Data visualisation with R

Statistical Computing and Empirical Methods
Unit EMATM0061, Data Science MSc

Rihuan Ke
rihuan.ke@bristol.ac.uk

Teaching Block 1, 2024

University of BRISTOL

What will we cover today

We discuss why visualisation is a crucial skill for data scientists.

We consider the difference between various visual cues within plots.

We will take a brief look at the ggplot2 library within R.

We will also think about basic data types and shapes.

The importance of visualisation

1. Exploring data

Many people are skilled at thinking visually.

Plotting data is often the fastest way to gain insights

- Identifying outliers
- Determining the “shape” of a data distribution
- Identifying relationships between variables
- Spotting trends over time

The importance of visualisation

2. Communicating your insights:

Data scientists must do more than understand and gain insight from data.

That insight must also be communicated to others within their organization.

Remember that your audience is often:

- very short on time
- from a non-technical background.

Effective visualisations often allow us to bridge that gap.

A case study: The Challenger

In January 1986 the Challenger rocket was due to be launched by NASA.

A group of engineers who designed motors for NASA requested a delay.

It was argued that the rubber O-rings would not withstand the cold.

The advice was disregarded with dire consequences.

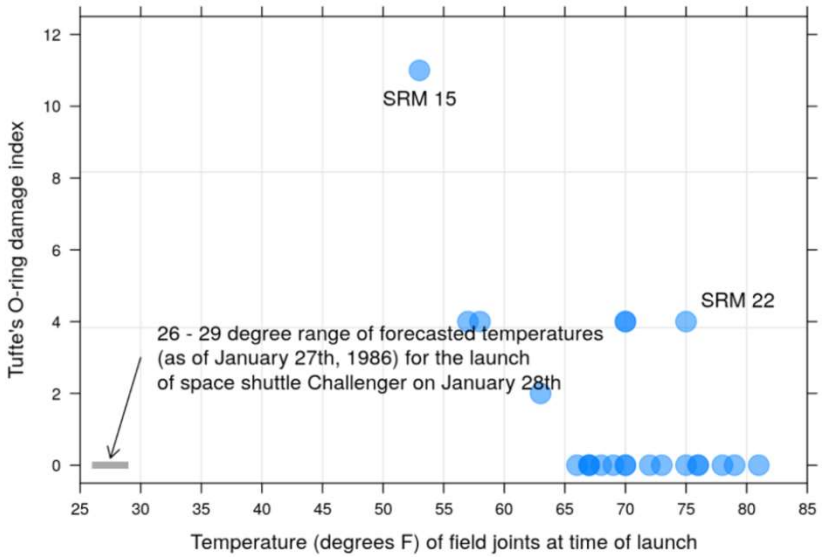
The rocket exploded 73 seconds after the launch.

A case study: The Challenger

Tufte (1997) has argued that this could have been avoided by a better presentation.

HISTORY OF O-RING TEMPERATURES (DEGREES - F)

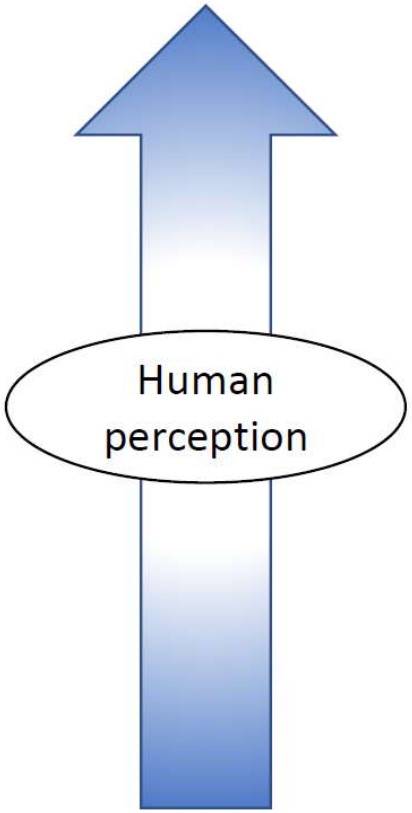
MOTOR	MBT	AMB	O-RING	WIND
QM-4	68	36	47	10 MPH
QM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH



Visual cues

Visual cues are components of a plot or graph which draw the attention of your audience.

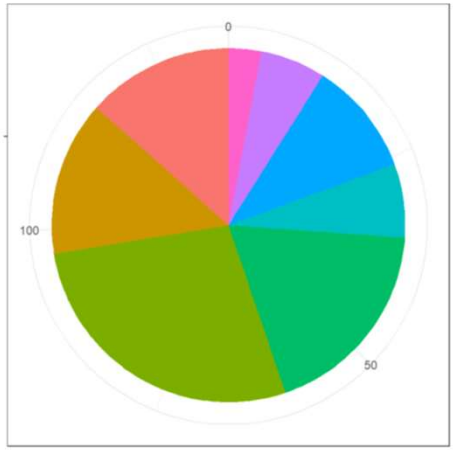
1. **Position** (numerical): Where in relation to other things?
2. **Length** (numerical): How large (in one dimension)?
3. **Angle** (numerical): How wide is
4. **Direction** (numerical): At what slope?
5. **Shape** (numerical): Which group?
6. **Area** (numerical): How big (in two dimensions)?
7. **Volume** (numerical): How big (in three dimensions)?
8. **Shade** (numerical or categorical): How dark is something?
9. **Colour** (numerical or categorical): What colour is something?



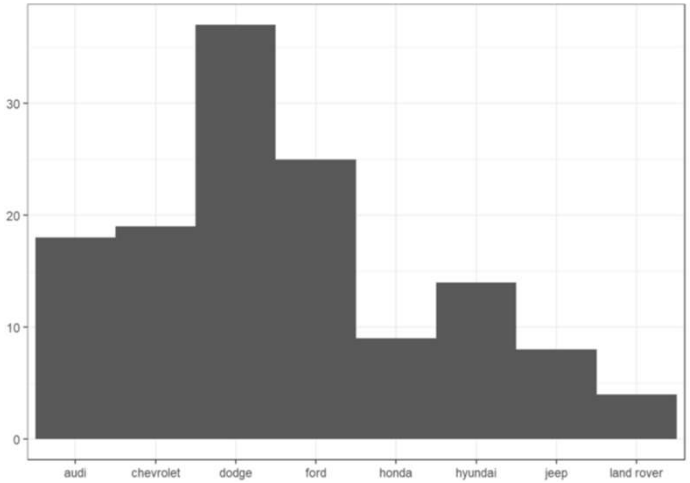
Visual cues

Visual cues are components of a plot or graph which draw the attention of your audience.

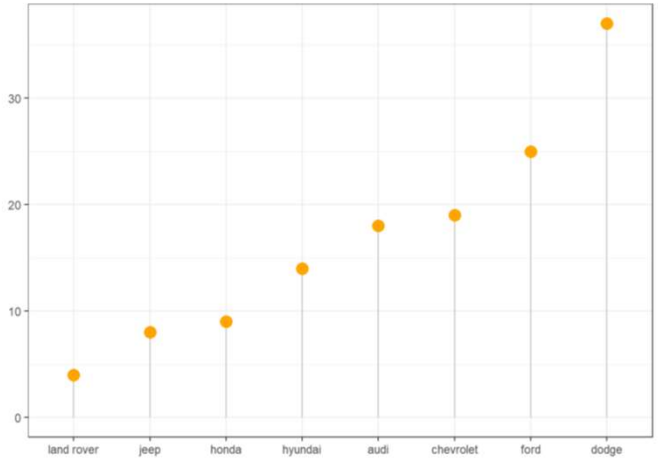
Three figures for the number of cars sold over a given month, broken down by the manufacturer.



Pie chart



Bar chart



Lollipop chart

Which of these plots do you think is easiest to interpret?

Visualisation in R with ggplot2

Hadley Wickham's **ggplot2** package allows us to quickly generate impressive plots within R.

The **ggplot2** package implements Leland Wilkinson's Grammar of Graphics:

1. An aesthetic is a mapping between a variable and a visual cue.
2. A glyph is a basic graphical element e.g. a mark or symbol.
3. A guide is an annotation which provides context.

