BA820 – Project M2

Cover Page

- **Project Title:** Exploring Latent Color Patterns and Similarity in Bob Ross Paintings
- **Section and Team Number:** B1 Team 3
- **Student Name:** Tzu-Jen(Stephanie)Chen

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

The domain question I personally investigated in this milestone builds on the first question from M1: **What are the potential structures in Bob Ross paintings based on how paint colors are combined, and do these structures reveal distinct palette styles?**

However, I refined the framing of this question to make it more specific and better aligned with the findings from the EDA conducted in M1: **Although most Bob Ross paintings use a very similar set of core colors, do these patterns correspond to a single dominant palette template, or to multiple distinct but closely related palette structures?**

The analysis validated the assumption that a shared core palette exists across most paintings, while also revealing that meaningful variation arises through selective use of accent and contrast paints. The updated framing better reflects this structure by emphasizing dominant patterns and controlled deviations rather than searching for entirely separate styles.

2. EDA & Preprocessing: Updates (~0.75 page)

The EDA conducted in M1 revealed that color usage in Bob Ross's paintings is highly structured and repetitive. Paint usage across the 18 indicators is strongly imbalanced, with a small core set of colors, especially Titanium White (0.9926), Alizarin Crimson (0.9429), and Van Dyke Brown (0.9206), appearing in nearly all paintings (Figure 1). These core colors also exhibit frequent co-occurence, suggesting the presence of a shared base palette underlying most works (Figure 2).

In addition, pairwise Jaccard Similarity analysis showed high overlap in palette usage across paintings(Figure 3). With a median above 0.64 and a large number of near-duplicate palette combinations. Several exact paint combinations appeared repeatedly across dozens of paintings. Temporal exploration further indicated little change in average paints usage across seasons, suggesting that palette structure is largely stable over time(Figure 4).

Together, these findings suggest that palette usage in Bob Ross's paintings is highly consistent and non-random. However, they do not determine whether the observed repetition reflects a single dominant palette template or multiple recurring but closely related palette structures. The analyses in M1 mainly focused on marginal frequencies, co-occurrence patterns, and pairwise similarity, which describe overlap but do not reveal whether the ways colors are combined across paintings follow one dominant pattern or several distinct patterns. This limitation motivates the use of clustering methods in M2 to examine the overall structure of palette usage.

No additional EDA or processing steps were introduced in M2 but I removed the part like correlation that is irrelevant to solve this domain away to make the notebook more concise. Since the existing EDA already established strong imbalance, repetition, and stability in color usage.

3. Analysis & Experiments (~1.5 page)

The palette data consist of binary variables indicating the presence (1) or absence (0) of individual paints in each painting. As a result, Euclidean-distance-based methods such as k-means are less appropriate, since they rely on continuous features and arithmetic means. Instead, I use the Jaccard distance, which measures similarity based on shared paint usage while ignoring joint absences—an important property given that most paints are absent in any single painting. Hierarchical clustering with Jaccard distance is therefore well suited for capturing similarity in sparse, categorical palette data. This approach is expected to reveal whether Bob Ross's paintings follow a single canonical palette or multiple closely related palette templates.

To apply hierarchical clustering, the pairwise Jaccard distance was first computed as a square distance matrix and then converted into a condensed distance vector to meet the requirements of the hierarchical linkage algorithm. Average linkage was selected, as it provides a balanced compromise between single and complete linkage when working with similarity-based distances. A dendrogram was then constructed to visualize the nested structure of palette similarity and to explore potential cluster separations at different levels.

To guide the choice of the number of clusters, silhouette scores were computed for values of k ranging from 2 to 8. The silhouette score was maximized at $k = 2$ and declined gradually as k increased. Combining insights from both the silhouette analysis and the dendrogram, multiple cluster solutions were examined. Ultimately, $k = 5$ was selected to balance cluster separation with interpretability.

Finally, cluster-level paint usage rates were computed by averaging the binary palette vectors within each cluster. Paints were categorized as core (usage rate ≥ 0.8), optional ($0.3 \leq$ usage rate < 0.8), or signature (usage substantially higher than the overall average), allowing each cluster to be interpreted as a distinct palette template.

