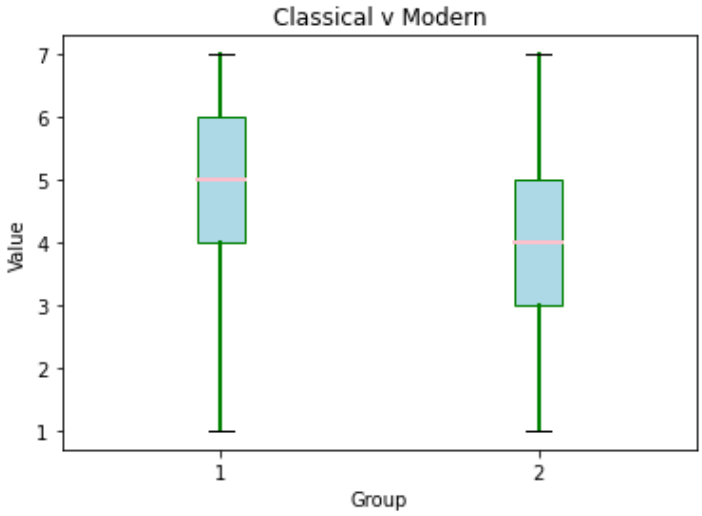
Kendall Brogle

Capstone Project Report

I handled pre processing on this data differently for each question, so as to best address each question's concerns when it comes to missing data. For the first few questions that required energy and preference ratings, I did not adjust the data because there were no missing values in these columns, and I did not want to reduce the quantity of data points where I did not have to. In further questions, I handled missing values by performing row wise removal of the specific chunk of data I was using. I did it this way to prevent row wise removal of every row that had any NA value in it. As a result of this method I also ended up reducing some of the partnering data for specific regressions and other analysis that required equal length data.

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

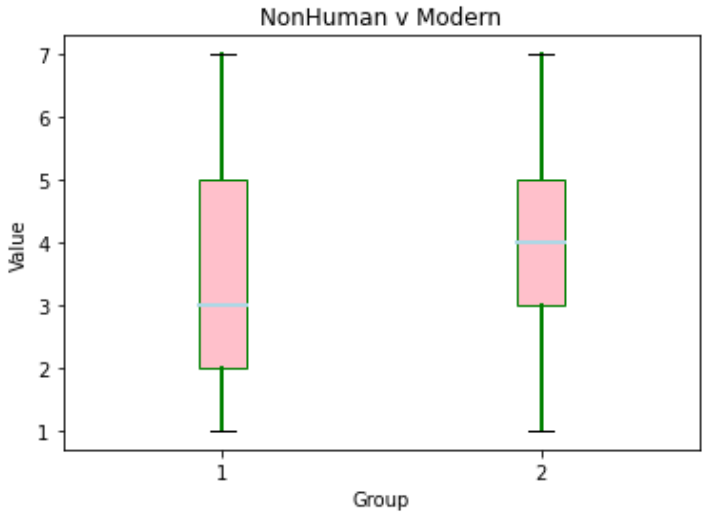
Overall, classical art is more well liked than modern art. Classical paintings had a median value of 5 and modern paintings had a median value of 4. Based on a Mann Whitney U test with a p value of 3.18 e-97 I rejected the null hypothesis with an alpha value of 0.05. I performed this test by splicing and flattening the data and utilizing scipy's stats function. I represented this data with a box plot and whisker figure as seen below, where group 1 represents classical paintings and group 2 represents modern paintings.



