# Streaming Media Consumption

## Group 2

Ishan Varshney, Rose Absin,
Shadi Bitarafhaghighi, Viraj Vijaywargiya

# Introducing Project Objective

Identifying evolution of user consumption patterns of media content such as movies and tv shows on streaming services during COVID

**End goal**: Predictions for actionable insights for film and tv show production and development

**Methodology**:

- Identifying top-performing genres
- Building predictive models to identify factors that maximize user engagement
- Verifying predictions through sentiment analysis and Latent Dirichlet Allocation (LDA) on IMDb dataset

# Data Sources

- Streaming Media Consumption Dataset (Middle-east streaming data)
- IMDB Reviews Dataset

# Streaming Consumption Dataset Description

| consumption_month | consumption_isoweek | consumption_day_shift | Content_type | content_language | content_genre | total_duration_sec | unique_users |
|---|---|---|---|---|---|---|---|
| Jan-19 | 4 | Afternoon | Season | English | Documentary | 486,611 | 269 |
| Jan-19 | 1 | Afternoon | Season | Arabic | Animation | 823,849 | 318 |
| Jan-19 | 4 | Evening | Movie | Arabic | Comedy | 118,629 | 59 |
| Jan-19 | 1 | Night | Season | English | Documentary | 250,794 | 109 |
| Jan-19 | 3 | Evening | Movie | English | Comedy | 3,861,637 | 1,452 |

## Variable description

**consumption_month**: month and year

**consumption_isoweek**: week of the year

**consumption_day_shift**: time of day

**Platform**: streaming device type

**total_duration_sec**: total number of seconds watched by the users

**content_type**: TV series or movie

**content_genre**: genre

**unique_users**: total number of distinct users that watched the media content

# IMDb Dataset Description: Reviews

**Positive Review:**

Bromwell High is nothing short of brilliant. Expertly scripted and perfectly delivered, this searing parody of a students and teachers at a South London Public School leaves you literally rolling with laughter. It's vulgar, provocative, witty and sharp. The characters are a superbly caricatured cross section of British society (or to be more accurate, of any society). Following the escapades of Keisha, Latrina and Natella, our three "protagonists" for want of a better term, the show doesn't shy away from parodying every imaginable subject. Political correctness flies out the window in every episode. If you enjoy shows that aren't afraid to poke fun of every taboo subject imaginable, then Bromwell High will not disappoint!

**Negative Review:**

Robert DeNiro plays the most unbelievably intelligent illiterate of all time. This movie is so wasteful of talent, it is truly disgusting. The script is unbelievable. The dialog is unbelievable. Jane Fonda's character is a caricature of herself, and not a funny one. The movie moves at a snail's pace, is photographed in an ill-advised manner, and is insufferably preachy. It also plugs in every cliche in the book. Swoozie Kurtz is excellent in a supporting role, but so what?<br /><br />Equally annoying is this new IMDB rule of requiring ten lines for every review. When a movie is this worthless, it doesn't require ten lines of text to let other readers know that it is a waste of time and tape. Avoid this movie.

# Connection between Datasets

## Streaming Consumption Dataset

1. Perform regression and predictive analysis on streaming consumption dataset
2. **Top performing genres** is one of the primary takeaway from this analysis

## IMDb Dataset

1. Determine which kind of movies have received positive reviews (even if movies were released 2011 and prior, they will still be available on streaming)
2. Tag and cluster these movies by genre through LDA

Cross-validate top genres derived from both dataset

# Database Management + Tools & Libraries

## Python

LDA/LLM libraries like Pandas, scikit, Transformers, spaCy, numpy, plotly, and etc.

