**BAIS: 3200**

**Project Report**

Bryce Hansen, Paul Hanson, Mason Heckman, Matthew Mandzen, and Pat Sheehy

**Introduction:**

Netflix is one of the world's largest streaming services and used to be a premier DVD rental service. Netflix has thousands of TV shows and movies to choose from, but which motion pictures are the best? In this project, we hope to identify which genre produces the highest-rated shows, which country produces the most popular movies, and which directors receive the highest ratings on Netflix. Our database application can be used by any Netflix customer to better navigate to the highest-rated genres, country of production, and directors that produce the highest quality motion pictures.

**Data:**

This project uses data from a 2021 Kaggle survey looking at productions that have been featured on Netflix over the years ( <https://www.kaggle.com/akashguna/netflix-prize-shows-information> ). The data features 9827 entries. We have immensely cut down the number of entries from the original data set. To ensure that our application has the most in-depth analysis, we wanted to use as many factors as possible. Additionally, by using motion pictures from multiple decades, we can begin to see trends in the data for the popularity of fields such as genre and director success.

*Table 1 Data Dictionary*

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| ProductionID | Numeric | Unique ID for each Production (e.g. 1201) |
| Title | Text | Title or Name of Production |
| Year | Numeric | Release Year of Production (e.g. 2002) |
| Rating | Numeric | Average Likeliness of Motion Picture voted on by Viewers (e.g. 7.7) |
| Votes | Numeric | Number of Votes to find the Average Rating (e.g. 81) |
| Runtime | Numeric | Length of the Production in Minutes (e.g. 105) |
| DirectorID | Numeric | Unique ID for Person in Charge of Film Production (e.g. D-0001) |
| FirstName | Text | First name of Director |
| MiddleName | Text | Middle name of Director |
| LastName | Text | Last name of Director |
| Genre(s) | Text | Category of Style/Form of Motion Picture |

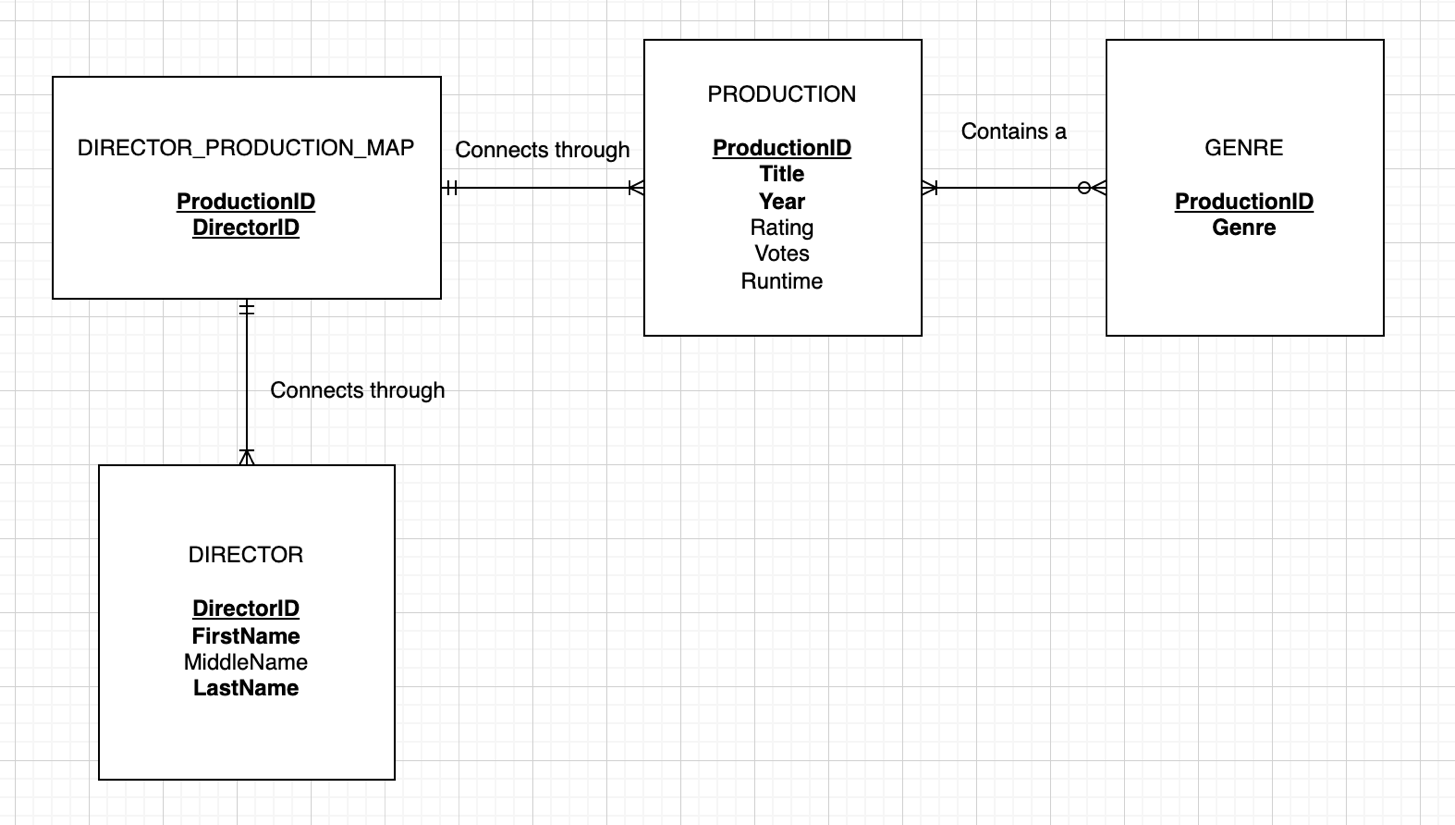
The primary entity of our data is Production, which is identified by the ProductionID. Attributes contained within the entity include Title, Year, Rating, Votes, and Runtime. The attributes Title, Year, and Runtime are mandatory, whereas Rating and Votes are optional.

