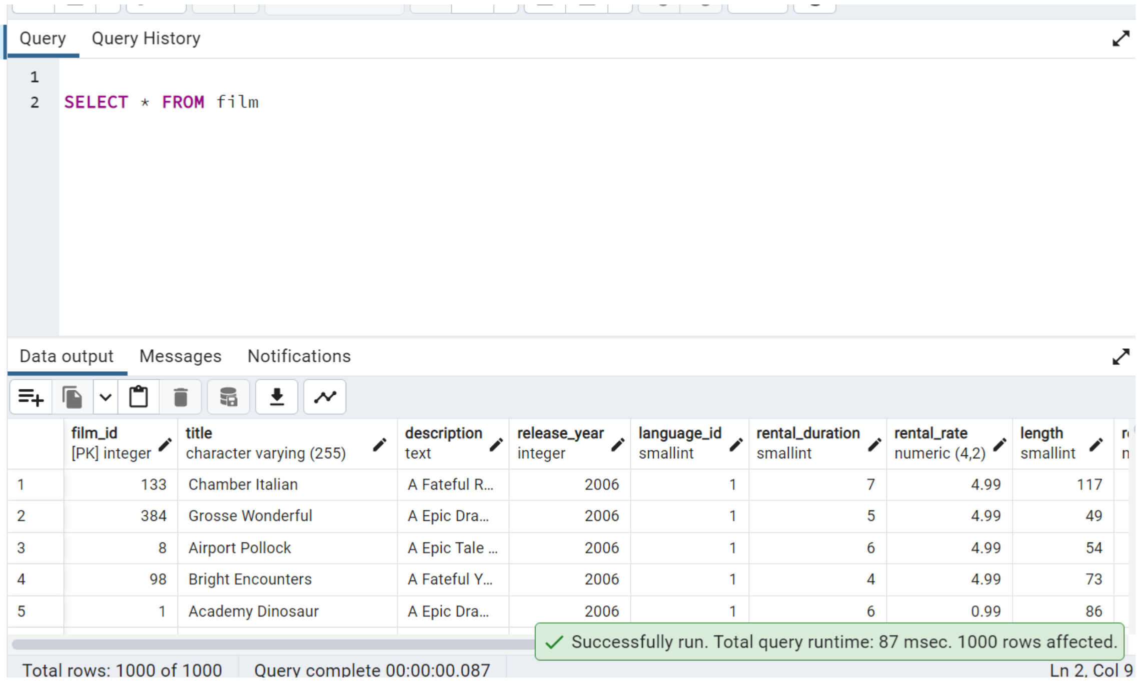
3.4: Database Querying in SQL

1.

Refining Your Query: You need to get some data from the "film" table and decide to use the query SELECT * FROM film.



