**DATA SCIENCE TOOLBOX: PYTHON PROGRAMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

***(NETFLIX DATA ANALYSIS)***

Submitted by

Kartikey Singh

Registration No: 12316555

Programme and Section B.Tech(CSE) K23EP

Course Code INT375

Under the Guidance of

**Dr. Tamina Thakur**

**Discipline of CSE/IT**

**Lovely School of Computer Science**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Kartikey Singh bearing Registration no. 12316555 has completed INT375 project titled, **“**NETFLIX DATA ANALYSIS**”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Dr. Tamina Thakur**

**Faculty**

**School of Computer Science**

Lovely Professional University

Phagwara, Punjab.

Date:

**DECLARATION**

I, Kartikey Singh, student of B.Tech CSE under 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: 12/04/2025 Signature

Registration No. 12316555 Kartikey Singh

**TABLE OF CONTENTS**

| **S. No.** | **Content** | **Page No.** |
| --- | --- | --- |
| 1 | Introduction | [Introduction](#Intro) |
| 2 | Source of Dataset | [Source](#dataset) |
| 3 | Data Pre Processing | [Data Preprocessing](#dataP) |
| 4 | Exploratory Data Analysis (EDA) Process | [EDA](#EDA) |
| 5 | Analysis on Dataset | [Analysis](#Objective) |
| 6 | Conclusion | [Conclusion](#conslu) |
| 7 | Future Scope | [Future](#FutureScope) |
| 8 | References | [Reference](#re) |
| 9 | LinkedIn Post | [Post](#LI) |

Introduction

This data science project is centred around analysing a movie dataset using Python to extract meaningful insights and patterns related to film characteristics and performance. The objective is to work with raw structured data and apply fundamental data science techniques to interpret trends in genres, ratings, revenues, and release timelines. By using Python libraries such as Pandas for data manipulation and Matplotlib and Seaborn for visualizations, the project demonstrates how raw data can be translated into understandable and actionable information.

Visualizations are used to support the findings and to present the data in a clear and informative manner. Bar charts, line plots, and distribution graphs help communicate the results effectively, making it easier to interpret trends and comparisons. Overall, this project showcases the practical application of Python in data analysis and highlights how structured programming and basic data science techniques can be used to make sense of large datasets in the entertainment domain.

Source of Dataset

<https://github.com/TheiScale/YouTube-Video-Notes/blob/main/New%20Netflix%20Data%20Analysis%20Project%202025/mymoviedb.csv>

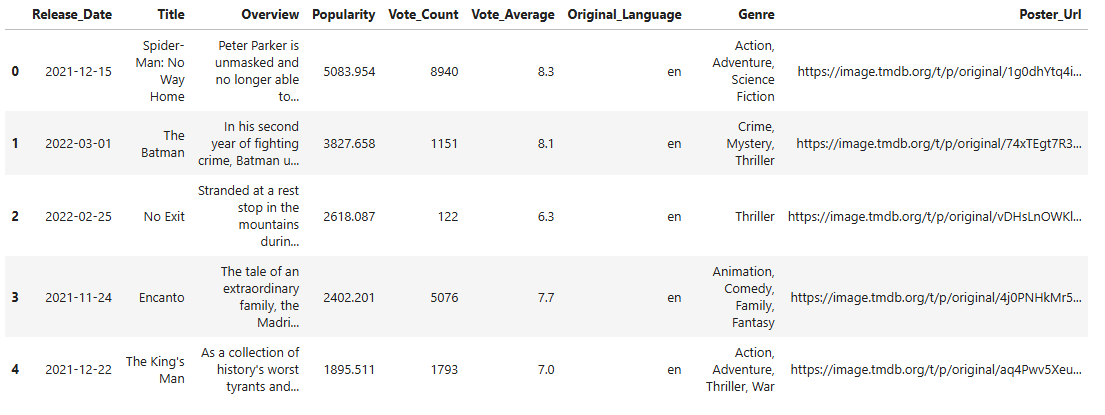
Data Preprocessing

As part of the data preprocessing process in this project, specific transformations were applied to make the dataset more suitable for analysis and visualization. One of the key modifications involved simplifying the date format. The original release date, which was in the yyyy-mm-dd format, was converted to only the year (yyyy). This helped in analysing trends over time more effectively and allowed for cleaner grouping and comparison across different years without being affected by specific months or days.

Another significant transformation focused on the Vote Average column, which represents the average rating received by each movie. To make this metric more interpretable for categorization and comparison, a custom classification was created. Ratings were segmented into four distinct popularity levels: movies with ratings in the minimum to 25th percentile were labelled as "Not Popular", those from the 25th to 50th percentile as "Below Average", from the 50th to 75th percentile as "Average", and from the 75th percentile to the maximum as "Popular". This transformation enabled a clearer understanding of how movies are distributed based on audience ratings and made it easier to visualize popularity trends across different genres and years. These steps played a key role in preparing the dataset for meaningful insights.