We created another table titled Director\_Production\_Map. This table allows us to join the Production table to the Director table while keeping it in 3NF form. Without the map table we would have a repeating primary key in the director table.

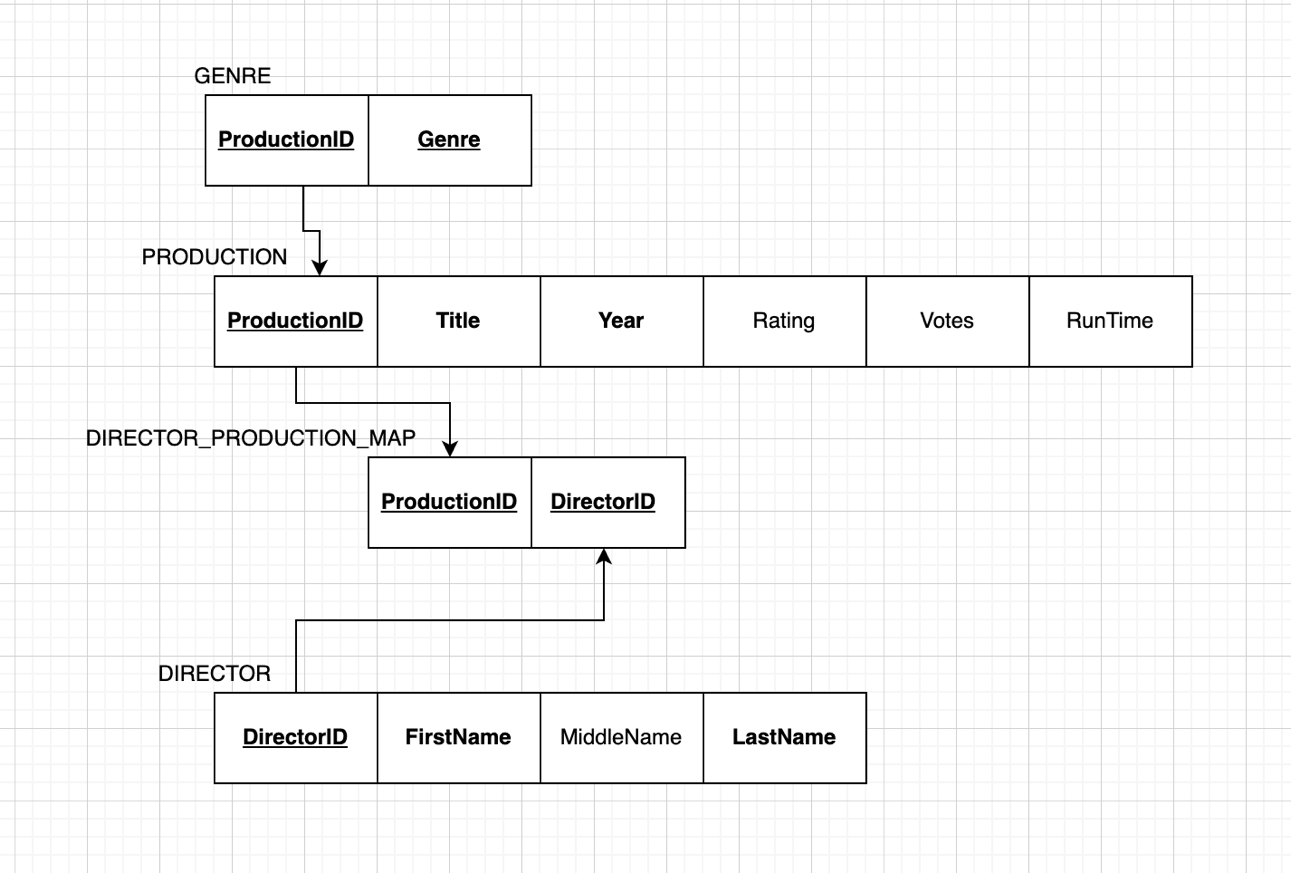
Director is its own entity. It is a weak entity. The Director table is identified by the DirectorID. Attributes contained within the entity include FirstName, MiddleName, and LastName. FirstName and LastName are mandatory, while MiddleName is optional.

Genre is its own entity. It is a weak entity. The attributes of this table, Genre and ProductionID, co-exist in the table as primary keys. ProductionID is also a foreign key referenced from the Production table.

**Diagram:**



*Fig. 1 Entity Relationship Diagram*



*Fig. 2 Graphical relational schema*

**Database Implementation**

We wrote create table commands in APEX to introduce our data into the database.

PRODUCTION

As our parent table, PRODUCTION was created first.

CREATE TABLE PRODUCTION (

ProductionID number(4, 0) NOT NULL,

Title varchar(50) NOT NULL,

Year number(4, 0) NOT NULL,

Rating number(2, 1),

Votes number(10, 0),

Runtime number(3, 0) NOT NULL,

CONSTRAINT ProductionID PRIMARY KEY (ProductionID)) ;

GENRE

