3 - Data Exploration

Felipe Melo

Nottingham Trent University - UK

You should know today

- Make questions to your data?
- Explore the basic features of your data
- Make simple exploratory graphics

Before we begin

- R and Rstudio installed
- Don’t panic
- Everything is reproducible
- You’ll have to train to fix the content

What questions should I make to the data?

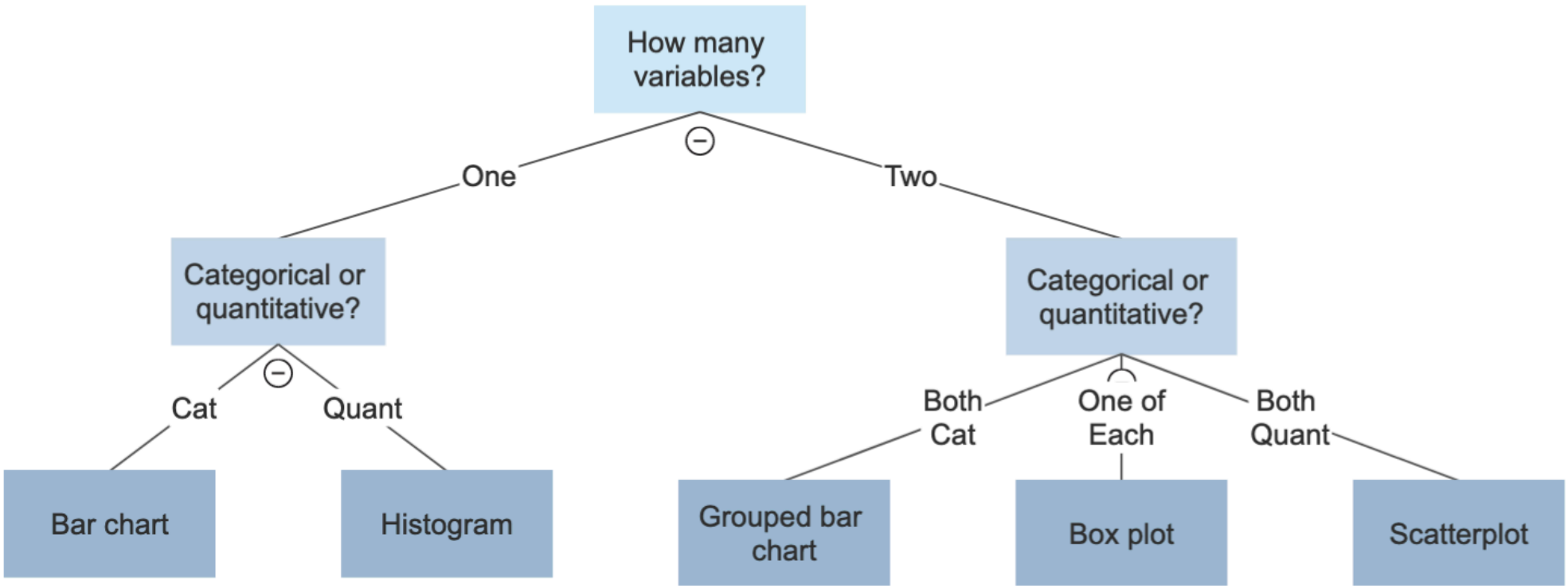
Back to Spreadsheets

The Penguins file

```
1 penguins_df<-read.csv("https://raw.githubusercontent.com/fplmelo/ecoaplic/r
2 penguins_df
```

	X	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm
1	1	Adelie	Torgersen	39.1	18.7	181
2	2	Adelie	Torgersen	39.5	17.4	186
3	3	Adelie	Torgersen	40.3	18.0	195
4	4	Adelie	Torgersen	NA	NA	NA
5	5	Adelie	Torgersen	36.7	19.3	193
6	6	Adelie	Torgersen	39.3	20.6	190
7	7	Adelie	Torgersen	38.9	17.8	181
8	8	Adelie	Torgersen	39.2	19.6	195
9	9	Adelie	Torgersen	34.1	18.1	193
10	10	Adelie	Torgersen	42.0	20.2	190
11	11	Adelie	Torgersen	37.8	17.1	186
12	12	Adelie	Torgersen	37.8	17.3	180
13	13	Adelie	Torgersen	41.1	17.6	182
14	14	Adelie	Torgersen	38.6	21.2	191
15	15	Adelie	Torgersen	34.6	21.1	198

