BA820 – Project M2

Cover Page

- **Project Title:Survival Strategies, Attrition Patterns, and Audience Engagement in the Alone TV Series**
- **Section and Team Number: A1 – Team 08**
- **Student Name:Marcus Shi, Xincheng Ding, Yangze Li, Yanlun Li**

1. Refined Problem Statement & Focus (~0.5 page)

In the proposal stage, our team outlined four domain questions examining survival strategies, exit dynamics, competition structure, and audience engagement in *Alone*. In this milestone, I focus on the first question: whether participants' loadout configurations give rise to distinct, implicit survival strategy types, and whether these strategy types are systematically associated with survival duration and exit outcomes.

This focus reflects a refinement of the original problem statement informed by early exploratory analysis. Initial EDA showed that loadout choices are highly heterogeneous and sparse, with many items appearing infrequently and no obvious, pre-defined strategy categories. At the same time, survival duration and exit outcomes exhibit substantial variation but do not naturally separate participants into clear outcome-based groups. These observations challenge the assumption that survival strategies can be directly defined or inferred prior to analysis.

As a result, the analytical focus of this milestone shifts from testing outcome differences across assumed strategies to first examining whether meaningful strategy types can be discovered from loadout configurations alone. Survival outcomes are therefore treated as conditional variables, evaluated only after identifying latent strategy groupings. This reframing motivates the use of unsupervised methods to uncover structure in initial planning decisions without imposing outcome-driven labels.

Overall, the refined problem emphasizes discovery before explanation: establishing whether stable and interpretable loadout-based strategy types exist, and only then assessing how these strategies relate to survival trajectories and exit pathways. This approach ensures that subsequent analysis is grounded in observed data structure rather than a priori assumptions.

2. EDA & Preprocessing: Updates (~0.75 page)

Exploratory analysis conducted in the proposal stage revealed two patterns that directly motivate the current analysis. First, survival duration exhibits substantial variability and a right-skewed distribution, without clear separation into outcome-based groups, suggesting that outcomes alone are insufficient for defining meaningful participant categories. Second, loadout item usage is highly uneven, with a small set of frequently selected items and many low-frequency items, indicating that individual items provide limited discriminatory power for defining survival strategies.

Building on these findings, additional EDA was introduced in this milestone to better assess whether loadout configurations encode latent strategic structure. Specifically, I examined the long-tail distribution of item usage on a log scale and analyzed pairwise similarity between participants' loadouts using Jaccard similarity. These analyses were not included in the proposal stage and were motivated by the need to evaluate whether strategy signals emerge at the configuration level rather than through single items or outcomes. The results suggest substantial

sparsity and heterogeneous overlap patterns, supporting an unsupervised approach to discovering implicit strategy types.

Preprocessing in this milestone focused on consistency rather than aggressive feature engineering. Loadout items were standardized using the coarse-grained item field to reflect functional strategy categories, with minor duplication treated as measurement noise due to category aggregation. Participants were uniquely identified by season and name. Although the show enforces a ten-item rule, minor inconsistencies exist in the number of unique item categories recorded per participant; given their limited extent, no imputation or filtering was applied. No additional dimensionality reduction was performed in order to preserve the sparse and high-dimensional structure of the loadout space for unsupervised analysis.

3. Analysis & Experiments (~1.5 page)

Method : Similarity-Based Strategy Discovery via Hierarchical Clustering

Our primary research question asks whether participants' loadout configurations give rise to distinct, implicit survival strategy types. Rather than treating individual items or outcomes as defining features, this analysis focuses on configuration-level similarity between participants' chosen loadouts. By representing each participant's loadout as a set of items and computing pairwise Jaccard similarity, we directly operationalize the notion of strategic overlap: two participants are considered similar if they select many of the same items, regardless of item order or frequency. Hierarchical clustering applied to this similarity structure allows me to explore whether participants naturally group into strategy types without imposing predefined labels or assumptions.

This approach is well suited to the data and analytical goal for several reasons. First, loadout data are inherently binary and set-based, with each participant selecting a small number of items from a shared catalog. Jaccard similarity is specifically designed for such sparse, binary representations and captures overlap in selections rather than numeric distance. Second, exploratory analysis showed that item usage follows a long-tail distribution and that most participants select a common core of survival necessities, making feature-level clustering unreliable. Hierarchical clustering further avoids the need to pre-specify the number of strategies and allows me to examine structure at multiple levels of granularity. I expected this approach to reveal whether strategy differentiation emerges at the configuration level, and if so, whether such differentiation is sharp or gradual.

