

BA820 – Project M4

Cover Page

- **Project Title: Bob Ross Paintings**
- **Section and Team Number: B1 Team05**
- **Student Name: Shengqi Wei**

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

Using hierarchical clustering and association rule mining, I identified recurring palette structures across paintings. Preliminary exploratory analysis (e.g., frequency distributions and co-occurrence summaries) provided initial descriptive support for this pattern.

In M4, the core question remains unchanged: Can Bob Ross paintings be grouped into repeatable palette style structures based on structural similarity in color usage? What changes is my focus—shifting from identifying structure to evaluating its robustness. Rather than only demonstrating the existence of recurring palette groupings, I now test whether these structures remain consistent under alternative similarity definitions, clustering configurations, and parameter settings.

This refinement tests an assumption from M2: that the recurring palette groupings are not driven by a particular configuration. In M4, I check whether the key conclusions hold under reasonable alternative specifications. Consistent results would validate the template interpretation; large shifts would suggest the findings are more method-dependent and should be qualified.

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

Building upon the findings from M2, my analysis continues to investigate whether recurring color combinations reflect structured palette templates rather than isolated color choices. Earlier exploratory patterns suggested three consistent signals: (1) a small set of core colors appears in most paintings, (2) palette sizes cluster within relatively narrow ranges, and (3) preliminary co-occurrence observations indicate that certain colors frequently appear together. These patterns motivated a more structural view of palette composition.

In M2, I refined the EDA layer through additional visualizations and structural mapping. First, I added a Lorenz curve to more clearly show how uneven color usage is and to test whether most paintings rely on a small core set of colors. Second, I evaluated palette complexity using density estimates of color counts per painting and added a regression line to the scatter plot to summarize the overall trend—showing that palette expansion does not systematically increase reliance on rare colors. Finally, I visualized color co-occurrence relationships using network diagrams to make recurring “pairing” tendencies more intuitive.

In M4, no major preprocessing revisions were required beyond the data transformations already implemented previously. The dataset remains fully usable, and the main updates in this milestone focus on refining the analysis through robustness checks rather than changing the underlying data preparation. In M4, the main update is not additional preprocessing but running controlled method variants for robustness comparison while keeping the data preparation consistent.

3. Analysis & Experiments (~2 pages)

To answer our question, I followed a clear sequence. First, I checked whether color co-usage patterns reflect real structure rather than just very common colors. Then I built a similarity space and applied clustering. After that, I tested robustness across different k values, linkage methods, and distance metrics. Finally, I used PCA to interpret the structure and examine whether the grouping reflects discrete styles or a continuous deviation from a core palette. The goal here is not to find balanced style families, but to test whether paintings consistently organize around repeatable structural patterns. The consistent dominance of a core cluster supports that interpretation.

I began by addressing a key concern: some colors appear very frequently across paintings. If two colors co-occur often, that might simply be because both are common. To separate real structure from frequency effects, I computed both **co-occurrence** and **lift**. Co-occurrence shows which colors appear together often. Lift adjusts for overall frequency and highlights pairs that occur together more than expected by chance. The lift results suggest that there are stable pairing patterns in the palette system. This means the similarity space is meaningful — paintings are not random combinations of colors.

Next, I defined a **core palette**, consisting of colors that appear in more than half of the paintings. Based on this, I computed a **deviation score** for each painting using Jaccard similarity to the canonical palette. Most paintings have relatively low deviation, and only a smaller group shows high deviation. The distribution is right-skewed. This suggests that instead of many equally distinct style families, the structure may look like a strong core with varying levels of departure from it.

With this intuition, I performed **hierarchical clustering** using Jaccard similarity on the binary palette matrix. When cutting the dendrogram at $k=4$, I observed one very large cluster and several much smaller ones. Importantly, when I compared clusters with deviation scores, the clusters were ordered by deviation: the large cluster had low deviation, while the smaller clusters had progressively higher deviation. This indicates that clustering is not arbitrary. It reflects structural distance from the canonical palette.

However, the question requires that grouping be **repeatable**, so I conducted robustness tests.

First, I examined **k sensitivity** ($k=2$ to 8). Silhouette is highest at $k=2$ (around 0.60) and gradually decreases as k increases. At the same time, the largest cluster proportion remains very high across k . This suggests that the strongest natural split is between core-like paintings and more peripheral ones. Increasing k mainly subdivides the peripheral region rather than revealing multiple balanced style families. This suggests that the strongest separation is a core-versus-

periphery split ($k=2$). We still report $k=4$ as a more descriptive view that breaks the peripheral region into finer layers.

Second, I tested **linkage sensitivity** (average, complete, single). While silhouette values change slightly, the overall pattern — one dominant cluster and smaller peripheral clusters — remains consistent. This indicates that the structure is not dependent on a specific linkage choice.

