Categorical data

Numeric data

```r
library(ade4)
data("clementines")
str(clementines)
```

```
'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

```r
library(fivethirtyeight)
str(food_world_cup[,1:12])
```

```
Classes 'tbl_df', 'tbl' and 'data.frame':   1373 obs. of  12 variables:
 $ respondent_id   : num  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  "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  "West South Central" "West South Central" "Pacific" "New
England" ...
 $ algeria         : chr  "N/A" "N/A" "3" "N/A" ...
 $ argentina       : chr  "3" "N/A" "4" "3" ...
 $ australia       : chr  "5" "3" "N/A" "N/A" ...
```

Warnings

- words are hard to work with!
- not a lot of options (esp. for 1 dimension): bar plot, Cleveland dot plot
- data cleaning takes more time
- main choices: *which* categories to plot, *order* of categories

Types of data

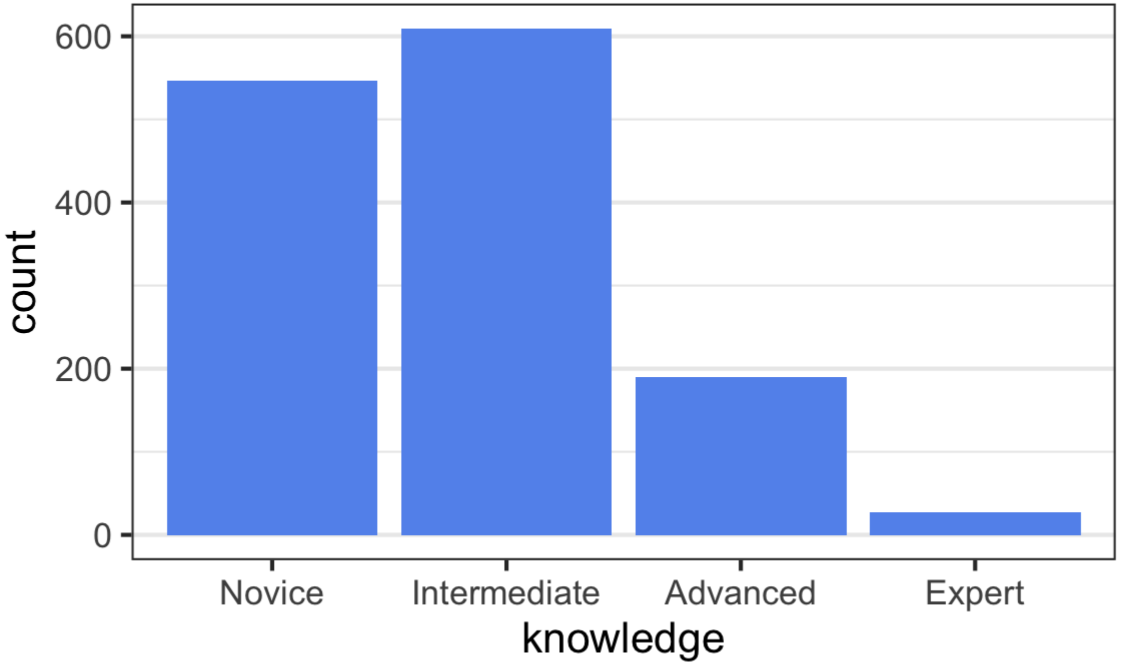
- nominal does not have a fixed category order
- ordinal does have a fixed category order
- (“real”) discrete, small ## of possibilities
- Not always clearcut: nominal vs. ordinal, ordinal vs. discrete, etc.
- Sometimes numbers = nominal, not discrete

Ordinal data

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

