BA820 – Project Proposal

Cover Page

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

1. Primary Dataset

1.1. Project Motivation and Preliminary Exploratory Data Analysis (EDA)

This project explores the Bob Ross paintings dataset. Our team is motivated to study this dataset because we are interested in the cultural impact and artistic legacy of Bob Ross, as well as the potential of quantitative analysis to reveal patterns and insights into his creative process. Understanding the common elements, color palettes, and structural similarities in his work is interesting and could yield significant benefits. Our potential stakeholders include:

- **Art critics**: To understand stylistic evolution and recurring themes.
- **Aspiring artists**: To learn common painting techniques and color combinations
- **Admirer of Bob Ross**: To gain a deeper appreciation for his work.
- **AI/ML researchers**: To train models on art generation or style transfer.

The dataset includes 403 paintings and 27 variables, with episode metadata and 18 paint indicators. Palette size is fairly concentrated (median num_colors = 11; IQR = 9–12) but shows clear extremes. For example, “Contemplative Lady” uses 1 color, while some paintings use a maximum of 15. Paint usage is highly uneven across the 18 indicators, with a small core set appearing in most paintings, especially Titanium White (0.9926), Alizarin Crimson (0.9429), and Van Dyke Brown (0.9206) (Figure A2). These core paints also co-occur frequently, consistent with a shared base palette (Figure A5).

Similarly, many paintings share a similar palette, as our Pairwise Jaccard similarity is relatively high, with a median above 0.64 and a large number of near-duplicate palette combinations. The most common exact paint recipe appears in dozens of paintings. Besides, our temporal exploration shows little change in average palette size across seasons, suggesting stylistic consistency over time rather than evolution.

Using a near-duplicate threshold of 0.95, we identified 2,330 highly similar pairs (Table A1), indicating substantial repetition or templating in palette composition. This tendency is also evident at the exact-recipe level, where the most frequent paint recipe appears in 46 paintings. In conclusion, the results suggest that palette choices follow a small set of recurring templates with limited variation, creating clear opportunities to define interpretable groupings and boundaries within an otherwise consistent instructional series.

1.2. Domain / Business Questions (Core Section)

Motivated by the observed imbalance, co-occurrence structure, and high palette similarity, this project proposes the following four domain/business questions.

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?
Our EDA shows strong color imbalance and frequent co-occurrence among a core set of paints, as well as meaningful variation in other colors. This suggests that paintings may cluster around a small number of palette templates, rather than being independent compositions. Art critics and

historians can interpret whether Bob Ross's work follows a small number of recurring stylistic patterns. Aspiring artists and AI researchers can use these potential structures as a foundation for learning, replication, or creative generation.

2. Do paintings that are highly similar in color composition form coherent groups over time, or are similar paintings evenly distributed across seasons?
The high Jaccard similarity and presence of many near-duplicate paint combinations motivate us to examine whether repeated palette structures are temporally localized or recur throughout the series. Identifying temporal grouping in similarity patterns can help explain whether Bob Ross reused instructional formats within localized periods.

3. Which paint colors contribute most to defining similarity between paintings, and which colors primarily introduce differentiation?
Our EDA suggests that some colors appear in almost all paintings, while others vary considerably. This implies that not all colors carry equal informational value when comparing paintings. Understanding which features drive similarity versus differentiation improves the representation and comparison of paintings. This could benefit researchers building representations of artistic style and practitioners designing simplified or focused learning tools.

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?
Our EDA identified paintings that use an unusually small or unusually large number of colors compared to the norm. These cases may represent intentional departures from Bob Ross's standard instructional approach rather than random variation. Identifying structurally atypical paintings clarifies the boundaries of Bob Ross's style and teaching methods and improves analytical understanding of variation and outliers within an otherwise consistent creative process.

2. Backup Dataset
2.1 Preliminary Exploratory Data Analysis (EDA)

This backup dataset examines domestic cat movement and behavior in the UK using GPS tracking data. Understanding cat movement is relevant for animal welfare, wildlife conservation, and urban planning, as outdoor cats interact with human infrastructure while also posing ecological and safety risks. So, our key stakeholders include:

- **Pet owners:** To understand their cats' outdoor routines and safety risks.
- **Conservation groups and wildlife biologists**: To monitor and reduce ecological impacts.
- **Veterinary professionals:** To provide informed advice on health and behavior.
- **Urban planners and local authorities**: To design shared human and animal spaces.
- **Animal behaviorists**: To study the drivers of feline movement and activity patterns.

