Anas Moazzam
Pascal Wallisch
Introduction to Data Science (DS-UA-112)
May 2023

Final Capstone Project

Let's begin with certain protocols I followed during the course of my analysis. Since there were columns with NaN values, I needed to deal with missing or incompatible data. To minimize loss of data, I created two dataframes from the Art and Data csv files in which for every question, a new dataframe was created with the necessary columns. After, I then dropped the NaN values from the dataframes specific to the question. Furthermore, when a PCA was needed to be done, I could properly conduct it without running into NaN errors.

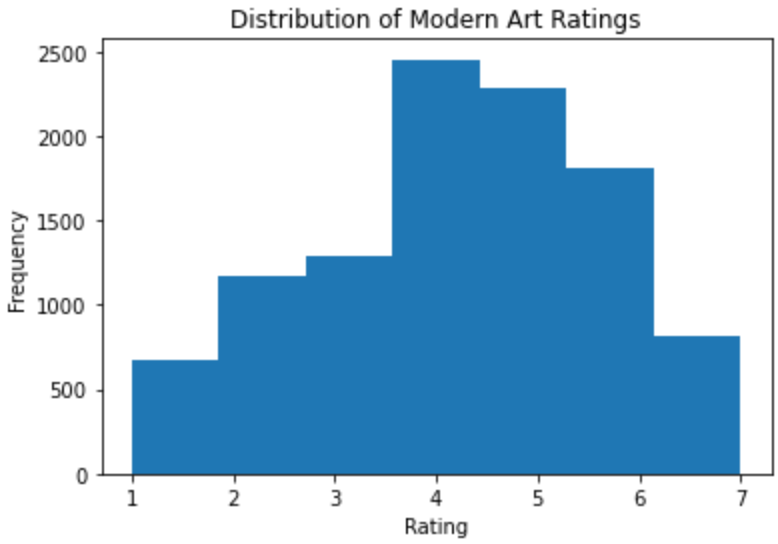
Since the Data csv file had no headers, I had to create headers myself. I assigned the first 91 columns with the number corresponding to the art piece, then did the same but specified it was 'energy' ratings, then added dark questions, action questions, self questions, and demographic questions at the end.

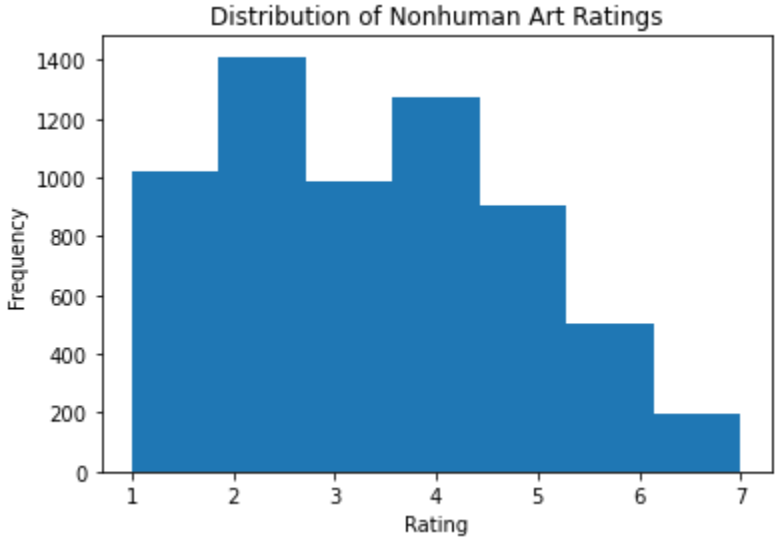
1. **Is classical art more well liked than modern art?**

I interpreted this question in a simple manner. I compared the mean and median of classical and modern art styles. I compared both to check in case the mean was affected by any outliers, although I didn't suspect any. To do this, I created a' classical' variable that contained any rows that were classified as 'classical', and the same for modern art. I realized that the first 35 columns were classical and the next 35 were modern. I used this to create two separate variables for classical and modern art and calculated the mean and median for each variable. Thus, I concluded that classical art is, on average, more well liked or higher rated than modern art.

