Comparative Analysis of Andy Warhol's Five "Shot Marilyn" Paintings

Yuanyuan Gao, Yirong Xu, Rachel How, Kayla Pham
University of California, Davis
STA 160 - Practice in Data Science
Professor Fushing Hsieh
June 11, 2023

Abstract

This report analyzes Andy Warhol's five shot Marilyn paintings, explicitly focusing on their color compositions and potential correlations with sold prices. The objective of this report is to explore Warhol's original ideas on colors and compare the images based on color-compositions, investigating relationships between color choices and artwork prices. The analysis mainly consists of four parts: relative conditional entropy, color analysis, cluster analysis, and correlation testing. Various methods, such as entropy analysis, statistical measures, histograms, ensemble learning, and K-means clustering, are employed to understand color usage, identify associations with pricing, and reveal patterns in Warhol's choices. By integrating these analytical methods, this report can contribute to a broader understanding of Warhol's artistic choices, highlight patterns in color compositions, and reveal the underlying factors that influenced the sale price of Marilyn's paintings. The results enhance our understanding of Warhol's work, his use of color, and the importance of color in the art market.

1 Introduction

Andy Warhol's "Shot Marilyn" is part of a five painting series, just by incorporating different colors allows viewers to interpret a series of distinctive emotions just from the virtue of colors and shadows in the paintings. In this report, the main objective is to gain insights into Warhol's original ideas about colors and examine any relationships between specific color choices and the artworks' selling prices.

To achieve these goals, the analysis will be divided into four parts: relative conditional entropy, color analysis, cluster analysis, and correlation testing. Various methods will be applied, including relative conditional entropy analysis, statistical measures and histograms for color analysis, classification techniques such as ensemble learning, cluster analysis to identify distinct color palettes, K-means clustering to determine the optimal K value, and correlation testing between quantitative image information and sold prices. These methods will provide a comprehensive understanding of the color usage in the Marilyn paintings and help identify any significant associations between colors and pricing.

By integrating these analytical approaches, the project aims to contribute to a broader understanding of Warhol's artistic choices, highlight patterns in color compositions, and uncover potential factors that influence the selling prices of the Marilyn paintings. The results will enhance our knowledge of Warhol's work, his use of colors, and the significance of color in the art market.

2 Research Questions

1. Each of the paintings were sold, and the orange and sage blue Marilyn were sold for the most. Is there a correlation between these paintings and their sold price?
2. What does K-Means Clustering tell us about the five Marilyn photos?
3. What are similarities and differences among the five Marilyn photos?

3 Dataset Description

Our dataset comprises Andy Warhol's iconic series of "Shot Marilyn" paintings from 1964, consisting of five distinct variations [4]. Our analysis will focus on comparing the colors and outlines used in these paintings. The five shots we will be examining are categorized by their predominant colors: Light Blue, Sage Blue, Orange, Red, and Turquoise.

(a) Shot Light Blue Marilyn

(b) Shot Sage Blue Marilyn

(c) Shot Orange Marilyn

(d) Shot Red Marilyn

(e) Shot Turquoise Marilyn

Figure 1: Andy Warhol's five "Shot Marilyn" paintings (1964)

4 Data Visualization

In Figure 2 and Figure 3, we can see significant observations regarding the color channels of the five Marilyn images. The light blue Marilyn image demonstrates high values in the blue color channel (B), while observing relatively low values in the red and green color channels (R and G). The sage blue Marilyn image displays high values in the blue color channel, moderate values in the red channel, and low values in the green channel. Moving on to the orange and red Marilyn images, they both have a high value in the red color channel, while they have a lower values in blue and green color channels.

In the case of the red Marilyn image, we can observe a sufficient presence of green as indicated by the visible yellow pattern on the graph. This yellow pattern arises from the combination of red and green colors in the image. Finally, the turquoise Marilyn image shows high values in the blue and green channels, while displaying a relatively low value in the red channel. However, it is important to note that an enough amount of red is present, allowing for the distinct visualization of the yellow pattern. This yellow pattern emerges as a result of the harmonious blend of red and green within the image.

These observations clarify the distinct color characteristics and channel dependencies exhibited by each Marilyn image, providing valuable insights into their color composition.

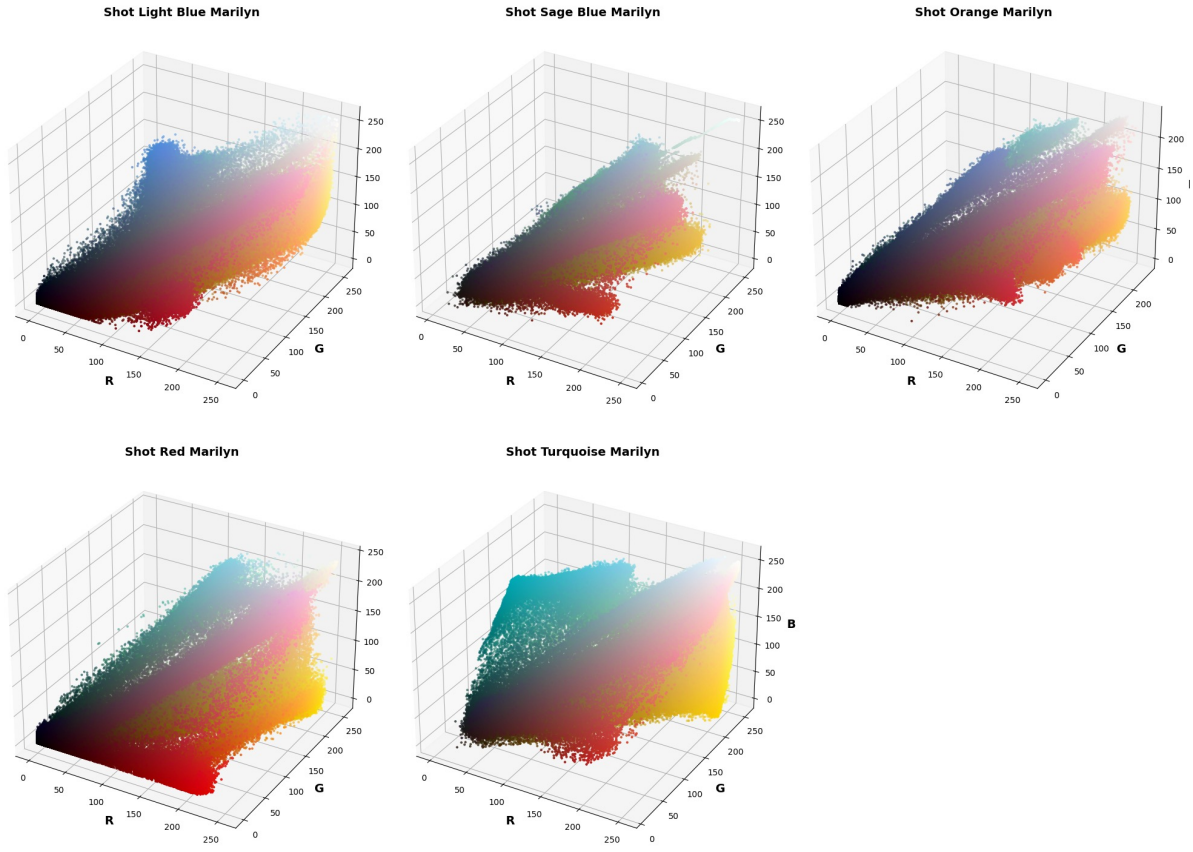


Figure 2: RGB 3D Scatter Plots in Light Blue, Sage Blue, Orange, Red, and Turquoise

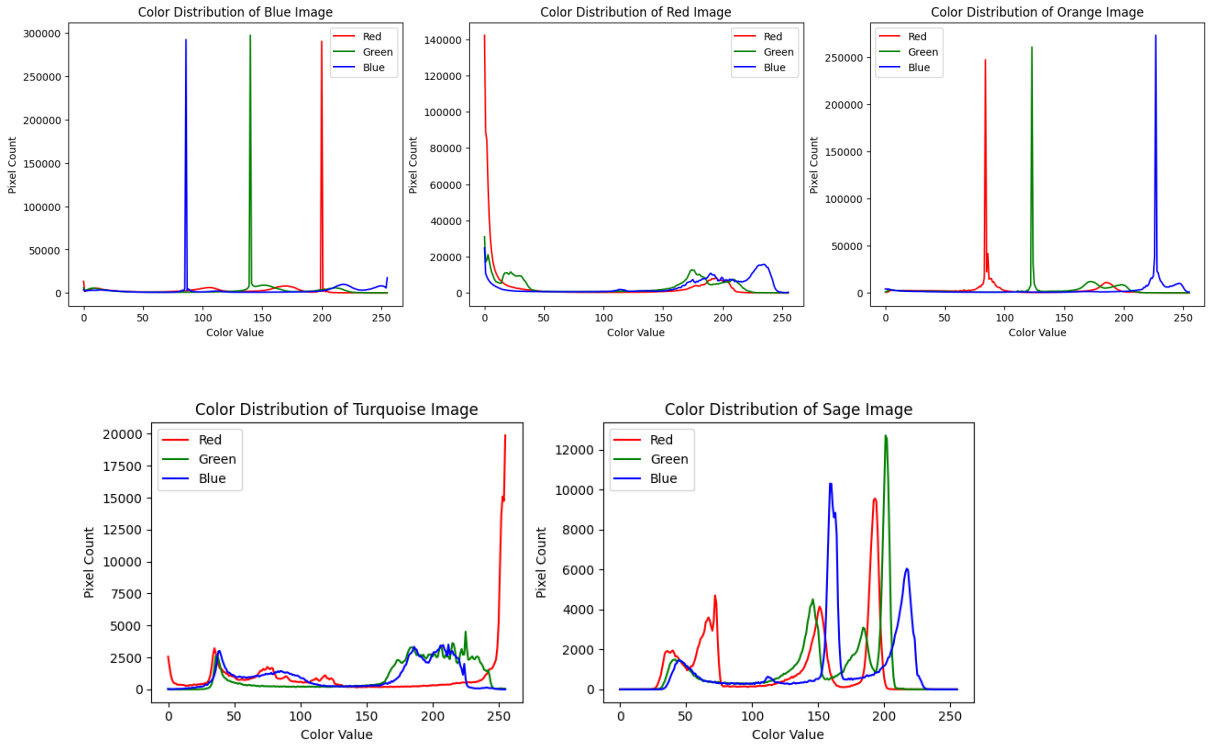


Figure 3: RGB 2D - Color Distribution of the Five Marilyn Paintings

5 Methodology

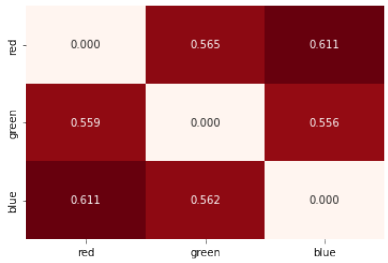
To analyze the five images, we have chosen to apply several key methods for this project: Relative Conditional Entropy, Color Analysis, Classification, Cluster Analysis, K-Means Algorithm, and Correlation Testing. These methods will be employed to gain insights and extract meaningful information from the images.

