CSE - 5334 - 001 : Data Mining

Assignment 1 - Exploratory Data Analysis - R

Instructor: Dr. Elizabeth D Diaz

Team 18: Urmi Manish Sheth – 1002064934

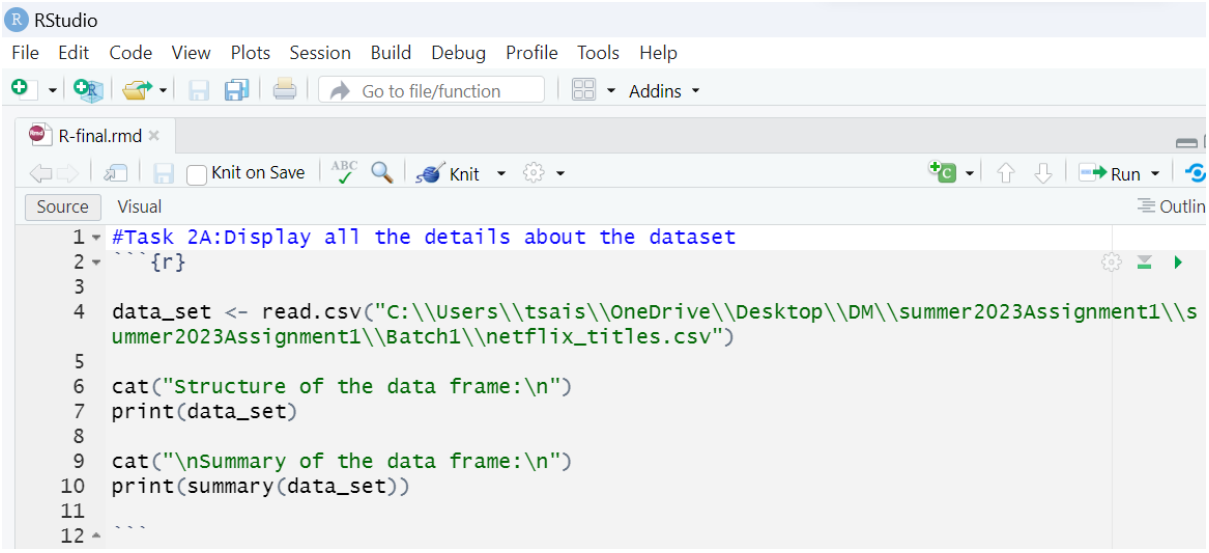
Harsh Navinbhai Shah – 1002057387

Sai Swetha Tadivaka – 1002112726

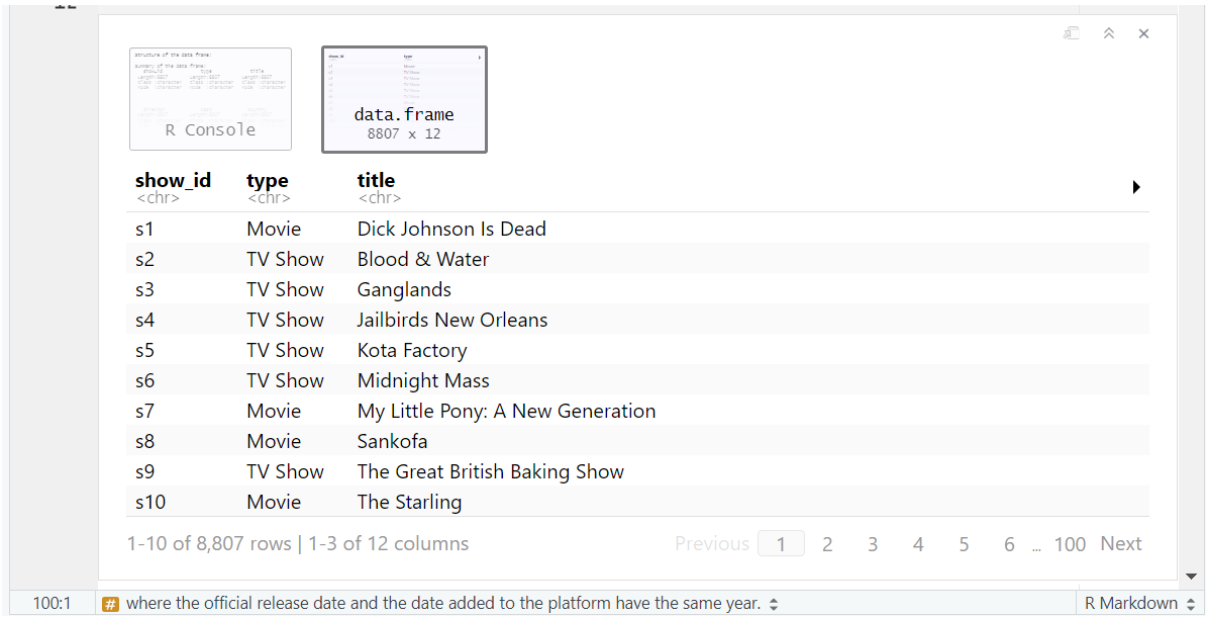
Part 2: R (40 Points) Perform all the tasks in R, using a different R notebook. The data and questions will be the same as those in Task 1.

Task 1A (2 points): Display all the details about the dataset.

Solution

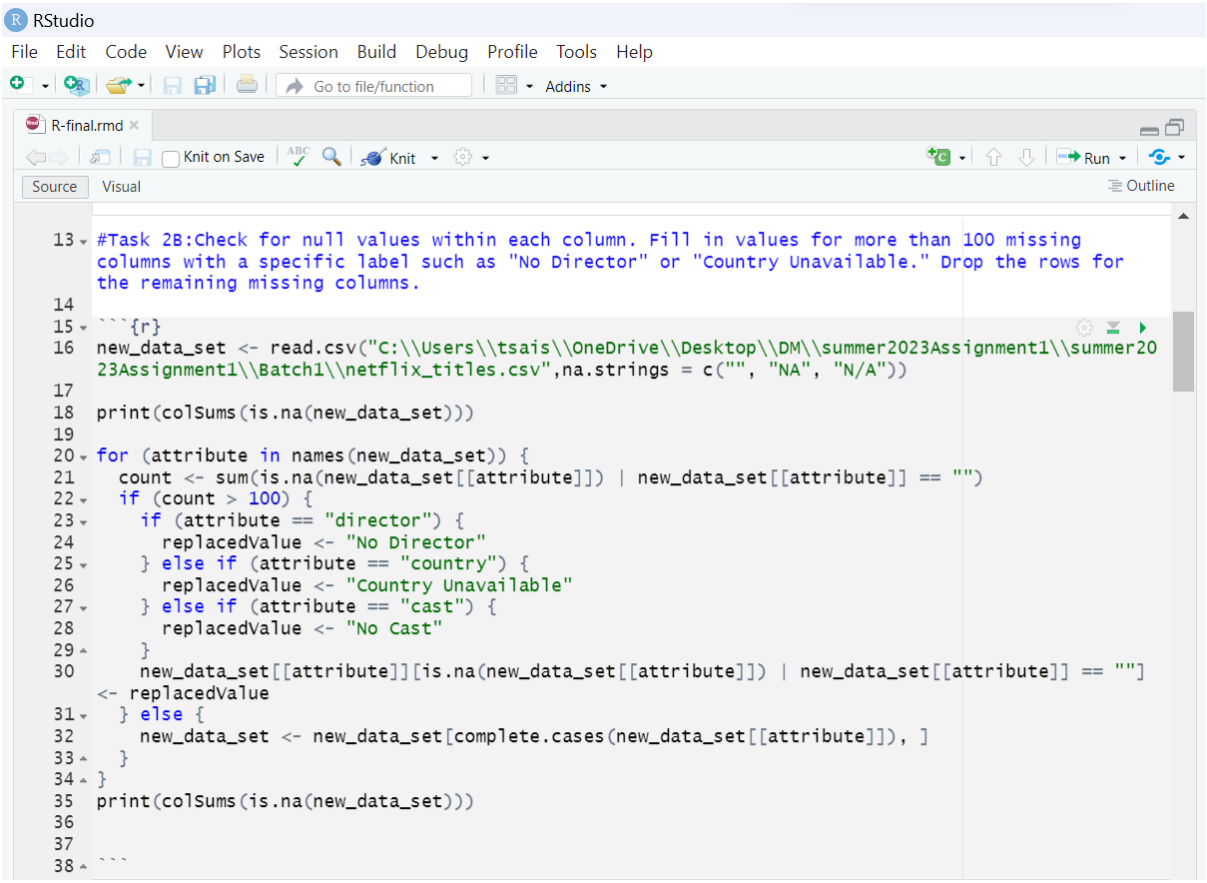


Output



Task 2B (3 points): Check for null values within each column. Fill in values for more than 100 missing columns with a specific label such as "No Director" or "Country Unavailable." Drop the rows for the remaining missing columns.

