BA820 - M2: Early Analysis & Insights

Team B02

Akbar Wibowo

1. Refined Problem Statement & Focus

1.1 Problem Statement

This milestone investigates the question: "Are IMDb ratings mostly consistent, or are there clear high- and low-rated episodes or seasons?" The objective is to identify latent structure in audience reception of the television series Alone by examining IMDb ratings across episodes and seasons. Rather than relying on simple summary statistics, the analysis evaluates whether ratings exhibit distinct regimes (e.g., consistently higher vs lower-rated content) across the show's timeline and season structure, suggesting meaningful differences in perceived quality. Key stakeholders include network or streaming and distribution teams for promotion and positioning, production and content strategy teams tasked with identifying recurring characteristics of highly rated content, and marketing teams targeting standout seasons or episodes for campaigns.

1.2 What Changed Since Milestone 1

The original project questions were primarily descriptive, emphasizing contestant attributes and survival outcomes through basic distributions and comparisons. Those questions did not explicitly leverage unsupervised learning to uncover deeper structure. Since Milestone 1, the scope has shifted toward understanding latent patterns in audience reception, specifically whether IMDb ratings separate into distinct groupings at the episode and/or season level. Preliminary EDA suggested non-trivial variation in ratings across the series, motivating the use of unsupervised methods to test whether this variation forms coherent clusters rather than reflecting random noise.

1.3 Assumptions Validated and Challenged

Initial assumptions treated IMDb ratings as relatively stable across the series. Early exploration indicated that ratings vary across episodes and seasons, but it remains unclear whether the variability is random or reflects consistent, interpretable segmentation. This milestone therefore shifts from assuming overall consistency to explicitly evaluating, through unsupervised structure discovery, whether identifiable rating regimes emerge across seasons and episodes.

2. EDA & Preprocessing: Updates

2.1 Relevant Findings from M1

IMDb ratings are available for 93 of 98 U.S. episodes, with five missing values (5.1%). Ratings are generally strong and moderately concentrated (mean = 7.82; median = 7.80; interquartile range = 7.6–8.1), while the full range spans 6.9 to 8.8. This combination—overall stability with meaningful extremes—motivates assessing whether ratings reflect coherent groupings across seasons or episodes rather than random fluctuation.

Episode-level inspection further suggests non-random structure. The lowest-rated episodes are concentrated in early Season 4 (episodes 1–4 rated 6.9–7.2), whereas several of the highest-rated episodes occur in Seasons 6–7 (e.g., Season 6, Episode 11 rated 8.8; multiple

Season 7 episodes rated 8.5–8.6). These concentrated pockets of low and high ratings provide preliminary evidence that audience reception may differ systematically across the series.

2.2 Changes Made in M2

To support season- and episode-level comparisons, analysis was restricted to the U.S. version of the dataset (version = 'US'). The air_date field was parsed into datetime format with no parsing failures, enabling chronological checks where appropriate. Data integrity checks confirmed no duplicate identifiers at the (season, episode) level. For rating-focused EDA, episodes with missing IMDb ratings were excluded (93 retained). Vote counts (n_ratings) were retained to contextualize extreme ratings; only one episode had fewer than 20 votes (Season 9, Episode 11), and despite a high rating (8.6), the low vote volume suggests this observation should be interpreted cautiously. A season × episode rating heatmap was also evaluated; however, differences in episode counts across seasons produce an irregular grid with systematic missing cells, and the relatively narrow rating range (6.9–8.8) limits visual contrast. Consequently, the heatmap was not retained as a primary EDA artifact in favor of season-level summaries and clustering-based segmentation.

3. Analysis & Experiments

3.1 Episode-Level Segmentation with K-Means (IMDb Rating + Vote Volume)

K-means clustering was applied to evaluate whether IMDb ratings for Alone are broadly consistent or whether episodes separate into distinct reception segments. This directly supports the targeted question by testing whether episodes form separable groups in a reception-oriented feature space rather than clustering tightly around a single typical value (Appendix Figure A1).

Episode-level clustering used imdb_rating as a measure of audience evaluation and n_ratings as a proxy for vote volume and rating stability. Features were standardized prior to clustering to ensure balanced distance computations. Candidate cluster counts were evaluated using the elbow plot (distortion/SSE) and silhouette visualization (Appendix Figures A4–A5). While the elbow curve suggested additional granularity beyond three clusters, k = 3 was selected to preserve interpretability and align with the objective of identifying clear high- and low-reception groupings.

The k-means solution produced three interpretable segments (Appendix Table B5). One cluster represented a lower-rated regime, including the most consistently weak pocket observed in earlier exploration. A second cluster represented a higher-rated regime and contained multiple standout episodes. A third cluster was distinguished primarily by substantially higher vote volume rather than rating extremes, indicating that latent structure is not exclusively driven by rating level but also reflects engagement or stability differences (Appendix Figure A6). Representative episodes within each cluster further support interpretability and provide concrete examples of “typical” and “extreme” episodes within each segment (Appendix Table B6).

Season-by-cluster cross-tabulations indicated non-random concentration across seasons (Appendix Tables B7–B8). Certain seasons concentrated heavily within the lower-rated cluster, while Season 7 concentrated within the higher-rated cluster, consistent with systematic differences in audience reception across production cycles. A key interpretive constraint is that extreme ratings are not equally reliable when vote counts are low; one highly rated episode has a very small number of votes and is therefore treated as a potentially unstable extreme (Appendix Table B4). Retaining n_ratings alongside ratings supports more credible interpretation of outliers and prevents over-weighting sparsely supported observations.

3.2 Hierarchical Clustering and Robustness Experiments (Episodes and Seasons)

Hierarchical clustering was applied as a complementary approach to assess whether similar structure emerges under a different clustering paradigm and to support season-level grouping. Ward linkage was applied to standardized features to prevent scale-dominant variables from driving merges. Dendrograms provide a multiscale representation of similarity and enable cluster selection through tree cuts rather than reliance on a single fixed partition (Appendix Figures A3 and A7).

At the episode level, hierarchical clustering on standardized imdb_rating and n_ratings yielded groupings broadly consistent with the k-means segmentation, supporting the conclusion that the presence of higher and lower-reception segments is not method-dependent (Appendix Figure A3). At the season level, episodes were aggregated into season profiles (mean_rating, std_rating, min_rating, max_rating, n_episodes, and mean_votes) and clustered after standardization. This season-profile approach aligns directly with the business question because it groups seasons according to reception characteristics rather than episode-level variability (Appendix Table B2; Appendix Figure A7).

To deepen the investigation, a robustness experiment used an expanded episode feature set (season, episode, viewers, imdb_rating, n_ratings) with median imputation for missing numeric values. Hierarchical clustering was repeated for k = 3, 4, and 5 to evaluate sensitivity to segmentation granularity. Results exhibited a stable small cluster across k values, while increased k primarily subdivided the largest cluster into medium-sized subgroups (Appendix Figures A8–A10; Appendix Table B9). This pattern suggests that core segmentation is not an artifact of a single arbitrary choice of k, although inclusion of season and episode introduces chronology as a potential confound. Accordingly, expanded-feature results are interpreted as a sensitivity check rather than the primary specification for reception segmentation.

