

# CS-591 Tools and techniques for Data Science

## Assignment # 03 : GIT & R

Name: Muhammad kawish sarfraz

Roll # MSDS25038

---

Use tables of ‘Sakila’ schema which were used during SQL lectures to complete following tasks. Use the Git and data.table package of R instead.

**1. Write a query to display all films that have a rating of PG and a rental duration greater than 5 days. [10 Marks]**

**Code:**

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

**Screenshot of running code:**

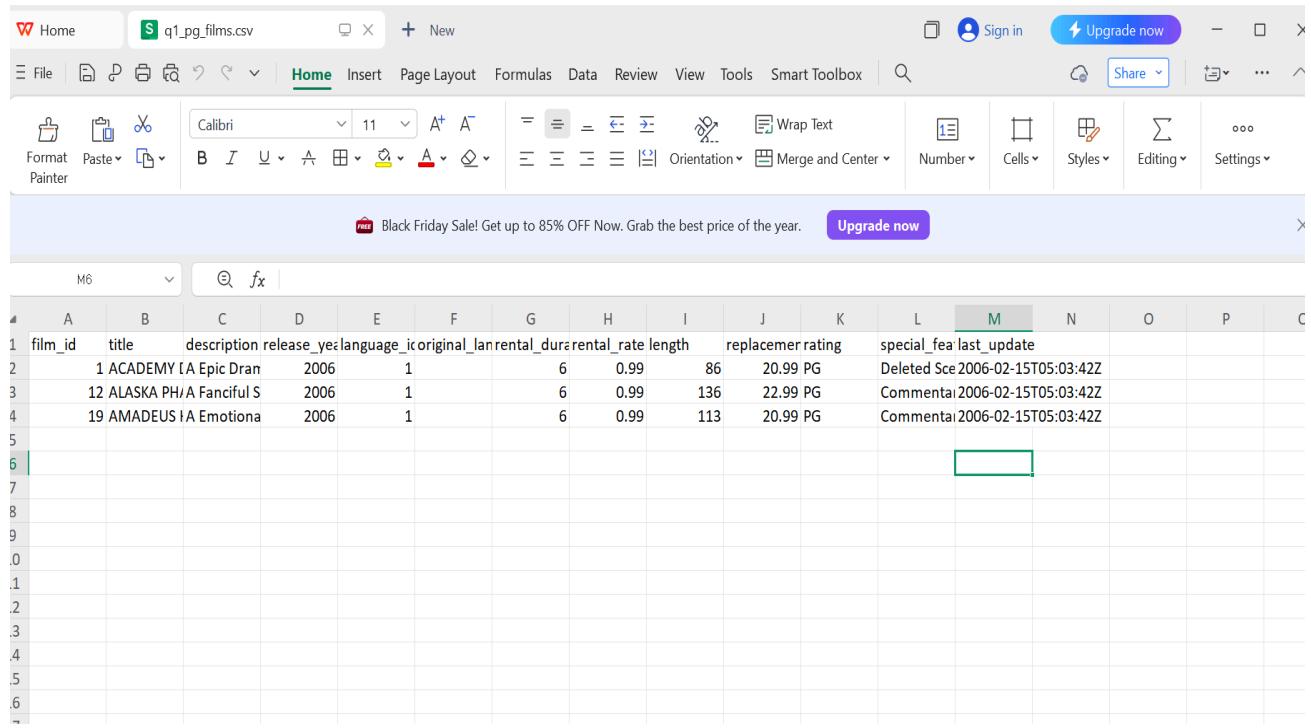
The screenshot shows the RStudio interface with the following details:

- File Menu:** File Edit Code View Plots Session Build Debug Profile Tools Help
- Toolbar:** Source on Save Run Addins
- Code Editor:** new.R\* analysis.R\*  
The code in new.R\* is:

```
26
27 # 1) Select films rated 'PG' with rental duration longer than 5 days
28 pg_long_rental_films <- films_dt[rating == "PG" & rental_duration > 5]
29 fwrite(pg_long_rental_films, "results/q1_pg_films.csv")
30 |
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
```
- Console:** R 4.5.2 · C:/Users/User/Desktop/new assignment 3/  
The console output is:

```
> # Load input CSV files (expect them in ./data)
> films_dt <- fread("data/film.csv")
> languages_dt <- fread("data/language.csv")
> customers_dt <- fread("data/customer.csv")
> stores_dt <- fread("data/store.csv")
> payments_dt <- fread("data/payment.csv")
> staff_dt <- fread("data/staff.csv")
> rentals_dt <- fread("data/rental.csv")
>
> # 1) Select films rated 'PG' with rental duration longer than 5 days
> pg_long_rental_films <- films_dt[rating == "PG" & rental_duration > 5]
> fwrite(pg_long_rental_films, "results/q1_pg_films.csv")
>
```
- Environment:** Global Environment  
Shows the following objects:
  - customers\_dt: 50 obs. of 9 variables
  - films\_dt: 50 obs. of 13 variables
  - languages\_dt: 6 obs. of 3 variables
  - payments\_dt: 50 obs. of 7 variables
  - pg\_long\_rental\_f...: 3 obs. of 13 variables
  - rentals\_dt: 50 obs. of 7 variables
  - staff\_dt: 2 obs. of 11 variables
  - stores\_dt: 2 obs. of 4 variables
- Files:** Shows the directory structure:
  - ..
  - data (218 B, Nov 16, 2025, 9:24 PM)
  - data.Rproj (0 B, Nov 16, 2025, 9:25 PM)
  - new.R
  - results

## Result (opened from result folder):



The screenshot shows a Microsoft Excel spreadsheet titled "q1\_pg\_films.csv". The ribbon is visible at the top with tabs like Home, Insert, Page Layout, etc. A promotional banner for a Black Friday sale is displayed across the top. The main content is a table with the following data:

#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	C
1	film_id	title	description	release_year	language_id	original_language_id	rental_duration	rental_rate	length	replacement_cost	rating	special_features	last_update				
2	1	ACADEMY	[A Epic Dram]	2006	1		6	0.99	86	20.99	PG	Deleted	Sce 2006-02-15T05:03:42Z				
3	12	ALASKA PHI	[A Fanciful S	2006	1		6	0.99	136	22.99	PG	Commentar	ta 2006-02-15T05:03:42Z				
4	19	AMADEUS	[A Emotiona	2006	1		6	0.99	113	20.99	PG	Commentar	ta 2006-02-15T05:03:42Z				
5																	
6																	
7																	
.0																	
.1																	
.2																	
.3																	
.4																	
.5																	
.6																	

