SQL: Structured Query Language

Part 3

GROUP BY and HAVING

- So far, we've applied aggregate operators to all (qualifying) tuples. Sometimes, we want to apply them to each of several *groups* of tuples.
- Consider: ***Find the age of the youngest sailor for each rating level.***
  - In general, we don't know how many rating levels exist, and what the rating values for these levels are!
  - Suppose we know that rating values go from 1 to 10; we can write 10 queries that look like this (!):

For $i = 1, 2, \ldots, 10$:

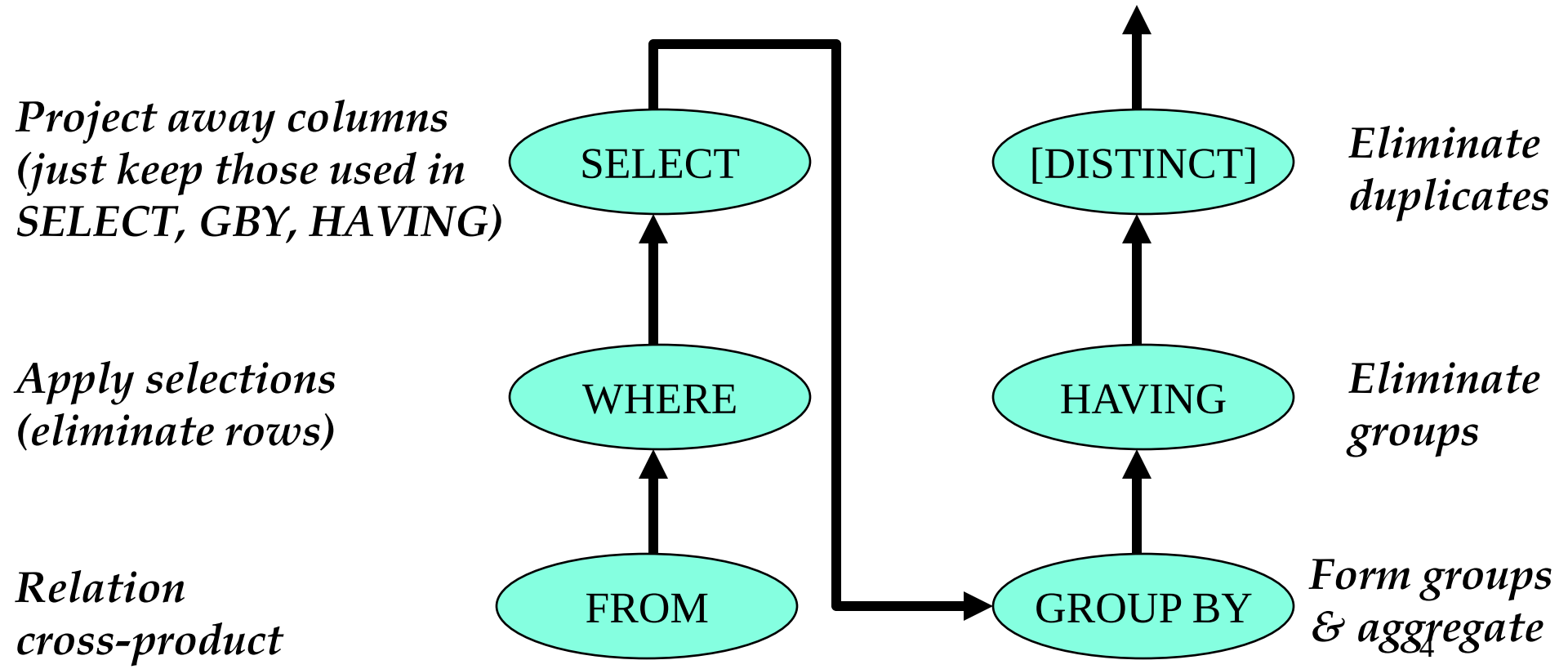
```
SELECT  MIN (S.age)
FROM  Sailors S
WHERE  S.rating = i
```

Find the age of the youngest sailor for each rating level with at least 2 sailors

```
SELECT  S.rating, MIN (S.age)
FROM  Sailors S
GROUP BY  S.rating
HAVING COUNT(*) >1
```

Conceptual Evaluation

SELECT [DISTINCT] *target-list*
FROM *relation-list*
WHERE *qualification*
GROUP BY *grouping-list*
HAVING *group-qualification*



Find the age of the youngest sailor for each rating level with at least 2 sailors

