**UGANDA CHRISTIAN UNIVERSITY**

**DATA MINING AND WRANGLING ASSIGNMENT**

**TEST TWO**

**MOVIE DATA WEB SCRAPING**

**MEMBERS**

| **NAME** | **ACCESS NUMBER** | **REG NUMBER** |
| --- | --- | --- |
| **Namatovu Liz Treasure** | **B20854** | **S23B38/015** |
| **Charles Kidega** | **B24691** | **S23B38/006** |
| **Lorraine Paula Arinaitwe** | **B20729** | **M23B38/004** |
| **Mwiine Irwin** | **B20814** | **M23B38/008** |

**Movie Data Scraping and Analysis Report**

**1. Introduction**

This report presents the findings from scraping movie data using the OMDB API. Our objective was to extract key movie details, analyze trends, and identify correlations with IMDb ratings, including the most common movie genres, the average IMDb rating of the dataset, and which parameters correlate the most with IMDb ratings.

**2. Why Use an API?**

An API (Application Programming Interface) allows structured data retrieval without the need for complex web scraping. Web scraping involves extracting data directly from websites, which can be unreliable due to frequent changes in web page structure and anti-scraping mechanisms. We chose the OMDB API because:

* Structured Data Output: It provides data in JSON format, making it easy to parse and analyze without needing complex HTML extraction.
* Reduced Risk of Blocking: Unlike web scraping, which can violate website terms of service, APIs are designed to provide data access within specified limits.
* API requests return only the needed data, reducing processing time compared to extracting data from entire web pages.
* Consistency: The API ensures a standardized format, reducing errors that can occur with web scraping due to website layout changes.

**3. Packages Used Several Python libraries were utilized in this project:**

* **requests**: To send HTTP requests and retrieve movie data from the OMDB API.
* **pandas**: For storing, organizing, and manipulating the structured data in a DataFrame.
* **matplotlib & seaborn:** Used to create graphs and charts to visualize trends in the dataset.
* **Numpy:** For numerical computations such as finding averages, correlations, and statistical summaries.

**4. Data Collection Process**

1. **API Requests:** The OMDB API was queried using movie titles to retrieve details such as genre, director, cast, IMDb rating, and box office revenue.
2. **Data Storage:** The retrieved JSON data was converted into a structured Pandas DataFrame, making it easy to manipulate and analyze.
3. **Exporting Data:** The DataFrame was saved as a CSV file, which allows easy storage, sharing, and future reanalysis of the data.
4. **Data Cleaning:** Missing or inconsistent values were identified and handled appropriately, ensuring that the dataset was ready for analysis.
5. **Exploratory Data Analysis (EDA):** Various statistical and visualization techniques were applied to uncover patterns and insights in the dataset.

**5. Data Exploration and Analysis**

* **Most Common Movie Genres:** The dataset revealed that Drama, Action, Adventure, Comedy, and Thriller were the most frequently occurring genres, showing their dominance in the film industry. A bar chart was used to visualize the distribution of genres.
* **Average IMDb Rating:** The average IMDb rating of the dataset was calculated as 8.31 (to be computed from the dataset), giving an overview of the general reception of the selected movies. A histogram was used to show the spread of IMDb ratings across different movies.
* **Trends Observed:**
  + Genre Trends: Drama and Action movies appear more frequently, reflecting their popularity among filmmakers and audiences. A bar chart was used to show how often each genre appeared.
  + Movie Runtime Effect: Longer movies tend to receive better IMDb ratings, possibly due to more detailed storytelling and character development. A scatter plot was used to visualize the relationship between runtime and IMDb rating.

**6. Correlation Analysis**

A correlation analysis was conducted to determine which parameters have the highest impact on IMDb ratings. Key findings included:

* Box Office Gross vs. IMDb Rating: A positive correlation was observed, suggesting that commercially successful movies tend to receive better audience reception. A scatter plot was used to visualize this trend.
* Runtime vs. IMDb Rating: Movies with longer runtimes generally had higher IMDb ratings, implying that longer movies allow for deeper storytelling. A scatter plot was used to illustrate this relationship.
* Genre vs. IMDb Rating: Certain genres, such as Drama and Sci-Fi, had higher IMDb ratings on average. A bar chart was used to compare IMDb ratings across genres.

**7. Visualization Techniques**

To better understand the dataset, several visualization methods were employed:

* **Bar Charts:** Used to represent the frequency of different movie genres and compare their average IMDb ratings.
* **Histograms:** Displayed the distribution of IMDb ratings, showing how ratings are spread across different movies.
* **Scatter Plots:** Illustrated relationships between numerical variables such as runtime and IMDb rating, revealing trends and outliers.
* **Heatmaps:** Displayed correlation matrices, helping to identify strong and weak relationships between factors such as box office earnings and IMDb rating.

**8. Key Findings**

* **Box Office Performance:** Movies that perform well at the box office generally receive higher IMDb ratings, indicating audience satisfaction and popularity. A scatter plot was used to compare box office revenue and IMDb ratings, showing a positive correlation.
* **Genre Influence**: Certain genres, such as Drama and Sci-Fi, tend to have higher average ratings, possibly due to their strong storytelling and immersive world-building. A bar chart was used to compare the average IMDb ratings of different genres.
* **Director Impact:** Renowned directors like Christopher Nolan and Steven Spielberg consistently produce highly rated movies, showing that an experienced director can significantly influence a film’s success. A scatter plot was used to compare IMDb ratings across different directors.

**Conclusion:**

This project successfully extracted and analyzed movie data, providing insights into the key factors influencing IMDb ratings and box office performance. The use of an API proved to be an efficient and structured method for retrieving movie data, reducing errors and inconsistencies associated with traditional web scraping. Visualizations played a crucial role in uncovering trends, making it easier to interpret relationships between various movie attributes.

Future improvements could include:

* Expanding the dataset with additional movie parameters such as audience reviews and critic scores to enhance analysis.
* Applying machine learning models to predict IMDb ratings based on historical data.
* Incorporating audience sentiment analysis from social media and review platforms to gain deeper insights into movie reception.