1. **Relative Conditional Entropy:** Relative Conditional Entropy is a method employed in this project to analyze the relationship between variables and determine the amount of information shared between them. By calculating the relative conditional entropy, we can assess the degree of correlation or dependence between different variables in the context of our analysis.
2. **Color Analysis:** Extract the color information from each image by converting the images to a suitable color space, such as RGB or HSV. We calculate statistical measures such as the mean, standard deviation, or color histograms for each color channel. By comparing these statistical measures across the different paintings, we aim to identify any variations or similarities in color usage.
3. **Cluster Analysis:** Cluster analysis involves applying clustering algorithms, such as K-means clustering, to group similar colors together within each painting. This technique helps identify distinct color palettes and patterns in the artworks. By clustering colors, we can determine if specific color combinations or arrangements are commonly observed throughout the series or if each painting has a unique color composition.
4. **Correlation Testing:** Correlation testing involves extracting specific color attributes from the images and calculating the correlation between these attributes and the painting's sold price. This analysis helps us determine any potential relationship between color characteristics and the artwork's monetary value.

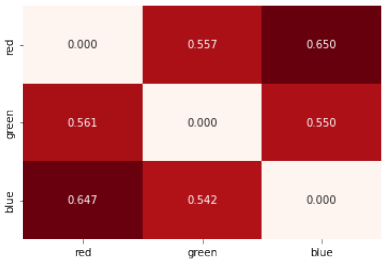
6 Relative Conditional Entropy

In this section, we computed the relative conditional entropy for the five Marilyn images to analyze the relationship between different color channels. We first determined the relative conditional entropy between pairs of colors, which quantifies the level of uncertainty in describing one color (red, green, or blue) given another color channel. Then we constructed a matrix to represent the relative conditional entropy, depicting the amount of shared information between the two colors on the x and y axes.

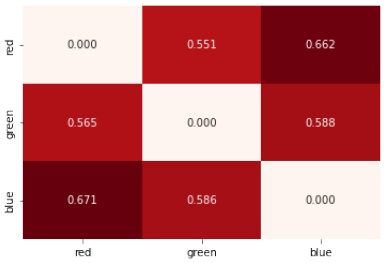
From the matrices in Figure 4 below, we can observe that all values fall within the range of 0 and 1. 0 refers to the dependence between two colors while 1 refers to the independence between two colors. With the exception of the diagonal elements, which represent the conditional entropy values of a color compared to itself and therefore have a value of 0, all other relative conditional entropy values for the five Marilyn images lie between 0.551 and 0.877. This indicates that knowing the value of one color channel (red, green or blue) provides a relatively low information towards the value of another color channels. Therefore, we can conclude that the colors are not predictable based on the values of other color channels.



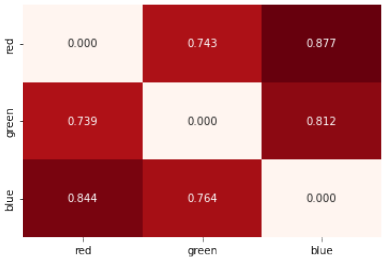
(a) Light Blue Marilyn



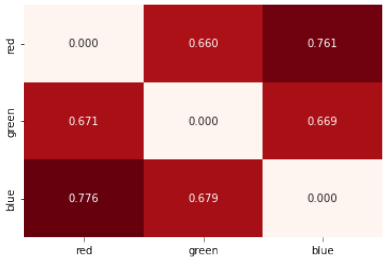
(b) Sage Blue Marilyn



(c) Orange Marilyn



(d) Red Marilyn



(e) Turquoise Marilyn

Figure 4: Relative conditional entropy between red, green, and blue coordinates of all pixels for the five Marilyn images

7 Color Analysis

7.1 Dominant Colors in the Five Images

In this section, we conducted an analysis of the dominant colors in the five Marilyn paintings using the K-means clustering algorithm. We present the findings through five donut charts, each representing the top 11 dominant colors in a respective Marilyn painting.

To provide a visual representation of the artwork being analyzed, the original image of each Marilyn painting is displayed on the left side. The corresponding donut chart, located on the right side, illustrates the proportions of the dominant colors within the painting. Each color segment in the donut chart represents a dominant color, and the size of the segment corresponds to its proportion in the overall image. To make it easier to understand, a legend is provided alongside the donut chart, displaying the hex code of each color segment and the percentage indicating its prevalence in the image. Furthermore, to offer a comprehensive understanding of the dominant colors, we have included five tables showcasing the RGB values for the top 11 dominant colors in each image. These tables provide specific information about the color composition of each Marilyn painting, offering more detailed insights.

By incorporating these visual representations and accompanying tables, we aim to provide a comprehensive and visually appealing analysis of the dominant colors in the Marilyn paintings. These visual elements help us interpret the proportions of different colors and contribute to a deeper understanding of Andy Warhol's artistic choices and the unique color compositions found in his iconic Marilyn series.

7.1.1 Light Blue Marilyn

From Figure 5, we can see that the most dominant color in the light blue Marilyn painting is light blue ($\#568cc7$) with 34.5% of the entire painting, the second dominant color is ($\#899a8$) with 18.3% of the painting, and the third dominant color is ($\#e4b558$), representing 12.2% of the overall composition. For the remain dominant colors, we repeated the same structure and provided the dominant colors and percentages for each.

Table 1 provides the RGB values of the dominant colors. For the most dominant color, we observe a relatively high blue value of 199.83, a moderate green value of 140.03, and a lower red value of 86.09. As for the second dominant color, we note a moderate red value of 67.27, accompanied by relatively lower values for green and blue, which are 41.66 and 38.47, respectively. The third dominant color exhibits relatively higher values for red and green, suggesting a yellow color that results from the combination of red and green.

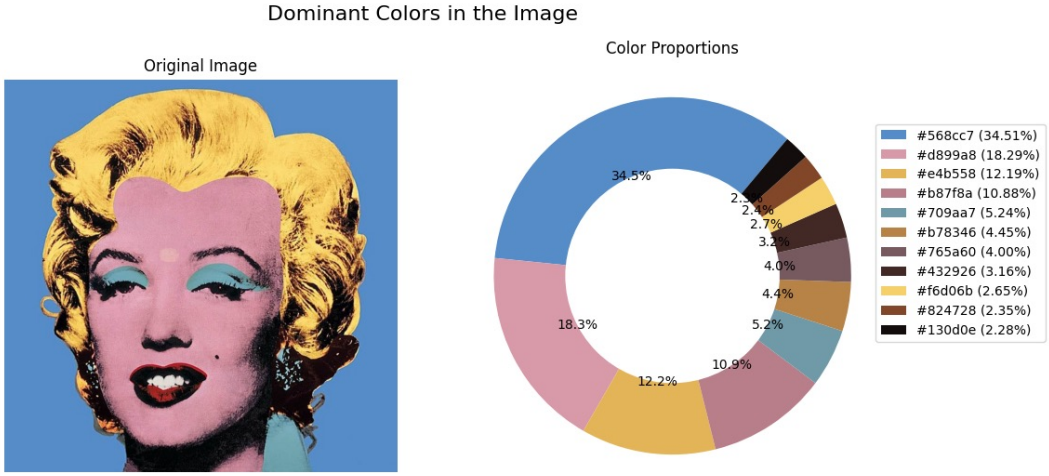


Figure 5: Top 11 Dominant Colors of Light Blue Marilyn

	Red values	Green Values	Blue Values
1	86.09232831	140.03425336	199.82613735
2	67.27196087	41.65788926	38.47335554
3	183.13499727	131.05993806	70.2244489
4	216.55820576	153.14570804	168.54143886
5	228.45244134	181.83901395	88.45006341
6	112.41961457	154.94772967	167.56480993
7	19.16223353	13.31819052	14.79811703
8	246.85576264	208.88385638	107.43812526
9	130.08888102	71.6180949	40.74858357
10	118.75862069	90.75783341	96.88127854
11	184.10819672	127.20899297	138.02744731

Table 1: RGB Values for Top 11 Dominant Colors of Light Blue Marilyn

7.1.2 Sage Blue Marilyn

From Figure 6, we can see that the most dominant color in the sage blue Marilyn painting is light blue ($\#d7b646$) with 33.4% of the entire painting, the second dominant color is ($\#a0c8bf$) with 17.9% of the painting, and the third dominant color is ($\#bd8185$), representing 13.9% of the overall composition. For the remain dominant colors, we repeated the same structure and provided the dominant colors and percentages for each.

From Table 2, we explore the RGB values of the dominant colors. For the most dominant color, we observe a relatively low blue value of 70.11, a relatively high green value of 182.45, and a large red value of 215.64. For the second dominant color, we note a significant red value of 160.73, accompanied by substantial values for green and blue, which are 200.30 and 191.83, respectively. As for the third dominant color, it exhibits relatively high values for red, green, and blue, specifically 189.18, 129.66, and 133.60, respectively.

