ODSC: ggplot2 Graph Gallery

Categories and distributions: distributions

by Martin Frigaard & Peter Spangler

Written: September 21 2021

Updated: April 03 2022

Resources:

The graphs

- The ggplot2 book by Hadley Wickham, Danielle Navarro, and Thomas Lin Pedersen

- Data Visualization: A Practical Introduction by Kieran Healy (2018)

- R Graphics Cookbook, 2nd edition by Winston Chang (2022)

Graph Categories

- Fundamentals of Data Visualization by Claus O. Wilke (2019)

- Data Visualisation: A Handbook for Data Driven Design by Andy Kirk (2019)

- Data Points by Nathan Yau (2013)

Graph Categories: *The 'CHRTS' Families of Chart Types*

From *"Data Visualisation: A Handbook for Data Driven Design"*, Andy Kirk (2019)

Comparing categories and distributions

Hierarchies/part-to-whole relationships

Correlations and connections

Trends and intervals over time

Maps, overlays, and/or distortions

Graph Categories: Directory of Visualizations

From *"Fundamentals of Data Visualization"*, Claus O. Wilke (2019)

Amounts

Distributions

Proportions

X–Y relationships

Geospatial Data

Uncertainty

Comprehensive Graph Gallery

Comparing categories and values

- Amounts
- *Distributions*

Hierarchies and proportions

- Part-to-whole relationships

Trends, correlations and connections

- X–Y relationships

Maps, overlays, and distortions

- Geospatial Data

Statistical measures

- Uncertainty

Data

Data come from the following packages:

- `palmerpenguins`
- `fivethirtyeight`
- `ggplot2movies`

Or created using `tribble()`

```
tribble(
  ~`variable 1`, ~`variable 2`,
           "a",              1,
           "b",              2,
           "c",              3)
```

variable 1 <chr>	variable 2 <dbl>
a	1
b	2
c	3

3 rows

Load data packages


```
library(palmerpenguins)
library(fivethirtyeight)
library(ggplot2movies)
```

palmerpenguins

palmerpenguins package website

```
palmerpenguins::penguins -> penguins
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
<fct>	<fct>	<dbl>	<dbl>	<int>	<int>	<fct>	<int>
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	*NA*	*NA*	*NA*	*NA*	*NA*	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	34.1	18.1	193	3475	*NA*	2007
Adelie	Torgersen	42.0	20.2	190	4250	*NA*	2007

1-10 of 344 rows

Previous **1** 2 3 4 5 6 … 35 Next

fivethirtyeight

fivethirtyeight package website

All datasets are listed below with descriptions

```
datasets("fivethirtyeight")
```

dataset <chr>
US_births_1994_2003
US_births_2000_2014
ahca_polls
airline_safety
antiquities_act
august_senate_polls
avengers
bachelorette
bad_drivers
bechdel

1-10 of 129 rows | 1-1 of 2 columns

Previous **1** 2 3 4 5 6 … 13 Next

ggplot2movies

ggplot2movies package website

We're using `movies_data` (derived version of the `ggplot2movies::movies`)

```
movies_data
```

title <chr>	year <int>	length <int>	budget <int>	rating <dbl>	mpaa <fct>
100 Mile Rule	2002	98	1100000	5.6	R
13 Going On 30	2004	98	37000000	6.4	PG-13
15 Minutes	2001	120	42000000	6.1	R
2 Fast 2 Furious	2003	107	76000000	5.1	PG-13
2046	2004	129	12000000	7.6	R
21 Grams	2003	124	20000000	8.0	R
25th Hour	2002	135	15000000	7.8	R
3000 Miles to Graceland	2001	125	62000000	5.4	R
40 Days and 40 Nights	2002	96	17000000	5.4	R
50 First Dates	2004	99	75000000	6.8	PG-13

1-10 of 751 rows | 1-6 of 7 columns

Previous **1** 2 3 4 5 6 … 76 Next

Comparing Categories and Distributions

Distributions

Distributions: *Histograms*

Histograms use bars, but the `x` axis is divided into 'bins' that cover the range of the variable. The standard number of bins is `30` (but you should experiment to see how many bins fit your variable's distribution). In `ggplot2`, the geom for creating histograms is `geom_histogram()`

Distributions: *Histograms*


```
penguins
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
<fct>	<fct>	<dbl>	<dbl>	<int>	<int>	<fct>	<int>
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	*NA*	*NA*	*NA*	*NA*	*NA*	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	34.1	18.1	193	3475	*NA*	2007
Adelie	Torgersen	42.0	20.2	190	4250	*NA*	2007

