SQL Programming

Set Functions

Page A-1: Intro

In this module we discuss one of the most powerful features of the SQL language: Group Functions.

Group functions are functions that apply to groups of rows.

Page B-1 Terminology

Now the ‘more better’ way to refer to these functions is to think of them as set functions, as this is the terminology that is used in the standard.

Set functions is a more precise term and harkens back to the set theoretical underpinnings of relational db theory. But, for now, let’s focus on developing your intuitions, and use the term group functions for just a little while longer.

Page B-2 Terminology

These group functions operate on a group of rows and return a single value (or single row) of results.

This ability to distill and aggregate data has prompted many users to refer to these functions as aggregate functions.

Page B-3 Aggregate Functions

The aggregate functions defined by the standard give the SQL programmer a handy way to gather and display statistics about the data.

Statistic	*Function*
tally	COUNT()
total	SUM()
high value	MAX()
low value	MIN()
average	AVG()

Page C-1 Count

This function is used to answer the question, ’How many?’

The COUNT function accepts one of two parameters, either the name of a column, or the special character ‘*’:

COUNT(*column-name*)

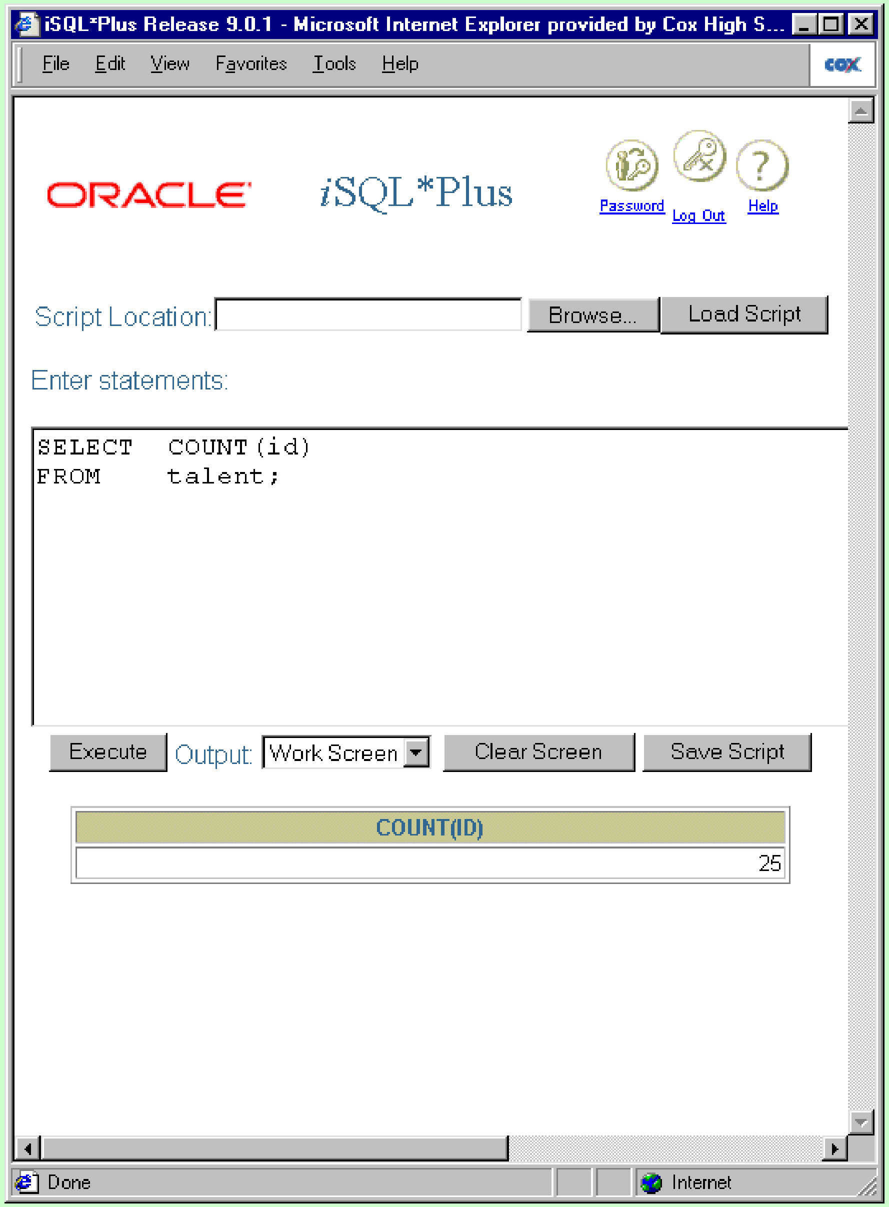
COUNT(*)

Count with the column-name parameter, counts the number of values in the specified column, and in this tally, NULL values are omitted.

Count with the asterisk parameter returns the count, or tally, of the number of rows.

Module 12: Set Functions

Page C-2 Count(*column-name*)



```
SELECT   COUNT(id)
FROM     talent;
```

```
COUNT(ID)
       25
```

Consider the question:

How many clients do we have?

Since the question is asking 'How many', we know that we want to use the count function.

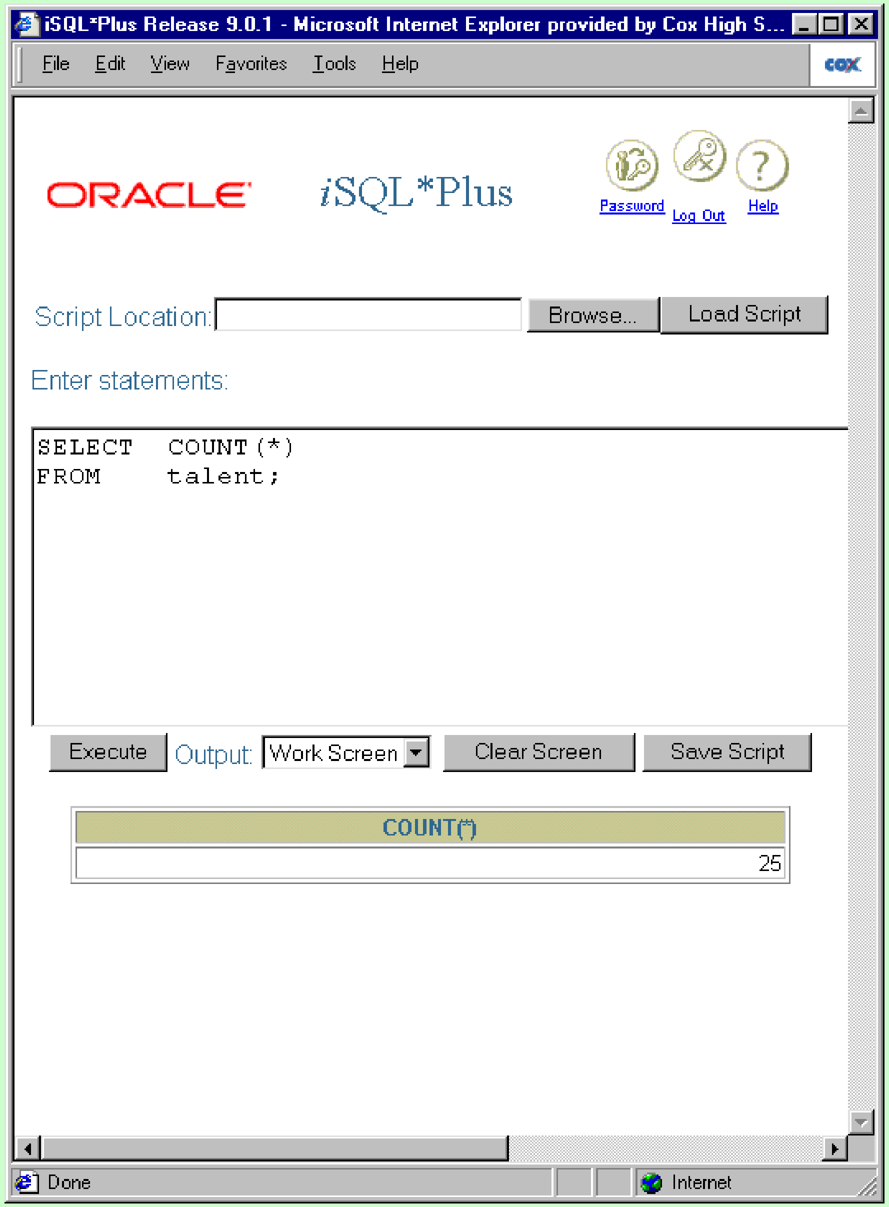
The next question for the programmer to answer then is, which column is it that we should be counting? In this problem, I ask myself, is there any column that I'm sure has a valid value for each and every one of our clients?

