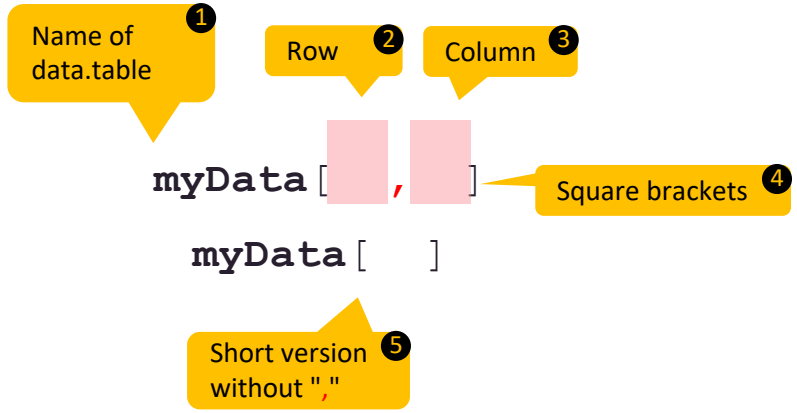
Selecting rows

By selecting data from our dataset, we can answer the following questions

- Which customers joined in 2015?
- Which customers spent the most on a single transaction?
- Which transactions had a purchase amount greater than 100?

General command structure for data.table objects

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
149236	12.06.2007	1	79.95	35.00
...	...	...	...	...



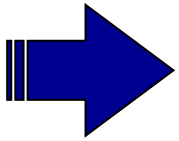
There are multiple ways of selecting rows

1. Selecting rows by row numbers
2. Selecting rows by conditions

Selecting rows by row numbers
Select the first row

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
149236	12.06.2007	1	79.95	35.00
…	…	…	…	…

Select the first row



Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00

1 Returns a data.table

2 Row numbers to be selected

```
myData[1, ]
 myData[1]
```

Sidenote: Selecting does not make changes to the original data.table

Select first row (1)

```
myData[1,]
```

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00

The output of select operations need to be stored via "<–" (2)

```
myData
```

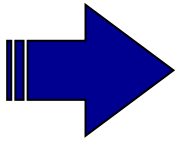
Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	12.06.2007	1	79.95	35.00

myData was not changed (3)

Selecting rows by row numbers
Select the first 3 rows

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
149236	12.06.2007	1	79.95	35.00
...	...	...	...	...

Select the first 3 rows



Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00

```
myData[1:3, ]
```

Row numbers to be selected.
":" generates a sequence for slicing

Sidenote: "1"-based indexing in R

R uses 1-based indexing, i.e. the **first index is 1** (not 0).

R Basics: Use the colon operator (:) to generate a sequence between 2 numbers

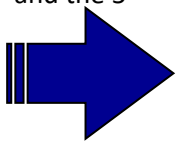
":" generates a regular sequence

```
> 1:5
 1 2 3 4 5
> 5:1
 5 4 3 2 1
> -2:2
 -2 -1  0  1  2
> 1:1
 1
```

Selecting rows by row numbers
Select the first 3 and the 5th row

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
149236	12.06.2007	1	79.95	35.00
…	…	…	…	…

Select the first 3 rows and the 5th



Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	12.06.2007	1	79.95	35.00

```
myData[c(1:3,5), ]
```

Combine integers in a **vector**

R Basics: Understand the dimensions of your data.table

Important functions to determine dimensions:

- Number of rows/columns:

```
nrow(myData)
ncol(myData)
```

- Length of a vector:

```
length(c(1,1))
```

- Length of string:

```
nchar("hello")
```

R Basics: Understand the dimensions of your data.table

Important functions to determine dimensions:

- Number of rows/columns:

```
nrow(myData)
ncol(myData)
```

- Length of a vector:

```
length(c(1,1))
```

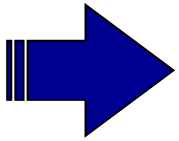
- Length of string:

```
nchar("hello")
```

Selecting rows by row numbers
Select the last row

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
149236	12.06.2007	1	79.95	35.00
...	...	...	...	...
199542	17.09.2012	1	39.95	10.50

Select the last row



Customer	TransDate	Quantity	PurchAmount	Cost
199542	17.09.2012	1	39.95	10.50

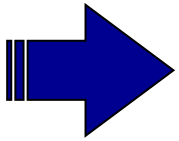
`.N` gives the number of rows and hence selects the last one

```
myData[.N, ]
tail(myData, 1)
myData[nrow(myData),]
```

Selecting rows by row numbers
Select the last row

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
149236	12.06.2007	1	79.95	35.00
...	...	...	...	...
199542	17.09.2012	1	39.95	10.50

