**IFT 598: Data Visualization & Reporting for IT**

Sushmitha Alagesan

**Information Technology, Arizona State University**

**Project Phase 3: Dashboard Implementation**

Dr. Asmaa Elbadrawy

Due date: 23 April 2023

# 

**Section 1: The Dashboard**

**The figure of the dashboard:**

The below screenshot represents the figure of the dashboard:

A picture containing chart

Description automatically generated

**Explanation of dashboard:**

With the advent of the internet and communication revolution, almost every human has a smartphone. This led to several businesses moving online. One such industry is the entertainment industry, which has adopted this mode of delivery extensively. It is evident from the sheer number of OTT (Over the top) platforms that are on the rise. Further the e-commerce giant Amazon has not backed up from the competition by introducing their very own OTT platform Amazon Prime.

This dashboard is primarily designed to provide useful actionable insights for the end users to extract information regarding the myriad of content available on this platform. The dashboard can provide insights into the Amazon TV and Movies shows that were released since 1913 to 2023. The dashboard is aimed at answering various questions related to the distribution of content across different genres, age certifications, production countries, and release years. It also looks at the top N countries and actors based on various criteria such as content production and IMDB score. Additionally, it provides a comparison of TV shows' popularity with seasons across different factors.

Overall, the dashboard is used for exploratory data analysis and to derive insights from the Amazon TV and Movies shows dataset. The insights generated from the dashboard can be used by content creators, producers, and other stakeholders to make informed decisions about content production, marketing, and distribution.

**Section 2: The Dataset**

**Explanation of the dataset:**

## Dataset Chosen: Amazon Prime Movies and TV Shows

The dataset is available on Kaggle.

**Kaggle**: <https://www.kaggle.com/datasets/dgoenrique/amazon-prime-movies-and-tv-shows>

This dataset, which details the American movies and TV episodes that are currently accessible on Amazon Prime, was compiled from Just Watch in March 2023. It consists of two CSV files: "titles.csv" and "credits.csv". We have merged the data from these two CSV files and used them in our project.

* **title.csv** - The "titles.csv" file contains more than 10k titles with 15 columns. These columns include the title ID, name of the title, show type (movie or TV show), a brief description, release year, age certification, runtime (for TV shows, the length of the episode), genres, production countries, number of seasons (if it's a TV show), IMDB ID, IMDB score, IMDB votes, TMDB popularity, and TMDB score.
* **credits.csv** - Over 140k credits of actors and directors are included in the "credits.csv" file, which has five columns that include the person ID on Just Watch, the title ID on Just Watch, the actor or director's name, the character name, and the role (either ACTOR or DIRECTOR).

This dataset offers insights into the different kinds of content that are available on Amazon Prime, including genre and runtime distributions, popular titles, and details on the producers—such as the actors and directors—who worked on the project.

**Dataset Attributes:**

The dataset title.csv has the following columns and are described as follows:

**1. Id**

* The datatype of the Id is Ordinal.
* The domain is tm100001 to ts99878.
* The variable type is the string.
* The column ID represents the title ID on just watch.

**2. Title**

* The datatype of the title is Categorical.
* The domain is None.
* The variable type is the string.
* The column title represents the name of the title within the dataset.

**3. Type**

* The datatype of the type is Categorical.
* The domain is None.
* The variable type is the string.
* The column type represents whether a particular title is a movie or TV show.

**4. Description**

* The datatype of the type is Categorical.
* The domain is None.
* The variable type is the string.
* The column type represents a brief description of the movie or a TV show.

**5. release\_year**

* The datatype of the release\_year column is Ordinal.
* The domain is from 1912 to 2023.
* The variable type is the integer.
* The column release\_year represents the year in which the movie or TV show is released.

**6. Age\_Certification**

* The datatype of the age\_certification column is Categorical.
* The domain is None.
* The variable type is the String.
* The column age\_certification represents the age certification of the title.

**7. Runtime**

* The datatype of the runtime column is Ratio.
* The domain is 0 to 940.
* The variable type is the integer.
* The column runtime represents the length of the movie or the TV show episodes in minutes.

