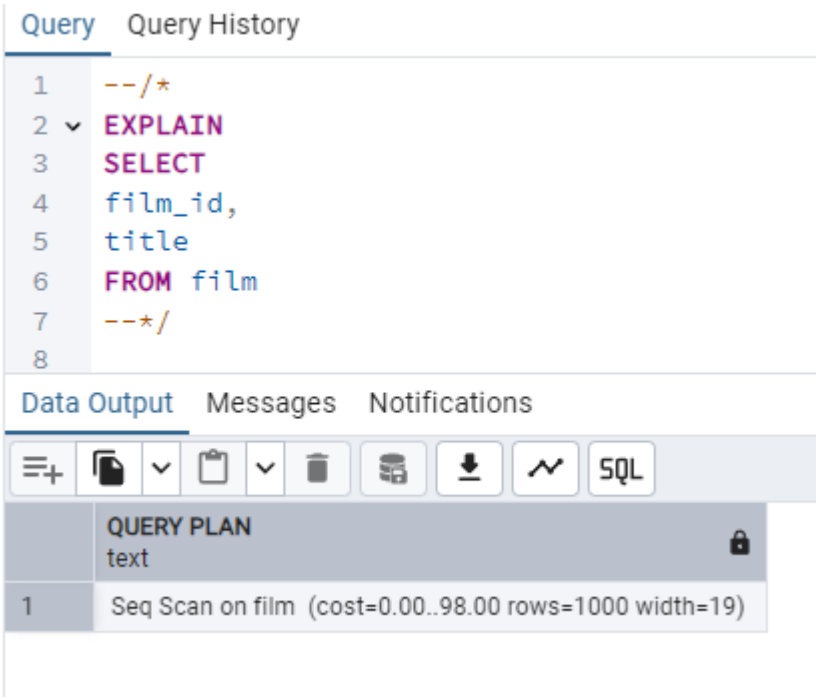
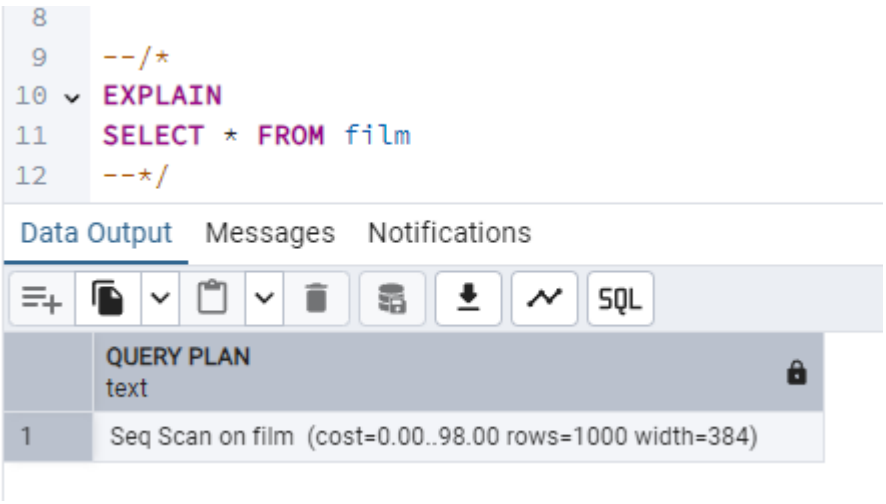
Answers 3.4

❖ **Refining Your Query**





- The cost for both the original query (**SELECT * FROM film**;) and the revised query (**SELECT film_id, title FROM film**;) remains the. However, the revised query has a significantly smaller width (**19 vs. 384**), meaning it processes less data by retrieving only two columns instead of all.

- **To optimize the query**:

✓ Avoid using **SELECT *** to reduce data retrieval overhead.

✓ Create an index on the film_id and title columns to improve query performance, especially for large tables.

❖ 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.

E3.4-title.csv E3.4-rental_rate.csv E3.4-release_year.csv E3.4-ALL.xlsx

❖ Grouping Data

- What is the average rental rate for each rating category?

	rating mpaa_rating	avg_rental_rate numeric
1	PG	3.05
2	PG-13	3.04
3	NC-17	3.00
4	R	2.94
5	G	2.89

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

	rating mpaa_rating	min_rental_duration numeric	max_rental_duration numeric
1	G	3.00	7.00
2	NC-17	3.00	7.00
3	PG	3.00	7.00
4	PG-13	3.00	7.00
5	R	3.00	7.00

❖ Database Migration

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

1. **Extract Data**: Use the external tool to collect user behavior data and export it in a standardized format (e.g., CSV, JSON).
2. **Transform Data**: Clean and preprocess the data to align with the schema of the data warehouse. This may include deduplication, data type conversions, and validation.
3. **Load Data**: Import the transformed data into the data warehouse using ETL tools like Talend, Informatica, or native database tools.
4. **Verification**: Validate the loaded data to ensure accuracy and consistency.
5. **Responsibilities**:
    - **Data Engineers**: Oversee the ETL process and ensure the pipeline is robust.
    - **Data Analysts**: Validate data quality post-migration.
    - **IT Team**: Handle any infrastructure or tool-related issues.

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

1. **Incomplete Data**: Analysis might be based on partial data, leading to inaccurate insights.
2. **Inconsistency**: Without preprocessing, the data may contain duplicates or errors, reducing reliability.
3. **Performance Issues**: Accessing raw external data directly can strain systems and increase latency.
4. **Integration Challenges**: Data might not align with existing datasets, complicating analysis.

❖ Bonus Task

```sql
SELECT
rating,
ROUND(MIN(replacement_cost),2) AS min_replacement_cost,
ROUND(MAX(replacement_cost),2) AS max_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'
END
```

Data Output | Messages | Notifications

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