Movie Rental Data Analysis

Name : Sakshi

Mentor : Jaya Pandey

Project Introduction :

Movie On rent, a leading chain of movie rental stores in a specific country, takes pride in its extensive collection of movies available in both DVD and Blu-ray formats. As part of its commitment to enhancing customer satisfaction, the management is keen on analyzing the rental trends to understand which types of movies are in high demand.

By delving into the preferred genres and the actors who appear in these popular films, Movie On rent aims to curate its inventory to align with the audience's preferences, ensuring a more enjoyable and tailored movie rental experience for its valued customers. Through this data-driven approach, Movie On rent aspires to elevate its business by consistently providing the movies that resonate most with its diverse customer base.

Description of the database

- Here is a description of the dataset's main components:
- Film: This table includes information about individual movies available for rental. It contains columns such as film id, title, description, release year, rental duration, rental rate, replacement cost, and rating. These attributes provide details about each film's unique identifier, title, plot description, release year, rental-related information (duration, rate, and replacement cost), and its rating.
- Actor: This table contains details about actors who have appeared in various movies. The columns include actor id, first name, and last name. Each row represents an actor and provides their unique identifier, first name, and last name.
- Film Actor: This table establishes the relationship between movies and actors, indicating which actors appeared in which films. It has columns such as actor id and film id, representing the unique identifiers of actors and films, respectively.
- Category: This table lists the categories or genres to which movies belong. It includes columns like category id and name, providing the unique identifier and name of each category.
- Film Category: This table links films to their respective categories, indicating the genres to which each movie belongs. It contains columns like film id and category id, representing the unique identifiers of films and categories, respectively.

• Brief on implementation of the capstone:

- The implementation involved writing SQL queries to retrieve and manipulate data from the database. The project consisted of several tasks:

1. Displaying the full names of all actors available in the dataset.
2. Analyzing the frequency of actor names to identify frequently appearing names and unique names.
3. Examining the frequency and uniqueness of last names of the actors.
4. Analyzing movies based on their ratings to determine suitability for different age groups.
5. Exploring the replacement cost of movie copies and sorting them based on cost.
6. Identifying the top three movies with the greatest number of actors.
7. Finding movies starting with the letters 'K' and 'Q' due to their increased popularity.
8. Listing the actors who appeared in the movie "Agent Truman."
9. Identifying movies categorized as family films for targeted promotion.
10. Analyzing rental rates and frequencies of movies to identify popular titles.
11. Determining the film categories where the difference between average replacement cost and rental rate exceeds $15.
12. Displaying film categories with more than 70 movies.

Task : Display Full names of the actors available in the dataset.

- **Technique used :**
- The technique used in the given query is string concatenation.
- CONCAT function is used to combine the first name and last name

- **Interpretation :**
- There are total 201 actors in the dataset.

Full_name
PENELOPE GUINESS
NICK WAHLBERG
ED CHASE
JENNIFER DAVIS
JOHNNY LOLLOBRIGIDA
BETTE NICHOLSON
GRACE MOSTEL
MATTHEW JOHANSSON
JOE SWANK
CHRISTIAN GABLE
ZERO CAGE
KARL BERRY
UMA WOOD
VIVIEN BERGEN
CUBA OLIVIER
FRED COSTNER
HELEN VOIGHT

Task 2 : a. Display the number of times each first name appears in the database.

b. What is the count of actors that have the unique first names in the database? Display the first names of all these actors.

Interpretation :

Techniques used :

GROUP BY is used to group the rows based on the first name column.

COUNT(*) is used to count the number of occurrences of each first name.

The **HAVING** clause is used to filter the groups based on a condition. In this case, the condition is `frequency = 1`

unique_firstname	frequency
JENNIFER	1
BETTE	1
GRACE	1
JOE	1
ZERO	1
KARL	1
UMA	1
FRED	1
HELEN	1
BOB	1
ELVIS	1
ALEC	1
SISSY	1
TIM	1
JUDY	1
VAL	1
GOLDIE	1

frequently_occuring_first_name	frequency
PENELOPE	4
NICK	3
ED	3
JENNIFER	1
JOHNNY	2
BETTE	1
GRACE	1
MATTHEW	3
JOE	1
CHRISTIAN	3
ZERO	1
KARL	1
UMA	1
VIVIEN	2
CUBA	3

Task . The management is interested in analyzing the similarity in the last names of the actors.

a. Display the number of times each last name appears in the database.

b. Display all unique last names in the database.