The **ggplot2** package is included in the tidyverse package. To use it, first:

```
library(tidyverse)
```

The Palmer penguins data set

First load the palmer penguins library

We can take a look at the data set by using the head function.

```
library(palmerpenguins)
head(penguins)
```

```
## # A tibble: 6 × 8
##   species island    bill_length_mm bill_depth_mm flipper_l…¹ body_…² sex    year
##   <fct>   <fct>              <dbl>         <dbl>       <int>   <int> <fct> <int>
## 1 Adelie  Torgersen           39.1          18.7         181    3750 male   2007
## 2 Adelie  Torgersen           39.5          17.4         186    3800 fema…  2007
## 3 Adelie  Torgersen           40.3          18           195    3250 fema…  2007
## 4 Adelie  Torgersen           NA            NA            NA      NA <NA>   2007
## 5 Adelie  Torgersen           36.7          19.3         193    3450 fema…  2007
## 6 Adelie  Torgersen           39.3          20.6         190    3650 male   2007
## # … with abbreviated variable names ¹flipper_length_mm, ²body_mass_g
```

Types of variables

```
## # A tibble: 6 × 8
##   species island    bill_length_mm bill_depth_mm flipper_l…¹ body_…² sex    year
##   <fct>   <fct>              <dbl>         <dbl>       <int>   <int> <fct> <int>
## 1 Adelie  Torgersen           39.1          18.7         181    3750 male   2007
## 2 Adelie  Torgersen           39.5          17.4         186    3800 fema…  2007
## 3 Adelie  Torgersen           40.3          18           195    3250 fema…  2007
## 4 Adelie  Torgersen           NA            NA            NA      NA <NA>   2007
## 5 Adelie  Torgersen           36.7          19.3         193    3450 fema…  2007
## 6 Adelie  Torgersen           39.3          20.6         190    3650 male   2007
## # … with abbreviated variable names ¹flipper_length_mm, ²body_mass_g
```

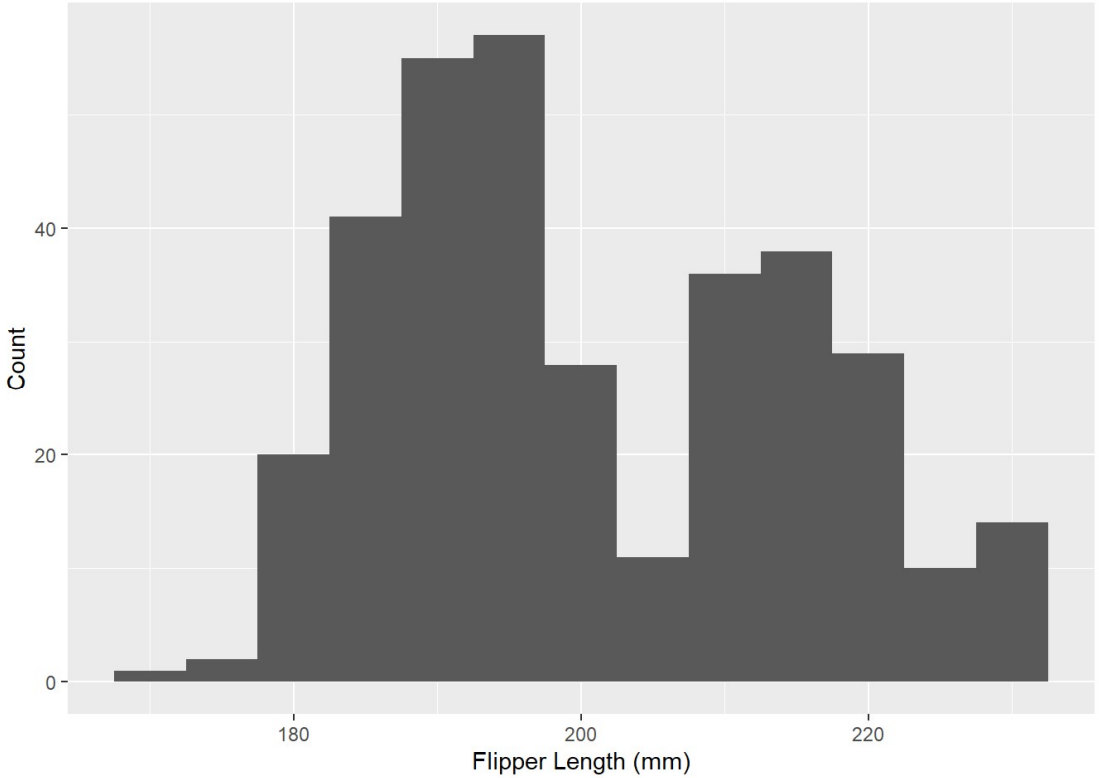
Continuous Numeric variables that can take any value on an interval e.g. Bill length, Bill depth

Discrete Numeric variables for which there is a minimum gap between possible values.
e.g. year the observation was recorded.

Categorical Variables that can take on only a specific set of values representing distinct categories
e.g. species, island, etc.

Univariate plots

```
univar_plot <- ggplot(data=penguins, aes(x=flipper_length_mm)) + xlab("Flipper Length (mm)")
univar_plot+geom_histogram(binwidth = 5)+ylab("Count")
```



Aesthetic

A mapping between a variable and a visual cue.

Flipper length → horizontal position.

Guide

An annotation which provides context.

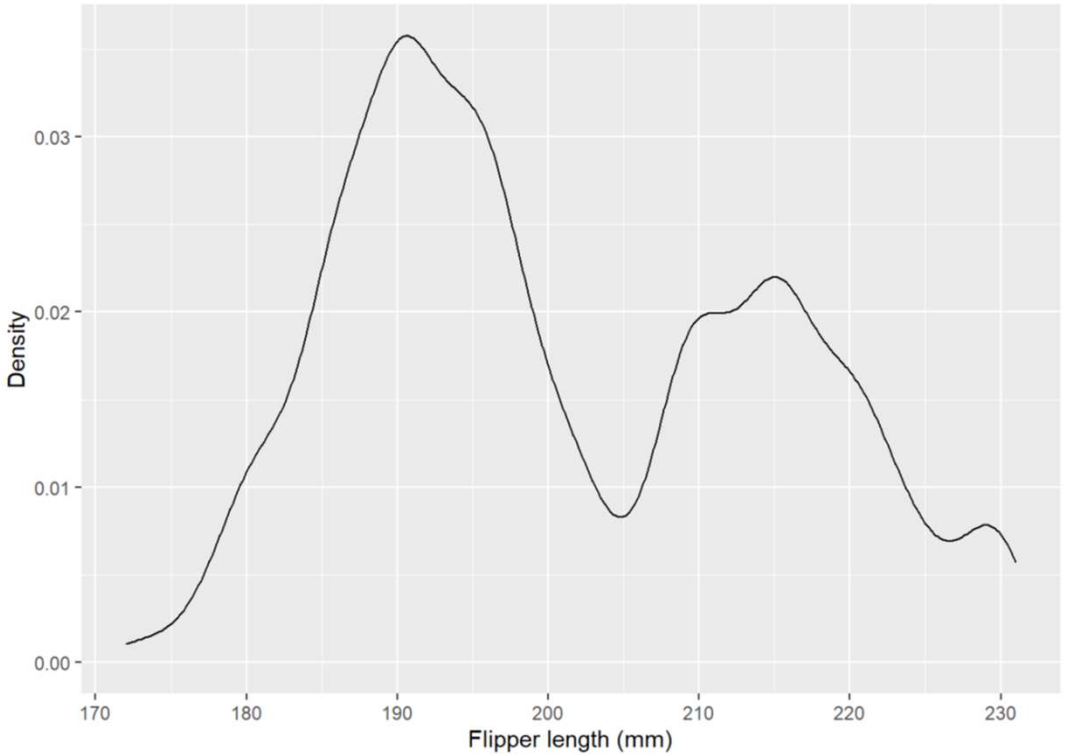
Glyph

A glyph is a basic graphical element.

Each bar represents the number of penguins with flipper lengths within the window.

Univariate plots

```
univar_plot+geom_density(adjust=0)+ylab('Density')
```



Aesthetic

A mapping between a variable and a visual cue.

Flipper length → horizontal position.