2. **Is there a difference in the preference ratings for modern art vs. non-human (animals and computers) generated art?**

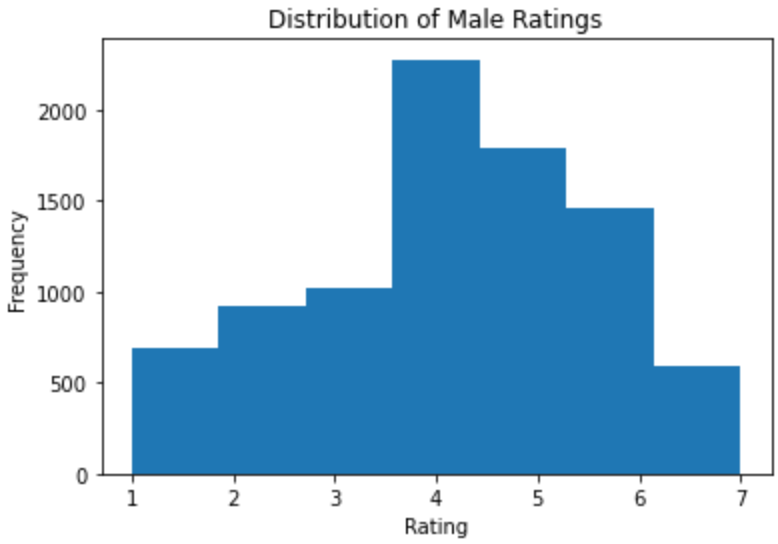
I started off by following the same process from question 1 and calculating the mean and median of nonhuman art and compared the two values. Afterwards, I decided to check the distribution of the ratings of modern and non-human art by graphing their ratings. This is to see whether they were normally distributed or not in which I found modern art seemed to be normally distributed but I couldn't tell if nonhuman art was distributed normally as well. Thus, I decided to use a Mann-Whittney U test to test whether the difference between preference ratings were statistically significant or not. Since I got a p-value which was essentially 0, I concluded that there is a statistically significant difference in the preference ratings for modern art vs. non-human (animals and computers) generated art.

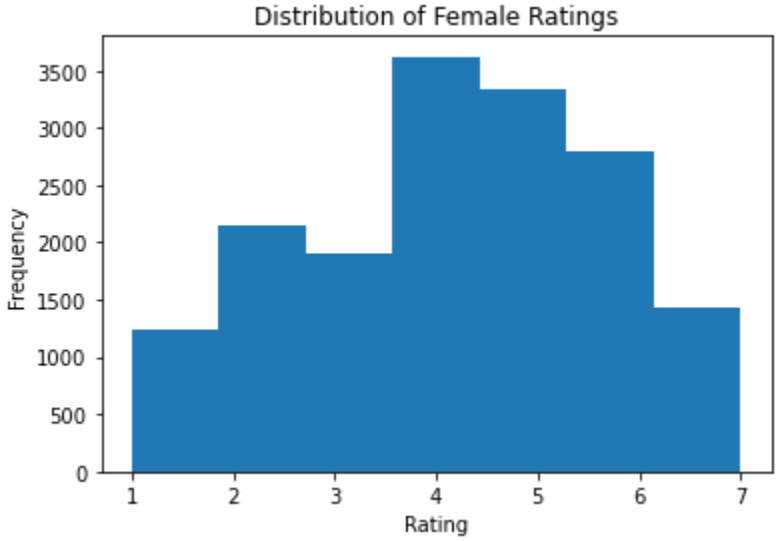




3. **Do women give higher art preference ratings than men?**

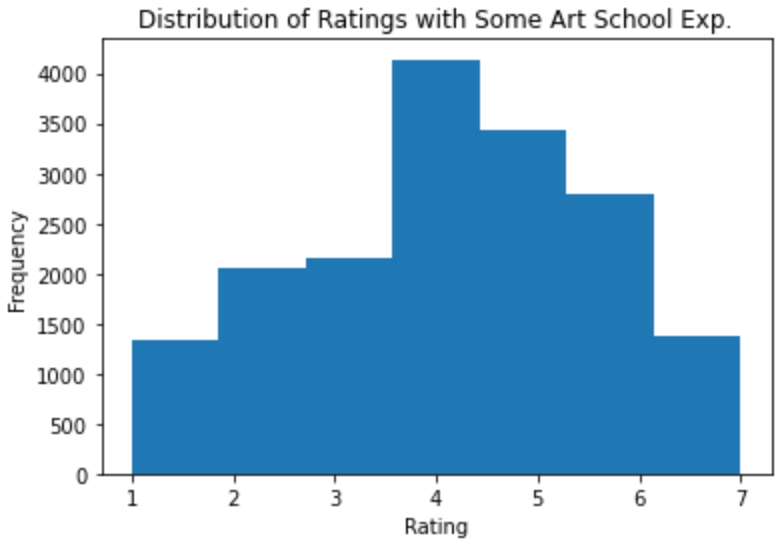
To get the necessary data, I created a new df with the art preference ratings as well as the gender demographic column. Although there were NaN values, I decided that none of my methods would be affected by them inside the df so I decided to leave them in. I created new dataframes for male and female ratings and then checked the distribution again. The shape was much clearer for these distributions, hence, I decided it would be appropriate to run a T-Test between the means of the distributions. In this specific instance, the mean was higher for preference ratings of women, but this was deemed not statistically significant. In general, I don't expect there to be a difference between the preference ratings of men and women.

