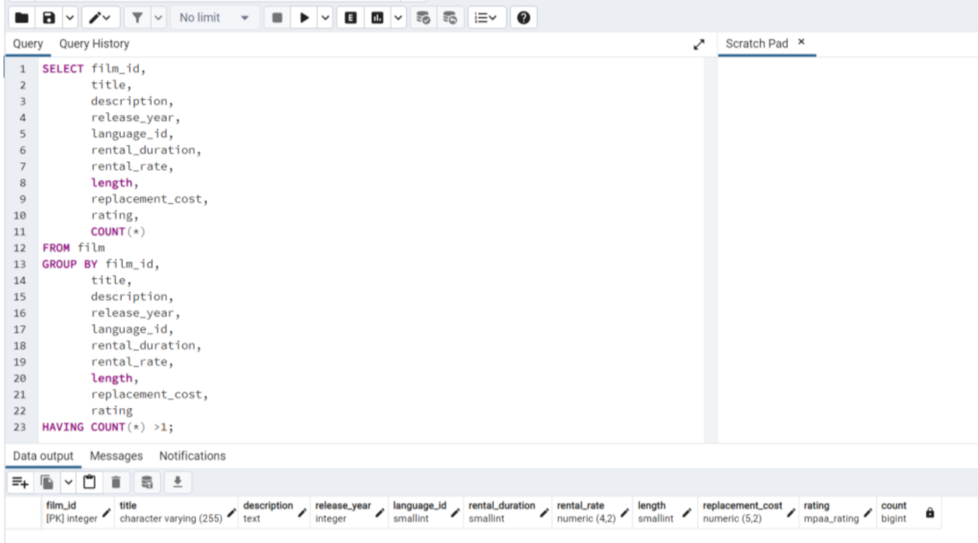
Answers 3.6

1. **Check for and clean dirty data:** Find out if the film table and the customer table contain any dirty data, specifically non-uniform or duplicate data, or missing values. Create a new "Answers 3.6" document and copy-paste your queries into it. Next to each query write 2 to 3 sentences explaining how you would clean the data (even if the data is not dirty).

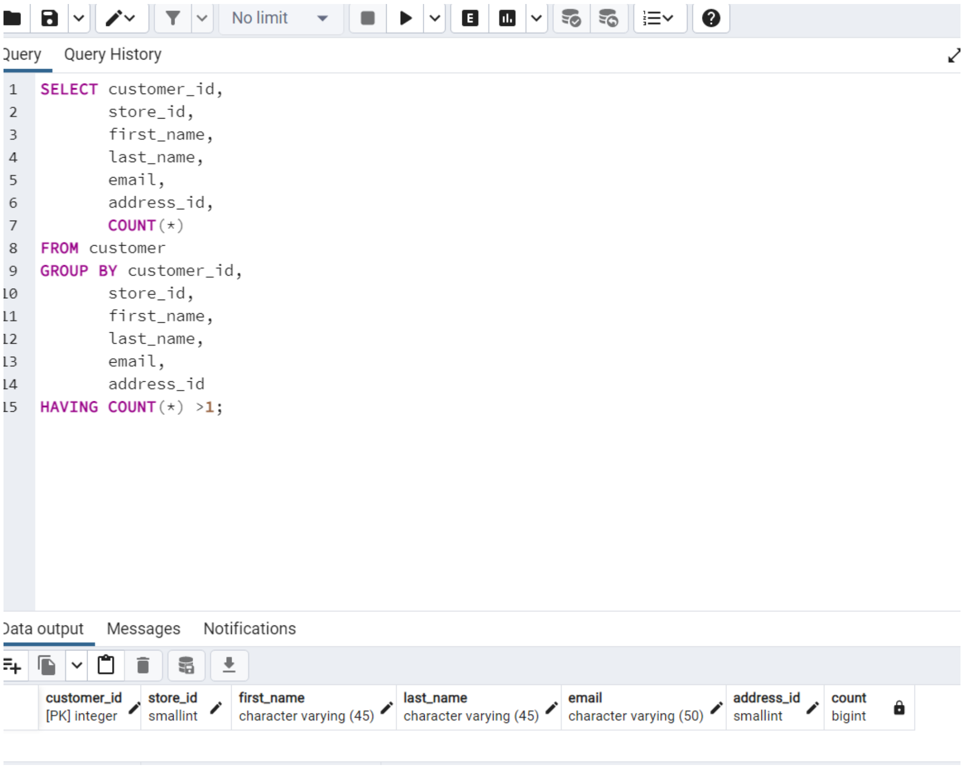
Checking for duplication in film



```sql
SELECT film_id,
       title,
       description,
       release_year,
       language_id,
       rental_duration,
       rental_rate,
       length,
       replacement_cost,
       rating,
       COUNT(*)
FROM film
GROUP BY film_id,
       title,
       description,
       release_year,
       language_id,
       rental_duration,
       rental_rate,
       length,
       replacement_cost,
       rating
HAVING COUNT(*) >1;
```

film_id [PK] integer	title character varying (255)	description text	release_year integer	language_id smallint	rental_duration smallint	rental_rate numeric (4,2)	length smallint	replacement_cost numeric (5,2)	rating mpaa_rating	count bigint

