Exploratory Data Analysis and visualisation (cont-d)

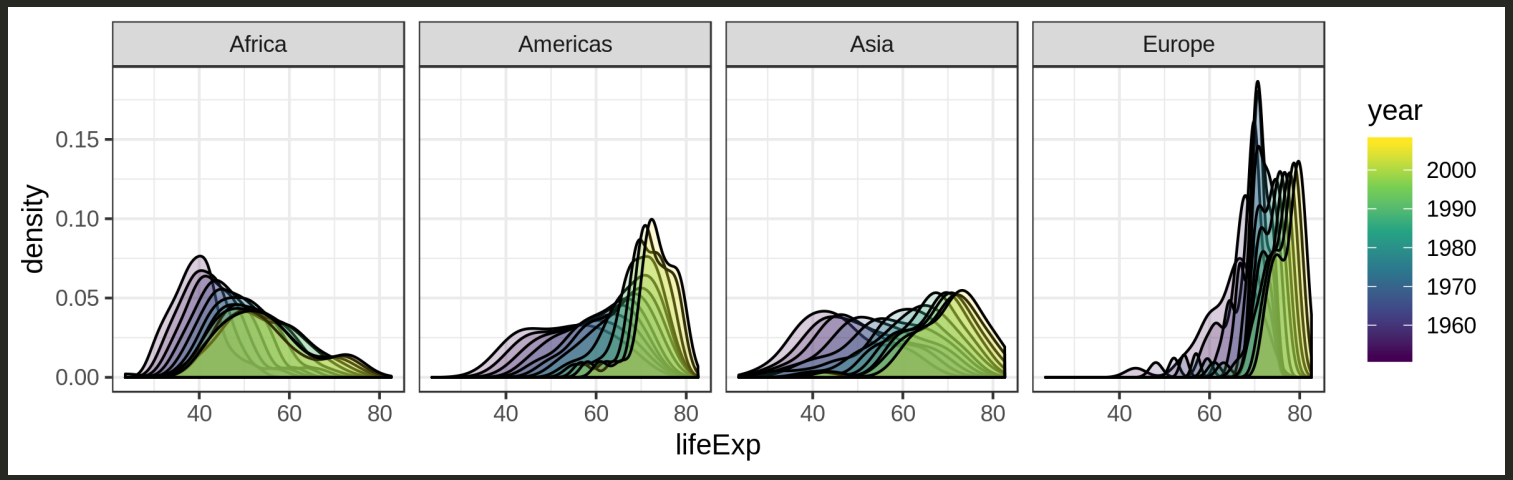
Kaspar Märtens

4 December 2018

Last time:

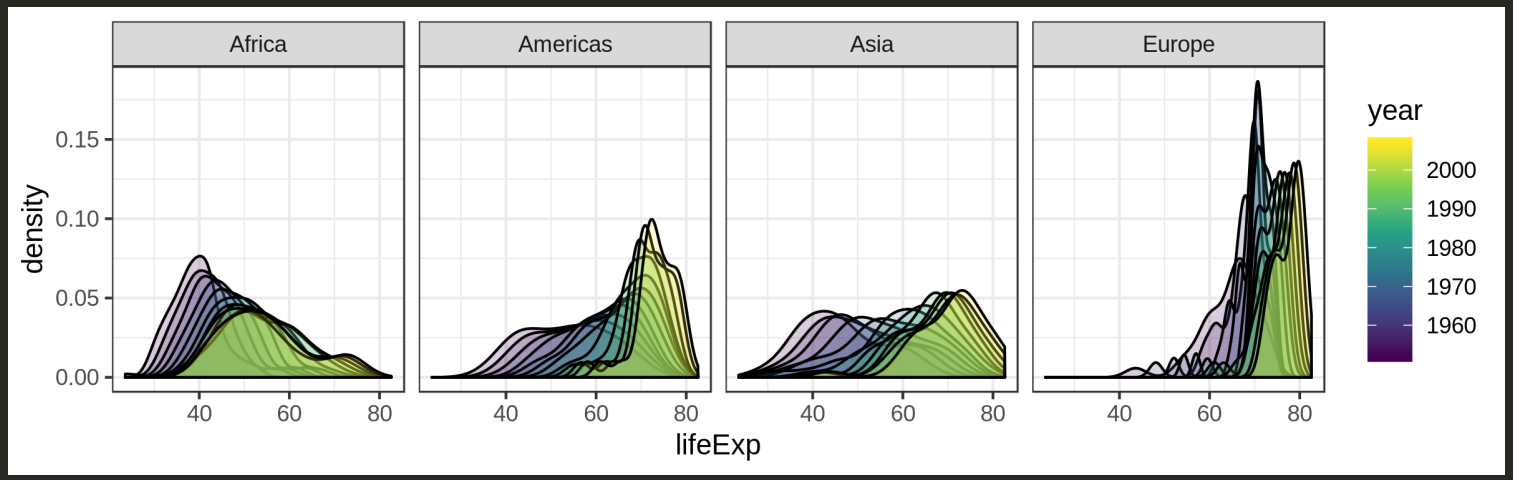
Last time:

Data visualisation and ggplot2 (with examples based on gapminder data)



Last time:

Data visualisation and ggplot2 (with examples based on gapminder data)



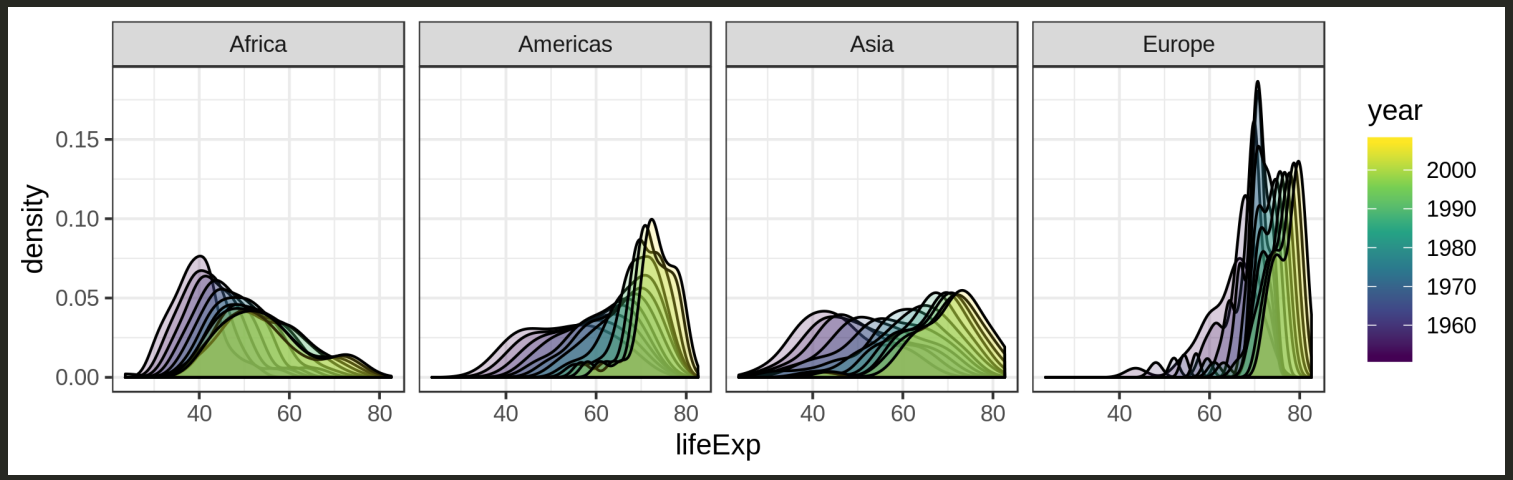
Today:

Exploratory data analysis (using tidyverse) in practice:

Let's try out some real data analysis on TCGA breast cancer data

Last time:

Data visualisation and ggplot2 (with examples based on gapminder data)



Today:

Exploratory data analysis (using tidyverse) in practice:

Let's try out some real data analysis on TCGA breast cancer data

If there is time, we might take a look at the gganimate package

TCGA breast cancer data set

- Phenotypes:
  - age at diagnosis
  - ER status (estrogen-receptor-positive or negative)
  - PAM50 cancer subtype
  - etc
- Gene expression data:
  - `log(x+1)` transformed expression for all genes