Glyph

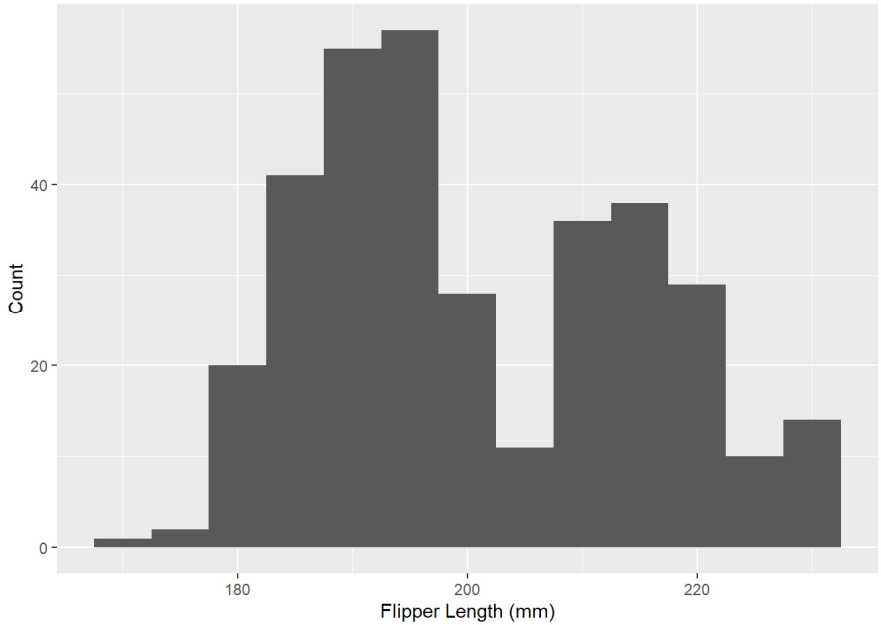
A glyph is a basic graphical element.

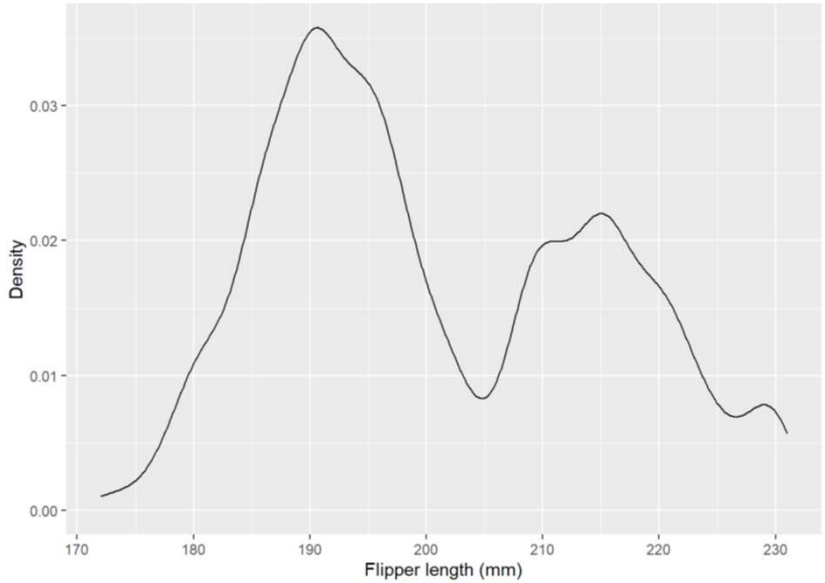
The line within the density plot.

A density plot is a smoothed analogue of a histogram.

Counts are replaced with smoothed bump functions i.e., kernels

Univariate plots

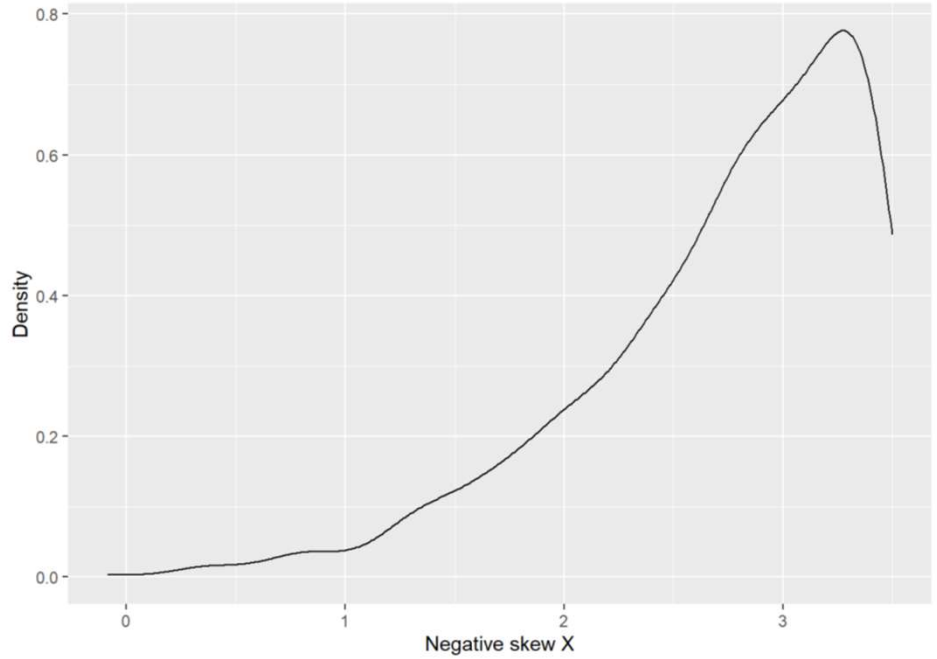




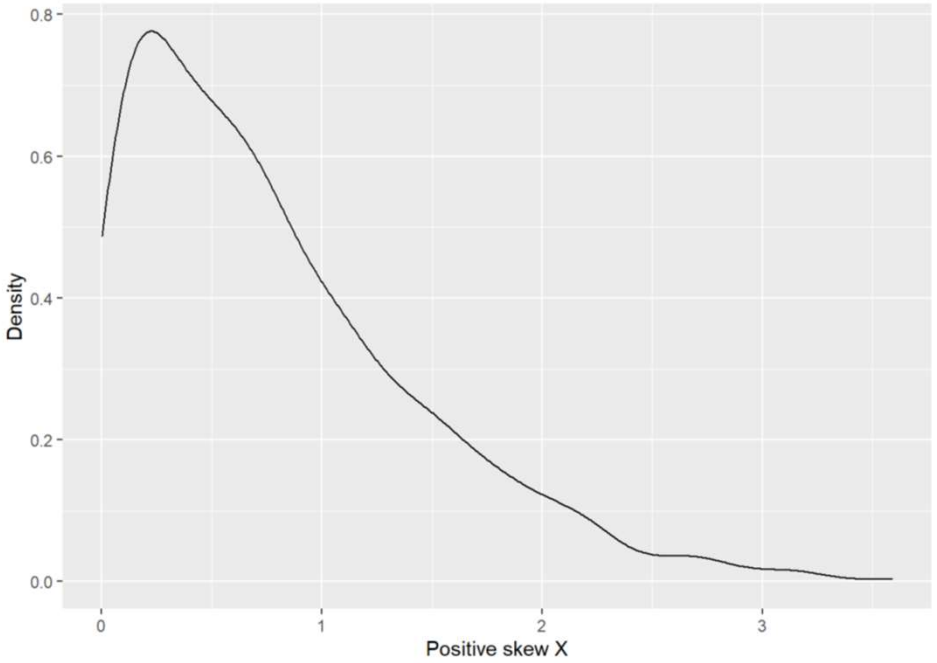
Histograms and density plots display the shape of the data distribution.

Skewness

Negatively skewed data occurs when there is a large **left tail** consisting of a relatively small number of relatively low values, but most of the data is towards the upper end of the plot.

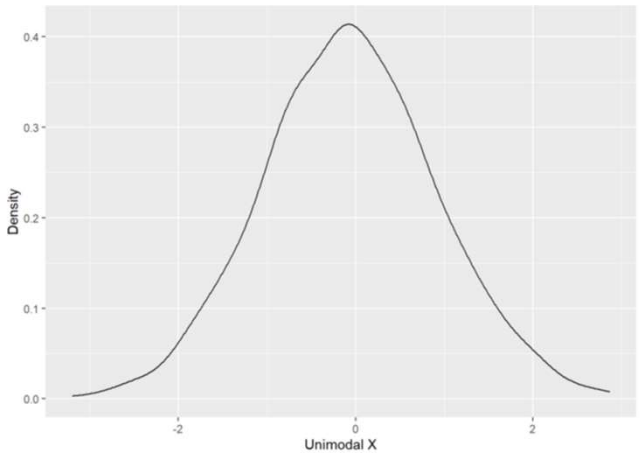


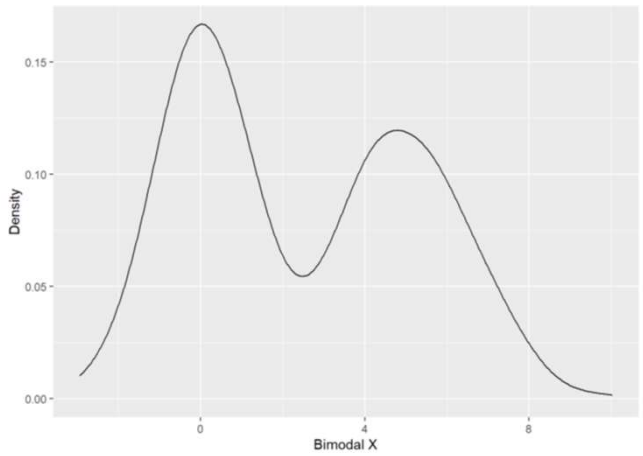
Positively skewed data occurs when there is a large **right tail** consisting of a relatively small number of relatively high values, but most of the data is towards the lower end of the plot.