Find the age of the youngest sailor for each rating level with at least 2 sailors

```
SELECT  S.rating, MIN (S.age)
FROM  Sailors S
GROUP BY  S.rating
HAVING COUNT(*) >1
```

Find the age of the youngest sailor for each rating level with at least 2 sailors

```
SELECT  S.rating, MIN (S.age)
FROM  Sailors S
GROUP BY  S.rating
HAVING COUNT(*) >1
```

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0

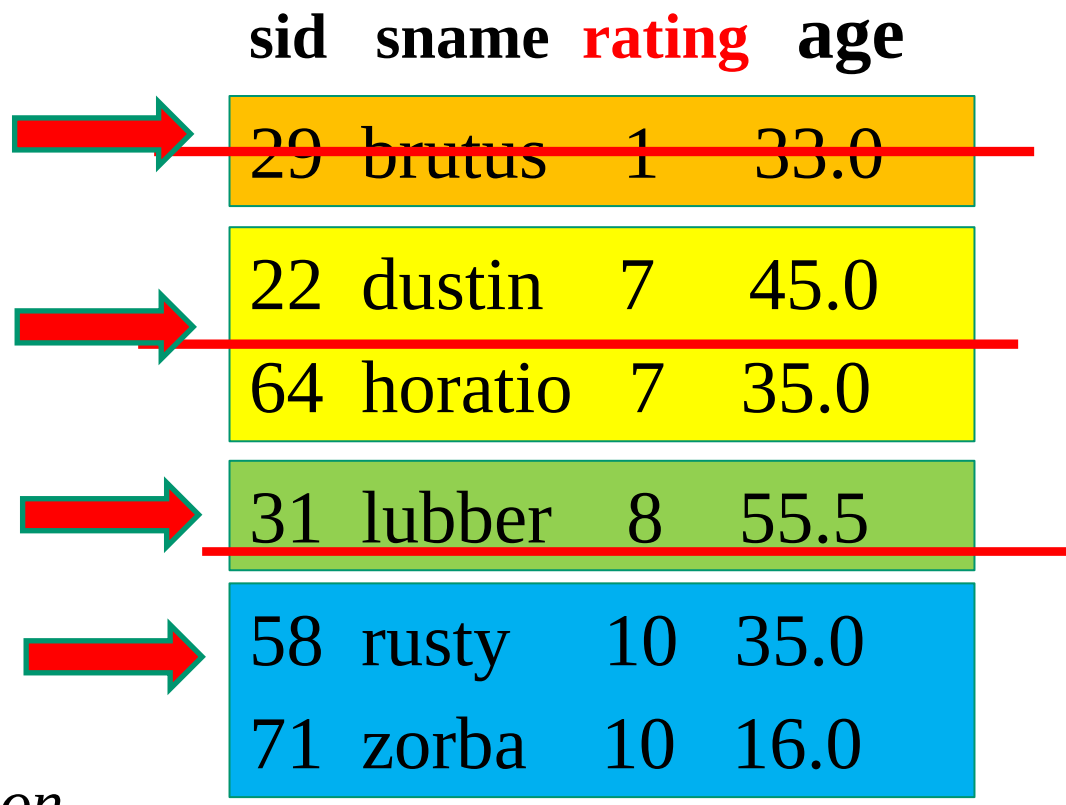
Expressions in *group-qualification* must have a *single value per group*!

One answer tuple is generated per qualifying group.

How about this change on the query?

```
SELECT  S.rating, MIN (S.age)
FROM  Sailors S
GROUP BY  S.rating
HAVING S.rating > 8
```

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0



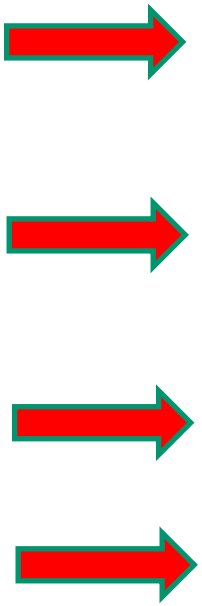
Expressions in *group-qualification* must have a *single value per group*!

One answer tuple is generated per qualifying group.

How about this change on the query?

```
SELECT  S.rating, MIN (S.age)
FROM  Sailors S
GROUP BY  S.rating
HAVING S.age > 30
```

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0



sid	sname	rating	age
29	brutus	1	33.0
22	dustin	7	45.0
64	horatio	7	35.0
31	lubber	8	55.5
58	rusty	10	35.0
71	zorba	10	16.0

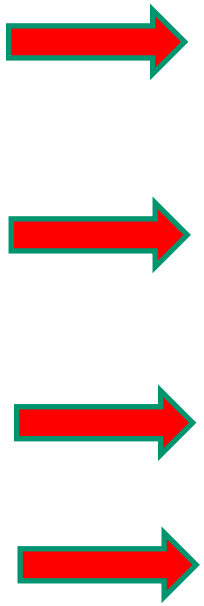
Expressions in *group-qualification* must have a *single value per group*!