CREATE TABLE GENRE (

Genre varchar(30) NOT NULL,

ProductionID number(4) NOT NULL,

CONSTRAINT Genre\_PK PRIMARY KEY (Genre, ProductionID),

CONSTRAINT Genre\_FK FOREIGN KEY (ProductionID) REFERENCES Production(ProductionID)) ;

DIRECTOR PRODUCTION MAP

CREATE TABLE DIRECTOR\_PRODUCTION\_MAP (

DirectorID CHAR (6) NOT NULL,

ProductionID NUMBER (4,0) NOT NULL,

CONSTRAINT DIRECTOR\_PRODUCTION\_MAP\_PK PRIMARY KEY (DirectorID, ProductionID)) ;

DIRECTOR

CREATE TABLE DIRECTOR (

DirectorID CHAR (6) NOT NULL,

FirstName VARCHAR (30) NOT NULL,

MiddleName VARCHAR (30) NULL,

LastName VARCHAR (30) NOT NULL,

CONSTRAINT DIRECTOR\_PK PRIMARY KEY (DirectorID)) ;

**Analysis:**

Our analysis is intended to show Netflix viewers what fields of our data affect the ratings that movies receive, allowing them to make more informed decisions of what they should watch. Research questions include:

* Who are the top 5 directors, based on their average rating?
* Dating back to 1920, what decade has the highest average rating on Netflix?
* What genres produce the highest number of votes on Netflix?
* For movies that are over two hours and movies under an hour, how does Runtime affect the rating of each movie?
* Which year produced the most movies?
* How many genres as each director explored?