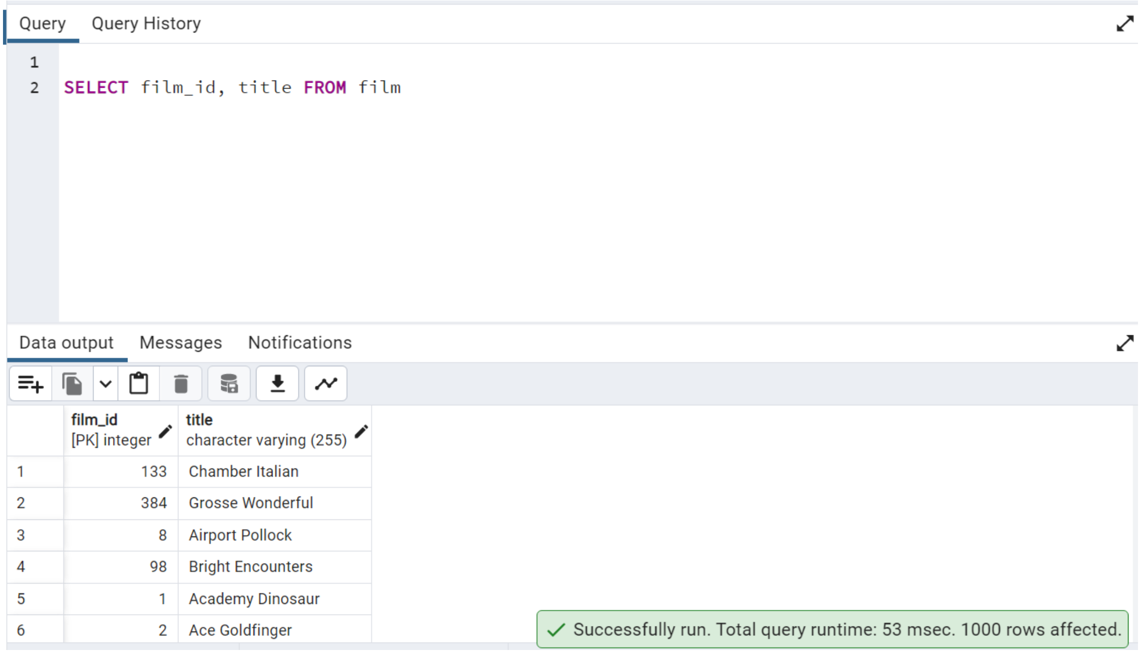
```
SELECT * FROM film
```

	film_id [PK] integer	title character varying (255)	description text	release_year integer	language_id smallint	rental_duration smallint	rental_rate numeric (4,2)	length smallint	r n
1	133	Chamber Italian	A Fateful R...	2006	1	7	4.99	117	
2	384	Grosse Wonderful	A Epic Dra...	2006	1	5	4.99	49	
3	8	Airport Pollock	A Epic Tale ...	2006	1	6	4.99	54	
4	98	Bright Encounters	A Fateful Y...	2006	1	4	4.99	73	
5	1	Academy Dinosaur	A Epic Dra...	2006	1	6	0.99	86	

✓ Successfully run. Total query runtime: 87 msec. 1000 rows affected.

Total rows: 1000 of 1000   Query complete 00:00:00.087   Ln 2, Col 9

- You realize that only the "film_id" and "title" columns are needed. Write a new query that selects only those 2 columns.

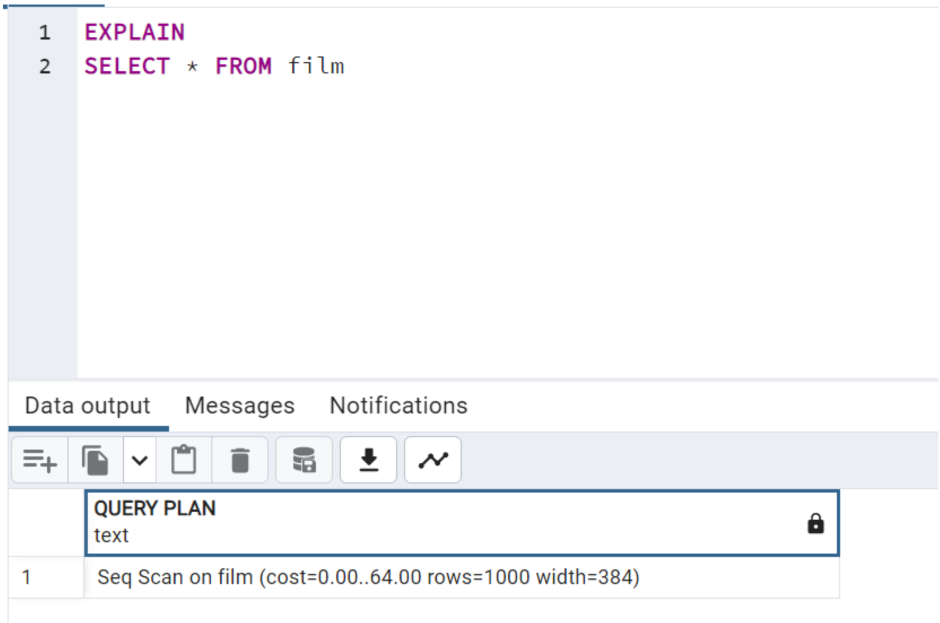


```
SELECT film_id, title FROM film
```

	film_id [PK] integer	title character varying (255)
1	133	Chamber Italian
2	384	Grosse Wonderful
3	8	Airport Pollock
4	98	Bright Encounters
5	1	Academy Dinosaur
6	2	Ace Goldfinger

