

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:** Yu-Jiun Zou

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

In this milestone, I focus on refining Question 1 from the M1 proposal, which examines whether Bob Ross paintings exhibit meaningful structural patterns based on how paint colors are combined. Compared to M1, the core question itself remains unchanged; however, the focus has been refined. Instead of broadly asking whether palette structures exist, in this milestone I specifically investigate whether the observed structures are driven by meaningful multi-color combinations or are artifacts of a small number of dominant, high-frequency paints. This refinement was motivated by initial EDA revealing extreme imbalance in color usage, with certain paints appearing in nearly all paintings.

In the original framing, Question 3 implicitly assumed that certain individual colors contribute more to defining similarity and differentiation between paintings. Early analysis showed that Question 3 is largely frequency-driven, indicating that it serves more as an underlying assumption rather than as a research question. To evaluate this assumption, M2 conducts robustness checks comparing similarity structures with and without near-universal colors. The results show that removing dominant paints does not materially alter the similarity distribution, challenging the assumption that individual high-frequency colors primarily drive structural similarity. Overall, these findings validate the refined focus on multi-color co-occurrence patterns as the primary source of palette structure and reinforce Question 1 as the central research question for M2.

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

In M1, the EDA revealed substantial imbalance in paint usage, with several colors appearing in the majority of Bob Ross paintings. Usage rate analysis showed that several paints, particularly Titanium White, are nearly ubiquitous, while many others appear far less frequently (Figure A1). In addition, co-occurrence analysis indicates that these high-frequency paints often appear together (Figure A2), suggesting the presence of recurring color combinations rather than independent color choices. Baseline pairwise Jaccard similarity analysis further showed that many paintings share highly overlapping color palettes (Figure A3).

Building on these findings, I targeted robustness checks during M2. I tested the similarity after excluding the most used color (Titanium White) to assess whether the high similarity structure observed in M1 was driven by this near-universal color. Comparing the Jaccard similarity distributions before and after removal shows that the overall shape of the similarity distribution remains largely unchanged (Figure A4). While the average similarity decreases slightly (Figure A5), the distribution continues to exhibit a strong right-skew, with a small number of painting pairs displaying very high similarity. As a supplementary exploratory step, M2 also included basic text-based EDA on painting titles, examining word frequency patterns and pairwise Jaccard similarity between title token sets (Figure A6). Aside from the targeted robustness checks and

supplementary text exploration introduced in M2, the core EDA findings and preprocessing steps established in M1 remained unchanged. Paint usage indicators, binary encoding, and the overall data structure were preserved to ensure consistency across analyses.

3. Analysis & Experiments (~1.5 page)

Method 1: Association rule mining

Association rule mining was applied to test whether recurring palette structures in Bob Ross paintings are driven by strong dependencies between individual paints or by broader multi-paint combinations. This analysis directly supports the domain question of whether Bob Ross palettes follow a small number of templates, and whether those templates arise from dominant color pairings or coordinated multi-color usage. The data consist of binary indicators of paint presence, which is the standard input format for Apriori-based association rule mining. Consistent with the EDA findings, Titanium White was excluded prior to analysis to avoid trivial associations caused by its near-universal presence.

Multiple threshold settings were tested to balance coverage and interpretability. A minimum support of 0.05 was selected to focus on color combinations that appear in a meaningful fraction of paintings, while confidence ≥ 0.6 was used to retain reliable conditional relationships. Lift thresholds were varied to distinguish genuine dependence from frequency-driven effects.

Using a minimal lift threshold (>1) produced many one-to-one color rules that were largely driven by marginal color frequencies rather than true dependence. Applying a stricter threshold of lift ≥ 2 eliminated these spurious pairwise associations, leaving no strong single paint-to-paint dependencies. Instead, the remaining associations involved multi-color combinations, indicating that palette structure in this dataset is characterized by coordinated co-occurrence among several paints rather than isolated color pairings.