- Technique used :
- The GROUP BY clause is used to group the rows based on the last name column.
- The COUNT(*) function is then used to count the number of occurrences of each last name
- **HAVING** clause is added to filter the groups based on a condition. In this case, the condition is **`frequency = 1`**,

unique_lastname	frequency
ASTAIRE	1
BACALL	1
BALE	1
BALL	1
BARRYMORE	1
BASINGER	1
BERGEN	1
BERGMAN	1
BIRCH	1
BLOOM	1
BRIDGES	1
BULLOCK	1
CARREY	1
CHAPLIN	1
CLOSE	1
COSTNER	1
CROWE	1

frequent_occuring_lastname	frequency
AKROYD	3
ALLEN	3
ASTAIRE	1
BACALL	1
BAILEY	2
BALE	1
BALL	1
BARRYMORE	1
BASINGER	1
BENING	2
BERGEN	1
BERGMAN	1
BERRY	3
BIRCH	1
BLOOM	1
BOLGER	2
BRIDGES	1

Task . The management wants to analyze the movies based on their ratings to determine if they are suitable for kids or some parental assistance is required.

a. Display the list of records for the movies with the rating 'R'.
(The movies with rating 'R' are not suitable for audience under age 17 years of age)

b. Display the list of records for the movies that are not rated 'R'.
C. Display the list of records for the movies that are suitable for audience below 13 years of age.

R_rated_movies
AIRPORT POLLOCK
ALONE TRIP
AMELIE HELLFIGHTERS
AMERICAN CIRCUS
ANACONDA CONFESSIONS
ANALYZE HOOSIERS
ANYTHING SAVANNAH
APOCALYPSE FLAMINGOS
ARMY FLINTSTONES
BADMAN DAWN
BANGER PINOCCHIO
BEAR GRACELAND
BEAST HUNCHBACK
BEVERLY OUTLAW
BOOGIE AMELIE
BOULEVARD MOB
BROOKLYN DESERT

Not_R_rated_movies
AIRPORT POLLOCK
ALONE TRIP
AMELIE HELLFIGHTERS
AMERICAN CIRCUS
ANACONDA CONFESSIONS
ANALYZE HOOSIERS
ANYTHING SAVANNAH
APOCALYPSE FLAMINGOS
ARMY FLINTSTONES
BADMAN DAWN
BANGER PINOCCHIO
BEAR GRACELAND
BEAST HUNCHBACK
BEVERLY OUTLAW
BOOGIE AMELIE
BOULEVARD MOB
BROOKLYN DESERT

PG13_rated_movies
AIRPLANE SIERRA
ALABAMA DEVIL
ALTER VICTORY
ANTHEM LUKE
APOLLO TEEN
ARACHNOPHOBIA ROLLERCOASTE
ARGONAUTS TOWN
ATTACKS HATE
ATTRACTION NEWTON
BACKLASH UNDEFEATED
BASIC EASY
BEETHOVEN EXORCIST
BERETS AGENT
BILKO ANONYMOUS
BINGO TALENTED
BLADE POLISH
BLINDNESS GUN

- **Technique used : WHERE clause is used to filter the rows from the film table.**
- **Interpretation : Interpretation: Movies that are not rated 'R' include those with ratings such as 'PG', 'PG-13', and 'G'. These ratings indicate that the content is more suitable for a wider range of audiences, including families and younger viewers.**

title	replacement_cost
ARABIA DOGMA	29.99
BALLROOM MOCKINGBIRD	29.99
BLINDNESS GUN	29.99
BONNIE HOLOCAUST	29.99
CHARIOTS CONSPIRACY	29.99
CLOCKWORK PARADISE	29.99
CLYDE THEORY	29.99
CRUELTY UNFORGIVEN	29.99
CUPBOARD SINNERS	29.99
DESPERATE TRAINSPOTTING	29.99
DIRTY ACE	29.99
DOCTOR GRAIL	29.99
EARTH VISION	29.99
EVERYONE CRAFT	29.99
EXTRAORDINARY CONQUER...	29.99
FANTASIA PARK	29.99
FEUD FROGMEN	29.99

movies_with_rc_upto11	replacement_cost
ALIEN CENTER	10.99
AMISTAD MIDSUMMER	10.99
ANACONDA CONFESSIONS	9.99
ARMAGEDDON LOST	10.99
ARTIST COLDBLOODED	10.99
BALLOON HOMEWARD	10.99
BLADE POLISH	10.99
CANDIDATE PERDITION	10.99
CATCH AMISTAD	10.99
CIDER DESIRE	9.99
COMA HEAD	10.99
CONTACT ANONYMOUS	10.99
CONTROL ANTHEM	9.99
COWBOY DOOM	10.99
DAISY MENAGERIE	9.99
DELIVERANCE MULHOLLAND	9.99
DOLLS RAGE	10.99