```r
library(tidyverse)
ggplot(food_world_cup, aes(knowledge)) +
    geom_bar(fill = "cornflowerblue") +
    ggtitle("Knowledge level of respondents") +
    theme_bw(16) +
    theme(panel.grid.major.x = element_blank())
```

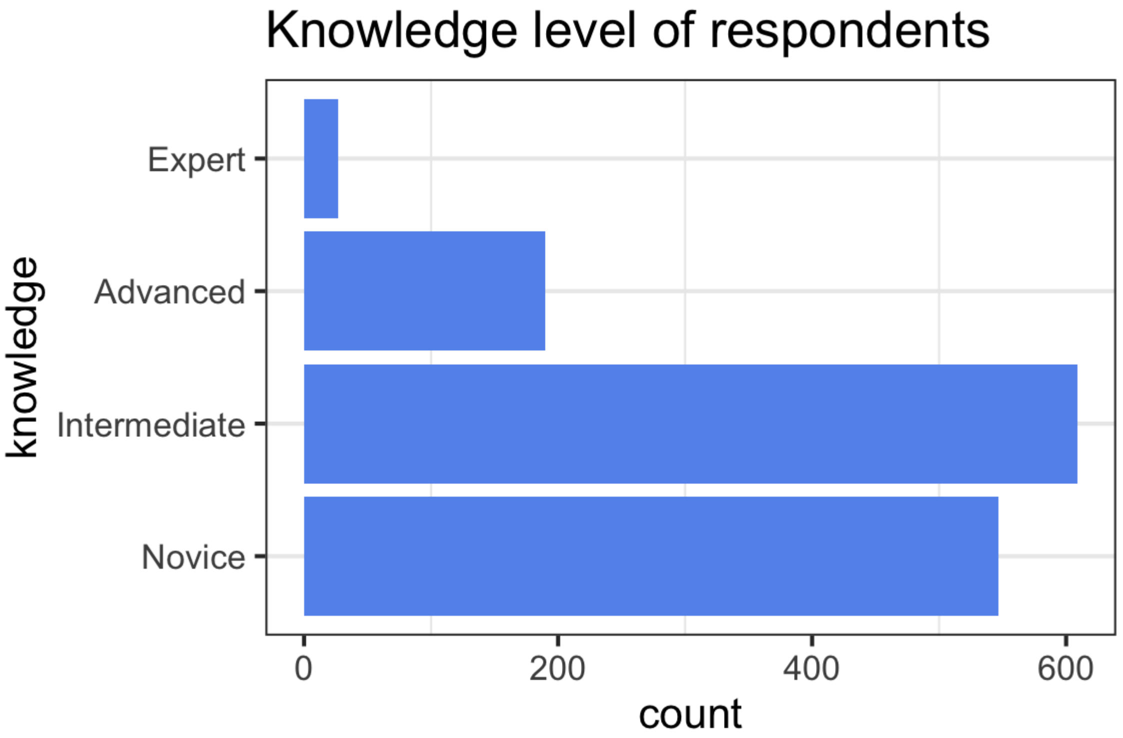
Knowledge level of respondents



Ordinal data, horizontal bars

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

