BA820 – Project M2

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)

In this phase, I conducted an in-depth study on Q2 from the proposal: Do recurring color combinations indicate that Bob Ross employed a fixed color palette template? The core research objective remains unchanged—to verify whether his color usage follows stable, repeatable patterns rather than random selection.

Through data exploration, I broke the question into several measurable dimensions: color usage frequency, number of colors per painting, proportion of rare colors, and color co-occurrence patterns. This shifted the research from broad stylistic description to more data-driven structural examination. Some initial assumptions were supported. Frequency analysis shows that BR relies heavily on a core set of colors, indicating a stable foundational palette. In addition, the number of colors used per painting clusters within a relatively narrow range, suggesting consistent palette complexity.

However, not all expectations were confirmed. I initially assumed that larger palettes would include rarer colors to increase richness. The scatter analysis suggests otherwise — as the total number of colors increases, the proportion of rare colors tends to decline. This implies that palette complexity is more often achieved by recombining common colors rather than introducing new ones. Finally, the co-occurrence matrix provides further evidence. Certain core colors frequently appear together, forming stable pairing relationships. This pattern suggests that color usage is structured rather than random.

In summary, while Q2's core question remains unchanged, this phase's analysis has deepened the perspective from macro-level style description to concrete examination of color combination structures and potential templates.

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

Building upon the findings from M1, the current analysis further investigates whether recurring color combinations reflect structured palette templates. Previous exploratory work revealed several key patterns: a small set of core colors appears in most paintings, palette sizes cluster within relatively narrow ranges, and preliminary co-occurrence observations indicate certain colors frequently appear together. These findings directly prompted this study to pursue a deeper structural exploration of palette composition, rather than treating color choices as isolated decisions.

In M2, the analytical approach was refined through additional visualizations and structural mapping. First, from M1 we saw that a small number of colors were used much more often than others. So in M2 we added a Lorenz curve to clearly show how uneven the color usage is and to check whether most paintings rely on a small core set of colors.

Second, the distribution of palette complexity was further evaluated through density estimates of color counts per painting. A regression line was added to the original scatter plot to better capture the overall trend, showing that palette expansion does not systematically increase reliance on rare colors. Finally, color co-occurrence relationships were visualized through network diagrams to more intuitively show which colors frequently appear together.

Beyond the data transformations already implemented in M1, no major preprocessing steps were required. The dataset remains fully usable, meaning M2's improvements primarily involve deepening the analytical layer.

3. Analysis & Experiments (~1.5 page)

To investigate whether recurring color combinations suggest structured palette templates, this study applies two complementary unsupervised methods: hierarchical clustering and association pattern mining. The former evaluates how paintings group based on overall palette similarity, while the latter identifies specific recurring color combinations and their structural relationships.

Hierarchical Clustering of Palette Structures

This approach helps identify whether paintings naturally group into similar palette structures without requiring a pre-set number of clusters.

Hierarchical clustering was conducted to examine whether paintings could be grouped by palette composition. Each painting was represented as a binary vector indicating the presence or absence of specific colors. Jaccard distance was used to measure similarity, and average linkage was applied to construct the clustering hierarchy.

The dendrogram shows that most paintings merge at relatively low distance levels, indicating highly similar palette structures. Rather than forming multiple distinct branches, the clustering pattern is dominated by a single dense backbone. This suggests that most works rely on closely related palette systems rather than independent palette families.

Cluster assignment results reinforce this interpretation. One dominant cluster contains most paintings, with only a small number appearing in peripheral clusters. This indicates that while palette variation exists, its overall range is limited. The dominant cluster is characterized by frequent use of core landscape colors such as Titanium White, Alizarin Crimson, Sap Green, and Cadmium Yellow, while smaller clusters reflect more specialized tonal compositions.

Before using hierarchical clustering, I also considered other clustering approaches. K-means was tested to group paintings based on palette composition. However, it did not work well for this dataset. Since the palette data is binary, K-means— which relies on averaging and Euclidean distance—was not suitable for capturing similarity in color combinations. The clustering results were unstable and difficult to interpret.

