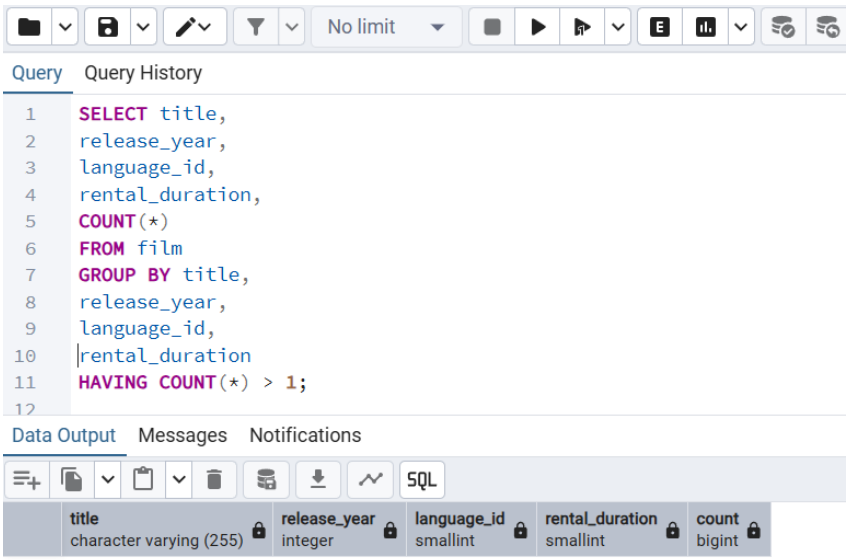
EXCERCE 3-6: Summarizing & Cleaning Data in SQL

1/Check for and clean dirty data for film table

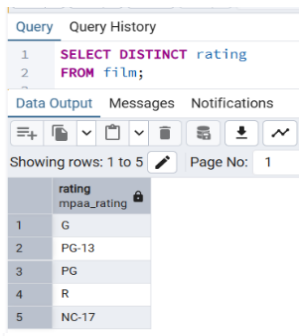
Duplicate value

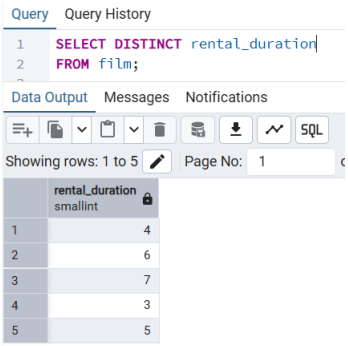


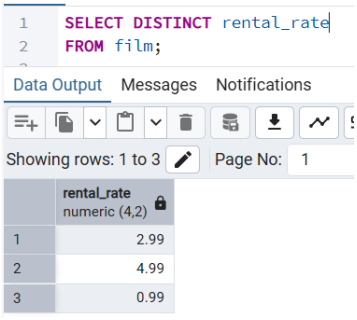
There is no duplicate data

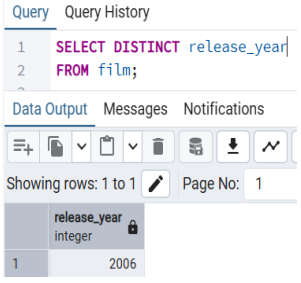
>>>I ll use SELECT DISTINCT to avoid counting the same value multiple times or I create the sql view, and I use DELETE

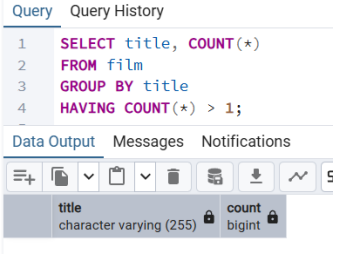
Non-uniform value











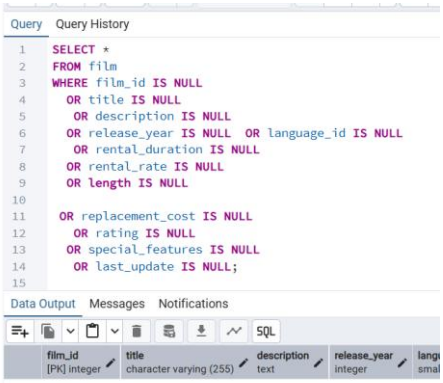
There is not non-uniform value

>>>If there is non-uniform value, I will use UPDATE with sql view :**UPDATE film**

