  
UNIVERSITY OF HERTFORDSHIRE  
School of Physics, Engineering and Computer Science

MSc Data Science and Analytics  
(7COM1039-0509-2024) Advanced Computer Science Master's Project  
14th July, 2025

Customer Behaviour Analysis For Retail Optimization Using K-Mode and Association Rule Mining

Name: Amobi Obidiebube  
Student ID: 23059829  
Supervisor: Paul Moggridge

Contents

[**Interim Progress Report: Customer Behaviour Analysis for Retail Optimization Using K-Mode and Association Rule Mining** 3](#_Toc203395779)

[**1.** Introduction and Overview 3](#_Toc203395780)

[**1.1** Research Question and Investigation 3](#_Toc203395781)

[**1.2** Technical Approach and Deliverables 4](#_Toc203395782)

[**1.3** Tools and Methodology 5](#_Toc203395783)

[**1.4** Ethical and Legal Considerations 5](#_Toc203395784)

[**2.** Progress to Date 6](#_Toc203395785)

[**2.1** Literature Review 6](#_Toc203395786)

[**2.2** Data Processing and Analysis Progress 7](#_Toc203395787)

[**2.3** K-Modes Clustering for Categorical Movie Segmentation 8](#_Toc203395788)

[**2.4** Challenges and Solutions 8](#_Toc203395789)

[**2.4** Alignment with Project Objectives 9](#_Toc203395790)

[**3.** Planned Work 9](#_Toc203395791)

[**3.1** Remaining Tasks and Timeline 9](#_Toc203395792)

[**3.2** Evaluation Strategy 10](#_Toc203395793)

[**4.** Conclusion 11](#_Toc203395794)

[Bibliography 13](#_Toc203395795)

[Appendix A: Code Snippets for Data Preprocessing 15](#_Toc203395796)

[**A.1. Library Imports and Setup** 15](#_Toc203395797)

[**A.2. Dataset Loading** 15](#_Toc203395798)

[**A.3. Initial Data Inspection and Screenshot of Output** 15](#_Toc203395799)

[**A.4. Timestamp Conversion** 16](#_Toc203395800)

[**A.5. Missing Value Check and Screenshot of Output** 16](#_Toc203395801)

[**A.6. Duplicate Tag Removal** 17](#_Toc203395802)

[**A.7. Data Merging** 17](#_Toc203395803)

[**A.8. Rating Distribution Summary and Screenshot of Output** 17](#_Toc203395804)

[**A.9. Unique Users and Movies Count and Screenshot of Output** 18](#_Toc203395805)

[**A.10. Top 10 Most Rated Movies and Screenshot of Output** 18](#_Toc203395806)

[**A.11. Distribution of Ratings Visualization and Screenshot of Output** 19](#_Toc203395807)

[**A.12. Number of Ratings per Movie Visualization and Screenshot of Output** 20](#_Toc203395808)

[**A.13. WordCloud of Movie Tags and Screenshot of the output** 21](#_Toc203395809)

[**A.14.**  **Installing the K-modes library and Preparing the data for clustering** 22](#_Toc203395810)

[**A.15.**  **Applying the K-modes clustering and Screenshot of the results** 23](#_Toc203395811)

[**A.16.**  **Installing the MCA Library for 2D Visualization of Clusters and Screenshot of Output** 26](#_Toc203395812)

# **Interim Progress Report: Customer Behaviour Analysis for Retail Optimization Using K-Mode and Association Rule Mining**

## **1.** Introduction and Overview

This report presents the progress made on an MSc Data Science project that focuses on analyzing customer behavior patterns using K-Mode clustering and association rule mining techniques. The project aims to develop a comprehensive framework for retail optimization by combining these analytical approaches to extract deeper insights than either method could provide independently. This research is important because it has the potential to change how companies understand and respond to consumer preferences, which will ultimately result in better recommendation systems and happier customers. Combining these two powerful analytical methods is an innovative approach that overcomes the shortcomings of current approaches and offers a scalable solution for modern retail settings.