**TOP DIRECTORS QUESTION 1**

Which directors produce the best films consistently? To answer this question, we constructed a query that finds the average rating for every director. Following this we returned the list in a descending order to find the best directors.

SELECT Director.DirectorID, round(avg(Rating),1) as Avg\_Rating, FirstName || ' ' || LastName as FullName

FROM director\_production\_map

RIGHT JOIN Director on Director.DirectorID = director\_production\_map.directorID

join production on director\_production\_map.productionID = production.productionID

WHERE Rating not like '%[^0-9]%'

Group by Director.DirectorID, FirstName || ' ' || LastName

order by round(avg(Rating),1) desc

;

The results below show the highest average moving rating amongst directors in this dataset is 9.2. Both T Burnett and Alex Orbison share the title of ‘highest average movie rating’ amongst directors. Surprisingly, directors such as Spielberg and Hitchcock don’t make an appearance on this list. We believe this because they’ve directed a lot more movies than other directors, possibly bringing down their average score.

A screenshot of a computer

Description automatically generated with medium confidence*Fig. 3*

**HIGHEST RATE DECADE QUESTION 2:**

For our second query we wanted to answer the question, which era produces the highest rated films? For this query we conducted a CASE query that would return the average rating for each from the 1920’s to the 2010’s. Also, we ordered the listing by descending to see which era was the highest rated and which era was the lowest rated.

SELECT CASE

WHEN Year < 1930 and Year >= 1920 THEN '1920s'

WHEN Year < 1940 and Year >= 1930 THEN '1930s'

WHEN Year < 1950 and Year >= 1940 THEN '1940s'

WHEN Year < 1960 and Year >= 1950 THEN '1950s'

WHEN Year < 1970 and Year >= 1960 THEN '1960s'

WHEN Year < 1980 and Year >= 1970 THEN '1970s'

WHEN Year < 1990 and Year >= 1980 THEN '1980s'

WHEN Year < 2000 and Year >= 1990 THEN '1990s'

WHEN Year < 2010 and Year >= 2000 THEN '2000s'

ELSE '2010s'

END AS Era,

ROUND(AVG(Rating),2) AS Era\_Rating

FROM PRODUCTION

GROUP BY CASE

WHEN Year < 1930 and Year >= 1920 THEN '1920s'

WHEN Year < 1940 and Year >= 1930 THEN '1930s'

WHEN Year < 1950 and Year >= 1940 THEN '1940s'

WHEN Year < 1960 and Year >= 1950 THEN '1950s'

WHEN Year < 1970 and Year >= 1960 THEN '1960s'

WHEN Year < 1980 and Year >= 1970 THEN '1970s'

WHEN Year < 1990 and Year >= 1980 THEN '1980s'

WHEN Year < 2000 and Year >= 1990 THEN '1990s'

WHEN Year < 2010 and Year >= 2000 THEN '2000s'

ELSE '2010s'

END

ORDER BY Era\_Rating DESC ;

Our results below show that the 1930’s had the highest average rating, and the 2010’s had the lowest average rating. There is correlation between the older movies and higher IMDB. We believe this is because old movie lovers are a niche fan base. If someone is to use IMDB and watch old movies, they have bias towards giving these older films a higher rating. On the other hand, more modern and “trending” movies will be seen by a larger audience revealing more films of a lower quality.

A screenshot of a computer

Description automatically generated with medium confidence*Fig. 4*

**MOST NUMBER OF VOTES BY GENRE QUESTION 3:**

For our third query we wanted to answer the question, what genre receives the most IMDB votes? To answer this question, we conducted a query that found the average votes for each genre as well as only using numeric cells to avoid the null fields. Finally, we grouped by genre and ordered the least to present the highest voted genre first.

