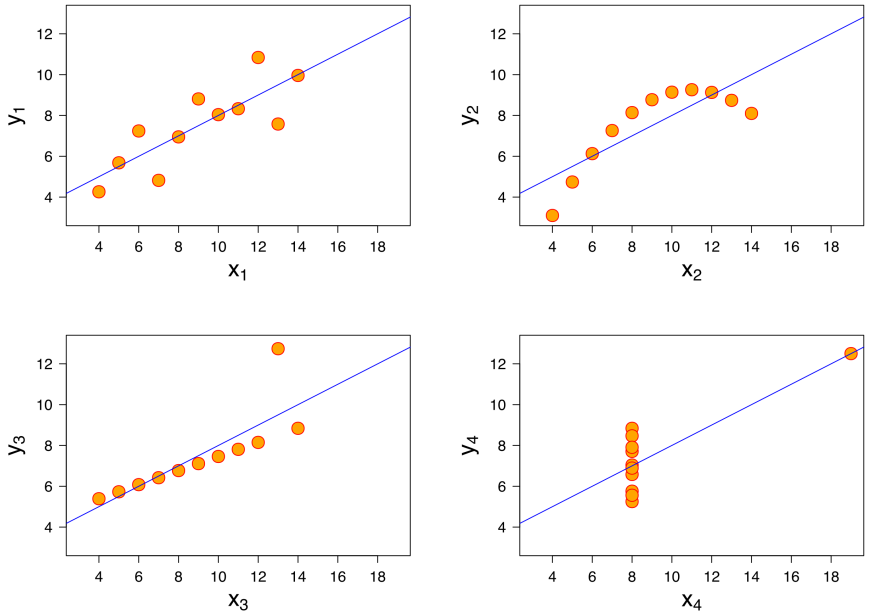
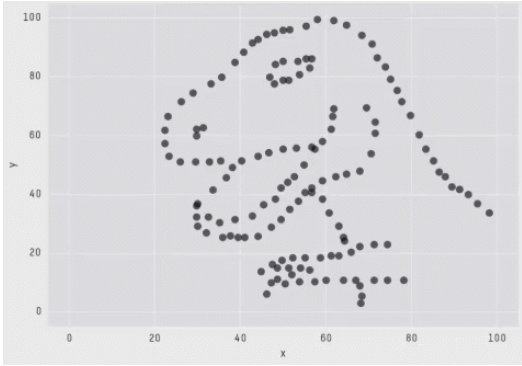
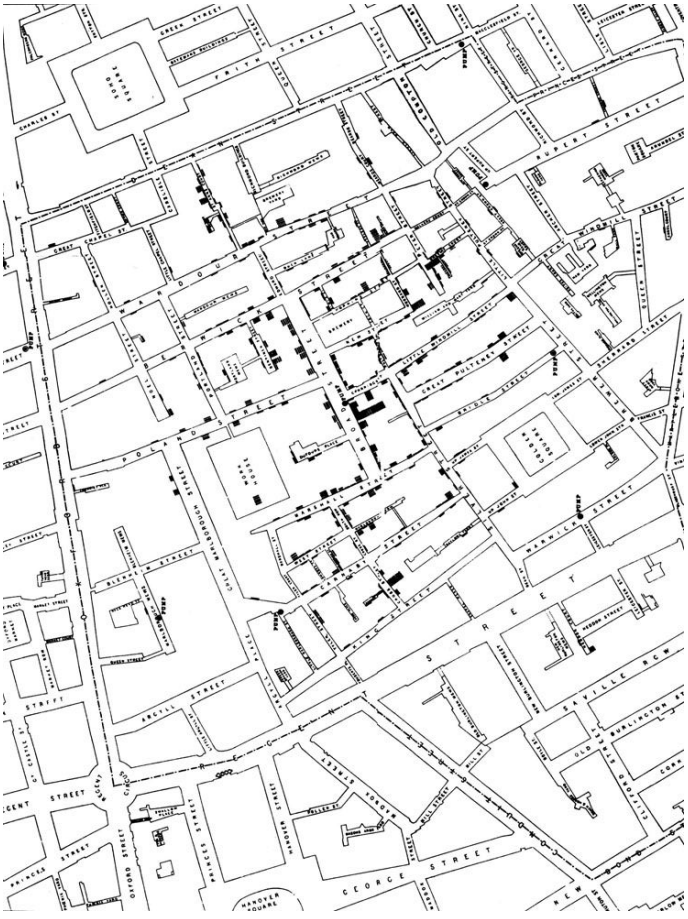
Each of the four data sets yields the same standard output from a typical regression program, namely

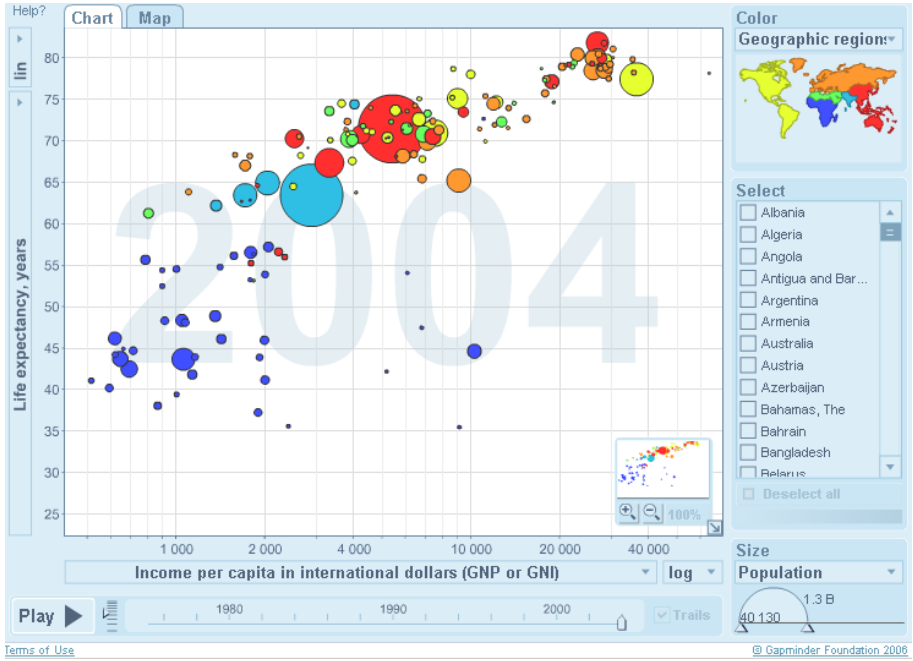
Number of observations $(n) = 11$
Mean of the x's $(\bar{x}) = 9.0$
Mean of the y's $(\bar{y}) = 7.5$
Regression coefficient (b_1) of y on $x = 0.5$
Equation of regression line: $y = 3 + 0.5\,x$
Sum of squares of $x - \bar{x} = 110.0$
Regression sum of squares $= 27.50$ (1 d.f.)
Residual sum of squares of $y = 13.75$ (9 d.f.)
Estimated standard error of $b_1 = 0.118$
Multiple $R^2 = 0.667$

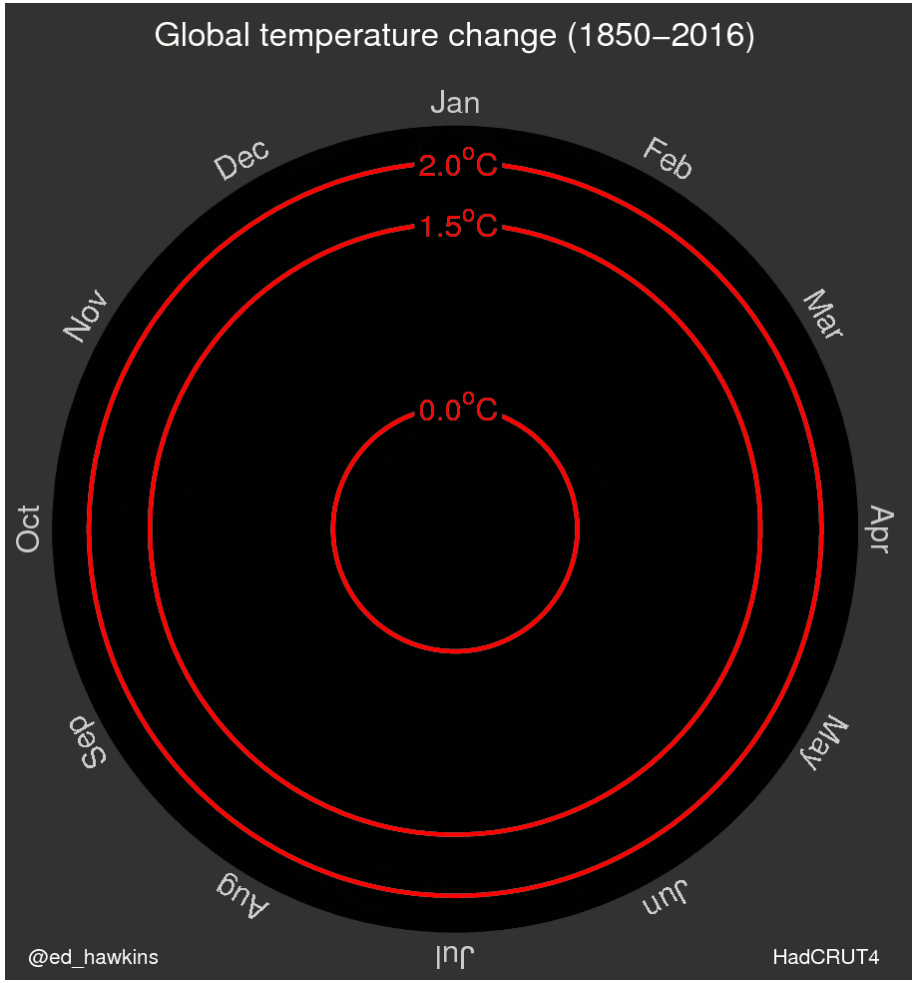

y1
x1
y2
x2
y3
x3
y4
x4
4
6
8
10
12
14
16
18

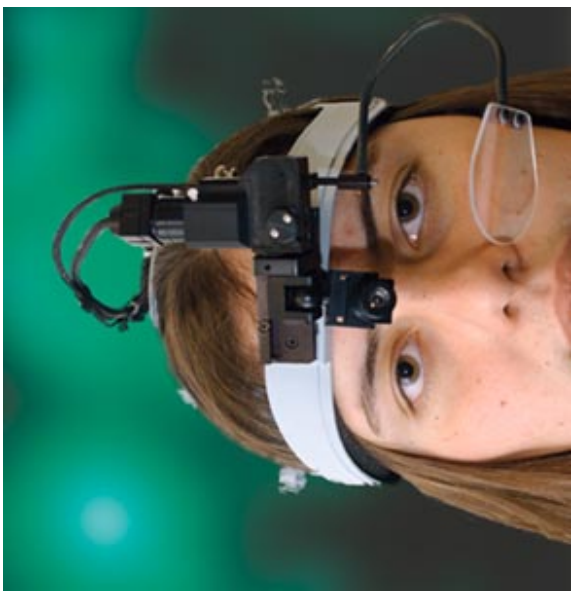


X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526


BROAD STREET
GEORGE STREET
GREEK STREET
PORTLAND STREET
RUPERT STREET
MARSHALL STREET
BERWICK STREET
BEAK STREET
GREAT MARLBOROUGH STREET
ARGYLL STREET
LITTLE MARLBOROUGH STREET
GOLDEN SQUARE
HANOVER SQUARE
OXFORD STREET
REGENT STREET


Help?
Chart
Map
lin
Life expectancy, years
80
75
70
65
60
55
50
45
40
35
30
25
2004
1 000
2 000
4 000
10 000
20 000
40 000
Income per capita in international dollars (GNP or GNI)
log
Play
1980
1990
2000
Trails
100%
Color
Geographic region
Select
Albania
Algeria
Angola
Antigua and Bar...
Argentina
Armenia
Australia
Austria
Azerbaijan
Bahamas, The
Bahrain
Bangladesh
Deselect all
Size
Population
1.3 B
40
130
Terms of Use
© Gapminder Foundation 2006


Global temperature change (1850–2016)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2.0°C
1.5°C
0.0°C
@ed_hawkins
HadCRUT4



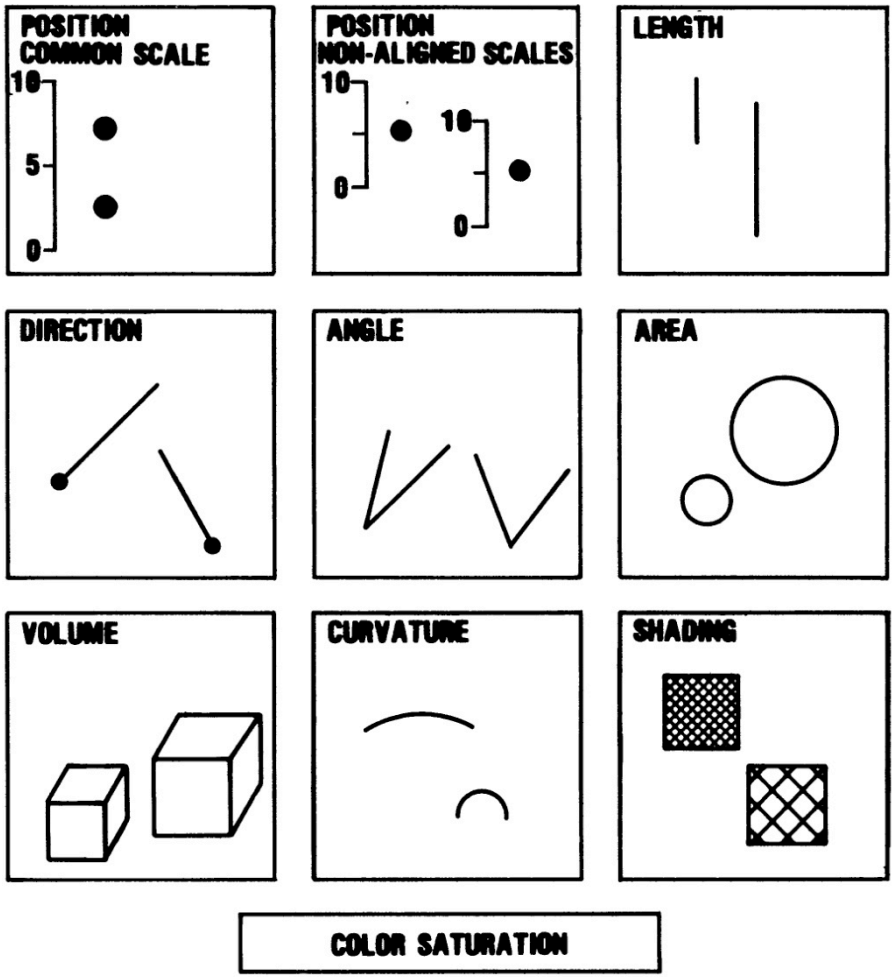


Figure 1. Elementary perceptual tasks.