movies_with_rc_between_11_and_20	replacement_
ACE GOLDFINGER	12.99
ADAPTATION HOLES	18.99
AGENT TRUMAN	17.99
AIRPORT POLLOCK	15.99
ALAMO VIDEOTAPE	16.99
ALONE TRIP	14.99
AMERICAN CIRCUS	17.99
ANALYZE HOOSIERS	19.99
ANGELS LIFE	15.99
ANNIE IDENTITY	15.99
ANONYMOUS HUMAN	12.99
ANTHEM LUKE	16.99
ANTITRUST TOMATOES	11.99
APACHE DIVINE	16.99
APOCALYPSE FLAMINGOS	11.99
APOLLO TEEN	15.99
ARGONAUTS TOWN	12.99

Task . The board members want to understand the replacement cost of a movie copy (disc - DVD/Blue Ray).

Display the list of records for all the movies where the replacement cost is up to $11.

b. Display the list of records for all the movies where the replacement cost is between $11 and $20.

c. Display the list of records for all the movies in descending order of their replacement costs.

- Interpretation : The number of movies with a replacement cost up to $11 is indicating a relatively affordable range.
- The range of replacement cost between $11 and $20 includes movies, suggesting a slightly higher cost category.
- The highest replacement cost among the movies is indicating the most expensive movie in the database.
- **Techniques used** :
- Comparison operator (<=) and BETWEEN in the WHERE clause to filter records.
- ORDER BY clause and the DESC keyword to sort the records in descending order.

Task . Display the names of top 3 movies with the greatest number of actors.

	title	actor_count
▶	LAMBS CINCINATTI	15
	CRAZY HOME	13
	RANDOM GO	13

- **Interpretation :**
- The movies with a higher number of actors involved might indicate popular or well-known films.
- Such movies often attract a large audience due to the presence of notable actors, which can contribute to their success.
- **Techniques used :**
- JOIN to combine data from multiple tables.
- COUNT(*) to calculate the number of actors associated with each movie.
- GROUP BY clause to group the data by movie title.
- ORDER BY clause and DESC to sort the results based on the actor count in descending order.
- LIMIT clause to restrict the output to the top three movies.

Task : films starting with the letters 'K' and 'Q' have also soared in popularity, Display the titles of the movies starting with the letters 'K' and 'Q' .

- Interpretation :
- **The films starting with letters 'K' and 'Q' are popular movies so inventories must be filled with these movies to increase revenues.**
- Techniques used :
- **The '%' wildcard in the query is used to represent any sequence of characters.**
- **The 'LIKE' operator is used to match patterns in the 'title' column.**

title
KANE EXORCIST
KARATE MOON
KENTUCKIAN GIANT
KICK SAVANNAH
KILL BROTHERHOOD
KILLER INNOCENT
KING EVOLUTION
KISS GLORY
KISSING DOLLS
KNOCK WARLOCK
KRAMER CHOCOLATE
KWAI HOMEWARD
QUEEN LUKE
QUEST MUSSOLINI
QUILLS BULL

Task 8. The 'Agent Truman' has been a great success. Display the names of the actors who appeared in this film.

fullname	title
KIRSTEN PALTROW	AGENT TRUMAN
SANDRA KILMER	AGENT TRUMAN
JAYNE NEESON	AGENT TRUMAN
WARREN NOLTE	AGENT TRUMAN
MORGAN WILLIAMS	AGENT TRUMAN
KENNETH HOFFMAN	AGENT TRUMAN
REESE WEST	AGENT TRUMAN

- **Interpretation :**
- Movies get successful because of the actors acting in them. So the inventories must be filled with movies of such actors who have acted in the movie 'Agent Truman'.
- **Technique used :**
- JOIN keyword to combine multiple tables.
- aliases to assign short names to the tables.
- The CONCAT() function is used to concatenate the first name and last name.

Task9 : Sales have been lagging among young families, so the management wants to promote family movies categorized
as family movies. Identify all the movies categorized as family films.

- **Interpretation :**
- By focusing on promoting movies that cater to the interests and preferences of young families, the management aims to improve sales among this category.

- **Techniques used :**
- JOIN operation to combine multiple tables.
- Aliases are used to provide shorter names to tables.

title	name
AFRICAN EGG	Family
APACHE DIVINE	Family
ATLANTIS CAUSE	Family
BAKED CLEOPATRA	Family
BANG KWAI	Family
BEDAZZLED MARRIED	Family
BILKO ANONYMOUS	Family
BLANKET BEVERLY	Family
BLOOD ARGONAUTS	Family
BLUES INSTINCT	Family
BRAVEHEART HUMAN	Family
CHASING FIGHT	Family
CHISUM BEHAVIOR	Family
CHOCOLAT HARRY	Family
CONFUSED CANDLES	Family
CONVERSATION DO...	Family
DATE SPEED	Family

Task . The management wants to observe the rental rates and rental frequencies (number of times the movie disc is rented).

a. Display the maximum, minimum and average rental rates of movies based on their ratings. The output must be sorted in descending order of the average rental rates.

