

Categorical data

`slides/05_categorical.pdf`

Numeric data

```
1 library(ade4)
2 data("clementines")
3 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 22 71 11 106 0 47 1
```

Categorical data

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

```
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"
```

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

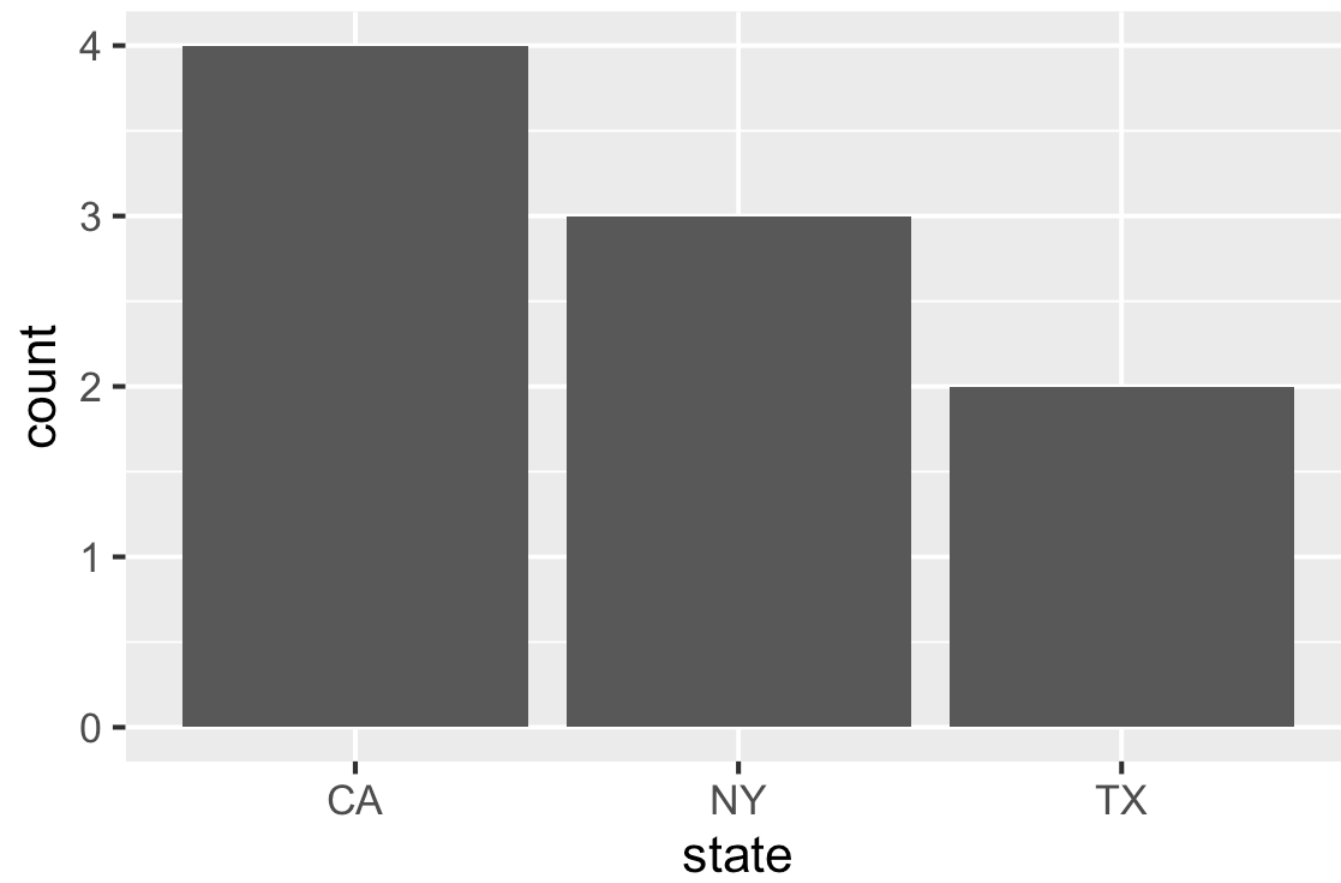
```
1 df_binned
```

```
state count
```