Collectively, the analyses indicate that IMDb ratings are not fully uniform across Alone: episodes and seasons exhibit identifiable groupings consistent with lower- and higher-reception regimes, while vote volume contributes an additional latent dimension relevant to interpretation of episode-level extremes and the stability of observed ratings (Appendix Table B5; Appendix Figure A6).

4. Findings & Interpretations

IMDb ratings for Alone are generally strong, but they are not fully uniform across the series. Ratings cluster within a relatively narrow band overall (Appendix Figure A1), yet meaningful variation exists across seasons (Appendix Figure A2; Appendix Table B2). This indicates that audience reception is broadly positive, while still containing identifiable high- and low-reception pockets rather than purely random episode-to-episode noise.

At the episode level, the analysis separates content into distinct reception segments. One segment corresponds to comparatively lower-rated episodes, including the most consistently weak pocket observed in the exploratory results—early Season 4 episodes appear repeatedly among the lowest-rated content (Appendix Table B3) and concentrate heavily in the lower-reception segment (Appendix Table B8). In contrast, a higher-reception segment contains multiple standout episodes and is strongly represented by Season 7 (Appendix Table B7; Appendix Table B8). Taken together, these season-level concentrations support the conclusion that certain production cycles delivered a more consistently strong viewer experience than others.

A second key insight is that audience engagement differs across episodes and should be considered when interpreting ratings. A distinct group of episodes is characterized by substantially higher vote volume, even when average ratings are not extreme (Appendix Table B5; Appendix Figure A6). This suggests that some episodes function as broader "attention drivers," and their ratings may be more stable indicators of audience reception due to higher participation. Conversely, a small number of high-rated episodes are supported by very few votes; for example, one highly rated episode in Season 9 has an unusually low vote count (Appendix Table B4). As a result, extreme values are more credible when supported by typical or high voting volume rather than sparse participation.

Robustness checks reinforce these interpretations. Hierarchical clustering and sensitivity experiments (varying the number of groups) consistently preserved a small, distinct segment across specifications, while higher granularity primarily subdivided the largest "typical" group into smaller subgroups (Appendix Figures A8–A10; Appendix Table B9). This pattern suggests that the core segmentation is not driven by an arbitrary choice of grouping level, but reflects persistent differences among subsets of episodes. For stakeholders, these findings have direct relevance: higher-reception segments and consistently strong seasons provide candidates for promotion and discovery placement, while the lower-reception pocket provides a diagnostic signal for production and content strategy teams seeking to understand what differentiated weaker periods of the series (Appendix Table B8).

5. Next Steps

Several items remain incomplete. The current work identifies episode and season-level rating segments but does not yet incorporate explanatory content features to determine why certain periods are rated higher or lower. Next steps include separating reception metrics (ratings, vote volume) from exposure/chronology effects, testing alternative clustering specifications for stability, and further examining the engagement-driven segment. Open questions include whether the high–vote volume cluster reflects true salience versus voting dynamics and what factors explain seasons concentrated in higher- versus lower-reception groups (Appendix Tables B7–B8). These extensions are justified because segmentation is non-random and engagement appears to shape grouping structure (Appendix Table B5; Appendix Figure A6).

Appendix

Shared GitHub Repository

https://github.com/MikaIsmayilov/B02_Alone_TVShow_Unsupervised_Project/tree/main/Individual/Akbar

Appendix A: Figures

Figure A1. Distribution of episode IMDb ratings (histogram + boxplot)

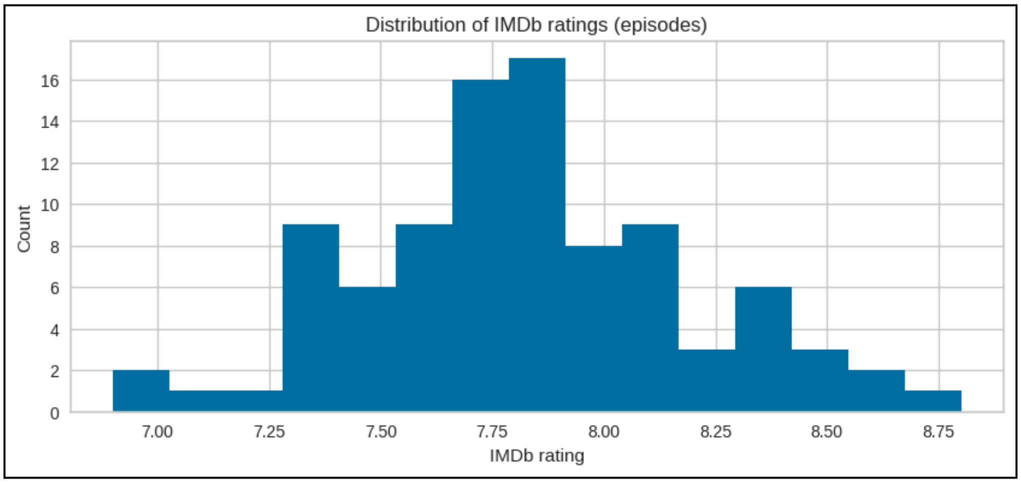


Figure A2. IMDb ratings by season (season-level boxplot)

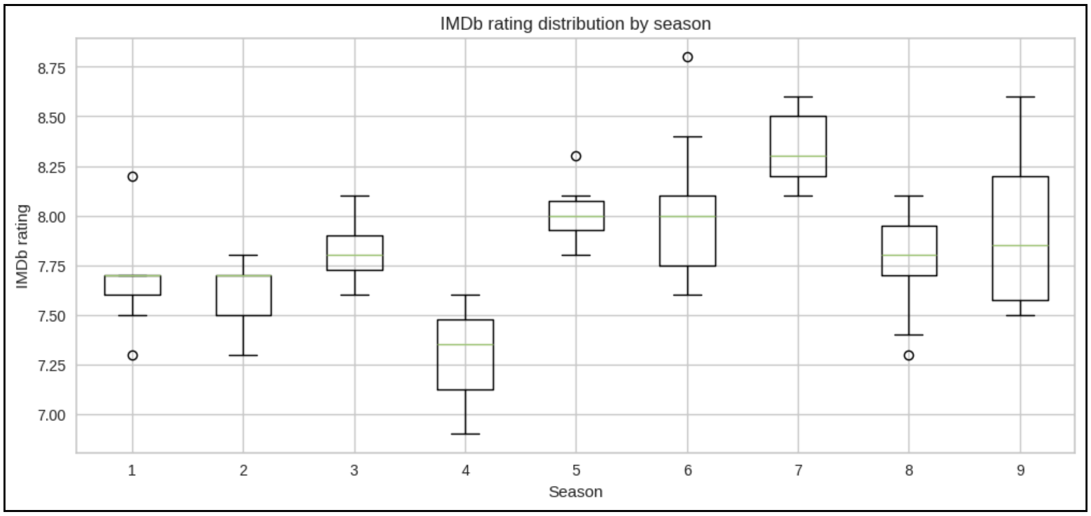


Figure A3. Hierarchical clustering dendrogram (episodes; Ward; standardized features)