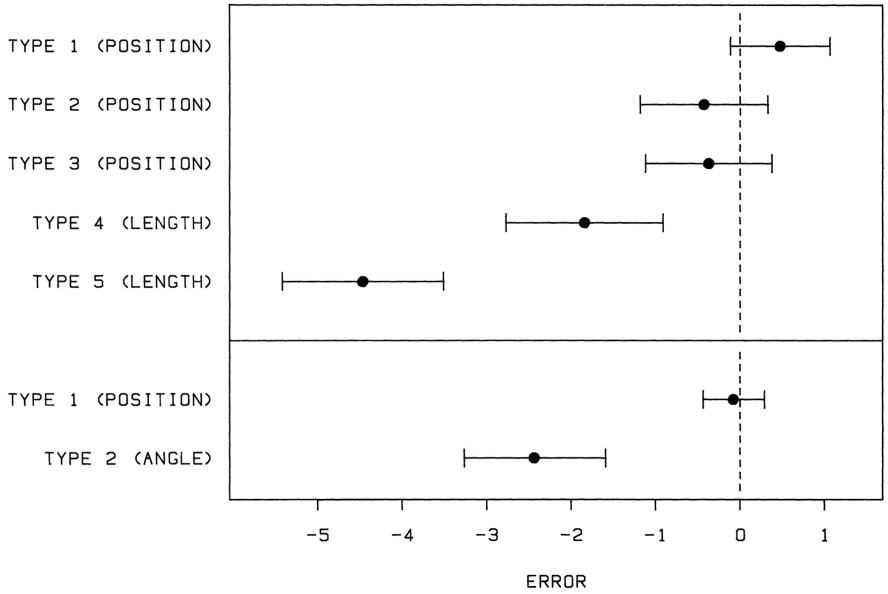


Figure 20. Error means and 95% confidence intervals for judgment types in position–length experiment (top) and position–angle experiment (bottom).

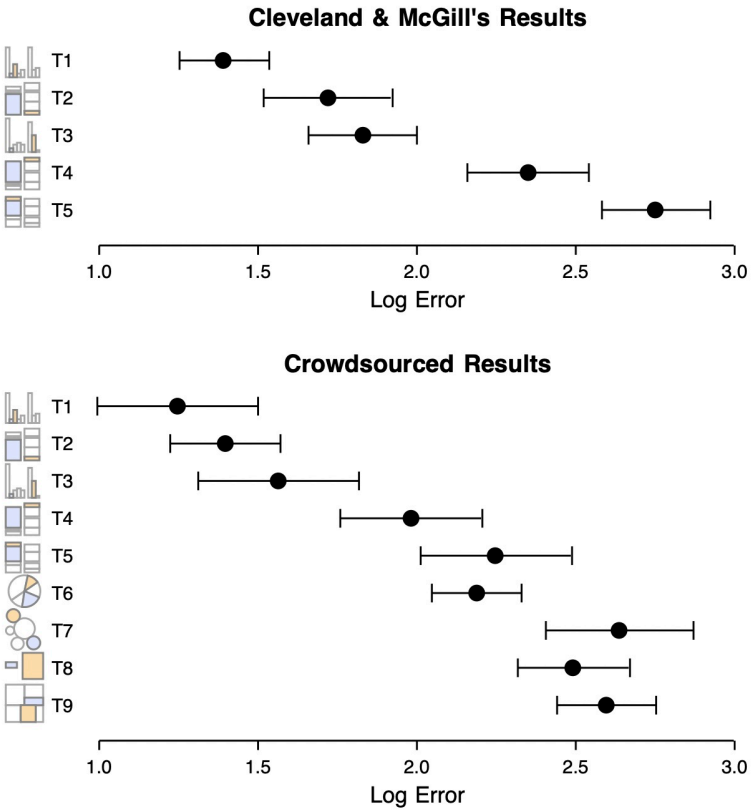
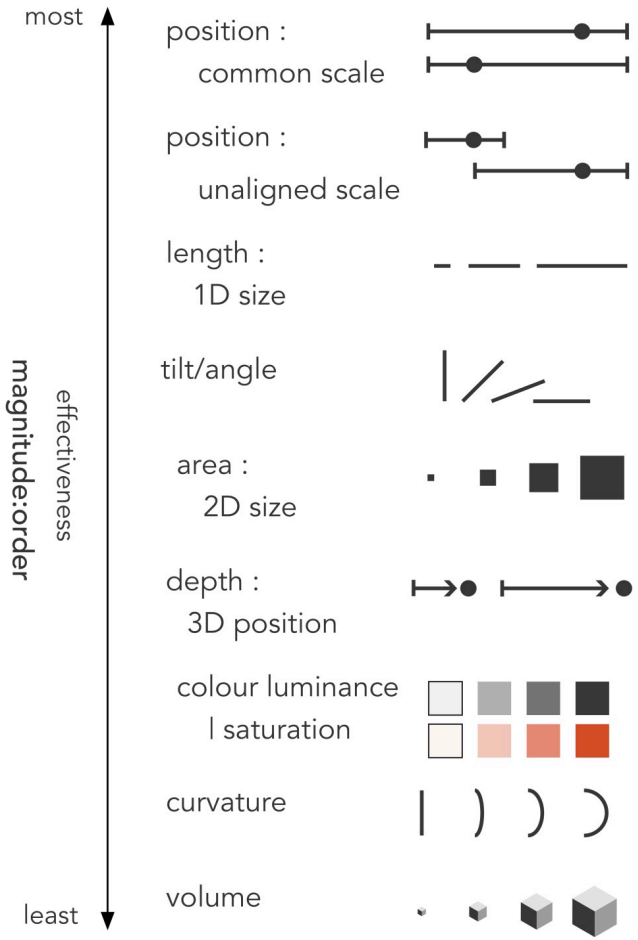
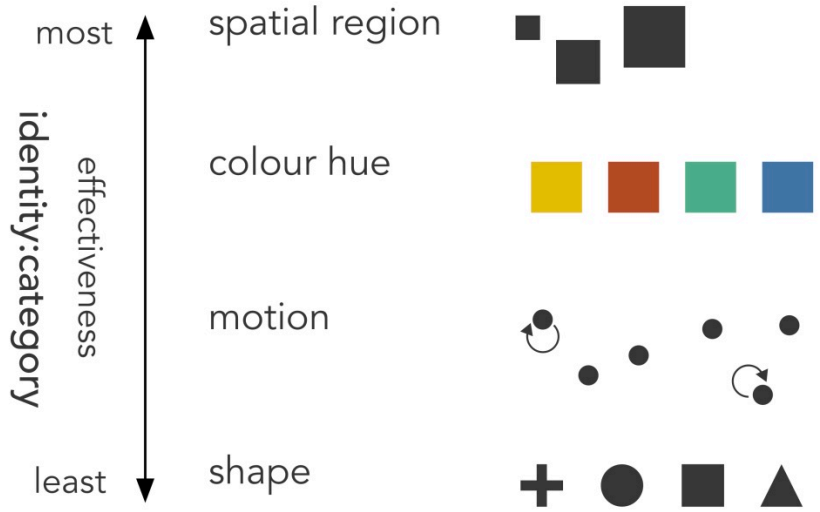


Figure 4: Proportional judgment results (Exp. 1A & B). Top: Cleveland & McGill's [7] lab study. Bottom: MTurk studies. Error bars indicate 95% confidence intervals.


most
effectiveness
magnitude:order
least
position :
common scale
position :
unaligned scale
length :
1D size
tilt/angle
area :
2D size
depth :
3D position
colour luminance
I saturation
curvature
volume


most
effectiveness
identity:category
least
spatial region
colour hue
motion
shape

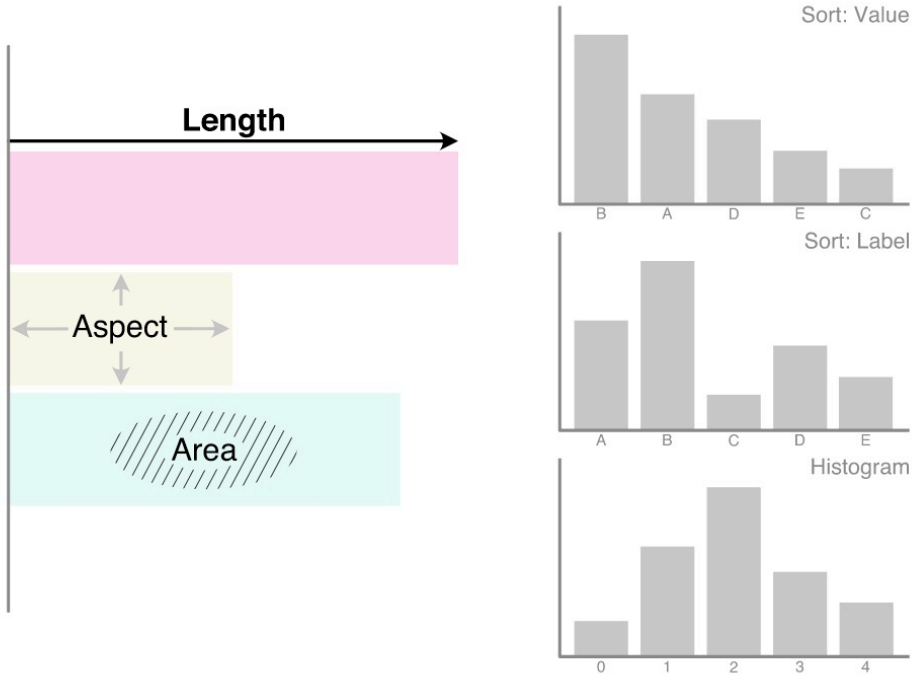


FIGURE 2. Bar charts encode values as their length, but also their area, aspect ratio, and overall shape. Sorting is often used for specific use cases.

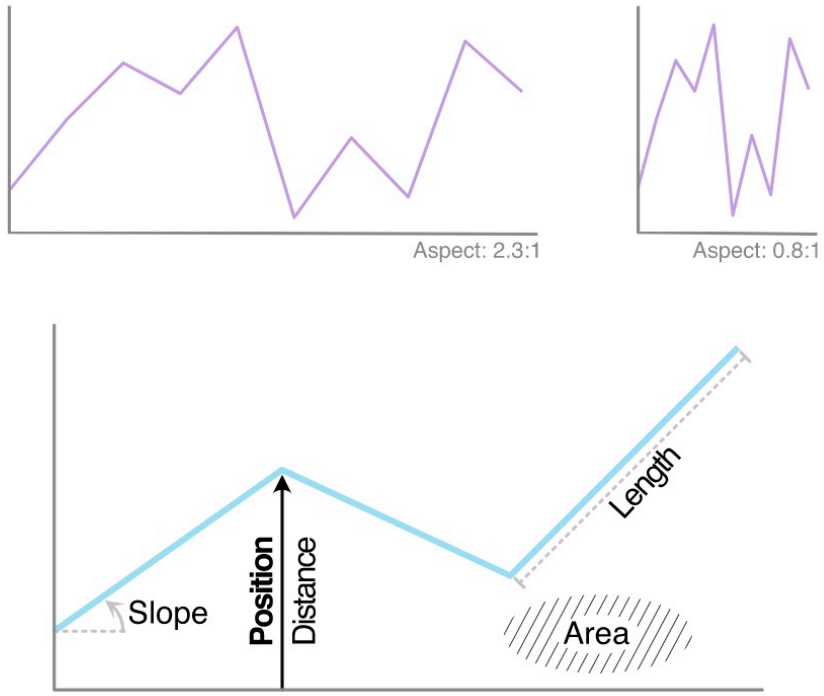


FIGURE 3. Line charts are specified by the location of the points connected by lines, but are read as slope and length, as well as area. Aspect ratio of the chart is also generally considered important.

