The Colors of College Basketball

Isaac Schmidt ISCHMIDT20 @ BERKELEY.EDU
Department of Statistics

Abstract

In this paper, the colors of NCAA Division I basketball programs are clustered in order to accurately approximate their wide variability in precise RGB specifications with just a small set of colors. This analysis may be useful in applications where team's colors need to be accurately represented but the entire color gamut is not available. In other words, college basketball clusters are quantized.

1. Introduction

Sports teams use many different colors in their logos, uniforms, and other branding. The advent of computer graphics has allowed teams to specify precise color codes using the RGB format, such as `#FFC72C`, as opposed to more generic expressions like "gold." The RGB color gamut contains $256^3 = 16,777,216$ colors, allowing computers to discern the tiny differences among tints and shades. However, such a fine range is not practical for many applications, such as those that require physical manifestations of colors—i.e. clothing. Therefore, in the context of sports, it is useful to reduce the entire gamut to a small palette, such that every color of every team can be accurately approximated by a member of the palette.

For the purposes of this analysis, college basketball programs were chosen, as opposed to professional teams in other sports or other countries. This is in part due to the large number of programs—there are currently 357 in NCAA Division I. Other top-level leagues and organizations have a much smaller number, generally no more than 32. Many "minor leagues," such as those in baseball or in lower-tier club football divisions, also have a large number of teams, but they lack sufficient data on the colors and brandings of the teams involved. Additionally, some minor teams may have intentionally unique color schemes as a form of self-promotion, so the colors of these teams may not be representative of sports teams in general. College basketball is a happy medium between the competing characteristics of a large population, and data availability and consistency.

As colors are represented as points in a feature space, it is natural to cluster them, such that colors that are close together—that is, colors that look similar—are placed in the same cluster. For each cluster, its center is itself a color, and that color can be used as a close approximation for all other colors within its cluster. Thus the set of all possible team colors can be reduced to the cluster centers.

2. Data

2.1 Data Collection and Cleaning

Official colors for all 357 NCAA Division I basketball programs were obtained from Wikipedia [1]. As the only official sources for program colors are the schools themselves, there did not exist any other complete listing of colors, to the best of the author's knowledge. The raw dataset contained color palettes for 830 schools, including all 357 Division I programs. Data for all non-Division I schools was discarded.

The record for each school included anywhere from 2 to 5 colors, all in hexadecimal format. Each color is denoted by a string of seven characters. The first character is a `#`, which is followed by three pairs, representing the red, green, and blue components of the color, respectively. Each component is in hexadecimal format, ranging from `00` = 0 to `ff` = 255. These strings were then converted to tuples of integers, such that `#FFC72C` becomes `(255, 199, 44)`.

The next step involved converting each color from the RGB color space to the CIELAB color space, with illuminant D-65. This was done because CIELAB is close to "perceptually uniform," whereas RGB is not. With a perceptually uniform color space, a given numerical change in the color's components generally leads to the same change in human perception, no matter the original color. Such color spaces give proper meaning to the "distance" between colors. CIELAB, like RGB, represents a color with three quantities: **L** representing lightness, **a** representing a green-red scale, and **b** representing a blue-yellow scale. **L** is bounded between 0 and 100, and **a** and **b** are unbounded, but both typically range between between -100 and 100. The conversion was done using the `scikit-image` module in Python [2] [3].

Finally, pure black (RGB: `(0, 0, 0)`, Lab: `(0, 0, 0)`) and pure white (RGB: `(255, 255, 255)`, Lab: `(100, 0, 0)`) were removed from the dataset. All schools had at least one of these two colors listed, and many schools incorporate both in their color schemes. The intended focus is on the "colors" of the teams, not the prominence of black or white. After all processing, every school has between 1 and 3 colors, totaling 634 across 357 teams. Of these colors, 453 are unique.

2.2 Exploratory Data Analysis

Figure 1 shows scatterplots for each pairwise combination of **L**, **a**, and **b**, with each point representing a color. Visually similar colors appear to be close together for each combination of variables, although not entirely. For example, note the greens interspersed among the reds in the second plot. The feature space contains both dense and empty regions, indicating the presence of clusters. Precisely how many clusters are appropriate is not immediately obvious from the scatterplots.

