A

Project Report

on

MOVIE GENRE ANALYSIS

For the Degree of

Bachelor of Technology in

Artificial Intelligence in Data Science

By

Tauheed Abdullah Khan (A26)

Shruti Hemant Konde (A28)

Faizan Irfan Mirza (A36)

Under the Guidance of

S.B.Waghmare

QUEST FOR EXCELLENCE

Department of Emerging Science and Technology

Marathwada Institute of Technology, Aurangabad

Maharashtra State, India

2023-2024

CERTIFICATE

This is to certify that, the project entitled **“MOVIE GENRE ANALYSIS”,** which has been submitted herewith in the partial fulfilment for the award of the **‘Bachelor of Technology’** in **‘Artificial Intelligence in Data Science’** of Dr. Babasaheb Ambedkar Technological University, Lonere, Raigad (M.S.). This is the result of the original work and contribution for **Project Part I** ,by **Tauheed Abdullah Khan, Shruti Hemant Konde , Faizan Irfan Mirza** under my supervision and guidance.

Place: Aurangabad
Date:

S.B.Waghmare
Guide

Dr. K. V. Bhosle
Head,
Department of Emerging Science and Technology

Dr. N.G. Patil
Director
MIT, Aurangabad (M.S.) - 431 010

1.Introduction

Provide an overview of the project, emphasizing the importance of analyzing movie genres. Highlight the significance of understanding audience preferences for movie recommendations and industry trends.

2.Project Overview

The proposed mini project focuses on leveraging R programming for an in-depth analysis of movie genres using a dataset sourced from IMDb in CSV format. The primary objective is to provide a comprehensive visualization of the distribution of movie genres within the dataset. The dataset comprises a list of movies with associated information such as title, release year, and genre. The analysis will primarily involve the extraction, cleaning, and transformation of relevant data using R programming techniques.

The core visualization components of this mini project include the creation of separate pie charts and histograms. The pie charts will vividly represent the proportional distribution of various movie genres within the dataset, offering an intuitive insight into the overall genre landscape. On the other hand, the histograms will provide a more detailed perspective by illustrating the frequency distribution of genres, allowing for a closer examination of genre prevalence. Through this mini project, participants will not only gain hands-on experience with R programming and data visualization but will also develop a practical understanding of the distribution patterns within the IMDb movie dataset.

3.Project Methodology

a) Data Acquisition:

- Obtain the IMDb movie dataset in CSV format.
- Load the dataset into R using appropriate functions (e.g., **read.csv**).

b) Data Cleaning and Pre-processing:

- Identify and handle missing or inconsistent data.
- Extract relevant columns related to movie titles, release years, and genres.
- Ensure data types are appropriate for analysis.
- Remove duplicates or irrelevant entries.

c) Data Exploration:

- Conduct exploratory data analysis to gain insights into the structure of the dataset.
- Calculate basic statistics such as the total number of movies, unique genres, and the range of release years.
- Generate summary tables to understand the distribution of genres.

d) **Genre Categorization:**

- Group movies by genre to prepare data for visualization.
- Create a genre-wise count of movies to use in pie charts and histograms.

e) **Pie Chart Visualization:**

- Use R's plotting libraries (e.g., **ggplot2**) to generate pie charts.
- Represent the distribution of movie genres as proportions in a visually appealing manner.
- Customize labels and colours for clarity and aesthetics.

f) **Histogram Visualization:**

- Utilize R's plotting capabilities to create histograms.
- Illustrate the frequency distribution of movie genres, allowing for a more granular analysis.
- Adjust bin sizes and labels for optimal interpretation.

g) **Documentation and Reporting:**

- Document the code using comments for clarity and reproducibility.
- Prepare a summary report detailing the findings and insights obtained from the visualizations.