One answer tuple is generated per qualifying group.

How about this change on the query?

SELECT S.rating, **MIN (S.age)**
FROM Sailors S
GROUP BY S.rating
~~HAVING S.age > 30~~

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0



sid	sname	rating	age
29	brutus	1	33.0
22	dustin	7	45.0
64	horatio	7	35.0
31	lubber	8	55.5
58	rusty	10	35.0
71	zorba	10	16.0

Expressions in *group-qualification* must have a *single value per group*!

One answer tuple is generated per qualifying group.

Find the age of the youngest sailor for each rating level with at least 2 sailors

```
SELECT  S.rating, MIN (S.age)
FROM  Sailors S
GROUP BY  S.rating
HAVING COUNT(*) >1
```

Equivalently…

```
SELECT  S.rating, MIN (S.age)
FROM  Sailors S
GROUP BY  S.rating
HAVING 1 < (SELECT  COUNT (*)
            FROM  Sailors S2
            WHERE  S.rating=S2.rating)
```

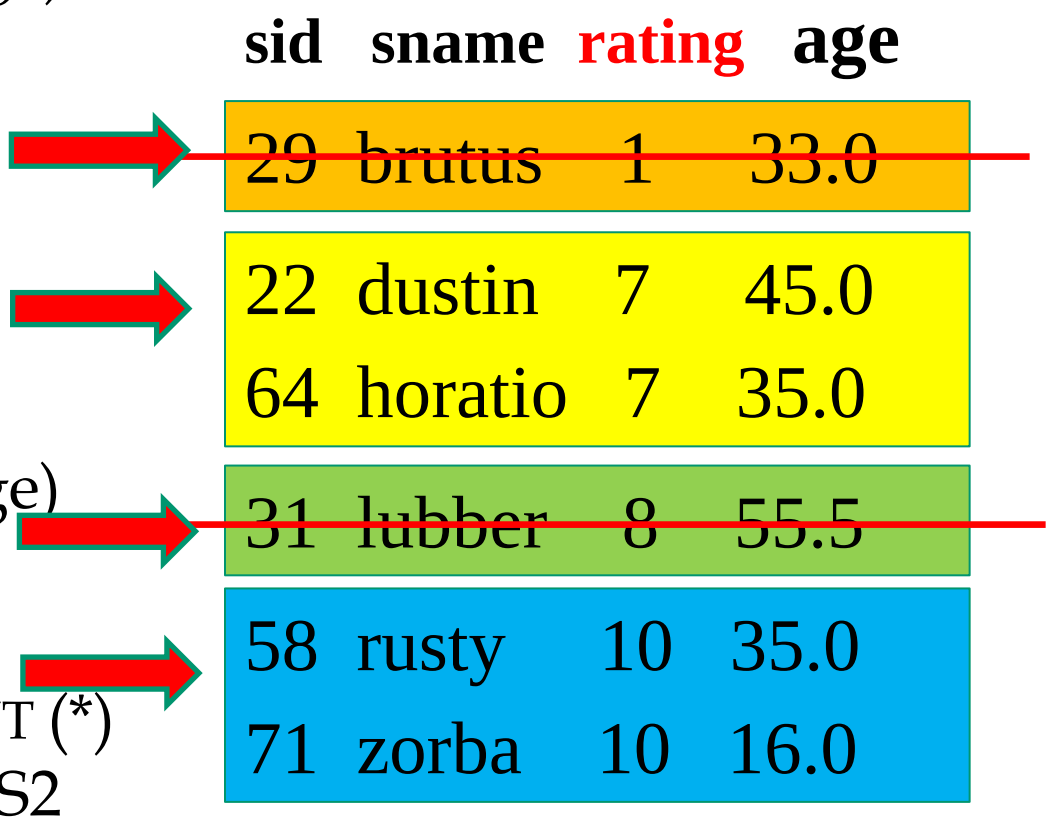
- Shows HAVING clause can also contain a subquery.