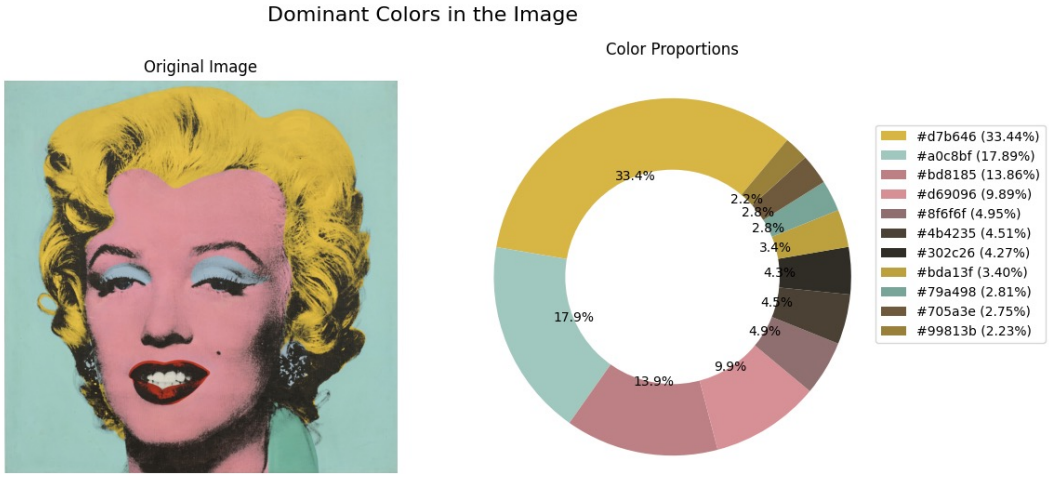


Figure 6: Top 11 Dominant Colors of Sage Blue Marilyn

	Red values	Green Values	Blue Values
1	215.63656593	182.45145091	70.10822433
2	160.73227346	200.29929854	191.83324404
3	189.18349702	129.66146227	133.59741459
4	143.47410581	111.96367362	111.77533532
5	75.70731231	66.25715123	53.22298155
6	189.9597303	161.00544934	63.12090145
7	121.49772865	164.85523925	152.66262871
8	48.86541537	44.83654575	38.89462166
9	214.38020347	144.80908158	150.43690498
10	112.95527859	90.23435973	62.06561584
11	153.37059611	129.70804222	59.13795154

Table 2: RGB Values for Top 11 Dominant Colors of Sage Blue Marilyn

7.1.3 Orange Marilyn

From Figure 7, we can see that the most dominant color in the Orange Marilyn painting is ($\#e27b54$) with 35.9% of the entire painting, the second dominant color is ($\#c2909c$) with 18.3% of the painting, and the third dominant color is ($\#f1c158$), representing 16.2% of the overall composition. For the remain dominant colors, we repeated the same structure and provided the dominant colors and percentages for each.

Table 3 provides us RGB values of the dominant colors. For the most dominant color, we observe a significant red value of 226.57, a moderate green value of 123.05, and a relatively lower blue value of 84.06. For the second dominant color, we note relatively large values for red, green, and blue, specifically 225.50, 172.56, and 185.37, respectively. The third dominant color displays small values for red, green, and blue, amounting to 10.95, 14.02, and 16.51, respectively.

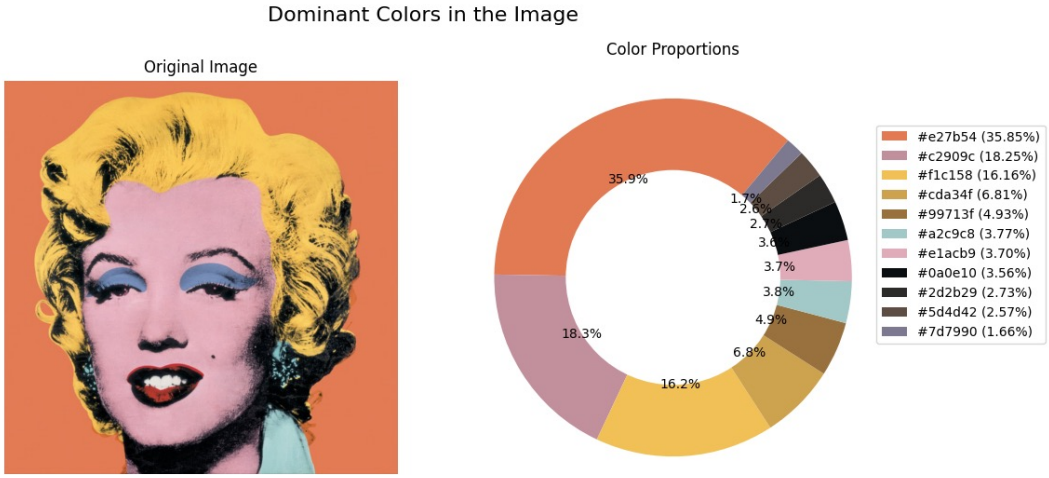


Figure 7: Top 11 Dominant Colors of Orange Marilyn

	Red values	Green Values	Blue Values
1	226.57151823	123.04626864	84.05552004
2	225.50170851	172.55597066	185.37290993
3	10.94779313	14.01840078	16.50770265
4	241.56248711	193.70368269	88.61060966
5	45.03856563	43.5	41.99399526
6	162.29178185	201.55993968	200.71852224
7	153.26804281	113.0911315	63.06544343
8	205.95343606	163.03722866	79.49488247
9	93.02794036	77.03511872	66.76322474
10	194.64586767	144.38768543	156.20261361
11	125.29720052	121.17285156	144.21842448

Table 3: RGB Values for Top 11 Dominant Colors of Orange Marilyn

7.1.4 Red Marilyn

From Figure 7, we can see that the most dominant color in the Red Marilyn painting is ($\#c70a02$) with 19.5% of the entire painting, the second dominant color is ($\#cd99a1$) with 18.2% of the painting, and the third dominant color is ($\#8d7516$), representing 15.7% of the overall composition. For the remain dominant colors, we repeated the same structure and provided the dominant colors and percentages for each.

Table 4 presents the corresponding RGB values of the dominant colors. For the most dominant color, we observe a significant red value of 205.05, a moderate green value of 153.24, and a relatively large blue value of 161.71. For the second dominant color, we note a relatively large red value of 178.03, while the green and blue values are small, amounting to 26.51 and 2.72, respectively. The third dominant color exhibits a relatively large value for red and green, measuring 227.65 and 204.01, respectively. However, the blue value is very small, with a value of 6.82.

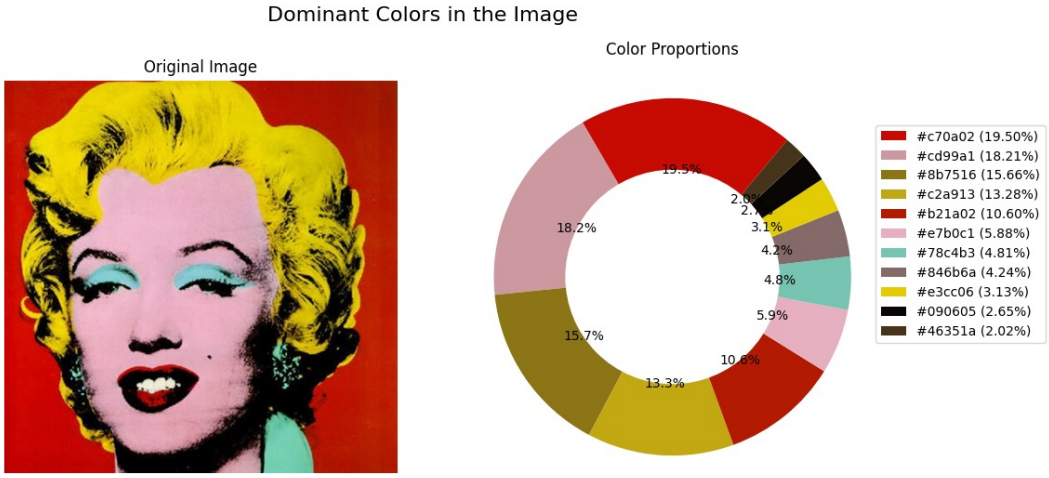


Figure 8: Top 11 Dominant Colors of Red Marilyn

	Red values	Green Values	Blue Values
1	205.04977141	153.23738463	161.70870353
2	178.03228168	26.51348692	2.71836464
3	227.64509506	204.00757273	6.82380076
4	231.40891199	176.5057759	193.88645018
5	139.58573134	117.04800424	22.41572736
6	132.41862382	107.49217141	106.53893696
7	194.9844479	169.67722324	19.91050474
8	70.3048637	53.19239066	26.8617376
9	120.1895117	196.98791019	179.40320301
10	199.58409966	10.30040085	2.1835652
11	9.73977112	6.48872864	5.03022417

Table 4: RGB Values for Top 11 Dominant Colors of Red Marilyn

7.1.5 Turquoise Marilyn

From Figure 9, we can see that the most dominant color in the Turquoise Marilyn painting is ($\#45babf$) with 21.2% of the entire painting., the second dominant color is $\#f7d5d2$ with 14.6% of the painting, and the third dominant color is $\#ceafa0$, representing 13.0% of the overall composition. For the remain dominant colors, we repeated the same structure and provided the dominant colors and percentages for each.

Table 5 presents the corresponding RGB values of the dominant colors. For the most dominant color, we observe a significant red value of 247.64, a relatively large green and blue values of 213.56 and 210.29 respectively. For the second dominant color, there are moderate red and green values of 142.58 and 122.63, and a relatively small blue value of 89.82. The third dominant color exhibits a relatively large value for red and green, measuring 249.98 and 219.1, respectively. However, the blue value is relatively small, with a value of 57.51.

