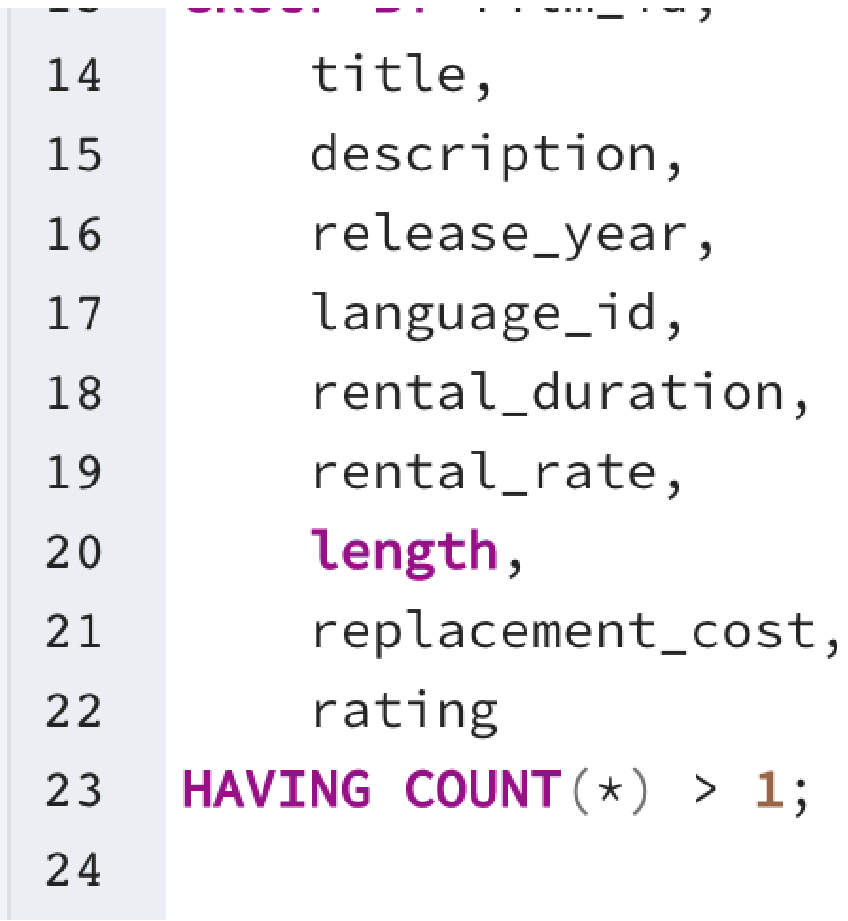
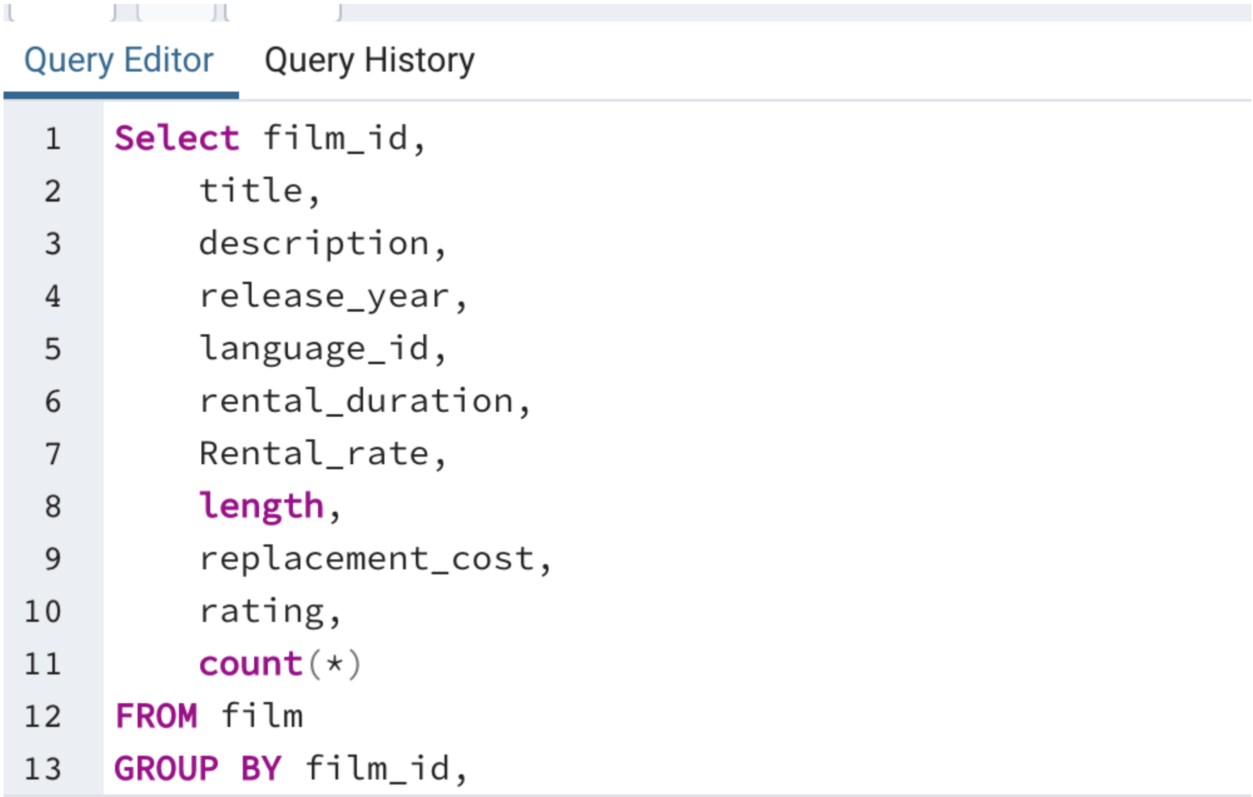
3.6 Summarizing & Cleaning Data in SQL

