Sakila Database Analysis Report

[Git and R]

Author: *Atique Ahmad*
Student ID: *MSDS25030*
Date: *11-Nov-2025*
Subject: CS 591: Tools and Techniques for Data Science

Assignment Overview

This project analyzes the **Sakila Database** using **R** and the **data.table** package. The purpose of this analysis is to extract insights from the database, including details about films, customers, rentals, and payments. Additionally, visualization is used to summarize the distribution of film ratings.

All queries were executed using R scripts, and the results were stored as `.csv` files for reproducibility.

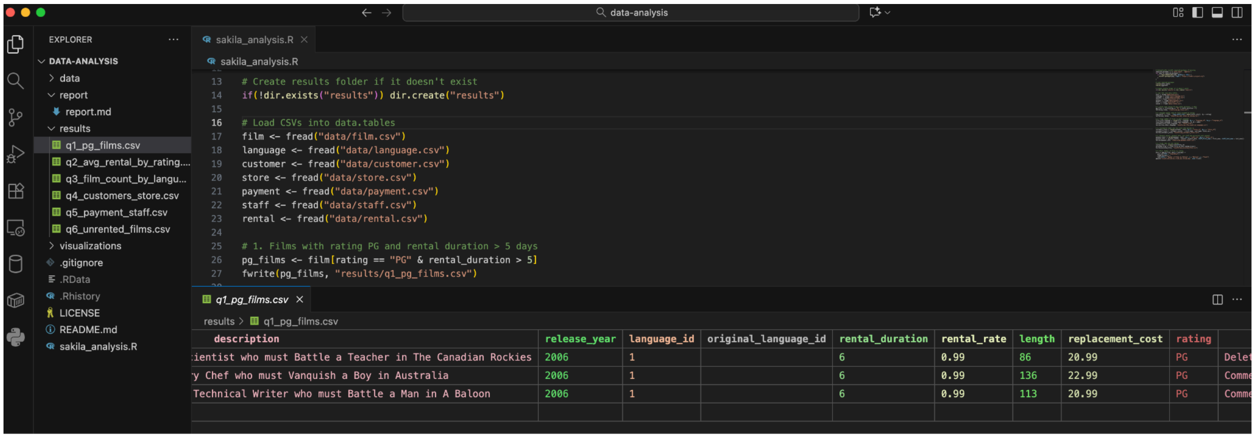
1. Films with Rating PG and Rental Duration > 5 Days

Objective: Identify films rated **PG** that have a rental duration greater than 5 days.

R Code:

```
pg_films <- film[rating == "PG" & rental_duration > 5]
fwrite(pg_films, "results/q1_pg_films.csv")
```

Output File: `results/q1_pg_films.csv`



```r
# Create results folder if it doesn't exist
if(!dir.exists("results")) dir.create("results")

# Load CSVs into data.tables
film <- fread("data/film.csv")
language <- fread("data/language.csv")
customer <- fread("data/customer.csv")
store <- fread("data/store.csv")
payment <- fread("data/payment.csv")
staff <- fread("data/staff.csv")
rental <- fread("data/rental.csv")

# 1. Films with rating PG and rental duration > 5 days
pg_films <- film[rating == "PG" & rental_duration > 5]
fwrite(pg_films, "results/q1_pg_films.csv")
```

description	release_year	language_id	original_language_id	rental_duration	rental_rate	length	replacement_cost	rating	
ientist who must Battle a Teacher in The Canadian Rockies	2006	1		6	0.99	86	20.99	PG	Delet
y Chef who must Vanquish a Boy in Australia	2006	1		6	0.99	136	22.99	PG	Comme
Technical Writer who must Battle a Man in A Baloon	2006	1		6	0.99	113	20.99	PG	Comme