Planning a data visualization



source: Andrew Gard

https://www.youtube.com/@EquitableEquations

We know this data

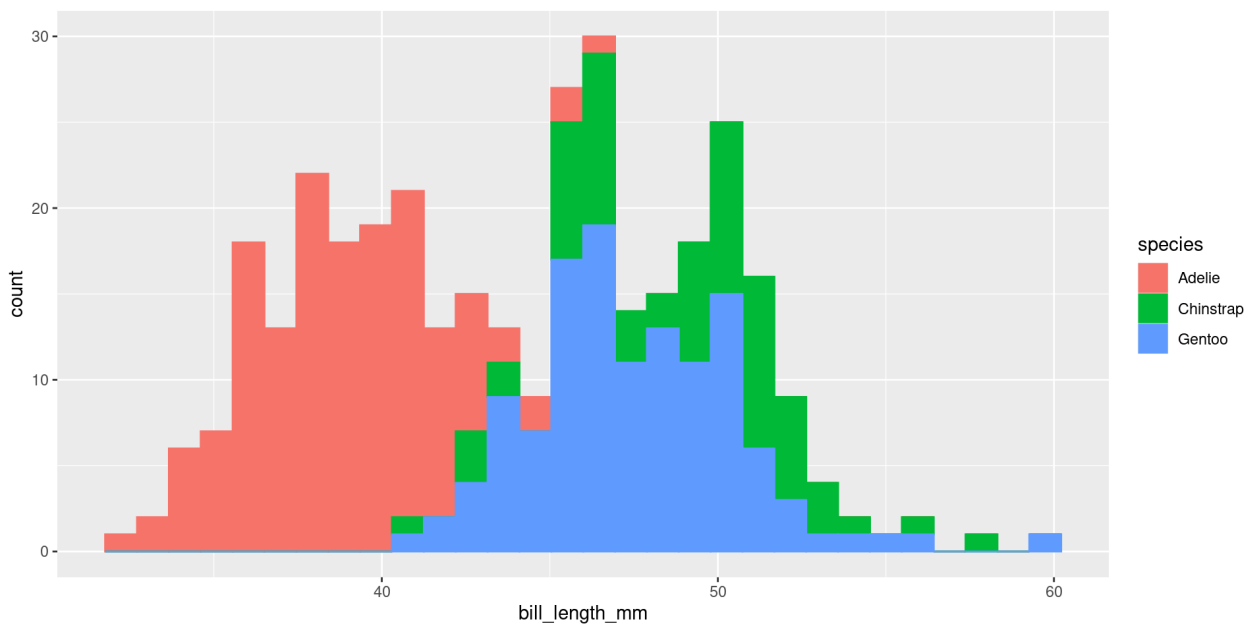
```r
library(tidyverse)
library(palmerpenguins)

data("penguins")
penguins %>%
  select(1:5)
```

```
# A tibble: 344 × 5
   species island    bill_length_mm bill_depth_mm flipper_length_mm
   <fct>   <fct>              <dbl>         <dbl>             <int>
 1 Adelie  Torgersen           39.1          18.7               181
 2 Adelie  Torgersen           39.5          17.4               186
 3 Adelie  Torgersen           40.3          18                 195
 4 Adelie  Torgersen           NA            NA                  NA
 5 Adelie  Torgersen           36.7          19.3               193
 6 Adelie  Torgersen           39.3          20.6               190
 7 Adelie  Torgersen           38.9          17.8               181
 8 Adelie  Torgersen           39.2          19.6               195
 9 Adelie  Torgersen           34.1          18.1               193
10 Adelie  Torgersen           42            20.2               190
# ℹ 334 more rows
```

How to visually check continuous variables?

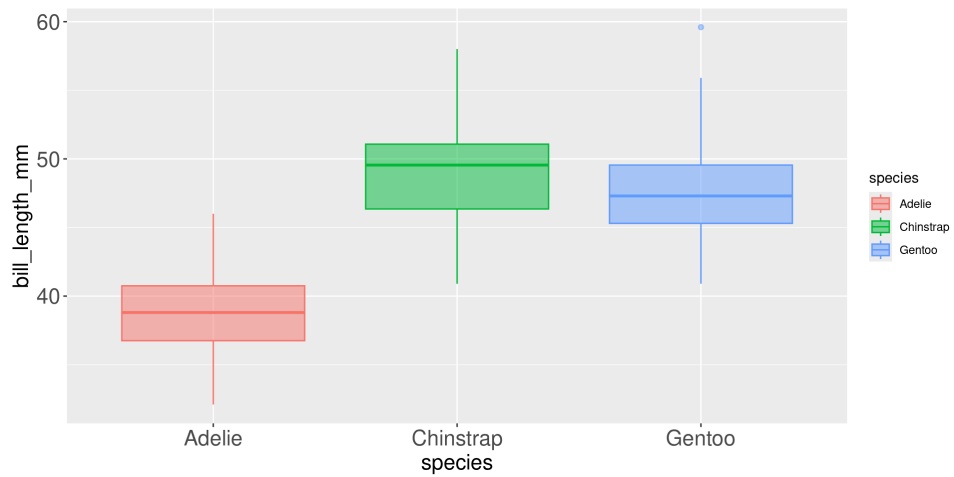
Histograms

```
1  library(tidyverse)
2  library(palmerpenguins)
3
4  data("penguins")
5  penguins %>%
6  group_by(species) %>%
7    ggplot(aes(x=bill_length_mm, color=species, fill=species))+
8    geom_histogram()
```



