**INT232 PROJECT REPORT**

(Project 4th Semester January-May 2024)

**DASHBOARD ON**

**BEST NETFLIX MOVIES AND SHOWS**

Submitted by:

Nilopant Pranav

Registration No. – 12205684

Program and Section – CSE and K22GX

Course Code – INT232 Under the Guidance of

**(Munish Katoch : 29317)**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that **Nilopant Pranav** bearing Registration no. 12205684 has completed INT232 project titled, **“Dashboard on Best Netflix Movies and Shows”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort, and study.

## Signature and Name of the Supervisor Designation of the Supervisor

**School of Computer Science and engineering**

Lovely Professional University Phagwara, Punjab.

Date: 25/04/2024

**DECLARATION**

I, **Nilopant Pranav**, student of Bachelor of Technology under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 25/04/2024 Signature

Registration No. - 12205684 Nilopant Pranav

**Acknowledgment**

I would like to thank Mr. **Munish Katoch Sir** for his vital cooperation and help in ensuring the successful completion of my project. His valuable guidance has been the one that helped me patch this project and make it full proof success.

Secondly, I'd like to thank Lovely Professional University for giving me this golden opportunity to learn many new things, to learn another aspect of life.

Finally, I want to convey my sincere thanks to my family and friends for their constant support and reassurance.

Nilopant Pranav 12205684

# Table of Content

|  |  |
| --- | --- |
| **Title** | **Page** |
| Cover Page | 1 |
| Declaration | 2 |
| Certificate | 3 |
| Acknowledgment | 4 |
| Table of Contents | 5 |
| Introduction | 6 |
| Objectives | 7 |
| Existing System | 8 |
| Source of Dataset | 9 |
| Packages or Library used | 10-11 |
| Analysis on Dataset | 12-18 |
| Final Dashboard | 19-20 |
| Future Scope and Summary | 21-22 |
| References and Bibliography | 23 |

**INTRODUCTION**

Welcome to our project report on the top Netflix movies and shows, presented via the lens of a R Flex dashboard. In this digital age, where entertainment is increasingly consumed via online streaming platforms, Netflix stands out as a major player, providing a vast selection of content to suit a wide range of tastes and inclinations. With this project, we hope to give a selected selection of the best Netflix services while employing R Flex Dashboard's data visualization and analytics capabilities.

Based on the R programming language, which is well known for its adaptability and efficiency in data processing and display, R Flex Dashboard is a dynamic application. It transforms unprocessed data into interactive, aesthetically pleasing dashboards that make complex statistics easy to analyze and comprehend. Using R Flex Dashboard, we have developed an intuitive user interface that makes it easy for customers to browse our carefully selected library of Netflix films and TV series.

Through in-depth data analysis and readily navigable visualization, we hope to deliver insightful information about Netflix's wide content offering that appeals to both casual and dedicated users alike.

You may find a wide selection of TV shows and movies in every genre, as well as intelligent reviews and suggestions. Our R Flex Dashboard-powered project offers plenty to offer everyone, regardless of your taste in movies. There is something for everyone, regardless of whether you are a movie buff hunting for obscure treasures or a casual watcher looking for your next series to binge on.

We cordially invite you to accompany us as we use R Flex Dashboard's insightful tools and visualizations to look more into the fascinating world of Netflix entertainment.

# OBJECTIVE

1. List the top Netflix movies and TV shows across a variety of genres, giving equal weight to both well-known and lesser-known titles.

1. Create interactive, aesthetically pleasing dashboards that make it easy to discover and navigate Netflix content by utilizing the showing data features of R Flex Dashboard.
2. Do comprehensive data research to provide insightful information on user ratings, trends, and preferences across Netflix's extensive content library.
3. To assist users in discovering fresh and engaging information, offer perceptive assessments and recommendations based on their preferences and areas of interest.
4. Make sure Netflix has an easy-to-use interface that will appeal to both die-hard viewers and casual viewers, making it engaging and available to a wide range of Netflix users.

# Drawbacks or limitations of an existing system

# “A language and environment for statistical computing and graphics.” It’s an open-source programming language often used as a data analysis and statistical software tool.

* + It's a challenging language. R requires a lot of learning. The ideal candidates for this language are those with prior programming experience.
  + It's not as safe. R lacks fundamental security features. As such, it's not a recommended option for developing web-safe applications. R cannot be included into web browsers either.
  + It moves slowly. R is not as fast as Python or MATLAB, among other programming languages.
  + It consumes a large amount of RAM. One of R's weaknesses is memory management. R’s data must be kept in physical memory. However, the increasing use of cloud-based memory may someday make this limitation moot.
  + The quality of the documentation and packages is inconsistent. Packages and documentation may be haphazard, inconsistent, or lacking.

It is crucial to remember that many of these restrictions can be overcome by utilising third- party packages or by integrating R Studio with other software tools. Furthermore, for many users, the advantages of R Studio—such as its adaptability and capacity for handling sophisticated statistical analyses—often exceed these drawbacks.

# Source of Data set

The dataset is taken from Kaggle. Kaggle is a platform that hosts a variety of datasets from different domains such as healthcare, finance, sports, and more. The datasets on Kaggle are contributed by users and organizations from all over the world.

To access datasets on Kaggle, you first need to create an account on the platform. Once you have an account, you can search for datasets using the search bar on the Kaggle homepage or browse through the datasets by category.

Some of the popular datasets on Kaggle include the Titanic dataset, the IMDB movie review dataset, the New York City Airbnb dataset, E-commerce sales dataset.

**About:** This dataset contains information on all of the movies and TV shows available on Netflix as of May 2022. In addition to basic information such as title, release year, and runtime, the dataset includes data on the cast and crew, IMDB score and number of votes, genres, production companies, and more. With this data, you can build models to find the best movies and TV shows on Netflix according to your own criteria.

Here are the details of my chosen data set.

* Name: Netflix TV Shows and Movies (2022 Updated)
* Link: https://www.kaggle.com/datasets/thedevastator/the-ultimate-netflix-tv-shows-and-movies-dataset/data
* Format: CSV
* No. of data sets: 2( in 1)
* Number of Rows: 633
* Number of columns: 8
  1. String: 3
  2. Decimal: 1
  3. Number: 4
  4. Other: 0
* Size: 31.3 Kb
* Date Fields:
  1. Index
  2. TITLE
  3. RELEASE\_YEAR
  4. SCORE
  5. NUMBER\_OF\_VOTES
  6. DURATION
  7. MAIN\_GENRE
  8. MAIN\_PRODUCTION

# Packages or Library Used

## Flexdashboard:

You can use R Markdown to create interactive dashboards using the flexdashboard R package. Built on top of the rmarkdown and shiny packages, the package provides a framework for rapidly creating dashboard layouts with a range of widgets, charts, and data tables.

With Flexdashboard, you may create a wide range of dashboard designs, including:

• A dashboard with live, dynamic content.

• An interactive dashboard with material and static charts; • A dashboard with a menu that allows the user to switch between views.

Flexdashboard supports a variety of chart types, including bar charts, line charts, scatter plots, heatmaps, and more. Additionally, tables, text, images, and other HTML elements can be used in your dashboard.

