Data and Vizualization

Visualizing different types of data

Click for PDF of slides

Identifying variables

Number of variables involved

- **Univariate data analysis**: distribution of single variable

- **Bivariate data analysis**: relationship between two variables

- **Multivariate data analysis**: relationship between many variables at once, usually focusing on the relationship between two while conditioning for others

Types of variables

- **Numerical variables** can be classified as **continuous** or **discrete** based on whether or not the variable can take on an infinite number of values or only non-negative whole numbers, respectively.
  - *height* is continuous
  - *number of siblings* is discrete

Types of variables

- **Numerical variables** can be classified as **continuous** or **discrete** based on whether or not the variable can take on an infinite number of values or only non-negative whole numbers, respectively.
  - *height* is continuous
  - *number of siblings* is discrete
- If the variable is **categorical**, we can determine if it is **ordinal** based on whether or not the levels have a natural ordering.
  - *hair color* is unordered
  - *year in school* is ordinal

Visualizing numerical data

Describing numerical distributions

- shape:
  - skewness: right-skewed, left-skewed, symmetric
  - modality: unimodal, bimodal, multimodal, uniform
- center: mean (**`mean`**), median (**`median`**), mode (not always useful)
- spread: range (**`range`**), standard deviation (**`sd`**), inter-quartile range (**`IQR`**)
- outliers: observations outside of the usual pattern

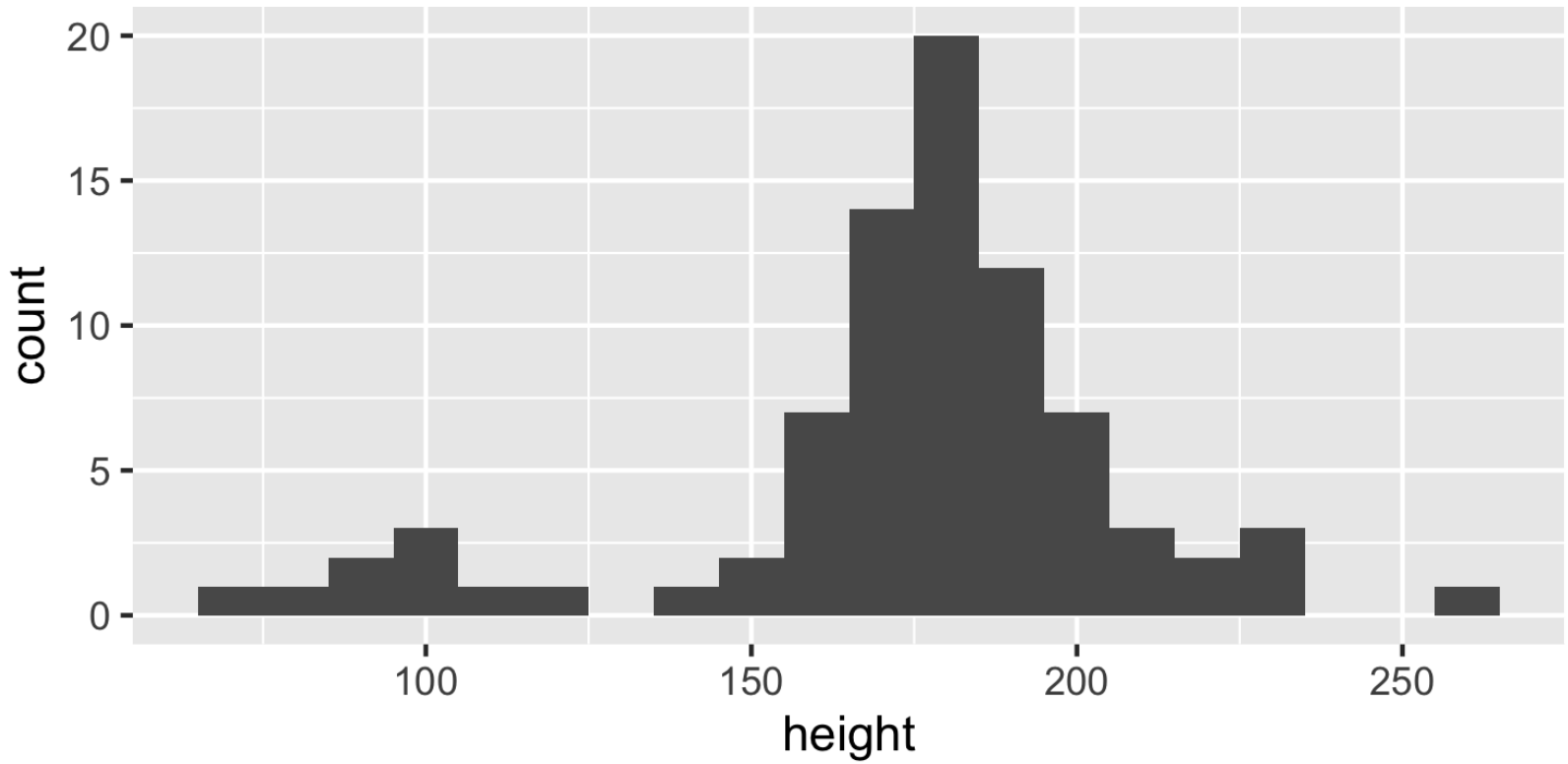
Starwars data

```
starwars
```

```
## # A tibble: 87 x 14
##    name  height  mass hair_color skin_color eye_color birth_year sex   gen
##    <chr>  <int> <dbl> <chr>      <chr>      <chr>           <dbl> <chr> <ch
##  1 Luke…    172    77 other      fair       blue             19   male  mas
##  2 C-3PO    167    75 none       gold       yellow          112   none  mas
##  3 R2-D2     96    32 none       white, bl… red              33   none  mas
##  4 Dart…    202   136 none       white      yellow           41.9 male  mas
##  5 Leia…    150    49 brown      light      brown            19   fema… fem
##  6 Owen…    178   120 brown      light      blue             52   male  mas
##  7 Beru…    165    75 brown      light      blue             47   fema… fem
##  8 R5-D4     97    32 none       white, red red              NA   none  mas
##  9 Bigg…    183    84 black      light      brown            24   male  mas
## 10 Obi-…    182    77 other      fair       blue-gray        57   male  mas
## # … with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>
```