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

Bar chart with binned data

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

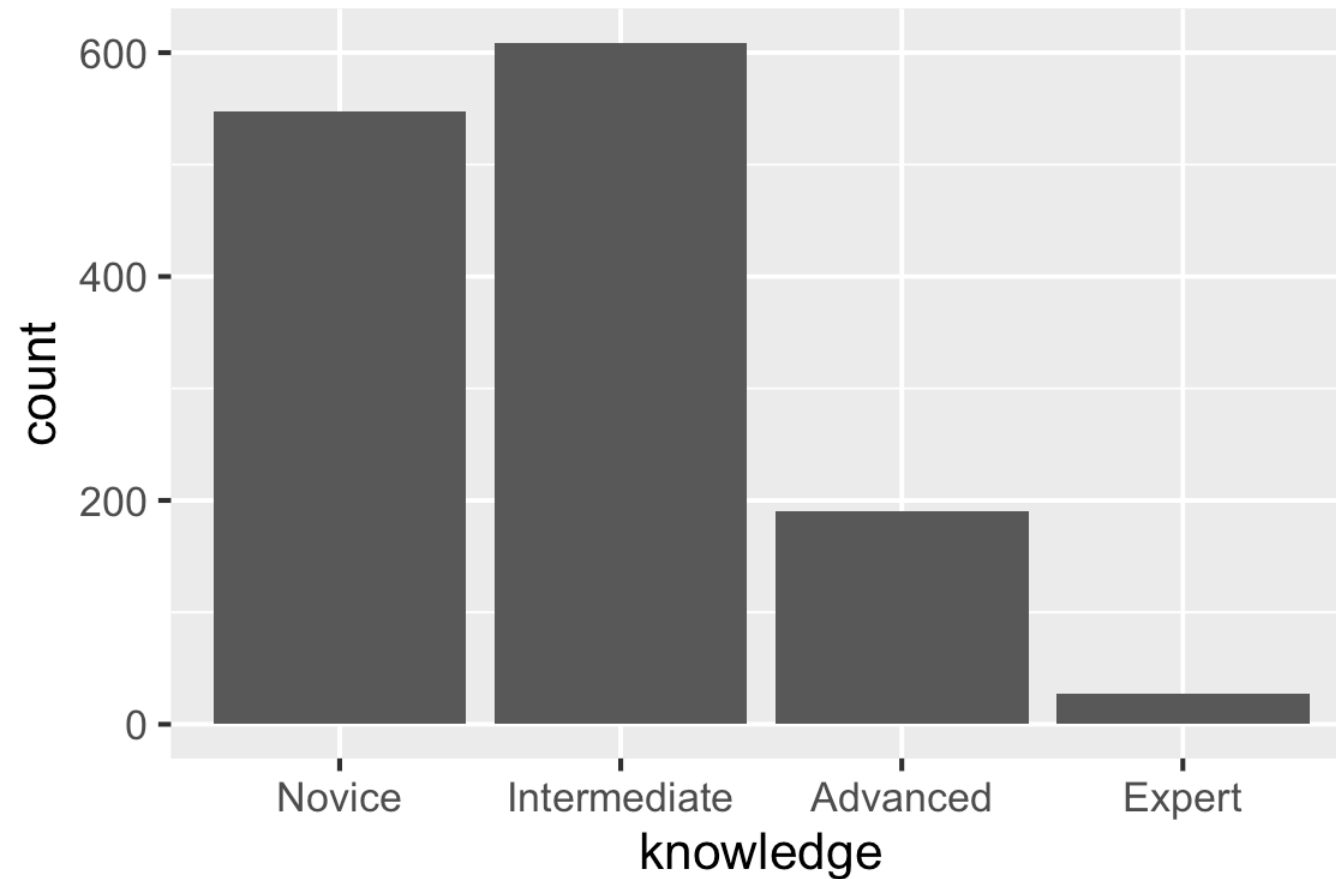


geom_bar()

