BA820 – Project M2

Analyzing Alone TV Show

Section and Team Number: Section B1, Team B02

Student Name: Burak Ataseven

1. Refined Problem Statement & Focus

The fundamental research question I focused on within this milestone is: “Can clustering identify content segments?”

However, early exploratory data analysis conducted in M1 provided a clearer understanding of the data structure and variable characteristics, making the problem framework more focused. Specifically, episode-based performance metrics (such as viewers, IMDb rating, and rating count) were found to be suitable indicators for content segmentation. Therefore, the analysis shifted towards segmentation based on episode performance and viewer engagement rather than content types.

One of my initial assumptions was that episodes could be naturally segmented based on viewing and rating metrics. The EDA results partially supported this assumption. Specifically, some episodes showed higher performance in both viewing and rating, supporting the existence of a “high-performing content” segment. However, it was also observed that not all episodes were clearly and sharply grouped, suggesting that segments may partially overlap.

The clustering method is suitable for this problem because:

- There are no predefined segments in the data.
- Natural groups are sought based on content performance.
- Numerical variables such as views, scores, and ratings are suitable for analyzing similarity structures.

Therefore, content segments based on the section were attempted to be determined using the hierarchical clustering method.

2. EDA & Preprocessing: Updates

The dataset consists of four main tables: episodes, seasons, contestants, and equipment. Clustering analysis focuses specifically on numerical variables at the episode level. The episode dataset includes performance metrics such as season, episode number, viewership, IMDb score, and rating.

The main numerical variables used for clustering are:

- season
- episode
- viewers
- imdb_rating

- n_ratings

During the EDA process, some missing values were observed in the dataset. Specifically, there were missing observations in the viewers, imdb_rating, and n_ratings variables. These missing values were filled with the median value to allow the clustering algorithm to function. This approach was preferred to avoid being affected by extreme values and to maintain data distribution.

Also, since the scales of different variables differ, normalization was applied using StandardScaler in some analyses. This process is necessary to prevent clustering results from being determined solely by large-scale variables.

The general data analysis steps performed in M1 have been largely preserved. However, in M2, the variable selection was narrowed down for clustering purposes, and only the numerical variables necessary for the analysis were used. These preprocessing steps enabled the clustering algorithm to produce more meaningful and balanced results.

3. Analysis & Experiments

The main method used in this study is Hierarchical Clustering (Ward linkage, default method). This method is suitable for discovering naturally occurring similarity groups within the data and does not require pre-defined labels. It also helps to identify content segments by analyzing the similarity structure between numerical variables.

Why were this method chosen?

- The aim is to discover the similarity structure between segments.
- The Ward linkage method creates balanced clusters by minimizing variance.
- The optimal number of clusters can be visually examined using a dendrogram.

The following tests were conducted during the analysis process:

- Hierarchical clustering with n_clusters = 3
- Hierarchical clustering with n_clusters = 4
- Hierarchical clustering with n_clusters = 5
- Comparison of data with and without normalization
- Dendrogram analysis
- 2D visualization (viewers vs imdb_rating)
- 3D visualization (viewers, imdb_rating, n_ratings)

Results with n_clusters = 3

When the dendrogram was examined, it was observed that there are approximately three main clusters in the data structure. Since these clusters merge at high distances, there is a certain difference between them.

- One cluster includes episodes with high viewership, high IMDb scores, and high ratings. This cluster represents popular episodes.
- Another cluster represents lower-performing or niche episodes.
- The middle cluster overlaps with the other two clusters and is not clearly separated.

The three-cluster structure generally captures the main content segments in the dataset. However, only the high-performance cluster stands out strongly.

Results with n_clusters = 4

- Two clusters represented mid-performing segments and overlapped significantly.
- One cluster included high-watching but relatively lower-rated segments.
- One cluster included low-watching but relatively high-rated segments.
- Some clusters represented small and niche segments.