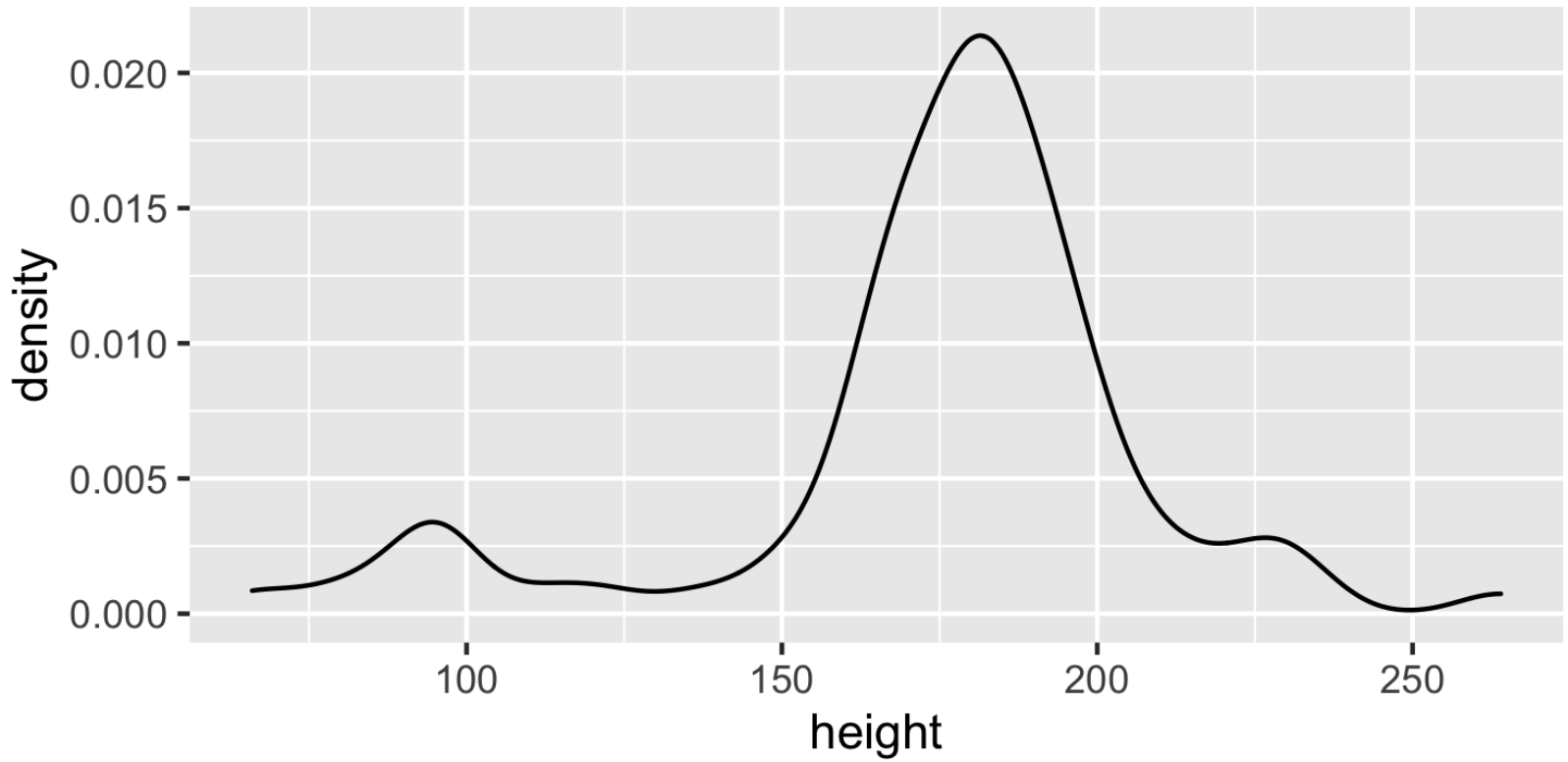
Histograms

```
ggplot(data = starwars, mapping = aes(x = height)) +
  geom_histogram(binwidth = 10)
```



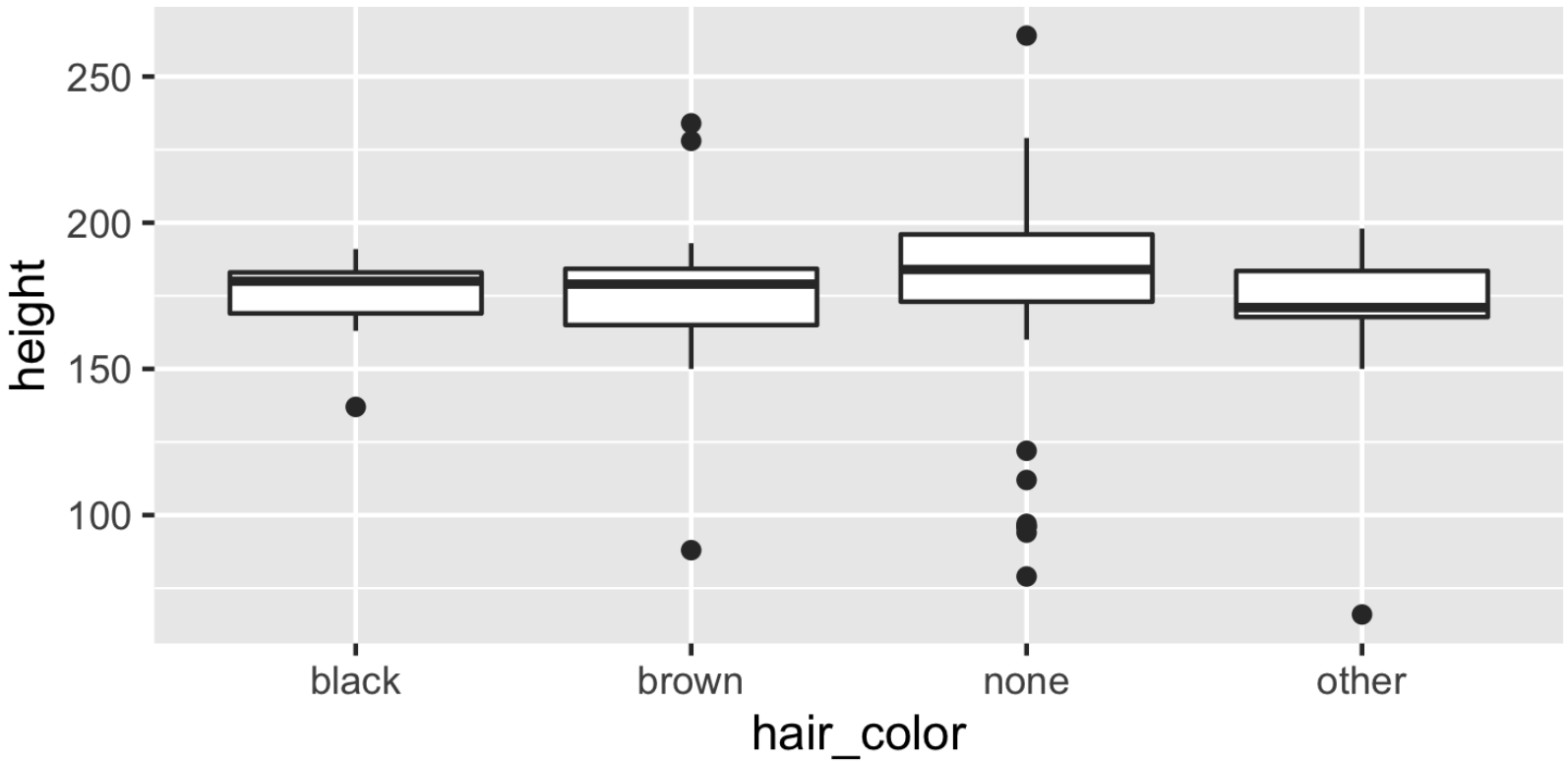
Density plots

```
ggplot(data = starwars, mapping = aes(x = height)) +
  geom_density()
```



