Task 4 Sorting and grouping data in SQL

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

New query:

SELECT film_id, title FROM film

- 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 optimize this query

To learn about the cost of the two query I used EXPLAIN.

Here are the results:

First query: "Seq Scan on film (cost=0.00..64.00 rows=1000 width=384)"

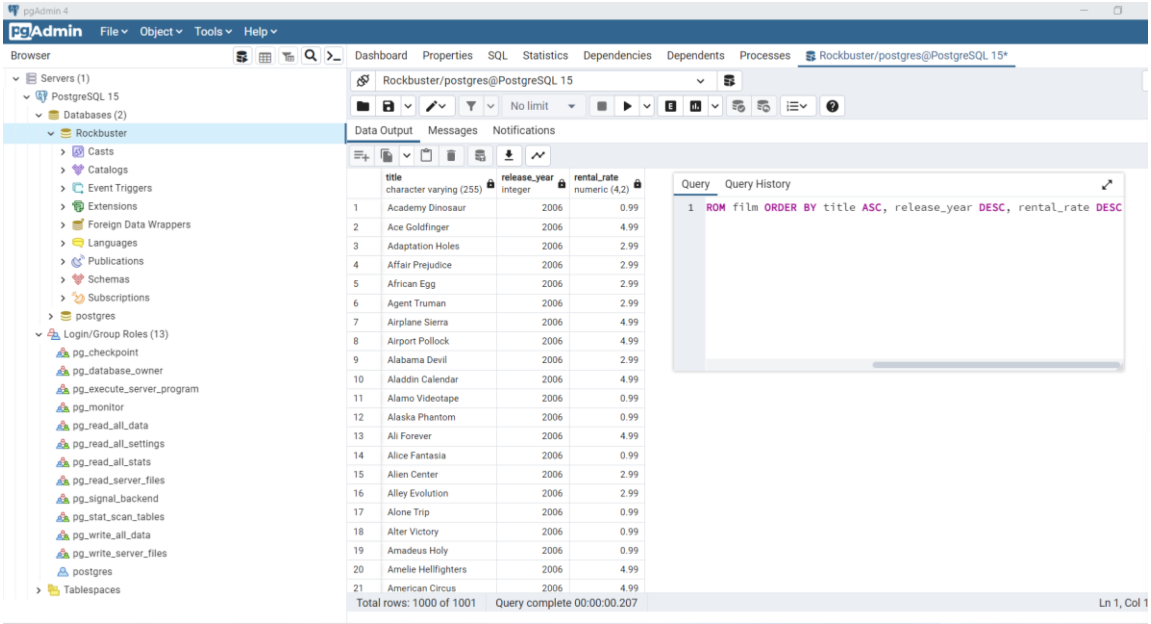
Second query: "Seq Scan on film (cost=0.00..64.00 rows=1000 width=19)"

This result means that the cost of the second query is the same as the first, although the second only retrieves 19 rows, so it is more efficient.

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.

Query:

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



```
ROM film ORDER BY title ASC, 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
20	Amelie Hellfighters	2006	4.99
21	American Circus	2006	4.99

Total rows: 1000 of 1001 Query complete 00:00:00.207 Ln 1, Col 1

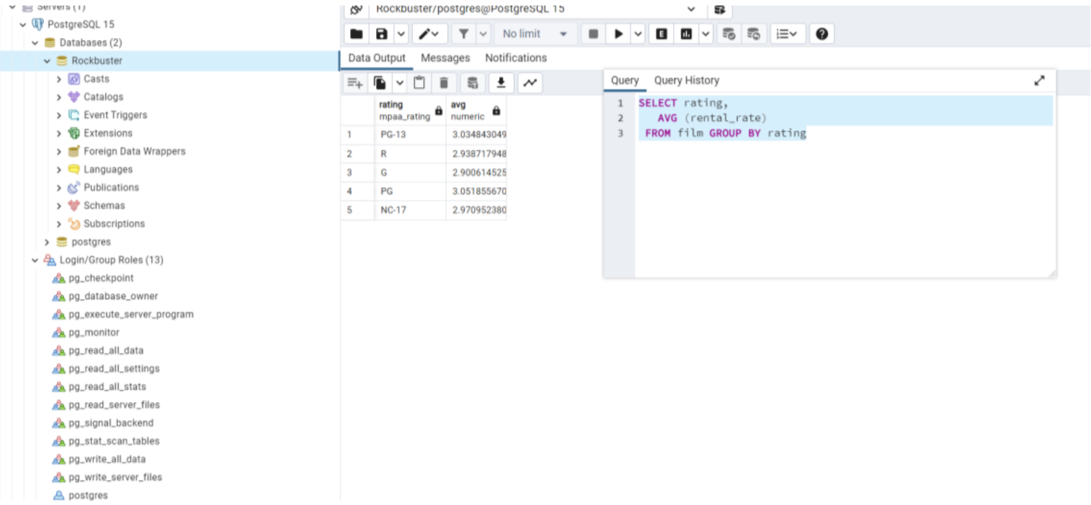
- Extract the data output of your query into a CSV file for the film collection department to analyze in Excel. To do this, click the button “Save results to file”:

File saved as an CSV

3. **Grouping Data:** The strategy department has asked you the questions 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, AVG (rental_rate) FROM film GROUP BY rating



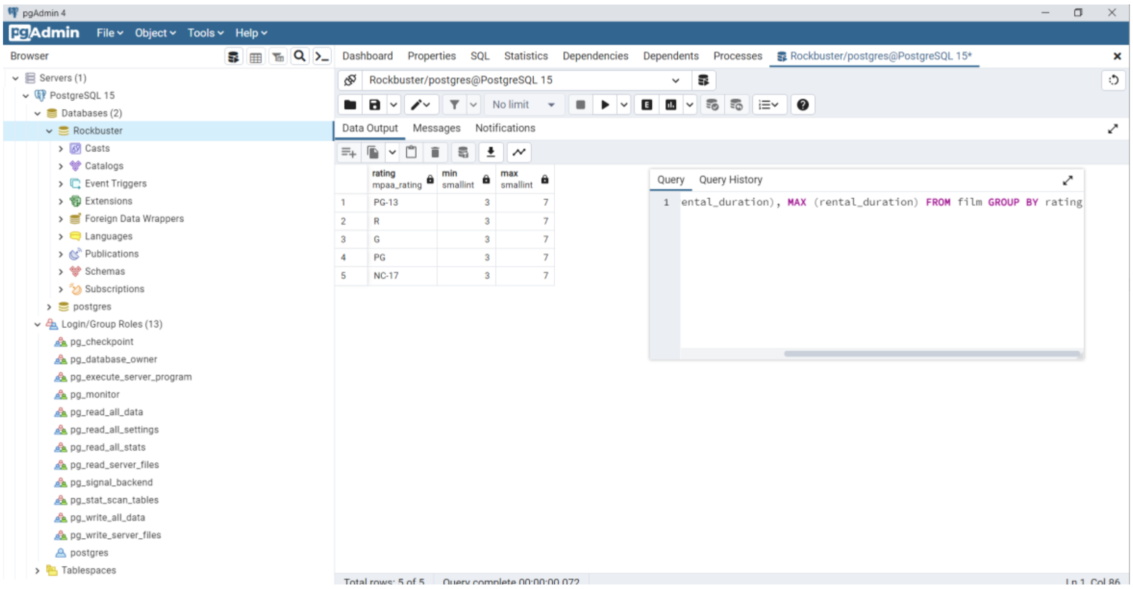
```
SELECT rating,
   AVG (rental_rate)
 FROM film GROUP BY rating
```

	rating mpaa_rating	avg numeric
1	PG-13	3.034843049
2	R	2.938717948
3	G	2.900614525
4	PG	3.051855670
5	NC-17	2.970952380

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

Command:

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



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

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

Total rows: 5 of 5 Query complete 00:00:00.072 Ln 1, Col 86

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

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

The procedure would ETL (extract, transform and load). For migrating the data the procedure is to first extract the data files from the Android app, second to convert the data into another format that can be then loaded into the data warehouse. For instance, data form the original source maybe in an image format, then it will need to be converted into a structured format (table of rows and columns) to able to be loaded into the data warehouse. This would be a job for a data engineer.

   ○ What problems do you foresee if you start analyzing the data before it's been loaded into the data warehouse?

