
ECOMMERCE RECOMMENDATION USING K-MEANS CLUSTERING

— Danylo Sovgut, Ke Wang, Shane Kim, Chris Korabik, Dean Q, Nathan Bywood, —
Vicki Yuan

About Us



**Danylo
Sovgut**



K Wang



Shane Kim



Chris Korabik



Dean Q



**Nathan
Bywood**



Vicki Yuan

A Smarter, Data-Driven Recommendation Engine

Problem:

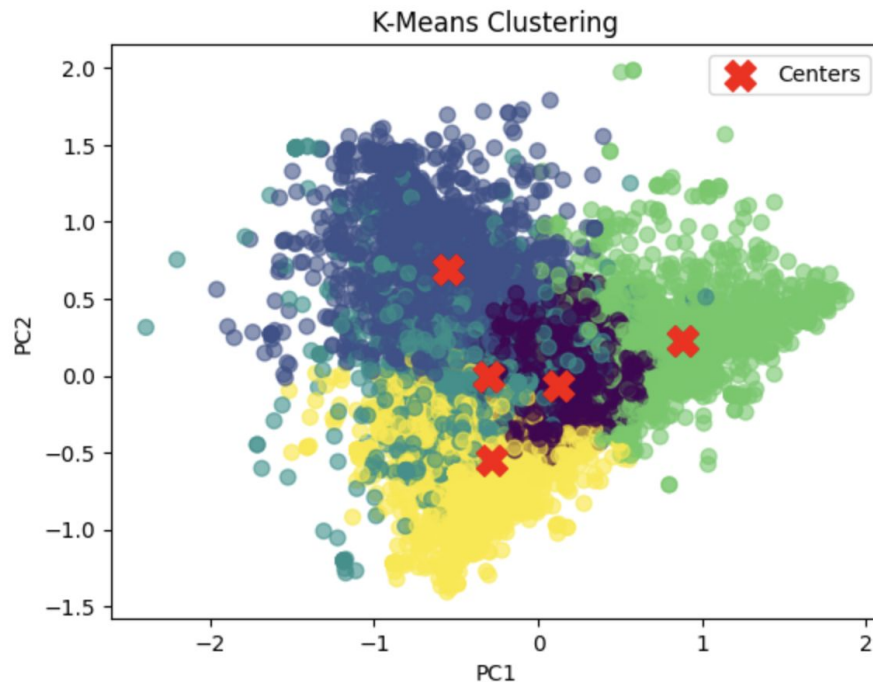
Our website currently recommends "most popular" items to all visitors, which lacks personalization and loses potential revenue.

Solution:

Our goal is to replace this one-size-fits-all approach with a smarter, data-driven recommendation engine that boosts revenue by **categorizing customers based on their behaviors**.

We Chose K-Means to Categorize Customers

- An **unsupervised clustering** model
- Allows us to effectively identify distinct segments by grouping customers based on their **buying frequencies & monetary values**






We Chose K-Means Based on Model Features



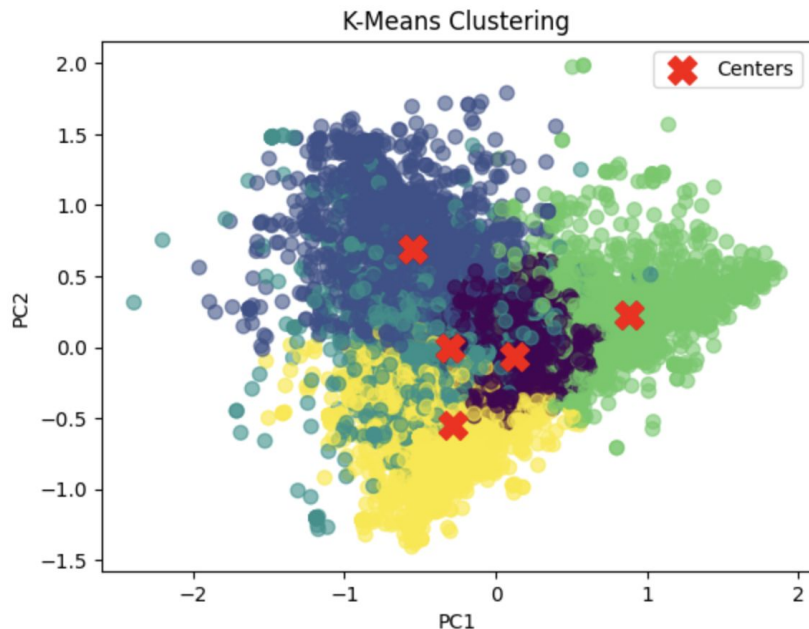
Models	Pros	Cons
K-Means	Simple and efficient for large datasets	Sensitive to outliers
DBSCAN	Automatically detects outliers as noise points	Does not perform well on high-dimensional data
GMM	Each point has a probability of belonging to a cluster	Computationally expensive
Hierarchical Clustering	Provides a dendrogram that helps understand how clusters form	Computationally expensive, sensitive to noise