Side-by-side box plots

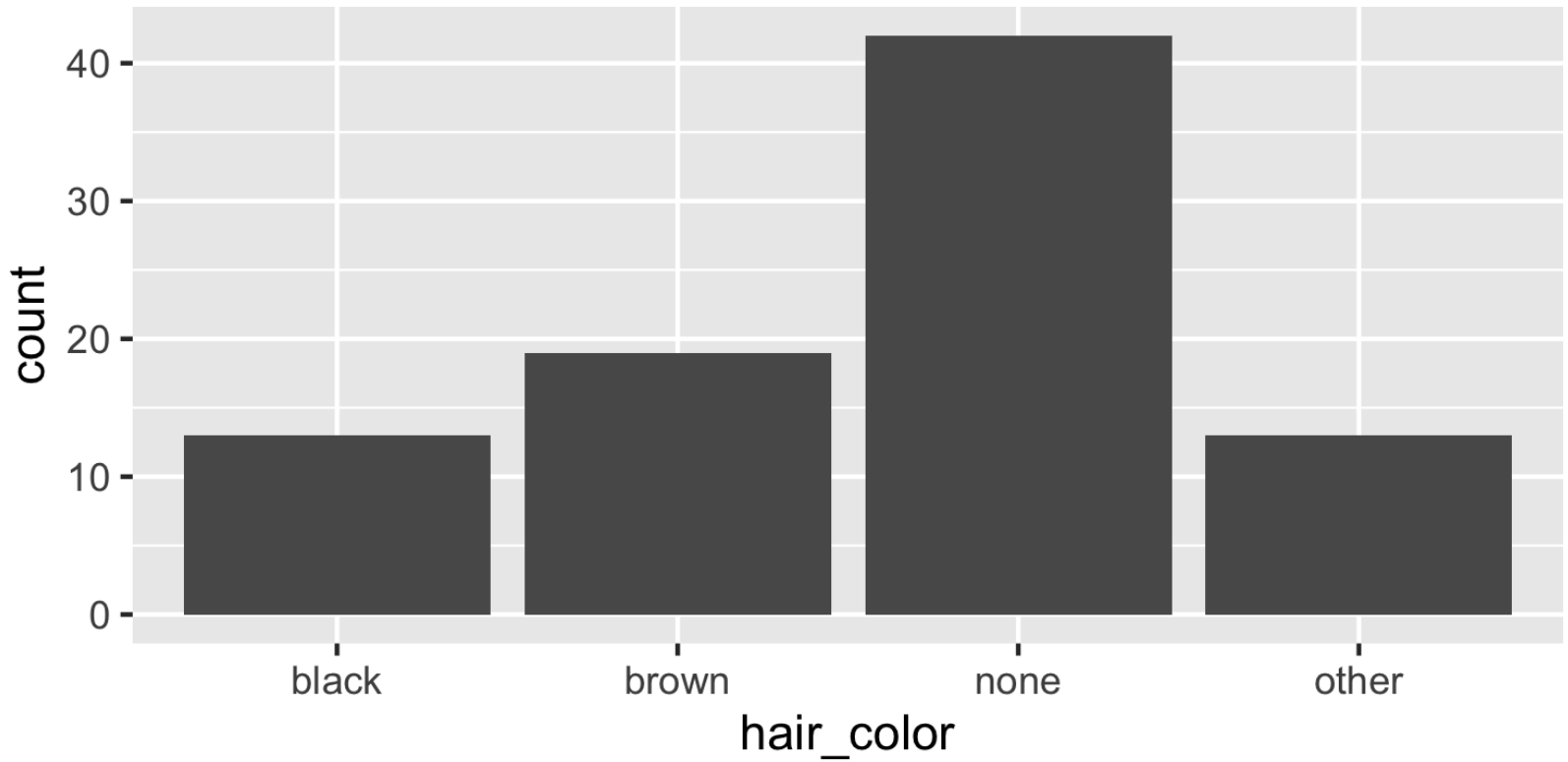
```
ggplot(data = starwars, mapping = aes(y = height, x = hair_color)) +
  geom_boxplot()
```



Visualizing categorical data

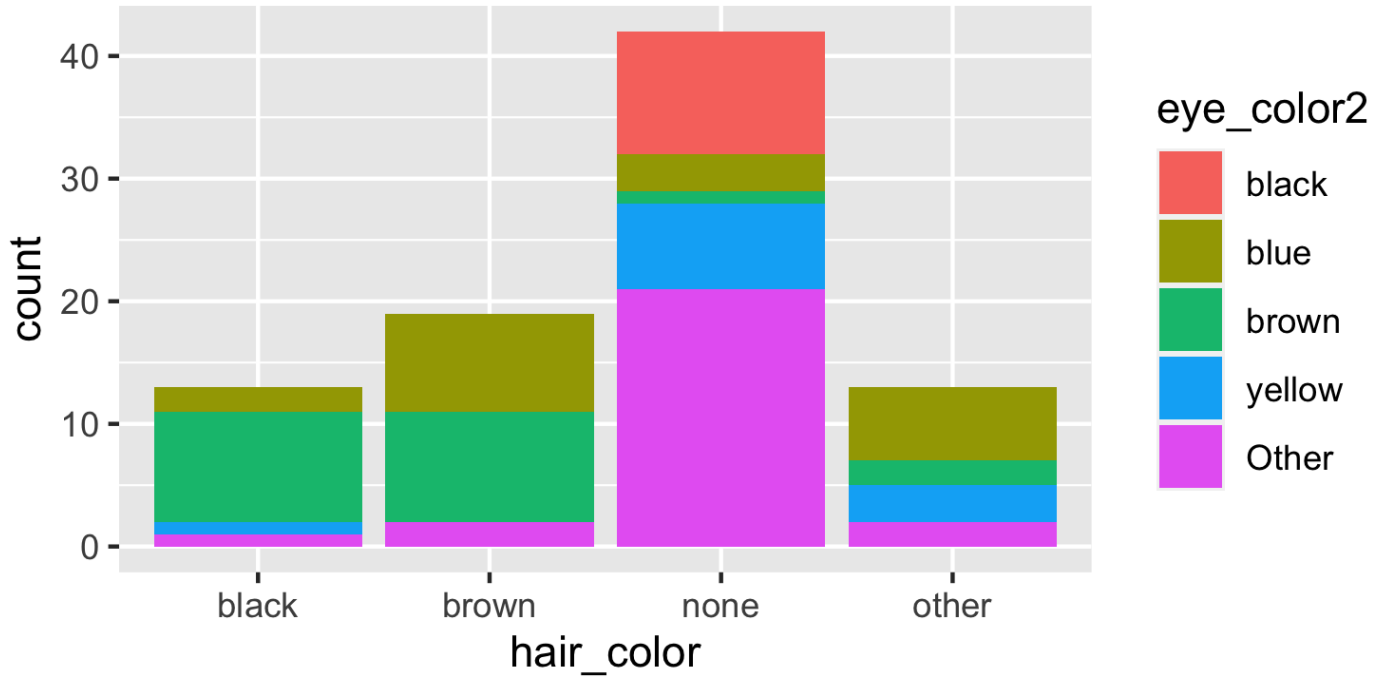
Bar plots

```
ggplot(data = starwars, mapping = aes(x = hair_color)) +
  geom_bar()
```



