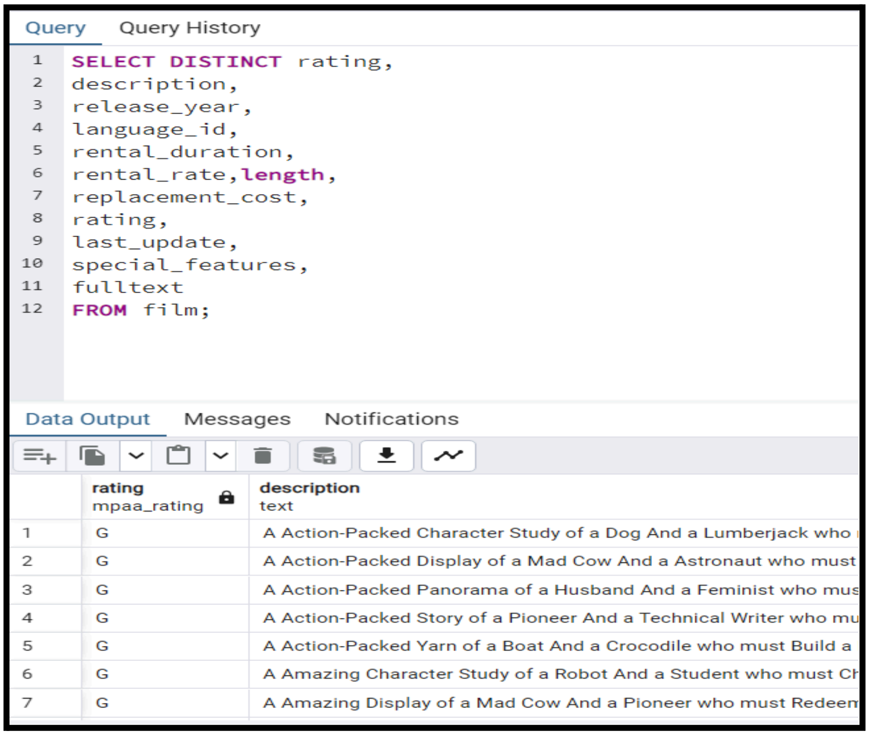
3.6 Answer Sheet
Katherine Lecce

1.

Non uniformed Data

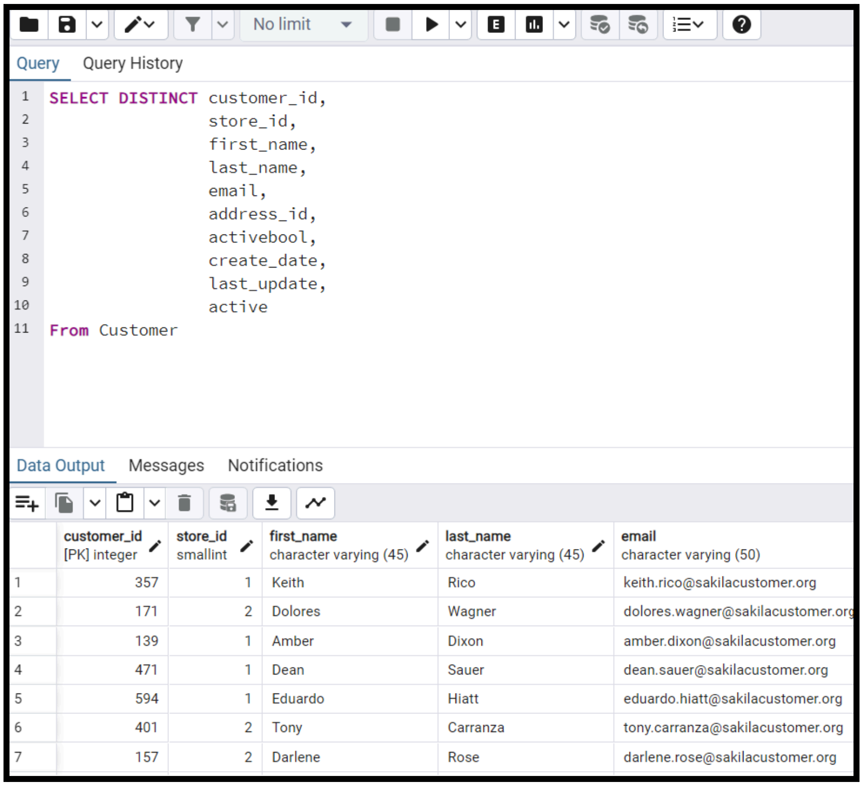
→ Film Table:



```sql
SELECT DISTINCT rating,
description,
release_year,
language_id,
rental_duration,
rental_rate,length,
replacement_cost,
rating,
last_update,
special_features,
fulltext
FROM film;
```