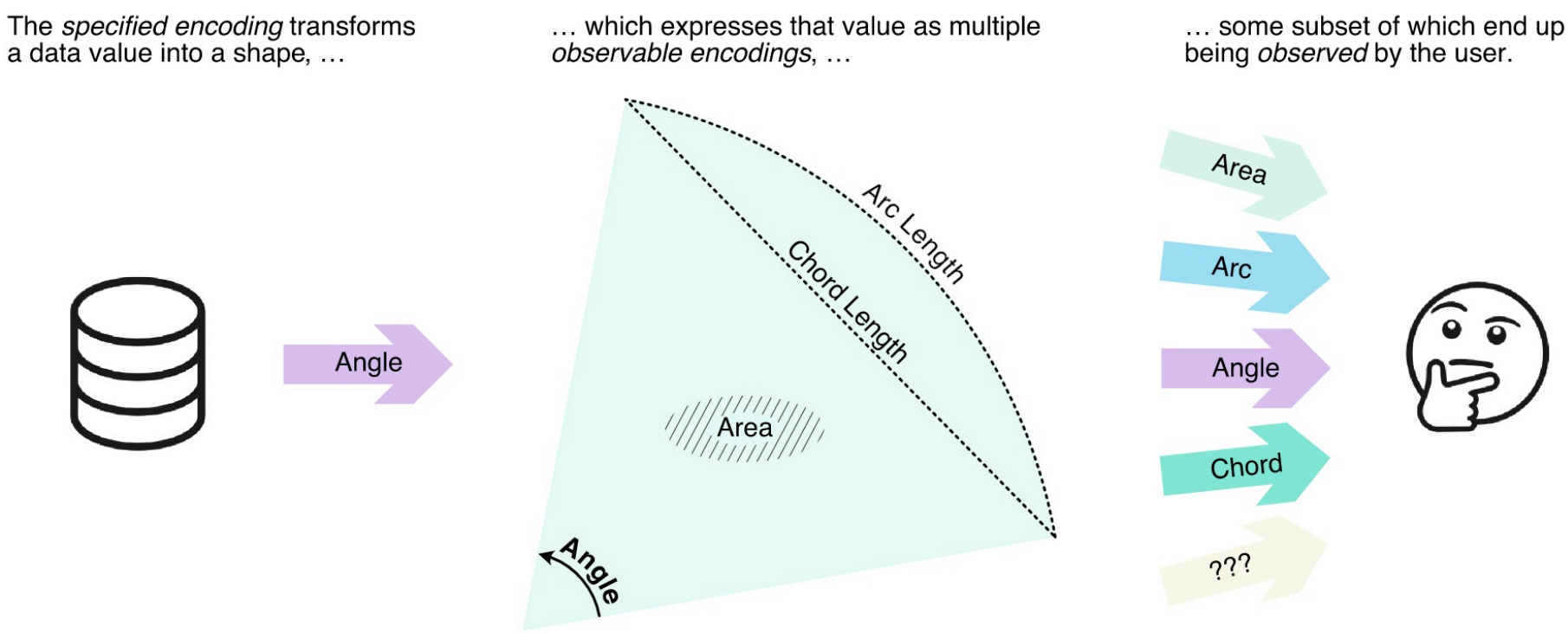
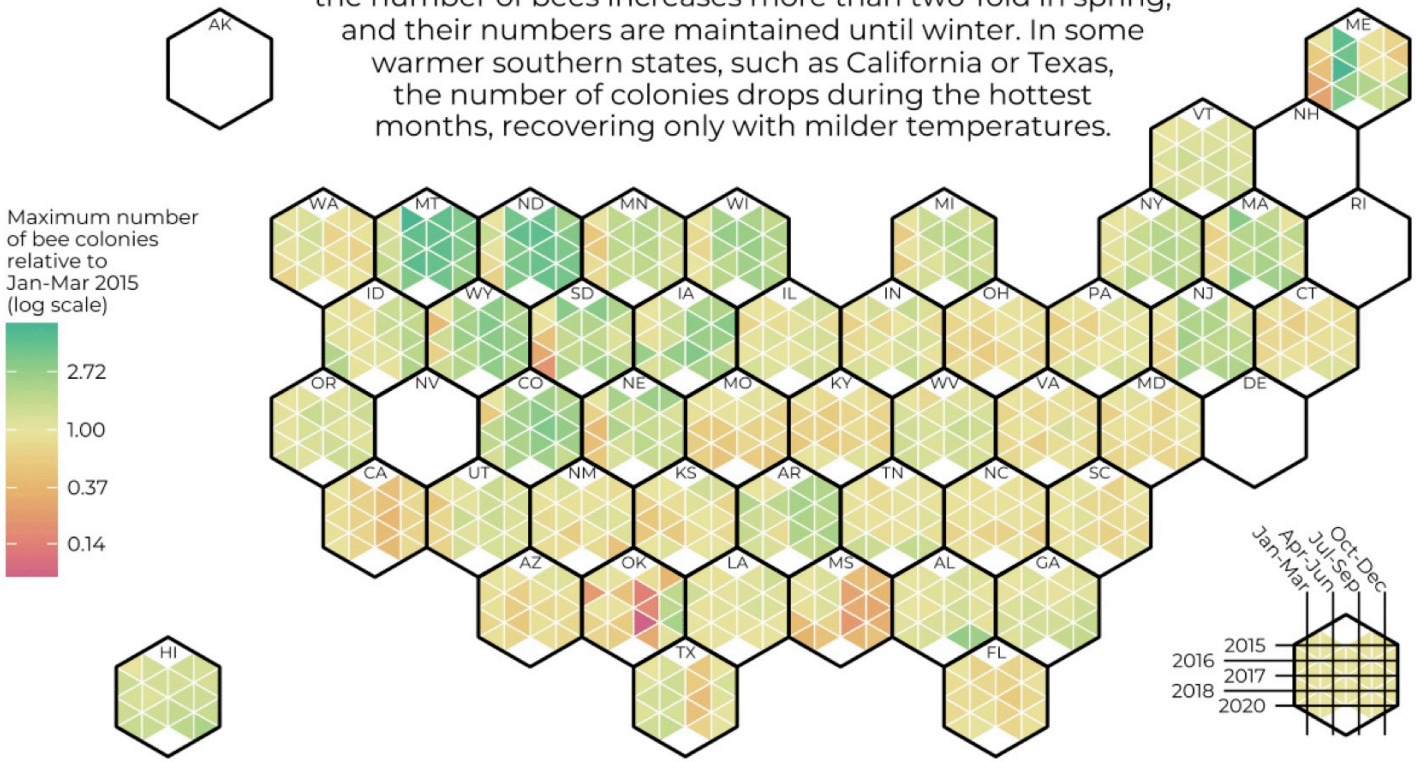


FIGURE 1. Pie charts are specified by angle, but may also be read by area, arc length, or even chord. Shape recognition is also likely for specific angles like 90°/25% and 180°/50%.

Data from USDA
Graph by @irg_bio

Seasonality in the number of bee colonies

The number of bee colonies in the US changes with the seasons depending on the state. In some northern states, the number of bees increases more than two-fold in spring, and their numbers are maintained until winter. In some warmer southern states, such as California or Texas, the number of colonies drops during the hottest months, recovering only with milder temperatures.



To extract accu

To quantatively

To find the largest

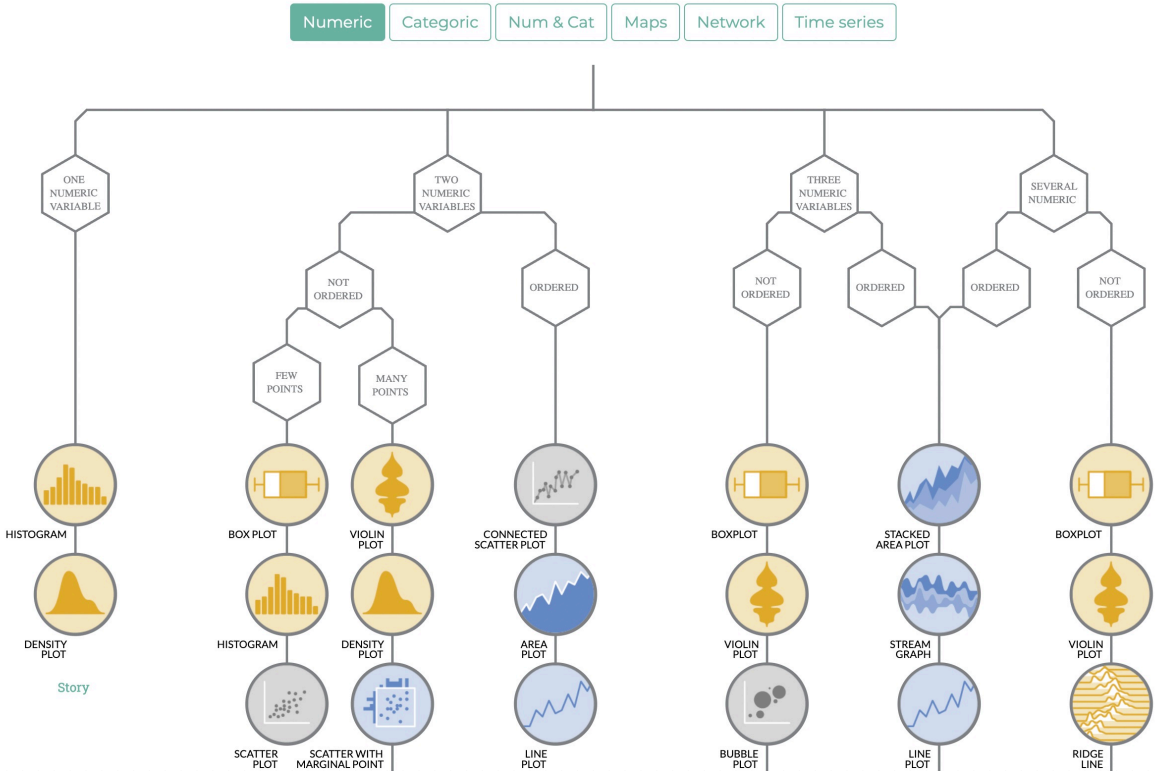
To find unusuvaal

You have a story you want to tel

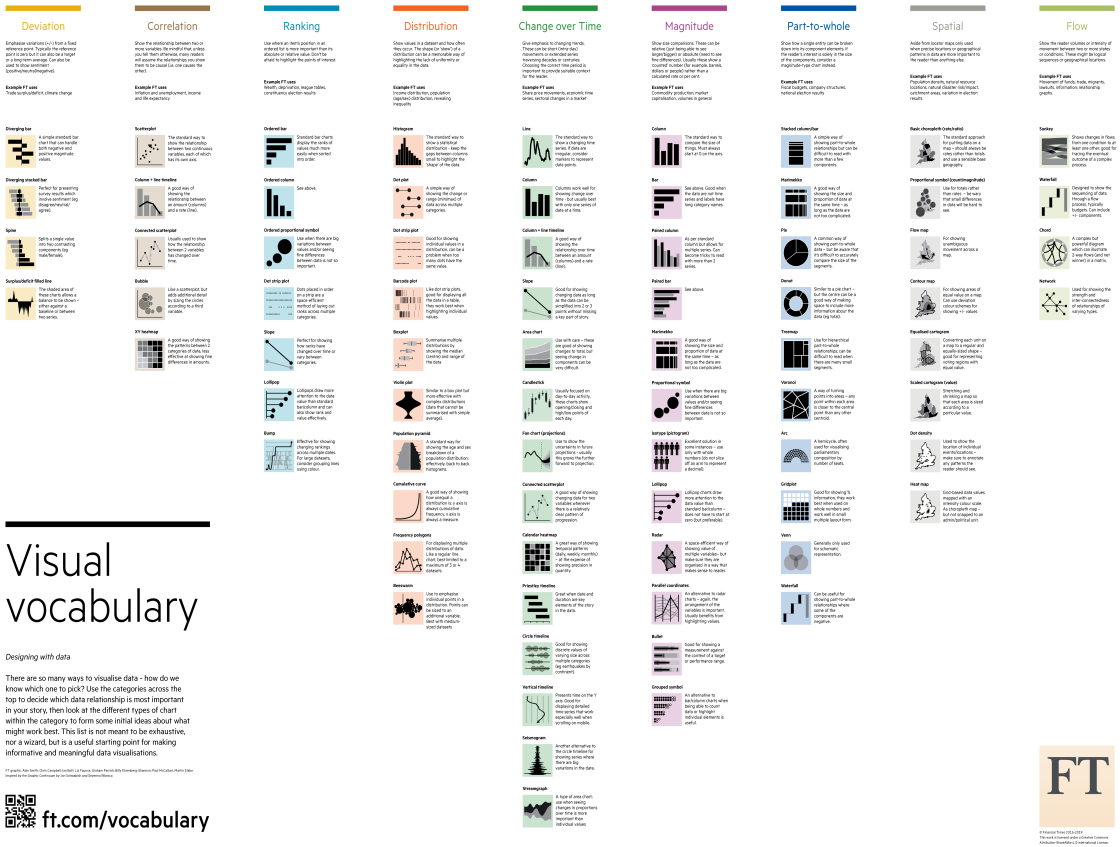
The reader want

The reader has a preconception ab

```
> msleep
# A tibble: 83 × 11
   name     genus vore  order conse…¹ sleep…² sleep…³ sleep…⁴ awake  brainwt  bodywt
   <chr>    <chr> <chr> <chr> <chr>     <dbl>   <dbl>   <dbl> <dbl>    <dbl>   <dbl>
 1 Cheetah  Acin… carni Carn… lc         12.1    NA      NA    11.9 NA         50
 2 Owl mo…  Aotus omni  Prim… NA         17       1.8    NA     7    0.0155     0.48
 3 Mounta…  Aplo… herbi Rode… nt         14.4     2.4    NA     9.6  NA         1.35
 4 Greate…  Blar… omni  Sori… lc         14.9     2.3     0.133 9.1  0.00029    0.019
 5 Cow      Bos   herbi Arti… domest…     4       0.7     0.667 20   0.423    600
 6 Three-…  Brad… herbi Pilo… NA         14.4     2.2     0.767 9.6  NA         3.85
 7 Northe…  Call… carni Carn… vu          8.7     1.4     0.383 15.3 NA        20.5
 8 Vesper…  Calo… NA    Rode… NA          7      NA      NA    17   NA         0.045
 9 Dog      Canis carni Carn… domest…    10.1     2.9     0.333 13.9 0.07      14
10 Roe de…  Capr… herbi Arti… lc          3      NA      NA    21   0.0982    14.8
# … with 73 more rows, and abbreviated variable names ¹conservation, ²sleep_total,
#   ³sleep_rem, ⁴sleep_cycle
# ℹ Use `print(n = ...)` to see more rows
```


Numeric
Categoric
Num & Cat
Maps
Network
Time series
ONE NUMERIC VARIABLE
TWO NUMERIC VARIABLES
THREE NUMERIC VARIABLES
SEVERAL NUMERIC
NOT ORDERED
ORDERED
ORDERED
ORDERED
NOT ORDERED
FEW POINTS
MANY POINTS
HISTOGRAM
DENSITY PLOT
BOX PLOT
VIOLIN PLOT
CONNECTED SCATTER PLOT
BOXPLOT
STACKED AREA PLOT
BOXPLOT
HISTOGRAM
DENSITY PLOT
AREA PLOT
VIOLIN PLOT
STREAM GRAPH
VIOLIN PLOT
SCATTER PLOT
SCATTER WITH MARGINAL POINT
LINE PLOT
BUBBLE PLOT
LINE PLOT
RIDGE LINE

Story



Visual vocabulary

Designing with data

There are so many ways to visualise data - how do we know which one to pick? Use the categories across the top to decide which data relationship is most important in your story, then look at the different types of chart within the category to form some initial ideas about what might work best. This list is not meant to be exhaustive, nor a wizard, but is a useful starting point for making informative and meaningful data visualisations.

