**Project Milestone 2: Early Analysis & Insights**

**Project Title: Uncovering Site Typologies, Feeding Strategies, and Participation Patterns in Project FeederWatch**

**BA820 B1**

**Yiheng Chen**

**Feb 9, 2026**

## Project Motivation and Problem Statement: What Changed and Why

Question: Feeding Strategy Segmentation (Intensity × Diversity × Seasonality). Do feeding sites cluster into distinct behavioral strategies based on feeder intensity, feeder type diversity, and seasonal consistency of feeding activity (e.g., light seasonal feeders vs. intensive year-round feeders)?
Who benefits / decisions: Outreach teams can identify which behaviors correlate with high engagement or richer observation value, and tailor guidance (e.g., education, retention) to different feeder strategies.

In the M1 proposal, our team raised several domain-related questions. Question 2 focused on whether feeding sites could be classified into different feeding strategies based on feeding frequency, feeder type variety, and seasonality. Because Project FeederWatch relies heavily on individual participant behavior, feeding sites vary widely across households. Some sites use many types of feeders and feed birds frequently, while others feed less often or only during certain months. Clearly identifying feeding strategies can help outreach teams provide more targeted guidance to different types of participants.

At this stage, my goal is to examine whether feeding sites differ meaningfully in feeding intensity, feeder diversity, and consistency of feeding throughout the year. The core research question remains unchanged. During M1, I combined site-level and checklist-level data to explore the overall dataset. However, further analysis showed that a single checklist usually represents feeding behavior from only one day and does not reflect a household's overall feeding pattern. To address this issue, I re-aggregated the data to the site–period level, so that each record summarizes overall feeding behavior within a project period.

Additionally, I initially assumed that "seasonality" would exhibit significant variation, with one group being seasonal foragers and another being year-round foragers. However, exploratory data analysis challenged this assumption. After converting foraging consistency to the number of foraging months, I observed that many locations clustered heavily within the consistency range (typically near the maximum value), indicating that year-round foraging is highly prevalent in this sample. Seasonal factors may play a smaller role than anticipated.

## EDA & Preprocessing: Updates

Based on the preliminary exploratory data analysis from Milestone 1, I found that additional preprocessing was needed to address scale differences and skewed distributions in the feeding-related variables. I first aggregated checklist-level observations to the site–period level because a single checklist usually reflects feeding behavior from only one day and does not represent a site's overall feeding pattern. After this aggregation, each row captures the overall feeding behavior of a site within a project cycle. Feeding intensity, diversity, and consistency were summarized using mean or sum measures, while checklist counts, and sampling duration were kept as descriptive indicators of participation.

When examining the distributions, I noticed that feeding intensity was highly right-skewed, with a small number of sites showing extremely high feeding frequencies. I think using the raw values would overly weight these extreme cases and obscure meaningful variation among most sites. To improve interpretability and prepare the data for clustering, I applied a log (1 + x) transformation. This reduced the influence of extreme values while still preserving relative differences among low- and medium-intensity feeding sites.

Finally, I examined the correlations among key foraging variables. Results revealed significant positive correlations between feeding intensity, forager diversity, and foraged species richness. In contrast, feeding stability showed markedly weaker correlations with other indicators and appeared to reflect factors beyond merely stronger feeding behavior. Therefore, in subsequent analyses, I treated feeding stability as an independent characteristic.

## Analysis & Experiments

The goal was not prediction, but to observe whether the data itself supported meaningful "strategy types." Thus, I employed two complementary clustering methods: hierarchical clustering and K-means clustering.

- **Hierarchical Clustering (Ward)**

My goal is to see whether feeding sites naturally form distinct "feeding strategies" based on feeding intensity, feeder variety, and feeding consistency. Hierarchical clustering helps me explore whether the data show clear grouping structure without committing to a single number of clusters at the start. This is useful for early-stage segmentation because it gives me a visual sense of how strongly sites separate.

I think Ward's method is appropriate here because it tends to create compact clusters by minimizing within-cluster variance. Since my features were on different scales, I standardized them first so that one variable would not dominate the distance calculation. With standardized features, Ward clustering gives a reasonable first look at whether "feeding strategy types" exist in a structured way.

