

# ETC1010: Introduction to Data Analysis

## Week 4, part A

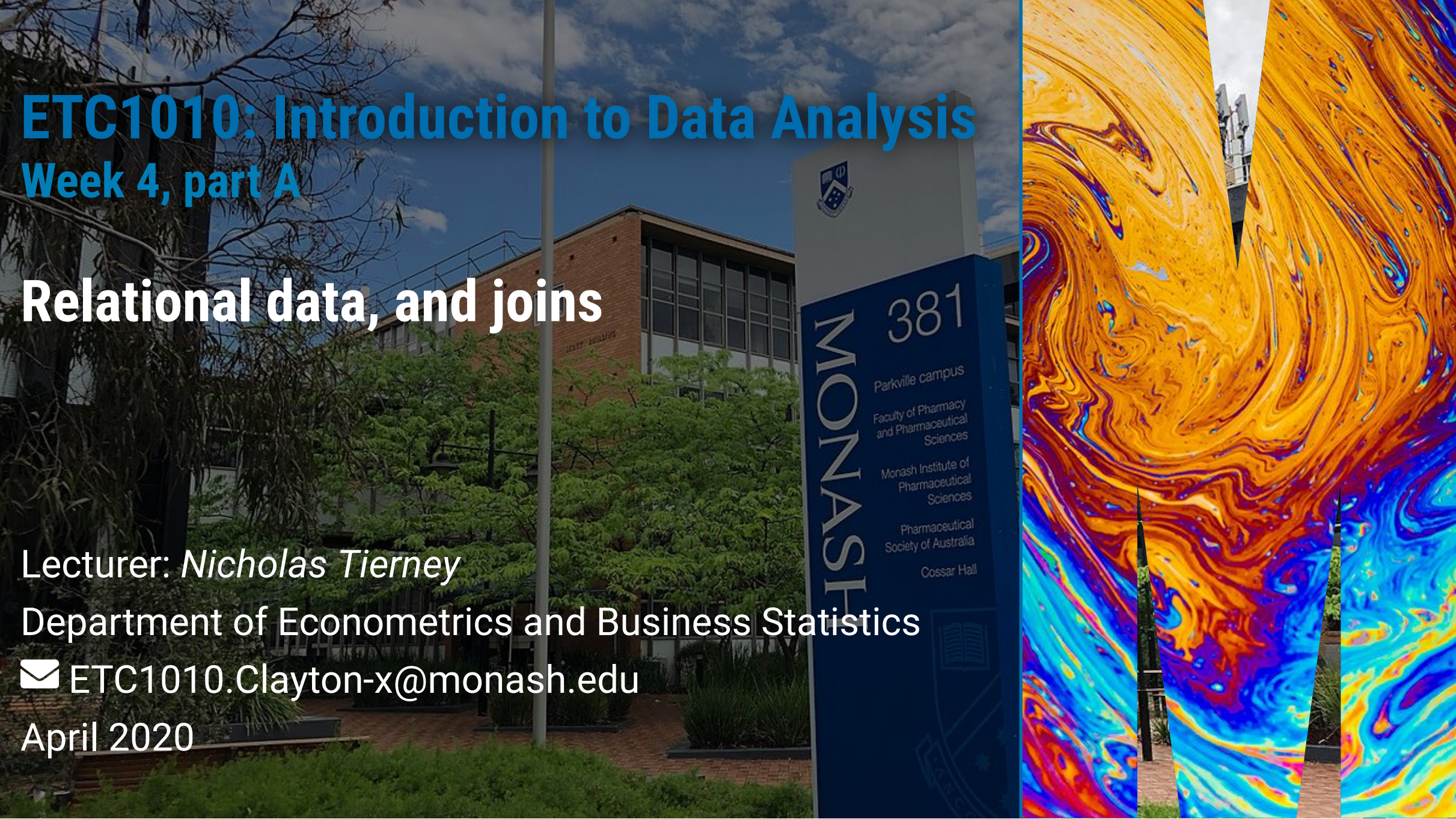
### Relational data, and joins

Lecturer: *Nicholas Tierney*

Department of Econometrics and Business Statistics

✉ [ETC1010.Clayton-x@monash.edu](mailto:ETC1010.Clayton-x@monash.edu)

April 2020





# Recap

- consultation hours
- ggplot
- tidy data
- drawing mental models

# Recap: ggplot mental model

- Note: take a moment to try this out yourself.

