BA820 – Project M2

Cover Page

- **Project Title: Unsupervised Analysis of Color Style Patterns Across Seasons in Bob Ross Paintings**
- **Section and Team Number: B1, Team05**
- **Student Name: Kefei Zhang**

1. Refined Problem Statement & Focus

I focus on the third domain question from our proposal: how Bob Ross's visual style changes across seasons(different seasons of the anime). However, since the feedback received indicated that this question was unclear, I will redefine and explain it.

I operationalize visual style using measurable palette features, including binary color usage vectors and palette complexity. Rather than treating seasonal style change as a subjective artistic shift, I define it as a measurable difference in feature structure and grouping patterns.

This refined question also has practical relevance beyond analysis. By identifying whether seasonal differences appear as shifts in cluster composition, the results support a more structured understanding of artistic style. For content platforms, archives, and media curators, cluster-based style segmentation enables more systematic categorization and grouping of paintings beyond simple season labels. For art educators and creative learning tools, understanding how style clusters are distributed across seasons can help translate artistic style into reproducible palette patterns and teaching templates. For creative software and digital art tools, these findings can inform palette recommendation systems and style-based search or tagging features.

EDA showed that core color usage and palette size are more stable across seasons than I initially expected, with most paintings using a similar number of colors and a consistent core palette. This challenged my initial assumption that seasonal style differences would appear primarily as large shifts in palette complexity or dominant colors. As a result, I narrowed the focus of the question: instead of asking whether overall style changes by season, I now investigate whether seasons differ in their mixture of underlying style clusters.

This reframing keeps the original motivation but makes the question measurable and directly supported by unsupervised analysis.

2. EDA & Preprocessing: Updates

In M1, the EDA for Q3 focused on seasonal differences in core color usage frequency. The results showed that the average number of colors used per painting varies moderately across seasons, while the overall usage rate of core colors remains relatively stable after the earliest seasons. In particular, Season 1 shows noticeably lower core-color usage and lower palette complexity, while mid-period seasons are more consistent. These findings suggest that Bob Ross's style maintains a stable core palette structure across most seasons.

While the M1 exploration suggested that core color usage is broadly stable across most seasons, those results were based mainly on averages and could not fully capture pattern-level variation or unusual paintings. To better understand whether deeper palette structure and style differences exist, we added several focused EDA visualizations in M2. These additional views look at full distributions, pattern structure, and unusually different works, and help build a stronger foundation for the clustering and deviation analysis that follows.

The palette size distribution plot was added to examine variation spread instead of only mean palette size. It shows that later seasons have a wider range of color counts, suggesting greater flexibility in palette choice. I also added a chart that shows how different each painting's color mix is compared to the overall typical palette. This reveals that most paintings are close to the typical style, but a smaller group deviates clearly. The chart showing how unusually different paintings are distributed across seasons further shows that these atypical paintings appear more often in early and some later seasons, while mid-period seasons are more consistent. Finally, the typical vs. high-deviation color profile chart was added to explain what drives deviation. It shows that unusual paintings usually keep the core colors but differ in secondary colors. Together, these added EDA views show that the style keeps a stable core palette while allowing structured variation and occasional atypical works, which motivates the clustering and deviation analysis used next.

From a preprocessing standpoint, I did not remove high-deviation paintings as outliers, because identifying and interpreting atypical style patterns is part of the Q3 objective. Instead, I treated them as a separate subset ("high-deviation group") and analyzed their seasonal distribution and color profile differences.

3. Analysis & Experiments

I employ two unsupervised analysis methods: K-Means clustering and distance-based unsupervised analysis. Before modeling, I first preprocessed and filtered the color features. The original color variables comprised 18 binary indicator variables. I calculated the usage frequency of each color across all works and removed extremely low-frequency color features with usage rates below 5%. Such features contribute minimally to distance calculations and clustering but increase dimensional sparsity and introduce noise, thereby reducing clustering stability.

Method 1: K-Means Style Clustering Analysis (Unsupervised Clustering)

- What problem does this method address, and how does it help answer research questions?

