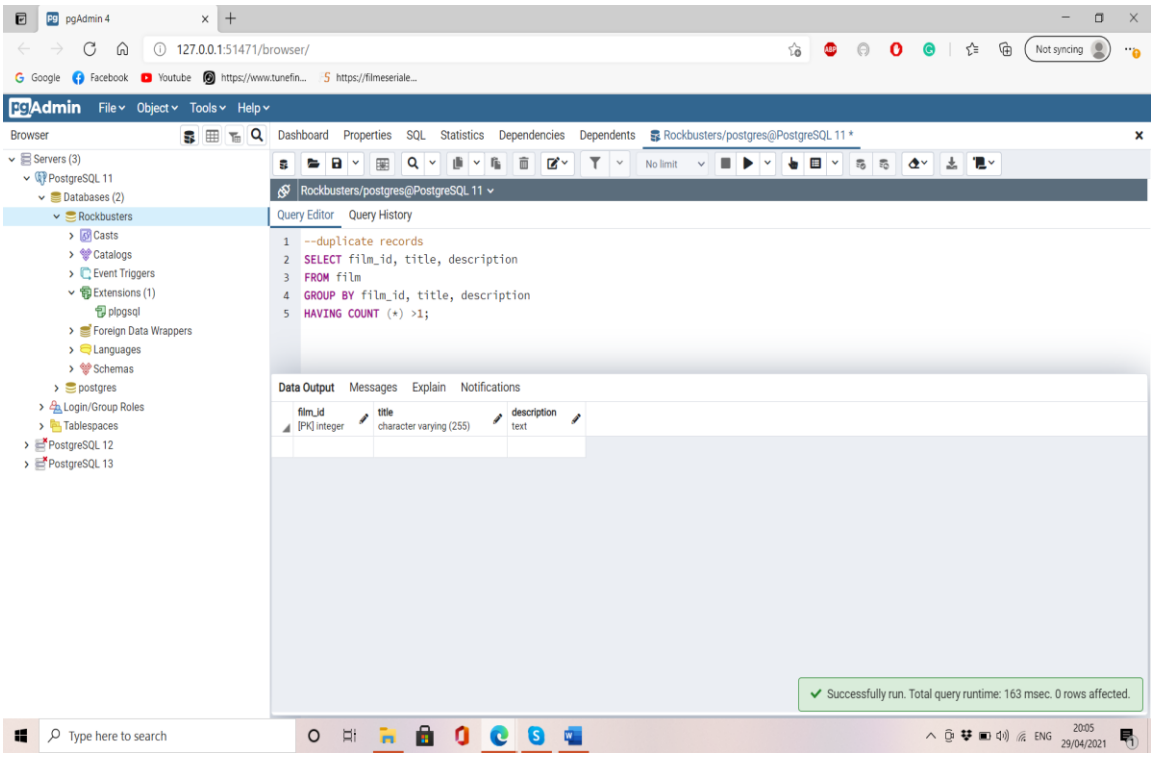
Directions

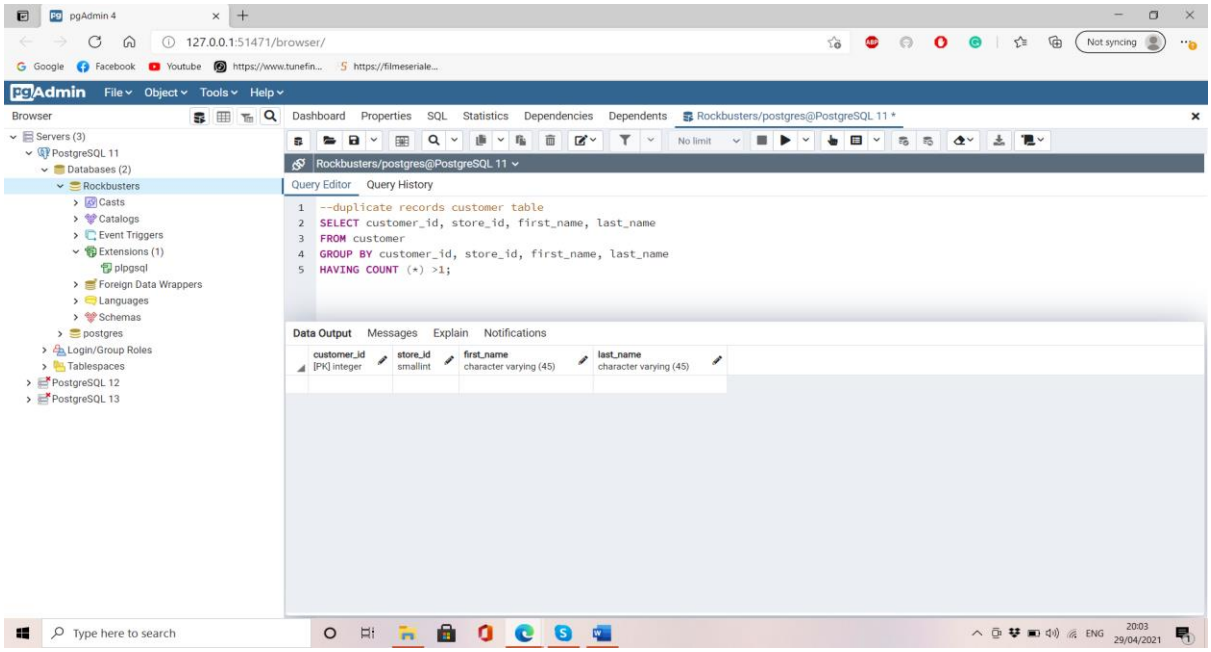
- Rockbuster's database engineers have loaded some new data into the database, and your manager has asked you to clean and profile it. Follow the instructions below to complete their request:

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.
- Create a new “Answers 3.6” document and copy-paste your queries into it. Next to each query write 2 to 3 sentences explaining how you would clean the data (even if the data is not dirty).

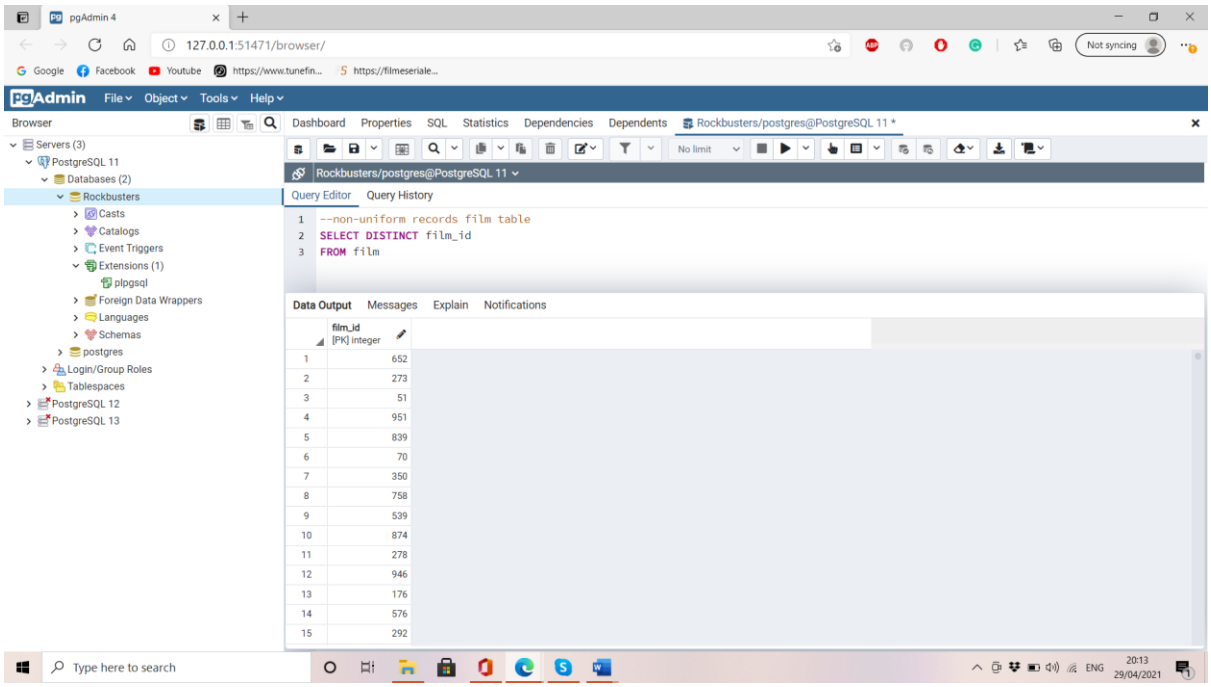
