National College of Ireland

Project Submission Sheet

Student Name:	Vanshika Sharma		
Student ID:	23198389		
Programme:	BSc (hons) in Data Science	**Year:**	2
Module:	Data Visualization		
Lecturer:	Jaswinder Singh		
Submission Due Date:	13th December 2024		
Project Title:	Evolution of Movie Trends Over Time		
Word Count:	3132 (without references)		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Vanshika Sharma

Date: 13th December 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

Data Visualization

CA 2

Your Name/Student Number	Course	Date
Vanshika(23198389)	BSc (hons) in Data Science	13-12-2024

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
ChatGPT	**It was used to check the grammar of my paragraph.**	www.chatgpt.com

Description of AI Usage

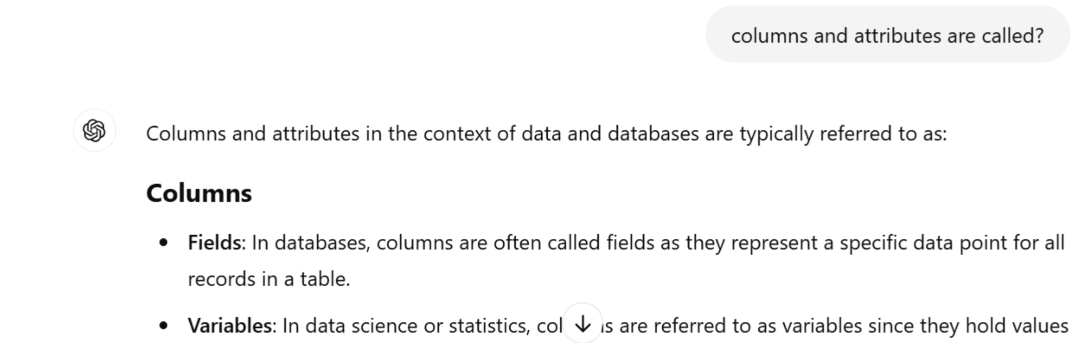
This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

ChatGPT	
It was used to check the grammar of my paragraph.	
Question: what are columns and rows called?	Answer: Features, fields,etc.

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:



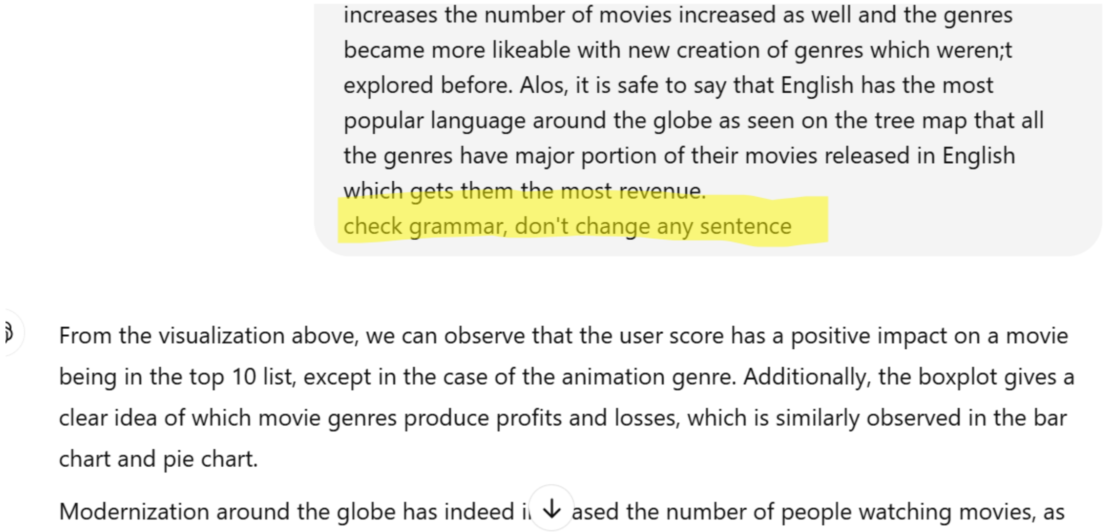
columns and attributes are called?