I constructed a participant-by-participant Jaccard similarity matrix and applied hierarchical clustering using average linkage as the primary method. To assess robustness, we also tested complete linkage, which enforces stricter within-cluster similarity. Cluster solutions were examined across multiple values of k, including k = 3, 4, and 5. For each configuration, I

evaluated cluster sizes, top loadout items within clusters, and distributions of survival duration. We explicitly explored whether increasing k revealed additional interpretable strategy types or merely fragmented existing groups.

Across linkage methods and cluster counts, the analysis consistently revealed one dominant cluster containing the majority of participants, a smaller secondary cluster, and a small number of highly distinctive individuals. This high-level structure was stable across both average and complete linkage, indicating that it is not an artifact of a specific clustering choice. The dominant cluster reflects a common survival configuration built around shared necessities (e.g., shelter, fire, and food acquisition), while the smaller cluster exhibits modest variation in emphasis rather than a fundamentally different strategy.

What did not work was increasing the number of clusters beyond three. For k = 4 and k = 5, hierarchical clustering primarily produced singleton or very small clusters, isolating idiosyncratic participants rather than uncovering new, population-level strategy types. These clusters offered no additional interpretability and did not meaningfully differentiate survival outcomes. This outcome suggests that strategy variation in the data is continuous rather than discretely segmented, with a long tail of individual-specific configurations.

One notable surprise was how consistently weak the separation remained even under methods specifically designed for sparse, set-based data. This reinforced the conclusion that the absence of sharp clustering is a property of the data itself rather than a modeling failure. Importantly, this analysis shifted our framing of "strategy types" away from discrete categories and toward overlapping patterns constrained by shared survival requirements.

K-means clustering was considered but ultimately not applied due to a fundamental mismatch between its assumptions and the structure of the data. K-means relies on Euclidean distance in a continuous feature space and assumes spherical clusters with interpretable centroids. Loadout configurations are binary, sparse, and set-based, and strategic similarity is driven by item overlap rather than numeric distance. Under one-hot encoding, high-frequency items dominate Euclidean distance while rare but potentially meaningful items are downweighted, leading to unstable and misleading clusters. Given this structural incompatibility, applying K-means would not meaningfully address the research question and could introduce artificial segmentation unsupported by the data.

4. Findings & Interpretations (~0.75 page)

This analysis examines whether participants' initial loadout configurations reflect distinct survival strategy types and whether such strategies meaningfully relate to survival duration. The results suggest that while loadout choices exhibit some structured similarity, they do not form clearly separable or consistently predictive strategy categories.

Across seasons, participants demonstrate strong overlap in loadout selections, driven by shared survival necessities such as shelter, fire-making, and food acquisition. This produces a dominant configuration pattern encompassing the majority of contestants. Differences between participants primarily reflect variations in emphasis rather than fundamentally different strategic approaches. In practical terms, most contestants begin the competition from a broadly similar baseline rather than from sharply differentiated philosophies of survival.

Clustering analysis reveals limited but interpretable structure. Participants consistently organize into one large group and one smaller subgroup, alongside a small number of highly distinctive individuals. Importantly, increasing the number of clusters does not uncover additional meaningful strategy types. Instead, it isolates idiosyncratic participants whose loadouts reflect personal preferences rather than repeatable, population-level strategies. This pattern indicates that loadout strategies are better understood as a continuous spectrum rather than discrete categories.

Differences in survival duration across loadout-based groupings are modest and unstable. While some clusters exhibit slightly higher median survival times, survival distributions overlap substantially and display high variability. No loadout configuration consistently dominates others in terms of survival outcomes. This suggests that initial item selection alone plays a constrained role in determining success, and that survival is more strongly influenced by execution, adaptability, and situational factors encountered during the competition.

Finally, the presence of highly distinctive loadouts highlights the limits of strategic categorization in this context. Although unconventional configurations exist, they do not generalize into broadly effective strategies. From a real-world perspective, these findings caution against overemphasizing "winning loadout strategies" and instead point to the importance of flexibility and in-situ decision-making in constrained environments.

5. Next Steps (~0.25 page)

The current analysis is limited to participants' initial loadout configurations and does not incorporate in-season behaviors or exit dynamics. In particular, we have not yet examined how survival unfolds over time, nor distinguished between different exit pathways such as voluntary withdrawal and medical evacuation.

The next stage of analysis will therefore shift toward process-oriented survival dynamics. Planned extensions include modeling time-to-exit, comparing exit types, and assessing how early departures shape season-level competition structure. These analyses aim to identify whether behavioral or situational factors explain survival variation beyond initial preparation.

