Basic techniques for aggregating observations

By aggregating data, we can answer the following questions

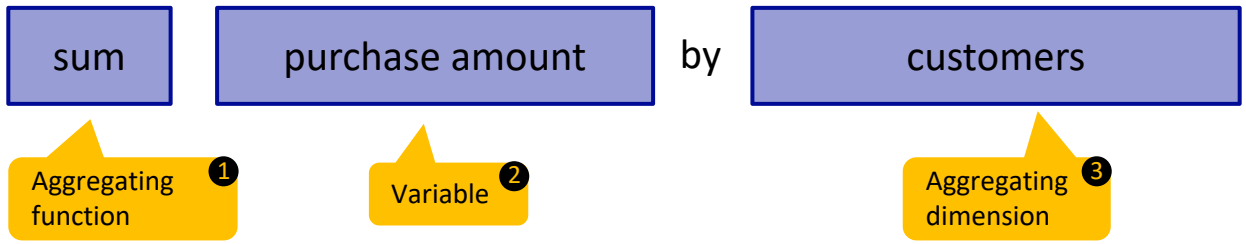
- What is the total value of purchases in 2015?
- How much has each customer spent per month?
- What is the average age of all male customers?

Aggregating means:
"do <<function>> to <<variable>> by <<dimension>>"

Aggregating has 2 components:

- Function and variable by which to aggregate.
- Dimension by which to aggregate.

For example:

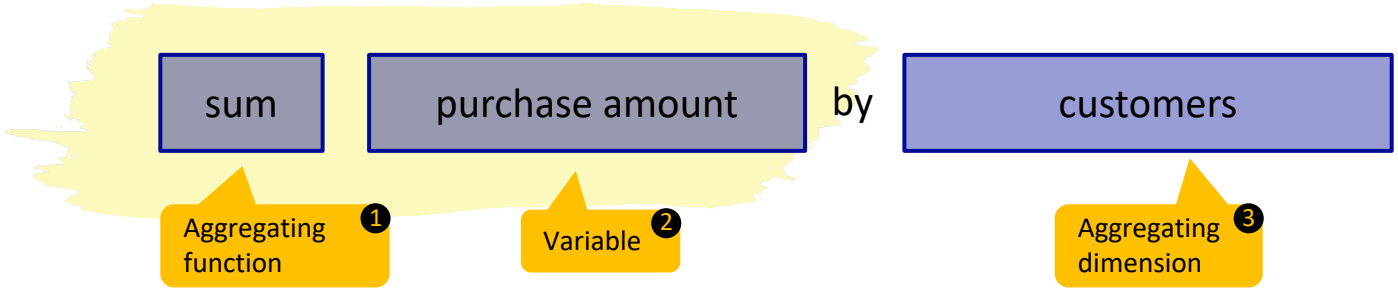


Aggregating means:
"do <<function>> to <<variable>> by <<dimension>>"

Aggregating has 2 components:

- Function and variable by which to aggregate.
- Dimension by which to aggregate.

For example:

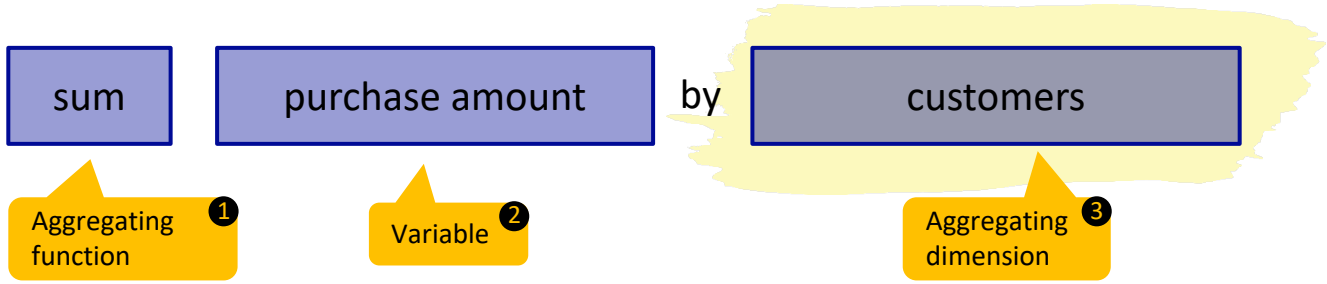


Aggregating means:
"do <<function>> to <<variable>> by <<dimension>>"