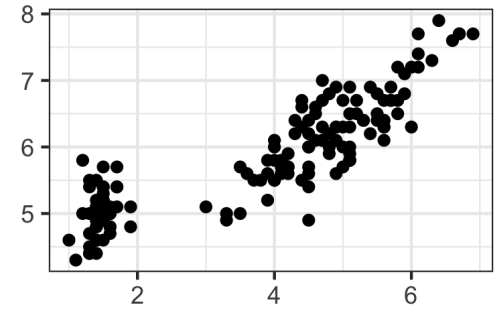
FT graphic: Alan Smith, Chris Campbell, Ian Bott, Liz Faunce, Graham Parrish, Billy Ehrenberg-Shannon, Paul McCallum, Martin Stabe
Inspired by the Graphic Continuum by Jon Schwabish and Severino Ribecca

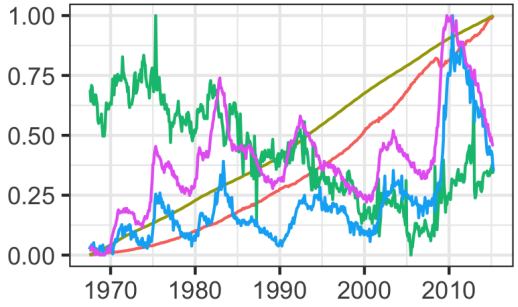
ft.com/vocabulary

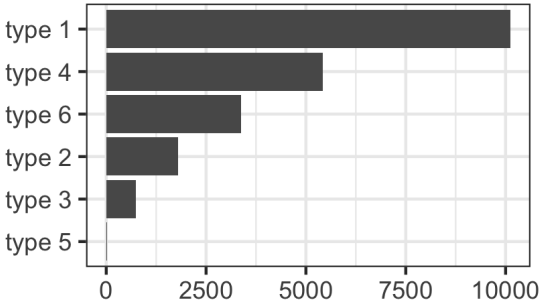
GGPLOT2 AS A CREATIVITY ENGINE

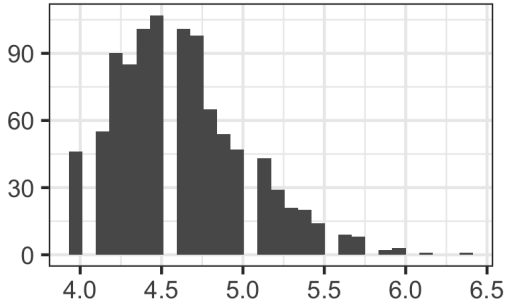
... AND OTHER WAYS R IS TRANSFORMING THE FT'S QUANTITATIVE JOURNALISM

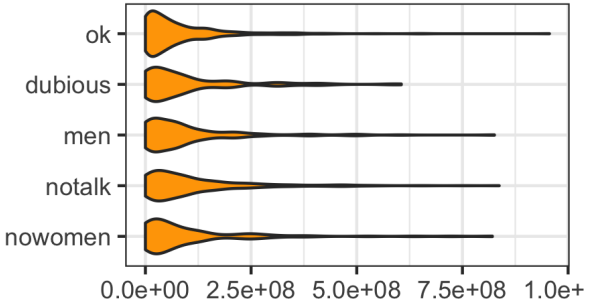
© Financial Times 2014-2019
Made in Illustrator and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

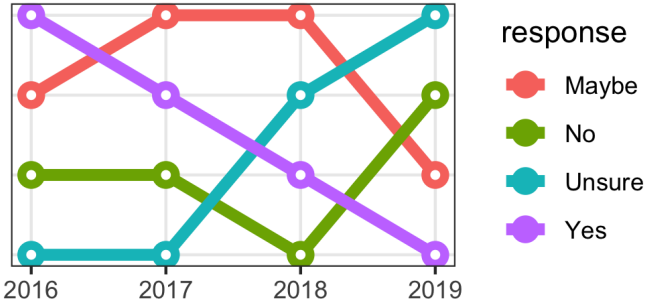

8
7
6
5
2
4
6


1.00
0.75
0.50
0.25
0.00
1970
1980
1990
2000
2010


type 1
type 4
type 6
type 2
type 3
type 5
0
2500
5000
7500
10000


90
60
30
0
4.0
4.5
5.0
5.5
6.0
6.5


ok
dubious
men
notalk
nowomen
0.0e+00
2.5e+08
5.0e+08
7.5e+08
1.0e+


2016
2017
2018
2019
response
Maybe
No
Unsure
Yes

Aesthetics

Geoms

Scales

Guides

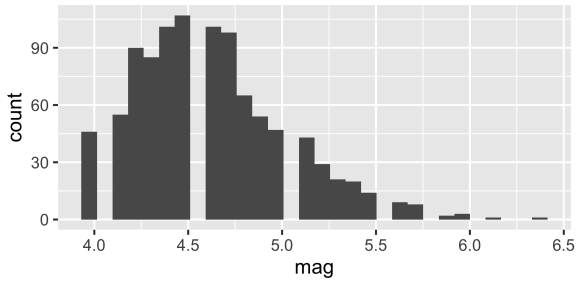
Theme

```r
msleep %>%
  ggplot() +
  aes(
    x = sleep_total,
    y = sleep_rem,
    colour = vore
  )
```

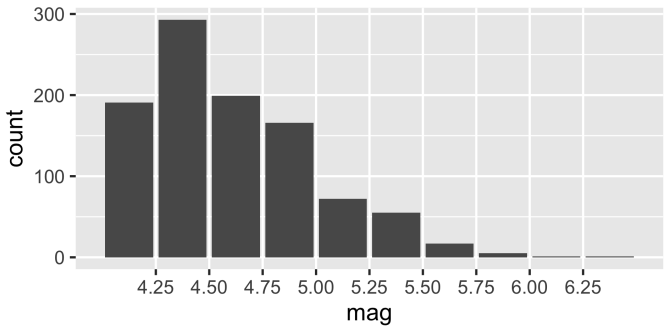
```r
msleep %>%
  ggplot() +
  aes(
    x = sleep_total,
    y = sleep_rem,
    colour = vore
  ) +
  geom_point()
```

```
geom_abline    geom_area      geom_bar       geom_bin2d     geom_blank
geom_boxplot   geom_col       geom_contour   geom_contour_filled
geom_count     geom_crossbar  geom_curve     geom_density   geom_density2d
geom_density_2d  geom_density_2d_filled  geom_density2d_filled
geom_dotplot   geom_errorbar  geom_errorbarh geom_freqpoly  geom_function
geom_hex       geom_histogram geom_hline     geom_jitter    geom_label
geom_line      geom_linerange geom_map       geom_path      geom_point
geom_pointrange  geom_polygon  geom_qq       geom_qq_line   geom_quantile
geom_raster    geom_rect      geom_ribbon    geom_rug       geom_segment
geom_smooth    geom_spoke     geom_step      geom_sf        geom_sf_label
geom_sf_text   geom_text      geom_tile      geom_violin    geom_vline
```

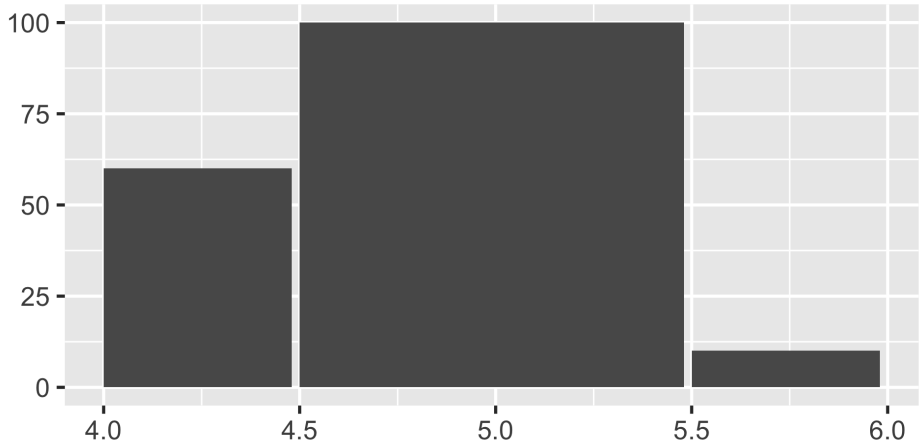
```
1 ggplot(quakes, aes(mag)) +
2   geom_histogram()
```




```
1 ggplot(quakes, aes(mag)) +
2   geom_bar() +
3   scale_x_binned()
```

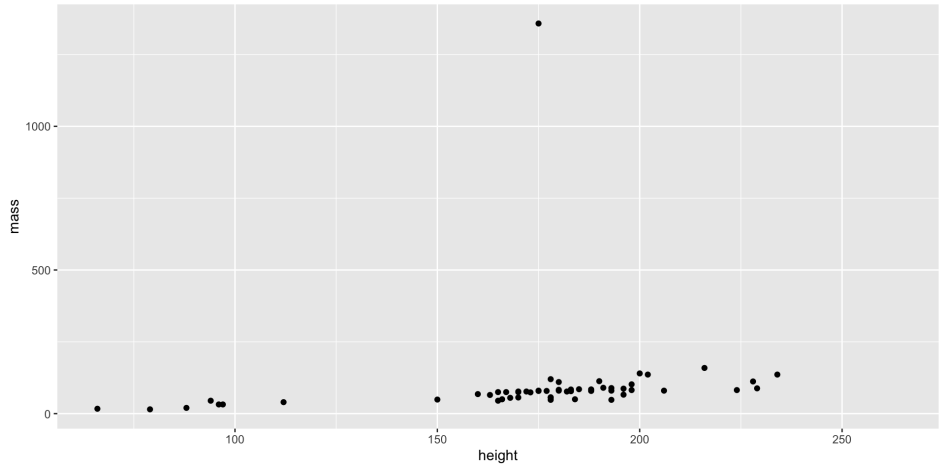


```r
rect_data <- tribble(
  ~x_min, ~x_max, ~y_min, ~y_max,
  4, 4.48, 0, 60,
  4.5, 5.48, 0, 100,
  5.5, 5.98, 0, 10
)
rect_data %>%
  ggplot() +
  geom_rect(aes(xmin = x_min,
                xmax = x_max,
                ymin = y_min,
                ymax = y_max)) +
    theme_gray(base_size = 25)
```



x y

```
starwars %>%
  ggplot() +
  aes(x = height,
      y = mass) +
  geom_point()
```