Boxplots

```
library(tidyverse)
library(palmerpenguins)

data("penguins")
penguins %>%
group_by(species) %>%
  ggplot(aes(x=species,
             y=bill_length_mm,
             color=species,
             fill=species))+
  geom_boxplot(alpha=0.5)+
  theme(axis.text=element_text(siz
        axis.title=element_text(si
```



Your turn

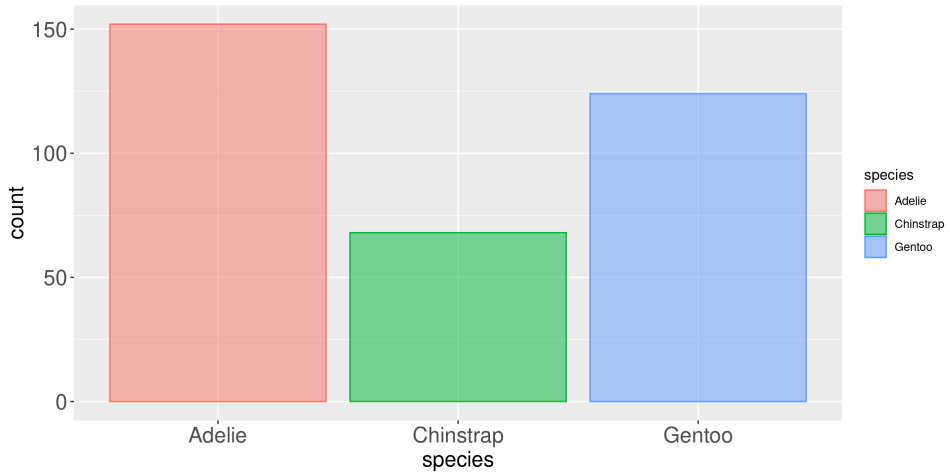
- Try to reproduce with any other continuous variable
- |Do a Historgram and a Boxplot

02:15

Checking categorical varibles

Species of penguin

```
 1 library(tidyverse)
 2 library(palmerpenguins)
 3
 4 penguins %>%
 5   ggplot(aes(x=species,
 6              color=species,
 7              fill=species))+
 8   geom_bar(alpha=0.5)+
 9   theme(axis.text=element_text(siz
10         axis.title=element_text(si
```

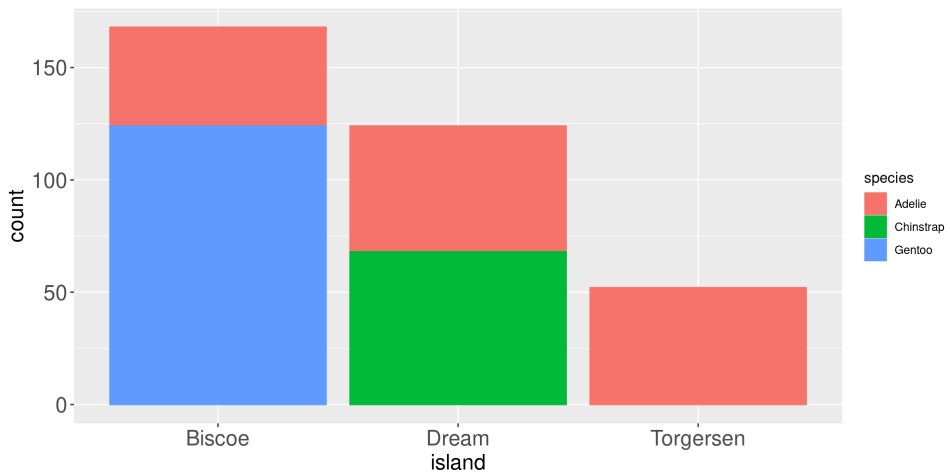


Observations per year

```
library(tidyverse)
library(palmerpenguins)

penguins %>%
  ggplot(aes(x=year,
             color=species,
             fill=species))+
  geom_bar()+
  theme(axis.text=element_text(siz
        axis.title=element_text(si
```

