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

Kevin Joseph Scaria

Sushmitha Alagesan

Swetha Anantha Murthy

Kiranmai Reddy Bommanapalli

**Information Technology, Arizona State University**

**Project Phase 2: Decision Making**

Dr. Asmaa Elbadrawy

Due date: 16 April 2023

# 

# Section 1: Used Visualization Tools

To develop an interactive dashboard for our chosen Dataset, "Amazon Prime Movies and Tv Shows," we chose to use Tableau software as our visualization tool.

**Reason for choosing Tableau:**

* Two CSV files, title.csv, and credit.csv, with roughly 10k and 140k records respectively, are part of the Amazon Prime dataset. When working with larger datasets in such situations, Tableau is frequently chosen over Bokeh. Because of Tableau's strong data management features, users can work more effectively and efficiently with massive datasets. Additionally, it provides a large variety of data connectors, making connecting to various data sources simpler.
* To quickly build new visualizations, Tableau provides a selection of pre-built templates, charts, and graphs. Analyzing a dataset and making fast prototypes can be helpful.
* When dealing with massive datasets like Amazon Prime, it is really challenging to identify patterns and trends in the data without a suitable visualization. With Tableau's customization features, we have the power to effectively communicate the most important data points from the data set.
* The user interface of Tableau is user-friendly and easy to use. It is possible to create interactive dashboards as they integrate with various data sources without any code. On the other hand, bokeh requires a lot of intensive coding to integrate with a few data sources and create dashboards which might be difficult.

**Section 2: Dataset Pre-processing**

The dataset is available on Kaggle.

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

* 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 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 scikit-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: Final Set of Questions**

* 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 counts across genre, production country, and release year?
* Who are the top N actors based on IMDB Score across countries and release years?

**Section 4: Dashboard Plots**

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

**Graphical user interface, text, application

Description automatically generated**

**Explanation of plot:**

The above plot is used to show the distribution of the content type (movies and TV Shows) across the different genre grouping, release years, and production countries. 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.

**Chart Used:** Pie-chart.

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

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

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

**Graphical user interface

Description automatically generated**

**Explanation of plot:**

The above plot is used to show the distribution of the title across different genre grouping, age certifications, and production countries. 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.

**Chart Used:** Line chart.

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

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

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

**A picture containing graphical user interface

Description automatically generated**

**Explanation of the plot:**

The above plot is used to show the distribution of the age certification across different genre grouping, release years, and production countries. 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.

**Chart Used:** Bar chart.

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

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

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

**Graphical user interface, text, application

Description automatically generated**

**Explanation of the plot:**

The above plot is used to show the top N countries based on content production across the age certification. 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.

**Chart Used:** Stacked bar chart.

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

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

* **Plot 5 - What is the distribution of title production across genres?**

Graphical user interface, application

Description automatically generated

**Explanation of the plot:**

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

**Chart Used:** Bar chart.

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

**Color:** Each color represents different genres.

* **Plot 6 - How does the temporal distribution of the top 5 genre production look like?**

Chart

Description automatically generated with medium confidence

**Explanation of the plot:**

The plot 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.

**Chart Used:** Line chart.

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

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

Graphical user interface, chart, bar chart

Description automatically generated

**Explanation of the plot:**

The plot presents 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.

**Chart Used:** 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

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

Graphical user interface, application

Description automatically generated

**Explanation of the plot:**

This plot compares 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.

**Chart Used**: Bar chart.

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

* **Plot 9 - Show the top N actors by count across genre, production country, and release year?**

Chart, bubble chart

Description automatically generated

**Explanation of the plot:**

This plot 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 bubble chart.

**Chart Used:** Bubble chart.

**Pre-attentive attributes:** size and color. Here size indicates the count and color encodes the top N actors.

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

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

Graphical user interface, chart, bar chart

Description automatically generated

**Explanation of the plot:**

This plot 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.

**Chart Used:** Bar chart.

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

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

**Section 5: Dashboard Interactivity**

Below are the interactivities used in the plots:

* **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.

**Plots connected:** Plots 4, 7, 9, and 10.

* **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.

**Plots connected:** Plots 1, 3, 6, 7, 8, 9, and 10

* **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.

**Plots connected:** Plots 2, 4, and 7

* **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.

**Plots connected:** Plots 1, 2, 4, 8, 9 and 10

* **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.

**Plots connected:** Plots 1, 2, 3, 7, 8, and 9.

* **Type 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 type of column. The options include ALL, MOVIES, and SHOWS as values.

**Plots connected:** Plots 4, 7, and 10

# Section 6: 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 16, 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 16, 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)