# INT232 PROJECT REPORT

(Project 4th Semester January-May 2024)

# DASHBOARD ON NETFLIX DATA ANALYSIS

Submitted by

MAKTHALA SAI TEJA GOUD

Registration No. - 12219303

Programme and Section – CSE and K22GB

Course Code – INT232

Under the Guidance of

**(Zeenat Zahra: 30447)**

**Discipline of CSE/IT**

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

# CERTIFICATE

This is to certify that **Makthala Sai Teja Goud**, bearing Registration No. 12001693, has completed the INT232 project titled "**Dashboard on Netflix Data Analysis**" under my guidance and supervision. To the best of my knowledge, the present work is the result of 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: 23rd April 2024

## DECLARATION

I, Makthala Sai Teja Goud, a student of Bachelor of Technology under the 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: 23rd April 2024 Signature

Registration No :12219303. Makthala Sai Teja Goud

# Acknowledgment

I would like to express my heartfelt gratitude to Ms. Zeenat Zahra for her vital cooperation and invaluable assistance in ensuring the successful completion of my project. Her guidance has been instrumental in shaping this project and ensuring its success.

Furthermore, I extend my sincere appreciation to Lovely Professional University for providing me with this invaluable opportunity to expand my knowledge and explore new horizons.

Lastly, I wish to convey my deepest thanks to my family and friends for their unwavering support and encouragement throughout this journey.

Makthala Sai Teja Goud
Registration No.: 12219303

# INTRODUCTION

Welcome to the Netflix Data Analysis R Dashboard project! In the digital age, streaming platforms like Netflix have revolutionized the way we consume entertainment. With a vast library of movies and TV shows, personalized recommendations, and original content production, Netflix has become a household name worldwide.

This project aims to delve into the rich ecosystem of Netflix by harnessing the power of data analysis and visualization using R programming language. By leveraging various data sources, including public datasets, this dashboard provides valuable insights into the Netflix platform's dynamics.

By providing a comprehensive overview of Netflix's data landscape, this project equips stakeholders, including content creators, marketers, and business strategists, with actionable insights to make informed decisions and drive growth in the ever-evolving streaming industry. So, fasten your seatbelts as we embark on a journey through the vast realm of Netflix data!

# OBJECTIVES

**Movie Analysis:** Identify the top 5 highest-rated and lowest-rated movies on Netflix to showcase the range of content quality and viewer preferences.

**Genre-wise Analysis:** Uncover the most popular genres among Netflix viewers to understand audience preferences and content consumption trends.

**Language-wise Analysis:** Examine the linguistic diversity of Netflix content by analyzing the distribution of movies across different languages, highlighting global appeal and localization efforts.

**Date-wise Analysis:** Analyze viewership patterns over time to identify seasonal trends, fluctuations in engagement, and the impact of new releases on audience behavior.

**Year-wise Analysis:** Provide an overview of Netflix content evolution across different years to discern overarching trends and shifts in content production and consumption habits.

# Drawbacks or Limitations of R Studio:

While R Studio is a robust and widely-used program for statistical computing and data analysis, it does come with several limitations and drawbacks that users should be aware of:

**Steep Learning Curve:** For individuals who are new to programming or statistical analysis, R Studio may present a steep learning curve. Getting started with the software can be challenging for beginners.

**Memory Limitations:** R Studio's performance can be constrained by memory limitations, particularly when dealing with large datasets. This may lead to software crashes or significant slowdowns during analysis.

**Limited Graphical Capabilities:** Although R Studio offers some basic graphical features, it may not be as powerful or versatile as other visualization tools like Tableau or PowerBI.

**Limited Big Data Support:** R Studio is not designed for processing and analyzing massive datasets or data streams. Handling large volumes of data may not be its strong suit.

**Limited Debugging Resources:** While R Studio provides some basic debugging tools, they may not be as comprehensive as those offered by other programming environments such as Python or Java.

**Limited Software Integration:** Users who are not familiar with R may encounter difficulties in exchanging data or analyses with other software applications. R Studio's compatibility with other software tools may be limited.

It's important to note that many of these limitations can be addressed by utilizing third-party packages or integrating R Studio with other software tools. Additionally, for many users, the benefits of R Studio—including its flexibility and capability for advanced statistical analyses—often outweigh these drawbacks.

## About Dataset:

This dataset provides a comprehensive overview of content available on the Netflix platform, including information on each title's genre, premiere date, runtime, IMDb score, language, and release year.

**Columns:**

1. **Title:** The title of the movie or TV show.
2. **Genre:** The category or type of content.
3. **Premiere:** The date when the movie or TV show was first released.
4. **Runtime:** The duration of the movie or TV show in minutes.
5. **IMDb Score:** The rating of the movie or TV show on IMDb.
6. **Language:** The primary language of the movie or TV show.
7. **Year:** The year when the movie or TV show was released.

Through this dataset, users can conduct various analyses, including exploring genre trends, evaluating content performance based on IMDb scores, and understanding language preferences among Netflix viewers. Whether for research, analysis, or content recommendation systems, this dataset offers a rich source of data to drive informed decision-making and enhance understanding of the streaming landscape.

**Details:**

Name: Netflix Content Analysis Dataset
Link: https://www.kaggle.com/datasets/yaminh/netflix-dataset-for-analysis
Format: CSV

Size: 2 MB

**Data Fields:**

i. Title
ii. Genre
iii. Language
iv. IMDb Score
v. Premiere
vi. Runtime
vii. Year

# Packages or Library Used

## 1. R Shiny:

R Shiny is a web application framework for R programming language that allows users to create interactive web applications directly from R scripts. With R Shiny, developers can build powerful, customizable, and interactive web applications without needing to learn web development languages such as HTML, CSS, or JavaScript.

