DATA VISUALIZATION WITH GGPLOT2

Bar Plots

Chapter Content

- Common pitfalls
- Best way to represent data

Bar plot

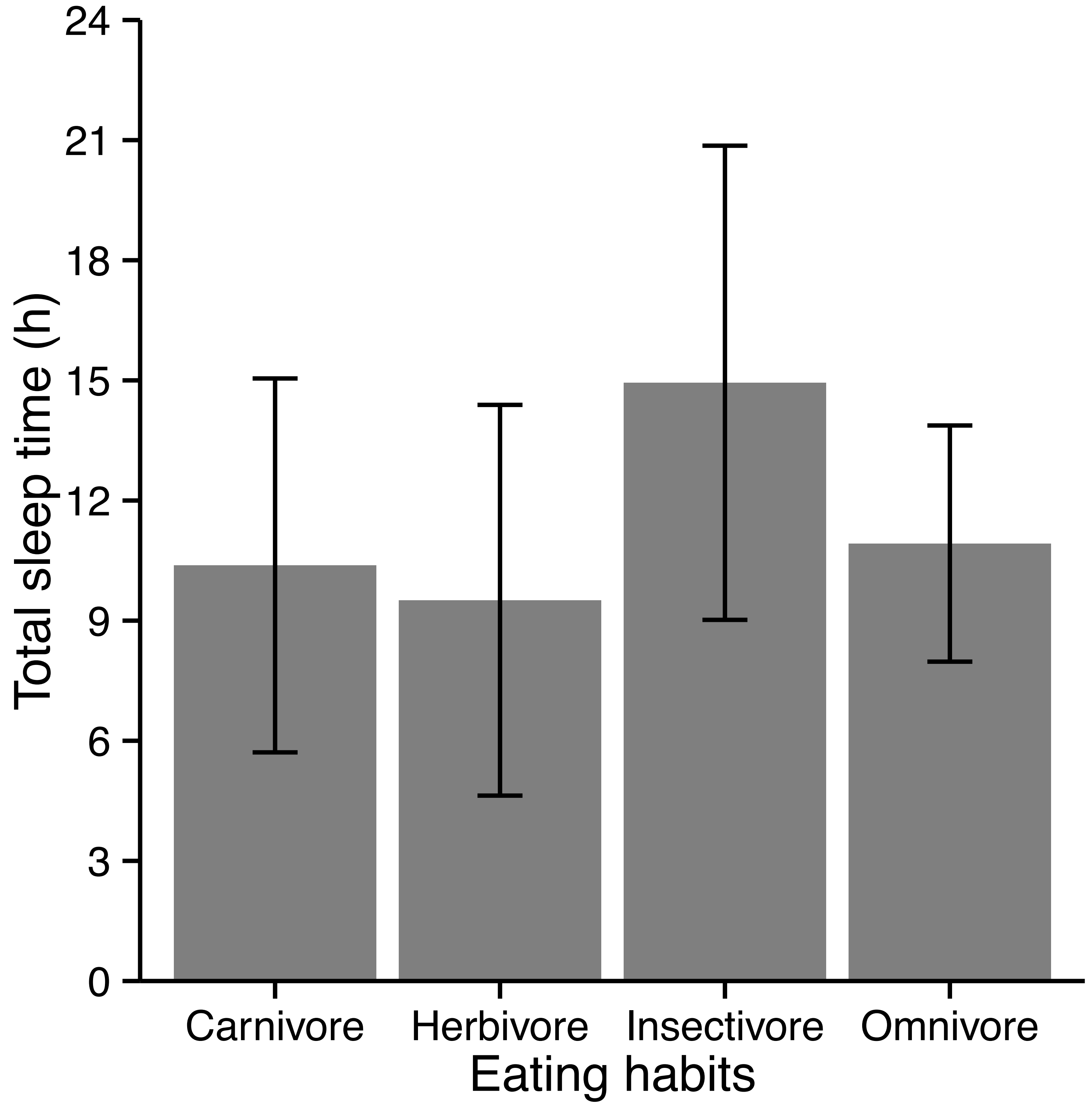
- Two types
  - Absolute values such as count per bin of a bar
  - Distribution

Mammalian sleep

```
> str(sleep)
'data.frame':76 obs. of  3 variables:
 $ vore : Factor w/ 4 levels "Carnivore","Herbivore",..: 1 4 2 ...
 $ total: num  12.1 17 14.4 14.9 4 14.4 8.7 10.1 3 5.3 ...
 $ rem  : num  NA 1.8 2.4 2.3 0.7 2.2 1.4 2.9 NA 0.6 ...
```

Dynamite plot

```
> d <- ggplot(sleep, aes(vore, total)) +
    scale_y_continuous("Total sleep time (h)",
                       limits = c(0, 24),
                       breaks = seq(0, 24, 3),
                       expand = c(0, 0)) +
    scale_x_discrete("Eating habits") +
    theme_classic()
> d +
    stat_summary(fun.y = mean, geom = "bar",
                 fill = "grey50") +
    stat_summary(fun.data = mean_sdl, mult = 1,
                 geom = "errorbar", width = 0.2)
```

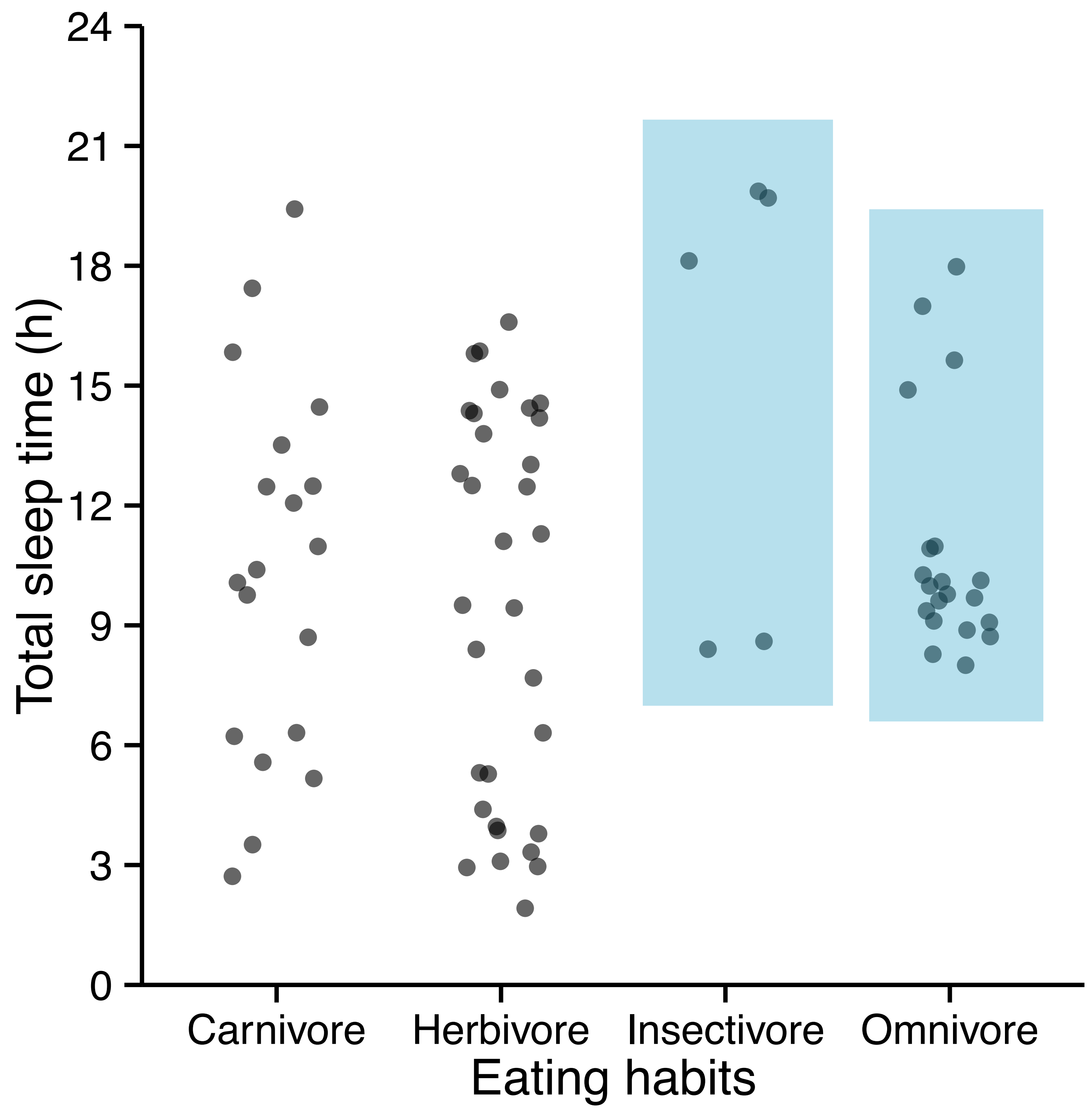


mean and SE suggest that data is normally distributed - we cannot know that. x scale suggests that there might be mammals who sleep 0 hours (impression there is data where there is none)
we don't know how many observations in each category - so we must add this.
no visuals above the mean!