I use K-Means to group paintings based on their color combinations without using labels. Then I examine how paintings from different seasons fall into these groups. If some seasons appear more often in certain groups, it suggests that style structure changes by season.

- Why is this method suitable for the current data and objectives?

Although the color features are binary indicators, K-Means clustering was still appropriate because each painting can be represented as a palette usage vector, and the goal is to discover recurring color combination templates. After feature scaling, Euclidean distance provides a reasonable similarity approximation.

I also considered hierarchical clustering, which has theoretical advantages for binary features. However, our goal is to identify clear and interpretable palette templates for each style group. K-Means directly produces cluster centers, which makes it easier to describe typical color patterns and link clusters with seasons and style deviation results.

- What I Tried (Parameters, Variants, Method Adjustments)

I systematically tested different clustering numbers k (2–10) and evaluated clustering quality using the silhouette score. We also experimented with two feature versions: one using all 18 color features, and another applying low-frequency color filtering (removing colors with usage rates <5%). Additionally, we compared versions with and without feature normalization. In the final setup, I grouped the paintings into 4 clusters.

- What unexpected findings and lessons learned

An unexpected finding was that while k=2 yielded one of the highest silhouette scores, its overly coarse structure provided low interpretive value. This prompted us to prioritize interpretability over single-metric optimization in model selection.

Method 2: Distance-based unsupervised analysis

- What problem does this method address?

I measure how different each painting's color mix is from the overall typical palette. By comparing these scores across seasons, we can see which seasons are more stable and which contain more style outliers.

- Why does this method suit the current data and objectives?

This method compares each painting to the average color pattern without using labels. It is useful for spotting typical versus unusual styles.

- What I Tried (Parameters, Variants, Method Adjustments)

I tested several ways to measure style difference, including using the original 0/1 color features versus scaled features, and checking whether removing very rare colors changes the results. I also tried different definitions of the "typical palette," using both averages and medians, and compared results across seasons using both average and maximum deviation. The overall patterns stayed similar across these choices, so we kept the average-based center because it is simpler and easier to explain.

- What worked, what didn't, unexpected findings, and lessons learned

This method revealed a clear time pattern: style differences are larger in early seasons, become much smaller in the middle period, and increase again later. I found that using unscaled features did not work well because very common colors dominated the distance measure, so that approach was dropped. One unexpected result was that highly unusual paintings do not appear only at the beginning or end of the series, but show up occasionally throughout. This suggests that unusual styles are often linked to specific themes rather than time alone.

4. Findings & Interpretations

This analysis shows that BR's painting style does not change randomly across seasons but follows a clear three-stage pattern: early exploration, middle stability, and later diversification. Clustering results show that early seasons are dominated by one main style group, middle seasons stay relatively consistent, and later seasons become more mixed, with multiple style structures appearing at the same time.

Deviation results support the same pattern. Average Style Deviation by Season shows that early paintings differ more from the typical style, middle seasons are closest to it, and later seasons become more varied again. This means style change is not linear. High-deviation paintings appear across several seasons rather than only one period, suggesting they are often driven by specific themes or creative choices. Typical vs High-Deviation Radar further shows that these differences mainly come from systematic shifts in key warm core colors.

These findings have practical relevance for several real-world groups. For **art educators and learning platforms**, the middle-stable phase can serve as a reference style baseline for teaching and beginner practice, while high-variation works can be used as advanced or creativity-focused case studies. For **content and recommendation platforms**, style stability and deviation metrics can support style-based content tagging and tiered recommendations, matching consistent-style works to beginners and higher-variation works to advanced users seeking inspiration.

5. Next Steps

There are still some limits in the current analysis. I only used binary color indicators and season/episode information, and did not measure color proportions, composition, or detailed themes. Because of this, some style differences may not yet be fully captured.