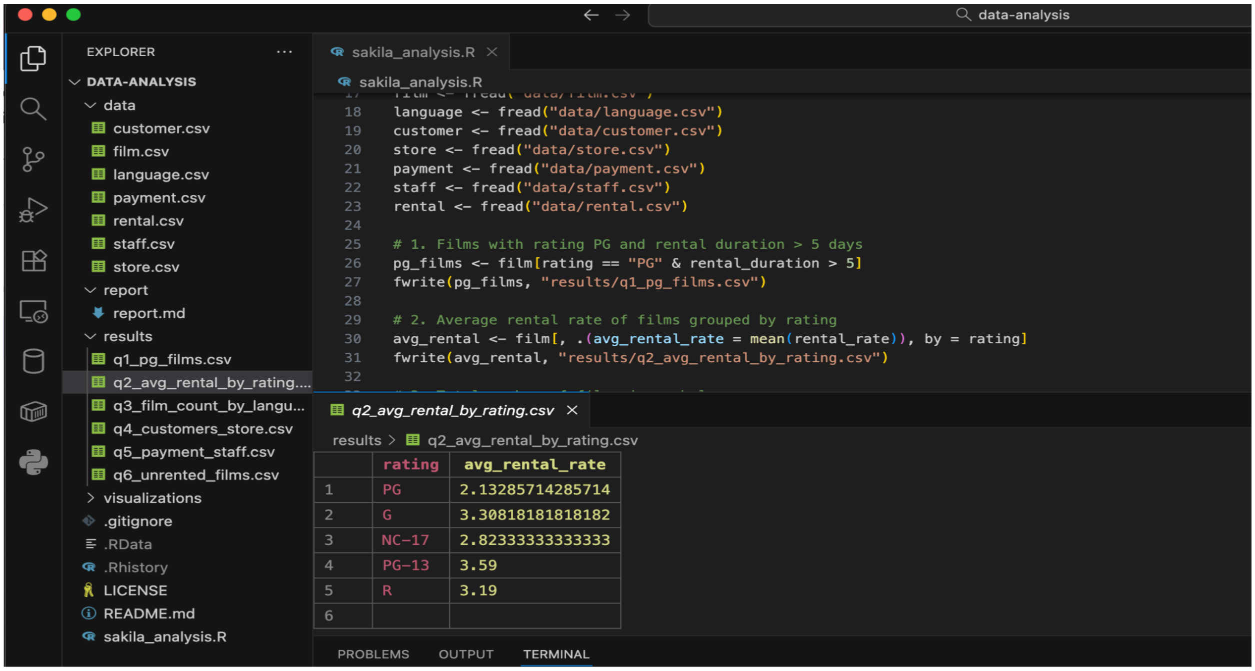
2. Average Rental Rate of Films Grouped by Rating

Objective: Calculate the average rental rate for each film rating category.

R Code:

avg_rental <- film[, .(avg_rental_rate = mean(rental_rate)), by = rating]
fwrite(avg_rental, "results/q2_avg_rental_by_rating.csv")

Output File: `results/q2_avg_rental_by_rating.csv`



```r
language <- fread("data/language.csv")
customer <- fread("data/customer.csv")
store <- fread("data/store.csv")
payment <- fread("data/payment.csv")
staff <- fread("data/staff.csv")
rental <- fread("data/rental.csv")

# 1. Films with rating PG and rental duration > 5 days
pg_films <- film[rating == "PG" & rental_duration > 5]
fwrite(pg_films, "results/q1_pg_films.csv")

# 2. Average rental rate of films grouped by rating
avg_rental <- film[, .(avg_rental_rate = mean(rental_rate)), by = rating]
fwrite(avg_rental, "results/q2_avg_rental_by_rating.csv")
```

	rating	avg_rental_rate
1	PG	2.13285714285714
2	G	3.30818181818182
3	NC-17	2.82333333333333
4	PG-13	3.59
5	R	3.19
6		