### **1.1** Research Question and Investigation

The primary research question informing this project is: "Can the combination of clustering and association rule mining reveal more useful knowledge for recommendation systems than applying each individually?" The question addresses a prime knowledge gap in current literature, in that few studies apply these methods individually from each other, as opposed to investigating their combined potential. The fusion is recognized in the research to respond that while each technique in its own has been helpful, their joint application of potentially unleashing hitherto unseen patterns and relationships among customer data. This research study seeks to investigate whether fusing such techniques could provide improved predictive capabilities and deeper insight into customers' behavior patterns.

The study targets the MovieLens dataset, which is high in categorical data like user ratings, movie genres, and user tags. The dataset is ideal for experimenting with applying retail analytics techniques to entertainment scenarios, and the findings should generalize across other domains where categorical behavioral data prevails. This dataset is chosen deliberately in that it encompasses the type of categorical data found in retail environments but is sophisticated enough to illustrate the value of the proposed analytical approach. The entertainment environment has unique challenges with a similar nature as retail environments, including diverse user interests, complex product taxonomies, and personalized recommendation needs.

### **1.2** Technical Approach and Deliverables

The technical implementation involves developing a machine learning framework that effectively combines K-Modes clustering with association rule mining. K-Modes is particularly suited for this project because it handles categorical data directly without requiring numerical transformation, preserving the meaningful relationships between categories that are often lost in traditional clustering approaches. This preservation of categorical integrity is crucial for maintaining the semantic meaning of customer preferences and behaviors. The framework will incorporate advanced optimization techniques to ensure scalability and robustness when applied to large-scale datasets typical of modern retail environments.

The planned deliverables include:

* A preprocessed MovieLens dataset optimized for categorical analysis
* K-Modes clustering models for user segmentation based on genre preferences and tagging patterns
* Association rules identifying frequent patterns in user behavior and content relationships
* An integrated analytical framework combining both approaches
* Comparative analysis demonstrating the investigation of the suitable integrated method
* Practical recommendations for enhancing recommender systems
* A generalizable framework applicable to similar problems across industries

### **1.3** Tools and Methodology

Python is being utilized as the primary development tool for this research, and it is specifically leveraging pandas and numpy for data manipulation, matplotlib and seaborn for plotting, and Jupyter Notebook for execution. The primary analytical techniques are K-Modes clustering for clustering categorical data and association rule mining using Apriori or FP-Growth algorithms for pattern discovery. Use of such tools is a sign of best practice in industry and enables reproducibility of the results with computational efficiency. The method incorporates the finest of machine learning and data mining best practices into its fold so that the analytical framework is kept within known scientific principles while new paradigms are incorporated into the data.

### **1.4** Ethical and Legal Considerations

The project addresses many important legal and ethical concerns that are of first priority in today's data science scholarship. Since the research is based on the publicly available, anonymized MovieLens dataset, formal ethics approval is not required; however, the research insists on rigorous attention to privacy issues and bias in recommendation. The ethical model emphasizes transparency surrounding data usage and algorithmic decision-making procedures. The research is more interested in enhancing user experience via personalization rather than manipulation, i.e., that model analyses should avoid reinforcing biases in data that might lead to discriminatory recommendations.

Legally, while the MovieLens dataset is openly accessible, any future use of this framework must be undertaken in compliance with existing data protection laws such as GDPR and CCPA. The study is being undertaken in full academic integrity, providing due transparency regarding methodologies and loopholes and following rigorous documentation standards. The legal considerations also relate to intellectual property rights and ethical provision of research outcomes, with the assurance that the system created can be used ethically in commercial purposes without infringing on user freedom or privacy.

## 

## **2.** Progress to Date

### **2.1** Literature Review

Comprehensive literature review has been undertaken that establishes the theoretical foundation of this project. The review highlights a number of important gaps in the current research that have been addressed by this project in its novel methodology. Systematic literature analysis has established core theoretical concepts and methodology that inform the project design and implementation strategy. The process of reviewing the literature involved examining more than 50 peer-reviewed papers, conference documents, and industry reports to have adequate coverage of applicable research areas.

Traditional clustering algorithms like K-Means are greatly hampered in the case of categorical data because they rely on numeric distance and mean measures, which cannot be meaningfully used for categorical variables (Ahmad & Dey, 2007). The K-Modes algorithm overcomes such handicaps by using mode-based centroids and simple matching dissimilarity measures, which are appropriate to use with categorical behavioral data analysis (Huang, 1998). Theoretical foundations of K-Modes provide a robust mathematical framework for dealing with the problem of complexity in categorical data analysis. The ability of this algorithm to preserve the discrete nature of the categorical variables while guaranteeing clustering effectiveness makes it particularly suitable for analysis of customer behavior in retail settings.