	rating mpaa_rating	description text
1	G	A Action-Packed Character Study of a Dog And a Lumberjack who
2	G	A Action-Packed Display of a Mad Cow And a Astronaut who must
3	G	A Action-Packed Panorama of a Husband And a Feminist who mus
4	G	A Action-Packed Story of a Pioneer And a Technical Writer who mu
5	G	A Action-Packed Yarn of a Boat And a Crocodile who must Build a
6	G	A Amazing Character Study of a Robot And a Student who must Ch
7	G	A Amazing Display of a Mad Cow And a Pioneer who must Redeem

→ Customer Table:

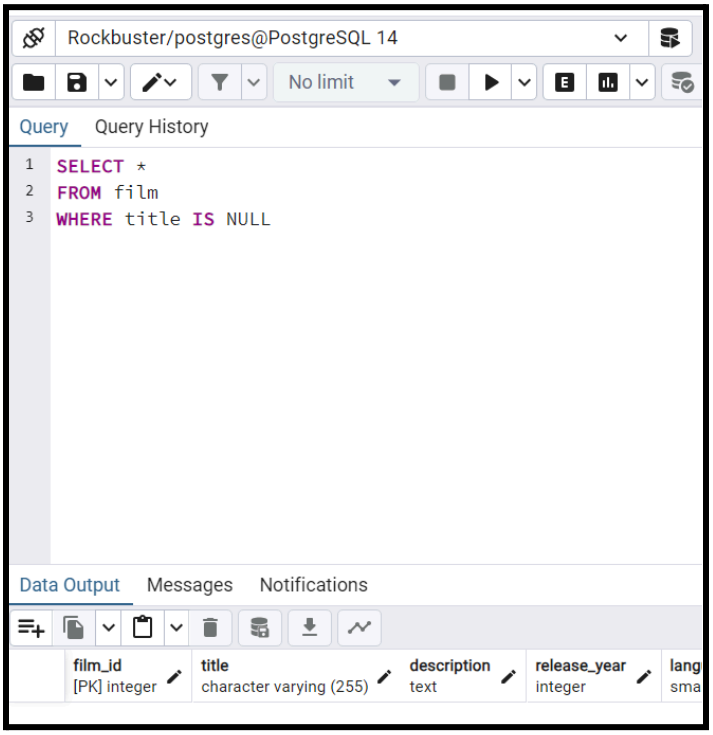


```sql
SELECT DISTINCT customer_id,
                store_id,
                first_name,
                last_name,
                email,
                address_id,
                activebool,
                create_date,
                last_update,
                active
From Customer
```

	customer_id [PK] integer	store_id smallint	first_name character varying (45)	last_name character varying (45)	email character varying (50)
1	357	1	Keith	Rico	keith.rico@sakilacustomer.org
2	171	2	Dolores	Wagner	dolores.wagner@sakilacustomer.org
3	139	1	Amber	Dixon	amber.dixon@sakilacustomer.org
4	471	1	Dean	Sauer	dean.sauer@sakilacustomer.org
5	594	1	Eduardo	Hiatt	eduardo.hiatt@sakilacustomer.org
6	401	2	Tony	Carranza	tony.carranza@sakilacustomer.org
7	157	2	Darlene	Rose	darlene.rose@sakilacustomer.org

Missing Data:

