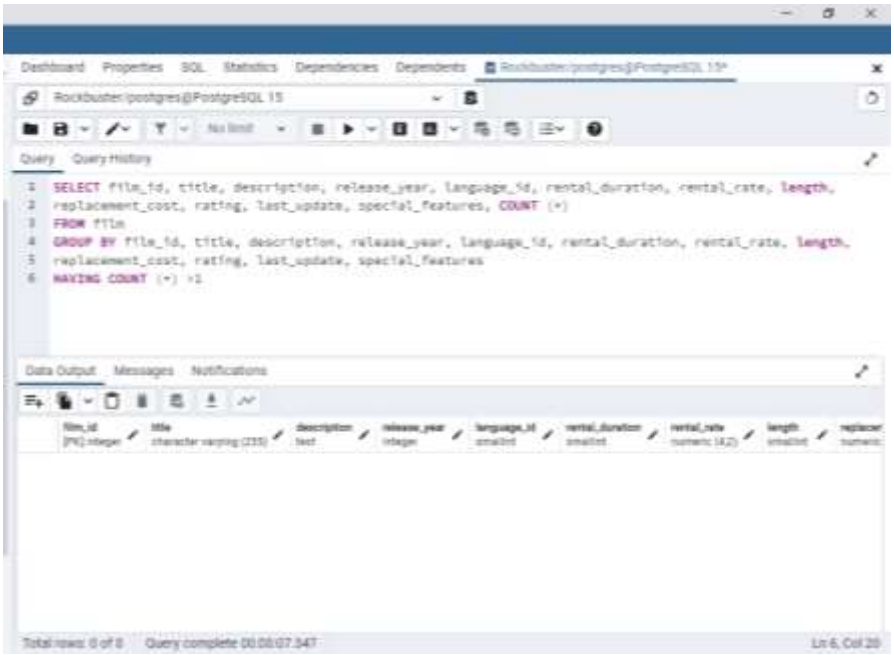
3.6 Task- Summarizing and Cleaning Data in SQL

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. Next to each query write 2 to 3 sentences explaining how you would clean the data (even if the data is not dirty).

-Duplicate Data for film table:



-Duplicate Data for customer table:

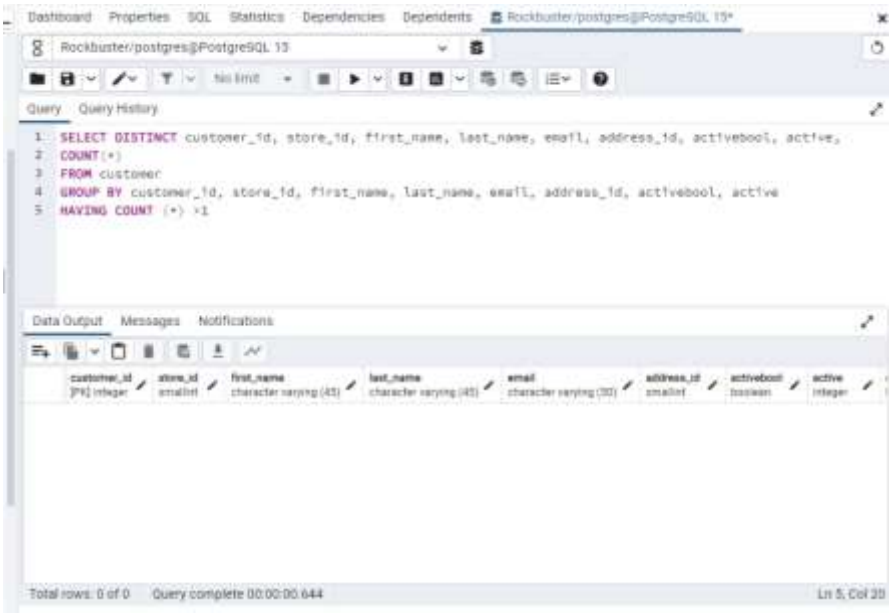
SELECT DISTINCT customer_id, store_id, first_name, last_name, email, address_id, activebool, active,

COUNT(*)