## Explanation:

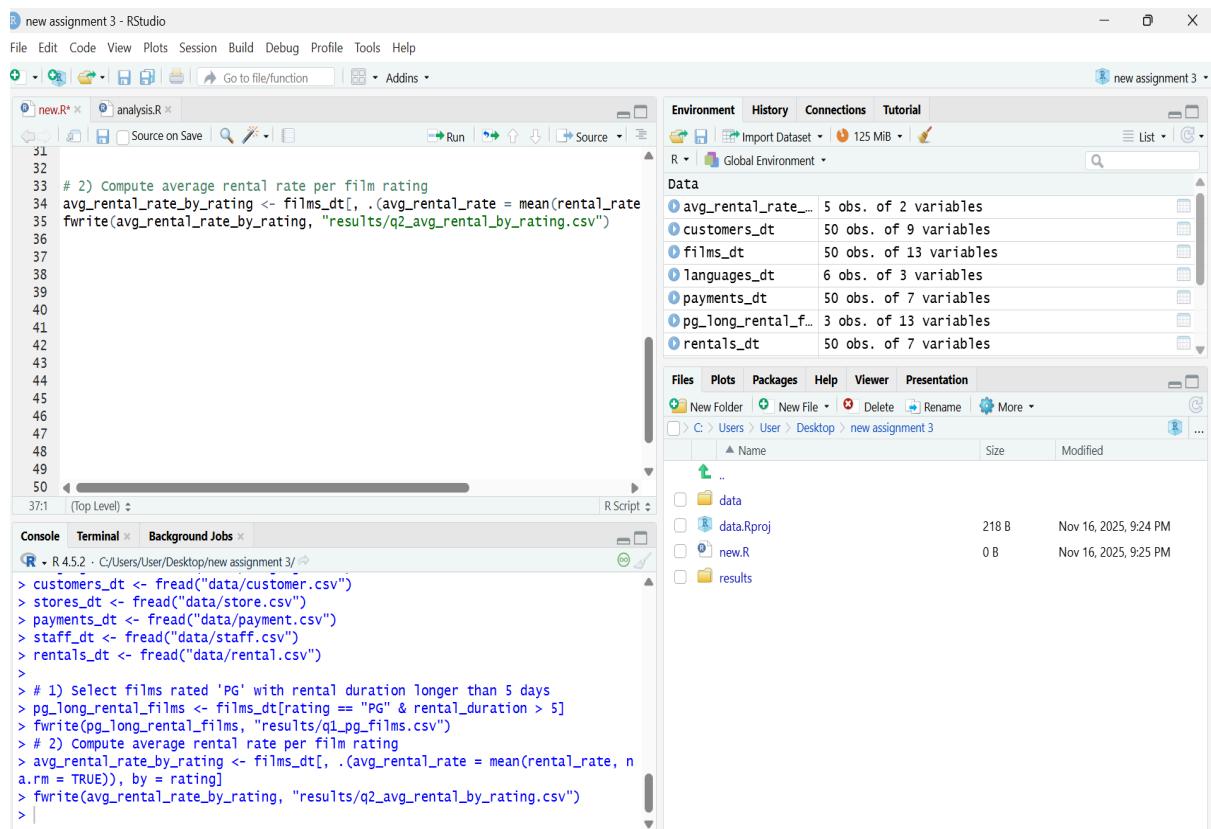
This query filters the film table to show only those movies that are rated PG and rented for more than 5 days. It basically applies two conditions together using AND. The result helps us identify longer-duration PG movies.

**2. Write a query to display the average rental rate of films, grouped by their rating. [10 Marks]**

**Code:**

```
avg_rental_rate_by_rating <- films_dt[, .(avg_rental_rate = mean(rental_rate, na.rm = TRUE)), by = rating]
fwrite(avg_rental_rate_by_rating,
"results/q2_avg_rental_by_rating.csv")
```

**Screenshot of running code:**



## Result (opened from result folder):

The screenshot shows a Microsoft Excel spreadsheet titled "q2\_avg\_rental\_by\_rating.csv". The table contains the following data:

	A	B	C	D	E	F	G	H	I	J
1	rating	avg_rental_rate								
2	PG	2.13285714285714								
3	G	3.30818181818182								
4	NC-17	2.82333333333333								
5	PG-13	3.59								
6	R	3.19								
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										

## Explanation:

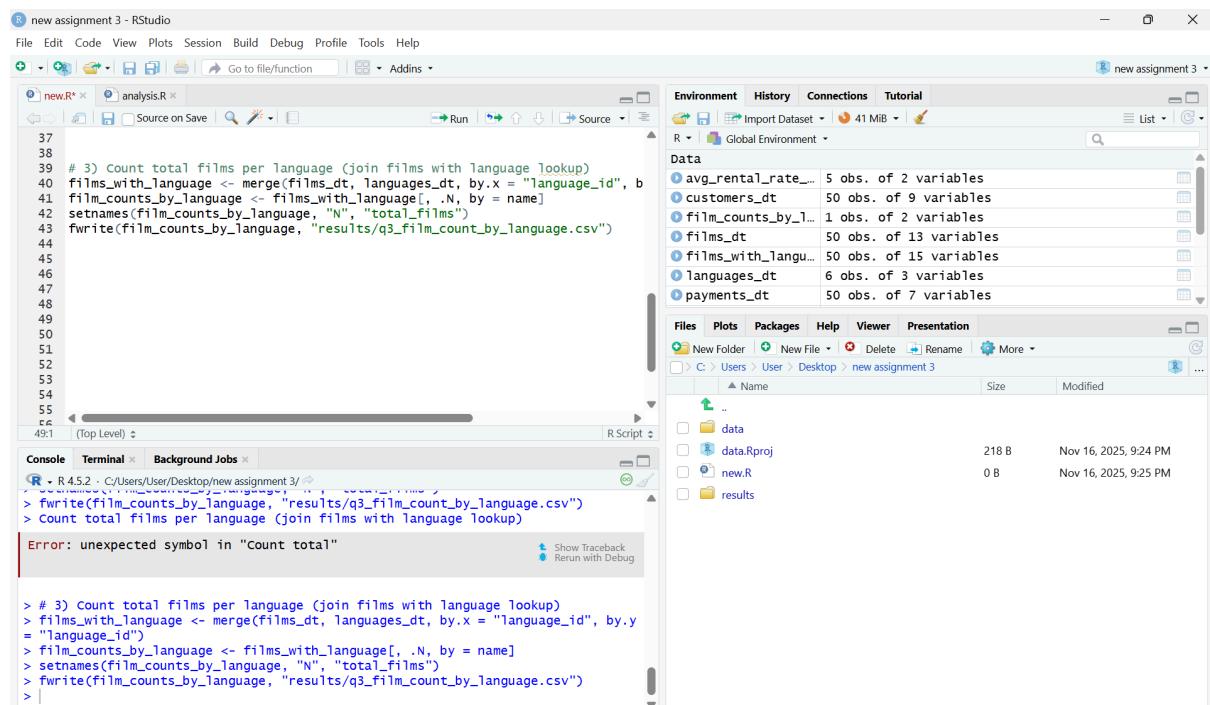
Here I grouped all films by their rating and calculated the average rental rate for each group. This helps us compare which ratings have higher or lower rental prices. It's similar to doing a summary for each category.

