Marilyn Monroe Image Analysis

Jalen Lee (918243072), Ryandeep Chawla (917335017) , Min Seo Park (916949129), Madelyn Joe (918245122)
STA160 Practice in Statistical Data Science

Abstract

The iconic images of Marilyn Monroe have been a subject of fascination for artists and art enthusiasts worldwide. Among those captivated by her timeless beauty was the renowned pop artist, Andy Warhol. In this data analysis project, we delve into the realm of color to uncover the essence of Warhol's original ideas as reflected in 5 different colored Marilyn Monroe pictures. To accomplish this, we begin by acquiring and curating a diverse dataset consisting of five image collections, each presenting Marilyn Monroe portraits with distinct color schemes. Employing image processing techniques, we extract the major colors present in each image data, enabling us to discern the dominant hues that compose these works of art.

Introduction

With the extracted color information at our disposal, we embark on a comprehensive analysis of the 5 images, meticulously comparing their color compositions. Through visual representations and statistical analysis, we unravel the similarities and differences in color palettes across the artworks. This examination sheds light on the artistic choices made by Warhol, hinting at his intentions and preferences regarding color usage. Furthermore, we engage in a thought-provoking discussion about the potential original ideas that may have guided Warhol's artistic vision when incorporating these colors into his Marilyn Monroe portraits. We explore theories and historical context surrounding Warhol's work, drawing insights from his broader body of art and cultural influences. By delving into the realm of color and meticulously examining the compositions of these iconic Marilyn Monroe images, we gain a deeper understanding of Warhol's artistic intentions and the significance of color in his works. This project serves as a tribute to Warhol's artistic legacy, offering new perspectives on his exploration of color and its role in shaping his artistic expression.

K-Means Clustering

K-means clustering is a popular unsupervised machine learning algorithm used for grouping similar data points into clusters based on distance. It aims to partition a dataset into a predetermined number of clusters, where each data point belongs to the cluster with the nearest mean (centroid). K-means clustering is an iterative algorithm that aims to minimize the within-cluster sum of squares, or inertia. It seeks to find the optimal positions for the centroids that minimize the total distance between data points and their assigned centroids. However, since the algorithm is sensitive to the initial centroid positions, multiple runs with different initializations are often performed, and the solution with the lowest inertia is selected. Additionally, the choice of K, the number of clusters, impacts the quality of segmentation or color quantization. Finding an optimal value for K can be challenging and may require experimentation or evaluation metrics specific to the task.

Image Segmentation: K-means clustering can be used to partition an image into distinct regions or segments based on color similarity. Each pixel in the image is considered a data point, and the RGB or

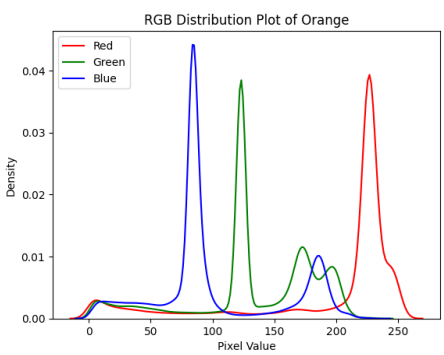
HSV color values of the pixels are used as features for clustering. By applying K-means clustering with K clusters, each pixel is assigned to one of the clusters based on the similarity of its color to the cluster centroids. This results in segmenting the image into regions that share similar colors.

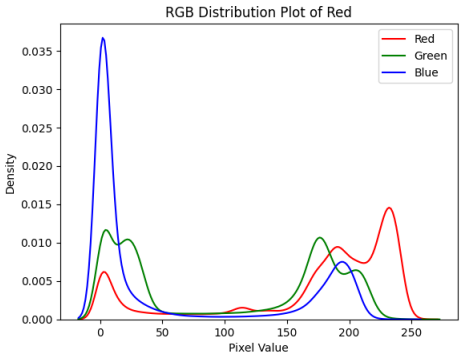
Color Quantization: K-means clustering can also be utilized to reduce the number of colors in an image while preserving its visual appearance. In this case, the RGB or HSV color values of all pixels in the image are flattened into a single feature vector. K-means clustering is then applied to cluster these color vectors into K clusters, where K represents the desired number of colors in the output image. The centroid of each cluster represents a representative color, and all pixels assigned to that cluster are replaced with the representative color. This process allows for reducing the color palette of the image while minimizing the perceptual loss.

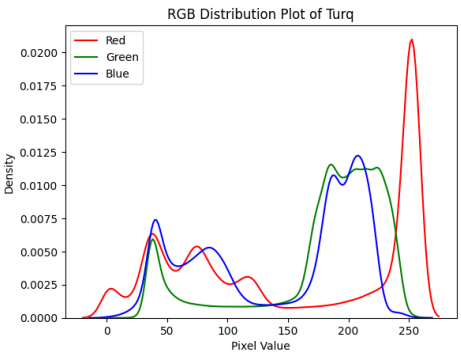
Color Distributions

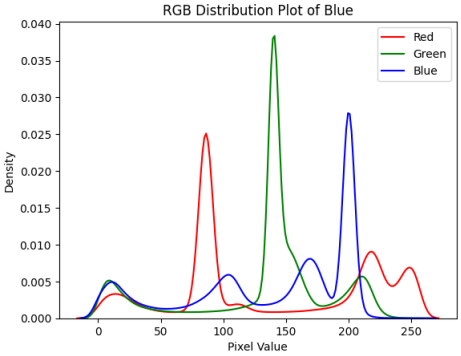
The RGB distribution plots provide valuable insights into the color compositions of the Marilyn Monroe images. By analyzing the distribution of RGB color channels, we can gain a deeper understanding of the dominant colors present in the images and their relative intensities. By analyzing the RGB distribution plots of the Marilyn Monroe images, we can unravel the color compositions and gain insights into the intentional use of color by the artist. This understanding enhances our appreciation of the artworks, revealing Andy Warhol's original ideas on colors and how he manipulated color to evoke specific emotions or convey his artistic vision. RGB distribution plots serve as a valuable tool for interpreting the color characteristics and compositions of the Marilyn Monroe images. Through these plots, we can decipher color dominance, balance, palette, and subtle variations, all of which contribute to the visual impact and convey the intended artistic expression.

