Project -1

Project Title: Netflix Shows and Movies Project

About Dataset

Netflix - TV Shows and Movies

This data set was created to list all shows available on Netflix streaming, and analyze the data to find interesting facts. This data was acquired in July 2022 containing data available in the United States.

Content

This dataset has two files containing the titles (**titles.csv**) and the cast (**credits.csv**) for the title.
This dataset contains **+5k** unique **titles on Netflix** with 15 columns containing their information, including:

- id: The title ID on JustWatch.
- title: The name of the title.
- show_type: TV show or movie.
- description: A brief description.
- release_year: The release year.
- age_certification: The age certification.
- runtime: The length of the episode (SHOW) or movie.
- genres: A list of genres.
- production_countries: A list of countries that produced the title.
- seasons: Number of seasons if it's a SHOW.
- imdb_id: The title ID on IMDB.
- imdb_score: Score on IMDB.
- imdb_votes: Votes on IMDB.
- tmdb_popularity: Popularity on TMDB.
- tmdb_score: Score on TMDB.

Business Problem: Netflix wants to gather useful insights on their shows and movies for their subscribers through their datasets. The issue is, they are working with too much data (approximately 82k rows of data combined) and are unsure how to effectively analyze and extract meaningful insights from it. They need a robust and scalable data analytics solution to handle the vast amount of data and uncover valuable patterns and trends.

How I Plan On Solving the Problem: In helping Netflix gather valuable insights from their extensive movies and shows dataset, I will be utilizing SQL and a data visualization tool like Tableau to extract relevant information, and conduct insightful analyses. By leveraging SQL's functions, I can uncover key metrics such as viewer ratings, popularity trends, genre preferences, and viewership patterns. Once the data has been extracted and prepared, I will leverage Tableau to present the findings. This will allow for interactive exploration of the data, enabling stakeholders at Netflix to gain actionable insights through visually appealing charts, graphs, and interactive visualizations. I plan on creating a dynamic dashboard in Tableau that enables users to delve into specific movie genres, viewer demographics, or geographical regions.

1. Which movies and shows on Netflix ranked in the top 10 and bottom 10 based on their IMDB scores?

- Top 10 Movies

```sql
SELECT title,
type,
imdb_score
FROM shows_movies.titles
WHERE imdb_score >= 8.0
AND type = 'MOVIE'
ORDER BY imdb_score DESC
LIMIT 10
```

Result:

title	type	imdb_score
Chhota Bheem & Krishna vs Zimbara	MOVIE	9.1
Major	MOVIE	9.1
David Attenborough: A Life on Our Planet	MOVIE	8.9
C/o Kancharapalem	MOVIE	8.9
Inception	MOVIE	8.8
Forrest Gump	MOVIE	8.8
Chhota Bheem & Krishna in Mayanagari	MOVIE	8.7
Bo Burnham: Inside	MOVIE	8.7
A Lion in the House	MOVIE	8.7
Chhota Bheem Neeli Pahaadi	MOVIE	8.7

- Top 10 Shows

```
SELECT title,
type,
imdb_score
FROM shows_movies.titles
WHERE imdb_score >= 8.0
AND type = 'SHOW'
ORDER BY imdb_score DESC
LIMIT 10
```

Result:

title	type	imdb_score
#ABtalks	SHOW	9.6
Khawatir	SHOW	9.5
Breaking Bad	SHOW	9.5
Our Planet	SHOW	9.3
Avatar: The Last Airbender	SHOW	9.3
Reply 1988	SHOW	9.2
Kota Factory	SHOW	9.1
The Last Dance	SHOW	9.1
My Mister	SHOW	9.1
Okupas	SHOW	9

- Bottom 10 Movies

```
SELECT title,
type,
imdb_score
FROM shows_movies.titles
WHERE type = 'MOVIE'
ORDER BY imdb_score ASC
LIMIT 10
```

Result:

title	type	imdb_score
He's Expecting	SHOW	2
Thomas & Friends: All Engines Go!	SHOW	2
Hype House	SHOW	2.1
A House of Blocks	SHOW	2.3
Until Dawn	SHOW	2.4
The Goop Lab	SHOW	2.5
Byron Baes	SHOW	2.6
First Class	SHOW	2.8
Bonus Family	SHOW	2.9
Richie Rich	SHOW	3