Find data and notebook in https://github.com/kasparmartens/2018_12_bham

```r
df_clinical <- readRDS("data/TCGA_clinical.rds")
head(df_clinical)
```

```
##             id stage PAM50 age_at_diagnosis ER_status PR_status pathology
## 1 TCGA-3C-AAAU                           55  Positive  Positive      <NA>
## 2 TCGA-3C-AALI                           50  Positive  Positive      <NA>
## 3 TCGA-3C-AALJ                           62  Positive  Positive      <NA>
## 4 TCGA-3C-AALK                           52  Positive  Positive      <NA>
## 5 TCGA-4H-AAAK                           50  Positive  Positive      <NA>
## 6 TCGA-5L-AAT0                           42  Positive  Positive      <NA>
```

```r
df_exprs <- readRDS("data/TCGA_exprs.rds")
head(df_exprs[, 1:5])
```

```
##             id SLC7A2_ENSG00000003989 HSPB6_ENSG00000004776
## 1 TCGA-3C-AAAU                8.605850                2.086391
## 2 TCGA-3C-AALI                3.017706                2.884330
## 3 TCGA-3C-AALJ                4.342539                3.086235
## 4 TCGA-3C-AALK                4.861960                2.754787
## 5 TCGA-4H-AAAK                4.298637                2.873187
## 6 TCGA-5L-AAT0                2.595548                2.472016
##   PDK4_ENSG00000004799 ZMYND10_ENSG00000004838
## 1             3.467477                3.891913
## 2             2.749256                2.385003
## 3             5.311276                2.766632
## 4             3.407671                4.407235
## 5             4.594217                4.634472
```

Differences between groups, focusing on a particular gene

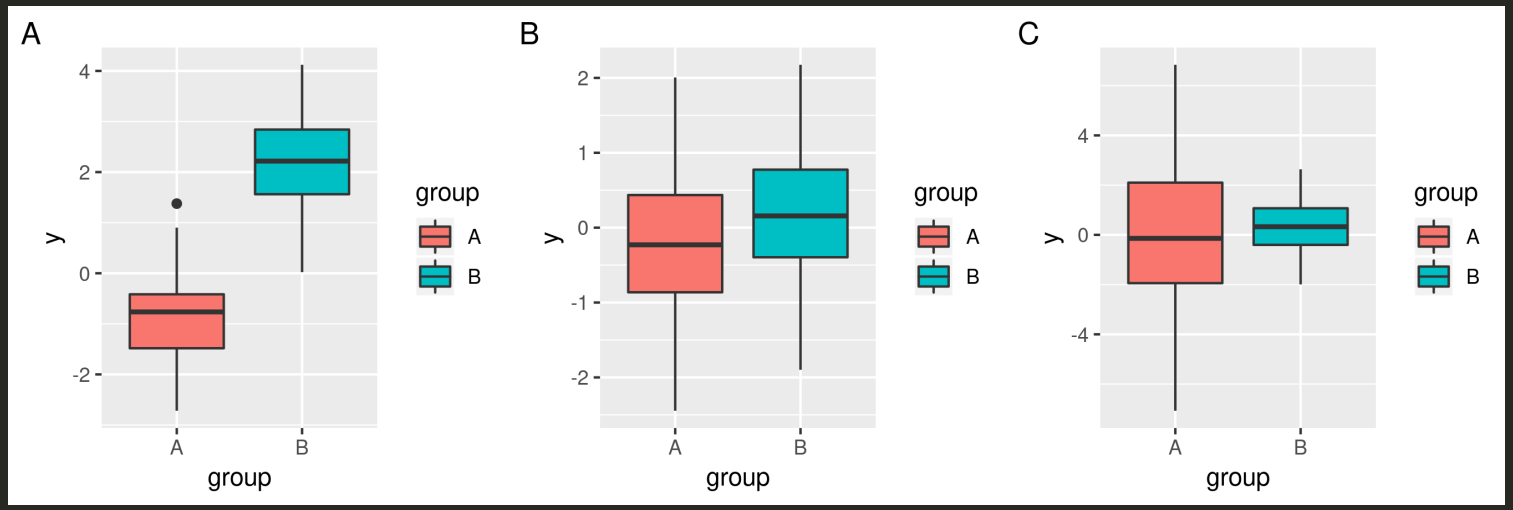
Q1: Do ER+ and ER- patients exhibit differences in the expression of a particular gene of interest (ESR1)?

Q2: Same question about PAM50 subtypes.

- How would you answer these questions
  - Using numerical quantification -- which model or test would you use?
  - Visually -- what type of plots would you consider?