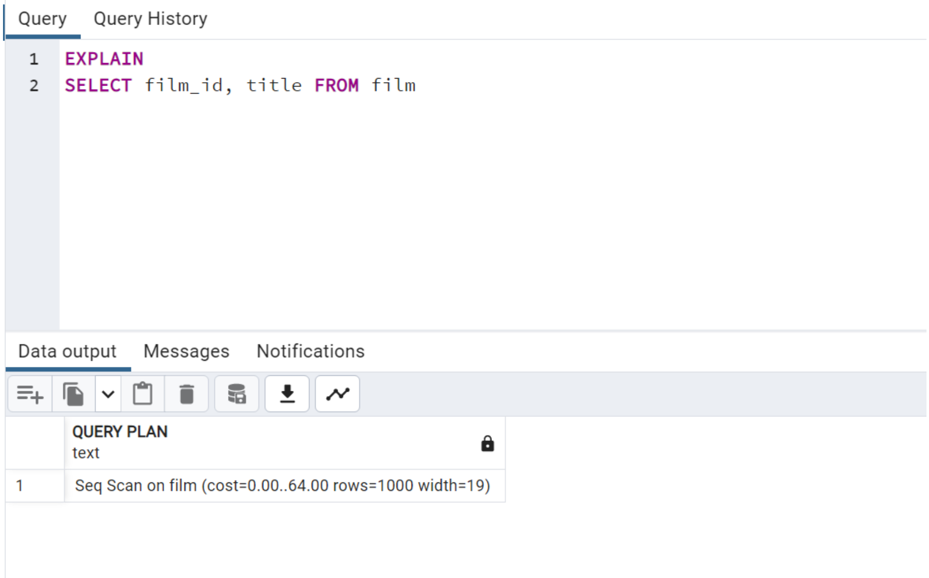
✓ Successfully run. Total query runtime: 53 msec. 1000 rows affected.

- 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?



```
EXPLAIN
SELECT * FROM film
```

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



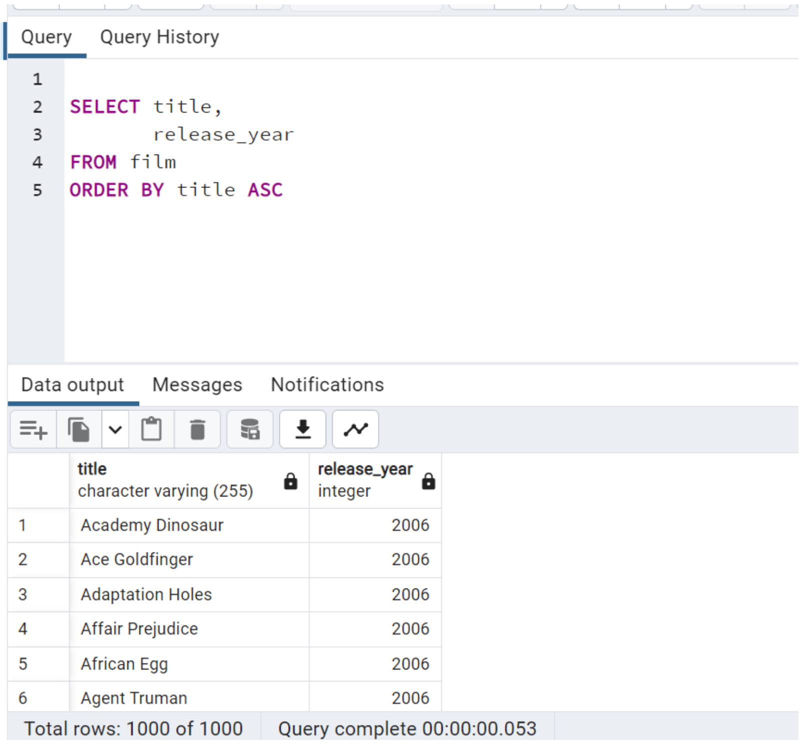
```
EXPLAIN
SELECT film_id, title FROM film
```

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

Of course if we reduce the information we are asking for the query should be optimized and answer faster. We can see as the width has reduced from 384 to 19. The more accurate we build the query, the more optimized will be the answer

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.