```r
ggplot(food_world_cup, aes(y = knowledge)) +
    geom_bar(fill = "cornflowerblue") +
    ggtitle("Knowledge level of respondents") +
    theme_bw(16) +
    theme(panel.grid.major.x = element_blank())
```



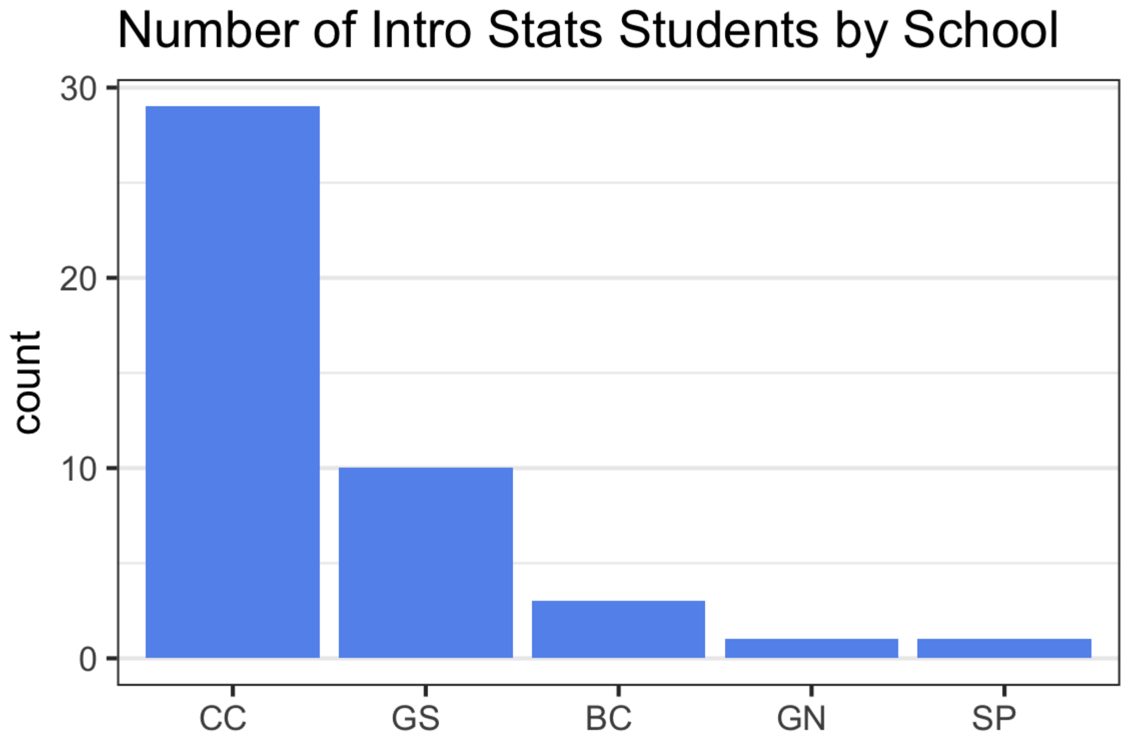
Knowledge level of respondents

Nominal data, vertical bars

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

```r
student <- read.csv("student_data.csv")
## See "School Codes and Descriptions" in SSOL help menu

