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:** Tzu-Jen(Stephanie) Chen

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

During the proposal stage of our project, we came up with four domain questions to explore:

1. What are the potential structures in Bob Ross paintings based on how paint colors are combined, and do these structures reveal distinct palette styles?
2. Do paintings that are highly similar in color composition form coherent groups over time, or are similar paintings evenly distributed across seasons?
3. Which paint colors contribute most to defining similarity between paintings, and which colors primarily introduce differentiation?
4. Are there structurally unusual paintings in terms of color usage, and how do these paintings differ from the majority of Bob Ross's work?

During M2 and our team integration in M3, my teammate and I attempted to address these questions using hierarchical clustering, threshold exploration to validate the clustering structure, and association rule analysis. However, after implementing these approaches, we observed the emergence of a dominant mega-cluster (385 out of 403 paintings), along with several very small clusters containing only one to five paintings.

For M4, my primary focus will be to explore alternative clustering methods (K-modes) that may produce more evenly distributed groups. In addition, I plan to incorporate text analysis and combine palette patterns with painting titles to better understand their relationship, and analyze this new question:

> **Whether palette structures show alignment with thematic content, and how palette usage differs across themes.**

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

For the EDA stage, I mainly revised Point 9, which focuses on text analysis. The updates include:

- Adding additional stopwords (e.g., *to, for, with, from, into, over, under*) to improve the word-cleaning process.
- Redefining the keyword groups used for thematic categorization.

To refine the keyword groups, I selected candidate words from the top 60 most frequently occurring terms identified through value counts on the cleaned title tokens (*words_clean*). Based on these high-frequency words, I created several theme categories, including **Seasonal, Mountain/Landscape, Woodland, Water, Sky/Light/Time,** and **Human Elements**.

Each theme is mapped to a set of relevant keywords extracted from painting titles, allowing the paintings to be categorized into meaningful thematic groups for further analysis.

3. Analysis & Experiments (~2 pages)

K-modes

To address the recurring mega-cluster issue observed in earlier stages, I explored the k-modes clustering method. Because our dataset is primarily categorical (binary indicators of whether a paint is used in a painting), k-modes is an appropriate alternative to hierarchical clustering. I applied k-modes to examine whether it could produce more evenly distributed clusters compared to the previous approach.

I set the number of clusters to five to maintain consistency with earlier analysis. After multiple initializations, the algorithm achieved the lowest cost on the seventh run and separated the paintings into five clusters with sizes of 60, 98, 71, 151, and 23 respectively.

To interpret the clustering outcome, I examined the cluster centroids, which represent the most common paint combinations within each group. The k-modes results indicate that paint usage in Bob Ross paintings is highly structured around a stable core palette. Rather than forming many sharply separated groups, most clusters vary gradually around a dominant canonical set of paints. This finding aligns with the earlier hierarchical clustering results, which also suggested the presence of a strong baseline palette rather than clearly isolated styles.

Clusters 3 and 4 appear closest to the canonical palette. Both clusters show consistently high usage across most core paints, suggesting that many paintings follow a similar foundational color structure. The differences between these clusters seem to arise from smaller variations in secondary paints and technical materials, rather than a complete shift in palette composition.

In contrast, Clusters 0, 1, and 2 highlight meaningful variations around the core structure. Cluster 0 reflects a more reduced and cool-dominant palette, characterized by stronger emphasis on blues and white and lower usage of warm accent colors. Cluster 1 maintains a relatively balanced palette but shows noticeably lower usage of *Prussian_Blue*, a color that appears prominently in most other clusters. Cluster 2 emphasizes green tones and darker materials, which may be associated with foliage-heavy scenes.

While hierarchical clustering revealed a dominant canonical palette, k-modes produced more evenly distributed clusters that highlight interpretable palette archetypes. Because these archetypes are easier to compare at the group level, the k-modes clusters are used as a practical reference for the subsequent text analysis.

Text Analysis

One limitation our team identified in M3 is that clustering based only on color usage allows us to see which paints are present, but does not capture how they are applied or what subjects are being painted. While this approach helps reveal structural regularity in palette construction, it does not fully reflect other important aspects of artistic style.