We Chose K-Means Based on Model Performances

Models	Silhouette Score The higher the better	DBI The lower the better	Variance Ratio Criterion The higher the better
K-Means	 0.30	 1.34	 5100.74
DBSCAN	0.03	1.40	78.14
GMM	0.006	3.95	739.24

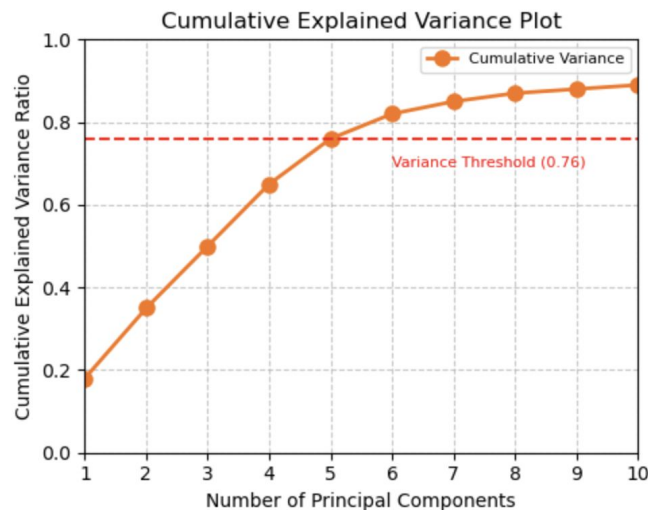
Our Key Metrics Provide Insight into Clustering Quality

- **Silhouette Score (0.30)**
 - Measures how similar a data point is to its own cluster compared to other clusters.
- **Davies-Bouldin Index (1.34)**
 - Evaluates cluster separation and compactness.
 - Lower values indicate better performance.
 - Values range from 0 to ∞
- **Variance Ratio Criterion (5100.74)**
 - Compares the ratio of intra-cluster dispersion to inter-cluster dispersion.
 - Higher values indicate better-defined clusters.
 - Values range from 0 to ∞



We Optimized Our Model Using A Variety of Techniques

- **Feature Extraction**
 - Selected “m” and all “F” columns
 - Divided all “F” columns by “f” (frequency)
- **Outlier Removal**
 - Removed using z-score method ($z > 3$)
- **MinMax Scaler**
 - K-Means and PCA require scale-sensitive data
 - Retains the relative distance between data points
- **Principal Component Analysis (PCA)**
 - Reduces dimensions into components
 - 5 components capture ~76% variance



Our Exhaustive Search Method Left Us With A Model We Are Confident In

Cols	Num Components	k	Inertia	Variance Ratio	Silhouette Score
[r]	5	2	16204.837706	0.742451	0.318002
[m, r]	5	2	16200.665454	0.742407	0.317944
[]	5	4	10435.758951	0.760873	0.309701
[m]	5	4	10433.912867	0.760833	0.309683
[m]	5	5	8517.901164	0.760833	0.301239
[]	6	2	17432.055222	0.801473	0.300878
[m]	6	2	17427.803346	0.801439	0.300869

- We iterated through **220 combinations** of k values, # of PCA components, and inclusion of extra columns.
- We ultimately chose model params that left us with relatively **low Inertia** and **high Silhouette Score**.