An IMDB score is a widely recognized measure of the overall quality and popularity of a movie or show. The top 10 movies and shows stood out for their exceptional IMDB scores, indicating that they are highly regarded by viewers. These titles have likely garnered significant acclaim and positive reviews, contributing to their high rankings within the Netflix library. Viewers who are seeking quality content would find these selections very appealing. On the other hand, the bottom 10 movies and shows had lower IMDB scores. While these entries may not have resonated as strongly with audiences, it's important to note that many factors influence these rankings such as individual preferences, weak plot, poor acting, and low-quality production. By uncovering the top and bottom performers based on IMDB scores, this project sheds light on the varying levels of audience reception and highlights titles that are likely to be well-received and those that may have room for improvement. These findings can provide valuable insights for viewers seeking highly-rated content and can serve as a basis for further analysis and decision-making for Netflix's audience recommendations.

2. How many movies and shows fall in each decade in Netflix's library?

```
SELECT CONCAT(FLOOR(release_year / 10) * 10, 's') AS decade,
        COUNT(*) AS movies_shows_count
FROM shows_movies.titles
WHERE release_year >= 1940
GROUP BY CONCAT(FLOOR(release_year / 10) * 10, 's')
ORDER BY decade;
```

Result:

decade	movies_shows_count
1940s	1
1950s	5
1960s	8
1970s	18
1980s	52
1990s	121
2000s	369
2010s	3304
2020s	1972

The results of the SQL query provide a fascinating insight into the distribution of movies and shows across different decades in Netflix's library. The data reveals a significant shift in content availability over time, with a notable increase in the number of titles from the 2000s onwards. Starting from the earlier decades, the 1940s-1980s showcase a small fraction of the total entries, suggesting that Netflix's collection from these decades is relatively limited. The 1990s demonstrate a large surge in offerings, with 121 titles. However, the true turning point in Netflix's library occurs in the 2010s with a remarkable 3,304 movies and shows from this decade. This abundance highlights Netflix's dedication to featuring contemporary content that aligns with current trends and audience preferences.

Even though the 2020s are still in progress, the dataset reveals an impressive count of 1,972 movies and shows, indicating a strong focus on acquiring and producing content from recent years. Overall, these findings shed light on Netflix's strategy of curating library that covers a wide range of decades. The significant increase in content availability from the 2000s onwards suggests a concerted effort to offer a diverse selection of titles. This collection spanning multiple decades allows Netflix's audience to explore a variety of movies and shows that reflect different eras.

3. How did age-certifications impact the dataset?

```
SELECT DISTINCT age_certification,
ROUND(AVG(imdb_score),2) AS avg_imdb_score,
ROUND(AVG(tmdb_score),2) AS avg_tmdb_score
FROM shows_movies.titles
GROUP BY age_certification
ORDER BY avg_imdb_score DESC
```

Result:

age_certification	avg_imdb_score
TV-14	6.71
TV-MA	6.62
TV-PG	6.37
TV-Y7	6.32
PG-13	6.2
TV-G	6.01
R	6
TV-Y	6
PG	5.96
NC-17	5.76
N/A	5.58
G	5.09

```sql
SELECT age_certification,
COUNT(*) AS certification_count
FROM shows_movies.titles
WHERE type = 'Movie'
AND age_certification != 'N/A'
GROUP BY age_certification
ORDER BY certification_count DESC
LIMIT 5;
```

Results:

age_certification	certification_count
R	556
PG-13	451
PG	233
G	124
NC-17	16

The first query focused on the average IMDB scores associated with each age certification, revealing interesting trends in audience ratings. According to the data, TV-14 emerges as the age certification with the highest average IMDB score of 6.71. This suggests that content designated for viewers aged 14 and older tends to receive relatively favorable ratings. The age certification TV-G obtains an average score of 6.01, signaling the appreciation for content suitable for all audiences. On the other hand, the age certifications R and TV-Y, each with average scores of 6, demonstrate that while they may have lower ratings, there is still a substantial audience that finds enjoyment in these respective categories.