Columns and attributes in the context of data and databases are typically referred to as:

Columns

- **Fields**: In databases, columns are often called fields as they represent a specific data point for all records in a table.
- **Variables**: In data science or statistics, col ↓ ıs are referred to as variables since they hold values

Additional Evidence:



increases the number of movies increased as well and the genres became more likeable with new creation of genres which weren;t explored before. Alos, it is safe to say that English has the most popular language around the globe as seen on the tree map that all the genres have major portion of their movies released in English which gets them the most revenue.
check grammar, don't change any sentence

From the visualization above, we can observe that the user score has a positive impact on a movie being in the top 10 list, except in the case of the animation genre. Additionally, the boxplot gives a clear idea of which movie genres produce profits and losses, which is similarly observed in the bar chart and pie chart.

Modernization around the globe has indeed i ↓ ased the number of people watching movies, as

Evolution of Movie Trends Over Time

Vanshika Sharma
School of Computing
National College of Ireland
Dublin, Ireland
x23198389@student.ncirl.ie

I. Introduction

Some people take movies so seriously that they watch shows on a first-come, first-serve basis and are very particular about it. Movies have been a part of us since the age of growing up; they are our ultimate source of entertainment. Since movies are so close to everyone's heart, and everyone has a type of movie they can relate to or feel themselves to be the character of, therefore, the topic chosen was related to movies. The question which arose after deciding the topic was: How has modernisation made watching movie easier? Did people in the previous century spend the same amount of money, which is now in billions just to make a single movie? Also, is it worth making movies of genres which are non-popular amongst the population? To tackle all the questions which arose, datasets were gathered from Kaggle[1], a popular platform that provides publicly available data. The datasets sourced were collected through web scraping by the authors, the dataset collected has both types of data qualitative and quantitative, but the table below provides better information about them.

	Dataset1[2]	***Dataset2[3]***	***Dataset3[4]***
Name on Kaggle	NetFlix.csv	Movies.csv	FimDetails.csv
Size	2833 KB	388 KB	1038 KB
N(Fields)	12	10	22
N(Features)	7788	9719	9719
Data Types	Integer, String, Location	String, DaeTime	String, Integer, ID
File Format	csv	csv	csv
Author	Senapati Rajesh	N/A	Hassan El Fattmi
Methodology	Web Scrapping	Web Scarping	Python3, Web Scrapping

II. Objectives

Through the visualizations, the main idea was to visualize how the selected datasets provide insights. The primary goal of this visualization is to extract meaningful information from the dataset. The audience for this research includes individuals working in the film industry, students conducting research, researchers, etc. It is not limited to one group of people.

The aimed insights include identifying which movie genre is most popular and profitable, trends in user scores over time, the geographical availability of Netflix and audience distribution, the relationship between movie budgets and revenues, and identifying genres that carry high risks or losses. The report begins with the evolution of movie trends, starting with Netflix's global accessibility and audience distribution. It explores genre-specific insights and highlights classic genres. Viewers learn about the relationships between budgets and revenues, identifying profitable and high-risk genres. Users can identify the popular language in which movies are mostly released in. The report concludes by showcasing visualizations that address the questions raised.

III. Design Process

The main goal was to present a report which would strictly adhere to the description of the CA and get the answers to the questions arose which gathering the dataset. Initially, the datasets were searched through Google datasets[5], which suggested Kaggle as the source to download them from. Later, on the Kaggle platform, the datasets were evaluated to determine if they would be useful for the research purpose. Then they were downloaded and stored together in a folder named CAII_DataVisualization as while working with large datasets it is necessary to organise them into a safe place so that it is convenient and accessible. The datasets were then cleaned; empty cells were deleted. Decisions were made about which visualizations would be suitable for each dataset and which variables to use. For example, a dataset with location as an attribute was visualized using a geographical map plot. All these plots were created on Tableau Public, published, and later discussed in the report. Initially, a different topic was chosen, but since the data was AI-generated and went against the rules of the CA description, the datasets were changed to strictly adhere to the description provided by the professor. Although the professor was kind enough to appreciate the effort, I did not want to take any risks due to personal goals. After discussing with my professor, I was able to select the right files and topic. The new datasets were not AI-generated but were generated after scrapping from web with most of the authors using python as a source for scrapping since python is easy to use.

