Demographics of Moviegoers:

Predicting Profitable Movies and Increasing Sales

Christopher Cooley, Angelo Kodra, Reyna Macabebe

Donovan A. Romaya, Emma Russo, Kleant Topalli, Mary Jabro

Wayne State University

BE2100 – Probability and Statistics

Sara Masoud

2/28/2024

Table of Contents

Abstract

The dynamic film industry mirrors the ever-evolving world, shaped by technological innovations and shifting audiences. This study delves into the intersection of demographics and movie preferences, investigating age groups, genres, preferred movie lengths, and viewing platforms. The research uses predictive analysis techniques to forecast individuals' tendency toward visiting movie theaters based on various demographic and behavioral variables. The methodology employs Google Forms and online datasets to gather comprehensive data and analysis using Python for predictive modeling and regression analysis. While the study initially hypothesized that younger demographics prefer shorter, mildly intense movies, the results reveal intriguing insights. While no significant correlation was found between age and movie length preferences in the survey dataset, a significant association emerged from Statista's broader dataset. This discrepancy underscores the importance of considering external data sources to understand audience preferences comprehensively. Our study confirmed our hypothesis that younger people enjoy mildly intense movies, but we also learned that younger people want short and long films equally. Overall, the study provides valuable insights for industry stakeholders, aiding in strategic decision-making regarding marketing, content creation, and audience engagement strategies.

Introduction

Like the world, the film industry is constantly changing, driven by new technologies and changing tastes among moviegoers. The rise of movies adapted from video games is a great example. Older adaptations like "Super Mario Bros" (1994), "Doom" (2005), and "Rampage" (2018) didn't do well with audiences and told movie executives that video game franchises weren't worth it. However, recent hits like "Sonic" (2020), "Five Nights at Freddy's" (2023), and the new "Super Mario Bros Movie" (2023) have tapped into a growing audience that loves seeing their favorite games on the big screen, mirroring the changing world and their changing preferences.

Movies play a significant role in shaping our culture and conversations online. That's why it's essential for the people making and selling movies to understand what viewers want. As viewers' habits and preferences change, we need a better understanding of who is watching what, how long they want their movies to be, and whether they prefer to watch them in theaters or at home.

This study uses data analysis and predictive modeling to determine which viewers are most likely to go to the movies using data collected through Google Forms and online datasets. We're looking closely at different age groups, what genres they like, and how long they want their movies to be. Our goal is to build models that can accurately identify who is likely to be a moviegoer and who isn't. This information could help movie makers and marketers make smarter choices that appeal more to their audiences.

We hypothesize that the best group to focus on might be younger viewers under 24. They generally have more free time and disposable income and might prefer fast-paced movies like action or thrillers that aren't too long. With this study, we want to provide solid data that helps the film industry engage these viewers more effectively, enhancing their experiences and bringing in better revenue.

Methodology

We used two primary methods to collect the data needed for our analysis: Google Forms and larger online datasets. By integrating responses from Google Forms with broader online datasets, we enhanced our study into profitable movie theater demographics. Our survey targeted vital variables such as age, gender, the average number of movies watched per year, preferred movie length, favorite genre, and the choice between watching movies at the theater or streaming. Combining these sources allowed us to compensate for the smaller sample size of our primary data, reducing bias and better representing a diverse public audience.

We used Python to conduct predictive modeling and generate visual data representations for data analysis. This approach helped simplify the interpretation and organization of our results. Python was instrumental throughout our study, enabling efficient data management and complex regression analyses to study the relationships between variables. We used various Python libraries, including pandas, NumPy, matplotlib, and seaborn, which were crucial for manipulating data, visualizing distributions, and performing statistical analyses. These tools helped us create detailed visualizations such as density plots, correlation matrices, and outputs from regression analyses, making it easier to understand and present our findings.

Our predictive analysis employed regression techniques to investigate the relationships between demographic factors and moviegoing behavior. Logistic regression models, in particular, were used to estimate the likelihood of individuals attending movies based on their demographic and preference data. This modeling provided valuable insights into how different factors influenced movie attendance, enhancing our ability to make informed predictions about audience behavior.

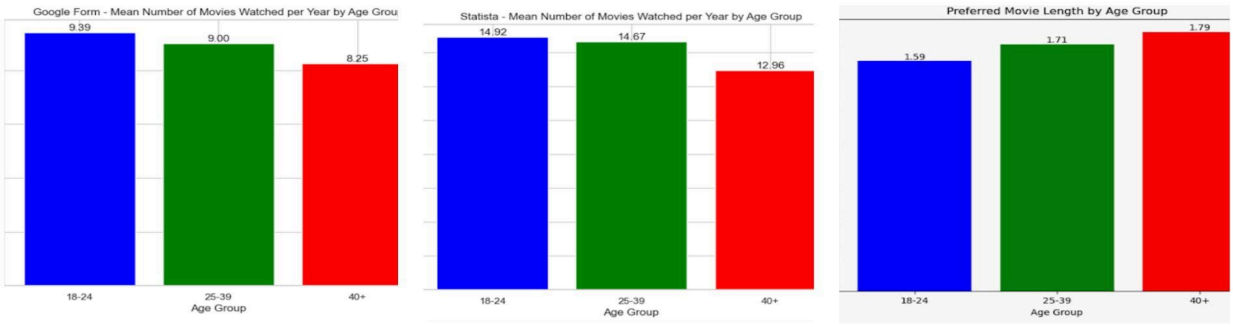
Additionally, we conducted a thorough correlation analysis to examine the interrelationships among the collected variables. Using Python's capabilities, we built a comprehensive correlation matrix and employed graphical tools to visualize these relationships. This part of our analysis was critical in understanding how different variables interacted with each other, influencing moviegoing preferences and behaviors.

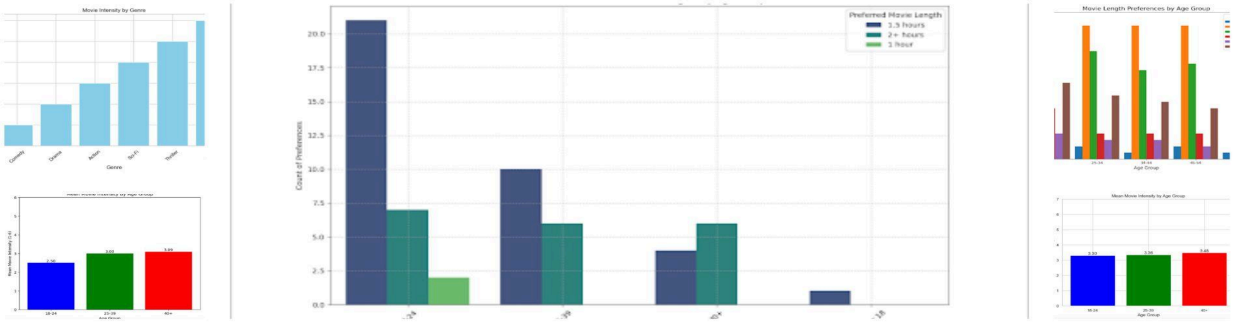
Results

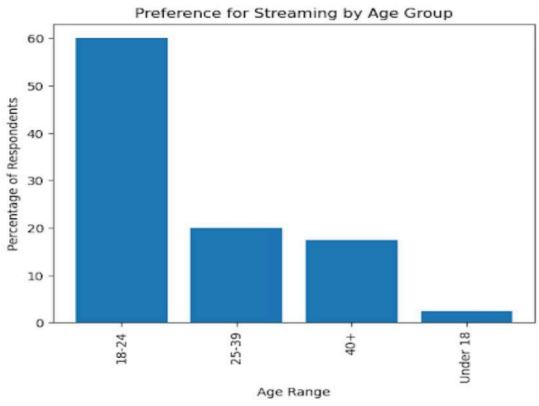
I. Descriptive Statistics

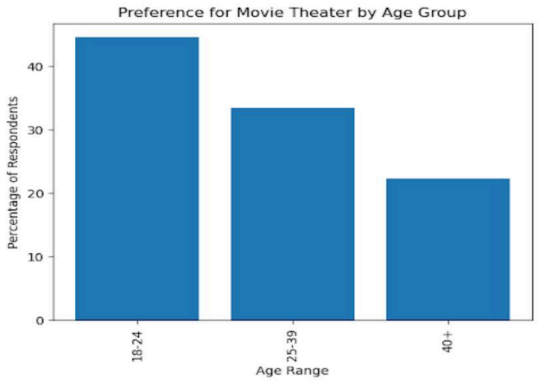
To find our statistics, we used two primary sources: statistics by Statista for the online datasheet and a Google form survey conducted by this group, which garnered around 60 respondents. We compared and contrasted both sources to make sure our collected data matched up with a larger dataset.

Representation of the distributions of each of the variables is shown below:









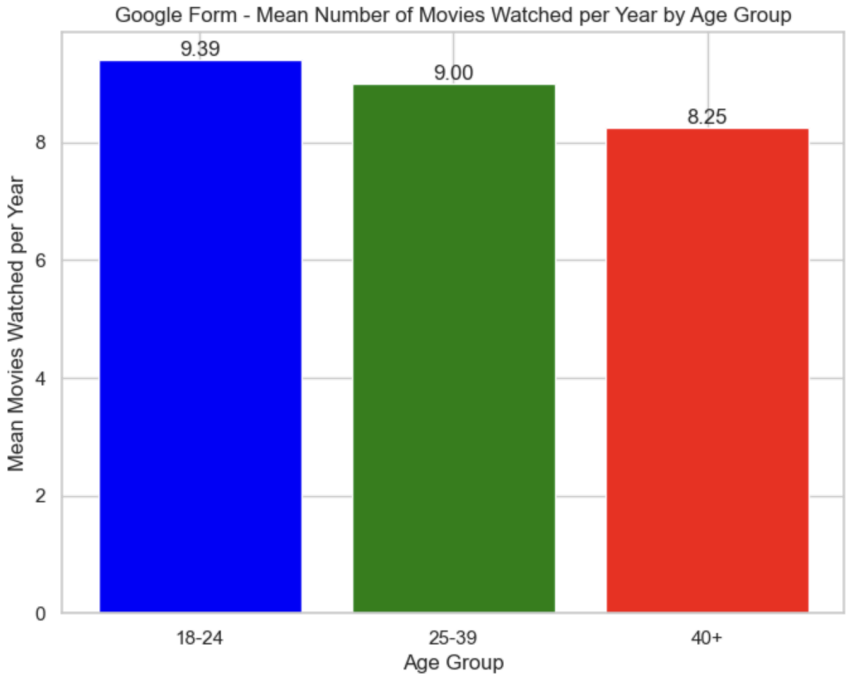


Figure 1: Google Form Survey- Age groups Mean Movie watching frequency

Our dataset reveals some surprising details about the film industry's business. For one, the mean of movie watchers is the highest in the age range of 18-24. This isn't all that surprising as people around 18-24 tend to be drawn towards movies as a primary form of entertainment. Along with this, it shows that people around the age of 40 and above are less likely to watch a movie. This is likely for many different reasons. Younger people are more likely to go out and socialize with others. This is probably because younger people have much more free time and disposable income than adults with a profession and a family. From our survey data, People between the ages of 25-39 are not as likely to watch movies as their younger counterparts. The survey contains a diverse group of different entries from all age ranges and groups.

