ggplot2 ecosystem & designing visualizations

Lecture 11

Dr. Colin Rundel

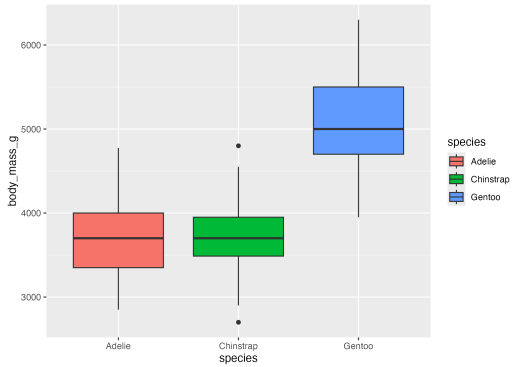
The wider ggplot2 ecosystem

ggthemes

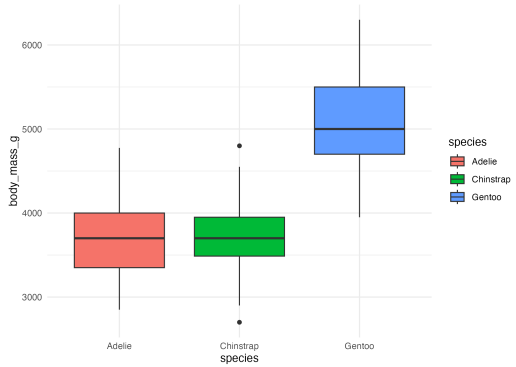
ggplot2 themes

```
g = ggplot( palmerpenguins::penguins, aes(x=species, y=body_mass_g, fill=species))
    geom_boxplot()
```

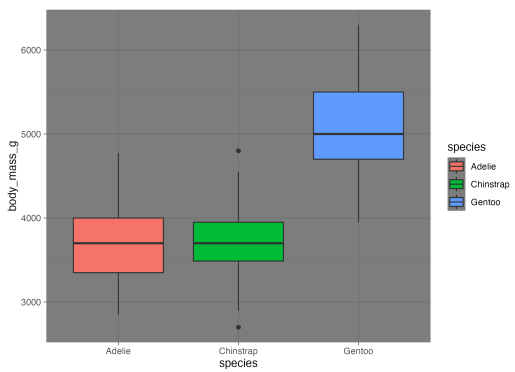
```
1 g
```



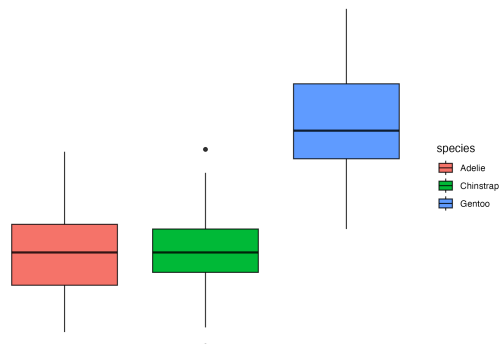

```
1 g + theme_minimal()
```




```
1 g + theme_dark()
```

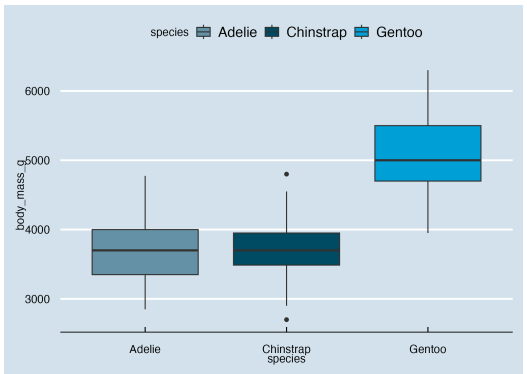



```
1 g + theme_void()
```

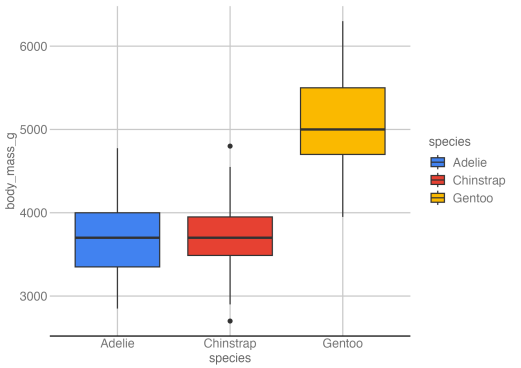


ggthemes

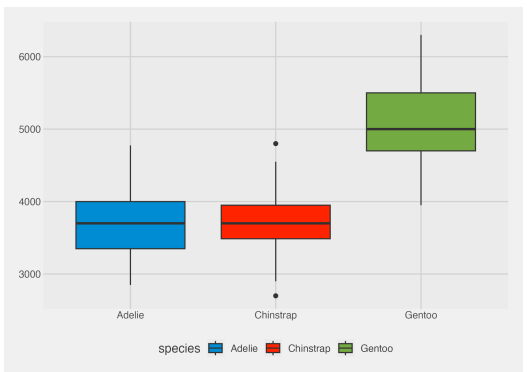
```
g + ggthemes::theme_economist() +
  ggthemes::scale_fill_economist()
```



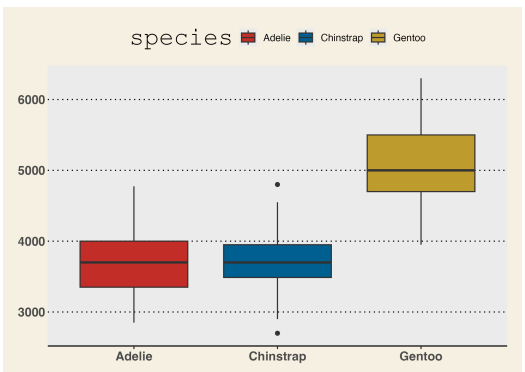

```
g + ggthemes::theme_gdocs() +
  ggthemes::scale_fill_gdocs()
```




```
g + ggthemes::theme_fivethirtyeight()
  ggthemes::scale_fill_fivethirtyeight
```

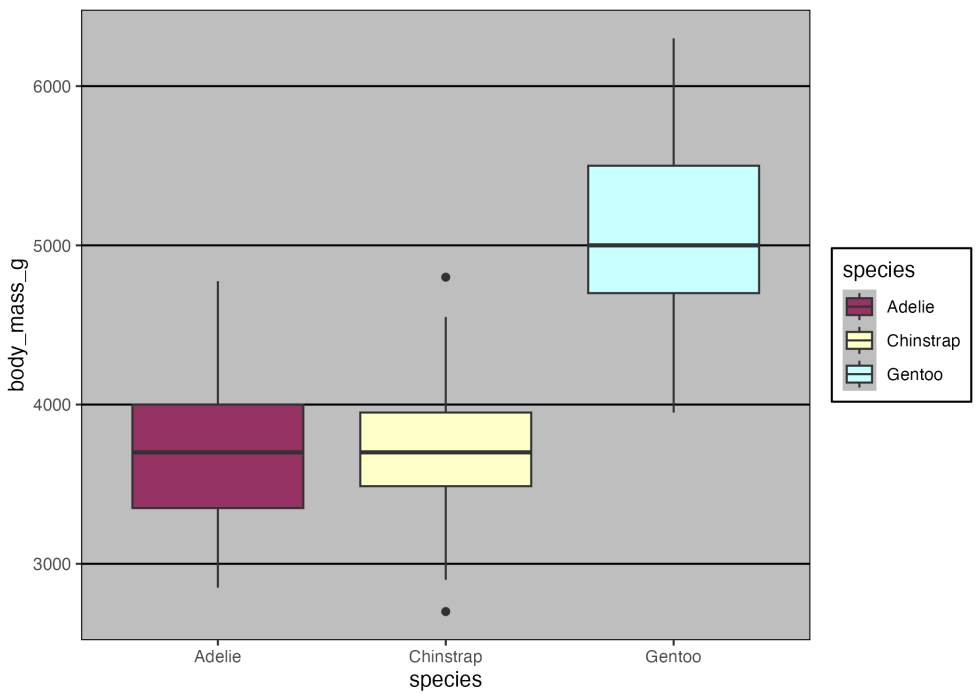



```
g + ggthemes::theme_wsj() +
  ggthemes::scale_fill_wsj()
```

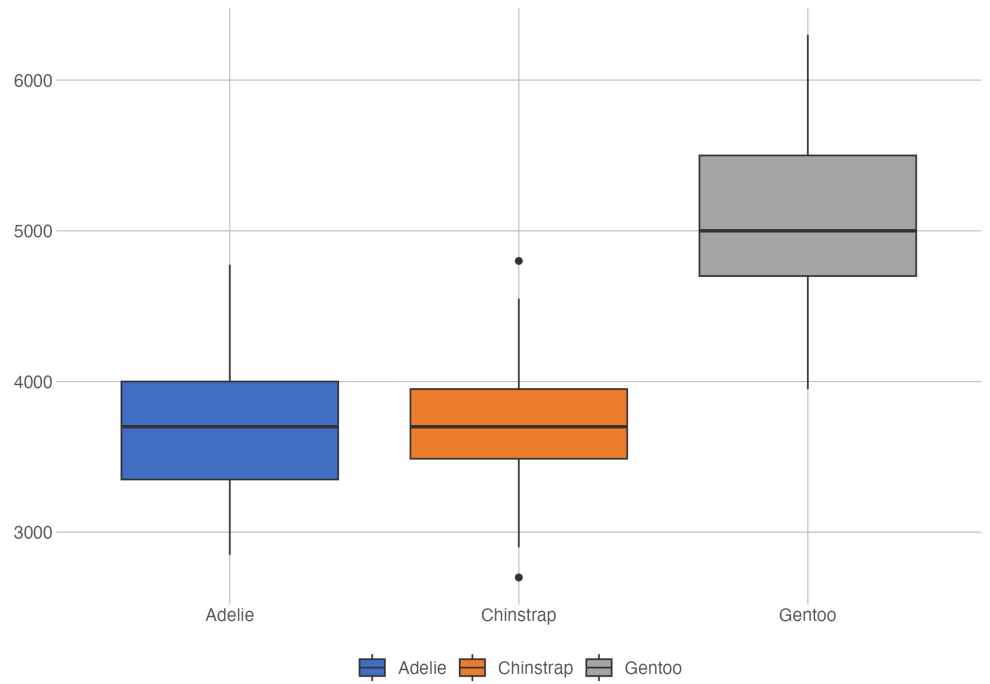


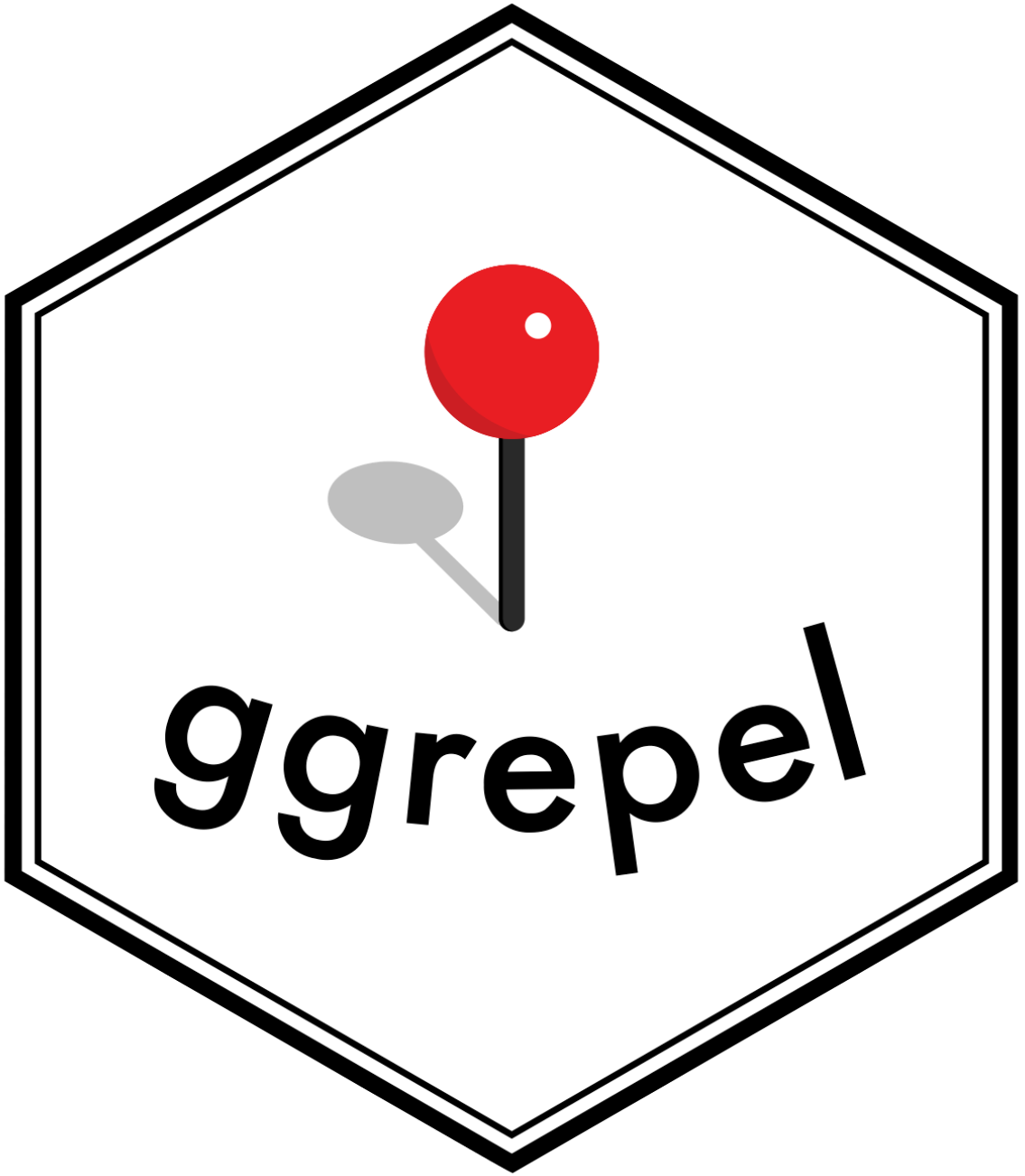
And for those who miss Excel

```
g + ggthemes::theme_excel() +
  ggthemes::scale_fill_excel()
```



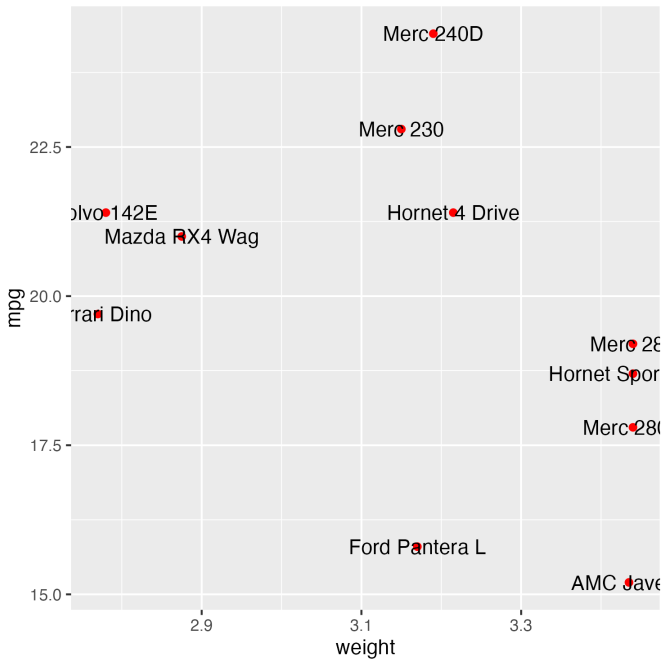

```
g + ggthemes::theme_excel_new() +
  ggthemes::scale_fill_excel_new()
```