IV. DESIGN CHOICES

In this section, the discussion is mainly related to the design choices made for the analysis. The main goal was to create a visualization that was self-explanatory to conclude the results and references in the conclusion. The visualizations are presented in the next section. This section talks about why a specific visualization was chosen, the layout behind it, its styling, interactivity features, etc. The table below describes which type of visualizations have been plotted:

Type of Visualization	*Name of Visualization*	*Visualizations*	*QR Code*
Interactive Visualization	Distribution of Netflix Content by Country (Dataset1)	Geographical Map	
	Trends in Movie Rating Over Time by Genre (Dataset3)	Scatter Plot	
	Distribution of Movie Genres and Languages by User Scores (Dataset2)	Tree Map	
Static Visualization	Genre Distribution of Movies (Dataset 2)	Pie Chart	N/A
	Exploring Budget and Revenue Trends by Genre (Dataset 3)	Box Plot	N/A
	Top 10 Highest Rated Movies (Dataset 2)	Bar Chart	N/A
	Movie Counts by Different Genres Over the Year (Dataset 2)	Stacked Bar Chart	N/A

A. Distribution of Netflix Content by Country (Geographical Map)[6]

Due to the location attribute in dataset1, a geographical map was plotted as it helped to visualize the access to movies/tv shows on an online platform called Netflix. The colour palette chosen was green-teal, as it complemented the map's theme, probably because we associate Earth with green-blue shades. This palette provided clarity with vivid colours, and borders were used to highlight the map’s shape. The interactive elements included zoom and pan, allowing users to focus on specific regions or pan across the map for a broader view. The boarders of the countries were highlighted to make it understandable.

B. Trends in Movie Rating Over Time by Genre (Scatter Plot)[7]

For the analysis of trends in genres (in the case of movies), a line graph was initially used, as it would help analyse data based on the time factor. It allowed us to change the view from years to months to days, providing deeper insights into the preferences of most people. First, the line graph was later changed to a dot plot, as the dot plot explained the data better. It provided information about the genre of the movie along with user scores. The dots in the graph represent the overall user scores, and a filter allows users to change the presentation to dates, months, a single year, five years, or a custom range of dates. Different colours were used to distinguish genres, with vibrant colours ensuring they differ in appearance. The line graph was harder to interpret as it did not provide specific points, and the thickness of the lines varied. The dot plot was better suited for clarity. Its interactive elements include a time slider, allowing data to be grouped by decades or other time intervals to observe trends over specific periods. Users can also select specific data points and highlight their trajectories over time, such as the average ratings of a specific genre. Trend lines were hidden because they caused overcrowding due to the number of genres, making the visualization harder to understand.

C. Distribution of Movie Genres and Languages by User Scores (Tree Map)[8]

This visualization was created using Dataset2, and it is an interactive visualization. This visualization was chosen to display the languages in which movies are released. The idea was to check if movies are accessible to different audiences. Movies being available to various people leaves an impact, as movies integrate us by allowing people from all over the world to relate to the characters. Back to the visualization, the colour palette chosen was blue-teal to ensure consistency while also providing differentiation in the boxes for the tree map. A tree map was selected to make the visualization interactive; with a tree map, it becomes easy to retrieve details by clicking on the icons. Moreover, it allows users to categorize attributes into high-level categories or break them down into subcategories. One more interactive feature is its flexibility with filters. It allows us to use filters and visualize the data for the necessary languages. Users can pick their preferred languages.

D. Genre Distribution of Movies (Pie Chart)

We know that what answers we are looking for in our analysis, but do we know how many genres we have? This is a basic yet important factor for conducting the analysis. In the original dataset, the total genres exceeded 40, as they were combinations of many listed genres, which made the data unrealistic. A filter was applied to ensure the genres stuck to authentic, original ones and weren't any combinations. A pie chart was plotted to display the portion of each genre based on its count in the dataset, making it easy to see which genres are most liked. The pie chart also shows the percentage of each genre, making it easy to understand. The colour gradient used was a blue-teal palette to maintain consistency and to differentiate between genres clearly. The interactive feature of this visualisation is that it allows to click on the graph and get information. This was done using tooltips in Tableau, which provide brief information about the genre, its percentage, and the total user count for each genre.