Duplicate Data



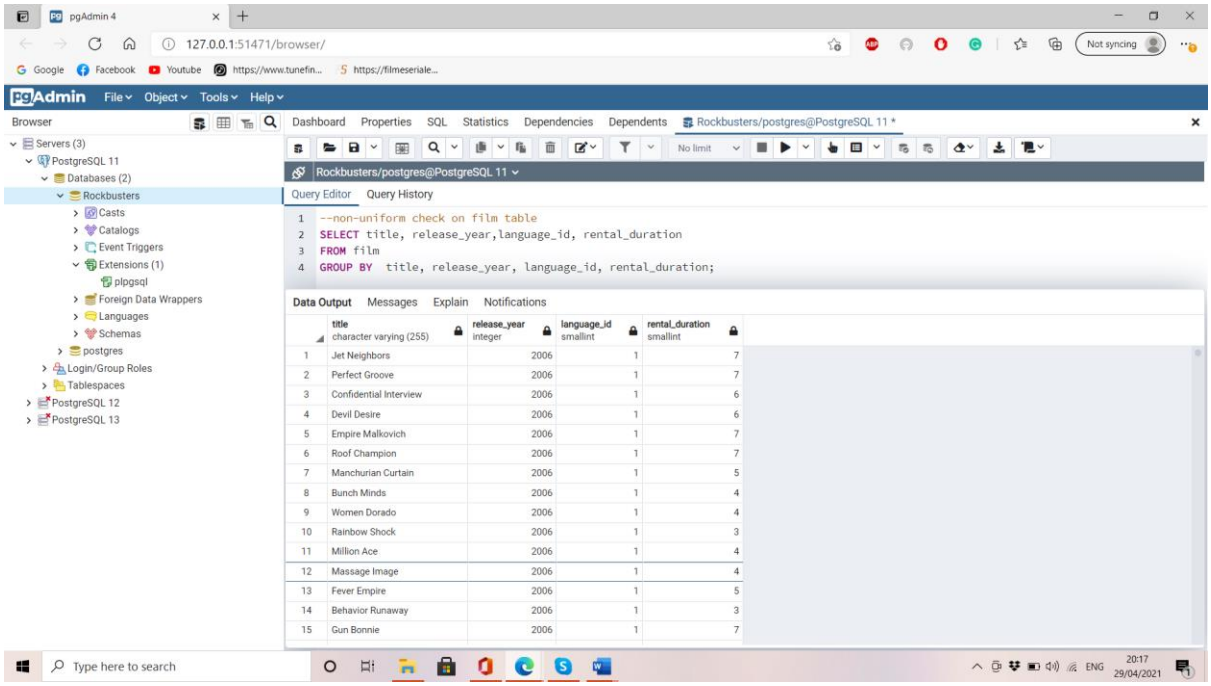


- This is the query that I had run, and it did not give any duplicate information.
- This means that the data is clean.
- I have selected the first few columns from each table to check for duplicates. There are none.