The overlap between clusters increased even further when normalized data was used. This indicates that it is difficult to divide the dataset into four clear segments.

Results with n_clusters = 5

Visualization results:

- One cluster represents high-performing segments.
- One cluster represents low-performing segments.
- Several clusters contain medium-performing segments and overlap significantly.
- One cluster is close to average values and has a mixed structure.

Cluster sizes have become unbalanced, and some clusters remain very small.

What worked?

- High-performing segments were clearly identified as distinct segments.
- Overall differences were observed between popular, average, and niche segments.
- It was understood that view count, IMDb score, and rating number determined segmentation.

What didn't work/challenges?

- Mid-range segments consistently overlapped.
- Segregation weakened as the number of clusters increased.
- Some clusters remained too small to be interpreted.

4. Findings & Interpretations

Cluster analysis shows that episodes can be segmented into specific content categories based on performance metrics. The most prominent segment consists of popular episodes with high viewership and high IMDb ratings. These episodes represent the highest viewer engagement.

Lower-performing or niche episodes form a separate segment. These episodes may appeal to a smaller audience. Mid-performing episodes form a broad segment and overlap with other clusters.

When comparing the number of clusters:

- Clusters 3 offer the clearest and most interpretable structure. (Figure 1- Pair Plot n_clusters=3 (normalized))
- Clusters 4 offer more detail but increase overlap. (Figure 8- Pair Plot n_clusters=4 (normalized))
- Clusters 5 create too much overlap and make interpretation more difficult. (Figure 12- Pair Plot n_clusters=5 (normalized))

In general, as the number of clusters increases, the differentiation weakens and analytical clarity decreases. Therefore, a three-cluster structure is considered the most meaningful content segmentation.

5. Next Steps

Things not yet done:

- Alternative clustering methods
- Cluster quality measurements
- Further feature engineering

Future analyses:

- Comparison with K-means clustering

Unanswered questions:

- Do content segments vary by season?
- Do some seasons perform better than others?

Current findings indicate that content segmentation can be partially identified using clustering. Future analyses will focus on making segmentation more precise.

Appendix

Shared GitHub Repository (Required)

- The **shared team GitHub repository**
- Ataseven Burak Project M2.docx and M2Alone_TV.ipynb

Supplemental Material (Highly Recommended)

