**Streaming Services Project**

Erik R. Lopez

**Purpose of the Project:**

This project aims to identify trends and patterns in the extensive data collected from various streaming platforms. By leveraging Microsoft SQL Server and Jupyter Notebook (Python), the goal is to pinpoint popular titles and genres within and across platforms. Specifically, I want to determine which platforms people use to watch different genres, and what types of genres and titles are most preferred overall. Additionally, I aim to identify the genres and titles that are popular among different generations (Gen Alpha, Gen Z, Millennials, and Gen X) to see if there are generational differences in movie or TV show preferences. Another focus is to explore what type of content is preferred by different countries compared to the U.S. Insights from this data will help stakeholders understand viewing preferences on their platform, identify which genres are favored by specific generations, and guide decisions on creating original content or acquiring licenses to target specific audiences.

**Limitations**

The dataset has several limitations that impact the comprehensiveness of the analysis:

1. **Lack of Gender Data:**
   * The absence of gender data restricts the ability to analyze genre and title preferences by gender across platforms and countries. This makes it difficult to understand how different genders engage with various types of content.
2. **No Data on Likes and Dislikes:**
   * Without data on likes and dislikes for content, it is challenging to gauge viewer enjoyment and to potentially avoid releasing content that receives significant negative feedback. This limits the ability to assess the audience’s reception of the content accurately.
3. **Missing Sales Data:**
   * The lack of sales data for each title hinders insights into the revenue generated by different movies or TV shows. Sales data would provide a clearer picture of the financial success and profitability of content, which is crucial for strategic decision-making.
4. **Age Group Trends:**
   * In my analysis, I missed examining trends and patterns across different age groups. This could have been addressed by filtering the data for years corresponding to different generational cohorts, which would have provided more detailed insights into the preferences of various age groups over time.

These limitations highlight the need for more comprehensive data collection, including gender information, viewer feedback, sales records, and detailed age group data. Addressing these gaps would significantly enhance the analysis and provide a more holistic view of audience preferences and content performance.

**Analysis Process**

To gain insights, I used several tools, including Microsoft Excel, Microsoft SQL Server, Looker Studio, and Jupyter Notebook. The datasets, obtained from Kaggle, contained data from Netflix, Hulu, HBO Max, Amazon Prime, and Apple TV. Initially, the files were cleaned by removing duplicates, with imdbId serving as the primary key, and by deleting rows with blank cells. Given the large volume of data, removing these rows did not significantly impact the analysis but rather improved its accuracy. Lastly, I removed a few rows that had titles that were not considered actual movies or TV shows; some of them were in the Amazon Prime original dataset, but I removed them when going over the cleaned dataset and sampled dataset.

Subsequently, the five datasets were imported into Microsoft SQL Server, where a column was added to indicate the platform (Netflix, Hulu, HBO Max, Amazon Prime, Apple TV) from which the data originated. This step was crucial for combining the datasets and simplifying the analysis. After identifying patterns and trends, the compiled dataset was exported as a CSV file and imported into Jupyter Notebook (Python) for further detailed analysis.

In Jupyter Notebook, I was concerned that the differing number of rows between each dataset could impact the regression and coefficient analysis. To overcome this, I executed a random sampling script to select data from each dataset to match the size of the dataset with the lowest number of rows. For instance, instead of Amazon Prime having close to 60,000 rows of data, a new file was created containing a little over 5,000 rows to match Hulu. Once this was done, these files were then downloaded into a CSV file, combined into a new file with all datasets, and re-uploaded into Jupyter to run regression and coefficient analysis.

Finally, visualizations were developed in Looker Studio to display some of the identified trends and patterns. These visualizations were prepared to be presented to stakeholders, providing a clear and comprehensive understanding of the analyzed data.

**Methodology Explanation**

During the initial analysis phase, I utilized Microsoft SQL Server to gather insights from the cleaned dataset, which combined multiple large datasets from different platforms to provide a comprehensive view of the data. For instance, the original files downloaded from Kaggle contained data from Netflix (18,685 rows), Amazon Prime (59,505 rows), Apple TV (16,041 rows), Hulu (8,637 rows), and HBO Max (5,314 rows). Given the varying sizes of these datasets, it was essential to balance them to avoid potential biases in the detailed analyses conducted in Python using Jupyter Notebook.

To achieve this balance, I employed a random sampling technique. This involved selecting a random subset of data from the larger datasets to match the size of the smallest dataset, HBO Max with 5,314 rows. By doing so, I ensured that each platform was represented equally, allowing for a fair and unbiased comparison across different platforms. This approach mitigated the risk of skewed results due to disproportionate data sizes and provided a robust foundation for accurate and reliable statistical analysis.

By maintaining consistency in the dataset size, I could produce more precise visualizations and perform detailed regression and coefficient analyses with greater confidence in the findings. This method of equalizing the number of rows across these datasets also helps in mitigating issues like outliers, ensuring a balanced comparison. The most effective method involved using scripts that randomly select data from each dataset to match the size of the smallest dataset.

Through this balanced approach, the datasets became more representative and less prone to biases, thereby facilitating more accurate and meaningful analysis. This comprehensive methodology allowed for the development of visualizations in Looker Studio to display the identified trends and patterns, providing stakeholders with clear and actionable insights.