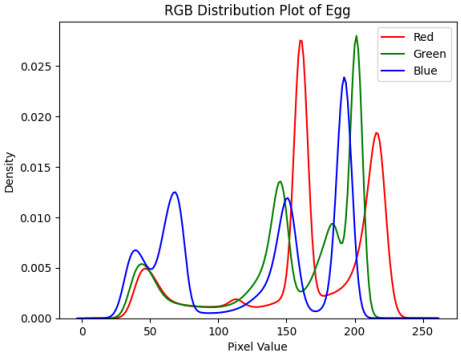
In the red and egg plots, we can see more of the RGB peaks overlapping which entails specific color combinations are prevalent and that colors are condensed into a reduced amount of clusters. In the egg and turq plots, we see an imbalance in color distribution since most of the plots are densely populated in the 150-250 pixel value range meaning it’s a brighter or intense region because of the higher values. We can also see that for the red plot, most of the plot is populated in the 0-50 pixel value range meaning it’s a lighter or dull region. Orange and blue plots have a decently uniform distribution across the entire color space, suggesting a well-balanced representation of colors in the image. Higher peaks suggest a higher density of pixels with that specific intensity, which helps identify the dominant colors in the image. Red displays more dominant colors with blue intensities and surprisingly Turq displays many red intensities.


RGB Distribution Plot of Orange
Red
Green
Blue
Density
Pixel Value


RGB Distribution Plot of Red
Red
Green
Blue
Density
Pixel Value


RGB Distribution Plot of Turq
Red
Green
Blue
Density
Pixel Value


RGB Distribution Plot of Blue
Red
Green
Blue
Density
Pixel Value


RGB Distribution Plot of Egg
Red
Green
Blue
Density
Pixel Value

Relative Conditional Entropy

As a way to analyze the associations between the colors of the paintings, we used the relative entropy between the red, green, and blue values within the RGB values of each pixel in the five images of Marilyn Monroe painting.

The relative entropy was computed using the function scipy.stats.entropy, with which the output is also known as the Kullback-Leibler divergence. This Kullback-Leiber divergence measures how much the first probability distribution is different from the second probability distribution, and it is calculated using the following equation:

$$D = \sum(p * log(p/q))$$

where p and q are the probability distributions. The outcome D is a non-negative number that can range from zero to infinity. This relative entropy is zero when the two distributions are identical, meaning the entropies for (red,red), (green,green), and (blue,blue) will always equal 0. Any non-zero entropy means there are some differences, or divergence, between the two probability distributions, and value closer to zero means this divergence is small, while higher value for this entropy means the divergence is big.

Methodology:

When calculating the relative entropy between two colors of each image, the following methods were used:

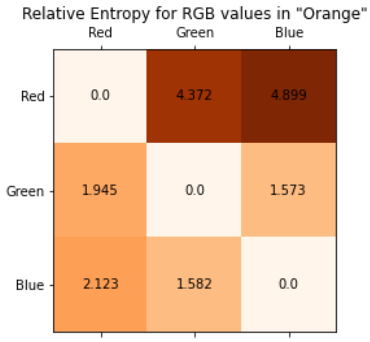
1. Record every RGB values of every pixel in the 750 x 750 pixel image (total 562,500 RGB values)
2. Separate all RGB values into three lists each containing the R, G, and B values of all pixels
3. Divide the full range of RGB values (0 to 255) into 16 bins (bin #1 with range 0 to 15, #2 with 16 to 31, until the last bin #16 with the range between 240 to 255)
4. For all three lists created in step 2, convert the values of RGB values into the bins accordingly (e.g. if the first element in the list of red values is between 0 and 15, then convert that number into the number of the bin with corresponding range, which is #1 in this case)
5. Count how many elements belong to each bins, and make this count into a list of 16 (the sum of all counts between 16 bins should add up to 562,500)
6. Find the relative entropy between two of these lists (length 16) using the function scipy.stats.entropy (e.g. relative entropy of Red values given Green values in “Orange” image is equal to 4.372)
7. Calculate the relative entropy between all combinations of red, green, and blue in all five images and make it as a matrix

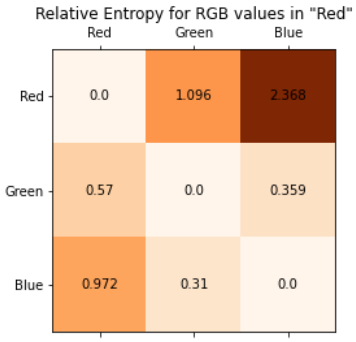
Interpretation:

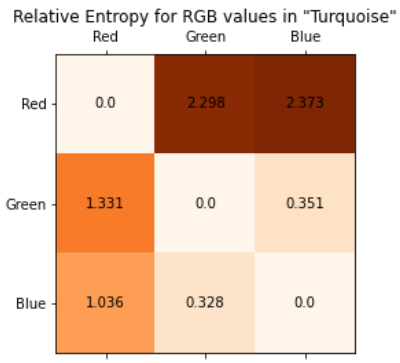
In the “Orange" image of Marilyn Monroe, the relative entropy of red given green, or D(R|G) is 4.372, while D(R|B) is 4.899. This means that when given the distribution of color red, it provides less information on both the distribution of green and blue colors; this is also true vice versa. In comparison, there are less ‘surprise’ when predicting the distribution of green given blue, or blue given green, since its relative entropy D(G|B) is 1.573, and D(B|G) is 1.582. This trend of color green and blue containing more information on the distribution of each other is apparent in all five images.