Find the age of the youngest sailor for each rating level with at least 2 sailors

```
SELECT  S.rating, MIN (S.age)
FROM  Sailors S
GROUP BY  S.rating
HAVING COUNT(*) >1
```

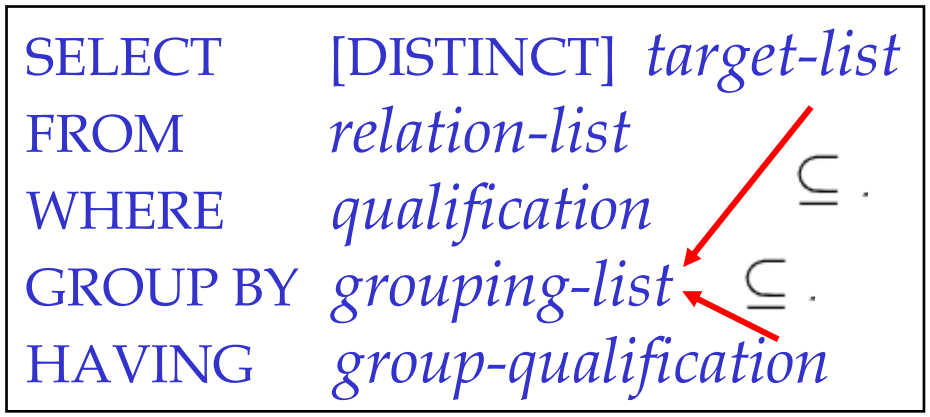
Equivalently…

```
SELECT  S.rating, MIN (S.age)
FROM  Sailors S
GROUP BY  S.rating
HAVING 1 < (SELECT  COUNT (*)
            FROM  Sailors S2
            WHERE  S.rating=S2.rating)
```



- Shows HAVING clause can also contain a subquery.

Queries With GROUP BY and HAVING



- The *target-list* contains (i) attribute names (ii) terms with aggregate operations (e.g., MIN (*S.age*)).
  - The attribute list (i) must be a subset of *grouping-list*. Intuitively, each answer tuple corresponds to a *group*, and these attributes must have **a single value per group**. (A *group* is a set of tuples that have the same value for all attributes in *grouping-list*.)

Conceptual Evaluation

- The cross-product of *relation-list* is computed, tuples that fail *qualification* are discarded, *`unnecessary'* fields are deleted, and the remaining tuples are partitioned into groups by the value of attributes in *grouping-list*.
- The *group-qualification* is then applied to eliminate some groups. Expressions in *group-qualification* must have a *single value per group*!
  - In effect, an attribute in *group-qualification* that is not an argument of an aggregate op also appears in *grouping-list*.
- One answer tuple is generated per qualifying group.

For each red boat, find the number of reservations for this boat

```
SELECT  B.bid,  COUNT (*) AS scount
FROM  Boats B, Reserves R
WHERE  R.bid=B.bid AND B.color='red'
GROUP BY  B.bid
```

- Grouping over a join of two relations.

For each red boat, find the number of reservations for this boat

Reserves

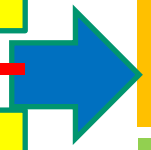
sid	bid	date
1	101	1/1/2017
1	102	15/1/2017
2	108	3/3/2016
1	101	5/6/2016
1	108	4/4/2017

Boats

bid	color
101	red
102	green
108	red

```
SELECT  B.bid,  COUNT (*) AS scount
FROM  Boats B, Reserves R
WHERE  R.bid=B.bid AND B.color='red'
GROUP BY  B.bid
```

sid	R.bid	date	B.bid	color
1	101	...	101	red
1	102		102	green
2	108		108	red
1	101		101	red
1	108		108	red



sid	R.bid		B.bid	color
1	101	…	101	red
1	101	…	101	red
2	108	…	108	red
1	108	…	108	red

For each red boat, find the number of reservations for this boat

```
SELECT  B.bid,  COUNT (*) AS scount
FROM  Boats B, Reserves R
WHERE  R.bid=B.bid AND B.color='red'
GROUP BY  B.bid
```

- Grouping over a join of two relations.
- What do we get if we remove *B.color='red'* from the WHERE clause and add a HAVING clause with this condition?

