BA820 – Project M2

Cover Page

- **Project Title:** Evaluating Structure in Bob Ross Paintings: Discrete Palette Styles or Gradual Variation?
- **Section and Team Number:** B1 Team 3
- **Student Name:** Kean Zhu

1. Refined Problem Statement & Focus

In this phase, I focused on whether Bob Ross's paintings share meaningful color palette patterns. Specifically, I looked at whether paintings can be grouped into consistent palette "templates", evaluating if the observed structure reflects some clearly separable palette styles or represents gradual variation around a shared core palette.

The overall question has not changed from M1, but my focus narrowed. Our M1 analysis did not reveal strong evidence of seasonal structure in palette usage. Season-level summaries showed that the average number of paints used per painting remains relatively stable across seasons, with only minor fluctuations. In addition, paint usage rates by season demonstrated consistent dominance of a core set of colors, and distributions of palette size and composition also exhibited substantial overlap across seasons, indicating that seasonal grouping is not meaningful. This motivated me to shift away from time-based analysis and toward a similarity-based exploration of global palette structure, allowing the analysis to more directly address whether recurring color combinations and potential palette styles exist across the full set of paintings.

Additionally, the rest of our analysis provides evidence that paintings frequently rely on similar combinations of colors. Pairwise Jaccard similarity between paintings is relatively high on average, with a substantial number of painting pairs exhibiting very high similarity and even exact matches in paint usage. This pattern indicates that Bob Ross often reuses nearly identical color combinations across episodes, supporting the idea that a small set of reusable palette "templates" underlies much of the observed variation. This led me to focus on clustering paintings by similarity instead of tracking changes over time.

2. EDA & Preprocessing: Updates

There is no additional data cleaning or transformation performed in M2 beyond what was completed in M1. The dataset is complete, contains no missing values or duplicate records, and the binary paint indicators directly reflect the analytical objective of comparing palette overlap. Because the data structure remained unchanged and prior preprocessing was sufficient for similarity-based analysis, further cleaning or modification was unnecessary and could have distorted the interpretation of genuine palette variation.

Our key findings from M1 directly informed the analytical choices in this milestone. The key takeaways from M1 that guided this phrase are:

- Bob Ross relies heavily on a small set of core colors that appear in almost every painting.
- Variation comes mainly from whether he includes certain accent colors, not from big shifts in his overall palette.
- Paintings tend to be very similar to each other in terms of color usage, which made similarity-based methods a better fit than distance-based ones.

Because the data was already prepared and the structure was clear, no additional transformations were needed for M2. Binary paint indicators were converted to integers only to allow distance and similarity calculations. There was also no normalization applied, since Jaccard distance does not depend on scale and is well suited for binary data. The similarity distributions and co-occurrence patterns from M1 were reused to motivate clustering choices, rather than repeated as exploratory results.

3. Analysis & Experiments

Hierarchical Clustering

Two unsupervised methods were applied to examine the potential structures in Bob Ross paintings based on how paint colors are combined, and to assess whether these structures correspond to distinct palette styles.

The primary analytical method applied is hierarchical clustering using Jaccard distance, in order to identify whether paintings group into recurring palette structures at different similarity levels. Jaccard distance aligns with the binary representation of paint usage, and average linkage was chosen to balance sensitivity to near-duplicate palettes with broader group-level similarity. This choice helps to reduce the influence of extreme pairs while capturing overall palette overlap across clusters. Moreover, it does not require specifying a single number of clusters in advance and allows structure to be examined at multiple levels of detail.

The dendrogram (Figure 1) was constructed and shows gradual merging rather than sharp breaks, indicating that palette similarity does not form cleanly separated groups. Smaller values of k collapse most paintings into a single dominant cluster, while larger values produce many extremely small clusters, often consisting of one or two paintings. Because there is no clear “natural” number of clusters and cluster boundaries became increasingly subjective as k increased, it was a key challenge to select k.

