Ronan Wade

Waderonan@yahoo.ie

A picture containing clipart

Description automatically generated



Top Rated Movies

Project Report

**Table of Contents**

[1. GitHub URL 2](#_Toc73365631)

[2. Abstract 2](#_Toc73365632)

[3. Introduction 2](#_Toc73365633)

[4. Dataset 2](#_Toc73365634)

[5. Implementation Process 2](#_Toc73365635)

[5.1. Importing libraries 2](#_Toc73365636)

[5.2. Importing Data 3](#_Toc73365637)

[5.3. Cleaning up data, analysing and replacing missing values 3](#_Toc73365638)

[5.4. Sorting, indexing and grouping 3](#_Toc73365639)

[5.5. Slicing 3](#_Toc73365640)

[5.6. Iterrows 3](#_Toc73365641)

[5.7. Numpy 3](#_Toc73365642)

[5.8. Matplotlib 3](#_Toc73365643)

[5.9. Seaborn 4](#_Toc73365644)

[5.10. sns.histplot(movies['runtime']) 4](#_Toc73365645)

[5.11. sns.stripplot(x="runtime",y="genre",data=movies) 4](#_Toc73365646)

[6. Results 4](#_Toc73365647)

[7. Insights 8](#_Toc73365648)

[8. References 9](#_Toc73365649)

# GitHub URL

<https://github.com/Waderonan/UCDPA_RonanWade/tree/master>

# Abstract

The main purpose of this report is to determine the factors that contribute to a movie becoming a box office success.

I will analyse the Top 1,000 rated movies by using a dataset from Internet Movie Database (IMDb). By using this data I hope to identify whether there is a trend in the success of a movie based on gross earning, popularity, genre etc.

# Introduction

Movies have become so prevalent in the last 18 months due to the Covid-19 outbreak. People have been forced to stay home due to lockdowns and as a result the demand home entertain has increased. Entertainment app services have either increased their offering or new services have been launched including Disney+, Prime Video, Hulu and Now TV to name a few.

The reason I decided to do this project is that I have always had a keen interest in movies and have a broad interest in different types of movies. I also felt that which the increased focus on home entertainment there could be some interesting insights available.

I was intrigued to see if there were any patterns in why some movies can be very popular without being a success with respect to earnings from the box office. I am hoping that this project will also determine if there is any common dominator which dictates a movie’s success from genre, runtime etc.

I will hopefully, from my analysis, be able to identify the characteristics that attract people to a certain movie and outline any trends I come across.

# Dataset

The dataset I used to assess what factors contribute to a box office hit is:

[www.kaggle.com/preetviradiya/imdb-movies-ratings-details](http://www.kaggle.com/preetviradiya/imdb-movies-ratings-details)

I selected this dataset from Kaggle as it had all the relevant information and data that I required to conduct my analysis. The dataset had information regarding the Top 1,000 movies as by voted by the public through the IMBD website and app. I expect that this data will give a more accurate reflection of public popularity for the movies produced in the last 20-25 years due to society becoming more digitalised in this time period.

# Implementation Process

## Importing libraries

I started by uploading the libraries I would need during the datareview:

* import pandas as pd
* import matplotlib.pyplot as plt
* import seaborn as sns
* import numpy as np

## Importing Data

I downloaded my dataset from Kaggle IMDB movie rating details (CSV) and converted it to a data frame using pd.read\_csv through Jupyter. I named the dataset “movies” so it would be easier to reference:

* pd.read\_csv('https://raw.githubusercontent.com/Waderonan/UCDPA\_RonanWade/main/IMDB\_movie\_reviews\_details.csv')
* movies = pd.read\_csv('https://raw.githubusercontent.com/Waderonan/UCDPA\_RonanWade/main/IMDB\_movie\_reviews\_details.csv')

Then printed to make sure it was updated correctly:

* print(movies.head())

## Cleaning up data, analysing and replacing missing values

To make sure the data was clean and every entry had a value I performed a certain code. As I was reviewing the data, I noticed that the “gross” column had the most missing values. I used a function to replace any values with “unknown” so I could avoid any problems further along in my analysis.

## Sorting, indexing and grouping

I decided to focus on the Top 20 rated movies in this dataset to see if the most popular movies would highlight any similar attributes.

## Slicing

To understand my data more I applied the .index, .describe and .columns functions to my already defined data of “movies**”.** From the results of applying this code I understood what columns were involved and what kind of data was used (integers, float,etc)

I decided to focus on the Top 20 most rated movies to get a better understanding of why these movies were rated so highly. By using the .iloc function I isolated the data for the Top 20 movies.

## Iterrows

I further analysed the data to see if there were any common traits by applying the function iterrows.

I tried unsuccessfully to apply it to 3 attributes of the movie’s including “year”, “runtime” and “genre” to see if it would reveal any similarities between the Top 20 rated movies.

## Numpy

I used numpy to confirm if there was any common value in the “year” column of my dataset or correlational between common values which would appear in the dataset. I seemed to have trouble getting a list of common numeric values but instead got a list of data types and the whole calculation.

