BA820 – Project Proposal

Cover Page

- **Project Title: Survival Strategies, Attrition Patterns, and Audience Engagement in the Alone TV Series**
- **Section and Team Number: A1 – Team 08**
- **Members: Yangze Li**

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

Paradigm Shift: From Demographics to Behavioral Strategy

In the initial phase (Milestone 1), our research framework was constrained by "demographic determinism," attempting to validate whether age structure or gender were the singular dominant variables predicting survival duration. However, this linear assumption was challenged by the Exploratory Data Analysis (EDA) in Milestone 2. The data revealed that even within identical demographic cohorts (e.g., males aged 30–40), survival performance exhibited massive variance, with duration gaps exceeding 60 days between peers. This significant heterogeneity suggests that physiological attributes are merely background features, insufficient to explain the divergence in survival outcomes.

Hypothesis Iteration and Validation:

Consequently, the analytical focus has shifted fundamentally from "who is competing" to "how they survive." The refined hypothesis postulates that the degree of specialization in equipment strategy—specifically how a participant constructs their survival system from the 10-item quota—is the decisive dimension. This study now focuses on utilizing unsupervised learning to identify latent strategic archetypes within the 27-item equipment matrix. We aim to verify whether "technical specialization" offers a statistically significant survival advantage over "risk-averse" or "generalist" approaches, thereby upgrading the research from static profiling to dynamic behavioral assessment.

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

Drivers for Analytical Pivot:

The Residual Analysis in M1 demonstrated that demographic variables possessed limited explanatory power, and the target variable (survival days) showed a heavy right-skew. To capture more granular behavioral patterns, the data architecture was reconstructed in M2 to expand the analytical dimension from the "person" to the "loadout."

Data Governance and Feature Reconstruction:

To ensure the mathematical robustness of the K-Means clustering algorithm, three critical data governance strategies were executed:

2.1. Construction of High-Dimensional Sparse Matrix:

The original loadouts.csv recorded equipment in a list format, unsuitable for distance calculation. I reconstructed this into a 27-dimensional Binary Feature Matrix. Each column represents a

specific tool (e.g., Gillnet, Pot, Bow), where a 0/1 encoding not only quantifies possession but mathematically defines the participant's "Strategic Space." This transformation converted unstructured text lists into computable vectors.

2.2. Strategic Pruning of Noise:

During the integrity audit of the merged dataset, the variables team and day_linked_up exhibited severe missingness rates of 85.1% and 91.5%, respectively. To maximize the signal-to-noise ratio, these variables were permanently removed to prevent the algorithm from being misled by non-informative data.

2.3. Feature Scaling and Alignment:

Merging demographic data with the equipment matrix introduced a scale mismatch: the magnitude of the Age variable (range 19–60+) far exceeded the binary equipment variables (0-1). Left unaddressed, Age would dominate the Euclidean distance calculations, masking the nuances of equipment choice. Therefore, Standard Scaling was applied to Age, ensuring it carries equal weight to the equipment features in the clustering algorithm.

3. Analysis & Experiments (~1.5 page)

Methodology: K-Means as a Strategy Discovery Tool
Faced with "survival strategies" that lack predefined labels, K-Means Clustering was selected as the core analytical engine. The objective is not numerical prediction but pattern discovery. By calculating Euclidean distances among 94 participants in a 28-dimensional space (27 items + Age), K-Means automatically groups participants employing similar survival means. This unsupervised approach directly addresses the business question: How can we identify mainstream survival schools of thought without prior knowledge?

Parameter Exploration and Model Iteration:

The search for the optimal number of clusters (k) involved a systematic sweep rather than reliance on a single metric, balancing statistical scores with business interpretability.

The experiment tested the parameter space from $k=2$ to $k=6$, using the Silhouette Score as the primary evaluation metric. The process encountered a typical "curse of dimensionality" challenge: due to the high sparsity of the equipment matrix (most participants do not carry niche items like the Scottish Dirk), data points are sparsely distributed, resulting in modest overall Silhouette Scores in the 0.13–0.14 range. When testing $k=6$, micro-clusters containing only 3–5 individuals appeared; this overfitting resulted in a loss of generalizability, marking a "dead end" in the analysis.

