 133 87.6 ...
 $ a6 : num  0 22.8 36.6 111.2 54.8 ...
 $ a7 : num  6.2 31 62 101.5 66.8 ...
 $ a8 : num  5 30.2 31.1 89.7 53.5 ...
 $ a9 : num  7.2 27 65 124.1 104.9 ...
 $ a10: num  0 25.8 60.8 69.5 81.9 ...
 $ a11: num  8 19.4 60.2 102.7 56.5 ...
 $ a12: num  15 38 71.4 106.9 67.4 ...
 $ a13: num  2.8 35.8 66.6 121.5 67.7 ...
 $ a14: num  4.4 35.4 48 120.7 41 ...
 $ a15: num  6.6 34.8 52 100.6 78 ...
```

Categorical data

</>

```
tibble [1,373 × 12] (S3: tbl_df/tbl/data.frame)
 $ respondent_id   : num [1:1373] 3308895255 3308891308 3308891135 3308879091 3308871671
...
 $ knowledge       : Ord.factor w/ 4 levels "Novice"<"Intermediate"<..: 2 1 2 1 1 3 1 3
1 1 ...
 $ interest        : Ord.factor w/ 4 levels "Not at all"<"Not much"<..: 3 3 4 2 2 4 3 4
2 3 ...
 $ gender          : chr [1:1373] "Male" "Male" "Male" "Male" ...
 $ age             : Factor w/ 4 levels "18-29","30-44",..: 1 1 2 3 2 2 3 3 2 NA ...
 $ household_income: Factor w/ 5 levels "$0 - $24,999",..: 4 4 3 1 2 3 NA 1 3 NA ...