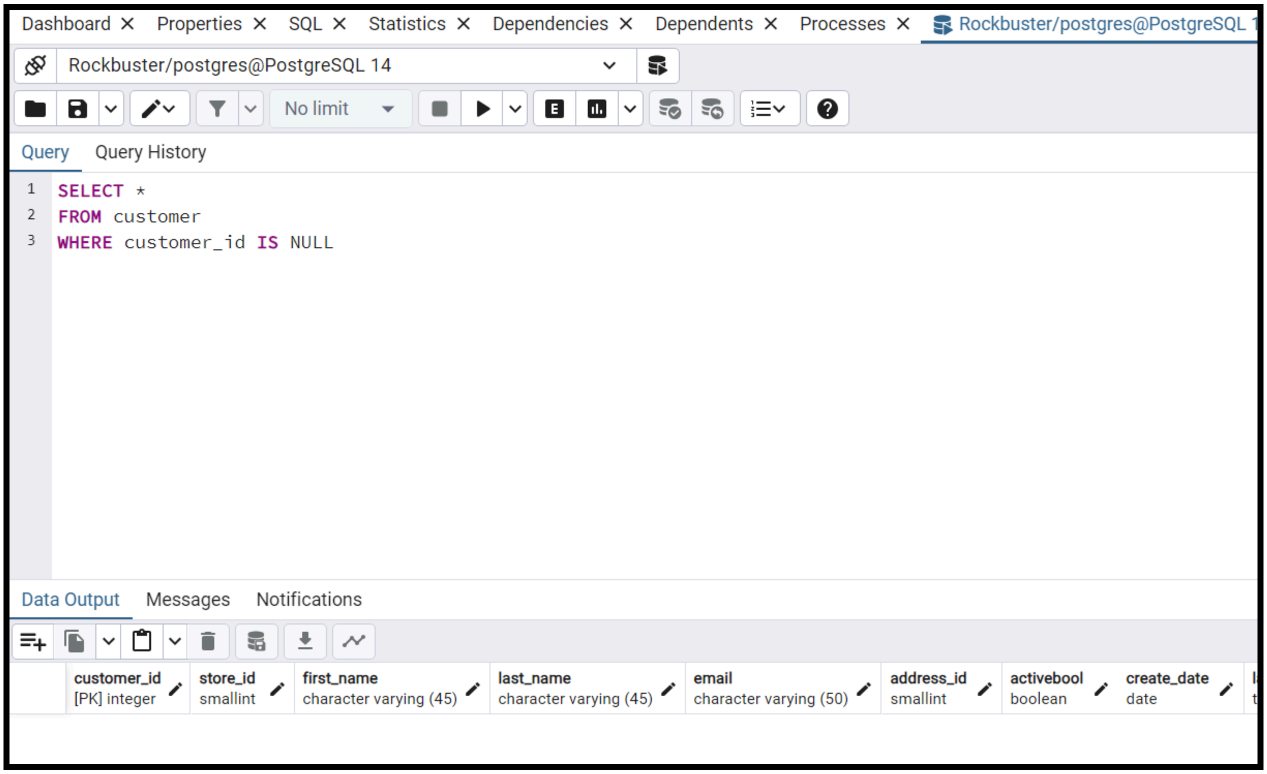
Film Data →



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

film_id [PK] integer	title character varying (255)	description text	release_year integer	lang sma

Customer Table →

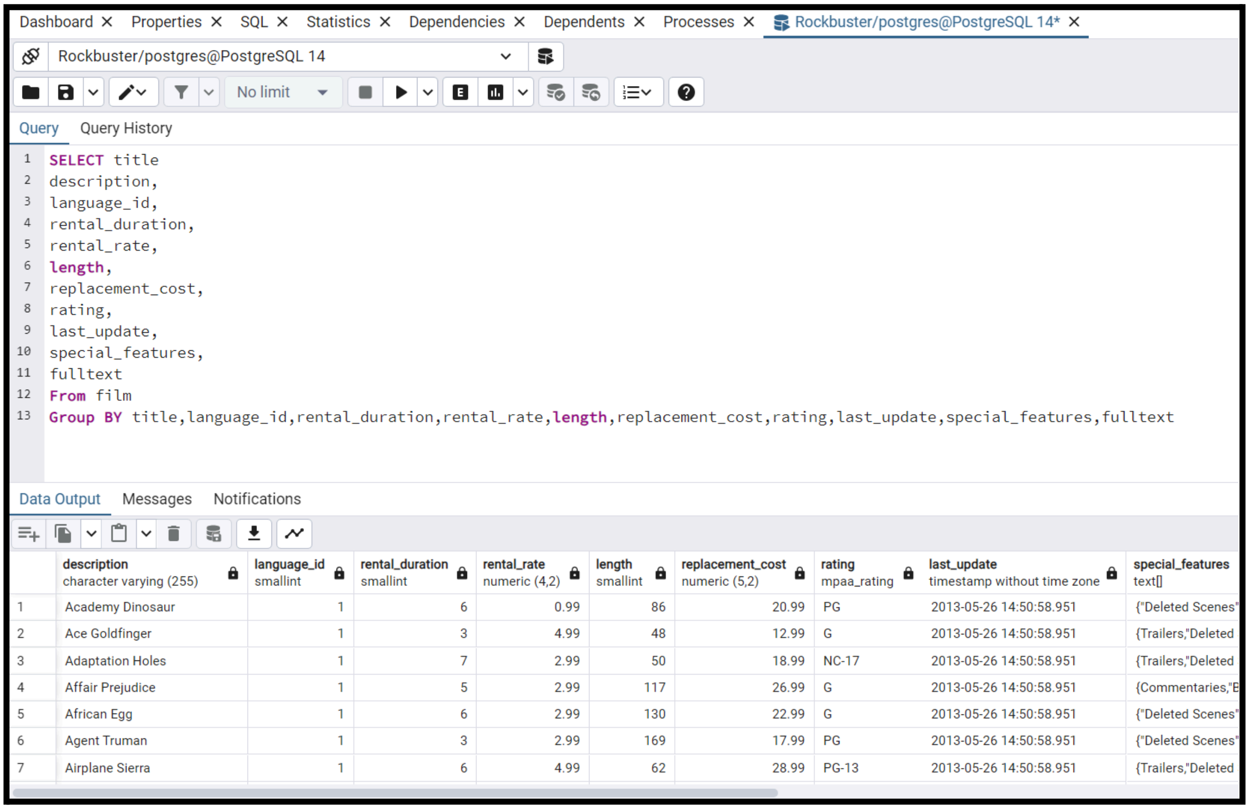


```
SELECT *
FROM customer
WHERE customer_id IS NULL
```