Third, I tested **distance metric sensitivity** (Jaccard, cosine, Hamming). Jaccard and cosine both produce moderate separation (silhouette around 0.47–0.52), while Hamming performs worse (around 0.36). This makes sense because we are comparing set overlap structure, and Jaccard/cosine better reflect structural similarity in this context. While the exact split strength changes by metric, the same high-level pattern persists: a dominant core group and a smaller peripheral set of paintings.

After confirming robustness, I used **PCA** for interpretation rather than for clustering. When projecting paintings into two dimensions, I colored the points by deviation score. A clear gradient appears, and PC1 is strongly correlated with deviation (Pearson $r = -0.92$, $p \ll 0.001$). This shows that the main direction of variation in the palette space is essentially “distance from the canonical palette.” In other words, the palette system appears organized along a dominant structural axis rather than several unrelated style axes.

To further understand dimensionality, I performed a **reconstruction test** using PCA. Reconstructing the palette matrix using only the first two principal components results in almost the same reconstruction error as using three components ($MSE \approx 0.076$ in both cases). This indicates that most structural variation is captured in very low dimensions. The reconstruction error distribution also shows that most paintings are well approximated by low-dimensional structure, while highly deviant paintings show larger errors. When examining the most extreme painting, the PCA reconstruction moves it closer to the canonical structure, suggesting that even extreme cases lie within the same structural space.

Based on these analyses, the answer is yes — Bob Ross paintings can be grouped into repeatable palette style structures based on structural similarity. However, the structure is not best described as multiple balanced style families. Instead, the evidence supports a strong canonical core with layered deviations around it. The grouping is robust across k values, linkage methods, and distance metrics, and PCA reveals that variation is largely organized along a single dominant axis of structural deviation.

4. Findings & Interpretations (~1 page)

Bob Ross paintings don't use colors in a random way. Most episodes rely on a familiar "home base" of colors, and that core shows up repeatedly. That's why the show feels so consistent and instantly recognizable. The interesting part is that variation still exists—but it usually looks like small, controlled moves away from the same core, not a full switch into completely different color styles.

So the paintings can be grouped, but not into several equal "style families." Instead, there is one large, stable mainstream style (the core palette structure), and then a smaller set of paintings that sit farther away—more experimental, more unusual, or simply less typical. If we try to split into more groups, we mostly just break the "unusual edge" into finer slices. The big picture stays the same: there's a center, and then there are layers moving outward.

This also matches what we see in the visual maps of the data: the paintings don't form many separate islands. They form more of a smooth spectrum—from very core-like paintings to more deviant ones. That tells us the system has one dominant logic behind it, rather than many unrelated logics. Even the most extreme paintings are not "a different world"—they are edge cases within the same overall structure.

Why this matters in the real world (business relevance)

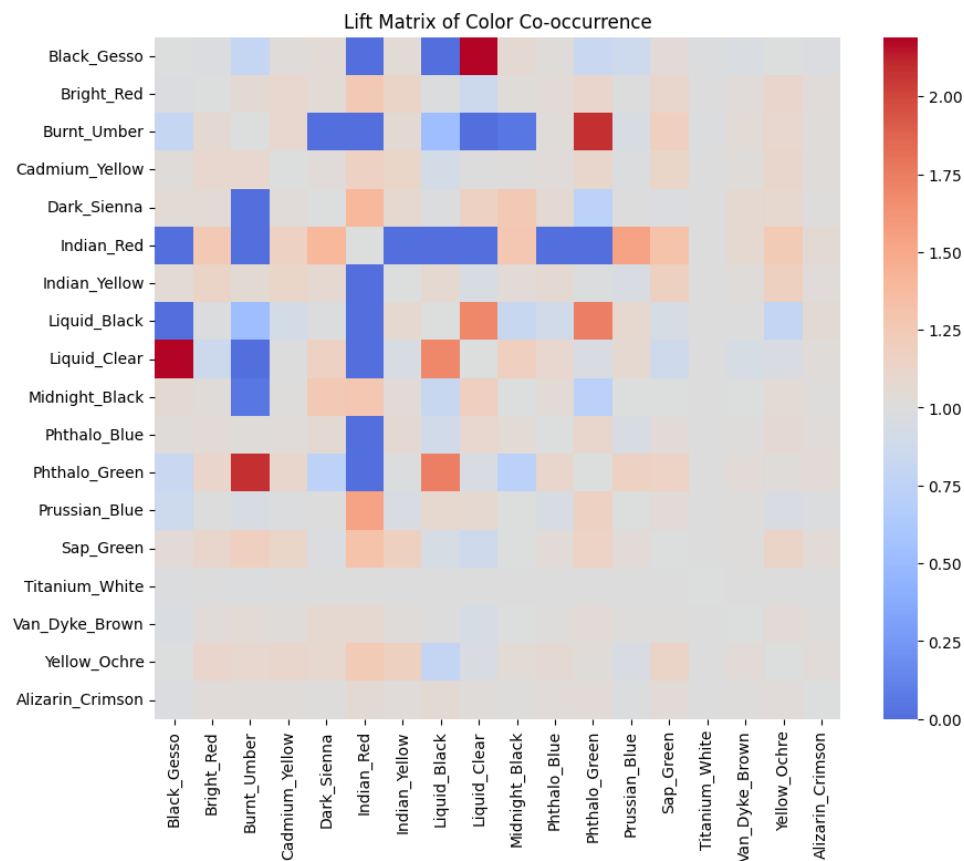
- This is basically the same problem many brands and content companies face: how do you stay consistent without being boring? Bob Ross's palette behaves like a simple, repeatable design system.
- Brand consistency: The stable core palette works like a brand guideline. Think of Coca-Cola's red, Apple's clean minimal style, or a company's design system colors. When the core stays stable, the audience feels "this is the same brand," even when details change.
- Scalable production: A repeatable palette structure makes it easier to produce content at scale. If the system is consistent and constrained, you can create hundreds of episodes without the style drifting. That's a real operational advantage for long-running series, marketing teams, or anyone producing frequent content.
- Controlled innovation: The smaller "outer" groups show how novelty happens without breaking identity. Businesses want this exact balance: keep the product recognizable but add just enough variation to keep people interested.
- Quality control / auditing: If we can detect which outputs are "core-like" versus "edge cases," this becomes a practical tool. A studio, brand team, or creator could use the same idea to flag off-brand designs, identify unusually experimental releases, or intentionally plan variety while staying within a coherent style.