### 3. Write a query to count the total number of films in each language. [10 Marks]

**Code:**

```
films_with_language <- merge(films_dt, languages_dt, by.x = "language_id", by.y = "language_id")
film_counts_by_language <- films_with_language[, .N, by = name]
setnames(film_counts_by_language, "N", "total_films")
fwrite(film_counts_by_language,
"results/q3_film_count_by_language.csv")
```

**Screenshot of running code:**



## **Result (opened from result folder):**

A screenshot of Microsoft Excel showing a table with two rows:

name	total_films
English	50

## Explanation:

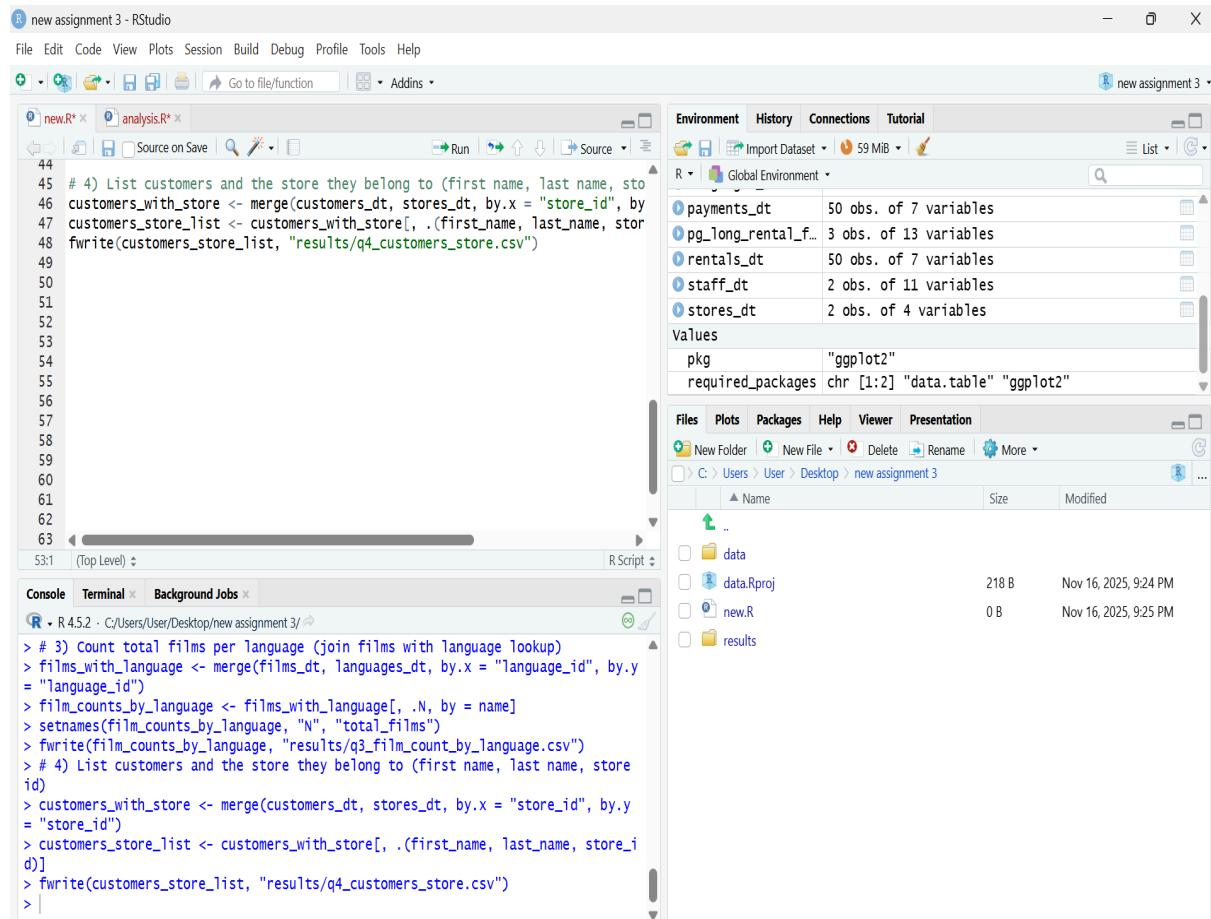
This query joins the films with the language table and then counts how many films exist for each language. It basically tells us which languages have more movies available. Grouping makes it easy to summarize by language.

#### 4. List the customers' names and the store they belong to. [10 Marks]

**Code:**

```
customers_with_store <- merge(customers_dt, stores_dt,
by.x = "store_id", by.y = "store_id")
customers_store_list <- customers_with_store[, 
.(first_name, last_name, store_id)]
fwrite(customers_store_list,
"results/q4_customers_store.csv")
```

**Screenshot of running code:**



## Result (opened from result folder):

The screenshot shows a Microsoft Excel spreadsheet titled "q4\_customers\_store". The table has three columns: "first\_name", "last\_name", and "store\_id". The data consists of 25 rows of customer names and their corresponding store ID, all of which are 1. The Excel ribbon is visible at the top, showing tabs for Home, Insert, Page Layout, Formulas, Data, Review, View, Tools, and Smart Toolbox. The "Home" tab is selected. The font is Calibri, size 11, and the alignment is left.