Recent advances have further enhanced K-Modes' applicability to modern data science challenges. Dinh et al. (2025) have developed scalable implementations for high-dimensional categorical data, while Thapaliya and Zhuang (2025) have improved initialization techniques for big data environments. These developments make K-Modes particularly suitable for large-scale applications like MovieLens analysis, where traditional clustering methods would fail to capture meaningful patterns. The evolution of K-Modes algorithms demonstrates the continued relevance and adaptability of this approach to contemporary data science challenges.

Association rule mining has proven effective in revealing co-occurrence patterns in transactional data across various domains. The mathematical foundation relies on key metrics including support, confidence, and lift to identify meaningful relationships (Tan et al., 2019). Recent applications in entertainment contexts have shown promise, with Hashad et al. (2024) demonstrating ARM's effectiveness in revealing genre combinations and customer interests. The theoretical underpinnings of association rule mining provide a solid framework for understanding complex relationships within customer behavior data, enabling the discovery of previously unknown patterns and preferences.

The integration of clustering and association rule mining is a high-level method that capitalizes on the complementary strengths of both methods. Nguyen et al. (2022) explored hybrid approaches in content recommendation, and Gupta & Goyal (2023) demonstrated uses in personalized marketing. Limited literature, however, explored this integration for categorical entertainment data, our project's new contribution. The synergy in merging these methods has the potential to open up deeper insights that each methodology could not produce on its own.

### **2.2** Data Processing and Analysis Progress

Data preprocessing and ingestion phase has been carried out to successful completion, setting the stage for analysis. The MovieLens dataset ratings.csv, movies.csv, tags.csv, and links.csv files have all been loaded and preprocessed for analysis in accordance with standard data preprocessing practice. The preprocessing pipeline has quality checks to ensure data integrity and consistency throughout the analytical process. This phase is a fundamental underpinning that enables the successful application of advanced analytical techniques in subsequent phases of the project.

Key preprocessing achievements include:

* Converting timestamp columns from Unix format to human-readable datetime format for potential temporal analysis
* Conducting thorough missing value analysis, identifying that only the tmdbId column contains missing values (which will not affect the analysis)See Figure 2
* Removing duplicate entries from the tags dataset to ensure data integrity
* Merging ratings and movies dataframes to create a comprehensive dataset for analysis
* Generating descriptive statistics showing a mean rating of 3.5 across 610 unique users and 9,724 unique movies. See Figure 3.

The exploratory analysis revealed interesting patterns in the data that provide valuable insights for the subsequent analytical phases. Rating distributions show a tendency toward higher ratings with a peak around 4.0, suggesting generally positive user experiences and indicating potential rating bias that must be considered in the analytical framework. The analysis of ratings per movie reveals the characteristic long-tail distribution where a few movies receive many ratings while most receive relatively few, reflecting typical consumption patterns in digital entertainment platforms. These patterns align with established theories of user behavior in digital media consumption and provide validation for the dataset's representativeness.

### **2.3** K-Modes Clustering for Categorical Movie Segmentation

After preprocessing the data, a K-Modes clustering model was executed to group movies based on their combinations of genres. As K-Means, utilized for numeric data, would be a bad choice for categorical attributes, the better-suited K-Modes was employed instead. The genres column in the movies dataset was split and one-hot encoded to convert every genre into a binary feature. This conversion was performed to enable the data for clustering without losing the categorical nature of the initial values.

The model was also fixed at five clusters, which meant groups of films with identical genre profiles. For the purpose of easy interpretation, Multiple Correspondence Analysis (MCA) was used to reduce the high-dimensional data to two components to ensure that the clusters could be easily visualized. The plot generated had distinct clusters, some made up of science fiction and action films and others made up of romantic dramas or family films.

This is the foundation of content-based recommendation systems, audience segmentation, and content acquisition strategy. It is also a forerunner to the sophisticated analysis phase of the project where machine learning algorithms are run on the data to generate actionable business intelligence.