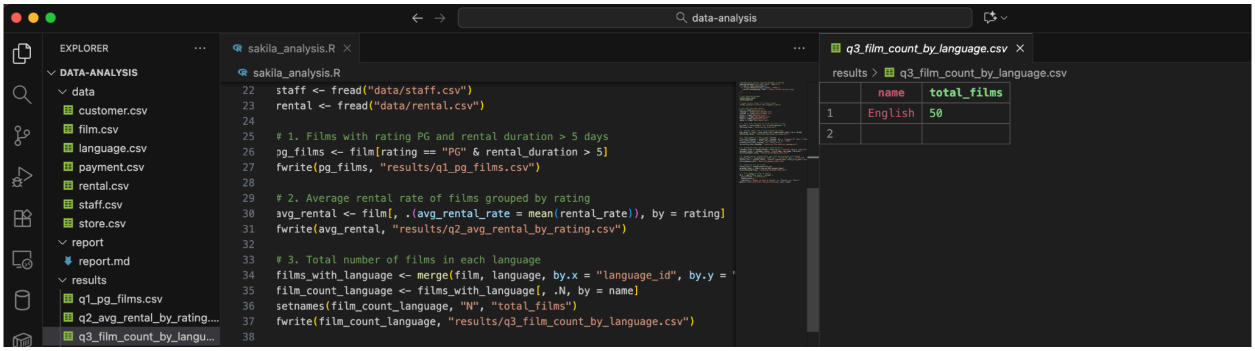
3. Total Number of Films in Each Language

Objective: Count how many films exist for each language.

R Code:

```
films_with_language <- merge(film, language, by = "language_id")
film_count_language <- films_with_language[, .N, by = name]
setnames(film_count_language, "N", "total_films")
fwrite(film_count_language, "results/q3_film_count_by_language.csv")
```

Output File: `results/q3_film_count_by_language.csv`



```
staff <- fread("data/staff.csv")
rental <- fread("data/rental.csv")

# 1. Films with rating PG and rental duration > 5 days
pg_films <- film[rating == "PG" & rental_duration > 5]
fwrite(pg_films, "results/q1_pg_films.csv")

# 2. Average rental rate of films grouped by rating
avg_rental <- film[, .(avg_rental_rate = mean(rental_rate)), by = rating]
fwrite(avg_rental, "results/q2_avg_rental_by_rating.csv")

# 3. Total number of films in each language
films_with_language <- merge(film, language, by.x = "language_id", by.y = '
film_count_language <- films_with_language[, .N, by = name]
setnames(film_count_language, "N", "total_films")
fwrite(film_count_language, "results/q3_film_count_by_language.csv")
```

	name	total_films
1	English	50
2		

---

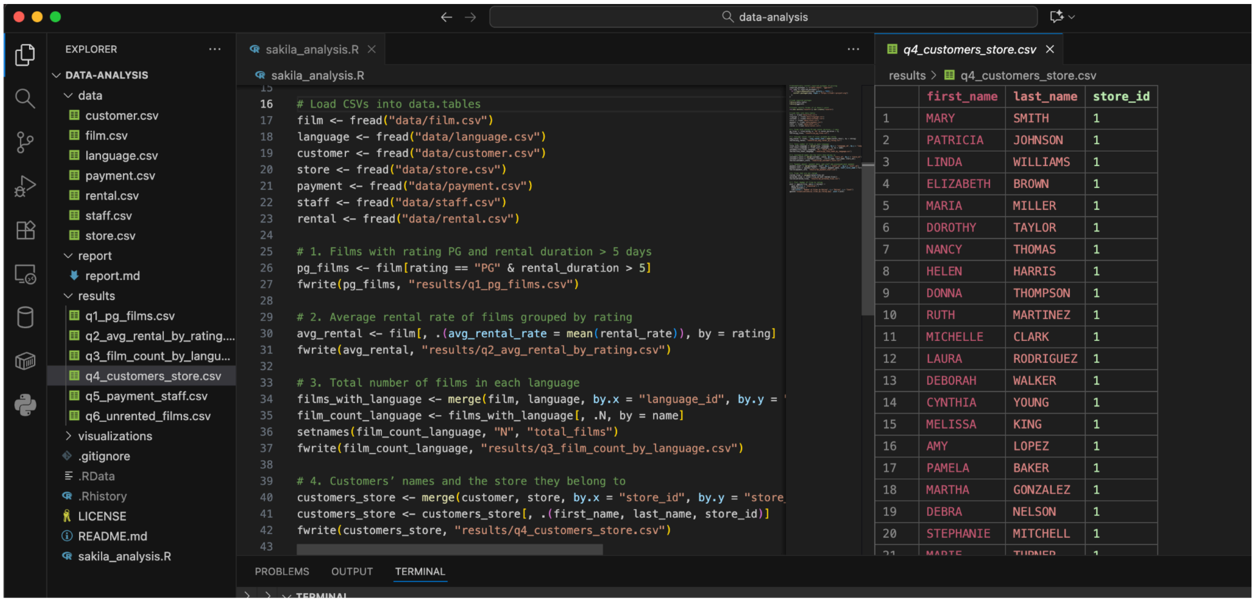
4. Customers’ Names and the Store They Belong To

Objective: Display each customer’s name along with their store ID.