	A	B	C	D	E	F	G	H	I	J	K
1	first_name	last_name	store_id								
2	MARY	SMITH	1								
3	PATRICIA	JOHNSON	1								
4	LINDA	WILLIAMS	1								
5	ELIZABETH	BROWN	1								
6	MARIA	MILLER	1								
7	DOROTHY	TAYLOR	1								
8	NANCY	THOMAS	1								
9	HELEN	HARRIS	1								
10	DONNA	THOMPSON	1								
11	RUTH	MARTINEZ	1								
12	MICHELLE	CLARK	1								
13	LAURA	RODRIGUEZ	1								
14	DEBORAH	WALKER	1								
15	CYNTHIA	YOUNG	1								
16	MELISSA	KING	1								
17	AMY	LOPEZ	1								
18	PAMELA	BAKER	1								
19	MARTHA	GONZALEZ	1								
20	DEBRA	NELSON	1								
21	STEPHANIE	MITCHELL	1								
22	MARIE	TURNER	1								
23	JANET	PHILLIPS	1								
24	FRANCES	PARKER	1								
25	ANN	EVANS	1								

## Explanation:

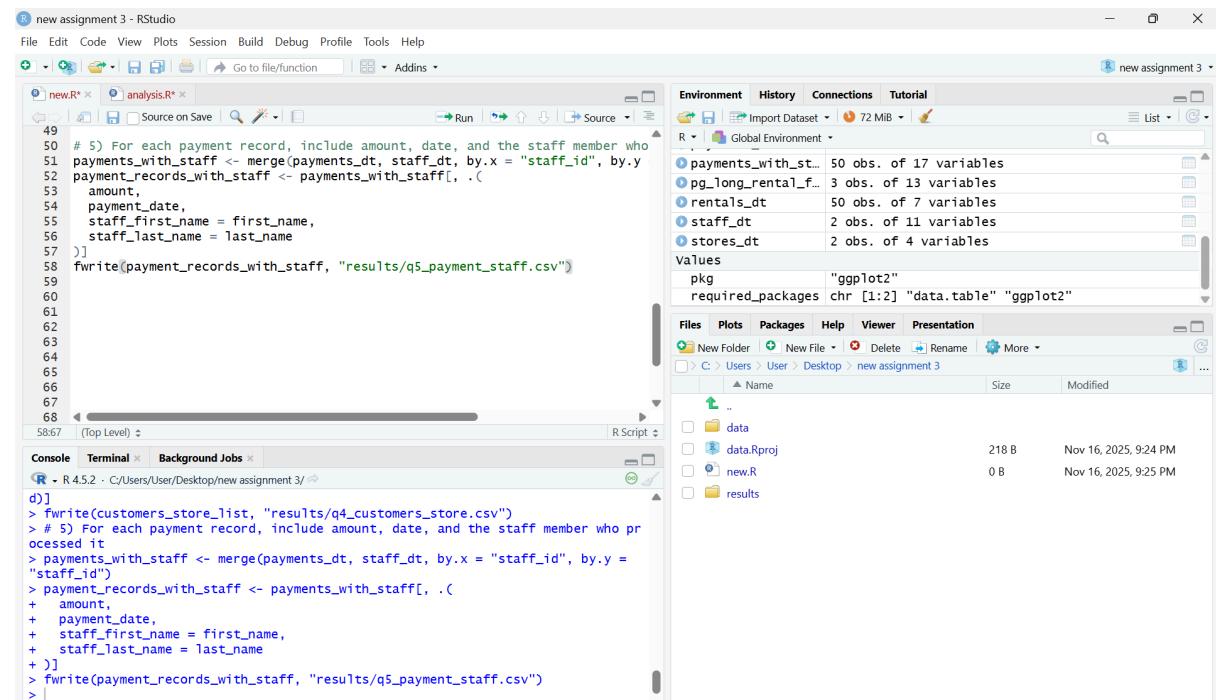
This merges customer information with store details using their store\_id. The result shows each customer along with the store they are registered in. It helps understand how customers are distributed across stores.

## 5. Display the payment amount, payment date, and the staff member who processed each payment. [10 Marks]

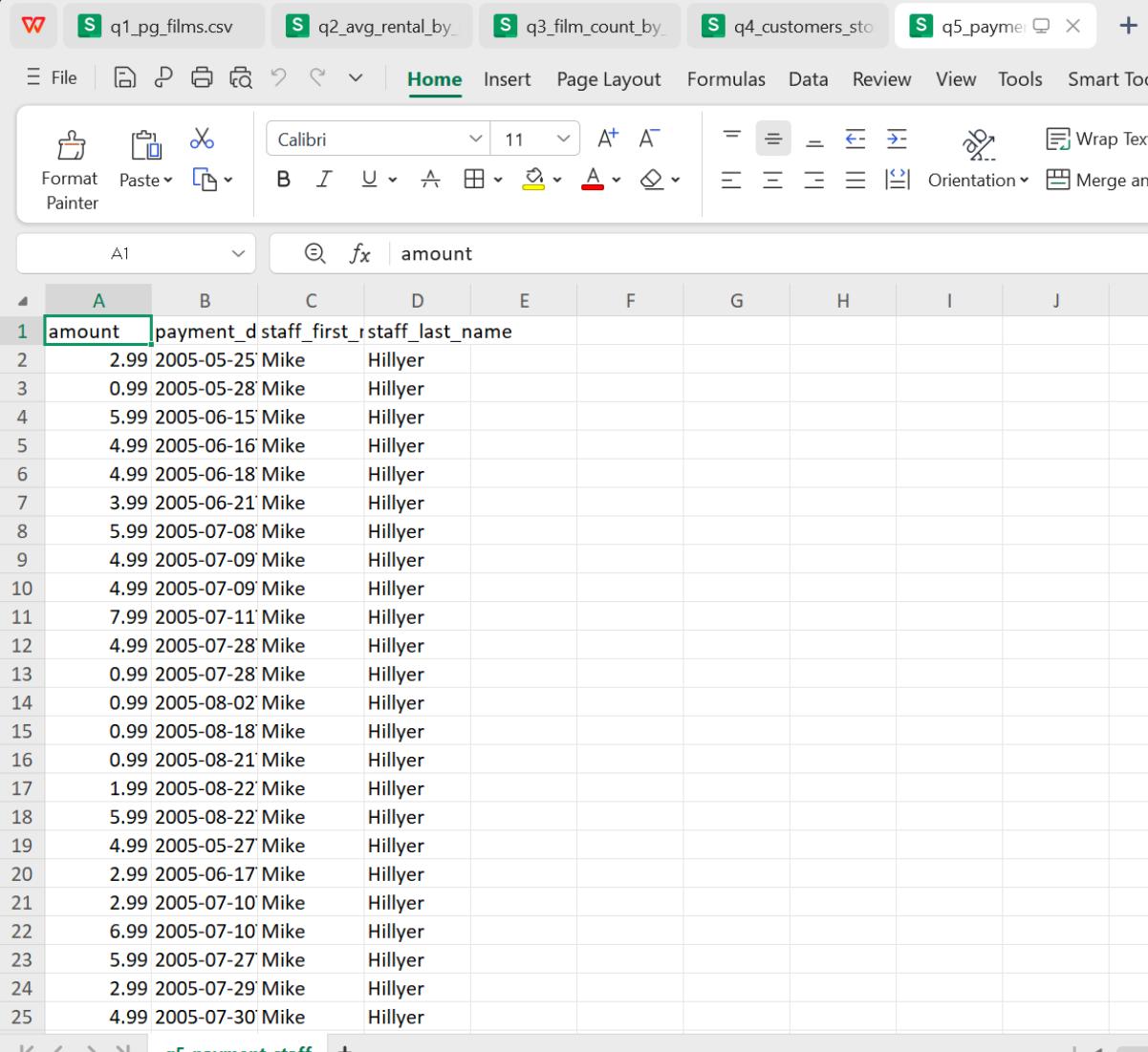
**Code:**

```
payments_with_staff <- merge(payments_dt, staff_dt, by.x =  
  "staff_id", by.y = "staff_id")  
payment_records_with_staff <- payments_with_staff[, .(  
  amount,  
  payment_date,  
  staff_first_name = first_name,  
  staff_last_name = last_name  
)]  
fwrite(payment_records_with_staff,  
 "results/q5_payment_staff.csv")
```

