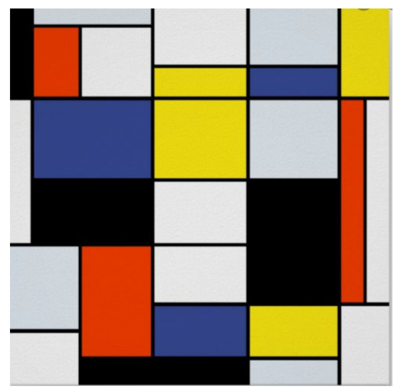
Introduction

The problem we have spent 10 weeks trying to address is whether it is possible to quantify artistic style, and if so, how would we do it. To break down this problem, we took artist Piet Mondrian's artwork and analyzed images of it. Mondrian's artwork is particularly appropriate for this analysis because the artistic style varies so much (from realism to abstracts). This gives us many more parameters through which we can quantify and categorically identify what artistic style is and how to apply algorithmic analysis to it. Additionally, there are many methods for analyzing artistic style and image data is well suited to those methods. Scans of Mondrian's artwork allow us to utilize those methods and facilitate analysis.

A realism piece by Mondrian



An abstract piece by Mondrian

The potential shortcomings are the quality of the data (we don't know how high-resolution the scans of the artwork are and if they are of a resolution that might impede in the analysis of the artwork). Additionally, I believe the data itself is a shortcoming in analyzing it. I believe a key factor in artwork is the intention behind the artwork, which is something that is difficult to identify with an algorithm or quantitative method. Therefore, analyzing an image, with no concrete notion of intention behind it, becomes complicated. In this project, however, we start by identifying what makes art abstract. By identifying those features before starting analysis, we were able to effectively isolate features of the images that we were then able to use in our analysis and quantification methods.

The Data Generation Process

The data originates from the Catalogue Resonae. There are no legal issues pertaining to accessing it. Looking at the robots.txt file for this website, we can see that all robots are permitted to access it, telling us that we are free to scrape from it.

The schema with which I stored my data is a dataframe containing information on where the image can be found (the image source) and other metadata for the image. This includes the year in which it was published, the material it was published on, and the title of the work. I think this is a good method of data storage because it allows image data to be understood in a tabular form, which is easy in terms of parsing the data and analyzing it.

I think this pipeline is applicable to similar data sources because of the way other sources that store Mondrian images are structured. For example, the artnome source structured data in a similar way and the data ingestion pipeline could be modified and used there as well. Other similar data sources would be other online catalogues of art data, including but not limited to sources that scan

Historical Context

Quantifying, or understanding, art had been widely approached in the past. Art historians and artists alike have been viewing art and attempting to derive meaning and understanding from it. This particular problem has been addressed with the same data, which I will attempt to replicate. However, I think a valiant attempt at deriving an answer to a deeper version of this problem, one that attempts to truly discern the meaning and intention behind artwork, has been happening for a very long time in the field of artistry. We are attempting to quantify Mondrian's art style, a problem that has been attempted in the past. We are generating data through an online catalogue of Mondran's pieces, through which I can get an image of each of his pieces and relevant metadata.

In the past, Manovich, Carey, and Wang have attempted to quantify artistry by using algorithms to analyze Mondrian's artwork. We will do the same, analyzing different works by Mondrian and attempting to come up with an algorithm to quantify the artistry behind it.

Pre-EDA: Scraping and Cleaning

Scraping and cleaning the data proved to be more challenging than I originally thought it would be. The website I scraped the data from (including both the primary image data and the secondary metadata) was organized but not completely consistent with the information that was presented. Isolating the data that needed to be scraped in a way that was consistent across all image pages proved to be challenging.

Another challenge was cleaning the scraped data, especially for the Year that the art was produced. Cleaning the data required extraction from a messy string, especially so for the "year" metadata. Extracting this was an interesting process that took many varied attempts. The method I finally went with was taking only the values that could be converted into integers from the entire messy string, giving me the year that the art was produced. Although this method was not perfect, it did produce the highest "value to NaN" ratio of all the methods that I tried.