E. Exploring Budget and Revenue Trends by Genre (Boxplot)

One quick question that came to mind was whether the budget and revenue generated by a movie are linked. To find answer for this question, the task was to test it but before that normality test had to be conducted. This was done by plotting the values, from which it was evident that the data was nowhere close to being normally distributed, which led us to use a non-parametric test. The colour gradient used in this analysis was a palette of green shades, ranging from light to dark, to differentiate between user scores, i.e., the ratings of the users for the movie. The size of the outliers was kept small to make the boxplot easy to understand. Initially, the visualization looked messy due to the numerous outliers, so a slider was added to filter the ranges (in USD). The interactive element included a detailed tooltip. While hovering over individual data points, users could see details such as genre, budget, revenue, and user score with exact numbers.

F. Top 10 Highest Rated Movies (Bar Chart)

An analysis related to movies done without seeing the top movies around the globe is incomplete. Therefore, a bar chart was plotted to display the top 10 movies most liked by viewers, as determined by user scores. The bar chart was chosen because it allowed various genres to be presented on a single page. Even after filtering the genres, there were still too many to fit into one visualization. Bar charts are easy to read, and with the sorting feature they provide, they were the most suitable option. The colours used were from the blue-teal palette, as this beautifully distinguished genres. The interactive features of this visualization included dynamic colour coding and the ability for users to sort the bars dynamically by different metrics.

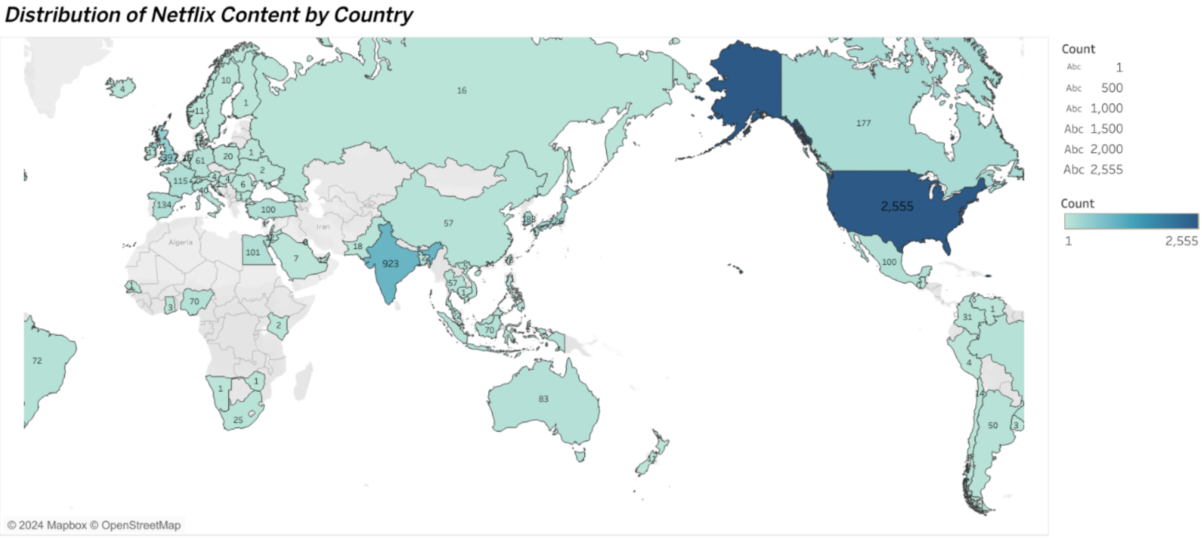
G. Movie Counts by Different Genres Over the Years (Stacked Bar Chart)

The stacked bar chart was chosen as it was easy to visualize the number of movies for a given year period briefly, as it provided the counts cumulatively. The colour palette chosen consisted of bright colours that were distinct from each other, mainly due to the large number of genres present. The interactive feature of the visualization is enabled by Tableau's ability to hover over any point on the chart, providing details such as user count, year, and genre. The stacked bars were kept in translucent mode so that one bar doesn't hide the other.