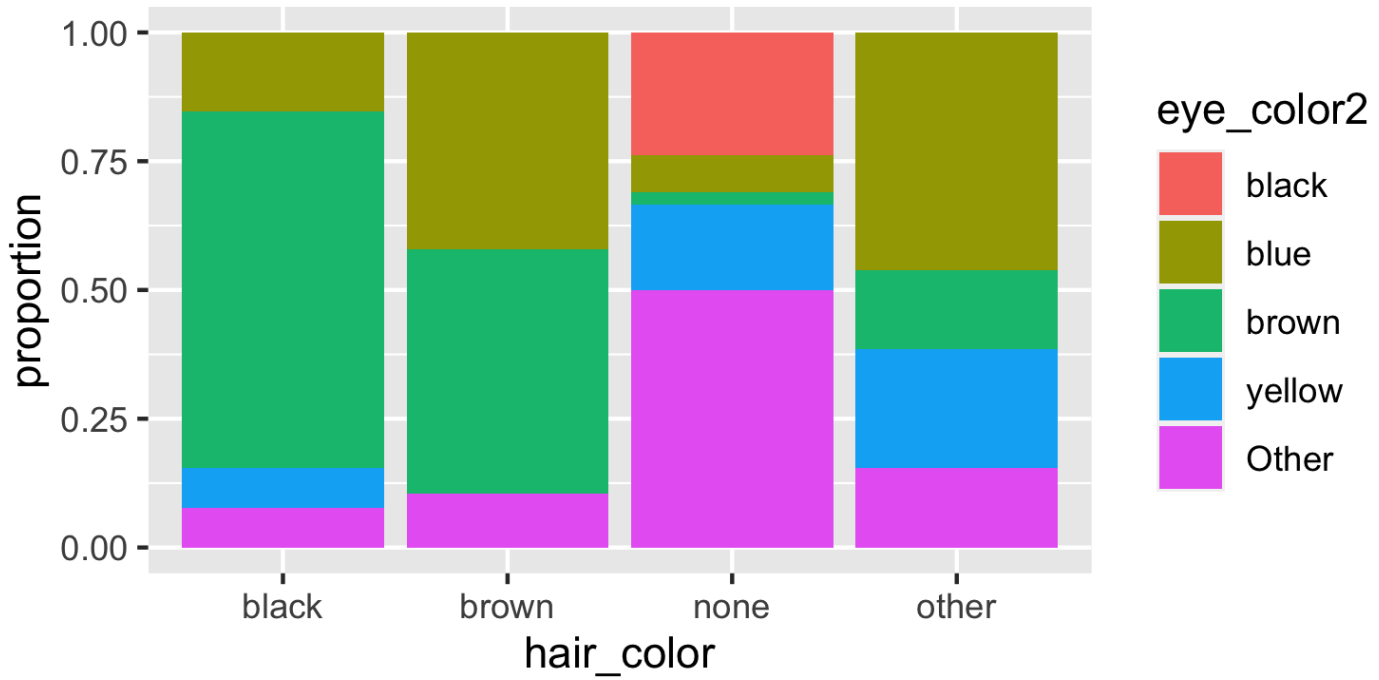
Segmented bar plots, counts

```
ggplot(data = starwars, mapping = aes(x = hair_color, fill = eye_color2)) +
  geom_bar()
```



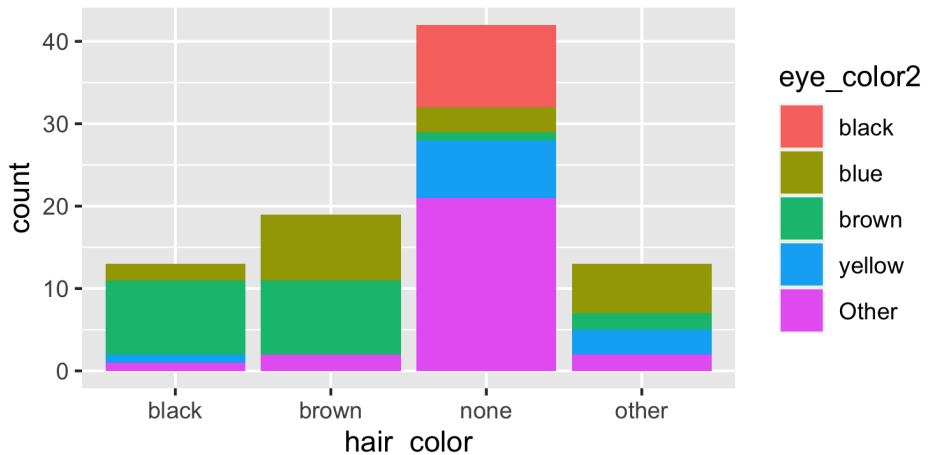
Segmented bar plots, proportions

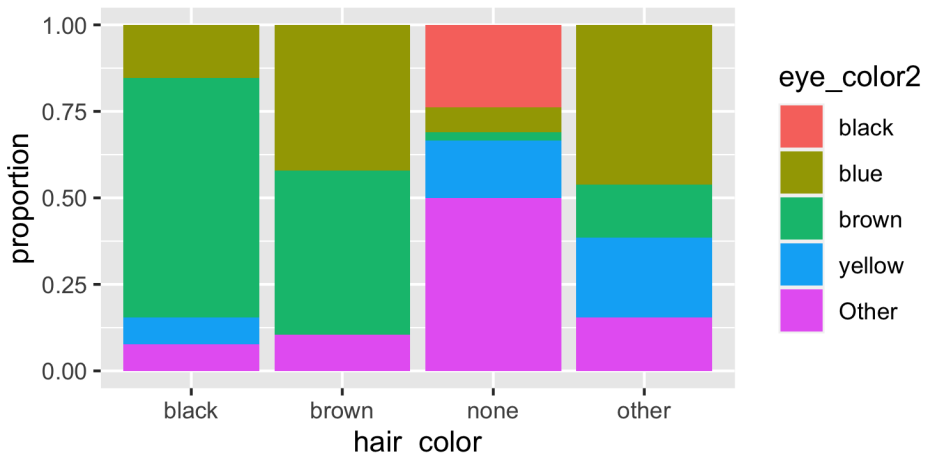
```
ggplot(data = starwars, mapping = aes(x = hair_color, fill = eye_color2)) +
         geom_bar(position = "fill") +
         labs(y = "proportion")
```



Which bar plot is more appropriate?