3. Methods

3.1 Clustering Algorithm

We want colors that are "close" to each other to be in the same cluster, so our clustering algorithm needs to leverage distance. The natural choice would be k-means, as it enforces

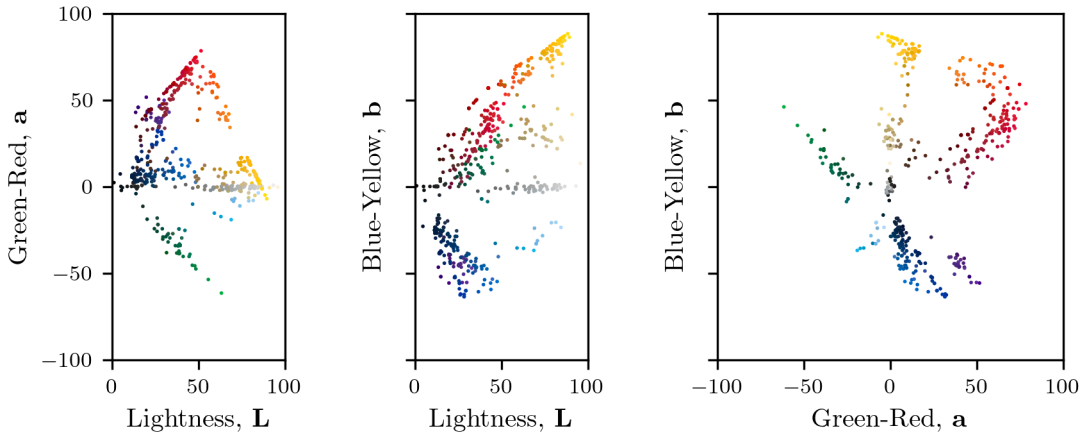

Figure 1: Scatterplot projections of the colors in CIELAB space.

that points get assigned to their closest cluster center. However, k-means relies on Euclidean distance, which is *not* the best metric for color difference, even though the CIELAB color space is roughly perceptually uniform. Instead of Euclidean distance, CIELAB has their own metrics, collectively named ΔE^*. This paper uses the most recent version, called CIEDE2000, with settings tuned for "imperceptibility" [4]. K-means does not allow for the use of custom distance metrics, but it has a natural extension that does, called k-medoids. K-medoids and k-means are very similar, in that both attempt to minimize the total (squared) distance between points and their cluster centers. The main difference is that k-medoids requires the "centers" to be actual data points, which allows a distance matrix to be precomputed, as no other distances will need to be calculated other than those between the data points themselves. This matrix, using the CIEDE2000 metric, was calculated with `scikit-image` and `scikit-learn`, and the actual clustering algorithm was implemented using the `cluster.KMedoids` method of the `scikit-learn-extra` module [5] [6].

3.2 Choosing k

One problem with k-medoids is that it requires k to be preselected. One popular method of choosing k is to use silhouette scores [7]. The larger a silhouette score for a given data point, the closer it is to points in its own cluster compared to points its closest neighboring cluster. The average silhouette score across all points in the dataset is a good indication of whether or not the chosen k is appropriate. To determine k, the procedure described below was run on 18 candidate values—all of the integers ranging from 3 to 20, inclusive.

K-medoids, like k-means, is highly susceptible to the randomness of the cluster initialization, so two mitigative measures were implemented. The first is that k-medoids++ was used to initialize the cluster centers. K-medoids++ is a modified implementation of k-means++, an algorithm for initializing cluster centers such that they tend to start far apart from each other. This reduces the chance that cluster centers start close together, which

often leads to "bad" clusterings. Secondly, for each candidate k, 10,000 clusterings were created, each using a seed chosen at random from the set of all 32-bit integers. Only the silhouette score of the clustering with the minimum inertia (across all 10,000 clusterings) was recorded. The same set of seeds was used for all candidate k. The results are shown below.

