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
```

- *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 cost of both queries is the same (64). The 2nd query took 65msec to retrieve the data vs. 141 msec with the 1st one. With large volume of data, the 2nd query is more efficient, it returns not all data, but the specific rows/columns ordered.

To optimize the query one can use LIMIT (or TOP) if it's necessary to restrict the number of rows to be pulled from the database.

Original query:

```
EXPLAIN SELECT *
FROM film
```

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

Revised query:

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

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

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 title, release_year, rental_rate
FROM film
ORDER BY title, release_year, rental_rate ASC
```

	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
5	African Egg	2006	2.99
6	Agent Truman	2006	2.99
7	Airplane Sierra	2006	4.99
8	Airport Pollock	2006	4.99
9	Alabama Devil	2006	2.99
10	Aladdin Calendar	2006	4.99
11	Alamo Videotape	2006	0.99

- *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.)*

Done: Task 3.4 data for film collection department.csv

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?*

     Task 3.4 Rental rate for stradegy department.xlsx

```
SELECT rating,
       AVG(rental_rate)
FROM film
GROUP by rating
```

	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?*

```
SELECT rating, MIN(rental_duration),
MAX(rental_duration)
FROM film
GROUP by rating
```

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

For the migrating process of data Data Engineers will use the procedure of ETL (Extract, Transform, Load). Firstly, they will collect the data about user behavior from the Rockbuster Android app. Secondly, the data will be converted into the necessary format in the warehouse. Lastly, the data will be loaded into the database for analysis.

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

Analyzing the data before it's been loaded into the data warehouse can be ineffective and time-consuming: the format of the data might not the same as in the warehouse or the data might not have structure at all that can make analysis hard or even impossible.

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?

```
SELECT rating,
MIN(replacement_cost) AS min_replacement_cost,
MAX(replacement_cost) AS max_replacement_cost
FROM film
GROUP BY rating
ORDER BY rating
```

	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