### **2.4** Challenges and Solutions

The preprocessing phase has proceeded as anticipated on the basis of the well-structured nature of the MovieLens data, but several issues of methodology have been clarified for upcoming phases. The main challenges ahead are applying Association rule mining in combination with K-modes clustering effectively to potentially enormous, sparse categorical data and deriving meaningful association rules that provide actionable knowledge rather than statistical abstractions. The research community has offered solutions for overcoming the said problems through careful parameter tuning and validation procedures. The other problems are to achieve a balance between computational ability vs. analytical frugality and to ensure the resulting integrated framework provides comprehensible results for practical use.

The literature review has identified several specific gaps that this project addresses through its innovative methodological approach:

* Limited integration of clustering and association rule mining in entertainment contexts
* Insufficient attention to specialized categorical data analysis techniques
* Lack of scalable, practical implementation frameworks
* Need for more interpretable and actionable analytical approaches

### **2.4** Alignment with Project Objectives

The effort thus far supports the overall objectives of the project and establishes a solid foundation for subsequent analysis stages. The comprehensive literature review has built the theoretical basis and identified specific research gaps that will be addressed by the current project within its new direction. The meticulous data preparation prepares the dataset for advanced analysis while ensuring data quality and integrity. This framework positions the project ideally to perform K-Modes clustering, association rule mining, and to develop the integrated analytical framework that represents the project's key contribution to the research field.

## **3.** Planned Work

### **3.1** Remaining Tasks and Timeline

The following major tasks will complete the project according to the established timeline and quality standards:

**Task 1: Association Rule Mining (1 week)**

* Apply Apriori or FP-Growth algorithms to identify significant relationships within user clusters
* Generate frequent itemsets and derive association rules
* Evaluate rules using support, confidence, and lift metrics

**Task 2: Integrated Analysis (3 days)**

* Combine clustering and association mining insights into unified recommendations
* Develop visualization frameworks for presenting integrated patterns
* Create actionable insights for retail optimization

**Task 3: Evaluation and Refinement (1 week)**

* Assess the effectiveness of integrated approach versus individual methods
* Refine models based on performance evaluation
* Document limitations and areas for future improvement

**Task 4: Final Report Writing (1 week)**

* Consolidate research findings into comprehensive final report
* Ensure academic writing standards and proper formatting

### **3.2** Evaluation Strategy

The project will be successful based on methodological rigor, generation of insights, practicability, and communicative clarity by the standards of current scholarly practice. The final artifact should be a sound analytical model for processing raw movie data, segmenting the users into descriptive categories, and discovering strong association rules in the segment. The assessment framework integrates quantitative measures with qualitative judgments to determine complete validation of research findings. The evaluation process will incorporate peer review components and external validation as needed to insure the research is of the highest academic quality.

Quality assessment will focus on:

* Appropriate application of analytical techniques to categorical data
* Novelty and depth of discovered patterns
* Practical utility of derived recommendations
* Technical implementation efficiency
* Critical reflection on outcomes and limitations

## **4.** Conclusion

This report has made significant progress in the theoretical and practical groundwork for customer behavior analysis through integrated clustering and association rule mining. The extensive literature review has identified significant gaps to be filled by this project through its novel methodology approach, and the careful preparation of data has set a solid stage for subsequent analysis. The success so far gained progress validates the feasibility of the project and guarantees the research team's capacity to make quality contributions to customer behavior study as a field of academics. The schedule and evaluation framework that have been established guarantee open doors to the successful completi on the project with maximum academic rigor.

The research activities conducted so far put the project in the best position to excel in the future stages of research and development. The combination of K-Modes clustering with association rule mining can potentially yield deep insights into customer behavior patterns that can be utilized to guide retail optimization strategies in various industries. The research methodology developed through the research activities will contribute to theoretical understanding as well as practical applications to the field of data science. As the project approaches completion, the research is expected to set up a framework that not only contributes to academic knowledge but also proves to be useful for real application within the industry for analysis of customer behavior and recommendation systems.

The second stage will focus on the application of fundamental analytical tools and demonstrating their combined strength in uncovering actionable insights for retail optimisation and recommendation systems. Integrating these methods marks a significant advance in the field and holds the potential to make important contributions to both theory and practice. The research framework developed in this project will serve as a foundation for future research into customer behaviour analysis and as a model for upcoming projects in related areas.