SET rental_rate = **

WHERE film_id = ***

Missing data



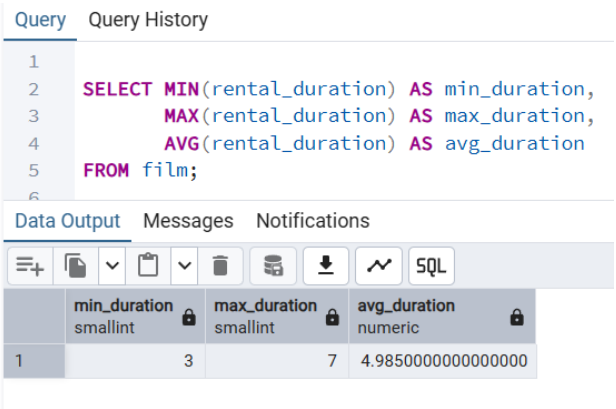
```
SELECT *
FROM film
WHERE film_id IS NULL
  OR title IS NULL
   OR description IS NULL
  OR release_year IS NULL  OR language_id IS NULL
   OR rental_duration IS NULL
  OR rental_rate IS NULL
  OR length IS NULL

 OR replacement_cost IS NULL
   OR rating IS NULL
  OR special_features IS NULL
   OR last_update IS NULL;
```

Data Output | Messages | Notifications

film_id [PK] integer	title character varying (255)	description text	release_year integer	langu smal

There is not missing value, if there is, I will use update with sql view

2/Summarizing data for film table (MIN; MAX; AVG; MODE)

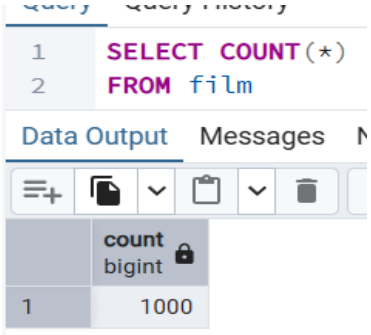


```
SELECT MIN(rental_duration) AS min_duration,
       MAX(rental_duration) AS max_duration,
       AVG(rental_duration) AS avg_duration
FROM film;
```

Data Output | Messages | Notifications

	min_duration smallint	max_duration smallint	avg_duration numeric
1	3	7	4.9850000000000000

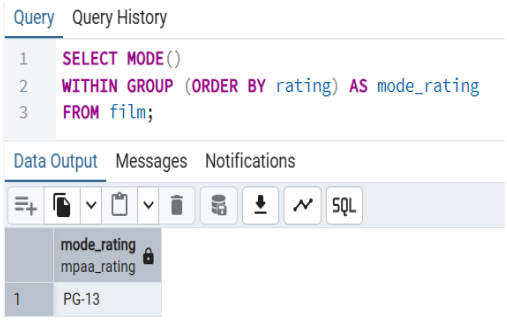
Colomn	MIN	MAX	AVG
rental duration	3	7	4.985
rental rate	0.99	4.99	2.98
length	46	185	115.272
replacement cost	9.99	29.99	19.984
film id	1	1000	500.5



```
SELECT COUNT(*)
FROM film
```

Data Output | Messages

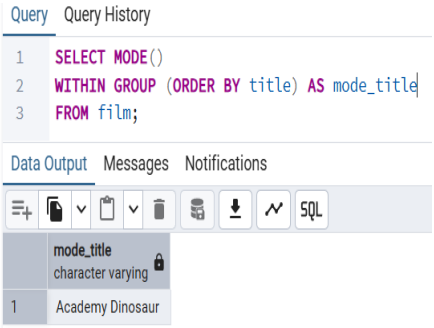
	count bigint
1	1000



```
SELECT MODE()
WITHIN GROUP (ORDER BY rating) AS mode_rating
FROM film;
```

Data Output | Messages | Notifications

	mode_rating mpaa_rating
1	PG-13



```
SELECT MODE()
WITHIN GROUP (ORDER BY title) AS mode_title
FROM film;
```