Another change that was made is that in genre 1 movie got multiple genre and that could have caused problems so each genre was extracted and new entries where made to.

Before Data Preprocessing



After Data Preprocessing

A screenshot of a computer

AI-generated content may be incorrect.

Exploratory Data Analysis

In this project, data analysis and visualization were central to uncovering patterns and insights from the movie dataset. Through thorough exploration, we examined trends across genres, release years, ratings, and financial metrics such as budget and box office revenue. The dataset was first cleaned and pre-processed to address missing values, normalize column names, and convert data types where necessary—for example, parsing release dates and converting numerical fields for consistency.

Data visualizations were instrumental in making these findings interpretable. Bar plots, scatter plots, histograms, and box plots were used to illustrate genre popularity, budget vs. revenue relationships, and the distribution of movie ratings across different production years. Tools like Matplotlib and Seaborn allowed us to quickly identify top-performing genres, revenue outliers, and how movie ratings evolved over time. These visual representations not only enhanced understanding but also laid the groundwork for future predictive modeling or strategic insights in film production and marketing.

Analysis on dataset

1. Analysing which type of movies are frequently made:

Introduction:

This section of the project uses data visualization techniques to analyze that which type of movies are getting produced in the market. This helps the makers to understand the current market trend.

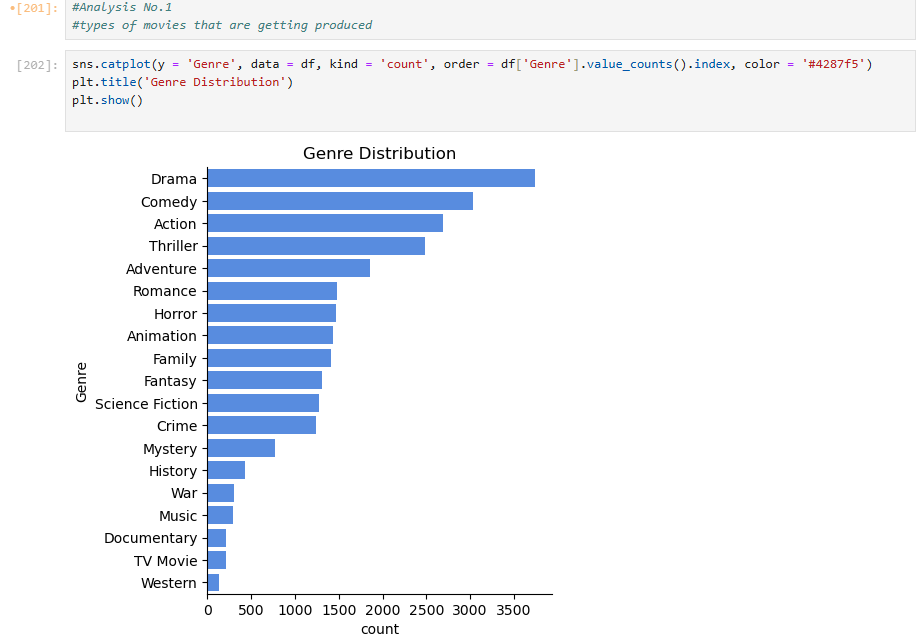
Specific Requirements:

The specific goal is to visualize that which type of movies are getting produced with the help of a column chart. This helps to understand the current trend easily.

Analysis Result:

The x-axis of the column chart shows the count of the genre, and the y-axis represents the genres.

Visualization:



1. Visualization of the Risk %

Introduction:

Sometimes people with less experience wants to produce a movie but the don’t know the risk in it. This visualization will help them to analysis the risk in make a movie.