Select Genre,ROUND(avg(Votes),0) as Average\_Vote\_For\_Each\_Genre

FROM GENRE JOIN PRODUCTION ON GENRE.ProductionID = PRODUCTION.ProductionID

WHERE Votes not like '%[^0-9]%'

Group by Genre

ORDER BY Average\_Vote\_For\_Each\_Genre DESC

;

In our results below we found that Sci-Fi was the most voted genre, and some of the lower voted genres are dramas and thrillers. Interestingly, dramas are the most frequent genre on the list. We believe that Animation and Sci-Fi movies are the most voted genres because they are typically the most expensive to make. Thus, making all levels of the film a high quality.

A screenshot of a computer

Description automatically generated with medium confidence*Fig. 5*

**RUNTIME EFFECT ON AVERAGE RATING QUESTION 4**

In our fourth query we wanted to answer the question, do shorter or longer movies yield the better rating? To do this we conducted a compound query that selected movies over two hours and movies under an hour. We only wanted the best movies, so we added that minimum rating would have to be above an 8.5. Lastly, we only wanted to numeric data, to avoid the nulls we added a where statement.

select Title, RunTime, Rating

from Production

WHERE (RunTime > 120)

AND Rating not like '%[^0-9]%'

and rating > 8.5

union

select Title, RunTime, Rating

from Production

WHERE (RunTime < 60)

AND Rating not like '%[^0-9]%'

and rating > 8.5

;

Our results below, present that idea that the longer movies typically receive a better rating than the very short movies. In fact, the only ‘movie’ that made the cut was a show from standup comedian Dave Chappelle. We believe our results are this way because the longer the movie is, the more developed plot becomes, and the movie will have more substance to it.

Background pattern

Description automatically generated*Fig. 6*

**WHICH YEAR PRODUCED THE MOST MOVIES QUESTION 5:**

In this query, we wanted to answer the question, which year produced the most movies? To solve this question, we counted how many movies were produced each year, and then grouped by each year. Finally, we wanted to see which year did the most so ordered our list in a descending order.

select year, count(productionid) as Movies\_Per\_Year

from production

group by year

order by count(productionid) desc

;

As our results show below the early 2000’s produced the most movies, followed by the 1990’s. If we take a look at our ratings per era we can see that the 2000’s and 90’s produced some of the worst averaged movies from any era. An observation would be the more movies produced in that year, the lower ratings would likely to be.

Background pattern

Description automatically generated*Fig. 7*

**Question 6: HOW MANY DIFFERENT GENRES HAS EACH DIRECTOR EXPLORED**

Within this query we wanted to know, which director have used the most genres in their movies? To do this we had to join all our tables together, then count how many genres were used by director. And finally, group by director and order by the number of genres used.

select FirstName || ' ' || LastName as FullName, Count(genre) as genre\_count

FROM director\_production\_map

RIGHT JOIN Director on Director.DirectorID = director\_production\_map.directorID

join production on director\_production\_map.productionID = production.productionID

join genre on production.ProductionID = genre.ProductionID

group by FirstName || ' ' || LastName

order by Count(genre) desc

;

Looking at the results below, we find that popular directors such as Spielberg and Allred Hitchcock make an appearance on this list. This list helps users identify how diverse a director is with their genre selection as well as how experienced they are with directing. For example, Spielberg has made a wide variety film, but has also made a lot of films. This informs the users that he is experienced, but he is also inclined to take risks and try out different genres.

Background pattern

Description automatically generated with medium confidence*Fig. 8*

**Web Design**

[https://apex.oracle.com/pls/apex/bmmpp\_project/r/netflix-project](https://apex.oracle.com/pls/apex/bmmpp_project/r/netflix-project/home?session=113124428798639)

Our home page briefly describes the data that we used as well as a link to find the original dataset. We also briefly describe what we have done to cut down that data. The picture on the home screen (Netflix and Chill Graffiti.jpg) shows a Netflix poster with a graffiti ‘chill’ underneath the poster text.