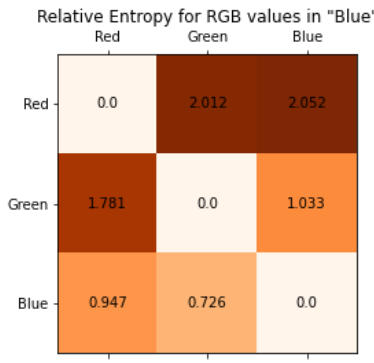
This could be due to the colors that Andy Warhol dominantly used in each of his paintings. The color red and orange, colors used in his “Red” and “Orange” image of Marilyn Monroe, has a high R value and relatively low G and B values. Therefore, the distribution of red is left-skewed in these images, while the distributions of blue and green are right-skewed. These differences in distributions could be the cause of low relative entropy between green and blue components in RGB values, and high relative entropy between red and green, as well as red and blue.

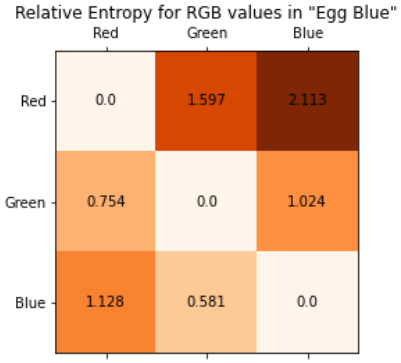
On the other hand, the “Turquoise,” “Blue,” and “Egg Blue” images use colors such as blue, light blue, or cyan. Since these colors have low values in ‘red’ components of RGB and high ‘green’ and ‘blue’ components of RGB, the distribution is the opposite. There are more values towards the first few bins for red, meaning it is right-skewed, while there are more values towards the last few bins for green and blue, therefore the distributions of ‘green’ and ‘blue’ are left skewed. This also results in similar results, where the relative entropy between green and blue are low since they have similar distributions, while the entropy between red & green or red & blue are high due to high divergence in their distributions.