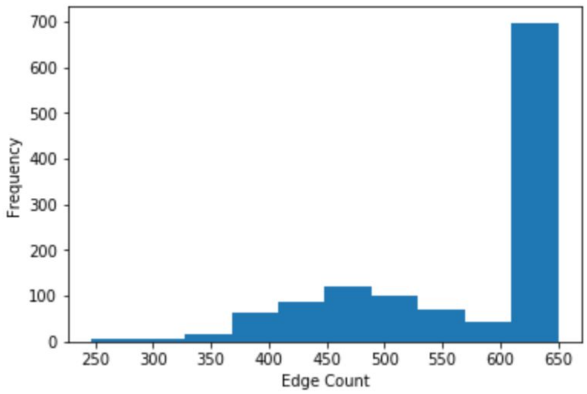
Pre-EDA: Feature Extraction

Once the data was scraped and cleaned, I was able to extract more features from the images that had been scraped. Using the scikit-image library, I went through each image and initially extracted the basic features that described it: the pixel count, saturation, brightness, edge count, and grayscale variance. Extracting these features from the images requires the images to be turned into integer arrays, which is something that the scikit-image library does. It was interesting seeing the conversion happen and to watch the quantification of images in real-time.

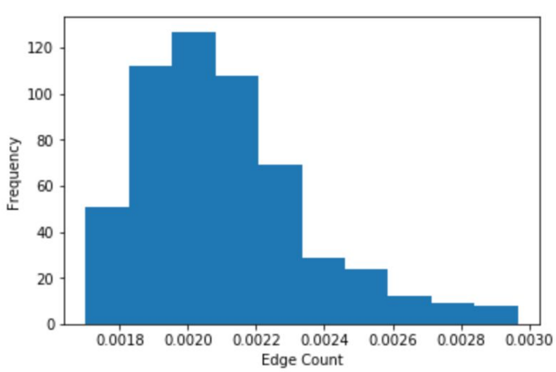
EDA: Univariate Analysis

Once the data was scraped and cleaned, and basic features were extracted, it was possible to look at trends in the data as a whole. Looking at the data, we can see a lot of variation in the Edge Count as well as the Brightness and Saturation. We can quantitatively distinguish between drawings and paintings through Saturation and Edges. Drawings have a lower saturation level as they often use fewer colors or are mostly grayscale in nature. While cleaning, I separated images with a low saturation level from those with higher saturation levels. The metadata I scraped for each image also contains information on what type of work every image is (painting or drawing), which gave me further confidence when splitting up the images. Most of the cleaning I did was making sure that what I had scraped was in an appropriate format for analysis.

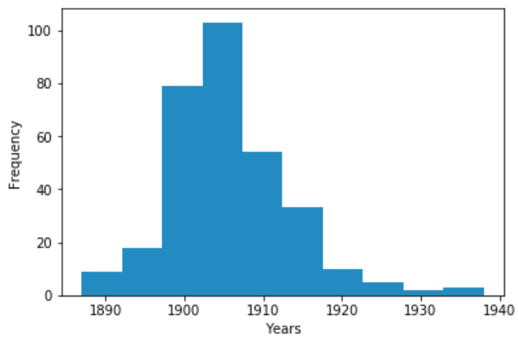
We can see that the original Edge Counts have a strange distribution.



With this in mind, and after some feedback, it became clear that I would need to normalize this value. I did this by dividing the edge count with the Pixel Count produces a better value. I filtered my dataframe with values from the middle 75% of this value. This produced a much more normalized Edge Count frequency.