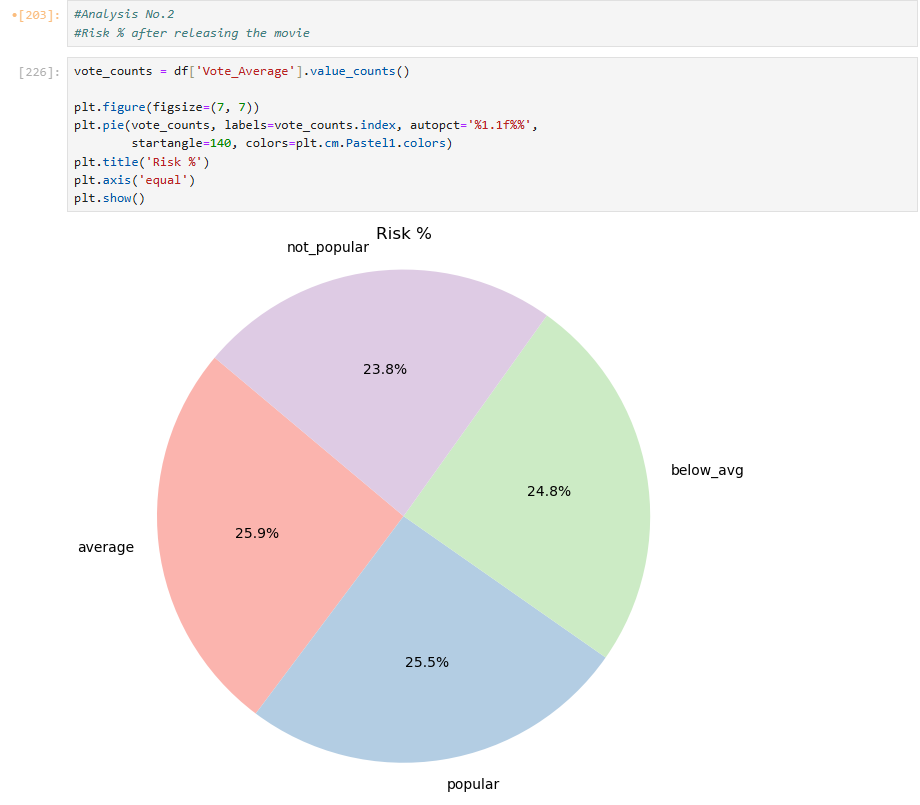
Specific Requirements:

The specific goal is to show the risk and help the maker to work perfecty on the projects.

Analysis Results:

The pie chart consist of % representation of vote average give by the user give to a particular movie.

Visualization:



1. Good movies are in which language

Introduction:

There are many movie makers who wants to explore different regions of the world and want to make move in different regions.

Specific Requirement:

This visualization will help them to see the popularity of movies with different languages and help them to decide the success rate for their projects.

Analysis Results:

The x-axis of the chart shows the languages shown in which the movies are shown and the y-axis show that how many viewers voted it as a popular movie.

Visualization:

A screen shot of a computer

AI-generated content may be incorrect.

1. Genre loved most by the public

Introduction:

Before making a movie, the makers must decide that they show make movie in which genre.

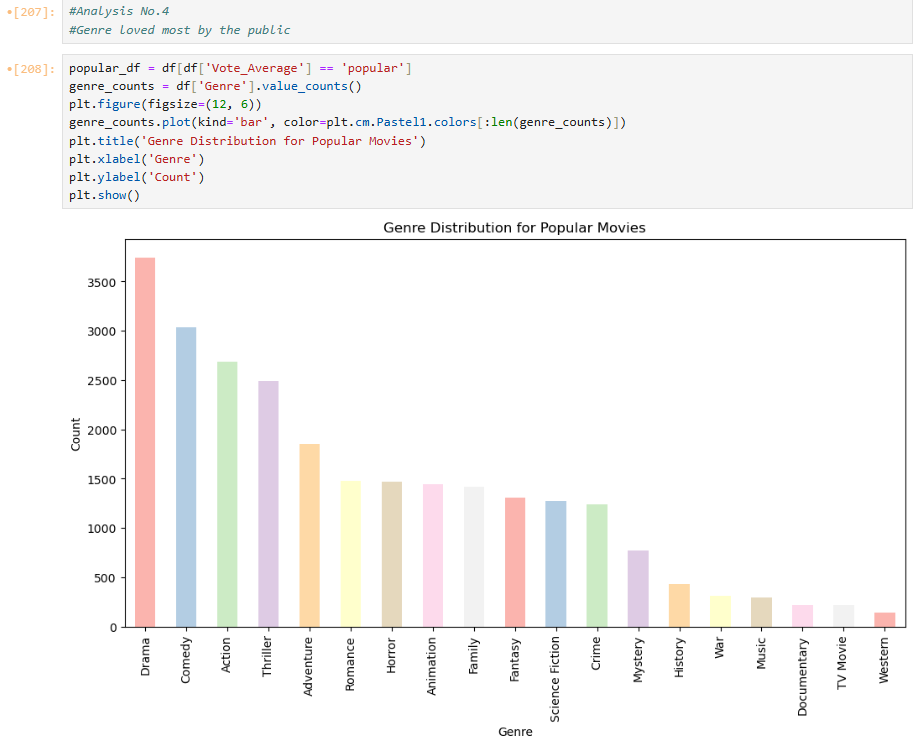
Specific Requirement:

This graph will tell us that in which genre they should produce the movie by see that what type of movie is loved by the viewers.

Analysis Results:

The x-axis of the graph shows the genres, and the y axis shows the vote counts of the public and the count is for the popular movie.