Graphical user interface, website

Description automatically generated*Fig. 9*

**Tables:**

We have created an interactive report for each data table. These reports allow users to filter through columns and select fields. We have also entered a brief description for each of the data tables to ensure users know the importance of each data table and the results they may be able to yield from them.

Graphical user interface

Description automatically generated*Fig. 10*

Graphical user interface, text, application

Description automatically generated*Fig. 11*

Graphical user interface, text, application

Description automatically generated*Fig. 12*

**Queries:**

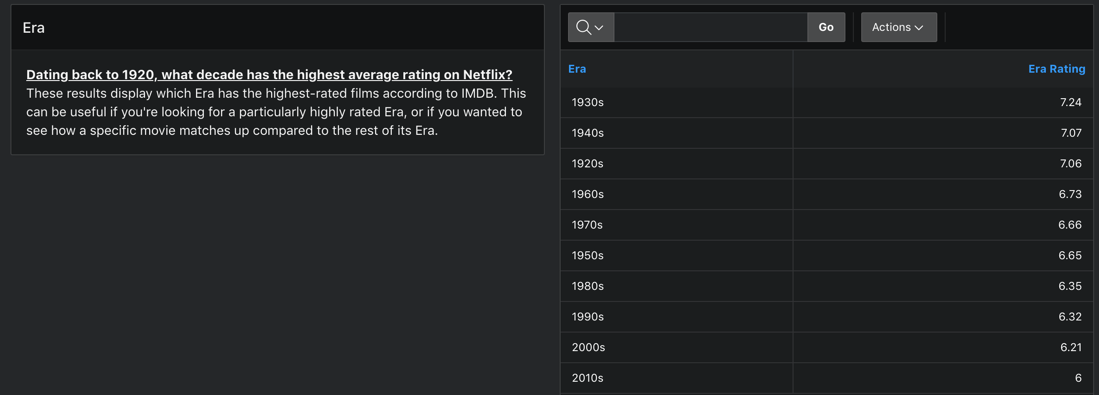
We have provided a picture result of our first query question (Figure 13). This query yields the average IMDB rating for each director, as well as orders them by the highest rating at the top. As shown below popular directors like Quentin Tarantino made the list. Users will be able to filter by the highest rated directors, as well as the ability to look up their favorite directors and see their average rating.