Appendix

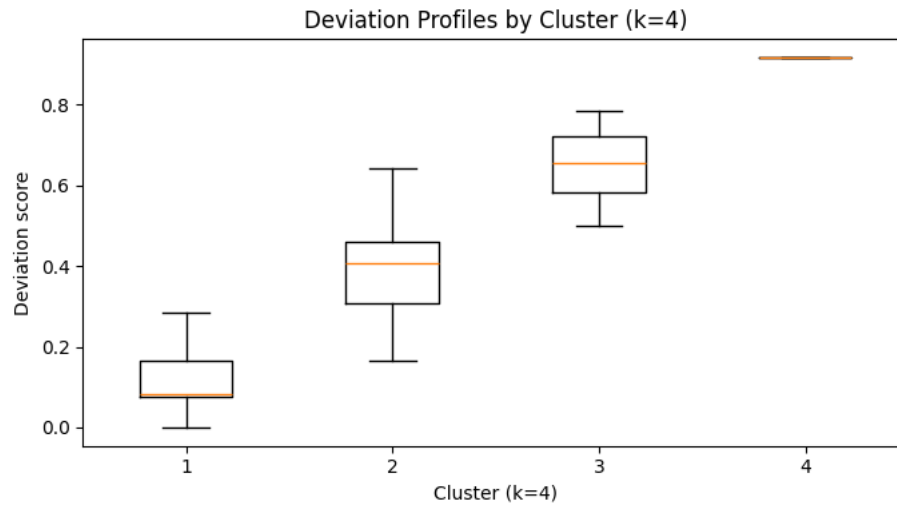
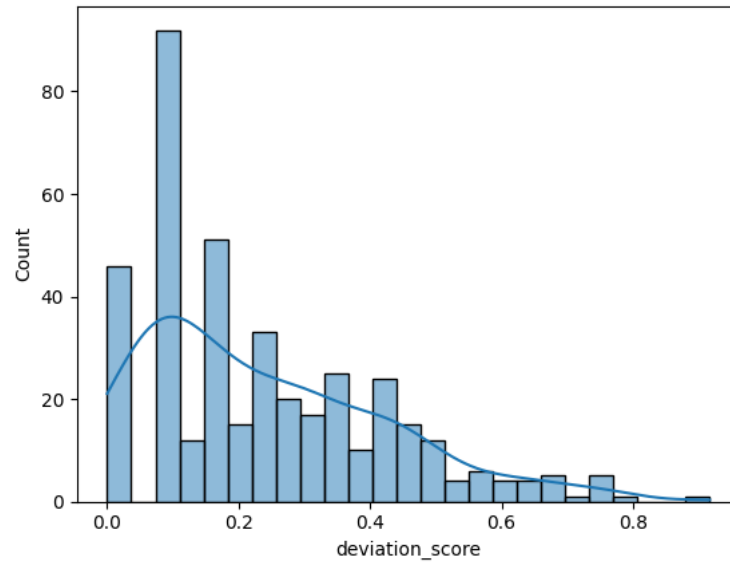
Shared GitHub Repository (Required)

