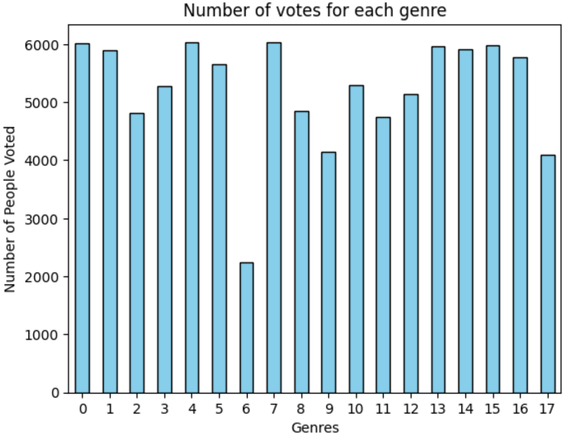
Movie Recommendation System

Course Project

Trupal Patel - IMT2021056
Kevin Adesara - IMT2021070
Anant Ojha - IMT2021102

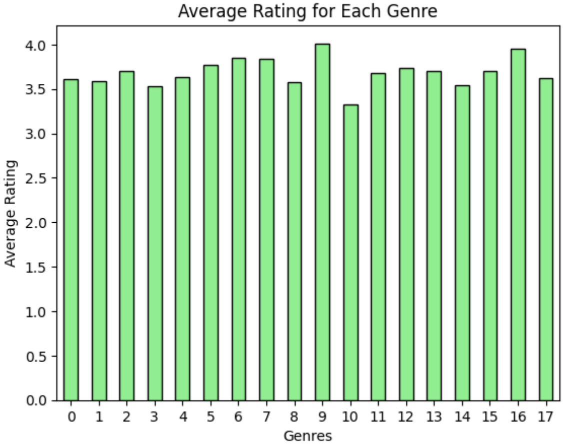
Data Analysis and EDA

1. Examined rating patterns and visualized votes per genre.



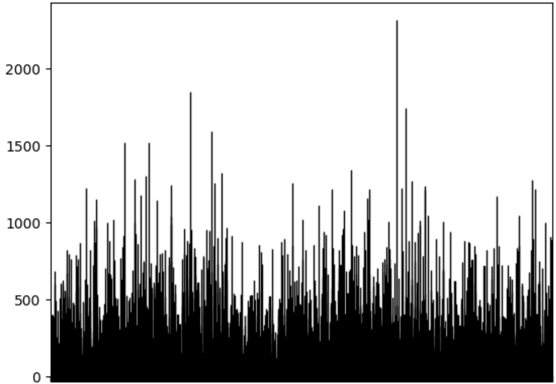
Data Analysis and EDA

2. Calculated genre-wise ratings and presented them in a bar plot.



Data Analysis and EDA

3. Determined the number of movies rated by each user and showcased the distribution using a bar plot.



Data Preprocessing

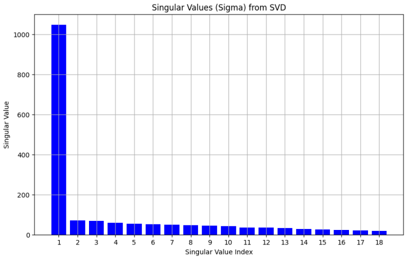
Enhanced data completeness by doing the following methods:

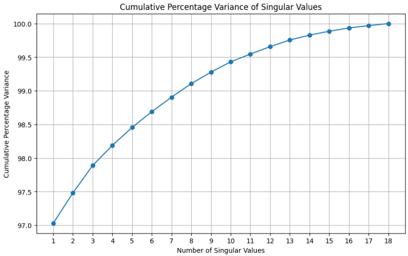
1. Replaced NaN values with a pre-defined number (say 2.5)
2. Replaced NaN values the mean value.
3. Left the NaN values as it is for Collaborative Filtering.

Singular Value Decomposition (SVD)

- SVD is implemented using QR decomposition.
- Performed SVD and reduced SVD on the user-genre matrix.
- Determined the number of singular values using the plot.

Singular Value Decomposition (SVD) - Images





K-means Clustering

- Implemented K-means with K-means++ initialization for enhanced convergence.
- Determined the optimal cluster count using elbow curves and test error plots.
- Introduced a novel recommendation approach: suggesting poorly rated movies from users in the farthest clusters, enhancing diversity in recommendations.

K-means Plots

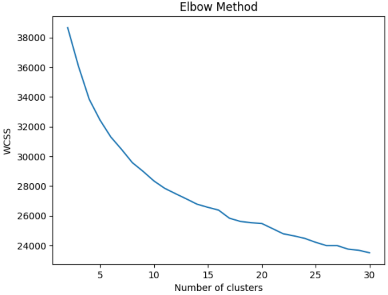


Figure 1: Elbow curve for optimal cluster count

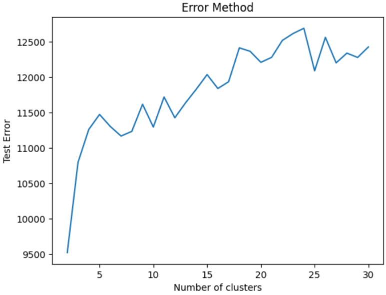


Figure 2: Test error plot for different cluster counts

Recommendation Algorithm

- Created a user-genre matrix to represent user preferences for different genres.
- Applied Singular Value Decomposition (SVD) to capture latent factors in the user-genre matrix.
- Utilized K-means clustering on the SVD outputs to group users with similar preferences into clusters.
- For a new user, calculated its user representation using the SVD outputs and determined the corresponding cluster.
- Within the identified cluster, employed Pearson similarity to identify the most compatible genres for the new user.
- Selected the top 5 movies from the top 3 genres as personalized recommendations.

Collaborative Filtering

- Utilized Pearson correlation to measure similarity between users.
- Identified users with similar preferences based on ratings for common items.
- Identified items rated by users with similar preferences for the target user.
- Calculated weighted average scores for recommended items.
- Ranked the recommended items based on the calculated scores.

Compare and Contrast : Neighborhood-based Collaborative Filtering

1. Uses Pearson correlation adjusted by IDF to measure user similarities directly.
2. Direct user comparisons can be computationally intensive.
3. Struggles with scalability and sparsity due to storing all pairwise similarities.

SVD-K-means Approach

1. Employs Singular Value Decomposition (SVD) to reduce data dimensionality.
2. Identifies user clusters using K-means, focusing on relevant users.
3. Incorporates localized similarity measures, refining predictions efficiently.

Novel Ideas Implemented

- Introduced a novel approach by assigning weights to Pearson similarity based on the age differences between users, making it more personalized.
- Further innovated by incorporating an alternate Pearson similarity with weights determined by occupation differences, offering a diversified perspective.
- Implemented a unique strategy for new users: recommended movies that were rated poorly by users in the farthest cluster from the new user.