To address this limitation, I focus on deeper text analysis by integrating painting titles with palette structure. This includes grouping title keywords into thematic categories and examining whether certain palette variants align with specific scene types.

For the text analysis, I extract words from painting titles and remove conjunctions or non-thematic words (such as *the, of, from, into,* etc.) to retain only meaningful descriptors. Based on frequently appearing words, I organize them into thematic keyword groups:

> **Seasonal, Mountain/Landscape, Woodland, Water, Sky/Light/Time,** and **Human Elements**.

Using these keyword groups, I analyze the k-modes clusters to explore whether color usage patterns show any relationship with thematic content.

Based on the heatmap results, each k-modes palette cluster demonstrates different thematic tendencies rather than a single dominant theme across all clusters. Cluster 0 shows a strong concentration in Seasonal (0.51), suggesting that this palette is more closely associated with seasonal scenes compared to other themes. Cluster 1 has the highest proportion in Sky / Light / Time (0.31), indicating that this palette may relate more to atmospheric or lighting-focused compositions. In contrast, Clusters 2 and 4 emphasize Water themes (0.33 and 0.29), implying that certain palette structures may align with water-related scenery.

Another notable pattern appears in Cluster 3, where Mountain / Landscape (0.33) becomes the dominant theme, suggesting a possible connection between landscape compositions and this palette configuration. Overall, instead of one universal theme dominating all clusters, the heatmap reveals subtle but meaningful differences in thematic distribution. These variations support the idea that palette usage may partially reflect scene context, even though the differences remain moderate rather than extreme.

4. Findings & Interpretations (~1 page)

The combined results from palette usage and theme distribution suggest that Bob Ross paintings follow a strong canonical color structure while showing moderate thematic variation. Clusters 3 and 4 appear closest to this canonical palette, with consistently high usage across most core paints. These clusters are not tied to a single theme but instead span landscape, water, and woodland categories. This indicates that the canonical palette serves as a stable foundation that supports many scene types, with only subtle adjustments in secondary paints or technical materials.

The theme heatmap further shows that each cluster leans toward different subjects rather than being dominated by one universal theme. Cluster 0 has a strong Seasonal concentration (0.51), supported by frequent keywords such as *winter* (37), *autumn* (18), and *spring* (3). This pattern suggests that cooler seasonal scenes rely on a more reduced and cool-dominant palette, with heavier emphasis on blues and white. Cluster 1 shows the highest proportion of Sky / Light / Time themes (0.31). Keywords such as *sunset* (13), *day* (10), *glow* (6), and *misty* (4) indicate a stronger focus on atmosphere and lighting rather than structural landscape elements, which helps explain why this cluster maintains a balanced palette while using less *Prussian_Blue* compared to others.

Clusters 2 and 4 emphasize Water themes (0.33 and 0.29), supported by frequent keywords such as *lake* (16), *waterfall* (11), *stream* (11), *river* (6), and *reflections* (6). The palettes in these clusters show stronger green tones and darker materials, suggesting that foliage and reflective environments influence how the core palette is adjusted. Cluster 3 stands out with Mountain / Landscape as the dominant theme (0.33), reinforced by high keyword counts such as *mountain* (45), *valley* (9), *view* (9), and *hills* (4). Even though its palette remains close to the canonical structure, these landscape keywords suggest that large-scale scenery often relies on a stable and predictable color foundation.

Overall, rather than introducing entirely new colors, paintings tend to adjust the balance of a shared core palette to match different scene contexts. These findings are important for the stakeholders we identified in M1 because they provide a structured way to understand Bob Ross's style. The results highlight how a consistent color foundation can support diverse themes, showing that different scenes are often created by shifting the balance of familiar paints rather than adding many new colors. In addition, combining palette structure with thematic information offers useful insights for building interpretable models in areas such as art generation, clustering, and style transfer.