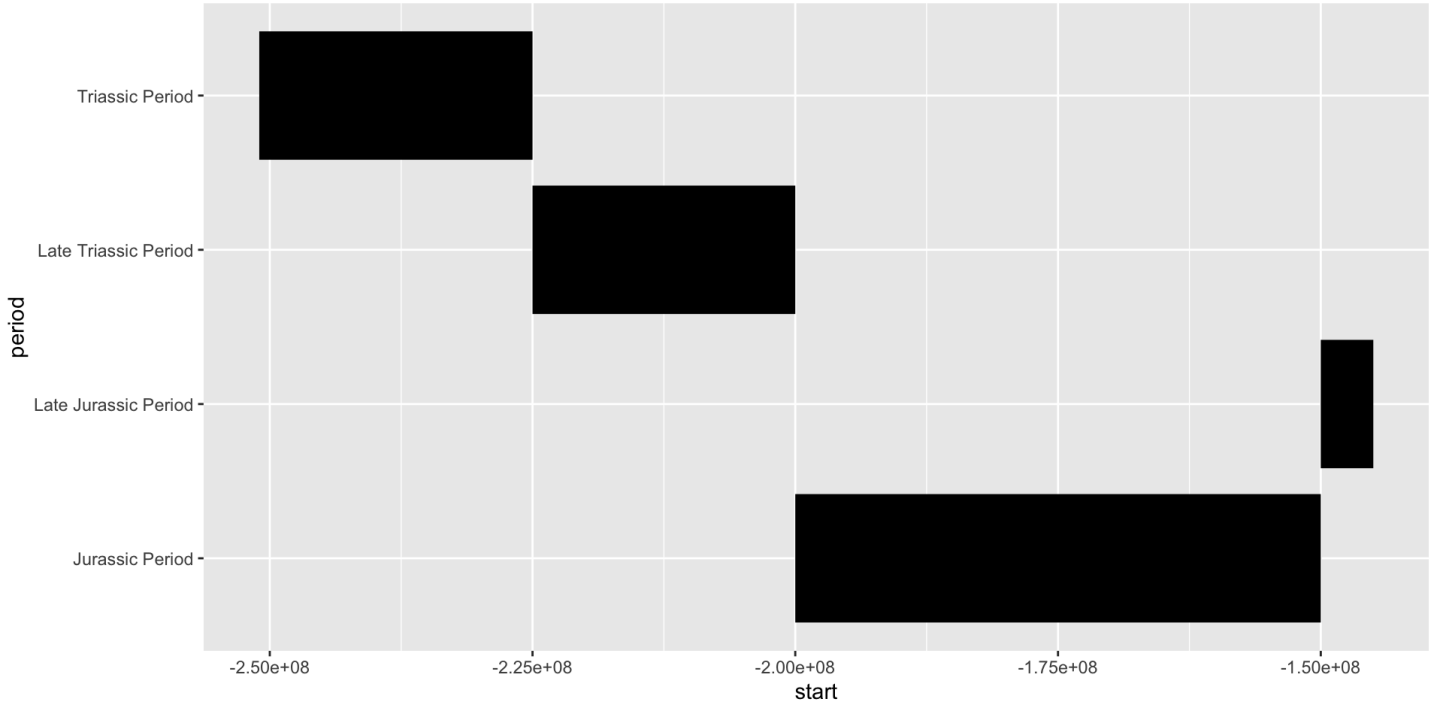


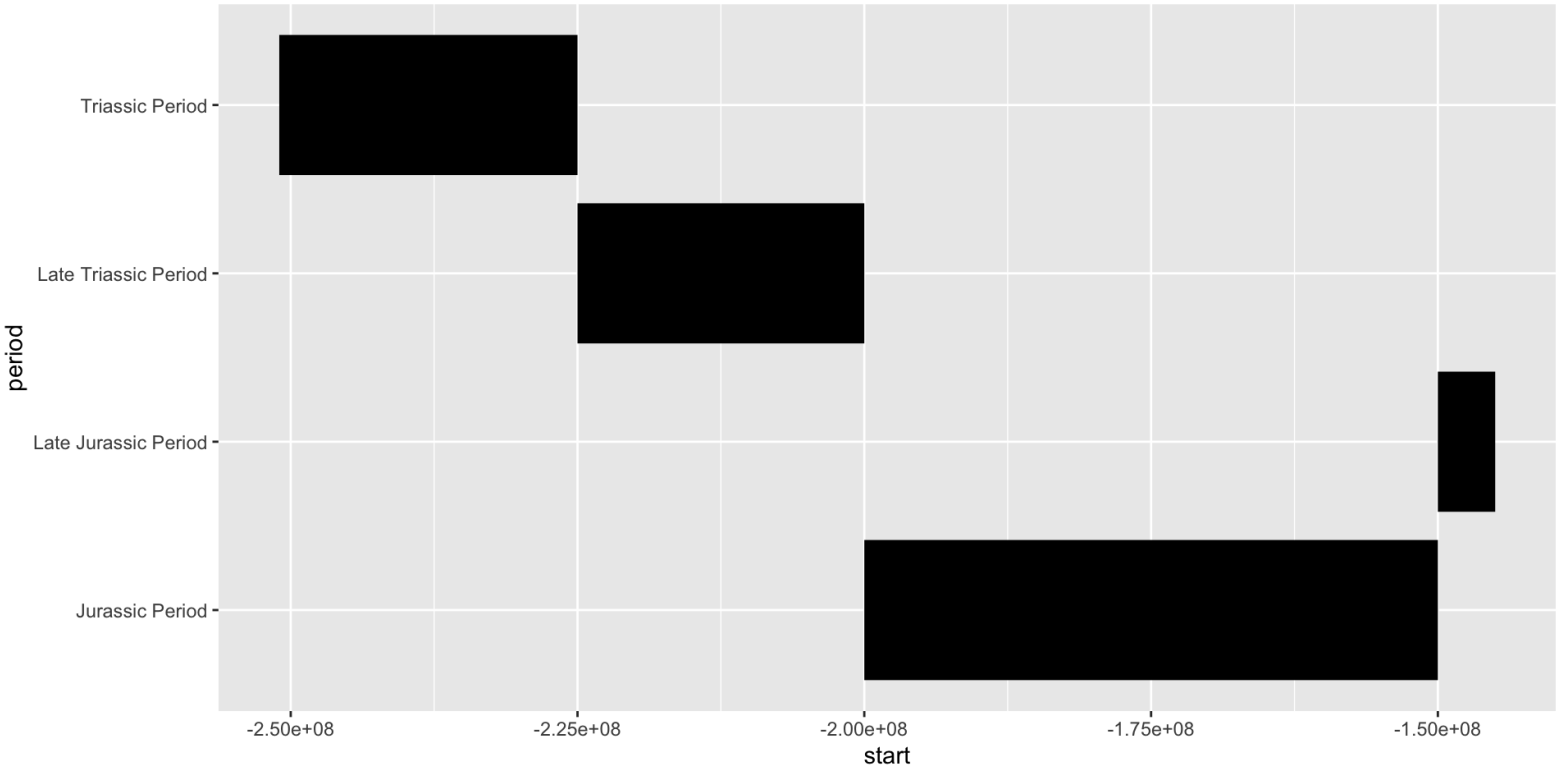
x y

```r
dinosaurs <- tribble(
  ~period, ~start, ~end,
  "Triassic Period", -251e6, -225e6,
  "Late Triassic Period", -225e6, -200e6,
  "Jurassic Period", -200e6, -150e6,
  "Late Jurassic Period", -150e6, -145e6
)
```

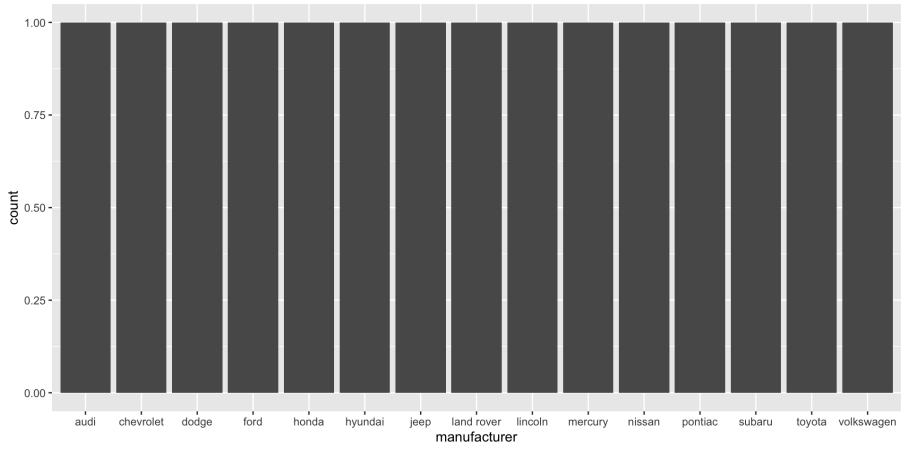
s i z e

```
dinosaurs %>%
  ggplot() +
  aes(x = start, xend = end,
      y = period, yend = period) +
  geom_segment(size = 30)
```

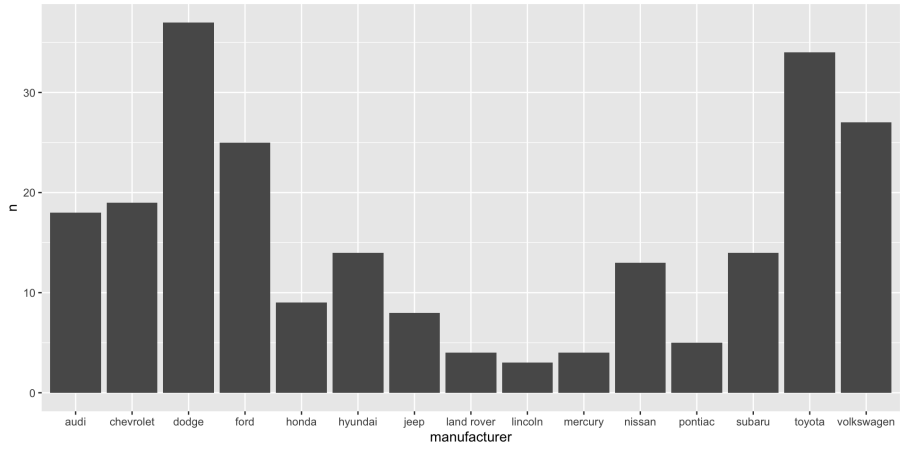



Triassic Period
Late Triassic Period
Late Jurassic Period
Jurassic Period
period
-2.50e+08
-2.25e+08
-2.00e+08
-1.75e+08
-1.50e+08
start

```r
mpg %>%
  count(manufacturer) %>%
  ggplot() +
  geom_bar(aes(manufacturer))
```

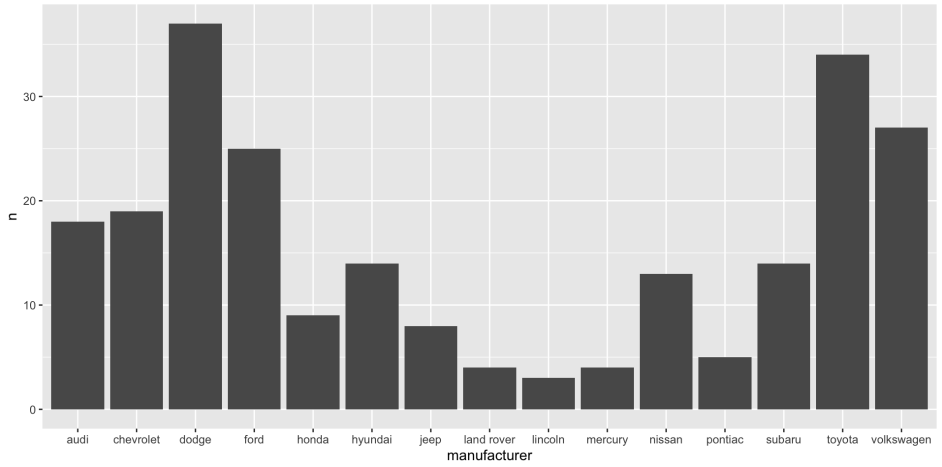



```r
mpg %>%
  count(manufacturer) %>%
  ggplot() +
  geom_col(aes(x = manufacturer, y = n))
```

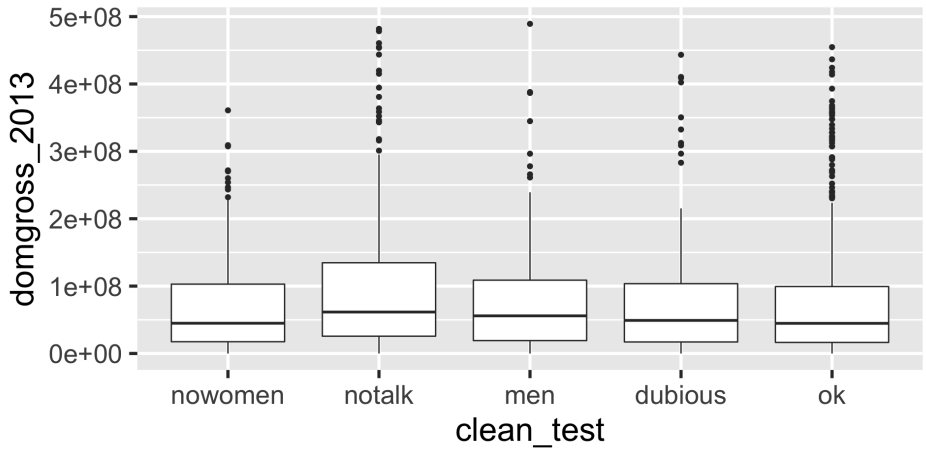


s t a t

```
mpg %>%
  count(manufacturer) %>%
  ggplot() +
  geom_bar(aes(x = manufacturer,
               y = n),
           stat = "identity")
```



```r
bechdel %>%
  filter(complete.cases(.),
         domgross_2013 < 0.5e9) %>%
  ggplot(aes(clean_test,
             domgross_2013)) +
  geom_boxplot() +
  theme_gray(base_size = 25)
```

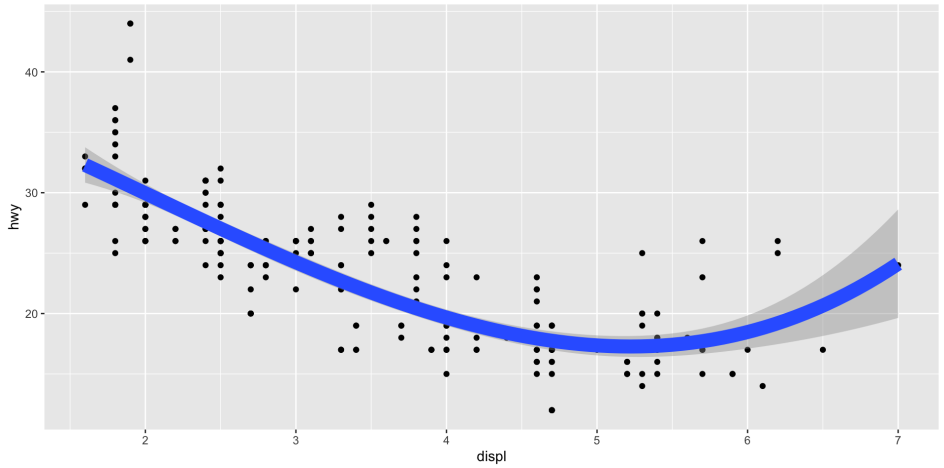


-
-
-

```r
gss_cat %>%
  count(relig, marital)
```

```
# A tibble: 78 × 3
   relig      marital           n
   <fct>      <fct>         <int>
 1 No answer  No answer         4
 2 No answer  Never married    22
 3 No answer  Separated         3
 4 No answer  Divorced         13
 5 No answer  Widowed           7
 6 No answer  Married           44
 7 Don't know Never married     6
 8 Don't know Separated         3
 9 Don't know Divorced          1
10 Don't know Married           5
# … with 68 more rows
```

```r
ggplot(mpg,
       aes(displ, hwy)) +
  geom_point() +
  geom_smooth(method = lm,
              formula = y ~ splines::bs(x, 3),
              size = 5)
```