Find the age of the youngest sailor with age ≥ 18, for each rating with at least 2 sailors (of any age)

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0

Find the age of the youngest sailor with age ≥ 18, for each rating with at least 2 sailors (of any age)

```
SELECT  S.rating,  MIN (S.age)
FROM  Sailors S
WHERE  S.age >= 18
GROUP BY  S.rating
HAVING  1  <  (SELECT  COUNT (*)
               FROM  Sailors S2
               WHERE  S.rating=S2.rating )
```

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0

Find the age of the youngest sailor with age ≥ 18, for each rating with at least 2 sailors (of any age)

```
SELECT  S.rating,  MIN (S.age)
FROM  Sailors S
WHERE  S.age >= 18
GROUP BY  S.rating
```

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0

Rating	age
1	33.0
7	45.5
7	35.0
8	55.5
10	35.0

Find the age of the youngest sailor with age ≥ 18, for each rating with at least 2 sailors (of any age)

```
SELECT  S.rating,  MIN (S.age)
FROM  Sailors S
WHERE  S.age >= 18
GROUP BY  S.rating
HAVING  1  <  (SELECT  COUNT (*)
               FROM  Sailors S2
               WHERE  S.rating=S2.rating )
```

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0

Rating	age
1	33.0
7	45.5
7	35.0
8	55.5
10	35.0

- Shows HAVING clause can also contain a subquery.

Find the age of the youngest sailor with age ≥ 18, for each rating with at least 2 sailors (of any age)

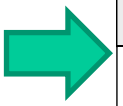
sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0

```
SELECT  S.rating,  MIN (S.age)
FROM  Sailors S
WHERE  S.age >= 18
GROUP BY  S.rating
HAVING  1  <  (SELECT  COUNT (*)
               FROM  Sailors S2
               WHERE  S.rating=S2.rating )
```

Rating	age
1	33.0
7	45.5
7	35.0
8	55.5
10	35.0

For rating=10
Count(*)
2

For rating=7
Count(*)
2



rating	
7	35.0
10	35.0

- Shows HAVING clause can also contain a subquery.

Find the age of the youngest sailor with age ≥ 18, for each rating with at least 2 such sailors

```
SELECT  S.rating,  MIN (S.age)
FROM  Sailors S
WHERE  S.age >= 18
GROUP BY  S.rating
HAVING  COUNT (*) > 1
```

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0

Find the age of the youngest sailor with age ≥ 18, for each rating with at least 2 such sailors

```
SELECT  S.rating,  MIN (S.age)
FROM  Sailors S
WHERE  S.age >= 18
GROUP BY  S.rating
HAVING  COUNT (*) > 1
```

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0

- Only S.rating and S.age are mentioned in the SELECT, GROUP BY or HAVING clauses; other attributes `*unnecessary*'.

Rating	age
7	45.0
8	55.5
7	35.0
1	33.0
10	35.0

Rating	age
1	33.0
7	45.5
7	35.0
8	55.5
10	35.0

rating	
7	35.0

Answer relation

Find the age of the youngest sailor with age ≥ 18, for each rating with at least 2 such sailors

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0

Find the age of the youngest sailor with age ≥ 18, for each rating with at least 2 such sailors

SELECT Temp.rating, Temp.minage
FROM

WHERE Temp.rcount > 1

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0

Assume a TEMP storing number & min age of sailors >=18 per rating

rating	rcount	minage
1	1	33.0
7	2	35.0
8	1	55.5
10	1	35.0

Find the age of the youngest sailor with age ≥ 18, for each rating with at least 2 such sailors

```
SELECT  Temp.rating,  Temp.minage
FROM
   (SELECT S.rating, COUNT(*) AS rcount, MIN(S.age) AS minage
    FROM Sailors S
    WHERE  S.age >= 18
    GROUP BY  S.rating) AS Temp
WHERE Temp.rcount > 1
```

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0

Assume a TEMP storing number & min age of sailors >=18 per rating

rating	rcount	minage
1	1	33.0
7	2	35.0
8	1	55.5
10	1	35.0

RECALL: Find name and age of the oldest sailor(s)

```
SELECT  S.sname, S.age
FROM  Sailors S
WHERE  S.age =
          (SELECT  MAX (S2.age)
           FROM  Sailors S2)
```

Find those ratings for which the average age is the minimum over all ratings

Find those ratings for which the average age is the minimum over all ratings

```
SELECT S.rating
FROM Sailors S
WHERE S.age = (SELECT MIN (AVG (S2.age)) FROM Sailors S2)
```

???