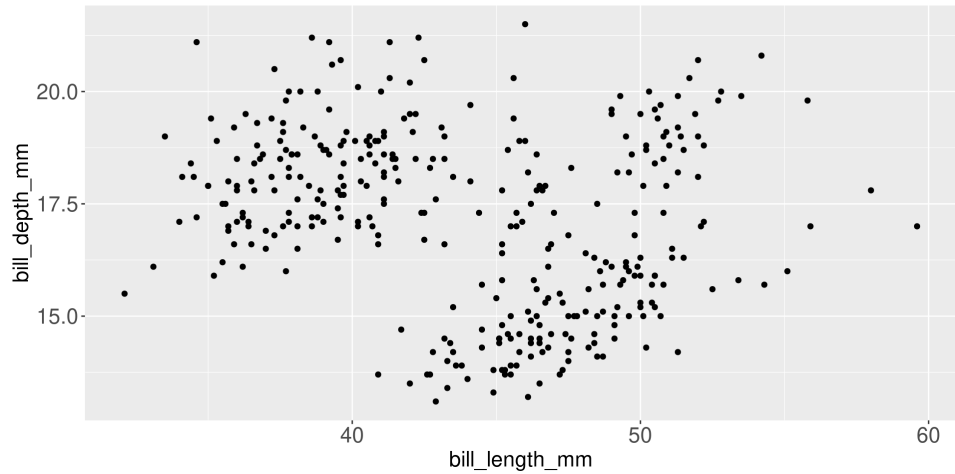
Observations per island

```
library(tidyverse)
library(palmerpenguins)

penguins %>%
  ggplot(aes(x=island,
             color=species,
             fill=species))+
  geom_bar()+
  theme(axis.text=element_text(siz
        axis.title=element_text(si
```



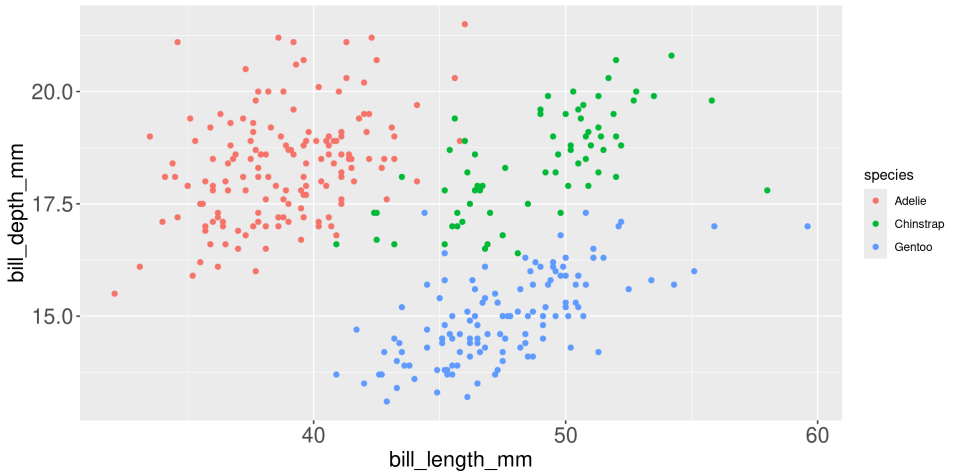
Visualising correlations

```
penguins %>%
  ggplot(aes(x=bill_length_mm,
             y = bill_depth_mm))+
  geom_point()+
  theme(axis.text=element_text(siz
        axis.title=element_text(si
```



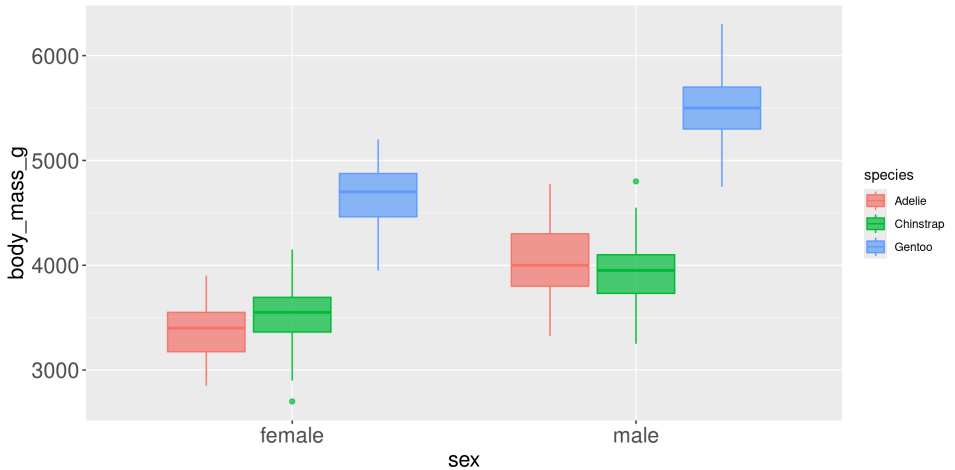
Visualising correlations per species

```
penguins %>%
  ggplot(aes(x=bill_length_mm,
             y = bill_depth_mm,
             color=species,
             fill=species))+
  geom_point()+
  theme(axis.text=element_text(siz
        axis.title=element_text(si
```



