EXTRACTING USEFUL INFORMATION FROM A DATASET OF MOVIES

A PROJECT REPORT

Submitted by

PRIYADHARSHINI J - 2303811724322085

in partial fulfillment of requirements for the award of the course

AGB1252-FUNDAMENTALS OF DATA SCIENCE USING R

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(An Autonomous Institution, affiliated to Anna University Chennai and Approved by AICTE, New Delhi)

SAMAYAPURAM – 621 112

JUNE- 2025

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY (AUTONOMOUS)

SAMAYAPURAM – 621 112

BONAFIDE CERTIFICATE

Certified that this project report titled **"EXTRACTING USEFUL INFORMATION FROM A DATASET OF MOVIES"** is the bonafide work of **PRIYADHARSHINI J(2303811724322085)** who carried out the project work during the academic year 2024 - 2025 under my supervision.

SIGNATURE

Dr.T. AVUDAIAPPAN, M.E.,Ph.D.,

HEAD OF THE DEPARTMENT

ASSOCIATE PROFESSOR

Department of Artificial Intelligence

K.Ramakrishnan College of Technology (Autonomous)

Samayapuram–621112.

SIGNATURE

Ms.S.MURUGAVALLI, M.E.,(Ph.D).,

SUPERVISOR

ASSISTANT PROFESSOR

Department of Artificial Intelligence

K.Ramakrishnan College of Technology (Autonomous)

Samayapuram–621112.

Submitted for the viva-voce examination held on …02.06.2025…….

INTERNAL EXAMINER

EXTERNAL EXAMINER

DECLARATION

I declare that the project report **"EXTRACTING USEFUL INFORMATION FROM A DATASET OF MOVIES"** is the result of original work done by me and best of my knowledge, similar work has not been submitted to **"ANNA UNIVERSITY CHENNAI"** for the requirement of Degree of **BACHELOR OF TECHNOLOGY**. This project report is submitted on the partial fulfilment of the requirement of the completion of the course **AGB1252 – FUNDAMENTALS OF DATA SCIENCE USING R.**

.

PRIYADHARSHINI J

Place: Samayapuram

Date: 02.06.2025

ACKNOWLEDGEMENT

It is with great pride that I express our gratitude and in-debt to our institution **"K.Ramakrishnan College of Technology (Autonomous)"**, for providing us with the opportunity to do this project.

I glad to credit honorable chairman **Dr. K. RAMAKRISHNAN**, **B.E.,** for having provided for the facilities during the course of our study in college.

I would like to express our sincere thanks to our beloved Executive Director **Dr. S. KUPPUSAMY, MBA, Ph.D.,** for forwarding to our project and offering adequate duration in completing our project.

I would like to thank **Dr. N. VASUDEVAN, M.Tech., Ph.D.,** Principal, who gave opportunity to frame the project the full satisfaction.

I whole heartily thanks to **Dr. T. AVUDAIAPPAN, M.E.,Ph.D.,** Head of the department, **ARTIFICIAL INTELLIGENCE** for providing his encourage pursuing this project.

I express our deep expression and sincere gratitude to our project supervisor **Ms.S.Murugavalli., M.E.,(Ph.D).,** Department of **ARTIFICIAL INTELLIGENCE,** for her incalculable suggestions, creativity, assistance and patience which motivated us to carry out this project.

I render our sincere thanks to Course Coordinator and other staff members for providing valuable information during the course.

I wish to express our special thanks to the officials and Lab Technicians of our departments who rendered their help during the period of the work progress.

INSTITUTE

Vision:

- To serve the society by offering top-notch technical education on par with global standards.

Mission:

- Be a center of excellence for technical education in emerging technologies by exceeding the needs of industry and society.
- Be an institute with world class research facilities.
- Be an institute nurturing talent and enhancing competency of students to transform them as all – round personalities respecting moral and ethical values.

DEPARTMENT

Vision:

- To excel in education, innovation, and research in Artificial Intelligence and Data Science to fulfil industrial demands and societal expectations.

Mission

- To educate future engineers with solid fundamentals, continually improving teaching methods using modern tools.
- To collaborate with industry and offer top-notch facilities in a conducive learning environment.
- To foster skilled engineers and ethical innovation in AI and Data Science for global recognition and impactful research.
- To tackle the societal challenge of producing capable professionals by instilling employability skills and human values.

