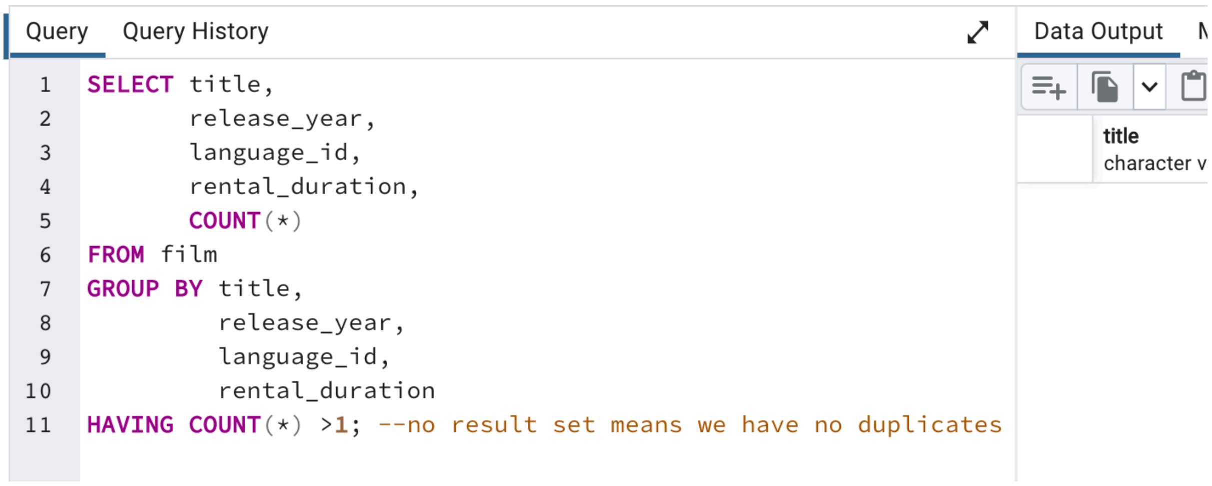
E.Christian Cotterman
Tuesday, 30. May 2023

Answers 3.6

3.6: Summarizing & Cleaning Data in SQL

- 1) No duplicates found.

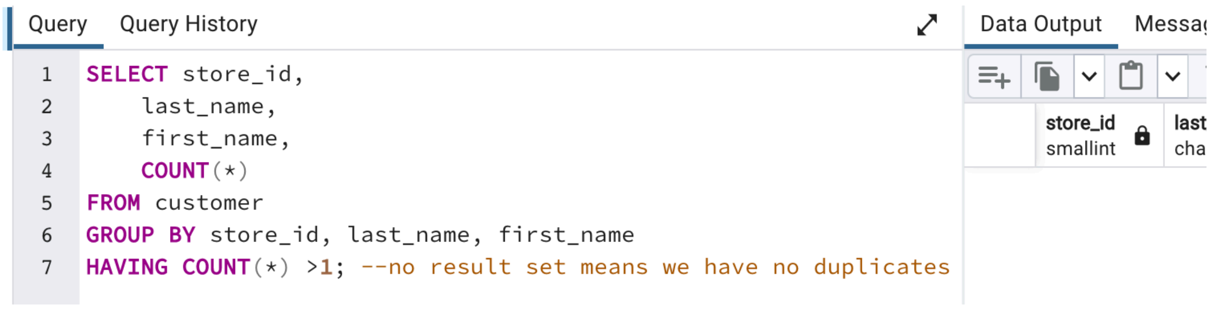


Query Query History

```
SELECT title,
       release_year,
       language_id,
       rental_duration,
       COUNT(*)
FROM film
GROUP BY title,
         release_year,
         language_id,
         rental_duration
HAVING COUNT(*) >1; --no result set means we have no duplicates
```

Data Output

title
character v



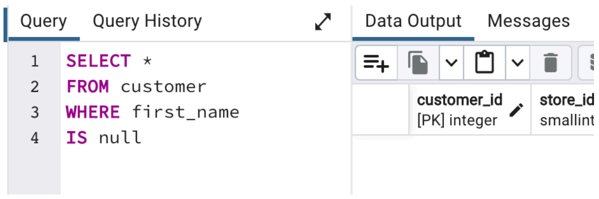
Query Query History

```
SELECT store_id,
    last_name,
    first_name,
    COUNT(*)
FROM customer
GROUP BY store_id, last_name, first_name
HAVING COUNT(*) >1; --no result set means we have no duplicates
```

Data Output Messag

store_id smallint | last cha

- No missing information found with 2 searches in both Customer and Film.