**Screenshot of running code:**



## Result (opened from result folder):



	A	B	C	D	E	F	G	H	I	J
1	amount	payment_date	staff_id	staff_first_name	staff_last_name					
2	2.99	2005-05-25	Mike	Hillyer						
3	0.99	2005-05-28	Mike	Hillyer						
4	5.99	2005-06-15	Mike	Hillyer						
5	4.99	2005-06-16	Mike	Hillyer						
6	4.99	2005-06-18	Mike	Hillyer						
7	3.99	2005-06-21	Mike	Hillyer						
8	5.99	2005-07-08	Mike	Hillyer						
9	4.99	2005-07-09	Mike	Hillyer						
10	4.99	2005-07-09	Mike	Hillyer						
11	7.99	2005-07-11	Mike	Hillyer						
12	4.99	2005-07-28	Mike	Hillyer						
13	0.99	2005-07-28	Mike	Hillyer						
14	0.99	2005-08-02	Mike	Hillyer						
15	0.99	2005-08-18	Mike	Hillyer						
16	0.99	2005-08-21	Mike	Hillyer						
17	1.99	2005-08-22	Mike	Hillyer						
18	5.99	2005-08-22	Mike	Hillyer						
19	4.99	2005-05-27	Mike	Hillyer						
20	2.99	2005-06-17	Mike	Hillyer						
21	2.99	2005-07-10	Mike	Hillyer						
22	6.99	2005-07-10	Mike	Hillyer						
23	5.99	2005-07-27	Mike	Hillyer						
24	2.99	2005-07-29	Mike	Hillyer						
25	4.99	2005-07-30	Mike	Hillyer						

## Explanation:

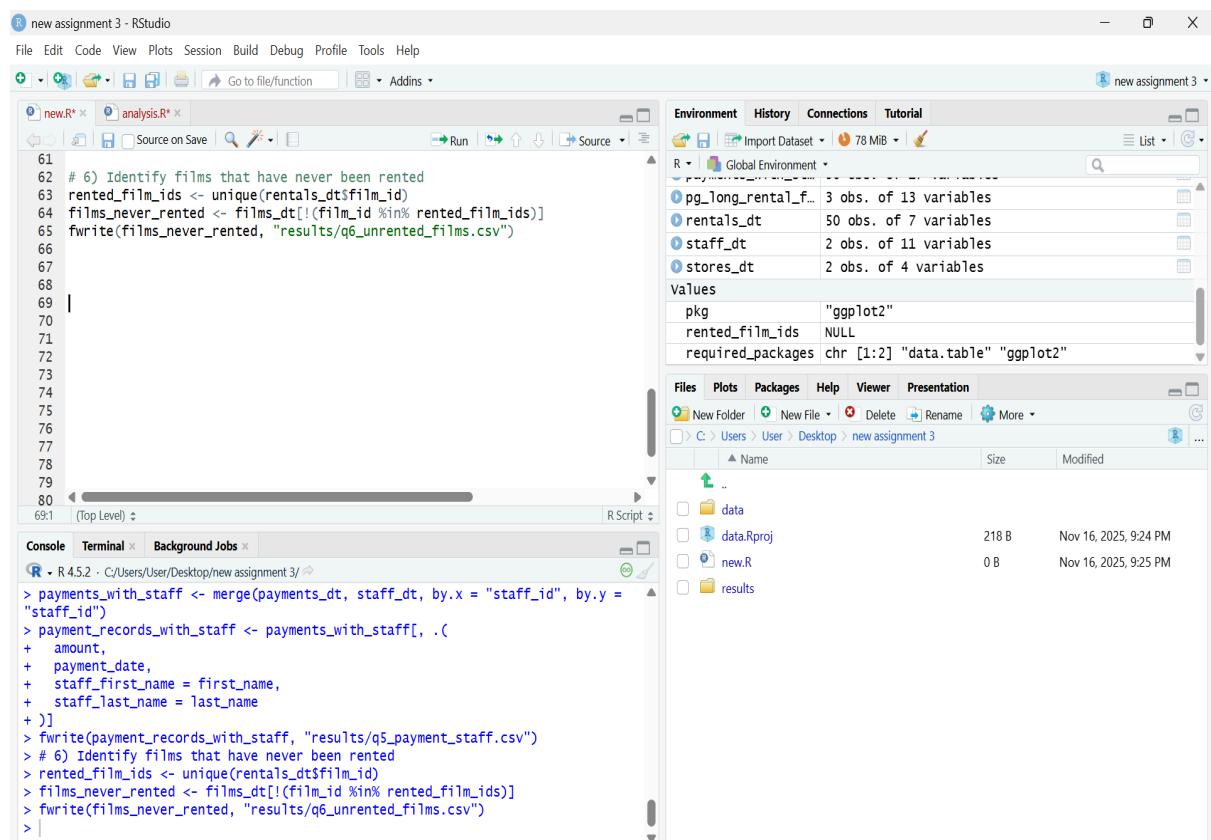
This joins the payment table with the staff table to show who handled each payment. The output includes the amount, the payment date, and the staff member's name. It helps track staff activity and payment processing.