Aggregating has 2 components:

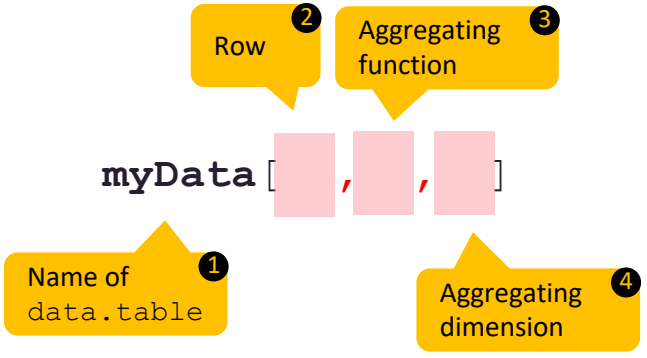
- Function and variable by which to aggregate.
- Dimension by which to aggregate.

For example:



General command structure for aggregating data.table objects on one dimension

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 aggregating

1. Apply an aggregating function to **a variable** by an aggregating dimension.
2. Apply an aggregating function to **multiple variables** by an aggregating dimension.
3. Apply **multiple aggregating functions** to a variable by an aggregating dimension.
4. Apply an aggregating function to a variable by **multiple aggregating dimensions**.
5. Apply an aggregating function to a variable by an aggregating dimension to a **selection of rows**.
6. Apply an aggregating function to the **whole dataset**.

1. Apply an aggregating function to a variable by an aggregating dimension (1/2)

Option 1: `sum( )` with direct aggregation procedure

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`

Customer	V1
149332	274.85
172951	889.80
120621	99.95
149236	119.90
...	...

5 Now summed

1 Name of `data.table`

3 Variable

```
myData[, sum(PurchAmount), by=Customer]
```

2 Aggregating function `sum()`

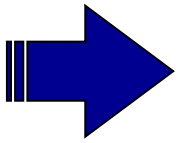
4 Aggregating dimension

1. Apply an aggregating function to a variable by an aggregating dimension (1/2)

Option 1: `sum( )` with direct aggregation procedure

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`



Customer	V1
149332	274.85
172951	889.80
120621	99.95
149236	119.90
...	...

5 Now summed

1 Name of `data.table`

3 Variable

```
myData[, sum(PurchAmount), by=Customer]
```

2 Aggregating function `sum()`

4 Aggregating dimension

1. Apply an aggregating function to a variable by an aggregating dimension (1/2)

Option 1: `sum( )` with direct aggregation procedure

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`

Customer	V1
149332	274.85
172951	889.80
120621	99.95
149236	119.90
...	...

5 Now summed

1 Name of `data.table`

3 Variable

```
myData[, sum(PurchAmount), by=Customer]
```

2 Aggregating function `sum()`

4 Aggregating dimension

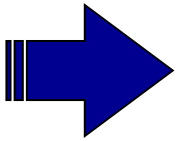
1. Apply an aggregating function to a variable by an aggregating dimension (2/2)

Option 2: `list( ) including renaming`

Using `list( )` enables multiple and different aggregation functions on columns (see later).

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 rename it `AggPurch`



Customer	AggPurch
149332	274.85
172951	889.80
120621	99.95
149236	119.90
...	...

3 Name of new column

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

1 Name of new column

2 Aggregating dimension

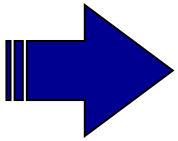
1. Apply an aggregating function to a variable by an aggregating dimension (2/2)

Option 2: `list( ) including renaming`

Using `list( )` enables multiple and different aggregation functions on columns (see later).

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 rename it `AggPurch`



Customer	AggPurch
149332	274.85
172951	889.80
120621	99.95
149236	119.90
...	...

3 Name of new column

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

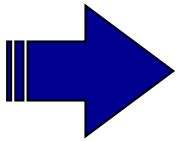
1 Name of new column

2 Aggregating dimension

2. Apply an aggregating function to multiple variables by an aggregating dimension

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` and `Quantity` by `Customer`



Customer	AggPurch	AggQuant
149332	199.95	1
172951	199.95	1
120621	99.95	1
149236	119.90	2
...	...	