4.**Diagrams**

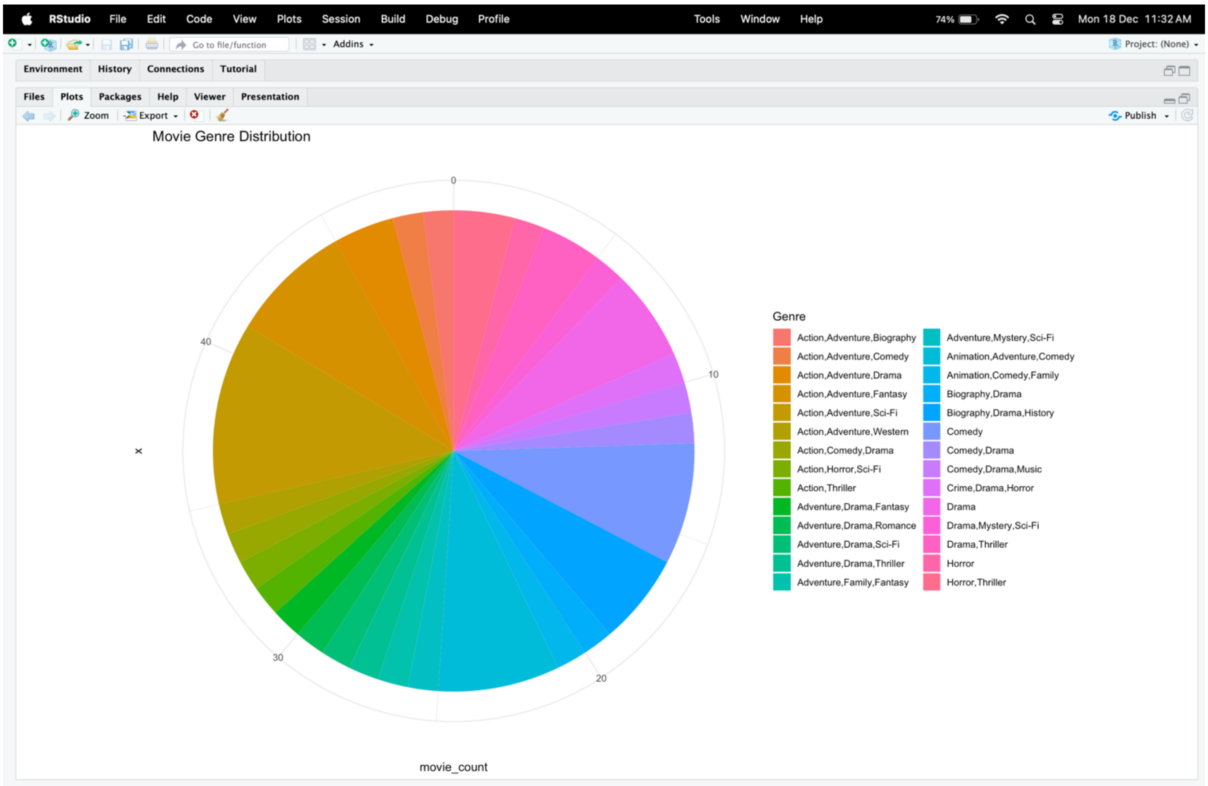
a) **Pie Charts:**

- *Purpose:* Pie charts are utilized to illustrate the proportional distribution of different movie genres within the dataset. Each slice of the pie represents a specific genre, and the size of each slice corresponds to the relative frequency or percentage of movies belonging to that genre.
- *Implementation:* R's plotting libraries, such as **ggplot2**, are employed to create aesthetically pleasing and informative pie charts. Colours and labels are chosen thoughtfully to enhance clarity and visual appeal.
- *Insights:* Pie charts offer a quick and intuitive overview of the overall genre landscape, highlighting which genres dominate the dataset and providing a clear comparison of genre prevalence.