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

Bar chart with unbinned data

```
1 ggplot(food_world_cup, aes(x = knowledge)) +  
2   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

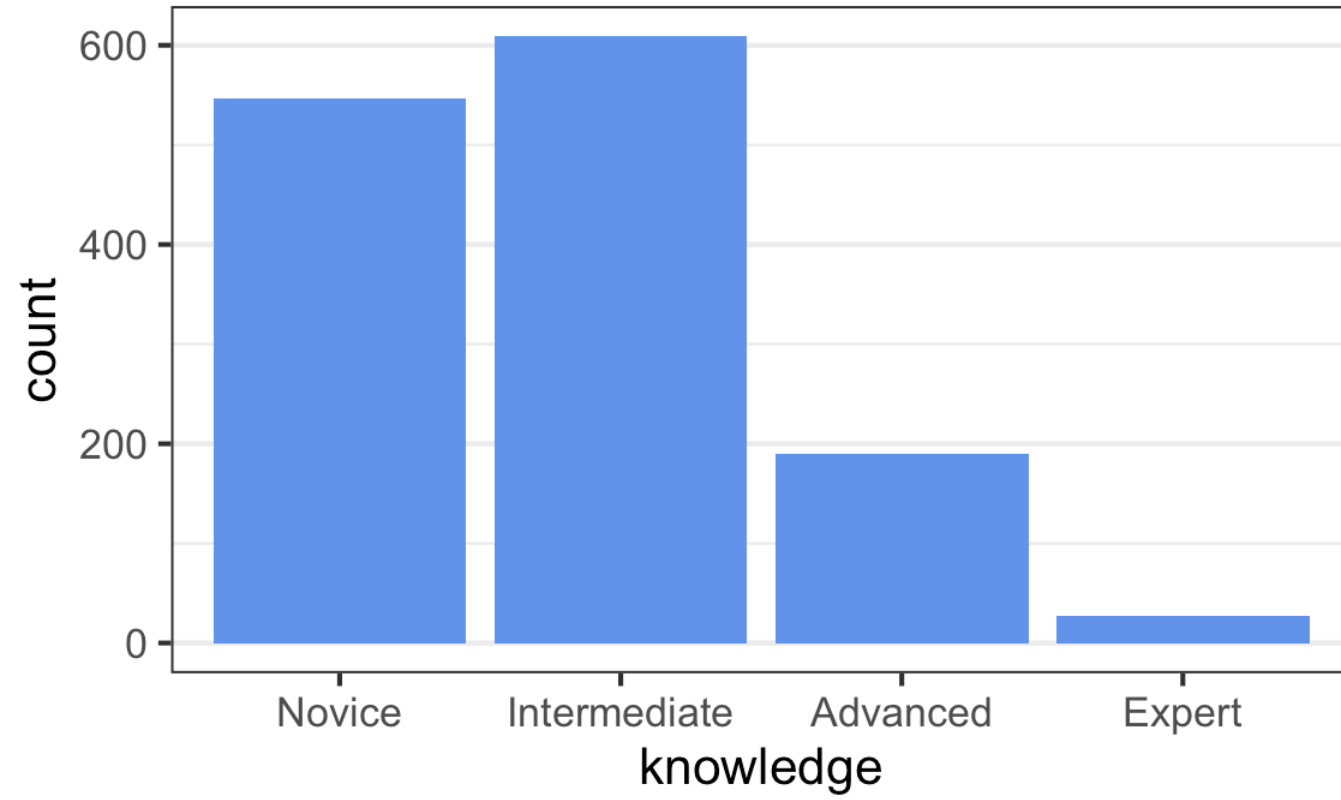
- nominal does not have a fixed category order
- ordinal does have a fixed category order
- (“real”) discrete, small number of possibilities
- Not always clearcut: nominal vs. ordinal, ordinal vs. discrete, etc.
- Sometimes numbers = nominal, not discrete

Ordinal data

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

