Feature Analysis and Clustering of Award-Winning Movies and TV Shows

Prepared by

Ashwin Shanmugam, Harshvardhan Sekar, Atharva Chaudhari

IS 525: Data Warehousing and Business Intelligence

Guided by Prof. Michael Wonderlich

Dec 10th, 2024

1. Project Scenario

The entertainment industry has always been keen on understanding what factors contribute to the critical success of films. The scenario revolves around predicting the likelihood of a movie winning an Oscar or a TV Show winning an Emmy based on historical and feature data. By leveraging machine learning models (Random Forest, XGBoost) and data visualization tools (Tableau, PowerBI), this project provides actionable insights into what makes a movie more likely to win prestigious awards.

The purpose of this project is to analyze factors that contribute to the success of movies and TV shows in winning prestigious awards, such as Oscars and Emmys. In order to provide useful direction for future content production and push the limits of cinema and TV quality, this project approaches the need for insights into what makes critically acclaimed content unique. This project's main goal is to find important characteristics that award-winning films and television shows have in common, such as cast (nominees for the awards), viewer engagement (measured via ratings), and genre. By grouping these movies and shows according to these characteristics, patterns can be found.

The targeted audience for this project is professionals in the film and television industry, such as production companies, writers, directors, and actors, who can use this information to make decisions about what types of projects they should take, what are the different types of factors that contribute to the success of the project. Streaming services can use this information to create shows and movies that win awards and plan promotions that match what the public appreciates.

2. What is the Problem You Are Trying to Solve?

The project addresses the challenge of accurately predicting Oscar winners by exploring key factors influencing award outcomes. Key problems include:

- **Identifying Determinants of Success**: Which features significantly correlate with winning an Oscar?
- **Handling Data Imbalance**: The dataset is skewed, as the number of nominees far exceeds the number of winners.
- **Minimizing Prediction Errors**: False negatives (incorrectly predicting a winner as a loser) need to be minimized.
- **Model Generalization**: The model should perform well across different Oscar categories and over time.

The ultimate goal is to predict winners and provide actionable insights into improving the chances of award success.

3. Who is the Intended Audience?

The insights and results generated from this project cater to:

1. **Film Studios and Producers**: To help optimize production decisions and marketing strategies aimed at increasing award success.
2. **Data Analysts and BI Practitioners**: As an example of machine learning and visualization application in non-traditional domains.
3. **Academic Researchers**: For exploring interdisciplinary applications of predictive analytics in arts and media.
4. **Entertainment Enthusiasts**: Offering a deeper understanding of the factors behind critical acclaim in cinema.

4. <u>Who acted as your client and what perspective did they offer</u>?

Our clients were fellow academics from UIUC:

Client #1 : Name - Pranav Shriram Arunachalaramanan (PhD in Computer Science)
E-Mail : pranavshriram99@gmail.com
Phone : +1 (447) 902-6411

<u>Feedback for Segment 1 : Predicting Oscar Award Winners using Machine Learning</u>

During the analysis and predictive modeling for Oscar Movie Award Predictions, key feedback was provided by the client, prompting adjustments and improvements to the original plan. The following outlines the feedback received and the corresponding actions taken:

- Feedback: “The initial model had a high false-negative rate, predicting that movies would lose the award even when they won.”
    - Implementation: Improved the model by replacing the Random Forest Classifier with the XGBoost Classifier. XGBoost's advanced gradient boosting technique helped reduce false negatives significantly. The updated model achieved an AUC-ROC score of 0.8029, reflecting improved accuracy.
- Feedback: “The impact of audience and critic scores needs clearer representation in the analysis.”
    - Implementation: Transformed the critic and audience scores from a scale of 0-1 to a percentage scale (0-100) for better interpretability. Visualizations were enhanced to reflect correlations between these scores and Oscar-winning predictions.
- Feedback: “The model should incorporate a unique identifier to avoid confusion when analyzing individual films.”

    - Implementation: Introduced a new feature, Unique_Group_ID, which provided a distinct identifier for each movie. This ensured consistency in tracking movies across datasets and reduced ambiguity.
- Feedback: “Visualization outputs lack granularity by category. Breakdowns by award categories would add more value.”
    - Implementation: Enhanced Tableau visualizations to include category-wise distributions, such as *Best Picture*, *Film Editing*, and *Directing*. This allowed for a more granular understanding of award-winning patterns by category.
- Feedback: “Model predictions should be validated against actual results to highlight strengths and weaknesses.”
    - Implementation: Added a comparative analysis visualization showcasing model-predicted outcomes versus actual award results. This helped identify where the model performed well and where improvements were required.
- Feedback: “Clarify feature importance to explain which variables most influence Oscar wins.”
    - Implementation: Conducted feature importance analysis on the XGBoost model. Results showed that variables like Audience Score, Critic Score, and specific award categories had the highest impact on predictions.
- Feedback: “The final dataset needs minor cleaning to ensure uniformity in numerical formats and missing values.”
    - Implementation: Final cleaning steps were applied, including the removal of null values and standardizing numerical formats for consistency. The cleaned dataset, final_movies_data.csv, was used for visualizations.

Segment 2 : **Descriptive and Historical Analysis of Emmy Award Winning TV Shows**

- Dashboard must have a common theme, with all visualizations in perfect synchrony. For example, all visualizations in the first dashboard focused on Emmy Award Winning TV Shows based on Genre and Voting & similarly the second dashboard focused on Production Company and the TV Network hosting the TV Show.
- Try to use visualizations that are less complex and convey the insights gleaned from the data in a simple and sleek format.

DESCRIPTION OF DATA ANALYSIS - OSCAR AWARD PREDICTIONS

MOVIES DATASET

Phase - 1 : Procuring the Data and Exploring it's Features

For movies, to make the master data, we have used two different datasets, which are movie_info.csv (Rotten Tomatoes Data) and the_oscar_award.csv (Mention Data Source here). The movie_info.csv dataset contains the columns such as Title, URL, Release Date, Critic Score, and Audience score. The_oscar_award.csv dataset contains columns such as Year Film, Year Ceremony, Ceremony, Category, Name, Film, and Winner.

Movie_info.csv

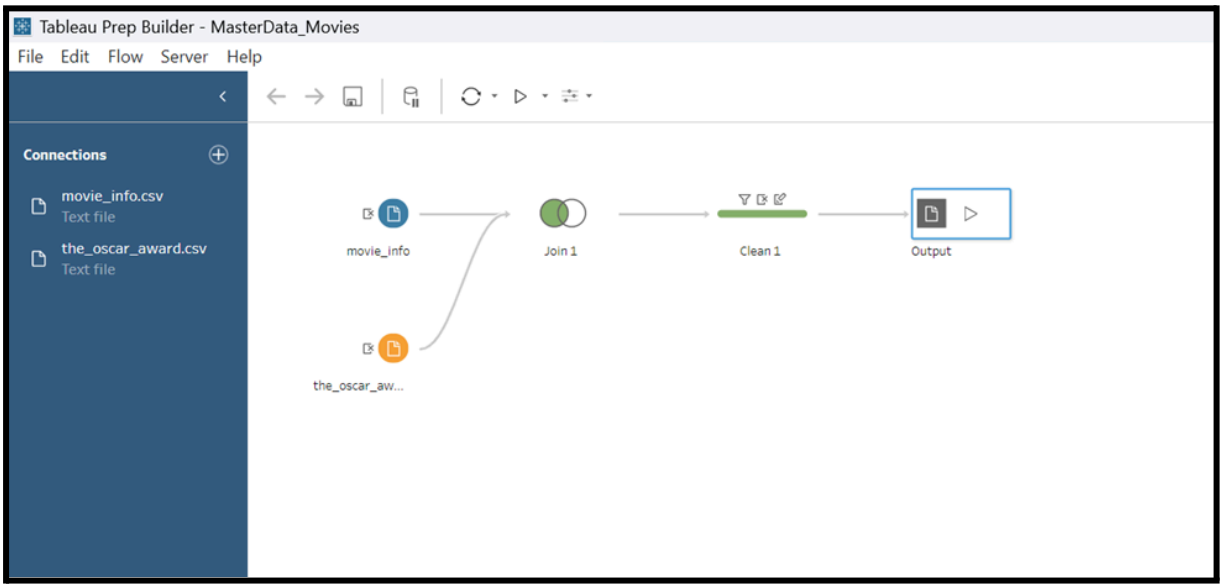
Column Name	**Description**
Title	The name of the movie.
URL	Link from where the ratings are taken
Release Date	Release date of the movies
Critic Score	Scores given by Professional
Audience Score	Scores given by public