Which plot is more useful for visualizing the relationship between hair color and eye color? Why?





Data visualization

What is data visualization?

Anything that converts data sources into a visual representation

- charts
- plots
- maps
- tables
- etc.

Source: https://guides.library.duke.edu/datavis

Why do we visualize?

Data: `datasaurus_dozen`

Below is an excerpt from the **`datasaurus_dozen`** dataset:

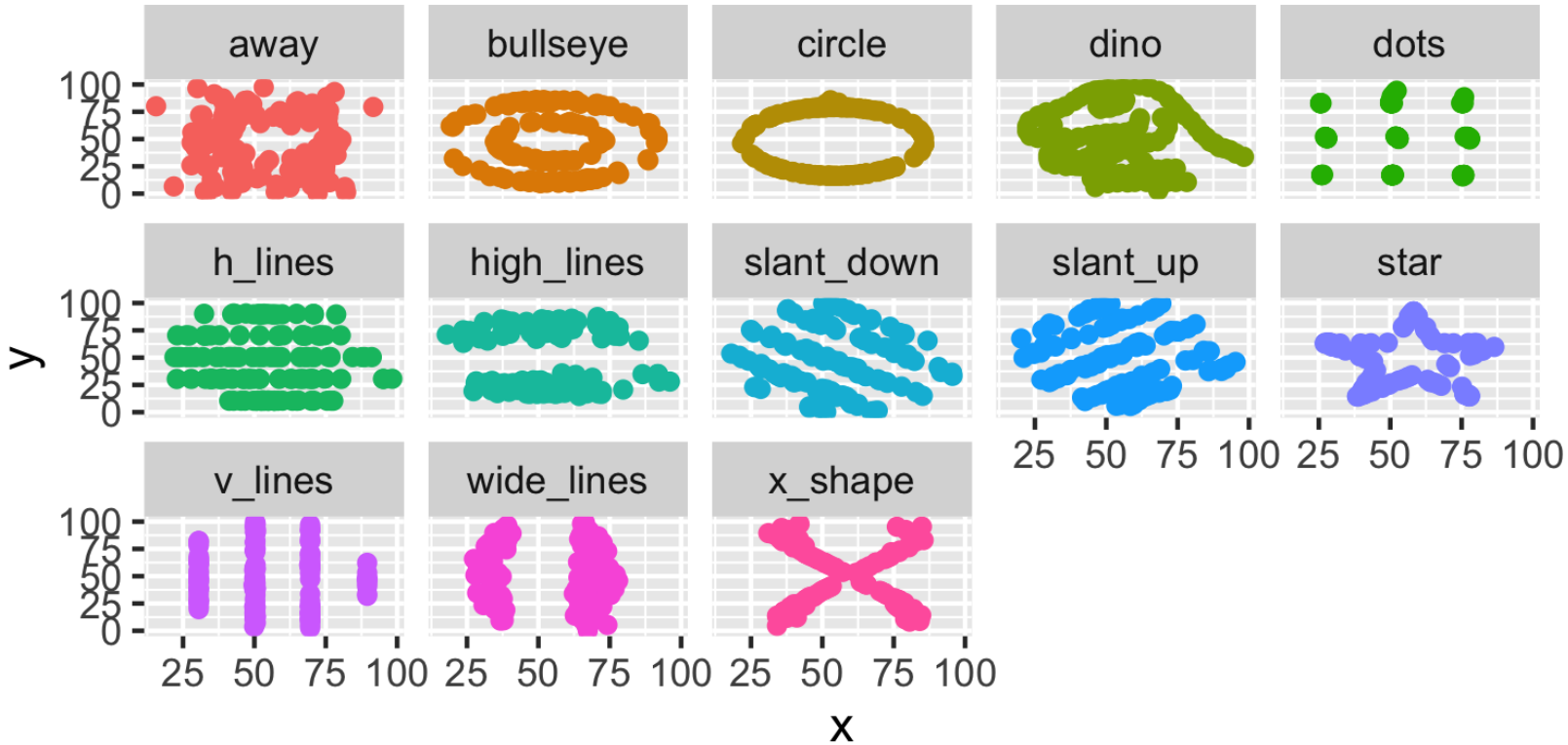
```
## # A tibble: 142 x 8
##    away_x away_y bullseye_x bullseye_y circle_x circle_y dino_x dino_y
##     <dbl>  <dbl>      <dbl>      <dbl>    <dbl>    <dbl>  <dbl>  <dbl>
##  1   32.3   61.4       51.2       83.3     56.0     79.3   55.4   97.2
##  2   53.4   26.2       59.0       85.5     50.0     79.0   51.5   96.0
##  3   63.9   30.8       51.9       85.8     51.3     82.4   46.2   94.5
##  4   70.3   82.5       48.2       85.0     51.2     79.2   42.8   91.4
##  5   34.1   45.7       41.7       84.0     44.4     78.2   40.8   88.3
##  6   67.7   37.1       37.9       82.6     45.0     77.9   38.7   84.9
##  7   53.3   97.5       39.5       80.8     48.6     78.8   35.6   79.9
##  8   63.5   25.1       39.6       82.7     42.1     76.9   33.1   77.6
##  9   68.0   81.0       34.8       80.0     41.0     76.4   29.0   74.5
## 10   67.4   29.7       27.6       72.8     34.6     72.7   26.2   71.4
## # … with 132 more rows
```

Summary statistics

```
datasaurus_dozen %>%
  group_by(dataset) %>%
  summarise(r = cor(x, y))
```

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

How similar do the relationships between **x** and **y** look based on the plots? Based on the summary statistics?



