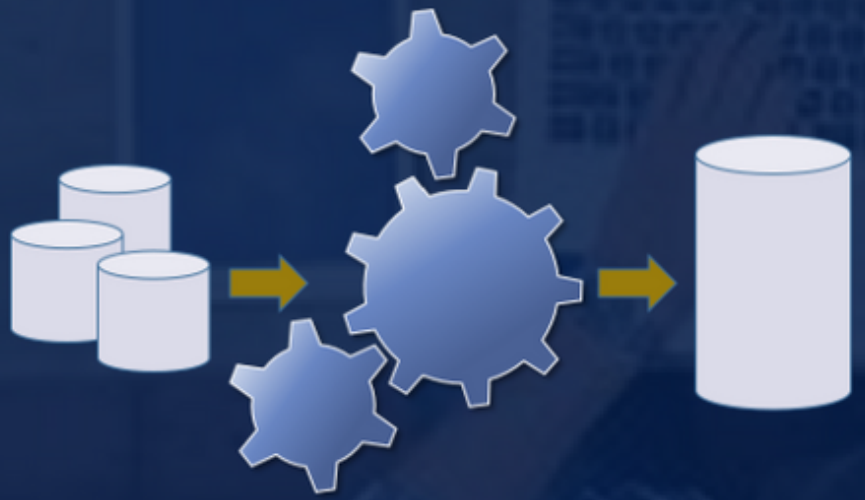
Rsquared Academy



Data Wrangling with dplyr

Agenda

- inner join
- left join
- right join
- semi join
- anti join
- full join

CASE STUDY

- details of customers who have placed orders and their order details
- details of customers and their orders irrespective of whether a customer has placed orders or not.
- get customer details for all orders
- get customer data, if available, for all orders
- details of customers who have not placed orders
- details of all customers and all orders

Case Study

Order
order_id
order_date
order_number
customer_id
amount

Customer
customer_id
first_name
last_name
city
phone number

Libraries

```
library(dplyr)
library(readr)
```

Data: Orders

```
order <- read_delim('https://raw.githubusercontent.com/rsquaredacademy/d
```

```
## # A tibble: 300 x 3
##       id order_date amount
##    <dbl> <chr>       <dbl>
##  1   368 7/2/2016     365.
##  2   286 11/2/2016   2064.
##  3    28 2/22/2017    432.
##  4   309 3/5/2017     480.
##  5     2 12/28/2016   235.
##  6    31 12/30/2016  2745.
##  7   179 12/21/2016  2358.
##  8   484 11/24/2016  1031.
##  9   115 9/9/2016    1218.
## 10   340 5/6/2017    1184.
## # ... with 290 more rows
```

Data: Customers

```
customer <- read_delim('https://raw.githubusercontent.com/rsquaredacadem
```

```
## # A tibble: 91 x 3
##       id first_name city
##    <dbl> <chr>      <chr>
##  1     1 Elbertine  California
##  2     2 Marcella   Colorado
##  3     3 Daria      Florida
##  4     4 Sherilyn   Distric...
##  5     5 Ketty      Texas
##  6     6 Jethro     California
##  7     7 Jeremiah   California
##  8     8 Constancia Texas
##  9     9 Muire      Idaho
## 10    10 Abigail    Texas
## # ... with 81 more rows
```