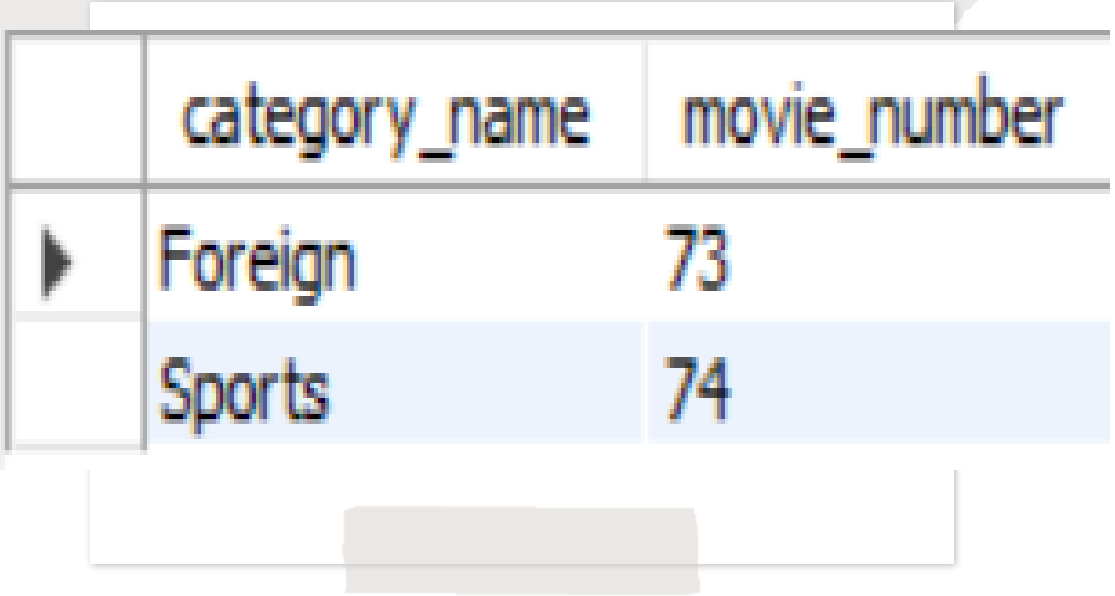
b. Display the movies in descending order of their rental frequencies, so the management can maintain more copies
of these movies.

title	rental_frequency	rating
WAR NOTTING	7	G
WASH HEAVENLY	7	R
WASTELAND DIVINE	7	PG
WESTWARD SEABISCUIT	7	NC-17
WOLVES DESIRE	7	NC-17
WON DARES	7	PG
WORKER TARZAN	7	R
ACADEMY DINOSAUR	6	PG
AFRICAN EGG	6	G
AIRPLANE SIERRA	6	PG-13
AIRPORT POLLOCK	6	R
ALADDIN CALENDAR	6	NC-17
ALAMO VIDEOTAPE	6	G
ALASKA PHANTOM	6	PG
ALICE FANTASIA	6	NC-17
ALLEY EVOLUTION	6	NC-17
ALTER VICTORY	6	PG-13

- **Interpretation :**
- **G rated movies have lowest average rental rates, has widest range of audience and have high rental frequency too so these movies must be stocked up.**
- **Technique used :**
- **The MAX(), MIN(), AVG() functions are used to find the maximum ,minimum and average rental rate.**
- **The GROUP BY clause is used to group the movies by their ratings.**
- **The ORDER BY clause, DESC is used to sort the output in descending order based on the average rental rates.**

rating	max_rental_rate	min_rental_rate	avg_rental_rate
PG	4.99	0.99	3.051856
PG-13	4.99	0.99	3.034843
NC-17	4.99	0.99	2.970952
R	4.99	0.99	2.938718
G	4.99	0.99	2.888876

Task. Display the film categories in which the number of movies is greater than 70.

