Alannah Fleet

Bismaad Minhas

Aashita Rakesh

Caitilin Beatty

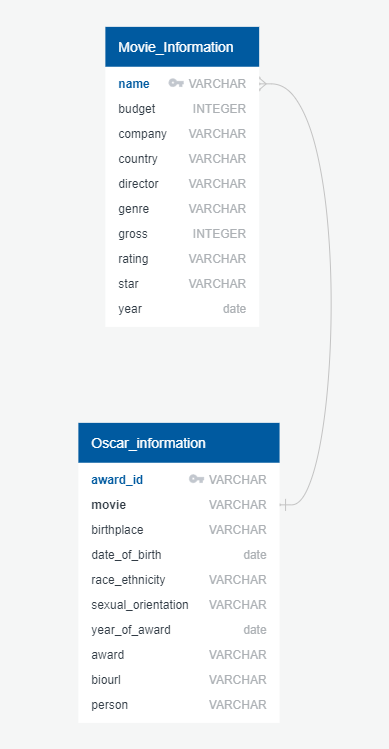
ETL Project Team 7

For this project, we were hired by a production house to use data to study trends that may help them produce successful movies as well as movies that could be contenders for the Oscars. Which factors can predict a movie winning an Oscar? This also illustrates trends of poor diversity of inclusion that the movie industry and the Academy have been criticized for. By focusing on aspects like budgets awarded to movies led by minorities and expanding access to roles may positively impact the diversity of Academy Award winners.

We started with ‘Movie Industry’ dataset from Kaggle.com which contained movies titles, budget, gross, ratings, and other key movie features. Next we obtained ‘Academy Awards Demographics’ from data.world. which tells us Oscar winners for movie, lead actor, and supporting actor. Both datasets included data from multiple decades, Oscar data as far back as 1928 and movie data from the 1986 to 2016.

First step was to clean each dataset. The initial form for each was a CSV file. Each was uploaded into a separate Jupyter notebook for cleaning. Cleaning included dropping null or empty values, dropping columns that were not needed due to their content, and cleaning column names where applicable. The Jupyter notebooks outputted clean CSV’s that were uploaded to a relational database.

We used an ERD to visualize the data and determine best path for moving forward. Using this we see we can join these tables on movie/name. There is an advantage to a relational database such as Postgres, as relational database approach enables a team to create meaningful information by joining the tables. In this case, joining tables will allow us the ability to understand the relationships between the Oscar winners and movie characteristics. The image below is the initial schema outline.



Using Postgres, we then created a schema structure and query script to create a new table joining the movies and Oscars tables. We called this new table ‘winning\_movies’. We can then see not only the Oscar winning movie and Oscar winning person, but also financial information, demographics of the persons involved, movie characteristics, and references to further information.

Final Production Tables

Oscars:



Movies:

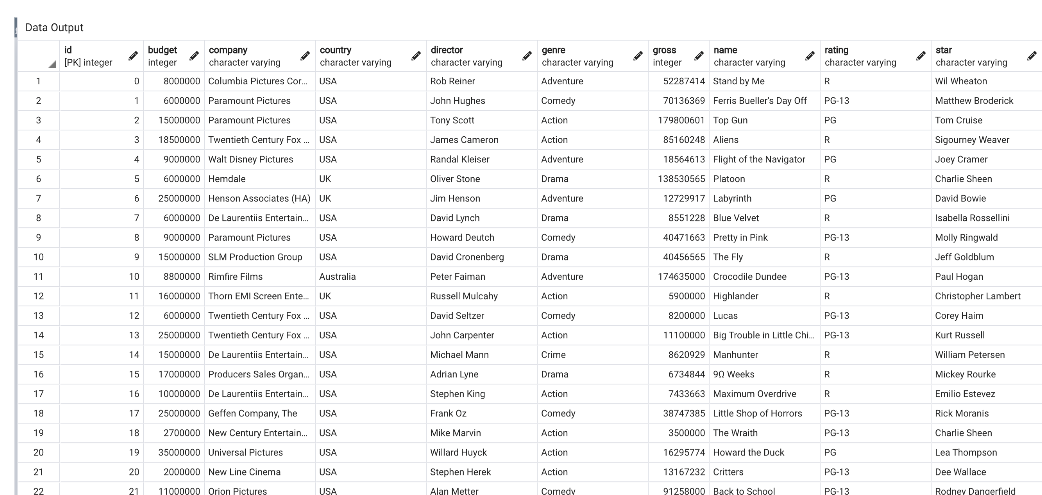
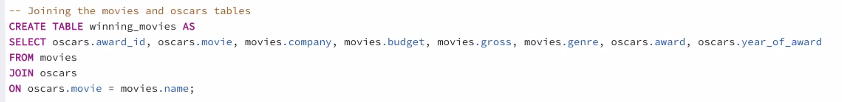
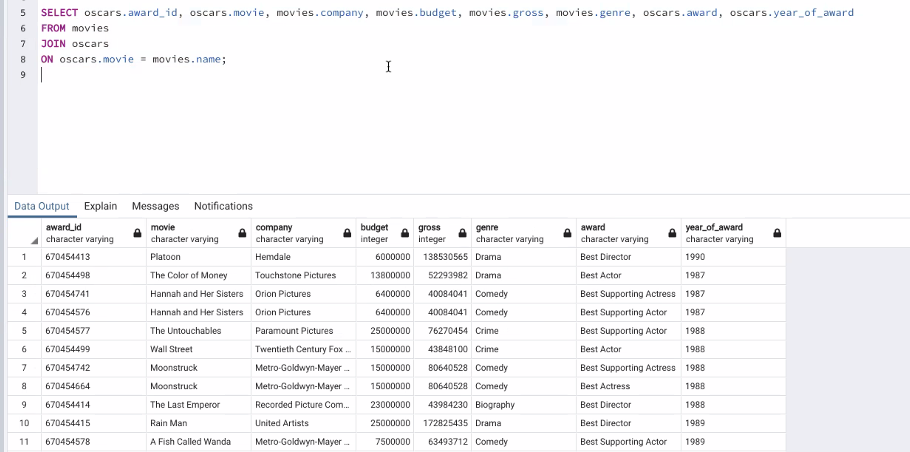