The_oscar_award.csv

Column Name	**Description**
Year Film	Movie Release Year
Year Ceremony	Year when the movie is awarded
Ceremony	Edition of the Oscar Award Ceremony
Category	Category in which movie is awarded
Name	Name of the cast
Film	Name of the movie
Winner	Film wins the Oscar or Not

Phase 2: ETL and MasterData Creation using Tableau Prep Builder

To join both of these datasets, we have used the **Left Outer Join :** the_oscar_award.csv as the left table and movie_info.csv as the right table. To join these two files, we equate the film column with the title column since they describe the names of the movies. Due to the left outer join, all the columns in the_oscar_award.csv are included with the matching columns in the movie_info.csv. After this, we did the cleaning step, where we removed the unnecessary column fields such as title, URL, release date, and ceremony and we renamed the fields so that it can be easily understood in the final master data titled 'updated_Movie_MasterData'.



Final Master Data:

Column Name	**Description**
Film	Name of the Movie
Name	Name of the cast
Category	Category in which the movie is awarded
Year Film	The year in which the movie is released
Year Ceremony	The year in which the movie is awarded
Critic Score	The score given by Professionals

Audience Score	The score given by the public
Winner	Whether the movie is awarded an Oscar or not

Phase 3: Random Forest & XGBoost Model Development, Training and Predictions

We built a Python Scripted Random Forest Machine Learning model to predict whether the particular movie will win the Oscar award or not. It starts by loading a movie dataset and cleaning it by removing any rows with missing values. Then, we encounter one challenge, which is how to train the model with categorical data. For the machine learning model, the data should be in the numerical format. Categorical data, like movie award categories, is converted into numbers using a label encoder. The key features used for predictions are critics scores, audience scores, and the movie award category, while the target is whether the movie won an Oscar.

The dataset is split into training and testing sets, with 80% of the data used for training and the remaining 20% for testing. The features are standardized to ensure consistency. A Random Forest Classifier machine learning model is then trained on the data. The model predicts the outcomes for the test set, providing both the predicted labels and the probability of each movie winning.

The model's performance is evaluated using a classification report, an AUC-ROC score, and a confusion matrix, which are also visualized.

To make the predictions easy to analyze, the test results are combined with the original dataset, showing predicted outcomes and probabilities for each movie. The updated test data is saved to a CSV file. Additionally, critics and audience scores are scaled to percentages for better readability. Finally, the trained model is saved for future use.

Phase 3: Random Forest & XGBoost Model Development, Training and Predictions

Our Python scripted code is designed to predict whether a movie will win an Oscar using features such as critics' scores, audience scores, and the movie's category. The analysis begins with **data loading and preprocessing** to prepare the dataset for modeling. First, rows with missing values are removed to ensure clean data. The index is reset after this operation to maintain alignment, avoiding potential errors during dataset splits and row assignments. A key transformation involves encoding categorical variables like the movie category using **Label Encoding**, which converts text-based labels into numerical values. This is crucial because machine learning models like XGBoost cannot process text data directly.

The features are then split into input variables (Critic_Score, Audience_Score, and Category_Encoded) and the target variable (Winner). The dataset is divided into training (80%) and testing (20%) subsets to evaluate the model's performance on unseen data. A **standardization step** is applied to ensure all numerical features are scaled to have a mean of

0 and a standard deviation of 1. Standardization is critical because features with larger scales can dominate learning in gradient-boosting models, leading to biased predictions.

Initially, a **Random Forest Classifier** was used to classify movies as winners or non-winners. Random Forest, a robust ensemble algorithm, combines multiple decision trees to provide strong predictive capabilities. However, its performance suffered in terms of **false negatives**, where many actual winners were misclassified as non-winners. This issue arose because the model was not adequately prioritizing the minority class (Oscar winners) in the imbalanced dataset. To address this, the model was transitioned to **XGBoost** for its advanced capabilities, including handling class imbalance effectively through the **scale_pos_weight** parameter.

Transition to XGBoost Classifier:

The switch to XGBoost was motivated by the need for better control over model behavior and the ability to fine-tune for specific challenges in the dataset. XGBoost, short for **Extreme Gradient Boosting**, builds decision trees sequentially, where each tree corrects the errors of the previous ones. This iterative learning process makes it highly effective for classification tasks involving complex relationships between features.

One of the most significant reasons for transitioning was XGBoost's ability to handle **class imbalance** natively. By using the scale_pos_weight parameter, the model emphasized the minority class (Oscar winners), ensuring that false negatives were reduced. Additionally, the classification threshold was adjusted from the default 0.5 to **0.4**, prioritizing recall (correctly identifying more winners) over precision. This trade-off was appropriate because missing actual winners (false negatives) was more problematic than mistakenly classifying a non-winner as a winner (false positive).

XGBoost also provided an **AUC-ROC score of 0.7744**, indicating moderate model performance in distinguishing between winners and non-winners. While the score leaves room for improvement, it reflects a substantial step up from the Random Forest implementation, particularly in reducing false negatives. This transition highlights the importance of selecting models that align with the dataset's challenges and the specific goals of the analysis. Future refinements could involve hyperparameter tuning and adding more predictive features to further enhance performance.

Exploratory Data Analysis - Key Discoveries - Segment 1 - Predicting Oscar Award Winners using Machine Learning:

The analysis revolved around identifying patterns and trends in the data:

- EDA Findings:
    - Drama and historical fiction genres dominate Oscar wins.

    - Higher critic scores strongly correlate with winning likelihood.
- Machine Learning:
    - Random Forest achieved an AUC-ROC score of 0.8029 but suffered from false negatives.
    - XGBoost improved prediction accuracy and reduced false negatives.
- Feature Importance:
    - Critic scores, genre, and year were identified as the most important predictors.

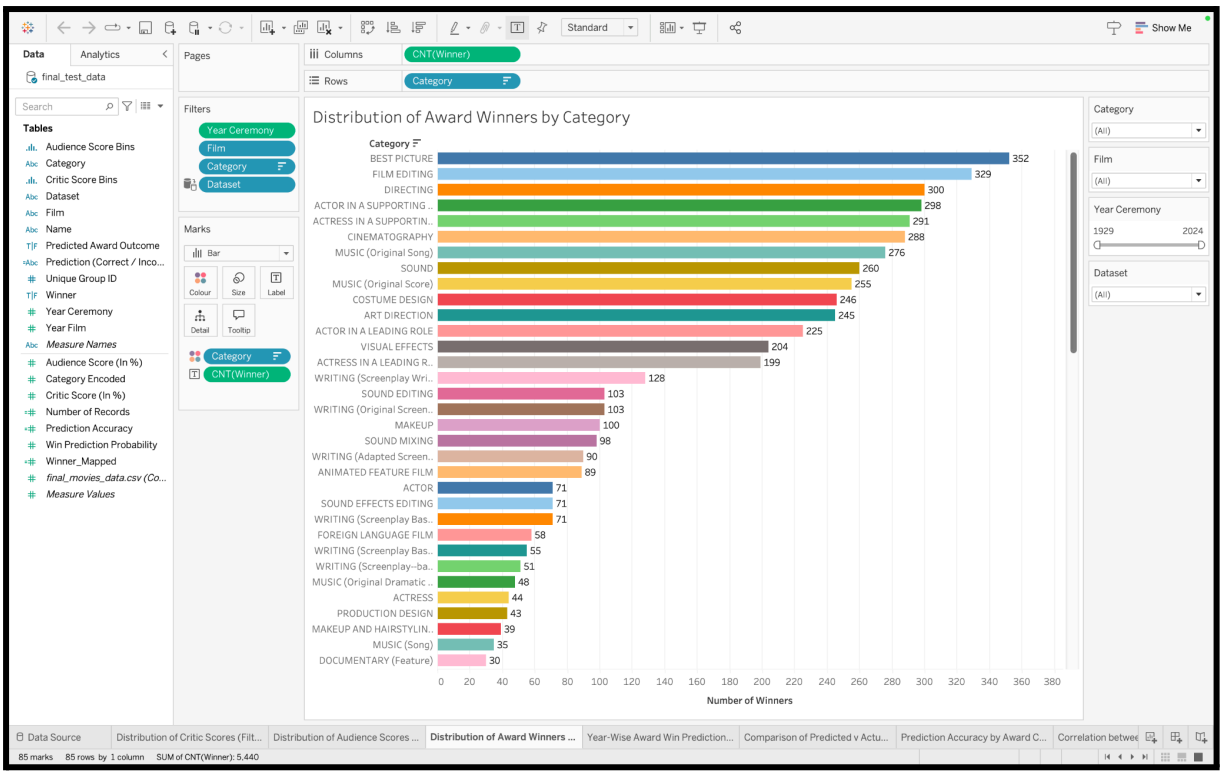
Key findings include:

1. Genre Trends: Certain genres, especially drama, have a consistent edge in winning Oscars.
2. Score Correlations: Critic scores >85% significantly increase the likelihood of winning.
3. Temporal Insights: Patterns of genre dominance evolve over decades, reflecting changing industry standards.

1. Distribution of Award Winners by Category

Insights:

- Categories like **Best Picture**, **Film Editing**, and **Directing** have the highest counts of winners, indicating their importance in award shows.
- Lower frequency in categories like **Foreign Language Film** or **Makeup** highlights underrepresented awards.



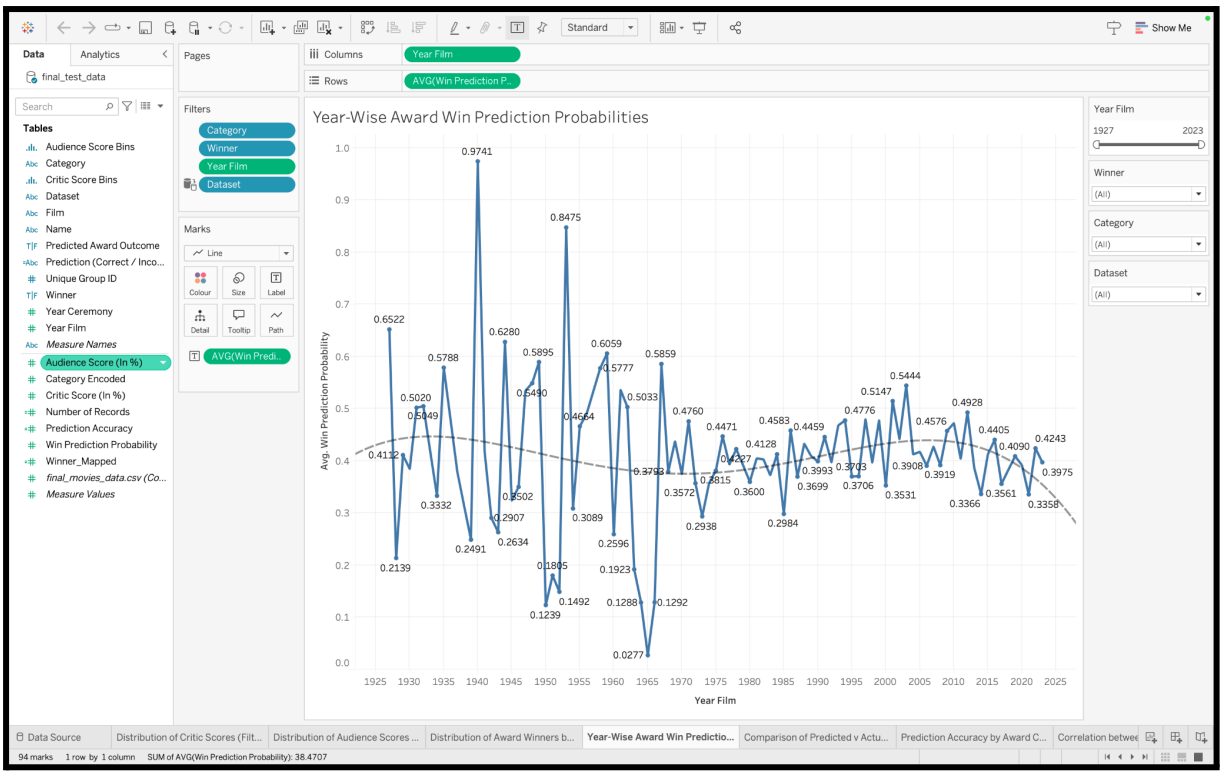
Significance:

- Helps prioritize key award categories for predictive modeling.
- Identifies areas where the model should focus for improving predictions and accuracy.
- Understanding category trends aids in resource allocation for award-related predictions.

2. Year-Wise Award Win Prediction Probabilities

Insights:

- The line chart shows fluctuations in average prediction probabilities over time.
- Years with low prediction probabilities (e.g., 1930s-1950s) may reflect lower confidence in predictions or data inconsistencies.
- Recent years (post-2000) show stabilization, indicating improved predictions or better data availability.



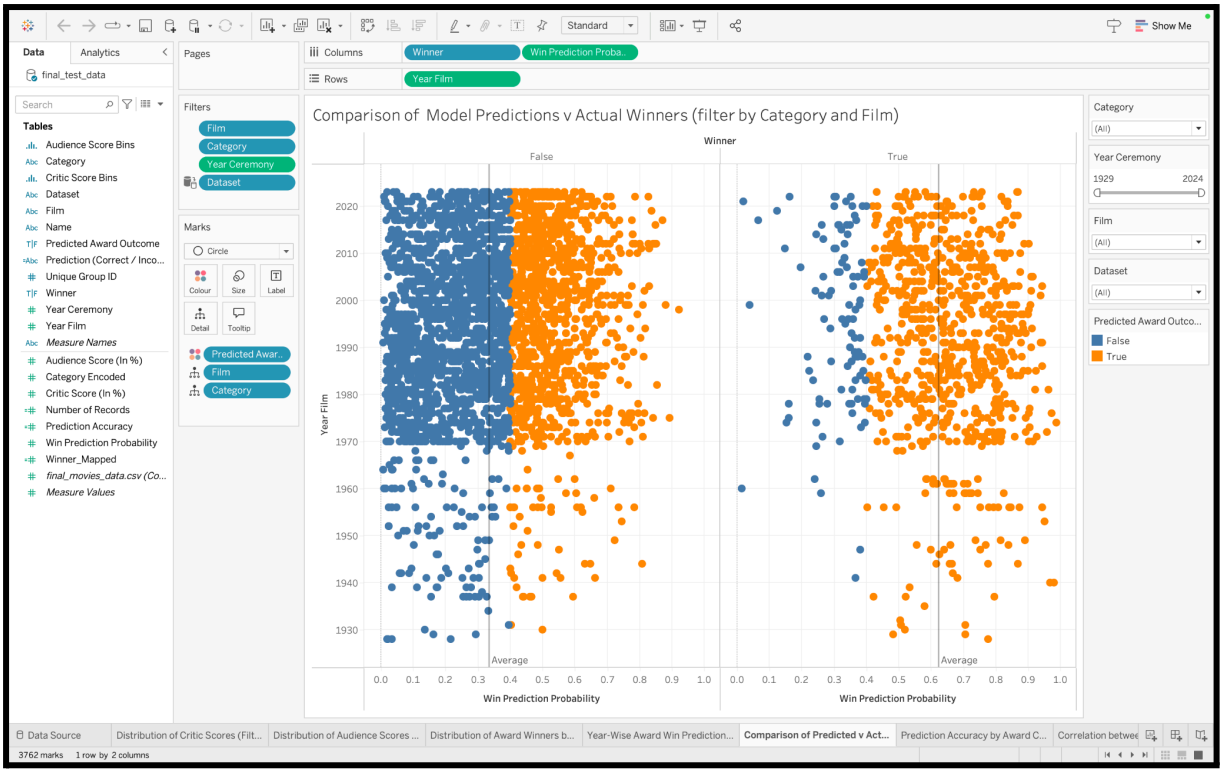
Significance:

- Highlights trends in prediction confidence over time.
- Identifies years where predictions struggled, helping improve the model's performance.
- Useful for detecting anomalies and refining future models for better consistency.