I used three core features for clustering: feeding intensity, feeding variety, and feeding consistency. I standardized these features and ran Ward hierarchical clustering. Then I examined the dendrogram and tested different cut points. I initially cut the tree into five clusters to see whether a more detailed segmentation would be interpretable. The dendrogram suggested there is real structure in the data, which supports the idea that feeding behavior is not random. However, when I cut the dendrogram into five clusters, I noticed the cluster sizes were uneven, with some clusters much smaller than others. This made me think that a five-cluster solution might be over-segmenting the data and creating clusters that are harder to interpret or less stable. This result pushed me to compare with a simpler clustering method to check whether the main patterns remain consistent.

- **K-Means**

The mean-centering clustering algorithm offers a more direct approach to generating practical segmentation results, as it assigns each site to a cluster with a distinct centroid. This facilitates the transformation of foraging behavior into a small number of describable and comparable strategy "types."

I believe the K-means clustering algorithm suits the data and objectives because it is simpler, more scalable, and easier to interpret, serving as a good complement to hierarchical clustering.

Furthermore, it forces me to explicitly specify the number of clusters, which is crucial for the goal of providing actionable segmentation results for promotion or guidance.

I attempted K-means clustering using the same standardized feature set across different k values, comparing results via the silhouette coefficient. Testing k values from 2 to 8 revealed how cluster separation evolves with increasing number of clusters. I also examined cluster balance and whether the final clustering aligns with behavioral principles.

I found the silhouette coefficient peaked at k = 2, suggesting a potential clear partition in the data. However, I considered the two-cluster scheme too broad. At k = 3, the silhouette coefficient declined but clusters maintained reasonable separability, making it easier to interpret as distinct strategy types. For larger k values, the silhouette coefficient continued to decrease, and the clustering results became increasingly difficult to interpret.

What surprised me was that feeding consistency showed no significant correlation with feeding intensity or diversity. This corroborates my observation during exploratory data analysis (EDA) that consistency may represent a distinct behavioral dimension rather than merely "more feeding." This reinforces my conviction that it should be treated as a separate feature rather than merged into a single intensity metric.

From this analysis, I learned that unsupervised methods are most useful for understanding structure rather than optimizing a single metric. Although simpler solutions sometimes score better numerically, they may hide important behavioral differences. I also learned that feeding behavior is multi-dimensional. This analysis showed me that initial assumptions, such as the importance of seasonality, should be revisited when the data suggests otherwise. Adjusting these assumptions helped refine, rather than weaken, the overall research direction.

## Findings & Interpretations

Based on the clustering results with K=3, I found that feeding sites can be categorized into three distinct feeding strategy types. These types exhibit significant differences in bird feeding frequency, diversity of feeder setups, and consistency of feeding throughout the year. This supports the view that bird feeding behavior is structured rather than random.

The first group represents low-intensity, low-diversity feeding sites. These locations exhibit infrequent feeding, use fewer types of feeders, and often display irregular feeding patterns throughout the year. Such sites may reflect more casual or occasional participation. For outreach efforts, these locations may most benefit from basic guidance and encouragement to support more consistent engagement.

The second group consists of moderate-intensity feeders exhibiting balanced behavior. These sites maintain moderate feeding frequency, utilize a wider variety of feeders, and exhibit relatively stable feeding patterns throughout the year. Participants in this group appear engaged but not extremely active. From a practical standpoint, these sites could enhance data quality through minor adjustments, such as optimizing feeder placement or recording methods.

The third category comprises high-intensity, highly stable feeding sites. These locations feed birds frequently, employ diverse feeder setups, and maintain feeding activity for most or all the year. This approach reflects significant dedication and extensive experience. Though fewer in number, these sites may yield exceptionally rich observational data. Project organizers could provide more in-depth recommendations or feedback to these sites to further enhance data value.

In this dataset, these findings indicate that only a few behavioral characteristics are required to identify and interpret feeding strategies. While this analysis does not aim to establish causality, it provides a useful framework for understanding participant behavior and offers practical guidance for designing differentiated outreach and support strategies.

## Next Steps

At this stage, I have not incorporated habitat or site environmental variables, nor have I tested whether the K=3 clustering remains stable across different feature sets or project periods. In subsequent phases, I plan to:

1) Add habitat and site characteristics to explore foraging strategy differences across environments

2) Conduct stability tests by comparing clustering results across different feature subsets and time periods.

