BA820 – Project M3

Cover Page

- **Project Title:** Exploring Latent Color Patterns and Similarity in Bob Ross Paintings
- **Section and Team Number:** B1 Team 3
- **Team Members:** Kean Zhu, Yihui(Irene) Tang, Yu-Jiun(Janice) Zou, Tzu-Jen(Stephanie) Chen.

1. Integrated Problem Framing and Updated Questions

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?

After re-evaluating our EDA during M2, our team agreed that Q1 best captures the main goal of the project, even though we worked independently. We felt it provides meaningful interpretability for art historians, artists, and researchers interested in style patterns.

Although we used different unsupervised methods, our results were very consistent. We all shared the insight that palette usage is strongly centered around a dominant core set of paints. In M3, our goal is to explore whether Bob Ross paintings show meaningful structure based on paint usage. Because the data is binary, we focused on similarity-based clustering and association rules, and looked at how the methods support or challenge each other instead of evaluating them separately.

2. Recap of Individual M2 Contributions

Kean used hierarchical clustering and DBSCAN and found that most paintings fall into one big core group, with only a few small variations. DBSCAN was very sensitive to the settings, which suggests that palette changes happen slowly and do not form clear-cut groups.

Irene used agglomerative clustering with Jaccard distance and saw a similar result at k=5: one large main cluster (about 95%) and a few very small ones. She also used association rules to understand how paints appear together, but the smallest groups didn't have enough data, so those patterns are more descriptive than strongly proven.

Stephanie also applied hierarchical clustering with Jaccard distance and confirmed that there is a stable core palette. The smaller clusters mainly show changes in accent or contrast colors, not totally different styles. Since silhouette scores kept pushing the solution toward k=2 and made the clusters uneven, we relied more on interpretation instead of only following the metrics.

Janice applied association rules and clustering to see whether the structure comes from specific paint pairs or larger color combinations. The results showed that the pattern comes from groups

of colors working together, not just one dominant paint like Titanium White. A small text-based EDA also suggested that paintings are more similar in color use than in their titles.

3. Integration Strategy and Synergy Effort

During the integration process, we aimed to consolidate overlapping methods. First, we retained Jaccard distance as the primary similarity metric because the paint indicators are sparse; Jaccard distinguishes paintings based on shared paints (1s), whereas Hamming is pulled toward similarity by shared absences (0s), producing a more compressed distance distribution. Second, we maintained hierarchical clustering as the central structural model because all members independently observed similar dominance patterns using this method. After discussion, we kept Irene's code as the base framework since it is the most structured and aligned with our interpretation using Jaccard distance. We made small adjustments for consistency, but the overall structure stayed the same.

We also kept association rule mining as a way to complementarily interpret the clusters. While hierarchical clustering showed one dominant cluster and a few small variants, it did not explain why those variants existed. Association rules allowed us to examine the internal paint co-usage patterns both globally and within clusters.

We also made several key modifications to improve coherence with feedback received. Instead of fixing the number of clusters using an arbitrary k value, we shifted to exploring distance thresholds in hierarchical clustering. We also then analyzed cluster size distributions across thresholds to evaluate imbalance. This allowed us to observe and assess if natural separation exists in the data. In addition, rather than applying association rule mining only once at the global level, we used it after clustering to interpret specific cluster identities, which allowed us to characterize how each cluster differs in terms of coordinated paint usage patterns. As shown in Figure 6, smaller clusters generated substantially more qualifying rules under the same thresholds than the canonical cluster, reinforcing that deviations are structurally concentrated within these smaller groups rather than uniformly distributed.

Some methods were removed during integration. We decided not to keep DBSCAN and K-means. DBSCAN was very sensitive to parameter choices and often labeled most paintings as noise, and K-means did not fit our data well because its Euclidean distance assumption is incompatible with binary data. To keep the analysis consistent, we focused only on Jaccard-based hierarchical clustering. We also relied on the dendrogram instead of the Silhouette score that was explored, as it mainly captured the separation of the dominant big cluster from the rest, which did not align with our goal of understanding finer structural variation.

4. Integrated Analysis and Results

Our integrated analysis combined hierarchical clustering using Jaccard distance with association rule mining to understand both the macro and micro view of the structure of paint usage patterns

in Bob Ross paintings. Previously, all members observed a dominant cluster through hierarchical clustering. In M3, we evaluated this pattern more in-depth by selecting an appropriate dendrogram threshold rather than pre-specifying k.