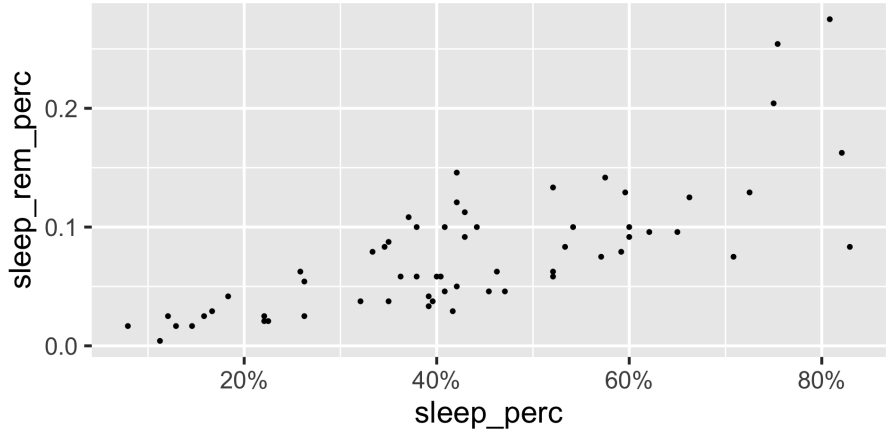
-
-

```
msleep %>%
  ggplot() +
  aes(
    x = sleep_total,
    y = sleep_rem,
    colour = vore
  ) +
  geom_point() +
  scale_colour_manual(
    values = c("carni" = "#c03728",
               "omni" = "#fd8f24",
               "insecti" = "#f5c04a",
               "herbi" = "#919c4c",
               "NA" = "#e68c7c")
  )
```

```r
percent(c(0.3, 0.5, 0.6))
```

```
[1] "30%" "50%" "60%"
```

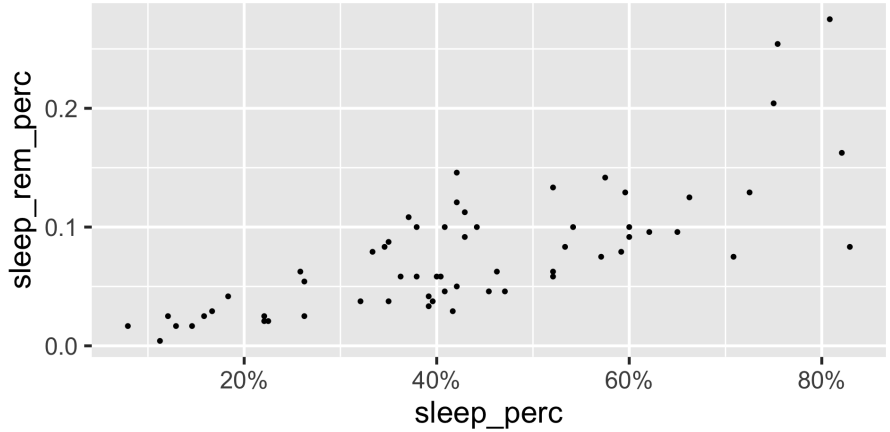
```r
msleep %>%
  mutate(sleep_perc = sleep_total / 24,
         sleep_rem_perc = sleep_rem / 24) %>%
  ggplot() +
  aes(x = sleep_perc,
      y = sleep_rem_perc) +
  geom_point() +
  scale_x_continuous(label = percent_format()) +
  theme_gray(base_size = 24)
```

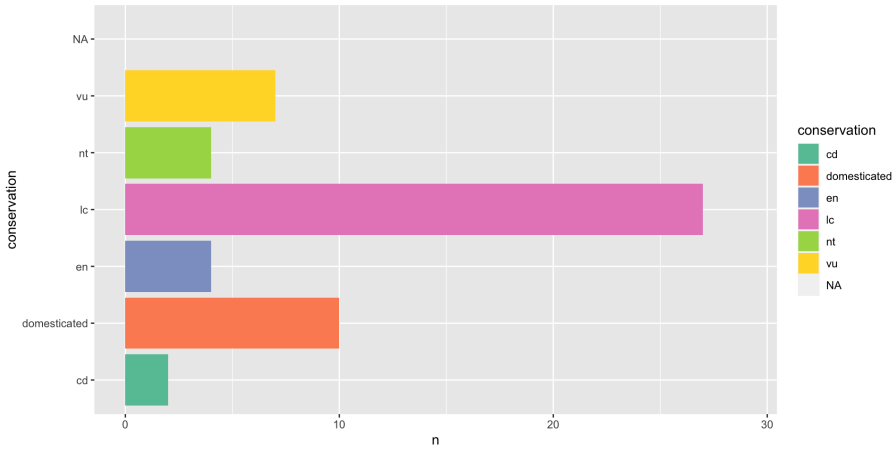


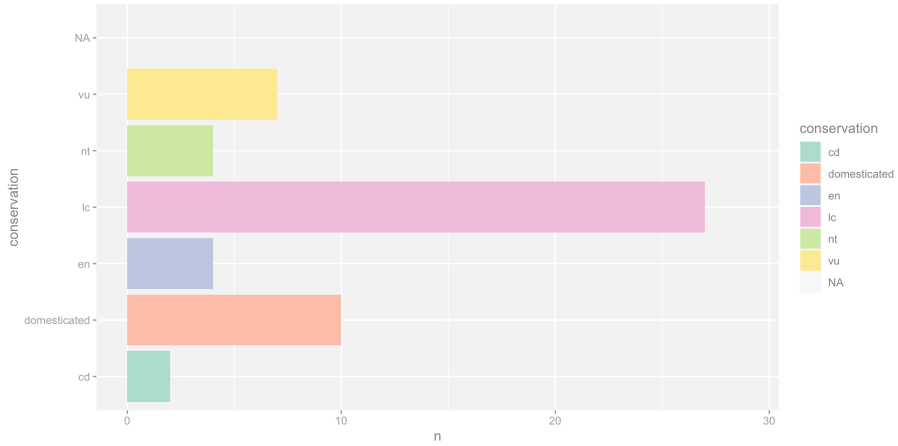
```r
label_percent()(c(0.3, 0.5, 0.6))
```

```
[1] "30%" "50%" "60%"
```

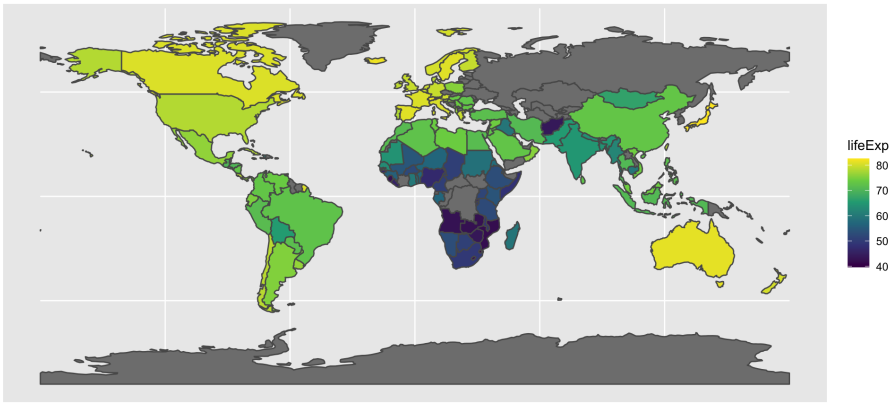
```r
msleep %>%
  mutate(sleep_perc = sleep_total / 24,
         sleep_rem_perc = sleep_rem / 24) %>%
  ggplot() +
  aes(x = sleep_perc,
      y = sleep_rem_perc) +
  geom_point() +
  scale_x_continuous(label = label_percent()) +
  theme_gray(base_size = 24)
```



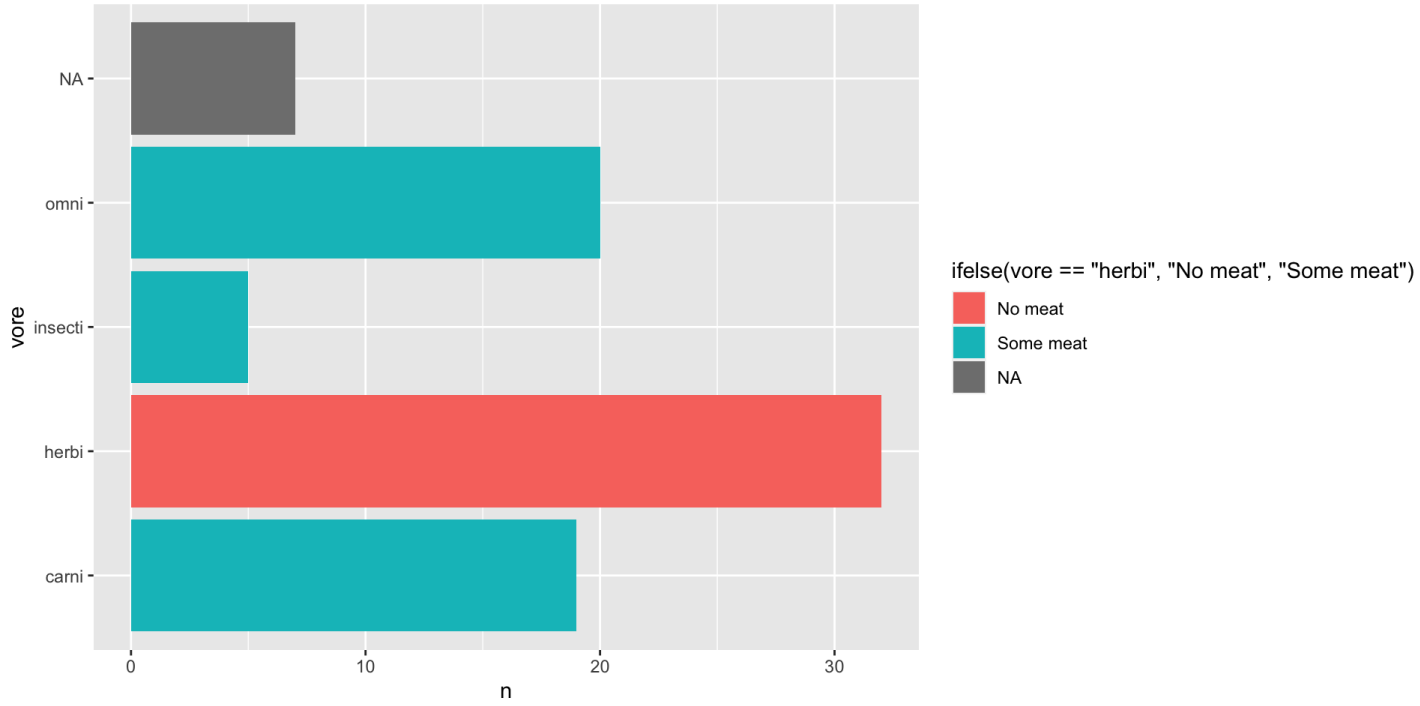

conservation
NA
vu
nt
lc
en
domesticated
cd
0
10
20
30
n
conservation
cd
domesticated
en
lc
nt
vu
NA



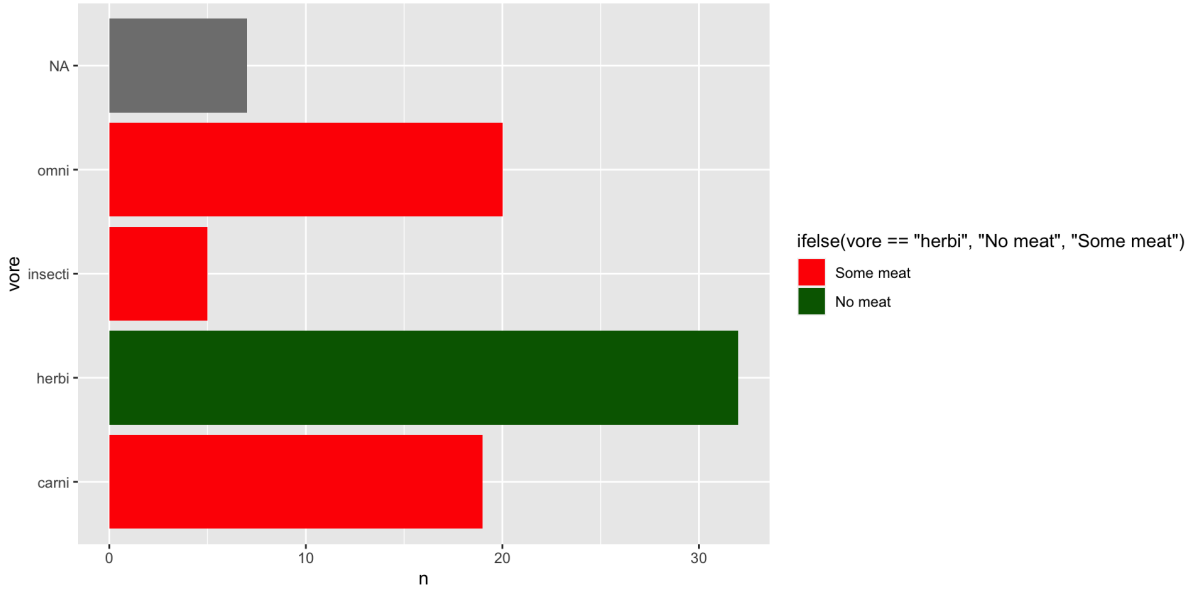

```
countries110 %>%
  st_as_sf() %>%
  left_join(filter(gapminder, year == 2007),
            by = c("name" = "country")) %>%
  ggplot() +
  geom_sf(aes(fill = lifeExp)) +
  scale_fill_viridis_c()
```



```r
msleep %>%
  count(vore) %>%
  ggplot() +
  aes(x = n,
      y = vore,
      fill = ifelse(vore == "herbi", "No meat", "Some meat")) +
  geom_col()
```