Visualization:



1. Movies released in 5 years

Introduction:

Sometime movie critics want to see the number of movies that were produced in a particular time. This helps them to make speculations about the upcoming time that how many movies might get released.

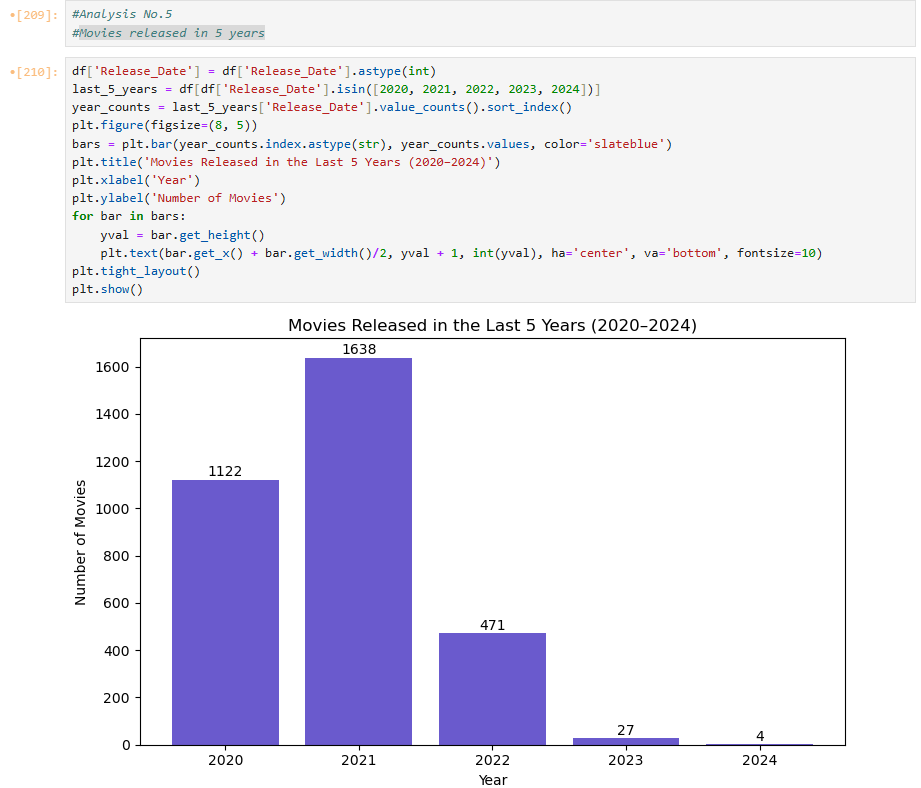
Specific Requirements:

The need for this graph is to check that how many movies where released and after seeing that they dig out the reasons what could be the reason and what could be the future results.

Analysis Results:

The x-axis of the bar chart represents year from 2020-2024 and the y-axis of the chart represents the count of the movies.

Visualization:



Conclusion

In conclusion, this project provided valuable insights into various aspects of the movie industry through effective data analysis and visualization. By cleaning and exploring the dataset, we were able to identify trends related to genres, ratings, budgets, and revenues, offering a clearer picture of what contributes to a film’s success. Visual tools like bar charts, scatter plots, and histograms made complex data more accessible and highlighted key patterns and relationships. These findings not only enhance our understanding of historical movie data but also set the stage for more advanced analytical applications such as predictive modeling and recommendation systems. Overall, this analysis demonstrates the power of data-driven decision-making in the entertainment sector and opens the door for further innovation and strategic planning.

Future Scope

The current analysis provides a solid foundation for deeper exploration and more advanced applications in the film industry. In the future, predictive modeling techniques could be employed to estimate a movie’s success based on factors like genre, budget, cast, and release date. Incorporating external data sources, such as social media sentiment or critic reviews, could enrich the dataset and offer more nuanced insights into audience preferences. Additionally, building a recommendation system based on user ratings and genre trends could enhance viewer engagement on streaming platforms. Time-series analysis could help identify seasonal trends or shifts in viewer interests over time, aiding in strategic release planning. Further, integrating geographic and demographic data may unlock more personalized marketing strategies. With continued data enrichment—such as adding director reputation, marketing spend, or award nominations—the analysis can evolve into a powerful decision-support tool for both creative and business sides of the film industry.

Reference

* Pandas Documentation: <https://pandas.pydata.org/docs>
* Matplotlib Documentation: <https://matplotlib.org/stable/contents.html>
* Seaborn Documentation: <https://seaborn.pydata.org>
* Python Official Website: <https://www.python.org>
* YouTube: <https://youtu.be/tjIWRqqMDaw?si=4sAZtKVb1d3kQSvp>

LinkedIn Post

