BA820 – Project M4

Cover Page

- **Project Title:** Exploring Latent Color Patterns and Similarity in Bob Ross Paintings
- **Section and Team Number:** B1 Team 3
- **Student Name:** Kean Zhu

1. Refined Problem Statement & Focus

In this milestone, I continued investigating whether Bob Ross's paintings exhibit meaningful structural patterns in how paint colors are combined. The central question remained grounded in our original motivation from M1- understanding whether palette usage reflects specific structures rather than random variation. However, based on integrated results in M3, especially the repeated presence of a single dominant cluster, my framing in M4 shifts away from searching for "clean separable styles" and toward explaining why a dominant palette regime persists and what forms of deviation are still meaningful within it.

In M4, rather than treating a big dominant cluster as an unsatisfactory outcome, I treated it as evidence of a strong shared "standard" palette, and evaluated whether the remaining variation is better understood as bounded deviations around the main cluster.

The validated assumptions remained that:

- Bob Ross relies heavily on a stable core set of paints
- Palette similarity across paintings is structurally high
- Variation occurs primarily through selective adjustments, not fully distinct regimes.

M4 therefore reframed the project from detecting clusters to interpreting structural consistency and meaningful deviation. More specifically, identifying which paint combinations define the backbone of Bob Ross's palette, and which combinations signal a handful of focused theme/style differences.

2. EDA & Preprocessing: Updates

There are no additional EDA or preprocessing updates in M4 beyond what was completed in M3. The dataset remains complete, binary, and directly aligned with the analytical objective of examining palette overlap. Because the data structure remained unchanged and prior preprocessing was sufficient for our similarity-based analysis, further cleaning or modification was unnecessary and could possibly alter our interpretability.

I believed the structural question I was interested in did not require new engineering. The only extension in M4 involves deeper integration of text analysis. Titles were tokenized and grouped to examine whether thematic patterns align with palette clusters. However, this step did not require additional preprocessing beyond standard token cleaning that was adopted in M3 from my teammate, as titles are short and structurally simple.

Thus, EDA and preprocessing remained unchanged because the existing structure sufficiently supports structural evaluation.

3. Analysis & Experiments

In M4, I focused on stress-testing whether the dominant baseline with small deviations pattern observed in M3 is structural or parameter-driven. Rather than adding new methods, I use targeted tuning and robustness checks to show how each conclusion was reached.

Hierarchical Clustering

In our project, hierarchical clustering using Jaccard distance remains the central structural model. We have explicitly compared Hamming distance vs Jaccard distance, and we proved that Jaccard displays a broader distribution compared to Hamming distance. After selecting Jaccard, we applied hierarchical clustering with average linkage. The results revealed to us that palette usage is dominated by a consistent shared core, with only minor deviations.

To avoid mistaking the dominant cluster as a failure, I tested two complementary dendrogram cuts. First was a fixed-k cut (k=5) to get a stable macro view, and second was distance-threshold cuts to check whether meaningful separations appear at any similarity scale. The threshold results showed no sharp separation point, reinforcing that the main structural signal is a dominant baseline with small, specific deviations.

I also revisited DBSCAN as a robustness check under a density-based assumption. Across settings (figure 3), small eps values produced high noise rates, while larger eps values reduced noise but collapsed to only 2-3 clusters with a dominant largest cluster. Even in the mid-range, DBSCAN produced only 3-7 clusters with a still-dominant largest cluster plus substantial noise. This pattern mirrored our hierarchical clustering and reinforced the 95% dominant structure as a persistent property of the palette space.

Association Rule Mining

In M3, after removing near-universal paints, we used a global rule set as a baseline, and compared it to within-cluster rules when sample size was sufficient. This separated the common dominant palette co-usage patterns from cluster-specific couplings. With the M3 setting (global_min_lift = 1.30) and (MIN_CLUSTER_N = 10) before mining within-cluster rules, it produced very few rules in the dominant cluster, which made it hard to summarize the baseline palette association in a clean way, although this strict setup reinforced us on the main structural interpretation that most paintings follow a baseline palette framework ,and what differentiates rare clusters is not random noise, but small and focused shifts where certain paint pairs show up together much more often.