Appendix

Shared GitHub Repository (Required)
Link: https://github.com/yujiunzou/BA820-Unsupervised-ML-Project

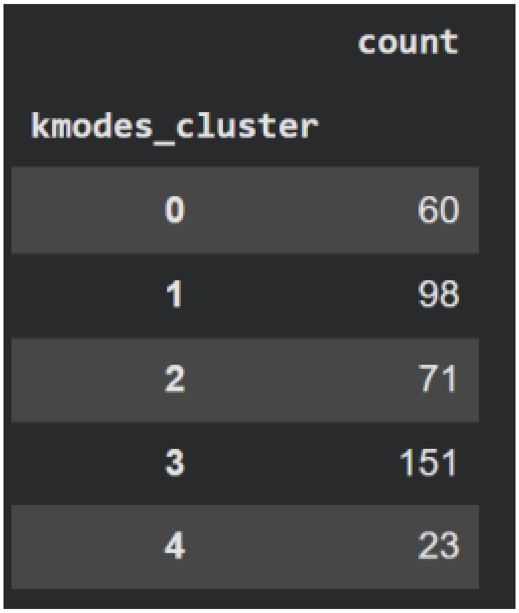
M4 work (where to find the files):

- Project Name: BA820-Unsupervised-ML-Project
- Branch: "*Main*"
- Folder: "*M4_individual_refinement*"
- Folder: "*M4_TzuJen_Chen*"
- Primary notebook: "*M4_Tzu-Jen(Stephanie)Chen_Bob_Ross_Paintings.ipynb*"
- Report PDF: "*M4_Tzu-Jen(Stephanie)Chen_Report.pdf* "

Notebook Link:
https://github.com/yujiunzou/BA820-Unsupervised-ML-Project/blob/c321f04b69ebb1c11b1fa9d5e9490ff171f4331b/m4_individual_refinement/M4_TzuJen_Chen/M4_Tzu_Jen(Stephanie)Chen_Bob_Ross_Paintings.ipynb

Supplemental Material

Method 3



	count
kmodes_cluster	
0	60
1	98
2	71
3	151
4	23

Figure 1. Numbers of paintings in each cluster

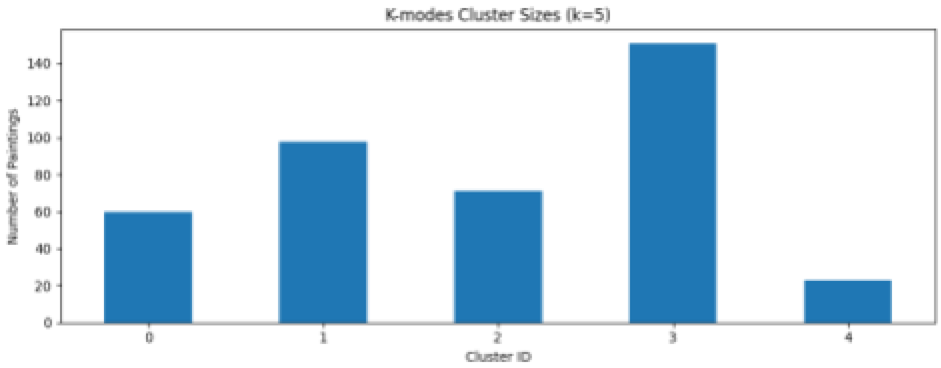


Figure 2. Cluster size bar chart (by K-modes)