Hierarchical clustering performed better. Using Jaccard distance allowed the analysis to directly measure how many colors paintings share, which better reflects palette structure. So, the clusters were more interpretable and aligned more closely to identify recurring palette patterns.

Association Rule Mining of Palette Structures

To directly examine recurring color combinations, this study applies association rule mining using binary color usage indicators. This method helps identify which colors frequently appear together and whether stable palette templates exist across paintings.

The findings reveal a highly concentrated palette structure. Titanium White appears in nearly all high-support combinations, suggesting it serves as a core base color in many palettes. Several two-color pairings recur frequently, most notably Titanium White combined with Alizarin Crimson or Van Dyke Brown. These pairings further extend into stable three-color groupings, indicating that certain palette structures are repeatedly reused across paintings. Larger color combinations do exist, but their lower support suggests that more complex palettes are typically formed by expanding familiar core combinations rather than introducing entirely new ones. Association rule analysis also reveals clear dependencies among colors. For example, darker tones such as Van Dyke Brown and Midnight Black often appear alongside landscape-supporting colors like Sap Green and Dark Sienna. Lift values greater than one indicate that these combinations occur more frequently than expected by chance, suggesting recurring structural relationships rather than random co-occurrence.

Before finalizing the method, different support thresholds were tested to balance pattern coverage and interpretability. Lower thresholds generated too many trivial or noisy combinations, while higher thresholds focused on more stable and meaningful templates. Adjusting this parameter helped refine the analysis and ensured that identified palette structures were both interpretable and representative.

In summary, association rule mining complements clustering analysis by shifting the focus from overall similarity to specific combinational structures, providing deeper insight into how recurring palette templates are formed.

Integrated Interpretation

Taken together, the two analytical approaches provide converging evidence of structured palette templates. Clustering reveals the dominance of a central palette regime, while association mining uncovers the compositional grammar underlying that regime. Smaller color combinations appear most frequently across paintings, forming stable foundational bundles. Larger palettes appear to emerge through extensions of core combinations, rather than forming entirely independent palette systems.

4. Findings & Interpretations (~0.75 page)

The analyses reveal a clear structural logic underlying Bob Ross’s palette construction. Rather than selecting colors independently, his paintings appear to be built upon a stable system of recurring palette templates. One key insight is the presence of a core palette foundation. A small set of colors—particularly Titanium White, Alizarin Crimson, Van Dyke Brown, and Cadmium Yellow—appears consistently across a large proportion of works. These colors function as structural anchors, forming the visual base upon which additional tones are layered. This pattern suggests that palette consistency is systematic rather than incidental.

A second insight concerns palette complexity. While some paintings employ more colors, the expansion of palette size does not systematically increase reliance on rare hues. This suggests that larger palettes are not necessarily driven by rare color introduction but instead reflect flexible extensions of existing palette structures.

From a practical perspective, these findings have implications for digital art and design applications. Palette recommendation systems, AI art generators, and creative design tools could benefit from incorporating template-based palette logic. By prioritizing historically coherent color combinations rather than random suggestions, such systems may improve visual harmony, stylistic consistency, and user design efficiency.