In M4, I tuned rule strictness to better balance two goals. I wanted to identify deviation signatures and describe a readable baseline palette grammar. I first lowered the trivial-paint cutoff to remove a broader set of near-universal paints before mining rules, making the remaining co-usage patterns more interpretable. Then I lowered the lift threshold to 1.0 as a diagnostic to surface clean 1->1 rules that strict lift filtering can hide, while still separating

cluster-only patterns for interpretation. I briefly tested a smaller MIN_CLUSTER_N to confirm that tiny clusters still do not produce stable, support-backed rules even under relaxed settings. This tuning process shifted the role of rules from only flagging statistically surprising pairs, to the global rules now better represent the foundation combinations Bob Ross repeatedly relies on, which could offer us more insight about a stable core palette. It also showed how conclusions change with parameters and reinforced that the dominant cluster reflects baseline structure, while smaller clusters concentrate distinctive couplings.

Overall, the M4 tuning process made the association-rule analysis more aligned with the business goal. The strict version was good for highlighting deviation signatures but could under describe the core palette structure. The relaxed lift improved baseline interpretability and enabled global 1 to 1 couplings. Together, these parameter experiments allowed me to look inside the co-usage structure, separating globally stable baseline pairings from cluster-specific couplings. It also strengthened the reframed conclusion that Bob Ross's palettes follow a dominant core 'grammar' with bounded deviations.

Text Analysis (Within Cluster)

Our dataset includes unstructured title texts, so I used within-cluster title analysis to test if palette deviations align with thematic cues in episode titles. I implemented BoW using CountVectorizer and summarized token usage within each cluster using relative frequencies.

The result showed that the dominant palette cluster shows broad, typical Bob Ross themes like landscapes and seasons, consistent with it functioning as the "default". Some smaller clusters show unusually concentrated tokens (e.g., "lady," "contemplative," "girl," "indian"), suggesting that at least a subset of palette deviations align with less common subject matter, rather than random variation. At the same time, the analysis revealed many clusters still share similar landscape vocabulary, so title text explains only part of the palette structure and should be treated as an interpretational lens rather than a primary driver of clustering.

Overall, the text analysis strengthened our narrative by adding a simple connection between palette clusters and title themes. It demonstrated that the baseline cluster reflects common Bob Ross landscape language, while some deviation clusters are marked by rarer subject terms. Again, this reinforced us that inside our dominant clusters, most paintings align with a broad baseline framework, while a few small clusters reflect focused deviations rather than fully separate styles

4. Findings & Interpretations

Stepping back from the mechanics of the methods, the most important domain insight we have drawn is that Bob Ross's color system behaves like a single, reusable palette framework with a small number of recognizable deviation templates, instead of a collection of distinct styles. We believe he maintained a stable baseline recipe and expressed variation through controlled,

repeatable adjustments, especially in the accent and specialty paints, while still preserving the core palette's overall identity. Since the structure is defined by a persistent baseline, the key task is interpreting what counts as a meaningful deviation.

The clustering results supported this interpretation because the dominant grouping is overwhelmingly large. This structure implied that the baseline palette is the norm, and the smaller groups exist as exceptions. The distance-threshold view added nuance. Many small clusters only appeared under very strict similarity thresholds, which means deviations exist but are narrow and specific. In deviation clusters, these relationships became more concentrated, as specific pairs showed up together far more often than they did in the baseline.

Association rules translated Bob Ross's painting into palette grammar. The baseline framework is supported by stable co-usage relationships that happen broadly across paintings, which reinforced the idea that Bob Ross relied on a disciplined internal structure, rather than improvising palettes from scratch. For example, strong pairings (Figure1) like {Dark Sienna & Midnight Black} (symmetrical) and {Phthalo Green -> PrussianBlue} appeared across the dataset, reinforcing the idea of a consistent internal palette grammar. The more interesting domain insight is what happens in deviation clusters. They showed concentrated dependencies, where certain paints are grouped together much more often inside those clusters. This suggests that deviation palettes are not simply "more colors" or "different colors" in Art. We do not need different overall systems, but coordinated design choices.