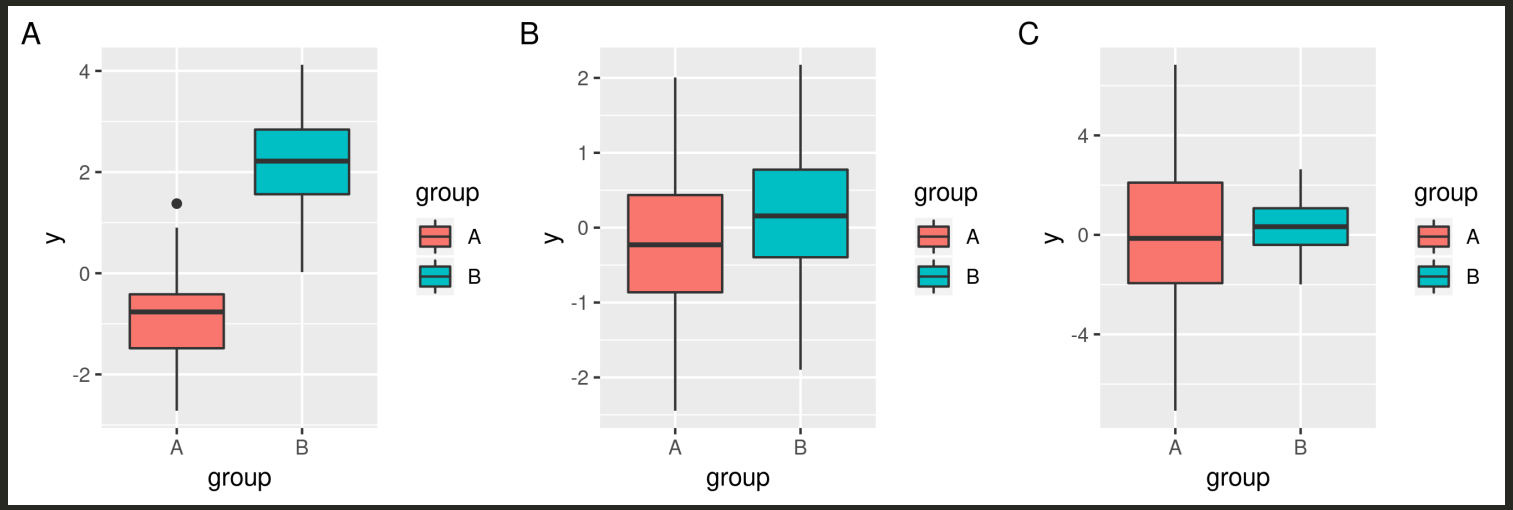
Boxplot quiz

Is there a significant difference between the groups?



Boxplot quiz

Is there a significant difference between the groups?