[demo]

# Recap: Tidy data

wide

id	x	y	z
1	a	c	e
2	b	d	f

long

id	key	val
1	x	a
2	x	b
1	y	c
2	y	d
1	z	e
2	z	f

# Recap: Tidy data - animation

wide

id	x	y	z
1	a	c	e
2	b	d	f

# Overview

- What is relational data?
- Keys
- Different sorts of joins
- Using joins to follow an aircraft flight path

# Relational data

- Data analysis **rarely involves** only a single table of data.
- To answer questions you generally need to combine many tables of data
- Multiple tables of data are called *relational data*
- It is the **relations**, not just the individual datasets, that are important.

# nycflights13

- Data set of flights that departed NYC in 2013 from <https://www.transtats.bts.gov> - a public database of all USA commercial airline flights. It has five tables:
  1. flights
  2. airlines
  3. airports
  4. planes
  5. weather



# flights

```
library(nycflights13)
flights
## # A tibble: 336,776 x 19
##   year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_
##   <int> <int> <int>   <int>         <int>         <dbl>   <int>         <int>
## 1  2013     1     1     517           515           2     830           819
## 2  2013     1     1     533           529           4     850           830
## 3  2013     1     1     542           540           2     923           850
## 4  2013     1     1     544           545          -1    1004          1022
## 5  2013     1     1     554           600          -6     812           837
## 6  2013     1     1     554           558          -4     740           728
## 7  2013     1     1     555           600          -5     913           854
## 8  2013     1     1     557           600          -3     709           723
## 9  2013     1     1     557           600          -3     838           846
## 10 2013     1     1     558           600          -2     753           745
## # ... with 336,766 more rows, and 10 more variables: carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <
## #   minute <dbl>, time_hour <dtm>
```

# airlines

```
airlines
```

```
## # A tibble: 16 x 2
```

```
##   carrier name
```

```
##   <chr>    <chr>
```

```
## 1 9E      Endeavor Air Inc.
```

```
## 2 AA      American Airlines Inc.
```

```
## 3 AS      Alaska Airlines Inc.
```

```
## 4 B6      JetBlue Airways
```

```
## 5 DL      Delta Air Lines Inc.
```

```
## 6 EV      ExpressJet Airlines Inc.
```

```
## 7 F9      Frontier Airlines Inc.
```

```
## 8 FL      AirTran Airways Corporation
```

```
## 9 HA      Hawaiian Airlines Inc.
```

```
## 10 MQ     Envoy Air
```

```
## 11 00     SkyWest Airlines Inc.
```

```
## 12 UA     United Air Lines Inc.
```

```
## 13 US     US Airways Inc.
```

```
## 14 VX     Virgin America
```

```
## 15 WN     Southwest Airlines Co.
```

```
## 16 YV     Mesa Airlines Inc.
```

# airports

```
airports
```

```
## # A tibble: 1,458 x 8
```

##	faa	name	lat	lon	alt	tz	dst	tzone
##	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
##	1 04G	Lansdowne Airport	41.1	-80.6	1044	-5	A	America/New_
##	2 06A	Moton Field Municipal Airport	32.5	-85.7	264	-6	A	America/Chic
##	3 06C	Schaumburg Regional	42.0	-88.1	801	-6	A	America/Chic
##	4 06N	Randall Airport	41.4	-74.4	523	-5	A	America/New_
##	5 09J	Jekyll Island Airport	31.1	-81.4	11	-5	A	America/New_
##	6 0A9	Elizabethton Municipal Airport	36.4	-82.2	1593	-5	A	America/New_
##	7 0G6	Williams County Airport	41.5	-84.5	730	-5	A	America/New_
##	8 0G7	Finger Lakes Regional Airport	42.9	-76.8	492	-5	A	America/New_
##	9 0P2	Shoestring Aviation Airfield	39.8	-76.6	1000	-5	U	America/New_
##	10 0S9	Jefferson County Intl	48.1	-123.	108	-8	A	America/Los_
##	#	... with 1,448 more rows						