Because the dataset consists of binary paint indicators, Jaccard distance was chosen as it measures similarity based on shared paint usage. Figure 1 shows that Jaccard distances display a broader distribution compared to Hamming distance. Average linkage hierarchical clustering allowed us to examine how clusters form across multiple thresholds (0.20, 0.40, 0.60, and 0.85). At lower thresholds, many small clusters appeared, but as the threshold increased, these clusters gradually merged into one dominant group that contains most paintings. Importantly, one large cluster remained consistently dominant across all thresholds, and the imbalance ratio consistently remained high. Such consistency indicates that the dominant cluster is a structural property of the data, not just a result of parameter choice, supporting the idea of a shared baseline palette. Under the retained five-cluster solution (Figure 3), Cluster 4 contains 385 of 403 paintings (≈95.5%), while the remaining clusters contain few. This extreme concentration confirms that the dominant cluster overwhelmingly represents the dataset rather than being only marginally larger.

To interpret the cluster structures, we performed association rule mining at both global and within-cluster levels. Global rules showed that certain paint combinations frequently co-occur, reinforcing the presence of a baseline palette. Within smaller clusters, higher lift values revealed coordinated paint combinations that differ from the global baseline, suggesting that these clusters represent structured deviations rather than random noise. Applying stricter lift thresholds helped filter out trivial high-frequency pairs and highlighted more meaningful multi-paint relationships. Additionally, conducting robustness checks such as removing the most used color Titanium White, showed that the global similarity structure remained stable, indicating that clustering results were not artificially driven by a single dominant color.

Overall, we have the association rule mining complementing hierarchical clustering by providing interpretability at the feature level in this milestone. Clustering identified a big structure of baseline compared to deviations, while rule mining explained the micro-level paint combinations that define those deviations.

5. Insights Gained Through Integration

Our integration process clarified that the primary structural feature of Bob Ross's paintings is not the presence of multiple similar-sized palette styles, but the dominance of a highly consistent standard color framework. Also, most paintings are variations built upon a shared foundational palette. Therefore, when we look at our domain question, the key insight is not how many distinct palette systems exist, but how strongly color usage is reflected around a stable compositional core.

Another important insight is centered on the nature of deviation. The smaller clusters provide additional nuance. They do not represent entirely separate styles, but rather more of coordinated

shifts in paint combinations relative to the dominant palette. These variations tend to emphasize certain tones like cool, warm, or high-contrast. Moreover, association rule analysis showed that these deviations are structured rather than random. This suggests that the stylistic diversity in Bob Ross's work shows through curated adjustments within a stable framework, rather than through fundamental shifts in palette structure. Basically variation builds upon consistency but not replacing it. For example, in Cluster 3 (n = 10), several rules achieved lift values up to 2.5 with 100% confidence at 30% support (Figure 5), indicating much stronger co-occurrence than expected under global frequencies. In contrast, many global rules involving high-frequency paints produced lift values closer to 1, reflecting baseline-driven usage rather than cluster-specific structure.

Integration also highlighted the degree to which color similarity is structurally stronger than we initially expected. Even under different thresholds and robustness checks, the majority of paintings remain closely connected in similarity space.When we looked more closely at the dominant colors within each cluster, we found that the smaller clusters are not built on completely different palettes, but instead reflect shifts in emphasis. For instance, the cool-toned cluster leans more heavily on blues and neutrals, the warm-accent cluster brings forward reds and earthy tones, and the high-contrast cluster highlights darker shades. Rather than representing entirely separate color systems, these clusters show how certain tones are accentuated while the overall palette framework remains recognizable. In contrast, the standard cluster combines these elements in a more balanced way. This pattern reinforces the idea that variation occurs through shifts in emphasis rather than through entirely new color systems.

Overall, the integrated clustering and rule analysis suggests that Bob Ross's artistic identity is defined by structural stability, instead of fragmented style divisions. The absence of sharply separated clusters reflects consistent palette architecture, not a limitation of the model. While meaningful deviations exist, they stay within recognizable boundaries and expand the familiar palette instead of moving away from it. Together, these findings provide a clearer answer to our guiding question: Bob Ross's color usage exhibits disciplined consistency with structured, bounded variation.