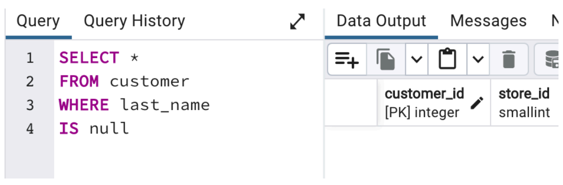


Query Query History

```
SELECT *
FROM customer
WHERE first_name
IS null
```

Data Output Messages

customer_id [PK] integer | store_id smallint



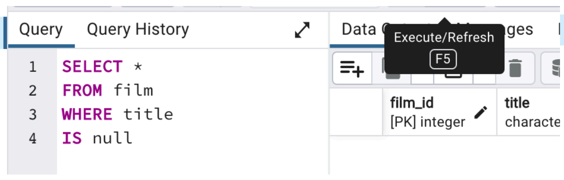
Query Query History

```
SELECT *
FROM customer
WHERE last_name
IS null
```

Data Output Messages

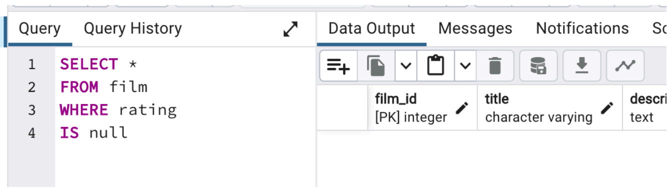
customer_id [PK] integer | store_id smallint

E.Christian Cotterman
Tuesday, 30. May 2023



```
SELECT *
FROM film
WHERE title
IS null
```

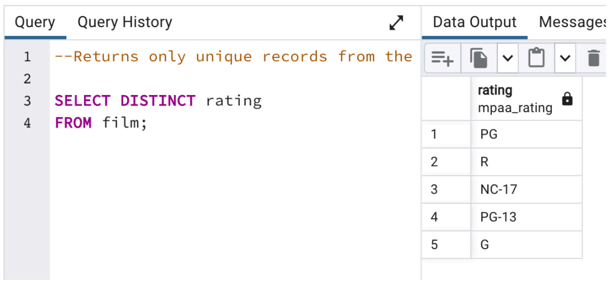
film_id [PK] integer	title characte



```
SELECT *
FROM film
WHERE rating
IS null
```

film_id [PK] integer	title character varying	descri text

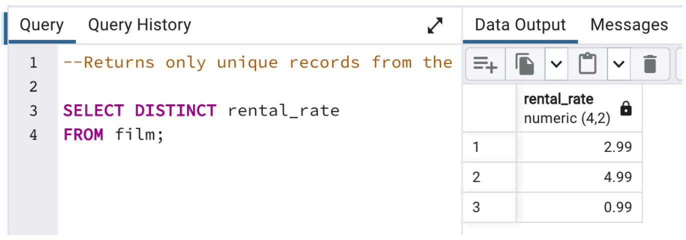
- Using DISTINCT I looked at ratings and rental_rate. There are 5 variables for rating and 3 for rental_rate.



```
--Returns only unique records from the

SELECT DISTINCT rating
FROM film;
```

	rating mpaa_rating
1	PG
2	R
3	NC-17
4	PG-13
5	G



```
--Returns only unique records from the

SELECT DISTINCT rental_rate
FROM film;
```