**8. Genres**

* The datatype of the genre’s column is Categorical.
* The domain is None.
* The variable type is the list of strings.
* The column genres represent the list of genres.

**9. Production\_countries**

* The datatype of the production\_countries column is Categorical.
* The domain is None.
* The variable type is the string.
* The column production\_countries represents the list of the countries that produced the movie or TV show title.

**10. Seasons**

* The datatype of the season’s column is Ordinal.
* The domain is 1.0 to 53.0.
* The variable type is float.
* The column seasons represents the number of seasons a TV show has. For movies it is null.

**11. imdb\_id**

* The datatype of the imdb\_id column is Categorical.
* The domain is None.
* The variable type is the string.
* The column seasons represents the Id of the title on imdb.

**12. imdb\_score**

* The datatype of the imdb\_score column is Ordinal.
* The domain is 1.0 to 10.0.
* The variable type is float.
* The column imdb\_Score represents the score on the imdb.

**13. imdb\_votes**

* The datatype of the imdb\_votes column is a ratio.
* The domain is 5.0 to 2081757.0.
* The variable type is float.
* The column imdb\_votes represents the number of votes on the imdb.

**14. tmdb\_popularity**

* The datatype of the tmdb\_popularity column is ordinal.
* The domain is 0.000152997726845 to 3187.531.
* The variable type is float.
* The column tmdb\_popularity represents the popularity score on the tmdb.

**15. tmdb \_score**

* The datatype of the tmdb\_score column is ordinal.
* The domain is 0.0 to 10.0.
* The variable type is float.
* The column tmdb\_score represents the score on the tmdb.

The dataset credits.csv has the following columns and are described as follows:

**1. person\_id**

* The datatype of the person\_id column is ordinal.
* The domain is 1 to 2771401.
* The variable type is the integer.
* The column tmdb\_score represents the person id on the just watch.

**2. name**

* The datatype of the name column is categorical.
* The domain is none.
* The variable type is the string.
* The column tmdb\_score represents the name of the actor or the director.

**3. character**

* The datatype of the character column is Categorical.
* The domain is none.
* The variable type is the string.
* The column tmdb\_score represents the name of the character within the movie or the TV show.

**4. role**

* The datatype of the role column is Categorical.
* The domain is none.
* The variable type is the string.
* The column tmdb\_score represents whether the role within the movie or the TV show is the actor or director.

**Pre-processing on the dataset:**

* The title.csv contains 10k records with 15 columns and credits.csv contains 140k records with 5 columns. Therefore, prior to creating the visualization, pre-processing should be done on such big datasets.
* **Libraries used for preprocessing the data:**
* ast - The Python library ast provides a way to parse and manipulate the source code of a Python program.
* MultiLabelBinarizer - The class is imported from sklearn.preprocessing module and is used to convert a list of labels into a binary matrix representation, and is often used in multi-label classification tasks.
* pandas - The Python library pandas are used for reading, manipulating, and analyzing data.
* NumPy - NumPy is used while working with arrays and matrices of numerical data.
* The below code reads in two CSV files using the panda’s library and assigns the resulting data frames to the variables tdf and cdf.

Graphical user interface, text, application, email, website

Description automatically generated

* Then we combined the two data frames, tdf, and cdf, based on the column id, and we assigned the combined data frame to the variable df. To combine two or more data frames into a single data frame based on a shared column or index, we used the pandas' merge () function.
* We converted the genres and production\_countries columns from strings to lists of strings using the ast.literal\_eval() function.

Text

Description automatically generated with medium confidence

* **Genre -** A list of genres for each title is contained in the genre’s column, which is converted into a binary encoded format using the MultiLabelBinarizer () function from the sci-kit-learn module. Using the pd.concat() function, the converted data is combined with the original data frame df in a new data frame called genre\_columns. The add\_prefix() function renames the column names in genre\_columns with the prefix "genre\_"
* **Production countries:** A list of countries connected to each title is found in the production\_countries column. Using the apply() and lambda functions, we are just selecting the first country from the list in this case. The value is set to an empty string if the list is empty.
* **Season and Age Certification:** The fillna() function is used to replace the missing values for the seasons and age\_certification columns with 0 and the 'NoVal' label, respectively.