6. Limitations, Open Questions, and Next Steps

M3 clarified the overall palette structure by identifying consistent patterns. However, several limitations became more apparent. First, our analysis is limited to binary paint usage data. It focuses only on which paints are used, but does not capture how they are used or what is being painted. While this allows us to understand structural regularity in palette construction, it does not reflect other important aspects of artistic style. Second, although dendrogram threshold exploration justified the presence of a dominant mega-cluster and showed that it is not driven by inappropriate parameter selection, the extreme imbalance in cluster sizes remains a limitation. Their limited sample sizes of these clusters make it difficult to draw strong statistical conclusions about rare palette types. Several questions remain open:

- How do palette variations relate to painting title themes?

So far, text information has only been explored at a basic level, we observed that title similarity is much weaker than palette similarity. To explore further, we could group titles into thematic categories, and test whether they are associated with palette variations.

- Are rare clusters truly stylistic archetypes or simply edge cases?

There are very small clusters that lack strong statistical support, it is difficult to determine whether they represent meaningful stylistic variants or as low frequency combinations. Future work could test under alternative clustering settings, or validate them using additional information such as painting themes.

Our next step will focus on deeper text analysis and integration between title and palette structure. This may include grouping title keywords into themes and testing whether certain palette variants align with specific scene types. In addition, we plan to incorporate contextual variables such as season and episode to determine whether smaller clusters reflect stable stylistic patterns or are simply tied to specific production periods.

Appendix

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

M3 work (where to find the files):

- Project Name: BA820-Unsupervised-ML-Project
- Branch: "*Main*"
- Folder: "*M3_integrated_methods*"
- Primary notebook:
  "*BA820-Unsupervised-ML-Project/m3_integrated_methods/M3_Bob_Ross_Paintings.ipynb*"
- Support file (report PDF):
  "*BA820-Unsupervised-ML-Project/m3_integrated_methods/M3_Bob_Ross_Paintings_Report.pdf*"
  "*BA820-Unsupervised-ML-Project/m3_integrated_methods/M3_ML_Chatlog.pdf*"
- Folder documentation:
  "*BA820-Unsupervised-ML-Project/m3_integrated_methods/README.md*"

Notebook Link:
https://github.com/yujiunzou/BA820-Unsupervised-ML-Project/blob/main/m3_integrated_methods/M3_Bob_Ross_Paintings.ipynb

Contribution Table

Team member	Kean Zhu	Irene Tang	Janice Zou	Stephanie Chen
M2 contributions	Hierarchical clustering and DBSCAN to test whether palettes form distinct styles; documented DBSCAN sensitivity to ε and unclear cluster boundaries.	Hierarchical clustering with Jaccard (k=5); documented extreme imbalance; ran association rule mining (global and within-cluster) to interpret co-usage patterns and small-cluster signatures.	Association rules and hierarchical clustering to examine whether structure is driven by strong pairs vs broader multi-paint sets; explored stricter lift thresholds; ran robustness check removing Titanium_White; text-based EDA on title similarity.	Hierarchical clustering with Jaccard(k=5); evaluated Silhouette-based k selection(k=2); found silhouette favored overly rough solutions under imbalance; use threshold to define the core, optional, signature paint within each cluster.

Team member	Kean Zhu	Irene Tang	Janice Zou	Stephanie Chen
M3 role in integration	Led structural refinement of clustering strategy and interpretations.	Led integration of methods & clustering strategy	Led expansion of EDA beyond color usage, and interpretation synthesis.	Led narrative consolidation, interpretation synthesis.
Integrated analyses (directly contributed)	Added distance-threshold exploration for hierarchical clustering to test whether dominant cluster persists under finer Jaccard cutoffs; reframed imbalance as structural finding.	Introduced lift vs baseline comparison to distinguish true co-usage effects from high-frequency paints; aligned rule mining results with clustering imbalance findings.	Implemented title-based text similarity exploration in EDA; examined whether painting themes aligned with palette-based clusters; analyzed whether clusters reflected stylistic coherence beyond paint usage.	Refactored and reorganized EDA for logical flow; synthesized cross-method interpretations; ensured conclusions directly addressed prior feedback.
Indirect contributions	Coordinated integration between clustering and rule interpretation through verbal group discussion.	Coordinated integration between clustering and rule interpretation through verbal group discussion.	Coordinated integration between clustering and rule interpretation through verbal group discussion.	Coordinated integration between clustering and rule interpretation through verbal group discussion.
Abandoned integration attempts	DBSCAN was tested but abandoned due to high sensitivity to ε and unstable cluster boundaries on sparse binary data.	Less relevant result explorations were abandoned, but both hierarchical clustering and association rules methods were retained as it provided the most interpretable structure for M3.	K-means was considered but discarded because its Euclidean distance assumption was not appropriate for binary Jaccard-based data.	Silhouette-based k selection was explored but abandoned because imbalance consistently favored overly coarse cluster solutions.