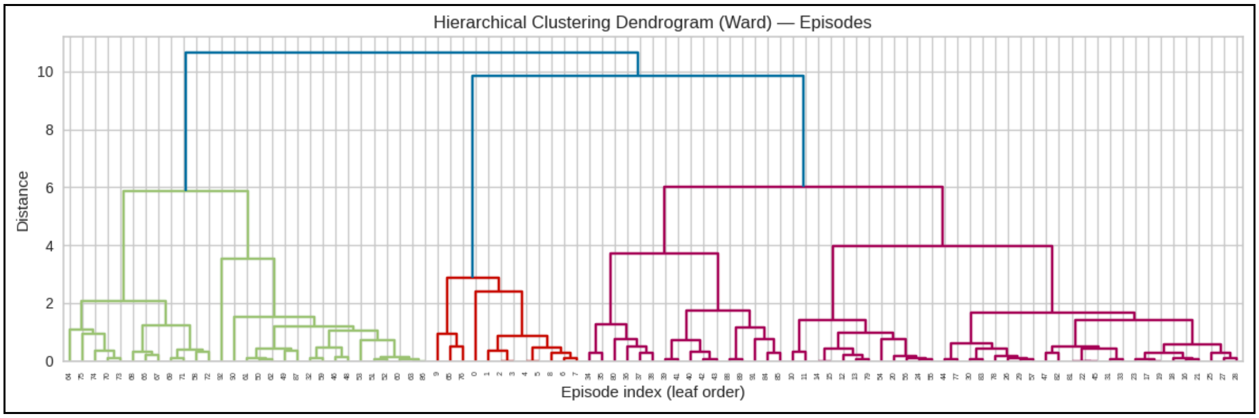


Figure A4. K-means elbow plot (distortion/SSE)

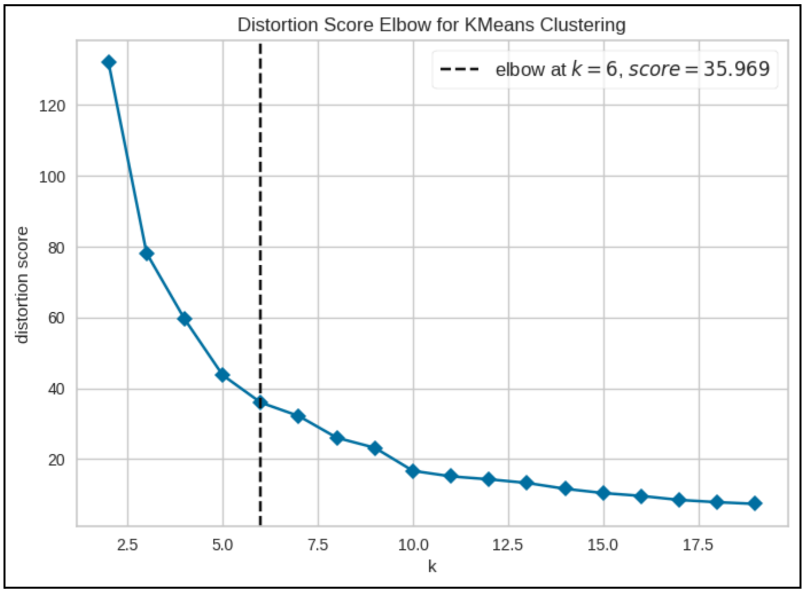


Figure A5. K-means silhouette plot (selected k)

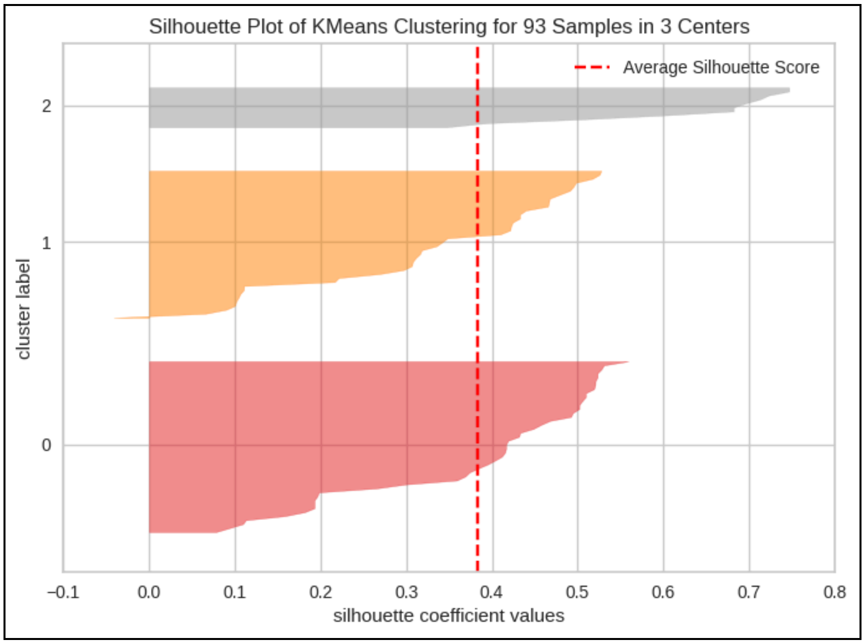


Figure A6. Scatter: n_ratings vs imdb_rating (episode-level)

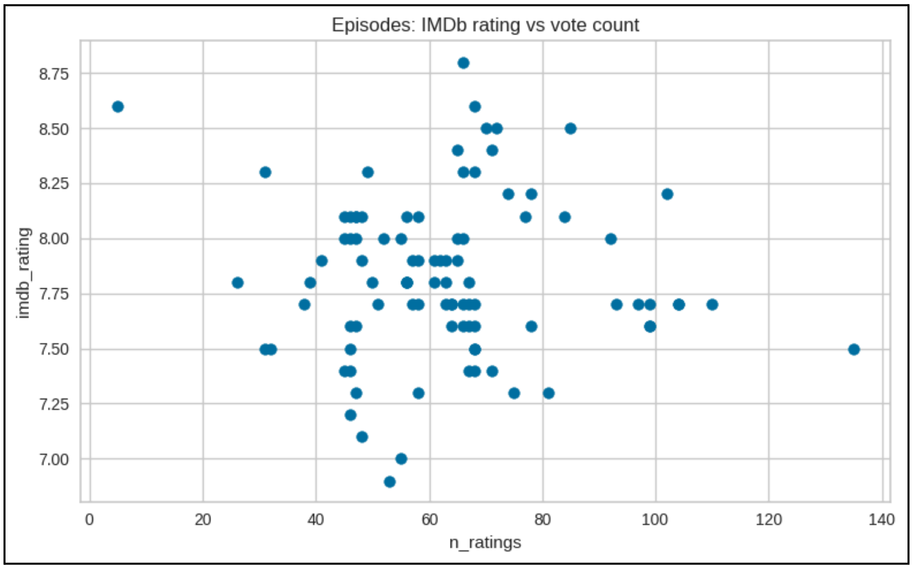


Figure A7. Hierarchical clustering dendrogram (seasons; Ward; standardized season profiles)