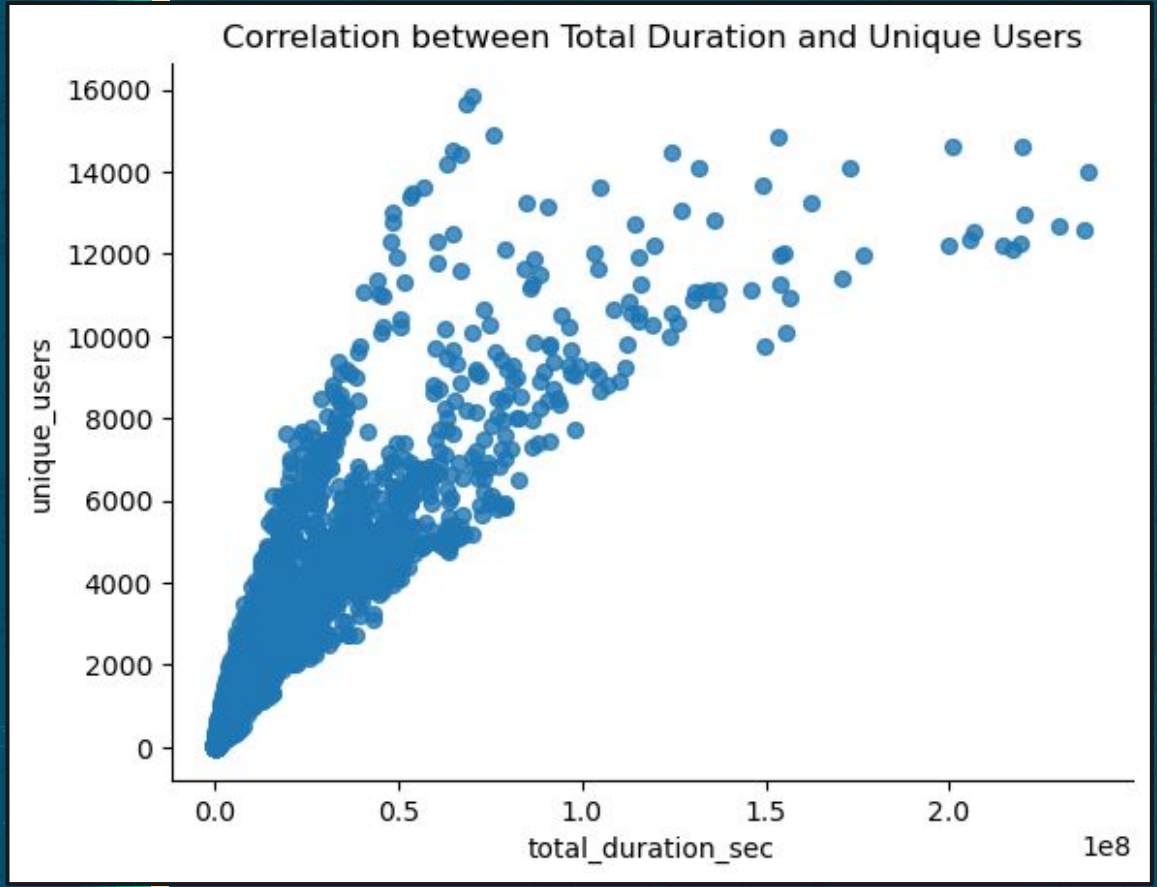
## R

Regression through LM and GLM for numerical data and predictions

## PostgreSQL

For data hosting and querying relevant subsets of data

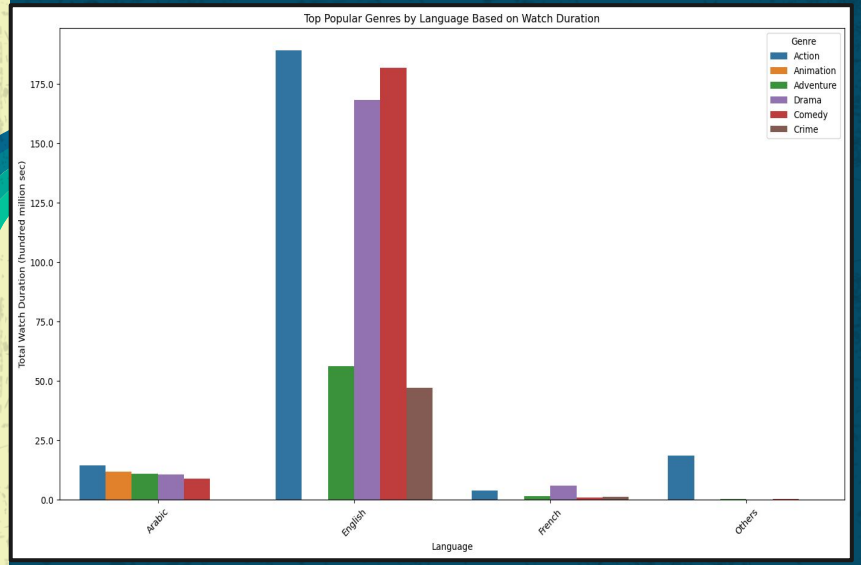
# EDA for Streaming Dataset



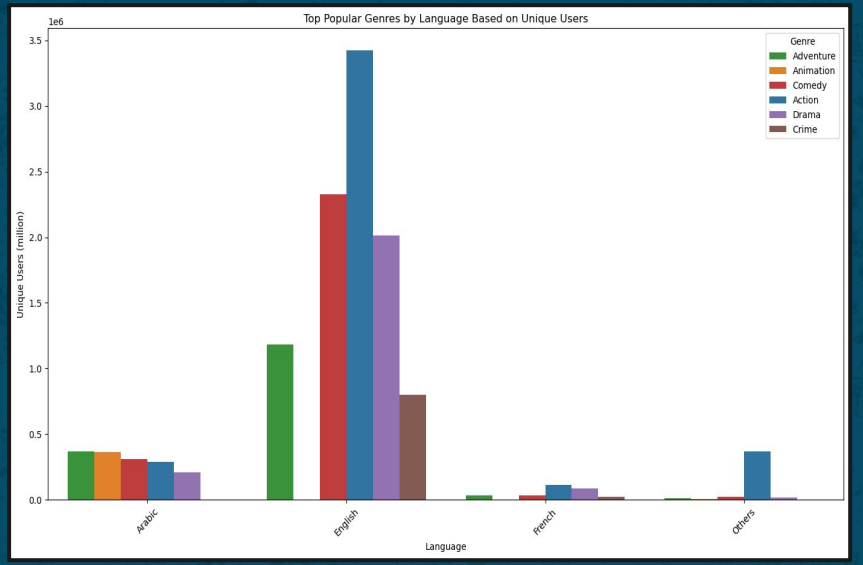

- positive correlation between total_duration_sec and unique_users.
- For high values of total duration, the number of unique users varies widely.
- This  Implies that while longer content has the potential to engage more users, there are other factors at play that determine whether it actually does.

# Visual Insights into Categorical Factors

## Watch Duration




## Unique Users



# Description of our Analysis Methods

**Regression & Prediction:**

- Media Consumption Dataset: Variable of interest is **unique_users**
- Identifying most significant factors to **maximize unique_users**
- Fitting LM and GLM regression models based on these factors
- Clustering Analysis for user engagement pattern

**LDA/LLM:**

- IMDb Dataset: Use LDA for topic modeling → automatically detect topics based on the words that appear in a review
- Can use this to explore sub-genres that are more specific than the overall listing
  - Ex: Christmas movies can be under Comedy, Disney, Romance, etc.,
- **Determine the genre associated with a time period and a positive review**

# Justification of our Analysis Methods

**Regression & Prediction:**

- Modelling and predicting User engagement..
- Identifying significant factors that affect our response, unique_users
- Multivariate Capability
- Forecasting and Validation

**LDA/LLM:**

- LDA:
  - Create SUB-"genres" that are not listed under main genres.
    - Ex: christmas movies during december are not seen through the consumption dataset, because they just fall under Disney/Comedy/Romance (since they don't have their own subgenre)
  -
- LLM:
  - More precise sentiment analysis on reviews to group it into good and bad reviews.
    - Identify use of untraditional words that correlate to good and bad feelings like "lit" or "wack"
    - Identify if users are watching for enjoyment and not for "trends"

# Identifying Significant Factors

Fit GLM model and isolated top variables with **significant p-value < 0.001**:

- consumption_month
- consumption_type
- consumption_language
- consumption_genre