Design and code:

```
SELECT   COUNT(id)
FROM     talent;
```

Module 12: Set Functions

Page C-3 Count(*)


iSQL*Plus Release 9.0.1 - Microsoft Internet Explorer provided by Cox High S...
File Edit View Favorites Tools Help
ORACLE iSQL*Plus
Password Log Out Help
Script Location:
Browse... Load Script
Enter statements:
Execute Output: Work Screen Clear Screen Save Script
Done Internet

```
SELECT   COUNT(*)
FROM     talent;
```

COUNT(*)
25

Instead of counting known values in columns, I could have simply counted rows:

Design and code:

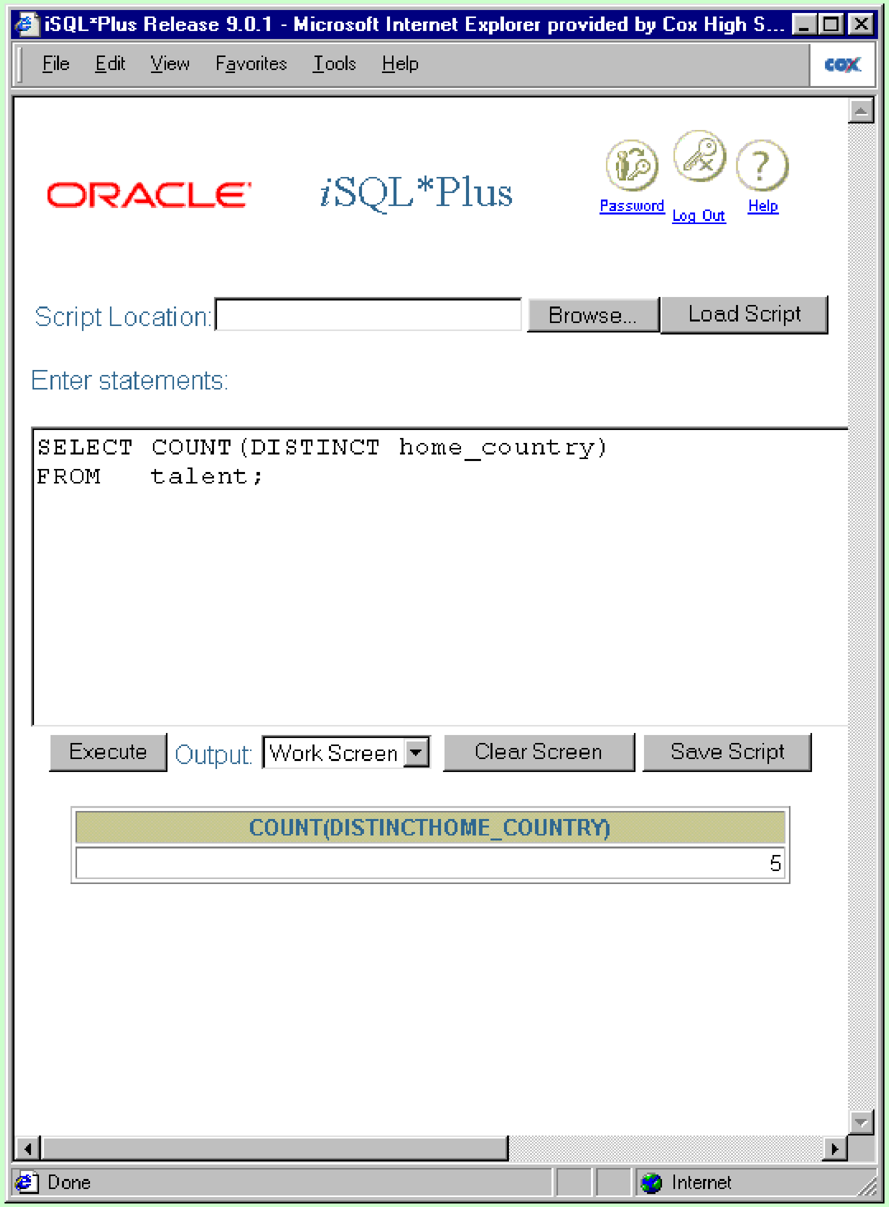
```
SELECT  COUNT(*)
FROM     talent;
```

Let's refine our terminology.

Count(*) counts rows, right?

Count(*column-name*) counts rows in that column with known values (ie. excludes NULLs), right?

Page C-4 Count and DISTINCT



```
SELECT  COUNT(DISTINCT  home_country)
FROM    talent;
```

```
COUNT(DISTINCTHOME_COUNTRY)
                          5
```

Consider this problem.

From how many different countries do we draw our clients?

Design and code:

We've got that magic phrase 'How many' so we know this problem will use the COUNT function.

That word 'different' suggests to me that we need to use the DISTINCT phrase.

Let's try:

```
SELECT COUNT(DISTINCT home_country)
FROM    talent;
```

Page C-4 NULLs and *

The COUNT function is the only function that can account for NULLs (no pun intended).

Hence, it is the only function that may take the asterisk as a parameter

Every other function disregards NULL values and does not factor them in to the operation.

Page D-1 MIN

This function is used to find the smallest value in a column.

Character columns return the ‘smallest’ value based on the collating sequence of the character set being used. In ASCII, the letter ‘a’ occurs earlier in the collating sequence than the letter ‘g’, hence ‘a’ is less than ‘g’.

Page D-2 Collating

To collate a list means to arrange the items in that list in some order.

A collating sequence is the rule for ordering a list of items.

A good example of collating is alphabetizing (ie. sorting). The collating sequence for the English alphabet is the list of letters arranged in this order:

abcdefghijklmnopqrstuvwxyz.