After controlling for observation effort, I will examine whether these foraging strategy clusters correlate with bird observation records and whether sites maintain consistent strategies over time. The current results provide clear initial segmentation and offer directional guidance for subsequent variable selection and validation steps.

## Use of Generative AI Tools

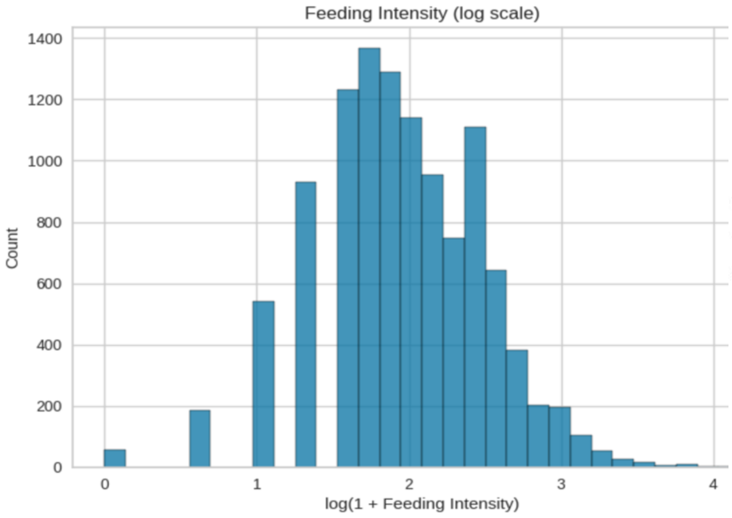
In this project, artificial intelligence tools served as supplementary resources. After conducting my own brainstorming session, I utilized AI to verify if any critical points were overlooked, and AI supplemented some of my brainstorming content. The primary objectives were to enhance efficiency while identifying and refining certain requirements. During the coding phase, whenever errors occurred, I consulted ChatGPT on how to modify the code.

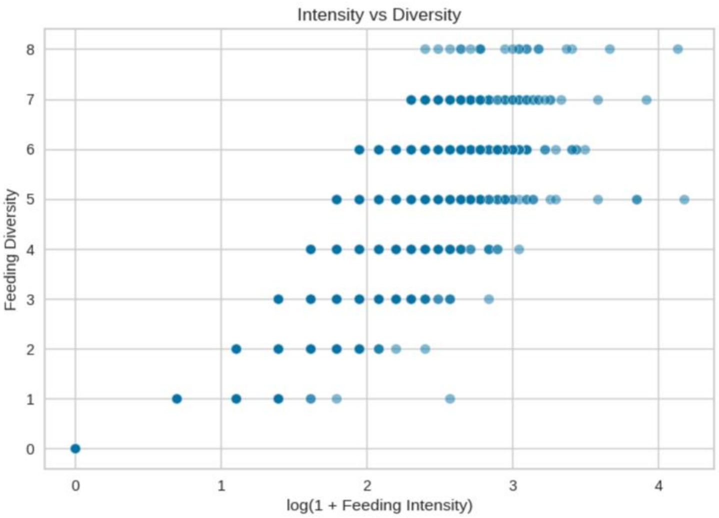
All data processing, exploratory analysis, modeling decisions, and result interpretation were completed independently by me. AI assisted with polishing the text and refining syntax, improving the overall coherence of the paper. For certain interpretations, I would first review ChatGPT's insights, but the final conclusions, methodological choices, and interpretations were conceived and finalized by me. I will include links to ChatGPT's responses.