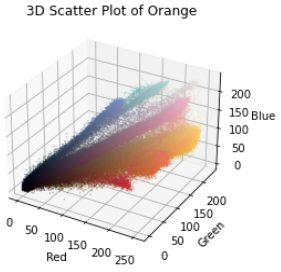


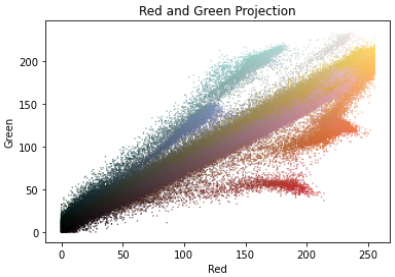


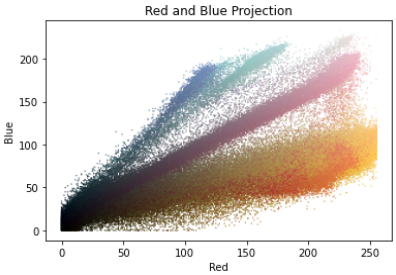


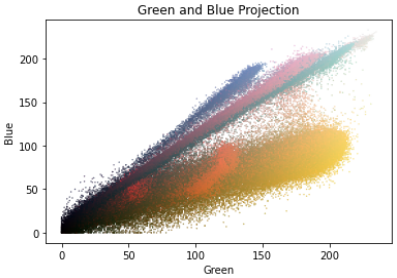


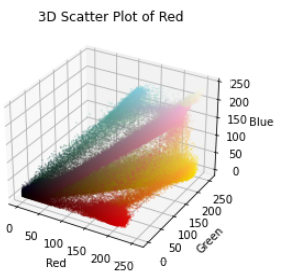
3D Scatterplots

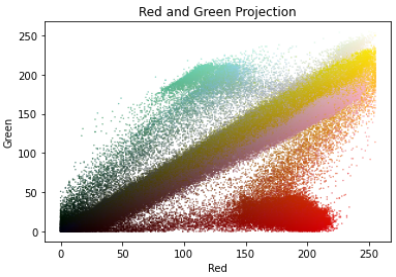


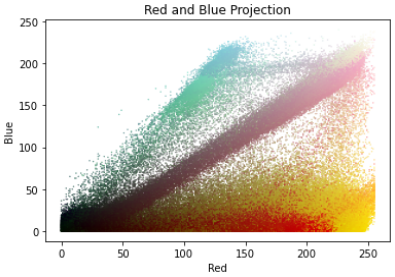


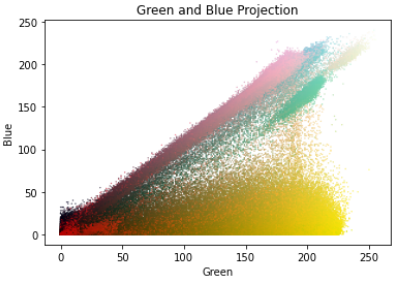


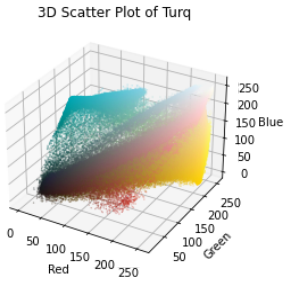


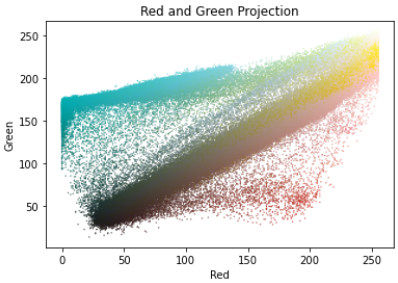


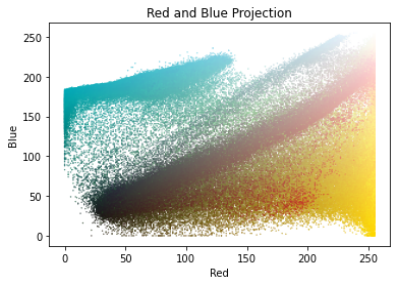


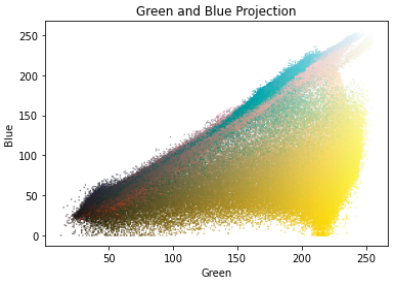


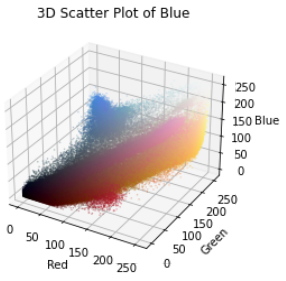


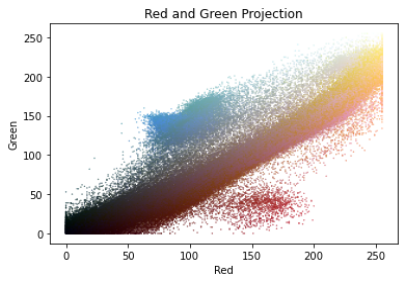


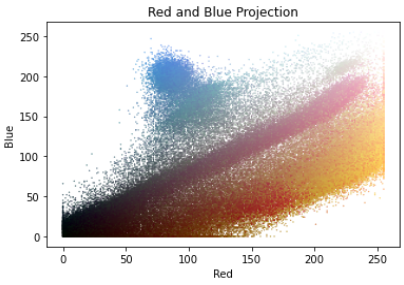


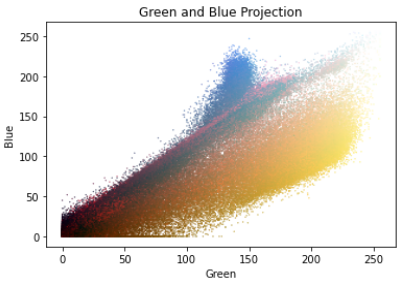


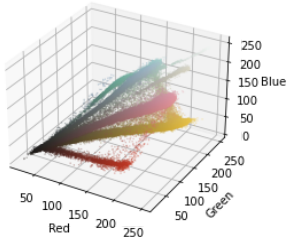


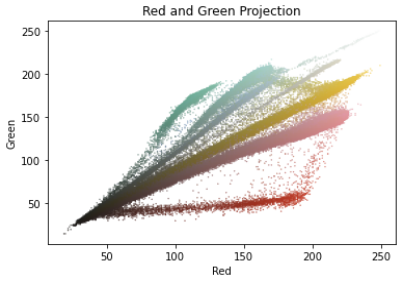


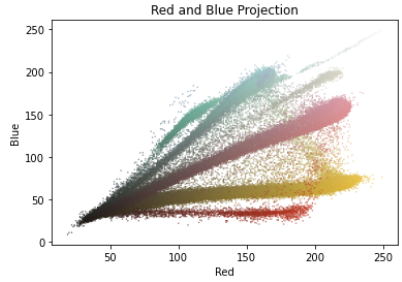


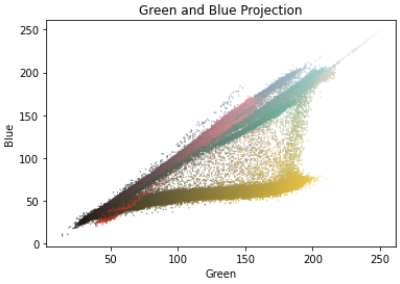












A 3D RGB scatterplot can provide insights into the distribution and composition of colors in an image. The plot can give you an overview of how colors are distributed in the image like the concentration and spread of different RGB values in the plot, and can assist in identifying the overall color palette of the image. In the projection plots, we can observe the five main background colors as they hold a high density in those specific regions. These colors contribute significantly to the overall composition of the images. Additionally, distinct patterns or clusters are visible in the projection plots, indicating the dominance of certain colors or color combinations in each image. For instance, in the red and egg plots, we can observe thin and concentrated color clusters. This suggests that there will be fewer overall clusters in the image, as the distances between the color clusters are relatively larger. Many points are grouped together or overlap into a main color cluster, indicating a stronger presence of that particular color. Conversely, in the orange, turquoise, and blue plots, we observe a more dispersed arrangement of points on the projection graphs, implying a higher number of clusters. This dispersion makes it more challenging to group these images into a reduced number of clusters, resulting in a larger number of distinct color clusters.