R Code:

```
customers_store <- merge(customer, store, by = "store_id")
customers_store <- customers_store[, .(first_name, last_name, store_id)]
fwrite(customers_store, "results/q4_customers_store.csv")
```

Output File: `results/q4_customers_store.csv`



```r
# Load CSVs into data.tables
film <- fread("data/film.csv")
language <- fread("data/language.csv")
customer <- fread("data/customer.csv")
store <- fread("data/store.csv")
payment <- fread("data/payment.csv")
staff <- fread("data/staff.csv")
rental <- fread("data/rental.csv")

# 1. Films with rating PG and rental duration > 5 days
pg_films <- film[rating == "PG" & rental_duration > 5]
fwrite(pg_films, "results/q1_pg_films.csv")

# 2. Average rental rate of films grouped by rating
avg_rental <- film[, .(avg_rental_rate = mean(rental_rate)), by = rating]
fwrite(avg_rental, "results/q2_avg_rental_by_rating.csv")

# 3. Total number of films in each language
films_with_language <- merge(film, language, by.x = "language_id", by.y = '
film_count_language <- films_with_language[, .N, by = name]
setnames(film_count_language, "N", "total_films")
fwrite(film_count_language, "results/q3_film_count_by_language.csv")

# 4. Customers' names and the store they belong to
customers_store <- merge(customer, store, by.x = "store_id", by.y = "store_
customers_store <- customers_store[, .(first_name, last_name, store_id)]
fwrite(customers_store, "results/q4_customers_store.csv")
```

	first_name	last_name	store_id
1	MARY	SMITH	1
2	PATRICIA	JOHNSON	1
3	LINDA	WILLIAMS	1
4	ELIZABETH	BROWN	1
5	MARIA	MILLER	1
6	DOROTHY	TAYLOR	1
7	NANCY	THOMAS	1
8	HELEN	HARRIS	1
9	DONNA	THOMPSON	1
10	RUTH	MARTINEZ	1
11	MICHELLE	CLARK	1
12	LAURA	RODRIGUEZ	1
13	DEBORAH	WALKER	1
14	CYNTHIA	YOUNG	1
15	MELISSA	KING	1
16	AMY	LOPEZ	1
17	PAMELA	BAKER	1
18	MARTHA	GONZALEZ	1
19	DEBRA	NELSON	1
20	STEPHANIE	MITCHELL	1
21	MARIE	TURNER	1

---

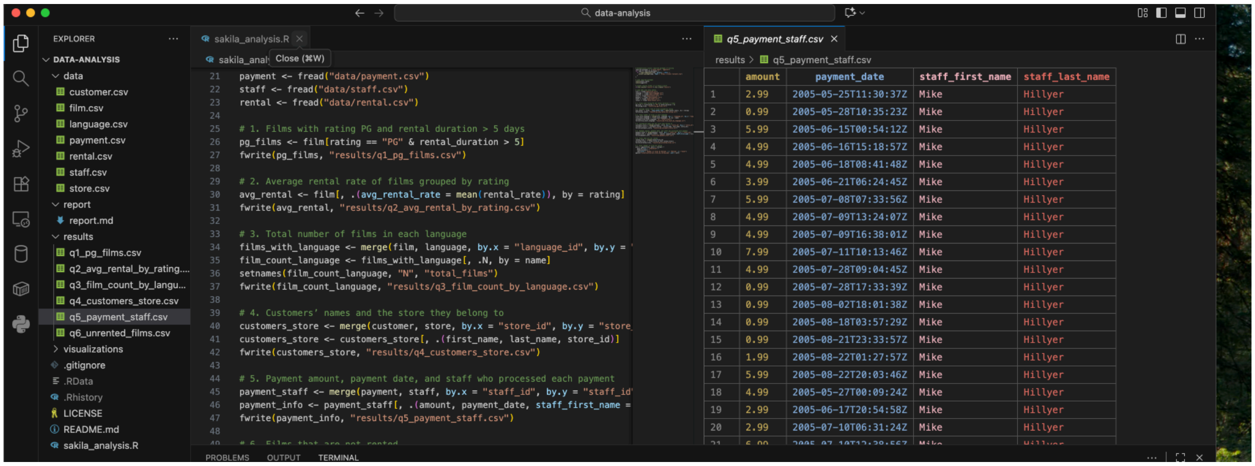
5. Payment Amount, Date, and Staff Who Processed It

Objective: Combine payment data with staff information to show transaction details.