Text

Description automatically generated

* **Character, imdb\_id , imdb\_score , imdb\_votes , tmdb\_score and description:** These columns have missing values , so the below code will remove the rows with the missing values.

**Text

Description automatically generated with medium confidence**

* Finally, as a last step, we verified whether the dataset has any null values.

Graphical user interface

Description automatically generated with low confidence

Graphical user interface

Description automatically generated with low confidence

* Pre-processed data:

Table

Description automatically generated

**Section 3: Dashboard Users**

# Below is the list of prospective users who will be using our dashboard/who will be interested in this dashboard:

* **Film enthusiasts** - People who are passionate about exploring the dataset related to movies and TV shows to gain more insights about it.
* **Movie critics** - Professionals who provide critics and reviews on movies to provide feedback to the audience based on the evaluated data.
* **Researchers and analysts** - Data scientists and researchers can use data to analyze trends and performance and derive insights about the movies.
* **Media companies** - Media companies use the data to derive informed decisions that can aid them in producing, distributing, and broadcasting movies and TV shows.
* **Marketing Team** - Marketing teams at Amazon Prime or production companies could use the dashboard to analyze user demographics and preferences, helping them to create targeted campaigns for specific audiences.
* **Amazon Prime Product Manager** - The product managers at Amazon Prime could use this dashboard to track user engagement and popularity of different titles, helping them to make decisions about which titles to acquire or promote.

**Section 4: Questions**

# Below is the list of questions that the dashboard will answer:

* What is the distribution of Content-Type Shows across genre and production countries?
* What is the distribution of Titles by Type across genre, age certification, production countries, and release year?
* What is the distribution of age certification across genres, production countries, and release years?
* What are the top N countries based on content production across age certification?
* What is the distribution of title production across genres?
* How does the temporal distribution of the top 5 genre production look like?
* What are the top N titles based on IMDB Score across type, genre, age certification, and release year?
* Comparison of TV Shows’ popularity with a few seasons across genre, production countries, and release year?
* Show the top N actors by count across genre, production country, and release year?
* Who are the top N actors based on IMDB Score across countries and release years?

**Section 5: Plots**

**Question 1: What is the distribution of Content-Type Shows across genre and production countries?**

The below plot is used to depict the distribution of the content type (movies and TV Shows) across the different genre grouping, release years, and production countries using a pie chart. We have used the multi-value drop-down filter for release year, genre grouping, and production countries. The distribution of the pie chart (size) will change based on the genre grouping, release year, and the country we select in the filter.

**The pre-attentive attribute used is -** color and size.

**Color:** Blue color represents the percentage of movies and green color represents the percentage of tv shows.

Graphical user interface, text, application

Description automatically generated

**Question 2: What is the distribution of Titles by Type across genre, age certification, production countries, and release year?**

The below plot is used to depict the distribution of the title across different genre grouping, age certifications, and production countries using a line chart. We have used the multi-value drop-down filter for age certification, genre grouping, and production countries. Using the filter, we can select one or more combinations of inputs to get the desired distribution of titles.

**The pre-attentive attribute used is** – position and color.

**Color:** Blue color represents movies and green color represents tv shows

Graphical user interface, application

Description automatically generated

**Question 3: What is the distribution of age certification across genres, production countries, and release years?**

The below plot is used to depict the distribution of the age certification across different genre grouping, release years, and production countries using a bar chart. We have used the multi-value drop-down filter for release year, genre grouping, and production countries. Using the filter, we can select one or more combinations of inputs to get the desired bar chart depicting the distribution of age certifications.