Method 2: Hierarchical Clustering

Hierarchical clustering was applied to explore whether Bob Ross paintings exhibit meaningful palette-level structure based on overall color composition similarity. Using Jaccard distance on binary paint usage vectors allows direct comparison of shared versus distinct color sets between paintings, which aligns with the domain question of whether recurring palette templates exist beyond individual color frequencies. Average linkage was selected to balance sensitivity to local similarity and global structure.

This approach is appropriate given the binary and sparse nature of the data, where Euclidean distance is less interpretable. Dendrogram suggested a small number of dominant clusters alongside a long tail of highly similar paintings, reflecting strong overall palette overlap. Several cut levels were tested, with a 5-cluster solution used for reporting and interpretation. However, cluster assignments remained highly imbalanced (Table A1), with most paintings grouped into a

single dominant cluster and only a few outliers forming small clusters. This indicates that while hierarchical clustering confirms high global similarity among palettes, it provides limited discrimination of fine-grained palette styles, highlighting the dominance of shared color combinations across paintings.

Method 3: K-means Clustering (Binary paint vectors)

K-means clustering was explored as a complementary method to assess whether palette structures could be captured through centroid-based partitioning of binary color usage patterns. Although K-means is not theoretically optimal for binary data, it was tested to evaluate whether it could yield interpretable group-level palette profiles that complement hierarchical results.

Multiple values of k were examined using elbow and silhouette diagnostics, with $k = 4$ chosen based on a balance between interpretability and clustering stability. To improve stability, the algorithm was initialized multiple times with different random starting points, and the solution with the lowest distortion was retained. Compared to hierarchical clustering, K-means produced more balanced cluster sizes and clearer differences in average paint usage across clusters (Table A3). Cluster-level heatmaps revealed distinct palette tendencies, suggesting that centroid-based aggregation can surface dominant color profiles despite the binary input. However, silhouette scores remained moderate reflecting overlap between clusters and reinforcing that palette boundaries are not sharply separated (Figure A7).

The main limitation of K-means is its mismatch with the binary and sparse nature of the data, which weakens the interpretability of distance-based centroids. Although the method highlights average palette tendencies, overlapping clusters and moderate silhouette scores suggest that the resulting groups should be viewed as descriptive summaries rather than distinct or well-separated palette styles.

4. Findings & Interpretations (~0.75 page)

This analysis reveals that Bob Ross's color palettes exhibit strong structural regularity, but that structure does not arise from rigid one-to-one color pairings. Instead, it emerges from the repeated use of coordinated sets of multiple paints. Many paintings share a large portion of their palettes, indicating a strong underlying structure rather than random or ad hoc color selection.

Across the dataset, many paintings share a large portion of their color sets. Pairwise Jaccard similarity analysis (Figure A3) shows a strong right-skewed distribution, indicating that a substantial number of painting pairs have highly overlapping palettes. Importantly, this pattern persists even after removing Titanium White, a nearly universal paint. While excluding this dominant color slightly reduces average similarity, the overall distributional shape remains stable (Figure A4). This suggests that palette similarity is not an artifact of a single ubiquitous paint, but rather reflects recurring combinations of several commonly used colors.

Association rule analysis further clarifies the nature of this structure. Under minimal thresholds, many apparent color-to-color relationships emerge; however, these are largely explained by marginal usage frequency rather than meaningful dependence. When stricter criteria are applied, no single paint pair exhibits a strong one-to-one dependency (Table A1). Instead, the strongest associations consistently involve multi-color combinations. This indicates that Bob Ross's palettes are organized around broader color sets rather than fixed pairings, supporting the interpretation that stylistic consistency is achieved through holistic palette composition.

Clustering results support this overall picture. Most paintings naturally fall into one large group with very similar color palettes, while only a small number stand out as clearly different (Table A2). When the paintings are grouped in alternative ways, the resulting clusters highlight general tendencies in color use rather than strict or isolated styles. Most paintings share similar colors, with groups differing mainly in which colors are emphasized rather than following distinct palette types.