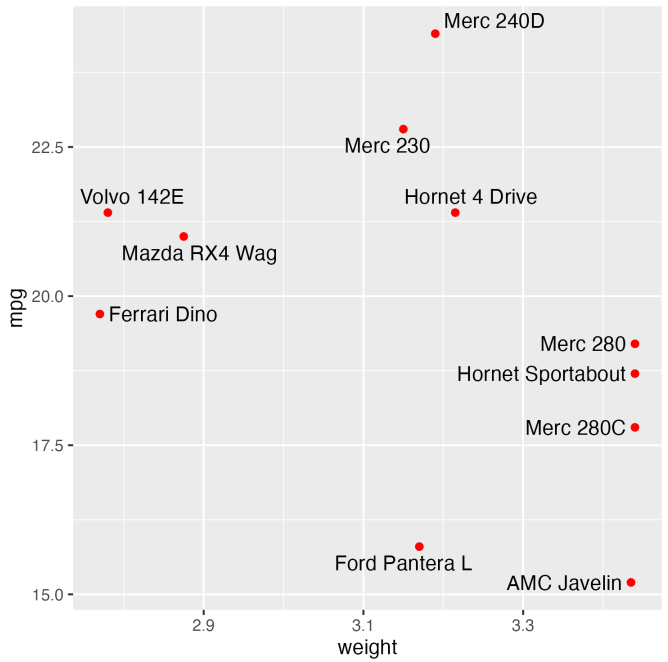

ggrepel

```r
d = tibble(
  car = rownames(mtcars),
  weight = mtcars$wt,
  mpg = mtcars$mpg
) |>
  filter(weight > 2.75, weight < 3.45)
```

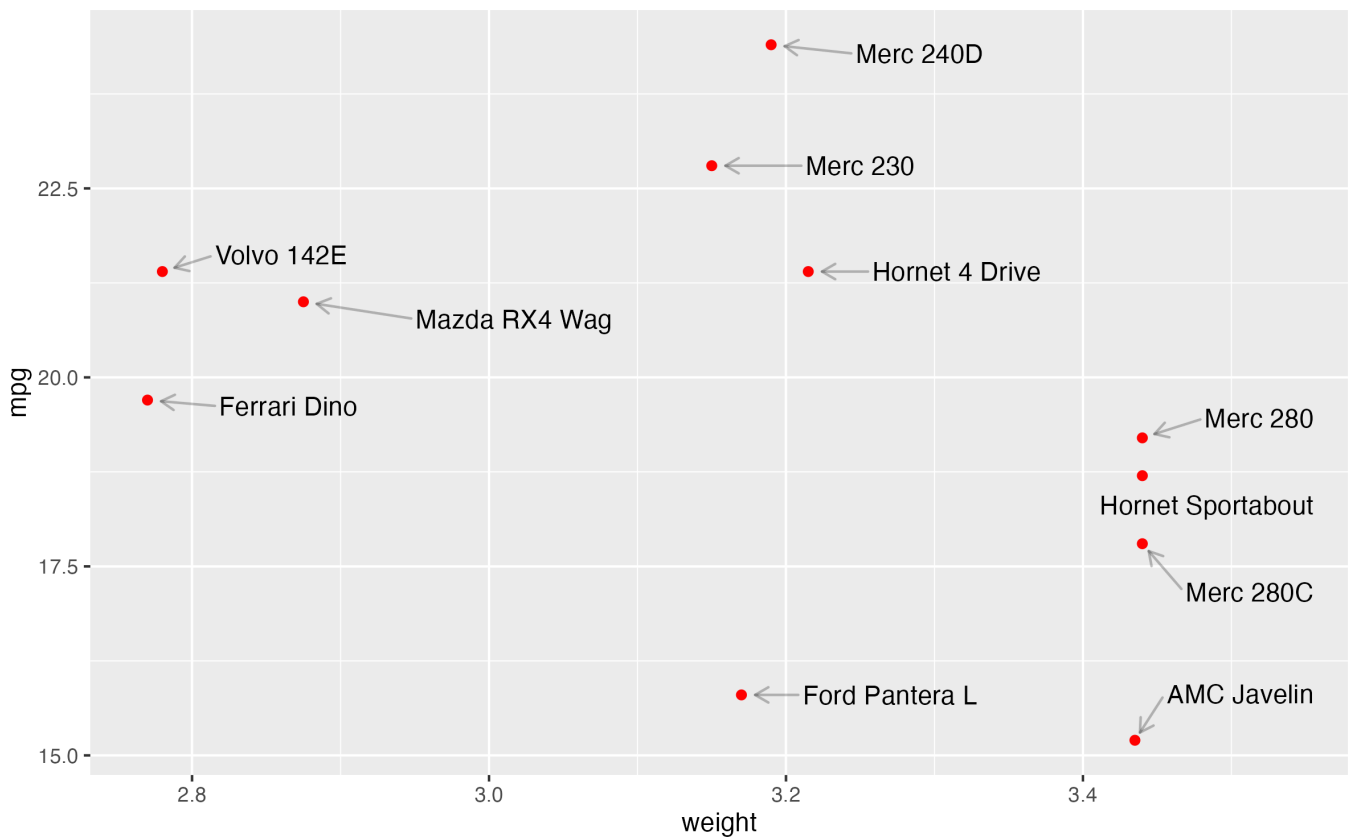
```r
ggplot(d, aes(x=weight, y=mpg)) +
  geom_point(color="red") +
  geom_text(
    aes(label = car)
  )
```




```r
ggplot(d, aes(x=weight, y=mpg)) +
  geom_point(color="red") +
  ggrepel::geom_text_repel(
    aes(label = car)
  )
```



```r
ggplot(d, aes(x=weight, y=mpg)) +
  geom_point(color="red") +
  ggrepel::geom_text_repel(
    aes(label = car),
    nudge_x = .1, box.padding = 1, point.padding = 0.6,
    arrow = arrow(length = unit(0.02, "npc")), segment.alpha = 0.25
  )
```




PATCH
WORK

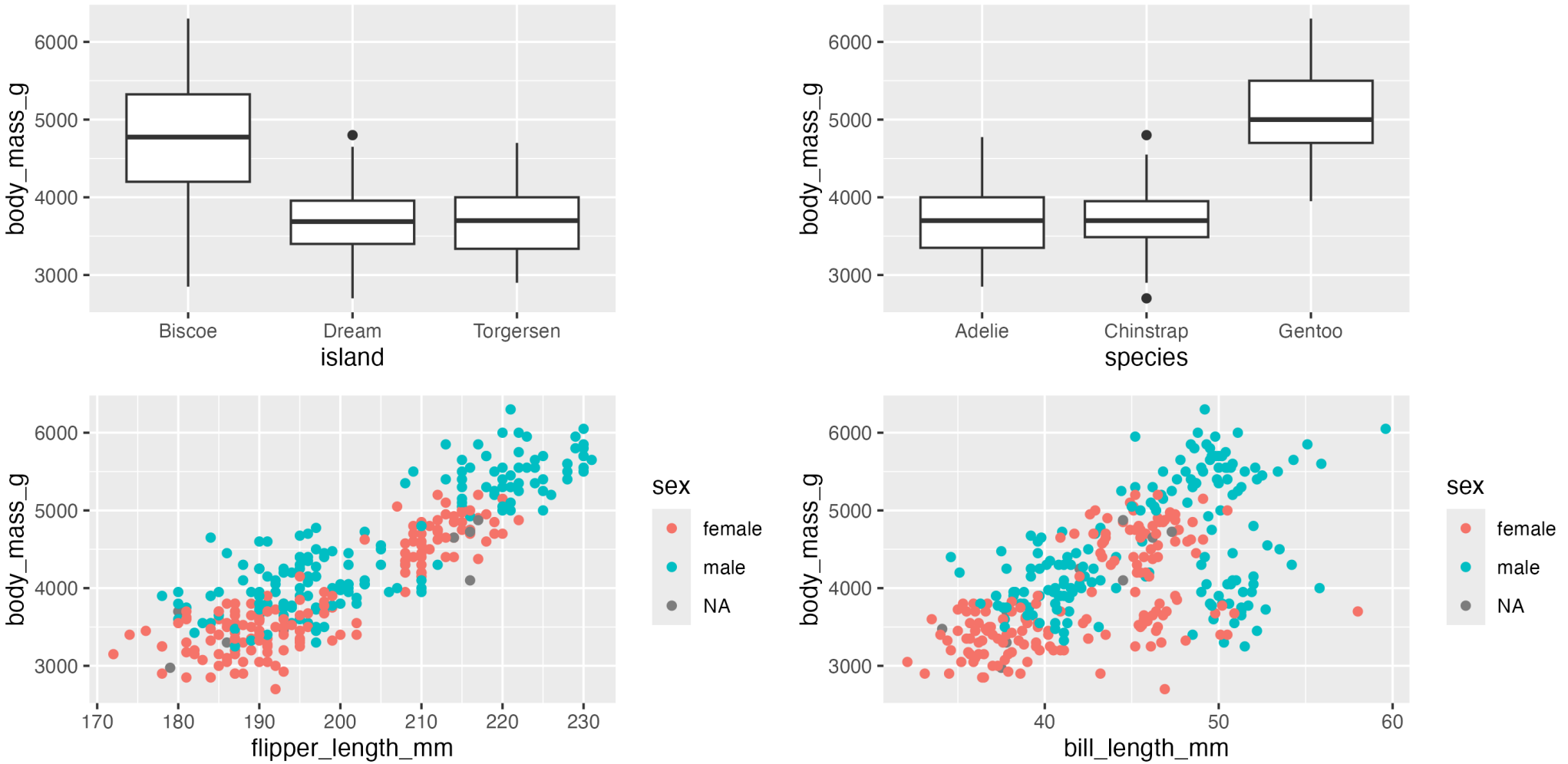
ggplot objects

```
library(patchwork)

p1 = ggplot(palmerpenguins::penguins) +
  geom_boxplot(aes(x = island, y = body_mass_g))

p2 = ggplot(palmerpenguins::penguins) +
  geom_boxplot(aes(x = species, y = body_mass_g))

p3 = ggplot(palmerpenguins::penguins) +
  geom_point(aes(x = flipper_length_mm, y = body_mass_g, color = sex

p4 = ggplot(palmerpenguins::penguins) +
  geom_point(aes(x = bill_length_mm, y = body_mass_g, color = sex))
```

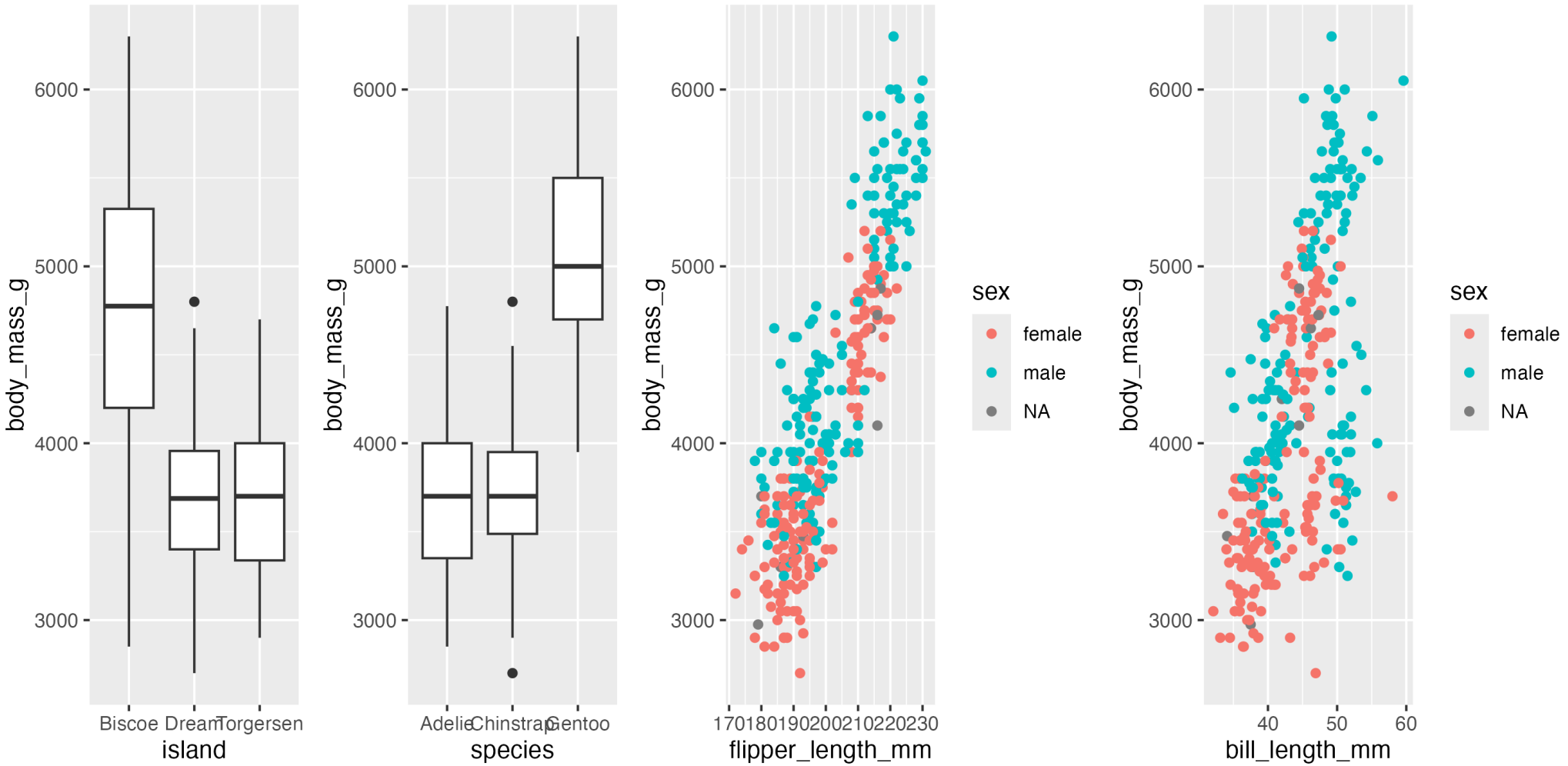
```
class(p1)
```

```
[1] "ggplot2::ggplot" "ggplot"          "ggplot2::gg"     "S7_object"
[5] "gg"
```

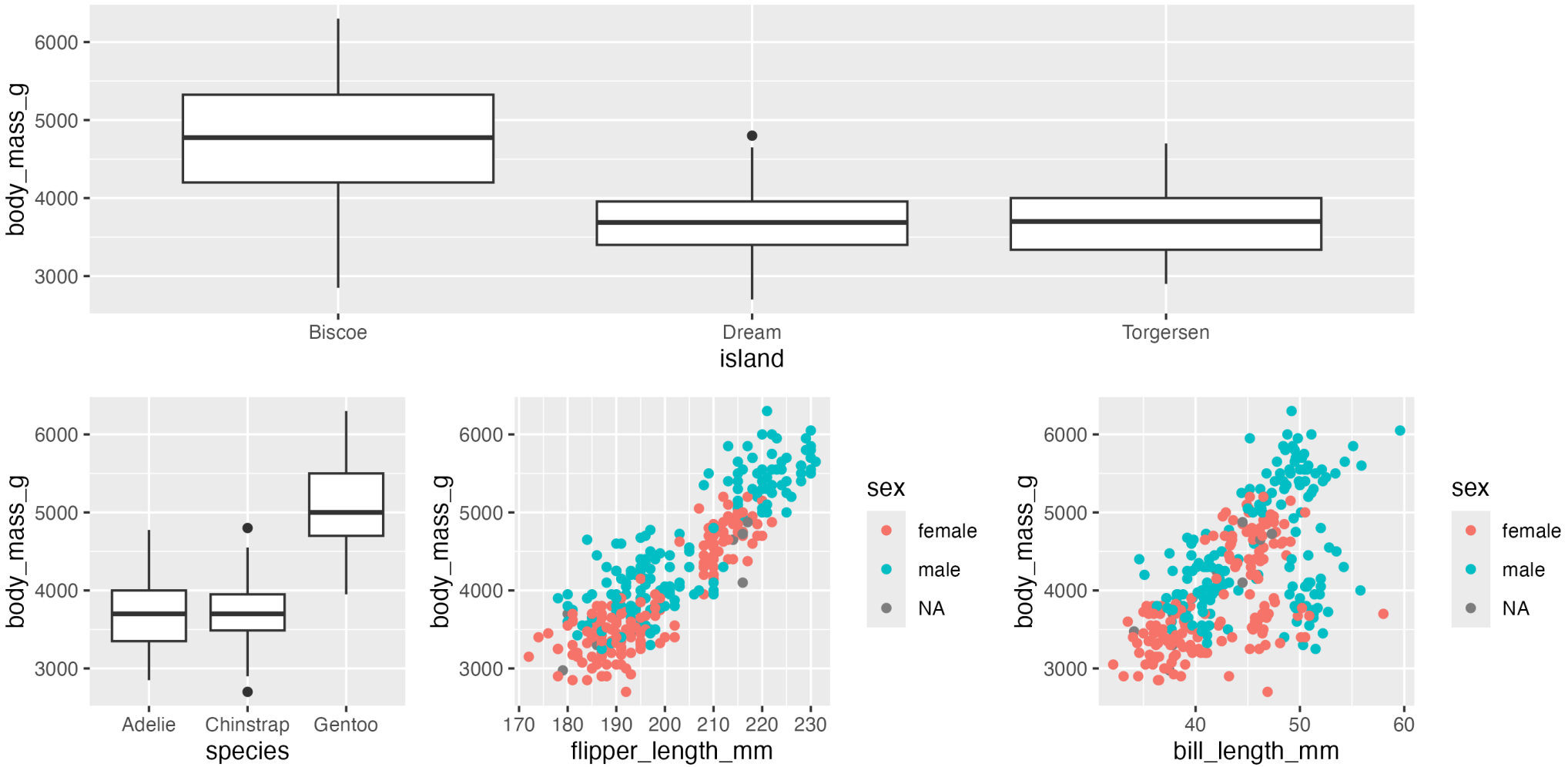
```
1 p1 + p2 + p3 + p4
```


body_mass_g
6000
5000
4000
3000
Biscoe
Dream
Torgersen
island
Adelie
Chinstrap
Gentoo
species
170
180
190
200
210
220
230
flipper_length_mm
40
50
60
bill_length_mm
sex
female
male
NA