customer_id [PK] integer	store_id smallint	first_name character varying (45)	last_name character varying (45)	email character varying (50)	address_id smallint	activebool boolean	create_date date	l t

Duplicate Data:

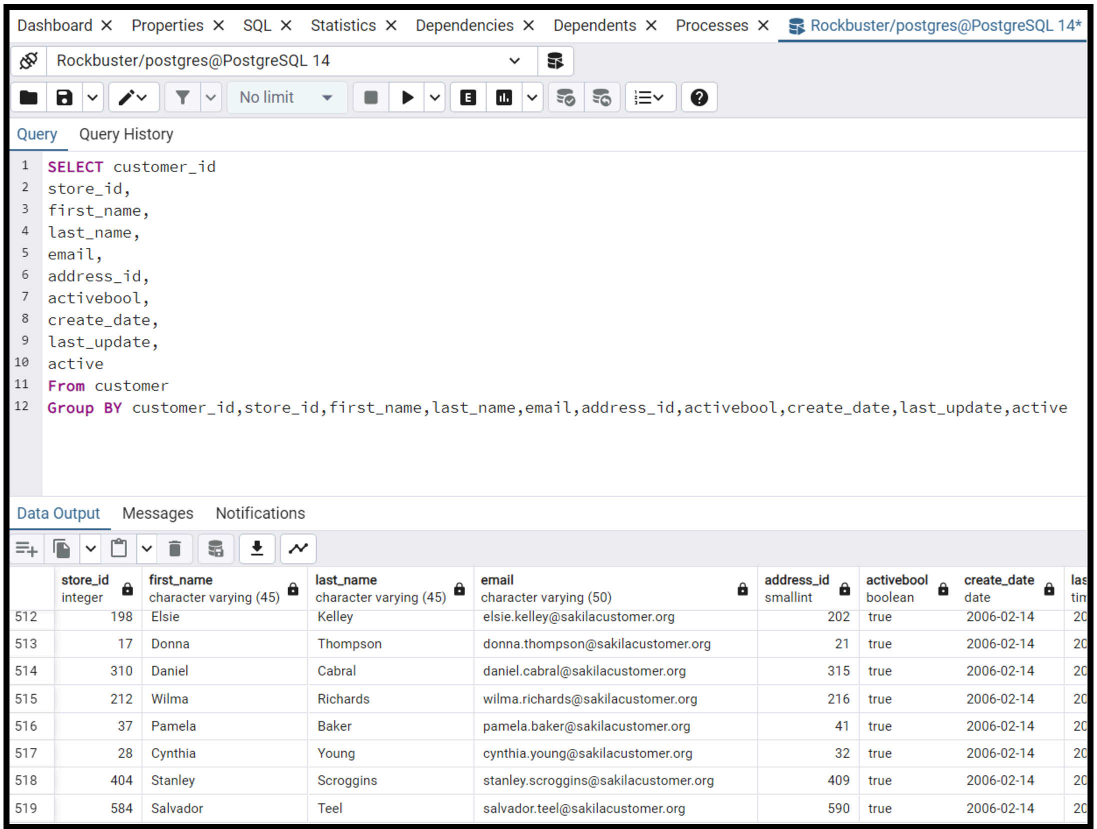
Film Table:



```sql
SELECT title
description,
language_id,
rental_duration,
rental_rate,
length,
replacement_cost,
rating,
last_update,
special_features,
fulltext
From film
Group BY title,language_id,rental_duration,rental_rate,length,replacement_cost,rating,last_update,special_features,fulltext
```

