Applied Data Science: Milestone 4 Project Report

Team 1: MovieLens Dataset

Soumen Mohanty, Zaeem Shahzad, Ritin Malhotra, Giorgi Kituashvili

Abstract:
This report presents the findings from an applied data science project focused on the MovieLens 1M dataset, aimed at unraveling the intricate relationships between movie preferences and demographic factors such as occupation, age, and gender. Through comprehensive data preparation, hypothesis formulation, exploratory data analysis, and predictive modeling, we sought to understand how these demographic characteristics influence genre preferences among viewers. Our analysis revealed significant correlations, particularly between occupation and movie genre preferences, highlighting nuanced patterns in how different occupational groups engage with movie genres. Advanced statistical methods, including ANOVA and correlation analysis, were employed to substantiate these findings. The project culminates in a predictive model that leverages these insights, offering potential applications in targeted marketing and personalized content recommendation in the movie industry. The implications of our findings extend to enhancing user experience on streaming platforms and guiding content creation decisions. This report distills these complex analyses into a coherent narrative, making the results accessible to a non-technical audience and demonstrating the power of data storytelling in data science.

In the ever-evolving landscape of the movie industry, understanding audience preferences has become increasingly crucial for filmmakers, streaming services, and marketers. With the rise of data-driven strategies, the ability to predict and cater to viewer tastes can significantly impact content creation and distribution. This project delves into the MovieLens 1M dataset to explore the complex interplay between viewer demographics - specifically occupation, age, and gender - and their movie preferences. Our primary objective is to investigate whether these demographic factors correlate with preferences for certain movie genres, thereby providing actionable insights for targeted marketing and personalized content recommendations.

The problem statement at the heart of this study posits that an individual's occupation, along with their age and gender, might influence their movie genre preferences. This hypothesis is grounded in the notion that an individual's professional background could reflect their interests and, by extension, their choice of movies. For instance, it's conceivable that individuals in creative professions might exhibit a predilection for genres like drama or art films, whereas those in

technical fields might prefer genres such as sci-fi or action. Similarly, age and gender might also play pivotal roles in shaping movie preferences, influenced by cultural, social, and psychological factors.

Existing literature in consumer behavior and media studies — such as studies on how popular streaming service Netflix recommends movies to its users (Chong, Steck et al.) — provides some evidence supporting these assumptions. However, there remains a gap in applying these insights specifically to the realm of movie preferences at a granular level of demographic categorization. By leveraging advanced data analysis techniques, including exploratory data analysis, statistical testing, and predictive modeling, this study aims to fill this gap. Our findings reveal intriguing patterns and correlations, offering a deeper understanding of audience preferences that can be leveraged in the movie industry for more nuanced and effective audience targeting. This report presents these findings and discusses their broader implications, weaving together the threads of data analysis into a coherent narrative accessible to both technical and non-technical audiences.

Curating and Exploring the MovieLens Dataset:

The MovieLens 1M dataset forms the bedrock of our study, offering a rich tableau of 1,000,209 anonymous movie ratings. This dataset, pertaining to approximately 3,900 movies reviewed by 6040 viewers, was collected by the University of Minnesota's GroupLens Research Project in 2000-2003.

The MovieLens dataset, essential for our movie recommendation system analysis, was initially segmented into three distinct files: movies.dat, ratings.dat, and users.dat. To facilitate a comprehensive analysis, these files were merged into a single dataframe:

Movies Dataset: Contained movie details like IDs, titles, and genres.
Ratings Dataset: Included user ratings for movies, with user and movie IDs.
Users Dataset: Held user demographics including age, gender, and occupation.

The merging process was twofold:
First, the ratings_df was merged with users_df using user id as the key, combining user ratings with their demographic details.

Next, this combined dataset was merged with movies_df using movie id, integrating movie details into the mix.