Solution




```r
#Task 2B:Check for null values within each column. Fill in values for more than 100 missing columns with a specific label such as "No Director" or "Country Unavailable." Drop the rows for the remaining missing columns.

```{r}
new_data_set <- read.csv("C:\\Users\\tsais\\OneDrive\\Desktop\\DM\\summer2023Assignment1\\summer2023Assignment1\\Batch1\\netflix_titles.csv",na.strings = c("", "NA", "N/A"))

print(colSums(is.na(new_data_set)))

for (attribute in names(new_data_set)) {
  count <- sum(is.na(new_data_set[[attribute]]) | new_data_set[[attribute]] == "")
  if (count > 100) {
    if (attribute == "director") {
      replacedValue <- "No Director"
    } else if (attribute == "country") {
      replacedValue <- "Country Unavailable"
    } else if (attribute == "cast") {
      replacedValue <- "No Cast"
    }
    new_data_set[[attribute]][is.na(new_data_set[[attribute]]) | new_data_set[[attribute]] == ""] <- replacedValue
  } else {
    new_data_set <- new_data_set[complete.cases(new_data_set[[attribute]]), ]
  }
}
print(colSums(is.na(new_data_set)))


```
```

Output: Before eliminating null values

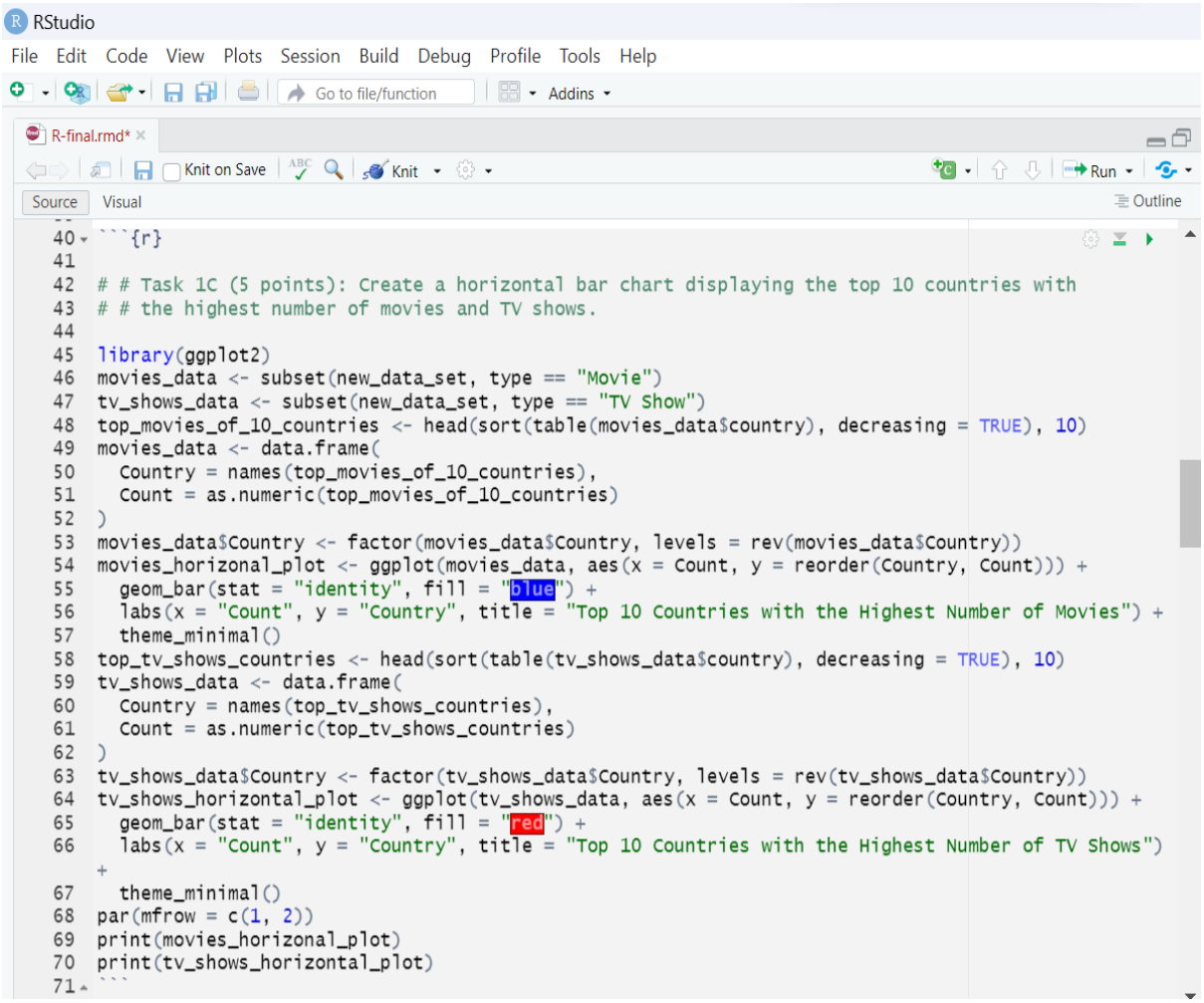
```
    show_id        type       title    director        cast
          0           0           0        2634         825
    country  date_added release_year      rating    duration
        831          10           0           4           3
 listed_in description
          0           0
```

Output: After eliminating null values

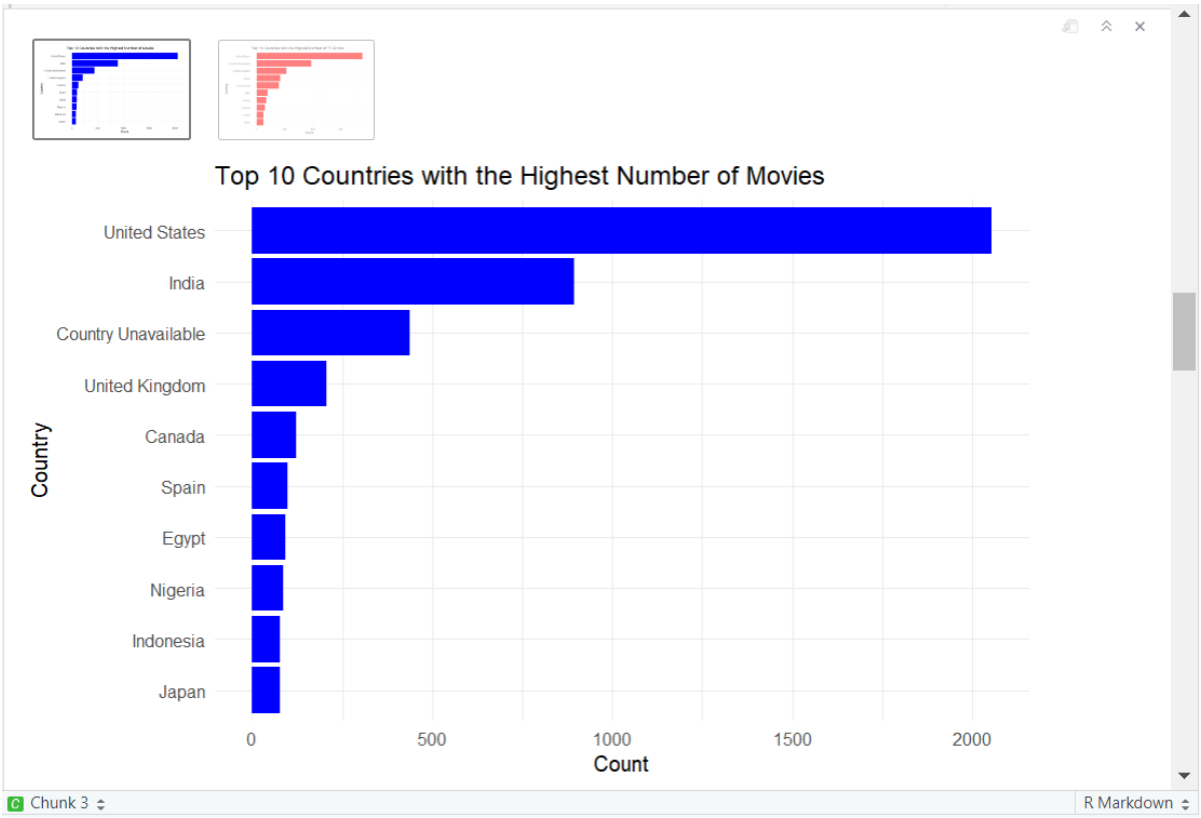
```
    show_id        type       title    director        cast
          0           0           0           0           0
    country  date_added release_year      rating    duration
          0           0           0           0           0
 listed_in description
          0           0
```