# print-planes

```
planes
```

```
## # A tibble: 3,322 x 9
```

```
##   tailnum year type      manufacturer      model    engines seats speed
##   <chr>   <int> <chr>      <chr>      <chr>    <int> <int> <int>
## 1 N10156  2004 Fixed wing multi en... EMBRAER      EMB-145...      2    55    NA
## 2 N102UW  1998 Fixed wing multi en... AIRBUS INDUSTRIE A320-214      2   182    NA
## 3 N103US  1999 Fixed wing multi en... AIRBUS INDUSTRIE A320-214      2   182    NA
## 4 N104UW  1999 Fixed wing multi en... AIRBUS INDUSTRIE A320-214      2   182    NA
## 5 N10575  2002 Fixed wing multi en... EMBRAER      EMB-145...      2    55    NA
## 6 N105UW  1999 Fixed wing multi en... AIRBUS INDUSTRIE A320-214      2   182    NA
## 7 N107US  1999 Fixed wing multi en... AIRBUS INDUSTRIE A320-214      2   182    NA
## 8 N108UW  1999 Fixed wing multi en... AIRBUS INDUSTRIE A320-214      2   182    NA
## 9 N109UW  1999 Fixed wing multi en... AIRBUS INDUSTRIE A320-214      2   182    NA
## 10 N110UW 1999 Fixed wing multi en... AIRBUS INDUSTRIE A320-214      2   182    NA
## # ... with 3,312 more rows
```

# weather

```
weather
```

```
## # A tibble: 26,115 x 15
```

```
##   origin  year month   day  hour  temp  dewp humid wind_dir wind_speed wind_gust p
```

```
##   <chr>   <int> <int> <int> <int> <dbl> <dbl> <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
```

```
## 1 EWR      2013     1     1     1  39.0  26.1  59.4     270     10.4     NA
```

```
## 2 EWR      2013     1     1     2  39.0  27.0  61.6     250      8.06     NA
```

```
## 3 EWR      2013     1     1     3  39.0  28.0  64.4     240     11.5     NA
```

```
## 4 EWR      2013     1     1     4  39.9  28.0  62.2     250     12.7     NA
```

```
## 5 EWR      2013     1     1     5  39.0  28.0  64.4     260     12.7     NA
```

```
## 6 EWR      2013     1     1     6  37.9  28.0  67.2     240     11.5     NA
```

```
## 7 EWR      2013     1     1     7  39.0  28.0  64.4     240     15.0     NA
```

```
## 8 EWR      2013     1     1     8  39.9  28.0  62.2     250     10.4     NA
```