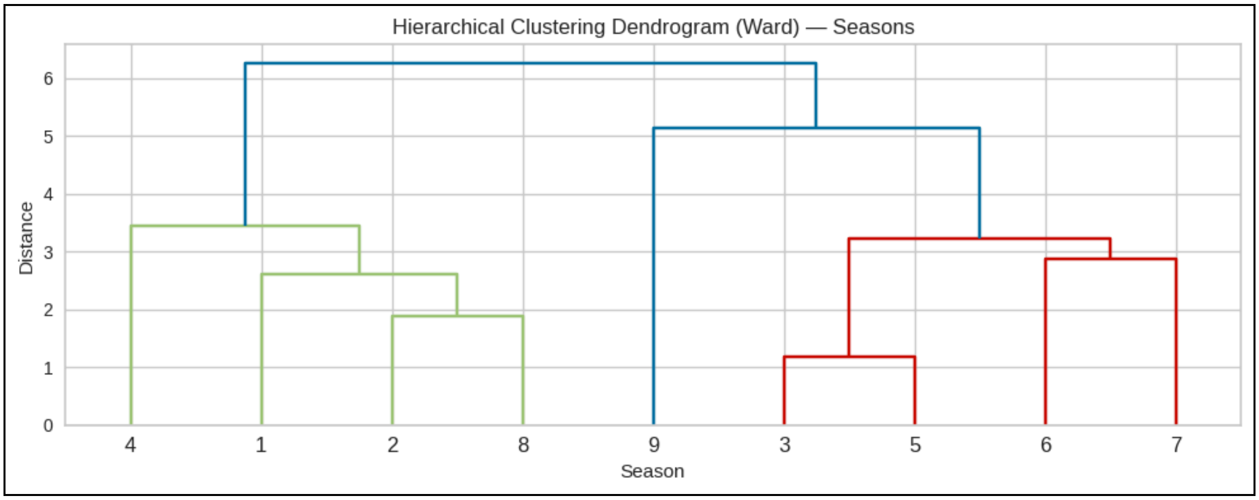


Figure A8. Sensitivity experiment: hierarchical clusters k=3 (expanded features) — viewers vs rating scatter

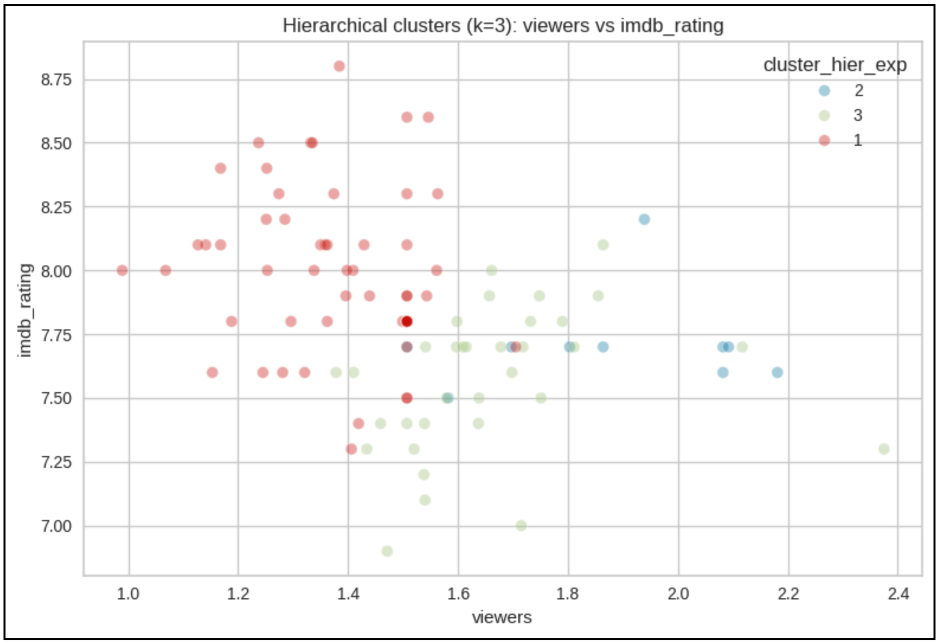


Figure A9. Sensitivity experiment: k=4 — viewers vs rating scatter

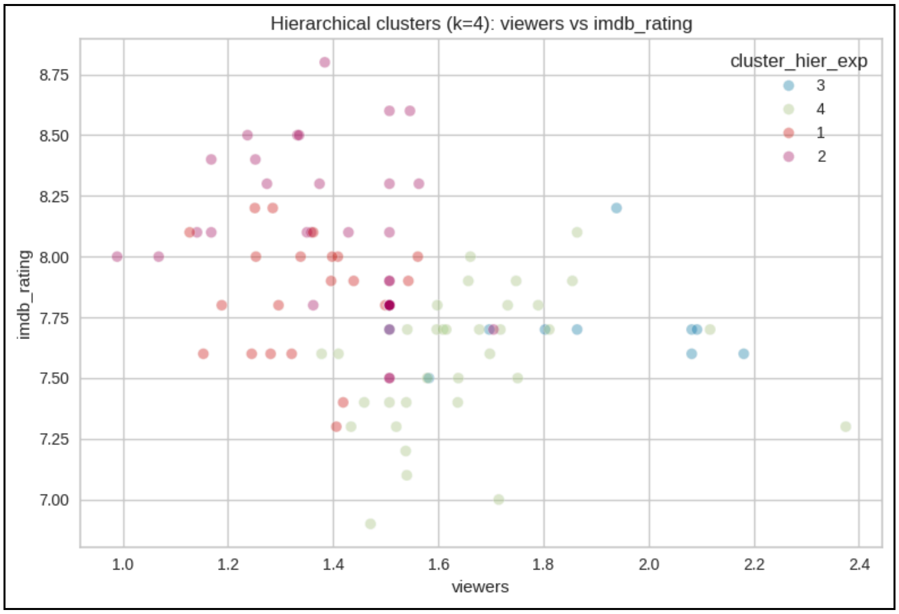
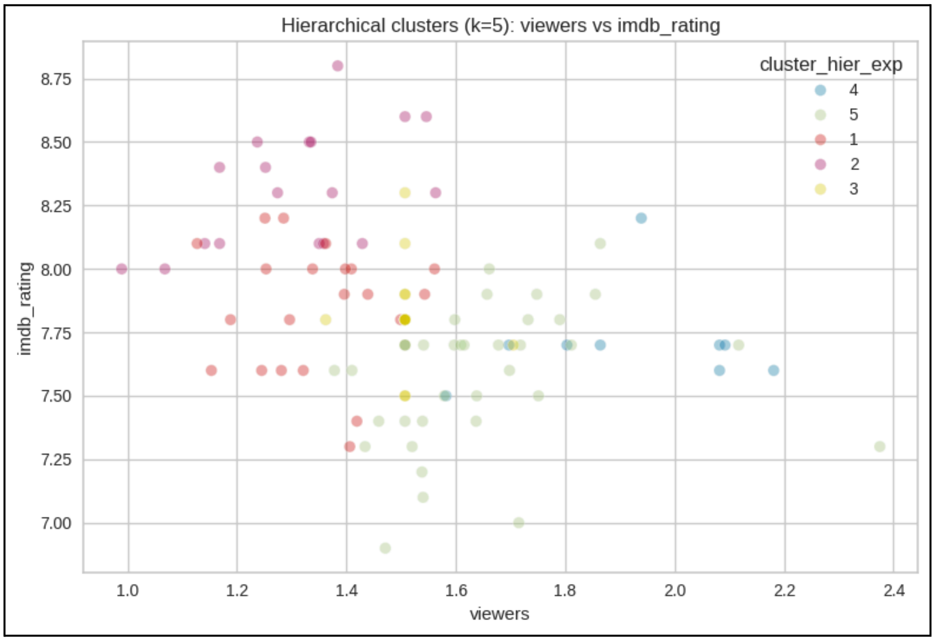
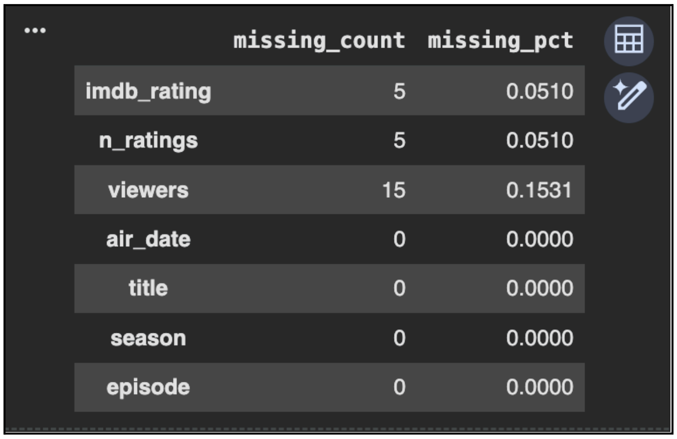