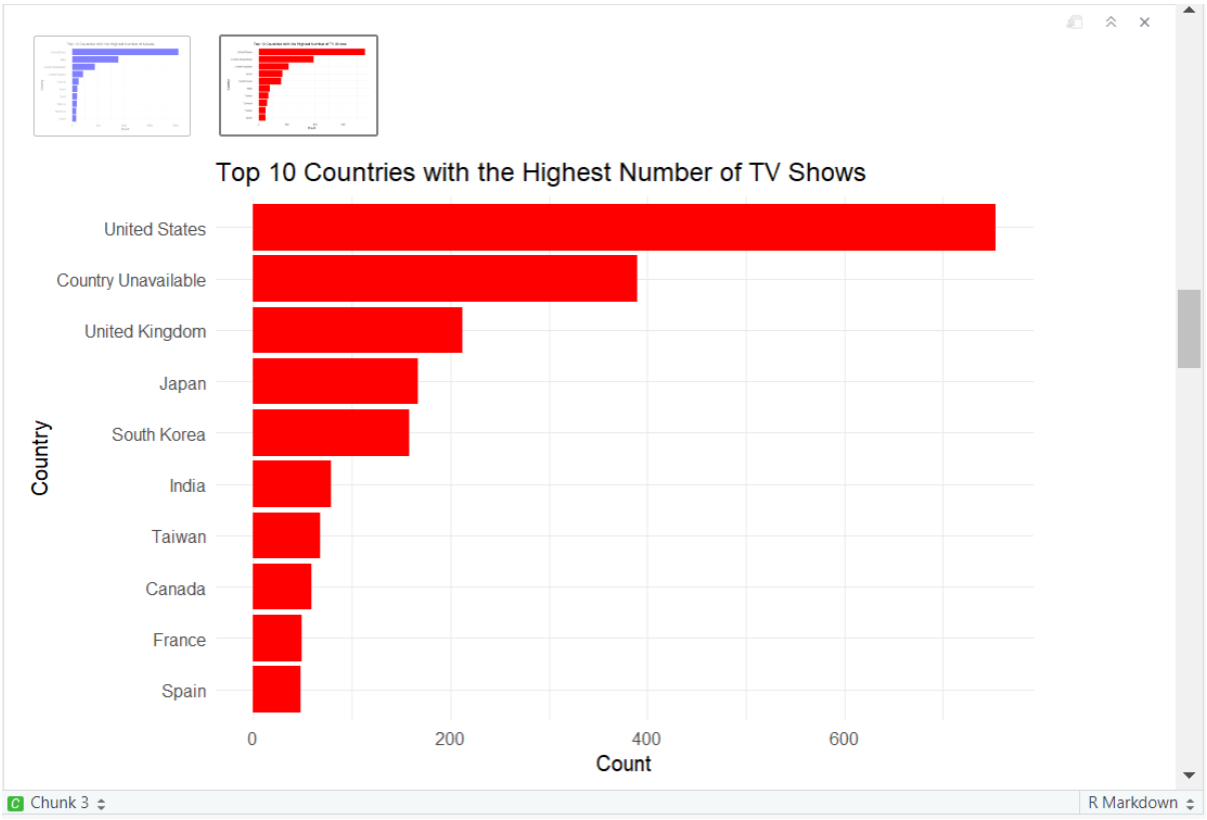
Task 2C (5 points): Create a horizontal bar chart displaying the top 10 countries with the highest number of movies and TV shows.

Solution



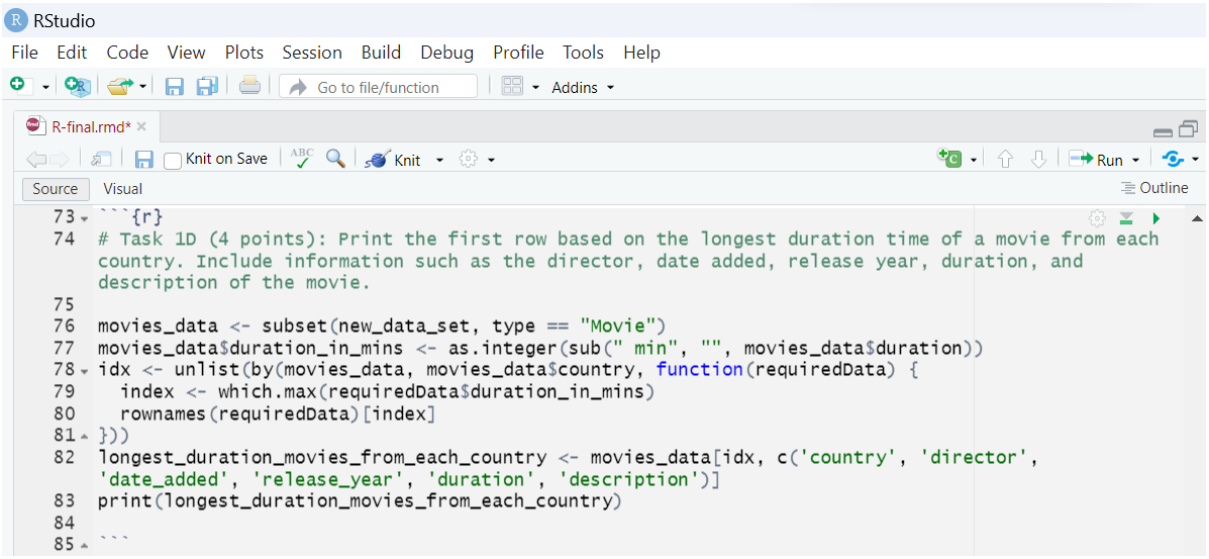
Output


Top 10 Countries with the Highest Number of Movies
United States
India
Country Unavailable
United Kingdom
Canada
Spain
Egypt
Nigeria
Indonesia
Japan
Country
0
500
1000
1500
2000
Count
Chunk 3
R Markdown