PROGRAM EDUCATIONAL OBJECTIVES (PEO)

- **PEO1:** Compete on a global scale for a professional career in Artificial Intelligence and Data Science.
- **PEO2:** Provide industry-specific solutions for the society with effective communication and ethics.
- **PEO3** Enhance their professional skills through research and lifelong learning initiatives.

PROGRAM SPECIFIC OUTCOMES (PSOs)

- **PSO1:** Capable of finding the important factors in large datasets, simplify the data, and improve predictive model accuracy.
- **PSO2:** Capable of analyzing and providing a solution to a given real-world problem by designing an effective program.

PROGRAM OUTCOMES (POs)

Engineering students will be able to:

1. **Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis:** Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences
3. **Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations
4. **Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions
5. **Engineering Tool Usage:** Create, select and apply appropriate techniques, resources and modern engineering & IT tools, including prediction and modelling recognizing their limitations to solve complex engineering problems.
6. **The Engineer and The World:** Analyze and evaluate societal and environmental aspects while solving complex engineering problems for its impact on sustainability with reference to economy, health, safety, legal framework, culture and environment.
7. **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

8. **Individual and team work:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
9. **Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
10. **Project management and finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
11. **Life-long learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

ABSTRACT

An interactive Shiny-based web application developed in R enables intuitive exploration and analysis of a structured movie dataset. The app allows users to upload movie data in CSV or Excel format and interactively engage in essential preprocessing tasks such as data viewing, editing, summary statistics, and visualization. Designed with modular components, the application simplifies the process of data-driven insights through dynamic UI controls and real-time rendering. Key functionalities include data upload and validation, editable tables for record management, descriptive statistical summaries, genre-based filtering, and multi-format visualizations such as bar charts, line graphs, pie charts, and heatmaps. The integration of popular R packages like shiny, shinydashboard, DT, ggplot2, and dplyr ensures robustness, interactivity, and visual clarity throughout the interface. A clean and responsive dashboard layout enhances usability, making the platform suitable for both beginner analysts and experienced data practitioners. Interactive widgets support tailored exploration of genre and revenue trends over time, empowering users to derive valuable insights from the dataset without writing code. By promoting efficient exploratory data analysis, the tool highlights the power of R and Shiny in building user-centric applications for structured data interpretation. This system supports the broader goal of accessible analytics and encourages best practices in dataset handling and visualization. It also serves as a strong foundational platform for potential future enhancements such as predictive analytics and recommendation systems.

ABSTRACT WITH POs AND PSOs MAPPING

CO 5 : BUILD DATABASES FOR SOLVING REAL-TIME PROBLEMS.

ABSTRACT	POs MAPPED	PSOs MAPPED
An interactive Shiny web application developed in R enables comprehensive exploration and analysis of movie datasets through an intuitive dashboard interface. The app supports uploading data in CSV or Excel format, live editing of records, and real-time visualization using bar charts, line graphs, pie charts, and heatmaps. Built with packages such as shiny, shinydashboard, DT, ggplot2, and dplyr, it allows users to perform genre-based filtering, view summary statistics, and export cleaned datasets—all without writing code. Designed to streamline exploratory data analysis, the application promotes accessibility, clarity, and ease of use, making it ideal for both educational and practical purposes. It also establishes a scalable foundation for integrating advanced features like predictive analytics and personalized recommendations in future development.	**PO1 -3** **PO2 -3** **PO3 -2** **PO4 -3** **PO5 -3** **PO6 -3** **PO7 -2** **PO8 -3** **PO9 -3** **PO10 -3** **PO11-3** **PO12 -3**	**PSO1 -2** **PSO2 -2** **PSO3 -3**

Note: 1- Low, 2-Medium, 3- High

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	viii
1	**INTRODUCTION**	
	1.1 Objective	1
	1.2 Overview	1
	1.3 Data Science related concepts	2
2	**PROJECT METHODOLOGY**	
	2.1 Proposed Work	3
	2.2 Block Diagram	4
3	**MODULE DESCRIPTION**	
	3.1 Data Upload & Preprocessing Module	5
	3.2 Data Editing Module	5
	3.3 Summary & EDA Module	5
	3.4 Visualization Module	6
	3.5 Report & Export Module	6
	3.6 UI & Dashboard Module	6
4	**CONCLUSION & FUTURE SCOPE**	7
5	**APPENDIX A SOURCE CODE**	9
	APPENDIX B SCREENSHOTS	14
	REFERENCES	16