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

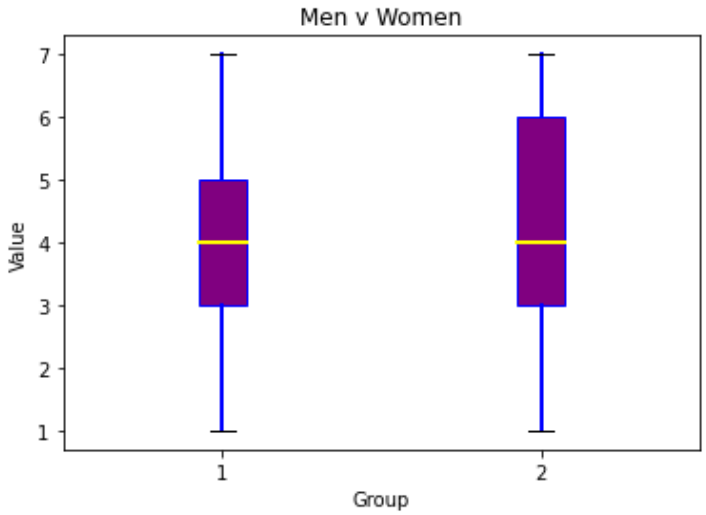
There is a statistically significant difference between the ratings for modern vs non human art; modern art was rated significantly higher. To establish this I performed another Mann Whitney U test based on the ordinal data of these two groups and found a p value of 8.74 e-264 and rejected the null hypothesis using an alpha of 0.05. I represented this data with a box plot and whisker

figure as seen below, where group 1 represents nonhuman paintings and group 2 represents modern paintings. The median preference rating of nonhuman paintings was 3.



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

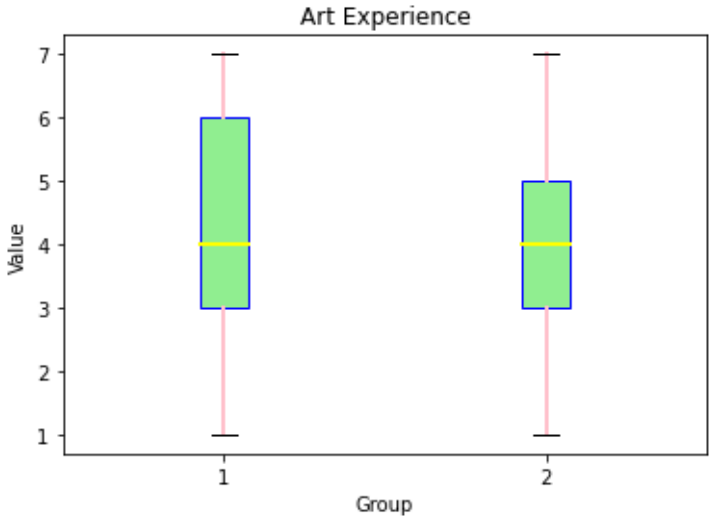
Women do not give significantly higher art preference ratings than men. I determined this by performing another Mann Whitney U test based on the ordinal data of two groups. I found a p value of 0.27 and when compared with an alpha level of 0.05 I accepted the null hypothesis. I represented this data with a box plot and whisker figure as seen below, where group 1 represents men's ratings and group 2 represents women's ratings. Both groups had a median rating of 4.



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

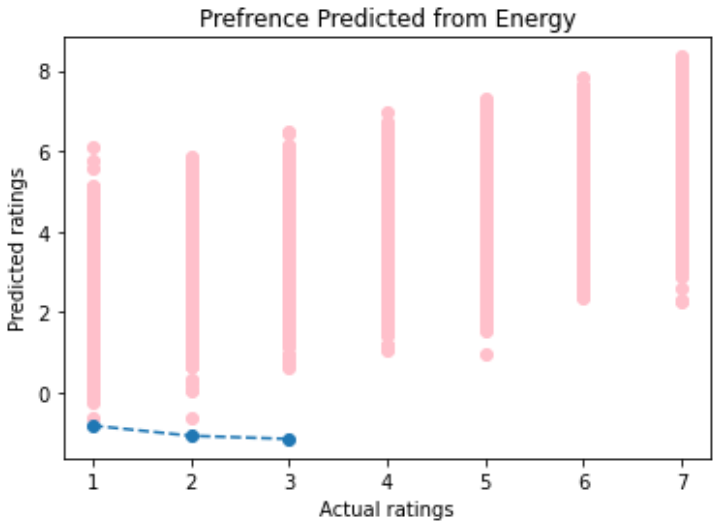
There is a statistically significant difference in the preference ratings of users with some art background versus those with none. I handled this question by sorting the data into two data sets,

one with people who recorded they had zero art background with a zero value in column 219, and another with those who had put any other value indicating any art education above none. This way of separating the data handled any NA values in itself, as no data that did not have a value of 0, 1, or 2 in it could have been added to my data frame as an NA value would not fall into one of these three categories. I then converted this data to a numpy array, flattened, and performed a Mann Whitney U test, based on the data being ordinal. I found a p value of 1.01e-08, which compared to an alpha level of 0.05 led me to reject the null hypothesis. I represented this data with a box plot and whisker figure as seen below, where group 1 represents people with no art background and group 2 represents those with some art background.