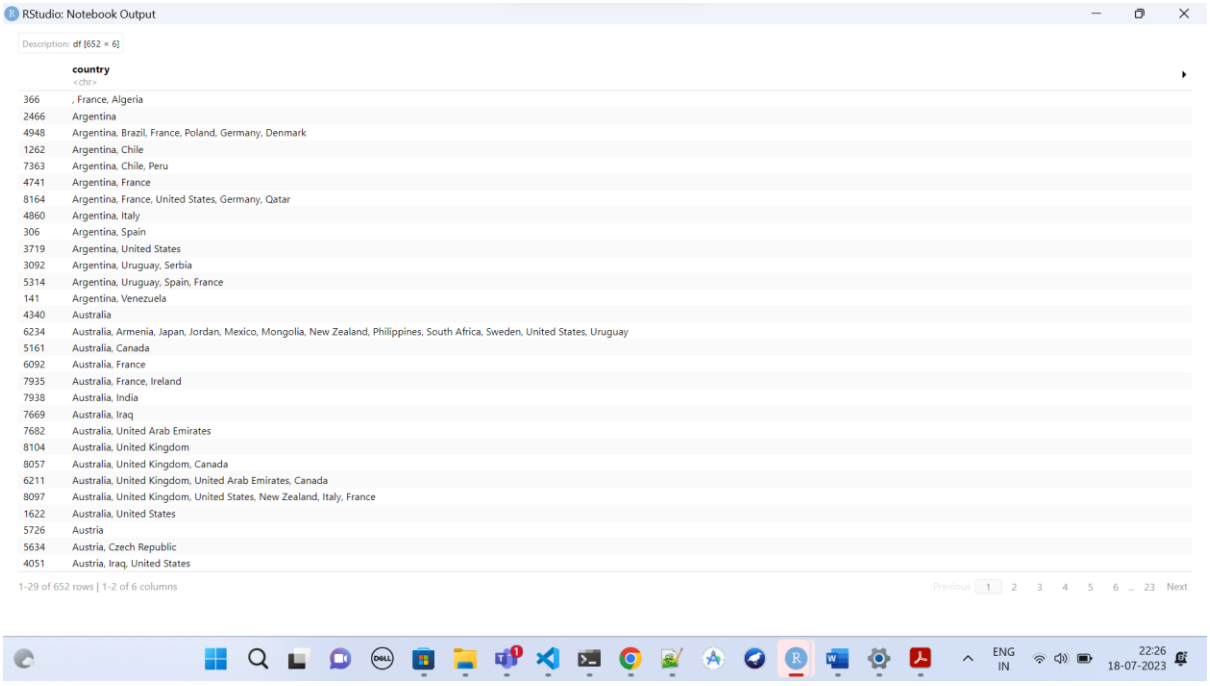

Top 10 Countries with the Highest Number of TV Shows
United States
Country Unavailable
United Kingdom
Japan
South Korea
India
Taiwan
Canada
France
Spain
Country
0
200
400
600
Count
Chunk 3
R Markdown

Task 2D (4 points): Print the first row based on the longest duration time of a movie from each country. Include information such as the director, date added, release year, duration, and description of the movie.

Solution



Output



RStudio: Notebook Output

Description: df [652 × 6]

director
Najwa Najjar
Alejandro Montiel
Diego Lerman
Pablo Larraín
Ticoy Rodriguez
Pablo Giorgelli
Natalia Garagiola
Mercedes Cordova
Marcelo Piñeyro
Miguel Cohan, Miguel Cohan
Emir Kusturica
Israel Adrián Caetano
Sebastián Schindel
Baljit Singh Deo
Matthew Salleh
Josh Wakely
Anne Fontaine
Luc Vinciguerra
Rohit Jugraj
Bill Guttentag, Michael Ware
Ben C. Lucas
Baz Luhrmann
Julius Avery
Michael Petroni
Alex Holmes
Simon Wincer
Kief Davidson, Richard Ladkani
Andreas Prochaska
Fritz Ofner

1-29 of 652 rows | 3-3 of 6 columns

Previous 1 2 3 4 5 6 .. 23 Next

RStudio: Notebook Output

Description: df [652 × 6]

date_added	release_year	duration
July 30, 2021	2014	103 min
May 28, 2020	2020	116 min
April 1, 2018	2017	96 min
February 26, 2021	2006	112 min
July 15, 2016	2015	100 min
August 1, 2018	2017	87 min
July 15, 2018	2017	108 min
May 16, 2018	2017	66 min
August 5, 2021	1993	124 min
June 21, 2019	2019	113 min
December 27, 2019	2018	73 min
September 1, 2017	2017	113 min
September 1, 2021	2014	100 min
December 1, 2018	2015	146 min
August 15, 2017	2017	101 min
November 21, 2017	2017	51 min
November 20, 2019	2013	111 min
December 20, 2015	2010	72 min
October 15, 2017	2016	145 min
November 15, 2018	2015	76 min
October 15, 2017	2017	96 min
January 1, 2020	1992	95 min
July 19, 2018	2014	109 min
December 25, 2018	2015	90 min
February 15, 2015	2014	100 min
December 1, 2020	1990	120 min
November 4, 2016	2016	112 min
January 15, 2017	2014	98 min
March 2, 2019	2018	89 min

1-29 of 652 rows | 4-6 of 6 columns

Previous 1 2 3 4 5 6 .. 23 Next

RStudio: Notebook Output

Descriptions: df [652 × 6]

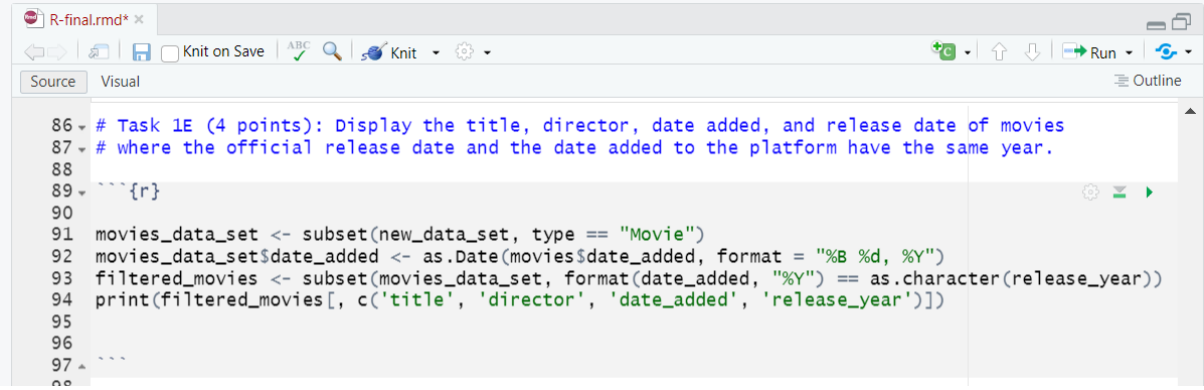
duration <chr>	description <chr>
103 min	After a decade in prison, a Palestinian man with a dark secret returns to the West Bank and searches for the daughter he lost long ago.
116 min	Police officer Pipa works on her first big case while simultaneously investigating her boss, who is suspected of murder. The prequel to "Perdida".
96 min	An Argentine doctor faces legal and ethical challenges when she travels to the countryside to pick up the infant she has been waiting to adopt.
112 min	A struggling music student discovers and tries to complete an unfinished symphony begun by a famous composer, whose career was ended by a past trauma.
100 min	Three men from different nations must find their way out of the Amazon jungle – on foot – when they get lost on the way to Brazil's World Cup Final.
87 min	After learning she's pregnant, a teen grapples with the consequences while trying to go about her life as if nothing has changed.
108 min	Animosity gives way to love and forgiveness when an angry teenager reunites with his estranged biological father in the wilds of Patagonia.
66 min	Epicureans and experts trace the migration of Italian cuisine and its evolution in the Americas, delving into the ideas of origin, diaspora and heritage.
124 min	One of the first rock stars in Argentina, Tango leads a full life during an era of social unrest – one of both glory and despair.
113 min	Family patriarch Elías begins to unravel after the death of his wife – which casts a suspicious light on her tragic accident.
73 min	In this intimate documentary, former Uruguayan President José "Pepe" Mujica talks about lessons he learned while in prison, his ideals and the future.
113 min	While settling the affairs of his murdered mother and brother, a man partners with a local thug in a life insurance scam, putting him on a dark path.
100 min	A lawyer defends an illiterate man whose exploitation by a cruel boss while working as a butcher in Buenos Aires led to tragedy. Based on a true case.
146 min	When a man is arrested on arrival in America because authorities mistake him for a wanted gangster, a new friend tries to help him prove his innocence.
101 min	A blend of cultural nuance and mesmerizing techniques adds flavor to this globe-hopping celebration of cooking, tradition and community.
51 min	After winning a local talent show, the Beat Bugs journey to compete on "The Bug Factor," a televised singing contest held in faraway Rocket Ship Park.
111 min	When lifelong friends Roz and Lil fall in love with each other's teenage sons, they must carry out their affairs in relative secrecy.
72 min	Young orphan Nicolas learns how to guide a sleigh, slide down chimneys and other duties he'll need when he becomes the next Santa Claus.
145 min	A simple farmer's life becomes complicated after he resolves to save his village from an economic crisis by traveling to Australia to earn money.
76 min	Given unprecedented access to soldiers and terrorists alike, journalist Michael Ware presents an epic, uncensored, intimate account of the Iraq War.
96 min	After inventing a drug that induces time-compressed virtual realities, young Ren grapples with partner Sam over how to use their powerful creation.
95 min	Dumped by his partner just before a major dance competition, gifted hoofer Scott Hastings is forced to take a graceless neophyte as his new partner.
109 min	A young petty thief who received protection from a notorious inmate in jail is forced to spring his defender from prison and help pull off a heist.
90 min	Haunted by visions of the dead, psychologist Peter returns home to confront a long-ago trauma. But will the trip exorcise or embolden the ghosts?
100 min	An athlete dupes the world with his tale of miraculous recovery from cancer, becoming a sports hero and beacon of hope, but can he outpace the truth?
120 min	After traveling to Australia for a job, a sharpshooting cowboy becomes a target himself when he refuses an immoral assignment from a corrupt landowner.
112 min	Filmmakers infiltrate the corrupt global network of ivory trafficking, exposing poachers and dealers as African elephants edge closer to extinction.
98 min	Tasked with unraveling the assassination of Archduke Franz Ferdinand, an honorable magistrate finds himself targeted by the shadowy perpetrators.
89 min	Speaking to retailers and owners of the Glock pistol, two filmmakers shed light on the politics of violence and its implications from Chicago to Kirkuk.

