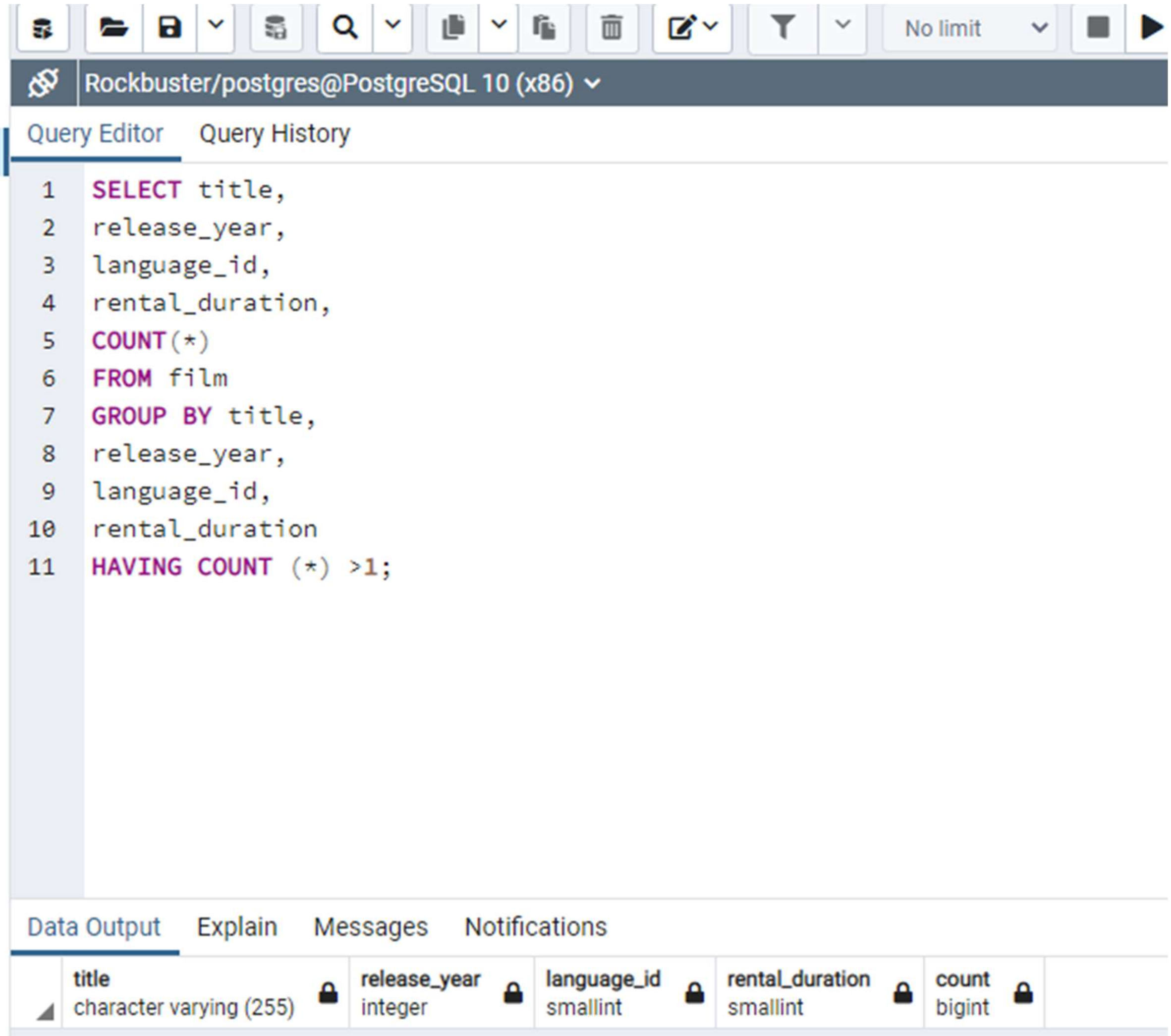
Summarizing & Cleaning Data in SQL

Question 1

Film Table

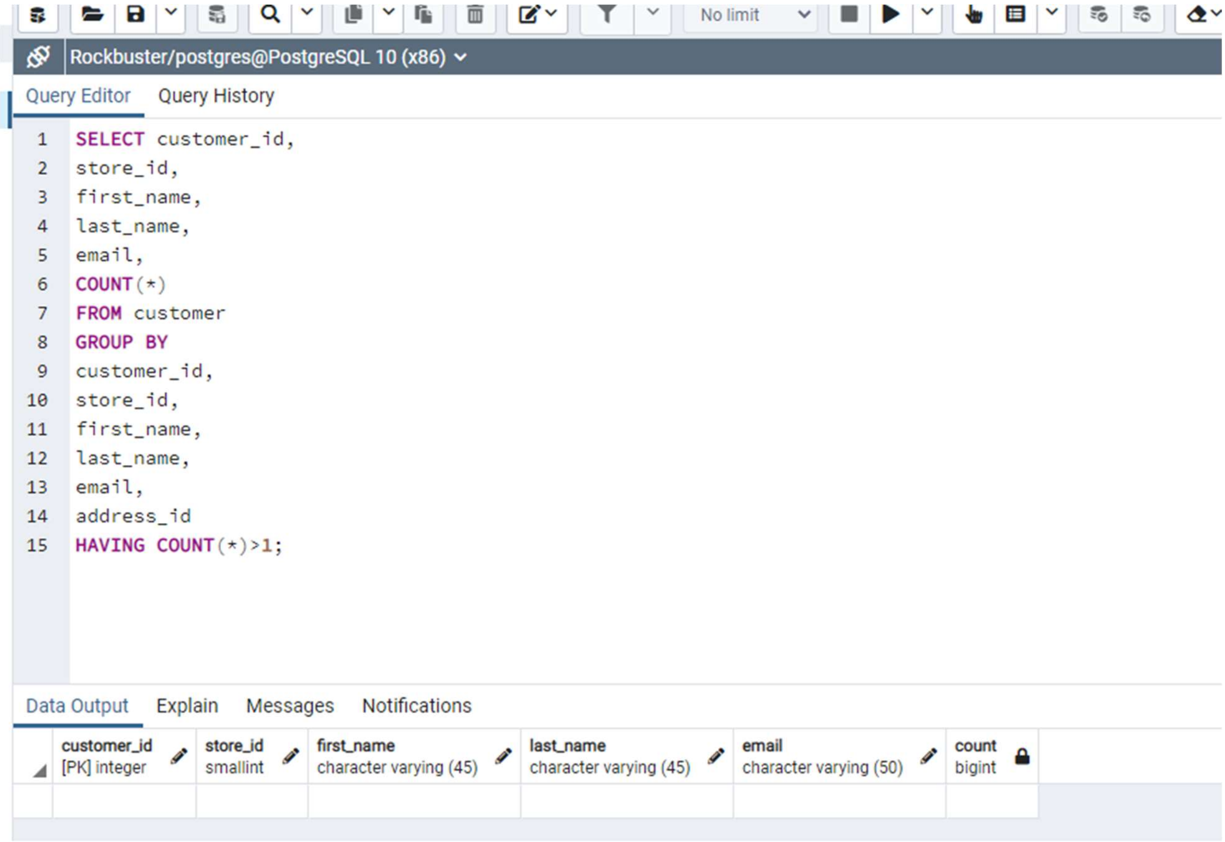


Rockbuster/postgres@PostgreSQL 10 (x86)

```sql
SELECT title,
release_year,
language_id,
rental_duration,
COUNT(*)
FROM film
GROUP BY title,
release_year,
language_id,
rental_duration
HAVING COUNT (*) >1;
```

title	release_year	language_id	rental_duration	count
character varying (255)	integer	smallint	smallint	bigint