At its core, R Shiny operates on the principle of reactive programming, where changes in user inputs automatically trigger updates in the application's outputs. This makes it easy to create dynamic and responsive applications that adapt to user interactions in real-time.

R Shiny applications consist of two main components: a user interface (UI) definition, which controls the layout and appearance of the application, and a server-side logic, which handles data processing, calculations, and responses to user inputs.

One of the key advantages of R Shiny is its seamless integration with R's extensive ecosystem of statistical and data analysis packages. This enables developers to leverage R's powerful capabilities for data manipulation, visualization, and modeling within their web applications.

Overall, R Shiny empowers R users to transform their analyses and models into interactive web-based tools, making it easier to share insights, collaborate with others, and deploy data-driven applications for a wide range of purposes.


## 2. Tidyverse:

The tidyverse is a comprehensive collection of R packages tailored for efficient and effective data science workflows, unified by a shared philosophy of data manipulation and visualization. It simplifies and streamlines data analysis tasks by providing a consistent and cohesive set of tools. The core packages within the tidyverse include:

**ggplot2:** Facilitates the creation of sophisticated graphics and visualizations.
**dplyr:** Enables efficient data manipulation and transformation operations.
**tidyr:** Provides tools for reshaping and tidying up messy data.
**readr:** Offers functions for reading and writing tabular data in various formats.
**purrr:** Supports functional programming paradigms for working with lists and vectors.
**tibble:** Introduces tibbles, a modern and user-friendly version of data frames.
**stringr:** Facilitates manipulation and processing of strings and text data.
**forcats:** Specializes in working with factors and categorical data.

The tidyverse promotes code that is easy to read and maintain due to its integrated nature, where packages seamlessly complement each other. The %>%, or pipe operator, simplifies chaining operations together, aligning with the tidyverse's emphasis on using functions that produce data frames as output.

Thanks to its straightforward syntax and the ability to accomplish complex data tasks with minimal code, the tidyverse has gained immense popularity among data scientists and analysts. It serves as an excellent resource for mastering advanced techniques in data analysis and visualization within the R ecosystem.


# Analysis of Data set

**Objective 1**: **To Identify the Top 5 Highest-Rated and Lowest-Rated Movies on Netflix.**

**Description:**

The objective of this code segment is to analyze the Netflix movie dataset and determine the top 5 highest-rated and lowest-rated movies based on IMDb scores. It involves sorting the movies based on their IMDb scores and selecting the top 5 highest-rated and lowest-rated movies for further analysis.

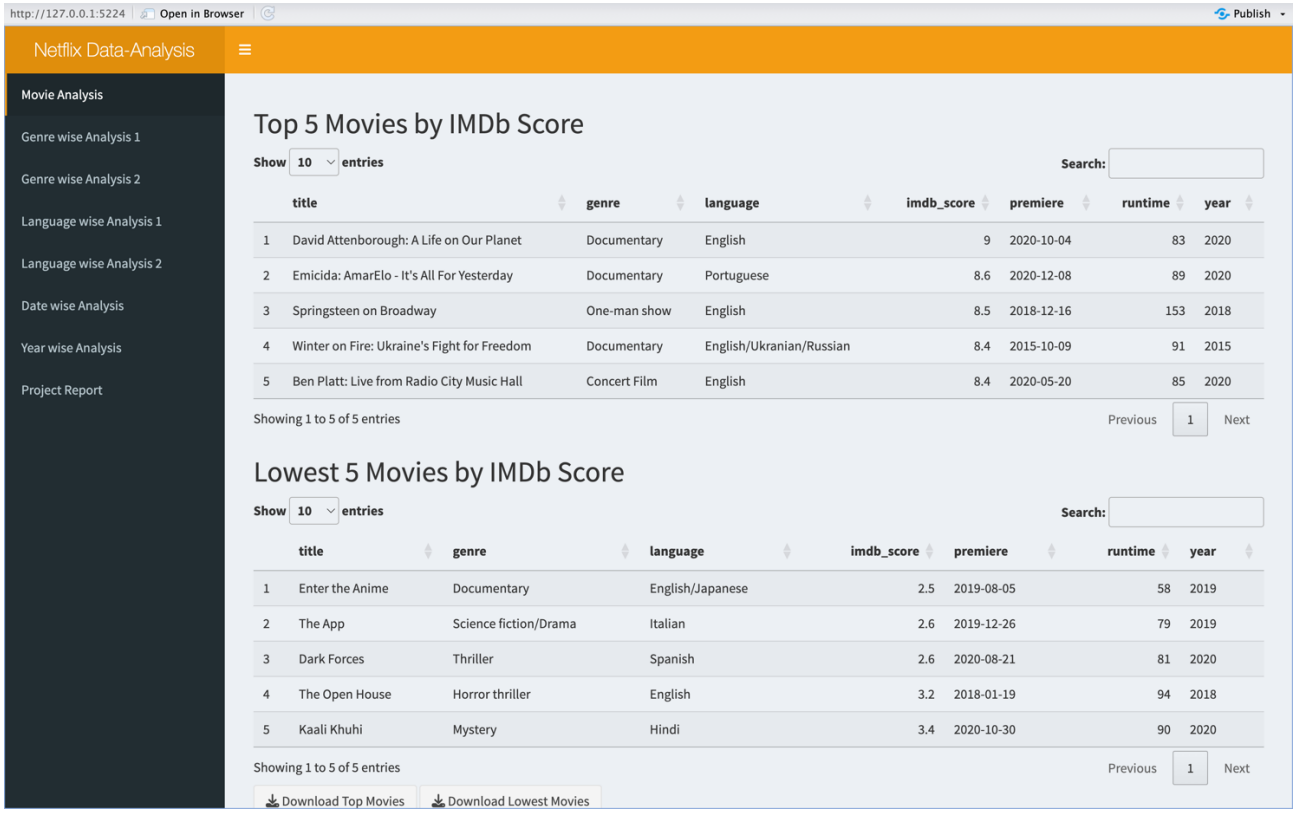
**Specification:**

1. **Rendering Top 5 Movies Table (output$topMovies):**
    a. The output$topMovies function renders a data table displaying the top 5 movies based on their IMDb scores.
    b. It retrieves the Netflix movie data from the movies_data dataset.
    c. The dataset is arranged in descending order of IMDb scores (imdb_score) using arrange(desc(imdb_score)).
    d. The head(5) function selects the top 5 rows from the dataset, representing the highest-rated movies.
    e. This table is dynamically updated and displayed in the Shiny application UI.