Individual data points

we can see how data look like - patterns:
- insectivores - little amt of data
- omnivores appear positively skewed

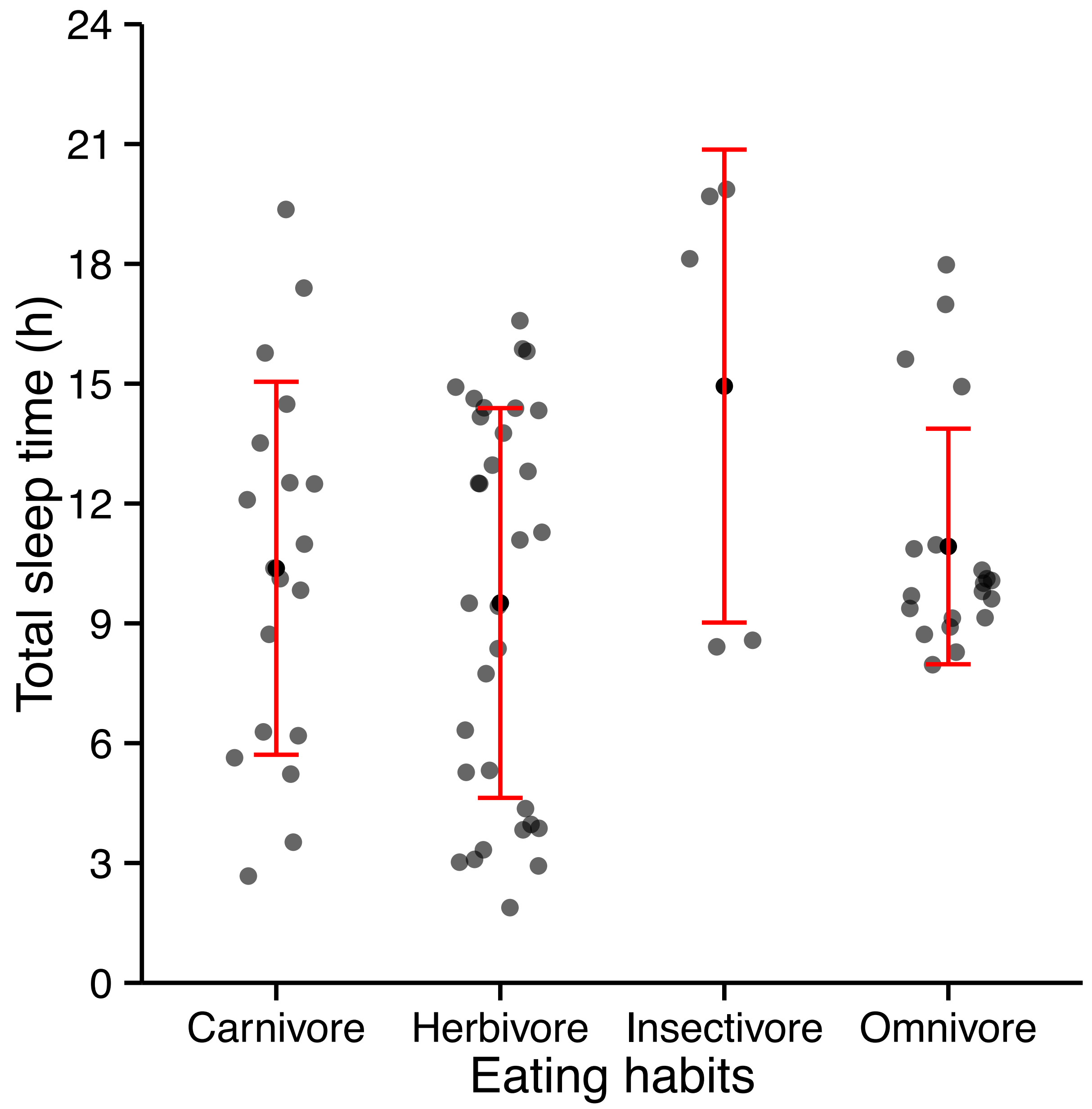
```
> d +
    geom_point(alpha = 0.6, position = position_jitter(width = 0.2))
```



we can add summary statistics to that with
- geom_errorbar()
- geom_pointrange()

errorbar

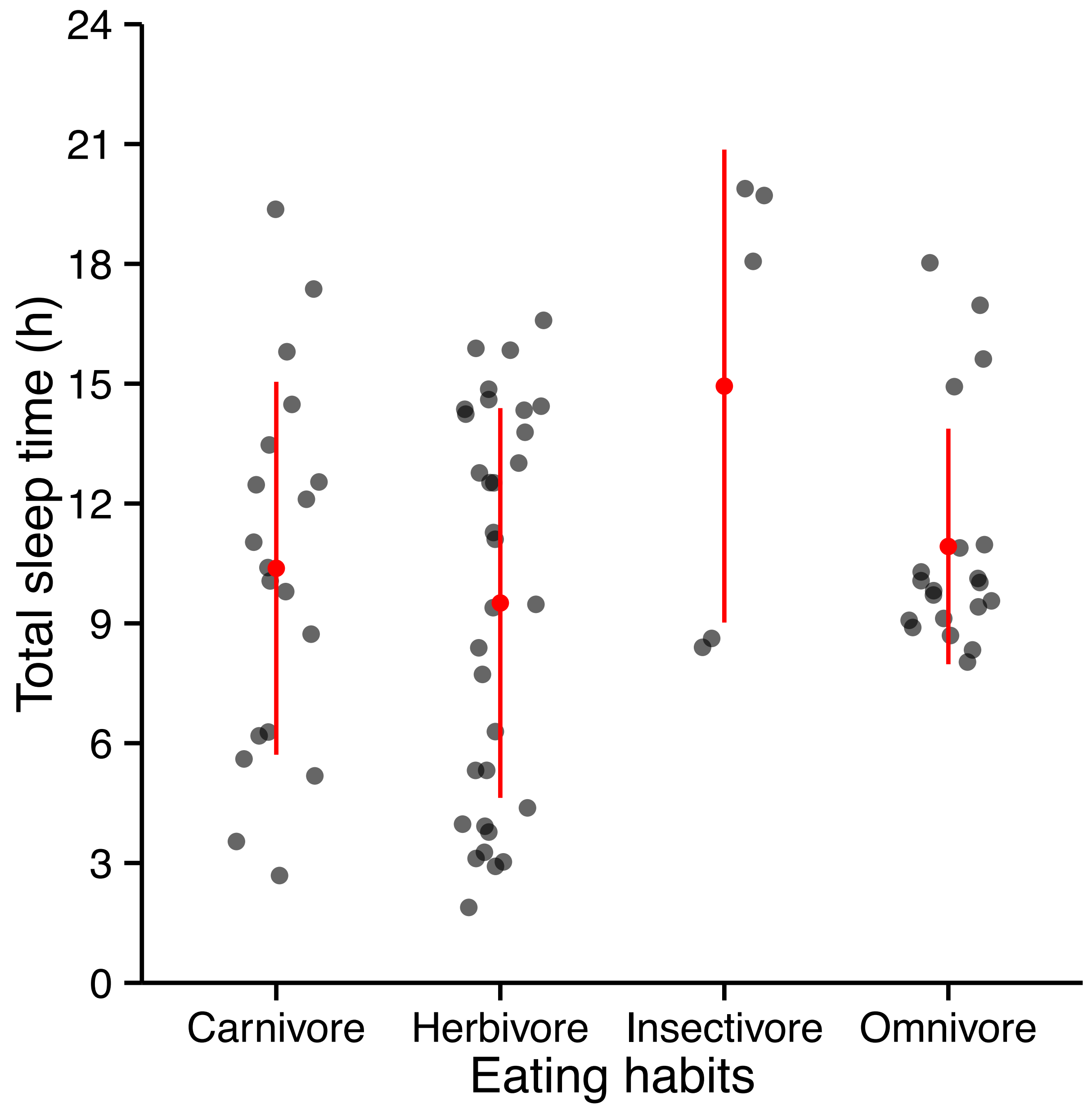
```
> d +
    geom_point(alpha = 0.6, position = position_jitter(width = 0.2)) +
    stat_summary(fun.y = mean, geom = "point", fill = "red") +
    stat_summary(fun.data = mean_sdl, mult = 1, geom = "errorbar",
                 width = 0.2, col = "red")
```



errorbars with points is much cleaner represenatation of data
the bars are simply not necessary