# Verification of Method Assumptions

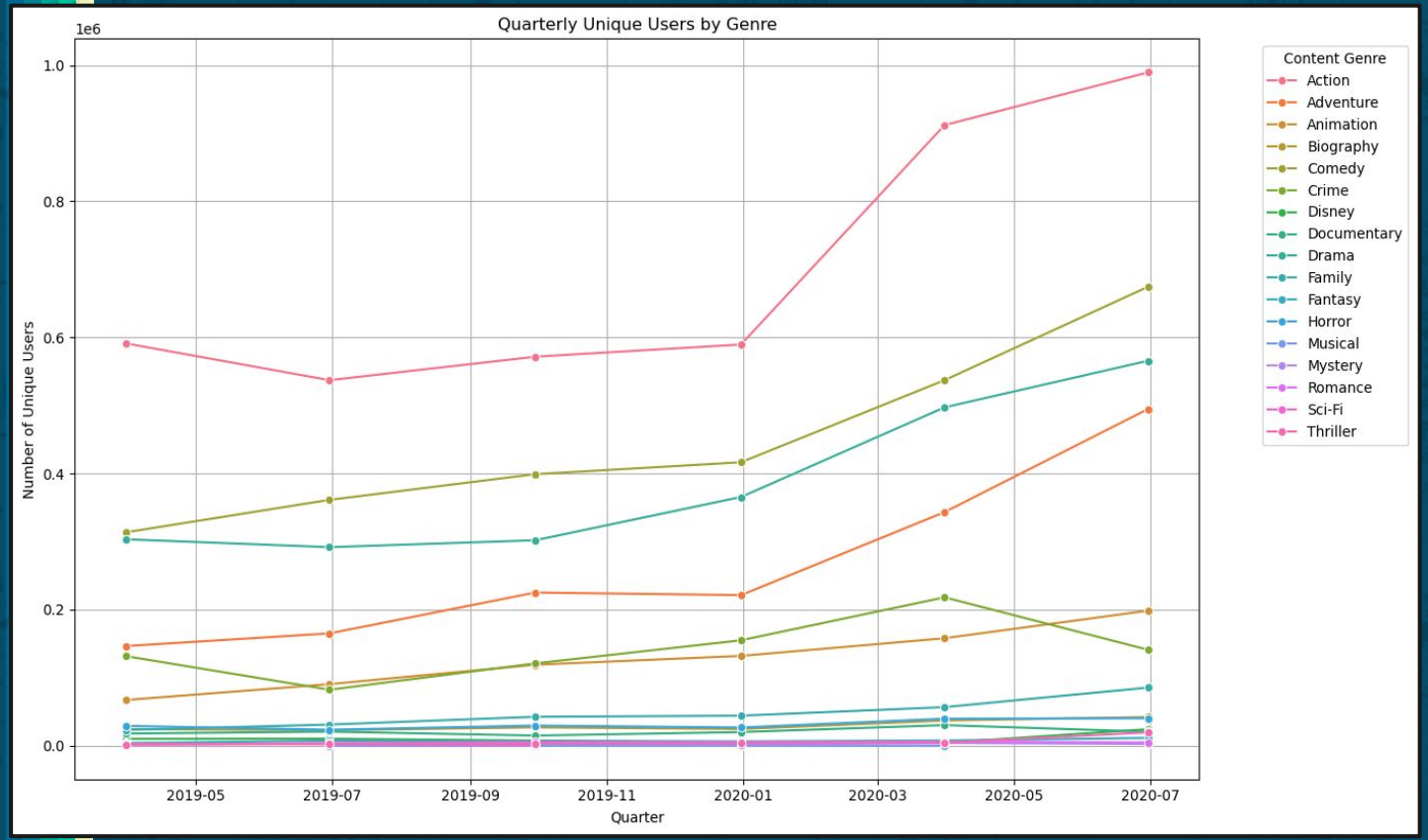
Regression Diagnostics:

| Model | R-squared value |
| --- | --- |
| unique_users ~ total_duration_sec + content_genre | Multiple R-squared: 0.8189<br>Adjusted R-squared: 0.8188 |
| unique_users ~ total_duration_sec + content_type + content_genre | Multiple R-squared: 0.8209<br>Adjusted R-squared: 0.8208 |
| unique_users ~ total_duration_sec + content_genre + content_language | Multiple R-squared: 0.8466<br>Adjusted R-squared: 0.8465 |
| unique_users ~ total_duration_sec + content_genre + content_language + content_type | Multiple R-squared: 0.849<br>Adjusted R-squared: 0.8489 |
| unique_users ~ total_duration_sec + content_genre + content_language + consumption_month | Multiple R-squared: 0.8504<br>Adjusted R-squared: 0.8502 |

# Final Results of Regression Analysis and Modelling

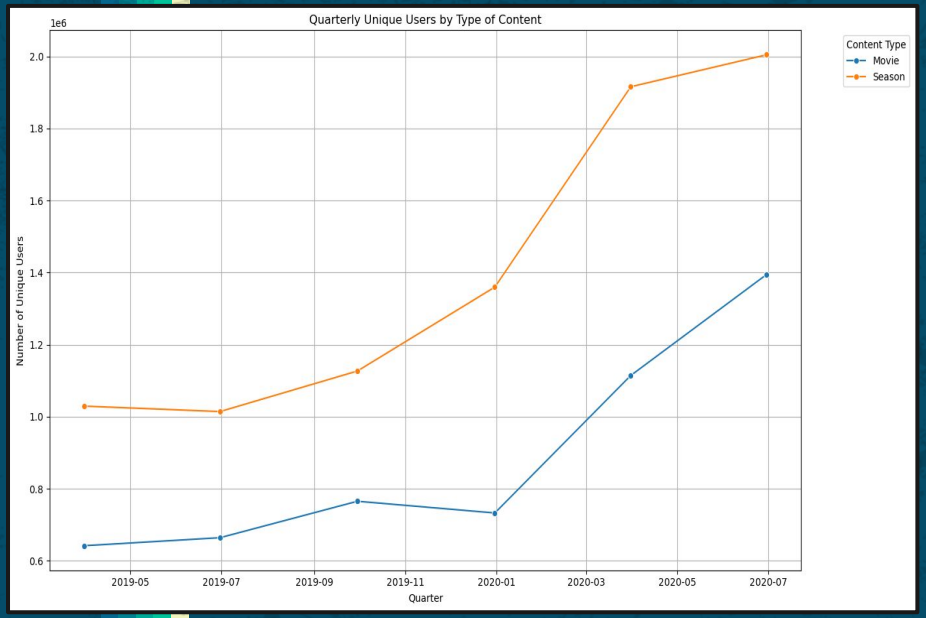
- consumption_month, consumption_type, consumption_language, consumption_genre
- Consumption genre has ***least p-value*** among these
- Prediction analysis and visualizations confirm these findings

