Workshop 5
The tidyverse and beyond

- I still haven't found what I'm looking for

Brendan Palmer,
Statistics & Data Analysis Unit,
Clinical Research Facility - Cork

Last leg with dplyr

- summarise() or summarize() changes the analysis from the overall dataset to individual specified groups

```
> flights %>%
+ summarise(mean(dep_delay, na.rm = TRUE))
# A tibble: 1 x 1
  `mean(dep_delay, na.rm = TRUE)`
                            <dbl>
1                        12.63907
> |
```

- works best in conjunction with group_by()

```
> flights %>%
+ group_by(month) %>%
+ summarise(mean(dep_delay, na.rm = TRUE))
# A tibble: 12 x 2
   month `mean(dep_delay, na.rm = TRUE)`
   <int>                           <dbl>
 1     1                        10.036665
 2     2                        10.816843
 3     3                        13.227076
 4     4                        13.938038
 5     5                        12.986859
 6     6                        20.846332
 7     7                        21.727787
 8     8                        12.611040
 9     9                         6.722476
10    10                         6.243988
11    11                         5.435362
12    12                        16.576688
> |
```

Worksheet A

Open ws5_script1_working_with_dplyr_partB.R

- Lets pick up where we left off last week and finish off examining dplyr with the join() and summarise() functions

Joining data frames

- Important to understand the chain of relations between the tables

- Variables used to connect each pair of tables are called **keys**

  - primary keys
  - foreign keys

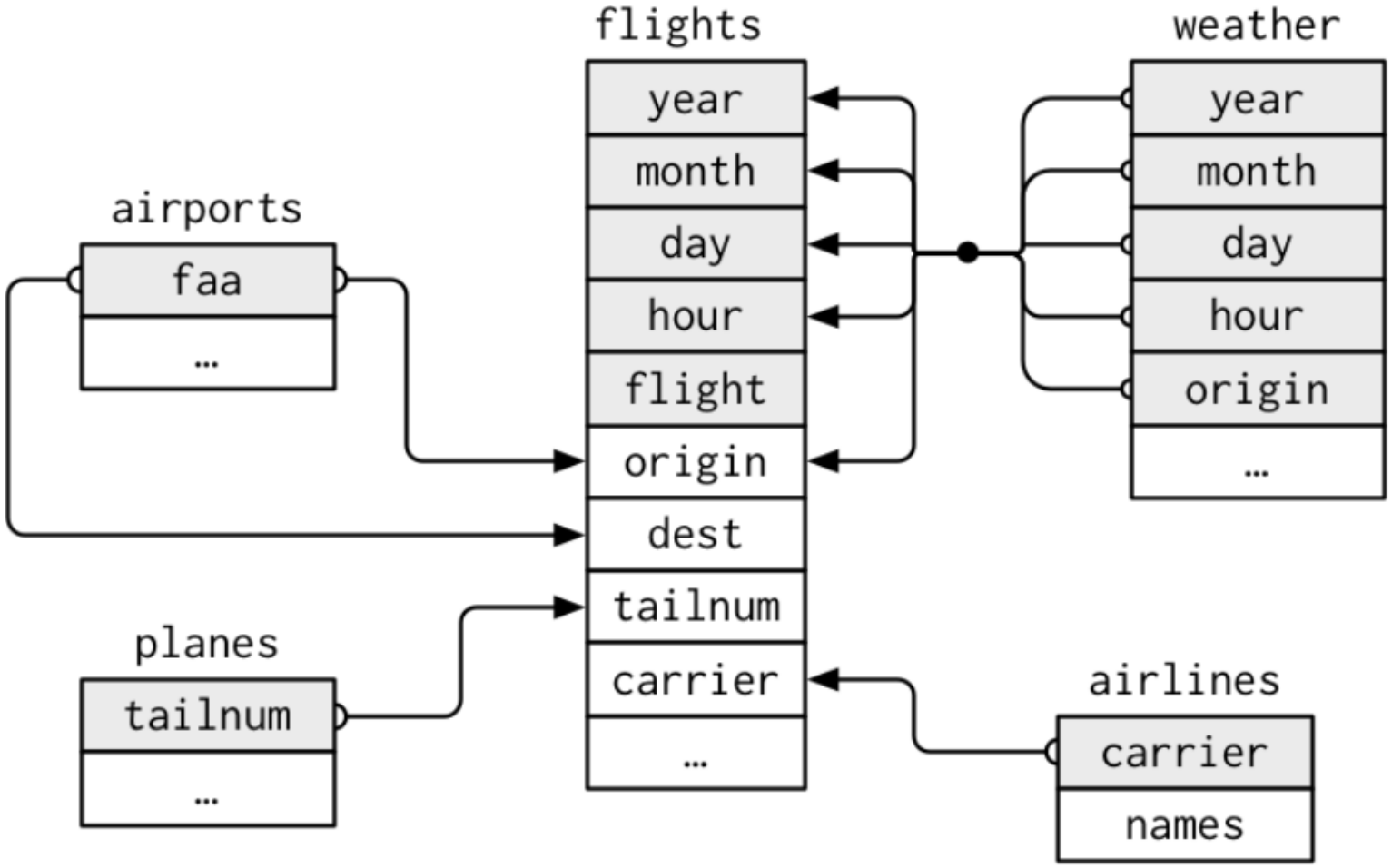
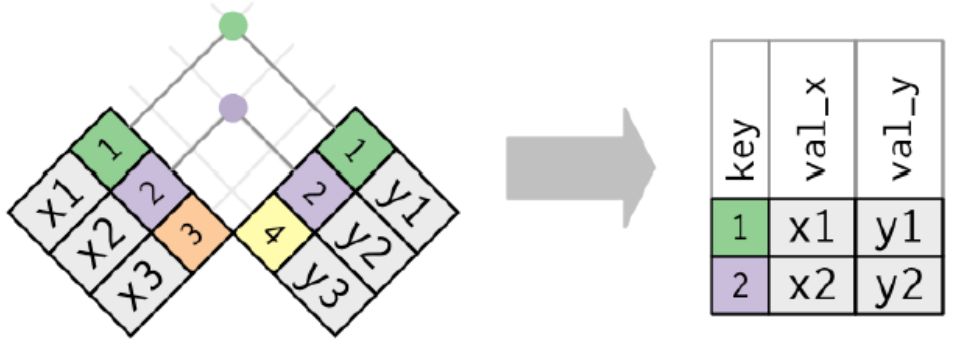