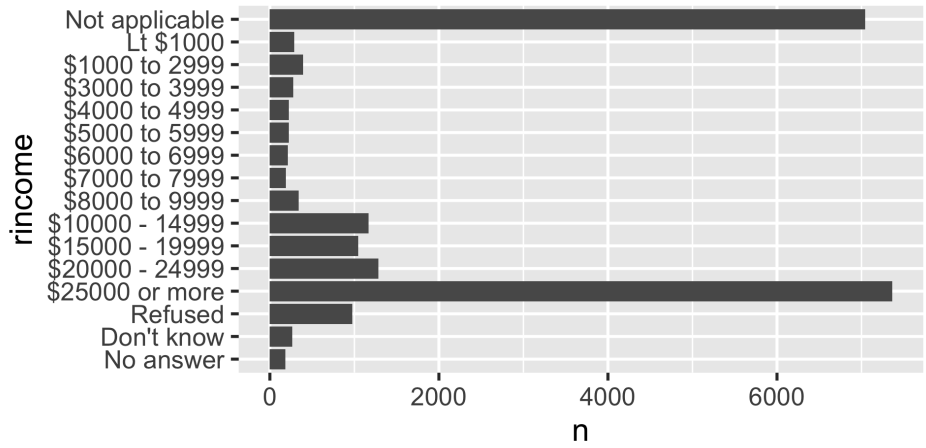
```r
msleep %>%
  count(vore) %>%
  ggplot() +
  aes(x = n,
      y = vore,
      fill = ifelse(vore == "herbi", "No meat", "Some meat")) +
  geom_col() +
  scale_fill_manual(values = c("Some meat" = "red",
                               "No meat" = "darkgreen"))
```



```r
gss_cat %>%
  head() %>%
  pull(rincome)
```

```
[1] $8000 to 9999  $8000 to 9999  Not applicable Not applicable Not applicable
[6] $20000 - 24999
16 Levels: No answer Don't know Refused $25000 or more ... Not applicable
```

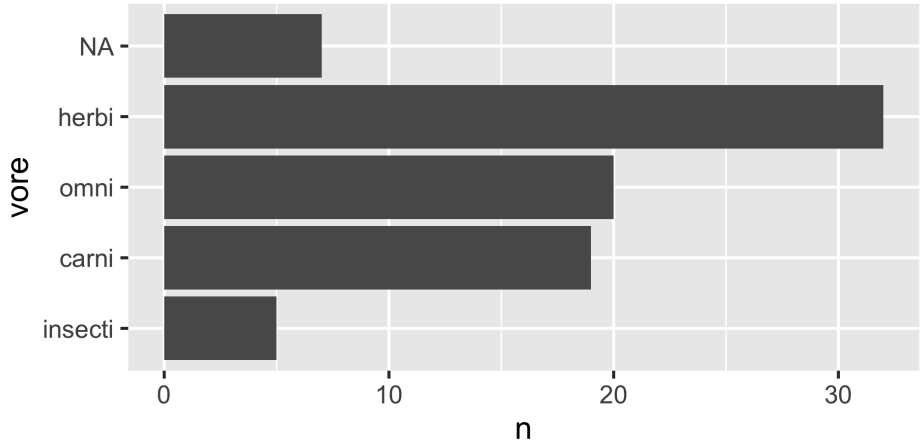
```r
gss_cat %>%
  count(rincome) %>%
  ggplot() +
  aes(x = n,
      y = rincome) +
  geom_col() +
  theme_gray(base_size = 24)
```



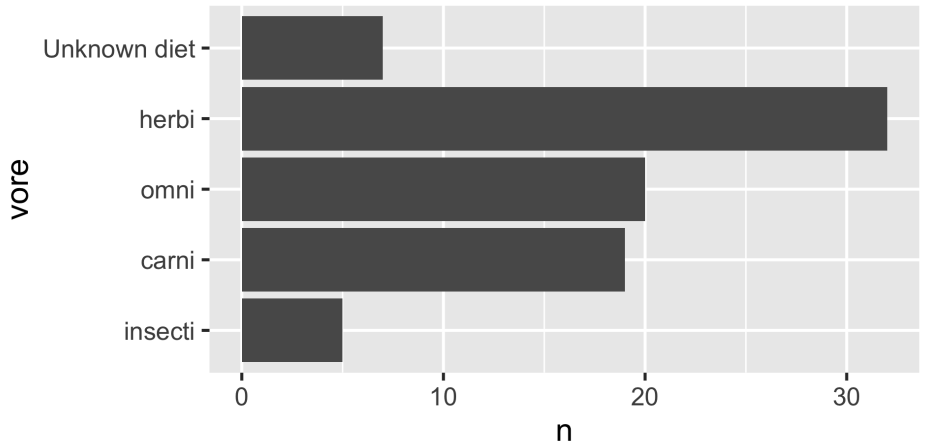
```r
msleep %>%
  count(vore)
```

```
# A tibble: 5 × 2
  vore        n
  <chr>   <int>
1 carni      19
2 herbi      32
3 insecti     5
4 omni       20
5 <NA>        7
```

```r
msleep %>%
  count(vore) %>%
  mutate(vore = fct_reorder(vore, n)) %>%
  ggplot() +
  aes(x = n,
      y = vore) +
  geom_col() +
  theme_gray(base_size = 24)
```

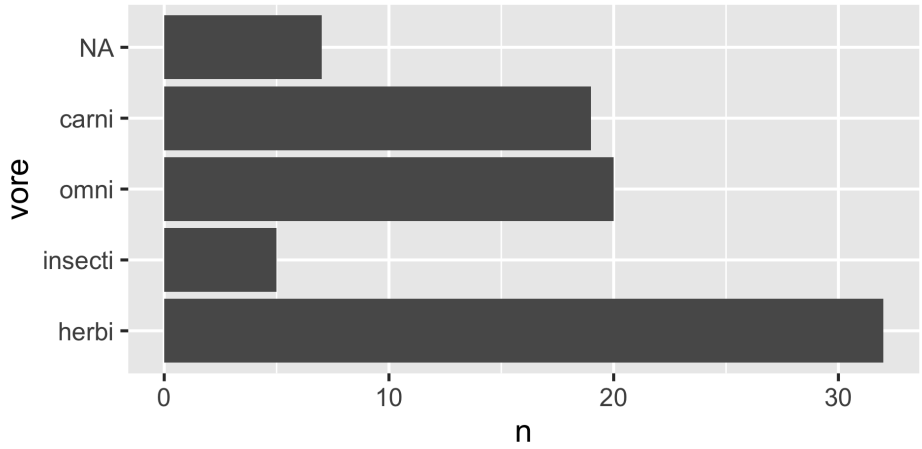


```r
msleep %>%
  count(vore) %>%
  mutate(vore = fct_reorder(vore, n),
         vore = fct_explicit_na(vore, "Unknown diet")) %>%
  ggplot() +
  aes(x = n,
      y = vore) +
  geom_col() +
  theme_gray(base_size = 24)
```

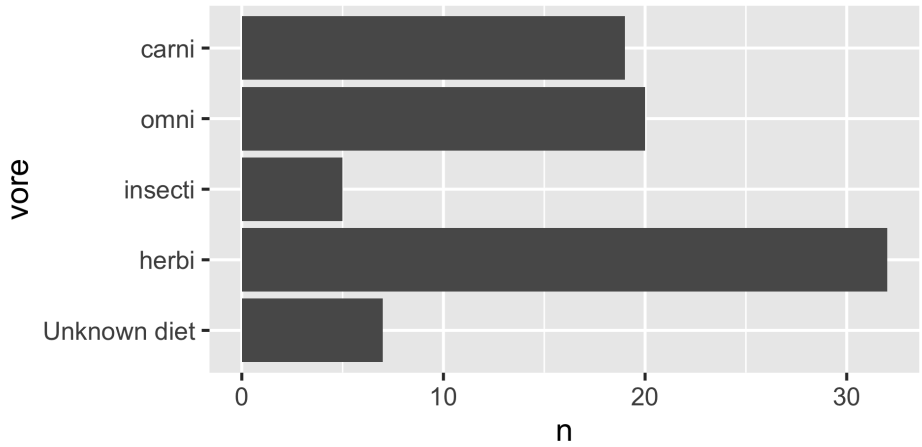


```r
order_vore <- c("carni", "omni", "insecti", "herbi")

msleep %>%
  count(vore) %>%
  mutate(vore = fct_relevel(vore, order_vore),
         vore = fct_rev(vore)) %>%
  ggplot() +
  aes(x = n,
      y = vore) +
  geom_col() +
  theme_gray(base_size = 24)
```



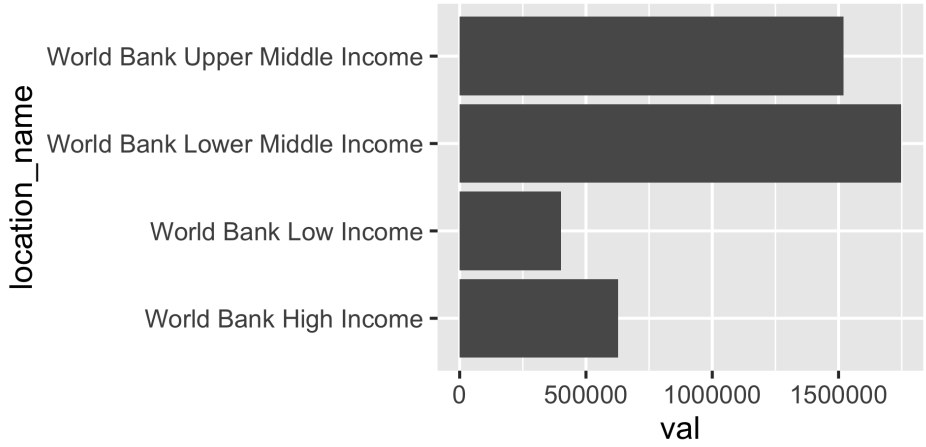
```r
msleep %>%
  count(vore) %>%
  mutate(vore = fct_relevel(vore, order_vore),
         vore = fct_rev(vore),
         vore = fct_explicit_na(vore, "Unknown diet"),
         vore = fct_relevel(vore, "Unknown diet", afte
  ggplot() +
  aes(x = n,
      y = vore) +
  geom_col() +
  theme_gray(base_size = 24)
```



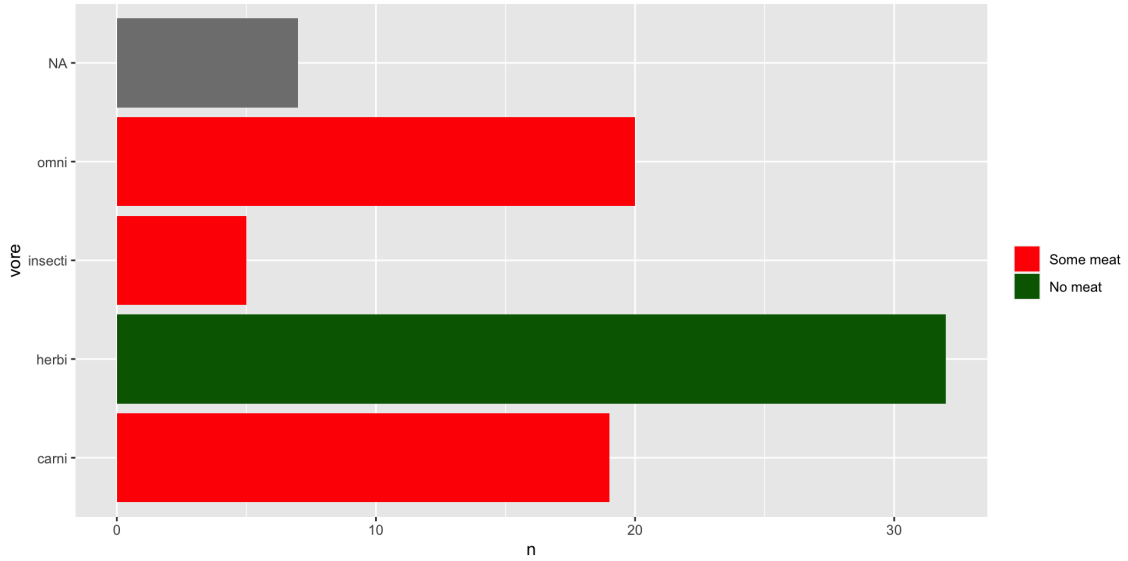
```
download.file("https://raw.githubusercontent.com/charliejhadley/eng7218_data-science-for-healthcare-applications_bcu-
              destfile = "data/global-burden-of-disease-data.csv")
```

-
-

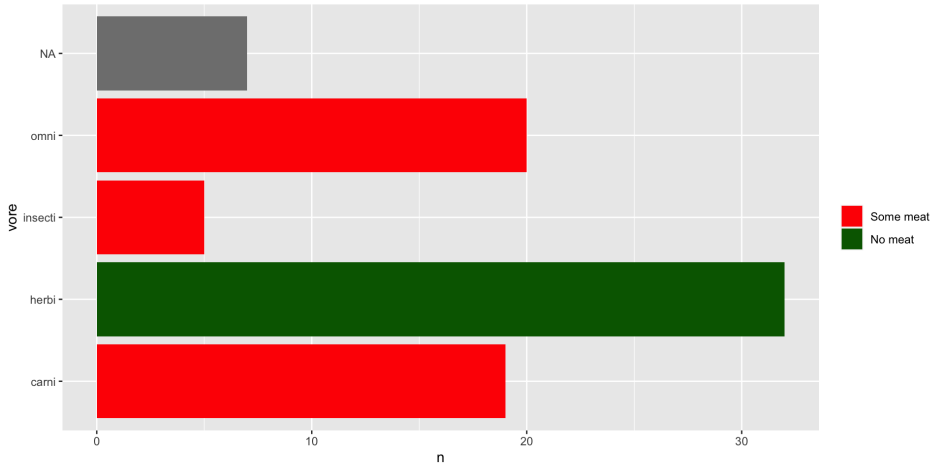
```
gdb_injuries %>%
  ggplot() +
  aes(x = val,
      y = location_name) +
  geom_col() +
  theme_gray(base_size = 24)
```