A brief exploration of painting titles provides complementary context. Title-based similarity is generally low (Figure A6), indicating that textual descriptions vary more freely than color usage. This contrast highlights that while Bob Ross's naming choices are diverse, his instructional color choices remain comparatively stable, reinforcing the idea of a consistent underlying palette framework.

To summarize, these findings suggest that Bob Ross's paintings follow a small number of recurring palette templates characterized by overlapping, multi-color combinations rather than isolated color rules. From a practical perspective, this insight may benefit art educators, digital content creators, and generative art systems by emphasizing the importance of coordinated palette selection over strict color pairing rules when replicating or learning Bob Ross's style.

5. Next Steps (~0.25 page)

Several aspects of the analysis remain open for refinement. While initial efforts were made to group title keywords into broader themes, this information has not yet been fully integrated into the clustering analysis. Future work could more systematically incorporate these structured text features to better understand how textual descriptions relate to palette usage.

In addition, the role of clustering warrants further clarification. Although clustering revealed consistent patterns in color usage, it remains unclear whether these groups represent stylistic templates, instructional patterns, or descriptive summaries of gradual palette variation. Finally, future analysis could focus on more deliberate integration of text-based and color-based information. Current findings suggest strong regularity in color usage alongside greater flexibility in titles, and further exploration of how these dimensions complement each other may provide deeper insight into how Bob Ross's style is constructed and communicated.

Appendix

Shared GitHub Repository (Required)

- Link to the shared team GitHub repository:
<https://github.com/yujiunzou/BA820-Unsupervised-ML-Project.git>
- Under Branch: *Yujiun_Zou*, Folder: *M2_Yujiun_Zou*, NotebookFile: *M2_Bob_Ross_Paintings_yujiunzou.ipynb*

Supplemental Material

- EDA figures and tables

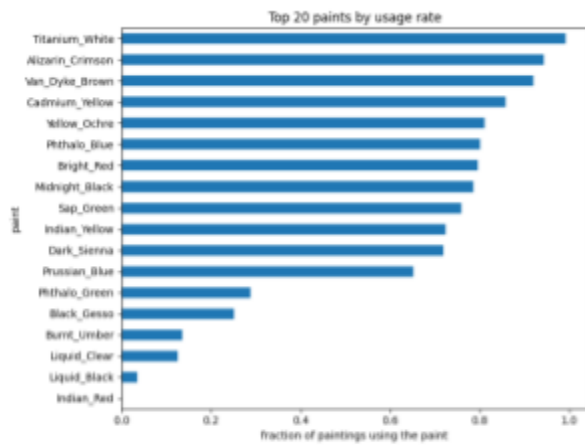


Figure A1. Top 20 paints by usage rate

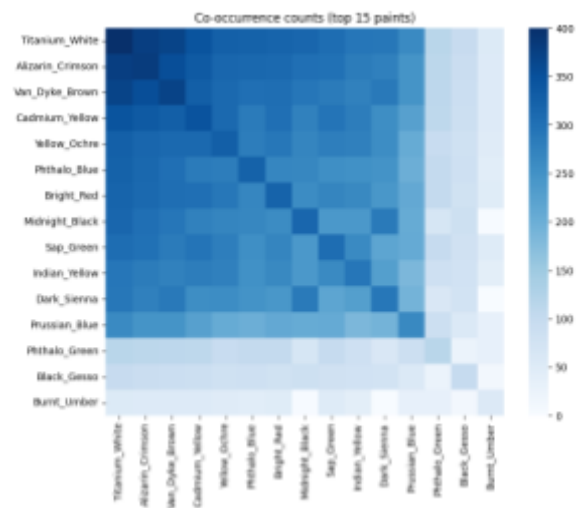


Figure A2. Paint co-occurrence heatmap (top 15 paints)