Our Model Provides Rankings of Most Frequently Purchased Categories in Each Cluster

History
Cont History
Music
Health
Travel

Cont History
Travel Guides
History
Music
Religion

Music
History
Health
Learning
Travel

History
Music
Cont Hist
Health
Travel

History
Music
Travel
Health
Non-Books

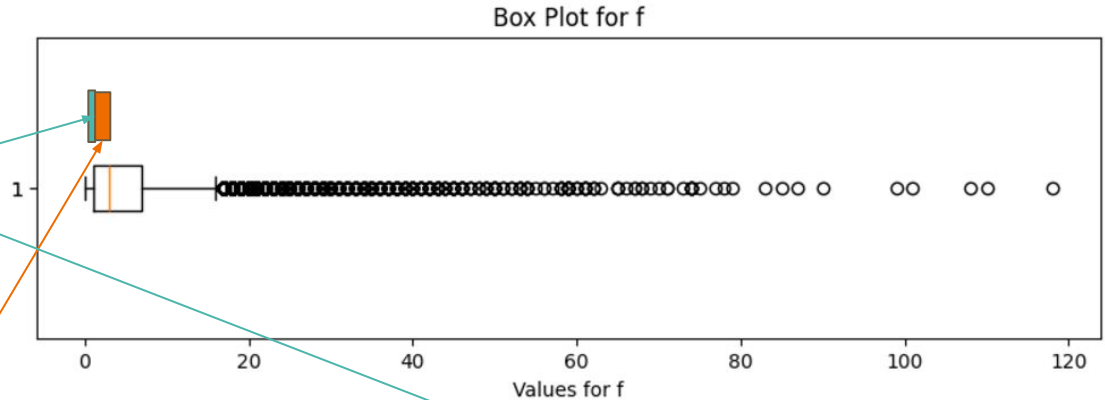
Our Clustering Model Provides Us with Recommendations Tailored to Each User

Test Set of 5 Real Users all assigned their own clusters/recommendations:

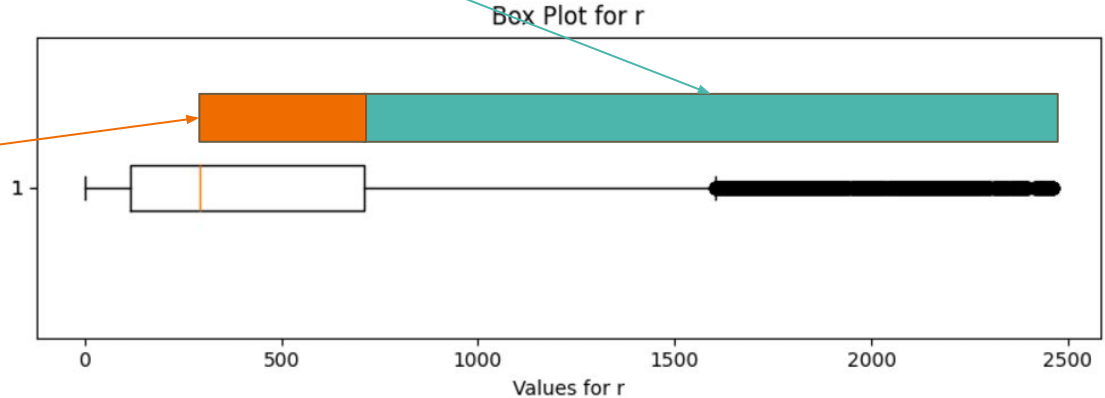
	r	f	m	tof	cluster	recs
27120	1272	2	164.559204	2210	2	Fhealth35, Fmusic14, Fhistory19, Flearning37, ...
4714	227	14	597.795898	2390	1	Fhistory19, Fconthist20, Ftravelguides31, Fmus...
16123	187	3	49.749969	634	4	Ftravelguides31, Fhistory19, Fmusic14, Fhealth...
27444	12	14	585.036133	2162	4	Ftravelguides31, Fhistory19, Fmusic14, Fhealth...
22196	110	1	33.739960	110	0	Fmusic14, Fhistory19, Fconthist20, Fhealth35, ...

We Also Provide Discount Offers Based on Quartiles in Recency, Frequency

- A **30% discount** is offered to users in Q4 for **r** or Q1 for **f**



- A **15% discount** is offered to users in Q3 for **r** or Q2 for **f**



We Calculated Expected Monetary Values for the Clustering and Base Model

Step 1

Calculate the **probability** for each book category:

$$\text{Probability} = \frac{\text{Category Mean Frequency}}{\text{Total Mean}}$$