- <https://github.com/Charles-Wei77/-ba820-bob-ross-team05>
- Branch: Shengqi Wei
- Report: M4 Shengqi Wei - BA820 - 2026.pdf
- Notebook: M4_Shengqi Wei_BA820.ipynb

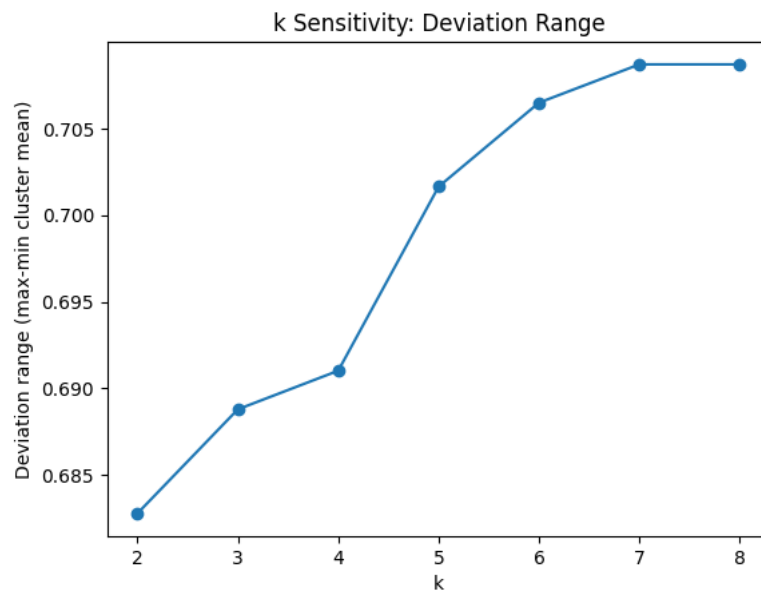
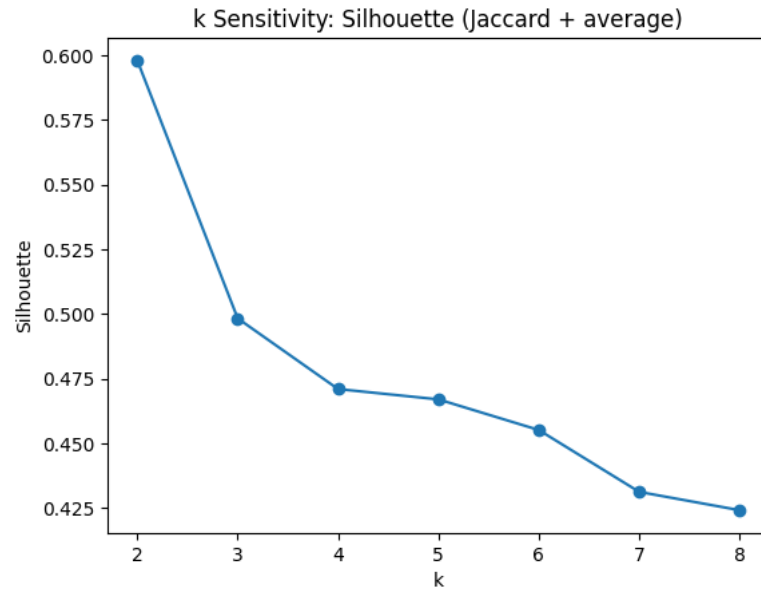
Supplemental Material (Highly Recommended)



