Final Project

By: Kate Ryan and Alexis Arthur

Introduction

We will be utilizing data from IMDb, *The Internet Movie Database*, to analyze the Top 250 movies rated by regular IMDb voters. Specifically, we will include the movie name, runtime, year, score, nominations, and winnings of each movie from IMDb. We will then dive into how many Oscars, *Academy of Motion Picture Arts and Sciences*, each movie has won to see how successful they were. In terms of what else we are trying to figure out, we want to see if movies with longer run times are more popular than movies with shorter run times or vice versa and if older films, before the year 2000, are more popular than newer films, after the year 2000. The year of the earliest movie made on this list is 1921 and the year of the most recent movie made on this list is 2024. So, movies on this list were made throughout those 103 years, which is something to think about in terms of popularity of films. Some people know a lot about most of these films, but we want to answer how the different analytics like runtime, year, and score contribute to the success of these films. We do not want to state what is already seen in the data. We will build upon that and describe the significance of the runtime and year in terms of the ranking it received in IMDb's Top 250 movies.

You may be reading this and wonder, why did they choose to analyze IMDb's Top 250 movies and Oscar's data? We both enjoy watching movies whether that is romantic comedies, horror movies, or dramas and we wanted to find significant data for a topic that we enjoy researching. Choosing to research movies has made this project fun and entertaining for both of us. Also, IMDb's Top 250 movies are updated frequently, with the most recent movie coming out in 2024. Since we have already looked at Oscars data in Data Wrangling, we were able to jog our memory and use some of those ideas and insights in our project. It is important to note that the top 250 list is voted on by regular IMDb users, so many of these people are avid movie watchers like us. As we answer our research questions, we get more clarity as to why all of these movies are so popular. Most of these movies have some overlap and are in the IMDb's Top 250 movies and Oscar's movie data.

Data

The “IMDb Top 250 Movies” is located under the *Movies* tab on the IMDb website, and lists every movie along with the year, runtime (in hours and minutes), rating (PG-13, R, etc.), and the score, in terms of stars and the number of people who have rated that specific movie. To get simple information about the movie, you can click on the circle with the “i” inside located on the far right of the movie listing. As you scroll down the list, you will see some information regarding the rankings. This reads, “The Top-Rated Movie list only includes feature films. Shorts, TV movies, and documentaries are not included. The list is ranked by a formula which includes the number of ratings each movie received from users, and value of ratings from regular users. To be included on the list, a movie must receive ratings from at least 25000 users” (IMDb Top 250 Movies, n.d.). The IMDb link that we used is, https://www.imdb.com/chart/top/?ref_=nv_mv_250. To learn more about how list ranking is determined, you can click on the link above and all the information described is on the website.

“The Oscar Award, 1927 – 2025” is a data set that we found on Kaggle and contains columns for the year the film was made, year the ceremony took place, number of the ceremony, category name, name of the nominee, title of the film, and finally, a column indicating if the nominee won or lost. This final column had the word “True” if a nominee won the award and the word “False” if they had lost. While looking at this data set as well as the Oscars website and comparing it to the IMDb data set, the earliest film year is 1927/1928 while the earliest film year from IMDb is 1921. Although there are only a few films between 1921 and 1927/1928, this may cause some concern in our data, so we will have to figure out if we want to keep those films or remove them. The Kaggle data set is a reliable source because its data is scraped from the Oscars Award Database (Fontes & Lu, 2025). The Kaggle link that we used is, https://www.kaggle.com/datasets/unanimad/the-oscar-award.

Data Dictionary

Field	Type	Description
Movie_Name	Text	Movie Title
Winner	Boolean	Did this movie win an Oscar or not (True or False)?
Run_Time_Minutes	Numeric	How long (in minutes) is the movie?
Year	Numeric	What year did the movie air?
Nominations	Numeric	How many nominations did the movie have?

Score	Numeric	What does IMDb rate the movie?
Winners	Numeric	How many times has the movie won an Oscar?