To construct a dashboard with Flexdashboard, you have to write R code inside of a R Markdown file. The R code generates the charts and other visualisations, while the Markdown syntax controls the dashboard's style and layout. When your dashboard is complete, you can use a variety of technologies, like GitHub Pages, Shiny Server, RStudio Connect, and others, to publish it online.

Overall, flexdashboard is a powerful tool for building interactive dashboards in R, even if you have no prior experience with web development or dashboard construction.

## Tidyverse:

A group of R packages for data science with a similar approach to data management and visualization is called the tidyverse. The tidyverse contains the following packages:

* + ggplot2: A package for creating graphics and data visualizations.
  + dplyr: A package for manipulating and transforming data.
  + tidyr: A package for reshaping and tidying data.
  + readr: A package for reading and writing tabular data in plain-text formats.
  + purrr: A package for functional programming tools for working with lists and vectors.
  + tibble: A package for creating and working with tibbles, a modern version of data frames.
  + stringr: A package for working with strings and text data.
  + forcats: A package for working with factors and categorical data.

## Because the tidyverse packages are designed to interface with one another easily and provide a common vocabulary for code, it might be easier to write code that is easy to read and maintain.

## data manipulation. The tidyverse method promotes the usage of functions that generate data frames as output, which makes it easy to link processes together with the pipe operator%>%.

## The tidyverse has become quite popular among data scientists and analysts, partly because of its straightforward syntax and ability to process large amounts of data with minimal code. This will teach you the best R techniques for data analysis and visualization.

## Highcharter:

The R package highcharter offers an interface to the Highcharts JavaScript library, a widely used tool for creating interactive charts and visualisations on the web. With highcharter, you may create a wide variety of charts, such as heatmaps, line charts, bar charts, scatter plots, and more.

The htmlwidgets package, the cornerstone of highcharter, allows R code to generate interactive HTML widgets that may be seen in a number of browser windows, the RStudio Viewer window, and even Shiny web applications.

Among the salient characteristics of highcharter are:

* + Easy-to-use syntax for creating charts and visualizations.
  + Support for a wide range of chart types and styles.
  + Interactive features such as zooming, panning, and hovering.
  + Support for customizing chart elements such as colors, fonts, and labels.
  + The ability to export charts to various file formats, including PNG, PDF, and SVG.

If you're searching for a strong and versatile tool for making web-based visualizations, or if you're already familiar with the Highcharts package, highcharter is an excellent choice for building interactive charts and visualizations in R.

1. **GT:**

With the gt package, anyone can make wonderful-looking tables using the R programming language. The gt philosophy: we can construct a wide variety of useful tables with a cohesive set of table parts. These include the table header, the stub, the column labels and spanner column labels, the table body, and the table footer.

Table data—whether it's a tibble or a data frame—is where it all starts. After that, you choose how to put together your gt table using the components and layout required for the job at hand. Lastly, the table can be produced by exporting to a file using gtsave(), printing it at the terminal, or including it into a R Markdown page. At the moment, gt is compatible with RTF, LaTeX, and HTML output formats.

There are twelve datasets provided by gt: countrypops, sza, gtcars, sp500, pizzaplace, exibble, towny, metro, constants, illness, rx\_adsl, and rx\_addv.

1. **DT:**

The R interface to the JavaScript library DataTables is provided by the DT package in R. HTML sites can display R data objects (matrices or data frames) as tables, and DataTables allows for filtering, pagination, sorting, and many other capabilities in the tables.

1. **ggplot2:**

ggplot2 is a popular data visualization package in the R programming language. It was developed by Hadley Wickham and is based on the principles of the “Grammar of Graphics,” which provides a systematic and structured approach to creating and understanding data visualizations. ggplot2 allows users to create a wide variety of high-quality and customizable statistical graphs, making it a valuable tool for data exploration and presentation.

ggplot2 in R is the latest version of the famous open-source data visualization tool ggplot for the statistical programming language R. The term ggplot2 relates to the package’s name. We use the function ggplot() to produce the plots when using the package. Therefore, ggplot() is the command, and the whole package is called ggplot2. It is a part of the R tidyverse, an ecosystem of packages designed with common APIs.

1. **ggrepel**:

ggrepel in ggplot2 for Non-Overlapping Text Labels

To position non-overlapping text labels on your plots automatically, use the R utility ggrepel in conjunction with ggplot2. Here's a brief summary:

What it performs:

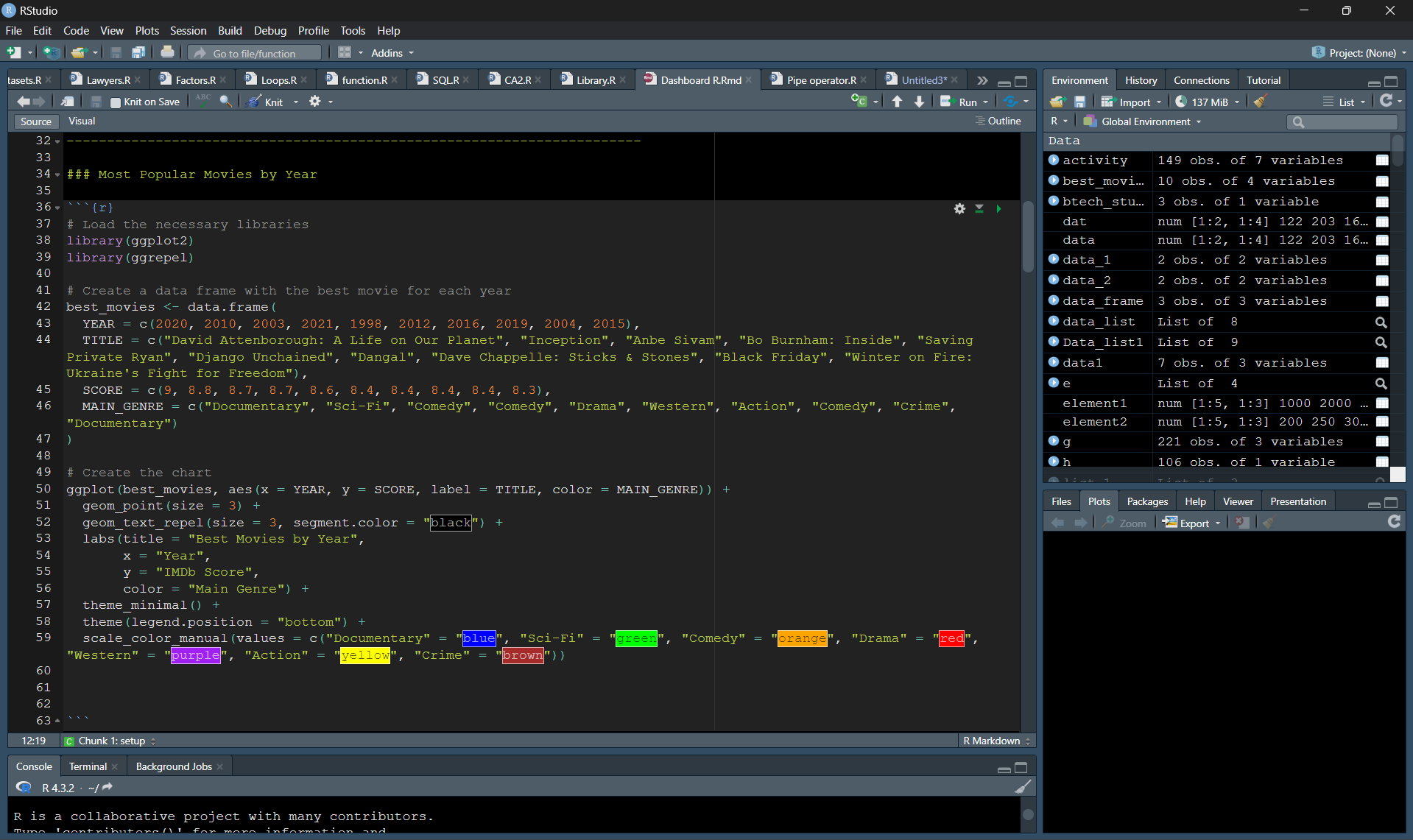
* gives two geoms
* The geom\_text\_repel() function adds text to the plot directly.
* Similar to geom\_text\_repel(), but with a rectangle drawn behind the text for easier reading, is geom\_label\_repel().
* To ensure a clean and uncluttered plot, these geoms automatically exclude overlapping text labels from data points and from each other.

