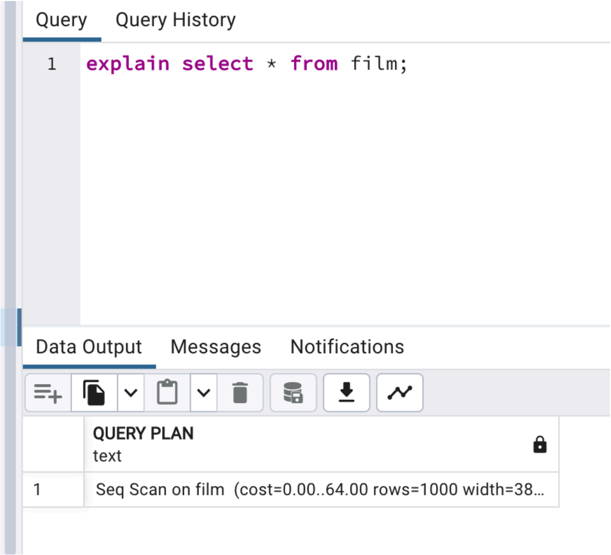
3.4: Database Querying in SQL

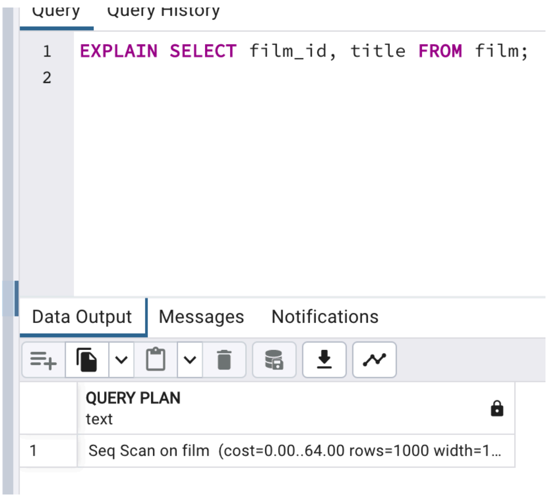
1. **Refining Your Query:** You need to get some data from the “film” table and decide to use the query SELECT * FROM film.
    - ○ You realize that only the “film_id” and “title” columns are needed. Write a new query that selects only those 2 columns.
    - ○ Compare the cost of the original query and the revised query, and write a few sentences explaining the comparison. Can you suggest any ways to optimize this query?

Answer:



```
explain select * from film;
```

	QUERY PLAN text
1	Seq Scan on film (cost=0.00..64.00 rows=1000 width=38…



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

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

SELECT * FROM film.
Seq Scan on film (cost=0.00..64.00 rows=1000 width=388)

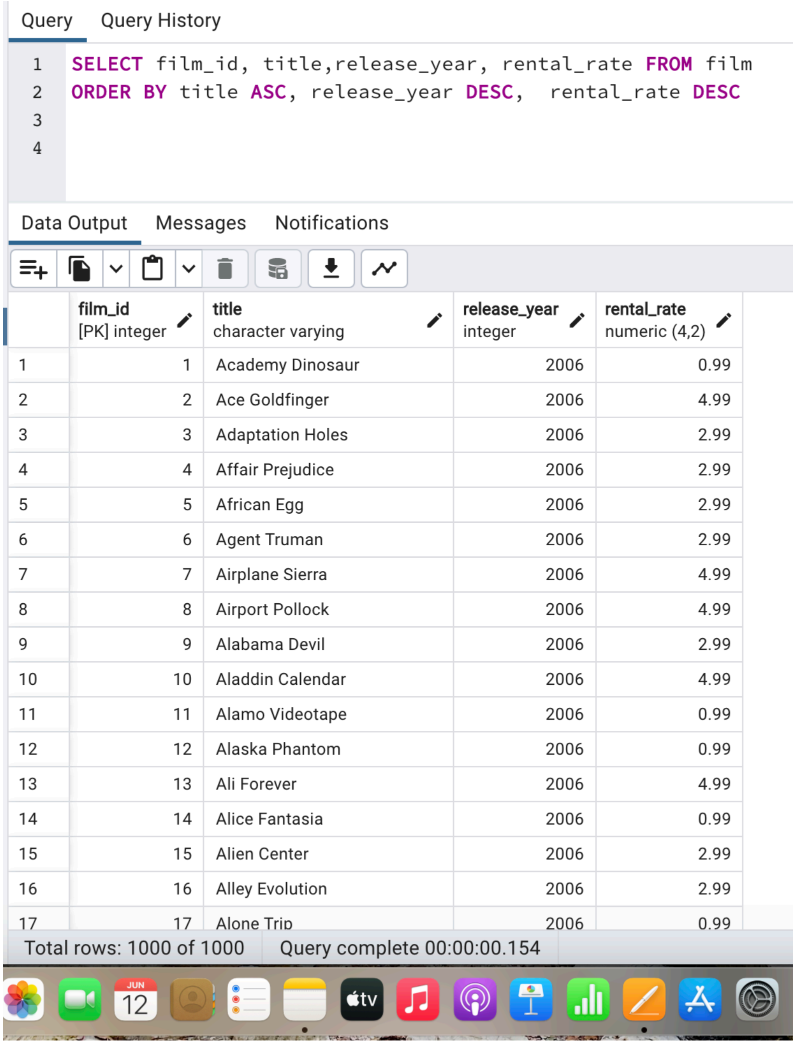
EXPLAIN SELECT film_id, title FROM film;
Seq Scan on film (cost=0.00..64.00 rows=1000 width=19)

The cost of two queries (the original query and the revised query) has the same cost. The revised query returns few columns as original query. We could also use the LIMIT command and put them in a particular order by ASC or DESC for better improvement. We can also use WHERE clause to improve the query performance by reducing the number of rows.