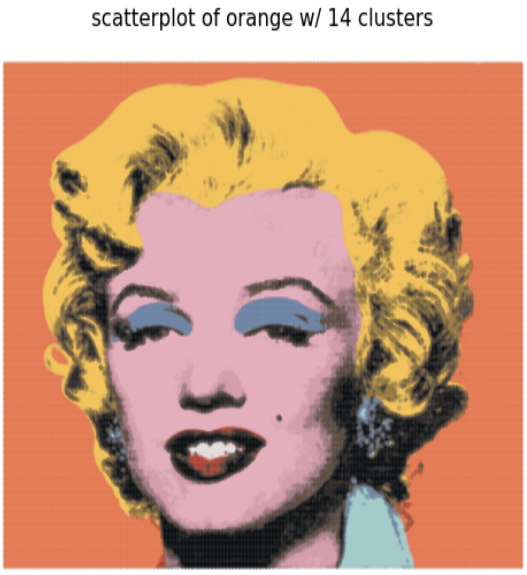
Painting Marilyn Monroe with K-means

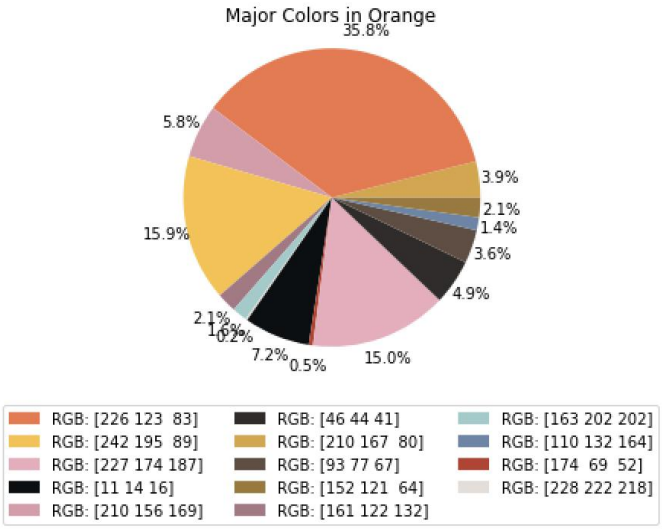
Kmeans is a clustering algorithm that groups points based on proximity. The Kmeans group can be random at the start of the algorithm since points are chosen randomly. However, we can always get the same results if we initialize our points so we leave no room for randomness. Additionally, by choosing the starting points we can pick colors that we might want to see at the end of the clustering process. This is especially useful for parts of Marilyn that take up a very small portion of the painting such as her teeth and her lips. By including a starting value that is close to white we can sway the chances of the final result having a center that includes her teeth. We tried to reduce the number of clusters as much as possible while trying to keep the color of the important features as close to the original image as possible. Some of these features included the background, face, hair, eyeshadow, collar, lips, and teeth. Some images contain fewer clusters than others, because the stopping point was chosen when reducing the clusters further will lead to compromising the colors of the important features. Our strategy to paint Marilyn based on these 5 images and their clusters was to use scatterplots with miniscule sized points and color these points based on their centroid value. We clustered the images solely on color. This means that the centroids will be the color that minimizes the distance to the other colors within the cluster. In other words, these following images of Marilyn are a 750 x 750 scatterplot colored by their cluster labels. For the color values of K-means, the RGB values were binned into 32 groups, each bin would range with eight numbers and the average of these ranges were the new values.

Orange Image of Marilyn Monroe

Original Image:

Scatterplot of Marilyn Monroe with clusters





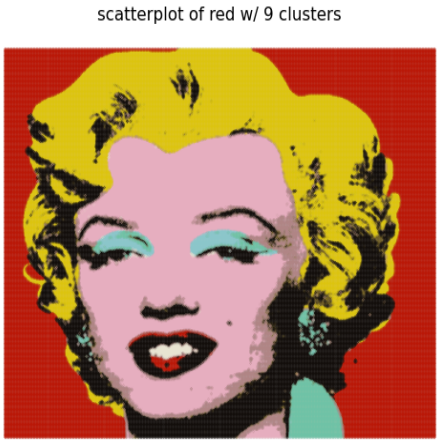
Pie chart of major colors based on K-means cluster with 14 clusters

For the orange image we are able to get the image with all the requirements back with 14 clusters. We see that RGB: [226 123 83] has the highest percentage at 35.8 % which makes sense because it is the background color of the image. Following that we see that the hair color which is made of three different colors, RBG: [242 195 89], RBG: [210 167 80], RBG: [152 121 64] has combined share of the second highest 21.9% and then there is the face color which is made with a shading of 3 different shades where majority of the face is shaded with, RGB: [227 174 187], then RGB:[210 156 169] and RGB:[161 122 132] combined share of 22.9%. The lips and teeth only comprise 0.7% together with RGB: [174 69 52] and RGB: [228 222 218]. The eye-shadow of Marilyn Monroe is 1.6% with RGB: [163 202 202] and the collar with the color RGB:[110 132 164] makes 1.4% of the image. We also see that the shading (RGB: [11 14 16] mixed with RGB: [46 44 41] and RGB: [93 77 67]) takes almost 15.7% combined which is quite a significant share of the color. Below is the segmentation of the different aspects of the image.

Red Image of Marilyn Monroe

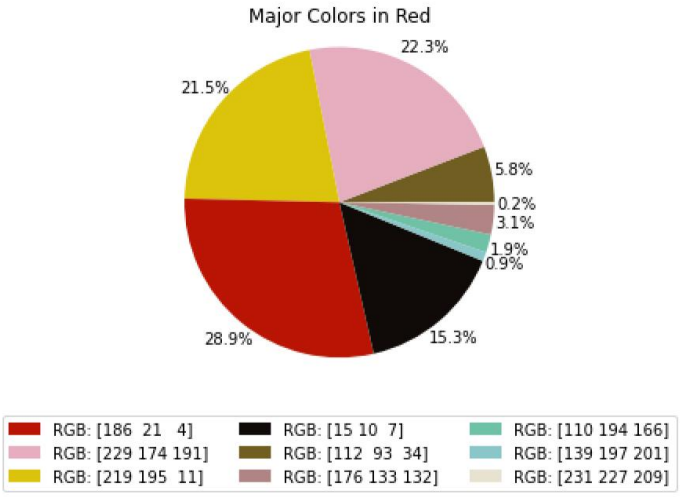
Original Image:

Scatterplot of Marilyn Monroe with clusters



Pie chart of major colors based on K-means cluster with 9 clusters

For the Red image, we see that the background color, RGB: [186 21 4] has the highest percentage at 28.9 %. Following that we see that the hair color which is made of just one color RBG: [219 195 11] taking 21.5% of major colors and then there is the face color which is made with a shading of 2 different shades where majority of the face is shaded with, RGB: [229 174 191], and RGB:[176 133 132], combining and taking approximately 25.4% of the image colors. The teeth only comprise 0.2% with RGB: [231 227 209]. We also see that the shading (RGB: [15 10 7] mixed with RGB: [112 93 34]) takes 21.1% combined which is quite a significant share of the color in the image. The eye shadow of Marylin Monroe in this image is similar to the collar shade as well. The eye-shadow (RGB: [139 197 201]) and collar (RGB: [110 194 166]) makes a total of 2.8% of image color. Below is the segmentation of the different aspects of the image.