1-10 of 344 rows

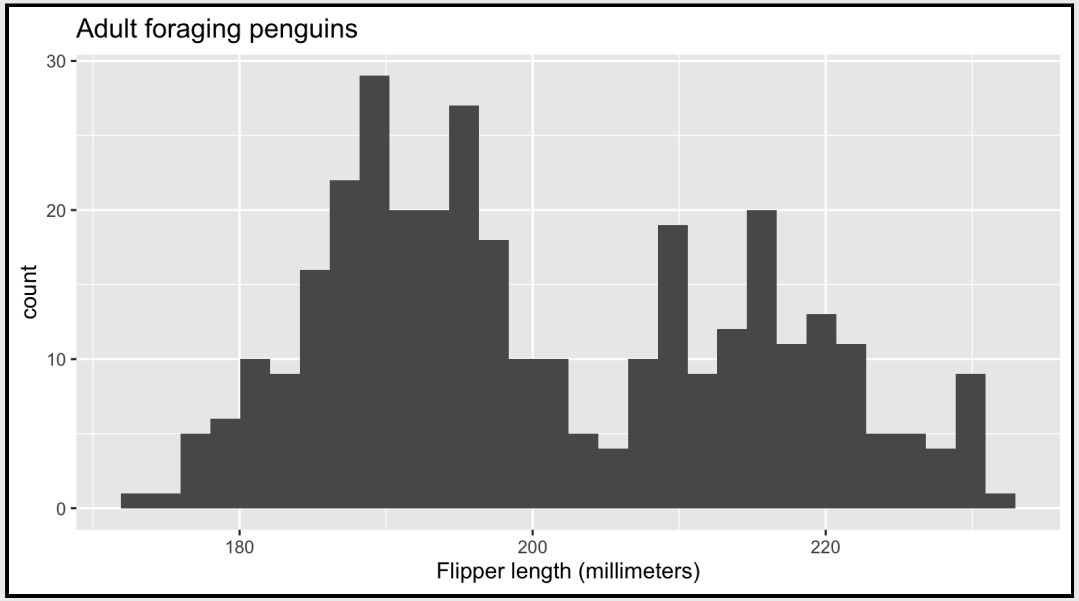
Previous **1** 2 3 4 5 6 … 35 Next

Distributions: *Histograms*

Map `flipper_length_mm` to the `x` axis, add the `geom_histogram()` layer and the labels

```
labs_histogram <- labs(
  x = "Flipper length (millimeters)",
  title = "Adult foraging penguins")
```

```
ggplot(data = penguins,
     aes(x = flipper_length_mm)) +
     geom_histogram() +
     labs_histogram
```



Distributions: *Frequency Polygon*

Frequency polygons (`geom_freqpoly()`) are similar to histograms, but use lines instead of bars to represent the variable distribution.

Distributions: *Frequency Polygon*


```
penguins
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
<fct>	<fct>	<dbl>	<dbl>	<int>	<int>	<fct>	<int>
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	*NA*	*NA*	*NA*	*NA*	*NA*	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	34.1	18.1	193	3475	*NA*	2007
Adelie	Torgersen	42.0	20.2	190	4250	*NA*	2007

1-10 of 344 rows

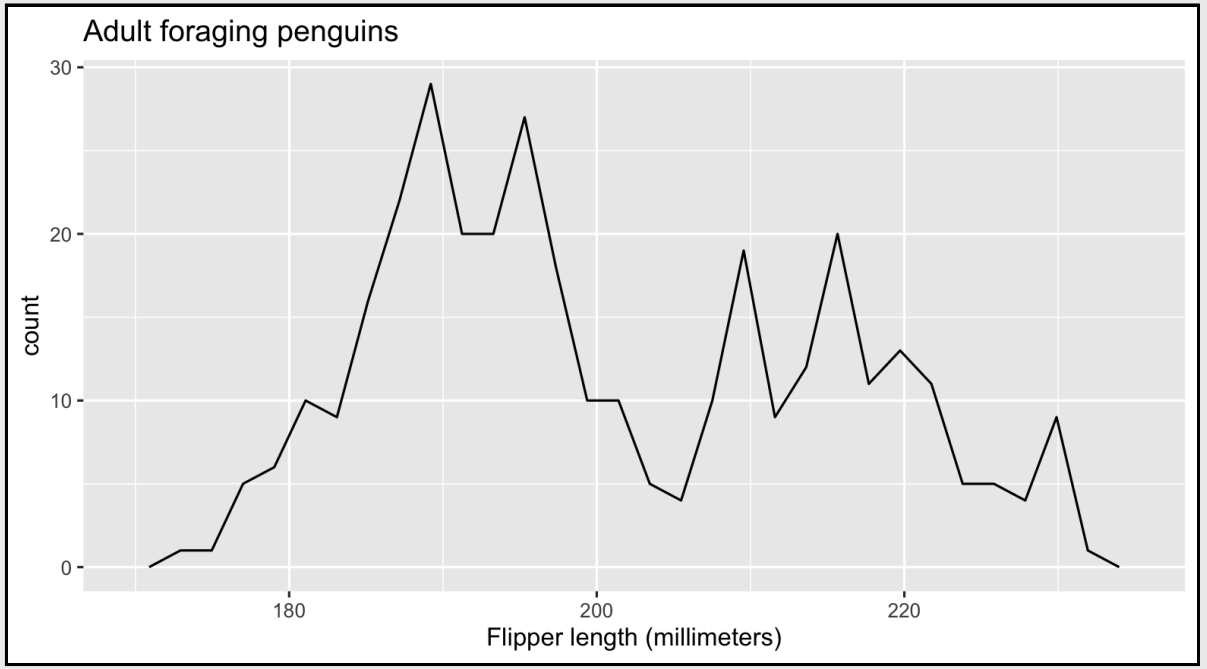
Previous **1** 2 3 4 5 6 … 35 Next

Distributions: *Frequency Polygon*

Map `flipper_length_mm` to the `x` axis, add the `geom_freqpoly()` layer and the labels

```
labs_freqpoly <- labs(
  x = "Flipper length (millimeters)",
  title = "Adult foraging penguins")
```