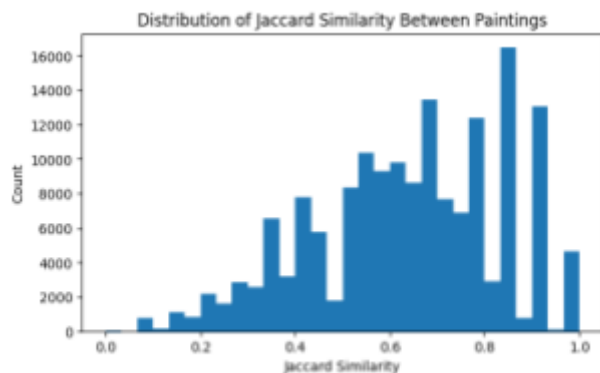


Figure A3. Distribution of Jaccard similarity between paintings

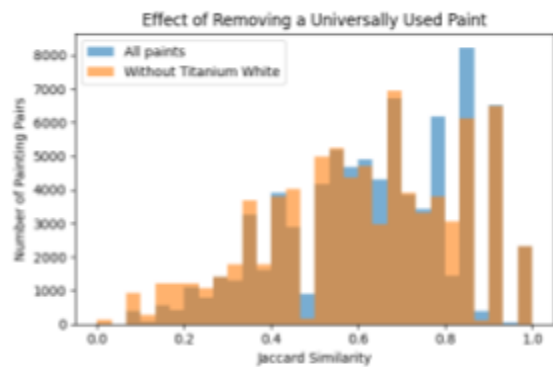


Figure A4. Effect of removing a universally used paint on Jaccard Similarity

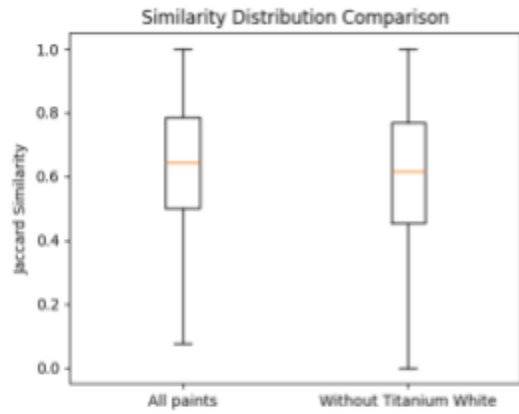


Figure A5. Similarity Distribution Comparison
(all colors vs without titanium white)