Example Data

Age

Name	Age
John	26
Jenny	24
Jacob	28

Height

Name	Height
John	170
Jenny	174
Janet	166

inner join

inner join

Age

Name	Age
John	26
Jenny	24
Jacob	28

Height

Name	Height
John	170
Jenny	174
Janet	166

Output

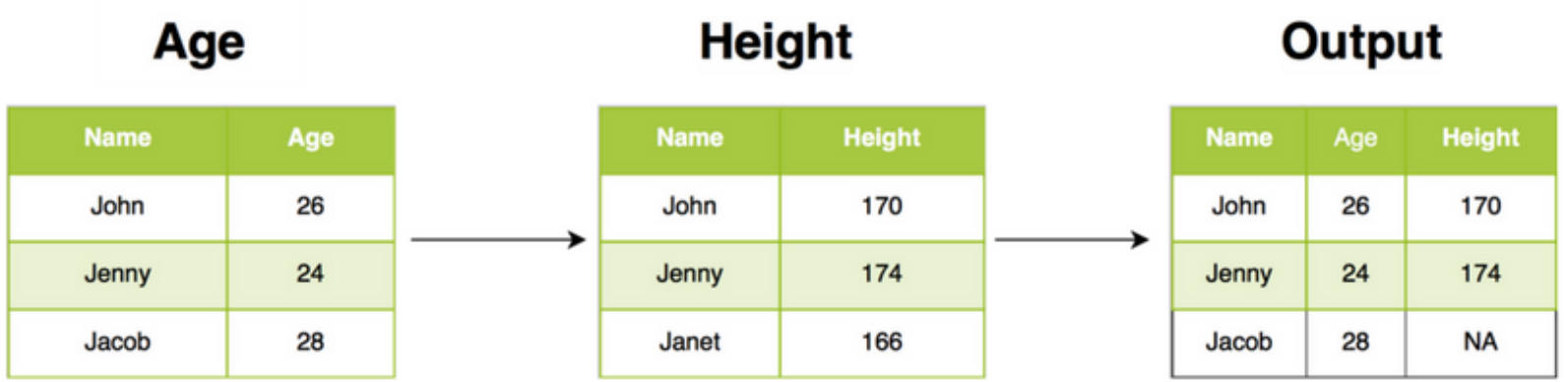
Name	Age	Height
John	26	170
Jenny	24	174

```
inner_join(customer, order, by = "id")
```

```
## # A tibble: 55 x 5
##       id first_name city       order_date amount
##    <dbl> <chr>      <chr>      <chr>       <dbl>
##  1     2 Marcella   Colorado   12/28/2016   235.
##  2     2 Marcella   Colorado   8/31/2016   1150.
##  3     5 Ketty      Texas      1/17/2017    346.
##  4     6 Jethro     California 1/27/2017   2317.
##  5     7 Jeremiah   California 6/21/2016    136.
##  6     7 Jeremiah   California 2/13/2017   1407.
##  7     7 Jeremiah   California 7/8/2016    1914.
##  8     8 Constancia Texas      11/5/2016   2461.
##  9     8 Constancia Texas      5/19/2017   2714.
## 10     9 Muire      Idaho      12/28/2016   187.
## # ... with 45 more rows
```

left join

left join


Age
Name
Age
John
26
Jenny
24
Jacob
28
Height
Name
Height
John
170
Jenny
174
Janet
166
Output
Name
Age
Height
John
26
170
Jenny
24
174
Jacob
28
NA

```
left_join(customer, order, by = "id")
```

```
## # A tibble: 104 x 5
##       id first_name city       order_date amount
##    <dbl> <chr>      <chr>      <chr>       <dbl>
##  1     1 Elbertine  California <NA>          NA
##  2     2 Marcella   Colorado   12/28/2016   235.
##  3     2 Marcella   Colorado   8/31/2016   1150.
##  4     3 Daria      Florida    <NA>          NA
##  5     4 Sherilyn   Distric... <NA>          NA
##  6     5 Ketty      Texas      1/17/2017    346.
##  7     6 Jethro     California 1/27/2017   2317.
##  8     7 Jeremiah   California 6/21/2016    136.
##  9     7 Jeremiah   California 2/13/2017   1407.
## 10     7 Jeremiah   California 7/8/2016    1914.
## # ... with 94 more rows
```