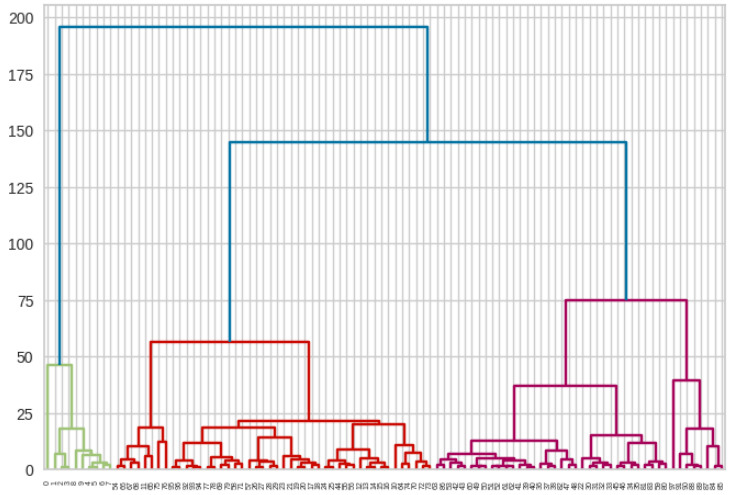

Figure 1- Dendrogram

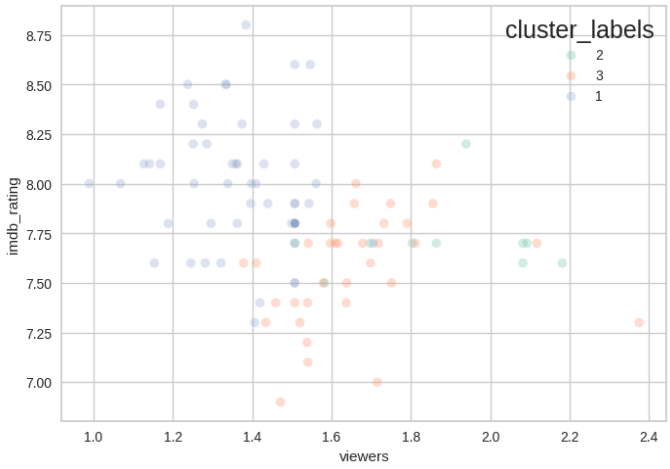

Figure 2- Scatter Plot n_clusters=3 (nonnormalized)

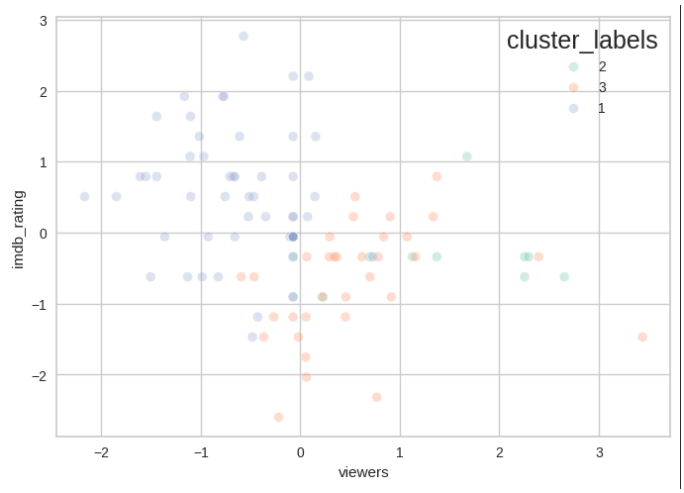

Figure 3- Scatter Plot n_clusters=3 (normalized)

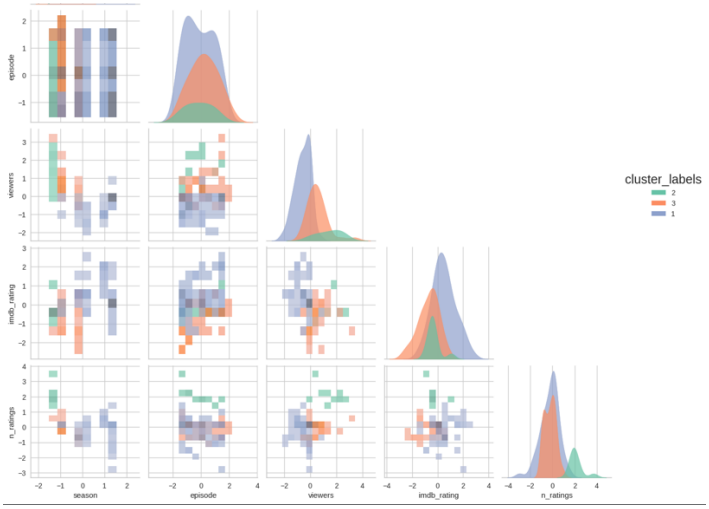

Figure 4- Pair Plot n_clusters=3 (normalized)

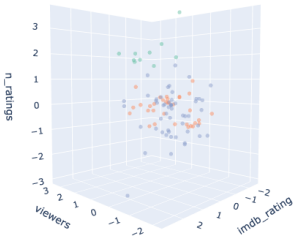

Figure 5- 3D Scatter Plot n_clusters=3 (normalized)