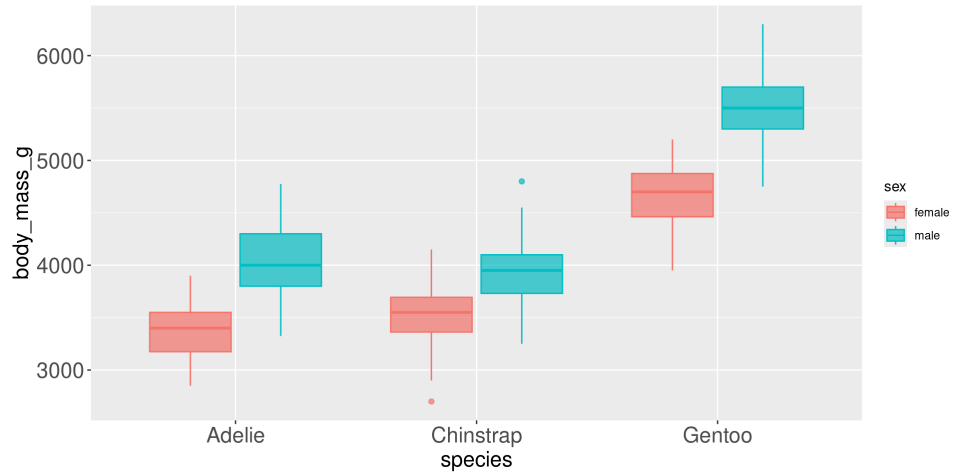
Body mass per sex

```
penguins %>%
  na.omit() %>%
  ggplot(aes(x=sex,
             y = body_mass_g,
             color=species,
             fill=species))+
  geom_boxplot(alpha=0.7)+
  theme(axis.text=element_text(siz
        axis.title=element_text(si
```



Body mass per sex (iverting groups)

```
penguins %>%
  na.omit() %>%
  ggplot(aes(x=species,
             y = body_mass_g,
             color=sex,
             fill=sex))+
  geom_boxplot(alpha=0.7)+
  theme(axis.text=element_text(siz
        axis.title=element_text(si
```



Your turn

- Can boby mass predict bill length?
- Do sex explain flipper length

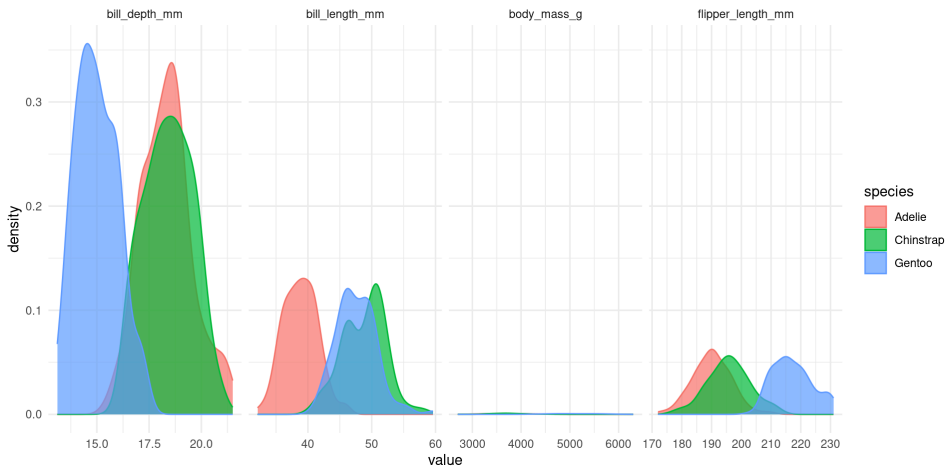
02:15

Exploring data is asking relevant questions

- This is not mining
- Don’t just correlate random things
- Start to imagine before coding

Check distributions

```
penguins %>%
  na.omit() %>%
  pivot_longer(bill_length_mm:body
  ggplot(aes(x=value,
         group=species,
         fill=species,
         color=species))+
  geom_density(alpha=0.7)+
  facet_grid(~trait, scales = "fre
  theme(axis.text=element_text(siz
        axis.title=element_text(si
  theme_minimal()
```