Supplemental Material

Method 1:

Figure 1: Jaccard vs Hamming distance distribution

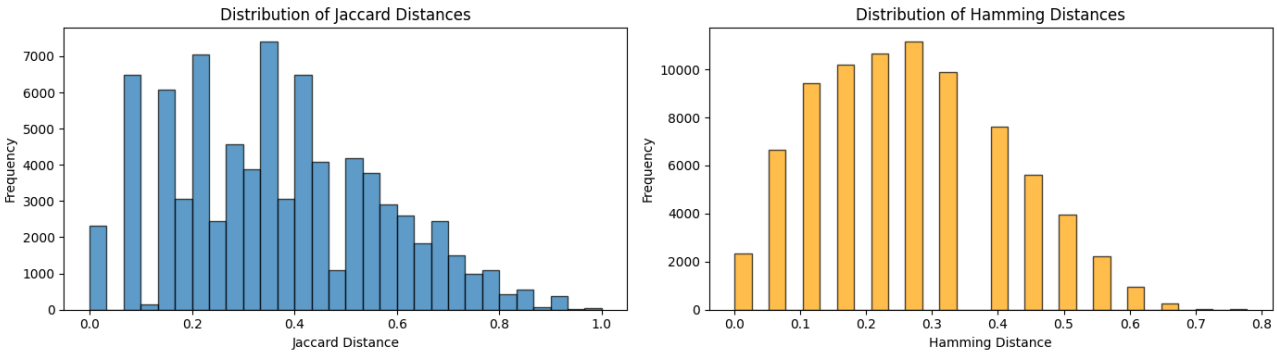


Figure 2: Dendrogram