Title text adds context but also helps rule out an intuitive explanation. If subject matter were driving palette structure, we would expect clusters to separate more clearly by title themes. Instead, titles remained dominated by broad landscape and seasonal language across clusters, and the baseline cluster spanned most themes (Figure2). The implication is that Bob Ross's palette consistency is not simply because he painted the same kinds of scenes, but because he maintained a stable color identity that could be adopted across many subjects. Although rare clusters did show focused tokens, they described what makes a deviation cluster feel different, rather than explaining the overall structure.

In conclusion, Bob Ross's artistic identity is defined by structural stability with bounded variation. The baseline palette provides a consistent visual signature that viewers recognize, and deviation templates allow novelty without breaking that signature. For art-focused audiences, this supports the idea that stylistic diversity in his work is expressed through controlled palette emphasis shifts instead of multiple competing styles. More broadly, this shows how a consistent core can support variety through structured variation

Appendix

1. Shared GitHub Repository

shared team GitHub repository:
https://github.com/yujiunzou/BA820-Unsupervised-ML-Project

Individual work location:
m4_individual_refinement/M4_Kean/

2. Supplemental Material

Figure 1: Association Rule Table - 1 antecedent and 1 consequent

Filtered rules with 1 antecedent and 1 consequent:

	antecedents	consequents	support	confidence	lift	antecedent_len	consequent_len
39	Dark_Sienna	Midnight_Black	0.709677	0.986207	1.253758	1	1
40	Midnight_Black	Dark_Sienna	0.709677	0.902208	1.253758	1	1
107	Burnt_Umber	Sap_Green	0.124069	0.909091	1.197267	1	1
110	Liquid_Clear	Midnight_Black	0.119107	0.941176	1.196511	1	1
141	Indian_Yellow	Sap_Green	0.650124	0.897260	1.181686	1	1
142	Sap_Green	Indian_Yellow	0.650124	0.856209	1.181686	1	1
163	Indian_Yellow	Yellow_Ochre	0.689826	0.952055	1.173327	1	1
164	Yellow_Ochre	Indian_Yellow	0.689826	0.850153	1.173327	1	1
167	Liquid_Clear	Dark_Sienna	0.106700	0.843137	1.171670	1	1
179	Phthalo_Green	Prussian_Blue	0.218362	0.758621	1.162449	1	1
223	Phthalo_Green	Sap_Green	0.250620	0.870690	1.146693	1	1
252	Indian_Yellow	Bright_Red	0.655087	0.904110	1.135066	1	1
253	Bright_Red	Indian_Yellow	0.655087	0.822430	1.135066	1	1
263	Sap_Green	Yellow_Ochre	0.697270	0.918301	1.131728	1	1
264	Yellow_Ochre	Sap_Green	0.697270	0.859327	1.131728	1	1