Hierarchical clustering proved effective in revealing a dominant canonical palette shared by the majority of Bob Ross paintings. Cluster-level paint usage profiles were highly interpretable and aligned with known characteristics of his style, such as the consistent use of Titanium White and Prussian Blue. In addition, the method successfully surfaced rare but meaningful palette variations, including warm-accent palettes, high-contrast dark palettes, and highly constrained palettes that represent stylistic edge cases.

However, the analysis also revealed important limitations. The silhouette score strongly favored k = 2, resulting in a highly imbalanced split in which 402 out of 403 paintings were assigned to a single cluster. While this solution maximized separation, it was too coarse to capture finer stylistic nuances. As k increased, the clusters remained highly imbalanced, with some containing very few paintings, including single-painting clusters. Consequently, traditional clustering metrics alone were insufficient for selecting the number of clusters, and interpretability played a critical role in identifying a more informative clustering solution.

One surprising result was that, regardless of the clustering configuration, a single cluster consistently dominated, containing over 95% of the paintings. This highlights how standardized Bob Ross's palette choices were across his work. Despite their small size, the remaining clusters exhibit internally consistent paint usage patterns, indicating that they capture meaningful stylistic variations rather than noise or outliers.

4. Findings & Interpretations (~0.75 page)

Analysis of paint usage across Bob Ross paintings reveals a clear balance between consistency and variation. The heatmap of palette templates shows that the vast majority of paintings rely on a shared core set of paints, most notably Titanium White and Prussian Blue, which appear consistently across nearly all clusters. This common foundation explains why Bob Ross's work feels immediately recognizable, even as individual paintings differ in mood, lighting, and composition(Figure 5).

Beyond this shared core, the clusters differ in how accent and contrast paints are used, revealing several distinct but related palette templates.

- **Cluster 1(n = 5) – Cool Palette**

  This cluster exhibits a highly constrained palette, relying primarily on Titanium White, Prussian Blue, and Van Dyke Brown, with very limited use of other paints. This palette suggest simpler or more instructional compositions, such as calm coastal or minimal landscapes.

- **Cluster 2(n = 2) – Warm Accent**

  This cluster is characterized by high usage of warm and earthy paints, especially Bright Red and Burnt Umber, with Cadmium Yellow appearing as a secondary accent. These palettes are commonly associated with sunset or autumn scenes and demonstrate how warmth is introduced without abandoning the core palette.

- **Cluster 3(n = 10) – Dark Accent with High-Contrast**

This cluster emphasizes darker tones, particularly Midnight Black. Moderate usage of Sap Green and Alizarin Crimson suggests compositions with strong contrast and visual depth, such as forest scenes or darker backgrounds.

- **Cluster 4(n = 385) – Canonical**

  This cluster contains the vast majority of paintings and represents the canonical Bob Ross palette. It shows a broad and balanced mix of cool tones, warm accents, and earth tones—such as Phthalo Blue, Sap Green, Cadmium Yellow, Bright Red, Yellow Ochre, and Van Dyke Brown—making it versatile enough to support a wide range of classic landscapes, including mountains, lakes, trees, and skies.

- **Cluster 5(n= 1) – Van Dyke Brown**

  This cluster consists of a single painting and represents an extreme, highly constrained palette dominated by Van Dyke Brown, with minimal use of other paints. Rather than a recurring template, this cluster reflects a rare stylistic edge case.

  Overall, these findings show that Bob Ross's paintings are structured around a dominant, standardized palette, with variation emerging through selective emphasis on warm tones, dark contrasts, or palette simplicity. This helps art critics understand recurring themes, gives aspiring artists practical guidance on using a limited set of paints effectively, deepens fans' appreciation of his creative choices, and provides AI and machine learning researchers with a structured way to model artistic style while preserving consistency and variation.