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

# Bibliography

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 207–216.<https://doi.org/10.1145/170035.170072>

Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering, 63*(2), 503–527.<https://doi.org/10.1016/j.datak.2007.03.016>

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems, 46*, 109–132.<https://doi.org/10.1016/j.knosys.2013.03.012>

Çano, E., & Morisio, M. (2017). Hybrid recommender systems: A systematic literature review. *Intelligent Data Analysis, 21*(6), 1487–1524.<https://doi.org/10.3233/IDA-163209>

Dinh, T., Wong, H., Fournier-Viger, P., Lisik, D., Ha, M. Q., Dam, H. C., & Huynh, V. N. (2025). Categorical data clustering: 25 years beyond Kmodes. *Expert Systems with Applications, 272*, Article 126608.

GroupLens. (n.d.). MovieLens dataset. University of Minnesota.<https://grouplens.org/datasets/movielens/>

Gupta, R., & Goyal, R. (2023). Hybrid recommendation systems: Combining clustering and association mining for personalized marketing. *Journal of Information Systems Research, 12*(1), 55–70.

Hashad, A. A., Khaw, K. W., Alnoor, A., & Chew, X. (2024). Exploratory analysis with association rule mining algorithms in the retail industry. *Malaysian Journal of Computing, 9*(1), 1746–1758.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery, 2*(3), 283–304.<https://doi.org/10.1023/A:1009769707641>

Jannach, D., Adomavicius, G., Tuzhilin, A., & Jugovac, M. (2016). Recommender systems: Challenges, insights and research opportunities. *ACM Transactions on Interactive Intelligent Systems, 7*(1), 1–42.<https://doi.org/10.1145/3004291>

Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender system application developments: A survey. *Decision Support Systems, 74*, 12–32.<https://doi.org/10.1016/j.dss.2015.03.008>

Nguyen, V., Zhao, X., & Tang, J. (2022). Hybrid recommendation models using clustering and association rule mining for content personalization. *International Journal of Data Science and Analytics, 15*(3), 321–338.<https://doi.org/10.1007/s41060-021-00289-w>

Tan, P. N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson.<https://books.google.com.ng/books/about/Introduction_to_Data_Mining_eBook_Global.html?id=274oEAAAQBAJ&redir_esc=y>

Thapaliya, S., & Zhuang, J. (2025). Fast clustering of categorical big data. arXiv.<https://arxiv.org/abs/2502.07081>

## 

## 

## 

## 

## 

## 

## 

## 

## 

# Appendix A: Code Snippets for Data Preprocessing

This appendix contains the Python code snippets used for the data ingestion and preprocessing steps, as referenced in Section 4 of this report.

### **A.1. Library Imports and Setup**

**import** pandas **as** pd

**import** numpy **as** np

**import** matplotlib**.**pyplot **as** plt

**import** seaborn **as** sns

**from** wordcloud **import** WordCloud

sns**.set(**style**=**'whitegrid'**,** palette**=**'pastel'**,** font\_scale**=**1.1**)**

### **A.2. Dataset Loading**

# Load the CSV datasets

ratings **=** pd**.**read\_csv**(**'ratings.csv'**)**

movies **=** pd**.**read\_csv**(**'movies.csv'**)**

tags **=** pd**.**read\_csv**(**'tags.csv'**)**

links **=** pd**.**read\_csv**(**'links.csv'**)**

### **A.3. Initial Data Inspection and Screenshot of Output**

**print(**"Ratings Sample:\n"**,** ratings**.**head**())**

**print(**"Movies Sample:\n"**,** movies**.**head**())**

**print(**"Tags Sample:\n"**,** tags**.**head**())**

**print(**"Links Sample:\n"**,** links**.**head**())**

A screen shot of a computer

AI-generated content may be incorrect.