1-29 of 652 rows | 6-7 of 6 columns

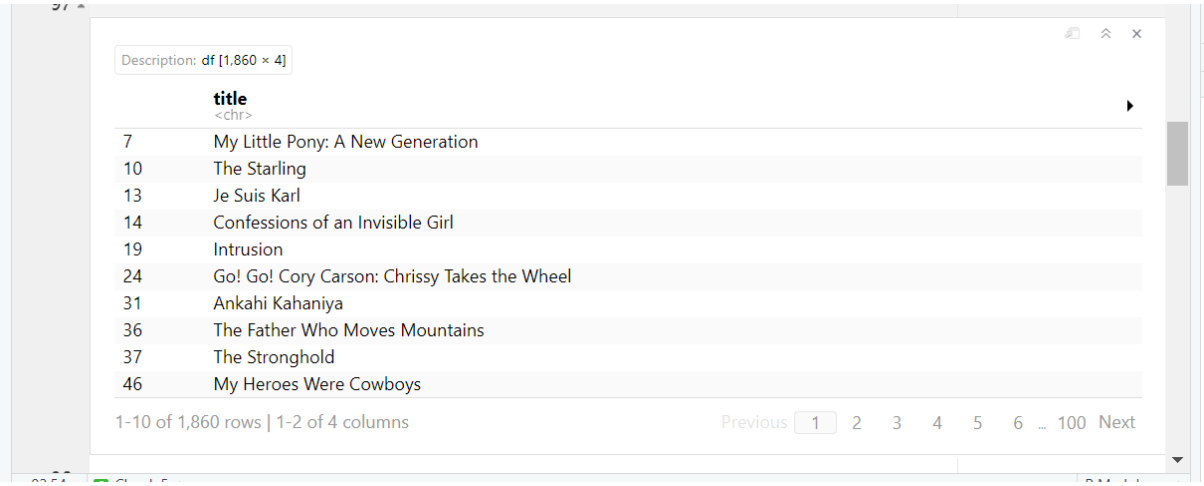
Previous 1 2 3 4 5 6 … 23 Next

Task 2E (4 points): Display the title, director, date added, and release date of movies where the official release date and the date added to the platform have the same year.

Solution



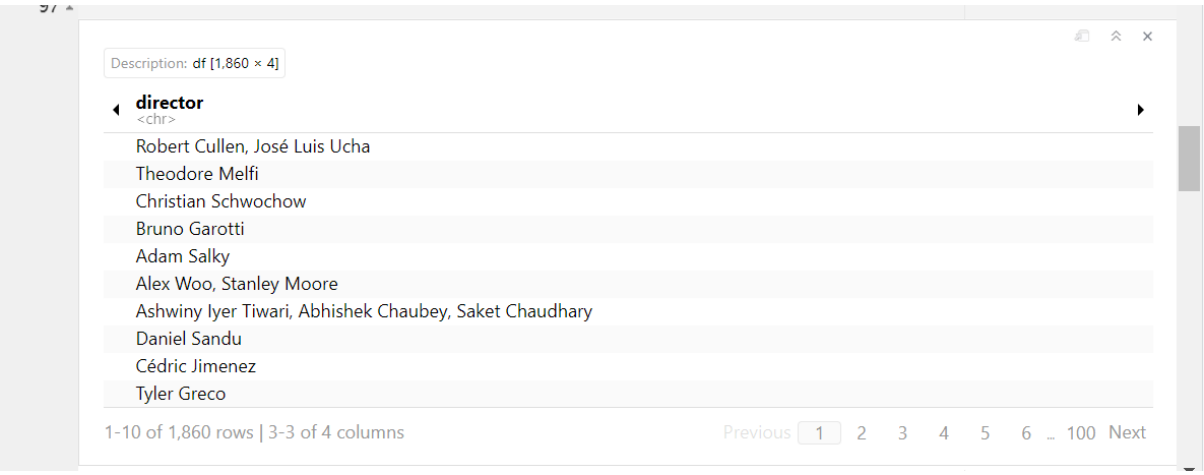
Output



Description: df [1,860 × 4]

	title <chr>
7	My Little Pony: A New Generation
10	The Starling
13	Je Suis Karl
14	Confessions of an Invisible Girl
19	Intrusion
24	Go! Go! Cory Carson: Chrissy Takes the Wheel
31	Ankahi Kahaniya
36	The Father Who Moves Mountains
37	The Stronghold
46	My Heroes Were Cowboys

1-10 of 1,860 rows | 1-2 of 4 columns Previous 1 2 3 4 5 6 … 100 Next



Description: df [1,860 × 4]

director <chr>
Robert Cullen, José Luis Ucha
Theodore Melfi
Christian Schwochow
Bruno Garotti
Adam Salky
Alex Woo, Stanley Moore
Ashwiny Iyer Tiwari, Abhishek Chaubey, Saket Chaudhary
Daniel Sandu
Cédric Jimenez
Tyler Greco

1-10 of 1,860 rows | 3-3 of 4 columns Previous 1 2 3 4 5 6 … 100 Next

Description: df [1,860 × 4]

	date_added <date>	release_year <int>
	2021-09-24	2021
	2021-09-24	2021
	2021-09-23	2021
	2021-09-22	2021
	2021-09-22	2021
	2021-09-21	2021
	2021-09-17	2021
	2021-09-17	2021
	2021-09-17	2021
	2021-09-16	2021

1-10 of 1,860 rows | 4-5 of 4 columns

Previous 1 2 3 4 5 6 ... 100 Next

Task 2F (4 points): Display the director, release year, and the number of movies and TV shows directed by each director within a year. Sort the results in descending order based on count

Solution

```r
# Task 1F (4 points): Display the director, release year, and the number of movies and
# TV shows directed by each director within a year. Sort the results in descending order
# based on the count.

movies_tvshows_data <- subset(new_data_set, type %in% c("Movie", "TV Show"))
director_count <- aggregate(cbind(count = type) ~ director + release_year, data =
movies_tvshows_data, FUN = length)
director_count <- director_count[order(-director_count$count), ]
desired_columns <- c("director", "release_year", "count")
print(director_count[desired_columns])
```