The resultant unified dataframe provided a comprehensive view of user demographics, movie details, and ratings. Some sample rows from our dataset look as follows:

	userid	movieid	rating	rating date	gender	age	zip-code	occupation	age_categories	title	genre	year
0	1	1193	5	2000-12-31	F	1	48067	K-12 student	Under 18	One Flew Over the Cuckoo's Nest	Drama	1975
1	2	1193	5	2000-12-31	M	56	70072	self-employed	56+	One Flew Over the Cuckoo's Nest	Drama	1975
2	12	1193	4	2000-12-30	M	25	32793	programmer	25-34	One Flew Over the Cuckoo's Nest	Drama	1975
3	15	1193	4	2000-12-30	M	25	22903	executive	25-34	One Flew Over the Cuckoo's Nest	Drama	1975
4	17	1193	5	2000-12-30	M	50	95350	educator	50-55	One Flew Over the Cuckoo's Nest	Drama	1975

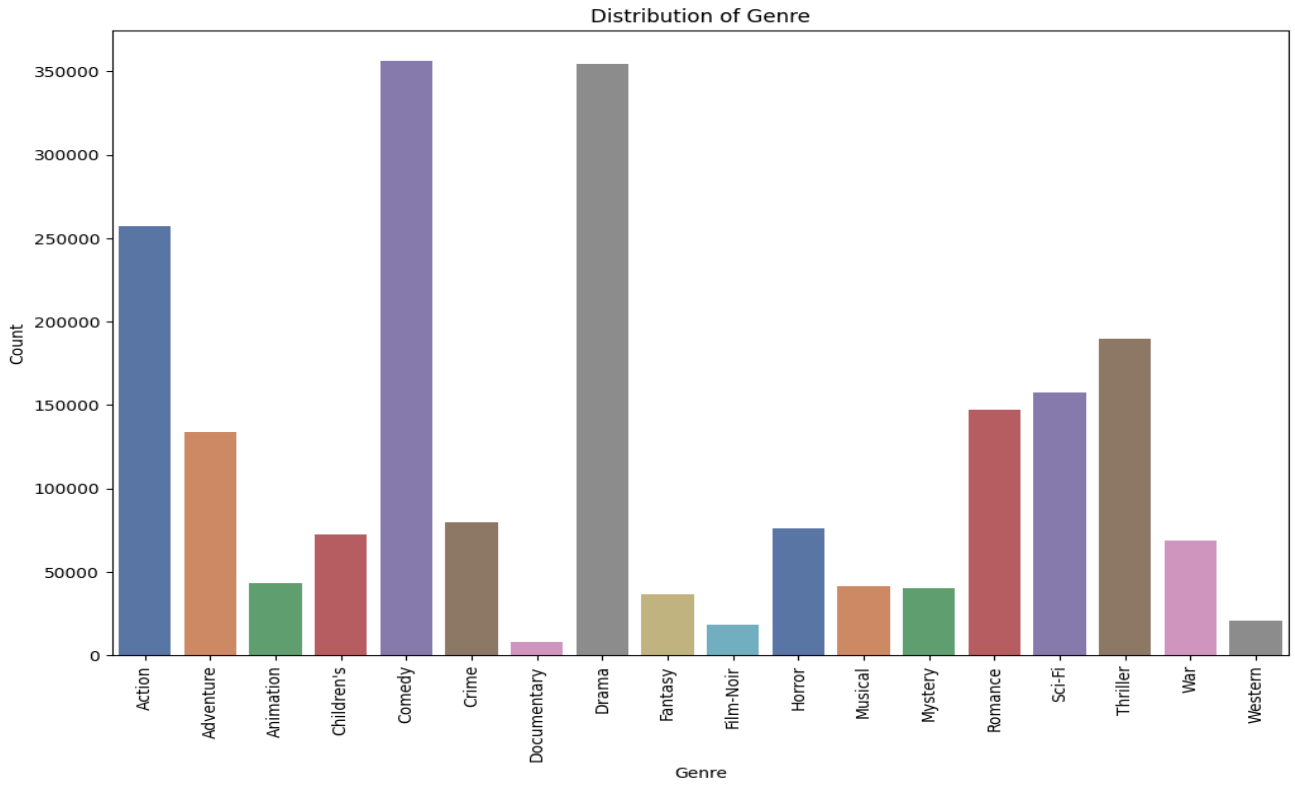
This format was crucial for our analysis, revealing insights into user preferences and behavior in relation to movie genres. The final merged dataset was saved in a CSV file within a cleaned_data folder for easy access in further stages of the project.