Figure A10. Sensitivity experiment: k=5 — viewers vs rating scatter



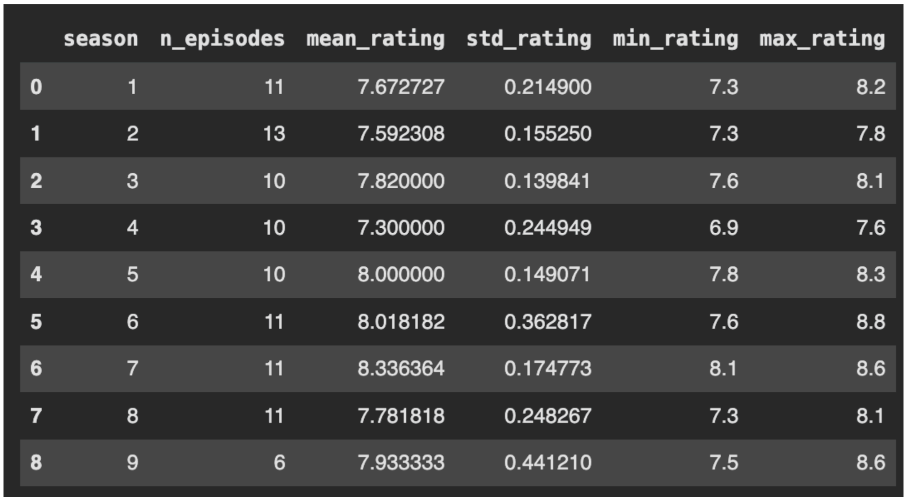
Appendix B: Tables

Table B1. Missingness summary for key fields (imdb_rating, n_ratings, viewers, air_date)



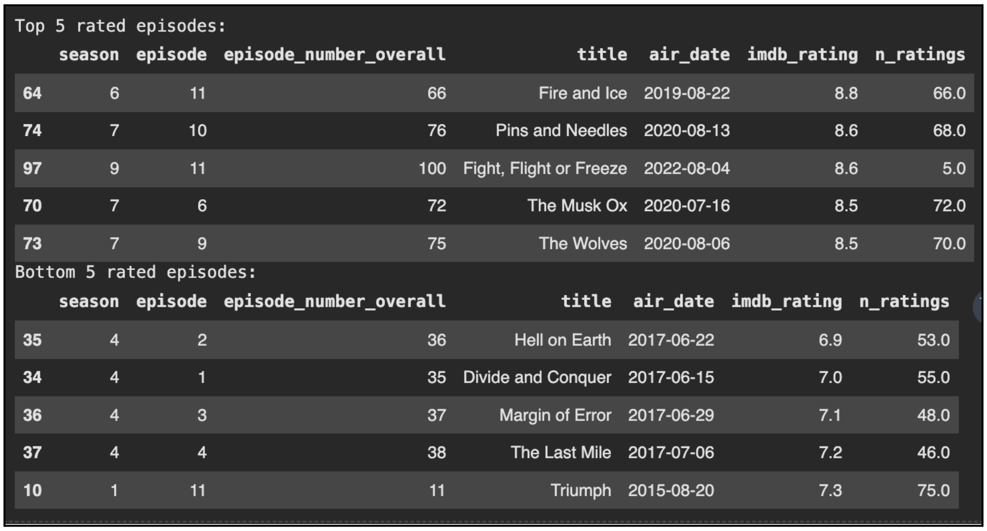
	missing_count	missing_pct
imdb_rating	5	0.0510
n_ratings	5	0.0510
viewers	15	0.1531
air_date	0	0.0000
title	0	0.0000
season	0	0.0000
episode	0	0.0000

Table B2. Season-level rating summary table (mean/std/min/max/count)



	season	n_episodes	mean_rating	std_rating	min_rating	max_rating
0	1	11	7.672727	0.214900	7.3	8.2
1	2	13	7.592308	0.155250	7.3	7.8
2	3	10	7.820000	0.139841	7.6	8.1
3	4	10	7.300000	0.244949	6.9	7.6
4	5	10	8.000000	0.149071	7.8	8.3
5	6	11	8.018182	0.362817	7.6	8.8
6	7	11	8.336364	0.174773	8.1	8.6
7	8	11	7.781818	0.248267	7.3	8.1
8	9	6	7.933333	0.441210	7.5	8.6

Table B3. Top 5 rated episodes + Bottom 5 rated episodes



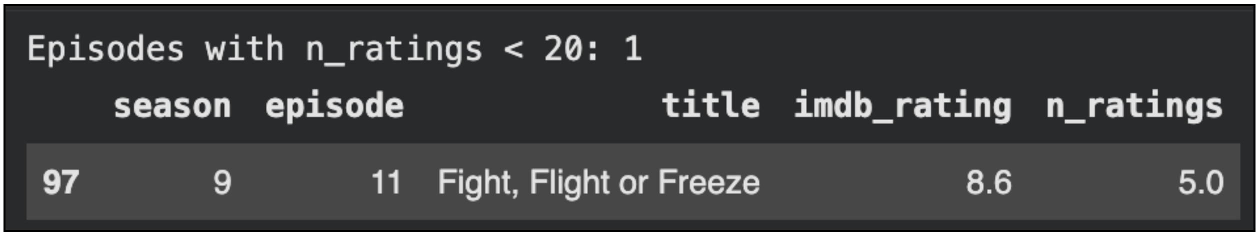
Top 5 rated episodes:

	season	episode	episode_number_overall	title	air_date	imdb_rating	n_ratings
64	6	11	66	Fire and Ice	2019-08-22	8.8	66.0
74	7	10	76	Pins and Needles	2020-08-13	8.6	68.0
97	9	11	100	Fight, Flight or Freeze	2022-08-04	8.6	5.0
70	7	6	72	The Musk Ox	2020-07-16	8.5	72.0
73	7	9	75	The Wolves	2020-08-06	8.5	70.0