Output

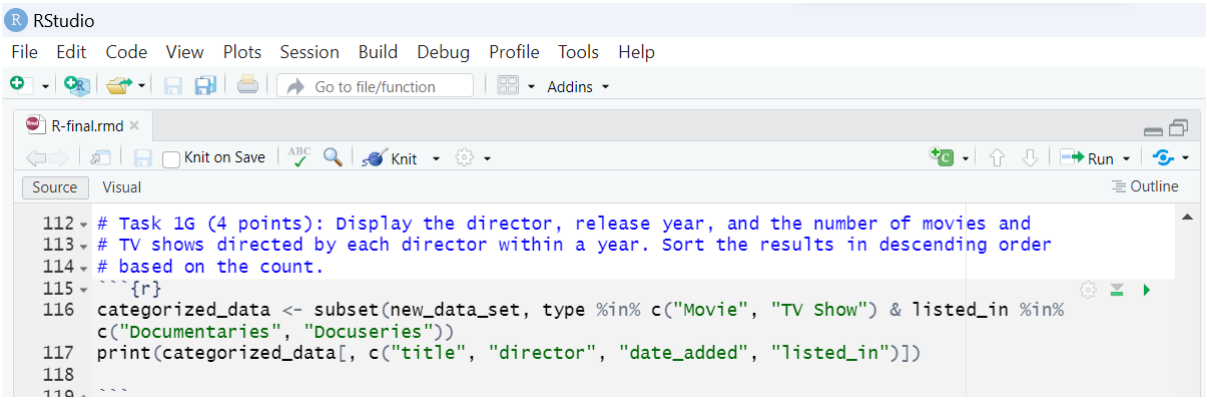
Description: df [5,948 × 3]

	director <chr>	release_year <int>	count <int>
5495	No Director	2020	405
4934	No Director	2019	401
4310	No Director	2018	386
5869	No Director	2021	295
3605	No Director	2017	258
2918	No Director	2016	248
2372	No Director	2015	157
2002	No Director	2014	85
1554	No Director	2012	61
1757	No Director	2013	61

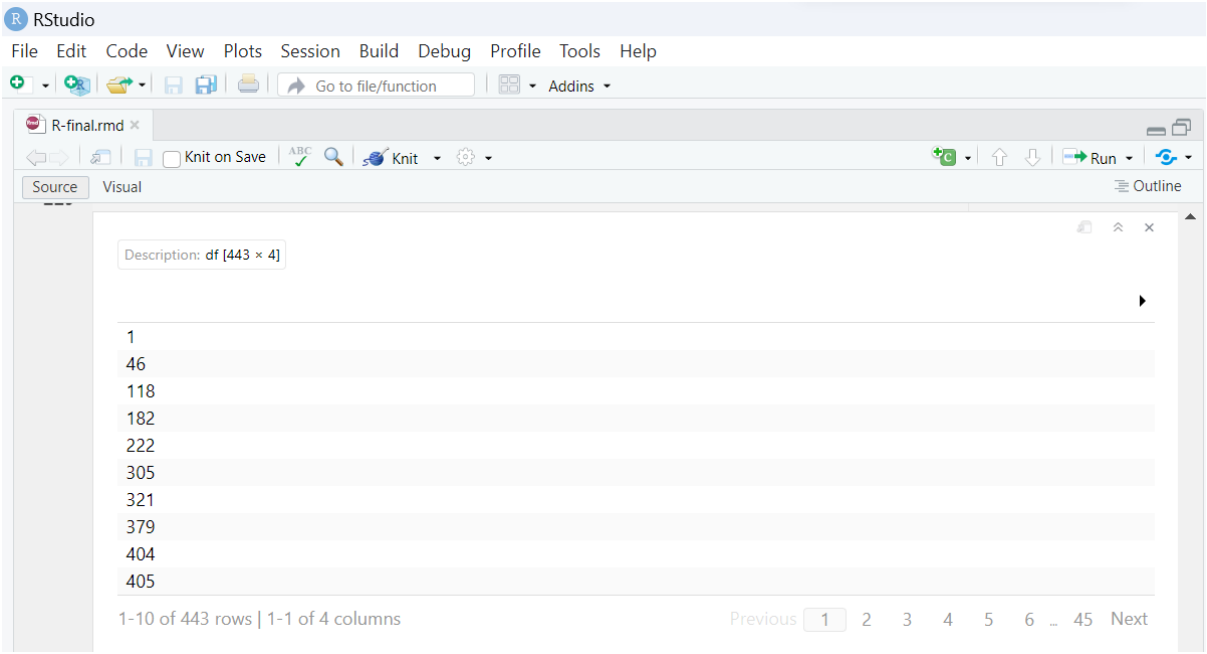
1-10 of 5,948 rows Previous 1 2 3 4 5 6 ... 100 Next

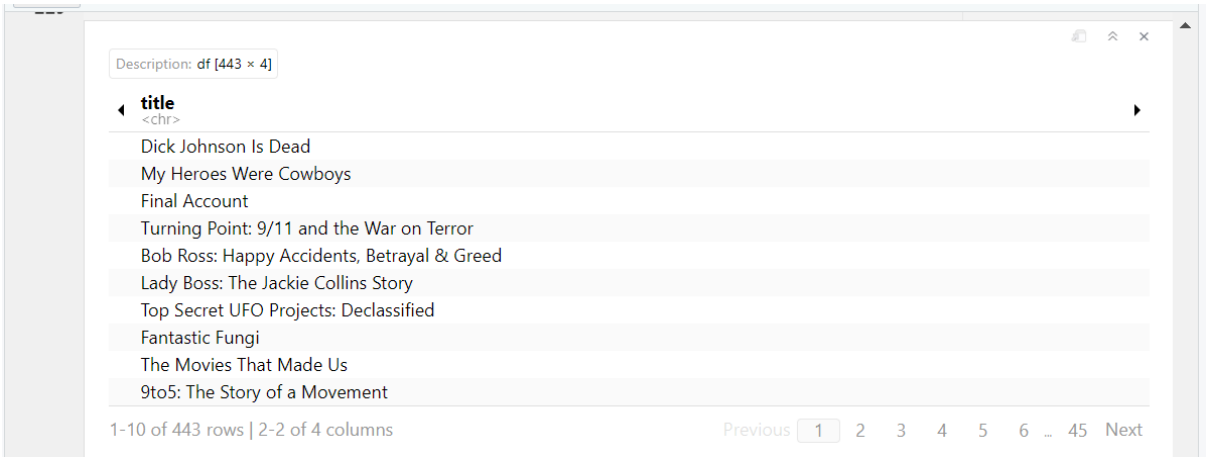
Task 2G (3 points): Display the title, director, date added, and category of movies or TV shows from the Documentary/Docuseries category.

Solution



Output





Description: df [443 × 4]

director <chr>
Kirsten Johnson
Tyler Greco
Luke Holland
No Director
Joshua Rofé
Laura Fairrie
No Director
Louie Schwartzberg
No Director
No Director

1-10 of 443 rows | 3-3 of 4 columns

Previous 1 2 3 4 5 6 ... 45 Next

Description: df [443 × 4]

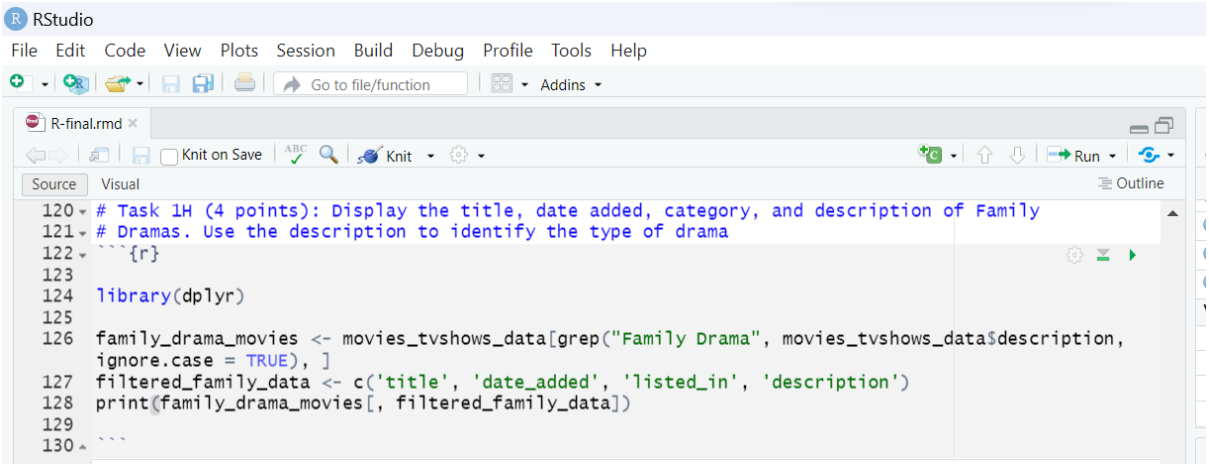
date_added <chr>	listed_in <chr>
September 25, 2021	Documentaries
September 16, 2021	Documentaries
September 2, 2021	Documentaries
September 1, 2021	Docuseries
August 25, 2021	Documentaries
August 5, 2021	Documentaries
August 3, 2021	Docuseries
July 28, 2021	Documentaries
July 23, 2021	Docuseries
July 22, 2021	Documentaries

1-10 of 443 rows | 4-5 of 4 columns

Previous 1 2 3 4 5 6 ... 45 Next

Task 2H (4 points): Display the title, date added, category, and description of Family Dramas. Use the description to identify the type of drama.