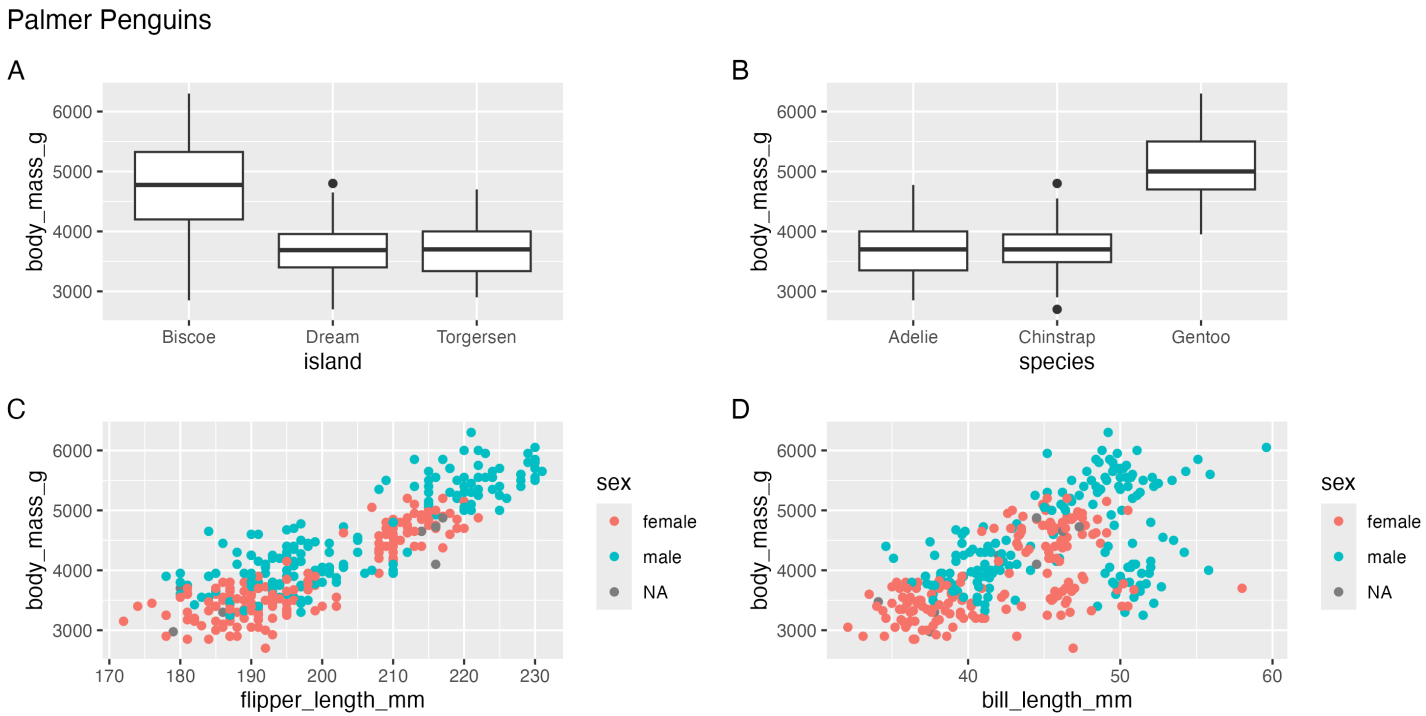
```
1 p1 + p2 + p3 + p4 + plot_layout(nrow=1)
```



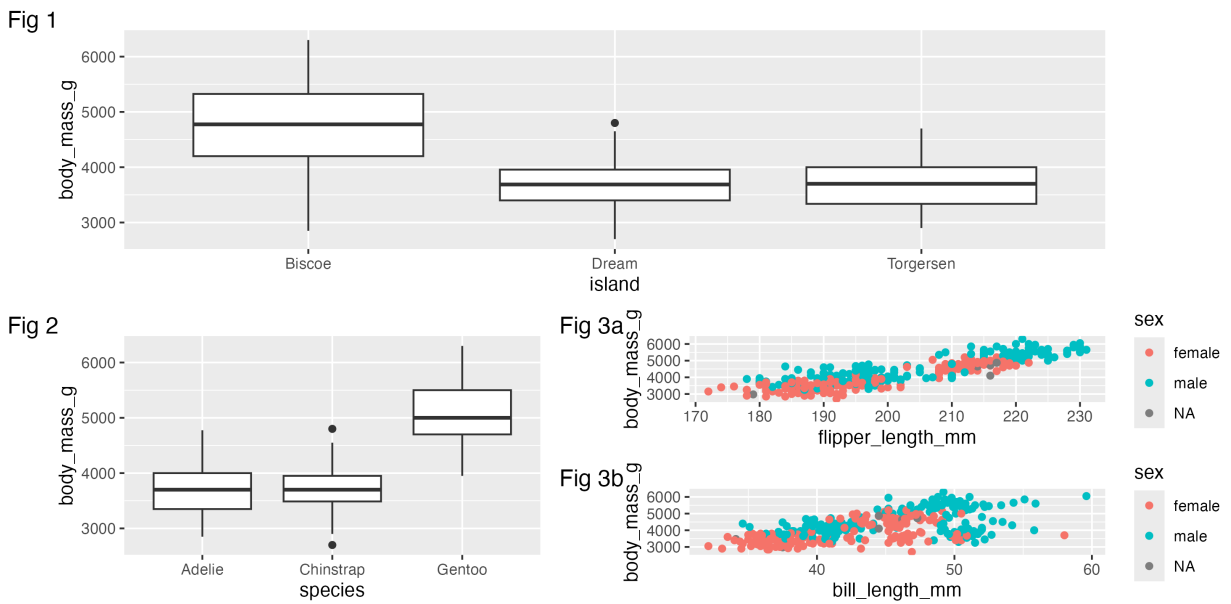
```
1 p1 / (p2 + p3 + p4)
```


body_mass_g
6000
5000
4000
3000
Biscoe
Dream
Torgersen
island
Adelie
Chinstrap
Gentoo
species
flipper_length_mm
170
180
190
200
210
220
230
bill_length_mm
40
50
60
sex
female
male
NA

```
p1 + p2 + p3 + p4 +
  plot_annotation(title = "Palmer Penguins", tag_levels = c("A"))
```

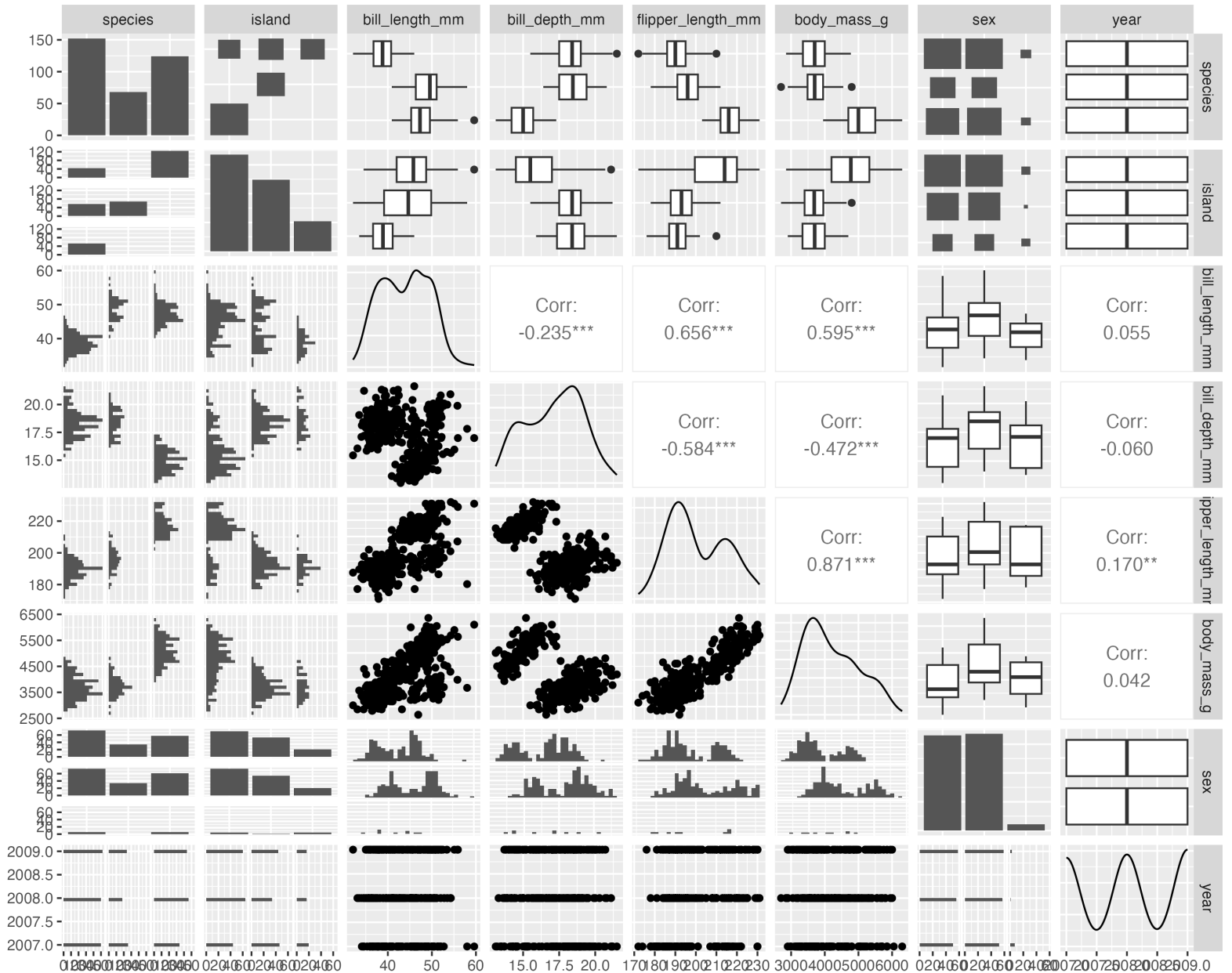


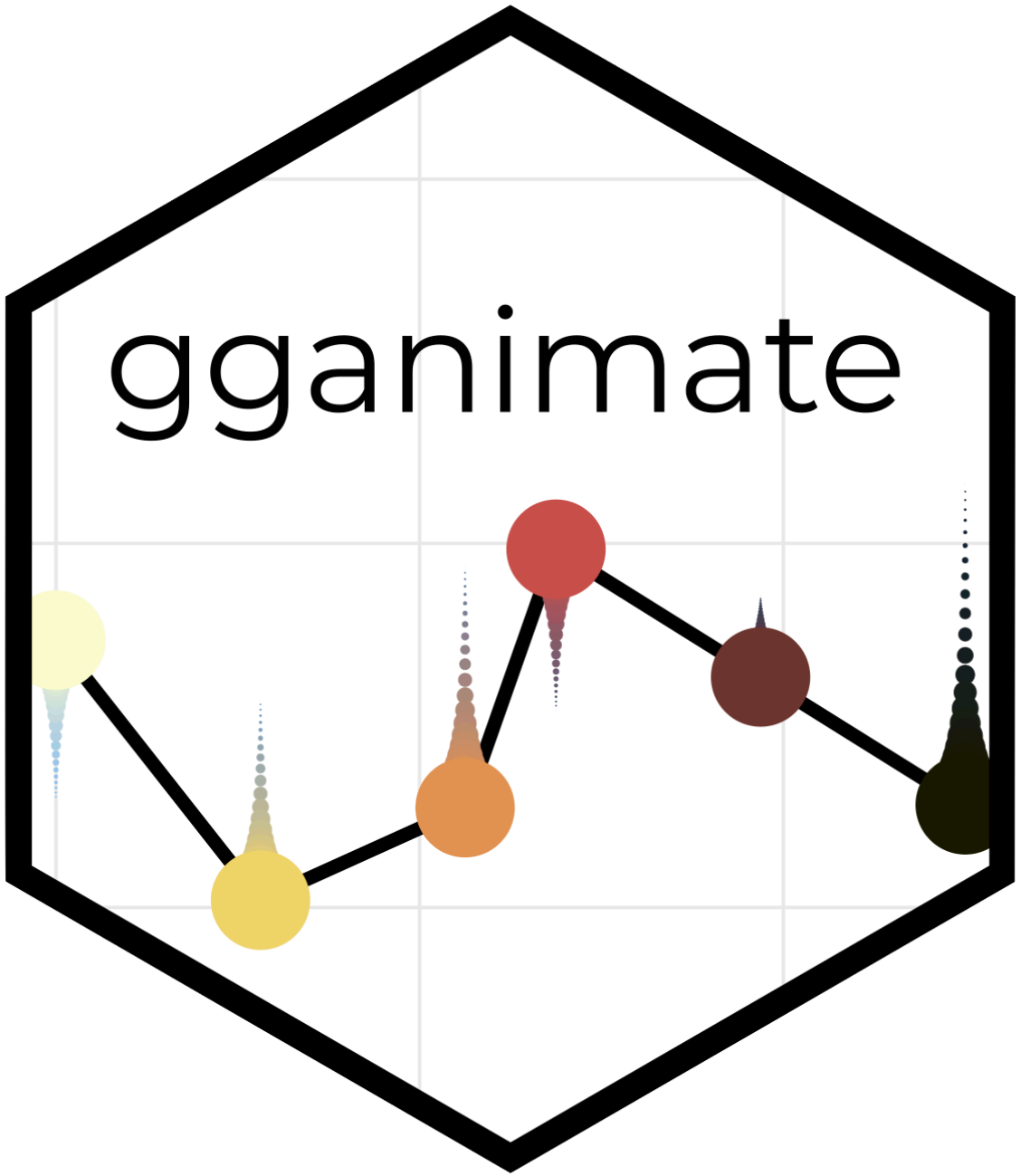
```r
p1 + {
  p2 + {
    p3 + p4 + plot_layout(ncol = 1) + plot_layout(tag_level = 'new')
  }
} +
  plot_layout(ncol = 1) +
  plot_annotation(tag_levels = c("1","a"), tag_prefix = "Fig ")
```