R Code:

payment_staff <- merge(payment, staff, by = "staff_id")
payment_info <- payment_staff[, .(amount, payment_date, staff_first_name = first_name, staff_last_name = last_name)]
fwrite(payment_info, "results/q5_payment_staff.csv")

Output File: `results/q5_payment_staff.csv`



```r
payment <- fread("data/payment.csv")
staff <- fread("data/staff.csv")
rental <- fread("data/rental.csv")

# 1. Films with rating PG and rental duration > 5 days
pg_films <- film[rating == "PG" & rental_duration > 5]
fwrite(pg_films, "results/q1_pg_films.csv")

# 2. Average rental rate of films grouped by rating
avg_rental <- film[, .(avg_rental_rate = mean(rental_rate)), by = rating]
fwrite(avg_rental, "results/q2_avg_rental_by_rating.csv")

# 3. Total number of films in each language
films_with_language <- merge(film, language, by.x = "language_id", by.y = '
film_count_language <- films_with_language[, .N, by = name]
setnames(film_count_language, "N", "total_films")
fwrite(film_count_language, "results/q3_film_count_by_language.csv")

# 4. Customers’ names and the store they belong to
customers_store <- merge(customer, store, by.x = "store_id", by.y = "store_
customers_store <- customers_store[, .(first_name, last_name, store_id)]
fwrite(customers_store, "results/q4_customers_store.csv")

# 5. Payment amount, payment date, and staff who processed each payment
payment_staff <- merge(payment, staff, by.x = "staff_id", by.y = "staff_id'
payment_info <- payment_staff[, .(amount, payment_date, staff_first_name =
fwrite(payment_info, "results/q5_payment_staff.csv")
```

	amount	payment_date	staff_first_name	staff_last_name
1	2.99	2005-05-25T11:30:37Z	Mike	Hillyer
2	0.99	2005-05-28T10:35:23Z	Mike	Hillyer
3	5.99	2005-06-15T00:54:12Z	Mike	Hillyer
4	4.99	2005-06-16T15:18:57Z	Mike	Hillyer
5	4.99	2005-06-18T08:41:48Z	Mike	Hillyer
6	3.99	2005-06-21T06:24:45Z	Mike	Hillyer
7	5.99	2005-07-08T07:33:56Z	Mike	Hillyer
8	4.99	2005-07-09T13:24:07Z	Mike	Hillyer
9	4.99	2005-07-09T16:38:01Z	Mike	Hillyer
10	7.99	2005-07-11T10:13:46Z	Mike	Hillyer
11	4.99	2005-07-28T09:04:45Z	Mike	Hillyer
12	0.99	2005-07-28T17:33:39Z	Mike	Hillyer
13	0.99	2005-08-02T18:01:38Z	Mike	Hillyer
14	0.99	2005-08-18T03:57:29Z	Mike	Hillyer
15	0.99	2005-08-21T23:33:57Z	Mike	Hillyer
16	1.99	2005-08-22T01:27:57Z	Mike	Hillyer
17	5.99	2005-08-22T20:03:46Z	Mike	Hillyer
18	4.99	2005-05-27T00:09:24Z	Mike	Hillyer
19	2.99	2005-06-17T20:54:58Z	Mike	Hillyer
20	2.99	2005-07-10T06:31:24Z	Mike	Hillyer

