IMDb and Oscars Analysis

By: Kate Ryan and Alexis Arthur

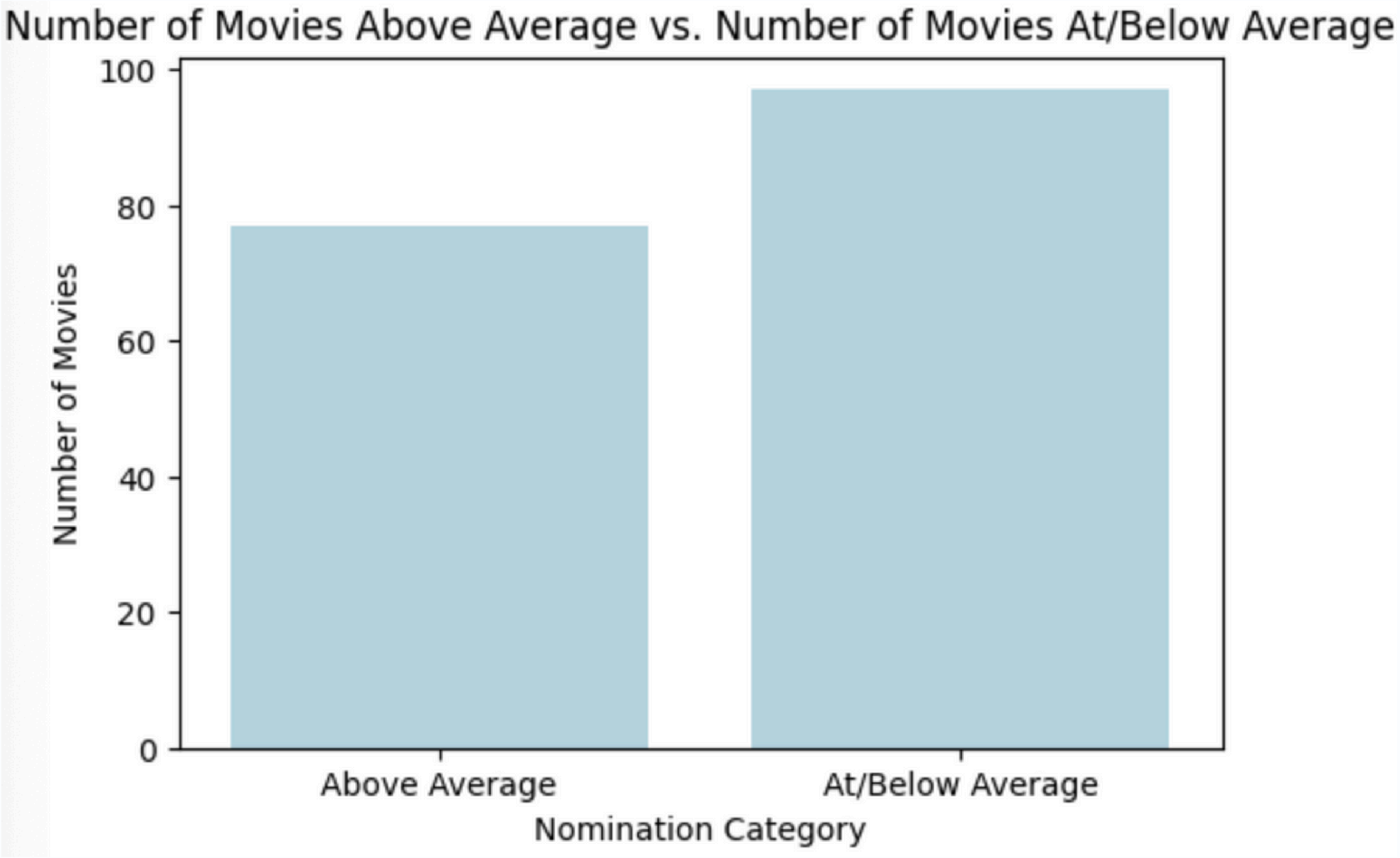
Initial Analysis

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