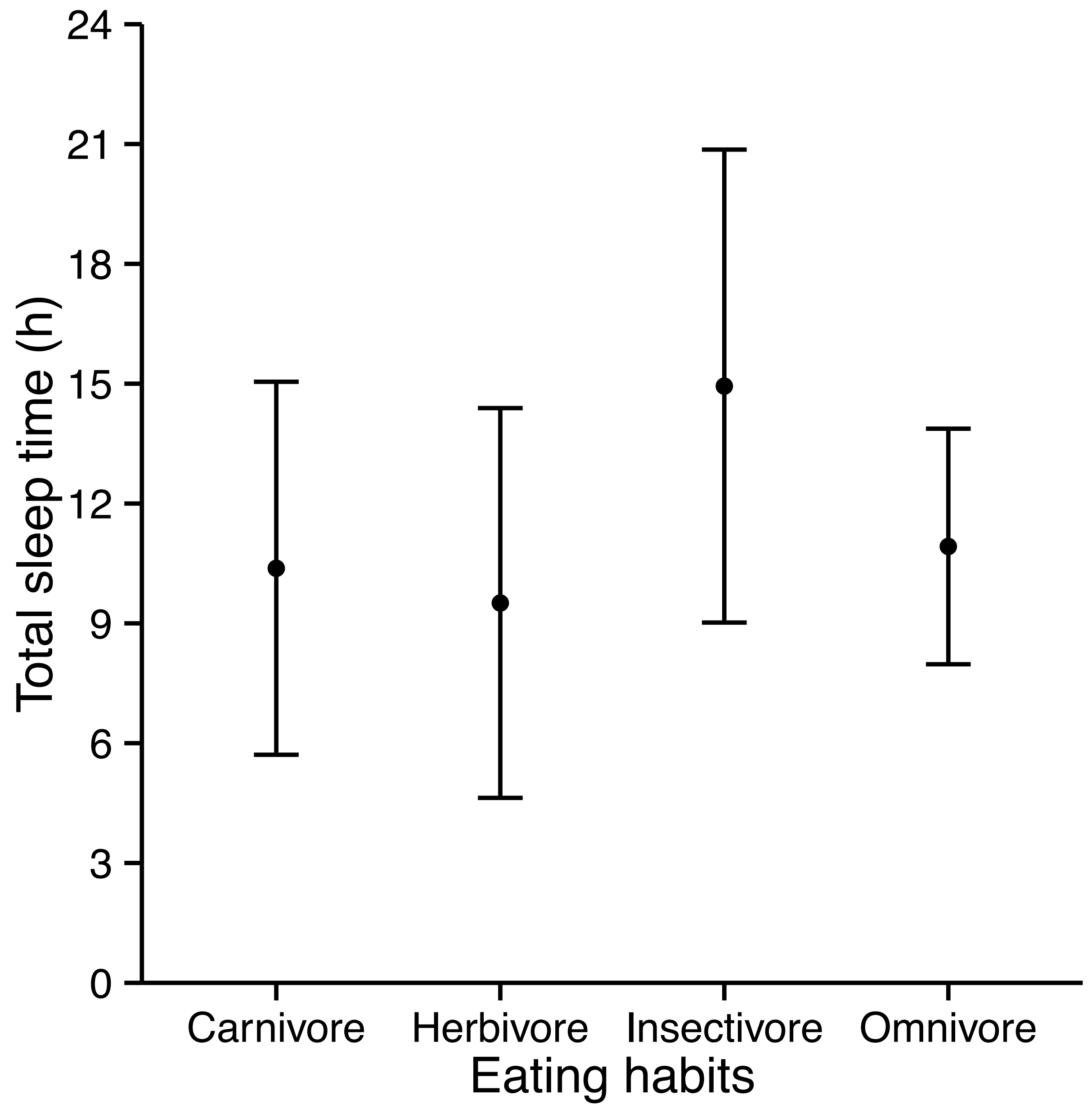
pointrange

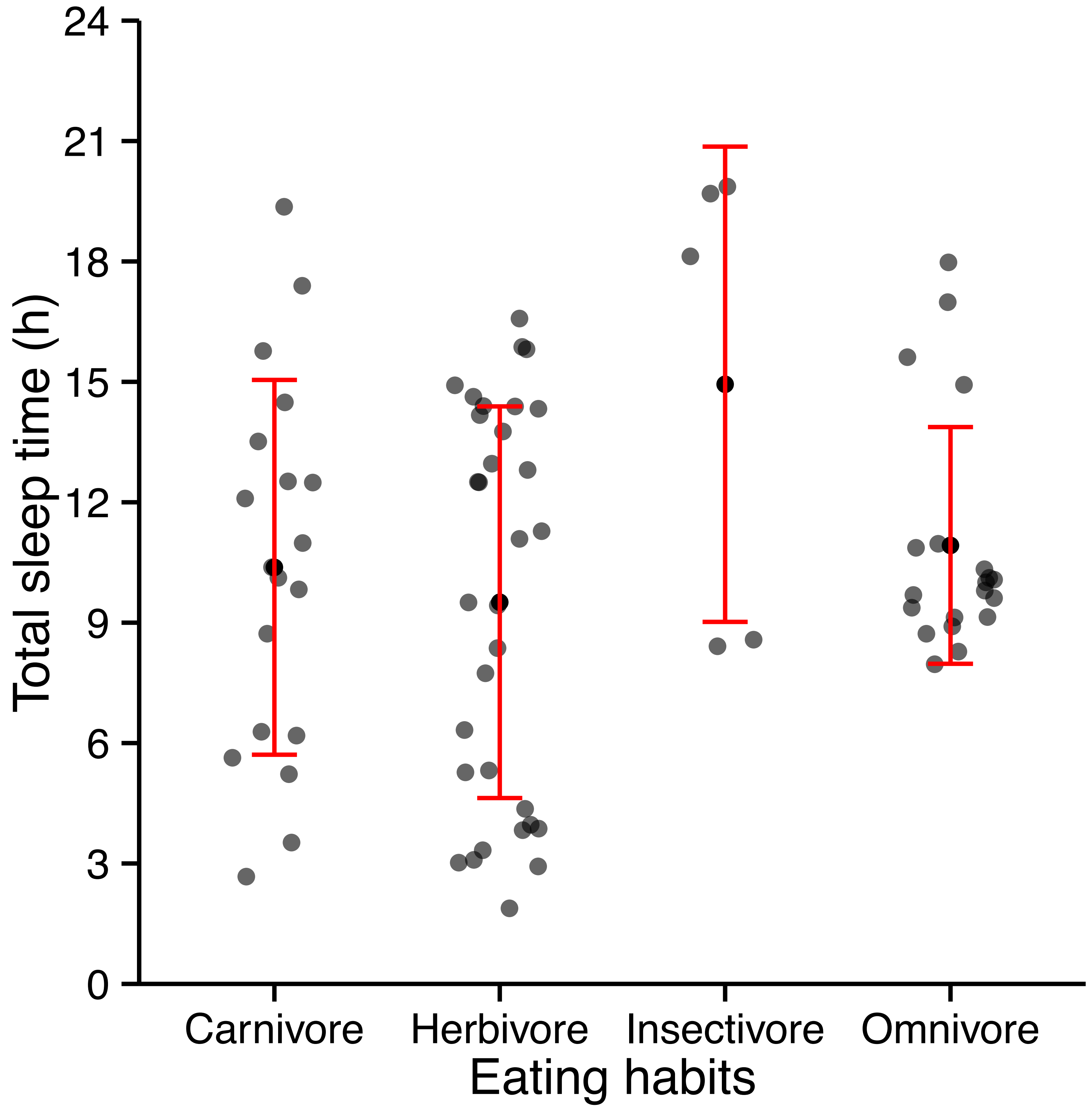
```
> d +
    geom_point(alpha = 0.6, position = position_jitter(width = 0.2)) +
    stat_summary(fun.data = mean_sdl, mult = 1, width = 0.2, col = "red")
```



Without data points

```
> d +
    stat_summary(fun.y = mean, geom = "point") +
    stat_summary(fun.data = mean_sdl, mult = 1,
                 geom = "errorbar", width = 0.2)
```

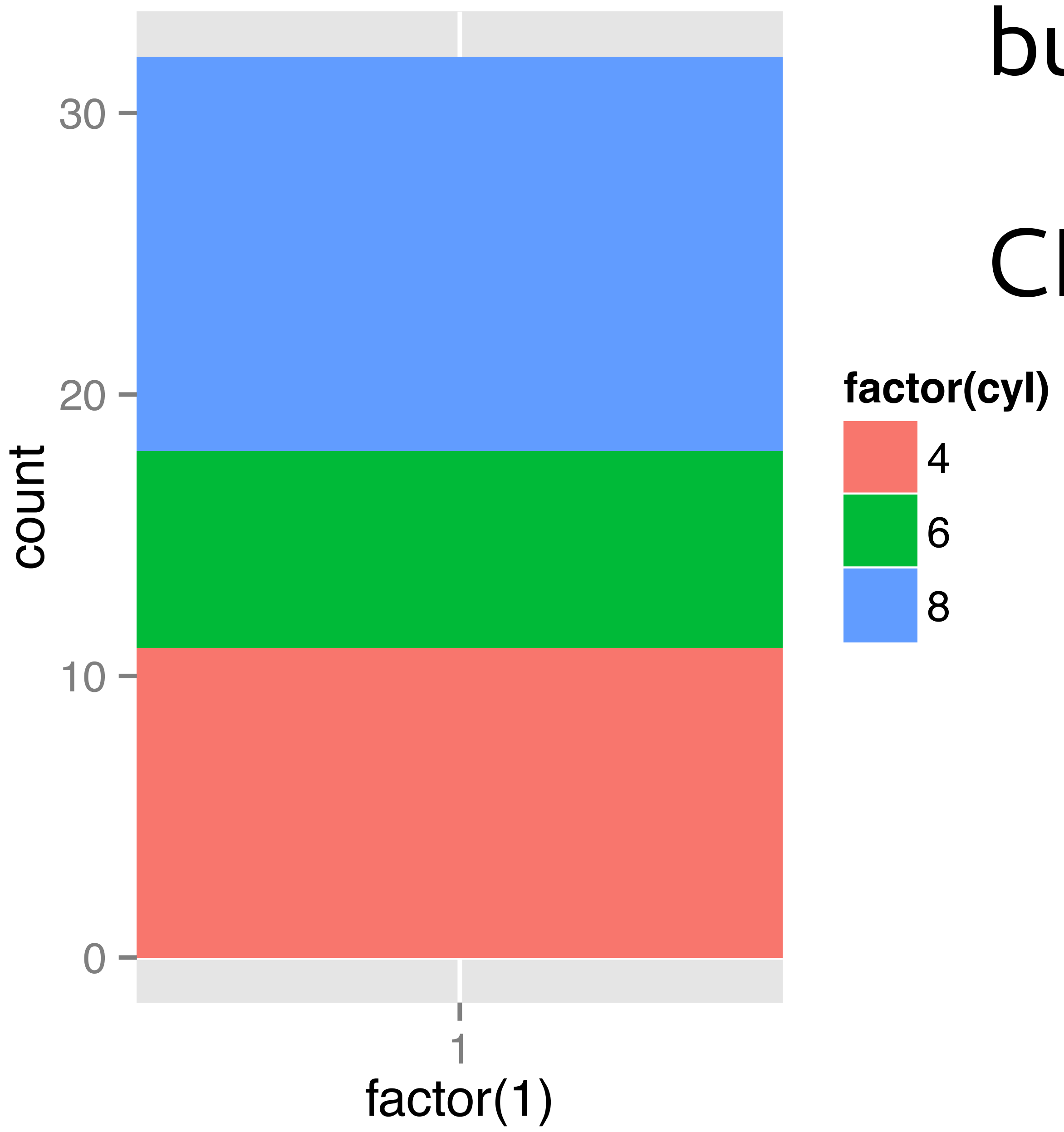



Total sleep time (h)
0
3
6
9
12
15
18
21
24
Carnivore
Herbivore
Insectivore
Omnivore
Eating habits

Pie Charts

Stacked bar chart ...

```
> ggplot(mtcars, aes(x = factor(1), fill = factor(cyl))) +
   geom_bar(width = 1)
```