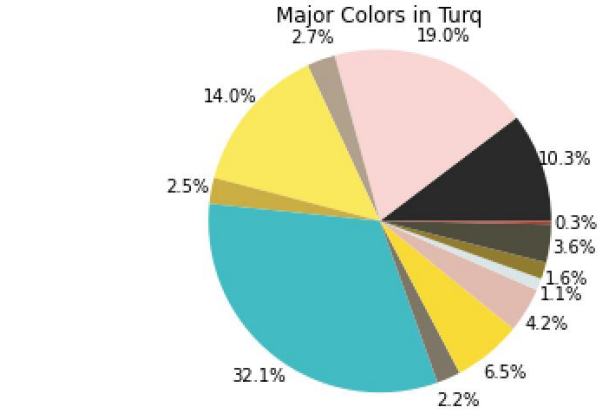


Turq Image of Marilyn Monroe

Original Image:

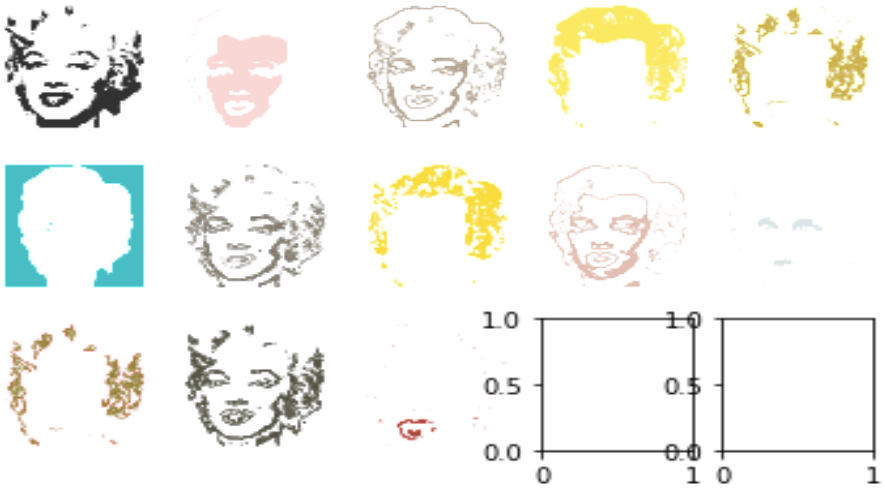
Scatterplot of Marilyn Monroe with clusters

Pie chart of major colors based on K-means cluster with 13 clusters



For the Turq image, we see that the background color, RGB: [67 187 193] has the highest percentage at 32.1 % and the collar here is the same color as the background. Following that we see that the hair color which is made of just 4 different shades with major color being RBG: [250 233 92] followed by RGB: [248 219 54], RGB: [203 173 68], RGB: [145 124 49] respectively,

taking 24.5% of colors percentage. The face color is made with 2 different shades where the majority of the face is shaded with, RGB: [249 214 211], and then RGB:[224 187 174], combining and taking approximately 23.2% of the image's major colors. The teeth and eye-shadow have the same color, RGB: [218 229 232]and lips, RGB: [163 66 54] together comprise 1.4%. We also see that the shading involves mixing 4 different colors namely RGB: [40 41 41] mixed with RGB: [79 77 61], RGB:[177 161 140] , RGB:[125 117 101] combining a total of 18.8%. Below is the segmentation of the different aspects of the image.



