BA820 – Project M4

- **Project Title: Alone TV Survivalist Profiling: What Types of People Survive Alone?**
- **Section and Team Number: Section B1, Team B02**
- **Student Name: Burak Ataseven**

1. Refined Problem Statement & Focus

The original M1 framing asked: What drives viewer engagement and episode quality in the Alone TV show? Early milestones explored this through episode-level clustering on IMDb ratings and viewership (M2), and through contestant gear and tap-out narrative analysis (M3).

M4 shifts the unit of analysis from episodes to people. The refined question is: can unsupervised clustering identify distinct archetypes of Alone contestants based on who they are, their age, survival duration, finishing placement, and medical evacuation status, and do those archetypes reveal systematic patterns about what types of people survive? This shift was motivated by a gap in prior milestones. M2 and M3 characterized episodes and gear, but neither asked whether survivalists themselves form predictable groups. The addition of unstructured text analysis on the profession column further extends the scope to ask whether professional background is associated with the survival archetype.

M1 assumed that survival outcomes are primarily driven by individual skill and randomness. The clustering results challenge this: survival duration and medical evacuation status produce well-separated groups with consistent demographic and professional signatures, suggesting that contestant type, not just chance, is a systematic predictor of outcome. In particular, the Resilient Medical Exits cluster (Cluster 4) invalidates the assumption that long-lasting contestants are always voluntary survivors: nine contestants lasted an average of 65 days or more and were removed by medical teams rather than choosing to leave.

2. EDA & Preprocessing: Updates

Earlier milestones established that the Alone dataset contains 94 survivalist records across 9 seasons, with four numerically complete features relevant to personal outcomes: age, days_lasted, result (finishing placement), and medically_evacuated. Prior EDA confirmed that there were no missing values in these columns. The StandardScaler normalization approach was established in the M3 phase and carried forward here without modification.

Two new preprocessing steps were introduced. First, the medically_evacuated column, originally a boolean, was encoded as a binary integer (0/1) to make it compatible with distance-based clustering algorithms. Second, the profession column required text normalization: all values were lowercased using str.lower() before TF-IDF vectorization, consistent with the text cleaning approach used in M3 Section F. No imputation was needed as both columns were complete across all 94 records. No other new EDA steps were required. Feature selection was deliberate, and a minimal four features were chosen for conceptual clarity and direct relevance to the research question, rather than for exploratory purposes.

3. Analysis & Experiments

3.1 Selecting the Number of Clusters (k)

Before fitting any clustering model, we needed to determine the optimal number of groups. We used two complementary tools that were introduced in the course and applied in M3. The KElbowVisualizer plots the distortion, the sum of squared distances from each point to its cluster centroid across k values from 2 to 9. The bend in the curve appeared clearly at k=5, indicating diminishing returns beyond that point. A parallel silhouette score sweep confirmed this: scores were highest in the k=5 to k=6 range (peak at k=6: 0.357), with k=5 chosen as the more parsimonious solution. The SilhouetteVisualizer for k=5 showed reasonably balanced cluster coefficients with no strongly negative values, confirming that k=5 produces coherent, non-overlapping groups.

3.2 Hierarchical vs. K-Means: Method Comparison

We ran both hierarchical clustering (Ward linkage) and K-Means at k=5 on the StandardScaler-normalized feature matrix. The Ward dendrogram visually supported five natural groupings, with a clear gap at the five-cluster cut level. Both methods were then evaluated by the silhouette score in Table 1- Method Comparison:

Hierarchical clustering marginally outperformed KMeans (0.347 vs. 0.342). The difference is small, but hierarchical clustering also produced more interpretable cluster sizes. Hierarchical labels were used for all downstream profiling.

3.3 Five Survivalist Archetypes

The five clusters produced by hierarchical clustering have clear, interpretable identities driven by the interplay of days_lasted, result, and medically_evacuated. Table 2 - Survivalist Archetypes (Hierarchical Ward) summarizes the archetypes:

Cluster 1 - Long-Haul Champions:

The largest and most competitive group. These contestants lasted deep into the game with no medical evacuations. The group includes multiple winners. The average finishing result of 2.3 confirms that these are consistently top performers. 77% male, evenly distributed across all 9 seasons.

Cluster 2 - Mid-Game Competitors:

The youngest group (avg age 30.6). These contestants are competitive but tap out in the middle rounds, primarily for family or personal reasons (10 of 16 exits). Notably, this is the only 100%

male cluster, so no female contestants fall here. Professions include survival instructors, trappers, and skilled tradespeople.

Cluster 3 - Early Voluntary Exits:

The oldest group (avg age 43.9), yet one of the shortest-lasting. These contestants chose to leave early, with 15 of 22 citing family or personal reasons, the highest voluntary exit rate in the dataset. Despite relevant backgrounds (law enforcement, military, instructors), age, or family pull, the decision to leave is ultimately driven by family. No medical evacuations.

Cluster 4 - Resilient Medical Exits (the most analytically significant cluster):

These nine contestants lasted nearly as long as the champions (averaging 65.3 days) but were entirely medically evacuated and never chose to leave voluntarily. Every single exit was medical (9/9). Critically, 67% of this cluster is female, making it the only cluster where women are the majority. **This suggests that female contestants in this show disproportionately persist past physical limits rather than self-evacuating.** Professions are strongly towards wildlife: Ancestral Living Skills Instructor, Homesteader & Carpenter, Wilderness Guide.

Cluster 5 - Early Medical Exits:

Short-duration medical exits. These contestants left early due to health reasons (16/16 medical exits). While sharing the medical evacuation characteristic with Cluster 4, the distinction is duration: these contestants did not build the physical or psychological resilience to last. Professions include blacksmith, homesteader, fisherman, and retired military.