Research Questions

We have developed five research questions to help us analyze the relationship between the two data sets. The research questions are as follows:

1. How many Oscars has each movie won and how many nominations has each movie received? Are movies with Oscars wins more popular than movies without wins?

This question helps analyze the relationship between public acclaim and institutional recognition. Year after year, many films beloved by audiences get passed over for awards like the Oscars and vice versa. Our research aims to unpack this problem and see if there is a disconnect between what audiences prefer and what the Academy rewards.

2. Are movies with longer run times more popular than movies with shorter run times or vice versa?

This research question explores evolving viewer preferences and storytelling criteria. If shorter movies appear to be more popular, it could be a testament to how viewer's attention spans have shifted over time, and they want storytelling to be straight to the point. However, if longer movies appear to be more popular, this could suggest that viewers appreciate a complex and meaningful plotline. We hope that with this analysis, we can help current and future filmmakers cater to the audience's needs and provide them with insight on how they should structure their storytelling.

3. Are older movies (before the year 2000) more popular than newer movies (after the year 2000)?

This research question explores the cultural impact and legacy that movies leave on their audiences. This analysis could significantly help filmmakers and production companies create successful movies. If older movies are preferred, this suggests that the audience appreciates nostalgia, and certain themes will always find appeal amongst audiences for years and years to

come. Whereas if newer movies are more popular, this could suggest that audience's preferences shift as the culture and world around them shifts.

4. What is the average number of Nominations across all movies? Do more movies fall above or below the average?

Our fourth research question aims to analyze the average number of nominations across the top 250 movies. We want to see if more movies in the top 250 fall above the average or if more movies fall at/below the average. This can give great insight into if movies with more nominations are more popular than movies that have less nominations and looks at the relationship between audience approval and critical acclaim.

5. What is the lowest and highest polarity score based on the movie name? Also, what is the sentiment over time for the year column?

Finally, our last research question goes along with our sentiment analysis. We wanted to know which films had the lowest polarity, lower sentiment and more negative view, compared to the films with the highest polarity, higher sentiment and a more positive view. We did that based on movie name, so the movies with a more negative connotation had a lower polarity score and the movies with a more positive connotation had a higher polarity score. Our graph depicts the sentiment over time based on the year. The chart will be shown in our analysis section below.

Horizontal Integration/Data Scraping

To gather the necessary information from the two data sets and to combine them together, a multitude of procedures was performed. To gather information from the first data set, "IMDb Top 250 Movies", we decided to scrape it. We wrote code in a Jupyter Notebook that utilized the website's URL and scraped for the name of the movie, the year it was made, the movie's run time, and the score the movie received. Once this data was collected, we put it in a DataFrame called "movies_df" which has 250 columns and 4 rows. We then saved this DataFrame to an excel file titled "IMDB_Top_250.xlsx."

For our second data set, "The Oscar Award, 1927 – 2025", we downloaded the data as a CSV file from Kaggle and loaded it into the same Jupyter Notebook. We decided to save it as an Excel file titled 'the_oscar_award.xlsx' and then loaded this into a DataFrame titled

“oscars_data_excel.” The original DataFrame had 11,110 rows and 8 columns but we performed some cleaning procedures and removed the rows that had null values in the ‘film’ column, which then brought the DataFrame down to 10,751 rows. As mentioned before, this data set has a column containing True or False values if a nominee won or lost the award they were nominated for. We wanted to use this column, called ‘winner’, and the ‘film’ column to create two new columns called ‘Nominations’ and ‘Winners’ that would contain numeric values. We wrote code to group by the ‘film’ column and create these two new columns. Finally, we dropped the rows that we deemed unnecessary to leave the final DataFrame with 10,751 rows and 4 columns and then saved this to an Excel file titled ‘cleaned_oscars_data.xlsx.’