ggplot(student, aes(x = fct_infreq(School))) +
    geom_bar(fill = "cornflowerblue") +
    labs(title = "Number of Intro Stats Students by School", x = NULL) +
    theme_bw(16) +
    theme(panel.grid.major.x = element_blank())
```

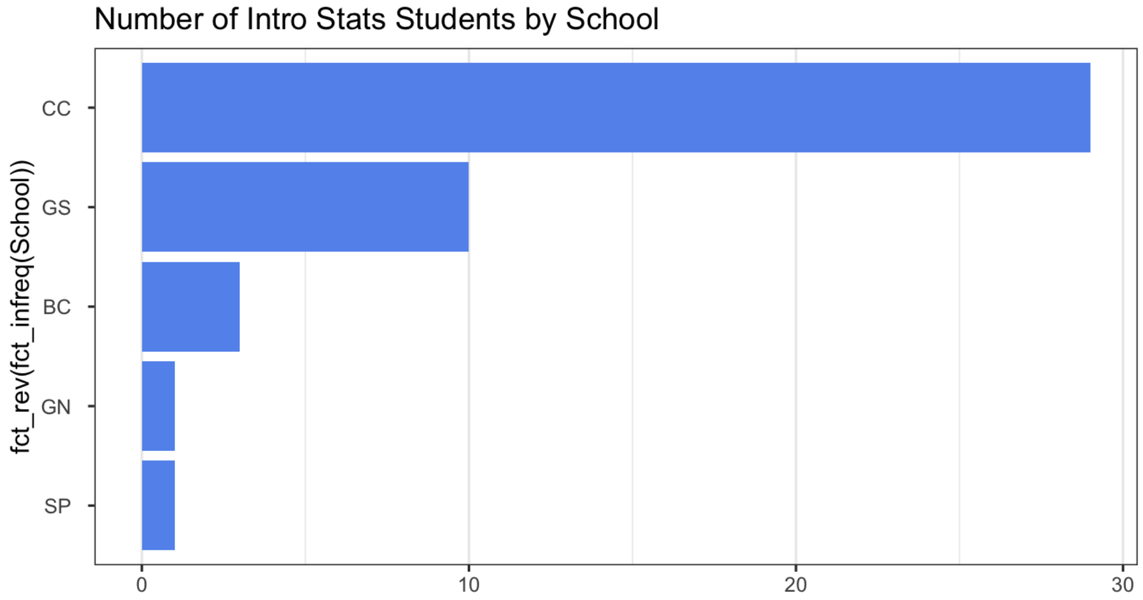

Number of Intro Stats Students by School
count
30
20
10
0
CC
GS
BC
GN
SP

Nominal data, horizontal bars

… or top to bottom

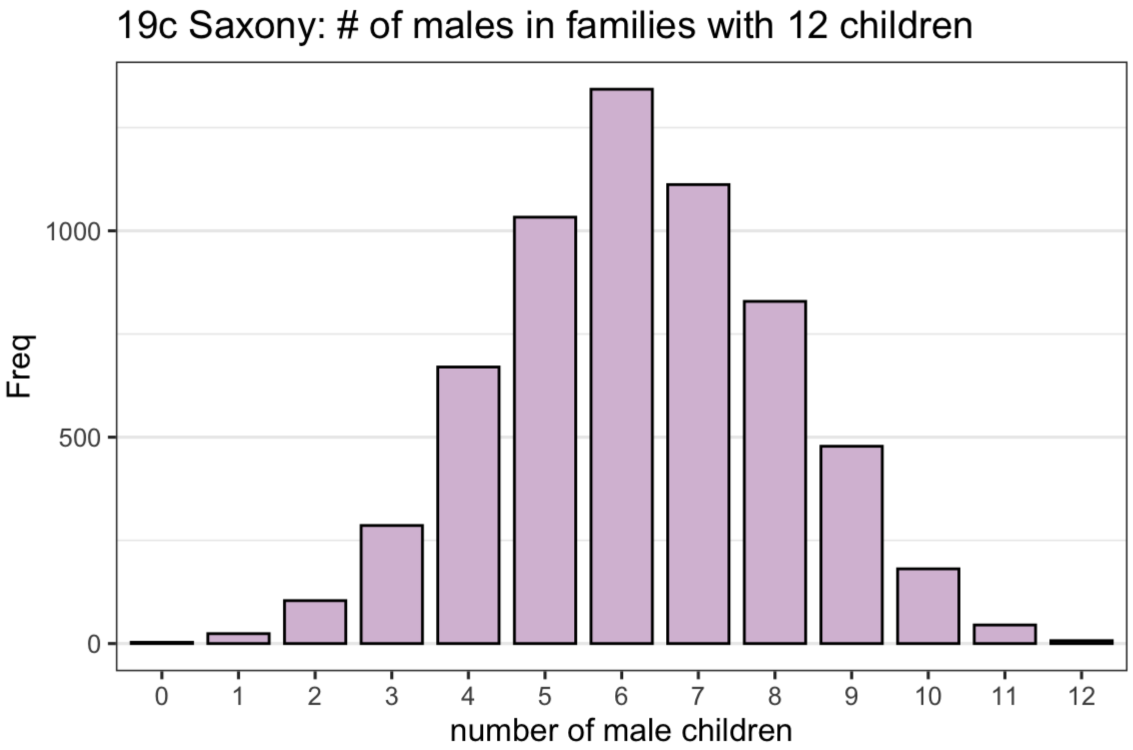
```r
student$School <- fct_recode(student$School,
                             `Barnard College`= "BC",
                             `Columbia College` = "CC",
                             `General Studies Post Bac` = "GN",
                             `General Studies` = "GS",
                             `School of Professional Studies` = "SP")


ggplot(student, aes(y = fct_rev(fct_infreq(School)))) +
  geom_bar(fill = "cornflowerblue") +
  labs(title = "Number of Intro Stats Students by School", x = NULL) +
  theme_bw(12) +
  theme(panel.grid.major.y = element_blank())
```


Number of Intro Stats Students by School
fct_rev(fct_infreq(School))
CC
GS
BC
GN
SP
0
10
20
30

Discrete data