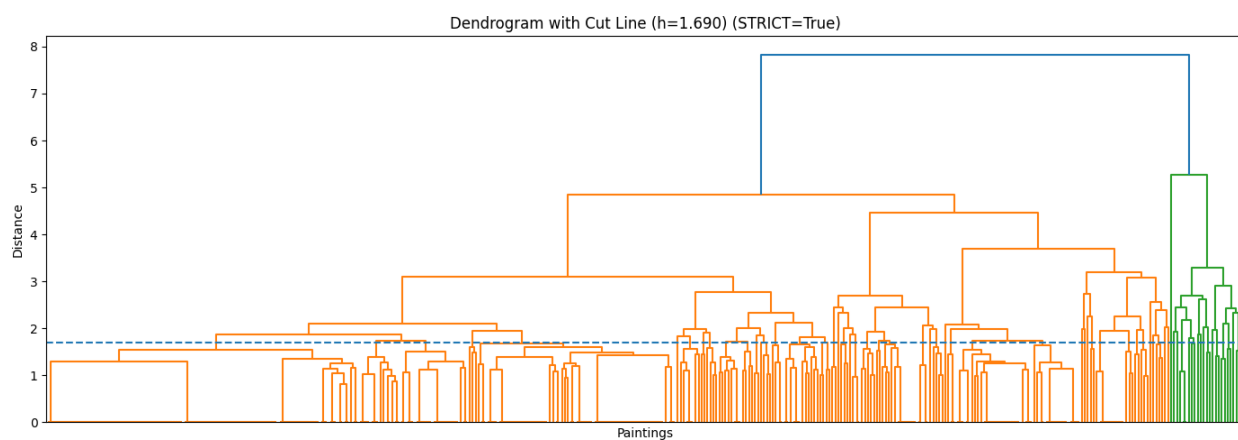
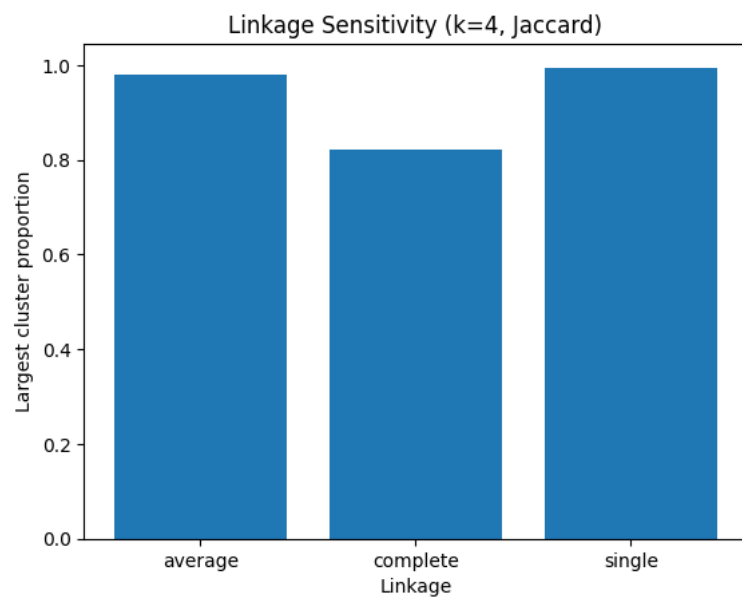
deviation_score	
count	403.000000
mean	0.235566
std	0.185986
min	0.000000
25%	0.083333
50%	0.200000
75%	0.357143
max	0.916667



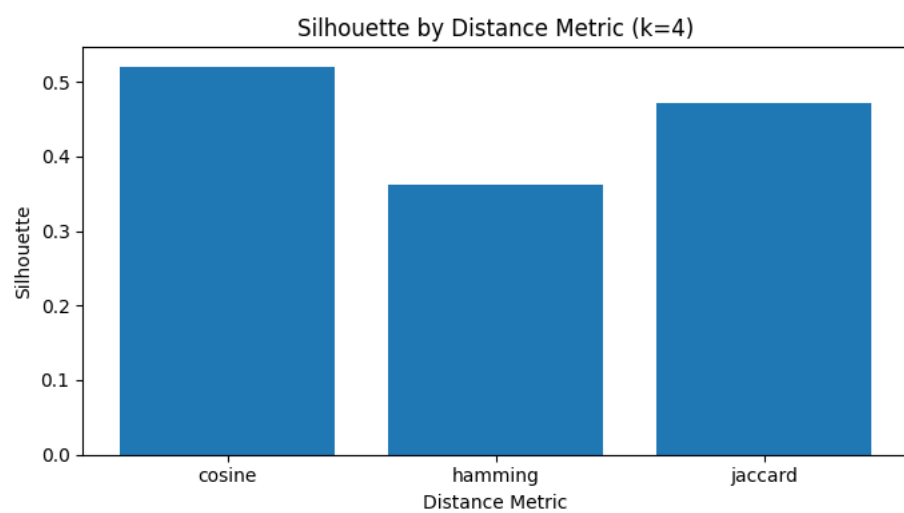
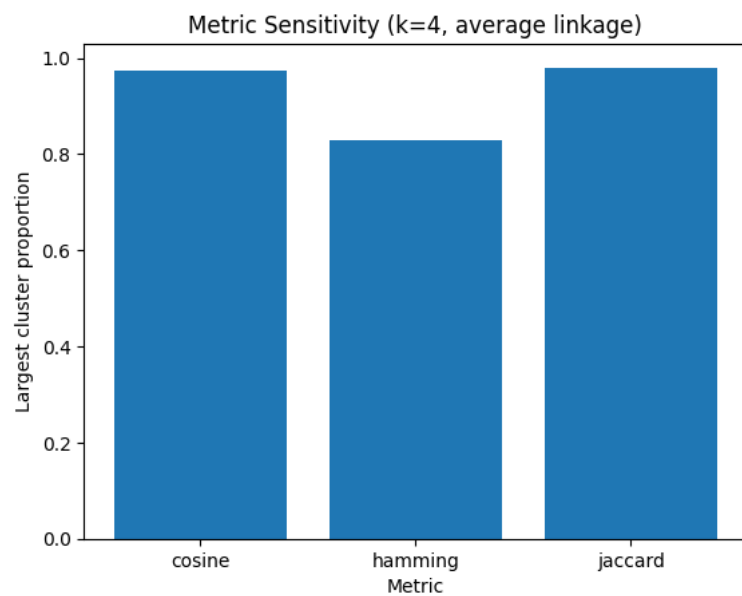
	metric	linkage	k	silhouette	largest_cluster_pct	smallest_cluster_pct	min_cluster_mean_deviation	max_cluster_mean_deviation	dev_range
0	jaccard	average	2	0.597944	0.997519	0.002481	0.233872	0.916667	0.682795
1	jaccard	average	3	0.498231	0.985112	0.002481	0.227872	0.916667	0.688795
2	jaccard	average	4	0.470907	0.980149	0.002481	0.225646	0.916667	0.691021
3	jaccard	average	5	0.466916	0.955335	0.002481	0.215030	0.916667	0.701636
4	jaccard	average	6	0.455069	0.945409	0.002481	0.210188	0.916667	0.706479
5	jaccard	average	7	0.431152	0.940447	0.002481	0.207977	0.916667	0.708690
6	jaccard	average	8	0.424071	0.940447	0.002481	0.207977	0.916667	0.708690