Feature Selection and Importance:
Our analysis focused on key variables deemed most informative for our hypothesis testing:

- **Occupation**: Different occupational categories, particularly the representation of 'grad students', were analyzed to explore their influence on genre preferences.
- **Genre**: With 'Comedy' being predominant, we examined how various genres align with occupational preferences.
- **Age Categories**: This demographic segmentation helped us to understand genre preferences across different age groups.
- **Gender:** Different genders may prefer different genres of movies. The abundance of male respondents compared to females might also skew our genre distribution data.
- **Rating**: The distribution of ratings, showing a slight positive skew, was crucial for examining variations in movie preferences among different occupations. This can also potentially act as a “label” for our predictive model.
- **Year of Movie Release**: This contextual element allowed us to assess whether preferences for genres are influenced by the era of the movie's release.

Limitations and Future Considerations:
While our dataset was comprehensive, certain limitations were noted. For instance, the dataset did not include detailed viewer feedback or specific reasons behind ratings, which could have provided deeper insights into user preferences. We also did not have detailed data on viewer interactions — such as genre preferences based on time of day, sudden interests in particular genres — that are utilized by modern streaming platforms like Netflix for their recommendation systems (Ostatnie). Additionally, the dataset's time frame (2000-2003) may not fully reflect current trends in movie consumption. Furthermore, there was a significant imbalance in the dataset with a disproportionately higher number of observations for females compared to males. This skew in demographic representation could influence the accuracy and applicability of our findings. Moreover, the distribution of genres within the dataset was uneven, with certain genres

being overrepresented in terms of the number of reviews while others were significantly underrepresented. This disparity in genre representation could potentially bias our analysis and limit the generalizability of our conclusions to all genres of movies. Future studies might benefit from incorporating more recent data and detailed user feedback and interaction data for a more nuanced understanding of genre preferences across demographics.

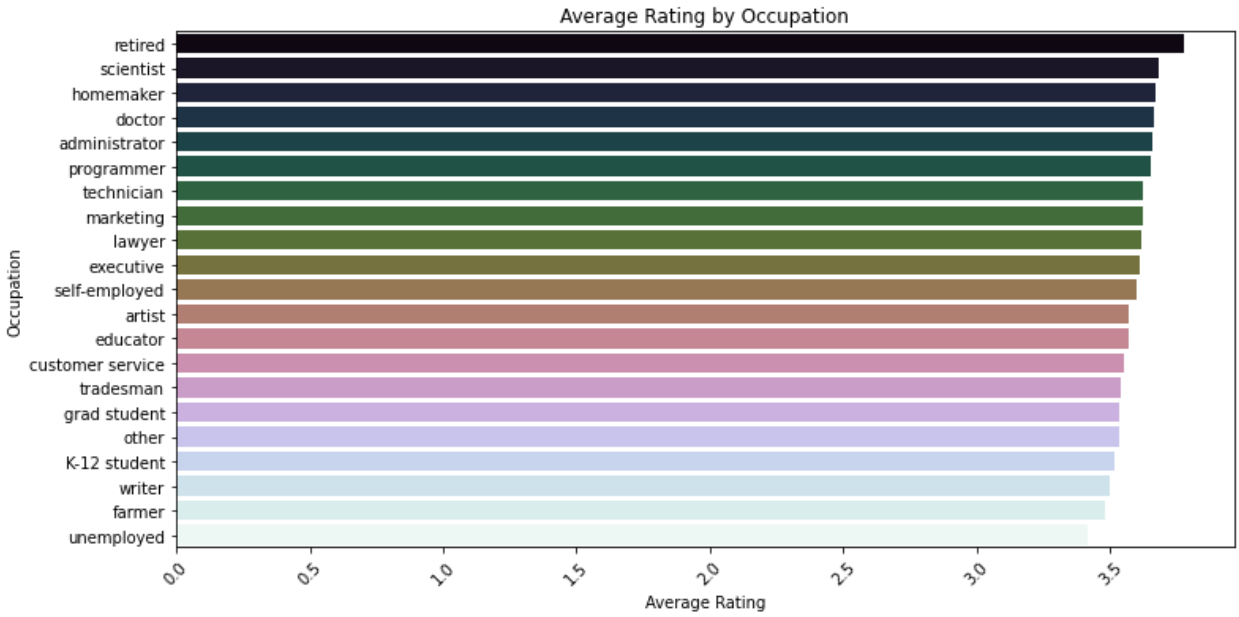


The graph clearly shows that genres such as Drama, Comedy, and Action have a significantly higher number of reviews compared to genres like Western, War, or Musical. This visual evidence supports the statement that there is an uneven representation of genres in the dataset, which underscores the potential for a biased analysis and suggests that the findings may not be uniformly applicable across all movie genres.