Find those ratings for which the average age is the minimum over all ratings

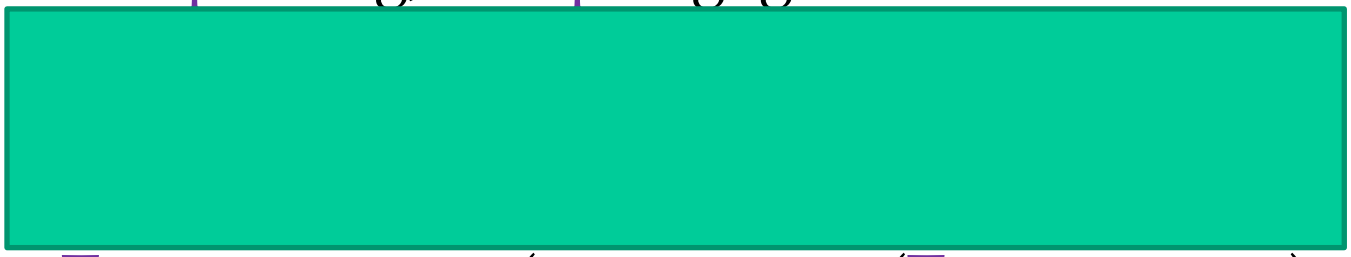
```
SELECT S.rating
FROM Sailors S
WHERE S.age = (SELECT MIN (AVG (S2.age)) FROM Sailors S2)
```

- Aggregate operations **cannot be nested**! WRONG:

Find those ratings for which the average age is the minimum over all ratings

- Correct solution (in SQL/92):

```
SELECT  Temp.rating, Temp.avgage
FROM
WHERE  Temp.avgage = (SELECT  MIN (Temp.avgage)
                      FROM  Temp)
```



OK to define once on pen&paper

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0

rating	avgage
1	33.0
7	40.0
8	55.5
10	25.5

Find those ratings for which the average age is the minimum over all ratings

- Correct solution (in SQL/92):

```
SELECT  Temp.rating, Temp.avgage
FROM  (SELECT  S.rating, AVG (S.age) AS avgage
        FROM  Sailors S
        GROUP BY  S.rating) AS Temp
WHERE  Temp.avgage = (SELECT  MIN (Temp.avgage)
                      FROM  Temp)
```

OK to define once on pen&paper

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
71	zorba	10	16.0
64	horatio	7	35.0
29	brutus	1	33.0
58	rusty	10	35.0

rating	avgage
1	33.0
7	40.0
8	55.5
10	25.5

Find those ratings for which the average age is the minimum over all ratings

- How about….?

```
SELECT  Temp.rating, MIN(Temp.avgage)
FROM  (SELECT  S.rating, AVG (S.age) AS avgage
          FROM  Sailors S
          GROUP BY  S.rating) AS Temp
GROUP BY Temp.rating
```

Temp

rating	avgage
1	33.0
7	40.0
8	55.5
10	25.5

Find those ratings for which the average age is the minimum over all ratings

- How about….?

```
SELECT  Temp.rating, MIN(Temp.avgage)
FROM  (SELECT  S.rating, AVG (S.age) AS avgage
        FROM  Sailors S
        GROUP BY  S.rating) AS Temp
GROUP BY Temp.rating
```

Temp

rating	avgage
1	33.0
7	40.0
8	55.5
10	25.5

If the SELECT clause uses an aggregate operation, then it must use *only* aggregate operations **unless** the query contains GROUP BY clause

```
SELECT  S.sname, MAX (S.age)
FROM  Sailors S
```

Find those ratings for which the average age is the minimum over all ratings

```
SELECT  Temp.rating, Temp.avgage
FROM  (SELECT  S.rating, AVG (S.age) AS avgage
        FROM  Sailors S
        GROUP BY  S.rating) AS Temp
WHERE  Temp.avgage = (SELECT  MIN (Temp.avgage)
                      FROM  Temp)
```

What if you can't use MIN?

Find those ratings for which the average age is the minimum over all ratings

rating	avgage
1	33.0
7	40.0
8	55.5
10	25.5

```
SELECT  Temp.rating, Temp.avgage
FROM  (SELECT  S.rating, AVG (S.age) AS avgage
        FROM  Sailors S
        GROUP BY  S.rating) AS Temp
WHERE  Temp.avgage = (SELECT  MIN (Temp.avgage)
                      FROM  Temp)
```

```
SELECT      S.rating, AVG (S.age) AS avgage
FROM        Sailors S
GROUP BY  S.rating
HAVING    AVG (S.age) <= ALL
                      (SELECT  AVG(S2.age)
                       FROM    Sailors S2
                       GROUP BY S2.rating)
```

Find those ratings for which the average age is the minimum over all ratings

rating	avgage
1	33.0
7	40.0
8	55.5
10	25.5

```
SELECT  Temp.rating, Temp.avgage
FROM  (SELECT  S.rating, AVG (S.age) AS avgage
        FROM  Sailors S
        GROUP BY  S.rating) AS Temp
WHERE  Temp.avgage = (SELECT  MIN (Temp.avgage)
                      FROM  Temp)
```

```
SELECT      S.rating, AVG (S.age) AS avgage
FROM        Sailors S
GROUP BY  S.rating
HAVING    AVG (S.age) <= ALL
                        (SELECT  AVG(S2.age)
                         FROM    Sailors S2
                         GROUP BY S2.rating)
```

33.0
40.0
55.5
25.5

Null Values

- Field values in a tuple are sometimes *unknown* (e.g., a rating has not been assigned) or *inapplicable* (e.g., no spouse's name).
  - SQL provides a special value *null* for such situations.