5. Next Steps (~0.25 page)

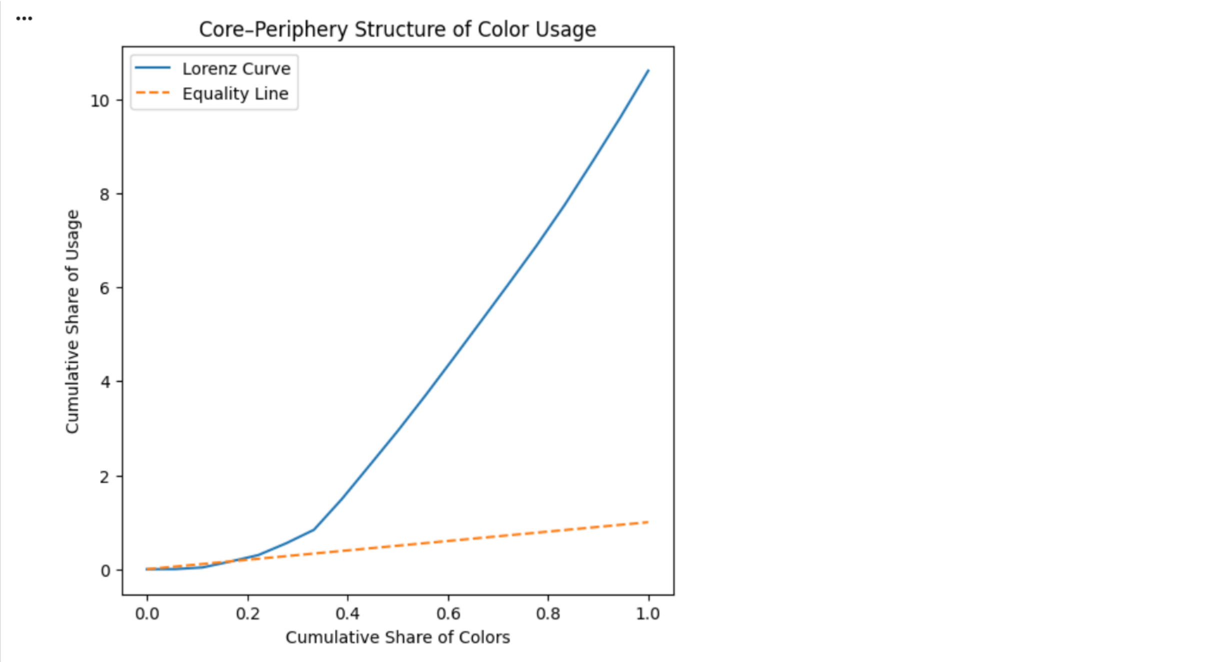
First, the current analysis has not yet examined whether different painting themes correspond to distinct palette templates. Linking palette structures to subject matter could provide deeper contextual insight. Second, while frequent itemsets successfully identify recurring color bundles, more advanced network-based approaches could further examine higher-level structural groupings among palettes. Some variation also remains to be better understood. A small number of paintings display palette compositions that deviate from the dominant template. At this stage, the drivers of these deviations remain unclear. A next step would be to investigate whether these outliers are systematically associated with observable contextual signals—such as inferred themes, environmental settings reflected in titles, or recurring subject elements—to better understand the flexibility of palette templates within the broader structural system.

Appendix

Shared GitHub Repository (Required)

- https://github.com/Charles-Wei77/-ba820-bob-ross-team05
- BA820-ProjectM2-Shengqi Wei.pdf
- 820_Bob Ross Paintings_Team05_Shengqi_Wei.ipynb

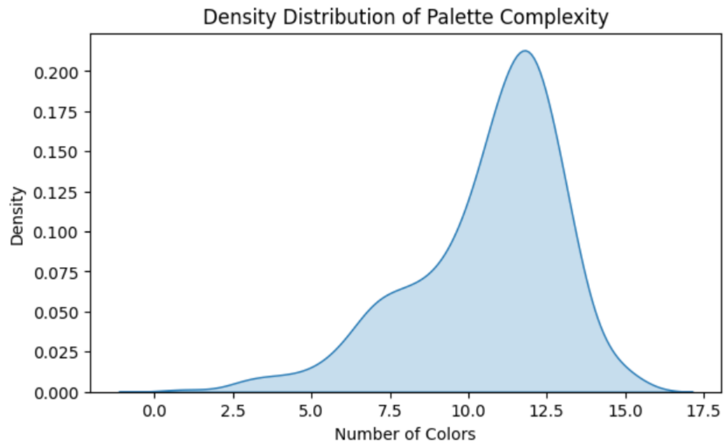
Supplemental Material (Highly Recommended)