The importance of distributions

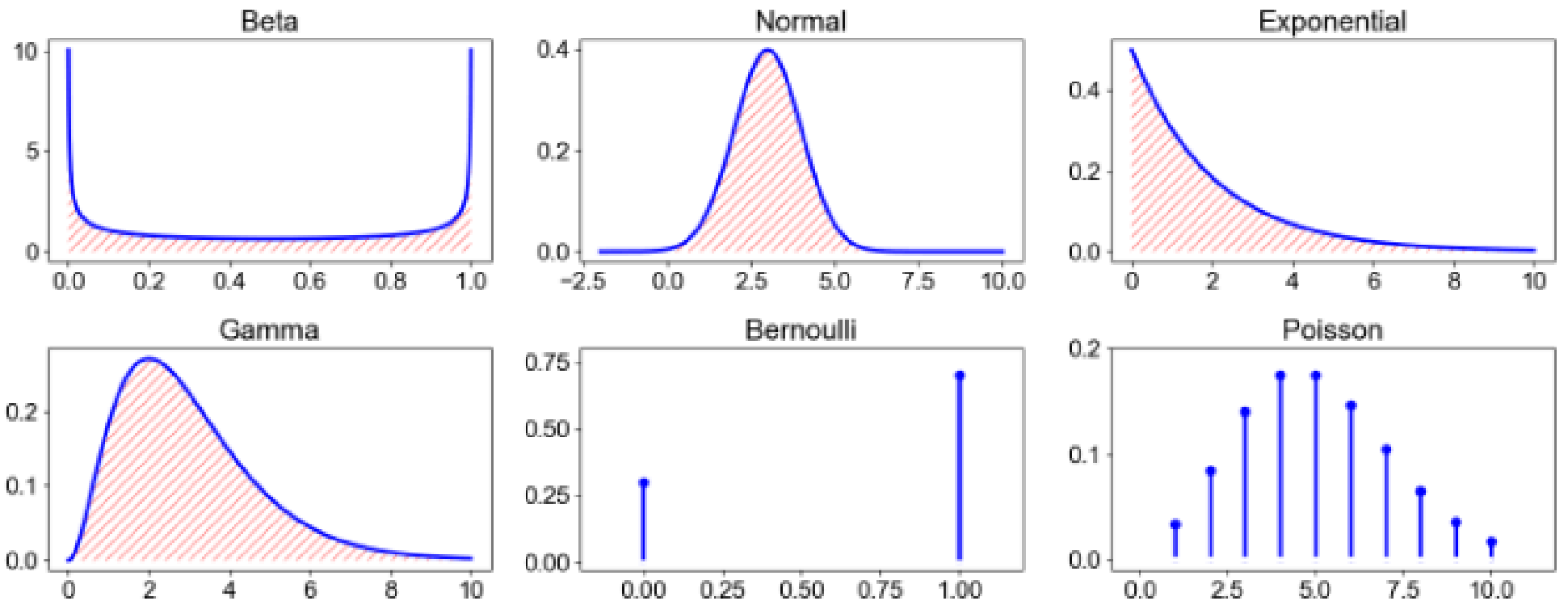
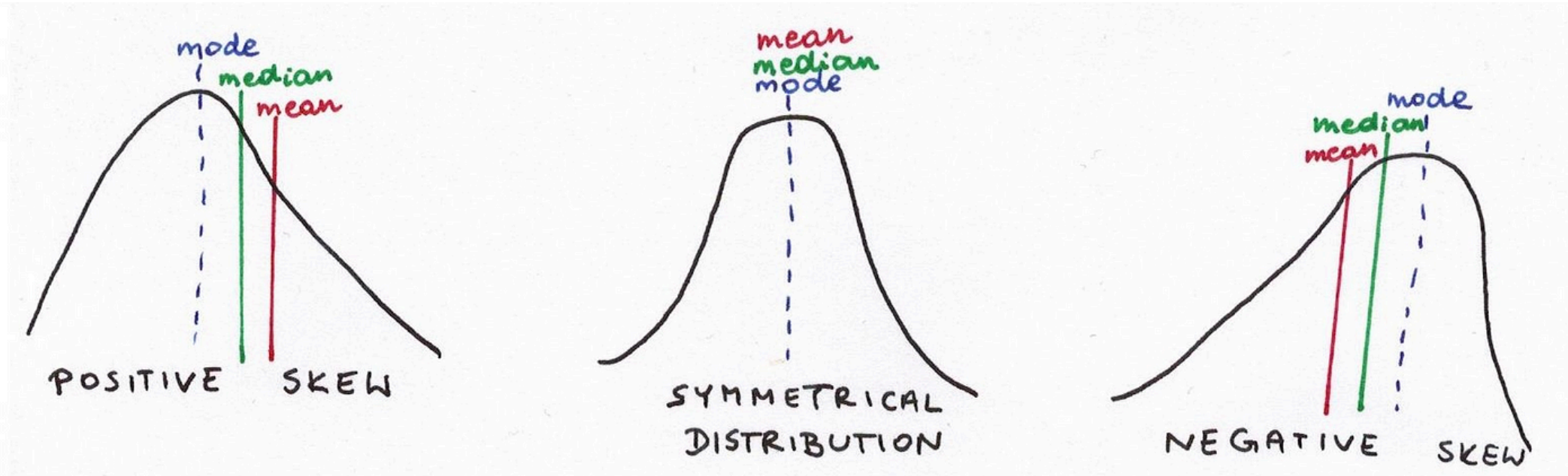


Figure 1. Beta, normal, exponential, gamma, Bernoulli, and Poisson distributions, each with a total mass of one.