```
ggplot(data = penguins,
       aes(x = flipper_length_mm)) +
    geom_freqpoly() +
    labs_freqpoly
```

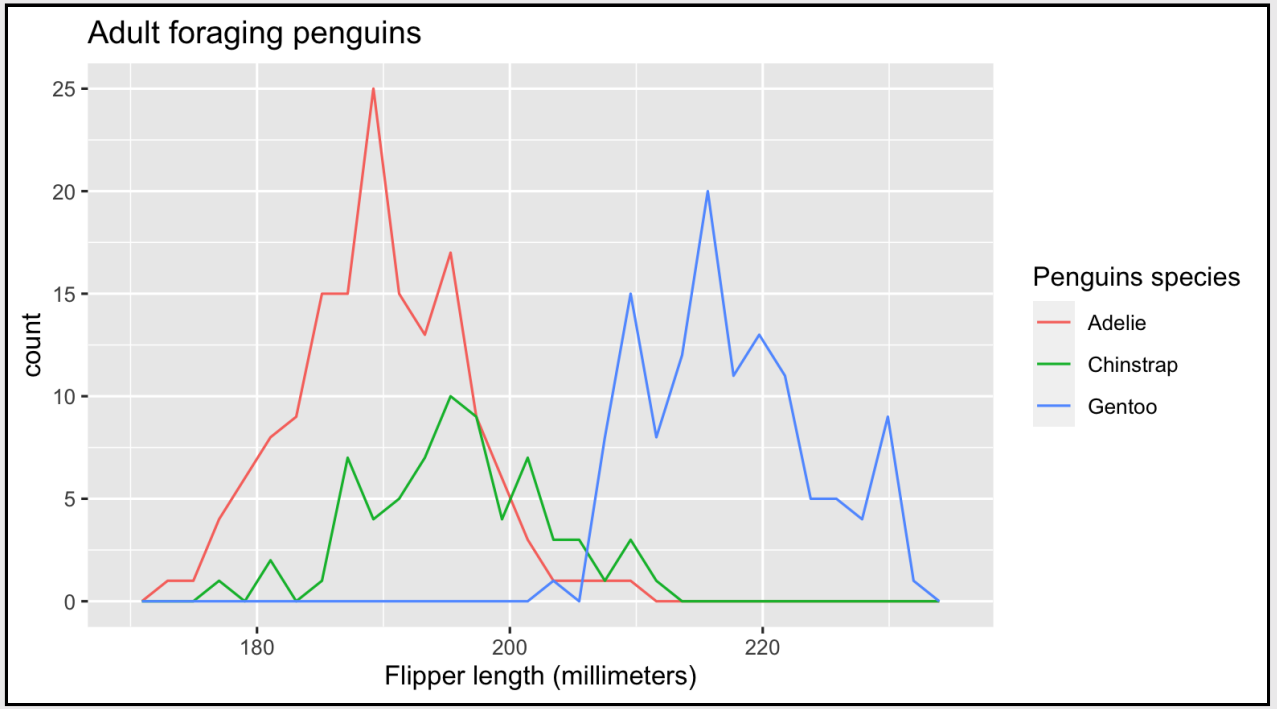


Distributions: *Frequency Polygon*

Frequency polygons are helpful when we want to look at a continuous variable across the levels of a categorical variable

```
labs_freqpoly_2 <- labs(
  x = "Flipper length (millimeters)",
  color = "Penguins species",
  title = "Adult foraging penguins")
```

```
ggplot(data = penguins,
       aes(x = flipper_length_mm)) +
  geom_freqpoly(
    aes(color = species,
        group = species)) +
  labs_freqpoly_2
```



Distributions: *Dot-Plots*

Dot-plots (`geom_dotplot()`) are similar to histograms and frequency polygons, except instead of using bars or lines, they use dots to represent the values of a given variable.

Distributions: *Dot-Plots*


```
penguins
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
<fct>	<fct>	<dbl>	<dbl>	<int>	<int>	<fct>	<int>
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	*NA*	*NA*	*NA*	*NA*	*NA*	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	34.1	18.1	193	3475	*NA*	2007
Adelie	Torgersen	42.0	20.2	190	4250	*NA*	2007

1-10 of 344 rows

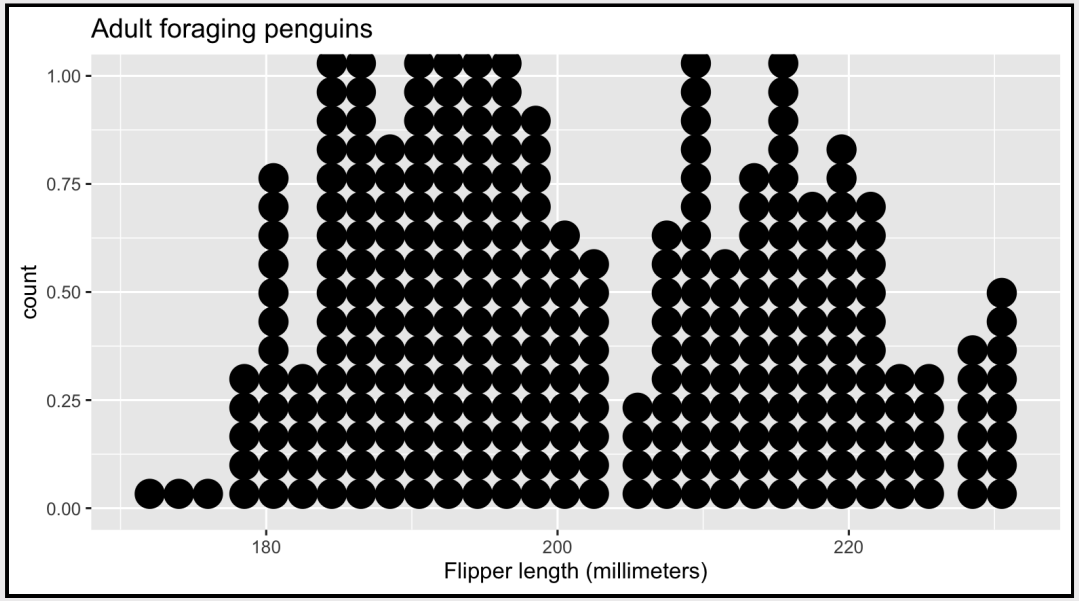
Previous **1** 2 3 4 5 6 … 35 Next

Distributions: *Dot-Plots*

Map `flipper_length_mm` to the `x` axis, add the `geom_dotplot()` layer and the labels

```
labs_dotplot <- labs(
  x = "Flipper length (millimeters)",
  title = "Adult foraging penguins")
```

```
ggplot(data = penguins,
    aes(x = flipper_length_mm)) +
    geom_dotplot() +
    labs_dotplot
```