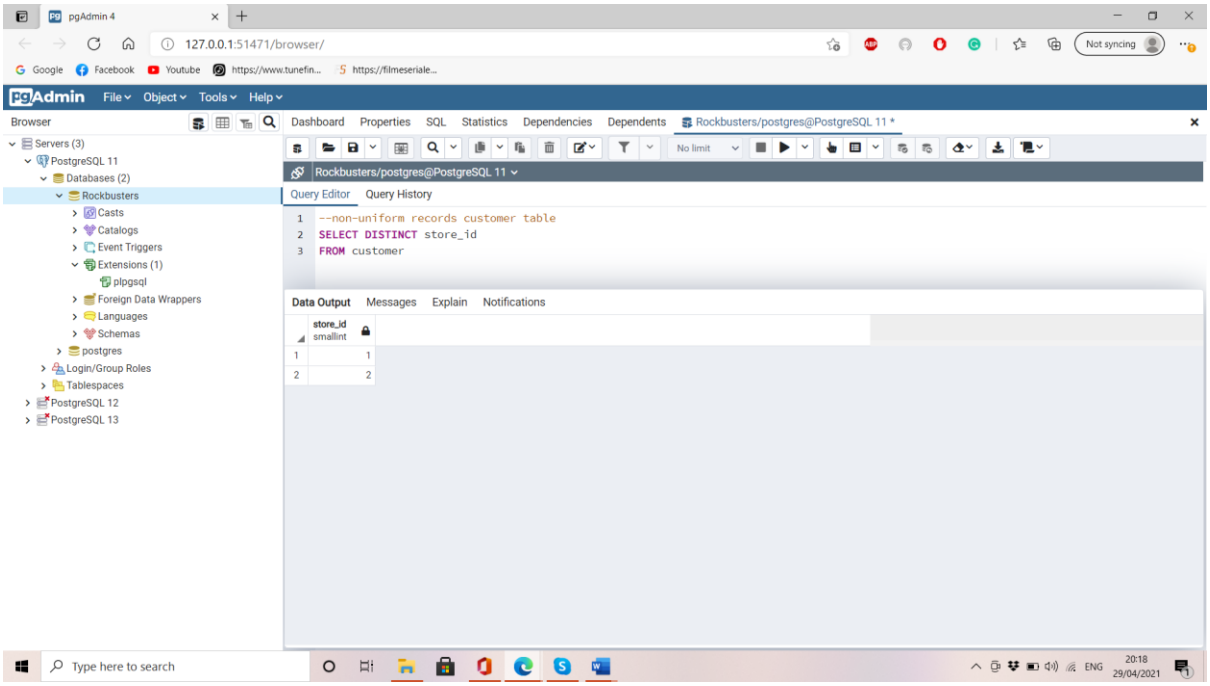
Non - uniform Values



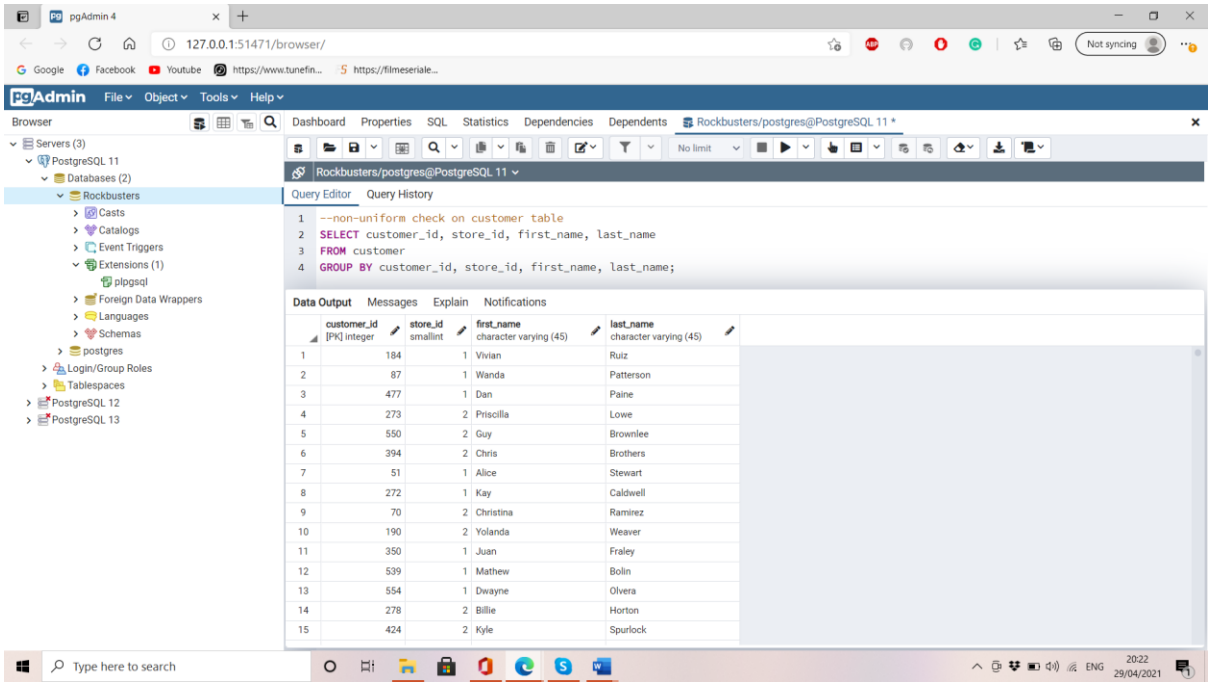
- Scanning through the distinct values for film_id nothing looks non-uniform.



- Visual check shows no obvious inconsistencies.