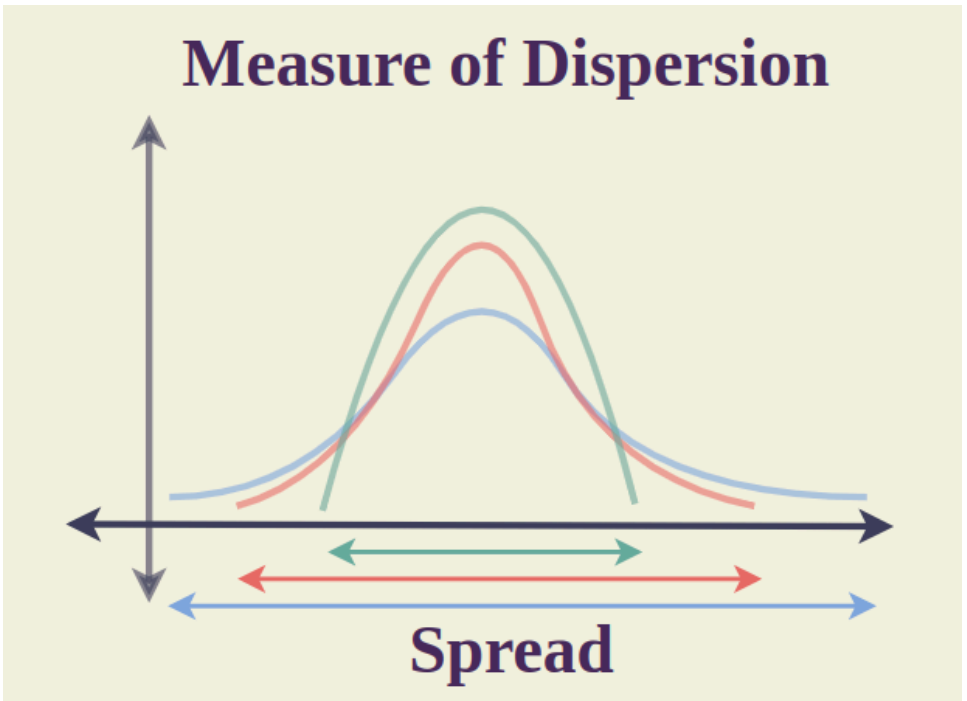
source:https://gregorygundersen.com/blog/2020/04/11/moments/