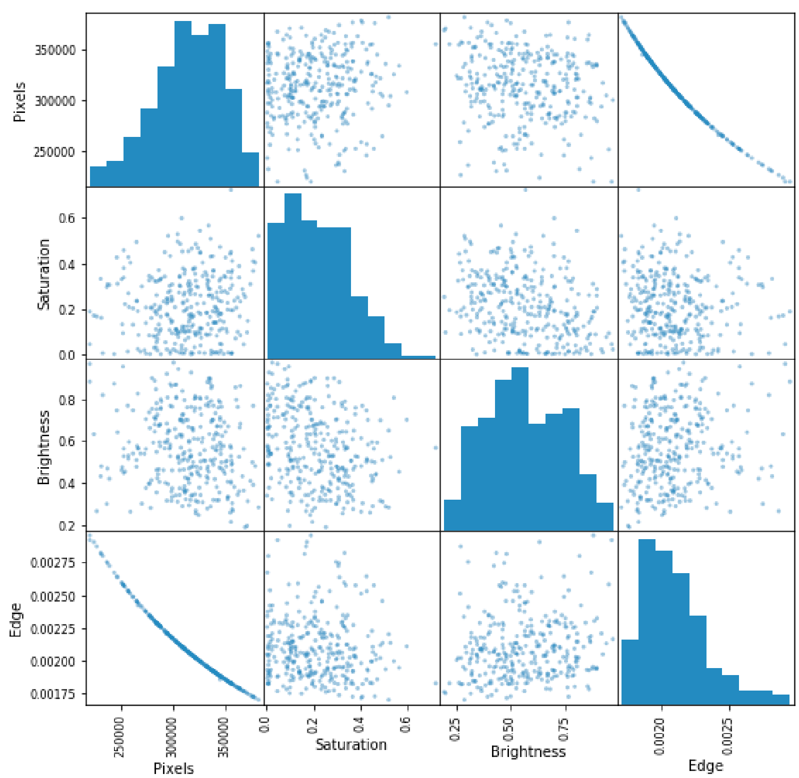


A visualization of the distribution of the "Years" metadata displays how accurate or inaccurate the scraping method had been. Looking at the distribution, we can see that most of the images are from 1900-1910, which is logical considering that that had been the mid/peak of Mondrian's career.

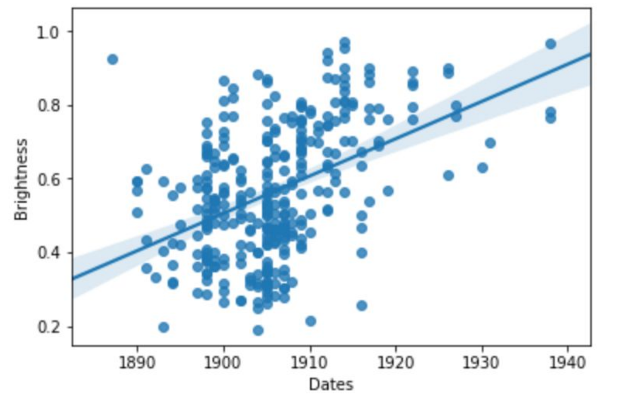


However, the scraping methods were imperfect in the sense that there are more images between 1920-1940 that were not scraped.

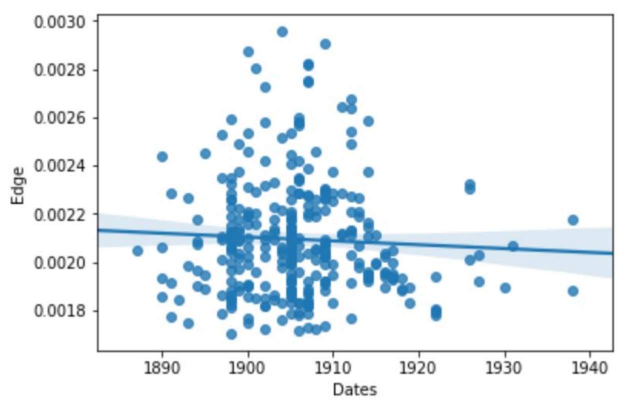
EDA: Bivariate Analysis



Looking at Dates vs Brightness, we can see that Mondrian's work tends to become brighter as time goes on. This makes sense thinking visually about the progress of Mondrian's work, which started with darker realism pieces and progressed to very bright, saturated abstract pieces.



Comparing Edges and Time shows a subtle downward trend, indicating that Mondrian produced works with fewer edges later in his career. This fits with what we know about his art as well. Realism pieces tend to incorporate more lines and contours, while the abstract pieces were more simple.



Data Inclusion Justification

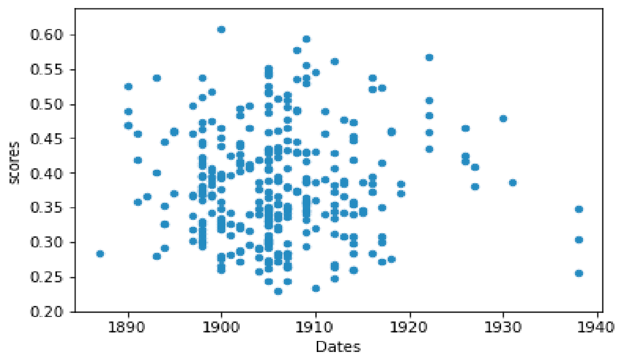
When thinking about resolution, it is difficult to discern a minimum resolution necessary to participate in this analysis because we are breaking down every image into a series of arrays. When quantifying an image, resolution seems to become less important. Furthermore, all images were scraped from the same online database and so have extremely comparable levels of resolution. I think to determine how low of a resolution is “too low”, it would depend on what type of analysis is being done. In our analysis, we are mostly looking at color variation, edges, and brightness variation. I believe these aspects are not hindered by the level of resolution that the paintings from the catalogue raisonne provides. However, as we saw with the edge count, normalizing values in the dataframe lend some more reliability to analysis. After normalizing the edge count value by dividing by pixel count, I was able to take the mid-75% of values based on resolution.