If the data is not in a structured format, it will be very difficult to analyse it. But even if it is in a structured format, data may have problems in terms of cohesiveness, such as different formats for the same column entries, for instance, which will result in an additional cost and time for the analyst. Without having it loaded it will not be possible to make queries, such as simple statistical queries as all as grouping and ordering data. So having the data uploaded to the data warehouse is fundamental to be able to interview the dataset and gain insights from the information stored.

Bonus Task

You've not yet covered custom sorting; however, let's imagine you've found the two resources below that explain it. Read each one, then try to write a query to answer the following question: What are the minimum and the maximum replacement costs for each rating category ordered by rating as follows: G, PG, PG-13, R, NC-17?

- SQL Server - Custom Sorting in ORDER BY Clause
- Custom Order By in SQL Server

Don't worry if you can't do this bonus task right now. To custom sort the data, you have to use the CASE statement, a concept that you'll cover in the next Exercise.

Command:

SELECT rating, MAX (replacement_cost), MIN (replacement_cost) FROM film GROUP BY rating ORDER BY CASE

when rating = 'G' then 1

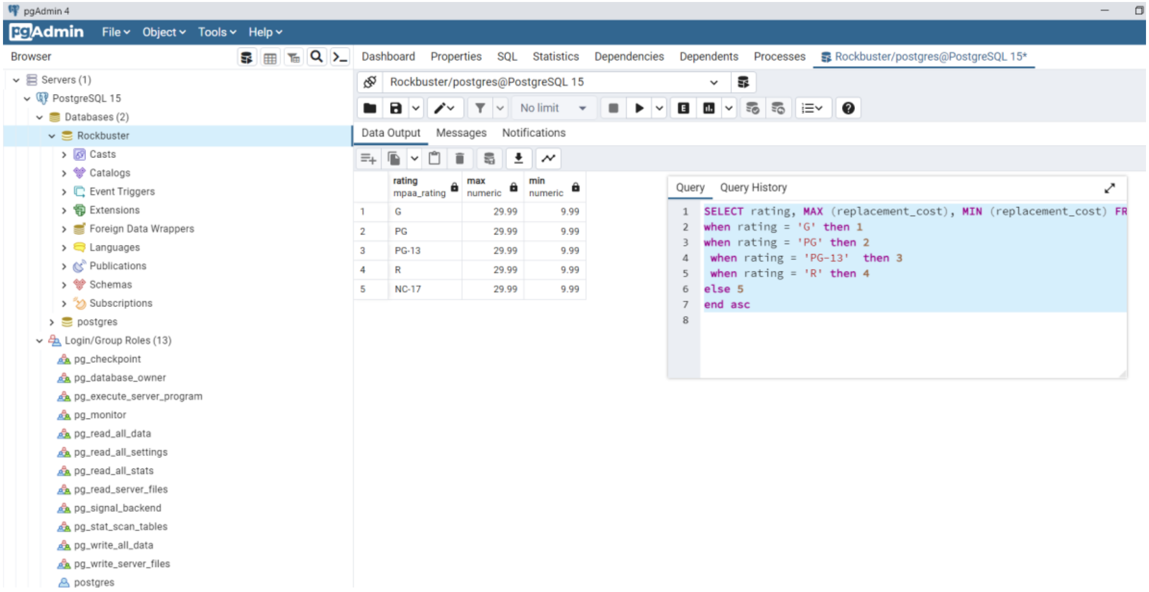
when rating = 'PG' then 2

 when rating = 'PG-13'   then 3

 when rating = 'R' then 4

else 5

end asc



Rockbuster/postgres@PostgreSQL 15

No limit

	rating mpaa_rating	max numeric	min numeric
1	G	29.99	9.99
2	PG	29.99	9.99
3	PG-13	29.99	9.99
4	R	29.99	9.99
5	NC-17	29.99	9.99

```
SELECT rating, MAX (replacement_cost), MIN (replacement_cost) FR
when rating = 'G' then 1
when rating = 'PG' then 2
 when rating = 'PG-13'  then 3
 when rating = 'R' then 4
else 5
end asc
```