 $ education       : Ord.factor w/ 5 levels "Less than high school degree"<..: 1 3 5 1 2
5 2 3 3 NA ...
 $ location        : chr [1:1373] "West South Central" "West South Central" "Pacific"
"New England" ...
 $ algeria         : chr [1:1373] "N/A" "N/A" "3" "N/A" ...
 $ argentina       : chr [1:1373] "3" "N/A" "4" "3" ...
```

Two geoms for bar charts

- Binned data (has a count column) `geom_col()`
- Unbinned data (no count column) `geom_bar()`

`geom_col()`

- Requires an **x** and **y**
- Intended to be used with one **continuous** and one **discrete** variables but other combinations may also work

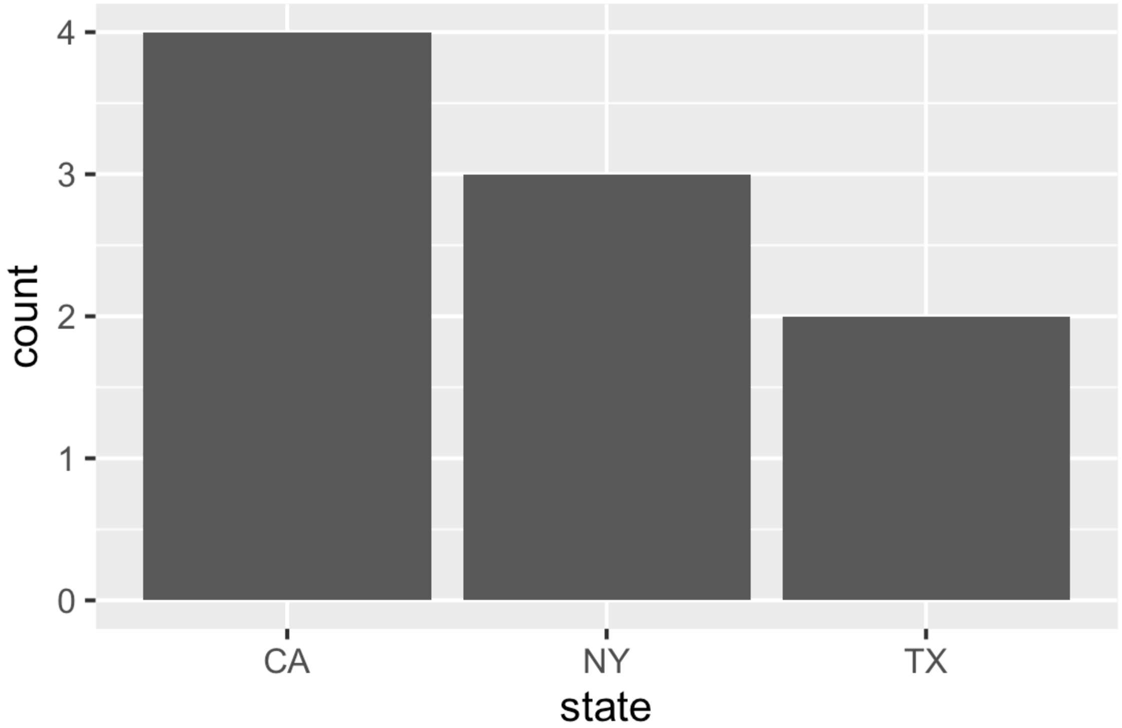
Look at the data

```
df_binned
```

```
  state count
1    CA     4
2    NY     3
3    TX     2
```

Bar chart with binned data

```r
ggplot(df_binned, aes(x = state, y = count)) +
  geom_col()
```

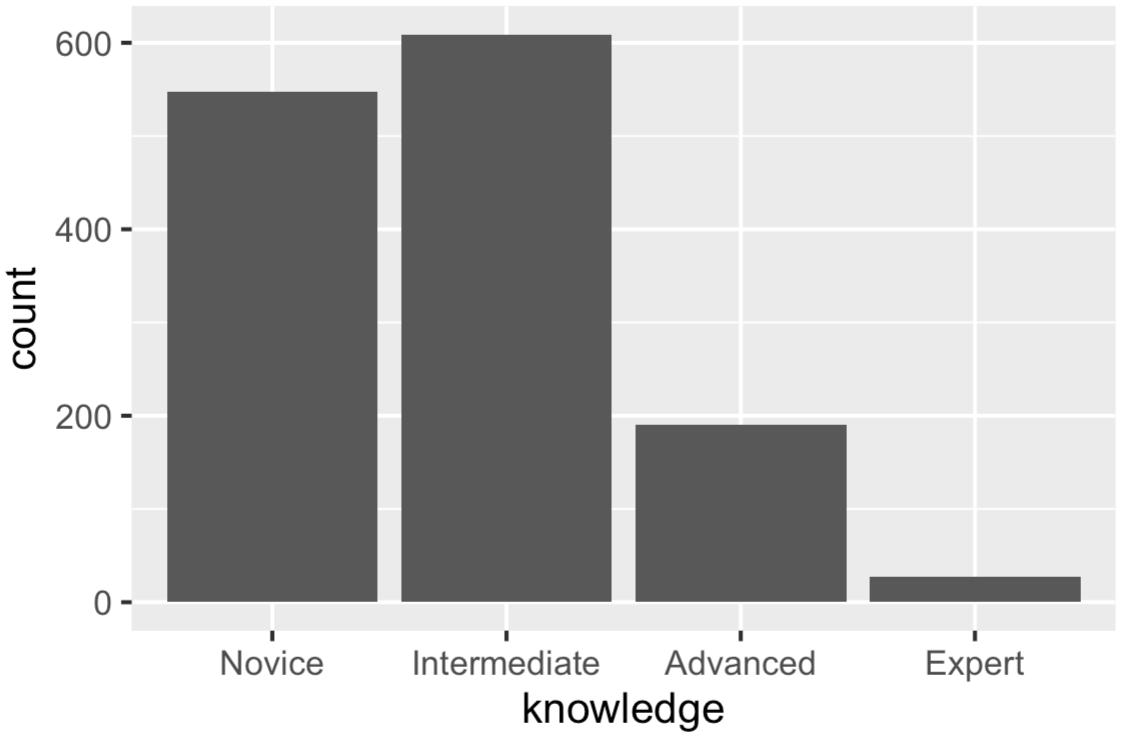


`geom_bar()`

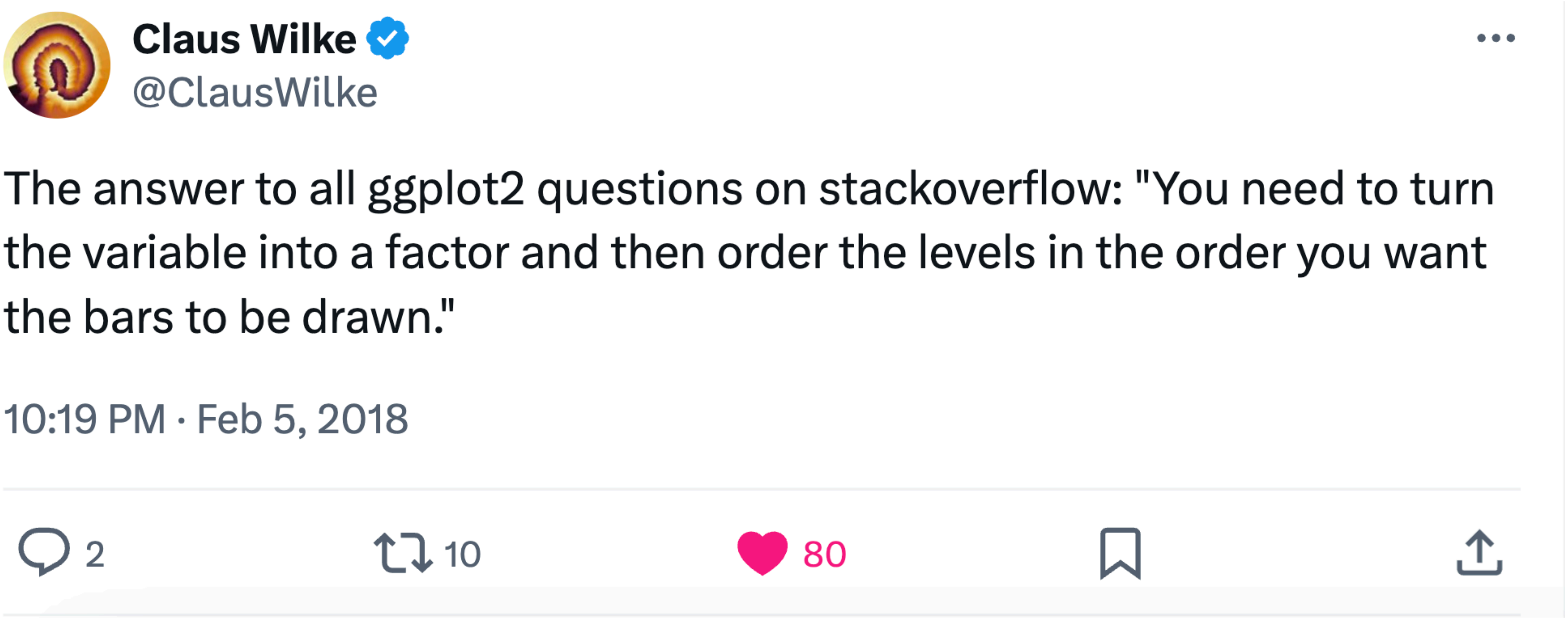
- Requires an **x** or **y**
- Intended to be used with one **discrete** variable (unbinned data)