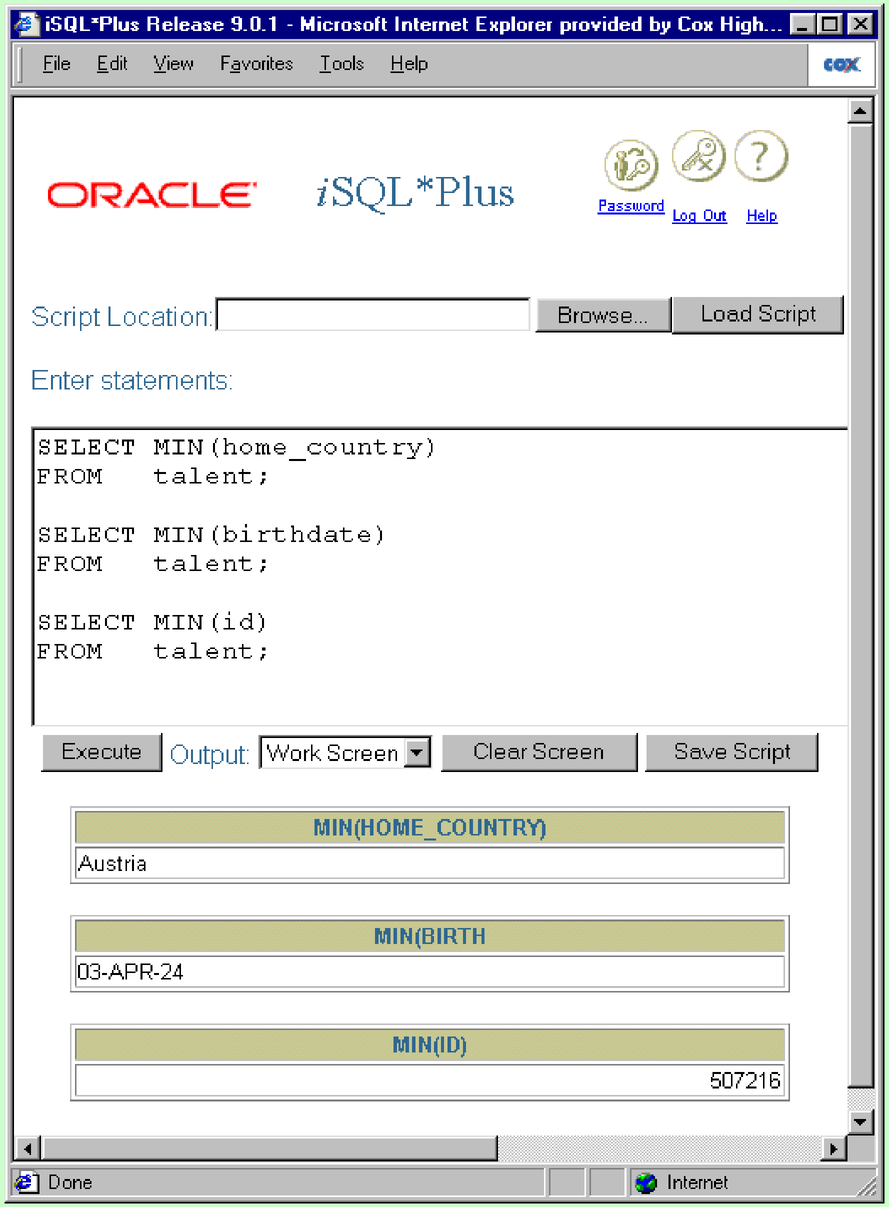
In computing we use the term collating rather than alphabetizing, because alphabetizing is restricted to only letters in the alphabet.

In computing we're concerned with sorting more than just the letters of the alphabet, we need to be able to arrange all of the characters in the character set according to some rule.

That rule is the collating sequence, and each character set has its own collating sequence.

And, on some platforms, the programmer can customize the collating sequence for her needs.

Page D-3 MIN w/ Date Values


iSQL*Plus Release 9.0.1 - Microsoft Internet Explorer provided by Cox High...
File Edit View Favorites Tools Help
ORACLE iSQL*Plus
Password Log Out Help
Script Location:
Browse... Load Script
Enter statements:
Execute Output: Work Screen Clear Screen Save Script
Done Internet

```
SELECT MIN(home_country)
FROM   talent;

SELECT MIN(birthdate)
FROM   talent;

SELECT MIN(id)
FROM   talent;
```

MIN(HOME_COUNTRY)
Austria

MIN(BIRTH
03-APR-24

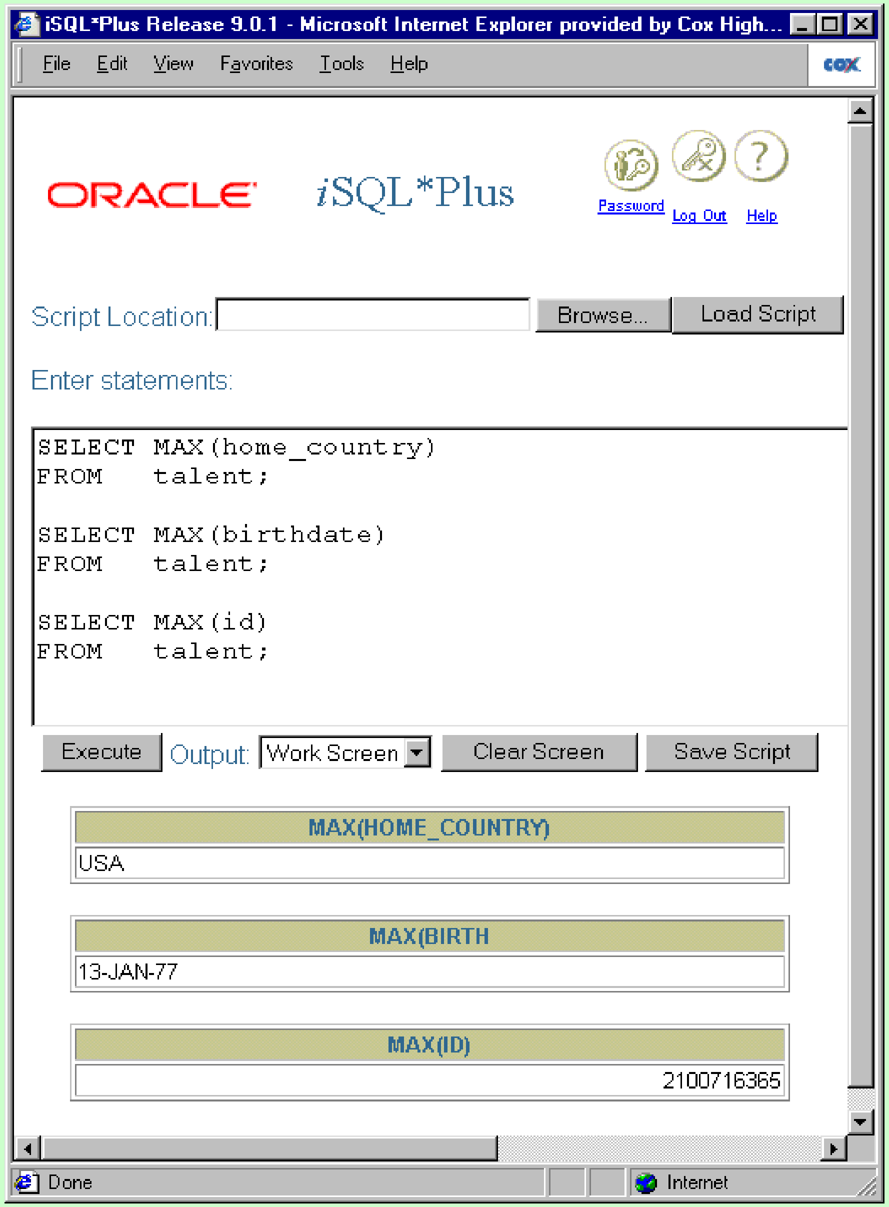
MIN(ID)
507216

When used with a date value, MIN returns the earliest date in that column.

When used with a numeric value, MIN returns the smallest value in the column.