CHAPTER 1

INTRODUCTION

1.1 OBJECTIVE

The objective is to develop an interactive, modular web application using R and Shiny that simplifies the exploration and analysis of structured movie datasets. The system enables users to upload data files, edit records, generate descriptive statistics, and visualize patterns through a responsive dashboard. By integrating intuitive UI components and real-time feedback mechanisms, the application promotes efficient, code-free data analysis for both learners and professionals. The design emphasizes accessibility, flexibility, and ease of use, supporting a smooth data exploration workflow.

1.2 OVERVIEW

This application provides a comprehensive Shiny-based dashboard that facilitates the end-to-end analysis of a movie dataset. Users begin by uploading their dataset in CSV or Excel format. Once uploaded, the data is displayed in an interactive table that supports live editing, allowing users to correct or update records on the spot. The summary statistics tab generates descriptive insights such as minimum, maximum, mean, and median values for numeric fields like revenue and release year. The visualization module enables users to generate bar charts, line graphs, pie charts, and heatmaps with real-time updates based on genre-based filters. A final module allows the export of cleaned and transformed data as a downloadable CSV file. Throughout the app, real-time responsiveness is maintained using reactive programming principles, ensuring seamless user experience. Built entirely with R and open-source packages, the system combines data management, visualization, and export capabilities into a unified interface.

1.3 DATA SCIENCE RELATED CONCEPTS

1. Exploratory Data Analysis (EDA)

A crucial step in data science that involves summarizing the main characteristics of the dataset using visualizations and descriptive statistics. In your app, this is done through summary tables, genre distributions, and revenue trend plots.

2. Data Cleaning & Preprocessing

The process of ensuring data quality before analysis by handling incorrect, inconsistent, or missing entries. Your app supports this through data upload validation, editable tables, and column-wise statistics.

3. Interactive Data Visualization

The use of dynamic charts to make data patterns visible and actionable. Bar charts, pie charts, line graphs, and heatmaps in your app allow users to visually interpret trends across genres and years.

4. User-Centered Dashboard Design

A core concept in data science product development. It involves designing interfaces that prioritize clarity, accessibility, and usability-reflected in your use of shinydashboard and organized tab-based layout.

5. Reactive Programming

A key concept in Shiny development that aligns with data science automation principles. It ensures that outputs update instantly in response to user inputs, enhancing real-time data exploration.

CHAPTER 2

PROJECT METHODOLOGY

2.1 PROPOSED WORK

The proposed work is to design a Shiny-based application that streamlines the process of exploratory data analysis for structured movie datasets. The application allows users to upload datasets in CSV or Excel format and perform essential operations such as data editing, summary generation, and visualization through an intuitive dashboard interface. It aims to provide a clean, modular environment where users can interact with the data in real time-without requiring programming skills-making it a practical tool for both beginners and data practitioners.

Data Collection and Input

- Users upload a structured movie dataset in CSV or Excel format
- App reads and displays the data in a reactive table with built-in editing features

Data Exploration

- Summary statistics (mean, median, min, max) for numeric fields like Revenue and Year
- Structural insights into column types and data distribution

Data Filtering and Editing

- Genre-based filtering using dynamic multi-select input
- Inline editing, deletion, and correction of data using interactive DT tables