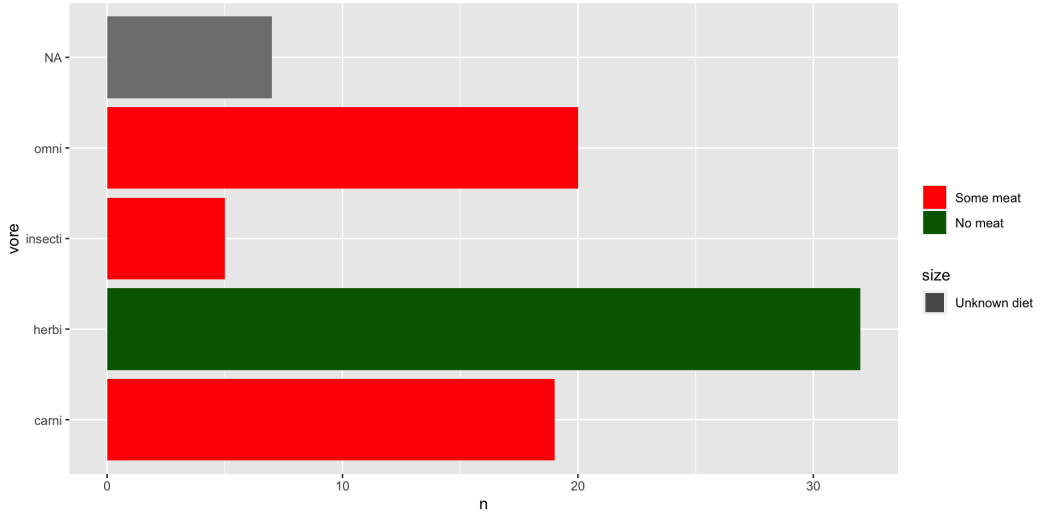
```r
ggplot() +
  geom_line(show.legend = FALSE) +
  guides(alpha = guide_legend()) +
  theme(legend.position = "bottom")
```



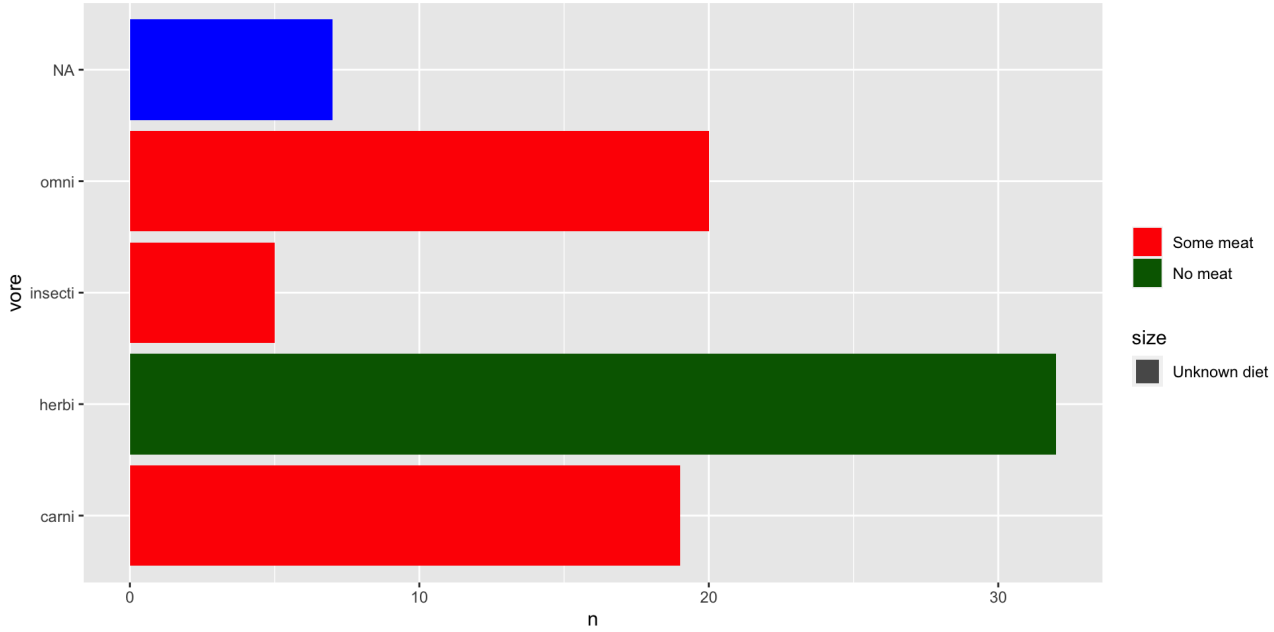
```r
msleep %>%
  count(vore) %>%
  ggplot() +
  aes(x = n,
      y = vore,
      fill = ifelse(vore == "herbi",
                    "No meat",
                    "Some meat")) +
  geom_col() +
  scale_fill_manual(values = c("Some meat" = "red",
                               "No meat" = "darkgreen"
                    name = "")
```



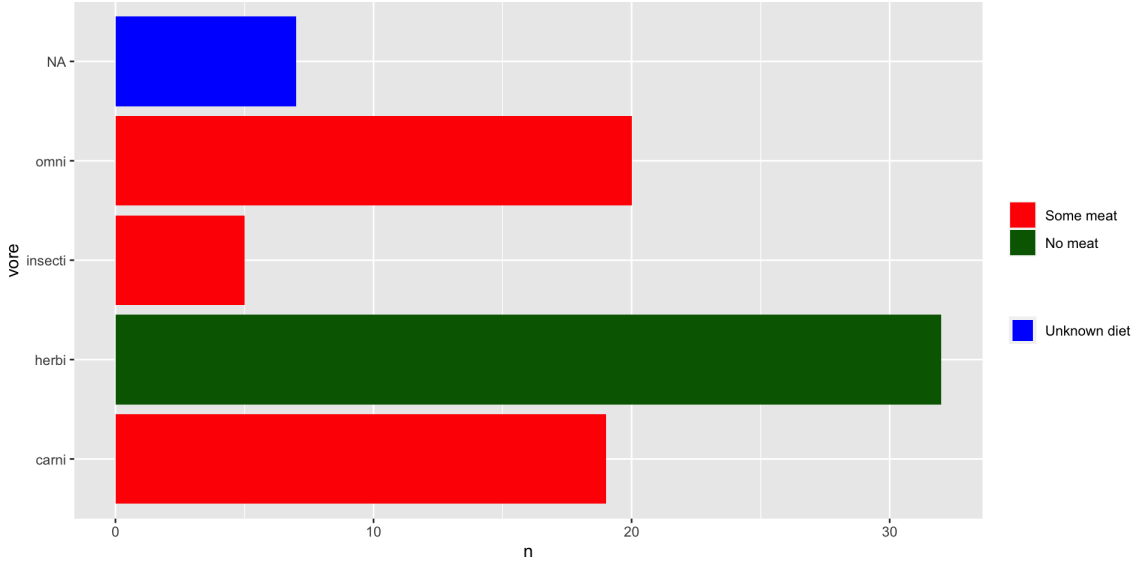
```r
msleep %>%
  count(vore) %>%
  ggplot() +
  aes(x = n,
      y = vore,
      fill = ifelse(vore == "herbi", "No meat", "Some meat")) +
  geom_col(aes(size = "Unknown diet")) +
  scale_fill_manual(values = c("Some meat" = "red",
                               "No meat" = "darkgreen"),
                    name = "")
```



```r
msleep %>%
  count(vore) %>%
  ggplot() +
  aes(x = n,
      y = vore,
      fill = ifelse(vore == "herbi", "No meat", "Some meat")) +
  geom_col(aes(size = "Unknown diet")) +
  scale_fill_manual(values = c("Some meat" = "red",
                               "No meat" = "darkgreen"),
                    name = "",
                    na.value = "blue")
```



```r
msleep %>%
  count(vore) %>%
  ggplot() +
  aes(x = n,
      y = vore,
      fill = ifelse(vore == "herbi", "No meat", "Some meat")) +
  geom_col(aes(size = "Unknown diet")) +
  scale_fill_manual(values = c("Some meat" = "red",
                               "No meat" = "darkgreen"),
                    name = "",
                    na.value = "blue") +
  guides(size = guide_legend(title = "",
                             override.aes = list(fill = "blue")))
```



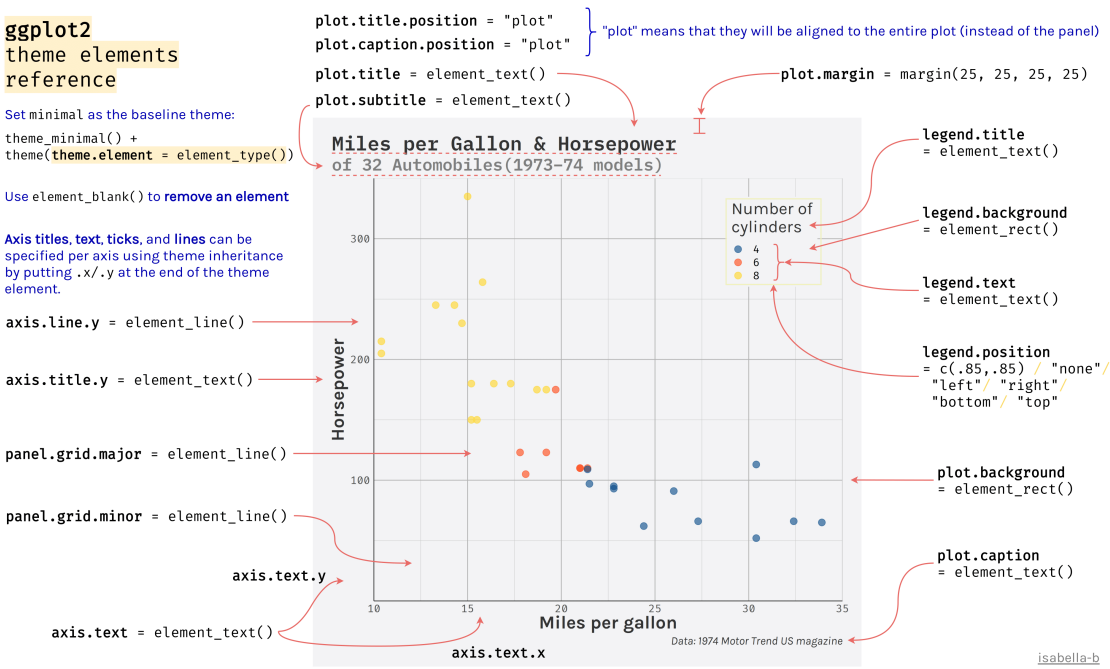
ggplot2 theme elements reference

Set `minimal` as the baseline theme:

```
theme_minimal() +
theme(theme.element = element_type())
```

Use `element_blank()` to **remove an element**

Axis titles, **text**, **ticks**, and **lines** can be specified per axis using theme inheritance by putting `.x`/`.y` at the end of the theme element.

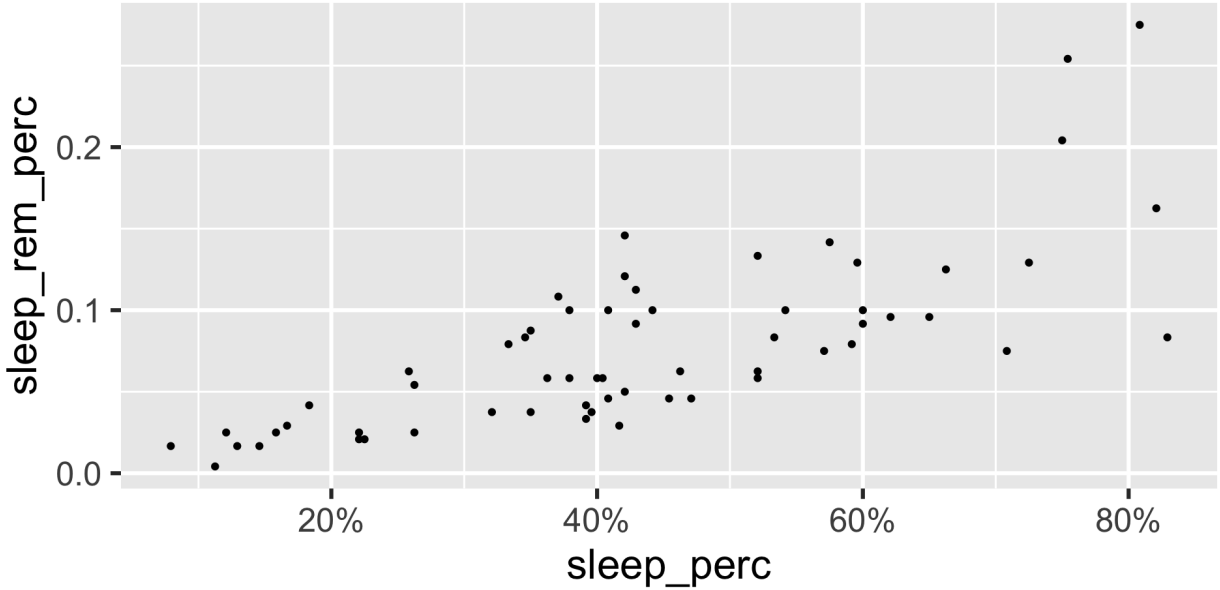


isabella-b

`text` = `element_text()` ← modifications will be applied to all text elements

Full list of elements at ggplot2.tidyverse.org/reference/theme

```r
msleep %>%
  mutate(sleep_perc = sleep_total / 24,
         sleep_rem_perc = sleep_rem / 24) %>%
  ggplot() +
  aes(x = sleep_perc,
      y = sleep_rem_perc) +
  geom_point() +
  scale_x_continuous(label = label_percent()) +
  theme_gray(base_size = 24)
```



```r
theme_fivethirtyeight() +
  theme(panel.grid.major = element_line(colour = "red"))
```

• • •

References

Management

Revive Quarterly

The American Statistician 27

Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems

On the mode of communication of cholera

Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army

The best stats you've ever seen

EuroVis 2016 - Short Papers

Journal of the American Statistical Association 79

Proceedings of the 28th international conference on Human factors in computing systems - CHI '01

ISPRS International Journal of Geo-Information 10

IEEE Computer Graphics and Applications 42

Twitter