BA820 – Project M4

Cover Page

- **Project Title: Unsupervised Analysis of Color Style Patterns Across Seasons in Bob Ross Paintings**
- **Section and Team Number: B1, Team05**
- **Student Name: Kefei Zhang**

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

In earlier milestones, I examined whether Bob Ross's visual style changes across the show's production seasons. My initial assumption was that differences between TV seasons would appear through shifts in palette complexity or dominant colors. However, early clustering showed limited separation under these assumptions.

Subsequent analysis using alternative clustering configurations and deviation measures revealed that core color usage and palette size remain highly stable over time. Instead, variation appeared in cluster composition patterns, suggesting a potential three-phase production structure: early exploration, middle stability, and later diversification.

While this pattern was consistent under the modeling choices used in M3, I now question whether it reflects a genuine structural evolution across production seasons or a consequence of specific methodological settings.

Therefore, in this milestone, I refine the research question to:

Is the observed three-phase production pattern robust across alternative clustering configurations and similarity metrics?

Rather than identifying new patterns, this milestone evaluates whether the previously observed structure persists under different modeling assumptions, thereby strengthening the credibility of earlier conclusions.

2. EDA & Preprocessing: Updates (~0.5 page)

Earlier exploratory analysis revealed three key structural patterns. First, palette size distributions remained highly stable across production seasons, with most paintings consistently using between 10 and 12 colors. This finding challenged the initial assumption that seasonal variation would manifest through changes in color richness.

Second, season-level core color heatmaps indicate that Bob Ross's visual system is anchored by a stable set of recurring pigments, while variation appears primarily in the selective use of accent colors. This suggests that stylistic change is structural rather than volumetric.

Third, distance-based deviation analysis revealed a non-linear pattern across production phases: early seasons exhibited higher dispersion from the global style center, middle seasons were most concentrated, and later seasons showed renewed diversification.

Given these findings, M4 introduces methodological refinements rather than new feature construction. The binary palette representation is retained to preserve interpretability. However, alternative similarity metrics and clustering configurations are explored to test whether the observed three-phase structure depends on specific modeling choices.

No additional feature engineering was introduced at this stage, as earlier preprocessing steps were deemed appropriate for isolating structural variation in palette usage.

3. Analysis & Experiments (~2 pages)

- **Evaluating Method Dependence**

The objective of this section is to evaluate whether the previously identified three-stage production pattern—early exploration, middle stability, and late diversification—is robust across alternative modeling assumptions and implementation refinements. Rather than introducing a new segmentation goal, the purpose is to test whether the structural pattern persists when similarity construction, clustering algorithms, and dimensional representations are varied.

- **Refining Similarity Construction and Hierarchical Clustering**

A first refinement concerned the construction of the Jaccard similarity matrix. Earlier versions relied on row-wise operations, which were computationally inefficient and potentially vulnerable to asymmetry or floating-point inconsistencies. This was replaced with a fully vectorized implementation, ensuring symmetry and numerical stability.

Hierarchical clustering was then re-implemented correctly using a condensed Jaccard distance matrix. This correction was necessary because the linkage function expects a condensed distance vector rather than a full square matrix; passing the wrong format can distort the merge structure and lead to misleading dendrograms. After correction, silhouette scores peaked at 0.598 (k=2), suggesting strong separation. However, this result proved misleading: the largest cluster contained approximately 99.8% of paintings, while the smallest cluster accounted for only 0.2%. Even at higher k values, imbalance remained severe (largest cluster consistently above 94%). Although the implementation was technically correct, the resulting segmentation offers little practical insight.

- **Spectral Clustering as a Structural Alternative**

To address the limitations of hierarchical merging, spectral clustering was introduced as a graph-based alternative. While silhouette scores were lower (approximately 0.414 at k=2, and ranging between 0.26 and 0.40 across k), cluster balance improved substantially. For example, at **k=4**, the largest cluster accounted for roughly **44.9%** of paintings, and the smallest cluster for about **12.2%**, producing interpretable and balanced partitions.

