Assignment 2 - Group 10

Ocean Wang Student ID 5689694
Onno van Wijngaarden Student ID 5313236
Nicholas Wu Student ID 5677815
Julia Schramm Student ID 6556809
Jaap de Ruiter Student ID 5408466

ABSTRACT

This design proposal outlines an analytics system that examines global and regional trends in Netflix movies and TV shows to understand how content consumption has evolved over time. Using datasets from Kaggle and IMDB, we analyze factors such as genre performance, budgets, viewership, and ratings. The goal is to identify which types of content perform best in specific regions and periods, helping filmmakers and streaming platforms make informed, data-driven decisions. Through interactive visualizations, timeline analyses, geographic overviews, and detailed comparisons of individual titles, this proposal presents a structured approach to uncovering meaningful insights into shifting viewer preferences and emerging industry trends.

1 DESCRIPTION OF TASK

Our main question, leading the later proposed design is:
Which identifiable trends or variables are there that filmmakers can use to increase viewership of films and TV shows on Netflix?
In an increasingly digitalized and globally connected world, it may seem as though content consumption has become uniform across regions. But is that truly the case? To answer this, we focus on a specific niche of global media consumption - movies and series on Netflix - and analyze how viewing preferences differ across regions and how these patterns have evolved over time.
Our goal is to identify trends and variables that filmmakers can use to optimize the success of their productions. For instance, a filmmaker aiming to create a bestseller in the USA might ask: What features should my movie have? Similarly, industry professionals may want to understand regional trends, identify which genres perform best in certain countries, or determine the most successful films and series within specific categories. Questions such as which genre generates the highest ratings or views, why certain movies succeed in particular markets (e.g., due to cast, studios, movie budget), and how much revenue each genre produces are central to this investigation.
The proposed approach involves analyzing both global and regional trends in Netflix's catalogue, exploring how content consumption varies across countries, and tracking changes over time. By enabling interactive selection of countries and examining timelines of ratings, and viewership preferences, we aim to uncover insights into the types of content that resonate most strongly in specific regions and periods. These insights can help filmmakers determine when where and why it is most effective to release particular movies or shows and reveal long-term shifts in viewer behaviour.

2 DESCRIPTION OF THE DATA

1. https://www.kaggle.com/datasets/utkarshx27/movies-dataset In this dataset the revenue and budget of movies is shown

2. https://developer.imdb.com/non-commercial-datasets/ The IMDB database is the most comprehensive database on movies and series, with some open source data capabilities. From this database multiple datasets can be accessed, ranging from data on movies and TV series to data on individual actors.

3. https://www.kaggle.com/datasets/konradb/netflix-engagement-report This dataset is used to determine the viewership of the specific movie/series.

3 RELEVANT ATTRIBUTES:

With the following attributes from the datasets, we can implement the visualizations in section 5. The Datasets we used are all of type Table and for each attribute, we also list the corresponding data type.

- **Viewership**. Equal to **Hours Viewed** divided by **Movie Duration**, found in the Netflix Engagement Report dataset. *Data type: Numerical (continuous).*
- **Genre**. The viewership and genre per movie are found in the IMDB database, which we can average ourselves. *Data type: Categorical (nominal, multi-label).*
- **Release year**. The release year can be found in the IMDB database. *Data type: Numerical (discrete).*
- **Country**. The country of origin is found in the IMDB database. *Data type: Categorical (nominal).*
- **Language**. The language is found in the IMDB database. *Data type: Categorical (nominal).*
- **Movie budget**. The movie budget is found in the Kaggle database. *Data type: Numerical (continuous, currency).*
- **Movie duration**. This is found in the IMDB database. *Data type: Numerical (discrete, minutes).*
- **Actor ratings**. In the IMDB database we can find the lead actors of the movie and their ratings. By averaging the ratings of the lead actors, we obtain the actor rating of the movie. *Data type: Numerical (continuous, score).*
- **Success factors of movies**:
  - **Movie ratings** (IMDB database). *Data type: Numerical (continuous, 0–10 scale).*
  - **Box office numbers** (Kaggle database). *Data type: Numerical (continuous, currency).*
  - **Revenue** (Kaggle database). *Data type: Numerical (continuous, currency).*

4 DATA AND TASK ABSTRACTION

4.1 Data abstraction

Dataset Type: Tables (joined on Movie ID from Kaggle, IMDB, Netflix Engagement datasets) **Attribute Types**:

- *Keys*: Movie ID, Title (identify items)
- *Quantitative*: Viewership, Budget, Revenue, Ratings, Duration (ordered, continuous)
- *Categorical*: Genre (multi-valued), Country, Language (nominal)
- *Temporal*: Release Year (ordered, discrete)

Derived Attributes: Viewership (Hours Viewed / Duration), Actor Ratings (average of leads), genre/region/time aggregations

4.2 Task abstraction

We map our domain questions to abstract tasks:

- **"What features should my movie have to succeed in USA?"**
  Filter (USA) → Identify (successful movies) → Correlate (success factors)
- **"What are trends in a specific region?"**
  Locate/Filter (region) → Identify trends (temporal patterns 1995-2025)
- **"Which genre performs best in which region?"**
  Filter (region) → Summarize (by genre) → Compare → Identify (extremum)
- **"Why do certain movies succeed?"**
  Identify (outliers) → Lookup (attributes) → Correlate (success factors)

Core Operations: Identify (trends, outliers), Compare (distributions across regions/genres/time), Correlate (attribute relationships), Filter (by region/genre/time), Summarize (aggregate metrics), Locate (find items), Lookup (retrieve details)

Typical Workflow: Summarize (global map) → Locate/Filter (select region) → Identify trends (timeline) → Correlate (scatter plot) → Compare (radar chart)

5 Description of Proposed Design

Our design follows the workflow identified in Section 4: users begin with a global summary, filter to regions of interest, identify temporal trends, and correlate success factors. Each visualization addresses specific abstract tasks: the world map enables **Summarize** and **Filter** operations, the timeline supports **Identify** (trends over time), and the deep dive view facilitates **Correlate** (scatter plot) and **Compare** (radar chart) tasks for detailed movie analysis.
We want to implement the following visualizations:

1. **Global Overview**:
   This visualization introduces the analytics with an interactive world map enhanced with glyphs. Each glyph represents a country and summarizes key attributes at a glance, allowing for the comparison of regions and the identification of patterns across the globe. This serves as a starting point and gives a high-level overview of the current situation across the globe. Depending on how many countries/regions will be meaningful to look at, the world map might be replaced by just glyphs being shown for each region. For example, 1 glyph can be shown for each continent. This might be more readable in the case that the world map does not add much use.

2. **Timeline View**:
   This visualization shows how key movie metrics evolve from 1995 to 2025. It tracks three averages: movie length, budget, and genre popularity, and reveals long-term trends at a glance. The region toggle allows filtering by country, making it possible to compare how industry developments differ across regions. It will also be possible to select multiple countries/regions in order to compare the change in movie industry between them. The line chart is a rather effective one-dimensional mark that enables to see the development over time. Overall, it provides a clear and high-level view of changes in the film industry over time.

3. **Deep Dive View**:
   This section provides a detailed view on specific movies of one or more regions using two interactive charts. The scatter plot shows the relationship between the budget and a selectable quality metric, such as rating or revenue. Each dot here represents a movie or series. It helps reveal patterns such as whether higher budgets correlate with better performance and highlights outliers. The colors indicate the genre of the movie. The radar chart compares selected titles from the scatter plot across attributes such as actor ratings, movie length, viewership, budget, and revenue, allowing quick comparison of general profiles. Together, these charts support deeper exploration of individual movies or series across multiple dimensions. The scatterplot also allows for selecting multiple movies by "drag and drop". This then returns an average radar graph of these instances. This could, for example, be used to drag and drop the top left corner of the scatterplot, to show what might make a low budget movie successful.

   The timeline view also adjusts when a selection is made in the deep dive view. This way the timeline view is only shown for the selection, and the change of this selection over time can be discovered. The other way around will work too, when a selection between years is made in the timeline view, the deep dive view will also show all the movies from this period.

   To add more ways to look into data in the deep dive, some selections will be possible. For example, it will be possible to select a specific actor, a specific director, only movies of a certain length, etc., and then the deep dive will show just those movies. Furthermore, it will be possible to compare multiple radar plots as there is a possibility of showing them overlap.

4. **Alternative ideas**:
   We were also discussing the possibility of comparing movies based on where their budget went to, see figure 3. This could create a pie chart that might be useful as a glyph representing a movie. However no useful database has been found for this yet.

   Another idea of a glyph that might give a useful representation of a movie is the colour bar-code combined with a line that shows the plot progression of a movie, see figure 3. The colour bar-code is the average colour of each scene of a movie combined together. The plot progression might be measurable by the intensity of the sound during the movie. However, to implement this, it would probably be necessary to have the downloaded file of a movie, and would therefore not be feasible/legal for more than a few movies.

   An extra idea that might be added if there is time, is an interactive part of our tool where a radar-graph similar to the one shown in the deep dive section can be drawn. This then returns the expected revenue/rating of a movie with those aspects. This will allow directors to get an idea of how successful their movie will be depending on what they focus on, for example should they focus on getting good actors, or should they choose a certain genre/movie length. A regression model would have to be created on our data to be able to predict this.

6 Design Sketches and Prototype

We made a Canva project where we visualized how we want to portray the data. https://moviemakeranalytics.my.canva.site/ Examples of these visualizations are shown in figure 3. Some intermediate ideas are shown in figure 5.