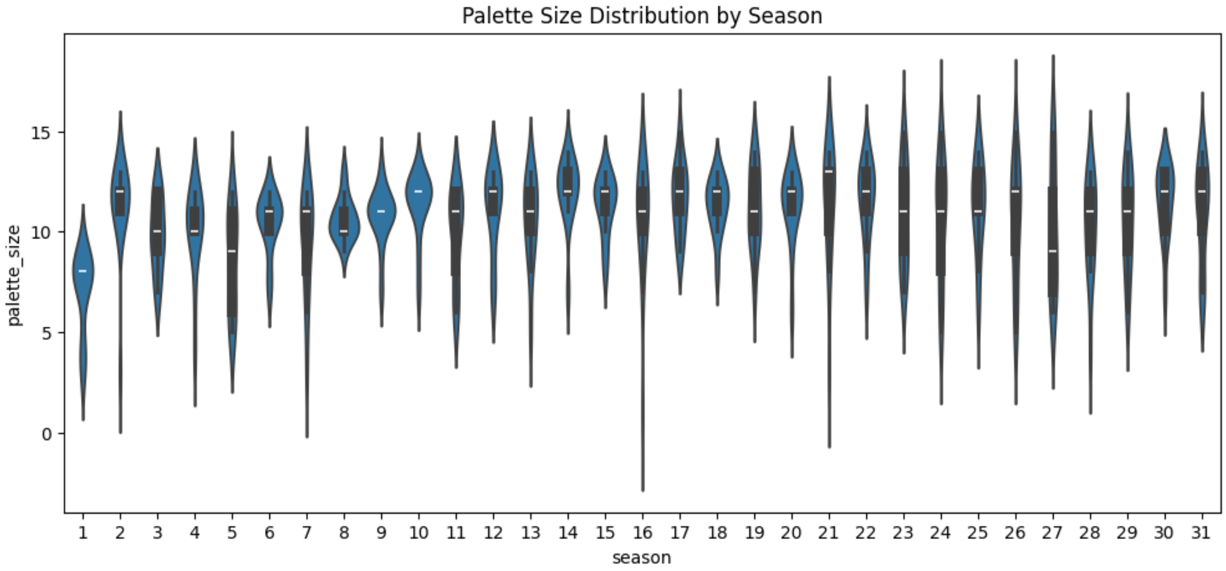
Next, I plan to add color proportion features and test other distance and clustering methods to check whether the style segments remain stable. I also plan to study episode-level changes within seasons instead of only seasonal summaries. Some questions remain open — I see clear stage-based style shifts, but I do not yet know what drives them. Since our results show real structure in style variation, further analysis is justified to better explain the causes behind these changes.