Customer Table



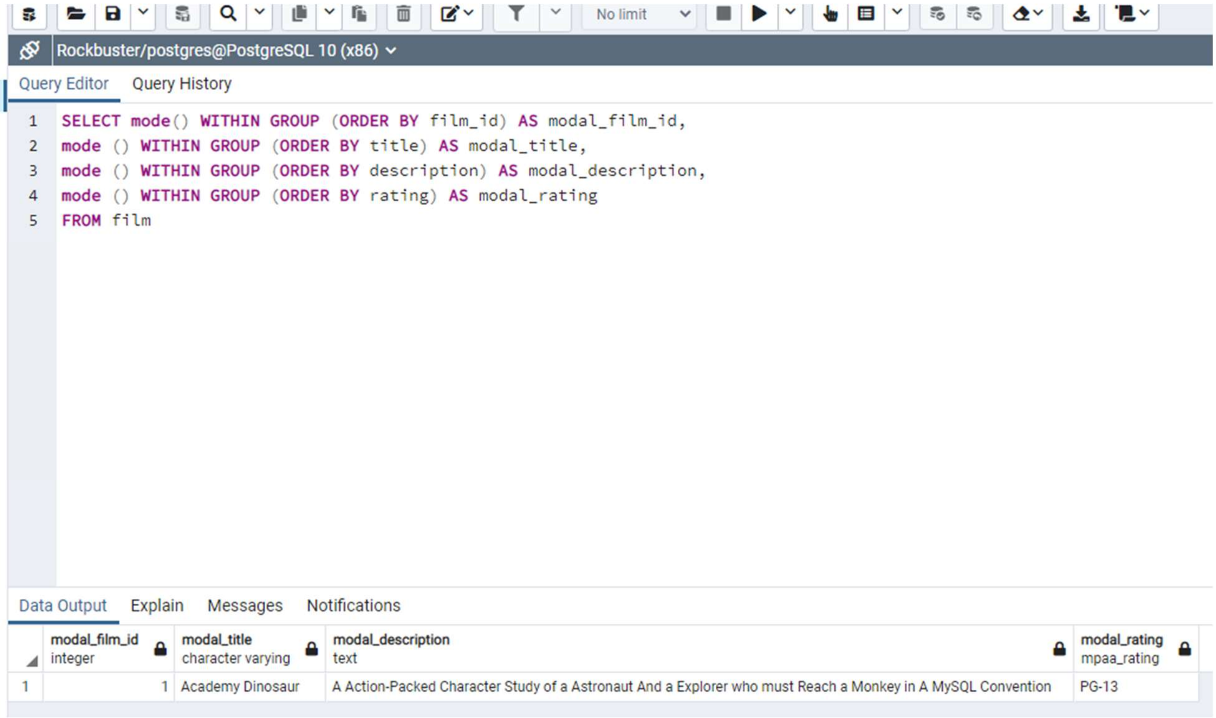
```sql
SELECT customer_id,
store_id,
first_name,
last_name,
email,
COUNT(*)
FROM customer
GROUP BY
customer_id,
store_id,
first_name,
last_name,
email,
address_id
HAVING COUNT(*)>1;
```

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

There are no duplicates in the above.