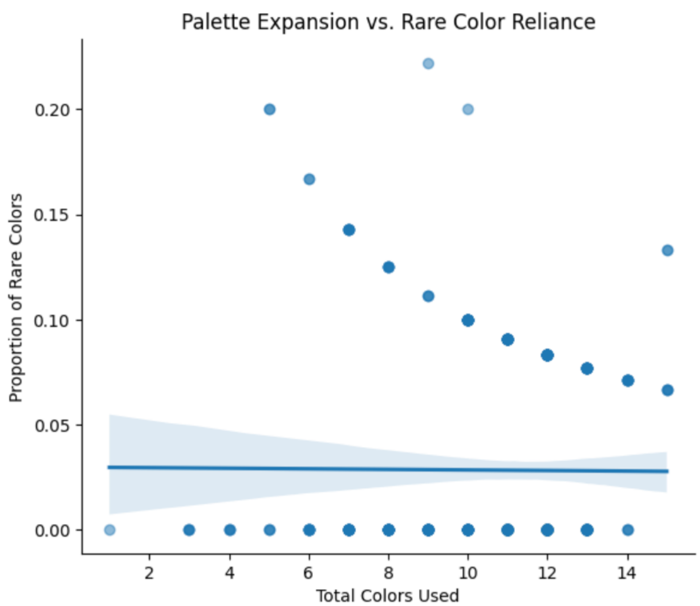
The Lorenz curve indicates that color usage is highly concentrated. A few core colors dominate most paintings, while the remaining colors are used more sparingly. This supports the idea that Bob Ross's palettes are built around stable base colors rather than evenly distributed color choices.

[11]

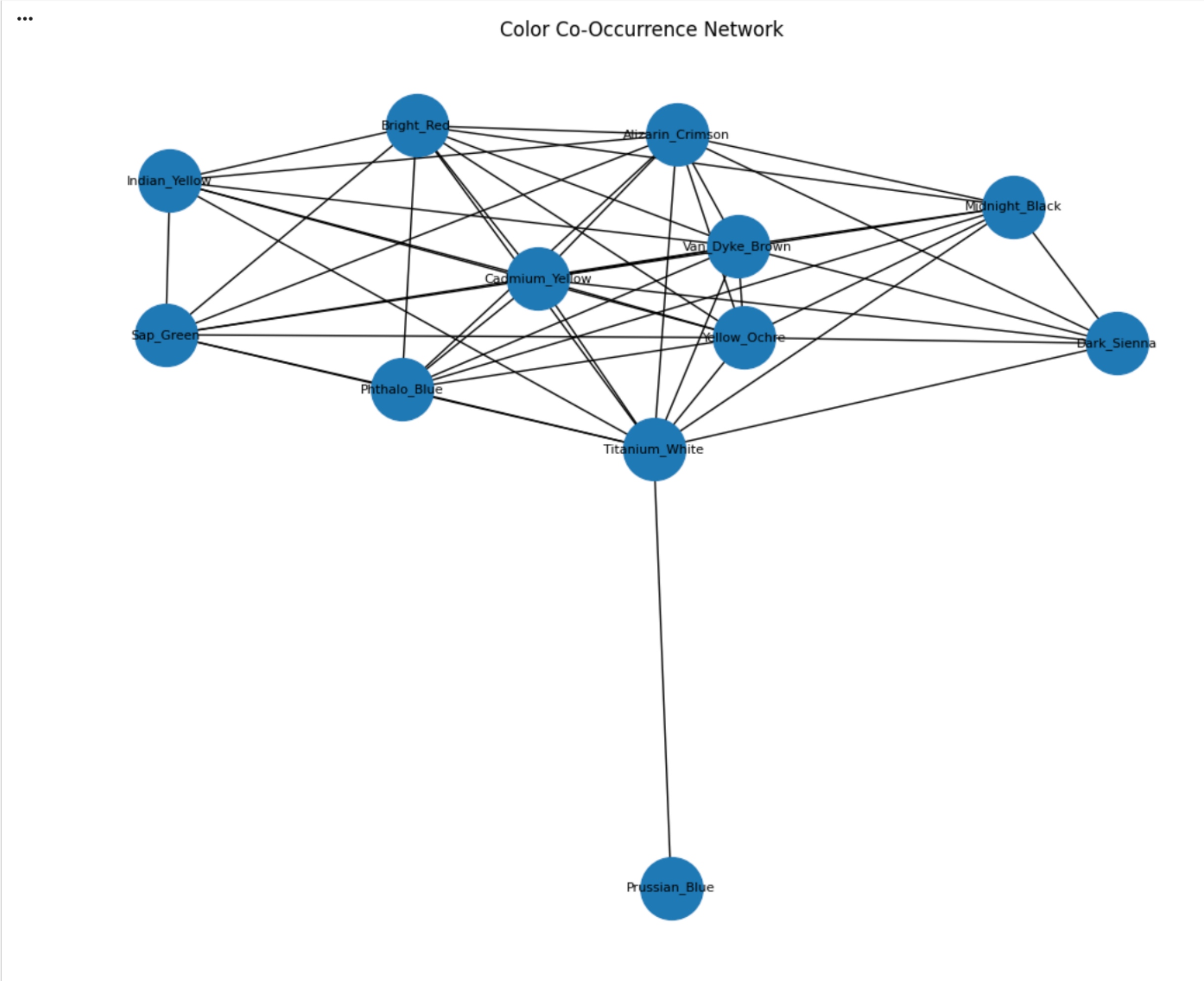
```
plt.figure(figsize=(7,4))
sns.kdeplot(bob_ross["num_colors"], fill=True)
plt.title("Density Distribution of Palette Complexity")
plt.xlabel("Number of Colors")
plt.ylabel("Density")
plt.show()
```