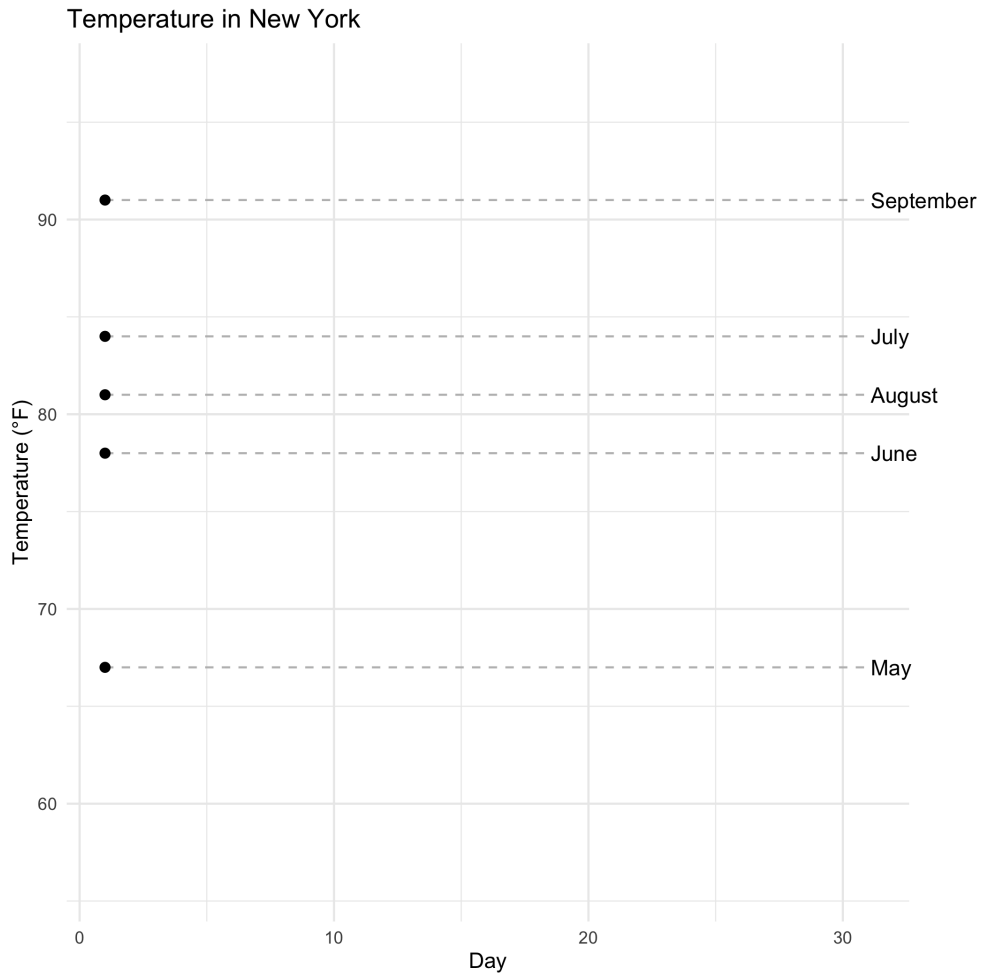
GGally

```
1 GGally::ggpairs(palmerpenguins::penguins)
```




gganimate

```r
airq = airquality
airq$Month = month.name[airq$Month]

ggplot(
  airq,
  aes(Day, Temp, group = Month)
) +
  geom_line() +
  geom_segment(
    aes(xend = 31, yend = Temp),
    linetype = 2,
    colour = 'grey'
  ) +
  geom_point(size = 2) +
  geom_text(
    aes(x = 31.1, label = Month),
    hjust = 0
  ) +
  gganimate::transition_reveal(Day) +
  coord_cartesian(clip = 'off') +
```



Some other notable packages

- `marquee` - add rendered markdown to your plots
- `thematic` & `brand.yml` - automatic theming of plots to match your app / site
- `ggridges` - creates ridgeline plots (stacked density plots)
- `ggdist` - visualizations and utilities for distributions and uncertainty (think bayesian model output)
- `legendry` - adds additional guides (legends) to ggplot2

More extensions

exts.ggplot2.tidyverse.org/gallery/

Why do we visualize?

Anscombe's Quartet

```
1 datasets::anscombe |> as_tibble()
```

```
# A tibble: 11 × 8
      x1    x2    x3    x4    y1    y2    y3    y4
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1    10    10    10     8  8.04  9.14  7.46  6.58
 2     8     8     8     8  6.95  8.14  6.77  5.76
 3    13    13    13     8  7.58  8.74 12.7   7.71
 4     9     9     9     8  8.81  8.77  7.11  8.84
 5    11    11    11     8  8.33  9.26  7.81  8.47
 6    14    14    14     8  9.96  8.1   8.84  7.04
 7     6     6     6     8  7.24  6.13  6.08  5.25
 8     4     4     4    19  4.26  3.1   5.39 12.5
 9    12    12    12     8 10.8   9.13  8.15  5.56
10     7     7     7     8  4.82  7.26  6.42  7.91
11     5     5     5     8  5.68  4.74  5.73  6.89
```

Tidy anscombe

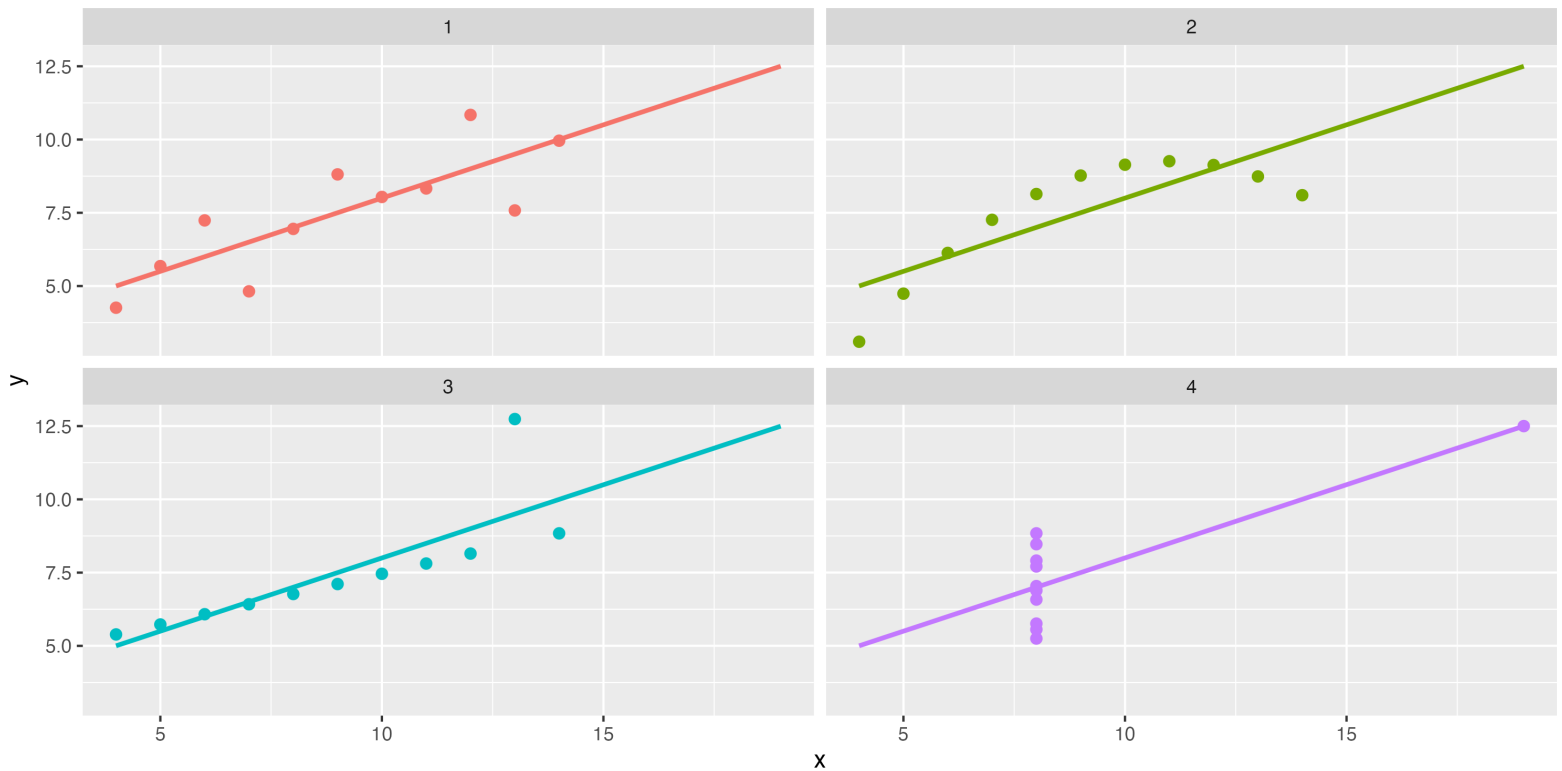
```
(tidy_anscombe = datasets::anscombe |>
  pivot_longer(everything(), names_sep = 1, names_to = c("var", "group")) |>
  pivot_wider(id_cols = group, names_from = var,
              values_from = value, values_fn = list(value = list)) |>
  unnest(cols = c(x,y)))
```

```
# A tibble: 44 × 3
   group     x     y
   <chr> <dbl> <dbl>
 1 1        10  8.04
 2 1         8  6.95
 3 1        13  7.58
 4 1         9  8.81
 5 1        11  8.33
 6 1        14  9.96
 7 1         6  7.24
 8 1         4  4.26
 9 1        12 10.8
10 1         7  4.82
# ℹ 34 more rows
```

```r
tidy_anscombe |>
  group_by(group) |>
  summarize(
    mean_x = mean(x), mean_y = mean(y),
    sd_x = sd(x), sd_y = sd(y),
    cor = cor(x,y), .groups = "drop"
  )
```

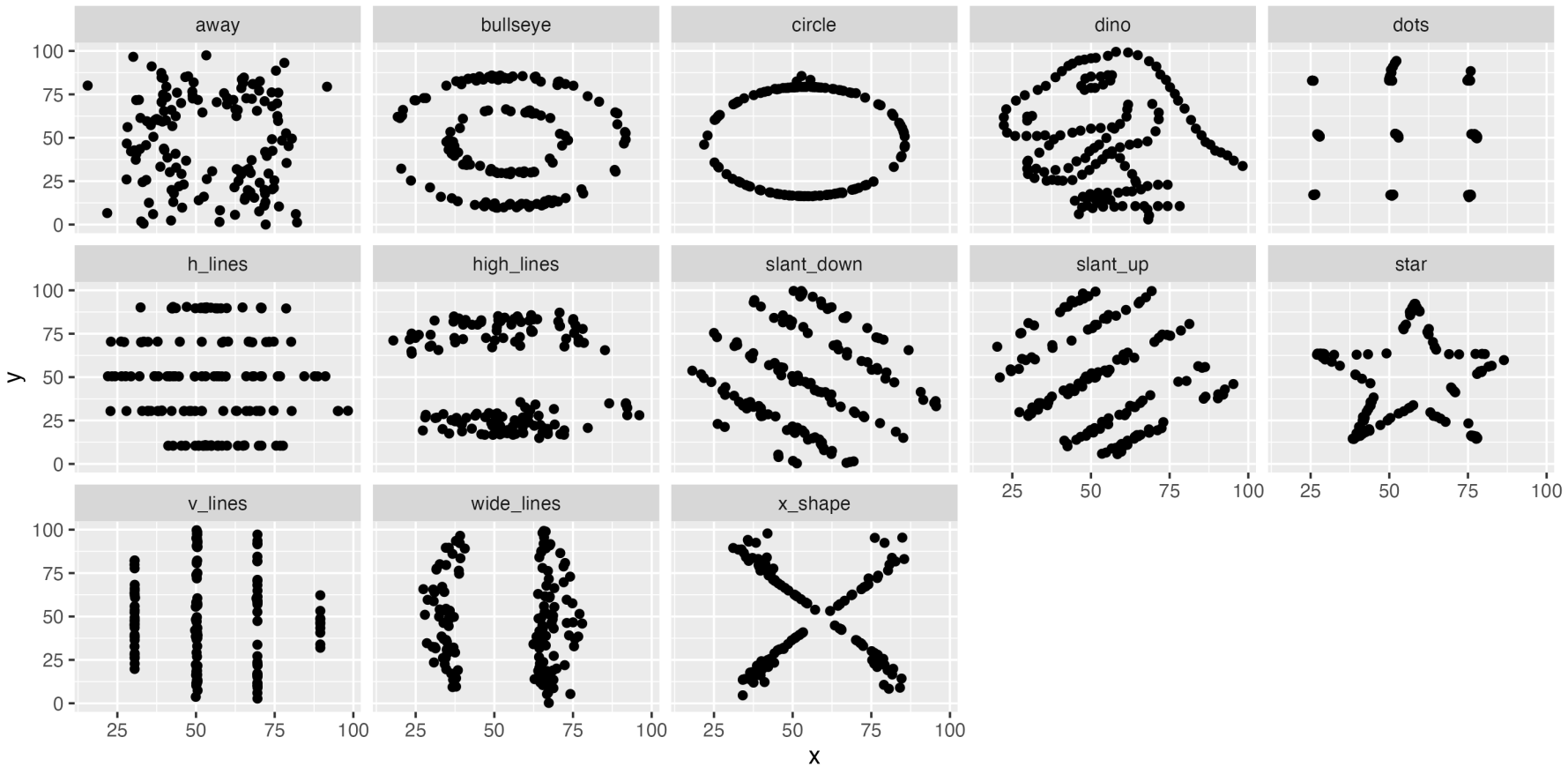
```
# A tibble: 4 × 6
  group mean_x mean_y  sd_x  sd_y   cor
  <chr>  <dbl>  <dbl> <dbl> <dbl> <dbl>
1 1          9   7.50  3.32  2.03 0.816
2 2          9   7.50  3.32  2.03 0.816
3 3          9   7.5   3.32  2.03 0.816
4 4          9   7.50  3.32  2.03 0.817
```

```r
ggplot(tidy_anscombe, aes(x = x, y = y, color = as.factor(group))) +
  geom_point(size=2) +
  facet_wrap(~group) +
  geom_smooth(method="lm", se=FALSE, fullrange=TRUE, formula = y~x) +
  guides(color="none")
```



DatasauRus

```
ggplot(datasauRus::datasaurus_dozen, aes(x = x, y = y)) +
  geom_point() +
  facet_wrap(~dataset, ncol=5)
```




bullseye
dino
star
100
75
50
25
0
y
20
40
60
80
100
x

```
datasauRus::datasaurus_
```

```
# A tibble: 1,846 × 3
   dataset     x     y
   <chr>   <dbl> <dbl>
 1 dino     55.4  97.2
 2 dino     51.5  96.0
 3 dino     46.2  94.5
 4 dino     42.8  91.4
 5 dino     40.8  88.3
 6 dino     38.7  84.9
 7 dino     35.6  79.9
 8 dino     33.1  77.6
 9 dino     29.0  74.5
10 dino     26.2  71.4
# ℹ 1,836 more rows
```

```
datasauRus::datasaurus_dozen |>
  group_by(dataset) |>
  summarize(mean_x = mean(x), mean_y = mean(y),
            sd_x = sd(x), sd_y = sd(y),
            cor = cor(x,y), .groups = "drop")
```