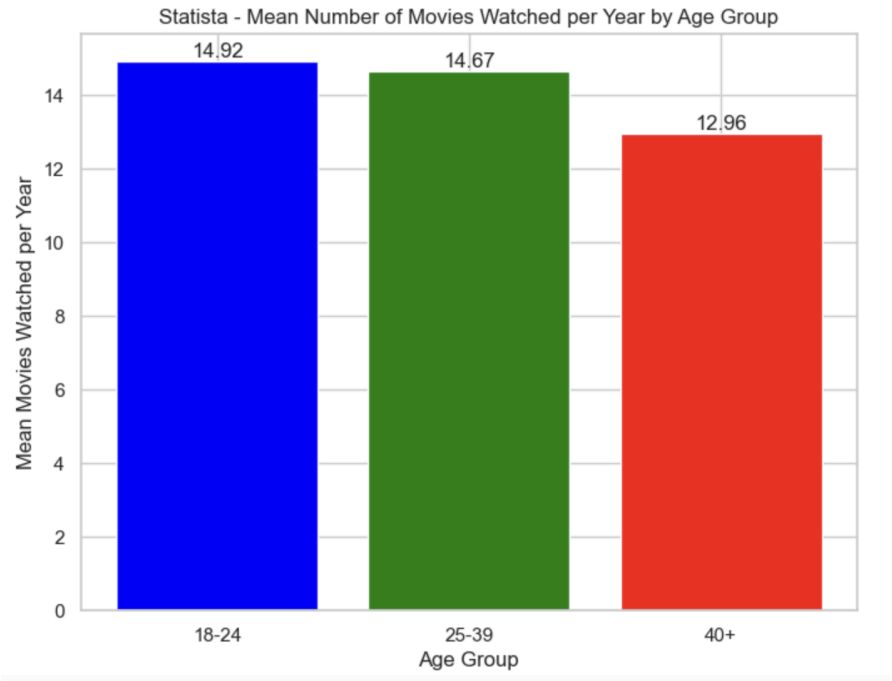


Figure 2: Statista Data - Mean frequency of watching movies among adults by age group

Based on the findings of the Google survey, the information provided by Statista follows the same trend as what we gathered. People in the 18-24 age range tend to watch more movies than those in older age groups. It agrees with the data we gathered from the Google Survey, which gave us a clear answer on what age group is likelier to watch a movie.

Overall, the age group of 18-24 is more likely to watch movies than those aged 25-39, 40 and above. With this information, the film industry can focus more on promoting movies that are oriented toward their intended fan base or audience. Movies are meant to be central stepping stones in pop culture; by promoting movies oriented towards younger age groups, we are pushing it further and causing it to grow for future generations.

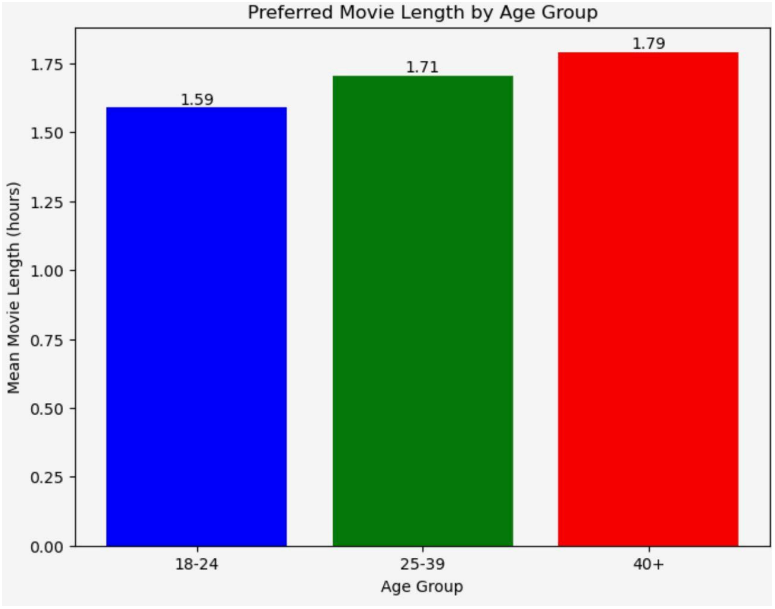


Figure 3: Preferred Movie Length by Age Group

We targeted specific questions to categorize preferences by age group to identify trends related to age groups and preferred movie lengths. Our central hypothesis is that younger people tend to watch shorter, around 1.5-hour-long movies and that we should cater to that demographic because they are the largest demographic of moviegoers. We want to know if there is a correlation between age and preferred movie length to determine if our hypothesis is correct. Due to the categorical nature of the data obtained, the best test to run to look for trends was a Chi-square test. This test would allow us to determine if a statistically significant association exists between the respondents' age and their preferences.

- The null hypothesis, H_0, is "there is no association between age group and movie length preference."
- The alternative hypothesis, H_a, is "there is an association between age group and movie length preference."

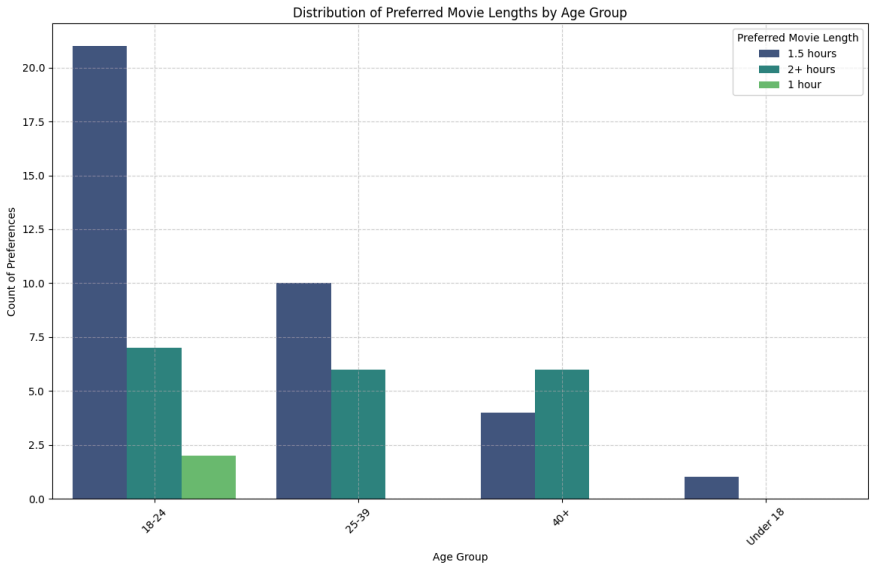


Figure 4: Distribution of Preferred Movie Lengths by Age Group

Based on our Google Form responses, this bar graph shows the distribution of preferred movie lengths by age group. The X-axis represents the "Age Group," while the Y-axis represents the "Count of Preferences." The dark blue bar represents a 1.5-hour movie length, a teal 2+ hour movie length, and a green 1-hour movie length. Visually analyzing the data shows that an older age group (25 and older) prefers a longer movie length slightly more than a younger age group (24 and younger).

Conducting the Chi-square test on our dataset resulted in these values:

- Chi-square Statistic: 6.5375000000000005
- P-value: 0.3657409666272405
- Degrees of Freedom: 6
- Expected

  [[1.05263158 18.94736842 10.]

  [0.56140351 10.10526316 5.33333333]

  [0.35087719 6.31578947 3.33333333]

  [0.03508772 0.63157895 0.33333333]]

The p-value is notably higher than the conventional thresholds for statistical significance (usually 0.05 or 0.01), suggesting no substantial evidence to reject the null hypothesis. The test results indicate no statistically significant association between the age groups of the respondents and their preferred movie lengths. The expected frequencies calculated under the null hypothesis were relatively close to the observed frequencies across different age groups and movie length categories. This leads to the conclusion that the variations in movie length preferences are likely due to chance rather than any underlying solid trends related to the age of the respondents.

Given the insights from our Chi-square analysis, which suggest no significant variation in movie length preferences across different age groups, it's intriguing to compare these findings with external data sources like Statista, which might offer a broader context. The dataset that we collected, “Preferred length of movies among consumers in the U.S. 2018, by age” from Statista, was nearly identical to the data that we collected in that it collected data about age range and preferred movie lengths (only in ranges, as opposed to our exact lengths for a couple of options.) They collected the data via an online survey and had 24,001 respondents. Due to the similarities in data collection, we also used a Chi-square test on this dataset.

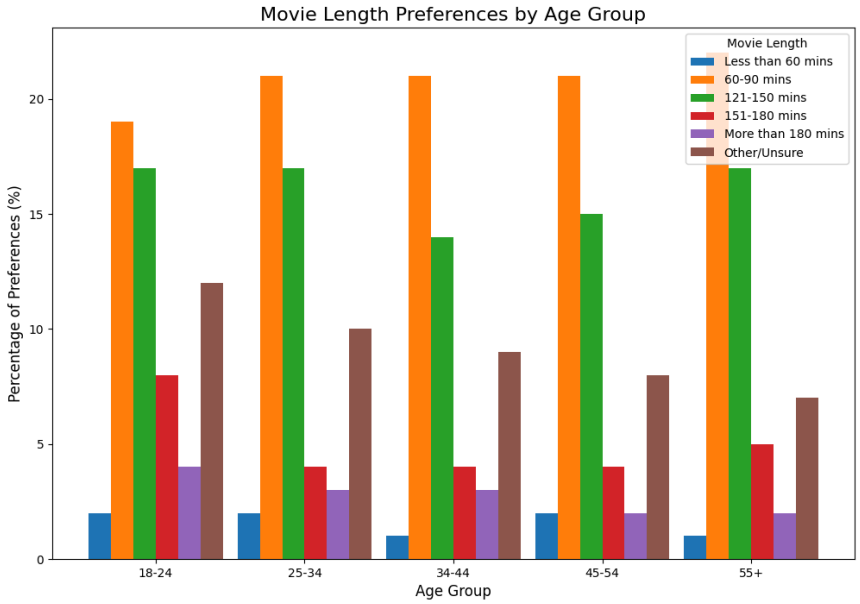


Figure 5:Statista Data - Movie Length Preferences by Age Group

This bar chart illustrates the percentage of respondents within each age group who prefer various movie lengths. Preferences are segmented into five categories: movies less than 60 minutes (blue), 60-90 minutes (orange), 121-150 minutes (green), 151-180 minutes (red), and more than 180 minutes (purple), as well as an 'Other/Unsure' category (brown). The x-axis categorizes respondents by age group, while the y-axis shows the percentage of preferences. The data suggests nuanced preferences for movie lengths, with the 18-24 age group showing a notable inclination towards movies that are 151-180 minutes long.

Based on the Statista survey data and the Chi-square test conducted:

- The Chi-square statistic is approximately 1200.20.
- The p-value is extremely small (around $6.78\times10-2426.78\times10-242$), far below any conventional significance level (e.g., 0.05 or 0.01).
- The degrees of freedom (df) for this test are 20.

Given the p-value, we reject the null hypothesis, which means a statistically significant association exists between the respondents' age groups and their preferred movie lengths. The differences in preferences are not due to random chance alone; there is a discernible pattern of preferences among different age groups regarding the size of movies they prefer to watch.

This is important because the Chi-square test conducted on our survey determined no significant

relation between age and preferred movie length, as opposed to the Chi-square test performed on the Statista dataset, which determined a significant relation between age and preferred movie length. This may be due to the small sample size for our Google form, as opposed to the much larger Statista population size. To further analyze the results from this test, we examined the adjusted standardized residuals. The results were as follows:

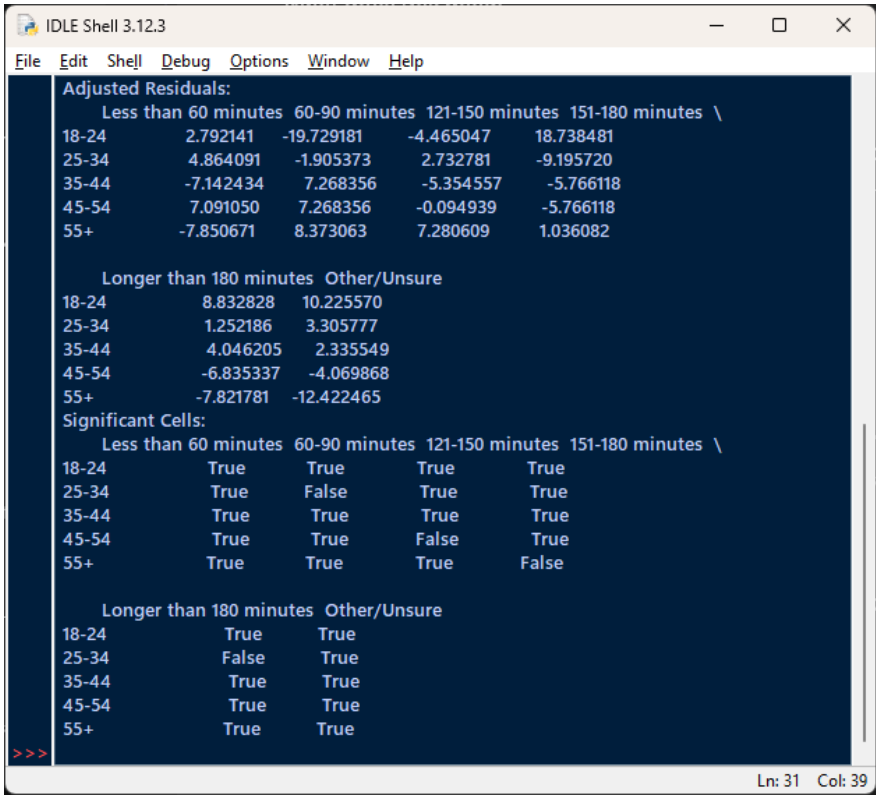


Figure 6: Adjusted Residuals Analysis

This figure is a screenshot of the output adjusted residuals and table of significant cells that allow us to determine which age group prefers which movie length. A positive residual indicates more observations than expected, while a negative residual indicates fewer. A residual greater than 2 or less than -2 is considered significant.

