IMDb and Oscars Analysis

By: Kate Ryan and Alexis Arthur

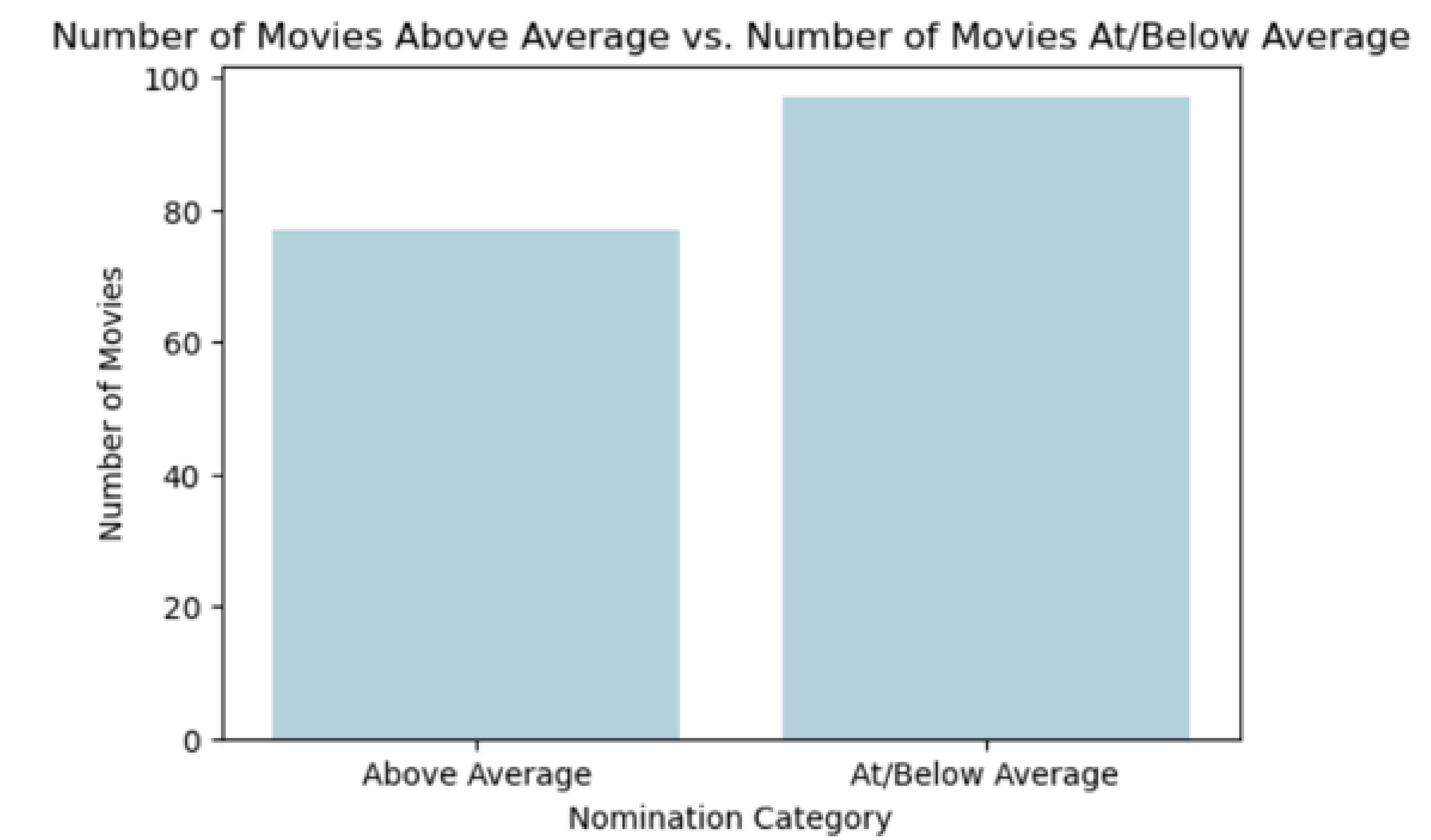
Initial Analysis

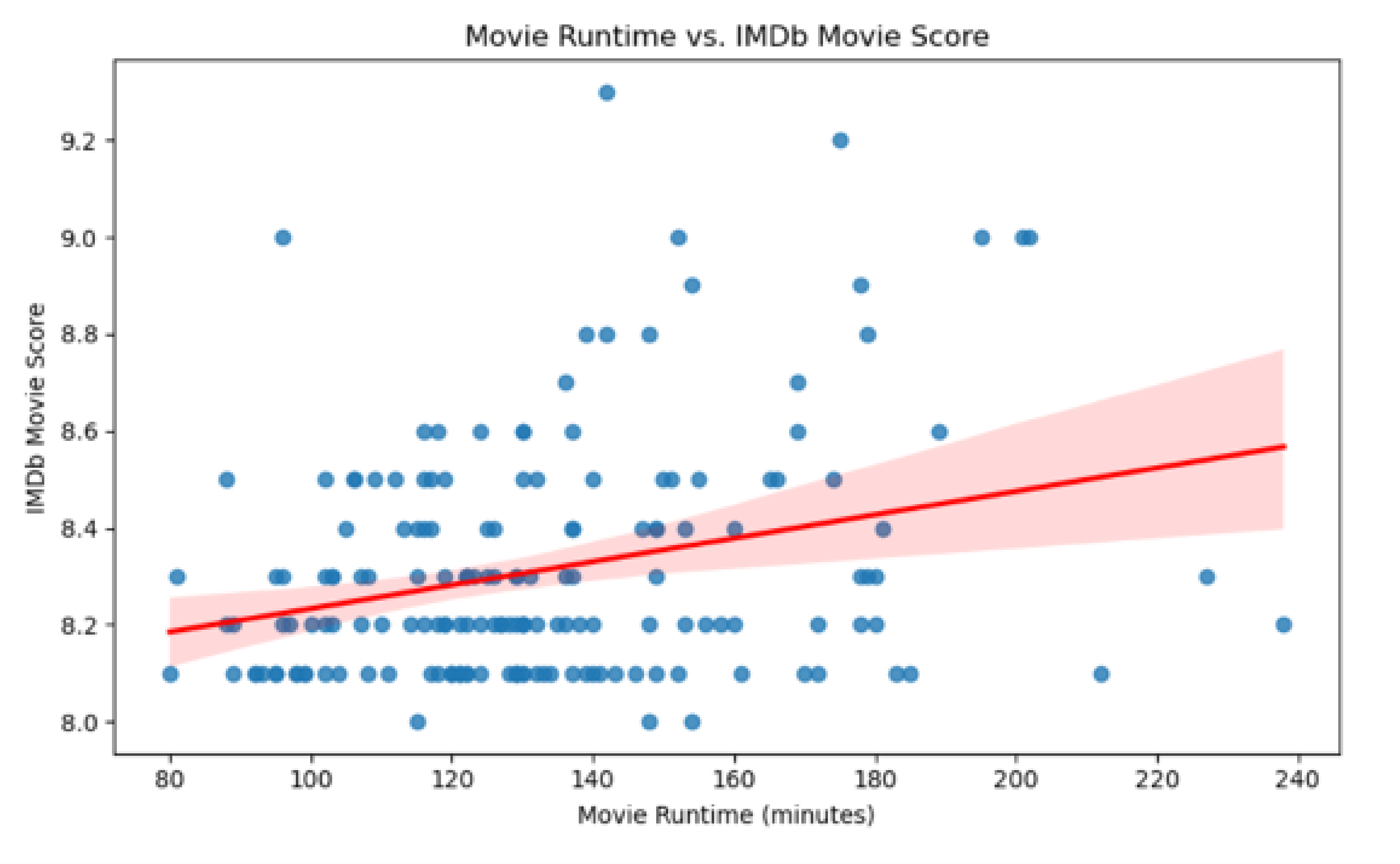
01 Horizontal Integration of the two datasets, 'IMDB_Top_250.xlsx' and 'the_oscar_award.xlsx'

02 Univariate, Bivariate, Hypothesis Test, Visualizations, Machine Learning, and Sentiment Analysis

UNIVARIATE

- We are going to look at the ‘Nominations’ column
- Question: What is the average number of nominations across all the movies? Do more movies fall above or at/below the average?
- We created a bar chart to show the number of movies above average vs. the number of movies at/below the average
- The average number of nominations across all movies is 5.33
- 97 movies are at/below the average
- 77 movies are above average