2. **Rendering Lowest 5 Movies Table (output$lowestMovies):**

a. Similar to the top movies table, the output$lowestMovies function renders a data table displaying the lowest 5 movies based on their IMDb scores.
b. The dataset is arranged in ascending order of IMDb scores (imdb_score) using arrange(imdb_score).
c. The head(5) function selects the lowest 5 rows from the dataset, representing the lowest-rated movies.
d. This table is also dynamically updated and displayed in the Shiny application UI.

3. **Download Handler for Top Movies (output$downTopMovies):**
    a. It creates a download handler for exporting the data of the top 5 movies as a CSV file.
    b. The createDownloadHandler function defines a handler that generates the data to be downloaded.
    c. It selects the top 5 movies based on IMDb scores and formats them into a CSV file.
    d. The filename for the downloaded file is set as "Top_Movies.csv".
4. **Download Handler for Lowest Movies (output$downLowestMovies):**
    a. Similar to the download handler for top movies, it creates a download handler for exporting the data of the lowest 5 movies as a CSV file.
    b. It selects the lowest 5 movies based on IMDb scores and formats them into a CSV file.
    c. The filename for the downloaded file is set as "Lowest_Movies.csv".




**Objective 2: To Analyze Movie Distribution by Genre and Language.**

**Description:**

The objective of this code segment is to analyze the distribution of movies based on selected genres and languages. It involves rendering a table of movies filtered by a selected genre and creating a histogram depicting the distribution of movies by language within the selected genre.