5. Next Steps (~0.25 page)

This analysis focuses only on binary paint usage and does not consider additional context such as painting themes, episode order, or stylistic evolution over time. Further work could explore more clustering methods such as k-modes or k-prototypes to assess the robustness of the identified palette templates. Given the strong dominance of a canonical palette observed here, future analyses may also examine whether stylistic variation is driven more by composition and technique than by color choice.

Appendix

Shared GitHub Repository (Required)

Link to Github's repository: https://github.com/yujiunzou/BA820-Unsupervised-ML-Project.git
Branch: Tzu-Jen(Stephanie)Chen → Folder: M2_Tzu-Jen Chen → File: Bob_Ross_Paintings_M2_Tzu_Jen(Stephanie)Chen.ipynb

Supplemental Material (Highly Recommended)

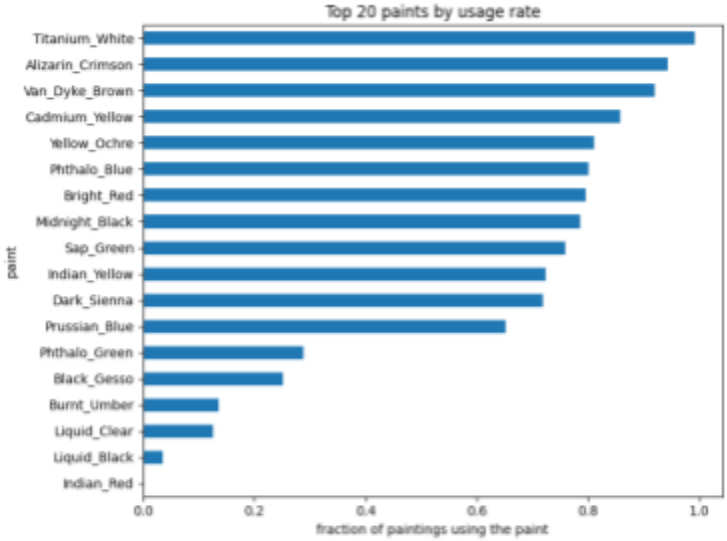

Figure 1. Top paints usage rate

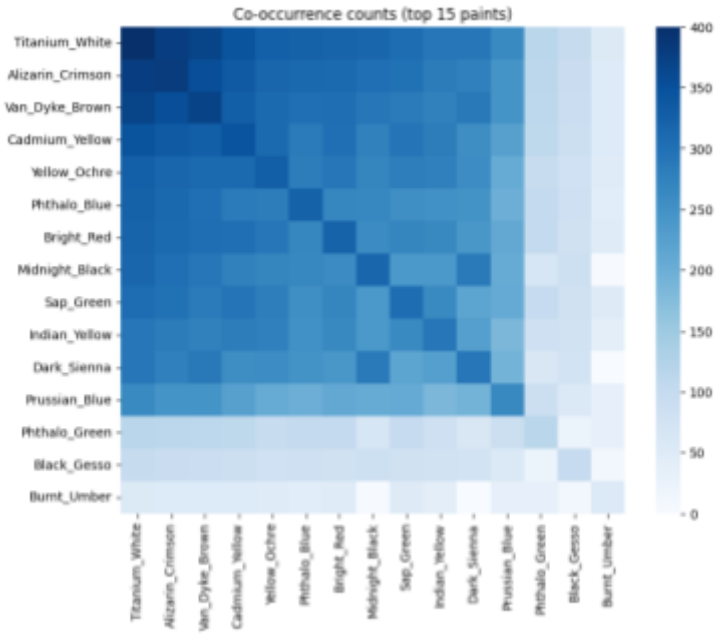

Figure 2. Paint co-occurrence heatmap