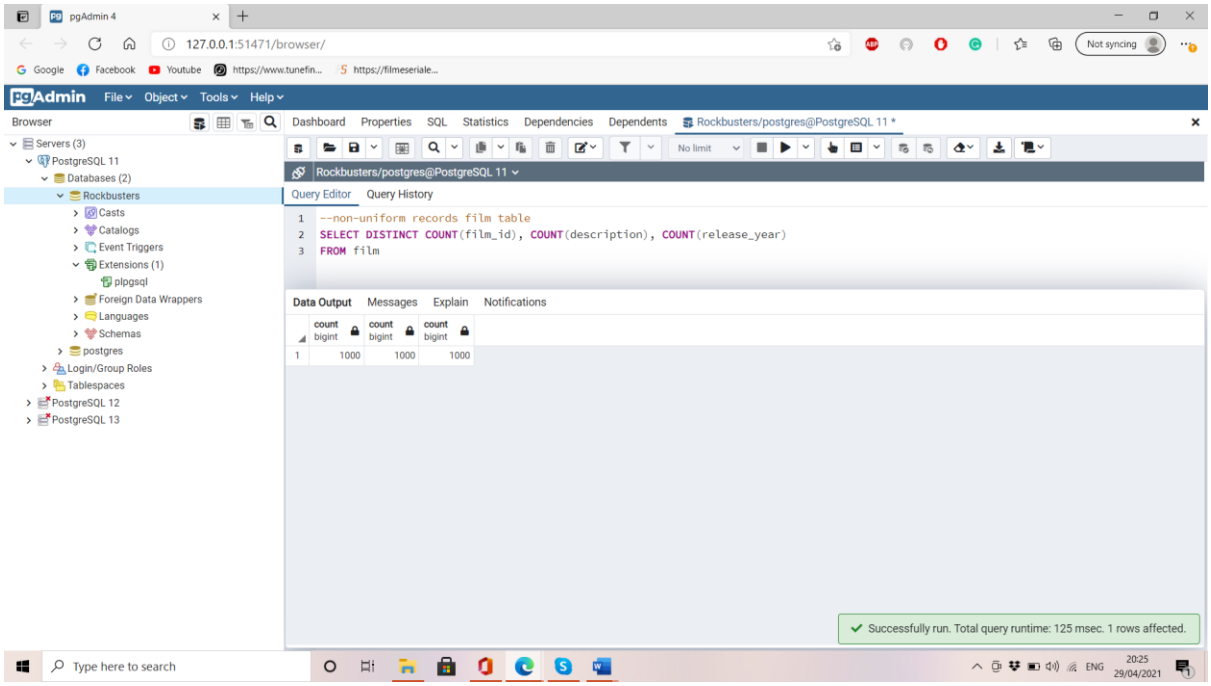


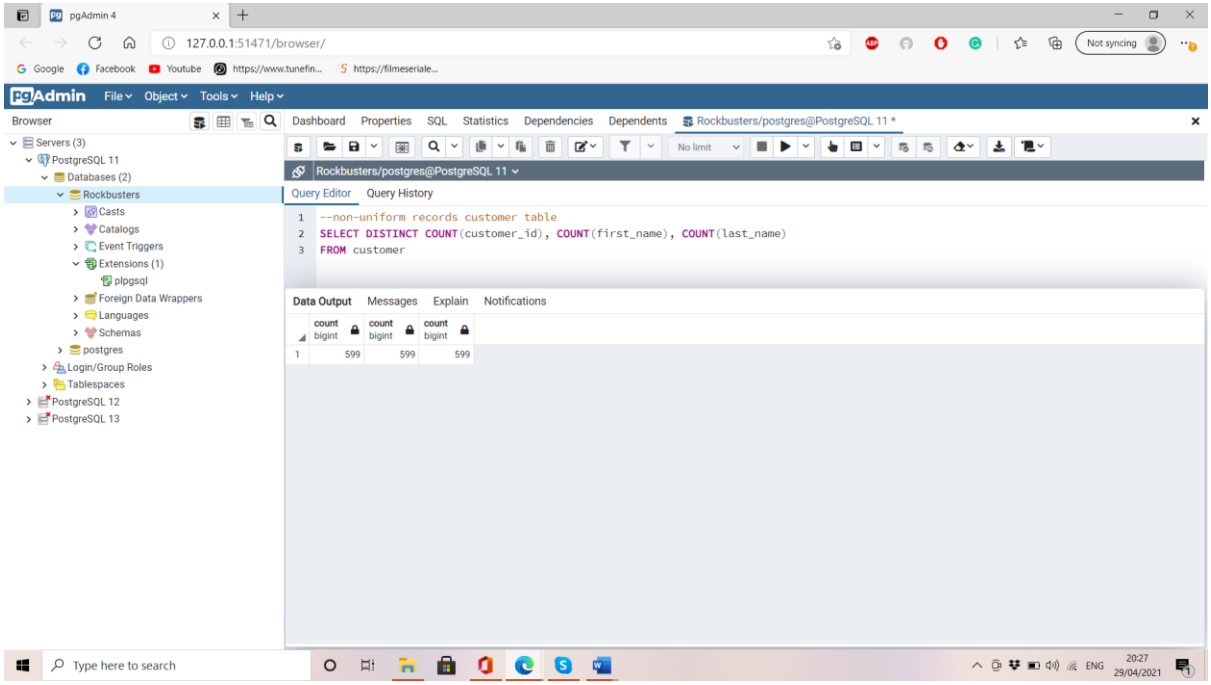
- Only two values for store_id – not non-uniform.