Distributions and Features of Interest:

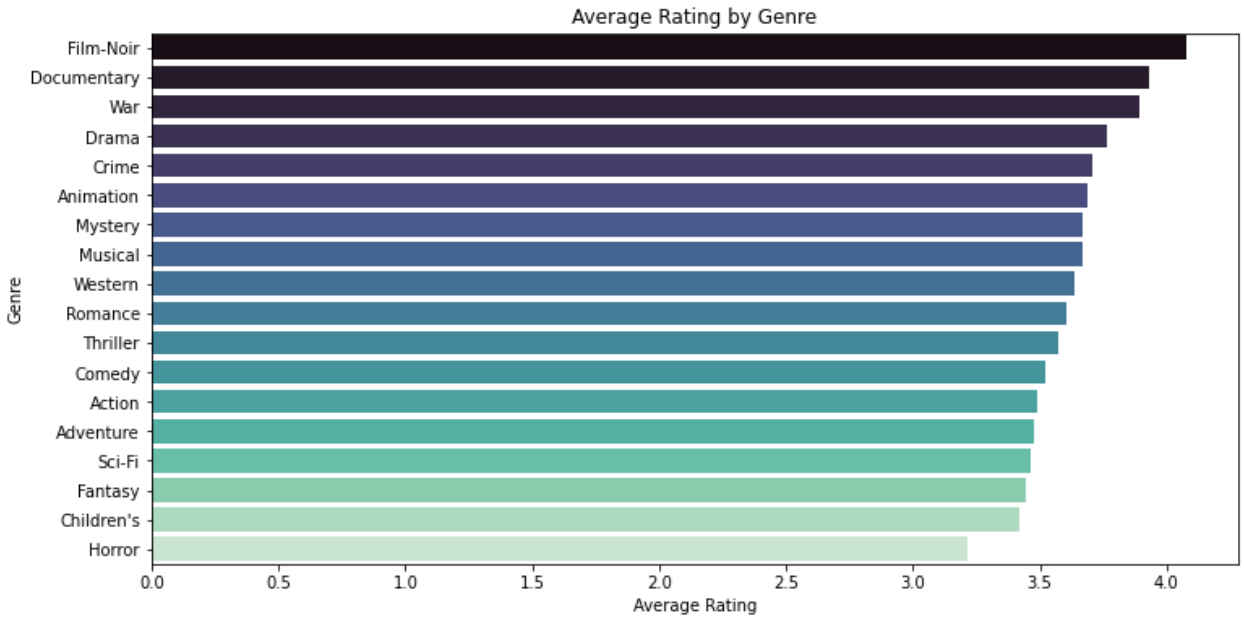
Given our understanding of the dataset, we can proceed to explore the underlying patterns and tendencies within the MovieLens dataset that may inform our understanding of viewer preferences. This section delves into how specific features, such as occupation and genre, distribute across our dataset and influence movie ratings, providing the empirical groundwork to test our hypotheses regarding the impact of occupation on genre preference.

Figure 1: Average Rating by Occupation



This visualization illustrates the average movie ratings distributed by the occupation of the raters. Although the differences in average ratings among occupations are not drastic, the graph reveals that retirees and scientists tend to rate movies more generously than other groups such as writers, farmers, and the unemployed. This finding is intriguing as it suggests that personal interests and lifestyle, which can be associated with one's occupation, might subtly influence their ratings. While the overall variance is small, these differences could provide valuable insights for a predictive model, potentially affirming the alternative hypothesis (H_a) that occupation correlates with genre preferences. It also raises questions about the influence of occupational identity on the perception and enjoyment of media, which could be critical for targeted marketing and content creation strategies.

Figure 2: Average Rating by Genre



The second visualization focuses on the average ratings that different movie genres receive. It highlights that genres such as film noir, documentary, and war have the highest average ratings, despite their lower popularity compared to genres like comedy or drama. This trend could suggest that these less popular genres are often rated by a niche audience that may have a particular affinity for them, leading to higher average ratings. The visualization supports the idea that people may self-select the genres they are inclined to watch and rate, indicating a bias that is worth considering in predictive modeling. The high ratings in these genres could be a result of a concentrated audience of enthusiasts, which might be a significant factor for the alternative hypothesis, suggesting that there is a correlation between the genre and the audience's occupation.