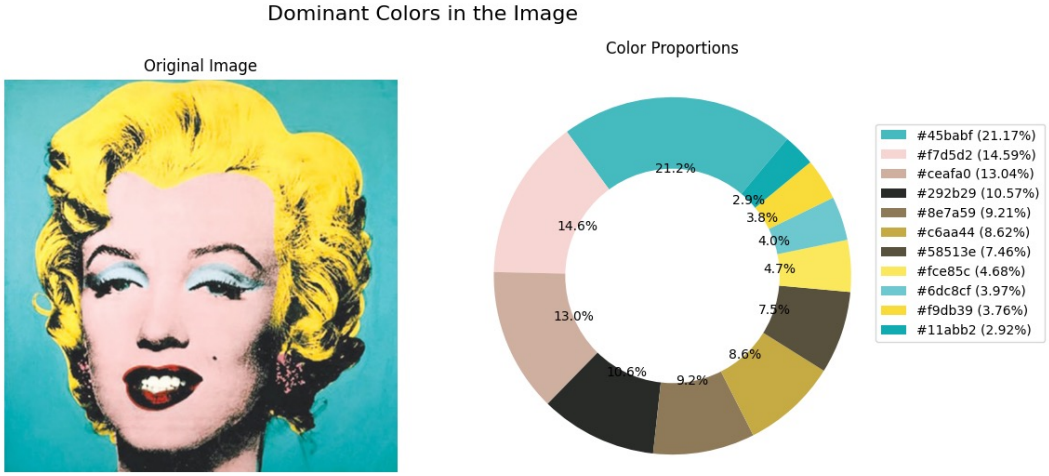


Figure 9: Dominant Colors of Turquoise Marilyn

	Red values	Green Values	Blue Values
1	247.64313949	213.55597947	210.28519735
2	142.57763975	122.63420586	89.8216504
3	249.97504466	219.09803484	57.5076485
4	252.73334185	232.97292551	92.80971233
5	17.41191823	171.12155043	178.1086946
6	206.87872852	175.28136408	160.37788265
7	198.66399657	170.13975111	68.59605207
8	69.15311046	186.14286122	191.40648642
9	88.79301301	81.68422157	62.13973982
10	109.99642469	200.95845402	207.83223208
11	41.422081	43.07018236	41.87874003

Table 5: RGB Values for Top 11 Dominant Colors of Turquoise Marilyn

7.2 RGB Pixel Visualization based on Regions of Interest

We chose to analyze the major colors presented in the five Marilyn images. We separated them into the parts of background, face, hair, and eye-shadow.

In this part, we first extracted the color of the desired part and then masked the images to obtain the regions of interests. When masking the images, we first found the HSV ranges, which are the hue range (in the range of (0,360)), saturation range (in the range of (0,255)), and the value range (in the range of (0,255)). The Hue range represents the wavelength of the light presented in color that helps to separate the basic categories of colors. Hue values/range represent different colors. Hue ranges over colors with familiar names like red, green, orange, yellow, green, blue, indigo, and violet, and it is often visualized to be on a color circle. It is usually measured in 360 degrees of a color circle, where red is at 0°, green is at 120°, and blue is 240°[1]. The saturation range represents the intensity of a color, where lower values/range shows that the color is relatively unsaturated and higher values/range shows that the color is more intense. Lastly, the value range shows the brightness of the color, where the higher the range, the brighter the color. After the process of calculating the HSV ranges, we would be able to obtain the masked images.

We then split the masked images into individual color channels (B, G, and R) and plotted the scatter plot for each of the Marilyn images to visualize the relationship between each color channel (red vs. green, red vs. blue, and green vs. blue).

7.2.1 Background

For the background part, we first extracted the background colors for each of the five Marilyn images using HSV, and then masked the images to obtain the regions of interests as shown in Figure 10.

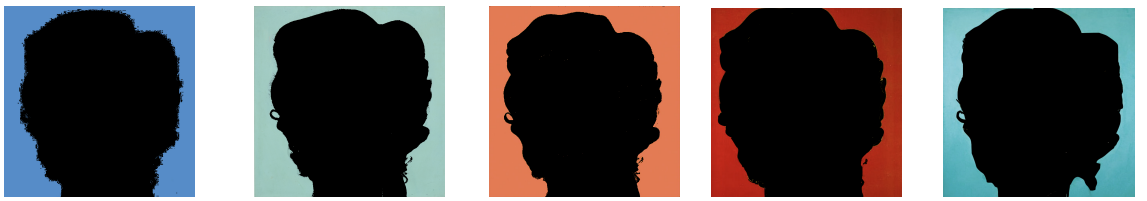
(a). For the background of the light blue Marilyn, we obtained a hue range of [105, 106], a saturation range of [140, 150], and a value range of [196, 202]. From the given ranges, we can say that the background color is within the range of green and the color has a relatively high intensity and a high brightness. After plotting a scatter plot on RGB space (see Figure 11 (a)), we can see that for the background color, blue and green has a higher color intensity than red, and blue has a higher color intensity than green.

(b). For the background of the sage blue Marilyn, we obtained a hue range of [60, 88], a saturation range of [2, 60], and a value range of [174, 250]. From the given ranges, we can say that the background color is within the range of green and the color has a relatively low intensity and a high brightness. After plotting a scatter plot on RGB space (see Figure 11 (b)), we can see that there appears to have a strong linear relationship between each pair of red, green, and blue colors, showing that for each pair of pixels, they have similar color intensities.

(c). For the background of the orange Marilyn, we obtained a hue range of [0, 14], a saturation range of [125, 185], and a value range of [201, 249]. From the given ranges, we can say that the background color is within the range of red and the color has a relatively high intensity and a high brightness. After plotting a scatter plot on RGB space (see Figure 11 (c)), we can see that red has a higher color intensity than green and blue, and that green has a higher color intensity than blue.

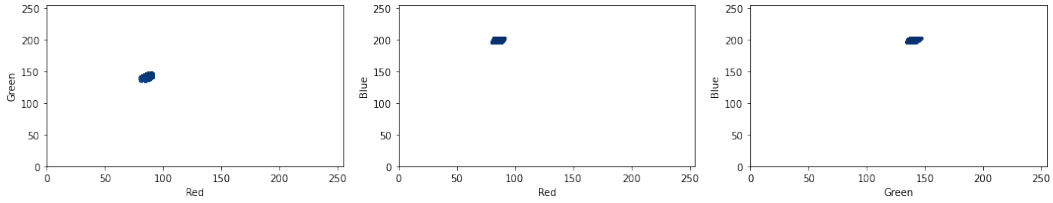
(d). For the background of the red Marilyn, we obtained a hue range of [0, 3], a saturation range of [238, 255], and a value range of [182, 218]. From the given ranges, we can say that the background color is red and the color has an extremely high intensity and a relatively high brightness. After plotting a scatter plot on RGB space (see Figure 11 (d)), we can see that this background color consists of mostly red, with a minimum amount of green and blue.

(e). For the background of the turquoise Marilyn, we obtained a hue range of [89, 94], a saturation range of [142, 255], and a value range of [150, 212]. From the given ranges, we can say that the background color is in the range of green and the color has a relatively high intensity and a relatively high brightness. After plotting a scatter plot on RGB space (see Figure 11 (e)), we can see that pixels have a similar blue and green intensity.

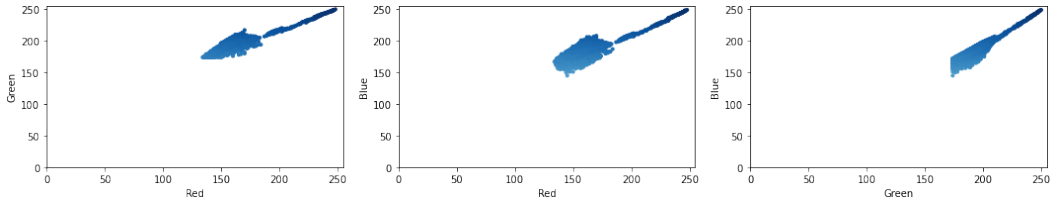


(a) Light Blue Marilyn **(b)** Sage Blue Marilyn **(c)** Orange Marilyn **(d)** Red Marilyn **(e)** Turquoise Marilyn

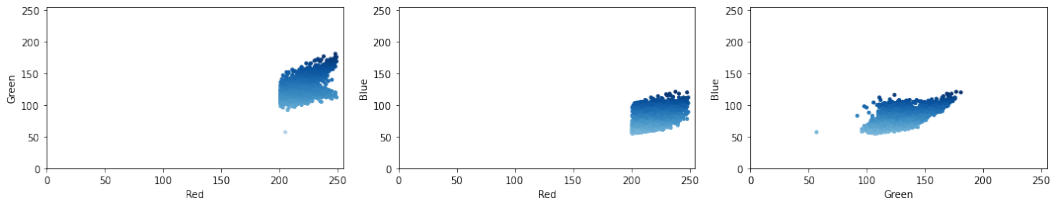
Figure 10: Region of Interest of the Background Portion of the Five Marilyn Images



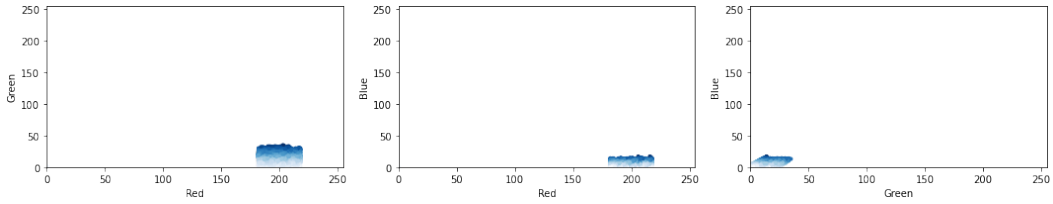
(a) Light Blue Marilyn



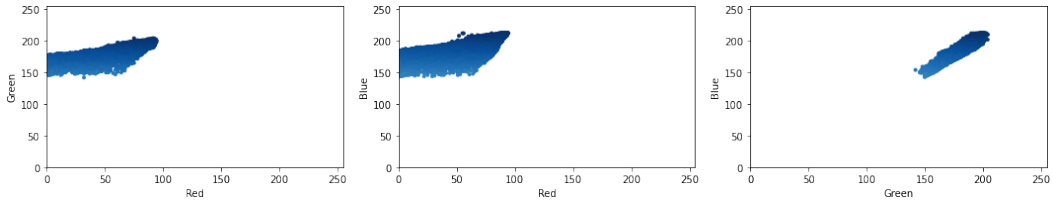
(b) Sage Blue Marilyn



(c) Orange Marilyn



(d) Red Marilyn



(e) Turquoise Marilyn

Figure 11: Red, Green, and Blue Space of the Background Portion of the Five Marilyn Images

7.2.2 Face

For the face part, we again extracted the face colors for each of the five Marilyn images using HSV first, and then masked the images to obtain the regions of interests as shown in Figure 12.