V. Development of design solution

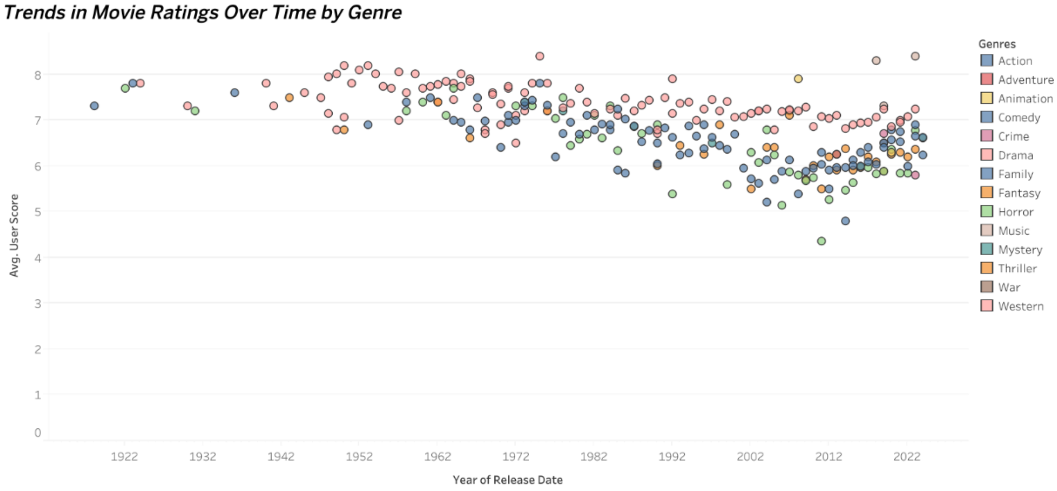
All the seven visualisations have been created using Tableau platform because of its simplicity and feasibility. Tableau, being an open-source tool, allows us to create many different visualisations all at once and makes it easy to change the data type or visualisation type

A. **Visualization 1**


Distribution of Netflix Content by Country

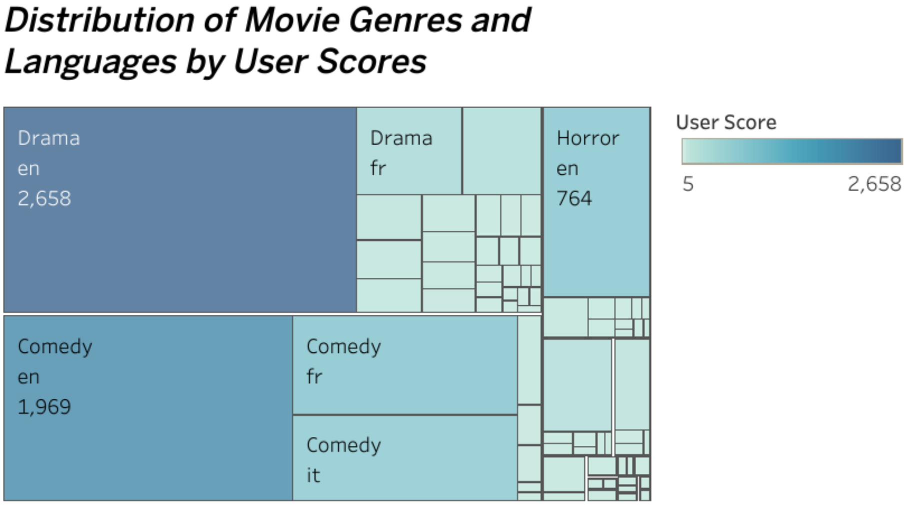
This is the first graph has been created using Dataset1. This visualisation shows where Netflix content is available. Netflix is a movie & TV show streaming platform which is popular and a unicorn company. From the graph, we can interpret that the streaming platform is available in most countries, except those in the continent of Africa. Also, the visualization provides the number of people watching Netflix, with the metric measured in millions. The solid fill colour for each country varies in shade, as seen in the visualization. The number on each country represents the count of user score count. The intensity of colour is determined by the user score for each country for movies and TV shows. The darker shades represent a higher density of people watching, with the United States being the highest, followed by India.

B. **Visualization 2**


Trends in Movie Ratings Over Time by Genre

This visualization has been created using Dataset2. From the graph, we can observe many outliers which can be seen from the deviations. The x-axis represents the years and y-axis represent average user score for movies of different genres. We can see that user scores are mostly clustered in recent years. In recent years, there are many user scores with a wide variety ranging from low to high. However, for years before 1952, user scores are fewer and only exist for a limited number of genres, indicating that very few genres existed at that time. Gradually, over time, the number of genres increased. It can be observed that the genres Drama, Action, and Horror are among the oldest genres and remain consistent and very popular, meaning that other genres are newly created. Over time, the classic genres have gained even more popularity, and a significant number of people have started watching movies.