We drew insights from one event-level movement dataset containing 18,215 GPS observations from 101 unique cats, and a reference dataset providing demographic and lifestyle attributes at the cat level. Our EDA shows that movement variables are highly skewed, with most

observations concentrated at low speeds and a long tail of extreme values (Figure B1). While cat-level average speeds are more stable, substantial variability across individuals suggests heterogeneous movement behaviors (Figure B2). The temporal analysis reveals clear daily activity patterns, with increased movement during specific hours (Figure B3). This finding is consistent with known feline activity rhythms.

The spatial analysis further highlights large differences in ranging behavior, with some cats covering much broader areas than others. At the cat level, maximum speed is strongly correlated with speed variability, suggesting that more active cats also move less consistently (Figure B5). Reproductive condition shows descriptive differences in movement behavior, but with substantial overlap between groups (Figure B4). Overall, movement varies widely across time, intensity, and space, and a handful of extreme values may reflect rare events or measurement noise, motivating the unsupervised questions that follow.

2.2. Domain / Business Questions (Core Section)

1. Can we identify cats with unusually high or low activity levels that may require special monitoring or intervention?

Our EDA revealed right-skewed distributions in movement, such as maximum speed and speed variability, with a small number of cats showing extreme values. Identifying these atypical patterns helps distinguish meaningful behavioral anomalies and support more targeted monitoring and intervention approaches for animal welfare and ecological research.

2. How do cats' activity levels vary across the day, and what does this imply for optimal outdoor access or care schedules?

Temporal EDA showed clear variation in movement intensity across hours of the day, indicating structured daily activity circles rather than uniform behavior. Understanding these differences could provide guidance for owners regarding outdoor access timing and help researchers study how cats interact with human schedules and urban environments.

3. Do cats exhibit routine-driven versus more irregular movement patterns over time?

Cat-level summaries represent big differences in speed variability and consistency of movement behavior across cats. Distinguishing between routine-driven and irregular movement patterns may provide insight into lifestyle differences, environmental exposure, or potential safety risks, which are relevant and beneficial for both animal welfare and behavioral research.

4. Does reproduction condition meaningfully relate to cats' movement behavior, and what does this imply about using biological status for behavioral segmentation?

Our EDA showed descriptive differences in movement behavior across reproductive conditions, such as spayed, neutered, and not fixed, but with substantial overlap between groups, suggesting that reproductive condition alone does not fully explain movement behavior. This motivates further unsupervised analysis to assess whether movement-based behavioral patterns align with reproductive conditions and helps stakeholders interpret movement data more accurately without over-relying on biological labels.

Appendix

GitHub Repository (Required)

link to project's GitHub repository:

https://github.com/yujiunzou/BA820-Unsupervised-ML-Project.git

Supplementary Material - Primary Dataset:

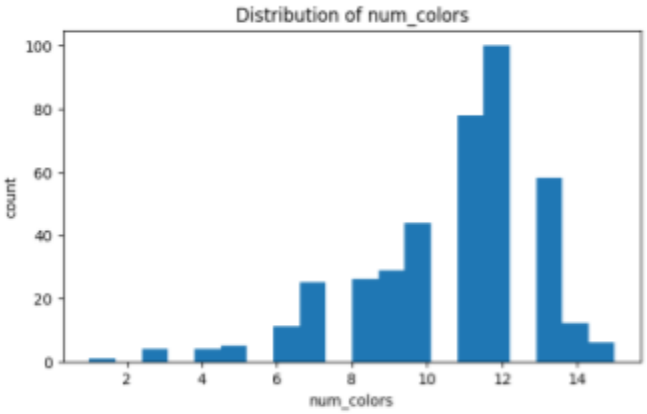

Figure A1. Distribution of number of colors used per painting

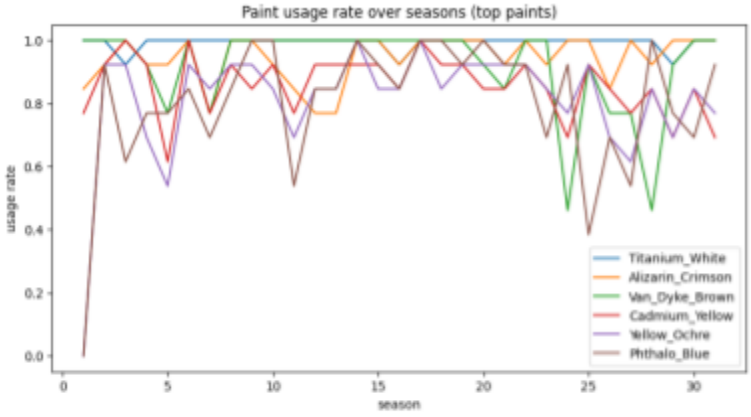

Figure A2. Top paint usage rates

Top 20 most similar pairs:

1 to 20 of 20 entries Filter

index	i	j	jaccard_similarity
70531	257	371	1.0
70528	257	368	1.0
70664	258	360	1.0
70516	257	356	1.0
70513	257	353	1.0
70511	257	351	1.0
70496	257	336	1.0
70492	257	332	1.0
70487	257	327	1.0
70479	257	319	1.0
34957	99	110	1.0
34954	99	107	1.0
34984	99	137	1.0
35007	99	160	1.0
35001	99	154	1.0
35056	99	209	1.0
35072	99	225	1.0
35069	99	222	1.0
35067	99	220	1.0
35064	99	217	1.0

Show 25 per page

Number of pairs with similarity >= 0.95: 2330

Table A1. Near-duplicate Pairs Counts

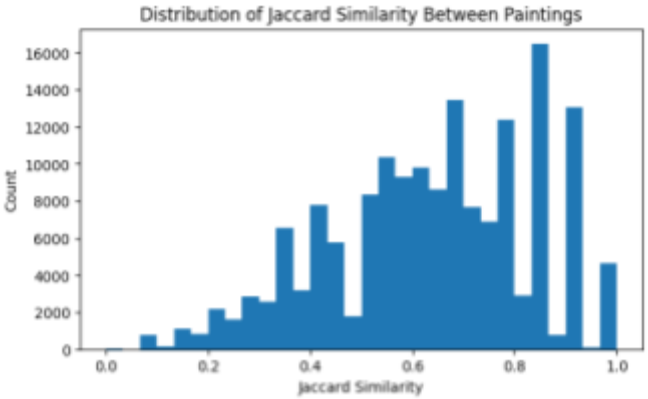

Figure A3. Distribution of Jaccard similarity between paintings

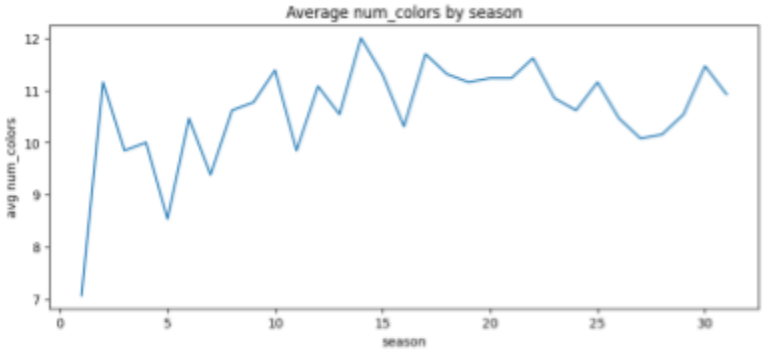

Figure A4. Average number of colors by season

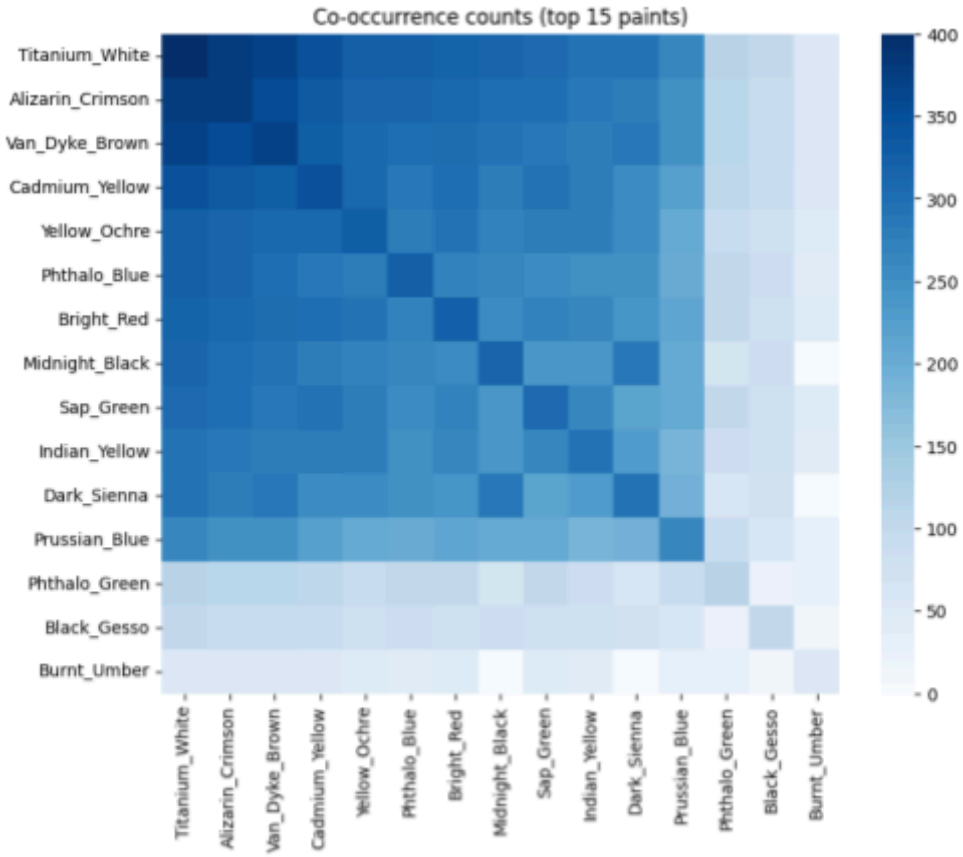

Figure A5. Paint co-occurrence heatmap (top 15 paints)

Supplementary Material - Backup Dataset:

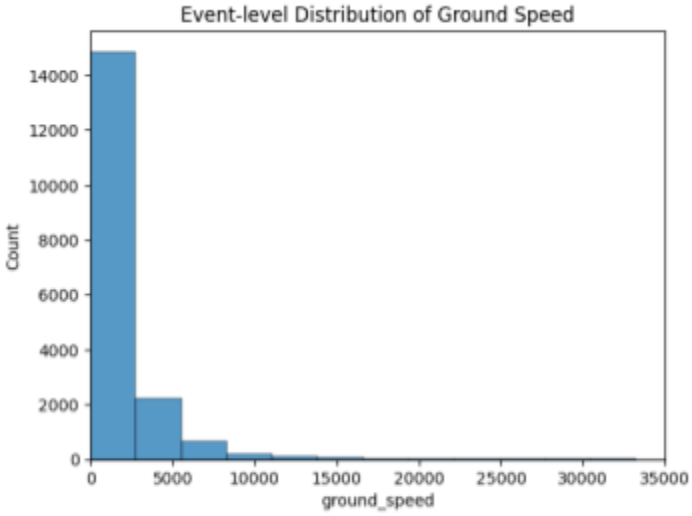

Figure B1:Distribution of event-level ground speed

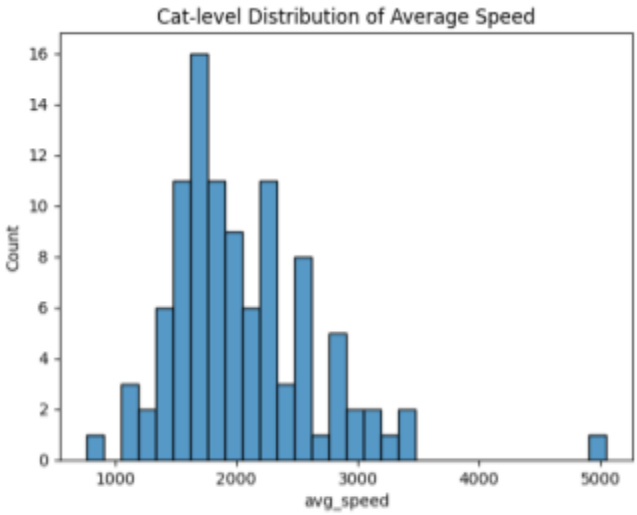

Figure B2: Distribution of cat-level average speed

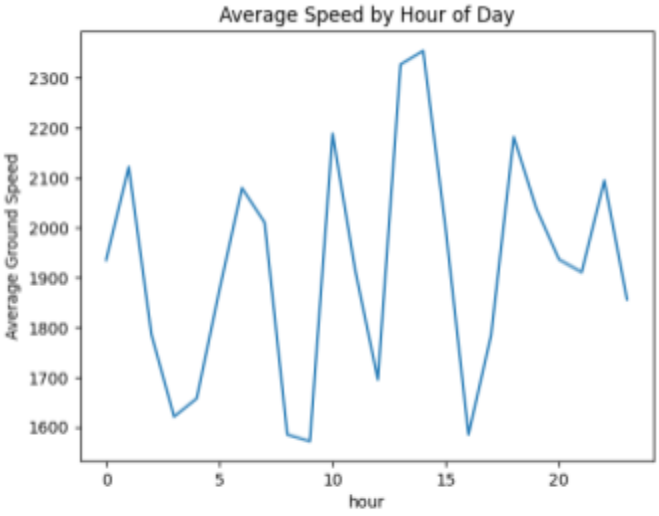

Figure B3:Average movement speed by hour of day

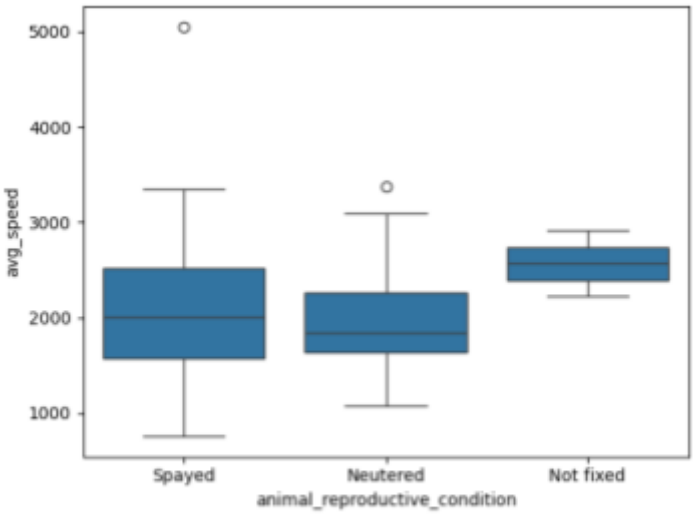

Figure B4:Average speed by reproductive condition

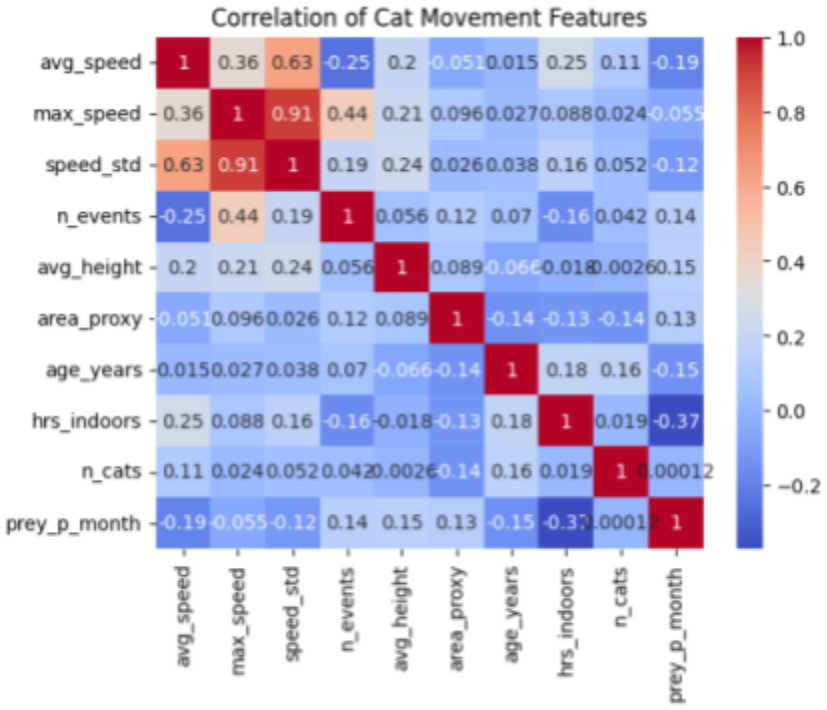

Figure B5: Correlation heatmap of cat-level movement features

Use of Generative AI Tools

Generative AI tools were used to assist with outlining and refining written sections of the proposal. All analytical decisions, exploratory analyses, and interpretations of results were conducted by the project team. AI-generated content was reviewed, edited, and integrated to reflect the team's own understanding and judgment.