	description	language_id	rental_duration	rental_rate	length	replacement_cost	rating	last_update	special_features
	character varying (255)	smallint	smallint	numeric (4,2)	smallint	numeric (5,2)	mpaa_rating	timestamp without time zone	text[]
1	Academy Dinosaur	1	6	0.99	86	20.99	PG	2013-05-26 14:50:58.951	{"Deleted Scenes"
2	Ace Goldfinger	1	3	4.99	48	12.99	G	2013-05-26 14:50:58.951	{Trailers,"Deleted
3	Adaptation Holes	1	7	2.99	50	18.99	NC-17	2013-05-26 14:50:58.951	{Trailers,"Deleted
4	Affair Prejudice	1	5	2.99	117	26.99	G	2013-05-26 14:50:58.951	{Commentaries,"B
5	African Egg	1	6	2.99	130	22.99	G	2013-05-26 14:50:58.951	{"Deleted Scenes"
6	Agent Truman	1	3	2.99	169	17.99	PG	2013-05-26 14:50:58.951	{"Deleted Scenes"
7	Airplane Sierra	1	6	4.99	62	28.99	PG-13	2013-05-26 14:50:58.951	{Trailers,"Deleted

Customer Table:

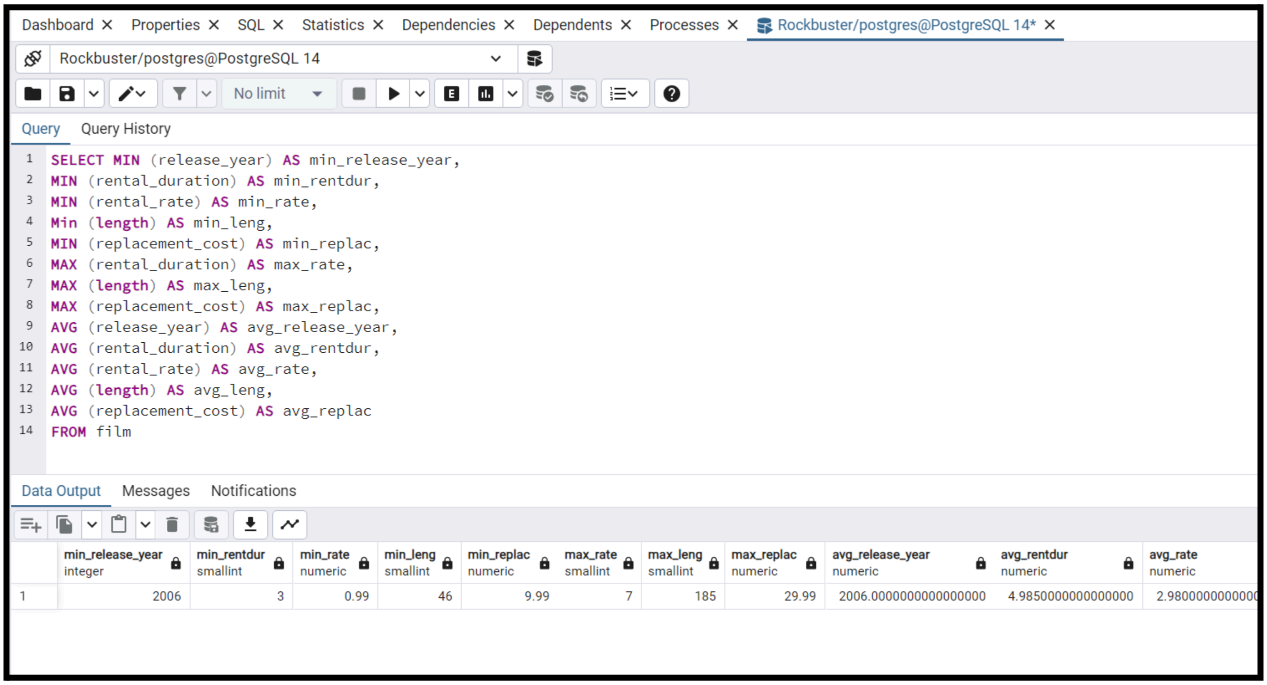


```sql
SELECT customer_id
store_id,
first_name,
last_name,
email,
address_id,
activebool,
create_date,
last_update,
active
From customer
Group BY customer_id,store_id,first_name,last_name,email,address_id,activebool,create_date,last_update,active
```