Module 12: Set Functions

Page D-3 MAX



```
SELECT  MAX(home_country)
FROM    talent;

SELECT  MAX(birthdate)
FROM    talent;

SELECT  MAX(id)
FROM    talent;
```

MAX(HOME_COUNTRY)
USA

MAX(BIRTH
13-JAN-77

MAX(ID)
2100716365

MAX is used to find the largest value in a column.

When used with character values, MAX returns the value from the column that occurs latest in the collating sequence.

When used with a date value, MAX returns the latest date in that column.

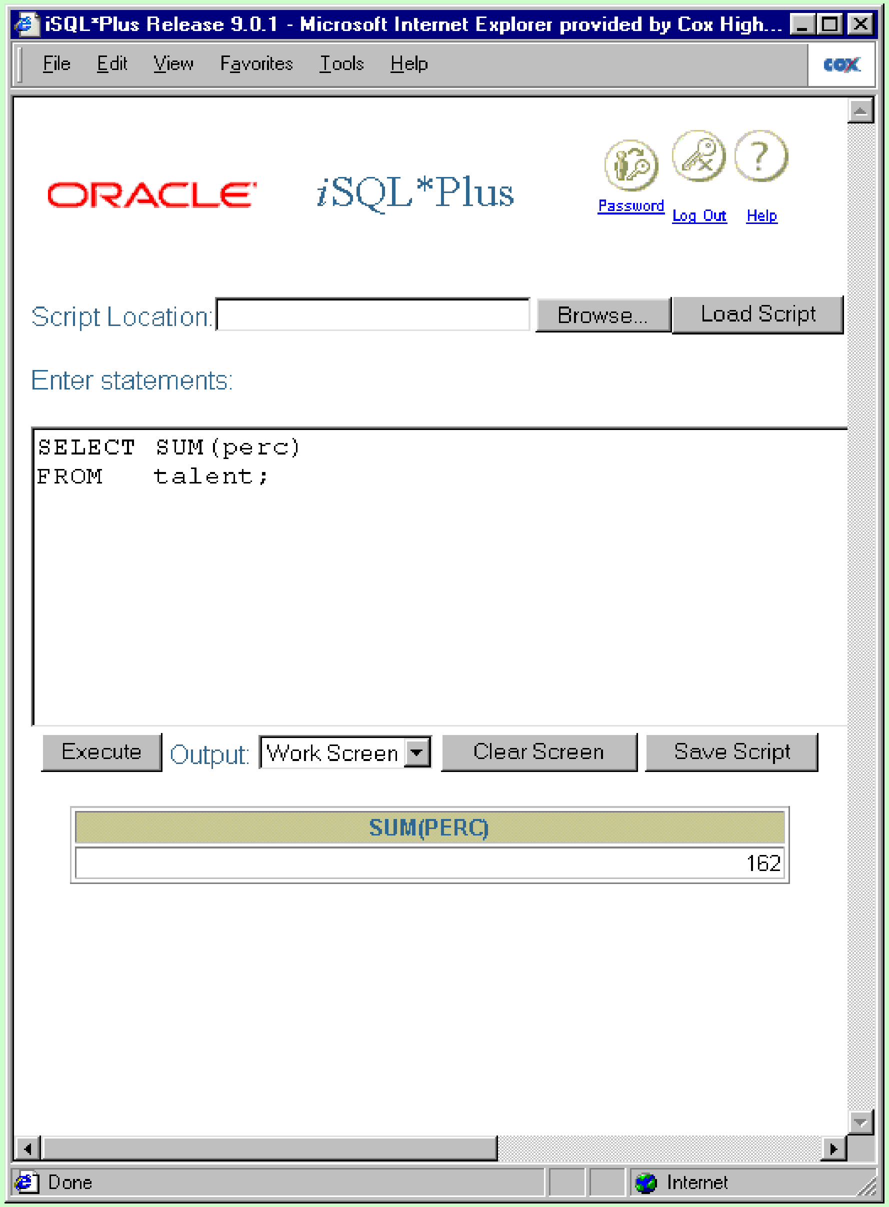
When used with a numeric value, MAX returns the largest value in the column.

Page D-3 Extrema Functions

MIN and MAX are referred to as extrema functions because they return the ‘extreme’ values, or ‘endpoints’ in the column.

Module 12: Set Functions

Page E-1 SUM


iSQL*Plus Release 9.0.1 - Microsoft Internet Explorer provided by Cox High...
File Edit View Favorites Tools Help
ORACLE iSQL*Plus
Password Log Out Help
Script Location:
Browse... Load Script
Enter statements:
Execute Output: Work Screen Clear Screen Save Script
Done Internet

```
SELECT  SUM(perc)
FROM    talent;
```

```
SUM(PERC)
      162
```

The SUM function is used to total, or add up, the values in a column.

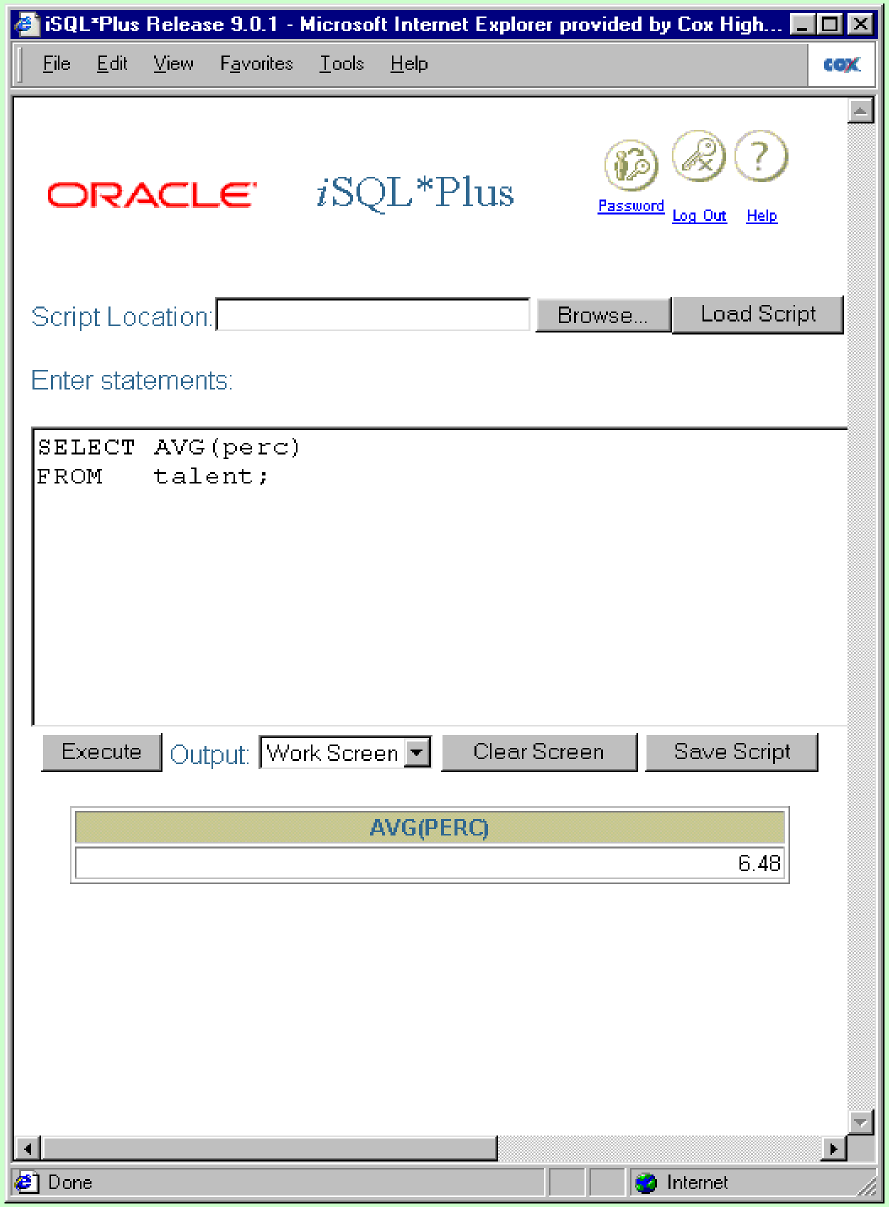
SUM can only be used with numeric data. You cannot SUM character values nor date values.