Distributions: *Dot-Plots*


```
penguins_histodot <- filter(penguins, !is.na(sex))
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
<fct>	<fct>	<dbl>	<dbl>	<int>	<int>	<fct>	<int>
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	41.1	17.6	182	3200	female	2007
Adelie	Torgersen	38.6	21.2	191	3800	male	2007
Adelie	Torgersen	34.6	21.1	198	4400	male	2007

1-10 of 333 rows

Previous **1** 2 3 4 5 6 … 34 Next

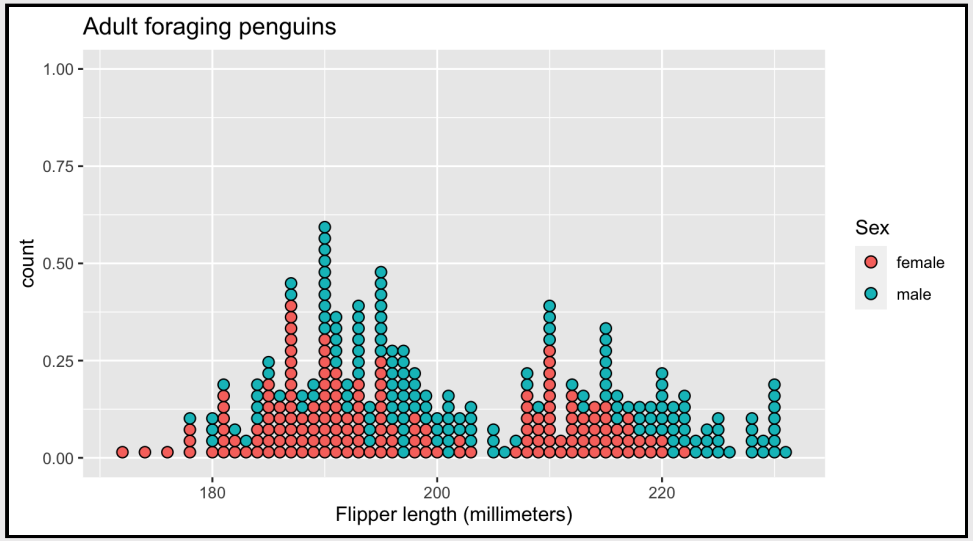
Distributions: *Dot-Plots*

We can also use dot-plots to look at the range of a continuous (numerical) variable across the levels of a categorical (character) variable (like `sex` below).

The default setting for the size of the dots is '1/30 of the range of the data.' We can adjust the size with `binwidth` (and `method = "histodot"`)

```
labs_histodot <- labs(
  x = "Flipper length (millimeters)",
  fill = "Sex",
  title = "Adult foraging penguins")
```

```
ggplot(data = penguins_histodot,
    aes(x = flipper_length_mm,
        fill = factor(sex))) +
  geom_dotplot(
        stackgroups = TRUE,
        binwidth = 1,
        method = "histodot") +
  labs_histodot
```



Distributions: *Bee-swarm Plots*

We can also use dots to show the spread of values for a particular variable with `bee-swarm` plots. These display the distribution of numeric values across the levels of a categorical variable.

Distributions: *Bee-swarm Plots*


```
penguins
```

title <chr>	year <int>	length <int>	budget <int>	rating <dbl>	mpaa <fct>
100 Mile Rule	2002	98	1100000	5.6	R
13 Going On 30	2004	98	37000000	6.4	PG-13
15 Minutes	2001	120	42000000	6.1	R
2 Fast 2 Furious	2003	107	76000000	5.1	PG-13
2046	2004	129	12000000	7.6	R
21 Grams	2003	124	20000000	8.0	R
25th Hour	2002	135	15000000	7.8	R
3000 Miles to Graceland	2001	125	62000000	5.4	R
40 Days and 40 Nights	2002	96	17000000	5.4	R
50 First Dates	2004	99	75000000	6.8	PG-13

1-10 of 751 rows | 1-6 of 7 columns

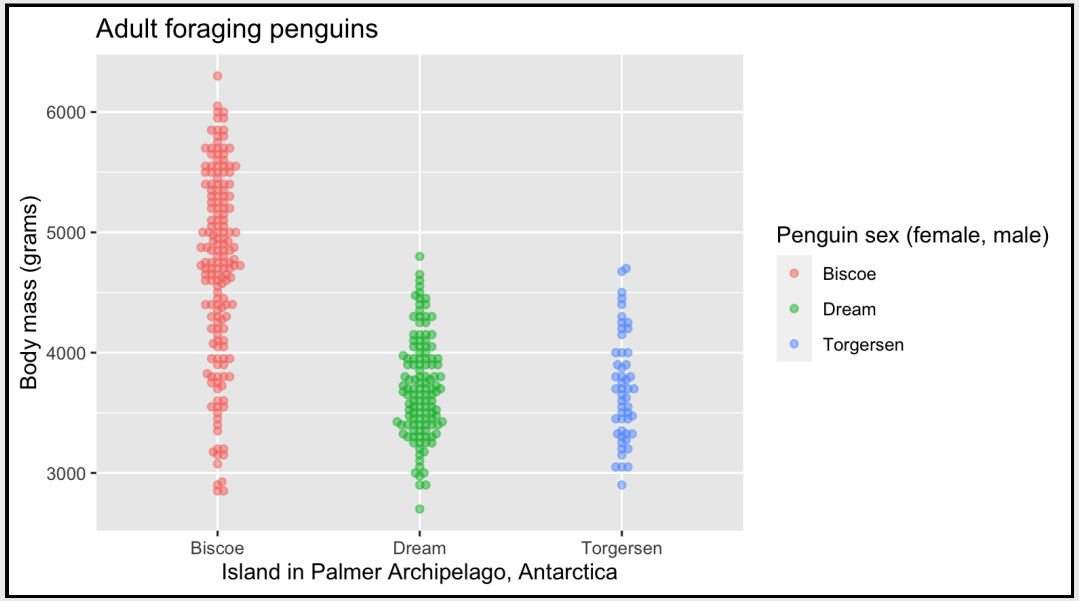
Previous **1** 2 3 4 5 6 … 76 Next

Distributions: *Bee-swarm Plots*

Map `island` to the `x` axis and `color`, `body_mass_g` to the `y` axis, the `geom_beeswarm()` layer (with `alpha`), and the labels