Figure 1 The output of executing data inspection. Notice that the dataset has been loaded successfully

### **A.4. Timestamp Conversion**

ratings**[**'timestamp'**]** **=** pd**.**to\_datetime**(**ratings**[**'timestamp'**],** unit**=**'s'**)**

tags**[**'timestamp'**]** **=** pd**.**to\_datetime**(**tags**[**'timestamp'**],** unit**=**'s'**)**

### **A.5. Missing Value Check and Screenshot of Output**

**print(**"\nMissing Values Summary:"**)**

**print(**ratings**.**isnull**().sum())**

**print(**movies**.**isnull**().sum())**

**print(**tags**.**isnull**().sum())**

**print(**links**.**isnull**().sum())**

A screen shot of a computer

AI-generated content may be incorrect.

Figure 2 The output of executing the missing value check, notice that the dataset has been checked for missing values and 8 missing values were found on the tmdbId column on the links dataset

### **A.6. Duplicate Tag Removal**

tags**.**drop\_duplicates**(**inplace**=True)**

### **A.7. Data Merging**

ratings\_movies **=** pd**.**merge**(**ratings**,** movies**,** on**=**'movieId'**)**

### **A.8. Rating Distribution Summary and Screenshot of Output**

**print(**"\nRatings Summary:"**)**

**print(**ratings**[**'rating'**].**describe**())**

A screenshot of a computer

AI-generated content may be incorrect.

Figure 3 The output of executing the rate of distribution, notice that a lot of values were calculated, including mean, std, etc

### **A.9. Unique Users and Movies Count and Screenshot of Output**

**print(**f"\nNumber of unique users: {ratings**[**'userId'**].**nunique**()**}"**)**

**print(**f"Number of unique movies: {ratings**[**'movieId'**].**nunique**()**}"**)**

A black background with white text

AI-generated content may be incorrect.

Figure 4 The output gotten from Calculating the number of unique users and number of movies in the dataset

### **A.10. Top 10 Most Rated Movies and Screenshot of Output**

top\_rated **=** ratings\_movies**[**'title'**].**value\_counts**().**head**(**10**)**

**print(**"\nTop 10 Most Rated Movies:\n"**,** top\_rated**)**

A screenshot of a computer

AI-generated content may be incorrect.

Figure 5 The output of executing the top ten most rated movies, notice that the top ten most rated were selected from the dataset

### **A.11. Distribution of Ratings Visualization and Screenshot of Output**

plt**.**figure**(**figsize**=(**8**,** 5**))**

sns**.**histplot**(**ratings**[**'rating'**],** bins**=**9**,** kde**=True)**

plt**.**title**(**'Distribution of Ratings'**)**

plt**.**xlabel**(**'Rating'**)**

plt**.**ylabel**(**'Count'**)**

plt**.**tight\_layout**()**

plt**.**show**()**

A graph with blue lines

AI-generated content may be incorrect.

Figure 6 The output of executing the Visualization of the Distribution Ratings

### **A.12. Number of Ratings per Movie Visualization and Screenshot of Output**

ratings\_per\_movie **=** ratings**[**'movieId'**].**value\_counts**()**

filtered **=** ratings\_per\_movie**[**ratings\_per\_movie **>=** 10**]**

plt**.**figure**(**figsize**=(**10**,** 5**))**

sns**.**histplot**(**filtered**,** bins**=**50**,** log\_scale**=True)**

plt**.**title**(**'Number of Ratings per Movie (Log Scale X)'**)**

plt**.**xlabel**(**'Number of Ratings'**)**

plt**.**ylabel**(**'Number of Movies'**)**

plt**.**tight\_layout**()**

plt**.**show**()**

A graph of rating

AI-generated content may be incorrect.

Figure 7 The Visualized output of executing the ratings per movie

### **A.13. WordCloud of Movie Tags and Screenshot of the output**

# Word Cloud from Tags

text **=** ' '**.**join**(**tags**[**'tag'**].**dropna**().**astype**(str).**values**)**

wordcloud **=** WordCloud**(**width**=**1000**,** height**=**500**,** background\_color**=**'white'**,** colormap**=**'viridis'**).**generate**(**text**)**

plt**.**figure**(**figsize**=(**15**,** 7**))**

plt**.**imshow**(**wordcloud**,** interpolation**=**'bilinear'**)**

plt**.**axis**(**'off'**)**

plt**.**title**(**'Word Cloud of Movie Tags'**)**

plt**.**tight\_layout**()**

plt**.**show**()**

A close up of words

AI-generated content may be incorrect.