But just because you ***can*** do something, doesn't mean that you ***should*** do something.

And as smart as SQL is, it won't warn you when you do something stooopid.

The SQL program in this example is syntactically correct, and as far as SQL can tell, is semantically correct. But it is utter nonsense. What meaning (other than a description of the activity) does it have?

Page E-2 AVG



```
SELECT AVG(perc)
FROM   talent;
```

```
AVG(PERC)
     6.48
```

The AVG function is used to calculate the average (arithmetic mean) of the values in a column.

And just as we saw with the SUM function, the AVG function can only operate on numeric domains.

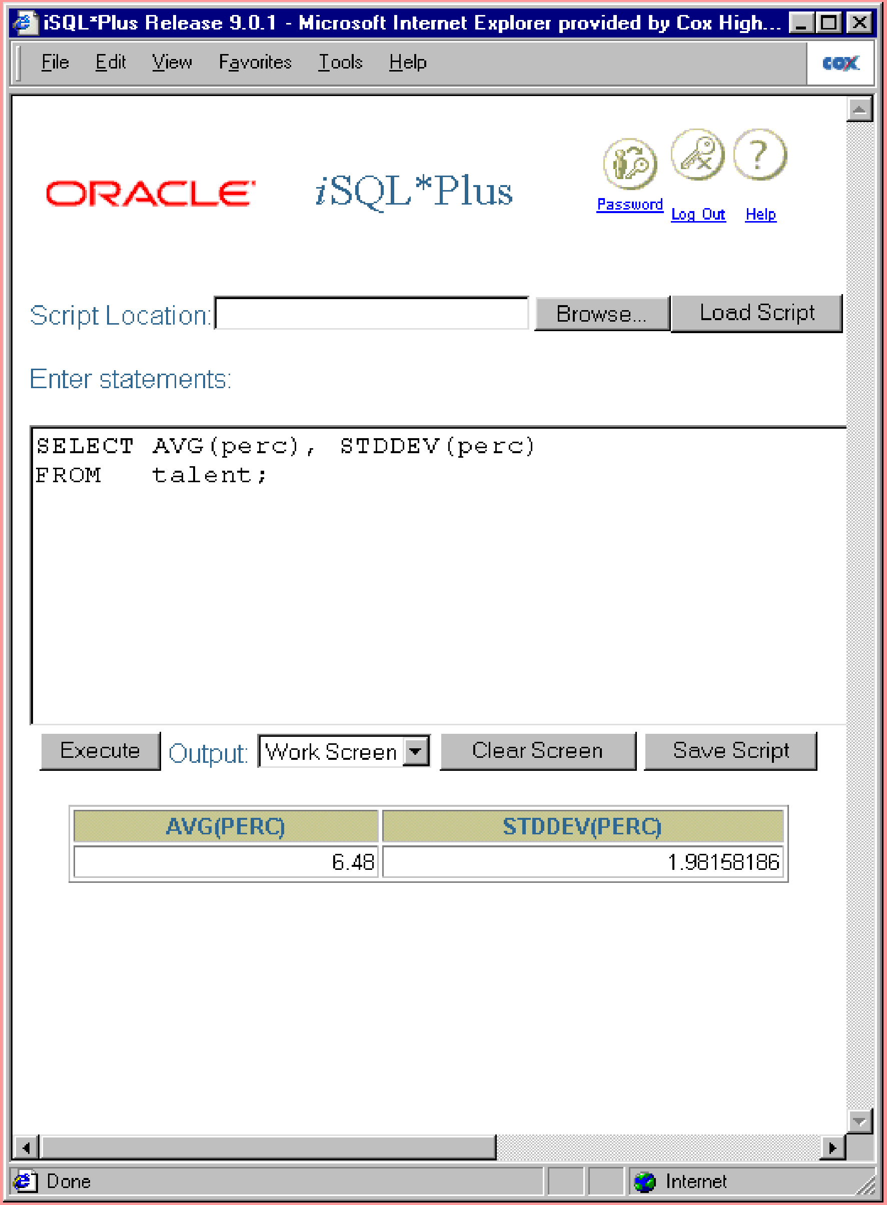
Page H-1 Vendor Extensions

I think of set functions as tools that I can use to gather and display statistics about the data in my database.

Both Oracle and MySQL provide some additional statistical functions.

Function	Oracle	MySQL
Standard Deviation	STDDEV	STD, STDDEV, STDDEV_POP
		STDDEV_SAMP
Variance	VARIANCE	VARIANCE VAR_POP
		VAR_SAMP

Page H-2 STDDEV


iSQL*Plus Release 9.0.1 - Microsoft Internet Explorer provided by Cox High...
File Edit View Favorites Tools Help
ORACLE iSQL*Plus
Password Log Out Help
Script Location:
Browse... Load Script
Enter statements:
Execute Output: Work Screen Clear Screen Save Script
Done Internet

```
SELECT AVG(perc), STDDEV(perc)
FROM   talent;
```