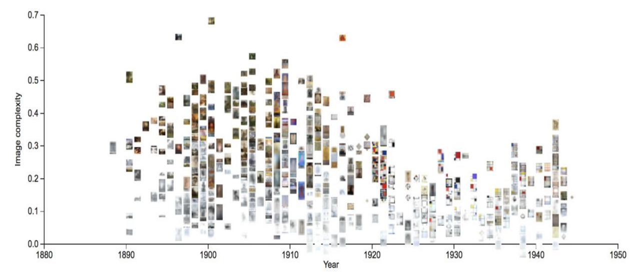
I initially considered using all the images that I had scraped. However, after looking at the distribution of different values, it became clear that I would need to normalize and take the mid 75% of certain values so as to create a more coherent dataset. Keeping this in mind, I initially considered leaving out the drawings and only working with the paintings. However, I realized that this would not assist in quantifying artistic expression at all as Mondrian’s drawings are so integral as a precursor to his abstract paintings. Mondrian’s stylistic development went from drawing to painting, so it would not be fair to not include both those types of work in the analysis of his stylistic development.

Replication: Explanation of Methods

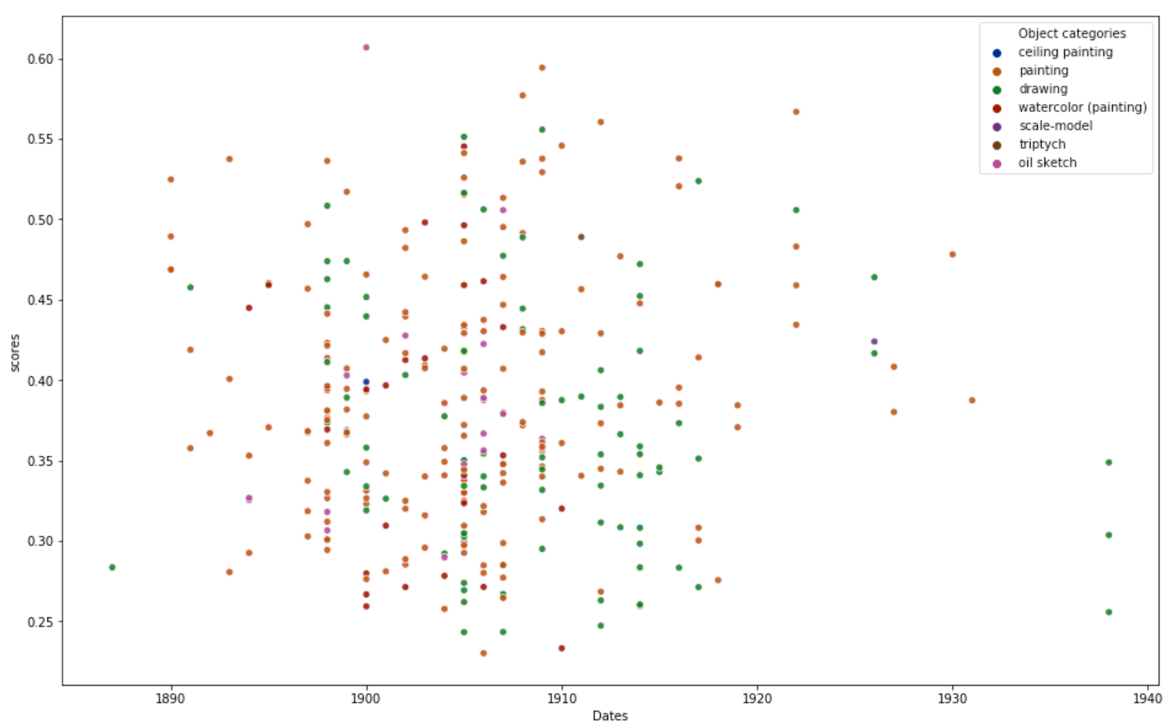
I attempted to replicate Jason Bailey’s quantification of abstract style. Bailey and his colleague decided on this formula to quantify abstraction: complexity_score = (color_score + variance_score + edge_score)/3 , where the color_score is the number of unique RGB values, the variance_score is the mean of the mean gray_scale variance by row, and the edge_score is the number of contours detected in an image. The variance score takes every row of an image and compares the row of values to the mean of that row. The final score takes the mean of all these variances, thereby finding the average variance in shade in the entire image by row. The edge score detects how many contours are there in an image, or how “lined” or “structured” an image is. I believe is a good measure of abstraction as it can be seen that Mondrian’s more abstract paintings have fewer lines or shapes in them. His realistic paintings, on the other, are more complicated, with a larger number of lines and shapes in them.