Solution



Output

Description: df [7 × 4]

	title <chr>	date_added <chr>
2276	Hook	July 5, 2020
2781	Curtiz	March 25, 2020
2881	I Am Not Okay With This	February 26, 2020
2998	October Faction	January 23, 2020
3177	Three Days of Christmas	December 6, 2019
6994	Hope Aur Hum	September 1, 2018
7040	I Am Me!	March 3, 2017

7 rows | 1-3 of 4 columns

Description: df [7 × 4]

listed_in <chr>
International TV Shows, Romantic TV Shows, TV Comedies
Dramas, International Movies
TV Comedies, TV Dramas, TV Sci-Fi & Fantasy
TV Action & Adventure, TV Dramas, TV Horror
International TV Shows, Spanish-Language TV Shows, TV Dramas
Children & Family Movies, Comedies, Dramas
International TV Shows, Spanish-Language TV Shows, TV Comedies

7 rows | 4-4 of 4 columns

Description: df [7 × 4]

description <chr>
Despite their fathers' rivalry, two university students form a friendship at a boxing gym as they tackle family drama, romanc...
Driven and arrogant, film director Michael Curtiz deals with studio politics and family drama during the troubled production...
Angsty Syd navigates high school awkwardness, family drama and an unrequited crush on her best friend while trying to rei...
The family drama gets downright monstrous as parents Fred and Deloris hunt the supernatural, and twins Viv and Geoff har...
Four sisters deal with family drama and secrets throughout three different time periods, all occurring on Christmas Day.
A multigenerational household navigates daily life, evolving relationships and clashes between old and new mindsets in this...
Aspiring hip-hop star Francisco builds a name for himself as a musical artist while his conscience helps him through family ...

7 rows | 5-5 of 4 columns

Task 2I (5 points): Plot the distribution of TV shows based on the number of seasons using a horizontal bar chart. Group the seasons into the following categories:

Less than 3 seasons

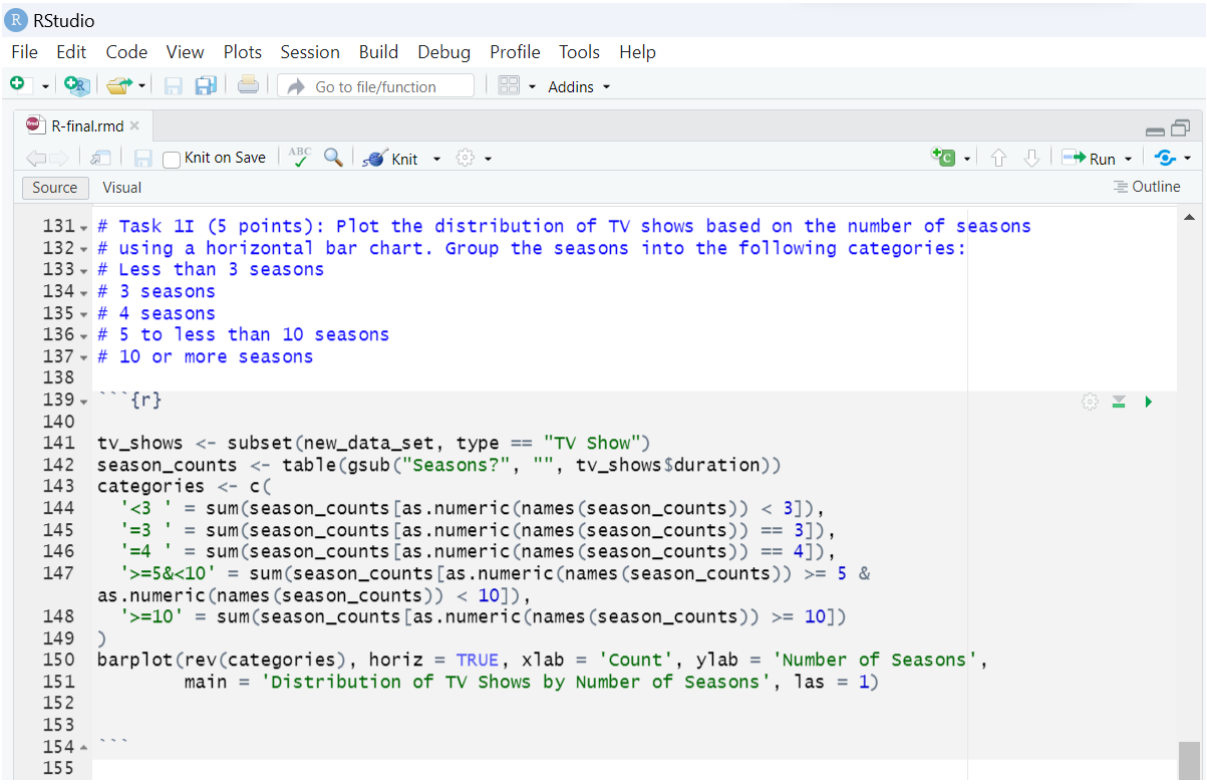
3 seasons

4 seasons

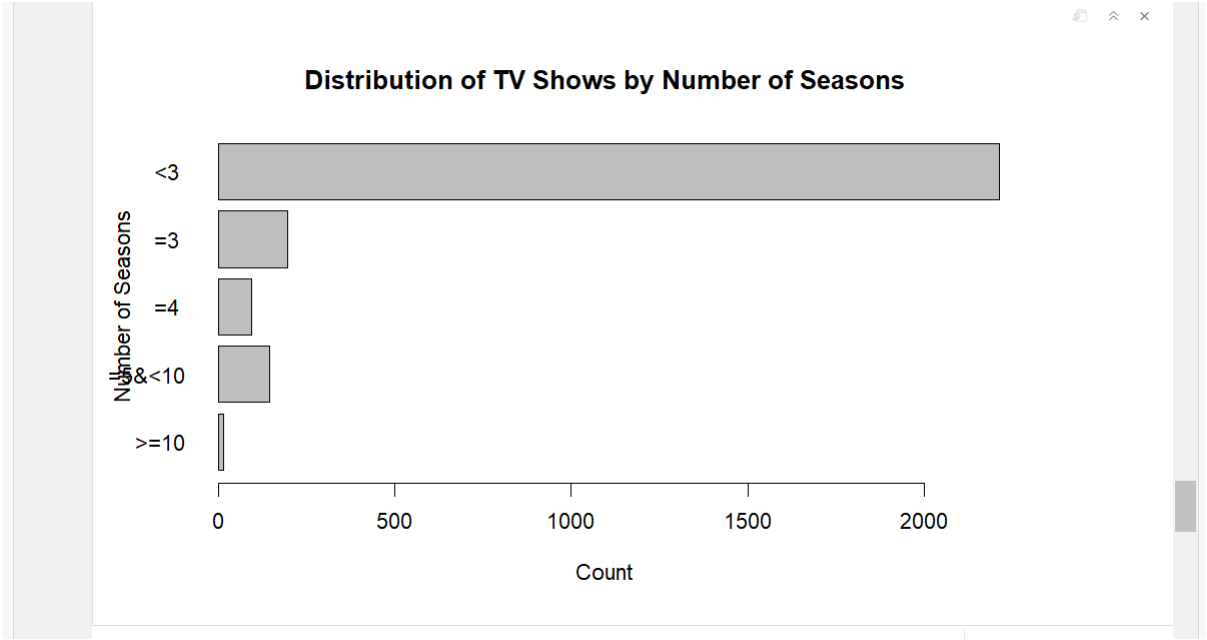
5 to less than 10 seasons

10 or more seasons

Solution



Output


Distribution of TV Shows by Number of Seasons
Number of Seasons
<3
=3
=4
>3&<10
>=10
0
500
1000
1500
2000
Count

Task 2J (6 points): Display a side-by-side pie chart showing the distribution of movie and TV show ratings.