- The presence of *null* complicates many issues.
  - E.g.: comparison

→Special operators "IS NULL" and "IS NOT NULL".

sid	sname	rating	age
21	Dan	NULL	38

Null Values: 3 valued logic

- Is *rating>8* true or false when *rating* is equal to *null*? What about AND, OR and NOT connectives?
- We need a 3-valued logic (true, false and *unknown*).

sid	sname	rating	age
21	Dan	NULL	38

  - *rating>8 can be* **true**, **false** or **unknown**
  - *rating > 8* OR *age <40 ???*
  - *rating > 8* OR *age >40 ???*
  - *rating > 8* AND *age >40 ???*

- Meaning of constructs must be defined carefully. (e.g., WHERE clause **eliminates rows** that **don't evaluate to true**.)

Null Values: 3 valued logic

- Is *rating>8* true or false when *rating* is equal to *null*? What about AND, OR and NOT connectives?
- We need a 3-valued logic (true, false and *unknown*).

sid	sname	rating	age
21	Dan	NULL	38

  - *rating>8 can be* **true**, **false** or **unknown**
  - *rating > 8* OR *age <40 ???*
  - *rating > 8* OR *age >40 ???*
  - *rating > 8* AND *age >40 ???*

OR TABLE		
T	T	T
T	F	T
F	T	T
T	Unknown	T
Unknown	T	T
F	F	F
F	Unknown	Unknown
Unknown	F	Unknown
Unknown	Unknown	Unknown

- Meaning of constructs must be defined carefully. (e.g., WHERE clause **eliminates rows** that **don't evaluate to true**.)

Null Values: Aggregate Operators

- COUNT(*) → counts also the null values
- All others (as below): discards NULL values

COUNT ([DISTINCT] A)
SUM ([DISTINCT] A)
AVG ([DISTINCT] A)
MAX (A)
MIN (A)

Outer Joins

- Left outer join for S and R:
  - Each S row without a matching R rows appears (once) in the result, with R columns including NULLs
  - SELECT S.sid, R.bid

    FROM Sailors S **NATURAL LEFT OUTER JOIN** Reserves R

Sailors

sid	sname	rating	age
22	Dustin	7	45.0
31	lubber	8	55.5
71	Zorba	10	16.0

Reserves

sid	bid	date
22	101	…
31	102	

sid	bid
22	101
31	102
71	null

Modifying Tables – Insert

- Inserting a single row into a table
  - Attribute list can be omitted if it is the same as in CREATE TABLE
  - NULL and DEFAULT values can be specified

INSERT INTO Transcript(*StudId*, *CrsCode*, *Semester*, *Grade*)
VALUES (12345, 'CSE305', 'S2000', NULL)

Bulk Insertion

- Insert the rows output by a SELECT

```
CREATE TABLE DeansList (
        StudId      INTEGER,
        Credits     INTEGER,
        CumGpa      REAL,
        PRIMARY KEY  (StudId))
```

```
INSERT INTO  DeansList (StudId, Credits, CumGpa)
SELECT       T.StudId, 3 * COUNT (*),  AVG(T.Grade)
FROM         Transcript T
GROUP BY     T.StudId
HAVING       AVG (T.Grade) > 3.5  AND
             COUNT(*) > 30
```

Recall: Transcript(*StudId*, *CrsCode*, *Semester*, *Grade*)

Modifying Tables – Delete

- Similar to SELECT except:
  - No project list in DELETE clause
  - **No Cartesian product** in FROM clause (only 1 table name)
  - Rows satisfying WHERE clause (general form, including subqueries, allowed) are deleted instead of output

```
DELETE FROM Transcript T
WHERE  T.Grade IS NULL  AND  T.Semester <> 'F2020'
```

Modifying Data - Update

```
UPDATE  Employee E
SET     E.Salary = E.Salary * 1.05
WHERE   E.Department = 'R&D'
```