- every film from the “film” table, with the movies sorted by title from A to Z

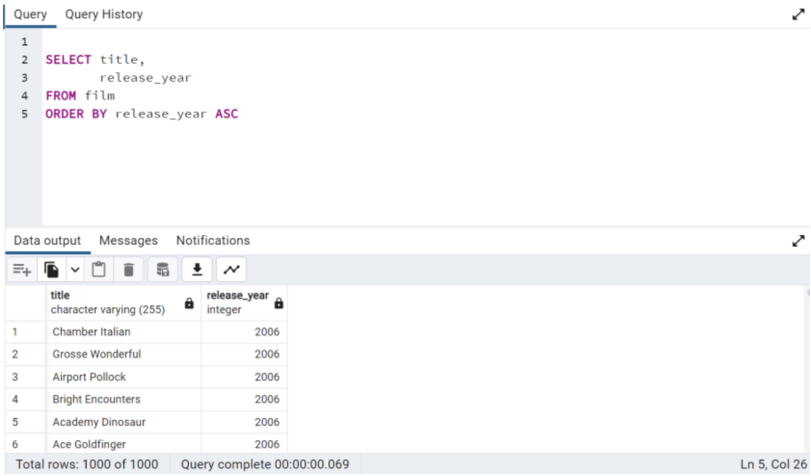


```
SELECT title,
       release_year
FROM film
ORDER BY title ASC
```

	title character varying (255)	release_year integer
1	Academy Dinosaur	2006
2	Ace Goldfinger	2006
3	Adaptation Holes	2006
4	Affair Prejudice	2006
5	African Egg	2006
6	Agent Truman	2006

Total rows: 1000 of 1000 Query complete 00:00:00.053

- every film from the “film” table, with the movies sorted by most recent release year

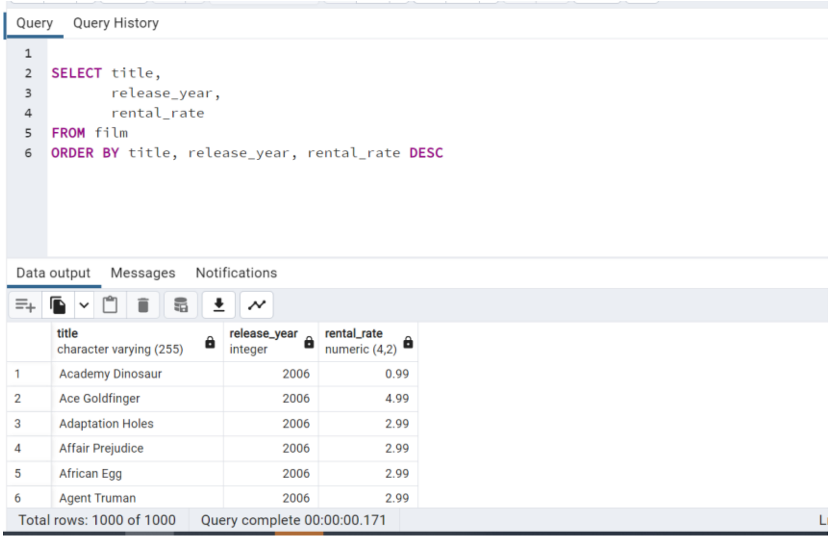


```
SELECT title,
       release_year
FROM film
ORDER BY release_year ASC
```

	title character varying (255)	release_year integer
1	Chamber Italian	2006
2	Grosse Wonderful	2006
3	Airport Pollock	2006
4	Bright Encounters	2006
5	Academy Dinosaur	2006
6	Ace Goldfinger	2006

Total rows: 1000 of 1000 Query complete 00:00:00.069 Ln 5, Col 26

- everything together:

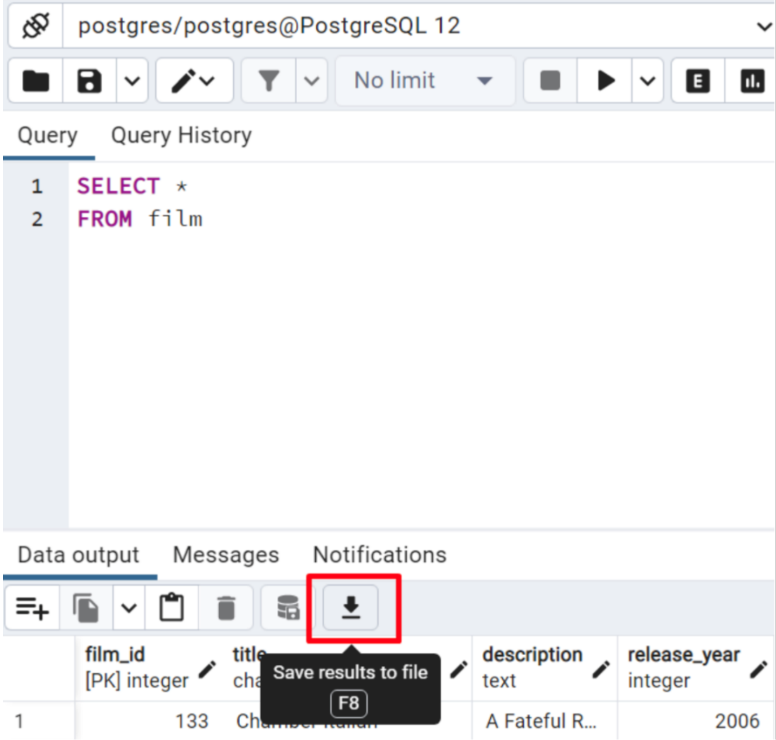


```
SELECT title,
       release_year,
       rental_rate
FROM film
ORDER BY title, release_year, rental_rate DESC
```

	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

Total rows: 1000 of 1000   Query complete 00:00:00.171

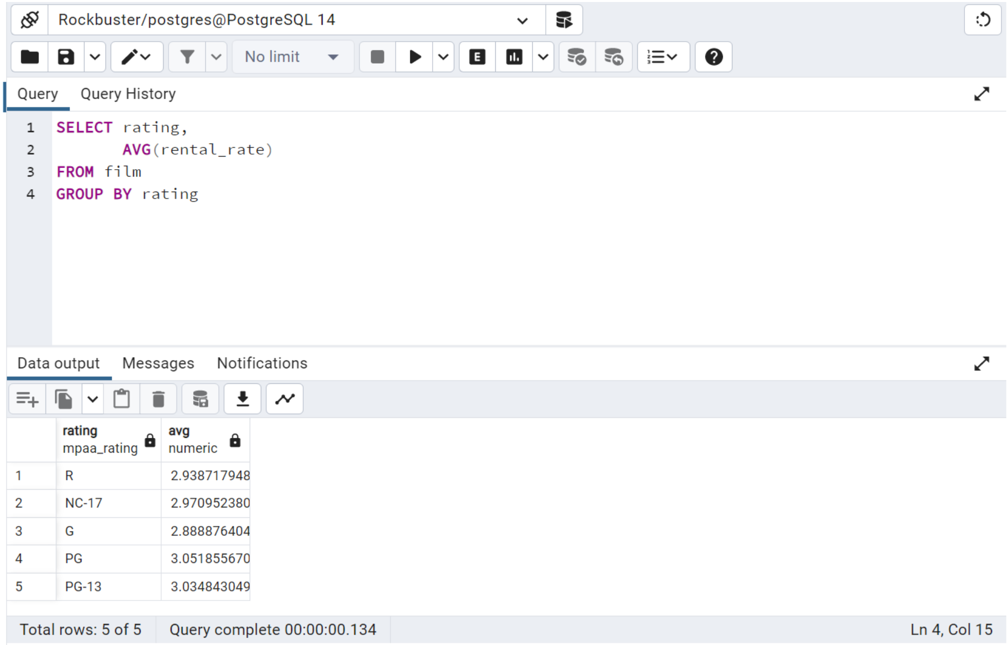
- Extract the data output of your query into a csv file for the film collection department to analyze in Excel. To do this, click the button “Save results to file”:



```
SELECT *
FROM film
```

	film_id [PK] integer	title cha	description text	release_year integer
1	133	Ch	A Fateful R...	2006

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?