Importantly, phase-level cluster composition revealed systematic differences. Early-phase paintings were disproportionately concentrated in one cluster, middle-phase paintings showed strong concentration in a dominant stable cluster, and late-phase paintings were distributed more broadly across clusters. The persistence of phase-level differentiation under spectral clustering indicates that the three-stage pattern is not an artifact of hierarchical aggregation.

- **PCA Axes and Structural Stability**

PCA provided an independent dimensional perspective. The first two principal components explain approximately **42% of total variance**, capturing interpretable axes related to palette richness and tonal contrast. Phase-level means show that early paintings occupy darker and less saturated regions of the style space, middle paintings cluster in richer and more structurally coherent regions, and late paintings exhibit greater dispersion.

To further test structural coherence, reconstruction error was computed using the first two principal components. The overall RMSE has a mean of **0.262**, a standard deviation of **0.089**, and a maximum of approximately **0.496**, with a right-skewed distribution indicating a subset of atypical paintings. Phase-level comparison shows that the middle phase has the lowest median reconstruction error and fewer extreme outliers, while early and late phases display higher dispersion. This independently supports the interpretation of middle-period structural consolidation.

- **Interpretation and Lessons Learned**

Several methodological challenges emerged. Silhouette scores proved unreliable under extreme imbalance. Hierarchical clustering tended to collapse structure in sparse binary space. Reconstruction error required careful interpretation to avoid conflating dimensional compression with stylistic deviation. Increasing PCA components reduced reconstruction error but weakened anomaly contrast.

Despite these challenges, the convergence of independent methods—corrected hierarchical clustering, spectral clustering, PCA axis differentiation, and reconstruction-based validation—consistently supports the same structural narrative. Early works appear exploratory and dispersed, middle works are cohesive and structurally stable, and late works reintroduce diversification.

The agreement across distance-based, graph-based, and dimensional frameworks indicates that the three-stage production pattern reflects a robust structural property of the data rather than a consequence of a specific modeling choice.

4. Findings & Interpretations

- **A Three-Stage Evolution in Style**

Across multiple analyses, a consistent pattern emerges: Bob Ross's palette-based style follows a three-stage evolution over the course of the show.

In the early seasons, paintings show greater variation and structural dispersion. Palette choices are less concentrated, and stylistic consistency is lower, suggesting a period of exploration before compositional habits fully stabilized. The middle period stands out as the most structurally coherent phase. Paintings from this stage display the strongest internal consistency and lowest stylistic deviation. The palette structure appears consolidated, indicating that a recognizable and repeatable "BR style" had fully formed and stabilized. In the later years, dispersion increases again. Although the overall style remains recognizable, variation in palette richness and cluster membership grows. This reflects diversification rather than instability — an expansion of stylistic expression following consolidation.

- **Stability Is Not Constant Over Time**

One of the most important insights is that stylistic stability is not uniform across seasons. The middle production period consistently shows: lower structural deviation, fewer atypical palette structures, higher internal coherence. In contrast, early and late periods show higher dispersion and more structural outliers.

This indicates that creative systems, even highly recognizable ones, move through phases of experimentation, stabilization, and diversification.

- **Why This Matters Beyond This Dataset**

First, it illustrates how recognizable style does not emerge instantly. It often requires an initial exploratory phase before a stable identity forms. Second, once established, creative systems can enter a phase of peak consistency. This may correspond to periods of strongest brand clarity and audience recognition. Finally, diversification does not necessarily imply decline. In this case, the late-stage increase in variation suggests experimentation within an established framework rather than loss of structure.

For media production, branding, and creative direction, this highlights an important principle: Strong identity often emerges through iteration, stabilizes at maturity, and later evolves through controlled diversification.

- **Implications for Understanding Creative Structure**

The findings also demonstrate that style is not simply defined by individual colors or elements. Instead, structure emerges from combinations and patterns across many dimensions.