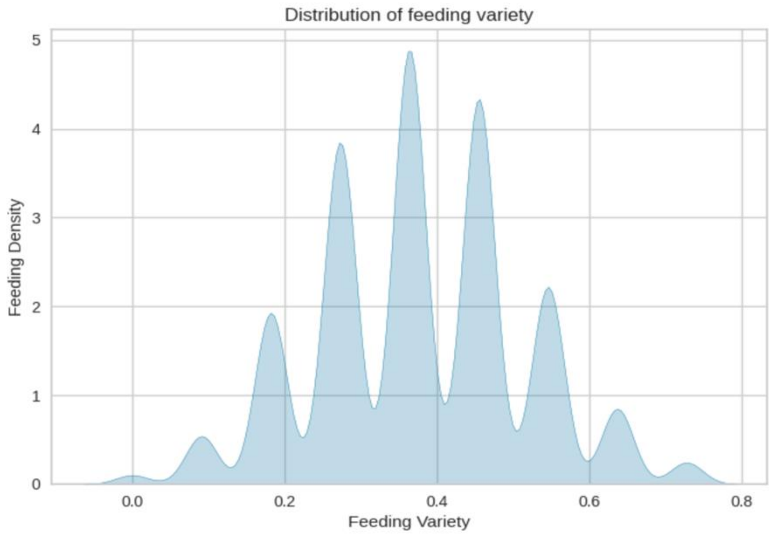
AI Link: https://chatgpt.com/c/698a47c1-deac-832b-9a0d-810d3cf997e3

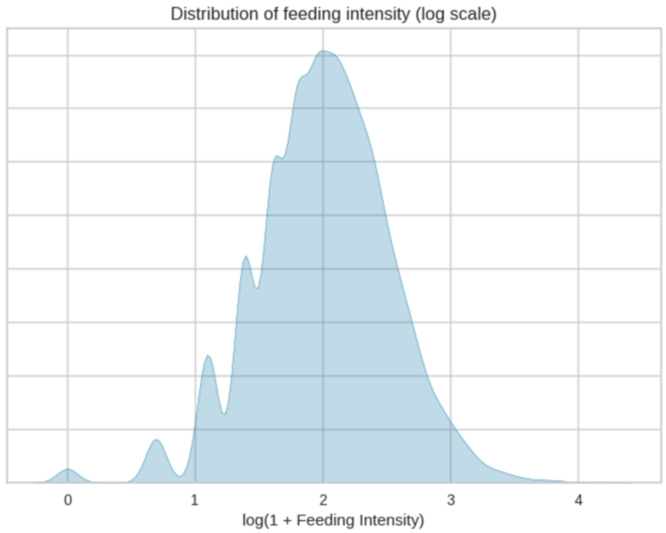
## Appendix

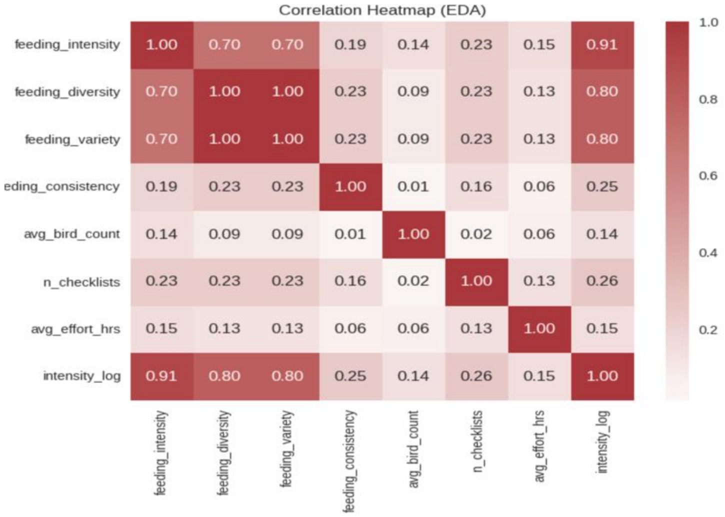
GitHub Link: https://github.com/ply00/820-team9.git
Branch: Yiheng-Chen-M2
Report: Yiheng BA820 M2.pdf
Notebook: Yiheng_of_820_birds_feeder(1).ipynb