Therefore, the number of k was selected based on interpretability, cluster size balance, and domain relevance rather than any optimization. The k=4 cluster was subjectively judged and selected as a compromise that preserves a dominant core cluster, while still capturing few meaningful palette variants. As shown in Figure 2, this choice results in one large core cluster alongside a small number of minor variants, rather than many fragmented groups. Overall, the clustering revealed one large core cluster and several smaller, less cohesive variants, rather than revealing clearly distinct palette categories. This was consistent with earlier similarity analysis and reinforced the idea that palette structure exists, but without sharp boundaries.

DBSCAN

To further assess the structure of Bob Ross’s palette usage, density-based clustering DBSCAN was applied. It was used as a comparison method to test whether Bob Ross paintings form very

tight, clearly separated groups based on color combinations. Unlike hierarchical clustering, it only groups paintings when many of them are extremely similar, making it useful for checking whether distinct palette styles truly exist .

The neighborhood radius ε defines how similar two paintings must be to be considered neighbors. Because Jaccard distance is used, ε corresponds directly to a similarity threshold: two paintings are neighbors if their Jaccard similarity is at least 1-ε. Smaller ε values therefore require paintings to be extremely similar to form a cluster, while larger ε values allow more loosely related palettes to be grouped together.

Multiple ε values were explored (ε = 0.05, 0.08, 0.10, 0.12, 0.15, 0.20). These values were chosen to span a range from very strict similarity requirements to more permissive thresholds.. This range was informed by the earlier Jaccard similarity distribution, which showed that many painting pairs cluster between moderate and high similarity values rather than forming sharply separated groups.

At very small ε values (0.05-0.08), DBSCAN only grouped palettes that were almost identical, while marking many paintings as noise. This suggests that only a small set of paintings are extremely similar to enough others to form tight clusters. At medium ε values (0.10-0.12), DBSCAN started to create a few clusters, but many paintings were still left out, and small changes in ε often changed which paintings belonged to which cluster. At larger ε values (0.15-0.20), the clusters quickly merged together, leaving just one or two large groups and making the results harder to interpret. Overall, DBSCAN failed to produce several clear, well-sized clusters without leaving many paintings unassigned. This sensitivity to the similarity threshold is also summarized in Figure 5, which shows how small changes in ε dramatically affect both cluster count and the number of noise points. This suggests that palette similarities in the dataset overlap heavily instead of forming distinct groups. Because DBSCAN relies on one fixed similarity cutoff, it is not the most suitable to capture the gradual shifts in Bob Ross's color choices. This shows that palette structure is more continuous. As a result, hierarchical clustering is more appropriate because it can examine patterns across multiple similarity scales without relying on a single cutoff.

4. Findings & Interpretations

Overall, the analysis reveals clear structure in how paint colors are combined across Bob Ross paintings, but this structure does not reflect sharply distinct palette styles. Instead, most paintings rely on a highly consistent core set of colors, with variation introduced gradually through the selective addition or omission of a small number of paints.

Hierarchical clustering helps clarify the nature of this structure. Beyond the primary cluster, smaller, more distinct groups were also identified. When compared to the overall color usage, these sub-clusters show deliberate deviations from the core palette. These deviations are visible

in the paint usage heatmap (Figure 3), which highlights differential use of specific accent colors across clusters. For example, one subset prominently shows usage of "Prussian Blue" and "Liquid Black," while markedly avoiding warmer tones like "Cadmium Yellow" and "Yellow Ochre." These variations reflect intentional, but subtle adjustments in color selection, and they are most likely respondent to specific compositional needs or thematic elements. Importantly, the boundaries between these groups are not sharply defined.

In terms of DBSCAN analysis, when requiring a high degree of similarity, the model classified most paintings as noise, indicating that identical palettes are rare. On the other hand, with a larger epsilon, nearly all paintings merged into one or two broad clusters. This sensitivity reinforces that Bob Ross's palette choices fall along a smooth range rather than into clear, separate styles. In practice, this means his color decisions changed gradually within a familiar setup, instead of shifting suddenly or sticking to strict phases.

In conclusion, my findings characterize Bob Ross as an artist who mastered consistency without missing nuances. The consistent core palette gave Bob Ross a recognizable look that viewers could trust and easily identify. At the same time, the small, gradual changes in color show a thoughtful flexibility within his established style. This balance between consistency and variation is useful beyond art. For any creator or brand, it highlights the importance of having a clear, recognizable identity while still leaving room to adapt. Bob Ross's approach allowed his work to feel familiar but never repetitive, helping maintain interest and strengthen his lasting reputation as a unique artistic voice.

