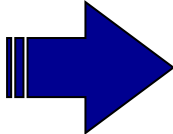


**Advanced techniques for  
aggregating observations**

# Aggregate a variable by a transformed aggregating dimension

Customer	TransDate	Quantity	PurchAmount	Cost	Sum PurchAmount by month of each year	Date	AggPurch
149332	15.11.2005	1	199.95	107.00		2005-11-01	313080.30
172951	29.08.2008	1	199.95	108.00		2008-08-01	197361.00
120621	19.10.2007	1	99.95	49.00		2007-10-01	268155.63
149236	14.11.2005	1	39.95	18.95		...	...
149236	12.06.2007	1	79.95	35.00			
...	...	...	...	...			

```
myData[, list(AggPurch=sum(PurchAmount)),
         by=list(Date=floor_date(TransDate,
                                unit="month"))]
```

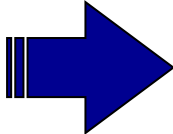
Use month to get a  
monthly summary

②

Command to set the  
date to the first of the  
month (lubridate  
package)

①

# Aggregate a variable by a transformed aggregating dimension

Customer	TransDate	Quantity	PurchAmount	Cost	Sum PurchAmount by month of each year	Date	AggPurch
149332	15.11.2005	1	199.95	107.00		2005-11-01	313080.30
172951	29.08.2008	1	199.95	108.00		2008-08-01	197361.00
120621	19.10.2007	1	99.95	49.00		2007-10-01	268155.63
149236	14.11.2005	1	39.95	18.95		...	...
149236	12.06.2007	1	79.95	35.00			
...	...	...	...	...			

```
myData[, list(AggPurch=sum(PurchAmount)),
        by=list(Date=floor_date(TransDate,
                                unit="month"))]
```

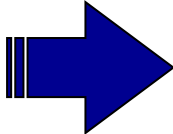
Use month to get a  
monthly summary

②

Command to set the  
date to the first of the  
month (lubridate  
package)

①

# Aggregate a variable by a transformed aggregating dimension

Customer	TransDate	Quantity	PurchAmount	Cost	Sum PurchAmount by month of each year	Date	AggPurch
149332	15.11.2005	1	199.95	107.00		2005-11-01	313080.30
172951	29.08.2008	1	199.95	108.00		2008-08-01	197361.00
120621	19.10.2007	1	99.95	49.00		2007-10-01	268155.63
149236	14.11.2005	1	39.95	18.95		...	...
149236	12.06.2007	1	79.95	35.00			
...	...	...	...	...			

```
myData[, list(AggPurch=sum(PurchAmount)),
  by=list(Date=floor_date(TransDate,
unit="month"))]
```

Use month to get a  
monthly summary

②

Command to set the  
date to the first of the  
month (lubridate  
package)

①

## Sidenote: Chaining saves memory and is faster

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

Sum  
PurchAmount by  
Customer and  
select customers  
with aggregated  
sums greater than  
100

Customer	AggPurch
149332	274.85
172951	889.80

Customer	AggPurch
149332	199.95
172951	199.95

```
myData[, list(AggPurch=sum(PurchAmount)), by=Customer][PurchAmount > 100]
```

Order is important: here, selection is done first and then aggregation on the selected customers only.

**Not the same as:** `myData_agg <- myData[, list(AggPurch=sum(PurchAmount)),by=Customer]`  
`myData_agg[PurchAmount > 100, ]`

## Sidenote: Chaining saves memory and is faster

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

Sum  
PurchAmount by  
Customer and  
select customers  
with aggregated  
sums greater than  
100

Customer	AggPurch
149332	274.85
172951	889.80

Customer	AggPurch
149332	199.95
172951	199.95

```
myData[, list(AggPurch=sum(PurchAmount)), by=Customer][PurchAmount > 100]
```

Order is important: here, selection is done first and then aggregation on the selected customers only.

**Not the same as:** `myData_agg <- myData[, list(AggPurch=sum(PurchAmount)),by=Customer]`  
`myData_agg[PurchAmount > 100, ]`

## Sidenote: Chaining saves memory and is faster

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

Sum  
PurchAmount by  
Customer and  
select customers  
with aggregated  
sums greater than  
100

Customer	AggPurch
149332	274.85
172951	889.80

Customer	AggPurch
149332	199.95
172951	199.95

```
myData[, list(AggPurch=sum(PurchAmount)), by=Customer][PurchAmount > 100]
```

Order is important: here, selection is done first and then aggregation on the selected customers only.

Not the same as: `myData_agg <- myData[, list(AggPurch=sum(PurchAmount)), by=Customer]`  
`myData_agg[PurchAmount > 100, ]`

## Sidenote: Pay attention to operation sequences

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

Sum  
PurchAmount by  
Customer and  
select customers  
with aggregated  
sums greater than  
100

Customer	AggPurch
149332	274.85
172951	889.80

Customer	AggPurch
149332	199.95
172951	199.95

`myData[, list(AggPurch=sum(PurchAmount)), by=Customer][PurchAmount > 100]`



Not the same as: `myData[PurchAmount > 100, list(AggPurch=sum(PurchAmount)), by=Customer]`