	store_id	first_name	last_name	email	address_id	activebool	create_date	las
	integer	character varying (45)	character varying (45)	character varying (50)	smallint	boolean	date	tin
512	198	Elsie	Kelley	elsie.kelley@sakilacustomer.org	202	true	2006-02-14	20
513	17	Donna	Thompson	donna.thompson@sakilacustomer.org	21	true	2006-02-14	20
514	310	Daniel	Cabral	daniel.cabral@sakilacustomer.org	315	true	2006-02-14	20
515	212	Wilma	Richards	wilma.richards@sakilacustomer.org	216	true	2006-02-14	20
516	37	Pamela	Baker	pamela.baker@sakilacustomer.org	41	true	2006-02-14	20
517	28	Cynthia	Young	cynthia.young@sakilacustomer.org	32	true	2006-02-14	20
518	404	Stanley	Scroggins	stanley.scroggins@sakilacustomer.org	409	true	2006-02-14	20
519	584	Salvador	Teel	salvador.teel@sakilacustomer.org	590	true	2006-02-14	20

2. Summarize your data:

Film Table: Numeric

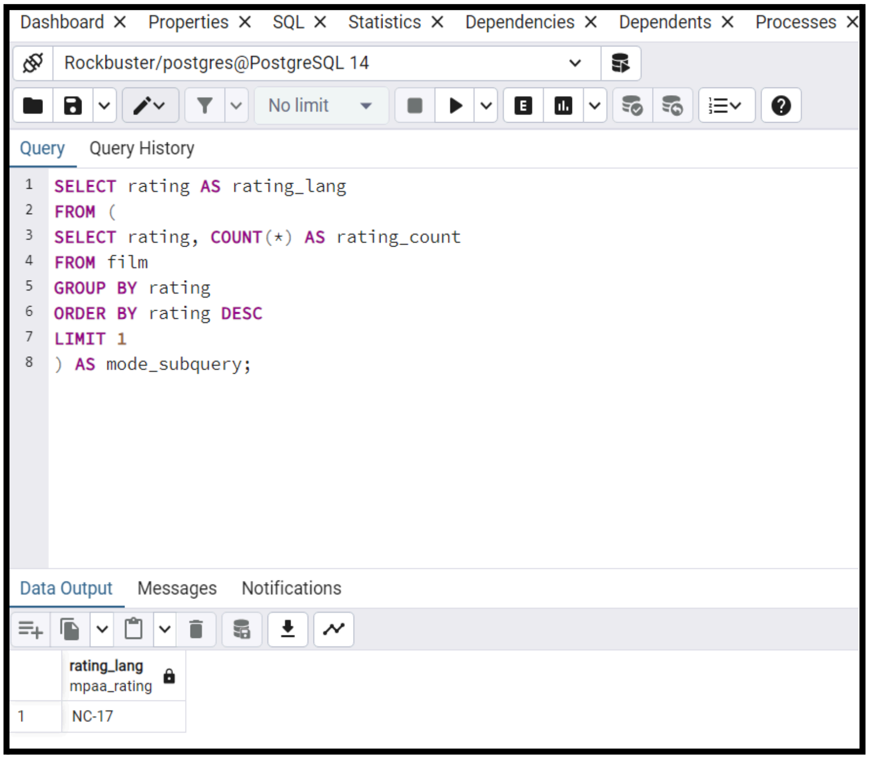


```sql
SELECT MIN (release_year) AS min_release_year,
MIN (rental_duration) AS min_rentdur,
MIN (rental_rate) AS min_rate,
Min (length) AS min_leng,
MIN (replacement_cost) AS min_replac,
MAX (rental_duration) AS max_rate,
MAX (length) AS max_leng,
MAX (replacement_cost) AS max_replac,
AVG (release_year) AS avg_release_year,
AVG (rental_duration) AS avg_rentdur,
AVG (rental_rate) AS avg_rate,
AVG (length) AS avg_leng,
AVG (replacement_cost) AS avg_replac
FROM film
```

	min_release_year	min_rentdur	min_rate	min_leng	min_replac	max_rate	max_leng	max_replac	avg_release_year	avg_rentdur	avg_rate
	integer	smallint	numeric	smallint	numeric	smallint	smallint	numeric	numeric	numeric	numeric
1	2006	3	0.99	46	9.99	7	185	29.99	2006.0000000000000000	4.9850000000000000	2.980000000000000