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.
- 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”:

Answer:



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

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

```
Total rows: 1000 of 1000    Query complete 00:00:00.154
```

-- Display film title, release year, rental_rate in descending and ascending order.

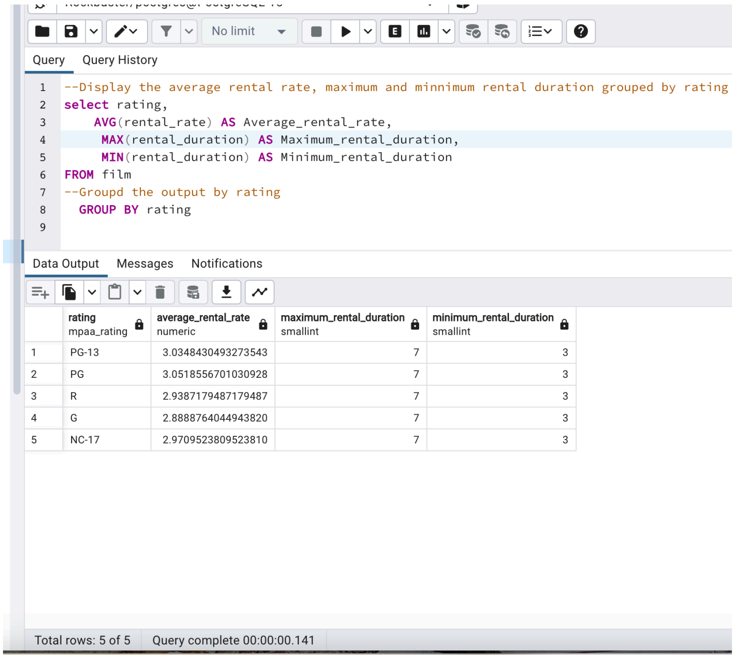
SELECT film_id, title, release_year, rental_rate FROM film

ORDER BY title ASC, release_year DESC, rental_rate DESC

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?

--Display the average rental rate grouped by rating



```
--Display the average rental rate, maximum and minnimum rental duration grouped by rating
select rating,
    AVG(rental_rate) AS Average_rental_rate,
     MAX(rental_duration) AS Maximum_rental_duration,
     MIN(rental_duration) AS Minimum_rental_duration
FROM film
--Groupd the output by rating
  GROUP BY rating
```

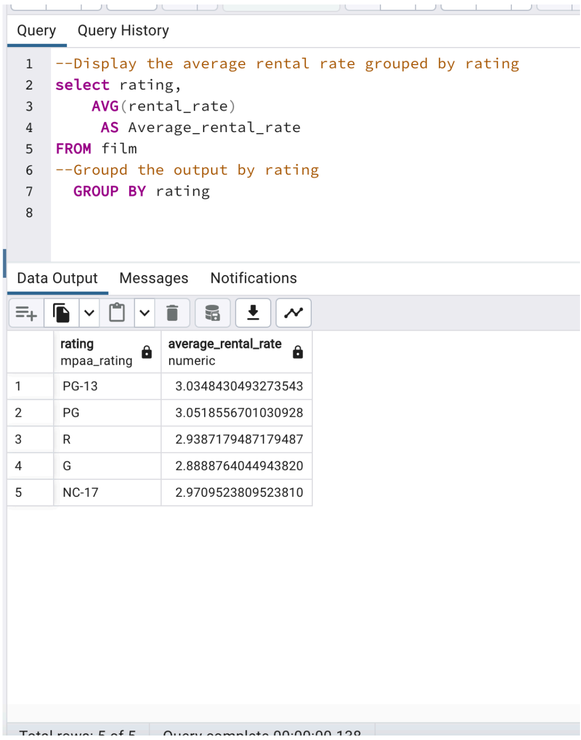
	rating mpaa_rating	average_rental_rate numeric	maximum_rental_duration smallint	minimum_rental_duration smallint
1	PG-13	3.0348430493273543	7	3
2	PG	3.0518556701030928	7	3
3	R	2.9387179487179487	7	3
4	G	2.8888764044943820	7	3
5	NC-17	2.9709523809523810	7	3

Total rows: 5 of 5   Query complete 00:00:00.141

```
SELECT rating,
   AVG(rental_rate)
      AS Average_rental_rate
FROM film
--Grouped the output by rating
  GROUP BY rating
```