3.4 Text Analysis: Do Professions Predict Survival Archetype?

After clustering on numeric features, we applied TF-IDF vectorization to the profession column, 94 short text strings such as 'Wilderness Guide' or 'Blacksmith', using stop_words='english' and ngram_range=(1,2) with min_df=2. This yielded a 94x48 term matrix. We then computed the mean TF-IDF vector per cluster to identify which profession terms are most associated with each archetype. The results showed a directional pattern consistent with the numeric profiling. Clusters 1 and 4, the two long-lasting groups, had profession terms weighted toward wilderness-specific and instructional roles: 'outdoor,' 'wilderness,' 'skills instructor,' 'builder.' Clusters 3 and 5, the shorter-duration groups, showed terms like 'army,' 'law enforcement,' 'blacksmith,' 'fisherman,' reflecting skilled but less wilderness-specific backgrounds. However, the profession TF-IDF signal is limited by the small vocabulary (48 terms after min_df filtering) and the high uniqueness of profession strings (80+ distinct values in the original data). The text analysis is best interpreted as suggestive rather than confirmatory. It adds a narrative layer to the numeric clusters rather than providing independent predictive power.

3.5 Cross-Validation

We used crosstabs to validate cluster identities against external variables not used in clustering. Three patterns are noteworthy:

1. Gender: Cluster 4 is 67% female; Cluster 2 is 100% male. All other clusters are 75-86% male, reflecting the overall dataset composition.
2. Tap-out reason: Clusters 4 and 5 show zero family/personal exits, so it's entirely medical. Clusters 1, 2, and 3 have mixed reasons, with Cluster 3 showing the highest family/personal proportion (15/22).
3. Season: All five clusters are present across all nine seasons, ruling out the possibility that any archetype is an artifact of a single season's casting.

4. Findings & Interpretations

Clustering 94 Alone contestants revealed five distinct survivalist archetypes, each with a coherent identity that goes beyond simple duration rankings. The most important finding is the role of medical evacuation in separating two groups that are otherwise very similar in how long they last: the Long-Haul Champions and the Resilient Medical Exits, both averaging over 65 days on the show, but they differ in how they exit. Champions chose to stay and ultimately won or finished near the top. The Resilient Medical Exits would not leave voluntarily; they were physically removed.

This distinction has a striking gender dimension. The Resilient Medical Exits cluster is 67% female, the only cluster in the dataset where women are the majority. This suggests that female contestants on Alone are disproportionately likely to push past physical breaking points rather than self-evacuate. For show producers and safety teams, this is a direct operational insight: long-lasting female contestants may require more proactive medical monitoring, not because they are more fragile, but precisely because they are less likely to tap out on their own.

Age, contrary to intuition, does not consistently predict longer survival. The oldest group (Cluster 3, avg age 43.9) actually exits the earliest on a voluntary basis, suggesting that for more experienced or older contestants, family and personal reasons outweigh the competitive drive to persist. Youth, on the other hand, characterizes the Mid-Game Competitors (avg age 30.6), who compete hard but do not reach the endgame. A professional background adds a supporting signal. Wilderness-specific professions, guides, instructors, and homesteaders appear most heavily in the two long-lasting clusters (1 and 4). More generalist trade and military backgrounds are concentrated in the shorter-duration clusters. A casting director looking to build a competitive season should weigh wilderness-specific professional experience over general outdoor or military credentials. Taken together, the five archetypes directly answer the central question of this milestone: survival on Alone is not random. Contestant type, defined by age, profession, gender, and psychological tendency to self-evacuate, predicts not just how long someone lasts, but how their story ends.

Appendix

Shared GitHub Repository (Required)

- GitHub Repository (github.com/MikaIsmayilov/B02_Alone_TVShow_Unsupervised_Project)
- TEAM02_BA820_M4_Report_Ataseven_Burak.pdf
- TEAM02_BA820_M4_Ataseven_Burak.ipynb

Additional figures, tables, or extended analysis:

Table 1- Method Comparison

Method	Silhouette Score	Cluster Sizes
Hierarchical (Ward)	0.347	31 / 16 / 22 / 9 / 16
KMeans	0.342	29 / 19 / 22 / 14 / 10

Table 2 — Survivalist Archetypes (Hierarchical Ward)

Cluster	Archetype Name	N	Avg Days	Avg Age	Med Evac %
1	Long-Haul Champions	31	66.6	38.5	0%
2	Mid-Game Competitors	16	27.1	30.6	0%
3	Early Voluntary Exits	22	16.1	43.9	0%
4	Resilient Medical Exits	9	65.3	39.9	100%
5	Early Medical Exits	16	14.3	34.9	100%

Process Overview

Data source: survivalists_data (94 records, 16 features)

Feature selection: age, days_lasted, result, medically_evacuated (no missing values)

Preprocessing: boolean encoding (medically_evacuated → 0/1), StandardScaler normalization

K selection: KElbowVisualizer (k=5 suggested) + silhouette score sweep (peak k=6, selected k=5)

Clustering: Hierarchical (Ward) silhouette=0.347; KMeans silhouette=0.342. Hierarchical selected

Profiling: groupby().agg() summary tables, pairplot with hue, boxplots by cluster

Text analysis: TF-IDF on profession (94 x 48 matrix), mean cluster vectors, bar charts

Cross-validation: pd.crosstab by gender, reason_category, season

Use of Generative AI Tools

ChatGPT log (https://chatgpt.com/share/699cca31-68cc-800c-858e-652f9fa9952f)

Grammarly is used to correct grammatical mistakes.

All lecture notebook's used as a code template.