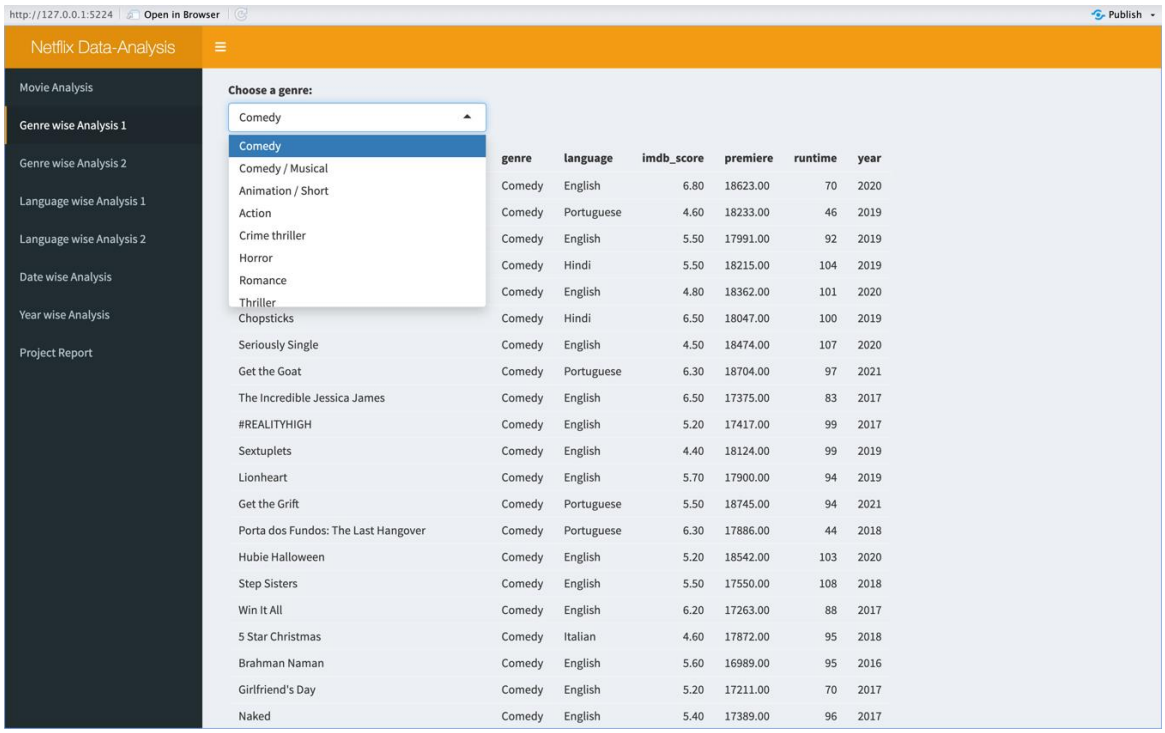
**Specification:**

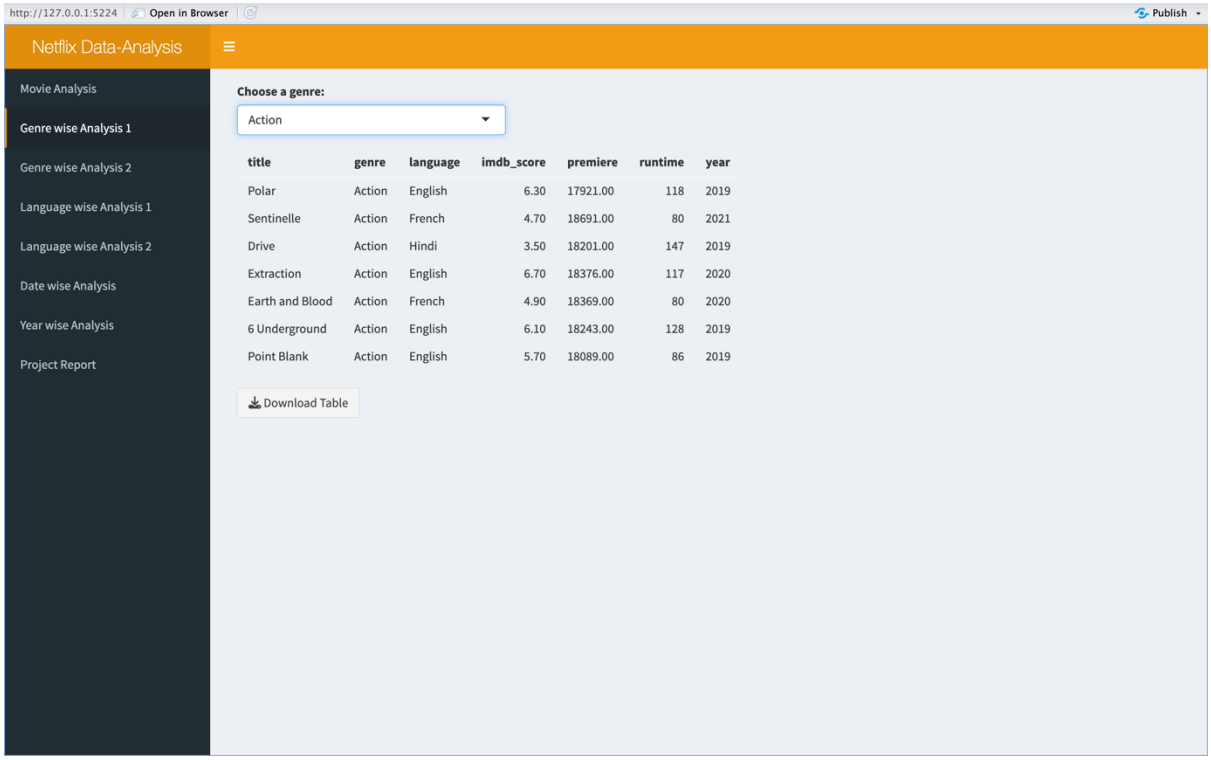
1. **Rendering Table Based on Selected Genre for Genre wise Analysis 1 tab (output$filteredTable1):**
    a. The output$filteredTable1 function renders a table displaying movies filtered by the selected genre input.
    b. It filters the movies_data dataset based on the selected genre input (input$genreInput1).
    c. The renderTable function is used to dynamically update and display the filtered movies table in the Shiny application UI.

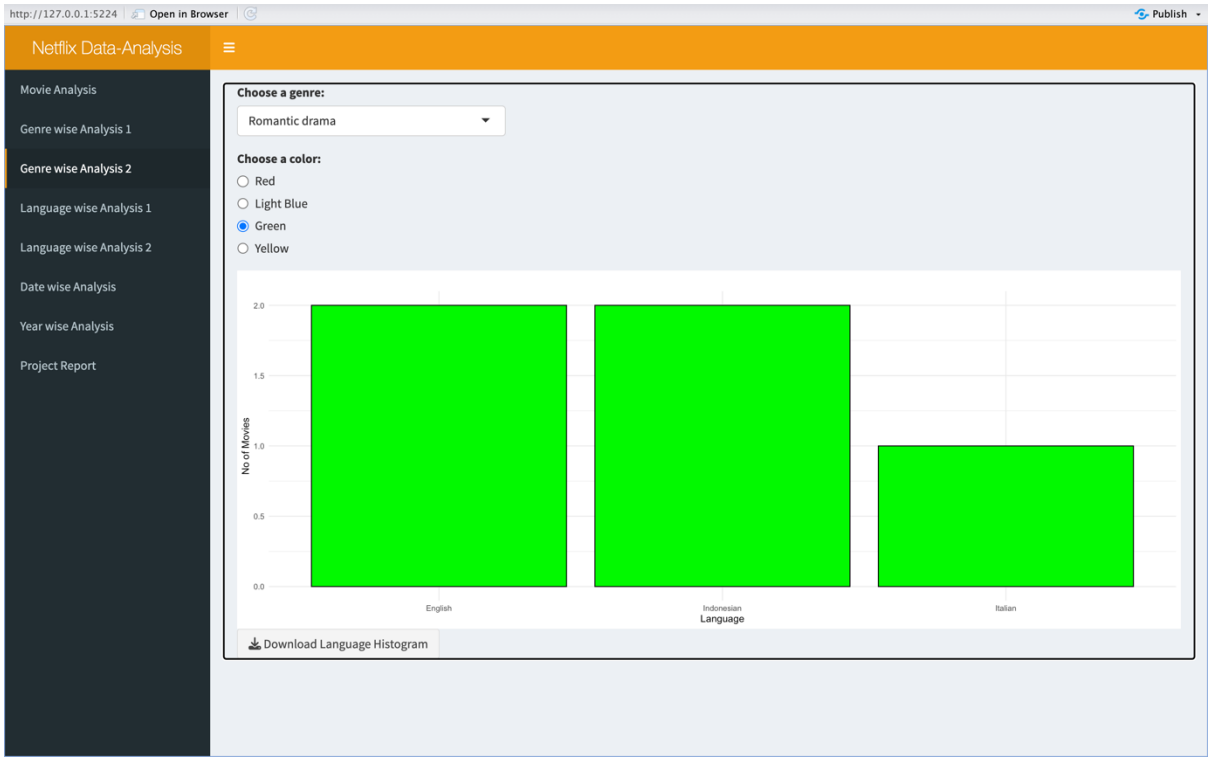
2. **Rendering Language Histogram Based on Selected Genre for Genre wise Analysis 2 tab (output$languageHistogram):**