Question 2

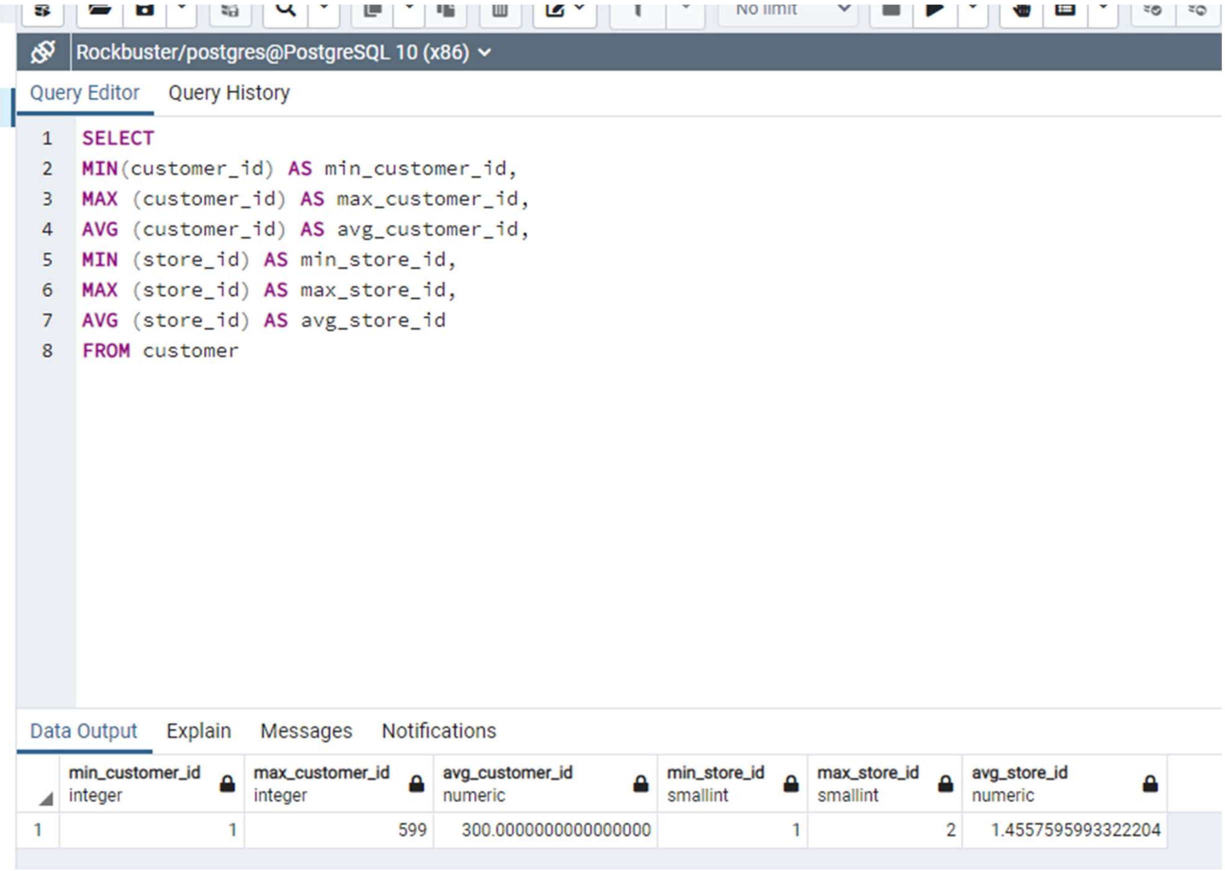
Film table with non-numerical columns



```sql
SELECT mode() WITHIN GROUP (ORDER BY film_id) AS modal_film_id,
mode () WITHIN GROUP (ORDER BY title) AS modal_title,
mode () WITHIN GROUP (ORDER BY description) AS modal_description,
mode () WITHIN GROUP (ORDER BY rating) AS modal_rating
FROM film
```

	modal_film_id integer	modal_title character varying	modal_description text	modal_rating mpaa_rating
1	1	Academy Dinosaur	A Action-Packed Character Study of a Astronaut And a Explorer who must Reach a Monkey in A MySQL Convention	PG-13