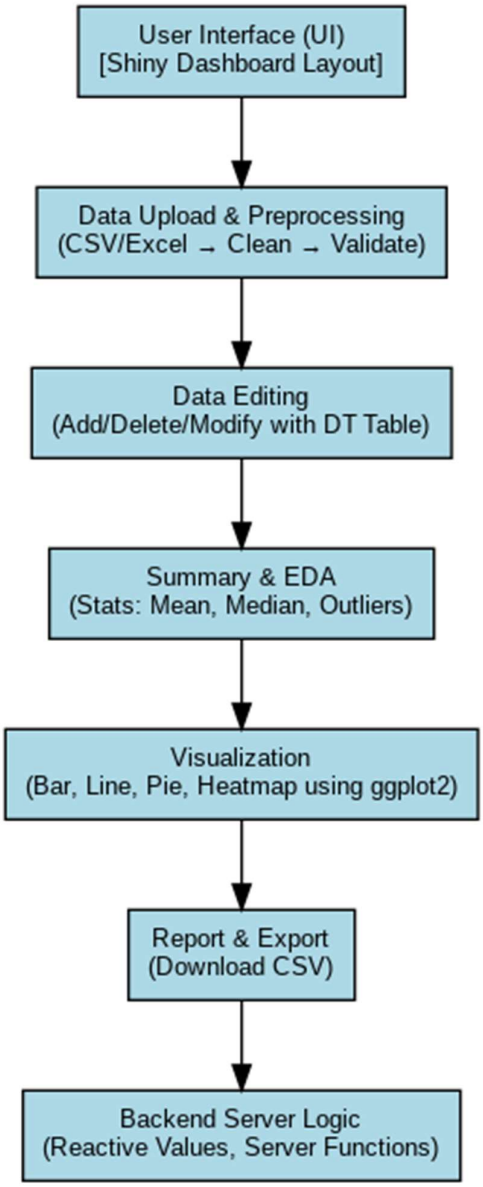
Visualization

- Allows filtering by genre to dynamically update visual outputs
- Real-time plots using `ggplot2` to reflect current dataset state

Report and Export

- Cleaned/edited data can be downloaded as a new CSV file
- Prepares the dataset for downstream reporting or analysis

2.2 BLOCK DIAGRAM



CHAPTER 3

MODULE DESCRIPTION

3.1 DATA UPLOAD & PREPROCESSING MODULE

This module is the initial step in the application workflow, allowing users to upload structured movie datasets in CSV or Excel format. Once uploaded, the system validates required fields (such as Title, Genre, Year, and Revenue) and displays the dataset in an interactive table. The module ensures correct formatting, handles file extension logic, and sets the foundation for exploration. It also previews the dataset structure to help users assess data completeness and relevance before moving forward.

3.2 DATA EDITING MODULE

This module provides in-app editing capabilities through an interactive, spreadsheet-like table using the DT package. Users can update values, delete rows, or make inline corrections to the data. All modifications are reflected immediately in the reactive data structure, ensuring that the updated dataset is consistently available across other modules like visualizations and exports. This flexibility allows for quick adjustments without leaving the app.

3.3 SUMMARY & EDA MODULE

The Summary and Exploratory Data Analysis (EDA) module offers users a high-level statistical overview of the uploaded dataset. It displays key statistics such as mean, median, and quartiles for numerical variables (e.g., Revenue, Year), and counts for categorical variables (e.g., Genre). This insight helps users understand the data distribution, identify potential outliers, and recognize patterns. It plays a crucial role in preparing the dataset for meaningful visualization and further analysis.

3.4 VISUALIZATION MODULE

This module is responsible for converting data into meaningful graphical representations. Users can generate various chart types such as bar charts (to analyze genre frequency), line graphs (to observe revenue trends by year), pie charts (for genre proportions), and heatmaps (to understand Genre-Year-Revenue intersections). The charts are reactive, adjusting based on user-selected filters like genre. Built with ggplot2, the visuals enhance interpretability and engagement.

3.5 PREDICTIVE IMPUTATION MODULE

The report and export module enables users to save the modified dataset for further use. After editing and analyzing the data, users can download it as a CSV file directly from the interface. This module ensures a smooth transition from interactive analysis to external reporting, making it suitable for integration into presentations, documentation, or additional modeling workflows.

3.6 UI & DASHBOARD MODULE

This module defines the structure and navigation of the application using shinydashboard. It organizes content into clean, tab-based layouts including Upload, Summary, Charts, and Export sections. Enhanced with packages like shinyWidgets, bslib, and shinyjs, the user interface provides a modern look and real-time responsiveness. This module ensures a smooth user experience by guiding users through each stage of analysis without requiring technical expertise.