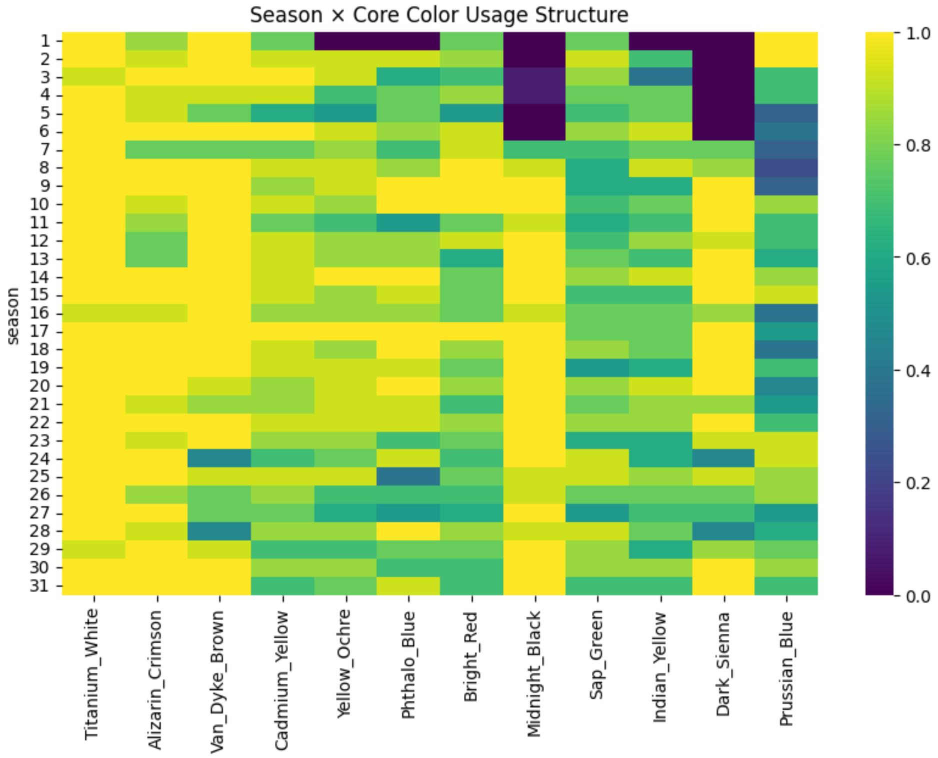
Appendix

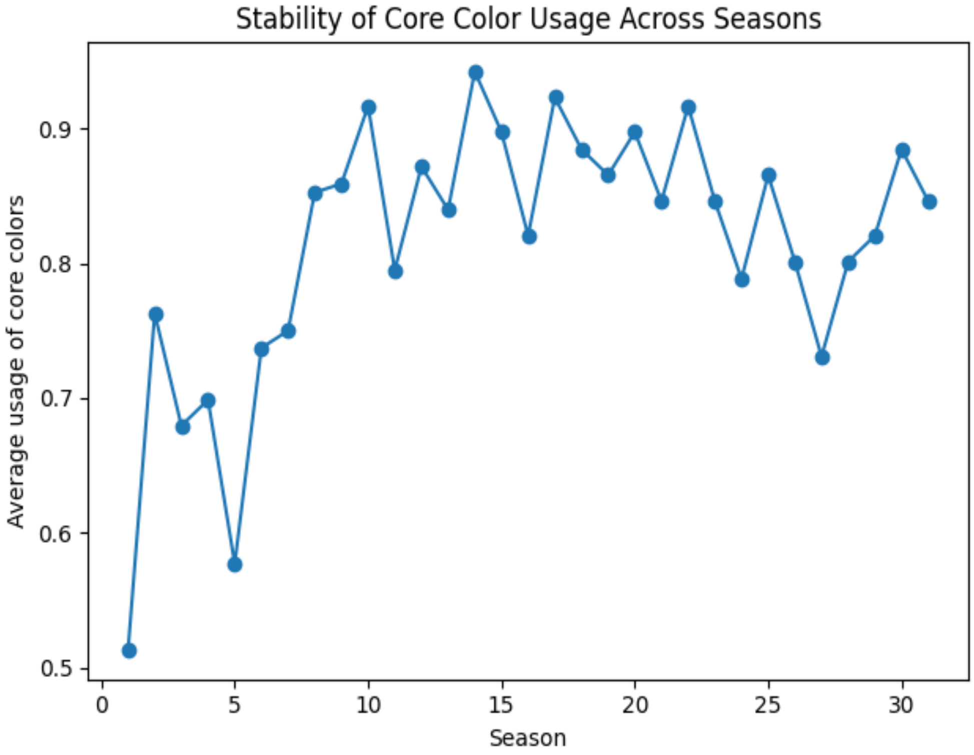
Shared GitHub Repository (Required)

