**BFI Idea**

At the beginning of this project, two ideas stood out to me for this small scale demonstration:

1. Developing an approach to gathering and link relevant external data to the dataset.
2. Overhaul the reporting format to make it more engaging and accessible.

The approach to the first idea was retrieving data from an open source platform, ‘**The Movie Database’ (TMDB),**  using their developer API. The data from this would enhance the dataset to allow users to have a better understanding of the films they are being presented with, the information retrieved would be:

1. Movie Poster
2. Movie Description
3. Genres
4. Trailer
5. Runtime
6. Rating
7. Release date

After researching, I discovered that IMDB’s Box Mojo is where the grossing information is retrieved, the approach to scrap the data from the website, store in a data frame.

The approach to the latter is the use of [Streamlit](https://streamlit.io/) to create a dashboard which would provide the information above as well as the gross values changes through a visualisation.

**How it works**

The demonstration operates in two main sections:

1. **Data Retrieval and Processing:**
   * **Source Data Extraction:**
     + The dataset is imported from a CSV file (bfi-weekend-box-office-report-2024-08-09-11.csv) that contains the weekend box office data for films in the UK.
     + Data is cleaned to extract relevant columns such as "Film," "Weekend Gross," and "Total Gross to Date." Rows with missing or irrelevant values are removed.
   * **Web Scraping:**
     + The code scrapes Box Office Mojo's website to fetch the latest weekend gross figures for films in the UK by entering the date you wish to retrieve grossing information.
     + It collects film-specific URLs and scrapes details for the top 15 films, including gross revenue, weekly changes, and other performance metrics.
   * **Data Enrichment with TMDB API:**
     + Additional movie details are retrieved using The Movie Database (TMDB) API. This includes movie posters, genres, overviews, trailers, ratings, and release dates.
     + The enriched data is compiled into a new DataFrame and exported to CSV files (gross.csv and movie\_data.csv) for visualization.
2. **Visualization and Reporting:**
   * **Streamlit Dashboard:**
     + A Streamlit app is created to provide a dynamic user interface. Users can select movies from a dropdown menu and see detailed movie information such as descriptions, ratings, and trailers.
   * **Data Visualization:**
     + The dashboard uses Matplotlib and Seaborn to generate visualizations like line and bar charts, showing the box office figures over time and the weekly changes.

**How it was built**

1. **Data Collection and Preparation**
   * Python’s Pandas library was used to handle data extraction, cleaning and preprocessing.
   * Web Scraping was performed using  **requests** library and  **BeautifulSoup** to collect data from **Box Office Mojo.**
   * The **TMDB API** was utilised to fetch additional metadata about the films.
2. **Visualisation and Reporting** 
   * The dashboard was created using **Streamlit,** an open-source app framework to enable interactivity and ease of use.
   * Visualisation Python libraries, **Matplotlib and Seaborn**  were used to build graphs to provide insights into the box office data.

**What Value it adds**

1. **Enhanced Data Completeness**: By integrating external data( e.g, movie descriptions, genres, trailers) from TMDB and Box Office Mojo, the dataset becomes more comprehensive allowing for richer insights.
2. **Improved User Engagement:** The interactive dashboard provide a more engaging and accessible way for users to explore the data.
3. **Information Decision-Making:** Visualisations such as weekly gross changes and total earnings, offer deeper insights and support data-driven decision-making.

**How you would develop it further**

1. **Automation Data Pipeline:** By building the process flow into functions, this would allow automation process, that could be scheduled for every start of the week, reducing the amount of time spent on working on pushing the weekly reports.
2. **Integration of Additional Data Sources:** For this project my decision to use TMDB’s API over other sources was it was readily available. To improve the service, additional sources from IMDb, Rotten Tomatoes could be included to provide user reviews, critic scores, for a holistic view of movie performance.
3. **Advanced Visualisations:** The implementation of more visualisations such as heatmaps, bubble charts, radar graphs to providing correlations between movie features and patterns from the additional sources.

The use of

1. **User authentication:** By adding user authentication allows customised reports and dashboards based on user preferences, this aspect allows further interaction with the website providing a better insight into each customer providing a better approach to each.