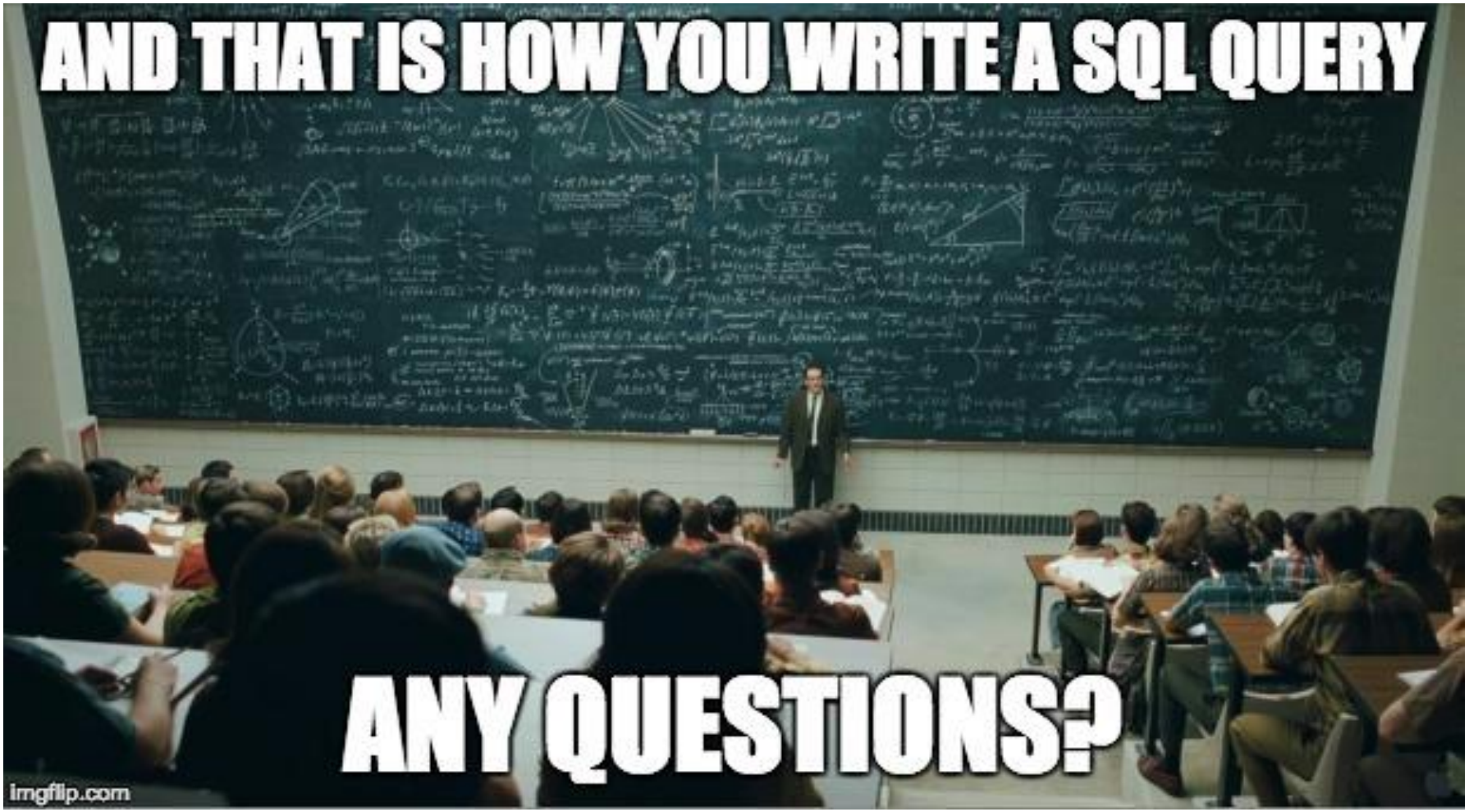
- Updates rows in a **single table**
- All rows satisfying WHERE clause (general form, including subqueries, allowed) are updated

Test first!

```
UPDATE  Employee E
SET     E.Salary = E.Salary * 5
WHERE   E.Salary <=  (SELECT AVG(E2.salary)
                      FROM Employee E2)
```

eid	ename	salary	age
1	Fred	1000	20
2	Jim	2000	39
9	Mike	500	20
4	Mary	3000	17
22	Fred	500	50
3	Nancy	1400	21

Avg: 8400/6 = 1400


AND THAT IS HOW YOU WRITE A SQL QUERY
ANY QUESTIONS?
imgflip.com