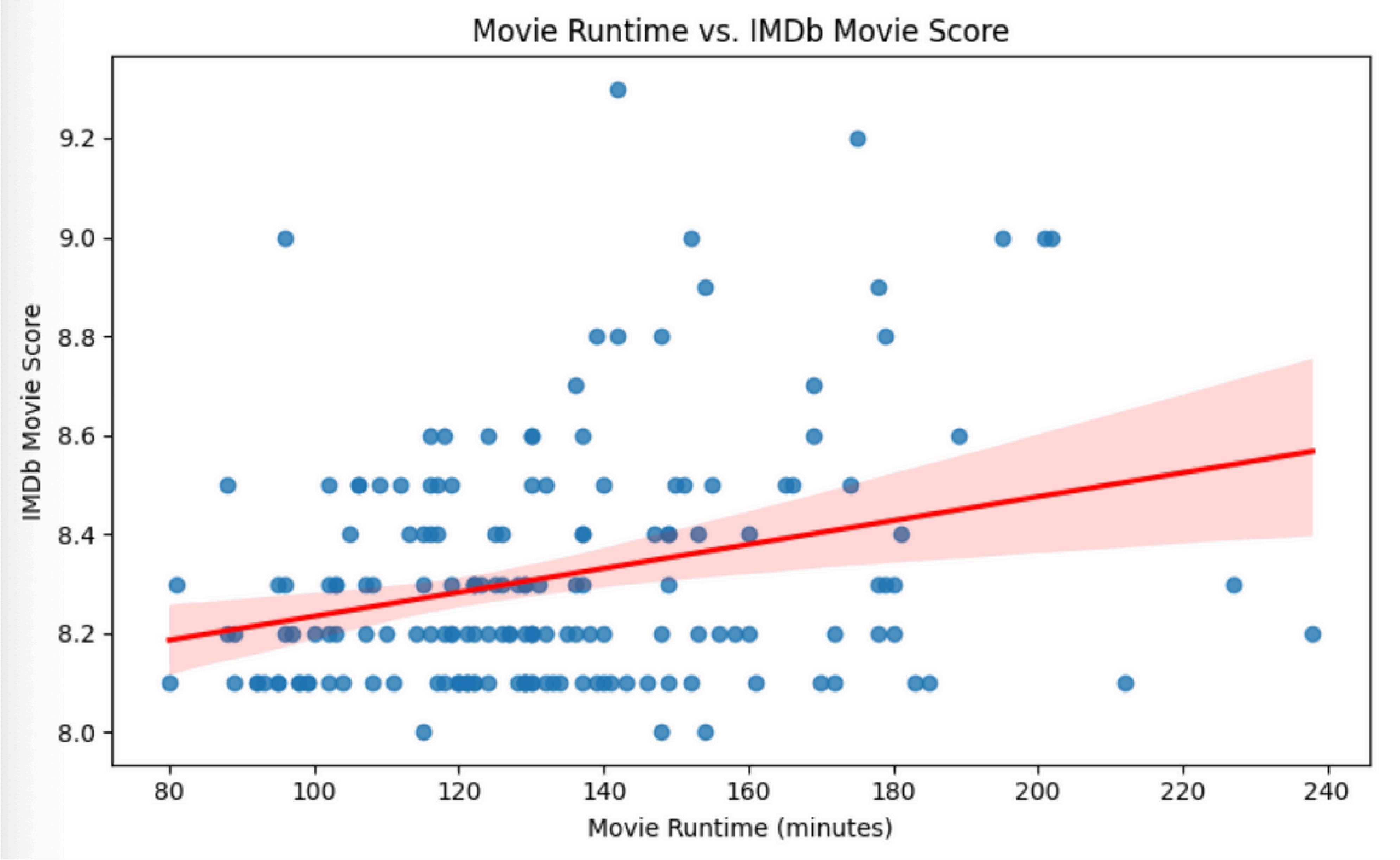
02 Univariate, Bivariate, Hypothesis Test, Visualizations, and Machine Learning