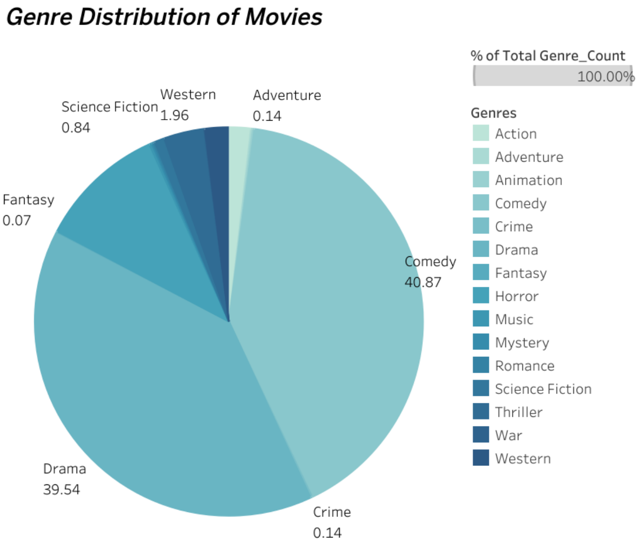
C. **Visualization 3**



Distribution of Movie Genres and Languages by User Scores

The tree map represents the languages of movies based on user scores. This is an interactive visualization, which allows users to click on the blocks and view details about the language, genre, and user score. The size of the blocks is based on the user score. Different languages are presented, such as English, French, etc. The intensity of each block is determined by the user score. Since it's a tree map, there is no x-axis or y-axis. From the visualization, we can observe that movies are versatile as they are released in many languages. We can see that the maximum number of movies are released in English, primarily in the Drama genre, followed by Comedy in English, and then Horror in English. The second most popular language is French, followed by Italian. The small boxes represent the least popular languages, such as Japanese and Russian.