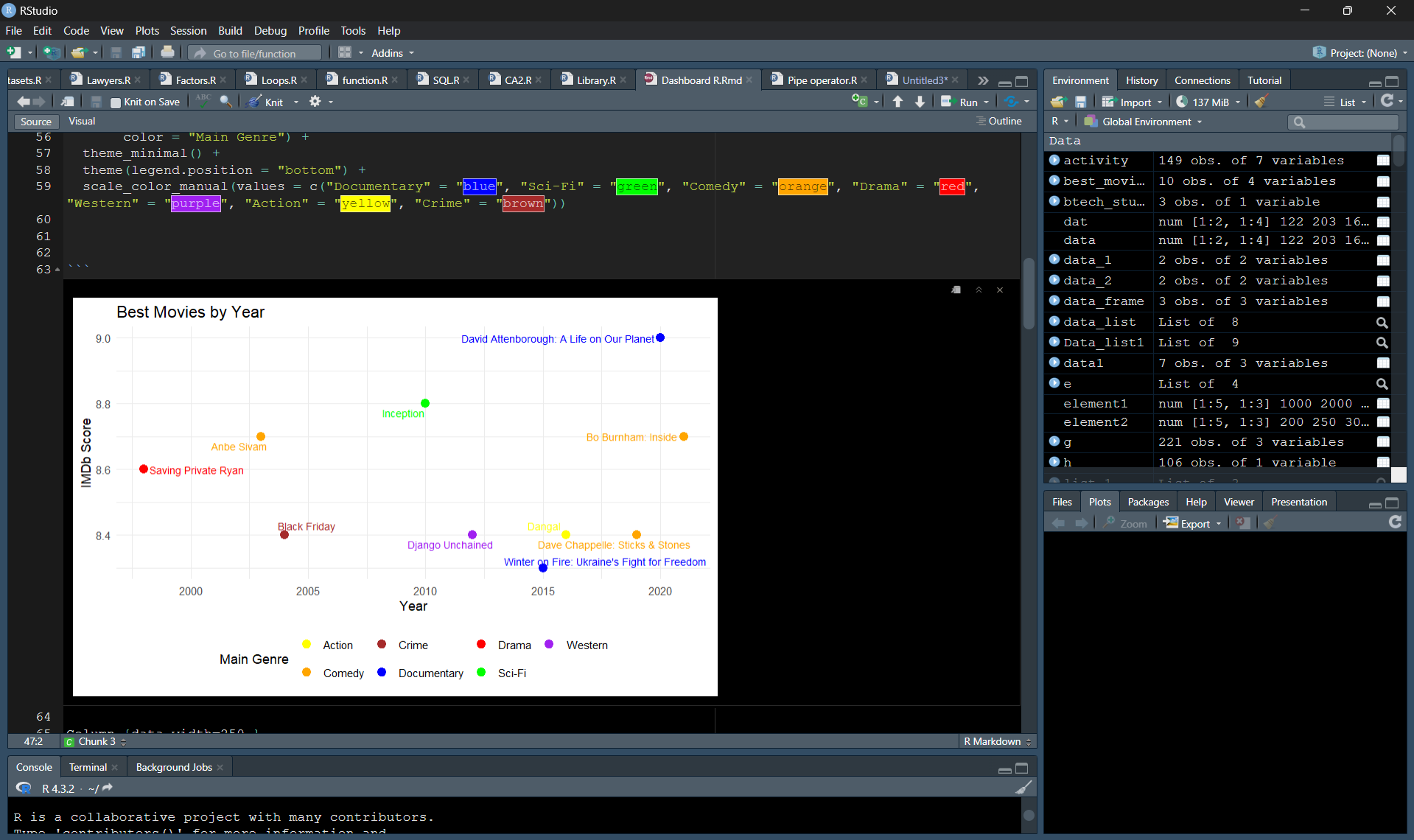
# Analysis of Data set

**Objective 1:** Identify the most popular movies by year on Netflix.

**Description:** This can be done by analyzing the "Best Movies by Year" section of the code, which creates a scatter plot showing the release year of each movie on the x-axis and its IMDb score on the y-axis.

## Specification:

****

****

This code performs the following tasks:

Data Preparation:

* It creates a data frame named best\_movies. This data frame contains information about the "best" movies for each year.
* The data frame has four columns:
* YEAR: The year the movie was released.
* TITLE: The title of the movie.
* SCORE: The IMDb score of the movie (on a scale of 1 to 10).
* MAIN\_GENRE: The main genre of the movie.

Visualization:

* It creates a scatter plot using the ggplot2 library. The plot visualizes the relationship between the year (YEAR) and the IMDb score (SCORE) for each movie.
* Here's a breakdown of the code used for visualization:
* ggplot(best\_movies, aes(x = YEAR, y = SCORE, label = TITLE, color = MAIN\_GENRE)): This line defines the basic structure of the plot. It specifies the data frame to be used (best\_movies) and the aesthetic mappings for the x-axis (YEAR), y-axis (SCORE), labels (TITLE), and color (MAIN\_GENRE).
* geom\_point(size = 3): This line adds points to the plot, representing each movie. The size argument sets the size of the points.
* geom\_text\_repel(size = 3, segment.color = "black"): This line adds labels to the data points, showing the movie titles (TITLE). The geom\_text\_repel function helps prevent overlapping labels.
* labs(title = "Best Movies by Year", ...): This line sets the title, labels for the x and y-axis, and the legend title.
* theme\_minimal() + theme(legend.position = "bottom"): These lines set the overall theme of the plot to a minimalist style and position the legend at the bottom.
* scale\_color\_manual(values = ...): This line defines the colors used for each genre in the plot.

**Objective 2:** Identify the most popular shows and movies on Netflix by score.

**Description:** This code creates a bar chart visualizing the IMDb scores of the most popular Netflix shows and movies, sorted by descending score.