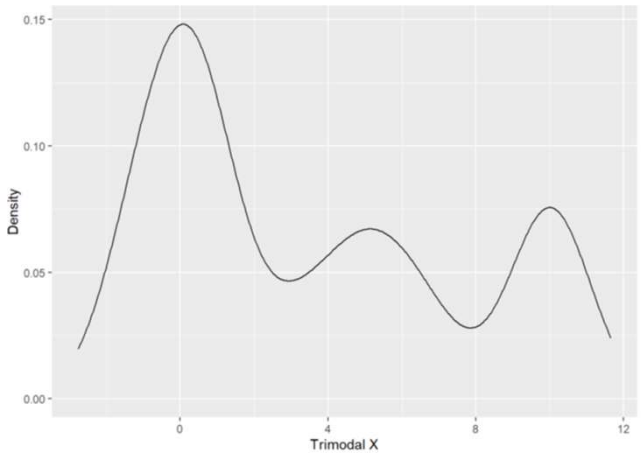


Unimodal vs. multi-modal

The number of **modes** refers to the number of peaks within the data.

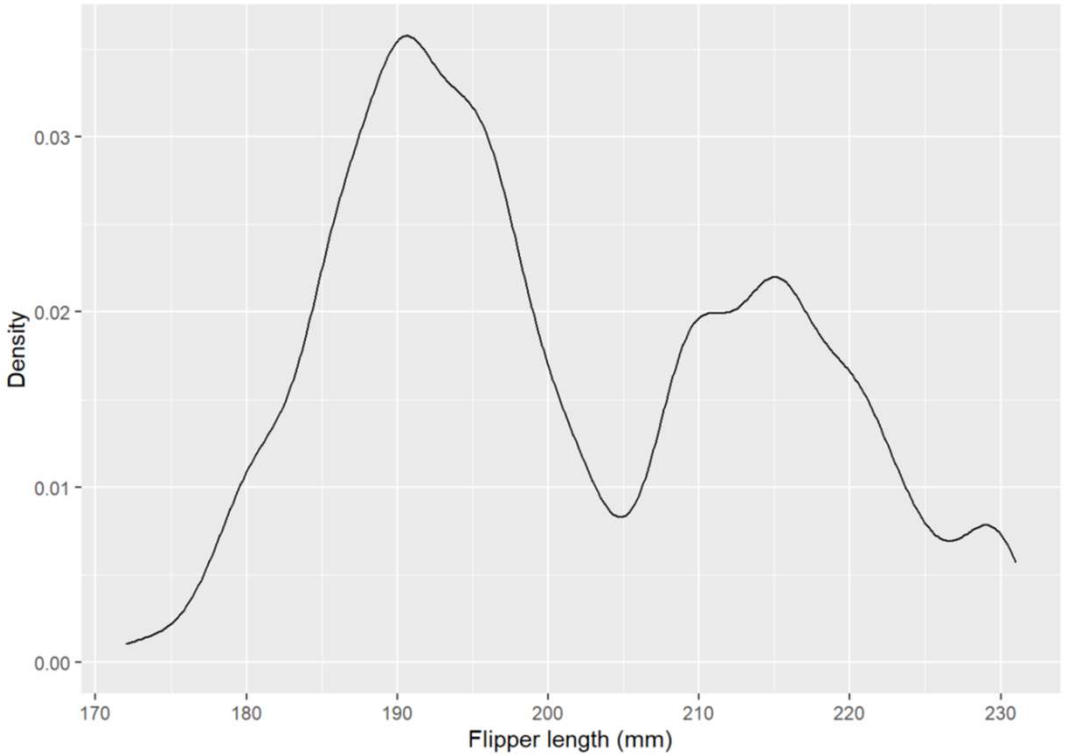






Univariate plots

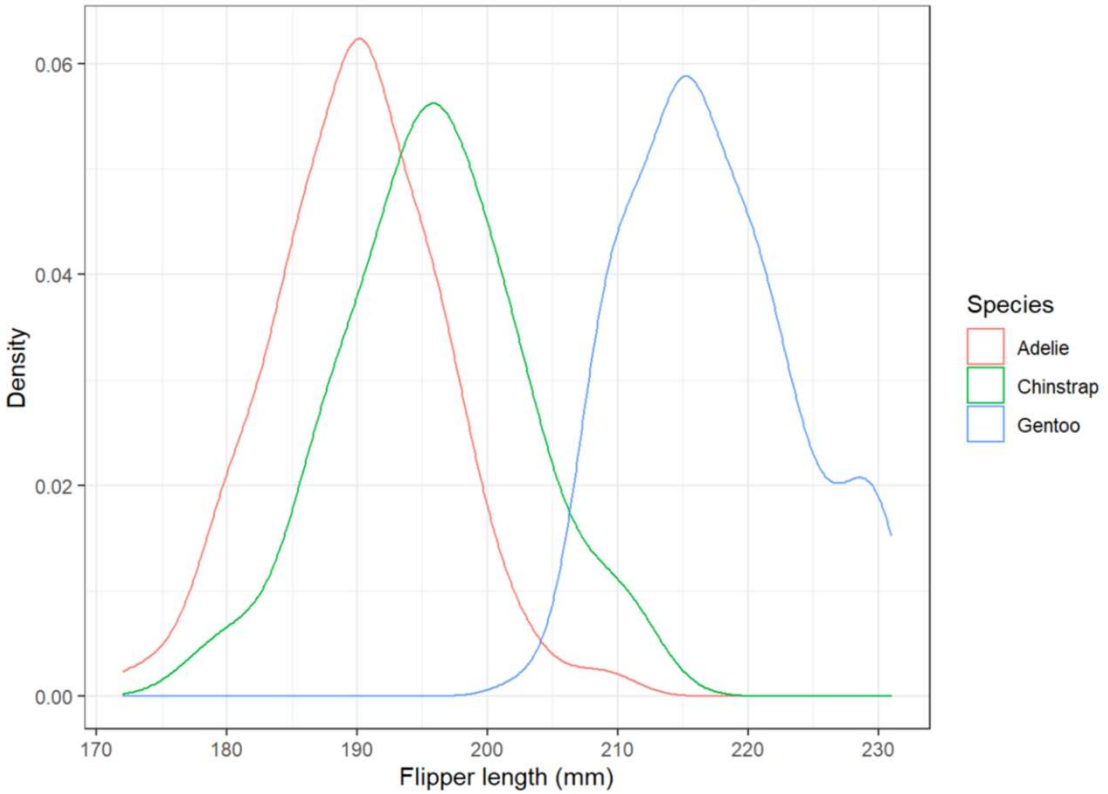
```
univar_plot+geom_density(adjust=0)+ylab('Density')
```



A bimodal distribution.