Image: Hadley Wickham R for Data Science Chapter 10

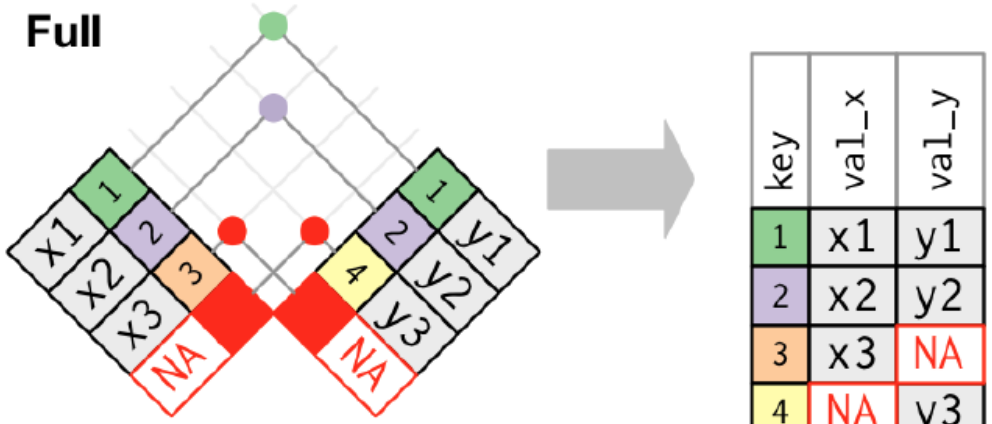
Types of join

- Inner join: Matches pairs of observations whenever their keys are equal
- Unmatched rows are not included



key	val_x	val_y
1	x1	y1
2	x2	y2

- Outer joins:
- left join keeps all the observations in x
- right join keeps all the observations in y
- full join keeps all the observations in a and y



key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA
4	NA	y3

Image: Hadley Wickham R for Data Science Chapter 10

Worksheet B

Open ws5_script2_working_with_dplyr_partB.R

- Lets pick up where we left off last week and finish off examining dplyr with the join() and summarise() functions

Missing values

- if you apply a calculation to a column with missing values, the output will be a missing value

$$1 + 2 + 3 = 6$$

$$1 + 2 + NA = NA$$

- in simple scenarios, NA's can be removed in advance of the calculation with na.rm argument

Implicit versus explicit missing data

- Implicit
    - The absence of a presence
    - simply not present in the data

```
x <- c(1, 2, 4)
```

- Explicit
    - The presence of an absence
    - flagged with NA

```
x <- c(1, 2, NA, 4)
```

- For each data set you will need to determine to nature of the missing data to decide how to proceed
    - remove
    - impute

- For next generation sequencing, zeros are a big issue

Worksheet C

Open ws5_script3_missing_data.R

Exploratory data analysis

- Generate questions about your data
- Search for answers by visualising, transforming and modelling the data
- Use this information to refine the question or generate new questions
- Understand the type of data you are working with

Variation

- If you measure any continuous variable twice, you will get slightly different results

- Categorical measurements may also vary if you take readings from different subjects

