Task 3.4 Database Querying in SQL

By Lee Heng Chuah

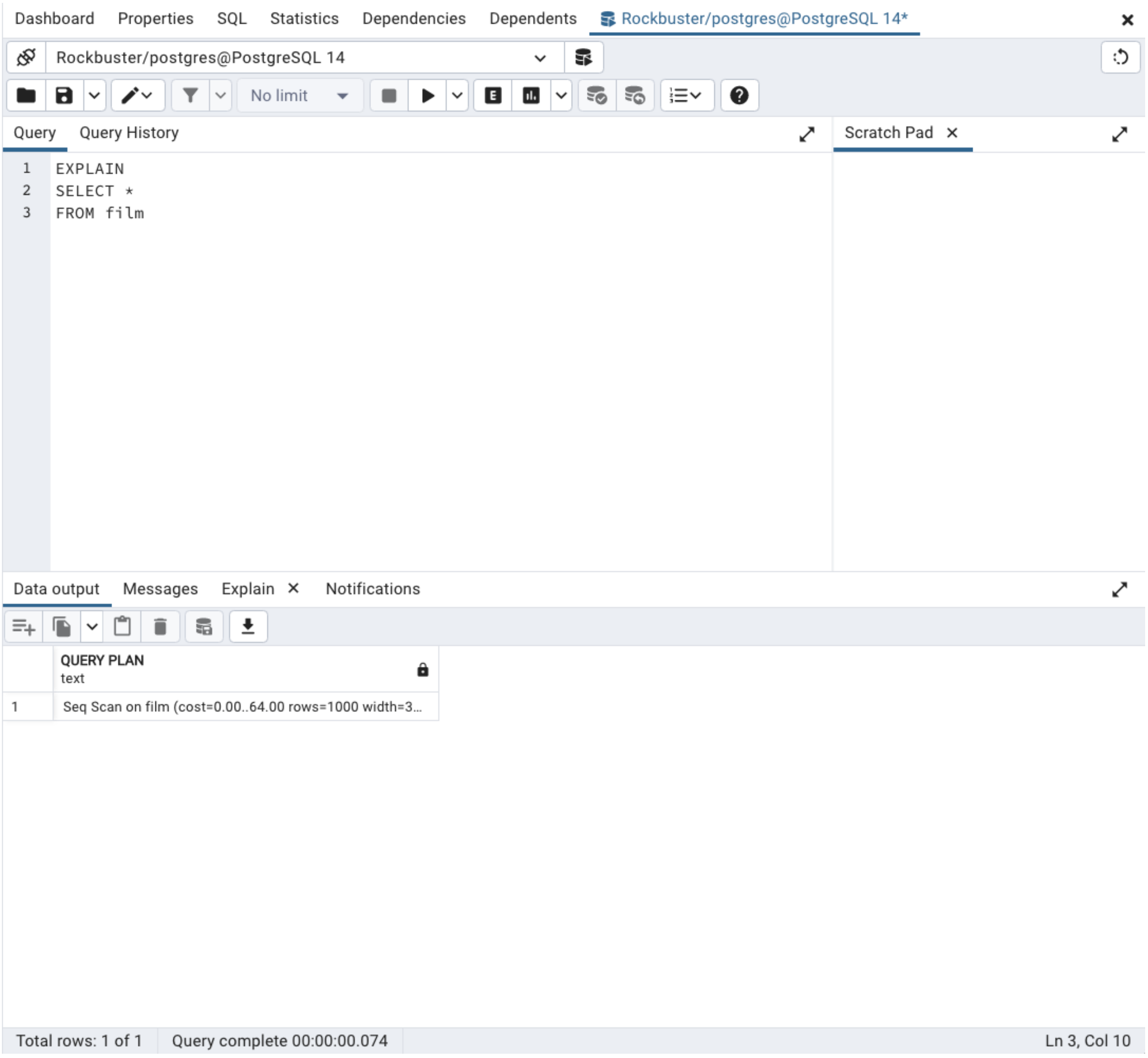
It is time to put what you have learned into practice. In this task you will be optimising queries, sorting and grouping data, and reflecting on the database migration process outlined in the exercise.

TIP: You will need to use aliases to give appropriate names to the aggregate columns you calculate in step 3 and the bonus task. For a recap of how to assign aliases, see Rules and Best Practices in Exercise 3.3: SQL for Data Analysts

Directions:

As you have done for previous tasks, create a new text document for your answers and call it "Answers 3.4".

1. **Refining Your Query:** You need to get some data from the "film" table and decide to use the query SELECT * FROM film.

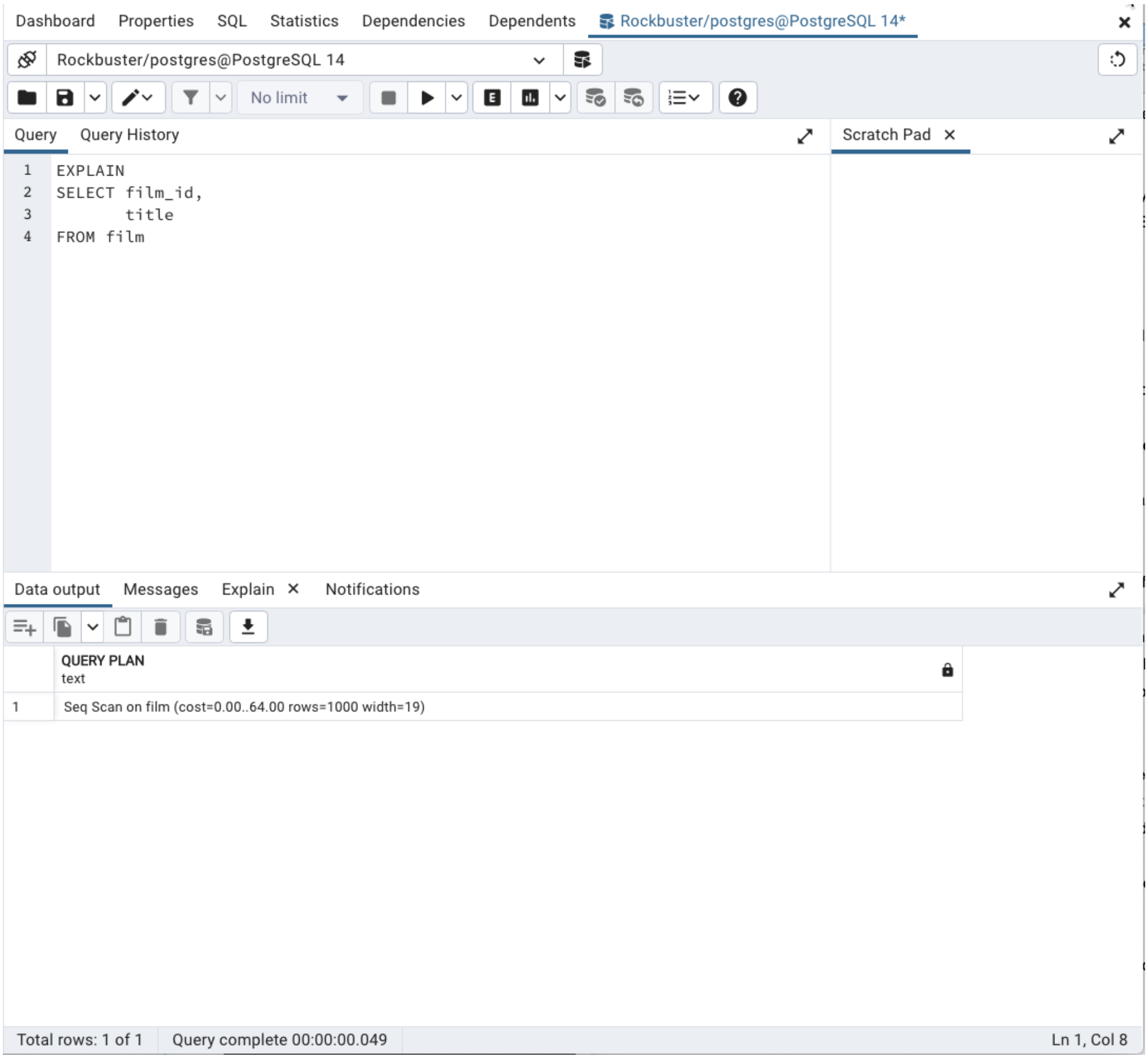


```
EXPLAIN
SELECT *
FROM film
```

	QUERY PLAN text
1	Seq Scan on film (cost=0.00..64.00 rows=1000 width=3...

Total rows: 1 of 1   Query complete 00:00:00.074   Ln 3, Col 10

- You realise that only the “film_id” and the “title” columns are needed. Write a new query that selects only those 2 columns.



```
EXPLAIN
SELECT film_id,
       title
FROM film
```

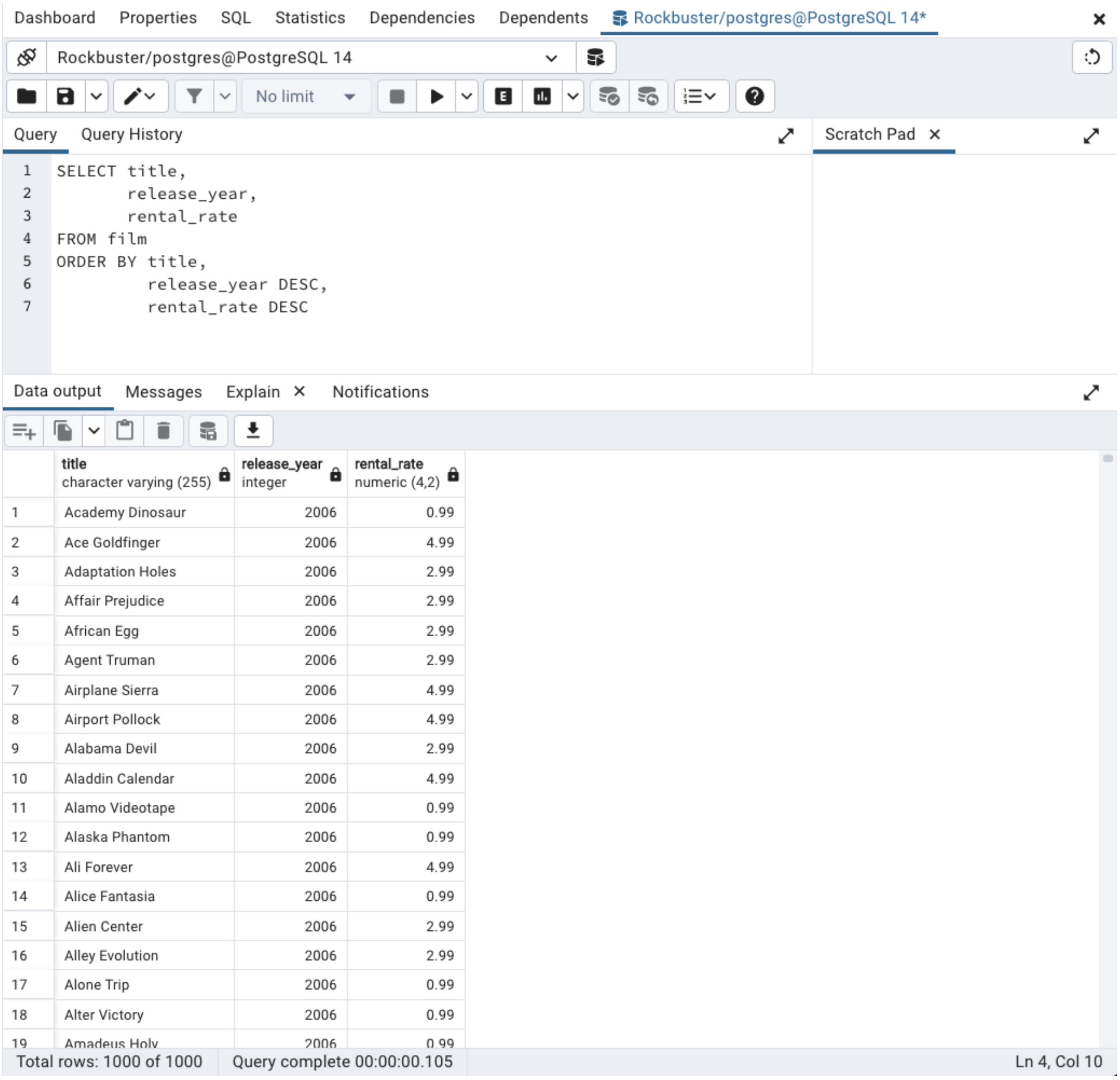
	QUERY PLAN text
1	Seq Scan on film (cost=0.00..64.00 rows=1000 width=19)

Total rows: 1 of 1 Query complete 00:00:00.049 Ln 1, Col 8

- Compare the cost of the original query and the revised query, and write a few sentences explaining the comparison. Can you suggest any ways to optimise this query?
  Both have same **cost** as we can see from the Data Output (cost=0.00…64.00). It means the “cost” or time, of retuning the first row is 0, but the cost of returning all the rows is 64.
  In this case, the query run time for the first query took 120 msec, while the query run time for the second query is 105 msec. The SQL becomes faster if we use the actual column names or more defined column in SELECT statement instead of than ‘*’.

2. **Ordering the Data:**

- In the PgAdmin Query Tool, run a query that selects every film from the “film’ table, with the movies sorted by title from A to Z, then by most recent release year, and then by highest to lowest rental rate



```
SELECT title,
       release_year,
       rental_rate
FROM film
ORDER BY title,
         release_year DESC,
         rental_rate DESC
```

	title character varying (255)	release_year integer	rental_rate numeric (4,2)
1	Academy Dinosaur	2006	0.99
2	Ace Goldfinger	2006	4.99
3	Adaptation Holes	2006	2.99
4	Affair Prejudice	2006	2.99
5	African Egg	2006	2.99
6	Agent Truman	2006	2.99
7	Airplane Sierra	2006	4.99
8	Airport Pollock	2006	4.99
9	Alabama Devil	2006	2.99
10	Aladdin Calendar	2006	4.99
11	Alamo Videotape	2006	0.99
12	Alaska Phantom	2006	0.99
13	Ali Forever	2006	4.99
14	Alice Fantasia	2006	0.99
15	Alien Center	2006	2.99
16	Alley Evolution	2006	2.99
17	Alone Trip	2006	0.99
18	Alter Victory	2006	0.99
19	Amadeus Holy	2006	0.99

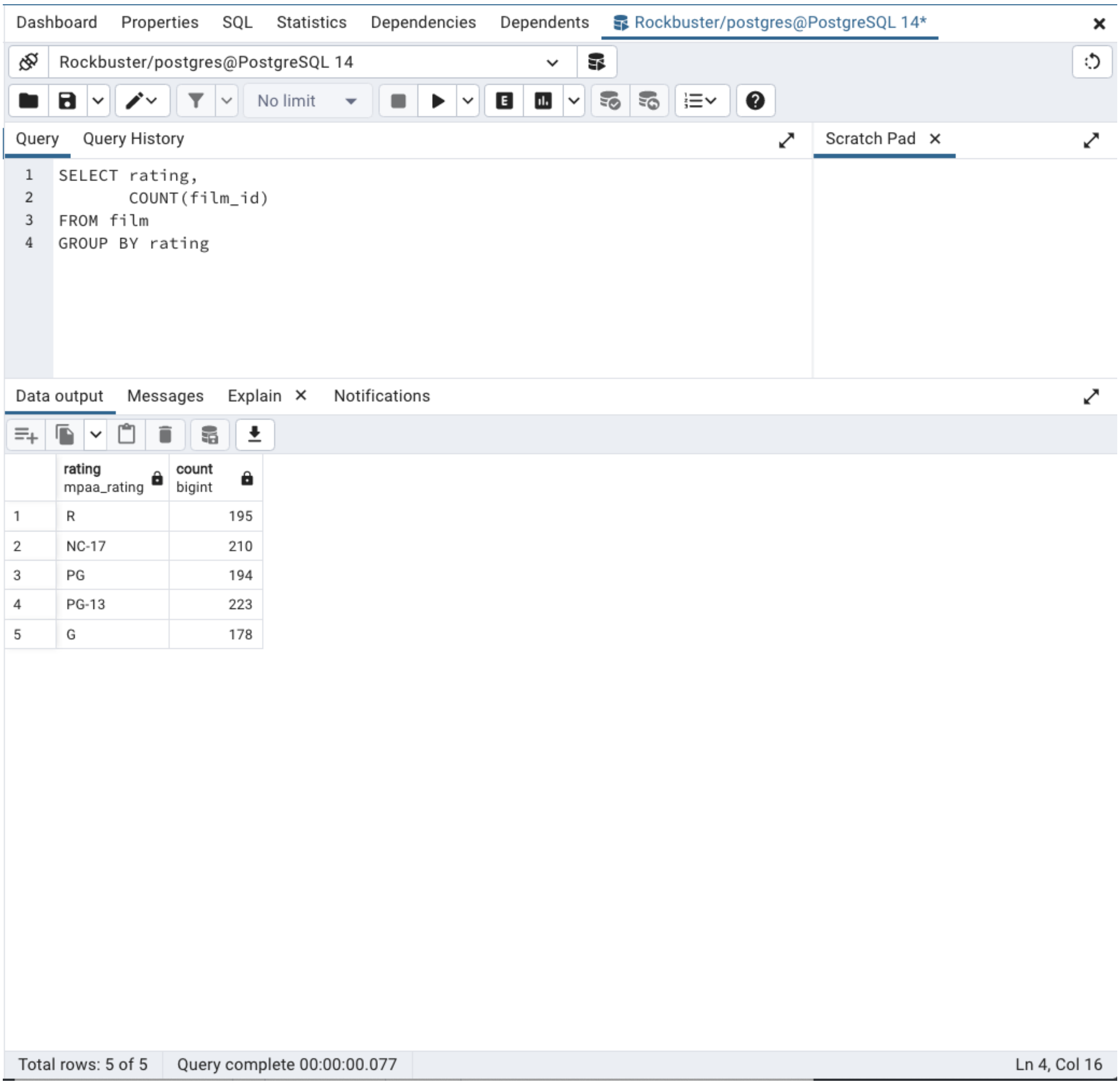
Total rows: 1000 of 1000 Query complete 00:00:00.105 Ln 4, Col 10

- Extract the data output of your query into a csv file for the film collection department to analyse in Excel. (You may need to explore how to save your output as a csv file in the Query Tool)

  See separate csv file. - Completed

3. **Grouping Data:** The strategy department has asked you the question below. Write a SQL query to retrieve the correct answers, then extract your results as a csv file.

- What is the average rental rate for each rating category?



```
SELECT rating,
       COUNT(film_id)
FROM film
GROUP BY rating
```

	rating mpaa_rating	count bigint
1	R	195
2	NC-17	210
3	PG	194
4	PG-13	223
5	G	178

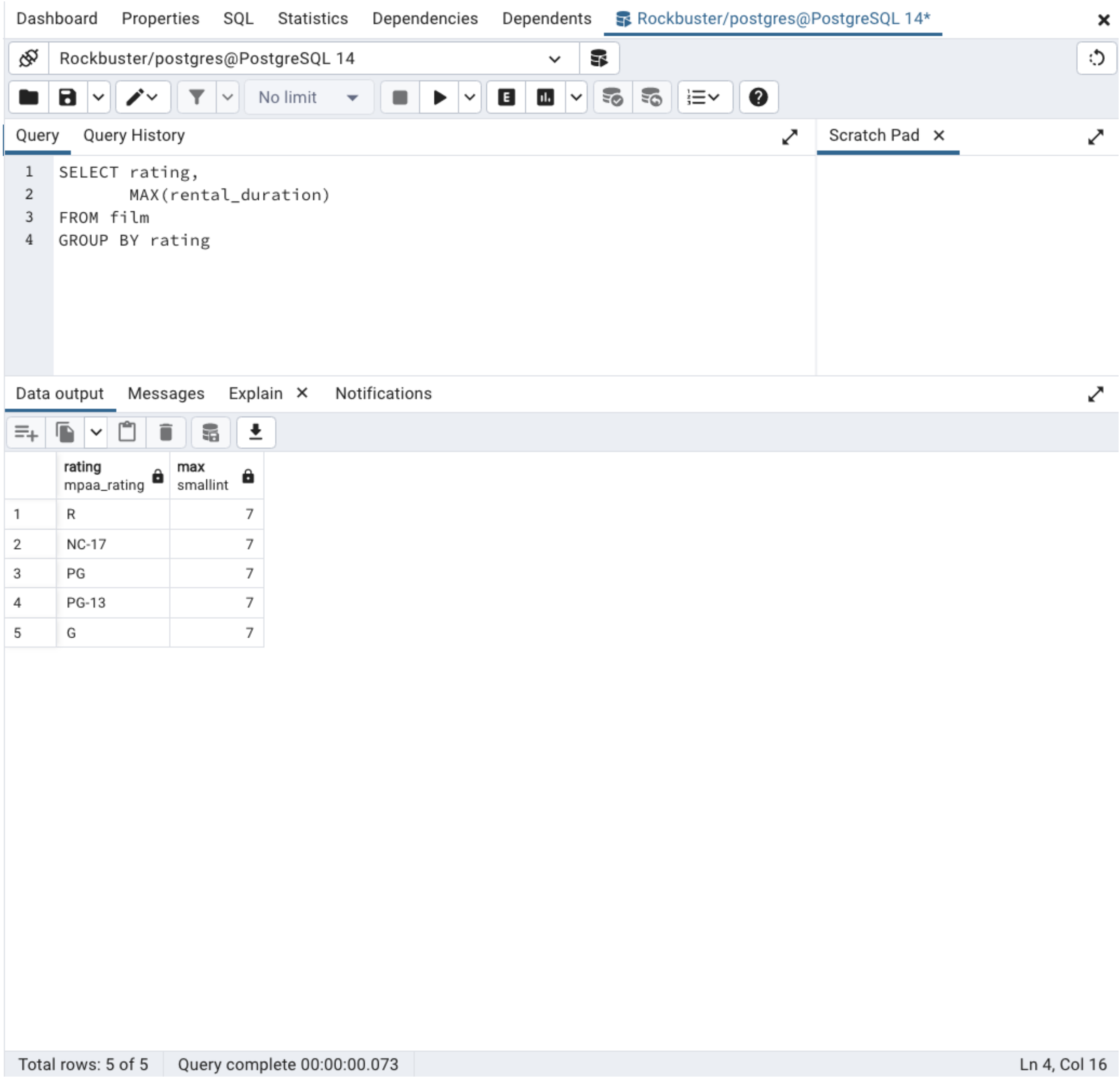
Total rows: 5 of 5 Query complete 00:00:00.077 Ln 4, Col 16

- What are the minimum and maximum rental durations for each rating category?

Maximum rental duration = 7

Minimum rental duration = 3

Maximum rental duration

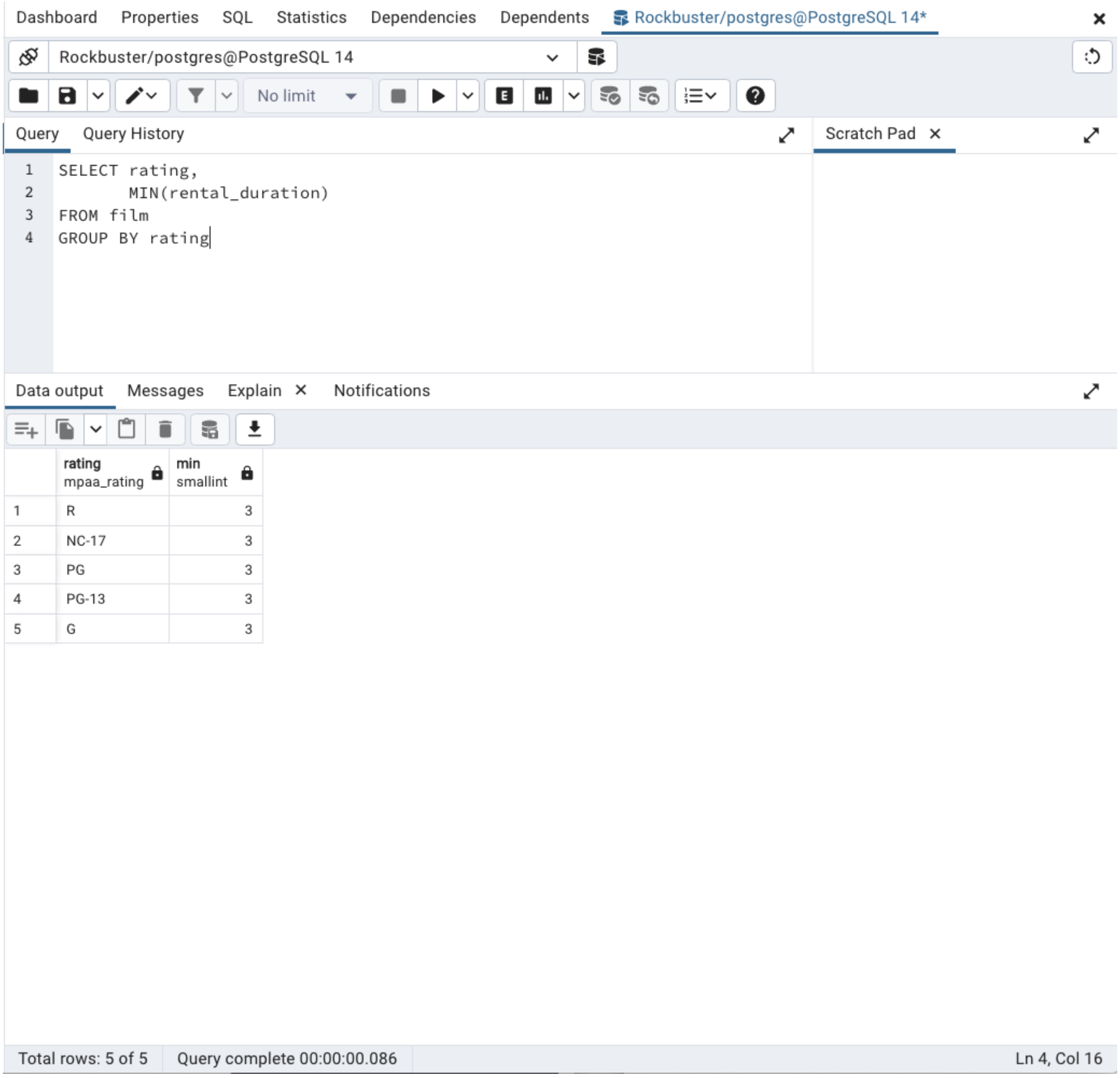


```
SELECT rating,
       MAX(rental_duration)
FROM film
GROUP BY rating
```

	rating mpaa_rating	max smallint
1	R	7
2	NC-17	7
3	PG	7
4	PG-13	7
5	G	7

Total rows: 5 of 5 Query complete 00:00:00.073 Ln 4, Col 16

Minimum rental duration



```
SELECT rating,
       MIN(rental_duration)
FROM film
GROUP BY rating
```

	rating mpaa_rating	min smallint
1	R	3
2	NC-17	3
3	PG	3
4	PG-13	3
5	G	3

Total rows: 5 of 5 Query complete 00:00:00.086 Ln 4, Col 16

4. **Database Migration:** Your team has decided to use an external tool to collect data on user behaviour in the new Rockbuster Android app. Data collected from this new source will need to be loaded into the data warehouse before you can analyse it.

   - Can you outline the procedure for migrating the data and who will be responsible for it?

     Data engineer would take the responsibility in migrating the data into data warehouse. Data engineer will start collecting the data from multiple data sources. Then, he or she will convert this data into another format which will be aligned with the new data warehouse format. The transformed data will then be inserted or loaded into the new database system. ETL process shall be followed in this case.

   - What problems do you foresee if you start analysing the data before it is been loaded into the data warehouse?

     The data from different systems typically doesn't play together very well. If we start analysing the data before it has been loaded into the data warehouse, the data could be full of inconsistencies and the key relationships across different data sources might be missed. The decision makers will lose faith in its reliability e.g., not bringing in the information the users need the most, failing to support mission critical reporting workflow and fail to anticipating future data needs.

5. Combine your “Answers 3.4” document and csv files into a single PDF and upload it here for your tutor to review.