1. Check for dirty data:

Film table:

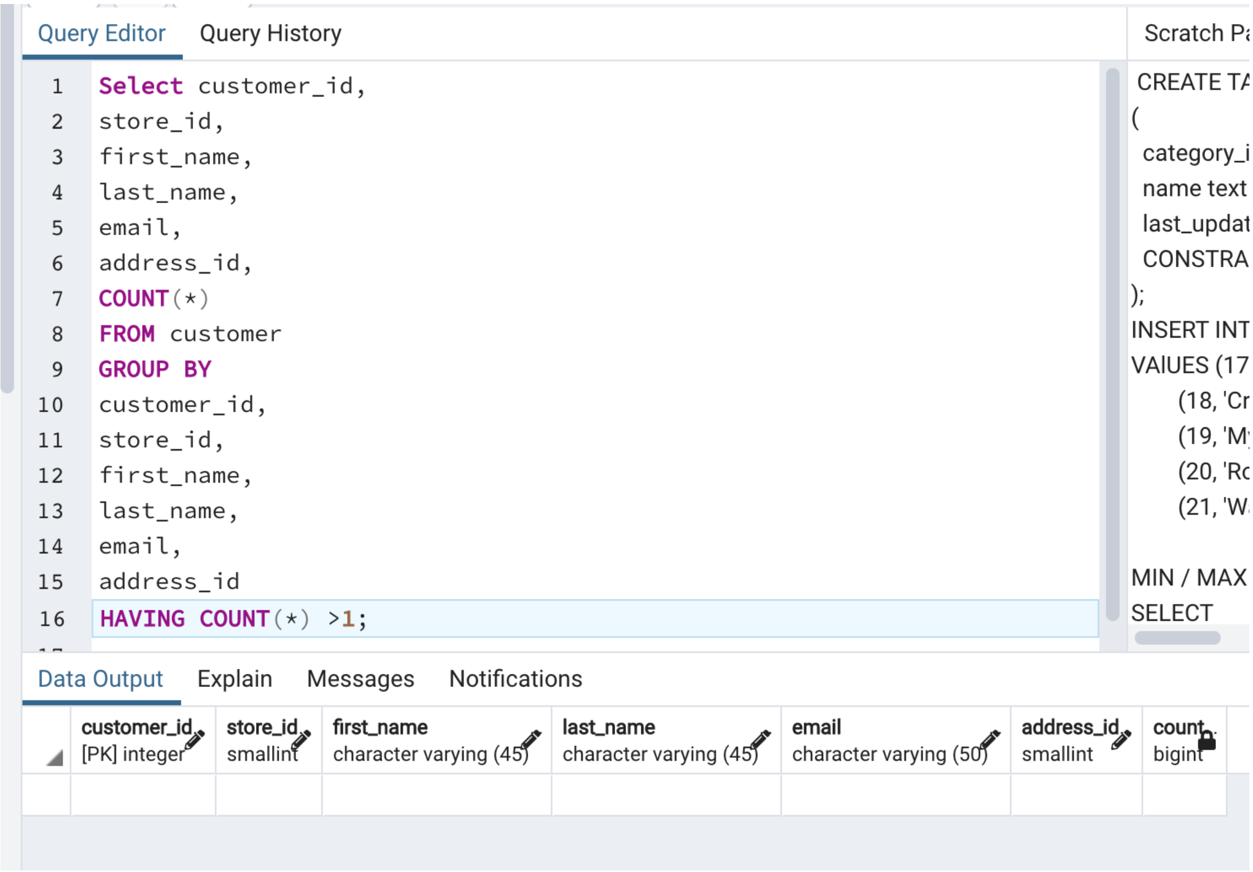


```sql
GROUP BY film_id,
    title,
    description,
    release_year,
    language_id,
    rental_duration,
    rental_rate,
    length,
    replacement_cost,
    rating
HAVING COUNT(*) > 1;
```



```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,
```