(a). For the background of the light blue Marilyn, we obtained a hue range of [0, 179], a saturation range of [62, 92], and a value range of [182, 242]. From the given ranges, the face color has a wide range, from red to cyan, and the color has a relatively low intensity and a relatively high brightness.

(b). For the background of the sage blue Marilyn, we obtained a hue range of [0, 179], a saturation range of [72, 91], and a value range of [193, 226]. From the given ranges, the face color has a wide range, from red to cyan, and the color has a relatively low intensity and a relatively moderate to high brightness.

(c). For the background of the orange Marilyn, we obtained a hue range of [2, 179], a saturation range of [48, 70], and a value range of [209, 241]. From the given ranges, the face color has a wide range, from red to cyan, and the color has a relatively low intensity and a high brightness.

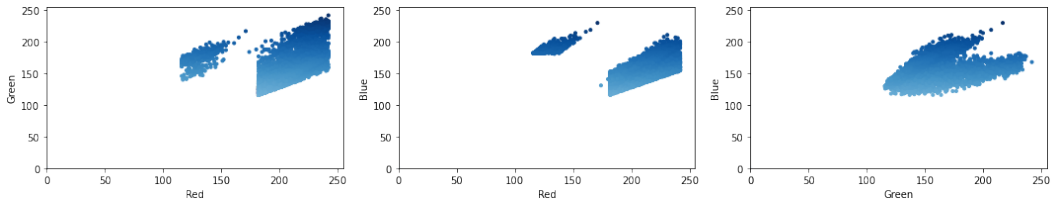
(d). For the background of the red Marilyn, we obtained a hue range of [165, 173], a saturation range of [49, 70], and a value range of [203, 245]. From the given ranges, we can say that the face color is closest to cyan and the color has a relatively low intensity and a high brightness.

(e). For the background of the turquoise Marilyn, we obtained a hue range of [0, 179], a saturation range of [26, 49], and a value range of [247, 255]. From the given ranges, the face color has a wide range, from red to cyan, and the color has a relatively low intensity and an extremely high brightness.

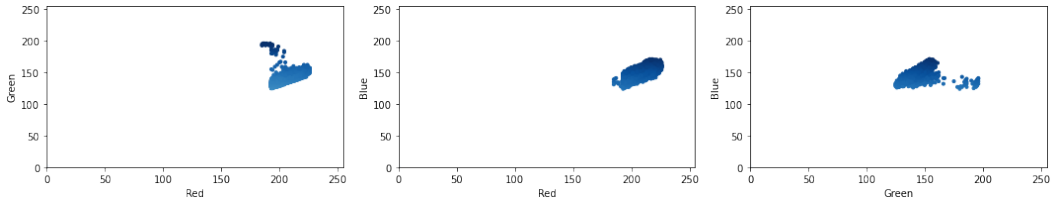
By viewing the face color in Figure 12, they all appeared to have a similar tanned color. From Figure 13 below, we can see that the face color of the five images all have a relatively linear relationship. For all the face portions of the five Marilyn images, the pixels all have a similar trend and all of them have higher red intensities compared to green and blue.

(a) Light Blue Marilyn **(b)** Sage Blue Marilyn **(c)** Orange Marilyn **(d)** Red Marilyn **(e)** Turquoise Marilyn

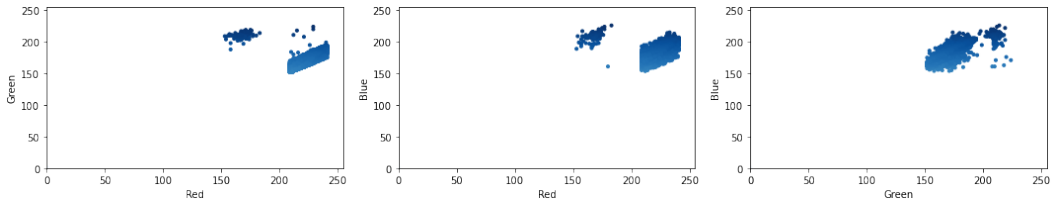
Figure 12: Region of Interest of the Face Portion of the Five Marilyn Images



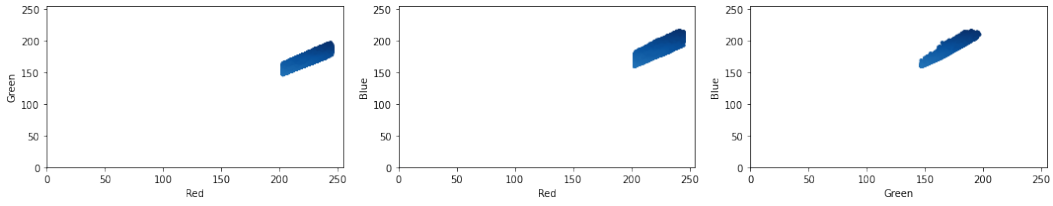
(a) Light Blue Marilyn



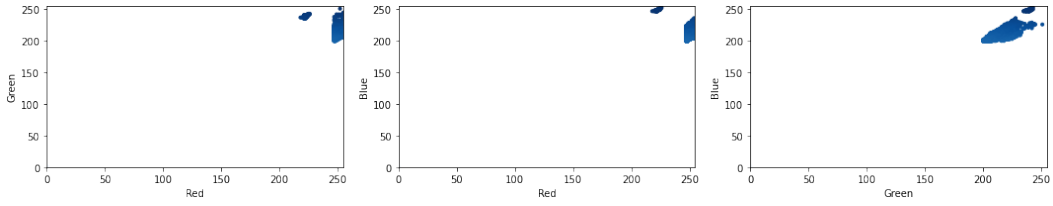
(b) Sage Blue Marilyn



(c) Orange Marilyn



(d) Red Marilyn



(e) Turquoise Marilyn

Figure 13: Red, Green, and Blue Space of the Face Portion of the Five Marilyn Images

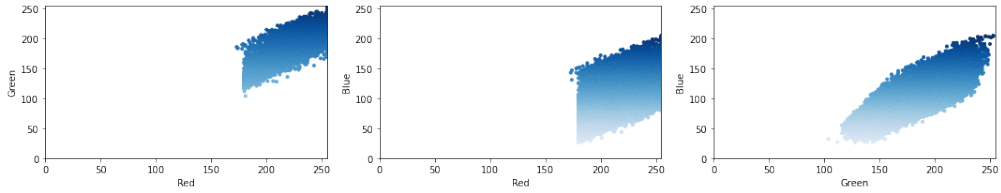
7.2.3 Hair

For the hair part, since they all have a yellow color, for the HSV range, we used a hue range of [15, 40], a saturation range of [50, 255], and a value range of [180, 255], and then masked the images to obtain the regions of interests as shown in Figure 14, we can see that all the masked parts are shown accurately.

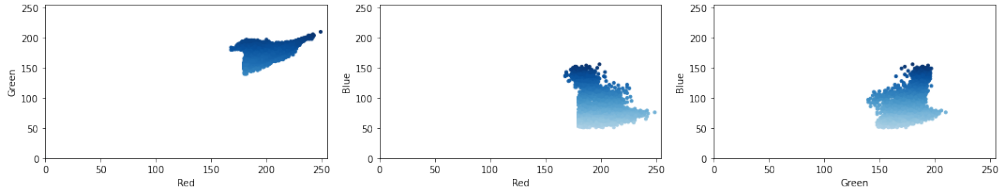
We then plotted scatter plots for the hair part of the five Marilyn images on RGB space. From Figure 15, we can see that the patterns are similar for all five scatter plots. The intensities of the red vs. green plot are constant, therefore the yellow color presented in each plot is constant as well. We can also conclude from the scatter plots that they all have a relatively low intensity of blue.

(a) Light Blue Marilyn **(b)** Sage Blue Marilyn **(c)** Orange Marilyn **(d)** Red Marilyn **(e)** Turquoise Marilyn

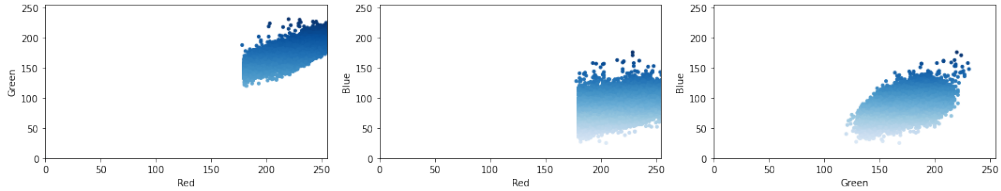
Figure 14: Region of Interest of the Hair Portion of the Five Marilyn Images



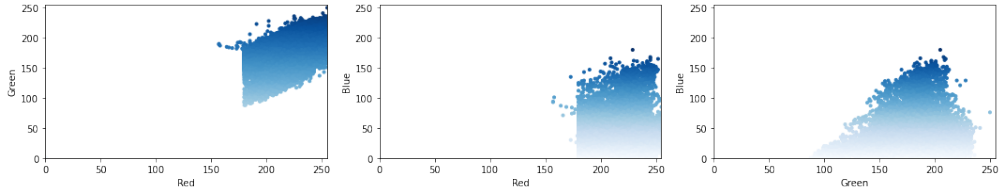
(a) Light Blue Marilyn



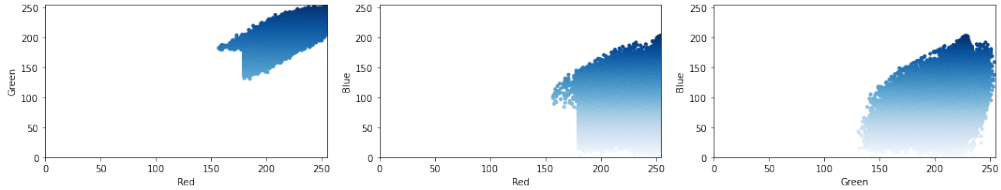
(b) Sage Blue Marilyn



(c) Orange Marilyn



(d) Red Marilyn



(e) Turquoise Marilyn

Figure 15: Red, Green, and Blue Space of the Hair Portion of the Five Marilyn Images

7.2.4 Eye-shadow

For the eye-shadow part, we again extracted the eye-shadow colors for each of the five Marilyn images using HSV first, and then masked the images to obtain the regions of interests as shown in Figure 16.