CHAPTER 4

CONCLUSION

4.1 CONCLUSION

The Movie Dataset Analysis App demonstrates the powerful capabilities of R and Shiny in delivering an end-to-end solution for interactive data analysis. The application offers a user-friendly interface where datasets can be uploaded, explored, edited, visualized, and exported without requiring any programming knowledge. Each module — from data import and editing to summary statistics and visualization — is designed to function independently while contributing to a cohesive workflow. By using widely adopted R packages like `shiny`, `shinydashboard`, `ggplot2`, `dplyr`, and `DT`, the system provides a smooth, reactive environment for performing exploratory data analysis. Real-time updates, genre-based filtering, and visually rich charts enhance the overall analytical experience. The app proves especially useful in academic settings, where learners can gain practical experience with structured data, and in professional environments, where quick insights and clean exports are essential.

4.2 FUTURE SCOPE

While the current version of the app serves as a strong foundation for data exploration and visualization, several enhancements can further extend its utility and intelligence:

1. Real-Time Machine Learning Integration

Incorporating ML models like linear regression, decision trees, or random forests could enable real-time predictions (e.g., revenue estimation) based on selected attributes such as genre and release year.

2. Recommendation System

Building a recommendation engine using collaborative or content-based filtering would allow users to explore related or top-rated movies, based on existing data patterns.

3. Sentiment Analysis from User Reviews

Integrating text mining and sentiment analysis on movie reviews could add another layer of insight, enabling users to understand public perception and critical reception.

4. Cloud or Database Integration

Migrating from local file uploads to connected SQL/NoSQL databases or cloud platforms like Firebase or Google Sheets could support persistent multi-user environments and real-time data updates.

5. Advanced Interactive Visualizations

Incorporating `plotly` or `highcharter` for interactive charts with tooltips, zooming, and animation could elevate the analytical depth and user engagement.

6. Public Web Deployment

Publishing the app on Shinyapps.io or hosting on RStudio Connect would allow broader access, making it usable from any device with internet connectivity.

APPENDICES

APPENDIX A – SOURCE CODE

```
library(shiny)
library(shinydashboard)
library(DT)
library(dplyr)
library(ggplot2)
library(bslib)
library(shinyWidgets)
library(shinyjs)

# UI Definition
ui <- dashboardPage(
  dashboardHeader(title = span("🎬 Movie Dataset Analysis App", style =
"color:white")),
  dashboardSidebar(
    sidebarMenu(
      menuItem("Upload & Edit", tabName = "upload", icon =
icon("upload")),
      menuItem("Summary & Stats", tabName = "summary", icon =
icon("chart-line")),
      menuItem("Charts", tabName = "charts", icon = icon("chart-pie")),
      menuItem("Export", tabName = "export", icon = icon("file-export"))
    )
  ),
  dashboardBody(
    useShinyjs(),
```

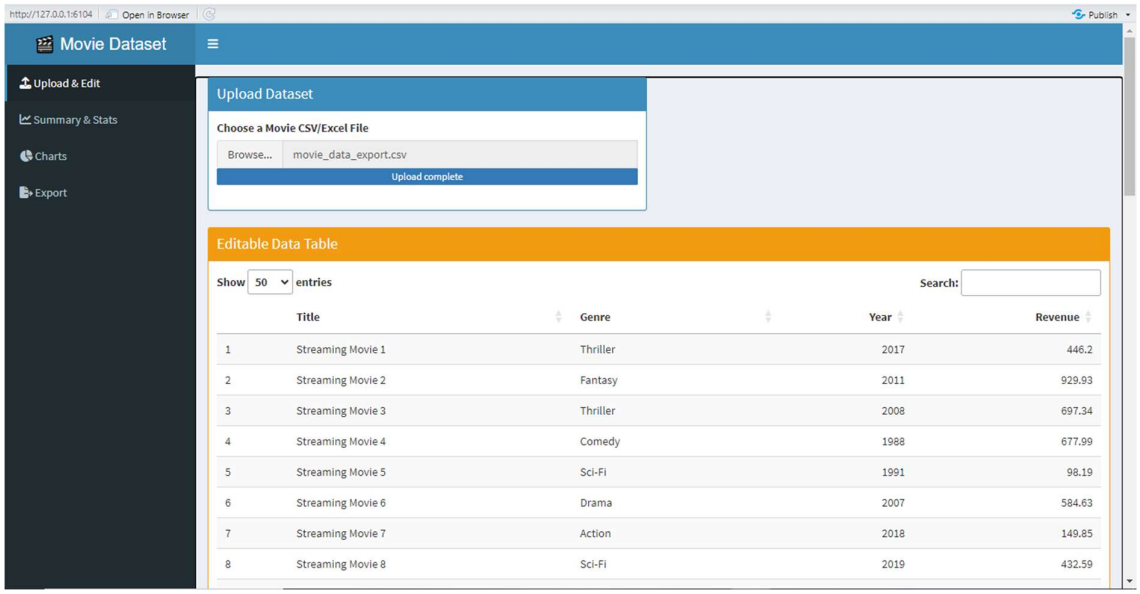
```
    theme = bs_theme(bootswatch = "cosmo"),
    tabItems(
      # Upload & Edit Tab
      tabItem(tabName = "upload",
              fluidRow(
                box(width = 6, status = "primary", solidHeader = TRUE, title =
"Upload Dataset",
                    fileInput("file", "Choose a Movie CSV/Excel File", accept =
c(".csv", ".xlsx"))
                ),
                box(width = 12, status = "warning", solidHeader = TRUE, title
= "Editable Data Table",
                    DTOutput("editableTable"))
              )
      ),

      # Summary Tab
      tabItem(tabName = "summary",
              fluidRow(
                box(width = 12, status = "info", solidHeader = TRUE, title =
"Summary Statistics",
                    verbatimTextOutput("summaryOutput"))
              )
      ),
      # Charts Tab
      tabItem(tabName = "charts",
              fluidRow(
                box(width = 4, status = "primary", solidHeader = TRUE, title =
"Chart Settings",
                    selectInput("chartType", "Select Chart Type",
```