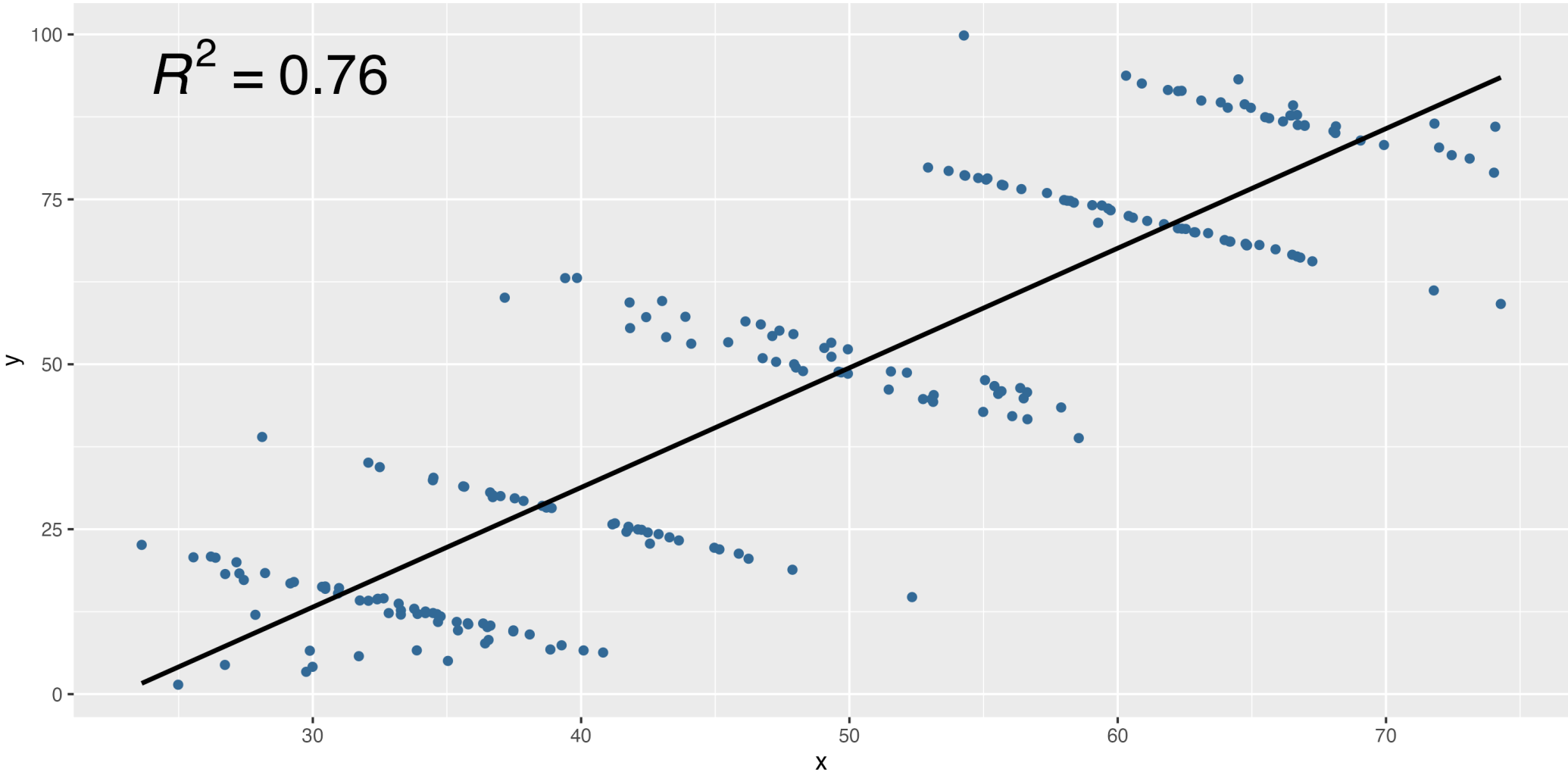
```
# A tibble: 13 × 6
   dataset    mean_x mean_y  sd_x  sd_y     cor
   <chr>       <dbl>  <dbl> <dbl> <dbl>   <dbl>
 1 away         54.3   47.8  16.8  26.9 -0.0641
 2 bullseye     54.3   47.8  16.8  26.9 -0.0686
 3 circle       54.3   47.8  16.8  26.9 -0.0683
 4 dino         54.3   47.8  16.8  26.9 -0.0645
 5 dots         54.3   47.8  16.8  26.9 -0.0603
 6 h_lines      54.3   47.8  16.8  26.9 -0.0617
 7 high_lines   54.3   47.8  16.8  26.9 -0.0685
 8 slant_down   54.3   47.8  16.8  26.9 -0.0690
 9 slant_up     54.3   47.8  16.8  26.9 -0.0686
10 star         54.3   47.8  16.8  26.9 -0.0630
11 v_lines      54.3   47.8  16.8  26.9 -0.0694
12 wide_lines   54.3   47.8  16.8  26.9 -0.0666
13 x_shape      54.3   47.8  16.8  26.9 -0.0656
```

Simpson's Paradox

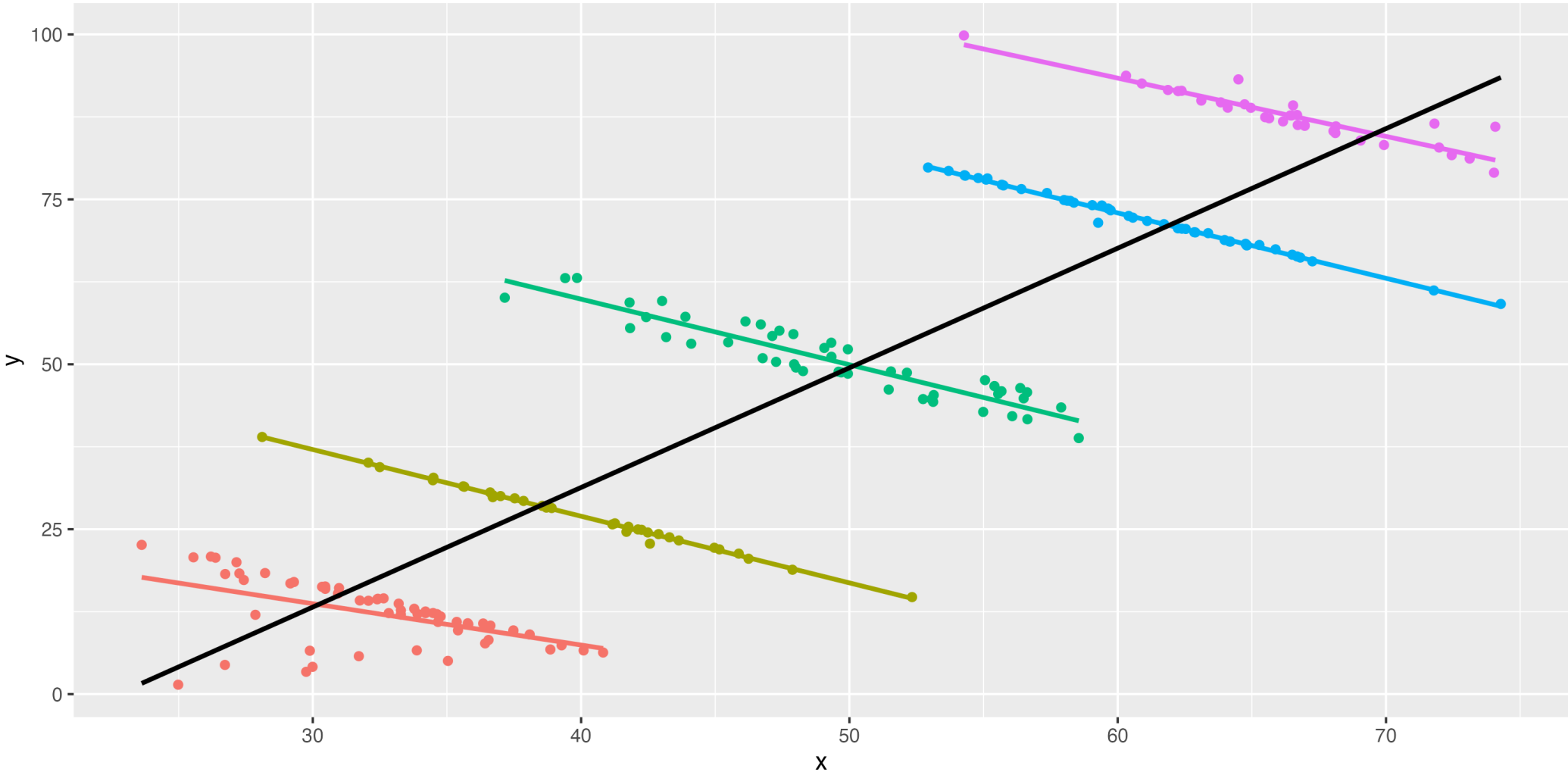
```r
lm(y~x, data=simpsons) |>
  summary()
```

```
Call:
lm(formula = y ~ x, data = simpsons)

Residuals:
    Min      1Q  Median      3Q     Max
-38.988 -10.208  -0.707   9.874  42.642

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -41.20220    3.51007  -11.74   <2e-16 ***
x             1.81324    0.06993   25.93   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.93 on 215 degrees of freedom
Multiple R-squared:  0.7577,    Adjusted R-squared:  0.7566
F-statistic: 672.2 on 1 and 215 DF,  p-value: < 2.2e-16
```

Simpson's Paradox Visually



Simpson's Paradox with groups



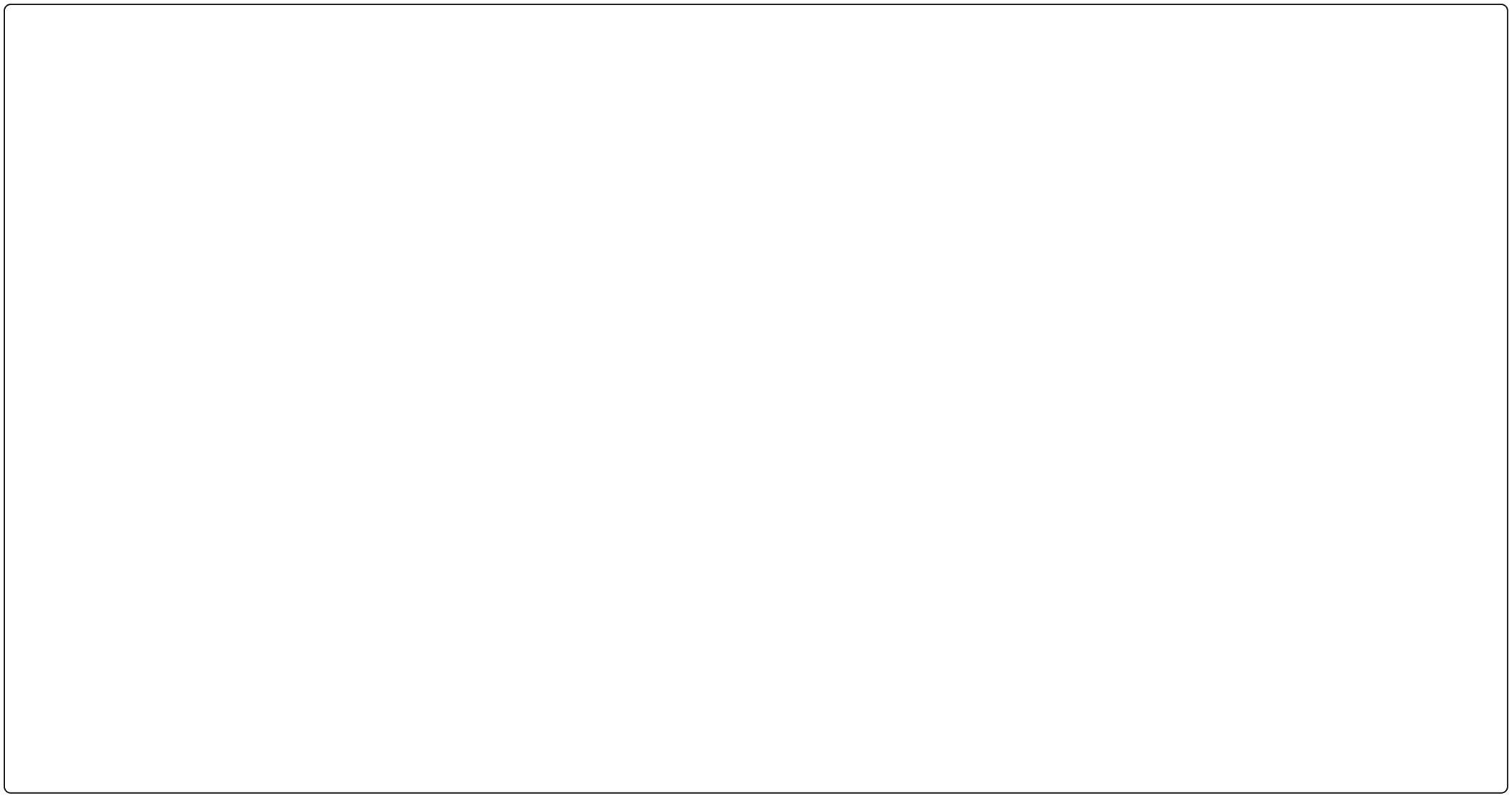
Revised model

```r
lm(y~x*group-1, data=simpsons) |>
  summary()
```

```
Call:
lm(formula = y ~ x * group - 1, data = simpsons)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.4264  -0.6137   0.0811   1.0448   5.0613 

Coefficients:
          Estimate Std. Error t value Pr(>|t|)    
x          -0.62658    0.07987  -7.846 2.27e-13 ***
group1     32.50512    2.61640  12.424  < 2e-16 ***
group2     67.38858    3.47010  19.420  < 2e-16 ***
group3     99.63330    3.34565  29.780  < 2e-16 ***
group4    132.39316    4.76158  27.804  < 2e-16 ***
group5    146.36456    6.78530  21.571  < 2e-16 ***
x:group2   -0.38394    0.11747  -3.268 0.001267 ** 
x:group3   -0.36743    0.10440  -3.519 0.000532 ***
x:group4   -0.36425    0.11146  -3.268 0.001268 ** 
x:group5   -0.25654    0.12950  -1.981 0.048917 *  
```