(a). For the eye-shadow of the light blue Marilyn, we obtained a hue range of [89, 156], a saturation range of [10, 110], and a value range of [141, 194]. From the given ranges, we can say that the eye-shadow color is within the range of green and the color has a relatively low intensity and a moderate brightness. After plotting a scatter plot on RGB space (see Figure 17 (a)), we can see that the pixels have higher green and blue intensities than red.

(b). For the eye-shadow of the sage blue Marilyn, we obtained a hue range of [99, 104], a saturation range of [34, 50], and a value range of [135, 203]. From the given ranges, we can say that the eye-shadow color is within the range for green and the color has a relatively low intensity and a relatively high brightness. After plotting a scatter plot on RGB space (see Figure 17 (b)), we can see that pixels have similar color intensities for green, red, and blue. Therefore, the eye-shadow color presented in Figure 16 (b) looks bright and close to white (when a similar amount of red, green, and blue are mixed together).

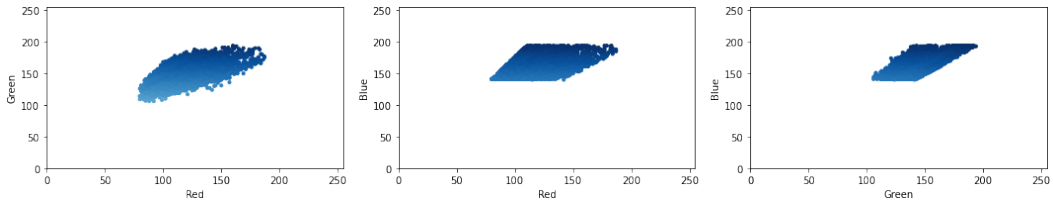
(c). For the eye-shadow of the orange Marilyn, we obtained a hue range of [106, 115], a saturation range of [77, 111], and a value range of [101, 189]. From the given ranges, we can say that the eye-shadow color is within the range of green and the color has a moderate intensity and a moderate to high brightness. After plotting a scatter plot on RGB space (see Figure 17 (c)), we can see that the pixels have higher green and blue intensities than red.

(d). For the eye-shadow of the red Marilyn, we obtained a hue range of [84, 99], a saturation range of [44, 131], and a value range of [123, 219]. From the given ranges, we can say that the eye-shadow color is within the range for green and the color has a moderate intensity and a relatively high brightness. After plotting a scatter plot on RGB space (see Figure 17 (d)), we can see that the pixels have higher green and blue intensities than red and that green has a higher intensity than blue.

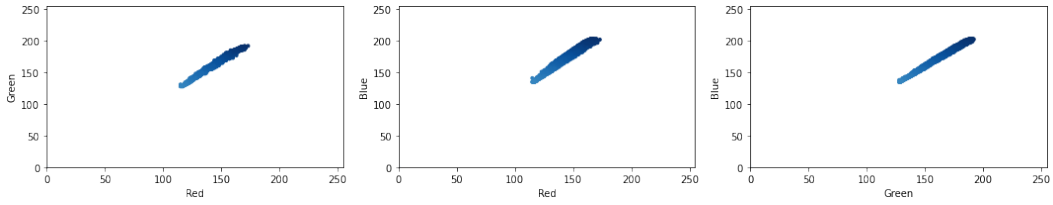
(e). For the eye-shadow of the turquoise Marilyn, we obtained a hue range of [0, 176], a saturation range of [2, 38], and a value range of [206, 246]. From the given ranges, we can say that the eye-shadow color has a wide range, from red to cyan, and the color has a low intensity and an extremely high brightness. After plotting a scatterplot on RGB space (see Figure 17 (e)), we can see that the pixels have similar color intensities for green, red, and blue. Therefore, the eye-shadow color presented in Figure 16 (b) looks bright and close to white (when a similar amount of red, green, and blue are mixed together).

(a) Light Blue Marilyn **(b)** Sage Blue Marilyn **(c)** Orange Marilyn **(d)** Red Marilyn **(e)** Turquoise Marilyn

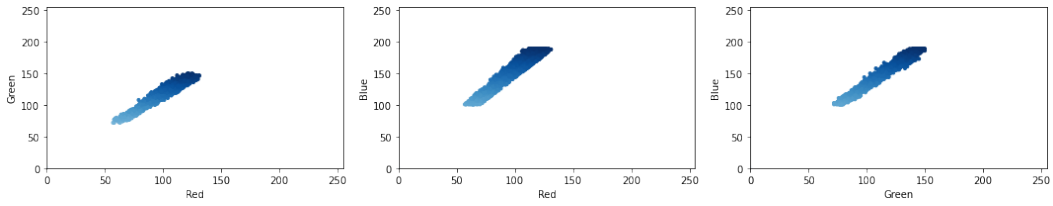
Figure 16: Region of Interest of the Eye-shadow Portion of the Five Marilyn Images



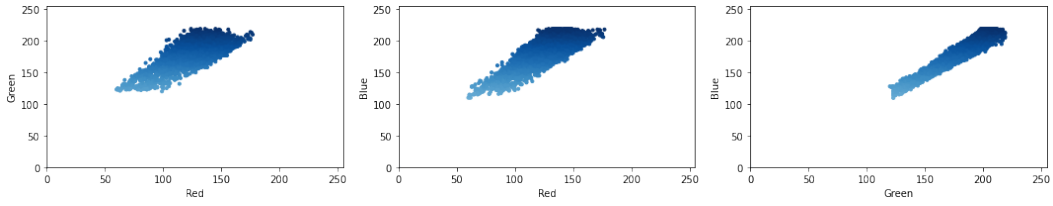
(a) Light Blue Marilyn



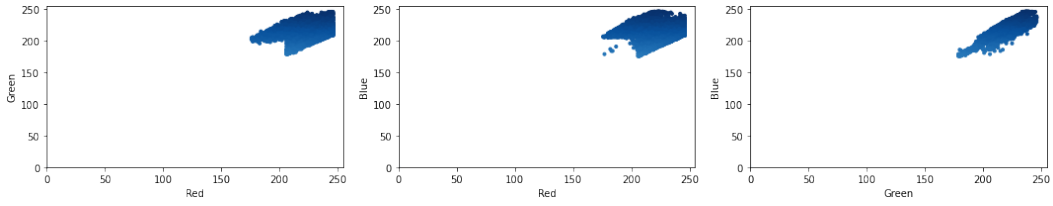
(b) Sage Blue Marilyn



(c) Orange Marilyn



(d) Red Marilyn



(e) Turquoise Marilyn

Figure 17: Red, Green, and Blue Space of the Eye-shadow Portion of the Five Marilyn Images

8 Cluster Analysis

8.1 K - Means Clustering

Separating colors from images is a process done through K-Means Clustering Algorithm. K-means clustering is part of Machine Learning algorithms. Through K-means algorithms, we identify the k number of centroids and then use that to allocate each data point to the nearest cluster. We cluster the pixels of an RGB image. Given an MxN size image, each consisting of three components: Red, Green, and Blue respectively. Using MxN pixels as data points and clustering those points with the K-means algorithm.

Image segmentation serves the purpose to pre-process before pattern recognition, extracting features, and image compression. In other words, the classification of images into different groups and the most popular method to perform is through K-Means clustering. In this case, we focused on two methodologies: **Elbow Method** and **Color Quantization using K-Means**.

8.1.1 Elbow Method

First, we load the different colors of Marilyn Monroe images and convert them into an image numpy array using *skimage.io.imread*. Since all images use RGB colors, the output will be MxNx3 (in 3D), and reshaping it into a 2D format allows us to move on to the next step. We decided to use 5 clusters to fit the image with K-means. Next, we will be looking at the Elbow Method as it gives us an optimal k number of clusters based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. The optimal k spot would be where SSE starts to flatten and form an elbow.

- **Inertia** The sum of squared distances of samples to their closest cluster center

We iterate the values of k from 1 to n and calculate the values of distortions for the number of k and calculate the distortion. Below are outputs for the inertia value for each k from 1 to 10 values.

ORANGE MARILYN

```
[5807847034.525303, 1979888189.2933738, 991582043.0862007, 628020031.0567105, 397914723.80024
42, 301509732.77330184, 254100141.23465735, 214111423.63525197, 184039934.83711052]
```

RED MARILYN

```
[10654178902.54937, 4617420955.633598, 2486517366.0084763, 898899672.3525758, 620791054.67899
13, 455341572.70740956, 374574729.2625363, 308156470.999598, 274051942.95113516]
```

LIGHT BLUE MARILYN

```
[5072722864.853047, 2047488416.0076227, 878122009.477684, 428861415.52970546, 266341898.73591
208, 209836011.31203902, 165901122.81421638, 127537411.14232743, 109204795.82137708]
```

SAGE BLUE MARILYN

```
[7691930446.7559185, 3674485250.216593, 1127885659.561123, 698224088.7437097, 402725867.6705
4, 315944856.0542464, 276238486.2050581, 240528461.88016295, 212274551.43393242]
```

TURQUOISE MARILYN

```
[9275531278.38882, 4304107498.654599, 2087472522.7380311, 1031915613.9329455, 732267105.83992
59, 528578805.96843356, 437416760.64822567, 380825512.5497427, 326542110.6340509]
```

Figure 18: Inertia Value for each k

We then plot those inertia values with respect to each k in a graph. There are five plots in total, each representing the distortion values on the x-axis and the number of clusters on the y-axis for each shot of Marilyn.