Finally, we horizontally integrated the two DataFrames together in the Jupyter Notebook. We checked the two DataFrames to make sure everything looked correct and analyzed factors such as the shape, date types, and null values for each. We also took care of the columns and renamed ‘film’ to ‘Movie_Name’ and ‘winner’ to ‘Winner’ in the “oscars_data_excel” DataFrame. We wrote code to check the columns and then added the missing columns ‘Winners’, ‘Winner’, and ‘Nominations’ to the “movies_df” DataFrame. We then wrote code to horizontally integrate the two DataFrames, drop rows from “oscars_data_excel” that aren’t in “movies_df”, drop duplicates, and then save the integrated DataFrames into a CSV file titled ‘Horizontal_Integration_Films.csv.’ Below is a picture of what the final integrated DataFrame looks like.

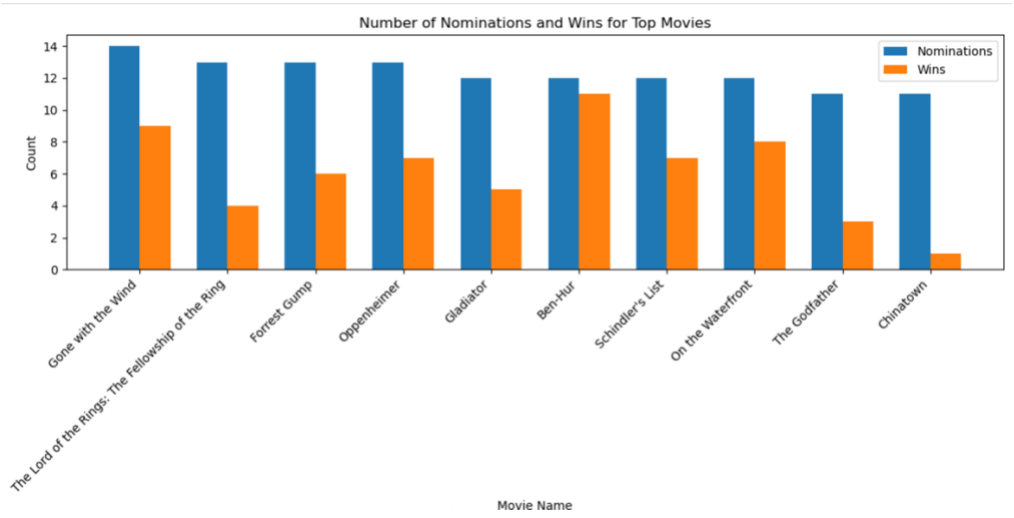
	Movie_Name	Year	Run_Time_Minutes	Score	Winner	Nominations	Winners
0	The Shawshank Redemption	1994	142	9.3	False	7	0
1	The Godfather	1972	175	9.2	True	11	3
2	The Dark Knight	2008	152	9.0	True	8	2
3	The Godfather Part II	1974	202	9.0	False	11	6
4	12 Angry Men	1957	96	9.0	False	3	0
...	...	...	...	...	...	...	...
169	The Grapes of Wrath	1940	129	8.1	False	7	2
170	Into the Wild	2007	148	8.0	False	2	0
171	The Help	2011	146	8.1	False	4	1
172	Amores Perros	2000	154	8.0	False	1	0
173	Rebecca	1940	130	8.1	False	11	2

174 rows × 7 columns

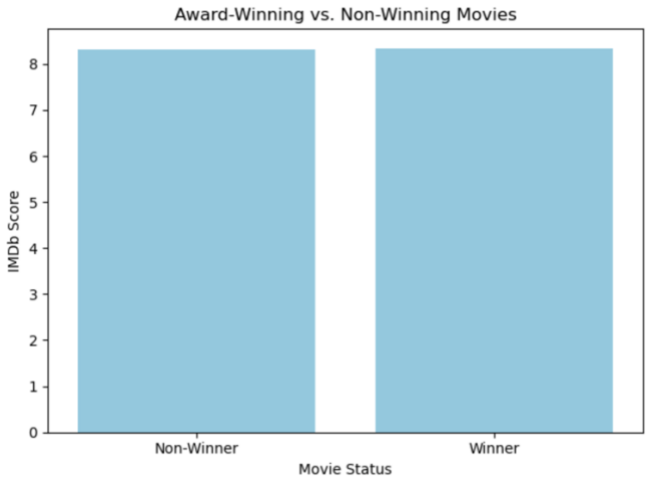
Analysis – Building On Our ‘Research Questions’