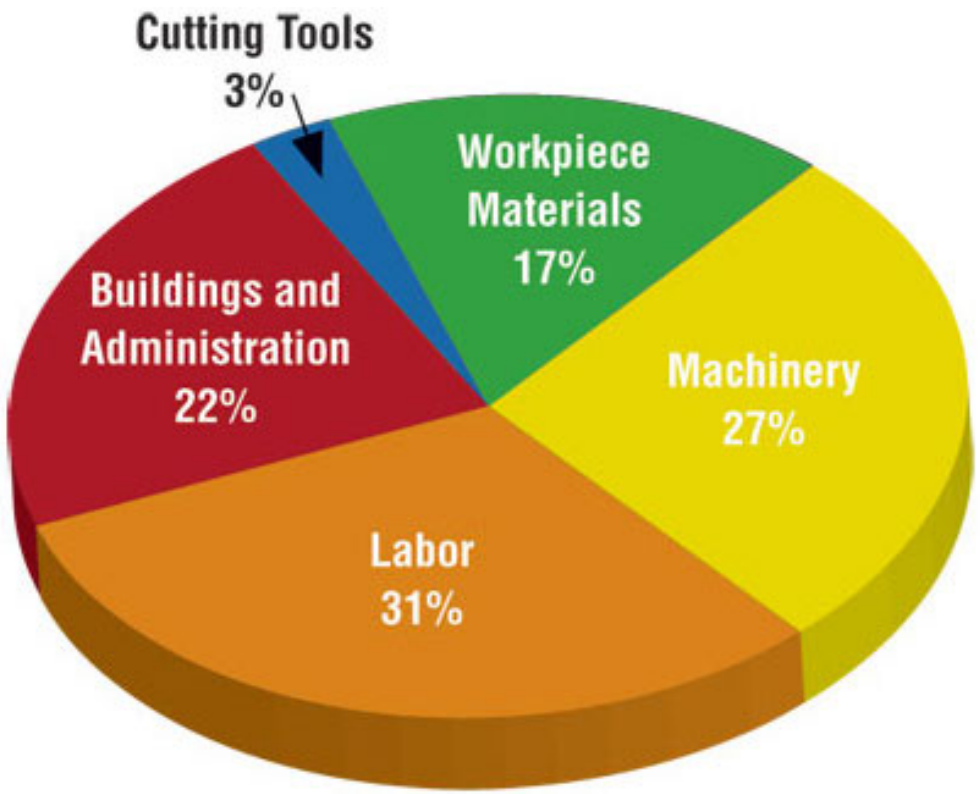
Designing effective visualizations

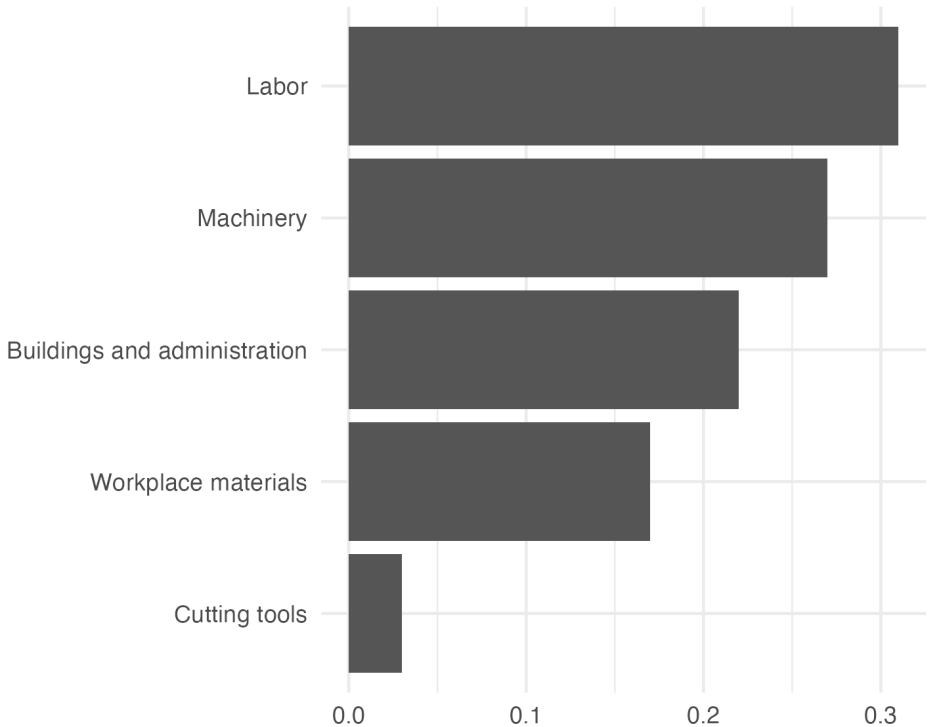
Gapminder



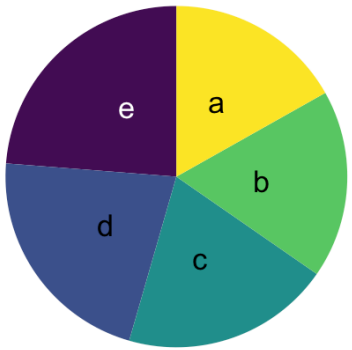
gapminder.org/dollar-street

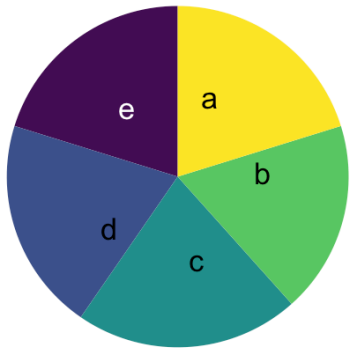
Keep it simple

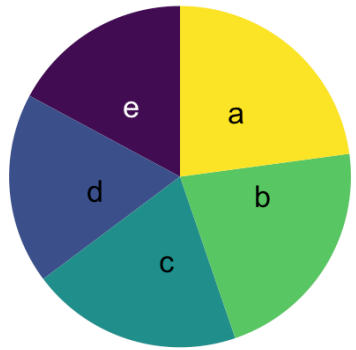


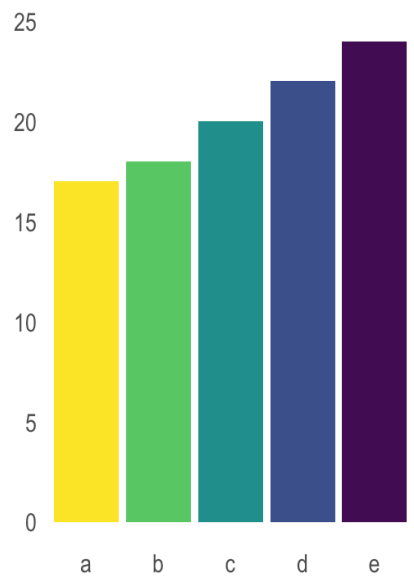


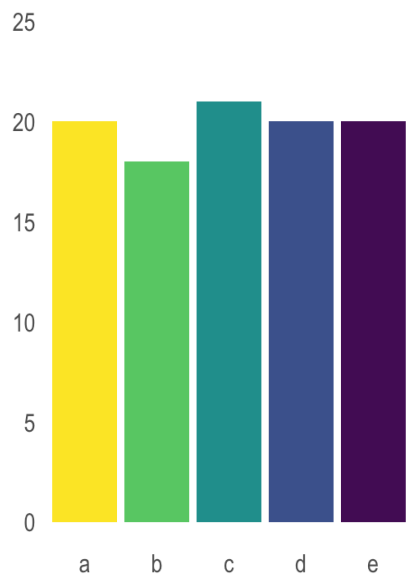
Judging relative area

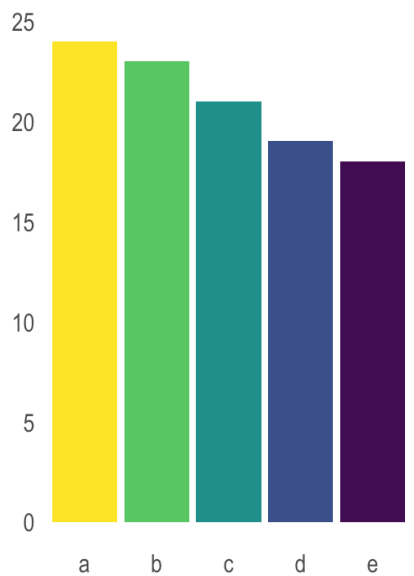




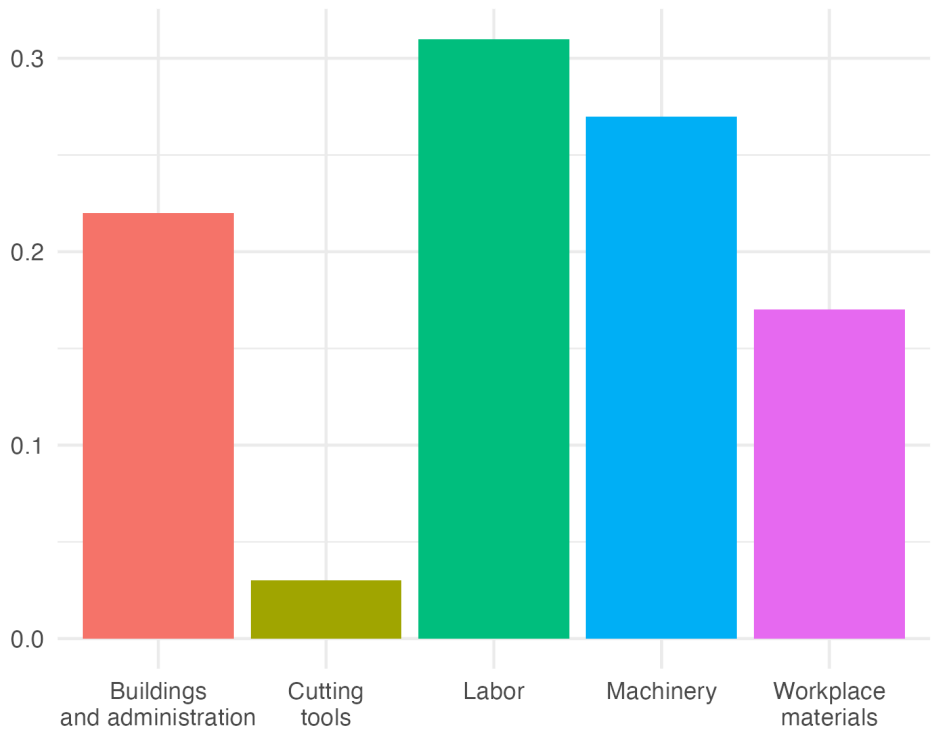


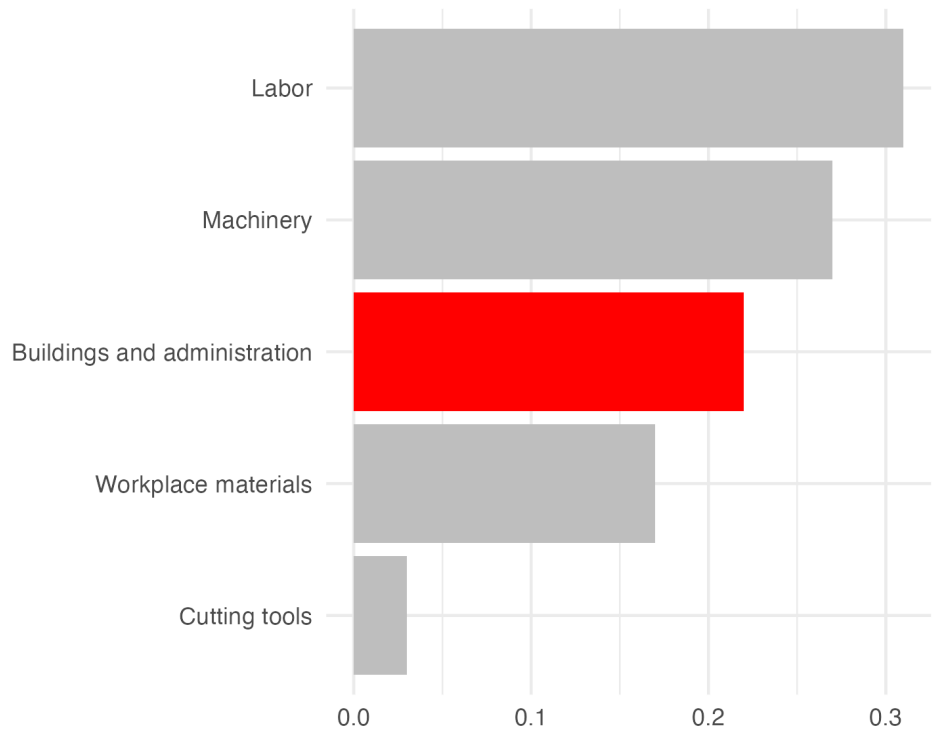




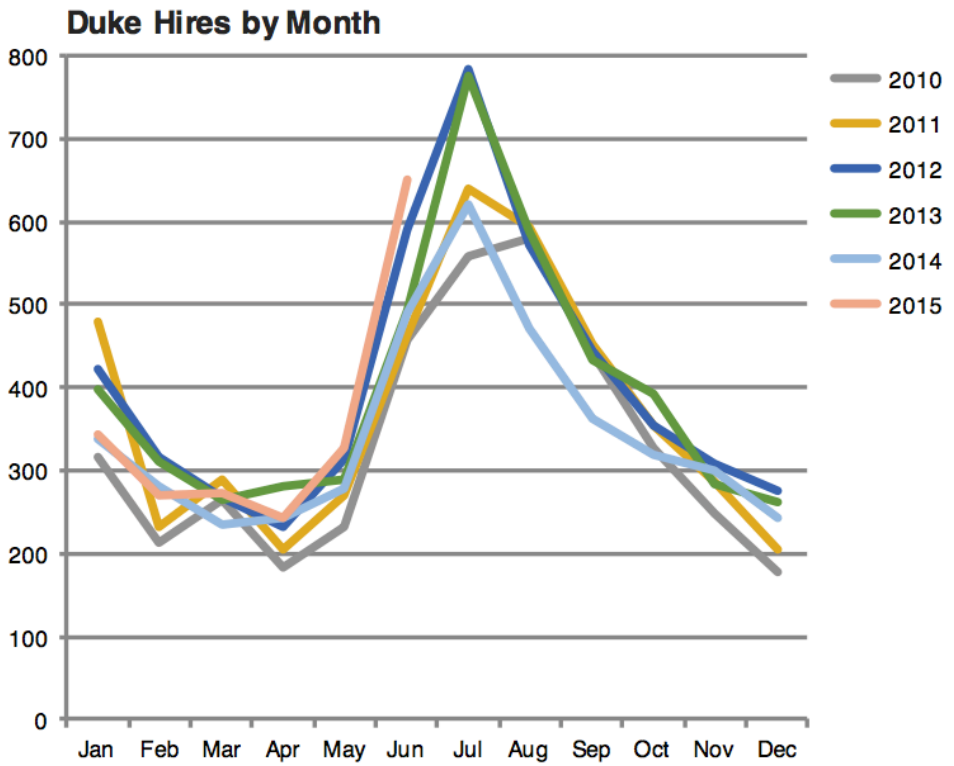


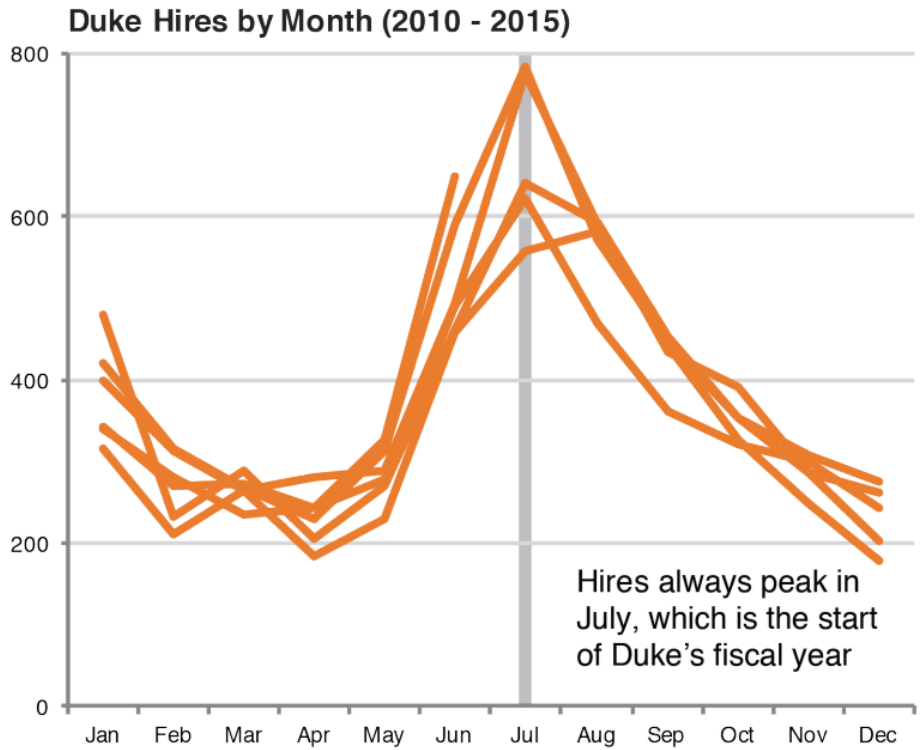
Use color to draw attention



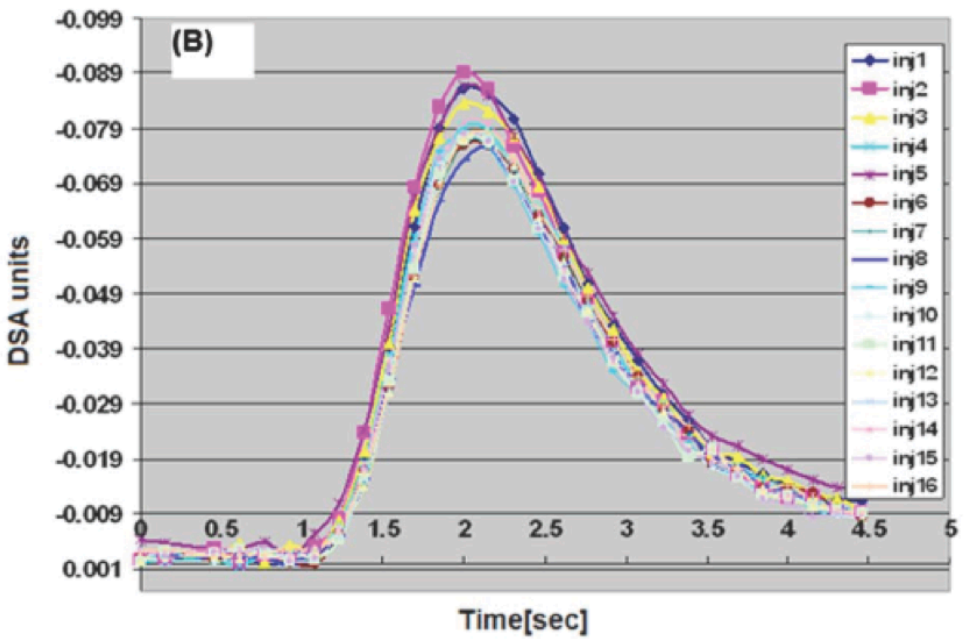


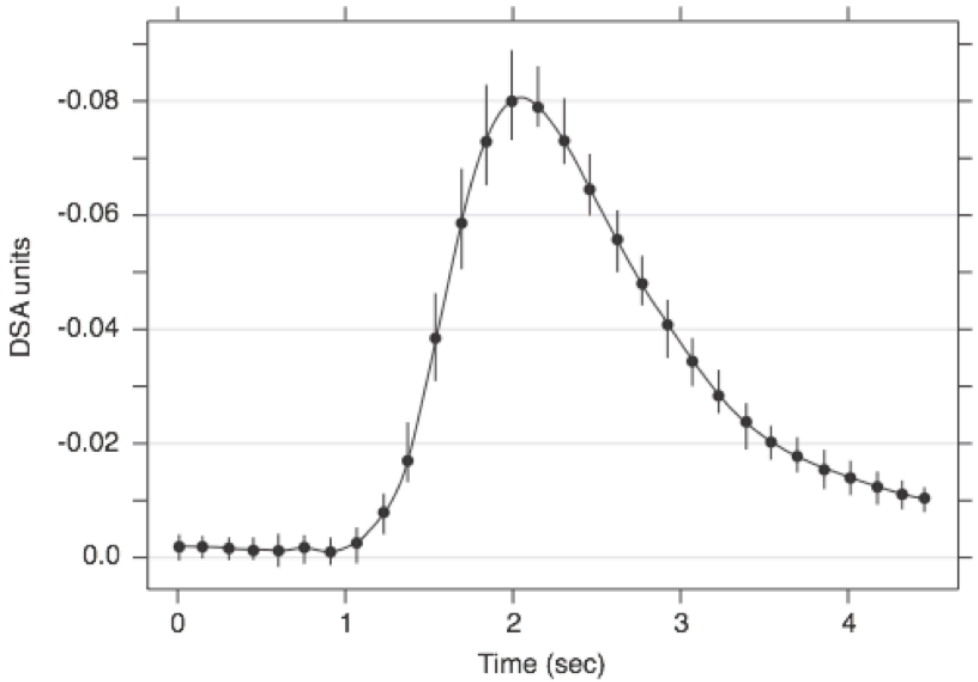
Tell a story


Duke Hires by Month
800
700
600
500
400
300
200
100
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2010
2011
2012
2013
2014
2015