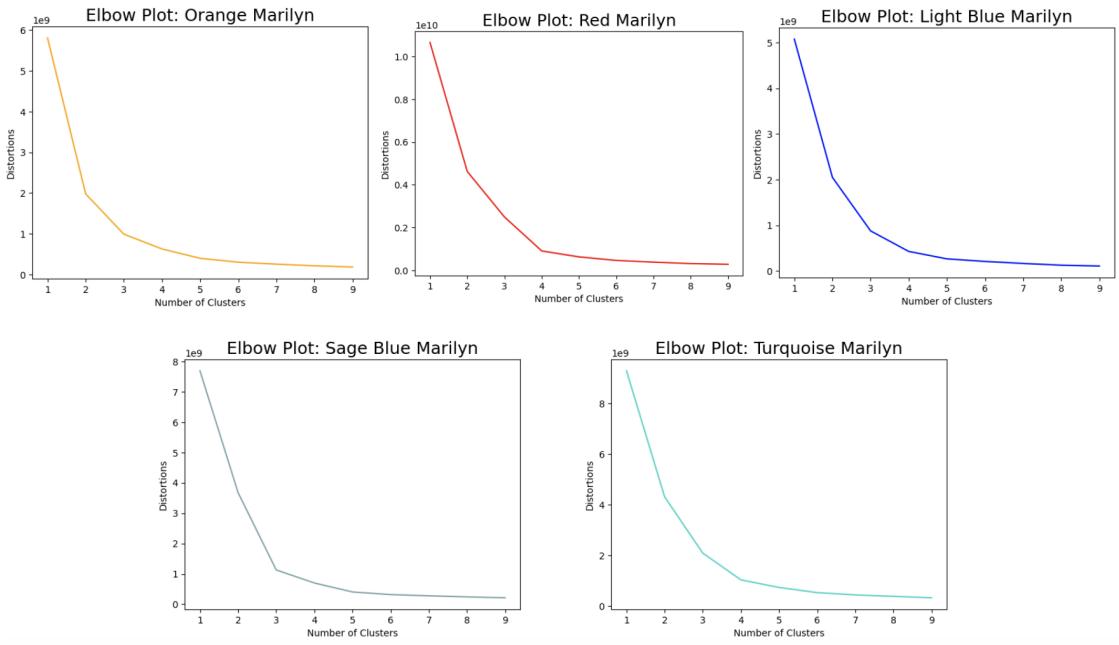


Figure 19: Elbow Plot

From the above plots, we can observe that the plot shows a gradual decrease in the distortion value as the number of clusters increases. However, at k = 4, the plot flattens and forms an elbow, indicating a significant drop in the distortion value. This suggests that k = 4 is the optimal value for clustering, as it provides a balance between capturing distinct color palettes and minimizing distortion.

8.1.2 Color Quantization using K-Means

Moving on to the next step, since we found the optimal k value through the K-mean algorithm, the next thing we could do is perform Color Quantization as a clustering problem. It reduces the number of colors used in an image while also maintaining the visual appearance of the original image. In other words, it is a form of cluster analysis. We would like to quantize the images such that they contain just K colors instead of RGB colors. To do so, we consider the R,G,B values and reshape the image array by flattening out the color arrays so each point represents RGB values as a data point. We perform color quantization for the images using the K-means clustering algorithm with different K values (K = 2, K = 4, and K = 8) (see Figure 20 below).

Below, we can see the original image and color-quantized images plotted next to each other to see the comparisons. There are significant differences in colors in different images with different K-values, the higher the K - value the closer it is to the original image. As the K-value increases, more distinct colors are captured, resulting in a finer representation of the original image. Conversely, lower K-values yield images with fewer colors, leading to a more simplified and less detailed representation. The significant differences in colors across the images highlight the impact of the K-value on the color quantization process.

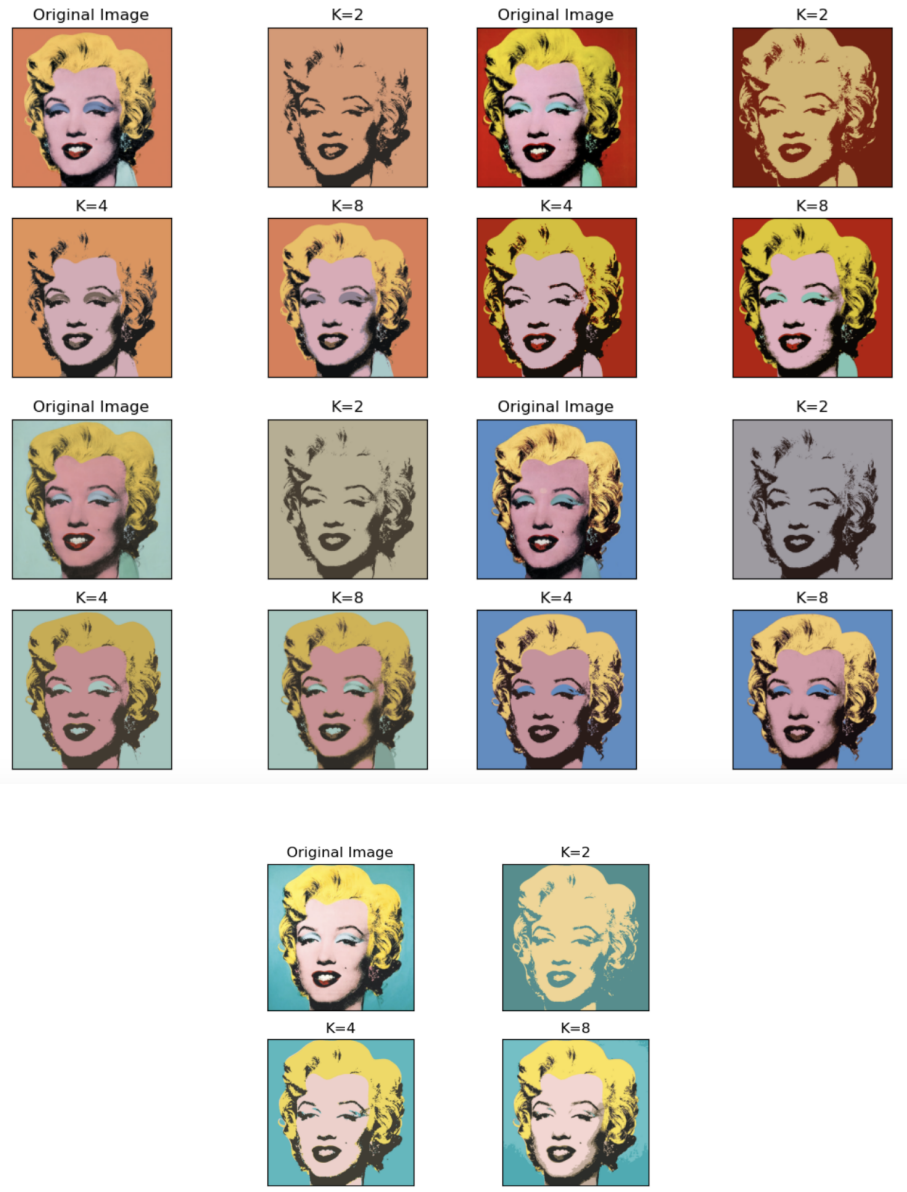


Figure 20: Color Quantization with K-Means: K=2, K=4, K=8

9 Correlation Testing

We first gathered data on the sold prices of the Marilyn paintings. These prices serve as a proxy for market valuation, reflecting the perceived worth of the artworks based on various factors such as artistic significance, demand, and market conditions. By aligning the quantitative features extracted from the paintings with their corresponding sold prices, we can investigate the relationship between these variables and see if there's a possible higher demand for the paintings based on their colors. Light Blue Shot Marilyn sold for $5,000 [2], Shot Red Marilyn sold for $4.1 million [3], Orange Marilyn sold for $17.3 million [8], Turquoise Marilyn sold for $80 million [5], and Shot Sage Blue Marilyn sold for $195 million [6].

To analyze the correlation, we extracted quantitative features from a collection of Marilyn painting images. In this case, we focused on three different methodologies: **Average Pixel Intensity, Haralick Texture Features,** and **Color Histograms**.

9.1 Average Pixel Intensity

The first approach measures the average pixel intensity of each Marilyn Monroe painting. By calculating the Pearson correlation coefficient, we determine the strength and direction of the linear relationship between the average pixel intensity and the sold prices.

The obtained correlation coefficient is 0.596779, and the p-value is 0.288043. The correlation coefficient of 0.596 suggests a moderate positive correlation between the average pixel intensity and the sold prices of Marilyn Monroe paintings. However, the p-value of 0.288 indicates that the correlation is not statistically significant at conventional levels, suggesting that other factors may contribute to the variation in sold prices.

Name	Value
Correlation coefficient	0.5967788188134844
p-value	0.28804298843678705

Table 6: Pearson correlation coefficient and P-value

9.2 Haralick Texture Features

The second approach utilizes Haralick texture features, which capture the texture properties of an image. Spearman correlation coefficient is computed to assess the monotonic relationship between the texture features and the sold prices. Because of the multiple quantitative values given by the texture features, the correlation is best described in matrix form.