5. Next Steps

Our analysis has not examined how palette variation relates to the content or themes of each painting. While clustering reveals subtle palette variants, the factors driving these differences remain unexplored.

My next planned step is to incorporate text-based information, such as episode titles or descriptive keywords, to assess whether certain palette variants align with specific scene types, such as mountains, water, winter settings. We could look into keyword frequency or theme grouping, examining whether color combinations are intentionally adapted to subject matter.

Another way to strengthen the analysis would be to see what happens when commonly used paints are removed, which might make differences driven by accent colors easier to spot. It is still unclear whether these small palette changes consistently line up with particular visual themes or teaching intentions.These next steps are justified by current findings, which show that palette usage follows a shared core with gradual variation. Text-based analysis and targeted robustness checks would build on this foundation by helping explain why certain color combinations are used, rather than only identifying how they are structured.

Appendix

Shared GitHub Repository (Required)
Team Repository: https://github.com/yujiunzou/BA820-Unsupervised-ML-Project/tree/main
My individual M2 analysis is presented in the **Kean-Zhu** branch of the project repository. The relevant files are:

- m2_initial_analysis/KZ_M2_Bob_Ross_Paintings.ipynb
- m2_initial_analysis/KZ_Project M2- BA820 - 2026

Supplemental Material

Figure 1: Hierarchical Clustering Dendrogram

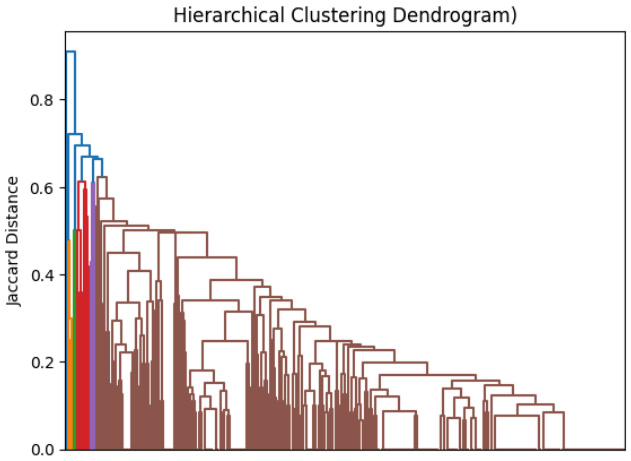


Figure 2: Cluster Size (k=4)