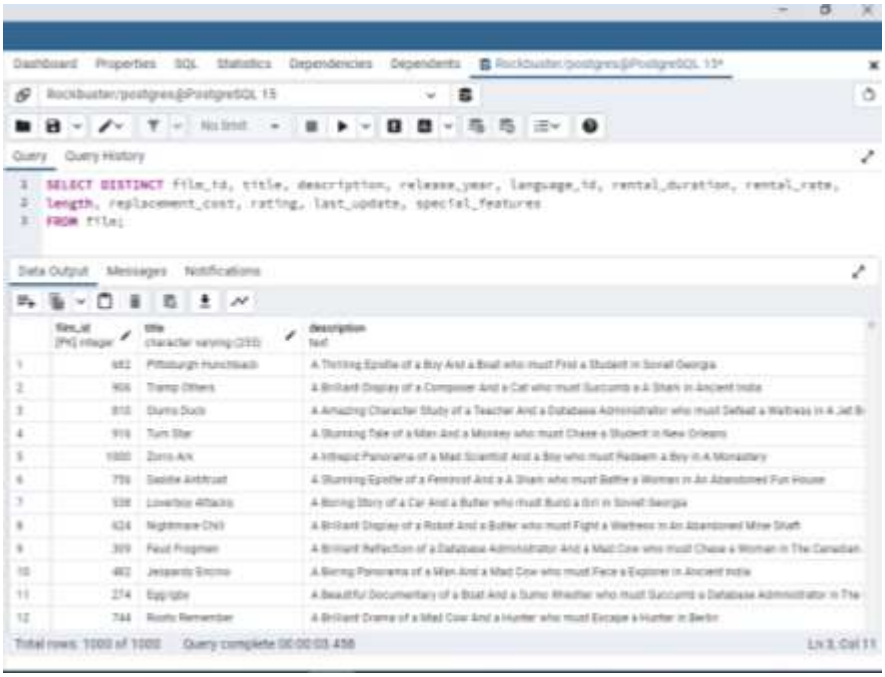
FROM customer

GROUP BY customer_id, store_id, first_name, last_name, email, address_id, activebool, active

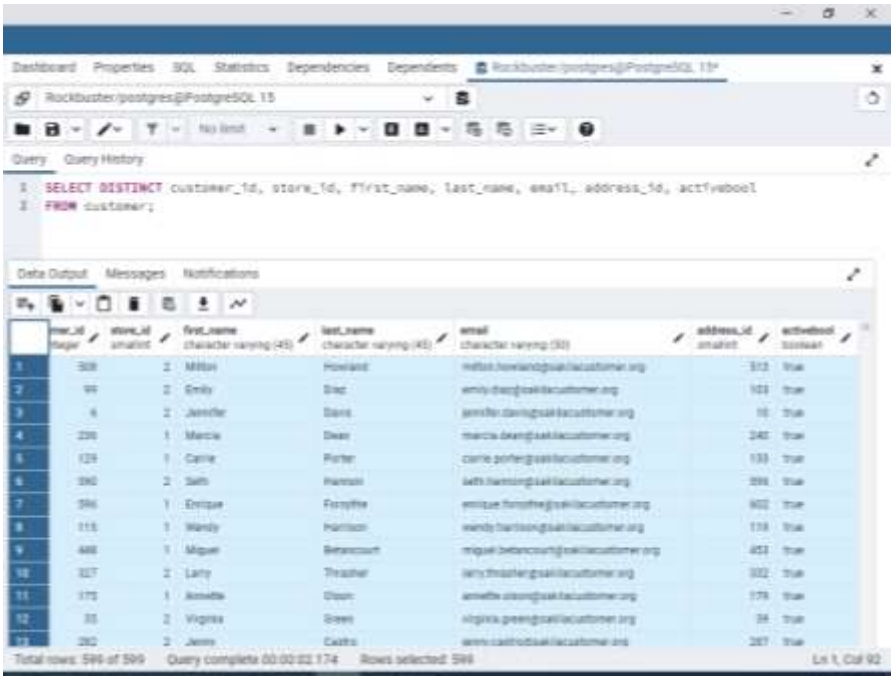
HAVING COUNT (*) >1



-Non-Uniform Data for film table:



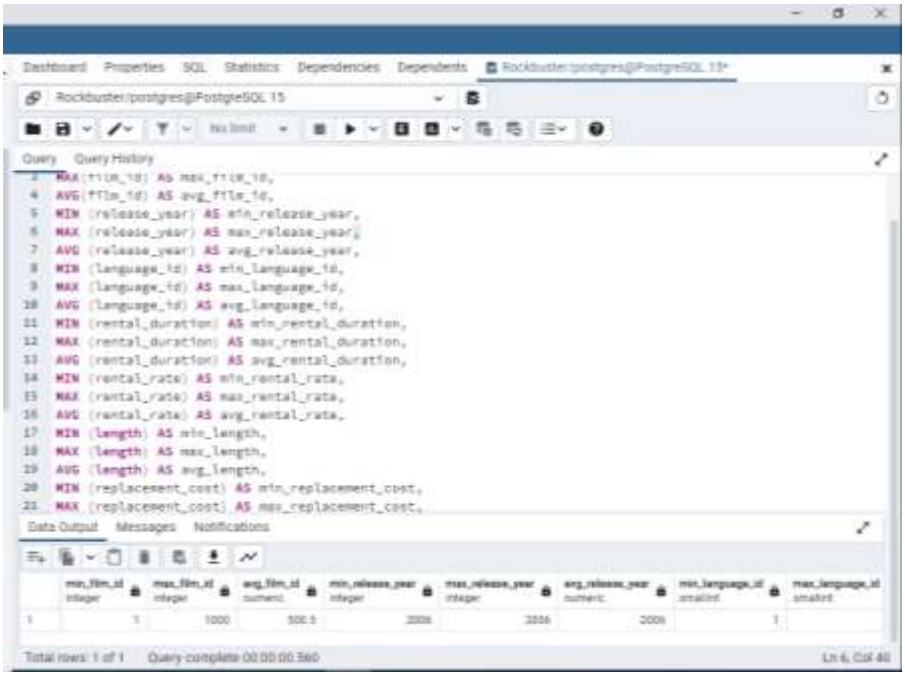
-Non-Uniform Data for customer table:



There were no duplicate values in the film or the customer table, but if there were and you were given permission to alter the database, you could create a view and select only the unique records, or delete the duplicate record from the table or view.

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

-Descriptive Stats for film table;



MIN (release_year) AS min_release_year,

MAX (release_year) AS max_release_year,

AVG (release_year) AS avg_release_year,

MIN (language_id) AS min_language_id,

MAX (language_id) AS max_language_id,

AVG (language_id) AS avg_language_id,

MIN (rental_duration) AS min_rental_duration,

MAX (rental_duration) AS max_rental_duration,

AVG (rental_duration) AS avg_rental_duration,

MIN (rental_rate) AS min_rental_rate,

MAX (rental_rate) AS max_rental_rate,

AVG (rental_rate) AS avg_rental_rate,

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,

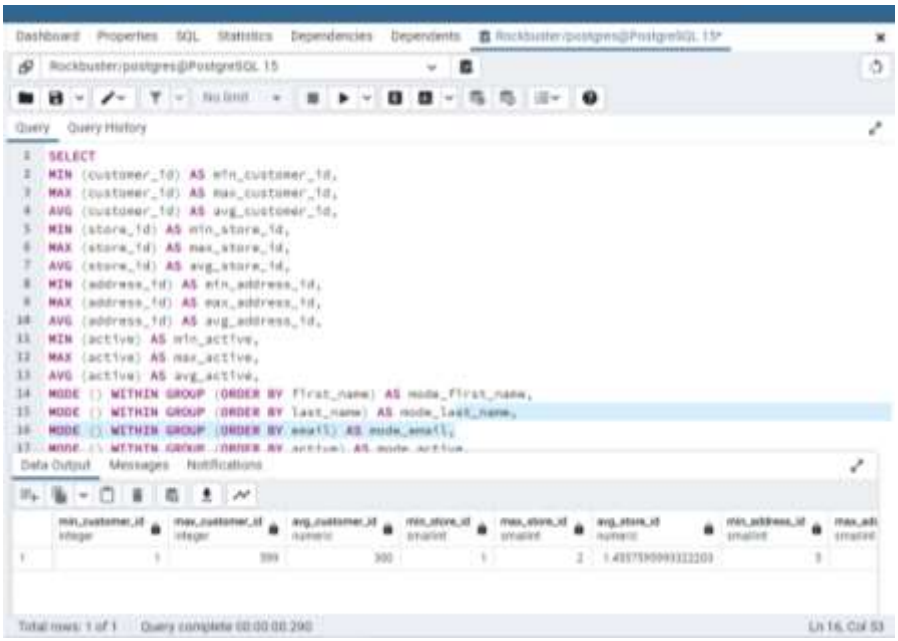
AVG (replacement_cost) AS avg_replacement_cost,

MODE () WITHIN GROUP (ORDER BY title) AS mode_title, MODE () WITHIN GROUP (ORDER BY description) AS mode_description, MODE () WITHIN GROUP (ORDER BY rating) AS mode_rating,

MODE () WITHIN GROUP (ORDER BY special_features) AS mode_special_features,

MODE () WITHIN GROUP (ORDER BY fulltext) AS mode_fulltext FROM film

- Descriptive or customer table :



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,
MIN (active) AS min_active,
MAX (active) AS max_active,
AVG (active) AS avg_active,
MODE () WITHIN GROUP (ORDER BY first_name) AS mode_first_name,
MODE () WITHIN GROUP (ORDER BY last_name) AS mode_last_name,
MODE () WITHIN GROUP (ORDER BY email) AS mode_email,
MODE () WITHIN GROUP (ORDER BY active) AS mode_active,
MODE () WITHIN GROUP (ORDER BY activebool) AS mode_activebool
FROM customer;

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.

When I have small data I feel more confident using exchel, But in case with large data I think it’s pretty easy using SQL.