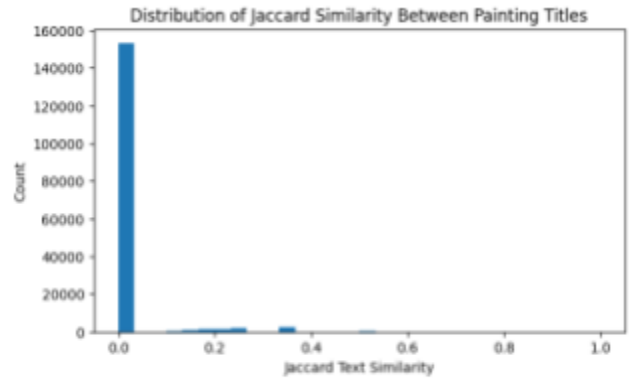


Figure A6. Distribution of Jaccard Similarity Between Painting Titles

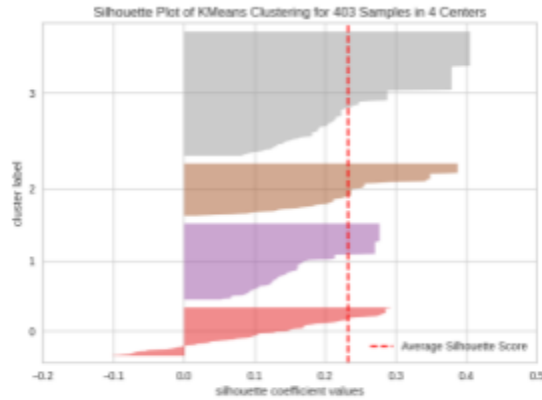


Figure A7. Silhouette Plot of K-means

```
rules_filtered = rules_filtered.copy()
rules_filtered["ante_len"] = rules_filtered["antecedents"].apply(len)
rules_filtered["cons_len"] = rules_filtered["consequents"].apply(len)
pair_rules = rules_filtered[(rules_filtered["ante_len"] == 1) & (rules_filtered["cons_len"] == 1)]
pair_rules.sort_values(by="lift", ascending=False)

pair_rules.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 0 entries
Data columns (total 10 columns):
 #  column          Non-Null Count  Dtype
---  ---
 0  antecedents      0 non-null      object
 1  consequents      0 non-null      object
 2  antecedent support  0 non-null      float64
 3  consequent support  0 non-null      float64
 4  support          0 non-null      float64
 5  confidence       0 non-null      float64
 6  lift            0 non-null      float64
 7  representativity  0 non-null      float64
 8  leverage        0 non-null      float64
 9  conviction       0 non-null      float64
10  zhangs_metric    0 non-null      float64
11  jaccard         0 non-null      float64
12  certainty       0 non-null      float64
13  kulczynski      0 non-null      float64
14  ante_len        0 non-null      int64
15  cons_len        0 non-null      int64
dtypes: float64(12), int64(2), object(2)
memory usage: 0.0+ bytes
```

Table A1. No one-to-one association after lift>2



Figure A8. K-means Cluster-Level Point Usage Profiles

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=n_clusters, random_state=42, n_init=10)

bob_ross["cluster_kmeans"] = kmeans.fit_predict(X).astype(str)
bob_ross["cluster_kmeans"].value_counts()
```

cluster_kmeans	count
3	166
1	102
2	70
0	65

Table A3. K-means clusters

```
n_clusters = 5
bob_ross["cluster_labels"] = fc
bob_ross["cluster_labels"].value
```

cluster_labels	count
4	385
3	10
1	5
2	2
5	1

Table A2. Hierarchical Clustering

- **Extended Analysis: Combined Text and Color Hierarchical Clustering**

To explore whether semantic information from painting titles complements color-based similarity, conducting an extended hierarchical clustering analysis that integrates both modalities. Textual similarity was represented using Jaccard distance on title keyword sets, while palette similarity was computed using Jaccard distance on binary paint usage vectors. Greater weight was assigned to paint-based distance (0.9) than to title-based distance (0.1). This choice reflects earlier findings that color usage exhibits strong and stable structure across paintings, whereas title similarity is generally weak and highly variable. The resulting dendrogram shows that most paintings cluster into one large, highly similar group, while a smaller number of paintings separate into secondary clusters. Examination of cluster-level title keywords suggests that these smaller clusters are associated with more specific themes, such as winter scenes or introspective subjects, while the largest cluster reflects broadly shared landscape motifs. Overall, the combined clustering reinforces the conclusion that Bob Ross's style is driven primarily by consistent palette usage, with titles offering supplementary but not defining structure.

count	
1	265
3	77
2	35
4	25
5	1

Cluster 1:	[('mountain', 42), ('autumn', 18), ('lake', 16), ('oval', 15), ('sunset', 13)]
Cluster 2:	[('winter', 20), ('oval', 5), ('cabin', 3), ('solitude', 3), ('barn', 3)]
Cluster 3:	[('winter', 8), ('snow', 3), ('grey', 2), ('night', 2), ('falls', 2)]
Cluster 4:	[('lady', 1), ('contemplative', 1)]

Use of Generative AI Tools

Generative AI tools (ChatGPT) were used as conceptual and technical support during the analysis. I asked ChatGPT early in the project how clustering can be applied to binary data, including whether K-means can be used for binary paint-usage vectors, what limitations this entails, and what alternative methods (e.g., k-modes) are more appropriate. After obtaining clustering results, I asked ChatGPT whether my interpretations, such as cluster imbalance, moderate silhouette scores, and overlap between clusters were reasonable and consistent with standard analytical understanding.

I also asked ChatGPT for guidance on text analysis, including how to write code to extract and preprocess painting titles, how to represent short text data for similarity analysis, and how to summarize keyword patterns. In the later stages, I asked ChatGPT to walk me step by step through how text-based features could be combined with binary paint features for clustering, including how to structure the code, compute separate distance matrices, and integrate them

using weighted distance fusion. ChatGPT was additionally used to help identify and fix coding errors and to refine the organization and clarity of written sections.

All data preprocessing, analysis design, coding, interpretation of results, and final writing were completed by the author. ChatGPT did not generate final code or analytical outputs, but was used to support understanding of concepts, implementation logic, and methodological choices.

Link: <https://chatgpt.com/share/69893104-d918-8003-9ac3-10d12f9c1268>

Examples:

is it correct to say that the average silhouette score is modest indicating that the clusters are only moderately separated rather than clearly distinct.

explain k-mode clustering for unsupervised machine learning and how it work with data and python code with resources

Yes — **that statement is correct**, and it is **well-phrased**.

Here is a **slightly refined version** that reads more academically polished (optional, but recommended):

The average silhouette score is modest, indicating that the resulting clusters are only moderately separated rather than clearly distinct.

If you want one step more precision (still safe and accurate), you could also write:

The modest average silhouette score suggests that while some cluster structure is present, the separation between clusters is limited and not sharply defined.