Anscombe's quartet

```
library(Tmisc)
quartet
```

```
##    set  x     y
## 1    I 10  8.04
## 2    I  8  6.95
## 3    I 13  7.58
## 4    I  9  8.81
## 5    I 11  8.33
## 6    I 14  9.96
## 7    I  6  7.24
## 8    I  4  4.26
## 9    I 12 10.84
## 10   I  7  4.82
## 11   I  5  5.68
## 12  II 10  9.14
```

```
##    set  x     y
## 23 III 10  7.46
## 24 III  8  6.77
## 25 III 13 12.74
## 26 III  9  7.11
## 27 III 11  7.81
## 28 III 14  8.84
## 29 III  6  6.08
## 30 III  4  5.39
## 31 III 12  8.15
## 32 III  7  6.42
## 33 III  5  5.73
## 34  IV  8  6.58
```

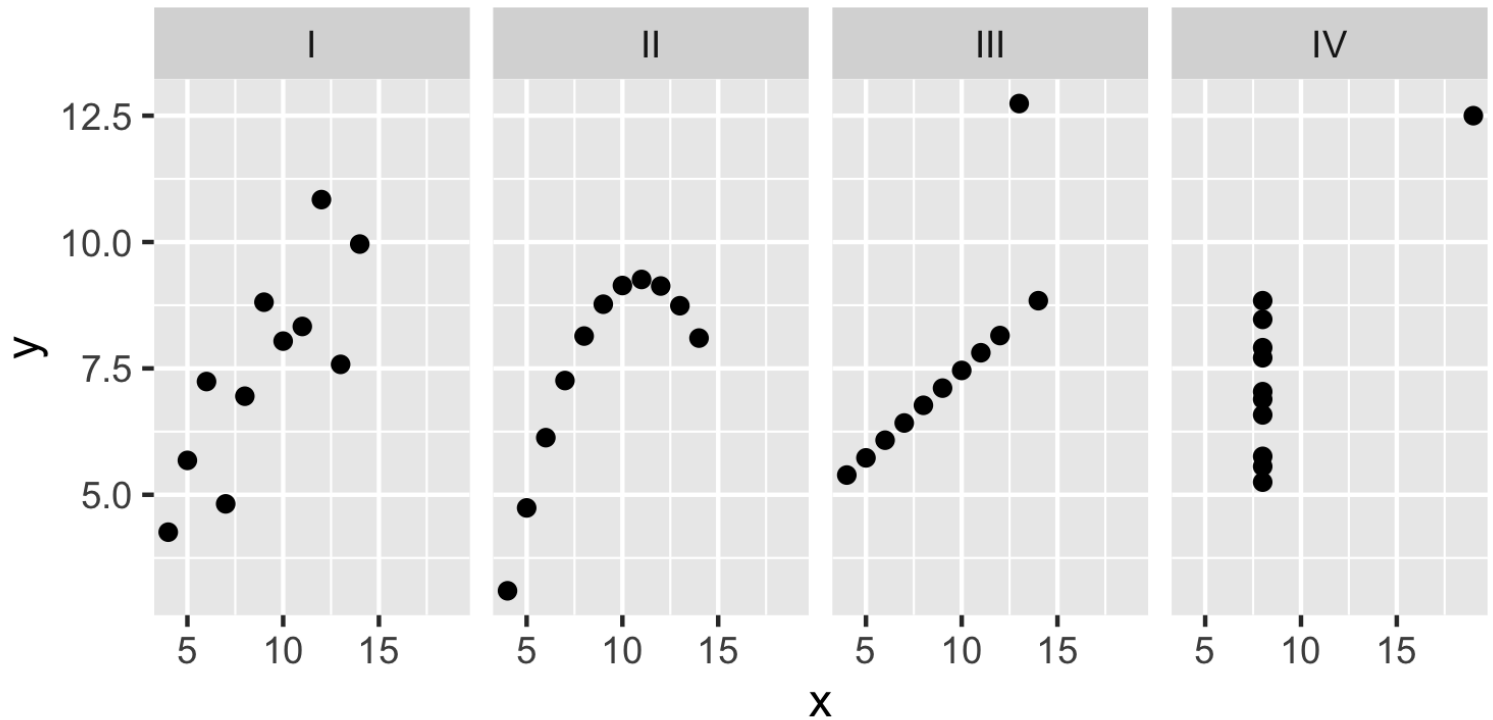
Summarising Anscombe's quartet

```
quartet %>%
  group_by(set) %>%
  summarise(
    mean_x = mean(x), mean_y = mean(y),
    sd_x = sd(x), sd_y = sd(y),
    r = cor(x, y)
  )
```

```
## # A tibble: 4 x 6
##   set   mean_x mean_y  sd_x  sd_y     r
##   <fct>  <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 I          9   7.50  3.32  2.03 0.816
## 2 II         9   7.50  3.32  2.03 0.816
## 3 III        9   7.5   3.32  2.03 0.816
## 4 IV         9   7.50  3.32  2.03 0.817
```

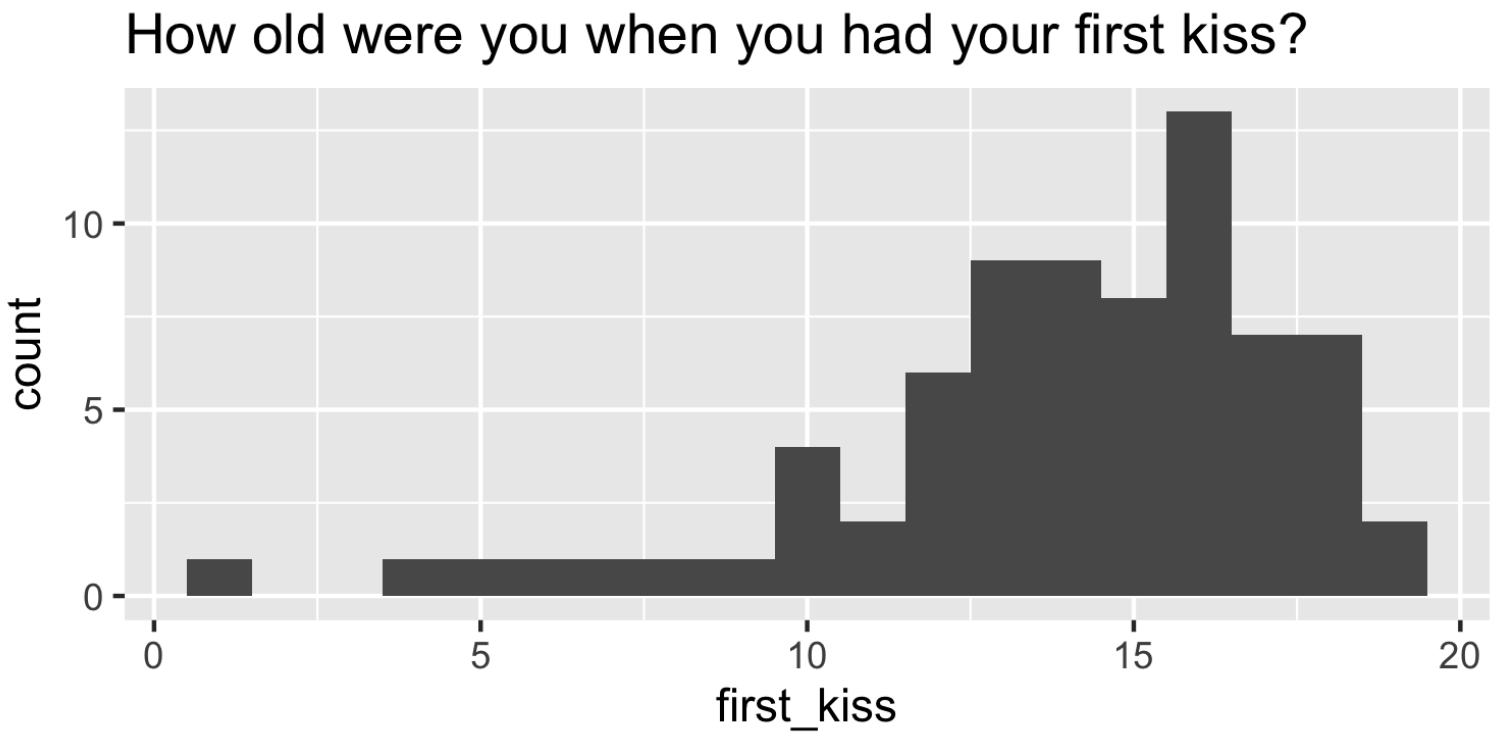
Visualizing Anscombe's quartet

```
ggplot(quartet, aes(x = x, y = y)) +
  geom_point() +
  facet_wrap(~ set, ncol = 4)
```