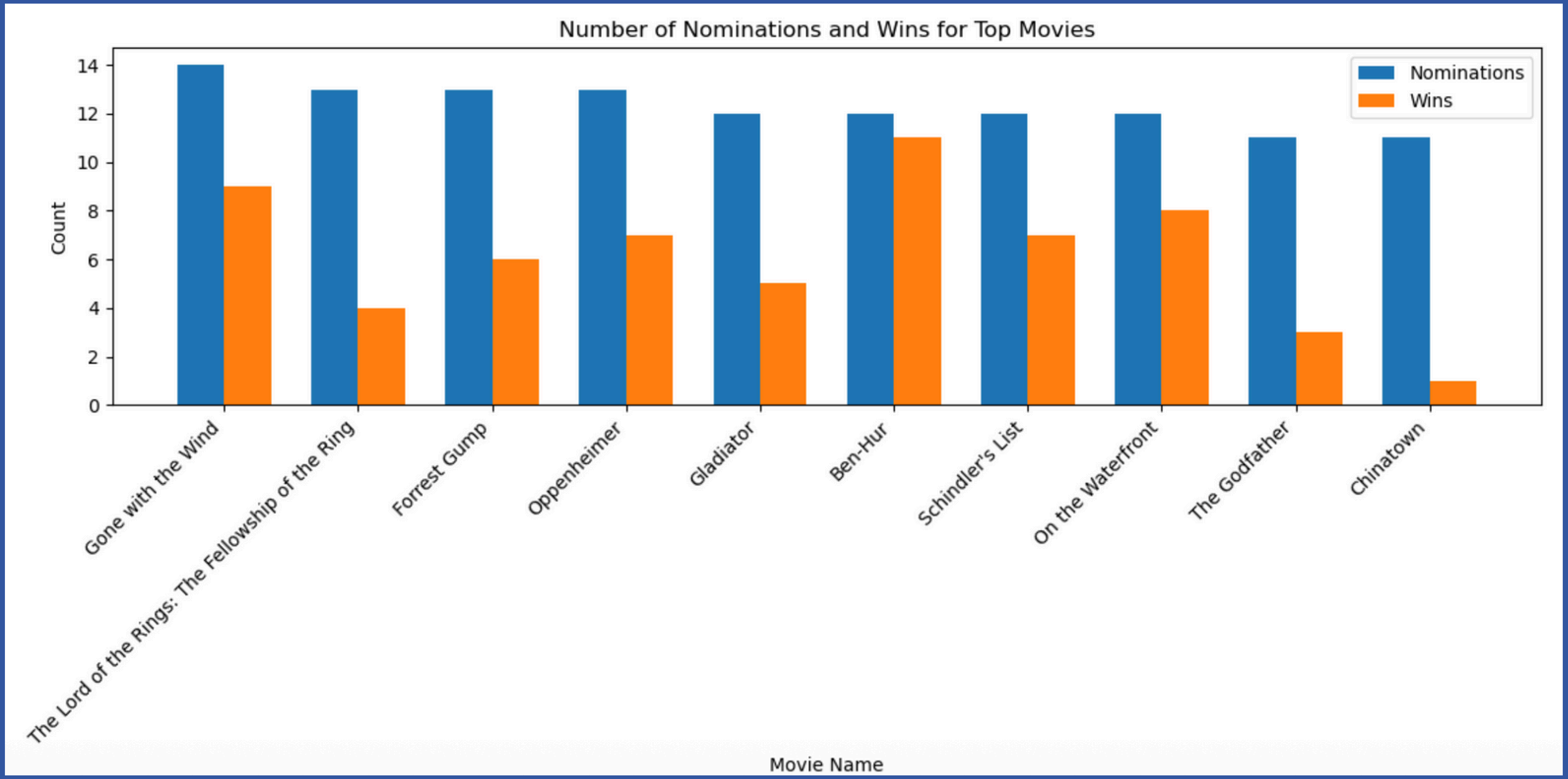


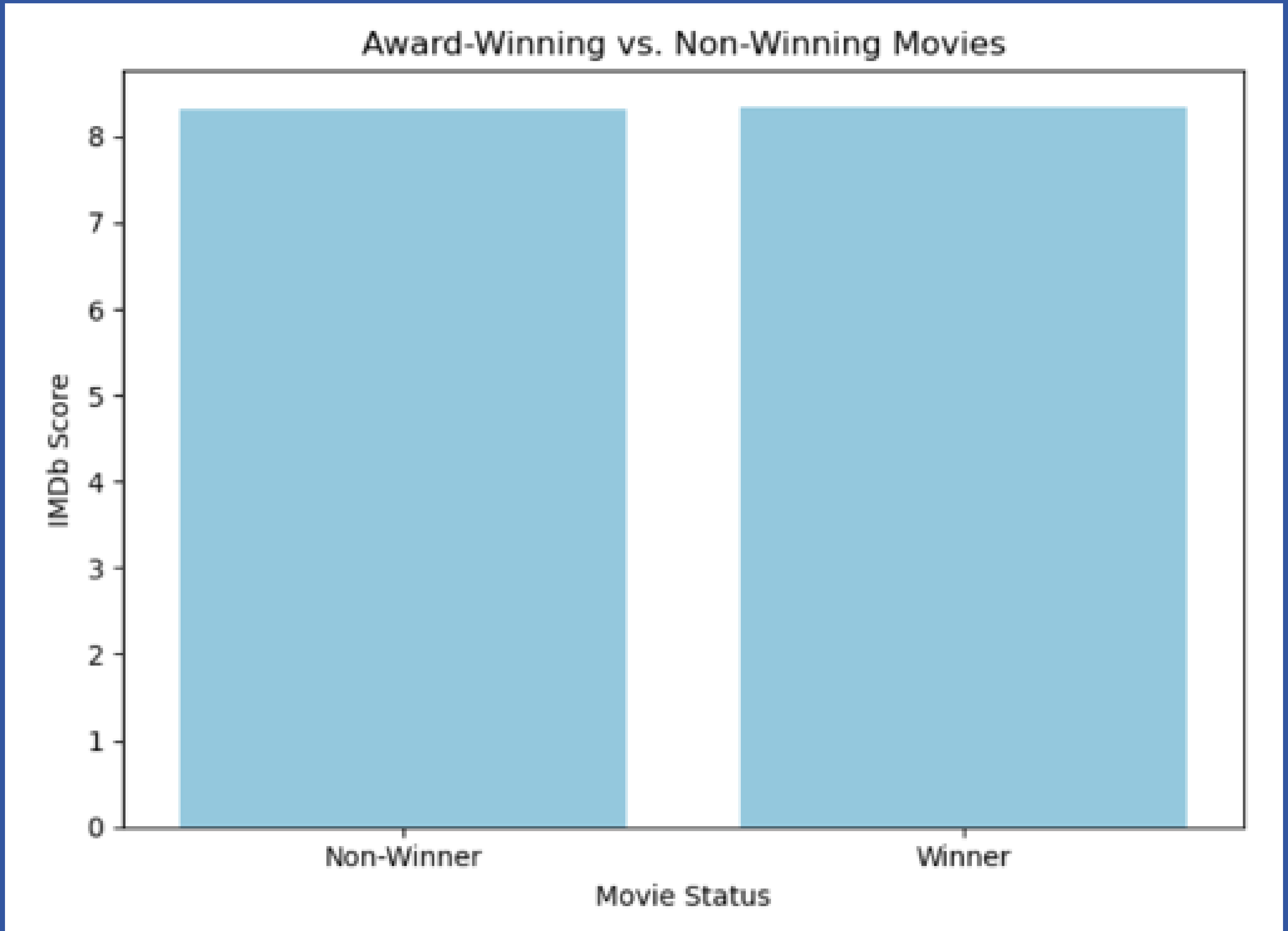
BIVARIATE

- Question: Are movies with longer run times more popular than movies with shorter run times or vice versa?
- We will analyze the ‘Run_TIme_Minutes’ and ‘Score’ columns
- The scatterplot shows that there is not a weak positive relationship, so we will have to do further analysis on this question
- The correlation coefficient is 0.28, which shows that there is not a strong correlation

HYPOTHESIS TEST: T-TEST

- First we made a chart to show how many wins and nominations each movie has
- The top 10 are shown in the chart below



