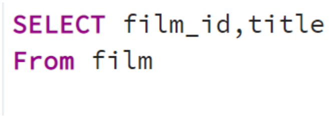
ANSWERS 3.4

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

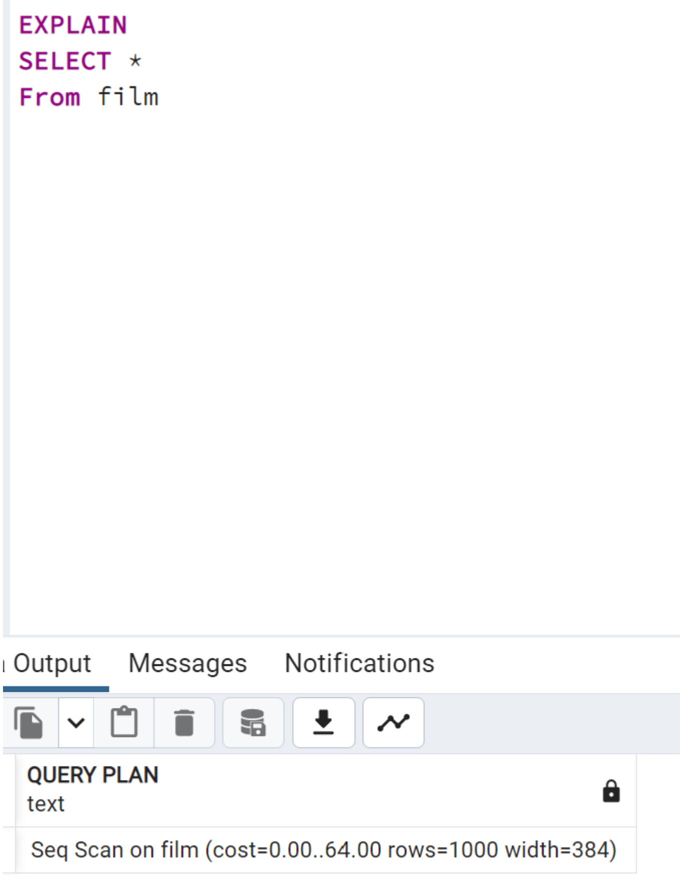
   - You realize that only the "film_id" and "title" columns are needed. Write a new query that selects only those 2 columns.



```
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?

     - COST OF ORIGINAL QUERY



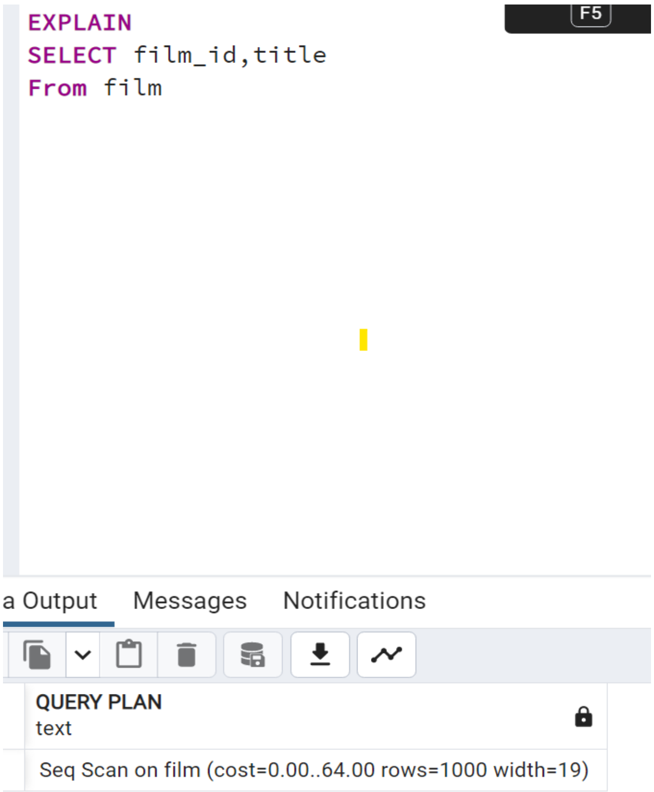
```
EXPLAIN
SELECT *
From film
```

QUERY PLAN
text

Seq Scan on film (cost=0.00..64.00 rows=1000 width=384)

ANSWERS 3.4

- COST OF NEW QUERY

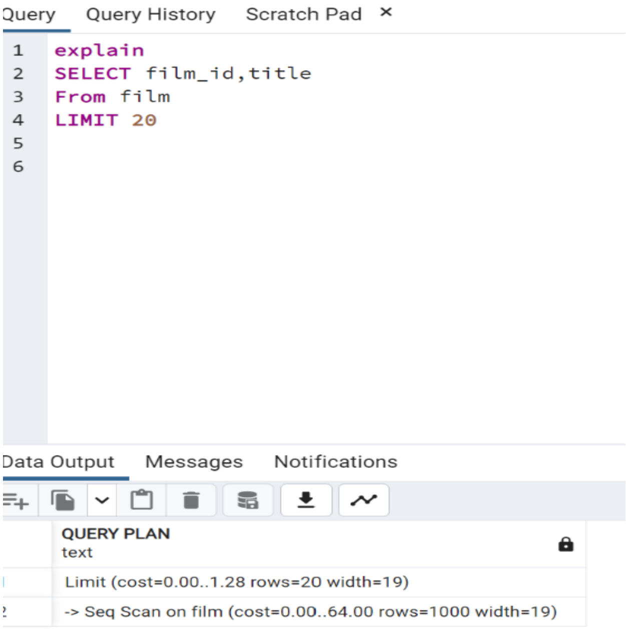


```
EXPLAIN
SELECT film_id,title
From film
```

QUERY PLAN
text

Seq Scan on film (cost=0.00..64.00 rows=1000 width=19)

We see from the above that there is no change in the cost when executing the two different queries. We can refine the syntax by adding limit to lower the cost.



```
explain
SELECT film_id,title
From film
LIMIT 20
```

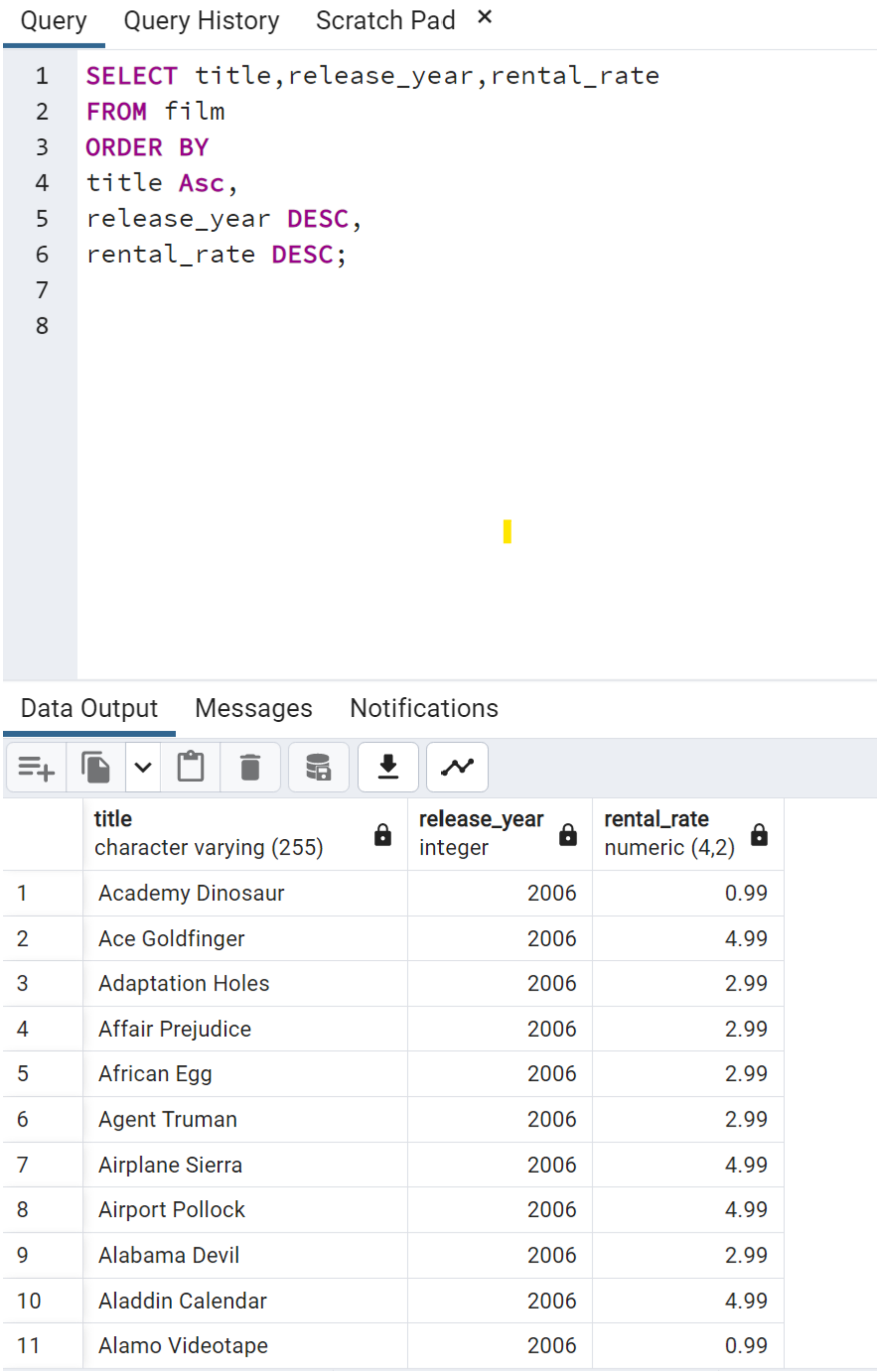
QUERY PLAN
text

Limit (cost=0.00..1.28 rows=20 width=19)
-> Seq Scan on film (cost=0.00..64.00 rows=1000 width=19)

Here we can see that by applying Limit, the cost has reduced to 1.28.

ANSWERS 3.4

2. Ordering the Data:



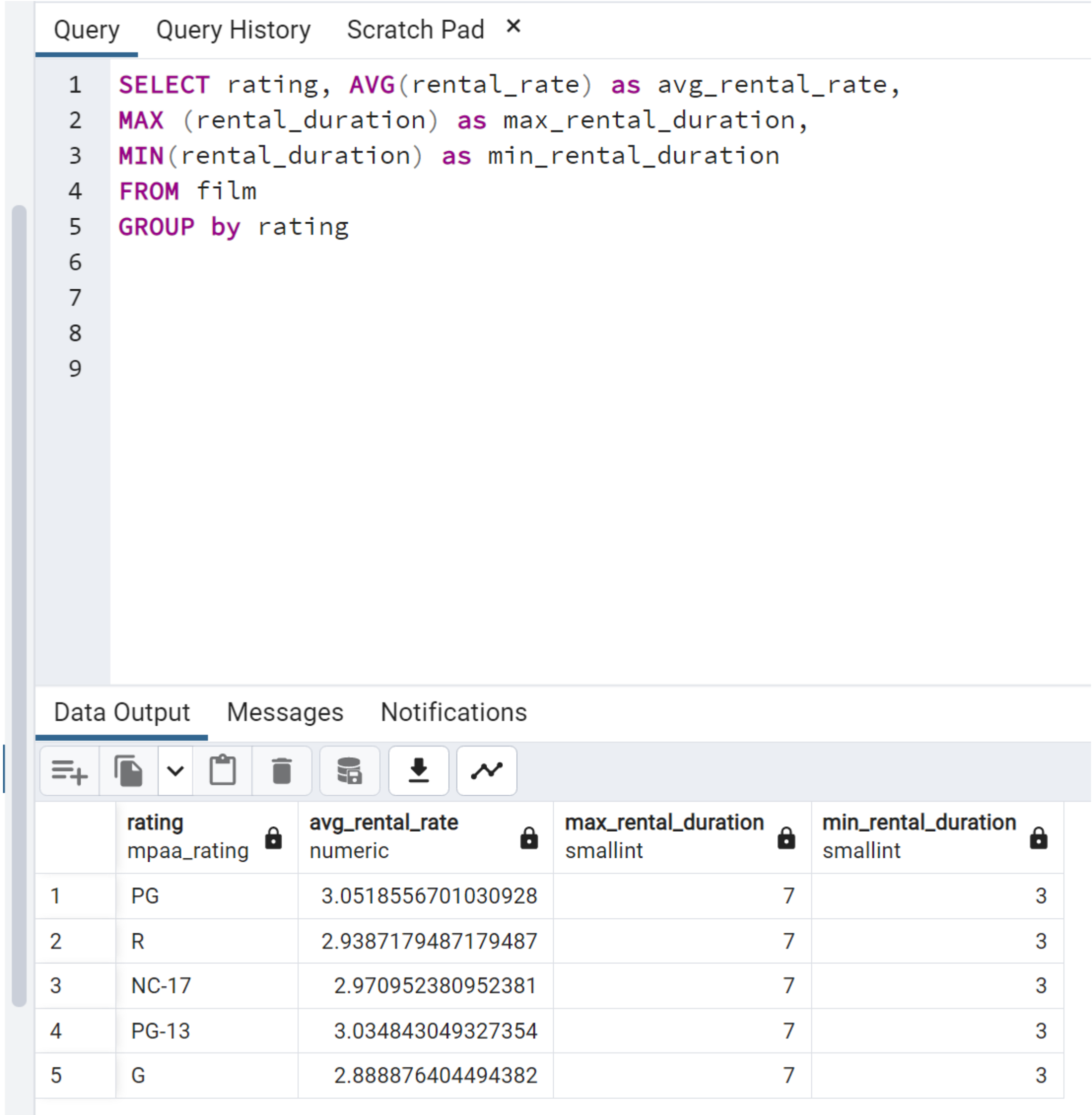
```
SELECT title,release_year,rental_rate
FROM film
ORDER BY
title Asc,
release_year DESC,
rental_rate DESC;
```

	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

ANSWERS 3.4

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



```sql
SELECT rating, AVG(rental_rate) as avg_rental_rate,
MAX (rental_duration) as max_rental_duration,
MIN(rental_duration) as min_rental_duration
FROM film
GROUP by rating
```

	rating mpaa_rating	avg_rental_rate numeric	max_rental_duration smallint	min_rental_duration smallint
1	PG	3.0518556701030928	7	3
2	R	2.9387179487179487	7	3
3	NC-17	2.970952380952381	7	3
4	PG-13	3.034843049327354	7	3
5	G	2.888876404494382	7	3

ANSWERS 3.4

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?

     - First step is to **EXTRACT** the data from the system sources
     - Then the extracted data is converted or **TRANSFORMED** into another format (as per need).
     - And finally, this transformed data is **LOADED** into the new database
     - ETL is primarily is a data engineer's job, but as a Data Analyst, one should have awareness about its basic concepts so that they are able to coordinate with Data Engineers and make sense of the timelines in the migration process.

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

     - If an analyst tries to begin analyzing data before it has been loaded into the data warehouse, the data will not be consistently formatted. As a result, it would be very difficult to retrieve and manipulate the data or draw any meaningful conclusions from it.