- Visual check shows no obvious inconsistencies.
- If I needed to correct non-uniform values, I would use the UPDATE command. For example, if store_id had been inputted incorrectly for some records (instead of 1 – ‘one’ or ‘uno’ had been entered) I would use the following command: UPDATE customer SET store_id = 1 WHERE store_id IN (‘one’, ‘uno’).

Missing Values



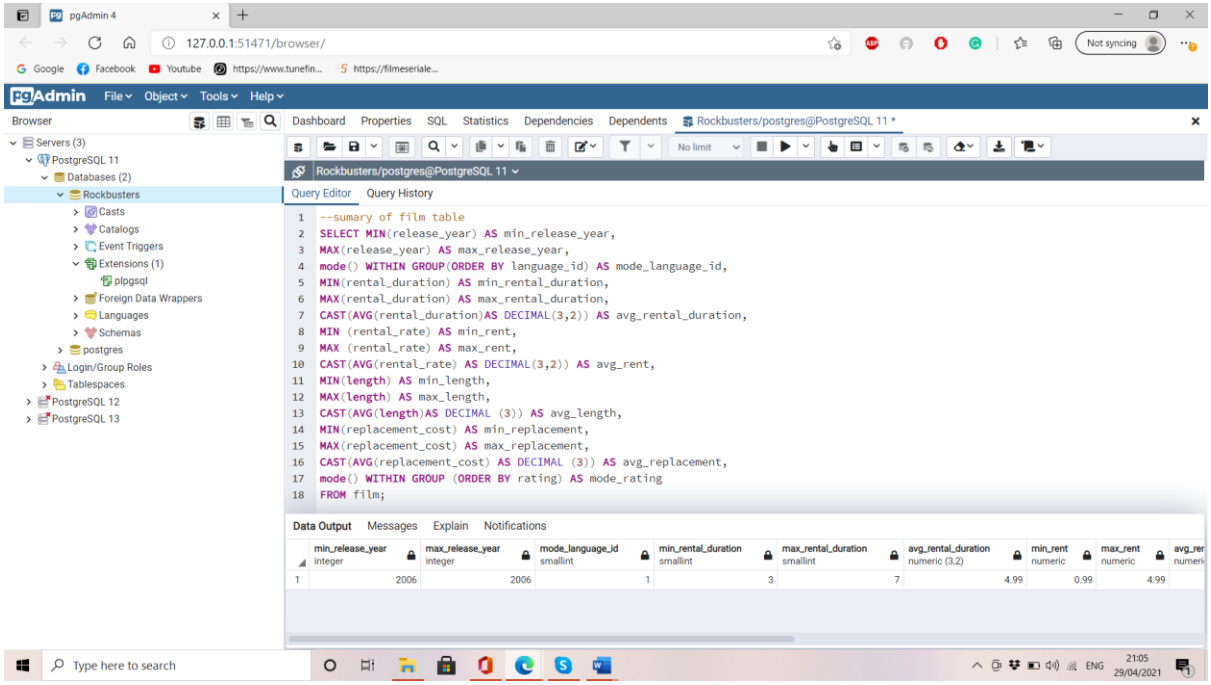


- To find non-uniform data it´s important to check some random values and look for potential anomalies or inconsistencies.
- To fill in missing data following options should be considered: - Ignore columns with a high percentage of missing values - Impute the missing values with estimates.