```r
library(vcd)
df <- data.frame(Saxony)
ggplot(df, aes(x = nMales, y = Freq)) +
  geom_col(color = "black", fill = "thistle", width = .8) +
  labs(title = "19c Saxony: # of males in families with 12 children",
       x = "number of male children") +
  theme_bw(12) +
  theme(panel.grid.major.x = element_blank())
```


19c Saxony: # of males in families with 12 children
Freq
0
500
1000
0
1
2
3
4
5
6
7
8
9
10
11
12
number of male children

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

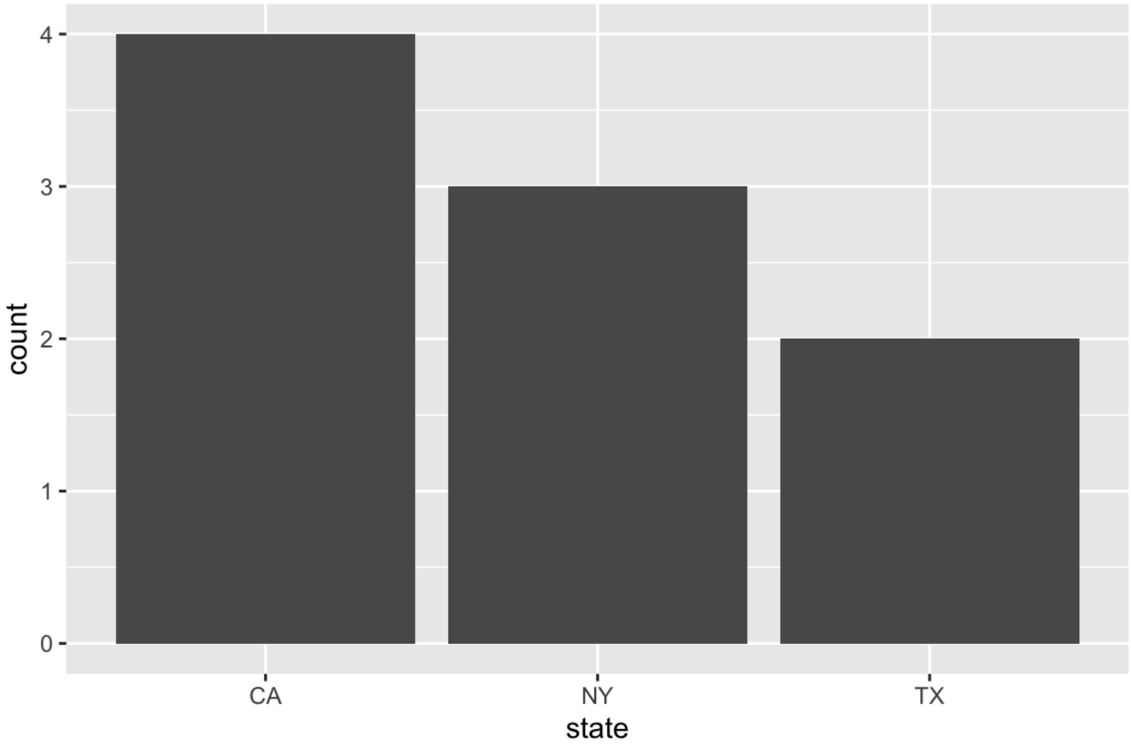
Look at the data

```r
df_binned <- data.frame(state = c("CA", "NY", "TX"),
                        count = c(4, 3, 2))
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