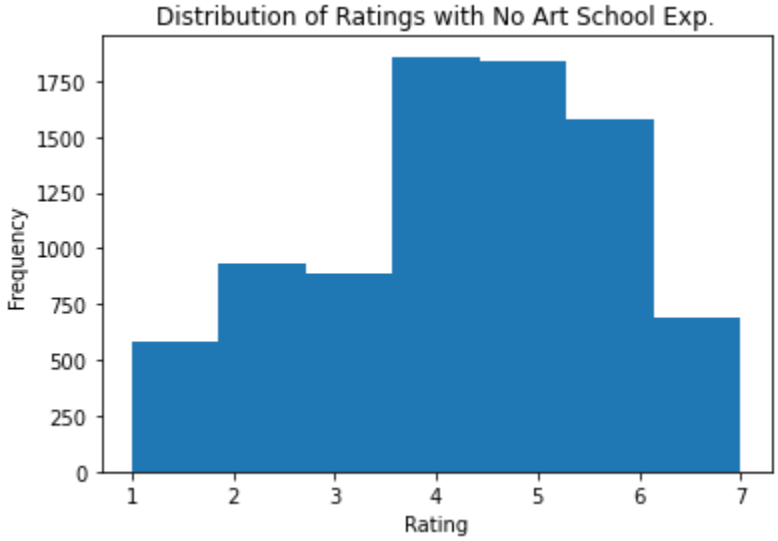




4. **Is there a difference in the preference ratings of users with some art background (some art education) vs. none?**

To get the necessary data, I created a new df with the art preference ratings as well as the artistic education demographic column. I created a variable of ratings with any artistic schooling (so any values of 1, 2, or 3) and calculated the mean. I did the same for ratings with no artistic schooling and also calculated the mean. I then graphed each distribution to check if they were distributed normally, which they both ended up being. This allowed me to use a T-Test to check whether the difference between the means were statistically significant. Since my p-value was essentially 0, I can conclude that there is no significant difference in the preference ratings of users with some art background (some art education) vs. none.





5. **Build a regression model to predict art preference ratings from energy ratings only. Make sure to use cross-validation methods to avoid overfitting and characterize how well your model predicts art preference ratings.**

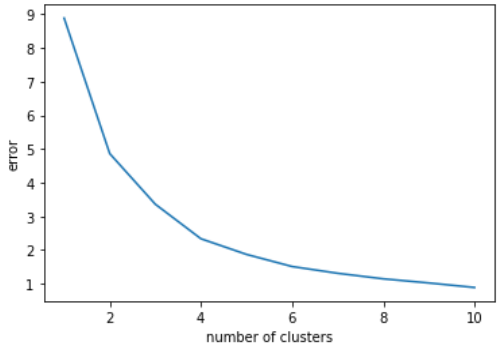
I began by creating a new dataframe in which I only included the art preference ratings and energy ratings. Next, I checked for any NaN values and in the new df, there were none. I decided to then use sklearn to create a test and train set from our dataframe with a test size of 20% of the data. I decided to run three different regression models: a normal regression model, a ridge regression, and lasso regression to minimize overfitting. I checked the performance of the models by checking the RMSE of each model. The RMSE of the normal regression model was 1.689, for the Ridge regression was 1.5144, and for the Lasso Regression it was 1.3653. Based on the RMSE of my regression models, the energy of the painting doesn't seem to be a great predictor of painting preference ratings. Our best RSME tells us that our model on average is off by about 1.37 units in ratings, which is substantial given the scale is 1-7.