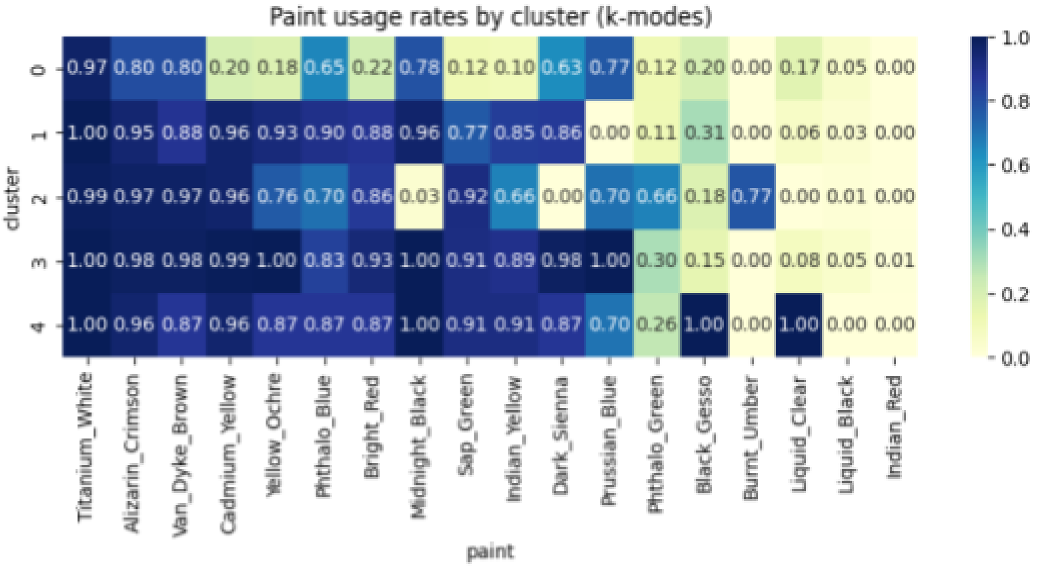


Figure 3. Paint usage rates heatmap (K-modes)

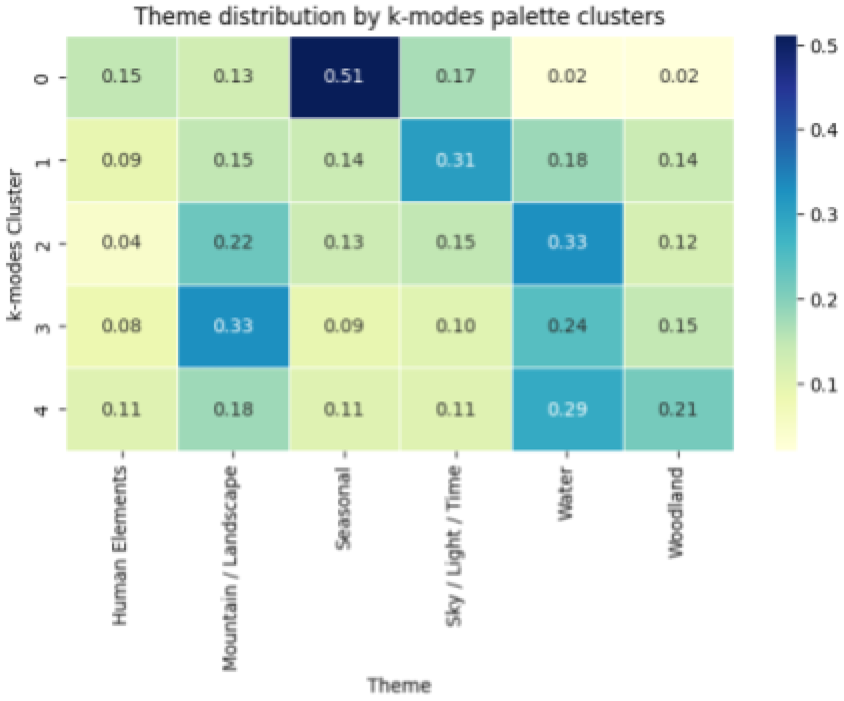


Figure 4. Theme distribution by K-modes palette clusters

Use of Generative AI Tools

Link: https://chatgpt.com/share/699d1d83-08f8-8010-8af3-0a61ee9033fe

I used ChatGPT during the brainstorming stage to explore possible solutions, including how to address the mega-cluster issue, whether k-modes would be an appropriate method, and how to implement k-modes and text analysis step by step. I also used it to help organize painting title keywords into broader thematic categories more efficiently.

In addition, I used ChatGPT as a revision tool. After drafting my own interpretations of charts and analytical results, I checked with ChatGPT to confirm whether my reasoning was logical. I also used it to refine grammar, correct spelling, and improve the clarity and flow of my writing while keeping the original ideas my own.