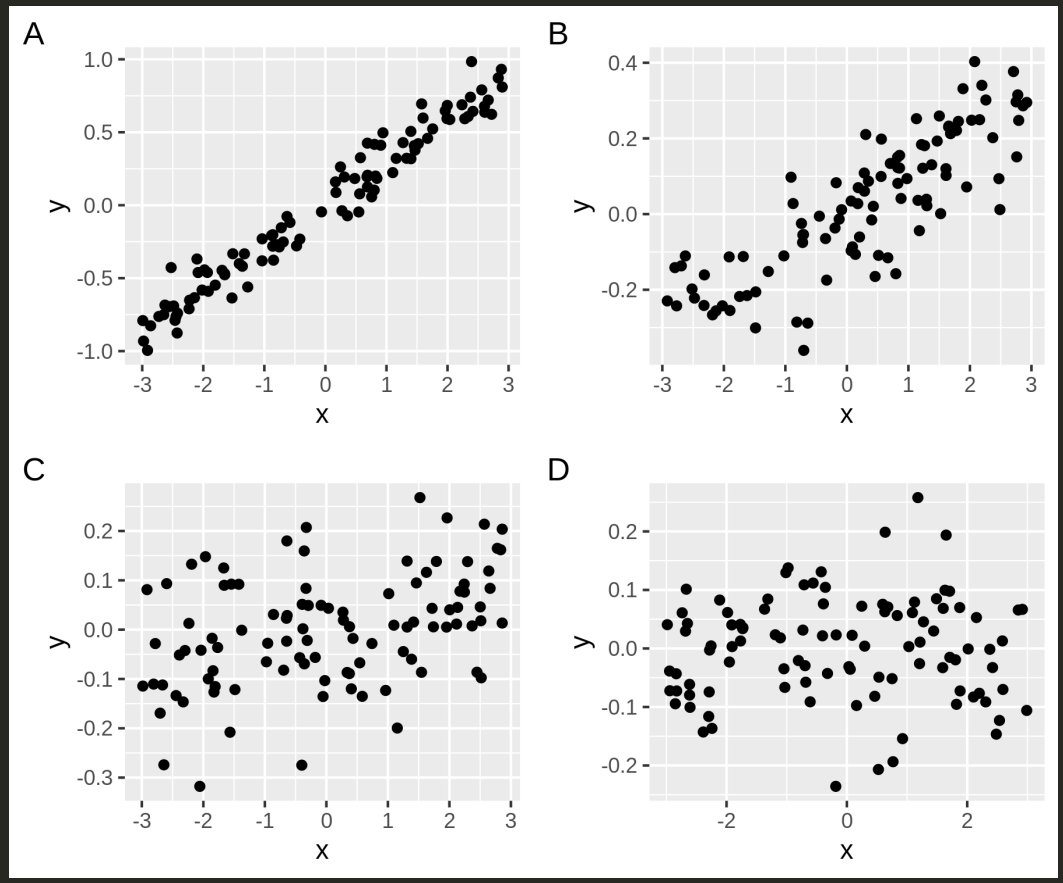
Do you think

- Yes, `p < 0.05` (i.e. group means are significantly different)
- Not enough information to decide
- No, `p > 0.05`

If you choose "cannot decide", what additional information would you need?

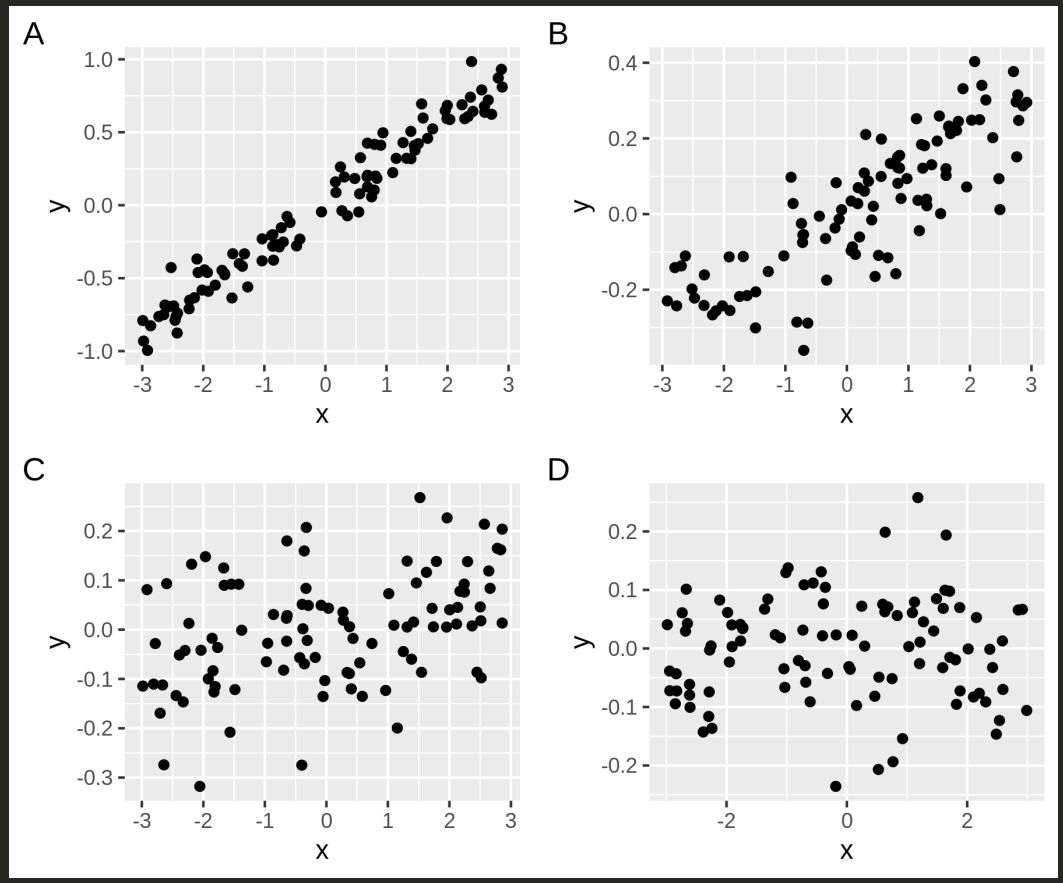
Correlation quiz

How strong is the correlation between the two variables? Is it significant?



Correlation quiz

How strong is the correlation between the two variables? Is it significant?