```
labs_beeswarm <- labs(
  x = "Island in Palmer Archipelago, Antarctica",
  y = "Body mass (grams)",
  color = "Penguin sex (female, male)",
  title = "Adult foraging penguins")
```

```
ggplot(data = penguins,
       aes(x = island,
           y = body_mass_g,
           color = island)) +
  ggbeeswarm::geom_beeswarm(
      alpha = 1/2) +
  labs_beeswarm
```



Distributions: *Density Plots*

Density plots are similar to frequency polygons and histograms, except the line has been 'smoothed.' Instead of dividing the x axis into discrete quantitative 'bins' to create groups for the variable values, density plots transform the distribution according to a 'bandwidth' parameter.

Distributions: *Density Plots*


```
penguins
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
<fct>	<fct>	<dbl>	<dbl>	<int>	<int>	<fct>	<int>
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	*NA*	*NA*	*NA*	*NA*	*NA*	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	34.1	18.1	193	3475	*NA*	2007
Adelie	Torgersen	42.0	20.2	190	4250	*NA*	2007

1-10 of 344 rows

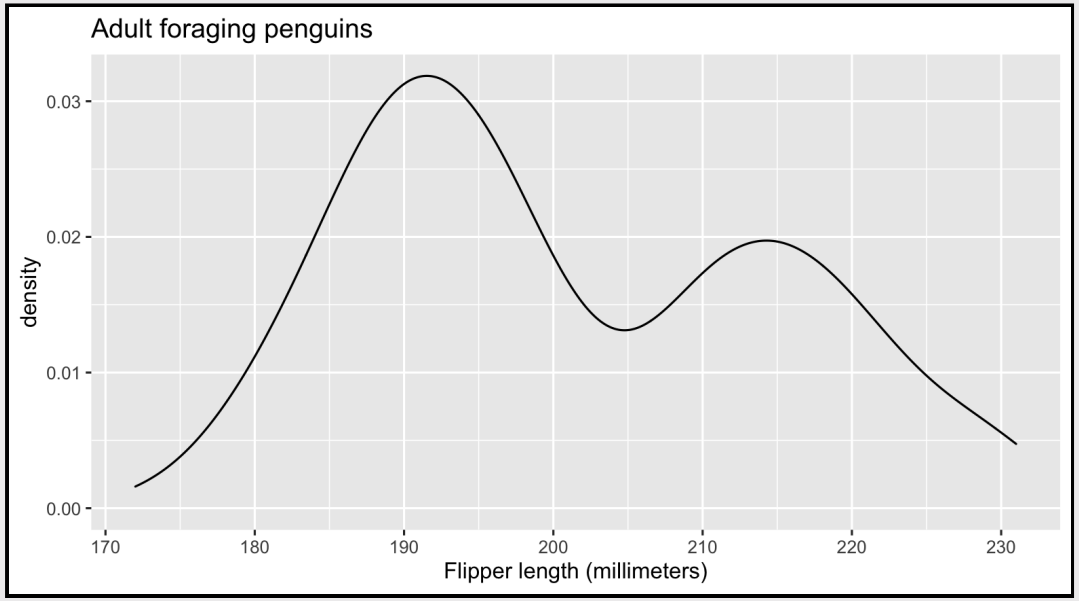
Previous **1** 2 3 4 5 6 … 35 Next

Distributions: *Density Plots*

Map `flipper_length_mm` to the `x` axis, add the `geom_density()` layer and the labels

```
labs_density <- labs(
  x = "Flipper length (millimeters)",
  title = "Adult foraging penguins")
```

```
ggplot(data = penguins,
       aes(x = flipper_length_mm)) +
       geom_density() +
       labs_density
```



