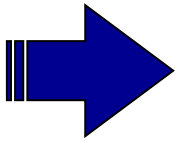
Combining select and aggregate operations

Select the first 3 purchases of each customer

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
172951	29.08.2008	1	249.95	162.50
172951	29.08.2008	1	39.95	18.95
172951	28.01.2009	1	79.95	35.00
…	…	…	…	…

Select the first 3 purchases of each customer



Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
149236	12.06.2007	1	79.95	35.00
172951	29.08.2008	1	199.95	108.00
172951	29.08.2008	1	249.95	162.50
172951	29.08.2008	1	39.95	18.95
…	…	…	…	…

1 Order the data by `TransDate`

```
setkey(myData,TransDate)

myData[, head(.SD,3), by=Customer]
```

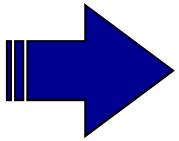
2 `.SD` – Subset of data.table

3 Select the first three entries of every customer

Select the first 3 purchases of each customer

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
172951	29.08.2008	1	249.95	162.50
172951	29.08.2008	1	39.95	18.95
172951	28.01.2009	1	79.95	35.00
…	…	…	…	…

Select the first 3 purchases of each customer



Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
149236	12.06.2007	1	79.95	35.00
172951	29.08.2008	1	199.95	108.00
172951	29.08.2008	1	249.95	162.50
172951	29.08.2008	1	39.95	18.95
…	…	…	…	…

1 Order the data by `TransDate`

```
setkey(myData,TransDate)

myData[, head(.SD,3), by=Customer]
```

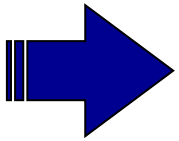
2 `.SD` – Subset of data.table

3 Select the first three entries of every customer

Select the first 3 purchases of each customer

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
172951	29.08.2008	1	249.95	162.50
172951	29.08.2008	1	39.95	18.95
172951	28.01.2009	1	79.95	35.00
…	…	…	…	…

Select the first 3 purchases of each customer



Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
149236	12.06.2007	1	79.95	35.00
172951	29.08.2008	1	199.95	108.00
172951	29.08.2008	1	249.95	162.50
172951	29.08.2008	1	39.95	18.95
…	…	…	…	…

1. Order the data by `TransDate`

```
setkey(myData,TransDate)
myData[, head(.SD,3), by=Customer]
```

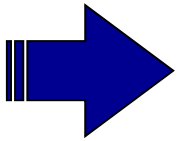
2. `.SD` – Subset of data.table

3. Select the first three entries of every customer

Select the last purchase of each customer

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
172951	29.08.2008	1	249.95	162.50
172951	29.08.2008	1	39.95	18.95
172951	28.01.2009	1	79.95	35.00
…	…	…	…	…

Select the last purchase of each customer



Customer	TransDate	Quantity	PurchAmount	Cost
172951	28.01.2009	1	79.95	35.00
…	…	…	…	…

1 `.N` counts the number of rows per customer

```
myData[, .SD[.N], by=Customer]
```

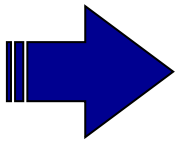
2 Two options to select the last purchase for each customer using either subset or tail.

```
myData[, tail(.SD,1), by=Customer]
```

Updating columns using an aggregating dimension (1/2)

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
...	...	...	...	...

Add a column counting the quantities per `Customer`



Customer	TransDate	...	Count
149332	15.11.2005		1
172951	29.08.2008		1
120621	19.10.2007		1
149236	14.11.2005		2
149236	12.06.2007		2
...	...	...	...

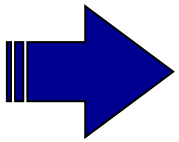
```
myData[, Count := .N, by=Customer]
```

`.N` counts the number of rows per customer

Updating columns using an aggregating dimension (2/2)

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
…	…	…	…	…

Add a column counting the transactions per `Customer`



Customer	TransDate	…	Count
149332	15.11.2005		1
172951	29.08.2008		1
120621	19.10.2007		1
149236	14.11.2005		2
149236	12.06.2007		2
…	…	…	…

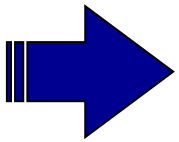
```
myData[, relDate := 1:.N, by=Customer]
```

Save output as variable `relDate`

Updating columns using an aggregating dimension (2/2)

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
...	...	...	...	...

Add a column counting the transactions per `Customer`



Customer	TransDate	...	Count
149332	15.11.2005		1
172951	29.08.2008		1
120621	19.10.2007		1
149236	14.11.2005		2
149236	12.06.2007		2
...	...	...	...

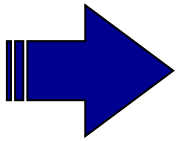
```
myData[, relDate := 1:.N, by=Customer]
```

Save output as variable `relDate`

Creating a lagged variable

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
...	...	...	...	...

Add a column with lagged cost by one observation aggregated by `Customer`



Customer	TransDate	...	Cost	CostLag
149332	15.11.2005	1	107.00	NA
172951	29.08.2008		108.00	NA
120621	19.10.2007		49.00	NA
149236	14.11.2005		18.95	NA
149236	12.06.2007		35.00	18.95
...	...	...	...	...

Only works for observation two and further (2)

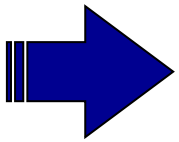
```
myData[, CostLag := shift(Cost), by=Customer]
```

Lag (1)

Cumulating variables

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
...	...	...	...	...

Add a column with cumulated cost aggregated by `Customer`



Customer	TransDate	...	PurchAmount	totSpend
149332	15.11.2005		199.95	199.95
172951	29.08.2008		199.95	199.95
120621	19.10.2007		99.95	99.95
149236	14.11.2005		39.95	39.95
149236	12.06.2007		79.95	119.9
...	...	...	...	...

```
myData[, totSpend := cumsum(PurchAmount), by=Customer]
```

Cumulates the `PurchAmount` variable