Bivariate plots

```
ggplot(data=rename(penguins, Species=species), aes(x=flipper_length_mm, color=Species))+
  geom_density()+theme_bw()+xlab("Flipper length (mm)")+ylab("Density")
```



Aesthetics

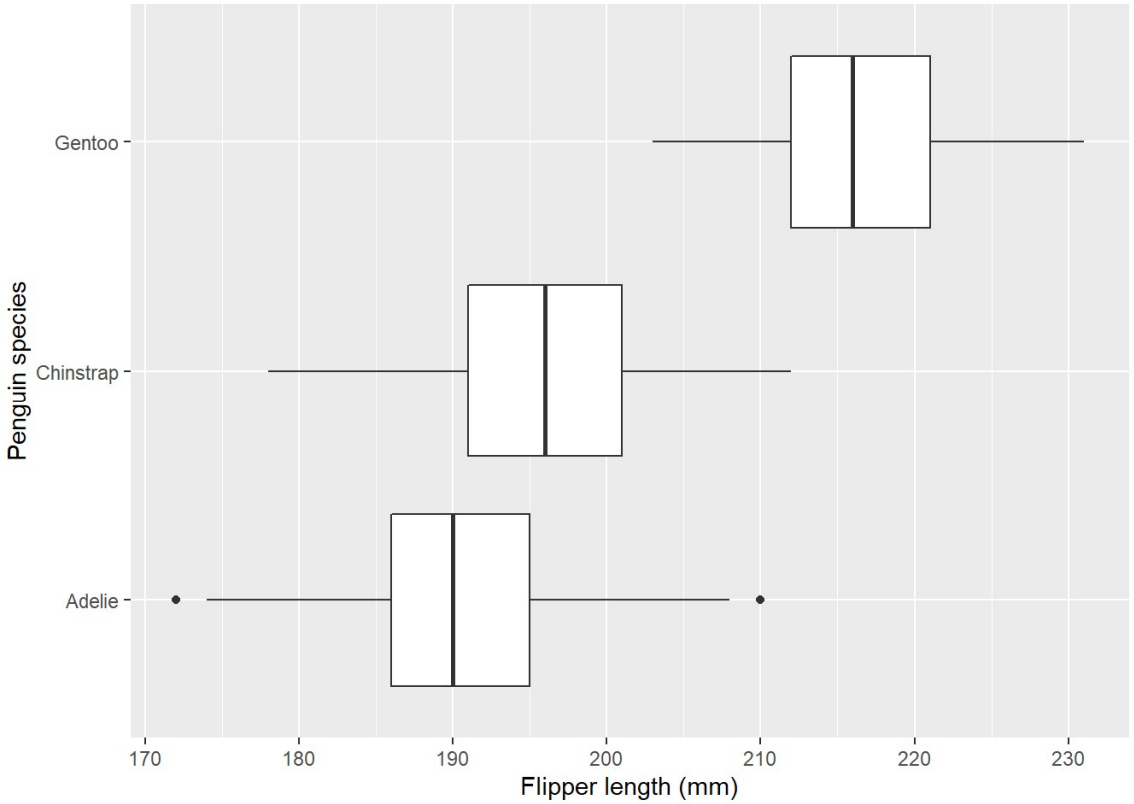
Mappings between a variable and a visual cue.

Flipper length → horizontal position.

Species → colour

Bivariate plots

```
ggplot(data=penguins, aes(x=flipper_length_mm, y=species))+geom_boxplot()+
  xlab('Flipper length (mm)') + ylab("Penguin species")
```



Aesthetics

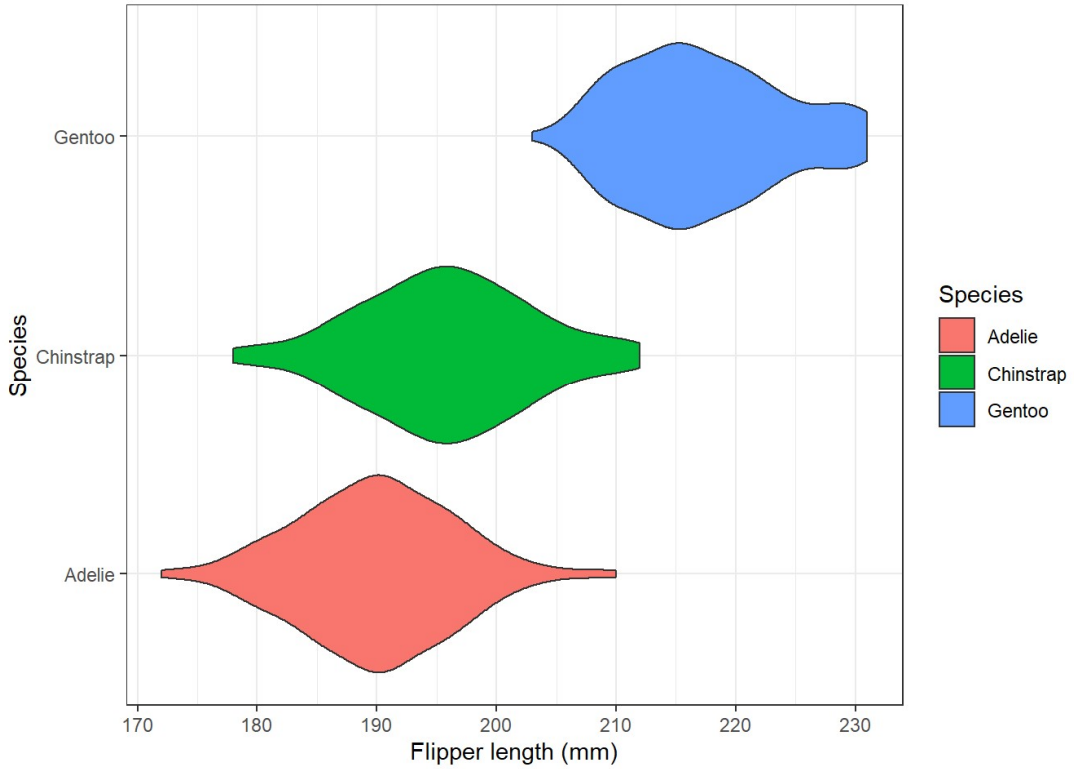
Mappings between a variable and a visual cue.

Flipper length → horizontal position.

Species → vertical position.

Bivariate plots

```
ggplot(data=rename(penguins, Species=species), aes(x=flipper_length_mm, y=Species, fill=Species))+geom_violin()+theme_bw()+x
lab("Flipper length (mm)")
```



Aesthetics

Mappings between a variable and a visual cue.

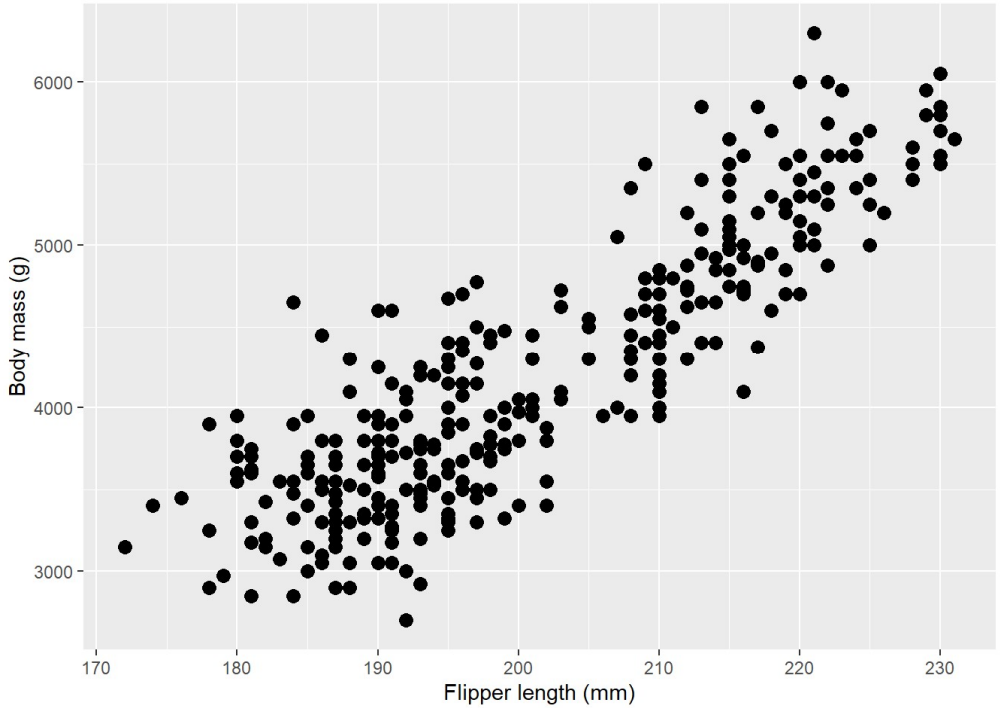
Flipper length → horizontal position.

Species → vertical position.

Species → colour

Bivariate plots

```
mass_flipper_scatter <- ggplot(data=penguins, aes(y=body_mass_g, x=flipper_length_mm))+
  xlab("Flipper length (mm)") + ylab("Body mass (g)")
mass_flipper_scatter+geom_point(size=3)
```



Aesthetics

Flipper length → horizontal position.