a. The output$languageHistogram function renders a histogram depicting the distribution of movies by language within the selected genre.
b. It filters the movies_data dataset based on the selected genre input (input$genreInput2).
c. The ggplot function is utilized to create the histogram, mapping the x-axis to the movie languages and the fill color to a selected color input (input$colorInput).
d. The geom_histogram function calculates and displays the count of movies for each language.
e. Additional aesthetic elements such as theme settings, axis labels, and legend adjustments are applied to enhance the plot.
f. This plot is dynamically updated and displayed in the Shiny application UI for Genre wise Analysis 2 tab.

These functionalities enable users to explore the distribution of movies within selected genres and languages, providing insights into the diversity and composition of Netflix's movie catalog.




http://127.0.0.1:5224
Open in Browser
Publish
Netflix Data-Analysis
Movie Analysis
Genre wise Analysis 1
Genre wise Analysis 2
Language wise Analysis 1
Language wise Analysis 2
Date wise Analysis
Year wise Analysis
Project Report
Choose a genre:
Action
title genre language imdb_score premiere runtime year
Polar Action English 6.30 17921.00 118 2019
Sentinelle Action French 4.70 18691.00 80 2021
Drive Action Hindi 3.50 18201.00 147 2019
Extraction Action English 6.70 18376.00 117 2020
Earth and Blood Action French 4.90 18369.00 80 2020
6 Underground Action English 6.10 18243.00 128 2019
Point Blank Action English 5.70 18089.00 86 2019
Download Table


http://127.0.0.1:5224
Open in Browser
Publish
Netflix Data-Analysis
Movie Analysis
Genre wise Analysis 1
Genre wise Analysis 2
Language wise Analysis 1
Language wise Analysis 2
Date wise Analysis
Year wise Analysis
Project Report
Choose a genre:
Romantic drama
Choose a color:
Red
Light Blue
Green
Yellow
No of Movies
2.0
1.5
1.0
0.5
0.0
English
Indonesian
Italian
Language
Download Language Histogram

**Objective 3: To Visualize Movie Distribution by Language Using a Pie Chart.**

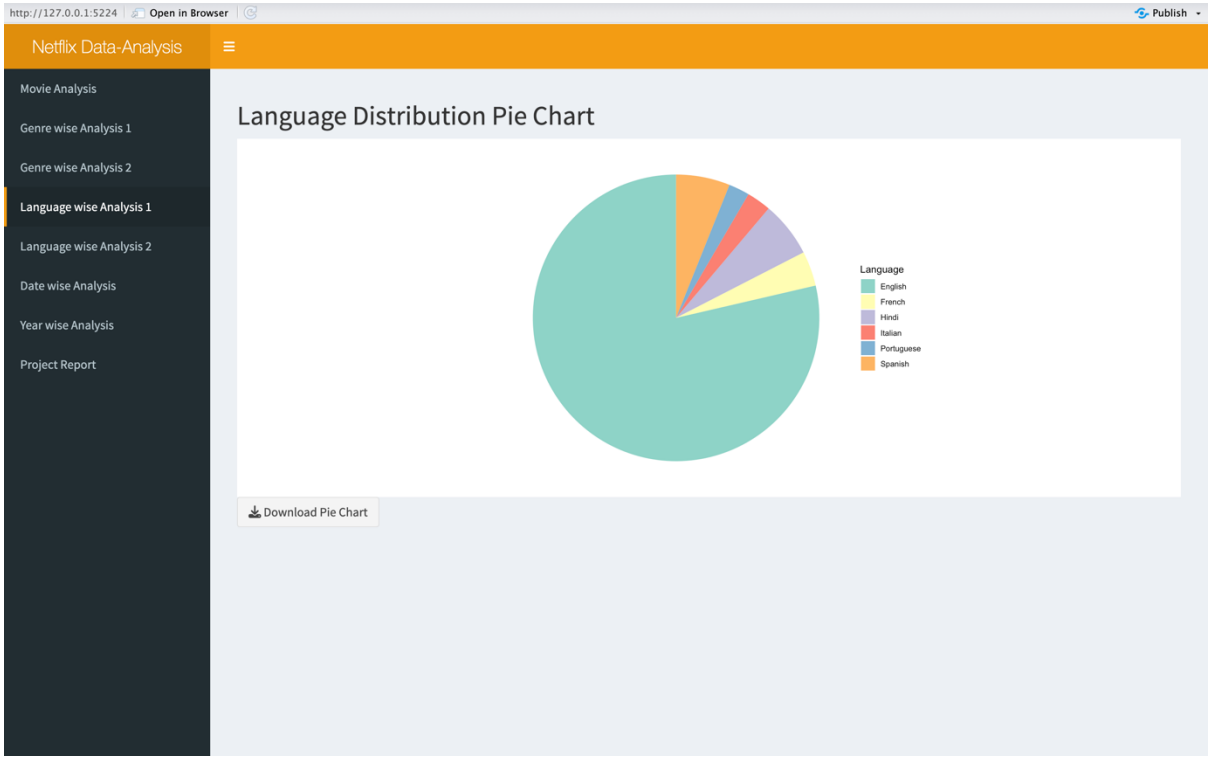
**Description:**

This code segment aims to visualize the distribution of movies by language using a pie chart. It calculates the count of movies for each language, selects the top languages, and creates a pie chart to represent their distribution.