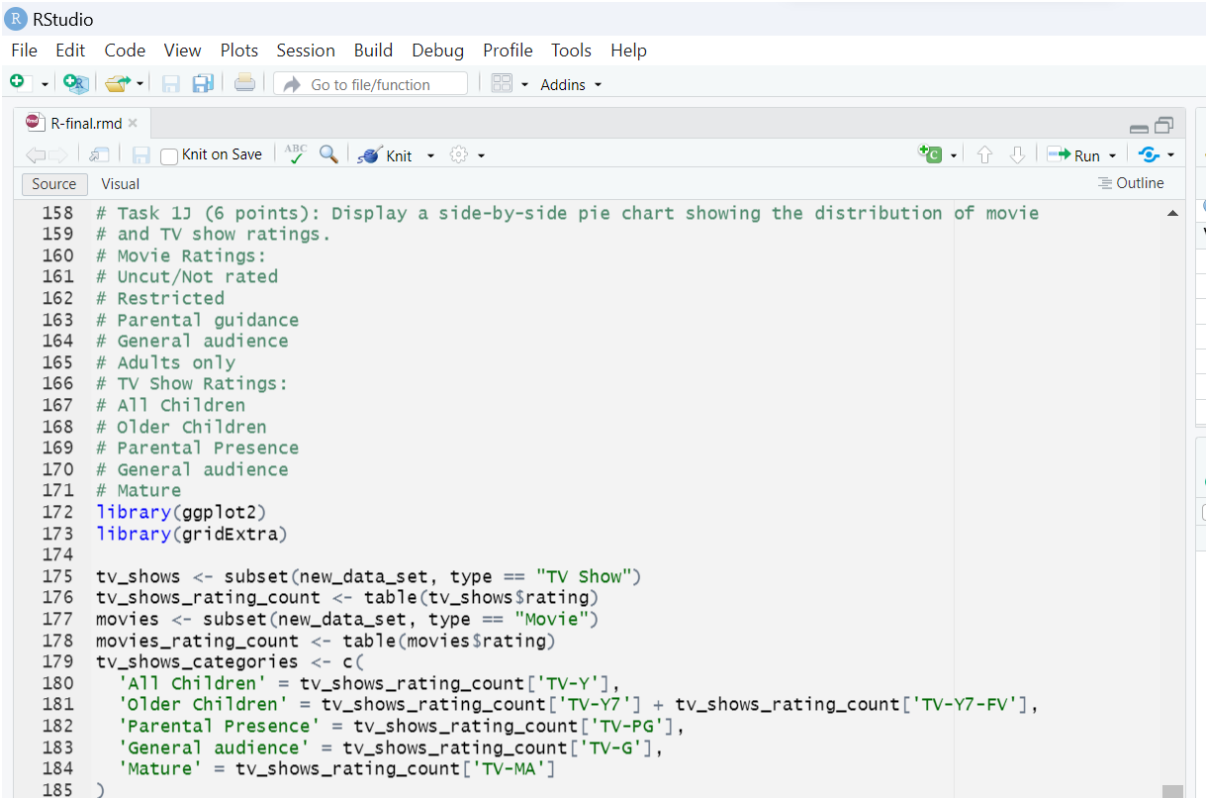
Movie Ratings:

- **Uncut/Not rated**
- **Restricted**
- **Parental guidance**
- **General audience**
- **Adults only**

TV Show Ratings:

- **All Children**
- **Older Children**
- **Parental Presence**
- **General audience**
- **Mature**

Solution

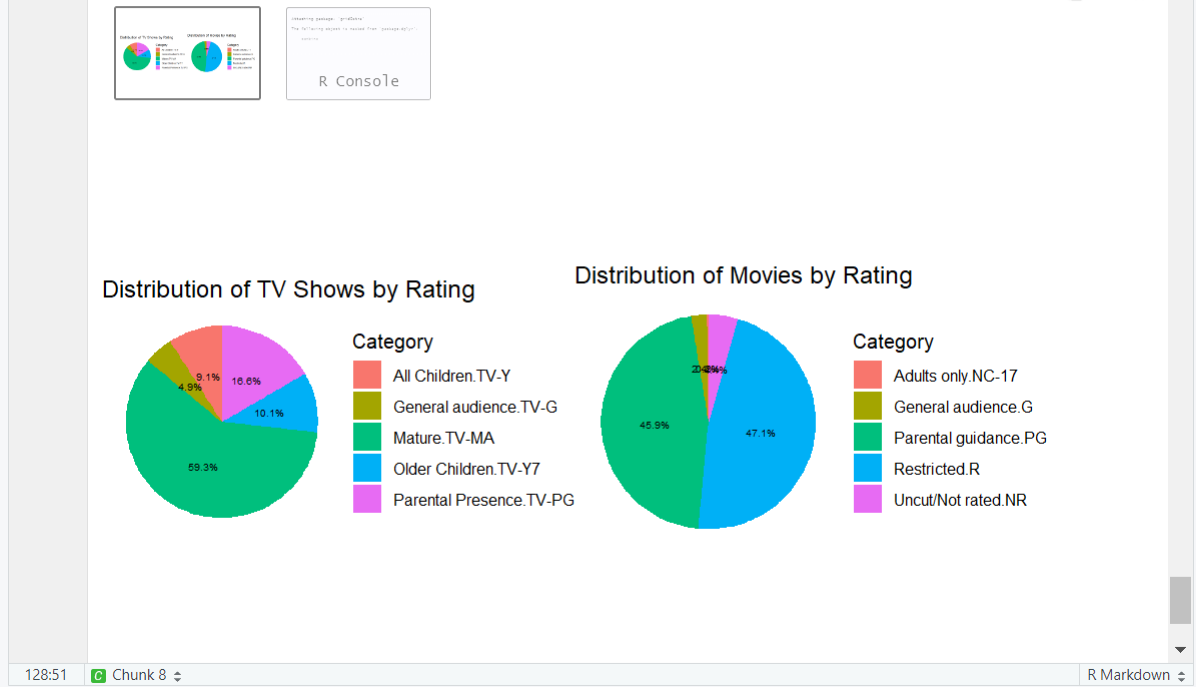


```r
movies_categories <- c(
  'Uncut/Not rated' = movies_rating_count['NR'],
  'Restricted' = movies_rating_count['R'],
  'Parental guidance' = movies_rating_count['PG'] + movies_rating_count['PG-13'],
  'General audience' = movies_rating_count['G'],
  'Adults only' = movies_rating_count['NC-17']
)

tv_shows_data <- data.frame(
  Category = names(tv_shows_categories),
  Count = as.numeric(tv_shows_categories)
)
movies_data <- data.frame(
  Category = names(movies_categories),
  Count = as.numeric(movies_categories)
)

```

```r

tv_shows_data$Percentage <- tv_shows_data$Count / sum(tv_shows_data$Count) * 100
movies_data$Percentage <- movies_data$Count / sum(movies_data$Count) * 100

tv_shows_plot <- ggplot(tv_shows_data, aes(x = "", y = Count, fill = Category)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Distribution of TV Shows by Rating") +
  theme_void() +
  geom_text(aes(label = paste0(sprintf("%.1f", Percentage), "%")),
            position = position_stack(vjust = 0.5),
            size = 2) # Adjust the size here

movies_horizontal_plot <- ggplot(movies_data, aes(x = "", y = Count, fill = Category)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Distribution of Movies by Rating") +
  theme_void() +
  geom_text(aes(label = paste0(sprintf("%.1f", Percentage), "%")),
            position = position_stack(vjust = 0.5),
            size = 2) # Adjust the size here

grid.arrange(tv_shows_plot, movies_horizontal_plot, ncol = 2)


```

Output

```r
```



128:51 Chunk 8 R Markdown

References:

R Language Definition (r-project.org)