## 6. Find the films that are not rented. [10 Marks]

Code:

```
rented_film_ids <- unique(rentals_dt$film_id)
films_never_rented <- films_dt[!(film_id %in%
rented_film_ids)]
fwrite(films_never_rented,
"results/q6_unrented_films.csv")
```

Screenshot of running code:



## Result (opened from result folder):

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	film_id	title	description	release_year	language_id	original_language	rental_duration	rental_rate	length	replacement_cost	rating	special_features	last_update	
2	1	ACADEMY A	Epic Drama	2006	1		6	0.99	86	20.99	PG	Deleted Sce	2006-02-15T05:03:42Z	
3	2	ACE GOLDF	Astounding	2006	1		3	5.49	48	12.99	G	Trailers,Del	2025-09-15T17:04:12Z	
4	3	ADAPTATION	Astounding	2006	1		7	2.99	50	18.99	NC-17	Trailers,Del	2006-02-15T05:03:42Z	
5	4	AFFAIR PRE	A Fanciful	2006	1		5	3.49	117	26.99	G	Comments,Del	2025-09-15T17:04:12Z	
6	5	AFRICAN EC	A Fast-Pace	2006	1		6	3.49	130	22.99	G	Deleted Sce	2025-09-15T17:04:12Z	
7	6	AGENT TRU	A Intrepid F	2006	1		3	2.99	169	17.99	PG	Deleted Sce	2006-02-15T05:03:42Z	
8	7	AIRPLANE S	A Touching	2006	1		6	4.99	62	28.99	PG-13	Trailers,Del	2006-02-15T05:03:42Z	
9	8	AIRPORT PCA	Epic Tale	2006	1		6	4.99	54	15.99	R	Trailers	2006-02-15T05:03:42Z	
10	9	ALABAMA EA	Thoughtf	2006	1		3	2.99	114	21.99	PG-13	Trailers,Del	2006-02-15T05:03:42Z	
11	10	ALADDIN Cr	A Action-Pa	2006	1		6	4.99	63	24.99	NC-17	Trailers,Del	2006-02-15T05:03:42Z	
12	11	ALAMO VIDA	Boring Ep	2006	1		6	1.49	126	16.99	G	Comments,Del	2025-09-15T17:04:12Z	
13	12	ALASKA PH	A Fanciful S	2006	1		6	0.99	136	22.99	PG	Comments,Del	2006-02-15T05:03:42Z	
14	13	ALI FOREVE	A Action-Pa	2006	1		4	4.99	150	21.99	PG	Deleted Sce	2006-02-15T05:03:42Z	
15	14	ALICE FANT	A Emotiona	2006	1		6	0.99	94	23.99	NC-17	Trailers,Del	2006-02-15T05:03:42Z	
16	15	ALIEN CENT	A Brilliant D	2006	1		5	2.99	46	10.99	NC-17	Trailers,Cor	2006-02-15T05:03:42Z	
17	16	ALLEY EVOLA	Fast-Pace	2006	1		6	2.99	180	23.99	NC-17	Trailers,Cor	2006-02-15T05:03:42Z	
18	17	ALONE TRIF	A Fast-Pace	2006	1		3	0.99	82	14.99	R	Trailers,Bel	2006-02-15T05:03:42Z	
19	18	ALTER VICT	A Thoughtf	2006	1		6	0.99	57	27.99	PG-13	Trailers,Bel	2006-02-15T05:03:42Z	
20	19	AMADEUS IA	Emotiona	2006	1		6	0.99	113	20.99	PG	Comments,Del	2006-02-15T05:03:42Z	
21	20	AMELIE HEIA	Boring Dr	2006	1		4	4.99	79	23.99	R	Comments,Del	2006-02-15T05:03:42Z	
22	21	AMERICAN A	Insightful	2006	1		3	4.99	129	17.99	R	Comments,Del	2006-02-15T05:03:42Z	
23	22	AMISTAD M	A Emotiona	2006	1		6	3.49	85	10.99	G	Comments,Del	2025-09-15T17:04:12Z	
24	23	ANACONDA A	Lacklustu	2006	1		3	0.99	92	9.99	R	Trailers,Del	2006-02-15T05:03:42Z	
25	24	ANALYZE HI	A Thoughtf	2006	1		6	2.99	181	19.99	R	Trailers,Bel	2006-02-15T05:03:42Z	

## Explanation:

This query finds all films whose `film_id` does not appear in the `rentals` table. It basically checks which movies have zero rental records. This helps identify films that are not being watched or rented.