Moments of centrality



Mean, median and mode

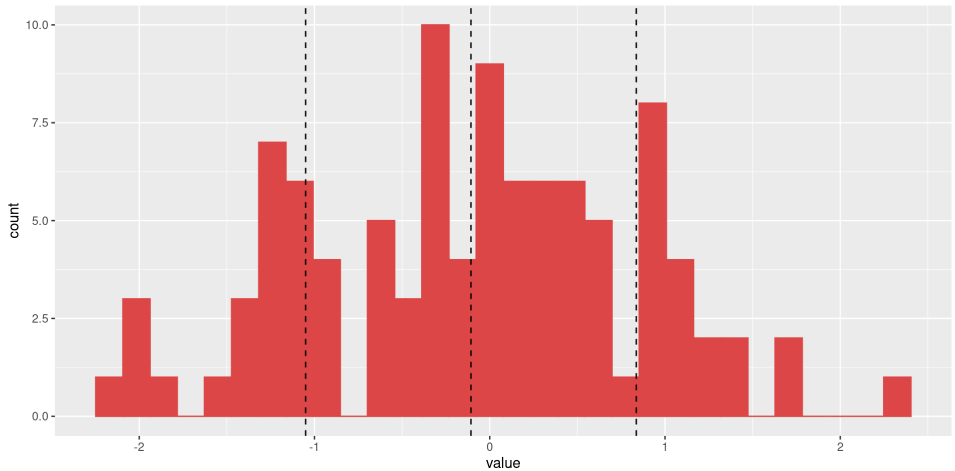
Moments of dispersion



- Variance
- Standard deviation
- Standard Error
- Range
- Quantiles

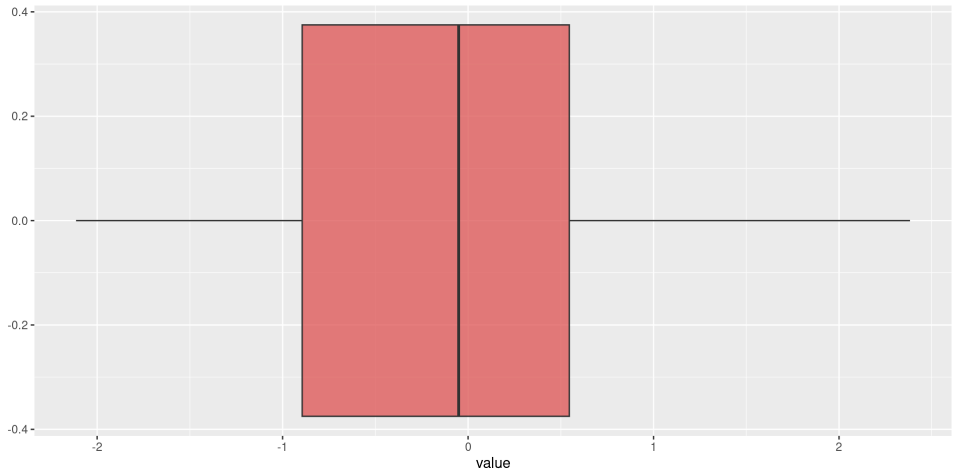
Checking via histogram

```
1  set.seed(999)
2  normal<-rnorm(100)
3  normal %>%
4    as.tibble() %>%
5    ggplot(aes(value))+
6    geom_histogram(color="#DD4A48",
7    geom_vline(xintercept=c(mean(nor
8                 linetype="dashed")
```

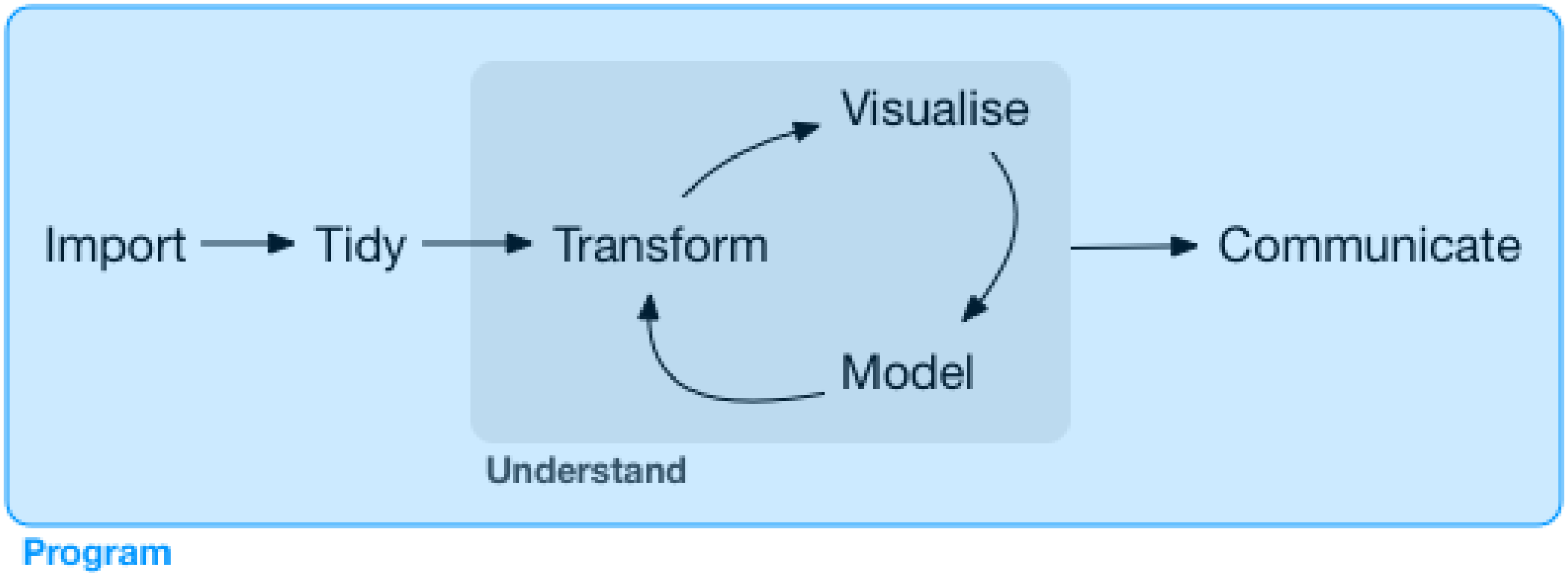


Checking via bokplot

```
set.seed(999)
normal<-rnorm(100)
normal %>%
  as.tibble() %>%
  ggplot(aes(value))+
  geom_boxplot(fill="#DD4A48",alph
```



Workflow


Import
Tidy
Transform
Visualise
Model
Understand
Communicate
Program

End of session on DA