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;
```

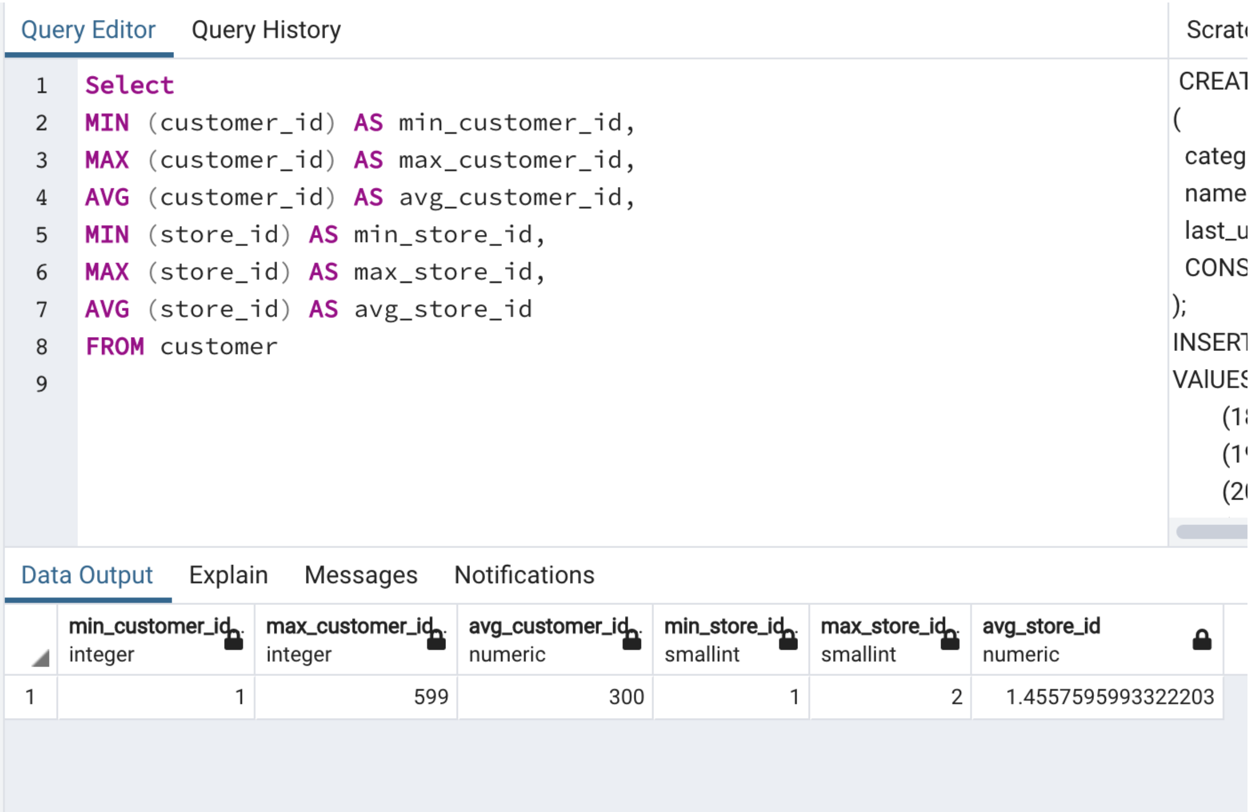
```
CREATE TA
(
 category_i
 name text
 last_updat
 CONSTRA
);
INSERT INT
VAlUES (17
       (18, 'Cr
       (19, 'M
       (20, 'Rc
       (21, 'W

