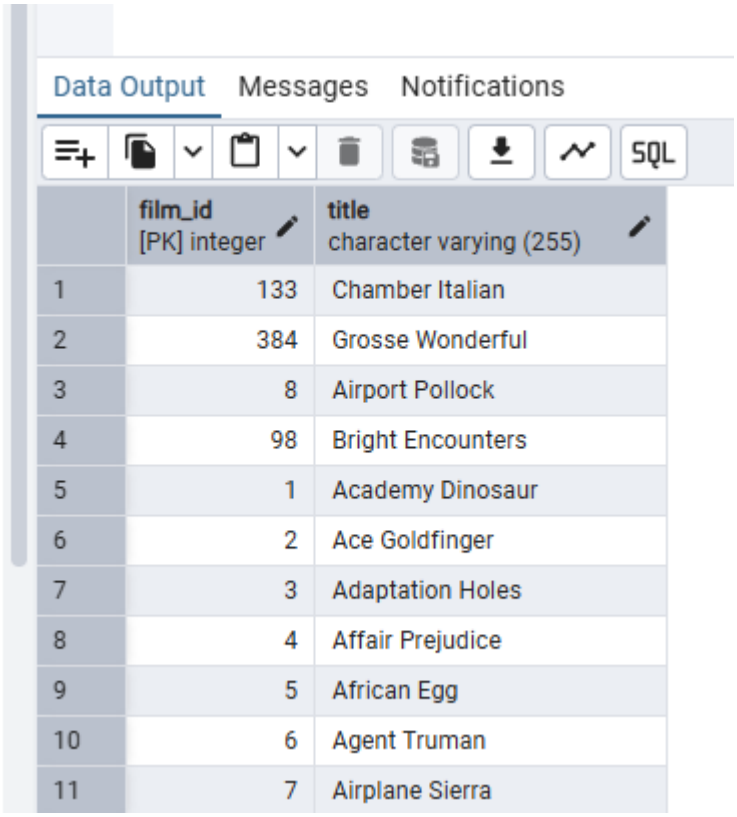
1. **Refining Your Query: You need to get some data from the "film" table and decide to use the query SELECT * FROM film.**
    - **You realize that only the "film_id" and "title" columns are needed. Write a new query that selects only those 2 columns.**

SELECT film_id, title FROM film



Data Output | Messages | Notifications

	film_id [PK] integer	title character varying (255)
1	133	Chamber Italian
2	384	Grosse Wonderful
3	8	Airport Pollock
4	98	Bright Encounters
5	1	Academy Dinosaur
6	2	Ace Goldfinger
7	3	Adaptation Holes
8	4	Affair Prejudice
9	5	African Egg
10	6	Agent Truman
11	7	Airplane Sierra

    - **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?**

EXPLAIN

SELECT * FROM film

Seq Scan on film (cost=0.00..98.00 rows=1000 width=384)

EXPLAIN

SELECT film_id, title FROM film

Seq Scan on film (cost=0.00..98.00 rows=1000 width=19)

Both costs appears to be the same, but in the second query, far less data is being drawn from.

It's unlikely that you ever need to use all of the films at once, there will normally be a reason to look at some of them – so you might want only those of one genre, or only the best rated ones etc.

I tried just:

SELECT DISTINCT title FROM film

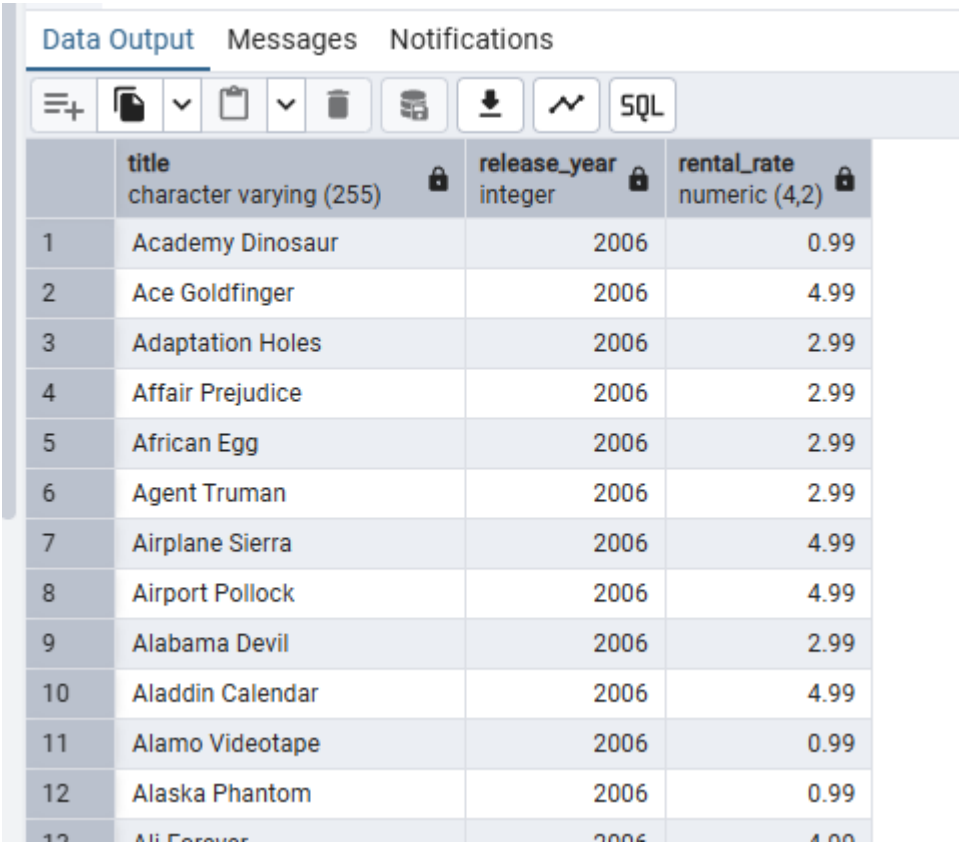
But that appears to be more expensive, and doesn't give the ID which may be necessary

HashAggregate (cost=100.50..110.50 rows=1000 width=15)

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 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

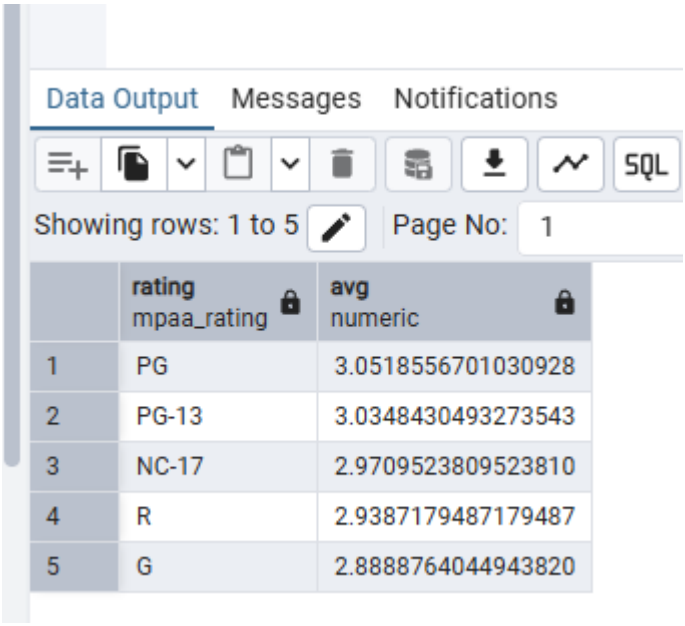
This doesn't make much sense as a query. Every title is unique, so nothing is being ordered after the title.

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

ORDER BY AVG(rental_rate) DESC



Data Output | Messages | Notifications

Showing rows: 1 to 5 | Page No: 1

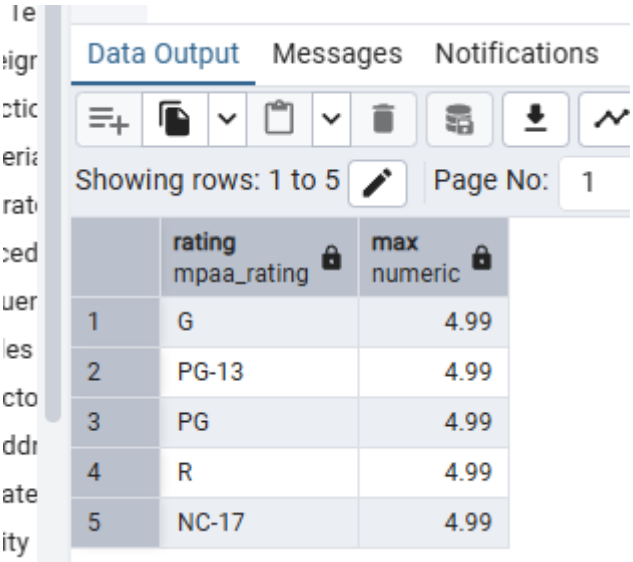
	rating mpaa_rating	avg numeric
1	PG	3.0518556701030928
2	PG-13	3.0348430493273543
3	NC-17	2.9709523809523810
4	R	2.9387179487179487
5	G	2.8888764044943820

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

SELECT rating, MAX(rental_rate) FROM film

GROUP BY rating

ORDER BY MAX(rental_rate) DESC

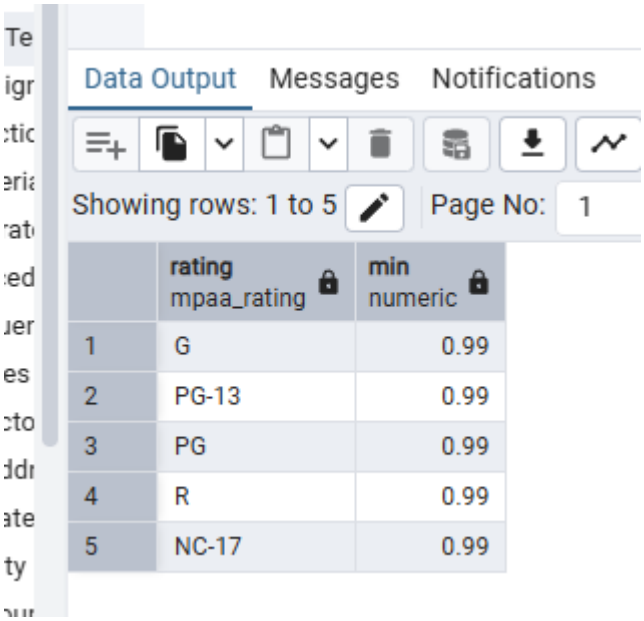


Data Output | Messages | Notifications

Showing rows: 1 to 5 | Page No: 1

	rating mpaa_rating	max numeric
1	G	4.99
2	PG-13	4.99
3	PG	4.99
4	R	4.99
5	NC-17	4.99

```
SELECT rating, MIN(rental_rate) FROM film
GROUP BY rating
ORDER BY MIN(rental_rate) DESC
```



Data Output | Messages | Notifications

Showing rows: 1 to 5 | Page No: 1

	rating mpaa_rating	min numeric
1	G	0.99
2	PG-13	0.99
3	PG	0.99
4	R	0.99
5	NC-17	0.99

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?

Extract: The initial stage is gathering data from that source

Transform: Then data is modified to fit the required format.

Load: Finally, the processed data is transferred and stored in the database.

A data Engineer should be responsible for this process.

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

You might start aggregating the date in ways which mean you can't then go into the detail you need alter on. Best to keep all the data, then you can use whatever you need.

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

Needed a bit of help from Co-Pilot for this

```
SELECT rating, 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
   WHEN rating = 'NC-17' THEN 5
   ELSE 6
END;
```

Query | Query History

```sql
SELECT rating, 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
    WHEN rating = 'NC-17' THEN 5
    ELSE 6
END;
```

Data Output | Messages | Notifications

Showing rows: 1 to 5 | Page No: 1 of 1

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