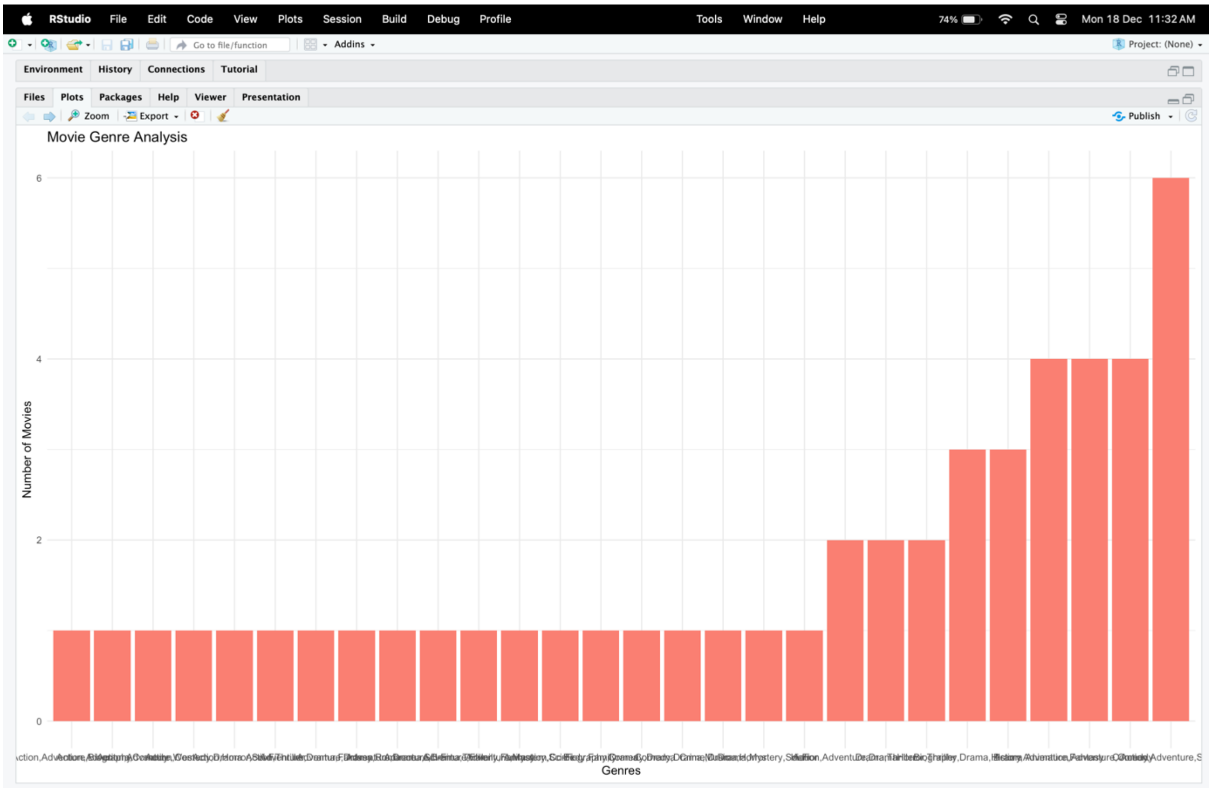
b) **Histograms:**

- *Purpose:* Histograms are employed to delve deeper into the distribution patterns of movie genres. They showcase the frequency of occurrence of each genre, providing a more granular understanding of how genres are distributed across the dataset.
- *Implementation:* R's plotting capabilities are leveraged to construct histograms, with the x-axis representing different genres and the y-axis indicating the frequency or count of movies within each genre. Adjustments to bin sizes and labels are made to optimize readability.
- *Insights:* Histograms allow for the identification of trends, outliers, and patterns in the dataset. They are particularly useful for exploring variations in genre popularity and uncovering potential insights into genre preferences over time or across different movie release periods.