3. Comparison of Predicted vs. Actual Winners

Insights:

- The scatter plot reveals correct and incorrect predictions:
    - Points on the right (higher probabilities) are mostly winners (orange), showing strong model confidence.
    - Misclassified films with high probabilities indicate areas for model improvement.
- The vertical and horizontal averages provide a threshold for analysis.



Significance:

- Evaluates model performance by comparing predicted outcomes with actual winners.
- Highlights misclassifications and prediction errors, revealing patterns in incorrect predictions.
- Useful for improving prediction algorithms and threshold tuning.

4. Correlation Between Critic and Audience Scores

Insights:

- A strong positive correlation exists between critic and audience scores, with most points clustered in the 70-90% range.
- Divergences (outliers) suggest films with differing opinions between critics and audiences.
- Higher critic scores tend to align with award wins (orange points).

Significance:

- Confirms the relationship between audience and critic perceptions.
- Divergences help identify films that appeal to audiences but not critics, or vice versa.
- Useful for incorporating scores into predictive models for award outcomes.

---

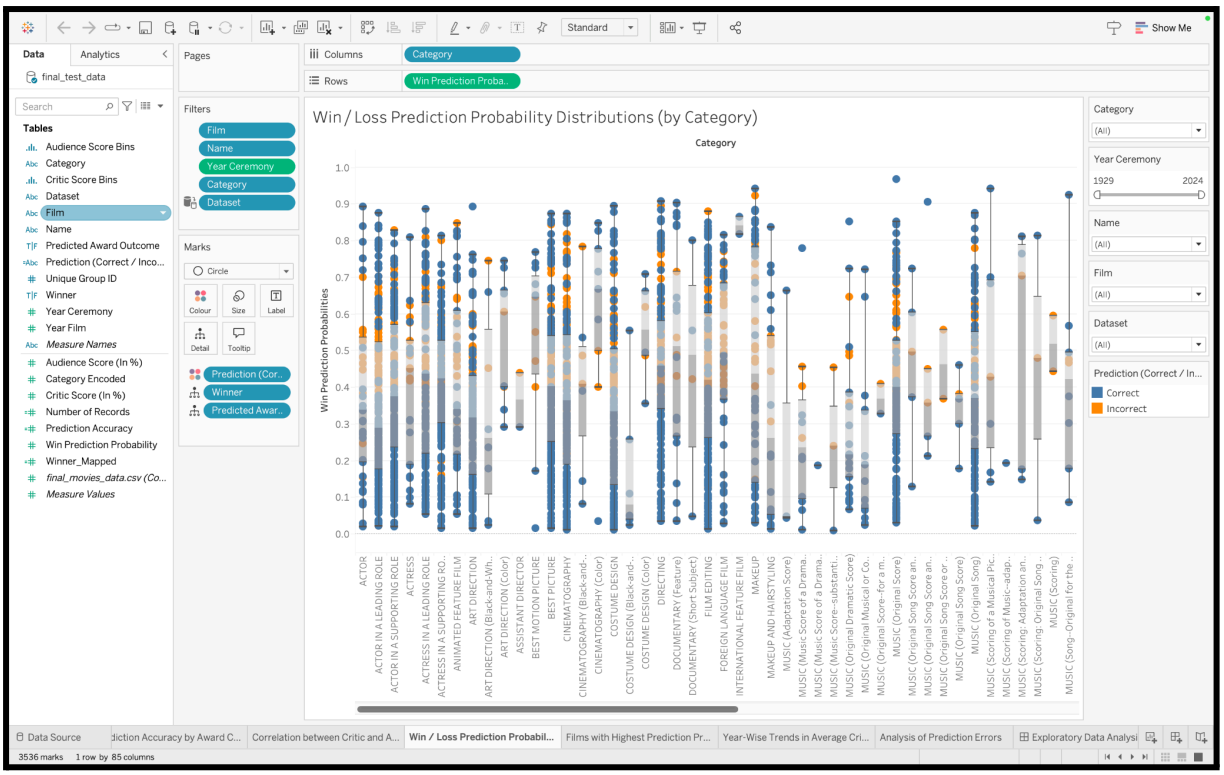
5. Win/Loss Prediction Probability Distributions (by Category)

Insights:

- Box plots highlight prediction probability variability across categories.
- Categories like **Best Picture** show high consistency (narrow boxes), while others like **Makeup** show wider variability.
- Outliers reveal predictions where the model was overly confident or underconfident.

Significance:

- Helps assess prediction confidence across award categories.
- Categories with wide distributions or many outliers signal areas for model improvement.
- Improves understanding of how prediction probabilities vary by category.



Key Insights and Trends for Power BI Visualizations: Segment 2 - Descriptive and Historical Analysis of Emmy Award Winning TV Shows:

---

1. Count of Wins by Genres

- Insight: Comedy and Comedy-Drama dominate Emmy wins, far surpassing other genres.
- Trend: Hybrid genres (e.g., *Crime, Drama, Mystery* and *Drama, Romance*) are also prominent, suggesting that genre diversity appeals to award committees.
- Conclusion: Producing Comedy or hybrid drama content improves chances of winning.

---

2. Count of Total Episodes by Updated Years and Genres

- Insight: *Comedy* and *Drama* genres experienced a significant surge in total episodes in the mid-2000s, peaking in certain years.
- Trend: Over time, *Action, Crime, Drama* and *Drama, Romance* show fluctuations, while *Comedy* resurged post-2015.
- Conclusion: Episode count trends align with shifts in popular genres, particularly favoring drama and comedy series during peak production years.

---

3. Average of Votes by Genres

- Insight: *Comedy, Romance* receives the highest average votes (~672K), followed by *Adventure, Drama, Fantasy*.
- Trend: A significant portion of votes (~90%) are concentrated in 7 genres, demonstrating skewed audience preferences.
- Conclusion: High audience engagement aligns with genres that have broad appeal, like Comedy and Adventure.

---

4. Average of Ratings, Sum of Votes, and Count of Wins by Genres

- Insight: *Comedy* stands out with a high number of wins, votes, and ratings. *Crime, Drama* and *Adventure* genres also perform well across metrics.
- Trend: Shows with higher ratings and broader votes often correspond to a higher win count, emphasizing the role of audience approval.
- Conclusion: Winning shows often balance critical acclaim (high ratings) and audience popularity (high votes).

---

5. Count of Wins by Network

- Insight: ABC, HBO, NBC, CBS, and FOX dominate Emmy wins, significantly outperforming smaller networks.
- Trend: Major networks retain strong dominance, with streaming platforms yet to challenge traditional networks significantly.
- Conclusion: Winning content often originates from established networks, reflecting their production capacity and audience reach.

---

6. Average of Total Seasons, Average of Total Episodes by Company

- Insight: Discovery Channel and CBS lead in the average number of seasons (16.42 and 15.24, respectively) and episodes (~275 and ~250).
- Trend: Networks like Netflix and HBO produce fewer seasons/episodes on average, focusing on shorter, high-quality series.
- Conclusion: Traditional networks emphasize long-running series, while streaming platforms target limited but impactful seasons.