My replication creates roughly the same shape as the Artnome analysis. I believe the issues with scraping the appropriate year from each image caused for data to be dropped between the years 1920-1930, which affected the analysis as there were fewer images during that time to begin with.





Color-coding the data by the image's category we can see that paintings done later in Mondrian's career have a high complexity score, while drawings done at the same time tend to have a very low complexity score.



Replication: Analysis of Shortcomings

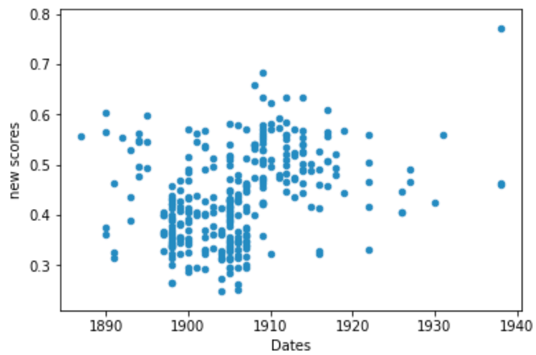
The color_score portion of the complexity_score seems a little off. Quantifying abstraction in part by looking at the number of unique colors in a work does not seem completely accurate. For example, a picture with slightly varying shades of one color would have a high color complexity score, which does not really make sense because having a lot of variations of the same color does not indicate a greater level of complexity. This is why I think some paintings received a higher complexity score than they perhaps should have been allotted. Furthermore, while I do not think that measuring greyscale variance is necessarily a bad measure of complexity, I do think that measuring entropy would be better. As noted in the documentation of image processing library, scikit-image, the entropy of an image "is related to the complexity contained in a given neighborhood." I believe entropy would be a better measure of "disorder", and therefore maybe abstraction, than the greyscale variance. Looking at entropy instead of the greyscale variance might produce a higher complexity score for a piece, which might be more accurate for that piece.

Moving Beyond Replication

Revisiting the complexity score, I wanted to improve upon some aspects that I thought were shortcomings in the original work. This included the color_score, which is made up of the count of unique RGB values in the image. Building on this, my color_score was made up of the average variance between each R, G, and B value. To do this, I found the unique R, G, and B values as arrays. I then found the variance within each, which was essentially finding the average and comparing the array of values to that average. This was done for each of the R, G, and B unique value arrays. Finding the average variance of each individual portion of what makes up a color allows, I believe, for more nuance in classifying the variance between colors. This solves the problem of counting similar, but slightly different, colors as equally variant to extremely different colors. I changed the calculation of the color_score to be this value.

Looking at the variance_score, and the shortcomings that came with it, I decided to implement my entropy idea to replace the original variance_score. I found the entropy of the image, using scikit-image's in-built entropy function and took the mean of that. This would give me the mean "state of disorder" or variability in the image, which I think makes a lot of sense when calculating how abstract an image is.

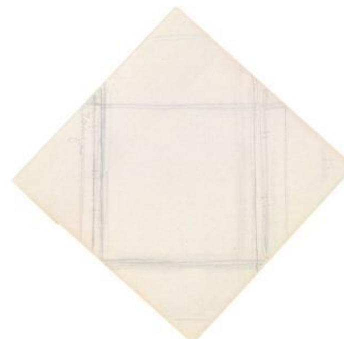
Therefore, my new abstraction score was: (mean_RGB_variance_score + entropy_score + edge_score)/3. This score would serve to determine how "abstract" an image is.



This was interesting as the scores were higher than before. Images that previously received a very low score now received a much higher one.

Color-coding by the category of the images shows us that a majority of Mondrian's work was paintings and drawings, which confirms what we have already observed by looking through his work. In direct opposition to the original complexity_score, though, we can see that paintings on average have a lower score than drawings.