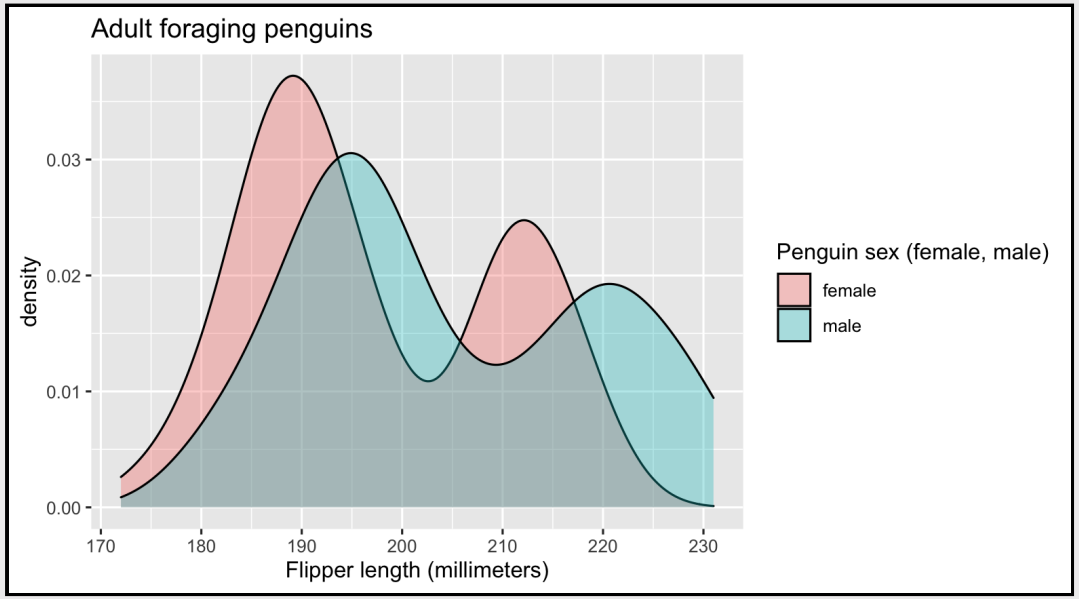
Distributions: *Density Plots*

Similar to frequency polygons, `geom_density()` is useful when we want to look at the distribution of a continuous variable across the levels of a categorical variable

We can set the `fill` argument to a categorical variable, and use the `alpha` to handle the overlapping areas.

```
labs_density_alpha <- labs(
  x = "Flipper length (millimeters)",
  fill = "Penguin sex (female, male)",
  title = "Adult foraging penguins")
```

```
# remove missing sex
penguins_density <- filter(penguins, !is.na(sex))
ggplot(data = penguins_density,
       aes(x = flipper_length_mm,
        fill = sex)) +
        geom_density(alpha = 1/3) +
  labs_density_alpha
```



Distributions: *Ridgeline Plots*

If we want to plot density curves but retain the interpretability of the axes, consider comparing multiple distributions using `ridgeline plots` (`geom_density_ridges()`)

Distributions: *Ridgeline Plots*


```
penguins
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
<fct>	<fct>	<dbl>	<dbl>	<int>	<int>	<fct>	<int>
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	*NA*	*NA*	*NA*	*NA*	*NA*	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	34.1	18.1	193	3475	*NA*	2007
Adelie	Torgersen	42.0	20.2	190	4250	*NA*	2007

1-10 of 344 rows

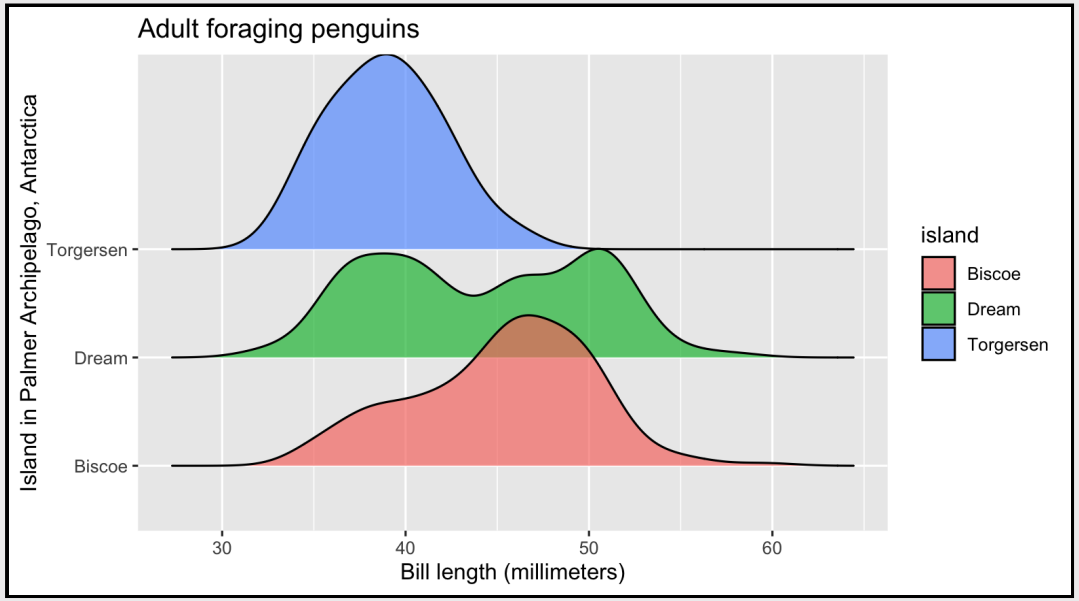
Previous **1** 2 3 4 5 6 … 35 Next

Distributions: *Ridgeline Plots*

Map `bill_length_mm` to the `x` axis, `island` to the `y` axis and `fill`, the `geom_density_ridges()` layer (with `alpha`) and the labels

```
labs_density_ridges <- labs(
  x = "Bill length (millimeters)",
  y = "Island in Palmer Archipelago, Antarctica",
  title = "Adult foraging penguins")
```

```
# remove missing island
penguins_density_ridges <- filter(penguins,
!is.na(island))
ggplot(data = penguins_density_ridges,
       aes(x = bill_length_mm,
           y = island,
         fill = island)) +
    # adjust alpha
  ggridges::geom_density_ridges(alpha = 2/3) +
  labs_density_ridges
```



Distributions: *Box-plots*

Box-plots (sometimes called box-and-whisker plots) are great because they display a collection of statistics in a single graph. We're going to build a box-plot of a single numeric variable and review it's contents.

Distributions: *Box-plots*