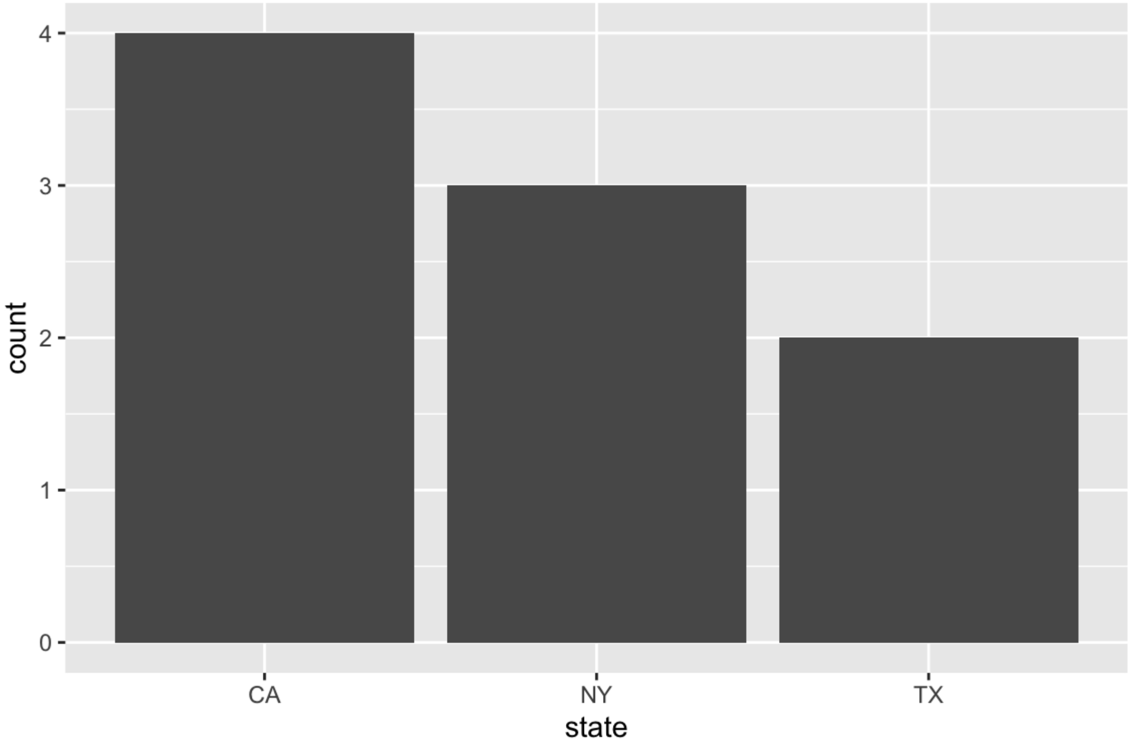
Look at the data

```r
df_unbinned <- data.frame(state = c("NY", "CA", "TX", "NY", "CA", "CA", "TX", "CA",
str(df_unbinned)
```

```
'data.frame':	9 obs. of  1 variable:
 $ state: chr  "NY" "CA" "TX" "NY" ...
```

Bar chart with unbinned data

```r
ggplot(df_unbinned, aes(x = state)) +
  geom_bar()
```