First variable 1

Second variable 2

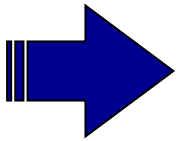
```
myData[, list(AggPurch=sum(PurchAmount), AggQuant=sum(Quantity)),
          by=Customer]
```

Aggregating dimension 3

2. Apply an aggregating function to multiple variables by an aggregating dimension

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` and `Quantity` by `Customer`



Customer	AggPurch	AggQuant
149332	199.95	1
172951	199.95	1
120621	99.95	1
149236	119.90	2
...	...	

First variable 1

Second variable 2

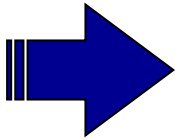
```
myData[, list(AggPurch=sum(PurchAmount), AggQuant=sum(Quantity)),
          by=Customer]
```

Aggregating dimension 3

3. Apply multiple aggregating functions to a variable by an aggregating dimension

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 and select max of `PurchAmount` by `Customer`



Customer	AggPurch	Purch_max
149332	274.85	199.95
172951	889.8	349.95
120621	99.95	99.95
149236	119.9	79.95
...	...	...
199542	39.95	39.95

1 Aggregated function `sum()`

```
myData[, list(AggPurch=sum(PurchAmount),
        Purch_Max=max(PurchAmount)), by=Customer]
```

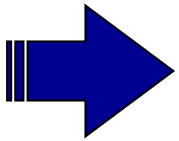
2 Aggregated function `max()`

3 Aggregating dimension

4. Apply an aggregating function to a variable by multiple aggregating dimensions

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` aggregated by `Customer` and `TransDate`



Customer	TransDate	PurchAmount
149332	15.11.2005	199.95
172951	29.08.2008	799.85
120621	19.10.2007	99.95
149236	14.11.2005	39.95
149236	12.06.2007	79.95
…	…	…

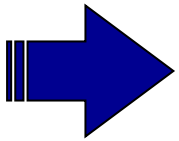
Multiple aggregating dimensions

```
myData[, list(PurchAmount=sum(PurchAmount)), by=list(Customer, TransDate)]
```

4. Apply an aggregating function to a variable by multiple aggregating dimensions

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` aggregated by `Customer` and `TransDate`



Customer	TransDate	PurchAmount
149332	15.11.2005	199.95
172951	29.08.2008	799.85
120621	19.10.2007	99.95
149236	14.11.2005	39.95
149236	12.06.2007	79.95
...	...	...

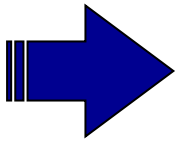
Multiple aggregating dimensions

```
myData[, list(PurchAmount=sum(PurchAmount)), by=list(Customer,
TransDate)]
```

5. Apply an aggregating function to a variable by an aggregating dimension to a selection of 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 rows 2 to 5 and sum `PurchAmount` by `Customer`



Customer	AggPurch
172951	199.95
120621	99.95
149236	119.90

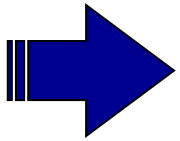
```
myData[2:5, list(AggPurch=sum(PurchAmount)), by=Customer]
```

Select rows

6. Apply an aggregating function to the whole dataset

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 all entries of `PurchAmount`



```
18784785
```

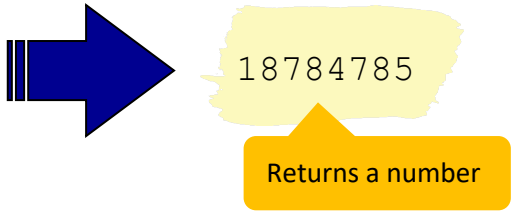
Returns a number

```
myData[, sum(PurchAmount)]
```