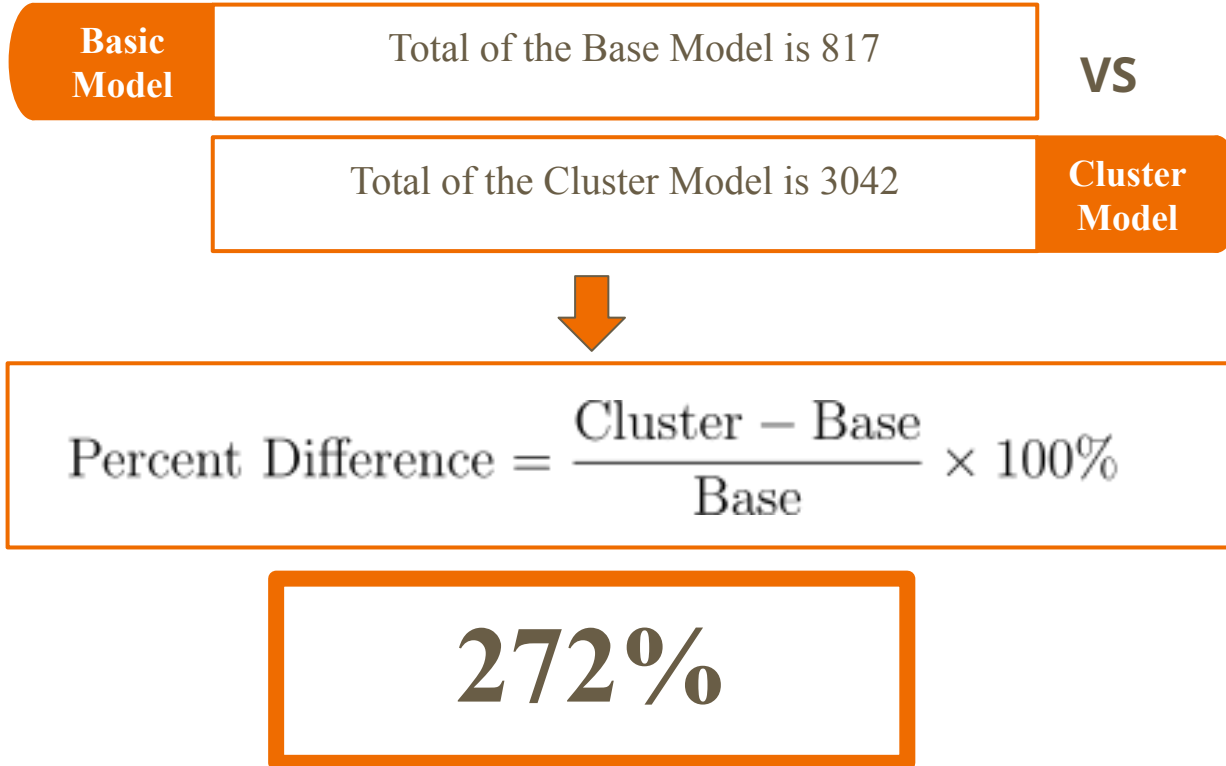
Step 2

Multiply it by the average price to get the **expected category value**:
Expected Category Value = Average Price x Probability