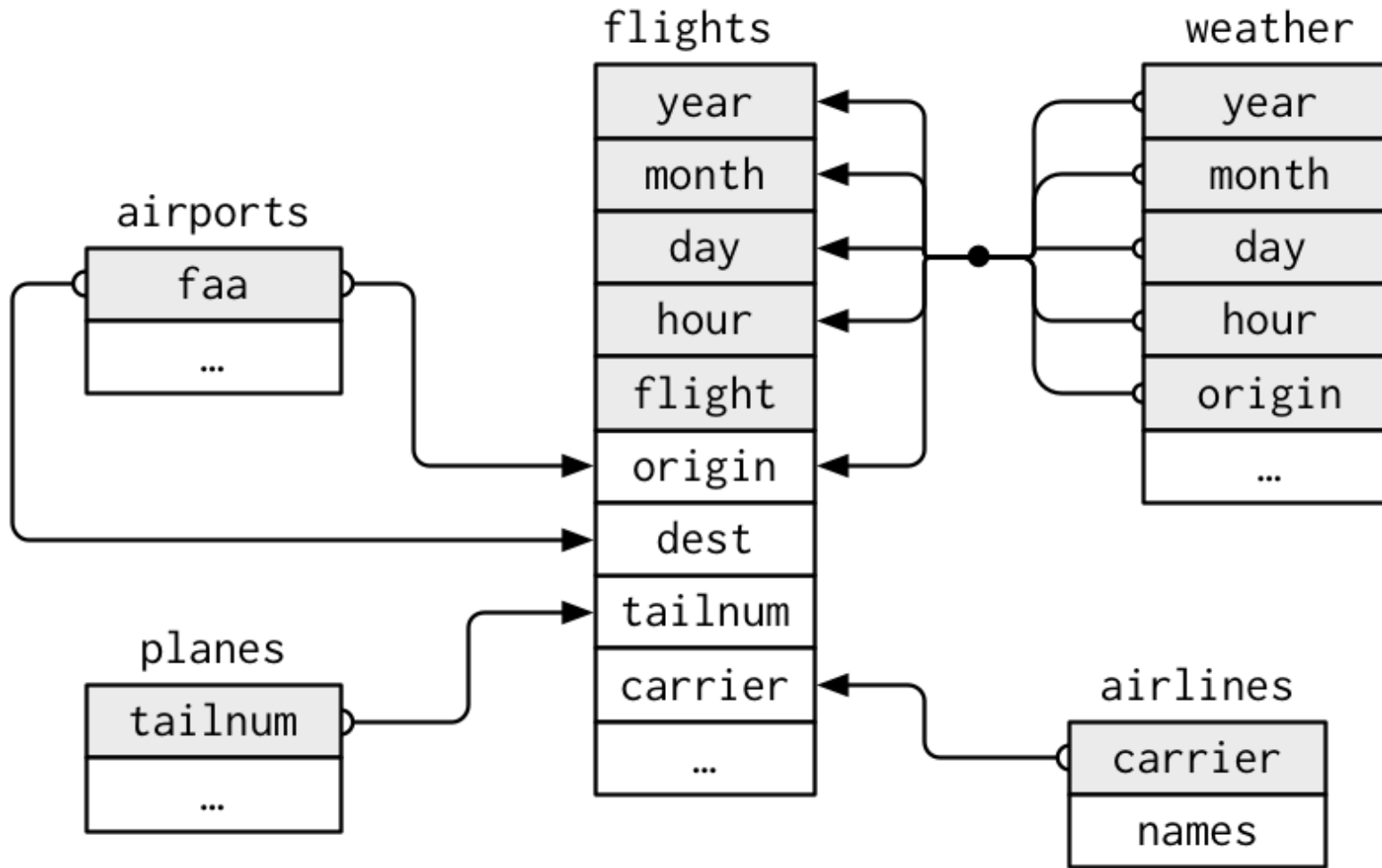
```
## 9 EWR      2013     1     1     9  39.9  28.0  62.2     260     15.0     NA
```

```
## 10 EWR      2013     1     1    10  41    28.0  59.6     260     13.8     NA
```

```
## # ... with 26,105 more rows, and 3 more variables: pressure <dbl>, visib <dbl>,
```

```
## #   time_hour <dtm>
```

# Concept map of tables and joins from the text





# Keys

- Keys = variables used to connect records in one table to another.
- In the `nycflights13` data,
  - `flights` connects to `planes` by a single variable `tailnum`
  - `flights` connects to `airlines` by a single variable `carrier`
  - `flights` connects to `airports` by two variables, `origin` and `dest`
  - `flights` connects to `weather` using multiple variables, `origin`, and `year`, `month`, `day` and `hour`.