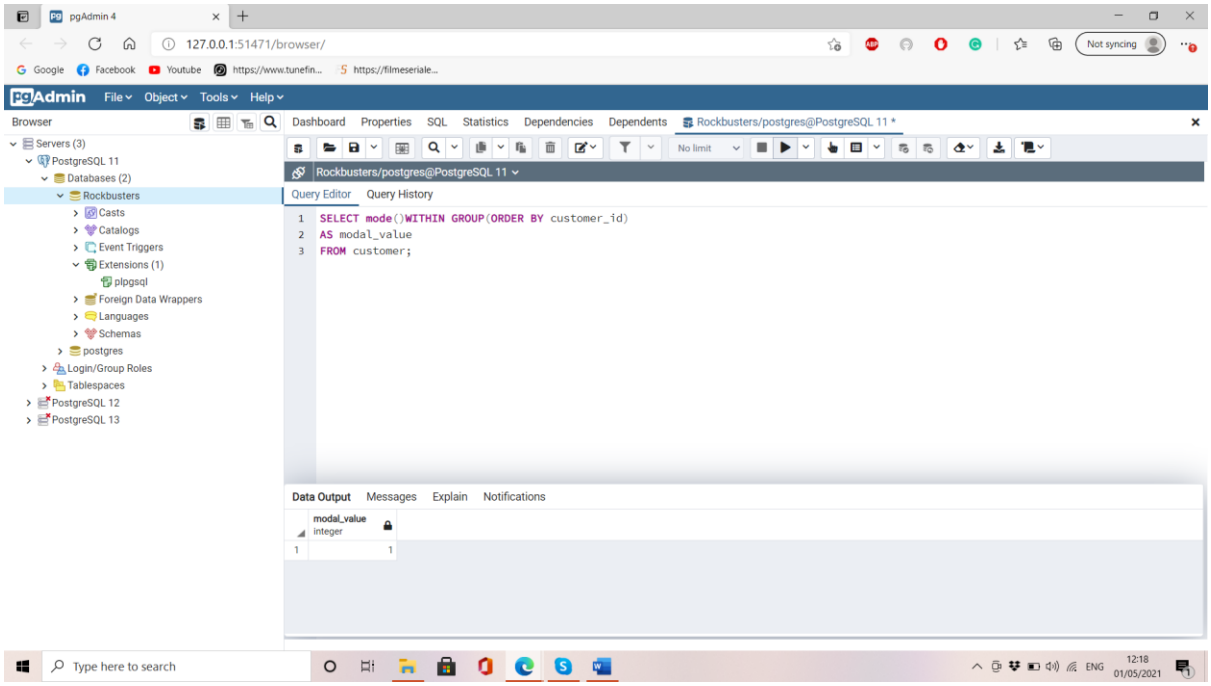
Summarize your data:

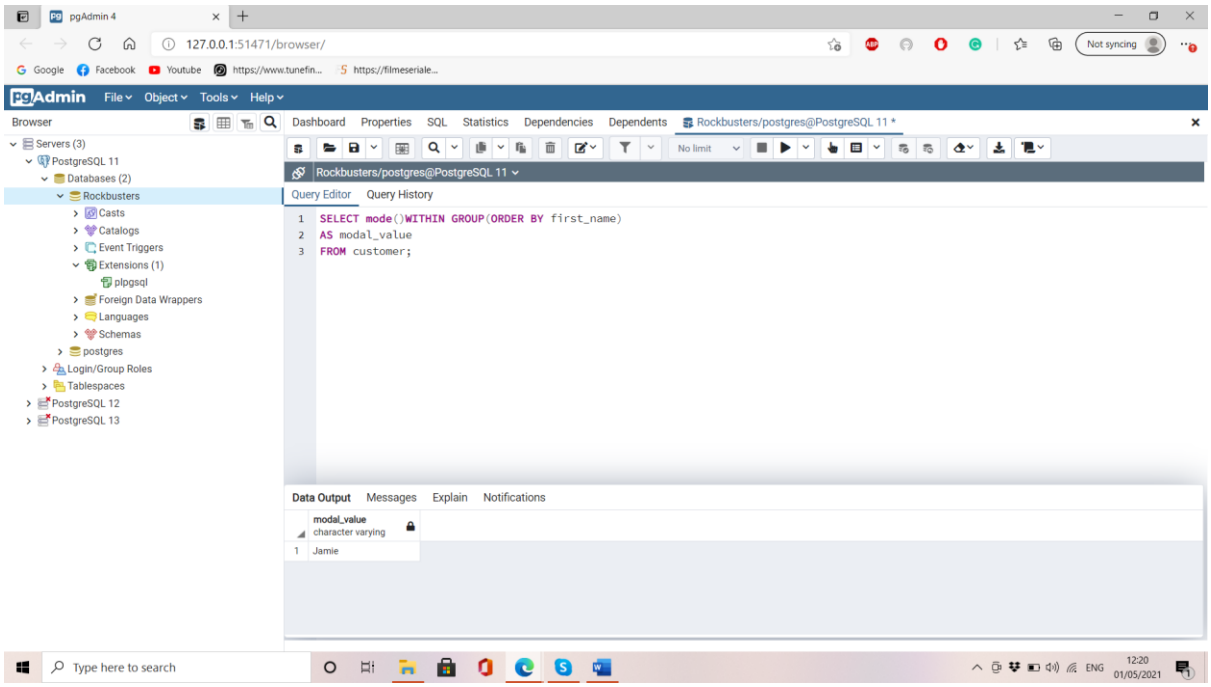
- Use SQL to calculate descriptive statistics for both the film table and the customer table.
- For numerical columns, this means finding the minimum, maximum, and average values. For non-numerical columns, calculate the mode value.
- Copy-paste your SQL queries and their outputs into your answers document.