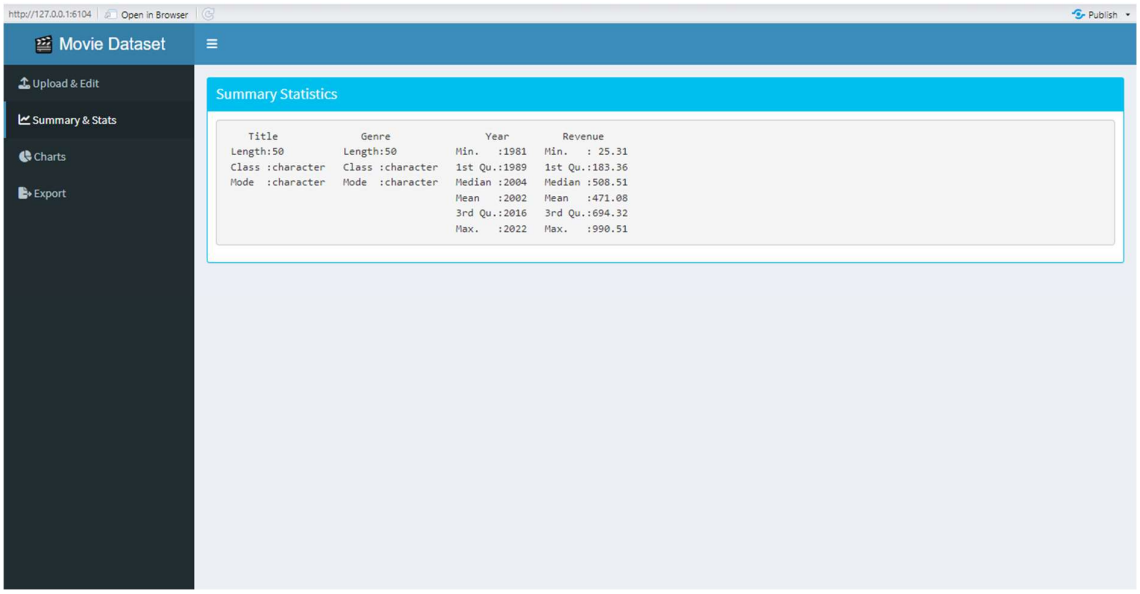
```
 choices = c("Bar Chart", "Line Chart", "Pie Chart", "Heatmap")),
                    pickerInput("selectedGenre", "Filter by Genre",
                                choices = NULL, multiple = TRUE, options =
list(`actions-box` = TRUE))
                ),
                box(width = 8, status = "success", solidHeader = TRUE, title =
"Visualization",
                    plotOutput("chartOutput", height = "400px"))
            )
        ),

        # Export Tab
        tabItem(tabName = "export",
            fluidRow(
                box(width = 6, status = "success", solidHeader = TRUE, title =
"Download Your Data",
                    downloadButton("downloadCSV", "📥 Download as CSV")
                )
            )
        )
    )
)

# Server Logic
server <- function(input, output, session) {

  # Reactive Data Store
  movie_data <- reactiveVal(data.frame())
```