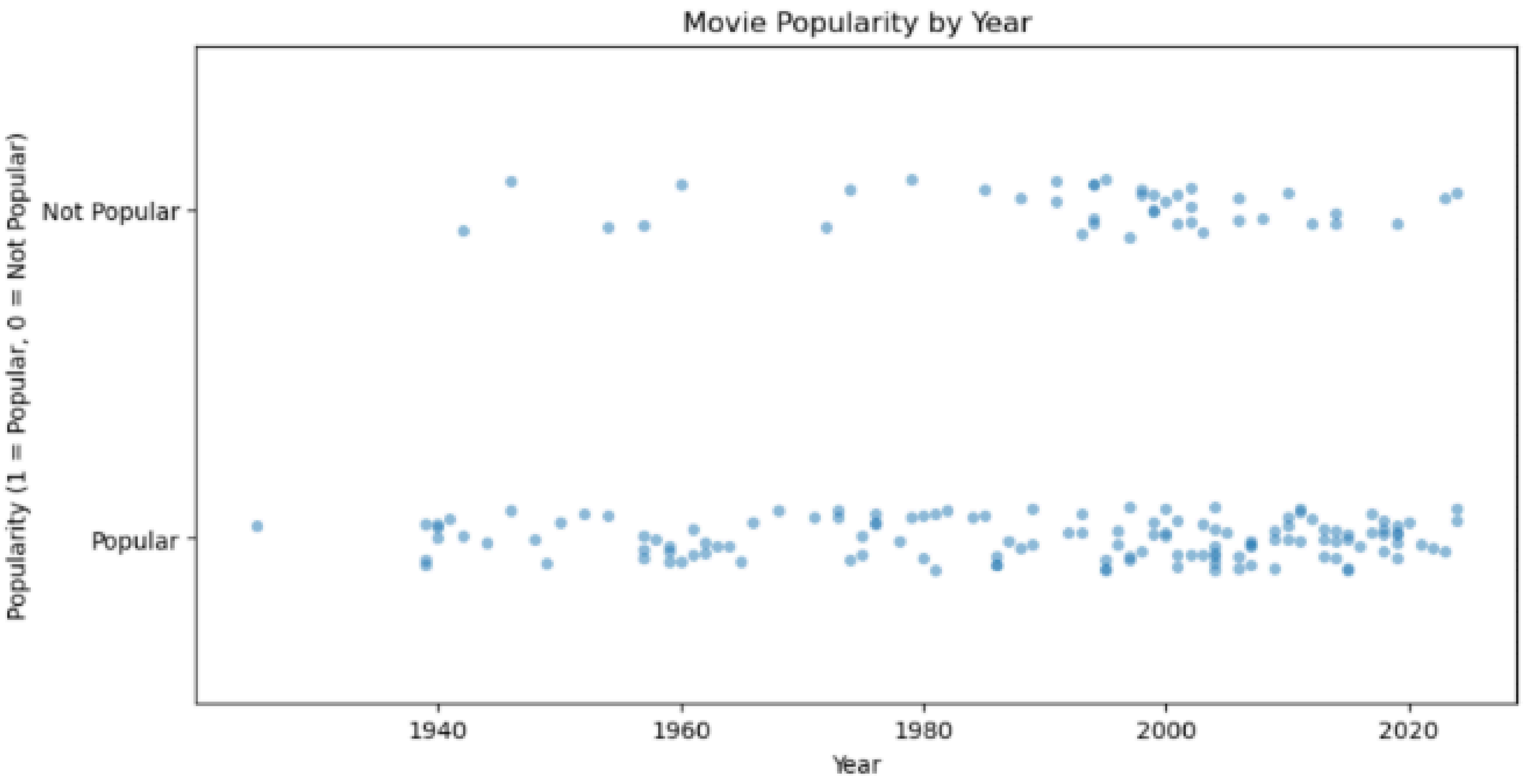


HYPOTHESIS TEST: T-TEST

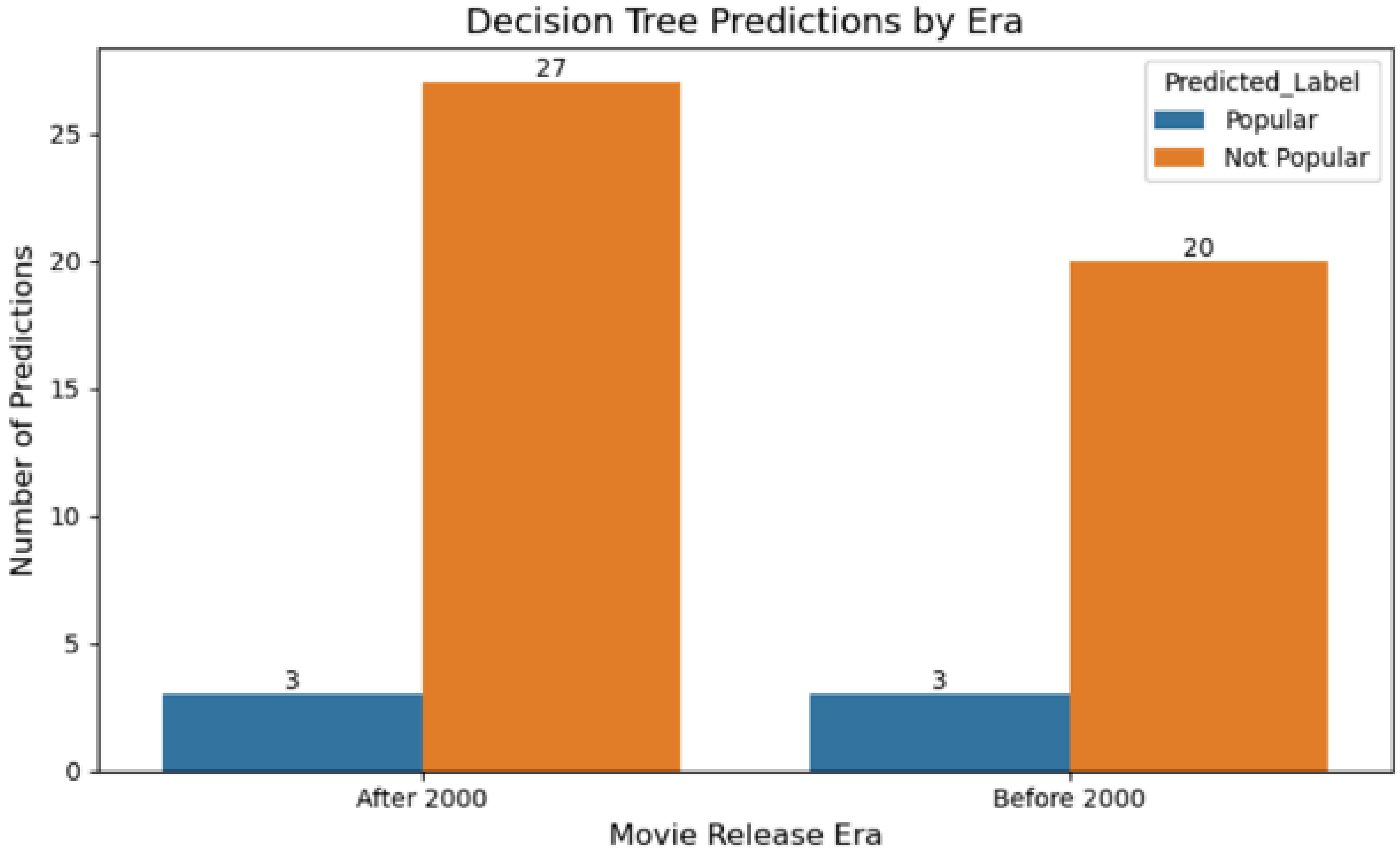
- The t-statistic is 1.0580
- The p-value is 0.2925
- The result is not statistically significant
- Award-winning movies and non-winners have a similar performance in terms of their IMDb score