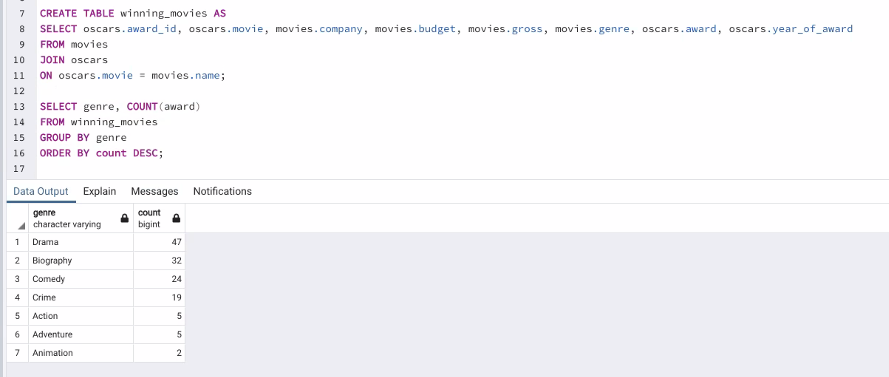
Table Join

Winning Movies:

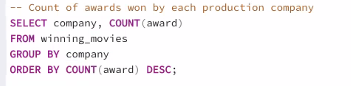


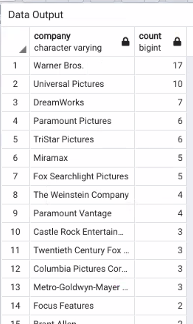


In the below example, we looked to see what movie genres won the most awards. We used a table joining Oscar information and movie information then performed a count to see how many awards were won per genre.

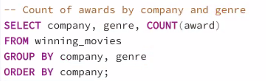


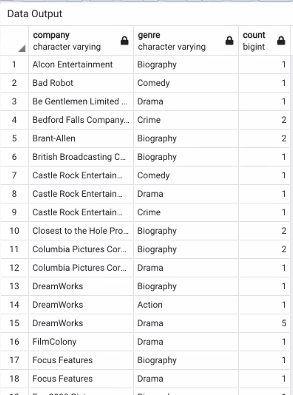
A further question we studied was to see which production company won the most Oscars using the created ‘winning\_movies’ table. We found Warner Bros. won the most Oscars.



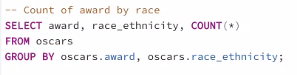


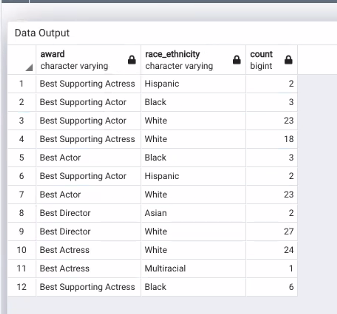
We can further break this down by seeing company and genre using the ‘winning\_movies’ table to see company from the Oscars table and genre from the movies table.



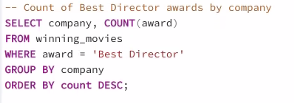
D

We were curious to see the breakdown of ethnicity across Oscar winning best actor, director, best actress, and supporting roles. We used the Oscar table to run a group by and count to determine this information.



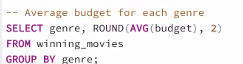


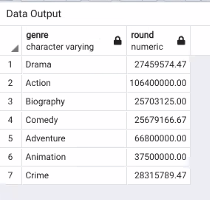
Again, using our created ‘winning\_movies’ table which was a join of Oscars and movies tables, we were able to have a view showing what company had the most ‘Best Director’ awards.





Finally, we used our ‘winning\_movies’ tale to display budget factors with Oscar winning movies. We used an average feature with the budget and genre column, which came from movies table, with the Oscar winners and categories, which came from Oscars table, to find average budget across genres.





References:

<https://www.kaggle.com/danielgrijalvas/movies>

<https://data.world/crowdflower/academy-awards-demographics/workspace/file?filename=Oscars-demographics-DFE.csv>