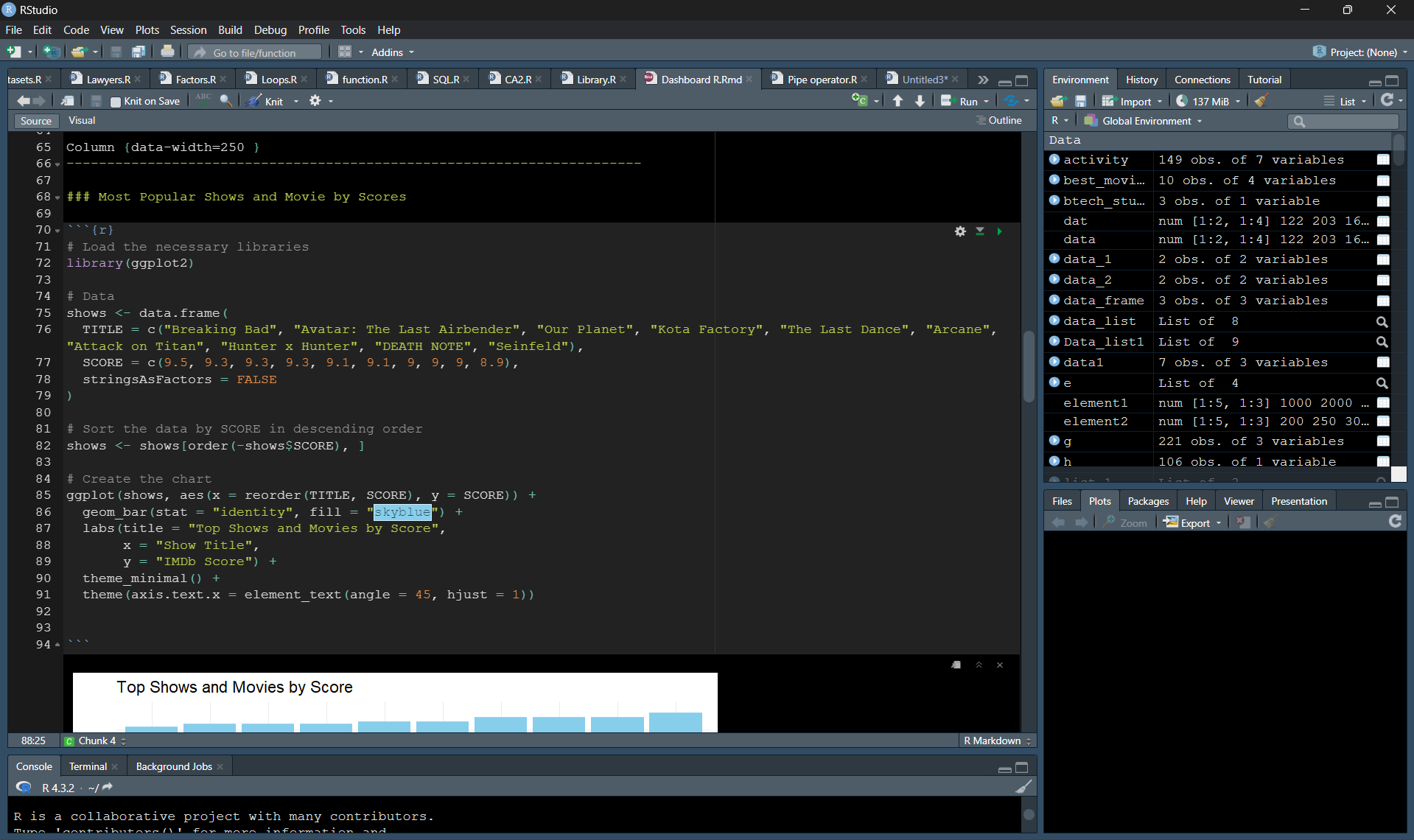
**Specification:**

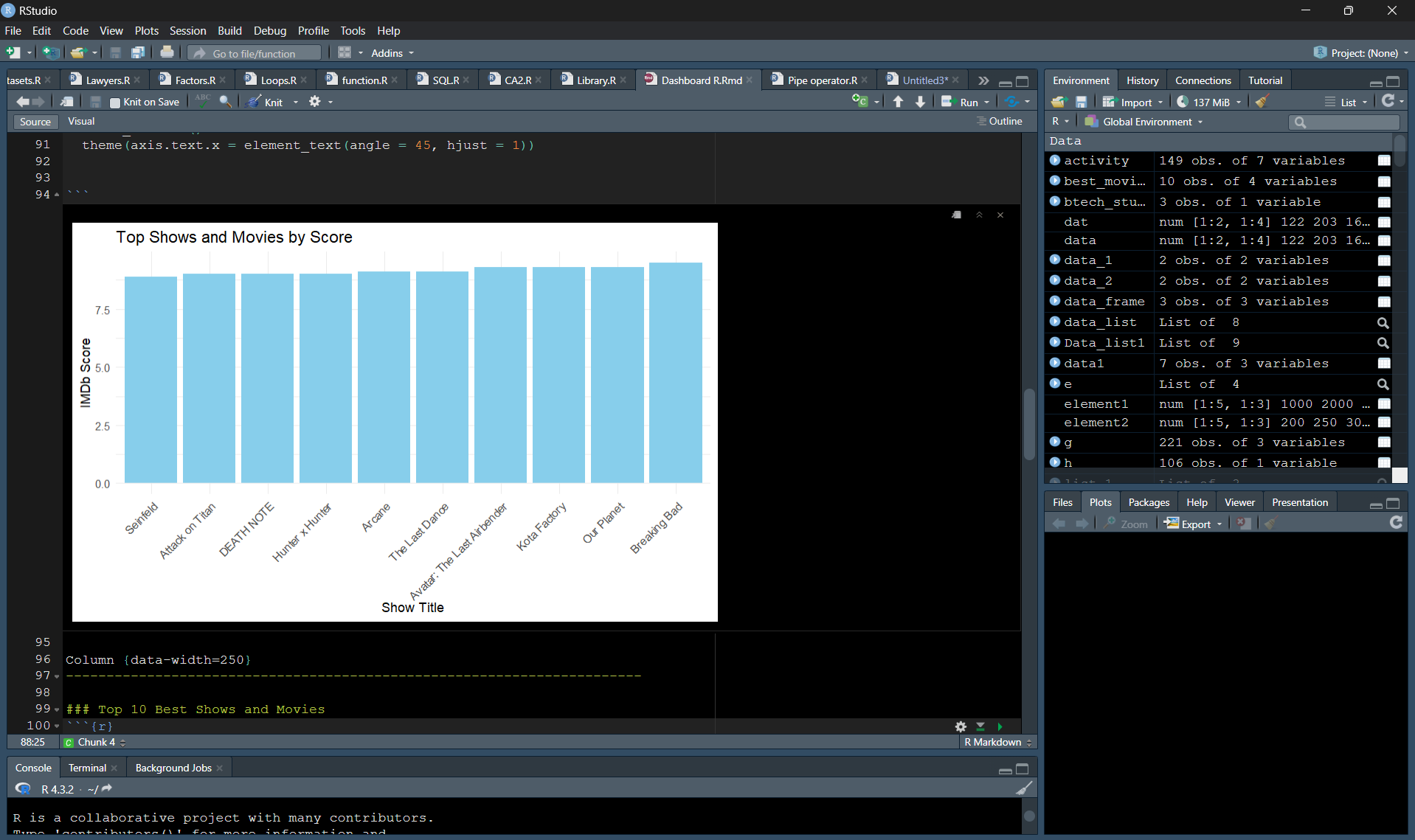
1. Data Preparation and Sorting:

* It defines a data frame named shows with columns for Title (TITLE) and Score (SCORE) of various shows and movies.
* It sorts the data frame shows by the SCORE column in descending order. This ensures the shows with the highest scores appear first.