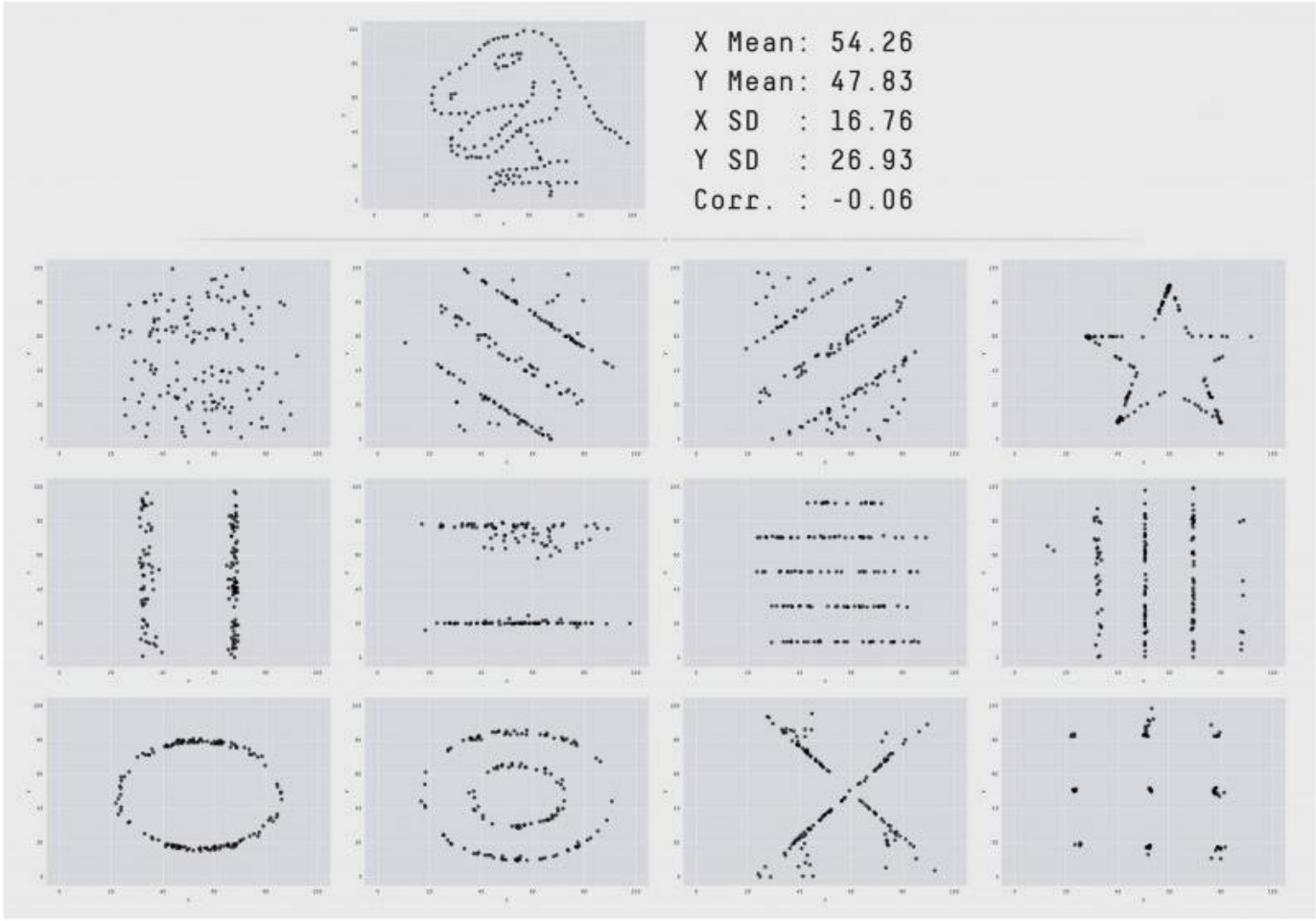
- Every variable has its own pattern of variation

- The best way to understand that pattern is to visualise the distribution of the variables values

Always visualise your data

- At the beginning of this course we used ggplot2 over and over

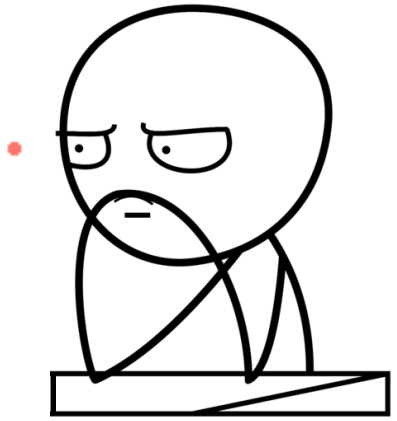
- Once you have tidied your data, you should always generate some visual outputs to check;
    - distribution
    - variance
    - subgroups
    - anomalies



The Datasaurus Dozen. While different in appearance, each dataset has the same summary statistics (mean, standard deviation, and Pearson's correlation) to two decimal places.

https://www.autodeskresearch.com/publications/samestats

Max wind speeds nycflights13 for February



What's going on here?

Mean wind speed 12th Feb at EWR including this value = 55.66898

Mean wind speed 12th Feb at EWR excluding this value = 12.50848

Worksheet D

Open ws5_script4_exploratory_data_analysis.R

Next week

Workshop 6: Don't look back in anger

- writing clear re-useable code

- We will go through some useful tips to help you along the way

- We will work with a template R script for you to populate with lines of code

- The script will be divided into sections with space allocated for readr, tidyr, dplyr and ggplot2 lines of code

- You will be given a data set that will import, tidy, transform and visualise

Other useful packages

- Click on the link below for a quick reference page to other, topic specific, R packages

https://support.rstudio.com/hc/en-us/articles/201057987-Quick-list-of-useful-R-packages