# Your turn: go to [rstudio.cloud](https://rstudio.cloud)

- Load the Lahman package, which contains multiple tables of baseball data.
- What key(s) connect the batting table with the salary table?
- Can you draw out a diagram of the connections amongst the tables?

04 : 00

# Joins

- "mutating joins", add variables from one table to another.
- There is always a decision on what observations are copied to the new table as well.
- Let's discuss how joins work using some [lovely animations](#) provided by [Garrick Aden-Buie](#).

# Example data

$x$

1	x1
2	x2
3	x3

$y$

1	y1
2	y2
4	y4

# Left Join (Generally the one you want to use)

All observations from the "left" table, but only the observations from the "right" table that match those in the left.

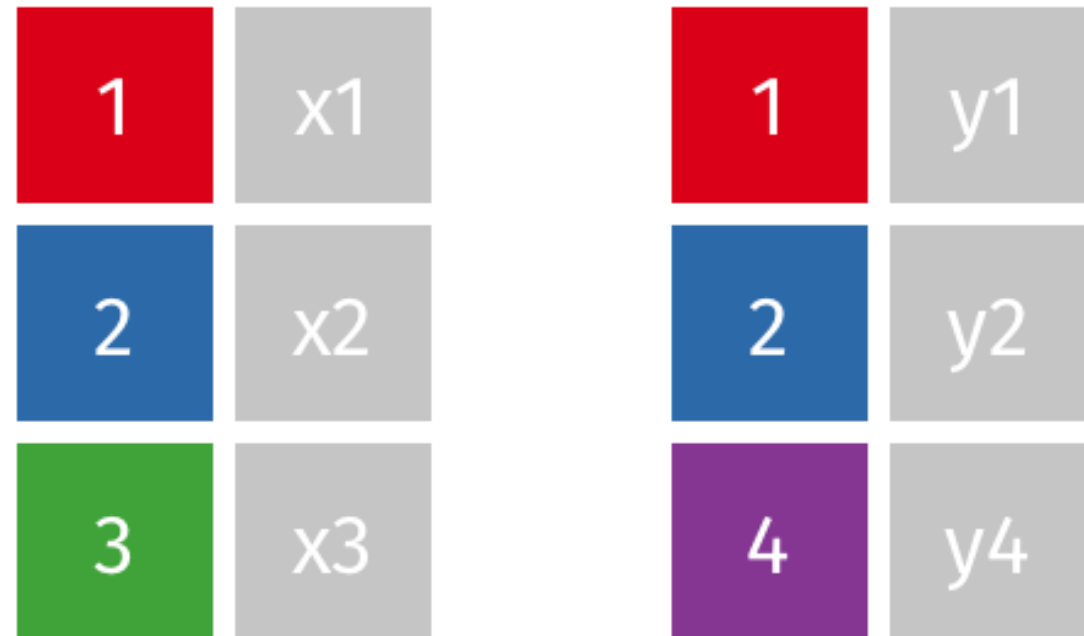
`left_join(x, y)`

1	x1	1	y1
2	x2	2	y2
3	x3	4	y4

# Right Join

Same as left join, but in reverse.

`right_join(x, y)`





# Inner join

Intersection between the two tables, only the observations that are in both

`inner_join(x, y)`

1	x1	1	y1
2	x2	2	y2
3	x3	4	y4

# Outer (full) join

Union of the two tables,  
all observations from  
both, and missing values  
might get added

`full_join(x, y)`

1	x1	1	y1
2	x2	2	y2
3	x3	4	y4

# Combine full airline name with flights data?

```
flights
```

```
## # A tibble: 336,776 x 19
```

```
##   year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_
```

```
##   <int> <int> <int>   <int>         <int>         <dbl>   <int>         <int>
```

```
## 1  2013     1     1     517           515         2     830           819
```

```
## 2  2013     1     1     533           529         4     850           830
```

```
## 3  2013     1     1     542           540         2     923           850
```

```
## 4  2013     1     1     544           545        -1    1004          1022
```

```
## 5  2013     1     1     554           600        -6     812           837
```

```
## 6  2013     1     1     554           558        -4     740           728
```

```
## 7  2013     1     1     555           600        -5     913           854
```

```
## 8  2013     1     1     557           600        -3     709           723
```

```
## 9  2013     1     1     557           600        -3     838           846
```

```
## 10 2013     1     1     558           600        -2     753           745
```

```
## # ... with 336,766 more rows, and 10 more variables: carrier <chr>, flight <int>,
```

```
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <
```

```
## #   minute <dbl>, time_hour <dtm>
```