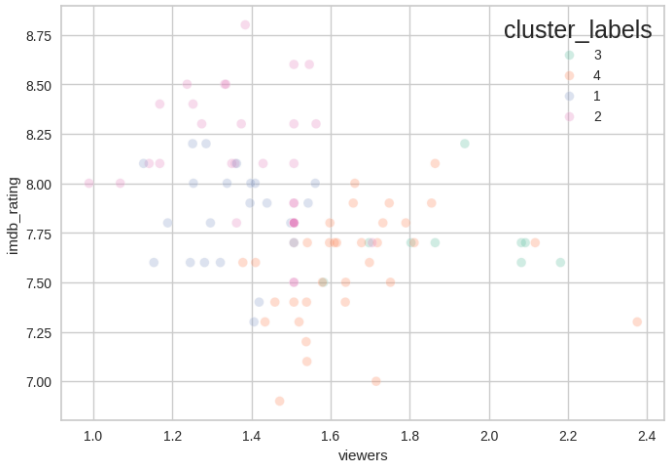

Figure 6- Scatter Plot n_clusters=4 (nonnormalized)

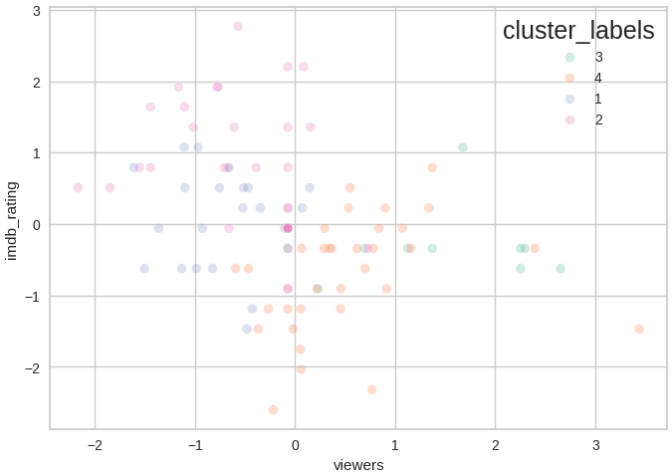

Figure 7- Scatter Plot n_clusters=4 (normalized)

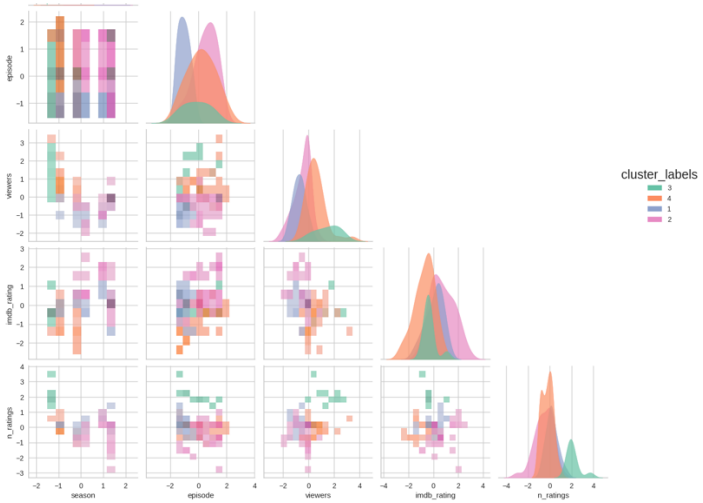

Figure 8- Pair Plot n_clusters=4 (normalized)

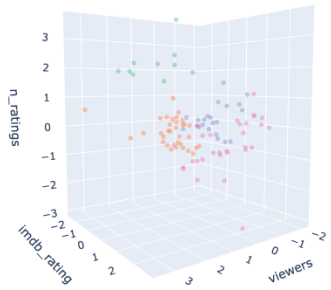

Figure 9- 3D Scatter Plot n_clusters=4 (normalized)

