### **Project Report: The-Numbers.com Movies Analysis**

#### **Agenda**

This project aimed to conduct a comprehensive analysis of over 1000 movies, exploring how key metrics—such as IMDb ratings, worldwide box office performance, profitability, genre, and production characteristics—influence a movie's success. The objective was to derive meaningful insights into trends and correlations in the movie industry, such as which genres generate the highest box office revenue, which regions and production companies perform best, and how IMDb ratings impact perceived movie quality.

#### **Workflow**

To achieve the project objectives, the following steps were taken:

1. **Data Collection**:
   * A **Selenium bot** was built to automatically navigate The-Numbers.com and extract data for at least 1000 movies. Since The-Numbers.com uses dynamic content with embedded frames, collecting this data required a carefully structured bot capable of loading multiple pages and dynamically locating elements on the page.
   * Complex **XPath paths** were crafted for each element, as traditional indexing wasn’t feasible. Many data points were nested within tags, requiring detailed XPaths to precisely target elements.
   * To streamline the collection process across different pages, the bot was programmed to open new tabs for each page, allowing for efficient multi-tab data scraping.
2. **IMDb Ratings Retrieval**:
   * IMDb ratings were sourced via **API calls** using a search-based approach to match movie titles with their IMDb data, as IMDb provides ratings data linked to unique movie IDs. This required careful title matching to accurately retrieve ratings and integrate them with the primary dataset.
3. **Data Structuring**:
   * After data collection, the data was organized into dataframes using **Pandas** and **NumPy**. This step involved converting data into arrays, preparing it for data manipulation and analysis.
   * The main dataframe included essential columns such as budget, worldwide gross, IMDb rating, genre, production company, and region of release.
4. **Data Transformation**:
   * **Handling Missing Values**: Missing values, particularly in budget and gross earnings columns, were managed using a **Random Forest-based imputation** method. This predictive approach uses patterns in other data columns to estimate missing values, offering a more accurate fill than simple mean or median substitution.
   * **String-to-Numeric Conversion**: Financial columns containing currency symbols (e.g., "$") were transformed into numeric values using regular expressions, preparing the data for numerical analysis.
   * **Addressing Data Quality Issues**: Some cells had values like "$0" in key financial columns, which were converted to NaN using regex. These NaN values were later imputed using the Random Forest method to maintain data consistency.
5. **Data Visualization and Analysis**:
   * The analysis was visualized using **Matplotlib**, **Seaborn**, and **Power BI**. Various plots and charts were created to capture and convey insights into the relationships between genres, profitability, regional performance, and ratings. Key visualizations included:
     + **Histogram**: A histogram of IMDb ratings provided insights into the general distribution of movie ratings, identifying the most common rating range.
     + **Correlation Matrix**: Displayed correlations between IMDb ratings, profitability (worldwide gross - budget), and genre as a categorical factor, highlighting genre-specific trends in ratings and revenue.
     + **Bar Graphs**: Created multiple bar graphs to analyze various aspects:
       - Worldwide gross by genre, segmented by top-performing countries.
       - Profits in high-performing countries, classified by production methods (e.g., live-action, animation).
     + **Line Chart**: Illustrated the top 20 countries with the highest overall performance in terms of movie earnings, showing market trends on a regional level.
     + **Geographic Mapping**: In Power BI, a geographic map was used to display the distribution of top-performing countries, highlighting dominant genres and the top production companies in each region.
     + **Pie and Donut Charts**: Analyzed the relationship between genre performance and source, providing insights into the popularity of sources like original screenplays and adaptations from fiction.
     + **Clustered Column Chart**: Highlighted the earnings of top-rated movies segmented by both source and genre.
     + **Box Plot**: Showed IMDb rating distribution by source and genre to identify outliers and rating variability within categories.
     + **Table and Bubble Chart**: Represented the overall movie distribution across genres, sources, and IMDb ratings, offering a summary of categorical breakdowns.

#### **Conclusion**

The analysis provided several significant insights:

1. **Genre Correlations**:
   * Drama was found to be highly correlated with IMDb ratings, suggesting that audiences often rate dramas favorably. In contrast, the adventure genre showed a strong correlation with profitability, likely due to the widespread appeal and visual appeal of these movies across international markets.
2. **IMDb Ratings Distribution**:
   * The histogram showed that most movies receive IMDb ratings between 6 and 7, indicating a central tendency for moderately favorable ratings, likely due to the industry's high volume of average-performing movies.
3. **Regional Performance**:
   * Action movies tend to perform best in **China**, reflecting a strong market preference for action-packed blockbusters. Additionally, both **China and the United Kingdom** emerged as major markets for movies with live-action and animated production methods, highlighting the regional affinity for diverse content types.
4. **Leading Production Companies**:
   * The analysis identified **Walt Disney Pictures, Warner Bros, Marvel Studios, Universal Pictures, Paramount Pictures, and DreamWorks** as the most successful production companies globally. These companies consistently produce high-grossing films, indicating robust production and marketing capabilities.
5. **Genre and Source Popularity**:
   * Action and adventure emerged as the top-performing genres, while the most successful sources for movies were original screenplays, followed by adaptations from fiction and comic novels. This trend highlights the enduring appeal of original ideas and popular literature in driving box office performance.

#### **Difficulties Encountered**

The project faced multiple challenges due to the complexity of the dataset and tools involved:

1. **Bot Development**:
   * Building a Selenium bot to handle complex, dynamically loaded content required precise handling of frames and elements across multiple pages. Using Selenium’s multi-tab functionality allowed more efficient data extraction from different pages.
2. **XPath Customization**:
   * Writing accurate XPath selectors was essential due to deeply nested HTML tags and non-standard indexing, adding complexity to the bot-building process.
3. **Multi-Tab Data Collection**:
   * Managing data collection from multiple tabs allowed for a more streamlined approach but required additional logic to handle tab-switching in Selenium.
4. **Chart Data Extraction**:
   * Extracting performance data for each country required parsing chart visuals with regex, which was challenging but necessary to obtain accurate region-specific insights.
5. **IMDb Rating Retrieval**:
   * IMDb’s API required movie IDs for rating retrieval, which necessitated an additional search-based matching method to ensure that ratings were accurately integrated with each title.
6. **Handling Placeholder Values**:
   * Placeholder values like "$0" in key columns were replaced using regex. The Random Forest model was then used to impute realistic values, preserving data integrity for analyses involving financial metrics.
7. **Random Forest Model Training**:
   * Training the Random Forest model required extensive preprocessing to handle categorical data, as the model can only interpret numeric inputs. This step was crucial to achieving accurate predictions for missing values in budget and box office columns.