The middle phase does not merely use “different colors” — it uses them in more consistent combinations. This suggests that style stability is fundamentally relational rather than component-based.

Recognizing that style emerges from consistent patterns rather than isolated elements is essential for assessing creative coherence, maintaining brand identity, and designing structured visual systems, which suggests that evaluating structure requires looking beyond surface-level features and focusing on pattern coherence.

- **Overall Interpretation**

Taken together, the evidence suggests that Bob Ross’s visual style did not remain static over time. Instead, it evolved through a recognizable lifecycle: exploration, consolidation, diversification.

Importantly, this pattern appears consistently across multiple independent analytical perspectives, increasing confidence that it reflects a genuine structural phenomenon rather than a modeling artifact.

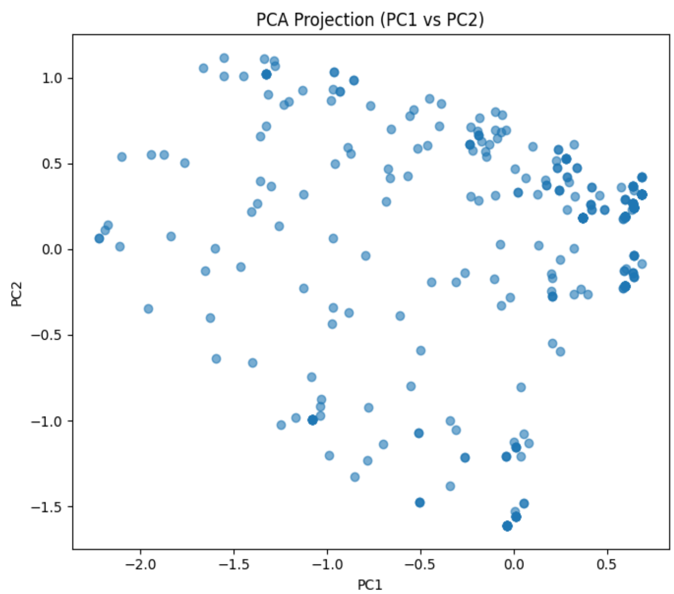
This reinforces a broader insight: Creative identity is dynamic. Stability is achieved, maintained, and eventually expanded — not assumed.