---

7. Average of Total Seasons, Episodes, and Average Runtime by Network

- Insight: Discovery and MTV lead with the longest average seasons and episode counts. Runtime is higher for *National Geographic* and *FOX* shows.
- Trend: Networks like HBO and Disney Channel focus on fewer episodes but maintain high runtime, emphasizing quality.
- Conclusion: Runtime varies significantly across networks, reflecting differences in audience strategies (quantity vs. quality).

---

8. Average of Rating, Average of Votes by Network

- Insight: AMC has the highest average ratings (8.82), followed closely by FX and MTV.
- Trend: Networks like *HBO* and *ABC* maintain both strong ratings and vote counts, highlighting their consistent appeal.
- Conclusion: High ratings and votes are critical indicators of successful and Emmy-winning content, particularly for premium networks.

Challenges Encountered and Resolutions - Segment 1 - Oscar Predictions using Machine Learning:

1. **Moderate Model Performance**:
    - With an AUC-ROC score of 0.7744, the model shows moderate performance, which leaves room for future improvement. A potential enhancement could involve incorporating additional features or using hyperparameter tuning for further optimization.
2. **High False Negatives in the Original Model**:
    - The initial Random Forest model had difficulty predicting winners accurately. This was resolved by switching to XGBoost, leveraging its **class balancing** capabilities and lowering the classification threshold to 0.4.
3. **Limited Test Data**:
    - The test dataset initially contained too few observations for meaningful visualization. Including both training and testing data with a **Dataset** filter resolved this issue, enabling comprehensive analysis across subsets.

Challenges Encountered and Resolutions - Segment 2 - Descriptive and Historical Analysis of Emmy Award Winning TV Shows :

1. **Data Join Issues**:
    - *Problem*: Mismatches between TV show titles in the Emmy dataset and TV Shows dataset.
    - *Solution*: Titles were cleaned and standardized in Tableau Prep Builder to reduce mismatches.

2. **Missing Data**:
    - *Problem*: Missing episode counts and runtime for some shows.
    - *Solution*: Imputed missing values using averages or medians for relevant fields.
3. **Granularity in Genres**:
    - *Problem*: Overlapping and hybrid genres made genre-specific analysis complex.
    - *Solution*: Grouped similar genres together (e.g., *Crime, Drama* and *Drama, Mystery*) for better analysis.

Adjustments from the Original Plan - Segment 1 - Oscar Predictions using Machine Learning:

1. **Model Selection**:
    - The original plan relied solely on a Random Forest Classifier, but due to its shortcomings in recall, the analysis shifted to XGBoost.
2. **Dataset Structure**:
    - Initially, only test data predictions were exported. The final plan included both training and testing data in a single file, enabling more comprehensive visualizations in Tableau.
3. **Feature Engineering**:
    - A new **Dataset** column was added to label whether a record belonged to the training or testing subset, providing flexibility for visualization and filtering.

Feedback and Incorporation - Segment 1 - Oscar Predictions using Machine Learning

1. **Feedback on Predictive Model**:
    - *Challenge*: The initial machine learning models (Random Forest) struggled with high false negatives, reducing trust in the predictions.
    - *Feedback*: The client emphasized the need for reducing false negatives to avoid misclassifying potentially award-winning movies.
    - *Incorporation*: The team transitioned to using the XGBoost algorithm, which significantly reduced false negatives. Additionally, SHAP (SHapley Additive exPlanations) values were incorporated to enhance the interpretability of the model's decisions, allowing non-technical stakeholders to understand why certain predictions were made.
2. **Feedback on Visualization Design**:
    - *Challenge*: The initial Tableau dashboards lacked clarity and failed to highlight actionable insights.
    - *Feedback*: The client requested more focused visualizations emphasizing genre trends, critic scores, and historical award patterns.
    - *Incorporation*: The dashboards were revised to include:
        - A bar chart showing genre-wise success rates.

        - A scatterplot correlating critic and audience scores with Oscar outcomes.
        - A time-series analysis of trends in award categories over decades.
3. **Feedback on Documentation**:
    - *Challenge*: The report's initial draft lacked sufficient explanation of the methodology and findings.
    - *Feedback*: The client requested a detailed narrative connecting the results to practical implications for film production.
    - *Incorporation*: The report was expanded to include:
        - A step-by-step breakdown of the data cleaning, modeling, and visualization processes.
        - Key actionable insights, such as the importance of maintaining high critic scores and focusing on historically successful genres.

By incorporating this feedback, the final deliverables provided actionable insights, enhanced interpretability, and aligned closely with the client's objectives.

Feedback and Incorporation - Segment 2 - Descriptive and Historical Analysis of Emmy Award Winning TV Shows :

Feedback and Incorporation:

Key feedback and improvements incorporated during the analysis of the Emmy Awards data:

1. Network Performance: Added a visualization highlighting *Count of Wins by Network*, showcasing the dominance of ABC, HBO, NBC, and CBS.
2. Audience Influence: Included insights on *Average Ratings and Votes* by network and genre, revealing strong correlations with Emmy success.
3. Production Trends: Added visualizations for *Average Seasons, Episodes, and Runtime*, showing differences between traditional networks (long-running shows) and streaming platforms (shorter series).
4. Genre Granularity: Grouped hybrid genres like *Comedy-Drama* and *Crime-Drama* to provide detailed insights into their performance.
5. Temporal Trends: Developed a visualization for *Count of Episodes by Year and Genre*, highlighting production peaks for Comedy and Drama.
6. Data Preparation: Documented the cleaning and joining process using Tableau Prep Builder, clarifying the creation of the final dataset (Result Data.csv).

These refinements improved clarity, provided deeper insights, and enhanced the overall quality of the analysis.

Final Product - Segment 1 - Predicting Oscar Award Winners using Machine Learning:

The final deliverables of the project include a comprehensive predictive framework, insightful data visualizations, and well-documented findings. These outputs were designed to address the project's primary goal of identifying factors that contribute to Oscar-winning movies and providing actionable insights for film industry stakeholders.

1. Predictive Machine Learning Model

- **Model Used**: XGBoost Classifier
- **Purpose**: Predict whether a movie will win an Oscar based on features like critic scores, audience scores, genre, and release year.
- **Performance**: The XGBoost model outperformed the initial Random Forest model by significantly reducing false negatives and improving overall accuracy.
- **Interpretability**: SHAP (SHapley Additive exPlanations) values were incorporated to enhance model transparency, enabling stakeholders to understand which features contributed most to the predictions.
- **Output**: The final predictions were stored in the dataset final_train_test_data.csv.

2. Interactive Tableau Dashboards

The cleaned and updated dataset was used to generate visualizations that highlight key trends and insights. The Tableau dashboards include:

- **Distribution of Award Winners by Category**:
    - Key Insight: Categories like *Best Picture*, *Film Editing*, and *Directing* have the highest number of winners, highlighting their significance in the Oscars.
- **Year-Wise Award Win Prediction Probabilities**:
    - Key Insight: The average prediction probabilities for winners fluctuate significantly across decades, with certain peaks indicating years where models confidently predicted outcomes.
- **Comparison of Model Predictions vs. Actual Winners**:
    - Key Insight: The scatterplot clearly separates the predicted winners and non-winners, demonstrating the model's accuracy while also identifying borderline cases where predictions were less confident.
- **Prediction Accuracy by Award Category**:
    - Key Insight: Categories like *Costume Design*, *Documentary*, and *Writing* show varying levels of prediction accuracy, reflecting the model's performance across diverse award categories.
- **Correlation Between Critic and Audience Scores**:
    - Key Insight: A positive correlation exists between critic scores and audience scores, with higher critic scores showing a stronger association with Oscar-winning outcomes.
- **Win/Loss Prediction Probability Distributions (by Category)**:

    - Key Insight: Categories exhibit wide variability in win prediction probabilities, with some categories showing consistent trends in prediction accuracy while others display outliers.
- **Films with Highest Prediction Probabilities of Winning (by Category)**:
    - Key Insight: Movies like *The Godfather*, *All the King's Men*, and *Hamlet* achieved the highest prediction probabilities, particularly in the *Actor* category, aligning with historical Oscar trends.
- **Year-Wise Trends in Average Critics & Audience Scores**:
    - Key Insight: Over time, both critic and audience scores have shown an upward trend, indicating improvements in movie quality and growing audience engagement.
- **Analysis of Prediction Errors**:
    - Key Insight: Errors in prediction are concentrated in certain categories like *Sound Editing* and *Cinematography*, suggesting areas where the model could be further refined.