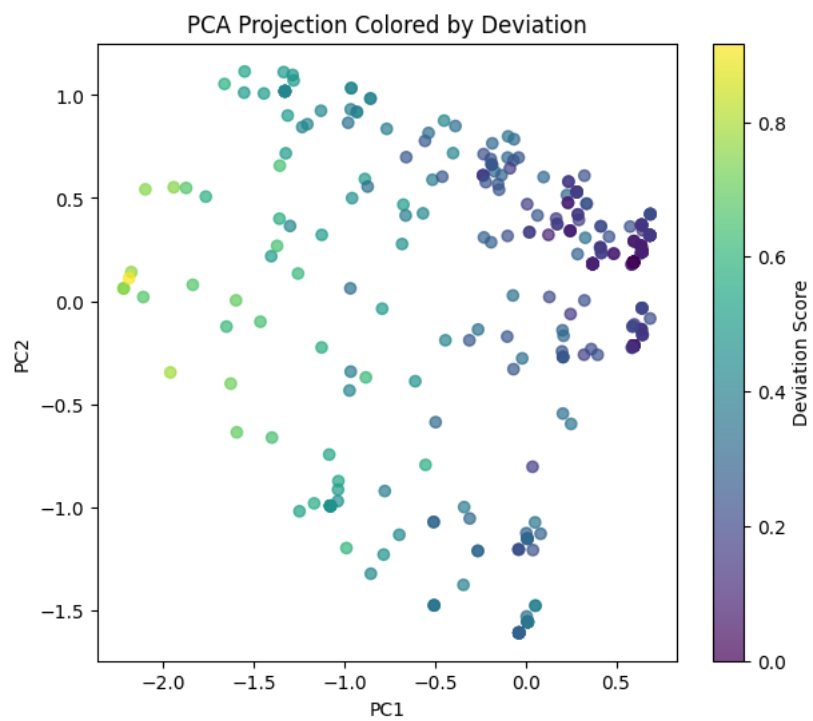
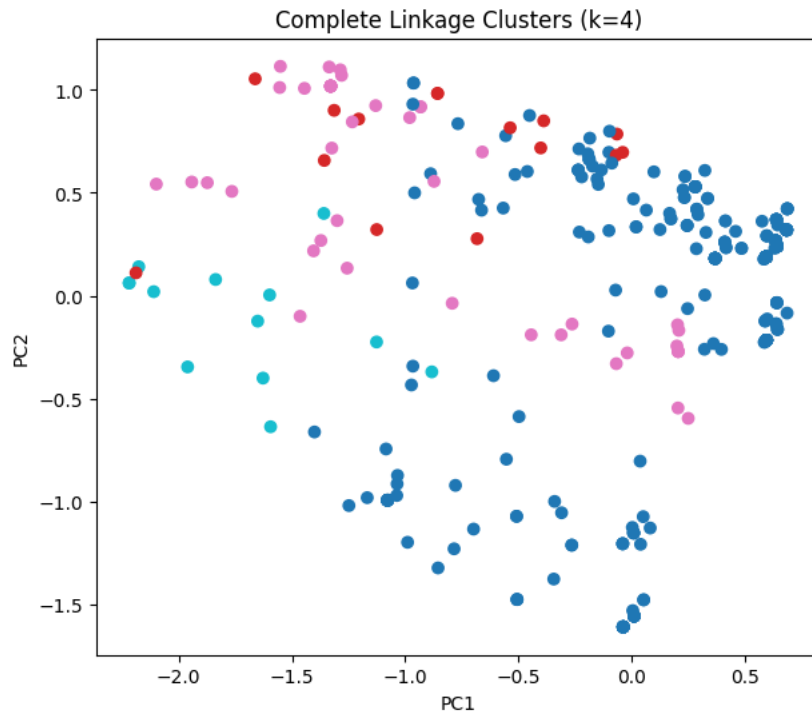
	metric	linkage	k	silhouette	largest_cluster_pct	smallest_cluster_pct	min_cluster_mean_deviation	max_cluster_mean_deviation	dev_range
0	jaccard	average	4	0.470907	0.980149	0.002481	0.225646	0.916667	0.691021
1	jaccard	complete	4	0.351004	0.821340	0.032258	0.182511	0.681600	0.499089
2	jaccard	single	4	0.498293	0.995037	0.002481	0.232729	0.916667	0.683938