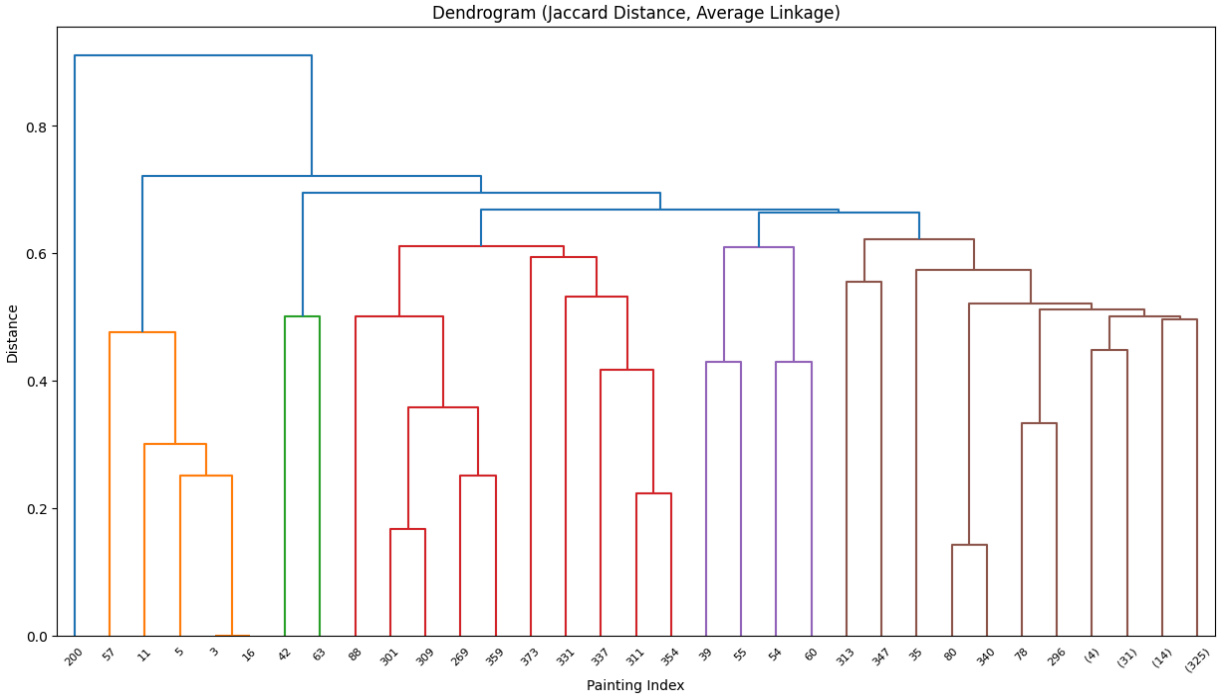


Figure 3: Cluster Size

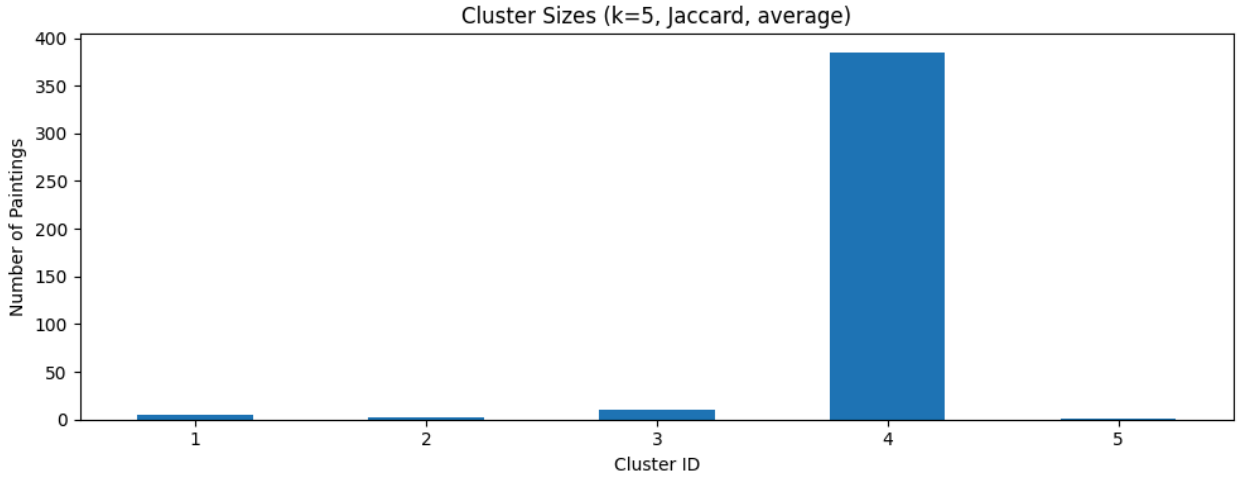
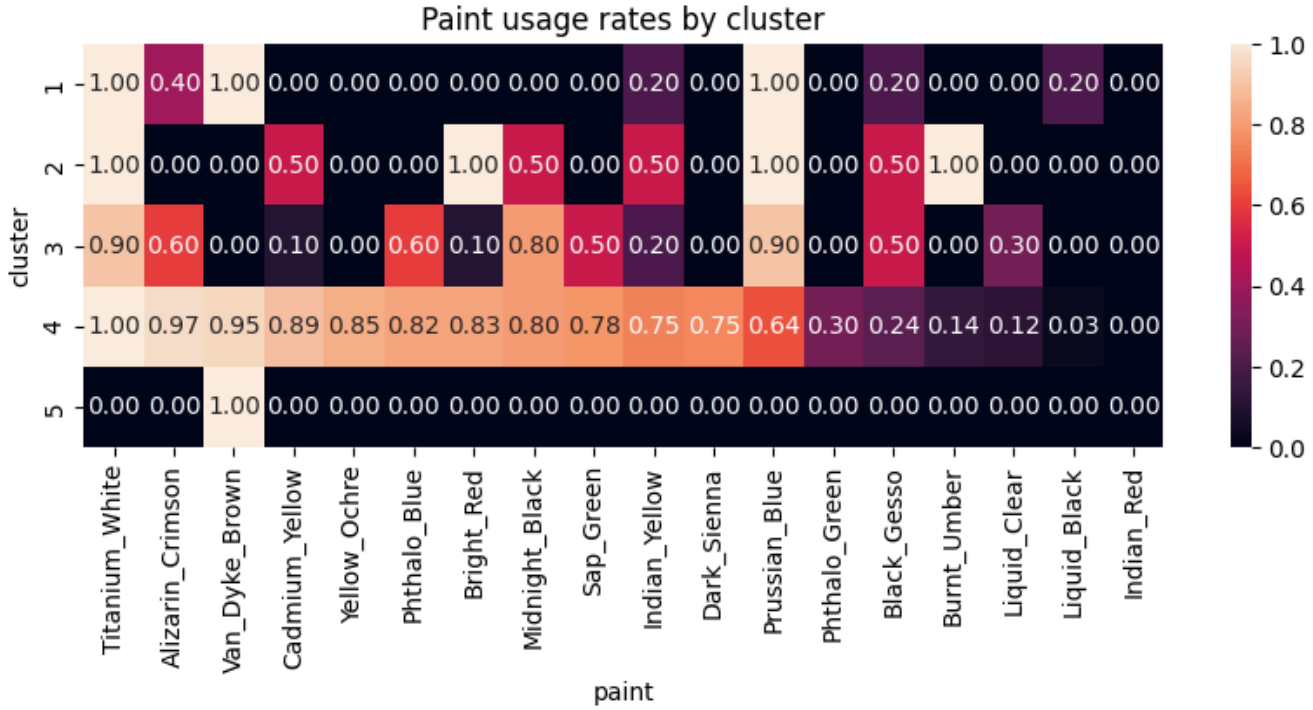