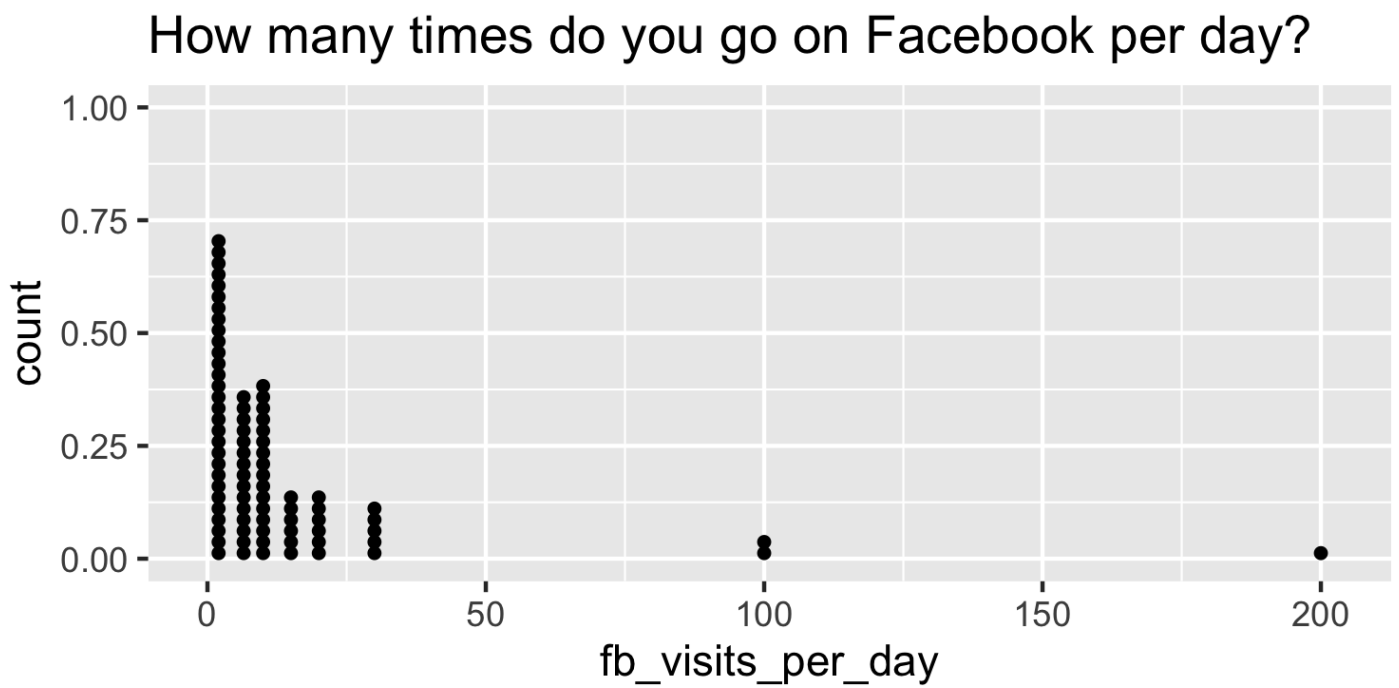
Do you see anything out of the ordinary?

```
ggplot(student_survey, aes(x = first_kiss)) +
  geom_histogram(binwidth = 1) +
  labs(title = "How old were you when you had your first kiss?")
```



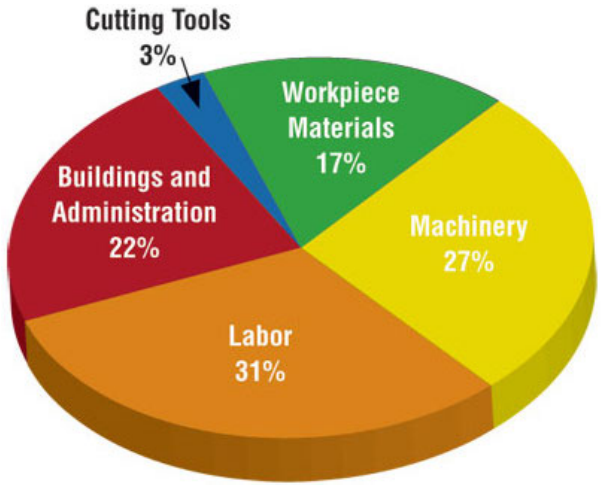
Reporting lower vs. higher values

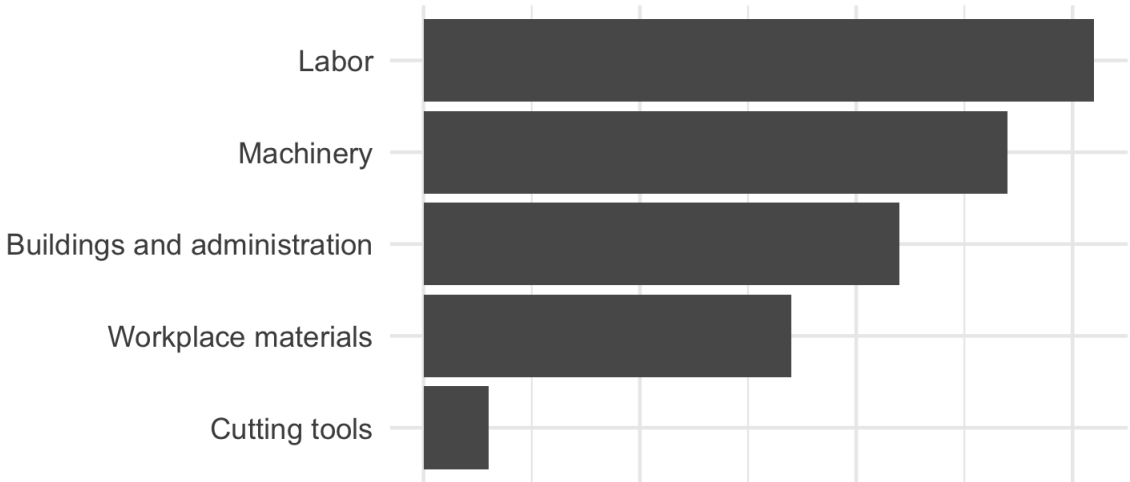
```
ggplot(student_survey, aes(x = fb_visits_per_day)) +
  geom_dotplot(binwidth = 5, dotsize = 0.4) +
  labs(title = "How many times do you go on Facebook per day?")
```