Film Table - Non Numeric

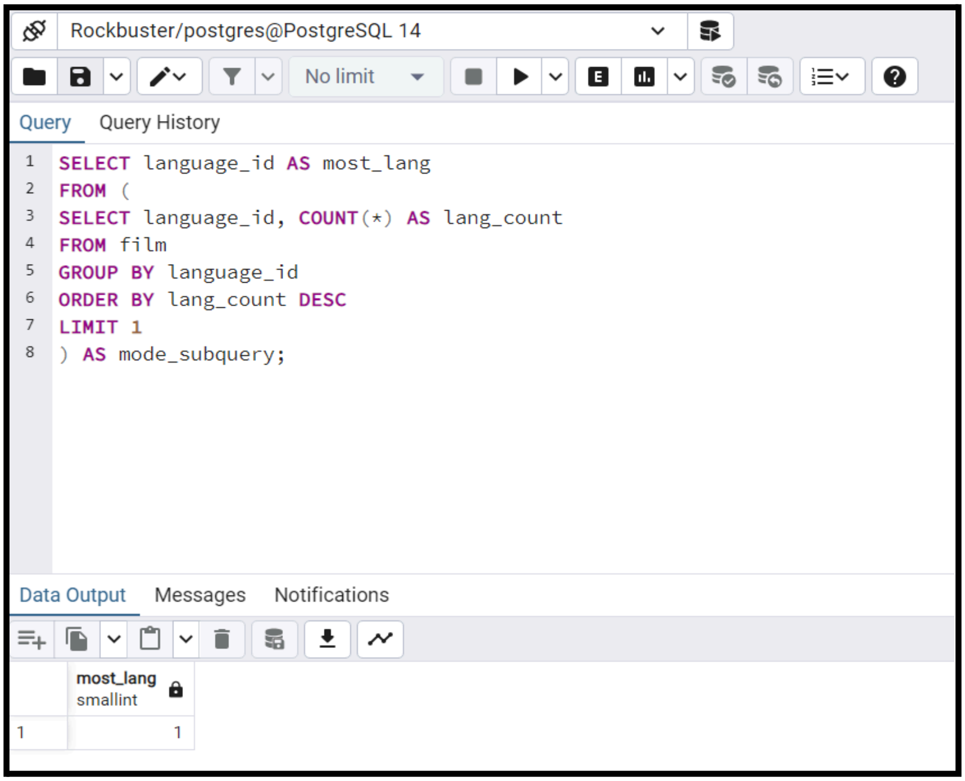
Movie Rating



```sql
SELECT rating AS rating_lang
FROM (
SELECT rating, COUNT(*) AS rating_count
FROM film
GROUP BY rating
ORDER BY rating DESC
LIMIT 1
) AS mode_subquery;
```