Hypothesis Test and Question 1: How many Oscars has each movie won and how many nominations has each movie received – and are movies with Oscars wins more popular than movies without wins?

For our hypothesis test and to answer our first question, we decided to do a two-sample t-test to see if movies with Oscars wins are more popular than movies without wins. To start, we created a chart showing how many wins and nominations each movie has received to provide some context before performing the test. As you can see, *Gone with the Wind* has 14 nominations, which is the highest number of nominations, and 9 Oscar wins.

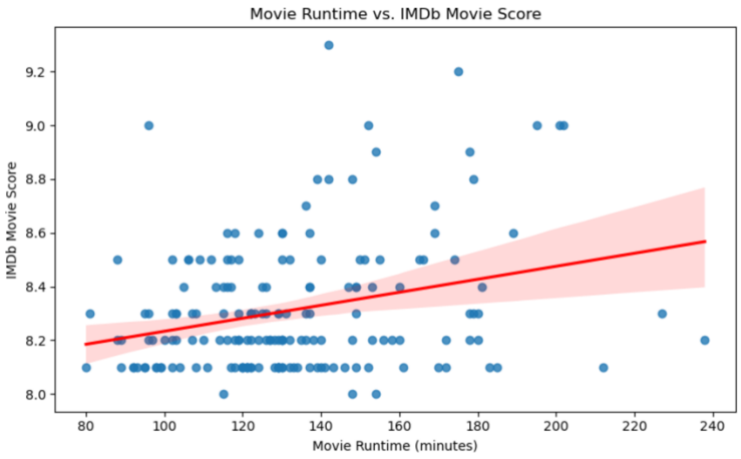


After we created that graph, we performed the two-sample t-test. We got a t-statistic of 1.0580 and a p-value of 0.2925 (alpha = 0.05, so 0.2925 > 0.05), which shows that the result is not statistically significant, and that award-winning films and non-award-winning films have a similar performance. This also means that the IMDb scores are not that different between award-winning films and non-award-winning films. We created a second bar chart that shows the minimal difference between these two categories.



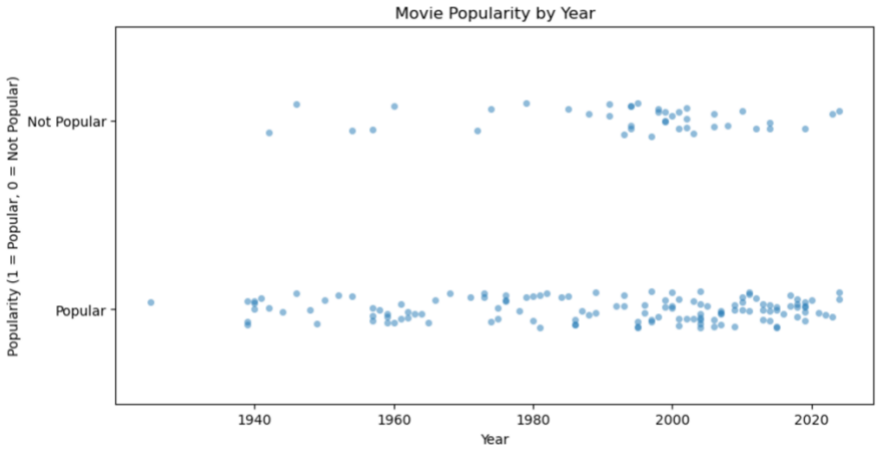
Bivariate and Question 2: Are movies with longer run times more popular than movies with shorter run times or vice versa?