right join

right join

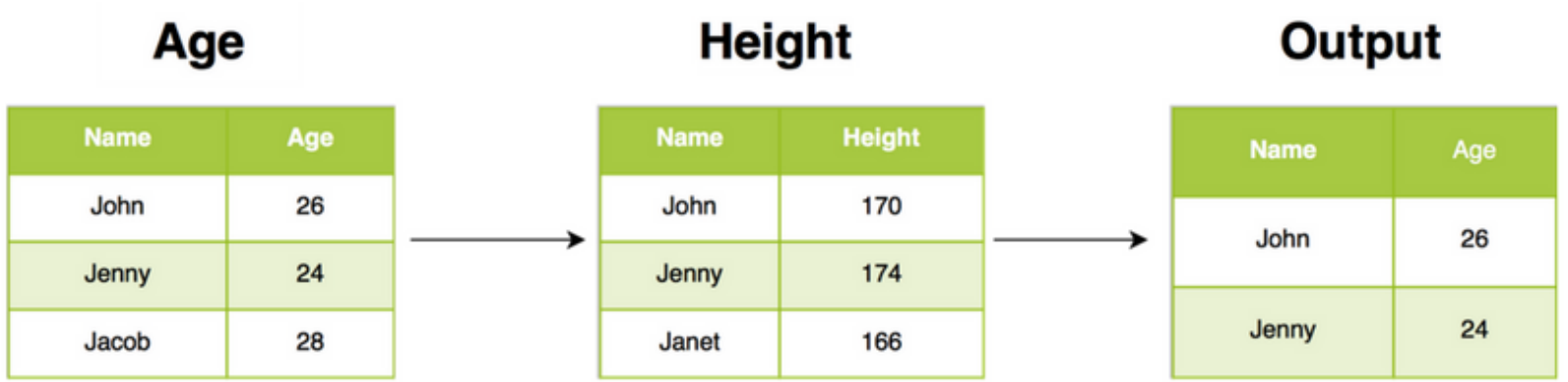

Age
Name
Age
John
26
Jenny
24
Jacob
28
Height
Name
Height
John
170
Jenny
174
Janet
166
Output
Name
Age
Height
John
26
170
Jenny
24
174
Janet
NA
166

```
right_join(customer, order, by = "id")
```

```
## # A tibble: 300 x 5
##       id first_name city      order_date amount
##    <dbl> <chr>      <chr>     <chr>       <dbl>
##  1   368 <NA>       <NA>      7/2/2016     365.
##  2   286 <NA>       <NA>      11/2/2016   2064.
##  3    28 Avrit      Texas     2/22/2017    432.
##  4   309 <NA>       <NA>      3/5/2017     480.
##  5     2 Marcella   Colorado  12/28/2016   235.
##  6    31 Clemmie    Tennessee 12/30/2016  2745.
##  7   179 <NA>       <NA>      12/21/2016  2358.
##  8   484 <NA>       <NA>      11/24/2016  1031.
##  9   115 <NA>       <NA>      9/9/2016    1218.
## 10   340 <NA>       <NA>      5/6/2017    1184.
## # ... with 290 more rows
```

semi join

semi join


Age
Name
Age
John
26
Jenny
24
Jacob
28
Height
Name
Height
John
170
Jenny
174
Janet
166
Output
Name
Age
John
26
Jenny
24

```
semi_join(customer, order, by = "id")
```

```
## # A tibble: 42 x 3
##       id first_name city
##    <dbl> <chr>      <chr>
##  1     2 Marcella   Colorado
##  2     5 Ketty      Texas
##  3     6 Jethro     California
##  4     7 Jeremiah   California
##  5     8 Constancia Texas
##  6     9 Muire      Idaho
##  7    15 Valentijn  California
##  8    16 Monique    Missouri
##  9    20 Colette    Texas
## 10    28 Avrit      Texas
## # ... with 32 more rows
```

anti join

anti join


Age
Name
Age
John
26
Jenny
24
Jacob
28
Height
Name
Height
John
170
Jenny
174
Janet
166
Output
Name
Age
Jacob
28

```
anti_join(customer, order, by = "id")
```

```
## # A tibble: 49 x 3
##       id first_name city
##    <dbl> <chr>      <chr>
##  1     1 Elbertine  California
##  2     3 Daria      Florida
##  3     4 Sherilyn   Distric...
##  4    10 Abigail    Texas
##  5    11 Wynne      Georgia
##  6    12 Pietra     Minnesota
##  7    13 Bram       Iowa
##  8    14 Rees       New York
##  9    17 Orazio     Louisiana
## 10    18 Mason      Texas
## # ... with 39 more rows
```

full join

full join

Age

Name	Age
John	26
Jenny	24
Jacob	28

Height

Name	Height
John	170
Jenny	174
Janet	166

Output

Name	Age	Height
John	26	170
Jenny	24	174
Jacob	28	NA
Janet	NA	166

```
full_join(customer, order, by = "id")
```

```
## # A tibble: 349 x 5
##       id first_name city       order_date amount
##    <dbl> <chr>      <chr>      <chr>       <dbl>
##  1     1 Elbertine  California <NA>          NA
##  2     2 Marcella   Colorado   12/28/2016   235.
##  3     2 Marcella   Colorado   8/31/2016   1150.
##  4     3 Daria      Florida    <NA>          NA
##  5     4 Sherilyn   Distric... <NA>          NA
##  6     5 Ketty      Texas      1/17/2017    346.
##  7     6 Jethro     California 1/27/2017   2317.
##  8     7 Jeremiah   California 6/21/2016    136.
##  9     7 Jeremiah   California 2/13/2017   1407.
## 10     7 Jeremiah   California 7/8/2016    1914.
## # ... with 339 more rows
```

Thank You

For more information please visit our website
www.rsquaredacademy.com