Select the last row



Customer	TransDate	Quantity	PurchAmount	Cost
199542	17.09.2012	1	39.95	10.50

`.N` gives the number of rows and hence selects the last one

```
myData[.N, ]
tail(myData, 1)
myData[nrow(myData),]
```

Sidenote: How to sort your data.table

- **To sort your DataFrame according to transaction dates (increasing), use:**

  `myData[order(TransDate)]` or `setkey(myData,TransDate)`

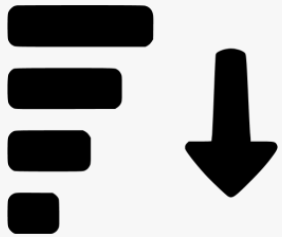
- **Order first according to transaction dates and then according to customers:**

  `myData[order(TransDate, Customer)]` or

  `setkey(myData,TransDate,Customer)`

- **Order decreasing:**

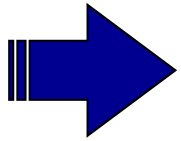
  `myData[order(TransDate, Customer, decreasing = TRUE)]`



Selecting rows by condition
Identify transactions greater than $100

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
149236	12.06.2007	1	79.95	35.00
...	...	...	...	...

Select transactions with value > 100



Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
...	...	...	...	...

```
myData[PurchAmount > 100, ]
```

Select all transactions > $100

R Basics: Logical Operators

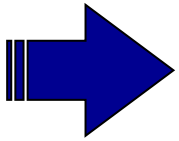
Sign	Description	Example
<	less than	`a < 0`
<=	less than or equal than	`a <= 3`
>	greater than	`a > 0`
>=	greater than or equal than	`a >= 3`
==	equal to	`a == 0`
!=	not equal to	`!= 0`
!	logical negotiation (NOT)	`!x`
&	logical AND	`x & y`
|	logical OR	`x | y`

Selecting rows by condition

Select the transactions of a single customer

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
149236	12.06.2007	1	79.95	35.00
...	...	...	...	...

Select transactions where `Customer` is `149332`



Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
...	...	...	...	...

Note: Variable `Customer` is of type integer (1)

```
myData[Customer == 149332, ]
```

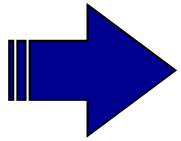
Selects all observations where `Customer` is equal to `149332` (2)

Selecting rows by condition

Select the transactions of multiple customers

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
149236	12.06.2007	1	79.95	35.00
...	...	...	...	...

Select transactions where `Customer` is `149332` or `172951`



Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
...	...	...	...	...
172951	29.08.2008	1	199.95	108.00
...	...	...	...	...

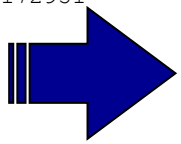
```
myData[Customer %in% c(149332, 172951), ]
```

Selects all observations where `Customer` is either `149332` **or** `172951`

Bang operator (!) precedes %in% to negate the condition

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
149236	12.06.2007	1	79.95	35.00
…	…	…	…	…

Select transactions where `Customer` is **NOT** `149332` or `172951`



Customer	TransDate	Quantity	PurchAmount	Cost
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
149236	12.06.2007	1	79.95	35.00
…	…	…	…	…

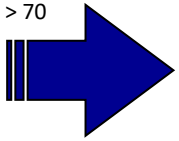
```
myData[!Customer %in% c(149332, 172951), ]
```

"Not in"

Combining conditions

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
149236	12.06.2007	1	79.95	35.00
…	…	…	…	…
140729	28.04.2012	1	89.95	35.00
…	…	…	…	…

Select transactions where date after 24.10.2007 **and** `PurchAmount` > 70



Customer	TransDate	Quantity	PurchAmount	Cost
140729	28.04.2012	1	89.95	35.00
…	…	…	…	…

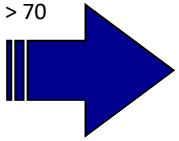
```
myData[TransDate > ymd("2010-12-24") & PurchAmount > 70, ]
```

Combine multiple conditions

Combining conditions

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
149236	12.06.2007	1	79.95	35.00
…	…	…	…	…
140729	28.04.2012	1	89.95	35.00
…	…	…	…	…

Select transactions where date after 24.10.2007 **and** `PurchAmount` > 70



Customer	TransDate	Quantity	PurchAmount	Cost
140729	28.04.2012	1	89.95	35.00
…	…	…	…	…

```
myData[TransDate > ymd("2010-12-24") & PurchAmount > 70, ]
```

Combine multiple conditions