2. Creating a Visualization:

* It uses the ggplot2 library to create a bar chart.
* The chart displays the show titles (TITLE) on the x-axis and their corresponding scores (SCORE) on the y-axis.
* The titles are reordered on the x-axis based on the descending score order, ensuring the highest-scoring titles appear to the right.
* The chart is customized with a title, labels for the axes, a sky blue fill color for the bars, and a minimalist theme.
* It rotates the x-axis text labels by 45 degrees for better readability when dealing with many titles.





**Objective 3:** Identify the top 10 best shows and movies on Netflix.

**Description:** Identifies the top 10 shows and movies on Netflix based on score.

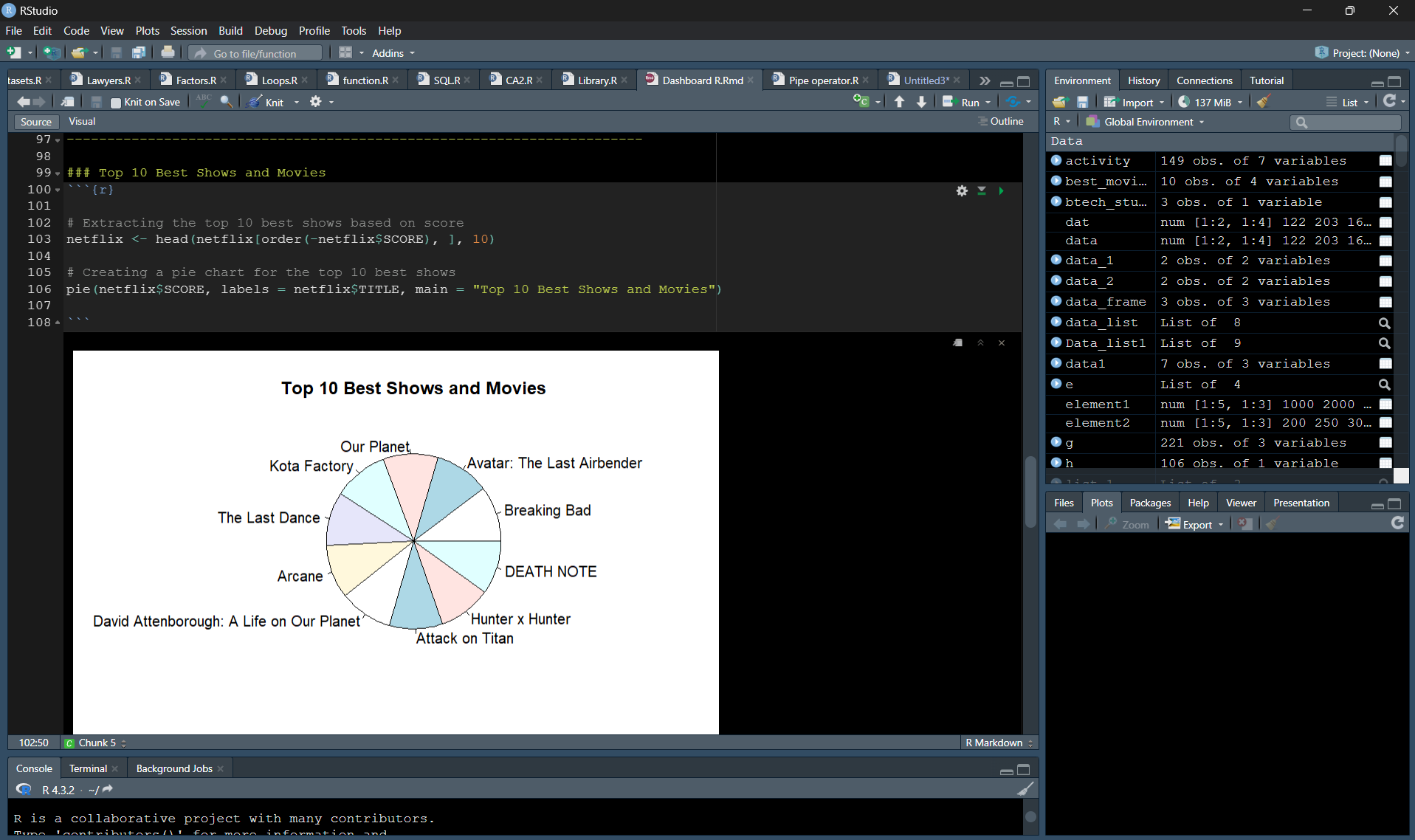
## Specification: This code performs the following tasks:

1. Identifying Top Shows

1. Sorting: The code first uses the order function to sort the netflix data frame based on the SCORE column in descending order (-netflix$SCORE). This ensures the shows with the highest scores appear first.
2. Head Selection: The head function then extracts the top 10 entries from the sorted data frame. This creates a new data frame, netflix, containing only the information for the top 10 shows (or movies) based on their scores.

2. Visualizing Scores with a Pie Chart

1. Data Selection: The code uses the pie function from likely a plotting library (not shown) to create a visualization. It selects two pieces of data from the netflix data frame:
2. Values: netflix$SCORE - This provides the numerical scores for each entry, which will determine the size of the pie slices.
3. Labels: netflix$TITLE - This provides the titles of the shows (or movies) which will be used as labels for each slice.
4. Pie Chart Creation: The pie function creates a pie chart where each slice represents a show (or movie), and the size of the slice corresponds to its score. The main argument sets the title of the pie chart to "Top 10 Best Shows and Movies".