Graphical user interface, text, application

Description automatically generated

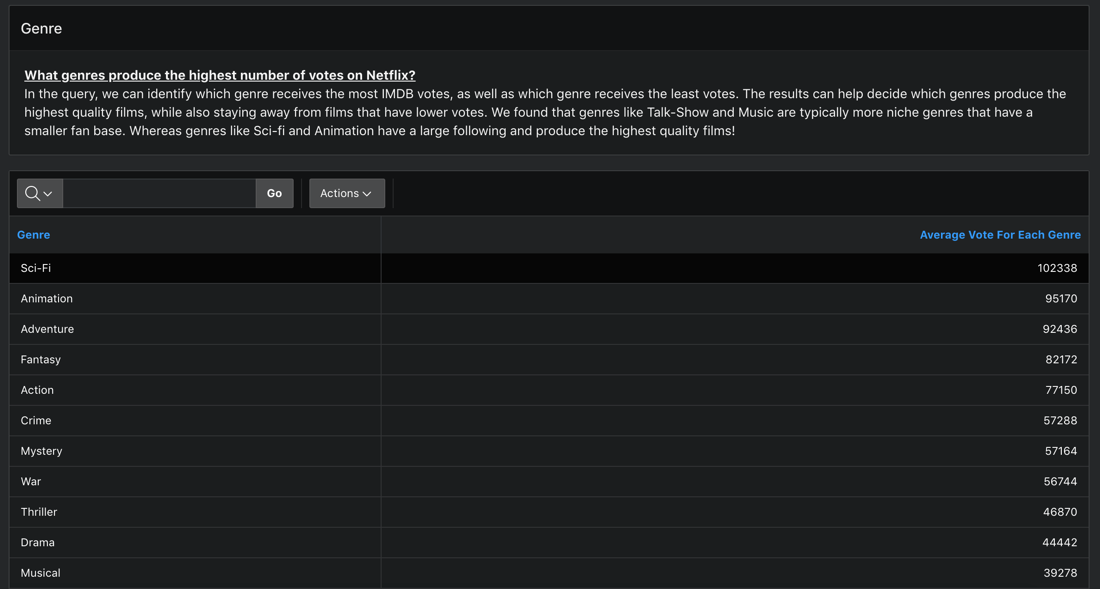
*Fig. 13 Best Directors*

Figure 14 below shows the results of our second query question. This query shows the average IMDB rating from each era starting at the 1920’s and ending in the 2010’s. Users will be able to sort by each era to find lowest and highest rated era.



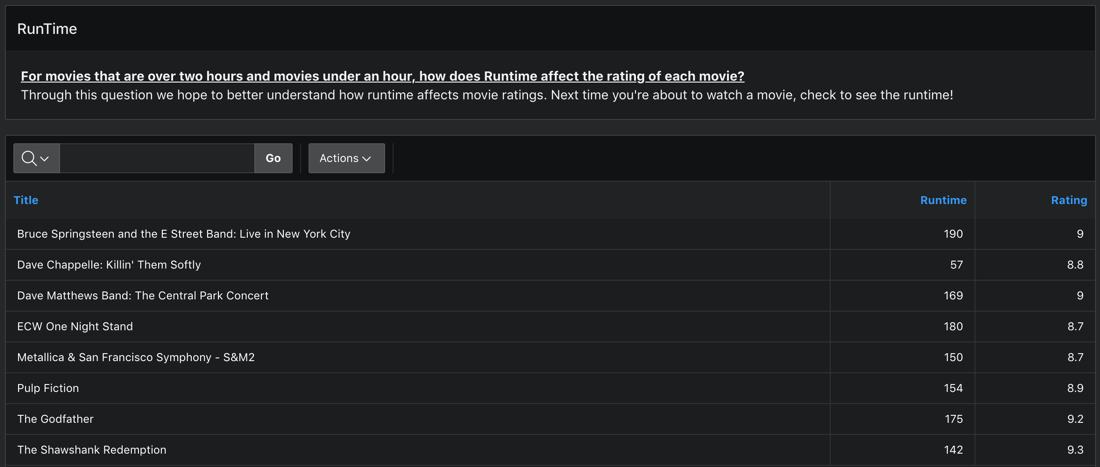
*Fig. 14 Era*

This query provides information on which genres receive the most IMDB votes in descending order from the most votes to the least. From this information, users will be able to see which genres have the biggest fan bases, as the genres with the most votes more than likely have the most people watching them. This will help users find out what genres are well-liked by the general population (Figure 15).