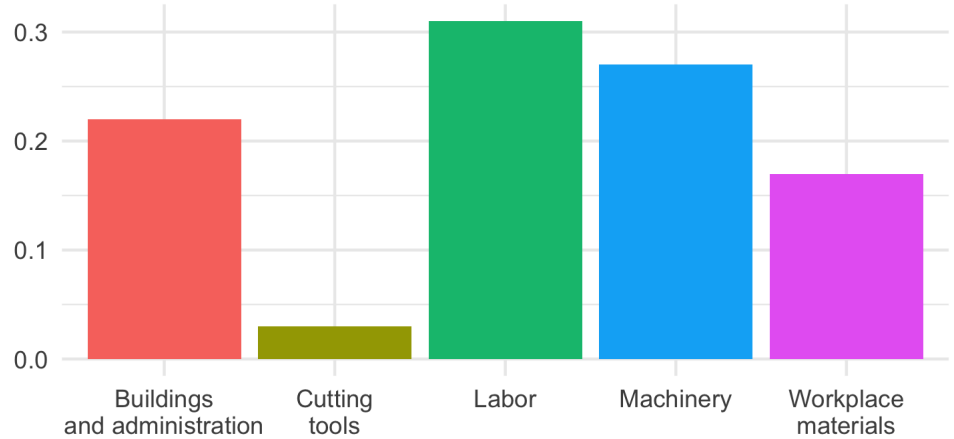
Designing effective visualizations

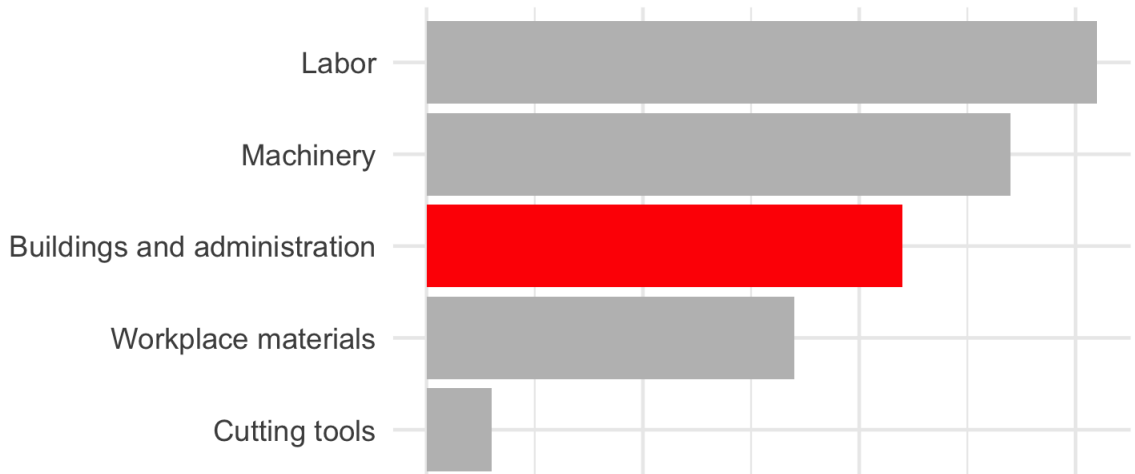
Keep it simple



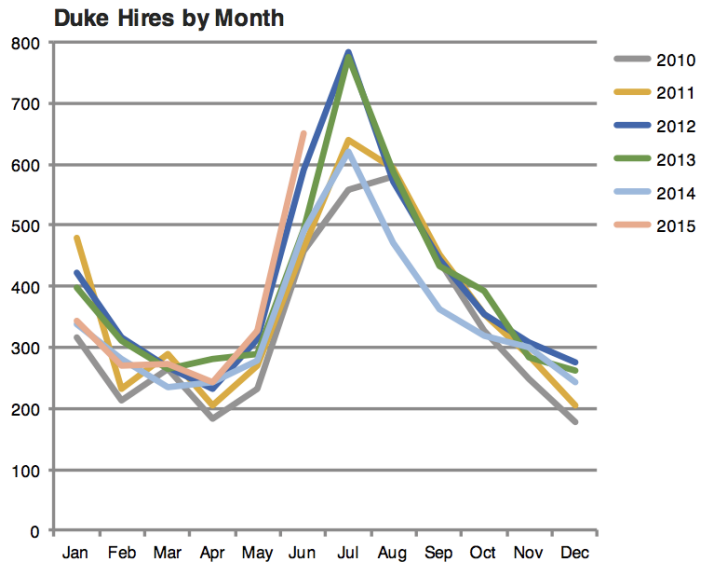


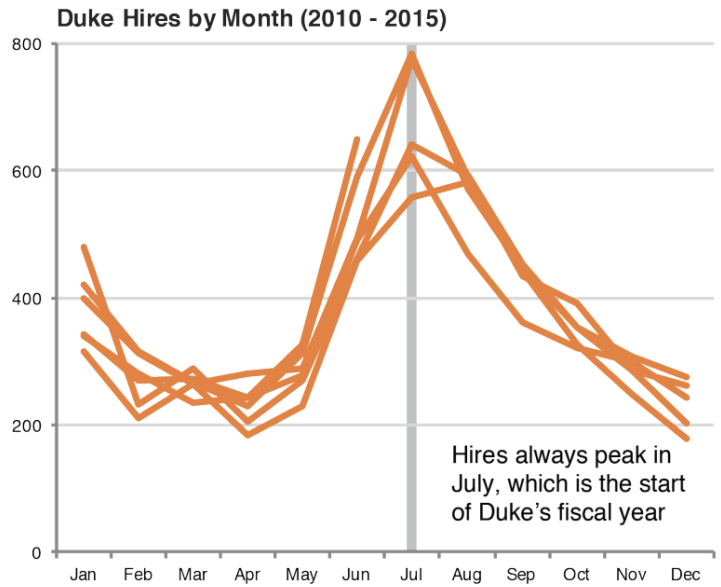
Use color to draw attention





Tell a story





Credit: Angela Zoss and Eric Monson, Duke DVS