MACHINE LEARNING: LOGISTIC REGRESSION

- Model accuracy is 0.47 or 47%
- Model coefficient for Year is approximately 0.0083
- The F1 Score is 0.46
- Looking at the accuracy/F1 score, our model suggests Year alone is not an accurate predictor, so we need to run more tests to come to a conclusion



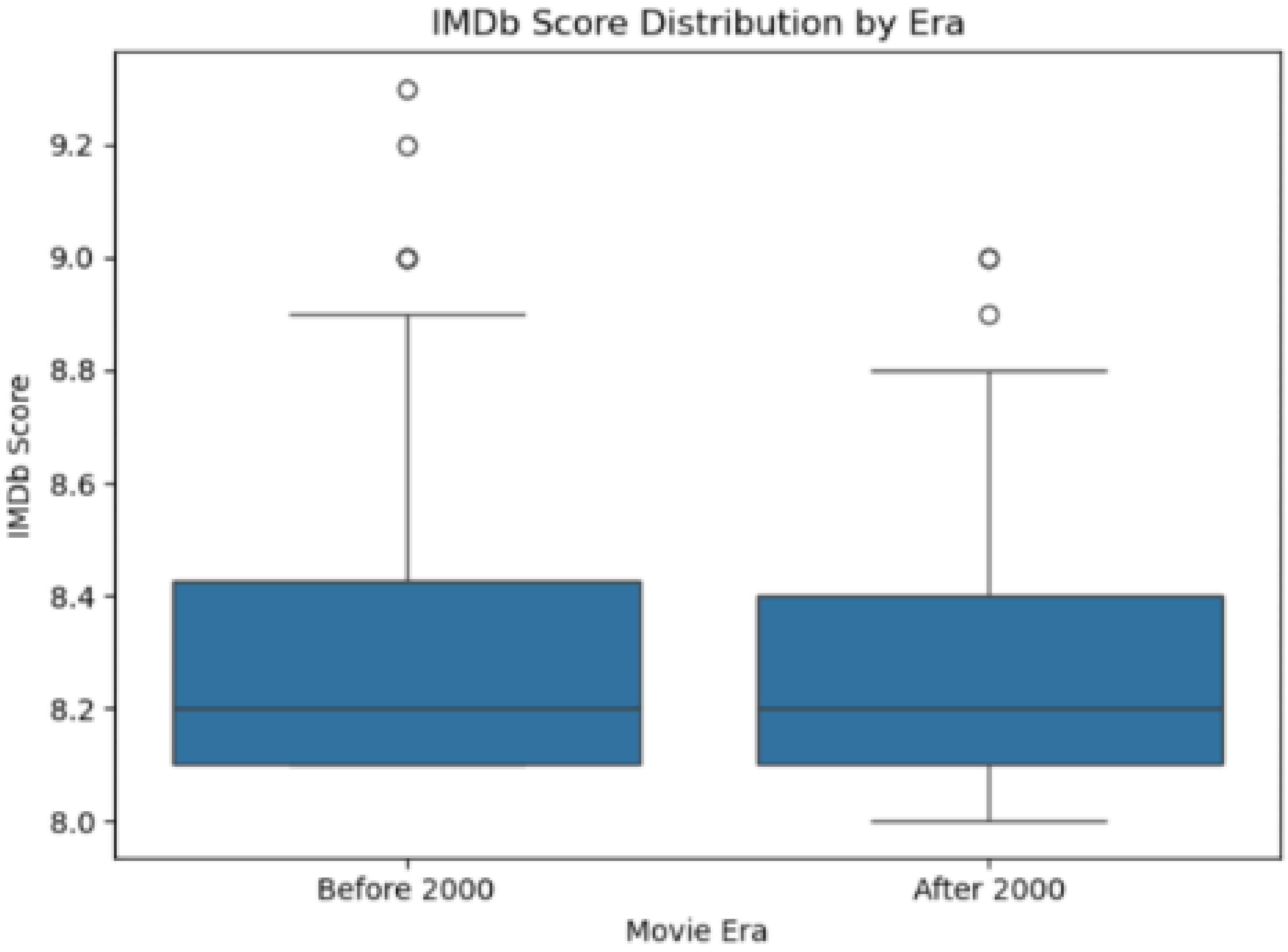
MACHINE LEARNING: DECISION TREE