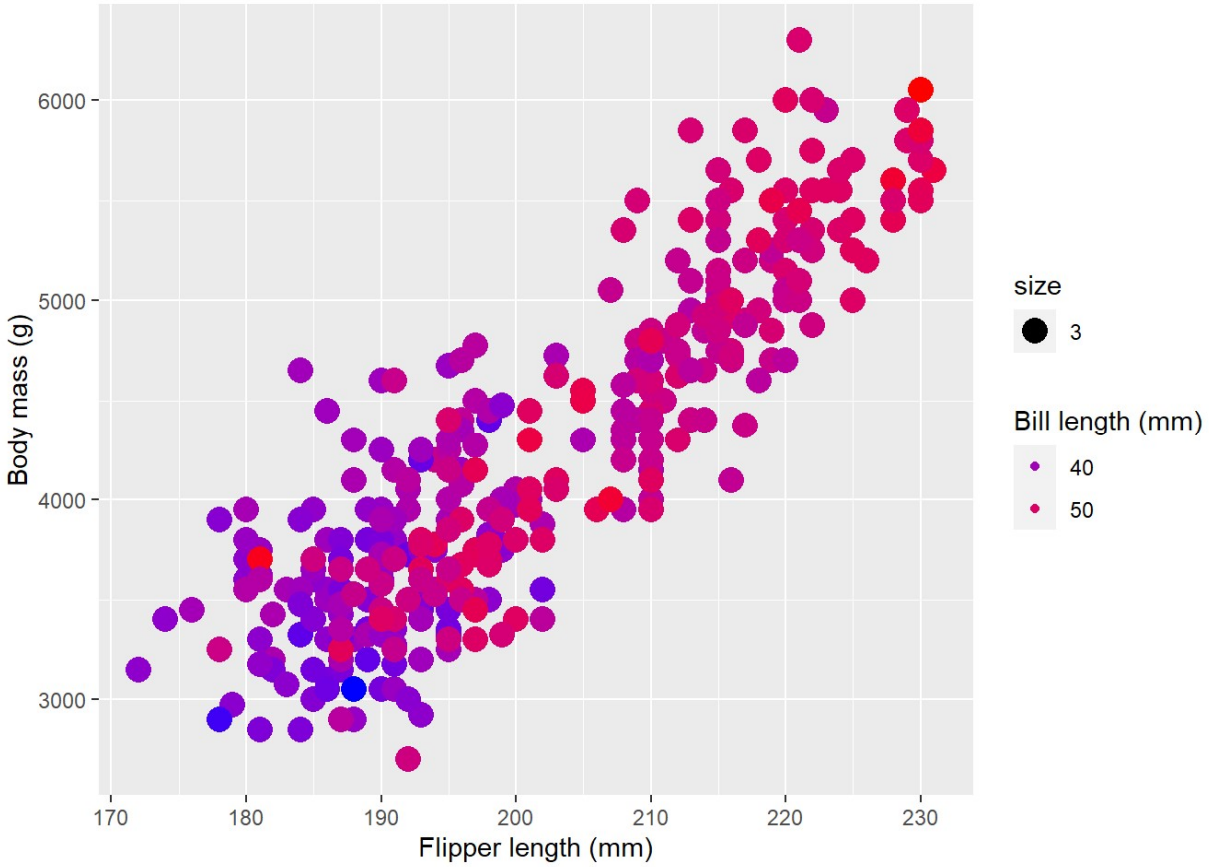
Body mass → vertical position.

Glyph

Points.

Multivariate plots

```
mass_flipper_scatter+geom_point(aes(color=bill_length_mm, size=3))+
  scale_color_gradient(low="blue", high="red")+guides(color=guide_legend("Bill length (mm)"))
```



Aesthetics

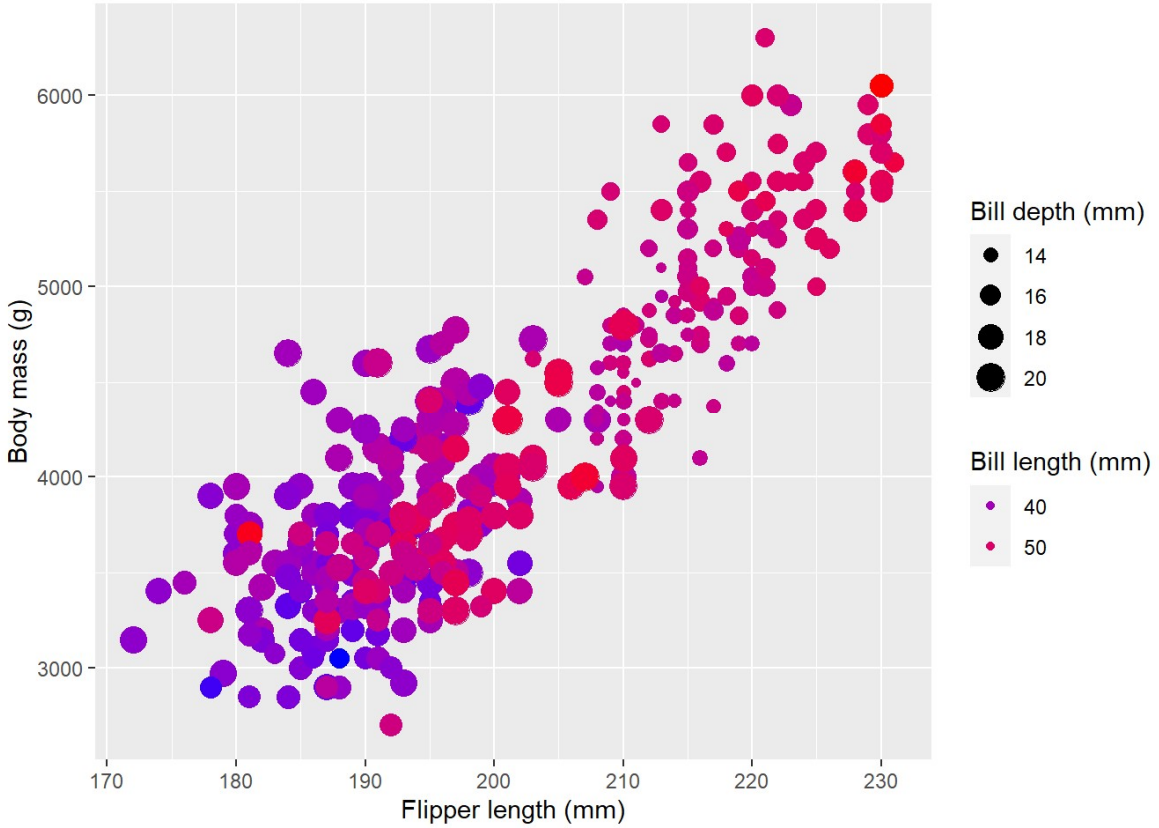
Flipper length → horizontal position.

Body mass → vertical position.

Bill length → colour

Multivariate plots

```
mass_flipper_scatter+geom_point(aes(color=bill_length_mm, size=bill_depth_mm))+
  scale_color_gradient(low="blue", high="red")+
  guides(color=guide_legend("Bill length (mm)"), size=guide_legend("Bill depth (mm)"))
```



Aesthetics

Flipper length → horizontal position.