Together, these visualizations contribute to a nuanced understanding of the relationship between raters' occupations and their movie preferences. They form an essential part of the narrative that leads us from the general occupation-based preferences to a more detailed genre-specific preference, providing a data-driven foundation to test our hypotheses.

Continuing from the informed observations derived from our visual data analysis, we embark on a statistical examination to quantify the relationships between demographic factors and movie ratings / genre preferences. Our correlation analysis, although indicating only weak associations between variables like gender, age, occupation, and genre, underscores the subtleties of our dataset. These findings, particularly the minimal correlations, sharpen our focus towards the potential influence of occupational backgrounds on genre preference, thus laying a statistical foundation for ANOVA testing. The ANOVA will further illuminate the nuances of our hypothesis, assessing the strength of occupation as a predictor for genre preference in movies.

```
ANOVA for K-12 student: F-Value: 57.009915935443, P-Value: 2.8640057550011396e-193
ANOVA for self-employed: F-Value: 115.59300497953507, P-Value: 0.0
ANOVA for programmer: F-Value: 120.34768208775675, P-Value: 0.0
ANOVA for executive: F-Value: 221.28225959637612, P-Value: 0.0
ANOVA for educator: F-Value: 250.09123608596312, P-Value: 0.0
ANOVA for administrator: F-Value: 56.2091889367738, P-Value: 8.318934383982504e-191
ANOVA for grad student: F-Value: 304.02599479800665, P-Value: 0.0
ANOVA for farmer: F-Value: 5.720454268720781, P-Value: 3.807369266901621e-13
ANOVA for technician: F-Value: 162.84356697805404, P-Value: 0.0
ANOVA for other: F-Value: 337.4632543377334, P-Value: 0.0
ANOVA for artist: F-Value: 100.2085253242381, P-Value: 0.0
ANOVA for homemaker: F-Value: 18.291862386737996, P-Value: 1.52669660135494e-55
ANOVA for unemployed: F-Value: 36.907804901837835, P-Value: 4.400925431812419e-121
ANOVA for tradesman: F-Value: 25.113945140450937, P-Value: 2.2529491149582612e-79
ANOVA for scientist: F-Value: 51.10269745027825, P-Value: 3.8082524882058005e-172
ANOVA for lawyer: F-Value: 45.949768085225465, P-Value: 1.45433612558538e-153
ANOVA for writer: F-Value: 143.3822789684093, P-Value: 0.0
ANOVA for retired: F-Value: 46.311816037201105, P-Value: 4.521454893566878e-154
ANOVA for customer service: F-Value: 38.52046075027297, P-Value: 3.0244530581095788e-127
ANOVA for marketing: F-Value: 93.77266300252542, P-Value: 0.0
ANOVA for doctor: F-Value: 102.41731467645317, P-Value: 0.0
```

Building on the visual insights and preliminary correlation analyses, our statistical journey brought us to the application of ANOVA tests across various occupational groups. The findings were striking—high F-values and extremely low P-values across the board signal that raters' occupations indeed correlate significantly with their preferences for movie genres. For instance, the pronounced rating patterns among programmers, executives, and writers strongly suggest that different occupations have distinct tastes in films, which challenges the null hypothesis of no correlation. These compelling ANOVA results weave into our narrative by supporting the alternative hypothesis and potentially guiding targeted marketing and content creation within the movie industry. They affirm the importance of occupation as a substantial factor in understanding movie preferences, setting a firm statistical ground for the predictive model and our study's conclusion.

The Predictive Model

To curate predictive models to align recommendations with user preferences, we delved deep into the realm of foresight, leveraging our dataset's rich demographic and rating information to anticipate user preferences. This section unfurls the methodologies and reasoning behind two distinct predictive models tailored for different stages of user interaction.

1. **Addressing the Cold Start Problem with a Multi-Label Classification Model**:

In tackling the cold start problem, where new users have yet to express their movie preferences through ratings, we developed a sophisticated multi-label classification model. This model is built on the foundation that demographic attributes – specifically age, occupation, and gender – can be predictive of a user's genre preferences.