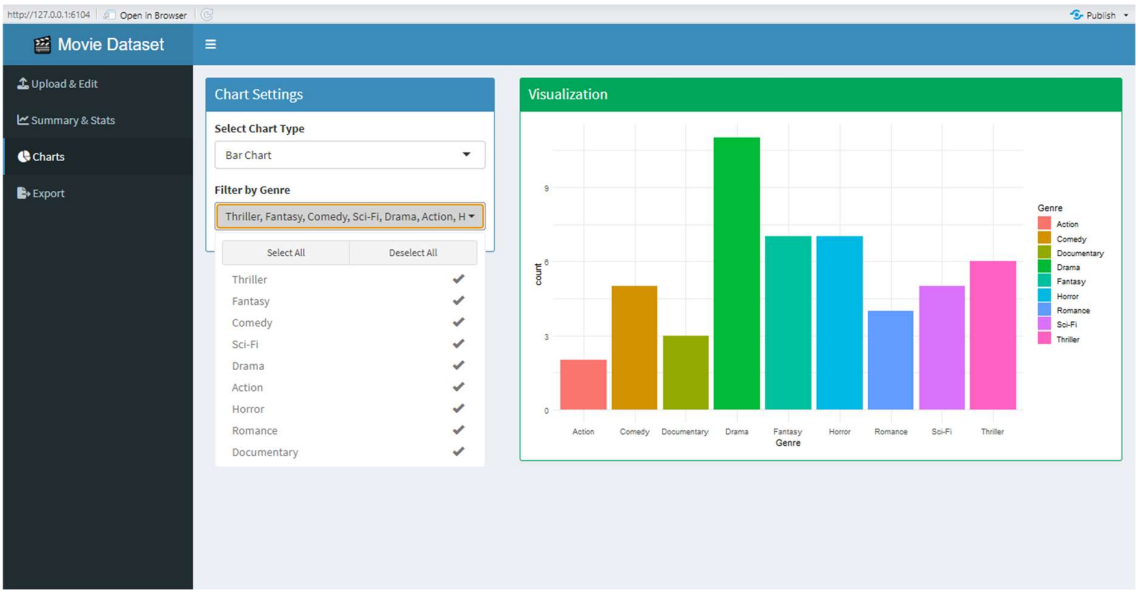
```
  # Load CSV or Excel File
  observeEvent(input$file, {
    ext <- tools::file_ext(input$file$name)
    df <- if (ext == "csv") {
      read.csv(input$file$datapath)
    } else if (ext == "xlsx") {
      readxl::read_excel(input$file$datapath)
    }
    req(all(c("Title", "Genre", "Year", "Revenue") %in% colnames(df)))
    movie_data(df)

    updatePickerInput(session, "selectedGenre", choices = unique(df$Genre),
selected = unique(df$Genre))
  })

  # Editable Table
  output$editableTable <- renderDT({
    datatable(movie_data(), editable = TRUE, options = list(pageLength =
10))
  })

  # Summary Statistics
  output$summaryOutput <- renderPrint({
    summary(movie_data())
  })

  # Chart Rendering
  output$chartOutput <- renderPlot({
    df <- movie_data()
```