MIN / MAX
SELECT
```

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

To clean the tables, I would create a new table in the view format and delete any duplicate records found using the having count above.

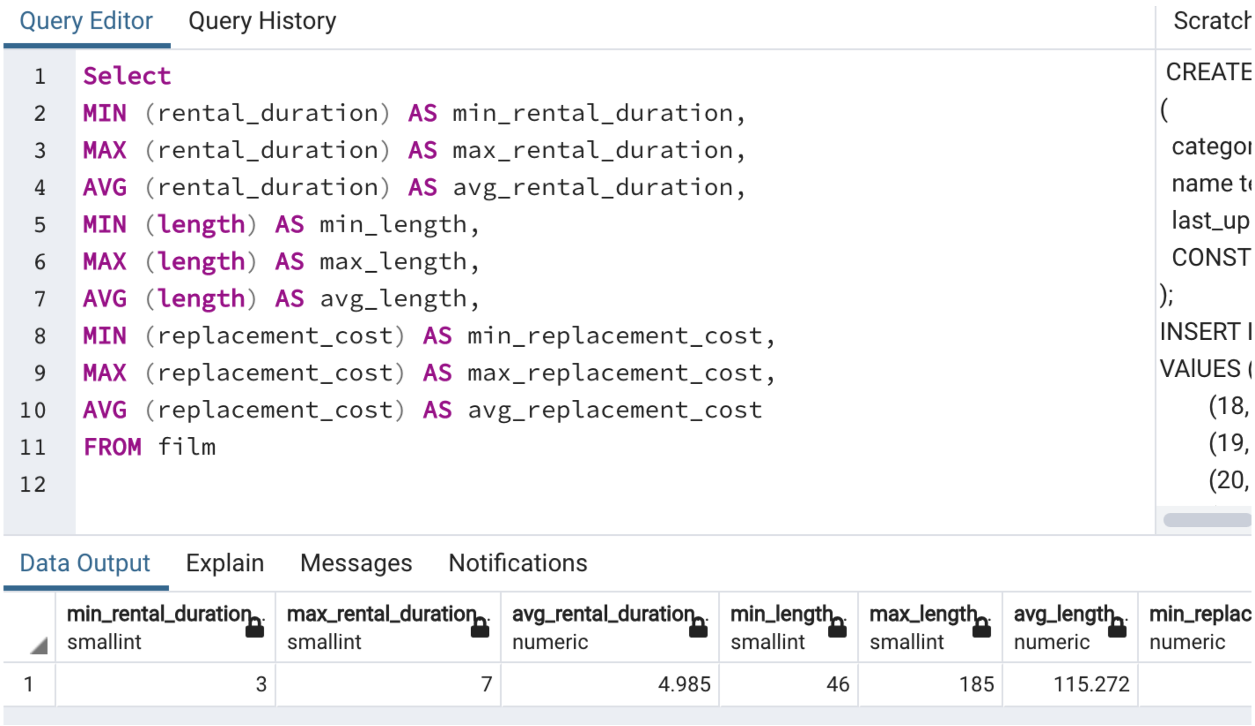
2. Summarize your data:



```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
```

```
CREAT
(
 categ
 name
 last_u
 CONS
);
INSERT
VAlUES
    (1
    (1
    (2
```

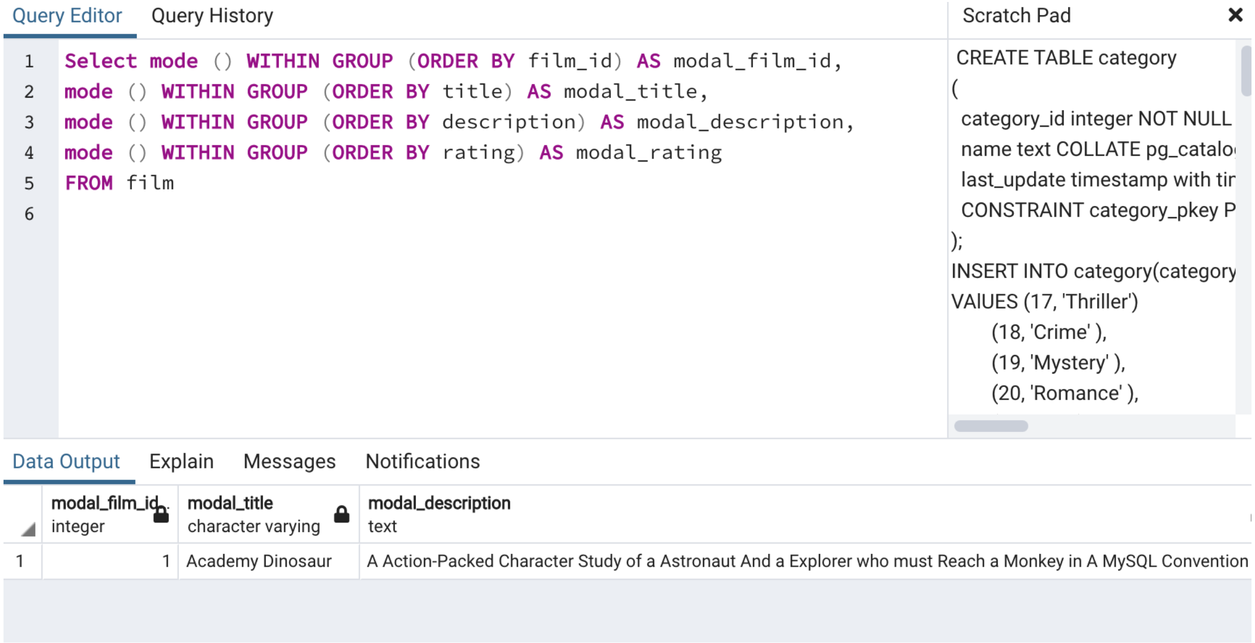
	min_customer_id integer	max_customer_id integer	avg_customer_id numeric	min_store_id smallint	max_store_id smallint	avg_store_id numeric
1	1	599	300	1	2	1.4557595993322203



```sql
Select
MIN (rental_duration) AS min_rental_duration,
MAX (rental_duration) AS max_rental_duration,
AVG (rental_duration) AS avg_rental_duration,
MIN (length) AS min_length,
MAX (length) AS max_length,
AVG (length) AS avg_length,
MIN (replacement_cost) AS min_replacement_cost,
MAX (replacement_cost) AS max_replacement_cost,
AVG (replacement_cost) AS avg_replacement_cost
FROM film
```

