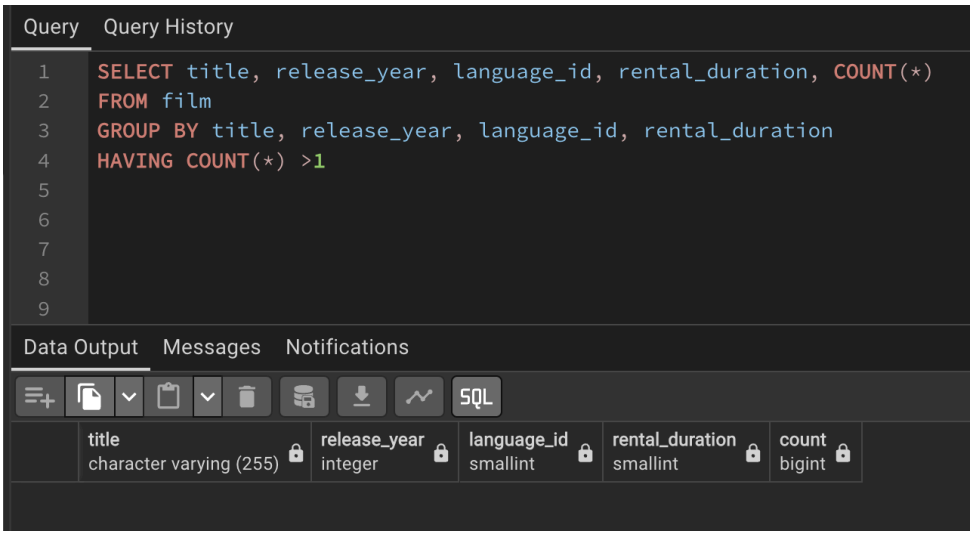
1. ## Check for and clean dirty data

Film table

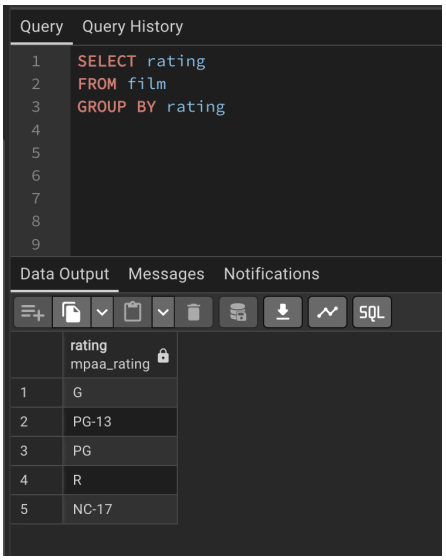
1. Duplicate data



No duplicates were found in the data. There are two ways to fix duplicates.

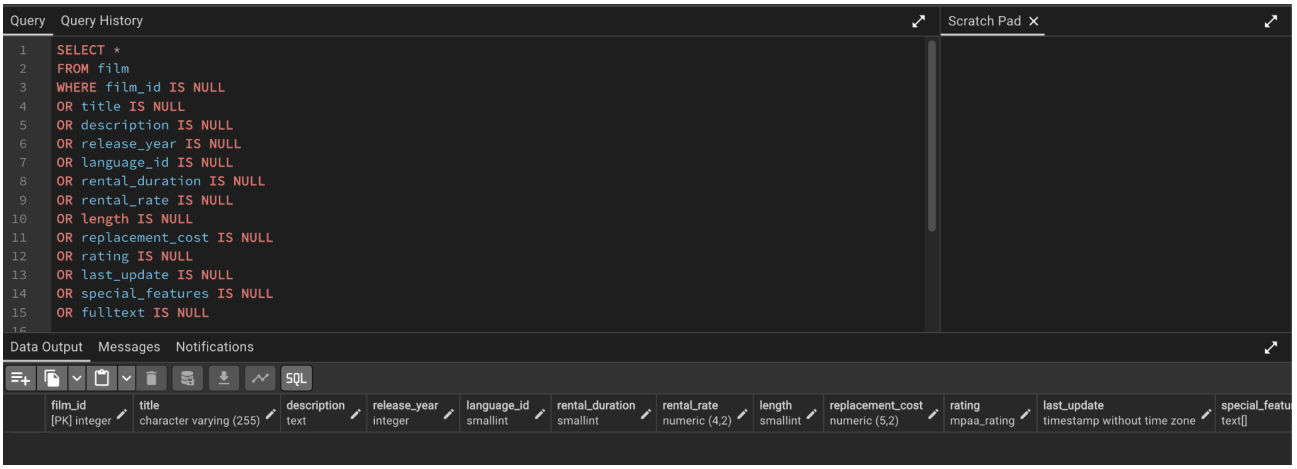
- Create a view table where we will select only unique records.
- Delete the duplicate records from the table or view.