The model's architecture is designed to recognize the multifaceted nature of individual tastes: users often enjoy a variety of genres rather than a single type. By adopting a multi-label approach, we enable the model to predict a range of genres that align with each user's unique demographic profile. This is a significant enhancement over traditional single-label models, which may oversimplify user preferences.

Key Aspects of the Multi-Label Model:

- **Data Refinement**: We focus exclusively on higher ratings (4 or above), ensuring that the model learns from genres that users genuinely enjoy.
- **Demographic-Based Grouping**: The model groups data by age, occupation, and gender, acknowledging the diverse influences these factors have on movie preferences.
- **Multi-Label Encoding**: Utilizing the MultiLabelBinarizer, we effectively handle multiple genres per user, transforming the problem into a multi-label classification scenario.
- **Algorithm Selection**: The model employs a OneVsRestClassifier paired with LogisticRegression, an optimal choice for multi-label classification tasks.
- **Evaluation Metrics**: We assess model performance using accuracy and F1 score metrics, both of which are crucial for evaluating multi-label models.

By incorporating these enhancements, our model becomes a more accurate and nuanced tool for genre recommendation, aptly catering to the diverse tastes of new users. It serves as a dynamic entry point into our platform, guiding users through the vast array of movies by aligning recommendations with their demographic characteristics.

2. **Beyond the Cold Start – The Content-Based Filtering Approach:**

Acknowledging the efficacy of content-based filtering models in real-world applications, we advanced to address the needs of existing users (who have rated at least 20 movies as per our dataset) through a collaborative filtering model. This model, grounded in the principles of collaborative filtering, predicts user preferences based on the collective patterns observed in user-item interactions. Collaborative filtering, uniquely, does not require metadata about the items but instead relies on the aggregate behaviors of users.

In our approach, we employed Singular Value Decomposition (SVD), a powerful matrix factorization technique rooted in linear algebra. SVD decomposes the user-item rating matrix into three key components: the User Matrix (U), representing the relationships between users and latent factors; the Singular Values (Σ), a diagonal matrix indicating the strength of each latent factor; and the Item Matrix (V^T), encapsulating the relationship between items and latent

factors. These matrices unveil latent structures in user-item interactions, such as shared preferences and behavioral patterns.

Our model bifurcates collaborative filtering into two strategies: User-Based, focusing on similarities among users, and Item-Based, emphasizing item similarities derived from user ratings. By utilizing SVD within this framework, we unraveled latent factors that capture the essence of user-item relationships, crucial for predicting and personalizing user preferences.

Methodologically, we constructed a user-item interaction matrix from user ratings, treating unrated items as zeros to signify non-interaction. To enhance computational efficiency, we converted this matrix into a sparse format. The model was developed using the TruncatedSVD class from sklearn, with the number of latent factors set at fifteen. This allowed the decomposition of the user-item matrix into a format conducive to predicting ratings for all movies, thereby enabling us to curate personalized movie recommendations, particularly prioritizing films with ratings of 4 or above.

This sophisticated SVD-based collaborative filtering model represents an evolution in our predictive capabilities, offering existing users highly personalized movie suggestions aligned with their historical preferences and unraveled latent interests.