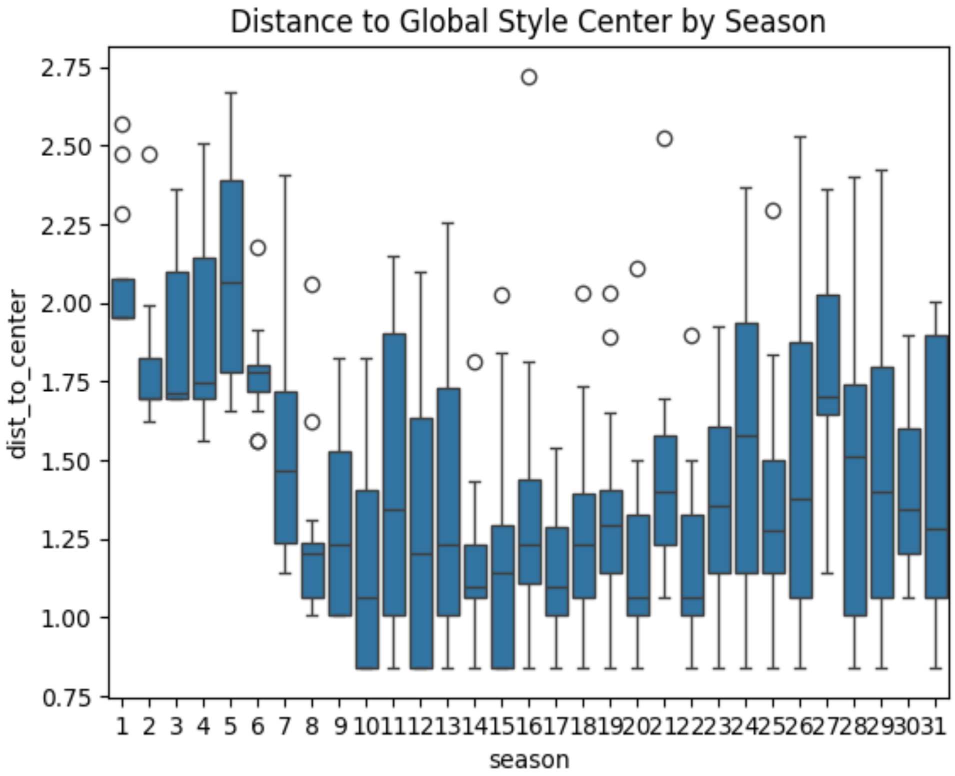
- https://github.com/Charles-Wei77/-ba820-bob-ross-team05
- Branch: Kefei-Zhang
- Report: BA820_M2_Kefei Zhang_Team05_2026.pdf
- Notebook: 820_Bob_Ross_Paintings_Team05_Kefei Zhang.ipynb

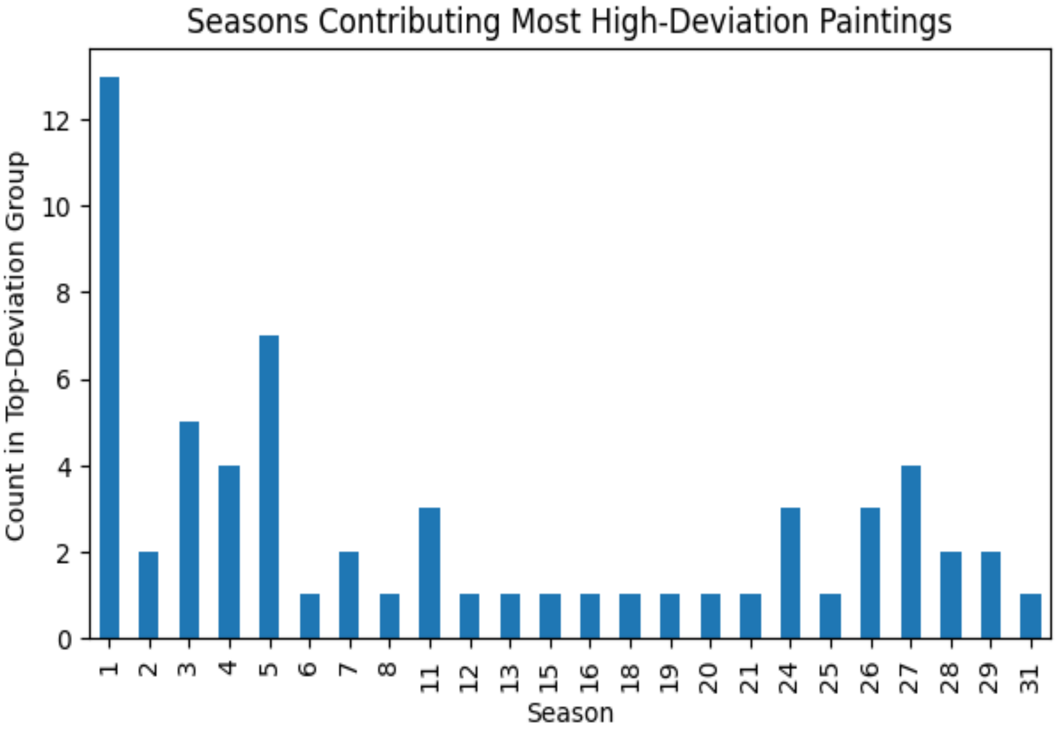
Supplemental Material (Highly Recommended)