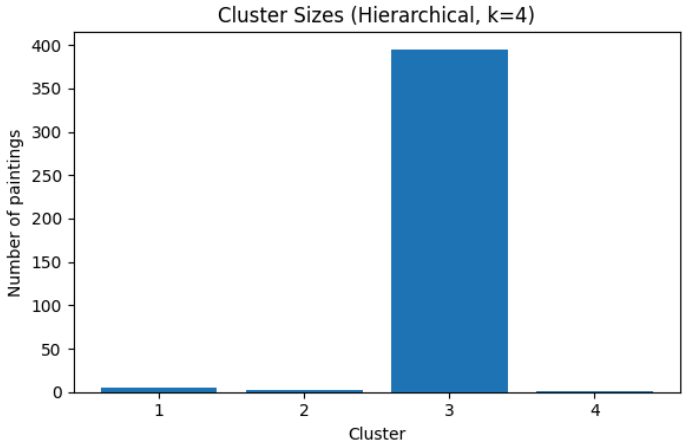


Figure 3: Paint Usage by Cluster Heatmap

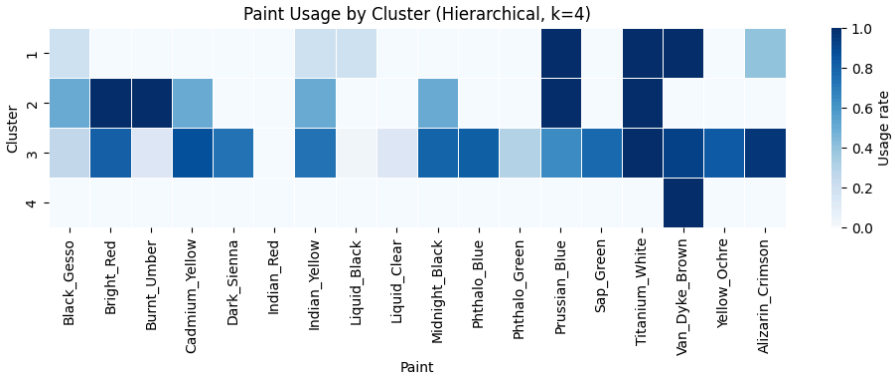


Figure 4: Similarity Comparison: Within-Cluster vs Overall

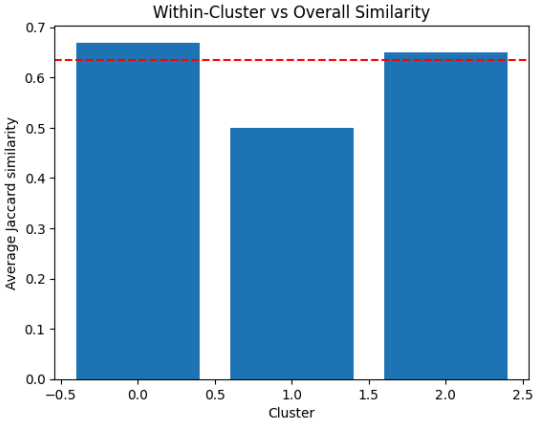
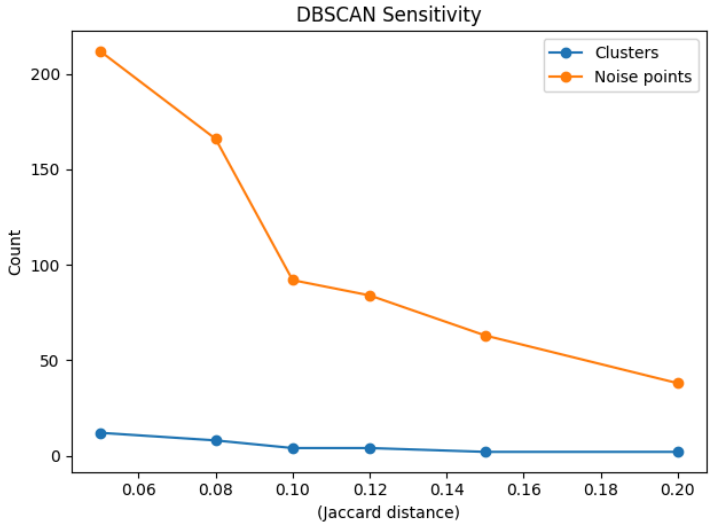


Figure 5 : DBSCAN Sensitivity: Cluster vs Noise



Process Overview

This milestone followed an iterative analysis workflow guided by insights from earlier exploration. Initial EDA was revisited , and domain focus was refined toward understanding global structure in how paint colors are combined.

Based on this reframed question, hierarchical clustering was selected as the primary method to explore potential palette structures without assuming distinct styles in advance. Clustering results were analyzed to assess group coherence and interpretability. A second method, DBSCAN, was then applied to test whether palette combinations form dense, clearly separated groups.

Findings from both approaches were synthesized and interpreted in a real-world context, leading to the conclusion that Bob Ross's palette usage reflects a shared core with gradual variation rather than distinct palette styles. These insights informed the design of next steps, including integrating text-based scene information to explain observed palette variation.

Use of Generative AI Tools

Generative AI tools were used to support conceptual understanding and coding beyond what was covered in class. Specifically, AI was consulted to discuss appropriate linkage methods for binary data, interpret within-cluster versus overall similarity, and explore clustering approaches suitable for non-Euclidean settings. AI assistance was also used to understand and implement DBSCAN, including its parameters and interpretation, which was not directly taught in class. All analytical decisions, results, and interpretations were independently evaluated and finalized by myself.

Link to chat: https://chatgpt.com/share/69893a3a-3410-8011-904c-60fc15a72fdf