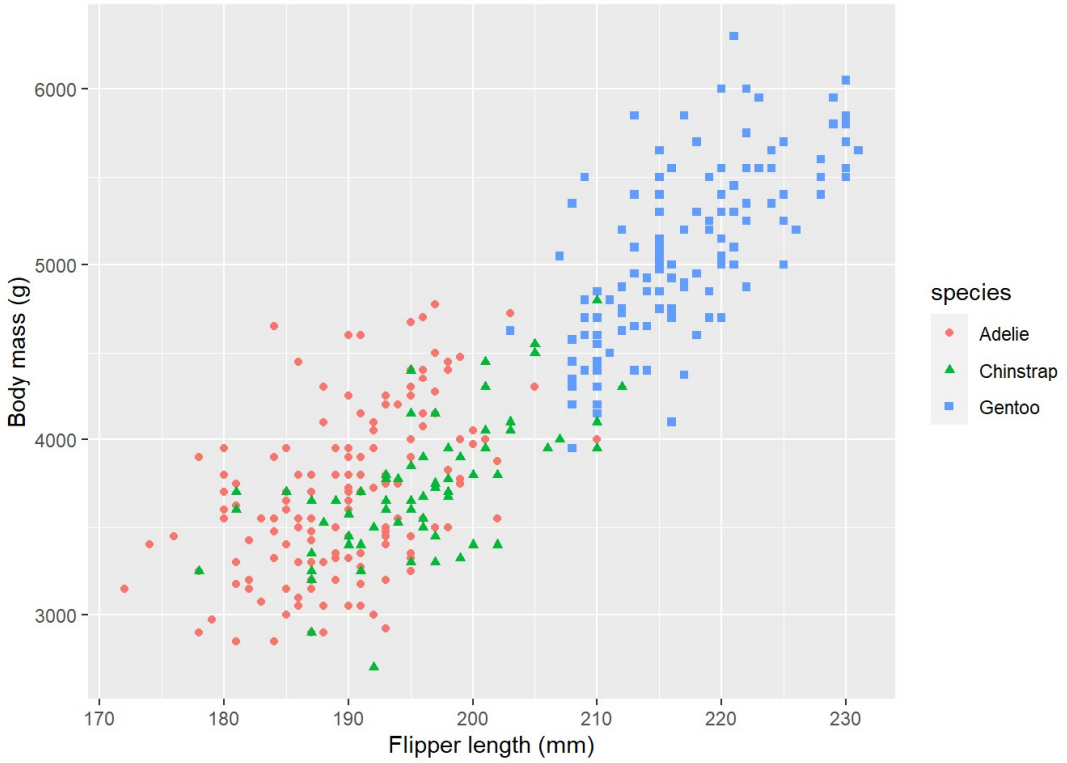
Body mass → vertical position.

Bill length → colour

Bill depth → size

Multivariate plots

```
mass_flipper_scatter+geom_point(aes(color=species, shape=species))
```



Aesthetics

Flipper length → horizontal position.

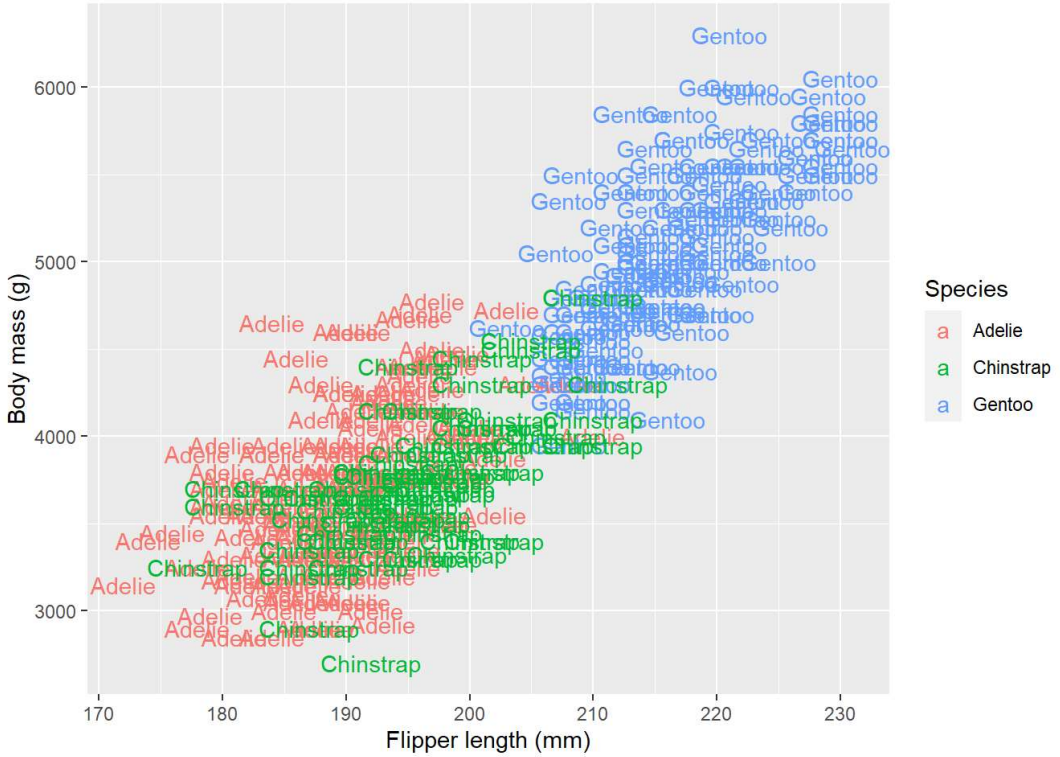
Body mass → vertical position.

Species → colour

Species → shape

Multivariate plots

```
mass_flipper_scatter + geom_text(aes(label=species, color=species)) +
  guides(color=guide_legend("Species"))
```



Aesthetics

Flipper length → horizontal position.

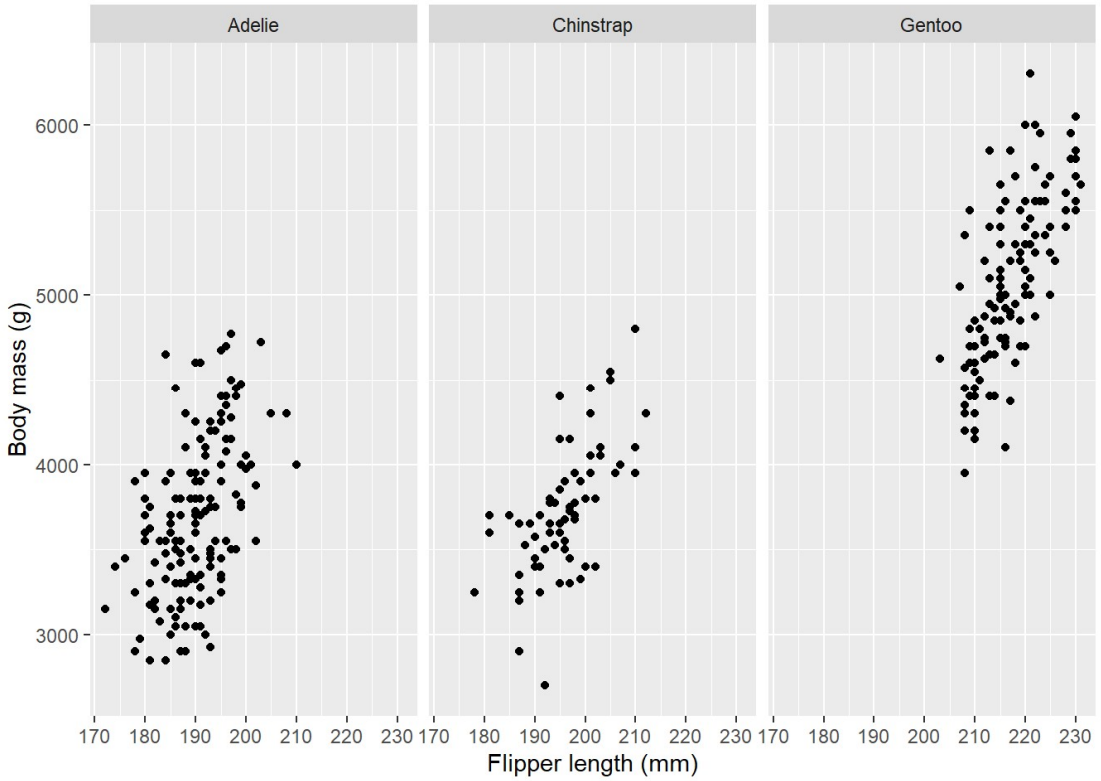
Body mass → vertical position.

Species → colour

Species → text

Facets

```
mass_flipper_scatter + geom_point() + facet_wrap(~species)
```



Aesthetics

Flipper length → horizontal position.