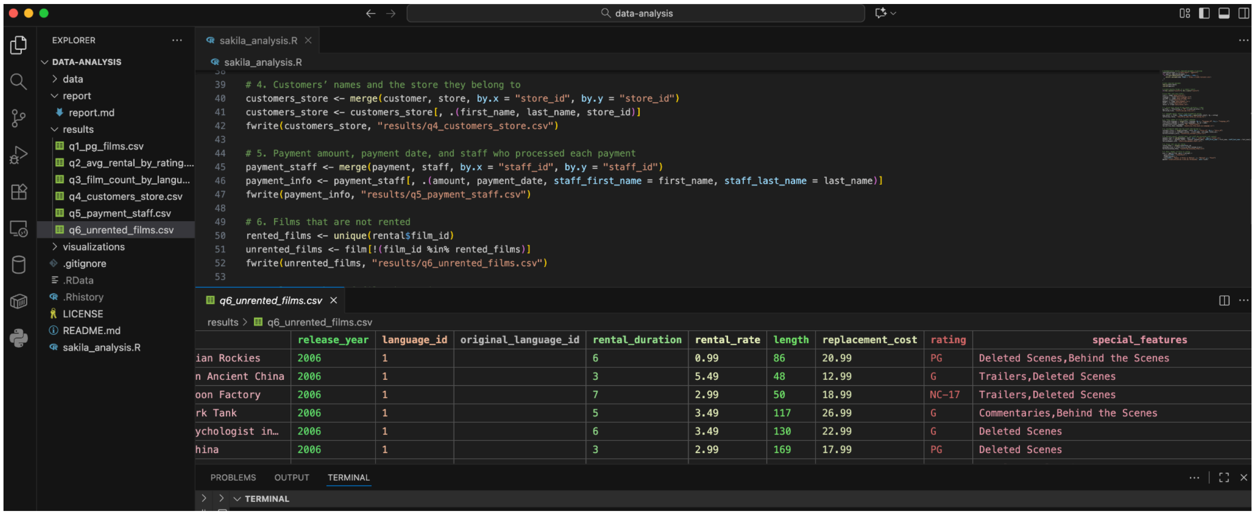
6. Films That Are Not Rented

Objective: Identify films that have never been rented.

R Code:

```r
rented_films <- unique(rental$film_id)
unrented_films <- film[!(film_id %in% rented_films)]
fwrite(unrented_films, "results/q6_unrented_films.csv")
```

Output File: `results/q6_unrented_films.csv`



```r
# 4. Customers’ names and the store they belong to
customers_store <- merge(customer, store, by.x = "store_id", by.y = "store_id")
customers_store <- customers_store[, .(first_name, last_name, store_id)]
fwrite(customers_store, "results/q4_customers_store.csv")

# 5. Payment amount, payment date, and staff who processed each payment
payment_staff <- merge(payment, staff, by.x = "staff_id", by.y = "staff_id")
payment_info <- payment_staff[, .(amount, payment_date, staff_first_name = first_name, staff_last_name = last_name)]
fwrite(payment_info, "results/q5_payment_staff.csv")

# 6. Films that are not rented
rented_films <- unique(rental$film_id)
unrented_films <- film[!(film_id %in% rented_films)]
fwrite(unrented_films, "results/q6_unrented_films.csv")
```

	release_year	language_id	original_language_id	rental_duration	rental_rate	length	replacement_cost	rating	special_features
ian Rockies	2006	1		6	0.99	86	20.99	PG	Deleted Scenes,Behind the Scenes
n Ancient China	2006	1		3	5.49	48	12.99	G	Trailers,Deleted Scenes
oon Factory	2006	1		7	2.99	50	18.99	NC-17	Trailers,Deleted Scenes
rk Tank	2006	1		5	3.49	117	26.99	G	Commentaries,Behind the Scenes
ychologist in…	2006	1		6	3.49	130	22.99	G	Deleted Scenes
hina	2006	1		3	2.99	169	17.99	PG	Deleted Scenes