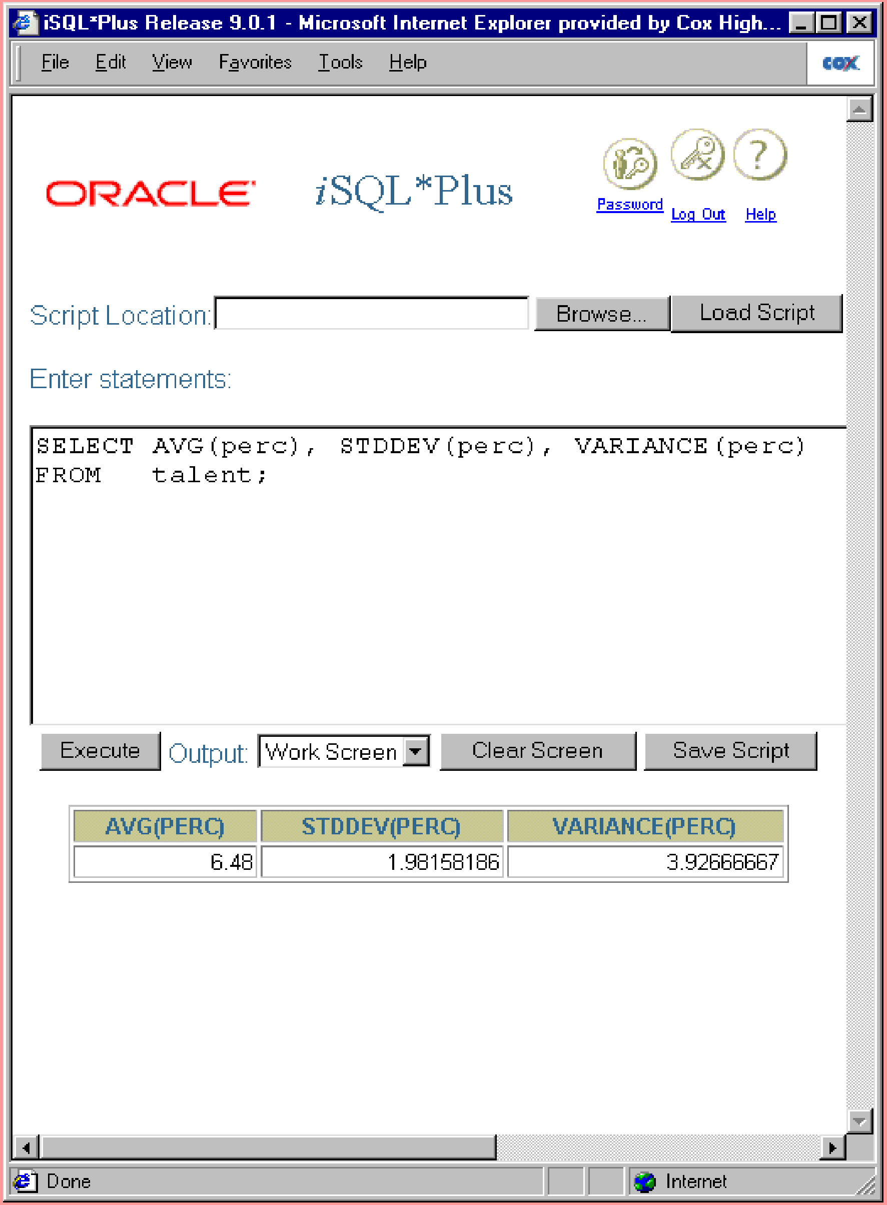
AVG(PERC)	STDDEV(PERC)
6.48	1.98158186

The standard deviation is a measure of variation from the mean. In other words, it helps you visualize how all of the data is clustered around the mean.

In my opinion, knowing the average value for any data set is almost meaningless unless it is accompanied by some information about the variation of the data.

I'm not going to delve any further into 'statistical realms'. If you understand and appreciate the value of metrics such as the standard deviation and the variance, just know that Oracle provides set functions that will easily calculate those values for you.

If you don't appreciate these metrics, then get thee to a stats class ;-)



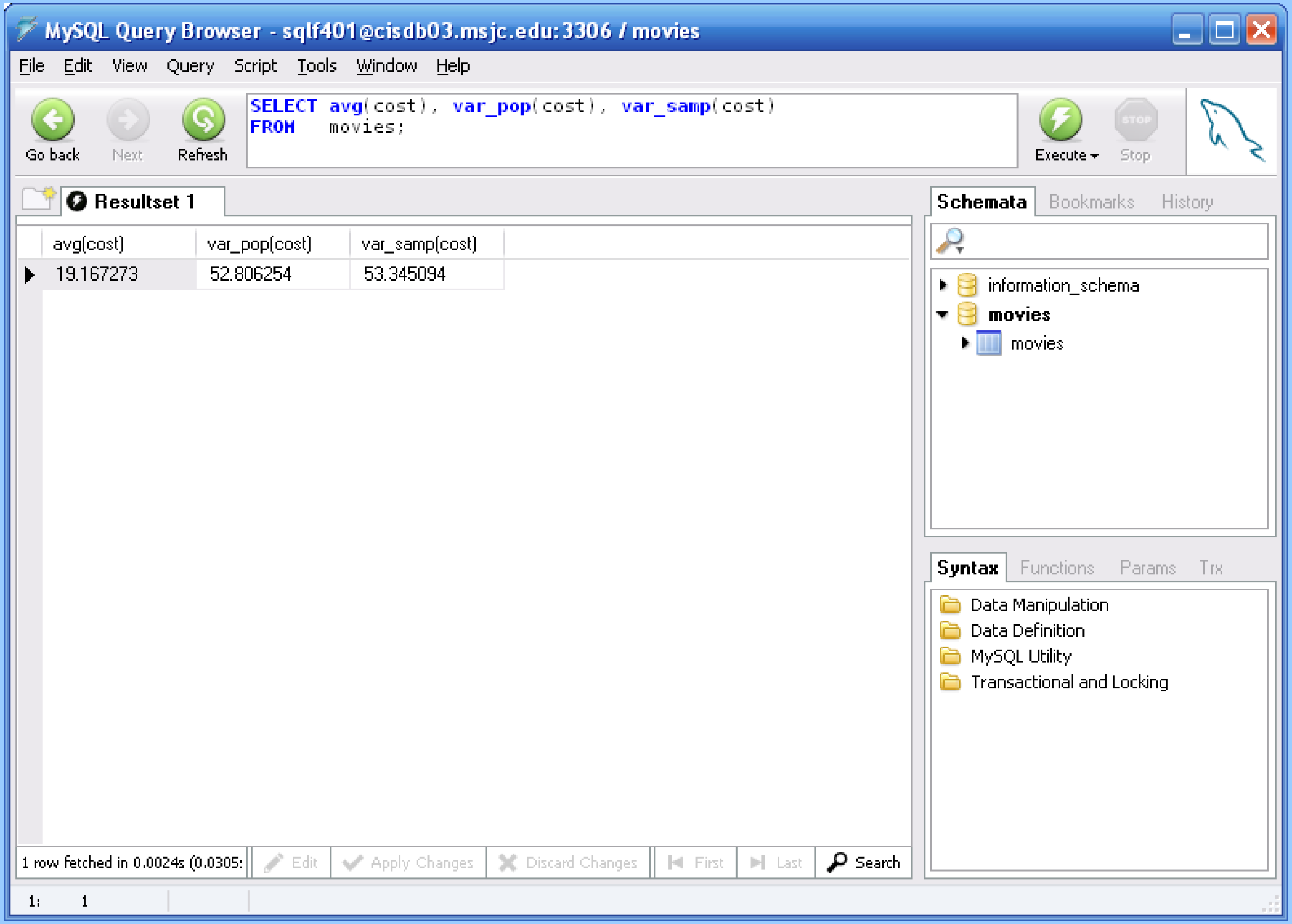
```
SELECT AVG(perc), STDDEV(perc), VARIANCE(perc)
FROM   talent;
```

AVG(PERC)	STDDEV(PERC)	VARIANCE(PERC)
6.48	1.98158186	3.92666667

VARIANCE is another measure of the variation of the data about the mean.

Module 12: Set Functions

Page H-4 VAR_POP


MySQL Query Browser - sqlf401@cisdb03.msjc.edu:3306 / movies
File Edit View Query Script Tools Window Help
Go back
Next
Refresh
Execute
Stop
Resultset 1
Schemata Bookmarks History
information_schema
movies
movies
Syntax Functions Params Trx
Data Manipulation
Data Definition
MySQL Utility
Transactional and Locking
Edit
Apply Changes
Discard Changes
First
Last
Search

```
SELECT avg(cost), var_pop(cost), var_samp(cost)
FROM   movies;
```

avg(cost)	var_pop(cost)	var_samp(cost)
19.167273	52.806254	53.345094

1 row fetched in 0.0024s (0.0305:

The statistics functions in Oracle are population statistics. MySQL gives you the option of calculating these statistics as either *population* or *sample* statistics.

The distinction between these two functions is based on whether we are treating the data as the entirety of the population of data (all of the data), or 'just' a sample of the data.