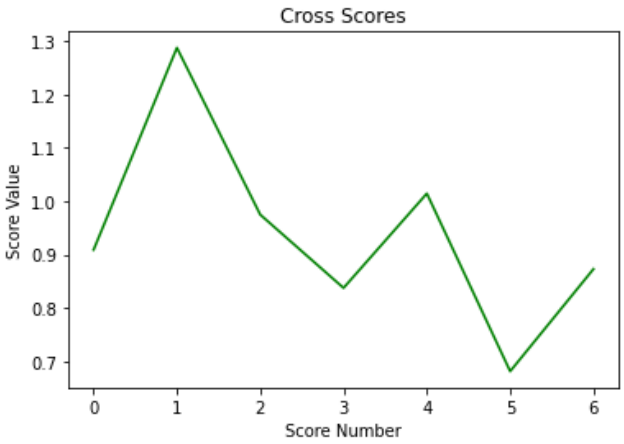
5) **Build a regression model to predict art preference ratings from energy ratings only. Make sure to use cross-validation methods to avoid overfitting and characterize how well your model predicts art preference ratings.**

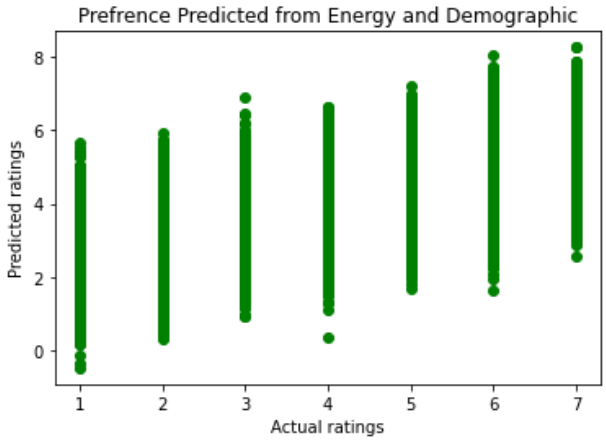
To complete this problem I performed a simple linear regression. To avoid overfitting I utilized sklearn's cross validation score feature. I computed the mean and standard deviation of the cross validation scores with K folds equal to 3. With this I found a mean absolute score of 0.92 and a standard deviation of 0.05. With this model I calculated a R squared value of 0.39 and a RMSE value of 1.15. Overall, this model can explain some, around 40%, of the variance in preference ratings. I represented this with a scatter plot of predicted versus actual preference ratings, and overlaid the cross validation scores.



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.**

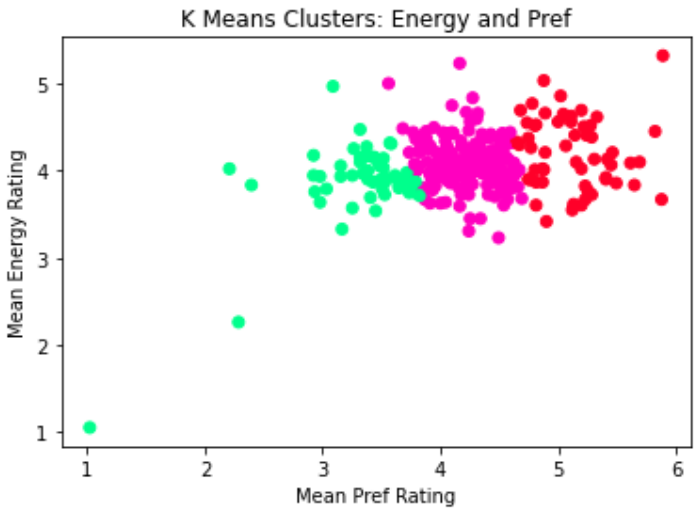
To answer this question I did another simple linear regression and used cross validation scores with a KFolds of 7 splits to address overfitting. This is the first question I explicitly handled NA values. I performed a row removal of NA values of the concatenated energy and demographic information. I then spliced the preference ratings to match the length of the predictive data. I found a mean absolute score of 0.94 and an average RMSE of 0.97 and a R squared value of 0.36. This means the model accounts for some of the variance in preference ratings but only 37% and predicts worse than the energy ratings only model. To represent this data I plotted the cross validation scores as well as a scatter plot of the predicted versus actual ratings.