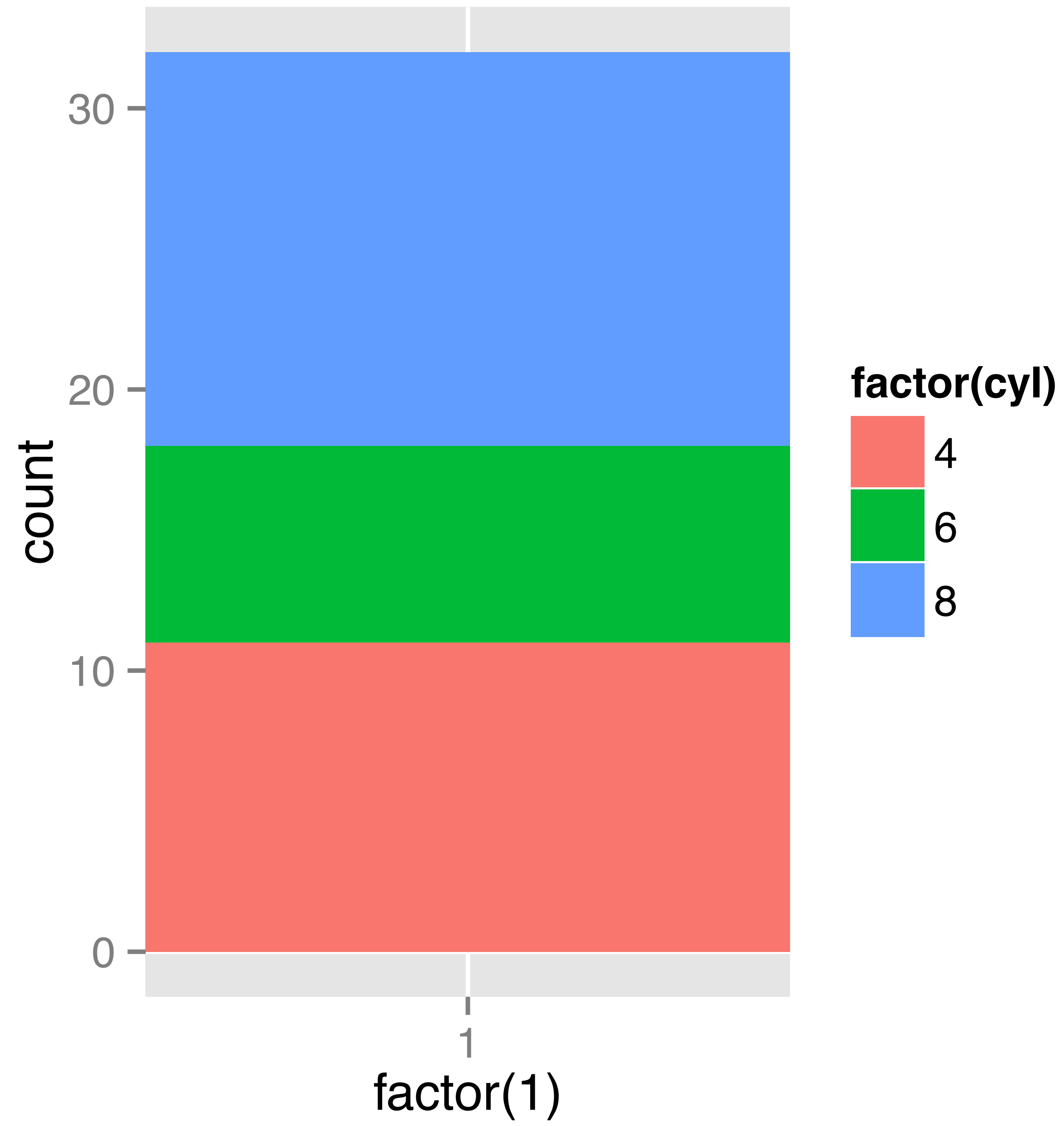


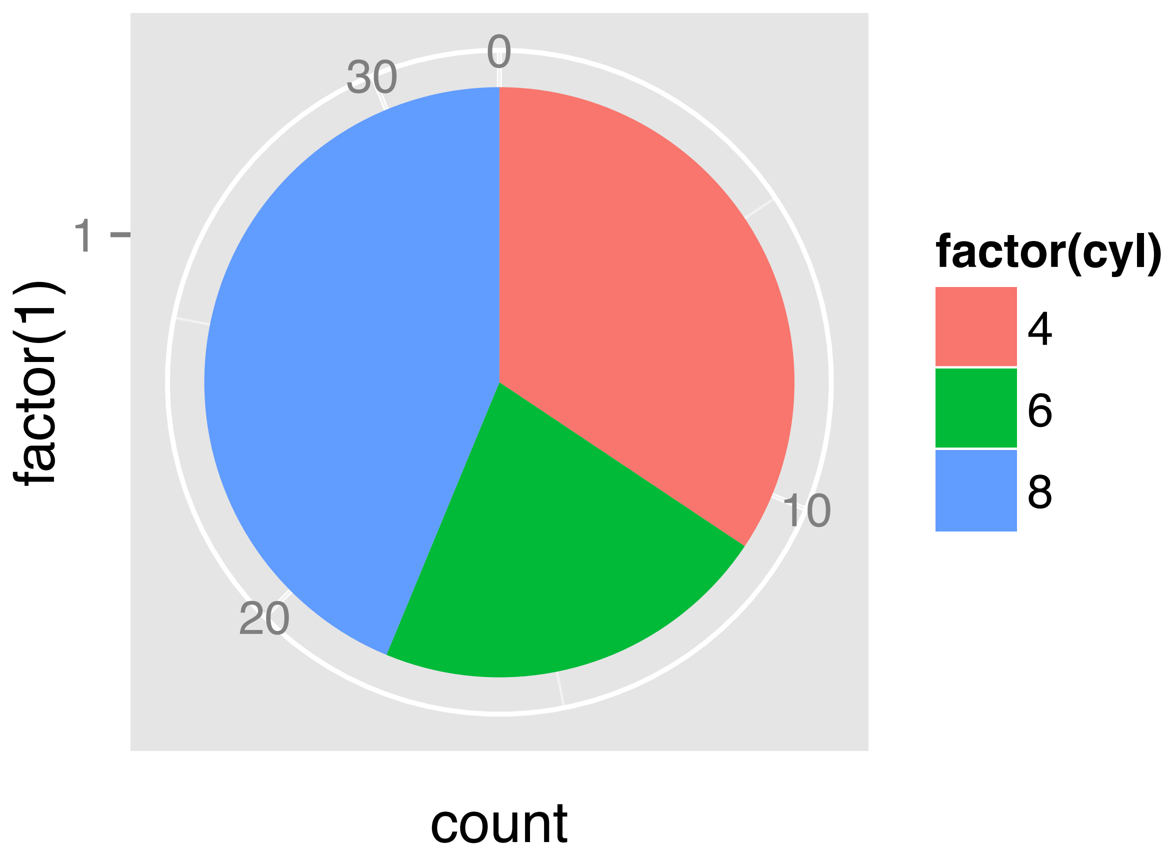
when making a piechart we are asking a question:
- what proportion of a categorical varianle is represented in each subgroup
but there remains a question - how subgroups are over or under represented

CIRCLE - a symbol of the whole - all possible oucomes appear included

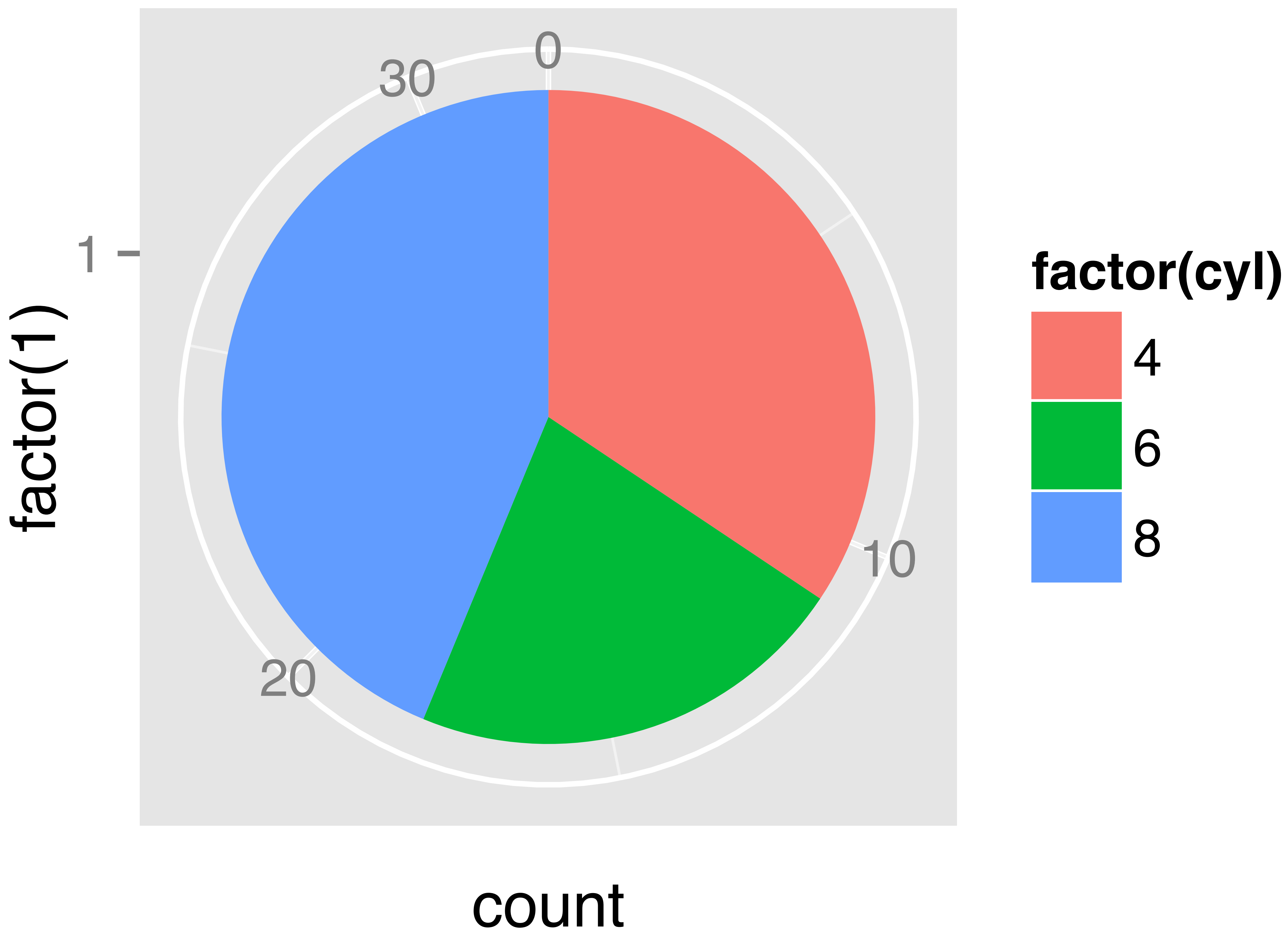
... pie chart

```
> ggplot(mtcars, aes(x = factor(1), fill = factor(cyl))) +
   geom_bar(width = 1)
> ggplot(mtcars, aes(x = factor(1), fill = factor(cyl))) +
   geom_bar(width = 1) +
   coord_polar(theta = "y")
```