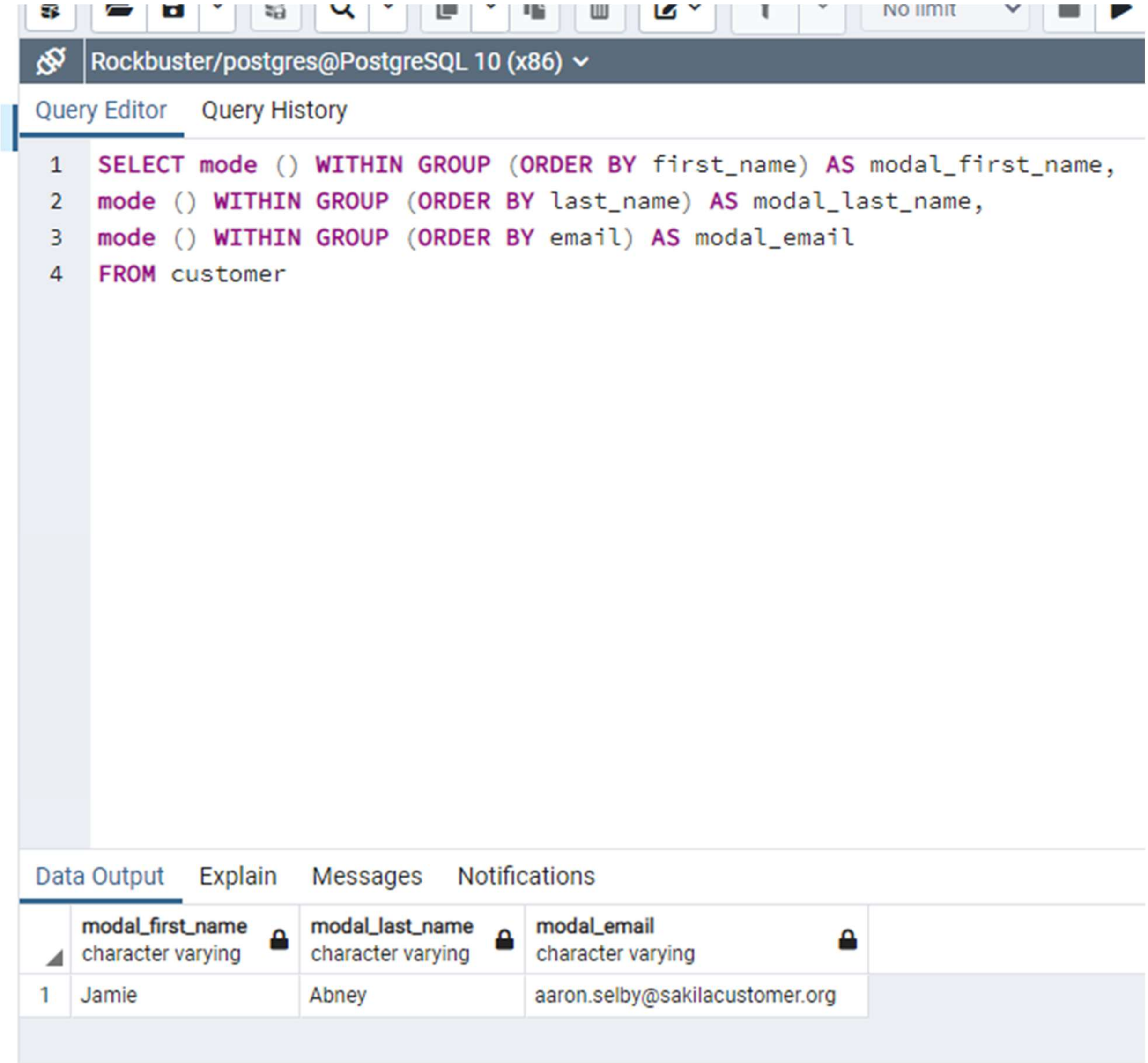
Customer table with numerical columns



```sql
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
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
1	1	599	300.0000000000000000	1	2	1.4557595993322204

Customer table with non-numerical columns



```
SELECT mode () WITHIN GROUP (ORDER BY first_name) AS modal_first_name,
mode () WITHIN GROUP (ORDER BY last_name) AS modal_last_name,
mode () WITHIN GROUP (ORDER BY email) AS modal_email
FROM customer
```

	modal_first_name character varying	modal_last_name character varying	modal_email character varying
1	Jamie	Abney	aaron.selby@sakilacustomer.org

Question 3

When it comes to Data profiling I think it is much each to use SQL dependent on the size of the Database.