Based on this, we can draw that:

- For the **18-24 age group**, there's a significant preference for movies 151-180 minutes long and a substantial aversion to movies 60-90 minutes long.
- The **25-34 age group** significantly prefers shorter movies (less than 60 minutes) and movies 121-150 minutes long, but not for the other categories.
- **35-44-year-olds** significantly prefer shorter movies, and significantly fewer prefer movies that are 60-90 and 121-150 minutes long.
- The **45-54 age group** has a significantly greater preference for shorter movies and markedly less for films over 180 minutes, and it is in the "Other/Unsure" category.

- For those **55+**, there's a significant aversion to movies 60-90 minutes long, a significant preference for movies 121-150 minutes long, and a significant aversion for the "Other/Unsure" category.

Our initial hypothesis included that younger groups would prefer a shorter movie length. However, according to the adjusted residuals from the Chi-square test, the data does not support our hypothesis. Contrary to expectations, the significant positive residual in the 18-24 age group for movies ranging from 151-180 minutes suggests a preference for longer movies, while the negative residual for the 60-90 minute category indicates less preference than expected for mid-length films. Similarly, the 25-34 age group displayed a significant preference for movies less than 60 minutes and those 121-150 minutes long, showing a more varied set of preferences that include shorter and longer films, not exclusively shorter ones. These findings challenge our initial hypothesis and suggest that younger viewers' preferences for movie lengths may not lean towards shorter films as anticipated. Instead, they demonstrate a nuanced pattern of preferences that includes a substantial inclination towards longer movies, which may reflect complexity in viewing habits not captured by the original assumption.

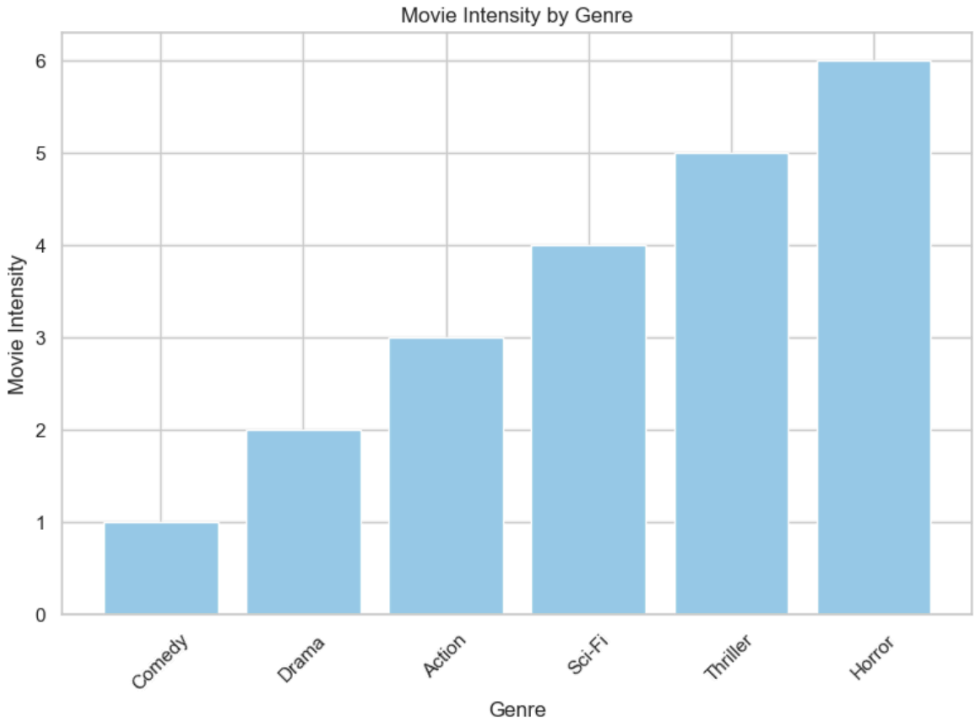


Figure 7: Movie Intensity Scale

We separated movie intensity into six different genres. We needed to generalize the genres into numerical values, so we measured them by intensity. Low-stake genres like comedy would be 1, and higher-stake films would increase to 6. Our genres include low-intensity comedy and drama (levels 1-2), middle-intensity action and sci-fi (levels 3-4), and finally, high-intensity thriller and horror (levels 5-6). We aim to see if movie intensity and age correlate to test our hypothesis that younger people enjoy high-intensity movies more than low-intensity ones.

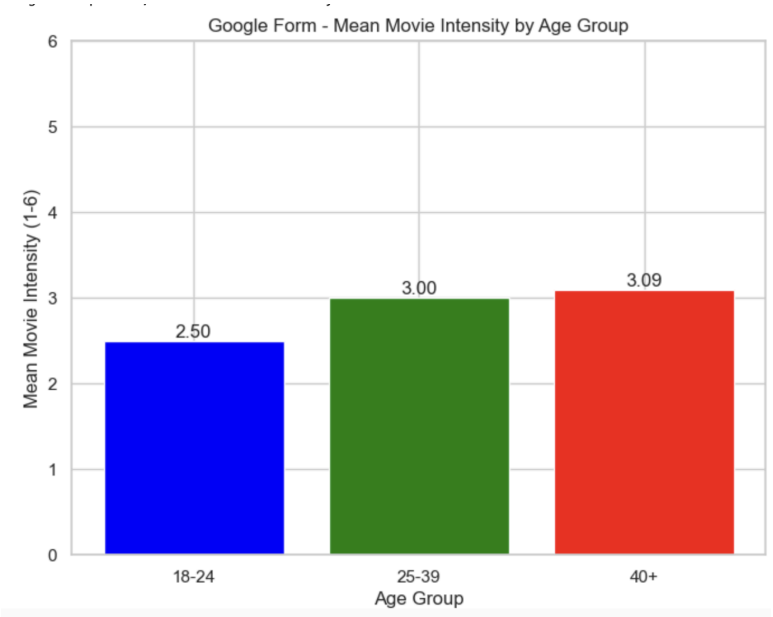


Figure 8: Google Form Survey - Age Groups Mean Movie Intensity Favorability

We asked each respondent about their favorite movie genre and generalized it to fit the numerical values. We collected all the respondents in each age group and found the mean intensity value for each group. The analysis of our dataset reveals a compelling trend in movie intensity preferences across age groups, supported by empirical evidence indicating a positive correlation between age and the inclination towards movies with higher intensity levels. Our comprehensive examination encompasses diverse age demographics, each associated with corresponding movie intensity ratings on a scale from 1 to 6, aligning with specific genres.

Upon closer examination, it becomes evident that younger age groups, particularly those aged 18 to 24, predominantly favor movies with moderate intensity levels, with mean ratings ranging from 2 to 3. This preference suggests a bias towards genres such as Drama and Action, known for engaging narratives without excessive intensity. Contrastingly, as age demographics progress towards older cohorts, there is a noticeable shift towards slightly higher movie intensity ratings. Individuals aged 40 and above consistently assign movie intensity ratings slightly above three on average, indicating a preference for genres characterized by heightened intensity, such as Thrillers and Action films.

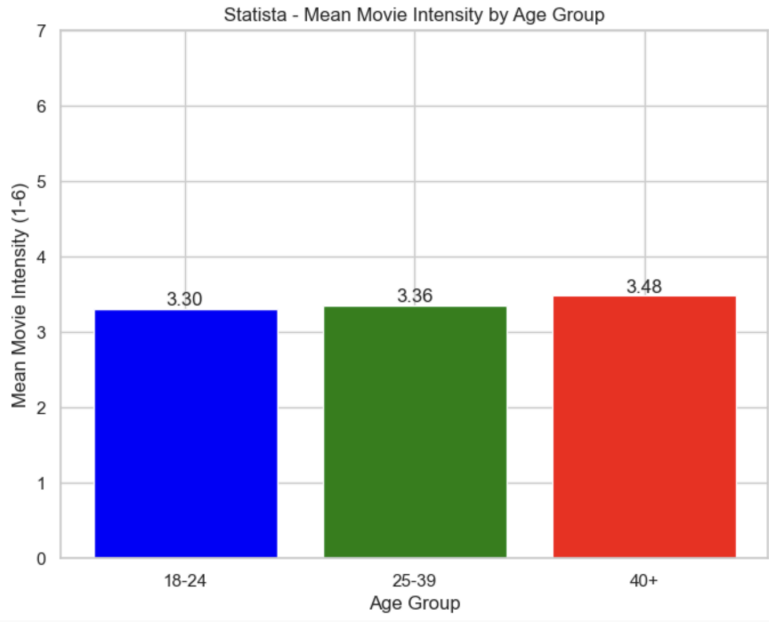


Figure 9: Statista Survey - Age groups Mean Movie Intensity favorability

Intriguing insights emerge when comparing our findings with the survey data from Statista (Figure 9). While our survey indicates a mean movie intensity of 2.5 for the 18-24 age group, Statista reports a slightly higher mean intensity of 3.3 for the same demographic. Similarly, the 25-39 age group in our survey exhibits a mean intensity of 3.00, while Statista's data suggests a slightly higher mean intensity of 3.36. Moreover, our survey indicates a mean intensity of 3.09 for the 40+ age group, while Statista's data reveals a somewhat higher mean intensity of 3.48. This observation underscores the dynamic nature of movie intensity preferences across age groups. These figures reinforce the positive correlation between age and movie intensity preferences. There is variability within each age group. However, the overall pattern supports the contention that as individuals mature, they gravitate towards more stimulating and emotionally engaging cinematic movies, typically associated with genres featuring higher intensity levels.

While we were able to prove that there is a correlation between age and preferred movie intensity, both our Google form data and the Statista data go against our hypothesis. We assumed that 18-to 24-year-olds would enjoy high-intensity films more than older generations, but that doesn’t seem true. We can see now that as people age, they enjoy higher intensity in movies. So, 18-to 24-year-olds enjoy a middle ground for intensity (3-4), while older generations slowly increase to high intensity.