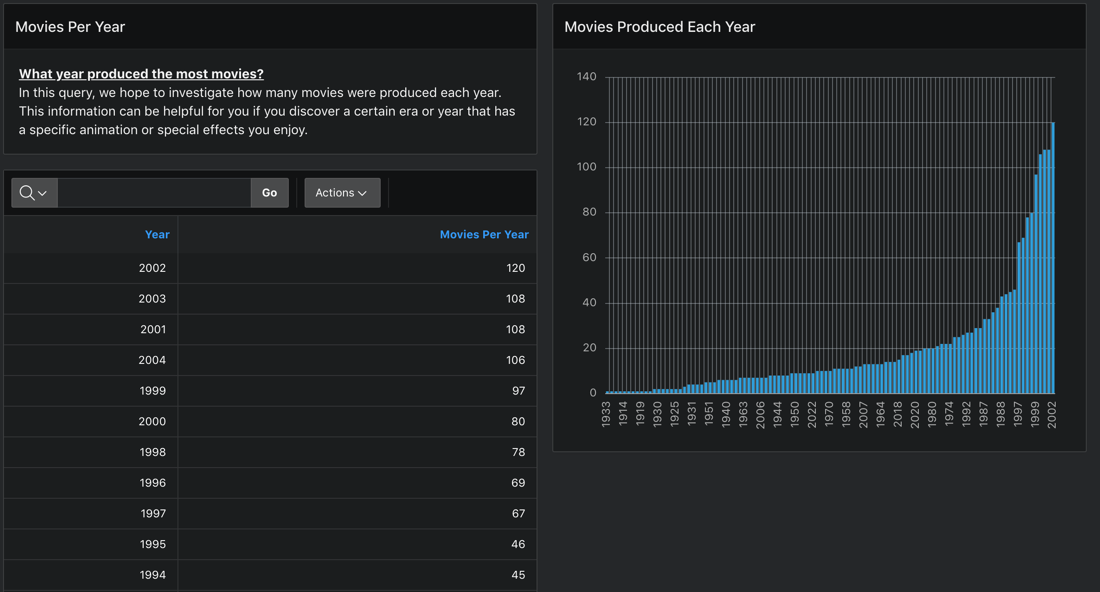
*Fig. 15 Genre*

The picture below shows the result for our fourth query question (Figure 16). The results show the highest rated productions for movies under an hour and movies over two hours. This can help portray to users that the longer a movie is higher the average rating will be. Additionally, this will help provide some quality films for users that want to watch movies over two hours.



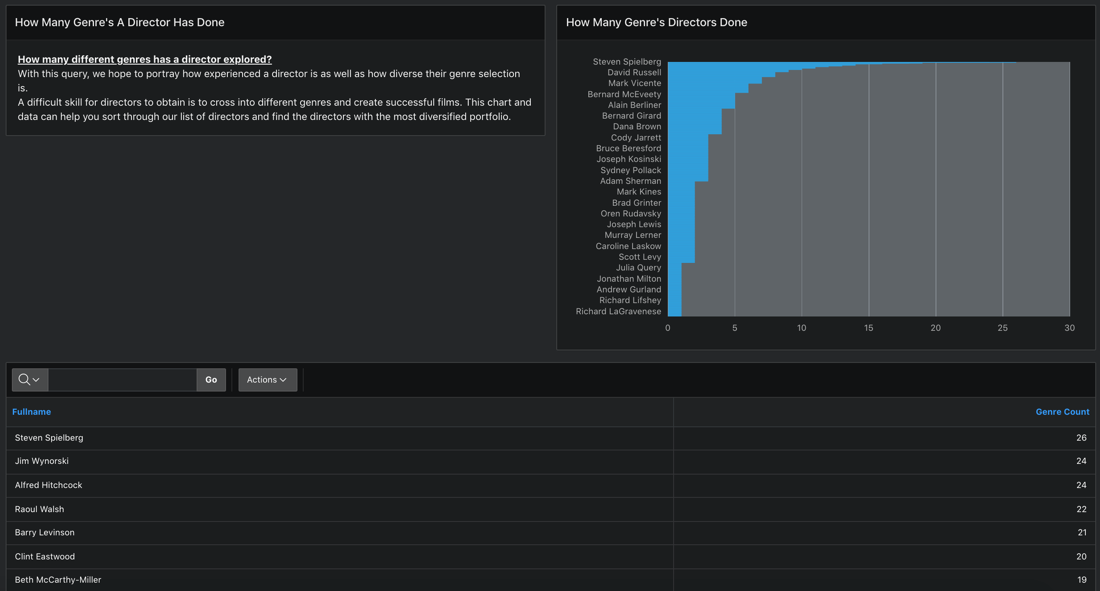
*Fig. 16 RunTime*

Figure 17 below shows the amount of movies on Netflix that were produced in each year. This information can be helpful to users to enjoy movies from a specific time period to see how many options they have to pick from on Netflix. In our graph on the right we display a visual depiction of our data, allowing users to examine the entirety of the data in a centralized location.



*Fig. 17 Movies Per Year*

The picture below shows the result for our sixth query question (Figure 18). The results yield how many genres each director has directed in their career. These results will allow users to filter by directors to find the most diverse and experienced directors. Additionally, users will be able to look up their favorite directors and see how many genres they have conducted in their career. In our graph we visually display the most diverse directors, while also displaying how many genres they have done.



*Fig. 18 How Many Genre’s A Director Has Done*