Parts-of-a-whole

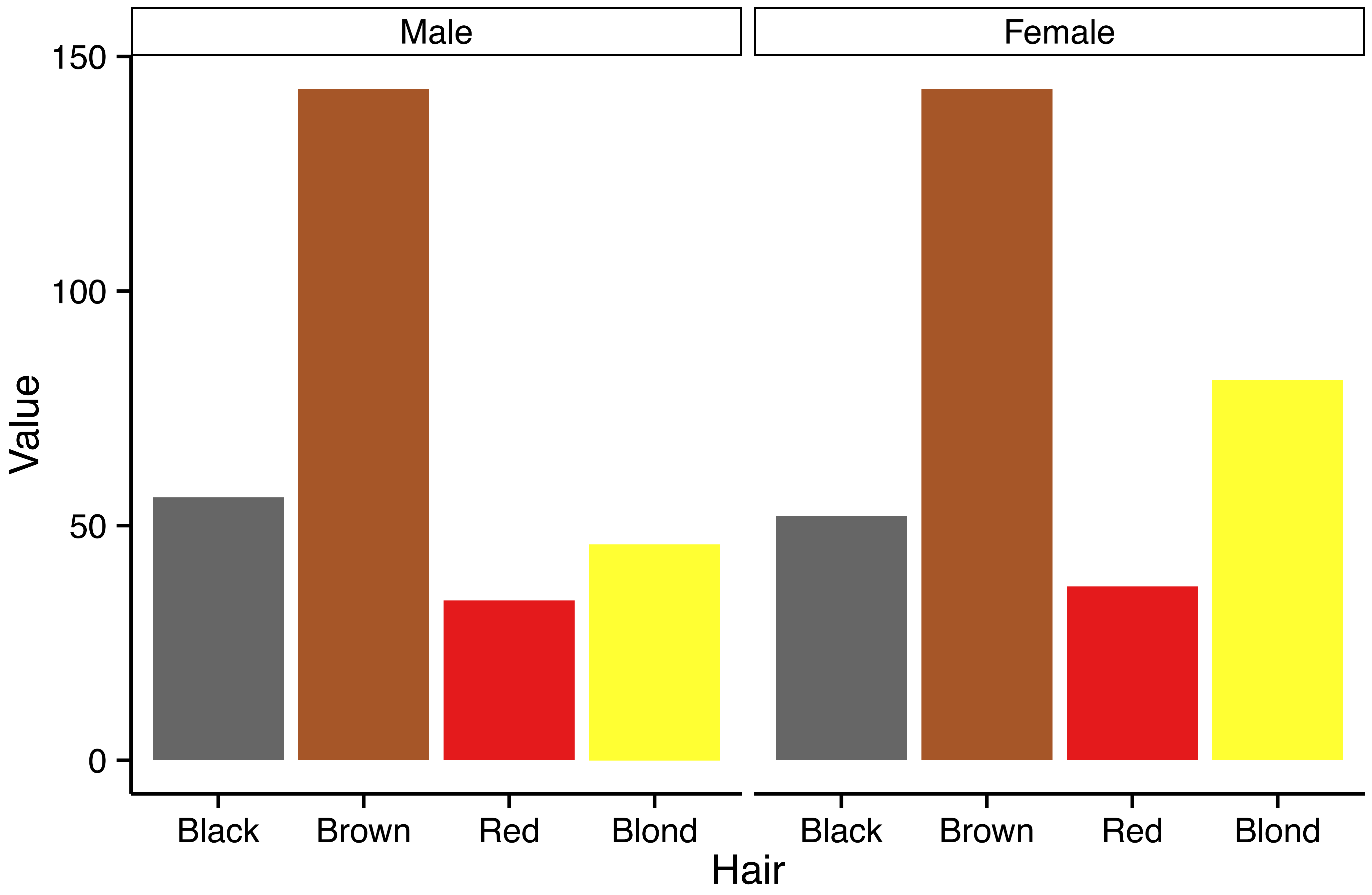


HairCol

```
> HairCol
   Hair    Sex Value  fillin   n     nprop
1 Black   Male    56 #666666 279 0.4712838
2 Brown   Male   143 #A65628 279 0.4712838
3   Red   Male    34 #E41A1C 279 0.4712838
4 Blond   Male    46 #FFFF33 279 0.4712838
5 Black Female    52 #666666 313 0.5287162
6 Brown Female   143 #A65628 313 0.5287162
7   Red Female    37 #E41A1C 313 0.5287162
8 Blond Female    81 #FFFF33 313 0.5287162
```

HairCol - Bar Charts

```
> ggplot(HairCol, aes(x = Hair, y = Value, fill = fillin)) +
    geom_bar(stat = "identity", position = "dodge") +
    facet_grid(. ~ Sex) +
    scale_fill_identity() +
    theme_classic()
```

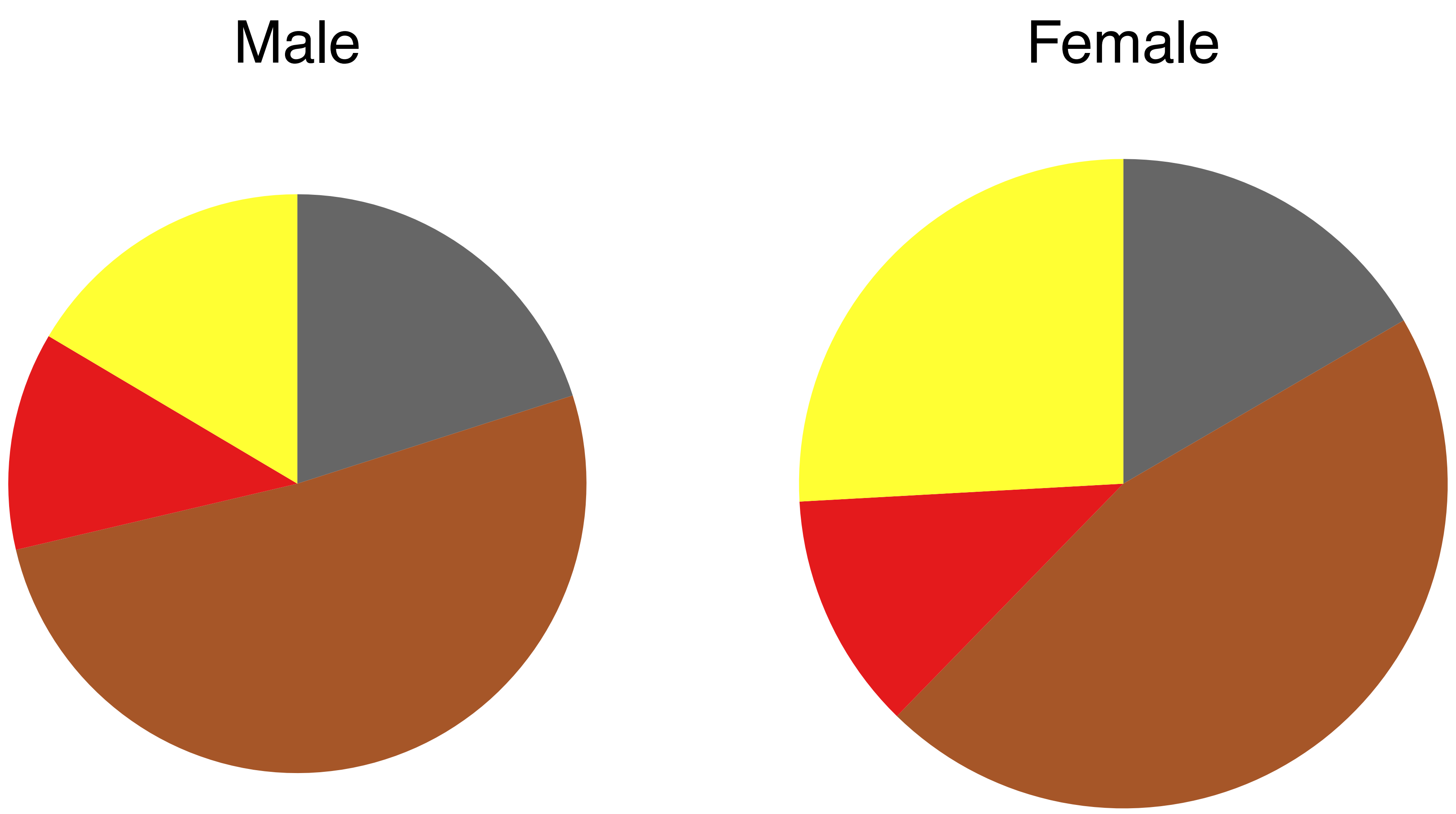


Hard to reveal interesting trends

Difference in total counts is unclear

HairCol - Pie Charts

```
> ggplot(HairCol, aes(x = n/2, y = Value,  fill = fillin, width = n)) +
    geom_bar(stat = "identity", position = "fill") +
    facet_grid(. ~ Sex) +
    scale_fill_identity() +
    coord_polar(theta = "y") +
    theme(...)
```