Figure 4: Cluster-level paint usage heatmap



Method 2:

Figure 5: Association Rules Table

```
 Cluster 1 | n=5
Too small for stable rules -> treat as rare palette case (skip rule mining).

 Cluster 2 | n=2
Too small for stable rules -> treat as rare palette case (skip rule mining).

 Cluster 3 | n=10
min_support=0.300, min_conf=0.60, min_lift=1.30
```

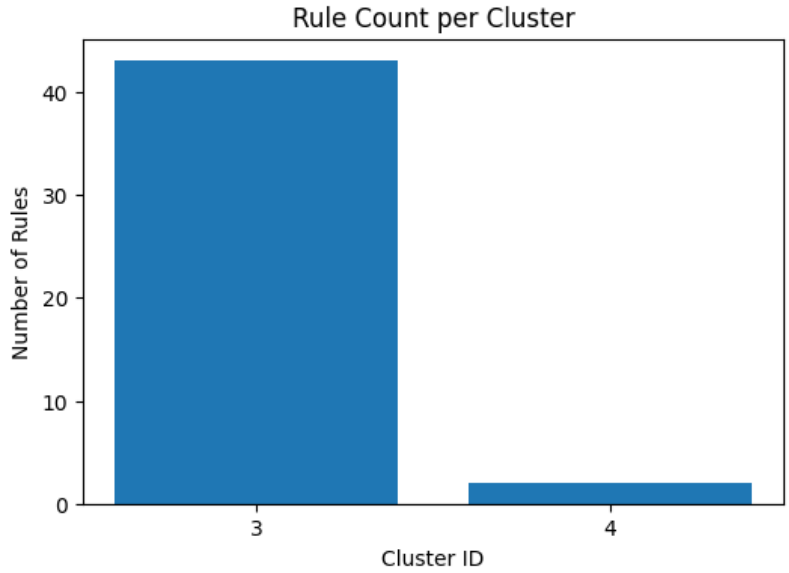
	antecedents	consequents	support	confidence	lift
0	Liquid_Clear	Black_Gesso, Sap_Green	0.3	1.00	2.5
1	Black_Gesso, Sap_Green	Liquid_Clear	0.3	0.75	2.5
2	Liquid_Clear	Black_Gesso	0.3	1.00	2.0
3	Liquid_Clear	Sap_Green	0.3	1.00	2.0
4	Liquid_Clear, Prussian_Blue	Black_Gesso	0.3	1.00	2.0
5	Liquid_Clear	Black_Gesso, Prussian_Blue	0.3	1.00	2.0
6	Liquid_Clear, Sap_Green	Black_Gesso	0.3	1.00	2.0
7	Black_Gesso, Liquid_Clear	Sap_Green	0.3	1.00	2.0
8	Black_Gesso, Midnight_Black	Sap_Green	0.3	1.00	2.0
9	Alizarin_Crimson, Black_Gesso	Sap_Green	0.3	1.00	2.0
10	Liquid_Clear, Prussian_Blue	Sap_Green	0.3	1.00	2.0
11	Liquid_Clear	Prussian_Blue, Sap_Green	0.3	1.00	2.0

```
 Cluster 4 | n=385
min_support=0.100, min_conf=0.60, min_lift=1.30
```

