Final Project Report

Movie Data Analysis

Submitted By:

|  |
| --- |
| *Thasneem Therodam Kandi* |
| *Amal Alappat* |
| *Lingeshwaran C* |
| *Jayalakshmi* |

Under the guidance of:

|  |
| --- |
| Shalini Kumari |
|  |

Table of Contents

[1 Introduction 4](#_Toc83550357)

[1.1 Purpose 4](#_Toc83550358)

[1.2 Background 4](#_Toc83550359)

[1.3 Data Description 4](#_Toc83550360)

[1.4 Problem Statement 5](#_Toc83550361)

[1.5 Software Requirement 5](#_Toc83550362)

[2 Data Analysis 5](#_Toc83550363)

[2.1 Data Exploration 5](#_Toc83550364)

[2.2 Data Cleaning 6](#_Toc83550365)

[2.3 Data Visualization 6](#_Toc83550366)

[2.4 Machine Learning 6](#_Toc83550367)

[3 Repository 6](#_Toc83550368)

[3.1 GITHUB Reference 6](#_Toc83550369)

[4 Conclusion 6](#_Toc83550370)

# 

## Purpose

This document describes the details of our final group project “Movie Data Analysis”.

## Background

A commercial movie is not only for entertaining the people but also for financial gain. The success of movies are also helps to thrive the industry which eventually helps to the people work in the movie industry and their dependents. A lot of factors such as good directors, actors, stories are considerable for creating good movies. However, these factors often bring an expected box office income, but cannot guarantee good movie rating.

## Data Description

The dataset is from Kaggle website. It contains

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| name | Name of the movie |
| rating | Content rating of the movie like PG, PG13, R etc |
| genre | Film categorization like ‘Animation’, ‘Comedy’, ‘Romance’, ‘Horror’, ‘Sci-Fi’, ‘Action’, ‘Family’ |
| year | The year in which the movie is released |
| released | Date and country of release |
| score | IMDB score of the movie |
| votes | Number of people who voted for the movie |
| director | Name of the Director of the Movie |
| writer | Name of the writer of the movie |
| star | Name of the main actor/actress of the movie |
| country | Country where the movie is produced |
| budget | Budget of the movie in Dollars |
| gross | Gross earnings of the movie in Dollars |
| company | The producer of the movie |
| runtime | Duration of the movie |

## Problem Statement

Based on the given information, what kind of movies are more successful than others, in other words, we would like to analyze what are the important factors that make a movie to get a high IMDB score. We also want to visualize the results.

In this project, we take score as response variable and focus on analyzing the rest of the variables in movie data.

We will be doing the data analysis and visualization in 5 various platforms, that are Python, R, SAS, Tableau and Excel. We will also use machine learning algorithm to predict the score.

## Software Requirement

|  |  |  |
| --- | --- | --- |
| **Language/Technology** | **Tools/IDE** | **Remarks** |
| Python | Jupyter Notebook | Python IDE |
| R | R Studio | R IDE |
| SAS | SAS Studio | SAS IDE online |
| Tableau | Tableau |  |
| Excel | Microsoft Excel |  |
| GitHub | GitHub | Code/File/Document Repository |

# Data Analysis

## Data Exploration

First we import all relevant packages which are required to analyse our data.

Then we read the data

## Data Cleaning

Our next step is to do the cleaning of data. We have reviewed the data and cleaned up unnecessary data. We have also removed the empty data, if that are impacting our whole data analysis. In some cases, we filled the empty data using the average values.

## Data Visualization

Here, we visualise our assumptions and finding using various graph such as histogram, line graph, box plot, bar graph, regression plots, etc

## Machine Learning

Since the score data is regressive, we use Supervised Learning. We used two models, which are linear regression and Logistic regression

# Repository

## GITHUB Reference

|  |  |
| --- | --- |
| **Item** | **Link** |
| CSV Data File | <https://github.com/ThasniTk/Edubridge/blob/main/TKRepo/Project/DataFiles/movies.csv> |
| Python | <https://github.com/ThasniTk/Edubridge/blob/main/TKRepo/Project/Python/movie_data_analysis.ipynb> |
| R | <https://github.com/ThasniTk/Edubridge/blob/main/TKRepo/Project/R/Movie%20Data%20Analysis.pdf>  <https://github.com/ThasniTk/Edubridge/blob/main/TKRepo/Project/R/MoviedataGroupProject.R> |
| SAS | <https://github.com/ThasniTk/Edubridge/blob/main/TKRepo/Project/SAS/Movie_Project.sas> |
| Tableau | <https://github.com/ThasniTk/Edubridge/blob/main/TKRepo/Project/SAS/Movie_Project.sas> |
| Excel | <https://github.com/ThasniTk/Edubridge/blob/main/TKRepo/Project/Excel/movies.xlsb> |

# Conclusion

1. Comedy, Action and Drama are mostly produced movie genre
2. United states hold the highest stake in movie production
3. Lebanon top the list of average movie score, surprisingly US not featuring in the top 10 list.
4. **Finally**, duration, number of votes and gross of the movie impact the score of a movie.