Checking for duplication in customer table



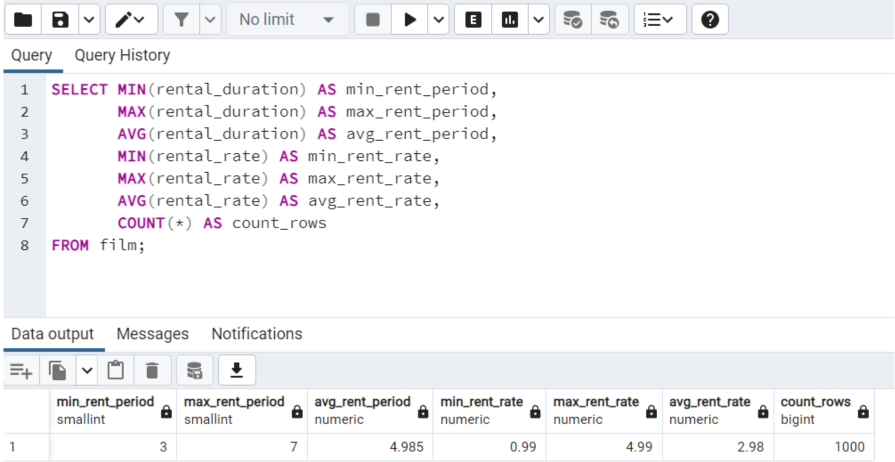
```sql
SELECT customer_id,
       store_id,
       first_name,
       last_name,
       email,
       address_id,
       COUNT(*)
FROM customer
GROUP BY customer_id,
       store_id,
       first_name,
       last_name,
       email,
       address_id
HAVING COUNT(*) >1;
```

customer_id [PK] integer	store_id smallint	first_name character varying (45)	last_name character varying (45)	email character varying (50)	address_id smallint	count bigint

There is no returned duplicate value. If there were duplication, there would have been two ways to deal with it. One, create a virtual table “view” where unique records can be selected. Two, Delete duplicate record from the table or view. If neither of those are permitted by the company, we can use GROUP BY or DISTINCT to select unique records.

2. **Summarize your data:** Use SQL to calculate descriptive statistics for both the film table and the customer table. For numerical columns, this means finding the minimum, maximum, and average values. For non-numerical columns, calculate the mode value. Copy-paste your SQL queries and their outputs into your answers document.

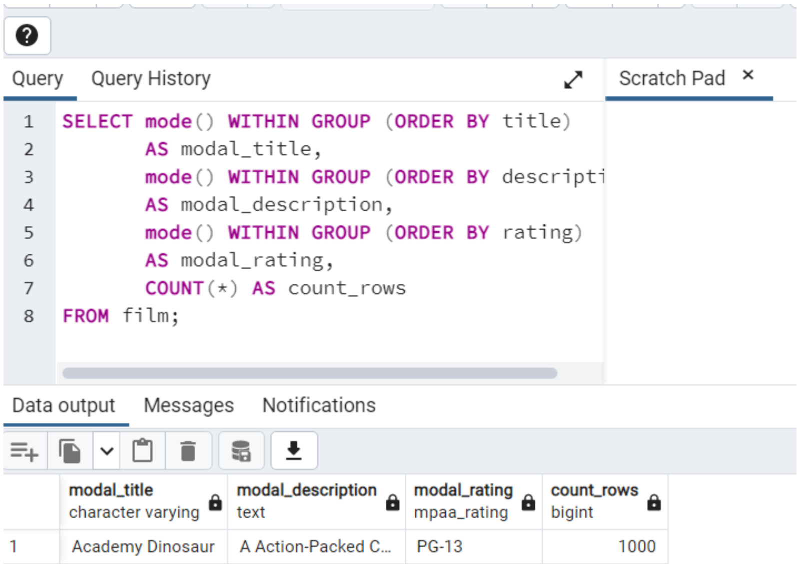
Summary for numeric columns in film



```
SELECT MIN(rental_duration) AS min_rent_period,
       MAX(rental_duration) AS max_rent_period,
       AVG(rental_duration) AS avg_rent_period,
       MIN(rental_rate) AS min_rent_rate,
       MAX(rental_rate) AS max_rent_rate,
       AVG(rental_rate) AS avg_rent_rate,
       COUNT(*) AS count_rows
FROM film;
```

	min_rent_period smallint	max_rent_period smallint	avg_rent_period numeric	min_rent_rate numeric	max_rent_rate numeric	avg_rent_rate numeric	count_rows bigint
1	3	7	4.985	0.99	4.99	2.98	1000