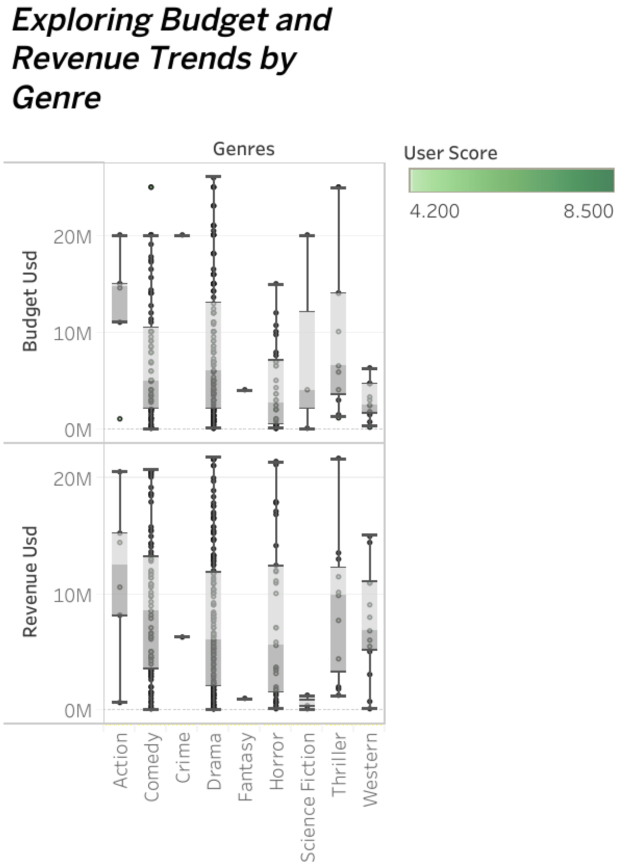
D. **Visualization 4**



Genre Distribution of Movies

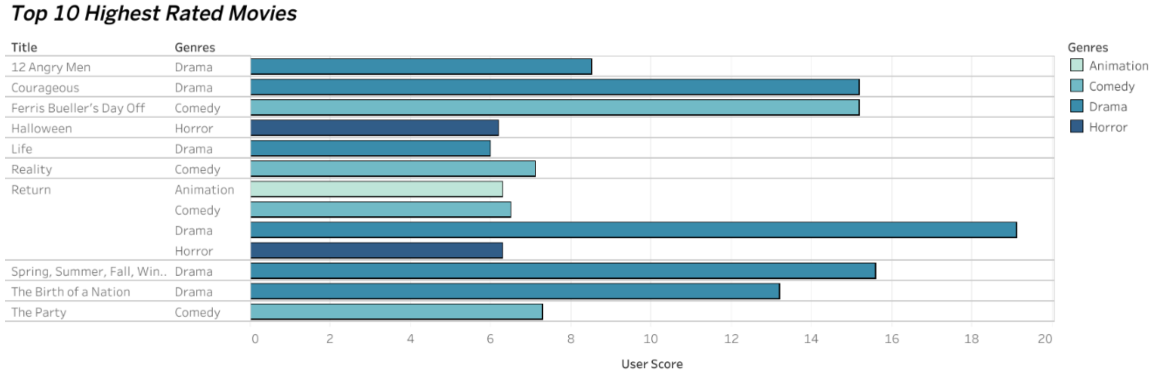
From the above visualization, it is easy to understand which genres are popular and which are less relevant. It's a pie chart, and as such, it doesn't have an x-axis or y-axis. Instead, there are checkboxes with colors representing each genre. We can observe that Comedy is the most popular genre (40.87%), followed by Drama (39.54%), and then Horror (10.57%). The least popular genres are Music (0.21%), Fantasy (0.07%), and Crime (0.14%). But the question is if this also determine the budget that should be assigned to a movie with a less popular genre? Moreover, can we conclude from these that less popular genres are irrelevant and movies with such genres are unsuccessful? To answer these questions, we will need to examine the visualization described below, which will be further discussed in the conclusion. This is a static visualization.

E. **Visualisation 5**



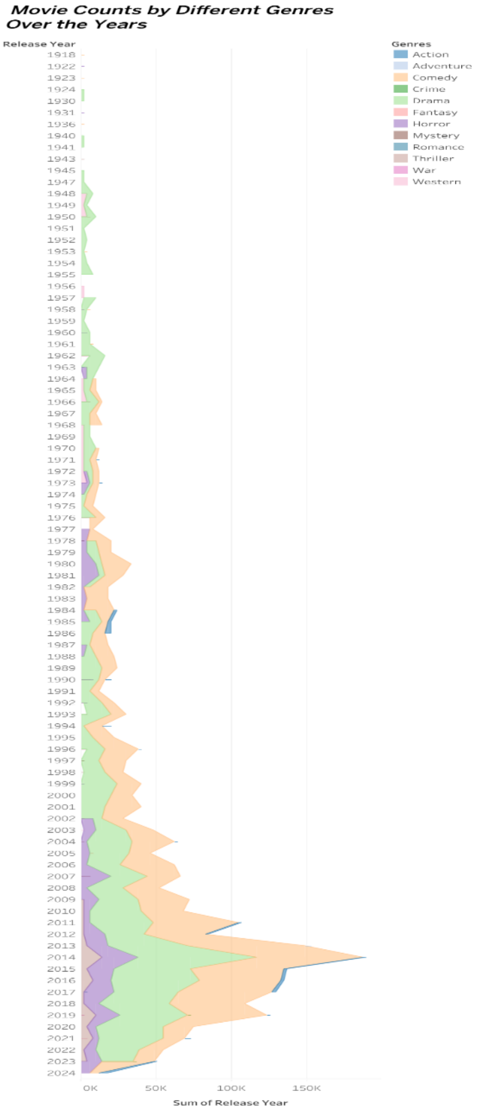
The fourth visualization explores the relationship between revenue generated and budget allocated for movies, categorized by genre. The x-axis represents the genres, and the y-axis has a dual representation of Revenue (USD) and Budget (USD). Genres such as Drama and Comedy are high-revenue genres with high budgets, indicating a potentially high-risk factor. Action and Horror movies appear to be the most profitable genres, while Thriller seems to follow a similar trend. The genre with the most loss is Science Fiction, which has a high budget but almost negligible revenue. For genres like Crime and Fantasy, there is little to observe, as they are not very popular.

