







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?

Query	Query History
1	EXPLAIN
2	SELECT *
3	FROM film

Data output	Messages	Notifications
     		
QUERY PLAN text		
1	Seq Scan on film (cost=0.00..64.00 rows=1000 width=388)	

Query	Query History
1	EXPLAIN
2	SELECT film_id, title
3	FROM film

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

The cost for the original query and the revised query are both the same (0.00..64.00), however, from an efficiency standpoint it is always better to refine the query so that only the information that is needed is displayed. This reduced the amount of time that data analyst have to sift through unneeded data.

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 DESC, rental_rate DESC
```

Query Query History

```

1 SELECT title, release_year, rental_rate
2 FROM film
3 ORDER BY title, release_year DESC, rental_rate DESC
4

```

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
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
12	Alaska Phantom	2006	0.99
13	Ali Forever	2006	4.99
14	Alice Fantasia	2006	0.99
15	Alien Center	2006	2.99
16	Alley Evolution	2006	2.99
17	Alone Trip	2006	0.99
18	Alter Victory	2006	0.99
19	Amadeus Holy	2006	0.99
20	Amelle Hellfighters	2006	4.99
21	American Circus	2006	4.99

Total rows: 1000 of 1000 Query complete 00:00:00.052

- 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”

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?

Query Query History

```

1 SELECT rating, AVG(rental_rate)
2 FROM film
3 GROUP BY rating

```

Data output Messages Notifications

	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?

```
1 SELECT rating, MIN(rental_duration), MAX(rental_duration)
2 FROM film
3 GROUP BY rating
```

Data output Messages Notifications				
	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?

The procedure for migrating the data is the Extract, Transform, and Load (ETL) procedure. Data is collected, converted into another format, and then inserted or loaded into another database. This process is completed by data engineers.

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

If data is analyzed before its been loaded into the data warehouse there may be inconsistencies with the data formatting, and accuracy as the entire ETL process has not been completed.

5. Save your "Answers 3.4" document as a pdf (with screenshots) and your csv files as a single .xlsx Excel file and upload it here for your tutor to review.