## 7. Plot any graph of your choice. [10 Marks]

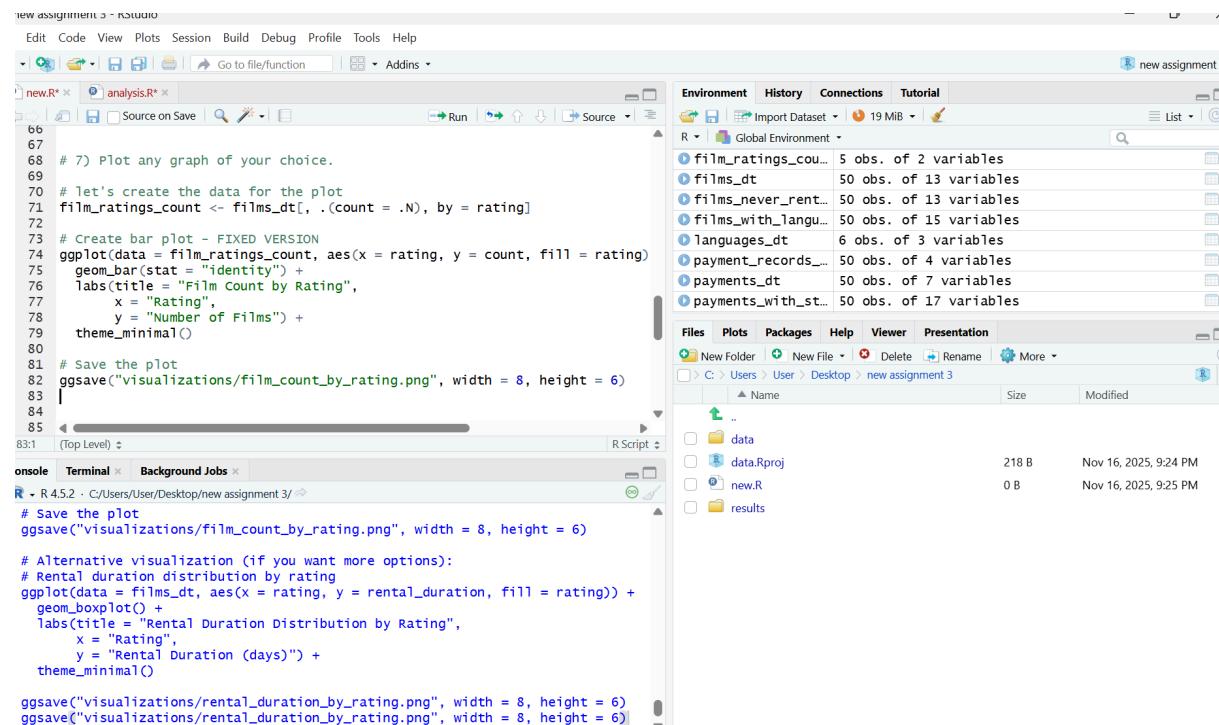
### Code:

```
# let's create the data for the plot
film_ratings_count <- films_dt[, .(count = .N), by =
rating]

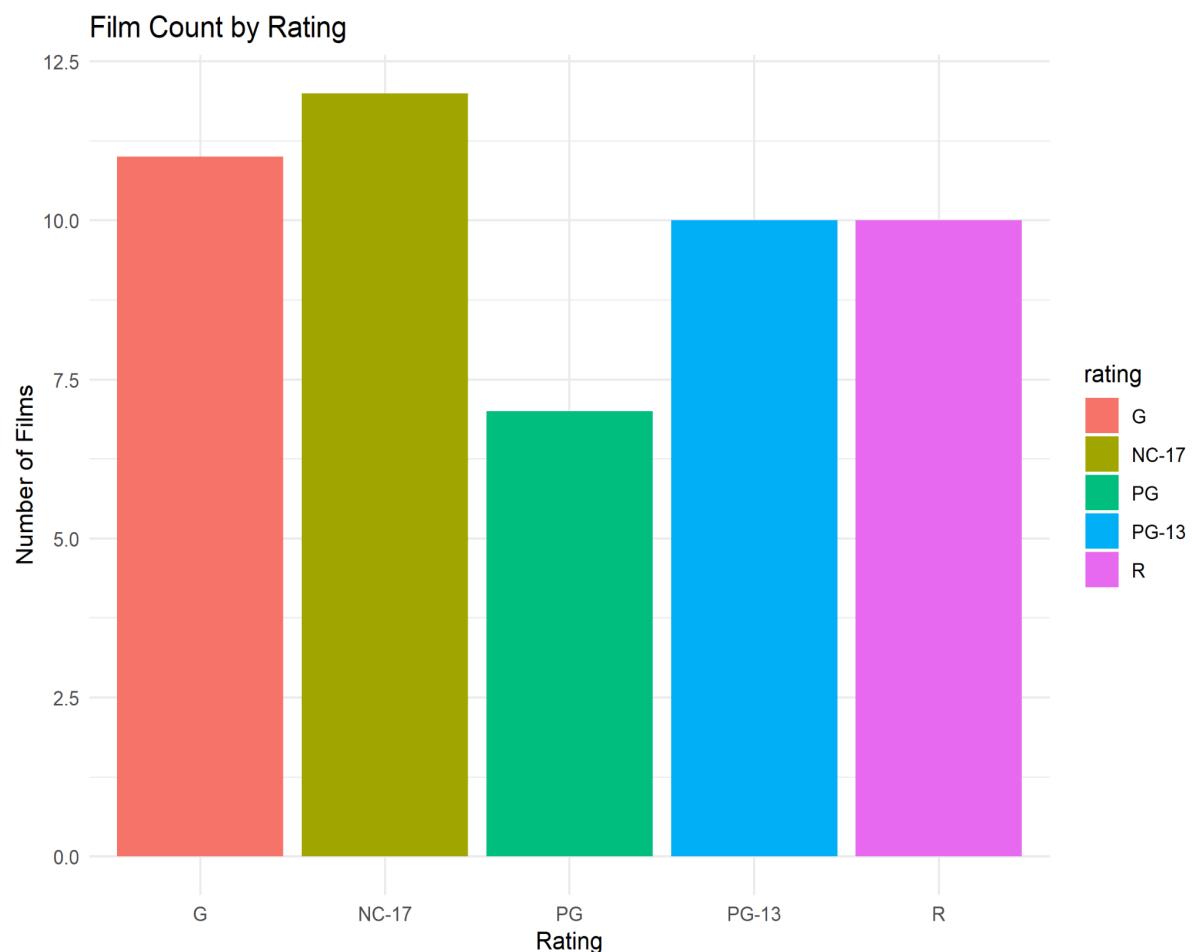
# Create bar plot - FIXED VERSION
ggplot(data = film_ratings_count, aes(x = rating, y =
count, fill = rating)) +
  geom_bar(stat = "identity") +
  labs(title = "Film Count by Rating",
       x = "Rating",
       y = "Number of Films") +
  theme_minimal()

# Save the plot
ggsave("visualizations/film_count_by_rating.png", width =
8, height = 6)
```

### Screenshot of running code:



**Visual/graph (from visualisation folder):**



**Explanation:**

For the plot, I drew a bar graph, counted how many films fall under each rating and visualized it using a bar chart. This helps quickly see which ratings have the highest number of films.