Pie-Chart Print

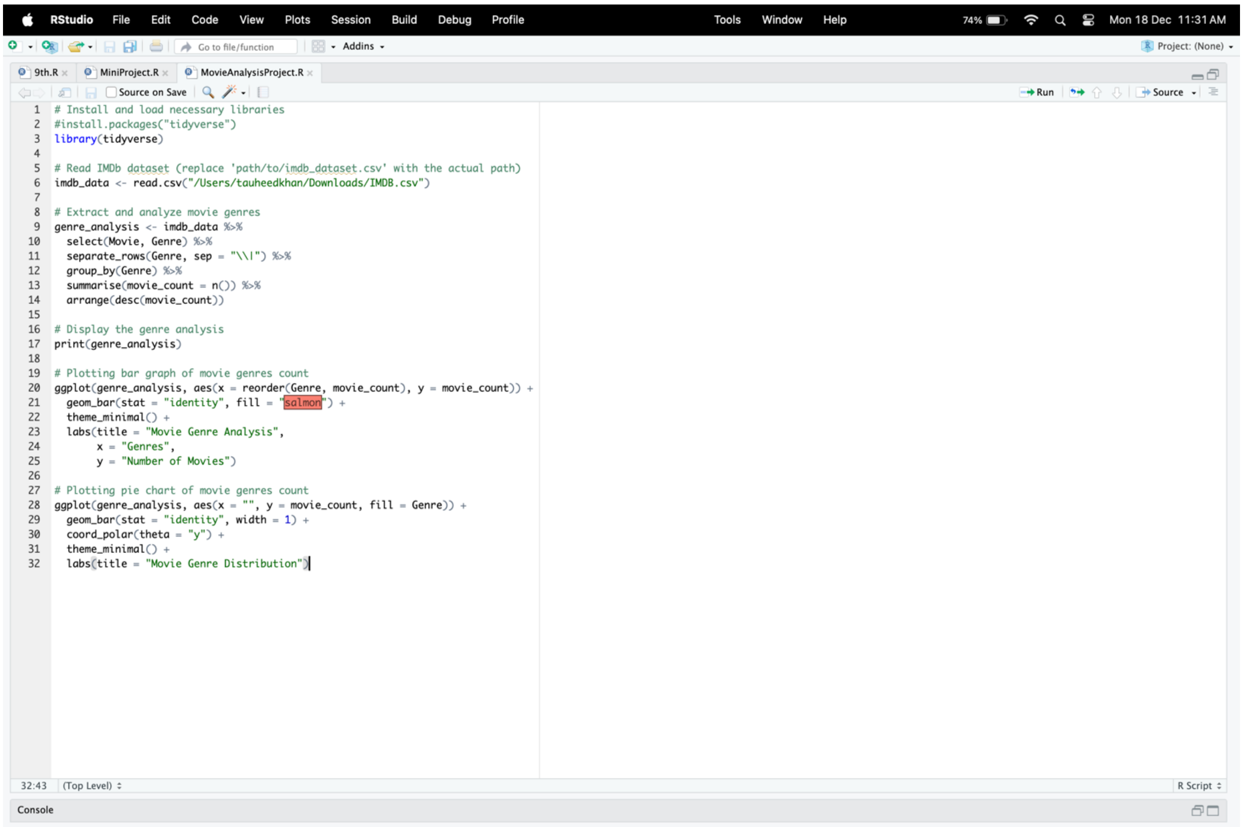


Histogram Print



5.**Program**

```
# Install and load necessary libraries
#install.packages("tidyverse")
library(tidyverse)
# Read IMDb dataset (replace 'path/to/imdb_dataset.csv' with the actual path)
imdb_data <- read.csv("/Users/tauheedkhan/Downloads/IMDB.csv")
# Extract and analyze movie genres
genre_analysis <- imdb_data %>%
  select(Movie, Genre) %>%
  separate_rows(Genre, sep = "\\|") %>%
  group_by(Genre) %>%
  summarise(movie_count = n()) %>%
  arrange(desc(movie_count))
# Display the genre analysis
print(genre_analysis)
# Plotting bar graph of movie genres count
ggplot(genre_analysis, aes(x = reorder(Genre, movie_count), y = movie_count)) +
  geom_bar(stat = "identity", fill = "salmon") +
  theme_minimal() +
  labs(title = "Movie Genre Analysis",
       x = "Genres",
       y = "Number of Movies")
# Plotting pie chart of movie genres count
ggplot(genre_analysis, aes(x = "", y = movie_count, fill = Genre)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  theme_minimal() +
  labs(title = "Movie Genre Distribution")
```



```
# Install and load necessary libraries
#install.packages("tidyverse")
library(tidyverse)

# Read IMDb dataset (replace 'path/to/imdb_dataset.csv' with the actual path)
imdb_data <- read.csv("/Users/tauheedkhan/Downloads/IMDB.csv")

# Extract and analyze movie genres
genre_analysis <- imdb_data %>%
  select(Movie, Genre) %>%
  separate_rows(Genre, sep = "\\|") %>%
  group_by(Genre) %>%
  summarise(movie_count = n()) %>%
  arrange(desc(movie_count))

# Display the genre analysis
print(genre_analysis)

# Plotting bar graph of movie genres count
ggplot(genre_analysis, aes(x = reorder(Genre, movie_count), y = movie_count)) +
  geom_bar(stat = "identity", fill = "salmon") +
  theme_minimal() +
  labs(title = "Movie Genre Analysis",
       x = "Genres",
       y = "Number of Movies")

# Plotting pie chart of movie genres count
ggplot(genre_analysis, aes(x = "", y = movie_count, fill = Genre)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  theme_minimal() +
  labs(title = "Movie Genre Distribution")
```

6.Program Explanation

1. **Library Installation and Loading:**
   - The code begins by installing and loading the tidyverse library, a collection of packages that facilitate data manipulation and visualization.

```
# Install and load necessary libraries #install.packages("tidyverse") library(tidyverse)
```

2. **Read IMDb Dataset:**
   - The IMDb dataset is read into R using the read.csv function. Replace "/Users/tauheedkhan/Downloads/IMDB.csv" with the actual path to your IMDb dataset CSV file.

```
# Read IMDb dataset (replace 'path/to/imdb_dataset.csv' with the actual path) imdb_data <-
read.csv("/Users/tauheedkhan/Downloads/IMDB.csv")
```