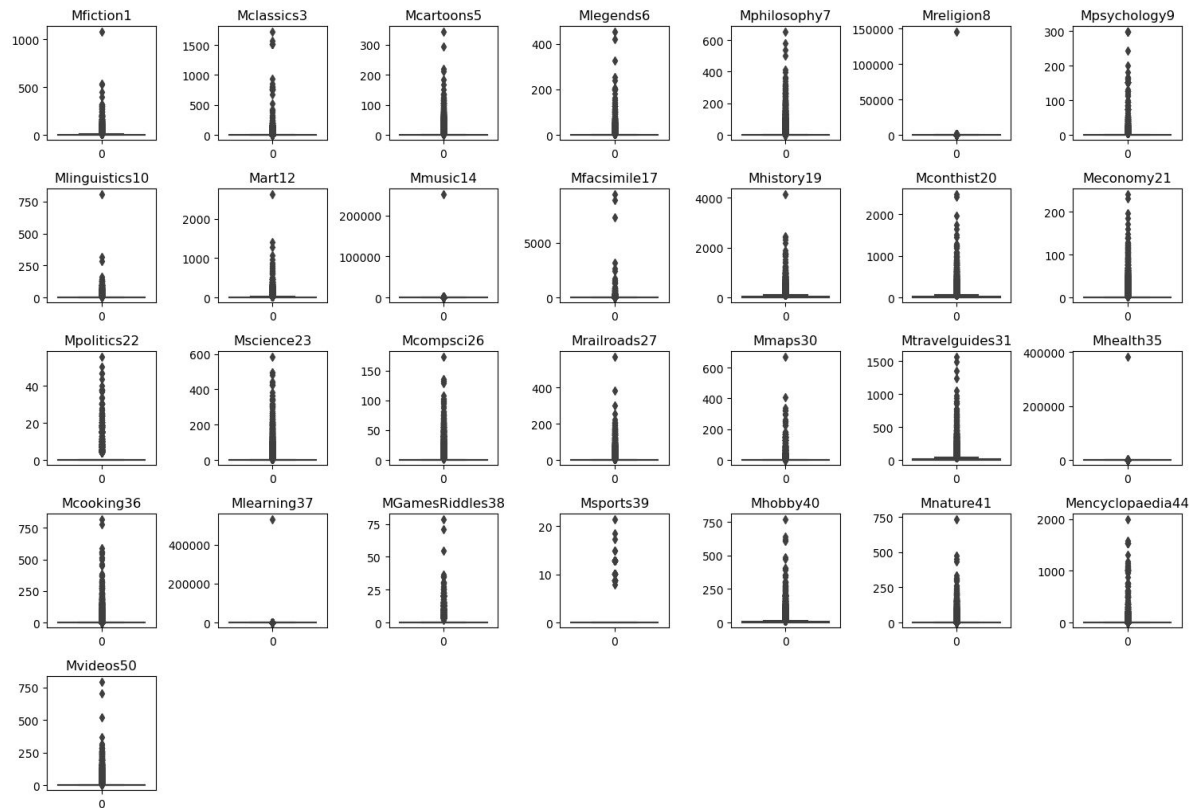
Step 3

Sum all the monetary values across all categories to get **total monetary index**
Total Monetary Index = Category 1 + Category 2 + ... Category 10

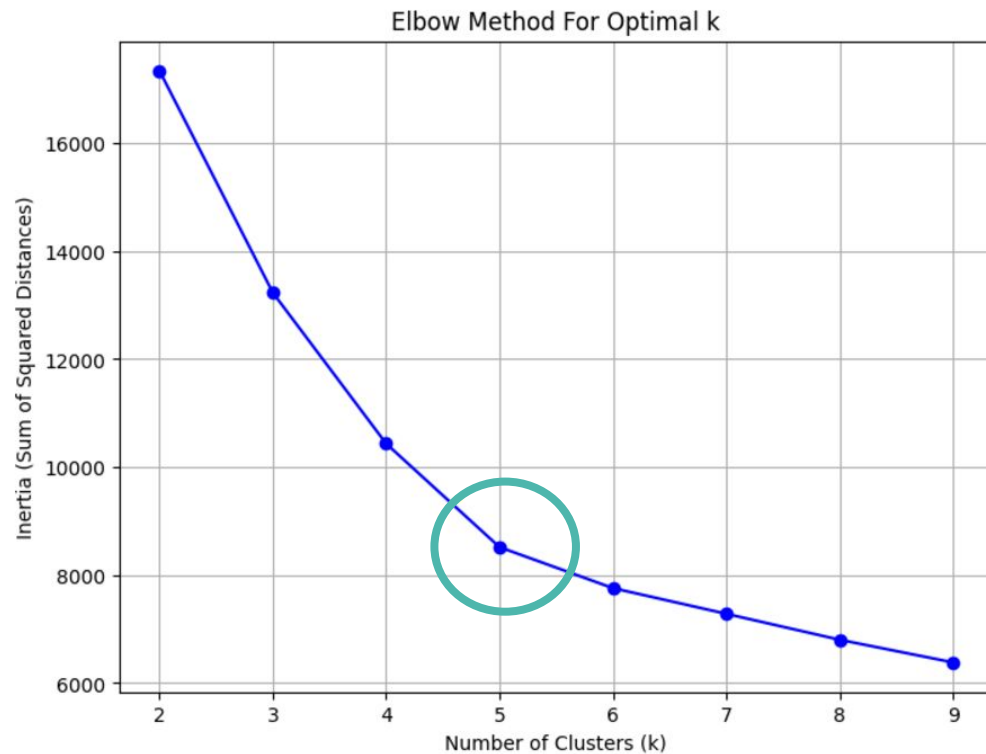
Our Model Is 272% Better than the Base Model



Appendix i



Appendix ii



Appendix iii

$$T = \sum_{i=1}^{10} \frac{M(f_i) \cdot \text{Mean}(f_i)}{\text{Total Mean}}, \text{Total Monetary Value of the Base Model}$$

$$CV = \sum_{i=1}^{10} \frac{M(f_i) \cdot \text{Mean}(f_i|C)}{\sum_{j=1}^{32} \text{Mean}(f_j|C)}, \text{Monetary Value for a Specific Cluster}$$

$$ACM = \sum_{C=1}^5 (P_C \cdot CV_C), \text{Adjusted Monetary Value for the Cluster Model}$$