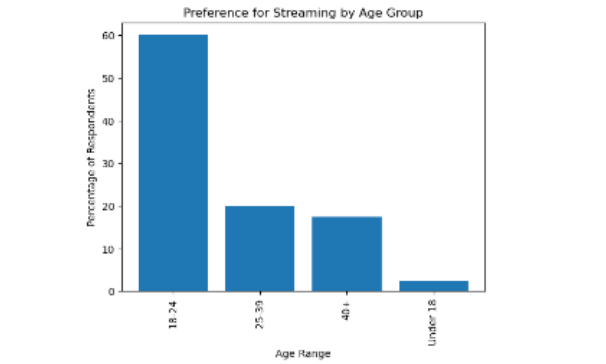


Figure 10: Preference for Movie Viewing via Streaming by Age Group

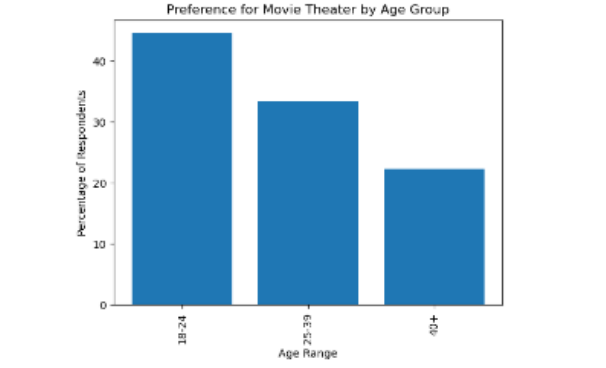


Figure 11: Preference for Movie Viewing via Theater by Age Group

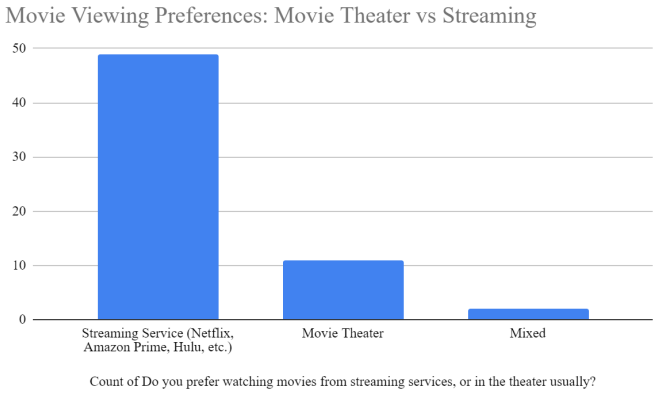


Figure 12: Most Popular Movie Viewing Preference

We collected data on how each age group prefers to watch movies in movie theaters, streaming, or both. This data is necessary for film executives, as the revenue made from movie theaters is needed to offset the costs of film production and garner profit from the movie itself. This study will help film producers anticipate how much revenue they will make by knowing how the overall population will consume their products.

Our collected data clearly shows a massive bias towards streaming services in the younger demographic. This can be due to many reasons; movie theater tickets can be expensive, especially for younger people. The ease of streaming can also be a factor, as younger people tend to prefer accessibility more than the experience itself. Older demographics also like watching films via streaming, but a much more significant percentage of older people prefer movie theaters more than the younger demographics. As mentioned, this could be due to older generations enjoying the experience of going to the movie theater more than just sitting at home. Streaming services are also a newer type of technology, which explains why younger people are more comfortable using it than older people.

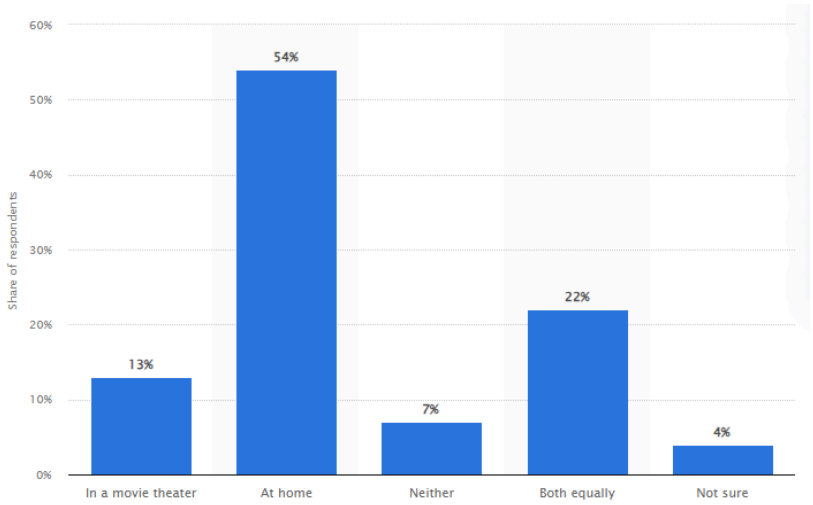


Figure 13: Statista Survey - Preferred Movie-Watching Locations Among Adults

Statistia's survey matches up with our own. Given the common trend among both our survey and the larger sample Statista survey, it is clear that there is a common preference for At-Home movie viewing. This could be used by movie production companies to market more movies towards streaming services after an in-theater release or even a straight-to-home release. This could also be used to justify a parallel release in movie theaters and on streaming services; doing so would target both people who prefer viewing movies in a theater and those who prefer streaming services.

Our data for movie-watching preferences among age groups shows that among those who prefer streaming services, most of these respondents are 18-24. However, for our portion of respondents who stated they preferred watching movies in movie theaters, the distribution of answers was much more equal among all age ranges. Using this information, movie companies can infer that as time passes, more and more of their target audience will prefer at-home movie viewing and market their releases to accommodate that preference.

II. Predictive Analysis

For predictive analysis, we ran tests to predict if a person goes to the movies based on age, preferred genre, and preferred movie length. We chose if a person goes to the movies as our target variable because movie theater profit depends on the number of people watching movies.

1. Establishing variables

In this initial phase of our data analysis, we started by loading and examining a dataset detailing information about moviegoers, encompassing factors like age, number of movies watched per year, movie length, and movie intensity. We employed descriptive statistics to grasp the dataset's structure and values, identified vital features, and defined our target variable as whether a person goes to the movies.

2. Confusion Matrix

A confusion matrix is generated to assess the model's performance, displaying the number of correct and incorrect predictions. Our confusion matrix is shown below:

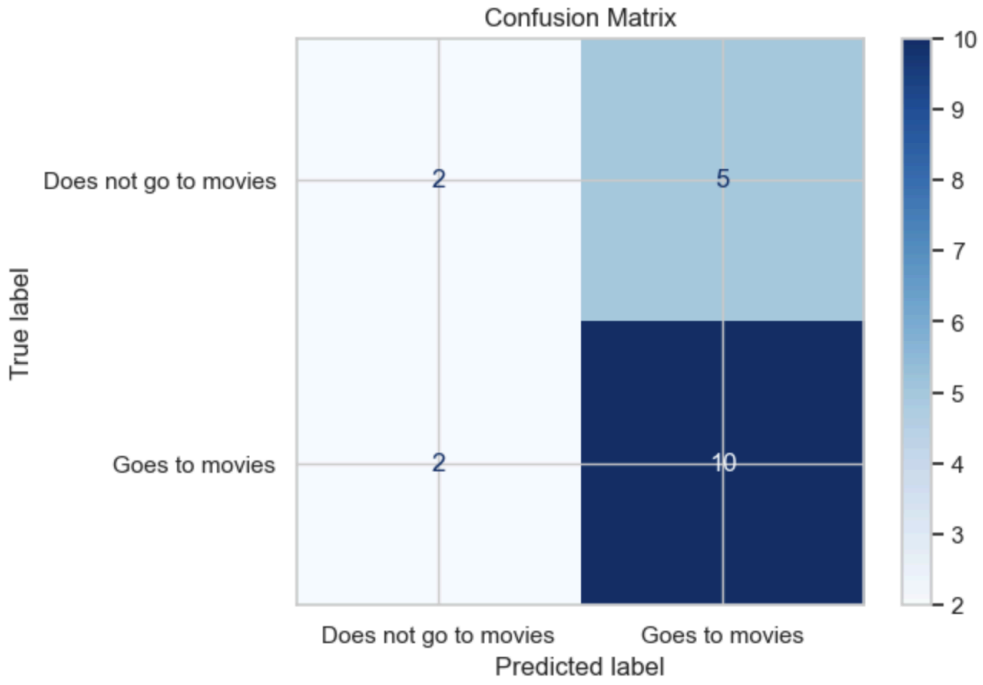


Figure 14: Confusion Matrix: Going to Movies Prediction

Metrics:

```
Precision: 0.67
Accuracy: 0.63
Recall: 0.83
F1 Score: 0.74
```

3. Decision Tree

We utilized a decision tree classifier to predict individuals' likelihood of going to the movies based on features such as age, movie intensity, and movie length. Understanding the factors influencing an individual's decision to visit movie theaters is paramount for businesses and marketers. Using machine learning techniques such as decision trees allows for exploring intricate patterns within demographic and behavioral data to predict whether a person is likely to frequent movie theaters.

We employ several Python libraries to conduct this analysis, including pandas for handling and preprocessing the dataset, Scikit-learn for building and evaluating the decision tree classifier, and Graphviz for visualizing the resulting decision tree. The dataset, containing information on gender, age, movie intensity, movie length, and movie attendance, is loaded into a DataFrame using pandas. The decision tree model is trained using the DecisionTreeClassifier from Scikit-learn, with features such as age, gender, movie intensity, and movie length as predictors of movie theater attendance. Performance metrics such as accuracy, precision, recall, and F1-score are calculated to evaluate the model's predictive capability. Finally, the resulting decision tree is visualized using Graphviz, providing insights into the key factors influencing movie theater attendance. Below is our decision tree displayed:

]:

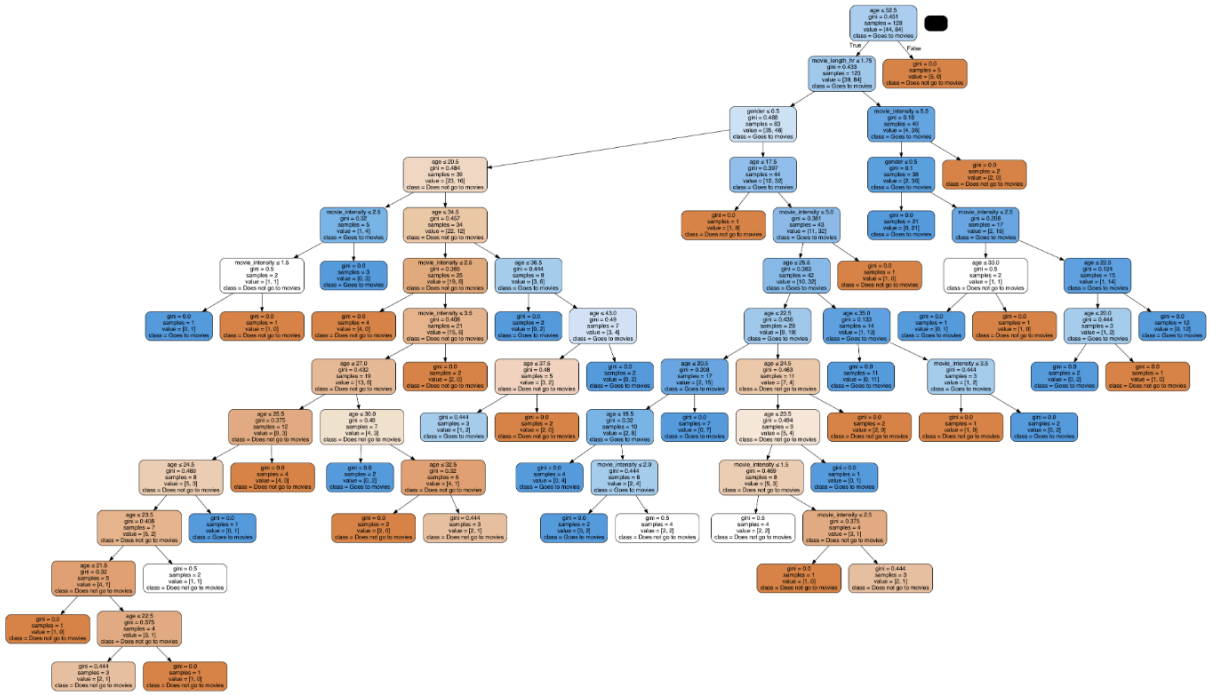


Figure 15: Decision Tree - Likeliness to Go to the Movie Theater

Mean Square Error (MSE):

```
Training MSE: 0.0703125
Test MSE: 0.23214285714285715
```

In addition to evaluating the model's performance metrics, a closer examination of specific nodes within the decision tree provides valuable insights into its predictive capabilities. Notably, node 6 emerges as a standout feature in the decision tree analysis. With an accuracy of 100%, node 6 represents a subset

of the data characterized by distinct criteria, which the model can confidently classify with perfect accuracy. Below is displayed Node 6 with its subsequent nodes:

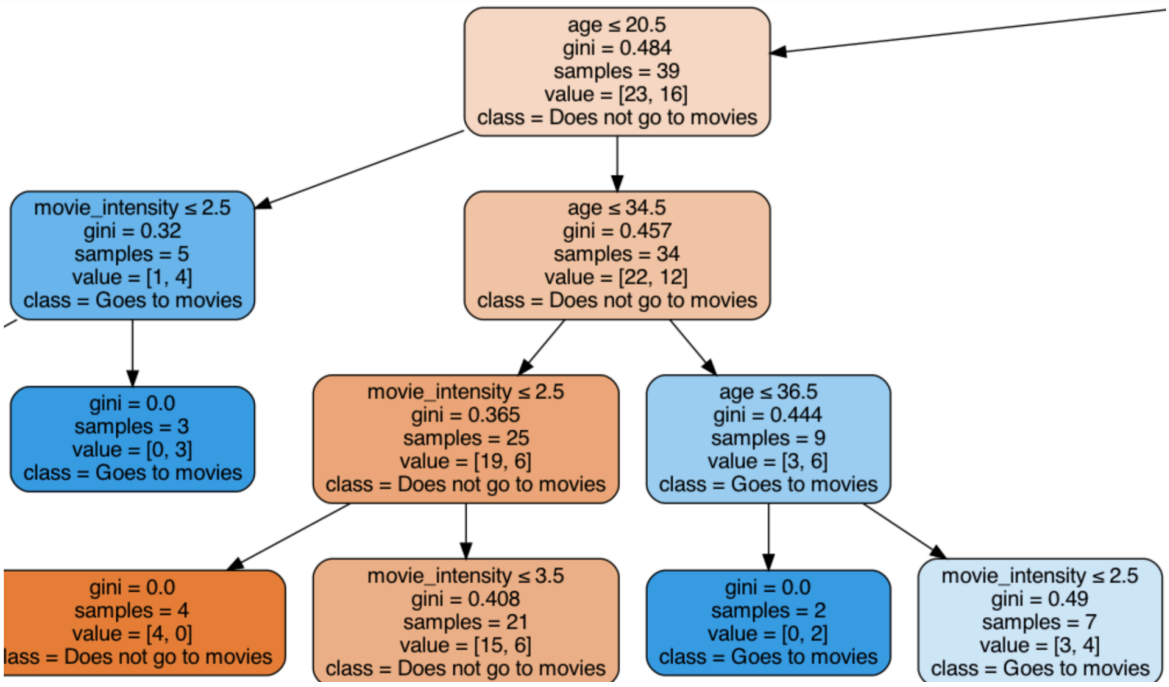


Figure 16: Close-up of Node 6 of the Decision Tree

4. ANN Model

Artificial Neural Networks (ANNs) are a vital machine learning technology inspired by how the human brain works. These models are made up of nodes, or "neurons," arranged in layers connected. ANNs can learn complex patterns and relationships in data, making them useful for classification, regression, and pattern recognition tasks.

Our study used a Sequential model architecture from TensorFlow's Keras API to build an ANN. This model type is designed for easy layering, creating a straightforward path for building feedforward neural networks. We trained the ANN with labeled data that included features like gender, age, movie intensity, and movie length to determine whether individuals are likely to go to the movies. We monitored the model's training process, structure, and performance metrics, such as accuracy and loss, over several training cycles. This provided valuable insights into how well the model was learning and performing.

Model: "sequential_19"

Layer (type)	Output Shape	Param #
dense_56 (Dense)	(None, 16)	80
dense_57 (Dense)	(None, 16)	272
dense_58 (Dense)	(None, 1)	17

Total params: 369 (1.44 KB)
Trainable params: 369 (1.44 KB)
Non-trainable params: 0 (0.00 B)

```
Epoch 1/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 17ms/step - accuracy: 0.6674 - loss: 0.6750 - val_accuracy: 0.7692 - val_loss: 0.6482
Epoch 2/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6466 - loss: 0.6672 - val_accuracy: 0.7692 - val_loss: 0.6423
Epoch 3/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6442 - loss: 0.6614 - val_accuracy: 0.7692 - val_loss: 0.6359
Epoch 4/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6307 - loss: 0.6644 - val_accuracy: 0.7692 - val_loss: 0.6305
Epoch 5/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6953 - loss: 0.6431 - val_accuracy: 0.7692 - val_loss: 0.6249
Epoch 6/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6255 - loss: 0.6590 - val_accuracy: 0.7692 - val_loss: 0.6208
Epoch 7/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6234 - loss: 0.6567 - val_accuracy: 0.7692 - val_loss: 0.6156
Epoch 8/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6463 - loss: 0.6413 - val_accuracy: 0.7692 - val_loss: 0.6090
Epoch 9/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6244 - loss: 0.6434 - val_accuracy: 0.7692 - val_loss: 0.6033
Epoch 10/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6515 - loss: 0.6333 - val_accuracy: 0.7692 - val_loss: 0.5986
Epoch 11/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6526 - loss: 0.6336 - val_accuracy: 0.7692 - val_loss: 0.5948
Epoch 12/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 5ms/step - accuracy: 0.6567 - loss: 0.6276 - val_accuracy: 0.7692 - val_loss: 0.5898
Epoch 13/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.6203 - loss: 0.6274 - val_accuracy: 0.7692 - val_loss: 0.5848
Epoch 14/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6536 - loss: 0.6180 - val_accuracy: 0.7692 - val_loss: 0.5804
Epoch 15/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6494 - loss: 0.6210 - val_accuracy: 0.7692 - val_loss: 0.5773
Epoch 16/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6838 - loss: 0.5983 - val_accuracy: 0.7692 - val_loss: 0.5746
Epoch 17/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6765 - loss: 0.5941 - val_accuracy: 0.7692 - val_loss: 0.5727
Epoch 18/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6297 - loss: 0.6106 - val_accuracy: 0.7692 - val_loss: 0.5697
Epoch 19/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6338 - loss: 0.6160 - val_accuracy: 0.7692 - val_loss: 0.5663
Epoch 20/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6328 - loss: 0.6091 - val_accuracy: 0.7692 - val_loss: 0.5627
Epoch 21/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6349 - loss: 0.5979 - val_accuracy: 0.7692 - val_loss: 0.5589
Epoch 22/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6776 - loss: 0.5908 - val_accuracy: 0.7692 - val_loss: 0.5558
Epoch 23/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6338 - loss: 0.6033 - val_accuracy: 0.7692 - val_loss: 0.5527
Epoch 24/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6270 - loss: 0.5957 - val_accuracy: 0.7692 - val_loss: 0.5488
Epoch 25/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6619 - loss: 0.5888 - val_accuracy: 0.7692 - val_loss: 0.5450
Epoch 26/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.7123 - loss: 0.5701 - val_accuracy: 0.7692 - val_loss: 0.5410
Epoch 27/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.7141 - loss: 0.5689 - val_accuracy: 0.8077 - val_loss: 0.5375
Epoch 28/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6849 - loss: 0.5956 - val_accuracy: 0.8077 - val_loss: 0.5349
Epoch 29/50
```

```
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.7086 - loss: 0.5698 - val_accuracy: 0.7692 - val_loss: 0.5319
Epoch 30/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6724 - loss: 0.5998 - val_accuracy: 0.7692 - val_loss: 0.5295
Epoch 31/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6633 - loss: 0.5968 - val_accuracy: 0.7692 - val_loss: 0.5268
Epoch 32/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6953 - loss: 0.5729 - val_accuracy: 0.7692 - val_loss: 0.5245
Epoch 33/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6933 - loss: 0.5797 - val_accuracy: 0.8077 - val_loss: 0.5241
Epoch 34/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.7201 - loss: 0.5619 - val_accuracy: 0.8077 - val_loss: 0.5237
Epoch 35/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6618 - loss: 0.5800 - val_accuracy: 0.7692 - val_loss: 0.5236
Epoch 36/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6568 - loss: 0.5841 - val_accuracy: 0.7692 - val_loss: 0.5227
Epoch 37/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6683 - loss: 0.5767 - val_accuracy: 0.7692 - val_loss: 0.5220
Epoch 38/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6716 - loss: 0.5604 - val_accuracy: 0.7692 - val_loss: 0.5209
Epoch 39/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6893 - loss: 0.5584 - val_accuracy: 0.7692 - val_loss: 0.5195
Epoch 40/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6768 - loss: 0.5444 - val_accuracy: 0.7692 - val_loss: 0.5173
Epoch 41/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6904 - loss: 0.5408 - val_accuracy: 0.7692 - val_loss: 0.5167
Epoch 42/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6779 - loss: 0.5585 - val_accuracy: 0.7692 - val_loss: 0.5157
Epoch 43/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6474 - loss: 0.5706 - val_accuracy: 0.7692 - val_loss: 0.5138
Epoch 44/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6701 - loss: 0.5619 - val_accuracy: 0.7692 - val_loss: 0.5112
Epoch 45/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6990 - loss: 0.5364 - val_accuracy: 0.7692 - val_loss: 0.5097
Epoch 46/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6896 - loss: 0.5572 - val_accuracy: 0.7692 - val_loss: 0.5093
Epoch 47/50
Epoch 48/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.7282 - loss: 0.5339 - val_accuracy: 0.7692 - val_loss: 0.5075
Epoch 49/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.7042 - loss: 0.5507 - val_accuracy: 0.7692 - val_loss: 0.5060
Epoch 50/50
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.6928 - loss: 0.5365 - val_accuracy: 0.7692 - val_loss: 0.5042
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 1ms/step - accuracy: 0.6235 - loss: 0.5876
ANN Test Accuracy: 0.6071428656578064
```

The ANN model then undergoes evaluation to assess its performance. During this evaluation, we analyze the loss function and the range of X, which represents the number of epochs, and subsequently visualize this data.

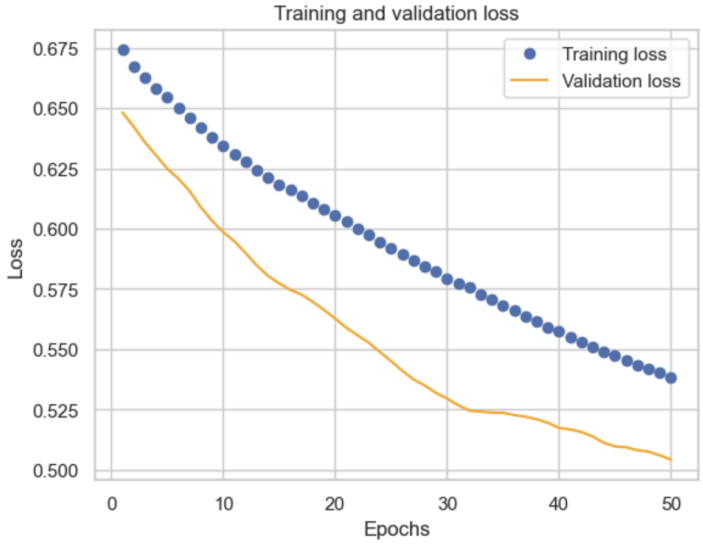


Figure 17: Training and Validation Loss

Following this, we test the accuracy of both the training and validation sets and generate plots to depict the model's behavior. We get the model shown below:

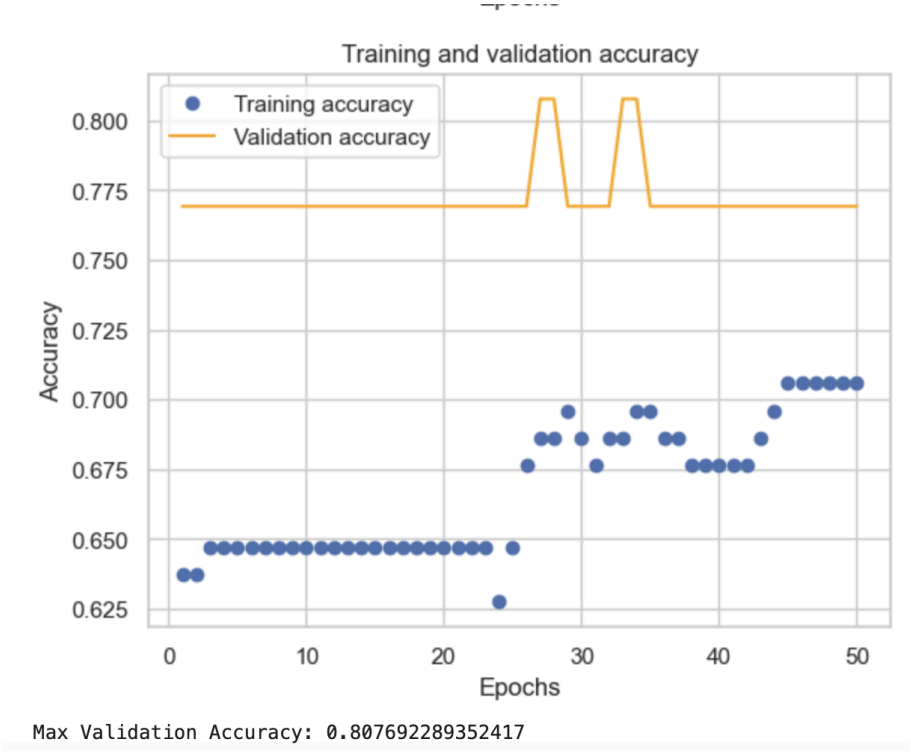


Figure 18: Training and Validation Accuracy

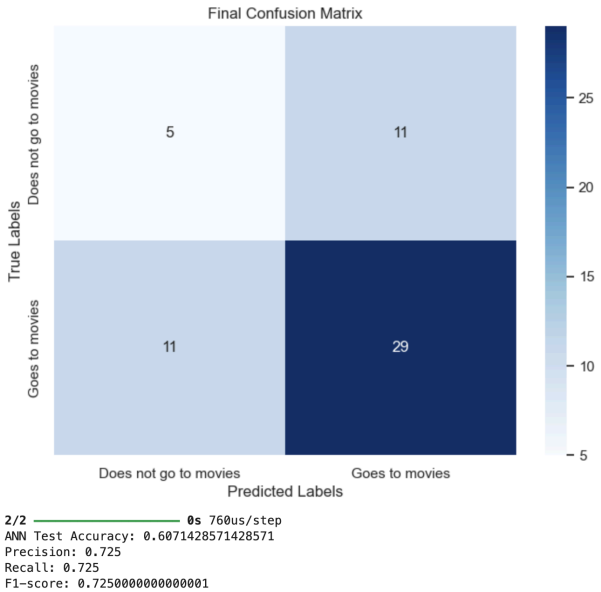


Figure 19: Confusion Matrix and Accuracy results of ANN model:

The metrics provided give a detailed overview of the ANN model's performance. These insights reveal the model's robustness and dependability in forecasting movie-going behavior. Each metric presents a distinct viewpoint, enabling stakeholders to evaluate various facets of the model's effectiveness and make well-informed decisions about its implementation.

III. Logistic Regression

We selected logistic regression for our analysis due to its effectiveness in handling binary classification tasks, such as determining whether individuals are likely to go to the movies based on various attributes. This method models the probability of membership in a specific class, yielding results that are easy to interpret in terms of probability. Moreover, logistic regression is computationally efficient, making it well-suited for our dataset size and complexity. Using logistic regression, we aim to understand the connections between our features and the primary outcome of movie-going, enabling precise predictions and deeper insights into what drives people to attend movies.

We have effectively implemented logistic regression in our dataset to explore how individual characteristics influence their probability of going to the movies. We trained the logistic regression model after preparing and dividing the data into training and testing sets. Evaluating the model with metrics such as accuracy, precision, recall, and F1-score has provided critical insights into its predictive strength. Furthermore, using a confusion matrix to visualize the model's performance clarifies its predictive abilities, enhancing our understanding of the factors contributing to movie-going behavior.

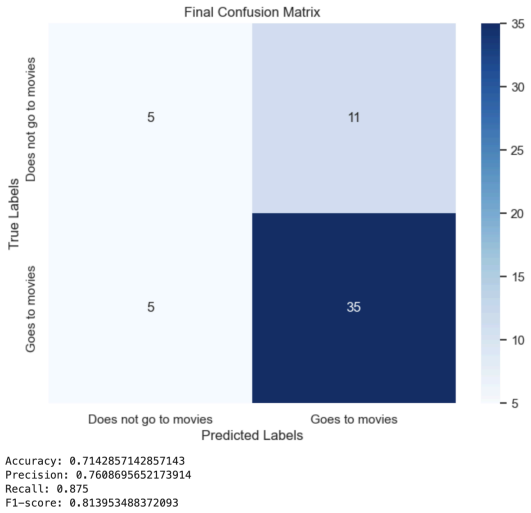
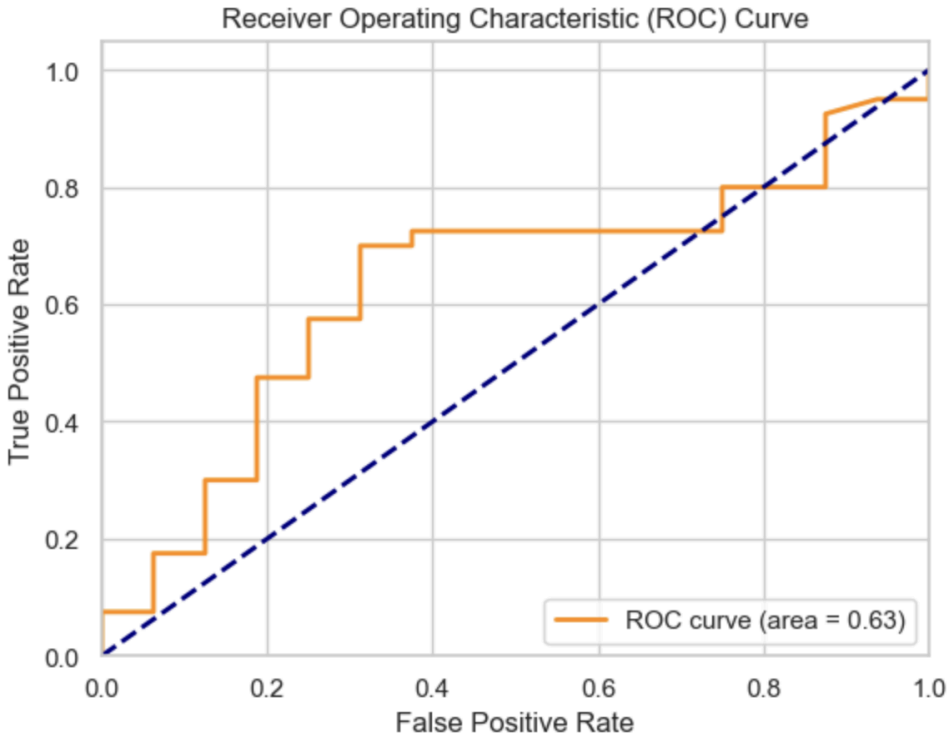


Figure 20: Results of Logistic Regression

To confirm the accuracy of our logistic regression results, we used two key evaluation metrics: the ROC curve and the Precision-Recall curve. These metrics help us understand the model's effectiveness beyond just essential accuracy. The ROC curve compares the actual positive rate to the false positive rate at different thresholds, showing how well the model distinguishes between positive and negative outcomes. On the other hand, the Precision-Recall curve measures the trade-off between precision and recall, which is particularly useful in situations where the classes are imbalanced. We calculated the Area

Under the Curve (AUC) for both curves, giving us a clear numerical value of the model's ability to classify correctly and balance between precision and recall. Both graphs are included below to illustrate these findings.



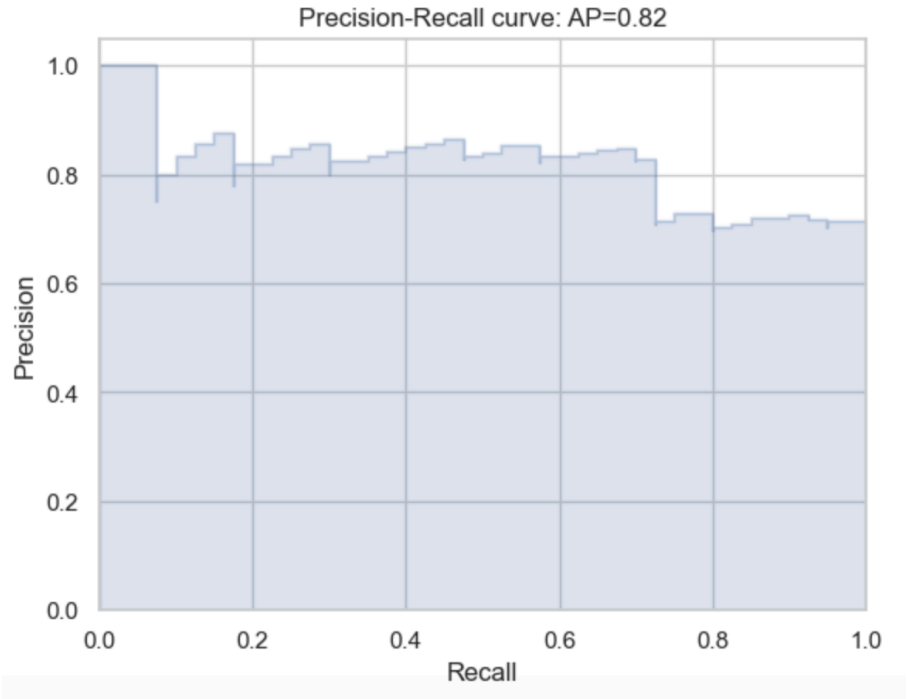


Figure 21 and 22: ROC Curve and Precision-Recall Curve

AUC Values:

```
ROC AUC: 0.63359375
Precision-Recall AUC: 0.8181325270228341
```

IV. Correlation Analysis

1. *Correlation Matrix Table*: A correlation matrix summarizes the relationships between multiple variables within a dataset. It offers insights into how variables change together, indicating both the direction and strength of their linear relationships. In Python, creating a correlation matrix utilizes the pandas and seaborn libraries. Below is our correlation matrix table:

	gender	age	goes_to_movies	movie_intensity	movie_length_hr
gender	1.000000	-0.184831	0.172033	-0.140247	-0.133147
age	-0.184831	1.000000	-0.155281	0.179146	0.262870
goes_to_movies	0.172033	-0.155281	1.000000	0.003361	0.184407
movie_intensity	-0.140247	0.179146	0.003361	1.000000	0.048293
movie_length_hr	-0.133147	0.262870	0.184407	0.048293	1.000000

2. *Heatmap:* The heatmap illustrates the correlation between different variables in our dataset, specifically focusing on gender, age, movie attendance, movie intensity preference, and preferred movie length. We used the data from our correlation matrix table to make our heatmap. Each cell in the heatmap displays the correlation coefficient between two variables, ranging from -1 to 1. A correlation coefficient close to 1 suggests a strong positive correlation, indicating that as one variable increases, the other also tends to increase. Conversely, a correlation coefficient near -1 indicates a strong negative correlation, implying that as one variable increases, the other decreases.