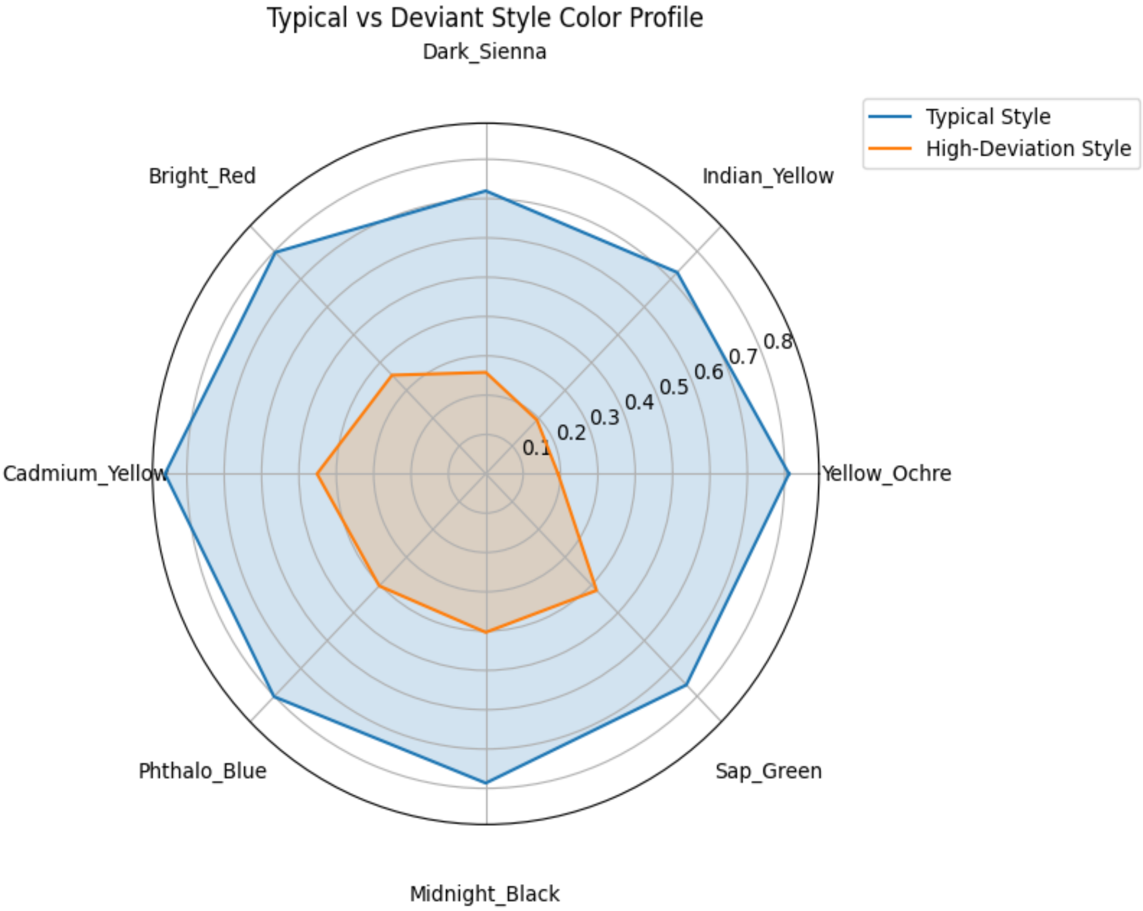


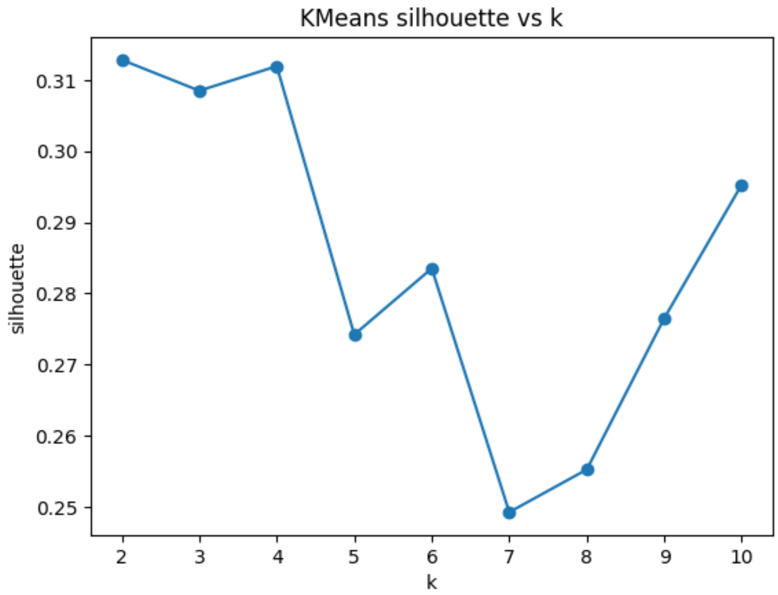



Stability of Core Color Usage Across Seasons
Average usage of core colors
0.5
0.6
0.7
0.8
0.9
0
5
10
15
20
25
30
Season

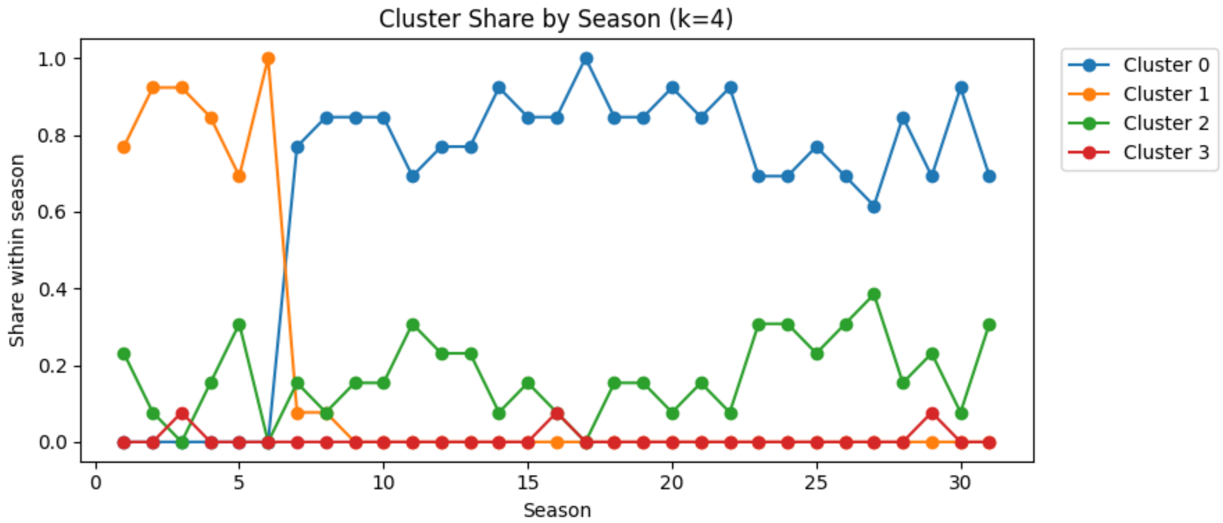

Distance to Global Style Center by Season
dist_to_center
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
1 2 3 4 5 6 7 8 9 10111213141516171819202122232425262728293031
season


Seasons Contributing Most High-Deviation Paintings
Count in Top-Deviation Group
0
2
4
6
8
10
12
Season
1
2
3
4
5
6
7
8
11
12
13
15
16
18
19
20
21
24
25
26
27
28
29
31