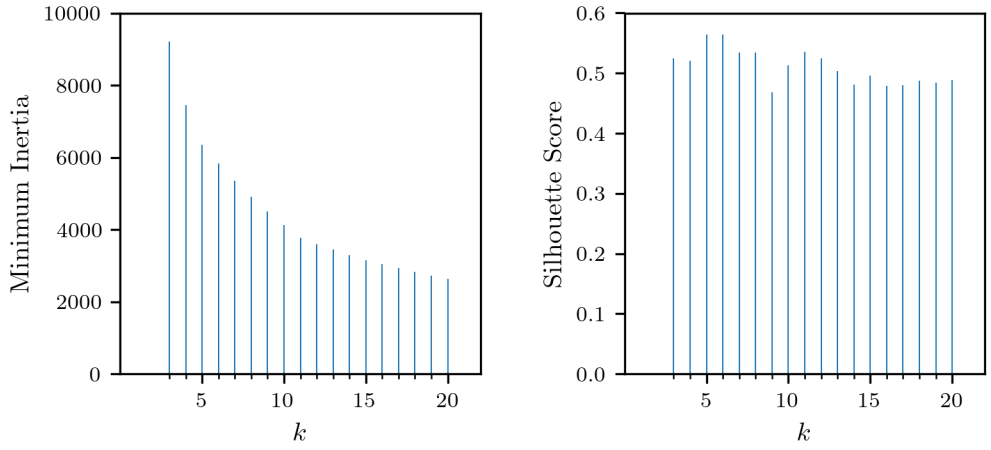

Figure 2: Inertia and silhouette score of optimal clustering for candidate k.

As expected, the plot of inertia is monotonically decreasing, because as the number of clusters grows, points will generally be closer to their cluster's medoid. More interestingly, the plot of silhouette scores is bimodal. There is one peak at $k = 6$, and another at $k = 11$, and these two values also correspond to kinks in the inertia plot. Both observations indicate that 6 and 11 are probably good choices for k.

Note that the selection process only considered the best possible clustering for each k, in terms of inertia. While this decision would be seen as "overfitting" in many other cases, the dataset is considered to be a population, not a sample of all possible team colors. Thus no statistical claims are made about the applicability of the resulting clusters to team colors in general, although we may consider them informally.

4. Clusters

Two clusterings are created, one with $k = 6$, and one with $k = 11$.

4.1 $k = 6$

The seed that provided the best clustering for $k = 6$ was 177264706. The 6 medoids returned by the algorithm are:

Cluster	RGB	Hex	Lab	Cluster Size
1	(6, 49, 91)	#06315B	(20.0, 3.9, -29.1)	201
2	(182, 30, 46)	#B61E2E	(39.7, 58.5, 31.5)	143
3	(241, 184, 45)	#F1B82D	(78.0, 8.8, 72.1)	125
4	(177, 179, 179)	#B1B3B3	(72.8, -0.7, -0.2)	81
5	(0, 103, 71)	#006747	(38.1, -34.3, 10.9)	45
6	(239, 91, 12)	#EF5B0C	(57.9, 54.3, 65.3)	39

Table 1: Colors for the optimal clustering with $k = 6$. Lab values rounded to one decimal place.

4.2 $k = 11$

The seed that provided the best clustering for $k = 11$ was 106384465. The 11 medoids returned by the algorithm are:

Cluster	RGB	Hex	Lab	Cluster Size
1	(0, 40, 85)	#002855	(16.4, 7.1, -30.9)	99
2	(254, 195, 37)	#FEC325	(82.0, 8.4, 78.1)	87
3	(200, 16, 46)	#C8102E	(42.5, 65.9, 35.7)	86
4	(174, 178, 181)	#AEB2B5	(72.3, -0.8, -2.0)	75
5	(132, 26, 43)	#841A2B	(29.0, 44.8, 18.0)	53
6	(183, 163, 105)	#B7A369	(67.5, -0.9, 32.8)	50
7	(0, 77, 159)	#004D9F	(33.6, 13.4, -50.0)	46
8	(0, 103, 71)	#006747	(38.1, -34.3, 10.9)	43
9	(239, 91, 12)	#EF5B0C	(57.9, 54.3, 65.3)	34
10	(34, 34, 34)	#222222	(13.2, 0.0, 0.0)	31
11	(79, 37, 130)	#4F2582	(25.6, 39.2, -45.0)	30

Table 2: Colors for the optimal clustering with $k = 11$. Lab values rounded to one decimal place.