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

--Display the average rental rate, maximum and minimum rental duration grouped by rating



```
--Display the average rental rate grouped by rating
select rating,
    AVG(rental_rate)
     AS Average_rental_rate
FROM film
--Groupd the output by rating
  GROUP BY rating
```

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

Total rows: 5 of 5   Query complete 00:00:00.128

```
SELECT rating,
   AVG(rental_rate) AS Average_rental_rate,
   MAX(rental_duration) AS
Maximum_rental_duration,
   MIN(rental_duration) AS
Minimum_rental_duration
FROM film
--Grouped the output by rating
  GROUP BY rating
```

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

Procedure:

The process of migrating the data is generally performed by Data Engineers. If we are aware of this process then this ETL process will allow us to coordinate with data engineers and make sense of the timelines in the migration process. Every step of the ETL process is very important to identify errors and problems in data. First the data should be extracted from other resources which is on user behaviour for the new Rockbuster Android app. Then this data would be transformed which converts the data into appropriate format. Finally the data loaded into the warehouse for analysis. Once the data is in the data warehouse, we can use it to answer important questions.

Problems:

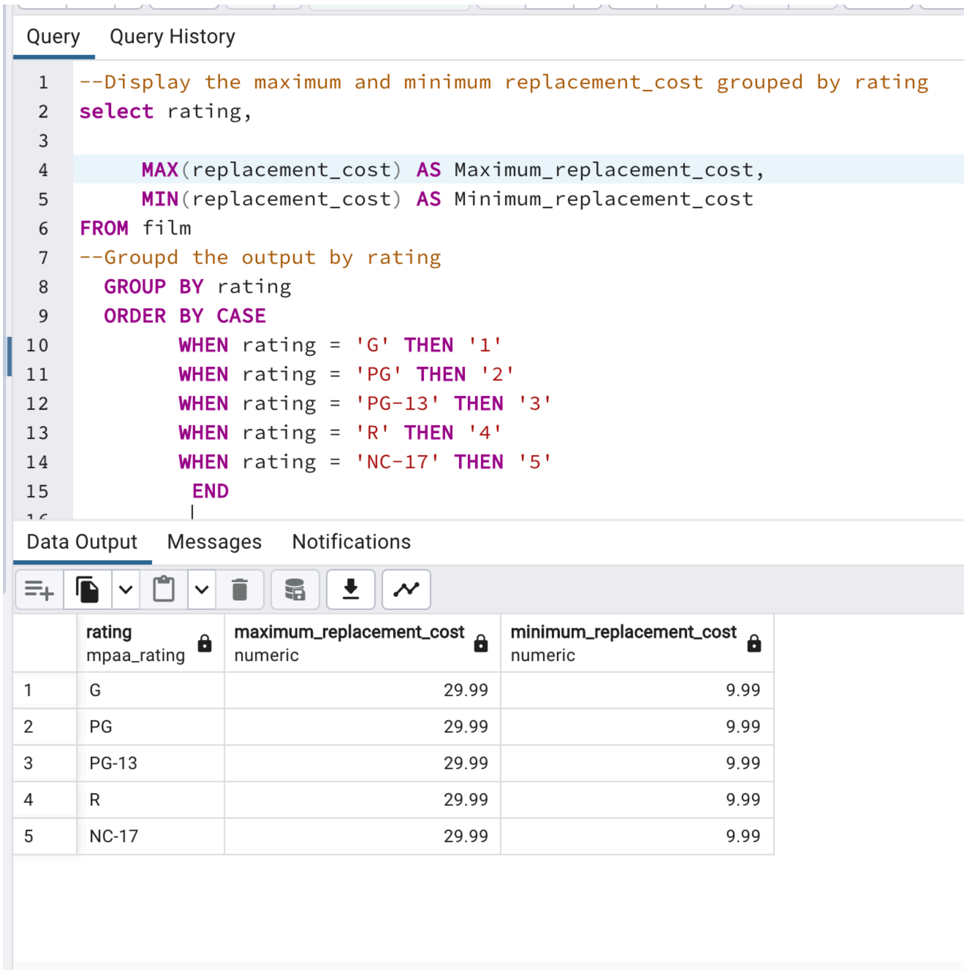
If we analyse data from one data source before it has been extracted and converted, we might face some missing data and problems In data. Also we may be less familiar with the format the data is in. There could also be potential errors in the data that need to be cleaned before being analyzed. the greater potential there is to have incompatible inform.

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?

```
--Display the maximum and minimum replacement_cost grouped by rating
SELECT rating,

          MAX(replacement_cost) AS Maximum_replacement_cost,
          MIN(replacement_cost) AS Minimum_replacement_cost
FROM film
--Grouped the output by rating
  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'
                    END
```



```sql
--Display the maximum and minimum replacement_cost grouped by rating
select rating,

     MAX(replacement_cost) AS Maximum_replacement_cost,
     MIN(replacement_cost) AS Minimum_replacement_cost
FROM film
--Groupd the output by rating
  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'
         END
```

	rating mpaa_rating	maximum_replacement_cost numeric	minimum_replacement_cost 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