```
1 ggplot(food_world_cup, aes(knowledge)) +  
2   geom_bar(fill = "cornflowerblue") +  
3   labs(title = "Knowledge level of respondents") +  
4   theme_bw(16) +  
5   theme(panel.grid.major.x = element_blank())
```

Knowledge level of respondents

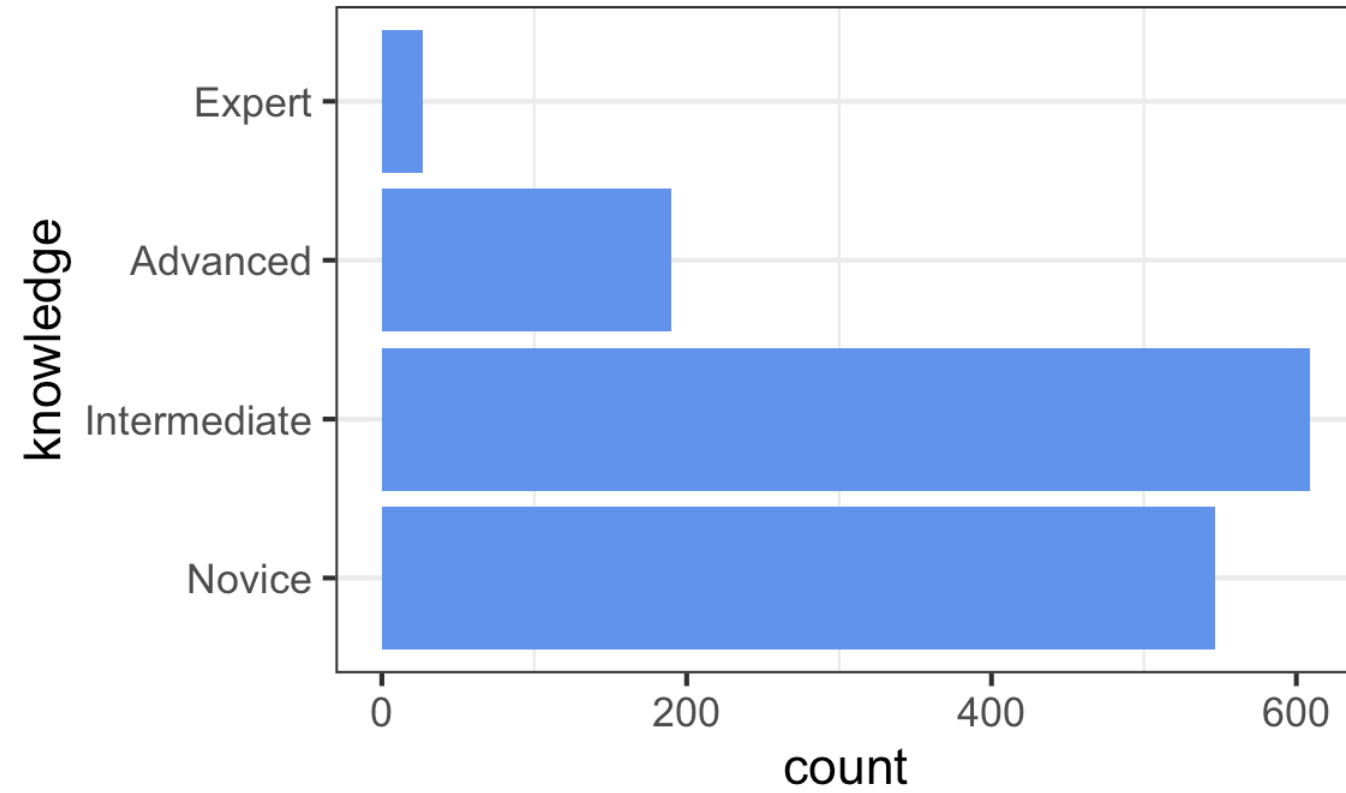


Ordinal data, horizontal bars

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