3. **Genre Analysis:**
   - The code then extracts and analyzes movie genres. It uses the %>% (pipe) operator to chain together a series of operations.
   - It selects the "Movie" and "Genre" columns from the dataset.
   - The separate_rows function is used to split entries in the "Genre" column that contain multiple genres separated by '|'.
   - The data is then grouped by genre, and the summarise function calculates the count of movies in each genre.
   - Finally, the results are arranged in descending order based on the movie count.

```
genre_analysis <- imdb_data %>% select(Movie, Genre) %>% separate_rows(Genre, sep =
"\\|") %>% group_by(Genre) %>% summarise(movie_count = n()) %>%
arrange(desc(movie_count))
```

4. **Display Genre Analysis:**
   - The genre analysis is printed to the console.

```
# Display the genre analysis print(genre_analysis)
```

5. **Bar Graph (Histogram) Visualization:**
   - The code creates a bar graph using the ggplot2 library to visualize the count of movies for each genre. It uses the geom_bar function to plot the bars, and theme_minimal for a clean appearance.

```
# Plotting bar graph of movie genres count ggplot(genre_analysis, aes(x = reorder(Genre,
movie_count), y = movie_count)) + geom_bar(stat = "identity", fill = "salmon") +
theme_minimal() + labs(title = "Movie Genre Analysis", x = "Genres", y = "Number of
Movies")
```

6. **Pie Chart Visualization:**
   - The code creates a pie chart using ggplot2 to visualize the distribution of movie genres. It uses the geom_bar function with polar coordinates (coord_polar) to create a pie chart.

```
# Plotting pie chart of movie genres count ggplot(genre_analysis, aes(x = "", y = movie_count,
fill = Genre)) + geom_bar(stat = "identity", width = 1) + coord_polar(theta = "y") +
theme_minimal() + labs(title = "Movie Genre Distribution")
```

This code provides a comprehensive analysis of movie genres from the IMDb dataset and offers visualizations to better understand the distribution of genres across the movies.

7.Applications

1. **Content Strategy for Streaming Services:**

   - Streaming platforms like Netflix, Hulu, and Amazon Prime Video can use the genre analysis results to refine their content strategy. By understanding the popularity of different genres, these services can make data-driven decisions on content acquisition, production, and recommendations to cater to diverse viewer preferences.

2. **Marketing and Promotion:**

   - Movie studios and distributors can leverage genre insights to tailor marketing and promotional campaigns. Understanding the popularity of specific genres allows for more targeted and effective advertising, helping to reach the intended audience and maximize the impact of promotional efforts.

3. **Audience Segmentation:**

   - The analysis can assist in audience segmentation, helping filmmakers and studios identify and target specific demographic groups based on their genre preferences. This information is valuable for designing targeted marketing campaigns and tailoring content to different audience segments.

4. **Production Planning:**

   - Filmmakers and production houses can use genre analysis to inform their decisions on upcoming projects. Identifying trends and popular genres can guide the development of new content, ensuring alignment with audience preferences and increasing the likelihood of success.

5. **Viewer Engagement and Recommendations:**

   - Streaming platforms can enhance their recommendation algorithms by incorporating genre preferences derived from the analysis. This can lead to more personalized and accurate content recommendations, improving the overall user experience and increasing user engagement.

6. **Industry Trends and Research:**

   - Researchers and industry analysts can utilize the genre analysis to identify broader trends in the film industry. This information can be valuable for understanding shifts in audience preferences over time, predicting future trends, and conducting market research for strategic decision-making.

Conclusion :

In conclusion, the movie genre analysis project utilizing R programming and the IMDb dataset has illuminated the multifaceted landscape of film preferences. Through meticulous data extraction, cleaning, and visualization, we gained a nuanced understanding of genre distribution. The compelling bar graphs and pie charts not only revealed the popularity of specific genres but also showcased the power of data visualization in conveying complex insights. This project underscores the practical applications of R programming in data analysis, offering participants valuable skills for future endeavours. The findings hold significance for industry stakeholders, influencing content strategies, marketing approaches, and investment decisions. As the entertainment industry embraces data-driven methodologies, this project serves as a testament to the pivotal role data plays in unravelling patterns and informing strategic choices within the dynamic world of filmmaking.

References:

- Cite all the sources referenced throughout the project, following a standard citation format.

Ensure that each section is detailed enough to provide a thorough understanding of the research-oriented movie genre analysis project in R Studio.