The dashboards are interactive, allowing stakeholders to filter data by year, genre, and category to explore insights relevant to their interests.

3. Supporting Documentation

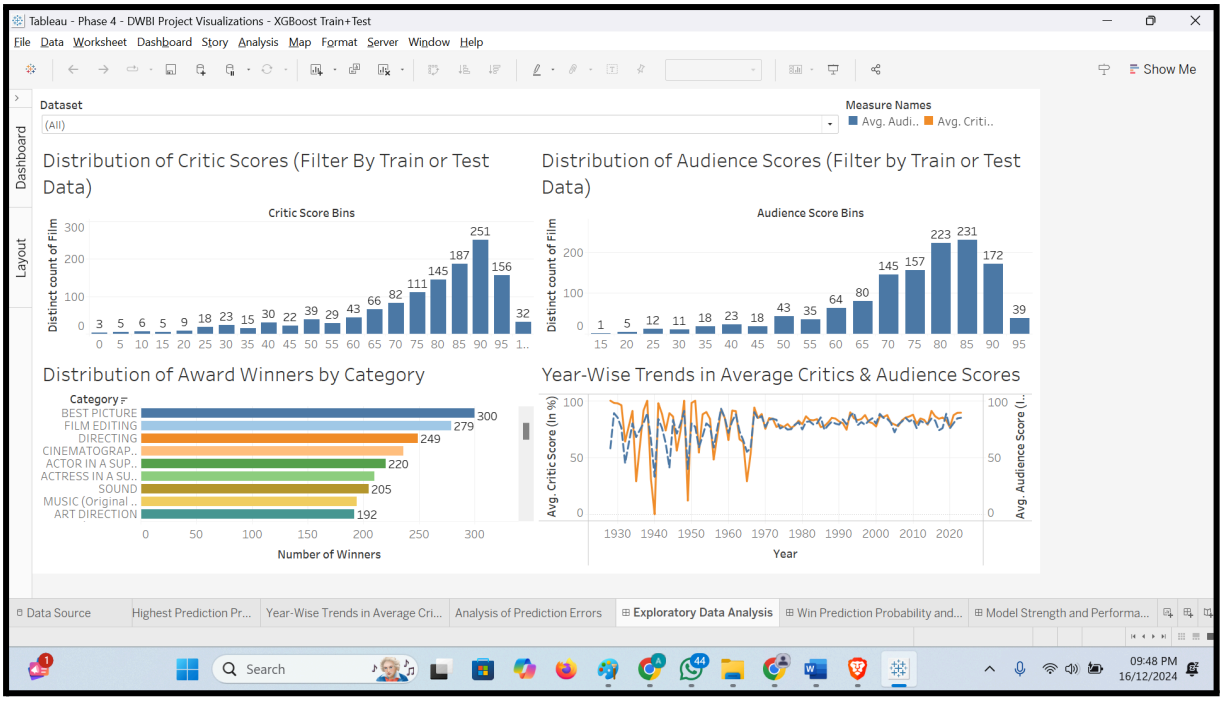
- **Project Report**: A detailed document describing the project lifecycle, methodologies, challenges, findings, and client feedback.
- **Code Notebooks**:
    - Movie_Dataset_Predictive_Analytics_XGBoost.ipynb: Contains the code for data preprocessing, model training, evaluation, and SHAP analysis.
    - Movie_Dataset_Predictive_Analytics (1).ipynb: Includes the initial Random Forest implementation and exploratory data analysis (EDA).
- **Datasets**:
    - final_train_test_data.csv: The final cleaned dataset used for predictions and visualizations.
    - Updated_Movie_Masterdata.csv: Enhanced dataset after feature engineering.