Also, see http://guessthecorrelation.com/

So far, we focused on a single gene

Now let's consider the expression of all genes

Before going further: a detour on tidy data (idea underlying the tidyverse)

Tidy data

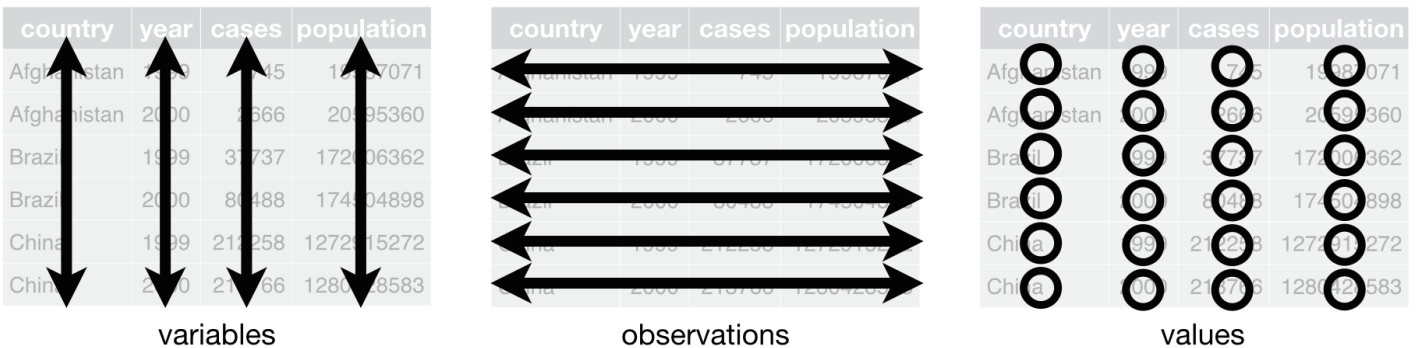
> “Tidy datasets are all alike, but every messy dataset is messy in its own way.” — Hadley Wickham

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.

Tidy data

> “Tidy datasets are all alike, but every messy dataset is messy in its own way.” — Hadley Wickham

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.



When unsure, we should ask ourselves: Is our data trapped in column names? Column headers should be variable names (but not values).

Tidy vs untidy data

Wide-format

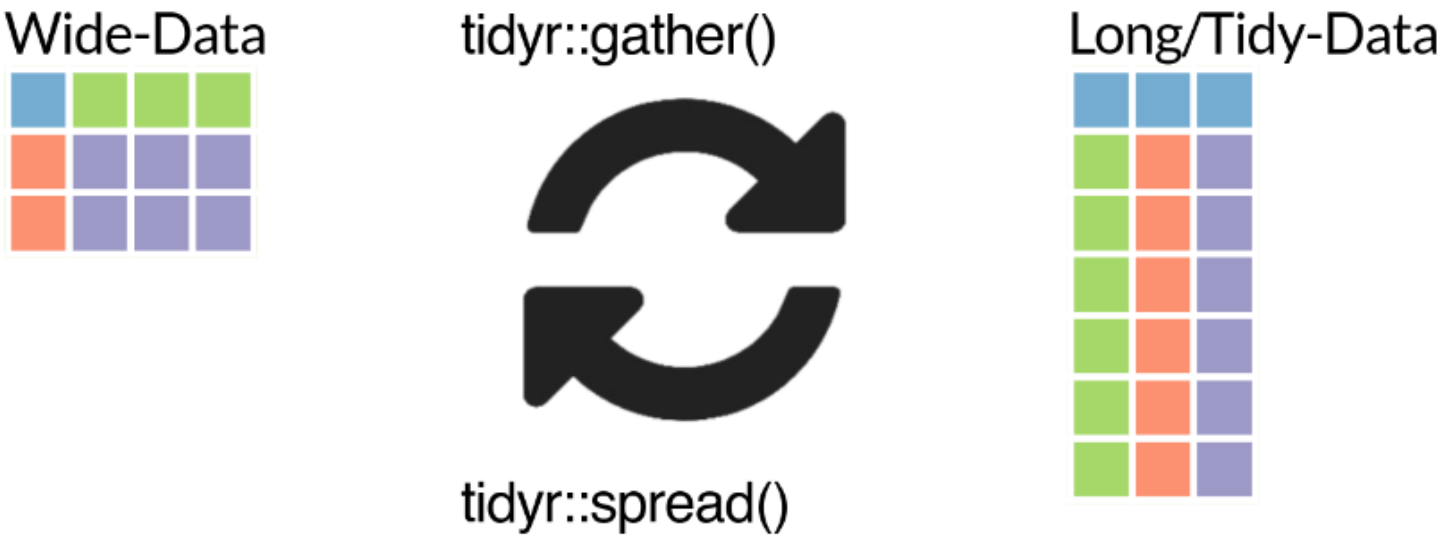
Year	Alice	Bob	Charlie
2010	105	100	90
2011	110	97	95