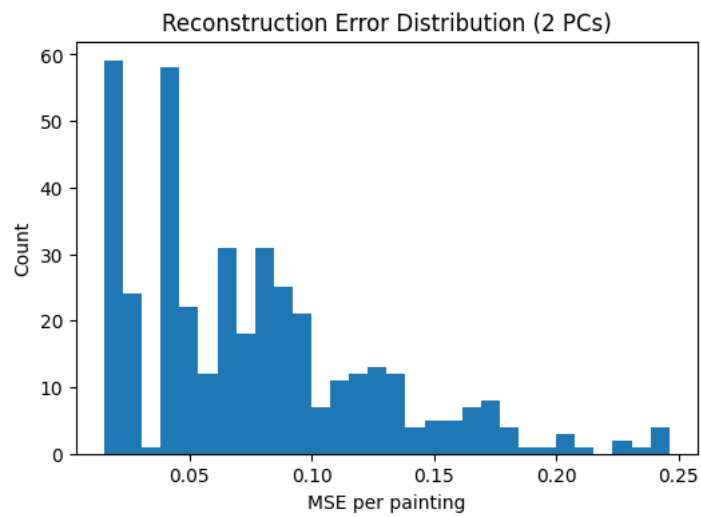
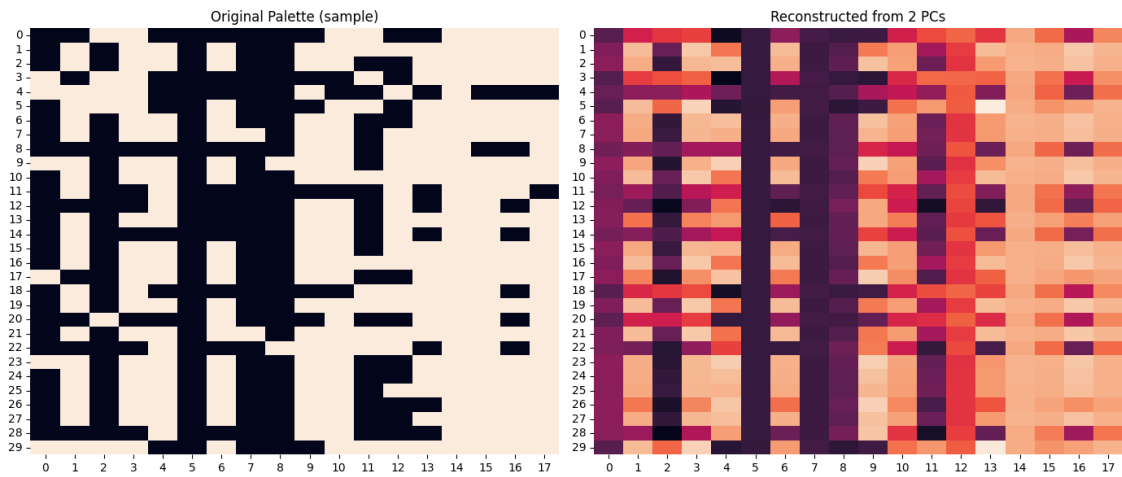
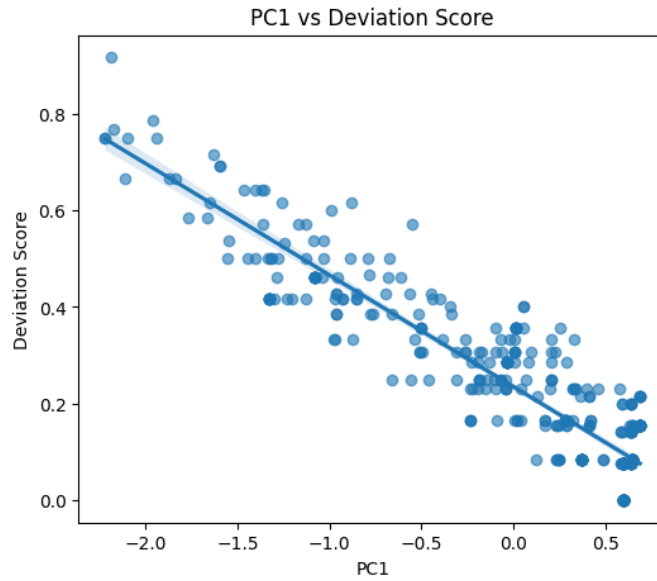