2. Non-Uniform data



There is no non-uniform data found. To fix non-uniform data, I will use UPDATE command to set the rating consistent.

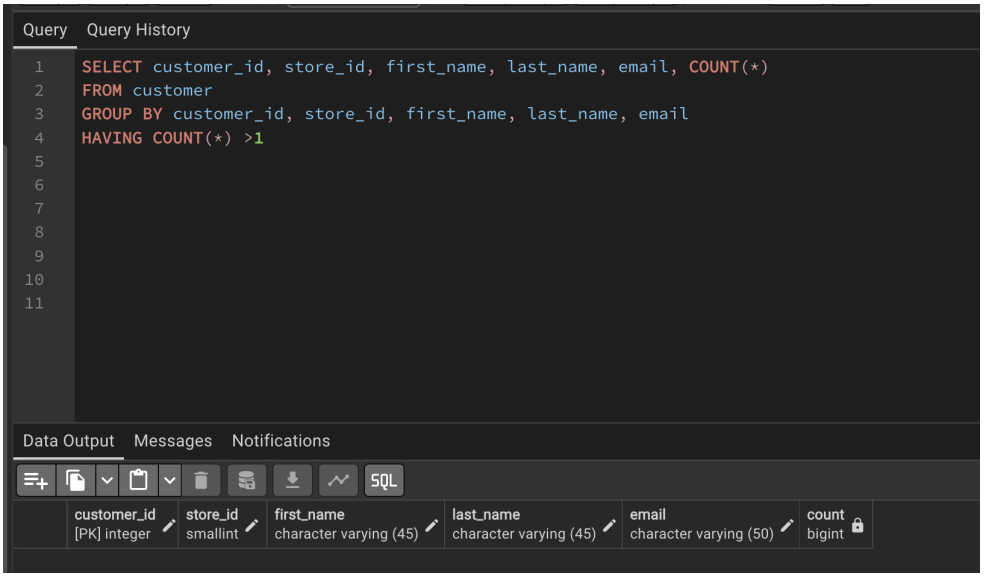
3. Missing data



No missing data was found. If there is data missing then we can simply ignore the columns with the high percentage of missing values. Another option is to impute the values using statistical methods. We can input the estimate or average values.