Bar chart with unbinned data

```
ggplot(food_world_cup, aes(x = knowledge)) +
  geom_bar()
```



Bar order



Claus Wilke
@ClausWilke

The answer to all ggplot2 questions on stackoverflow: "You need to turn the variable into a factor and then order the levels in the order you want the bars to be drawn."

10:19 PM · Feb 5, 2018

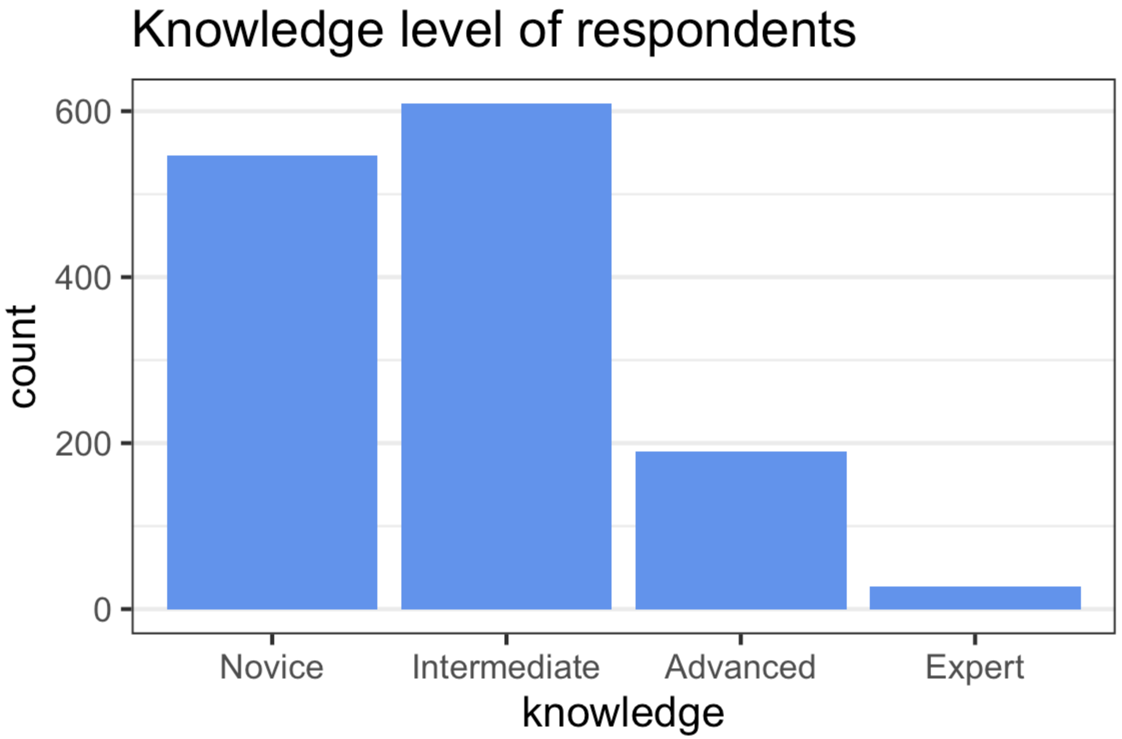
Types of data

- nominal does not have a fixed category order
- ordinal does have a fixed category order
- (“real”) discrete, small number of possibilities