```
stopwords = ["the", "of", "a", "in", "at", "and", "to", "with", "by", "for"]
title_sets = [
    setross["painting_title"]
    .str.lower()
    .str.replace(r"[^\w]", "", regex=True)
    .str.split()
    .apply(lambda x: {w for w in x if w not in stopwords})
]
```

i had this what is the next step

You're **exactly at the right place**.

From `title_sets` (sets of words per painting), the **next steps are Jaccard → clustering**.

Below is the **minimal, correct continuation**, step by step.

Step 2 — Compute Jaccard distance matrix

Clustering needs `distance = 1 - similarity`.

```
python
import numpy as np

def jaccard(a, b):
    return len(a & b) / len(a | b) if (a | b) else 0

n = len(title_sets)
D = np.zeros((n, n))

for i in range(n):
    for j in range(i + 1, n):
        dist = 1 - jaccard(title_sets.iloc[i], title_sets.iloc[j])
        D[i, j] = dist
```

Below is a **clear, exam-ready explanation** of **K-modes clustering**, how it works with data, and **Python code** you can actually run. This is framed for **unsupervised ML coursework / interviews**, especially when data is **categorical**.

1. What is K-modes clustering?

K-modes is an unsupervised clustering algorithm designed for **categorical data**.

- It is an extension of **K-means**
- **K-means** → numerical data
- **K-modes** → categorical data

Examples of suitable data:

- Gender: (Male, Female)

Ross's style is driven primarily by consistent palette usage, with titles offering supplementary but not defining structure.

Here is a **polished, academically strong revision**. I've kept your meaning, tightened the language, and fixed tense/flow issues while avoiding overclaiming.

Refined version

To examine whether semantic information from painting titles complements color-based similarity, we conducted an extended hierarchical clustering analysis that integrates both modalities. Textual similarity was measured using Jaccard distance computed on sets of title keywords, while palette similarity was quantified using Jaccard distance on binary paint-usage indicators. Greater weight was assigned to paint-based distance (0.9) than to title-based distance (0.1), reflecting earlier findings that color usage exhibits strong and stable structure across paintings, whereas title similarity is sparse and highly variable.