The density curve indicates that most paintings use a similar number of colors, with complexity concentrated around a central range. This suggests that palette size follows a stable working range rather than forming clearly distinct complexity regimes.



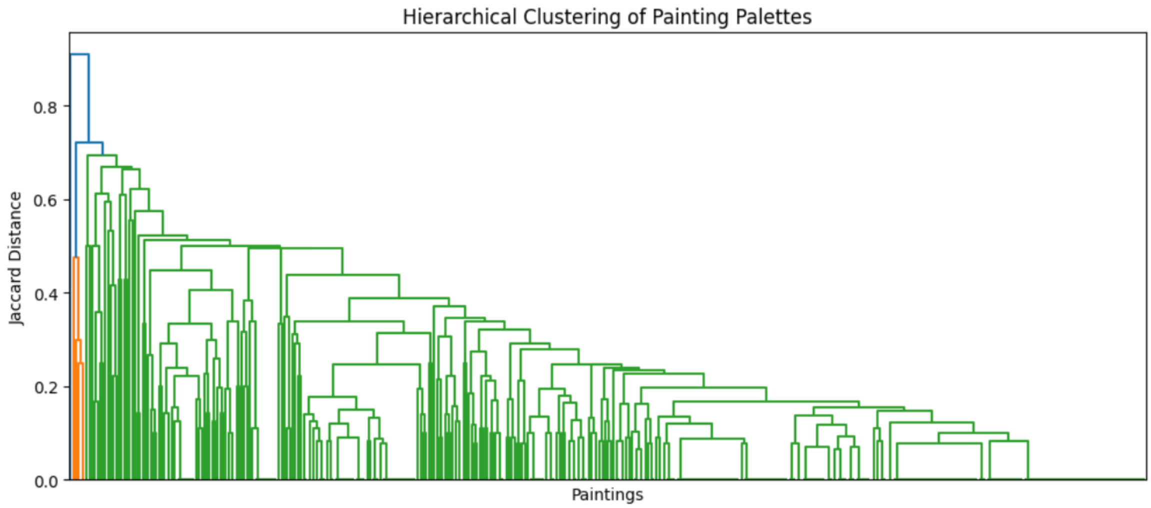
The plot indicates that using more colors does not necessarily mean relying more on rare ones. Paintings with larger palettes still depend primarily on commonly used colors. This suggests that palette expansion is more likely driven by recombining familiar colors rather than introducing many new ones.



The co-occurrence network reveals a highly interconnected core palette, where several foundational colors consistently appear together. Surrounding this core are supporting colors that extend palette variation while maintaining structural cohesion. Peripheral colors show limited integration, suggesting occasional rather than systematic use.