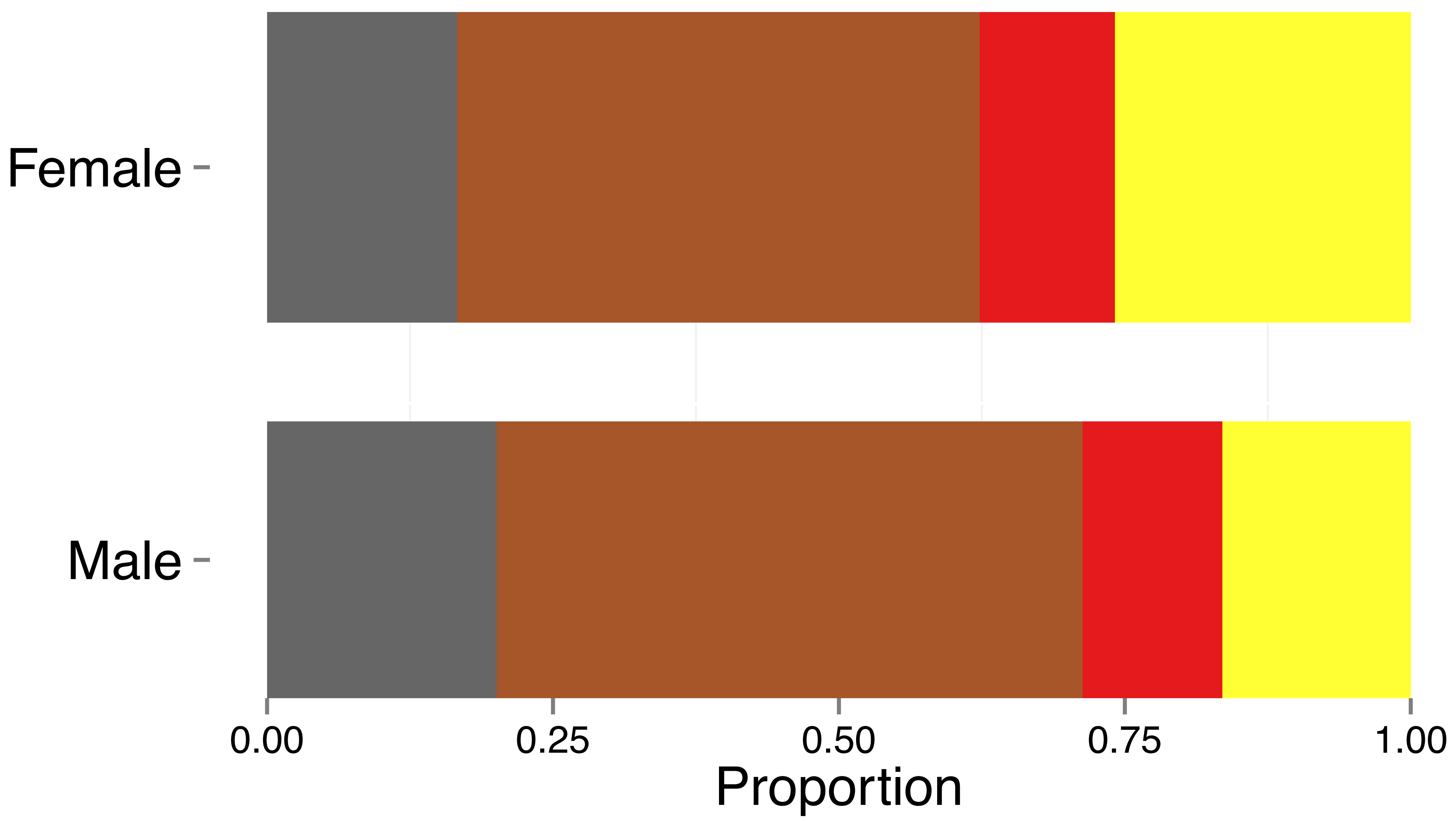


angle, area, length
mediocre encoding elements

3:
Use piecharts for encoding at most THREE variables when representing large quantitative differences

Alternative

```
> ggplot(HairCol, aes(x = Sex, y = Value, fill = fillin, width = nprop)) +
    geom_bar(stat = "identity", position= "fill") +
    scale_y_continuous("Proportion") +
    scale_x_discrete("", expand = c(0, 0)) +
    scale_fill_identity() +
    coord_flip() +
    theme(...)
```



Here we see proportions on a common scale

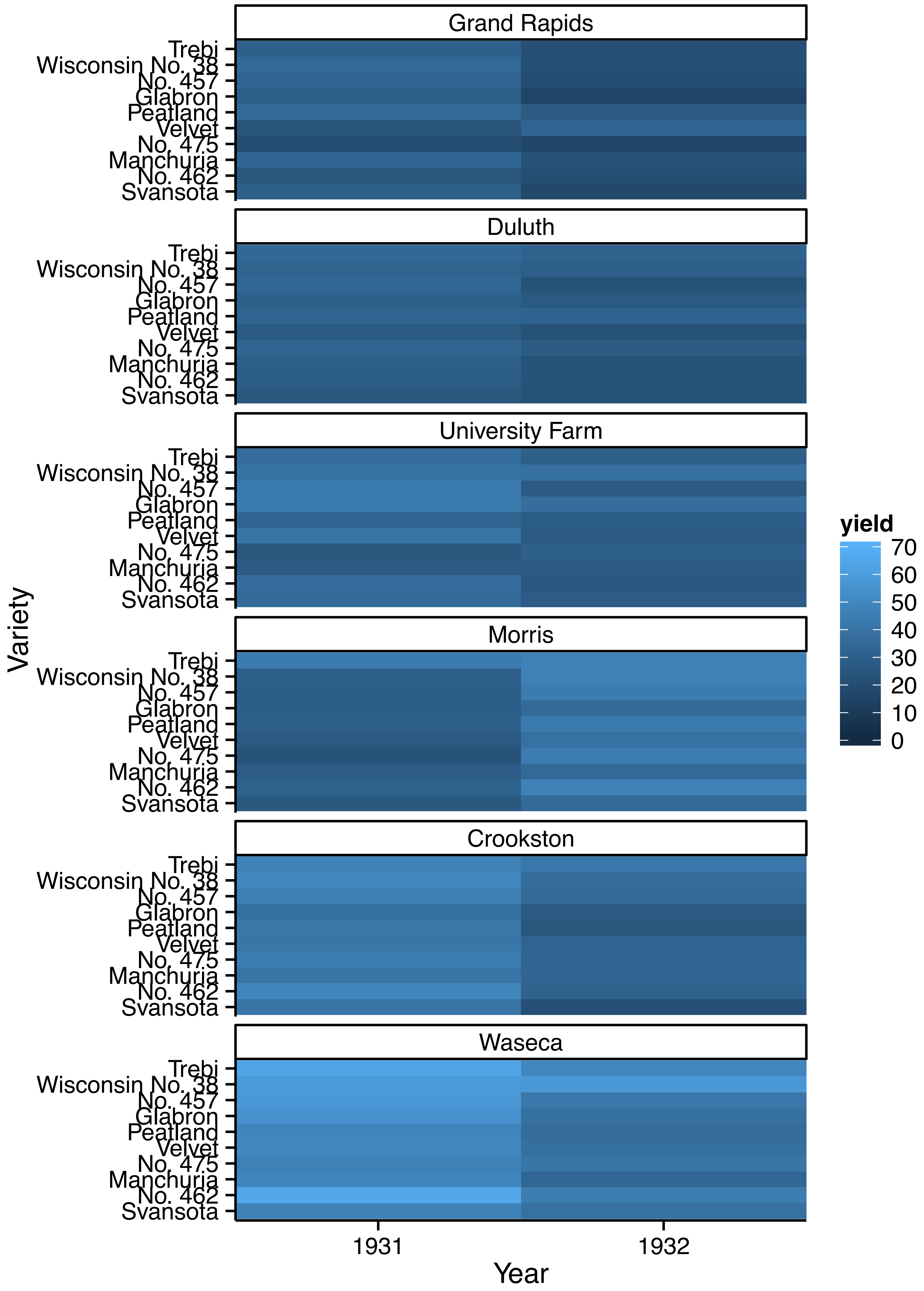
Consider parallel plots too as alternative to piecharts