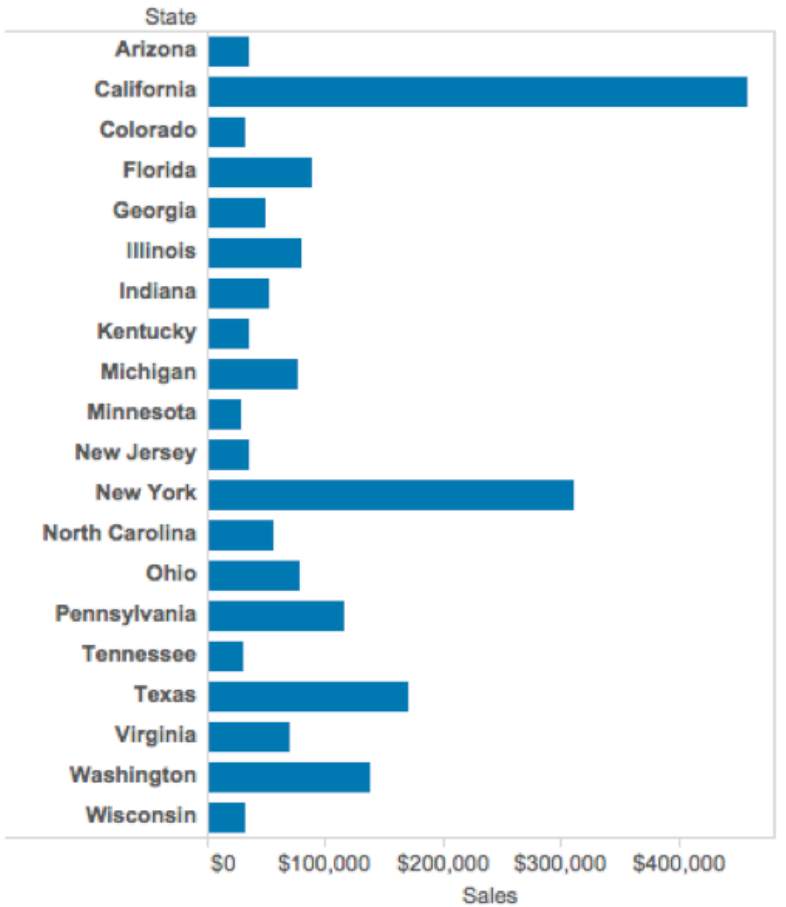

Duke Hires by Month (2010 - 2015)
800
600
400
200
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Hires always peak in July, which is the start of Duke's fiscal year

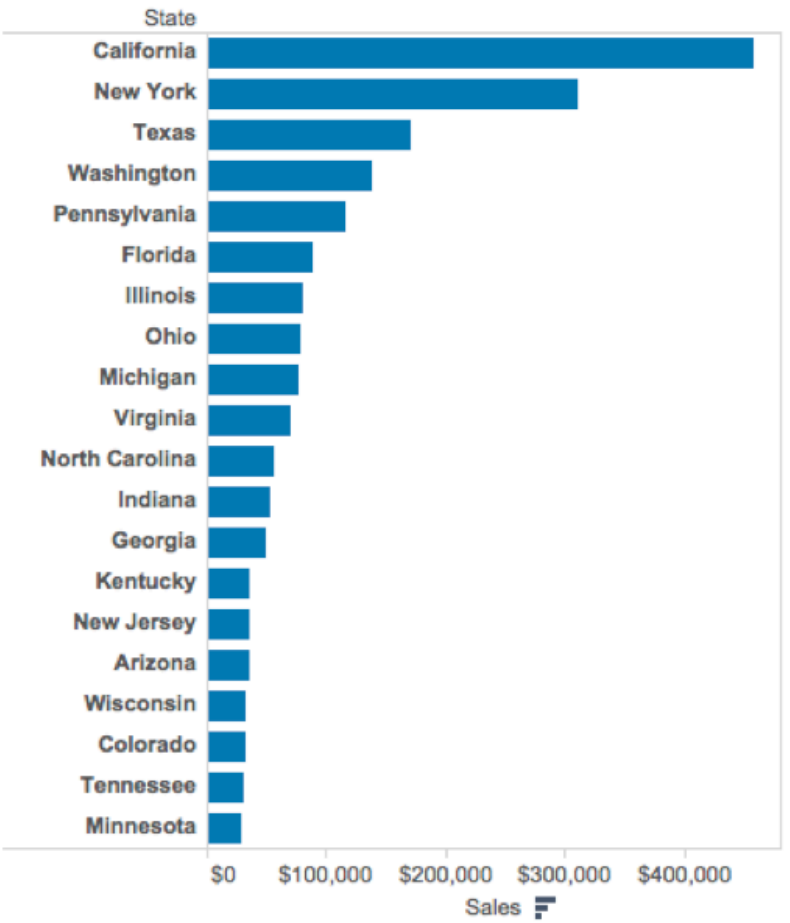
Leave out non-story details





Ordering matters





Clearly indicate missing data

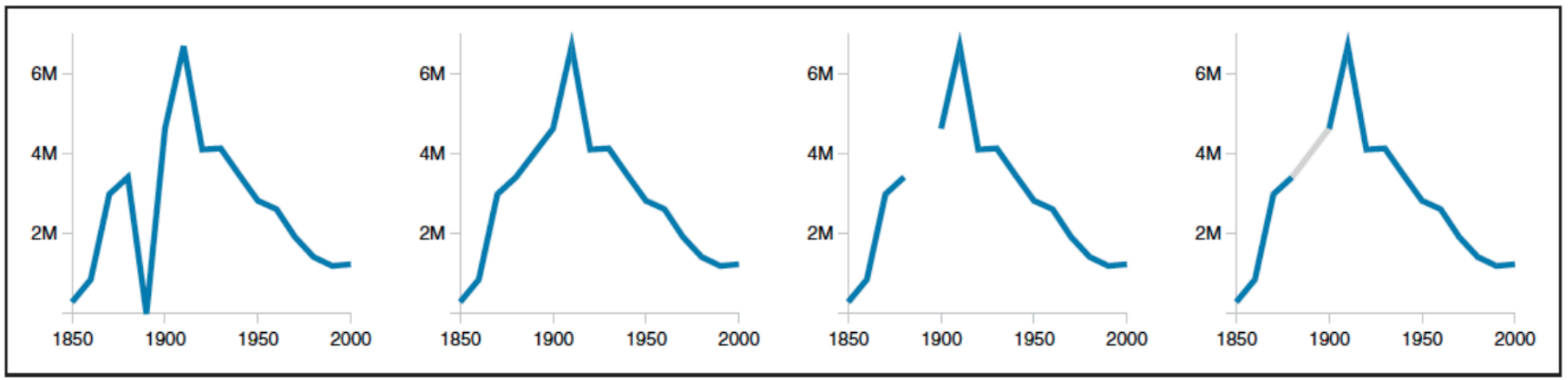
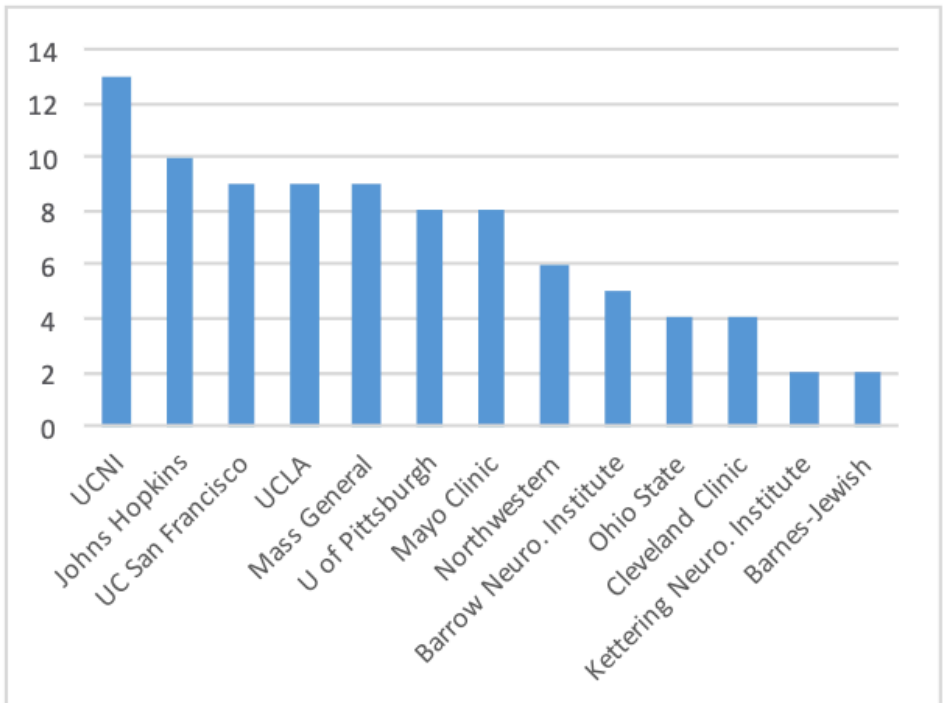
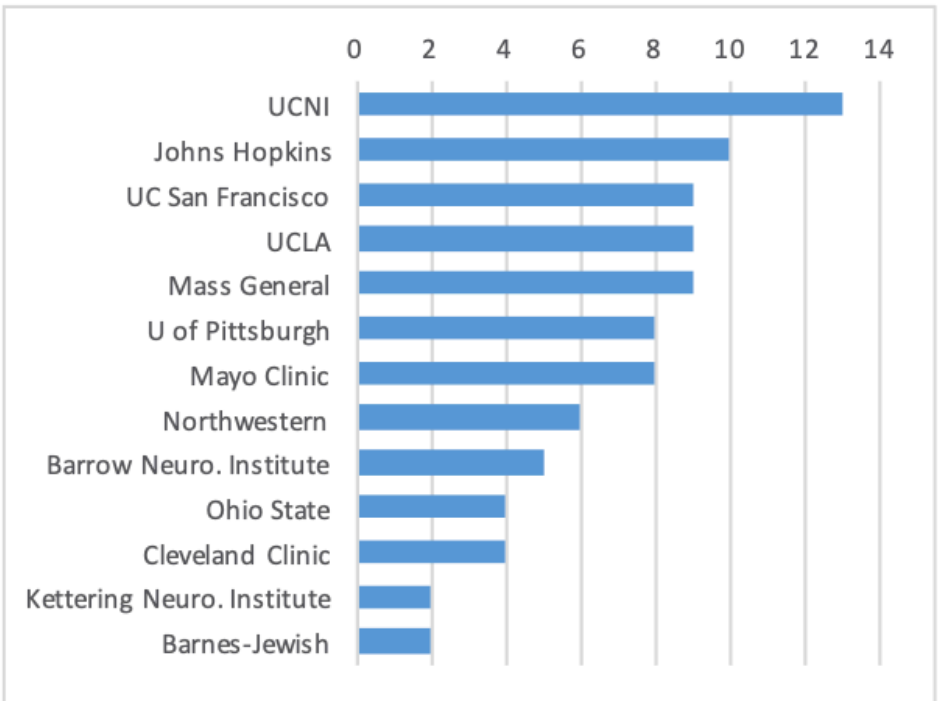


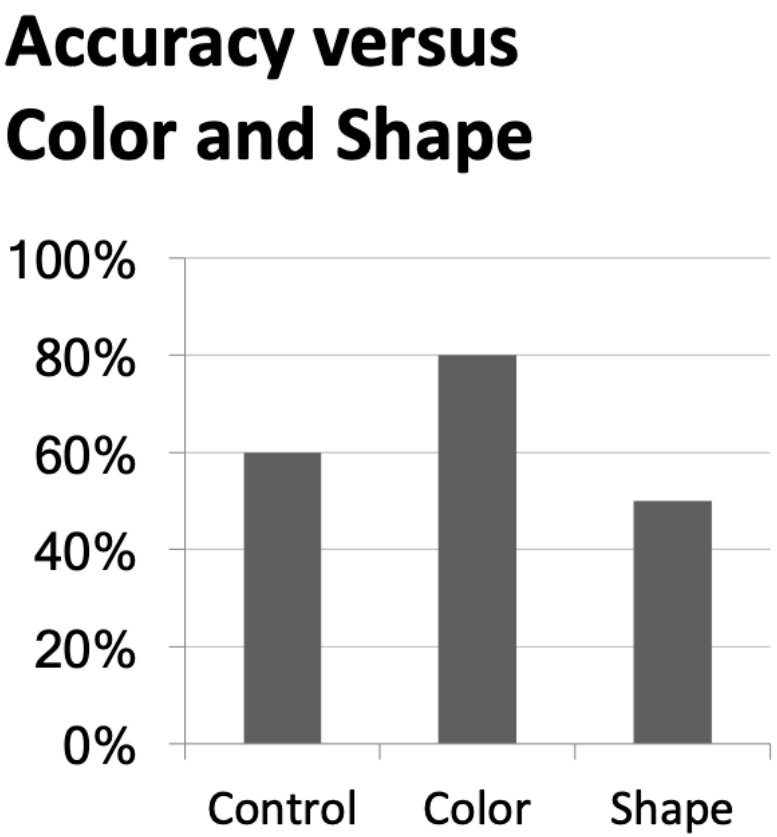
Figure 4. Alternative representations of missing data in a line chart. The data are U.S. census counts of people working as 'Farm Laborers'; values from 1890 are missing due to records being burned in a fire. (a) Missing data is treated as a zero value. (b) Missing data is ignored, resulting in a line segment that interpolates the missing value. (c) Missing data is omitted from the chart. (d) Missing data is explicitly interpolated and rendered in gray.