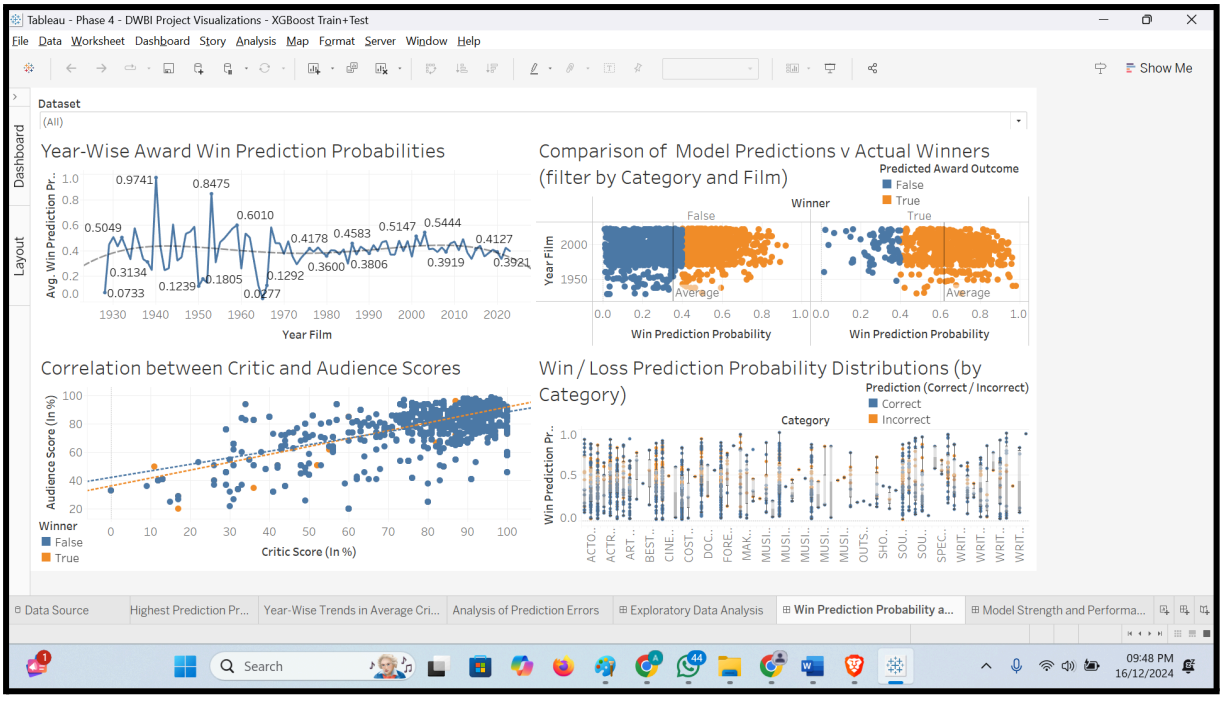
4. Key Insights

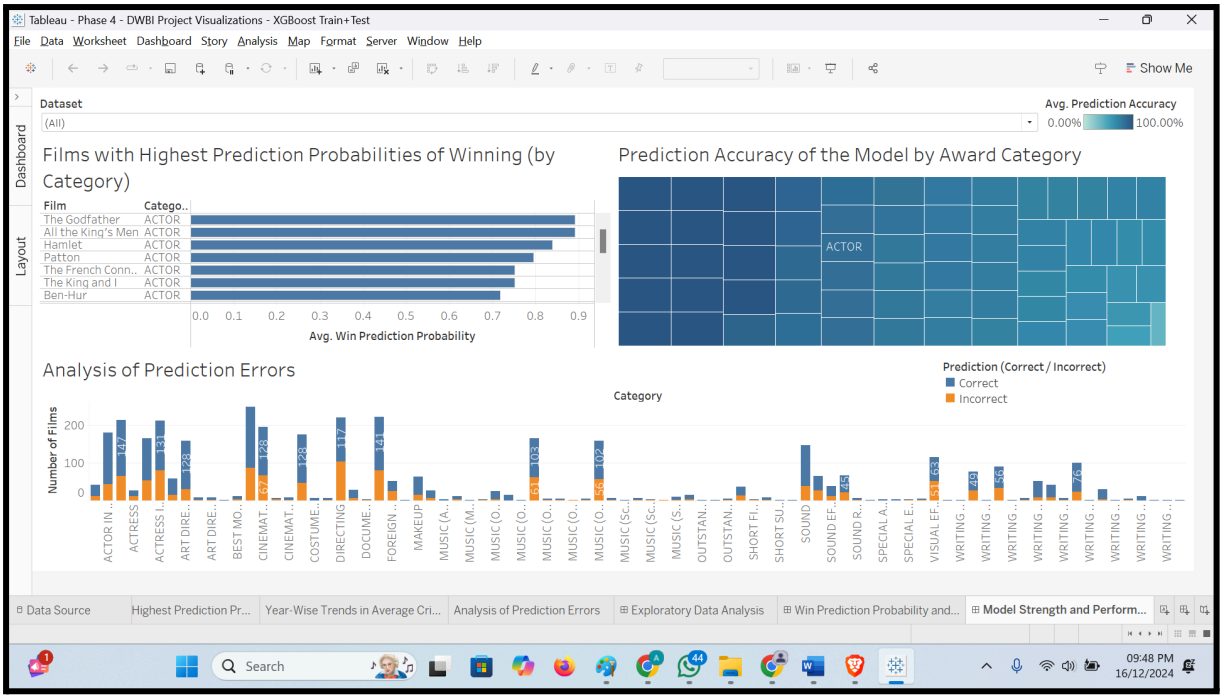
The final product successfully addressed the project objectives by providing:

- A reliable predictive model to forecast Oscar winners.
- Actionable insights into genre preferences, score correlations, and historical trends.
- Clear and interactive visualizations that allow stakeholders to derive insights specific to their requirements.

By combining machine learning with business intelligence tools, the project delivers a practical and insightful framework for understanding success factors in the film industry.







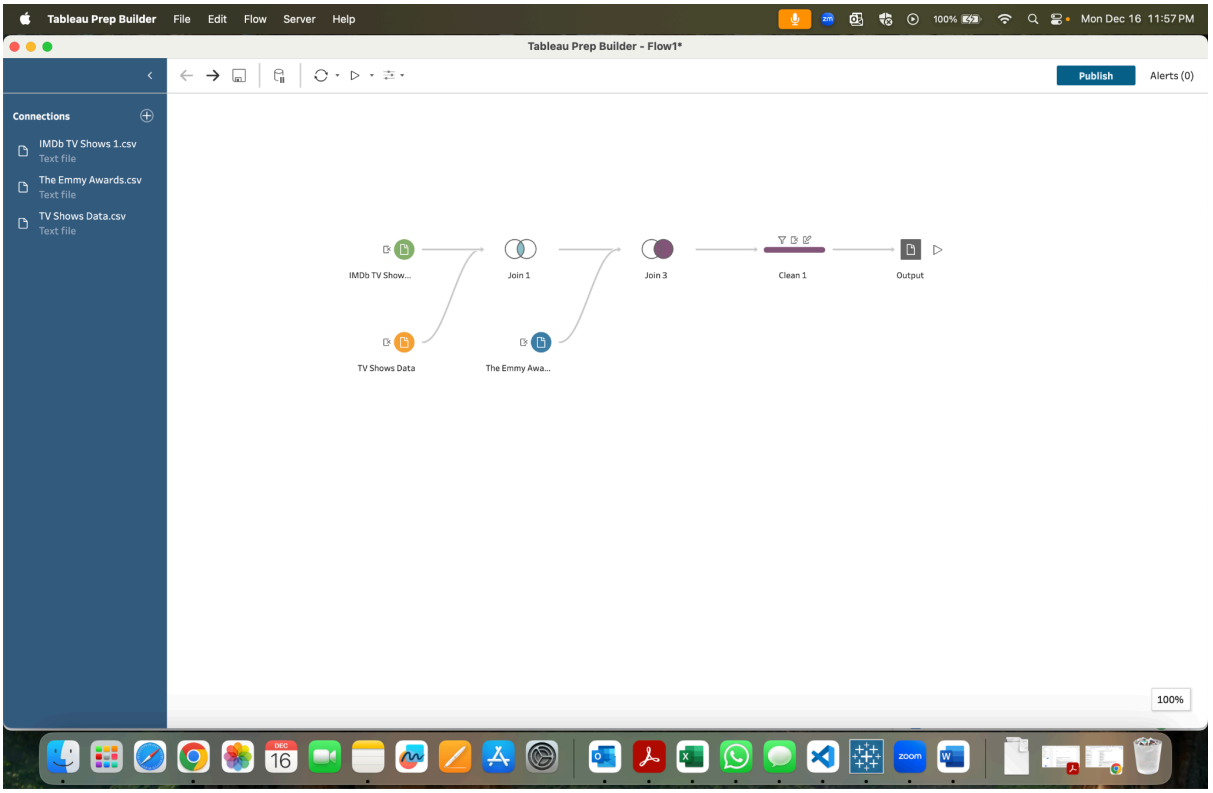
Phase - 1 : Procuring the Data and Exploring it's Features

The dataset "IMDb TV Shows" consists of various columns that provide key information about movies, shows, or episodes. The Title column contains the name of the content, while EpisodeDuration indicates its runtime, typically in minutes or hours and minutes. The Genres column specifies the category or categories, such as Drama, Comedy, or Action, that the content belongs to. The Actors column lists the primary actors or characters involved, showcasing their key roles. The Rating column reflects the average viewer rating on a scale, helping to assess the content's popularity. Votes show the total number of votes cast by viewers, indicating the level of engagement. Lastly, the Updated Years column tracks the year or years when new content or updates were released, providing insight into its recency and relevance.

Column Name	Description
Title	The name of the movie, show, or episode.
EpisodeDuration	The duration of the episode in minutes or hours and minutes.
Genres	The category or categories the movie/show belongs to (e.g., Drama, Comedy, etc.).
Actors	The list of main actors or characters featured in the movie/show.
Rating	The average rating given to the movie/show by viewers (e.g., on a scale of 1-10).
Votes	The total number of votes the movie/show has received from viewers.
Updated Years	The year(s) when the movie/show was last updated or when new content was released.

The dataset “TV Shows Data” provides comprehensive details about various movies, shows, and series. The **Name** column includes the title of the content, while **Average Runtime** gives the average length of each episode or movie, usually in minutes or hours. The **Updated Dates** track the dates or years when the content was updated or when new seasons or episodes were released. **Genres** describe the content’s category, like Drama, Action, or Comedy, and **Type** specifies whether it is a movie, TV show, or documentary. The **Language** column indicates the language in which the content is primarily produced, and **Network** reveals the platform or network hosting the content. **Rating** reflects the viewer’s average rating of the content, and **Total Seasons** and **Total Episodes** provide a count of the series' seasons and episodes. Lastly, the **Person Names** column lists key individuals associated with the content, such as actors, directors, and producers.