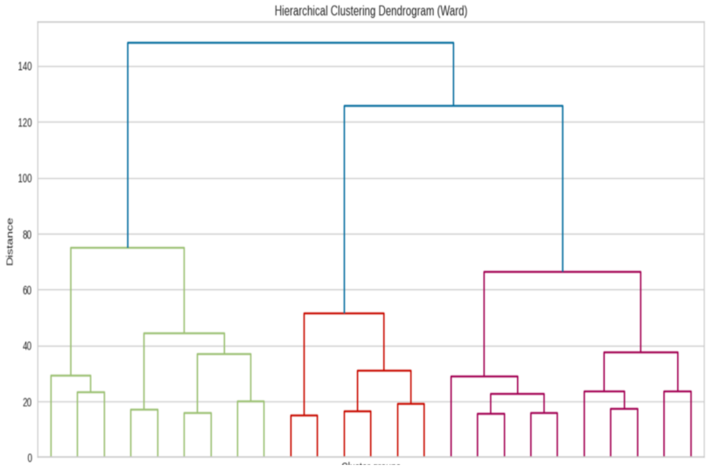


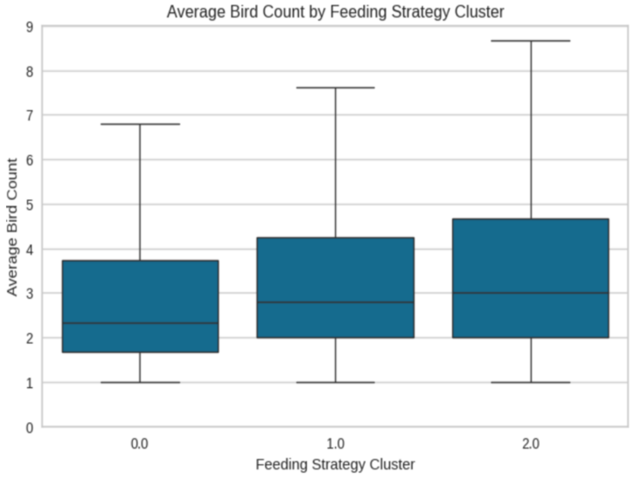


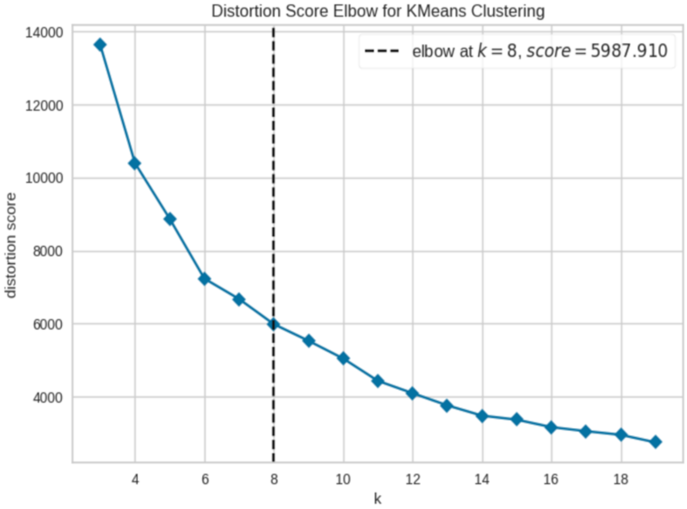





Correlation Heatmap (EDA)
feeding_intensity 1.00 0.70 0.70 0.19 0.14 0.23 0.15 0.91
feeding_diversity 0.70 1.00 1.00 0.23 0.09 0.23 0.13 0.80
feeding_variety 0.70 1.00 1.00 0.23 0.09 0.23 0.13 0.80
eding_consistency 0.19 0.23 0.23 1.00 0.01 0.16 0.06 0.25
avg_bird_count 0.14 0.09 0.09 0.01 1.00 0.02 0.06 0.14
n_checklists 0.23 0.23 0.23 0.16 0.02 1.00 0.13 0.26
avg_effort_hrs 0.15 0.13 0.13 0.06 0.06 0.13 1.00 0.15
intensity_log 0.91 0.80 0.80 0.25 0.14 0.26 0.15 1.00
feeding_intensity
feeding_diversity
feeding_variety
feeding_consistency
avg_bird_count
n_checklists
avg_effort_hrs
intensity_log
1.0
0.8
0.6
0.4
0.2


Hierarchical Clustering Dendrogram (Ward)
Distance
0
20
40
60
80
100
120
140


Average Bird Count by Feeding Strategy Cluster
Average Bird Count
0
1
2
3
4
5
6
7
8
9
0.0
1.0
2.0
Feeding Strategy Cluster


Distortion Score Elbow for KMeans Clustering
elbow at $k = 8$, $score = 5987.910$
distortion score
4000
6000
8000
10000
12000
14000
4
6
8
10
12
14
16
18
k