5. Discussion

The two clusterings reveal the most prevalent colors among college basketball teams. The medoid of the largest cluster, in both cases, is navy blue. This makes sense, as it is a popular alternative to black for many teams. Even in the finer clustering, with lighter blues shunted off into their own cluster, the navy blue cluster is still the largest. One notable absence from both lists is sky blue, such as #7BAFD4. This color is commonly associated with college basketball as it is worn by one of the sport's premier programs, but apparently it was not popular or distinct enough to earn its own cluster. Further analysis might give larger weighting to teams that are more prominent than others.

There are two medoids that appeared exactly the same in both clusters—#006747 and #EF5B0C. Another common color was red, which split into plain red and a darker crimson

in the finer clustering. Gray was also popular, as many teams use this in place of black or white, or as a neutral third color. A surprising inclusion is #B7A369, a shade of tan. Finally, the last two colors in the finer clustering were off-black, and purple.

6. Other Results

The details in this section primarily refer to the 11-medoid clustering.

6.1 Color Schemes

Of the 357 schools in the dataset, 94 used one color, 249 used two colors, and 14 used three colors, not including black and white.

The most common color scheme across all schools was just red, #C8102E (in combination with either black or white), which belonged to more than 30 schools. The next three highest schemes all involved navy blue, #002755, with gold (#FEC325), gray (#AEB2B5), and red as the counterparts. Only two schemes with three colors were used by multiple schools, which were navy and gray with either red or gold. One of the schools with the latter scheme, however, had a third color that was clearly light blue, which was reduced to gray as a result of the clustering. Only one school, Providence, had a scheme with gray as the only color, and zero schools used only orange.

In terms of color difference, the most contrasting two-color scheme was navy-gold, which 18 schools used. Closely behind was purple (#4F2582) and gold, used by 8 schools. Among all of the schemes used by any team, the one with the least contrast was navy and purple, used by two schools. In the 11-cluster case, one school actually used two different colors that both fell within the same cluster. Old Dominion, in addition to its primary navy blue, used a gray and a light blue, which both fell into the gray cluster.

Schools tended to prefer darker colors as their primary colors. For example, navy was used as the primary color in 85 schools, about 86 percent of its overall usage, but gold was only used as the primary color by 11 schools, which was the third-lowest total and second-lowest percentage of usage. Crimson, #841A2B, was the most likely to be a primary color given that a school used it—only one out of the 53 schools with crimson used it as a secondary color.

6.2 Distant Colors

Even though the algorithm assigns each cluster to its closest medoid, this clustering method can be prone to "mistakes," e.g. light blues reducing to grays.

The most unique color, which is defined as the color with the largest distance to its closest neighbor, is #FDF2D9, of Nebraska. A distance of more than 9.7 away from anything else, this color was reduced to gray. The color farthest away from its medoid was the #00B141 of Marshall, which was a distance of over 27. Tan was its closest medoid as determined by CIELAB2000, although a human might consider this color's "proper" cluster to be green.

7. Conclusion

This document presents an example of k-medoids clustering applied on a real world dataset, for the purposes of quantization. However, the results provided here are not perfect, as many colors were assigned to a cluster whose center was not all that perceptually similar. The use of k-medoids++ encouraged clusters to be far apart from each other, but perhaps there are better clusterings with the groups closer together. Additionally, the CIELAB color space and the CIEDE2000 distance metric may not be perfect representations of what humans actually see.

Despite these flaws, there are still practical uses for this analysis. For example, an intramural youth basketball program, that wishes to have the color schemes of its teams closely represent those of a "real-world" league, can use the colors returned by this clustering algorithm.

References

[1] "Module:College color/data," Dec. 2020. Page Version ID: 996787440.

[2] "scikit-image: Image processing in Python."

[3] S. v. d. Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "scikit-image: image processing in Python," *PeerJ*, vol. 2, p. e453, June 2014. Publisher: PeerJ Inc.

[4] I. C. on Illumination and C. T. C. 1-47, eds., *Improvement to industrial colour-difference evaluation*. No. CIE 142-2001 in Technical report, Vienna, Austria: CIE Central Bureau, 2001. OCLC: ocn223775480.

[5] "scikit-learn: A set of python modules for machine learning and data mining."

[6] "scikit-learn-extra: A set of tools for scikit-learn.."

[7] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987.