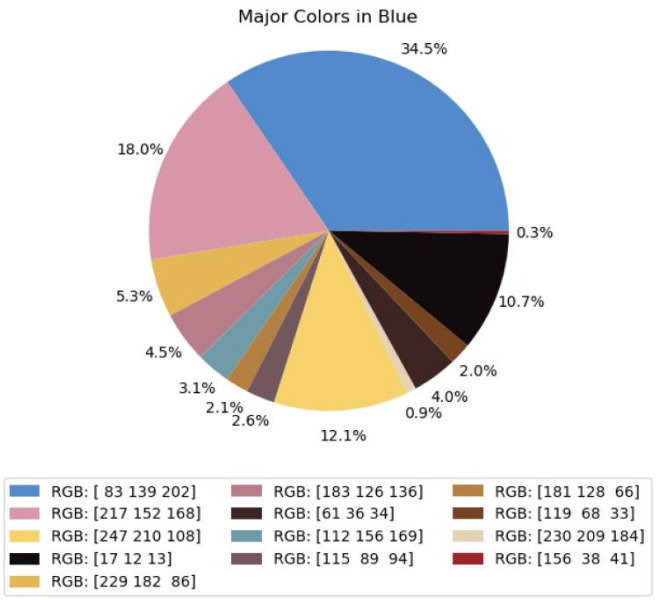
Blue Image of Marilyn Monroe

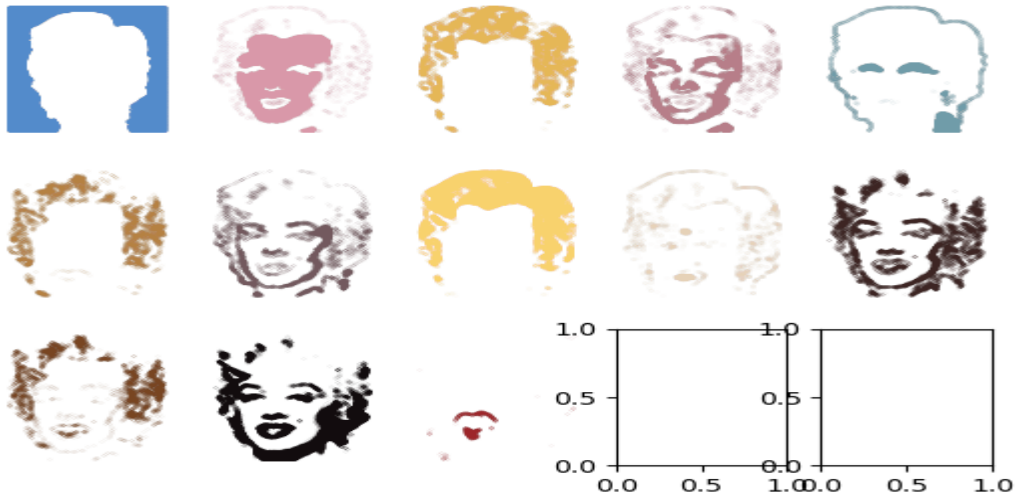
Original Image:

Scatterplot of Marilyn Monroe with clusters

Pie chart of major colors based on K-means cluster with 13 clusters

For the Blue image, we see that the background color, RGB: [83 139 202] has the highest percentage at 34.5% . The hair color is made of just 4 different shades with the major color being RBG: [247 152 168] followed by RGB: [229 182 86], RGB: [181 128 66], RGB: [119 68 33] respectively, taking 21.5% of color percentage. The face color is made with 2 shades where the majority of the face is shaded with, RGB: [217 152 168], and then RGB:[183 126 136], combining and taking approximately 22.5% of the image's major colors. The teeth, and the dot on the forehead have color RGB: [230 209 184] and the lip has RGB:[156 38 41] with a total percentage of 1.2%. The eye-shadow and the collar have the same color, RGB: [112 156 169] which has 3.1% of major color share. We also see that the shading involves mixing 3 different colors namely RGB: [17 12 13] mixed with RGB: [61 36 34], RGB:[115 89 94] making a total of 17.2%. Below is the segmentation of the different aspects of the image

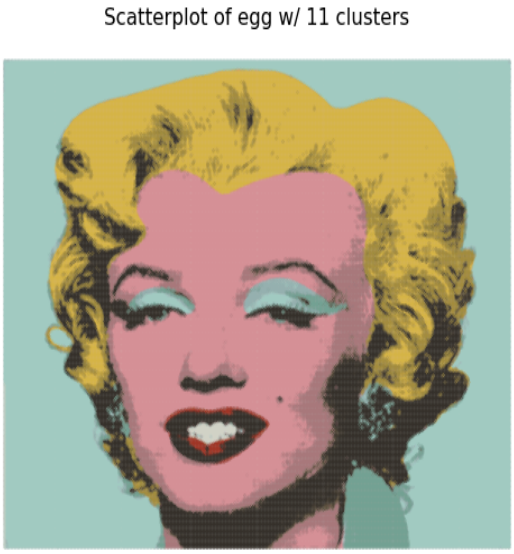




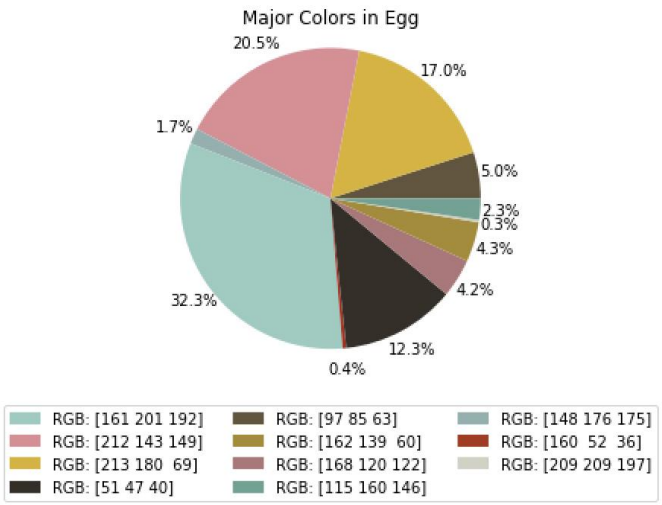
Eggblue Image of Marilyn Monroe

Original Image:

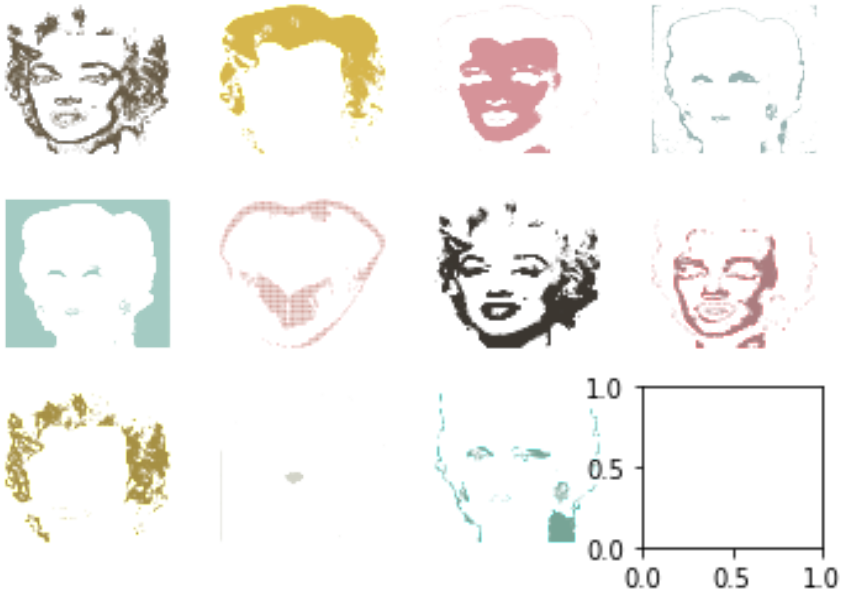
Scatterplot of Marilyn Monroe with clusters