Bottom 5 rated episodes:

	season	episode	episode_number_overall	title	air_date	imdb_rating	n_ratings
35	4	2	36	Hell on Earth	2017-06-22	6.9	53.0
34	4	1	35	Divide and Conquer	2017-06-15	7.0	55.0
36	4	3	37	Margin of Error	2017-06-29	7.1	48.0
37	4	4	38	The Last Mile	2017-07-06	7.2	46.0
10	1	11	11	Triumph	2015-08-20	7.3	75.0

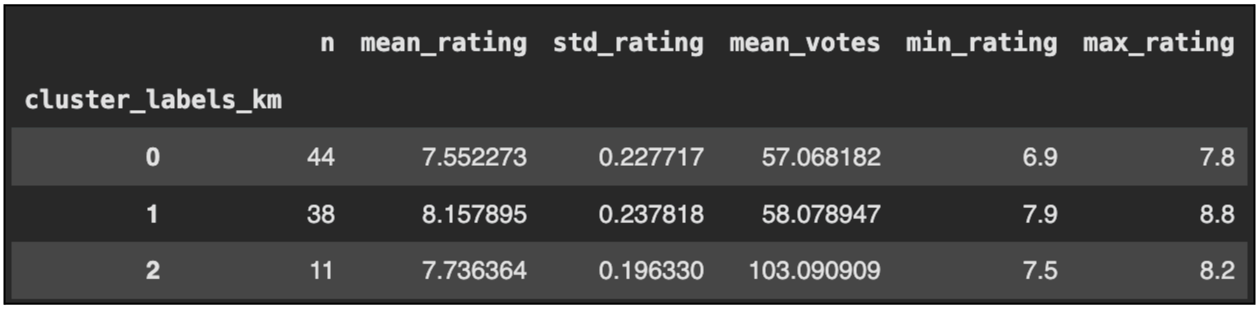
Table B4. Low-vote episodes (n_ratings < 20)



Episodes with n_ratings < 20: 1

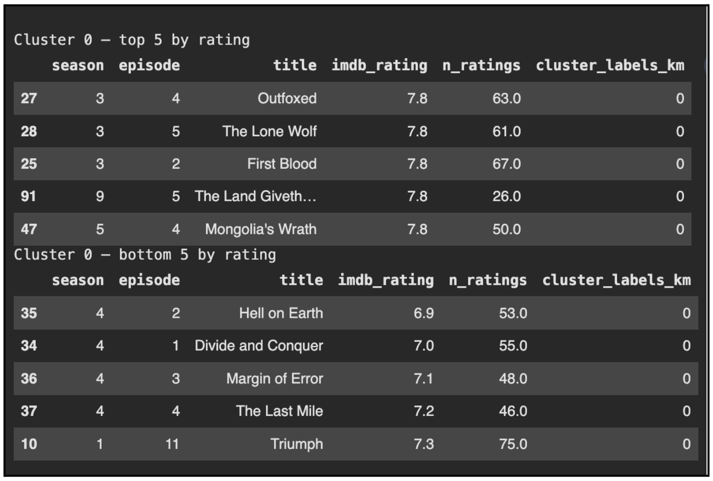
	season	episode	title	imdb_rating	n_ratings
97	9	11	Fight, Flight or Freeze	8.6	5.0

Table B5. K-means episode cluster summary (n, mean_rating, mean_votes, min/max rating)



	n	mean_rating	std_rating	mean_votes	min_rating	max_rating
cluster_labels_km						
0	44	7.552273	0.227717	57.068182	6.9	7.8
1	38	8.157895	0.237818	58.078947	7.9	8.8
2	11	7.736364	0.196330	103.090909	7.5	8.2

Table B6. Representative episodes per K-means cluster (top/bottom within cluster)

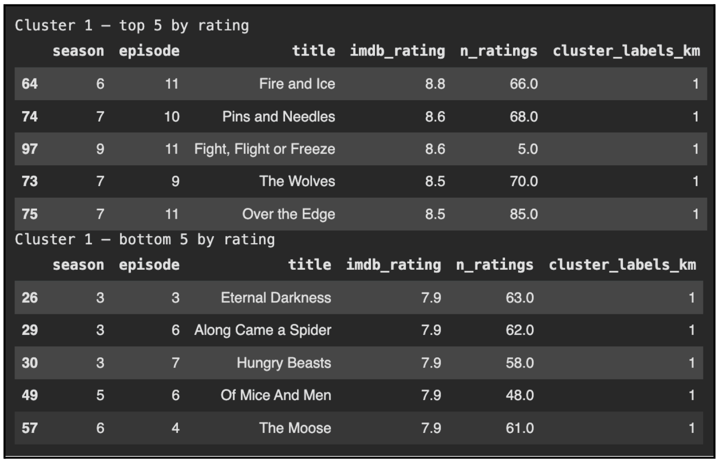


Cluster 0 — top 5 by rating

	season	episode	title	imdb_rating	n_ratings	cluster_labels_km
27	3	4	Outfoxed	7.8	63.0	0
28	3	5	The Lone Wolf	7.8	61.0	0
25	3	2	First Blood	7.8	67.0	0
91	9	5	The Land Giveth…	7.8	26.0	0
47	5	4	Mongolia's Wrath	7.8	50.0	0

Cluster 0 — bottom 5 by rating

	season	episode	title	imdb_rating	n_ratings	cluster_labels_km
35	4	2	Hell on Earth	6.9	53.0	0
34	4	1	Divide and Conquer	7.0	55.0	0
36	4	3	Margin of Error	7.1	48.0	0
37	4	4	The Last Mile	7.2	46.0	0
10	1	11	Triumph	7.3	75.0	0