# Combine full airline name with flights data?

```
airlines
## # A tibble: 16 x 2
##   carrier name
##   <chr>    <chr>
## 1 9E      Endeavor Air Inc.
## 2 AA      American Airlines Inc.
## 3 AS      Alaska Airlines Inc.
## 4 B6      JetBlue Airways
## 5 DL      Delta Air Lines Inc.
## 6 EV      ExpressJet Airlines Inc.
## 7 F9      Frontier Airlines Inc.
## 8 FL      AirTran Airways Corporation
## 9 HA      Hawaiian Airlines Inc.
## 10 MQ     Envoy Air
## 11 00     SkyWest Airlines Inc.
## 12 UA     United Air Lines Inc.
## 13 US     US Airways Inc.
## 14 VX     Virgin America
## 15 WN     Southwest Airlines Co.
## 16 YV     Mesa Airlines Inc.
```

# Combine airlines & flights using left\_join()

```
flights %>%  
  left_join(airlines,  
            by = "carrier") %>%  
  glimpse()
```

```
## Observations: 336,776  
## Variables: 20  
## $ year          <int> 2013, 2013, 2013, 2013, 2013,  
## $ month         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1,  
## $ day           <int> 1, 1, 1, 1, 1, 1, 1, 1, 1,  
## $ dep_time      <int> 517, 533, 542, 544, 554, 555, 556, 557, 558,  
## $ sched_dep_time <int> 515, 529, 540, 545, 600, 605, 606, 607, 608,  
## $ dep_delay     <dbl> 2, 4, 2, -1, -6, -4, -5, -4, -3,  
## $ arr_time      <int> 830, 850, 923, 1004, 812, 844, 858, 917, 854,  
## $ sched_arr_time <int> 819, 830, 850, 1022, 837, 840, 858, 917, 854,  
## $ arr_delay     <dbl> 11, 20, 33, -18, -25, 12, 12, 12, 12,  
## $ carrier       <chr> "UA", "UA", "AA", "B6", "DL", "AA", "B6", "DL", "DL",  
## $ flight        <int> 1545, 1714, 1141, 725, 468, 1545, 1714, 1141, 725,  
## $ tailnum       <chr> "N14228", "N24211", "N619", "N14228", "N24211", "N619",  
## $ origin        <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "LGA", "JFK", "JFK",  
## $ dest          <chr> "IAH", "IAH", "MIA", "BQN", "IAH", "IAH", "MIA", "BQN", "IAH",  
## $ air_time      <dbl> 227, 227, 160, 183, 116, 227, 227, 160, 183,  
## $ distance      <dbl> 1400, 1416, 1089, 1576, 759, 1400, 1416, 1089, 1576,  
## $ hour          <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6,
```