The correlation matrix of the Haralick Texture Features and Marilyn Painting Sold Price summarizes the relationships between 14 different texture features and the corresponding sold prices of the paintings. These texture features include Contrast, Dissimilarity, Homogeneity, Energy, Correlation, Entropy, Autocorrelation, Cluster Shade, Cluster Prominence, Haralick's Correlation, Maximum Probability, Sum Average, Sum Variance, and Sum Entropy. The figure of the correlation matrix analyzes the correlation between these texture features and the prices of the five Marilyn paintings.

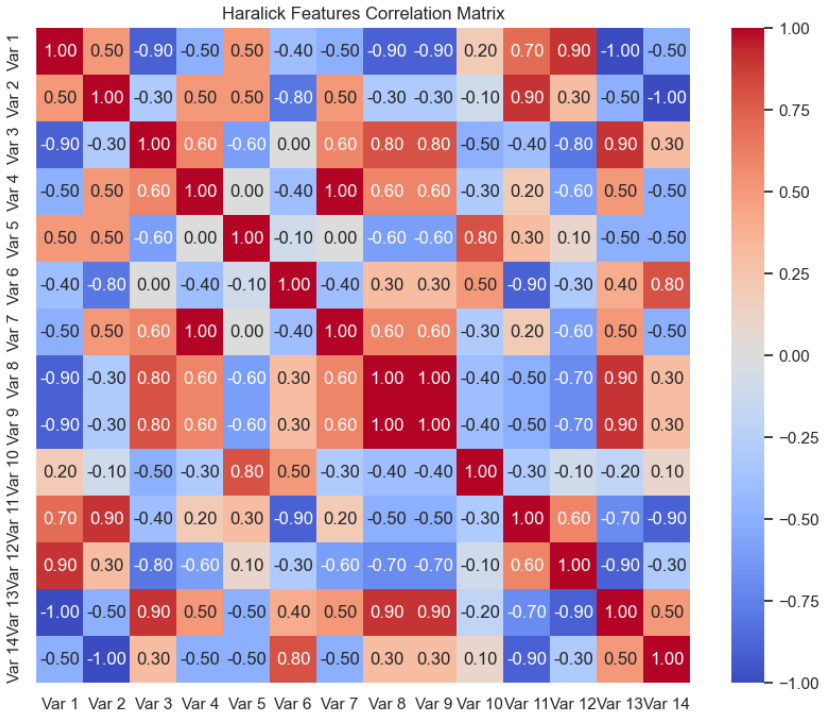


Figure 21: Correlation Matrix of Haralick Texture Features and Painting Sold Price

The Spearman correlation coefficient matrix reveals various correlations between the Haralick texture features and the sold prices. Positive coefficients indicate a positive monotonic relationship, while negative coefficients suggest a negative monotonic relationship. The associated p-values provide a measure of the statistical significance of the correlations.

9.3 Color Histograms

The third approach involves analyzing the color histograms of Marilyn Monroe paintings. Spearman correlation coefficient is used to evaluate the monotonic relationship between the color histograms and the sold prices.

The Spearman correlation coefficient matrix for the color histograms resulted in many missing values (NaN). This suggests that certain pairs of paintings did not have computable correlations. This could be due to factors such as identical or constant color histograms, resulting in undefined correlation coefficients.

Analyzing the correlation between quantitative values extracted from Marilyn paintings and their sold prices offers valuable insights into the relationship between the intrinsic characteristics of the artworks and their market value. By employing techniques such as color histogram analysis and Spearman correlation, we can identify patterns and associations that shed light on the factors influencing the pricing of these paintings.

9.4 Discussion of Correlation Analysis

Overall, the analysis of the correlation between quantitative values and sold prices of Marilyn Monroe paintings provides interesting findings. The first approach, measuring the average pixel intensity, reveals a moderate positive correlation of 0.596 between the average pixel intensity and the sold prices. However, the lack of statistical significance with a p-value of 0.288 suggests that other factors may contribute to the variation in sold prices. The second approach focuses on Haralick texture features, capturing the texture properties of the images. The correlation matrix of the Haralick Texture Features and Marilyn Painting Sold Price shows various correlations between the 14 texture features and the corresponding sold prices. Analyzing these correlations provides insights into the relationship between texture characteristics and the pricing of the paintings. The third approach involves analyzing the color histograms of the Marilyn Monroe paintings. However, the Spearman correlation coefficient matrix for the color histograms resulted in missing values (NaN), indicating that certain correlations were not computable due to factors like identical or constant color histograms.

10 Conclusion of Data Analysis

Each of the paintings were sold, and the orange and sage blue Marilyn were sold for the most. Is there a correlation between these paintings and their sold price?

The analysis of the correlation between quantitative values and sold prices of Marilyn Monroe paintings yields mixed results. While the average pixel intensity shows a moderate positive correlation, the statistical significance is not confirmed. The Haralick texture features and color histograms offer additional insights into the relationship, but the presence of missing data and non-significant correlations necessitates further investigation. While the average pixel intensity and some Haralick texture features show correlations, the lack of statistical significance and missing correlations emphasize the need for further investigation and consideration of additional factors. Therefore, from the information we gathered on the five Marilyn Monroe paintings, we cannot draw a correlation between the price that each of these paintings sold for and the actual painting. The sold price could vary due to the time that each painting was sold, inflation, or etc.; however, that would require a deeper analysis into the selling of these paintings.

What does K-Means Clustering tell us about the five Marilyn photos?

The analysis of K-Means clustering in image processing gives an insight into the color distribution and grouping of pixels within each image. It enables different applications such as color quantization, image segmentation, compression, and enhancement. Leveraging the algorithm's ability to distinguish similarities and differences in image colors is an effective way to analyze and manipulate for a wide range of purposes. Finally, the analysis above concluded that by applying the K-Means clustering algorithm and Elbow Method we found that the optimal number of clusters for all shots of Marilyn Monroe is 4. This is indicated by the formation of an elbow in the distortion value plot. In the color quantization process, a high K value will resemble the original image, while a low K value render with fewer colors to the original image. Overall, the result highlights the effectiveness of the K-Means algorithm in separating colors in images.

What are similarities and differences among the five Marilyn photos?

The color analysis provides insight into the RGB color distribution for the five Marilyn photos. Through the analysis of the dominant colors in the Marilyn series, we can see that the colors reflect Warhol's intentional artistic decisions, which are using vibrant and bold colors to explore the mass culture and consumerism. The specific colors chosen, such as sage blue, light blue, orange, red, and turquoise, reflect Warhol's affinity for visually engaging and impact works, and his unique style of design. Furthermore, by separating the whole image into different regions of interest, we would be able to analyze all the major colors individually. For the face, hair, and eye-shadow parts, just by viewing them, we may believe that all the colors are similar; however, differences can be seen when plotting them onto the RGB scatter plots. For instance, we observed that the background color for light blue Marilyn and sage blue Marilyn contains a large amount of green. This method of separating images into regions of interest allows us to gain a deeper understanding of the underlying structures of red, green, and blue colors within each individual color. It is important to note that there may be some inaccuracies in the RGB scatter plots due to the uneven borders between colors.

In summary, the relationship between color and market value should be understood in the context of subjectivity and personal preference. Different audiences and collectors may interpret and prefer specific color choices differently. However, Warhol's use of vibrant colors in his Marilyn paintings remained consistent with the broader aesthetic preferences associated with his style, contributing to their enduring appeal in the art market. The visual impact and intentional color compositions in Warhol's works played a significant role in establishing their market value and securing their place as iconic pieces of art.

Acknowledgments

We would like to thank our TA, Shuting Liao, for her guidance and support throughout this project. Additionally, we extend our appreciation to our groupmates for their invaluable contributions towards the success of this project.

References

[1] Color Terms. (n.d.). John December. Retrieved June 11, 2023, from https://johndecember.com/html/spec/colorterms.html

[2] Peter Brant on Josh Smith, Urs Fischer and David AltmejdlCollector's Eye. (2011, May 7). The Wall Street Journal. Retrieved June 8, 2023, from https://www.wsj.com/articles/SB10001424052748703834804576301351377559400

[3] Reif, R. (1989, May 4). A Warhol 'Red Marilyn' Sets Record at Christie's (Published 1989). The New York Times. Retrieved June 8, 2023, from https://www.nytimes.com/1989/05/04/arts/a-warhol-red-marilyn-sets-record-at-christie-s.html

[4] Schmidt, B. (2022, May 10). A Visual Critique of Warhol's Shot Sage Blue Marilyn, 1964 — The Interior Review. The Interior Review. Retrieved June 8, 2023, from https://www.theinteriorreview.com/story/2022/5/10/critically-assessing-warhols-shot-sage-blue-marilyn

[5] Steve Cohen, billionaire hedge fund manager, is lending works to US and UK museums. (2007, September 30). The Art Newspaper. Retrieved June 8, 2023, from https://www.theartnewspaper.com/2007/10/01/steve-cohen-billionaire-hedge-fund-manager-is-lending-works-to-us-and-uk-museums

[6] Ulaby, N. (2022, May 9). A Warhol 'Marilyn' brings a record auction price, $195 million. NPR. Retrieved June 8, 2023, from https://www.npr.org/2022/05/09/1096617152/a-warhol-marilyn-brings-a-record-auction-price-195-million

[7] Why Do People Like Andy Warhol So Much? (2023, March 15). Wayne Arthur Gallery. Retrieved June 11, 2023, from https://www.waynearthurgallery.com/why-do-people-like-andy-warhol-so-much/

[8] The Wild History of the Warhol Marilyn That's Set to Fetch $200 Million. (2022, May 2). Artsy. Retrieved June 8, 2023, from https://www.artsy.net/article/artsy-editorial-wild-history-warhol-marilyn-set-fetch-200-million

[9] Colour Quantization Using K-Means Clustering and OpenCV. (n.d.). Retrieved June 11, 2023, from https://www.analyticsvidhya.com/blog/2021/07/colour-quantization-using-k-means-clustering-and-opencv

[10] How to Use the Elbow Method in Python to Find Optimal Clusters .... (n.d.). Retrieved June 11, 2023, from https://www.statology.org/elbow-method-in-python