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?**

Using K means clustering I identified 3 clusters in the 2D space of average preference ratings versus average energy ratings. I found this by manually calculating the silhouette score of two through six clusters and identified that three clusters was the best with a silhouette score of 0.41.I represented this with a scatter plot with the three clusters color coded. I am presuming that the central cluster of data is most closely related to modern art because the overall mean of all modern art preference ratings is 4.26 which is close to the centroid of the central cluster. The left cluster can be loosely associated with non-human art, as it was rated significantly lower, while the right can be loosely associated with classical art which was rated higher than modern art.

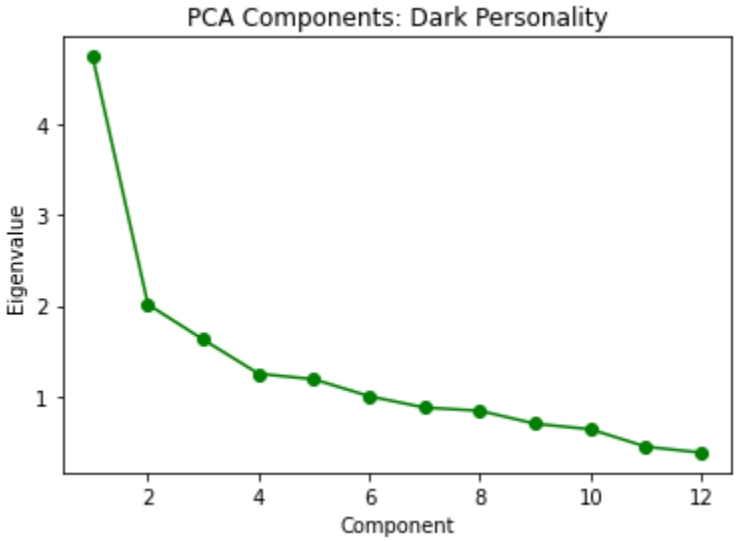


8) **Considering only the first principal component of the self-image ratings as inputs to a regression model – how well can you predict art preference ratings from that factor alone?**

When only considering the first principal component of the self image ratings in a regression model the preference rating was not predicted very well. I found this result by performing a one component PCA and then doing a train test split with a 80, 20 split. I then used a linear regression model. I calculated the absolute R squared value of this data at 0.02, meaning the model could account for 2% of the variation in the outcome. I also found an RMSE of 1.48. I represented this data with a loading graph to illustrate the strength between the original variables and the principal components. I found that the most significant component to be “All in all, I am inclined to feel that I am a failure.”