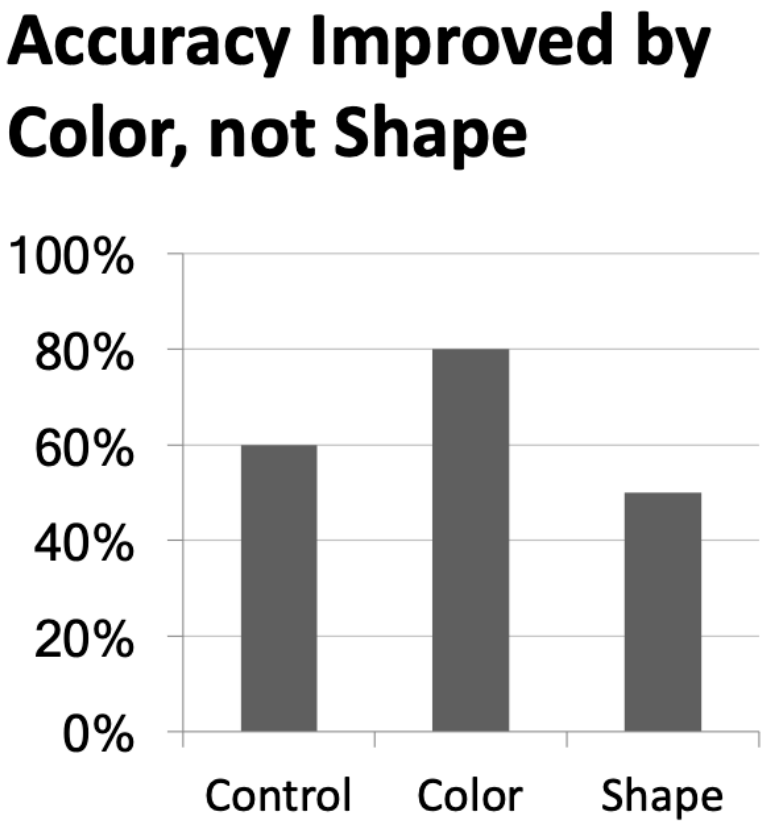
Reduce cognitive load





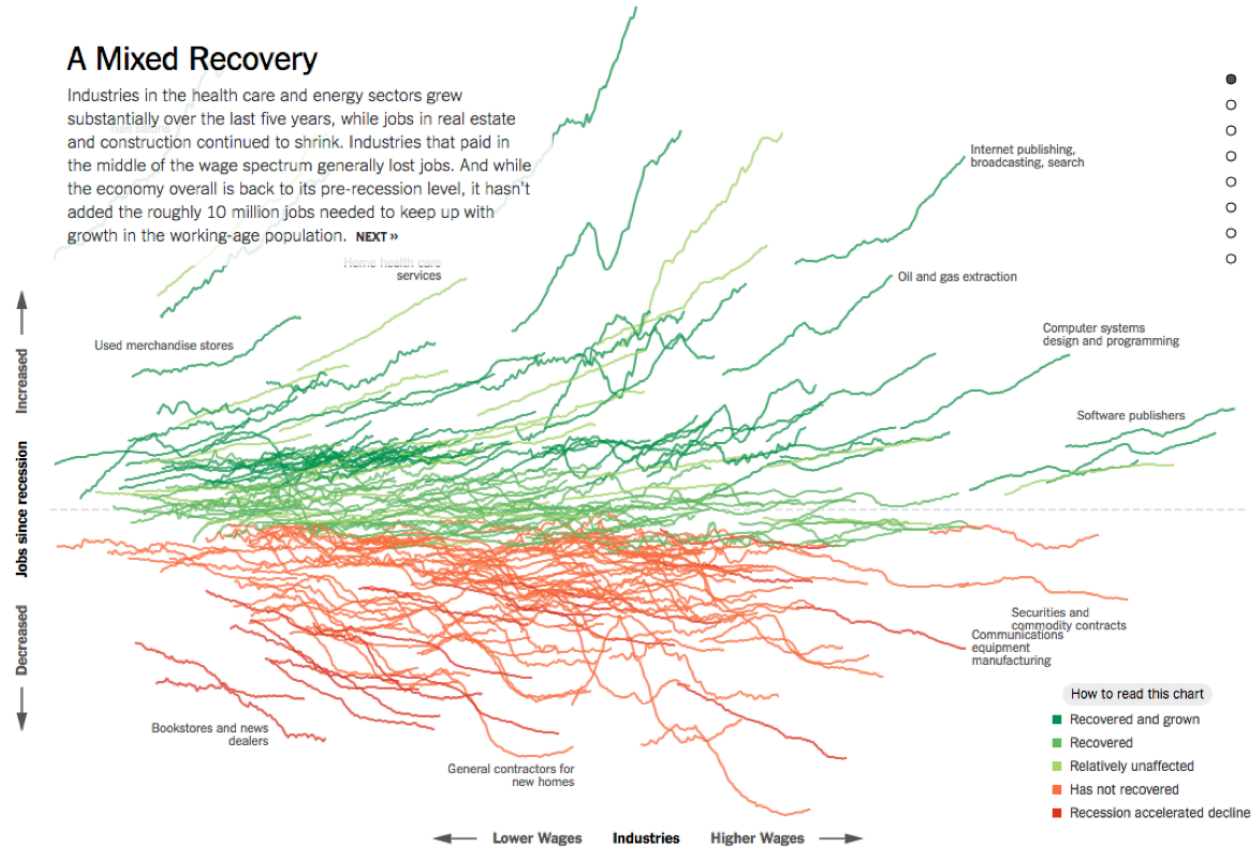
Use descriptive titles



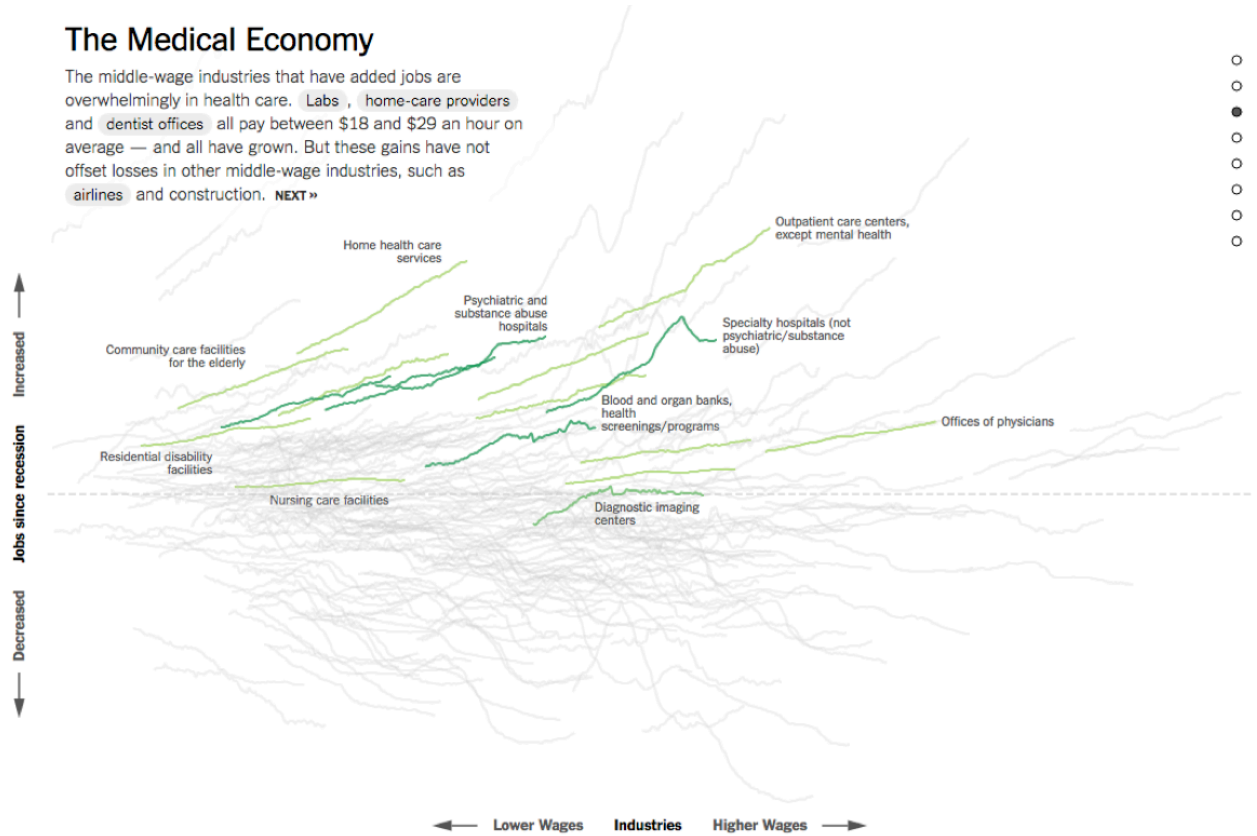


Annotate figures

All of the data doesn't tell a story



All of the data doesn’t tell a story



All of the data doesn’t tell a story

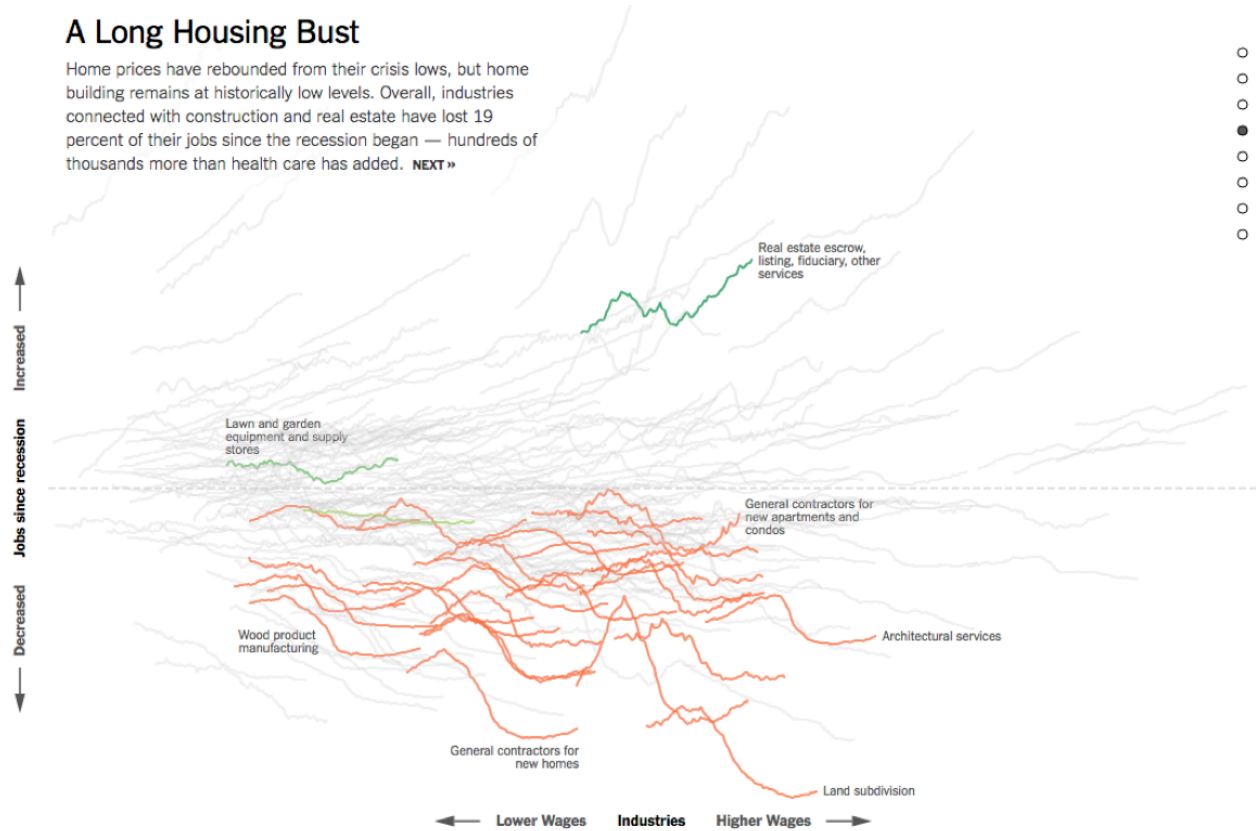
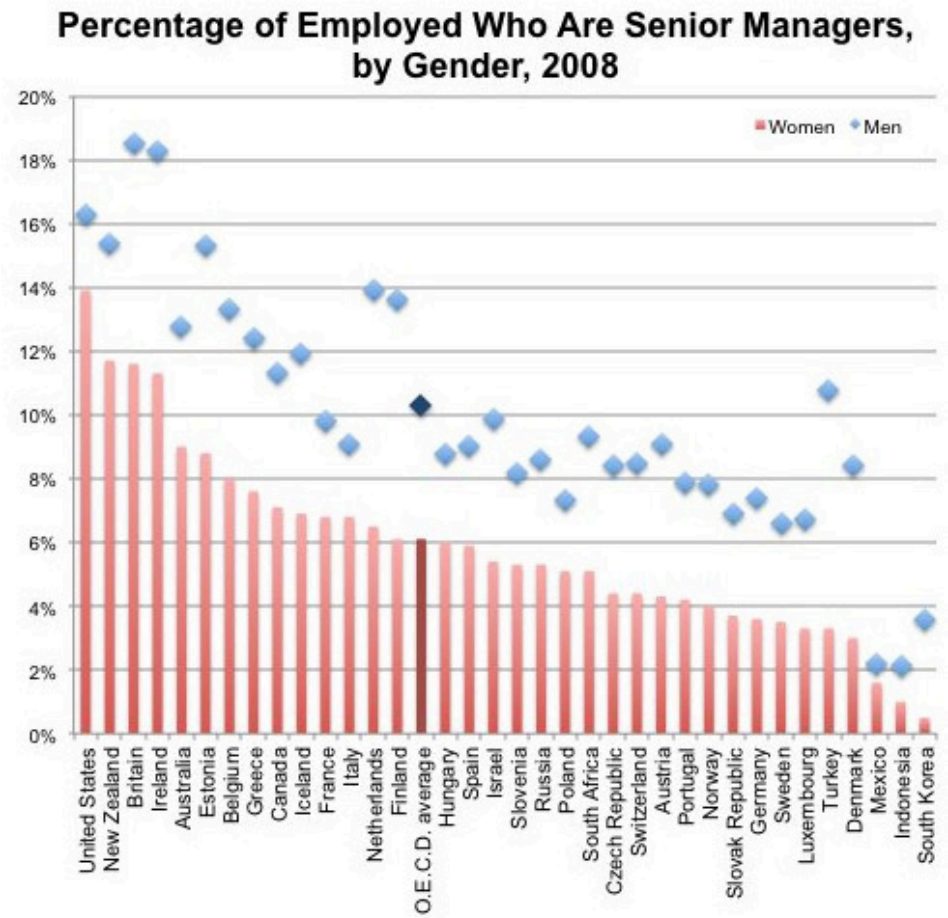


Chart Remakes / Makeovers

The Why Axis - Gender Gap



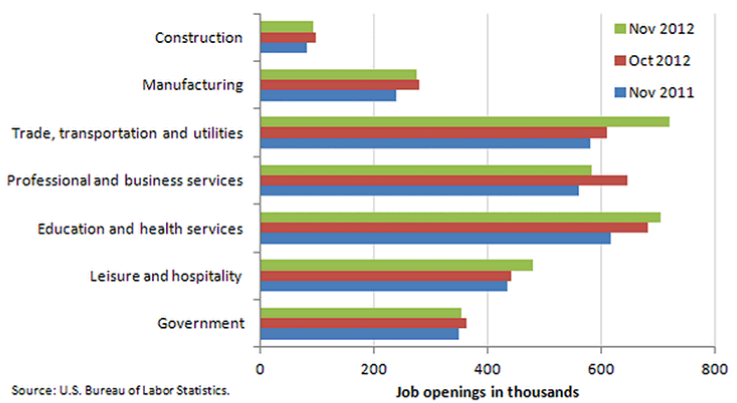
The Why Axis - BLS

Job openings in November 2012

JANUARY 11, 2013

There were 3.7 million job openings on the last business day of November 2012, unchanged from October 2012. In November 2011 there were 3.3 million job openings.

CHART IMAGE | CHART DATA



From November 2011 to November 2012, job openings increased most in retail trade (144,000, within the trade, transportation and utilities industry) and health care and social assistance (91,000, within the education and health services industry).

Government job openings increased the least, by 6,000.

These data are from the Job Openings and Labor Turnover Survey. Data for the most recent month are preliminary and subject to revision. For additional information, see Job Openings and Labor Turnover — November 2012" (HTML) (PDF), news release USDL-13-0015. More charts featuring data on job openings, hires, and employment separations can be found in Job Openings and Labor Turnover Survey Highlights: November 2012 (PDF).

Other Resources

- Duke Library - Center for Data and Visualization Sciences - https://library.duke.edu/data/
- Tidy Tuesday - https://github.com/rfordatascience/tidytuesday
- Twitter / Bluesky / Mastodon - #dataviz, #tidytuesday
- Books:
  - Wickham, Navarro, Pedersen. *ggplot2: Elegant Graphics for Data Analysis*. 3rd edition. Faller, 2021.
  - Wilke. *Fundamentals of Data Visualization*. O'Reilly Media, 2019.
  - Healy. *Data Visualization: A Practical Introduction*. Princeton University Press, 2018.
  - Tufte. *The visual display of quantitative information*. 2nd edition. Connecticut Graphics Press, 2015.

Acknowledgments

Above materials are derived in part from the following sources:

- Visualization training materials originally developed by Angela Zoss and Eric Monson
- Duke Center for Data and Visualization Sciences