In this case I found numpy wasn’t the best application to be used on my dataset**.**

## Matplotlib

I used Matplolib on my full datataset to see could I find the most popular trends by using “genre” “runtime” and “year”. I wanted to see what was topping the charts in each of these categories and compare it against the Top 20 rated movies. This will show, for example, if the movies made most falling within a particular genre in the Top 20 rated movies voted by the public:

* Ranking of Movies by genre
* Ranking movies by year
* Top 10 Runtimes.

## Seaborn

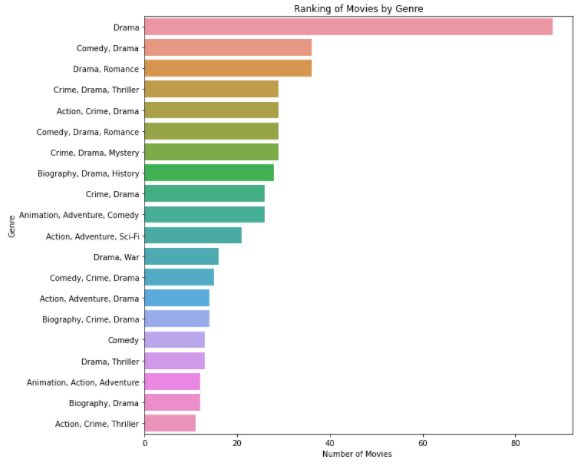
I used Seaborn to plot a histogram to show the most popular runtime being used by production companies and to determine whether the length of a movie has an effect on the popularity of the movie.

## sns.histplot(movies['runtime'])

I also plotted a chart for genre by putting limitations on the data against the Top 10 genres and then comparing it against runtime to see whether the runtime has an impact on the popularity of the movies

## sns.stripplot(x="runtime",y="genre",data=movies)

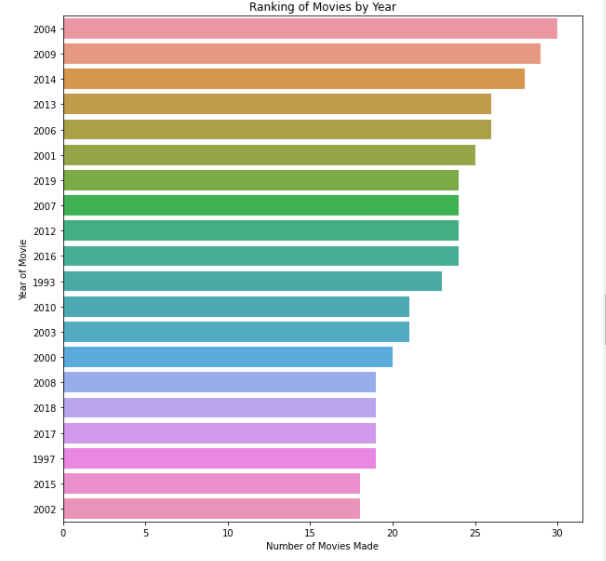
# Results



After reviewing all my data and presenting it in the easiest form to understand. You can see in the above bar chart that it is very clear that the most popular genre of movie is Drama and that over double the amount of movies fall within this genre compared to the second most popular genre. The top three genres are:

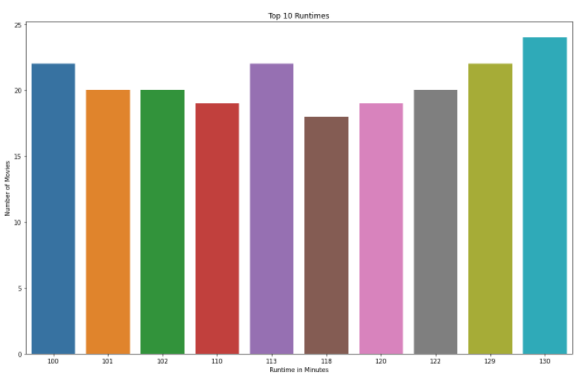
* Drama
* Comedy Drama
* Drama Romance.

This is very interesting to see as “drama” features in each of these genres and with so many issues in the world due to Covid-19 it suggests the people are seeking comfort in situations perhaps similar to there own rather than seeking an escape through a genre such as “comedy”.



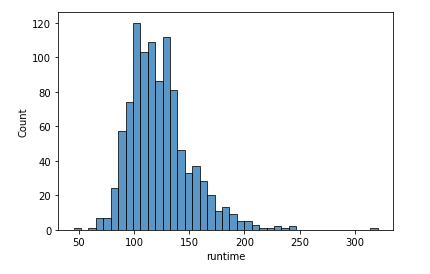
The next step I took was to use the graph above “Ranking of Movies by Year” to determine what the most popular year for movies was. You can see from the above that the year 2004 has over 30 plus movies in the Top 1,000 rated. The top three movies by year are:

* **2004 – 30 movies made**
* **2009 - 28 movies made**
* **2014 – 27 movies made.**