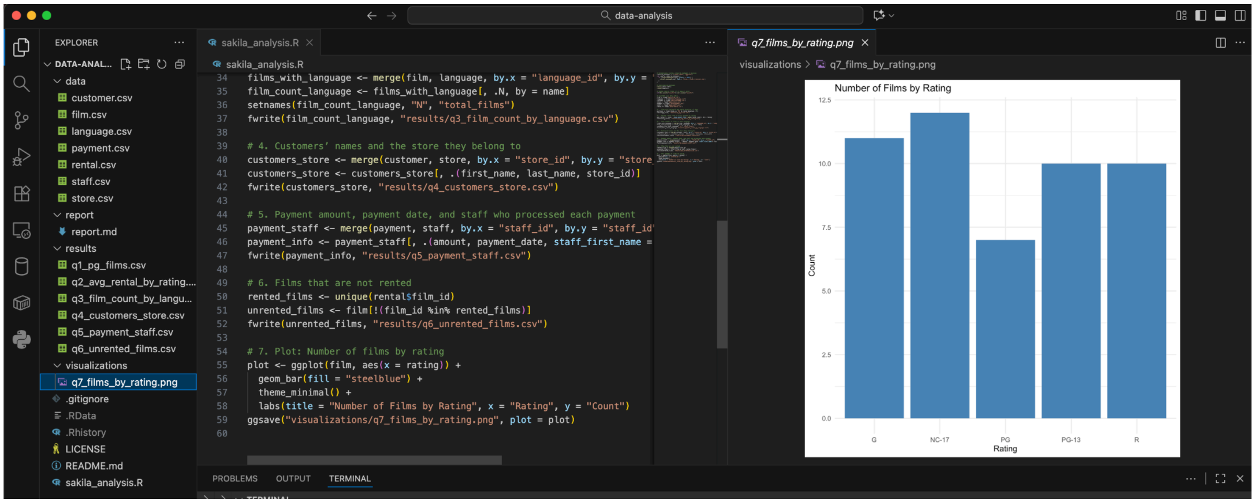
7. Visualization — Number of Films by Rating

Objective: Create a bar chart to visualize the number of films per rating.

R Code:

```
plot <- ggplot(film, aes(x = rating)) +
  geom_bar(fill = "steelblue") +
  theme_minimal() +
  labs(title = "Number of Films by Rating", x = "Rating", y = "Count")
ggsave("visualizations/q7_films_by_rating.png", plot = plot)
```

Generated Plot:
Number of Films by Rating
(Stored in `visualizations/q7_films_by_rating.png`)



```
films_with_language <- merge(film, language, by.x = "language_id", by.y = '
film_count_language <- films_with_language[, .N, by = name]
setnames(film_count_language, "N", "total_films")
fwrite(film_count_language, "results/q3_film_count_by_language.csv")

# 4. Customers’ names and the store they belong to
customers_store <- merge(customer, store, by.x = "store_id", by.y = "store_
customers_store <- customers_store[, .(first_name, last_name, store_id)]
fwrite(customers_store, "results/q4_customers_store.csv")

# 5. Payment amount, payment date, and staff who processed each payment
payment_staff <- merge(payment, staff, by.x = "staff_id", by.y = "staff_id'
payment_info <- payment_staff[, .(amount, payment_date, staff_first_name =
fwrite(payment_info, "results/q5_payment_staff.csv")

# 6. Films that are not rented
rented_films <- unique(rental$film_id)
unrented_films <- film[!(film_id %in% rented_films)]
fwrite(unrented_films, "results/q6_unrented_films.csv")

# 7. Plot: Number of films by rating
plot <- ggplot(film, aes(x = rating)) +
  geom_bar(fill = "steelblue") +
  theme_minimal() +
  labs(title = "Number of Films by Rating", x = "Rating", y = "Count")
ggsave("visualizations/q7_films_by_rating.png", plot = plot)
```

8. Use of Git in the entire assignment.

Git was used throughout the project to manage version control, ensure collaboration, and maintain a clean development workflow. The project repository is publicly available on GitHub.

Repository Link: https://github.com/atique-ahmad-01/data-analysis/

Documentation:
A detailed project report is available in Markdown format, maintained through Git commits.

Report Link: https://github.com/atique-ahmad-01/data-analysis/blob/main/report/report.md

Conclusion

The Sakila database analysis provided insights into:

- Films with **PG** rating and longer rental durations.
- **Average rental rates** grouped by film rating.
- Film distribution across **languages**.
- Association between **customers and stores**.
- Tracking of **payments and responsible staff**.
- Identification of **unrented films**.
- A visualization summarizing the **film ratings distribution**.

All outputs are saved in the `results/` and `visualizations/` folders for future reference.