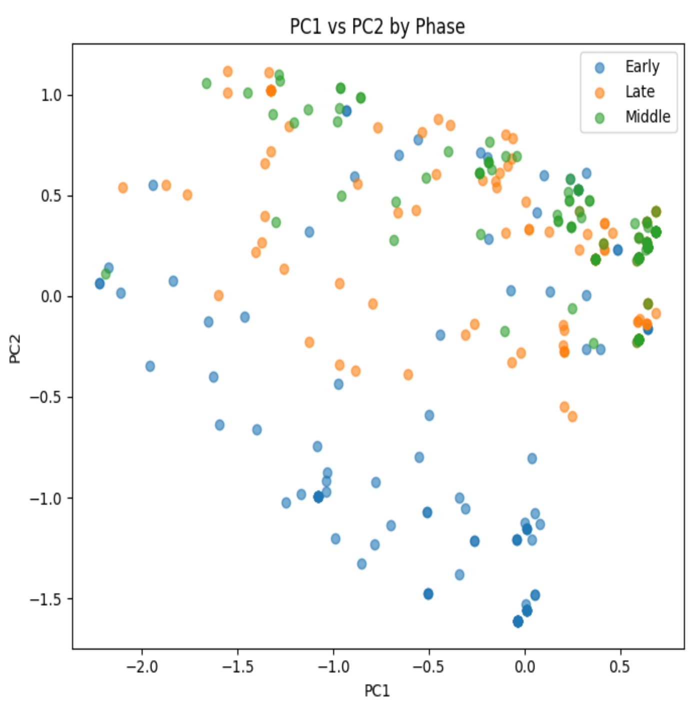
Appendix

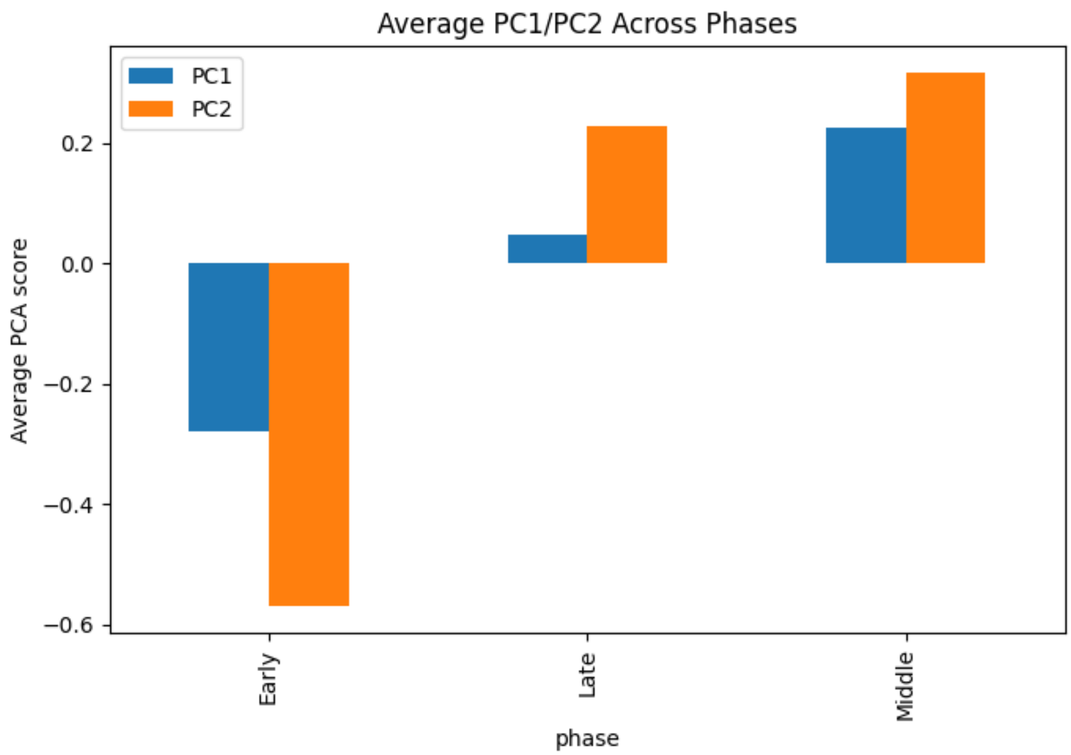
Shared GitHub Repository (Required)