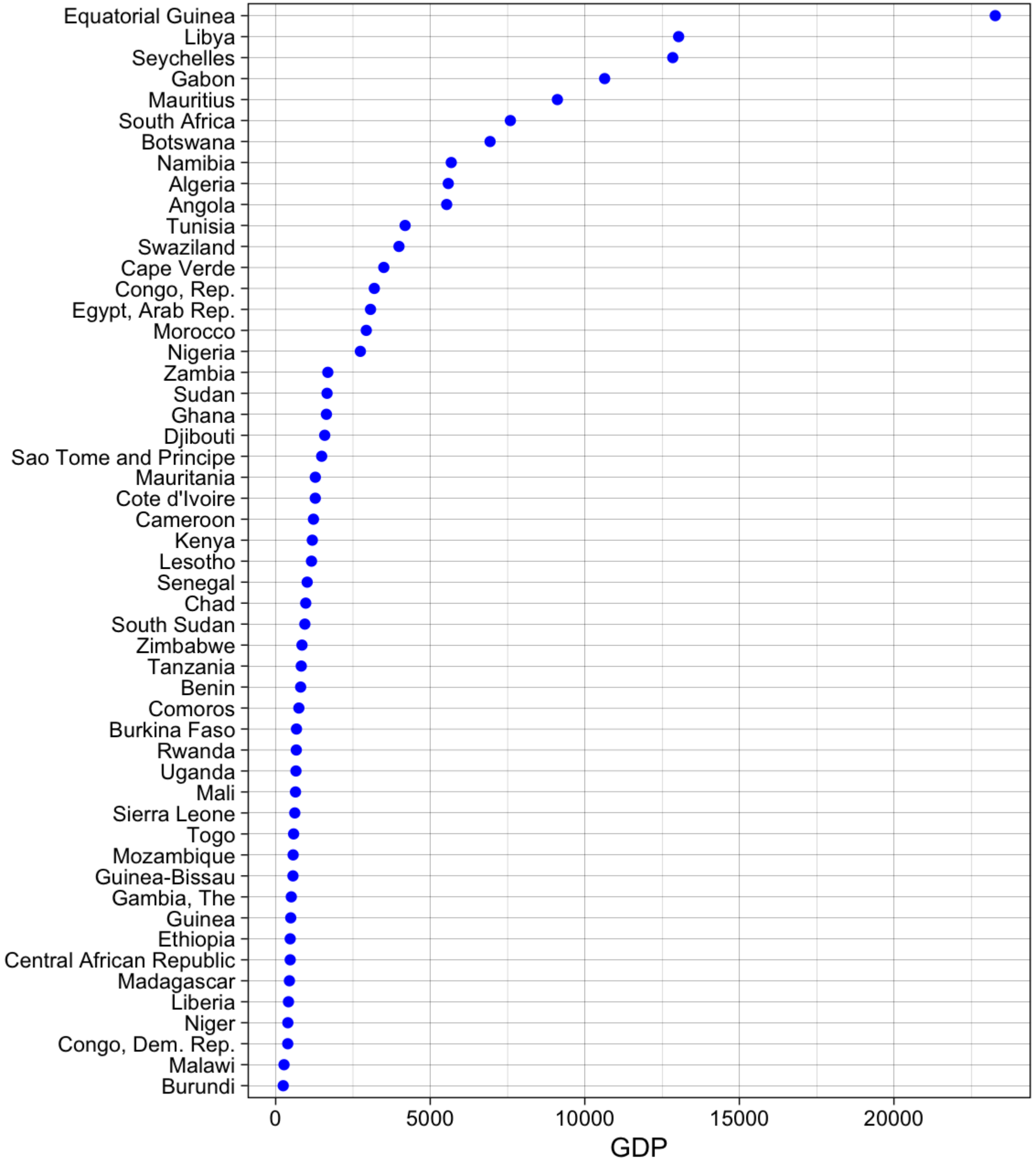


Cleveland dot plot

Cleveland dot plot

```r
world <- read_csv("countries2012.csv")
africa <- world |>
  filter(CONTINENT == "Africa")
ggplot(africa, aes(x = GDP, y = fct_reorder(COUNTRY, GDP))) +
  geom_point(color = "blue") +
  labs(title = "Africa: GDP per capita, 2012", y = NULL) +
  theme_linedraw() ## works well for dotplots
```