## 8. Use of Git in the entire assignment. [30 Marks]

### Initial commit:

```
C:\ Command Prompt X + v - D X

:\Users\User\Desktop\Assignment_3_MSDS25038
:\Users\User\Desktop\Assignment_3_MSDS25038>git init
initialized empty Git repository in C:/Users/User/Desktop/Assignment_3_MSDS25038/.git/
:\Users\User\Desktop\Assignment_3_MSDS25038>git add .
warning: in the working copy of '.proj.user\FAACAT71\pcs\files-pane.ppr', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of '.proj.user\FAACAT71\pcs\source-pane.ppr', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of '.proj.user\FAACAT71\pcs>windowlayoutstate.ppr', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of '.proj.user\FAACAT71\pcs\workbench-pane.ppr', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of '.proj.user\FAACAT71\sources\prop\B9306586', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of '.proj.user\FAACAT71\sources\prop\B97903F1', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of '.proj.user\FAACAT71\sources\prop\INDEX', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of '.proj.user\FAACAT71\sources\session-fcc45762\441ED9A4', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of '.proj.user\FAACAT71\sources\session-fcc45762\441ED9A4-contents', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of '.proj.user\FAACAT71\sources\session-fcc45762\9298046E', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of '.proj.user\FAACAT71\sources\session-fcc45762\9298046E-contents', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'data\customer.csv', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'data\film.csv', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'data\language.csv', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'data\payment.csv', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'data\rental.csv', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'data\staff.csv', LF will be replaced by CRLF the next time Git touches it
warning: in the working copy of 'data\store.csv', LF will be replaced by CRLF the next time Git touches it

:\Users\User\Desktop\Assignment_3_MSDS25038>git commit -m "Initial Commit"
[master (root-commit) 76af15a] Initial Commit
 20 files changed, 484 insertions(+)
create mode 100644 .Rproj.user\FAACAT71\pcs\files-pane.ppr
create mode 100644 .Rproj.user\FAACAT71\pcs\source-pane.ppr
create mode 100644 .Rproj.user\FAACAT71\pcs>windowlayoutstate.ppr
create mode 100644 .Rproj.user\FAACAT71\pcs\workbench-pane.ppr
create mode 100644 .Rproj.user\FAACAT71\sources\prop\B9306586
create mode 100644 .Rproj.user\FAACAT71\sources\prop\B97903F1
create mode 100644 .Rproj.user\FAACAT71\sources\prop\INDEX
create mode 100644 .Rproj.user\FAACAT71\sources\session-fcc45762\441ED9A4
create mode 100644 .Rproj.user\FAACAT71\sources\session-fcc45762\441ED9A4-contents
create mode 100644 .Rproj.user\FAACAT71\sources\session-fcc45762\9298046E
create mode 100644 .Rproj.user\FAACAT71\sources\session-fcc45762\9298046E-contents
create mode 100644 .Rproj.user\FAACAT71\sources\session-fcc45762\lock_file
create mode 100644 .Rproj.user\shared\notebooks\patch-chunk-names
create mode 100644 data\customer.csv
create mode 100644 data\film.csv
create mode 100644 data\language.csv
create mode 100644 data\payment.csv
create mode 100644 data\rental.csv
create mode 100644 data\staff.csv
create mode 100644 data\store.csv

:\Users\User\Desktop\Assignment_3_MSDS25038>git remote add origin https://github.com/kawishbinsarfraz-debug/Assignment_3_MSDS25038.git
:\Users\User\Desktop\Assignment_3_MSDS25038>git push -u origin main
error: src refspec main does not match any
error: failed to push some refs to 'https://github.com/kawishbinsarfraz-debug/Assignment_3_MSDS25038.git'
:\Users\User\Desktop\Assignment_3_MSDS25038>git push -u origin main
error: src refspec main does not match any
error: failed to push some refs to 'https://github.com/kawishbinsarfraz-debug/Assignment_3_MSDS25038.git'
:\Users\User\Desktop\Assignment_3_MSDS25038>git branch -M main
:\Users\User\Desktop\Assignment_3_MSDS25038>git push -u origin main
info: please complete authentication in your browser...
  numerating objects: 30, done.
  counting objects: 100% (30/30), done.
  delta compression using up to 12 threads
  compressing objects: 100% (26/26), done.
  writing objects: 100% (30/30), 50.12 KiB | 1.67 MiB/s, done.
  total 30 (delta 3), reused 0 (delta 0), pack-reused 0 (from 0)
  remote: Resolving deltas: 100% (3/3), done.
o https://github.com/kawishbinsarfraz-debug/Assignment_3_MSDS25038.git
* [new branch]      main -> main
branch 'main' set up to track 'origin/main'.

:\Users\User\Desktop\Assignment_3_MSDS25038>
```

## Committing R script and its results:

```
C:\Users\User\Desktop\Assignment_3_MSDS25038>git status
On branch main
Your branch is up to date with 'origin/main'.

Untracked files:
  (use "git add <file>..." to include in what will be committed)
    data.Rproj
    new.R
    results/
    visualizations/

nothing added to commit but untracked files present (use "git add" to track)

C:\Users\User\Desktop\Assignment_3_MSDS25038>git add .

C:\Users\User\Desktop\Assignment_3_MSDS25038>git commit -m "Adding R script file and output result files and visualizations"
[main 97bff7b] Adding R script file and output result files and visualizations
 10 files changed, 178 insertions(+)
 create mode 100644 data.Rproj
 create mode 100644 new.R
 create mode 100644 results/q1_pg_films.csv
 create mode 100644 results/q2_avg_rental_by_rating.csv
 create mode 100644 results/q3_film_count_by_language.csv
 create mode 100644 results/q4_customers_store.csv
 create mode 100644 results/q5_payment_staff.csv
 create mode 100644 results/q6_unrented_films.csv
 create mode 100644 visualizations/film_count_by_rating.png
 create mode 100644 visualizations/rental_duration_by_rating.png

C:\Users\User\Desktop\Assignment_3_MSDS25038>git push
Enumerating objects: 15, done.
Counting objects: 100% (15/15), done.
Delta compression using up to 12 threads
Compressing objects: 100% (12/12), done.
Writing objects: 100% (14/14), 46.18 KiB | 2.72 MiB/s, done.
Total 14 (delta 0), reused 0 (delta 0), pack-reused 0 (from 0)
To https://github.com/kawishbinsarfraz-debug/Assignment_3_MSDS25038.git
 76af15a..97bff7b main -> main

C:\Users\User\Desktop\Assignment_3_MSDS25038>
```

## Final commit:

In the final commit I added this report in my local repo and then committed these changes into the Github repository.

## Github repo:

[https://github.com/kawishbinsarfraz-debug/Assignment\\_3\\_MSDS25038](https://github.com/kawishbinsarfraz-debug/Assignment_3_MSDS25038)