UNIVARIATE

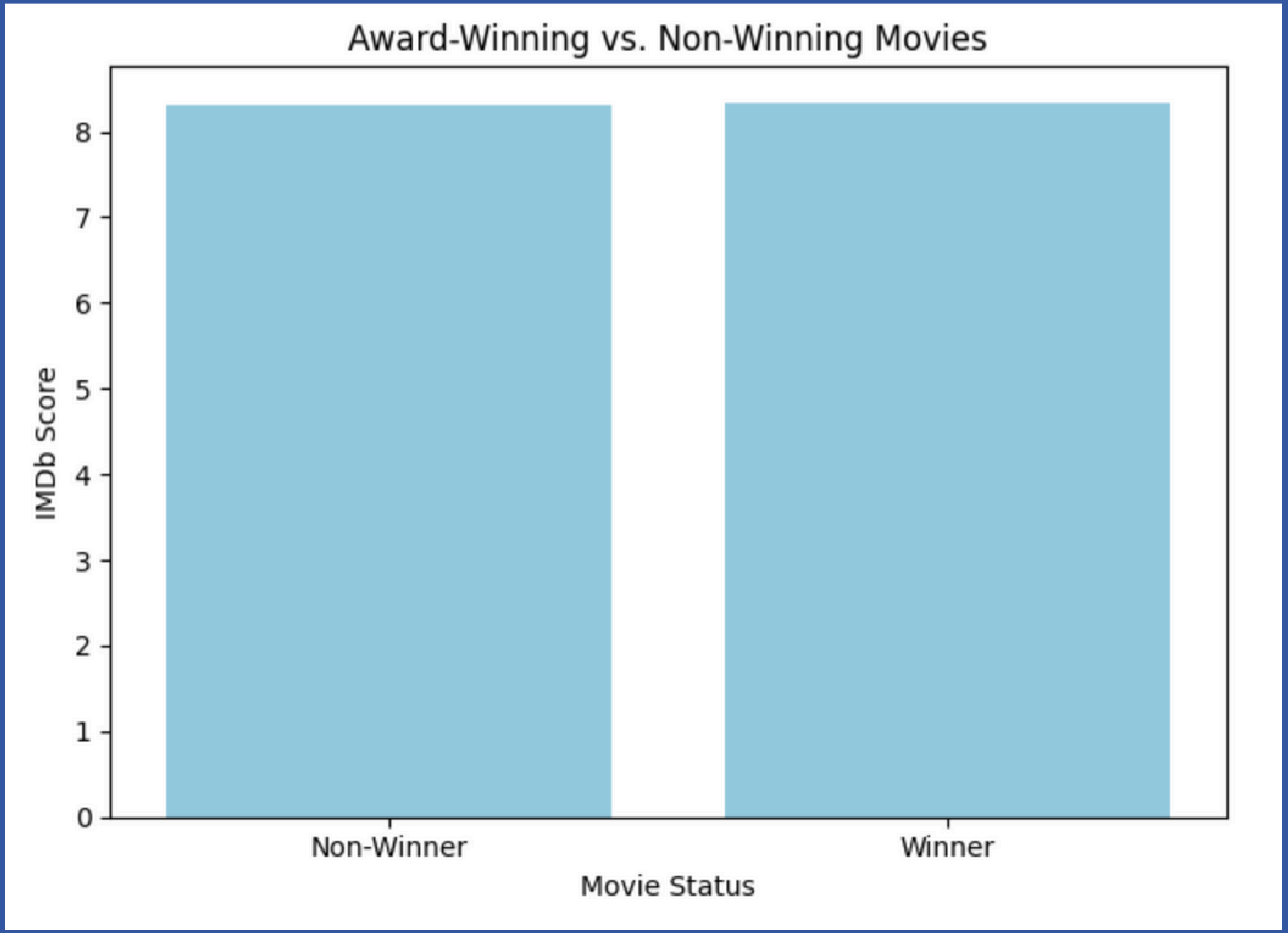
- We are going to look at the 'Nominations' column
- Question: What is the average number of nominations across all the movies? Do more movies fall above or at/below the average?
- We created a bar chart to show the number of movies above average vs. the number of movies at/below the average
- The average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