Figure 1: Global Distribution Map

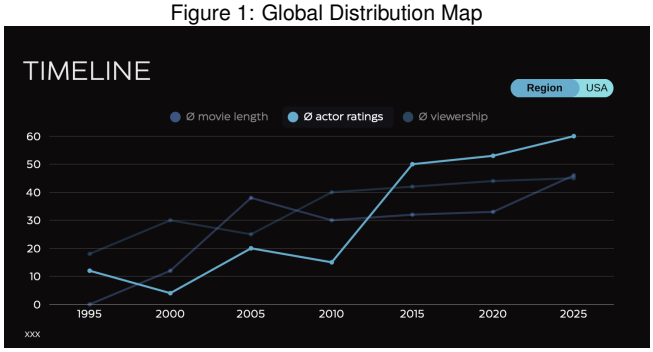

Figure 2: Timeline Overview

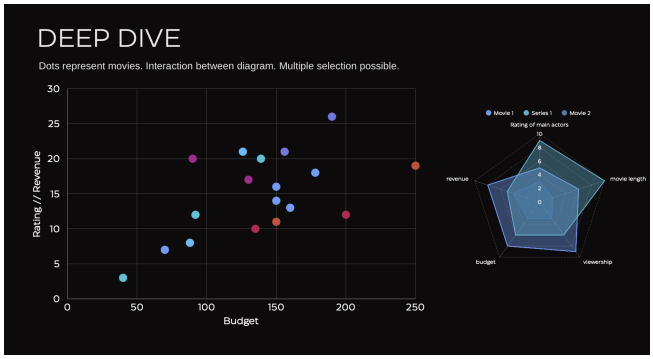

Figure 3: Movie Deep Dive

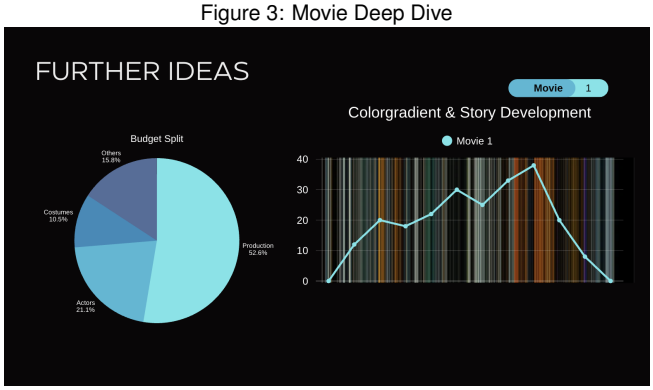

Figure 4: Further Ideas

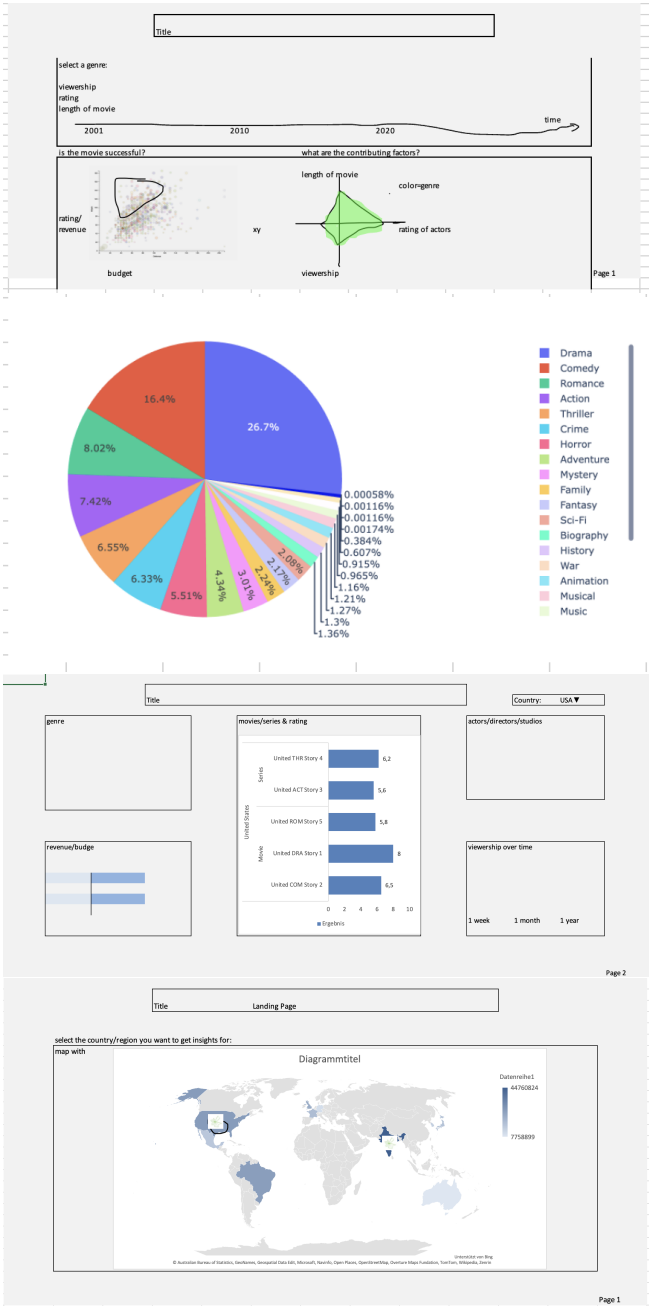

Figure 5: Previous Design Drafts