```
movies_data
```

title <chr>	year <int>	length <int>	budget <int>	rating <dbl>	mpaa <fct>
100 Mile Rule	2002	98	1100000	5.6	R
13 Going On 30	2004	98	37000000	6.4	PG-13
15 Minutes	2001	120	42000000	6.1	R
2 Fast 2 Furious	2003	107	76000000	5.1	PG-13
2046	2004	129	12000000	7.6	R
21 Grams	2003	124	20000000	8.0	R
25th Hour	2002	135	15000000	7.8	R
3000 Miles to Graceland	2001	125	62000000	5.4	R
40 Days and 40 Nights	2002	96	17000000	5.4	R
50 First Dates	2004	99	75000000	6.8	PG-13

1-10 of 751 rows | 1-6 of 7 columns

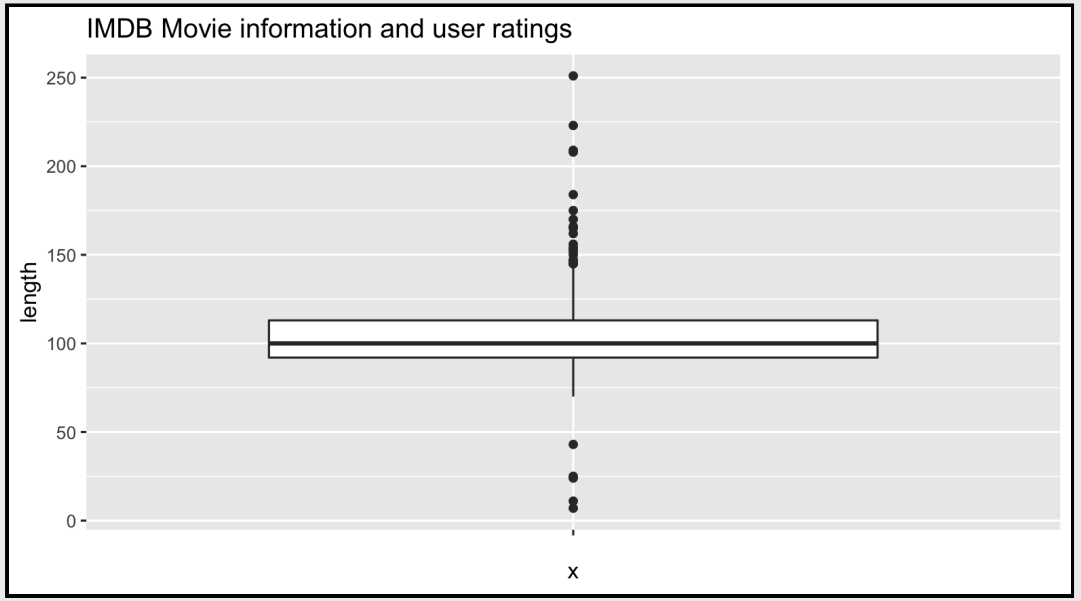
Previous **1** 2 3 4 5 6 … 76 Next

Distributions: *Box-plots*

Map a blank character string (`" "`) to the `x` axis, `length` to the `y` axis, the `geom_boxplot()` layer, and the labels

```
labs_boxplot <- labs(
  y = "length",
   title = "IMDB Movie information and user ratings")
```

