DV-VA Final Project Draft: A Comprehensive Visualization System for The Movies Dataset

Group #1: 0716021 張家豪、0816137 王培碩、0816153 陳琮方

In our final project, we would like to build a **comprehensive visualization system** for **The Movies Dataset** from Kaggle.
First, we will introduce the dataset that we are going to visualize, describe each attribute and their distribution. Then, we describe what goals we want to achieve with our visualization. Lastly, we provide a draft of the interface of our design.

A. Dataset: The Movies Dataset

This dataset from Kaggle contains metadata for over 45,000 movies released between 1974 and 2017. The dataset is collected from The Movie DB (TMDB) and the GroupLens research group. Below, we list and describe the attributes that we consider useful. For each movie, we have:

title	`string` the title of the movie	revenue	`number` The total revenue of the movie
genres	`list` The list of genres that this movie belongs to, each item in the list is an object with the id and name of genre	vote_average	`number` The vote (rating) of the movie from TMDB
budget	`number` The total budget for the production of the movie	vote_count	`number` The number of votes that contributed to the average above
popularity	`number` The popularity of the movie from TMDB	keywords	`list` A list of keywords regarding the movie plot, each keyword has an index
date	`date` The release date of the movie	casts	`list` A list containing cast information
countries	`list` The countries that took part in the production	crew	`list` A list containing crew information
companies	`list` The list of companies in charge of the production, each item in the list is an object that contains the ID and name of the company		

This dataset also contains 26 million ratings from 270,000 users. For each rating, we have:

movieId	`number` The movie id that this rating corresponds to
rating	`number` The rating ranging from 1 to 5
timestamp	`number` When this rating was given

B. The Goal of Our Project

We would like to build a **dashboard** that enables users to: **(1)** Get a **big picture** of the movies during a period of time **at a glance (2)** Easily **compare trends and distributions** between different subsets **(3)** Dive deep by analyzing all of the movies by each actor/director **(4)** Figure out the **connections and relations** between each entity (movies, actors)

C. The Interface Design

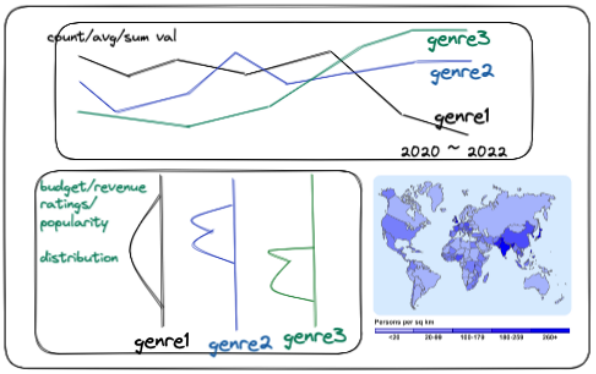
We have designed four pages to meet our goals. In this section, we illustrate what components and visualizations would be put on each page and their goals respectively.

I. Dataset Overview Page (Goal 1)

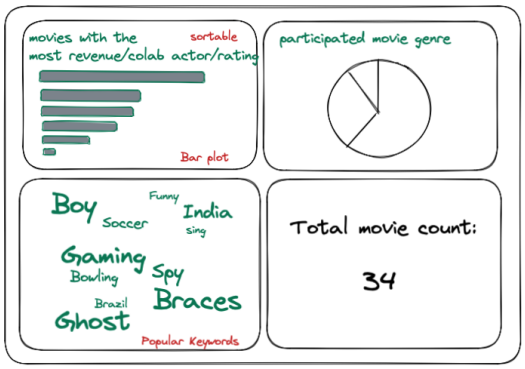
For our overview page, a **word graph** is shown on the top left, which indicates the most frequent plot keywords among all movies. The **line chart** on the top right indicates the number of movies produced each year. The two **bar charts** in the middle are used to show the **top** information of the dataset. One to show the actors/directors with the most movie counts, the other to show the movies that are the most popular or bring the most revenue. The **histogram** on the bottom left shows the number of movies within each revenue/budget bucket, which lets users observe the distribution. The **map** on the bottom right shows the number of movies produced by each country.
A sidebar is provided which enables users to create filters to visualize a subset of the dataset. Users should be able to filter out data through brushing or clicking on the elements.

II. Subset Comparison Page (Goal 2)

In this page, users could select different subsets (e.g., genre) to visualize and compare. The **line chart** on top shows the number of movies of each genre produced each year. The **violin plot** on the bottom shows the distribution of budgets/revenues/ratings of each genre, which could be visually compared.
Also, we will draw different attributes' distributions of each subsets on the worldmap using different colors.



III. Single Entity Analysis Page (Goal 3)



This page allows us to explore **detailed information and statistics** of a specific actor or director. On the **bar chart**, users can choose to view the top movies of the entity, sorted by either revenue or rating. On the **pie chart** on the top right, the pie chart shows the percentage of his/her works in different genres. The **word graph** shows the most frequent keywords in his or her works; we could know the representative word for this actor/director. Lastly, we show the total number of movies by this director or actor.

IV. Connection and Relation Page (Goal 4)

This page aims to present the connections and relations between each entity (movies, actors) through an **interactive relationship diagram**. Each node represents a movie or actor, and the edge could represent as keyword, cast, or movies. Movies positioned in nearby areas on the graph indicate they're highly related. If a user is interested in some of the nodes, he/she could even dive into specific nodes and look up the details.