	antecedents	consequents	support	confidence	lift
0	Phthalo_Green	Prussian_Blue, Sap_Green	0.215584	0.715517	1.363733
1	Phthalo_Green	Phthalo_Blue, Prussian_Blue	0.200000	0.663793	1.324147

```
 Cluster 5 | n=1
Too small for stable rules -> treat as rare palette case (skip rule mining).
```

Figure 6: Rule Count per Cluster



Process Overview

1. **Data loading:** Load the Bob Ross dataset (403 paintings) and select the 18 paint indicator columns.
2. **Data validation**: Confirm paint indicators are binary, with no missing values/duplicates in the feature matrix.
3. **EDA**:
    a. Summarize global paint usage rates and assess palette similarity using pairwise Jaccard similarity to confirm recurring palette templates.
    b. Feature setup: Use the 18 paint indicators as binary vectors; treat each painting's palette as a set of paints.
4. **Method 1: Clustering**:
    a. Compute the Jaccard distance between paintings
    b. Run agglomerative clustering (average linkage)
    c. Clustering identified macro-level structure (baseline vs. deviations)
    d. Clustering with distance threshold instead of maxclust
    e. **Cluster interpretation**:
        i. Report cluster sizes
        ii. Summarize cluster paint profiles
5. **Method 2: Association rules (global & within-cluster)**:
    a. Remove trivial paints
    b. Mine the global rules as a baseline
    c. Mine the within-cluster rules for clusters with $n \geq 10$ and compare against the global baseline
6. **Insight Generation**: Translate outputs into a non-technical interpretation: a dominant "standard palette" archetype plus a small

Use of Generative AI Tool

Link: Not applicable. We used ChatGPT's group chat, which does not support shareable links. Full conversation documentation is provided in a separate PDF (M3_ML_Chatlog.pdf).

Summary:

In M3, we used ChatGPT to assist with analysis strategy clarification, coding support, and writing refinement. Specifically, we asked about whether the dominant cluster could be treated as a baseline and whether smaller clusters could be interpreted using lift comparisons. GPT suggested it could be reasonable and recommended several precautions, including checking cluster characteristics and ensuring stability when interpreting very small clusters.

We asked what alternative analytical approaches could help to uncover meaningful structure beyond the dominant baseline. ChatGPT suggested trying different distance measures and linkage methods, reweighting common items. We also asked whether instead of forcing a fixed number of clusters using "maxclust", we could cut the dendrogram using thresholds to explore alternative cluster structures. And also asked how to better organize our cluster size calculations when testing multiple distance thresholds. Based on the AI's suggestions and coding guidance, we implemented distance thresholds clustering, varied the Jaccard distance threshold and computed summary metrics for each cut. This allowed us to examine how cluster structure changed as the cut increased.

Additionally, we asked AI to suggest interpretable names and several key traits for our clusters, in order to verify that our interpretation of the paint-usage heatmap was reasonable. We also used it to review our report for grammar and spelling errors and to suggest corrections where needed.

Here is the code structure we used:

```
thresholds = [0.3, 0.5, 0.7, 1.0]
results = []

for t in thresholds:
    labels = fcluster(Z, t=t, criterion="distance")

    cluster_sizes = np.bincount(labels)[1:]  # ignore zero index
    n_clusters = len(cluster_sizes)

    largest_pct = cluster_sizes.max() / len(labels)
    smallest_pct = cluster_sizes.min() / len(labels)
    imbalance_ratio = cluster_sizes.max() / cluster_sizes.min()

    results.append({
        "threshold": t,
        "n_clusters": n_clusters,
        "largest_pct": largest_pct,
        "smallest_pct": smallest_pct,
        "imbalance_ratio": imbalance_ratio
    })

summary_df = pd.DataFrame(results)
```

Our use of AI was not extensive in M3, as most of the conceptual questions had already been explored during M2. In the previous stage, we individually asked the AI to clarify methodological approaches and certain coding implementation details. As a result, the main analytical framework had already been established prior to this phase.