PROFITABILITY ANALYSIS AND CLUSTERING OF MOVIES

Di Kang, Estefania Padilla Gonzalez, Fang Li
Jiuqin Wei, Muhammad Usman Taj

Overview

- Movie database analysis
- Producers always emphasize on profits
- Apply ML algorithms to analyze if certain features can predict profitability
- Identify clusters of similar movies based on profit (Revenue over Budget ratio)
- Methods used: K-means

Data

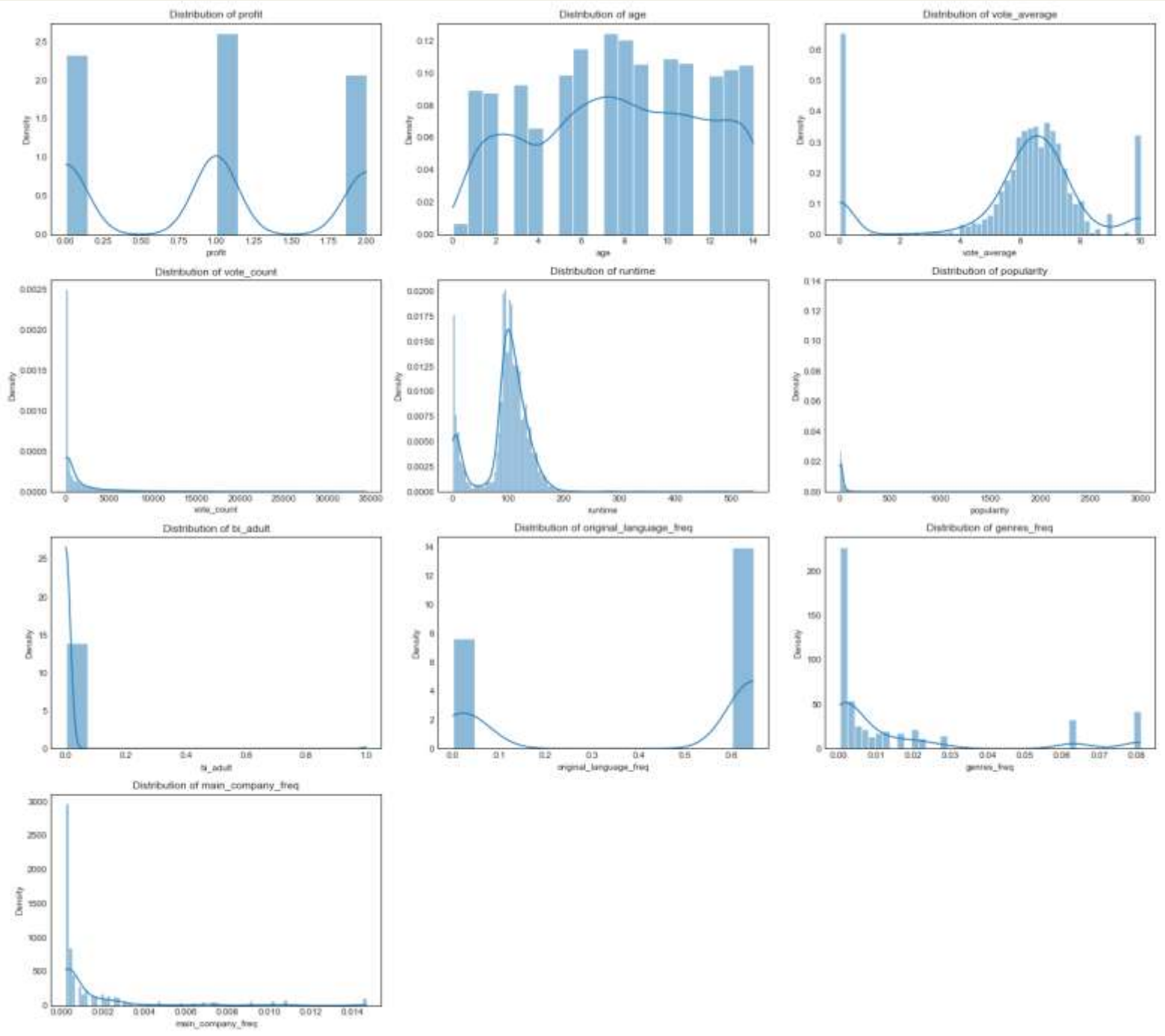
- The TMDB (The Movie Database) is a comprehensive database
- Includes information of around 1,000,000 movies
- Most movies released in the last 40 years
- Constitute important features such as release date, budget, revenue, production company, production language, vote average, genre, etc.