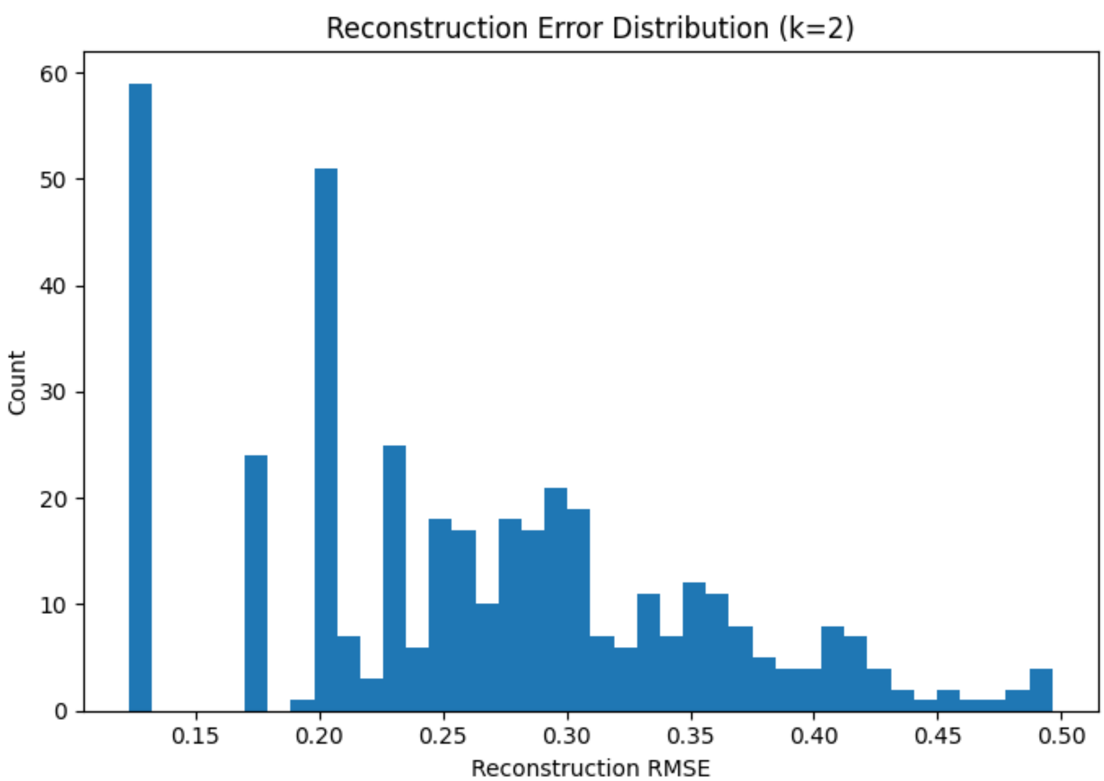
- { HYPERLINK "https://github.com/Charles-Wei77/-ba820-bob-ross-team05" \h }
- Branch: Kefei Zhang, Report: M4 - Kefei Zhang - BA820 - 2026.pdf, Notebook: M4_Kefei Zhang_BA820.ipynb