Final Decision: Prioritizing Interpretability

Although $k=2$ achieved a marginally higher statistical score (0.148), I definitively selected $k=4$ (Score: 0.147) as the optimal model. This decision was driven by "Business Logic":

- $k=2$ offered a crude binary split (essentially "Bow" vs. "No Bow"), losing critical strategic nuance.
- The $k=4$ model successfully deconstructed the population into four groups with distinct tactical signatures (e.g., "Professional Hunters" vs. "Minimalists").
  In exploratory analysis, a specific, actionable insight is far more valuable than a negligible statistical margin. We chose to sacrifice 0.001 in score to gain deep resolution into strategy types.

4. Findings & Interpretations (~0.75 page)

Based on the $k=4$ model, we mapped four distinct "Survivalist Personas." The data validates that strategy is not random; there is a strong causal suggestion between equipment choice and survival outcome.

Insight 1: The Victory of Active Aggression (Active vs. Passive)

The data powerfully demonstrates the dominance of "High Risk, High Reward" strategies.

- Cluster 2 (Technical Hunters) emerged as the clear top-performing cohort, with an average survival duration of 46 days. The defining characteristic of this group is an aggressive loadout: 84% carried a Bow and Arrow, frequently paired with Gillnets.
- In sharp contrast, Cluster 0 (Minimalist Foragers), despite having a similar average age (approx. 33 years), relied on passive tools (rations, pots) and exited significantly earlier (average 30 days).
- Interpretation: In calorie-deficient environments, "Revenue Generation" (active hunting) is far more effective than "Cost Cutting" (passive rationing). Strategies that actively control energy sources yield a substantial survival dividend.

Insight 2: Experience as Intangible Capital

- Cluster 3 (Senior Veterans) debunks the myth that "youth equals capability." This group, with the highest average age (47 years), achieved an impressive 43 days of survival, leveraging balanced gear and (inferred) psychological resilience.
- This performance decisively outperformed the youngest group, Cluster 1 (Generalists, Avg Age 32), who averaged only 29 days—the lowest in the dataset. Cluster 1's failure implies that a "middle-of-the-road" strategy lacking tactical focus is highly risky.

- Interpretation: For talent selection, this suggests that "Hard Metrics" (Physical Fitness) should not overshadow "Soft Metrics" (Experience/Patience), as the latter often dictates who endures the longest.

5. Next Steps (~0.25 page)

While the current model successfully maps the correlation between strategy and result, two "causal black boxes" remain, forming the roadmap for Milestone 3.

First, the model currently cannot distinguish between "Strategic Failure" and "Force Majeure." Is Cluster 1's low survival rate due to flawed tactics, or did they simply suffer more random injuries? The next step involves integrating the "Reason for Tap-out" variable from episodes.csv to separate medical evacuations from voluntary withdrawals. This will isolate the true efficacy of the strategies by filtering out bad luck.

Second, the current analysis treats all seasons as a uniform backdrop, ignoring environmental context. The damp rainforests of Vancouver Island and the arid winds of Patagonia demand vastly different gear. Future work will introduce Location as a control variable to determine if specific strategies are environmentally dependent, providing a more rigorous causal inference.