Appendix iv

Explanation of Each Formula and Variable:

1. **Base Model Formula:** $[T = \sum_{i=1}^{10} \frac{M(f_i) \cdot \text{Mean}(f_i)}{\text{Total Mean}}]$

- **T:** Total monetary value of the base model.
- **M(f_i):** Monetary value assigned to feature (f_i).
- **Mean(f_i):** Average frequency of feature (f_i) across all users.
- **Total Mean:** The sum of the mean values for the top 10 features.

2. **Cluster Model Formula:** $[CV = \sum_{i=1}^{10} \frac{M(f_i) \cdot \text{Mean}(f_i \mid C)}{\sum_{j=1}^{32} \text{Mean}(f_j \mid C)}]$

- **CV:** Monetary value for a specific cluster (C).
- **M(f_i):** Monetary value of feature (f_i).
- **Mean(f_i | C):** Average frequency of feature (f_i) within cluster (C).
- **$\sum_{j=1}^{32} \text{Mean}(f_j \mid C)$:** Sum of average frequencies of all 32 features in cluster (C).

3. **Adjusted Cluster Model Formula:** $[ACM = \frac{\sum_{C=1}^5 CV}{5}]$

- **ACM:** Adjusted monetary value, averaged across all 5 clusters.
- **CV:** Monetary value for each cluster (C).

4. **Percent Difference Formula:** $[\text{Percent Difference} = \frac{100 \cdot (ACM - T)}{T}]$

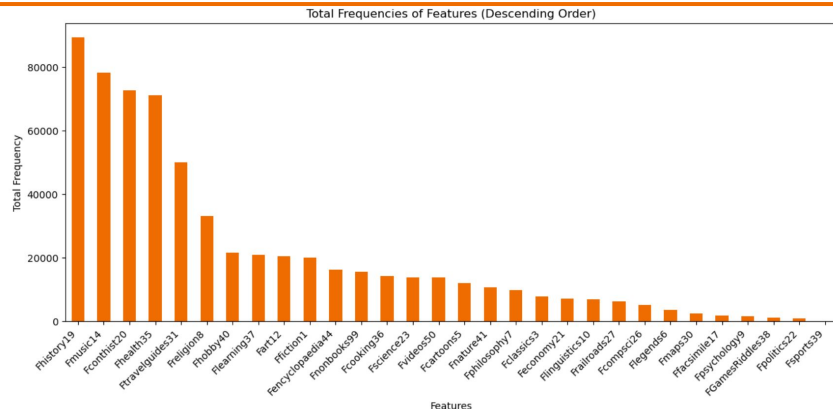
- **Percent Difference:** The relative improvement of the cluster model over the base model, expressed as a percentage.
- **ACM:** Adjusted cluster model value.
- **T:** Base model total value.

Appendix v

Step 1

Extract the **top 10 categories** based on total frequencies

$$T = \sum_{i=1}^{10} \frac{M(f_i) \text{Mean}(f_i)}{\text{Total Mean}}$$



Step 2

Calculate the **Weighted Mean** for each Feature

Step 3

Multiply the mean by the monetary value for each category and **Sum them up**

Appendix vi

Monetary Index	The Base Model total (T) is assumed to be an aggregated monetary index , not an actual revenue figure.	Each feature's monetary value is assumed to reflect its financial contribution accurately .	Feature Value
Equal Influence	All users are assumed to be equally influenced by the most popular features.	The monetary index (T and ACM) is valid as a relative measure , even if it doesn't represent real revenue or customer value.	Index Validity
User Interaction	Assumes users will interact with or purchase based on recommendations , without accounting for conversion rates.		
Behavior Consistency	User behavior within each cluster is assumed to be consistent and reflective of preferences .	Monetary values and user preferences are static and do not change over time or external factors.	Static Preferences

An Analogy for Our Clustering Model

- Have you ever been pressured into watching a movie you didn't want to?
- **Our solution: a personalized movie night**, where attendees are divided into **five genre groups** based on their movie preferences.
- We conduct a survey to analyze each attendee's favorite genres and past viewing habits.
- Each group gets its own movie screening of their preferred genre.
- Using this information, we group attendees into **five distinct movie-watching groups: Music, History, Health, Contemporary History, and Travel/Adventure.**