# Example: flights joining to airports

```
flights %>%  
  left_join(  
    airports,  
    by = c("origin" = "faa")) %>%  
  glimpse()
```

```
## Observations: 336,776  
## Variables: 26  
## $ year          <int> 2013, 2013, 2013, 2013, 2013,  
## $ month         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1,  
## $ day           <int> 1, 1, 1, 1, 1, 1, 1, 1, 1,  
## $ dep_time      <int> 517, 533, 542, 544, 554, 555, 555, 555, 555,  
## $ sched_dep_time <int> 515, 529, 540, 545, 600, 605, 605, 605, 605,  
## $ dep_delay     <dbl> 2, 4, 2, -1, -6, -4, -5, -5, -5,  
## $ arr_time      <int> 830, 850, 923, 1004, 812, 837, 845, 854, 917,  
## $ sched_arr_time <int> 819, 830, 850, 1022, 837, 845, 854, 863, 917,  
## $ arr_delay     <dbl> 11, 20, 33, -18, -25, 12, 10, 19, 12,  
## $ carrier       <chr> "UA", "UA", "AA", "B6", "DL", "AA", "B6", "DL", "DL",  
## $ flight        <int> 1545, 1714, 1141, 725, 468, 1545, 1714, 1141, 725,  
## $ tailnum       <chr> "N14228", "N24211", "N619", "N14228", "N24211", "N619",  
## $ origin        <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "LGA", "JFK", "JFK",  
## $ dest          <chr> "IAH", "IAH", "MIA", "BQN", "IAH", "IAH", "MIA", "BQN", "IAH",  
## $ air_time      <dbl> 227, 227, 160, 183, 116, 227, 227, 160, 183,  
## $ distance      <dbl> 1400, 1416, 1089, 1576, 706, 1400, 1416, 1089, 1576,  
## $ hour          <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 5, 5, 5, 5, 6, 5, 6, 6, 6,  
## $ minute        <dbl> 15, 29, 40, 45, 0, 58, 0, 15, 29, 40, 45, 0, 58, 0,
```



# Airline travel, ontime data

```
plane_N4YRAA <- read_csv("data/plane_N4YRAA.csv")
```

```
glimpse(plane_N4YRAA)
```

```
## Observations: 145
```

```
## Variables: 8
```

```
## $ FL_DATE    <date> 2017-05-26, 2017-05-02, 2017-05-05, 2017-05-11, 2017-05-03, 2017-
```

```
## $ CARRIER   <chr> "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA",
```

```
## $ FL_NUM     <dbl> 2246, 2276, 2278, 2287, 2288, 2291, 2297, 2297, 2297, 2297, 2302,
```

```
## $ ORIGIN     <chr> "CVG", "DFW", "DFW", "STL", "IND", "CHS", "DFW", "DFW", "MKE", "MK
```

```
## $ DEST       <chr> "DFW", "IND", "OKC", "ORD", "DFW", "DFW", "MKE", "MKE", "DFW", "DF
```

```
## $ DEP_TIME   <chr> "0748", "2020", "0848", "0454", "0601", "0807", "0700", "0659", "1
```

```
## $ ARR_TIME   <chr> "0917", "2323", "0941", "0600", "0719", "0947", "0905", "0909", "1
```

```
## $ DISTANCE   <dbl> 812, 761, 175, 258, 761, 987, 853, 853, 853, 853, 447, 447, 761, 8
```

# Airline travel, airport location

```
airport_raw <- read_csv("data/airports.csv")

airport_raw %>%
  select(AIRPORT,
         LATITUDE,
         LONGITUDE,
         AIRPORT_STATE_NAME) %>%
  glimpse()
## Observations: 13,094
## Variables: 4
## $ AIRPORT      <chr> "01A", "03A", "04A", "05A", "06A", "07A", "08A", "09A",
## $ LATITUDE     <dbl> 58.10944, 65.54806, 68.08333, 67.57000, 57.74528, 55.554
## $ LONGITUDE    <dbl> -152.90667, -161.07167, -163.16667, -148.18389, -152.882
## $ AIRPORT_STATE_NAME <chr> "Alaska", "Alaska", "Alaska", "Alaska", "Alaska", "Alask
```

# Our Turn: Joining the two tables to show flight movements

- Go to [rstudio.cloud](https://rstudio.cloud) and open "flight-movements.Rmd" and complete exercise - the aim is to show flight movement on the map
- Next: Open "nycflights.Rmd"

# Learning more

- The coat explanation of joins: Different types of joins explained using a person and a coat, by [Leight Tami](#)

# References

- Chapter 13 of R4DS