Long/Tidy-Data

Name	Year	Sales
Alice	2010	105
Alice	2011	110
Bob	2010	100
Bob	2011	97
Charlie	2010	90
Charlie	2011	95

Source: slides by David Zimmermann

tidyr for converting between the two formats



Wide-format

Year	Alice	Bob	Charlie
2010	105	100	90
2011	110	97	95

Long/Tidy-Data

Name	Year	Sales
Alice	2010	105
Alice	2011	110
Bob	2010	100
Bob	2011	97
Charlie	2010	90
Charlie	2011	95

Converting from wide format to tidy

```
data %>%
  gather(key = "Name", value = "Sales", -year)
```

Your turn

Is gene expression matrix in a tidy format?

Your turn

Is gene expression matrix in a tidy format?

```
df_exprs <- readRDS("data/TCGA_exprs.rds")
```

id	SLC7A2_ENSG00000003989	HSPB6_ENSG00000004776	PDK4_ENSG00000004799	ZMYND10_ENSG00000004838
TCGA-3C-AAAU	8.6058498	2.0863915	3.4674773	3.8919134
TCGA-3C-AALI	3.0177057	2.8843299	2.7492562	2.3850030
TCGA-3C-AALJ	4.3425394	3.0862352	5.3112761	2.7666318
TCGA-3C-AALK	4.8619597	2.7547870	3.4076714	4.4072346
TCGA-4H-AAAK	4.2986373	2.8731872	4.5942167	4.6344716
TCGA-5L-AAT0	2.5955476	2.4720156	3.1656451	3.4055321

Convert this gene expression data into a tidy format using `tidyr::gather()`.

dplyr for joining data frames

Joining data frames

Source: slides by David Zimmermann

inner_join()

inner_join(df_a, df_b,
by = "id")

Output Data

full_join()

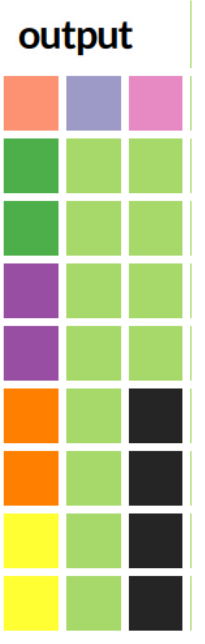

```
full_join(df_a, df_b,
          by = "id")
```

Output Data

left_join()


```
left_join(df_a, df_b,
by = "id")
```

Output Data



Getting insights into high-dimensional data

Now considering *all genes*.

Q1: Do ER+ and ER- patients have generally quite different expression profiles, or are they quite similar?

Q2: Same question about PAM50 subtypes.

How would you answer these questions?

Briefly discussed last time:

How to visually explore data if there are more than two or three variables?

That is, how to visualise high-dimensional data?

Briefly discussed last time:

How to visually explore data if there are more than two or three variables?

That is, how to visualise high-dimensional data?

Options include:

- visualise a subset of variables,
    - randomly selected
    - selection based on summary statistics
- compute summary statistics and visualise those instead
- apply a dimensionality reduction method such as PCA

Why you should first visualise your data before launching your favourite ML model

Why you should first visualise your data before launching your favourite ML model

You should have a rough idea about:

- Data quality
  - Are there any batch effects?
  - Are there possible sample mislabellings?
  - Outliers?
- Distributions - should you apply a transformation?
- Is there a feature which explains majority of variation in the data (e.g. gender, age, batch)

Why you should first visualise your data before launching your favourite ML model

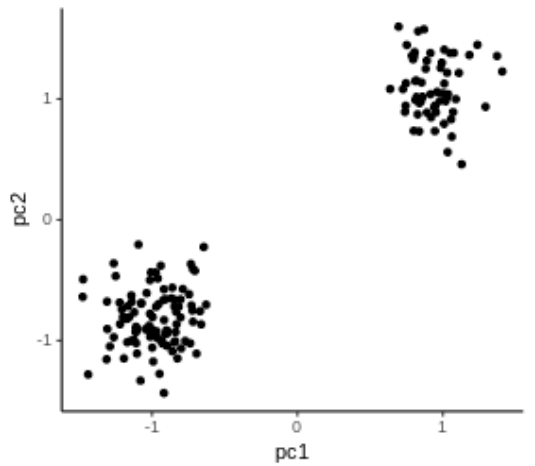
You should have a rough idea about:

- Data quality
  - Are there any batch effects?
  - Are there possible sample mislabellings?
  - Outliers?
- Distributions - should you apply a transformation?
- Is there a feature which explains majority of variation in the data (e.g. gender, age, batch)

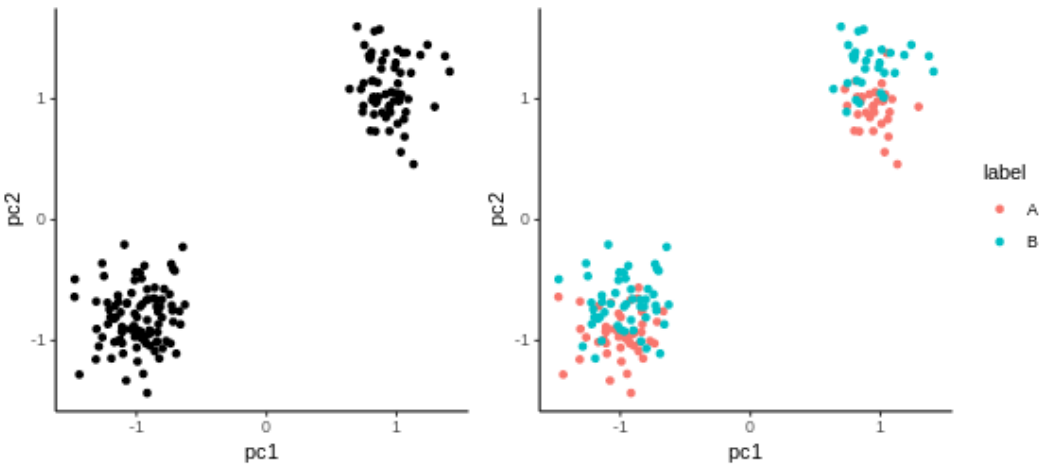
Before fitting a fancy complicated model, you should consider starting with a simple one, to have a rough idea about:

- Expected model fit:
  - Is there strong (linear) signal?
  - Your expected prediction accuracy (will it be close to random or close to 100%)

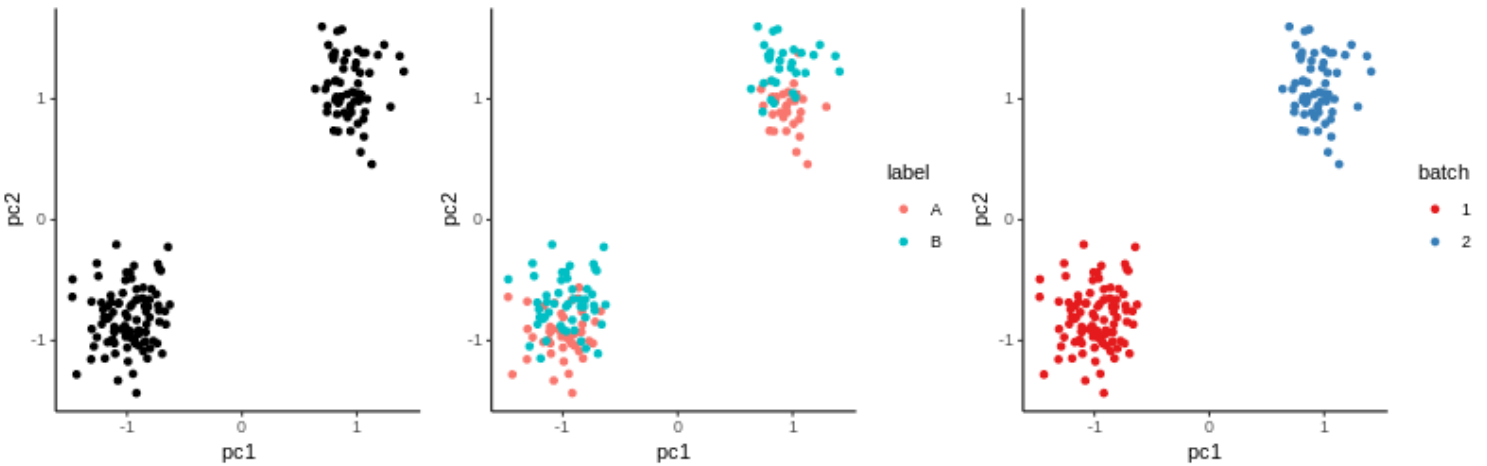
Why you should first visualise your data before launching an ML algorithm



Why you should first visualise your data before launching an ML algorithm



Why you should first visualise your data before launching an ML algorithm



Data science workflow within tidyverse