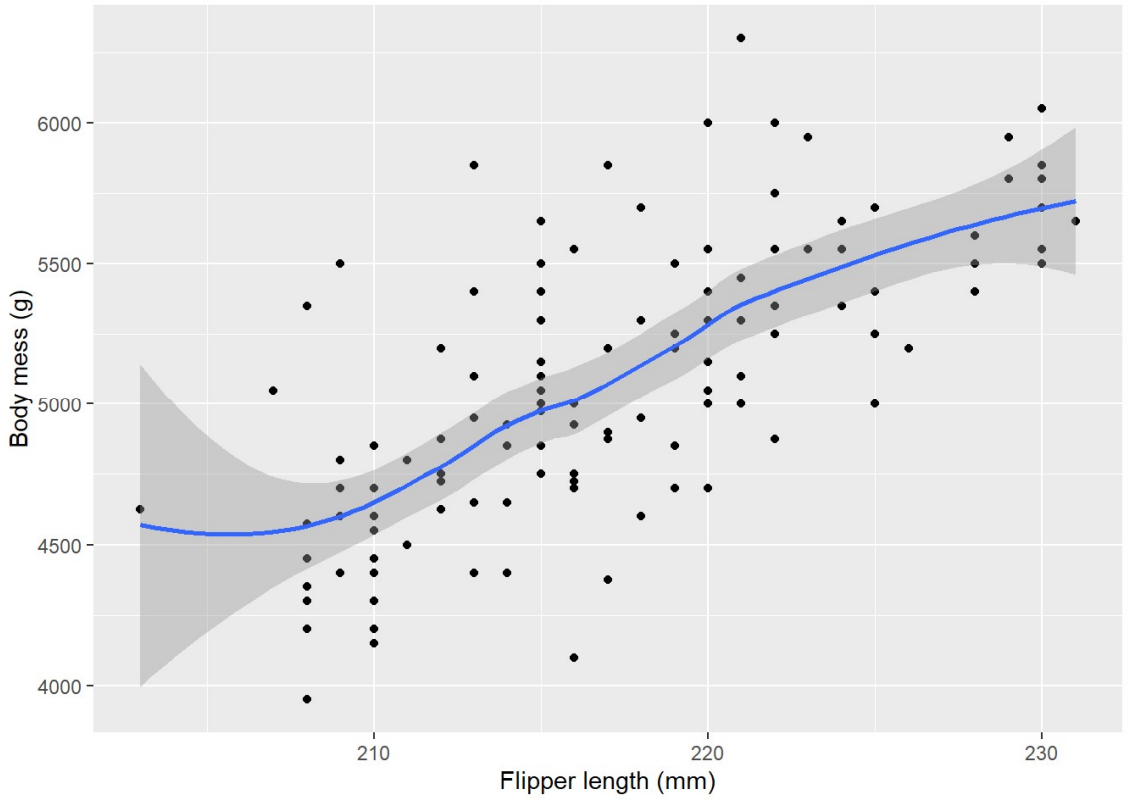
Body mass → vertical position

Facets

Species

Trend lines

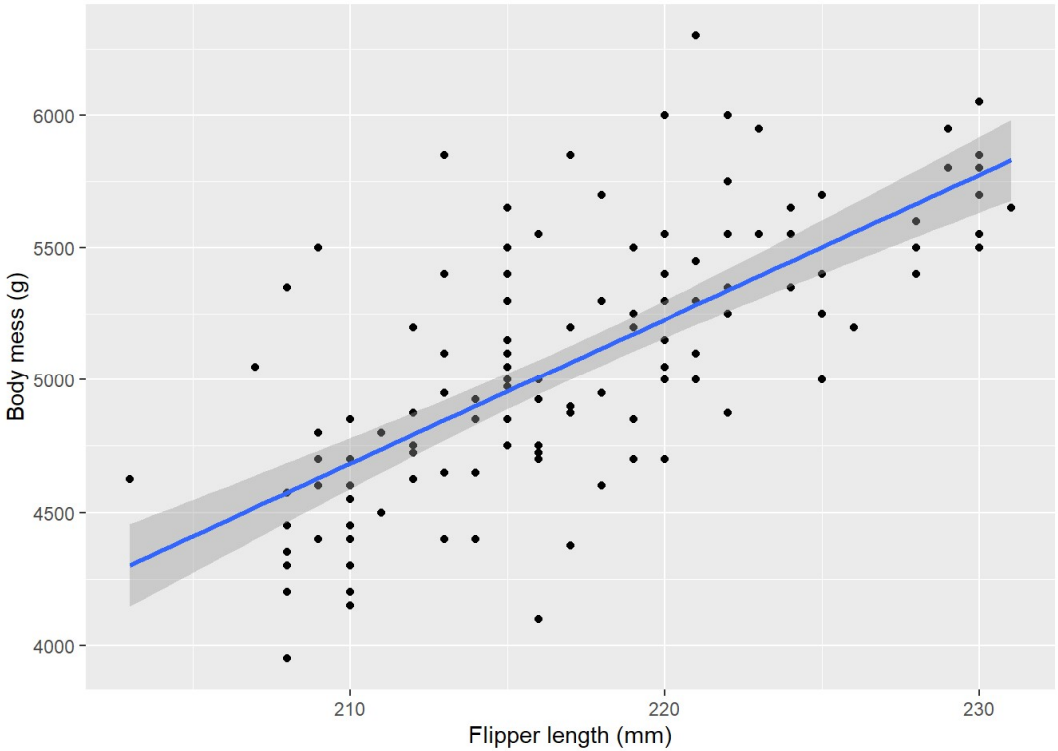
```
trend_plot <- ggplot(data=filter(penguins, species=='Gentoo'), aes(y=body_mass_g, x=flipper_length_mm))  + xlab('Flipper len
gth (mm)') + ylab('Body mess (g)') + geom_point()
trend_plot + geom_smooth()
```



Trend lines illustrate the relationship between two variables.

Trend lines

```
trend_plot+geom_smooth(method="lm")
```

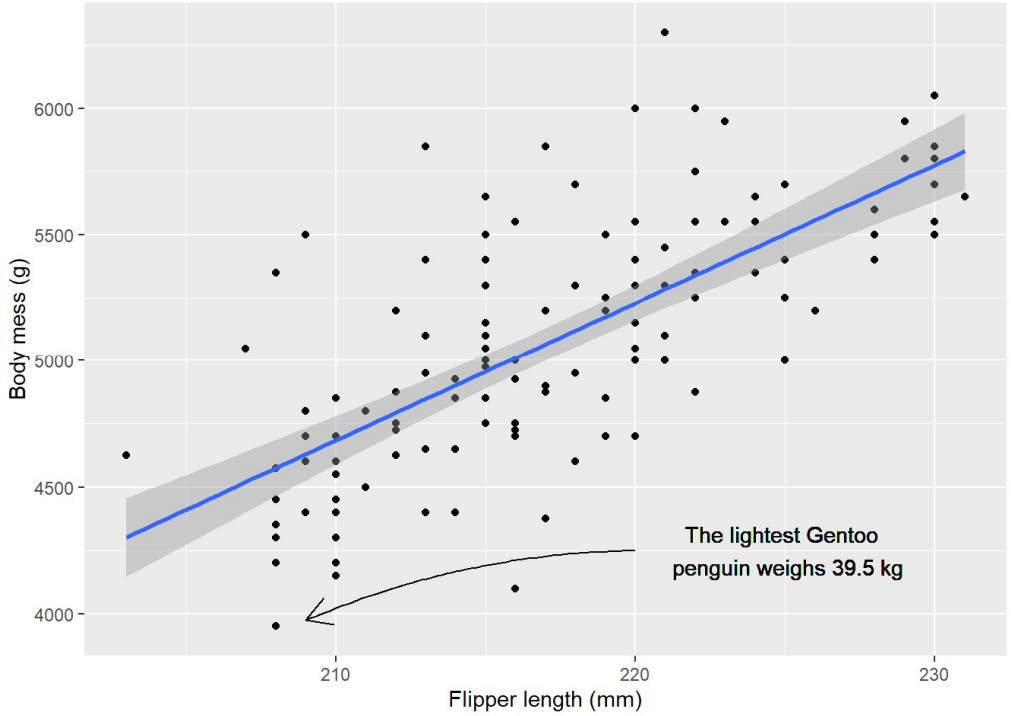


Annotation

```
min(filter(penguins, species=='Gentoo')$body_mass_g, na.rm=TRUE)
```

```
## [1] 3950
```

```
trend_plot + geom_smooth(method="lm") +
  geom_curve(x=220, xend=209, y=4250, yend=3975, arrow=arrow(length=unit(0.5, 'cm')), curvature=0.1) +
  geom_text(x=225, y=4250, label="The lightest Gentoo \n penguin weighs 39.5 kg")
```



GGplot2 gallery:
https://exts.ggplot2.tidyverse.org/gallery/

What have we covered?

We discussed the importance of visualisations for data science:

- To explore data
- To explain your insights to colleagues.

We have discussed the difference between various visual cues.

We have had a brief look at the power of the ggplot2 library within R.

Try the examples yourself?

The illustration, codes, and examples are included in the R Markdown file **LectureDataVisualisation.Rmd** which can be downloaded via the course webpage.

University of BRISTOL

Thanks for listening!

Dr. Rihuan Ke

rihuan.ke@bristol.ac.uk

Statistical Computing and Empirical Methods
Unit EMATM0061, MSc Data Science