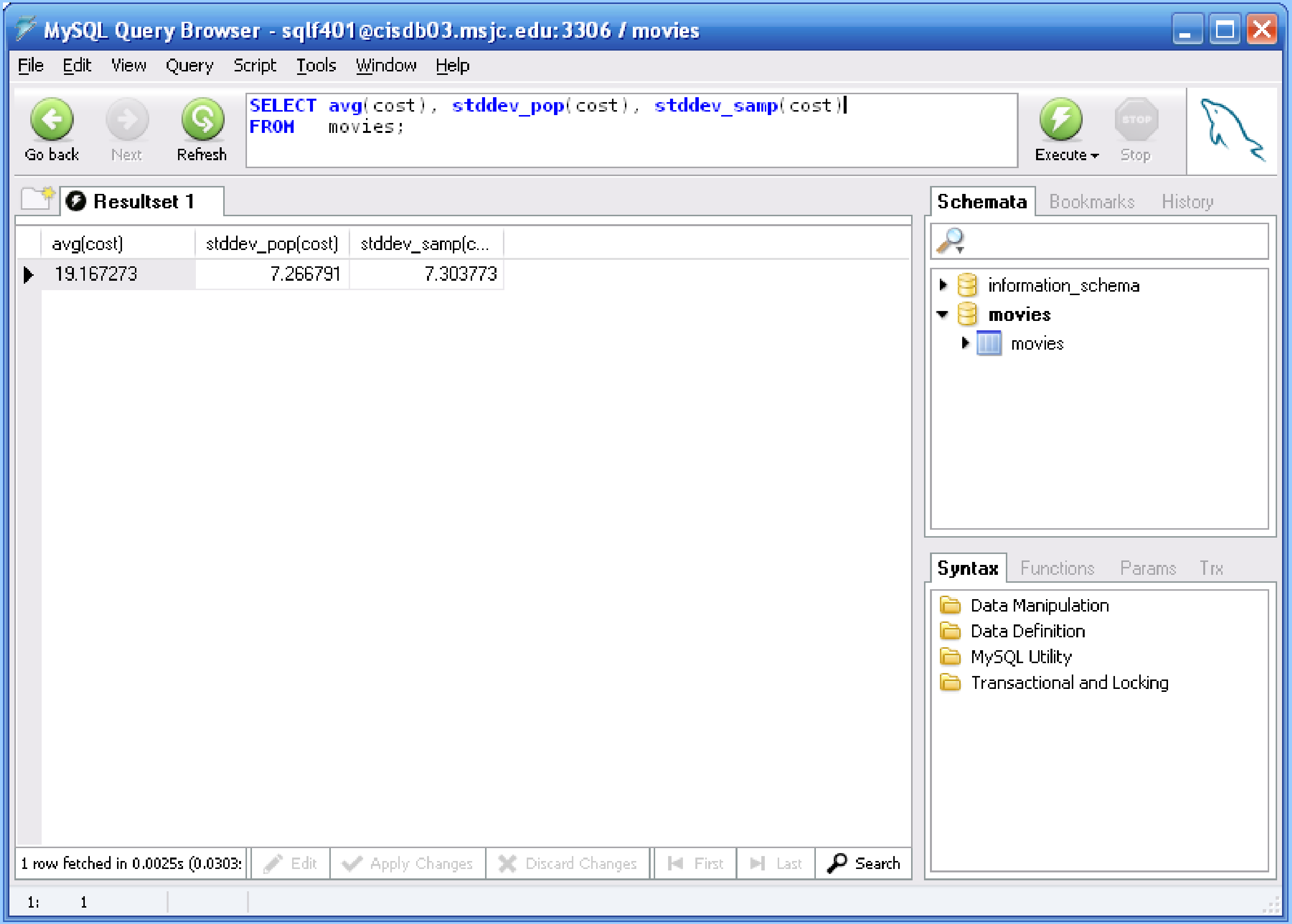
For the population calculation, the denominator is the number of rows.

For the sample calculation, the denominator is the number of rows, less 1.

Module 12: Set Functions

Page H-5 STDDEV_POP

As we saw with the VARIANCE statistic, there are also population and sample statistics for the standard deviation.



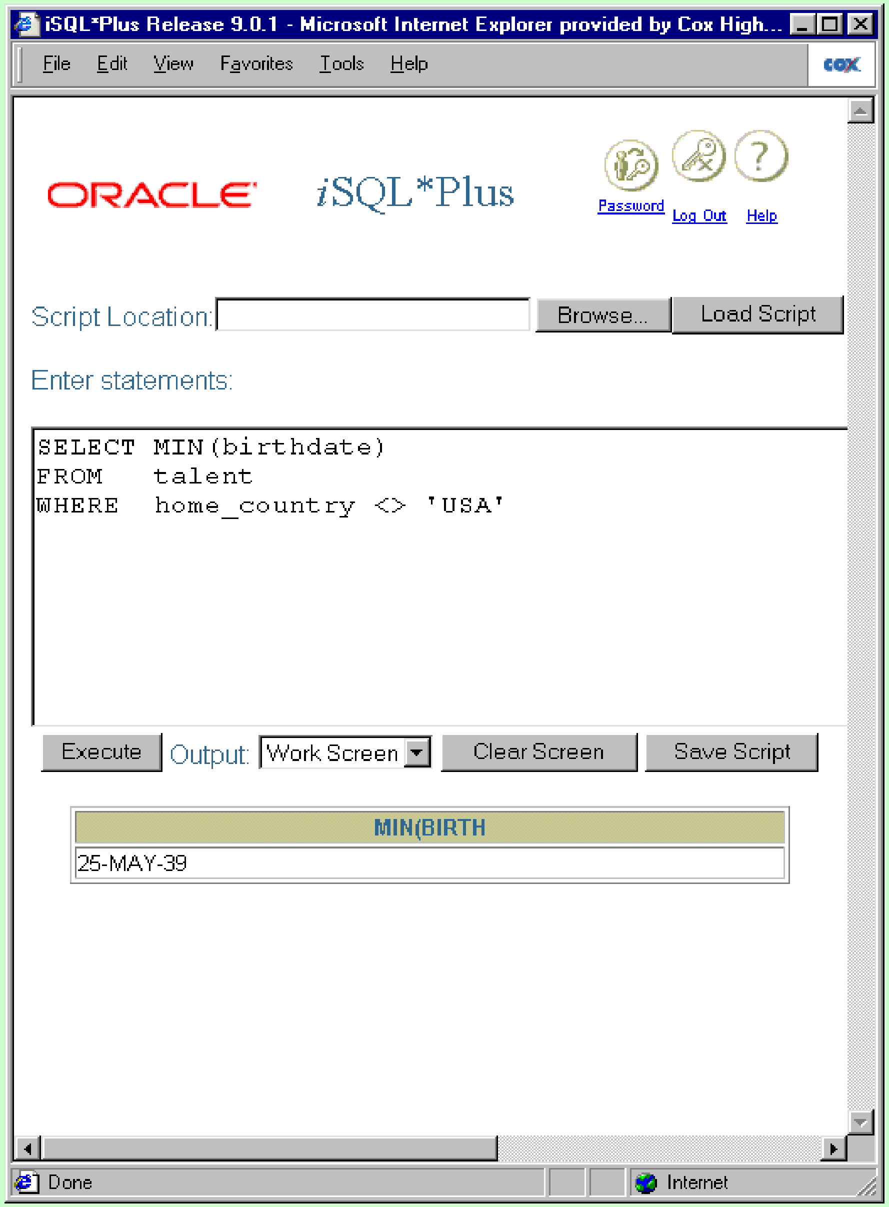
```
SELECT avg(cost), stddev_pop(cost), stddev_samp(cost)
FROM   movies;
```

```
avg(cost)   stddev_pop(cost)  stddev_samp(c...
19.167273           7.266791          7.303773
```

1 row fetched in 0.0025s (0.0303s)

Module 12: Set Functions

Page I-1: WHERE clause


iSQL*Plus Release 9.0.1 - Microsoft Internet Explorer provided by Cox High...
File Edit View Favorites Tools Help
ORACLE iSQL*Plus
Password Log Out Help
Script Location:
Browse... Load Script
Enter statements:
Execute Output: Work Screen Clear Screen Save Script
Done Internet

```
SELECT MIN(birthdate)
FROM   talent
WHERE  home_country <> 'USA'
```

```
MIN(BIRTH
25-MAY-39
```

We can use the where clause in tandem with the column functions.