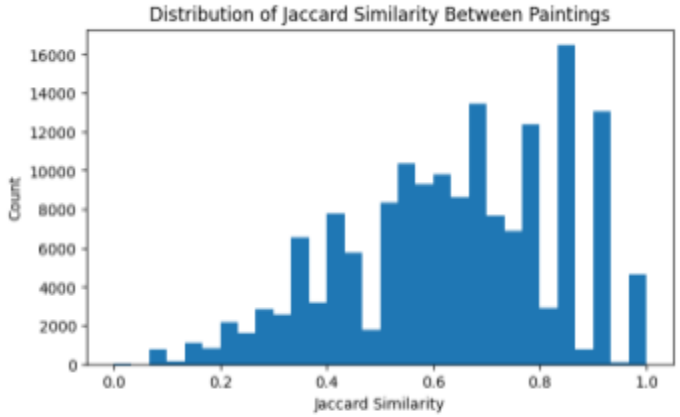

Figure 3. Distribution of Jaccard Similarity between paintings

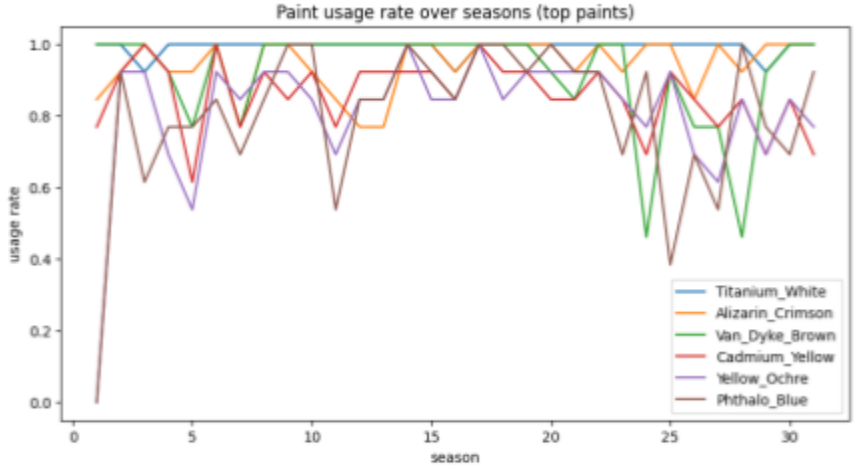

Figure 4. Temporal analysis paint usage rate over season

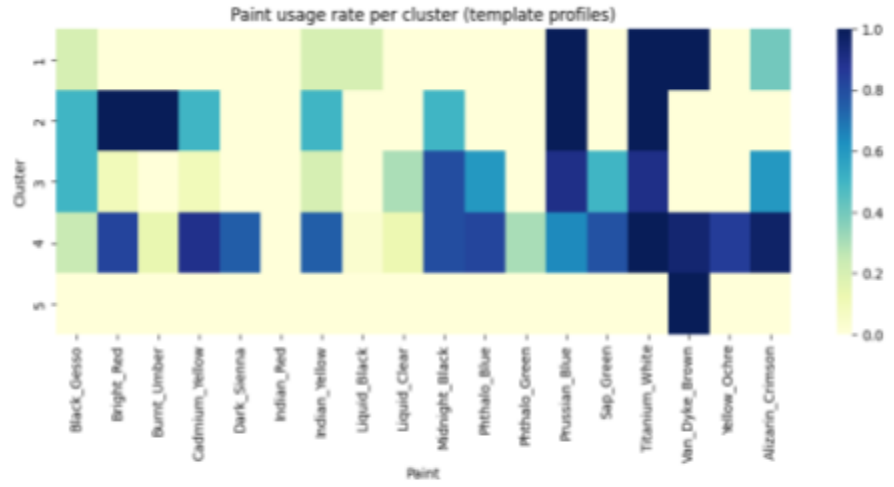

Figure 5. Template profiles of 5 clusters

Use of Generative AI Tools

For this project, I used ChatGPT as a support tool during the exploratory and planning stages. I shared my M1 notebook to ask whether the existing EDA required changes and to seek guidance on potential directions for refining the research question. I also consulted ChatGPT for suggestions on appropriate clustering methods and distance metrics for binary paint usage data, including how to compute Jaccard distance and apply it within hierarchical clustering. In addition, ChatGPT was used to help revise and clarify written explanations after I drafted my initial ideas, improving clarity and organization without generating analysis or results.

LINK