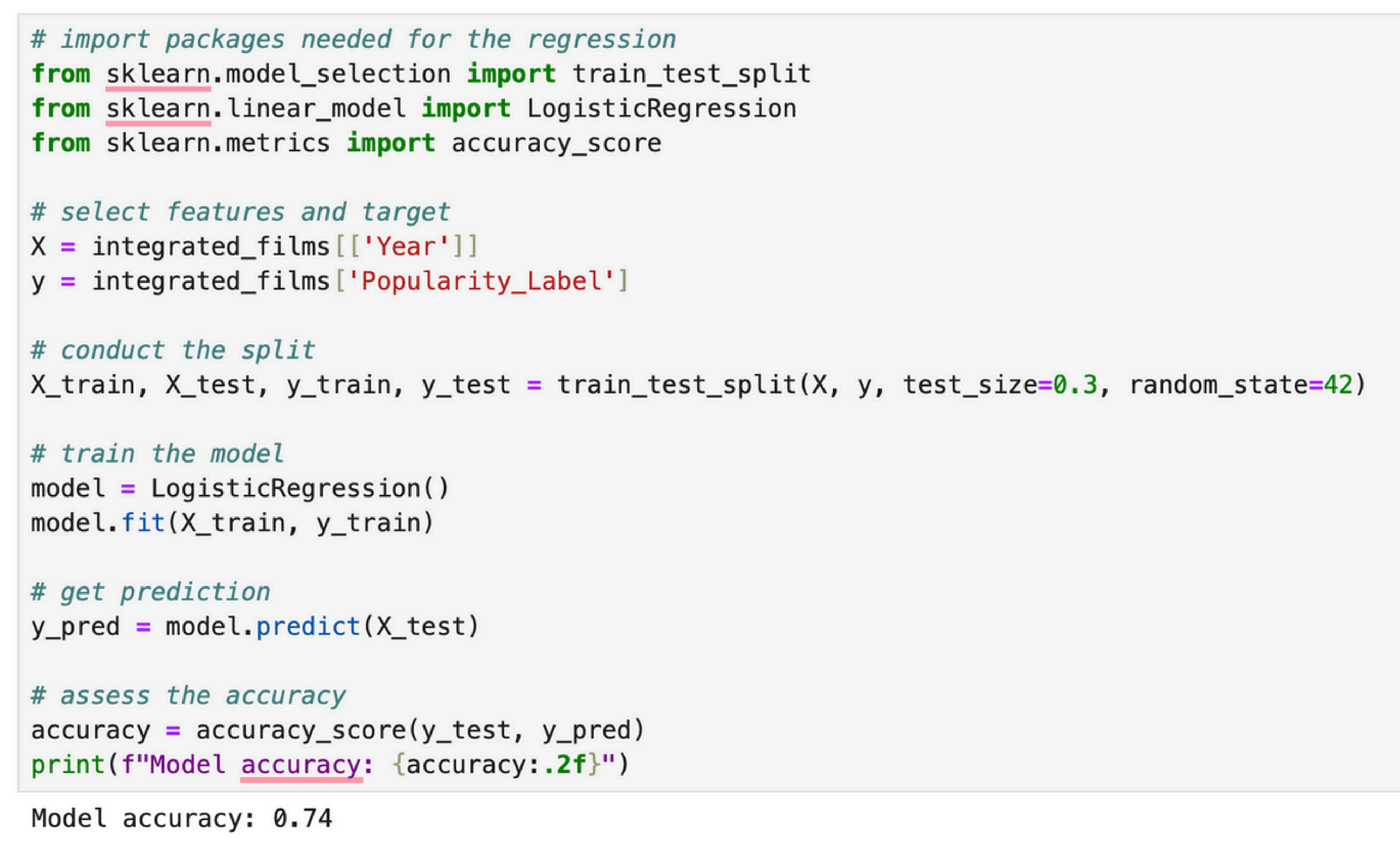


HYPOTHESIS TEST: T-TEST

- The t-statistic is 1.03
- The p-value is 0.30
- The result is not statistically significant
- Award-winning movies and non-winners have a similar performance in terms of their IMDb score