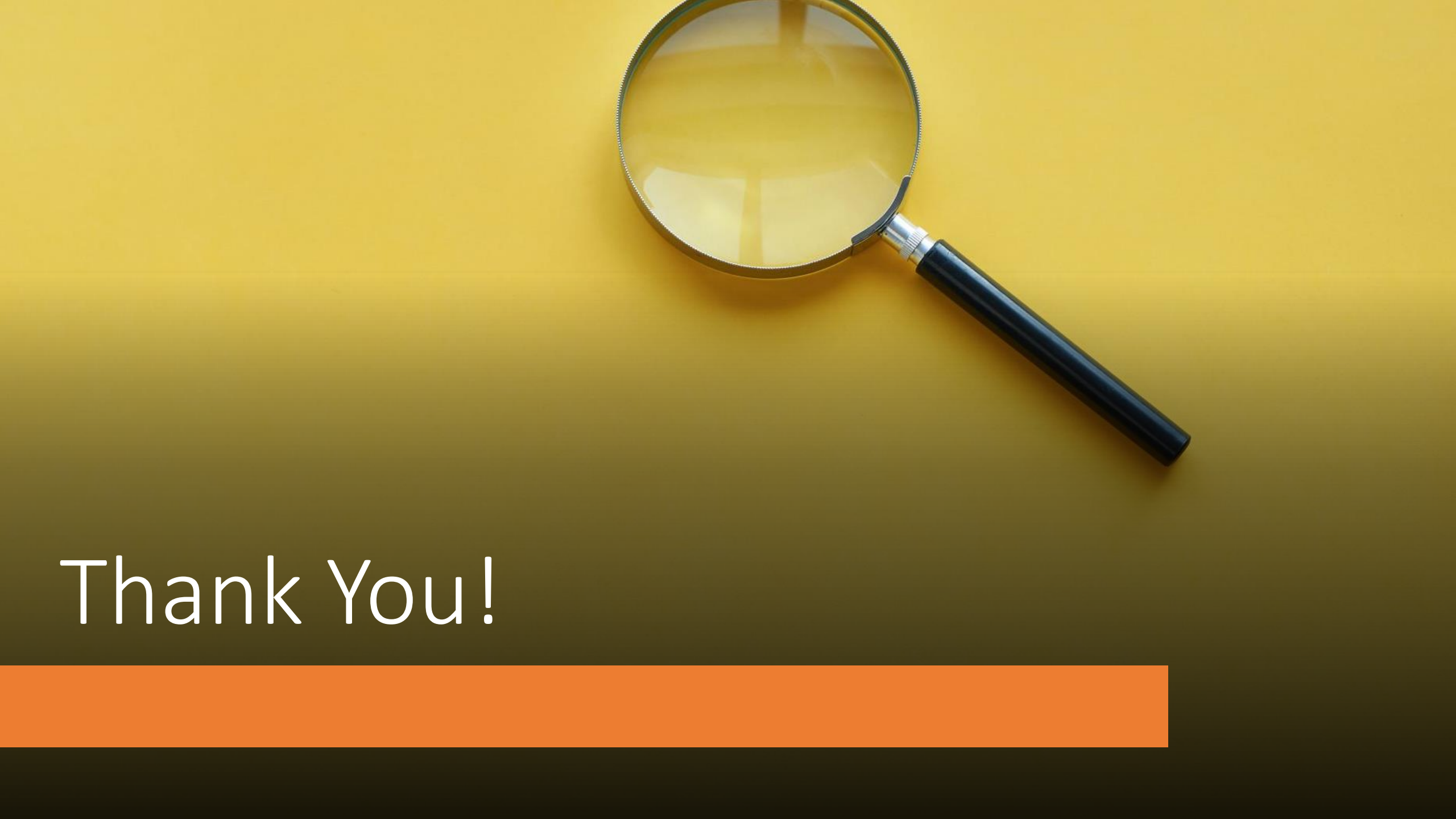


category_name	movie_number
Foreign	73
Sports	74

- **Interpretation :**
- By focusing on categories with a higher number of movies, the chances of attracting a larger audience can be improved.
- **Techniques used :**
- JOIN operations to combine the table.
- GROUP BY clause is used to group the result set by the category name.
- The COUNT() function is used to calculate the number of movies
- The HAVING clause filters the grouped result set based on the condition **`movie number > 70`**

Conclusion :

- In conclusion, The queries cover a range of topics including actor analysis, movie ratings, replacement costs, rental rates, and film categories. They allow the management to understand the frequency of actor names, similarities in last names, and identify popular movies. The analysis of movie ratings helps determine if they are suitable for different age groups. The examination of replacement costs helps understand the expenses associated with movie copies. The rental rate analysis provides insights into the popularity of movies and can guide decisions on maintaining more copies. Additionally, the queries explore film categories and identify those with a significant number of movies. Overall, these queries provide a comprehensive view of the dataset, enabling data-driven decision-making and a better understanding of the film industry.


Thank You!