```
CREATE
(
 categor
 name t
 last_up
 CONST
);
INSERT I
VAlUES (
    (18,
    (19,
    (20,
```

	min_rental_duration smallint	max_rental_duration smallint	avg_rental_duration numeric	min_length smallint	max_length smallint	avg_length numeric	min_replac numeric
1	3	7	4.985	46	185	115.272	

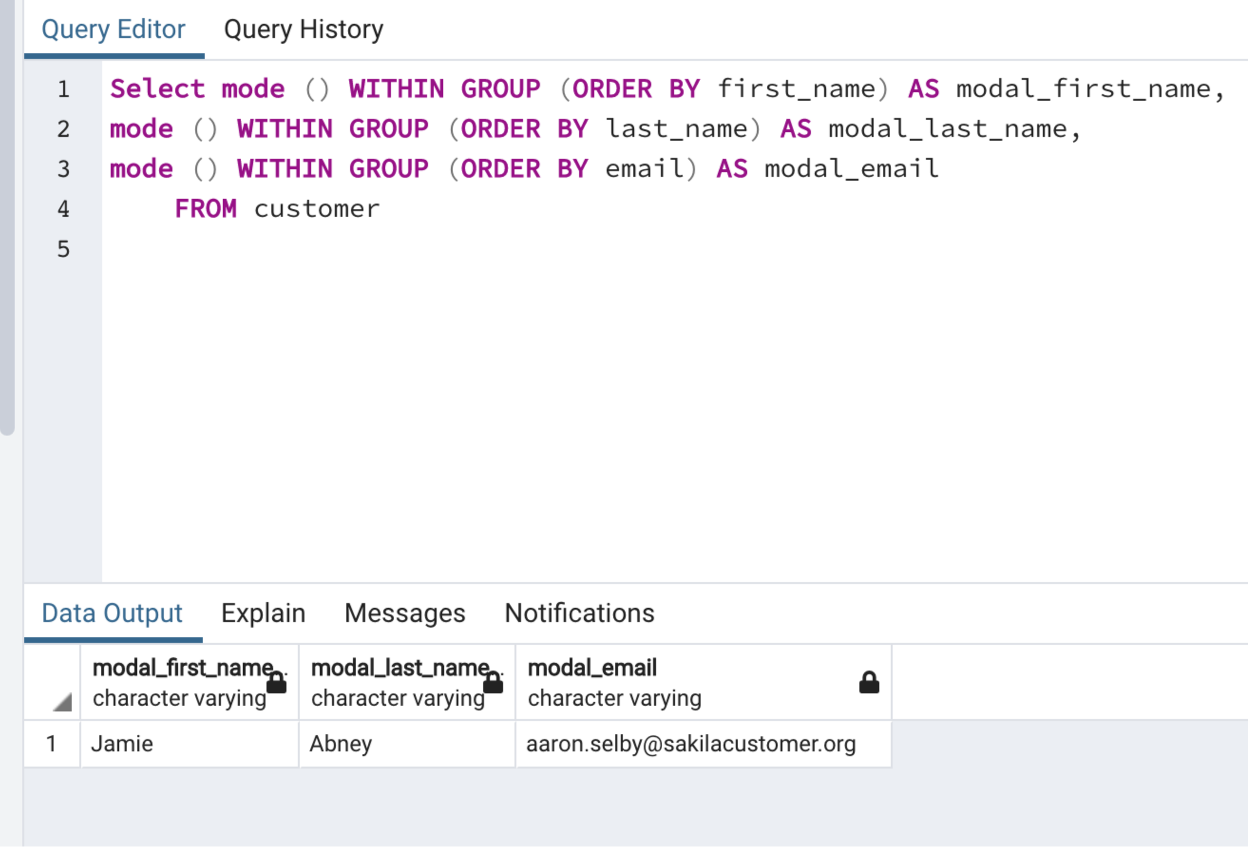


```sql
Select mode () WITHIN GROUP (ORDER BY film_id) AS modal_film_id,
mode () WITHIN GROUP (ORDER BY title) AS modal_title,
mode () WITHIN GROUP (ORDER BY description) AS modal_description,
mode () WITHIN GROUP (ORDER BY rating) AS modal_rating
FROM film

```

```
CREATE TABLE category
(
 category_id integer NOT NULL
 name text COLLATE pg_catalo
 last_update timestamp with tir
 CONSTRAINT category_pkey P
);
INSERT INTO category(category
VALUES (17, 'Thriller')
       (18, 'Crime' ),
       (19, 'Mystery' ),
       (20, 'Romance' ),
```

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



```sql
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

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.

Larger databases are best suited for SQL and smaller data sets are best suited for excel. I think the speed depends on which program you are more comfortable with and how large your data set is. SQL takes me a little longer due to my lack of experience with it, although I can see how fast it could be once you do have to background in it.