```
1 ggplot(food_world_cup, aes(y = knowledge)) +  
2   geom_bar(fill = "cornflowerblue") +  
3   labs(title = "Knowledge level of respondents") +  
4   theme_bw(16) +  
5   theme(panel.grid.major.x = element_blank())
```


Knowledge level of respondents

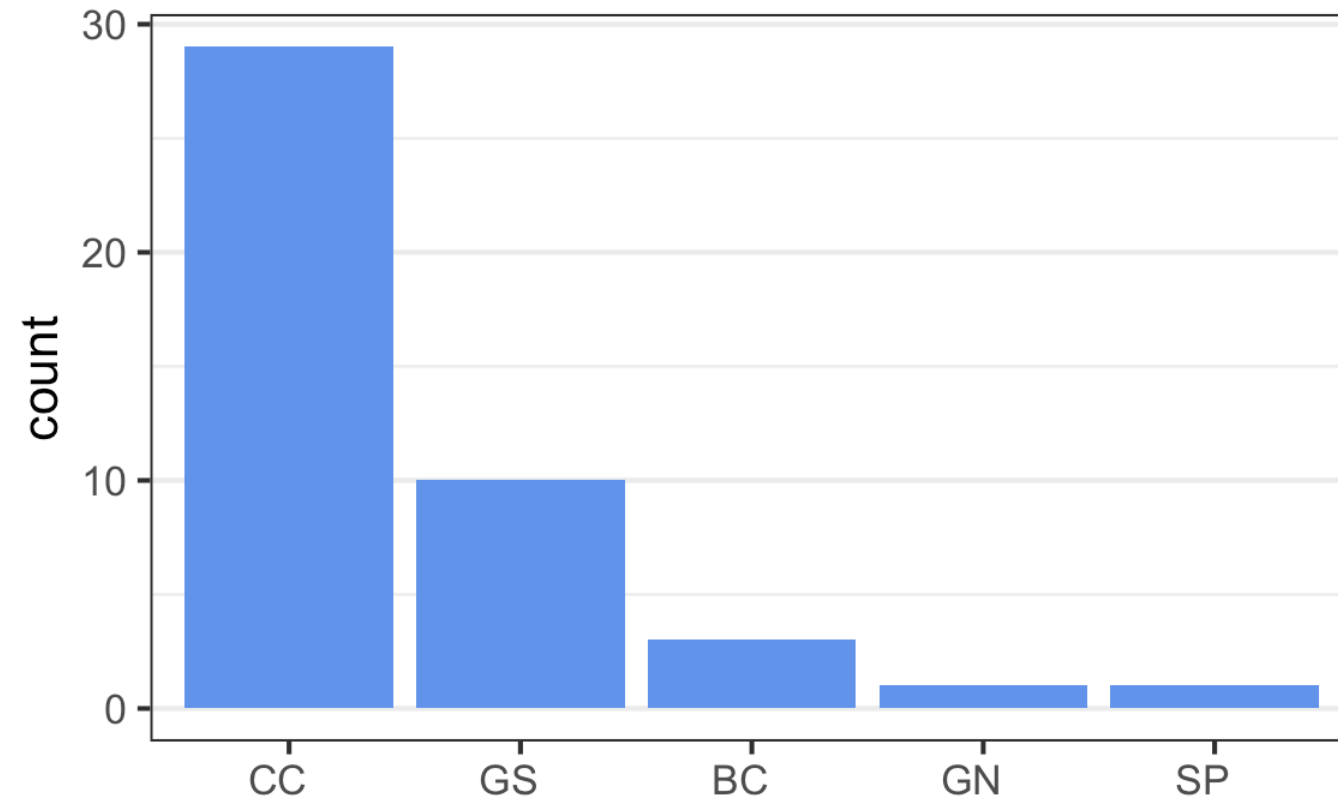


Nominal data, vertical bars

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