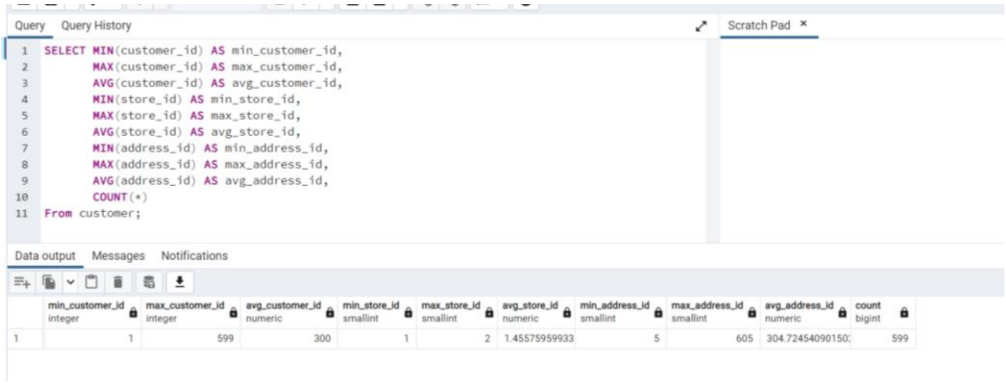
Summary for non-numeric columns in film table



```
SELECT mode() WITHIN GROUP (ORDER BY title)
       AS modal_title,
       mode() WITHIN GROUP (ORDER BY descripti
       AS modal_description,
       mode() WITHIN GROUP (ORDER BY rating)
       AS modal_rating,
       COUNT(*) AS count_rows
FROM film;
```

	modal_title character varying	modal_description text	modal_rating mpaa_rating	count_rows bigint
1	Academy Dinosaur	A Action-Packed C...	PG-13	1000

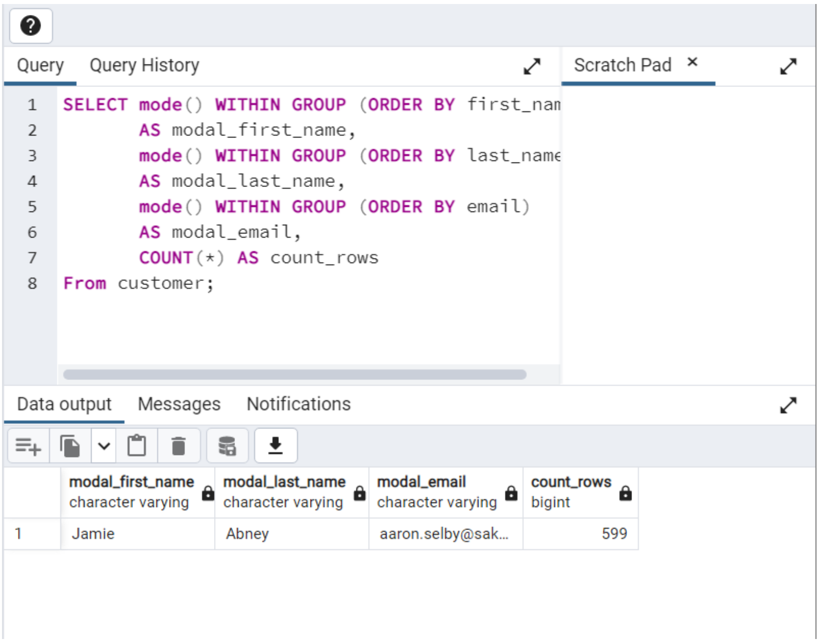
Summary for numeric columns in customer table



```
SELECT MIN(customer_id) AS min_customer_id,
       MAX(customer_id) AS max_customer_id,
       AVG(customer_id) AS avg_customer_id,
       MIN(store_id) AS min_store_id,
       MAX(store_id) AS max_store_id,
       AVG(store_id) AS avg_store_id,
       MIN(address_id) AS min_address_id,
       MAX(address_id) AS max_address_id,
       AVG(address_id) AS avg_address_id,
       COUNT(*)
From customer;
```

	min_customer_id integer	max_customer_id integer	avg_customer_id numeric	min_store_id smallint	max_store_id smallint	avg_store_id numeric	min_address_id smallint	max_address_id smallint	avg_address_id numeric	count bigint
1	1	599	300	1	2	1.45575959933	5	605	304.72454090150	599

Summary of non-numeric columns in customer table



```
SELECT mode() WITHIN GROUP (ORDER BY first_nam
       AS modal_first_name,
       mode() WITHIN GROUP (ORDER BY last_name
       AS modal_last_name,
       mode() WITHIN GROUP (ORDER BY email)
       AS modal_email,
       COUNT(*) AS count_rows
From customer;
```

	modal_first_name character varying	modal_last_name character varying	modal_email character varying	count_rows bigint
1	Jamie	Abney	aaron.selby@sak...	599

3. **Reflect on your work:** Back in Achievement 1 you learned about data profiling in Excel. Based on your previous experience, which tool (Excel or SQL) do you think is more effective for data profiling, and why? Consider their respective functions, ease of use, and speed. Write a short paragraph in the running document that you have started.

   Excel works best with smaller data, while using pivot tables are easy, it is harder where there is a massive about of data. SQL is easier to manipulate with large data, which is also faster. You can also answer specific questions, more detailed questions and with the right query answers pop up much quicker.