```
ggplot(data = movies_data,
    # place an empty string in the
    # x axis
           aes(x = " ",
    # place the length on the y
               y = length)) +
    geom_boxplot() +
    labs_boxplot
```



Distributions: *Box-plots*

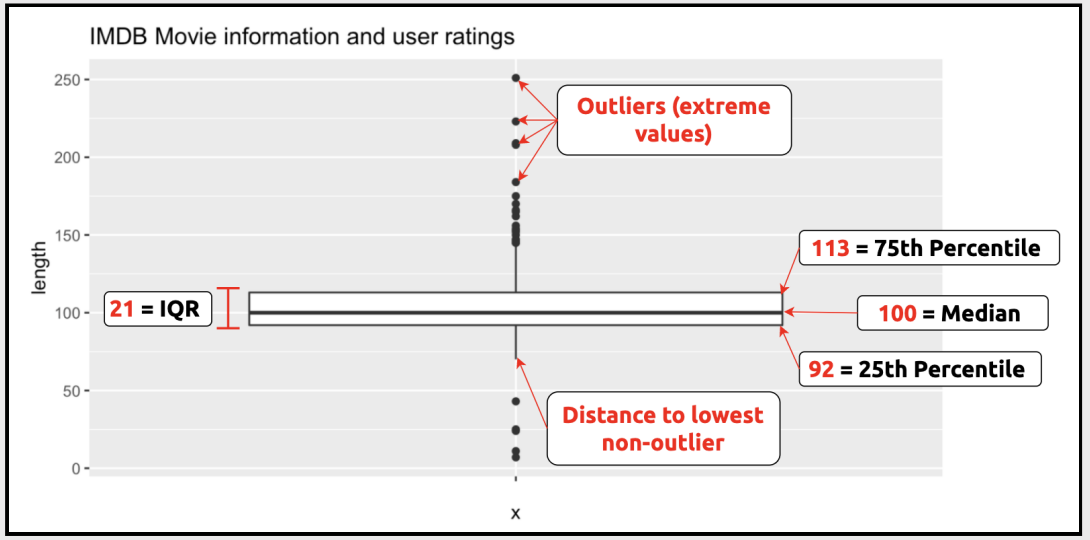
The table below shows the 25th percentile, the median, the 75th percentile, the IQR, and a histogram of the `length` column from the `movies_data` dataset.

25th	Median	75th	IQR	Histogram
<dbl>	<dbl>	<dbl>	<dbl>	<chr>
92	100	113	21	▁▇▃▁

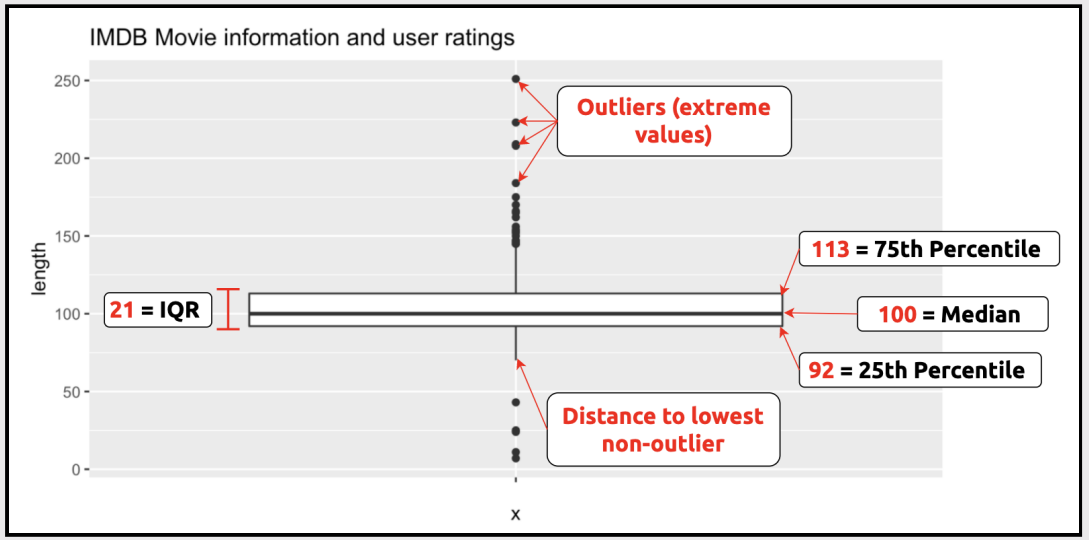
1 row

These three horizontal lines give us a picture of the 'spread' of the data. If there is equal distance on either side of the middle (`Median`) line, this tells us the distribution is symmetrical.

Use these numbers to help you interpret the structure of the box-plot.



Distributions: *Box-plots*



As we can see, the box-plot combines multiple summary statistics.

The 25th percentile (first quartile), the median (50th percentile or second quartile), and the 75th percentile (third quartile) values are common to all box-plots.

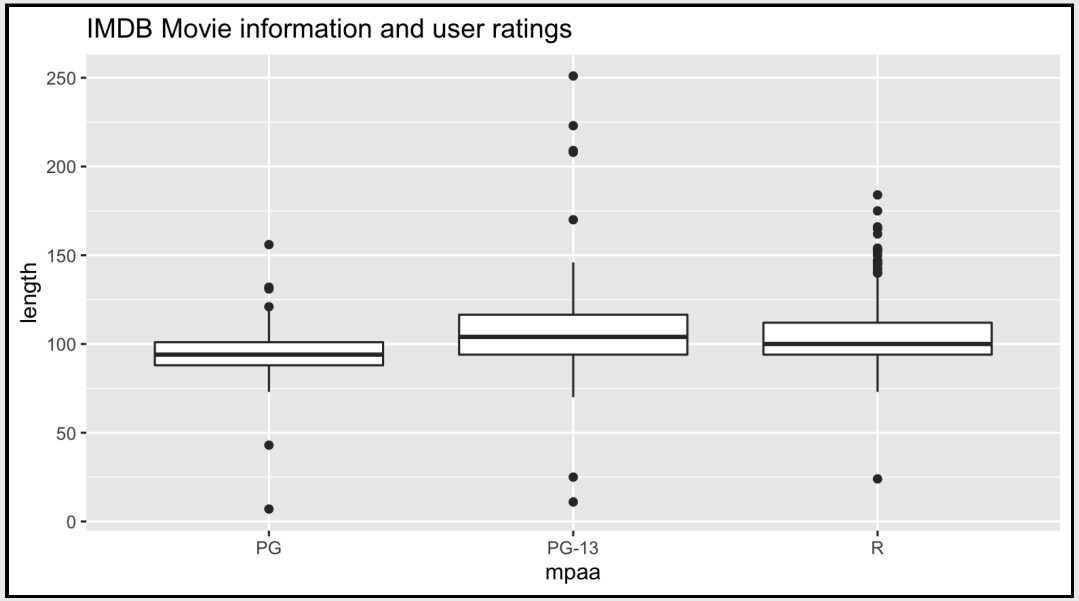
In `ggplot2`, values that fall more than 1.5 times the IQR are displayed as individual points (aka outliers). The lines extending from the bottom and top of the main box represent the last non-outlier value in the distribution.

Distributions: *Box-plots*

Because box-plot provides so many helpful statistical measures, they are also helpful for viewing how a continuous variable varies across a categorical variable

```
labs_boxplots <- labs(
  x = "mpaa",
  y = "length",
   title = "IMDB Movie information and user ratings")
```

```
ggplot(data = movies_data,
    # place an empty string in the
    # x axis
           aes(x = mpaa,
    # place the length on the y
               y = length)) +
    geom_boxplot() +
    labs_boxplots
```



Distributions: *Box-plots*

Compare the four graphs of `length` from `movie_data` below to the box-plot: