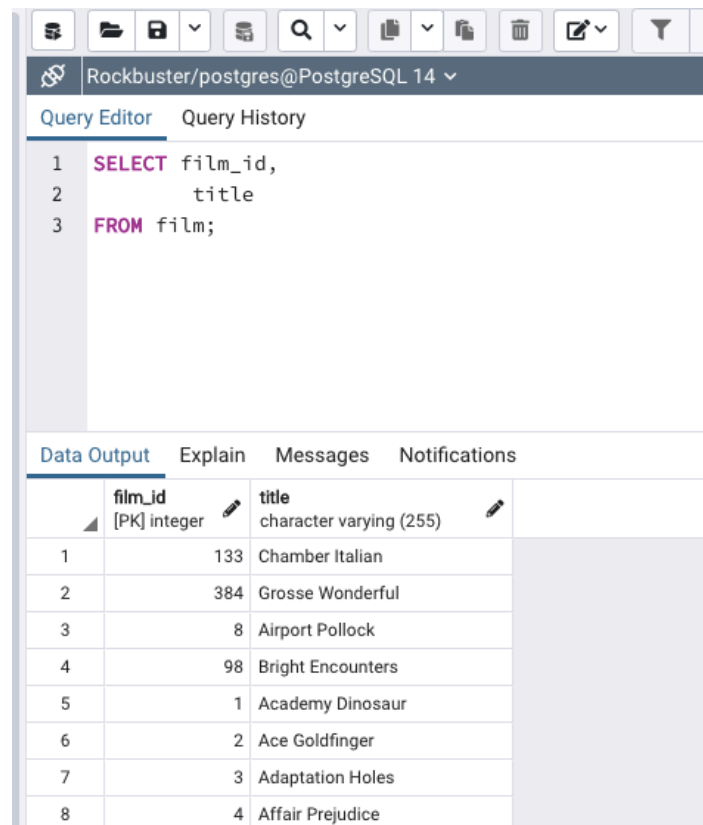


### 3.4 Database Querying in SQL

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.



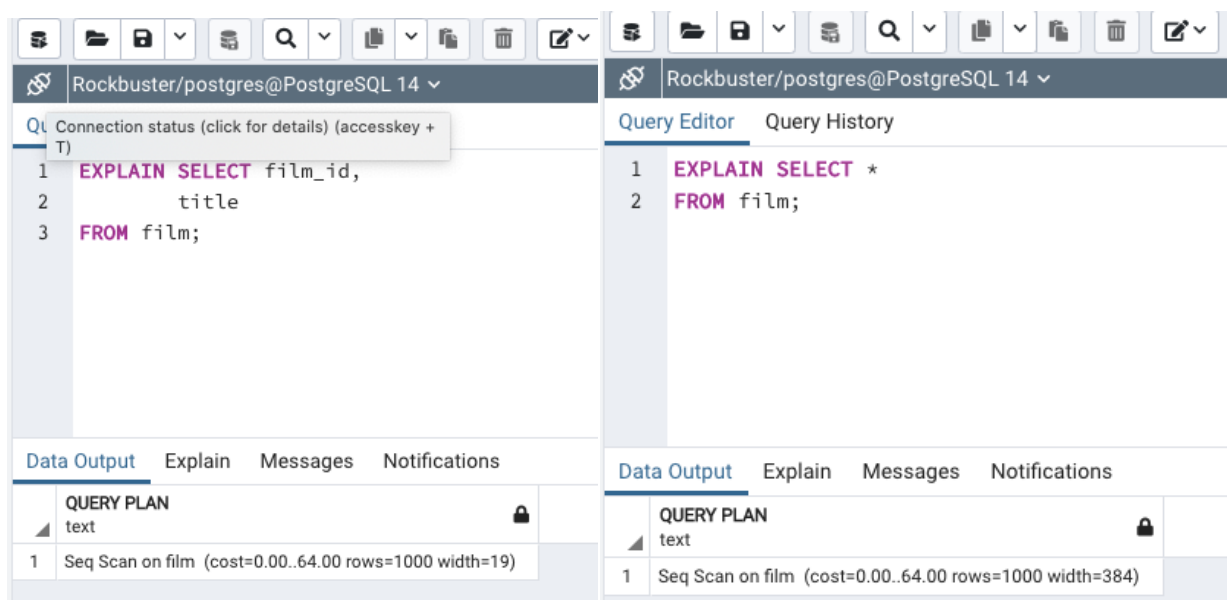
The screenshot shows a PostgreSQL query editor interface. The top toolbar contains icons for file operations, search, and execution. The connection bar shows 'Rockbuster/postgres@PostgreSQL 14'. The 'Query Editor' tab is active, displaying the following SQL query:

```
1 SELECT film_id,  
2       title  
3 FROM film;
```

Below the query editor, the 'Data Output' tab is selected, showing the results of the query in a table format. The table has two columns: 'film\_id' (integer) and 'title' (character varying (255)).

	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

- 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?



The screenshot shows two side-by-side PostgreSQL query editor windows. Both windows show the same connection: 'Rockbuster/postgres@PostgreSQL 14'.

The left window displays the following SQL query:

```
1 EXPLAIN SELECT film_id,  
2       title  
3 FROM film;
```

The right window displays the following SQL query:

```
1 EXPLAIN SELECT *  
2 FROM film;
```

Both windows show the 'Data Output' tab with the 'QUERY PLAN' section expanded. The query plan for both queries is a 'Seq Scan on film' with a cost of 0.00..64.00, 1000 rows, and a width of 19 (left) or 384 (right).

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

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

The two queries cost of the returning the first row is 0 and all the rows is 64. Therefore, they have the same cost.

## 2. Ordering the Data:

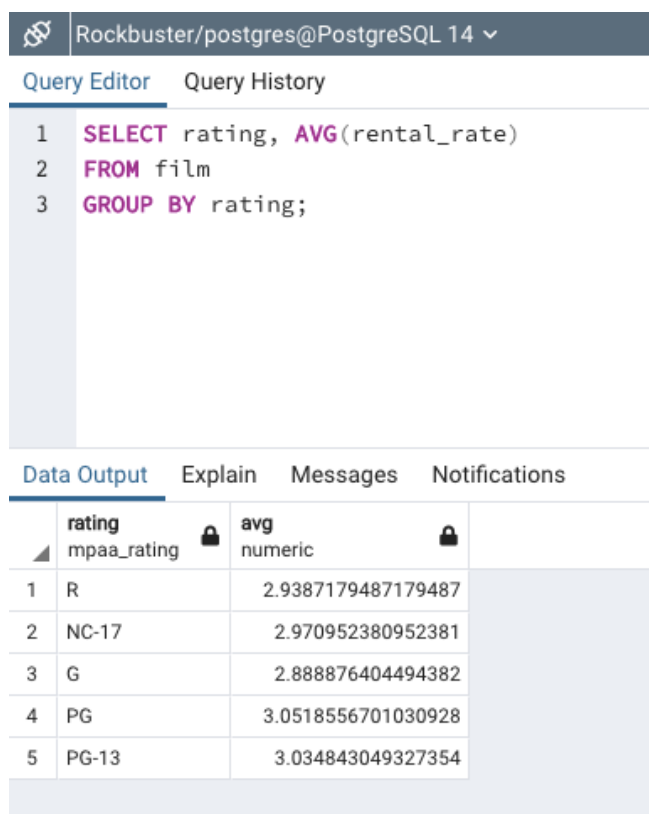
- In the pgAdmin Query Tool, run a query that selects every film from the “film” table, with the movies sorted by title from A to Z, then by most recent release year, and then by highest to lowest rental rate.

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

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

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?



The screenshot shows the pgAdmin Query Tool interface. At the top, the connection is 'Rockbuster/postgres@PostgreSQL 14'. Below the connection bar, there are two tabs: 'Query Editor' and 'Query History'. The 'Query Editor' tab is active, showing a SQL query:

```
1 SELECT rating, AVG(rental_rate)  
2 FROM film  
3 GROUP BY rating;
```

Below the query editor, there are four tabs: 'Data Output', 'Explain', 'Messages', and 'Notifications'. The 'Data Output' tab is active, showing the results of the query in a table. The table has two columns: 'rating' (mpaa\_rating) and 'avg' (numeric). The results are as follows:

	rating mpaa_rating	avg numeric
1	R	2.9387179487179487
2	NC-17	2.970952380952381
3	G	2.888876404494382
4	PG	3.0518556701030928
5	PG-13	3.034843049327354

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

Rockbuster/postgres@PostgreSQL 14

Query Editor Query History

```

1 SELECT rating, MIN(rental_duration)
2 FROM film
3 GROUP BY rating;
4

```

Data Output Explain Messages Notifications

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

Rockbuster/postgres@PostgreSQL 14

Query Editor Query History

```

1 SELECT rating, MAX(rental_duration)
2 FROM film
3 GROUP BY rating;
4

```

Data Output Explain Messages Notifications

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

**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 for migrating data is called ETL (Extract, Load and Transfer) and it is primarily the responsibility of the Data Engineer. The user behaviour data is collected (Extract), then transform into appropriate format before loaded into the data warehouse.

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

It will be impossible to analyse the data in relation to other acquire data due to no interconnectivity. Therefore, drawing insight from the data might be difficult thereby reducing productivity of decision making