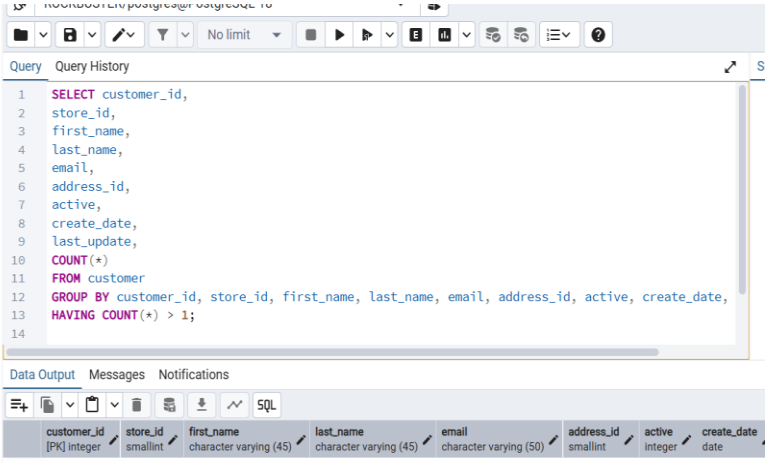
Data Output | Messages | Notifications

	mode_title character varying
1	Academy Dinosaur

Column	Mode
title	Academy Dinosaur
rating	PG-13
special_features	Trailers

Check for and clean dirty data for Customer table

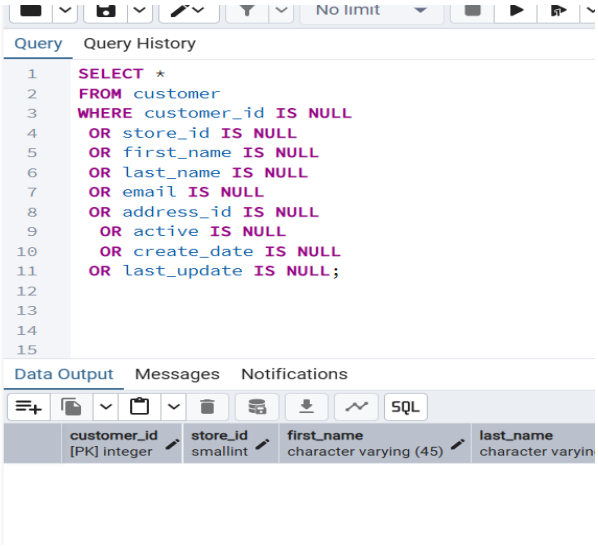
Duplicate value



There is no duplicate data

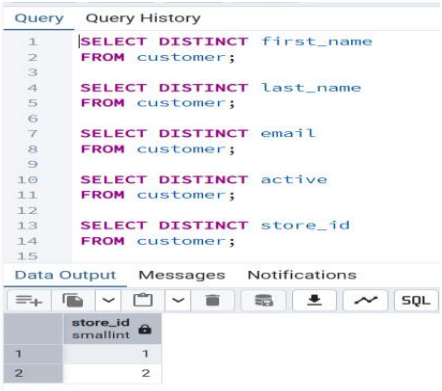
>>>I ll use SELECT DISTINCT to avoid counting the same value multiple times or I create the sql view, and I use DELETE

Missing value



There is no missing value, if there is, I will use update with sql view

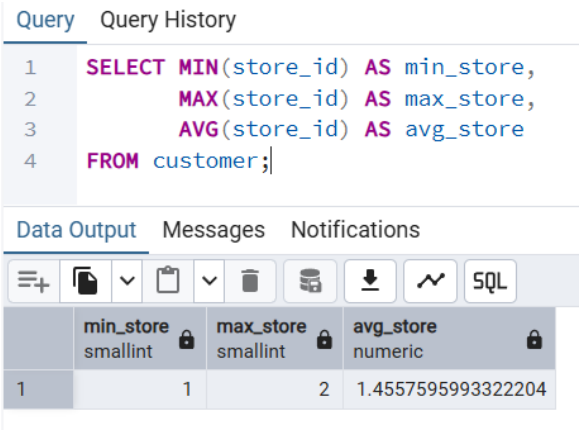
Non-uniform data



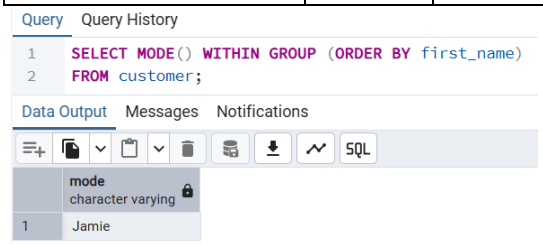
There is no non-uniform value

>>>If there is non-uniform value, I will use UPDATE with SQL View

Summarizing data for Customer table (MIN; MAX; AVG; MODE)



Colomn	MIN	MAX	AVG
custumer id	1	599	300
store id	1	2	1.45576
adress id	5	605	304.7245



Column	Mode
first_name	Jamie
last_name	Abney
email	aaron.selby@sakila.customer.org

Reflect on your work

I feel that with SQL, the work is simpler and faster; you can insert three functions at the same time for example. You just need to master the SQL commands and syntax, which is easier with Excel, where with a few clicks you get what you want.