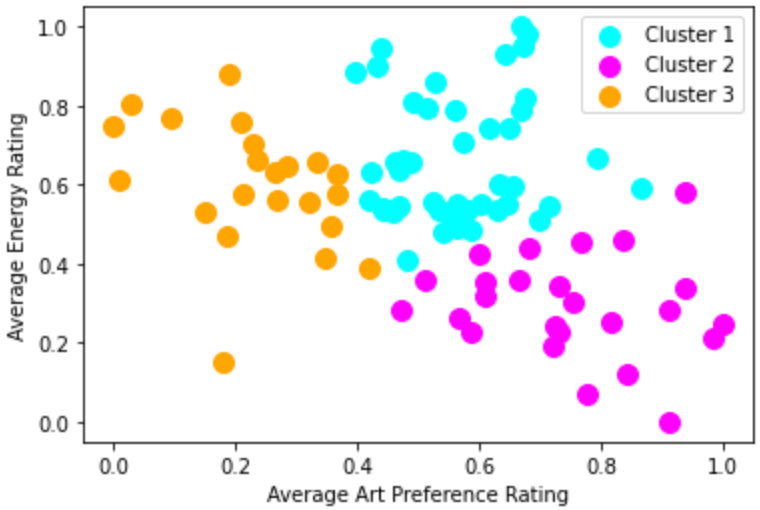
6. **Build a regression model to predict art preference ratings from energy ratings and demographic information. Make sure to use cross-validation methods to avoid overfitting and comment on how well your model predicts relative to the "energy ratings only" model.**

Similar to Question 5, I created a df with art preference ratings, energy ratings, but this time, included demographic columns at the end. I checked for NaN values and saw that there were about 20 rows with NaN values within the demographic rows. I decided to drop these rows and proceeded. I then created another test and train set from this dataframe with a test size of 20%. I also ran three regression models to prevent overfitting and find the best model. I used the RMSE again to judge model performance. The RMSE for the regular regression model is 2.1146, the RMSE for Ridge Regression was 1.8342, and the RMSE for Lasso Regression was 1.5563. Based on the RMSE of my regression models, energy ratings and demographic ratings don't seem to be great predictors of painting preference ratings. Our best RSME tells us that our model on average is off by about 1.5563 units in ratings, which is substantial given the scale is 1-7. This model does worse than our previous models most likely due to the increased amount of factors introduced in this model. These variables may be correlated with each other and thus not provide the best estimates.

7. **Considering the 2D space of average preference ratings vs. average energy rating (that contains the 91 art pieces as elements), how many clusters can you – algorithmically - identify in this space? Make sure to comment on the identity of the clusters – do they correspond to particular types of art?**

I started off by creating a new df by getting the columns the first 182 columns from my first data df. I then created separate variables for art and energy ratings, in which I calculated both the mean artistic preference and energy rating for each painting. I decided that to properly identify the clusters, I will be using K Means clustering. I first normalized the ratings for artistic preference and energy ratings. Then, I used the elbow method to calculate the optimal grouping for my K Means which turned out to be 3 clusters. This makes intuitive sense as we already have three different types of artistic styles. After running my K Means clustering, I labeled the original data with the cluster that was assigned and plotted their distribution. Although not exact, I can assume that Cluster 1 is approximately modern art, Cluster 2 is classical art, and Cluster 3 is nonhuman art. This may provide us evidence that there are certain innate traits within these artistic styles that cultivate certain emotions and sensations.





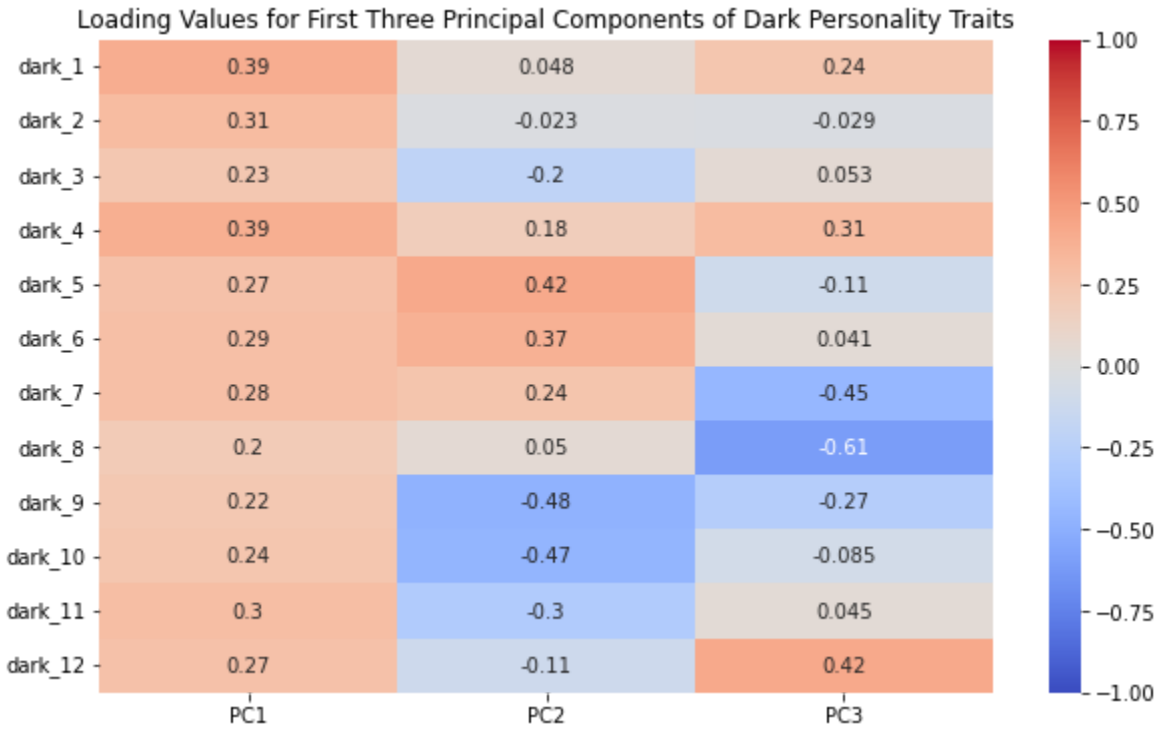
8. **Considering only the first principal component of the self-image ratings as inputs to a regression model – how well can you predict art preference ratings from that factor alone?**