**Objective 4:** Create a table of Netflix movies.

**Description:** Display a table of Netflix movies from a CSV file.

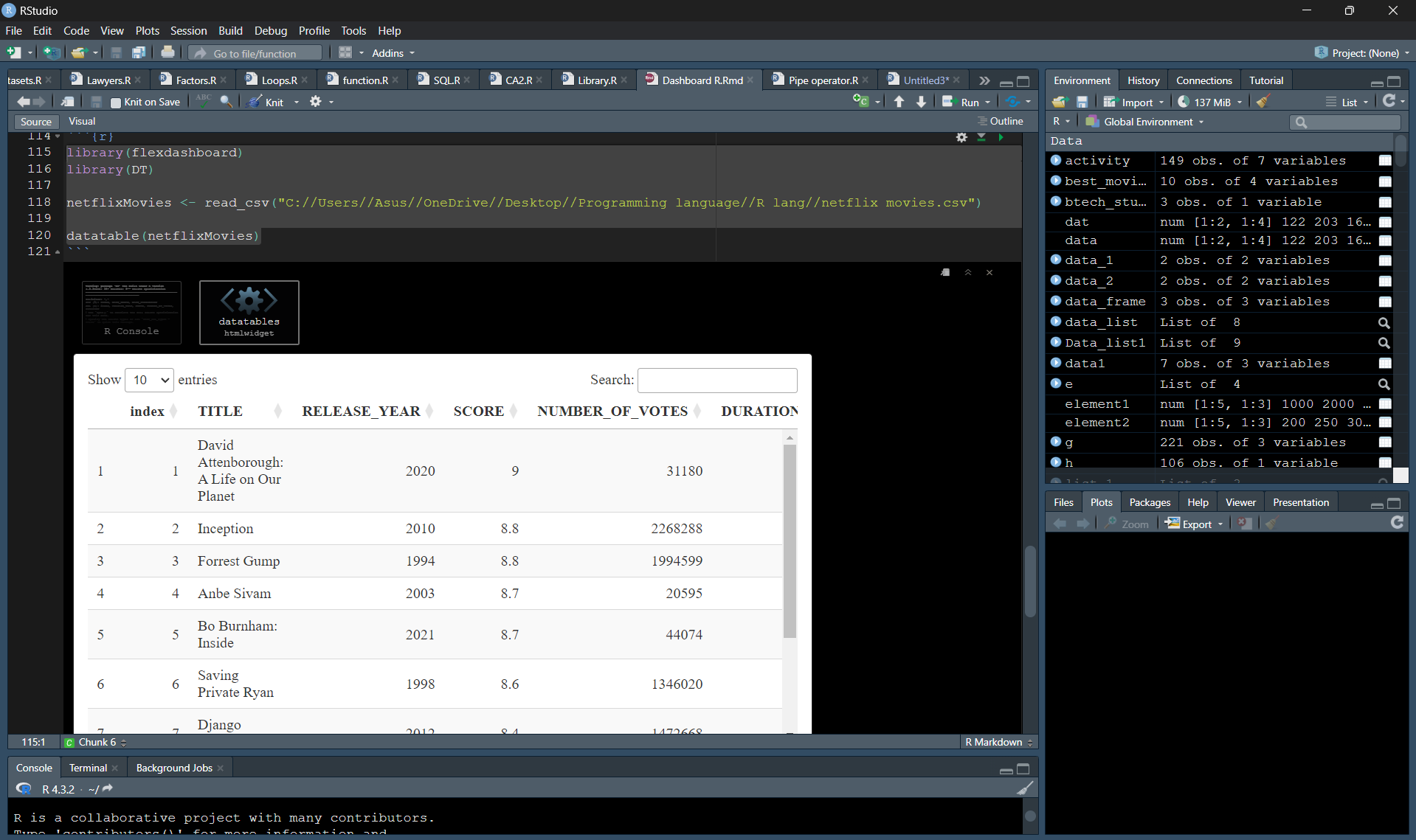
**Specification:** This code performs the following tasks:

1. Identifying Top Shows

* Sorting: The code first uses the order function to sort the netflix data frame based on the SCORE column in descending order (-netflix$SCORE). This ensures the shows with the highest scores appear first.
* Head Selection: The head function then extracts the top 10 entries from the sorted data frame. This creates a new data frame, netflix, containing only the information for the top 10 shows (or movies) based on their scores.

2. Visualizing Scores with a Pie Chart

* Data Selection: The code uses the pie function from likely a plotting library (not shown) to create a visualization. It selects two pieces of data from the netflix data frame:
* Values: netflix$SCORE - This provides the numerical scores for each entry, which will determine the size of the pie slices.
* Labels: netflix$TITLE - This provides the titles of the shows (or movies) which will be used as labels for each slice.
* Pie Chart Creation: The pie function creates a pie chart where each slice represents a show (or movie), and the size of the slice corresponds to its score. The main argument sets the title of the pie chart to "Top 10 Best Shows and Movies".



**Objective 5:** Create a table of Netflix shows.

**Description:** The objective is to check the regions which have total number of confirmed cases.

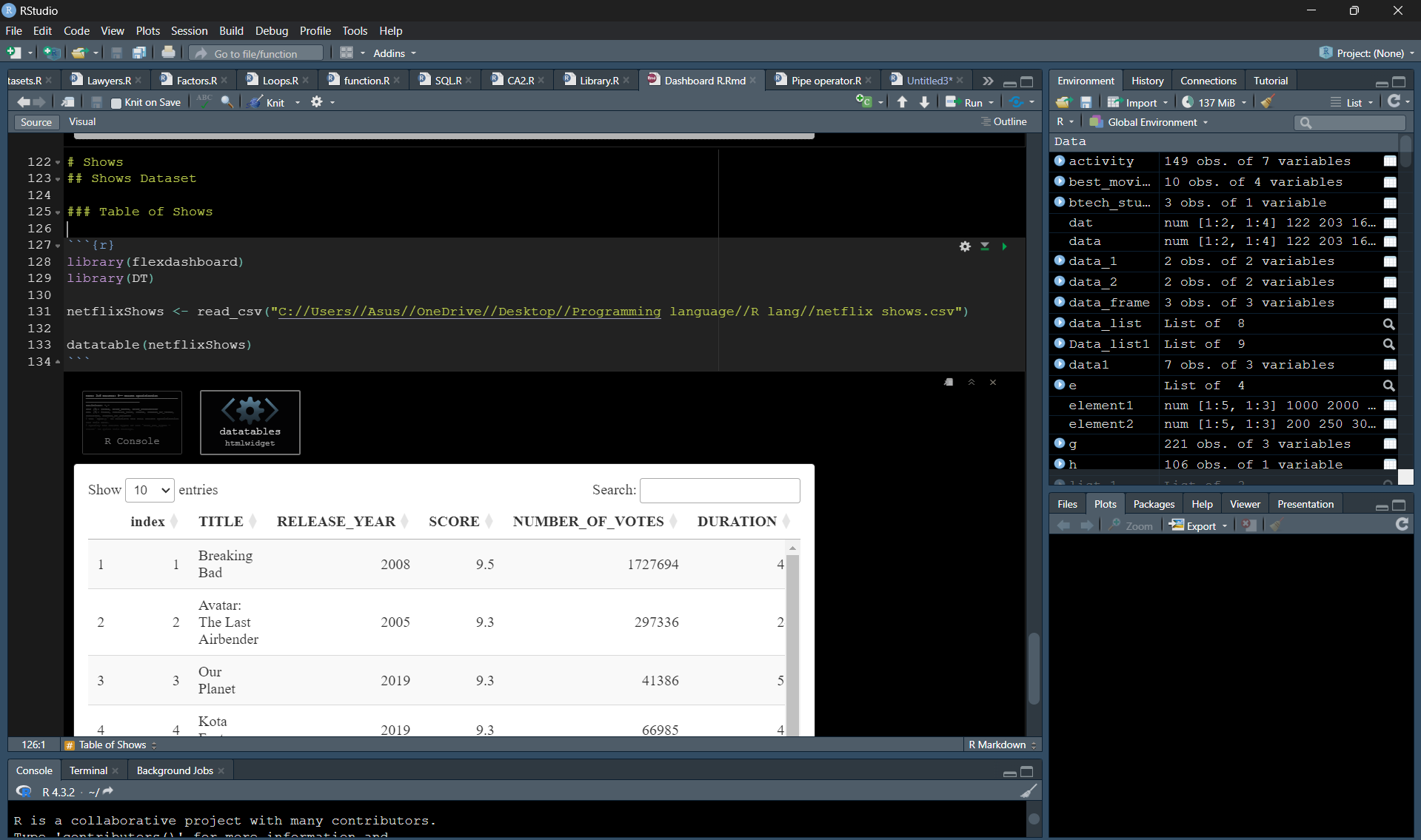
**Specification:**

1.Reading Data:

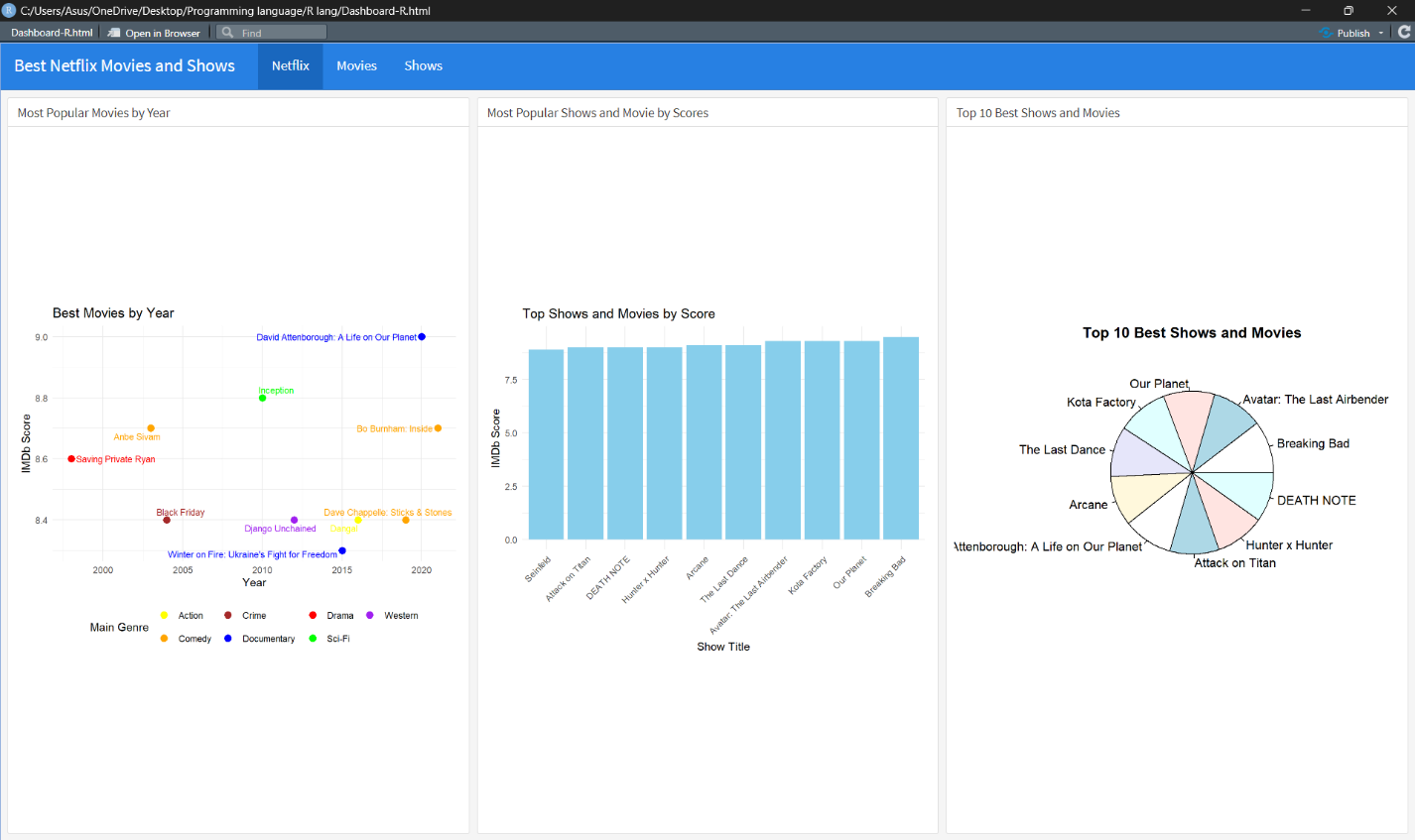
* netflixShows <- read\_csv("C://Users//Asus//OneDrive//Desktop//Programming language//R lang//netflix shows.csv"): This line reads data from a CSV file named "netflix shows.csv" located on a specific path on your computer.
* The read\_csv function from the base R package is used for this purpose. It assumes the file is comma-separated and parses its contents into an R data frame named netflixShows.
* This data frame likely contains information about various Netflix shows, such as titles, genres, release dates, etc. (The exact columns would depend on the structure of your CSV file.)

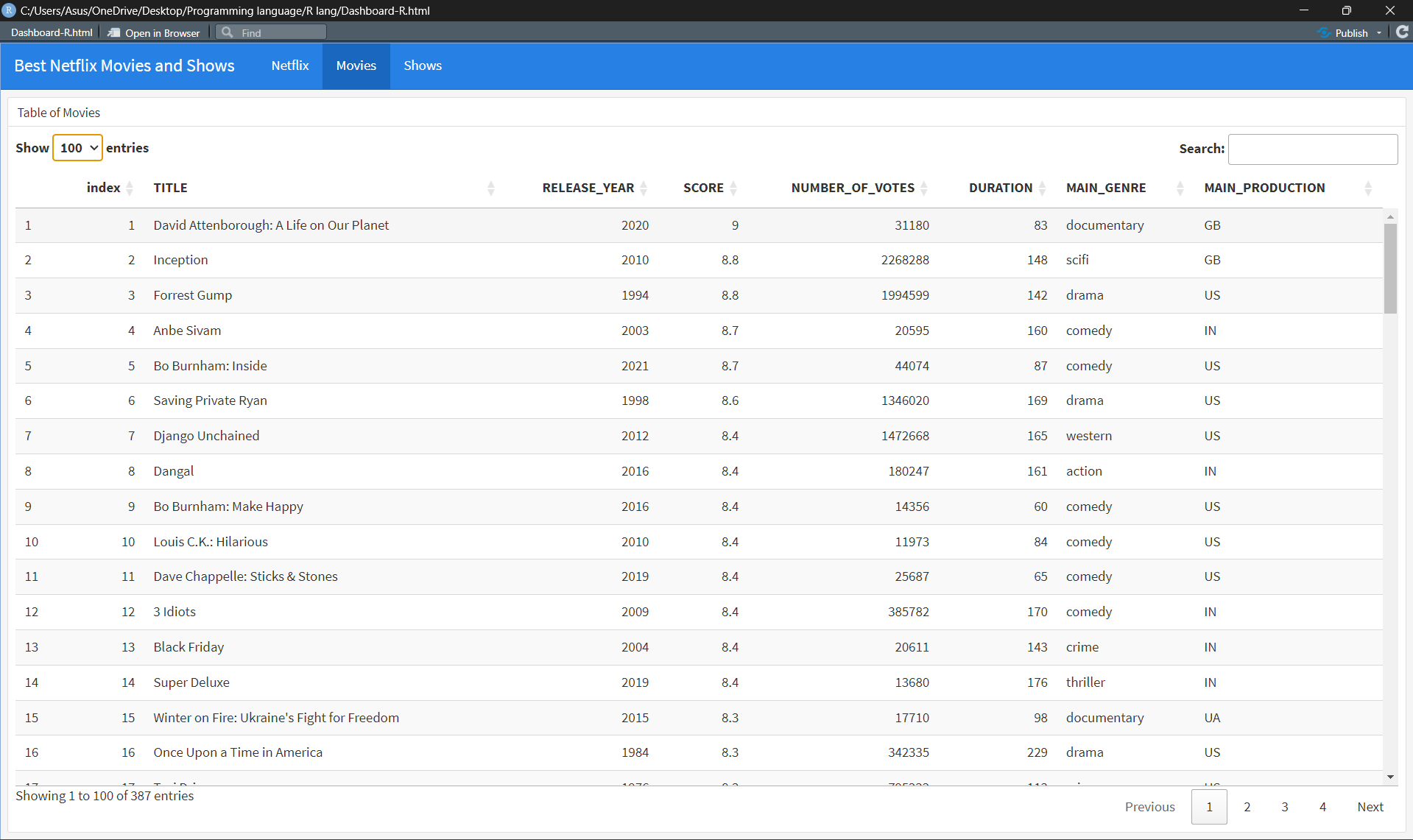
2. Creating a Data Table:

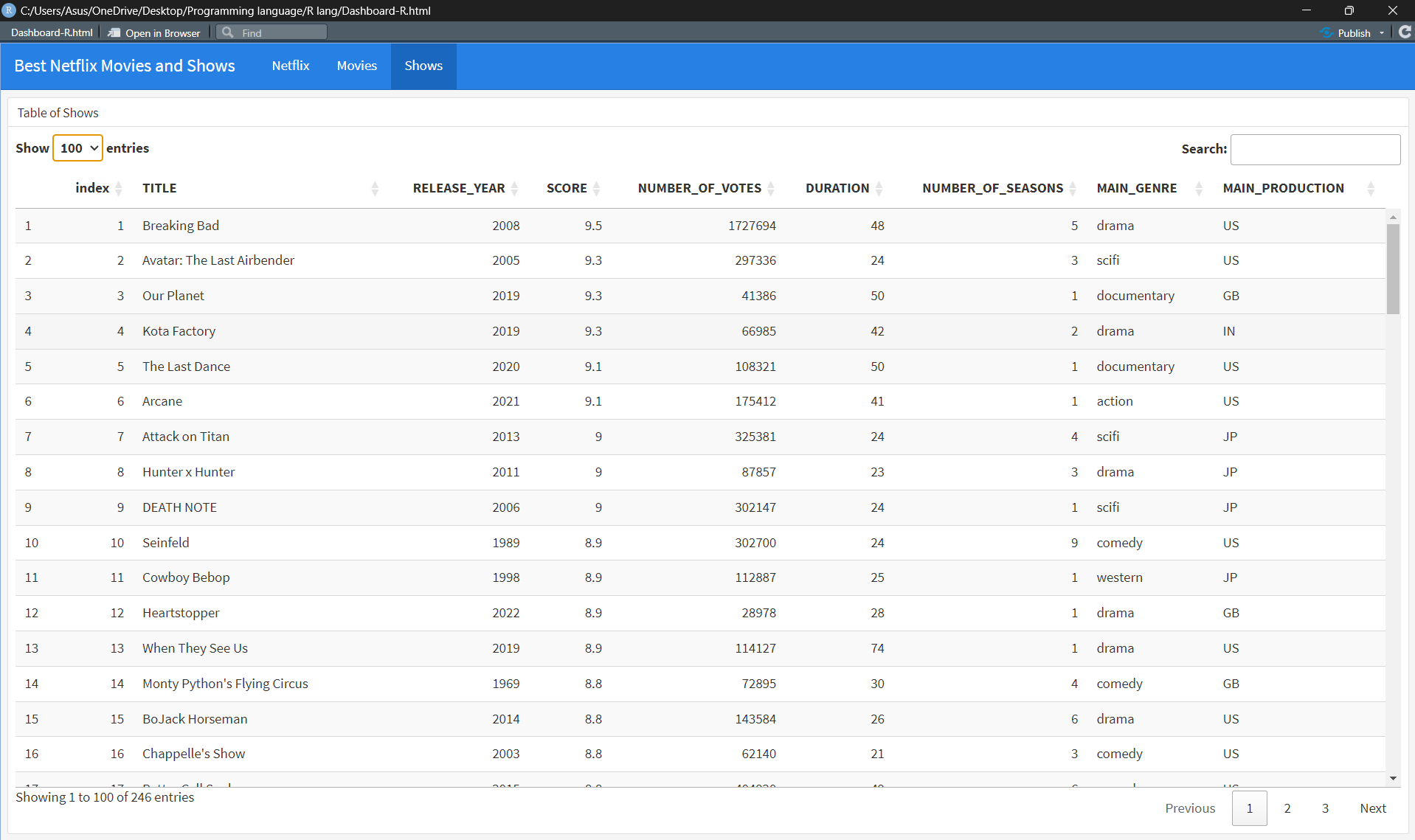
* datatable(netflixShows): This line creates a DataTables object from the netflixShows data frame using the datatable function from the DT package.
* This function essentially transforms the data frame into an interactive table that can be displayed in the R environment.
* You can interact with this table by clicking on column headers to sort, typing in the search bar to filter, or using the pagination controls to navigate through large datasets.



**Final Dashboard**







# Future scope and Summary

A excellent place to start when exploring Netflix content is using this dashboard! Here are some suggestions for enlarging its future reach:

* User Interaction: Give consumers the option to filter data according to release year, genre, or other specifications.
* Data Integration: For a more comprehensive comparison, incorporate data from streaming services outside Netflix.
* Create a recommendation engine by taking into account the viewing history and preferences of the user.
* Dynamic Data Updates: To receive automatic updates on popular content, link the dashboard to real-time data sources.
* Advanced Visualizations: To uncover more nuanced information, including interactive charts such as scatter plots or heatmaps.
* User Reviews: For a more complete picture, include user reviews in addition to scores.
* Content Categorization: Make genre categories more precise (e.g., romantic comedy vs. dark comedy, subgenres).

Using this R Markdown code, a Flexdashboard for Netflix content exploration is created. It shows the following information:

* The most popular films by year are displayed in a scatter plot along with their genre and score for each year.
* Most Popular Movies and Series by Scores: An IMDb score-based bar chart displaying the best films and series.
* Top 10 Greatest TV Series and Films: A pie chart showcasing the top 10 selections according to rating.
* Television Series & Films Datasets: Individual DataTables with comprehensive details on Netflix films and series.
* Although it provides a quick overview of Netflix's most popular content, this dashboard has the potential to develop into a more thorough and interactive resource for discovering other streaming options.

# References

1. https://www.kaggle.com/datasets/thedevastator/the-ultimate-netflix-tv-shows-and-movies-dataset/data
2. https://pkgs.rstudio.com/flexdashboard/
3. https:/[/www.youtube.com/watch?v=](http://www.youtube.com/watch?v=_a4S4tq62OE)\_[a4S4tq62OE](http://www.youtube.com/watch?v=_a4S4tq62OE)
4. https://r-spatial.org/r/2018/10/25/ggplot2-sf.html
5. https:/[/www.r](http://www.r-bloggers.com/2021/09/draw-a-trend-line-using-ggplot-quick-guide/)-[bloggers.com/2021/09/draw-a-trend-line-using-ggplot-quick-guide/](http://www.r-bloggers.com/2021/09/draw-a-trend-line-using-ggplot-quick-guide/)
6. <https://cran.r-project.org/web/packages/ggrepel/vignettes/ggrepel.html>