```
1 student <- read.csv("student_data.csv")
2 ## See "School Codes and Descriptions" in SSOL help menu
3
4 ggplot(student, aes(x = fct_infreq(School))) +
5   geom_bar(fill = "cornflowerblue") +
6   labs(title = "Number of Intro Stats Students by School", x = NULL) +
7   theme_bw(16) +
8   theme(panel.grid.major.x = element_blank())
```

Number of Intro Stats Students by School

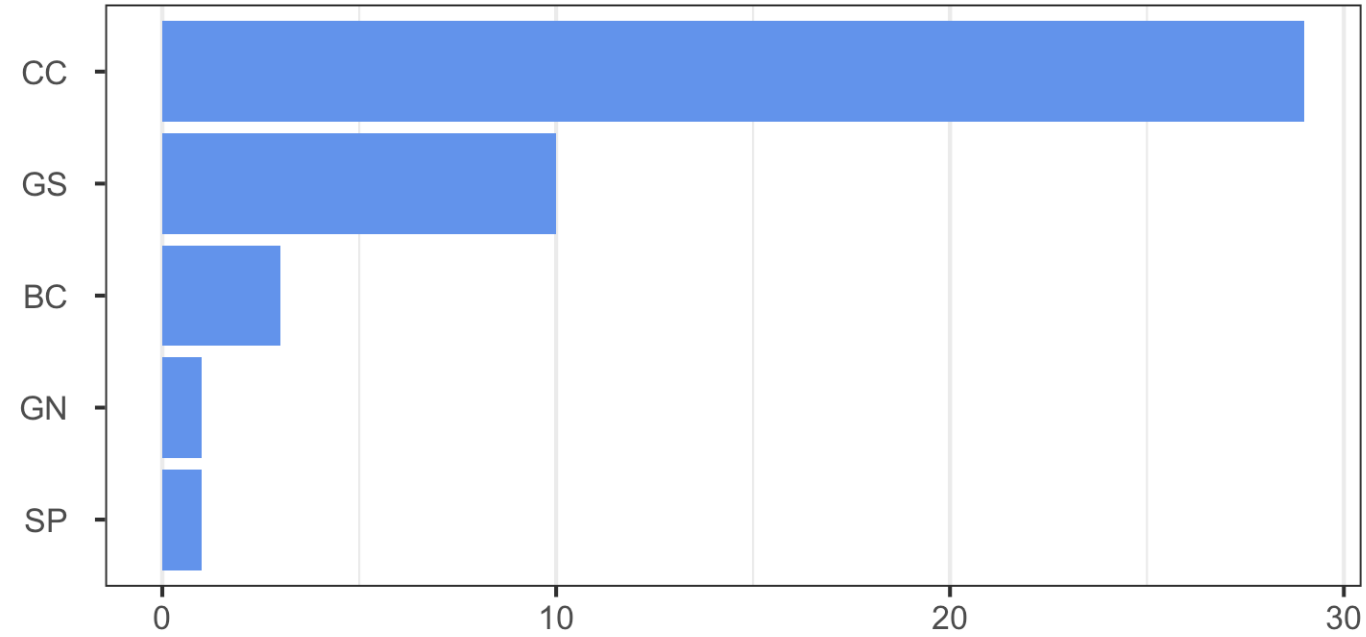


Nominal data, horizontal bars

... or top to bottom