Figure 2: Word Distribution Across Clusters

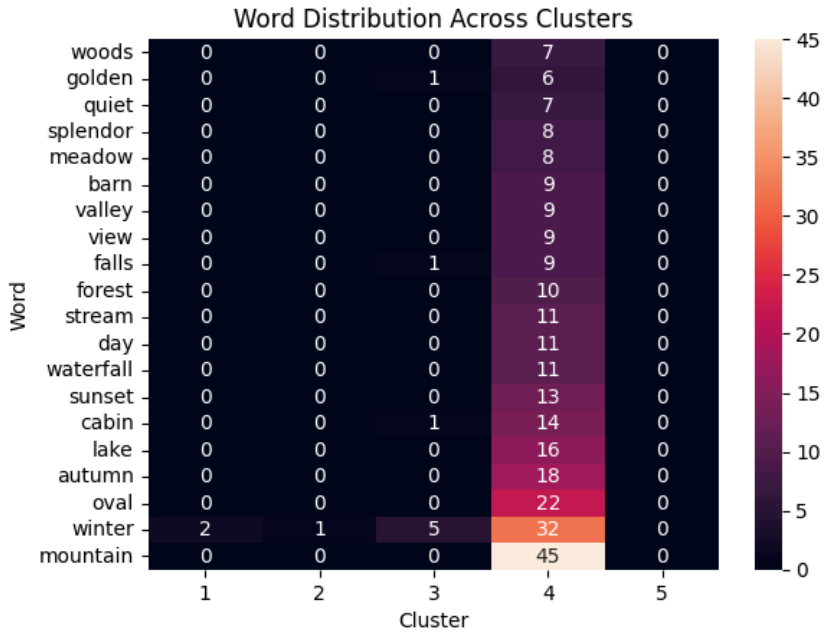
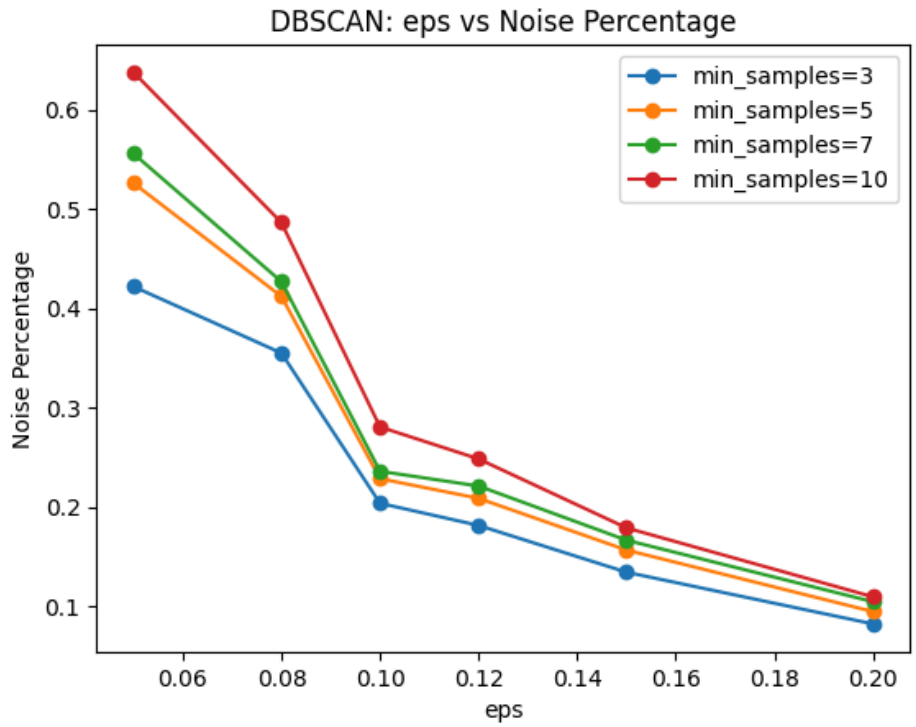


Figure 3: DBSCAN Epsilon vs Noise Line Chart



3. Process Overview

1. I refined the M4 focus from "finding distinct clusters" to testing whether Bob Ross's palettes are organized around a dominant baseline with meaningful, bounded deviations.

2. I revised EDA and preprocessings to confirm nothing required updates.

3. I evaluated hierarchical clustering (fixed-k and distance-threshold cuts) to examine whether segmentation emerges at any similarity scale or whether patterns merge gradually.

4. I used association rule mining (global baseline vs within-cluster rules with threshold and trivial-paint tuning) to identify stable global paint couplings and cluster-specific dependencies that characterize deviations.

5. I revisited DBSCAN from M2 and adopted it as a density-based robustness check to assess whether alternative clustering assumptions reveal clearer regimes or reinforce the dominant-baseline interpretation.

6. I incorporated within-cluster title text analysis as an interpretation layer and synthesized evidence across methods to reach a unified structural conclusion.

4. Use of Generative AI Tools

I used ChatGPT to assist with coding refinement, implementation clarification, and visualization design throughout this milestone. Specifically, I asked help on structuring pandas operations on filtering 1-to-1 association rules, improving text vectorization and tokenization decisions, computing frequency-based metrics for cluster interpretation, and organizing DBSCAN grid search outputs (including noise percentage analysis and visualization). All analytical decisions, parameter selections , interpretation of clustering and association rule results, and integration into the business/domain narrative were conducted and validated independently. The AI was used as a technical support tool to improve code clarity and presentation, not to generate results or substitute analytical reasoning.

Chat Link:https://chatgpt.com/share/699cd305-ef68-8007-b715-120bd12aad7d