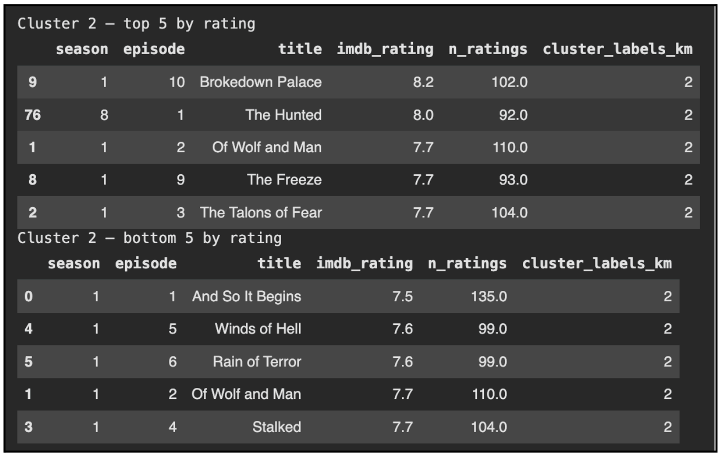


Cluster 1 — top 5 by rating

	season	episode	title	imdb_rating	n_ratings	cluster_labels_km
64	6	11	Fire and Ice	8.8	66.0	1
74	7	10	Pins and Needles	8.6	68.0	1
97	9	11	Fight, Flight or Freeze	8.6	5.0	1
73	7	9	The Wolves	8.5	70.0	1
75	7	11	Over the Edge	8.5	85.0	1

Cluster 1 — bottom 5 by rating

	season	episode	title	imdb_rating	n_ratings	cluster_labels_km
26	3	3	Eternal Darkness	7.9	63.0	1
29	3	6	Along Came a Spider	7.9	62.0	1
30	3	7	Hungry Beasts	7.9	58.0	1
49	5	6	Of Mice And Men	7.9	48.0	1
57	6	4	The Moose	7.9	61.0	1



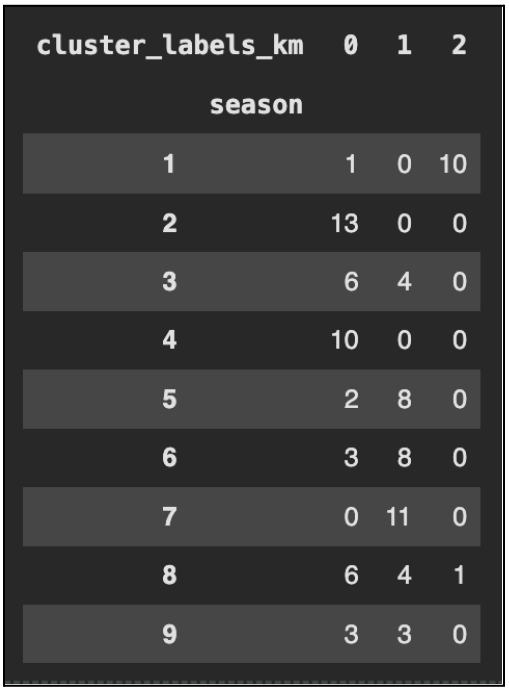
Cluster 2 — top 5 by rating

	season	episode	title	imdb_rating	n_ratings	cluster_labels_km
9	1	10	Brokedown Palace	8.2	102.0	2
76	8	1	The Hunted	8.0	92.0	2
1	1	2	Of Wolf and Man	7.7	110.0	2
8	1	9	The Freeze	7.7	93.0	2
2	1	3	The Talons of Fear	7.7	104.0	2

Cluster 2 — bottom 5 by rating

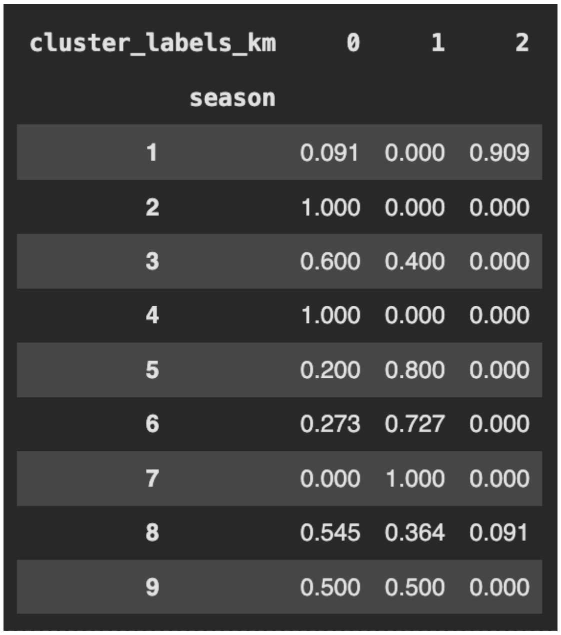
	season	episode	title	imdb_rating	n_ratings	cluster_labels_km
0	1	1	And So It Begins	7.5	135.0	2
4	1	5	Winds of Hell	7.6	99.0	2
5	1	6	Rain of Terror	7.6	99.0	2
1	1	2	Of Wolf and Man	7.7	110.0	2
3	1	4	Stalked	7.7	104.0	2

Table B7. Season × cluster crosstab (counts)



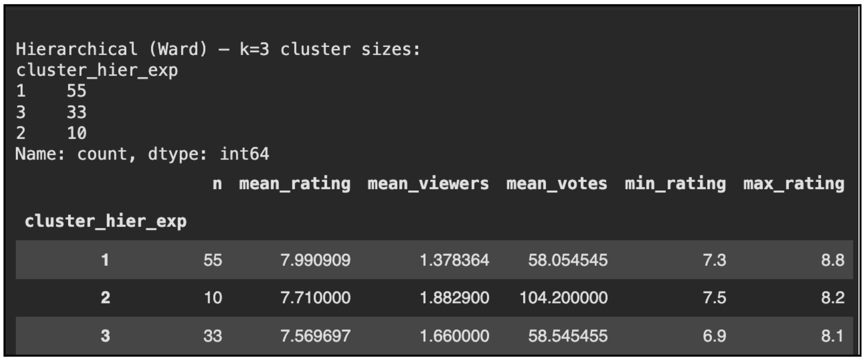
cluster_labels_km	0	1	2
season			
1	1	0	10
2	13	0	0
3	6	4	0
4	10	0	0
5	2	8	0
6	3	8	0
7	0	11	0
8	6	4	1
9	3	3	0

Table B8. Season × cluster crosstab (row-normalized percentages)



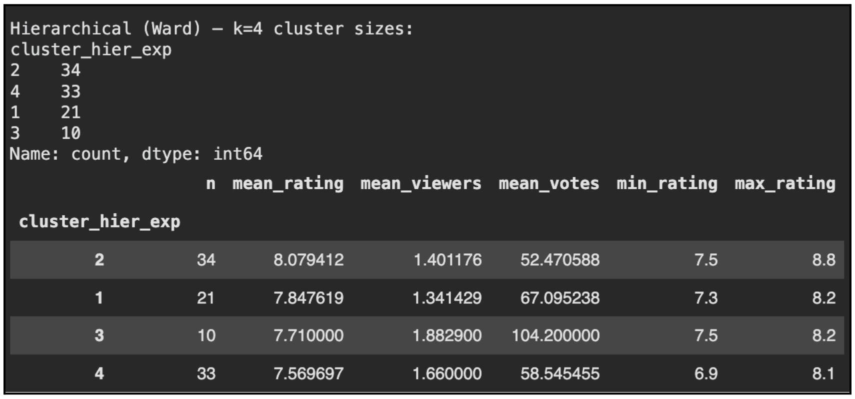
cluster_labels_km	0	1	2
season			
1	0.091	0.000	0.909
2	1.000	0.000	0.000
3	0.600	0.400	0.000
4	1.000	0.000	0.000
5	0.200	0.800	0.000
6	0.273	0.727	0.000
7	0.000	1.000	0.000
8	0.545	0.364	0.091
9	0.500	0.500	0.000