**The pre-attentive attribute used is** – length and color.

**Color:** Blue color represents movies and green color represents tv shows

A screenshot of a computer

Description automatically generated

**Question 4:** **What are the top N countries based on content production across age certification?**

The below plot is used to show the top N countries based on content production across the age certification using a stacked bar chart. We have used the multi-value drop-down filter for the age certification and production countries, the multi-value list filter to select one or more content types, and the top N filter to select the required top n values to be displayed on the chart. Using the filter, we can select one or more combinations of inputs to get the desired bar chart depicting the top N countries based on content production across age certification.

**The pre-attentive attribute used is** – length and color.

**Color:** Blue color represents movies and green color represents tv shows

Graphical user interface, text, application

Description automatically generated

**Question 5: What is the distribution of title production across genres?**

The below plot is used to depict the distribution of titles across the different genres using a bar chart. We have not used any of the interaction controls here and the bar chart is static.

**The pre-attentive attribute used is** – length and color.

**Color:** Blue color represents movies and green color represents tv shows

Graphical user interface, application

Description automatically generated with medium confidence

**Question 6: How does the temporal distribution of the top 5 genre production look like?**

The below line chart presents the distribution of the top 5 genres based on production count, across release years. We have used the multi-value drop-down filter for the release year. Using the filter, we can select one or more combinations of inputs to get the desired line chart.

**Pre-attentive attributes:** position, length, and color where position encodes count of the genre.

Chart

Description automatically generated

**Question 7: What are the top N titles based on IMDB Score across type, genre, age certification, and release year?**

The below bar chart represents the top N titles based on IMDb scores and their comparison across the production count, across release years. We have used the multi-value drop-down filter for the age certification release year and genre grouping. Using the filter, we can select one or more combinations of inputs to get the desired bar chart.

**Pre-attentive attributes**: position, length, and color where the length of the bar encodes the IMDB scores.

**Color:** Blue color represents movies and green color represents tv shows

Graphical user interface, application

Description automatically generated

**Question 8:** **Comparison of TV Shows’ popularity with a few seasons across genre, production countries, and release year?**

The below charts compare the TV show’s popularity along with the count of each season across genre, production country, and release year. We have used the multi-value drop-down filter for the production countries, release year, and genre grouping. Using the filter, we can select one or more combinations of inputs to get the desired bar chart.

**Pre-attentive attributes:** position, length, and color where the length of the bar encodes the IMDB scores.

Graphical user interface

Description automatically generated

**Question 9: Show the top N actors by count across genre, production country, and release year.**

The below bar chart presents the top N actors by a count of titles they were involved with across genre, production country, and release year. We have used the muti-value drop-down filter for release year, production countries, and genre grouping along with the top N filter to show the desired bar chart.

**Pre-attentive attributes:** length and color

**Color:** Blue color intensity has been used for the count of titles.

Graphical user interface, application

Description automatically generated

**Question 10: Who are the top N actors based on IMDB Score across countries and release years?**

The below bar chart shows the top N actors based on IMDb scores of the titles they acted in across the release year. We have used the multi-value drop-down menu for the release year and production countries, a multi-value list to select one or more content types, and the top N filter to select the required top n values to be displayed on the bar chart.

**Pre-attentive attributes:** position, length, and color where the length of the bar encodes the IMDb scores.

**Color:** Blue color represents movies and green color represents tv shows

Graphical user interface, application

Description automatically generated

**Section 6:**

Below are the interactivities used in the dashboard:

* **Top N parameter:**

This parameter is used to filter out top N dimensions. The Top N value ranges are populated from the dimension column selected.

**The top N filter will be used in the below plots:**

* Top N Countries based on content produced age certification.
* Top N titles based on IMDB score across type, genre, age certification, and release year.
* Top N actors by counts across genre, production countries, and release year.
* Top N actors based on IMDB score across countries and release year.
* **Release year filter:**

This filter is a multi-value dropdown selector that adds the interactive ability to wrangle the data and update the plots based on the values selected. The filter value range is populated from the release year column. The options include ALL, 1993, … 2023, etc. as values and range from 1912 to 2023.