- Not always clearcut: nominal vs. ordinal, ordinal vs. discrete, etc.
- Sometimes numbers = nominal, not discrete

Ordinal data

Sort in logical order of the categories (left to right)

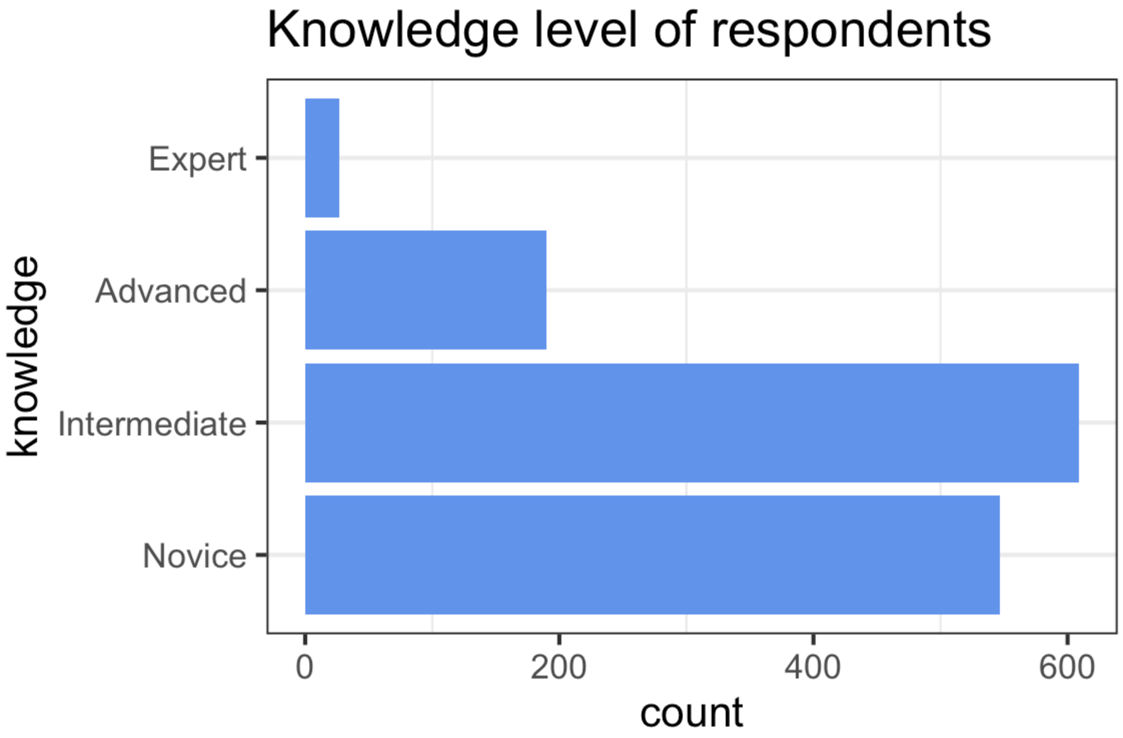
</>



Ordinal data, horizontal bars

Sort in logical order of the categories (starting at bottom OR top)

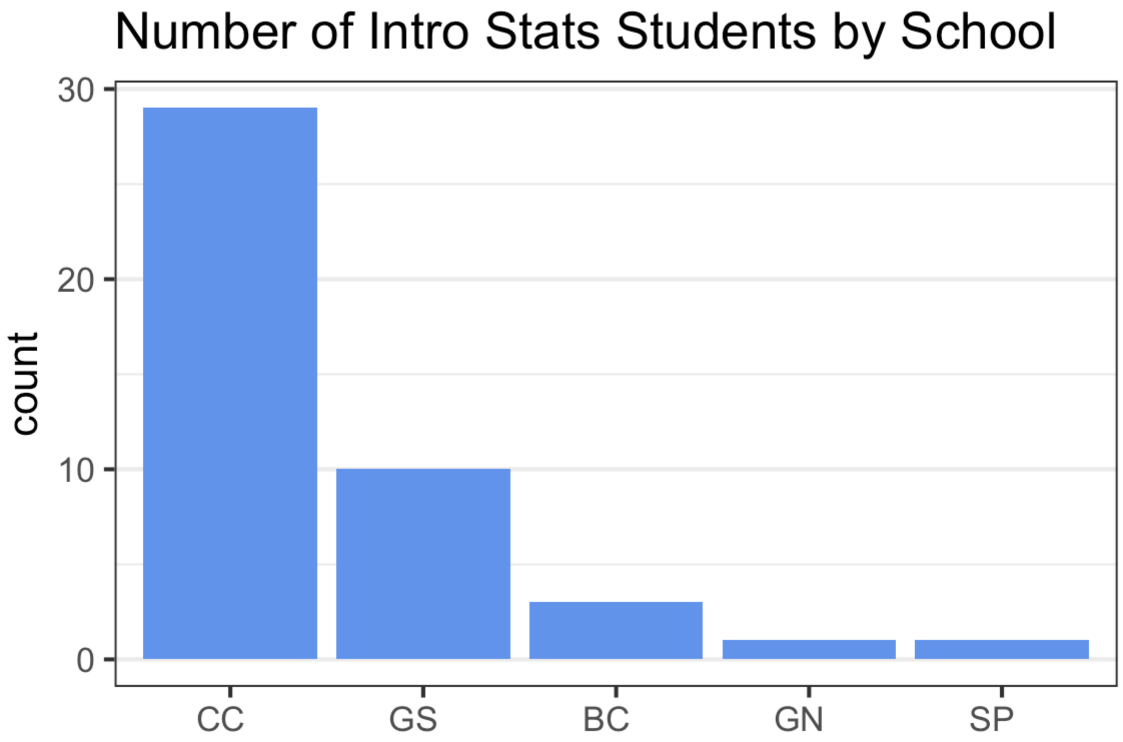
</>



Nominal data, vertical bars

Sort from highest to lowest count (left to right, or top to bottom)

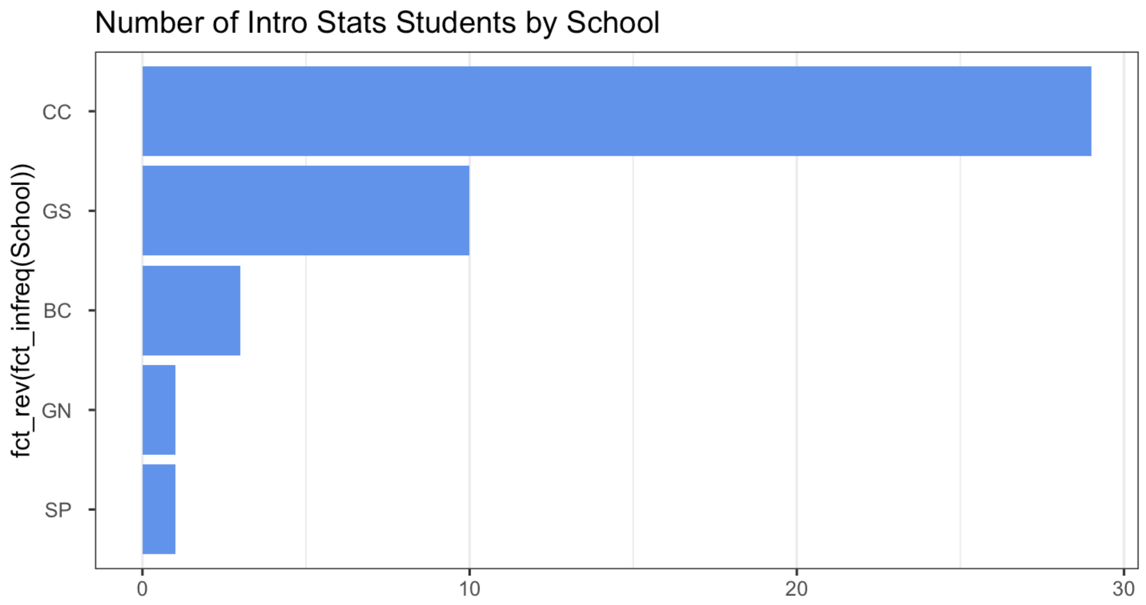