We ran a bivariate test to go along with this analysis question. A bivariate test looks at two variables and, in this test, we are looking at runtime (in minutes) and the IMDb movie score. We created a scatter plot to look at the relationship between these two variables and found that it does show a weak positive relationship between movie runtime and IMDb score, but more precise calculations are needed to determine if this trend is accurate. We decided to calculate the correlation coefficient and came up with these criteria; if the correlation coefficient is closer to +1 then longer movies will score higher, if the correlation coefficient is closer to -1 then longer movies tend to score lower, and if the correlation coefficient is closer to 0 then there is no clear relationship. The correlation coefficient between movie runtime and IMDb score is 0.28 and confirms what the scatter plot illustrates; there is no strong or meaningful relationship between runtime and score. This means that longer movies are not necessarily more popular than shorter movies and vice versa and runtime alone cannot accurately predict higher IMDb ratings.

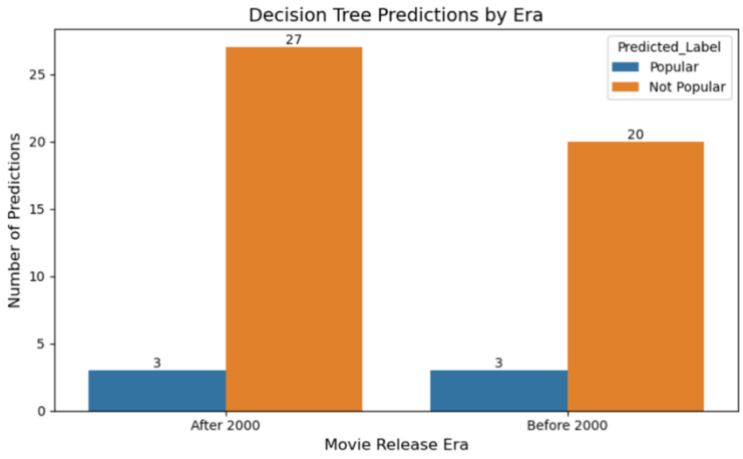


Machine Learning and Question 3: Are older movies (before the year 2000) more popular than newer movies (after the year 2000)?

To properly analyze this question, we decided to run multiple machine learning tests including logistic regression, decision tree, and two KNN tests. We first created a label to establish if a movie is popular or not popular. Movies with a score greater than or equal to 8.5 were considered “popular” and movies with a score less than 8.5 were considered “not popular.” We performed the logistic regression test and got an accuracy score of 0.47, which means our model is 47% accurate at testing if newer or older movies are more popular. 47% is not good as we would prefer to have a score above 60% and as close to 100% as possible. From this test, we can conclude that year alone is not a good predictor of popularity. We then tested for the model coefficient and F1 score to make sure our assumption was accurate. We came up with these criteria for coefficient; a positive coefficient means newer movies are more popular because as year increases, the probability of being "popular" goes up and a negative coefficient means older movies are more popular because as year increases, the probability of being "popular" goes down. We got a coefficient of approximately 0.0083 which means that, according to year, newer movies are slightly more popular. However, this coefficient is very small and suggests a weak relationship between year and score. Finally, our F1 score turned out to be 0.46 which means the model has moderate ability to identify which movies are more popular. However, this F1 score is still extremely low which means the year itself may not be an accurate predictor of popularity. All of this is pictured in the scatter plot below.