```
1 student$School <- fct_recode(student$School,
2                               `Barnard College` = "BC",
3                               `Columbia College` = "CC",
4                               `General Studies Post Bac` = "GN",
5                               `General Studies` = "GS",
6                               `School of Professional Studies` = "SP")
7
8
9 ggplot(student, aes(y = fct_rev(fct_infreq(School)))) +
10   geom_bar(fill = "cornflowerblue") +
11   labs(title = "Number of Intro Stats Students by School", x = NULL, y = NULL) +
12   theme_bw(16) +
13   theme(panel.grid.major.y = element_blank())
```

Number of Intro Stats Students by School

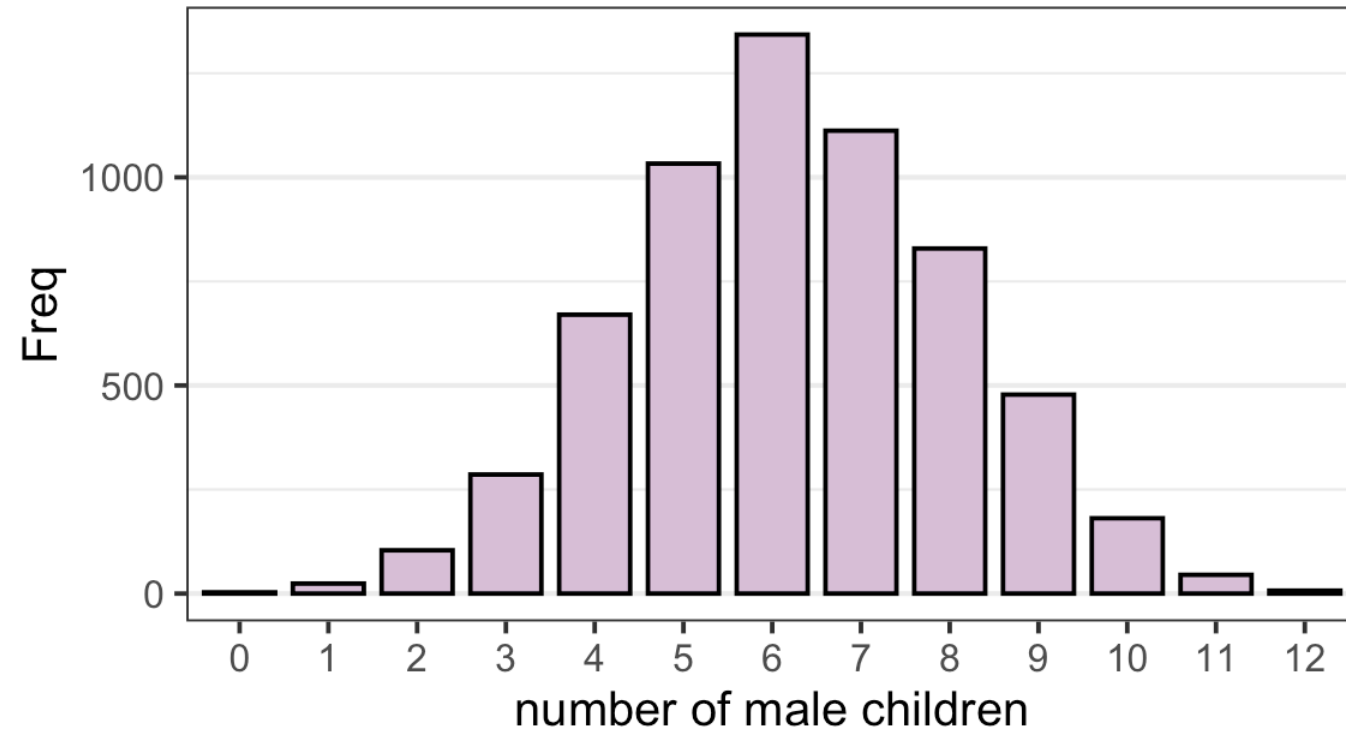


Discrete data