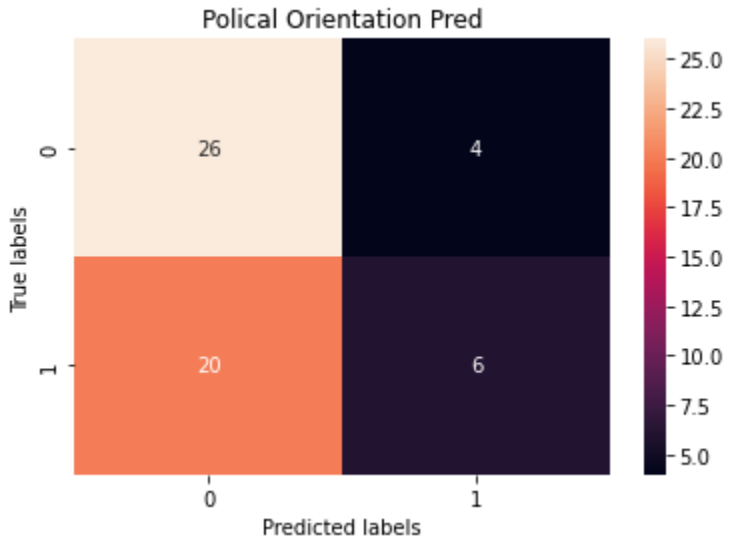
9) **Consider the first 3 principal components of the "dark personality" traits – use these as inputs to a regression model to predict art preference ratings. Which of these components significantly predict art preference ratings? Comment on the likely identity of these factors (e.g. narcissism, manipulativeness,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

To answer this question I set up a PCA, fit it to my data, and calculated the explained variance (eigen values) for the length of the data. I then plotted this, which is shown below, to identify the elbow curve as to the best number of components for this PCA. I noted that two was the elbow value and I think performed a PCA with two components. I used a k fold with ten splits, did a 80, 20 test train split and then did a linear regression. This model gave me an absolute R squared value of 0.03 and a RMSE of 1.45. I also calculated the loading score of each of the factors in this model to see what features were most significant. I found that the top three, in order, were; "I tend to manipulate others to get my way", "I tend to lack remorse", and "I tend to be cynical". The likely identity of these factors are, manipulativeness, resourcefulness, and cynicism.



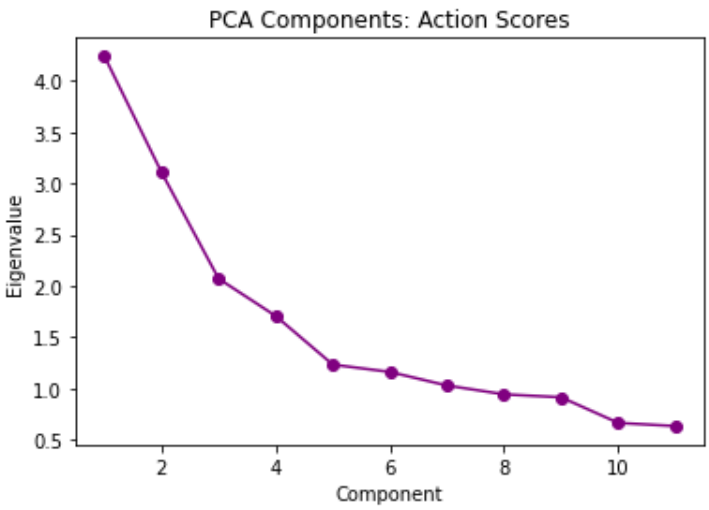
10) **Can you determine the political orientation of the users (to simplify things and avoid gross class imbalance issues, you can consider just 2 classes: "left" (progressive & liberal) vs. "nonleft" (everyone else)) from all the other information available, using any classification model of your choice?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

To address this question I sorted the political orientation data into two bins, left and non left, I then used all of the remaining data and did an 80, 20 test train split and used these values to do a random forest classifier with 100 estimators. With this I found an accuracy rating of around 57%, which I represented with a confusion matrix below.



Extra Credit: Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions.

For extra credit I examined the components of Action Preferences and how well they predicted the preference ratings of the paintings. Through graphing the explained variance (eigen values) of the components I found that there was an elbow at five components, but it was not as sharp as the Dark personality traits component analysis. I then used a train test split of 80, 20 and used a linear regression model to predict values. I found an absolute R squared value of 0.04 indicating that the action preferences do not account for very much if any of the variance in the outcome. The top five components were "I like to take walks in the forest" and "I like to play video games", " I like to do yoga", "I like to ski" and "I like amusement parks". Overall, action preferences are not a great predictor of preference ratings for art, and action preferences are not reduced into components as efficiently as dark personality traits.


PCA Components: Action Scores
Eigenvalue
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
2
4
6
8
10
Component