Data Preprocessing

- Cleaning of raw data (dropping missing values)
- Drop variables such as overview of the movie and tagline as they are not helpful in revenue generation
- Keep movies of 2010 onwards to observe recent revenue trends
- For movies with multiple production companies, we only consider the 1st parent company for simplicity

Profit

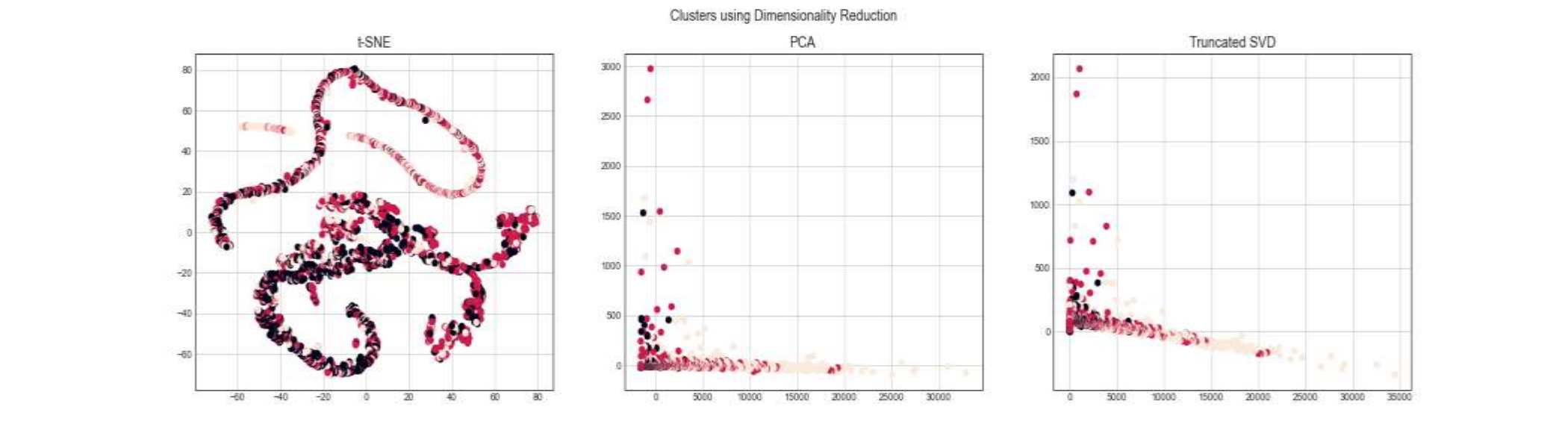
- We are interested in the profits of movies where Profit = Revenue / Budget
- If the ratio is > 3, profit is labeled as 2
- If the ratio is between 1 and 3, profit is labeled as 1
- If the ratio is less than 1 or negative, profit is labeled as 0


Distribution of profit
Density
profit
Distribution of age
Density
age
Distribution of vote_average
Density
vote_average
Distribution of vote_count
Density
vote_count
Distribution of runtime
Density
runtime
Distribution of popularity
Density
popularity
Distribution of bi_adult
Density
bi_adult
Distribution of original_language_freq
Density
original_language_freq
Distribution of genres_freq
Density
genres_freq
Distribution of main_company_freq
Density
main_company_freq