**The release year filter will be used in the below plots:**

* Distribution of content types across genre and production countries
* Distribution of Age certification across genre grouping, production countries, and release year.
* Top N titles based on IMDB score across type, genre, age certification, and release year.
* Top N actors by counts across genre, production countries, and release year.
* Top N actors based on IMDB score across countries and release year.
* **Genre grouping filter:**

This filter is a multi-value dropdown selector that adds the interactive ability to wrangle the data and update the plots based on the values selected from the genre grouping column. The options include ALL, Action, Animation, etc. as values.

**The genre grouping filter will be used in the below plots:**

* Distribution of content types across genre and production countries
* Distribution of Titles by type across genre, age certification, production countries, and release year.
* Distribution of Age certification across genre grouping, production countries, and release year.
* Distribution of Title production across genres.
* Temporal distribution of top 5 genre productions.
* Top N titles based on IMDB score across type, genre, age certification, and release year.
* Comparison of TV Shows popularity with the number of seasons across genre grouping, production countries, and release year.
* Top N actors by counts across genre, production countries, and release year.
* **Age certification filter:**

This filter is a multi-value dropdown selector that adds the interactive ability to wrangle the data and update the plots based on the values selected. The filter value range is populated from the age certification column. The options include ALL, PG, R, etc. as values.

**The age certification filter will be used in the below plots:**

* Distribution of Titles by type across genre, age certification, production countries, and release year.
* Distribution of Age certification across genre grouping, production countries, and release year.
* Top N Countries based on content produced age certification.
* Top N titles based on IMDB score across type, genre, age certification, and release year.
* **Production countries filter:**

This filter is a multi-value dropdown selector that adds the interactive ability to wrangle the data and update the plots based on the values selected from the production countries column. The options include ALL, AU, US, etc. as values.

**The production countries filter will be used in the below plots:**

* Distribution of content types across genre and production countries
* Distribution of Titles by type across genre, age certification, production countries, and release year.
* Distribution of Age certification across genre grouping, production countries, and release year.
* Top N Countries based on content produced age certification.
* Comparison of TV Shows popularity with the number of seasons across genre grouping, production countries, and release year.
* Top N actors by counts across genre, production countries, and release year.
* Top N actors based on IMDB score across countries and release year.
* **Type filter:**

This filter is a multi-value list filter that adds the interactive ability to wrangle the data and update the plots based on the values selected from the type of column. The options include ALL, MOVIES, and SHOWS as values.

**The type filter will be used in the below plots:**

* Top N Countries based on content produced age certification.
* Top N titles based on IMDB score across type, genre, age certification, and release year.
* Top N actors based on IMDB score across countries and release year.

**Dashboard Submission:**

We have published our tableau dashboard and below is the link to access it:

<https://public.tableau.com/views/IFT598_FinalDashboard_SwethaKevinKiranSushmitha/AmazonPrimeMoviesTVShows-Dashboard?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link>

# References

*Wexler, S., Shaffer, J., & Cotgreave, A. (2017). The big book of dashboards visualizing your data using real-world business scenarios. John Wiley & Sons, Inc.*

*Enrique, D. (2023, March 13). Amazon prime movies and TV shows. Kaggle. Retrieved April 23, 2023, from* [*https://www.kaggle.com/datasets/dgoenrique/amazon-prime-movies-and-tv-shows*](https://www.kaggle.com/datasets/dgoenrique/amazon-prime-movies-and-tv-shows)

*KKSS. mural.co. (n.d.). Retrieved April 23, 2023, from* [*https://app.mural.co/t/ift598dvprojectspring20235871/m/ift598dvprojectspring20235871/1680555750325/93c7e8dcb4cd3a044e6242d8b39d8dc7fad83654?sender=u32d73f61785512dd02250914*](https://app.mural.co/t/ift598dvprojectspring20235871/m/ift598dvprojectspring20235871/1680555750325/93c7e8dcb4cd3a044e6242d8b39d8dc7fad83654?sender=u32d73f61785512dd02250914)