```
1 library(vcd)
2 df <- data.frame(Saxony)
3 ggplot(df, aes(x = nMales, y = Freq)) +
4   geom_col(color = "black", fill = "thistle", width = .8) +
5   labs(title = "Number of males in families with 12 children", subtitle = "19th ce
6         x = "number of male children") +
7   theme_bw(15) +
8   theme(panel.grid.major.x = element_blank())
```

Number of males in families with 12 children

19th century Saxony

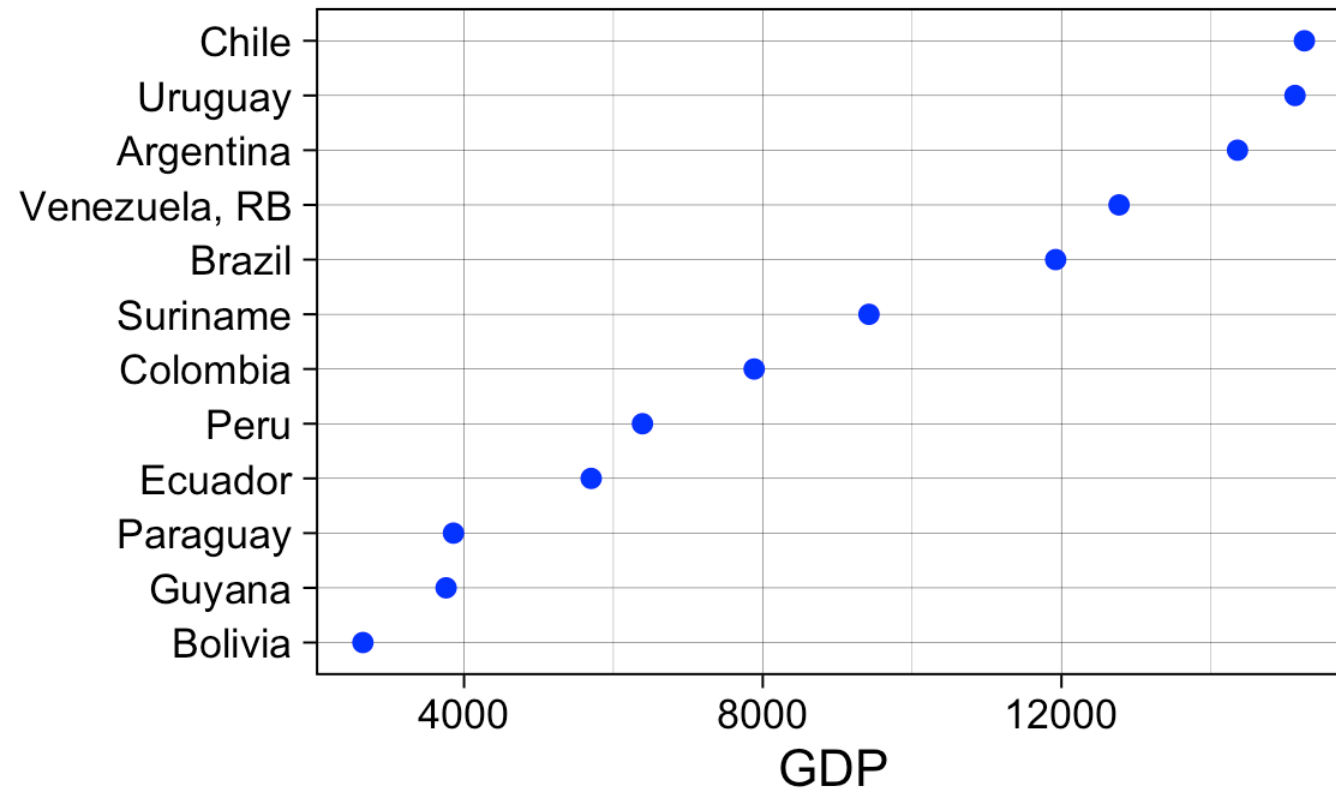


Cleveland dot plot

Cleveland dot plot

```
1 world <- read_csv("countries2012.csv")
2 sa <- world |>
3   filter(CONTINENT == "South America")
4 ggplot(sa, aes(x = GDP, y = fct_reorder(COUNTRY, GDP))) +
5   geom_point(color = "blue") +
6   labs(title = "South America: GDP per capita, 2012", y = NULL) +
7   theme_linedraw(16) ## works well for dotplots
```

South America: GDP per capita, 2012



Cleveland dot plot with multiple dots

```
1 library(AER)
2 data("USSeatBelts")
3 belts <- USSeatBelts |>
4   filter(year %in% c(1983, 1997)) |>
5   select(state, year, fatalities)
6
7 ## `fct_reorder2` --double sort: year, then fatalities
8 ggplot(belts, aes(x = fatalities,
9                   y = fct_reorder2(state, year == 1997, fatalities, .desc = FALSE)
10                  color = year)) +
11   geom_point() +
12   labs(title = "# of fatalities per million traffic miles", y = NULL) +
13   guides(color = guide_legend(reverse=TRUE)) +
14   theme_linedraw() +
15   theme(legend.position = "top")
```

of fatalities per million traffic miles