These next steps are directly motivated by the present findings: because loadout-based strategies exhibit weak differentiation and limited explanatory power, understanding survival outcomes requires moving beyond static inputs toward dynamic decision-making and adaptation during the competition.
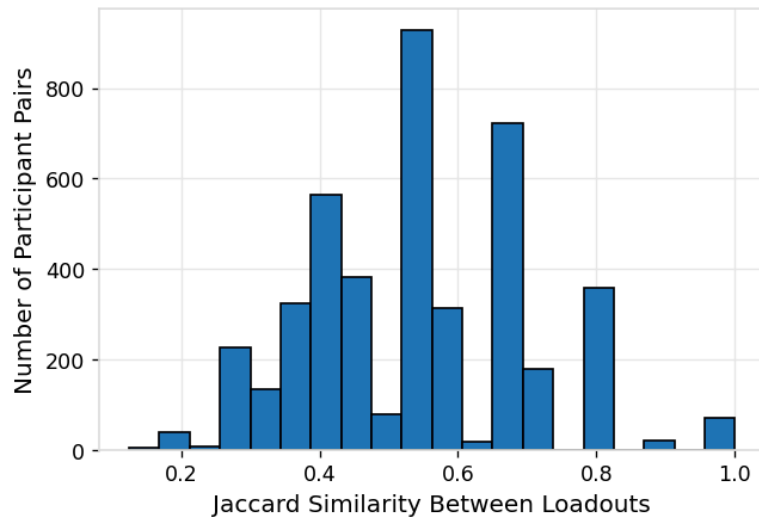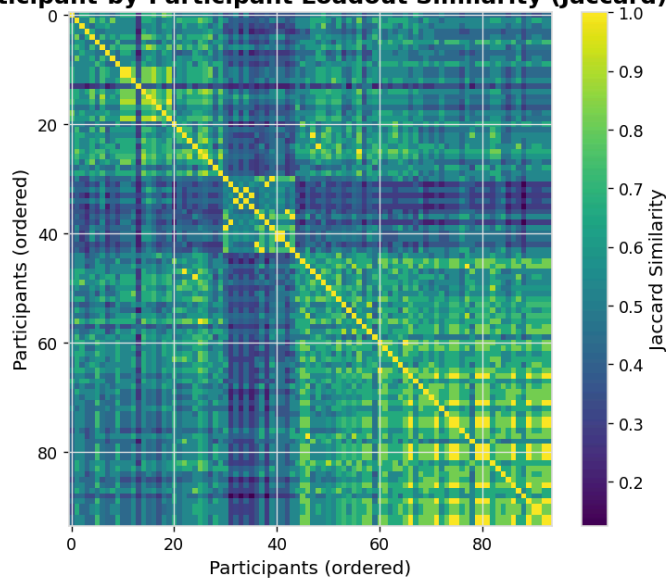
Appendix

Shared GitHub Repository (Required)

- [Shared Repo Link](#)

