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.

```
select film_id,
title
from film
```

(Personally, I like to add some order, so I add an 'order by' statement with film_id asc)

    - 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?

Seq Scan on film (cost=0.00..64.00 rows=1000 width=384) - select * from film

Seq Scan on film (cost=0.00..64.00 rows=1000 width=19) - select film_id, title from film

a. The costs are the same, the output is just simply "thinner" in that there is a much smaller width. One suggestion for looking at these titles would be to include an order by and filter the 'film_id'. – In this case, the film_id was put into this data base associated with the alphabetical order of the titles; a useful tool when simply looking at a title and film_id selection. Replacing film_id for title is also a valid method for alphabetizing the list

    1. One thing of note, this additional command, order by, does increase the cost value:

        i. Index Scan using film_pkey on film (cost=0.28..92.92 rows=1000 width=19)

        ii. This suggests that utilizing this function does put a bit more strain on the computing power and resources as well as time to complete

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 film_id,
title
from film
order by title asc

select film_id,
title
from film
order by release_year asc

select film_id,
title
from film
order by 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?
        a. "G" 2.8888764044943820
           "R" 2.9387179487179487
           "NC-17" 2.9709523809523810
           "PG-13" 3.0348430493273543
           "PG" 3.0518556701030928
        b. Syntax:

```
SELECT rating,
avg (rental_rate)
FROM film
group by rating
```

order by avg asc

- What are the minimum and maximum rental durations for each rating category? (this is consistent across all rating categories).
    a. Min rental_duration: 3
    b. Max rental_duration: 7

```
SELECT rating,
min (rental_duration)
FROM film
group by rating

SELECT rating,
max (rental_duration)
FROM film
group by rating
```

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?
        a. Generally speaking, the procedure for this is extract, transform, and load or ETL. The data source where these data are extracted are the new app, the transformation is altering source syntax and output from the app to a format that a DBMS like pgadmin4 could utilize, and the loading is moving that data into a DBMS. Data engineering team would be responsible for the extraction and transformation while the data analyst would load and analyze data from that file.
    - What problems do you foresee if you start analyzing the data before it's been loaded into the data warehouse?
        a. Whilst I do not have direct experience with this, I imagine that the major issues that could come about are related to the translation step – namely in syntax. If the syntax for the app were different for an outside DBMS like pgadmin4, I suspect that there could be various errors – both code and human entry errors – that could arise. But once there was a successful transformation and translation,

in general, I image it would work well enough for an analyst to extract data and make sense of it.

b. With the above in mind, as long as there is an open dialogue with the engineering team and the analysts, there shouldn't be too many problems.

5. Combine your "Answers 3.4" document and csv files into a single PDF and upload it here for your tutor to review.