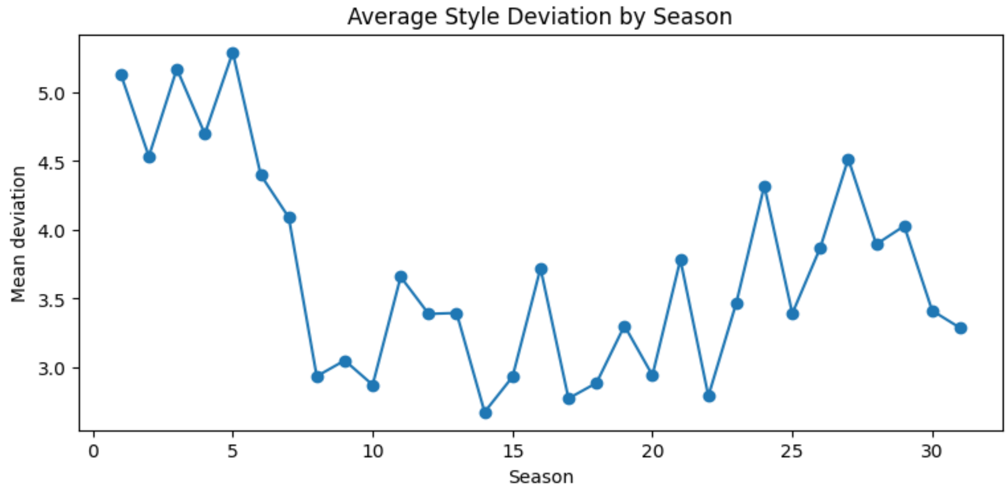

Typical vs Deviant Style Color Profile
Typical Style
High-Deviation Style
Dark_Sienna
Indian_Yellow
Yellow_Ochre
Sap_Green
Midnight_Black
Phthalo_Blue
Cadmium_Yellow
Bright_Red
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8



KMeans silhouette vs k



Cluster Share by Season (k=4)



Average Style Deviation by Season

Process Overview

- Conceptual Flow
    - Painting palette data
    - Identify usable color features
    - Clean and refine style indicators
    - Explore seasonal palette patterns
    - Detect natural style segments
    - Measure style deviation from the typical pattern
    - Compare seasons by structure and deviation
    - Generate style evolution insights
    - Translate findings into practical relevance

This conceptual flow shows how the analysis moves from raw descriptive palette information to interpretable style-structure insights and real-world creative or content applications.

- Technical Flow
    - Raw dataset (palette indicators + season/episode)
    - Binary color feature detection
    - Color frequency calculation
    - Low-frequency color filtering (≥5%)
    - Feature matrix construction (paintings × colors)
    - Feature standardisation
    - Unsupervised Method 1 — K-Means clustering
        - k parameter testing (2–10)
        - silhouette evaluation
        - cluster–season distribution
    - Unsupervised Method 2 — Style deviation distance
        - global style center
        - distance computation
        - high-deviation identification
    - Season-level aggregation
    - Feature-level deviation profiling
    - Insight interpretation and validation

Use of Generative AI Tools

Further exploration of Q3's EDA led to a brainstorming session with AI, particularly the investigation into “Which color features most strongly differentiate high-deviation paintings from the typical style profile” provided me with visualizations and production ideas to choose from. However, I personally made modifications to the presentation.

After consulting ChatGPT on whether hierarchical or k-means clustering was more suitable for my question, I ultimately chose k-means based on my understanding from class.

I concluded that association rule mining was unsuitable for my in-depth inquiry. Association rule mining focuses on recurring color combinations, useful for identifying co-occurrence patterns. However, Q3 requires measuring overall style structure and deviation across seasons. Since association rules describe local feature combinations rather than global style similarity or deviation, they are not suitable as the primary method for Q3. Therefore, I asked ChatGPT if other methods were appropriate, and it recommended distance-based analysis.

{ HYPERLINK "https://chatgpt.com/share/698aa007-bafc-8008-a34f-c1b222824656" }