- Contribute to the branch Yanlun-M2 all contents and the main branch
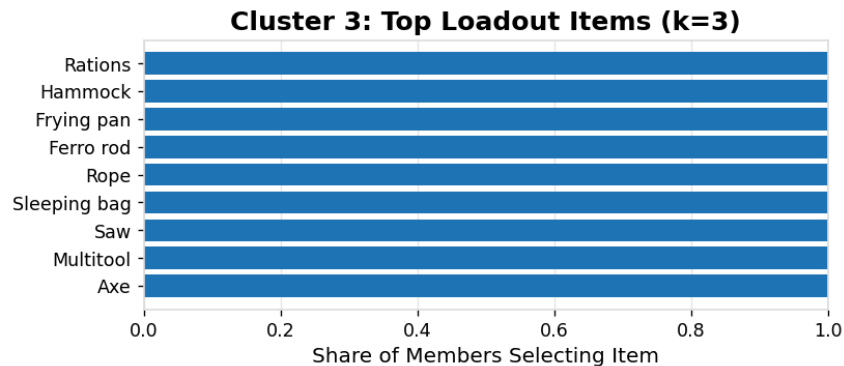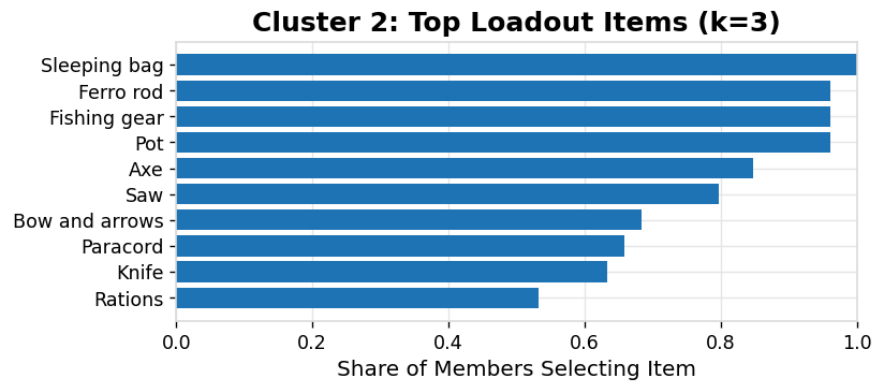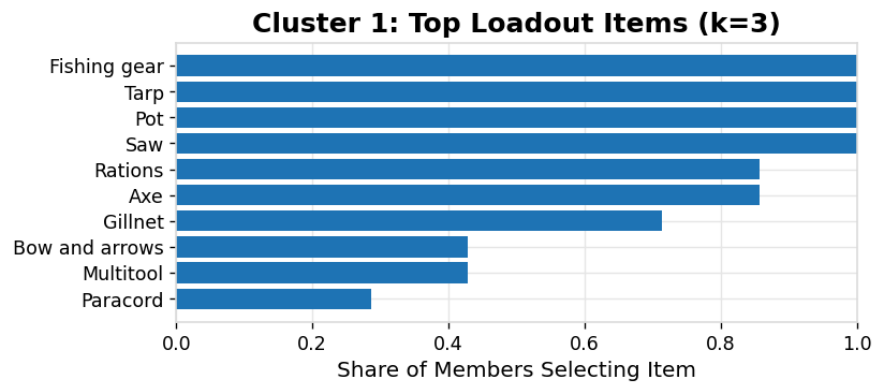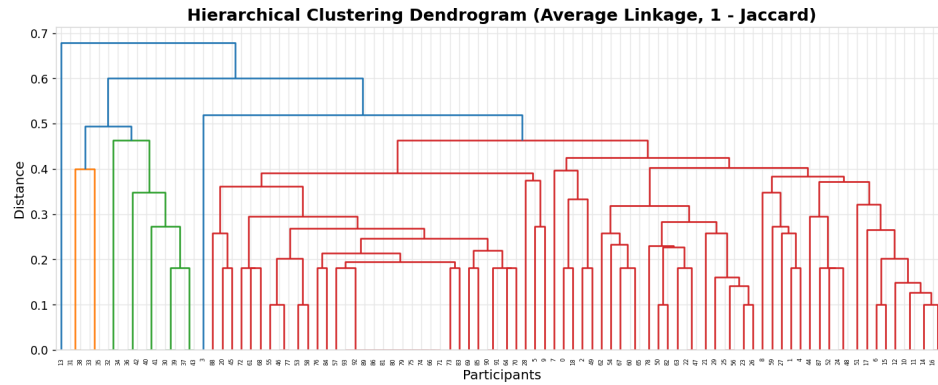 BA820_A1_08_Alone_TV_Show.ipynb for M1

Supplemental Material (Highly Recommended)

Distribution of Pairwise Loadout Similarity

Participant-by-Participant Loadout Similarity (Jaccard)

Hierarchical Clustering Dendrogram (Average Linkage, 1 - Jaccard)

Cluster 1: Top Loadout Items (k=3)

Share of Members Selecting Item

Cluster 2: Top Loadout Items (k=3)

Share of Members Selecting Item

Cluster 3: Top Loadout Items (k=3)

Share of Members Selecting Item

Process Overview

The analytical workflow begins by representing each participant's initial loadout as a set of selected items, enabling direct comparison of preparation choices without imposing artificial ordering or weighting. This representation naturally motivates a similarity-based approach aligned with the discrete and constrained nature of the data.

Exploratory analysis is first used to assess overall overlap in loadout composition and survival outcomes, establishing that participants share a large common core of items. Pairwise Jaccard similarity is then computed between participants' loadouts, followed by hierarchical clustering using distance defined as one minus similarity. Multiple linkage methods and cluster resolutions are examined to evaluate the stability and interpretability of any emergent structure.

Finally, cluster-level item profiles and survival duration distributions are analyzed post hoc to interpret whether identified groupings correspond to distinct preparation strategies or meaningful outcome differences. This process allows the analysis to directly assess the extent to which static loadout configurations alone can explain variation in survival performance.

Use of Generative AI Tools

AI tools were used in a limited and supportive capacity, primarily to assist with language refinement, clarification of concepts not explicitly covered in class, and high-level brainstorming of analytical approaches. All data analysis, methodological decisions, code implementation, and interpretation of results were completed independently by myself.

ChatGPT Link