**Specification:**

1. **Rendering Pie Chart (output$pieChart):**
    a. The output$pieChart function renders a pie chart depicting the distribution of movies by language.
    b. It calculates the count of movies for each language using the table function applied to the movies_data$language column.
    c. The resulting language counts are converted into a data frame format (language_data) for further processing.
    d. The top languages are selected by sorting the language data frame in descending order of frequency (Freq) and selecting the top 6 languages using the head function.
    e. The ggplot function is used to create the pie chart, mapping the x-axis to an empty string (""), the y-axis to the frequency of movies (Freq), and the fill color to the language variable (Var1).
    f. The geom_bar function with stat = "identity" parameter creates a bar chart representing the count of movies for each language.
    g. The coord_polar function with "y" parameter sets the polar coordinate system for the pie chart.
    h. Additional aesthetic elements such as theme settings, axis labels, and color palette adjustments are applied to enhance the plot.
    i. This pie chart is dynamically updated and displayed in the Shiny application UI.

This functionality allows users to visually explore the distribution of movies by language, providing insights into the linguistic diversity of Netflix's movie catalog.




**Objective 4: To Display Movie Data Based on Selected Language**.

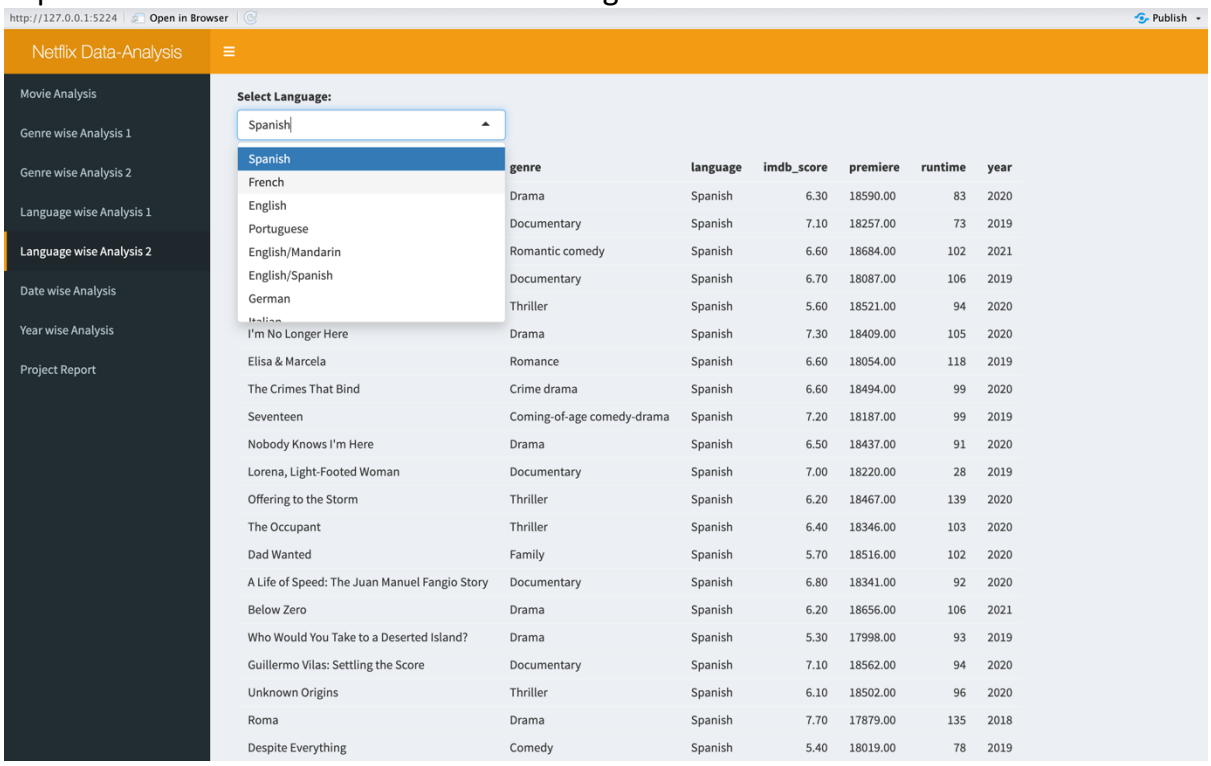
**Description:**

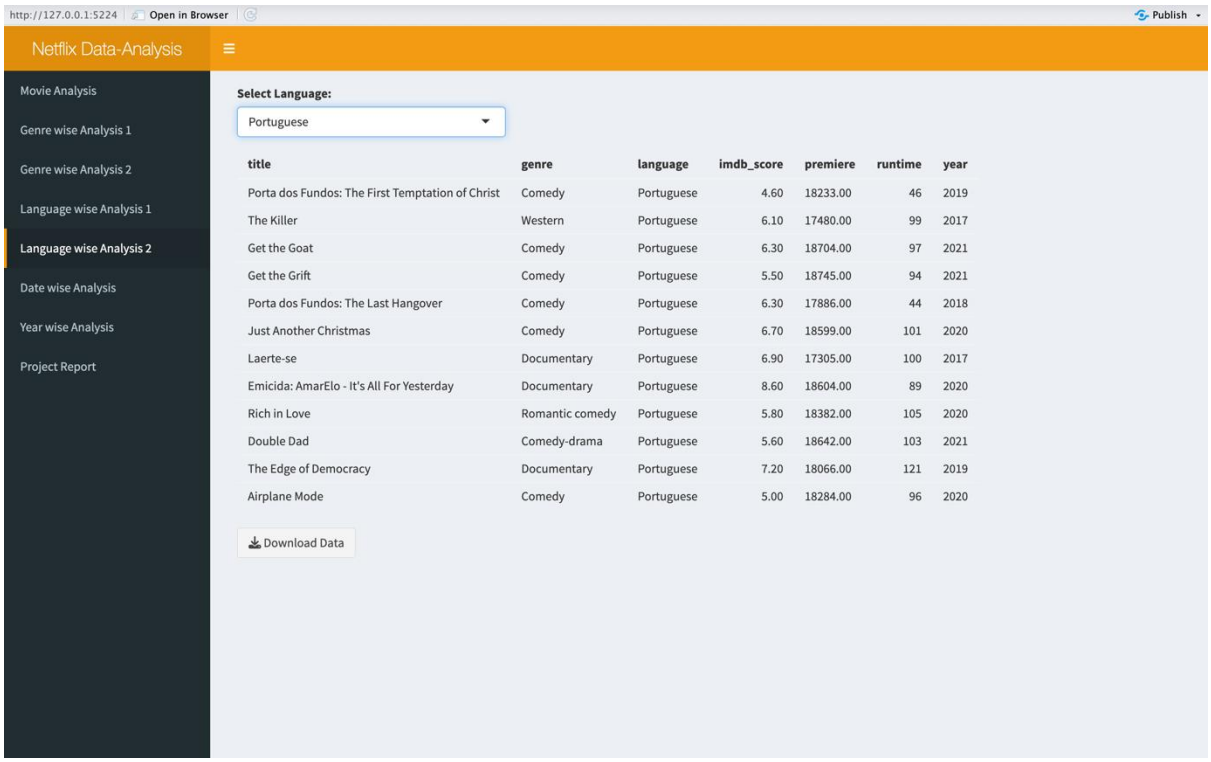
This code segment aims to render a table displaying movie data filtered by the selected language input. It allows users to explore the details of movies available in a specific language, facilitating analysis and understanding of language-wise content distribution.

**Specification:**

1. **Rendering Table Based on Selected Language for Language wise Analysis 2 tab (output$dataDisplay):**