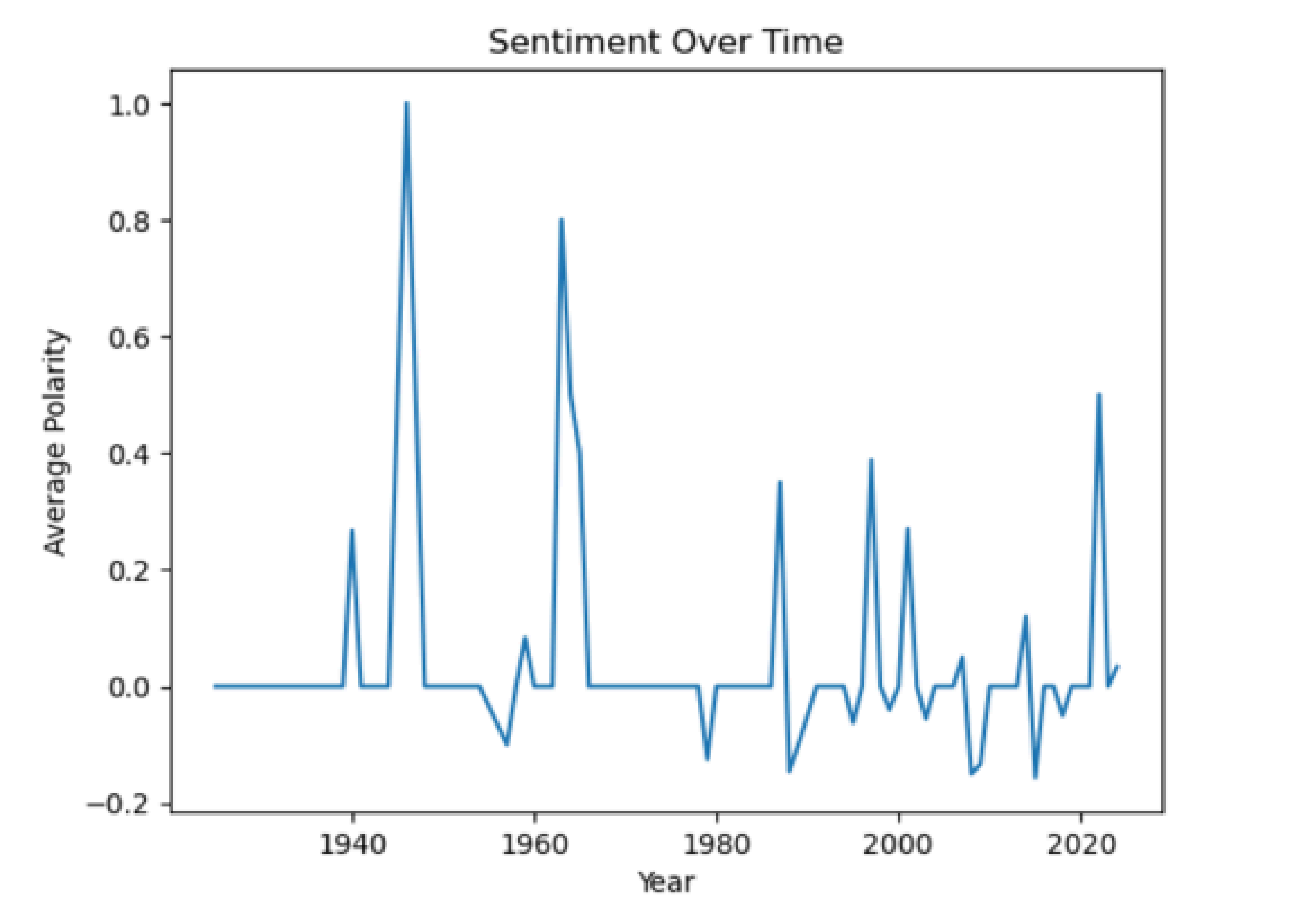
- The accuracy of the model is 0.70 or 70%
  - This is above what we got doing logistic regression but we really need to look at the F1 Score
- The F1 score is 0.20 which is very low and shows it could be predicting a lot of “Popular” movies incorrectly
- The results from this test are still showing that Year alone is not a good predictor of movie popularity