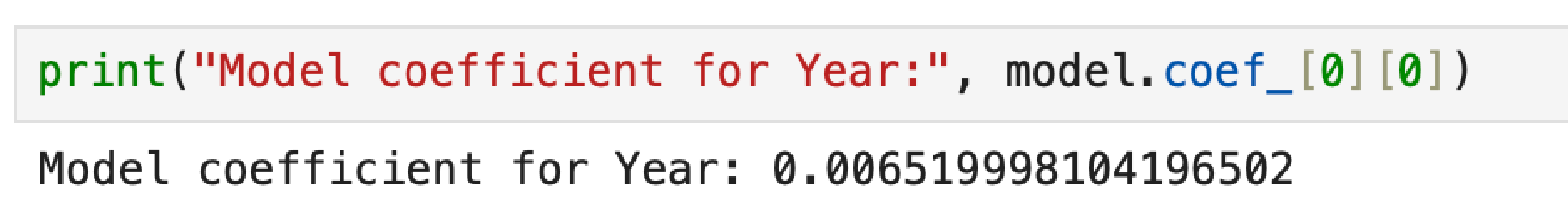
MACHINE LEARNING: LOGISTIC REGRESSION

- Model accuracy is 0.74 or 74%
- Model coefficient for Year is approximately 0.00652
- The regression illustrates that our model was able to predict, with 74% accuracy, that newer movies are slightly more popular than older movies



```
# import packages needed for the regression
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

# select features and target
X = integrated_films[['Year']]
y = integrated_films['Popularity_Label']

# conduct the split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# train the model
model = LogisticRegression()
model.fit(X_train, y_train)

# get prediction
y_pred = model.predict(X_test)

# assess the accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f"Model accuracy: {accuracy:.2f}")
```

```
Model accuracy: 0.74
```



```
print("Model coefficient for Year:", model.coef_[0][0])
```

```
Model coefficient for Year: 0.006519998104196502
```

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?

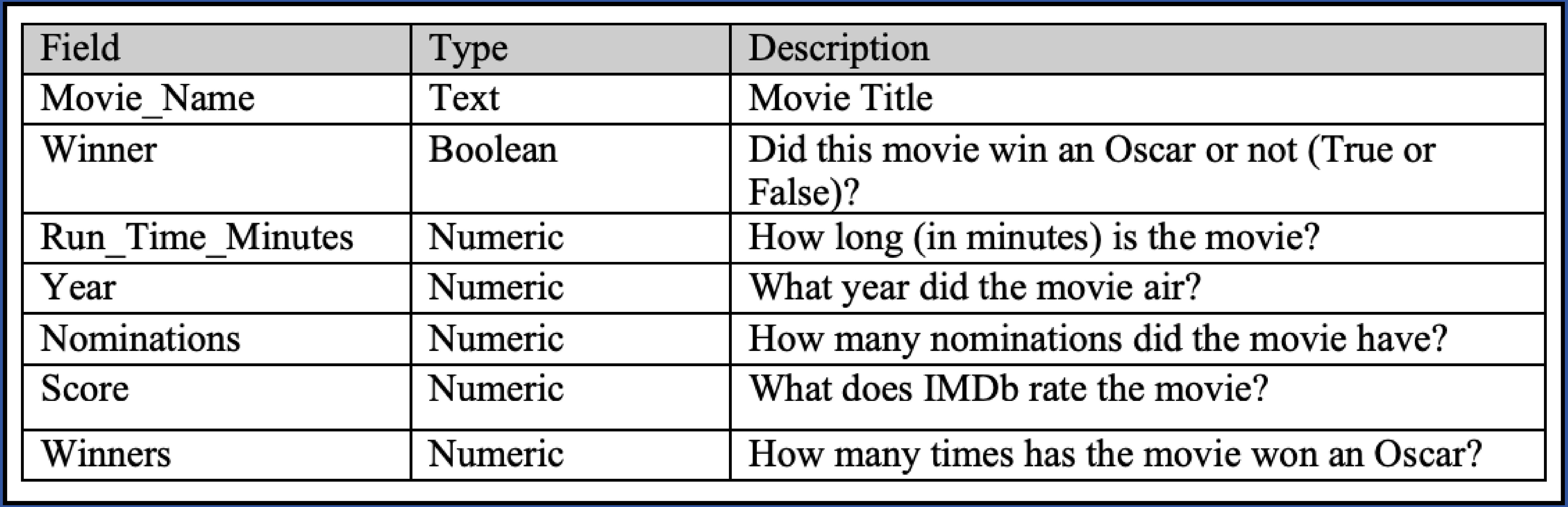
Challenges:

- 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

ROAD MAP

- Answer the question, are older movies (before the year 2000) more popular than newer movies (after the year 2000)?
    - Display the bar graph and utilize ‘Year’ and ‘Nominations’
- In terms of our hypothesis test, is it relevant even though award-winning movies and non-winning movies perform the same in terms of IMDb score? (may be a challenge)
    - Should we do a different hypothesis test?
- Since the project is due Friday, May 9, we will answer our final question and resolve the challenges we stated in the previous slide
    - Also, we will clean up our Jupyter Notebook, so it can be easily read because there is a lot of information we provided on the 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?