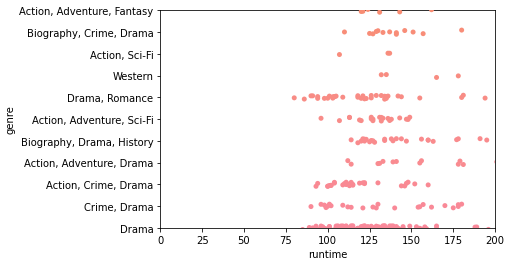
I then used the above graph for the “Top 10 Runtimes”. This demonstrates what production companies favoured in terms of movie length. You can see that 130 minutes is the most popular runtime from the section of my dataset that I examined. For this analyse I looked at the Top 250 movies and as you can see above 24 of these had a runtime of 130 minutes. The three most popular movie runtimes are:

* 130 minutes (2hrs 10 mins) –24 movies made with this runtime
* 113 minutes (1hr 53 mins) – 22 movies made with this runtime
* 100 minutes (1hr 30mins) – 22 movies made with this runtime.



I created the above histogram to determine where 130 minutes is also the most common runtime when the Top 1,000 movies are analysed. I found this to be untrue and in fact the most common runtime when the full dataset is analysed is 100 minutes. The top three runtimes in the Top 1,000 movies are:

* 100 (1hrs 30 mins) –119 movies made with this runtime
* 135(2hr 15mins) – 116 movies made with this runtime
* 130(2hr 10mins) – 114 movies made with this runtime.



I used the above scatter chart to see if there is any corelation between genre and runtime when movie producers are considering making a movie. From the graph above you can see that there is an even spread between Top 10 genres and runtimes. The data points are condensed around the most popular genre but still no defining runtime for certain genres (drama, comedy drama**,** drama romance) can be seen in the chart.

# Insights

With the information I learned from the analyse of the Top 1,000, using code and graphs, I decided to compare my information with that of the Top 20 rated movies in IMDB for my insights. The below table shows the information for the Top 20 movies:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Number** | **Name** | **Year** | **Runtime** | **Genre** | **Rating** |
| 0 | The Shawshank Redemption | 1994 | 142 | Drama | 9.3 |
| 1 | The Godfather | 1972 | 175 | Crime, Drama | 9.2 |
| 2 | Soorarai Pottru | 2020 | 153 | Drama | 9.1 |
| 3 | The Dark Knight | 2008 | 152 | Action, Crime, Drama | 9 |
| 4 | The Godfather: Part II | 1974 | 202 | Crime, Drama | 9 |
| 5 | 12 Angry Men | 1957 | 96 | Crime, Drama | 9 |
| 6 | The Lord of the Rings: The Return of the King | 2003 | 201 | Action, Adventure, Drama | 8.9 |
| 7 | Pulp Fiction | 1994 | 154 | Crime, Drama | 8.9 |
| 8 | Schindler's List | 1993 | 195 | Biography, Drama, History | 8.9 |
| 9 | Inception | 2010 | 148 | Action, Adventure, Sci-Fi | 8.8 |
| 10 | Fight Club | 1999 | 139 | Drama | 8.8 |
| 11 | The Lord of the Rings: The Fellowship of the Ring | 2001 | 178 | Action, Adventure, Drama | 8.8 |
| 12 | Forrest Gump | 1994 | 142 | Drama, Romance | 8.8 |
| 13 | The Good, the Bad and the Ugly | 1966 | 178 | Western | 8.8 |
| 14 | The Lord of the Rings: The Two Towers | 2002 | 179 | Action, Adventure, Drama | 8.7 |
| 15 | The Matrix | 1999 | 136 | Action, Sci-Fi | 8.7 |
| 16 | Goodfellas | 1990 | 146 | Biography, Crime, Drama | 8.7 |
| 17 | Star Wars: Episode V - The Empire Strikes Back | 1980 | 124 | Action, Adventure, Fantasy | 8.7 |
| 18 | One Flew Over the Cuckoo's Nest | 1975 | 133 | Drama | 8.7 |
| 19 | Parasite | 2019 | 132 | Comedy, Drama, Thriller | 8.6 |

My insights are:

* It is very interesting to see on a graph that drama is the clear top genre especially with so many issues in the world today due to Covid-19. This suggests that people are seeking drama to unwind during their time off;
* That the most popular year for movies made and voted on was 2004, 2009 and 2014 but none of these make the Top 20;
* The most produced movie runtime is 100 minutes but except for “12 angry men” which was made in 1957. Every movie in Top 20 rated movies exceeds this runtime;
* By doing this project reviewing all the data it seems that production companies do understand how people are drawn to drama as the top 3 genres (Drama, Comedy Drama**,** Drama Romance) all include it. Maybe its human nature to seek out drama and its an easy money maker for Hollywood;
* You can see from the Top 20 movies that besides all the data I reviewed for genre, runtime and year that the only common trend is Drama. That the other factors had nothing to do with creating “A great movie”.

# References

[www.kaggle.com/preetviradiya/imdb-movies-ratings-details](http://www.kaggle.com/preetviradiya/imdb-movies-ratings-details)

[JupyterLab](http://localhost:8888/lab)

<https://github.com/Waderonan/UCDPA_RonanWade>

<https://git-scm.com/downloads>