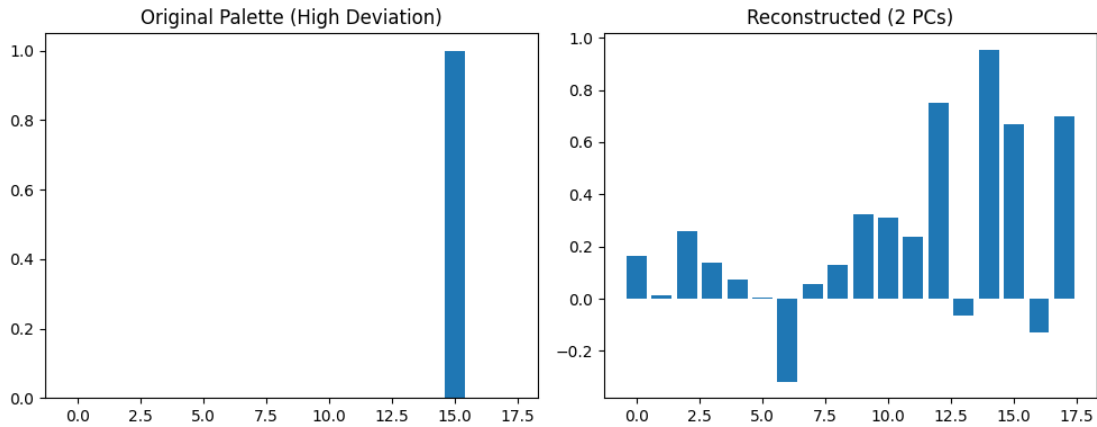
	metric	linkage	k	silhouette	largest_cluster_pct	smallest_cluster_pct	min_cluster_mean_deviation	max_cluster_mean_deviation	dev_range
2	cosine	average	4	0.520221	0.975186	0.004963	0.222977	0.750000	0.527023
1	hamming	average	4	0.361921	0.828784	0.002481	0.179339	0.609890	0.430551
0	jaccard	average	4	0.470907	0.980149	0.002481	0.225646	0.916667	0.691021



	metric	linkage	k	silhouette	largest_cluster_pct	smallest_cluster_pct	min_cluster_mean_deviation	max_cluster_mean_deviation	dev_range
10	cosine	average	4	0.520221	0.975186	0.004963	0.222977	0.750000	0.527023
11	hamming	average	4	0.361921	0.828784	0.002481	0.179339	0.609890	0.430551
0	jaccard	average	2	0.597944	0.997519	0.002481	0.233872	0.916667	0.682795
1	jaccard	average	3	0.498231	0.985112	0.002481	0.227872	0.916667	0.688795
2	jaccard	average	4	0.470907	0.980149	0.002481	0.225646	0.916667	0.691021







Process Overview

I first converted each painting into a simple yes/no color profile so that palette usage could be compared in a consistent and structured way.

I then examined whether certain colors tend to appear together in stable patterns, rather than just co-occurring because some colors are common. This step ensured that the similarity I measure reflects real structure.

Next, I defined a canonical core palette and calculated how far each painting deviates from it. This allowed me to interpret variation as movement away from a shared center rather than as separate styles.

I grouped paintings based on structural similarity and tested whether the grouping remained stable when I changed key settings, such as the number of groups, the grouping method, and the similarity definition. This helped confirm that the pattern was repeatable and not dependent on a single parameter choice.

Finally, I used a simple two-dimensional visualization to interpret what the grouping means in practical terms—whether the paintings form distinct style families or mostly cluster around a dominant core with layered deviations—and connected this structure to real-world ideas like brand consistency, controlled variation, and scalable creative systems.

Use of Generative AI Tools

Link: <https://chatgpt.com/share/699d0a46-0260-8003-a92a-ad8240fb1197>

I asked ChatGPT for an overall in-depth approach and whether there are any other new methods to recommend.

I inquired about approaches to testing model stability and what aspects can be evaluated.

In the previous milestone, we explored the co-occurrence relationships between colors. However, it's possible that colors appear frequently individually, leading to high simultaneous occurrence rates. To address this issue, I consulted ChatGPT. Among the approaches it suggested was using

lift to determine whether the probability of colors appearing together exceeds what would be expected under random independence. This concept was also covered in our earlier coursework, prompting us to upgrade the heatmap.

Choosing different linkages involves trade-offs. I'm torn between complete and average. Using average leads to extreme imbalance, while using complete feels like it's due to noise issues. I'm consulting AI for a solution approach.