Figure 8 The output of executing world cloud and notice that the larger the text is, the more it is mentioned throughout the list of all the movies in the dataset

### **A.14. Installing the K-modes library and Preparing the data for clustering**

# Install kmodes if not already installed

**!**pip install kmodes

# Import necessary library for K-Modes clustering

**from** kmodes**.**kmodes **import** KModes

# Expand genres into individual rows per movie

movies\_expanded **=** movies**.**copy**()**

movies\_expanded**[**'genres'**]** **=** movies\_expanded**[**'genres'**].str.**split**(**'|'**)**

movies\_expanded **=** movies\_expanded**.**explode**(**'genres'**)**

# Create one-hot encoded genre columns using crosstab (more reliable than pivot\_table here)

genre\_ohe **=** pd**.**crosstab**(**movies\_expanded**[**'movieId'**],** movies\_expanded**[**'genres'**])**

# Reset index to bring 'movieId' back as a column

genre\_ohe**.**reset\_index**(**inplace**=True)**

# Merge one-hot genre data with movie titles

movies\_with\_titles **=** pd**.**merge**(**movies**[[**'movieId'**,** 'title'**]],** genre\_ohe**,** on**=**'movieId'**)**

# Prepare the features for K-Modes clustering (exclude ID and title)

X **=** movies\_with\_titles**.**drop**(**columns**=[**'movieId'**,** 'title'**])**

### **A.15. Applying the K-modes clustering and Screenshot of the results**

# Apply K-Modes clustering

**from** kmodes**.**kmodes **import** KModes

km **=** KModes**(**n\_clusters**=**5**,** init**=**'Huang'**,** n\_init**=**5**,** verbose**=**1**)**

clusters **=** km**.**fit\_predict**(**X**)**

# Assign clusters back to the main DataFrame

movies\_with\_titles**[**'Cluster'**]** **=** clusters

# View cluster distribution

**print(**"Number of movies per cluster:\n"**,** movies\_with\_titles**[**'Cluster'**].**value\_counts**().**sort\_index**())**

# Preview titles per cluster

**for** i **in** **range(**5**):**

**print(**f"\nSample movies from Cluster {i}:\n"**)**

**print(**movies\_with\_titles**[**movies\_with\_titles**[**'Cluster'**]** **==** i**][**'title'**].**head**(**10**))**

A screenshot of a computer program

AI-generated content may be incorrect.

Figure 9 The output of executing number of movie per K-mode clustering, notice that we had 5 iterations.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 10 The output of the movie sample from the First and Second clustering

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 11 The output of the movies sample with a title from the third and fourth clustering

### **A.16. Installing the MCA Library for 2D Visualization of Clusters and Screenshot of Output**

# Install the MCA package (prince)

**!**pip install prince

# Import MCA from prince

**import** prince

**import** matplotlib**.**pyplot **as** plt

**import** seaborn **as** sns

# STEP 6: Visualize Clusters using MCA

# Re-apply MCA to the feature space (X)

# NOTE: X is one-hot encoded genre data

mca **=** prince**.**MCA**(**n\_components**=**2**,** random\_state**=**42**)**

mca\_coords **=** mca**.**fit\_transform**(**X**)**

# Add MCA components and cluster labels to a new DataFrame

mca\_df **=** mca\_coords**.**copy**()**

mca\_df**[**'Cluster'**]** **=** movies\_with\_titles**[**'Cluster'**]**

# Plot clusters in 2D MCA space

plt**.**figure**(**figsize**=(**10**,** 6**))**

sns**.**scatterplot**(**data**=**mca\_df**,** x**=**0**,** y**=**1**,** hue**=**'Cluster'**,** palette**=**'tab10'**,** s**=**60**,** alpha**=**0.7**)**

plt**.**title**(**'K-Modes Clustering of Movies (MCA-reduced)'**)**

plt**.**xlabel**(**'MCA Component 1'**)**

plt**.**ylabel**(**'MCA Component 2'**)**

plt**.**legend**(**title**=**'Cluster'**)**

plt**.**grid**(True)**

plt**.**tight\_layout**()**

plt**.**show**()**

A diagram with many colored dots

AI-generated content may be incorrect.

Figure 12 The output of the 2D visualization of the k mode clustering