Africa: GDP per capita, 2012



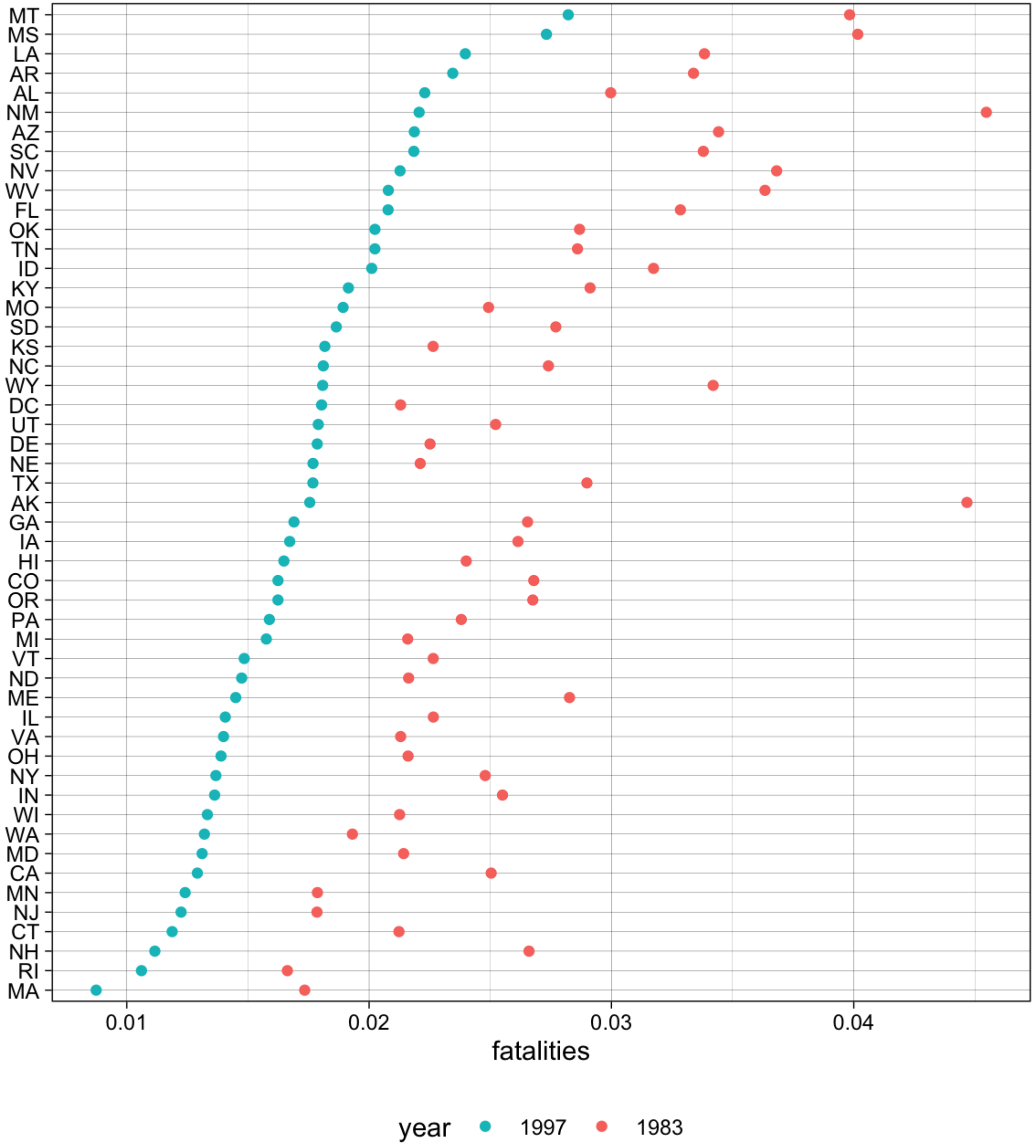
Cleveland dot plot with multiple dots

Sorted by 1997 fatality rate

```r
library(AER)
data("USSeatBelts")
belts <- USSeatBelts |>
  filter(year %in% c(1983, 1997)) |>
  select(state, year, fatalities)

## `fct_reorder2` --> double sort: year, then fatalities
ggplot(belts, aes(x = fatalities,
                  y = fct_reorder2(state, year == 1997, fatalities, .desc = FALSE),
                  color = year)) +
  geom_point() +
  labs(title = "# of fatalities per million traffic miles", y = NULL) +
  guides(color = guide_legend(reverse=TRUE)) +
  theme_linedraw() +
  theme(legend.position = "bottom")
```

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