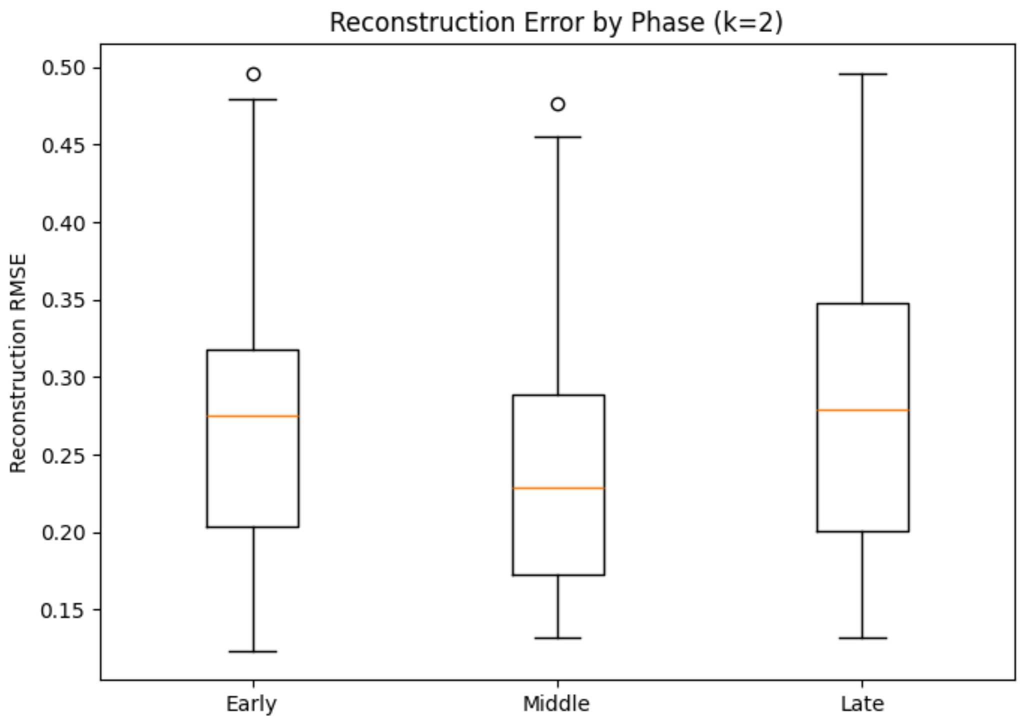
Supplemental Material

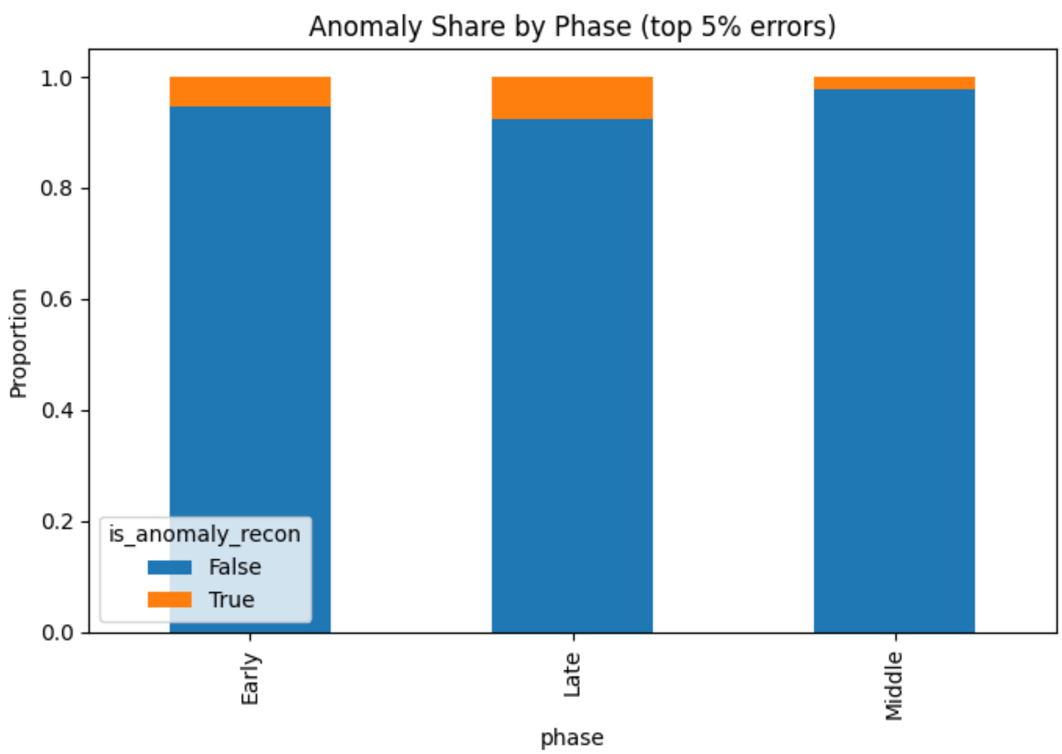


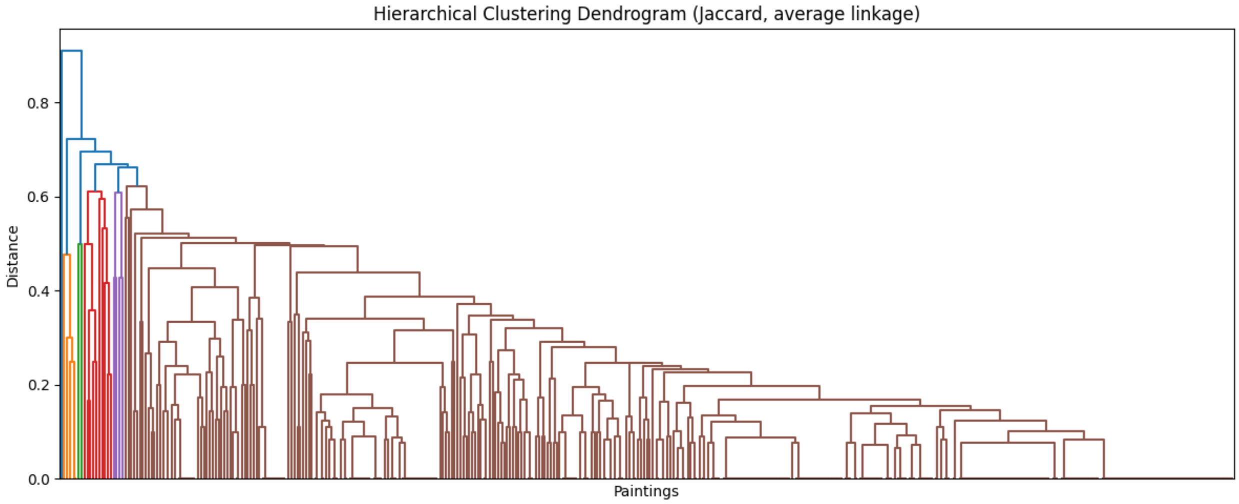



Average PC1/PC2 Across Phases
PC1
PC2
0.2
0.0
−0.2
−0.4
−0.6
Average PCA score
Early
Late
Middle
phase


Reconstruction Error Distribution (k=2)
60
50
40
30
20
10
0
Count
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
Reconstruction RMSE