6. Apply an aggregating function to the whole dataset

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

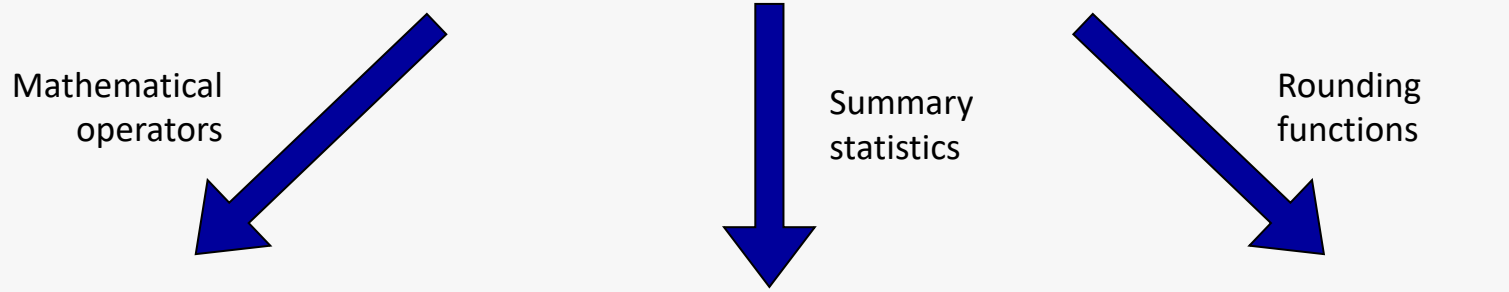
Sum all entries of `PurchAmount`




```
myData[, sum(PurchAmount)]
```

R Basics: a short list of aggregating functions

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



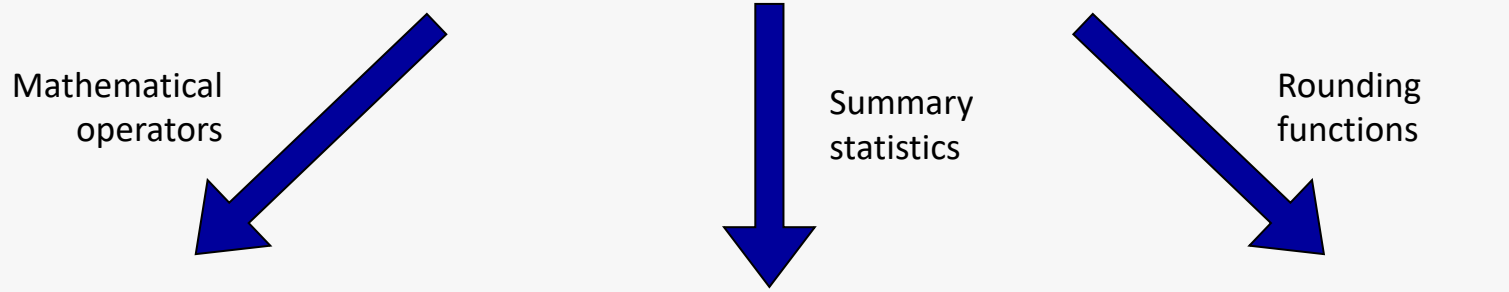

```
sum( )
min( )
max( )
```

```
mean(x, )
median(x, )
sd(x, )
```

```
round(x)
floor(x)
ceiling(x)
```

R Basics: a short list of aggregating functions

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( )
min( )
max( )
```

```
mean(x, )
median(x, )
sd(x, )
```

```
round(x)
floor(x)
ceiling(x)
```

Sidenote: Create new columns in the original data.table with " :="

```
myData[, AggPurch := sum(PurchAmount)]
```

Customer	TransDate	Quantity	PurchAmount	Cost	AggPurch
149332	15.11.2005	1	199.95	107.00	18784785
172951	29.08.2008	1	199.95	108.00	18784785
120621	19.10.2007	1	99.95	49.00	18784785
149236	14.11.2005	1	39.95	18.95	18784785
149236	12.06.2007	1	79.95	35.00	18784785
...	...	...	...	...	...

1 Creates a new column in `myData`

```
myData[, list(AggPurch = sum(PurchAmount))]
```

AggPurch
18784785

2 Summarizes the `myData` without making any changes to `myData`

Sidenote: Create new columns in the original data.table with " :="

```
myData[, AggPurch := sum(PurchAmount)]
```

Customer	TransDate	Quantity	PurchAmount	Cost	AggPurch
149332	15.11.2005	1	199.95	107.00	18784785
172951	29.08.2008	1	199.95	108.00	18784785
120621	19.10.2007	1	99.95	49.00	18784785
149236	14.11.2005	1	39.95	18.95	18784785
149236	12.06.2007	1	79.95	35.00	18784785
...	...	...	...	...	...

1 Creates a new column in `myData`

```
myData[, list(AggPurch = sum(PurchAmount))]
```

AggPurch
18784785

2 Summarizes the `myData` without making any changes to `myData`