Project Proposal

CIS 5500

Team Members

Brendan Brett - bbt@seas.upenn.edu - @brendanbrett
Edoardo Palazzi - palazzi@seas.upenn.edu - @EdoardoPalazzi
Eitan Jacob - eitanj@seas.upenn.edu - @EitanJacob

Git Repository

https://github.com/brendanbrett/cis5500-project
Note: this repository is private. We will add our project TA to this repository once identified.

Application Idea

Our web application will be used to serve as an *EGOT* database and tracker. EGOT is an acronym for '**E**mmy, **G**rammy, **O**scar, **T**ony', which are generally recognized as the four major entertainment awards that can be won by artists and performers. Each of these awards serve to honor individuals across the arts:

Emmy award: television
Grammy award: music
Oscar award: movies
Tony award: broadway theater

While winning just one of these awards is an incredible achievement, winning all four awards is a rare feat, and thus the term EGOT was created to refer to those individuals who managed to pull off the impossible. As of February 2024, only 19 artists have managed to obtain the coveted EGOT designation.

The application will have two main purposes:

1. Provides a history of each of these award shows, by providing users with a visual timeline for each award show, allowing users to see the history of entertainment.
2. Provides an overview of current EGOTs, and allows the user to find information about the awards that each EGOT has earned. Beyond that, it will provide some fun statistics and trivia, such as identifying potential future EGOT award winners.

Data Sets

Sourcing & Location

The below datasets represent our current primary sources of awards data. After performing our initial EDA, we did not find the data sets available to us online to be high quality or comprehensive enough, and so we scraped three of the four major datasets ourselves directly from the official award sources, which will greatly improve our ability to create an accurate EGOT database.

Dataset	Original Source	Description	Repository Location
Emmy Award Winners	Kaggle (1949-2019) Project team disaggregated nominees	Comprehensive history of the Emmy Awards. Includes the year, award category, all nominees for the award, the winner of the award, and the title for which they won.	data/emmys/processed/emmy_award_history.csv

Grammy Award Winners	Grammys.com - scraped by project team (code in repo)	Comprehensive history of the Grammy Awards. Includes the year, award category, all nominees for the award, the winner of the award, and the title for which they won.	data/grammys/processed/grammy_award_history.csv
Oscar Award Winners	awardsdatabase.oscars.org/ - scraped by project team (code in repo)	Comprehensive history of the Oscar Awards. Includes the year, award category, all nominees for the award, the winner of the award, and the title for which they won.	data/oscars/processed/oscar_award_history.csv
Tony Award Winners	ibdb.com/awards/ - scraped by project team (code in repo)	Comprehensive history of the Tony Awards. Includes the year, award category, all nominees for the award, the winner of the award.	data/tonys/processed/tony_award_winners.tsv
EGOT Winners	wikipedia.org/wiki/List_of_EGOT_winners - manual download	Dataset of the 19 EGOT winners which includes their name, year in which they won each award, age, and award categories won.	data/egot/egot_winners.csv

Summary & Statistics

Detailed EDA and summary statistics can be found in our Colab notebook.

Emmy Award Dataset:	62,054 rows
Grammy Award Dataset:	62,023 rows
Oscar Award Dataset:	10,563 rows
Tony Award Dataset:	7,387 rows
EGOT Winners Dataset:	19 rows

```
Emmy Award DB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62054 entries, 0 to 62053
Data columns (total 9 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   origin_id  62054 non-null  int64
 1   year       62054 non-null  int64
 2   category   62054 non-null  object
 3   title      61898 non-null  object
 4   nominee    61352 non-null  object
 5   role       37493 non-null  object
 6   company    61509 non-null  object
 7   producer   30130 non-null  object
 8   win        62054 non-null  bool
dtypes: bool(1), int64(2), object(6)
memory usage: 3.8+ MB
None
```

```
Grammy Award DB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62023 entries, 0 to 62022
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   year      62023 non-null  int64
 1   category  62023 non-null  object
 2   title     62023 non-null  object
 3   nominee   62023 non-null  object
 4   win       62023 non-null  bool
dtypes: bool(1), int64(1), object(3)
memory usage: 2.0+ MB
None
```

```
Oscar Award DB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10563 entries, 0 to 10562
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   year      10563 non-null  int64
 1   category  10563 non-null  object
 2   nominee   10563 non-null  object
 3   win       10563 non-null  bool
 4   film      10445 non-null  object
dtypes: bool(1), int64(1), object(3)
memory usage: 340.5+ KB
None
```

```
Tony Award DB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7387 entries, 0 to 7386
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   year      7387 non-null   int64
 1   category  7387 non-null   object
 2   nominee   7387 non-null   object
 3   win       7387 non-null   object
dtypes: int64(1), object(3)
memory usage: 231.0+ KB
None
```

Questions To Be Answered

1. What is the history of the Emmy awards for each year? The Grammy's? The Oscars? The Tony's? Provide an award-by-award presentation for each year.
    a. Identify the award year, award category, award winner, and award nominees
2. Who are the current EGOT award winners?
    a. Identify individuals who were indicated as winners in all 4 primary award datasets
3. What awards did the current EGOTs win?
    a. Identify each award for which they won.
4. What individuals are closest to becoming the next EGOT winners?
    a. Which individuals have been winners in 3 of the 4 primary award datasets.
5. Who is up and coming?
    a. Identify individuals that have won > 2 awards in the past 20 years.

SQL Queries

1. **Finding the EGOT winners:** Among all winners in each awards dataset, who are the artists that won all 4 (EGOT)?
2. **Multiple Winner Timeline**: Among all three-award nominees, is the gap between the first and second awards longer or between the second and third? How about EGOT winners?
3. **Crossover Award Analysis**:
    a. **Overlapping Skillset**: What is the most prevalent combination of two awards among nominees nominated for more than one type of award?
    b. **Trajectory**: Which crossover, such as from Emmy to Grammy or from Tony to Grammy, is the most common in proportion to the combinations found in "2a"?
    c. **Diversity Within Each Award:** How many of these nominees above have also won multiple awards within each of the respective award categories?
4. **Longevity Analysis**: Among the most awarded nominees (e.g., top 1000), how many won all their awards within a span of only five years? How many won their awards across a span of longer than twenty years?
5. **Temporal Engagement Analysis**: Which decade boasts the highest average of consecutive awards among nominees who have been nominated in consecutive years? How about winners?
6. **Resilient Performers**: Who among the top 100 nominees was nominated the most times without winning before finally receiving an award? Over how many years did these losses span, and how many years did the subsequent wins span if there were multiple victories?
7. **Top Award Indicator Analysis**: Identify the most common additional awards nominated for winners of the most prestigious award within each category (e.g., Best Picture for Oscars, Album of the Year for Grammy, Best Musical/Play for Tony, etc.).