# Time Series Analysis

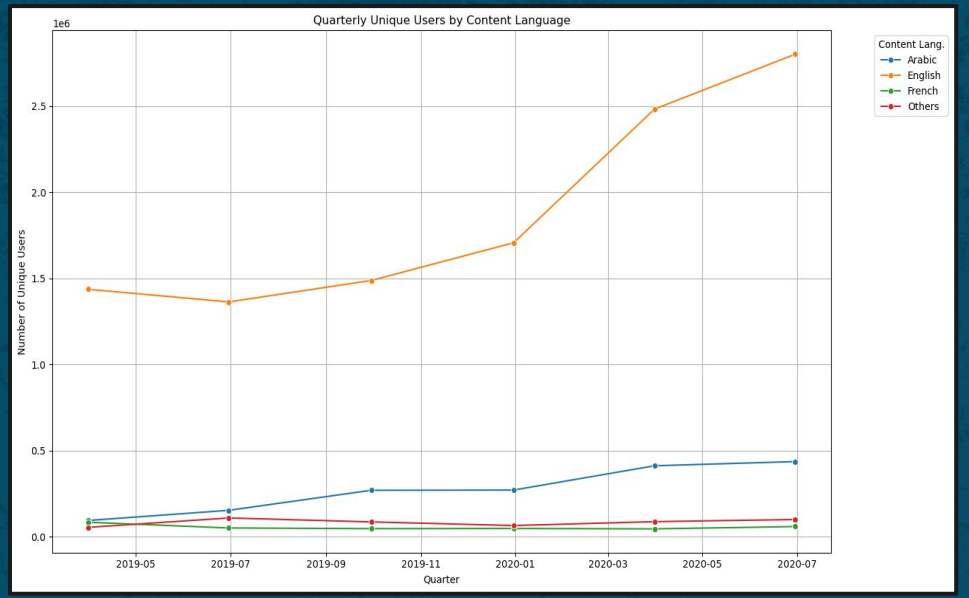



- Most popular genre is Action over time
- ***Significant increase*** in content consumption can be seen after start of Covid-19 Pandemic in March 2020