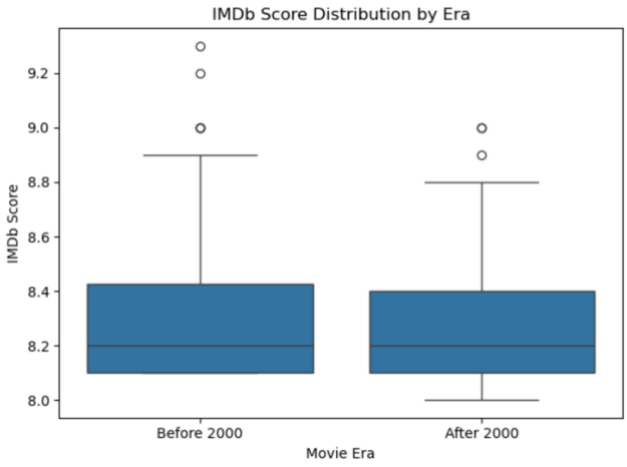


Next, we ran a decision tree test, and our accuracy came out to be 0.70. This is significantly higher than the 46% from the previous logistic regression test and is above 60%, so it could indicate that year plays a role in determining movie popularity. Before we jump to conclusions, we need to look at the F1 score which came out to be 0.20. This is a very low F1 score, which tells us that our model got some predictions correct but could be predicting a lot of "popular" movies incorrectly, again indicating that year alone isn't a good predictor. The chart below illustrates what number of movies the test found to be "popular" and "not popular" and as you can see, it predicted a lot more movies to be "not popular." We decided to run two more tests to make our final determinations.



Finally, we decided to run a KNN test to really see if year alone can predict the popularity of movies. Our accuracy score is 75%, which is now our highest out of the three tests and does indicate that there is some pattern between year and popularity. To be sure, we check the F1

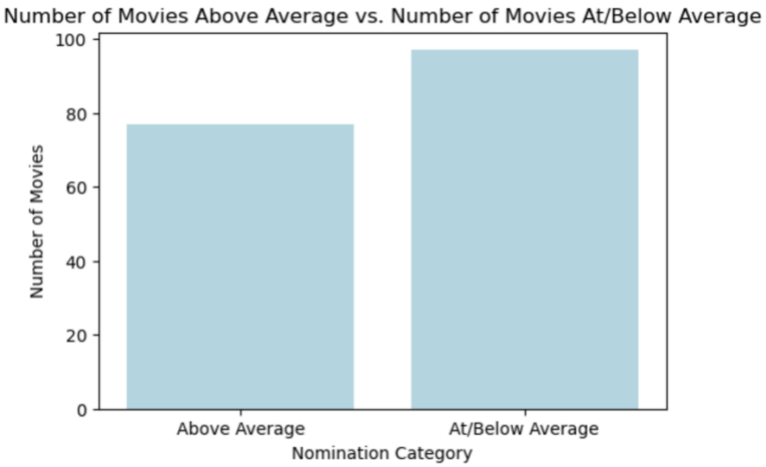
score which is 0.38, a low score but the highest one yet. This illustrates that our model could be imbalanced so we "balanced" the data and ran the KNN test again. The accuracy score from this test is 70%, which is lower than the first one but expected considering we balanced the data. This is a decent accuracy score, but it's the F1 score that will tell us the real story. We got 0.47, an increase from the previous F1 score, but still relatively low. This tells us that the model is fairer and better at identifying "popular" movies but because the score is still low, we can still conclude that year alone is not the most reliable predictor of popularity. We wrap up our KNN tests with a boxplot that shows similar medians, IQR's, whiskers, and outliers between the two groups. Because these plots are so similar, we are able to firmly conclude that there is no significant difference in IMDb score between each category, which means that year alone is not a substantial predictor of movie popularity.



Univariate and Question 4: What is the average number of Nominations across all movies? Do more movies fall above or below the average?