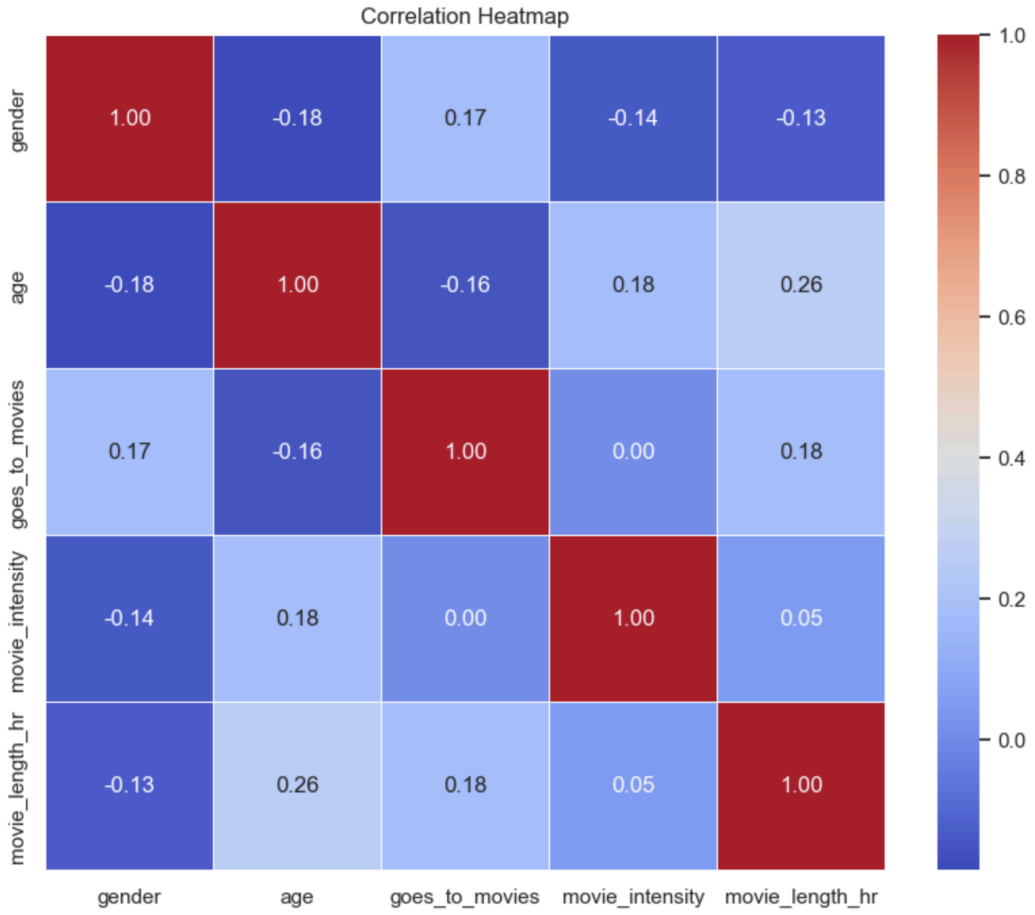


Figure 23: Correlation Heatmap

3. *Pair Plots:* The pair plot visualization was created using the Seaborn library in Python. By utilizing Seaborn's pair plot function, we generated a grid of scatterplots for all pairs of variables in our dataset, along with histograms along the diagonal. This allowed us to visualize the relationships between different variables and explore potential correlations.

   In the context of our movie dataset, the pair plot offers a comprehensive view of the relationships between different variables. For example, we can examine how age relates to movie intensity or movie length preference or how gender influences movie-going behavior. Each scatterplot in the pair plot provides insight into the potential correlations between these variables. Additionally, the graphs along the diagonal give information on the distribution of each variable individually, helping us understand the data's overall characteristics. Below are our pair plots:

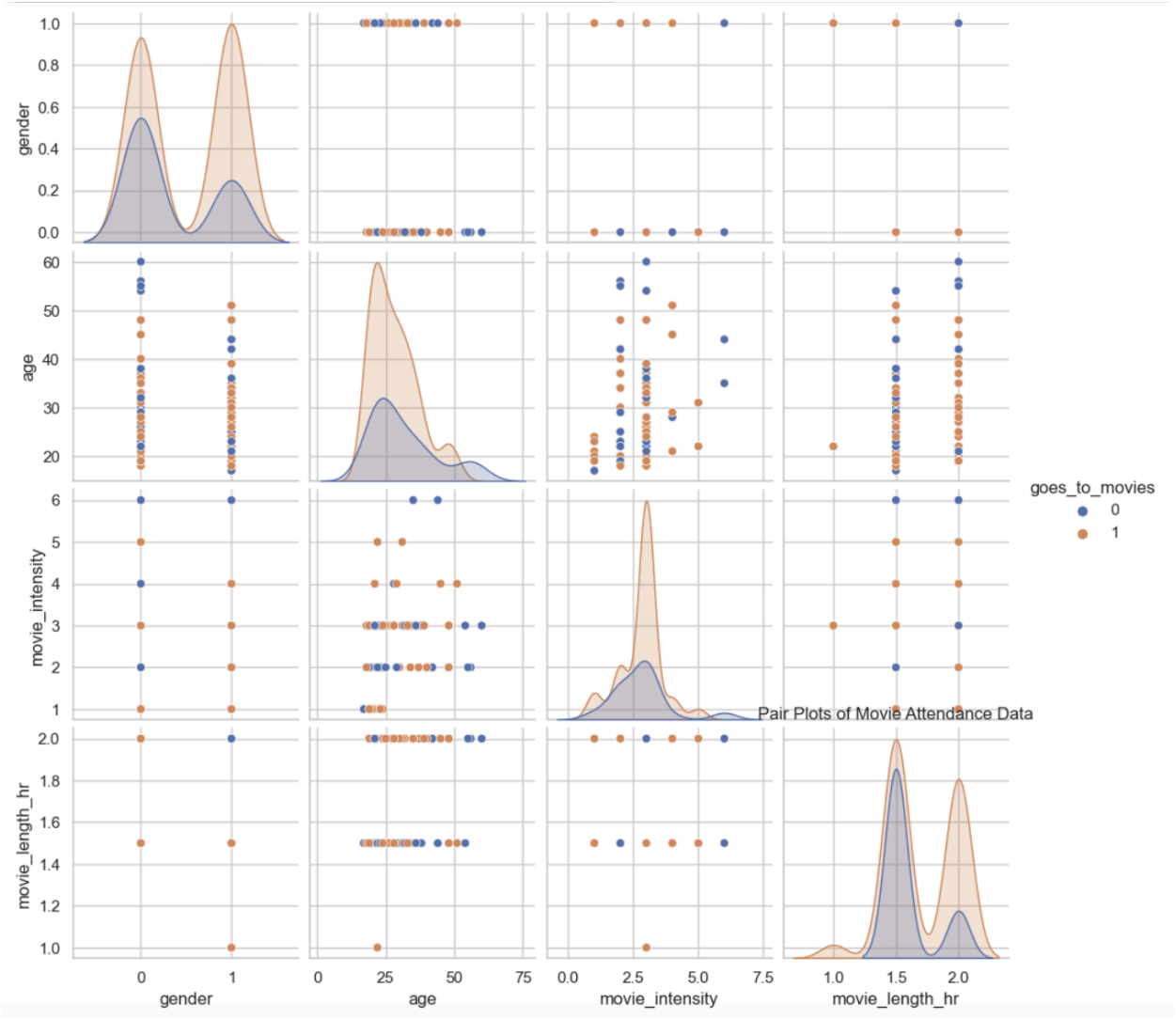
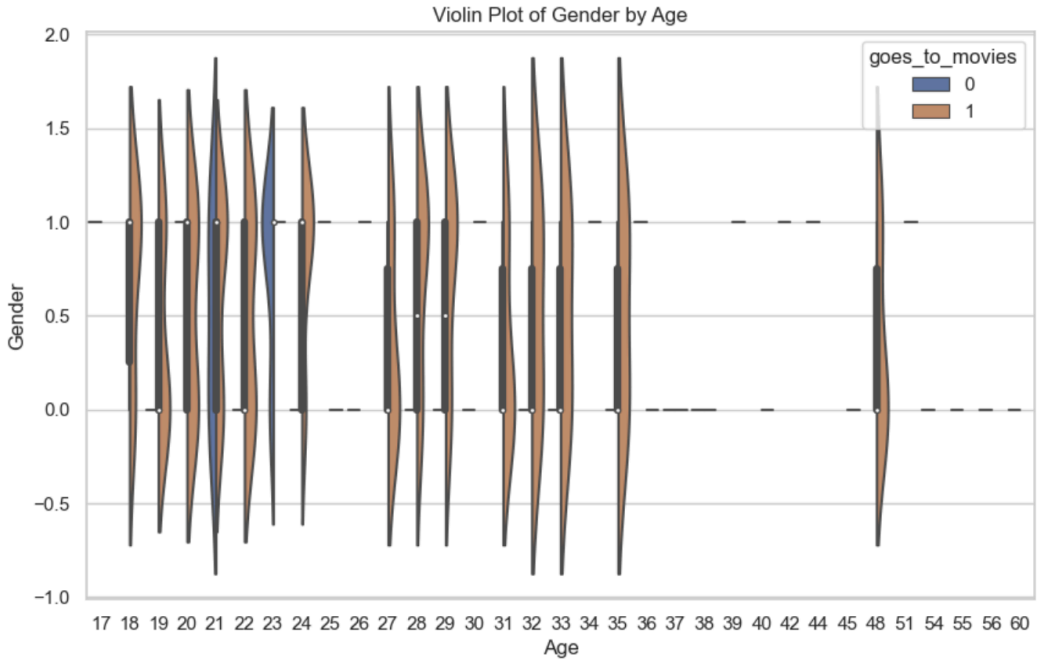
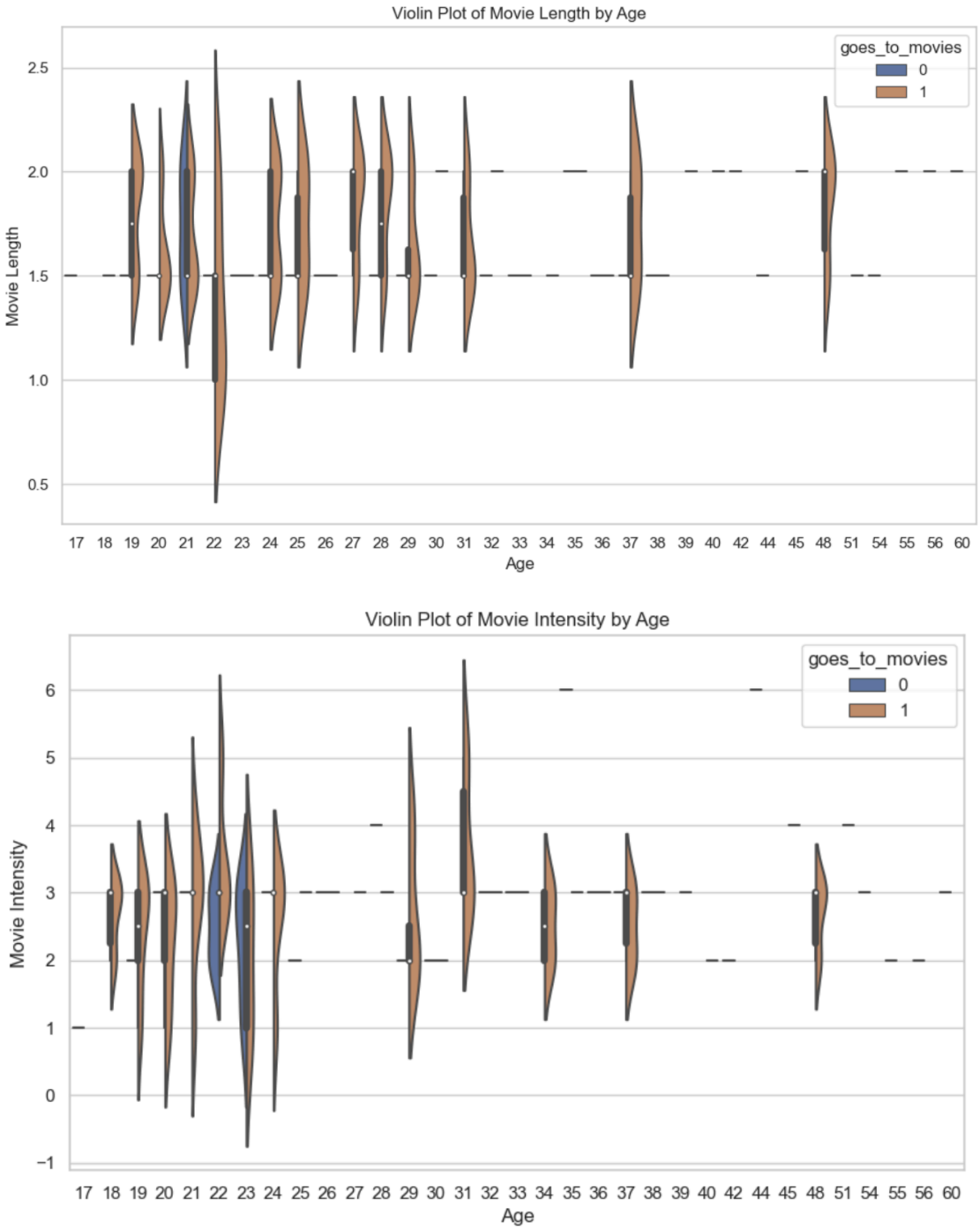


Figure 24: Pair Plots

4. *Violin Plots:* We utilized the matplotlib and seaborn libraries in Python to create violin plots to visually explore the distribution of various factors across different age groups and their association with movie-going behavior. These plots provide a comprehensive view of the data distribution and allow for easy comparison between age groups. Violin plots are particularly useful for displaying the distribution of continuous variables, such as movie intensity and movie length, across categorical variables like age and gender. By splitting the violins based on movie-going behavior, we can observe how preferences vary across age groups. Additionally, violin plots can handle multimodal data and help identify outliers, providing insights into the underlying patterns within the dataset.