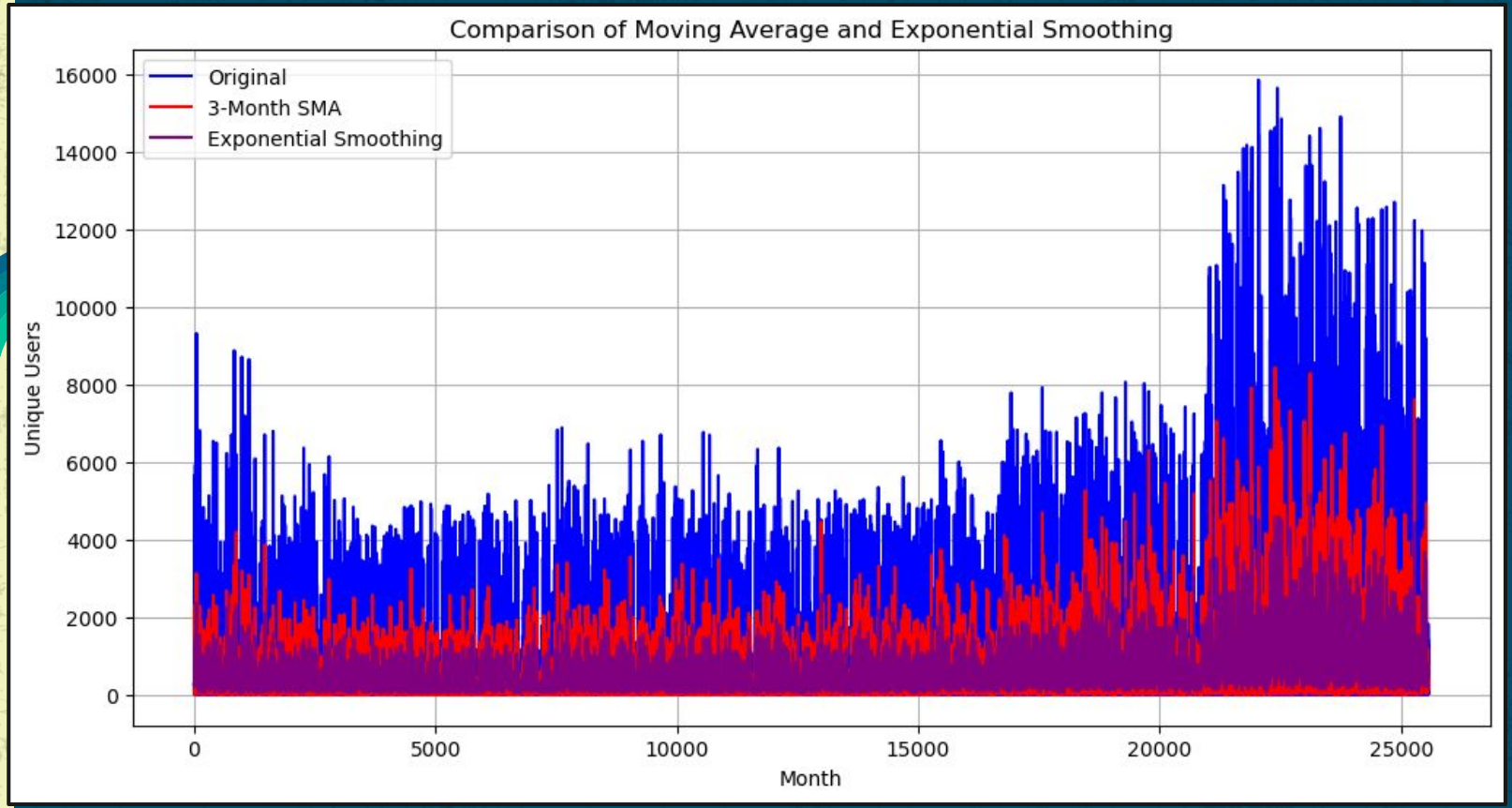
# Content Type



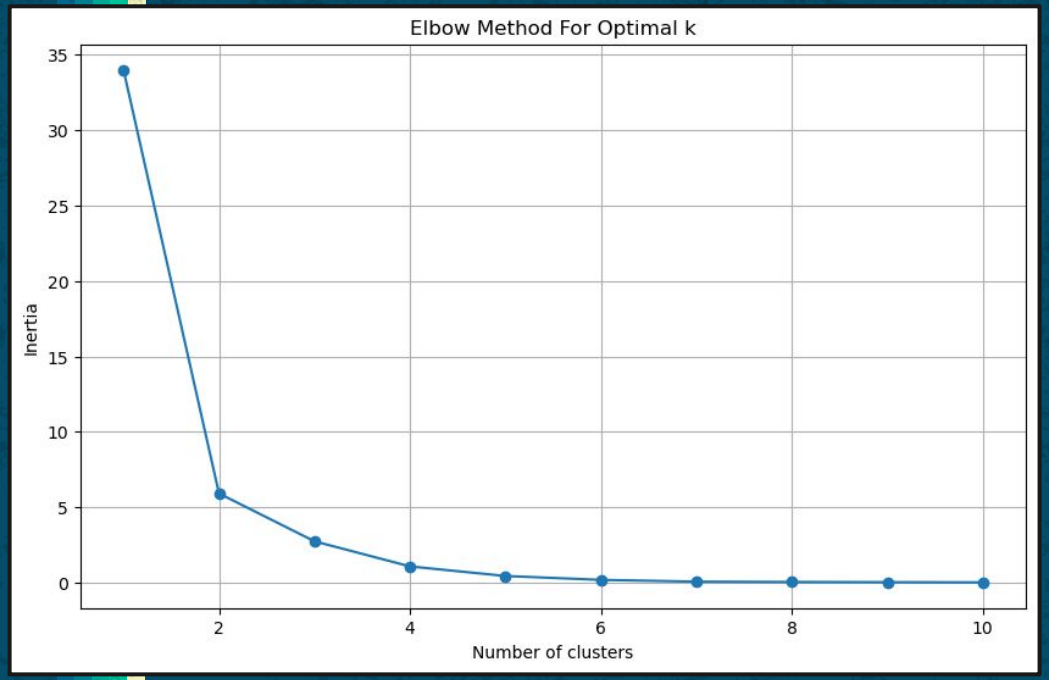

# Content Language



# Model fitting and forecasting

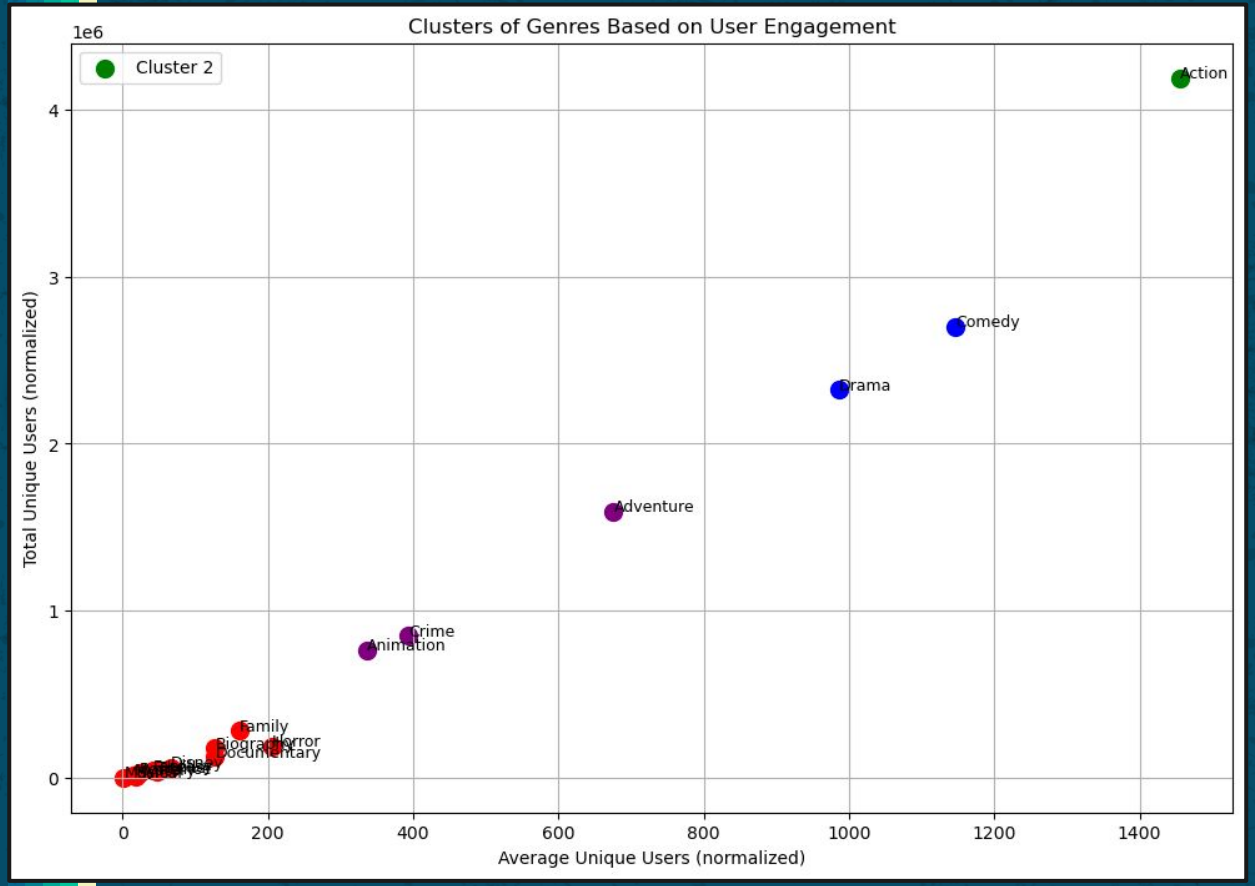


# Interpretation of Prediction Results




| | Genre | Average_Unique_Users | Total_Unique_Users | Cluster |
|---|---|---|---|---|
| 16 | Thriller | 47.663043 | 35080 | 0 |
| 14 | Romance | 23.578302 | 24993 | 0 |
| 13 | Mystery | 15.165029 | 15438 | 0 |
| 3 | Biography | 127.165706 | 176506 | 0 |
| 12 | Musical | 1.606061 | 106 | 0 |
| 11 | Horror | 205.463308 | 187588 | 0 |
| 6 | Disney | 66.049073 | 60567 | 0 |
| 7 | Documentary | 127.075335 | 123136 | 0 |
| 15 | Sci-Fi | 18.661597 | 4908 | 0 |
| 9 | Family | 161.191648 | 281763 | 0 |
| 10 | Fantasy | 42.231237 | 41640 | 0 |
| 8 | Drama | 986.262622 | 2324621 | 1 |
| 4 | Comedy | 1146.617148 | 2701430 | 1 |
| 0 | Action | 1455.614102 | 4190713 | 2 |
| 2 | Animation | 335.565466 | 763747 | 3 |
| 1 | Adventure | 675.825847 | 1594949 | 3 |
| 5 | Crime | 393.336427 | 847640 | 3 |

# Cluster Analysis Results




- Each point represents a genre, plotted according to its normalized average and total unique users.
- Different colors represent different clusters, showing which genres have similar patterns of user engagement.