When examining the distribution of movies and shows across age certifications, the second query showcases the varying prevalence of different certifications within Netflix's dataset. R emerges as the most prevalent age certification, with 556 titles falling under this category. PG-13 closely follows with 451 titles, reflecting a significant number of movies and shows targeted at mature audiences. The age certification PG accounts for 233 titles, indicating a considerable selection suitable for general audiences. The dataset also includes 124 titles classified as G, which mostly caters to a younger audience. Lastly, the least represented certification is NC-17, with only 16 titles available. These findings highlight the diverse range of age certifications present in Netflix's movies and shows dataset and provide valuable insights into both audience preferences and content distribution. The higher average scores associated with

TV-14, TV-MA, and TV-PG certifications suggest that content aligned with these age categories tends to resonate positively with viewers.

4. Which genres are the most common?

- Top 10 most common genres for MOVIES

```
SELECT genres,
COUNT(*) AS title_count
FROM shows_movies.titles
WHERE type = 'Movie'
GROUP BY genres
ORDER BY title_count DESC
LIMIT 10;
```

Result:

genres	title_count
['comedy']	384
['documentation']	230
['drama']	224
['comedy', 'documentation']	100
['comedy', 'drama']	84
['drama', 'romance']	76
['drama', 'comedy']	65
['comedy', 'romance']	60
['romance', 'comedy']	48
['comedy', 'drama', 'romance']	44

- Top 10 most common genres for SHOWS

```
SELECT genres,
COUNT(*) AS title_count
FROM shows_movies.titles
WHERE type = 'Show'
GROUP BY genres
ORDER BY title_count DESC
LIMIT 10;
```

Result:

genres	title_count
['reality']	113
['drama']	104
['comedy']	100
['documentation']	99
['comedy', 'drama']	51
['drama', 'romance']	48
['drama', 'comedy']	44
['documentation', 'crime']	35
['animation']	30
[]	25

- Top 3 most common genres OVERALL

```
SELECT t.genres,
COUNT(*) AS genre_count
FROM shows_movies.titles AS t
WHERE t.type = 'Movie' or t.type = 'Show'
GROUP BY t.genres
ORDER BY genre_count DESC
LIMIT 3;
```

Result:

genres	genre_count
['comedy']	484
['documentation']	329
['drama']	328

By analysing the frequency of genres, we can gain a better understanding of the content that dominates the platform and the preferences of its audience. Starting with movies, the first query reveals the top 10 most common genres. Comedy emerges as the most popular genre with a total of 384 movies, reflecting its widespread appeal. Following closely behind are documentation with 230 movies and drama with 224 movies, indicating the significance of these genres in Netflix's movie collection. Combinations of genres also feature prominently, with comedy + documentation and comedy + drama occupying the fourth and fifth positions respectively. The presence of drama + romance, drama + comedy, and comedy + romance further emphasizes the audience's likeness for movies that blend multiple genres. These findings highlight the diverse range of movie genres available on Netflix and the platform's commitment to catering to a wide array of preferences.

Shifting the focus, the second query presents the top 10 most common genres for shows. Reality takes the lead with 113 shows, showcasing the popularity of this genre among Netflix viewers. Drama follows closely behind with 104 shows. Comedy and documentation also emerge as prevalent genres with 100 shows each. Similar to movies, combinations of genres such as comedy + drama and drama + romance are present, indicating viewer interest in multi-genre shows.

Combining the results from both movies and shows, the third query provides an overview of the top three most common genres overall. Comedy takes the lead with a total of 484 entries, reaffirming its position as a very popular genre among Netflix subscribers. Documentation follows closely behind with 329 entries, reflecting the popularity of informative content. Finally, drama secures the third

spot with 328 entries. Overall, these findings shed light on the genres that dominate Netflix's library, showcasing the platform's efforts to cater to a diverse range of viewer preferences.

Conclusion

By exploring various aspects of the dataset, a comprehensive understanding of Netflix's content landscape was gained. The analysis revealed the top 10 and bottom 10 movies and shows based on their IMDB scores, which highlighted the titles that garnered high praise and those that received lower ratings. This information can assist viewers in making informed choices and highlight areas for potential improvement in content quality. The examination of movies and shows distributed across different decades showed significant shifts in content availability over time. Notably, the dataset showcased a substantial increase in offerings from the 2000s onwards, emphasizing Netflix's commitment to featuring newer content that resonates with current trends and audience preferences.