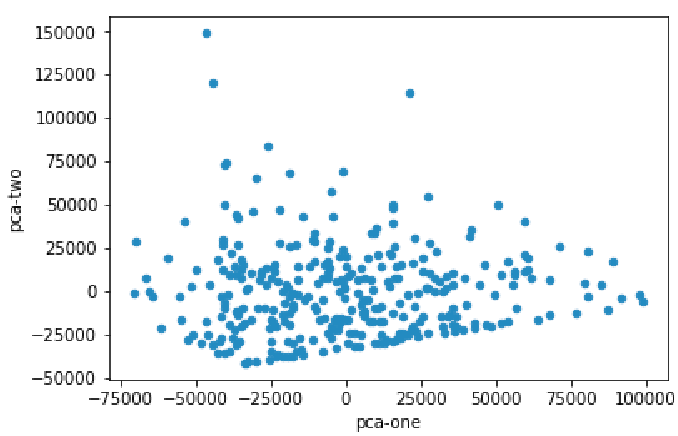
Looking at an image that has the lowest abstraction_score (left) and comparing it to the image with the highest abstraction_score (right), we can see that the low scoring image is very realistic while the high scoring image is quite abstract. This is an indicator that the changes made to the complexity_score resulted in a better indicator of when an image is abstract or not.



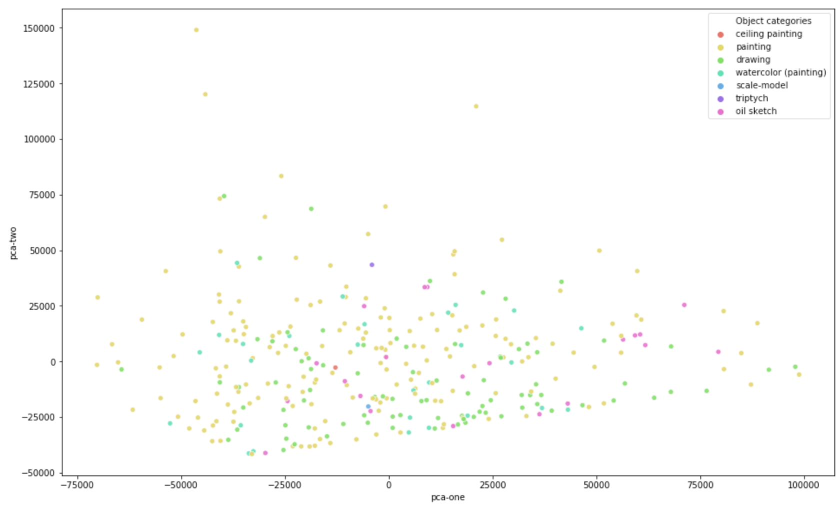
Lowest abstraction_score *Highest abstraction_score*

Dimensionality Reduction

Utilizing dimensionality reduction to the set of integer features, I extract the more “important” or prominent features. I ran a Principle Component Analysis on the set of integer features, which was made up of the Pixel Count, Saturation, Brightness, Edge Count, Greyscale Variance, Unique Colors, and Entropy Variance.



I then visualized the results of the PCA as well as the results color-coded by the category of the image.



Conclusion

Our major goal behind this project was to quantify abstraction. Going into this project, we discussed what made up abstraction and whether it was able to be quantifiable at all. The mixture between art and quantitative methods is fascinating because it forces the quantifier to consider how much of art is tangible and how much of it is left to interpretation or relies on human thought and connection.

With this replication, we can see that quantification of art is possible. With the complexity_score (Artnome replication), we were able to identify particular features of images that would contribute to how complex an image is, and therefore how abstract it may be. Building on this, the abstraction_score (my variation on the complexity_score) modified the complexity_score to isolate features that I believe lent themselves well to how abstract an image is.

However, I believe that in attempting to quantify abstraction, it also became very clear which aspects of the image were innately human. A person looking at an image might "rate" its abstractness by asking themselves how confusing or unrealistic the image was to them. When an image does not look like a scene we have seen in real life, we immediately think of it as "abstract". The complexity_score we replicated and the abstraction_score I extrapolated aimed to be able to do this. While these scores did a good job at assessing the features that make art abstract, I believe I can conclude from this project that no amount of feature extraction and quantification of imagery will be able to quantify art with complete accuracy.