**Dataset Columns:**

* title – the title of the content displayed on the streaming service.
* type – the type of content (movie/tv).
* genres – the list of genre(s) the content falls onto (e.g., Drama, Action, etc.).
* releaseYear – the year the content was released.
* imdbId – the primary key that identifies the content.
* imdbAverageRating – the average rating given to each title.
* imdbNumVotes – the number of votes given to each title.
* platform – the streaming platform that the title was shown on (Netflix, Hulu, HBO Max, Amazon Prime, and Apple TV).
* availableCountries – the country or countries that have had access to streaming the title.

**Major Insights from the Streaming Services Data Analysis**

Based on the various statistical analyses conducted using Microsoft SQL Server, Jupyter Notebook, and Looker Studio, the following five major insights were identified. These findings can provide stakeholders with a deeper understanding of the data collected from multiple streaming platforms.

1. **Popular Movies Across All Platforms**:
   * The top five movies consistently popular across all platforms are **The Dark Knight, The Lord of The Rings: The Fellowship of the Ring, The Lord of The Rings: The Return of the King, The Shawshank Redemption**, and **Inception**. This indicates that over the years, the genres of **Action, Adventure**, and **Drama** have remained highly favored among viewers.
2. **Generational Preferences**:
   * **Generation Alpha's** top five titles are **Inception** (action, adventure, sci-fi), **Game of Thrones** (action, adventure, drama), **Interstellar** (action, drama, sci-fi), **The Dark Knight Rises** (action, drama, thriller), and **Django Unchained** (comedy, drama, western). *Understanding these preferences helps platforms create and promote content that resonates with younger audiences.*
   * **Generation Z's** top five titles are **The Dark Knight** (action, crime, drama), **Inception** (action, adventure, sci-fi), **Fight Club** (drama), **Game of Thrones** (action, adventure, drama), and **Breaking Bad** (crime, drama, thriller). *Recognizing the favorites of Generation Z can guide content strategy and acquisitions targeting this demographic.*
   * **Millennials'** top five titles are **The Shawshank Redemption** (drama), **Forrest Gump** (drama, romance), **Pulp Fiction** (crime, drama), **Se7en** (crime, drama, mystery), and **The Silence of the Lambs** (crime, drama, thriller). *Insights into Millennial preferences are valuable for understanding the enduring appeal of classic dramas and thrillers.*
   * **Generation X's** top five popular movies are **The Godfather** (crime, drama), **The Godfather Part II** (crime, drama), **The Shining** (drama, horror), **One Flew Over the Cuckoo’s Nest** (drama), and **Alien** (horror, sci-fi). *Knowing Generation X's preferences helps in curating classic and nostalgic content that appeals to this age group.*
3. **Platform Quality Ratings**:
   * Among the platforms, **HBO Max** has the highest average ratings, followed by **Hulu, Netflix, Apple TV**, and **Amazon Prime**. This suggests that HBO Max generally offers higher-quality content compared to the other platforms, with Amazon Prime rated the lowest.
4. **Improvement in Content Quality Over Time**:
   * The quality of content has shown a linear increase over time, as indicated by the rising average ratings and the number of votes titles receive each year. This trend can be attributed to two factors: the significant increase in movie releases in the 2000s compared to the early to mid-1900s, and the decreased demand for older movies.
5. **Geographical Distribution of Streaming Titles**:
   * The **United States** leads in the number of titles streamed, followed by **Japan**. The most popular content across these countries includes **The Shawshank Redemption, The Dark Knight**, and **Inception**. Specifically, in the United States, the dominant genre over the years has been **Drama**, highlighting a consistent preference for this type of content.

These insights can help stakeholders better understand audience preferences and trends across different platforms and demographics, enabling them to make informed decisions about content creation and acquisition.

**Recommendations**

Based on the analysis of trends and patterns, several key insights emerged:

1. **Popular Genres**:
   * Content falling under the genres of **Drama, Action,** and **Adventure** are notably popular across different countries. This indicates a universal appeal of these genres, which can guide content strategies.
2. **High-Rated Genres**:
   * When averaging rating scores, genres such as **Family, Talk Shows, Animation,** and **Drama** consistently rank among the top. This highlights the potential for success in these genres beyond just movies and suggests opportunities for diversification.
3. **Top Performing Movies**:
   * Movies like **The Dark Knight, Inception, Game of Thrones, The Shawshank Redemption,** and similar titles consistently receive high average ratings and a large number of votes. This suggests that focusing on content within these genres can attract significant viewership and engagement.
4. **Cross-Country Appeal**:
   * The data shows that the popularity of genres such as **Drama, Action,** and **Adventure** extends beyond the U.S. to other countries. This cross-country appeal underscores the importance of these genres in global content strategies. However, it's important to note the limitation due to the lack of likes and dislikes data, which means these conclusions are based on available ratings and votes.
5. **Recommendations for Content Strategy**:
   * To appeal to a broad audience across generations and countries, companies should focus on releasing movies and shows in genres such as **Drama, Action, Adventure, Horror,** and **Crime**. These genres have shown consistent popularity and high ratings.
   * To enhance the analysis and decision-making process, it is recommended to request additional data such as sales records for released titles, likes and dislikes for each type of content, and the frequency of content availability on platforms. This would provide a more comprehensive understanding and enable more precise coefficient and regression analyses to account for variables impacting audience preferences.

These recommendations aim to guide stakeholders in creating and promoting content that aligns with audience preferences, thereby maximizing engagement and viewership across different platforms and demographics.