	rental_rate numeric (4,2)
1	2.99
2	4.99
3	0.99

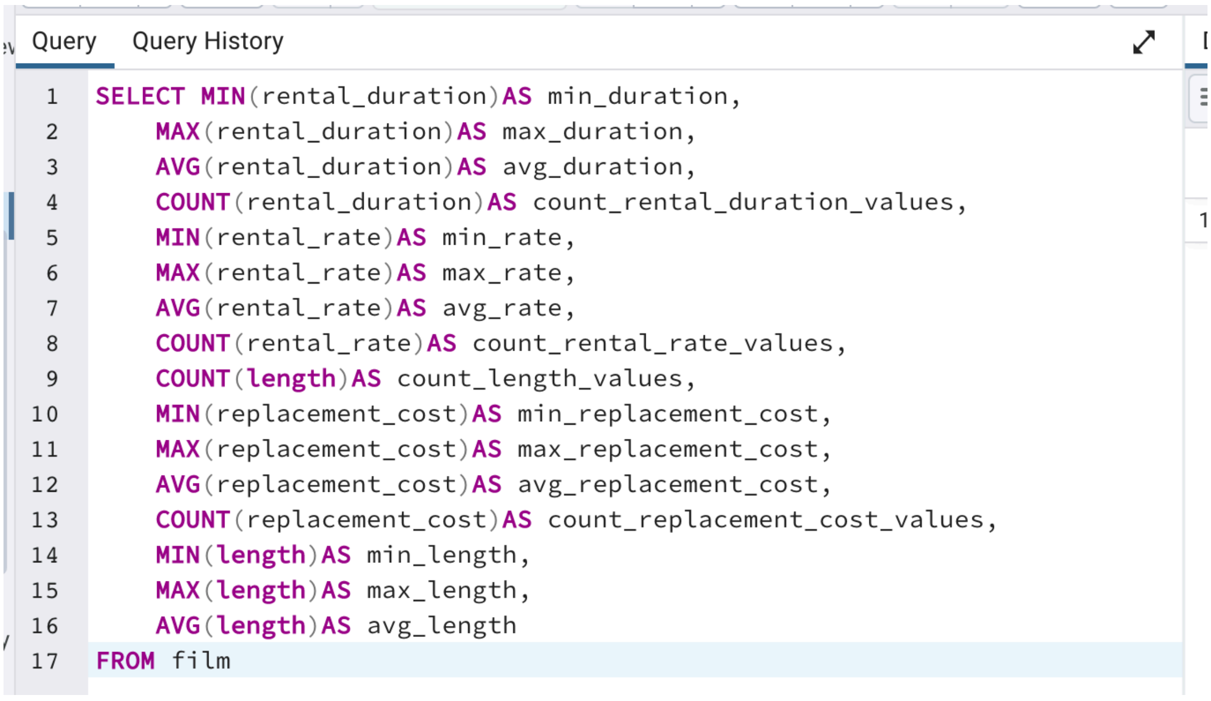
— Cleaning the affected units would require permissions firstly. To change the duplicates, we could GROUP the duplicates.

—As we would likely not be able to DELETE data, we may be able to add an ACTIVE or INACTIVE filter to the accounts with duplicate names (for Customers).

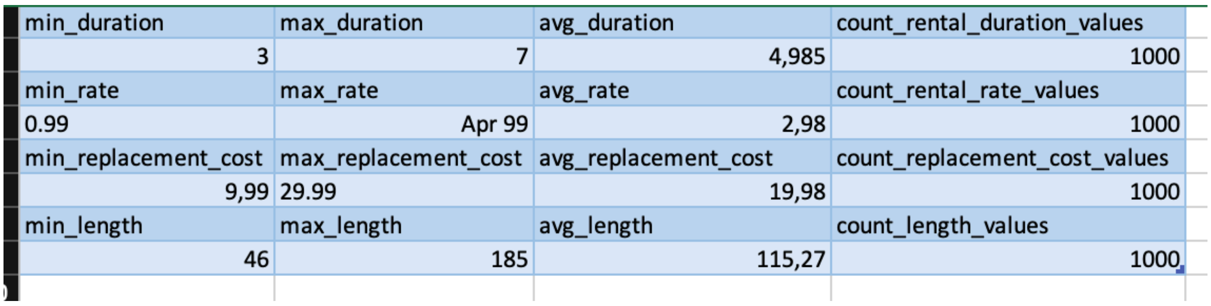
—For the missing data, we can use an educated guess for that. (*New releases are 4.99; less than 5 years old 2.99; over 5 years old .99¢*)

E.Christian Cotterman
Tuesday, 30. May 2023

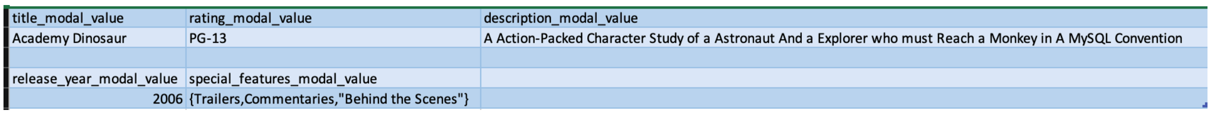
- 2) Film Data summaries.