DATA VISUALIZATION WITH GGPLOT2

Heat Maps

barley.s

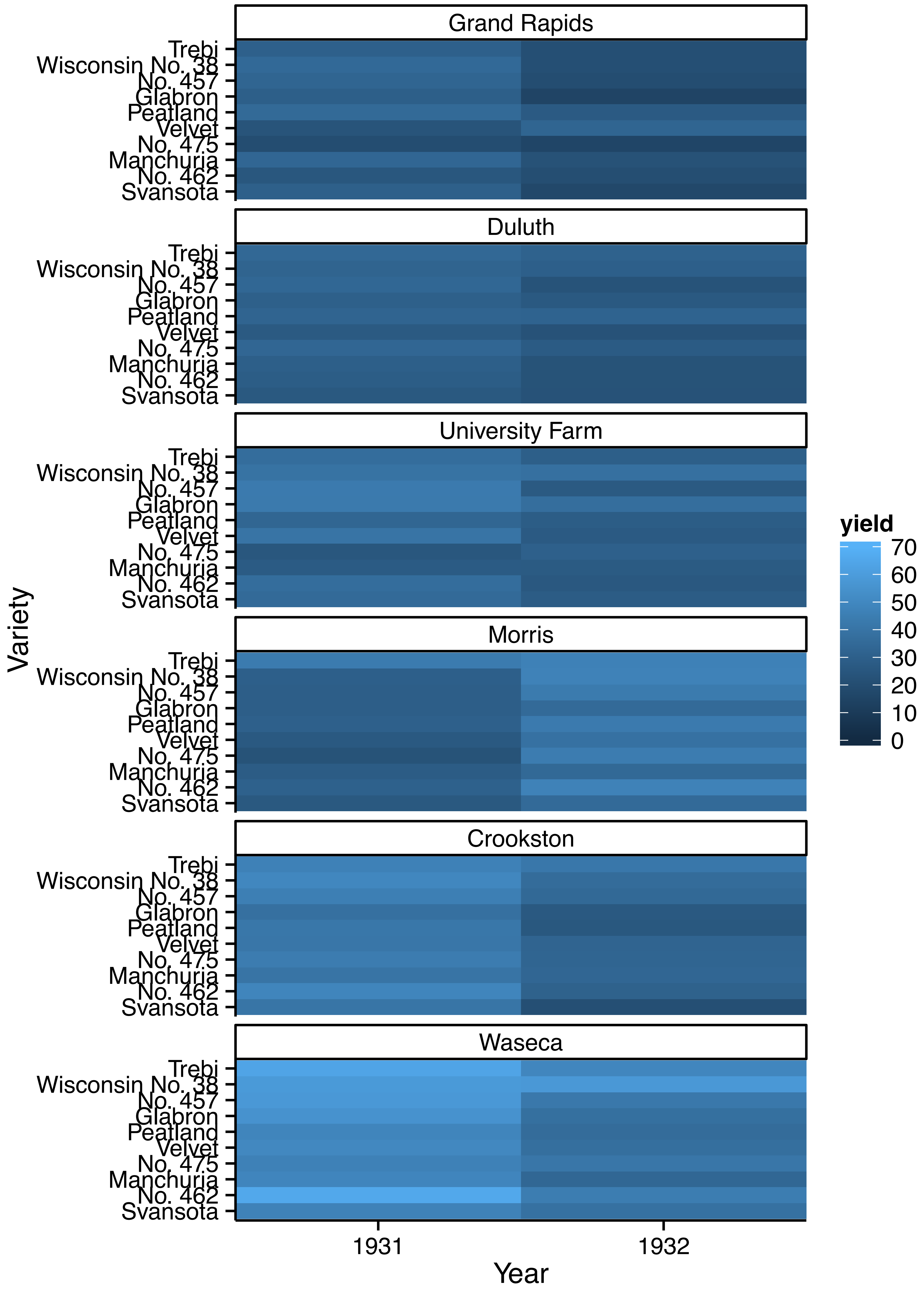
```
> head(barley.s, 15)
     variety            site     1932     1931
1   Svansota    Grand Rapids 16.63333 29.66667
2   Svansota          Duluth 22.23333 25.70000
3   Svansota University Farm 27.43334 35.13333
4   Svansota          Morris 35.03333 25.76667
5   Svansota       Crookston 20.63333 40.46667
6   Svansota          Waseca 38.50000 47.33333
7    No. 462    Grand Rapids 19.90000 24.93334
8    No. 462          Duluth 22.50000 28.10000
9    No. 462 University Farm 25.56667 36.60000
10   No. 462          Morris 47.00000 30.36667
11   No. 462       Crookston 30.53333 48.56666
12   No. 462          Waseca 44.70000 65.76670
13 Manchuria    Grand Rapids 22.13333 32.96667
14 Manchuria          Duluth 22.56667 28.96667
15 Manchuria University Farm 26.90000 27.00000
```

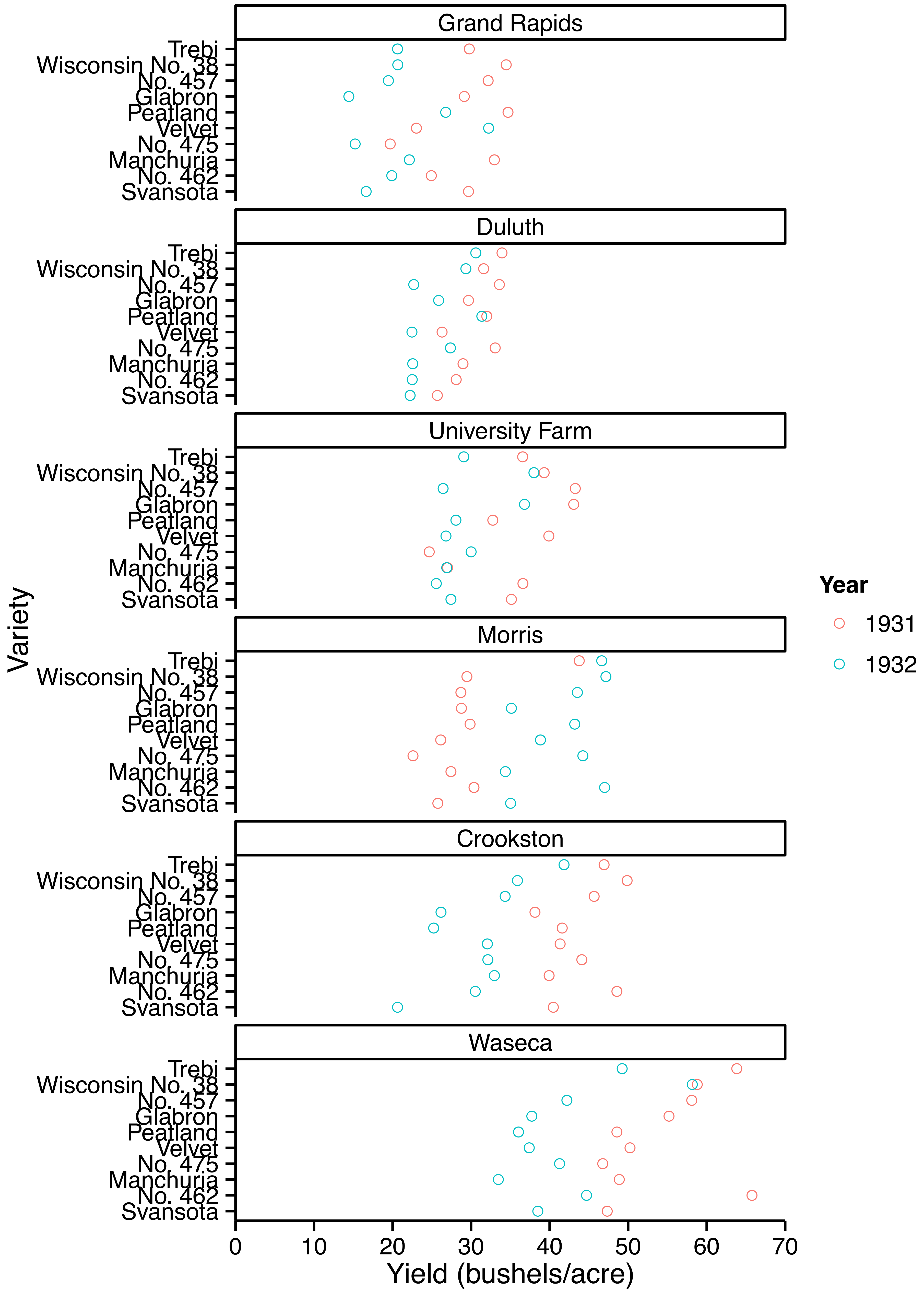


- very hard to understand
in addition eye perceives colour gradations
depending on what other shades are around it

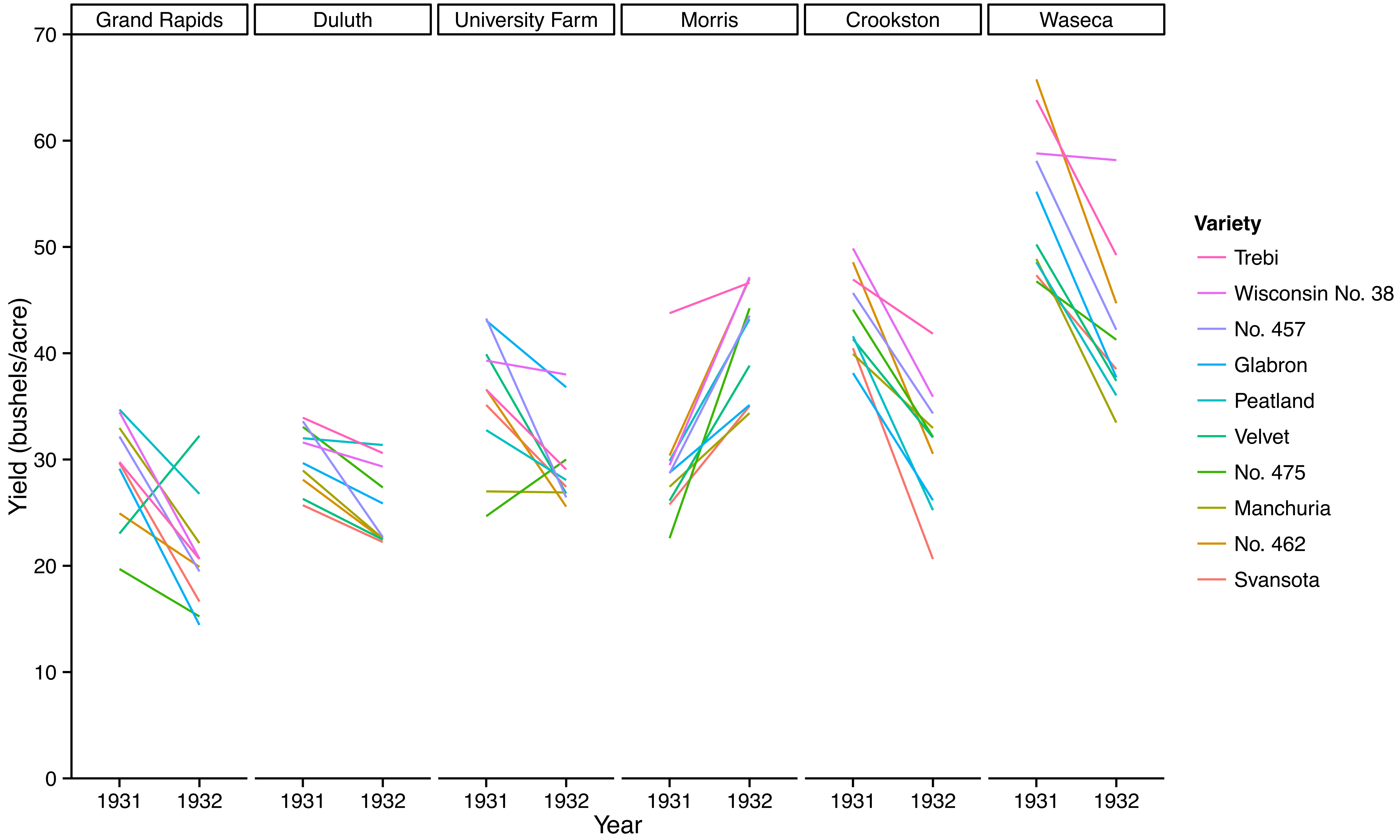
barley

```
> head(barley, 15)
      yield   variety year            site
1  27.00000 Manchuria 1931 University Farm
2  48.86667 Manchuria 1931          Waseca
3  27.43334 Manchuria 1931          Morris
4  39.93333 Manchuria 1931       Crookston
5  32.96667 Manchuria 1931    Grand Rapids
6  28.96667 Manchuria 1931          Duluth
7  43.06666   Glabron 1931 University Farm
8  55.20000   Glabron 1931          Waseca
9  28.76667   Glabron 1931          Morris
10 38.13333   Glabron 1931       Crookston
11 29.13333   Glabron 1931    Grand Rapids
12 29.66667   Glabron 1931          Duluth
13 35.13333  Svansota 1931 University Farm
14 47.33333  Svansota 1931          Waseca
15 25.76667  Svansota 1931          Morris
```