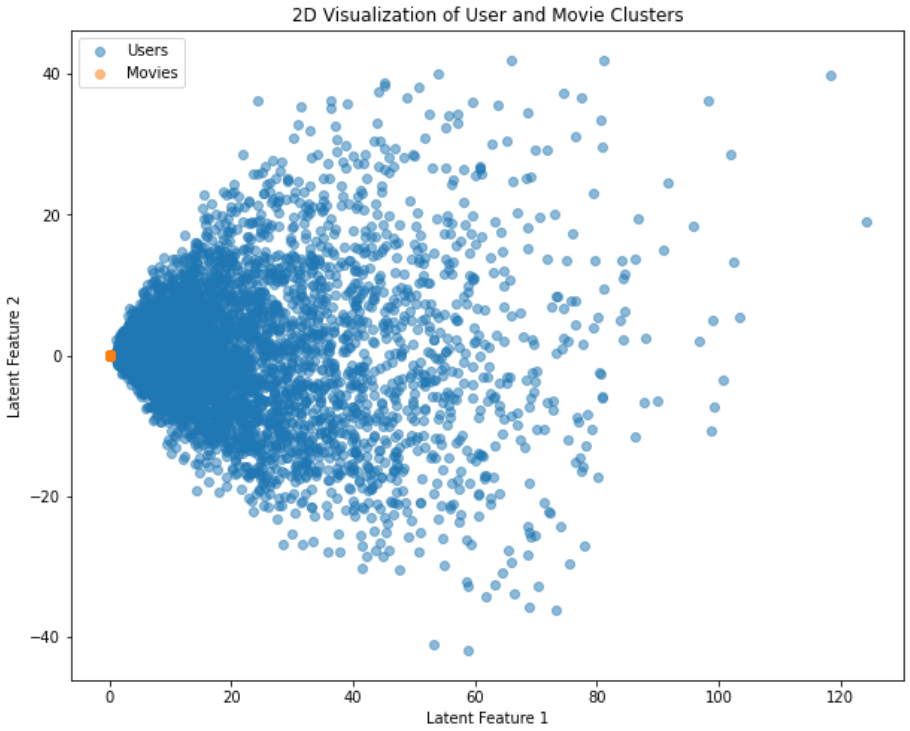
Results of our SVD-based Approach

Our validation process compared the user's historical preference for highly-rated movies against the recommended list. The congruence of genres within these two sets corroborated the accuracy of our model's suggestions.

```
Top 10 Movie Recommendations for User ID 23 :
Star Wars: Episode IV - A New Hope  - {'Action', 'Sci-Fi', 'Adventure',
'Fantasy'}
Total Recall  - {'Action', 'Sci-Fi', 'Adventure', 'Thriller'}
Terminator 2: Judgment Day  - {'Action', 'Sci-Fi', 'Thriller'}
Matrix, The  - {'Action', 'Sci-Fi', 'Thriller'}
Men in Black  - {'Comedy', 'Action', 'Sci-Fi', 'Adventure'}
Alien  - {'Action', 'Sci-Fi', 'Thriller', 'Horror'}
Terminator, The  - {'Action', 'Sci-Fi', 'Thriller'}
Twelve Monkeys  - {'Drama', 'Sci-Fi'}
Aliens  - {'Action', 'Sci-Fi', 'Thriller', 'War'}
Blade Runner  - {'Sci-Fi', 'Film-Noir'}
```

```
Genres rated 4 or higher by User ID 23 :
Sci-Fi       56
Action       50
Comedy       42
Thriller     36
Drama        29
Adventure    26
```

```
Name: genre, dtype: int64
```



The visualizations illustrated the clusters formed by users and movies within the latent feature space. Despite the complexity of multi-dimensional data, these visualizations offered an intuitive understanding of the model's recommendation mechanics.

The Bottom Line

In our project, we explored two distinct approaches to build movie recommender systems. The first is a naive model that conducts multilabel classification for each unique combination of user demographics, aiming to discern their preferred genres. This model underscores the significance of demographic factors in predicting movie preferences. Our second approach delves into collaborative filtering, sidestepping content attributes to focus on the power of user-item interaction data. Throughout our exploratory data analysis (EDA) and statistical assessments, we've uncovered and leveraged inherent correlations within the data, reinforcing the validity of collaborative filtering techniques. These correlations serve as the empirical backbone of our project, substantiating the collaborative filtering model's ability to recommend movies based on user rating patterns without relying on content metadata.

These methodologies, while distinct in application, both illuminate the multifaceted nature of recommendation systems. By first proving the correlations through EDA, we established a concrete base for employing collaborative filtering, which has become a mainstay in industry practices for its effectiveness in capturing the nuanced preferences of users. As we proceed, the insights gained from our EDA will inform the refinement of our models, ensuring they remain attuned to the intricate dynamics of user preferences. The project encapsulates a comprehensive study into the algorithms that define modern recommender systems, setting a foundation for the future exploration of more intricate, personalized recommendation engines in the entertainment sector.

References

Chong, David. “Deep Dive into Netflix’s Recommender System.” *Medium*, 24 Sept. 2021, https://towardsdatascience.com/deep-dive-into-netflixs-recommender-system-341806ae3b48.

Steck et al. *Netflix Research*. https://research.netflix.com/publication/%20Deep%20Learning%20for%20Recommender%20Systems%3A%20A%20Netflix%20Case%20Study.

Ostatnie Wpisy. *Netflix Recommendation System: How It Works | RecoAI*. https://recoai.net/netflix-recommendation-system-how-it-works/.