Violin Plot of Movie Intensity by Age
goes_to_movies
0
1
Movie Intensity
6
5
4
3
2
1
0
−1
17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 42 44 45 48 51 54 55 56 60
Age


Violin Plot of Gender by Age
goes_to_movies
0
1
Gender
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 42 44 45 48 51 54 55 56 60
Age



Figures 25- 28: Violin Plots

5. *Data Density Plot:* We generated a data density plot for all variables in our dataset, separating age from the rest of the variables. We utilized the Seaborn library to create kernel density estimate plots for each variable. The KDE plot represents the probability density function of each variable's distribution. By visualizing the density of each variable's values, we can gain insights into their distributions and how they relate. The legend indicates which curve corresponds to each variable, allowing for straightforward plot interpretation.

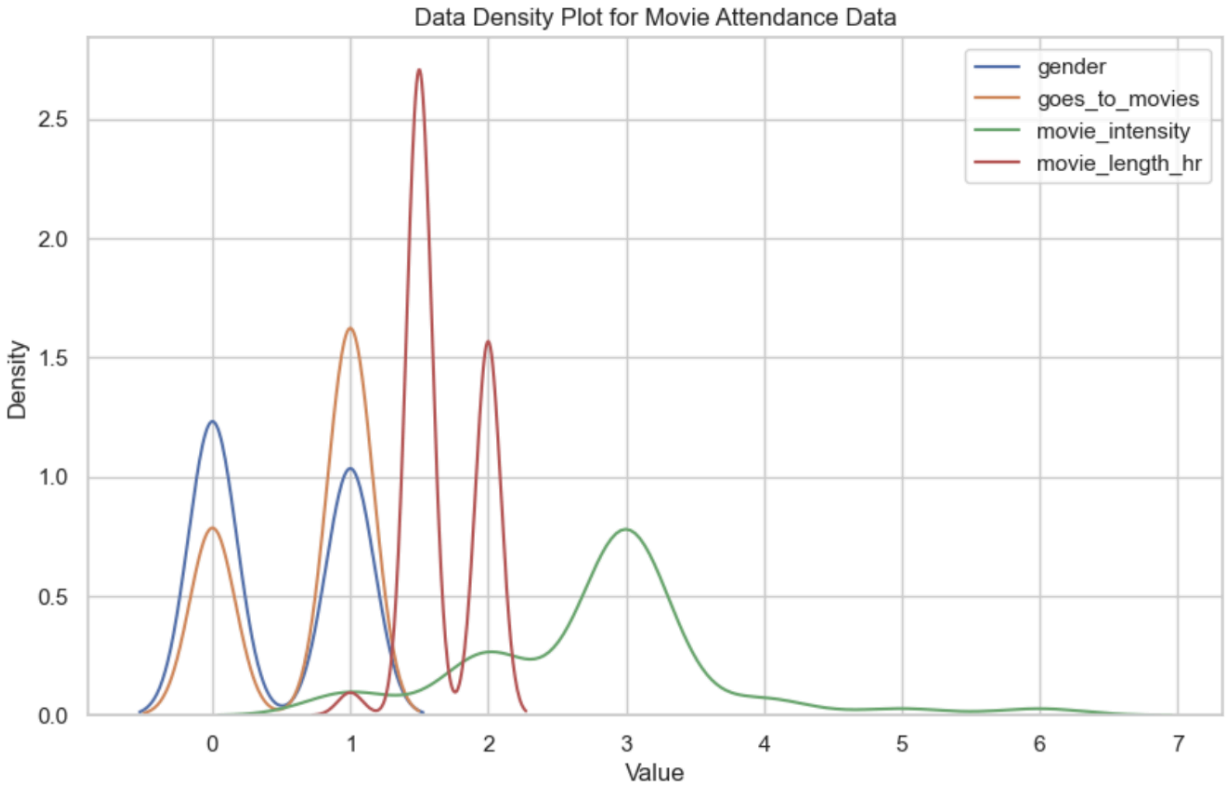


Figure 29: Data Density Plot For Movie Attendance Data

Splitting the variables from age in our density plot serves multiple purposes. Primarily, it ensures visualization clarity by accommodating the wide range of age values and preventing distortion in the plots of other variables. Additionally, it allows for a more focused analysis of age-specific patterns and their influence on the distribution of each variable. By separating age, we can scale each plot appropriately, facilitating comparative analysis and providing insights into how different variables vary across age groups.

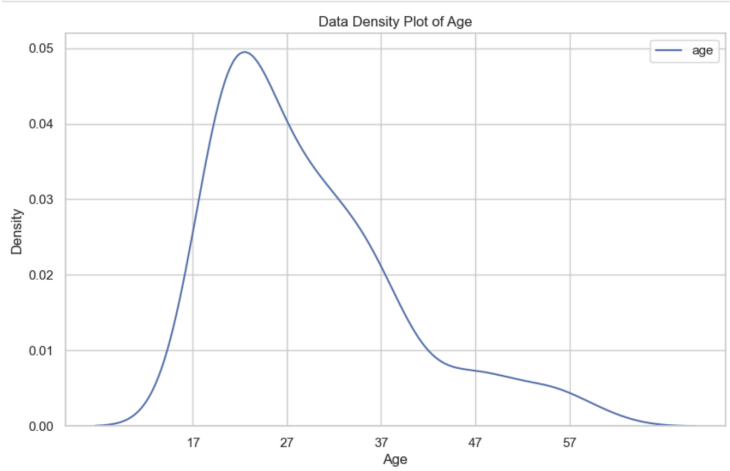


Figure 30: Data Density Plot For Age

Discussion

Our study was designed to predict movie attendance based on demographic and preference factors. This analysis is crucial for optimizing marketing strategies and enhancing profitability in the film industry. We employed various predictive models, each offering unique insights into the dataset's characteristics and the effectiveness of different predictive techniques.

We started with simple descriptive statistics. Using Google Forms and online datasets, we measured statistics for all variables used throughout the testing process. Through analysis and chi-square tests, we confirmed our hypothesis that there is a correlation between age, preferred movie genre, and movie-watching experience while also learning there is little correlation between age and preferred movie length.

Delving deeper into our hypothesis, we created a decision tree to find the best demographic to cater to. The decision tree model demonstrated a practical capacity for classifying individuals based on their likelihood of going to the movies. With a training mean squared error (MSE) of 0.0703125, the model showed high accuracy in fitting the training data, suggesting that it could effectively generalize the patterns in these data. However, a slightly higher test MSE of 0.23214285714285715 indicated some challenges in generalizing these predictions to unseen data, though the error margin was still reasonable. Notably, the decision tree's node 6 emerged as particularly influential, perfectly classifying a subset of data. Node 6 was a perfect sample with 100% precision and could predict our model.

The ANN model, structured with three dense layers, was employed to predict movie attendance through a more complex, layered learning approach. This model showed a progressive improvement over 50 training epochs, yet it faced limitations in validation accuracy, peaking at around 60.7%. This suggests that while the ANN could capture intricate patterns, its generalization outside the training set was limited. The metrics such as precision, recall, and the F1-score showed a balanced performance but highlighted the need for further optimization to enhance predictive accuracy.

Logistic regression outperformed other models with a higher accuracy of 71.4%, precision of 76.1%, and recall of 87.5%. ROC AUC and Precision-Recall AUC scores further supported this model's effectiveness in binary classification, which indicated good discriminative ability and firm performance in precision and recall, respectively. The logistic regression model proved to be a robust method for our binary classification problem, utilizing the relationship between features like age, gender, preferred movie intensity, and movie length to predict attendance. The success of this model indicates that the variables selected for the study, including demographic and preference factors, correlate with movie attendance.

Across all models, logistic regression consistently provided the best performance, suggesting that simpler models may be more effective and interpretable for this binary classification problem. This insight is precious for marketing strategies as it allows for clear communication of the factors that most predict movie attendance, which can be leveraged in targeted advertising and promotional activities.

The correlation analysis, including heatmap and pair plots, revealed significant relationships between variables such as age and movie preference, which are critical for understanding consumer behavior in the movie industry. These visual tools helped confirm the importance of these variables in predicting movie attendance, supporting the findings from the predictive models.

Lastly, our overall analysis demonstrates the value of predictive modeling in understanding and influencing movie-goer behavior. We gained comprehensive insights into the factors driving movie attendance by integrating decision trees, ANN, and logistic regression models. These findings enhance our understanding of consumer behavior and offer insights for the film industry to refine their marketing strategies and potentially increase profitability.

Conclusion

Throughout our study, we aimed to predict movie attendance using a range of factors, including age, movie preferences, and viewing habits, and employing machine-learning techniques like logistic regression and decision trees. These methods helped us uncover clear patterns in our data, revealing that younger audiences, especially those aged 18-24, are likelier to attend movies. This demographic prefers going to the cinema and favors longer movie formats, contrary to our initial assumptions about their preferences for shorter movies. Additionally, we found a significant trend towards streaming among this age group, which is crucial for the industry to consider in their distribution strategies. The study on the relationship between age and preferred movie genre/intensity also gave us helpful insight. While we initially hypothesized that younger people preferred high-intensity genres like horror, we can see now that the younger demographic prefers mildly intense genres like action and sci-fi.

Our analysis highlighted the effectiveness of logistic regression in accurately predicting movie-going behaviors, outperforming other models such as ANN. This reinforced the importance of choosing suitable modeling techniques that align with the data characteristics and research objectives. Moreover, the decision trees and logistic models we developed provided insights essential for tailoring marketing strategies targeting young audiences to enhance cinema attendance.

However, our study faced limitations due to the small and potentially biased sample size derived mainly from self-reported data. Despite these challenges, the diverse data sources from online datasets and analytical methods we employed strengthened our findings.

Given these insights, we recommend that the movie industry focus its marketing efforts on younger demographics and consider producing longer films that cater to their preferences, available both in theaters and through streaming platforms. Focusing on mildly intense genres for films widens the demographic, as each age group enjoys genres such as action or drama. For future research, expanding the demographic scope to include more varied groups and exploring additional factors like socio-economic status and geographic location could provide deeper insights into movie-going habits. These efforts can help the industry better understand and engage its audience, optimizing content creation and distribution strategies to meet evolving viewer preferences.

References

Cooley, Chris, Kodra, Angelo, Jabro, Mary, Macabebe, Reyna, Romaya, Donovan, Russo, Emma, Topalli, Kleant. “Moviegoers Survey”, Google Forms, Google, 18 March 2024,

https://docs.google.com/spreadsheets/d/17HMGXKOnKcxZNN2xMP9-bBMLbsMedfB8LDTxI6ti5BM/edit?usp=sharing

“Preferred Movie Length in the U.S. by Age 2018.” *Statista*, 5 Jan. 2023,

www.statista.com/statistics/860072/preferred-movie-length-age/.Accessed 28 Apr. 2024.

“Preferred Place of Movie Consumption in the U.S. 2018.” *Statista*, 5 Jan. 2023,

www.statista.com/statistics/264399/preferred-place-of-movie-consumption-in-the-us/. Accessed 28 Apr. 2024.

Stoll, Julia. “Movie Streaming Frequency U.S. by Age 2021.” *Statista*, 12 Jan. 2023,

www.statista.com/statistics/935493/movies-watching-streaming-frequency-us-by-age/. Accessed 28 Apr. 2024.

“U.S.: Moviegoing Frequency by Generation 2022.” *Statista*, 28 June 2023,

www.statista.com/statistics/538259/frequency-going-to-the-movies-age-usa/. Accessed 28 Apr. 2024.

“U.S.: Moviegoing Frequency by Generation 2022.” *Statista*, 28 June 2023,

www.statista.com/statistics/538259/frequency-going-to-the-movies-age-usa/. Accessed 28 Apr. 2024.

Cooley, Chris, Kodra, Angelo, Jabro, Mary. “Code Appendix”, Python.

Code Appendix