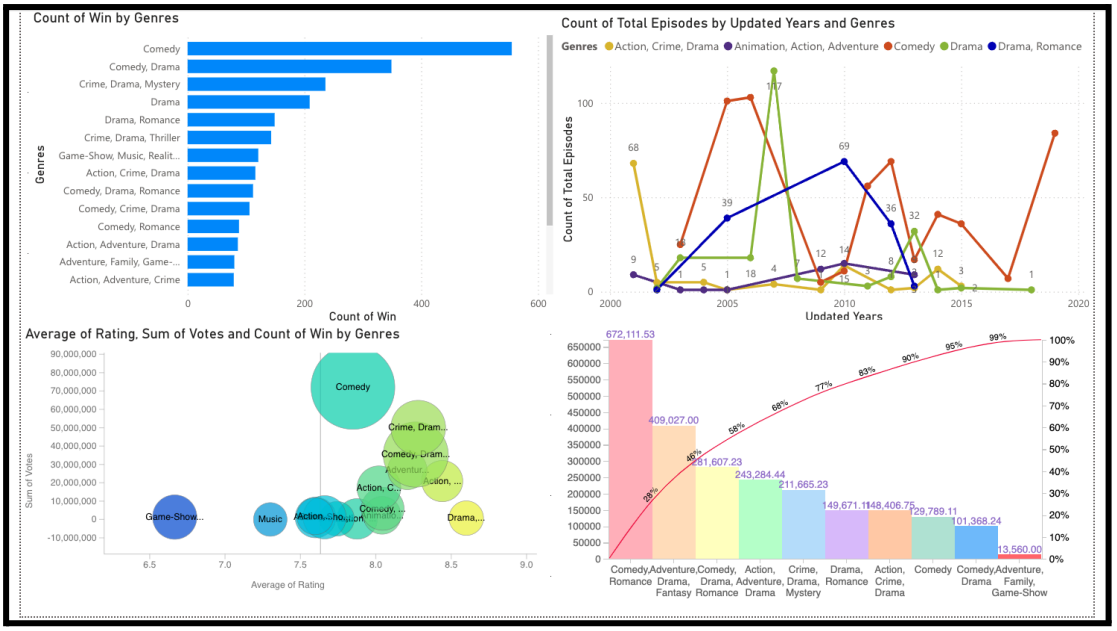
Column Name	Description
Name	The title of the TV show or movie.
Average Runtime	The average duration (in minutes) of each episode or the full runtime for a movie.
Updated Dates	The dates when the information about the TV show or movie was last updated.
Genres	A list of genres associated with the TV show or movie (e.g., Drama, Comedy, Thriller).
Type	The type of content, such as "TV Show" or "Movie."
Language	The primary language in which the TV show or movie is produced or broadcast.
Network	The production or broadcasting network (e.g., HBO, Netflix, ABC).
Rating	The average audience or critic rating for the TV show or movie, typically on a scale (e.g., 1–10).
Total Seasons	The total number of seasons for a TV show (if applicable).
Total Episodes	The total number of episodes produced for a TV show (if applicable).
Person Names	The names of key cast members or contributors (e.g., actors, directors, or writers).

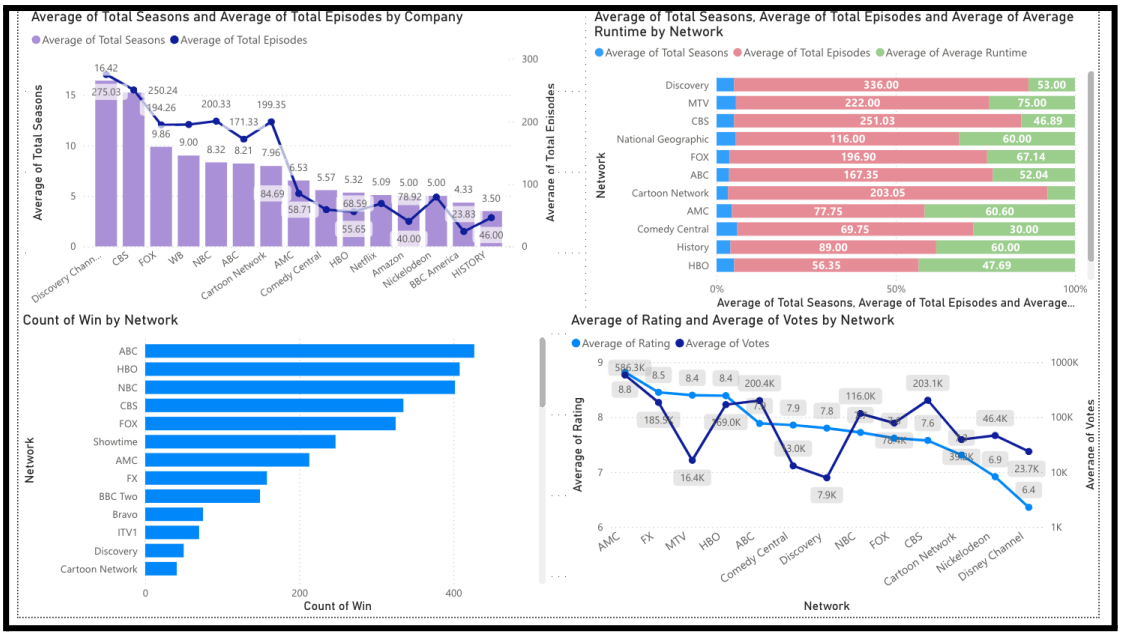


Final Product - Segment 2 - Descriptive and Historical Analysis of Emmy Award Winning TV Shows:

The deliverables for the **TV Shows Emmy Award Analysis** include:

1. **Power BI Dashboards**:
    - **Genre Insights**: Comedy and hybrid genres dominate Emmy wins.
    - **Network Analysis**: ABC, HBO, NBC, and CBS emerge as leading networks.
    - **Audience Trends**: High ratings and vote counts strongly influence Emmy success.
    - **Production Trends**: Traditional networks favor longer seasons, while streaming platforms produce shorter series.
2. **Supporting Documentation**:
    - **Report**: Comprehensive documentation of project lifecycle, analysis, findings, and challenges.
    - **Data Files**: Cleaned and merged dataset (Result Data.csv) used for analysis.


Count of Win by Genres
Comedy
Comedy, Drama
Crime, Drama, Mystery
Drama
Drama, Romance
Crime, Drama, Thriller
Game-Show, Music, Realit...
Action, Crime, Drama
Comedy, Drama, Romance
Comedy, Crime, Drama
Comedy, Romance
Action, Adventure, Drama
Adventure, Family, Game-...
Action, Adventure, Crime
Genres
Count of Win
Count of Total Episodes by Updated Years and Genres
Genres Action, Crime, Drama Animation, Action, Adventure Comedy Drama Drama, Romance
Count of Total Episodes
Updated Years
Average of Rating, Sum of Votes and Count of Win by Genres
Sum of Votes
Average of Rating
Comedy
Crime, Dram...
Comedy, Dram...
Action, C...
Game-Show...
Music
Drama,...
672,111.53
409,027.00
261,607.23
243,284.44
211,665.23
149,671.11
148,406.75
129,789.11
101,368.24
13,560.00
Comedy,Adventure,Romance
Comedy, Drama, Fantasy
Comedy, Drama, Romance
Action, Adventure, Drama
Crime, Drama, Mystery
Drama, Romance
Action, Crime, Drama
Comedy
Comedy,Adventure, Drama
Family, Game-Show


Average of Total Seasons and Average of Total Episodes by Company
Average of Total Seasons
Average of Total Episodes
Discovery Chann...
CBS
FOX
WB
NBC
ABC
Cartoon Network
AMC
Comedy Central
HBO
Netflix
Amazon
Nickelodeon
BBC America
HISTORY
Average of Total Seasons, Average of Total Episodes and Average of Average Runtime by Network
Average of Total Seasons
Average of Total Episodes
Average of Average Runtime
Discovery
MTV
CBS
National Geographic
FOX
ABC
Cartoon Network
AMC
Comedy Central
History
HBO
Network
Average of Total Seasons, Average of Total Episodes and Average...
Count of Win by Network
ABC
HBO
NBC
CBS
FOX
Showtime
AMC
FX
BBC Two
Bravo
ITV1
Discovery
Cartoon Network
Network
Count of Win
Average of Rating and Average of Votes by Network
Average of Rating
Average of Votes
AMC
FX
MTV
HBO
ABC
Comedy Central
Discovery
NBC
FOX
CBS
Cartoon Network
Nickelodeon
Disney Channel
Network