# Final Results of STAT/ML Methods + Interpretation

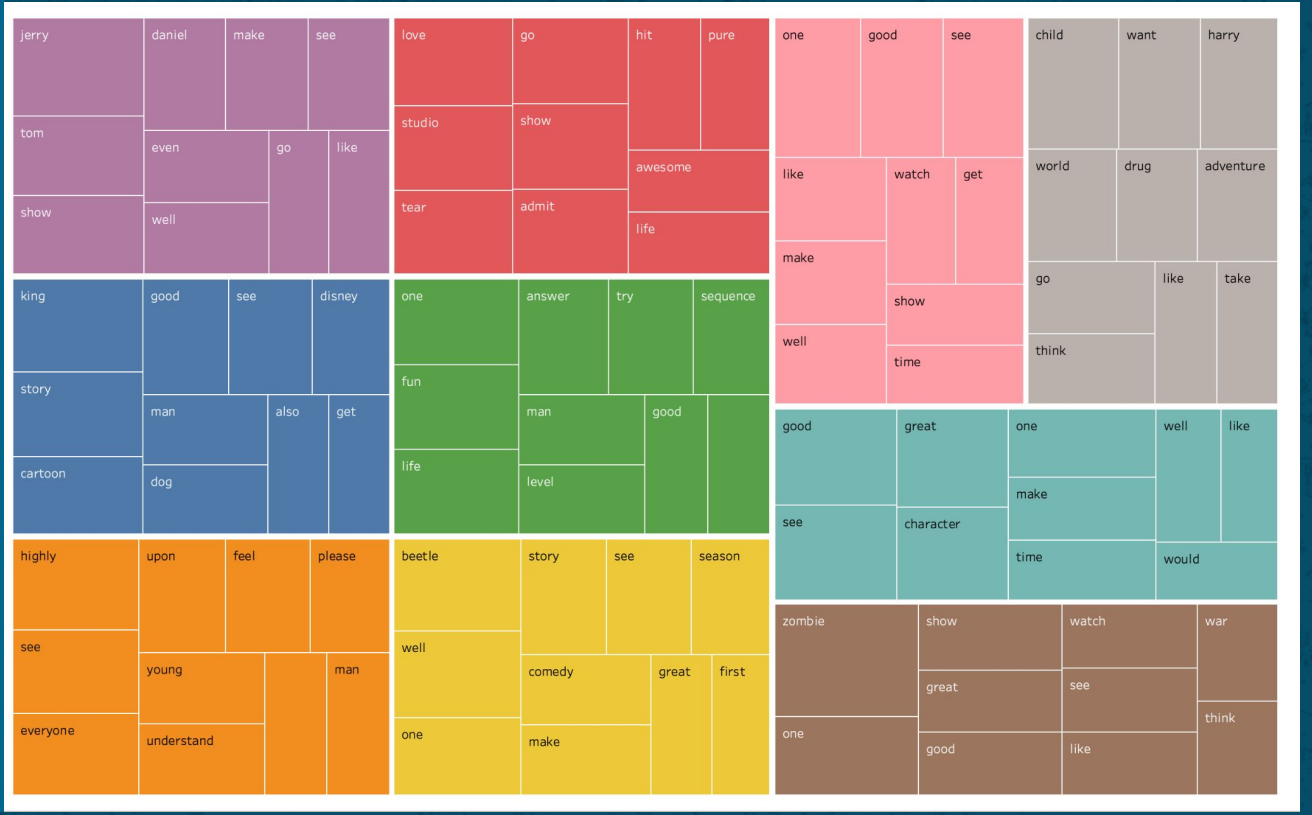
LDA Analysis with Training Set

- Results using 1300 reviews
- Fed our starting genres (adventure, comedy) to see if they tend to have positive vs. negative reviews

**Interpretation:**

- Can see that comedy (Yellow) also has “great”, “well”
- Adventure (grey) has “like”, “want”
- Disney (blue) has “good”

**Need to do some fine tuning and include as many reviews as possible.**



# Conclusion

**Final Results & Major Takeaways:**

- We learned that there was a big spike in viewership and media consumption when the pandemic hit (January - March 2020), and continued to grow overall until mid-2020 (end of data)
- We found that Genre followed by Language were the most significant predictors
- Top 5 Genres during this period were Action, Comedy, Drama, Adventure, and Crime
- Top Languages were English, Arabic, and French
- English Action had the highest user engagement
- LDAs can give both general and specific results based on what we feed it.
- Able to see if certain genres tend to get positive reviews vs negative reviews.

# Conclusion

**Recommendations for Production Houses:**

- Investing and creating **action-centric content** as Action genre tends to have the most unique viewers
- Focusing on **specific periods/seasons** for releasing their content on streaming platforms, such as family movies in the Thanksgiving - Christmas seasons, to **maximize viewership**

**Future Scope:**

- Expand results to look at after 2020 (Netflix yearly reports), want to see whether the trends continued once the quarantine period was over
- Analyze these trends **globally** beyond the Middle East to see if there are similarities and differences

# Thank You