I began by creating a df with artistic preference rating columns and the self image rating columns. Afterwards, I checked for NaN values in which there were varying numbers of NaN values in some of the self image columns. To deal with this, I got rid of any row with NaN values. Once the data was cleaned, I created new variables that separated the artistic preference ratings from the self image ratings as I would need to treat these dataframes differently. To run a PCA on self image ratings, I first normalized the ratings by making them into Z-Scores. I could've used a different method but I had previous code doing PCA and decided this would be the most convenient without compromising the sanctity of the data. After transforming the data with the PCA, I ran all three regression models but since they were very similar in RMSE, I kept the regular regression. This model has a RMSE of 1.4819. As such, this shows that perhaps the self image ratings were not the best predictors of art ratings. This may imply that art has a universal effect on the viewers, regardless of how they feel about themselves.

9. **Consider the first 3 principal components of the "dark personality" traits – use these as inputs to a regression model to predict art preference ratings. Which of these components significantly predict art preference ratings? Comment on the likely identity of these factors (e.g. narcissism, manipulativeness, callousness, etc.).**

Similar to Question 8, I created a df with artistic preference ratings and dark personality trait ratings. I checked for NaN values, which there were, and dropped them. I normalized the dark personality ratings and then continued on to do a PCA to the third component. I then created a heatmap of the loadings to check the effect of each question on the ratings. According to my heatmap, in the first component, all the questions had a similar positive impact, with the first and fourth question having the greatest impact with values of .39. In the second component, the results varied a little more, with question 5 and 6 being positively impacting the ratings, while questions 9 and 10 impacted the ratings negatively. In the Third Component, questions 7 and 8 had the largest absolute impact although negative, and question 12 had a positive impact. This

implies that people with darker traits seem to be more narcissistic and manipulative as they are more worried about how to manipulate others for their own benefit. They don;t seem to be concerned too much with how others view them as the heatmap suggests those questions have a strong negative impact, but want to just be better off overall. With that being said, our regression model was not the best predictor of artistic preference ratings. As I ran both a regular regression and a lasso regression, the RMSE was similar for both, with the former being 1.5112 and the latter being 1.5249.



10. Can you determine the political orientation of the users (to simplify things and avoid gross class imbalance issues, you can consider just 2 classes: “left” (progressive & liberal) vs. “nonleft” (everyone else)) from all the other information available, using any classification model of your choice? Make sure to comment on the classification quality of this model.

To begin this question, I created a copy of my original data df. I then checked for NaN values, and since there were some in the demographic values, I decided to drop those rows. I then had to binarize the political column as originally it has very specific information on political affiliation. However, we want to simplify this. I decided to set any value with 1 or 2 as 0 representing the left spectrum, and then anything else 1 for the non left. To make it more convenient, I moved the new political column to the end, and separated the data from the political data and the rest of the data. I decided to use a Random Forest model as it is robust to overfitting and will be more accurate than other models in this situation. I set my test set to be 30% as I wanted there to be more data to work with. The overall accuracy of the model was not that great, being able to

predict correctly only 58% of the time, however, it was better at predicting the left versus the non left. This is expected as we are combining so many types of political affiliations into the non left.

	precision	recall	f1-score	support
0	0.54	0.93	0.68	41
1	0.77	0.24	0.36	42
accuracy			0.58	83
macro avg	0.66	0.58	0.52	83
weighted avg	0.66	0.58	0.52	83