6. Appendix：

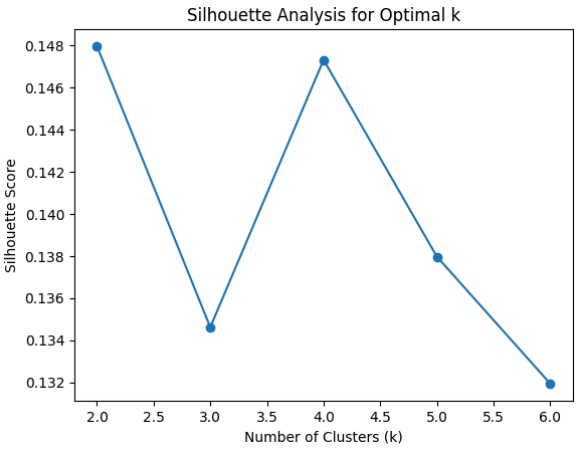
Shared GitHub Repository (Required)

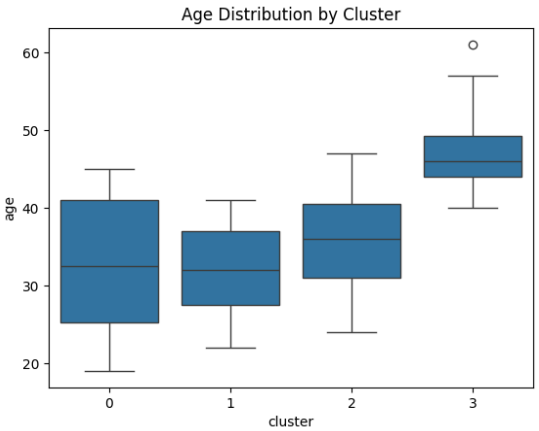
GitHub Repository: https://github.com/Marcusshi/BA820-A1-08

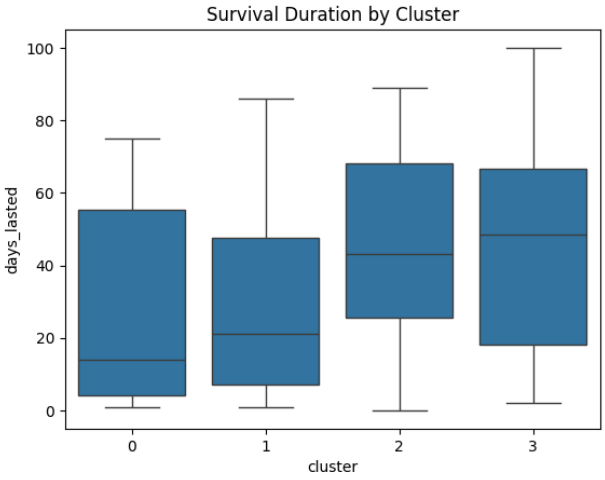
My individual work for Project M2 is contained in the following files:

- ➢ M2/Yangze Li/BA820_M2_TeamA1_08_YangzeLi.ipynb
- ➢ M2/Yangze Li/BA820_M2_TeamA1_08_YangzeLi_Report.pdf

EDA :







Process Overview:

1. **Data Integration:** Merged survivalists.csv (demographics) with loadouts.csv (equipment) to create a unified analytical dataset.
2. **Preprocessing:** Constructed a 27-item binary matrix for equipment; Removed variables with >80% missing data; Standardized Age to ensure scale consistency with binary data.
3. **Modeling:** Applied K-Means Clustering algorithm, iteratively testing cluster counts from $k=2$ to $k=6$.
4. **Validation:** Evaluated model performance using Silhouette Scores, identifying $k=4$ as the optimal balance between statistical cohesion and interpretability.
5. **Profiling:** Mapped the resulting 4 clusters against survival_days to identify distinct strategic archetypes.

Use of Generative AI Tools

I utilized generative AI tools as a supportive assistant and critical reviewer during this project. Specifically, I used them to troubleshoot Python syntax errors in my notebook and to refine the grammatical flow of this report to ensure clarity. I also engaged the AI to audit my analytical logic against the project rubric, ensuring my pivot from demographics to equipment strategy remained aligned with the assignment's goals. Additionally, I used it to brainstorm potential interpretations of cluster characteristics. However, all data preprocessing decisions, the specific selection of the $k=4$ K-Means model, and the strategic insights regarding survival duration reflect my own analytical judgment.

AI Link: https://chatgpt.com/share/698a8925-07c8-800f-ab41-6eb8ed4be4e0