    a. The output$dataDisplay function renders a table displaying movie data filtered by the selected language input.
    b. It filters the movies_data dataset based on the selected language input (input$languageInput).
    c. The renderTable function is used to dynamically update and display the filtered movie data table in the Shiny application UI.

This functionality enables users to explore detailed information about movies available in a particular language, aiding in language-wise analysis and content exploration within the Netflix movie catalog.



**Objective 5: To Analyze Movie Data Based on Selected Date Range**.

**Description:**

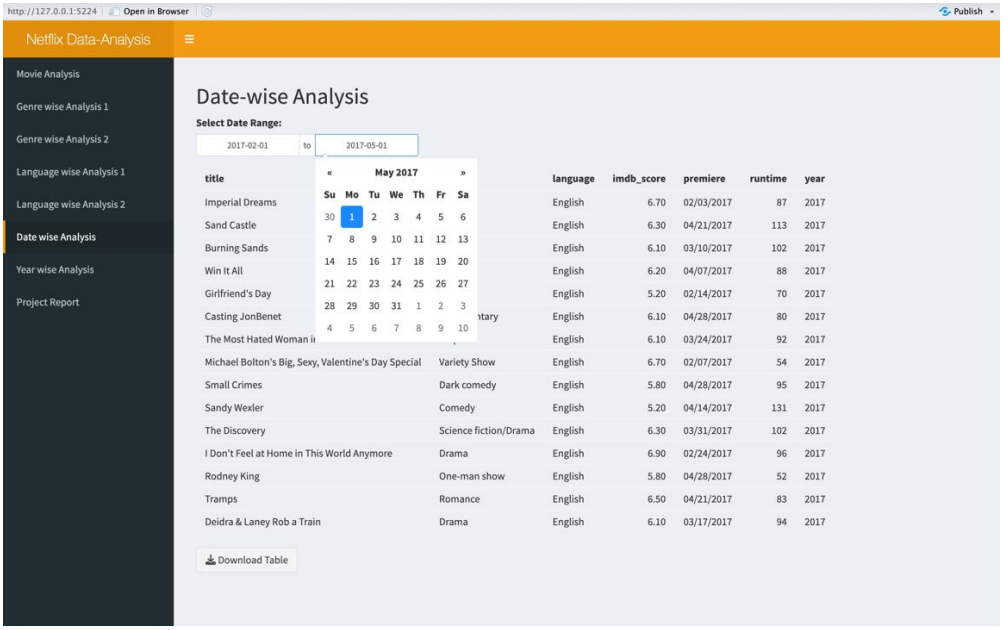
This code segment aims to render a table displaying movie data filtered by the selected date range input. It allows users to explore movie data within a specific period, facilitating analysis and understanding of temporal trends and patterns in movie releases.

**Specification:**

1. **Rendering Table Based on Selected Date Range for Date wise Analysis tab (output$dateAnalysis):**

    a. The output$dateAnalysis function renders a table displaying movie data filtered by the selected date range input.
    b. It filters the movies_data dataset based on the selected date range input (input$dateRange[1] and input$dateRange[2]).
    c. The filter function from the dplyr package is used to subset the dataset based on the specified date range.
    d. The mutate function is employed to format the "premiere" column to display dates in the format "%m/%d/%Y" for better readability.
    e. The renderTable function dynamically updates and displays the filtered movie data table in the Shiny application UI.

This functionality empowers users to explore and analyze movie data within a specific time frame, facilitating insights into temporal trends, seasonal patterns, and the evolution of content over time within the Netflix movie catalog.