In this example we're looking for the oldest foreign-born client.

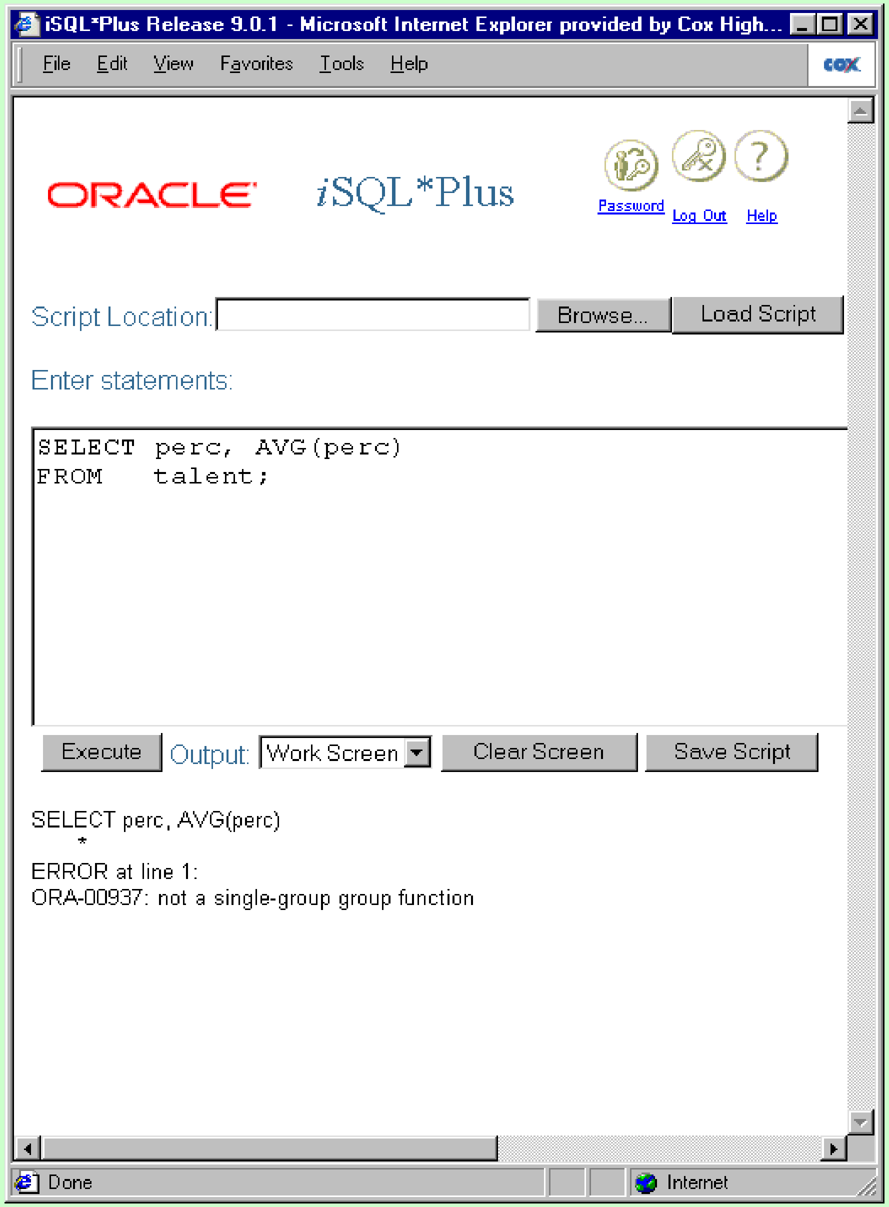
In a subsequent module we'll re-examine the SQL processing model and try to better understand how all of the pieces fit together. Until then, you should 'give it a go'.

Which comes first, now?

SELECT
FROM
WHERE
ORDER BY

Module 12: Set Functions

Page S-1: Summary


iSQL*Plus Release 9.0.1 - Microsoft Internet Explorer provided by Cox High...
File Edit View Favorites Tools Help
ORACLE iSQL*Plus
Password Log Out Help
Script Location:
Browse... Load Script
Enter statements:
Execute Output: Work Screen Clear Screen Save Script
Done Internet

```
SELECT perc, AVG(perc)
FROM   talent;
```

```
SELECT perc, AVG(perc)
       *
ERROR at line 1:
ORA-00937: not a single-group group function
```

The set functions that we examined in this module can be used to gather and display statistical information about the data base.

This statistical information is, by it’s nature, a summary or an aggregation about some aspect of the data.

Summary information cannot be combined with detail information, else SQL will throw an error. As the slide to the left demonstrates.

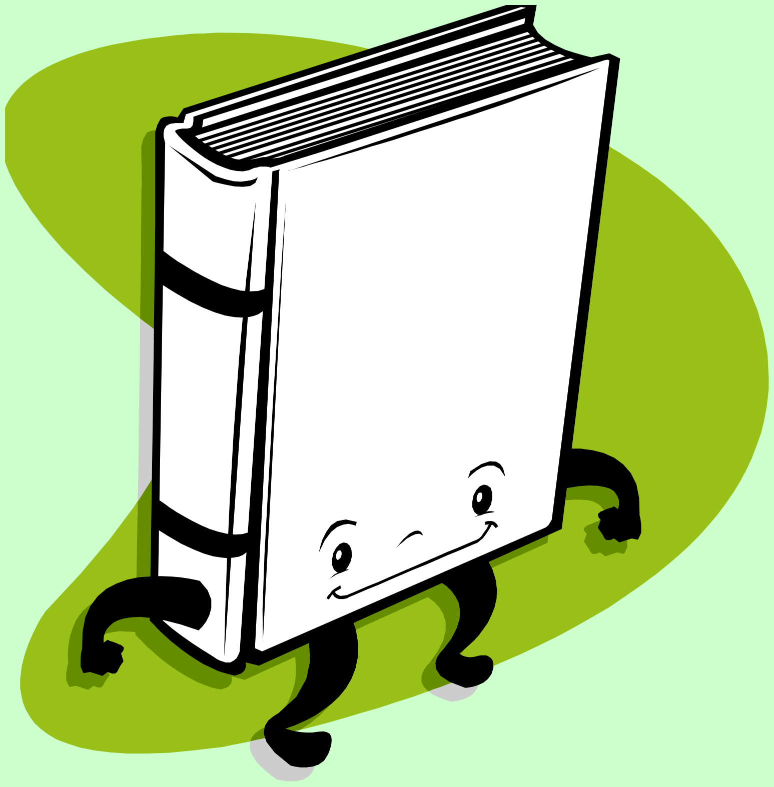
Page T-1: Terminology

Group function, set function, aggregate function

COUNT(column-name), COUNT(*)
SUM()
MIN()
MAX()
AVG()
Tally

STDDEV()
VARIANCE()

Collating sequence
Character set
extrema



Page Z-1: End Notes

Please drop me an email if you noticed any errors in this module. I'd also appreciate reading your comments, criticisms, and or suggestions as to how this module could be improved.

Thanks,

bil

That’s All