	rating_lang
	mpaa_rating
1	NC-17

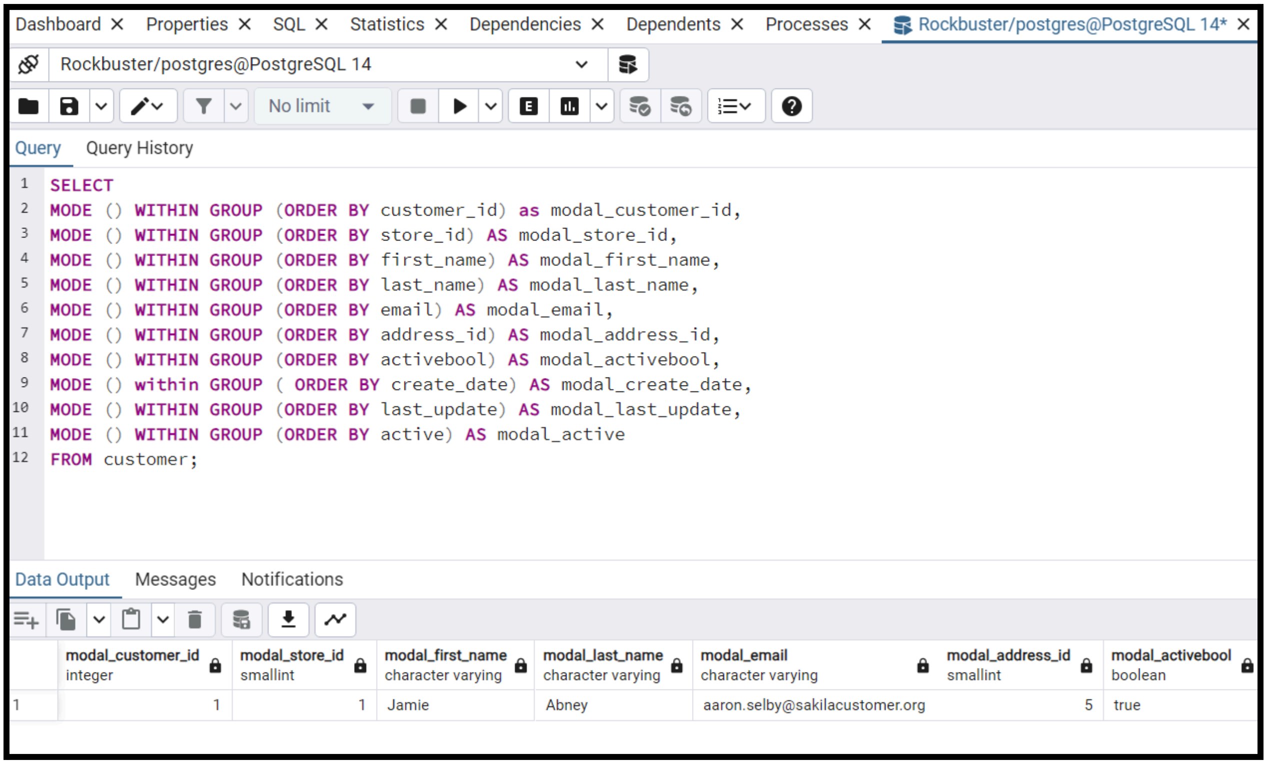
Language



```
SELECT language_id AS most_lang
FROM (
SELECT language_id, COUNT(*) AS lang_count
FROM film
GROUP BY language_id
ORDER BY lang_count DESC
LIMIT 1
) AS mode_subquery;
```

	most_lang smallint
1	1

Customer Table:



```
SELECT
MODE () WITHIN GROUP (ORDER BY customer_id) as modal_customer_id,
MODE () WITHIN GROUP (ORDER BY store_id) AS modal_store_id,
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,
MODE () WITHIN GROUP (ORDER BY address_id) AS modal_address_id,
MODE () WITHIN GROUP (ORDER BY activebool) AS modal_activebool,
MODE () within GROUP ( ORDER BY create_date) AS modal_create_date,
MODE () WITHIN GROUP (ORDER BY last_update) AS modal_last_update,
MODE () WITHIN GROUP (ORDER BY active) AS modal_active
FROM customer;
```

	modal_customer_id integer	modal_store_id smallint	modal_first_name character varying	modal_last_name character varying	modal_email character varying	modal_address_id smallint	modal_activebool boolean
1	1	1	Jamie	Abney	aaron.selby@sakilacustomer.org	5	true

1. How would you clean each set of data?

A. Missing Data: Save Data, then you'll want to use the 'SELECT' statement to omit any columns you'd like to skip over when checking this data (make note in query). You can also use the 'Impute' or fill in where there are missing sections of the data.

B. Duplicate Data: First and foremost, backup your data before starting to edit the query. Then use the 'Group By' function to identify the records that you want to check in a specific column of the data. Then, if duplicates are found, use the 'Delete' statement to remove any information in the set that is invalid or 'Merge' multiple statements into one single record by using the 'update' function to join the multiple sections you'd like to combine.

C. Non-Uniformed Data: After saving your data, use the GROUP By and DISTINCT functions to check a few areas in your data to see if your outputs have any inconsistencies. If you you find areas that need to be uniformed use the

UPDATE (name of data set)
SET (section you'd like the format to change FROM) = (section name you'd like the area to be formatted TO)
WHERE (column) IN (section you'd like the format to change FROM)

3 - 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.

Because both systems have different functions it's important to first consider how large the data sets you're looking at are. If this is a task that only one or two people are looking at altering any type of data within the spreadsheet. While SQL is more practical for large amounts of data across a range of multiple operations and avenues within a business, company, or program. SQL is better for tasks that require the data to alter and change into migrating into one system format rather than keeping it in a spreadsheet format. Personally, because I am more familiar with spreadsheets, it is more comfortable and easier for me to profile data when Excel and more familiar with using functions such as filter, pivot tables, and highlighting to more clearly see where information might need to be altered.