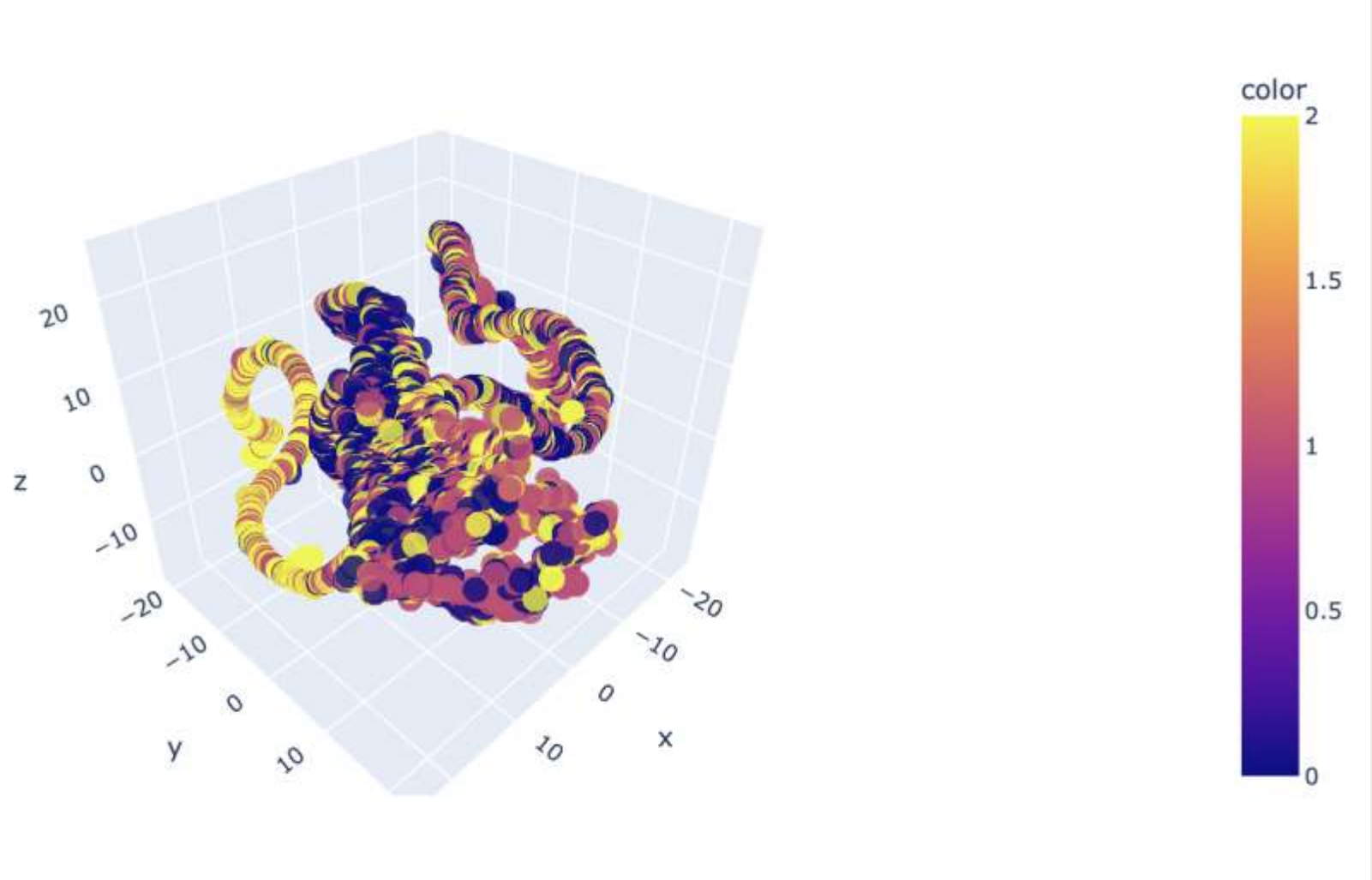
Dimensionality Reduction

- T–SNE (t – Distributed Neighbor Embedding) : Converts similarities between data points in the high dimension into probabilities. It helps to maintain the local structure of the data in the lower-dimensional space.
- PCA (Principal Component Analyst): Identifies the directions (principal components) that maximize the variance in the data. Then, projects the data onto these components to reduce the dimensionality.
- Truncated SVD (Truncated Singular Value Decomposition): Similar to PCA but is specifically designed to sparse data. It approximates the original matrix by using only the largest singular values.