Pie chart of major colors based on K-means cluster with 11 clusters



For the Eggblue image, we see that the background color, RGB: [161 201 192] has the highest percentage at 32.3% . The hair color is made of just 2 different shades with the major color being RBG: [213 180 169] followed by RGB: [4.3], taking 21.3% of color percentage. The face color is made with 3 shades where the majority of the face is shaded with, RGB: [212 143 149], and then RGB:[168 120 122] and with light shading of the cheeks with RGB: [160 52 36] combining and taking approximately 24.7% of the image's major colors. The teeth have color RGB: [209 209 197] and the lip has RGB:[160 52 36] with a total percentage of 0.7%. The eye-shadow and the collar are mixed with 3 different shades including the same shade as the background, the other two shades are RGB:[115 160 146] and RGB:[148 176 175] which makes a total of 4% of major color share. We also see that the shading is with RGB:[51 47 40] mixed with RGB:[97 85 63] having a total of 17.3%. Below is the segmentation of the different aspects of the image.



Below is a table which summarizes and compares the 5 images:

Image	Background Color	Hair Color	Face Color	Teeth and Lips	Eye-shadow and Collar	Shading
Orange	RGB: [226 123 83] (35.8%)	RGB: [242 195 89], [210 167 80], [152 121 64] (21.9%)	RGB: [227 174 187], [210 156 169], [161 122 132] (22.9%)	RGB: [174 69 52], [228 222 218] (0.7%)	RGB: [163 202 202] (1.6%)	RGB: [11 14 16], [46 44 41], [93 77 67] (15.7%)
Red	RGB: [186 21 4] (28.9%)	RGB: [219 195 11] (21.5%)	RGB: [229 174 191], [176 133 132] (25.4%)	RGB: [231 227 209] (0.2%)	RGB: [139 197 201], [110 194 166] (2.8%)	RGB: [15 10 7], [112 93 34] (21.1%)
Turq	RGB: [67 187 193] (32.1%)	RGB: [250 233 92], [248 219 54], [203 173 68], [145 124 49] (24.5%)	RGB: [249 214 211], [224 187 174] (23.2%)	RGB: [218 229 232] (1.4%)	-	RGB: [40 41 41], [79 77 61], [177 161 140], [125 117 101] (18.8%)

Blue	RGB: [83 139 202] (34.5%)	RGB: [247 152 168], [229 182 86], [181 128 66], [119 68 33] (21.5%)	RGB: [217 152 168], [183 126 136] (22.5%)	RGB: [230 209 184] (1.2%)	RGB: [112 156 169] (3.1%)	RGB: [17 12 13], [61 36 34], [115 89 94] (17.2%)
Eggblue	RGB: [161 201 192] (32.3%)	RGB: [213 180 169] (21.3%)	RGB: [212 143 149], [168 120 122], [160 52 36] (24.7%)	RGB: [209 209 197] (0.7%)	RGB: [115 160 146], [148 176 175], [161 201 192] (4%)	RGB: [51 47 40], [97 85 63] (17.3%)

When comparing the five images numerically, we can observe variations in key color components. The background colors range from RGB [67 187 193] with the highest percentage at 32.1% in the turq image to RGB [83 139 202] with 34.5% in the blue image. In terms of hair color, the red image stands out with a single shade, RGB [219 195 11], accounting for 21.5% of the major colors. The orange image has the most diverse hair color, comprising three shades, and taking up 21.9% of the image. The face color demonstrates variations as well, with the turq image showing a combination of shades, RGB [249 214 211] and [224 187 174], accounting for 23.2% of the major colors. The lips and teeth have the lowest overall percentage in all images, ranging from 0.2% to 1.4%, suggesting minimal prominence. The eye-shadow and collar colors differ, with the blue image having the most prominent share at 3.1%. Shading constitutes a significant portion of the images, ranging from 15.7% in the orange image to 21.1% in the red image. Overall the five images display noticeable differences in their color compositions. Each image has its own unique background color, ranging from warm tones like RGB [226 123 83] in the orange image to cooler tones like RGB [67 187 193] in the turq image. The hair colors vary, with the orange image exhibiting multiple shades, the red image having a single shade of RGB [219 195 11], and the turq and blue images featuring multiple shades as well. The face colors also differ among the images, showcasing various skin tones and shading. Teeth and lips exhibit distinct colors in each image, while eye-shadow and collar colors vary in combination and intensity. Shading plays a significant role in the images, with different combinations of shades creating variations in overall appearance. The five images showcase unique and diverse color compositions, resulting in visually distinct images.

Recreation of the Original Image of Marilyn Monroe

In order to attempt to recreate the potential of what the original image of Marilyn might look like we looked at the five previous images. We looked at the important features of Marilyn, again being the hair, face, eyeshadow, teeth, lips, background, and collar. Then we took the average of the five images for each of the important features and set that as the new color. Next, we took the orange scatterplot image of Marilyn since it had the most clusters so the segmentation will be more precise and we changed the color of the labels to the new averaged color and ran that as the new image. After taking the averages of these important features, we can assume that the hair, face, lips, and teeth make sense as to why they are this color. Marilyn was blond so her hair being close to yellow makes sense and pink was a common trend with her face color based on the images. Additionally, her lips are illustrated as red and her teeth seem pretty white and realistic. The three issues with this method are the eyeshadow, collar, and background. We have to guess the color of these parts solely based on what information we have. Judging on those images we do see the trend that the eyeshadow is a similar color to the collar, a lighter or a darker version. However, in the recreated image, this isn't the case as the colors are distinct from each other. Since the backgrounds for the five images are so different, when the RGB values of the images are averaged we get a certain gray color.