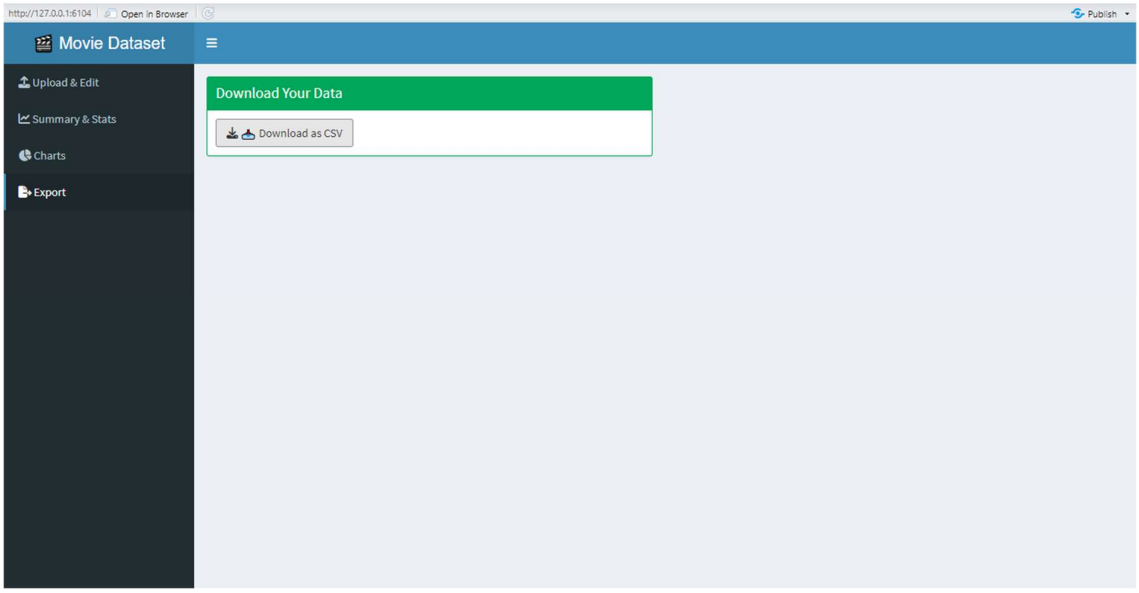
```
      if (!is.null(input$selectedGenre)) {

df <- df[df$Genre %in% input$selectedGenre, ]
      }
      if (input$chartType == "Bar Chart") {
        ggplot(df, aes(x = Genre, fill = Genre)) + geom_bar() +
theme_minimal()
      } else if (input$chartType == "Line Chart") {
        ggplot(df, aes(x = Year, y = Revenue, color = Genre)) + geom_line() +
theme_minimal()
      } else if (input$chartType == "Pie Chart") {
        pie(table(df$Genre), col = rainbow(length(unique(df$Genre))))
      } else if (input$chartType == "Heatmap") {
        ggplot(df, aes(x = Genre, y = Year, fill = Revenue)) +
          geom_tile(color = "white") + theme_minimal()
      }
  })

  # Download as CSV
  output$downloadCSV <- downloadHandler(
    filename = function() {"movie_data_export.csv"},
    content = function(file) {
      write.csv(movie_data(), file, row.names = FALSE)
    }
  )
}

# Run the App
shinyApp(ui, server)
```

APPENDIX B - SCREENSHOTS






http://127.0.0.1:6104
Open in Browser
Publish
Movie Dataset
Upload & Edit
Summary & Stats
Charts
Export
Chart Settings
Select Chart Type
Bar Chart
Filter by Genre
Thriller, Fantasy, Comedy, Sci-Fi, Drama, Action, H
Select All
Deselect All
Thriller
Fantasy
Comedy
Sci-Fi
Drama
Action
Horror
Romance
Documentary
Visualization
count
Genre
Action
Comedy
Documentary
Drama
Fantasy
Horror
Romance
Sci-Fi
Thriller


http://127.0.0.1:6104
Open in Browser
Publish
Movie Dataset
Upload & Edit
Summary & Stats
Charts
Export
Download Your Data
Download as CSV

REFERENCES

- **Allaire, J.J., et al.** (2023). *shiny: Web Application Framework for R*. RStudio Documentation. https://shiny.posit.co/

- **Chang, W., et al.** (2023). *shinydashboard: Create Dashboards with Shiny*. RStudio. https://rstudio.github.io/shinydashboard/

- **Grolemund, G., & Wickham, H.** (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media. https://r4ds.had.co.nz

- **Kuhn, M., & Silge, J.** (2022). *Tidy Modeling with R*. O'Reilly Media. https://www.tmwr.org/

- **Sievert, C.** (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. CRC Press. https://plotly-r.com/

- **Wickham, H.** (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

- **Wickham, H., François, R., Henry, L., & Müller, K.** (2023). *dplyr: A Grammar of Data Manipulation*. Tidyverse Documentation. https://dplyr.tidyverse.org