F. **Visualization 6**



In this visualization, we can see the movies most liked by people, along with their titles and genres, providing insight into which genres are popular. The x-axis displays the user scores, which were used as a numerical attribute to determine the length of the bars in the chart. A filter was applied to show only the top 10 movies. The y-axis consists of the titles of the movies and their genres. We can observe that Drama is the most liked genre, with '12 Angry Men' being the most popular movie among all. The last movie on this list is 'Life', which also belongs to the Drama genre. For most of the movies in the top 10 list, it is evident that the most liked genres among people are Drama, Horror, Comedy, and Animation. One movie, ‘Return’, has all these genres, meaning it is a combination of many genres. Moreover, it is the only movie in the list with Animation genre.

G. **Visualization 7**



Through this visualization, we can clearly see that the highest level of movie count over the years is for the Action genre, followed by Adventure, and then Comedy. The genre with the least movie count is Thriller. The y-axis represents the years, and the x-axis represents the movie counts. The genres included in the graph are Crime, Drama, Fantasy, etc. For the years until the early 2000s, the movie count was low, but it increased significantly from 2000 to 2024. It peaked in the year 2014, indicating a high count of movies released. Additionally, during the period 1918–1945, the only movies released were of the Drama genre.

VI. CONCLUSION

Modernization around the globe has indeed increased the number of people watching movies, as seen in the scatter plot and geographical map. An article by Rogers[9] justifies the fact that modernization has brought people together and increased the global craze for movies. From the visualizations above, we can observe that the user score has a direct relationship with a movie being a hit movie. This is supported by a study done by Barrio[10] reveals that there is a direct influence of content on box-office revenues, meaning that the content liked by most of the population has a higher success rate than less popular genres. Additionally, the boxplot gives a clear idea of which movie genres produce profits and losses, which is similarly observed in the bar chart and pie chart. Also, with the stacked bar chart, it is easy to predict that as time increases, the number of movies increases as well, and the genres become more likable with the creation of new genres that weren’t explored before. Also, it is safe to say that English is the most popular language around the globe, as seen on the tree map, where all the genres have a major portion of their movies released in English, which generates the most revenue.

REFERENCES

[1] "Find Open Datasets and Machine Learning Projects | Kaggle." Accessed: Oct. 21, 2023. [Online]. Available: https://www.kaggle.com/datasets

[2] S. Rajesh, "Latest Netflix TV shows and movies." Accessed: Dec. 12, 2024. [Online]. Available: https://www.kaggle.com/datasets/senapatirajesh/netflix-tv-shows-and-movies

[3] "Daily Global Box Office Performance of Movies." Accessed: Dec. 12, 2024. [Online]. Available: https://www.kaggle.com/datasets/thedevastator/daily-global-box-office-performance-of-movies

[4] H. El Fattmi, "Which movie should I watch today?" Accessed: Dec. 12, 2024. [Online]. Available: https://www.kaggle.com/datasets/hassanelfattmi/which-movie-should-i-watch-today

[5] "Dataset Search." Accessed: Oct. 14, 2023. [Online]. Available: https://datasetsearch.research.google.com/search?src=0&query=flight%20prices&docid=L2cvMTF0eGxsM2c2dA%3D%3D

[6] "VanshikaSharmaCADV1 | Tableau Public." Accessed: Dec. 13, 2024. [Online]. Available: https://public.tableau.com/app/profile/vanshika.sharma7307/viz/VanshikaSharmaCADV1/Sheet1?publish=yes

[7] "VanshikaSharmaCADV2 | Tableau Public." Accessed: Dec. 13, 2024. [Online]. Available: https://public.tableau.com/app/profile/vanshika.sharma7307/viz/VanshikaSharmaCADV2/Sheet1?publish=yes

[8] "VanshikaSharmaCADV3," Tableau Public. Accessed: Dec. 13, 2024. [Online]. Available: https://public.tableau.com/app/profile/vanshika.sharma7307/viz/VanshikaSharmaCADV3/Sheet6

[9] A. Rogers, *Cinematic Appeals: The Experience of New Movie Technologies*. Columbia University Press, 2013. doi: 10.7312/roge15916.

[10] P. Barrio, "Do movie contents influence box-office revenues?: Applied Economics: Vol 49, No 17." Accessed: Dec. 13, 2024. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00036846.2016.1223828