Customer table

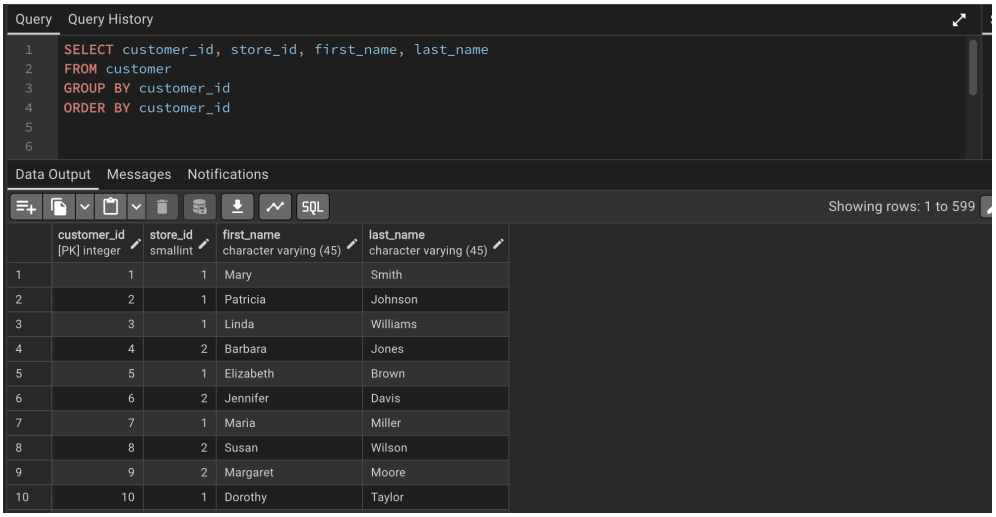
1. Duplicate data



No duplicates were found in the data. There are two ways to fix duplicates.

- Create a view table where we will select only unique records.
- Delete the duplicate records from the table or view.

2. Non-Uniform data

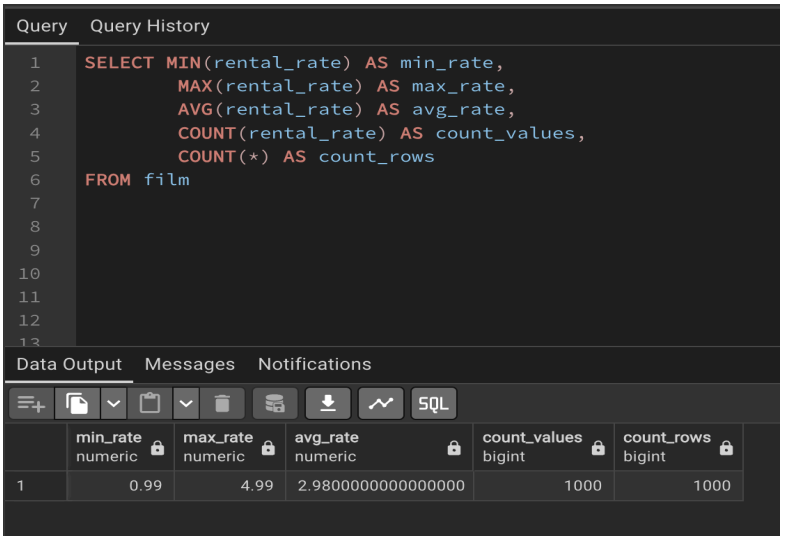


No non-uniform data found. To fix non-uniform data, i will use GROUP BY or DISTINCT commands.

3. Missing data

No missing data was found. If there is data missing then we can simply ignore the columns with the high percentage of missing values. Another option is to impute the values using statistical methods. We can input the estimate or average values.

2. **Summarize your data**



Film table

	MIN	MAX	AVG	COUNT	COUNT _ROWS	MODE
film_id	1	1000	500.50	1000	1000	
title				1000	1000	Academy Dinosaur
description				1000	1000	
release_year	2006	2006	2006	1000	1000	

language_id	1	1	1	1000	1000	
Rental _duration	3	7	4.99	1000	1000	
length	46	185	115.27	1000	1000	
replacement_ cost	9.99	29.99	19.98	1000	1000	
rating				1000	1000	PG-13
last_update				1000	1000	2013-05-26
special_featu re				1000	1000	Trailers, Commenta ries
fulltext				1000	1000	Balloon 19 confront

Customer table

	MIN	MAX	AVG	COUNT	COUNT_ Rows	MODE
customer_id	1	599	300	599	599	
store_id	1	2	1.455	599	599	
first_name				599	599	Jamie
email				599	599	Aaron.selby @sakilacusto mer.org
address_id	5	605	304.72	599	599	
activebool				599	599	true
create_date				599	599	2006-02-14
last_update				599	599	2013-05-26
active	0	1	0.97	599	599	

3. **Reflect on work**

Between Excel and SQL, SQL is more effective for data profiling. SQL can handle very large datasets quickly and lets you write queries to check for missing values, duplicates, or unusual patterns without slowing down. It is also more reliable because the results come directly from the database. Excel, on the other hand, is easier to use for small datasets and quick visual checks, but it becomes slow and hard to manage when the data is large or complex. Overall, SQL is faster, more powerful, and better for professional data profiling, while Excel is good for small or simple checks.