</>



Nominal data, horizontal bars

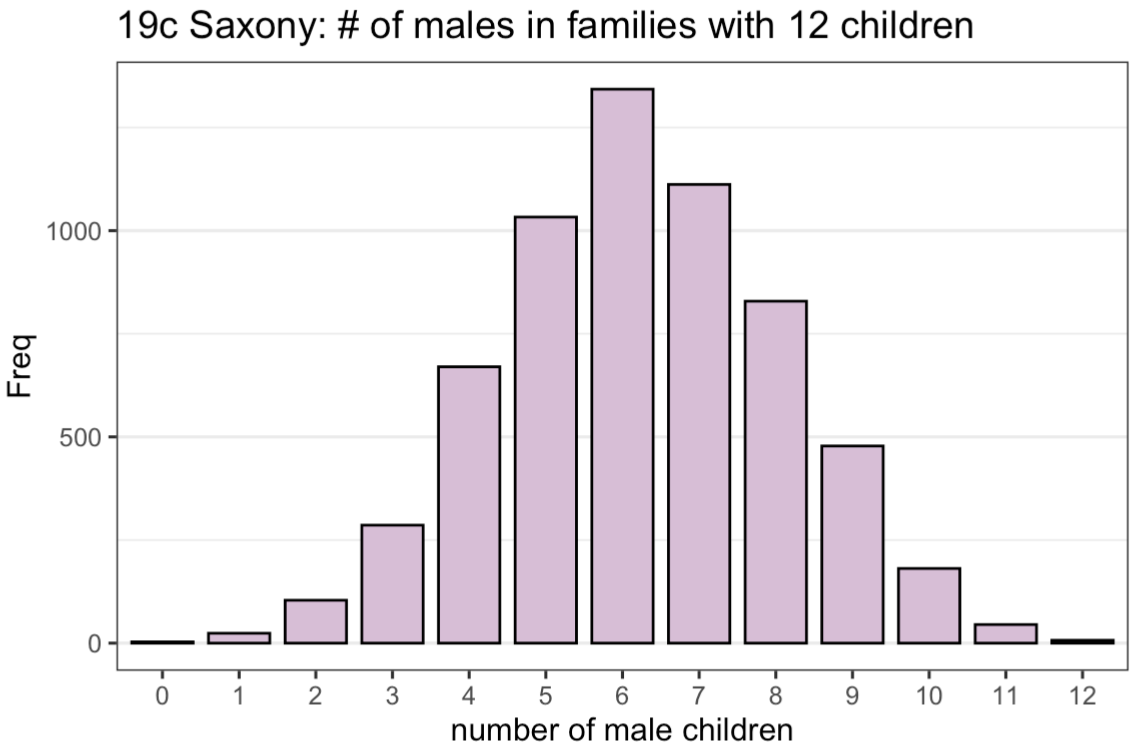
… or top to bottom

</>

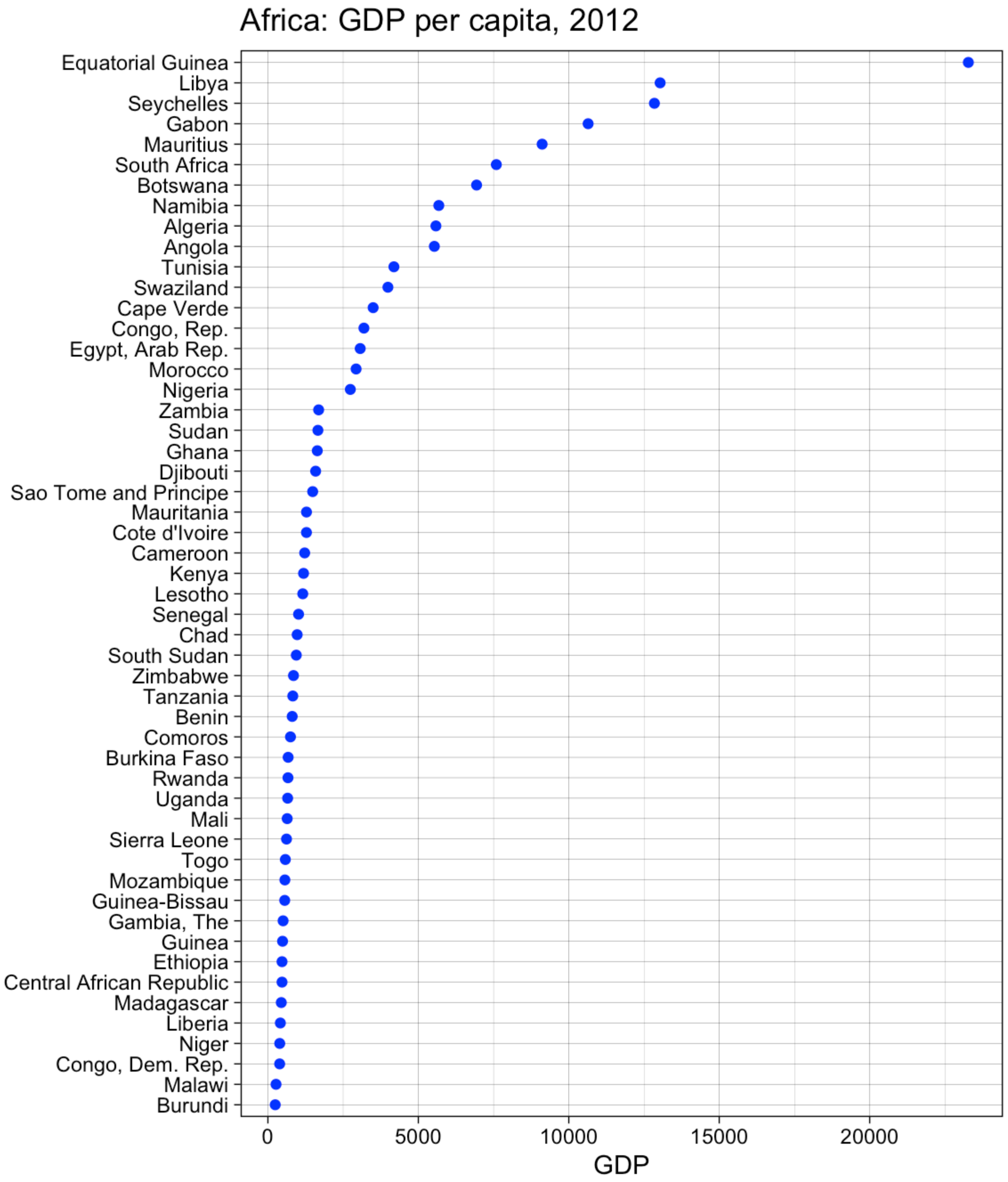


Discrete data

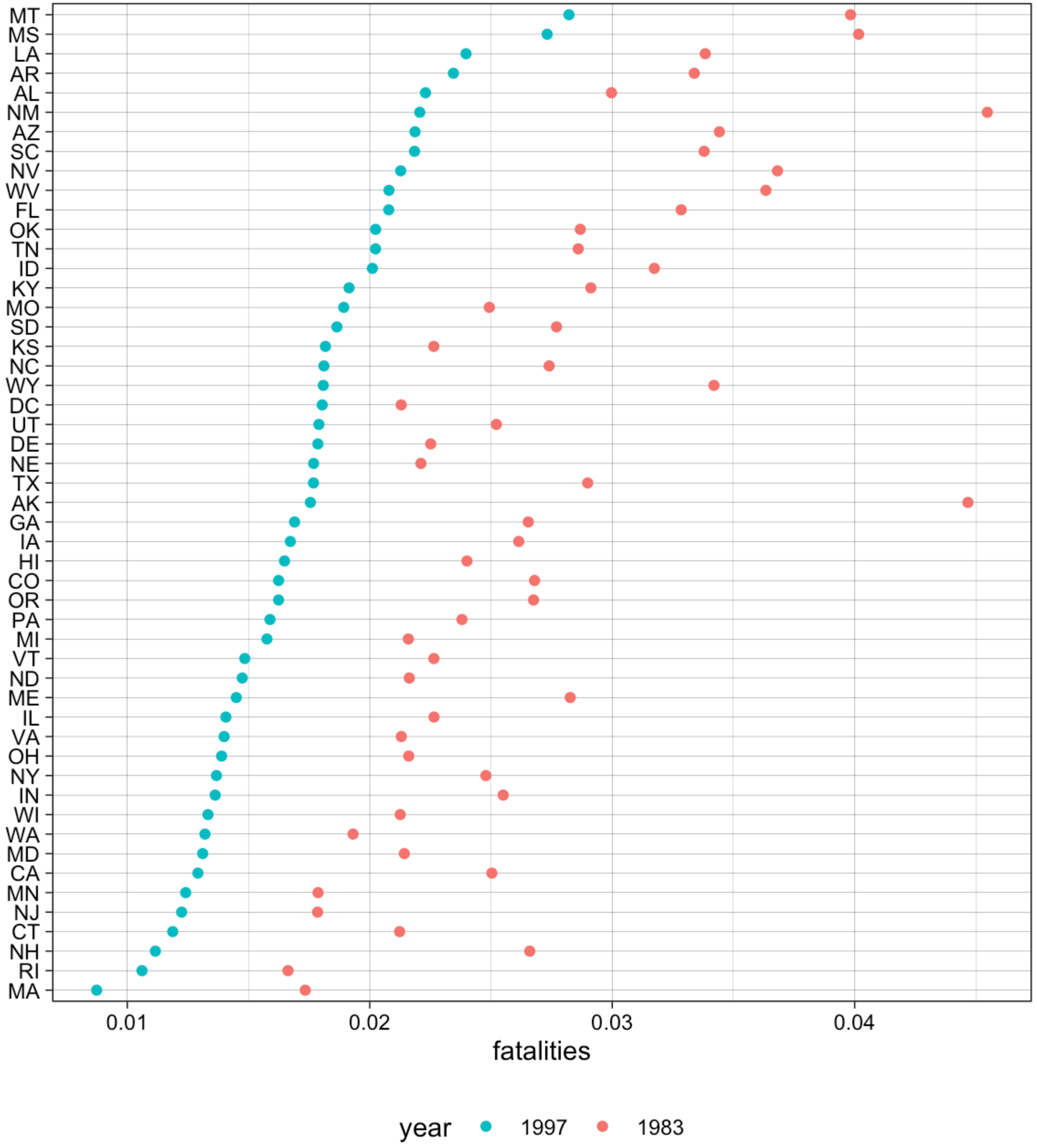
</>



Cleveland dot plot


Africa: GDP per capita, 2012
Equatorial Guinea
Libya
Seychelles
Gabon
Mauritius
South Africa
Botswana
Namibia
Algeria
Angola
Tunisia
Swaziland
Cape Verde
Congo, Rep.
Egypt, Arab Rep.
Morocco
Nigeria
Zambia
Sudan
Ghana
Djibouti
Sao Tome and Principe
Mauritania
Cote d'Ivoire
Cameroon
Kenya
Lesotho
Senegal
Chad
South Sudan
Zimbabwe
Tanzania
Benin
Comoros
Burkina Faso
Rwanda
Uganda
Mali
Sierra Leone
Togo
Mozambique
Guinea-Bissau
Gambia, The
Guinea
Ethiopia
Central African Republic
Madagascar
Liberia
Niger
Congo, Dem. Rep.
Malawi
Burundi
0
5000
10000
15000
20000
GDP

of fatalities per million traffic miles


MT
MS
LA
AR
AL
NM
AZ
SC
NV
WV
FL
OK
TN
ID
KY
MO
SD
KS
NC
WY
DC
UT
DE
NE
TX
AK
GA
IA
HI
CO
OR
PA
MI
VT
ND
ME
IL
VA
OH
NY
IN
WI
WA
MD
CA
MN
NJ
CT
NH
RI
MA
fatalities
year
● 1997
● 1983