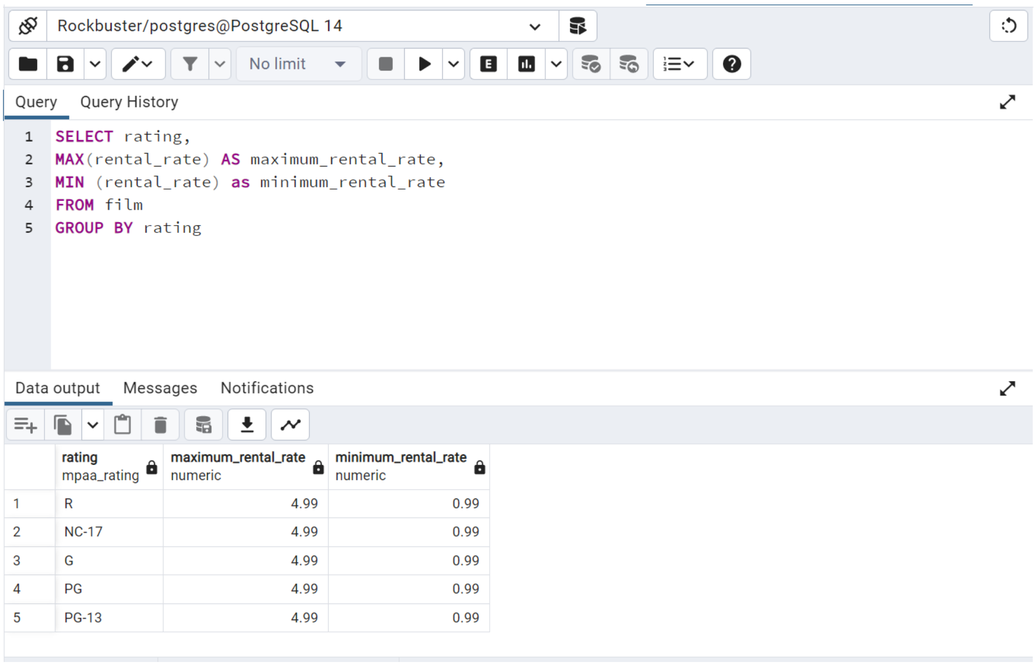


```sql
SELECT rating,
       AVG(rental_rate)
FROM film
GROUP BY rating
```

	rating mpaa_rating	avg numeric
1	R	2.938717948
2	NC-17	2.970952380
3	G	2.888876404
4	PG	3.051855670
5	PG-13	3.034843049

Total rows: 5 of 5 Query complete 00:00:00.134 Ln 4, Col 15

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



```sql
SELECT rating,
MAX(rental_rate) AS maximum_rental_rate,
MIN (rental_rate) as minimum_rental_rate
FROM film
GROUP BY rating
```

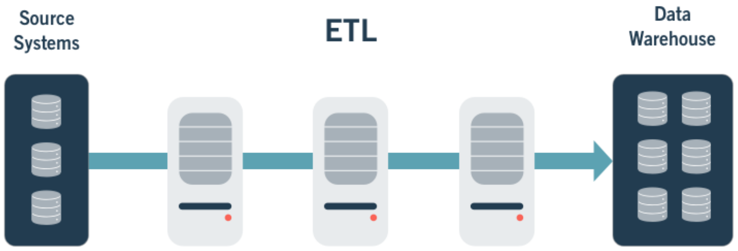
	rating mpaa_rating	maximum_rental_rate numeric	minimum_rental_rate numeric
1	R	4.99	0.99
2	NC-17	4.99	0.99
3	G	4.99	0.99
4	PG	4.99	0.99
5	PG-13	4.99	0.99

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?

     As it´s a new data base to add to the previous information we already have on the warehouse, we should analize and normalize so we can after that use and mix the information to give an answer and correlate to the insights we may extract also from this new new source

This migration should follow the classic procedure of ETL:

- **Extract:** The first step involves collecting the data from multiple data sources.
- **Transform:** During 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:** At this point 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?
  Precisely if we haven´t done the previous analysis and normalization, we can find incoherences on the information so we can´t extract conclusion from it

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.