Query Query History

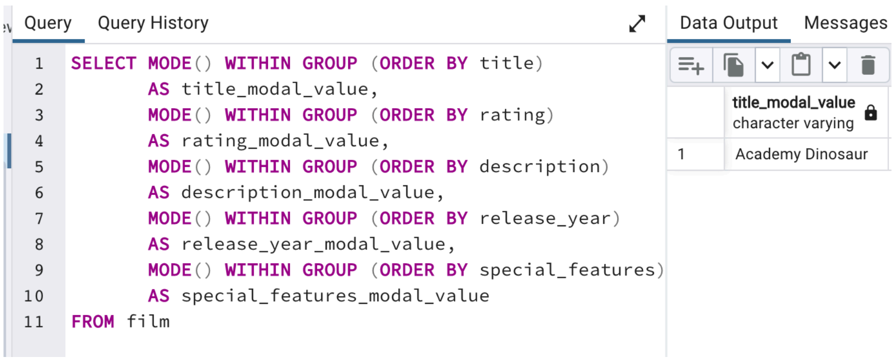
```
SELECT MIN(rental_duration)AS min_duration,
     MAX(rental_duration)AS max_duration,
     AVG(rental_duration)AS avg_duration,
     COUNT(rental_duration)AS count_rental_duration_values,
     MIN(rental_rate)AS min_rate,
     MAX(rental_rate)AS max_rate,
     AVG(rental_rate)AS avg_rate,
     COUNT(rental_rate)AS count_rental_rate_values,
     COUNT(length)AS count_length_values,
     MIN(replacement_cost)AS min_replacement_cost,
     MAX(replacement_cost)AS max_replacement_cost,
     AVG(replacement_cost)AS avg_replacement_cost,
     COUNT(replacement_cost)AS count_replacement_cost_values,
     MIN(length)AS min_length,
     MAX(length)AS max_length,
     AVG(length)AS avg_length
FROM film
```



min_duration	max_duration	avg_duration	count_rental_duration_values
3	7	4,985	1000
min_rate	max_rate	avg_rate	count_rental_rate_values
0.99	Apr 99	2,98	1000
min_replacement_cost	max_replacement_cost	avg_replacement_cost	count_replacement_cost_values
9,99	29.99	19,98	1000
min_length	max_length	avg_length	count_length_values
46	185	115,27	1000



title_modal_value	rating_modal_value	description_modal_value
Academy Dinosaur	PG-13	A Action-Packed Character Study of a Astronaut And a Explorer who must Reach a Monkey in A MySQL Convention
release_year_modal_value	special_features_modal_value	
2006	{Trailers,Commentaries,"Behind the Scenes"}	


Query Query History

```
SELECT MODE() WITHIN GROUP (ORDER BY title)
       AS title_modal_value,
       MODE() WITHIN GROUP (ORDER BY rating)
       AS rating_modal_value,
       MODE() WITHIN GROUP (ORDER BY description)
       AS description_modal_value,
       MODE() WITHIN GROUP (ORDER BY release_year)
       AS release_year_modal_value,
       MODE() WITHIN GROUP (ORDER BY special_features)
       AS special_features_modal_value
FROM film
```

Data Output Messages

	title_modal_value character varying
1	Academy Dinosaur

E.Christian Cotterman
Tuesday, 30. May 2023

- 3). I am slowly becoming more comfortable with SQL. Thinking about how to best clean and organise data, Excel seems slightly ahead with its ability to filter the data within each column. If you want PG-13 films from 1994 for example. SQL is easier because I just had to copy-paste the queries one after the other, changing what it was I was looking for, so I could create larger/longer strings. (*See the MIN/MAX example).* SQL may end up being more useful for me (in the sense of I could use it more easily) than Excel, but I assume it's all about using both daily for long enough to really grasp the strengths and weaknesses of each.