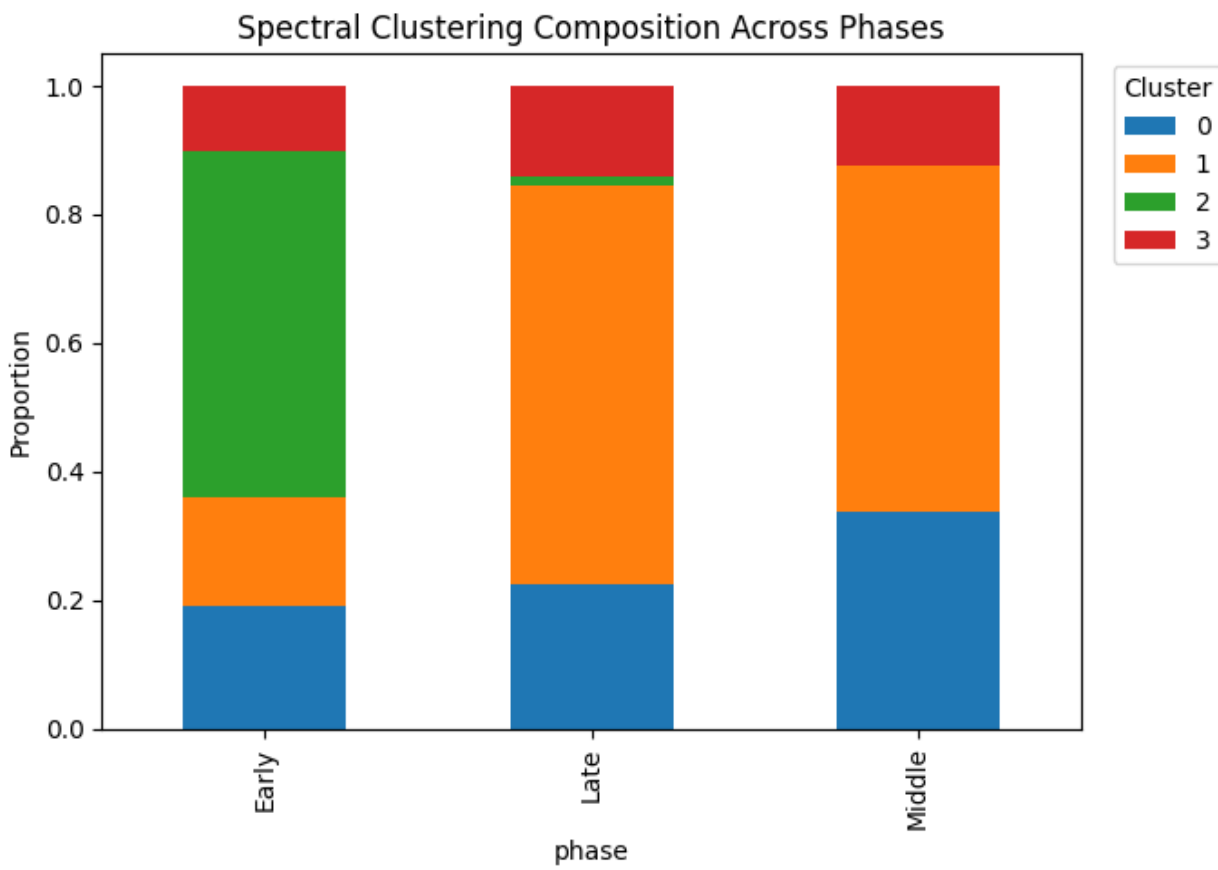

Reconstruction Error by Phase (k=2)
Reconstruction RMSE
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
Early
Middle
Late


Anomaly Share by Phase (top 5% errors)
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
is_anomaly_recon
False
True
Early
Late
Middle
phase



	k	silhouette(precomputed)	largest_cluster_%	smallest_cluster_%	deviation_range(cluster_means)
0	2	0.598	99.8	0.2	0.683
1	3	0.498	98.5	0.2	0.689
2	4	0.471	98.0	0.2	0.691
3	5	0.467	95.5	0.2	0.702
4	6	0.455	94.5	0.2	0.706
5	7	0.431	94.0	0.2	0.709
6	8	0.424	94.0	0.2	0.709

	k	silhouette	largest_cluster_%	smallest_cluster_%
0	2	0.414	85.1	14.9
1	3	0.405	67.7	13.9
2	4	0.263	44.9	12.2
3	5	0.275	42.7	7.9
4	6	0.217	36.2	2.2
5	7	0.261	32.0	3.5
6	8	0.258	34.5	1.0



Process Overview

The analysis began by representing each painting as a binary palette vector and grouping seasons into broader production phases (Early, Middle, Late). After constructing a reliable Jaccard similarity matrix, exploratory analysis examined palette dispersion and deviation patterns across phases. Clustering methods were then used to assess structural segmentation, with hierarchical clustering serving as a baseline and spectral clustering as the primary model. In parallel, PCA provided a continuous style-space representation, and reconstruction error was used to evaluate structural stability. Insights were generated by comparing results across clustering and dimensional analyses to determine whether the three-stage pattern was consistent and robust.

Data → Binary Encoding → Similarity Construction → Clustering & PCA → Cross-Method Validation → Insight Generation

Use of Generative AI Tools

- Link: **{** HYPERLINK "https://chatgpt.com/share/699cd847-c8a8-8008-b11c-07985b36a093" \h }
- After trying k-means and hierarchical clustering, I asked ChatGPT what other methods might work for my dataset. It introduced me to spectral clustering, and I asked it for guidance on how to apply it.

- I asked ChatGPT how I should approach conducting a robustness analysis on existing methods and where to begin.
- Regarding the linkage issue, I consulted Chatgpt's perspective. Feeding the NxN distance matrix directly into the linkage algorithm is problematic, as it results in our Jaccard score being “double-processed.”
- I used AI to think through whether PCA-based season analysis required reconstructing the original palette data. In discussion, we clarified that reconstruction is not strictly necessary for analyzing phase differences in principal component space, since PCA scores themselves already capture structural positioning. However, I ultimately chose to include reconstruction error as an additional structural validation step. Rather than serving as a required procedure, reconstruction was used deliberately to assess stylistic stability and identify potential structural outliers across phases.
- Previously, manually calculating similarity line by line led to issues like asymmetry and errors. I inquired about the underlying principles of ChatGPT and its implications for our future work.