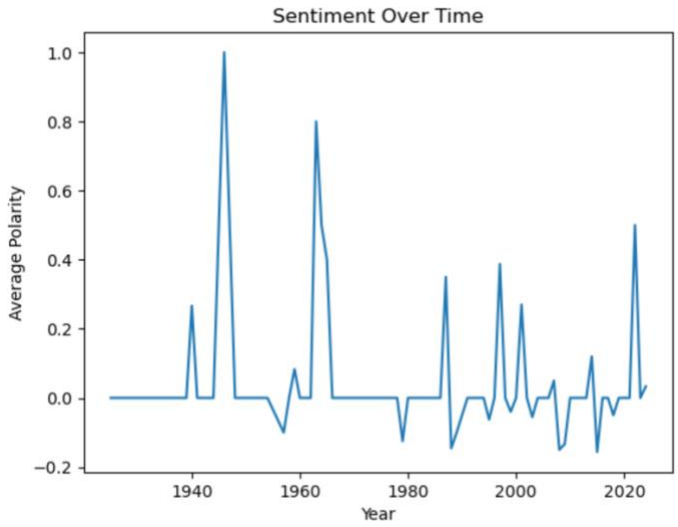
We chose to combine our univariate analysis and our question 4. Univariate means that we are only focusing on one variable, and we chose to focus on the nominations. We found that the average number of nominations across all movies is 5.33. We created a column that categorizes nominations based on the average we found. The movie would be considered 'above average' if the number of nominations was greater than 5.33 and 'at/below average' if the number of nominations was less than 5.33. We plotted the bar graph below and found the exact number of movies for 'above average' and 'at/below average', and those numbers are 77 and 97, respectively. Looking at the chart and additional calculations, we found that more movies are

at/below the average of 5.33 nominations compared to movies above the average of 5.33 nominations. There is a difference of 20 movies between the two categories.



Sentiment Analysis and Question 5: What is the lowest and highest polarity score based on the movie name? Also, what is the sentiment over time for the year column?

For text analytics, we decided to do sentiment analysis. First, we looked at the polarity and subjectivity of the "Movie_Name" column. For the sake of this question, we will focus on just polarity. The lowest polarity score is –0.625 for *Mad Max: Fury Road 2015* and the highest polarity score is 1.0 for *It's a Wonderful Life* and *The Best Years of Our Lives.* The polarity score of –0.625 means that it has a negative sentiment, which means that it has a negative connotation. On the other hand, the polarity score of 1.0 means that it has a positive sentiment, which means that it has a more positive connotation. We found that the average polarity for our integrated_films DataFrame is about 0.036, which proves that most of the films had a positive sentiment. Also, we looked at the sentiment over time for the year column and we got the chart below. It seems that around 1945 the average polarity spiked to 1.0, which is the highest polarity we found. The line did dip below 0 and into the negatives near the end of the 1950s, beginning of the 1980s, around the 1990s and 2010s.



Conclusion

We conclude that since all these movies are a part of IMDb's Top 250 movie list, they all have high scores and ratings, and that non-winning films and winning films have a similar performance in terms of IMDb scores. We also found that the average number of nominations was 5.33 for films, which is high, but it makes sense because of the popularity of all these films. We did notice that some of the movies changed as we kept running our code. It seemed like every day the list was updated since the IMDb voters cast their votes whenever. This did create some limitations because our numbers kept changing, but we were able to get the most up to date data from the website. In the future, we would suggest that groups pick a website to scrape that will not update every single day. It just gets confusing because your data may change, which could completely change your analysis for the project. We did enjoy learning more about popular movies and we suggest that groups pick a topic that fascinates or interests them, so they are involved in the project and fully understand what is going on. If we had more time, we probably would have devoted more time to the sentiment analysis because that was the one visualization that we were not sure what to do with. We couldn't do a word cloud because we don't have repeating data, and it wouldn't serve as a purpose for our specific project.

Works Cited

Awards Databases. (n.d.). Oscars.org. Retrieved April 2, 2025, from https://www.oscars.org/oscars/awards-databases

Fontes, R., & Lu, D. (2025, March 9). *The Oscar Award, 1927 - 2025*. Kaggle.com. Retrieved April 2, 2025, from https://www.kaggle.com/datasets/unanimad/the-oscar-award

IMDb Top 250 Movies. (n.d.). IMDb. Retrieved April 2, 2025, from https://www.imdb.com/chart/top/?ref_=nv_mv_250