MACHINE LEARNING: KNN

- The model has an accuracy of 0.75 or 75%, the highest so far
    - There could be a pattern between Year and popularity but we need to check the F1 score
- The F1 score is 0.38 and indicates that the model could be imbalanced
- We ran another KNN test and balanced the data
- The accuracy of this model is 0.70 or 70%, lower than the first KNN test
    - This shows that the model is predicting both classes fairly
- The F1 score is 0.47 which is an increase from the previous KNN test
    - This shows the second test is better at identifying “Popular” movies but the score is still low meaning Year alone isn’t a good predictor



SENTIMENT ANALYSIS



- In our analysis, we answered the question, what is the lowest and highest polarity score based on the movie name? Also, what is the sentiment over time for the year column?
- The lowest polarity score was -0.625 for Mad Max: Fury Road 2015
- The highest polarity score was 1.0 for It’s a Wonderful Life and The Best Years of Our Lives
- Around 1945, the average polarity score spiked to 1.0, which is the highest polarity score that 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.

Questions & Challenges

1. How many Oscars has each movie won and how many nominationshas each movie received - and are movies with more Oscars wins more popular than movies without wins (do they have a higher IMDb scores)?
2. Are movies with longer run times more popular than movies with shorter run times or vice versa?
3. Are older movies (before the year 2000) more popular than newer movies (after the year 2000)?
4. What is the average number of Nominations across all movies? Do more movies fall above or below the average?
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?

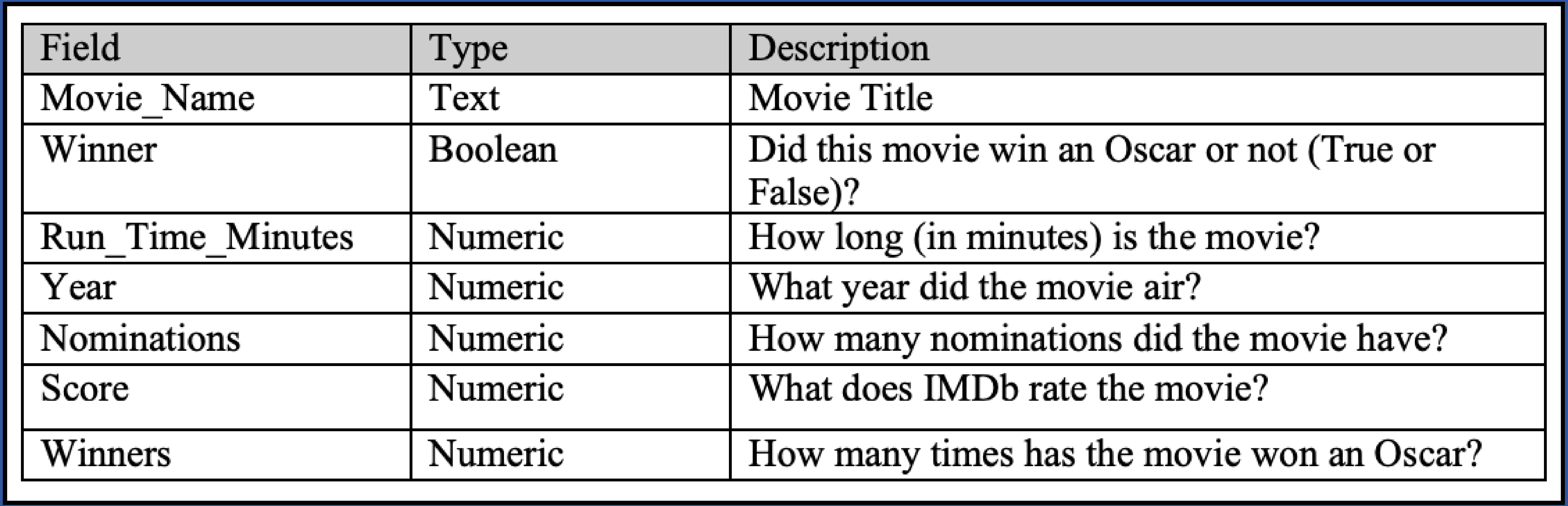
Challenges:

- When trying to compare if 'Winner' and 'Score' have a correlation, it is hard because these movies are all rated very high because they are apart of IMDb's Top 250 movies of all-time

FINAL CONCLUSIONS

- We have cleaned up our Jupyter Notebook and included more graphs in terms of machine learning and sentiment analysis
- We did notice that the IMDb Top 250 movie list updates frequently, so we did our best to get the most up to date graphs and dataframes
- For future groups, we suggest that they pick a website that does not update as frequently as the IMDb website
    - This is because our data would be different every day, so we included the latest dataframe that we ran in our Jupyter Notebook

UP-TO-DATE 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?