**Objective 6: To Visualize the Number of Movies Released Per Year.**

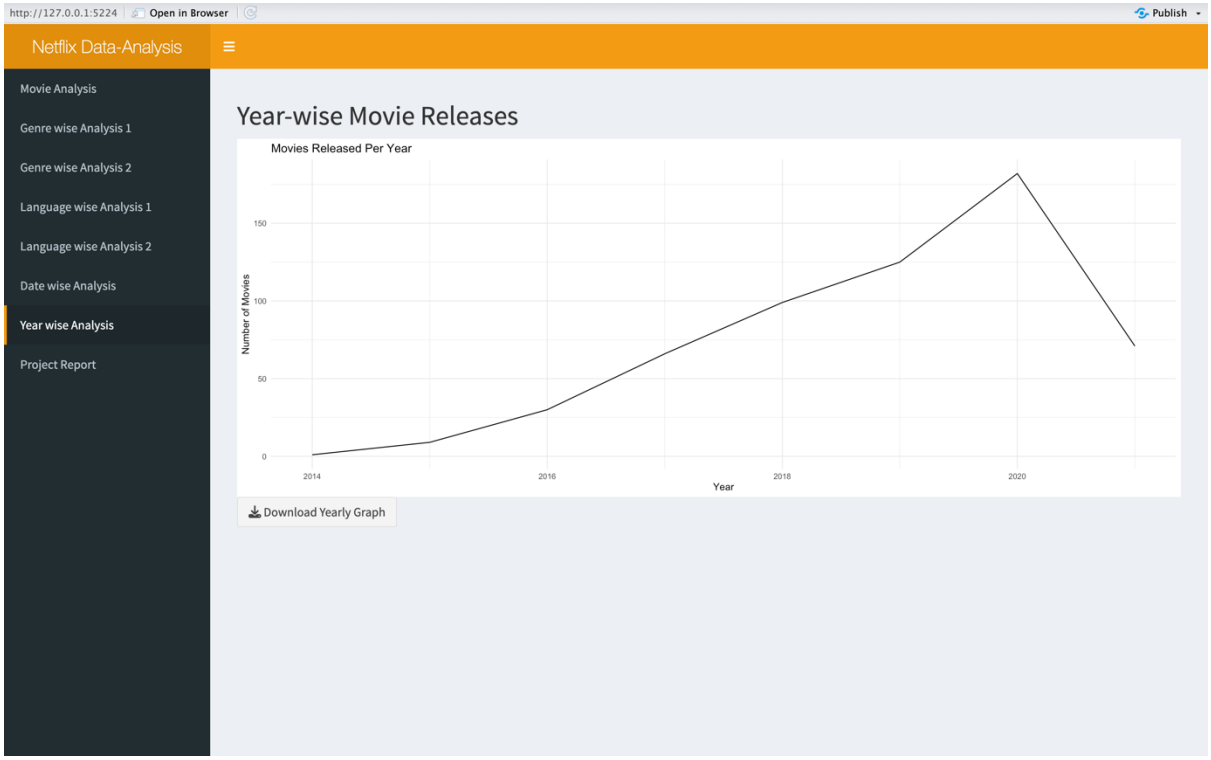
**Description:**

This code segment aims to render a line graph depicting the number of movies released per year. It allows users to explore the trend of movie releases over time, facilitating analysis and understanding of the evolution of Netflix's movie catalog.

**Specification:**

1. **Rendering Yearly Graph for Year wise Analysis tab (output$yearlyGraph):**
    a. The output$yearlyGraph function renders a line graph displaying the number of movies released per year.
    b. It aggregates movie data by year using the group_by function from the dplyr package.
    c. The summarise function calculates the count of movies for each year using the n() function.
    d. The resulting data frame (yearly_data) is arranged in ascending order of year.
    e. The ggplot function is utilized to create the line graph, mapping the x-axis to the numeric representation of the year (as.numeric(year)) and the y-axis to the count of movies.
    f. The geom_line function adds a line connecting data points to visualize the trend of movie releases over time.
    g. Additional aesthetic elements such as title, axis labels, and theme settings are applied to enhance the plot.
    h. This line graph is dynamically updated and displayed in the Shiny application UI for the Year wise Analysis tab.

This functionality enables users to visualize and analyze the temporal distribution of movie releases, providing insights into the growth and trends within the Netflix movie catalog over the years.



## Summary

The Netflix Data Analysis Dashboard offers a comprehensive platform for exploring and understanding the intricate dynamics of Netflix's movie catalog. Through an array of interactive visualizations and analytical tools, stakeholders gain valuable insights into various facets of content consumption and audience preferences.

At its core, the dashboard provides a detailed analysis of movie ratings, genres, languages, release dates, and temporal trends. By identifying top-rated and lowest-rated movies based on IMDb scores, stakeholders can discern patterns in content quality and audience reception. Furthermore, dissecting the movie catalog by genre and language unveils nuanced insights into viewer demographics and cultural preferences, guiding content creation and marketing strategies.

The dashboard's temporal analysis feature visualizes the distribution of movie releases over the years, enabling stakeholders to track trends, anticipate demand, and plan content strategies accordingly. Looking forward, the dashboard holds immense potential for further expansion and refinement. Future iterations could leverage advanced analytics techniques such as predictive modeling and sentiment analysis to forecast viewership trends and gauge audience sentiment towards specific content.

Moreover, integration of geospatial data could unveil regional variations in content preferences, facilitating targeted localization efforts and audience engagement initiatives. As the digital entertainment landscape continues to evolve, the Netflix Data Analysis Dashboard stands poised as a pivotal tool for driving strategic initiatives, fostering innovation, and enhancing the overall viewing experience for audiences worldwide.

Link to my Dashboard: https://saiteja123.shinyapps.io/Netflix_Data_Analysis_Dashboard/

# References

[1] https://www.kaggle.com/datasets/yaminh/netflix-dataset-for-analysis

[2] https://shiny.posit.co/r/getstarted/shiny-basics/lesson1/index.html

[3] https://www.r-project.org/about.html