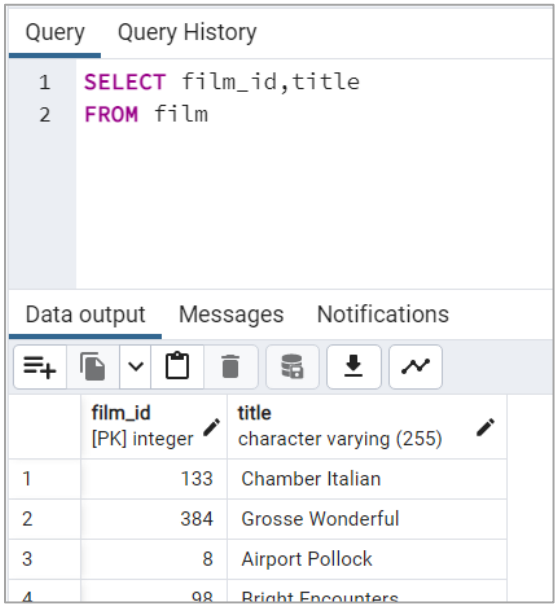
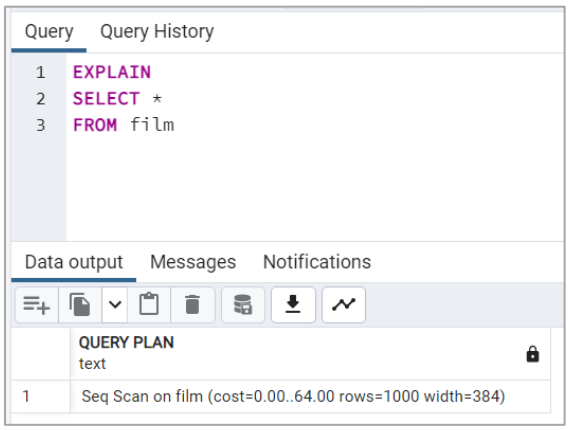
Exercise 3.4 - Database Querying in SQL

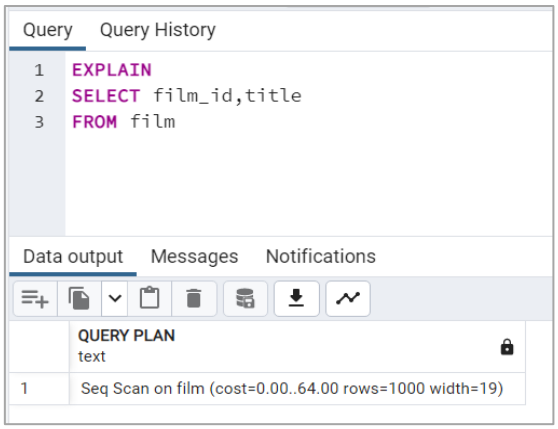
**Step 1: Refining Your Query**

- 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?




→ From the data output message, the cost of the original query and the revised query are the same (cost=0.00..64.00). Although they have different runtimes as shown in each of their output. The best way to optimise and safe cost is to create a script.

**Step 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.

Query | Query History

```
1  SELECT title,
2  release_year, rental_rate
3  FROM film
4  ORDER BY title ASC,
5  release_year DESC,
6  rental_rate DESC
```

Data output | Messages | Notifications

| | 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 |

- 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”:

→Done.

**Step 3: Grouping Data**

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

Query | Query History

```
1  SELECT rating,
2  AVG(rental_rate)
3  AS average_rental_rate
4  FROM film
5  GROUP BY rating
```

Data output | Messages | Notifications

| | rating mpaa_rating | average_rental_rate 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?

Query | Query History

```
SELECT rating,
MAX(rental_rate)
AS maximum_rental_rate,
MIN(rental_rate)
AS nminimum_rental_rate
FROM film
GROUP BY rating
```

Data output | Messages | Notifications

| | rating mpaa_rating | maximum_rental_rate numeric | nminimum_rental_rate numeric |
|---|---|---|---|
| 1 | R | 4.99 | 0.99 |
| 2 | NC-17 | 4.99 | 0.99 |
| 3 | G | 4.99 | 0.99 |
| 4 | PG | 4.99 | 0.99 |
| 5 | PG-13 | 4.99 | 0.99 |

**Step 4 : Database Migration**

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

→The migration will be done via ETL (Extract, Transform, Load). This is basically carried out by data engineers. What each step entails is described as follows:

- Extract: This the first step and it involves collection of data from various data sources
- Transform: In this step, the extracted data is converted into another format. This could mean calculating ages from dates of birth or combining multiple data points like area codes and telephone numbers to get a contact number, for example.
- Load: In this step, the transformed data is inserted or loaded into the new database.

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

→They might be a problem of cohesiveness in the data, such as formatting issues, inclusion of irrelevant data etc. This will result to time and cost consuming for the analyst.