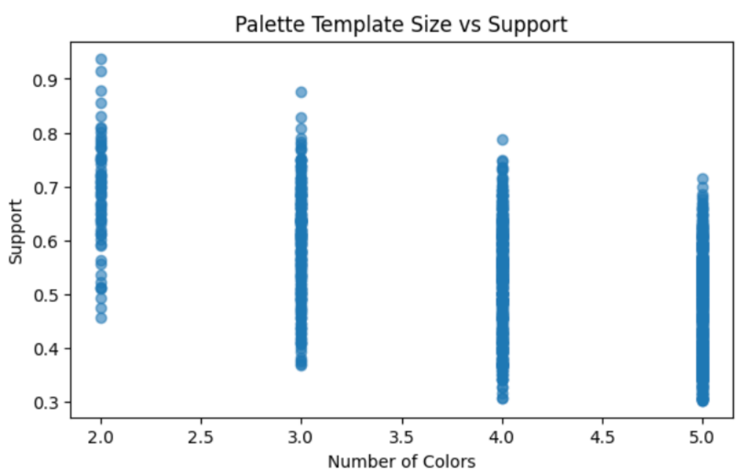
```python
from scipy.cluster.hierarchy import dendrogram
import matplotlib.pyplot as plt

plt.figure(figsize=(12,5))
dendrogram(Z, no_labels=True, color_threshold=0.7)
plt.title("Hierarchical Clustering of Painting Palettes")
plt.xlabel("Paintings")
plt.ylabel("Jaccard Distance")
plt.show()
```



Hierarchical clustering based on Jaccard distance reveals that most paintings merge at relatively low dissimilarity levels, indicating high similarity in palette composition. This suggests that a large proportion of works rely on closely related color structures. However, a small number of branches only merge at higher distance thresholds, implying occasional deviations from dominant palette patterns. Overall, the dendrogram exhibits a single primary backbone with limited branching, supporting the interpretation that Bob Ross's palette system is organized around a dominant template with minor variations rather than multiple fully distinct regimes.

...



The relationship between template size and support shows a clear pattern in how color combinations are reused. Combinations that include only two or three colors appear most frequently across paintings, suggesting that these basic sets form the foundation of many palettes. As more colors are added to a combination, the frequency of that exact combination appearing begins to decrease. This indicates that larger palettes are usually created by adding extra colors onto a stable core set rather than forming entirely new combinations. In other words, palette complexity grows by building on familiar color groups instead of introducing completely different ones.

Process Overview

- Data Collection
- Data Structuring (Binary Color Matrix)
- Preprocessing & Feature Setup
- Exploratory Data Analysis
- Unsupervised Methods
- Insight Integration & Interpretation
- Business / Creative Implications

The analytical process begins with structuring the dataset into a binary color matrix, where each painting is represented by the presence or absence of specific colors. This format enables systematic comparison of palette composition across works.

Next, EDA was then performed to identify foundational patterns, including color frequency, palette complexity, and co-occurrence relationships. These visuals provided early evidence of recurring palette structures.

Building on these insights, two unsupervised methods were applied. Hierarchical clustering was used to identify broader palette regimes, while frequent itemset and association rule analyses uncovered recurring color bundles and their dependencies.

Finally, insights from both approaches were integrated to interpret palette construction logic and assess the presence of structured palette templates, along with their potential creative and commercial relevance.

Use of Generative AI Tools

I asked ChatGPT on how to conduct an in-depth EDA for Q2, particularly regarding the use of Lorenz curves to determine whether color usage is dominated by a few core colors.

At first, I wasn't entirely sure which clustering method best suited Q2, especially between k-means and hierarchical clustering. With AI's guidance, I confirmed my suspicion that k-means was unsuitable.

All analytical decisions, interpretations, and conclusions were completed independently based on course materials and my own understanding of the data. AI tools were not used to generate results or replace the core analytical work.

https://chatgpt.com/share/698aa4da-3d1c-8003-8b9f-196ef9d69df6