Table(s) B9. Sensitivity experiment cluster profiles (k=3/4/5)



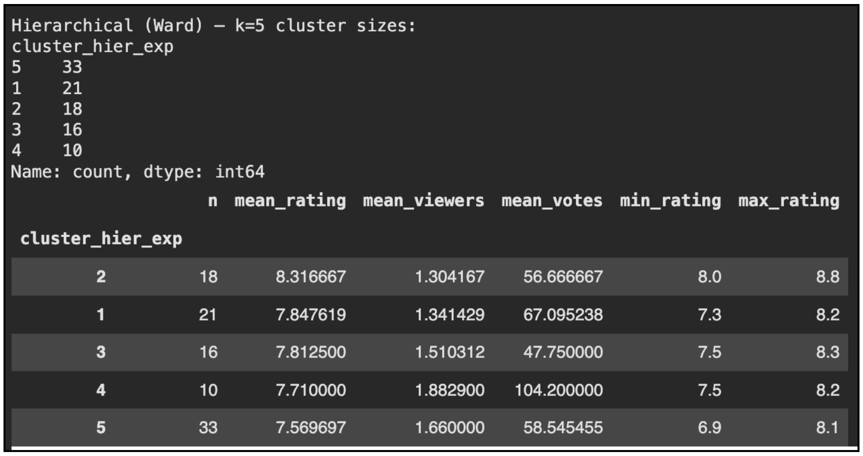
```
Hierarchical (Ward) — k=3 cluster sizes:
cluster_hier_exp
1    55
3    33
2    10
Name: count, dtype: int64
```

cluster_hier_exp	n	mean_rating	mean_viewers	mean_votes	min_rating	max_rating
1	55	7.990909	1.378364	58.054545	7.3	8.8
2	10	7.710000	1.882900	104.200000	7.5	8.2
3	33	7.569697	1.660000	58.545455	6.9	8.1



```
Hierarchical (Ward) — k=4 cluster sizes:
cluster_hier_exp
2    34
4    33
1    21
3    10
Name: count, dtype: int64
```

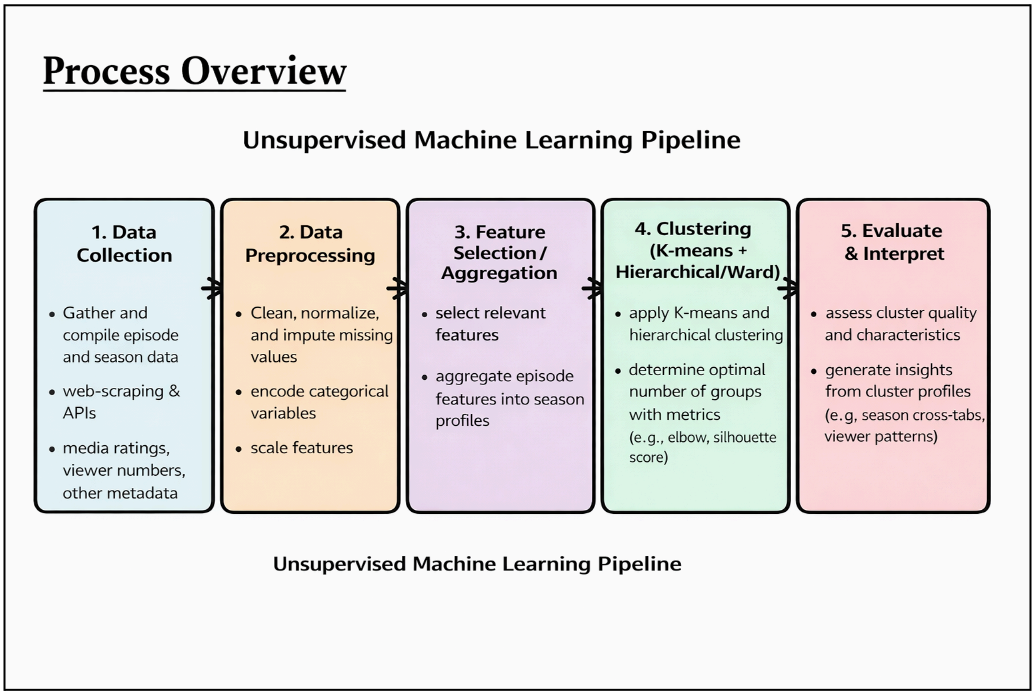
cluster_hier_exp	n	mean_rating	mean_viewers	mean_votes	min_rating	max_rating
2	34	8.079412	1.401176	52.470588	7.5	8.8
1	21	7.847619	1.341429	67.095238	7.3	8.2
3	10	7.710000	1.882900	104.200000	7.5	8.2
4	33	7.569697	1.660000	58.545455	6.9	8.1



```
Hierarchical (Ward) — k=5 cluster sizes:
cluster_hier_exp
5    33
1    21
2    18
3    16
4    10
Name: count, dtype: int64
```

cluster_hier_exp	n	mean_rating	mean_viewers	mean_votes	min_rating	max_rating
2	18	8.316667	1.304167	56.666667	8.0	8.8
1	21	7.847619	1.341429	67.095238	7.3	8.2
3	16	7.812500	1.510312	47.750000	7.5	8.3
4	10	7.710000	1.882900	104.200000	7.5	8.2
5	33	7.569697	1.660000	58.545455	6.9	8.1

Appendix C: Process Overflow



Process Overview

Unsupervised Machine Learning Pipeline

1. Data Collection

- Gather and compile episode and season data
- web-scraping & APIs
- media ratings, viewer numbers, other metadata

→

2. Data Preprocessing

- Clean, normalize, and impute missing values
- encode categorical variables
- scale features

→

3. Feature Selection / Aggregation

- select relevant features
- aggregate episode features into season profiles

→

4. Clustering (K-means + Hierarchical/Ward)

- apply K-means and hierarchical clustering
- determine optimal number of groups with metrics (e.g., elbow, silhouette score)

→

5. Evaluate & Interpret

- assess cluster quality and characteristics
- generate insights from cluster profiles (e.g, season cross-tabs, viewer patterns)

Unsupervised Machine Learning Pipeline

Appendix D: Use of Generative AI Tool

The usage of Generative AI Tools in this specific milestone (M2) would be ChatGPT. The chatbot's function was to rectify any coding errors and mistakes that occurred during the Exploratory Data Analysis and Unsupervised and Unstructured Methods experimentation done during sections two and three. The code was taken from the in-class exercise notebooks as both template and inspiration. Another Generative AI Tool I used was to guide me in uploading my files to the main, shared GitHub repository for the group that my teammate, Mika Ismayili, created.

BA820 M2 Code Editor and Help:
https://chatgpt.com/share/698aa0bb-60f8-8007-aded-26fa00cd533e

GitHub Folder Creation Guide:
https://chatgpt.com/share/698a7e69-7bb8-8002-a335-6ef5c1fe1376