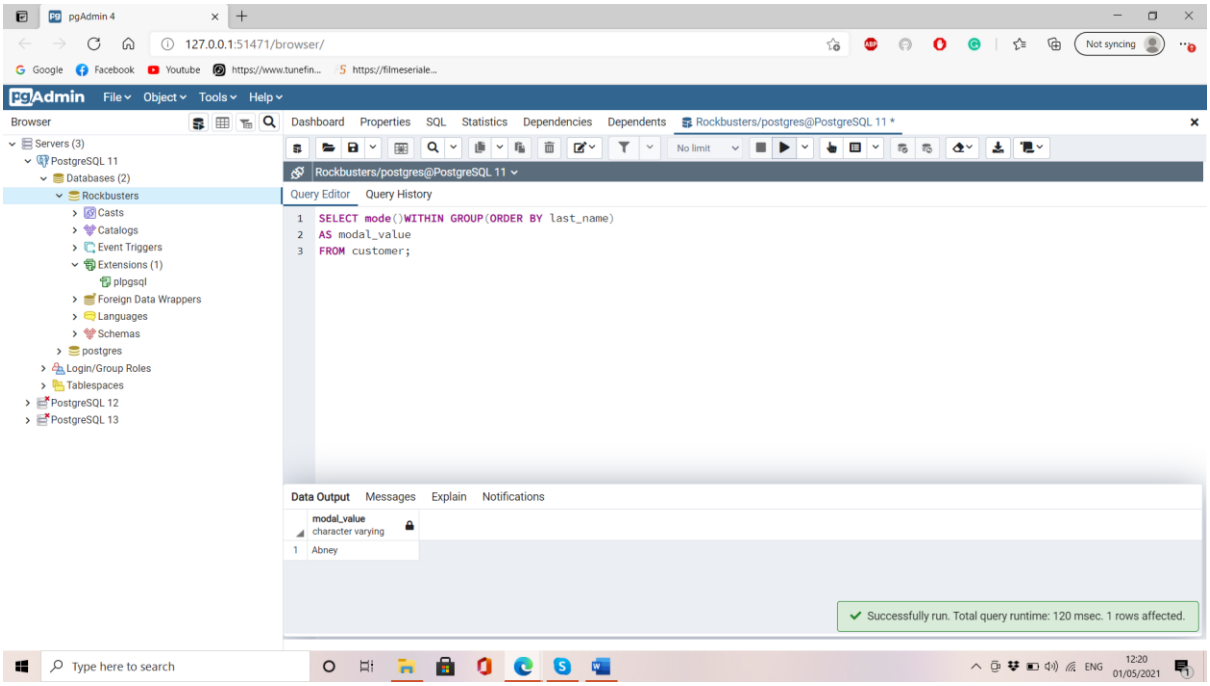
Film table



Customer table







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.

- I think that the initial filtering and summarizing of data is easier in excel using pivot tables and filters however this is only the case if you are working with small data sets.
- SQL is amazingly fast once you know the correct queries to use and is remains very fast when working with large datasets which is not the case with excel.
- If I were only working with relatively small datasets, and no other user needed access to the data – I would use excel. Otherwise, SQL is the right choice.