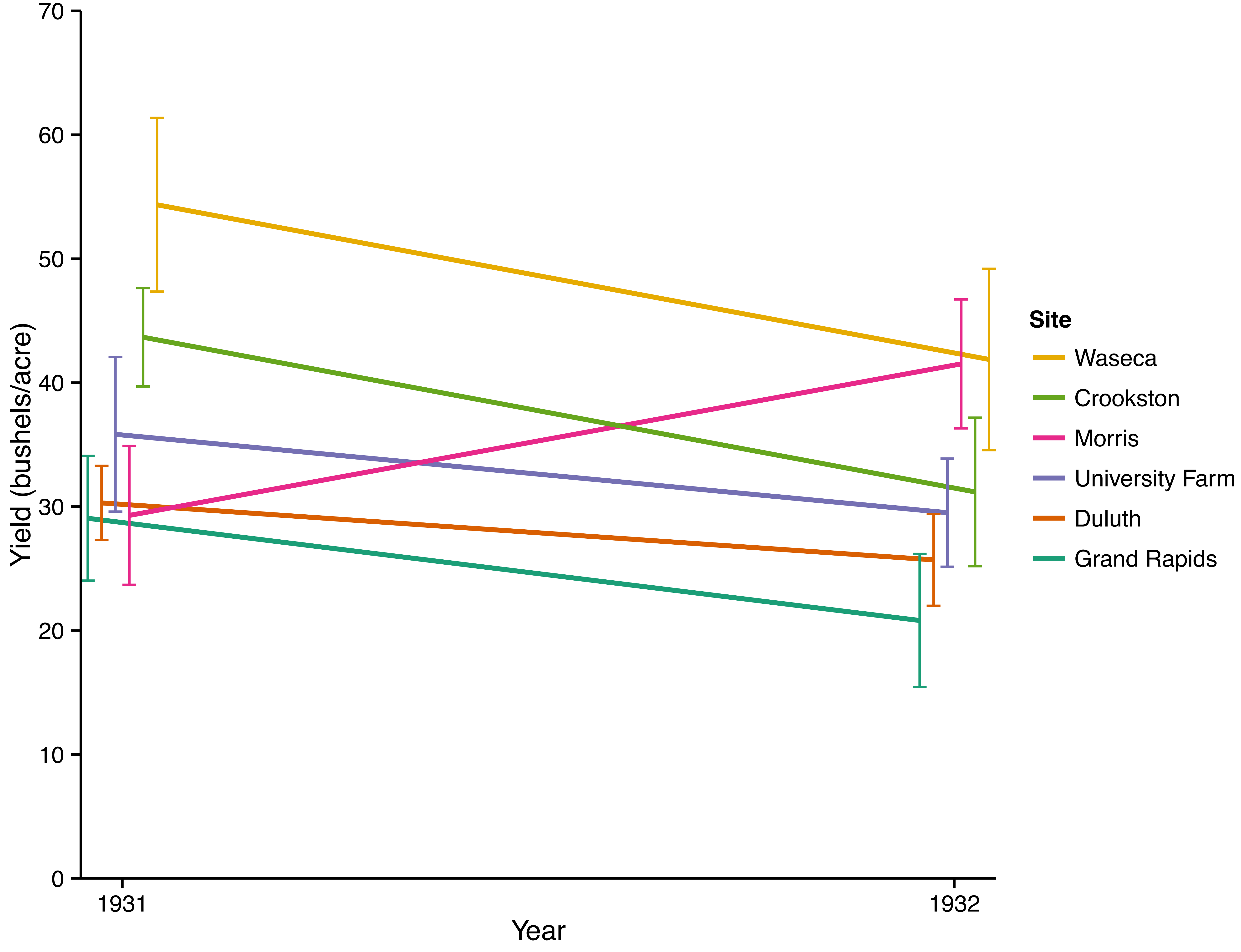

Grand Rapids
Duluth
University Farm
Morris
Crookston
Waseca
Trebi
Wisconsin No. 38
No. 457
Glabron
Peatland
Velvet
No. 475
Manchuria
No. 462
Svansota
Variety
1931
1932
Year
yield
70
60
50
40
30
20
10
0

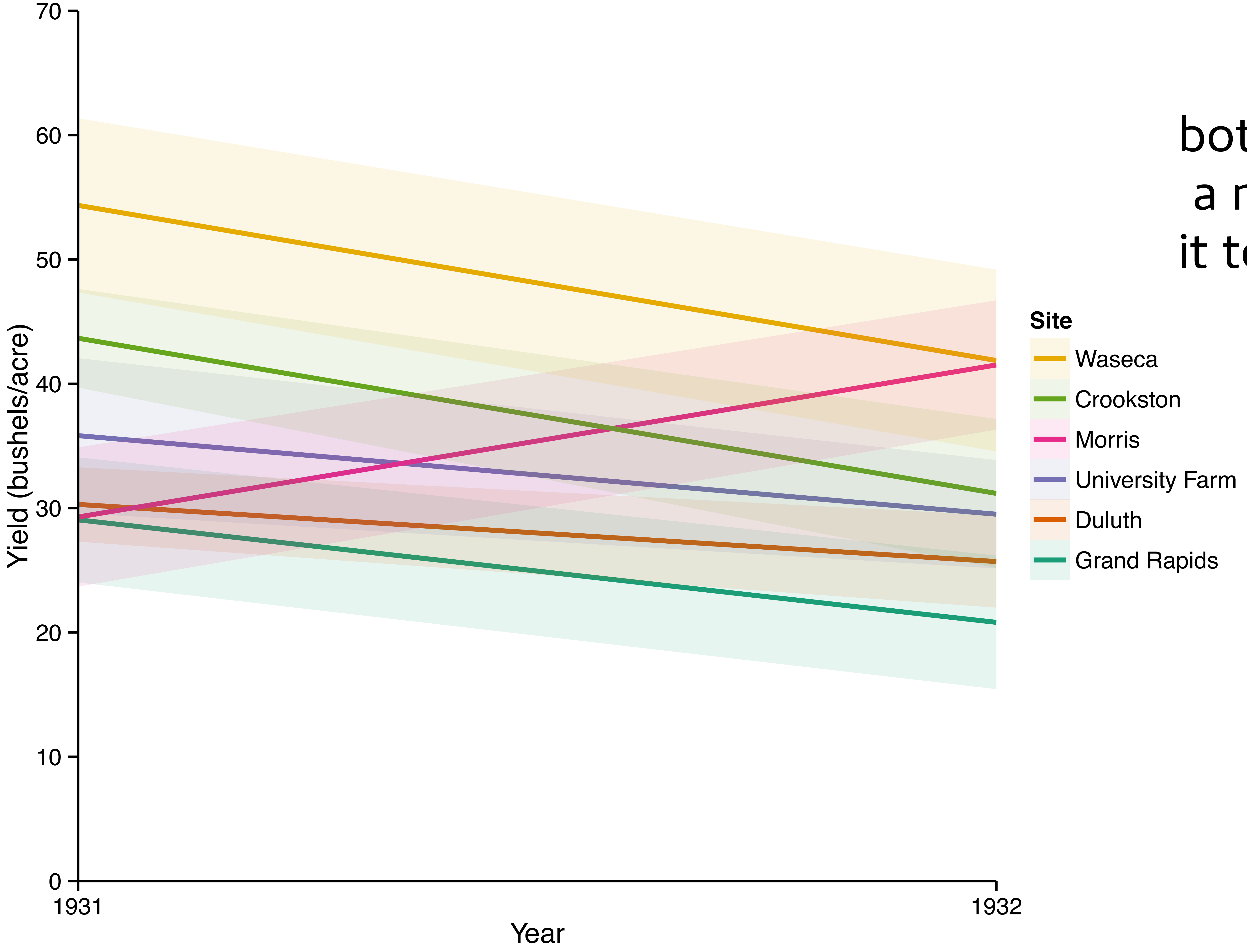


where we have data points it is much easier to see trends - from year to year
and between differing sorts and places



here trends are more clear
but colors are a bit
hard to distinguish


70
60
50
40
30
20
10
0
Yield (bushels/acre)
1931
1932
Year
Site
Waseca
Crookston
Morris
University Farm
Duluth
Grand Rapids



both trends from year to year and CIs
a nice summary and we can easily imagine
it to be good if more years added