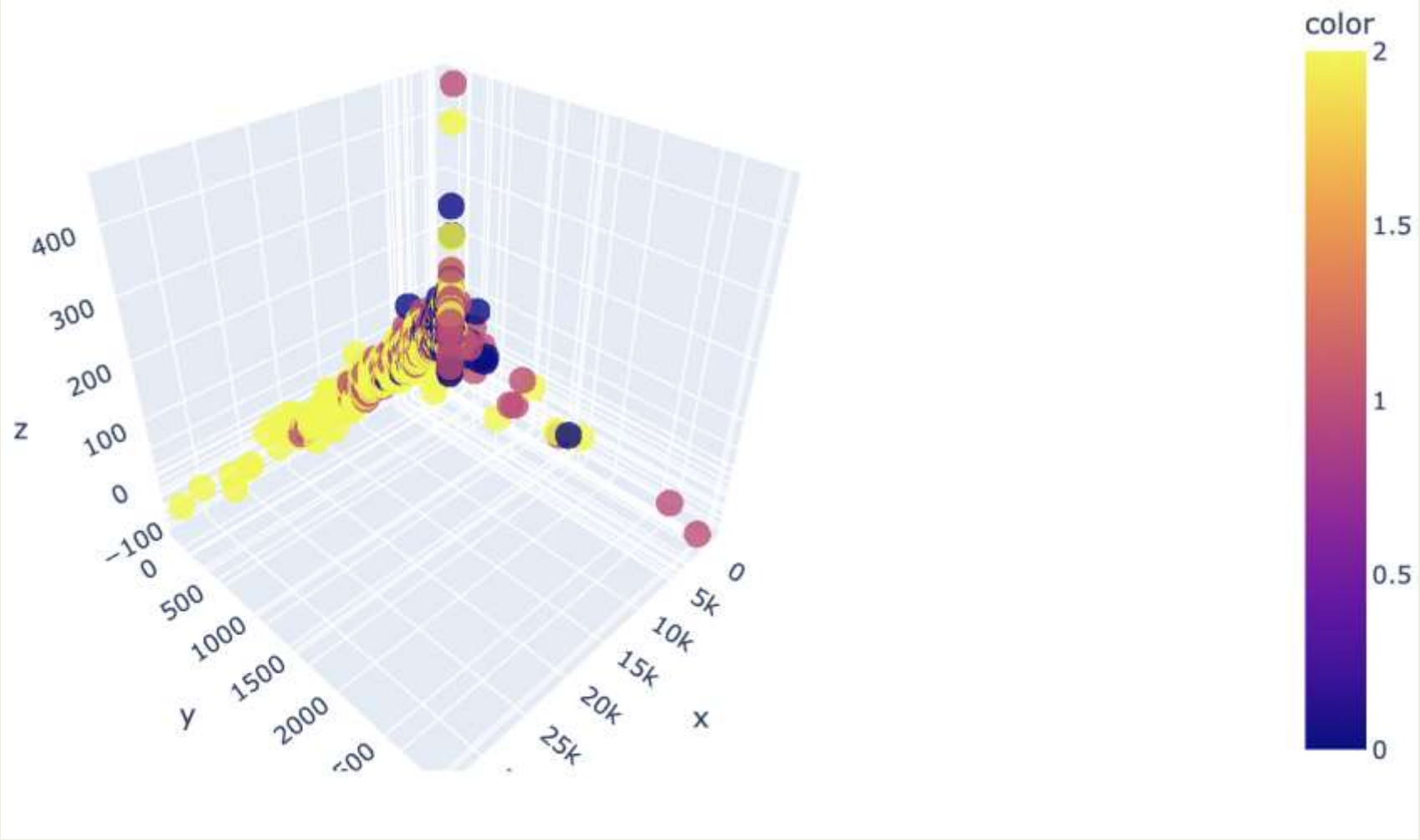
2-D


Clusters using Dimensionality Reduction
t-SNE
PCA
Truncated SVD

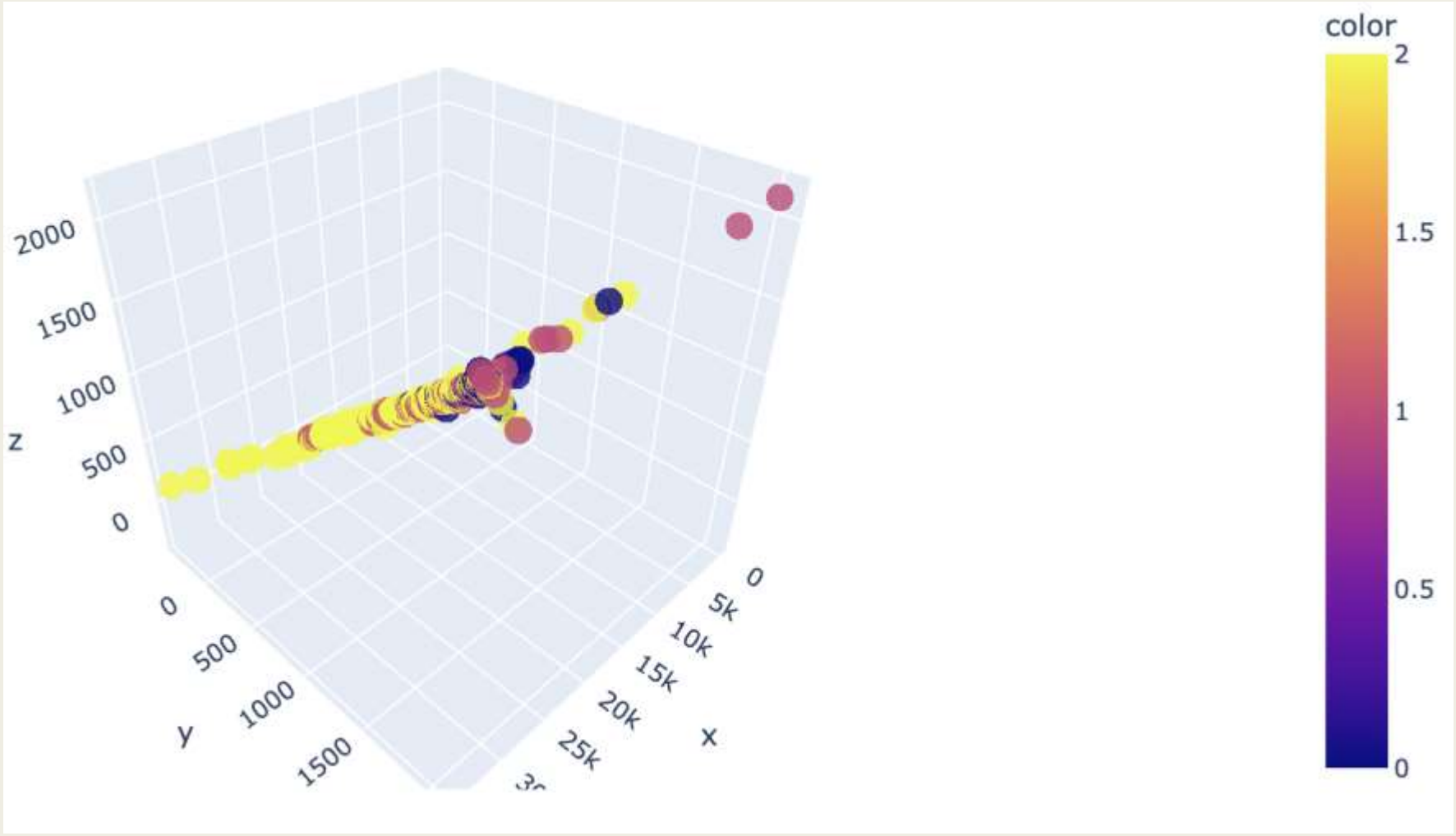
3-D (T-SNE)



3-D (PCA)


color
2
1.5
1
0.5
0
z
400
300
200
100
0
−100
y
0
500
1000
1500
2000
x
0
5k
10k
15k
20k
25k

3-D (SVD)


color
2
1.5
1
0.5
0
z
2000
1500
1000
500
0
y
0
500
1000
1500
x
0
5k
10k
15k
20k
25k

K - Means

Features	Feature Importance Score
Vote Count	12902.98
Popularity	27.17
Run time	24.68
Age (in years)	1.27
Vote Average (rating)	1.14
Original Language Frequency	0.13
Adult Category	0.01
Genre Frequency	0.01
Main Company Frequency	0.002

Performance Indicator

- Silhouette Score: It measures how similar an object is to its own cluster as compared to other clusters. It ranges from -1 to 1.
  - *Values close to 1 indicate that the data point is well matched to its own cluster and poorly matched to neighboring clusters*
  - *Values close to 0 indicate that the data point is very close to the decision boundary between two neighboring clusters*
  - *Values less than 0 indicate that the points might have been assigned to the wrong cluster*

Adjusted Rand Index & Adjusted Manual Index

- **Adjusted Rand Index**:
  - *This index measures the similarity between the true labels and the cluster assignment. A higher ARI indicates better clustering, with a range from -1 to +1.*
- **Adjusted Manual Index**:
  - *Similar to ARI but includes adjustment for matching that could occur by chance.*

Performance Score

Performance Indicators	Score
Silhouette Score	0.78
Adjusted Rand Index	0.03
Adjusted Mutual Score	0.07

Conclusion

- This project provides clustering for profitability analysis
- Valid information for producers, investors, and streaming platforms
- Results show that vote count and popularity are the most important features
- For future work, need more relevant features such as targeted population for each movie, average seating capacity of cinemas where the movie will be released, etc.