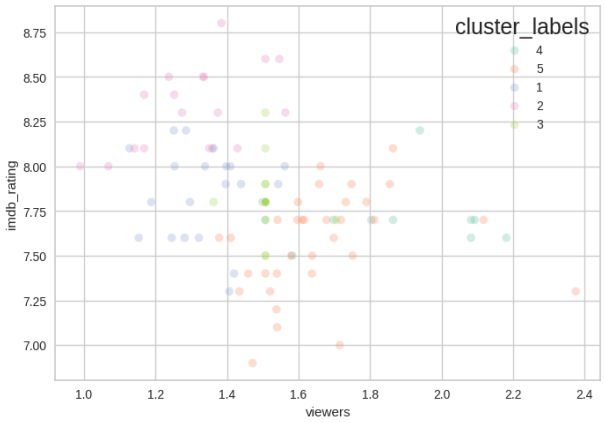

Figure 10- Scatter Plot n_clusters=5 (nonnormalized)

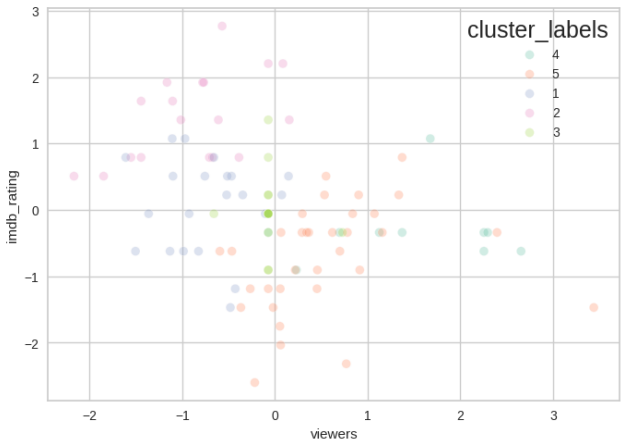

Figure 11- Scatter Plot n_clusters=5 (normalized)

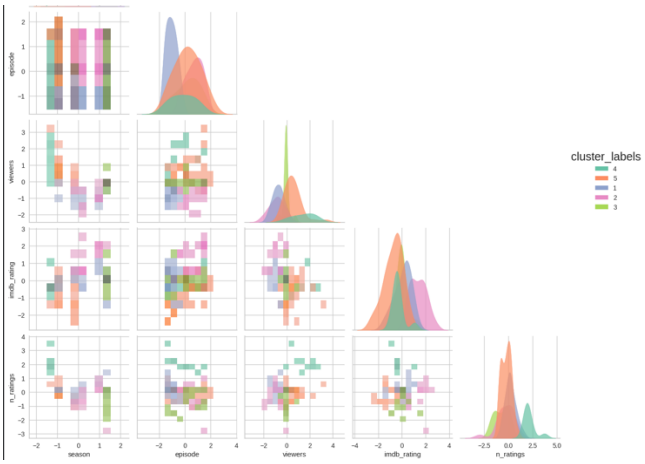

Figure 12- Pair Plot n_clusters=5 (normalized)

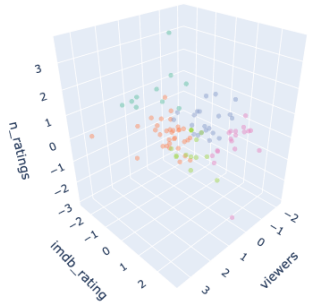

Figure 13- 3D Scatter Plot n_clusters=5 (normalized)

Process Overview

The analytical process follows these steps: data loading, data cleaning, feature selection, feature scaling, hierarchical clustering, visualization of results, and interpretation of insights.

Use of Generative AI Tools

https://chatgpt.com/share/6988f084-ce04-800c-9ca7-3e7236c11a0f

Asked about how to analyze figures (especially 3D). Reason is, I did not want to do any mistakes when interpreting the results (as you know the most important specification for Data Analyst). Also used the Gemini tool in the Google Collab notebook for fixing the clustering codes but couldn't upload the text log. Mostly used Basic_Clustering.ipynb and Bank_Transaction_Analysis.ipynb file (lecture content) when creating clustering settings.